## Computational Tools for CNV Detection Using Probe-level Analysis of Affymetrix SNP Arrays

### Application to the Study of CNVs in Follicular Lymphoma

by

Noushin R. Farnoud

M.A.Sc. Electrical Engineering, Ryerson University, 2004 B.Sc. Computer Software Engineering, Shahid Beheshty University, 2000

## A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

**Doctor of Philosophy** 

in

### THE FACULTY OF GRADUATE STUDIES

(Genetics)

The University Of British Columbia

(Vancouver)

August 2012

© Noushin R. Farnoud, 2012

# Abstract

Copy number variants (CNVs) account for both variations among normal individuals and pathogenic variations. The introduction of DNA microarrays had a significant impact on the resolution of detectable CNVs and yielded a new perspective on the submicroscopic CNVs. Oligonucleotide microarrays, such as Affymetrix SNP arrays, have been commonly used for genome-wide CNV analysis. Despite the improvements in the technology, a major concern of using microarrays is how a putative CNV is defined. A disadvantage of oligonucleotide arrays is the poor signal-to-noise ratio of the data that leads to considerable variation in reported intensity readouts. Such variation will lead to false positive and false negative results, regardless of how the data are analysed. The most common approach to circumvent this problem is looking for abrupt ratio intensity shifts in several consecutive markers (e.g., SNP probes). However this approach reduces the overall resolution and mitigates the sensitivity of detecting CNVs with fewer probes. This limitation emphasizes the importance of designing methods that can identify noisy readouts at the probe-level.

The main goals of this work were to study the scale of the variability in Affymetrix SNP arrays and to develop computational tools that can improve the resolution of CNV detection. By using simulated data, it was shown that the proposed method improved the accuracy and precision of detecting CNVs with fewer probes compared to standard methods. This approach was also applied to tumor/normal pairs from 25 follicular lymphoma patients and 286 candidate CNVs were found, from which 261 (91.2%) were also seen by other array-based method(s). Importantly, from 32

novel deletions, undetected by other array-based methods, at least 15 (47%) were real based on sequence-based validation. An example of a novel discovery was a partial deletion of the extracellular domain of the KIT proto-oncogene that may lead to constitutive activation of this gene. Gain of function mutations of KIT has been previously reported in several other hematologic cancers through other mechanisms such as point mutations.

In conclusion, CNV discovery contributes to our understanding of complex diseases and the methods presented here should provide means for better detection of CNVs and their interpretation.

# Preface

Portions of the statistical methods used to analyse the copy number data in Chapters 3 and 4 have been described: T. J. Pugh, A. D. Delaney, N. Farnoud, S. Flibotte, M. Griffith, H. I. Li, H. Qian, P. Farinha, R. D. Gascoyne and M. A. Marra, "Impact of whole genome amplification on analysis of copy number variants". Nucleic Acids Res. 2008; 36, 13:e80. I was involved in the computational aspects of this study and wrote the relevant sections of the paper. Specifically, I performed all computational analysis described in 'Sequence analysis of copy number data and investigations of how various factors, such as the GC content and repetitive sequences influence this analysis. I also generated Figure 4 of the paper.

The tools described in Chapters 3 and 4 were also used in the following manuscripts and conference proceedings: J. M. Friedman, A. Baross, A. D. Delaney, A. Ally, L. Arbour, J. Asano, D. K. Bailey, S. Barber, P. Birch, M. Brown-John, M. Cao, S. Chan, D. L. Charest, N. Farnoud, N. Fernandes, S. Flibotte, A. Go, W. T. Gibson, R. A. Holt, S. J. M. Jones, G. C. Kennedy, M. Krzywinski, S. Langlois, H. I. Li, B. C. McGillivray, T. Nayar, T. J. Pugh, E. Rajcan-Separovic, J. E. Schein, A. Schnerch, A. Siddiqui, M. I. Van Allen, G. Wilson, S.-L. Yong, F. Zahir, P. Eydoux, and M. A. Marra. "Oligonucleotide Microarray Analysis of Genomic Imbalance in Children with Mental Retardation". American Journal of Human Genetics. September 2006; 79(3): 500513. M. Rahmani, M. Earp, P. Pannu, N. Farnoud, J. Wu, L. Akhabir, J. Halaschek-Wiener, B. Munt, C. Thompson, S. Mitropanopoulos, D. Craig, P. Par, B. McManus, and A. Brooks-Wilson. "Identification of novel risk loci for calcific aortic valve stenosis on chromosome 1 by a genome-wide scan of 1,000,000 single nucleotide polymorphisms". Proceedings of National Research Forum for Young Investigators in Circulatory and Respiratory Health, May 2009. N. Farnoud, S. Chan, S. Flibotte, A. Delaney, J.M. Friedman and M. A. Marra. "DLOH: A novel bioinformatics tool for detection of copy-number deletions using LOH data". Proceedings of Advances in Genome Biology and Technology conference, February 2008.

A manuscript based on Chapter 3 is in preparation: Noushin Farnoud, Stephane Flibotte, Inanc Birol, P. Eydoux, Robert A. Holt, J. M. Friedman and Marco A. Marra. "Detecting DNA copynumber variations based on probe-level analysis of Affymetrix SNP array data". This manuscript reports the details of OPAS copy number analysis approach explained in Chapter 3. I developed all the methods described in the manuscript and made relevant figures and tables with input from my supervisor M.A.M. and members of my supervisory committee J.M.F and R.A.H. J.M.F. provided samples and P.E. performed FISH experiments. Figures 3.1, 3.11, 3.15, 3.9 of Chapter 3 and Table I.1 are directly taken from this manuscript draft. Figures 3.6, 3.7, 3.12 of Chapter 3 are from the supplementary material. S.F. and I.B. provided guidance for statistical analysis. M.A.M, R.A.H., P. E. and J.M.F. provided guidance for biological interpretation of the CNV results. M.A.M. helped in data interpretation and provided supervisory support, including manuscript revision.

A version of Chapter 4 is in preparation: Noushin Farnoud, Andy J. Mungall, Susana Ben-Neriah, Andy Chu, Martin Krzywinski, Inanc Birol, Jacqueline Schein, Randy Gascoyne and Marco A. Marra. "Integrated genome-wide DNA copy number and expression analysis of follicular lymphoma genomes". The samples and SNP array data were provided by R.G. at the Centre for Lymphoid Cancers. The fluorescent in situ hybridization (FISH) validation results in this manuscript were performed by S.B.-N. at Clinical Cancer Genetics Laboratory at the BC Cancer Agency. The DNA fingerprint profiling of the samples was conducted by J.S. and sequencing validation experiments were performed by A.J.M. at the Genome Sciences Center. A.C. implemented the Tumordb database and also helped to obtain the data from multiple platforms used in the integrated analysis. For this study, I performed all the array copy number analysis of follicular lymphoma samples (detailed in Sections 4.3.3-4.3.6). I also implemented several additional scripts to summarize the CNV findings in large-scale studies (automated using Matlab and PERL), such as scripts to obtain recurrently affected regions across multiple patients (used to generate Figures 4.9 and 4.17; both taken from the manuscript draft), and a script to automatically generate UCSC tracks for representation of the CNV results with comprehensive graphical details (such as shown in Figure 4.25c). I also made the figures, tables, interpreted the results and wrote the sections of the manuscript (a version of Sections 4.2-4.4 of the thesis), with the exception of the FISH images that were provided by S.B.-N. (Figures 4.3b, 4.28b, 4.32). I also performed the pathway analysis using Ingenuity software (Figure 4.21 from the manuscript). M.A.M. supervised all aspects of this study.

I also conducted an integrative analysis of copy number and expression data. The integrated analysis was performed using an available software package (DR-Integrator) and a script that I designed to integrate copy-number/expression data results a single sample. The latter method is not included in the thesis since this work was done following thesis preparation. Under the supervision of M.A.M., I combined the data from different platforms and interpreted the results. M.A.M, A.J.M. and J.S. provided guidance for biological interpretation of the CNV results and R.G. helped in conception of integrating gene expression and copy number results. A.J.M. implemented the experiments for Illumina sequence validation of several CNVs (Section 4.3.8.3, adapted from the manuscript), and was also involved in designing a computational model to investigate a hypothetical fusion between *HVCN1* and *PPTC7* genes (Section 4.3.8.1 and Figure 4.29, also from the manuscript). Figures 4.7, 4.10, 4.17, 4.21, 4.29, 4.34 and Table 4.2, 4.4, 4.7 are taken from the manuscript draft and Figures 4.9, 4.15, 4.31, 4.32, 4.33, J.1 are from the supplementary material.

# **Table of Contents**

21	• • • • • •	• • • •	••	••	••	••	••	•	•	••	••	••	•	••	•	••	•	•	•	•	•••	•	ii
	•••••	••••	••	••	••	••	••	•	•	••	••	••	•	••	•		•	•	•	•		•	iv
f Conte	nts	••••	••	••	••	••	••	•	•	••	••	••	•	••	•		•	•	•	•	••	•	vii
Tables	••••	••••	••	••	••	••	••	•	•	••	••	••	•	••	•		•	•	•	•		•	xiv
Figures	••••	••••	••	••	••	••	••	•	•	••	••	••	•	••	•	••	•	•	•	•		•	xvi
vledgme	ents		••	••	••	••	••	•	•	••	•••	••	•	••	•	• •	•	•	•	•		•	xxiii
hods a	nd Strateg	gies Fo	r A	naly	sin	g C	ору	v N	um	ber	· Va	ria	ntic	on	Us	sing	g I	DN	JA	N	/li	-	
irrays			••	••	••	••	••	•	•	••	••		•	••	•		•	•	•	•		•	1
Introd	uction																						1
Techn						• •	• •	•	•	•••								•					-
Ittelli	ologies .					· ·	· ·	•	•••	· ·	•••							•		•			3
1.2.1	ologies . First Gei	····	· ·	 chni	· ·	••• ••• s••	· ·	•••	•	· ·	•••	 	•	 	•		•	•	•	•	•••		3
1.2.1	ologies . First Ger 1.2.1.1	neration	· · n Teo moso	 chni ome	 ques Bar	s . ndin	•••• •••	• •	•	· · · ·	· · ·	· ·	•	· ·	•	· ·	•	•	•	•	•••		3 3 3
1.2.1	ologies . First Ger 1.2.1.1 1.2.1.2	neration Chro Spect	 n Teo moso tral 1	 chni ome Kary	ques Bai yoty	· · · s . ndin ping	 g.	SKY	() a	   und	   	  		  Fl		  res		nt	in	S	  itu		3 3 3
1.2.1	ologies . First Ger 1.2.1.1 1.2.1.2	neration Chro Spect Hybr	 n Teo moso tral ∃ idiza	 chni ome Kary ation	ques Bai yoty n (M	s . ndin ping	· · · · · g · g (S	SKY	() analy	· · · · · · · · ·	   	  		  Fli	uoi	•••• ••• res		· · nt	· · · in	S	  	1	3 3 3 4
1.2.1 1.2.2	ologies . First Ger 1.2.1.1 1.2.1.2 Second (	neration Chro Spect Hybr Generat	 m Teo moso tral 1 idiza tion	 chni ome Kary ation Higl	ques Bai yoty n (M h-Re	s . ndin ping -FIS	· · · · · · g · (S SH) utic	SKY Au	() analy	   und vsis	   	•••• •••• ••••	· · · · ·	   	uo:	  				ı S	  	· · 1	3 3 3 4 5
	f Conte Tables Figures vledgmo thods an arrays Introd	f Contents Tables Figures vledgments thods and Strateg arrays Introduction	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents	f Contents

			1.2.2.2	DNA Ol	igonucleotide Arrays	6
			1.2.2.3	Genotyp	ing Arrays	7
			1	.2.2.3.1	Affymetrix SNP Arrays	8
			1	.2.2.3.2	Illumina SNP arrays	9
			1	.2.2.3.3	Application of SNP arrays for CNV Detection	9
		1.2.3	Third Ge	eneration I	High-Resolution Techniques	10
	1.3	Metho	ds to Disco	over CNV	s Using SNP-array Data	12
		1.3.1	Normaliz	zation Me	hods	12
		1.3.2	CNV Ca	lling Meth	ods	13
	1.4	Limita	tions of C	urrent CN	V Detection Algorithms for SNP Data Analysis	17
	1.5	Thesis	Objective	s and Hyp	othesis	18
	1.6	Chapte	er Summar	ies		20
	1.7	Figure	s and Tabl	es		26
2	Ana	lysing V	ariability	in Micro	array Data	30
	2.1	Introdu	uction			30
		2.1.1	Variabili	ty in Micr	oarray Data	31
		2.1.2	Biologica	al Variabil	ity	32
		2.1.3	Technica	l Variabili	ty	32
	2.2	Metho	ds for Mea	asuring an	d Quantifying Microarray Variability	34
		2.2.1	Quantify	ing Techn	ical Variability in Affymetrix SNP Arrays	35
			2.2.1.1	Log-tran	sformation	35
			2.2.1.2	Coefficie	ent of Variation (CV)	35
		2.2.2	A Link B	etween th	e CV and the Probability of Observing k-fold Disparities	
			Between	Replicate	Measurements	37

		2.3.1	Affymetr	ix 10K Repl	cate Experiment	38
		2.3.2	Quantifyi	ng Technica	l Variability in Affymetrix SNP Arrays	39
		2.3.3	Assessing	chip Varia	pility in the Replicate Dataset (10K)	41
		2.3.4	Assessing	, Labeling V	ariability in the Replicate Dataset	41
		2.3.5	Analysing	g the Relatio	nship Between Oligo-level and SNP-level Variabilities	43
			2.3.5.1	The Accept	able Range of Variability Between Replicate Oligos .	43
			2.3.5.2	Finding a L	ink Between Oligo-level CV to the Changes in SNP-	
				level LR Va	lues	45
			2.3.5.3	Evaluating	the Impact of Oligo-level Variability on the Extent	
				and Freque	ncy of Noisy SNPs in Replicate Experiments	48
	2.4	Conclu	sions			49
		E				50
	2.5	Figures	s and Table	28		52
3	2.5 Algo	orithm f	or Oligon	s .cleotide Pr	obe-level Analysis of Signal Intensities (OPAS)	52 66
3	2.5 Algo 3.1	orithm f	or Oligon	s	obe-level Analysis of Signal Intensities (OPAS)	<b>66</b>
3	2.5 Algo 3.1 3.2	orithm f Introdu Methoo	or Oligona action	s	obe-level Analysis of Signal Intensities (OPAS)	<b>66</b> 66 69
3	<ul><li>2.5</li><li>Algo</li><li>3.1</li><li>3.2</li></ul>	Prigures prithm f Introdu Methoo 3.2.1	or Oligona action ds Algorithm	s	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>66</li> <li>69</li> <li>69</li> </ul>
3	2.5 Algo 3.1 3.2	Prigures prithm f Introdu Methoo 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre-	Icleotide Pr	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>52</li> <li>66</li> <li>69</li> <li>69</li> <li>70</li> </ul>
3	2.5 Algo 3.1 3.2	Angures Prithm f Introdu Method 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre- 3.2.2.1	Icleotide Pr	obe-level Analysis of Signal Intensities (OPAS)       .	<ul> <li>66</li> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> </ul>
3	2.5 Algo 3.1 3.2	Anticological Antices Programs of the product of th	or Oligona action ds Algorithm SNP Pre- 3.2.2.1 3.2.2.2	Incleotide Pr Incleotide Pr In	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> <li>70</li> </ul>
3	2.5 Algo 3.1 3.2	Prigures prithm f Introdu Methoo 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre- 3.2.2.1 3.2.2.2 3.	Incleotide Pr Incleotide Pr In	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> <li>70</li> <li>71</li> </ul>
3	2.5 Algo 3.1 3.2	Prigures prithm f Introdu Methoo 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre- 3.2.2.1 3.2.2.2 3. 3. 3.	In Design And Andrew Clustering 1 2.2.2.1 C 2.2.2.2 C	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> <li>70</li> <li>71</li> <li>72</li> </ul>
3	2.5 Algo 3.1 3.2	orithm f Introdu Methoo 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre- 3.2.2.1 3.2.2.2 3. 3.2.2.2 3. 3.2.2.3	Incleotide Presented in the second se	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> <li>70</li> <li>71</li> <li>72</li> <li>72</li> </ul>
3	2.5 Algo 3.1 3.2	orithm f Introdu Method 3.2.1 3.2.2	or Oligona action ds Algorithm SNP Pre- 3.2.2.1 3.2.2.2 3. 3.2.2.2 3. 3.2.2.3 3.2.2.4	Incleotide Presented in the second se	obe-level Analysis of Signal Intensities (OPAS)	<ul> <li>66</li> <li>69</li> <li>69</li> <li>70</li> <li>70</li> <li>70</li> <li>71</li> <li>72</li> <li>72</li> <li>73</li> </ul>

	3.2.3	Alternati	ve Approach for SNP Pre-processing Based on Naive Bayes Clas-	
		sification		74
	3.2.4	Post-proc	cessing and CNV Calling	75
		3.2.4.1	PCR Fragment Length Normalization	75
		3.2.4.2	Circular Binary Segmentation (CBS) Algorithm	75
	3.2.5	OPAS Vi	sualization and Other Features	76
	3.2.6	Simulate	d Data for Comparative Analysis of CNV Calling Algorithms	76
	3.2.7	Analysin	g the Effect of Noise on CNV Calling Performance	77
	3.2.8	Compara	tive Analysis of OPAS and Circular Binary Segmentation	78
3.3	Results	5		78
	3.3.1	Patterns	of LR Intensity Fluctuations in SNP Array Data	78
	3.3.2	Analysin	g the Impact of the Size of the Reference Set on the Estimated	
		SNP Sign	nals	79
		3.3.2.1	Impact of the Size of the Reference Set on the Estimated LR Values	79
		3.3.2.2	Impact of the Size of the Reference Set on the Number of CNV-	
			affirmative Oligos	80
	3.3.3	Results o	of OPAS Pre-processing Phase	81
		3.3.3.1	Clustering PM Oligos in SNP Probe-sets	81
		3.3.3.2	Performing Oligonucleotide Probe-level Analysis of the SNP Ar-	
			ray Data	81
		3.3.3.3	SNPs Classification and LR Estimation	82
		3.3.3.4	Comparison of OPAS Pre-processing and Naive Bayes Classifi-	
			cation	82
		3.3.3.5	An Example of the Impact of SNP Pre-processing on Improving	
			CNV Data Quality	84

		3.3.4	Results of OPAS Post-processing Phase    85
		3.3.5	Assessing the Effect of Noise on CNV Calling Performance
		3.3.6	Comparative Analysis of CNV-calling Algorithms
		3.3.7	Comparing OPAS and CBS Accuracy
	3.4	Conclu	asions
	3.5	Figure	s and Tables
4	Ana	lysing (	CNVs in Follicular Lymphoma Genomes
	4.1	Introdu	uction
	4.2	Materi	als and Methods
		4.2.1	Samples and Cytogenetic Analysis
		4.2.2	BAC Arrays and SNP Arrays
		4.2.3	Fingerprint Profiling (FPP)
		4.2.4	Ingenuity Pathway Analysis Software
	4.3	Result	s
		4.3.1	Magnitude of Copy Number Changes in FL Genomes
		4.3.2	Spectrum of Somatic CNVs in FL Genomes
		4.3.3	Category 1: CNVs Affecting Whole-chromosomes or Chromosome Arms
			(WCA) 123
			4.3.3.1 Recurrent WCA CNVs
		4.3.4	Category 2: CNVs Affecting the Distal Ends of Chromosomes
			4.3.4.1 Functional Analysis of the Genes Affected by Distal CNVs 126
		4.3.5	Category 3: Other CNVs
		4.3.6	Candidate Focal CNVs in FL Genomes and Functional Analysis of the
			Affected Genes
		4.3.7	Comparison of OPAS Generated CNVs with Other Methods

		4.3.8	Examples of Sequence Validated Novel (OPAS-exclusive) CNV Findings .	132
			4.3.8.1 <i>HVCN1</i>	132
			4.3.8.2 <i>CDKN2A</i>	134
			4.3.8.3 <i>KIT</i>	135
	4.4	Conclu	sions	136
	4.5	Figure	and Tables	138
5	Con	clusions	and Future Directions	187
	5.1	Summa	ıry	187
	5.2	Signifi	ance and Contribution to Field of Study	194
	5.3	Potenti	al Applications and Future Directions	196
Bi	bliogr	aphy .		198
A	Арр	endix A	Probability Distribution Function and Cumulative Distribution Function	245
B	Арр	endix B	Normal and Standard Normal Distributions	248
С	Арр	endix C	<b>Proof of the Relationship Between</b> $p(k)$ <b>and CV</b>	252
D	Арр	endix I	: Estimating SNP-array Reproducibility Using Analysis of Variance	
	(AN	OVA).	••••••••••••••••••••••••••••••••••	255
E	Арр	endix E	Description of Boxplots (in the thesis)	263
F	Арр	endix F	Comparative Analysis of the SNP Array Normalization Techniques	265
	F.1	Introdu	ction	265
	F.2	Metho	and Results	265
G	Арр	endix G	: Description of QDA	274

Η	Appendix H: Measures of Predicting the Accuracy
Ι	Appendix I: List of Validated CNVs in 146 MR Patients
J	Appendix J: Relationship Between Hybridization Intensity Noise and the Number         of Predicted CNVs in FL Genomes       281
K	Appendix K: Description of FPP Events
L	Appendix L: Supplementary Material for Chapter 4

# **List of Tables**

Table 1.1	Aneuploidies and large-scale CNVs associated with human disease 2	8
Table 1.2	Comparison of array specifications in 4 generations of Affymetrix SNP array 2	8
Table 1.3	A partial list of Affymetrix SNP array data (raw .CEL files) that are publicly	
	available	9
Table 2.1	The relationship between CV and predicted replicate oligos that are $\geq k$ -fold	
	different	4
Table 2.2	The impact of oligo-level variability across replicate measurements on the ex-	
	pected SNP-level variability	5
Table 3.1	Three sets of hypothesis tests used in null-likelihood phase	2
Table 3.2	The impact of the size of reference sets on the estimated SNP signals 11	3
Table 3.3	Comparison of the performance of Naive Bayes and OPAS QDA-based SNP	
	pre-processing methods	4
Table 3.4	Comparing CBS and OPAS performance in detecting known CNVs with respect	
	to the number of SNP probes in the CNV regions	5
Table 4.1	Spectrum of LR deviation of all estimated DNA regions from 25 FL genomes . 17	8
Table 4.2	Summary of candidate somatic copy number changes in the FL dataset 17	9
Table 4.3	Frequency of candidate distal CNVs (category 2) in each FL chromosome 18	0

Table 4.4	Summary of all candidate somatic focal CNVs ( $\leq 150$ kb) that affect at least 1
	gene in an FL patient
Table 4.5	List of 19 new candidate OPAS amplifications that were not previously detected
	by SNP data analysis
Table 4.6	List of OPAS-exclusive deletions that overlap with FPP or Illumina sequence
	validated events
Table 4.7	Partial list of some important genes that have been linked to cancer and their
	frequency across the FL dataset
Table D.1	Results of assessing chip and labeling variability in 8 samples based on two-way ANOVA analysis
Table F.1	List of CNVs and large-scale copy number alteration in 8 studied follicular lym-
	phoma patients
Table F.2	Comparing the results of normalization methods in 8 cancer samples (FL) 273
Table I.1	Validation of OPAS sensitivity in detecting previously known CNVs in 146 mental retardation patients
Table L.1	Ingenuity Pathway Analysis list of known genes that were affected by candidate facel CNV $a$ ( $< 150$ kb)
	$10car Civvs ( 130 KU) \dots 293$

# **List of Figures**

Figure 1.1	Karyotype and M-FISH chromosome analysis of a patient with follicular lym-	
	phoma	26
Figure 1.2	Structure of SNP probe sets in Affymetrix GeneChip <sup>®</sup> SNP arrays	27
Figure 2.1	Sources of variability in SNP microarray experiments for identification of copy	
	number variations (CNVs)	52
Figure 2.2	Common sources of microarray technical variability	53
Figure 2.3	Impact of log-transformation on the distribution of raw signal intensities from	
	Affymetrix SNP arrays	54
Figure 2.4	Schematic representation of Affymetrix 10K replicate study	55
Figure 2.5	Schematic representation of assessing technical variability in 10K SNP array	
	replicate study	56
Figure 2.6	Global normalization of log-transformed intensity readouts from 3 replicate	
	arrays	57
Figure 2.7	MA-plots of chip variability	58
Figure 2.8	Histograms of deviation (error) in replicate arrays	60
Figure 2.9	Estimated chip and labeling variability of the Affymetrix 10K replicate exper-	
	iment	61

Figure 2.10	Relationship between array variability and the probability of estimating oligos
	with k-fold difference in their intensity readout
Figure 3.1	Flowchart of the OPAS algorithm
Figure 3.2	Flowchart of the alternative approach for SNP pre-processing based on Naive
	Bayes classification
Figure 3.3	Schematic representation of OPAS input/output data
Figure 3.4	The Nsp signal of chromosome 14 of a follicular lymphoma patient that har-
	bors a deletion on 14q32.33 (~644 kb; 9 SNPs)
Figure 3.5	Example of oligo-level and SNP-level variability in SNP arrays (100K data) 98
Figure 3.6	Comparing the impact of the size of the reference set on estimated log2-ratio
	intensity readouts of SNP probe sets within a deleted region
Figure 3.7	Comparison of CDFs of all oligos in SUFU deleted region with 6 reference
	sets with varying sizes
Figure 3.8	Boxplots of the average number of CNV-affirmative PM oligos per SNP probe
	set $(\overline{\theta})$ , with respect to 3 reference sets with varying sizes
Figure 3.9	Fuzzy-Kmeans clustering of PM oligos in 5 SNPs within the SUFU deleted
	region
Figure 3.10	Schematic representation of oligo-clustering and likelihood estimation mod-
	ules of OPAS default pre-processing
Figure 3.11	Distribution of PM log2-ratio intensities before and after pre-processing (500K
	data)
Figure 3.12	The relationship between SNP log2-ratio intensities and PCR fragment length,
	before and after LOWESS normalization
Figure 3.13	Comparing estimated LR values of 567 SNPs in 19 validated regions of copy
	number loss based QDA and Naive Bayes classifications

Figure 3.14	Examples of OPAS detected CNVs in simulated signals with different magni-
	tudes of LR deviations
Figure 3.15	Comparing the accuracy and precision of CNV calling algorithms 109
Figure 3.16	Number of false negative deletion calls of the IGH locus, plotted against in-
	creasing noise of the simulated data
Figure 3.17	Results of comparing CBS and OPAS performance in detecting a known deletion 111
Figure 4.1	Example of an FPP event on chromosome 4q12 in FL patient 20 that is proved
	to be a deletion by Illumina sequencing
Figure 4.2	Distribution of LR intensity measurements of all regions across FL dataset 139
Figure 4.3	Deletion on 1p36 chromosomal region of ht-9 with slight signal deviation (LR
	= -0.13), validated by FISH
Figure 4.4	Probability Density Function (PDF) of z-scores from all regions with gain
	$(LR>0) \mbox{ or loss } (LR<0) \mbox{ of log2-ratio signal intensity (in FL dataset) } 141$
Figure 4.5	Boxplot of z-scores of all regions with loss or gain of log2-ratio signal intensi-
	ties (in FL dataset)
Figure 4.6	Candidate deletions on chromosome 4 of patient 29 with slight signal devia-
	tions but significant z-scores
Figure 4.7	Pie charts of the frequency of candidate amplifications and deletions in 25 FL
	patients
Figure 4.8	Examples of two real whole chromosome gains in an FL patient (ht-11) with
	slight log-ratio deviations from base-line (LR = $0.12$ and $0.13$ )
Figure 4.9	Frequency of WCA events per chromosome across all FL patients
Figure 4.10	Chromosome ideogram view of 48 WCA CNVs in the FL dataset 147
Figure 4.11	The only two WCA events that were not directly validated by cytogenetic anal-
	ysis (slight gains of chromosomes 7 and X in ht-29)

Figure 4.12	Example of a distal CNV on chromosome 22 of an FL patient (ht-22) 150
Figure 4.13	Two candidate amplifications on chromosome 5 of an FL patient (ht-12), in-
	cluding a distal copy number gain on 5 q-end
Figure 4.14	Candidate distal deletion on chromosome 1p36 (ht-7) with slight LR deviation
	but a significant z-score (LR = $-0.12$ ; z-score = $-1$ )
Figure 4.15	Recurrent distal deletion of chromosome 1p36 in the FL dataset
Figure 4.16	Most significantly associated gene networks with candidate distal CNVs of the
	FL dataset
Figure 4.17	Frequency of candidate CNVs in each chromosome among 25 FL patients 155
Figure 4.18	Validated focal deletion ( $\sim$ 38.7 kb) on 13q21.33, detected by OPAS and SMD
	but not aCGH results
Figure 4.19	Candidate focal deletions on chromosome 2 that potentially affect the first in-
	tron of <i>ERBB4</i> gene in two FL patients (OPAS-exclusive)
Figure 4.20	OPAS-exclusive candidate deletion on chromosome $3q13.33$ of patient $20 (\sim 97)$
	kb) that is adjacent to an FPP inversion event
Figure 4.21	The most significant gene network associated with OPAS candidate focal CNVs
	$(\leq 150 \text{ kb})$
Figure 4.22	Venn diagram comparing predicted copy number amplifications in FL samples,
	generated by 3 methods (OPAS, SMD and aCGH) 161
Figure 4.23	Candidate OPAS-exclusive focal amplification on 9p13.2 (~76 kb; 14 SNPs)
	that encompasses 2 exons of <i>PAX5</i> (OPAS-exclusive)
Figure 4.24	Venn diagram comparing predicted copy number deletions in FL samples, gen-
	erated by 3 methods (OPAS, SMD and aCGH) 163
Figure 4.25	Multiple OPAS candidate deleted regions on chromosome 10 of an FL patient
	(ht-19), that align with FPP complex events

Figure 4.26	Candidate OPAS-exclusive deletion on chromosome 14q32.33 of patient 14,
	mapping to an FPP 'coverage-gap' (~230 Kb; 4 SNPs)
Figure 4.27	Candidate OPAS-exclusive deletion on 15q11.2 (ht-21), mapping to a region
	with several 'multi fpp' events
Figure 4.28	Validated OPAS-exclusive focal deletion on chromosome 12 of patient 6 ( $\sim$ 143
	kb; 4 SNP probes), affecting 4 genes including <i>HVCN1</i> and <i>PPTC7</i> 169
Figure 4.29	Analysis of a putative fusion between HVCN1 and PPTC7 genes in FL patient
	6 as the result of a focal deletion (145,148 bp) on 12q24.11
Figure 4.30	Candidate ~1.4 Mb deletion on 12q24 in patient 9, encompassing the HVCN1
	gene
Figure 4.31	Sequence validated OPAS-exclusive focal deletion on 9p21.3 in FL patient 16,
	encompassing CDKN2A gene
Figure 4.32	Recurrent deletion of 9p21.3 chromosomal region in 4/25 (16%) FL patients,
	suggesting a potentially important role of CDKN2A tumor suppressor in FL 173
Figure 4.33	Sequence validated OPAS-exclusive focal deletion on 4q12 affecting the KIT
	gene (ht-20)
Figure 4.34	Analysis of the impact of 4q12 deletion (136,811 bp) in FL patient 20 on the
	<i>KIT</i> gene
Figure A.1	Probability density function (PDF)
Figure A.2	A graphical representation of the relationship between the PDF and CDF 247
Figure B.1	PDF and CDF of normal distribution
Figure D.1	Schematic representation of two-way ANOVA test performed on each 10K
	SNP

Figure D.2	Measuring the effect of chip and labeling variability in 10K replicate experi-
	ment using two-way ANOVA analysis
Figure D.3	Frequency of inconsistent SNPs across 8 studied DNA samples
Figure E.1	Boxplot and a probability density function (PDF) of a dataset with standard
	normal distribution
Figure F.1	Comparing the results of 5 normalization techniques on the estimated probe-
	level log2-ratio intensity readouts
Figure J.1	Relationship between variation of hybridization signal intensities and CNV
	counts in 25 FL genomes
Figure L.1	Candidate distal deletion on chromosome 8p23.3 of patient 25 (~21 kb) con-
	taining 3 SNP probe markers (OPAS-exclusive)
Figure L.2	Candidate OPAS distal deletion on chromosome 1p36 (ht-18), that includes an
	array CGH predicted deletion
Figure L.3	Candidate distal amplification on 14p36 (ht-7) with slight gain of signal inten-
	sity (LR = +0.18) but a significant z-score (+1.1) $\dots \dots \dots$
Figure L.4	Candidate OPAS distal deletion on chromosome 6 of patient 20 that includes
	11 SNP probes with significant loss of signal intensity (LR = $-0.42$ ; z-score =
	-1.50)
Figure L.5	Examples of FISH validated 1p36 deletions in 4 FL patients
Figure L.6	Candidate OPAS-exclusive focal deletions on chromosome 19 of patient 4 291
Figure L.7	Slight gain of signal intensity of a region within a deleted chromosome arm,
	predicted to represent an amplification (LR = $0.03$ ; z-score = $+0.6$ )

Figure L.8	Candidate OPAS-exclusive focal deletion on chromosome 4 of FL patient 19
	$(\sim 10 \text{ kb}; 4 \text{ SNPs})$ , adjacent to an FPP translocation site between 4q28.3 and
	2p25.1

# Acknowledgments

I owe many thanks to my PhD thesis supervisor Dr. Marco Marra for the opportunity to pursue studies in his lab, as well as for his guidance and support. He was an inspiration, and always enthusiastic about teaching biology and fostering interdisciplinary science. Besides being a great supervisor, he was an excellent mentor, and someone I would like to emulate in my career.

My sincere thanks also go to the members of my supervisory committee: Dr. Jan Friedman and Dr. Robert Holt, who were instrumental in my training throughout the course of my degree. I have also enjoyed the support of many fellow graduate students and post-doctoral fellows including Trevor Pugh, Suganthi Chittaranjan, Ryan Morin, Malachi Grifith, Sorana Morrissy, Claire Hou, Jaswinder Khattra, Monica Sleumer, Maria Mendez-Lago, Rodrigo Goya, and Obi Griffith. I especially extend my sincere thanks to Olena Morozova who has not only been a true friend and an invaluable emotional support throughout my PhD studies, but also immensely helpful in reviewing my thesis in more ways than I could ever list here. I also wish to express my gratitude to Dr. Inanc Birol and Dr. Stephane Flibotte who were abundantly helpful and offered invaluable assistance in the statistical aspects of my thesis.

I am grateful for the funding received from the National Cancer Institute of Canada, the University of British Columbia (Department of Genetics), Genome Canada, Genome British Columbia and the British Columbia Cancer Foundation. Most of the laboratory validation work described in this thesis would not have been possible without the help of Dr. Andy Mungall, whose exceptional dedication to science has motivated my work and who has always answered my questions about biology with patience. I must also extend my thanks to many other colleagues at the GSC, specially Martin Krzywinski, Irene Li, Andy Chu, Matthew Field, Dr. Allen Delaney and Jacquie Schein as well as collaborators outside the Genome Sciences Centre , including Dr. Horsman, Susana Ben-Neriah, Dr. Patrice Eydoux and Dr. K-John Cheung, who helped me in many aspects of data analysis and biological interpretation of the results. I also extend appreciation and thanks to all the members of the Genome Sciences Centre who I do not mention by name but who helped me by

creating an open and exciting atmosphere of scientific collaboration.

The path I have chosen in science and my immense passion in the application of mathematics in solving biological problems would have not been ignited if it were not for the enthusiasm and inspiration of certain people that I have been blessed to meet and collaborate with in my life. I am forever indebted to Dr. Michael Kolios, my Master's thesis mentor, who introduced me to the fascinating field of biophysics and provided me with the first opportunity to experience the exciting field of biomedical data analysis. It is also difficult to overstate my gratitude to the memory of two people who also had major impacts on my passion for bioinformatics: Dr. Sam Roweis, whose immense passion and enthusiasm for new ideas and new perspective on interdisciplinary science sparked my passion for bioinformatics; and my very good friend and fellow graduate student Adrian Quayle, who I will always be grateful to have known.

On a personal level, I cannot forget to express thanks to all my friends and my brothers for believing in me and especially Ali who has been there for me every step of the way. Finally, thanks to my father who dedicated so much of his life to my happiness and who I miss dearly. To my mother, I dedicate this work. She continues to be the only constant parameter of inspiration and love among all other variabilities of my life.

## Chapter 1

# Methods and Strategies For Analysing Copy Number Variation Using DNA Microarrays

### 1.1 Introduction

Genetic variants resulting in gains or losses of DNA segments are collectively termed copy number variants or CNVs and are found both in human and other mammals such as chimpanzee [1–3]. From the earliest days of cytogenetics-based chromosomal analysis, scientists were able to identify chromosomal variants and in many cases were able to associate them with certain human diseases [4–6] (Table 1.1). Association of copy number variation with a phenotype goes back as early as 1936 when Bridges identified duplication of the *BAR* gene in Drosophila melanogaster as the cause of the 'Bar eye phenotype' [7]. In 1959, Jérôme Lejeune discovered the first chromosomal disorder in humans, an extra copy of chromosome 21 (trisomy 21) that was associated with Down syndrome [8], more than 90 years after Down syndrome was first described by John Langdon Down in 1866 [9]. This discovery of the first human condition definitely attributable to chromosome copy number variation was regarded as a turning point in cytogenetics. Since then many other syndromes were associated with large deletions or duplications of chromosomal regions, which were visible using chromosome microscopy [4–6]. In addition to discovery of copy number variants in human disease, seemingly benign chromosomal variants were also identified among normal individuals (often referred to as copy number polymorphisms). These events were

frequently detected in regions of heterochromatin on chromosomes 1, 9 and 16 and in the short arm of the acrocentric chromosome 6 [10, 11].

Later, the hybridization of molecular probes to human chromosomes, particularly with fluorescence in situ hybridization (FISH) [12], provided an effective tool for detection of subtle DNA gains and losses, as well as other chromosomal rearrangements such as inversions and translocations. The advances in molecular analysis technique paved the way for discovery of numerous new genetic variants including short tandem repeats [13] and single nucleotide polymorphisms (SNPs) [14–17]. As the result of these discoveries, it became clear that the scale of variation in the human genome ranged from single base pairs (SNPs<sup>1</sup>) to regions as large as several megabases in size [18–25]. Since then, our understanding of chromosomal variants both in human disease and normal populations has been profoundly expanded as the result of genome-wide chromosome analysis techniques that have allowed us to interrogate the DNA sequence and discover submicroscopic CNVs (< 1Mb), much smaller than the earlier cytogenetics analysis (> 5 – 10 Mb).

We now know that CNVs are common characteristics of human diseases such as mental retardation [26-30], autism [31, 32] and cancer [33-36]. The copy number variants can directly cause disease through altering the abundance of dosage-sensitive genes [37], as in micro-deletion or micro-duplication disorders [26, 38, 39], or affect gene expression, either directly by affecting the genes that are harboured within CNVs, or indirectly through altering upstream or downstream sequences that are involved in gene regulation [40-43]. Furthermore, multiple recent studies have indicated the importance of copy number alteration in susceptibility to human complex diseases such as Alzheimer disease [44–46], Crohn's disease [47, 48], autism [31, 32, 49, 50], psoriasis [51], Parkinson's disease [52–54], schizophrenia [55–59] and glomerulonephritis [60]. Importantly, over the past several years emerging evidence has shown the significance of smaller copy number variants associated with human disease [61-65] as well as normal genome diversity [20, 66-68]. However, the resolution of detecting CNVs is not only dependent on the resolution of the technology, but also on the sensitivity and precision of the computational methods that are used to for CNV analysis. The capacity to reliably detect small copy number variations (below 100 kb in size) using early clinical microarray platforms appeared to be limited [69–71], suggesting that there are yet undetected and potentially disease causing CNVs that required higher resolution methods of genome analysis.

As the importance of copy number variants is well established, it is important to note that the power to detect these variants depends on two main factors: 1) the resolution of the technology

<sup>&</sup>lt;sup>1</sup>SNPs are individual base positions in the genome that show natural variation in a population with more than 1% frequency (according to the Single Nucleotide Polymorphism database; dbSNP).

being used to study the sample, and 2) the sensitivity and specificity of the computational methods that are applied to analyse the data. In the rest of this Chapter, I will first review the technologies that have been used for CNV detection and then will discuss some of the computational approaches that are employed for CNV analysis.

## **1.2** Technologies

The techniques to detect chromosomal abnormalities can be broadly categorized into 3 groups based on their detection resolution and genome coverage: techniques with low to moderate resolution with limited genome coverage (first generation), high-resolution chromosomal microarray analysis (second generation), and massively parallel sequencing technologies (third generation).

### **1.2.1** First Generation Techniques

#### 1.2.1.1 Chromosome Banding

For more than 50 years, the standard clinical method for detection of chromosomal abnormalities was chromosomal cytogenetic analysis using karyotyping, a microscopic method that requires highly skilled interpretation. The conventional process for karyotyping, known as "chromosome banding", involves adding a dye to metaphase chromosomes that provides a visual image of different regions of each chromosome by its unique pattern (usually as a black-and-white staining pattern). The most commonly used chromosome banding, G-banding, uses Giesma stains to visualize transverse bands on a chromosome<sup>1</sup>. Each chromosome has a characteristic banding pattern that helps to identify it and both members of a chromosome pair have the same banding pattern (see Figure 1.1). This method typically produces between 400-800 unique bands which can be distinguished by the order and size of each band. Karyotypes are arranged with the paired chromosomes ordered by size, the short arm of the chromosome on top and the long arm on the bottom [72]. Comparing each chromosome's banding pattern to its normal pattern enables cytogeneticists to recognize chromosomal abnormalities such as numerical changes as well as deletions, duplications and translocations if they are of sufficient size (e.g., larger than 4-5 Mb; Figure 1.1). A major challenge of using G-banding to stain chromosomes is that subtle deletions and translocations near the telomeres are extremely difficult to identify by this method (since Giesma staining produces

<sup>&</sup>lt;sup>1</sup>In general, heterochromatic regions, which tend to be AT-rich stain more darkly in G-banding, in contrast to less condensed GC-rich chromatin which incorporates less stain and thus appear as light bands in G-banding.

light bands for most of chromosomes tips), and thus other staining methods such as R-banding<sup>1</sup> have been used to detect telomere-specific aberrations [73–75]. Despite their limitations, G-banded karyotypes are still routinely used in clinical applications to diagnose a wide range of large-scale chromosomal abnormalities, including trisomy 21 in Down syndrome (written as 47, XX, +21).

# 1.2.1.2 Spectral Karyotyping (SKY) and Multiple Fluorescent in Situ Hybridization (M-FISH) Analysis

To improve resolution and efficiency of conventional chromosome banding, Spectral Karyotyping (SKY) and Multiplex Fluorescent in Situ Hybridization (M-FISH) chromosome analysis methods were developed. These methods use labeled chromosome-specific paints to provide simultaneous visualization of chromosomes in different colors [76, 77]. SKY uses multiple fluorochromes to measure the spectrum of each image pixel, simultaneously, by means of an interferometer. M-FISH, on the other hand, generates separate images for each of 5 employed fluorochromes through application of special filters and later superimposes these images automatically to obtain a single image in full color (see Figure 1.1b). The main advantage of SKY and M-FISH is characterizing changes with respect to their origin, such as translocation. SKY and M-FISH have been used to detect and characterize chromosomal abnormalities in different cancers including those of breast [78, 79], colon [80, 81], bladder [82–84], lung [85, 86] and cervix [79, 83, 84, 87].

In contrast to conventional karyotyping, both SKY and M-FISH generate a digital image in full color (instead of simple black-and-white pattern), which enhances the observation of structural aberrations in the entire genome, and provides insight into the chromosomal composition of several ambiguous marker chromosomes<sup>2</sup> [88]. Furthermore, working with digital images allowed scientists to use computers to analyse the "painted" chromosomes and automate identification of structural abnormalities. This significantly reduced the cost and time of conventional labour-intensive karyotyping, and improved the accuracy of finding true abnormalities by minimizing the human errors of interpreting black-and-white karyotypes. As a result, SKY and M-FISH have been used to detect chromosomal rearrangements [89, 90]. However, the resolution of these techniques is estimated to be approximately ~1-2 Mb for both SKY [91, 92] and M-FISH [93].

<sup>&</sup>lt;sup>1</sup>R-banding is a staining method in which chromosomes are heated in a phosphate buffer, then treated with Giesma stain to produce a banding pattern that is the reverse of that produced in G-banding. Thus, the dark regions are gene-rich euchromatic and light bands are heterochromatic (tightly packed form of DNA).

<sup>&</sup>lt;sup>2</sup>A marker chromosome is an abnormal chromosome that is distinctive in appearance but not fully identified.

#### **1.2.2** Second Generation High-Resolution Techniques

#### **1.2.2.1** Comparative Genome Hybridization (CGH)

In cytogenetic CGH, which was first developed by Kallioniemi et al. [94], the patient and normal samples (also referred to as "test" and "reference" samples, respectively) are labeled with different fluorescent tags and co-hybridized to normal metaphase chromosomes. The hybridization fluorescence intensities from test and normal samples are converted into quantitative ratio measurements that represent the gains and losses in the test (patient) genome relative to the reference genome [94]. Cytogenetic CGH has been a popular tool to characterize chromosome imbalances in many different clinical applications [94–97] including mental retardation [97–99] and cancer [94, 100–106]. For example application of CGH in ovarian cancer resulted in the discovery of several copy number variants that affected some of the key genes in ovarian tumorigenesis including loss of 17pter-q21 that harbors *p53*, gain of 17q that results in amplification of *HER2/neu (ERBB2)* and amplification of 8q24 including the *MYC* oncogene [107–109]. CGH has also been successfully applied to analyse hematological cancers such as leukemia and lymphoma [103, 106, 110].

Although cytogenetic CGH technology had a huge impact on cytogenetics analysis of human disease, these methods were very labor intensive and required the use of metaphase chromosomes, which led to limited resolution, typically about 5-10 Mb [103, 111, 112]. Completion of the Human Genome Project [113], where large-insert clone libraries were developed and assembled into overlapping contigs for sequencing, initiated a major improvement to CGH techniques. In an attempt to overcome the aforementioned limitations associated with cytogenetic CGH, investigators developed a method that combined the principles of CGH with the use of microarrays [114]. Instead of using metaphase chromosomes, this method, which is known as array CGH (aCGH), used DNA clones that accurately mapped to known regions of the genome and were robotically spotted onto array glass slides or glass capillaries [115]. In array CGH labelled samples are applied to a slide containing thousands or millions of DNA probes [116]. The resolution of such arrays depends on the size of the genomic fragments that are used as DNA probes (e.g., ~150 kb for BAC aCGH), the density of the array (e.g., Agilent ultra-dense 1M array CGH platform includes 1 million probes) and the structure of the genome sequence being analysed. Similarly to cytogenetic CGH, in array CGH technology, test and reference DNAs are labeled with different fluorescent dyes and then are co-hybridized to the array. Relative gains and losses of signal intensities are subsequently measured and reported as copy number deleted or amplified regions of the genome. This technique has revolutionized the study of CNVs in many different applications such as cancer studies [116–121].

Among the genomic representations of DNA that are used as the probes in array CGH platforms, Bacterial Artificial Chromosomes or BACs were heavily used initially, especially for studying CNVs in cancers [36, 122] and mental retardation syndromes [26, 123–125]. Using BACs for CGH (known as BAC aCGH) provided genome resolution and coverage that was unprecedented before that time, as shown in a study by Krzywinski et al. [126] which indicated that more than 99% of the entire human genome can be represented by a set of 32,000 BAC clones with an average intermarker distance of 76 kb between BAC clones. Despite the unprecedented resolution and genome coverage of BAC-aCGH, the empirical resolution of this technology to detect chromosomal abnormalities was still limited by the average size of BAC clones [126, 127] and by technical difficulties of producing high density, highly reproducible BAC arrays in large numbers required for clinical applications.

Following the success of aCGH technology to detect structural aberrations up to one-tenth the size of those detectable by conventional cytogenetics [36, 122, 126, 128–132], using DNA arrays with shorter probe sequences became increasing popular in recent years. Thus in the next section I will focus on describing oligonucleotide arrays as an important tool for high-resolution whole-genome analysis developed during the past decade.

### 1.2.2.2 DNA Oligonucleotide Arrays

Oligonucleotide arrays consist of an arrayed series of thousands or millions of microscopic spots of DNA oligonucleotides, called features, each containing a specific DNA sequence (or probe) [133–138]. The length of oligonucleotide probes, or oligos, varies between different array types and vendors but typically is in the range of 25 (used by Affymetrix ) to 60 (used by Agilent and Illumina) nucleotides [137–139]. The intensity of target-probe hybridization is then translated into measurements representing the abundance of DNA in the test sample relative to the reference. Oligonucleotide arrays provide the highest potential resolution for microarrays. However, in practice the effective resolution of oligonucleotide microarrays is dependent on several factors, such as the length of the oligonucleotide probes (or "oligos"), the density of the probes on the array and the coverage of the genome. Another source of data variability in microarrays is the array manufacturing technology. Based on the manufacturing technology, microarrays can be broadly categorized into "spotted" and "in situ synthesized" arrays.

In earlier arrays the probes were synthesized prior to deposition on the array and then spotted on the array surface by means of a spotting robot (known as "spotted arrays"). Spotting technology allowed a maximum of about  $\sim 60,000$  oligos to be printed on any given array [139] and consequently the density of the spotted arrays was limited to approximately a single probe per 50 Kb sequence [139]. The next major manufacturing technology, in situ hybridization, was fundamentally different from robotic spotting as the oligos were synthesized, base-by-base, directly on the array surface. Over the past decade, many developments have been made in array technology, and in particular there has been a significant trend toward increased numbers of features (probes) and toward shorter DNA sequences as hybridization targets [139], both of which have impacts on the resolution at which CNVs can be detected. Despite their improved density and detection resolution, a major disadvantage of oligo arrays is the relatively poor signal-to-noise hybridization intensities that leads to considerable variability in the reported number and size of CNVs [139–142]. To improve the signal-to-noise ratio (SNR) limitation, Lucito et al. [115] developed a method that is based on reducing the complexity of the genomic DNA that being is hybridized on oligo arrays, known as Representational Oligonucleotide Microarray, or ROMA using 70-mer oligos as genome representations. Briefly, in ROMA the genomic DNA is digested using a restriction enzyme (often BgIII) and the resultant fragments are then ligated to adapters and PCR amplified using universal primers. Because of preferential PCR amplification of smaller segments the final amplification product would be depleted in larger fragments, leading to a reduction in the complexity of the sample [115]. These products were then hybridized to an oligonucleotide array consisting of probes that were selected to match the reduced set of restriction fragments. But ROMA also suffers from poor signal-to-noise ratio measurements. Furthermore, it presents other potential problems for CNV detection studies. First, as the result of the complexity reduction step (explained above) different regions in the DNA could have different representations due to their sequence content (not their sequence abundance), and this different representation could be mistakenly interpreted as copy number variation. Second, different individuals will have different restriction digestion patterns, and it is possible that some individual probe ratios may be related to restriction fragment size differences rather than to true copy number changes. However, the main limitation of ROMA arrays is their low signal-to-noise ratio compared to BAC arrays, and thus typically 3 probes are averaged to improve the associated variance of the array data leading to lower CNV detection resolution compared to BAC aCGH [139].

#### 1.2.2.3 Genotyping Arrays

During the past two decades, Single Nucleotide Polymorphisms (SNPs) have been recognized as a major source of human genetic variation [14–17, 143–145]<sup>1</sup>. These findings have been made

<sup>&</sup>lt;sup>1</sup>http://www.hapmap.org

possible largely by the development of high-throughput array technologies for SNP genotyping from commercial vendors such as Affymetrix and Illumina. Although these arrays were originally developed for genotyping SNPs, the intensity information from these arrays can be used to detect copy number variants providing both SNP genotypes and copy number estimates from a single experiment simultaneously. Since the introduction of this technology, extensive research has focused on studying CNVs in human disease, such as mental retardation [29, 30, 146], schizophrenia [147], autism [148], cancer [149–160], and normal polymorphisms [69, 153, 158, 161–164]. In the next two sections I present a brief description of SNP arrays from Affymetrix and illumina.

**1.2.2.3.1 Affymetrix SNP Arrays:** The first generation of commercial SNP arrays known as "HuSNP" was produced by Affymetrix and became available more than a decade ago [17]. The early HuSNP arrays were capable of genotyping 1,494 SNPs in a single experiment, and since then Affymetrix has continued to release newer arrays with increased numbers of features including 10,000, 100,000, 500,000 and now with ~1 million SNPs (www.affymetrix.com; see Table 1.2).

Briefly, in this technology, total genomic DNA (250 ng) is digested with a restriction enzyme (such as Nsp I or Sty I in 500K arrays) and ligated to adaptors for PCR amplification<sup>1</sup>. The PCR conditions have been optimized to preferentially amplify fragments in the 200 to 1,100 base pairs (bp) size range<sup>2</sup>. This preferential amplification reduces the complexity of the hybridization by incorporating the smaller fragments. Finally, the amplified DNA is fragmented, labeled, and hybridized to a chip.

In Affymetrix technology used in 10K, 100K and 500K platforms, each SNP sequence is interrogated by a set of 25-mer oligonucleotide probes that target the SNP site and its surrounding base pairs, as shown in Figure 1.2. Each SNP on the array is represented by a collection of probe quartets, also known as the SNP probe set. A probe quartet consists of a set of 25-mer oligonucleotide "probe pairs" for two most common alleles (known as 'A' and 'B') and for both forward and reverse strands (antisense and sense) for the SNPs. Each probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe (see Table 1.2). The number of designated oligonucleotides in a probe set varies in each generation of the arrays. In 10K and 100K arrays [135, 137], a probe set consisted of 40 different oligonucleotide probes, while in 500K analysis this number was generally reduced to 20 oligonucleotides, although a subset of SNPs in 500K Affymetrix arrays still retain 40 features [165]. For CNV detection, the signal intensities of the probes (see Figure 1.2) are compared with values from another individual (or group of individuals) and the relative copy number

<sup>&</sup>lt;sup>1</sup>http://media.affymetrix.com:80/support/technical/datasheets/500k\_datasheet.pdf

<sup>&</sup>lt;sup>2</sup>http://www.affymetrix.com/support/help/faqs/gw\_human\_snp5/faq\_4.jsp

per locus is determined [166, 167]. The technology used in the latest Affymetrix genotyping array, SNP 6.0, has major differences with the three previous generations of Affymetrix SNP arrays. The SNP 6.0 array consists of probes for both SNPs and copy number variation. The copy number variation probes were selected based on Toronto Database of Genomic Variants (DGV) [168]. In SNP 6.0 platform each 'A' and 'B' allele of a SNP probe are presented by 3-4 replicate PM oligonucleotide probes resulting in 6-8 oligos per SNP.

The Affymetrix Genome-Wide Genotyping<sup>®</sup> arrays have been widely used for high-throughput SNP genotyping (in population genetics [169], linkage disequilibrium analysis [170], and whole-genome association studies [171]), and copy number analysis [149–157].

**1.2.2.3.2 Illumina SNP arrays:** Similar to Affymetrix technology, Illumina SNP arrays have also been increased in capacity from 100,000 SNP probes in Human-1 array to ~1.2 million probes in Infinium HD BeadChip that consists of both SNP and CNV probes [172, 173]. Both Affymetrix and Illumina technologies share the same underlying principle for identifying CNVs through using the array intensity data and both enable simultaneous analysis of genotypes and copy number data [174]. Despite their similarities, the two products have substantial differences [172, 173, 175]. For example, Illumina arrays use 50-mer oligos compared to Affymetrix 's 25-mers. Also, Illumina has 1 or 2 replicate probes per SNP allele whereas Affymetrix has about 4-6 probes per allele (Affymetrix 10K-500K arrays have 10-40 oligos to interrogate a SNP locus, however, these oligos do not have the exact same sequence<sup>1</sup>). In the context of SNP genotyping, the Illumina Infinium assay, which runs on its 1M-Duo chip, uses single-base extension with a labeled base to call a SNP genotype [172, 175]. Nonetheless despite their SNP genotyping approaches, in the context of chromosome copy number discovery, the signal-intensity output from both platforms present similar analysis and interpretation problems [173, 175].

**1.2.2.3.3 Application of SNP arrays for CNV Detection:** Initial genotyping arrays provided unprecedented resolution for identifying chromosome copy number aberrations both in normal and disease states and the results have improved with the subsequent developments of the technology. Even so, the high level of associated noise has been the main computational challenge of interpreting the array signal intensity for CNV detection [176]. Since the introduction of this technology, various methods have been developed to reduce the noise and improve the sensitivity and specificity of CNV calling [69, 149, 150, 161, 178–183]. Also, various copy number detection

<sup>&</sup>lt;sup>1</sup>The new design strategy in Affymetrix 6.0 arrays uses replicate oligonucleotide probes to interrogate each SNP (see Table 1.2).

algorithms have been developed to aid CNV detection [161, 168, 184–193]. However, for robust CNV detection, most of these methods require significant concordant ratio shifts in several probes sequentially located along the genome, which consequently lowers "effective resolution" of these arrays [139, 141]. More importantly, the distribution of the probes is not uniform across the entire genome (e.g., in Affymetrix GeneChip 100K, SNPs have a median spacing of 8.5 kb but a mean intermarker distance of 23.6 kb [167]) and, thus, CNV calling approaches based on ratio shifts in several consecutive SNPs will inevitably limit sensitivity for small CNVs and CNVs in genomic regions sparsely populated by probes (often methods require at least 8 - 10 probes to identify a CNV).

### 1.2.3 Third Generation High-Resolution Techniques

More recently with the advent of next-generation sequencing technologies, a few groups have applied massively parallel sequencing platforms with the aim of improving the sensitivity of CNV detection to base-pair resolution [194–197]. The existing algorithms for sequencing-based CNV analysis can broadly be categorized into two groups. The first category is primarily based on paired-end read mapping (PEM), as was previously reported by Tuzun et al. [24] and Korbel et al. [198] (both obtained by 454 technology). In the PEM approach, 3 kb paired end reads are computationally mapped to the human reference genome. The mapped pattern of the 3 kb reads is then analysed to detect regions of structural variations [198]. Therefore, deletions are identified by paired ends spanning a genomic region in the reference genome longer than the known fragment length. Similarly, insertions are predicted through paired ends that span a region shorter than the reference genome would predict, or pair ends that cross chromosomes. It is reported that PEM-based detection methods have several limitations and demonstrate particularly poor performance in complex genomic regions that are rich in segmental duplications and have limited ability to detect insertions larger than the average insert size of the library [24].

The second category uses the depth of the sequencing coverage to predict CNVs [194]. Evan Eichler's group used this approach to develop the mrFAST algorithm. This algorithm measures the depth of the coverage of whole-genome shotgun sequencing (WGS) reads that are aligned to the human reference genome by checking every locus in the genome and matching them to reads with at least 94% identity. The algorithm consequently uses the average depth of the aligned reads to detect regions of copy number aberration [194]. The performance of this approach was tested using 3 sequenced human genomes, including Yoruban [199], Han Chinese [187] and the Watson genome [200]. The result of this study indicated that the number of reads sampled from a given

region, referred to as read depth, was proportional to the number of times the region appeared in the corresponding genome. To further test this hypothesis, Alkan et al. [194] studied read depth in 961 autosomal duplications and concluded that at 20-fold sequence coverage, > 90% of all segmental duplications larger than 20 kb could be accurately identified by analysing the sequencing read depth. By selecting regions with increased sequence coverage, Alkan et al. [194] identified 725 non-overlapping large segmental duplications. Nearly all of these 725 detected segmental duplications were present in the sequenced genomes of all three subjects. A similar approach for detecting structural variations in next generation sequencing data has been posed by Yoon et al. [195]. This approach, known as event-wise testing or EWT, also uses read depth (RD) of the coverage and relies on statistical testing of 100-bp RD intervals to identify potential region of increased or decreased RD coverage, and then uses this information to infer regions of copy number variation. The plots of averaged RD data across the chromosome lengths resemble copy number scatterplots from oligonucleotide arrays, and there seems to be a significant variation in the estimated RD measurements [195]. Furthermore, the windowing approach towards smoothing the RD data has certain limitations, notably, the criteria that are used to select the optimal window size is rather experimentspecific and may not be applicable for other experiments (a detailed discussion of the pros and cons of windowing-based smoothing of the data is presented in Section 1.3). The basis of the method proposed by Yoon et al. [195] is that the scatterplot of RD coverage along the genome will follow the normal distribution after averaging the RD measurements in 100-bp intervals; however, this hypothesis may not hold true for different experiments or even different parts of the genome with uneven coverage [197]. Therefore, in practice, a 100-bp windowing of RD data may not generate a normal distribution and thus none of the downstream statistical analysis in EWT [195] would be applicable, since they are strictly based on the assumption of a normal distribution. Substituting non-parametric methods to analyse the sequence data could reduce variation in RD data and may be a potent tool to improve CNV identification at high resolution. It is clear that new computational approaches are needed to systematically detect copy number variants from sequence data [201].

Oligonucleotide arrays are also being used in parallel with sequencing data. For instance, Affymetrix 500K SNP arrays were used to assess the accuracy of the known sequence-derived SNPs from Watson genome [200] and to provide a map of CNVs of the Craig Venter genome [163]. Additionally, as the costs of array production, labelling, and hybridization continue to fall, these arrays are becoming more accessible and, therefore, the range of their applications is growing [202, 203]. These factors emphasize the significance of developing highly accurate computational methods that can improve the sensitivity/specificity of current CNV detection algorithms. Additionally, although next generation DNA sequencers may become the dominant technologies for CNV detec-

tion [202], computational biologists have already begun borrowing methods initially developed for oligonucleotide arrays to analyse sequence data. The next section details the history of computational advances associated with SNP arrays, which have been the focus of my thesis.

## **1.3** Methods to Discover CNVs Using SNP-array Data

The underlying principle of all SNP array copy number analysis algorithms is to compare the fluorescence intensity ratios along the length of each chromosome to identify regions of candidate copy number loss or gain in the test sample, relative to the reference sample. A major concern for the detection of CNVs using oligonucleotide array technology is how a putative CNV is defined computationally. There is a plethora of different methods being used to call significant changes in relative intensity ratio from arrays. Relative to Illumina SNP arrays, more methods have been developed and evolved to analyse Affymetrix SNP array data, since these arrays have both more data redundancy per SNP locus and have been commercially available longer than Illumina SNP arrays. Typically, each copy number analysis algorithm involves 2 main steps: (1) normalization, and (2) CNV calling. With each new version of Affymetrix SNP array technology, these modules have evolved and been modified to improve the sensitivity and specificity of CNV detection [69, 149, 150, 161, 178–183].

#### **1.3.1** Normalization Methods

As previously mentioned, SNP array data analysis is based on processing the relative signal ratio changes in a test array against one or more samples known as the "reference set". By doing so, the analysis inevitably incorporates the variability that exists between different arrays (chip-to-chip variability) [204]. Therefore, to perform an accurate data analysis, it is crucial to first normalize raw intensities to correct for the variation that exists between different chips and different hybridization experiments [205] (also known as between-array normalization). Several commonly used methods for normalizing SNP microarray data have been adopted from expression arrays, including global normalization [206], invariant-set normalization [207], and LOWESS [208]. The current consensus, however, is based on Quantile Normalization [205], a non-parametric approach developed by Terry Speed's group for SNP oligonucleotide array data normalization. Quantile normalization guarantees all samples in an analysis have similar intensity distributions (instead of just focusing on the mean/median of intensities across the chips, as in global normalization).
#### **1.3.2** CNV Calling Methods

Statistical methods for analysing copy number data are necessary for identification of CNVs. The development of methods that can accurately identify CNV regions has been a major challenge for microarray-based copy number analysis during the past several years [175, 176]. Therefore a variety of statistical analysis and visualization tools have been developed for Affymetrix SNP array platforms [69, 149, 150, 161, 175, 176, 178–183]. Despite their algorithmic differences, the logical structure underlying these approaches typically belongs to one of the following statistical models: (1) hidden Markov models (HMMs), such as QuantiSNP [185], PenCNV [186], HMMSeg [184] and dChipSNP [188, 209], (2) segmentation algorithms such as DNAcopy [189], GLAD [210], Circular Binary Segmentation (CBS) [189] and FACADE [211] (3) t-tests and standard deviation-based thresholding of the log2-ratio intensity measurements, as in [210, 212, 213]. Many of these methods, such as CBS and HMMs, were initially designed for aCGH and later adopted for SNP arrays [139, 214]. Below is a brief description of the fundamental basics of each of these algorithm categories.

One of the simplest approaches to identifying shifts in the array intensity outputs is based on analysing the standard deviation (also known as SD) of log2-ratio intensities using thresholds for identifying putative regions with significant log2-ratio deviation from the baseline, as in [212, 215]. These methods were originally developed and widely used for copy number data analysis in aCGH platforms (such as BAC aCGH [213] and cDNA-based aCGH [34]). Despite their simplicity and speed, in the presence of non-specific variation in the signal intensities (or noise) the thresholding-based methods perform poorly in detecting true regions of copy number aberration [175, 176].

In an attempt to overcome limitations of thresholding-based CNV detection algorithms, Pollack et al. [216] used a modified thresholding approach in an aCGH platform to detect CNVs in 44 primary breast tumors and 10 breast cancer cell lines. In this approach, the data were first smoothed by averaging over a window of optimal size, and a statistic was calculated for each probe in the window. Next, based on these statistics, CNV false discovery rates (FDRs) were estimated using the Benjamini and Hochberg method [217] and applied to determine the thresholds of DNA copy number gains and losses in the corresponding breast cancer data set [216]. The windowing based t-test methods have since been widely used in numerous copy number detection studies [29, 30, 162, 215], but these methods suffer from major drawbacks. The main challenge of windowing approaches is the ambiguous nature of determining an optimal window length (the same limitation that was described earlier for EWT method, proposed by Yoon et al. [195]; p. 11). In fact, genomics was not the first field of study that applied windowing to improve signal variation and recognized

its limitations. The drawbacks of windowing-based smoothing had been previously brought to attention in neurology when scientists tried to identify normal and abnormal patterns of brain EEG  $(electroencephalography)^1$  and knee VAG  $(vibroarthrographic)^2$  signals among individuals. The common problem of windowing based methods is the optimal length of the defined window (or window size). Choosing a large window size results in a greater degree of smoothing but would inevitably hide small copy number changes. On the other hand, if the selected window size is too small, the presence of a few sporadic noisy probes would be sufficient to generate a false positive CNV readout. Another drawback of this approach is that it suppresses the magnitude of signal intensities for both noisy and informative probes. Although such data suppressing reduces the variation of signal intensities and lowers the overall standard deviation of the signal, it also reduces the magnitude of true signal aberrations. This limitation can reduce the sensitivity of CNV detection, particularly for small CNVs or CNVs with fewer SNP probe markers [218–221]. A fundamental drawback of the above approach to data smoothing and thresholding is handling tumor heterogeneity in cancer, which refers to the presence of different cell subpopulations in a sample [218]. In such cases the combination of normal and copy number aberrated cells results in log2-ratio measurements that are well-below the predetermined thresholds of calling a CNV. Thus methods that are solely based on thresholding cannot detect such changes unless the CNV is present in the majority of the cells to generate a significant shift in signal intensities away from the baseline.

To overcome the limitations of thresholding-based CNV calling algorithms to handle noisy data (particularly in oligonucleotide arrays), model-based algorithms were developed that focused on statistical analysis of candidate CNV regions to improve the ability to recognize the difference between non-specific (noise) and informative hybridization signal intensities. One of the earliest model-based approaches for copy number data was proposed by Hodgson et al. (2001) [33] in an aCGH study in mice, where a three-component mixture model was fit to islet tumor data in which each component represented one state of copy number data corresponding to copy number gain, loss or neutral states. They subsequently used the information from these Gaussian models to determine the thresholds above or below which aCGH ratios should be considered as increased or decreased. In 2003, Snijders et al. (2003) [35] developed a heuristic method to fit a Gaussian hidden Markov model (HMM) to array CGH copy number data, and since then HMMs have been routinely used to detect CNV regions and to predict the actual ploidy of the regions [184]. A common assumption by HMMs is that observed intensities are related to an unobserved copy number state

<sup>&</sup>lt;sup>1</sup>electrical activity along the scalp produced by the firing of neurones within the brain

<sup>&</sup>lt;sup>2</sup>vibration signals emitted during movement of the knee

at each locus that can be defined by an emission distribution (often assumed to be Gaussian).

Like windowing methods, HMMs have a long history in other applications, mainly in speech recognition [222], and were later adopted by bioinformaticians initially for DNA sequence alignment [223]. Another important assumption of an HMM model is that copy number states follow a pattern, so neighbouring SNPs have similar copy number states, and thus the transitions between copy number states can be predicted through a transition matrix that describes the probability of moving from one state to another. This transition matrix is learned directly from the data (also known as the training set) by applying another statistical method, such as Expectation Maximization (EM) [175, 224]. After the training phase of the model, the HMM can be applied for CNV detection in a new experiment, where each log2-ratio possibility is assigned a state and the Viterbi algorithm is used to predict the state for each observed events [175].

Since the introduction of HMM models to array genomic hybridization, many different methods have been developed for analysing array copy number data from different platforms, such as HMMSeg [184] (in aCGH), QuantiSNP [185], dChipSNP<sup>1</sup> [188, 209], and PennCNV [186]. But despite their popularity and various publications that used this technique to identify novel CNVs, HMM models have their own limitations. First, training an HMM model requires proper initialization of both transition<sup>2</sup> and emission<sup>3</sup> matrices, with transition matrix being particularly sensitive to initialization values. This implies that the transition probabilities, and thus the estimated HMM results, are sensitive to the assumptions about the patterns of fluctuations of log2-ratio intensity data between neighbouring probes. Therefore, training an HMM in one particular disease yields a model that is sensitive to the CNV patterns in that particular disease.

The HMM approach for CNV detection has a number of other limitations. Estimating putative regions of CNVs depends of the initial hypothesis about present number states in the array data. Often two states are assumed in the data (one for amplification and another for deletion), however it can be difficult to assign a single state to a genomic region, particularly in cancer studies. For example, if a fraction of the tumor cells have lost a particular DNA segment while others have not (CNV heterogeneity), or if the size of the lost region varies between tumor cells, we would observe slight gains or losses of signal intensities, that do not necessarily fall into a certain HMM predefined state. There is an increasing body of evidence to indicate that tumor heterogeneity is an important characteristic of most cancers [225]. Another factor that can result to an ambiguous state is sample impurity, for example, when the pathologist has not been successful in removing

<sup>&</sup>lt;sup>1</sup>http://biosun1.harvard.edu/complab/dchip/

<sup>&</sup>lt;sup>2</sup>Transition probability of an HMM state describes the probability of moving from the current state to a new state.

<sup>&</sup>lt;sup>3</sup>Emission probability describes the likelihood of a certain output given the current HMM state

surrounding normal tissue, or if the tumor sample itself is an admixture of cancer and normal cells.

The above properties of the biological samples which makes it difficult, if not impossible, to accurately define certain copy number states and the dependency of the HMM model on the predefined number of copy number states suggest that HMM-based models may not be suitable for CNV analysis in cancer studies. Also, as previously explained, the HMM predictions are dependant on the transition probabilities that are estimated based on the training data. However, results from cytogenetics and other chromosomal analysis techniques have shown that the patterns of copy number changes are substantially different between disease (e.g., lymphoma versus mental retardation). Therefore, it is reasonable to assume that an HMM that is trained on validated copy number gains and losses from a particular training set would be more sensitive towards identifying CNVs in samples with similar properties (for example, with similar sample heterogeneity). Furthermore, if the transition probabilities are not properly initialized, there is a high risk of the algorithm getting stuck in a local minimum resulting in an improper training of the HMM, which further complicates HMM initialization.

The aforementioned limitations of HMM (training dependancy and initializations of states), emphasize that prior to using a computational method to analyse biological data, we need to have a thorough knowledge of the biological properties of the data as well as the requirements of computational methods to select a model that is likely to generate more accurate results.

Another popular approach for CNV calling is segmentation of log2-ratio copy number data [226, 227]. As is the case with HMMs, there are different kinds of segmentation models, many of which were originally developed for CNV analysis of cDNA and BAC CGH arrays. The common assumption underlying all copy number segmentation methods is that CNVs occur in contiguous regions of the chromosomes, often spanning multiple probes. Based on this hypothesis, segmentation methods attempt to split the chromosomes into regions of equal copy number. The average (or median) log2-ratio of each segment is then used to identify candidate CNVs. Various segmentation methods have been proposed, such as SW-Array (Price et al. [228]), and CGHseg (Picard et al. [229]). A popular segmentation method is Circular Binary Segmentation (CBS; Olshen et al. [189]). This non-parametric method is a modification of Binary segmentation (developed by Sen and Srivastava; [230]), with improved sensitivity towards small variants that may otherwise be obscured within larger segments. CBS assumes that copy number data may be noisy, and as a result, some probes do not reflect the true copy number in the test sample. Since this algorithm does not make any assumption regarding the distribution of the data (non-parametric), it provides a natural way to segment a chromosome into contiguous regions by recursively applying a statistical test to detect significant breakpoints in the data, and continues to divide a region into segments

until it no longer finds a segment that is different from the neighboring regions (or until it reaches a maximum number of permutations). The CBS change-point detection method is designed to identify all the places which partition the chromosome into segments with the same (log2-ratio) copy number. Due to the complete non-parametric treatment of the array data, CBS is potentially one of the most robust CNV calling algorithms [231, 232]. A comparison of the performance of segmentation algorithms by Lai et al. [231] using 11 different methods for analysis of aCGH data found that CBS was among the top 2 algorithms with the best performance under various conditions. Another independent study by Willenbrock and Fridlyand [232] compared 3 Bioconductor packages: DNAcopy<sup>1</sup> [189] (based on CBS segmentation), aCGH software<sup>2</sup> (based on HMM) and GLAD<sup>3</sup> [210] (based on adaptive weights smoothing) segmentation methods and found that CBS-based DNAcopy [189] had the best performance in terms of its sensitivity and FDR for breakpoint detection [232].

## 1.4 Limitations of Current CNV Detection Algorithms for SNP Data Analysis

Regardless of the algorithm used, the variation in hybridization intensity measurements from the array can impact the reliability of CNV detection. The high rate of variability in signal intensity outputs increases the number of regions that are mistakenly identified as CNVs, resulting in increased false positive rates. An excellent example of this undesirable effect is the overpopulation of apparent CNVs that have been observed in the Database of Genomic Variants<sup>4</sup> (DGV) in several independent publications [3, 141, 168]. In addition to increasing false positive rates, such variability can hamper our sensitivity to detect true CNVs and subsequently increases the false negative rate of CNV identification. Parametric approaches mitigate high rates of false positives by applying more stringent criteria, often requiring that a significant shift in signal intensity outputs must be detected in multiple neighbouring SNP probes before CNV is identified. This approach inevitably under-identifies small CNVs or CNVs that occur in regions of the genome with low probe density, such as segmental duplications [141]. The main caveat of such parametric approaches is that the discovery procedure emphasizes specificity over sensitivity and as a result, despite a general reduction in false positive calls, the detection power is dependent on probe counts and the computational methods are not sensitive enough to detect small CNVs.

<sup>&</sup>lt;sup>1</sup>http://watson.nci.nih.gov/bioc\_mirror/packages/2.3/bioc/html/DNAcopy.html

<sup>&</sup>lt;sup>2</sup>http://bioconductor.org/packages/2.6/bioc/html/aCGH.html

<sup>&</sup>lt;sup>3</sup>http://www.bioconductor.org/packages/2.4/bioc/html/GLAD.html

<sup>&</sup>lt;sup>4</sup>http://projects.tcag.ca/variation

A study of CNV hotspots in a general population published by Itsara et al. [141] provides a good example of describing the impact of the aforementioned computational limitations on CNV measurements by suggesting that their reported CNV findings significantly underestimated the number and size of small copy number aberrations. They further elaborated that this shortcoming was due to the dependency of their CNV detection algorithm on the number of probes [141]. Generally, while large CNVs (> 4 Mb) are routinely identified by most of the available algorithms, when it comes to identifying small CNVs or variants located in regions with reduced probe density (< 8 - 10 SNP probes), these algorithms are no longer consistent. The choice of reference set sample size (number of required reference samples for a particular algorithm) is another bioinformatics challenge, especially in cancer analysis where often it is desirable to perform a pairwise analysis of tumor and matching normal DNA to identify somatic CNVs. Meanwhile, some methods (such as CNAT [121] and GLAD [210]) depend on smoothing the variation of the intensity data by averaging the reference signal over multiple normal individuals and therefore do not allow pair-wise analysis.

It is clear that developing a method that can reduce both false positive and false negative CNV calls will have a major impact on the accuracy and reliability of CNV findings using SNP array data. As mentioned earlier, controlling false positive/negative rates also depends on improving the associated noise (non-specific variation) in the SNP intensity outputs. These factors imply that optimal CNV detection requires an algorithm with the following two components: (1) a preprocessing phase that filters unwanted noise from the array raw signal intensity readouts, and (2) a non-parametric CNV calling approach that can translate the intensity information into relative copy number change and apply statistical methods to identifying locations of gains or losses of copy number. Nonetheless, very little work has been done to address the associated noise as an independent module [218, 233, 234]. Instead, most often algorithms tend to adjust the impact of the unwanted variation by modifying the downstream CNV calling algorithm, which consequently results in high false positive/negative rates as discussed earlier [140]. In conclusion, the high level of noise associated with SNP oligonucleotide data are still a major limitation of identifying true CNVs and their boundaries [175, 235].

### **1.5** Thesis Objectives and Hypothesis

Copy number gains and losses are shown to be associated with complex human diseases such as developmental abnormalities [26–32] and cancer [33–36] as well as increased susceptibility to several diseases (such as Parkinson's Disease and HIV) [52, 236]. In addition to their role in human

disease, CNVs are also a major source of genome diversity in unaffected human populations [18–20, 22, 22, 24, 25, 66, 69, 161, 194, 237]. It follows that identifying and characterizing these variants are critical to our understanding of genome structure and function and the application of genomics to human disease, including the development of personalized genomic tools.

Based on literature reviews of copy number detection methods and algorithms, it is well accepted that oligonucleotide microarray noise is a major source of false positive and negatives in CNV results, and thus a key factor for underestimation of small CNVs [139, 141, 175, 176]. Considering the current limitations of available CNV calling algorithms (discussed in Section 1.4), it is hypothesized that the current available methods have significant statistical biases that results in lower CNV detection accuracy. This leads to lower sensitivity to detect small aberrations, an issue which has been addressed by several independent groups [141, 238, 239]. A study of global variation in 270 normal human genomes by Redon et al. [69] using Affymetrix 500K SNP array found that on average 206 kb of genome is affected by copy number variations in each individual. Comparing the latter finding with emerging evidence that highlights the prevalence of small CNVs, between 10-100 kb [20, 22, 24, 240] emphasizes that many smaller variants are missed at the current effective resolution of the arrays. An underlying assumption of this thesis is that a portion of these events can be successfully identified if proper statistical methods are used to analyse the array data. The ability to detect such events carries the potential to discover small CNVs that are associated with human disease or predisposition. For instance, it has been shown that small deletions between 70 bp to 7 kb of *MTUS1* tumor suppressor gene are associated with a decreased risk of familial and high-risk breast cancer [238, 241].

The general aim of this thesis was to develop new computational methods to facilitate analysis of Affymetrix SNP microarray data and design a method with improved accuracy for identifying copy number variant regions. In particular, I focused on developing tools that facilitate identification of CNVs based on non-parametric approaches towards improving the quality of SNP array data at the individual oligonucleotide probe-level. To address these challenges, I took advantage of developments in CGH microarray technology and non-parametric statistical methods to develop a novel approach to identify CNVs, and applied this method to study CNVs in follicular lymphoma patients. By developing these non-parametric probe-specific methods, my goal was to improve the accuracy of CNV detection, particularly for smaller events [141].

The main hypotheses of this thesis were as follows: (1) detection of candidate CNVs using current SNP microarray methods is greatly dependent not only on the density of the probes but also the number of probes within the candidate CNVs, and therefore (2) current analysis methods largely underestimate the extent of small CNVs. Furthermore (3) the number of candidate CNVs

reported using current methods is largely dependent on the level of array-wide variation of signal intensities (SD), which yields an increased chance of false positive calls. Finally (4) non-parametric probe-level analysis of SNP arrays allows identification of true CNVs that may be important in disease or disease progression.

## **1.6 Chapter Summaries**

A brief summary of the analysis that I developed and implemented to identify CNVs (Chapters 2 and 3) and two examples of applying this method to identify CNVs in human cancer (follicular lymphoma, Chapter 4) is provided as below.

In **Chapter 2**, I focused on evaluating the variability of hybridization intensity outputs from Affymetrix GeneChip<sup>®</sup> SNP arrays. To perform this analysis, I first assessed the technical variability of SNP arrays by analysing the intensity readouts from 11,564 SNP probes (10K array) in a replicate study consisting of 72 experiments from 8 individuals. The result of this analysis indicated that Affymetrix SNP array technology is highly reproducible ( $CV^1 = 5.16\%$  for chip variability, and CV = 6.3% for labeling variability). Next, I combined statistical theories with the reproducibility measured from empirical data to predict the likelihood of observing 2 or more random probes on the array with k-fold differences in their hybridization intensity ( $k \ge 2$ ). The aim of this chapter was to detect possible sources of variation in SNP array data, and based on the results presented in this chapter, I concluded that the non-specific variation between the performance of individual oligonucleotide probes is a major contributor to the overall microarray noise. The replicate experiment that was used in this Chapter was designed and performed by collaborators at the Affymetrix Company.

The aim of the work presented in **Chapter 3** was to develop and implement an enhanced algorithm to reduce the probe-level variability in SNP array data, in order to assess whether such an approach would provide improved accuracy for detecting CNVs. The specific aims of this chapter were (1) to develop an optimized approach for estimating SNP signal intensity, and (2) to implement an accurate CNV calling model to improve the sensitivity and precision of predicted CNVs. To develop this model, I took advantage of several non-parametric statistical methods that had previously been used in different applications, such as speech signal processing, geology and array CGH technology. The resulting algorithm, called Oligonucleotide Probe-level Analysis of Signal or OPAS, involves two major components: probe-level analysis and SNP-level analysis. I then designed and implemented OPAS visualization software and an associated pipeline that

<sup>&</sup>lt;sup>1</sup>Coefficient of Variability

facilitates automatic sample import and high-throughput sample analysis. An advantage of this approach is that most of the individual modules can be easily extracted from the source code and applied to analyse data from other platforms, for example, Illumina SNP arrays and sequence-based read depth data.

In the pre-processing phase of the algorithm, probe-level analysis is conducted. In this phase, the intensity of each PM oligonucleotide probe (or oligo) is analysed to identify noisy probes and subsequently eliminate them from data analysis. To achieve this goal, within each SNP probe set the PM oligos are categorized into groups or clusters with similar intensity patterns.

To facilitate probe set clustering of PM oligos, I proposed a new clustering approach, referred to as Fuzzy-Kmeans Clustering, based on combining two well-known clustering methods: *k*-means optimization-based and subtractive fuzzy-logic based clustering algorithms. Next, non-parametric KS-test is applied on each determined cluster of oligos (referred to as "oligo cluster") to evaluate the likelihood that the intensity pattern of the PM oligos, that are involved in the oligo cluster, represent a significant shift in the signal intensity. These KS-generated probabilities and the oligo cluster information are then passed to a machine-learning algorithm to identify the most significant oligo cluster(s) within each SNP probe set. Subsequently the mean log2-ratio intensity of the oligos in the most significant oligo cluster(s) is estimated and used to represent the SNP log-ratiometric value.

In the post-processing phase of the algorithm, SNP-level analysis is performed. In this step, I first apply GC-fragment length normalization to minimize the effect of fragment length biases on the estimated SNP readouts through a non-linear LOWESS normalization method. Next, in order to identify regions of the genome with copy number alterations, the algorithm applies Circular Binary Segmentation (CBS) [189, 242] non-parametric CNV calling method on the pre-processed SNP data. It is important to note that, while many of the components of this algorithm had been previously developed in other applications (e.g., CBS was originally designed in aCGH), the adaptation of these methods to SNP microarray data required a largely novel implementation to accommodate a different data type.

To facilitate high-throughput data analysis, I designed and implemented OPAS software that automatically generated a relevant record of the sample analysis. Upon sample import, OPAS software generates separate image and data folders for each sample that are named according to the sample file name (and date of analysis, if it already exists). During sample analysis, OPAS automatically creates a comprehensive catalogue of the graphs and data for each step of the sample processing, from normalization and mean/intensity (MA) plots to visualization of CNVs in each chromosome along the chromosome ideogram, and subsequently saves these images and data in the pre-designated sample folders (see page 95). The purpose of this enhanced sample recording is to provide a useful sample tracking method for future follow ups.

Furthermore, despite the fact that all the components of the OPAS algorithm were based on nonparametric analytical techniques, there were still a few assignable parameters that could affect the algorithm performance. Thus, I extensively tested each module to better assign these parameters and to improve the quality and speed of OPAS. The samples used in the work presented in this Chapter were provided by Dr. Jan Friedman and the "wet-lab" (sample preparation and Affymetrix experiments) was performed at the Genome Sciences Centre. Throughout this Chapter, I also use validated CNV results of mental retardation project by Dr. Friedman that has already been published [29, 30, 215] to test the performance of my proposed method in detecting known CNVs.

In **Chapter 4**, I presented the results of applying OPAS to detect somatic CNVs (present in the tumor but not the matching normal DNA) in 25 follicular lymphoma (FL) patients. Follicular lymphoma (FL) is the most prevalent type of non-Hodgkin lymphoma (NHL)<sup>1</sup> (cancers of the lymph nodes). In Canada, non-Hodgkin lymphomas accounted for about 7,500 new cases of cancer in 2010 (making them the fifth most common cancer) and 3,200 estimated deaths<sup>2</sup>. The statistics also indicate an increasing rate of incidence of NHL among young women aged 20-39 [243]. More importantly, follicular lymphomas frequently transform to a more rapidly progressive invasive and lethal cancer, diffuse large B-cell lymphoma or DLBCL (10 year survival < 20%).

Cytogenetic abnormalities are common characteristics of FL genomes [244–248]. A genetic hallmark of follicular lymphoma is the recurrent chromosomal translocation t(14;18)(q32;q21), which is present in approximately 85-90% of the patients [245, 249, 250]. As a result of this translocation, a part of chromosome 14 involving the enhancer of the immunoglobulin heavy chain (IGH) locus moves to chromosome 18 and into the proximity of the *BCL2* anti-apoptotic gene, resulting in *BCL2* over-expression [245, 249–251]. However, transgenic mice with *BCL2* over-expression do not develop lymphoma [252, 253], and t(14;18) bearing lymphocytes have also been reported in healthy individuals [254, 255]. These findings suggest that t(14;18) alone is not sufficient to produce clinical FL [256–258]. I hypothesized that the OPAS algorithm could improve CNV discovery in FL patients and could enhance our understanding of FL genetics. To test this hypothesis, I looked for candidate somatic aberrations in 25 FL genomes by performing a pairwise analysis of Affymetrix 500K data from tumor matched normal DNA.

In total, I identified 286 somatic CNVs (11.4 per patient) of which 18.5% (53/286) were smaller than 150 kb and 14.3% (41/286) encompassed fewer than 10 SNP probe markers. To assess the ac-

<sup>&</sup>lt;sup>1</sup>http://www.lymphoma.org

<sup>&</sup>lt;sup>2</sup>http://www.cancer.ca

curacy of the putative CNV calls, I compared OPAS findings in 23/25 patients with the results from several alternative technologies and methods that were applied to study the same patients, including Significance Mean Distance method (SMD [259]) based on 500K SNP arrays, BAC aCGH, BAC-end fingerprint profiling (FPP), and Illumina sequencing data. This comparison indicated that 87.9% of all OPAS predicted CNVs are seen by at least one additional dataset, while most of the remaining candidate events had no corresponding FPP or sequence information that could reject them. One of the sequence validated CNVs that was below the detection power of standard SNP calling methods and thus was previously undetected with SNP arrays was a  $\sim 104$  kb somatic deletion on 9p21.3 which included only 8 SNP probes. This deletion harbored the CDKN2A gene which is frequently deleted in many cancers [260, 261], for instance, in an aggressive subset of cutaneous T-cell lymphomas [262]. Another interesting finding was detecting a deletion, approximately 143 kb in length, on chromosome 12q24.11 that encompassed only 4 Nsp SNP probes and was validated using BAC end sequence data. This deletion affected the HVCN1 and PPTC7 genes. A recent publication on the voltage-gated proton channel HVCN1 gene suggested that it modulates the B cell antigen receptor [263]. The deletion of 12q24.11 removes all but the first exon of HVCN1 and the first 3 coding exons of PPTC7, juxtaposing the HVCN1 promoter to the remaining PPTC7 exons. Therefore, this deletion can potentially create a novel fusion gene between the 5' end of HVCN1 and 3' end of PPTC7. Both of the deletions found (CDKN2A and HVCN1-PPTC7) were validated, although neither of these deletions was previously identified when the same 500K data were analysed using the SNP analysis method described in [29, 30]. These findings confirmed that the accuracy of calling CNVs can be improved by performing a probe-level analysis of the array data and applying non-parametric data mining methods to analyse the data.

**Other Advantages:** During the past several years, there has been a major breakthrough in massively parallel sequencing technologies, which offer the promise of detecting whole-genome structural aberrations at base-pair resolution. The two main aspects of my research that I believe are relevant to emerging sequencing technologies are (1) methods that are currently developed for detecting CNVs from sequencing data are in their infancy, and many non-parametric modules that I developed and/or implemented in this work can be easily extracted and modified for detecting CNVs using sequencing data (for example for analysing RD coverage data, Section 1.2.3, p. 10), and (2) during the past several years, thousands of projects have been completed using SNP arrays to study CNVs in thousands of individuals affected with cancers, developmental abnormalities, rare diseases as well as unaffected "normal" individuals. The raw array data files (.CEL) of many of these projects are accessible online (for example HuRef [163] replicate 500K arrays can be

downloaded from ftp://ftp.jcvi.org/pub/data/huref/; see Table 1.3 for more examples). Re-analysis of these data sets with an enhanced algorithm, such as OPAS, can result in discovering small recurrent and potentially pathogenic aberrations, well below the sensitivity thresholds of other algorithms. Considering the relatively low cost of computational array analysis with respect to sample preparation and processing, OPAS re-analysis of the extant data sets will create new discovery opportunities at minimal cost.

Throughout my research, I have generated many tools to facilitate data acquisition, sequence interpretation and statistical inference and visualization of the genomic data. In order to keep my developed tools reusable and traceable, I generated comprehensible Graphical User Interfaces (GUIs) for most of these tools. For instance, DLOH software and GUI were designed to interrogate SNP genotype data in order to identify regions of loss-of-heterozygosity (LOH). It also generates a statistical score for each LOH region to predict potential regions of copy number deletion. I presented DLOH at the Advances in Genome Biology and Technology (AGBT) conference in 2008.

In addition to the work described in this thesis, I have been involved in several other collaborative projects at the Genome Sciences Centre (GSC) which have resulted in publications or presentations. I used non-parametric statistical tests to assess the impact of whole genome amplification on analysis of copy number variants (Pugh et al. [264]). In this work, I compared the distribution of data from paired pre- and post-whole genome amplified (WGA) samples, for 3 separate DNA samples and identified apparent WGA-induced over- and under-amplifications in each of the three comparisons of amplified versus unamplified material.

I also collaborated with Dr. Maziar Rahmani in analysing data and preparing genome wide visualization of markers in several genome wide association studies (GWAS). Based on my results several novel risk loci were chosen in patients with calcific aortic valve stenosis, which were presented at the Canadian Human Genetics Conference [265] and National Research Forum for Young Investigators in Circulatory and Respiratory Health [265]. More recently, a large-scale study has been initiated by Dr. Rahmani at the Genome Sciences Centre and I will apply the same approach to identify potentially significant markers which will be then genotyped for validation and presented in a future manuscript.

In conclusion, whole-genome analysis techniques provided us with the means to understand the extent of the variation in normal and disease genomes. Nonetheless, the past several years have yielded rapid developments in sequencing technology, creating a field of investigation that is transforming both our concept of the human genome and the application to clinical practice. I believe that merging the vast amount of data available from microarrays and recent sequence based studies carries the promise of understanding the mechanisms by which disease and normal genomes diverge. However, gaining this knowledge is largely dependent on the use of improved computational methods that can provide accurate analysis of these data.

## 1.7 Figures and Tables



(a) Karyotype of an FL patient (48,XY,+X)



(b) M-FISH of the same patient

**Figure 1.1: Karyotype and M-FISH chromosome analysis of a patient with follicular lymphoma (FL).** The karyotype and M-FISH analysis of an FL patient is shown in panels (a) and (b), respectively. Both techniques indicate several microscopically detectable chromosomal alterations in this patient, including extra copies of chromosomes 7 and X (48,XY,+X, +7). Karyotype and M-FISH results can also detect balanced translocation, such as t(14;18)(q32;q21) in the above FL patient (figures provided by Dr. Horsman's lab at the BCCRC).



Figure 1.2: Structure of SNP probe sets in Affymetrix GeneChip<sup>®</sup> SNP arrays. In Affymetrix technology, each SNP on the array is represented by a collection of probe quartets, also known as a SNP probe set. A probe quartet consists of a set of 25-mer oligonucleotide "probe pairs" for two most common alleles (known as 'A' and 'B') and for both forward and reverse strands (antisense and sense) for the SNPs. Each probe pair consists a perfect match (PM) probe and a mismatch (MM) probe. The Affymetrix chip design strategy is to use a set of these PM/MM probe pairs to interrogate the surrounding bases of SNPs for the forward and or reverse target for both alleles. At the top of the above figure, the 'A' and 'B' alleles of a given SNP are shown by pink and yellow beads in the middle of the green sequence. The 4 rectangles below the depicted sequence, represent the probe quartets for forward and reverse strands of allele 'A' (red rectangles) and allele 'B' (yellow rectangles). As shown in this figure by blue beads ('G' and 'A'), the sequence of MM probes is the same as the PM probes, except for a single nucleotide difference. Furthermore, within a probe set there are additional probes to interrogate the neighboring bases of the SNP locus, known as "offset probes". The position of the probes within a probe set is typically denoted by (-4, -2, -1, 0, +1, +2, +4), where the integer *n* refers to the base position being interrogated by the offset probes relative to the SNP position (n = 0). The bottom panel illustrates two offset probes (n = -1 and n = +4) of a given SNP sequence. The number of designated oligonucleotide probes in a probe set varies in each generation of the Affymetrix GeneChip SNP arrays. In 10K and 100K arrays, a probe set consists of 40 PM/MM oligonucleotide probes. Thus, for example, the 100K SNP array has in total more than 4.5 million features (oligonucleotide probes) on the array. The Nsp (~262,000 SNPs) and Sty (~238,000 SNPs) arrays from Affymetrix 500K SNP dual array-set have a density of more than 500,000 SNP sites with 20 or 40 PM/MM probe pairs per SNP locus, resulting to > 3 million features on each array and a total of ~6,071,040 features.

Name of Abnormality	Туре	Chromosomal Variation	Cytogenetics Representation	
Turner syndrome	aneuploidy	loss of entire X chromosome	(45, X) or (45, X0)	
Klinefelter syndrome*	aneuploidy	gain of entire X chromosome	(47, XXY)	
Edwards syndrome	aneuploidy	trisomy of chromosome 18	(47, +18)	
Down syndrome	aneuploidy	trisomy of chromosome 21	(47, +21)	
Patau syndrome	aneuploidy	trisomy of chromosome 13	(47, +13)	
Cri du chat	large-scale CNV	loss of the short arm of chromosome 5	46,XX,del(5)(p15.2)	
1p36 Deletion syndrome	large-scale CNV	loss of a region on the short arm of chromosome 1	deletions on 1p36	
Angelman syndrome	large-scale CNV	loss of $\sim 4$ Mb of the long arm of chromosome $15^{\dagger}$	46,XX,del(15)(q11-q13)	

Table 1.1: Aneuploidies and large-scale CNVs associated with human disease

\*the most common male chromosomal disease

<sup>†</sup>observed in 50% of patients with Angelman syndrome

Fable 1.2: Comparison of array specifications	s in 4 generations of Affymetrix SNP array
-----------------------------------------------	--------------------------------------------

	Enzymes	SNPs	Probe Pairs	Quartets	Features	Median IMD* (kb)	Average IMD (kb)	Coverage
10 K	XbaI	10,204	40	14	647,080	113	258	at least 1 SNP per 100 kb.
100K	XbaI, HindIII	116,204*	40	10	4,648,160	8.5	23.6	92% of genome within 100 kb of a SNP; 40% within 10 kb of a SNP.
500K	NspI, StyI	500,568 <sup>†</sup>	24, 40	6-10	12,013,632	2.5	5.8	85% of genome within 10 kb of a SNP.
SNP 6.0	NspI/StyI	<ul> <li>SNP 6.0 has major differences with the three previous generations of Affymetrix SNP arrays:</li> <li>combines the NspI and StyI fractions that were previously assayed on two separate arrays</li> <li>contains 906,600 SNP probes in addition to 945,826 non-polymorphic copy number (CN) variation probes:</li> <li>(~744,000 were selected for their spacing and ~202,000 were based on known copy number changes<sup>‡</sup>)</li> <li>3-4 replicate perfect match (PM) probes per SNP probe (CN probes have no replicate)</li> <li>median inter marker distance = 2180 bp (CN), 1270 bp (SNP), 680 bp (SNP + CN)</li> <li>average inter marker distance = 3160 bp (CN), 3230 bp (SNP), 1600 bp (SNP + CN)</li> </ul>						

\*IMD: Inter Marker Distance

\*\*There are 58,960 SNP probes on XbaI and 57,244 on HindIII arrays.

 $^\dagger There are {\sim}262{,}000$  SNP probes on NspI and  ${\sim}238{,}000$  on StyI arrays.

<sup>‡</sup>based on Toronto Database of Genomic Variants (DGV) [168]

## Table 1.3: A partial list of Affymetrix SNP array data (raw .CEL files) that are publicly available

270 samples from HapMap project	http://hapmap.ncbi.nlm.nih.gov/downloads/ raw_data/affy500k/
9 Tumor/Normal pairs derived from several human cancer cell lines, including adeno- carcinoma, non-small cell lung carcinoma and primary ductal carcinoma	http://www.affymetrix.com/support/technical/ sample_data/copy_number_data.affx)
GlaxoSmithKline (GSK) has made available the 500K SNP data for over 300 cancer cell lines from 30 different tissue types in a wide range of cancers including small cell lung carcinoma, neuroblastoma, lymphoma and glioblastoma	https://cabig.nci.nih.gov/caArray_GSKdata/
Marshall et al. (2008; [50]) provided the 500K .CEL files of 1318 individuals with Autism Spectrum Disorder (ASD)	http://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE9222
Chiang et al. (2009; [266]) provided the Mapping250K-Sty array data for 77 repli- cates of HCC1143 (breast ductal carcinoma), 69 replicates of HCC1143BL (matched normal), 42 replicates of HCC1954 (breast ductal carcinoma), 36 replicates of HCC1954BL (matched normal) and 1 replicate of NCI-H2347 (lung adenocarcinoma)	http://www.broadinstitute.org/cgi-bin/ cancer/publications/pub_paper.cgi?mode= view&paper_id=182
Solomon et al. (2008; [267]) provided Affymetrix Mapping250K-Nsp array data for 58 glioblastoma multiforme tumor samples	http://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE13021
161 primary breast cancer samples (500K array) reported by Kadota et al. (2009; [268])	http://www.ncbi.nlm.nih.gov/geo/query/acc. cgi?acc=GSE16619
141 gliomas and 33 normal tissue samples (100K array), used by Beroukhim et al. (2007; [269])	http://www.broadinstitute.org/cgi-bin/ cancer/publications/pub_paper.cgi?mode= view&paper_id=162&p=t
768 Affymetrix Mapping250K-Sty data for 384 tumor/normal lung adenocarcinoma pairs for Weir et al. (2007; [270])	http://www.broadinstitute.org/cancer/pub/ tsp/

## Chapter 2

# Analysing Variability in Microarray Data

## 2.1 Introduction

Oligonucleotide microarrays are powerful tools that enable high-throughput measurement of DNA copy number alterations across the entire genome. One application of oligonucleotide microarrays is copy number profiling, which has led to significant advances in the understanding of complex diseases and discovering submicroscopic aberrations in genes that are associated with or causative to diseases, such as mental retardation [271, 272] and cancer [273].

A major concern for the identification of CNVs using oligonucleotide microarray technology is how a putative copy number change is defined. There is a plethora of different methods that have been used to call changes in DNA copy number values from relative ratio intensity outputs of these arrays, ranging from simple preset thresholds [213] to complex statistical modelling [212]. Despite their differences, the fundamental principle underlying these approaches often include one (or a combination) of hidden Markov models (HMMs) [185, 186], segmentation algorithms [189, 210, 242], or t-tests and standard deviations (SDs) of the log2-ratio intensities (referred to as LR in this thesis) [212]. Regardless of how the data are analysed, nonspecific variations that are due to assay variability will mitigate the accuracy of the downstream CNV results.

One drawback of such non-specific variation is that the incorporated noise forces the measured signal intensities to pass over a predefined threshold of CNV calling, leading to false positive and false negative CNV results. To circumvent the false positive problem, algorithms often apply

more stringent CNV calling criteria (for example, by raising the LR thresholds of calling significant deviations from the base-line and/or increasing the minimum number of consecutively shifted SNPs for calling CNVs). This, on the other hand, mitigates the ability to identify other real aberrations that do not necessarily meet more stringent constraints, and consequently leads to increased false negative rate.

Therefore, despite the importance of technological advances in microarray platform design and array processing during the past decade, it has become increasingly clear that the capability to assess variability associated with array outputs is of paramount significance and is an essential factor for developing analytical tools that can maximize the utility of these platforms [139, 274]. Nonetheless, investigating variability and its possible sources have been largely unexplored in SNP arrays [140, 233, 275], unlike expression arrays where a great effort has been invested to understand the sources of variability among probes for each transcript [207, 234, 276–281]. As a result, the impact of oligonucleotide probe-level variability on SNP data reproducibility that directly influences CNV analysis and interpretation has also been largely unexplored.

To address this issue, in this Chapter I explore different sources of variability in Affymetrix SNP arrays. The specific aims of the work presented in this chapter are: (1) to estimate relative magnitudes of different sources of variation in Affymetrix SNP array data, and (2) to assess the variability associated with each probe set by modelling variance as a function of intensity. To perform these analyses, I first examine common causes of variation among array results by assessing the reproducibility of Affymetrix GeneChip<sup>®</sup> 10K SNP array platform using a replicate data set consisting of 69 arrays from 8 individuals. Next, I present a mathematical model to study the relationship between the empirical variability (%CV) and the theoretical fraction of individual oligos that are expected to differ in their log2-ratio intensity readouts by at least some given factor. The aim of this model is to incorporate the theoretical probabilities and the empirical variabilities (%CV) to determine the acceptable range of variability across replicate oligos on 10K SNP arrays.

In the rest of this Chapter, I first explain some of the major sources of variability in microarray data and then discuss the methods I use to analyse variability in Affymetrix SNP arrays and the corresponding results.

#### 2.1.1 Variability in Microarray Data

A typical microarray experiment, regardless of the array platform used, has many different sources of variation [204, 282, 283]. Figure 2.1 illustrates some of the common sources of variability in Affymetrix SNP genotyping arrays [204, 282]. As seen in Figure 2.1, the sources of variation

detected by microarrays can be broadly attributed to biological and technical causes. These sources of variability are briefly explained in the following sections (2.1.2-2.1.3).

#### 2.1.2 Biological Variability

At the highest level of variation hierarchy in Figure 2.1 is the population variability or biological variability, a well-known source of variation that exists among normal individuals. Such variability is independent of the microarray experimental process [204]. This type of intrinsic copy number variation among individuals, known as copy number polymorphisms or CNPs, is responsible for a substantial amount of human phenotypic variation and genome diversity [19, 66, 69, 163, 284–287]. Other studies of CNPs have also discovered several associations between CNPs affecting genes and biological functions related with immunity [46, 288, 289], sensory reception [290, 291] and other phenotypes, such as predisposition to HIV infection [236], and susceptibility to Crohn's disease [47]. Early studies suggested that about 12-18% of the human genome is involved in copy number polymorphisms [19, 69]. However, more recently, several independent large-scale studies of CNPs have consistently concluded that the frequency and size of these variations were largely overestimated in the initial studies [66, 168]. Nonetheless, these biological variations are a major source of variability among normal individuals and a fascinating field of population genetics [292].

#### 2.1.3 Technical Variability

During microarray experiments, many factors can lead to unwanted variation or noise in the generated data that are commonly referred to as experimental or technical variation. Such variation in array data affects our ability to identify real copy number changes in the downstream analysis. Technical variability is, therefore, fundamentally different from biological variability, which is an indicator of genetic diversity (Sec. 2.1.2). Multiple sources of variation in a microarray experiment can impact the overall experimental variability, including array manufacturing process, DNA or RNA isolation method, sample preparation, target labeling, hybridization and scanning. These sources of technical variability are not platform-specific and are relevant to all commercially available microarrays, albeit to a different degree [204, 293–297].

While the choice of the array platform used can affect the quality of obtained results [205, 298, 299], the ability to compare findings within the same platform requires systematic assessment of technical variability. An accurate estimation of different sources of technical variability is essential not only for understanding how well a microarray platform performs, but also for calibrating the array output signal and improving the sensitivity and specificity of the findings. An extensive

amount of research has been carried out in the past decade to discover, understand and quantify unwanted sources of technical variation. These studies have motivated the development of many advanced computational techniques for microarray data normalization and filtering [180, 205, 296, 300–302]. Figure 2.2 depicts some of the major components of technical variation in microarrays (applicable to both expression and SNP genotyping arrays). The first component depicted in this figure is referred to as "the variation between Cy5 and Cy3 colors" which is specific to two-color arrays such as CGH-based platforms. In such arrays two samples (e.g., "test" and "reference") are labeled with different fluorophores (usually cyanine-3 (Cy3) and cyanine-5 (Cy5)) and hybridized together on a single microarray, and thus the difference between the two dies can affect the integrity of the resultant readouts (refer to Chapter 1 for more details). Such variation does not exist in onecolor arrays, such as Affymetrix SNP genotyping chips, where in each experiment only a single sample is hybridized to an array after it has been labeled with a single fluorophore (such as Cy3 or Cy5). This implies that one-color arrays, such as Affymetrix SNP chips, are impacted by fewer sources of variability compared to two-color arrays.

The next source of variability denoted in Figure 2.2 is labelling variability which is due to the difference in sample labelling reactions. Such variability has also been addressed in other studies as the difference in the labelling efficiency across multiple arrays. However, in this Chapter labelling variability refers to the technical variation that is due to both sample preparation and labelling process. Another aspect of technical variation, commonly observed in large-scale studies, occurs when samples are divided into different groups and each group is processed independently. Such variation, commonly referred to as batch effects [303–308], has been observed from the earliest microarray experiments [309] and can be caused by many factors including the batch of amplification reagent used or the hybridization reaction. In contrast to random non-specific variation (noise) in microarrays, batch effects exclusively describe systematic technical differences when samples are processed and measured in different batches [303, 307]. In large scale studies, practical considerations limit the number of samples that can be amplified and hybridized at one time, so samples may be generated several days or months apart and, therefore, batch effects will inevitably confound the results of such large-scale studies. The batch effects can potentially mask the real biological events, particularly when the average variation between different batches is significantly larger than the biological variation within each batch [307, 308]. Thus, proper monitoring and managing of batch effects is essential for extracting relevant biological information [303, 307, 308].

The variability between hybridization to different arrays, referred to as chip variability, is another major source of technical variability that is illustrated in Figure 2.2. This variability represents the underlying chip-to-chip variation of the intensity readouts between multiple arrays (of the same platform). As shown in Figure 2.2, chip variability is intrinsically dependent on two other components of technical variability: (1) variability between arrays in relation to the manufacturing process (also referred to as manufacturing variability), and (2) variability between different hybridization reactions. As the result of intercorrelation between different sources of microarray variability, as illustrated in Figure 2.2, measuring the independent contribution of each sources of technical variability is a very difficult, if not impossible, task.

The rationale of the work presented in Chapters 2 and 3 is that an experimental design that takes into account the impact of experimental variability at the level of individual oligonucleotide probes can identify and exclude the noisy oligos and minimize the impact of noise in the downstream CNV detection process. In the next section (Section 2.2), I will discuss some of the methods that were used to estimate two major components of variability, chip and labelling variability, among replicate Affymetrix SNP arrays (described in Sections 2.3.3-2.3.4).

## 2.2 Methods for Measuring and Quantifying Microarray Variability

A common approach for measuring microarray data variability, regardless of the platform, is to perform replicate studies. Microarray experiments can be replicated at biological and technical levels to assess the two main categories of variation which were discussed in the previous sections (Sections 2.1.2-2.1.3).

**Biological replicates:** are replicates taken at the level of the population being studied (e.g., copy number polymorphism among normal individuals). Such replicate data sets include samples that are independently obtained from replicate sources (such as multiple cell lines, multiple biopsies or multiple patients). The purpose of studying biological replicates in copy number studies is to evaluate the extent of DNA copy number diversity among normal individuals [19, 66, 69, 163, 284–286, 310], and to measure the possible functional or phenotypical implications of such variations on normal individuals [46, 47, 236, 288–291].

**Technical replicates:** are replicates generated at the level of the experimental process [204]. The purpose of studying technical replicates in Affymetrix SNP arrays is to evaluate technical variability in the generated log2-ratio intensity readouts from a SNP array, which can directly affect the consistency and reliability of both genotyping and CNV results in the downstream data analysis. As discussed in detail in Section 2.1.3, technical variability consists of several further components, such as labeling and chip variability. Estimating these aspects of technical variation and their impact on the data integrity is an important task for evaluating the performance of any microarray experiment.

The focus of the remainder of this Chapter is to present the methods that were used and/or developed to measure technical variability of Affymetrix SNP arrays and the results of applying these methods to real data from a replicate SNP array experiment.

#### 2.2.1 Quantifying Technical Variability in Affymetrix SNP Arrays

Quantifying the levels of experimental (or technical) variability is a common practice before undertaking any large-scale microarray project [204]. Such variability is often evaluated by performing replicate experiments that aim to assess the systematic reproducibility of a particular technology (as explained in Section 2.2) [204]. Below is the description of some of the main computational aspects of analysing and quantifying variability in microarray experiments. It is important to note that these general methods are not platform-specific and can be applied to a variety of different microarrays, such as Affymetrix and Illumina SNP chips or expression arrays [301, 306].

#### 2.2.1.1 Log-transformation

Often the first step in estimating technical variability in any microarray platform, including Affymetrix SNP arrays, is to transform raw intensity data of the replicate study into log intensities [204, 311, 312]. It has been shown that generally in all microarray experiments larger intensities tend to have larger variations [301, 313, 314]. This bias leads to inconsistent variance across a measured range of intensity. Such inconsistency in variance, which is also described in statistics as "heteroskedasticity", imposes a serious challenge for analysing variability in any application, including microarray experiments [315, 316]. One of the primary reasons for performing log-transformation is that it circumvents the former issue by converting asymmetric distribution of (raw) intensities to a symmetric and Gaussian-like distribution, as illustrated in the example of Figure 2.3. Furthermore, the latter transformation would enable us to apply powerful statistical methods to stabilize variance or measure the underlying variability in a microarray experiment (see the following Section) [301, 311, 312].

#### 2.2.1.2 Coefficient of Variation (CV)

In statistics, it is common to represent the variability in a data set by evaluating the standard deviation (SD) of the data. In microarray applications, however, it is well-known that the standard deviation of intensities is positively correlated with the mean signal intensity of the array [204, 317]. If variance is proportional to raw signal intensity of the data, the application of log transformation produces a constant variance across the range of signal intensities on the logarithm scale. The most important advantage of log-normal assumption is that it allows different levels of variability to be expressed as a percentage known as the "coefficient of variability" (CV) [204], defined by:

$$CV = \frac{\sigma}{\mu} \times 100 \tag{2.2.1}$$

where  $\sigma$  and  $\mu$  represent the standard deviation and mean of the input signal, respectively. Based on Equation (2.2.1), the CV is estimated by finding the ratio of standard deviation of the signal to its mean and serves as an indicator of data variability.

A key advantage of using coefficient of variation (CV) in SNP arrays, instead of simple standard deviation (SD), is that the standard deviation (or variance) of microarrays are generally proportional to the mean of signal intensities. Therefore, dividing the standard deviation ( $\sigma$ ) by mean ( $\mu$ ), as shown in Equation (2.2.1), removes the intensity-specific dependencies of the estimated variability. This feature makes CV particularly useful for quantifying the variability in replicate microarray studies. Based on Equation (2.2.1), a lower value of CV represents a higher microarray reproducibility. However, the acceptable range of CV varies between different studies and groups [318–321]. For example, the Institute of Food and Research<sup>1</sup> (IRF) microarray facility in the UK uses median CV of 5% and 10% as the critical values for technical and biological replicates [321]; and MicroArray Quality Control (MAQC) project has reported 5-20% CV for six<sup>2</sup> commercially available microarray platforms for gene expression analysis [318].

Based on theoretical statistics, it is reasonable to assume that if the ratio of mean to standard deviation in a normal distribution is  $\geq 3$  the experiment is not reproducible. By using this hypothesis Johnson and Welch et al. had previously stated that 33% CV constitutes a permissible upper limit of CV, implying that any experiment with CV > 33% is not reproducible [322]. The main reason behind the debate about the critical value of the CV is that there is no fundamental analysis that can relate the assay variability with the reliability (or precision) of the biological interpretation of array findings.

Understanding the link between the measured variability and its impact on the expected frequency of inconsistent readouts is crucial to determine the acceptable extent of technical variation in microarray experiments. Nonetheless, no previous study has explored the relationship between the technical variation and the expected frequency of erroneous readouts in CNP arrays. To address this limitation, the focus of the rest of this Chapter is to adapt the previous studies of CV and variability in other applications [279, 323] to develop a comprehensive model that can relate oligo-

<sup>&</sup>lt;sup>1</sup>http://www.ifr.ac.uk/safety/microarrays/

<sup>&</sup>lt;sup>2</sup>Applied Biosystems; Affymetrix; Agilent Technologies; GE Healthcare; Illumina, and Eppendorf.

specific variability in Affymetrix SNP arrays (as measured by the CV) to the frequency of potential SNP-level errors. Such detailed analysis of variability and its impact on SNP data quality allows us to determine the critical value of CV that defines the acceptable range of technical variation in SNP microarrays.

### 2.2.2 A Link Between the CV and the Probability of Observing k-fold Disparities Between Replicate Measurements

Two studies by Wood et al. [323] and Reed et al. [279] had previously focused on developing a link between the estimated variation in replicate measurements and the expected fraction of pairs of those measurements that differ by a given factor, in the context of serological assays. Wood et al. [323] showed a mathematical relationship between the error frequency and the magnitude of the SD, under the assumption that the logarithm of measurements is normally distributed and based on using a maximum acceptable variability of 2 fold change (k = 2). Alternatively, Reed et al. [279] extended Wood's approach and derived a mathematical relationship between the CV and the expected frequency of any k-fold disparate results in serological assays. However, there is no current publication that links the estimated variation in SNP arrays to their performance and error rates, similar to Wood et al. [323] treatment of SD or Reed et al. [279] treatment of CV in serological assays. To address this limitation, I used Reed's model [279] to correlate the extent of oligo-specific CV measurements that have k-fold difference in their signal intensity readouts in a replicate Affymetrix SNP array experiment. I then further expanded this model and used empirical results of CV to estimate the effect of PM oligo-specific variability on the measured SNP log2-ratio (LR) values (described in Section 2.3.5).

According to Reed's model [279], in a given assay with log-normal data distribution and with a known value of CV, the likelihood that two replicate measurements differ by at least a factor of k (p(k)) is estimated by Equation (C.9), restated here:

$$p(k) = 2\Phi\left[\frac{-\log(k)}{\sqrt{2\log_e(CV^2 + 1)}}\right]$$
(2.2.2)

where  $\Phi$  denotes the cumulative density function (CDF) of "standard normal distribution" (see Appendix A and B for more detail). The complete mathematical proof of this equation is presented in Appendix C. In the context of SNP arrays, the knowledge of this probability (p(k)) helps to determine what magnitudes of difference between oligos can be expected by chance alone when a particular coefficient of variation (%CV) is in effect. This information plays a pivotal role in understanding whether an estimated CV represents a reproducible or non-reproducible microarray experiment.

It must be added that the CV is a common standard for assessing microarray data reproducibility and has been used for microarray platform comparisons in MicroArray Quality Control (MAQC) project and several other studies [318, 324, 325]. Nonetheless, Analysis of Variance (ANOVA) is another statistical approach for assessing microarray reproducibility. The description of this method and the results of its application on Affymetrix SNP array data are detailed in Appendix D.

## 2.3 Results

To estimate the technical variability of Affymetrix SNP arrays, a replicate study was performed using 69 samples studied on Affymetrix GeneChip<sup>®</sup> 10K SNP arrays. It is important to note that the aforementioned experiment was carried out at the Affymetrix company (part of a collaboration between Affymetrix and the Genome Sciences Centre). My role in this study was to apply computational techniques to quantify variability in the data generated from this experiment and to determine if Affymetrix 10K SNP arrays were reproducible. The results of my analyses are explained in the following sections.

#### 2.3.1 Affymetrix 10K Replicate Experiment

A schematic representation of the replicate experiment is shown in Figures 2.4-2.5. For this experiment, 8 individuals with normal karyotypes were selected by our collaborators at Affymetrix. The DNA from each subject was divided into 3 batches of approximately 750 ng. Each batch was labeled and hybridized to 3 Affymetrix SNP chips (10K), providing a total of 9 replicate arrays for each sample (Figure 2.4). The only exception in this design was sample #1 for which only 6 replicate arrays were available (2 chips per labelled batch). For each subject in this study, sample processing and hybridization were performed according to the Affymetrix GeneChip<sup>®</sup> 10K SNP assay [326], resulting in 69 arrays for the entire replicate data set. The .CEL files for these 69 arrays were obtained from Affymetrix for further analysis of technical variation. These .CEL files contain the raw intensity readouts of all (~462,400) oligonucleotide probes on the 10K chips. This data set is referred to as "10K rep-test" in the remainder of this chapter.

#### 2.3.2 Quantifying Technical Variability in Affymetrix SNP Arrays

In Section 2.2 several computational aspects of evaluating technical variability in replicate microarray data were discussed. Below is the summary of the steps that were performed to quantify technical variability in "10K rep-test" analysis (adapted and modified from Stekel et al. [204]).

- 1. Log-transform raw probe intensities (obtained from .CEL files) of the 69 arrays in "10K rep-test". Here, I have used natural logarithms for data transformation (see Figure 2.3).
- 2. Apply global normalization on the arrays by bringing the overall mean intensity of the SNP arrays to the same level (see Figure 2.6).
- 3. Estimate the mean log-intensity value of each oligonucleotide probe feature among the replicate arrays<sup>1</sup> (10K SNP array has ~462,400 oligos). This value is commonly known as *A* in the context of MA-plots, which will be discussed in Step 5.
- 4. For each oligo in the replicate arrays, estimate the deviation from the mean. This is also known as the error between replicate features [204, 205], and is typically denoted by *M* in MA-plots (see Step 5).
- 5. Generate MA-plots [204, 205] to visualize the relationship between the estimated error (or *M*, obtained in Step 4) versus mean log-intensities (or *A*, measured in Step 3). The MA-plots help to examine whether the magnitude of the error is independent from the signal intensity (see Figure 2.7).
- 6. If the data in MA-plot are not symmetrical around a horizontal line, it implies that the error is reliant on intensity and based on the shape of this plot linear normalization (e.g., linear regression) or non-linear normalization (e.g., LOWESS) techniques can be applied to correct for the error biases [205]. Otherwise, proceed to the next step (Step 7).
- 7. Calculate the standard deviation of error distribution ( $\sigma_e$ ). If the MA-plot suggests that the variation is dependent on the signal intensity (asymmetrical MA shape), the data can be partitioned into subsets with different intensity ranges so that the oligo readouts within each

<sup>&</sup>lt;sup>1</sup>In this algorithm, the term replicate arrays refers to the chips that are compared together to determine chip or labelling variability for a particular sample. For example, as seen in Figure 2.5 the first DNA batch (b = 1) of individual #5 (s = 5) is hybridized to 3 separate chips (c1, c2 and c3). Therefore, to measure chip variability of this particular batch (s = 5, b = 1), the intensity hybridization between c1,c2 and c3 must be compared. These 3 arrays are, therefore, referred to as replicate arrays for analysing chip variability for the above DNA batch (s = 5, b = 1).

subset are intensity-independent. In such circumstances, a separate  $\sigma_e$  must be evaluated for each partition, independently.

- 8. A common model for errors in microarray experiments is the log-normal assumption which assumes that the deviations in replicate experiments follow a normal distribution in the log-scale. This can be verified by plotting a histogram of the estimated deviation values (see Figure 2.8).
- 9. One advantage of the log-normal model of errors is that the coefficient of variation (CV) relates to the standard deviation of the errors in the log-scale ( $\sigma_e$ ) by the following formula (solved in Appendix C):

$$CV = \sqrt{e^{\sigma^2} - 1}$$
(2.3.1)

If log-normal assumption is validated, evaluate the coefficient of variation (CV) by substituting  $\sigma_e$ 's (obtained in Step 7) in Equation (2.3.1). For dataset with multiple partitions, as described in Step 7, estimate one CV for each partition, independently.

The above algorithm is a general method and can be used to estimate variability in the data from other biological platforms, when applicable.

As shown in Figure 2.3, the log-transformed intensity readouts of PM oligos from "10K reptest" arrays follow a normal distribution. This finding is consistent across all 69 arrays in this dataset (each image has been inspected separately, but to avoid redundancy only a subset of the images are shown in Figure 2.3). Next, according to Step 2, I applied global normalization to bring the overall (mean) intensity in these arrays to the same level, as shown in Figure 2.6. In the subsequent step (Step 5), sample-specific MA-plots of deviation of oligo-level hybridization intensities were generated. As seen in the MA-plots depicted in Figure 2.7, the symmetrical shape of the data indicate that the oligo-level errors (deviations) in these arrays are mainly independent from their mean intensity values. This finding is consistent across all other samples in "10K reptest" dataset, suggesting that there is no need to perform any further sophisticated normalization technique that would have been otherwise required to remove intensity-dependent deviations (as described in Step 6). Next, according to Step 7, I estimated the standard deviation of the errors ( $\sigma_e$ ) and then plotted the histograms of these errors to investigate whether the log-normal assumption was true in the "10K rep-test" dataset. As shown in Figure 2.8, the histogram of chip-specific deviations for all 8 samples in "10K rep-test" dataset followed a normal distribution, and this observation was consistent across all samples for both chip and labelling variabilities. The latter finding confirmed the log-normal assumption of the samples in "10K rep-test" dataset.

#### 2.3.3 Assessing Chip Variability in the Replicate Dataset (10K)

As shown in Figure 2.5, each DNA sample in the "10K rep-test" study was hybridized to a total of 9 separate 10K chips (except for sample #1 which had 6 replicate arrays). To measure chip variability in "10K rep-test" dataset, the hybridization intensities between these chips were analysed for each DNA sample, independently. To estimate chip variability for each individual oligonucleotide probe on the 10K SNP array, the mean (A) and deviation from the mean (M) were calculated based on the following formulas:

$$A_p(s,B) = \frac{1}{3} \sum_{i=1}^{3} \log I_p(s,B,\mathbf{C_i})$$
 (2.3.2a)

$$M_p(s,B,C) = \log I_p(s,B,C) - A_p(s,B)$$
 (2.3.2b)

where  $s \in \{1,...,8\}$  corresponds to the DNA sample number,  $B \in \{1,2,3\}$  represents the index of the same DNA batch, and  $C_i$ , denotes the *i*-th replicate chip of a particular DNA batch  $(i \in \{1,2,3\}^1)$ . Also,  $I_p(s,B,C_i)$  represents the intensity of oligonucleotide probe p in the *i*-th replicate array ( $C_i$ ) of a particular DNA batch (B) of sample s. For example,  $I_{100}(5,1,C_3)$  indicates the intensity of the 100-th oligo of the 3rd replicate chip (c = 3) of the first batch of sample #5 (b = 1, s = 5; see left panel of Figure 2.5). Based on the above definitions, Equation (2.3.2b) evaluates the mean (A) of oligonucleotide probe-level signal intensity ( $I_p$ ) of the same DNA batch (c = 3) of the first batch of sample #5 (d = 1, s = 5; see left panel of Figure 2.5). Based on the above definitions, Equation (2.3.2b) evaluates the mean (A) of oligonucleotide probe-level signal intensity ( $I_p$ ) of the same DNA batch across replicate arrays, and Equation (2.3.2b) uses the estimated mean value (A) to assess the error (deviation; M) of chip variability across replicate arrays.

As described in page 39, a separate MA-plot was generated for each DNA batch by using the evaluated oligo-specific M and A values. The aim of this plot was to assess whether oligo-level variations were dependent on the average intensity across replicate arrays. These values were then used in Equation (2.3.1) to assess the %CV of chip variability for all 8 samples of "10K rep-test" dataset, generating 3 estimates of CV per sample (one for each DNA batch) and a total of 24 CV's for the entire dataset. As depicted in Figure 2.9a the estimates of  $CV_l$  varied between 4 to 7% among all samples of the "10K rep-test" dataset, with the mean chip variability ( $CV_c$ ) of 5.16%.

#### 2.3.4 Assessing Labeling Variability in the Replicate Dataset

To assess labelling variability, it is necessary to measure the difference in hybridization intensity of different labelling reactions of the same DNA sample. Similar to assessing chip variability, log-

<sup>&</sup>lt;sup>1</sup> with the exception of sample #1 (s = 1), where  $i \in \{1, 2\}$ .

transformation, global normalization and other steps described in Section 2.3.2 were applied on the replicate arrays to evaluate the variability that was associated with labeling reactions.

According to the experimental design of the "10K rep-test" dataset and as seen in Figure 2.5, each DNA sample was hybridized on 9 SNP arrays (with the exception of sample #1 which was applied to 6 arrays). Considering that the aim of this analysis was to find the signal intensity deviation between different labelling reactions, I first estimated the average intensity for each batch and then compared the average batch intensities across the 3 labelling reactions, as illustrated in Figure 2.5. The formula for M and A for analysing labeling variability is given by:

$$A_p(s) = \frac{1}{9} \sum_{j=1}^{3} \sum_{i=1}^{3} \log I_p(s, \mathbf{B_j}, \mathbf{C_i})$$
(2.3.3a)

$$M_{p}(s,B) = \bar{I}_{p}(s,B) - A_{p}(s)$$
(2.3.3b)  
$$\bar{I}_{p}(s,B) = \frac{1}{3} \sum_{i=1}^{3} \log I_{p}(s,B,\mathbf{C}_{i})$$

where  $s \in \{1, ..., 8\}$  represents the sample number,  $B_j$ ,  $j \in \{1, 2, 3\}$ , denotes the batch number of a specified sample (s), and  $C_i \in \{1, 2, 3\}$  indicates the replicate chip for each batch. Also, as described before,  $I_p(s, B, C)$  corresponds to the intensity of a given oligonucleotide probe p in chip (C) which is associated with batch B of sample s (For a detailed description of these parameters see Section 2.3.3, p. 41). In Equation (2.3.3b),  $M_p(s, B)$  denotes the deviation of the average signal intensity of probe  $p(\bar{I}_p(s,B))$  in the specified DNA batch (s,B) from the mean intensity of the same probe  $(A_p(s))$  across the entire set of replicate arrays of the same sample (evaluated by Eq. (2.3.3a)). Next, the standard deviation of the estimated errors associated with labelling variability ( $\sigma_e$ ) were measured and, subsequently, the coefficients of variability for different labelling reactions  $(CV_l)$ were generated based on Equation (2.3.1) (Section 2.3.2). Figure 2.9b illustrates the  $CV_l$  results in the "10K rep-test" dataset. This Figure indicates that  $CV_l$  varies between 6-7% across the 8 DNA samples in this study, with an average  $CV_l$  of 6.36% per sample. Here, we observe that the overall estimate of labelling variability is slightly larger than the chip variability ( $CV_c = 5.16\%$ ;  $\Delta CV = |CV_l - CV_c| \approx 1.20\%$ ). The fact that a component of labelling variability is the chip-to-chip variation between arrays with the same labelled DNA explains this marginal increase of labelling variability compared to chip variability.

#### 2.3.5 Analysing the Relationship Between Oligo-level and SNP-level Variabilities

As discussed in Section 2.2.1.2, the coefficient of variability (CV) is preferred over the standard deviation (SD) as a means of quantifying the reproducibility of SNP arrays in "10K rep-test" dataset. However, understanding the extent to which technical variation influences the measured LR values requires a reliable formulation that links the CV to assay performance. To characterize this relationship in Affymetrix SNP arrays and to investigate how it influences the log2-ratio values of the oligos, I performed the following analysis. Initially, I estimated the acceptable *k*-fold variation between intensity measurements of the same oligonucleotide probe across replicate arrays, and then I used Reed's model (Equation (2.2.2)) to relate the CV to the probability of two disparate oligos (i.e., oligonucleotide probes that differ at least by *k*-fold between replicate arrays) in "10K rep-test" dataset.

In the final analysis, I derived a mathematical formulation to link the oligo-specific variability to the extent of SNP-level LR variations, under the assumption that the average log2-ratio signal intensity from *n* PM oligonucleotide probes (oligos) in the same probe set is used to measure the SNP log2-ratio intensity values (throughout this Chapter, I will refer to the SNP log2-ratio intensity estimate as the SNP LR values). In conclusion, I applied the above method to the real data obtained from the "10K rep-test" experiment to evaluate the associated oligo-level and SNP-level variability in this dataset. The detailed description of these analyses is presented in the following sections.

#### 2.3.5.1 The Acceptable Range of Variability Between Replicate Oligos

According to Reed's model (discussed in Section 2.2.2) the probability that two independent measurements from the same sample will differ by a factor of k or more is given by Eq. (2.2.2), under the assumption that the data from these samples are normally distributed after logarithmic transformation [279].

As discussed in the previous section, histograms of chip and labeling errors confirmed the log normal hypothesis of the "10K rep-test" dataset (p. 40 and Figure 2.3), thus satisfying the prerequisite of Equation (2.2.2) [209, 275, 327]. By applying this model, I generated Figure 2.10, which depicts the probability of observing > k fold difference in the raw intensities of two replicate oligos. The curves in this graph reflects the aforementioned relationship based on a variable range of k and CV. To interpret the results from this nomogram (Figure 2.10), we need to understand the critical value of k, which defines the maximum acceptable range of variation between replicate oligos in the context of copy number data analysis. For copy number analysis, the permissible variability corresponds to the range of variation between individual PM oligos that does not affect

the SNP's overall LR values. To find the proper value of k in 10K SNP arrays, I performed the following analysis. Let A and R be the readouts of the same normal oligonucleotide probe (LR = 0) in 'test' and 'reference' arrays, respectively. Therefore:

$$\log_2 \frac{A}{R} = 0$$

$$\Rightarrow \ \log_2 A - \log_2 R = 0$$
(2.3.4)

Now assume that in a replicate array the raw signal intensity of this oligo is increased by *k*-fold  $(A' \rightarrow kA, k > 1)$ . Thus the log2-ratio readout of the same oligo in the replicate array is:

$$\log_2 \frac{kA}{R} = \log_2 kA - \log_2 R$$
$$= \log_2 k + \log_2 A - \log_2 R \qquad (2.3.5)$$

Substituting the result of Eq. (2.3.4) in (2.3.5) would result in:

$$\log_2 \frac{kA}{R} = \log_2 k + 0 = \log_2 k \tag{2.3.6}$$

The result of Equation (2.3.6) suggests that k-fold increase in the raw intensity readout of a normal oligonucleotide probe (LR = 0) would result in  $\log_2(k)$  increase in the corresponding log2-ratio (LR) estimate (assuming that the same reference set is used to analyse the data from replicate arrays). In practice, regardless of the CNV calling method used, not every deviation of LR from the theoretical baseline (LR = 0) is reported as a significant copy number change. Instead, often default deletion and amplification thresholds are used to select significant deviations of intensity as potential regions of copy number aberration [328–330]. For instance, let assume log2-ratio of  $\pm 0.5$  is used to call amplifications and deletions. It must be noted that the theoretical log2 ratios of one copy gain and loss are +0.58 and -1, respectively. However, to compensate the effect of noise often a smaller magnitude of aberration is used to determine significant DNA gains and losses. Here, the selected values (LR =  $\pm 0.5$ ) are arbitrary thresholds that are used to show the probability of observing random noisy oligos that could shift a normal SNP above a deletion or an amplification threshold. As seen later in this Chapter, the presented results are generated based on several conditions that can be used to study the aforementioned probability with different thresholds. With the above assumption (LR =  $\pm 0.5$  as CNV thresholds), the acceptable technical variability is -0.5 < LR < +0.5. Therefore, the critical value of k estimated by Equation (2.3.4)

$$\log_2 k = 0.5 \Rightarrow \boxed{k = 1.4142 \approx 1.4} \tag{2.3.7}$$

This result indicates that in the context of copy number analysis, an approximate 1.4-fold difference in intensity measurements of the same oligonucleotide probe across replicate arrays can be regarded as the upper limit of acceptable variability. By substituting the measured CV from the empirical data ("10K rep-test" dataset), the frequency of replicate oligos that differ by  $\ge$  1.4 fold in Affymetrix 10K SNP arrays is estimated as the following:

$$p(1.4) = 2\Phi\left\{\frac{-\log_e(1.4)}{\sqrt{2\log_e\left[(7/100)^2\right) + 1}}\right\} = 6.66e - 04$$

where an upper estimate of CV ( $\approx 7\%$ ) was used to assess p(k = 1.4). Alternatively, this value may also be approximated by inspection of the nomogram presented in Figure 2.10 which plots the probabilities by using appropriate k (1.4) and CV (7%) values. This nomogram helps to understand the link between coefficient of variation (CV) and the probability of observing k fold disparate results in a replicate experiment.

This analysis reveals that based on estimated variability in 10K Affymetrix SNP arrays in the "10K rep-test" experiment, the p-value (*P*) of observing  $\ge 1.4$  fold<sup>1</sup> difference in the raw oligo-level intensities between replicate oligos is *P* = 6.66e-04. This means that based on estimated CV values in "10K rep-test" dataset, by random chance, about 308 oligos from 462,200 oligos on 10K SNP arrays are expected to be significantly different ( $\ge 1.4$ -fold) across replicate experiments. In general according to Figure 2.10, any variability below 10% would essentially have a p(k) probability of zero for observing  $\ge 1.4$  fold difference in SNP-level LR outputs. However, the visible difference between k = 1.1 and k = 1.4 in this figure emphasizes on the fact that, just by random chance, there is a much larger probability of observing disparate results (in replicate oligos) between 1.1-1.4 fold apart (corresponding to ~ 0.25 – 0.48 difference in log2-intensity values). A careful inspection of Figure 2.10 shows that there is ~50% probability that log2-ratio readouts of the same oligo in replicate arrays differ by 0.26 (in log2-scale), even in a highly reproducible experiment with CV as low as 10%.

#### 2.3.5.2 Finding a Link Between Oligo-level CV to the Changes in SNP-level LR Values

Understanding the extent to which these highly variable oligos affect the SNP-level RL readouts requires an analytical approach to identify how many noisy oligos in each 10K SNP probe set can

is:

<sup>&</sup>lt;sup>1</sup>equivalent to a difference of at least 0.5 between log2-intensity measurements of a replicate oligo

make a significant change in the LR value of a corresponding SNP probe (each 10K SNP readout is based on the information from 20 PM oligos). Here, I assumed that the SNP log2-ratio copy number (LR) values are calculated by averaging the log2-ratio intensity measurements from the PM oligos that belong to the SNP probe set. So the LR value of SNP "A" consisting of *n* perfectmatch oligos,  $A = \{A_1, A_2, ..., A_n\}$ , is estimated by:

$$S(A,R) = \frac{1}{n} \left( \log_2 \frac{A_1}{R_1} + \log_2 \frac{A_2}{R_2} + \dots + \log_2 \frac{A_n}{R_n} \right)$$
(2.3.8)

where n = 20 for Affymetrix 10K SNP array; and  $R_i$  denotes the intensity of the same PM oligonucleotide probe (*i*-th probe) in the reference set. The measurement of the same SNP probe, that is obtained from a replicate array, is represented by:

$$A' = \{A'_1, A'_2, \dots, A'_n\} = \{k_1 A_1, k_2 A_2, \dots, k_n A_n\}, \ k > 1$$

where  $k_i$  indicates the magnitude of intensity fold-change (of the *i*-th PM oligo). Similar to Equation (2.3.8), the average log2-ratio signal intensity of the replicate SNP (A') is assessed by:

$$S(A',R) = \frac{1}{n} \left( \log_2 \frac{k_1 A_1}{R_1} + \log_2 \frac{k_2 A_2}{R_2} + \dots + \log_2 \frac{k_n A_n}{R_n} \right)$$
(2.3.9)

To understand how  $k_i$ 's can make a significant change in the estimated LR values, the following analysis was performed. It was assumed that LR = -0.5 and LR = +0.5 are theoretical thresholds of loss and gain of DNA copy number, respectively. Thus, for a given normal SNP with LR = 0, the maximum acceptable random variability between replicate arrays is |LR| < 0.5. This means that the SNP readouts in replicate experiments are inconsistent (or disparate) if:

$$\bar{S}(A',R) = \frac{1}{n} (\log_2 \frac{A'_1}{R_1} + \log_2 \frac{A'_2}{R_2} + \dots + \log_2 \frac{A'_n}{R_n}) \ge 0.5$$
  
=  $\log_2 \frac{k_1 A 1}{R_1} + \log_2 \frac{k_2 A 2}{R_2} + \dots + \log_2 \frac{k_n A n}{R_n} \ge 0.5n$  (2.3.10)

By expanding the logarithm in Equation (2.3.10) and grouping the resultant terms, we have:

$$\begin{split} \bar{S}(A',R) &= (\log_2 k_1 + \log_2 k_2 + \ldots + \log_2 k_n) + (\log_2 A_1 + \log_2 A_2 + \ldots + \log_2 A_n) - \\ &\quad (\log_2 R_1 + \log_2 R_2 + \ldots + \log_n R_n) \geqslant 0.5n \\ &= (\log_2 k_1 + \log_2 k_2 + \ldots + \log_2 k_n) + (\log_2 \frac{A_1}{R_1} + \log_2 \frac{A_2}{R_2} + \ldots + \log_2 \frac{A_2}{R_2}) \geqslant 0.5n \\ &= \log_2 \prod_{i=1}^n k_i + \bar{S}(A,R) \geqslant 0.5n \end{split}$$

Assuming that SNP *A* represents a normal region of the genome with LR = 0 ( $\bar{S}(A, R) = 0$ ), and by substituting this in the previous inequality, we have:

$$S(A',R) = \log_2 \prod_{i=1}^n k_i + 0 \ge 0.5n \Rightarrow \boxed{\prod_{i=1}^n k_i \ge 2^{(0.5n)} \quad (k_i = 1 \text{ when } A'_i = A_i)}$$
(2.3.11)

where parameter *n* denotes the total number of PM probes in the probe set. If the above inequality (2.3.11) holds true, it implies that the variation in raw oligonucleotide signal intensities (denoted by  $k_i$ ) are sufficient to increase the average SNP LR measurement by at least 0.5 units (in log2-scale). Assuming that *m* represents the number of variable oligos (i.e., oligos that have at least *k* fold-change across replicate arrays) in a SNP probe set (*m* < *n*), the left hand side of inequality (2.3.11) can be substituted by  $(\bar{k}_m)^m$ . Therefore:

$$\prod_{i=1}^{m} \bar{k}_{m} \ge 2^{(0.5n)} \Rightarrow (\bar{k}_{m})^{m} \ge 2^{(0.5n)}$$
(2.3.12)

where *m* denotes the number of variable oligos in the SNP probe set that differ, on average  $\geq \bar{k}_m$  fold across replicate arrays; and *n* is the total number of oligos per SNP probe set. The extent to which a random *k*-fold change in *m* oligos (of the same SNP probe set) would affect the overall SNP LR value, can be measured by subtracting Equation (2.3.9) from Equation (2.3.8):

$$S(A',R) - S(A,R) = \frac{1}{n} \log_2 \prod_{i=1}^n k_i = \Delta S$$
(2.3.13)

where  $k_i$  is the variability of each PM oligonucleotide probe that belong to the same SNP probe set. Assuming that a given SNP probe set has only *m* variable (or disparate) oligos (*m* < *n*), with an average  $\bar{k}_m$  fold-change, the above equation becomes:

$$\Delta S = \frac{1}{n} (\log_2 \prod_{i=1}^m \bar{k}_m) = \frac{1}{n} \log_2 (\bar{k}_m)^m = \frac{m}{n} \log_2 \bar{k}_m$$
(2.3.14)

Equations (2.3.13) and (2.3.14) represent the magnitude of the difference between LR measurements of the same SNP in two replicate experiments ( $\Delta S$ ), depending on the oligo-level variability of the individual PM oligos in the SNP probe set. Estimating  $\Delta S$  for hypothetically varied values of *m* and  $\bar{k}_m$  can help to understand variations of SNP LR readouts based on the variability of their PM oligos.

## 2.3.5.3 Evaluating the Impact of Oligo-level Variability on the Extent and Frequency of Noisy SNPs in Replicate Experiments

The work presented in this section aimed to (1) assess what proportion of PM oligonucleotide probes on an Affymetrix SNP array are expected to vary by a minimum of k fold between replicate experiments, and (2) what proportion of Affymetrix SNP probe sets are affected by such oligolevel variations. Table 2.1 presents the results of evaluating oligos that vary by  $\geq k$  fold (aim #1). The first column, %CV, indicates an arbitrary range of CV values. Columns 2-6 denoted by  $p(k), k \in \{1.2, 1.41, 1.5, 2, 3\}$  represent the probability (p) that an oligonucleotide probe would differ by at least k fold across replicate experiments. The k and CV values used in this table are selected from the curves in Figure 2.10. As described in Section 2.3.5.1, k = 1.4 corresponds to the critical value of k in SNP data analysis estimated by Equation (2.3.7). The last 5 columns in this table (Table 2.1) denote the predicted frequency of oligos in 10K SNP arrays that differ  $\geq k$ across replicate experiments. It is observed that as the CV increases, the probability of observing variable oligos also increases. Furthermore, in spite of relatively small p(k) values, we observe a considerable number of oligos that are highly variable among replicate experiments. For example, for an experimental platform with a CV = 20%, the probability of an oligonucleotide exhibiting  $\geq$  1.4 fold difference across replicate experiments is only 0.148 (p(1.5)). However, under these circumstances the data from a 10K SNP replicate array is expected to contain 68,266 variable oligos. The impact of oligo-level variability on SNP-level variability is also shown in Table 2.2.

It is observed that for CV between 10-20%, there are between 445 to 3,065 SNPs that have at least two oligos that vary  $\geq 1.2$  fold across replicate experiments. As evident from this table, the predicted frequency of observing  $\Delta S \approx 0.5$  in the log2-ratio intensity measurement from the same SNP probe set across replicate arrays is dependent on the number of variable PM oligos across
replicate measurements (*m*) and the extent of this difference ( $\bar{k}_m$ ). In addition to these parameters, the frequency of such potentially unreliable SNPs is directly affected by the coefficient of variation (CV) of the experimental platform.

In summary, in this section I applied a mathematical model on the empirical data from a SNP array replicate experiment to predict the frequency of SNP readouts with significant non-specific variability between replicate arrays (more than 1.42-fold difference in raw signal intensities, which corresponds to more than 0.5 intensity difference in log2-scale). The results of Tables 2.1 and 2.2 indicate that based on the estimated CV in "10K rep-test" experiment (CV  $\simeq 7\%$ ), it is very unlikely that the experimental variation in this dataset would result in any significant change ( $\Delta S$ ) in the SNP LR values. However, it is important to note that the assessed CV is particularly low and the same data quality may not be reproduced in a typical laboratory conditions. Also, it must be pointed out that even based on the same CV value, there are still many oligos that may have LR changes less than 0.5 (see the intersection of CV = 7% with k = 1.1 (shown in black) and k = 1.2 (shown in blue) curves in Figure 2.10).

## 2.4 Conclusions

In this Chapter, I used a replicate data set obtained from 8 normal individuals to analyse probespecific contributions to technical variability in Affymetrix GeneChip<sup>®</sup> 10K SNP arrays. To quantify technical variability, a general step-by-step procedure was presented based on using the coefficient of variation or CV (Section 2.3.2). By applying this procedure on a 10K replicate dataset consisting of 69 samples from 8 individuals (Section 2.3.1), the average chip and labeling CV values were estimated to be 5.15% and 6.36%, respectively (Figure 2.9). These estimates indicated the measure of technical variability between intensity readouts of the same oligonucleotide probes across replicate SNP arrays, but they did not provide any information regarding the extent of variation of the SNP log2-ratios ( $\Delta S$ ) across replicate arrays (p. 43). This link is critical to understanding the impact of estimated CVs to the array performance in the context of copy number analysis. To address this issue, I used the mathematical model proposed by Reed et al. [279] to find the relationship between the CV and the probability of observing k-fold random difference (p(k))in intensity measurements of replicate oligos (Section 2.3.5.1). Next, I further expanded this model and developed a method to quantify the contribution of probe set-specific technical variability on the estimated SNP log2-ratio values ( $\Delta S$ ; Section 2.3.5.2). Then, I used the estimated CV values (shown in Figure 2.9) from "10K rep-test" dataset in the aforementioned model and found that under these conditions there was only a slight chance (p = 6.66e-04) that two replicate oligos in

Affymetrix 10K SNP arrays would differ significantly ( $k \ge 1.4$ ; p. 45).

The development of the relationship between the CV and p(k) enhances the usefulness of CV in biological interpretation of SNP array findings, which has never been previously investigated in the context of microarrays. This model allows generation of a dimensionless index of variability which is universally applicable to assess the performance of SNP array platforms in different laboratory conditions. Furthermore, the model can be easily adapted for other applications, for example, to find the probability of any specified k-fold differences in expression data that occur by random chance between replicate arrays.

By expanding Reed's model and applying it to SNP array data (Section 2.3.5), I was able to accomplish two important applications of the mathematical formulation that linked CV and p(k): (1) to assess whether or not the difference in raw intensity readouts of the same oligonucleotide probes between replicate arrays is due to random variation (by analysing the oligo-level variability; Table 2.1 and Figure 2.10); and (2) to assess whether the variation in a set of individual PM oligos can make a significant bias ( $\Delta S$ ) in the resultant SNP log2-ratio measurement (Table 2.2). In clinical applications, this model can also provide a quality control tool through which it can be determined whether the current estimated variability exceeds what has been established from past experiments. It is important to note that although in this analysis I used the critical value of k as minimum 1.4-fold difference in raw intensities (corresponding to  $\sim 0.5$  difference in log2ratio intensity readouts), the choice of k can also be easily adjusted to optimize the sensitivity and specificity of a specific analysis. For example, one could use this model to predict the frequency of observing a difference of 0.25 units across replicate experiments by setting k = 1.19 in Equation (2.2.2) and Figure 2.10. One important observation of the nomogram in Figure 2.10 is that any variability below 10% would essentially have p(k) probability of zero for observing  $k \ge 1.4$ fold difference in SNP-level LR outputs. However, the visible difference between k = 1.1 and 1.2 curves with k = 1.4 curve in Figure 2.10 implies that, just by random chance, there is a much larger probability of observing replicate oligos that are less than 1.4 fold different (k = 1.1 and k = 1.2 correspond to log2-ratio intensity changes of ~0.25 and ~0.48, respectively). A careful inspection of Figure 2.10 also indicates that there is  $\sim 50\%$  probability that two oligos would have at least 0.26 difference in their log2-ratio intensity readouts between replicate arrays, even in a highly reproducible experiment with a CV as low as 10%.

In this Chapter, I have outlined an approach to generate a reliable estimate of variability for Affymetrix SNP arrays and then developed a model to use the measured CVs to determine the quality of SNP data. The provided Equations, the nomogram visualization in Figure 2.10 and the critical-value tables provided in this Chapter (Tables 2.1-2.2) are simple tools that extend the

understanding of the CV and increase its usefulness in interpretation of technical variability of SNP arrays. The main conclusion of the work presented in this Chapter is that although Affymetrix 10K SNP arrays are shown to be highly reproducible, there is still a significant likelihood of the occurrence of noisy oligos just by random chance (Section 2.3.5.3). Such noisy oligos may affect SNP-level LR results and ultimately affect the quality of the downstream CNV findings. Based on the results of this Chapter, I concluded that in order to improve the quality of CNV results, it is important to develop a CNV detection method that is based on analysing probe-level variability of SNP array data.

## 2.5 Figures and Tables



**Figure 2.1: Sources of variability in SNP microarray experiments for identification of copy number variations (CNVs).** At the highest level of variability, there is a biological variation in the DNA copy number values that exists among different individuals in a population, also known as copy number polymorphism (CNP). At the experimental level, there is a variability in preparation and labeling of the same sample (labeling variability), as well as a variability in the hybridization of a sample to different arrays (chip variability). The last source of variability is a variability between different features on the same array, which includes both the variability between probes that target different SNPs loci, as well as the variability between individual oligonucleotide probes within the same SNP probe set.



Figure 2.2: Common sources of microarray technical variability: There are several components of technical variability in microarray experiments. The first component, "variability between Cy5 and Cy3 colors", is specific to two-color arrays, such as CGH-based platforms, where two samples (e.g., test and reference) are labeled with different fluorophores (usually Cy3 and Cy5 dyes) and co-hybridized on a single microarray. The other denoted factors are common among all arrays including Affymetrix SNP arrays, Illumina SNP arrays and aCGH technology. Labeling variability refers to the variation between intensity readouts of the sample that has been prepared and labelled by separate labeling reactions. The batch effects are often observed in large-scale studies where due to practical considerations the number of samples that can be prepared and hybridized at one time is limited. Under such circumstances, samples are often divided into groups or batches, and the samples in each group are analysed together in separate experiments that may be conducted several days or months apart. This procedure introduces systematic batch effects between these separate groups that makes it difficult, if not impossible, to compare between batches. The next source of technical variability is the variation between hybridization to different arrays, referred to as chip variability. As indicated in the above diagram, chip variability is dependent on two other components (1) variability between arrays in relation to the manufacturing process, also known as manufacturing variability; and (2) variability between different hybridization reactions.



(a) Histogram of Raw PM Intensities, Sample 5



(b) Histogram of Log-transformed PM Intensities, Sample 5

**Figure 2.3: Impact of log-transformation on the distribution of raw signal intensities from Affymetrix SNP arrays.** Panel (**a**) denotes the histogram of raw intensity readouts from all perfect match (PM) oligos on a 10K SNP array (total of 231,200 PM oligos), for a given 10K sample (S5 in replicate experiment, previously discussed in Figure 2.4). It is clear from this graph that the raw intensity data is not normally distributed. Panel (**b**) demonstrates the histogram of PM signal intensities of the same sample (S5) after logtransformation (natural logarithm). The red curve illustrates the estimated normal fit to the log-data. The bell-shape of the histogram presented in (b) indicates that the raw (PM) intensity readouts tend to follow a normal distribution after log-transformation.



**Figure 2.4:** Schematic representation of Affymetrix 10K replicate study. The replicate experiment, which is referred to as "10K rep-test", was designed and performed by our collaborators at Affymetrix<sup>®</sup> according to the process outlined below. The study included 8 normal individuals, referred by index 'S' as depicted in **Step A**. None of these individuals had any previously known genetic abnormality. Total genomic DNA obtained from each subject (such as S5, shown above) was divided into 3 batches<sup>\*</sup>. Each batch was then prepared and labelled independently, as illustrated in **Step B**. Following labeling process, each batch was hybridized onto 3 separate Affymetrix GeneChip<sup>®</sup> 10K SNP arrays, as demonstrated in **Step C**. Thus, the experiment resulted in 9 replicate arrays for each individual 'S' (except for S1 that had 6 replicate arrays<sup>\*</sup>). The "10K rep-test" experiment, therefore, generates a dataset consisting of the data from 69 replicate 10K arrays from 8 individual normal samples.

<sup>\*</sup>The only exception is S1, where only 2 (instead of 3) batches were available for the analysis.



Figure 2.5: Schematic representation of assessing technical variability in 10K SNP array replicate study. Panel (a) demonstrates how chip-to-chip variability is estimated in "10K rep-test" experiments. The "chip variability" is evaluated based on the difference of oligo-level intensity measurements of the same labelled sample hybridized to 3 replicate arrays. In the depicted example, chip variability of batch 1 of sample 5 (S5, b = 1), is estimated by comparing the microarray signal intensity outputs from chips c = 1, c = 2 and c = 3. Panel (b) depicts how "labeling variability" is measured in "10K rep-test" datasets. This variability, defined as the variation between separate labeling reactions of the same sample, is evaluated by comparing the oligo-level signal intensity outputs from arrays that have been hybridized by different batches of the same sample. In the depicted example (in panel (b)) the labeling variability of sample 5 (S5) is evaluated by comparing the variation in hybridization intensity output from 9 arrays, denoted by c = 1 to c = 9.



(a) Probe-level Distributions Before Normalization



(b) Probe-level Distributions After Normalization

Figure 2.6: Global Normalization of log-transformed intensity readouts from 3 replicate arrays. The boxplots are generated to compare the mean and spread of oligo-specific intensity readouts from 3 replicate chips of a sample from "10K rep-test" experiment (S8, b = 3). On each box, the central red line is the 50th percentile (median), the edges of the box are the 25th and the 75th percentiles and outliers are depicted by red '+' markers (see Appendix E for more information about boxplot visualization). Panel (a) shows the boxplots of log-transformed data before global normalization; and panel (b) denotes the boxplots of the same data after global normalization. It is evident from these plots that the overall mean intensity in these arrays have been brought to the same level as the result of normalization.













**Figure 2.7: MA-plots of chip variability.** Each dot represents the relationship between deviation in the log-intensity readouts of a particular oligo across 3 replicate arrays (same labelled sample hybridized to 3 separate 10K arrays, evaluated for 462,400 oligos on 10K array). The *x*-axis denotes the average intensity between replicate oligos ('A'), and the *y*-axis represents the deviation of intensity readouts of the same oligo between replicate arrays ('M'). Panels (A)-(C) show the MA-plots for 3 different batches of sample S2, and panels (D)-(F) indicate the MA-plots of 3 different batches of sample S5, described previously in Figure 2.5.a. These plots indicate that all depicted MA-plots are symmetrical around the *x*-axis, suggesting that the overall array deviation is not intensity-specific and, thus, there is no need to apply a more sophisticated normalization technique prior to quantifying labeling and chip variabilities.



**Figure 2.8: Histograms of deviation (error) in replicate arrays.** The above plots show the histogram of oligo-specific deviation of signal intensities of 4 samples in "10K rep-test" (S1, S3, S4 and S5). The associated histogram of oligo-level deviation is shown in blue and the predicted normal fit to each distribution is superimposed on the histogram plot by red curves. These bell-shaped distributions indicate that log-transformed intensity deviation (error) follows a normal distribution in these samples. Similar results have been obtained for the rest of samples in "10K rep-test" dataset (4 other samples; not shown here). This finding proves that the errors from "10K rep-test" experiment are normally distributed, an assumption which is the prerequisite for several statistical methods that were applied on this dataset throughout this chapter (such as Reed's mathematical model of relating the CV to the probability of random different readouts in replicate experiments; as described in Section 2.2.2).



(a) Chip Variability in SNP 10K Replicate Dataset



(b) Labeling Variability in SNP 10K Replicate Dataset

Figure 2.9: Estimated chip and labeling variability of the Affymetrix 10K replicate experiment. Panel (a) denotes the results of 10K chip variability in 8 DNA samples of the "10K rep-test" experiment. The *x*-axis ('S') represents the DNA sample index, and the *y*-axis (%CV) denotes the measured coefficient of variation for each specified batch of a given DNA sample. The estimated chip variability for 3 DNA batches of each sample is represented red, blue and green bars, respectively. The horizontal dashed line depicts the mean chip variability index across all samples and all batches in "10K rep-test" dataset ( $\overline{CV}_c = 5.16\%$ ). Panel (b) denotes labeling variability, evaluated by assessing the variation in hybridization intensity of the same sample that was prepared through 3 separate labeling variability across all samples ( $\overline{CV}_l = 6.36\%$ ). 1\*: denotes the only exception, S1, where only 2 (instead of 3) batches were available for this analysis.



Figure 2.10: Relationship between array variability and the probability of estimating oligos with k-fold difference in their intensity readout. This nomogram depicts the probability that two measurements from the same oligo will differ by a factor of k or more across replicate Affymetrix SNP arrays. The probability curves in this figure were generated by using a hypothetically varied range of CV and k values in Eq. (2.2.2). This graph allows monitoring the expected quality and consistency of the data generated from an array experiment with known variability (%CV).

The red solid curve (labelled as  $k = 1.4^*$ ) presents the probability of observing 0.5 units difference between log2-intensity readouts of the same oligo across replicate arrays. This magnitude of difference between log2-signal intensities (0.5), which is equivalent to 40% difference in raw signal intensities, is assumed as the maximum acceptable oligo-level variability between replicate measurements in this chapter. By finding the intersection of k = 1.4 curve with the estimated %CV of "10K rep-test" data set (max CV = 6.36%), that was presented in the previous figure (Fig. 2.9), the above nomogram indicates that there is a very low chance of observing oligos that differ  $\ge 1.4$  fold between the 10K replicate arrays. Thus, we can accurately conclude that the data from "10K rep-test" experiments were highly reproducible.

%CV	7	$p(k)$ : Probability that oligos differ $\geq k$ -fold					Predicted No. k-fold different oligos in 10K <sup>†</sup>				
-	<i>p</i> (1.2)	$p(1.41)^*$	<i>p</i> (1.5)	<i>p</i> (2)	<i>p</i> (3)	$F^{\ddagger}[1.2]$	<i>F</i> [1.41]	<i>F</i> [1.5]	F[2]	<i>F</i> [3]	
5	0.010	9.38E-07	9.60E-09	1.03E-22	1.68E-54	4566	0.43	0.004	4.77E-17	7.78E-49	
6.36	0.042	1.15E-04	6.41E-06	1.22E-14	2.24E-34	19619	53	3	5.63E-09	1.03E-28	
7	0.065	4.56E-04	4.12E-05	2.38E-12	1.11E-28	30129	211	19	1.10E-06	5.11E-23	
10	0.196	1.40E-02	4.05E-03	8.95E-07	6.82E-15	90689	6481	1872	0.41	3.15E-09	
12	0.281	0.040	0.016	4.15E-05	8.20E-11	129854	18680	7624	19	3.79E-05	
15	0.387	0.100	0.055	0.001	1.91E-07	179073	46411	25235	470	0.09	
20	0.515	0.216	0.148	0.013	8.76E-05	238062	99807	68266	6160	40	
25	0.601	0.320	0.244	0.047	0.002	277578	147720	112892	21503	742	
30	0.661	0.404	0.329	0.095	0.008	305303	186656	151944	43908	3762	
40	0.738	0.525	0.457	0.203	0.044	341056	242523	211111	93962	20224	
50	0.785	0.604	0.544	0.299	0.100	362789	279132	251386	138415	46252	
60	0.816	0.659	0.605	0.377	0.161	377227	304376	279689	174136	74522	
70	0.838	0.698	0.650	0.438	0.219	387431	322601	300345	202287	101053	
80	0.855	0.728	0.684	0.486	0.269	394980	336263	315934	224581	124509	
90	0.867	0.750	0.710	0.525	0.313	400764	346824	328039	242461	144764	

<sup>†</sup>Affymetrix GeneChip<sup>®</sup> 10K SNP array

\*1.41 is the critical value of k in 10K SNP arrays, as estimated by Equation (2.3.7).

<sup>‡</sup>'F' denotes the frequency of observing *m* variable oligos with  $\bar{k}_m$  fold-differences in the same SNP probe set, across replicate 10K SNP arrays

**Table 2.1:** The relationship between CV and predicted replicate oligos that are  $\geq k$ -fold different. This table summarizes the probability of obtaining a specific value of fold change, p(k), for various assumed values of CV. This uses the oligo-level variation information to predict the frequency of SNP probe sets in Affymetrix 10K SNP arrays that are affected by such variable oligos. The *m* and  $\bar{k}_m$  denote the number of variable oligos in a SNP probe set and the average fold-difference of variable oligos across replicate experiments, respectively. Columns 2-6 denoted by  $p(k), k \in \{1.2, 1.41, 1.5, 2.3\}$  represent the probability (*p*) that an oligonucleotide probe would differ by at least *k* fold across replicate experiments (see Figure 2.10). The last 5 columns of this table denote the predicted frequency of oligos in 10K SNP arrays that differ  $\geq k$  fold across replicate experiments.

In this table we observe that as the CV increases, the probability of observing variable oligos increases. Furthermore, in spite of relatively small p(k) values, there can be a considerate number of oligos on the 10K SNP array that differ  $\ge k$  fold among replicate experiments. For example, for an experimental platform with CV = 20%, the probability of an oligonucleotide to have  $\ge 1.4$  fold difference across replicate experiments is only 0.148 (p(1.5)). However, under these circumstances the data from a 10K SNP array are expected to contain 68,266 oligonucleotide intensity readouts that are likely to differ by  $\ge 1.5$  fold across replicate experiments.

F prob	PM Oligo be set that	s in the SNP t differ $\geq k$ -fold	Predicted No. of SNPs with $\Delta S$ difference in Affymetrix GeneChip <sup>®</sup> 10K SNP array						
$m^*$	$ar{k}_m^{**}$	$\Delta S^\dagger$	CV = 7%	CV = 10%	CV = 15%	CV = 20%	CV = 40%		
2	1.2	0.03	49	445	1734	3065	6292		
5	1.2	0.07	0.01	3	101	419	2528		
10	1.2	0.13	1.60E-08	9.77E-04	0.88	15	553		
15	1.2	0.20	1.88E-14	2.84E-07	0.01	0.55	121		
18	1.2	0.24	5.22E-18	2.15E-09	4.47E-04	0.08	49		
2	1.4	0.05	0.01	3	142	609	3330		
5	1.4	0.12	1.52E-12	1.68E-05	0.19	7	515		
10	1.4	0.24	2.00E-28	2.43E-14	3.20E-06	4.71E-03	23		
15	1.4	0.36	2.62E-44	3.52E-23	5.32E-11	3.00E-06	1		
18	1.4	0.44	7.76E-54	1.75E-28	7.21E-14	3.64E-08	0.16		
2	1.5	0.06	1.96E-05	0.19	34	252	2411		
5	1.5	0.15	1.37E-18	1.26E-08	0.01	0.81	230		
10	1.5	0.29	1.62E-40	1.37E-20	2.72E-09	5.71E-05	5		
15	1.5	0.44	1.91E-62	1.50E-32	1.32E-15	4.01E-09	0.09		
18	1.5	0.53	1.33E-75	9.94E-40	2.15E-19	1.29E-11	0.01		
2	2	0.10	6.52E-20	9.25E-09	0.01	2	478		
5	2	0.25	8.75E-55	6.62E-27	1.26E-11	4.86E-06	4		
10	2	0.50	6.63E-113	3.79E-57	1.37E-26	2.04E-15	1.39E-03		
15	2	0.75	5.02E-171	2.17E-87	1.49E-41	8.60E-25	4.84E-07		
18	2	0.90	6.74E-206	1.56E-105	1.56E-50	2.04E-30	4.06E-09		

\***m**: number of variable (PM) oligos that belong to the same SNP probe set ( $m \le n$ ; n = 20).

\*\* $\mathbf{\bar{k}}_{m}$ : average fold-change of log2-ratio intensity measurements of *m* variable oligos within the same SNP probe set (p. 47).

 $^{\dagger}\Delta$ S: magnitude of the difference in the LR value of a SNP due to oligo-level variability (estimated according to Eq. (2.3.14)).

**Table 2.2:** The impact of oligo-level variability across replicate measurements on the expected SNP-level variability. This table summarizes the effect of variable PM oligos on the estimated variability of SNP log2-ratio readouts ( $\Delta S$ ) and the frequency of such SNPs. The *m* (1st column) denotes the number of PM oligos in the SNP probe set that are different between replicate arrays ( $m \le 20$ , in Affymetrix 10K SNP arrays). The  $\bar{k}_m$  (2nd column) denotes the average fold-change, *k*, of the corresponding *m* variable oligos (explained in page 47). The magnitude of the resultant difference in the SNP LR values between replicate arrays is evaluated by Equation (2.3.14) and is shown in the 3rd column of the above table ( $\Delta S$ ). The last 5 columns indicate the predicted frequency of SNPs on Affymetrix 10K SNP array that are expected to differ  $\ge k$  fold between replicate arrays. For example, for k = 1.4, the expected frequency of SNP probe sets that have 2 PM oligos that are at least 1.2-fold different across replicate experiments is 49, when the SNP platform CV is 7% (the approximate value of CV that was estimated for the "10K rep-test" dataset). Under such circumstances, the SNP probe set mean signal could vary between replicate experiments by  $\Delta S = 0.03$  (in log2-scale). We also observe that for CV between 10% and 20%, 445 to 3065 SNP probe sets are expected to contain 2 oligos that vary  $\ge 1.2$  fold across replicate experiments, and therefore the SNP probe set readouts could differ by  $\Delta S = 0.03$ . However, this number increases to 6,292 if the CV of the experimental platform is 40%.

## **Chapter 3**

# Algorithm for Oligonucleotide Probe-level Analysis of Signal Intensities (OPAS)

## 3.1 Introduction

In recent years, high-density SNP genotyping arrays have become increasingly popular for copy number detection application, since these arrays can serve a dual role for both SNP genotyping as well as copy number analysis [135, 136, 149, 162, 209]. The most prominent SNP arrays are from two commercial vendors; Affymetrix and Illumina (explained previously in Section 1.2.2.3). As described in Chapter 1, the Affymetrix SNP array construction involves synthesizing 25-mer oligonucleotide probes (which I also refer to as "oligos" in this thesis) corresponding to a perfect match and mismatch of the two SNP alleles. This probe quartet, commonly known as a SNP "probe set", is the basic unit for downstream computational analysis (see Figure 1.2 for more details). The hybridization reaction of target DNA to the above oligos generates signal intensity measurements, which can then be converted by computational tools to infer SNP genotypes as well as regions of copy number variation [161, 172].

Successful application of this technology for copy number analysis has discovered a number of interesting CNVs with relationships to complex disease. For example, rare CNVs have been linked to schizophrenia [57] in a study where micro-deletions and duplications were shown to be responsible for disrupting genes involved in neurodevelopment. The *UGT2B17* gene on 4q13.2

was also linked to osteoporosis in a case-control study of 350 affected individuals in a Chinese population [331]. Despite these advances, a major disadvantage of oligonucleotide arrays is their poor signal-to-noise that can affect the quality of the predicted CNVs by increasing the rates of false positive and false negative CNV calls [141, 168, 332]. Several algorithms have been developed to improve the signal-to-noise ratios in these arrays by taking into account the length and GC content of the probes, and various algorithms for copy number detection have been developed to aid CNV detection using approaches often based on either Hidden Markov Models (such as QuantiSNP [185], PenCNV [186]), Segmentation (such as CBS Segmentation [189]) or t-tests [210, 212, 213]. Nonetheless, often algorithms depend on multiple sequential SNPs with significant intensity change to call a putative CNV. As a consequence, several studies (such as Itsara et al. [141]) have shown that CNV calling algorithms often emphasize specificity over sensitivity and as a result the CNV detection power is dependent on the number of SNP probes markers in a region of interest. Therefore, while large CNVs (> 4 Mb) are routinely identified by most of the available algorithms, when it comes to small CNVs (< 100 - 150 kb) or CNVs in regions with less probe density (< 8 - 10 SNP probes), the results from different algorithms are often no longer consistent [141, 215].

As discussed earlier, each Affymetrix SNP is represented by a collection of oligos that interrogate a SNP locus and its surrounding sequence in the genome. Such design strategy is inherited from Affymetrix expression arrays, where the relative expression of each gene was estimated by a set of probes in a probe set. In expression arrays, it was shown that although ideally all probes within the same probe set should represent the expression of the same gene, there was often a remarkable variation between their intensity readouts [207, 276].

In this Chapter, I first provide evidence that (1) there is a significant variation both between SNP probe sets in regions with known copy number values and among the oligonucleotide probes in the same SNP probe set; and (2) such variation exists regardless of the array density (e.g., 100K or 500K) or the size of the normal reference set that is used to evaluate intensity ratios, and thus I hypothesize that (3) the accuracy of predicted CNVs can be improved by distinguishing between true signal and noisy oligos and incorporating the true information to the downstream CNV calling method. Therefore, the underlying hypothesis of the work presented in this Chapter is that by improving the quality of SNP readouts, which are the input data to the downstream CNV calling algorithm, the impact of noise in Affymetrix SNP arrays can also be improved and, consequently, the CNV detection process would have higher sensitivity and specificity. Based on this hypothesis, I developed an algorithm which utilizes nonparametric statistical model-based methods for pre-processing SNP array data by analysing the presence of nonspecific binding and oligonucleotide

probe-specific effects on the estimated noise in the array readout. I will show in this Chapter that the pre-processing was able to dramatically improve the oligonucleotide probe level variation (noise) within SNP probe sets by improving the SD values from 1.22 (before pre-processing) to 0.41 (after pre-processing). In addition, I will provide further evidence indicating that despite improvements in the SD values, the magnitudes of copy number aberrations are significantly improved after the pre-processing phase (Section 3.3.3.5). This provides evidence that the proposed algorithm has been able to address one of the major issues in analysing SNP array data. As described in Chapter 1, other methods often improve the SD by smoothing the data or incorporating large windows of neighboring SNPs, which reduces the magnitude of both noisy oligos and true copy number aberrations. Thus, apparent noise reduction in such algorithms comes at the cost of reduced detection sensitivity, particularly CNVs in DNA regions with low density of SNP markers [220, 221]. In contrast to such methods, the proposed non-parametric approach I implemented distinguishes between noisy and informative oligonucleotide probes prior to excluding the noisy oligos from the array readouts.

## Contribution

The samples used in the work presented in this Chapter were provided by Dr. Jan Friedman and the "wet-lab" work (sample preparation and Affymetrix experiments) was performed at the Genome Sciences Centre. Throughout this Chapter, I also use validated CNV results of mental retardation (MR) project that have already been published [29, 30, 215] to test the performance of OPAS in improving the noise and detecting the real CNVs . The reported MR CNVs in the Friedman et al. study [29] were based on 100K SNP array data [29]. The CNVs were first selected by using CNAG [180] and dChipSNP [209] software packages with additional t-test statistics, as explained in [29] (this analysis was performed by a separate group at the GSC). The candidate CNVs were then examined by FISH analysis performed by Dr. Patrice Eydoux's lab at the Children's and Women's Hospital in Vancouver. The reported CNVs were a subset of all de novo events (i.e., events that were present in the affected child and not in either of the normal parents) that were successfully validated by FISH analysis.

The second MR-related CNV study was based on Affymetrix 500K SNP arrays [30]. In the latter study, chip-to-chip normalization, standardization to a reference set, genotype detection and copy number estimation were all performed using Affymetrix Power Tools (version 1.6.0) software suite (http://www.affymetrix.com). This analysis was performed by a separate group at the GSC. Estimation of CNV boundary positions was done using Significance of Mean Difference (SMD)

method, developed by Delaney et al. [259]. In SMD, the mean of SNP copy number estimates (or log2-ratios) within a candidate CNV region was compared to the mean of those SNPs on the rest of the chromosome. The probability of accepting the null hypothesis of Student's t-test (i.e., that the means were from the same distribution) was then calculated and used to identify the most significant putative CNV regions [30]. The CNVs reported in the Friedman et al. 500K array publication [30] were selected based on SMD hits with at least 10 contiguous SNPs and with a p-value less than 10e-8. Similar to the 100K study [29], all reported MR-related CNVs in 500K publication [30] were validated de novo events. Throughout this Chapter all references to known CNVs in MR study refer to the results from the aforementioned studies [29, 30].

In this Chapter, I present an algorithm for CNV detection based on probe-level data analysis using Affymetrix GeneChip Mapping 10K, 100K and 500K arrays. This algorithm, called Oligonucleotide Probe level Analysis of Signal intensities (OPAS), makes no assumptions about the performance of individual oligos within a SNP probe set; instead, OPAS uses fuzzy-logic theory to analyse the relationship between individual perfect match (PM) oligos in a SNP probe set to determine how many possible groups of such oligos exist in each probe set. The decision about which group(s) of oligos are informative is made through non-parametric statistical tests and a machine learning classifier. The validated CNVs in the aforementioned publications in MR [29, 30] are used to analyse the oligo-level variabilities in known CNV regions and also to examine the performance of OPAS in detecting real copy number aberrations. This approach for identifying individual noisy oligos in each SNP probe set is novel and has not yet been reported in other copy number detection algorithms [168, 180, 184, 185, 188, 333, 334].

## **3.2** Methods

#### 3.2.1 Algorithm Design

The OPAS design is divided into two main levels, as illustrated in Figure 3.1. The first module, "SNP pre-processing", aims to find the most informative subset of PM oligos in each SNP probe set and to generate improved SNP log-ratio readouts; the improvements are detailed in Sections 3.3.6-3.3.7. The next phase, "SNP post-processing", applies a normalization method to correct for PCR-induced biases and to partition each chromosome into regions where copy number changes between neighbouring segments.

The main underlying hypothesis of OPAS algorithm is that the noise in Affymetrix SNP arrays, in addition to being SNP dependent, also depends on the oligonucleotide probes within the SNP

probe sets. Therefore, the data reliability of individual oligos can directly impact the quality of the estimated SNP log2-ratio (LR) values and consequently the accuracy of the downstream CNV calls. Detailed explanation of different parts of this algorithm are provided in the following sections.

#### 3.2.2 SNP Pre-processing Phase

As previously described in Chapter 1, Affymetrix genotyping arrays interrogate each SNP genotype with a set of 25-mer oligonucleotide probes (or oligos), known as a SNP probe set. The oligos in a probe set are designed to query both DNA strands at multiple offsets with respect to the SNP position (Figure 1.2). The generated intensity data from these arrays can also be compared to a reference set to determine the relative abundance of DNA at the specified SNP sites for CNV analysis.

#### 3.2.2.1 Quantile Normalization

In the first step of pre-processing, a quantile normalization [205, 335] technique is applied on the log-transformed fluorescent intensity data from test and reference SNP arrays (.CEL files) to enable differentiation between real variations in DNA copy number and variations due to experimental differences between multiple arrays. The quantile normalization method was originally designed to normalize intensities in Affymetrix high-density oligonucleotide expression arrays [205, 335, 336]. However, the same general method can be adapted to normalize the data from SNP arrays. This normalization method is based on the assumption that the data from all samples (being compared) follow a common underlying distribution [336]. A possible theoretical problem with this approach is the risk of removing some of the signal in the tails of the distribution; however, several studies have shown that empirical evidence does not indicate that quantile normalization leads to such errors in practise [205, 302]<sup>1</sup>. The OPAS default normalization approach is a modified version that adjusts the data in extreme tail values to allow for greater differentiation [336].

#### 3.2.2.2 Clustering Individual Oligonucleotide Probes in a SNP Probe Set

The goal of OPAS pre-processing phase is to improve the quality of the estimated SNP signal by finding a subset of oligonucleotide probes with the most informative log2-ratio intensity in each SNP probe set. The OPAS strategy to find such informative subset of oligos is to first cluster

<sup>&</sup>lt;sup>1</sup>Appendix F provides a comparative study of several normalization techniques using 500K array data from cancer samples. The results of this analysis also indicates that background adjusted quantile normalization does not suppress the magnitude of real CNVs, empirically.

PM oligos in each SNP probe set into groups with similar LR values. One of the most popular methods for data clustering is the *k*-means algorithm [337, 338], a successful method that has been widely used in several applications, such as gene expression profiling [339–342] and expressed sequence tag (EST) analysis [343–345]. Despite the popularity of *k*-means algorithm for data clustering, one of the limitations of this method is that it requires the number of clusters (*k*) and the approximate centroid values ( $\mu_k$ ) as input parameters. If no prior information is available for cluster centers, random initialization is often employed. However, *k*-means is particularly sensitive to these initialization values. Another drawback of random initialization is that clustering the same SNP probe set multiple times would not necessarily generates the same clusters.

These limitations imply that in order to apply *k*-means for clustering oligos, first the appropriate initialization values must be determined in a non-random manner. To accomplish this task, I designed a two-stage clustering algorithm that first predicts the optimal clustering parameters (*k* and  $\mu_k$ ) using a non-parametric approach based on fuzzy logic theory (subtractive cluster analysis) [346] and then uses these values to initialize an optimization-based *k*-means clustering. The details of this clustering approach is described in the following two sections.

#### 3.2.2.2.1 Cluster Prediction: Fuzzy Subtractive Clustering

In step one of clustering, a fuzzy-logic subtractive algorithm [346–348] was applied on PM oligonucleotide probe level data in each Affymetrix SNP probe set. Subtractive method is a fast, one-pass algorithm for estimating the optimal number of clusters (k) and cluster centers ( $\mu_k$ ) in a set of data [346–349]. Assuming that  $P = \{p_i \mid i = 1, ..., n\}$  denotes a SNP probe set with 'n' number of PM oligos, the likelihood that the *i*-th PM oligo ( $p_i$ ) from *P* is a cluster center is defined by [350]:

$$D_i = \sum_{j=1}^n \exp\left(-\frac{\|P_i - p_j\|^2}{(r_a/2)^2}\right), \quad i = 1, 2, \dots, n, \ i \neq j$$
(3.2.1)

Where radius  $r_a > 0$  is the cluster radius (the best  $r_a$  default values are usually between 0.2 and 0.5). In the following step, the PM oligo that is associated with the largest *D* likelihood value is chosen as the first cluster center. Subsequently, the density measure of the rest of the PM oligos in this probe set (*P*) are revised by:

$$D_i = D_i - D_{c_1} \exp\left(-\frac{\|P_i - p_{c_1}\|^2}{(r_b/2)^2}\right)$$
(3.2.2)

Where,  $c_1$  is the first cluster center and  $r_b$  is usually set at  $1.5 \times r_a$  to avoid obtaining closely spaced cluster centers. The subtractive clustering algorithm iterates between Equations (3.2.1)-(3.2.2) until

all of the PM oligos in a SNP probe set are within the radii of a cluster center [350]. Notably, fuzzy clustering alone did not perform well at clustering border line oligos, and therefore, we combined it with *k*-means optimization based clustering as described below.

#### 3.2.2.2.2 Cluster Estimation: k-means Clustering

In the second phase of clustering, for each SNP probe set the number of clusters (k) and cluster centers ( $\mu_k$ ) that were obtained by fuzzy subtractive analysis are used to initialize a *k*-means clustering algorithm (Sec. 3.2.2.2). The *k*-means partitions the PM oligos in each SNP probe set into *k* mutually exclusive clusters, such that the oligos within each cluster are as close to each other as possible and as far from oligos in the other clusters as possible. In the rest of this thesis, I refer to a cluster of similar PM oligos as an "oligo cluster".

The *k*-means method iteratively updates cluster centers to minimize a cost function, until there is no significant change in the cluster centers or when it exceeds the maximum number of iterations. The *k*-means cost function f is defined by:

$$f = \underset{\mathscr{C}}{\operatorname{arg\,min}} \sum_{i=1}^{k} \sum_{p_j \in c_i} \|(p_j - \mu_i)\|^2, \quad j = 1, 2, \dots, n$$
(3.2.3)

Where  $c_i$  represents the *i*-th oligo cluster in a SNP probe set with centroid  $\mu_i$ ; and *k* is the number of predicted oligo clusters in the SNP probe set ( $1 \le k \le n$ ). When clustering finds more than one oligo cluster in a SNP probe set, further analysis is required to distinguish between noisy and informative oligo clusters. This process, referred to as SNP classification, is described in the following section (Section 3.2.2.3).

#### 3.2.2.3 SNP Classification

The goal of SNP classification is to identify which subset of oligos in a SNP probe set represents the true SNP signal intensity. The OPAS default SNP classification approach consists of two modules. First, the likelihood estimation phase tests the null-hypothesis that the log2-ratio intensity values of each oligo cluster are significantly different from the normal baseline or not. In the next step, prediction phase, the likelihood of these null-hypothesis tests are used as input data to a machine learning classifier to find the most informative oligo clusters in each SNP probe. For each SNP, the center of the informative oligo cluster is then used as the OPAS-estimated LR value.

#### 3.2.2.4 Likelihood Estimation: Kolmogorov-Smirnov (KS) test

In this phase, three sets of non-parametric Kolmogorov-Smirnov tests are separately applied to each oligo cluster of the SNP probe sets; as listed in Table 3.1. Each KS-test in this table makes a statement about how the population of the oligos in an oligo cluster (X) is related to a specified baseline distribution (e.g.,  $y_0$ ). The values of the alternative tests, shown in Table 3.1 ("two.sided", "less" and "greater"), define the null hypothesis that the cumulative distribution function (CDF) of oligo cluster data X is equal to, not less than or not greater than the cumulative distribution function of the background, respectively. For instance, KS-test set #1 tests the alternative hypothesis that the CDFs of the oligo cluster data and the background population are not equal ( $H_1 : X \neq y_0$ ).

For an oligo cluster X with k number of PM oligos and a standard deviation of  $\sigma_x$ , the background distribution for test set #1,  $y_0$ , is defined as a distribution of k randomly generated numbers with mean zero and the same standard deviation of the oligo cluster data being compared to  $(\sigma_x)$ . Nonetheless, when the same background distribution  $(y_0)$  was used in KS-tests 2-3, often none of these mutually exclusive null hypothesis tests were rejected. The implication of this observation was that often we could not determine whether the oligo cluster data is likely smaller or larger than the zero-mean baseline<sup>1</sup>. To circumvent this issue, different baseline distributions were used for the aforementioned test sets (in Tab. 3.1). The baseline distributions for tests 2 and 3 is generated with the same strategy used to create  $y_0$ , but with different mean population values ( $\mu(y_1) = -0.5$ ,  $\mu(y_2) = +0.6$ ). These values were set following the analysis of > 1300 oligo clusters in more than 600 SNP probe sets from known deleted and amplified regions (from mental retardation project). These background distributions  $(y_1 \text{ and } y_2)$  provided the greatest distinguishability between the results of the aforementioned 3 mutually exclusive tests. Although, it must be added that still  $\sim 20\%$ of SNP probe sets did not provide any relevant copy number information, mainly not because the tests failed to generate accurate results, but the fact that the SNP probe set readouts were inconsistent with the known DNA copy number of the underlying region. Some of the examples of inconsistent SNPs were previously described in Section 3.3.3.1 and also seen in Figure 3.6 (e.g., SNP<sub>80</sub> and SNP<sub>81</sub>). The schematic representation of clustering and likelihood estimation analysis

<sup>&</sup>lt;sup>1</sup>This statement requires further clarification. In biology often a hypothesis test can have one of two outcomes: either the researcher can "accept" the hypothesis or "reject" it. For instance, if the scientist discovers that a particular gene does not appear in the genome sequence, then it is deduced that it has been deleted. However, in statistics there is a major issue with the notion of accepting the null hypothesis. Instead, the failure to reject the initial hypothesis implies that the data are not sufficiently persuasive to prefer the alternative hypothesis over the null hypothesis. Therefore, even if we reject the null hypotheses that oligo cluster data is equal to the baseline and also reject that is it smaller than the baseline but fail to reject that is it larger that the latter distribution, we can not directly conclude that the oligo data is in fact larger than the baseline.

for a given SNP probe set is presented in Figure 3.10.

#### 3.2.2.5 Machine Learning Classifier Based on Discriminant Analysis

In the next phase of SNP classification, a machine learning-based classifier was implemented to discover the most informative oligo cluster for each SNP by utilizing the information that was obtained in the likelihood estimation phase. A quadratic discriminant analysis (QDA) was implemented to perform this classification task. The main inputs to the QDA classifier are the estimated p-values of the hypothesis tests described in the previous section, the number of PM oligos in each oligo cluster and the cluster centers. Further description of QDA classifier is presented in Appendix G.

It is hypothesized that the SNP data manipulations described in the SNP pre-processing phase, can decrease the overall standard deviation of the array (noise) while increasing the magnitude of true signal aberrations (results shown in Section 3.3.3.5). Such improvements can directly influence the quality of the downstream CNV calling process; as detailed in Sections 3.3.6-3.3.7.

## 3.2.3 Alternative Approach for SNP Pre-processing Based on Naive Bayes Classification

It has been suggested that instead of clustering oligos into groups with similar LR values and using likelihood estimation and QDA to independently classify each SNP, a classifier should be directly applied on the SNPs probe sets. To implement this alternative approach I trained Naive Bayes classifiers using SNPs from known deleted, amplified and normal regions (the same training data used for QDA<sup>1</sup>). The underlying assumption is that a Naive Bayes classifier can directly determine the SNP copy-number status (i.e., normal, deleted or amplified) by comparing the log2-ratio intensity pattern of the oligos in the SNP probe set with those from DNA regions with known copy-number values (Figure 3.2). Two Naive Bayes classifiers were implemented for this analysis, one with normal (gaussian) distribution and another with kernel smoothing density estimation.

The main difference between the Naive Bayes approach and OPAS default SNP classification is that all the PM oligos in a SNP probe set are used as the input data to the Naive Bayes classifiers. However, as explained in the previous section, the OPAS SNP classification approach is based on QDA analysis of the information that is obtained from oligo clusters (compare the flowcharts in Figures 3.1 and 3.2). The comparison between the classification performance using the QDA and Naive Bayes approaches is detailed in Section 3.3.3.4.

<sup>&</sup>lt;sup>1</sup>The same 324 SNPs from known classes (deleted, normal and normal regions) that were used for QDA training.

#### 3.2.4 Post-processing and CNV Calling

The goal of the OPAS post-processing phase is to partition the whole genome into regions where the copy number changes between contiguous segments. The post-processing phase includes two main sub-modules (Figure 3.1). In the first step, a second phase of normalization is applied to SNP log2-ratio data (generated by pre-processing phase) to correct for biases that are due to the PCR process (Section 1.2.2.3). Next, a non-parametric CNV calling algorithm (Circular Binary Segmentation) is applied to identify putative regions of copy number change in Affymetrix SNP data. I hypothesized that incorporating high quality SNP readouts with a non-parametric CNV calling approach can improve the quality of ultimate CNV calls. In this section, I describe the OPAS post-processing modules.

#### 3.2.4.1 PCR Fragment Length Normalization

As mentioned in Chapter 1, Affymetrix genotyping assay involves a Polymerase Chain Reaction (PCR) process to amplify the target DNA sample. It has been shown that locus-specific array intensity data may be correlated with PCR fragment length [180]. This is due to the fact that longer fragments usually generate fewer amplified products, which reduces the material available for labeling and hybridization and results in weaker signals [180, 351]. The magnitudes of such PCR-induced biases may vary between arrays, so they do not necessarily cancel each other out in the estimated test versus reference LR ratios [180, 351–353]. To correct for such biases, a non-linear LOWESS regression method was used [208, 354, 355]. This method has been used in a wide range of microarray applications, such as adjusting the waves in microarray signal intensities [140, 356] and normalizing Illumina Infinium SNP data [357].

#### 3.2.4.2 Circular Binary Segmentation (CBS) Algorithm

The Circular Binary Segmentation (CBS) algorithm [189] is a modification of binary segmentation [230], a well-known method for the change-point problem<sup>1</sup> in statistics. The basic idea of this entirely non-parametric approach is to recursively split chromosomes into segments based on a maximum p-value that is estimated by permutation analysis. A study by Lai et al. [231] comparing 11 different CNV calling algorithms such as mixture model, HMM, maximum likelihood, regression, wavelets and genetic algorithms, concluded that CBS appears to perform consistently well. Another comparison study by Willenbrock and Fridlyand et al. [232] found that a CBS-based CNV calling method (DNAcopy software) [189] had the highest sensitivity rate and the lowest false de-

<sup>&</sup>lt;sup>1</sup>http://biostats.bepress.com/cobra/ps/art44

tection discovery rate (FDR) of CNV breakpoints compared to Gaussian-based GLAD [210] and an HMM-based method [358].

A key advantage of CBS method is that it does not have a strict limitation on the minimum number of probes in a candidate CNV region. The main disadvantage of this algorithm is its low speed. However, the Venkatraman et al. [242] modification of the original algorithm has alleviated this problem to some extent. Due to its non-parametric nature and proven accuracy, CBS is one of the most powerful statistical-based algorithms for CNV detection<sup>1</sup> [164, 232, 359–362].

#### **3.2.5 OPAS Visualization and Other Features**

Several visualization and computational tools have been developed to facilitate OPAS data analysis and CNV interpretation. Figure 3.3 presents a schematic representation of the data that is generated during OPAS analysis. As noted in this graph, the OPAS software uses the raw Affymetrix .CEL files and generates 3 main output files (.jpg, .BED and .txt files). The .jpg files show the OPAS results of segmented data in each chromosome along the chromosome ideograms based on banding patterns from USCS genome browser (hg17, hg18 or hg19)<sup>2</sup>. The .txt file is a list of all estimated segments in all chromosomes following the CBS analysis.

An additional script has been implemented that allows OPAS to generate UCSC custom tracks for any list of candidate CNVs, based on .BED formatting<sup>3</sup>. This tool can also generate tracks with color spectra for better representation of the type of predicted events (e.g., deletion versus amplification) or the magnitudes of the estimated LR values. Other computational tools have also been implemented to allow finding overlapping CNVs across multiple samples; and comparing a list of putative CNVs with known copy number polymorphisms in Toronto Database of Genomic Variants (DGV)<sup>4</sup>. These tools can help to obtain further information about the generated CNV results.

#### 3.2.6 Simulated Data for Comparative Analysis of CNV Calling Algorithms

A simulated data set was generated based on the Affymetrix 250K Nsp array to evaluate the performance of OPAS algorithm. To create the simulated signals, first a distribution of random Gaussian data was generated. The SNP level and oligonucleotide level standard deviations of the generated

<sup>&</sup>lt;sup>1</sup>http://www.broadinstitute.org/cancer/software/genepattern

<sup>&</sup>lt;sup>2</sup>The OPAS default genome build is hg18; however, it can be modified through the user interface.

<sup>&</sup>lt;sup>3</sup>http://genome.ucsc.edu/FAQ/FAQformat.html#format1

<sup>&</sup>lt;sup>4</sup>http://projects.tcag.ca/variation/

distribution were set to equal those of a Follicular Lymphoma sample<sup>1</sup> to demonstrate performance on real data (SD= 0.08 and 0.04). Next, 200 non-overlapping simulated regions of copy number change were randomly scattered throughout the simulated signal. These simulated CNVs had 8 different alteration sizes with 2, 4, 8, 10, 15, 25, 100, and 200 data points. For each alteration size (referred to as *w*), 25 non-overlapping random CNV regions were generated, resulting in total of 200 ( $8 \times 25 = 200$ ) distinct simulated CNVs throughout the entire signal. This data vector is referred to as a template signal. To implement the range of alterations expected in a typical SNP array experiment, a constant log2-ratio magnitude of 0.11, 0.2, 0.4, 0.6, 0.8 and 1.0 was then added to the simulated CNV regions of the template signal. This model is used in comparing the sensitivity and precision of CNV calling algorithm; detailed in Section 3.3.6.

#### 3.2.7 Analysing the Effect of Noise on CNV Calling Performance

The second simulation analysis used an artificial but biologically inspired model to generate synthetic data to evaluate OPAS CNV detection sensitivity in the presence of added noise. This simulated model is based on Affymetrix Nsp SNP array data (250K) from a follicular lymphoma patient. The genome of this patient, referred to as ht-17, has been thoroughly analysed for both sequence level mutations and structural aberrations using several different platforms, including array CGH, 500K SNP arrays, fingerprint profiling (FPP) [363] and whole transcriptome shotgun sequencing (WTSS) [364] (all these data have been provided by other groups at the GSC).

The FPP results found a deletion on chromosome 14 immunoglobulin heavy locus (IGH@), spanning approximately 870 Kb (14q32; Figure 3.4). This deletion was validated by BAC end sequencing (BES) of an FPP clone that harbored this deletion (clone HTa17-0164B08, denoted in Figure 3.4b). No other CNV was detected in this chromosome by either of the above methods<sup>2</sup>. Therefore, it can be hypothesized that the aforementioned deletion on 14q32 (105,163,197-106,035,402) is likely the only CNV in chromosome 14 of the above patient. To generate simulated chromosome 14 signals based on real data, I added random Gaussian noise to the original Nsp data of the this patient. Based on this model, the simulated noisy signal *Y* is defined by:

$$Y_i = X_i + \varepsilon_i \quad 1 \le i \le n$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
(3.2.4)

where *n* represents the sample size and X is the original Nsp data of chromosome 14. The term  $\varepsilon$ 

<sup>&</sup>lt;sup>1</sup>This sample is obtained from follicular lymphoma project, described in Chapter 4.

<sup>&</sup>lt;sup>2</sup>http://map02.bcgsc.ca:5000/anomaly/

denotes the added gaussian noise which follows a normal distribution with standard deviation of  $\sigma$  and mean of zero. Based on this model, signal *Y* includes the IGH deletion (on 14q32) but is corrupted with a known magnitude of noise ( $\varepsilon$ ).

#### 3.2.8 Comparative Analysis of OPAS and Circular Binary Segmentation

As explained in Section 3.2.4, in OPAS post-processing phase Circular Binary Segmentation (CBS) algorithm is applied to detect regions where log2-ratio intensity changes between neighbouring segments. The aim of the following experiment is to investigate whether analysing a sample using OPAS has any advantage for CNV detection over application of CBS segmentation alone. The 100K array data from a mental retardation (MR) patient with a known (validated) deletion is used to perform this experiment. The aforementioned deletion spans approximately 294 Kb (chromosome 2p16.3) and includes 17 Xba SNPs. Next, a series of simulated signals are generated by randomly selecting N SNPs from the deleted region and eliminating them from the original signal. This process is repeated multiple times for each value of N (see Section 3.3.7 for more details). The performance of OPAS and CBS is compared by assessing the sensitivity of each method to detect the known deletion with fewer SNP probe markers. The results of this analysis are described in Section 3.3.7.

### 3.3 Results

#### 3.3.1 Patterns of LR Intensity Fluctuations in SNP Array Data

To study the extent of variabilities in SNP array data, I examined more than 950 SNP probe sets in known regions of copy number variation using FISH validated CNVs from Friedman et al. [29] and Taylor et al. [365] publications. The results of this analysis found different fluorescent intensity responses to DNA copy number abundance, both across SNPs in a region with the same copy number (SNP-level variability), and among perfect-match oligos that are in the same probe set (oligo-level variability). Figure 3.5 demonstrates examples of such SNP-level and oligo-level variabilities in log2-ratio fluorescent intensities. The top panel (Fig. 3.5a) presents a scatterplot of a region on chromosome 10 that includes a known deletion, approximately 353 kb in size (10q24.31-q25.1; denoted by red arrows). The sample is from a patient with global developmental delay and a desmoplastic medulloblastoma that was previously reported by Taylor et al. [365]. The phenotypes in this patient were associated with the disruption of the *SUFU* gene that is located on the deleted region. In the rest of this Chapter, I will refer to the aforementioned deletion in this patient as the

"SUFU deleted region".

The highlighted SNPs in Figure 3.5a (SNPs 79, 81, 83, 93 and 100) are all within the SUFU deleted region. As seen in this figure, these deleted SNPs have noticeably different LR values and standard deviations (Figures 3.5b-3.5e;  $-1.34 \leq \overline{LR} \leq -0.05$ ,  $0.3 \leq SD \leq 1.3$ ). Figure 3.5e shows an example of the impact of oligo-level variability on the quality of the estimated SNP signal. The depicted SNP (SNP<sub>100</sub>) has mean LR value of -0.5; however, 5/20 perfect-match (PM) oligos of this SNP indicate increased signal intensities (with mean log2-ratio value of +0.2; marked by blue dots) and can be considered as noisy oligos. In the same SNP probe set, 11 other PM oligos indicate a relative loss of signal intensity (with mean log2-ratio intensity of -0.52; marked by green dots). The remaining 4 PM oligos in this probe set denote a significant loss of signal intensity with mean log2-ratio of -1.3 (marked by red dots). The latter subset of oligos represents 2.6-fold increase in the magnitude of copy number loss compared to the original mean signal of  $SNP_{100}$  $(\overline{LR} = -0.5)$ ; and an LR estimate that is approximately equal to the theoretical log2-ratio value of one copy number loss  $(\log_2(1/2) = -1)$ . Such findings, that are commonly observed in SNP arrays, prove that using a subset of PM oligos to evaluate log2-ratio values can improve the quality of the SNP signal. It is also reasonable to speculate whether the SNP signal can be improved simply by increasing the number of reference samples. The analysis of the impact of reference set size on the estimated log2-ratio intensities is presented in the following section (Section 3.3.2).

## **3.3.2** Analysing the Impact of the Size of the Reference Set on the Estimated SNP Signals

To assess whether a larger reference set improves the quality of SNP signals, it is important to measure its impact on both the magnitude of estimated LR values and the number of informative probes in regions with known copy number values. These two analyses are described in Section 3.3.2.1 and Section 3.3.2.2, respectively.

#### 3.3.2.1 Impact of the Size of the Reference Set on the Estimated LR Values

For this analysis, the relative copy number ratios of all PM oligos in the "SUFU deleted region" were compared using 6 different reference sets with varying sizes. Three of these reference sets were the normal father ( $R_{father}$ ), normal mother ( $R_{mother}$ ) and the mean of both parents ( $R_{parents}$ ) of the affected MR patient. The other three reference sets consisted of 24 ( $R_{24}$ ), 99 ( $R_{99}$ ) and 150 ( $R_{150}$ ) normal individuals, all with normal karyotypes<sup>1</sup>. The results of this analysis are

<sup>&</sup>lt;sup>1</sup>These samples were part of a larger data set of normal parents in a trio study by Friedman et al. [29].

presented in Figures 3.6-3.7 and Table 3.2.

Figure 3.7 presents a comparison between the distribution of log2-ratio intensity readouts of PM oligos in the *SUFU* deleted region (500) with all other PM oligos on the array (> 1.1 mil), using the cumulative density plots of these two populations. As seen in this figure, in all cases the distribution of *SUFU* deleted oligos falls on the left side of the baseline CDF (CDF of the entire array; shown in black). The latter observation implies that the distribution of log2-ratio intensities of the PM oligos in *SUFU* deleted region is consistently smaller than the rest of the array, regardless of the size of the reference set that was used to calculate LR values. The distance between the CDFs of the deleted oligos from the rest of the array. Figure 3.7 also denotes that at  $F(x) = 0.5^1$ , the  $\delta$  value with respect to the largest reference set (*CDF*<sub>150</sub>; yellow curve) is smaller than the  $\delta$  with respect to a single reference set does not necessarily improve the magnitude of LR deviation of real CNVs.

#### 3.3.2.2 Impact of the Size of the Reference Set on the Number of CNV-affirmative Oligos

The Xba SNP array data (50K) of 995 PM oligos in 22 regions of known copy number deletion (previously reported in [29]) was analysed to investigate whether using a larger reference set improves the proportion of CNV-affirmative oligos in these CNVs. In this analysis, the proportion of CNV-affirmative oligos (referred to as  $\theta$ ) is defined as the number of PM oligonucleotide probes in a deleted SNP probe set that indicate loss of signal intensity (LR< 0) divided by the total number of SNPs in the corresponding deleted region. The boxplots in Figure 3.8 illustrate the distributions of the average  $\theta$  values (average number of affirmative oligos per SNP probe set) across the 22 known deletions with respect to 3 separate reference sets ( $R_{24}$ ,  $R_{99}$  and  $R_{150}$ )<sup>2</sup>. As seen in this figure, the reference set with 99 normal samples ( $R_{99}$ ) has a better  $\theta$  rate compared to the reference set with 24 samples ( $\theta_{R_{24}} = 13.3/20 \simeq 66\%$ ,  $\theta_{R_{99}} = 15.3/20 = 77\%$ ). However, the proportion of affirmative oligos using the largest reference set with 150 samples is lower than that of 99 samples ( $\theta_{R_{150}} = 14.3/20 \simeq 71\% < \theta_{R_{99}} = 77\%$ ).

Similar analysis was applied on 5 SNPs in *SUFU* deleted region (SNPs 79, 81, 83, 93 and 100; shown in Fig. 3.5) to study the impact of the size of reference set on the estimated SNP log2-ratio values and the number of CNV-affirmative oligos in these 5 SNPs. Comparing the data generated by  $R_{150}$  and  $R_{father}$  in Table 3.2 reveals that using  $R_{father}$  leads to larger magnitudes of copy number

 $<sup>{}^{1}</sup>F(x)$  is defined as the proportion of X values that are less than or equal to x.

<sup>&</sup>lt;sup>2</sup>These reference sets were previously described in page 79.

loss and approximately the same overall number of informative oligos for these 5 deleted SNPs. In summary, the results of the analyses presented in Sections 3.3.2.1-3.3.2.2 suggest that there is no substantial evidence to support the hypothesis that increasing the size of a reference set improves the magnitude of real copy number aberrations<sup>1</sup>.

### 3.3.3 Results of OPAS Pre-processing Phase

#### 3.3.3.1 Clustering PM Oligos in SNP Probe-sets

The raw array data (from .CEL files) was normalized using quantile normalization method, as explained in Section 3.2.2.1. Fuzzy-Kmeans clustering was applied to each SNP probe set according to the model described in Section 3.2.2.2. Figure 3.9 illustrates examples of fuzzy-kmeans clustering of 5 SNPs in *SUFU* deleted region that were previously shown to have noticeably different LR values (see Figure 3.5). Each predicted oligo cluster is shown by a different color in these plots (red, blue and green). As observed, 3/5 SNPs (Figs. 3.9a-3.9c) are predicted to have 2 separate oligo clusters (k = 2), while the remaining 2 SNPs (Figs. 3.9d-3.9e) each have 3 separate oligo clusters (k = 3).

As observed excluding noisy oligos (such as the blue oligo cluster in Figure 3.9c) can lead to significant improvements in the estimated SNP signal intensities. However, filtering oligos does not always improve the magnitude of LR deviation. For instance, excluding the blue oligo cluster from the probe set of  $SNP_{81}$  (Figure 3.9b) does not significantly improve the estimated SNP LR value (compare LR and LR' values in Fig. 3.9b). Figure 3.6 also confirms that  $SNP_{81}$  does not provide any deletion information to begin with (regardless of the reference set). Therefore, the oligo-level processing cannot have a significant impact on the quality of the estimated signal from this SNP.

#### 3.3.3.2 Performing Oligonucleotide Probe-level Analysis of the SNP Array Data

As explained in Section 3.2.2.4, for each oligo cluster in a SNP probe set three null-hypothesis tests were applied to assess whether the oligo cluster indicates a significant LR deviation from the baseline or not. Figure 3.10 provides an example of clustering and likelihood estimation for a given SNP on the Xba array (SNP\_A-1740765, panel (A)). The fuzzy-kmeans clustering found 2 separate PM oligo clusters in the SNP probe set, which are denoted by red and blue colors in

<sup>&</sup>lt;sup>1</sup>In fact, based on empirical data, I observed that using a larger reference set can have a negative impact on the ability to detect somatic changes in cancer samples.

Panel (B). Panels (D) and (E) show the results of null hypothesis tests for each of the above 2 oligo clusters (see Sec. 3.2.2.4 for the description of these tests). The generated data from these tests are passed as input data to the downstream QDA classifier; detailed in the next section.

#### 3.3.3.3 SNPs Classification and LR Estimation

The clustering of SNPs in multiple arrays revealed that less than 5% of SNPs have only 1 oligo cluster (k = 1). For these SNPs no further pre-processing is applied and the center of the only predicted oligo cluster is used as the SNP-level LR readouts. The remaining SNPs that have more than 1 oligo cluster are passed through QDA classifiers (Equation (G.2)). The QDA classifiers were trained using 108 SNPs from validated regions of copy number loss [29] and the same number of SNPs from validated regions of copy number gain, as well as 108 SNPs from putative normal regions<sup>1</sup>. As more regions of deletion and amplification are validated, the user can add the validated SNP data to the training set. This feature enables the algorithm to potentially improves its own decision boundaries. The input data to the QDA classifiers include the p-values of the KS-tests from likelihood estimation phase, the number of PM oligos in each oligo cluster and their centroid values.

#### 3.3.3.4 Comparison of OPAS Pre-processing and Naive Bayes Classification

As explained earlier in Section 3.2.3, an alternative approach was suggested for preprocessing Affymetrix SNP probes based on Bayes classification and without oligo clustering. The same dataset that was used to train QDA classifier (324 SNPs from known deleted, amplified and normal regions<sup>2</sup>) is also used for training the Naive Bayes classifiers. Table 3.3 presents the comparisons between the performance of QDA and 2 Naive Bayes classifiers (described in Section 3.2.8), using 567 SNPs in 19 regions of copy number deletion in follicular lymphoma samples that were validated by Illumina sequencing<sup>3</sup> (500K data). The aims of this analysis were: 1) to compare the the proportion of the 567 SNPs that were accurately detected as deleted by each classifier; and 2) to compare the magnitudes of estimated LR values by each approach. The summary of the results, presented in the last row of Table 3.3, indicates that the OPAS number of false negative deletions was 151/567 (27%); however, both of the Naive Bayes classifiers showed lower deletion sensitivity with 217/567 (38%; kernel fit) and 208/567 (37%; normal fit) false negative deletion calls (these

<sup>&</sup>lt;sup>1</sup>The exact copy number values for these regions were not experimentally determined, however all these SNPs had  $LR \approx 0$  and manually inspected prior to adding them to the training set.

 $<sup>^{2}108 \</sup>times 3 = 324$ 

<sup>&</sup>lt;sup>3</sup>These data are from analysis of Follicular Lymphoma CNVs, described in Chapter 4.

are deletion calls at SNP level and not CNV level. It is also observed that for some of the deleted regions QDA has a noticeably better sensitivity compared to Naive Bayes classifiers. For instance, for region #4 in Table 3.3, QDA predicted 101/152 (66%) SNPs as deleted; however, Naive Bayes classifiers predicted 71-72 (~47%) deleted SNPs.

The suggested OPAS modification excludes oligo clustering phase (see Fig. 3.2); therefore, the mean log2-ratio intensity of all PM oligos with LR < 0 were used as LR estimate of the SNPs that were classified as deleted by Naive Bayes classifiers<sup>1</sup>. The mean Naive Bayes estimated LR values of all SNPs within the deleted region is shown in columns 10-11 (LR<sub>NBN</sub> and LR<sub>NBK</sub>) of Table 3.3. This table also shows the original LR values based on average log2-ratio intensity of all PM oligos in the SNPs probe sets (column 9;  $\overline{LR}$ ), as well as the OPAS-estimated LR values (column 12; LR<sub>OPAS</sub>). Comparing the estimated LR values (columns 9-12) in Table 3.3 reveals that for all 19 sequence-validated deletions, the OPAS pre-processed LR values ( $\overline{LR}_{OPAS} = -1.01$ ) have noticeably larger magnitude of copy number loss compared to the mean probe set LR measurements before pre-processing ( $\overline{LR} = -0.43$ ). The magnitude of OPAS-generated LR values are also larger than the LR estimates based on Naive Bayes classifiers (LR<sub>NBN</sub>  $\approx$  LR<sub>NBK</sub> = -0.59). The boxplots in Figure 3.13 also demonstrate noticeable differences between LR values of the same deleted regions based on different approaches. These boxplots indicate that OPAS-estimated LR values (boxplot 4) have larger magnitudes of copy number loss compared to LR measurements based on Naive Bayes classifiers (boxplots 2-3). It is also evident from this figure that the LR values of the deleted regions are significantly improved after OPAS SNP pre-processing analysis (compare boxplots 1 and 4 in Fig. 3.13). The latter observation confirms the initial hypothesis that oligo-level analysis of SNPs can lead to improve LR estimates.

It is also important to investigate whether the higher QDA sensitivity to call deleted SNPs compared to Naive Bayes classifiers is because QDA approach has a poor false positive rate for deletion SNP calls. This can be assessed by randomly selecting SNPs from regions that are putatively normal and comparing the number of such SNPs that are predicted as deleted by each classifier. To implement the above analysis, 567<sup>2</sup> SNPs were randomly selected from the autosomes of two FL samples (any known CNV regions were excluded from these autosomes prior to the random SNP selection). In the rest of this analysis, these 567 randomly selected SNPs are referred to as 'putative normal SNPs'. The QDA classification of OPAS and the Naive Bayes classifiers were then applied on these putative normal SNPs the number of deletion calls were estimated. The results of

<sup>&</sup>lt;sup>1</sup>For those SNPs that were classified as normal by Naive Bayes methods, the mean of all PM oligos in the probe set with  $-0.1 \le LR \le +0.1$  was used as the SNP's LR measurement.

<sup>&</sup>lt;sup>2</sup>the same number of SNPs in 19 deleted regions of Table 3.3

OPAS analysis suggest that 35.4% (201/567) of the aforementioned putative normal SNPs appear to indicate loss of signal intensity. The Naive Bayes classifiers with normal distribution and kernel fit, respectively, predicted 31.2% (177/567) and 39.5% (224/567) of these putative normal SNPs as deleted. These findings indicate that all 3 classifiers generated almost the same proportion deleted SNP calls among 567 putative normal SNPs (31.2%, 35.4% and 39.5%).

In conclusion, Table 3.3, Figure 3.13, and the findings of the latter analysis collectively suggest that using Naive Bayes for SNP pre-processing phase provides no substantial advantage for SNP pre-processing. Nonetheless, the two Naive Bayes classifiers have also been added as optional features of OPAS that can be selected by the user instead of the default pre-processing method.

#### 3.3.3.5 An Example of the Impact of SNP Pre-processing on Improving CNV Data Quality

As described in Chapter 1, a major limitation of microarray data smoothing is that such approaches often tend to smooth the entire signal aggressively in order to eliminate noisy artifacts; and therefore, may also suppress the magnitude of true CNVs [220, 221]. To investigate if such problems exist in OPAS outputs, the distribution of log2-ratio fluorescent intensity readouts of 540 PM oligos from 45 Nsp SNPs in a known deleted region were analysed before and after pre-processing (this deletion has been detected in an MR patient and was validated by FISH [30]).

Figure 3.11 illustrates the CDF plot of these 540 deleted oligos before and after pre-processing, shown by black and red curves, respectively. This figure also depicts the CDF of all oligos on the array (blue curve) that is considered as a baseline distribution for the copy number analysis<sup>1</sup>. As observed, the CDF of deleted oligos after pre-processing (red curve) falls further left of the distribution of the same oligos before pre-processing (black curve).

In summary, Figure 3.11 indicates that SNP pre-processing improved both the proportion of informative oligos and the magnitude of LR deviation from the baseline in the aforementioned known CNV region. As the result, there is a wider separation between the LR distributions of deleted and baseline oligos. Such improvements can consequently lead to better CNV detection accuracy in the downstream copy number analysis. Section 3.3.6 details the impact of such oligo-level improvements on the downstream CNV calling accuracy based on a comparative analysis using simulated data.

<sup>&</sup>lt;sup>1</sup>No other large-scale CNV or an euploidy was discovered in this MR patient; therefore, it is reasonable to consider that the distribution of whole-genome microarray data (> 3 mil) represents a copy number normal baseline.
#### 3.3.4 Results of OPAS Post-processing Phase

Analysing estimated SNP signals from different experiments found that sometimes a considerable correlation exists between the LR values and PCR fragment length data, such as shown in Figure 3.12a (r = 0.9, with p-value P = 2e-5). To correct for such biases LOWESS non-parametric normalization is applied on SNP pre-processed data (described in Section 3.2.4.1). The effect of LOWESS normalization in removing the fragment length bias is shown in Figure 3.12 [180].

Next, to find regions with multiple SNPs that exhibit a statistically different LR measurements, Circular Binary Segmentation algorithm is applied on the modified SNP data (Fig. 3.1). As described in Section 3.2.4.2, this entirely non-parametric method splits the chromosomes into contiguous regions of equal copy number by modelling discrete copy number gains and losses through permutation analysis [189, 242]. In OPAS design, a default of 10,000 permutations are used to estimate the p-values that define the significance of a segment split.

Although, in theory any detected segment with LR deviation from the baseline (|LR| > 0) presents a putative CNV, in practise other factors can contribute to fluctuations in LR intensities that are statistically significant but do not represent a real biological event (e.g., DNA quality). Therefore, in a classification based CNV calling method a second filter is used to choose segments which likely indicate a real copy number change; and this filter often depends on the LR value. For example, in a study of ovarian tumor samples, regions with log2-ratio > +0.3 or < -0.3 were used as candidate CNVs [366] (direct use of LR cut offs for CNV calling); and in another study, z-scores of the segmented read depth coverage data were used to call putative CNVs from genome sequence data [195] (z-scores also depend on the distribution of LR values). From the biological perspective, we also have some prior knowledge about the frequency and extent of genomic CNVs in different samples. It is well-known that cancer DNA is frequently affected by genome rearrangements compared to copy number polymorphisms among normal individuals [18, 19, 24, 25, 37, 69]. Therefore, instead of using predefined cut-offs to label segments as candidate CNVs, OPAS provides the list of all segments in the genome (Fig. 3.3) and basic analysis of the distribution of these segments (meta-data). The proposed approach is to use both the OPAS-generated data and the biological information to assign proper statistical cut-offs for each separate study, depending on the type of samples that are being analysed (e.g., cancer or mental retardation).

#### 3.3.5 Assessing the Effect of Noise on CNV Calling Performance

To evaluate the robustness of the CNV calling in the presence of noise, the biologically inspired model described in Section 3.2.7 was used to generate a dataset of simulated noisy signals. These

simulated noisy signals represent data from chromosome 14 of a sample with known IGH deletion<sup>1</sup> (ht-17).

Figure 3.16 presents the effect of increasing the  $\sigma$  of the underlying noise on the performance of the CNV calling algorithm based on 9 different levels of noise ( $\sigma_{noise} = .01, .03, .05, .09, .13, .15, .17, .19, .21$ ). For each specified  $\sigma_{noise}$ , 10 random simulated signals were generated and analysed using OPAS method. The number of times that the known IGH deletion was not detected in a simulated signal with  $\sigma_{noise}$  divided by the total number of generated signals with the same  $\sigma_{noise}$  (n = 10) is referred to as the detection error rate ( $\eta$ ). The signal-to-noise ratio (SNR) at each corresponding noise level is also generated based on the estimated amplitude of the original signal and the known added noise.

The result of this analysis is shown in Figure 3.16. As expected, it is observed that the detection error rate ( $\eta$ ) increases with the standard deviation of the noise. This figure also denotes that when  $\sigma_{noise}$  reaches 0.17 (marker 'B' in Fig. 3.16) the IGH deletion is no longer detected in any of the simulated signals ( $\eta = 1$ ). Further study of this particular data point revealed that at  $\sigma_{noise} = 0.17$  the SNR is approximately equal to 1 (SNR=1.03); and the noise and the original signal have approximately the same standard deviation ( $\sigma_{signal} = 0.1694$  and  $\sigma_{noise} = 0.17$ ). This finding implies that when the magnitude of the simulated noise is equal to or greater than the original signal, the algorithm cannot distinguish between the distribution of the deleted SNPs and the rest of the data.

It must be noted that the above analysis found no false positive CNV calls in any of the simulated signals (n = 90). Therefore, I continued this experiment by increasing the magnitude of the added noise in simulated chromosome 14 signals and analysing the estimated CNV calls. The first false positive CNV call was detected when the amplitude of the added noise was more than two times larger than that of the original signal ( $\sigma_{noise} = 0.35$  and  $\sigma_{signal} \simeq 0.17$ ). In conclusion, the findings of these analyses suggest that despite the increased level of noise, the CNV calling algorithm yields reasonable results.

<sup>&</sup>lt;sup>1</sup>The validated *IgH* deletion is located on chromosome 14q32.33, as shown in Figure 3.4.

#### 3.3.6 Comparative Analysis of CNV-calling Algorithms

A simulated model was implemented according to the method described in Section 3.2.6 (see Figure 3.14). OPAS, SMD and GLAD were run with default parameters, as most users will do. For OPAS and GLAD, all regions with LR < 0 and LR > 0 were selected as candidate deletions and amplifications, respectively. Based on SMD recommendations, two different p-value cut-offs (p-values  $\leq 1e-6$  and  $\leq 1e-8$ ) were used to generate two separate lists of SMD results<sup>1</sup>. The results from each dataset were compared to the known simulated CNV regions for each alteration size (w = 2, 4, 8, 10, 15, 25, 100, 200) and LR ratio response ( $\delta = 0.11, 0.2, 0.4, 0.6, 0.8, 1.0$ ).

The results from each dataset were compared to the known simulated CNV regions for each alteration size (*w*) and LR ratio response ( $\delta$ ). True positives, false positives and false negatives were aggregated for each algorithm and simulation to evaluate the sensitivity (true positive rate; TPR) and precision (true positive predictive value; PPV) estimates. As shown in Figure 3.15, all algorithms offered similar sensitivity and precision for alterations with log2-ratio shifts larger than 0.6 ( $\delta \ge 0.6$ ) that had more than 10 to 15 SNP probe markers (panels D-E and I-J; Figure 3.15). However, detecting CNVs with fewer number of SNPs (w < 10) was largely dependent on the magnitude of the LR deviation ( $\delta$ ) and the method used to analyse the data. For example, Figure 3.15.A shows that all methods had substantially low sensitivity for detecting simulated CNVs with slight LR deviations which also contained fewer than 10 SNP probe markers ( $\delta = 0.11$  and w < 10). This plot also denotes that the OPAS sensitivity for simulated CNVs with LR deviation of 0.11 improved when these CNV regions contained at least 10-15 SNP probe markers (Panel (A)). Panel (F) shows that the estimated precision (PPV = TP/TP + FP) of detecting CNVs with only slight LR deviation ( $\delta = 0.11$ ) was also low, regardless of the algorithm used to analyse the simulated data.

As expected, in each plot the detection sensitivity and precision increased with the number of SNP probe markers within the simulated regions (*w*). The main conclusion of the analysis presented in Figure 3.15 is that in mid-range LR deviations ( $\delta = 0.2 - 0.6$ ) OPAS provides a noticeably better sensitivity and precision for detecting simulated CNVs with fewer than 10 SNP probes (see panels (B-C) and (G-H)).

<sup>&</sup>lt;sup>1</sup>The SMD documentation indicates that generated putative CNVs with p-values  $\leq 1e-8$  have the lowest false positive rate but a higher rate of false negatives; on the contrary, results with  $p \leq 1e-6$  have better CNV detection sensitivity but  $\sim 40\%$  false positive rate.

#### 3.3.7 Comparing OPAS and CBS Accuracy

To investigate whether the observed improvements in CNV detection accuracy with fewer SNP probe markers are due to the OPAS approach in dealing with noisy oligos and not CBS segmentation alone, the performance of OPAS and CBS were compared using the model previously described in Section 3.2.8. A series of simulated signals was generated by randomly selecting *N* oligos from a known deleted region (on chromosome 2p16.3 of an MR patient) and eliminating them from the Xba SNP array data (the deletion includes 17 Xba SNP probe markers). Fourteen different *N* values ( $N = \{2, 3, ..., 15\}$ ) were used to generate these simulated signals. This process was repeated 100 times for each value of *N*, resulting in a total of 1,400 simulated signals with 2 to 15 SNP probe markers within the known deleted locus ( $\iota$ ).

The simulated signals were analysed with OPAS and CBS, independently, and segments with log2-ratio < 0 were considered for further analysis. The boundaries of these segments (with LR < 0) were then compared with the known 2p16.3 deleted region and those with > 60% coverage were considered as putative correct calls. Table 3.4 and Figure 3.17 show the results of this analysis. As observed, OPAS detected almost all deletions that had 8 or more SNP probe markers (except for 1 false negative call in 800 signals with  $t \ge 8$ ; see row 7 of Tab. 3.4), and CBS detected almost all deletions with at least 9 SNP probe markers (except for 2 false negative calls; see rows 5-6 of Tab. 3.4). The results of this analysis also indicate that as the number of remaining SNPs in the deleted region (t) drops below 9, the methods start to show increasingly different CNV detection accuracies (see the pink dashed line in Fig. 3.17a).

As previously described in page 85, the accuracy of detecting CNVs using a segmentation based approach not only depends on the ability of the algorithm to find a segment that maps to the real CNV event, but also its estimated magnitude of LR change. Therefore, to perform an accurate comparison between OPAS and CBS, the estimated LR values of the known deleted region (within the simulated signals) are also presented in Figure 3.17b and Table 3.4. Comparing OPAS and CBS estimated LR values shows significant improvements in the magnitude of copy number loss as the result of OPAS-analysis.

For instance, the analysis results in Figure 3.17a and Table 3.4 suggest that OPAS detected the known deletion in 92 of the total 100 (92%) simulated signals that had only 3 deleted SNP probes (t = 3). However, CBS detected this deletion in 58 (58%) of these simulated signals. The LR data of the same simulated signals (t = 3), presented in Figure 3.17b and Table 3.4, reveal that CBS and OPAS estimated average LR of the deleted regions is approximately -0.1 and -0.82,

respectively<sup>1</sup> (see columns 6 and 4 of Table 3.4). These findings suggest that eventhough CBS found an overlapping segment in 58% of the aforementioned  $\iota = 3$  signals, unlike OPAS results, these segments do not appear as significant and reliable deletion calls. Collectively, these data indicate that OPAS approach resulted in significant loss of signal intensities for the known deleted region in more than 90% of the signals that had only 3 SNP probe markers within the deleted boundaries (2p16.3).

In summary, the results of this experiment confirm the initial hypothesis that OPAS oligo-level data processing has a major impact on the accuracy of finding CNVs with fewer SNP probe markers and that these improvement are not the result of the CBS segmentation alone (Table 3.4 and Figure 3.17). This conclusion is also supported by the results of CNV analysis in follicular lymphoma patients, presented in the next Chapter, where OPAS detected several real CNVs (validated by FPP or Illumina sequencing) that were not identified with several alternative methods.

# 3.4 Conclusions

To investigate the sources of variability in SNP arrays, in this Chapter I studied several factors that influence signal intensity readouts in Affymetrix GeneChip SNP arrays (Sections 3.3.1-3.3.2), such as analysing the impact of reference set size on the magnitude of copy number deviation. This analysis revealed that in contrast to what may be expected, a larger reference set does not necessarily yield a better CNV detection sensitivity (Tab. 3.2). Based on the observed results and the results of analysing SNP probe sets (Fig. 3.5), I hypothesized that processing the SNP array data at the oligo-level could improve the accuracy of the downstream CNV detection. To implement this idea, I developed the algorithm for Oligonucleoytide Probe-level Analysis of Signal intensities or OPAS (Fig. 3.1).

In the first step of OPAS, the raw signal intensities between test and reference samples are log-transformed and normalized using quantile normalization (described in Section 3.2.2.1 and Appendix F). As mentioned in Section 3.2.2.1, a possible theoretical problem with this approach is the risk of removing some of the signal in the tails of the distribution; however, several studies have shown that empirical data does not support this hypothesis [205, 302]. To investigate this problem, a comparative analysis of the impact of normalization on the estimated log2-ratio values was performed; as described in Appendix F. This analysis applies 5 normalization methods on SNP array data (500K) from 8 follicular lymphoma samples and compares the average LR values of 50 sequence-validated CNVs in these samples. The results of this comparison did not provide

<sup>&</sup>lt;sup>1</sup>in both cases  $LR_{baseline} \simeq 0$ 

any substantial evidence to support the hypothesis that quantile-based normalization method suppresses the magnitude of real CNVs in the aforementioned follicular lymphoma samples. As seen in Table F.2, contrary to what was expected, the quantile-based OPAS normalization often showed an increase in the magnitude of real CNVs in these patients. Nonetheless, the 5 normalization approaches in Appendix F have been added to the OPAS software and the user can change the default pre-processing normalization method or replace it by a new user-defined function.

Following normalization, for each SNP on the Affymetrix array, the PM oligos within the SNP probe set were separated into groups with similar LR values based on a two-level clustering approach (Fig. 3.10; panels A-B). The proposed clustering method, referred to as fuzzy-kmeans clustering (Section 3.2.2.2), first uses a fuzzy inference system (subtractive algorithm) to model the probe set data behavior through a minimum number of rules and then uses this information to initialize a k-means optimization-based clustering algorithm (Fig. 3.9). Each estimated cluster of PM oligos (i.e., oligo cluster) is subsequently passed through a hypothesis testing model (Section 3.2.2.4) that performs KS-tests to test the null-hypothesis that the distribution of the oligo cluster is around the normal copy number baseline (panels C-E in Fig. 3.10). These estimates are then passed to a QDA machine learning classifier (Section 3.2.2.5) to identify the SNP's oligo cluster with the most informative signal and subsequently the center of this oligo cluster is used as the SNP-level log2-ratio (LR) measurement. In the post-processing phase, first a LOWESS non-parametric method is applied on the estimated SNP LR values to remove the potential PCRfragment length biases (Sec. 3.2.4.1, Fig 3.12). The generated OPAS meta-data can also help to identify some potential issues for specific samples. Analysing the meta-data can help building quality assurance controls.

It is important to note that in more than 600 analysed samples (from mental retardation, follicular lymphoma, as well as normal individuals) only about 5% of the cases showed a strong PCR-fragment length bias ( $r \ge 0.9$ ; such as shown in Figure 3.12a). Possible causes for such correlation is that experimental issues with PCR process or poor quality of the sample genomic data (problems at the experiment-level) have resulted in such bias. Although this correlation is computationally improved (Fig. 3.12b), there is no guarantee that the quality of the generated data in such experiments is at the same level of samples that showed no such experimental biases. The generated OPAS meta-data, such as intensity scatter plots before and after PCR-fragment length normalization that are automatically generated and stored during sample analysis, can potentially help to track down problematic samples. Although building proper quality assurance controls is beyond the scope of this thesis, analysing the generated meta-data can provide further information about the samples and may potentially facilitate quality control process. The final module of OPAS, shown in Fig. 3.1 flowchart, is non-parametric CBS segmentation (Section 3.2.4.2). The aim of applying CBS segmentation is to identify neighboring regions of DNA that exhibit a statistically significant difference in their average signal intensities.

To assess the CNV calling accuracy, OPAS, GLAD and SMD results were compared using a simulated Nsp data set (described in Section 3.3.6). The results of this analysis indicated that all methods had similar sensitivity and precision for LR shifts larger than 0.6 ( $\delta > 0.6$ ), and regions with more than 10-15 SNP probe markers (Sec. 3.3.6). The analysis also found that for midrange shifts ( $0.2 \leq \delta \leq 0.6$ ), the performance of OPAS for detecting CNVs with fewer than 10 SNP probes was noticeably higher than the other methods (Fig. 3.15). The latter finding implies that OPAS has a better accuracy in detecting CNVs with fewer probes compared to GLAD and SMD. To investigate if this improved performance was due to OPAS oligo-level pre-processing analysis and not the CBS segmentation alone, another experiment was implemented to compare the accuracy of both methods in detecting a known CNV with variable number of SNP probe markers (Sec. 3.2.8 and Sec. 3.3.7). For this analysis 1,400 signal were generated with 2-15 SNP probe markers within a known deleted region on chromosome 2p16.3 of a patient with mental retardation. The results of this analysis (Figure 3.17, Table 3.4) also supported the hypothesis that OPAS oligo-level processing of SNP data has a major impact on the accuracy of finding CNVs with fewer SNP probe markers, which is not achieved by using CBS segmentation alone.

In addition to the simulated data analysis, described in Sec. 3.3.6, OPAS was applied on data from 146 patients with mental retardation and the predcited CNVs were compared to a list of 30 validated CNVs in the same patients that were previously found by integrating the results of 7 alternative copy number algorithms [215]. The results of this analysis, presented in Table I.1, revealed that OPAS detected all of these 30 validated CNVs in addition to 52 extra putative CNVs in these MR patients. While there is no biological verification for the new putative events, Ingenuity Pathway Analysis found genetic disorder, neurological disease and behavior are the most significant functions associated to these candidate MR CNVs. The pathway analysis results provide additional confidence to OPAS findings in the MR dataset in the absence of experimental validation.

Furthermore, I also used a biologically inspired model to evaluate the performance of CNV calling in the presence of controlled added noise (on chromosome 14 IGH locus; Fig. 3.4). The results of this analysis showed that the average error rate ( $\eta$ ) increased with the magnitude of simulated noise and reached 100% when the signal-to-noise-ratio was greater than 1 (Fig. 3.16). The first false positive CNV call among these simulated signals was observed when SNR  $\geq$  3. Based on these findings, it can be speculated that the CNV calling algorithm yields reasonable sensitivity and specificity to detect real CNVs.

In the next Chapter (Chapters 4), I will apply OPAS to study CNVs in 25 patients with follicular lymphoma and will use several alternative data sets to compare the predicted CNV results. This analysis will confirm high sensitivity of the OPAS non-parametric approach in identifying small CNVs with only a few SNP probes that were otherwise cryptic to alternative SNP CNV analysis methods (for instance, OPAS detected several deletions with 4-8 SNPs that were validated by sequencing; p. 132). Such findings support the underlying hypothesis of OPAS design that a probe-level copy number analysis approach can improve CNV detection accuracy in Affymetrix SNP arrays, particularly for those events with fewer SNP probe markers.

# 3.5 Figures and Tables



**Figure 3.1:** Flowchart of the OPAS algorithm: The OPAS algorithm has 2 main modules, pre-processing (shown by red arrow) and post-processing (shown by yellow arrow). The first phase, pre-processing, aims to detect the most informative PM oligos within each SNP probe set in order to improved SNP log2-ratio intensity (LR) readouts. The main modules in SNP pre-processing are quantile normalization (Section 3.2.2.1), fuzzy K-means clustering (Section 3.2.2.2), likelihood estimation (Section 3.2.2.4), and QDA-based machine learning classification (Section 3.2.2.5). The goal of the next phase, post-processing, is to partition the whole genome into regions where the copy number changes between contiguous segments. In this phase, the SNP-level LR intensities are first subjected to a LOWESS normalization to remove the systematic biases induced by impact of PCR fragment length on the estimated SNP signal intensities (Section 3.2.4.1). Subsequently, the LR values are passed to CBS segmentation algorithm (Section 3.2.4.2). Analysing the mean log2-ratio values of these segments or their z-scores can help to identify candidate CNV regions.



**Figure 3.2: Flowchart of the alternative approach for SNP pre-processing based on Naive Bayes classification.** The modifications suggested to OPAS pre-processing phase are superimposed on the default algorithm flowchart (previously shown in Fig. 3.1). The blue boxes on the left of the main flowchart denote the modifications. The red boxes (with greyed-out text) are the steps of default OPAS pre-processing phase that are being replaced by these modifications. As seen in this plot, clustering is not applied on SNP probe sets; and instead, all PM oligos in each probe set are passed to a Naive Bayes classifier. The aim of this classifier is to determine whether the log2-ratio intensity of each SNP probe set is similar to those SNPs from known deleted, normal or amplified regions. Since clustering information in not available in this approach (blue modules), log2-ratio values of SNPs are estimated based on the method explained in Section 3.3.3.



**Figure 3.3:** Schematic representation of OPAS input/output data. As shown in this figure, the Affymetrix GeneChip raw .CEL intensity files are the only user-input data to the OPAS algorithm (OPAS allows both single and batch imports). During sample analysis, several different meta-data are generated which are automatically stored on the server or the designated storage space. The output .txt file includes a list of all segments in each 23 chromosomes of the analysed sample. The generated CNV plots (.jpg files) provide a means to visualize CNVs along chromosome ideograms that are generated based on the UCSC genome browser (the visualization also accommodates switching between different versions of UCSC genome builds). An additional script enables generating UCSC custom tracks for any list of OPAS putative CNVs, according to BED formatting (.BED files). During sample analysis, a series of other figures and data (referred to as meta-data) are also generated and automatically saved at a pre-allocated space on the server (such as PCR fragmentation length normalization plots; and probability density plots of the estimated LR values). These generated meta-data and graphs provide a means to facilitate CNV interpretation and analysis, as discussed in Section 3.4.



(a) OPAS visualization output



(**b**) UCSC screenshot of the deleted region with FPP results

**Figure 3.4: The Nsp signal of chromosome 14 of a follicular lymphoma patient that harbors a deletion on 14q32.33** (~644 kb; 9 SNPs) and is used as the template to generate simulated noisy signals. Panel (a) shows OPAS visualization output of chromosome 14 of a follicular lymphoma patient (patient 17; obtained from Chapter 4 CNV analysis). The black arrow highlights a predicted deletion, approximately 644 kb with 9 Nsp SNP probe markers, located on chromosome 14 IGH locus. Panel (b) demonstrates the screenshot of UCSC genome browser illustrating the fingerprint profiling (FPP) alignment of BAC clones in chromosome 14q32.33 of this sample (ht-17) to the reference human genome (hg18). BACs with linear alignments to the reference genome are coloured blue and the ones with split alignments are coloured green (see Appendix K for more information). The two blue arrows indicate that the ends of BAC clone HTa17-0164B08 align to different loci on chromosome 14, suggesting that the region between these two arrows may have been deleted. The OPAS predicted region of copy number deletion in the same sample is shown in pink. The black vertical dashed lines denote the concordance between OPAS and FPP predicted boundaries of IGH deletion. The black arrow in (b) indicates that the above deletion in patient 17 was also detected by an alternative CNV calling algorithm (SMD [259]). These observations suggest that the predicted deletion of IGH locus in patient 17 is a real CNV event.



(a) SNP Scatterplot





Figure 3.5 (*previous page*): Example of oligo-level and SNP-level variability in SNP arrays (100K data). Panel (a) denotes the SNP scatterplot of a region on chromosome 10 in a child with developmental abnormalities that harbors a known deletion, highlighted by the red dashed lines. This deletion is  $\sim$ 353 kb in length and includes 25 Xba SNP probes. The deletion disrupts several genes, including the *SUFU* gene which has been associated with the observed phenotypes in this patient [365].

Panels (b)-(f) illustrate the probe sets of 5 SNPs in *SUFU* deleted region (SNPs 79, 81, 83, 93 and 100). The *x*-axis denotes the index of PM oligonucleotide probes in the SNP probe set (1, 2, ..., 20); and the *y*-axis demonstrates their fluorescent log2-ratio intensity readouts. The probe set plots (b-f) reveal that although all of the aforementioned 5 SNPs are within the *SUFU* deleted region, there is a wide range of variability between their estimated LR values and standard deviations  $(-1.34 \le \overline{LR} \le -0.05, 0.3 \le SD \le 1.3)$ . Some of these SNPs have similar probe sets and LR values, such as SNP<sub>83</sub> (3.5c) and SNP<sub>93</sub> (3.5d) which both indicate significant loss of signal intensities (LR<sub>83</sub> = -1.34, LR<sub>93</sub> = -1.33). Compared to these 2 SNPs, SNP<sub>100</sub> (3.5e) has a relatively smaller magnitude of copy number loss (LR<sub>100</sub> = -0.5; Panel 3.5e). The colored dots in panel 3.5e denote the differences between the log2-ratio readouts of PM oligos in SNP<sub>100</sub> probe set; as detailed in page 79 (red and green dots show oligos with LR < -0.5 and oligos with  $0 < LR \le -0.5$ , respectively; while the blue dots show noisy oligos with LR > 0). The last plot in this figure (3.5f) illustrates SNP<sub>81</sub> probe set. As observed, the log2-ratio intensity readouts of the PM oligos in this probe set provide no substantial evidence to conclude that SNP<sub>81</sub> is deleted (LR<sub>81</sub>  $\simeq$  0).



Figure 3.6: Comparing the impact of the size of the reference set on estimated log2-ratio intensity readouts of SNP probe sets within *SUFU* deleted region. These plots demonstrate the probe sets of 10 SNPs in *SUFU* deleted region. Three of these 10 SNPs were also shown in the previous figure, including the non-informative  $SNP_{81}$  illustrated in Fig. 3.5f.

The plots in each row illustrate the PM oligonucleotide probe sets of the same SNP. Each column represents estimated log2-ratio measurements with respect to a separate reference set. These 6 reference sets include the mother of the affected patient (column 1), his father (column 2), the mean of both parents (column 3), and three other reference sets with 24, 99 and 150 normal samples ( $R_{24}$ ,  $R_{99}$ ,  $R_{150}$ ; columns 4-6). This figure indicates that, regardless of the reference set used to estimate LR values, there is a remarkable variation in the overall pattern of probe set LR intensities. It is also observed that using a larger reference set does not improve the probe set response of uninformative SNPs, such as SNP<sub>81</sub> (Fig. 3.5f) or SNP<sub>96</sub>.



**Figure 3.7: Comparison of cumulative density functions (CDFs) of all oligos in** *SUFU* deleted region with 6 reference sets with varying sizes. This figure illustrates the CDF of log2-ratio fluorescent intensities from 500 Xba PM oligos (representing 25 Xba SNPs) in *SUFU* deleted region (previously shown in Fig. 3.5). Six reference sets with varying sizes (Fig. 3.6) were used to estimate log2-ratio intensities of the PM oligos in this region. Each CDF is depicted by a different color, as detailed in the figure legend (see Section 3.3.2.1 for the description of these reference sets).

In addition to the CDF of deleted oligos, the CDF of all PM oligos on the Xba array (> 1.1 mil PM oligos) is also shown by the black curve (referred to as base-line CDF). The fact that all CDFs that represent *SUFU* deleted region fall on the left side of the base-line CDF emphasizes that no matter what reference set was used to estimate test versus normal ratios, the log2-ratio intensity distribution of the *SUFU* deleted region is smaller than, and distinct from, the LR distribution of the entire array. However, the extent of this distinguishability, which is determined by the deviation between *SUFU* and base-line CDFs, varies depending on the reference set that was used to estimate the LR values. It is observed that using the largest reference set ( $R_{150}$ ; denoted by the orange arrow) does not improve the distinguishability of *SUFU* deleted oligos from the rest of the array, compared to single-array reference set of  $R_{father}$  (denoted by the blue arrow).



Figure 3.8: Boxplots of the average number of CNV-affirmative PM oligos per SNP probe set  $(\overline{\theta})$ , with respect to 3 reference sets with varying sizes. Three reference sets with 24, 99 and 150 samples were used to estimated log2-ratio values of the SNPs from 22 regions of known copy number loss in mental retardation patients [29, 30]. Each boxplot presents the average number PM oligos in each SNP probe set with LR < 0 across 22 aforementioned deletions with respect to a separate reference set (the reference sets are denotes in the *x*-axis).

These boxplots indicate that when  $R_{24}$  is used, on average, 13.3 out of 20 PM oligos in each Xba SNP probe set indicate loss of signal intensity ( $\theta_{24} = 13.3/20 \simeq 67\%$ ). Using a larger reference set with 99 normal samples ( $R_{99}$ ) improves the number of CNV-affirmative oligos per SNP to 15.4/20 ( $\theta = 77\%$ ). However, when the largest reference set ( $R_{150}$ ) is used the average number of CNV-affirmative PM oligos per SNP probe set drops to 14.3/20 ( $\theta_{150} = 71\%$ ). This observation implies the number of PM oligos in the probe set of SNP within a deleted region that indicate loss of signal intensity, does not necessarily improve by increasing the number of reference samples.



Figure 3.9: Fuzzy-Kmeans clustering of PM oligos in 5 SNPs within the *SUFU* deleted region. Each plot depicts the probe set of an Xba SNP in *SUFU* deleted region that was previously shown in Figure 3.5 (each consisting of 20 PM oligos). The fuzzy-kmeans predicted oligo clusters in each SNP probe set are shown by different colors (red, blue and green). The mean of each oligo cluster is shown by a horizontal dashed line, with the same color as the corresponding oligo cluster. The red oligo cluster in each plot indicates a subset of PM oligos with the largest magnitude of copy number loss; while the blue oligo cluster denotes PM oligos with either mediocre loss of signal (as in 3.9c) or oligos that show positive mean log2-ratio values (noisy oligos). The reported  $\overline{LR}$  value for each SNP, is the average log2-ratio intensity measurement of 20 PM oligos in the SNP probe set; and  $\overline{LR}'$  presents the average SNP LR value after excluding the blue oligos from the probe set.



**Figure 3.10:** Schematic representation of oligo-clustering and likelihood estimation modules of OPAS default pre-processing. Panel (A) shows the probe set of a given Xba SNP, consisting of 20 PM oligos. This SNP is from a known deleted region in an MR patient. Panel (B) shows the result of fuzzy-kmeans clustering (Section 3.2.2.2) that found 2 separate oligo clusters in this SNP probe set (denoted by red and blue colors, respectively). These oligo clusters are then separately analysed using likelihood estimation module; described in Section 3.2.2.4. The results of the likelihood estimation phase are presented in Tables (D) and (E). The first 3 rows of these tables are the p-values of the null-hypothesis tests that were applied on each oligo cluster (Section 3.2.2.4). The last 2 rows of (D) and (E) are the number of PM oligos in each oligo cluster and the estimated cluster center. These data are passed to the downstream QDA classifier (Section 3.2.2.5) to determine the most informative subset of PM oligos in this SNP probe set.



(a) SNP scatterplot of a selected region in chromosome 6 that includes a known deletion (250 K Nsp array)



Variation of PM Probe-Level Intensity in 45 Deleted SNP on Nsp Array Before and After Pre-processing

(b) Boxplots of the standard deviation (SD) of PM oligonucleotide probes in the deleted SNP probe sets, before and after pre-processing



(c) Comparison of the distribution of oligo-level LR intensities before and after pre-processing

**Figure 3.11:** Distribution of PM log2-ratio intensities before and after pre-processing (500K data). Panel (a)-left shows the SNP scatter-plot of chromosome 6 of a patient with mental retardation that includes a validated deletion. This deletion (highlighted in red) spans ~353 kb and contains 45 Nsp SNP probes. Panel (a)-right denotes the oligo-level and SNP-level variabilities in 4 selected SNPs within this deleted region (these Nsp SNPs have 12 PM oligos). It is evident that there is a wide range of variability both between the SNPs and among the individual PM oligos in the same SNP probe set (similar to the observation in the 100K array data, shown in Fig. 3.5).

Panel (b) Shows the probe set variability of 45 SNPs within the deleted region, before and after OPAS pre-processing phase. This variability is estimated by assessing the standard deviation between PM oligos that belong to the same SNP probe set, before and after pre-processing (see Appendix E for description of boxplot visualization). These boxplots show more than 4.5-fold improvement in the oligo-level variabilities as the result of eliminating noisy oligos in pre-processing phase.

Panel (c) shows the CDF of all PM oligos in the aforementioned deletion ( $45 \times 12 = 540$  PM oligos), before (black curve) and after (red curve) pre-processing. Comparing these CDFs indicates that the rates of oligos with LR  $\leq -0.5$  before and after pre-processing is equal to ~91% and 71%, respectively. This indicates an improvement of approximately 20% in the rate of informative PM oligos as the result of SNP pre-processing analysis.



(a) Before PCR Fragment Length Normalization



(b) After PCR Fragment Length Normalization

Figure 3.12: The correlation between SNP log2-ratio intensities and PCR fragment length, before and after LOWESS normalization. Panel (a) illustrates the SNP log2-ratio intensities in an Xba array plotted against the corresponding PCR fragment length data. As seen in this figure, there is a significant correlation between the estimated LR intensity (y-axis) and PCR fragment length (x-axis) (r = 0.9, P = 2e-5). Panel (b) shows the same plot after correcting for PCR-induces biases using LOWESS non-parametric normalization. The symmetrical shape of this scatter-plot around the horizontal axis implies that the dependency of log2-ratio values on fragment length is remarkably improved after applying LOWESS fragment-length normalization.



Comparison of LR Estimates of 19 Regions of Copy Number Loss\*

Figure 3.13: Comparing estimated LR values of 567 SNPs in 19 validated regions of copy number loss based QDA and Naive Bayes classifications. These boxplots illustrate the log2-ratio intensity measurements of 567 SNPs in 19 Illumina sequence-validated deletions in follicular lymphoma samples (500K data). The first boxplot on the left shows the mean LR values of all SNPs (567) before pre-processing generated by averaging the log2-ratio intensities of all PM oligos in the SNP probe sets (after between-array normalization of test and reference samples). The remaining 3 boxplots represent the LR values of the aforementioned SNPs (567) based on different SNP pre-processing methods, as indicated below each bar (OPAS default QDA-based method, Naive Bayes with normal distribution and Naive Bayes with kernel fit). All 3 pre-processed data sets (bars 2-4) indicate improvements in the magnitudes of copy number loss compared to the original LR data before pre-processing (bar 1). It is also observed that OPAS pre-processing of the SNP-level data from deleted regions, in overall, has the largest magnitude of copy number loss compared to Naive Bayes classification results (LR<sub>OPAS</sub> = -1.09; LR<sub>NaiveBayes</sub>  $\simeq -0.59$ ).

<sup>\*</sup> validated by Illumina Sequencing



(a) OPAS estimated segments in a simulated chromosome with  $\delta = 0.11$ 



(b) OPAS estimated segments in a simulated chromosome with  $\delta = 0.2$ 

Figure 3.14: Examples of OPAS detected CNVs in simulated signals with different magnitudes of LR deviations. Panel (a) shows the OPAS-detected CNVs on a simulated chromosome 1 dataset that includes 11 predefined synthetic CNV regions with 2, 8, 10, 15 and 200 SNP probes (w). The results of OPAS analysis in detecting these simulated CNVs are denoted by pink and yellow circles. These colors indicate whether the simulated CNV was detected by OPAS (yellow) or not (pink). As seen in (a), from the 11 pre-defined CNVs in this signal ( $\delta = 0.11$ ), only 4 were detected by OPAS and they all had  $w \ge 15$ . Panel (b) shows the detection result of the same CNV regions in a simulated signal with  $\delta = 0.2$ . This figure (3.14b) shows that 10/11 simulated CNVs were detected by OPAS (the only case that was not detected had only 2 probes).



Figure 3.15: Comparing the accuracy and precision of CNV calling algorithms. Sensitivity (A-E) and precision (F-J) are obtained by applying OPAS, SMD (Delaney et al. [259]) and GLAD (Hupe et al. [210]) on simulated Nsp 250K array data (described in Sec. 3.2.6). This simulated dataset includes synthetic CNV regions with varying alteration sizes (*w*; *x*-axis) and log2-ratio intensity shifts ( $\delta$ ; denoted at the top of each plot). It is observed that all algorithms offer similar sensitivity and precision for alterations at log2-ratio shifts larger than 0.6 (D-E and I-J) and with more than 10 – 15 SNP probe markers. However, detecting CNVs with fewer number of SNP probe markers (smaller *w*) was largely dependent on the magnitude of the LR deviation ( $\delta$ ) and the method used to analyse the data.



Figure 3.16: Number of false negative deletion calls of the IGH locus, plotted against increasing noise of the simulated data. The bottom *x*-axis (shown in black) denotes the standard deviation (SD) of the random gaussian noise ( $\sigma_{noise}$ ) that was used to generate simulated chromosome 14 signals with known IGH deletions (Figure 3.4). The corresponding signal-to-noise ratio (SNR) for each value of  $\sigma_{noise}$  is also displayed at the top *x*-axis (shown in red).

The average error rate  $\eta$  (y-axis) at any given  $\sigma_{noise}$  is defined as the average number of cases that IGH deletion was not detected in simulated signals with the specified magnitude of noise (x-axis). This plot indicates that the error rate  $\eta$  increases with standard deviation of the noise ( $\sigma_{noise}$ ). At  $\sigma_{noise} = 0.17$  (marker 'B') the IGH deletion is not detected in any of the 10 simulated signals ( $\eta = 1$ ). Investigating the top red axis, reveals that at this level of noise (marker 'B') the signal-to-noise ratio is approximately equal to 1 (SNR = 1.03). This observation implies that when the amplitude of the added noise is equal to or larger than the original signal, the known IGH deletion is not detected in any of the simulated noisy signals.



Comparing OPAS and CBS Accuracy in Detecting a Known Deletion



(b) Comparison of the estimated LR values

**Figure 3.17: Results of comparing CBS and OPAS performance in detecting a known deletion.** Panel (a) compares the performance of OPAS and CBS in detecting a known deletion with respect to the number of SNPs in the deleted region. The *x*-axis denotes the number of SNP probes within the known deleted region after *n* random SNPs from the original 17 SNPs in this deleted region were excluded from the Xba data (n = 1, 2, ..., 15). The *y*-axis in Panel (a) denotes the total number of cases in 100 trials that an algorithm found a segment that mapped to the known deleted region (with at least 60% coverage of both regions). This plot shows that OPAS and CBS both present similar performance in detecting the known deletion with more than 9 - 10 SNP probe markers; however, there is a noticeable difference between the accuracy of the methods when there are fewer than 9 SNP probe markers in the region.

Panel (b) shows the LR values of the deleted region based on the original mean SNP signal (black curve), as well as OPAS (red curve) and CBS (blue curve) methods. The *x*-axis is the same as the previous plot and the *y*-axis represents the average estimated LR value of the known deleted region across 100 trials. The green horizontal dashed line illustrates the copy number normal base-line (LR=0). These plots indicate that the magnitude of the generated copy number loss decreases as more probes are removed from the known deleted region; however, OPAS shows a consistently larger magnitude of copy number loss, compared to CBS segmentation approach.

Set	Null Hypothesis (H <sub>0</sub> )	Alternative Hypothesis (H <sub>1</sub> )	
1	$X = y_0$	$X \neq y_0$	"two.sided"
2	$X \leq y_1$	$X > y_1$	"greater"
3	$X \ge y_2$	$X < y_2$	"less"

Table 3.1: Three sets of hypothesis tests used in null-likelihood phase. Test set #1 tests the alternative hypothesis that the CDFs of the oligo cluster data and the background population are not equal  $(H_1 : X \neq y_0)$ . The next KS-test examines the alternative hypothesis that the CDF of oligo cluster data (X) is larger than the CDF of the specified background  $(H_1 : X > y_1)$ , and the last test examines the alternative hypothesis that the CDF of X is smaller than that of the background  $(H_1 : X < y_2)$ . The significance level  $\alpha = 0.05$  is used in all of the above tests.

	<b>R</b> <sub>father</sub>			R	mother		R <sub>parents</sub>		
SNP #	LR	$n_1$	<i>n</i> <sub>2</sub>	LR	$n_1$	$n_2$	LR	$n_1$	<i>n</i> <sub>2</sub>
SNP <sub>79</sub>	-0.89	11	10	0.14	10	1	-0.38	10	10
SNP <sub>81</sub>	-0.05	10	2	0.05	8	0	0.00	11	0
SNP <sub>83</sub>	-1.34	19	12	-1.26	16	12	-1.30	19	12
SNP <sub>93</sub>	-1.33	17	12	-0.90	11	10	-1.12	12	10
SNP <sub>100</sub>	-0.51	14	9	-0.54	14	10	-0.52	18	9
				(a)					

	R <sub>24</sub>				R99	R <sub>150</sub>			
SNP #	LR	$n_1$	<i>n</i> <sub>2</sub>	LR	$n_1$	<i>n</i> <sub>2</sub>	LR	$n_1$	$n_2$
SNP <sub>79</sub>	-0.40	10	10	-0.85	16	10	-0.74	13	10
SNP <sub>81</sub>	0.14	10	6	0.16	9	6	0.14	10	6
SNP <sub>83</sub>	-1.00	17	13	-1.40	19	14	-1.32	19	13
SNP <sub>93</sub>	-0.60	10	10	-0.61	10	10	-0.59	10	10
SNP100	-0.86	17	11	-1.14	17	11	-1.01	17	11

 $n_1$ : number of PM oligos in the SNP probe set with LR < 0

 $n_2$ : number of PM oligos in the SNP probe set with LR < -0.5

**(b)** 

**Table 3.2:** The impact of the size of reference sets on the estimated SNP signals. These tables present mean log2ratio ( $\overline{LR}$ ) probe set intensities of 5 SNPs in *SUFU* deleted region that were previously shown in Figure 3.5 (SNPs 79, 81, 83, 93 and 100). Six reference sets with variable number of reference samples (1, 2, 24, 99 and 150 samples) were used to estimate the reported  $\overline{LR}$  values (see p. 79 for description of these reference sets). The tables also indicate the total number of PM oligos within each SNP probe set with LR < 0 ( $n_1$ ; CNV-affirmative oligos) and those with LR < -0.5 ( $n_2$ ).

Table (a) illustrates the impact of small reference sets (with 1-2 reference samples) on the estimated SNP LR values ( $\overline{LR}$ ) and the number of PM oligos in the SNP probe sets that indicate loss of signal intensity ( $n_1$ ,  $n_2$ ). Table (b) shows the mean LR,  $n_1$  and  $n_2$  estimates of the same 5 deleted SNPs based on larger reference sets, consisting of 24, 99 and 150 samples. The red highlighted data denote the estimated log2-ratio intensities of SNP<sub>81</sub> that was previously displayed in Figure 3.5f (as a non-informative SNP in *SUFU* deleted region). The data in the above table reveals that the estimated LR value of this SNP (SNP<sub>81</sub>) does not improve by any reference set. This is an example of a SNP that does not support copy number deletion in this experiment, regardless of the reference set used to estimate LR intensities.

	N	Naive Bayes Normal Func.		Naive Bayes Kernel Fit		QDA (default)		Mean Signal of the Deleted Region (LR)			
Del. id		# Deletion Calls	# False Neg- atives (β)	# Deletion Calls	β	# Deletion Calls	β	$\overline{LR}^*$	$\mathrm{LR}^{\dagger}_{\mathit{NBN}}$	$\mathrm{LR}^{\dagger}_{NBK}$	LR <sub>OPAS</sub> :
1	30	22	8	20	10	24	6	-0.48	-0.61	-0.58	-1.05
2	24	18	6	16	8	20	4	-0.44	-0.56	-0.50	-1.02
3	4	4	0	3	1	4	0	-0.68	-0.55	-0.55	-1.33
4	152	72	80	71	81	101	51	-0.44	-0.36	-0.37	-0.95
5	13	9	4	10	3	13	0	-0.62	-0.84	-0.84	-1.39
6	28	23	5	23	5	22	6	-0.49	-0.50	-0.69	-1.15
7	152	96	56	98	54	110	42	-0.41	-0.55	-0.56	-1.09
8	5	5	0	5	0	5	0	-0.60	-0.79	-0.79	-1.36
9	5	5	0	5	0	4	1	-0.53	-0.75	-0.75	-1.20
10	5	4	1	4	1	3	2	-0.44	-0.56	-0.56	-1.07
11	5	4	1	4	1	3	2	-0.37	-0.52	-0.52	-0.86
12	14	12	2	11	3	12	2	-0.46	-0.79	-0.79	-1.18
13	5	4	1	2	3	5	0	-0.37	-0.50	-0.35	-1.03
14	5	4	1	3	2	3	2	-0.25	-0.73	-0.73	-1.09
15	9	8	1	8	1	9	0	-0.65	-1.03	-1.09	-1.50
16	9	9	0	8	1	8	1	-0.51	-0.90	-0.90	-1.19
17	11	2	9	2	9	1	10	+0.32	+0.26	+0.23	-0.19
18	80	52	28	53	27	61	19	-0.42	-0.62	-0.61	-1.11
19	11	6	5	4	7	8	3	-0.26	-0.29	-0.21	-0.95
Sum:	567	359	208	350	217	416	151				
					1	Average LR Va	alues:	-0.43	-0.589	-0.587	-1.09

**Comparing Naive Bayes and QDA Classification Results** 

\* mean LR value of all SNPs in the deleted region

† mean LR values of all SNPs in the deleted region based on Naive Bayes classification with normal distribution (LR<sub>NBN</sub>) and kernel fit (LR<sub>NBK</sub>)

‡ OPAS-estimated mean LR value of all SNPs in the deleted region (based on QDA classification)

Table 3.3: Comparison of the performance of Naive Bayes and OPAS QDA-based SNP pre-processing methods. This table presents the summary of analysing SNPs in 19 regions of copy number deletion in follicular lymphoma samples (500K data) that were validated by Illumina sequencing (Chapter 4). The rows correspond to the results of each analysed deleted region. The number of Nsp SNPs within sequencevalidated boundaries of each deletion (N) is reported in the 2nd column of the table. The rest of the columns present the number of SNPs that were successfully predicted as deleted by 3 different methods (Naive Bayes classifiers with normal distribution and kernel fit; and OPAS QDA-based classification). The number of the remaining SNPs in the deleted region (those that were not classified as deleted) is also reported for each method ( $\beta$ ; number of false negatives). The last 3 columns of the table summarize the estimated LR values of the deleted regions based on each aforementioned method (LR<sub>NBN</sub>, LR<sub>NBK</sub> and LR<sub>OPAS</sub>). The original mean LR values of these regions are also reported in this table ( $\overline{LR}$ ; column 9). The summary of the comparison across 19 deleted regions, presented in the last 2 rows of this table, shows that in overall Naive Bayes classifier with normal distribution has better sensitivity to detect deleted SNPs compared to kernel fit; however, OPAS QDA-based classification appears to outperform both of these Naive Bayes classifiers (in addition to the above comparison of true positives, the analysis of false positive deletion calls of these 3 methods is presented in page 83).

		OPAS Resul	ts	CBS Results	
# SNPs (1)	Mean LR	# mapped segments with LR<0	LR <sub>OPAS</sub>	# mapped segments with LR<0	LR <sub>CBS</sub>
15	-0.312	100	-0.992	100	-0.362
14	-0.311	100	-0.987	100	-0.354
13	-0.301	100	-0.976	100	-0.357
12	-0.285	100	-0.944	100	-0.351
11	-0.262	100	-0.941	99	-0.339
10	-0.247	100	-0.945	99	-0.333
9	-0.253	99	-0.911	100	-0.331
8	-0.245	100	-0.894	90	-0.322
7	-0.183	96	-0.899	91	-0.289
6	-0.173	98	-0.893	86	-0.262
5	-0.150	94	-0.864	78	-0.235
4	-0.127	94	-0.837	78	-0.228
3	-0.047	92	-0.820	58	-0.119
2	-0.083	89	-0.809	63	-0.145

Table 3.4: Comparing CBS and OPAS performance in detecting known CNVs with respect to the number of SNP probes in the CNV regions. This table presents the results of CBS and OPAS in detecting a known deletion on chromosome 2p16.3 in a patient with mental retardation using Xba SNP array data (this deletion includes 17 Xba SNP probe markers). The first column indicates the number of SNP probes in the deleted region after *N* SNPs were randomly excluded from this region ( $2 \le N \le 15$ ). The 3rd and 5th columns indicate the average number of times that 2p16.3 deletion was detected by CBS and OPAS methods (100 simulated signals were generated for each listed value of *N*). The LR value of the known deleted region was estimated based on both CBS and OPAS methods and shown in the 4th (LR<sub>OPAS</sub>) and 6th (LR<sub>CBS</sub>) columns of this Table. This table also presents the original estimated (mean) LR value of the deleted region across 100 simulated signals (Mean LR; column 2). The visual representations of the data in the above Table are provided in Figure 3.17.

# **Chapter 4**

# Analysing CNVs in Follicular Lymphoma Genomes

# 4.1 Introduction

Lymphoma comprises more than 67 subtypes of two related cancers that affect the lymphatic system, Hodgkin lymphoma and non-Hodgkin lymphoma<sup>1</sup> [367]. In Canada, non-Hodgkin lymphomas accounted for about 7,500 new cases of cancer in 2010, making them the fifth most commonly diagnosed cancers; and 3,200 estimated deaths, which makes them the sixth most common cancer related mortality<sup>2</sup>. Follicular lymphoma (FL) is the second most common lymphoma and comprises about 20-30% of all non-Hodgkin lymphomas [368]. Cytogenetic abnormalities are a common characteristics of most FL cases [248], including frequent gains of 1q, 2p, 6p, 7, 9p, 12, 17q, 18, X and losses of 6q and 10q [244–248]. Additionally, more than 85% of FL cases are associated with a specific translocation, t(14;18)(q32;q21); however, this translocation is not sufficient to produce clinical FL [254, 369–371]. Therefore, other genetic aberrations may play a role in lymphoma tumorigenesis.

The advent of high-resolution microarray techniques provided the capability to detect submicroscopic DNA copy number gains and losses in FL and led to novel CNV discoveries and improvements in the characterization of known CNVs in FL [244, 247]. An example of novel relatively smaller CNVs in FL is 1p36 deletion which was found to be present in 25.5% of 108

<sup>&</sup>lt;sup>1</sup>http://www.lymphoma.org

<sup>&</sup>lt;sup>2</sup>http://www.cancer.ca/Canada-wide/About%20cancer/Cancer%20statistics/Canadian%20Cancer%20Statistics.aspx?sc\_lang=en

FL patients in a study by Cheung et al. [244] using array CGH. Nonetheless, most of the known altered regions in FL span several megabases and contain many genes, making it very difficult to identify specific genes that may play significant role in FL.

The goal of the study presented in this Chapter was to perform copy number analysis of the SNP array data from a cohort of 25 matched tumor/normal FL samples using the OPAS approach. These samples had been previously studied at the Genome Sciences Centre (GSC) by a multi-platform approach, including BAC array CGH<sup>1</sup> (BAC aCGH), finger print profiling (FPP) and targeted Il-lumina sequencing. In addition to OPAS, the 500K SNP array data from these samples have also been analysed by an alternative CNV calling algorithm known as Significant Mean Distance (SMD) method [29, 30, 259, 372] (performed by an independent group at the GSC). The specific aims of this Chapter were (1) to identify candidate somatic CNVs (i.e., CNVs that are present in the tumor but not the matching normal DNA) in these FL patients, (2) to profile the candidate CNVs and investigate the frequency, size and proportion of DNA gains and losses; and (3) to investigate whether the use of the OPAS method resulted in detecting novel CNVs, particularly smaller events that were not previously detected in these genomes. The latter analysis would indicate the usefulness of using OPAS in the context of detecting cancer-related CNVs.

In this Chapter, I focused on using the data from Affymetrix 250K Nsp array (consisting of 3,035,520 PM oligos) that is part of the Affymetrix 500K dual SNP array set<sup>2</sup>. The reason for using only Nsp arrays was that it has been shown Sty arrays have lower genotype call rates and higher genotyping errors compared to Nsp arrays [373]. Such biases could mitigate the genotyping call rates as well as the accuracy of CNV analysis since the same probe intensity data are used for both of these analyses. Even with half of the initial data points, an inter-platform comparison, explained in Section 4.3.7, revealed that the presented analysis (Nsp) found several novel events smaller than 150 kb that were not previously identified by analysing 500K array (Nsp and Sty) data with an alternative method.

# 4.2 Materials and Methods

#### 4.2.1 Samples and Cytogenetic Analysis

The 25 FL tumor and normal specimens were collected at the British Columbia Cancer Agency (BCCA) in Vancouver, British Columbia. In each case, a lymph node biopsy with FL morphol-

<sup>&</sup>lt;sup>1</sup>BAC array CGH experiments were performed at the BC Cancer Agency.

<sup>&</sup>lt;sup>2</sup>Affymetrix GeneChip<sup>®</sup> 500K SNP array consists of NspI and StyI arrays, each with ~250K SNPs.

ogy was paired with a peripheral blood sample as the normal DNA (used as the reference set in CNV analysis). Cytogenetic analysis of lymph node specimens was performed at the Center for Lymphoid Cancer (CLC) at BCCA, according to the method previously described by Horsman et al. [374].

### 4.2.2 BAC Arrays and SNP Arrays

The array CGH (aCGH) analysis was performed at the CLC using submega base resolution tiling (SMRT) arrays containing 26,819 BAC clones [121, 375]. Copy number analysis of aCGH data was also performed at the CLC using the Hidden Markov Model (HMM) program CNA-HMMer  $v0.1^1$  [244, 376] and visual inspection.

The Affymetrix GeneChip<sup>®</sup> 500K SNP array experiments were performed at the GSC using 500 ng samples of tumor and constitutional DNA according to manufacturer's protocol [166]. In addition to analysing the 250K Nsp array data using OPAS, the data from Nsp and Sty arrays were independently analysed at the GSC using Significant Mean Distance (SMD) method, which has been used in several other publications [29, 30, 259, 372]. In SMD analysis, putative CNVs including at least 10 contiguous SNPs called by the SMD software with a p-value below  $1 \times 10^{-8}$  were selected and passed through a manual inspection phase (at the GSC). At the end of this process, the analytical and manual inspection of SMD results generated a list of 211 putative somatic CNVs in 25 FL patients (provided by Dr. Allen Delaney, GSC).

# **4.2.3** Fingerprint Profiling (FPP)

The finger printing profiling (FPP) and validating candidate rearrangements were performed at the GSC. Briefly, BACs from each library were subjected to restriction digest fingerprinting, as previously described by Krzywinski et al. [126]. The FPP method maps each fingerprinted BAC to the reference genome (hg18), allowing the identification of the differences in the restriction fragment pattern between patient and the reference genome, such as shown in Figure 4.1. These differences were then converted by a computational algorithm into a list of 271 candidate rearrangements in 23 FL patients, consisting of 132 deletions, 14 insertions, 13 duplications, and 112 other events (35 translocations, 47 inversions, and 30 complex rearrangements). To validate the FPP candidate rearrangements, the BAC clones that captured these events were subjected to paired-end sequencing. In order to determine the exact rearrangement breakpoints, a subset of candidate rearrangements in 20 FL patients (354), supported by at least two BACs, were chosen for complete Illumina sequenc-

<sup>&</sup>lt;sup>1</sup>available at http://www.cs.ubc.ca/?sshah/acgh/

ing (Figure 4.1). A PCR assay was also performed to identify the origin of these rearrangements (somatic or germline). The final list of sequence and PCR validated rearrangements, including 193 deletions and 43 insertions/duplications, was compared to candidate CNVs in Section 4.3.7 as a source of validation. These 236 events (193 + 43) were a subset of all validated rearrangements that were proven to be somatic (by PCR) or had "undetermined source"<sup>1</sup>.

The sequence validation and PCR experiments were performed by Dr. Andy Mungall at the GSC. The detected rearrangements in the FL dataset from FPP and Illumina sequencing is accessible through an internal database at the GSC, called Follicular Lymphoma Tumour BAC Fingerprint Database at http://map02.bcgsc.ca:5000/anomaly/, thereafter, referred to as "Tumordb".

#### 4.2.4 Ingenuity Pathway Analysis Software

Ingenuity pathways analysis software (Ingenuity<sup>®</sup> Systems, www.ingenuity.com) was used to examine genes identified by OPAS analysis for their relevance to currently known biological functions and canonical pathways. The p-values are calculated with the right-tailed Fisher's Exact Test  $(P \le 0.05 \text{ indicates a statistically significant, nonrandom association})$ . The Benjamini-Hochberg multiple testing correction was also used where appropriate.

# Contributions

The Affymetrix experiments for the 25 FL samples were performed at the Genome Sciences Centre (GSC). The SMD copy number analysis was performed by an independent group at the GSC. Putative CNVs including at least 10 contiguous SNPs called by the SMD software with a p-value below  $1 \times 10^{-8}$  were selected and passed through a manual inspection phase by the same group. At the end of this process, the analytical and manual inspection of SMD results generated a list of 210 putative somatic CNVs in 25 FL patients (provided by Dr. Allen Delaney). Throughout this Chapter, all references to the SMD detected CNVs indicate regions that have been reported in the aforementioned list of 210 annotated CNVs which are available at the Tumordb website. Additionally, Tumordb also contains the CNV results from array CGH analysis of these samples, which was performed at the CLC lab at BCRC (Section 4.2.2). The aCGH results in Tumordb were also generated by coupling analytical approach (CNA-HMMer) and visual inspection.

The finger print profiling (FPP) was performed by Mapping group at the GSC. The sequencing and PCR validation of the source of the events were performed by Dr. Andy Mungall at the GSC.

<sup>&</sup>lt;sup>1</sup>If it was not possible to confirm whether the breakpoints were somatic or germline through PCR (for instance, when it was not possible to design unique PCR primers), the event was called a rearrangement with "undetermined source".

The FPP and sequence data are also available at the Tumordb website.

To identify somatic copy number aberrations, I analysed the SNP 250K Nsp raw signal intensity data (.CEL files) using the OPAS algorithm (described in Chapter 3). The candidate somatic CNVs were then selected by choosing a subset of all OPAS predicted regions that had log2-ratio intensity (LR) less then -0.2 or greater than +0.2, or CNVs that had a lower magnitude of LR but indicated a significant deviation with respect to their surrounding regions (or other chromosomes) based on their estimated z-scores. The OPAS default parameters were used to filter noisy oligos, normalize the data by two-level normalization process and split each chromosome into contiguous regions with different relative copy number estimates. The default parameters included 10,000 permutations to measure segmentation p-values, and a statistical significance of 0.01 to accept change points in segmentation phase. Candidate somatic CNVs were then selected from the list of all OPAS generated regions that indicated a significant copy number change based on their estimated LR or p-values, as described in Section 4.3.1.

# 4.3 Results

# 4.3.1 Magnitude of Copy Number Changes in FL Genomes

In a sample with predominantly diploid chromosome numbers, the expectation would be that a copy number of 2 corresponds to an LR of 0 ( $\log_2(2/2) = 0$ ). Thus, LR of -1 ( $\log_2(1/2)$ ) and +0.58 ( $\log_2(3/2)$ ) represent single copy loss and gain, respectively. However in practice, these values are compressed by the level of standard deviation of microarray hybridization intensities (noise) that can vary significantly between different experiments. In addition to microarray noise, CNV heterogeneity, described as copy number amplification or deletion that is present only in a subset of cells, further lowers the magnitude of log2-ratio changes of real CNV events. CNV heterogeneity and aneuploidy are common observations in cancers, further complicating CNV analysis in cancerrelated studies. Therefore, particularly in cancer-related studies, lower LR magnitudes are used for CNV calling. For example, Berger et al. [377] used LR = ±0.15 to detect CNVs in lung adenocarcinoma, and Haverty et al. [366] used LR = ±0.3 to report putative CNVs in breast and ovarian cancers.

To define significant log2-ratio intensity changes, I analysed the distribution of LR values from all OPAS generated regions with more than 2 snps in 25 FL patients (1931 regions), as depicted in Figure 4.2. Inspection of Figure 4.2a reveals two change points in the CDF curve that represent approximately 7% and 93% of all the OPAS estimated regions (among 25 FL samples). The region
between these two LR values, highlighted by the green box, seems to follow a normal distribution, as shown in the histogram of Figure 4.2b. Thus, this analysis indicates that about 7% of all OPAS regions appear to have a distribution that is not consistent with the rest of the data. Therefore, the cutoffs for significant LR changes representing CNV gains and losses were chosen as the 93% and 7% of the data, corresponding to regions with LR  $\leq -0.2$  or  $\geq 0.2$  (Table 4.1 provides a detailed summary of the frequency of predicted deletions and amplifications with varying range of LR cut-off values that were used in this study (LR = ±0.2) resulted in total of 251 candidate CNVs (134+118).

One limitation of using only LR cut-offs to determine whether an OPAS-estimated region is a candidate CNV is that depending on the overall spread of the data in a particular chromosome or sample, sometimes estimated regions with apparent gains or losses of signal intensity do not pass the LR significance thresholds. For instance, in Figure 4.3, the entire chromosome 1 of patient 9 (ht-9) has log2-ratio of zero, except for a region approximately 7.6 Mb on 1p36 that shows a clear loss of signal intensity, although its corresponding LR is only -0.13. To circumvent the LR cutoff problem, z-scores<sup>1</sup> were used to identify regions with slight gain or loss of signal intensity that may reflect a significant deviation with respect to the chromosome they belong to (or other chromosomes of the corresponding sample). Figures 4.4 and 4.5 compare the distribution of zscores of all regions in the FL dataset with significant LR deviation ( $|LR| \ge 0.2$ ), and regions with slight LR change (|LR| < 0.2). As seen in Figures 4.4a and 4.5a, the distributions of z-scores of regions with significant loss of signal intensity (referred to as "X") overlaps with regions with slight loss of signal intensity (referred to as "L"). In a linearly non-separable case with overlapping distributions, such as in Figure 4.4a, misclassification is inevitable. In practise, regions with z-score less than or equal to -0.9 often represent real deletions (see the estimated z-scores in Table 4.3). However, the data from validated deletions reveals that larger deletions may have lower magnitudes of z-scores (e.g., Fig. 4.3). The selected z-score of -0.6, shown by black arrow in Figure 4.4a, is an arbitrary value between -0.9 and 0 to select regions with slight loss of intensity that may represent real copy number deletions (z-score  $\leq -0.6$ ). The visual inspection of such cases (regions with -0.2 < LR < 0 and z-score  $\leq -0.6$ ) can add more confidence to the selected candidate deletions.

Figures 4.4b and 4.5b show similar analysis to compare the distribution of z-scores for regions with significant increase of signal intensity (LR  $\ge$  +0.2; referred to as "Y") with regions with slight gain of signal intensities (0 < LR < 0.2; referred to as "G"). As observed in these plots z-scores of regions with intensity gain ("Y" and "G") have a smaller overlap compared to z-scores of regions

<sup>&</sup>lt;sup>1</sup>The z-score is generated using  $z = (x - \mu)/\sigma$ , where *x* is the OPAS estimated LR of the region of interest. Also,  $\mu$  and  $\sigma$  are the mean and standard deviation of LR values of all SNPs in the corresponding chromosome(s).

with loss of intensity ("X" and "L"). In fact, selecting z-score = +0.6 almost completely separates the two distributions. The latter suggests that OPAS regions with slight gains of signal intensity but with z-scores  $\ge +0.6$  can also be considered as candidate regions of copy number amplification.

Based on the above z-score selection criteria, 34 OPAS detected regions with slight loss (19) or gain (15) of LR values were also added to list of candidate somatic CNVs in the FL dataset (the original list included 251 OPAS regions with  $|LR| \ge 0.2$ ).

#### 4.3.2 Spectrum of Somatic CNVs in FL Genomes

Genome instability in cancer results in a wide range of copy number rearrangements in these genomes, ranging from small focal CNVs that target specific genes (such as tumor suppressors or oncogenes) to aneuploidies that alter the number of chromosomes in the cells (e.g., trisomy 8 in acute myeloid leukemia<sup>1</sup>).

In addition to the variation in the size of CNVs in cancer, studies have also shown certain patterns of CNVs with respect to their location on the chromosome. For instance in colon cancer, it has been shown that there is a significant abundance of CNVs near centromeres [360] and in glioblastoma and melanoma there is an increased frequency of CNVs near chromosome ends [360]. In this Chapter any detected copy number variation in FL samples, regardless of its size, is referred to as a candidate CNV.

In total, 286 candidate somatic CNVs were found in 25 FL patients, with an average of 11.4 variants per individual and a range of 2-26 variants per patient. The predicted 286 CNV regions had a median size of approximately 895 kb. Analysis of the size of candidate CNVs in FL patients suggested that 20/25 (84%) patients had at least one CNV  $\leq$  150 kb, and 22/25 (88%) had at least one CNV  $\leq$  2 Mb. The 286 candidate somatic events included 153 (53%) deletions with a global median of ~672 kb, spanning between 753 bp-191 Mb in length (with median of 48 SNP probes). The predicted amplifications accounted for 133 (46%) of all candidate CNVs with global median of ~5.4 Mb, ranging between 8 kb and 242 Mb in size (with median of 269 SNP probes), as described in Table 4.2. The pie chart in Figure 4.7.(a) indicates that the frequency of candidate somatic amplifications). However, as shown in panel (b), at smaller sizes (~150 kb or less) candidate deletions are ~1.8 times more frequent than amplifications is two times higher than that of deletions. In conclusion, this figure suggests that although the overall frequency of candidate deletions and

<sup>&</sup>lt;sup>1</sup>http://atlasgeneticsoncology.org//Anomalies/tri8ID1017.html

amplifications is similar in the FL dataset (25 patients), the proportion of large-scale candidate copy number gains is likely greater than that of large-scale deletions, with the opposite holding true for smaller candidate CNVs. The relative enrichment of deletions at smaller sizes (compare (b) and (c) in Figure 4.7) may reflect higher rates of small acquired copy number loss events in lymphoma patients. Such small deletions may disrupt the function of specific gene(s) that may be important in FL.

In terms of the location of the detected CNVs with respect to the rest of the chromosome, there were 3 main patterns of events across 25 FL patients in this study. These categories can be summarized as the following: (1) CNVs that affect whole-chromosomes<sup>1</sup> or chromosome arms, (2) CNVs that affect the distal ends of chromosomes, or (3) CNVs that affect other regions of the genome. These three categories of candidate events are analysed separately in this study, as described in Sections 4.3.3-4.3.5. In Section 4.3.6 it is also shown that several small (< 150 kb) OPAS deletions that are validated (by Illumina sequencing) affect known cancer-related genes, such as the *DKN2A* tumor suppressor. Another interesting example, discussed in Section 4.3.8.3 is a deletion that removes 3 exons of *KIT*, a known proto-oncogene. Detailed analysis of this gene shows that the deleted region corresponds to the extracellular region of *KIT* which acts as ligand binding site for this gene. Therefore, the small OPAS detected deletion (which is also validated by Illumina sequencing) may contribute to constitutive activation of this proto-oncogene by removing the ligand binding site of *KIT* (see Section 4.3.8.3).

# 4.3.3 Category 1: CNVs Affecting Whole-chromosomes or Chromosome Arms (WCA)

In this study, large-scale events that span the entire length of chromosomes or chromosome arms are referred to as WCA events (e.g., Figure 4.8). In total, 48 WCA events were found in 17<sup>2</sup> FL patients (Table 4.2). The summary of the chromosomes that are affected by WCA events in the FL dataset and the patients that harbor those events are presented in Figures 4.9 and 4.10. These data indicate that WCA events are frequently observed in FL genomes and that these events are substantially enriched for copy number amplifications compared to deletions (copy number amplifications accounted for 42 of the total 48 WCA events). It was also observed that gains of entire chromosomes were the most prevalent observation in this category (22/48 events were gains of whole chromosomes; Figure 4.10). Previous studies of rearrangements in FL genomes using

<sup>&</sup>lt;sup>1</sup>As mentioned previously, the gain or loss of entire chromosome(s) is a condition known as an euploidy. However, in this analysis any predicted change in the DNA copy number, regardless of its size, is referred to as a candidate CNV.  $\frac{29}{25}$  EL patients did not have any WCA super (action to 2 - 10, 12, 17, 10, 20, and 20)

<sup>&</sup>lt;sup>2</sup>8/25 FL patients did not harbor any WCA events (patients 3, 6, 10, 13, 17, 19, 20, and 26).

different methods have consistently found gains of whole chromosomes or chromosome arms to be the most frequent chromosomal copy number abnormality in FL, which is consistent with the high frequency of WCA events that was detected in the current FL dataset (48/286 of all candidate FL CNVs).

From the 48 large-scale WCA events, 46 were directly validated by the results from cytogenetics analysis (karyotyping and/or MFISH). The remaining two candidate WCA events were slight gains of chromosomes X (LR  $\simeq 0.08$ ) and 8 (LR  $\simeq 0.07$ ) in the same patient (ht-29), shown in Figure 4.11. In addition to OPAS, aCGH analysis also indicated a slight shift in log2-ratio intensity readouts of the above two chromosomes in ht-29 (Figures 4.11e-4.11f). SMD analysis also reported one of these two candidate WCA events (+X; ht-29, Figure 4.11d). One possible explanation for these observations is that whole chromosome gains were only present in a subpopulation of cells in this patient, which was not detected by conventional cytogenetic analysis.

#### 4.3.3.1 Recurrent WCA CNVs

The 48 WCA CNVs included 22 distinct events that were observed between 1-6 times in the FL dataset (Fig. 4.9). The most frequently recurrent events (those found in 3 patients or more) among 25 FL samples were +1q (6/25; 24%), +6p (3/25; 12%), -6q (5/25; 20%), +7 (4/25; 16%), and +X (5/25; 20%). All these events are known to be frequent chromosomal rearrangements in FL that have been consistently found by other groups (gains of 1q, 2p, 6p, 7, 9p, 12, 17q, 18, X and losses of 6q and 10q have been reported as frequent CNVs in FL by several independent studies [244–248]).

Next, I examined whether there was an association between the detected WCA events among the 17 patients by using the Fisher exact test. This hypothesis was rejected with a p-value of 0.16, concluding that there was no significant evidence for co-occurrence of the WCA events among 25 FL samples in thus study. Nonetheless, visual inspection of WCA events presented in Figure 4.9 suggests that there may be a recognizable pattern for such large-scale CNVs among patients with the highest number of WCA events. As seen in this figure, +1q, +X, and -6q seem to be present almost in all patients with  $\geq 5$  WCA events.

#### 4.3.4 Category 2: CNVs Affecting the Distal Ends of Chromosomes

Inspection of the 575 OPAS chromosome plots revealed another frequent pattern of somatic CNVs in the FL dataset: copy number changes that localized near the ends of chromosomes, such as shown in Figure 4.12. Subtelomeric regions, regions proximal to chromosome ends, are gene-rich and therefore even small CNVs in these chromosomal loci can affect the function of genes that may

be important in FL. For instance, deletions of 1p36 subtelomeric region that have been reported in several cancers, such as neuroblastoma [378], disrupts the function of multiple known and putative tumor suppressor genes including *TP73*. This gene is significantly under-expressed in lymphomas and leukemias [379]. These results suggest that frequent deletions of chromosome 1p36 in FL genomes that encompasses *TP73* may also play an important role in FL.

After observing a pattern of CNVs, particularly losses of signal intensities, that appeared to encompass chromosome ends, I decided to analyse these events in a separate category. This category of events, thereafter referred to as 'distal CNVs', consists of candidate OPAS-detected CNVs that include either the first SNP probe on the p-arm or the last SNP probe on the q-arm of a chromosome (excluding CNVs in WCA category). Therefore, all distal CNVs encompass subtelomeric regions and may also include telomeres. Based on this definition, 29 candidate distal CNVs were found in 15/23 FL patients, ranging in size between  $\sim 21$  kb (Supplementary Figure L.1) to 59.6 Mb (see Table 4.2). Table 4.3 presents a list of distal events in each FL chromosome and whether each event was also detected by aCGH and/or SMD (these data are available in Tumordb). It is evident from this table that in contrast to WCA events, FL distal events were enriched in copy number losses with deletions accounting for 20 of the total 29 distal events ( $\sim 69\%$ ). The relative enrichment of deletions that affected chromosome ends may be suggestive of losses of sequences that affect the functional status of the telomeres. It is important to note that cytogenetic analysis has particularly poor resolution for detecting CNVs proximal to chromosome ends ( $\sim$ 5-10 Mb). Therefore, most of the distal CNVs reported in Table 4.3 were not detected by cytogenetics analysis, even though these events were mainly larger than 2 Mb (Figure 4.12). As seen in Table 4.3, from the 29 OPAS candidate distal CNVs, 24 (83%) were also detected by aCGH and SMD results in Tumordb (http://map02.bcgsc.ca:5000/), and 26/29 (~90%) were detected by aCGH or SMD (examples shown in Figure 4.13 and Supplementary Figures L.2-L.3).

There are three remaining OPAS candidate distal CNVs that are not seen by aCGH or SMD results. These 3 cases are shown in Figure 4.14 (~8.6 Mb) and Supplementary Figures L.1 (~21.3 kb) and L.4 (~78 kb). One of these putative OPAS-specific events occurs on 1p36 region (ht-7) that is frequently deleted in FL samples (Figure 4.14). For the other two detected events that are exclusively seen by OPAS (Supplementary Figures L.1 and L.4), the FPP data in Tumordb report coverage gaps (i.e., regions of the genome with no FPP coverage), but no further data was available to directly validate these putative distal CNVs. Table 4.3 also shows that, in contrast to WCA events, distal CNVs do not frequently affect the same regions, with the exception of 1p36.33 that was deleted in 32% (8/25; Figure 4.15) of patients in the FL dataset (examples shown in Figure 4.3, and Supplementary Figures L.2 and L.5). Cheung et al. [244] had previously reported

deletions of 1p36 in 25.5% of 106 FL patients in an independent study using array CGH technology (Figure 4.15).

#### 4.3.4.1 Functional Analysis of the Genes Affected by Distal CNVs

Analysis of the gene content of candidate distal CNVs found 1,691 unique genes that were potentially affected by distal deletions and 1,653 unique genes that were potentially affected by distal amplifications (Table 4.2). Further analysis of the function of genes in candidate distal CNV regions was performed using Ingenuity<sup>®</sup> Pathway Analysis software. Two separate analyses were performed to investigate candidate distal deletions and amplifications. The analysis of candidate distal deletions showed that four biological function categories- DNA replication, recombination and repair, cell-to-cell signalling, and cellular assembly and organization- were significantly associated with the corresponding genes, after adjusting for false discovery rate (Benjamini-Hochberg multiple testing correction,  $P \le 0.02$ ). As seen in Figure 4.16a, the top gene network associated with candidate distal deletions includes several important cancer-related genes and protein families. For instance, Caspase, highlighted by the red arrow in Figure 4.16a, refers to a family of proteins that are major regulators of apoptosis [380].

Similar analysis was performed using 1,653 candidate distal amplified genes. This analysis identified that cancer, immunological disease and cell cycle were the top affected networks associated with candidate distal amplifications (shown in Figure 4.16b). This analysis also revealed that lymphoid tissue structure and development, and hematological system function and development were the top physiological functions that were significantly associated with the above candidate amplifications. The top gene network associated with candidate amplified distal CNVs, illustrated in Figure 4.16b, also includes several known cancer-related genes, such as MYC. This gene, denoted by the blue arrow in Figure 4.16b, is a proto-oncogene that is often up-regulated in many cancers. It has been hypothesized that MYC over-expression stimulates gene amplification, possibly through DNA over-replication [381]. The alterations of MYC have also been linked to the variety of hematopoietic tumors, leukemias and lymphomas, such as Burkitt's Lymphoma. Importantly, it is hypothesized that MYC regulates the expression of 15% of all genes [382], such as those associated with cell proliferation (e.g., p21) and growth (e.g., DHFR) [383]. Therefore, amplification of MYC may result in constitutive up-regulation of this gene and may lead to cancer formation.

## 4.3.5 Category 3: Other CNVs

There were in total 209 candidate CNVs that were not part of WCA or distal events, consisting of 127 (61%) deletions with a global mean size of 2.13 Mb and 82 (39%) amplifications with a global mean size of 2.15 Mb (see Table 4.2). As seen in Figure 4.17, the frequency of this category of CNVs (shown by pink bars) varies remarkably between FL chromosomes. The observed high frequency of predicted CNVs on chromosomes 14 and 7 is mainly due to the common copy number changes in the *IgH* and T-cell receptor genes that are located on these chromosomes. These known CNV sites include 14q32 (*IgH*), 14q11.2 (T-cell receptor  $\alpha$  and  $\delta$ ), 7p14 (T-cell receptor  $\gamma$ ) and 7q34 (T-cell receptor  $\beta$ ). Since rearrangements in T-cell receptors occur in normal T-cells and are not lymphoma-specific, these events are excluded from further analysis. Therefore, the most frequently affected chromosomes in this category of events, other than functional immunoglobulin and T-cell receptor rearrangements, are chromosomes 3, 4, 6, 7, 9,10,12 and 17 (Figure 4.17).

The regions presented in the previous two categories (WCA and distal CNVs) were generally large (all 48 WCA CNVs were larger than 37 Mb and 25 of the total 27 distal CNVs were larger than 500 kb) and contained many genes. Therefore, it was not possible to identify specific gene(s) that may have been disrupted as the result of targeted CNVs related to FL. Studying the subset of putative somatic CNVs that are small can help to identify specific genes that are associated with FL initiation and progression. Therefore, in the following section I further analyse all candidate somatic events in the FL data set that are smaller than 150 kb (referred to as candidate focal CNVs).

## 4.3.6 Candidate Focal CNVs in FL Genomes and Functional Analysis of the Affected Genes

From the total 286 candidate CNVs, 53 (18.5%) events were  $\leq 150$  kb and are referred to as candidate "focal" CNVs. The focal CNVs were enriched in copy number losses with deletions accounting for 34/53 (64%) and amplifications accounting for 19/53 (36%) events. The difference between the relative proportion of predicted focal deletions and amplifications was further increased when the apparent candidate amplifications that overlapped with T-cell receptor genes ( $\gamma$  and  $\alpha$ ) were excluded from the list of focal events prior to gene analysis (11 of these 53 predicted small events overlapped with T-cell receptors). After this modification, the candidate focal deletions and amplifications accounted for 81% (34/42) and 19% (8/42) of all (42) candidate focal CNVs, respectively. From these 42 putative focal amplifications and deletions, less than half (20/42) overlapped with at least one gene. It is important to note that some of the other 20 focal candidate events that do not harbor any gene may also be involved in FL. For instance, a copy number alteration may affect a region downstream or upstream of a cancer-related gene that is involved in its regulation. Figure 4.18 presents an example of a sequence validated  $\sim$ 40 kb deletion (somatic) on chromosome 13 (ht-21) that does not overlap with any gene. Despite the importance of all focal events, in the rest of this section I focused on functional analysis of a subset of these CNVs that affect at least one gene (the 20 regions presented in Table 4.4). This table reveals that in total 30 unique genes were affected by these 20 putative focal CNVs. It also indicates that the most of these candidate focal CNVs (14/20) affect only a single gene.

From the 30 unique genes in this table, 28 (93%) overlapped with candidate focal deletions and only 2 (7%) genes, *PAX5* and *RNF217*, overlapped with candidate focal amplifications. The only gene in Table 4.4 that was observed in more than one FL patient was *ERBB4*. This gene is affected by two small candidate deletions in patient 4 ( $\sim$ 24 kb; ht-4) and patient 29 ( $\sim$ 4 kb; ht-29), that partially affect the same intron of this gene (shown in Figure 4.19). The ERBB4 gene ( $\sim 1.6$ Mb; 28 exons) is a member of growth factor receptors that regulate signals of cell differentiation, proliferation, migration and survival. Disruptions in the activation of these receptors have been associated with tumor development and malignancy in many cancers, such as breast cancer [384, 385]. Furthermore, a study by Soung et al. [386] had found somatic intronic mutations in *ERBB4* in several human cancers and suggested that such mutations affect *ERBB4* ligand binding [386]. Therefore, although the above candidate focal deletions (in ht-4 and ht-29) do not directly remove any coding exon of *ERBB4*, they may affect the function of this gene through altering its ligand binding domain. Figure 4.19a shows that in ht-4 there is an FPP detected fragment hole that aligns to the OPAS predicted  $\sim 24$  kb deletion on *ERBB4* (a fragment hole is referred to as a fragment in the fingerprint data that does not match to the alignment region in the reference genome; see Appendix K for more details). This Figure also indicates that in ht-29 there is a fragment hole next to the OPAS ~4 kb predicted region of copy number loss on ERBB4. No further data were available for direct validation of these putative CNVs, however, the FPP detected events suggest that the above *ERBB4* intron may harbor real focal deletions in 2 of the FL patients. As noted in Table 4.4, 13/20 (65%) gene-affecting focal CNVs overlap with FPP or SMD detected events (e.g., Figures 4.19-4.20 and Supplementary Figure L.6), or are directly validated by Illumina sequencing (e.g., Figure 4.32a). Nonetheless, except the two putative *ERBB4* intronic deletions and known T-cell receptor rearrangements, none of the other reported focal CNVs were recurrent among FL patients. One speculation of this finding is that the focal amplifications and deletions may target not necessarily the same genes, but a variety of genes that participate in the same critical pathway. Based on this assumption, I studied the function of the aforementioned 30 genes using Ingenuity Pathways Analysis (IPA). This analysis found that cell death, cell cycle and cellular development

were the most significantly enriched networks among the analysed 30 genes (p-value  $\leq 3.28e-02$ ; Figure 4.21)<sup>1</sup>. Some of the genes that are affected by focal CNVs, such as *CDKN2A* and *KIT*, are known cancer-related genes and have been reported to be frequently mutated or altered in a variety of cancers (Cancer Gene Census database).

In the next Section, I will show that several of these focal OPAS predicted events are real based on Illumina sequencing results. However, the CNV results from aCGH and 500K SNP array data analysed by SMD had previously failed to detect at least 3 of these known affected genes in the FL dataset, including deletions of *CDKN2A* (ht-16, ~104 kb), *KIT* (ht-20; ~119 kb) and *HVCN1* (ht-12; ~143 kb). More discussion about the putative effect of the copy number changes on the function of these genes is presented in Section 4.3.8.

#### 4.3.7 Comparison of OPAS Generated CNVs with Other Methods

Comparing candidate OPAS CNV calls with the results from other platforms applied to the same dataset can provide insight into the proportion of the events that are most likely real events. OPAS results were compared to the manually curated CNV results from aCGH and 500K Affymetrix SNP arrays<sup>2</sup>. To analyse the extra OPAS and SMD predicted CNVs (CNVs that were exclusively detected by OPAS or SMD), these events were compared with FPP events or Illumina sequence validated deletions and amplifications. It must be noted that 2 of the 25 FL patients were excluded from the comparison analysis, since these arrays were not evaluated by SMD or aCGH (ht-3 and ht-5). The following sections describe the aforementioned analysis on candidate deletions and amplifications.

#### **Analysis of Candidate Amplifications**

The Venn diagram in Figure 4.22 illustrates the overlap, similarities and differences between SMD, aCGH and OPAS predicted amplifications in 23/25 FL patients. As seen in this figure, the majority of the OPAS candidate amplifications (81/124) fell within the intersection of all 3 datasets; i.e., they are simultaneously observed by aCGH and SMD, as well as OPAS (denoted by 'A' in Figure 4.22). These amplifications ranged in size between 83 kb (9 SNP probes) to 242 Mb (21,382 SNP probes). The fact that all events in group 'A' are independently detected by 3 methods from two independent platforms (SNP array and array CGH) suggests that these are likely real copy number gains. There were in total 19 OPAS candidate amplifications that were not reported in

<sup>&</sup>lt;sup>1</sup>Other details of the IPA analysis are presented in Supplementary Table L.1.

 $<sup>^{2}</sup>$ Manual curation of CNV results had been performed by independent groups that generated the aCGH and SMD data. The final results of these analyses are available at Tumordb.

Tumordb by the alternative SNP array dataset (SMD results from 500K SNP array), referred to as group 'B' in Figure 4.22. Table 4.5 provides the list of all the aforementioned 19 OPAS candidate amplifications ('B'). As seen in this Table, 8 of these extra 19 events in group 'B' overlap with T-cell receptors (T-cell  $\gamma$  on 7p14,  $\beta$  on 7q34, and  $\alpha$  on 14q11.2). These apparent amplifications reflect the known rearrangements of the T-cell receptors that are frequently found in the normal DNA control.

Another known region of copy number gain that was missed in SMD results was a gain of the *IgH* locus on chromosome 14 (14q32) that is also known to be frequently rearranged in FL samples [387]. Subsequently, all regions in group 'B' were compared to sequence validated amplifications that confirmed another OPAS amplification on chromosome 6q16 region of patient 14 (shown in Supplementary Figure L.7). Additionally, 4 other candidate amplifications in this group ('B') were also observed by aCGH, implying that these candidate amplifications may also be real.

I also compared the amplifications that were only found by SMD (5 events in group 'C', denoted in Figure 4.22) and aCGH (11) with Illumina sequence validated amplifications, but no overlap was found. As shown in Figure 4.22, SMD and aCGH both reported one amplification event that was missed by OPAS. This amplification is approximately 150 kb and affects the *IgH* locus on chromosome 14 (14q32.33). Although this amplification was not validated by sequencing, the location of this candidate copy number gain and the fact that it was seen by both aCGH and SMD datasets imply that it is likely a real event that was missed by OPAS.

In summary, the data in Table 4.5 reveals that from the 19 amplifications that were seen by OPAS and not 500K results in Tumordb (SMD), at least 11 regions (58%) are most likely real events (9 overlapped with known copy number changes in T-cell and *IgH* loci, one was seen by both aCGH and OPAS and another one was validated by Illumina sequencing data), leaving only 8 OPAS amplifications in 'B' (42%) that can not be validated by available alternative datasets (referred to as OPAS-exclusive amplifications). As seen in Table 4.5, 3 of these 8 OPAS-exclusive candidate amplifications potentially affected a single gene, such as a small candidate amplification that affects *PAX5*. This focal amplification on 9p13.2 in patient 21 (ht-21) is approximately ~76 kb and encompasses 2 exons of *PAX5*, as shown in Figure 4.23. It is hypothesized that *PAX5* may play an important role in B-cell differentiation and regulation of the *CD19* gene that is a B-lymphoid-specific target gene [388]. In normal cells *PAX5* encodes the B-cell lineage specific activator protein that is expressed at early, but not late stages of B-cell differentiation. It has been reported that *PAX5* is consistently overexpressed in human follicular lymphomas [389]. *PAX5* is also involved in the t(9;14)(p13;q32) translocation that is recurrently observed in a subtype of B-cell lymphomas (small lymphocytic lymphomas) as well as large-cell lymphomas. This translocation juxtaposes an

enhancer of  $I_gH$  gene to *PAX5* promoter, suggesting that the changes in the *PAX5* gene transcription likely play a significant role in the aforementioned lymphomas. Although the FPP data did not detect any translocation at this site in patient 21 and no other dataset found this amplification, the OPAS-detected significant gain of signal intensity at this locus (LR = 0.33; z-score  $\simeq +2.4$ ), depicted in Figure 4.23, suggests that the above partial amplification of *PAX5* may be a real event. Sequencing data would have provided ultimate confirmation of this event, however, since this event was not seen by FPP analysis sequencing was not performed. Furthermore, the bootstrapping analysis found that it is unlikely that the candidate amplification of *PAX5* exons was obtained by random chance (p-value P = 2.2e-308; bootstrapping method with 100,000 permutations). Based on the above descriptions, it can be speculated that the candidate partial amplification on 9p13.2 chromosomal region may have triggered a mechanism that leads to expression of *PAX5* in FL patient 21. However, this event does not reoccur in any other 24 FL patients in this study.

In conclusion, based on the Venn diagram of Figure 4.22, ~83% (104/124) of OPAS candidate amplifications were observed by at least one alternative array-based dataset and ~64% (81/124) were observed by both array-based results. Furthermore, based on the information in Table 4.5 at least 58% (11/19)<sup>1</sup> of OPAS candidate amplifications that were not detected by 500K SNP results of Tumordb (region 'B') and 47% (7/15)<sup>2</sup> of amplifications that were exclusively detected by OPAS are likely real copy number gains.

#### **Analysis of Candidate Deletions**

Similarly to the approach that was used to analyse candidate amplifications, an analysis of candidate deletions in 23/25 FL patients was also performed. The Venn diagram of Figure 4.24 shows that 51 of 132 (39%) OPAS candidate deletions were also observed by aCGH and SMD and ranged in size between 519 kb-113 Mb. The estimated concordance between SMD and OPAS deletion results and the reciprocal comparison was 95% and 71%, respectively. In contrast, there was a relatively poor concordance between aCGH and SNP array-based results (43% concordance for aCGH and OPAS, and 53% for aCGH and SMD).

The apparently reduced concordance between OPAS and aCGH (43%) compared to SMD and aCGH (53%) is due to the fact that there were more deletion events called by OPAS. However, it is important to assess whether the additional events predicted by OPAS are real. Therefore, in

<sup>&</sup>lt;sup>1</sup>8 of the 19 OPAS candidate amplifications in 'B' overlapped with T-cell receptors, 1 overlapped with IgH locus, 1 was validated by Illumina sequencing and another one was seen by aCGH.

 $<sup>^{2}</sup>$ 6 of the 15 OPAS-exclusive candidate amplifications overlapped with T-cell receptors, and 1 was validated by Illumina sequencing.

the next step I used the data from FPP and Illumina sequence validated deletions (165) to assess whether any of the OPAS-exclusive candidate deletions (32) would likely reflect a real copy number loss that was missed by these alternative datasets. Similar analysis was performed to study SMD exclusive CNV calls. Each FPP event that mapped to a candidate deletion was further analysed by manual inspection of the predicted FPP event in Tumordb to increase the possibility that the reported FPP event was likely a real copy number loss<sup>1</sup> (such as examples shown in Figures 4.25-4.28 and Supplementary Figure L.8). As summarized in Table 4.6 and indicated in Figure 4.24, from the 32 regions that were exclusively seen by OPAS, 11 overlapped with FPP events. These 11 OPAS-exclusive deletions included 5 candidate deletions that mapped to FPP fragment holes<sup>2</sup> (e.g., Figure 4.19), 2 candidate deletions that aligned to FPP coverage gaps (e.g., Figures 4.25-4.26), and 4 other candidate deletions that mapped to regions that harbored complex FPP events (such as in Figures 4.27, 4.20 and Supplementary Figure L.8). Furthermore, 4 other OPAS-exclusive deletions mapped to Illumina sequence validated copy number losses, (e.g., Figure 4.28). Three of these cases are discussed in more detail in the next section (Section 4.3.8). These sequence validated deletions had 4, 8 and 13 SNP probe markers and were expected by OPAS to be between 104-143 kb. These findings indicate that from the 32 extra regions that were exclusively found by OPAS, at least 4 were real copy number losses, and 11 others mapped to FPP events but were not studied by sequencing; the remaining 17 regions may also be real copy number losses. As seen in Table 4.6, most of these OPAS exclusive deletions only affect a few genes (5 affect a single gene and 3 others affect between 2-3 genes). In the next Section, I will discuss about 3 genes that were affected by validated OPAS-exclusive deletions (HVCN1, CDKN2A and KIT) and the significance of these CNV findings in the context of cancer.

### 4.3.8 Examples of Sequence Validated Novel (OPAS-exclusive) CNV Findings

#### 4.3.8.1 HVCN1

One of the OPAS exclusive CNVs that was validated by sequencing data was a deletion on 12q24.11 chromosomal region of FL patient 6 (ht-6) that included only 4 SNP probes, shown in Figure 4.28. The exact deletion breakpoints found by sequencing indicated that this deletion is 145,148 bp (OPAS predicted this deletion to be 143,373 bp; Fig. 4.28a). Based on sequence-validated breakpoints this deletion removes all but the first exon of *HVCN1* and the first 3 coding exons of *PPTC7*,

<sup>&</sup>lt;sup>1</sup>I used the feedback from Dr. Andy Mungall and Matthew Field at the GSC to interpret the FPP results.

<sup>&</sup>lt;sup>2</sup>see page 283 for description of these events.

juxtaposing the *HVCN1* exon to the remaining *PPTC7* exons (Fig. 4.29a). Therefore, the above deletion can potentially create a novel fusion gene between the 5' end of *HVCN1* and 3' end of *PPTC7*. It must be added that the probability of detecting a similar deletion (143.4 kb) that overlaps with *HVCN1* just by random chance is estimated to be P = 3.0E-05 (based on bootstrapping method with 100,000 permutations; see Table 4.7). Thus, it is unlikely that the detected somatic partial deletion of *HVCN1* in ht-6 was just a random observation.

*HVCN1* is a newly discovered protein that was identified as the result of a proteomic analysis of B-cell plasma membranes that were isolated from patients with mantle cell lymphoma (MCL) [390]. A recent study by Capasso et al. [263] using expression data and function analysis showed that *HVCN1* is a key modulator of B-cell antigen receptor (BCR)<sup>1</sup> signalling. This group also indicated that absence of *HVCN1* blunts the immune response [263]. Based on this evidence, it can be speculated that loss of *HVCN1* may disrupt the role of immune system in recognizing and destroying cancer cells. Therefore, based on these observations, *HVCN1* could either act as an oncogene or a tumor suppressor in B-cell malignancies, and further studies are needed to delineate its function. Finding of a deletion of *HVCN1* supports the putative tumor suppressor role of this gene in evading the host immune response.

One speculation about the impact of the focal deletion on 12q24.11 chromosomal region of ht-6 is the possibility of a fusion gene, as mentioned earlier. To further investigate this possibility using computational tools, I joined the sequence of the 5' end single remaining exon of *HVCN1* and the three remaining 3' end exons of *PPTC7* (as depicted in Fig. 4.29). The hypothetical protein product of this synthetic sequence was then generated using an online tool (Six Frame Translation of Sequence). As shown in Figure 4.29, the *HVCN1/PPTC7* hypothetical fusion appeared to be in-frame and, thus, could result in a translated protein product. These speculations need to be experimentally validated to deduce whether the aforementioned deletion has led to a novel fusion gene in ht-6. If proved, functional analysis would be required to characterize the function of the novel protein and whether it contributes to cancer.

In addition to the above OPAS-exclusive partial deletion of *HVCN1* in ht-6, this gene was entirely deleted in another FL patient (ht-9) as the result of a larger deletion (~1.4 Mb), shown in Figure 4.30. The latter deletion in ht-9 is not OPAS-specific and SMD (Fig. 4.30b) and aCGH (Fig. 4.30c) results also detected this copy number loss event in patient 9.

<sup>&</sup>lt;sup>1</sup>In B-cells the stimulation of B cell antigen receptors results in production of ROS that participate in B-cell activation [391].

#### 4.3.8.2 CDKN2A

Another example of a sequence validated OPAS-exclusive copy number loss is a deletion located on chromosome 9p21.3 in patient 16 (ht-16) that has 8 SNP probe markers, as shown in Figure 4.31. The exact breakpoints found by sequencing revealed that this deletion was 124,010 bp (OPAS data had estimated this deletion to be ~104.4 kb). As summarized in Table 4.6, this deletion encompasses the *CDKN2A* known tumor suppressor gene. This table also indicates that the probability of detecting a similar CNV (104,434 bp) that overlaps with *CDKN2A* by random chance is P = 4.0E-05 (based on bootstrapping method with 100,000 permutations). The latter suggests that it is unlikely that the observed deletion of *CDKN2A* in ht-16 is a random finding.

CDKN2A is a known cancer-related gene<sup>1</sup> that is frequently mutated or deleted in a wide variety of cancer cell lines as well as primary tumors, such as lung, breast, brain, bone and lymphocyte [260, 261, 392-394]. A study in childhood acute lymphoblastic leukemia (ALL) found that although mutations of CDKN2A deletion were rare in childhood ALL, deletion of this gene was a significant secondary abnormality in ALL and was strongly correlated with the observed phenotype and genotype. A recent study by Jardin et al. [395] has also showed that patients with diffuse large B-cell lymphoma or DLBCL (i.e., an aggressive form of B-cell lymphoma) with CDKN2A deletion, have a distinct gene expression pattern and poor prognosis. The CDKN2A gene encodes two important cell cycle regulatory proteins, the p16(INK4) protein that interacts with the RB1 pathway; and the p14(ARF) protein which regulates the p53 pathway. The RB1 and p53 pathways are two of the major tumour suppressor pathways in human carcinogenesis. A deletion of CDKN2A would therefore disturb both pathways. The impact of the loss of the CDKN2A proteins were previously studied in mice models by Serrano et al. [396]. His study showed that mice deficient for both p16(INK4a) and p19(ARF) were viable but highly prone to tumors and died early in life due to lymphomas and fibrosarcomas [396]. Another study by Schmitt et al. [397] using p16/p19 null mice indicated that in these mice lymphomas were formed rapidly, were highly invasive and displayed apoptotic defects. Furthermore, a study by Lossos et al. [398] had found inactivation of CDKN2A and its adjacent gene, CDKN2B, as one of the most commonly identified genetic alterations that were associated with transformation of FL to DLBCL. The body of evidence on CDKN2A deletions and mutations emphasize that this gene plays a significant role in cancer and may even have a more profound impact in B-cell lymphomas and in transformation of FL to DLBCL.

In addition to the above OPAS-exclusive deletion of *CDKN2A* in patient 16 (ht-16), the OPAS results also indicated deletions on chromosome 9p21.3, encompassing *CDKN2A* gene, in 3 other

<sup>&</sup>lt;sup>1</sup>http://www.sanger.ac.uk/genetics/CGP/Census/

FL patients (ht-18, ht-22 and ht-24). These results are shown in Figure 4.32 and Table 4.7. Further FISH experiments validated the above 4 deletions and also indicated that the *CDKN2A* deletions were homozygous in 3 cases (ht-16 (Fig. 4.32b), ht-18 (Fig. 4.32c) and ht-24 (Fig. 4.32e)), and heterozygous in 1 case (ht-22; Fig. 4.32d). Nonetheless, aCGH results in the FL dataset found only one of the above *CDKN2A* deletions (ht-24), and SMD detected 3 of these 4 real deletions (ht-18, ht-22 and ht-24).

In conclusion, 16% (4/25) of all patients in the FL dataset have somatic copy number loss of *CDKN2A*, whereas no amplification was detected in this gene in any FL patient. This finding emphasizes that the focal deletion of this gene in these patients maybe part of a mechanism in FL genomes that tends to specifically remove the function of *CDKN2A*.

#### 4.3.8.3 KIT

A deletion on chromosome 4q12 in FL patient 20, depicted in Figure 4.33, is another example of an OPAS predicted deletion that was not previously detected by aCGH or SMD results (OPAS-exclusive deletion). Illumina sequencing validated this CNV and estimated that the exact size of this deletion was 136,811 bp (OPAS data had estimated this deletion to be  $\sim$ 119.5 kb).

As illustrated in Figure 4.32a, this focal deletion occurs within a larger copy number amplified region. Array CGH results in Tumordb did not indicate any CNV at this locus, and SMD results indicated the entire region is amplified (shown by green track in Figure 4.32b). Based on the sequencing information this deletion is 136,811 bp in length and appears to target the *KIT* gene (OPAS results had predicted the deletion to be ~120 kb), as shown in Figure 4.32a. The probability of detecting a similar event (119,507 bp) that overlaps with *KIT* just by random chance is estimated to be P = 1.0e-04 (based on bootstrapping method with 100,000 permutations; see Table 4.6). The latter implies that it is unlikely to detect a random ~120 kb deletion that overlaps with the *KIT* gene.

*KIT* is a proto-oncogene that encodes a receptor tyrosine kinase (RTK) that is crucial to melanogenesis, hematopoiesis and gametogenesis [399]. Gain-of-function mutations of *KIT* (i.e., mutations that cause constitutive activation of the *KIT* tyrosine kinase) have been associated with several cancers including acute myelogenous leukemia (AML), sinonasal T-cell lymphomas and gastrointestinal stromal tumor (GIST) [400] (the sites of common mutations of *KIT* are denoted in Figure 4.34b). A study by Kitayama et al. [401] in transgenic mice showed that mice with a specific mutation developed acute leukemia or malignant lymphoma. Since *KIT* has an oncogene-like function, it is expected that in cancer *KIT* would be amplified and not deleted. However, further inspection of *KIT* deletion site in FL patient 6 indicates that the 136,811 bp deletion does not remove *KIT* entirely, but only affects the first 3 exons of this gene (as shown in Figure 4.34a). Figure 4.34b illustrates the structure of the KIT protein that includes an extracellular domain (exons 1-9), a transmembrane domain (exon 10), and an intracellular domain (exons 11-21). It has been shown that the deletion of the extracellular domain of RTKs facilitates the ligand-independent activation of this family of genes, since this event removes negative regulatory constraints that are imposed by the extracellular domain [402]. However, it was also shown that such deletions were not sufficient for constitutive activation of RTKs and the oncogenic transformation required additional activating mutations [402, 403]. Introduction of KIT with a deletion of the ligand-binding domain into mice and cell line models did not result in the constitutive activation of the kinase, while the deletion of ligand-binding domain coupled with an activating mutation in the kinase domain did have such an effect [403]. This observation suggests that amplification of the kinase domain of KIT seen in patient 20 might act similarly to the previously reported gain-of-function point mutations in the kinase domain of KIT (in GIST tumors and leukemias). Therefore, the amplification of the kinase domain may act synergistically with the deletion of the ligand-binding domain (exons 1-3) to result in the constitutive activation of KIT. This finding may represent the first report of such a mechanism of KIT activation in lymphomas.

## 4.4 Conclusions

The study presented in this Chapter described high-resolution SNP array analysis of candidate somatic CNVs in 25 follicular lymphoma patients (FL). An analytical approach was used to detect candidate regions of copy number variation in these genomes. This approach involved statistical analyses of OPAS results from 250K Nsp arrays, as described in Section 4.3.1. In total, 286 candidate CNVs were identified among 25 FL patients (11.4 per patient) of which 53 (18%) were smaller than 150 kb. The profiling of these events found 3 main categories of candidate CNVs among FL patients (Sections 4.3.3-4.3.5). A separate analysis was performed to study putative focal events in the FL dataset (CNVs  $\leq$  150 kb) which found 53 (18.5%) such events in the FL dataset (Section 4.3.6). Gene analysis of candidate focal CNVs revealed that they affected more than 30 genes that were found to be most significantly related to cellular mechanisms that are particularly important in cancer, including cell death, cell cycle and cellular development (Figure 4.21). Some of the genes that were affected by these candidate focal events were *CDKN2A*, *ERBB4*, *KIT* and *HVCN1* (Tab. 4.4; all except *ERBB4* focal CNVs were also validated by Illumina sequencing). All these events, except a candidate small intronic amplification of *ERBB4*, were seen only in one FL patient. The latter finding suggests that CNV events in FL do not always affect the same genes, but instead, they likely affect a variety of genes that may have important roles in FL tumorigenesis or progression.

To assess the accuracy of the putative CNV calls, I compared OPAS findings with the results from several alternative technologies that were applied to the same data set, including BAC aCGH, 500K SNP array results from SMD analysis, BAC fingerprint profiling (FPP) and Illumina sequence validated data (Section 4.3.7). This comparison, which was performed in 23/25 FL samples, indicated that ~80% (204/256) of all OPAS candidate CNVs in these 23 samples were seen by at least one other array-based dataset (aCGH or SMD) (Figures 4.22 and 4.24). Importantly, from the remaining 47 putative OPAS-exclusive putative CNVs, at least 22 (46.8%) regions are likely real based on the comparison to FPP events or Illumina sequence validated CNVs (6/22 regions were not seen by FPP, but these were apparent gains of T-cell receptors which are common in B-cell versus T-cell comparisons, and thus were included in this list). Table 4.6 indicated that more than half of the candidate OPAS-exclusive deletions (Figure 4.24) had fewer than 10-15 SNP probes, among which were sequence-validated deletions on chromosomes 9p21.3 (~104 kb; 8 SNPs), 12q24.11 (~143 kb; 4 SNPs), and 4q12 (~120 kb; 8 SNPs), that affected important cancer-related genes (Section 4.3.8). These findings emphasize that a CNV calling approach that uses a fixed SNP threshold as a requirement to call copy number aberrations (for instance, a minimum of 10 deviated SNPs) would inevitably fail to detect such important CNVs.

In summary, the CNV analysis in 25 FL genomes presented this Chapter has observed several large-scale FL CNVs that had been previously found in other studies (Section 4.3.3) and showed that copy number losses at distal regions of chromosomes are frequent observations in FL (Section 4.3.4). Additionally, this work provided further insight into the extent of much smaller CNVs that can potentially affect important genes in FL (see Table 4.4). An example of a new validated discovery of this study was detecting a focal deletion of *CDKN2A* tumor suppressor gene in FL patient 16 that was not previously reported by other array-based datasets in Tumordb. *CDKN2A* (which is deleted in 4/25 FL patients) is important in many cancers and may be involved in progression of FL to DLBCL [398] (Section 4.3.8.2). Another example of a new validated focal CNV was a partial deletion on 12q24.11 (ht-6) that may result in a fusion event between the *HVCN1* and *PPTC7* genes (*HVCN1* is a newly discovered gene that is a key modulator of B-cell antigen receptor signalling, and *PPTC7* is a T-cell activation protein) (Section 4.3.8.1). Another interesting new discovery was detecting a small partial deletion of the *KIT* extracellular domain (ht-20) that, combined with the amplification of the intracellular domain of this gene, may result in constitutive activation of this proto-oncogene (Section 4.3.8.3).

## 4.5 Figures and Tables



**Figure 4.1: Example of an FPP event on chromosome 4q12 in FL patient 20 that is proved to be a deletion by Illumina sequencing.** (a) Screenshot of Tumordb illustrating FPP alignment of BAC clones from chromosome 4q12 of FL patient 20 to the reference human genome (hg18). BACs with linear alignments to the reference are coloured blue and those with split alignments are coloured green ('multi fpp'; See K). The red arrows indicate the ends of BAC clone HTa20-0033M06 have been aligned to two distinct locations on 4q12. The latter suggests that the region between these two positions, highlighted by the pink rectangle, may have been deleted in this patient. (b) The PCR assay proves that the event is somatic. Three sequence tagged sites were designed; one to interrogate the breakpoint ('bpt'), and two controls in the flanking sequences, located centromeric ('cen') and telomeric ('tel') of the breakpoint. Each sequence tagged site was used to amplify tumour DNA and matched peripheral DNA. Since the breakpoint exists in the tumor but not peripheral blood, it is concluded that the deletion of 4q12 is a somatic event. c) Capillary sequencing of patient 20 tumour DNA validated the deletion and identifies the exact deletion breakpoints (55,123,271-55,260,082). The red rectangle indicates the part of the sequence which has been deleted based on the sequencing results (this information is used in Section 4.3.8 to confirm a deletion that partially affects the *KIT* gene).





(**b**) PDF of predicted LR values

**Figure 4.2: Distribution of LR intensity measurements of all regions across FL dataset.** Panel (a) denotes the cumulative density function (CDF) of all OPAS detected regions (with more than 2 SNPs) from FL dataset with 25 samples (total of 1931 regions). The arrows labelled 'A' and 'B' mark two apparent change points in the CDF curve, corresponding to 7% and 93% of all OPAS regions (denoted by the red dashed lines). The concavity of the CDF curve appears to change at these two points ( $\pm 0.2$ ). However, between these two markers (area highlighted in green), the CDF is approximately normal. Panel (b) illustrates the PDF of the same data (LR measurements of 1931 regions. The PDF plot also suggests that OPAS regions in the green area can be approximated by a normal distribution, however, the regions outside this area seem to have different distribution(s). Based on the observations from (a) and (b), it can be inferred that OPAS regions that fall outside the green area are statistically different from the rest of the regions. Thus, it is speculated that these regions constitute candidate copy number changes (CNVs) of the FL dataset. These putative CNVs consist of the bottom 7% of all OPAS results, corresponding to regions with LR  $\leq -0.2$ , and the top 7%, corresponding to regions with LR  $\geq +0.2$ .



(a) OPAS scatterplot of chromosome 1 of patient 9, denoting a deletion on 1p36 (~kb; SNPs)



(b) FISH results of 1p36 deletion in ht-9 (22%)

Figure 4.3: Deletion on 1p36 chromosomal region of ht-9 with slight signal deviation (LR= -0.13), validated by FISH. This figure presents the OPAS scatterplot of chromosome 1 of patient 9 (ht-9). As observed, there is a region on 1p36 that has a lower signal intensity, compared to the rest of the chromosome 1 (LR = -0.13). The same region has been detected as a putative deletion by aCGH and SMD results of Tumordb, as shown in Table 4.3. It is observed that the OPAS estimated LR is -0.13, and z-score is -0.7. Based on the z-score and visual inspection, the above region in ht-9 was selected as a putative deletion. Panel (b) shows the FISH analysis performed to analyse the above candidate deletion in ht-9 (1p36.23). The FISH experiment confirmed this deletion. Furthermore, it indicated that the copy number loss was present in approximately 22% of the cells (data provided by Dr. Horsman's lab, BCCRC). The latter finding, which highlights the CNV heterogeneity in this case, explains the relatively small magnitude of signal intensity loss that was detected in this region (LR = -0.13).



(b) Probability Density Plot of z-scores of all regions with gain of signal intensity



Figure 4.4: Probability Density Function (PDF) of z-scores from all regions with gain (LR > 0) or loss (LR < 0) of log2-ratio signal intensity (in FL dataset). Panel (a) shows the histogram of regions with LR values  $\leq -0.2$  (referred to as "X"), compared to other OPAS-estimated regions with slight loss of signal intensity with -0.2 < LR < 0 (referred to as "L"). The approximate normal fit to the histograms is shown by black (for "L") and red (for "X") dashed lines. The red dashed line denotes the z-score that is used to call candidate regions with slight loss of signal intensity that may represent significant changes based on their z-scores (z-score  $\leq -0.6$ ). Panel (b) shows similar analysis to compare the distribution of z-scores for regions with significant gain of signal intensity (LR  $\geq +0.2$ ; referred to as "Y") with regions with slight gain of signal intensity (0 < LR < +0.2; referred to as "G"). The approximate normal fit to these data is shown by black (for "G") and blue (for "Y") dashed lines. As observed, the two distributions have relatively smaller overlap compared to deletion analysis in (a). This plot shows that the z-score of +0.6 approximately separates these two datasets ("G" and "Y"). The latter finding suggests that OPAS regions with slight gain of signal intensity but z-scores  $\geq +0.6$  can also be considered as candidate regions of copy number amplification.



Figure 4.5: Boxplot of z-scores of all regions with loss or gain of log2-ratio signal intensities (in FL dataset). Panel (a) shows the boxplots of z-scores in OPAS-generated regions with loss of signal intensity from 25 FL samples (see Appendix E for more information about boxplot visualization). The boxplot on the left, denoted by 'X', represents the distribution of z-scores of candidate deleted regions with LR  $\leq -0.2$ . The boxplot on the right, denoted by 'L', indicates the z-scores of all other OPAS regions with slight loss of signal intensity (-0.2 < LR < 0). Similarly, the boxplots in panel (b) indicate the distribution of z-scores of OPAS regions with significant gain (LR  $\geq 0.2$ ; 'Y') and slight gain of signal intensities (0 < LR < +0.2; 'G'). The red arrow in (a) and blue arrow in (b) indicate regions with slight loss or gain of signal intensity that have significantly different z-scores (compared to the rest of the regions with slight LR deviation). These outliers likely include a subset of real CNVs that have slight magnitudes of signal aberration.

As seen in these plots, for both amplification and deletion analyses, the distribution of z-scores of regions with significant LR values (|LR| > 0.2; 'X' and 'Y') overlaps with z-scores of regions with slight signal deviations (0 < |LR| < 0.2). However, these 2 figures indicate an important difference. For amplification analysis in panel (b), only the outliers of distribution 'G', denoted by the blue arrow, overlap with the z-scores of regions with significant LR increase ('Y'). Therefore, all regions with slight signal intensity gain that have significantly different amplification z-scores (subset of 'G' with z-scores  $\ge +0.6$ ) were also added to the list of candidate somatic amplifications of the FL dataset. Although, as seen in panel (a), the overlap of the z-scores of regions with significant loss ('X') and slight loss ('L') of signal intensity includes more that just the outliers of 'L' (red arrow). Following manual inspection of the OPAS visualization plots, I decided to add regions with slight loss of signal intensity that had z-scores equal or less than -0.6 (shown by the red dashed line in (a)) to the list of candidate copy number deletions (subset of 'L' regions with z-scores  $\le -0.6$ ).



Figure 4.6: Candidate deletions on chromosome 4 of patient 29 with slight signal deviations (LR = -0.16 and -0.13) but significant z-scores (-0.84 and -1.1). This figure illustrates the OPAS scatterplot of chromosome 4 of patient 29 (ht-29). As seen there are two predicted regions of copy number deletion with slight loss of signal intensity, with LR = -0.16 and LR = -0.13. The estimated z-scores of these candidate deletions are -1.1 and -0.84, respectively, indicating strong losses of signal intensity compared to the rest of SNPs on chromosome 4 of patient 29. None of these putative deletions was detected by other CNV datasets, however, the observed patterns of copy number losses in these regions and the estimated z-scores suggest that these candidate events may be real deletions.



Figure 4.7: Pie charts of the frequency of candidate amplifications and deletions in 25 FL patients. Panel (a) indicates that the frequency of candidate somatic amplifications is approximately comparable to that of deletions (53% deletions versus 46% amplifications). Panel (b) illustrates that at smaller sizes, ~150 kb or less, candidate deletions are ~1.8 times more frequent than amplifications (64% deletions compared to 36% amplifications). Panel (c) shows that in contrast to small events, for large CNVs  $\geq$  8 Mb, the frequency of amplifications is more than two times higher than deletions (70% amplifications versus 30% deletions). These observations suggest that although the overall frequency of candidate deletions and amplifications is similar in the FL dataset, the proportion of large-scale putative amplifications is greater than that of large-scale deletions. Additionally, the frequency of small deletions in the FL data set is likely more than small amplifications.



Figure 4.8: Examples of two real whole chromosome gains in an FL patient (ht-11) with slight log-ratio deviations from base-line (LR = 0.12 and 0.13). The red horizontal dashed line in the top panel denotes the single OPAS estimated region for chromosome 7 in FL patient 11 (ht-11). This line indicates a clear shift of log2-ratio intensities from the baseline (LR = 0; shown by the blue line), but only with a slight deviation (LR = 0.12; z-score = +0.8). The bottom panel shows similar observation in chromosome 18 of the same patient (LR = 0.13; z-score = +0.84). Similar to chromosome 7, the magnitude of the apparent gain of chromosome 18 is also small relative to the theoretical LR value of one copy gain (LR = 0.13 compared to log 2(3/2) = 0.58). Despite the slight LR deviations, as shown by the G-banded karyotype, these events reflect real chromosome duplication (+7 and +18) in the above patient (ht-11).

WCA Event Matrix



**Figure 4.9: Frequency of WCA events per chromosome across all FL patients.** Each row corresponds to a chromosome or chromosome arm that is affected by at least one WCA event in the FL dataset. The columns represent the 17 FL patients that carry such events within the FL dataset. The numbers at the top row indicate the sum of all such events per FL sample, sorted from left-to-right based on patients with the most (ht-21) and the least (ht-7) number of candidate WCA events. In total 17 of the 25 FL patients had at least one WCA event and, thus, there are 17 patient columns in this graph. The last column on the left represents the sum of all WCA events in a corresponding chromosome or chromosome arm. As shown in this graph, there is an evident abundance of large-scale WCA gains (shown in blue) in the FL dataset. The patients with high number of WCA events also seem to have similar patterns of CNVs, including gains of 1q and X and losses of 6q which are present in all patients with more than 4 WCA events.



Figure 4.10: Chromosome ideogram view of 48 WCA CNVs in the FL dataset. Regions of copy number gain are shown by the red lines to the right side of each associated chromosome and regions of copy number loss are denoted by the green lines on the left side of the chromosomes. In addition to gains and losses of chromosomes and chromosome arms, another observed pattern of whole chromosomal alteration in the FL dataset was iso-chromosome<sup>1</sup> 6p (i(6p)) that was detected in 3/25 (12%) FL patients. This event includes simultaneous gain of the short arm (+6p) and loss of the long arm (-6q) in the same patient, which is a known recurrent event in FL genomes<sup>2</sup> As seen, in total, there are 48 WCA CNVs in the FL dataset, the majority of which are copy number gains (42/48 = 87.5%). This plot also shows that WCA events of chromosomes 1q, 2p, 6, 7, 12, 18, 21 and X are observed in at least 3 patients ( $\geq 12\%$ ) of the FL dataset.



(a) OPAS result; Chr 8, patient 29



(c) mFISH result, patient 29



(b) OPAS result; Chr X, patient 29



(e) aCGH result, Chr 8, ht-29



(f) aCGH result, Chr X, ht-29

Figure 4.11 (*previous page*): The only two WCA events that were not directly validated by cytogenetic analysis (slight gains of chromosomes 7 and X in ht-29). Panels (a) and (b) present the OPAS scatterplot of chromosomes 7 and 8 of patient 29 (ht-29). The horizontal red lines represent OPAS estimated region(s) in these chromosomes. These plots indicate a slight positive shift of the red line with respect to the base line (LR = 0) that is denoted by the blue dashed line. Panel (a) shows that the entire chromosome has an LR of ~0.07. Panel (b) shows that the entire chromosome X has LR of ~0.08. As indicated in panel (c), none of these putative gains are observed by M-FISH. However, as shown in panel (d), SMD results also detected a slight gain of whole chromosome X in this patient (reported in tumordb). Panels (e) and (f) present the BAC aCGH results of chromosomes 8 and X of ht-29, respectively. These aCGH plots, show a slight increase of signal intensity in both of these chromosomes (aCGH observations are not reported in tumordb, since these slight gains do not pass the significance threshold). Therefore, the slight gain of chromosome 8 in ht-29 is observed by both OPAS and SMD results; and slight gain of chromosome X in this patient is observed by 3 datasets from two separate platforms. The latter implies that although these candidate events are not supported by cytogenetic analysis, they may indicate real copy number amplifications that are presents in a small subpopulation of the cells in ht-29.



(a) OPAS scatterplot of chromosome 22 in FL patient 22



(**b**) MFISH results of patient 22

(c) aCGH result of chromosome 22 of patient 22

**Figure 4.12: Example of a distal CNV on chromosome 22 of an FL patient (ht-22).** Panel (a) represents the OPAS scatterplot of a distal CNV affecting chromosome 22 (q13.2-q13.33) in FL patient 22 (ht-22). The deleted region, denoted by the red arrow, is approximately 8.6 Mb (718 SNPs ) and includes the distal end of chromosome 22. Panel (b) shows the MFISH result of this patient (ht-22). As seen, the MFISH analysis detects a translocation between chromosomes 17 and 22, and added material to chromosome 22, however, it does not detect the predicted deletion in this chromosome. Panel (c) shows the array CGH result of chromosome 22 of ht-22. It is observed that aCGH also identifies the same deletion that was seen by OPAS (22q13.2-q13.33). Based on observed consistent patterns of deletion in (a) and (c) it can be concluded that the candidate distal deletion in ht-22 is a real event; however, this event was not identified by MFISH. The latter is due to the fact that conventional cytogenetic analysis has a particularly low resolution in detecting CNVs in regions proximal to chromosome ends.



(a) OPAS scatterplot of chr 5 of patient 12



(b) aCGH results of chr 5 of patient 12

Figure 4.13: Two candidate amplifications on chromosome 5 of an FL patient (ht-12), including a distal copy number gain on 5 q-end, detected by OPAS, SMD and aCGH results. Panel (a) presents the OPAS scatterplot of chromosome 5 of patient 12 (ht-12). The blue and red arrows indicate two large regions, approximately 15 and 22 Mb, that have slight gains of signal intensities. As shown in this figure, region #1 (5q23.3-q31.3) has LR = 0.08 and a corresponding z-score of 0.64. Similarly, region #2 (5q34-q25) has LR = 0.07 and z-score = 0.6. Both of these slight gains have z-scores  $\geq$  0.6 and indicate patterns that support copy number amplifications. Based on the estimated z-scores and the observed pattern, these regions (regions #1 and #2) were selected as putative copy number amplifications in ht-12. In fact, both of these candidate amplifications are also detected by aCGH and SMD results in Tumordb (both aCGH and SMD results in Tumordb are generated by coupling computational analysis and visual inspection). Panel (b) illustrates the aCGH results of chromosome 5 of FL patient 12. The two blue vertical lines represent aCGH detected copy number gains on 5q23.3-q31.2 (CNV #1) and 5q34-q35 (CNV #2). These aCGH amplified region correspond to the same OPAS candidate copy number gains that were shown in (a). The similar findings of aCGH and OPAS imply that the aforementioned putative slight gains on chromosome 5 (ht-12) are real amplifications.



Patient 7 - Chromosome 1p36.33-p33

Figure 4.14: Candidate distal deletion on chromosome 1p36 (ht-7) with slight LR deviation but a significant z-score (LR = -0.12; z-score = -1). This figure presents the OPAS scatterplot of a portion of chromosome 1 of patient 7 (ht-7) that encompasses a candidate distal deletion (denoted by arrow) on chromosome 1p36 (1p36.33-p36.32). The 1p36 region is a known deletion hotspot of FL genomes [244]. The denoted deletion has a low magnitude of loss of signal intensity, with LR = -0.12. However, when the aforementioned LR value (-0.12) is compared to the distribution of signal intensities of the entire chromosome 1 in this patient (ht-7), the estimated z-score is -1. Additionally, the loss of signal intensity in this region also seems to be visually recognizable from the scatterplot. Therefore, these observations indicate that the above deletion is a putative copy number loss in ht-7. As shown in Table 4.3, aCGH and SMD datasets did not report this deletion, however, SMD indicated a copy number neutral LOH ("cnnloh") at this region in ht-7. In Affymetrix SNP arrays, deleted regions are predominantly called homozygous by genotyping algorithms. Therefore, observing a copy number neutral LOH for 1p36 region of patient 7 by SMD, may reflect a real deletion that was not detected by this method. Collectively, these data and the above figure suggest that the OPAS detected 1p36 loss of signal intensity in ht-7, is a candidate deletion.



**Figure 4.15: Recurrent distal deletion of chromosome 1p36 in the FL dataset.** The above ideogram of chromosome 1 illustrates the predicted recurrent deletions in 1p36 in the FL dataset (red lines), compared to 1p36 deletion hotspot that has been reported by Cheung et al. [244] (blue line). Cheung et al. [244] found ~11 Mb region on 1p36.22-p36.33 as the most frequently altered region in 106 FL samples, with 25.5% deletion rate. As indicated by the red lines in this figure, deletions of 1p36 chromosomal regions were detected in 8/25 (32%) patients in the FL dataset. From these 8 predicted deletions, 7 were confirmed by FISH (FISH experiments were performed by Dr. Horsman's group at the BCCRC; see Supplementary Figure L.5). For one patient, ht-7, there was no FISH experiment to identify whether this candidate deletion was real or not.



(a) Most significant gene network, associated with candidate distal deletions



(b) Most significant gene network, associated with candidate distal amplifications

**Figure 4.16:** Most significantly associated gene networks with candidate distal CNVs of the FL dataset. These results are based on IPA analysis ( $P \le 0.02$ ), after correcting for CNV false discovery rate using the Benjamini and Hochberg method. Panel (a) shows the top network of genes that is associated with candidate distal deletions. Several of these molecules are important in a variety of cellular processes. For instance, the red arrow denotes Caspase, a family of proteins that are vital in apoptosis (source: OMIM). Panel (b) shows the top gene network associated with candidate amplified distal CNVs among 25 FL patients. Similar to the gene network of distal deletion (a), the network of distal amplifications (b) also includes several known cancer-related genes such as *MYC*. This gene, marked by the blue arrow, is a proto-oncogene that is often up-regulated in many cancers and has been linked to a variety of hematopoietic tumors, leukemias and lymphomas (source: Entrez).



Figure 4.17: Frequency of candidate CNVs in each chromosome among 25 FL patients. The red and blue bars represent the number of candidate WCA (category 1) and distal (category 2) CNVs per chromosome among 25 FL patients, respectively. The number of all other candidate CNVs (category 3) in each chromosome is also denoted by the pink bars. The noticeably high frequency of putative CNVs in chromosomes 14 and 7 is due to the copy number changes of IgH and T-cell receptor genes in these chromosomes, that is common in follicular lymphoma.



(a) OPAS scatterplot of chromosome 13 of patient 21,(b) SMD plot of chromosome 13 of ht-21, also detecting detecting a deletion on 13q21.33 (~38.7 kb; 8 SNPs) 13q21.33 deletion



(c) Tumordb screenshot of 13q21.33 chromosomal region of ht-21

Figure 4.18: Validated focal deletion ( $\sim$ 38.7 kb) on 13q21.33, detected by OPAS and SMD but not aCGH results. Panel (a) shows the OPAS scatterplot of chromosome 13 of FL patient 21 (ht-21). The highlighted region indicates an OPAS predicted focal deletion on 13q21.33,  $\sim$ 38.7 kb, including 8 Nsp probe markers and LR $\approx$ -0.64. In addition to OPAS, SMD results also found a deletion in the same region (13q21.33), shown in panel (b) (the y-axis in (b) indicates the SMD predicted copy number and the x-axis is the relative position in the chromosome in megabase scale). This deletion encompasses only one predicted Ensemble pseudogene (ENSG00000216426). Panel (c) demonstrates the screenshot of Tumordb illustrating the FPP alignment of BAC clones in 13q21.33 to the reference human genome (hg18). This deletion was predicted by OPAS and SMD, shown by red and pink arrows, respectively (aCGH results did not report this deletion). The black and orange arrows indicate two BACs with split alignments to the reference (HTa21-0150H21, shown in green; and HTa21-0108G10, shown in orange). The non-contiguous alignment of these BACs to the reference suggests that they may capture a deletion event on chromosome 13q21.33. This deletion was validated by Illumina sequencing, as denoted by the blue arrow in (c) (40,466 bp).


(a) Schematic representation of 2 candidate somatic deletions in FL patients 4 and 29 that affect the first intron of *ERBB4* 



(b) OPAS scatterplot of chromosome 2 in patient 29

Figure 4.19: Candidate focal deletions on chromosome 2 that potentially affect the first intron of **ERBB4** gene in two FL patients (OPAS-exclusive). Panel (a) shows a schematic representation of 2 candidate somatic OPAS deletions on chromosome 2q34 in FL patients 4 and 29. Both of these OPASexclusive deletions, denoted by the red lines in (a), affect the first intron of the ERBB4 gene (although these candidate CNVs do not overlap). The candidate intronic *ERBB4* deletion in patient 4 (ht-4) is  $\sim$ 25 kb (13 SNPs; LR = -0.58) and the candidate deletion of patient 29 (ht-29) is ~4 kb (4 SNPs; LR = -0.39). None of these events has been reported in Tumordb, thus, no evidence is available to directly validate these candidate CNVs. However, there is an FPP "fragment hole" in ht-4 that aligns with the OPAS predicted deletion (~25 kb) in this patient. The OPAS candidate deletion in ht-29 (~4 kb) is also near an FPP detected "fragment hole". The blue bars in (a) illustrate FPP clones that align to the above events. The fragment holes that occur within these clones are denoted by thinner lines (see p. 283 for description of "fragment holes"). Although aligning with an FPP fragment hole does not validate these putative events, it increases the confidence of these OPAS-exclusive results. Panel (b) shows an example of *ERBB4* intronic deletion in ht-29. As indicated, this putative deletion is estimated to be  $\sim$ 4 kb and includes only 4 Nsp SNP probe markers. The significant deletion z-score of -2.8 and LR value of approximately -0.4 indicate that the above deletion in ht-29 is statically significant, however, further experiments are required to investigate this putative event.



(a) OPAS scatterplot of chromosome 3 of patient 20, denoting an OPAS-exclusive focal deletion on 3q13.33 (~97 kb; 8 SNPs)



(b) Tumordb screenshot of 3q31.33 chromosomal region of ht-20

Figure 4.20 (*previous page*): OPAS-exclusive candidate deletion on chromosome 3q13.33 of patient 20 (~97 kb) that is adjacent to an FPP inversion event. Panel (a) denotes the OPAS scatterplot of chromosome 3 of patient 20 (ht-20). The yellow highlighted region indicates two OPAS candidate deletions that are ~97 kb (CNV #1) and ~9 Mb (CNV #2), and are about 1.8 Mb apart. The larger deletion (CNV #2; ~9 Mb) is observed by other datasets in Tumordb (aCGH and SMD). However, the smaller deletion (CNV #1; ~97 kb) is an OPAS-exclusive candidate CNV that includes only 8 SNP probe markers (3q13.33; LR = -0.38). Panel (b) displays the Tumordb screenshots of chromosome 3q13.33 (top) and chromosome 3q22 (bottom) of ht-20. The FPP split alignment ("multi fpp"; see Appendix K) was detected in a single clone (20-31J21) and validated by BAC end sequencing (BES) in two clones (shown by red and blue arrows in top panel). The bottom panel shows that the other end of clone 20-31J21 (shown in green) is aligned to chromosome 3q22, as shown by red dashed lines. The FPP and BES analysis have concluded that the above FPP event in ht-20 is an inversion, approximately 11 Mb (3q13.33 and 3q22).

Top panel of (b) demonstrates that the breakpoint of the FPP inversion event in 3q13.33 partially overlaps with the OPAS-exclusive predicted deletion (CNV #1) in this region. Furthermore, another portion of CNV #1 aligns with an FPP "coverage gap" (p. 283), shown by red dashed lines. Although overlapping with these FPP events does not directly validate the OPAS-exclusive ~97 kb deletion (CNV #1), it increases the confidence of this CNV prediction. The bottom panel shows that the larger deletion (CNV #2; ~9 Mb) is detected by both SMD and OPAS, and overlaps with the FPP detected inversion in chromosome 3q22.



Figure 4.21: The most significant gene network associated with OPAS candidate focal CNVs ( $\leq 150$  kb). Ingenuity pathways analysis software was used to examine 30 genes that overlapped with candidate OPAS focal CNVs. This figure denotes the IPA result of the most significant gene network in this datasets (P $\leq$  3.28e-02). The molecules that overlapped with candidate focal deletions and amplifications were shown by red and blue colors, respectively. The green color represents a complex deletion/amplification event that affects the *KIT* gene (discussed in more detail in Section 4.3.8.3). This network includes several cancerrelated genes. For instance, *ERBB4*, is a member of growth factor receptors that is linked with cancer development and malignancy, and *PAX5* is involved in early stages of B-cell differentiation and has also been linked to B-cell and Hodgkins lymphomas (see pp. 130-131 for more detail). The IPA analysis also indicated that 3 biological function categories (cell death, cell cycle and cellular development) were significantly associated with the genes that overlapped with candidate focal OPAS CNVs.



### Comparison of Candidate Amplifications in 23/25 FL Patients

Figure 4.22: Venn diagram comparing predicted copy number amplifications in FL samples, generated by 3 methods (OPAS, SMD and aCGH). These datasets include candidate amplification results from aCGH (96; purple), 500K array results based on SMD (105; green), and 250K (Nsp) array results based on OPAS (125; orange). The pairwise concordance coefficients are shown by two arrows near the corresponding datasets. Here, the concordance percentages was defined as the proportion of predicted amplifications in one dataset (e.g., OPAS) that corresponded to candidate amplifications in another dataset (e.g., aCGH). The direction and color of these arrows correspond to the datasets being compared. For instance, the purple arrow at the top-left indicates that 85% of aCGH amplification results are also observed by SMD (also in Tumordb); and the green arrow, in the opposite direction, indicates that 78% of SMD amplification results are also observed by aCGH results. These diagrams and their corresponding concordance coefficients indicate that there is a high similarity between SMD and OPAS results (two SNP array-based datasets), with 95% and 80% concordance rates, respectively. This is also evident from the diagram since the majority of the predicted amplification calls generated by OPAS and SMD fall within the intersection area between these two datasets (SMD  $\cap$  OPAS = 100). It is also observed that 85%-89% of aCGH putative amplifications are observed by SNP array-based datasets. However, relatively, fewer OPAS and SMD amplifications are seen by aCGH.

<sup>\*</sup> due to the difference in the size of detected CNVs, a few SMD and aCGH predicted amplifications overlap with more than 1 OPAS predicted amplifications. Therefore, the intersection segment includes 86 OPAS predicted amps, 80 aCGH amps and 81 SMD amps.



Figure 4.23: Candidate OPAS-exclusive focal amplification on 9p13.2 (~76 kb; 14 SNPs) that encompasses 2 exons of *PAX5* (OPAS-exclusive). Panel (a) shows the OPAS scatterplot of chromosome 9 of patient 21, indicating a candidate OPAS-exclusive focal amplification on 9p13.2, approximately 76 kb (76,133 bp; LR = 0.33; z-score = +2.16). This putative copy number gain contains 14 SNP probe markers and potentially affects 2 exons of *PAX5* gene, as shown in panel (b). It has been speculated that *PAX5* plays an important role in B-cell differentiation and regulation of B-lymphoid-specific target gene, *CD19* [388, 405]. The estimated p-value of detecting a 76,133 bp CNV that overlaps with *PAX5* just by random chance is close to zero (P = 2.2e-308; based on bootstrap analysis using 100,000 permutations). Since there is no supporting data available in Tumordb to confirm this putative event, further experiments are required to investigate this potentially important OPAS-exclusive candidate amplification (in ht-21).



### Comparison of Candidate Deletions in 23/25 FL Patients

**Figure 4.24:** Venn diagram comparing predicted copy number deletions in FL samples, generated by 3 methods (OPAS, SMD and aCGH). The datasets include candidate deletions from aCGH (112 regions; purple), 500K array results based on SMD (99 regions; yellow), and 250K (Nsp) array results based on OPAS (132 regions; red). The arrows denote the concordance percentage of pairwise comparisons between these datasets (71% and 95% concordance between OPAS and SMD versus ~50% concordance compared to aCGH). These findings suggest an increased similarity between the two datasets from SNP array platform (SMD and OPAS), compared to results from array CGH platform. The area marked by blue dashed lines, denoted by 'F', indicates that 38 candidate deletions were observed by OPAS but not SMD. From these 38 putative deletions, 32 were not detected by aCGH either (referred to as "OPAS exclusive" candidate deletions). Similarly, the area marked by black dashed lines, denoted by 'E', indicates that 5 candidate deletions were detected by SMD but not OPAS. The candidate deletions in 'E' and 'F' regions are compared in Section 4.3.7.

The intersection with aCGH suggests that 6/38 candidate deletions in 'F' and 1/5 candidate deletions in 'E' are likely real events. From the remaining 32 regions in 'F', also referred to as OPAS-exclusive events, 4 aligned to Illumina sequence validated deletions and 11 aligned to FPP events that could represent copy number deletions. Thus, 15/32 (47%) OPAS exclusive events and 21/38 (55%) events in 'F' (seen by OPAS but not SMD) are likely real copy number deletions. There is no additional information that can be used to prove or reject the remaining 17 OPAS exclusive candidate deletions in 'F'.







(b) Screenshot of Tumordb (10q22.2 region; ht-19)

(c) Screenshot of Tumordb (10q23.32-q23.33; ht-19)

Figure 4.25: Multiple OPAS candidate deleted regions on chromosome 10 of an FL patient (ht-19), that align with FPP complex events. Panel (a) denotes the OPAS scatterplot of chromosome 10 of patient 19 (ht-19). The yellow highlighted region demonstrates 2 OPAS candidate deletions in chromosome 10q22.2. One of these candidate deletions (CNV #1) is an OPAS-exclusive finding (~226 kb, LR = -0.5) that contains 14 SNP probe markers. The second event (CNV #2) is approximately 449 kb (19 SNPs; LR = -0.58) and is also found by SMD and aCGH. In addition to these two candidate events in 10q22.2, there is another 1.4 Mb candidate deletion on 10q23.32-q23.33 (CNV #3) that is also detected by SMD and aCGH datasets. Panel (b) shows the screenshot of Tumordb, illustrating the alignment of FPP BAC clones to the reference genome (hg18). FPP analysis detected a complex event between 10q22.2-q23.32. This figure denotes an FPP event that was captured in one BAC (19-157G11; shown in green) and validated by BES in two BACs (denoted by red and blue arrows). As observed, the candidate OPAS-exclusive deletion (CNV #1) is adjacent to the FPP validated breakpoint in 10q22.2. Furthermore, the yellow highlighted area in (b) shows that the depth of FPP coverage data is particularly low in CNV #2 region. These findings suggest that CNV #2 (~449 kb) is likely a real copy number loss, although, it is not directly validated by FPP. (*continued on the proceeding page*)

(continued): Panel (c) shows that the other end of FPP BAC clone 19-157G11 (that was shown in (b)) is aligned to chromosome 10q23.32. This BES validated breakpoint is adjacent to OPAS candidate CNV #3 in chromosome 10. There are no FPP events to directly validate candidate CNV #3, which is also detected by SMD and aCGH. However, the yellow area in (c) illustrates that the FPP depth of coverage is particularly low in the predicted region of copy number loss (CNV #3).

Based on these figures, it can be speculated that the three aforementioned candidate deletions on chromosome 10 of ht-19, including the OPAS-exclusive event on 10q22.2 (CNV #1), are likely real deletions.



(a) OPAS scatterplot of chromosome 14 of patient 14, indicating a candidate deletion on 14q32.33 (~230 kb; 4 SNPs; LR = -0.99)



(b) Tumordb screenshot of 14q32.33 chromosomal region of ht-14

Figure 4.26: Candidate OPAS-exclusive deletion on chromosome 14q32.33 of patient 14, mapping to an FPP 'coverage-gap' (~230 Kb; 4 SNPs). Panel (a) denotes the OPAS scatterplot of chromosome 14 of FL patient 14 (ht-14). The yellow highlighted region denotes an OPAS candidate copy number loss, approximately 230 kb, with 4 SNP probe markers. This candidate deletion is located near the q-end of chromosome 14 (14q32.33) and demonstrates a strong loss of signal intensity (LR = -0.99; z-score = -7.2). This putative deletion is not detected by SMD or aCGH results in Tumordb. Panel (b) shows the screenshot of Tumordb, illustrating the alignment of FPP BAC clones of chromosome 14q32.33 to the reference human genome (hg18). The red arrow denotes that FPP data detected a 'coverage gap' (see p. 283), overlapping with OPAS predicted deletion in (a). Although this FPP event does not validate the aforementioned OPASexclusive candidate deletion, it increases the confidence of this finding.



(a) OPAS scatterplot of chromosome 15 of patient 21, indicating an OPAS-exclusive deletion on 15q11.2 (~1.9 Mb; 58 SNPs)



(b) Screenshot of Tumordb (chr 15q11.2; ht-21)

Figure 4.27: Candidate OPAS-exclusive deletion on 15q11.2 (ht-21), mapping to a region with several 'multi fpp' events. Panel (a) shows the OPAS scatter plot of chromosome 15 of patient 21 (ht-21). The highlighted region indicates an OPAS candidate deletion, ~1.9 Mb (58 SNP probes) on 15q12 with LR = -0.2 (z-score = -0.83). Panel (b) illustrates the screenshot of Tumordb, representing the alignment of FPP BAC clones of chromosome 15q11.2 to the reference genome (hg18). The OPAS predicted region of deletion is shown by the black bar and denoted by the pink arrow. The red arrows indicate two 'multi fpp' events (see p. 283) that have been captured by clones Hta21-0055E19 and Hta21-0019A15. Furthermore, the candidate OPAS-exclusive deletion is adjacent to an FPP translocation event between 15q11.2 and 2q37.3, that has been captured by BAC clone HTa21-0128J21 (shown by blue arrow). In summary, these FPP events suggest that the aforementioned OPAS predicted deletion (a) may encompass one or several smaller deletions. Further experiments are required to investigate this event.





(a) OPAS scatterplot of chromosome 12 of ht-6, detecting a focal deletion ( $\sim$ 143 kb) on 12q24.11 that partially affects *HVCN1* and *PPTC7* genes

(b) FISH validation of 12q24.11 deletion in ht-6



(c) Screenshot of Tumordb, illustrating FPP alignment of BAC clones in 12q24.11-24.12 chromosomal region of ht-6 to the reference human genome (hg18)

Figure 4.28 (previous page): Validated OPAS-exclusive focal deletion on chromosome 12 of patient 6 (~143 kb; 4 SNP probes), affecting 4 genes including HVCN1 and PPTC7 (confirmed by Illumina sequencing). Panel (a) illustrates the OPAS scatterplot of chromosome 12 of FL patient 6 (ht-6). The region highlighted by yellow indicates an approximately 143 kb deletion on chromosome 12q24.11 with  $LR \simeq -0.6$  that includes 4 Nsp SNP probes. This deletion was also detected by FPP analysis (shown in panel (c)) and Illumina sequencing data. However, aCGH and SMD results in Tumordb both failed to detect this deletion. Panel (b) shows the result of the FISH experiment that was performed to validate the aforementioned OPAS-exclusive focal deletion. The FISH experiment was designed and performed by Susana Ben-Neriah at Dr. Horsman's laboratory at the BCCRC. Two FISH probes were designed in this experiment, one to interrogate the 12q24.11 predicted deletion (shown by red dots) and one control in the flanking sequence (shown by green dots). The two green (control) and one red (test) signals in the depicted nucleus indicate that 12q24.11 deletion is heterozygous in this patient (ht-6). Panel (c) displays the screenshot of Tumordb illustrating FPP alignment of BAC clones in ht-6 to the reference human genome (hg18). A custom track is added to this plot to show the OPAS boundaries of the aforementioned deletion (109,501,013-109,644,386), denoted by the blue arrow. The black arrows indicate that the ends of BAC clone HTa06-0209N22 are aligned to two distinct locations on 12q24.11 ('multi fpp' event; described in Appendix K). This FPP event suggests that the region between the denoted two BAC ends, shown by pink dashed lines, may have been deleted in ht-6. An Illumina sequencing experiment confirmed this event and identified the exact deletion boundaries (109,465,967-109,611,115). As seen in this graph, the above deletion affects 4 genes including T-cell activation protein PPTC7 and voltage-gated proton channel HVCN1 that is highly expressed in immune tissues (denoted by red arrows). The possible impact of this focal deletion on these genes is further discussed in the next figure.



LSFVYKLTAVKIYIFGSVFFLH\*LGGRKKYILRNP\*IKAMFYI\*VR\*HWC

Output of 6-frame translation of the synthetic HVCN1/PPTC7 fusion

Figure 4.29: Analysis of a putative fusion between HVCN1 and PPTC7 genes in FL patient 6 as the result of a focal deletion (145,148 bp<sup>1</sup>) on 12q24.11 (OPAS-exclusive deletion that was validated by Illumina sequencing). Panel (a) shows the UCSC screenshot of chromosome 12q24.11 deletion breakpoints in patient 6 (ht-6), generated by Illumina sequencing (denoted by red dashed lines). As seen in this plot the deletion of 12q24.11 removes all but the first exon of HVCN1 and the first 3 coding exons of PPTC7, juxtaposing the HVCN1 first exon to the remaining 3 PPTC7 exons. Therefore, the above deletion can potentially create a novel fusion gene between the 5' end of HVCN1 and 3' end of PPTC7. To generate a synthetic fusion between HVCN1 and PPTC7 genes, the sequence from the first exon of HVCN1 is merged with the sequence of the remaining 3 exons of PPTC7. This synthetic fusion sequence is then translated to obtain its corresponding protein product, using an online tool (available at http://searchlauncher.bcm.tmc.edu/seq-util/Options/sixframe.html). Panel (b) shows the results of the translation of PPTC7/HVCN1 synthetic fusion. The fusion point is marked by the red arrow. As indicated, the sequence on the right of this fusion represents the protein product of the 3 remaining PPTC7 exons and the sequence on the left of this fusion point (highlighted in blue) indicates the protein product of the only remaining HVCN1 exon. It is observed that HVCN1/PPTC7 fusion appears to be in-frame. This finding suggests that the deletion on chromosome 12q24.11 in FL patient 6 could result in a translated protein product.

<sup>&</sup>lt;sup>1</sup> based on the exact boundaries of the deletion, generated by Illumina sequencing



(a) OPAS scatterplot of chromosome 12 of patient 9, indicating a deletion on 9q24 ( $\sim$ 1.4 Mb)



Figure 4.30: Candidate ~1.4 Mb deletion on 12q24 in patient 9, encompassing *HVCN1* gene. Panel (a) shows the OPAS scatterplot of chromosome 12 of patient 9 (ht-9), denoting a ~1.4 Mb candidate deletion in this patient (45 SNPs; LR = -0.39) that encompasses the *HVCN1* gene. This putative copy number loss was not detected by FPP. Panel (b) denotes the SMD result of chromosome 12 of ht-9 that also found a deletion in the aforementioned region (shown by black arrow). Panel (c) illustrates a partial screenshot of array CGH (aCGH) results of chromosome 12 of ht-9. The blue arrow indicates that aCGH analysis also identified a deletion in the same region (12q24.11-q24.12). Collectively, these plots suggest that although FPP did not detect the aforementioned candidate deletion in ht-9, it is likely a real CNV.



(a) OPAS scatterplot of chromosome 12 of patient 9, detecting *CDKN2A* deletion (~104 kb; 8 SNPs)



(b) Tumordb screenshot of 9p21.3 chromosomal region of ht-16

Figure 4.31: Sequence validated OPAS-exclusive focal deletion on 9p21.3 in FL patient 16, encompassing *CDKN2A* gene (~104 kb; 8 Nsp). Panel (a) denotes the OPAS LR scatterplot of chromosome 9 of FL patient 16 (ht-16). The yellow highlighted region indicates an OPAS-exclusive deletion, ~104 kb with 8 Nsp SNP probe markers (LR $\approx$  -0.95). Panel (b) shows a screenshot of Tumordb illustrating the FPP alignment of BAC clones in 9p21.3 region of ht-16 to the reference human genome (hg18). The FPP coverage data detects a deletion event in this region, captured by two BAC clones that are denoted by green (HTa16-0157E02) and yellow (HTa16-0092C10), respectively. As seen in (b), there is no FPP coverage between the highlighted regions of the chromosome, emphasizing that the FPP event is a deletion. This event confirms the aforementioned OPAS-exclusive deletion (9p21.3; ht-16) that encompasses the *CDKN2A* gene. The grey region in "manual affy annotations" track in (b) shows that SMD reported a copy number neutral loss-of-heterozygosity (LOH) that contains the aforementioned real deletion. In addition to FPP, this OPAS-exclusive deletion has also been validated by Illumina sequencing (Section 4.3.8.2) as well as FISH analysis (Figure 4.32b). The FISH analysis determined that the *CDKN2A* loss in patient 16 is a homozygous deletion. The latter finding is consistent with the significant loss of signal intensity (LR $\approx$  -0.95) of 9p21.3 loss.



(a) OPAS results indicating deletions of CDKN2A gene (9p21.3) in 4 FL patients

Figure 4.32: Recurrent deletion of 9p21.3 chromosomal region in 4/25 (16%) FL patients, suggesting a potentially important role of *CDKN2A* tumor suppressor in FL. OPAS results identified deletions on 9p21.3 in 4 FL patients (patients 16, 18, 22 and 24), all encompassing the *CDKN2A* gene. Panel (a) illustrates the OPAS scatterplots of chromosome 9 in the aforementioned patients (black arrows denote the predicted deletion that included *CDKN2A*). The FISH experiments, shown in panels (b)-(e), confirmed all of the above 4 predicted deletions in 9p21 and also identified whether these losses were homozygous (-/-) or heterozygous (-/+) (as shown in the yellow label at the top of each OPAS plot). Two probes were designed in each FISH experiment, one to interrogate the 9p21 predicted deletion (test) and one in the flanking 9q33 sequence (control). In plots (c)-(e), the red dots represent the probes for 9p21 and the green dots represent the control probe on 9q33. In plot (b), an opposite color combination is used (red for control and green for test probes). Nuclei showing homozygous or heterozygous deletion of 9p21 are indicated by the arrows. FISH images in (b)-(e) confirm all 4 cases of OPAS predicted 9p21 deletions in the FL dataset that include *CDKN2A* gene. As noted in these plots, the 9p21 deletion is homozygous in 3/4 FL patients, ht-16 (panel (b); LR = -0.95), ht-18 (panel (c); LR = -1.3) and ht-24 (panel (e); LR  $\simeq -1$ ) (continued on the next page).



(b) FISH validation; patient 16

(c) FISH validation; patient 18



(d) FISH validation; patient 22



(continued): The only heterozygous deletion was detected in ht-22 (panel (d); LR = -0.43). As indicated at the top of each image, the number of FL cells that carry 9p21 deletion differs among these patients, varying between 30-80% of the cells. It is important to note that SMD [259] and aCGH results in Tumordb failed to detect ~104 kb deletion in ht-16 that included 8 Nsp SNP probes, shown in panel (b). The latter OPAS-specific deletion was also investigated in more detail by Illumina sequencing in Figure 4.31.



(a) OPAS LR scatterplot of chromosome 4, indicating a complex CNV pattern in 4q12 region of patient 20



(b) Tumordb screenshot of 4q12 chromosomal region of ht-20

Figure 4.33 (previous page): Sequence validated OPAS-exclusive focal deletion on 4q12 affecting the *KIT* gene (ht-20). Panel (a) denotes the OPAS scatterplot of chromosome 4 of FL patient 20 (ht-20). The yellow highlighted region illustrates a complex structure of several CNVs in 4q12 chromosomal region of ht-20. CNV #1 is a relatively large deletion that contains an apparently amplified region (CNV #2; highlighted in red). However, within this apparently amplified region, there is a focal deletion (CNV #3), approximately 119 kb that is indicated by blue arrows. The latter focal deletion (CNV #3) that contains 13 SNP probe markers (LR = -0.38) is exclusively identified by OPAS.

Panel (b) shows the screenshot of Tumordb illustrating the FPP alignment of BAC clones in 4q12-q13 region of ht-20 to the reference human genome (hg18). As indicated by the red arrows, the ends of FPP BAC clone Hta20-0033M06 are aligned to two distinct locations on 4q12, suggesting that there may be a deletion in this region. This FPP event ('multi fpp'; see Appendix K) aligns with the aforementioned OPAS-exclusive focal deletion (CNV #3) in ht-20. This deletion is also validated by Illumina sequencing (shown in Fig. 4.1). The Illumina sequencing of this region determined that the above deletion (CNV #3) is 136,811 bp (OPAS analysis had estimated this deletion to be 119,507 bp). Further analysis of this event, described in Section 4.3.8.3, reveals that CNV #3 affects the extracellular portion of the *KIT* gene, while the adjacent amplification impacts the intracellular region of this proto-oncogene.



(a) The boundaries of the deletion on chromosome 4 of ht-20 (based on Illumina sequencing data) and its impact on the KIT gene



**Figure 4.34:** Analysis of the impact of 4q12 deletion (136,811 bp) in FL patient 20 on the *KIT* gene. Panel (a) denotes the sequence validated boundaries of a deletion on chromosome 4 of patient 20 (shown by the black bar). As observed, this deletion removes 3 exons of the *KIT* gene. The OPAS results also denote an amplification in the 3' end of this gene (SMD results detected the entire gene to be amplified). Panel (b) illustrates the structure of the *KIT* protein that includes an extracellular domain, a transmembrane domain and an intracellular domain. Importantly, the ligand binding domain of *KIT*, exons 1-3, have been previously implicated in constitutional activating mutations of *KIT*. Several mutations in *KIT* have also been linked in other cancers that are noted on this graph with blue the legends [403].

LR	$\leqslant -0.1$	$\leqslant -0.2$	$\leqslant -0.3$	$\leqslant -0.4$	$\leqslant -0.5$	$\leqslant -0.58$	$\leqslant -0.6$	$\leqslant -0.7$	$\leqslant -0.8$	$\leqslant -0.9$	≤ −1
# Losses	236	134	94	53	31	21	18	13	10	5	3
LR	$\geqslant 0.1$	$\geqslant 0.2$	≥ 0.3	≥ 0.4	≥ 0.5	≥ 0.58	≥ 0.6	≥ 0.7	$\geqslant 0.8$	≥ 0.9	≥1
# Gains	213	118	58	32	23	21	21	17	11	7	4

Distribution of OPAS regions with respect to their LR values (with more than 2 SNPs; total = 1931)

Table 4.1: Spectrum of LR deviation of all estimated DNA regions from 25 FL genomes. Various level of LR cut-offs are used to estimate the frequency of OPAS regions (with at least 2 SNPs) with increase/decrease of signal intensity. It is observed that based on the theoretical value of one copy loss  $(LR \leq \log_2 \operatorname{ratio}(1/2) = -1)$ , only 3 OPAS regions in the entire dataset would be called as putative deletions (shown in red). Similarly, if the theoretical value of one copy gain  $(LR \geq \log_2(3/2) = 0.58)$  is used, only 21 candidate regions would be considered as putative amplifications (shown in blue). Therefore, these theoretical cut-offs would result in a total of 24 candidate CNVs in 25 FL patients, less than one CNV per patient. Considering that copy number variations are one of the hallmarks of cancers, and are particularly frequent in follicular lymphoma, it can be concluded that using the theoretical LR values to call putative CNVs in this dataset would significantly underestimate the true extent of CNVs in this study.

Event Category	Desc.	Freq.	Mean Size (min-max)	Median #SNPs (min-max)	Median GC%	# Unique Genes
	Amps	42	102 Mb (37-242.6 Mb)	7162 (2010-21382)	0.38	21,122
WCA	Dels	6	115 Mb (76-191 Mb)	10610 (7437-18384)	0.38	2,825
	All	48	104 Mb (37-242.6 Mb)	7658.50	0.38	22,213
	Amps	9	20 Mb (569 kb-59.6 Mb)	1969 (33-5511)	0.41	1,653
Distal CNVs	Dels	20	9.4 Mb (21 kb-30 Mb)	523 (3-4214)	0.44	1,691
	All	29	13 Mb (21 kb-59.6 Mb)	581 (3-5511)	0.41	3,343
	Amps	82	2.15 Mb (8 kb-27.2 Mb)	36 (2-1558)	0.39	2,766
Other CNVs	Dels	127	2.13 Mb (753 bp-45 Mb)	31 (3-4469)	0.38	2,413
	All	209	2 Mb (753 bp-45 Mb)	22 (2-4469)	0.38	5,142

List of Candidate Somatic FL CNVs

**Table 4.2: Summary of candidate somatic copy number changes in the FL dataset.** This table presents a summary of all CNVs detected in 25 FL patients in this study. The specified CNV categories are explained in Sections 4.3.3 (WCA), 4.3.4 (distal CNVs) and 4.3.5 (all other CNVs). The first and second rows within each category summarize candidate amplification (Amps) and deletion (Dels) events, respectively. The third row presents the summary of all candidate CNVs, collectively (both amplifications and deletions).

Chr	p-ter	q-ter	Size (bp)	LR	z-score	Cytoband	Patient id	Found by aCGH?	Found by SMD?
1	del		7,743,477	-0.39	-2.0	1p36.33-p36.22	ht-24	✓	$\checkmark$
	del		8,517,646	-0.37	-2.2	1p36.33-p36.22	ht-12	$\checkmark$	$\checkmark$
	del		9,369,740	-0.43	-2.6	1p36.33-p36.22	ht-29	$\checkmark$	$\checkmark$
	del		3,146,252	-0.46	-1.6	1p36.33-p36.22	ht-6	$\checkmark$	$\checkmark$
	del		8,604,770	-0.12 <sup>1</sup>	-1	1p36.33-p36.32	ht-7	-	(cnnloh) <sup>2</sup>
	del		1,937,560	-0.23	-1	1p36.33-p36.32	ht-18	$\checkmark$	-
	del		12,871,761	-0.42	-1.6	1p36.33-p36.21	ht-28	$\checkmark$	$\checkmark$
	del		2,570,969	-0.13 <sup>3</sup>	-0.7	1p36.33-p36.23	ht-9	$\checkmark$	$\checkmark$
5		amp	21,854,461	+0.07 4	0.6	5q33.3-q35.3	ht-12	$\checkmark$	$\checkmark$
6	amp		17,091,068	+0.22	1	6p15.5-p15.1	ht-11	$\checkmark$	$\checkmark$
	del		78,003	-0.42	-1.5	7p25.3-p22.3	ht-20	-	-
7		del	8,318,241	-0.44	-2.4	7q36.1-q36.3	ht-20	$\checkmark$	$\checkmark$
8	del		21,309	-0.23	-1.6	8q24.12-q24.3	ht-25	-	-
	del		30,341,791	-0.39	-1	8p23.3-p12	ht-24	$\checkmark$	$\checkmark$
		amps	24,341,677	+0.21	1.3	8q24.12-q24.3	ht-25	$\checkmark$	$\checkmark$
		amps	59,649,419	+0.25	0.9	8q21.2-q24.3	ht-24	$\checkmark$	$\checkmark$
		amps	20,717,223	+0.25	1.2	8q24.13-q24.3	ht-28	$\checkmark$	$\checkmark$
9	del		5,319,718	-0.38	-1.3	9p24.3-p24.1	ht-24	$\checkmark$	$\checkmark$
13		del	13,148,807	-0.26	-0.8	13q33.1-q34	ht-8	$\checkmark$	$\checkmark$
14		amp	569,948	+0.18 5	1.1	14q23.33	ht-24	$\checkmark$	-
14		del	1,117,776	-0.40	-1.9	14q32.33	ht-21	$\checkmark$	$\checkmark$
15		del	27,827,063	-0.30	-1	15q24.1-q26.3	ht-29	$\checkmark$	$\checkmark$
17	del		2,783,451	-0.35	-1.6	17p13.3	ht-24	$\checkmark$	$\checkmark$
	del		7,962,118	-0.21	-1.3	17p13.3-p13.1	ht-10	$\checkmark$	$\checkmark$
		amp	26,045,703	+0.26	0.9	17q22-q25.3	ht-22	$\checkmark$	$\checkmark$
18		del	17,164,386	-0.28	-0.8	18q21.33-q23	ht-24	$\checkmark$	$\checkmark$
22		del	8,613,952	-0.41	-1	22q13.2-q13.33	ht-22	$\checkmark$	$\checkmark$
X	amp		2,570,969	+0.84	3.2	Xp22.23	ht-23	$\checkmark$	$\checkmark$
Х		amp	8,505,076	+0.34	1.1	Xq25-q28	ht-5	$\checkmark$	$\checkmark$
Tota	d = 29 (2)	0 deletio	ns and 9 amplif	ications)					

List of candidate distal CNVs in each FL chromosome

(1) Figure 4.14 (ht-7; del(1p36.33-p36.32))

(2) copy number neutral loss-of-heterozygosity (LOH)

(3) Figure 4.3 (ht-9; del(1p36.33-p36.23))

(4) Figure 4.13 (ht-12; del(5q33.3-q35.3))

(5) Supplementary Figure L.3 (ht-24; del(14q23.33))

**Table 4.3: Frequency of candidate distal CNVs (category 2) in each FL chromosome.** Each row indicates a candidate distal CNV in an FL patient ('patient id'). The second and third columns ('p-ter' and 'q-ter') specify whether the predicated CNV event is near the end of the short arm (p-ter) or the long arm (q-ter) of the denoted chromosome (column 1). Columns 4-6 present further information about the candidate distal CNV, including the size, LR value and z-score of the predicted event. The last two columns denote whether the OPAS candidate distal CNV was also detected by aCGH or SMD results.

Chr	Cytoband	Size (bp)	LR	ID.	Symbol	Function	Validation/ present in other datasets?
1	1p34.3	59,105	-0.35	ht-21	SFPQ	splicing factor proline/glutamine-rich	-
2	2q34 <sup>1</sup>	24,855	-0.58	ht-4	ERBB4	v-erb-a erythroblastic leukemia viral onco- gene homolog 4 (avian)	FPP "fragment hole"
2	2q34	4,127	-0.39	ht-29	ERBB4	v-erb-a erythroblastic leukemia viral onco- gene homolog 4 (avian)	<ul> <li>(SMD detects this region as "cnnloh")</li> </ul>
3	3p14.2	48,959	-0.34	ht-12	FHIT	fragile histidine triad gene	near an "FPP hole"
3	3q13.33 <sup>2</sup>	97,410	-0.38	ht-20	STXBP5L	may play a role in vesicle trafficking and ex- ocytosis (potential)	FPP "multi fpp"
4	4q11 <sup>3</sup>	119,507	-0.38	ht-20	KIT	mast/stem cell growth factor receptor Precur- sor	Sequencing
6	6p25.3	78,003	-0.42	ht-20	AL035696.1	4cDNA FLJ43763 fis, clone TESTI2048603	-
6	6p12.3- q25.1	113,423	-0.23	ht-7	CENPQ MUT C6orf141	centromere protein Q methylmalonyl CoA mutase chromosome 6 open reading frame 141	-
6	6q22.31	80,350	+0.29	ht-12	RNF217	probable E3 ubiquitin-protein ligase RNF217	_
8	8p23.3	21,309	-0.23	ht-25	ZNF596	zinc finger protein 596	_
8	8q12.1	126,232	-0.37	ht-12	LYN	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog	SMD
8	8q24.3	87,205	-0.63	ht-18	SLC45A4 DENND3	solute carrier family 45, member 4 DENN/MADD domain containing 3	SMD
9	9p21.3 <sup>4</sup>	104,434	-0.95	ht-16	MTAP CDKN2A	methylthioadenosine phosphorylase cyclin-dependent kinase inhibitor 2A (melanom	Sequencing a, p16, inhibits CDK4)
9	9p13.2 <sup>5</sup>	76,133	+0.33	ht-21	PAX5	paired box protein Pax-5 (B-cell-specific transcription factor)(BSAP)	-
12	12q24.11 <sup>6</sup>	143,373	-0.82	ht-6	HVCN1 PPP1CC PPTC7 TCTN1	voltage-gated hydrogen channel 1 protein phosphatase 1 protein phosphatase PTC7 homolog (T-cell activation protein phosphatase 2C) regulator of Hedgehog (Hh), required for both activation and inhibition of the Hh path- way in the patterning	Sequencing, FISH
13	13q21.33	133,291	-0.34	ht-23	KLHL1	kelch-like 1 (Drosophila)	-
15	15q24.1	115,472	-0.65	ht-20	CLK3	CDC-like kinase 3	SMD, FPP "fragment hole"
					ARID3B	AT rich interactive domain 3B	
17	17p12	121,102	-0.46	ht-22	ТЕКТЗ	tektin-3 (function not known)	SMD

List of candidate focal CNVs (  $\leq 150~{\rm kb}$  ) that affect at least a single gene (20 regions, 30 unique genes)

(continued)										
Chr	Cytoband	Size (bp)	LR	ID.	Symbol	Function	Validation			
19	19p13.2 <sup>1</sup>	65,280	-0.39	ht-4	ZNF559	zinc finger protein 559	"fragment hole" (com- plex event)			
					ZNF177	zinc finger protein 177				
19	19p13.2	90,114	-0.43	ht-4	ZNF440 ZNF491 ZNF441	zinc finger protein 440 zinc finger protein 491 zinc finger protein 441	(same as above)			

(1): Figure 4.19.

(2): Figure 4.20.

(3): Figure 4.33.

(4): Figure 4.31.

(5): Figure 4.23.

(6): Figure 4.28.

(7): Supplementary Figure L.6.

Table 4.4: Summary of all candidate somatic focal CNVs ( $\leq 150$  kb) that affect at least 1 gene in an FL patient. This table provides the list of 20 candidate focal CNVs that affect at least one gene in an FL patient. The first 5 columns indicate specifications of the OPAS predicted focal CNVs. The 6th column indicates the name of the gene(s) that overlap with a corresponding candidate focal CNV. This list includes 30 such unique genes that are affected by 20 candidate focal events (generated by OPAS). The 7th column ('Function') presents a brief description of the gene function, based on Ensemble, Entrez or SwissProt databases. The last column indicates whether the reported event was validated by FPP or Illumina sequencing or was seen by other Tumordb datasets, such as aCGH or SMD (the description of FPP events, such as 'fragment hole', can be found in Appendix K).

No.	Chr	Start	End	Size (bp)	LR	# SNPs	ID.	Validation	Number of Genes
1*	1	10,488,859	10,724,172	235,313	0.26	15	ht-29	_	2 (PEX14, CASZ1)
2*	1	162,032,628	162,074,256	41,628	0.21	11	ht-7	-	_
3*	4	155,025,583	155,110,310	84,727	0.16 <sup>1</sup>	16	ht-7	-	_
5*	5	104,556,024	104,581,898	25,874	0.26	6	ht-7	-	-
6	6	33,076,317	33,599,035	522,718	0.15 <sup>2</sup>	50	ht-25	aCGH	22 (e.g., HLA-DOA)
6*	6	98,729,854	101,127,689	2,397,835	0.03 <sup>3</sup>	260	ht-24	Seq.	11 (e.g., FBXL4)
7*	6	125,296,761	125,377,111	80,350	0.29	6	ht-12	-	1(RNF217)
8*	7	38,269,645	38,318,500	48,855	0.53	7	ht-16	-	T-cell receptor $\gamma$ site
9*	7	38,285,864	38,318,500	32,636	0.91	5	ht-8	-	T-cell receptor $\gamma$ site
10*	7	38,285,864	38,337,859	51,995	0.99	6	ht-29	-	T-cell receptor $\gamma$ site
11*	7	141,928,232	142,213,198	284,966	0.38	26	ht-8	-	T-cell receptor $\beta$ site
12*	7	141,945,722	142,191,578	245,856	0.25	19	ht-16	-	T-cell receptor $\beta$ site
13*	8	19,824,065	20,524,551	700,486	0.16	83	ht-29	-	4 (e.g., <i>LPL</i> )
14*	9	36,864,168	36,940,301	76,133	0.33	14	ht-21	-	$1 (PAX5)^4$
15*	13	30,565,684	30,573,673	7,989	0.42	5	ht-5	-	1 (RP11-173P16.2)
16	14	21,398,106	21,633,557	235,451	0.29	43	ht-29	aCGH	T-cell receptor $\alpha$ site
17	14	21,398,846	21,557,538	158,692	0.23	36	ht-18	aCGH	T-cell receptor $\alpha$ site
18*	14	22,038,694	22,069,902	31,208	0.75	7	ht-28	-	T-cell receptor $\alpha$ site
19	14	105,786,534	106,356,482	569,948	0.18 <sup>5</sup>	33	ht-24	aCGH	Immunoglobulin heavy chain site

All OPAS candidate amplifications that were not previously reported by 500K SNP Data

\* OPAS exclusive amplifications (not detected by aCGH or SMD results).

(1): z-score = +1.3

(2): z-score = +1.3

(3): z-score = +0.6; Supplementary Figure L.7

(4): Figure 4.23

(5): z-score = +1.08

Table 4.5: List of 19 new candidate OPAS amplifications that were not previously detected by SNP data analysis. The first 8 columns indicate the specifications of OPAS candidate amplifications that were not reported by 500K SNP array results in Tumordb (SMD). The 'Validation' column denotes whether the candidate OPAS amplification was reported by array CGH or Illumina sequencing results in Tumordb. The '-' mark in validation column means that no information was available to validate (or reject) the corresponding OPAS amplification. The last column indicates the number of genes that overlap with the candidate amplification. This table shows that from the 19 amplifications found by OPAS that were not previously detected by 500K SNP data analysis, at least 11 (58%) are real (9 overlap with known copy number changes in T-cell receptor and *IgH*, 1 is seen by both aCGH and OPAS and 1 is validated by Illumina sequencing data). The remaining 8 candidate amplifications (with no available validation data) range in size from ~7.9 kb up to 700 kb. Furthermore, 3 of these events potentially affect a single gene, such as ~76 kb putative amplification in chromosome 9 (event #14) that may affect *PAX5*. The *PAX5* gene is known to be involved in lymphomagenesis (see p. 130 for more information), suggesting that OPAS detected amplification in patient 21 may be a real CNV event that is related to FL tumorigenesis (Figure 4.23).

		(	OPAS Predicte	1 Deletion					FPP/Sequencing Anomaly		
ID.	Chr	Start	End	Size (bp)	#SNPs	LR	Start	End	Overlap Status	# Genes Affected (examples)	
ht-13	1	142,756,696	143,782,024	1,025,328	13	-0.24	120,728,444 143,056,638	143,744,590 143,236,834	CPLX-MATCH SEQ-MATCH	7	
ht-4	2	212,870,593	212,895,448	24,855	13	-0.58	212,862,613	212,878,937	FRAG-MATCH	1 (ERBB4)	
ht-20	3	122,499,369	122,596,779	97,410	8	-0.38	122,566,181	133,895,826	CPLX-MATCH <sup>1</sup>	1 (STXBP5L)	
ht-20	4	55,172,333	55,291,840	119,507	13	-0.38	55,123,271	55,260,082	SEQ-MATCH <sup>2</sup>	1 ( <i>KIT</i> )	
ht-19	4	131,663,499	131,674,302	10,803	4	-0.72	131,671,379	131,791,464	CPLX-MATCH <sup>3</sup>	0	
ht-16	9	21,884,299	21,988,733	104,434	8	-0.95	21,855,442	21,979,452	SEQ-MATCH <sup>4</sup>	2 (MTAP, CDKN2A)	
ht-19	10	75,378,912	75,604,515	225,603	14	-0.50	75,249,961 75,573,524	75,387,178 75,658,579	COV-GAP <sup>5</sup> FPP-MATCH	3 (VCL, AP3M1, ADK)	
ht-20	13	100,531,649	100,535,000	3,351	3	-1.14	100,531,649	100,534,365	FRAG-MATCH	1 (VGCNL1)	
ht-6	12	109,501,013	109,644,386	143,373	4	-0.82	109,465,967	109,611,115	SEQ-MATCH <sup>6</sup>	5 (PPTC7, HVCN1, TCTN1, PPP1CC)	
ht-14	14	105,169,391	105,399,872	230,481	4	-0.99	105,296,718	105,476,362	COV-GAP <sup>7</sup>	7 (IGHG2, IGHA1, IGHEP1, IGHV4-31, IGHG3, IGHD, IGHM)	
ht-25	14	105,169,391	105,786,534	617,143	7	-0.41	105,087,747	105,496,075	CPLX-MATCH	80	
ht-21	14	106,137,584	106,356,482	218,898	22	-0.15*	106,278,489	106,360,586	COV-GAP	38	
ht-21	15	18,427,103	20,335,459	1,908,356	58	-0.20	18,721,630	20,143,540	CPLX-MATCH <sup>8</sup>	32 (e.g., VSIG7, HERC2P3, POTEB, VSIG6)	
ht-9	16	82,677,316	88,690,776	6,013,460	488	-0.25	87,210,550	87,296,384	FRAG-MATCH	107	
ht-22	18	51,189,731	51,368,213	178,482	21	-0.27	51,245,287	51,248,206	FRAG-MATCH	1 ( <i>TCF4</i> )	
ht-4	19	9,315,173	9,380,453	65,280	6	-0.39	9,347,111	9,350,236	FRAG-MATCH	2 VZNF559, ZNF177)	
<ul> <li>(1): Figure 4.20</li> <li>(4): Figure 4.31</li> <li>(7): Figure 4.26</li> </ul>			<ul><li>(2): Figure</li><li>(5): Figure</li><li>(8): Figure</li></ul>	4.33 4.25 4.27		(3): Supplem (6): Figure 4 * z-score = -(	entary Figure I .28 ).61	2.8			

List of OPAS-exclusive deletions that overlap with FPP or Illumina sequence validated events

184

**Table 4.6:** List of OPAS-exclusive deletions that overlap with FPP or Illumina sequence validated events. The first 7 columns indicate specifications of deletions that were exclusively detected by OPAS. The next 3 columns indicate the FPP events or sequence-validated deletions that overlap with the corresponding OPAS deletion. The 10th column ('Overlap Status') indicates the type of OPAS/FPP overlap (described in the following). Each reported FPP event in this list has been separately analysed to select only cases that more likely represented copy number losses. The last column reports the number of genes that overlap with a candidate OPAS-exclusive deletion, and example(s) of such genes where applicable.

### **Description of FPP 'overlap status' (from Appendix K):**

**SEQ-MATCH:** The CNV was validated by sequencing the region.

**FPP-MATCH:** There is a clone with 'multi fpp' alignment representing the event but no sequence information was available at the time of preparing this table.

**CPLX-MATCH:** A breakpoint inside the OPAS candidate deletion was identified, but only some aspects of the predicted event was captured.

**FRAG-MATCH:** There is a 'fragment hole' that overlaps with the OPAS predicted deletion. Fragment holes are suspected small rearranged regions within the span of a BAC clone. However, these events have not been validated (at the time of preparing this table).

Gene	Cytoband	Gene Size (bp)	ID	LR	Del(-) or Amp(+)	CNV Start	CNV End	CNV Size	p-value
CDKN2A	9p21.3	26,739	ht-22	-0.43	-	20,503,135	22,158,464	1,655,329	9.60E-04
			ht-16	-0.95	-	21,884,299	21,988,733	104,434	4.00E-05
			ht-18	-1.29	-	21,884,495	22,088,574	204,079	1.00E-04
			ht-24	-1.02	-	21,802,349	22,474,016	671,667	3.30E-04
CDKN2B	9p21.3	6410	ht-18	-1.29	-	21884495	22088574	204,079	7.00E-05
			ht-22	-0.43	-	20503135	22158464	1,655,329	8.40E-04
			ht-24	-1.01	-	21802349	22474016	671,667	4.70E-04
TP53	17p13.1	11551	ht-10	-0.2	-	18901	7981019	79,62,118	1.70E-03
MDM2	12q15	39258	ht-12	0.19	+	61880	74683372	74,621,492	2.66E-02
			ht-18	0.27	+	61880	87779978	87,718,098	2.64E-02
			ht-25	0.18	+	53502682	94095312	40,592,630	1.51E-02
			ht-29	0.28	+	65526302	74259224	8,732,922	3.64E-03
HVCN1	12q24.11	35,643	ht-6	-0.82	-	109,501,013	109,644,386	143,373	3.00E-05
			ht-9	-0.39	-	108,796,615	110,200,759	1,404,144	1.60E-04
ERBB4	2q34	1,163,119	ht-4	-0.58	-	212,870,593	212,895,448	24,855	6.70E-04
			ht-29	-0.39	-	212,897,542	212,901,669	4,127	6.30E-04
KIT	4q12	82,784	ht-10	-0.29	-	52,972,432	97,996,020	45,023,588	1.80E-02
			ht-20	-0.38	-	55,172,333	55,291,840	119,507	1.00E-04
			ht-20	0.33	+	55,291,840	55,527,502	235,662	1.60E-04
HERC2P3	15q11.2	122,885	ht-21	-0.2	-	18,427,103	20,335,459	1,908,356	3.00E-05
ALK	2p23.2-	728792	ht-21	0.24	+	21305332	83451071	62,145,739	1.10E-02
	p23.1		ht-22	0.26	+	24049	242650580	242,626,531	1.12E-02
			ht-28	0.24	+	24049	84234990	84,210,941	1.13E-02
NBS1 (NBN)	8q21.3	51335	ht-23	0.24	+	86614119	104514185	17,900,066	6.10E-03
TP73	1p36.32	43231	ht-6	-0.46	-	1258710	9863480	8,604,770	2.30E-04
			ht-12	-0.37	-	1743546	4719663	2,976,117	1.60E-04
			ht-24	-0.39	-	775852	8519329	7,743,477	2.00E-04
			ht-28	-0.42	-	775852	13647613	12,871,761	1.70E-04
			ht-29	-0.43	-	775852	10145592	9,369,740	1.90E-04
WAF1/p21	6p21.31	8622	ht-14	0.27	+	119769	36754465	57,316,797	1.52E-02
BCL2	18q21.33	1426	ht-21	0.21	+	36754465	36763087	12,886,590	5.69E-03
			ht-24	-0.28	-	36754465	36763087	17,164,386	7.10E-03
BCL6	3q27.3	15711	ht-14	0.26	+	188921859	188937570	26,212,744	9.14E-03
BCL10	1p22.3	12123	ht-8	-0.3	-	85504048	85516171	526,795	3.00E-04

Partial list of some important genes that have been linked to cancer

**Table 4.7:** Partial list of some important genes that have been linked to cancer and their frequency across the **FL dataset.** The first 3 columns present the name of the candidate affected genes, their size and cytoband information. The 4th column is the FL patient id. The next 5 columns specify the FL candidate CNVs that overlap with the reported gene. The last column ('p-value') indicates the gene specific confidence limits for the respective CNVs calculated by bootstrap analysis with 100,000 permutations.

## Chapter 5

# **Conclusions and Future Directions**

### 5.1 Summary

It is now more than 50 years since Lejeune et al. discovered the first microscopic human chromosomal abnormality, trisomy 21, in Down syndrome patients (1959; [406]). In the following years, classical genetics paved the way to discovering several other whole chromosome changes (aneuploidy) that were associated with human disease, such as monosomy X in Turner syndrome [407]and XXY in Kleinfelter syndrome [408]. In 1991 Lupski et al. [409] discovered a submicroscopic gain of DNA copy number, approximately 1.5 Mb in size, on chromosome 17p12 that caused Charcot-Marie-Tooth type 1 (CMT1) disease. However, it wasn't before the completion of the Human Genome Project that scientists began to realize the extent of submicroscopic copy number variations among normal populations (2004; [18]). Upon this discovery, scientists began to speculate that submicroscopic copy number changes may account for both normal variations among humans as well as pathogenic variations [26, 410–412]. Discoveries that associated submicroscopic DNA copy number changes to susceptibility to diseases such as HIV infection, neurological disorders and leukaemia also emphasized the significance of these events in disease predisposition [18, 19, 46, 60, 236]. Thus, identification and characterization of such submicroscopic DNA copy number changes is important for both the basic understanding of complex diseases and their diagnosis.

The advent of microarray technologies initiated an era of high-resolution whole-genome analysis techniques and provided the means to investigate genome wide chromosome copy number variations on a sub-chromosomal scale [117, 413, 414]. Oligonucleotide microarrays, such as Affymetrix SNP arrays [29, 30, 135, 137, 146], have been commonly used for high-resolution genome-wide copy number detection studies. However, one of the major challenges for accurate CNV discovery using the data generated from the arrays is the high level of associated noise with microarray signal intensity outputs, particularly in oligonucleotide arrays [140, 176, 415]. More recently the advent of next-generation sequencing technologies offer the potential to detect copy number variations at the base-pair resolution. Although, the computational methods for analysing sequencing data to infer CNV regions are still in their infancy [194–198]. In fact, current methods used to analyse CNVs from sequencing data are often adapted from methods that were originally used for microarray data analysis [266, 416]. For instance, CNV-seq [416], a method for CNV detection using high-throughput sequencing data, is conceptually derived from array CGH platform.

For SNP oligonucleotide arrays, numerous methods have been developed to improve the accuracy of CNV calling [69, 150, 161, 178–183, 185, 186, 188]. However, most algorithms attempt to apply fixed parametric thresholds to reduce the number of false positive CNV calls. Most notably, often a candidate aberration is called significant only if changes of log2-ratio intensities are observed in at least a certain number of consecutive SNP probes. Therefore, such approaches would often emphasize specificity over sensitivity [141]. As a result, several recent publications have acknowledged underestimating the true frequency of small CNVs (that generally have fewer SNP probe markers relative to larger CNVs) [141, 146]. Applying more stringent CNV calling criteria seems to be the only way to avoid calling thousands of putative CNV hits, many of which turn out to be false positives [146, 259]. Thus, the effective power of finding CNVs not only depends on the resolution of the technology used but also on the computational methods that are applied to analyse the data.

The main goals of this thesis were to investigate the sources of variability that mitigate the accuracy of CNVs detected by SNP array data analysis (see page 19) and also to design and implement computational tools that use this information to improve CNV detection accuracy. The underlying assumption was that by using a non-parametric data analysis approach, we can achieve better CNV detection sensitivity while attaining a reasonable specificity. As hypothesized in the introduction of this thesis, such improvements would ultimately allow us to detect small CNVs that may contain only a few SNP probe markers. I also hypothesized that developing further visualization and analysis tools that could summarize CNV findings is necessary for interpreting and understanding these results.

In **Chapter 1**, I provided a literature review of the past and current methods that have been used to analyse copy number variations and elaborated the limitations of each technology with a particular emphasis on Affymetrix SNP arrays (Sections 1.1-1.2). This Chapter also discussed the limitations of some of the common CNV calling methods for analysis of SNP array data (Sec-

tion 1.4).

In Chapter 2, I focused on studying the reproducibility of the data from Affymetrix SNP arrays. To perform this task, I analysed raw signal intensities at the level of individual probes from 72 Affymetrix 10K replicate SNP arrays from 8 individuals and found that the SNP level data are highly reproducible (CV = 5.16% for chip variability and 6.3% for labeling variability; Sections 2.3.3-2.3.4). Next, I applied a theoretical approach to estimate the frequency of observing random oligonucleotide probes that vary by at least k-fold in their intensities between replicate measurements (Section 2.3.5). The main motivation behind the latter analysis was to investigate whether the occurrence of random noisy oligonucleotide probes could be the reason behind nonspecific variations that are commonly observed in Affymetrix SNP array data. The concept of the theoretical model that was used in this analysis was derived from similar work that had been previously performed in agriculture [417] and later in the context of biological assays [279] (Section 2.2.2). To adapt this model for analysing SNP array variability, I coupled the theoretical model with the empirical data obtained from the aforementioned reproducibly experiment and further expanded the mathematical formulations to adjust to Affymetrix SNP array data (pp. 45-48). Based on this approach, I concluded that the variation between individual PM oligonucleotide probes in the same SNP probe set is likely the main contributor to non-biological variations in SNP array readouts (Section 2.4).

The aim of **Chapter 3** was to elucidate the complexity of SNP array data by developing a novel method for CNV detection that was based on oligonucleotide probe level analysis of signal intensities. The analysis of SNP readouts in CNV regions indicated that average SNP probe set log2-ratio (LR) signal intensities varied significantly among SNPs that were located within the same (validated) CNVs (Section 3.3.1). Furthermore, I noticed a remarkable difference between the proportion of informative PM oligos across the SNPs. This observation was consistent with the conclusion of the theoretical analysis of noisy oligos that was discussed in Chapter 2 (p. 51). The nature of the observed noisy readouts was so complex that it was clear that averaging the oligos and applying a global filtering strategy would not be able to minimize the impact of noisy oligos on the predicted CNV calls (e.g., Figure 3.6).

In order to test whether the observed undesirable variation in oligonucleotide readouts was due to the choice of reference samples, I studied the distribution of non-informative oligos using several reference sets with varying number of normal reference samples (ranging from 1 to 150 samples; Section 3.3.2). This analysis revealed that, in contrast to what may be expected, a larger reference-set does not necessarily yield better CNV detection accuracy. For instance, the example provided in Chapter 3 showed that using a large reference set ( $R_{150}$ ) suppressed the magnitude of

a real pathogenic copy number loss in a patient with mental retardation (Figures 3.7-3.8). This finding implies that improving the detection sensitivity cannot be achieved by simply increasing the size of the reference samples.

Based on these results, I went on to develop and implement an algorithm for CNV detection using non-parametric approaches towards analysing oligonucleotide probe level signal intensities. The implemented algorithm, called Oligonucleotide Probe-level Analysis of Signal (OPAS), involves two major components, probe-level analysis and SNP-level analysis. The aim of probe level analysis is to analyse individual oligonucleotide probes within a SNP probe set and identify the most significant group of such oligos to generate the SNP signal (Figure 3.1).

In the next phase of the algorithm, post-processing, I first applied a non-linear LOWESS normalization technique to minimize the effect of fragment length biases on the estimated SNP readouts (Section 3.2.4.1) and then applied non-parametric Circular Binary Segmentation (CBS) [189, 242] to identify regions that likely have different copy number values (Section 3.2.4.2). It is important to note that while many of the components of this algorithm (such as CBS) had been previously developed in other applications, the adaptation of these methods on SNP microarray data required a largely novel implementation in order to accommodate a different data type.

The underlying assumption is that using the non-parametric data mining approaches that aim to improve the quality of SNP readouts would ultimately provide better CNV detection accuracy. To test the above hypothesis, I generated simulated Nsp signals (Sec. 3.2.6) with known CNV regions with varying number of probes (w = 2, 4, 8, 10, 15, 25, 100, 200) and amplitudes of signal aberration ( $\gamma = 0.11, 0.2, 0.4, 0.6, 0.8, 1.0$ ). I then applied three different algorithms to identify CNVs in these signals (OPAS, SMD and GLAD). The results from each dataset were then compared to the known simulated CNV regions for each alteration size and LR ratio response to estimate the corresponding number of true positives and false positives (Section 3.3.6). These values were then used to evaluate the sensitivity (TPR) and precision (PPV) of each algorithm, respectively (described in Appendix H). The result of this analysis indicated that while all algorithms showed similar performance for detecting CNVs that contained more than 25-100 SNP probes, OPAS had a noticeably higher sensitivity and precision in detecting CNVs with less than 10 SNP probe markers (Figure 3.15). The latter finding confirms the initial hypothesis that using a non-parametric approach results in better CNV detection accuracy for events with fewer SNP probe markers.

To test the impact of additive noise on the performance of CNV calling, I developed a biologically inspired simulated model by adding white gaussian noise with varying amplitudes to a real Nsp signal that harboured a known deletion (chromosome 14 of FL patient 17; Section 3.3.5). The exact breakpoints of this deletion were determined by sequencing. I then assessed whether the simulated signals with added noise were capable of identifying the aforementioned known deletion. Based on the total misclassification rate from 10 trials, I concluded that the CNV calling process would fail when signal-to-noise ratio is larger than one (SNR > 1; Figure 3.16). In other words, the algorithm is incapable of distinguishing between specific and non-specific variations of log2-ratio signal intensities when the signal and noise have approximately the same amplitude.

Another approach to study the impact of noise in CNV analysis is to examine the correlation between the total number of generated CNVs with respect to the overall standard deviation (or noise) of hybridization signal intensities from the corresponding arrays. I studied this correlation using OPAS-predicted CNVs in 25 lymphoma patients and the standard deviation of the Nsp arrays that were used to generate the CNV calls (Section J). This analysis indicated that the aforementioned 2 parameters (SD and number of CNV calls) were independent (p-value of Pearson's correlation  $\simeq 0.59$ ; Figure J.1)<sup>1</sup>. The latter finding suggests that the number of OPAS false positive calls do not significantly increase with the standard deviation of array hybridization intensities. This is likely due to the fact that in OPAS raw signal intensities are processed by data mining strategies that aim to identify and remove the effect of noise at the oligonucleotide probe-level prior to applying the CNV calling algorithm.

In **Chapter 4**, I focused on finding and profiling somatic CNVs in 25 follicular lymphoma (FL) patients by using data from Affymetrix 250K Nsp arrays. These samples have also been analysed using several other technologies such as FPP, BAC aCGH, karyotyping and BAC end sequencing (Sections 4.2.2-4.2.3). In additional to the results from these alternative platforms, I also compared OPAS results with CNVs generated from both Nsp and Sty arrays (500K) by another computational method (SMD) that has already been used in several other studies [30, 259, 372]. The availability of these alternative datasets provided the opportunity to compare and cross-validate OPAS predicted CNVs in FL patients.

In total, I identified 286 candidate somatic CNVs (11.4 per patient) from which 53 (18.5%) were smaller than 150 kb and 41 (14.3%) contained fewer than 10 SNP probe markers. Furthermore, from all (286) somatic events 133 (47%) were candidate amplifications and 153 (53%) were candidate deletions (Figure 4.7). To provide a comprehensive profile of somatic CNVs in the FL dataset and in order to facilitate data interpretation, candidate CNVs were subdivided into 3 groups based on their size and/or location in the chromosomes. This profiling indicated that large-scale CNVs that affected entire chromosomes (aneuploidy) or chromosome arms (WCA events; Section 4.3.3) were the most recurrent copy number alterations in the FL dataset (22 unique CNVs

<sup>&</sup>lt;sup>1</sup>A similar study by Itsara et al. [141] showed a strong CNV-SD correlation among their analysed samples and concluded that the rate of their false positives is proportional to the SD of the analysed arrays.

resulting in 48 WCA events). Examples of such events include +1q (6/25), -6q (5/25), +7 (4/25) and +X (5/25). All detected recurrent WCA events have been previously reported in follicular lymphoma by several independent studies [244–248].

The profiling also showed that 29 of the total (286) CNVs were located proximal to chromosome ends, referred to as distal CNVs in the FL dataset (Section 4.3.4). One important observation in this category was the relative enrichment of deletions compared to amplifications (20 deletions and 9 amplifications). The latter finding may reflect losses of sequences that affect the functional status of the telomeres. Even if the telomere is intact, regions proximal to chromosome ends (subtelomeric regions) are gene-rich, and thus, even small aberrations in these regions can disrupt the function of important gene(s) that may be significant in FL. For instance, 32% (8/25) of the FL patients had deletions near the short (p) end of chromosome 1 (1p36; Figure 4.15)<sup>1</sup>. This region contains multiple known or putative tumor suppressor genes such as *TP73* that has been shown to be significantly under-expressed in lymphomas and leukemias [379].

Section 4.3.6 focused on the analysis of focal candidate CNVs (CNVs  $\leq$  150 kb) that accounted for 18.5% (53/286) of all somatic CNVs in FL dataset. As explained in Section 4.3.6, after excluding amplifications that overlapped with T-cell receptor genes, there were 42 candidate focal events (34 focal deletions and 8 focal amplifications). The gene-content analysis of these events identified 30 unique genes that were affected by 20 focal events, listed in Table 4.4 (21 other focal CNVs did not overlap with any gene). Examples of the genes that were disrupted by these putative focal CNVs are *CDKN2A*, *MTAP*, *PAX5*, *ERBB4* (the only gene affected by 2 focal deletions<sup>2</sup>) and *HVCN1*.

In Section 4.3.7, I compared OPAS candidate CNVs with the results from several alternative technologies that were applied on the same FL samples, including results from BAC array CGH, BAC fingerprint profiling (FPP) and Illumina sequencing data. Additionally, I compared the CNV results from my analysis with candidate CNVs generated from 500K SNP arrays based on SMD analysis. These alternative datasets are part of Follicular Lymphoma Tumour BAC Fingerprint Database (Tumordb). This comparison revealed that 91.2% (261/286) of all OPAS predicted CNVs were seen by at least one other CNV dataset or mapped to an FPP event (Section 4.4). I also compared 32 OPAS-exclusive deletions (i.e., deletions that were detected by OPAS but not detected by other array-based results in Tumordb) with FPP and Illumina sequence-validated data. This analysis found that from these 32 exclusive deletions, 16 (50%) overlapped with Illumina sequence-validated deletions or FPP events (Figure 4.24). An important finding of this analysis

<sup>&</sup>lt;sup>1</sup>At least 7 of the 8 reported 1p36 deletions are real based on FISH analysis (see Fig. 4.15 for more detail).

<sup>&</sup>lt;sup>2</sup>These candidate focal events are non-overlapping deletions on the same intron (intron 1) of *ERBB4* (Fig. 4.19b).
was validation of 4 OPAS-exclusive deletions (4/32) based on Illumina sequencing data, including 3 focal deletions on 12q24.11 (ht-6; ~104 kb with 4 SNPs), 9p21.3 (ht-16; ~124 kb with 8 SNPs) and 4q12 (ht-20; ~143 kb with 8 SNPs).

Section 4.3.8 provided a description of 3 genes that were disrupted by the above 3 validated deletions. This section also provided a brief overview as how the focal deletions of these genes may be important in cancer (2 genes had partial deletions and 1 gene was entirely deleted). One of these sequence-validated small events was a deletion on 12q24.11 in FL patient 6 (145,148 bp based on sequencing validated breakpoints, 4 SNPs). Two genes located at the breakpoint boundaries were affected by this deletion (HVCN1 and PPTC7; Section 4.3.8.1). The voltage-gated proton channel HVCN1 gene had been previously reported to be highly expressed in immune tissues [418] and, more recently, it was shown that HVCN1 is involved in modulating the B cell antigen receptor [263]. The aforementioned sequence-validated focal deletion on 12q24.11 (OPAS-exclusive), removed all but the first exon of HVCN1 and the first 3 coding exons of PPTC7, bringing the HVCN1 single remaining exon to the 3 remaining PPTC7 exons. Therefore, it was speculated that the above deletion may create a novel fusion gene between the 5' end of HVCN1 and 3' end of *PPTC7.* A simulated fusion experiment was then performed to investigate the above hypothesis. This experiment was implemented by joining the HVCN1 and PPTC7 remaining exons and translating the resultant sequence using an online tool (Six Frame Translation of Sequence). As shown in Figure 4.29, the hypothetical HVCN1/PPTC7 fusion appeared to be in-frame. Therefore, it was concluded that the focal OPAS-exclusive deletion on 12q24.11 (in FL patient 6) could, in theory, result in a gene fusion event with a translated protein product.

Another important OPAS-exclusive finding was a focal deletion  $(124,010 \text{ bp})^1$  on chromosome 9p21.3 in patient 16 (8 SNPs) that included *CDKN2A* tumor suppressor gene (Section 4.3.8.2). Further analysis indicated that *CDKN2A* was deleted in 16% (4/25) of the patients in the FL dataset<sup>2</sup> (Figure 4.32). The observed frequent *CDKN2A* deletions emphasize the importance of tumor suppressor role of this gene in FL and highlights the significance of detecting the above OPAS-exclusive deletion.

Section 4.3.8.3 discussed the novel discovery of another small deletion  $(136,811 \text{ bp})^3$  on 4q12 chromosomal region of FL patient 20 (8 SNPs) that affects the extracellular domain of the *KIT* proto-oncogene (Figures 4.33 and 4.34). The intracellular domain of this gene was also amplified. However, the aCGH results of Tumordb did not detect either of these events and the SMD results

<sup>&</sup>lt;sup>1</sup>based on sequencing-validated breakpoints

<sup>&</sup>lt;sup>2</sup>All the reported 4 *CDKN2A* deletions were also validated by FISH (Figure 4.28).

<sup>&</sup>lt;sup>3</sup>based on sequencing-validated breakpoints

reported the entire region as one larger amplification event (Figure 4.32b). *KIT* is a proto-oncogene that encodes a receptor tyrosine kinase (RTK) and is crucial to melanogenesis, hematopoiesis and gametogenesis. Gain-of-function mutations of *KIT* have been associated with several cancers including AML, sinonasal T-cell lymphomas and GIST (Figure 4.34b). It has also been shown that the deletion of the extracellular domain of RTKs facilitates the ligand-independent activation of these genes by eliminating the negative regulatory constraints that are imposed by the extracellular domain [402]. Based on these data, I hypothesized that amplification of the kinase domain of *KIT* may act synergistically with the deletion of the ligand-binding domain (exons 1-3) and may result in the constitutive activation of this gene. This finding may represent the first report of such a mechanism of *KIT* activation in follicular lymphoma.

In conclusion, the work presented in Chapter 4 has provided a comprehensive profile of candidate somatic chromosome copy number changes in 25 FL genomes and generated a list of 286 candidate somatic CNVs in these patients<sup>1</sup>. Furthermore, this study provided further insight into the extent of small CNVs that could potentially affect important genes in follicular lymphoma (such as deletion of *CDKN2A* and partial deletions of *KIT* and *HVCN1*). Several additional tools have also been developed to facilitate CNV data visualization and comparison between samples that are important for interpretation of the CNV results, particularly for large-scale studies.

## 5.2 Significance and Contribution to Field of Study

One hypothesis of this thesis was that developing non-parametric methods based on oligonucleotide probe data analysis can improve the accuracy of identifying smaller CNVs from Affymetrix SNP arrays. This was investigated first based on simulated data in Chapter 3 and then by using real data from 25 follicular lymphoma patients in Chapter 4. In the latter analysis, in addition to known large-scale FL-related CNVs, several small somatic CNVs were discovered that were not previously detected in these samples by other array-based methods. Validation of focal deletions, such as those affecting *CDKN2A*, *KIT* and *HVCN1*, supported the initial hypothesis that using better analysis tools can result in discovering novel small events that may be important in follicular lymphoma.

It was also mentioned in Chapter 1 that the frequency of small CNVs is largely underestimated in studies that use stringent CNV selection criteria. The analysis presented in Chapter 4 showed that  $\sim$ 18.5% (53/286) of all candidate somatic CNVs in 25 follicular lymphoma samples were smaller

<sup>&</sup>lt;sup>1</sup>Here, the term CNV is used to refer to all chromosome copy number changes, regardless of their size. Therefore, the reported 286 candidate events includes both aneuploidies and deletions/amplifications of segments of the DNA.

than 150 kb and 14.3% (41/286) had fewer than 10 SNP probe markers.

One of the important results of Chapter 4 was that the small somatic CNVs in FL did not always affect the same genes. However, the pathway analysis of the affected genes indicated important cancer-related networks that may have been disrupted as the result of these events. Based on these observations, it can be speculated that CNVs in FL do not necessarily affect the same gene(s), but instead, these CNVs disrupt a variety of genes that participate in the same critical pathways.

In summary, in this thesis I described an approach for assessing the variation of microarray data and finding candidate regions of copy number aberration from Affymetrix GeneChip SNP arrays. Another aspect of this work focused on representing the results of large-scale copy number studies that are important for the interpretation of the CNV findings. It is important to note that during the past decade hundreds of projects have been conducted using Affymetrix SNP array platform for both genotyping and copy number applications and thousands of raw data files (.CEL) from these projects are now publicly available (see Table 1.3). For instance, the National Cancer Institute (NIH) has released 500K SNP array data for over 300 cancer cell lines from 30 different tissue types in a wide range of cancers, including small cell lung carcinoma, neuroblastoma, lymphoma and brain glioblastoma<sup>1</sup>. Also, Weir et al. [270] have provided 384 matched tumor/normal data sets ( $2 \times 384 = 768$  files) from lung adenocarcinoma<sup>2</sup>, which is one of the most common types of lung cancer (see Table 1.3). OPAS analysis of the samples of interest from such publicly available datasets may result in identifying novel small and/or recurrent CNVs that are below the sensitivity thresholds of other algorithms and have not been previously discovered.

Also, the work presented in this thesis was the result of generating hundreds of computer programs that aimed to facilitate data acquisition, sequence interpretation, statistical inference and visualization of the genomic data. In order to keep these tools reusable and traceable, I have developed Graphical User Interface (GUI) programs for many of these scripts, some of which are already being used at the GSC. For instance, one of the interactive visualization tools has been used by Dr. Maziar Rahmani at the GSC and the application of this tool helped to discover significant markers in calcific aortic valve stenosis [265].

<sup>&</sup>lt;sup>1</sup>GlaxoSmithKline (GSK) database, available at https://cabig.nci.nih.gov/caArray\_GSKdata/ <sup>2</sup>http://www.broadinstitute.org/cancer/pub/tsp/

## **5.3** Potential Applications and Future Directions

One area of future research will be the modification of the developed methods to facilitate their application to massively parallel DNA sequencing data. The non-parametric testing strategies and the proposed approach for clustering small number of data points with unknown behaviour based on combining fuzzy logic theory with optimization based clustering (in OPAS pre-processing phase) have the potential to be incorporated into other applications. Some of the tools that were implemented in this thesis have already been adapted for representation and analysis of the read depth (RD) data from whole-genome sequencing platforms (in collaboration with Dr. Jill Mwenifumbo).

Another potential application is coupling genome sequencing data with SNP array CNV information which may allow us to normalize the sequencing data. For instance, comparing two validated homozygous deletions (-/-) of the *CDKN2A* gene from FL patients that also had transcriptome sequencing data revealed an apparent transcriptome coverage across these homozygously deleted regions. Additionally, depth of coverage data from both genome and transcriptome show patterns of fluctuation very similar to the "wave" artifact in microarrays [195, 419]. Therefore, it is reasonable to assume that copy number information can be incorporated for background subtraction and possibly normalization of the depth of coverage data from sequencing platforms.

As large scale studies are implemented to characterize copy number variations among normal and disease populations, the next important subject is how such knowledge can influence the clinical interpretation of such findings. In the past few years, several studies revealed apparently acquired CNVs as likely mechanisms that are employed for adaptation to environmental conditions, similar to some nucleotide level changes. Even a more daunting task is to verify whether the phenotypic consequences of such CNVs vary under different environmental circumstances.

One aspect of this thesis was to develop computational tools that can help us to improve the accuracy of finding CNVs from SNP array platform and to catalogue acquired events in a population of follicular lymphoma patients. However, another major challenge that exists is the connections between such genomic observations and clinical implications, a task which is not straight forward and requires complex models that can verify the implications of such alterations.

As discussed earlier, based on the observations in Chapter 4 it can be speculated that the CNVs in FL most likely do not target the same gene(s) but, instead, they affect a variety of genes that participate in critical pathways. This implies the importance of shifting our mindset from a single gene strategy to other more sophisticated networks of interconnected factors such as gene pathways or even protein pathways. Increasing the FL sample size and combining genes that are affected by different mechanisms, such as CNVs, point mutations and expression analysis, may lead to the

discovery of important pathways or genes that are recurrently affected as the result of such alterations. An example of the significance of gene pathway analysis is a recent study in autism [148] that discovered new candidate autism pathways based on genes that were enriched in a cohort of 859 affected patients.

Despite these findings, interpretation of the predicted CNVs is a challenging task and much more work is required to translate the research findings to the clinical context. Meanwhile enhanced computational techniques are particularly significant tools that can help to provide more accurate results regarding CNVs and their impact on the affected chromosomes and genes. Thus, the main goal of large-scale efforts for CNV analysis in complex diseases is that the accumulation of such data will uncover patterns that correlate these findings to the disease state. I believe that the work presented in this thesis would facilitate identification and characterization of many novel small changes that can ultimately help to accomplish this goal.

## **Bibliography**

- [1] J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, N. P. Carter, S. W. Scherer, and C. Lee, "Copy number variation: new insights in genome diversity," *Genome Res.*, vol. 16, pp. 949–961, Aug 2006. 1
- [2] G. H. Perry, J. Tchinda, S. D. McGrath, J. Zhang, S. R. Picker, A. M. Caceres, A. J. Iafrate, C. Tyler-Smith, S. W. Scherer, E. E. Eichler, A. C. Stone, and C. Lee, "Hotspots for copy number variation in chimpanzees and humans," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 8006–8011, May 2006.
- [3] G. H. Perry, F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, A. C. Stone, M. E. Hurles, C. Tyler-Smith, E. E. Eichler, N. P. Carter, C. Lee, and R. Redon, "Copy number variation and evolution in humans and chimpanzees," *Genome Res.*, vol. 18, pp. 1698–1710, Nov 2008. 1, 17
- [4] J. Kunze, "Neurological disorders in patients with chromosomal anomalies," *Neuropedi*atrics, vol. 11, pp. 203–249, Aug 1980. 1
- [5] J. Lejeune, J. Lafourcade, R. Berger, and M. O. Rethor, "[The crying cat syndrome and its reciprocal]," Ann. Genet., vol. 8, pp. 11–15, 1965.
- [6] H. O. Sedano, R. A. Look, C. Carter, and M. M. Cohen, "B group short-arm deletion syndrome," *Birth Defects Orig. Artic. Ser.*, vol. 7, pp. 89–97, Jun 1971. 1
- [7] C. B. Bridges, "The "BAR" Gene a Duplication," *Science*, vol. 83, pp. 210–211, Feb 1936.
  1
- [8] P. A. Jacobs, A. G. Baikie, W. M. Court Brown, and J. A. Strong, "The somatic chromosomes in mongolism," *Lancet*, vol. 1, p. 710, Apr 1959. 1
- J. L. Down, "Observations on an ethnic classification of idiots (1866).," *Ment Retard*, vol. 33, pp. 54–56, Feb 1995.

- [10] C. C. Morton, L. A. Corey, W. E. Nance, and J. A. Brown, "Quinacrine mustard and nucleolar organizer region heteromorphisms in twins," *Acta Genet Med Gemellol (Roma)*, vol. 30, pp. 39–49, 1981. 2
- [11] R. S. Verma, H. Dosik, and H. A. Lubs, "Size variation polymorphisms of the short arm of human acrocentric chrosomes determined by R-banding by fluorescence using acridine orange (RFA)," *Hum. Genet.*, vol. 38, pp. 231–234, Sep 1977. 2
- [12] J. G. Bauman, J. Wiegant, P. Borst, and P. van Duijn, "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA," *Exp. Cell Res.*, vol. 128, pp. 485–490, Aug 1980. 2
- [13] A. Edwards, A. Civitello, H. A. Hammond, and C. T. Caskey, "DNA typing and genetic mapping with trimeric and tetrameric tandem repeats," *Am. J. Hum. Genet.*, vol. 49, pp. 746– 756, Oct 1991. 2
- [14] P. Y. Kwok, Q. Deng, H. Zakeri, S. L. Taylor, and D. A. Nickerson, "Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs," *Genomics*, vol. 31, pp. 123–126, Jan 1996. 2, 7
- [15] K. U. Mir and E. M. Southern, "Sequence variation in genes and genomic DNA: methods for large-scale analysis," *Annu Rev Genomics Hum Genet*, vol. 1, pp. 329–360, 2000.
- [16] P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, and P. Y. Kwok, "Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms," *Genome Res.*, vol. 8, pp. 748–754, Jul 1998.
- [17] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander, "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome," *Science*, vol. 280, pp. 1077–1082, May 1998. 2, 7, 8
- [18] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nat. Genet.*, vol. 36, pp. 949–951, Sep 2004. 2, 19, 85, 187
- [19] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Mnr, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler, "Large-scale copy number polymorphism in the human genome," *Science*, vol. 305, pp. 525–528, Jul 2004. 32, 34, 85, 187

- [20] D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard, "A high-resolution survey of deletion polymorphism in the human genome," *Nat. Genet.*, vol. 38, pp. 75–81, Jan 2006. 2, 19
- [21] D. P. Locke, A. J. Sharp, S. A. McCarroll, S. D. McGrath, T. L. Newman, Z. Cheng, S. Schwartz, D. G. Albertson, D. Pinkel, D. M. Altshuler, and E. E. Eichler, "Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome," *Am. J. Hum. Genet.*, vol. 79, pp. 275–290, Aug 2006.
- [22] S. A. McCarroll, T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly, and D. M. Altshuler, "Common deletion polymorphisms in the human genome," *Nat. Genet.*, vol. 38, pp. 86–92, Jan 2006. 19
- [23] D. A. Hinds, A. P. Kloek, M. Jen, X. Chen, and K. A. Frazer, "Common deletions and SNPs are in linkage disequilibrium in the human genome," *Nat. Genet.*, vol. 38, pp. 82–85, Jan 2006.
- [24] E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler, "Fine-scale structural variation of the human genome," *Nat. Genet.*, vol. 37, pp. 727–732, Jul 2005. 10, 19, 85
- [25] A. J. Sharp, D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel, and E. E. Eichler, "Segmental duplications and copy-number variation in the human genome," *Am. J. Hum. Genet.*, vol. 77, pp. 78–88, Jul 2005. 2, 19, 85
- [26] C. Shaw-Smith, R. Redon, L. Rickman, M. Rio, L. Willatt, H. Fiegler, H. Firth, D. Sanlaville, R. Winter, L. Colleaux, M. Bobrow, and N. P. Carter, "Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features," *J. Med. Genet.*, vol. 41, pp. 241–248, Apr 2004. 2, 6, 18, 187
- [27] D. J. McMullan, M. Bonin, J. Y. Hehir-Kwa, B. B. de Vries, A. Dufke, E. Rattenberry, M. Steehouwer, L. Moruz, R. Pfundt, N. de Leeuw, A. Riess, O. Altug-Teber, H. Enders, S. Singer, U. Grasshoff, M. Walter, J. M. Walker, C. V. Lamb, E. V. Davison, L. Brueton, O. Riess, and J. A. Veltman, "Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: A multicenter study," *Hum. Mutat.*, Mar 2009.
- [28] L. Edelmann and K. Hirschhorn, "Clinical utility of array CGH for the detection of chromosomal imbalances associated with mental retardation and multiple congenital anomalies," *Ann. N. Y. Acad. Sci.*, vol. 1151, pp. 157–166, Jan 2009.
- [29] J. M. Friedman, A. Baross, A. D. Delaney, A. Ally, L. Arbour, L. Armstrong, J. Asano, D. K. Bailey, S. Barber, P. Birch, M. Brown-John, M. Cao, S. Chan, D. L. Charest, N. Farnoud,

N. Fernandes, S. Flibotte, A. Go, W. T. Gibson, R. A. Holt, S. J. Jones, G. C. Kennedy, M. Krzywinski, S. Langlois, H. I. Li, B. C. McGillivray, T. Nayar, T. J. Pugh, E. Rajcan-Separovic, J. E. Schein, A. Schnerch, A. Siddiqui, M. I. Van Allen, G. Wilson, S. L. Yong, F. Zahir, P. Eydoux, and M. A. Marra, "Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation," *Am. J. Hum. Genet.*, vol. 79, pp. 500–513, Sep 2006. 8, 13, 22, 23, 68, 69, 78, 79, 80, 82, 101, 117, 118, 187

- [30] J. Friedman, S. Adam, L. Arbour, L. Armstrong, A. Baross, P. Birch, C. Boerkoel, S. Chan, D. Chai, A. D. Delaney, S. Flibotte, W. T. Gibson, S. Langlois, E. Lemyre, H. I. Li, P. MacLeod, J. Mathers, J. L. Michaud, B. C. McGillivray, M. S. Patel, H. Qian, G. A. Rouleau, M. I. Van Allen, S. L. Yong, F. R. Zahir, P. Eydoux, and M. A. Marra, "Detection of pathogenic copy number variants in children with idiopathic intellectual disability using 500 K SNP array genomic hybridization," *BMC Genomics*, vol. 10, p. 526, 2009. 2, 8, 13, 22, 23, 68, 69, 84, 101, 117, 118, 187, 191, 280
- [31] A. Noor, P. J. Gianakopoulos, B. Fernandez, C. R. Marshall, P. Szatmari, W. Roberts, S. W. Scherer, and J. B. Vincent, "Copy number variation analysis and sequencing of the X-linked mental retardation gene TSPAN7/TM4SF2 in patients with autism spectrum disorder," *Psychiatr. Genet.*, Mar 2009. 2
- [32] L. A. Weiss, Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu, and M. J. Daly, "Association between microdeletion and microduplication at 16p11.2 and autism," *N. Engl. J. Med.*, vol. 358, pp. 667–675, Feb 2008. 2, 18
- [33] G. Hodgson, J. H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D. G. Albertson, D. Pinkel, C. Collins, D. Hanahan, and J. W. Gray, "Genome scanning with array cgh delineates regional alterations in mouse islet carcinomas.," *Nat Genet*, vol. 29, pp. 459–464, December 2001. 2, 14, 18
- [34] M. M. Weiss, A. M. Snijders, E. J. Kuipers, B. Ylstra, D. Pinkel, S. G. Meuwissen, P. J. van Diest, D. G. Albertson, and G. A. Meijer, "Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization," *J. Pathol.*, vol. 200, pp. 320–326, Jul 2003. 13
- [35] A. M. Snijders, M. E. Nowee, J. Fridlyand, J. M. Piek, J. C. Dorsman, A. N. Jain, D. Pinkel, P. J. van Diest, R. H. Verheijen, and D. G. Albertson, "Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma," *Oncogene*, vol. 22, pp. 4281–4286, Jul 2003. 14
- [36] D. G. Albertson, B. Ylstra, R. Segraves, C. Collins, S. H. Dairkee, D. Kowbel, W. L. Kuo, J. W. Gray, and D. Pinkel, "Quantitative mapping of amplicon structure by array CGH iden-

tifies CYP24 as a candidate oncogene," *Nat. Genet.*, vol. 25, pp. 144–146, Jun 2000. 2, 6, 18

- [37] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," Annu. Rev. Med., vol. 61, pp. 437–455, 2010. 2, 85
- [38] J. R. Lupski and P. Stankiewicz, "Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes," *PLoS Genet.*, vol. 1, p. e49, Dec 2005. 2
- [39] S. B. Inoue, M. C. Siomi, and H. Siomi, "Molecular mechanisms of fragile X syndrome," J. Med. Invest., vol. 47, pp. 101–107, Aug 2000. 2
- [40] E. T. Dermitzakis and B. E. Stranger, "Genetic variation in human gene expression," *Mamm. Genome*, vol. 17, pp. 503–508, Jun 2006. 2
- [41] A. Reymond, C. N. Henrichsen, L. Harewood, and G. Merla, "Side effects of genome structural changes," *Curr. Opin. Genet. Dev.*, vol. 17, pp. 381–386, Oct 2007.
- [42] C. N. Henrichsen, E. Chaignat, and A. Reymond, "Copy number variants, diseases and gene expression," *Hum. Mol. Genet.*, vol. 18, pp. 1–8, Apr 2009.
- [43] M. L. Coll, "Noncoding DNA copy number variation links to gene expression in stem cells," *Nature Reports Stem Cells*, March 2009. 2
- [44] A. Rovelet-Lecrux, D. Hannequin, G. Raux, N. Le Meur, A. Laquerriere, A. Vital, C. Dumanchin, S. Feuillette, A. Brice, M. Vercelletto, F. Dubas, T. Frebourg, and D. Campion, "APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy," *Nat. Genet.*, vol. 38, pp. 24–26, Jan 2006. 2
- [45] K. Sleegers, N. Brouwers, I. Gijselinck, J. Theuns, D. Goossens, J. Wauters, J. Del-Favero, M. Cruts, C. M. van Duijn, and C. Van Broeckhoven, "APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy," *Brain*, vol. 129, pp. 2977–2983, Nov 2006.
- [46] M. Fanciulli, P. J. Norsworthy, E. Petretto, R. Dong, L. Harper, L. Kamesh, J. M. Heward, S. C. Gough, A. de Smith, A. I. Blakemore, P. Froguel, C. J. Owen, S. H. Pearce, L. Teixeira, L. Guillevin, D. S. Graham, C. D. Pusey, H. T. Cook, T. J. Vyse, and T. J. Aitman, "FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity," *Nat. Genet.*, vol. 39, pp. 721–723, Jun 2007. 2, 32, 34, 187
- [47] K. Fellermann, D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, C. L. Bevins, W. Reinisch, A. Teml, M. Schwab, P. Lichter, B. Radlwimmer, and E. F. Stange, "A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon," *Am. J. Hum. Genet.*, vol. 79, pp. 439–448, Sep 2006. 2, 32, 34

- [48] R. W. Bentley, J. Pearson, R. B. Gearry, M. L. Barclay, C. McKinney, T. R. Merriman, and R. L. Roberts, "Association of higher DEFB4 genomic copy number with Crohn's disease," *Am. J. Gastroenterol.*, vol. 105, pp. 354–359, Feb 2010. 2
- [49] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, pp. 445–449, Apr 2007. 2
- [50] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari, and S. W. Scherer, "Structural variation of chromosomes in autism spectrum disorder," *Am. J. Hum. Genet.*, vol. 82, pp. 477–488, Feb 2008. 2, 29
- [51] E. J. Hollox, U. Huffmeier, P. L. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P. C. van de Kerkhof, H. Traupe, G. de Jongh, M. den Heijer, A. Reis, J. A. Armour, and J. Schalkwijk, "Psoriasis is associated with increased beta-defensin genomic copy number," *Nat. Genet.*, vol. 40, pp. 23–25, Jan 2008. 2
- [52] J. Simon-Sanchez, S. Scholz, M. d. e. l. M. Matarin, H. C. Fung, D. Hernandez, J. R. Gibbs, A. Britton, J. Hardy, and A. Singleton, "Genomewide SNP assay reveals mutations underlying Parkinson disease," *Hum. Mutat.*, vol. 29, pp. 315–322, Feb 2008. 2, 18
- [53] D. W. Miller, S. M. Hague, J. Clarimon, M. Baptista, K. Gwinn-Hardy, M. R. Cookson, and A. B. Singleton, "Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication," *Neurology*, vol. 62, pp. 1835–1838, May 2004.
- [54] A. B. Singleton, M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, R. Nussbaum, S. Lincoln, A. Crawley, M. Hanson, D. Maraganore, C. Adler, M. R. Cookson, M. Muenter, M. Baptista, D. Miller, J. Blancato, J. Hardy, and K. Gwinn-Hardy, "alpha-Synuclein locus triplication causes Parkinson's disease," *Science*, vol. 302, p. 841, Oct 2003. 2
- [55] S. E. McCarthy, V. Makarov, G. Kirov, A. M. Addington, J. McClellan, S. Yoon, D. O. Perkins, D. E. Dickel, M. Kusenda, O. Krastoshevsky, V. Krause, R. A. Kumar, D. Grozeva, D. Malhotra, T. Walsh, E. H. Zackai, P. Kaplan, J. Ganesh, I. D. Krantz, N. B. Spinner, P. Roccanova, A. Bhandari, K. Pavon, B. Lakshmi, A. Leotta, J. Kendall, Y. H. Lee, V. Vacic, S. Gary, L. M. Iakoucheva, T. J. Crow, S. L. Christian, J. A. Lieberman, T. S.

Stroup, T. Lehtimaki, K. Puura, C. Haldeman-Englert, J. Pearl, M. Goodell, V. L. Willour, P. Derosse, J. Steele, L. Kassem, J. Wolff, N. Chitkara, F. J. McMahon, A. K. Malhotra, J. B. Potash, T. G. Schulze, M. M. Nothen, S. Cichon, M. Rietschel, E. Leibenluft, V. Kustanovich, C. M. Lajonchere, J. S. Sutcliffe, D. Skuse, M. Gill, L. Gallagher, N. R. Mendell, N. Craddock, M. J. Owen, M. C. O'Donovan, T. H. Shaikh, E. Susser, L. E. Delisi, P. F. Sullivan, C. K. Deutsch, J. Rapoport, D. L. Levy, M. C. King, and J. Sebat, "Microduplications of 16p11.2 are associated with schizophrenia," *Nat. Genet.*, vol. 41, pp. 1223–1227, Nov 2009. 2

- [56] T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M. C. King, and J., "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia," *Science*, vol. 320, pp. 539–543, Apr 2008.
- [57] J. L. Stone, M. C. O'Donovan, H. Gurling, G. K. Kirov, D. H. Blackwood, A. Corvin, N. J. Craddock, M. Gill, C. M. Hultman, P. Lichtenstein, A. McQuillin, C. N. Pato, D. M. Ruderfer, M. J. Owen, D. St Clair, P. F. Sullivan, P. Sklar, S. M. Purcell, J. L. Stone, D. M. Ruderfer, J. Korn, G. K. Kirov, S. Macgregor, A. McQuillin, D. W. Morris, C. T. O'Dushlaine, M. J. Daly, P. M. Visscher, P. A. Holmans, M. C. O'Donovan, P. F. Sullivan, P. Sklar, S. M. Purcell, H. Gurling, A. Corvin, D. H. Blackwood, N. J. Craddock, M. Gill, C. M. Hultman, G. K. Kirov, P. Lichtenstein, A. McQuillin, M. C. O'Donovan, M. J. Owen, C. N. Pato, S. M. Purcell, E. M. Scolnick, D. St Clair, J. L. Stone, P. F. Sullivan, P. Sklar, M. C. O'Donovan, G. K. Kirov, N. J. Craddock, P. A. Holmans, N. M. Williams, L. Georgieva, I. Nikolov, N. Norton, H. Williams, D. Toncheva, V. Milanova, M. J. Owen, C. M. Hultman, P. Lichtenstein, E. F. Thelander, P. Sullivan, D. W. Morris, C. T. O'Dushlaine, E. Kenny, J. L. Waddington, M. Gill, A. Corvin, A. McQuillin, K. Choudhury, S. Datta, J. Pimm, S. Thirumalai, V. Puri, R. Krasucki, J. Lawrence, D. Quested, N. Bass, D. Curtis, H. Gurling, C. Crombie, G. Fraser, S. L. Kwan, N. Walker, D. St Clair, D. H. Blackwood, W. J. Muir, K. A. McGhee, B. Pickard, P. Malloy, A. W. Maclean, M. Van Beck, P. M. Visscher, S. Macgregor, M. T. Pato, H. Medeiros, F. Middleton, C. Carvalho, C. Morley, A. Fanous, D. Conti, J. A. Knowles, C. P. Ferreira, A. Macedo, M. H. Azevedo, C. N. Pato, J. L. Stone, D. M. Ruderfer, J. Korn, S. A. McCarroll, M. Daly, S. M. Purcell, P. Sklar, S. M. Purcell, J. L. Stone, K. Chambert, D. M. Ruderfer, J. Korn, S. A. McCarroll, C. Gates, M. J. Daly, E. M. Scolnick, and P. Sklar, "Rare chromosomal deletions and duplications increase risk of schizophrenia," Nature, vol. 455, pp. 237-241, Sep 2008. 66
- [58] H. Stefansson, D. Rujescu, S. Cichon, O. P. Pietilinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson,

T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. J. Mller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Mhleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nthen, L. Peltonen, D. A. Collier, D. St Clair, K. Stefansson, R. S. Kahn, D. H. Linszen, J. van Os, D. Wiersma, R. Bruggeman, W. Cahn, L. de Haan, L. Krabbendam, and I. Myin-Germeys, "Large recurrent microdeletions associated with schizophrenia," *Nature*, vol. 455, pp. 232–236, Sep 2008.

- [59] B. Xu, J. L. Roos, S. Levy, E. J. van Rensburg, J. A. Gogos, and M. Karayiorgou, "Strong association of de novo copy number mutations with sporadic schizophrenia," *Nat. Genet.*, vol. 40, pp. 880–885, Jul 2008. 2
- [60] T. J. Aitman, R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A. J. Marshall, E. Petretto, M. D. Hodges, G. Bhangal, S. G. Patel, K. Sheehan-Rooney, M. Duda, P. R. Cook, D. J. Evans, J. Domin, J. Flint, J. J. Boyle, C. D. Pusey, and H. T. Cook, "Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans," *Nature*, vol. 439, pp. 851–855, Feb 2006. 2, 187
- [61] S. J. Diskin, C. Hou, J. T. Glessner, E. F. Attiyeh, M. Laudenslager, K. Bosse, K. Cole, Y. P. Mosse, A. Wood, J. E. Lynch, K. Pecor, M. Diamond, C. Winter, K. Wang, C. Kim, E. A. Geiger, P. W. McGrady, A. I. Blakemore, W. B. London, T. H. Shaikh, J. Bradfield, S. F. Grant, H. Li, M. Devoto, E. R. Rappaport, H. Hakonarson, and J. M. Maris, "Copy number variation at 1q21.1 associated with neuroblastoma," *Nature*, vol. 459, pp. 987–991, Jun 2009. 2
- [62] I. Rudan, "New technologies provide insights into genetic basis of psychiatric disorders and explain their co-morbidity," *Psychiatr Danub*, vol. 22, pp. 190–192, Jun 2010.
- [63] M. Shinawi, C. P. Schaaf, S. S. Bhatt, Z. Xia, A. Patel, S. W. Cheung, B. Lanpher, S. Nagl, H. S. Herding, C. Nevinny-Stickel, L. L. Immken, G. S. Patel, J. R. German, A. L. Beaudet, and P. Stankiewicz, "A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes," *Nat. Genet.*, vol. 41, pp. 1269–1271, Dec 2009.
- [64] W. Bi, T. Sapir, O. A. Shchelochkov, F. Zhang, M. A. Withers, J. V. Hunter, T. Levy, V. Shinder, D. A. Peiffer, K. L. Gunderson, M. M. Nezarati, V. A. Shotts, S. S. Amato, S. K. Savage, D. J. Harris, D. L. Day-Salvatore, M. Horner, X. Y. Lu, T. Sahoo, Y. Yanagawa, A. L. Beaudet, S. W. Cheung, S. Martinez, J. R. Lupski, and O. Reiner, "Increased LIS1 expression affects human and mouse brain development," *Nat. Genet.*, vol. 41, pp. 168–177, Feb 2009.

- [65] Y. S. Fan, P. Jayakar, H. Zhu, D. Barbouth, S. Sacharow, A. Morales, V. Carver, P. Benke, P. Mundy, and L. J. Elsas, "Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization," *Hum. Mutat.*, vol. 28, pp. 1124–1132, Nov 2007. 2
- [66] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tuzun, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, and E. E. Eichler, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, pp. 56–64, May 2008. 2, 19, 32, 34
- [67] J. S. Beckmann, X. Estivill, and S. E. Antonarakis, "Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability," *Nat. Rev. Genet.*, vol. 8, pp. 639–646, Aug 2007.
- [68] L. V. Wain, J. A. Armour, and M. D. Tobin, "Genomic copy number variation, human health, and disease," *Lancet*, vol. 374, pp. 340–350, Jul 2009. 2
- [69] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, D. T. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. Macdonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, November 2006. 2, 8, 9, 12, 13, 19, 32, 34, 85, 188
- [70] S. A. McCarroll and D. M. Altshuler, "Copy-number variation and association studies of human disease," *Nat. Genet.*, vol. 39, pp. 37–42, Jul 2007.
- [71] J. Y. Hehir-Kwa, M. Egmont-Petersen, I. M. Janssen, D. Smeets, A. G. van Kessel, and J. A. Veltman, "Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis," *DNA Res.*, vol. 14, pp. 1–11, Feb 2007. 2
- [72] L. G. Shaffer, M. L. Slovak, and L. J. Campbell, ISCN 2009: An international system for human cytogenetic nomenclature. S. Karger Publishers, USA, 2009. 3

- [73] C. Desmaze, C. Alberti, L. Martins, G. Pottier, C. N. Sprung, J. P. Murnane, and L. Sabatier,
  "The influence of interstitial telomeric sequences on chromosome instability in human cells," *Cytogenet. Cell Genet.*, vol. 86, pp. 288–295, 1999. 4
- [74] L. Xiao, H. Y. Zhou, Z. C. Luo, and J. Liu, "Telomeric associations of chromosomes in patients with esophageal squamous cell carcinomas," *World J. Gastroenterol.*, vol. 4, pp. 231– 233, Jun 1998.
- [75] J. F. Mattei, M. G. Mattei, M. A. Baeteman, and F. Giraud, "Trisomy 21 for the region 21q223: identification by high-resolution R-banding patterns," *Hum. Genet.*, vol. 56, pp. 409–411, 1981. 4
- [76] C. Lee, W. Rens, and F. Yang, "Multicolor Fluorescence In Situ Hybridization (FISH) approaches for simultaneous analysis of the entire human genome," *Curr Protoc Hum Genet*, vol. Chapter 4, p. Unit4.9, May 2001. 4
- [77] E. Schrck, T. Veldman, H. Padilla-Nash, Y. Ning, J. Spurbeck, S. Jalal, L. G. Shaffer, P. Papenhausen, C. Kozma, M. C. Phelan, E. Kjeldsen, S. A. Schonberg, P. O'Brien, L. Biesecker, S. du Manoir, and T. Ried, "Spectral karyotyping refines cytogenetic diagnostics of constitutional chromosomal abnormalities," *Hum. Genet.*, vol. 101, pp. 255–262, Dec 1997. 4
- [78] A. Adeyinka, S. Kytola, F. Mertens, N. Pandis, and C. Larsson, "Spectral karyotyping and chromosome banding studies of primary breast carcinomas and their lymph node metastases," *Int. J. Mol. Med.*, vol. 5, pp. 235–240, Mar 2000. 4
- [79] M. B. Watson, H. Bahia, J. N. Ashman, H. K. Berrieman, P. Drew, M. J. Lind, J. Greenman, and L. Cawkwell, "Chromosomal alterations in breast cancer revealed by multicolour fluorescence in situ hybridization," *Int. J. Oncol.*, vol. 25, pp. 277–283, Aug 2004. 4
- [80] W. M. Abdel-Rahman, K. Katsura, W. Rens, P. A. Gorman, D. Sheer, D. Bicknell, W. F. Bodmer, M. J. Arends, A. H. Wyllie, and P. A. Edwards, "Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 2538–2543, Feb 2001. 4
- [81] R. Melcher, S. Koehler, C. Steinlein, M. Schmid, C. R. Mueller, H. Luehrs, T. Menzel, W. Scheppach, H. Moerk, M. Scheurlen, J. Koehrle, and O. Al-Taie, "Spectral karyotype analysis of colon cancer cell lines of the tumor suppressor and mutator pathway," *Cytogenet. Genome Res.*, vol. 98, pp. 22–28, 2002. 4
- [82] H. M. Padilla-Nash, W. G. Nash, G. M. Padilla, K. M. Roberson, C. N. Robertson, M. Macville, E. Schrock, and T. Ried, "Molecular cytogenetic analysis of the bladder carcinoma cell line BK-10 by spectral karyotyping," *Genes Chromosomes Cancer*, vol. 25, pp. 53–59, May 1999. 4

- [83] M. I. Stamouli, A. D. Panani, A. D. Ferti, C. Petraki, R. T. Oliver, S. A. Raptis, and B. D. Young, "Detection of genetic alterations in primary bladder carcinoma with dual-color and multiplex fluorescence in situ hybridization," *Cancer Genet. Cytogenet.*, vol. 149, pp. 107–113, Mar 2004. 4
- [84] J. C. Strefford, D. M. Lillington, M. Steggall, T. M. Lane, A. M. Nouri, B. D. Young, and R. T. Oliver, "Novel chromosome findings in bladder cancer cell lines detected with multiplex fluorescence in situ hybridization," *Cancer Genet. Cytogenet.*, vol. 135, pp. 139– 146, Jun 2002. 4
- [85] J. N. Ashman, J. Brigham, M. E. Cowen, H. Bahia, J. Greenman, M. Lind, and L. Cawkwell, "Chromosomal alterations in small cell lung cancer revealed by multicolour fluorescence in situ hybridization," *Int. J. Cancer*, vol. 102, pp. 230–236, Nov 2002. 4
- [86] M. Grigorova, R. C. Lyman, C. Caldas, and P. A. Edwards, "Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series analyzed with spectral karyotyping," *Cancer Genet. Cytogenet.*, vol. 162, pp. 1–9, Oct 2005. 4
- [87] T. Knutsen, V. Gobu, R. Knaus, H. Padilla-Nash, M. Augustus, R. L. Strausberg, I. R. Kirsch, K. Sirotkin, and T. Ried, "CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence," *Genes Chromosomes Cancer*, vol. 44, pp. 52–64, Sep 2005. 4
- [88] B. R. Haddad, E. Schrck, J. Meck, J. Cowan, H. Young, M. A. Ferguson-Smith, S. du Manoir, and T. Ried, "Identification of de novo chromosomal markers and derivatives by spectral karyotyping," *Hum. Genet.*, vol. 103, pp. 619–625, Nov 1998. 4
- [89] B. Stark, M. Jeison, I. Bar-Am, L. Glaser-Gabay, J. Mardoukh, D. Luria, M. Feinmesser, Y. Goshen, J. Stein, A. Abramov, R. Zaizov, and I. Yaniv, "Distinct cytogenetic pathways of advanced-stage neuroblastoma tumors, detected by spectral karyotyping," *Genes Chromosomes Cancer*, vol. 34, pp. 313–324, Jul 2002. 4
- [90] T. Veldman, C. Vignon, E. Schrck, J. D. Rowley, and T. Ried, "Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping," *Nat. Genet.*, vol. 15, pp. 406–410, Apr 1997. 4
- [91] E. Schrck, S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, D. H. Ledbetter, I. Bar-Am, D. Soenksen, Y. Garini, and T. Ried, "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, pp. 494–497, Jul 1996. 4
- [92] Y. S. Fan, V. M. Siu, J. H. Jung, and J. Xu, "Sensitivity of multiple color spectral karyotyping in detecting small interchromosomal rearrangements," *Genet. Test.*, vol. 4, pp. 9–14, 2000.

- [93] S. M. Jalal and M. E. Law, "Utility of multicolor fluorescent in situ hybridization in clinical cytogenetics," *Genet. Med.*, vol. 1, pp. 181–186, 1999. 4
- [94] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.," *Science*, vol. 258, pp. 818–821, October 1992. 5
- [95] D. Wells and B. Levy, "Cytogenetics in reproductive medicine: the contribution of comparative genomic hybridization (CGH)," *Bioessays*, vol. 25, pp. 289–300, Mar 2003.
- [96] B. Levy, T. M. Dunn, S. Kaffe, N. Kardon, and K. Hirschhorn, "Clinical applications of comparative genomic hybridization," *Genet. Med.*, vol. 1, pp. 4–12, 1998.
- [97] J. Xu and Z. Chen, "Advances in molecular cytogenetics for the evaluation of mental retardation," Am J Med Genet C Semin Med Genet, vol. 117C, pp. 15–24, Feb 2003. 5
- [98] M. Kirchhoff, H. Rose, J. Maahr, T. Gerdes, M. Bugge, N. Tommerup, Z. Tumer, J. Lespinasse, P. K. Jensen, J. Wirth, and C. Lundsteen, "High resolution comparative genomic hybridisation analysis reveals imbalances in dyschromosomal patients with normal or apparently balanced conventional karyotypes," *Eur. J. Hum. Genet.*, vol. 8, pp. 661–668, Sep 2000.
- [99] J. K. Inlow and L. L. Restifo, "Molecular and comparative genetics of mental retardation," *Genetics*, vol. 166, pp. 835–881, Feb 2004. 5
- [100] A. Gutenberg, J. S. Gerdes, K. Jung, B. Sander, B. Gunawan, H. C. Bock, T. Liersch, W. Bruck, V. Rohde, and L. Fuzesi, "High chromosomal instability in brain metastases of colorectal carcinoma," *Cancer Genet. Cytogenet.*, vol. 198, pp. 47–51, Apr 2010. 5
- [101] S. C. Santos, I. J. Cavalli, E. M. Ribeiro, C. A. Urban, R. S. Lima, L. F. Bleggi-Torres, J. D. Rone, B. R. Haddad, and L. R. Cavalli, "Patterns of DNA copy number changes in sentinel lymph node breast cancer metastases," *Cytogenet. Genome Res.*, vol. 122, pp. 16–21, 2008.
- [102] I. Salaverria, S. Bea, A. Lopez-Guillermo, V. Lespinet, M. Pinyol, B. Burkhardt, L. Lamant, A. Zettl, D. Horsman, R. Gascoyne, G. Ott, R. Siebert, G. Delsol, and E. Campo, "Genomic profiling reveals different genetic aberrations in systemic ALK-positive and ALK-negative anaplastic large cell lymphomas," *Br. J. Haematol.*, vol. 140, pp. 516–526, Mar 2008.
- [103] H. Zitzelsberger, L. Lehmann, M. Werner, and M. Bauchinger, "Comparative genomic hybridisation for the analysis of chromosomal imbalances in solid tumours and haematological malignancies," *Histochem. Cell Biol.*, vol. 108, pp. 403–417, 1997. 5
- [104] S. Knuutila, A. M. Bjorkqvist, K. Autio, M. Tarkkanen, M. Wolf, O. Monni, J. Szymanska, M. L. Larramendy, J. Tapper, H. Pere, W. El-Rifai, S. Hemmer, V. M. Wasenius, V. Vidgren, and Y. Zhu, "DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies," *Am. J. Pathol.*, vol. 152, pp. 1107–1123, May 1998.

- [105] S. Knuutila, K. Autio, and Y. Aalto, "Online access to CGH data of DNA sequence copy number changes," Am. J. Pathol., vol. 157, p. 689, Aug 2000.
- [106] S. Struski, M. Doco-Fenzy, and P. Cornillet-Lefebvre, "Compilation of published comparative genomic hybridization studies," *Cancer Genet. Cytogenet.*, vol. 135, pp. 63–90, May 2002. 5
- [107] K. Kudoh, M. Takano, T. Koshikawa, S. Yoshida, M. Hirai, Y. Kikuchi, I. Nagata, M. Miwa, and K. Uchida, "Comparative genomic hybridization for analysis of chromosomal changes in cisplatin-resistant ovarian cancer," *Hum. Cell*, vol. 13, pp. 109–116, Sep 2000. 5
- [108] K. Kudoh, M. Takano, T. Koshikawa, M. Hirai, S. Yoshida, Y. Mano, K. Yamamoto, K. Ishii, T. Kita, Y. Kikuchi, I. Nagata, M. Miwa, and K. Uchida, "Gains of 1q21-q22 and 13q12q14 are potential indicators for resistance to cisplatin-based chemotherapy in ovarian cancer patients," *Clin. Cancer Res.*, vol. 5, pp. 2526–2531, Sep 1999.
- [109] N. Wang, "Cytogenetics and molecular genetics of ovarian cancer," Am. J. Med. Genet., vol. 115, pp. 157–163, Oct 2002. 5
- [110] C. A. Werner, H. Dohner, S. Joos, L. H. Trumper, M. Baudis, T. F. Barth, G. Ott, P. Moller, P. Lichter, and M. Bentz, "High-level DNA amplifications are common genetic aberrations in B-cell neoplasms," *Am. J. Pathol.*, vol. 151, pp. 335–342, Aug 1997. 5
- [111] L. James and J. Varley, "Preparation, labelling and detection of dna from archival tissue sections suitable for comparative genomic hybridization," *Chromosome Research*, vol. 4, pp. 163–164, March 1996. 5
- [112] O. Haas, T. Henn, K. Romanakis, S. du Manoir, and C. Lengauer, "Comparative genomic hybridization as part of a new diagnostic strategy in childhood hyperdiploid acute lymphoblastic leukemia," *Leukemia*, vol. 12, pp. 474–481, Apr 1998. 5
- [113] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931–945, Oct 2004. 5
- [114] J. Schein, T. Kucaba, M. Sekhon, D. Smailus, R. Waterston, and M. Marra, "Highthroughput BAC fingerprinting," *Methods Mol. Biol.*, vol. 255, pp. 143–156, 2004. 5
- [115] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Nguyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler, "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation," *Genome Res.*, vol. 13, pp. 2291–2305, Oct 2003. 5, 7
- [116] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter, "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," *Genes Chromosomes Cancer*, vol. 20, pp. 399–407, Dec 1997. 5

- [117] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nat. Genet.*, vol. 20, pp. 207–211, Oct 1998. 187
- [118] J. M. Lage, J. H. Leamon, T. Pejovic, S. Hamann, M. Lacey, D. Dillon, R. Segraves, B. Vossbrinck, A. Gonzlez, D. Pinkel, D. G. Albertson, J. Costa, and P. M. Lizardi, "Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH," *Genome Res.*, vol. 13, pp. 294–307, Feb 2003.
- [119] J. A. Veltman, E. F. Schoenmakers, B. H. Eussen, I. Janssen, G. Merkx, B. van Cleef, C. M. van Ravenswaaij, H. G. Brunner, D. Smeets, and A. G. van Kessel, "High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization," *Am. J. Hum. Genet.*, vol. 70, pp. 1269–1276, May 2002.
- [120] K. K. Mantripragada, P. G. Buckley, T. D. de Sthl, and J. P. Dumanski, "Genomic microarrays in the spotlight," *Trends Genet.*, vol. 20, pp. 87–94, Feb 2004.
- [121] A. S. Ishkanian, C. A. Malloff, S. K. Watson, R. J. DeLeeuw, B. Chi, B. P. Coe, A. Snijders, D. G. Albertson, D. Pinkel, M. A. Marra, V. Ling, C. MacAulay, and W. L. Lam, "A tiling resolution DNA microarray with complete coverage of the human genome," *Nat. Genet.*, vol. 36, pp. 299–303, Mar 2004. 5, 18, 118
- [122] A. N. Jain, K. Chin, A. L. Borresen-Dale, B. K. Erikstein, P. Eynstein Lonning, R. Kaaresen, and J. W. Gray, "Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 7952–7957, Jul 2001. 6
- [123] S. R. Ghaffari, E. Boyd, J. L. Tolmie, Y. J. Crow, A. H. Trainer, and J. M. Connor, "A new strategy for cryptic telomeric translocation screening in patients with idiopathic mental retardation," *J. Med. Genet.*, vol. 35, pp. 225–233, Mar 1998. 6
- [124] L. E. Vissers, B. B. de Vries, K. Osoegawa, I. M. Janssen, T. Feuth, C. O. Choy, H. Straatman, W. van der Vliet, E. H. Huys, A. van Rijk, D. Smeets, C. M. van Ravenswaaij-Arts, N. V. Knoers, I. van der Burgt, P. J. de Jong, H. G. Brunner, A. G. van Kessel, E. F. Schoenmakers, and J. A. Veltman, "Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities," *Am. J. Hum. Genet.*, vol. 73, pp. 1261–1270, Dec 2003.
- [125] N. Harada, E. Hatchwell, N. Okamoto, M. Tsukahara, K. Kurosawa, H. Kawame, T. Kondoh, H. Ohashi, R. Tsukino, Y. Kondoh, O. Shimokawa, T. Ida, T. Nagai, Y. Fukushima, K. Yoshiura, N. Niikawa, and N. Matsumoto, "Subtelomere specific microarray based comparative genomic hybridisation: a rapid detection system for cryptic rearrangements in idiopathic mental retardation," *J. Med. Genet.*, vol. 41, pp. 130–136, Feb 2004. 6

- [126] M. Krzywinski, I. Bosdet, D. Smailus, R. Chiu, C. Mathewson, N. Wye, S. Barber, M. Brown-John, S. Chan, S. Chand, A. Cloutier, N. Girn, D. Lee, A. Masson, M. Mayo, T. Olson, P. Pandoh, A. L. Prabhu, E. Schoenmakers, M. Tsai, D. Albertson, W. Lam, C. O. Choy, K. Osoegawa, S. Zhao, P. J. de Jong, J. Schein, S. Jones, and M. A. Marra, "A set of BAC clones spanning the human genome," *Nucleic Acids Res.*, vol. 32, pp. 3651–3660, 2004. 6, 118
- [127] B. P. Coe, B. Ylstra, B. Carvalho, G. A. Meijer, C. Macaulay, and W. L. Lam, "Resolving the resolution of array CGH," *Genomics*, vol. 89, pp. 647–653, May 2007. 6
- [128] B. C. Ballif, S. G. Sulpizio, R. M. Lloyd, S. L. Minier, A. Theisen, B. A. Bejjani, and L. G. Shaffer, "The clinical utility of enhanced subtelomeric coverage in array CGH," Am. J. Med. Genet. A, vol. 143A, pp. 1850–1857, Aug 2007. 6
- [129] B. C. Ballif, S. A. Hornor, E. Jenkins, S. Madan-Khetarpal, U. Surti, K. E. Jackson, A. Asamoah, P. L. Brock, G. C. Gowans, R. L. Conway, J. M. Graham, L. Medne, E. H. Zackai, T. H. Shaikh, J. Geoghegan, R. R. Selzer, P. S. Eis, B. A. Bejjani, and L. G. Shaffer, "Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2," *Nat. Genet.*, vol. 39, pp. 1071–1073, Sep 2007.
- [130] L. G. Shaffer and J. R. Lupski, "Molecular mechanisms for constitutional chromosomal rearrangements in humans," Annu. Rev. Genet., vol. 34, pp. 297–329, 2000.
- [131] B. C. Ballif, W. Yu, C. A. Shaw, C. D. Kashork, and L. G. Shaffer, "Monosomy 1p36 breakpoint junctions suggest pre-meiotic breakage-fusion-bridge cycles are involved in generating terminal deletions," *Hum. Mol. Genet.*, vol. 12, pp. 2153–2165, Sep 2003.
- [132] B. C. Ballif, K. Wakui, M. Gajecka, and L. G. Shaffer, "Translocation breakpoint mapping and sequence analysis in three monosomy 1p36 subjects with der(1)t(1;1)(p36;q44) suggest mechanisms for telomere capture in stabilizing de novo terminal rearrangements," *Hum. Genet.*, vol. 114, pp. 198–206, Jan 2004. 6
- [133] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor, "Lightgenerated oligonucleotide arrays for rapid DNA sequence analysis," *Proc. Natl. Acad. Sci.* U.S.A., vol. 91, pp. 5022–5026, May 1994. 6
- [134] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, pp. 767–773, Feb 1991.
- [135] G. C. Kennedy, H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. Fodor, and K. W. Jones, "Large-scale genotyping of complex DNA," *Nat. Biotechnol.*, vol. 21, pp. 1233–1237, Oct 2003. 8, 66, 187

- [136] H. Matsuzaki, S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. Fodor, and R. Mei, "Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays," *Nat. Methods*, vol. 1, pp. 109–111, Nov 2004. 66
- [137] H. Matsuzaki, H. Loi, S. Dong, Y. Y. Tsai, J. Fang, J. Law, X. Di, W. M. Liu, G. Yang, G. Liu, J. Huang, G. C. Kennedy, T. B. Ryder, G. A. Marcus, P. S. Walsh, M. D. Shriver, J. M. Puck, K. W. Jones, and R. Mei, "Parallel genotyping of over 10,000 SNPs using a oneprimer assay on a high-density oligonucleotide array," *Genome Res.*, vol. 14, pp. 414–425, Mar 2004. 6, 8, 187
- [138] G. J. Upton and J. C. Lloyd, "Oligonucleotide arrays: information from replication and spatial structure," *Bioinformatics*, vol. 21, pp. 4162–4168, Nov 2005. 6
- [139] N. P. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nat. Genet.*, vol. 39, pp. 16–21, Jul 2007. 6, 7, 10, 13, 19, 31
- [140] J. C. Marioni, N. P. Thorne, A. Valsesia, T. Fitzgerald, R. Redon, H. Fiegler, D. T. Andrews, B. E. Stranger, A. G. Lynch, E. T. Dermitzakis, N. P. Carter, S. Tavare, and M. E. Hurles, "Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization," *Genome Biology*, vol. 8, October 2007. 18, 31, 75, 188
- [141] A. Itsara, G. M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R. M. Krauss, R. M. Myers, P. M. Ridker, D. I. Chasman, H. Mefford, P. Ying, D. A. Nickerson, and E. E. Eichler, "Population analysis of large copy number variants and hotspots of human genetic disease," *Am. J. Hum. Genet.*, vol. 84, pp. 148–161, Feb 2009. 10, 17, 18, 19, 67, 188, 191, 282
- [142] I. Ionita-Laza, A. J. Rogers, C. Lange, B. A. Raby, and C. Lee, "Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis," *Genomics*, vol. 93, pp. 22–26, Jan 2009. 7
- [143] The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299–1320, Oct 2005. 7
- [144] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee,

A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart, "A second generation human haplotype map of over 3.1 million SNPs," Nature, vol. 449, pp. 851–861, Oct 2007.

- [145] S. Myles, D. Davison, J. Barrett, M. Stoneking, and N. Timpson, "Worldwide population differentiation at disease-associated SNPs," *BMC Med Genomics*, vol. 1, p. 22, 2008. 7
- [146] L. Bernardini, V. Alesi, S. Loddo, A. Novelli, I. Bottillo, A. Battaglia, M. C. Digilio, G. Zampino, A. Ertel, P. Fortina, S. Surrey, and B. Dallapiccola, "High-resolution SNP arrays in mental retardation diagnostics: how much do we gain?," *Eur. J. Hum. Genet.*, vol. 18, pp. 178–185, Feb 2010. 8, 187, 188
- [147] J. Sebat, D. L. Levy, and S. E. McCarthy, "Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders," *Trends Genet.*, vol. 25, pp. 528–535, Dec 2009. 8
- [148] J. T. Glessner, K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, M. Imielinski, E. C. Frackelton, J. Reichert, E. L. Crawford, J. Munson, P. M. Sleiman, R. Chiavacci, K. Annaiah, K. Thomas, C. Hou, W. Glaberson,

J. Flory, F. Otieno, M. Garris, L. Soorya, L. Klei, J. Piven, K. J. Meyer, E. Anagnostou, T. Sakurai, R. M. Game, D. S. Rudd, D. Zurawiecki, C. J. McDougle, L. K. Davis, J. Miller, D. J. Posey, S. Michaels, A. Kolevzon, J. M. Silverman, R. Bernier, S. E. Levy, R. T. Schultz, G. Dawson, T. Owley, W. M. McMahon, T. H. Wassink, J. A. Sweeney, J. I. Nurnberger, H. Coon, J. S. Sutcliffe, N. J. Minshew, S. F. Grant, M. Bucan, E. H. Cook, J. D. Buxbaum, B. Devlin, G. D. Schellenberg, and H. Hakonarson, "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes," *Nature*, vol. 459, pp. 569–573, May 2009. 8, 197

- [149] G. R. Bignell, J. Huang, J. Greshock, S. Watt, A. Butler, S. West, M. Grigorova, K. W. Jones, W. Wei, M. R. Stratton, P. A. Futreal, B. Weber, M. H. Shapero, and R. Wooster, "High-resolution analysis of DNA copy number using oligonucleotide microarrays," *Genome Res.*, vol. 14, pp. 287–295, Feb 2004. 8, 9, 12, 13, 66
- [150] J. Huang, W. Wei, J. Chen, J. Zhang, G. Liu, X. Di, R. Mei, S. Ishikawa, H. Aburatani, K. W. Jones, and M. H. Shapero, "CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays," *BMC Bioinformatics*, vol. 7, p. 83, 2006. 9, 12, 13, 188
- [151] M. Kadota, H. H. Yang, B. Gomez, M. Sato, R. J. Clifford, D. Meerzaman, B. K. Dunn, L. M. Wakefield, and M. P. Lee, "Delineating genetic alterations for tumor progression in the MCF10A series of breast cancer cell lines," *PLoS ONE*, vol. 5, p. e9201, 2010.
- [152] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y. J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M. S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson, "The land-scape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, pp. 899–905, Feb 2010.
- [153] H. Caren, H. Kryh, M. Nethander, R. M. Sjoberg, C. Trager, S. Nilsson, J. Abrahamsson, P. Kogner, and T. Martinsson, "High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset," *Proc. Natl. Acad. Sci.* U.S.A., vol. 107, pp. 4323–4328, Mar 2010. 8
- [154] D. Capello, M. Scandurra, G. Poretti, P. M. Rancoita, M. Mian, A. Gloghini, C. Deambrogi, M. Martini, D. Rossi, T. C. Greiner, W. C. Chan, M. Ponzoni, S. M. Moreno, M. A. Piris, V. Canzonieri, M. Spina, U. Tirelli, G. Inghirami, A. Rinaldi, E. Zucca, R. D. Favera,

F. Cavalli, L. M. Larocca, I. Kwee, A. Carbone, G. Gaidano, and F. Bertoni, "Genome wide DNA-profiling of HIV-related B-cell lymphomas," *Br. J. Haematol.*, vol. 148, pp. 245–255, Jan 2010.

- [155] E. Coustan-Smith, C. G. Mullighan, M. Onciu, F. G. Behm, S. C. Raimondi, D. Pei, C. Cheng, X. Su, J. E. Rubnitz, G. Basso, A. Biondi, C. H. Pui, J. R. Downing, and D. Campana, "Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia," *Lancet Oncol.*, vol. 10, pp. 147–156, Feb 2009.
- [156] W. Liu, J. Sun, G. Li, Y. Zhu, S. Zhang, S. T. Kim, J. Sun, F. Wiklund, K. Wiley, S. D. Isaacs, P. Stattin, J. Xu, D. Duggan, J. D. Carpten, W. B. Isaacs, H. Gronberg, S. L. Zheng, and B. L. Chang, "Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer," *Cancer Res.*, vol. 69, pp. 2176–2179, Mar 2009.
- [157] J. J. Yang, C. Cheng, W. Yang, D. Pei, X. Cao, Y. Fan, S. B. Pounds, G. Neale, L. R. Trevino, D. French, D. Campana, J. R. Downing, W. E. Evans, C. H. Pui, M. Devidas, W. P. Bowman, B. M. Camitta, C. L. Willman, S. M. Davies, M. J. Borowitz, W. L. Carroll, S. P. Hunger, and M. V. Relling, "Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia," *JAMA*, vol. 301, pp. 393–403, Jan 2009. 9
- [158] C. G. Mullighan and J. R. Downing, "Global genomic characterization of acute lymphoblastic leukemia," *Semin. Hematol.*, vol. 46, pp. 3–15, Jan 2009. 8
- [159] M. Chen, Y. Ye, H. Yang, P. Tamboli, S. Matin, N. M. Tannir, C. G. Wood, J. Gu, and X. Wu, "Genome-wide profiling of chromosomal alterations in renal cell carcinoma using high-density single nucleotide polymorphism arrays," *Int. J. Cancer*, vol. 125, pp. 2342– 2348, Nov 2009.
- [160] B. G. Barwick, M. Abramovitz, M. Kodani, C. S. Moreno, R. Nam, W. Tang, M. Bouzyk, A. Seth, and B. Leyland-Jones, "Prostate cancer genes associated with TMPRSS2-ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts," *Br. J. Cancer*, vol. 102, pp. 570–576, Feb 2010. 8
- [161] D. Komura, F. Shen, S. Ishikawa, K. R. Fitch, W. Chen, J. Zhang, G. Liu, S. Ihara, H. Nakamura, M. E. Hurles, C. Lee, S. W. Scherer, K. W. Jones, M. H. Shapero, J. Huang, and H. Aburatani, "Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays," *Genome Res.*, vol. 16, pp. 1575–1584, Dec 2006. 8, 9, 10, 12, 13, 19, 66, 188
- [162] J. Huang, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones, and M. H. Shapero, "Whole genome DNA copy number changes identified by high density oligonucleotide arrays," *Hum. Genomics*, vol. 1, pp. 287–299, May 2004. 13, 66

- [163] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. Macdonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. Mcintosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y.-H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and C. J. Venter, "The diploid genome sequence of an individual human," *PLoS Biology*, vol. 5, October 2007. 11, 23, 32, 34
- [164] T. H. Shaikh, X. Gai, J. C. Perin, J. T. Glessner, H. Xie, K. Murphy, R. O'Hara, T. Casalunovo, L. K. Conlin, M. D'Arcy, E. C. Frackelton, E. A. Geiger, C. Haldeman-Englert, M. Imielinski, C. E. Kim, L. Medne, K. Annaiah, J. P. Bradfield, E. Dabaghyan, A. Eckert, C. C. Onyiah, S. Ostapenko, F. G. Otieno, E. Santa, J. L. Shaner, R. Skraban, R. M. Smith, J. Elia, E. Goldmuntz, N. B. Spinner, E. H. Zackai, R. M. Chiavacci, R. Grund-meier, E. F. Rappaport, S. F. Grant, P. S. White, and H. Hakonarson, "High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications," *Genome Res.*, vol. 19, pp. 1682–1690, Sep 2009. 8, 76
- [165] G. Kirov, I. Nikolov, L. Georgieva, V. Moskvina, M. J. Owen, and M. C. O'Donovan, "Pooled DNA genotyping on Affymetrix SNP genotyping arrays," *BMC Genomics*, vol. 7, p. 27, 2006. 8
- [166] "Genechip human mapping 500k array set: Data sheet," tech. rep. http://media.affymetrix. com:80/support/technical/datasheets/500k\_datasheet.pdf. 9, 118
- [167] "Genechip human mapping 100k array set: Data sheet," tech. rep. http://media.affymetrix. com:80/support/technical/datasheets/100k\_datasheet.pdf. 9, 10
- [168] S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler, "Integrated detection and population-genetic analysis of SNPs and copy number variation," *Nat. Genet.*, vol. 40, pp. 1166–1174, Oct 2008. 9, 10, 17, 28, 32, 67, 69
- [169] T. Santiago-Sim, S. R. Depalma, K. L. Ju, B. McDonough, C. E. Seidman, J. G. Seidman, and D. H. Kim, "Genomewide linkage in a large Caucasian family maps a new locus for intracranial aneurysms to chromosome 13q," *Stroke*, vol. 40, pp. 57–60, Mar 2009. 9
- [170] Y. Y. Teo, K. S. Small, A. E. Fry, Y. Wu, D. P. Kwiatkowski, and T. G. Clark, "Power consequences of linkage disequilibrium variation between populations," *Genet. Epidemiol.*, vol. 33, pp. 128–135, Feb 2009. 9
- [171] H. Ling, D. M. Waterworth, H. A. Stirnadel, T. I. Pollin, P. J. Barter, Y. A. Kesaniemi, R. W. Mahley, R. McPherson, G. Waeber, T. P. Bersot, J. C. Cohen, S. M. Grundy, V. E. Mooser, and B. D. Mitchell, "Genome-wide linkage and association analyses to identify

genes influencing adiponectin levels: the GEMS Study," *Obesity (Silver Spring)*, vol. 17, pp. 737–744, Apr 2009. 9

- [172] D. A. Peiffer, J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A. Shaw, J. Belmont, S. W. Cheung, R. M. Shen, D. L. Barker, and K. L. Gunderson, "High-resolution genomic profiling of chromosomal aberrations using Infinium wholegenome genotyping," *Genome Res.*, vol. 16, pp. 1136–1148, Sep 2006. 9, 66
- [173] J. Perkel, "SNP genotyping: six technologies that keyed a revolution," *Nature Methods*, vol. 5, pp. 447–453, May 2008. 9
- [174] J. Ragoussis, "Genotyping technologies for genetic research," Annu Rev Genomics Hum Genet, vol. 10, pp. 117–133, 2009. 9
- [175] L. Winchester, C. Yau, and J. Ragoussis, "Comparing CNV detection methods for SNP arrays," *Brief Funct Genomic Proteomic*, vol. 8, pp. 353–366, Sep 2009. 9, 13, 15, 18, 19
- [176] A. E. Dellinger, S. M. Saw, L. K. Goh, M. Seielstad, T. L. Young, and Y. J. Li, "Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays," *Nucleic Acids Res*, Feb 2010. 9, 13, 19, 188
- [177] "Illumina genome-wide dna analysis beadchips: Data sheet," tech. rep. http://www.illumina. com/Documents/products/datasheets/datasheet\_infiniumhd.pdf. 9
- [178] A. Herr, R. Grutzmann, A. Matthaei, J. Artelt, E. Schrock, A. Rump, and C. Pilarsky, "High-resolution analysis of chromosomal imbalances using the Affymetrix 10K SNP genotyping chip," *Genomics*, vol. 85, pp. 392–400, Mar 2005. 9, 12, 13, 188
- [179] H. R. Slater, D. K. Bailey, H. Ren, M. Cao, K. Bell, S. Nasioulas, R. Henke, K. H. Choo, and G. C. Kennedy, "High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs," *Am. J. Hum. Genet.*, vol. 77, pp. 709– 726, Nov 2005.
- [180] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D. K. Bailey, G. C. Kennedy, and S. Ogawa, "A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays," *Cancer Res.*, vol. 65, pp. 6071–6079, Jul 2005. 33, 68, 69, 75, 85
- [181] X. Zhao, B. A. Weir, T. LaFramboise, M. Lin, R. Beroukhim, L. Garraway, J. Beheshti, J. C. Lee, K. Naoki, W. G. Richards, D. Sugarbaker, F. Chen, M. A. Rubin, P. A. Janne, L. Girard, J. Minna, D. Christiani, C. Li, W. R. Sellers, and M. Meyerson, "Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis," *Cancer Res.*, vol. 65, pp. 5561–5570, Jul 2005.

- [182] R. Beroukhim, M. Lin, Y. Park, K. Hao, X. Zhao, L. A. Garraway, E. A. Fox, E. P. Hochberg, I. K. Mellinghoff, M. D. Hofer, A. Descazeaud, M. A. Rubin, M. Meyerson, W. H. Wong, W. R. Sellers, and C. Li, "Inferring loss-of-heterozygosity from unpaired tumors using highdensity oligonucleotide SNP arrays," *PLoS Comput. Biol.*, vol. 2, p. e41, May 2006.
- [183] J. C. Ting, Y. Ye, G. H. Thomas, I. Ruczinski, and J. Pevsner, "Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan," *BMC Bioinformatics*, vol. 7, p. 25, 2006. 9, 12, 13, 188
- [184] N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, and W. S. Noble, "Unsupervised segmentation of continuous genomic data," *Bioinformatics*, vol. 23, pp. 1424– 1426, Jun 2007. 10, 13, 14, 15, 69
- [185] S. Colella, C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis, "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data," *Nucleic Acids Res.*, vol. 35, pp. 2013–2025, 2007. 13, 15, 30, 67, 69, 188
- [186] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Res.*, vol. 17, pp. 1665–1674, Nov 2007. 13, 15, 30, 67, 188
- [187] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, J. Zhang, Y. Guo, B. Feng, H. Li, Y. Lu, X. Fang, H. Liang, Z. Du, D. Li, Y. Zhao, Y. Hu, Z. Yang, H. Zheng, I. Hellmann, M. Inouye, J. Pool, X. Yi, J. Zhao, J. Duan, Y. Zhou, J. Qin, L. Ma, G. Li, Z. Yang, G. Zhang, B. Yang, C. Yu, F. Liang, W. Li, S. Li, D. Li, P. Ni, J. Ruan, Q. Li, H. Zhu, D. Liu, Z. Lu, N. Li, G. Guo, J. Zhang, J. Ye, L. Fang, Q. Hao, Q. Chen, Y. Liang, Y. Su, A. San, C. Ping, S. Yang, F. Chen, L. Li, K. Zhou, H. Zheng, Y. Ren, L. Yang, Y. Gao, G. Yang, Z. Li, X. Feng, K. Kristiansen, G. K. Wong, R. Nielsen, R. Durbin, L. Bolund, X. Zhang, S. Li, H. Yang, and J. Wang, "The diploid genome sequence of an Asian individual," *Nature*, vol. 456, pp. 60–65, Nov 2008. 10
- [188] C. Li, R. Beroukhim, B. A. Weir, W. Winckler, L. A. Garraway, W. R. Sellers, and M. Meyerson, "Major copy proportion analysis of tumor samples using SNP arrays," *BMC Bioinformatics*, vol. 9, p. 204, 2008. 13, 15, 69, 188
- [189] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostat*, vol. 5, no. 4, pp. 557–572, 2004. 13, 16, 17, 21, 30, 67, 75, 85, 190
- [190] G. M. Cooper, T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson, "Systematic assessment of copy number variant detection via genome-wide SNP genotyping," *Nat. Genet.*, vol. 40, pp. 1199–1203, Oct 2008.

- [191] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. 24, pp. 309–318, Feb 2008.
- [192] G. Rigaill, P. Hupe, A. Almeida, P. La Rosa, J. P. Meyniel, C. Decraene, and E. Barillot, "ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays," *Bioinformatics*, vol. 24, pp. 768–774, Mar 2008.
- [193] L. Franke, C. G. de Kovel, Y. S. Aulchenko, G. Trynka, A. Zhernakova, K. A. Hunt, H. M. Blauw, L. H. van den Berg, R. Ophoff, P. Deloukas, D. A. van Heel, and C. Wijmenga, "Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays," *Am. J. Hum. Genet.*, vol. 82, pp. 1316–1333, Jun 2008. 10
- [194] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler, "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nat. Genet.*, vol. 41, pp. 1061–1067, Oct 2009. 10, 11, 19, 188
- [195] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Res.*, vol. 19, pp. 1586–1592, Sep 2009. 11, 13, 85, 196
- [196] M. C. Wendl and R. K. Wilson, "Aspects of coverage in medical DNA sequencing," BMC Bioinformatics, vol. 9, p. 239, 2008.
- [197] M. C. Wendl and R. K. Wilson, "Statistical aspects of discerning indel-type structural variation via DNA sequence alignment," *BMC Genomics*, vol. 10, p. 359, 2009. 10, 11
- [198] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder, "Paired-end mapping reveals extensive structural variation in the human genome," *Science*, vol. 318, pp. 420–426, Oct 2007. 10, 188
- [199] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann,

A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," Nature, vol. 456, pp. 53–59, Nov 2008. 10

- [200] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg, "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, pp. 872–876, Apr 2008. 10, 11
- [201] A. V. Dalca and M. Brudno, "Genome variation discovery with high-throughput sequencing data," *Brief. Bioinformatics*, vol. 11, pp. 3–14, Jan 2010. 11
- [202] A. Coombs, "The sequencing shakeup," *Nat. Biotechnol.*, vol. 26, pp. 1109–1112, Oct 2008. 11, 12
- [203] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Res.*, vol. 37, pp. 4181–4193, Jul 2009. 11
- [204] D. Stekel, *Microarray bioinformatics*. Cambridge University Press, 2003. 12, 31, 32, 34, 35, 36, 39, 266

- [205] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.," *Bioinformatics*, vol. 19, pp. 185–193, January 2003. 12, 32, 33, 39, 70, 89, 266, 267
- [206] T. Park, S. G. Yi, S. H. Kang, S. Lee, Y. S. Lee, and R. Simon, "Evaluation of normalization methods for microarray data," *BMC Bioinformatics*, vol. 4, p. 33, Sep 2003. 12
- [207] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 31–36, Jan 2001. 12, 31, 67
- [208] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal* of the American Statistical Association, vol. 74, no. 368, pp. 829–836, 1979. 12, 75
- [209] X. Zhao, C. Li, J. G. Paez, K. Chin, P. A. Jänne, T. H. Chen, L. Girard, J. Minna, D. Christiani, C. Leo, J. W. Gray, W. R. Sellers, and M. Meyerson, "An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays.," *Cancer Res*, vol. 64, pp. 3060–3071, May 2004. 13, 15, 43, 66, 68
- [210] P. Hupe, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot, "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions," *Bioinformatics*, vol. 20, pp. 3413–3422, Dec 2004. 13, 17, 18, 30, 67, 76, 109
- [211] B. P. Coe, R. Chari, C. Macaulay, and W. L. Lam, "FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data," *Nucleic Acids Res*, Jun 2010. 13
- [212] H. Fiegler, R. Redon, D. Andrews, C. Scott, R. Andrews, C. Carder, R. Clark, O. Dovey, P. Ellis, L. Feuk, L. French, P. Hunt, D. Kalaitzopoulos, J. Larkin, L. Montgomery, G. H. Perry, B. W. Plumb, K. Porter, R. E. Rigby, D. Rigler, A. Valsesia, C. Langford, S. J. Humphray, S. W. Scherer, C. Lee, M. E. Hurles, and N. P. Carter, "Accurate and reliable high-throughput detection of copy number variation in the human genome," *Genome Res.*, vol. 16, pp. 1566–1574, Dec 2006. 13, 30, 67
- [213] J. R. Vermeesch, C. Melotte, G. Froyen, S. Van Vooren, B. Dutta, N. Maas, S. Vermeulen, B. Menten, F. Speleman, B. De Moor, P. Van Hummelen, P. Marynen, J. P. Fryns, and K. Devriendt, "Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis," *J. Histochem. Cytochem.*, vol. 53, pp. 413–422, Mar 2005. 13, 30, 67
- [214] Y. Lai and H. Zhao, "A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data," *Comput Biol Chem*, vol. 29, pp. 47– 54, Feb 2005. 13, 277

- [215] A. Baross, A. Delaney, I. H. Li, T. Nayar, S. Flibotte, H. Qian, S. Chan, J. Asano, A. Ally, M. Cao, P. Birch, M. B. John, N. Fernandes, A. Go, G. Kennedy, S. Langlois, P. Eydoux, J. Friedman, and M. Marra, "Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data," *BMC Bioinformatics*, vol. 8, no. 1, 2007. 13, 22, 67, 68, 91
- [216] J. R. Pollack, T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Borresen-Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, pp. 12963–12968, Oct 2002. 13
- [217] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, pp. 289–300, 1995. 13
- [218] M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra, "Preprocessing and downstream analysis of microarray DNA copy number profiles," *Brief Bioinform*, Feb 2010. 14, 18
- [219] V. M. Aris, M. J. Cody, J. Cheng, J. J. Dermody, P. Soteropoulos, M. Recce, and P. P. Tolias, "Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer," *BMC Bioinformatics*, vol. 5, p. 185, Nov 2004.
- [220] G. Yavas, M. Koyuturk, M. Ozsoyoglu, M. P. Gould, and T. LaFramboise, "An optimization framework for unsupervised identification of rare copy number variation from SNP array data," *Genome Biol.*, vol. 10, p. R119, 2009. 68, 84
- [221] G. Wong, C. Leckie, K. L. Gorringe, I. Haviv, I. G. Campbell, and A. Kowalczyk, "Exploiting sequence similarity to validate the sensitivity of SNP arrays in detecting fine-scaled copy number variations," *Bioinformatics*, vol. 26, pp. 1007–1014, Apr 2010. 14, 68, 84
- [222] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," pp. 267–296, 1990. 15
- [223] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999. 15
- [224] J. Solomon, J. A. Butman, and A. Sood, "Segmentation of brain tumors in 4D MR images using the hidden Markov model," *Comput Methods Programs Biomed*, vol. 84, pp. 76–85, Dec 2006. 15

- [225] J. Staaf, G. Jonsson, M. Ringner, J. Vallon-Christersson, D. Grabau, A. Arason, H. Gunnarsson, B. A. Agnarsson, P. O. Malmstrom, O. T. Johannsson, N. Loman, R. B. Barkardottir, and A. Borg, "High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer," *Breast Cancer Res.*, vol. 12, p. R25, 2010. 15
- [226] D. Lipson, "Interval scores for quality annotated cgh data.," *EEE InternationalWorkshop on Genomic Signal Processing and Statistics (GENSIPS05) Newport*, 2005. 16
- [227] D. Lipson, Y. Aumann, A. Ben-Dor, N. Linial, and Z. Yakhini, "Efficient calculation of interval scores for DNA copy number data analysis," *J. Comput. Biol.*, vol. 13, pp. 215–228, Mar 2006. 16
- [228] T. S. Price, R. Regan, R. Mott, A. Hedman, B. Honey, R. J. Daniels, L. Smith, A. Greenfield, A. Tiganescu, V. Buckle, N. Ventress, H. Ayyub, A. Salhan, S. Pedraza-Diaz, J. Broxholme, J. Ragoussis, D. R. Higgs, J. Flint, and S. J. L. Knight, "SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data," *Nucl. Acids Res.*, vol. 33, no. 11, pp. 3455–3464, 2005. 16
- [229] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin, "A statistical approach for array cgh data analysis.," *BMC Bioinformatics*, vol. 6, 2005. 16
- [230] A. Sen and M. S. Srivastava, "On tests for detecting change in mean," *The Annals of Statistics*, vol. 3, pp. 98–108, 1975. 16, 75
- [231] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol. 21, pp. 3763–3770, Oct 2005. 17, 75
- [232] H. Willenbrock and J. Fridlyand, "A comparison study: applying segmentation to array cgh data for downstream analyses," *Bioinformatics*, vol. 21, pp. 4084–4091, November 2005. 17, 75, 76, 277
- [233] Z. Wu and R. A. Irizarry, "Preprocessing of oligonucleotide array data," *Nat. Biotechnol.*, vol. 22, pp. 656–658, Jun 2004. 18, 31
- [234] Z. Wu, R. Irizarry, R. Gentleman, F. M. Murillo, and F. Spencer, "A model based background adjustment for oligonucleotide expression arrays," Johns Hopkins University Dept. of Biostatistics Working Paper Series 1001, Berkeley Electronic Press, July 2004. 18, 31
- [235] I. Ionita-Laza, G. H. Perry, B. A. Raby, B. Klanderman, C. Lee, N. M. Laird, S. T. Weiss, and C. Lange, "On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test," *Genet. Epidemiol.*, vol. 32, pp. 273–284, Apr 2008. 18

- [236] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, R. J. O'connell, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja, "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, vol. 307, pp. 1434–1440, Mar 2005. 18, 32, 34, 187
- [237] K. K. Wong, R. J. deLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng, D. E. Horsman, C. MacAulay, R. T. Ng, C. J. Brown, E. E. Eichler, and W. L. Lam, "A comprehensive analysis of common copy-number variations in the human genome," *Am. J. Hum. Genet.*, vol. 80, pp. 91–104, Jan 2007. 19
- [238] A. Shlien and D. Malkin, "Copy number variations and cancer," *Genome Med*, vol. 1, p. 62, 2009. 19
- [239] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution," *Annu Rev Genomics Hum Genet*, vol. 10, pp. 451–481, 2009. 19
- [240] R. Khaja, J. Zhang, J. R. MacDonald, Y. He, A. M. Joseph-George, J. Wei, M. A. Rafiq, C. Qian, M. Shago, L. Pantano, H. Aburatani, K. Jones, R. Redon, M. Hurles, L. Armengol, X. Estivill, R. J. Mural, C. Lee, S. W. Scherer, and L. Feuk, "Genome assembly comparison identifies structural variants in the human genome," *Nat. Genet.*, vol. 38, pp. 1413–1418, Dec 2006. 19
- [241] B. Frank, J. L. Bermejo, K. Hemminki, C. Sutter, B. Wappenschmidt, A. Meindl, M. Kiechle-Bahat, P. Bugert, R. K. Schmutzler, C. R. Bartram, and B. Burwinkel, "Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk," *Carcinogenesis*, vol. 28, pp. 1442–1445, Jul 2007. 19
- [242] E. S. Venkatraman and A. B. Olshen, "A faster circular binary segmentation algorithm for the analysis of array cgh data.," *Bioinformatics*, January 2007. 21, 30, 76, 85, 190
- [243] "Cancer care ontario: Cancer in young adults in canada, toronto, canada," 2006. http://www. phac-aspc.gc.ca/publicat/cyac-cjac06/index-eng.php. 22
- [244] K. J. Cheung, S. P. Shah, C. Steidl, N. Johnson, T. Relander, A. Telenius, B. Lai, K. P. Murphy, W. Lam, A. J. Al-Tourah, J. M. Connors, R. T. Ng, R. D. Gascoyne, and D. E. Horsman, "Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances," *Blood*, vol. 113, pp. 137–148, Jan 2009. 22, 116, 117, 118, 124, 125, 152, 153, 192, 287
- [245] H. Tilly, A. Rossi, A. Stamatoullas, B. Lenormand, C. Bigorgne, A. Kunlin, M. Monconduit, and C. Bastard, "Prognostic value of chromosomal abnormalities in follicular lymphoma," *Blood*, vol. 84, pp. 1043–1049, Aug 1994. 22

- [246] M. Bentz, C. A. Werner, H. Dohner, S. Joos, T. F. Barth, R. Siebert, M. Schroder, S. Stilgenbauer, K. Fischer, P. Moller, and P. Lichter, "High incidence of chromosomal imbalances and gene amplifications in the classical follicular variant of follicle center lymphoma," *Blood*, vol. 88, pp. 1437–1444, Aug 1996.
- [247] A. Viardot, P. Moller, J. Hogel, K. Werner, G. Mechtersheimer, A. D. Ho, G. Ott, T. F. Barth, R. Siebert, S. Gesk, B. Schlegelberger, H. Dohner, and M. Bentz, "Clinicopathologic correlations of genomic gains and losses in follicular lymphoma," *J. Clin. Oncol.*, vol. 20, pp. 4523–4530, Dec 2002. 116
- [248] H. Avet-Loiseau, M. Vigier, A. Moreau, M. P. Mellerin, F. Gaillard, J. L. Harousseau, R. Bataille, and N. Milpied, "Comparative genomic hybridization detects genomic abnormalities in 80% of follicular lymphomas," *Br. J. Haematol.*, vol. 97, pp. 119–122, Apr 1997. 22, 116, 124, 192
- [249] J. J. Yunis, G. Frizzera, M. M. Oken, J. McKenna, A. Theologides, and M. Arnesen, "Multiple recurrent genomic defects in follicular lymphoma. A possible model for cancer," N. Engl. J. Med., vol. 316, pp. 79–84, Jan 1987. 22
- [250] W. B. Graninger, M. Seto, B. Boutain, P. Goldman, and S. J. Korsmeyer, "Expression of Bcl-2 and Bcl-2-Ig fusion transcripts in normal and neoplastic cells," *J. Clin. Invest.*, vol. 80, pp. 1512–1515, Nov 1987. 22
- [251] T. W. McKeithan, J. D. Rowley, T. B. Shows, and M. O. Diaz, "Cloning of the chromosome translocation breakpoint junction of the t(14;19) in chronic lymphocytic leukemia," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 84, pp. 9257–9260, Dec 1987. 22
- [252] T. J. McDonnell, N. Deane, F. M. Platt, G. Nunez, U. Jaeger, J. P. McKearn, and S. J. Korsmeyer, "bcl-2-immunoglobulin transgenic mice demonstrate extended B cell survival and follicular lymphoproliferation," *Cell*, vol. 57, pp. 79–88, Apr 1989. 22
- [253] T. J. McDonnell and S. J. Korsmeyer, "Progression from lymphoid hyperplasia to high-grade malignant lymphoma in mice transgenic for the t(14; 18)," *Nature*, vol. 349, pp. 254–256, Jan 1991. 22
- [254] J. Limpens, R. Stad, C. Vos, C. de Vlaam, D. de Jong, G. J. van Ommen, E. Schuuring, and P. M. Kluin, "Lymphoma-associated translocation t(14;18) in blood B cells of normal individuals," *Blood*, vol. 85, pp. 2528–2536, May 1995. 22, 116
- [255] G. Dolken, G. Illerhaus, C. Hirt, and R. Mertelsmann, "BCL-2/JH rearrangements in circulating B cells of healthy blood donors and patients with nonmalignant diseases," J. Clin. Oncol., vol. 14, pp. 1333–1344, Apr 1996. 22

- [256] J. C. Reed, Y. Tsujimoto, S. F. Epstein, M. Cuddy, T. Slabiak, P. C. Nowell, and C. M. Croce, "Regulation of bcl-2 gene expression in lymphoid cell lines containing normal #18 or t(14;18) chromosomes," *Oncogene Res.*, vol. 4, pp. 271–282, 1989. 22
- [257] V. G. Dyomin, N. Palanisamy, K. O. Lloyd, K. Dyomina, S. C. Jhanwar, J. Houldsworth, and R. S. Chaganti, "MUC1 is activated in a B-cell lymphoma by the t(1;14)(q21;q32) translocation and is rearranged and amplified in B-cell lymphoma subsets," *Blood*, vol. 95, pp. 2666– 2671, Apr 2000.
- [258] C. Kalla, H. Nentwich, M. Schlotter, D. Mertens, K. Wildenberger, H. Dohner, S. Stilgenbauer, and P. Lichter, "Translocation t(X;11)(q13;q23) in B-cell chronic lymphocytic leukemia disrupts two novel genes," *Genes Chromosomes Cancer*, vol. 42, pp. 128–143, Feb 2005. 22
- [259] A. D. Delaney, H. Qian, J. M. Friedman, and M. A. Marra, "Use of Affymetrix mapping arrays in the diagnosis of gene copy number variation," *Curr Protoc Hum Genet*, vol. Chapter 8, p. Unit 8.13, Oct 2008. 23, 69, 96, 109, 117, 118, 174, 188, 191
- [260] A. M. Krasinskas, D. L. Bartlett, K. Cieply, and S. Dacic, "CDKN2A and MTAP deletions in peritoneal mesotheliomas are correlated with loss of p16 protein expression and poor survival," *Mod Pathol*, Jan 2010. 23, 134
- [261] C. Cox, G. Bignell, C. Greenman, A. Stabenau, W. Warren, P. Stephens, H. Davies, S. Watt, J. Teague, S. Edkins, E. Birney, D. F. Easton, R. Wooster, P. A. Futreal, and M. R. Stratton, "A survey of homozygous deletions in human cancer genomes," *Proc. Natl. Acad. Sci.* U.S.A., vol. 102, pp. 4542–4547, Mar 2005. 23, 134
- [262] E. Laharanne, E. Chevret, Y. Idrissi, C. Gentil, M. Longy, J. Ferrer, P. Dubus, T. Jouary, B. Vergier, M. Beylot-Barry, and J. P. Merlio, "CDKN2A-CDKN2B deletion defines an aggressive subset of cutaneous T-cell lymphoma," *Mod Pathol*, Jan 2010. 23
- [263] M. Capasso, M. K. Bhamrah, T. Henley, R. S. Boyd, C. Langlais, K. Cain, D. Dinsdale, K. Pulford, M. Khan, B. Musset, V. V. Cherny, D. Morgan, R. D. Gascoyne, E. Vigorito, T. E. DeCoursey, I. C. MacLennan, and M. J. Dyer, "HVCN1 modulates BCR signal strength via regulation of BCR-dependent generation of reactive oxygen species," *Nat. Immunol.*, vol. 11, pp. 265–272, Mar 2010. 23, 133, 193
- [264] T. J. Pugh, A. D. Delaney, N. Farnoud, S. Flibotte, M. Griffith, H. I. Li, H. Qian, P. Farinha, R. D. Gascoyne, and M. A. Marra, "Impact of whole genome amplification on analysis of copy number variants," *Nucleic Acids Res.*, vol. 36, p. e80, Aug 2008. 24
- [265] M. Rahmani, M. Earp, P. Pannu, N. Farnoud, J. Wu, L. Akhabir, J. Halaschek-Wiener, B. Munt, C. Thompson, S. Mitropanopoulos, D. Craig, P. Par, B. McManus, and A. Brooks-Wilson, "Identification of novel risk loci for calcific aortic valve stenosis on chromosome 1

by a genome-wide scan of 1,000,000 single nucleotide polymorphisms," 2009. Proceedings of National Research Forum for Young Investigators in Circulatory and Respiratory Health/ 2nd Annual Canadian Human Genetics Conference. 24, 195

- [266] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander, "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nat. Methods*, vol. 6, pp. 99–103, Jan 2009. 29, 188
- [267] D. A. Solomon, J. S. Kim, J. C. Cronin, Z. Sibenaller, T. Ryken, S. A. Rosenberg, H. Ressom, W. Jean, D. Bigner, H. Yan, Y. Samuels, and T. Waldman, "Mutational inactivation of PTPRD in glioblastoma multiforme and malignant melanoma," *Cancer Res.*, vol. 68, pp. 10300–10306, Dec 2008. 29
- [268] M. Kadota, M. Sato, B. Duncan, A. Ooshima, H. H. Yang, N. Diaz-Meyer, S. Gere, S. Kageyama, J. Fukuoka, T. Nagata, K. Tsukada, B. K. Dunn, L. M. Wakefield, and M. P. Lee, "Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of PIK3CA," *Cancer Res.*, vol. 69, pp. 7357–7365, Sep 2009. 29
- [269] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers, "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, pp. 20007–20012, Dec 2007. 29
- [270] B. A. Weir, M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, I. I. Wistuba, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson, "Characterizing the cancer genome in lung adenocarcinoma," *Nature*, vol. 450, pp. 893–898, Dec 2007. 29, 195
- [271] J. Wagenstaller, S. Spranger, B. Lorenz-Depiereux, B. Kazmierczak, M. Nathrath, D. Wahl, B. Heye, D. Glaser, V. Liebscher, T. Meitinger, and T. M. Strom, "Copy-number variations"
measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation," *Am. J. Hum. Genet.*, vol. 81, pp. 768–779, Oct 2007. 30

- [272] F. Zahir and J. M. Friedman, "The impact of array genomic hybridization on mental retardation research: a review of current technologies and their clinical utility," *Clin. Genet.*, vol. 72, pp. 271–287, Oct 2007. 30
- [273] D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C. S. Healey, R. Bowman, K. B. Meyer, C. A. Haiman, L. K. Kolonel, B. E. Henderson, L. Le Marchand, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C. Y. Shen, P. E. Wu, H. C. Wang, D. Eccles, D. G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M. R. Stratton, N. Rahman, G. Chenevix-Trench, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, D. J. Hunter, S. E. Hankinson, D. G. Cox, P. Hall, S. Wedren, J. Liu, Y. L. Low, N. Bogdanova, P. Schurmann, T. Dork, R. A. Tollenaar, C. E. Jacobi, P. Devilee, J. G. Klijn, A. J. Sigurdson, M. M. Doody, B. H. Alexander, J. Zhang, A. Cox, I. W. Brock, G. MacPherson, M. W. Reed, F. J. Couch, E. L. Goode, J. E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J. L. Hopper, M. McCredie, M. Southey, G. G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y. D. Ko, A. B. Spurdle, J. Beesley, X. Chen, A. Mannermaa, V. M. Kosma, V. Kataja, J. Hartikainen, N. E. Day, D. R. Cox, and B. A. Ponder, "Genome-wide association study identifies novel breast cancer susceptibility loci," Nature, vol. 447, pp. 1087-1093, Jun 2007. 30
- [274] K. D. Tsuchiya, L. G. Shaffer, S. Aradhya, J. M. Gastier-Foster, A. Patel, M. K. Rudd, J. S. Biggerstaff, W. G. Sanger, S. Schwartz, J. H. Tepperberg, E. C. Thorland, B. A. Torchia, and A. R. Brothman, "Variability in interpreting and reporting copy number changes detected by array-based technology in clinical laboratories," *Genet. Med.*, vol. 11, pp. 866–873, Dec 2009. 31
- [275] B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry, "Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data," *Biostatistics*, vol. 8, pp. 485–499, Apr 2007. 31, 43
- [276] A. C. Cambon, A. Khalyfa, N. G. Cooper, and C. M. Thompson, "Analysis of probe level patterns in Affymetrix microarray data," *BMC Bioinformatics*, vol. 8, p. 146, 2007. 31, 67
- [277] V. Budhraja, E. Spitznagel, W. T. Schaiff, and Y. Sadovsky, "Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays," *BMC Biol.*, vol. 1, p. 1, 2003.
- [278] T. M. Chu, B. Weir, and R. Wolfinger, "A systematic statistical linear modeling approach to oligonucleotide array experiments," *Math Biosci*, vol. 176, pp. 35–51, Mar 2002.

- [279] G. F. Reed, F. Lynn, and B. D. Meade, "Use of coefficient of variation in assessing variability of quantitative assays," *Clin. Diagn. Lab. Immunol.*, vol. 9, pp. 1235–1239, Nov 2002. 36, 37, 43, 49, 189
- [280] B. D. Jovanovic, S. Huang, Y. Liu, K. N. Naguib, and R. C. Bergan, "A simple analysis of gene expression and variability in gene arrays based on repeated observations," *Am J Pharmacogenomics*, vol. 1, pp. 145–152, 2001.
- [281] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, pp. 509–519, Jun 2001. 31
- [282] S. O. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. E. Scheirer, R. S. Parrish, D. B. Allison, and G. P. Page, "Sources of variation in Affymetrix microarray experiments," *BMC Bioinformatics*, vol. 6, p. 214, 2005. 31
- [283] S. Huang, H. R. Qian, C. Geringer, C. Love, L. Gelbert, and K. Bemis, "Assessing the variability in GeneChip data," *Am J Pharmacogenomics*, vol. 3, pp. 279–290, 2003. 31
- [284] J. I. Kim, Y. S. Ju, H. Park, S. Kim, S. Lee, J. H. Yi, J. Mudge, N. A. Miller, D. Hong, C. J. Bell, H. S. Kim, I. S. Chung, W. C. Lee, J. S. Lee, S. H. Seo, J. Y. Yun, H. N. Woo, H. Lee, D. Suh, S. Lee, H. J. Kim, M. Yavartanoo, M. Kwak, Y. Zheng, M. K. Lee, H. Park, J. Y. Kim, O. Gokcumen, R. E. Mills, A. W. Zaranek, J. Thakuria, X. Wu, R. W. Kim, J. J. Huntley, S. Luo, G. P. Schroth, T. D. Wu, H. Kim, K. S. Yang, W. Y. Park, H. Kim, G. M. Church, C. Lee, S. F. Kingsmore, and J. S. Seo, "A highly annotated whole-genome sequence of a Korean individual," *Nature*, vol. 460, pp. 1011–1015, Aug 2009. 32, 34
- [285] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," Nat. Rev. Genet., vol. 7, pp. 85–97, Feb 2006.
- [286] M. C. O'Donovan, G. Kirov, and M. J. Owen, "Phenotypic variations on the theme of CNVs," *Nat. Genet.*, vol. 40, pp. 1392–1393, Dec 2008. 34
- [287] F. F. Parl, "Glutathione S-transferase genotypes and cancer risk," *Cancer Lett.*, vol. 221, pp. 123–129, Apr 2005. 32
- [288] C. McKinney, M. E. Merriman, P. T. Chapman, P. J. Gow, A. A. Harrison, J. Highton, P. B. Jones, L. McLean, J. L. O'Donnell, V. Pokorny, M. Spellerberg, L. K. Stamp, J. Willis, S. Steer, and T. R. Merriman, "Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis," *Ann. Rheum. Dis.*, vol. 67, pp. 409–413, Mar 2008. 32, 34
- [289] Y. Yang, E. K. Chung, Y. L. Wu, S. L. Savelli, H. N. Nagaraja, B. Zhou, M. Hebert, K. N. Jones, Y. Shu, K. Kitzmiller, C. A. Blanchong, K. L. McBride, G. C. Higgins, R. M. Rennebohm, R. R. Rice, K. V. Hackshaw, R. A. Roubey, J. M. Grossman, B. P. Tsao, D. J.

Birmingham, B. H. Rovin, L. A. Hebert, and C. Y. Yu, "Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans," *Am. J. Hum. Genet.*, vol. 80, pp. 1037–1054, Jun 2007. 32

- [290] J. M. Young, R. M. Endicott, S. S. Parghi, M. Walker, J. M. Kidd, and B. J. Trask, "Extensive copy-number variation of the human olfactory receptor gene family," *Am. J. Hum. Genet.*, vol. 83, pp. 228–242, Aug 2008. 32
- [291] S. Giglio, K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, J. L. Weber, D. H. Ledbetter, and O. Zuffardi, "Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements," *Am. J. Hum. Genet.*, vol. 68, pp. 874–883, Apr 2001. 32, 34
- [292] Z. Jiang, H. Tang, M. Ventura, M. F. Cardone, T. Marques-Bonet, X. She, P. A. Pevzner, and E. E. Eichler, "Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution," *Nat. Genet.*, vol. 39, pp. 1361–1368, Nov 2007. 32
- [293] K. K. Dobbin, D. G. Beer, M. Meyerson, T. J. Yeatman, W. L. Gerald, J. W. Jacobson, B. Conley, K. H. Buetow, M. Heiskanen, R. M. Simon, J. D. Minna, L. Girard, D. E. Misek, J. M. Taylor, S. Hanash, K. Naoki, D. N. Hayes, C. Ladd-Acosta, S. A. Enkemann, A. Viale, and T. J. Giordano, "Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays," *Clin. Cancer Res.*, vol. 11, pp. 565–572, Jan 2005. 32
- [294] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nat. Methods*, vol. 2, pp. 345–350, May 2005.
- [295] J. E. Larkin, B. C. Frank, H. Gavras, R. Sultana, and J. Quackenbush, "Independence and reproducibility across microarray platforms," *Nat. Methods*, vol. 2, pp. 337–344, May 2005.
- [296] T. A. Patterson, E. K. Lobenhofer, S. B. Fulmer-Smentek, P. J. Collins, T. M. Chu, W. Bao, H. Fang, E. S. Kawasaki, J. Hager, I. R. Tikhonova, S. J. Walker, L. Zhang, P. Hurban, F. de Longueville, J. C. Fuscoe, W. Tong, L. Shi, and R. D. Wolfinger, "Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project," *Nat. Biotechnol.*, vol. 24, pp. 1140–1150, Sep 2006. 33
- [297] A. de Reynies, D. Geromin, J. M. Cayuela, F. Petel, P. Dessen, F. Sigaux, and D. S. Rickman, "Comparison of the latest commercial short and long oligonucleotide microarray technologies," *BMC Genomics*, vol. 7, p. 51, 2006. 32

- [298] M. Kollegal, S. Adak, R. Shippy, and T. Sendera, "Considerations in making microarray cross-platform correlations," 2005 IEEE Computational Systems Bioinformatics Conference - Workshops, vol. 0, pp. 101–102, 2005. 32
- [299] K. Hao, E. E. Schadt, and J. D. Storey, "Calibrating the performance of SNP arrays for whole-genome association studies," *PLoS Genet.*, vol. 4, p. e1000109, Jun 2008. 32
- [300] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush, "Within the fold: assessing differential expression measures and reproducibility in microarray assays," *Genome Biol.*, vol. 3, p. research0062, Oct 2002. 33
- [301] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18 Suppl 1, pp. 96–104, 2002. 35
- [302] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, Apr 2003. 33, 70, 89, 267
- [303] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron, "Adjustment of systematic microarray data biases," *Bioinformatics*, vol. 20, pp. 105–114, Jan 2004. 33
- [304] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, pp. 9309–9314, Jun 2004.
- [305] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, pp. 118–127, Jan 2007.
- [306] H. Hong, Z. Su, W. Ge, L. Shi, R. Perkins, H. Fang, J. Xu, J. J. Chen, T. Han, J. Kaput, J. C. Fuscoe, and W. Tong, "Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples," *BMC Bioinformatics*, vol. 9 Suppl 9, p. S17, 2008. 35
- [307] A. Scherer, Batch Effects and Noise in Microarray Experiments. Wiley, December 2009. 33
- [308] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, pp. 1929–1935, Sep 2006. 33
- [309] E. S. Lander, "Array of hope," Nat. Genet., vol. 21, pp. 3-4, Jan 1999. 33

- [310] E. E. Eichler, D. A. Nickerson, D. Altshuler, A. M. Bowcock, L. D. Brooks, N. P. Carter, D. M. Church, A. Felsenfeld, M. Guyer, C. Lee, J. R. Lupski, J. C. Mullikin, J. K. Pritchard, J. Sebat, S. T. Sherry, D. Smith, D. Valle, and R. H. Waterston, "Completing the map of human genetic variation," *Nature*, vol. 447, pp. 161–165, May 2007. 34
- [311] S. Huang, A. A. Yeo, L. Gelbert, X. Lin, L. Nisenbaum, and K. G. Bemis, "At what scale should microarray data be analyzed?," *Am J Pharmacogenomics*, vol. 4, pp. 129–139, 2004. 35
- [312] N. Rabbee and T. P. Speed, "A genotype calling algorithm for affymetrix SNP arrays," *Bioin-formatics*, vol. 22, pp. 7–12, Jan 2006. 35
- [313] D. M. Rocke and B. Durbin, "A model for measurement error for gene expression arrays," *J. Comput. Biol.*, vol. 8, pp. 557–569, 2001. 35
- [314] B. Durbin and D. M. Rocke, "Estimation of transformation parameters for microarray data," *Bioinformatics*, vol. 19, pp. 1360–1367, Jul 2003. 35
- [315] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, "Model-based variance-stabilizing transformation for Illumina microarray data," *Nucleic Acids Res.*, vol. 36, p. e11, Feb 2008. 35
- [316] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, vol. 18 Suppl 1, pp. S105– 110, 2002. 35
- [317] M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho, "Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum," *Bioinformatics*, vol. 20, pp. 3575–3582, Dec 2004. 35
- [318] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha,

M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nat. Biotechnol.*, vol. 24, pp. 1151–1161, Sep 2006. 36, 38

- [319] F. Raymond, S. Metairon, R. Borner, M. Hofmann, and M. Kussmann, "Automated target preparation for microarray-based gene expression analysis," *Anal. Chem.*, vol. 78, pp. 6299– 6305, Sep 2006.
- [320] J. Jaeger and R. Spang, "Selecting normalization genes for small diagnostic microarrays," *BMC Bioinformatics*, vol. 7, p. 388, 2006.
- [321] "What is the median coefficient of variation for our microarray experiments?." Institute of Food and Research (IRF), UK. http://www.ifr.ac.uk/safety/microarrays/. 36
- [322] N. L. Johnson and B. L. Welch, "On the calculation of the cumulants of the κ-distribution," *Biometrika*, vol. 31, no. 1/2, pp. 216–218, 1939. 36
- [323] R. J. Wood, "Alternative ways of estimating serological titer reproducibility," J. Clin. Microbiol., vol. 13, pp. 760–768, Apr 1981. 36, 37
- [324] P. J. Park, Y. A. Cao, S. Y. Lee, J. W. Kim, M. S. Chang, R. Hart, and S. Choi, "Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference," *J. Biotechnol.*, vol. 112, pp. 225–245, Sep 2004. 38
- [325] F. Sato, S. Tsuchiya, K. Terasawa, and G. Tsujimoto, "Intra-platform repeatability and interplatform comparability of microRNA microarray technology," *PLoS ONE*, vol. 4, no. 5, p. e5540, 2009. 38
- [326] "Genechip human mapping 10k snp array: Data sheet," tech. rep., 2004. http://media. affymetrix.com:80/support/technical/datasheets/10k2\_datasheet.pdf. 38
- [327] W. Wang, B. Carvalho, N. D. Miller, J. Pevsner, A. Chakravarti, and R. A. Irizarry, "Estimating genome-wide copy number using allele-specific mixture models," *Journal of Computational Biology.*, vol. 15, pp. 857–866, Sep 2008. 43
- [328] H. Matsuzaki, P. H. Wang, J. Hu, R. Rava, and G. K. Fu, "High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians," *Genome Biol.*, vol. 10, p. R125, 2009. 44
- [329] J. S. Reis-Filho, S. Drury, M. B. Lambros, C. Marchio, N. Johnson, R. Natrajan, J. Salter, P. Levey, O. Fletcher, J. Peto, A. Ashworth, and M. Dowsett, "ESR1 gene amplification in breast cancer: a common phenomenon?," *Nat. Genet.*, vol. 40, pp. 809–810, Jul 2008.

- [330] K. Nakao, K. R. Mehta, J. Fridlyand, D. H. Moore, A. N. Jain, A. Lafuente, J. W. Wiencke, J. P. Terdiman, and F. M. Waldman, "High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization," *Carcinogenesis*, vol. 25, pp. 1345–1357, Aug 2004. 44
- [331] T. L. Yang, X. D. Chen, Y. Guo, S. F. Lei, J. T. Wang, Q. Zhou, F. Pan, Y. Chen, Z. X. Zhang, S. S. Dong, X. H. Xu, H. Yan, X. Liu, C. Qiu, X. Z. Zhu, T. Chen, M. Li, H. Zhang, L. Zhang, B. M. Drees, J. J. Hamilton, C. J. Papasian, R. R. Recker, X. P. Song, J. Cheng, and H. W. Deng, "Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis," *Am. J. Hum. Genet.*, vol. 83, pp. 663–674, Dec 2008. 67
- [332] G. H. Perry, A. Ben-Dor, A. Tsalenko, N. Sampas, L. Rodriguez-Revenga, C. W. Tran, A. Scheffer, I. Steinfeld, P. Tsang, N. A. Yamada, H. S. Park, J. I. Kim, J. S. Seo, Z. Yakhini, S. Laderman, L. Bruhn, and C. Lee, "The fine-scale and complex architecture of human copy-number variation," *Am. J. Hum. Genet.*, vol. 82, pp. 685–695, Mar 2008. 67
- [333] J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler, "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nat. Genet.*, vol. 40, pp. 1253–1260, Oct 2008. 69
- [334] L. Wan, K. Sun, Q. Ding, Y. Cui, M. Li, Y. Wen, R. C. Elston, M. Qian, and W. J. Fu, "Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation," *Nucleic Acids Res.*, vol. 37, p. e117, Sep 2009. 69
- [335] T. Speed, "Normalization of affymetrix chips," 2001. http://www.bea.ki.se/staff/reimers/ Web.Pages/Affymetrix.Normalization.htm. 70
- [336] B. Bolstad, "Probe level quantile normalization of high density oligonucleotide array data." Unpublished Manuscript (PDF file), 2001. http://bmbolstad.com/stuff/qnorm.pdf. 70, 266
- [337] J. B. Macqueen, "Some methods of classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967. 71
- [338] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982. 71
- [339] F.-X. Wu, W. Zhang, and A. Kusalik, "A genetic k-means clustering algorithm applied to gene expression data," p. 994, 2003. 71
- [340] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 ed., Nov. 2001.

- [341] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, Jan 2002.
- [342] P. D'haeseleer, "How does gene expression clustering work?," Nat. Biotechnol., vol. 23, pp. 1499–1501, Dec 2005. 71
- [343] N. Keng-Hoong, P. Somnuk, and H. Chin-Kuan, "Evaluating the significance of global and local features in expressed sequence tag: A clustering quality perspective," *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, vol. I, March 18 - 20 2009. 71
- [344] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat. Genet.*, vol. 22, pp. 281–285, Jul 1999.
- [345] B. Kutlu, A. K. Cardozo, M. I. Darville, M. Kruhøffer, N. Magnusson, T. Ørntoft, and D. L. Eizirik, "Discovery of gene networks regulating cytokine-induced dysfunction and apoptosis in insulin-producing INS-1 cells," *Diabetes*, vol. 52, pp. 2701–2719, Nov 2003. 71
- [346] K. Demirli, S. X. Cheng, and P. Muthukumaran, "Subtractive clustering based modeling of job sequencing with parametric search," *Fuzzy Sets Syst.*, vol. 137, no. 2, pp. 235–270, 2003.
   71
- [347] S. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, Sep 1994.
- [348] H. Li, C. L. P. Chen, and H.-P. Huang, Fuzzy Neural Intelligent Systems: Mathematical Foundation and the Applications in Engineering. Boca Raton, FL, USA: CRC Press, Inc., 2000. 71
- [349] R. Yager and D. Filev, "Generation of fuzzy rules by mountain clustering," *Journal of Intelligent and Fuzzy Systems*, vol. 2, pp. 209–219, 1994. 71
- [350] K. Hammouda, "Tools of intelligent systems design: A comparative study of data clustering techniques." Electronic File (PDF). http://www.jsbi.org/pdfs/journal1/GlW09/Poster/ GlW09P061.pdf. 71, 72
- [351] H. Bengtsson, R. Irizarry, B. Carvalho, and T. P. Speed, "Estimation and assessment of raw copy numbers at the single locus level," *Bioinformatics*, vol. 24, pp. 759–767, March 2008.
   75
- [352] T. Kanagawa, "Bias and artifacts in multitemplate polymerase chain reactions (PCR)," J. *Biosci. Bioeng.*, vol. 96, pp. 317–323, 2003.

- [353] S. Jacobs, E. R. Thompson, Y. Nannya, G. Yamamoto, R. Pillai, S. Ogawa, D. K. Bailey, and I. G. Campbell, "Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays," *Cancer Res.*, vol. 67, pp. 2544–2551, Mar 2007. 75
- [354] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Res.*, vol. 30, p. e15, Feb 2002. 75
- [355] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988. 75
- [356] S. J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang, "Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms," *Nucl. Acids Res.*, vol. 36, no. 19, pp. e126–, 2008. 75
- [357] J. Staaf, J. Vallon-Christersson, D. Lindgren, G. Juliusson, R. Rosenquist, M. Hoglund, A. Borg, and M. Ringner, "Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios," *BMC Bioinformatics*, vol. 9, p. 409, 2008. 75
- [358] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain, "Hidden markov models approach to the analysis of array cgh data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 132 – 153, 2004. Special Issue on Multivariate Methods in Genomic Data Analysis. 76
- [359] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Goransson, G. Juliusson, R. Rosenquist, M. Hoglund, A. Borg, and M. Ringner, "Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays," *Genome Biol.*, vol. 9, p. R136, 2008. 76
- [360] J. Camps, M. Grade, Q. T. Nguyen, P. Hormann, S. Becker, A. B. Hummon, V. Rodriguez, S. Chandrasekharappa, Y. Chen, M. J. Difilippantonio, H. Becker, B. M. Ghadimi, and T. Ried, "Chromosomal breakpoints in primary colon cancer cluster at sites of structural variants in the genome," *Cancer Res.*, vol. 68, pp. 1284–1295, Mar 2008. 122
- [361] B. Hickey, "Detection of correlated breakpoints in cancer." Electronic File (PDF), 2009. http://www.cs.brown.edu/research/pubs/theses/masters/2009/hickey.pdf.
- [362] A. J. Aguirre, C. Brennan, G. Bailey, R. Sinha, B. Feng, C. Leo, Y. Zhang, J. Zhang, J. D. Gans, N. Bardeesy, C. Cauwels, C. Cordon-Cardo, M. S. Redston, R. A. DePinho, and L. Chin, "High-resolution characterization of the pancreatic adenocarcinoma genome," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, pp. 9067–9072, Jun 2004. 76

- [363] M. Krzywinski, I. Bosdet, C. Mathewson, N. Wye, J. Brebner, R. Chiu, R. Corbett, M. Field, D. Lee, T. Pugh, S. Volik, A. Siddiqui, S. Jones, J. Schein, C. Collins, and M. Marra, "A bac clone fingerprinting approach to the detection of human genome rearrangements," *Genome Biology*, vol. 8, October 2007. 77
- [364] R. D. Morin, N. A. Johnson, T. M. Severson, A. J. Mungall, J. An, R. Goya, J. E. Paul, M. Boyle, B. W. Woolcock, F. Kuchenbauer, D. Yap, R. K. Humphries, O. L. Griffith, S. Shah, H. Zhu, M. Kimbara, P. Shashkin, J. F. Charlot, M. Tcherpakov, R. Corbett, A. Tam, R. Varhol, D. Smailus, M. Moksa, Y. Zhao, A. Delaney, H. Qian, I. Birol, J. Schein, R. Moore, R. Holt, D. E. Horsman, J. M. Connors, S. Jones, S. Aparicio, M. Hirst, R. D. Gascoyne, and M. A. Marra, "Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin," *Nat. Genet.*, vol. 42, pp. 181–185, Feb 2010. 77
- [365] M. D. Taylor, L. Liu, C. Raffel, C. C. Hui, T. G. Mainprize, X. Zhang, R. Agatep, S. Chiappa, L. Gao, A. Lowrance, A. Hao, A. M. Goldstein, T. Stavrou, S. W. Scherer, W. T. Dura, B. Wainwright, J. A. Squire, J. T. Rutka, and D. Hogg, "Mutations in SUFU predispose to medulloblastoma," *Nat. Genet.*, vol. 31, pp. 306–310, Jul 2002. 78, 98
- [366] P. M. Haverty, L. S. Hon, J. S. Kaminker, J. Chant, and Z. Zhang, "High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors," *BMC Med Genomics*, vol. 2, p. 21, 2009. 85, 120
- [367] "Getting the facts: Non-hodgkin lymphoma," 2008. http://www.lymphoma.org/atf/cf/ %7B0363CDD6-51B5-427B-BE48-E6AF871ACEC9%7D/NON-HODGKIN.PDF. 116
- [368] G. Ott and A. Rosenwald, "Molecular pathogenesis of follicular lymphoma," *Haematolog-ica*, vol. 93, pp. 1773–1776, Dec 2008. 116
- [369] F. Schuler, C. Hirt, and G. Dolken, "Chromosomal translocation t(14;18) in healthy individuals," *Semin. Cancer Biol.*, vol. 13, pp. 203–209, Jun 2003. 116
- [370] G. Dolken, L. Dolken, C. Hirt, C. Fusch, C. S. Rabkin, and F. Schuler, "Age-dependent prevalence and frequency of circulating t(14;18)-positive cells in the peripheral blood of healthy individuals," *J. Natl. Cancer Inst. Monographs*, pp. 44–47, 2008.
- [371] F. Schuler, L. Dolken, C. Hirt, T. Kiefer, T. Berg, G. Fusch, K. Weitmann, W. Hoffmann, C. Fusch, S. Janz, C. S. Rabkin, and G. Dolken, "Prevalence and frequency of circulating t(14;18)-MBR translocation carrying cells in healthy individuals," *Int. J. Cancer*, vol. 124, pp. 958–963, Feb 2009. 116
- [372] R. Chao, L. Nevin, P. Agarwal, J. Riemer, X. Bai, A. Delaney, M. Akana, N. JimenezLopez, T. Bardakjian, A. Schneider, N. Chassaing, D. F. Schorderet, D. FitzPatrick, P. Y. Kwok, L. Ellgaard, D. B. Gould, Y. Zhang, J. Malicki, H. Baier, and A. Slavotinek, "A male with

unilateral microphthalmia reveals a role for TMX3 in eye development," *PLoS ONE*, vol. 5, p. e10565, 2010. 117, 118, 191

- [373] J. T. Herbeck, G. S. Gottlieb, K. Wong, R. Detels, J. P. Phair, C. R. Rinaldo, L. P. Jacobson, J. B. Margolick, and J. I. Mullins, "Fidelity of SNP array genotyping using Epstein Barr virus-transformed B-lymphocyte cell lines: implications for genome-wide association studies," *PLoS ONE*, vol. 4, p. e6915, 2009. 117
- [374] D. E. Horsman, J. M. Connors, T. Pantzar, and R. D. Gascoyne, "Analysis of secondary chromosomal alterations in 165 cases of follicular lymphoma with t(14;18)," *Genes Chro*mosomes Cancer, vol. 30, pp. 375–382, Apr 2001. 118
- [375] R. J. de Leeuw, J. J. Davies, A. Rosenwald, G. Bebb, R. D. Gascoyne, M. J. Dyer, L. M. Staudt, J. A. Martinez-Climent, and W. L. Lam, "Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes," *Hum. Mol. Genet.*, vol. 13, pp. 1827–1837, Sep 2004. 118
- [376] S. P. Shah, X. Xuan, R. J. DeLeeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy, "Integrating copy number polymorphisms into array CGH analysis using a robust HMM," *Bioinformatics*, vol. 22, pp. e431–439, Jul 2006. 118
- [377] A. H. Berger, M. Niki, A. Morotti, B. S. Taylor, N. D. Socci, A. Viale, C. Brennan, J. Szoke, N. Motoi, P. B. Rothman, J. Teruya-Feldstein, W. L. Gerald, M. Ladanyi, and P. P. Pandolfi, "Identification of DOK genes as lung tumor suppressors," *Nat. Genet.*, vol. 42, pp. 216–223, Mar 2010. 120
- [378] M. Kaghad, H. Bonnet, A. Yang, L. Creancier, J. C. Biscan, A. Valent, A. Minty, P. Chalon, J. M. Lelias, X. Dumont, P. Ferrara, F. McKeon, and D. Caput, "Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers," *Cell*, vol. 90, pp. 809–819, Aug 1997. 125
- [379] S. Kawano, C. W. Miller, A. F. Gombart, C. R. Bartram, Y. Matsuo, H. Asou, A. Sakashita, J. Said, E. Tatsumi, and H. P. Koeffler, "Loss of p73 gene expression in leukemias/lymphomas due to hypermethylation," *Blood*, vol. 94, pp. 1113–1120, Aug 1999. 125, 192
- [380] J. Wang and M. J. Lenardo, "Roles of caspases in apoptosis, development, and cytokine maturation revealed by homozygous gene deficiencies," *J. Cell. Sci.*, vol. 113 (Pt 5), pp. 753– 757, Mar 2000. 126
- [381] N. Denis, A. Kitzis, J. Kruh, F. Dautry, and D. Corcos, "Stimulation of methotrexate resistance and dihydrofolate reductase gene amplification by c-myc," *Oncogene*, vol. 6, pp. 1453–1457, Aug 1991. 126
- [382] J. Gearhart, E. E. Pashos, and M. K. Prasad, "Pluripotency redux —advances in stem-cell research," *New England Journal of Medicine*, vol. 357, pp. 1469–1472, 10 2007. 126

- [383] C. V. Dang, "c-Myc target genes involved in cell growth, apoptosis, and metabolism," *Mol. Cell. Biol.*, vol. 19, pp. 1–11, Jan 1999. 126
- [384] D. Tvorogov, M. Sundvall, K. Kurppa, M. Hollmen, S. Repo, M. S. Johnson, and K. Elenius, "Somatic mutations of ErbB4: selective loss-of-function phenotype affecting signal transduction pathways in cancer," J. Biol. Chem., vol. 284, pp. 5582–5591, Feb 2009. 128
- [385] M. Sundvall, K. Iljin, S. Kilpinen, H. Sara, O. P. Kallioniemi, and K. Elenius, "Role of ErbB4 in breast cancer," *J Mammary Gland Biol Neoplasia*, vol. 13, pp. 259–268, Jun 2008. 128
- [386] Y. H. Soung, J. W. Lee, S. Y. Kim, Y. P. Wang, K. H. Jo, S. W. Moon, W. S. Park, S. W. Nam, J. Y. Lee, N. J. Yoo, and S. H. Lee, "Somatic mutations of the ERBB4 kinase domain in human cancers," *Int. J. Cancer*, vol. 118, pp. 1426–1429, Mar 2006. 128
- [387] M. N. Nikiforova, E. D. Hsi, R. M. Braziel, M. L. Gulley, D. G. Leonard, J. A. Nowak, R. R. Tubbs, G. H. Vance, and V. M. Van Deerlin, "Detection of clonal IGH gene rearrangements: summary of molecular oncology surveys of the College of American Pathologists," *Arch. Pathol. Lab. Med.*, vol. 131, pp. 185–189, Feb 2007. 130
- [388] M. O'Riordan and R. Grosschedl, "Transcriptional regulation of early B-lymphocyte differentiation," *Immunol. Rev.*, vol. 175, pp. 94–103, Jun 2000. 130, 162
- [389] D. Cozma, D. Yu, S. Hodawadekar, A. Azvolinsky, S. Grande, J. W. Tobias, M. H. Metzgar, J. Paterson, J. Erikson, T. Marafioti, J. G. Monroe, M. L. Atchison, and A. Thomas-Tikhonenko, "B cell activator PAX5 promotes lymphomagenesis through stimulation of B cell receptor signaling," *J. Clin. Invest.*, vol. 117, pp. 2602–2610, Sep 2007. 130
- [390] R. S. Boyd, R. Jukes-Jones, R. Walewska, D. Brown, M. J. Dyer, and K. Cain, "Protein profiling of plasma membranes defines aberrant signaling pathways in mantle cell lymphoma," *Mol. Cell Proteomics*, vol. 8, pp. 1501–1515, Jul 2009. 133
- [391] J. M. Dal Porto, S. B. Gauld, K. T. Merrell, D. Mills, A. E. Pugh-Bernard, and J. Cambier, "B cell antigen receptor signaling 101," *Mol. Immunol.*, vol. 41, pp. 599–613, Jul 2004. 133
- [392] A. L. Reed, J. Califano, P. Cairns, W. H. Westra, R. M. Jones, W. Koch, S. Ahrendt, Y. Eby, D. Sewell, H. Nawroz, J. Bartek, and D. Sidransky, "High frequency of p16 (CDKN2/MTS-1/INK4A) inactivation in head and neck squamous cell carcinoma," *Cancer Res.*, vol. 56, pp. 3630–3633, Aug 1996. 134
- [393] P. Cairns, T. J. Polascik, Y. Eby, K. Tokino, J. Califano, A. Merlo, L. Mao, J. Herath, R. Jenkins, and W. Westra, "Frequency of homozygous deletion at p16/CDKN2 in primary human tumours," *Nat. Genet.*, vol. 11, pp. 210–212, Oct 1995.

- [394] P. B. Illei, V. W. Rusch, M. F. Zakowski, and M. Ladanyi, "Homozygous deletion of CDKN2A and codeletion of the methylthioadenosine phosphorylase gene in the majority of pleural mesotheliomas," *Clin. Cancer Res.*, vol. 9, pp. 2108–2113, Jun 2003. 134
- [395] F. Jardin, J. P. Jais, T. J. Molina, F. Parmentier, J. M. Picquenot, P. Ruminy, H. Tilly, C. Bastard, G. A. Salles, P. Feugier, C. Thieblemont, C. Gisselbrecht, A. de Reynies, B. Coiffier, C. Haioun, and K. Leroy, "Diffuse large B-cell lymphomas with CDKN2A deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: a GELA study," *Blood*, vol. 116, pp. 1092–1104, Aug 2010. 134
- [396] M. Serrano, H. Lee, L. Chin, C. Cordon-Cardo, D. Beach, and R. A. DePinho, "Role of the INK4a locus in tumor suppression and cell mortality," *Cell*, vol. 85, pp. 27–37, Apr 1996. 134
- [397] C. A. Schmitt, J. S. Fridman, M. Yang, S. Lee, E. Baranov, R. M. Hoffman, and S. W. Lowe, "A senescence program controlled by p53 and p16INK4a contributes to the outcome of cancer therapy," *Cell*, vol. 109, pp. 335–346, May 2002. 134
- [398] I. S. Lossos and R. Levy, "Higher grade transformation of follicular lymphoma: phenotypic tumor progression associated with diverse genetic lesions," *Semin. Cancer Biol.*, vol. 13, pp. 191–202, Jun 2003. 134, 137
- [399] G. Rothschild, C. M. Sottas, H. Kissel, V. Agosti, K. Manova, M. P. Hardy, and P. Besmer, "A role for kit receptor signaling in Leydig cell steroidogenesis," *Biol. Reprod.*, vol. 69, pp. 925–932, Sep 2003. 135
- [400] R. Roskoski, "Structure and regulation of Kit protein-tyrosine kinase-the stem cell factor receptor," *Biochem. Biophys. Res. Commun.*, vol. 338, pp. 1307–1315, Dec 2005. 135
- [401] H. Kitayama, T. Tsujimura, I. Matsumura, K. Oritani, H. Ikeda, J. Ishikawa, M. Okabe, M. Suzuki, K. Yamamura, Y. Matsuzawa, Y. Kitamura, and Y. Kanakura, "Neoplastic transformation of normal hematopoietic cells by constitutively activating mutations of c-kit receptor tyrosine kinase," *Blood*, vol. 88, pp. 995–1004, Aug 1996. 135
- [402] A. Uren, J. C. Yu, M. Karcaaltincaba, J. H. Pierce, and M. A. Heidaran, "Oncogenic activation of the alphaPDGFR defines a domain that negatively regulates receptor dimerization," *Oncogene*, vol. 14, pp. 157–162, Jan 1997. 136, 194
- [403] T. Tsujimura, K. Hashimoto, H. Kitayama, H. Ikeda, H. Sugahara, I. Matsumura, T. Kaisho, N. Terada, Y. Kitamura, and Y. Kanakura, "Activating mutation in the catalytic domain of c-kit elicits hematopoietic transformation by receptor self-association not at the ligandinduced dimerization site," *Blood*, vol. 93, pp. 1319–1329, Feb 1999. 136, 177
- [404] D. Larizza, G. Abbati, R. Lorini, A. Salvatoni, and F. Severi, "The Turner phenotype and the different types of human X isochromosome," *Hum. Genet.*, vol. 62, p. 93, 1982.

- [405] X. Zhang, Z. Lin, and I. Kim, "Pax5 expression in non-Hodgkin's lymphomas and acute leukemias," J. Korean Med. Sci., vol. 18, pp. 804–808, Dec 2003. 162
- [406] J. Lejeune, M. Gautier, and R. Turpin, "Étude des chromosomes somatiques de neuf enfants mongoliens," C. R. Hebd. Seances Acad. Sci., vol. 248, pp. 1721–1722, Mar 1959. 187
- [407] C. E. Ford, K. W. Jones, P. E. Polani, J. C. De Almeida, and J. H. Briggs, "A sexchromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome)," *Lancet*, vol. 1, pp. 711–713, Apr 1959. 187
- [408] P. A. Jacobs and J. A. Strong, "A case of human intersexuality having a possible XXY sex-determining mechanism," *Nature*, vol. 183, pp. 302–303, Jan 1959. 187
- [409] J. R. Lupski, R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, D. F. Barker, J. M. Killian, C. A. Garcia, A. Chakravarti, and P. I. Patel, "DNA duplication associated with Charcot-Marie-Tooth disease type 1A," *Cell*, vol. 66, pp. 219–232, Jul 1991. 187
- [410] C. Shaw-Smith, A. M. Pittman, L. Willatt, H. Martin, L. Rickman, S. Gribble, R. Curley, S. Cumming, C. Dunn, D. Kalaitzopoulos, K. Porter, E. Prigmore, A. C. Krepischi-Santos, M. C. Varela, C. P. Koiffmann, A. J. Lees, C. Rosenberg, H. V. Firth, R. de Silva, and N. P. Carter, "Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability," *Nat. Genet.*, vol. 38, pp. 1032–1037, Sep 2006. 187
- [411] R. Visser, O. Shimokawa, N. Harada, A. Kinoshita, T. Ohta, N. Niikawa, and N. Matsumoto, "Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion," *Am. J. Hum. Genet.*, vol. 76, pp. 52–67, Jan 2005.
- [412] C. Lee, A. J. Iafrate, and A. R. Brothman, "Copy number variations and clinical cytogenetic diagnosis of constitutional disorders," *Nat. Genet.*, vol. 39, pp. 48–54, Jul 2007. 187
- [413] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nat. Genet.*, vol. 23, pp. 41–46, Sep 1999. 187
- [414] A. Arslantas, S. Artan, U. Oner, R. Durmaz, H. Muslumanoglu, M. A. Atasoy, N. Baaran, and E. Tel, "Detection of chromosomal imbalances in spinal meningiomas by comparative genomic hybridization," *Neurol. Med. Chir. (Tokyo)*, vol. 43, pp. 12–18, Jan 2003. 187
- [415] M. A. van de Wiel, R. Brosens, P. H. Eilers, C. Kumps, G. A. Meijer, B. Menten, E. Sistermans, F. Speleman, M. E. Timmerman, and B. Ylstra, "Smoothing waves in array CGH tumor profiles," *Bioinformatics*, vol. 25, pp. 1099–1104, May 2009. 188

- [416] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, vol. 10, p. 80, 2009. 188
- [417] J. K. Patel, N. M. Patel, and R. L. Shiyani, "Coefficient of variation in field experiments and yardstick thereof an empirical study," *Current Science*, vol. 81, pp. 1163–1164, 2001. 189
- [418] T. Schilling, A. Gratopp, T. E. DeCoursey, and C. Eder, "Voltage-activated proton currents in human lymphocytes," J. Physiol. (Lond.), vol. 545, pp. 93–105, Nov 2002. 193
- [419] M. Snyder, A. Abyzov, A. Eckehart Urban, and G. M., "Discovering CNVs from read depth analysis of next generation sequencing data." Electronic File (PDF). http://www.jsbi.org/ pdfs/journal1/GIW09/Poster/GIW09P061.pdf. 196
- [420] B. W. Lindgren, Statistical Theory. MacMillan Publishing Company, 3rd ed., July 1976. 252
- [421] R. Hogg and J. Ledolter, *Applied statistics for engineers and physical scientists*. Statistics Mathematics, Macmillan, 1992. 255, 256
- [422] G. A. Churchill, "Using ANOVA to analyze microarray data," *BioTechniques*, vol. 37, pp. 173–175, Aug 2004. 255, 256
- [423] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988. 257, 259
- [424] Y. Gao, L. K. Wolf, and R. M. Georgiadis, "Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison," *Nucleic Acids Research*, vol. 34, pp. 3370–3377, 2006. 258
- [425] E. Carlon, T. Heim, J. K. Wolterink, and G. T. Barkema, "Comment on "Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays"," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 73, Jun 2006. 258
- [426] C. Li and W. Hung Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biol.*, vol. 2, 2001. 266
- [427] C. J. Best, J. W. Gillespie, Y. Yi, G. V. Chandramouli, M. A. Perlmutter, Y. Gathright, H. S. Erickson, L. Georgevich, M. A. Tangrea, P. H. Duray, S. Gonzalez, A. Velasco, W. M. Linehan, R. J. Matusik, D. K. Price, W. D. Figg, M. R. Emmert-Buck, and R. F. Chuaqui, "Molecular alterations in primary prostate cancer after androgen ablation therapy," *Clin. Cancer Res.*, vol. 11, pp. 6823–6834, Oct 2005. 266
- [428] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, July 2003. 274

- [429] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, pp. 589–609, September 1968. 275
- [430] Z. M. Sori and H. A. Jalil, "Financial ratios, discriminant analysis and the prediction of corporate distress," *Journal of Money, Investment and Banking*, no. 11, 2009. 275
- [431] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, pp. 195–200, January 2003. 275
- [432] A. C. Lorena and A. C. Carvalho, "Evaluation of noise reduction techniques in the splice junction recognition problem," *Genetics and Molecular Biology*, vol. 27, no. 4, 2004. 275
- [433] A. Sharma and K. K. Paliwal, "Cancer classification by gradient lda technique using microarray gene expression data," *Data Knowl. Eng.*, vol. 66, no. 2, pp. 338–347, 2008. 275

### Appendix A

## **Probability Distribution Function and Cumulative Distribution Function**

In probability and statistics, the probability density function (PDF) and cumulative density function (CDF) give a complete description of the probability distribution of a random variable.

- probability density function<sup>1</sup> (abbreviated as PDF) of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space (often denoted by f(x)).
- The cumulative distribution function<sup>2</sup> (abbreviated as CDF) of a random variable X evaluated at a number x, is the probability that the random variable X takes on a value less than or equal to x (often denoted by F(x)).

More detailed description of these two functions and their mathematical relationship are given as the following.

**Definition of Probability Distribution Function (PDF):** If X is a continuous random variable the probability of X falling within a given set is given by the integral of its probability distribution function (PDF) over the set. As illustrated in Figure A.1 this probability is equal to the area under the density function within the given range, which is estimated by Equation (A.1).

$$P(a \le x \le b) = \int_{a}^{b} f(x) \, dx \tag{A.1}$$

<sup>&</sup>lt;sup>1</sup>also known as 'probability distribution function' or 'probability mass function'.

<sup>&</sup>lt;sup>2</sup>also known as 'cumulative density function'.





Figure A.1: Probability Density Function (PDF): the probability that X takes on a value in the interval [a,b] is the area under the density function from a to b.

**Definition of Cumulative Density Function (CDF):** For every real number x, the CDF of a real-valued random variable X, which is often denoted by F(x), is given by:

$$x \mapsto F_x(x) = P(X \le x) \tag{A.2}$$

The CDF of continuous random variable X can, therefore, be defined in terms of its PDF (f(x)) by the following equation:

$$F(x) = \int_{-\infty}^{x} f(t)d(t)$$
(A.3)

Based on the above equation, for a given value *x* the CDF (F(x)) is the probability that the observed value of *X* will be at most *x*. The relationship between CDF and PDF is illustrated in Figure A.2.



**Figure A.2: A graphical representation of the relationship between the PDF and CDF.** The probability of *x* being equal or less than 3 ( $P(x \le 3)$  is to the highlighted area under the PDF (P = 0.84), which is equivalent to the CDF of *x* at X = 3 (CDF(x = 3) = 0.84). As shown in the above figure the value of the CDF at *x* is the area under the probability density function up to *x*.

#### **Appendix B**

## Normal and Standard Normal Distributions

Normal distributions are symmetrical, bell-shaped distributions that are useful in describing realworld data (Fig. B.1). The **standard normal distribution**, also represented by Z, is the simplest form of normal distribution that has a mean of 0 ( $\mu = 0$ ) and a variance of 1 ( $\sigma^2 = 1$ ) (the red PDF curve in Figure B.1a depicts a standard normal distribution). The density or **PDF function of a standard normal distribution** (which is often denoted by  $\phi$ ) is given by:

$$f(x) = \phi = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$
(B.1)

By substituting the above density formula in Equation (A.3), the **CDF of standard normal distribution** (which is often shown by  $\Phi(x)$ ) is computed as:

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \phi(x) dt$$
  
=  $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^{2}/2} dt$  (B.2)

Figure B.1b show the CDF of several normal distributions, including the CDF of a standard normal distribution that is shown in red.



Figure B.1: PDF and CDF of normal distribution. (a) The PDF is shown for several normal distributions with varied mean ( $\mu$ ) and standard deviation ( $\sigma$ ). The red line corresponds to the PDF of the standard normal distribution. Similarly panel (b) denotes the CDF of the normal distributions shown in (a), and the CDF of standard normal distribution is show in red color.

**Transforming a Normal Distribution to Standard Normal Distribution:** It is possible to relate all normal random variables to the *standard* normal. For example if X is normal with mean  $\mu$  and variance  $\sigma^2$ ,  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then:

$$Z = \frac{X - \mu}{\sigma} \tag{B.3}$$

where the transformed variable Z has a standard normal distribution (mean 0 and variance 1),  $Z \sim \mathcal{N}(0, 1)$ .

If *X* has a normal distribution, the PDF of *X* can be estimated by first transforming it into a *standard* normal distribution (by Equation (B.3)) and then using Equation (B.1) to estimate the PDF. Thus the **PDF of a normal distribution** can be estimated by:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$
  
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$
(B.4)

$$\Rightarrow f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$
(B.5)

Similarly, by substituting Equation (B.4) in cumulative density equation (Eq. (A.3)) the **CDF of a normal distribution** is defined by:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt$$
 (B.6)

$$\Rightarrow F(x) = \Phi(\frac{x-\mu}{\sigma})$$
(B.7)

**Properties of Normal Distribution:** Some of the properties of normal distribution which have been used in this thesis are as the following:

1. If *X* is normally distributed with mean  $\mu$  and variance  $\sigma^2$ , then a linear transform aX + b (for some real numbers  $a \neq 0$ ) is also normally distributed:

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

2. If  $X_1$ ,  $X_2$  are two independent normal random variables, with means  $\mu_1$ ,  $\mu_2$  and standard deviations  $\sigma_1$ ,  $\sigma_2$ , then their linear combination will also be normally distributed:

$$aX_1+bX_2\sim \mathcal{N}(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2)$$

3. The converse of the above property (2) is also true: if  $X_1$  and  $X_2$  are independent and their sum  $X_1 + X_2$  is distributed normally, then both  $X_1$  and  $X_2$  must also be normal.

### Appendix C

# **Proof of the Relationship Between** p(k)and **CV**

Let *X* represent the array raw intensity values from a test sample ( $\mathscr{T}$ ) that follow an unknown distribution with mean  $\mu$ , variance  $\sigma^2$  and CV of v ( $v = \sigma/\mu$ ). Assume that after log-transformation,  $\log_b(X)$  is normally distributed with mean *m* and variance of  $s^2$ , where *b* is the base of the logarithm function. Therefore:

$$\log_b(x) \sim \mathcal{N}(m, s^2) \tag{C.1}$$

Lindgren [420] has previously shown that the relationships between the mean and standard deviation of a distribution  $(\mu, \sigma)$  and its log-transformation (m, s) is defined by:

$$\mu = \exp\left\{m\log_e(b) + \frac{s^2}{2}[\log_e(b)]^2\right\}$$
(C.2a)

$$\sigma^{2} = \exp\left\{2m\log_{e}(b) + s^{2}[\log_{e}(b)]^{2}\right\} \left[\exp\{s^{2}[\log_{e}(b)]^{2}\} - 1\right]$$
(C.2b)

If natural logarithm is used to transfer the data (b = e), equations (C.2a) and (C.2b) would become:

$$\mu = \exp\{m + \frac{s^2}{2}\}$$
$$\sigma^2 = \exp\{2m + s^2\}(\exp s^2 - 1)$$

Therefore the coefficient of variation (CV) will be:

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\exp\{2m + s^2\}\{\exp s^2 - 1\}}}{\exp\{m + s^2/2\}}$$
  
=  $\sqrt{\exp s^2 - 1}$  (C.3)

where *s* is the standard deviation of the errors (normally distributed). Solving the above equation for *s* gives:

$$s = \sqrt{\log_e(\mathrm{CV}^2 + 1)} \tag{C.4}$$

Next, let assume that an additional independent experiment is performed using the same test sample  $\mathscr{T}$ , and lets refer to the results of this second assay by random variable *Y*. Therefore, *X* and *Y* are two independent random variables from a replicate study (using sample  $\mathscr{T}$ ). Similarly, we assume that *Y* would also follow a normal distribution after log-transformation with mean *m* and variance  $s^2$  ( $Y \sim \mathscr{N}(m, s^2)$ ). As the result of this hypothesis, the probability that the readout from these two replicate arrays differ by *k* fold or more is:

$$p(k) = P(Y|X \ge k \text{ or } X|Y \ge k) = 2P(Y|X \ge k)$$
(C.5)

where the right-hand side represents the probability that the ratio of the two arrays (X/Y) takes on a value greater than or equal to k. This probability implies :

$$Y/X \ge k \to \boxed{\log_b(X) - \log_b(Y) \ge \log_b(k)} \tag{C.6}$$

The left hand side of (C.6) is also normally distributed with mean equal to  $(\mu_x - \mu_y)$  and variance of  $(\sigma_x^2 + \sigma_y^2)$ . Since  $\mu_x = \mu_y = m$  and  $\sigma_x = \sigma_y = s$ , we have:

$$\log_{b}(X) - \log_{b}(Y) \sim N(0, 2s^{2})$$

$$\Rightarrow P(\log_{b}(X) - \log_{b}(Y) \ge \log_{b}(k))$$

$$\Rightarrow 1 - \Phi\left(\frac{\log_{b}(k) - 0}{\sqrt{2s^{2}}}\right) = \Phi\left(\frac{-\log_{b}(k)}{\sqrt{2s}}\right) = P(Y/X) \quad (C.7)$$

Substituting the above estimated P(Y/X) (Eq. C.7) in Equation (C.5) results in:

$$p(k) = 2\Phi\left[\frac{-\log_b(k)}{s\sqrt{2}}\right] \tag{C.8}$$

By substituting the formula of s as denoted in Equation (C.4), Equation (C.8) would become:

$$p(k) = 2\Phi\left[\frac{-\log_e^2(k)}{\sqrt{2\log_e(\mathrm{CV}^2+1)}}\right]$$
(C.9)

Equation (C.9) measures the probability that two replicate measurements in an experiment with a specific CV, differ by k-fold.

#### **Appendix D**

## **Estimating SNP-array Reproducibility Using Analysis of Variance (ANOVA)**

In general, the aim of analysis of variance (ANOVA) is to find out whether there is a significant difference in the means of the tested data from several groups. The one-way ANOVA test measures significant effects of only one independent variable (or factor) on the estimated response, whereas, the two-way ANOVA measures the effect of two factors, simultaneously [421, 422]. In order to evaluate labelling and chip variabilities of each DNA sample, two-way ANOVA model was applied on the log10-transformed intensity of each SNP on the 10K Affymetrix SNP array, respectively. The results of ANOVA analysis conducted in this section are used to address the following tasks: (1) to evaluate whether chip or labeling factors have main effects on the estimated SNP signal, (2) to find the number of SNPs that show statistically significant differences across replicate measurements due to the aforementioned factors (i.e., labeling or chip variabilities), and (3) to determine whether a subset of these SNPs show recurrent inconsistencies. In general, these analyses help to assess the impact of both chip and labeling factors on the estimated signal from Affymetrix SNP arrays.

Figure D.1 denotes a schematic representation of the input data (X) of the two-way ANOVA test for a given SNP, such as SNP<sub>a</sub>. The columns of the input matrix X (Figure D.1) represent the estimated SNP signal from 3 separate labelled batches of the same sample (factor A of two-way ANOVA). The data in different rows represent SNP signal from replicate chips<sup>1</sup> (factor B of two-way ANOVA). Here, the p-value of two-way ANOVA includes two components:

<sup>&</sup>lt;sup>1</sup>as previously described in Section 2.3.1 and Figure 2.5, each labelled DNA batch was applied on 3 replicate 10K chips.

- 1.  $p_A$  is p-value of the null-hypothesis  $H_{0A}$  that all samples from factor A (column data) are drawn from the same population. For example, in Figure D.1  $p_A$  is the p-value of the null-hypothesis that the replicate signal intensity readouts of SNP<sub>a</sub> from 3 separate labeling reactions (columns of matrix X) have the same overall mean.
- 2.  $p_B$  is the p-value of the null-hypothesis  $H_{0B}$  that all samples from factor B (row data) are drawn from the same population. In Figure D.1,  $p_B$  is the p-value of the null-hypothesis that signal intensity measurements of SNP<sub>a</sub> from 3 replicate chips (rows of matrix X) have the same overall mean.

A sufficiently small  $p_A$  rejects  $H_{0A}$  null hypothesis, suggesting that at least one column-sample mean is significantly different from the others. For instance, in Figure D.1 a sufficiently small  $p_A$  (e.g.,  $p_A < 0.05$  or  $p_A < 0.01$ ) suggests that the estimated SNP<sub>a</sub> signal in at least one of the three labelled batches is significantly different from the other batches of the same sample (i.e., there is a main inconsistency in the pattern of SNP<sub>a</sub> intensities across replicate measurements due to labeling variability) [421, 422]. A sufficiently small  $p_B$  suggests that at least one row-sample mean is significantly different from the other row-sample means [421, 422]. In the context of the ANOVA test illustrated in Figure D.1, a sufficiently small  $p_B$  (e.g.,  $p_B < 0.05$ ) suggests that the estimated SNP<sub>a</sub> signal intensity is significantly different in at least one of the three replicate chips (i.e., there is a main inconsistency in the pattern of SNP<sub>a</sub> intensities across replicate measurements that is due to chip-to-chip variability). In the rest of this Section, SNPs with significant chip or labeling variabilities across replicate measurements are referred to as "inconsistent SNPs".

#### Results

As explained in Section D, to assess chip and labeling variabilities two-way ANOVA was performed on each SNP on Affymetrix 10K SNP array. This experiment was repeated for each DNA sample from "10K rep-test" experiment that was previously described in Section 2.3.1. Therefore, the total of 92,480 (=  $11560 \times 8$ )  $p_A$  and 92,480  $p_B$  values were generated to assess the impact of labeling and chip variabilities among 8 DNA samples.

Prior to ANOVA analysis, log-transformation was applied on the raw intensity data from 10K Affymetrix arrays (oligo-level) and global normalization was performed to bring the overall (oligo-level) mean signal intensities from all the arrays to the same level. The signal intensity of each SNP was then defined as the average log10-intensity measurement of 40 oligos in the corresponding SNP probe set. Next, two-way ANOVA was applied for each SNP on the 10K array, as shown

in Figure D.1. The ANOVA p-values were then adjusted for multiple testing correction based on Hochberg method [423] with FDR of 5% (significance threshold= 0.05). Figure D.2 and Table D.1 present the summary of two-way ANOVA tests that were performed to analyse 8 DNA samples in the replicate experiment. As denoted in Table D.1 (column 2) and Figure D.2a, the average p-value of the two-way ANOVA tests of chip variability across 8 samples is  $\overline{p_B} \simeq 0.98$ , ranging between 0.96 to 0.99 (the mean  $p_B$  values across all 8 samples is denoted by the green dashed line in Fig. D.2a). It is also observed that the average p-value of labeling variability across all 8 samples is  $\overline{p_A} \simeq 0.97$  (denoted by the red dashed line in Fig. D.2a), ranging between 0.941 to 0.998 (column 5 of Table D.1).

Therefore, these data indicate that both labeling and chip related ANOVA p-values are noticeably larger than 0.05 significance threshold that is required to reject the  $H_{0A}$  and  $H_{0B}$  hypotheses<sup>1</sup> ( $p_A \ge 0.941$  and  $p_B \ge 0.962$ ).

As described earlier (p. 255), a main reason of ANOVA analysis of replicate data in this Section was to assess the number of SNPs that indicate statistically significant variabilities across replicate measurements (task 2; see page 255). The summary of this analysis (based on 0.05 significance threshold) is shown in Table D.1 and Figure D.2b. As seen in Table D.1, in average ~1.5% (0.6-3.8%) of all SNPs on the 10K array indicate statistically significant chip-related (column 4) and ~3.14% (0.2-5.9%) show statistically significant labeling-related (column 7) differences across replicate measurements in each studied DNA sample. The bars in Figure D.2b provide a visual representation of the proportion of 10K SNPs that show inconsistent readouts due to chip (green bars) or labeling (red bars) effects in each DNA sample.

Next, a further analysis was performed to assess the frequency of inconsistent SNPs for both labeling and chip effects (task 3; see page 255). The results of this analysis are depicted in Figure D.3. Here, the frequency of each inconsistent SNP, referred to as f, is defined as the total number of samples where the SNP indicate inconsistent intensity measurements across replicate data (of the corresponding DNA sample). Therefore, in theory, the value of f in "10K rep-test" experiment can vary between 1 to 8. However, it is later shown that the empirical value of f does not exceed 4 in this experiment.

The above analysis found that there are in total 2,440 unique SNPs that indicate statistically significant variabilities across replicate arrays due to labeling effects (Figure D.3.A). Further analysis of these 2,440 inconsistent SNPs found that 409 (16.7%) of them show inconsistency in at least 2 samples (equivalent to  $\sim$ 3.5% of all 10K SNPs). However, none of the SNPs has a frequency

<sup>&</sup>lt;sup>1</sup>It is observed that the mean p-value of labeling-related variability is just marginally less than chip variability (average p-values of chip and labeling variabilities are 0.98 and 0.97, respectively).

greater than four and only 2 SNPs show inconsistency in four independent samples. Similar analysis was performed to study the frequency of SNPs with chip related inconsistencies across "10K rep-test" samples. As shown in the pie chart of Figure D.3.B, in total, there are 1,327 SNPs with statistically significant chip-related inconsistencies among 8 studied samples, from which 92 (~7%) show inconsistencies in at least 2 samples ( $f \ge 2$ ). From these 92 chip-related inconsistent SNPs, only 1 inconsistent SNP was found in four separate samples (see red arrow in Fig. D.3.B). Similar to the labeling analysis, none of the SNPs that showed significant differences due to chip effects had f > 4.

The analysis of the frequency of inconsistencies across 10K SNPs indicated that > 98% of labeling-related and > 99% of chip-related inconsistent SNPs have  $f \le 2$  (Figure D.3). The latter finding implies that SNPs that show statistically significant inconsistencies across replicate measurements in "10K rep-test" experiment are generally observed either once or twice.

In the final part of the analysis, all SNPs that indicated labeling and chip related inconsistencies in at least 2 samples ( $f \ge 2$ ; 409 SNPs with labeling-related (Fig. D.3.A) and 92 SNPs with chiprelated (Fig. D.3.B) inconsistencies) were compared to assess whether a subset of these SNPs are observed in both datasets. The intersection of these two sets found that based on 0.05 significance level there are only 6 SNPs (6/11560 = 5.19e-04) that are inconsistent across both labeling and chip replicates in at least 2 independent samples of "10K rep-test" experiment. Therefore, it can be concluded that the hybridization intensity of these 6 SNPs exhibit a large degree of variability in both chip and labeling replicate measurements. This result is expected since other groups have shown that probe level variability in hybridization based technologies, is in part caused by experimental factors and in part determined by sequence properties of the probe (e.g., sequence affinities and DNA secondary structure) [424, 425]<sup>1</sup>.

In conclusion, similar to the result of analysing Coefficient of Variability (CV) in "10K rep-test" experiment, the ANOVA analysis also found that the data from Affymetrix GeneChip<sup>®</sup> 10K SNP arrays are highly reproducible and less than 6% of all SNPs on the 10K array indicate statistically significant differences across replicate measurements at FDR of 5% (Table D.1).

<sup>&</sup>lt;sup>1</sup>Analysis of the details of sequence based effects on the detected variability is beyond the scope of this thesis. However, later in Chapter 3, I will show examples of the impact of the length of PCR products on the efficiency of amplification and the resultant nonspecific variation of signal intensities.

	Chip Variability			La	Labeling Variability		
Sample	adjusted $\overline{p_B}^{\dagger}$	# inconsistent SNPs (N)	%	adjusted $\overline{p_A}^{\dagger}$	<pre># inconsistent SNPs</pre>	%	
1	0.969	354	3.1	0.998	19	0.2	
2	0.991	108	0.9	0.983	199	1.7	
3	0.991	106	0.9	0.947	614	5.3	
4	0.993	84	0.7	0.941	677	5.9	
5	0.988	139	1.2	0.944	653	5.6	
6	0.962	437	3.8	0.959	475	4.1	
7	0.994	72	0.6	0.990	113	1.0	
8	0.989	128	1.1	0.987	146	1.3	
average	0.98	178.5	1.5	0.97	362	3.14	
min	0.962	72	0.6	0.941	19	0.20	
max	0.994	437	3.8	0.998	677	5.90	

ANOVA results of chip and labeling variability

† based on Hochberg multiple testing correction

Table D.1: Results of assessing chip and labeling variability in 8 samples based on two-way ANOVA analysis. The rows illustrate the summary of ANOVA tests for all SNPs on the 10K array for each separate DNA sample in "10K rep-test" experiment. The 2nd column of this table indicates the average p-values ( $\overline{p_B}$ ) of 11560 SNPs on the 10K array for chip-related variabilities, after correcting for multiple testing based on Hochberg method [423]. The number of SNPs that have significantly small  $\overline{p_B}$  values based on FDR=5% ( $p_B < 0.05$ ) are reported in the 3rd column (N) of this table (i.e., there is a main effect on the estimated signal intensities of these SNPs due to chip-specific variabilities and these SNPs are considered to have inconsistent readouts). The 4th column of this table indicates the proportion of 10K SNPs that show inconsistencies due to chip effects (= (N/11560) × 100). The next 3 columns show the results of labeling variability for each tested DNA sample (8). As depicted in Figure D.1, the p-value of labeling variability ( $p_A$ ) of each SNP corresponds to the column p-value of the two-way ANOVA test of input matrix data X. Next the average p-values of all SNPs, number of inconsistent SNPs and percentage of inconsistent SNPs on the 10K array were estimated and shown in column 5 to 7, respectively.

The results of this table indicate that based on significance threshold of 0.05, in average 1.5% of all 10K SNPs show significant differences in their estimated hybridization intensities across replicate measurements due to chip variability ( $p_B < 0.05$ ), and ~3% of all 10K SNPs show such statistically significant differences due to sample preparation or labeling effects ( $p_B < 0.05$ ).



Figure D.1: Schematic representation of two-way ANOVA test performed on each 10K SNP. This figure represents a schematic representation of the input data to two-way ANOVA test that is performed on a each SNP on the 10K array (e.g.,  $SNP_a$ ). The columns of the input data X represent the estimated signal of  $SNP_a$  from 3 separate labelled batches of the same sample (i.e., factor A of two-way ANOVA). The data in the rows of X indicate the measured  $SNP_a$  signal from 3 replicate chips of the same labelled DNA batch (i.e., factor B of two-way ANOVA).



(a) p-values of two-way ANOVA tests per DNA sample





Figure D.2: Measuring the effect of chip and labeling variability in 10K replicate experiment using two-way ANOVA analysis. Panel (a) shows the average of 11,560 p-values for all SNPs on the 10K array. The green and red dots indicate the average p-values for labeling variability ( $\overline{p_A}$ ) and chip variability ( $\overline{p_B}$ ), respectively. This graph indicates that the average p-values are > 0.9 in all 8 DNA samples, which is noticeably larger than 0.05 significance level. Another observation is that the labeling-related p-values are marginally smaller than p-values of chip variability (the only exception is sample #1 where labeling variability is greater than chip variability. However, as previously mentioned, only 2 labeling batches were available for this particular sample (instead of 3). Thus, the p-values related to labeling effects for sample #1 are not as reliable as the other samples).

Panel (b) shows the proportion of SNPs on the 10K array that show inconsistent patterns of signal intensity across replicate measurements. The SNPs that exhibit inconsistencies related to chip and labeling variabilities are shown by the green and red bars, respectively. This figure indicates that in overall < 6% of all SNPs on the 10K array indicate statistically significant differences across replicate measurements at FDR=5%.



#### Frequency of SNPs that showed inconsistent intensity patterns

Figure D.3: Frequency of inconsistent SNPs across 8 studied DNA samples. The above pie charts depict the frequency (f) of SNPs that have inconsistent readouts in replicate arrays due to labeling (panel A) or chip (panel B) effects across DNA samples. The denoted percentage data represents the fraction of inconsistent SNPs with the specified frequency with respect to the total number of inconsistent SNPs in the corresponding category. For example, in panel (B) there are 84 inconsistent SNPs that occur in 2/8 DNA samples (f = 2; pink region) which accounts for 6.3% (= 84/1327) of all SNPs that indicate chip-related inconsistencies in panel B (see \*).

As seen, in both (A) and (B) the majority of inconsistent SNPs are detected only once (f = 1; blue regions). Furthermore, the red arrow denote that only 1-2 SNPs from all SNPs on the 10K array (11,560) show statistically significant variabilities in 4/8 studied samples (f = 4), and none of the inconsistent SNPs had f > 4.

Therefore, this analysis indicate that < 1% (92/11560 = 0.008  $\approx$  0.8%) of all SNPs on the 10K array have inconsistent patterns of intensity across replicate measurements with a frequency (*f*) greater than 1. Similar analysis found that the inconsistent SNPs due to labeling variability that have f > 1 account for ~3.5% (409/11560 = 0.035) of all 10K SNPs.

Further analysis of a subset of inconsistent inconsistent SNPs from 'A' and 'B' that have  $f \ge 2$  found that only 6 SNPs are common among these two groups.

#### **Appendix E**

### **Description of Boxplots (in the thesis)**



Figure E.1: Boxplot and a probability density function (PDF) of a dataset with standard normal distribution  $(Z \sim \mathcal{N}(0, 1))$ 

On each box, the central red line is the 50th percentile (median), the edges of the box are the 25th and the 75th percentiles. The whiskers extend to the most extreme data value that is not an outlier.

Points are drawn as outliers if they are larger than 1.5 interquartile range (IQR) of the upper quartile Q3, or smaller than  $1.5 \times IQR$  of the lower quantile Q1. Thus, data point *x* is an outlier if:

$$x > Q3 + 1.5 \times (Q3 - Q1)$$
 or (E.1)  
 $x < Q1 - 1.5 \times (Q3 - Q1)$ 

where Q1 and Q3 are the 25th and 75th percentiles of the data, respectively. The default of 1.5 corresponds to approximately  $\pm 2.7\sigma$  and 99.3% coverage if the data are normally distributed. Each outlier data point is plotted with a red '+' marker.
### Appendix F

# **Comparative Analysis of the SNP Array Normalization Techniques**

#### F.1 Introduction

To study the impact of between-array normalization on Affymetrix probe level data several methods were applied to normalize raw Nsp array data from 8 follicular lymphoma (FL) cancer patients. These samples have 50 Illumina sequence-validated regions of copy number loss, ranging in size between 9 bp to 1.7 Mb (with average size of 439.8 kb). The aforementioned 50 copy number deletions were initially identified by fingerprint profiling (FPP) of these patients, as described in Chapter 4. The Illumina sequencing of the paired tumor/normal samples determined both the origin of these events (all 50 aforementioned events are somatic) and the exact deletion breakpoints. The sequence-validated breakpoints were then used to locate the Nsp SNP probes that lied within the exact deletion boundaries in each sample.

The aim of the analysis described in this section was to determine which normalization method improves the quality of Affymetrix array data the most. The specific aim of this study was to investigate whether quantile normalization, which is the default normalization method of OPAS, suppresses the magnitude of CNV aberrations in cancer samples.

#### F.2 Method and Results

For each tumor/normal pairs, five methods were applied to normalize the log2-ratio Nsp array data. Therefore, a total of 40 (=  $8 \times 5$ ) datasets were generated for this comparative analysis,

each consisting of more than 3 million PM log2-ratio intensity values. The normalization methods used in this analysis include (1) global mean normalization [204, 205], (2) median scaling normalization [204, 205], (3) Median Absolute Distance normalization (MAD) [426, 427], (4) quantile normalization and (5) background adjusted quantile normalization [336] (OPAS default).

It must be added that the ability to identify regions of real copy number change not only depends on the estimated LR readouts of the probes within the CNV regions, but also on the magnitude of their LR deviation from the array's baseline signal. The underlying assumption of this analysis is that in each sample the LR distribution of the PM oligos within the validated deleted regions (D) is smaller than the distribution of the PM oligos from the rest of the array  $(BL)^1$ . Therefore, a normalization technique that leads to a wider separation between these two distributions ( $\Delta = D - BL$ ) improves the LR magnitudes real deletions.

A comparison of the results of the normalization methods is shown in Figure F.1 and Table F.2. For instance, in Figure F.1(a) the top panel illustrates the distribution of 228 PM oligos from 19 Nsp SNPs that fall within 10 sequence-validated deletions of FL patient 6 (ht-06). The five boxplots in the top figure present the distributions of the normalized log2-ratio readouts of these deleted oligos (228) based on the aforementioned 5 normalization methods, respectively. As seen in this plot (top panel of Fig. F.1(a)), global normalization and MAD both result in slight increase in the signal from the deleted oligos in ht-06 compared to the theoretical copy number baseline of zero ( $D_1 = +0.2$ ,  $D_3 = +0.46$ ). It is also observed that there is no significant change in the signal intensities of the deleted oligos as the result of median scaling normalization ( $D_2 = 0.01$ ). However, as seen in this boxplot, both quantile and background adjusted quantile normalization methods result in the decrease of the signal intensity of deleted oligos compared to the theoretical copy number baseline of zero ( $D_4 = -0.15$  and  $D_4 = -0.18$ ). The distributions of more than 3.03 million PM oligos that are located on the rest of the genome of ht-06 (empirical baseline, BL) are presented in the middle panel of Figure F.1(a). The distance between D and BL oligos in this patient is shown in the bottom plot of this figure (also detailed in the first rows of Table F.2).

The data presented in Figure F.1 and Table F.2 reveal that background adjusted quantile normalization leads to the largest magnitude of copy number loss in these 8 FL cancer patients. Another observation from this analysis is that in some cases the magnitude of copy number loss signal is significantly smaller than the theoretical LR value of one copy loss (-0.58) regardless of the normalization algorithm used. Such examples include the deleted regions of ht-06 (min( $\Delta$ ) = -0.15), ht-20 (min( $\Delta$ ) = -0.3) and ht-25 (min( $\Delta$ ) = -0.32). This observation points out to a major com-

<sup>&</sup>lt;sup>1</sup>This assumption cannot hold true if most of the patient's genome is deleted. The list of chromosomal gains/losses in these FL patients, presented in Table F.1, suggests that it is safe to assume D < BL in this experiment.

plexity of CNV analysis in cancer samples, emphasizing that detecting smaller magnitudes of signal aberration that represent real copy number changes in cancer genomes is not necessarily a downside of the choice of normalization techniques. As discussed in Chapter 4, some of the main reasons of observing lower magnitudes of copy number aberration in cancer samples may include CNV heterogeneity of cancer cells and also sample heterogeneity (admixed cancer and normal cells), both of which are common challenges in cancer CNV analysis.

In conclusion, despite the fact that using a quantile normalization approach poses the risk of removing some of the signals in the tails of the distribution, the empirical evidence presented in this analysis showed that such problem does not exist in practice (at least in the case of follicular lymphoma samples analysed in this study). It must be added that follicular lymphoma is one of the highly unstable cancer genomes, where DNA is often affected by large-scale copy number alterations (see Table F.1). Therefore, the absence of any apparent downside to using quantile method for normalizing FL data suggests that it is unlikely that quantile normalization would mitigate the data quality in other cancers. The result of this experiment is consistent with the findings of several other microarray studies that have shown empirical evidence does not indicate quantile normalization leads to problems in practise [205, 302].Nonetheless, all of the above 5 normalization techniques have been added to the new version of OPAS, and the user can change the default quantile normalization to any of the other 4 methods or replace it by a new user-defined normalization function.



1: Global normalization (mean scaling)

2: Median scaling normalization3: Median Absolute Deviation normalization (MAD)

3: Quantile normalization

4: Background-adjusted quantile normalization









Each bar in this boxplot, marked by 1 to 5, depicts the distribution of the deleted oligos using a separate normalization method, as detailed in the figure legend. The middle boxplots show the LR distributions of all of the remaining PM oligos (*BL*) based on the aforementioned five normalization techniques. The bottom panel shows the median of the deleted (*D*) and baseline (*BL*) populations, denoted by red and blue dashed lines respectively. The distance between these two lines ( $\Delta$ ) is reported in Table F.2.

Sample	Gains	Losses				
ht-06	none	1p36.2-p36.3, 14q32.33				
ht-08	+18 (18p11-q21), +21q, +Xp, 1q21.3, 14q32.33	1p22.3, 2p11.2, 3p21.31, 5q23.1, 7p21.3, 10q23-q25, 13q33, 14q32.33, 22q11.22				
ht-13	none	del(10)(q?22q24), 3q26, 3q27, 10q23-q25				
ht-20	3p14.2	6q11.2-q15, 7q36, 14q32.33, Xp22.33, 1p36.2-p36.3				
ht-21	+X, +18, +17, +11p, dup(17)(q11q23), der(1), dup(2)(p12p24.1),+der(11)t(1;11;3)(q23.2;q13.1;p26), 1q23-q44	1p36.2-p36.3, 2p11.2, 6q15.1-q27, 14q32.33				
ht-22	+2, add(22)(q13) 7q21.13, 16p11.2, 16p12.1,17q none	-X, -11 7q21.13, 22q13.2-q13.33, Xq28, 1p36.2-p36.3, 14q32.33				
ht-24	+1q, +6p, +7, +9q, +X 1q21.3-q44, 8q21.2-q24.3, 14q32.33, 17q23.3	-6q 1p36.2-p36.3, 8p12-p23.3, 9p242-p24.3, 9p23, 14q32.33, 17p13.3, 18q21.33-18q23, 22q11.22				
ht-25	+12, 6p21.3,7p14.1,7q34, 8q24.12-q24.3, 14q11.2, 19p13.2	14q32.3				

**Table F.1: List of CNVs and large-scale copy number alteration in 8 studied follicular lymphoma patients.** This table presents all CNVs and chromosomal alterations that were detected in these patients based on cytogenetics and BAC array CGH analyses (for more information please see Chapter 4). As seen here, these FL genomes include multiple aneuploidies (e.g., whole chromosome 2 gain in ht-22), large-scale gains and losses of entire chromosome arms (e.g., deletion of the long arm of chromosome 6 in ht-24), as well as relatively smaller CNVs that affect specific chromosomal regions (e.g., 1p22.3 deletion in ht-08).

					<b>Difference Between</b> $D$ and $BL(\Delta)$				
Patient	Ν	Size (kb)	# PM	# SNPs	$\Delta_1$	$\Delta_2$	$\Delta_3$	$\Delta_4$	$\Delta_5$
ht-06	10	112.4	228	19	0.06	0.06	0.06	-0.13	-0.15
ht-08	15	1748	27,204	2,267	-0.31	-0.31	-0.31	-0.29	-0.42
ht-13	1	62.7	12	1	-0.8	-0.8	-0.8	-0.79	-0.95
ht-20	2	540	480	40	-0.21	-0.21	-0.21	-0.21	-0.3
ht-21	8	93.6	528	44	-0.3	-0.3	-0.3	-0.32	-0.45
ht-22	4	520.8	2,748	229	-0.36	-0.36	-0.36	-0.35	-0.53
ht-24	3	209.2	216	18	-0.5	-0.5	-0.5	-0.5	-0.8
ht-25	7	231.4	888	74	-0.2	-0.2	-0.2	-0.2	-0.32
Ave.					-0.32	-0.32	-0.32	-0.34	-0.49

1: Global mean normalization

2: Median scaling

3: Median Absolute Difference (MAD)

4: Quantile normalization

5: Background adjusted quantile normalization

Table F.2: Comparing the results of normalization methods in 8 cancer samples (FL). The first 3 columns indicate the sample id, number of sequence-validated deletions in the sample (N) and the average size of these deletions. Columns 4-5 denote the number of PM oligonucleotide probes and the number of Nsp SNPs that fall within the Illumina sequence-validated deletion breakpoints. The last 5 columns of this table represent the difference between the median LR values of the deleted PM oligos and those form the rest of the genome ( $\Delta = D - BL$ ). Numbers 1 to 5 at the top of these columns refer to the methods that were used for between-array normalization, as detailed in the table legend.

## Appendix G

## **Description of QDA**

According to Bayes theorem, the knowledge of class posteriors Pr(G|c) can help us to perform an optimal classification [428]. Suppose that X represents the data that we wish to classify into k possible classes, that for convenience are labeled  $G = \{1, 2, ..., K\}$ . The Bayes rule is given by:

$$P(G=k|X=c) = \frac{P(c|k) P(G=k)}{P(X=c)}$$

Which can be rewritten as:

$$P(G = k | X = c) = \frac{f_k(c)\pi_k}{\sum_{l=1}^{K} f_l(c)\pi_l}$$
(G.1)

Where  $f_k(c)$  indicates the class-conditional probability that oligo cluster c belongs to class k and  $\pi_k$  represents the prior probability of class k (P(G = k)). Many classifier techniques are based on models for solving the above class densities, including Naive Bayes and discriminant analysis methods [428]. The main difference between these models is their specific approach towards solving Equation (G.1). In general, discriminant analysis assumes that each class density,  $f_k(c)$ , follows a multivariate Gaussian distribution defined by:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k))$$
(G.2)

Two of the most common types of discriminant analysis classifiers are linear and quadratic discriminant analyses. Linear Discriminant Analysis (LDA) assumes that the underlying classes have a common covariance matrix (all  $\Sigma_k$  values are equal) and utilizes this assumption to solve Equation (G.2). In contrast, Quadratic Discriminant Analysis (QDA) rejects a common covariance assumption and solves Equation G.2 by assigning a separate covariance matrix to each class ( $\Sigma_k$ ). Both LDA and QDA are widely used with high success rates in a diverse range of applications<sup>1</sup>, including economics [429, 430], face recognition [431], splice junction recognition [432] and cancer prediction models based on gene expression data [433].

<sup>&</sup>lt;sup>1</sup>http://www.is.umk.pl/projects/datasets-stat.html

### **Appendix H**

## **Measures of Predicting the Accuracy**

- **true positive (TP)**: simulated CNV regions that are also detected by a CNV calling algorithm.
- **false positive (FP)**: predicted CNVs that do not correspond to a pre-defined simulated CNV region.
- false negative (FN): is ambiguous in the context of comparing CNV-calling algorithms. In theory, every single SNP probe that does not fall within a pre-defined simulated CNV region and does not fall within a predicted CNV call (by the algorithm) is a false negative (FN).

Sensitivity and Specificity are the two most widely used measures of accuracy. These parameters are usually defined by :

$$TPR = \frac{TP}{(TP+FN)}$$
 (sensitivity or true positive rate) (H.1)

$$SPC = \frac{TN}{(FP+TN)}$$
 (specificity or true negative rate) (H.2)

In the context of simulated CNV analysis, sensitivity (TPR) is the proportion of simulated CNV regions that have been correctly predicted as a copy number aberration, and specificity is the proportion of normal SNPs (SNPs that are located in non-CNV regions) that have been correctly predicted as normal. However, since the frequency of normal SNPs in a high-resolution array experiment is much greater than the frequency of copy number aberrated SNPs, sensitivity tends

to be much larger than specificity, and thus SPC, as computed above, systematically produces very large non-informative values. For instance in a 250K array, even if the algorithm finds 10,000 false positive SNPs (FP), there would still be ~240,000 SNPs normal probes (true negatives; TN). Therefore, the estimated specificity would be close to 1 [214], despite the obvious abundance of false positive calls. The impact of this drawback is observed in several papers that compare CNV analysis methods, both for SNP arrays and array CGH platforms, [214, 232]. As the result of this problem, the ROC curves (sensitivity plotted versus specificity) tend to quickly converge towards 100% specificity. Thus, sensitivity overshadows specificity and the true extent of false positive calls is not clearly understood by using the above approach (ROC curves that use specificity).

To circumvent this problem, in Section 3.3.6 of this thesis the impact of false positive calls in simulated data analysis is computed as:

$$PPV = \frac{TP}{(FP+TP)} \qquad \text{(precision or positive predictive value)} \tag{H.3}$$

where PPV is the proportion of predicted CNV calls that are actually part of predefined simulated CNV regions. The estimated PPV value represents the 'precision' of a CNV calling algorithm This value not only depends on the true positive CNV calls (TP), but also the extent of the false positives (FP). Thus, if an algorithm finds numerous false positive calls in addition to all known true positives, it would have a high sensitivity (SPC) but a poor precision (PPV). In contrast, if an algorithm does not find all simulated CNVs (TP) but, at the same time, does not find many other false positives either, it would have a low sensitivity but a high precision. Therefore, analysing the sensitivity and precision at the same time can provide an insight into the performance of an algorithm with respect to both true positive and false positive calls.

## Appendix I

# List of Validated CNVs in 146 MR Patients

			Friedman	et al 2006 and 20	009 (SMD)	OPAS Result				
Chr	Del/Amp	Cytoband	Start	End	Size	Start	End	Size	#SNPs	LR
1	Deletion	1p36.32-p36.33	769,185	3,581,308	769,185	775,852	3,684,971	2,909,119	56	-0.42
2	Deletion	2p16.3	50,829,675	51,120,302	290,627	50,829,963	51,130,072	300,109	35	-0.42
2	Duplication	2q37	231,577,285	242,663,303	11,086,018	231,577,285	242,650,580	11,073,295	756	0.23
4	Deletion	4p16.1-p16.3	13,255	8,472,657	8,459,402	19,099	9,057,838	9,038,739	649	-0.41
4	Deletion	4p16.3	190,631	3,277,436	3,086,805	344,051	3,339,443	2,995,392	150	-0.35
4	Deletion	4p16.3	1,346,924	2,846,261	1,499,337	1,324,721	2,876,251	1,551,530	62	-0.43
6	Deletion	6p21.33	29,937,087	30,026,517	89,430	30,024,232	30,045,241	21,009	2	-0.95
6	Deletion	6 q21-q22.31	111,979,175	121,506,916	9,527,741	111,895,841	121,703,987	9,808,146	941	-0.40
7	Deletion	7p22.2-p22.1	3,657,805	6,165,597	2,507,792	3,576,829	7,089,167	3,512,338	210	-0.37
7	Deletion	7p15.3	14,141,506	24,950,414	10,808,908	14,154,194	25,014,570	10,860,376	1341	-0.32
7	Deletion	7q22.1	98,211,585	100,553,755	2,342,170	98,318,717	100,719,050	2,400,333	67	-0.33
8	Duplication	8p23.1-p23.3	180,568	6,898,076	6,717,508	180,568	6,986,630	2,038,410	224	0.27
8	Duplication	8 q12	58,388,614	65,306,097	6,917,483	58,377,571	65,346,519	6,968,948	681	0.25
8	Duplication	8q23.2-q23.3	111,442,951	113,003,770	1,560,819	111398963	113,077,495	1,678,532	128	0.17
9	Deletion	9p11.2-p13.3	33,702,471	44,744,675	11,042,204	33,396,609	66,273,146	32,876,537	356	-0.40
9	Deletion	9 p13.3	34,144,847	38,736,451	4,591,604	33,406,380	68,171,592	34,765,212	357	-0.28
9	Deletion	9q34.3	139,516,033	139,814,485	298,452	139,721,466	139,874,940	153,474	9	-0.52
9	Mosaic Trisomy 9		whole chi	romosome	140,007,236	140,524	140,147,760	27,641	11466	0.10
9	Mosaic Trison	ny 9	whole chi	romosome	140,007,236	140,524	140,147,760	27,641	11466	0.06
10	Deletion	10q26.13	126,415,527	134,032,911	7,617,384	125,279,327	135,272,495	9,993,168	959	-0.40
12	Deletion	12q14.2-q15	63,362,084	66,737,699	3,375,615	63,387,151	66,796,852	3,409,701	376	-0.43
13	Deletion	13q12.11-q12.12	18,876,037	24,330,232	5,454,195	18,958,454	24,384,515	5,426,061	664	-0.39
14	Deletion	14q11.2	20,741,117	20,988,716	247,599	20,767,781	20,998,178	230,397	18	-0.44
14	Deletion	14q11.2	19,592,409	21,256,822	1,664,413	19,496,544	21,284,915	1,788,371	170	-0.39
16	Duplication	16p13.3	2,681,813	3,927,524	1,245,711	2,814,497	3,844,547	1,030,050	47	0.27
17	Deletion	17q21.31	41,049,321	41,564,451	515,130	41,097,235	41,587,072	489,837	50	-0.45
22	Deletion	22q11.2	19,062,809	19,785,125	722,316	19,072,450	19,773,283	1,807,286	47	-0.30
22	Duplication	22q11.21	19,429,297	19,791,607	362,310	19,444,601	20,258,916	814,315	19	0.26
22	Deletion	22q12.1	26,293,416	27,462,458	1,169,042	26,319,777	27,505,873	1,186,096	71	-0.40
x	Duplication	Xq12-q21.1	67,088,023	76,204,344	9,116,321	65,158,451	76,235,884	11,077,433	259	0.19

List of Validated De novo CNVs in 146 MR Children

**Table I.1: Validation of OPAS sensitivity in detecting previously known CNVs in 146 mental retardation patients.** This table presents the list of validated CNVs in 146 MR patients that had been previously reported by Friedman et al. in [30] (all these CNVs were experimentally validated). The first 5 columns present the CNV data, as reported in [30]; and the last 3 columns indicate an OPAS-estimated CNV that overlaps with the corresponding region.

## Appendix J

# **Relationship Between Hybridization Intensity Noise and the Number of Predicted CNVs in FL Genomes**

The aim of this analysis is to investigate whether the number of predicted CNV calls are correlated with the overall standard deviation of raw signal intensities from the arrays. To perform this analysis, I studied OPAS-predicted regions of copy number change in 25 Follicular Lymphoma patients using 250K Nsp array data and plotted their frequency against the SD of the raw PM Affymetrix Nsp array data (PM oligo-level data obtained from .CEL files), as shown in Figure J.1.

The results of the analysis reveals that there no proof of significant correlation between the number of OPAS CNV calls (blue curve) and the standard deviation of hybridization intensities of the arrays (noise; red curve) at 0.05 significance level (Pearson's linear correlation coefficient r = -0.1157, with p-value P = 0.59). This finding suggests that the frequency of OPAS predicted CNV calls does not increase or decrease proportional to the overall SD (noise) of the array.



Figure J.1: Relationship between variation of hybridization signal intensities and CNV counts in 25 FL genomes. The left *y*-axis (shown in blue) denotes the number of candidate somatic CNVs in 25 patients from the follicular lymphoma study (Chapter 4). A low stringency ( $|LR| \ge 0.15$ ) was used to call candidate somatic CNVs in 25 follicular lymphoma patients.

The x-axis indicates the patient id, sorted from left to right based on increasing number of CNV counts per patient. Thus, patient 17 with only 2 CNVs and patients 21 and 24 with 35 predicted CNVs have the least and the most number of candidate somatic CNVs in this analysis. The right y-axis (shown in red) demonstrates the standard deviation (SD) of PM log-ratio intensities in the raw array data (250K Nsp array) for each corresponding patient. This analysis did not find a significant correlation between the SD of hybridization intensities in the raw array data and the number of predicted CNV calls (p-value of Pearson correlation p = 0.59). This finding reveals an important advantage of OPAS over other methods, such as in [141], that had previously reported a significant correlation between the above parameters and found a significant correlation between standard deviation of the sample intensities. As added by Itsara et al. in [141] when the aforementioned parameters are positively correlated, the rates of false positive CNV calls proportionally increase with the standard deviation (noise) of the raw array data.

## Appendix K

## **Description of FPP Events**

- Green color in Tumordb indicates clones with candidate rearrangements that were sequenced or sent for sequencing.
- Gold denotes clones with candidate rearrangements based on FPP alignments.
- Grey-blue BACs indicate that the FPP mapping is not particularly strong, i.e., the location is not particularly confident.
- **FPP hole** or **Fragment hole**: One or more fragments in the fingerprint does not match to the alignment region in the reference genome, resulting in a gap (or a hole) in the FPP alignment. There could be many reasons for such a gap, such as deletion, SNP, or FPP mapping error.
- **SNP hole**: An FPP-hole that could be explained by a snp in the fragment(s) that do not match to the reference genome. So candidate FPP holes that overlap with known SNPs are referred to as SNP holes.
- **Multi FPP**: the clone alignment is split to multiple regions (as opposed to a contiguous region). Such events can represent a deletion, translocation, inversion or duplication. However, further investigation is required to assess the exact source of the event.
- **SEQ-MATCH:** the CNV was validated by sequencing an FPP BAC clone that represented the event.
- **FPP-MATCH:**There is a clone with an FPP alignment representing the event but no sequence contig was available at the time of preparing this table.

- **CPLX-MATCH:** A breakpoint near the OPAS candidate event was identified, but only some aspects of the event was captured in a BAC clone. Also refers to events where several likely real events occurred within the OPAS estimated breakpoints.
- **FRAG-MATCH:** There is a fragment hole which overlaps with the OPAS predicted event. Fragment holes are suspected small rearranged regions within the span of a BAC clone. However, these events have not been validated (at the time of preparing this table).

Appendix L

## **Supplementary Material for Chapter 4**



Figure L.1: Candidate distal deletion on chromosome 8p23.3 of patient 25 ( $\sim$ 21 kb) containing 3 SNP probe markers (OPAS-exclusive). Illustrates the OPAS scatter plot of chromosome 8 of patient 25 (ht-25). The highlighted region indicates a candidate deletion (only detected by OPAS and not SMD or aCGH). This putative deletion is  $\sim$ 21.3 kb with LR = -0.23 and a significantly small z-score (z-score = -1.57). The latter region is located on chromosome 8p23.3 starting within 180 kb of the p-ter and includes only 3 Nsp SNP probe markers. The FPP data cannot be used to investigate this putative CNV, as FPP does not cover regions close to chromosome ends.



(a) OPAS scatterplot of chromosome 1 of patient 18, indicating a candidate distal deletion on 1p36.33-p36.32



(b) UCSC screenshot, comparing OPAS, aCGH and SMD results of 1p36 region of patient 18

Figure L.2: Candidate OPAS distal deletion on chromosome 1p36 (ht-18), that includes an array CGH predicted deletion. Panel (a) represents the OPAS scatterplot indicating a candidate deletion, approximately 1.9 Mb (775,852-2,713,412) in patient 18 (ht-18), which includes 22 Nsp SNP probe markers. This candidate deletion has LR = -0.23 and affects 1p36 region that is known to be frequently deleted in FL genomes [244]. Panel (b) indicates that OPAS candidate deletion in ht-18 is not confirmed by SMD. This putative deletion contains an aCGH predicted copy number loss in this patient. FPP results cannot be used to verify this putative deletion (since it is close to chromosome end). Although, there is no direct data to validate this candidate deletion, since it occurs at a commonly deleted chromosomal region and the fact that it also includes and aCGH predicted deletion, emphasize that this candidate CNV may be a real deletion in ht-21.



Figure L.3: Candidate distal amplification on 14p36 (ht-7) with slight gain of signal intensity (LR = +0.18) but a significant z-score (+1.1). This figure represents the OPAS scatterplot of chromosome 14 of patient 24 (ht-24). As seen there is a region with apparent gain of signal intensity, approximately 570 kb (denoted by arrow), at the distal end of the chromosome 14q (14q32.33). This putative copy number gain has a low magnitude of signal intensity deviation from baseline, LR = 0.18. However, compared to the distribution of the entire chromosome 14, the z-score of this gain is significant (z-score = +1.1). Based on the estimated z-score and visual inspection of OPAS plot, this region was selected as a candidate copy number gain in the FL dataset. The above candidate amplification was also detected by aCGH results in Tumordb (but not SMD; aCGH image was not available but the underlying data can be accessed through Tumordb).



Figure L.4: Candidate OPAS distal deletion on chromosome 6 of patient 20 that includes 11 SNP probes with significant loss of signal intensity (LR = -0.42; z-score = -1.50). This figure displays the OPAS scatter plot of chromosome 6 of FL patient 20. The image highlights a candidate subtelomeric deletion proximal to the p-end of chromosome 6 with LR = -0.42 and significant deletion z-score of -1.50. The latter candidate event is ~78 kb and includes 11 SNP probe marker. This candidate CNV has not been detected by aCGH or SMD results in Tumordb (OPAS-exclusive event). FPP data was also not available to investigate this putative deletion which starts within 119 kb of the p-ter of chromosome 6 (FPP does not cover regions close to chromosome ends). Nonetheless, the loss of log2-ratio intensity in the aforementioned region that is evident from OPAS chromosome plot (LR = -0.42) and its significant estimated z-score (z-score = -1.50) emphasize that the above putative CNV is likely a real copy number loss that affects the subtelomeric region of chromosome 6 in patient 20. Further experiments are required to investigate this candidate event.



**Figure L.5: Examples of FISH validated 1p36 deletions in 4 FL patients.** These plots illustrate the results of FISH experiments that had been performed (by Dr. Horsman's lab at the BCCRC) to verify 1p36 deletion in several FL patients of this study. The results shown here confirm 1p36 deletion in 4 FL patients (ht-24, ht-29, ht-28 and ht-9). The red dots represent probes that were designed to target 1p36 region, and the green dots represent control probes that were designed to hybridize to a normal region on 1q32.3. As observed, 1p36 region was deleted in all of these patients, although the deletion heterogeneity varied between 22%-94% among these 4 patients.



(a) OPAS scatterplot of chromosome 14 of patient 14, indicating a deletion on 14q32.33 (~230 kb; 4 SNPs)



(b) Tumordb screenshot of 14q32.33 chromosomal region of ht-14



(c) Tumordb screenshot of 14q32.33 chromosomal region of ht-14

**Figure L.6: Candidate OPAS-exclusive focal deletions on chromosome 19 of patient 4.** Panel (a) shows the OPAS scatter plot of chromosome 19 of patient 4. This plot indicates two candidate focal deletions on 6p13.2 that are ~65 kb (6 SNPs; CNV #1) and ~90 kb (9 SNPs ; CNV #2). Both of these candidate deletions indicate strong losses of signal intensity and significant deletion z-scores with LR = -0.39 (z-score = -1.9) for CNV #1, and LR = -0.43 (z-score = -2.1) for CNV #2. However, none of these deletions have been reported by SMD or aCGH results of Tumordb. The FPP alignments of these regions are shown in panel (b) (for candidate CNV #1) and panel (c) (for candidate CNV #2). The red dashed lines in (b) and (c) indicate the boundaries of OPAS-predicted copy number deletions. As observed there is no FPP event that can directly validated these candidate deletions. However, for CNV #1 there is only 1 BAC that spans the entire region (b); and for CNV #2 there are two BACs that span the entire region and a third BAC, shown in grey, that indicates a potential rearrangement although the expected event has not been validated (c). These observations may represent hemizygous deletions in these regions.



(a) OPAS scatter plot of chromosome 6 of patient 24



292

Figure L.7 (previous page): Slight gain of signal intensity of a region within a deleted chromosome arm, predicted to represent an amplification (LR = 0.03; z-score = +0.6). Panel (a) shows the OPAS scatterplot of chromosome 6 of patient 24. It is clear that the short arm is amplified and the long arm is deleted in this chromosome (an event known as iso-chromosome). There is also a clear deletion in the amplified p-arm that is detected by all three datasets (denoted by \* in OPAS, aCGH (b) and SMD (c) results). The red arrow in OPAS plots highlights another potential amplification, located on the predominantly deleted q-arm of this chromosome. This region (referred to as region M) has LR  $\approx$  0.03 with a z-score of +0.6. As seen in panel (b), aCGH result did not detect the increase of signal intensity in this region and reported that the entire q-arm is deleted. The SMD results, shown in panel (c), identified the pattern of signal intensity, however, it detected M as a copy number normal region. Based on the observed pattern of intensity in region M, it can be speculated that this region is likely amplified (z-score = +0.6). This candidate amplification (M) includes several genes, such as FBXL4 (highlighted in yellow). According to GNF Atlas expression track of UCSC, FBXL4 is highly expressed in several leukemias and lymphomas (e.g., Raji-Burkitt's lymphoma cell line)<sup>1</sup>. The observed amplification may also have a similar impact and may lead to increased expression of FBXL4.

<sup>&</sup>lt;sup>1</sup> http://genome.ucsc.edu/cgi-bin/hgGene?hgg\_gene=uc003ppf.1&hgg\_prot=Q9UKA2&hgg\_chrom=chr6&hgg\_start=99428321&hgg\_end=99502570&hgg\_type=knownGene&db=hg18&hgsid=183794759



(a) OPAS scatterplot of chromosome 4 of patient 19, indicating a candidate OPAS-exclusive focal deletion on 4q28.3 ( $\sim$ 10 kb; 4 SNPs)



(b) Screenshot of Tumordb query interface

(c) OPAS scatterplot of chromosome 2 of patient 19

Figure L.8: Candidate OPAS-exclusive focal deletion on chromosome 4 of FL patient 19 (~10 kb; 4 SNPs), adjacent to an FPP translocation site between 4q28.3 and 2p25.1. Panel (a) shows the OPAS scatterplot of chromosome 4, indicating a candidate OPAS-exclusive focal deletion (~10 kb) on chromosome 4q28.3 in patient 19 (ht-19). This putative deletion includes 4 SNP probe markers and has LR = -0.72. Panel (b) displays a screenshot of Tumordb query, showing events in ht-19 that overlap with the above OPAS predicted deletion (OPAS event: 131,663,499-131,674,302). As seen this query finds an FPP translocation event between chromosome 4 (chr4: 131,671,379-131,791,464). It is observed that the predicted translocation event is adjacent to the aforementioned candidate OPAS-exclusive deletion in 4q28.3 (FPP event: 131,663,499-131,674,302).

The other end of this translocation on chromosome 2 (chr2: 56,670,655-56,736,729) is also adjacent to another focal deletion (~132 kb) in this chromosome. However, the second deletion is detected by both OPAS (shown in panel (c)) and SMD (data can be obtained though Tumordb). Based on these observations, it can be speculated that a part of the rearranged sequence may have been lost during the t(2p; 4q) translocation in ht-19. Also, since both of these candidate focal deletions are adjacent to translocation sites, another possibility is that these small deletions increased the FL genome instability in these chromosomal loci, which resulted in a subsequent translocation (t(2p; 4q)). Further experiments are required to investigate whether the OPAS exclusive deletion in 4q28.3 is a real CNV.

Symbol	Description	Location	Type of product
ANKS1B	ankyrin repeat and sterile alpha motif domain containing 1B	Nucleus	other
ARID3B	AT rich interactive domain 3B (BRIGHT-like)	unknown	other
C6ORF141	chromosome 6 open reading frame 141	unknown	other
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	Nucleus	transc reg.*
CENPQ	centromere protein Q	unknown	other
CLK3	CDC-like kinase 3	Nucleus	kinase
DENND3	DENN/MADD domain containing 3	unknown	other
ERBB4**	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	Plasma Membrane	kinase
FHIT	fragile histidine triad gene	Cytoplasm	enzyme
HVCN1	hydrogen voltage-gated channel 1	unknown	ion channel
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	Plasma Membrane	kinase
KLHL1	kelch-like 1 (Drosophila)	Cytoplasm	other
LYN	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog	Cytoplasm	kinase
MTAP	methylthioadenosine phosphorylase	Nucleus	enzyme
MUT	methylmalonyl CoA mutase	Cytoplasm	enzyme
PAX5	paired box 5	Nucleus	transc reg.*
PPP1CC	protein phosphatase 1, catalytic subunit, gamma isozyme	Cytoplasm	phosphatase
PPTC7	PTC7 protein phosphatase homolog (S. cerevisiae)	unknown	phosphatase
SFPQ	splicing factor proline/glutamine-rich	Nucleus	other
SLC45A4	solute carrier family 45, member 4	unknown	other
STXBP5L	syntaxin binding protein 5-like	Cytoplasm	other
TCTN1	tectonic family member 1	unknown	other
TEKT3	tektin 3	unknown	other
ZNF177	zinc finger protein 177	Nucleus	other
ZNF440	zinc finger protein 440	Nucleus	other
ZNF441	zinc finger protein 441	Nucleus	other
ZNF491	zinc finger protein 491	Nucleus	other
ZNF559	zinc finger protein 559	Nucleus	other
ZNF596	zinc finger protein 596	unknown	other

Table L.1: Ingenuity Pathway Analysis list of known genes that were affected by candidate focal CNVs ( $\leqslant$  150 kb)

Total number of genes =  $29^{\dagger}$ .

\*: stands for transcription regulator.

\*\*: *ERBB4* is affected by two distinct putative small deletions.

 $^{\dagger}:$  IPA database did not identify one of the 30 genes in Table 4.4.