Mutation Discovery in Regions of Segmental Cancer Genome Amplifications from Next Generation Sequencing of Tumours

by

Anamaria Crisan

B.Comp.(hons. – SSP Biomedical Computing), Queen’s University, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

The Faculty of Graduate Studies
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

October 2010

© Anamaria Crisan, 2010
Abstract

Next generation sequencing has now enabled a cost-effective enumeration of the full mutational complement of a tumour genome - in particular single nucleotide variants (SNVs). However, most current computational and statistical models for analyzing next generation sequencing data do not account for cancer-specific biological properties, including somatic segmental copy number alterations (CNAs), which require special treatment of the data.

Here we present CoNAn-SNV (Copy Number Annotated –SNV): a novel algorithm for the inference of single nucleotide variants (SNVs) that overlap copy number alterations. The method is based on modelling the notion that genomic regions of segmental duplication and amplification induce an extended ‘genotype space’ where a subset of genotypes will exhibit heavily skewed allelic distributions in SNVs (and therefore render them undetectable by methods that assume diplody). CoNAn-SNV introduces the concept of modelling allelic counts from sequencing data using a panel of Binomial mixture models where the number of mixtures for a given locus in the genome is informed by a discrete copy number state given as input. We applied CoNAn-SNV to a previously published whole genome shotgun data set obtained from a lobular breast cancer and show that it is able to detect 24 experimentally revalidated somatic non-synonymous mutations that were not detected using copy number insensitive SNV discovery algorithms. Importantly, ROC analysis shows that the increased sensitivity of CoNAn-SNV does not result in disproportionate loss of specificity. Our results indicate that in genomically unstable tumours, copy number annotation for SNV detection will be critical to fully characterize the mutational landscape of cancer genomes.

The Binomial mixture model framework, however, is unable to fully cope with the complexity of tumour sequence data. We explore substituting the Binomial mixture model framework with the Beta-Binomial framework to overcome this limitation. The effectiveness of this substitution is compared against the lobular breast carcinoma and the 30 exon capture data sets all derived from triple negative breast cancers. The performance of Binomial and Beta-Binomial mixture model is evaluated against a cohort of exon capture test cases and we report ROC and f-measures.
Preface

The following lists the publication(s) arising from the work and reports the contributions of each of the authors.


Contributions to methods and analysis: The biological motivation and CoNAn-SNV model were devised jointly by Dr. Sam Aparicio and Dr. Sohrab Shah. Anamaria Crisan implemented as well as evaluated the model and contributed to refinements of the original concepts. Rodrigo Goya implemented the model in C. Anamaria Crisan summarised validation data and helped devise and implement novel methods for handling targeted ultra deep amplicon sequencing data. Interpretations of results in the context of biological implication were done jointly by Anamaria Crisan, Dr. Sohrab Shah and Dr. Sam Aparicio. Additional authors identified were involved in various stages of sample preparation, sequencing, validation, and analysis.

Contributions to paper: Anamaria Crisan created all the figures and tables as well as first several drafts of the paper. Dr. Sohrab Shah and Dr. Sam Aparicio suggested refinements to figures and tables and had the most creative control of content in later drafts. Content of this paper appears in Chapter 2 and throughout the Introduction.
Table of Contents

Abstract ................................................................................................................................. ii
Preface .................................................................................................................................. iii
Table of Contents ................................................................................................................ iv
List of Tables ....................................................................................................................... vi
List of Figures ...................................................................................................................... vii
Abbreviations ..................................................................................................................... viii
Acknowledgments ............................................................................................................... ix
Dedications ............................................................................................................................ x

Chapter 1 | Introduction .......................................................................................................... 1

1.1 Single Nucleotide Variants and Copy Number Alterations in Cancer ............................. 2
1.2 Discovering the Cancer Mutational Landscape with NGS .............................................. 4
1.3 Extending Genotype Definition: the Allele-Specific Copy Number Genotype .................. 7
1.4 SNV Discovery Algorithms for NGS Data ................................................................... 9
1.5 Research Goals ............................................................................................................. 11
1.6 Thesis Summary .......................................................................................................... 11

Chapter 2 | Mutation discovery in regions of segmental cancer genomic amplifications
with CoNAn-SNV: a mixture model for next generation sequencing of tumours ......... 12

2.1 The CoNAn-SNV Model ............................................................................................. 13
2.2 Methods ....................................................................................................................... 19
   2.2.1 Single nucleotide variant discovery and validation .............................................. 19
   2.2.2 Performance evaluation with OncoSNP and CRLMM ..................................... 21
2.3 Results .......................................................................................................................... 23
2.3.1 Experimental validation of the CoNAn-SNV model .................................................. 23
2.4 Discussion .................................................................................................................. 31
2.5 Conclusion .................................................................................................................. 35

Chapter 3 | Beta-Binomial extension of the Binomial Framework applied to Whole Genome Shotgun Sequencing and Exon Capture ................................................. 36

3.1 Introduction ................................................................................................................. 37
  3.1.1 The Beta-Binomial distribution ............................................................................. 37
  3.1.2 Exon capture ........................................................................................................ 41
3.2 Methods ..................................................................................................................... 42
3.3 Results ......................................................................................................................... 44
3.4 Discussion ................................................................................................................... 56
3.5 Conclusions ................................................................................................................. 59

Chapter 4 | Conclusions .................................................................................................... 60

4.1 Summary of Contributions ....................................................................................... 60
4.2 Future Work ............................................................................................................... 62
4.3 Implications for Cancer Treatment .......................................................................... 63

Bibliography ..................................................................................................................... 64

Appendix A | Exon Capture Processing .............................................................................. 69

Appendix B | SeqVal01(INX021): Validation of 200 SNVs in lobular breast cancer ............. 73
Appendix C | Description of additional files ......................................................................... 75
List of Tables

TABLE 1.1 SUMMARY OF DATA SETS APPLIED TO OUR METHODS.................................11

TABLE 2.1 HYPERPARAMETER INITIALISATION VALUES...........................................18

TABLE 2.2 SOMATIC SINGLE NUCLEOTIDE VARIANTS DISCOVERED BY CONAN-SNV.........26

TABLE 2.3 GERMLINE VARIANTS PROXIMAL TO SOMATIC SNVS EXHIBIT ALLELIC
SKEW.................................................................................................................................27

TABLE 3.1 COMPARING THE BEST PERFORMING ALPHA AND BETA PARAMETERS OF
THE BINOMIAL AND BETA-BINOMIAL MIXTURE MODEL AGAINST 5 VALIDATION
CASES WITHHELD FROM CV TRAINING.............................................................................51

TABLE 3.2 EVALUATING MODEL PERFORMANCE AT DEFAULT PARAMETERS..................51

TABLE 3.3 RECOVERY OF 24 VALIDATED SOMATIC MUTATIONS BY THE BINOMIAL
AND BETA-BINOMIAL IMPLEMENTATION OF CONAN-SNV..............................................55

TABLE 3.4 TOTAL NUMBER OF CALLS MADE BY EACH MODEL FROM ALL VALIDATION
DATA....................................................................................................................................55

TABLE A1.1. COMPARISON OF DIFFERENT ALIGNMENT REGIMENS...........................71
List of Figures

Figure 1.1 MUTATION ACCUMULATION IN A PHENOTYPICALLY NORMAL CELL LINEAGE........3

FIGURE 1.2 SCHEMATIC DIAGRAM FOR NEXT GENERATION SEQUENCING PIPELINE.........6

FIGURE 1.3 SNV AND CNA DISCOVERY FROM NGS GENOME DATA....................................7

FIGURE 1.4. MODELLING ALLELE-SPECIFIC COPY NUMBER GENOTYPES FROM USING AFFYMETRIX SNPCHIPL........................................................................................................8

FIGURE 1.5 NOVEL SOMATIC VARIANTS DETECTED IN ALLELE-SPECIFIC AMPLIFICATION ON CHROMOSOME 19Q ARM........................................................................................................10

FIGURE 2.1 OVERVIEW OF CONAN-SNV MODEL, INPUTS AND OUTPUTS.......................12

FIGURE 2.2 SNV DISCOVERY PIPELINE USING CONAN-SNV........................................22

FIGURE 2.3 VENN DIAGRAM OF PREDICTIONS MADE BY MAQ, SNVMIX1, CONAN-SNV SEPARATED BY CNA STATE........................................................................................................24

FIGURE 2.4 ONCOSNP AND CRLMM RECEIVER OPERATOR CHARACTERISTIC CURVES.................................................................................................................................28

FIGURE 3.1 EFFECTS OF $\alpha$ AND $\beta$ ON THE SHAPE OF THE BETA DISTRIBUTION.........38

FIGURE 3.2 BETA-BINOMIAL AND BINOMIAL MIXTURE MODEL COMPARISON WITH SIMULATED DATA................................................................................................................45

FIGURE 3.3 MIXED WEIGHT HYPERPARAMETERS..................................................................46

FIGURE 3.4 ROC FOR THE BINOMIAL AND BETA-BINOMIAL MIXTURE MODELS ON WGSS DATA............................................................................................................................47

FIGURE 3.5 DISITRIBUTION OF AUCS FOR THE BINOMIAL AND BETA-BINOMIAL MIXTURE MODEL AT STRENGTH1 OVER A DEFINED SPACE..............................49

FIGURE 3.6 EFFECTS OF CHANGING STRENGTHS ON THE BETA-BINOMIAL DISTRIBUTION..........................................................................................................................50

FIGURE 3.7 COMPARISON OF THE BINOMIAL AND BETA-BINOMIAL IMPLEMENTATIONS OF CONAN-SNV........................................................................................................53
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNA</td>
<td>Copy Number Alteration</td>
</tr>
<tr>
<td>CoNAn-SNV</td>
<td>Copy Number Annotated—Single Nucleotide Variant</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximisation</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a posterior</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>SNV</td>
<td>Single Nucleotide Variant</td>
</tr>
<tr>
<td>TSG</td>
<td>Tumour Suppressor Gene</td>
</tr>
<tr>
<td>WGSS</td>
<td>Whole Genome Shotgun Sequencing</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to acknowledge my supervisors and committee members for preparing me and guiding me through my MSc; my labs mates, in particular Andrew McPherson and Rodrigo Goya, for their help and knowledge; lastly, the Bioinformatics Trainee Program for the opportunity to pursue my interests.

I was supported throughout my research by a CIHR-MSFHR Bioinformatics Trainee scholarship as well as by large donations from the Bank of Mom and Dad.
Dedications

I would like to dedicate this work to my family. For my mother, who first planted the passion of genomics in my 7th grade brain and has nourished my interest ever since. To my father who, despite not understanding what I do, is an endless source of support and pride. To my uncle, who brought us to this country; without him I would not be here. My grandparents, aunts, uncles and cousins who despite being so far away are proud of my accomplishments and support me however they can. Lastly, to Renee who has been a cheerful companion to me since I was child and whose curious little life is greatly missed.

I would also like to dedicate this work to Matthew Brehmer, without whom I would be a crazier person and who has weathered the brunt of this more than anyone else.

Finally to my brother, whose own passion inspires me to pursuits beyond my comforts and whose drive, maturity and discipline I admire.
Chapter 1

Introduction

The seminal finding of the first cancer-causing single point mutation in HRAS established that cancer is a disease whose pathophysiology is linked with its genetic mutational landscape (Reddy, 1982; Tabin, 1982). Following this, numerous targeted and belaboured laboratory measures made slow progress towards elucidating cancer driving mutational events. The mosaic of the cancer mutational landscape grew to encompass various somatic aberrations including: copy number alterations (CNAs), gene fusions, translocations, inversions, as well as insertions and deletions (Stratton, 2009). It also became evident, that multiple types of somatic aberrations occurred together, such as point mutations and CNAs, and that the combination of these events had oncogenic implications (Bianchi, 1989). As the relevance of newer findings were becoming more important for diagnostic and therapeutic purposes, there was an increased emphasis on the need for a more rapid, efficient and accurate means of searching for cancer driving mutations (Bell, 2010). Next generation sequencing (NGS) is one such technology because it has allowed for a cost effective, high quality enumeration of entire genomes at single nucleotide resolution. The algorithms developed to interpret NGS data, however, are often targeted towards stable diploid genome structures that exist in healthy individuals. Such methods are ill-equipped when applied to cancer genomes, which are often aneuploid (Greenman, 2010). Thus, the cancer mutational landscape may be incompletely characterised unless algorithms are developed for the interpretation of NGS data that are able to account for the unstable genomic architectures existing in cancer genomes.
1.1 Single Nucleotide Variants and Copy Number Alterations in Cancer

Cancer arises from a mosaic of mutations which have a selective advantage for growth and proliferation (Greenman, 2007). In this study, we focus on two features of the aforementioned mutational landscape: single nucleotide variants (SNVs) and copy number alternations (CNAs). Both of these features have been implicated, through multiple experimental approaches, in driving cancer pathophysiology and influencing patient therapeutic response.

Single nucleotide variants are point mutations that, in cancer, present in two forms: inherited germline polymorphisms that may make an individual more susceptible to cancer in their lifetime and somatic variants that arise spontaneously (Stratton, 2009). Although inherited polymorphisms play important roles in understanding cancer development, this study focuses only on somatically arising variants. Somatically acquired SNVs occur commonly in all organisms due to environmental exposures to radiation and chemicals or by intrinsic factors such as errors in DNA replication. Cellular DNA repair machinery corrects for these mutations or initiates cell death to eliminate them from the organism (Hoeijmakers, 2009). Some somatic mutations affect the function of genes associated with cellular repair response or apoptosis. Mutations in these genes, known as Tumour Suppressor Genes (TSGs) can remain in the cell and contribute to a genomic instability that in some cases may result in the accumulation of further somatic mutations in other genes (Stratton, 2009). Oncogenes are genes that when mutated lead to cancer; TSGs work actively to counter the effects oncogenes and the loss of TSG function, due to somatic mutations, can lead to cancer over time (Croce, 2008). This is represented in Figure 1.1. Cancer is identified by its hallmark of neoplastic growth: rapid proliferation of mutant cells lacking an apoptotic response (Stratton, 2009). Neoplastic growth of a tumour is dependent upon the accumulation and evolution of somatic mutations that provide an advantage to the tumour’s further development. As a result, randomly occurring somatic SNVs are subject to selective pressure for tumour survival and growth. A tumour is comprised of populations of cells each with different mutations that have a selective advantage (Heppner, 1984; Greenman, 2007). The heterogeneity of the tumour complicates treatment, as not all populations may be responsive to same pharmaceutical regimen or new populations may emerge that are resistant to it (Mullighan, 2008). The full characterisation of somatically arising SNVs can play critical roles for understanding tumour aetiology, pathogenesis and response to therapeutics.
Figure 1.1 Mutation accumulations in a phenotypically normal cell lineage from Stratton et al. (2009). [Reproduced with permission] This figure illustrates mutations accumulated over an individual’s lifetime, and how exogenous as well as intrinsic events may contribute to the onset and eventually progression of cancer. Somatic mutations are classified as driver or passenger mutations, drivers are responsible for its pathogenesis where passengers play no role in cancer evolution. This figure highlights that few driver mutations contribute to the cancer phenotype. Absent in the illustration is that different cells within a tumour may evolve to have different driver mutations. This causes intra-tumoural heterogeneity and can affect patient outcome.

DNA copy number alterations (CNAs) are genomic segments, ranging in size from 1 kilobase (1 Kb) to whole chromosomes, where the number of copies of DNA is higher or lower than the expectant diploid genome, which has two copies. Raw copy numbers can be classified into different levels of amplification and loss for ease of interpretation. DNA copy number variation is a feature of both normal and cancer genomes, however somatic events involving oncogenes play a critical role in disease onset and progression (Sebat, 2005; Beroukhim, 2010). Although it is not clear for all cases, different levels amplification can be associated with higher levels of expression for certain genes (Neve, 2006). The amplification, and subsequent increased expression, of HER2 in breast cancers is one such example. This oncogene is significant because a routine clinical testing for HER2 over expression can result in using a targeted pharmaceutical, Trastuzumab®, which is highly effective against the cancer (Salamon, 1987; Vogel, 2002).

Multiple types of somatic aberration have been reported to occur together: for example, Kadota et al. (2009) observed recurrent mutations of PIK3CA in breast cancer with allele specific amplifications of the mutant allele, and suggested that PIK3CA point mutations with concomitant CNA amplification resulted in synergistic oncogenic effects. In instances of an allele-specific amplification, one allele is preferentially amplified over the other. The proposed mechanism for this
event in Bianchi et al. (1990) is that a point mutation can confer instability that makes a genomic segment susceptible to amplification; in such instances the mutant harbouring segment or chromosome is amplified. LaFramboise et al. (2005) showed allele specific amplification of EGFR mutant alleles in a lung cancer cell line and examples of amplification co-occurring with somatic mutations in MYC (Herrick, 2005), HRAS (Bianchi 1990), and MET (Zhuang, 1998) have been observed. Generally, allele-specific amplification and their functional impact are poorly understood (LaFramboise, 2005). In this study, we speculatively suggest that it is possible that amplification occurs first and somatic mutations develop later in tumour evolution.

The co-occurrence of single nucleotide variants in regions of segmental copy number amplification poses special problems for variant discovery methods that assume diploidy because unknown mixtures of allele abundances could result from the process of allele-specific or uneven amplifications. This was quantified using whole genome single nucleotide polymorphism (SNP) chip microarray platforms (Section 1.3). At a given position, the mixture of alleles may be skewed resulting in a departure from the theoretical frequency (0.5) for heterozygous variants expected in diploid genomes. Therefore, we hypothesize that the extent of allelic skew induced by CNAs will confound the ability to enumerate the mutational landscape of cancers. This topic is discussed further in chapter 2.

1.2 Discovering the Cancer Mutational Landscape with NGS

Recent advances in massively parallel genome short-read sequencing methods (so-called next generation sequencing) have placed the goal of complete delineation of the cancer genome landscape at single nucleotide resolution within practical reach. The general experimental protocol is to fragment whole genomes into smaller segments, sequence the fragments independently in parallel (to produce millions of short “reads”) and finally reconstruct the whole genome using bioinformatic algorithms; this is depicted schematically in Figure 1.2.

The reconstruction of the genome is achieved by aligning the genomic reads to a pre-existing reference genome (Li H., 2008). At a given position in the genome, the reads can be collapsed into allelic counts that measure the total number of reads that match the reference genome, $a$, or that do not match the reference genome, $b$ (Figure 1.3). Furthermore, the data can be genotyped on the
basis of total number of reads that match the reference genome: \textit{aa} homozygous reference, implies the majority of reads match the reference allele; \textit{ab} heterozygous implies approximately half of the reads match the reference; and \textit{bb} homozygous non-reference (Goya, 2010). These genotypes are assigned on the premise that diploid genomes will harbour two copies of each allele. Genotypes containing the \textit{ab} and \textit{bb} genotypes represent a SNV because they contain instances of a variant allele. Furthermore, it is possible to use the depth (sum of the allelic counts) to calculate raw copy number, which can be summarised further as a CNA state (loss, neutral/diploid, gain, amplification, or high level amplification) and used in analysis (Shah, 2009). The schematic representation of SNV and CNA discovery from NGS data are shown Figure 1.3. Further details of CNA discovery, particularly the methods employed in this study, are described in Shah et al. (2009).

Several cancer genomes have now been deeply sequenced with NGS and analyzed for CNAs and SNVs independently using bioinformatic approaches followed by targeted validation to confirm somatic alterations. These studies have revealed novel somatic point mutations in acute myeloid leukaemia (Ley, 2008; Mardis, 2009), breast cancer (Shah, 2009; Ding, 2010), ovarian cancer (Shah (b), 2009), melanoma (Pleasance, 2010), lymphoma (Morin, 2010) and lung cancer (Pleasance, 2010). Work by Pleasance et al. (2010), Chiang et al. (2008) and Shah et al. (2009) suggest that CNAs can be inferred from sequence data. These studies demonstrated a range of raw copy number up to an including 8 copies that can occur extensively in a both a primary tumour and cell line. None of these studies, however, have used algorithms that explicitly integrate CNA information to leverage SNV discovery; instead they employ methods that assume diploidy. As it has been previously described, SNVs and CNAs can occur concomitantly and this will cause the heterozygous ratio of reference to non-reference allele to deviate from heterozygous diploid ratio of 0.5, potentially rendering some variants undetectable (Figure 1.3B; Figure 1.5). New methods for the analysis of short-read sequence data are needed, in particular those that capable of coping with the complex genomic landscapes of tumours. We demonstrate in chapter 2 how the incorporation of CNA information in SNV discovery in cancer genome sequence data yields additional novel somatic mutations that were undetectable using copy number naïve SNV prediction algorithms. Furthermore, CNA data is important for the interpretation of mutational heterogeneity in tumours. Recent studies by Ding et al. (2010) and Shah et al. (2009) have used targeted validation methods to
estimate the frequency of mutations in the population of tumour cells in order to detect sub-dominant cell populations that result due to intra-tumoural heterogeneity (see 1.1). In these studies, mutations that exhibit a low frequency have been associated with so called sub-dominant populations. Skewed allelic ratios resulting from SNVs harboured within CNAs may confound the interpretation of intra-tumoural heterogeneity because they can also exhibit low frequency mutations due to allelic skew.

Figure 1.2 Schematic Diagram for Next Generation Sequencing Pipeline. A) The pipeline begins with the excision and sample collection of the tumour DNA from the patient. The sample is purified, fragmented into variable lengths by sonication, size selected for an ideal length (approx. 200 base pairs) and incorporated in a gDNA library. Fragments are prepared according to the regimen required by the sequencing technology and sequenced. B) Each NGS platform has a specific sequencing chemistry; the diagram shows the simplified method used by the Illumina GenomeAnalyzer II. This platform outputs a volume of coloured images (where each colour represents a different base) and requires further image processing to output the nucleotide base (A,C,T,G) in the sequence. C) The fragments are not sequenced to their entirety; instead a specific base pair length from either end of the molecule is sequenced. This is known as paired end sequencing. D) Paired end reads are aligned using a short read alignment algorithm to a reference genome. There are three possible scenarios for the alignment process: i) both paired ends align uniquely to the reference genome ii) only one of the paired ends aligns to the reference sequence and the other is aligns ambiguously to multiple locations iii) both paired ends align ambiguously and cannot be placed. It is possible to use i) and ii) for further analysis, however no methods currently exist that incorporate ambiguously aligned paired end reads for SNV and CNA discovery. E) The genome is represented as a space of aligned reads that can be collapsed into allelic counts for genotyping (Figure 1.3) F) Using a varying panel of existing bioinformatic tools it is possible to devise a list of somatic variants in the tumour genome for validation or further inquiry.
1.3 Extending Genotype Definition: the Allele-Specific Copy Number Genotype

The effects of CNAs on SNV discovery is an important consideration for the complete and accurate enumeration of the cancer mutational landscape. Existing SNV discovery tools for NGS data such as MAQ (Li H., 2009), SNVMix (Goya, 2010) and SOAPSNP (Li R., 2009) assume a diploid genomic architecture and are incapable of managing the distribution of allelic ratios that can result from CNAs. Algorithms developed for the whole genome single nucleotide polymorphism (SNP) chip microarray (SNPChip) platforms suggest, however, that applying the paradigm of diploid genomes to cancer genomes, which are more complex and aneuploid, is inappropriate (Greenman, 2010).
Figure 1.4 Modelling allele-specific copy number genotypes from the SNPChip in Wang et al. (2007). [Reproduced with permission] This figure shows the different levels of CNAs can exhibit a range of genotypes in the affected genomic segment. The genotypes are determined a priori and expand naturally based on the level of amplification.

Whole genome single nucleotide polymorphism arrays have allowed for quantification of allele-specific CNAs by mitigating the incorporation of copy number state with allelic genotype information. Algorithms such as QuantiSNP (Colella 2007), VanillaICE (Scharpf 2008), Birdsuite (Korn 2008), PennSNV (Wang 2007), OncoSNP (Yau, unpublished) and PICNIC (Greenman 2010) model CNAs by extending the genotype state space from the conventional three diploid genotypes. The base definition for the genotype space within different CNA levels was established by Colella et al. (2007) and has been adopted by other algorithms with little or no modification. The CNA specific genotypes represent a discrete space that is established a priori by the authors and the data are assigned to one of these genotypes. The genotypes naturally expand from the diploid definition such that for “gain” regions, for example a triploid chromosome or segmental gain, could have the following genotypes: aaa, aab, abb and bbb (Figure 1.4). This extended genotype state space is refered to as the allele-specific copy number genotype.

Unlike the NGS platforms, SNPChip technologies do not have the capacity to resolve the genome at a single nucleotide resolution and additionally are limited to the SNPs and co-ordinates that are affixed to the array. These limitations make SNPChip platforms a suboptimal method for interrogating cancer genomes, however it is possible to use the notion of an allele-specific copy number genotype, developed on these platforms, to improve NGS variant discovery algorithms.
1.4 SNV Discovery Algorithms for NGS Data

Next generation sequencing technologies produce as output millions of short reads that must be
organised and interpreted prior to further analysis. Bioinformatics algorithms have been developed
concomitantly with NGS platforms to resolve this task; as highlighted in Figure 1.2, this involves
two phases: an alignment phase that is followed by variant discovery. Although the selection of the
appropriate alignment tool does impact SNV discovery, we do not focus on this topic in this work.
In this study we use the same base genome alignment for all the variant discovery tools, thereby
negating the effect of aligner selection in our results.

A simplistic approach for SNV discovery would be to establish thresholds requiring a certain
amount of read support for the variant allele in order to call a SNV. Such thresholds, however, are
difficult to objectively determine. A Bayesian approach for SNV discovery in genomic data was used
in traditional Sanger Sequencing and continues to be the framework for contemporary NGS variant
discovery algorithms (Marth, 1999). The advantage of using a Bayesian approach is that it outputs a
probabilistic confidence measure for the predicted variants, which can help to better select
candidates for further study. The MAQ, SOAPsnp and SNV-Mix1 algorithms all use an underlying
Bayesian framework that models the three diploid genotypes as a panel of Binomial distributions.
SNV-Mix1 stands apart from these methods because it infers the parameters of the distributions as
opposed to using handset parameters, and is shown to improve variant discovery because it models
the properties of the data (Goya, 2010).

We show in Figure 1.5 an example of chromosome 19 from a lobular breast carcinoma (Shah,
2009) that demonstrates a skew in the allelic frequency away from heterozygosity on 19q. Both B-
allele frequency analysis in the array data and allelic ratio analysis in the NGS data suggest allele-
specific amplification on 19q in this genome. This event harbours eight co-existing somatic
mutations (annotated on the karyogram), validated by orthogonal methods, that are undetectable by
analytical methods that assume diploidy (chapter 2). Accurate and sensitive variant calling methods
may therefore require conceptual inclusion of co-existing CNAs in SNV discovery.
Figure 1.5 Novel somatic variants detected in allele-specific amplification on chromosome 19q arm. Genes annotated in the karyogram contain point mutations that were undetectable by methods assuming diploidy, but resolved with CoNAn-SNV. This is further investigated in chapter 2. A) and B) indicate raw log copy number and b allele log R ratios respectively, for normal DNA (from the same patient) on Affymetrix SNPChip 6.0 array. Blue colour indicates diploid (neutral) copy number state; the brighter the colour of red the higher the level of amplification. The three distinct bands in (b) indicate the presence of the alleles harbouring one of the three diploid genotypes: AA, AB and BB. C) and D) shows metastatic tumour copy number and b allele intensity respectively. The high level amplification on the 19q arm is accompanied by B allele intensities that show an absence of the AB heterozygous (middle) band that was present in the normal. E) shows allelic counts from next generation sequencing for the positions represented on the array as a proportion of depth; the allelic ratio is calculated by summing the total number of reads containing a variant at each position divided by the total depth at that position. Also noteworthy is that the 19p arm does not show such a distinct segregation of the counts as in the 19q arm F) shows the raw copy from the NGS data annotated with the amplification information and indicates the same sites of amplification revealed by orthogonal array platform. The normal genome was not processed by next generation sequencing.
1.5 Research Goals

This study seeks to demonstrate that the explicit incorporation of CNA state into SNV discovery is necessary for the complete characterisation of the cancer mutational landscape from NGS data. We demonstrate this goal by developing a panel of Binomial mixture models (described in Chapter 2) that is capable of modelling allele-specific copy number genotypes from NGS tumour data. The model is applied to the lobular breast carcinoma derived from whole genome shotgun sequencing (WGSS) that was previously published in Shah et al. (2009). We verify results derived from the WGSS data set by targeted ultra deep amplicon sequencing. The limitations of the Binomial framework, described in chapter 2, are further studied using the aforementioned WGSS data as well as 30 triple negative breast cancers derived from exon capture sequencing modality. Part of this work is in the process of publication and is presented in chapter 2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Reference</th>
<th>Validated Results</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobular Breast Carcinoma</td>
<td>WGSS</td>
<td>Shah et al. (2009)</td>
<td>yes</td>
<td>2/3</td>
</tr>
<tr>
<td>30 Triple Negative Tumours</td>
<td>EXCAP</td>
<td>n/a</td>
<td>no</td>
<td>3</td>
</tr>
</tbody>
</table>

1.6 Thesis Summary

Chapter 2 of this thesis introduces and describes the CoNAn-SNV model. We demonstrate with revalidated positions that this model improves the sensitivity of SNV discovery. Additionally, we describe how tumour evolution and morphology may be better understood using the allele-specific copy number genotypes. Chapter 3 addresses the limitations of the binomial mixture model. Namely, the maximum a posteriori frame work used by SNVMix and CoNAn-SNV models tends to overfit to training data and cannot cope with the full range of variability existing in genomic data, making it susceptible to outliers. The Beta-Binomial distribution is used as a substitute because it allows a true Bayesian approach and it belongs to a family of distributions that make it more robust to outliers and reduce the tendency to overfit. Chapter 4 highlights the strengths and weakness of CoNAn-SNV and the Beta-binomial framework. It concludes with future work and remarks on the future role of NGS clinical settings.
Chapter 2
Mutation discovery in regions of segmental cancer genomic amplifications with CoNAn-SNV: a mixture model for next generation sequencing of tumours

SUMMARY

In this chapter we present CoNAn-SNV (Copy Number Annotated –SNV): a novel algorithm for the inference of single nucleotide variants (SNVs) that overlap copy number alterations. Section 2.1 shows the method is based on modelling the notion that genomic regions of segmental duplication and amplification induce an extended ‘genotype space’ where a subset of genotypes will exhibit heavily skewed allelic distributions in SNVs (and therefore render them undetectable by methods that assume diploidy). In section 2.3 we applied CoNAn-SNV to a previously published WGSS data set of a lobular breast cancer and show that it is able to detect 24 experimentally revalidated somatic non-synonymous mutations that were not found using copy number insensitive SNV discovery algorithms. A ROC analysis based on ground truth data from an orthogonal assay, shows that the increased sensitivity of CoNAn-SNV does not result in disproportionate loss of specificity. Our results in section 2.4 indicate that in genomically unstable tumours, copy number annotation for SNV detection will be critical to fully characterize the mutational landscape of cancer genomes. Work in this section is from the manuscript Crisan et al. (2010)\(^1\).

---

\(^1\) Crisan et al. (2010) *Mutation discovery in regions of segmental cancer genome amplifications with CoNAn-SNV: a mixture model for next generation sequencing of tumours*. Publication pending at this time.
2.1 The CoNAn-SNV Model

To address the problem of varying allelic states in regions of copy number aberration, we developed a new model, CoANan-SNV (Copy Number Annotated SNV discovery), designed to incorporate knowledge of copy number state at individual positions. Depicted schematically in Figure 2.1A, and as a generative probabilistic graphical model in Figure 2.1B, the model uses a hierarchical Bayes (Gelman, 2004) conditional independence framework for parameter estimation and inference. CoANan-SNV relates to the SNVMix1 model described in Goya et al. (2010), but with important differences; namely that SNVMix1 is incapable of encoding copy number changes commonly found in cancer genomes (such as the 19q amplification shown in Figure 1.5). To overcome this limitation, CoANan-SNV inputs a set of allelic counts and a discrete copy number state for each position in the data. An example of the inputs and output is shown in Figure 2.1C. The objective is to predict which, out of a fixed number of genotypes (informed by the copy number state), would be most likely to have given rise to the observed allelic counts at a given position. The allelic counts are represented as the number of reads $a_i$ at each position $i \in \{1, 2, \ldots, I\}$ that match the reference genome, where $I$ is the total number of positions in the input. We let $N_i$ represent the total number of reads aligned to position $i$ (the depth) in the input. We introduce $C_i$ as the copy number state at position $i$, and we assume $C_i$ is known at run time and is given as input by the user. Theoretically, the full space of allele states could be inferred with knowledge of absolute copy number, however methods for determination of absolute copy number from NGS data remain problematic and in practice it is unlikely that all states could be resolved even with the current sampling depths of NGS (see Discussion). Therefore to a first approximation, we have defined copy number state, $C_i \in \{LOSS, NEUT, GAIN, AMP, HLAMP\}$ where LOSS corresponds to a deletion, NEUT is copy number neutral (diploid), GAIN approximates to low level duplication, AMP approximates to low-intermediate amplification and HLAMP is a high-level copy number amplification.
Figure 2.1 Overview of CoNAn-SNV model, inputs and outputs. A) CoNAn-SNV genotype statespace expansion shown schematically. As higher levels of amplification are encountered, a larger genotype state-space is required to accommodate the different events that could have given rise to amplifications (examples in Figure S1). B) CoNAn-SNV generative probabilistic graphical model. Circles represent random variables, and rounded squares represent fixed constants. Shaded nodes indicate observed data, such as allelic counts, while white nodes indicate quantities that are inferred during training though expectation maximisation. \( C_i \in \{ \text{NEUT, GAIN, AMP, HLAMP} \} \) represents the CNA states of a segment (defined by the HMM describe in Shah et al. (2009)) that spans position \( i \); \( G_i^c \) represents the genotype, which varies depending on CNA state; \( N_i^c \in \{ 0, 1, \ldots \} \) is the number of reads and \( a_i \in \{ 0, 1, \ldots, N_i \} \) is the number of reference reads; \( \pi^c \) is prior existing over the genotypes and extends to accommodate CNA states; and \( \mu_k^c \) is the genotype-specific Binomial parameter for genotype \( k \) in CNA state \( C_i \). C) Example of CoNAn-SNV input and output. CoNAn-SNV takes allelic counts and as well is CNA segment data as input, while SNVMix1 requires only allelic counts. The same positions and counts are provided to both algorithms, with different results. In some cases CoNAn-SNV will call a variant with an aaaa or aab genotype, which would otherwise be missed by SNVMix1; also, however, CoNAn-SNV will also genotype a positions with aabbb rather than bb (as SNVMix1 would), which allows for better interpretation of events.

Here we use the HMM-based method described by Shah et al. (2009). They key intuition in the CoNAn-SNV model is that \( C_i = c \) informs the state space of possible genotypes \( G_i^c = k \) at position \( i \) as follows:

\[
G_i^c = k, k \in \begin{cases} 
\{ aa, ab, bb \} & \text{if } c \in \{ \text{LOSS, NEUT} \} \\
\{ aaaa, aab, abb, bbb \} & \text{if } c = \text{GAIN} \\
\{ aaaa, aaaa, aab, abb, aabbb, bbbb \} & \text{if } c = \text{AMP} \\
\{ aaaa, aaaa, aabbb, aabbb, aabbb, bbbb \} & \text{if } c = \text{HLAMP}
\end{cases}
\]

(2.1)

Loss segments are analysed with a neutral state-space because they present challenges that require considerations which are separate from amplifications and in fact may even need a complimentary normal genome to resolve.
Accounting for copy number gains is especially important when such changes are allele specific, and when the allele that is amplified is the reference allele. For example, consider the case where \( C_i = AMP \), this will induce a genotype state space of \{aaaa, aaab, aabb, abbb, bbbb\}. Our model is therefore theoretically capable of detecting variants with allelic distributions skewed away from heterozygosity (i.e. aaab). We let \( \mu_k^c \) represent the parameter of the Binomial distribution that encodes the expected proportion of reads matching the reference sequence, where for a given position \( i \), \( k \) is the genotype state given by \( G_i^c = k \). We can therefore express the likelihood term for observing the number of reference reads given the depth, the copy number state, the genotype and the model parameters as follows:

\[
p(a_i \mid N_i, G_i^c = k, C_i = c, \mu_k^c) = \text{Binomial}(a_i \mid \mu_k^c, N_i)
\]  

(2.2)

thereby assuming that \( a_i \) is distributed according to the state-specific Binomial distribution indexed by genotype and copy number. We encode a copy-number specific prior over genotypes \( \pi^c \), assuming that the genotypes for copy number state \( c \) are distributed according to a Multinomial distribution with parameter \( \pi^c : p(G_i^c = k) = \text{Multinomial}(G_i^c = k \mid \pi^c, 1) \) for all \( i \in (1, 2, \ldots, I_c) \), where \( I_c \) is the total number of positions which copy number state \( c \). We use Bayes’ rule to compute the posterior probability that genotype \( k \) gave rise to the observed data with the explicit encoding of copy number state:

\[
p(G_i^c = k \mid a_i, N_i, C_i = c, \mu_k^c, \pi^c) = \frac{\pi_i^c \text{Binomial}(a_i \mid \mu_k^c, N_i)}{\sum_{h=1}^{k^c} \pi_h^c \text{Binomial}(a_i \mid \mu_h^c, N_i)}
\]  

(2.3)

where \( k^c \) is the number of possible genotypes for copy number state \( c \) (see Equation 2.1). Given \( p(G_i^c = k \mid a_i, N_i, C_i = c, \mu_k^c, \pi^c) \), we can then choose to compute:

\[
p(SNV_i) = \sum_{v \in V} p(G_i^c = v \mid a_i, N_i, C_i = c, \mu_k^c, \pi^c)
\]  

(2.4)

where \( v \in V \) represents any variant genotype state (i.e. any state that is not aa, aaaa, aaad, etc... as the case may be) to represent a single probability that a position encodes a SNV.
**Algorithm 1 — Training** From a pileup file the algorithm acquires the allelic counts a and b and, through a pre-processing step using the chromosome number and position, assigns them to a copy number c. Hand set α and β parameters are inputted. Finally, the allowable step size and the maximum number of iterations for the expectation maximisation algorithm are supplied.

**Require**: a, b, c, step, maxiter, α, β

1: /*Initialise Parameters*/
2: \( \pi = p(G_j = k) = \text{Multinomial}(G_j = k|\pi, 1) \)
3: \( \mu = \text{Beta}(\alpha, \beta) \)
4: /*Train each copy number state individually*/
5: for each unique(c) do
6: while \( (\ell(\hat{\theta}, \hat{\theta}) \leq \text{step}) \) and \( (\text{iter} \leq \text{maxiter}) \) do
7: if exists(\( \ell(\hat{\theta}) \)) then
8: \( \ell(\hat{\theta}) = \ell(\hat{\theta}) \)
9: end if
10: for i = 1 to \( I_k \) do
11: \( N_i = a_i + b_i \)
12: /* E STEP */
13: \[ p(G_i = k|a_i, N_i, C_i = c, \mu_k, \pi_k) = \frac{\pi_k \cdot \text{Binom}(a_i|\mu_k, N_i)}{\sum_{k=1}^{\nu} \pi_k \cdot \text{Binom}(a_i|\mu_k, N_i)} \]
14: end for
15: /* M STEP */
16: \( \hat{\mu} = \frac{\sum_{i \in I_k} a_i \cdot I(G_i = k) + \alpha - 1}{\sum_{i \in I_k} N_i \cdot I(G_i = k) + \alpha_k + \beta_k - 2} \)
17: \( \hat{\pi} = \frac{\sum_{i \in I_k} I(G_i = k) + \delta_k}{\sum_{j \in I_k} \sum_{i \in I_k} I(G_i = j) + \delta_j} \)
18: /*calculate the log likelihood*/
19: \( \ell(\hat{\theta}) = \sum_{i \in I_k} \sum_k I(G_i = k)[\log \pi_k + \log p(a_i|\mu_k)] \)
20: \( \ell(\theta, \hat{\theta}) = \sum_{i \in I_k} \sum_k p(G_i, \mu_k) \log [\pi_k \cdot p(a_i|\mu_k)] \)
21: iter++
22: end while
23: end for

*a pileup file is a summarisation of the sequence data following alignment to a reference genome. Each line of the pileup has the chromosome number, position, reference bases, an indication of the nucleotide contained in other reads that align to that position (and whether they match or not to the reference) and finally a mapping and base quality. The pileup format is specific to MAQ aligner (Li H., 2008).*
Algorithm 2 — Classification From a pileup file the algorithm acquires the allelic counts \( a \) and \( b \) and, through a pre-processing step using the chromosome number and position, assigns them to a copy number \( c \). Furthermore, the \( \mu \) and \( \pi \) parameters from the Binomial mixture model, pre-determined in training, must also be provided. Finally, a threshold should be assigned, either objectively through alternative means (mentioned in the methods) or subjectively selected.

Require: \( a, b, c, \pi, \mu, \text{threshold} \)
1: for \( i = 1 \) to \( T \) do
2: \( N_i = a_i + b_i \)
3: /*classification*/
4: \[ p(G_i^c = k|a_i, N_i, C_i = c, \mu_i^c, \pi^c) = \frac{\pi_i^c \text{Binom}(a_i|\mu_i^c, N_i)}{\sum_{k=1}^{\infty} \pi_i^c \text{Binom}(a_i|\mu_k^c, N_i)} \]
5: /*Evaluate if position contains an SNV*/
6: \[ p(\text{SNV}_i) = \sum_{v \in V} p(G_i^c|a_i, N_i, C_i = c, \mu_v^c, \pi^c) \]
7: if \( p(\text{SNV}_i) \geq \text{threshold} \) then
8: \( \text{isSNV =TRUE} \)
9: else
10: \( \text{isSNV =FALSE} \)
11: end if
12: end for

Hyperpriors and hyperparameters

We assume \( \pi^c \) is distributed according to a conjugate Dirichlet distribution with parameter \( \delta^c \). This is a user-defined parameter. In our study we set \( \delta^c \) in order to favour non-variant states since most positions in the genome will be homozygous for the reference sequence (i.e. wild-type). We assume \( \mu_k^c \) is distributed according to a conjugate Beta distribution with parameters \( \alpha_k^c, \beta_k^c \). We set \( \alpha_k^c, \beta_k^c \) using the biological intuition that homozygous reference positions will be nearly ‘pure’, with decreasing proportion towards homozygous variant positions. All hyperparameter settings are given in Table 2.1 and were intuitively set. Multiple parameter initialisations showed that the performance CoNAn-SNV is unaffected so long as there is reasonable separation between the genotype-specific \( \alpha \) and \( \beta \) values.
Table 2.1 Hyperparameter Initialisation Values

<table>
<thead>
<tr>
<th>Param</th>
<th>CN State</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Loss/Neut</td>
<td>$1000,500,1$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Gain</td>
<td>$1000,500,200,1$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Amp.</td>
<td>$1000,350,150,50,1$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>HL. Amp</td>
<td>$1000,350,200,100,10,1$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Loss/Neut</td>
<td>$1,500,1000$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Gain</td>
<td>$1,200,500,1000$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Amp.</td>
<td>$1,50,150,350,1000$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>HL. Amp</td>
<td>$1000,350,200,100,10,1$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Loss/Neut</td>
<td>$95,4,1$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Gain</td>
<td>$93,4,2,1$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Amp.</td>
<td>$90,4,3,2,1$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>HL. Amp</td>
<td>$85,5,4,3,2,1$</td>
</tr>
</tbody>
</table>

Model fitting and parameter estimation

Given the free model parameters $\theta = (\mu_k^c, \pi^c)$, we show how to use Equations 2.3 and 2.4 to infer $p(SNV_i)$ for all $i$ in the input data. As stated in Goya et al. (2010), it is advantageous to fit the model to the data using expectation maximization (EM) to learn $\theta$. For CoNaAn-SNV, we treat the data in each copy number state separately and run EM for each set of data independently (see Methods). We describe it briefly here. Let $I_c$ represent the complete set of positions in the input data annotated with copy number state $c$. Iterating over the copy number states $c \in \{\text{LOSS/NEUT, GAIN, AMP, HLAMP}\}$, the E-step consists of computing $p(G_i^c = k \mid a_i, N_i, C_i = c, \mu_k^c, \pi^c)$ using Equation 2.3 for each position $i \in I_c$, and the current estimates of $\mu_k^c, \pi^c$. The M-step re-estimates $\mu_k^c, \pi^c$ with standard conjugate updating:

$$
\hat{\mu}_k^c = \frac{\sum_{i \in I_c} a_i^{I(G_i^c = k)} + \alpha_k^c - 1}{\sum_{i \in I_c} N_i^{I(G_i^c = k)} + \alpha_k^c + \beta_k^c - 2}
$$

$$
\hat{\pi}_k^c = \frac{\sum_{i \in I_c} I(G_i^c = k) + \delta_k^c}{\sum_{j \in I_c} \sum_{i \in I_c} I(G_i^c = j) + \delta_j^c}
$$
The algorithm continues until the complete data log posterior no longer increases. The full training procedure is shown in Algorithm 1. Classification of new data, proceeding training, is described in Algorithm 2.

2.2 Methods

The CoNAn-SNV algorithm was applied to a lobular breast carcinoma and a subset of all predictions were validated using targeted ultra deep amplicon sequencing. Short read sequences obtained from the Illumina Genome Analyzer GA_i were aligned and analysed using the full analytical pipeline described in Figure 2.2. Lobular breast carcinoma WGSS and WTSS sequence reads were aligned using MAQ as outlined in the supplementary methods in Shah et al. (2009). All raw data for this study are available subject to material transfer agreement from the European Genome-Phenome archive (http://www.ebi.ac.uk/ega) under accession number: EGAS00000000054. The data are publically available through this material transfer agreement while also protecting the patient’s identity.

2.2.1 Single nucleotide variant discovery and validation

The model parameters for CoNAn-SNV were estimated by expectation maximization (EM) using 14,649 positions with high confidence SNP calls established as a ground truth standard in Goya et al. (2010). We fit a separate model for each of loss/neutral, gain, amplification and high level amplification sets of positions using expectation maximization in a maximum a posteriori (MAP) framework (Algorithm 1) with hyperparameter settings shown in

Table 2.1. Given the model parameters, we then applied CoNAn-SNV on the full set of WGSS lobular breast carcinoma data. To compute the probability of the presence of a SNV, we summed the posterior probabilities of the variant-containing genotypes (equation 2.4). We then filtered out any positions where $p(\text{SNV})<0.77$, as determined in Shah et al. (2009) to give a false positive rate of less than 1% (Figure 2.2). We used this threshold for accurate comparison against early SNV calls reported in Shah et al. (2009). Remaining positions were filtered against MAQ and SNVMix1 calls leaving only CoNAn-SNV specific predictions for further analysis. In the absence of a normal whole
genome sequence, we relied upon dbSNP positions to filter out germline variants leaving only
coding and non-synonymous candidate somatic variants. Further filtering of polymorphic loci such
as HLA genes and immunoglobulin genes resulted in a candidate list of 200 SNVs that were
submitted for validation using targeted ultra deep amplicon sequencing on the Illumina
GenomeAnalyzer\textsubscript{i} sequencer. Samples of the patient’s normal DNA, primary tumour DNA, and
metastatic tumour DNA were available and these were used to test for the presence of predicted
somatic variants as was done in Shah et al. (2009).

All validation sequence reads were aligned using MAQ to a custom reference genome created
solely from the primer coordinates used to generate the amplicons; the reference is available as
Supplemental Data in Fasta format. A substantial portion of the human genome contains repetitive
sequence and thus reads can align ambiguously to more than one location. It is possible to reduce
the ambiguity by aligning solely to custom genome. The uniqueness of the primer design ensures the
sequence is collected from the targeted genomic co-ordinates. Full details of the primer construction
and lists are in Appendix B. A one-tailed Binomial exact test using the R statistical package was used
to evaluate target positions for presence of the SNV against a null distribution designed to capture
the background error rate. Allelic counts for the five positions immediately flanking the both sides
of the target position were used to establish the null distribution. P-values were corrected using
Benjamini-Hochberg correction, and those with a corrected p-value $\leq 0.05$ were considered to be
present. We use Benjamini-Hochberg correction because it is less stringent than standard Bonferroni
 correction and additionally corrects against type 1 error. This procedure was applied to the normal,
primary and metastatic genomic data from the same patient (Shah, 2009) in order to classify the
predicted mutations. Positions were considered somatic mutations if they were not present in the
normal data, but existed in the tumour data; and germline SNPs if present in the normal and
metastatic tumour genome data. Positions at which a large discrepancy between the metastatic and
normal depth was observed, despite Binomial exact test results, were considered inconclusive and
were not considered. Some germline variants exhibited allelic skew; we objectively evaluated this by
using a chi-squared test comparing the allelic counts of these positions in both metastatic tumour
against the normal genomic data. Positions were considered skewed if the Benjamini-Hochberg
corrected p-values were $\leq 0.05$ with the additional requirement that the frequency of the non-reference allele between the normal and the metastatic had a disparity of at least 10%.

2.2.2 Performance evaluation with OncoSNP and CRLMM

Performance evaluation was completed using an orthogonal Affymetrix SNPChips 6.0 array of the lobular carcinoma. First, we used a well-characterized set of 14649 CRLMM algorithm (Carvalho, 2007) calls described in Goya et al. (2010). In addition, we analysed the SNP array using OncoSNP (Yau et al. unpublished) in order to benchmark CoNA-SNV against an analysis capable of detecting allele-specific CNAs (albeit limited to arrays). Overall, there was a large concordance between CRLMM and OncoSNP genotype calls (498,984 SNP positions) where 15,757 positions were confirmed to be a SNP by both algorithms. A total of 11,369 genotype calls were unique to OncoSNP and mainly, although not entirely, represent allele-specific amplifications where the reference allele is amplified; 4,457 were unique to CRLMM; there were no results for 338,755 positions and these were excluded from analysis. CRLMM unique calls may be due to OncoSNP calibration (see below). We moved forward with 530,567 OncoSNP calls that were filtered prior to performance analysis. SNPChips data reports genotypes based on major and minor alleles, however, it is necessary to translate this to reference and non-reference alleles that is used in the next generation data. If this is not done, the genotypes are inaccurately assigned in situations where the reference-allele is the minor allele. To qualify for further analysis, all positions were required to have a minimum depth of 2, with a minimum mapping and base quality of 10 and 20 respectively. Finally, some positions that are called a variant by OncoSNP, however the NGS data at the corresponding genomic coordinate lacked evidence of any variant reads. These positions either represented a missed call from OncoSNP or an under-sampling of the allele in the sequence data and thus these positions are removed from analysis so as not to artificially augment the false negative rate. Ultimately 12,588 positions passed all criteria of which 4,235 were SNVs and 8,353 were not.
Figure 2.2 SNV discovery Pipeline Using CoNan-SNV
2.3 Results

2.3.1 Experimental validation of the CoNAn-SNV model

To determine the sensitivity and specificity of CoNAn-SNV, we applied the model to the metastatic lobular carcinoma previously published in Shah et al. (2009) and subsequently re-sequenced all the novel, coding, non-synonymous predictions made by the model to establish its accuracy. The genome was segmented into discrete CNA segments using a hidden Markov model as described in Shah et al. (2009) and exhibited a variable CNA landscape. As reported previously, 30.2% of the genome was predicted as loss/netural, 44.5% was gain, 19.1% amplification and 4.2% high-level amplification (see Table S1 and Figure S1a). Figure 1.5 shows chromosome 19 and highlights an example of a somatic high level amplification on the 19q arm that also demonstrates a skew in the allelic frequency, away from heterozygosity, due to an allele-specific copy number amplification. Both B-allele frequency analysis in the array data and allelic ratio analysis in the NGS data support a mono-allelic amplification on 19q in this genome. A re-analysis of the genome with CoNAn-SNV made a total of 63,113 SNV calls in exonic regions of the genome (NCBI build 36.1, Ensembl v51 annotations); compared against 55,357 predictions by SNVMix1 and 56,347 by MAQ (Li H., 2008). Figure 2.3 shows how CoNAn-SNV, MAQ and SNVMix1 predictions are distributed. A total of 52,493 predictions were common to all three methods suggesting reasonable overall agreement. However, 4,308 predictions were CoNAn-specific. In contrast, only 406 positions were specific to MAQ and 0 were specific to SNVMix1. Neutral regions harboured 786 CoNAn-specific predictions while Gain, Amplification and High Level Amplifications harboured 1381, 777 and 1364 CoNAn-specific predictions respectively. Thus there was a substantial enrichment of CoNAn-specific SNVs in CNA amplification states. From the complete list of 4,308 CoNAn-specific predictions, we filtered out positions that were present in dbSNP v130. A set of 200 protein coding, non-synonymous substitution SNVs were identified (see Methods) as candidates for validation by targeted, ultra deep amplicon sequencing (shown schematically in Figure 2.2) in the metastatic and primary (from nine years earlier) tumour genome DNA as well as the normal buffy coat genomic DNA from the same patient.
**Figure 2.3** CoNAn-SNV is capable of retrieving more validated true positives compared with MAQ and SNVMix1. a) Venn diagram of predictions made by Maq, SNVMix1, CoNAn-SNV separated by CNA state. CoNAn-SNV makes more overall predictions especially in high-level amplifications.

A total of 75 SNVs could not be resolved due to PCR amplicon failure during validation, leaving 125 remaining for further analysis. Table 2.2 shows 24/125 (19.2%) novel, coding, non-synonymous somatic mutations that were validated by targeted ultra deep amplicon sequencing. For all of these somatic variants, their predicted genotypes were highly skewed towards the reference allele and had a most probable genotype of aab, aaab or aaaaab (Table 2.2). These amplicons generated an average of 17.4±8.63% reads representing the mutant allele in the metastatic genome (with a mean depth of coverage of 96,669) whereas the normal genome for the amplicons had an average mutant allele frequency of 0.630±0.950% and a mean depth of coverage of 71,963. Note that only one somatic mutation, K187M in ZNF607, a zinc finger protein putatively involved in transcriptional regulation, was also confirmed in the primary tumour. This supports the conclusion from Shah et al. (2009) that few mutations present in the metastatic tumour were present in the primary at diagnosis, and thus were candidate drivers of tumourigenesis. Additionally, we identified 49 (39.2%) germline variants, where the SNV was present in both the normal and metastatic DNA. Finally, 43 (34.7%) positions failed to validate as SNVs and were considered false positive predictions. Nine positions (7.2%) were inconclusive. A full summary of all 200 positions is available in Table S2.

Subheterozygous allele abundance could result from sub-dominant populations of cells or unequal allele amplification in regions of copy number aberration. Notably the mean abundance of the novel somatic SNVs from the validation experiments was 17.38±8.6% with five mutations (affecting genes TDRD5, NCF2, IPO9, ZNF480 and ZSCAN22) exhibiting a proportion of less than 10%. The explicit incorporation of CNA status in SNV discovery allowed such events to be recovered. Without information linking somatic SNVs to germline SNVs, however, interpreting
these heterozygous ratios of alleles is problematic; germline allelic ratios could help confirm whether the copy number amplification involved is predominantly mono-allelic. We examined the allelic ratios for all informative positions in the CNA segments analysed. We found seventeen of the 47 validated germline variants also exhibited substantial allelic skew, as highlighted in Table 2.3. Notably, germline variants at positions chr19: 40691038, chr19:42074256, chr19:50869860 and chr19:59415177 within the high level amplicon on chr19 had allelic distributions in the tumour that were skewed significantly away from their normal distribution (Chi Sq test, q < 0.01). These germline SNPs are proximal to the somatic mutations K187M in ZNF607, R1380H in PRX, E24*2 in PRR19, Q311* in ALDH16A1, E16Q in ZNF480, V328M in LILRA2, Q467H in ZFP28, and G348E in ZSCAN22. The most parsimonious explanation for these findings is that the somatic mutations were a later event, however it is not known if they occur on the amplified or unamplified sister chromosome. A different validation procedure would be required to make this inference. Allele-specific amplification is supported by an additional 463 SNVs within the 19q high level amplicon (chr19: 24301089-63793263 – see Table S3) that were predicted to be either aaaaab or abbbbb by CoNaSnV but were not sent for revalidation. In addition, the OncoSNP algorithm predicted an unbalanced amplification spanning chr19:32439833-63789666. This segment was predicted by OncoSNP, using the orthogonal array data, to contain 638 aaaaab variants, and 591 abbbbbb variants, supporting the conclusion of an allele-specific amplification in 19q. Interestingly, the allelic frequency of K187M in ZNF607, the only somatic variant found in the primary tumour (16.67%) was consistent in the metastatic tumour (15.25%), suggesting that the other 19q mutations occurred later in the tumour evolution.

\(^2\) * represents a stop codon. Thus, the wild type amino acid is converted to a stop codon.
Table 2.2. Somatic single nucleotide variants discovered by CoNAn-SNV

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>AA Mutation</th>
<th>Gene Name</th>
<th>Functional Impact</th>
<th>Ref Base</th>
<th>Mut Base</th>
<th>Depth</th>
<th>Nref count</th>
<th>p(snv)</th>
<th>Genotype</th>
<th>Depth</th>
<th>Frequency of minor allele</th>
<th>Depth</th>
<th>Frequency of minor allele</th>
<th>Transcriprome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31870778</td>
<td>S177L</td>
<td>PEF1</td>
<td>1.11 (M)</td>
<td>G</td>
<td>A</td>
<td>17</td>
<td>3</td>
<td>0.8459</td>
<td>aab</td>
<td>11190</td>
<td>0.0441</td>
<td>617</td>
<td>0.0340</td>
<td>G 2 N 0</td>
</tr>
<tr>
<td>1</td>
<td>177905078</td>
<td>Q872*</td>
<td>TDRD5</td>
<td>Truncating</td>
<td>C</td>
<td>T</td>
<td>105</td>
<td>12</td>
<td>0.9999</td>
<td>aaab</td>
<td>38705</td>
<td>0.0022</td>
<td>2474</td>
<td>0.0020</td>
<td>C 18 N 0</td>
</tr>
<tr>
<td>1</td>
<td>181813423</td>
<td>Q100H</td>
<td>NCF2</td>
<td>0.975 (L)</td>
<td>C</td>
<td>G</td>
<td>68</td>
<td>10</td>
<td>0.9998</td>
<td>aaab</td>
<td>39290</td>
<td>0.0039</td>
<td>1733</td>
<td>0.0069</td>
<td>C 25 N 0</td>
</tr>
<tr>
<td>1</td>
<td>200099346</td>
<td>R539T</td>
<td>IPO9</td>
<td>2.025 (H)</td>
<td>G</td>
<td>C</td>
<td>87</td>
<td>9</td>
<td>0.9913</td>
<td>aaab</td>
<td>18800</td>
<td>0.0026</td>
<td>274</td>
<td>0.0036</td>
<td>C 84 C 11</td>
</tr>
<tr>
<td>2</td>
<td>100965216</td>
<td>E525T</td>
<td>NPA52</td>
<td>1.68 (M)</td>
<td>A</td>
<td>T</td>
<td>45</td>
<td>6</td>
<td>0.8536</td>
<td>aaab</td>
<td>131465</td>
<td>0.0017</td>
<td>15627</td>
<td>0.0022</td>
<td>A 29 T 3</td>
</tr>
<tr>
<td>4</td>
<td>255249422</td>
<td>S-&gt;C</td>
<td>AC133961.3</td>
<td>Truncating</td>
<td>C</td>
<td>G</td>
<td>9</td>
<td>2</td>
<td>0.7930</td>
<td>aab</td>
<td>10999</td>
<td>0.0025</td>
<td>443</td>
<td>0.0045</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>4</td>
<td>175726007</td>
<td>E68K</td>
<td>ARL10</td>
<td>0.55 (L)</td>
<td>G</td>
<td>A</td>
<td>38</td>
<td>6</td>
<td>0.9575</td>
<td>aaab</td>
<td>35722</td>
<td>0.0011</td>
<td>5911</td>
<td>0.0008</td>
<td>G 1 N 0</td>
</tr>
<tr>
<td>4</td>
<td>176953851</td>
<td>E152*</td>
<td>TMED9</td>
<td>Truncating</td>
<td>G</td>
<td>T</td>
<td>46</td>
<td>7</td>
<td>0.9787</td>
<td>aaab</td>
<td>83887</td>
<td>0.0110</td>
<td>40283</td>
<td>0.0109</td>
<td>G 111 T 9</td>
</tr>
<tr>
<td>6</td>
<td>44361861</td>
<td>E222K</td>
<td>TCTE1</td>
<td>0.955 (L)</td>
<td>C</td>
<td>T</td>
<td>24</td>
<td>4</td>
<td>0.7858</td>
<td>aaab</td>
<td>63261</td>
<td>0.0054</td>
<td>4076</td>
<td>0.0064</td>
<td>G 0 N 0</td>
</tr>
<tr>
<td>6</td>
<td>111800869</td>
<td>N1794K</td>
<td>REV3L</td>
<td>0.345 (L)</td>
<td>G</td>
<td>T</td>
<td>35</td>
<td>6</td>
<td>0.9933</td>
<td>aaab</td>
<td>91581</td>
<td>0.0016</td>
<td>54683</td>
<td>0.0020</td>
<td>G 18 T 3</td>
</tr>
<tr>
<td>6</td>
<td>157570350</td>
<td>R2115Q</td>
<td>ARID1B</td>
<td>1.845 (M)</td>
<td>G</td>
<td>A</td>
<td>52</td>
<td>7</td>
<td>0.9353</td>
<td>aaab</td>
<td>304781</td>
<td>0.0024</td>
<td>118051</td>
<td>0.0022</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>7</td>
<td>139446250</td>
<td>L561V</td>
<td>JHDMD1D</td>
<td>0.615 (L)</td>
<td>G</td>
<td>C</td>
<td>37</td>
<td>6</td>
<td>0.9647</td>
<td>aaab</td>
<td>305</td>
<td>0.0000</td>
<td>1</td>
<td>0.0000</td>
<td>C 91 C 30</td>
</tr>
<tr>
<td>11</td>
<td>2383399</td>
<td>I109F</td>
<td>TRPM5</td>
<td>-0.08 (N)</td>
<td>T</td>
<td>A</td>
<td>13</td>
<td>3</td>
<td>0.7705</td>
<td>aaab</td>
<td>100659</td>
<td>0.0045</td>
<td>33904</td>
<td>0.0104</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>11</td>
<td>93999360</td>
<td>V359V</td>
<td>SERPINA9</td>
<td>0.28 (L)</td>
<td>G</td>
<td>A</td>
<td>24</td>
<td>4</td>
<td>0.7858</td>
<td>aaab</td>
<td>61006</td>
<td>0.0219</td>
<td>8291</td>
<td>0.0226</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>13</td>
<td>100417938</td>
<td>V982I</td>
<td>RTL1</td>
<td>0.0805 (L)</td>
<td>C</td>
<td>T</td>
<td>19</td>
<td>4</td>
<td>0.9063</td>
<td>aaab</td>
<td>107685</td>
<td>0.0135</td>
<td>6172</td>
<td>0.0146</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>15</td>
<td>6403457</td>
<td>G313S</td>
<td>SLCL25A23</td>
<td>1.83 (M)</td>
<td>C</td>
<td>T</td>
<td>9</td>
<td>2</td>
<td>0.7930</td>
<td>aab</td>
<td>46019</td>
<td>0.0048</td>
<td>6579</td>
<td>0.0050</td>
<td>C 1 T 2</td>
</tr>
<tr>
<td>19</td>
<td>42881337</td>
<td>K187M</td>
<td>ZNF607</td>
<td>NA</td>
<td>T</td>
<td>A</td>
<td>65</td>
<td>7</td>
<td>0.9540</td>
<td>aaab</td>
<td>2722</td>
<td>0.0026</td>
<td>174</td>
<td>0.1667</td>
<td>T 15 A 1</td>
</tr>
<tr>
<td>19</td>
<td>45591960</td>
<td>R1380H</td>
<td>PRX</td>
<td>0.55 (L)</td>
<td>C</td>
<td>T</td>
<td>33</td>
<td>5</td>
<td>0.9146</td>
<td>aaab</td>
<td>14090</td>
<td>0.0076</td>
<td>35487</td>
<td>0.0063</td>
<td>N 0 N 0</td>
</tr>
<tr>
<td>19</td>
<td>47506592</td>
<td>E24*</td>
<td>PRR19</td>
<td>Truncating</td>
<td>C</td>
<td>T</td>
<td>42</td>
<td>5</td>
<td>0.8074</td>
<td>aaab</td>
<td>47838</td>
<td>0.0018</td>
<td>2712</td>
<td>0.0026</td>
<td>C 5 N 0</td>
</tr>
<tr>
<td>19</td>
<td>56848470</td>
<td>Q311*</td>
<td>ALDH16A1</td>
<td>Truncating</td>
<td>G</td>
<td>T</td>
<td>35</td>
<td>5</td>
<td>0.8964</td>
<td>aaab</td>
<td>75066</td>
<td>0.0036</td>
<td>1935</td>
<td>0.0078</td>
<td>C 5 N 0</td>
</tr>
<tr>
<td>19</td>
<td>57509248</td>
<td>E16Q</td>
<td>ZNF480</td>
<td>1.67 (M)</td>
<td>G</td>
<td>C</td>
<td>51</td>
<td>9</td>
<td>0.9998</td>
<td>aaab</td>
<td>16867</td>
<td>0.0033</td>
<td>1133</td>
<td>0.0071</td>
<td>G 12 C 1</td>
</tr>
<tr>
<td>19</td>
<td>59779115</td>
<td>V328M</td>
<td>LILRA2</td>
<td>1.91 (M)</td>
<td>G</td>
<td>A</td>
<td>42</td>
<td>6</td>
<td>0.9682</td>
<td>aaab</td>
<td>145106</td>
<td>0.0029</td>
<td>60245</td>
<td>0.0028</td>
<td>G 6 N 0</td>
</tr>
<tr>
<td>19</td>
<td>61757367</td>
<td>Q467H</td>
<td>ZFP28</td>
<td>0.805 (L)</td>
<td>G</td>
<td>C</td>
<td>64</td>
<td>7</td>
<td>0.9583</td>
<td>aaab</td>
<td>279</td>
<td>0.0036</td>
<td>7</td>
<td>0.1429</td>
<td>G 4 C 2</td>
</tr>
<tr>
<td>19</td>
<td>63542071</td>
<td>G348E</td>
<td>ZSCAN22</td>
<td>2.99 (H)</td>
<td>G</td>
<td>A</td>
<td>66</td>
<td>7</td>
<td>0.9493</td>
<td>aaab</td>
<td>279784</td>
<td>0.0023</td>
<td>64866</td>
<td>0.0021</td>
<td>G 1 N 0</td>
</tr>
</tbody>
</table>
Table 2.3 Germline variants proximal to somatic SNVs exhibit allelic skew

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>144932587</td>
<td>F218C</td>
<td>AL139152.7</td>
<td>17928</td>
<td>0.3169</td>
<td>18017</td>
<td>0.2164</td>
<td>T</td>
<td>55</td>
<td>G</td>
<td>3</td>
<td>1.27E-102</td>
</tr>
<tr>
<td>1</td>
<td>149999951</td>
<td>I213V</td>
<td>MRPL9</td>
<td>5387</td>
<td>0.2046</td>
<td>8770</td>
<td>0.0409</td>
<td>T</td>
<td>154</td>
<td>C</td>
<td>28</td>
<td>4.29E-211</td>
</tr>
<tr>
<td>1</td>
<td>150543396</td>
<td>R3530S</td>
<td>FLG</td>
<td>61790</td>
<td>0.6191</td>
<td>78410</td>
<td>0.3981</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>146033676</td>
<td>A76V</td>
<td>ZNF7</td>
<td>92012</td>
<td>0.4499</td>
<td>147007</td>
<td>0.2683</td>
<td>C</td>
<td>2</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>33375641</td>
<td>C-&gt;F</td>
<td>AQP7</td>
<td>24722</td>
<td>0.2781</td>
<td>22104</td>
<td>0.1985</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>1.12E-89</td>
</tr>
<tr>
<td>10</td>
<td>29823914</td>
<td>M1259T</td>
<td>SVIL</td>
<td>128591</td>
<td>0.3867</td>
<td>110884</td>
<td>0.4808</td>
<td>A</td>
<td>6</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>390124</td>
<td>N477K</td>
<td>PKP3</td>
<td>37172</td>
<td>0.4601</td>
<td>57560</td>
<td>0.2907</td>
<td>C</td>
<td>11</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>17499485</td>
<td>R357Q</td>
<td>USH1C</td>
<td>101208</td>
<td>0.5595</td>
<td>58749</td>
<td>0.1548</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>65860057</td>
<td>A79T</td>
<td>RIN1</td>
<td>75400</td>
<td>0.4044</td>
<td>97848</td>
<td>0.1738</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>116569101</td>
<td>R710C</td>
<td>SITD2</td>
<td>260320</td>
<td>0.5342</td>
<td>237372</td>
<td>0.1390</td>
<td>C</td>
<td>51</td>
<td>T</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>124827464</td>
<td>E358Q</td>
<td>FEZ1</td>
<td>249388</td>
<td>0.5259</td>
<td>171924</td>
<td>0.1372</td>
<td>C</td>
<td>0</td>
<td>G</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>122455439</td>
<td>R279P</td>
<td>STED8</td>
<td>208542</td>
<td>0.3071</td>
<td>175257</td>
<td>0.4182</td>
<td>G</td>
<td>17</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>36535811</td>
<td>S-&gt;P</td>
<td>KRTAP4-15</td>
<td>1774</td>
<td>0.3207</td>
<td>4409</td>
<td>0.1851</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>1.51E-30</td>
</tr>
<tr>
<td>19</td>
<td>40691038</td>
<td>R-&gt;Q</td>
<td>DMKN</td>
<td>209119</td>
<td>0.5478</td>
<td>247223</td>
<td>0.1696</td>
<td>C</td>
<td>5</td>
<td>T</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>42074256</td>
<td>H426R</td>
<td>ZNF829</td>
<td>6402</td>
<td>0.4531</td>
<td>10867</td>
<td>0.1214</td>
<td>T</td>
<td>1</td>
<td>C</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>50869860</td>
<td>R190Q</td>
<td>GIPR</td>
<td>70793</td>
<td>0.4878</td>
<td>90262</td>
<td>0.1843</td>
<td>G</td>
<td>26</td>
<td>A</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>59415177</td>
<td>R-&gt;K</td>
<td>LILRB3</td>
<td>34753</td>
<td>0.1592</td>
<td>46500</td>
<td>0.0642</td>
<td>N</td>
<td>0</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

CoNAn-SNV retrieves more true positives without compromising overall accuracy

We assessed performance by evaluating the area under receiver operator characteristic curve (AUC) for CoNAn-SNV and SNVMix1. A comparison to MAQ on these same positions was presented in Goya et al. (2010) and showed the SNVMix1 performs better. The positions used as the ground truth were obtained from an Affymetrix SNP 6.0 positions genotyped using CRLMM and additionally with OncoSNP (see Methods). It is important to note that CRLMM assumes diploidy and its calls will therefore be enriched for heterozygous positions that approach expected allelic distributions for diploid genomes. OncoSNP, conversely, extends its state-space to accommodate genotypes induced by CNA events and can therefore capture allele-specific amplifications. As previously noted, OncoSNP calls were concordant with the NGS data and supported that notion that chromosome 1 and 19 have allele-specific amplifications (Table S4 and Figure S1b)
Figure 2.4 OncoSNP and CRLMM receiver operator characteristic curves. A) OncoSNP  B) CRLMM Graphs are shown separated by performance copy number status to show strengths and weakness of each model. Overall, CoNAn-SNV and SNVMix1 perform similarly with slight tradeoffs between the false positive rate and true positive rate. The true positive rate, when the false positive rate is 5%, is highlighted for high level amplification events. Similar model performance is not unexpected as SNVMix is capable of retrieving positions that CoNAn-SNV also finds, with exceptions, however using a diploid genotype space omits important contextual information (See Discussion).
The ROC results for OncoSNP suggest that CoNAn-SNV and SNVMix1 perform similarly, except in regions of high-level amplification. The AUCs for SNVs in regions of GAIN was 0.998 for SNVMix1 and 0.999 for CoNAn-SNV. For amplification and high-level amplification, the AUCs were (0.998, 0.999) and (0.991, 0.998) respectively. Examination of the breakdown of the calls we determine that CoNAn-SNV calls more true positives overall, compared with SNVMix1, but is also subject to calling more false positives. This is supported by the AUC and ROC. The proximity of the AUC measurements suggests that the false positives introduced by CoNAn-SNV do not outweigh the additional true positives retrieved. In fact, for high level amplifications CoNAn-SNV has a higher true positive rate compared to SNVMix1; in Figure 2.4 we highlight this difference in performance at a false positive rate of 5%, which is the permissible false positive allowance in the validation data (see Methods 2.2.1). It should additionally be noted that at lower false positive rates, such as 1% and especially 2% there is a larger disparity in performance CoNAn-SNV and SNVMix1 and there is a lack of stabilisation that appears at the 5% level. At these more stringent false positive thresholds CoNAn-SNVs performance is better than SNVMix1 performance. The ROC for HLAMP is different from the others, due to SNPs harboured in the allele-specific CNA regions of chromosome 1 and 19 that could not be detected by SNVMix1.

CRLMM results are a benchmark for variants that are easy to detect by SNVMix1. Area under ROC curve calculations indicated that CoNAn-SNV performs similarly to SNVMix1 for these positions (Figure 2.5B). The AUC for SNVs in regions of GAIN was 0.979 for SNVMix1 and 0.975 for CoNAn-SNV. For amplification and high-level amplification, the AUCs were (0.991, 0.990) and (0.911, 0.928) respectively. This suggests that the increased sensitivity gained by CoNAn-SNV does not compromise its overall accuracy compared to SNVMix1, which was also suggested using OncoSNP to assess performance. Again, the ROC for HLAMP is different, suggesting here that SNVMix1 performs best. A closer inspection of these calls reveals that CRLMM is capable of calling few SNVs in regions of HLAMP and those it is capable of calling are easy for SNVMix to detect. At a false positive rate of 5% SNVMix1 performs better than CoNAn-SNV however the difference in performance is very slight. Again, as with OncoSNP performance at lower false positive thresholds there is more instability in the true positive rate; however this time SNVMix1 has a higher true positive rate. OncoSNP, which is capable of calling SNVs where one
allele is preferentially amplified compared to the other, challenges the CRLMM results and suggests CoNAn-SNV performers better because the data set is not longer biased towards SNPs detectable by assuming diploidy (see Methods). A comparison to MAQ on these same positions was presented in Goya et al. (2010).

Functional impact of mutations

We assessed the potential functional impact of each of the 24 somatic mutations using MutationAssessor [http://mutationassessor.org] and by comparing the allelic distributions to those seen in RNASeq whole transcriptome shotgun sequencing data obtained on the lobular carcinoma genome (Table 2.2). Truncating mutations were identified in the TDRD5, TMED9, PRR19 and ALDH16A1 genes. In the RNASeq data, the TDRD5 (18 reads matched the reference vs 0 mutant reads) and PRR19 (5 vs 0) genes expressed only the reference allele. TMED9 (111 vs 9) exhibited skewed allelic expression where the reference allele was more highly represented and the ALDH16A1 had 0 reads covering the position of the mutation. In addition, we observed skewed allelic expression in JHDM1D (91 vs 30), IPO9 (84 vs 11) and REV3L (18 vs 3), which matched allelic skew found in the genome (37 vs 6, 87 vs 9 and 35 vs 6 respectively) suggesting that these mutations are being expressed in the transcriptome of the tumour.

The mutation with the highest functional impact, as evaluated by MutationAssessor, was the ZSCAN22, G348E, mutation (FI score 2.99), which affects a C2H2-type zinc finger domain of this protein involved in transcriptional regulation. The IPO9, R539T, mutation also had high functional impact (FI score 2.025). Its role in cancer is not well understood, however it is known to regulate ribosomal protein S7, which couples with MDM2 to stabilise the p53 protein (Chen, 2007). Intriguingly, another somatic mutation with significant functional impact (1.845) is ARID1B R2115Q (predicted to affect its Armadillo-type fold and hence substrate binding). ARID1B is a member of the BAF205 subunit of the SWI/SNF complex involved in chromatin remodelling and plays a role in tumorigenesis, disease progression and therapeutic resistance (Weissman and Knudsen, 2009). Somatic mutations in SLC25A23 and LILRA2 had high functional impact (FI=1.83, and 1.91) although their role in cancer is not well understood.
2.4 Discussion

In this study we showed that the explicit integration of CNA information with SNV discovery is an essential step towards the goal of comprehensive mutational profiling by next generation sequencing of cancer genomes. Unbalanced segmental copy number alterations are very frequent in tumour genomes and the presence of an unbalanced amplification or deletion of DNA would result in altered allelic ratios in randomly sampled sequence. Without incorporating this copy number information, probabilistic models of SNV detection cannot adjust their sensitivity accordingly. CoNAn-SNV incorporates copy number information into a Bayesian mixture model framework, using a reduced copy number space with 6 states. The number of possible allelic states naturally expands with increasing copy number, however at the same time, the number of reads required to reliably distinguish all states, will also increase. At high copy number states, distinction between higher order states differing by one allele would be highly impractical. A pragmatic approach is therefore to reduce copy number to 6 states, in our case inferred by a previously published HMM method (Shah, 2009). To test the sensitivity and specificity of CoNAn-SNV, we first analysed in silico the behaviour of the model in comparison with non-CNA aware SNV callers, using as ground truth the genome of a metastatic lobular breast cancer where many somatic and germline variants have been validated by independent methods. Using the CoNAn-SNV predictions, we validated 24 novel somatic non-synonymous coding mutations predicted that were not identified in the original analysis of this genome (Shah, 2009). All of these variants had allelic skew resulting from copy number amplifications of the reference allele and thus their predictions in this analysis can be directly attributed to the extension of the model to consider CNAs in the inference of SNVs. MAQ and SNVMix1 are capable of identifying allelic skew towards the non-reference allele although would likely characterise such an event with the \( bb \) genotype; this may be considered a loss of information. CoNAn-SNV may classify similar events as \( abb, abbb, \) or \( abbbb \) which may provide a more informative description of the genomic landscape at that location and flag certain events as potentially interesting for validation. Moreover skewed allelism in simpler models might be misconstrued as loss of heterozygosity. CoNAn-SNV rather allows the investigator to infer skewed heterozygosity caused by allele-specific CNAs. Overall, CoNAn-SNV is capable of calling more variants in highly amplified CNAs compared with SNVMix1 and MAQ. Performance metrics
indicate that the false positives introduced by CoNAn-SNV do not outweigh the true positives gained. Of the 200 high confidence CoNAn-SNV calls, we resolved that approximately 62.3% of predicated variants (excluding inconclusive results) successfully validated. It was observed that more variants from high level CNA states validated. It is possible that this arises from difficulty in establishing the boundaries of the CNA segments which may be too broadly defined; some small lengths of lower level amplifications may exist within other CNAs and the extended state-space is applied where it is not needed and detects noise.

Noise and errors will reduce as the capacity of alignment algorithms to accurately map a read improves. Often true variants existing in the data may cause ambiguous alignments that render reads unusable or incorrectly placed. As aligners continue to progress, we expect the false positive rate and true positives rates of CoNAn-SNV will improve and return more accurate results. CoNAn-SNV additionally calls more SNVs in the neutral states compared with SNVMix1 despite sharing a common framework. We propose that explicit consideration of CNAs in training procedures allows for better estimation of parameters which would otherwise be influenced by allelic skew in amplified regions (see section 2.1).

**Implications for inference of mutational heterogeneity, tumour evolution and LOH**

Our results show, on a genome-wide basis, how somatic point mutations can overlap with somatic CNAs in a manner that affects their detection and interpretation. Sub heterozygous somatic SNV allele ratios can arise from sub-dominant populations of cells or from masking of the somatic SNV by amplification of the wildtype allele. To resolve this situation, comparisons of tumour genome evolution are required, as shown by us and others (Shah, 2009; Ding, 2010). In the latter cases, sub dominant clonal evolution could be inferred because subdominant alleles became prevalent in diploid regions of the genomes or where copy number was not altered during progression. Without the possibility of comparison over time and evolution, skewed allelism in regions of CNA must be cautiously interpreted. Our validation data showed germline events in CNAs that exhibited allelic skew, as would be expected of an allele specific copy number aberration. Without appropriate consideration of amplification status, these events may have been misconstrued.
as loss of heterozygosity when in fact the data show that the imbalance results from the amplification of the reference allele rather than hemizygous deletion or copy-neutral LOH events.

**Implications for tumour evolution**

Consideration of the copy number status of the novel somatic variants allows for deeper interpretation of the tumour evolution. Without further validation it is only possible to further speculate upon the plausible events that would have given rise to these somatic aberrations; it is not possible to derive a definitive interpretation of these results, or their impact, without further validation. Allele-specific copy number alterations can arise when a somatic mutation confers a structural instability, leading to a selective advantage for tumour development through amplification (Bianchi, 1989 – see Introduction). The wild-type allele in the sister chromatid is unaffected because its structural stability is not compromised. The resultant genotype for such a driving event is $abb$, $abb$, or $a$bb. The somatic variants observed in this study, however, have $aaa$, $aab$, or $aab$ genotype suggesting that the amplification occurred first (since there is no presence of the variant in the normal DNA - Table 2.2) and that the somatic variants were a later event. The presence of PALB2 mutation indentified in the initial study by Shah et al. (2009) confers genomic instability that may have lead to the amplification event. It is uncertain where the somatic variants occur; whether they exist in the unamplified sister chromatid or the amplified arm or are mixed throughout both. Knowing where the somatic variants occur may have interesting implications for CNA biology. For example, if the all the variants occurred on the amplified arm, then there may be further genomic instability caused by the amplification event that makes its copies susceptible to further somatic mutations. Development of a validation protocol that is capable of revealing these events may be an interesting feature of future work, as well as further investigation of the relationship between CNAs and SNVs.

**Limitations and Future Work**

If the depth at a given position in the genome is large, the Binomial likelihood distribution exhibits very narrow peaks. As such, small deviations from the expected values in regions with substantial depth can produce extremely low likelihoods and result in uninformative likelihoods for
all genotypes. In such cases, the prior probability on genotypes ($\pi$) can dominate the calculation of the posterior and over-influence the final SNV call. The prior probabilities are distributed such that the majority of the probability mass is skewed towards the homozygous reference genotypes. As a result, some true SNVs may not be correctly classified. The rigidness of the model will cause overfitting; although the models are trained on a SNPChip array of the same tumour, the positions on this array may not be representative of the tumour’s overall genomic architecture. The rigidness is due to MAP estimation approach, which is not a fully Bayesian approach as it draws only a point estimate from the Beta distribution governing the $\mu$ parameter. Rather than using a point estimate it is better to generate a joint posterior distribution of the Beta and Binomial distribution. This reduces the tendency to overfit because the Beta distribution is an objective distribution and uninfluenced by the nature of the data. By setting appropriate strengths of the hyperparameters governing the Beta distribution it is possible to encode more variance in the model, to further prevent overfitting (this is further described in Chapter 3). This true Bayesian approach, however, requires computationally intensive frameworks, such as Markov Chain Monte Carlo methods, to carry out the calculations; whereas it is simple and efficient to use a MAP estimate. Furthermore, both methods, MAP estimation and a full Bayesian approach, are susceptible to outliers. It is possible to resolve the issues with computational efficiency, overfitting and susceptibility to outliers by using a Beta-Binomial distribution instead (Chapter 3).

Moreover, the CoNAn-SNV model is restricted to the possible state space of genotypes provided in the input data. Joint and simultaneous inference of copy number and genotype is a theoretically more attractive approach since genotype could influence the estimation of copy number and vice versa, however such a model is substantially more complex and is beyond the scope of this contribution.

Finally, the data presented here are done using only one dataset. At the time of this manuscript it was not possible to retrieve additional data sets to re-iterate the results of the CoNAn-SNV analysis. Obtaining further datasets is a critical item of future work.
2.5 Conclusion

The primary objective of this study was to explore how the consideration of CNA annotation in SNV discovery impacts the analysis and interpretation of NGS data from genomically unstable tumour genomes. We show that explicit integration of copy number information into algorithms of SNV detection not only increases sensitivity, but allows the significance of somatic mutation frequency in diploid and non-diploid regions to be more appropriately interpreted. The discovery of 24 new somatic mutations in the lobular breast cancer reveals how incorporation of CNAs into SNV analysis is essential to approaching comprehensive characterization of the somatic mutational landscape tumours by next generation sequencing technology.
Chapter 3
Beta-Binomial extension of the Binomial Framework applied to Whole Genome Shotgun Sequencing and Exon Capture

SUMMARY

In this chapter experiments that address limitations of the Binomial mixture model, expressed in the previous chapter, are presented. Section 3.1 proposes a solution by substituting the Binomial likelihood distribution for a Beta-Binomial distribution. The Beta-Binomial model is able to resolve issues of overfitting, and outliers without increasing computational complexity. Section 3.2 describes the methods for comparing the Binomial and Beta-Binomial mixture models with simulated and real data. Real data for comparisons includes the lobular breast carcinoma WGSS data (Shah, 2009) and 30 exon capture sequenced tumours. For simplicity and ease of interpretation, initial comparisons of the Binomial and Beta-Binomial are done using only the three diploid genotypes; however comparisons with CoNAn-SNV are also performed. For the exon capture data, a cross validation method is used to explore the parameter space of the diploid Binomial and Beta-Binomial models. Complementary material for exon capture is presented in appendix A. In section 3.3 the results of the simulated and real data tests are compared. It is observed that the overall performance, evaluated by area under the receiver operator characteristic curve, of the Binomial and Beta-Binomial models is similar, with the Beta-Binomial performing marginally better. Section 3.4 concludes by highlighting the strengths and weakness of using the Binomial compared to the Beta-Binomial mixture model and identifies scenarios where use of one over the other may be favourable.
3.1 Introduction

This introduction is divided into two orthogonal parts. The first part introduces the Beta-Binomial distribution, focusing on its derivation and differences from the Binomial distribution. The second section is a brief introduction of the exon capture sequencing modality. The goal of this chapter is to understand the utility of the Beta-Binomial in overcoming the limitations observed in the Binomial distribution (Section 2.4) and evaluate these differences in real data from WGSS and exon capture data sets. There are, however, significant contributions to the analysis of the exon capture derived data, presented in appendix A, that also effect the interpretation of results and their outcome.

3.1.1 The Beta-Binomial distribution

The dominant and recessive alleles of living organisms are an exemplarily model of the Binomial distribution. The values of the parameters for these distributions are intuitively assigned based on the biology of a diploid genotype. As described in the introduction of this work and in chapter 2, the genotypes from next generation sequencing are a measure of the purity of reference or non-reference allele at a given position in the genome. For the homozygous reference genotype, $aa$, the expectation is little to no presence of the non-reference allele and that $\mu_{aa} \equiv 1$; whereas for the heterozygous genotype, $ab$, we expect to see a 1:1 ratio of the reference and non-reference alleles, $\mu_{ab} \equiv 0.5$ and so on. Due to its ease of use and natural applicability the Binomial distribution was incorporated into variant discovery tools for next generation sequencing.

NGS data are variable due to differences in mapability, quality of base call and alignment. Initialising a singular value for $\mu_k$ may be limiting because few positions would model it exactly. Section 2.4 describes a scenario where high depth positions produce low likelihoods for allelic ratios that are only slightly different from the expected ratio; the result is a reliance on the prior, which is biased towards $aa$, to genotype the data. Such a scenario would be better handled by long tailed distributions, allowing for clearer assignment of genotypes by appropriately weighting the observed allelic ratios (Gelman, 2004). Specifically, positions with ratios that are outlying would be assigned to a more appropriate genotype, which would rescue some positions from being genotyped as $aa$ since the observed allelic ratios would be treated as an outlying event for $aa$ and not for $ab$. 

37
Figure 3.1. The effects of $\alpha$ and $\beta$ on the shape of the Beta distribution for genotype $ab$. The y-axis demarcates $p(\mu_k|\alpha, \beta)$, and the x-axis is the values of $\mu_k$ from 0 to 1. As the $\alpha$ and $\beta$ values have a greater strength the distribution becomes more peaked. Broader Beta distributions spread the probability mass to encode more variance, with the converse being true of more peaked distributions. The highest point on the distribution, reflects the posterior mode, and is indicated by the arrow; in this case the posterior mode is $\mu = 0.5$

The ultimate goal of SNV discovery is to predict whether a given position contains a variant; the parameter $\mu_k$ is a nuisance parameter and we would rather not worry about its value. It is possible to use a statistical technique called marginalisation to avoid the calculation of the nuisance parameter. Marginalising out the value of $\mu_k$ has the advantage of using a true Bayesian solution to variant prediction, over using a point estimate (see Limitations, Section 2.4). It also reduces overfitting and is robust to outliers without increasing computation complexity. To understand the advantage of this technique, it is essential to re-evaluate how the $\mu_k$ is set using the expectation maximization (EM) algorithm, as described in chapter 2.

The Binomial mixture model described in chapter 2 attempts to predict the genotype at a given position in the genome through the Bayesian tenant that $posterior_k \propto prior_k \times likelihood_k$. The prior referred to here is the prior probability of each genotype. In the discussion that follows, we focus on the likelihood, which is the Binomial distribution parameterised by $\mu_k$, where $\mu_k$ itself is represented by a Beta distribution parameterised with the hyperparameters $\alpha$ and $\beta$: $p(\mu_k|\alpha, \beta) = Beta(\mu_k | \alpha, \beta)$ (Figure 3.1). The Beta distribution describes the continuous variable $\mu_k$; although it is intuitive to consider $\mu_k$ as constant since it is straightforward to select, realistically the values of $\mu_k$ vary on a continuous scale depending on the context of problem. In context of sequence data it is
position dependant: at a given position in the genome the value of $\mu_k$ would be different depending on depth and number of reads that match the reference. Pooling all possible values of $\mu_k$ for each genotype, assuming that the genotype is known a priori, would result in a different Beta distribution for each of the genotypes. Rather than doing this, we set values of $\alpha$ and $\beta$ that reflect the expected shape of the Beta distribution for a given genotype; this is done in Figure 3.1 for the $\alpha\beta$ genotype. The strength of $\alpha$ and $\beta$ hyperparameters influences the shape the of the Beta distribution and as $\alpha \to \infty$ and $\beta \to \infty$ the distribution become more highly peaked. The width of the distribution encodes variance; the wider the distribution the mode variance is encoded (Gelman, 2004). The most frequently occurring $\mu_k$, if one samples randomly from the distribution, is the posterior mode (Figure 3.1). A true Bayesian approach is to resolve a posterior density from joint probability of the Binomial and conjugate prior (Beta), however this task may be non-trivial and different approximation methods may need to be used. One approach is to calculate the posterior mode from the training data, which is done with SNVMix1; although this approach may result in over fitting and is susceptible to the influence of outliers, it is computationally less expensive.

Given all of the training data, it is possible to express $p(\mu_k|a_T) \propto Beta(a_1^{1:T} + \alpha, b_1^{1:T} + \beta)$, where $a_i$ is the number of reads supporting a reference allele and $b_i$ is the number of reads supporting the non-reference allele at position $i$, and $T$ is the total number of training positions. As $T \gg \alpha$, the $\alpha$ and $\beta$ parameters have a diminished influence on estimating $\mu_k$ and training data has greater influence on setting its value. For each iteration of EM, we update $\mu_k$ equations described in Chapter 2 using the maximum a posterior (MAP) estimate:

$$\hat{\mu}_k = \frac{\sum_{i \in T} a_i^{(G_i = k)} + \alpha_k - 1}{\sum_{i \in T} N_i^{(G_i = k)} + \alpha_k + \beta_k - 2}$$ (3.1)

We summarise $\hat{\mu}_k$ updated by the training data as $p(\hat{\mu}_k|D)$. We would like to inform our analysis with training data in order to predict variants, via selecting the appropriate genotype, from new data, $a$: $p(a|D)$. It is necessary to use $\hat{\mu}_k$ in order to make these predictions, such that:

$$p(a|D) = p(a|\hat{\mu}_k)p(\hat{\mu}_k|D)$$ (3.2)

39
However, it is possible to use the technique of marginalizing out \( \mu_k \) to avoid explicitly calculating its value (\( \hat{\mu}_k \)) by integrating it out of the equations that define the Binomial and Beta distributions.

\[
p(a|D) = \int p(a|\mu_k)p(\mu_k|D)d\mu_k
\]

(3.3)

We use integration because \( \mu_k \) is a continuous variable described by Beta distribution that spans the interval [0,1]; we integrate over all values in the interval [0,1] as follows (Murphy, 2010):

\[
p(a|D) = \int p(a|\hat{\mu}_k)p(\hat{\mu}_k|D)d\hat{\mu}_k
\]

\[
= \int_0^1 \text{Binomial}(a|\hat{\mu}_k,N) \text{Beta}(\hat{\mu}_k|\alpha,\beta)d\hat{\mu}_k
\]

\[
= \int_0^1 (N)\hat{\mu}_k^a (1-\hat{\mu}_k)^{N-a} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\hat{\mu}_k^{\alpha-1}(1-\hat{\mu}_k)^{\beta-1}d\hat{\mu}_k
\]

\[
\equiv \text{BetaBinom}(a|\alpha,\beta,N) = \frac{\text{Beta}(\alpha,a,N-a+\beta)}{\text{Beta}(\alpha,\beta)}(N)
\]

What has changed through the integration is that instead of resolving the posterior mode, \( \hat{\mu}_k \), we now have a posterior predictive density that is a weighted average the predictions over all values of \( \mu_k \) in the interval [0,1](Murphy, 2010); this resulting distribution is called the Beta-Binomial. It can be thought of as a Beta mixture of binomials, where the allelic counts are binomial observations with unequal probabilities that follow a Beta distribution (Gelman, 2004). It can be shown that this weighted average is better than selecting a singular point from the distribution of \( \hat{\mu}_k \) and in section 3.3 we will demonstrate this with real and simulated data (Murphy, 2010). The advantage of incorporating the \( \alpha \) and \( \beta \) hyperparameters for calculating the posterior density, rather than just for EM updating, is that it becomes possible to encode variance into prediction and diminish overfitting to the training data (Gelman, 2004). This is because in MAP estimation, we do not evaluate the value \( p(\mu_k) \) when calculating the posterior density, instead we only consider the likelihood of observing \( a \) given \( \mu_k \), effectively assigning \( p(\mu_k) = 1 \). This can be resolved by using a true Bayesian approach, such as a Markov Chain Monte Carlo framework, however this approach is not robust against outliers and is computationally expensive. The final advantage of the Beta-Binomial distribution, in addition to being a true Bayesian approach, is that it belongs to the family of long-tailed distributions, which are able to appropriately downweight the effects of outliers (Gelman, 2004).
It is important to remember that the posterior predictive density mentioned here comprises the likelihood, and that it must still be multiplied by the prior in order to evaluate the posterior density for the genotype \( p(G_i = k | a_i, N_i, \pi) \). By integrating out \( \mu_k \) it is no longer to use training data to approximate it and we can use just using hand-set values \( \alpha \) and \( \beta \) to parameterise the Beta-Binomial distribution and perform further calculations although it is still possible to use inference to approximate \( \pi \). It is possible to define a regimen that would update \( \alpha \) and \( \beta \) given training data; however this task is non-trivial and is not the subject of this work.

### 3.1.2 Exon capture

The advent of next generation sequencing allowed for the enumeration of disease causing mutations in entire genomes. The high-cost of entire genome sequencing, however, limits its availability to the scientific community (Hodges, 2007). Furthermore, whole genome sequencing studies often focus on interpreting results only in the context of genomic coding regions. To address limitations in cost while promoting high throughput interrogation of regions of interest, exon capture was developed as cost-effective platform that sequences only the protein coding subset of an entire genome (Hodges, 2007).

Exon capture sequencing operates by pulling down exons of coding regions from a fragmented whole genome. It does this by hybridising the fragments to a pre-designed oligonucleotide array to capture desired exonic regions (Hodges, 2007). The fragments that do not hybridise are washed away, and those remaining are sequenced using the Illumina GenomeAnalyzerII platform. Following sequencing, the reads are aligned to a genome and submitted for further processing by variant discovery algorithms. In an idealized scenario, only exons targeted by exon capture will be pulled down for sequencing. In reality, the data are nosier. Appendix A describes a novel alignment regimen for exon capture data that improves sensitivity. It highlights how different alignments for the exon capture data affect results. In this study, all of the exon capture cases are from triple negative breast cancer. This type of breast cancer is characterised by the absence of ER, HER2 and PR amplification, making it and aggressive and difficult to treat (Dent, 2007).
3.2 Methods

Resolving an integral is often not an easy or intuitive task. There exist alternatives to approximate this integration or avoid it altogether. The SNVMix framework, as previously mentioned, uses MAP estimation to calculate the posterior mode of the Beta distribution. Through iterations of EM the posterior mode is better approximated, based on the data, and the updating relies on the values of $\alpha$ and $\beta$ to guide the approximation. Other methods for approximating the integral includes Markov Chain Monte Carlo frameworks, which more faithfully attempt to simulate effects the integration by sampling multiple points from the Beta distribution. The more points sampled the better the approximation; however sampling will never produce the exact target posterior distribution (Gelman, 2004). The Beta-Binomial framework, presented herein, would be superior to both of these alternatives; however in this work the comparison is made only to the SNVMix, MAP estimation, framework as it is the basis of CoNAn-SNV.

Data were simulated across various depths ($N$) and the normalised posterior probability (section 2.1) is calculated for each $\alpha \in [0, N]$ (where is $a$ the number of reads that match the reference allele) from the Binomial and Beta-Binomial distributions. In this manner the full transitional range of the distributions, given a set of $\alpha$ and $\beta$ values, can be explored. Initial values of $\alpha$ and $\beta$ are set by hand and with adjustments to the strength of $\alpha$ and $\beta$ in order to understand its effect on prediction. The $\alpha$ and $\beta$ values are first set according to the initialised values of SNVMix described in Goya et al. (2010): $\alpha = [1000, 500, 1]$ and $\beta = [1500, 1000]$ and $\tau = 100$. The strengths, $\tau$, are sequentially decreased such that $\alpha$ and $\beta$ values are: $\alpha = [100, 50, 1]$ and $\beta = [150, 100]$ and $\tau = 10$; $\alpha = [10, 5, 1]$ and $\beta = [15, 10]$ and $\tau = 1$. Evaluating the effects of different mixed $\tau$ was done using $\alpha = [50, 5, 1]$ and $\beta = [1, 5, 10]$.

This exercise is repeated with the lobular breast carcinoma WGSS data. A list of candidate SNPs are called, as described in section 2.2 of the previous chapter, by using an orthogonal array to generate a list of 14,649 positions with a high confidence genotype calls from CRLMM. For each set of parameters, we compute the threshold defined at a false positive rate (FPR) < 0.1% as well as the F-measure (Goya, 2010). The WGSS data serves as an additional quality benchmark for the exon capture data, which is used in the subsequent step. The comparison between the two sequencing modalities is important because exon capture remains a novel and experimental data
type. It may be ambiguous to determine whether performance of the Binomial and Beta-Binomial mixtures models is attributed to the pre-processing of the data or the probabilistic framework of the mixture models. Thus, benchmarking against better understood, and possibly more stable, WGSS data gives important contextual information about model performance.

Finally, the broader effects of \( \alpha \) and \( \beta \) values are explored by permuting over all of the possible combinations of parameters in a defined space. The hyperparameter space is intuitively assigned, such that the \( \alpha \) and \( \beta \) values, for the Binomial and Beta-Binomial distributions, can range as follows: \( 9.9 \alpha_{aa} \geq 7.0 \) and \( 0.1 \leq \beta_{aa} \leq 0.3 \); \( 6.5 \alpha_{ab} \geq 3.5 \) and \( 3.5 \leq \beta_{ab} \leq 6.5 \); finally, \( 0.1 \leq \alpha_{bb} \leq 3.0 \) and \( 0.1 \leq \beta_{bb} \leq 3.0 \). These ranges are established such that the allowable combinations of \( \alpha \) and \( \beta \) for each of the genotypes are biologically interpretable (namely \( \mu_{ab} \geq \mu_{ab} \geq \mu_{bb} \) for the Binomial mixture model). The \( \tau \) of the parameters are set such that \( [a_k, b_k] \times 10^{exp} \) where \( exp \in [0, 1, 2] \). A total of 343 combinations are tested at 3 different strengths for a grand total of 1029 combinations of \( \alpha \) and \( \beta \) values; mixed weights are not explored within this defined space. To explore the parameter space the values of \( \mu_k \) are not updated for the Binomial mixture model but simply set to \( \frac{a_k}{a_k + b_k} \). Due to this, the Binomial mixture model will be invariable to the strength of the hyperparameters.

The data are tested on 30 exon capture triple negative breast cancer cases with paired Affymetrix SNPChip 6.0 array data. Important details on the alignment of the exon capture data are presented in appendix A. Candidate SNP lists for each case are called using OncoSNP (Section 2.2). Twenty-five of the 30 cases are used in 5-fold cross validation (CV). For each iteration of CV the f-measure, area under the receiver operator characteristic curve, and threshold where the FPR \( \leq 1\% \) are evaluated for each case in the validation set for the Binomial and Beta-Binomial mixture model. Lastly, the update equation for \( \pi_k \) is the same between the Binomial and Beta-Binomial mixture models and its value is estimated using the training data and EM. A total of five cases out of thirty are withheld and serve as the final validation set against which to test the best performing \( \alpha \) and \( \beta \) values as evaluated by the highest average AUC and smallest standard deviation between the 25 CV cases. Knowing the best performing parameters, it is possible to compare against the intuitively assigned parameter performance. Furthermore, with the Binomial mixture model it is possible to evaluate whether the EM algorithm will update to the optimal performing parameters starting when initialised to intuitive parameters.
It follows naturally to evaluate the utility of the Beta-Binomial as the underlying framework for CoNaAn-SNV. For this comparison the lobular breast carcinoma WGSS data is used because there is reliable and validated ground truth information. The parameters used for the Binomial and Beta-Binomial CoNaAn-SNV implementations are those in Table 2.1 (pg 17); there is no EM updating employed for the Binomial CoNaAn-SNV framework. The Beta-Binomial implementation is tested at $\tau = [1, 10, 100]$. For all runs, the total number of SNVs and the total number of validated somatic variants called is evaluated.

### 3.3 Results

To initially understand the differences between the Binomial and Beta-Binomial mixture models, data are simulated to varying depths and for each $a \in [0, N]$ the normalised posterior probability is calculated (Section 2.1). The values of $N$ are chosen to reflect varying depths expected to be encountered in exon capture data, and for each depth we observe the transition between the three diploid genotypes given changes in number of reads that match the reference allele (Figure 3.1). The $\alpha$ and $\beta$ parameters, as well as the genotype priors, are manually set to explore specific aspects of the distribution’s behaviour with respect to changes in the parameters. The $\alpha$ and $\beta$ values are first set according to the initialised values of SNVMix describe in Goya et al. (2010): $\alpha = [1000, 500, 1000]$ and $\beta = [1, 500, 1000]$. At this strength, $\tau$, the Beta-Binomial is unable to distinguish between the genotypes at depths above 50 (Figure 3.1A). For the Binomial mixture model, the transition between the genotypes becomes very sharp and peaked, however the three do not converge as with the Beta-Binomial. Additionally, at depth of 50, the transitions between the genotypes are not as peaked compared with the Beta-Binomial. The $\tau$ of the hyperparameters is decreased by an order of 10 such that $\alpha = [100, 50, 1]$, $\beta = [1, 50, 10]$; the results are shown in Figure 3.1B. The Beta-Binomial can disambiguate between genotypes at higher depths and, compared against Binomial mixture model, does not as have steep transitions between the genotypes. Relaxing the $\tau$ of the parameters by an additional order of 10 to $\alpha = [10, 5, 1]$, $\beta = [1, 5, 10]$ results in a broader genotype transitions for the Beta-Binomial.
Figure 3.2. Beta-Binomial and Binomial mixture model comparisons with simulated data. The X-axis is the total number of reads that match the reference from 0 to N. The Y-axis is the normalised posterior probability.  

A) $\alpha = [1000,500,1]$ and $\beta = [1,500,1000]$  
B) $\alpha = [100,50,1]$ and $\beta = [1,50,100]$  
C) $\alpha = [10,50,1]$ and $\beta = [1,50,100]$
Figure 3.3. Beta-Binomial and Binomial mixture model comparisons with simulated data. The X-axis is the total number of reads that match the reference from 0 to N. The Y-axis is the normalised posterior probability. \( \alpha = [50,5,1] \) and \( \beta = [1,5,10] \)

Using an asymmetric set of \( \alpha \) and \( \beta \) values \( (\alpha = [50,5,1] \) and \( \beta = [1,5,10] \)) can shift the distributions as shown in Figure 3.3. At lower depths the Binomial mixture models shows a slight shift toward the right, compared to Figure 3.2, however this is absent as depths become higher. Conversely the Beta-Binomial mixture model has a distinct shift to the right at all depths that causes a sharper transition between the \( ab \) and \( aa \) genotypes. The shift is caused by the mixture in \( \tau \). The sharp transition between the \( aa \) and \( ab \) genotypes is reflected in Figure 3.2B, whereas the broadness between the \( bb \) and \( ab \) is a consequence of lower \( \tau \) \( \alpha \) and \( \beta \) values as reflected by Figure 3.2C.

The Binomial and Beta-Binomial mixture models were assessed against real data from the WGSS lobular breast carcinoma (Shah, 2009). We use 14,649 training positions for which we know the genotype from an orthogonal Affymetrix SNPChip array. The depth of the positions varies from 2 to 160 and there were a total of 4730 SNVs.
The exercise with the simulated data is repeated with real data from the WGSS lobular breast carcinoma. The best performance is achieved using the Beta-Binomial mixture model at $\tau = 1$ and 10, as well as the mixed strengths (Figure 3.4). Two different sets of prior probabilities are used; the first is set are uniform priors and the second are priors trained from the WGSS lobular breast carcinoma using SNVMixI (Shah, 2009). There is no difference in performance using the uniform or trained priors from for $\tau = 1, 10$ and mixed strengths, however at higher strengths the uniform priors perform better, although the difference is not large. The trained probabilities perform only marginally better than the uniform probabilities using Binomial distribution. At $\tau = 100$, it can be expected that the Beta-Binomial distribution would have difficulties making SNV calls, as suggested in Figure 3.2. This is not the case because the mean depth of the WGSS data is 34, based on Figure 3.2 the Beta-Binomial distribution would encounter difficulties at depths greater than 50. Generally,

3 The prior probabilities are a multinomial distribution used to distinguish between genotypes and is not related directly to $\mu_k$, which is governed by a separate, Beta, prior.
all of the models perform similarly well. It can be summarised that the lower strengths and mixed strengths have a lower false positive rate and a high true positive rate. As with the ROC curves shown in chapter 2, the greatest differences in performance are a lower false positive thresholds. Finally, we explore a defined space of \( \alpha \) and \( \beta \) values with various permutations. For each set of \( \alpha \) and \( \beta \) values, we use a training set (see Methods in 3.2) to train the prior \( \pi \). Figure 3.5A shows the changes in the AUC calculations only for parameter sets at the beginning [1-49], middle [148-196] and end [295-343] for the Beta-Binomial mixture model at \( \tau \) of 1. There is no significant difference in the performance of the Beta-Binomial mixture model for the different parameter settings, although the highest AUC values are observed as \( \alpha_{as} \) approaches 9.9 and \( \beta_{as} \) approaches 0.1. Figure 3.5B shows the Binomial mixture model performance over the same parameter sets in Figure 3.5A also at strength 1. The AUC values for the Binomial distribution are not as high, or as tight as they are for the Beta-Binomial at the different parameter settings, although the highest AUCs still occur in the end parameter sets. Figure 3.6 takes the end [295-343] parameters sets and shows their performance at \( \tau = [1, 10, 100] \). The performance of the Beta-Binomial distribution decreases as \( \tau \) of is increased. The most significant decrease in performance begins at set 310 and continues periodically until parameter set 331. In this interval, the parameters are: \( \alpha = [9.9, y, z] \) and \( \beta = [0.1, w, x] \), where \( y \in [4.5, 5.5], w \in [5.5, 4.5], x \in [0.1, 3.0], \) and \( z \in [7.0, 9.9] \); the dip in performance occurs as \( \mu_k \) approaches 0.5. This may be due to the biased representation of alleles by exon capture. The Binomial is unaffected by changes in strength, and at strength 100 surpasses the performance of the Beta-Binomial at certain \( \alpha \) and \( \beta \) settings.

The best performing set of parameters is determined by the highest average AUC with the lowest standard deviation over the 25 cross validation cases (each case will serve as part of the validation sets in the 5 fold cross validation). For the Beta-Binomial, those parameters are: \( \alpha = [9.9, 6.5, z] \) and \( \beta = [0.1, 3.5, x] \), where \( x \in [0.1, 3.0] \) and \( z \in [7.0, 9.9] \), where all sets of parameters have an average AUC of 0.9902 and a standard deviation of 0.0047. The best performing set of parameters for the Binomial mixture is \( \alpha = [9.0, 6.5, 3.0] \) and \( \beta = [1.0, 3.5, 7.0] \) with an average AUC of 0.9763 and standard deviation of 0.0103.
Beta-Binomial Mixture Model

B

Binomial Mixture Model

Figure 3.5 Distribution of AUCs for the Binomial and Beta-Binomial mixture model at strength 1 over a defined parameter space. Shown are box plots of the AUC across the training cases. A box plot shows, from bottom to top, the smallest observed value, the lower quartile, median and upper quartile followed by the largest observation; red crosses indicate outlier values. Notches are comparison intervals, if notches on boxes do not overlap it is possible to conclude with 95% confidence that the means differ. Only the beginning, middle and end of the parameter sets are shown here as performance is relatively similar throughout. For sets 1 to 7, $\alpha_{aa}$ is held constant at 7.0 while the $\alpha_{ab}$ and $\alpha_{bb}$ permute over all allowable values (the same is true for the $\beta$ value). At set 295 $\alpha_{aa} = 0.8$ and at set 295 $\alpha_{aa} = 0.99$. Note the difference in scale between the Binomial and Beta-Binomial plots, this is done to more clearly show the Beta-Binomial AUC values.
Figure 3.6 Effects of changing parameter strength on the Beta-Binomial mixture model. Shown are box plots of the AUC across the training cases. A box plot shows, from bottom to top, the smallest observed value, the lower quartile, median and upper quartile followed by the largest observation; red crosses indicate outlier values. Notches are comparison intervals, if notches on boxes do not overlap it is possible to conclude with 95% confidence that the means differ. Arrow indicates the intuitive (default) parameters. The decrease in performance at $\tau=100$ may be attributed to bias allelic representation by exon capture.
Table 3.1. AUC comparison of the best performing $\alpha$ and $\beta$ hyperparameters of the Binomial and Beta-Binomial mixture model against 5 validation cases withheld from CV training.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\pi$</th>
<th>Model</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
<th>Case 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9,6.5,2.5</td>
<td>0.1,3.5,7.5</td>
<td>0.618,0.121,0.261</td>
<td>BetaBin</td>
<td>0.9932</td>
<td>0.9904</td>
<td>0.9931</td>
<td>0.9893</td>
<td>0.9903</td>
</tr>
<tr>
<td>99,65,25</td>
<td>1,35,75</td>
<td>0.615,0.151,0.233</td>
<td>BetaBin</td>
<td>0.9844</td>
<td>0.9807</td>
<td>0.9817</td>
<td>0.9754</td>
<td>0.9796</td>
</tr>
<tr>
<td>990,650,250</td>
<td>10,350,750</td>
<td>0.616,0.146,0.238</td>
<td>BetaBin</td>
<td>0.9749</td>
<td>0.9662</td>
<td>0.9762</td>
<td>0.9716</td>
<td>0.9747</td>
</tr>
<tr>
<td>99,65,25</td>
<td>1,35,75</td>
<td>0.614,0.155,0.231</td>
<td>Binom</td>
<td>0.9753</td>
<td>0.9643</td>
<td>0.9754</td>
<td>0.9724</td>
<td>0.9760</td>
</tr>
<tr>
<td>9.0,6.5,3.0</td>
<td>1.0,3.5,7.0</td>
<td>0.672,0.040,0.289</td>
<td>BeatBin</td>
<td>0.9896</td>
<td>0.9860</td>
<td>0.9867</td>
<td>0.9813</td>
<td>0.9840</td>
</tr>
<tr>
<td>90,65,30</td>
<td>10,35,70</td>
<td>0.634,0.111,0.255</td>
<td>BetaBin</td>
<td>0.9846</td>
<td>0.9803</td>
<td>0.9795</td>
<td>0.9739</td>
<td>0.9705</td>
</tr>
<tr>
<td>900,650,300</td>
<td>100,350,700</td>
<td>0.635,0.111,0.254</td>
<td>BetaBin</td>
<td>0.9789</td>
<td>0.9763</td>
<td>0.9772</td>
<td>0.9699</td>
<td>0.9561</td>
</tr>
<tr>
<td>90,65,30</td>
<td>10,35,70</td>
<td>0.631,0.119,0.250</td>
<td>Binom</td>
<td>0.9824</td>
<td>0.9785</td>
<td>0.9780</td>
<td>0.9730</td>
<td>0.9639</td>
</tr>
</tbody>
</table>

Table 3.2. Evaluating model performance at default parameters with AUC. For the Binomial distribution, in addition to assessing performance at default parameters, it is also possible to assess performance of using expectation maximisation to update the model parameters. For the Beta-Binomial distribution, only the values at $\tau=100$ are shown. The results of the other parameter strengths can be inferred from Figure 3.6.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\pi$</th>
<th>Model</th>
<th>Set1</th>
<th>Set2</th>
<th>Set3</th>
<th>Set4</th>
<th>Set5</th>
</tr>
</thead>
<tbody>
<tr>
<td>990,500,10</td>
<td>0.695,0.13,0.282</td>
<td>BetaBin</td>
<td>0.9435</td>
<td>0.9373</td>
<td>0.9408</td>
<td>0.9355</td>
<td>0.9324</td>
</tr>
<tr>
<td>990,500,10</td>
<td>0.618,0.221,0.161</td>
<td>Binom</td>
<td>0.9763</td>
<td>0.9657</td>
<td>0.9745</td>
<td>0.9730</td>
<td>0.9701</td>
</tr>
<tr>
<td>9.9,5,0.1</td>
<td>0.838,0.042,0.119</td>
<td>Binom EM</td>
<td>0.5161</td>
<td>0.5137</td>
<td>0.5032</td>
<td>0.5132</td>
<td>0.5032</td>
</tr>
<tr>
<td>99,50,1</td>
<td>0.838,0.042,0.119</td>
<td>Binom EM</td>
<td>0.7290</td>
<td>0.7390</td>
<td>0.7138</td>
<td>0.7113</td>
<td>0.8116</td>
</tr>
<tr>
<td>990,500,10</td>
<td>0.838,0.042,0.119</td>
<td>Binom EM</td>
<td>0.7296</td>
<td>0.7390</td>
<td>0.7130</td>
<td>0.7119</td>
<td>0.8126</td>
</tr>
<tr>
<td>9900,5000,100</td>
<td>0.838,0.042,0.120</td>
<td>Binom EM</td>
<td>0.7331</td>
<td>0.7457</td>
<td>0.7162</td>
<td>0.7130</td>
<td>0.8178</td>
</tr>
</tbody>
</table>

Of the 30 cases, only 25 have been used until this point to evaluate the properties of the Binomial and Beta-Binomial mixture model and suggest an optimally performing set of $\alpha$ and $\beta$ parameters. The best performing parameters are then used to predict variants in the remaining 5 validation cases left out from cross validation (Table 3.1). For the Beta-Binomial mixture model $\alpha = [9.9,6.5,2.5]$ and $\beta=[0.1,3.5,7.5]$ were used to assess model performance and $\alpha = [9.0,6.5,3.0]$ and
$\beta = [1.0, 3.5, 7.0]$ is used for the Binomial distribution. The Beta-Binomial distribution had a marginal advantage in performance against the Binomial distribution. If $\tau$ were increased to 1000 or more, the Beta-Binomial distribution would not disambiguate between the genotypes, as established in Figure 3.2A, and the Binomial mixture model would perform better. Recall that in Figure 3.3 it was demonstrated that the Beta-Binomial mixture model, specifically at lower depths, performed better than the Binomial mixture. Table 3.1 summarises a similar trend from the exon capture data. This check is necessary due to high depths and bias for the reference allele in exon capture data and it is possible that the models would have had trouble detecting variants because their representation would be skewed.

Finally, it is possible to compare the optimal parameters, determined by cross validation, against intuitively set hand parameters. Compared against the intuitively set parameter values, the optimal parameters performed better for the Beta-Binomial distribution, especially at higher $\tau$. The results are summarised in Table 3.2; from Figure 3.6 it is observed that at lower $\tau$ the performance of the optimal parameters is not significantly different then the intuitively set parameters (indicated by the arrow) at $\tau = [1, 10]$ with a slight decrease in performance at $\tau = 100$. Furthermore the performance of the Beta-Binomial continues to be better than the Binomial mixture model at lower $\tau$. The large difference in performance arises when attempting to update the Binomial mixture model with expectation maximisation. The performance of the Binomial mixture model has an AUC around approximately 0.7300 and is as low as 0.5032, which is a large decrease in performance compared to the hand set or optimal parameters. Finally, the $\mu$ values were updated to: $\mu_{aa} = 0.7445, \mu_{ab} = 0.0018$, and $\mu_{bb} = 0.0001$, which is very different from the optimal performing parameters.
Figure 3.7 Comparison of Binomial and Beta-Binomial implementations of CoNAn-SNV. A) Total number of SNVs called from each of the different implementations, shown at different copy number states. The Binomial implementation calls a total of 69,308 SNVs. This number is different than in Figure 2.2 because the model has not been trained; due to the rigidity of the Binomial mixture model, especially at great depths, slight changes in parameters can have a large effect on the number of calls made. The total number of calls made by the Beta-Binomial implementation of CoNAn-SNV from $\tau$ 1 to 100 is: 62,254, 66,776 and 68,471 respectively. All of the 68,471 calls made at $\tau$=100 are a subset of the calls made by the Binomial implementation of CoNAn-SNV. B) Overlap of the calls made by different $\tau$ in the Beta-Binomial implementation of CoNAn-SNV. The calls made by the different $\tau$ are not all subsets of each other. In NEUT/LOSS and GAIN states the lower $\tau$ models call variants that are not detected at $\tau$ 100. As the CNA level increases however, the higher $\tau$ hyperparameters yield better results.

Lastly the Beta-Binomial framework is applied to CoNAn-SNV and tested against the WGSS lobular breast carcinoma. The WGSS data, as opposed to the exon capture, is used because there exists validation data for which to assess the model performance. Furthermore, the validation data covers more diverse regions of the genome. Due to the strategic design of the chip, SNP calls made from the Affymetrix SNPChip array may come from less complex regions of the genome and additionally the CRLMM or OncoSNP calls may not be accurate. It is possible that the similarity of the Binomial and Beta-Binomial mixture model performance that was observed in the exon capture data may be attributed to the stability of the regions targeted by the SNPChip array.
An initial appraisal of the model performance shows that the Binomial implementation of CoNAAn-SNV outperforms the tested $\tau$ of the Beta-Binomial implementation (Figure 3.7A). As the $\tau$ of Beta-Binomial implementation increases, the Beta-Binomial performs more similarly to the Binomial mixture model; exceeding $\tau$ of 100 may cause the model to fail calling anything as it cannot disambiguate between the genotype calls (Figure 3.2). All of the SNV calls made by the Beta-Binomial implementation at $\tau$ 100 are a subset of the calls made by the Binomial implementation. The calls made by the different Beta-Binomial $\tau$, however, are not a perfect subset of each other (Figure 3.7B). At lower $\tau$, there are more novel calls made in copy neutral regions; however it is evident in Figure 3.7B that higher strengths perform better in regions of higher level CNAs. This is further supported by the recovery of somatic variants in the Table 3.3. At a $\tau$=1, the majority of the validated somatic variants fail to be recovered, however at $\tau$=10 and above all somatic variants are recovered with a variable value of $p$(SNV) (Table 3.3). Overall, $\tau$=10 makes fewer calls than the higher $\tau$ Beta-Binomial implementation. The Binomial implementation, however is equally successful in recovering validated somatic variants. From the ROC results in Figure 3.4 it may be possible to infer that Beta-Binomial at $\tau$=10 calls more true positives, with a lower false positive rate. This notion remains ambiguous when comparing against false positives from the validated data (Table 3.4), where the Binomial and Beta-Binomial implementation (at $\tau$ = 10 or higher) perform equally well. The Beta-Binomial at $\tau$=10 would encode a lower $p$(SNV) for the potential false positives, however ultimately there is sufficient support for the alignment to call a variant with a confidence that passes the threshold.
Table 3.3 Recovery of the 24 validated somatic mutations by the Binomial and Beta-Binomial implementations of CoNaN-SNV. The Beta-Binomial implementation of CoNaN-SNV is evaluated at different $\tau$. The shaded boxes indicate positions that fail to be called SNVs. The threshold for $p$(-SNV) is 0.77 as described in chapter 2.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Ref</th>
<th>Ref Count</th>
<th>Nref</th>
<th>Nref Count</th>
<th>CN State</th>
<th>Beta-Binomial</th>
<th>Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PS(NV) Strength 1</td>
<td>PS(NV) Strength 10</td>
</tr>
<tr>
<td>1</td>
<td>31870778</td>
<td>G</td>
<td>17</td>
<td>A</td>
<td>3</td>
<td>3</td>
<td>0.813</td>
<td>0.964</td>
</tr>
<tr>
<td>1</td>
<td>177905078</td>
<td>C</td>
<td>105</td>
<td>T</td>
<td>12</td>
<td>5</td>
<td>0.402</td>
<td>0.919</td>
</tr>
<tr>
<td>1</td>
<td>181813423</td>
<td>C</td>
<td>68</td>
<td>G</td>
<td>10</td>
<td>5</td>
<td>0.489</td>
<td>0.962</td>
</tr>
<tr>
<td>1</td>
<td>200099346</td>
<td>G</td>
<td>87</td>
<td>C</td>
<td>9</td>
<td>5</td>
<td>0.368</td>
<td>0.841</td>
</tr>
<tr>
<td>2</td>
<td>100965216</td>
<td>A</td>
<td>45</td>
<td>T</td>
<td>6</td>
<td>4</td>
<td>0.275</td>
<td>0.858</td>
</tr>
<tr>
<td>4</td>
<td>25524942</td>
<td>C</td>
<td>9</td>
<td>G</td>
<td>2</td>
<td>3</td>
<td>0.842</td>
<td>0.949</td>
</tr>
<tr>
<td>5</td>
<td>175726007</td>
<td>G</td>
<td>38</td>
<td>A</td>
<td>6</td>
<td>4</td>
<td>0.319</td>
<td>0.917</td>
</tr>
<tr>
<td>5</td>
<td>176953851</td>
<td>G</td>
<td>46</td>
<td>T</td>
<td>7</td>
<td>4</td>
<td>0.313</td>
<td>0.936</td>
</tr>
<tr>
<td>6</td>
<td>44361861</td>
<td>C</td>
<td>24</td>
<td>T</td>
<td>4</td>
<td>4</td>
<td>0.321</td>
<td>0.816</td>
</tr>
<tr>
<td>6</td>
<td>111800869</td>
<td>G</td>
<td>35</td>
<td>T</td>
<td>6</td>
<td>3</td>
<td>0.830</td>
<td>0.993</td>
</tr>
<tr>
<td>6</td>
<td>157570350</td>
<td>G</td>
<td>52</td>
<td>A</td>
<td>7</td>
<td>4</td>
<td>0.280</td>
<td>0.899</td>
</tr>
<tr>
<td>7</td>
<td>139446250</td>
<td>G</td>
<td>37</td>
<td>C</td>
<td>6</td>
<td>4</td>
<td>0.327</td>
<td>0.924</td>
</tr>
<tr>
<td>11</td>
<td>2383399</td>
<td>T</td>
<td>13</td>
<td>A</td>
<td>3</td>
<td>4</td>
<td>0.399</td>
<td>0.805</td>
</tr>
<tr>
<td>14</td>
<td>93999360</td>
<td>G</td>
<td>24</td>
<td>A</td>
<td>4</td>
<td>4</td>
<td>0.321</td>
<td>0.816</td>
</tr>
<tr>
<td>14</td>
<td>100417938</td>
<td>C</td>
<td>19</td>
<td>T</td>
<td>4</td>
<td>4</td>
<td>0.390</td>
<td>0.887</td>
</tr>
<tr>
<td>19</td>
<td>6403457</td>
<td>C</td>
<td>9</td>
<td>T</td>
<td>2</td>
<td>3</td>
<td>0.842</td>
<td>0.949</td>
</tr>
<tr>
<td>19</td>
<td>42881337</td>
<td>T</td>
<td>65</td>
<td>A</td>
<td>7</td>
<td>5</td>
<td>0.377</td>
<td>0.801</td>
</tr>
<tr>
<td>19</td>
<td>45591960</td>
<td>C</td>
<td>33</td>
<td>T</td>
<td>5</td>
<td>5</td>
<td>0.482</td>
<td>0.872</td>
</tr>
<tr>
<td>19</td>
<td>47506592</td>
<td>C</td>
<td>42</td>
<td>T</td>
<td>5</td>
<td>5</td>
<td>0.402</td>
<td>0.845</td>
</tr>
<tr>
<td>19</td>
<td>54648470</td>
<td>G</td>
<td>35</td>
<td>T</td>
<td>5</td>
<td>5</td>
<td>0.462</td>
<td>0.984</td>
</tr>
<tr>
<td>19</td>
<td>57509248</td>
<td>G</td>
<td>51</td>
<td>C</td>
<td>9</td>
<td>5</td>
<td>0.560</td>
<td>0.886</td>
</tr>
<tr>
<td>19</td>
<td>59779115</td>
<td>G</td>
<td>42</td>
<td>A</td>
<td>6</td>
<td>5</td>
<td>0.467</td>
<td>0.808</td>
</tr>
<tr>
<td>19</td>
<td>61757367</td>
<td>G</td>
<td>64</td>
<td>C</td>
<td>7</td>
<td>5</td>
<td>0.382</td>
<td>0.795</td>
</tr>
<tr>
<td>19</td>
<td>63542071</td>
<td>G</td>
<td>66</td>
<td>A</td>
<td>7</td>
<td>5</td>
<td>0.373</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Table 3.4 Total number of calls made by each model of all validated data. The Binomial and Beta-Binomial implementation (at $\tau$=10 or higher) perform equally calling the same number of true positives and false positives at a threshold of $p$(SNV) >0.77.

<table>
<thead>
<tr>
<th></th>
<th>Somatic (/24)</th>
<th>Germline (/17)</th>
<th>False Pos. (/43)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>24</td>
<td>17</td>
<td>43</td>
</tr>
<tr>
<td>BB Strength 1</td>
<td>4</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>BB Strength 10</td>
<td>24</td>
<td>17</td>
<td>42</td>
</tr>
<tr>
<td>BB Strength 100</td>
<td>24</td>
<td>17</td>
<td>43</td>
</tr>
</tbody>
</table>
3.4 Discussion

This chapter explores using a Beta-Binomial mixture model framework for SNV discovery over the currently used SNVMix framework. The conventional Binomial mixture model framework is employed in genome variant discovery because it is highly applicable and it is intuitive to set the values of $\mu_k$ due to the biological distribution of alleles in diploid genomes. Explicitly setting the value of $\mu_k$ is unnecessary and it is possible to avoid its calculation by marginalizing it out and resolving a weighted average of the predictions.

Initial exploration with handset $\alpha$ and $\beta$ parameters show the Beta-Binomial mixture model is sensitive to the strength of the parameters ($\tau$). At high $\tau$ and greater depths the Beta-Binomial mixture model is incapable of distinguishing between the three diploid genotypes (shown in Figure 3.2A). This is because the relationship of the $\alpha$ and $\beta$ parameters with $N$; where if $\alpha$ and $\beta \gg N$, the Beta-Binomial’s equation would be decomposed to:

$$\text{BetaBinom}(a|\alpha, \beta, N) = \frac{\text{Beta}(\alpha, \beta)}{\text{Beta}(\alpha, \beta)} \left( \frac{N}{a} \right) \text{ if } \alpha, \beta \gg N$$

making the likelihood be the same for all genotypes, given $N$ and $a$. As the depths grow larger, but are still insignificant compared to $\alpha$ and $\beta$, $\left( \frac{N}{a} \right)$ has a larger value and $\pi$ has a decreased effect compared to lower depths. Eventually, the posterior probability for all the genomes will resolve to be the same (due to the limits in machine precision) making it impossible to distinguish between the genotypes. The Binomial mixture model framework has steep transitions between the genotypes, however is capable of distinguishing between them regardless of the depth. As $\tau$ is decreased the Beta-Binomial framework has broader genotype transitions compared with the Binomial framework. Setting asymmetric values for $\alpha$ and $\beta$ shifts the distribution of posterior probabilities toward the strengthened parameters (in this case for $G = aa$).

The broadness of the transitions between genotypes encodes for ambiguity. The Binomial framework has steep transitions, and a difference of 1 or 2 more reads supporting the variant is sufficient to call the variant with a high probability, with the converse being true if 1 or 2 fewer reads supported the variant. This is an unfavourable quality because false positives that barely have sufficient reads to be called a variant, by the model, may be called with a high probability, and vice
versa. Conversely, this blunt approach can also call more true positives by chance. With the Beta-Binomial, the same position with barely enough reads to be called a variant will be encoded with a lower posterior probability and will likely be removed when subsequent thresholds are applied to the data. If the goal is to call the most true positives, the blunt approach of the Binomial mixture model may be favourable. With the WGSS data, the Beta-Binomial consistently calls more true positives, without introducing more false positives. In exon capture data, due to alignment to the targeted genome space, has high depths and due to the hybridisation step tends to be biased towards the reference allele. This may an unfavourable environment to apply the Beta-Binomial mixture model because the distribution of the allelic counts will be skewed such that a variant is disproportionately represented compared to the reference. Although the Binomial mixture model will also have difficulties, the steep transitions are actually beneficial here and may lead to more variants being called.

Contrary to this logic, the Beta-Binomial mixture model out performs the Binomial mixture model on exon capture data, when \( \tau \) are either 1 or 10, evaluated both by AUC and f-measure. As the parameters strengths increase up to 1000, the Beta-Binomial cannot distinguish between either of the diploid genotypes. Exploring a defined space of \( \alpha \) and \( \beta \) parameters can approximate the parameter setting that will yield an optimal performance. Even at parameters settings where the Binomial distribution performs optimally, the Beta-Binomial still performs better. Using the default and intuitively assigned parameters yields a good performance, although missing true positives. The optimal performing parameter sets are those with the values for where the parameters defining the \( ab \) genotype are shifted away from 0.5 and are resting at \( \alpha=6.5 \) and \( \beta=3.5 \). This is because of the skew in the exon capture data; the default parameters do not capture this skew adequately and as a result will miss some true positives. Starting at default parameters, it was evaluated if the EM updating could approximate the optimal parameters indentified by exploring the defined space. At low strengths, the performance of the Binomial mixture model, with EM updating, is poor and although the performance increases with higher \( \alpha \) and \( \beta \) values it is still lower than the optimal performing parameters. Furthermore, the parameters at EM updates to are different then the optimal performing parameters. The \( \mu_{aa} \) is only 0.75, considerably lower than 0.90 for the optimal performing parameters. In this scenario, the skew induced by a bias for the reference allele affects the
ability to distinguish between the $aa$ and $ab$ genotypes. This is further emphasised by the value of
\[ \mu_{ab}, \] which is only 0.0014 due the incorporation of skewed genotypes into the updating of $\mu_{aa}$.

The defined space did not encounter such difficulties because the possible combinations of
parameters were constrained to a biologically interpretable space. The same could be done for the
EM updating using stronger priors to constrain the range of the parameters.

Applying the Beta-Binomial framework to CoNAn-SNV yields results that were different from
what was previously observed using only the diploid genotypes. First, the binomial CoNAn-SNV
implementation calls more variants overall (Figure 3.7), and calls all of the somatic variants with
high probability (Table 3.3). It may be inferred, using results from Figure 3.4, that the Binomial
implementation calls more true positives with the consequence of also calling more false positives.
The Beta-Binomial implementation of CoNAn-SNV at $\tau=10$ is interesting because it calls all the
true positives, with lower probability compared to the Binomial implementation, while making
fewer overall calls. Again, referring to Figure 3.4, the fewer calls may be at the sacrifice of false
positives while retaining true positives; this is a favourable result. Comparing against the false
positives from the validated data however shows little difference in the performance (Table 3.4).
Furthermore, it is evident in Figure 3.7B that in higher CNA states, the higher strength Beta-
Binomial implementation (which is a complete subset of the Binomial calls) makes significantly
more calls overall. It maybe all those additional calls are false positives; a larger ground truth data set
that covers diverse genomic regions is required to make this assessment.

**Limitations**

A primary limiting factor in this analysis is that $a$ and $\beta$ are not updated through a training
regimen. The performance of the Beta-Binomial mixture model is relatively similar to the Binomial
mixture model, however a larger difference is exposed when updating with expectation
maximisation. It is necessary to further evaluate whether the Beta-Binomial is more or less robust
with a defined training regimen. It is relevant to the question if the training regimen for the Beta-
Binomial is even necessary or appropriate, given that the Beta distribution should be agnostic to the
data in order to guard against overfitting. It is important to contrast these scenarios.
An additional limitation is the exon capture data itself and the available positions for evaluating the models. The complimentary SNPChip positions, that are used to generate a ground truth data set, may come from highly stable genomic regions. It is more challenging to find an objective method of evaluating model performance outside of these regions without experimental validation. Until a sufficiently rigours validation method is determined for these regions, it is also challenging to accurately gauge the benefits of using a Beta-Binomial mixture model for CoNAn-SNV.

3.5 Conclusions

The performance of the Beta-Binomial distribution on multiple data sets supports substituting the Binomial mixture model with a Beta-Binomial mixture model. This should be done with caution, as sequence data can have variable depth that makes selecting the appropriate $\alpha$ and $\beta$ challenging. It is also necessary to define an updating regimen of the Beta-Binomial mixture model to fully judge performance against the Binomial mixture model.
Chapter 4

Conclusions

The advent of next generation sequencing has ushered in era where it is possible to enumerate the complete mutational landscape of a patient’s tumour. The corollary to this advance is that the data produces millions of short reads that must be organised and interrogated in order to explore this mutational landscape. The final representation of this landscape, thus, is a projection of what can be reconstructed using bioinformatic alignment algorithms and variant discovery tools. In this work we investigate how different strategies of variant discovery affect the ability to provide a more complete catalogue of the cancer mutational landscape.

4.1 Summary of Contributions

In the second chapter of this work we introduce the CoNAn-SNV discovery tool. CoNAn-SNV extends the diploid framework that most NGS variant discovery tools employ. The extension of genotypes state space in regions of copy number amplification is topic that was explored from algorithms that are developed for SNPChip platforms. We show that it is possible to apply the allele-specific copy number genotypes to data obtained through next generation sequencing. We show that explicit incorporation of CNA status for SNV discovery, as is done with CoNAn-SNV, improves the yield of true positives by validating an additional 24 novel somatic variants in lobular breast carcinoma. Furthermore, we are able to detect germline variants which have a low frequency of the reference allele and are proximal to the discovered somatic variants. With this information we
are able to identify an allele specific amplification in the genome, particularly in chromosome 1 and 19. Orthogonal analysis confirms using OncoSNP confirms the events. Furthermore variants, germline or somatic, which have the non-reference allele amplified would be classified as \textit{bb} using diploid genotypes. This may be considered a loss of information, as allele-specific copy number genotypes such as \textit{abb, abbb} or \textit{abbb} genotype also gives contextual information that may be used to understand tumour evolution or aetiology. Finally, allele-specific amplification can affect frequency of the variant allele. Low frequency variants have been interpreted by Shah \textit{et al.} (2009) and Ding \textit{et al.} (2010) to represent intra-tumoural heterogeneity. However, we demonstrate the knowledge of CNA status and support from germline variants can distinguish an allele-specific amplification event from intra-tumoural heterogeneity.

In chapter three we explore substituting the Binomial mixture model for a Beta-Binomial mixture model. An advantage of the Beta-Binomial mixture model is that it is not necessary to explicitly define the parameter \( \mu_k \) and instead we use a weighted average of the predictions of \( \mu_k \) over the interval [0,1] by marginalising out \( \mu_k \). The advantage of the Beta-Binomial over the Binomial distribution is that it is capable of encoding variance when calling SNVs and is robust to outliers, allowing for clearer genotype segregation especially when depths are higher (Section 3.1.1). We compared the Binomial and Beta-Binomial distributions using both real and simulated data. Simulated data showed that the Beta-Binomial is sensitive to the strength of the parameters selected and we must be careful to select parameters for \( \alpha \) and \( \beta \) that are not significantly greater than the depth, \( N \). With sequence data, where \( N \) can vary significantly across the entire data set, it can be difficult to select the appropriate parameters. The exon capture data set shows itself to a challenging environment to train the Binomial mixture model, due to inherit bias for the reference allele in the data. The trained parameters for the Binomial distribution perform poorly, and the \textit{aa} and \textit{ab} genotypes become merged due to biased representation of the reference allele observed in this data type. It is necessary to define an appropriate updating regimen for the Beta-Binomial distribution to evaluate if the Beta-Binomial distribution is more robust to updating than the Binomial distribution. Application of the Beta-Binomial framework to CoNAn-SNV yielded encouraging results, suggesting that at a modest strength the Beta-Binomial can call as many true positives as the
Binomial implementation of CoNAn-SNV without calling more false positives. It is difficult to fully assess the validity of that notion, and more complex validation methods are required to support it.

Variant discovery tools will only be as effective as the initial alignment to a reference genome. As the alignment algorithms improve it will be possible to compile a variant list that is more complete. These algorithms, however, face the challenge of resolving ambiguous reads derived from repetitive regions and homologous regions. Targeted technologies, such as exon capture, still introduce considerable noise in the analysis and can produce ambiguous alignments. In order to better resolve the cancer mutational landscape is it therefore necessary to advance bioinformatic and alignment tools synergistically with technology and protocol to improve the quality and consistency of the data.

4.2 Future Work

The areas for future work can fall into two separate categories: refinements to the probabilistic graphical model itself and better interpretation of the biological events the model discovers. First the Binomial Mixture model has a number of limitations that are partly address by using the Beta-Binomial Distributions. As previously stated, a training regimen for the Beta-Binomial distribution should be devised, however it is important to question how training the Beta-Binomial framework improves results or negatively affects them. Additionally refinements of the model could include borrowing strength across the positions to infer the genotype. The advantage would be a clearer definition of allele-specific genotypes, however there may be a loss of information resulting from smoothing over somatic variants (i.e. bbb, bbbbbb ect.) in regions of allele-specific amplification. A final consideration is the using a binary variable to represent sequence data maybe overly simplified and it may be necessary to consider the Multinomial distribution, which is capable of modelling the four nucleotide base pairs, however there are challenges associated with this: the state-space may become unmanageable.

With respect to the biological properties of the data, one important area would be to understand the effect of an allele-specific amplifications on gene expression, where one could evaluate if there is enriched expression of the somatic variant. In this study (chapter 2) we made some observations regarding the expression of the somatic variants, however it is not possible to make full conclusions without obtaining a normal sample; this is not a simple task. Furthermore, understanding how the
somatic variants are distributed among the sister chromatids may have important implications in understanding the effects of CNAs. If, for example, the somatic variants described in chapter 2 occur on the amplified arm of the chromosome, then it is possible that redundancy makes CNA segments more mutable. It may be possible to make this inference by using longer reads when validating to capture germline variants that are in the vicinity of the somatic SNVs. If the somatic SNV appears only in reads harbouring the amplified germline variant it would be possible to highlight specific CNA events for further inquiry.

4.3 Implications for Cancer Treatment

Cancer is a disease whose genetic landscape influences its aetiology, aggressiveness and capacity to respond to therapeutics. It is important to develop technologies and algorithms that are able to extract these aberrant genomic variations. This manuscript shows that naive applications of diploid paradigms to cancer genomes are insufficient in characterising the whole of the cancer mutational landscape. Next generation sequencing technologies are novel enough that a clear standard for interrogating this type of data has not yet emerged. The application of different alignment tools and variant calling tools, each with algorithmic and statistical frameworks, will yield different results. Still, analysis of data sets with all available methods is not a cogent approach to capture the most variants. Regardless, early discoveries using next generation sequencing, with contemporary bioinformatic tools, have yielded important implications for the treatment of granulosa cell ovarian cancers (Shah(b), 2009) and lymphomas (Morin, 2010) by identifying initiating mutations for these cancers. Results from this work reveals even more novel targets and provides contextual information for them in ways that was previously unavailable. As sequencing technology matures and becomes more reliable, algorithms that are capable of enumerating the complete cancer mutational landscape will have increasingly critical roles. In this work we presented a methodology that advances this goal by demonstrating how modelling allele-specific copy number genotypes resolves additional novel somatic mutations and contextualizes SNVs in order to better understanding the cancer mutational landscape.
Bibliography


Appendix A

Exon Capture Processing

Conventionally, data obtained through the exon capture platform are aligned to the entire human genome. In this appendix, however, we describe the alignment of the exon capture data to a compacted, targeted genome that has shown to improve the quality of the results and overall SNV calling. We compare results from the lobular carcinoma obtained through whole genome shot-gun sequencing (described in chapter 2) against exon capture data derived from the same tumour DNA. We resolve, based on these findings described herein, that alignment to the targeted reference genome improves sensitivity of calls and depth as well as efficiently reduces alignment time.

Exon capture is detailed in section 3.1.2; it operates by pulling down specific exons by hybridising genomic fragments to an oligonucleotide array. In an idealised scenario exon capture will pull down only targeted sequences. Thus it is not necessary to align to the entire human genome since sequenced regions would be obtained only from a subset of the genome. It should be possible to align only to the genome space that is targeted by the platform. In reality the data are noisy. Non-target data can be introduced into the analysis in two ways: the first is through cross-hybridization, where the annealed fragment only partially matches the target probe; the second is competitive-hybridization where highly similar genomic regions will hybridize to the same probe. Data sequenced due to the first scenario will align somewhere in the whole genome outside the targeted regions. Data that are hybridized because they are highly similar to the targeted region, often different by only a few bases, will introduce ambiguity because these reads and the desired target reads can be placed by the aligner in multiple regions of the whole genome. Some of this ambiguity may be removed when the alignment space is restricted to only the positions meant to be targeted by exon capture; however such sequences will also introduce artefacts due to pseudogenes, which would have been ambiguously mapped in the whole genome but may be forced to map when using constricted alignment space. Conversely, this reduction in ambiguity when aligning to the targeted genome space, despite the introduction of artefacts, may also resolve more true positives. What this appendix will demonstrate is that aligning to the targeted space does indeed resolve more true positives at higher depths, compared to the aligning to the whole genome.
The initial appraisal of the exon capture data compared SNV calls made by MAQ, SNVMix and CoNAn-SNV. We find that there is some concordance between calls, but it is less compared to the WGSS data. When we profile the depths of the variants recovered by these three SNV discovery methods, it is possible to see that the majority of SNVs called have depths lower than 10, even with three lanes of sequence data. We were able to identify that the positions with poor coverage resulted from ambiguous alignment. To resolve this issue we created a custom reference genome which consisted of only the targeted genome sequences. We add flanking base pairs on either side to the target regions to account for the fragment overhanging on the probes. Although within the targeted space there are highly similar genomic regions, in general the majority of the ambiguity we observed was introduced by genomically similar sequence in the target space and a non-target region in the genome. It can be reasoned that, although it is possible to capture noise in the exon capture experiment, it is possible that there may still exist enrichment for the desired captured regions. As such there may be an increase in sensitivity and depth by reduction of ambiguous alignments.

We investigate the issues further using only SNVMix1 to calls SNVs in the different alignments. The parameters supplied to SNVMix play an important role in variant discovery as well, so we additionally consider how the selection of the model parameters affects variants discovery. In total, we use three different sets of model parameters: the first is derived from the WGSS lobular breast carcinoma; the second from the WTSS lobular breast carcinoma; and the third is from the exon capture data set. The parameters are fit to the WGSS and WTSS data as described in chapter 2 methods and in Goya et al. (2010). There is, however, insufficient overlap between the Affymetrix array positions and exon capture positions to properly train the model. Instead, we use dbSNP positions that overlap with the exon capture targeted regions. Although the patient will not necessarily have SNPs at all those positions, there is an increased probability of SNP representation compared to randomly sampling.

Table A1.1 summarises the results of aligning the different alignment regimens. Alignments to the entire human genome are abbreviated as HA and alignments to the targeted genome space are abbreviated TS. Furthermore, we can call variants from either all known exons (only possible with HA alignment) or only those meant to be targeted by exon capture. For HA alignments it may be worthwhile considering if true variants are captured outside of the target regions.
Table A1.1 Comparison of different alignment regimens

<table>
<thead>
<tr>
<th>Model</th>
<th>Alignment</th>
<th>Pileup Positions</th>
<th>Raw Calls</th>
<th>Validated Calls</th>
<th>Known</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGSS</td>
<td>TS</td>
<td>TS</td>
<td>55,797</td>
<td>19</td>
<td>23,953</td>
<td>31,844</td>
</tr>
<tr>
<td>RNASeq</td>
<td>TS</td>
<td>TS</td>
<td>48,125</td>
<td>16</td>
<td>23,066</td>
<td>25,059</td>
</tr>
<tr>
<td>DBSNP</td>
<td>TS</td>
<td>TS</td>
<td>103,173</td>
<td>34</td>
<td>27,611</td>
<td>75,562</td>
</tr>
<tr>
<td>WGSS</td>
<td>HA</td>
<td>TS</td>
<td>26,818</td>
<td>20</td>
<td>16,904</td>
<td>5,514</td>
</tr>
<tr>
<td>RNASeq</td>
<td>HA</td>
<td>TS</td>
<td>22,418</td>
<td>18</td>
<td>20,634</td>
<td>1,784</td>
</tr>
<tr>
<td>DBSNP</td>
<td>HA</td>
<td>TS</td>
<td>46,073</td>
<td>36</td>
<td>23,675</td>
<td>22,398</td>
</tr>
<tr>
<td>WGSS</td>
<td>HA</td>
<td>All Exonic</td>
<td>33,247</td>
<td>-</td>
<td>23,721</td>
<td>9,526</td>
</tr>
<tr>
<td>RNASeq</td>
<td>HA</td>
<td>All Exonic</td>
<td>24,720</td>
<td>-</td>
<td>22,441</td>
<td>2,279</td>
</tr>
<tr>
<td>DBSNP</td>
<td>HA</td>
<td>All Exonic</td>
<td>56,758</td>
<td>-</td>
<td>26,731</td>
<td>30,027</td>
</tr>
</tbody>
</table>

We evaluate the goodness of the model parameters and alignment based on the recovery of the validated somatic variants from Shah et al. (2009) and those made by CoNAn-SNV; a total of 52 validated somatic variants. Additionally, we evaluate the total number of dbSNP variants are also found.

Based on the observations drawn from Table A1.1 we draw several conclusions: if the criteria is to resolve the most possible somatic variants, aligning to the targeted genome space is a better choice. Not only does the data have more depth, but it also calls more dbSNP positions compared to the human all alignment. This alignment finds the most novel variants, although some of those will be true somatic variants, many may be artefacts introduced by the alignment. Aligning to the entire genome will recover the same total number of validated variants, however these variants come from conserved regions so there is little difference in their representation in the HA and TS alignments. The difference is that HA will call fewer dbSNP variants. This is worthwhile if the goal is to reduce the total number of false positives, however using the exon capture derived model parameters the HA alignment will still introduce a large number false positives into the analysis. To make the potentially large space of false positives more manageable, use the WGSS or WTSS derived parameters, however at the cost recovering only half of the validated somatic variants. Although not extensively studied in this work, it may be possible to use a paired normal genome of the tumour to filter out false positives. Presumably the normal genome will be subject to the introduction of the same artefacts from highly similar regions of the genome that are forced to align.
to the target space. A joint model capable of simultaneously considering support of the variant in both normal and tumour sequence data to evaluate somatic variant could fulfil such a purpose.
Appendix B

SeqVal01 (INX021): Validation of 200 SNVs in lobular breast cancer

Genomic DNA was prepared as described (Shah, 2009). Automated primer design was performed using Primer3 (Rozen and Skaletsky, 1998) and custom scripting. Primer pairs were designed to place the variant position within 75bps of either end of the amplicon and to be between 50-300bp in length. A total of 202 unique variant positions were targeted of which 192 yielded primer pairs that met the specified criteria. Primer pairs were independently validated by in silico PCR followed by BLAT against the human genome to ensure that the correct target was generated and that resulting amplicon was unique within the genome. DNA primers were synthesized in 96-well plates at a 25nmol scale with standard desalting (IDT Coralville, IA USA). Polymerase cycling reactions were set up in 96-well plates and comprised of 0.5 µM forward primer, 0.5 µM reverse primer, 1-2 ng of gDNA template, 5X Phusion HF Buffer, 0.2 µM dNTPs, 3% DMSO, and 0.4 units of Phusion DNA polymerase (NEB, Ipswich, MA, USA). Reaction plates were cycled on a MJR Peltier Thermocycler (model PTC-225) with cycling conditions of a denaturation step at 98 ºC for 30 sec, followed by 35 cycles of [98ºC for 10 sec, 69ºC for 15 sec, 72ºC for 15 sec] and a final extension step at 72ºC for 10 min. PCR reactions were visualized on 3% agarose (NuSieve) gels for 2hrs at 170V to assess PCR success. Successful reactions were manually pooled (4ul per well) by template and subjected to Illumina library construction using a modified paired-end protocol (Illumina, Hayward, USA). This involved A-tailing of the amplicons and ligation to Illumina PE adapters. Adapter-ligated products were purified on Qiaquick spin columns (Qiagen, Valencia, CA, USA) and PCR-amplified using Phusion DNA polymerase (NEB, Ipswich, MA, USA) in 10 cycles using PE primer 1.0 (Illumina) and a custom multiplexing PCR Primer [5’-CAAGCAGAAGACGGCATACGAGATBNNNNNNCGGTCTCGGATATTCCCTGC TGAACCGCCTTCCGATCT-3’] where “NNNNNN” was replaced with the following unique fault tolerant hexamer barcodes: pleural effusion metastatic tumor, “CGTGAT”; buffy coat, “ACATCG”; and primary tumor “GCCTAA”. PCR products of the desired size range were purified away from adapter ligation artifacts using 8% PAGE gels, pooled and DNA quality was
assessed and quantified using an Agilent DNA 1000 series II assay (Agilent, Santa Clara CA, USA) and Nanodrop 7500 spectrophotometer (Nanodrop, Wilmington, DE, USA) and subsequently diluted to 10nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen, Carlsbad, CA, USA). For sequencing, clusters were generated on the Illumina cluster station using v4 cluster reagents and paired-end 75bp reads generated using v4 sequencing reagents on the Illumina GAii platform following the manufacturer’s instructions. Between the paired 75bp reads a third 7 base pair read was performed using the following custom sequencing primer [5- GATCGGAAGAGCGGTTCAGAGGAATGCGAGACCG] to sequence the hexamer barcode. Image analysis, base-calling and error calibration was performed using v1.60 of Illumina’s Genome analysis pipeline.
Appendix C

Description of additional files

All files are on accompanied CD. The additional files we include more a FASTA data file of the reference genome used in the validation (described in Methods). Figures and tables that support the manuscript are also provided:

Table S1. CNA segment input to CoNAn-SNV. CoNAn-SNV takes as input CNA segments in addition to allelic counts. A line of input indicates a chromosome number, segment start and end site, and lastly a numerical encoding of the CNA state. The numbers are: 2(NEUT/LOSS); 3 (GAIN); 4(AMP); and 5(HLAMP). The model can receive input from any segmentation algorithm so long as it is provided in the same format as this table. Additionally, CoNAn-SNV is not constrained to the state-space used in this paper, and is flexible to other levels of amplification so long as they can be encoded numerically. There are, however, important considerations that should be made if choosing to extend the state-space beyond what has been described in this manuscript. Further instruction for using the model is available on the download page.

Table S2. Summary of the 200 positions submitted for validation

Tables S3. Genomic Positions with skewed allelic genotypes. This table indicates positions in the entire genome that harbour the extreme allelic skews such as aaaa and aaaaa.

Table S4. OncoSNP CNA segment predictions

Tables S5. Breakdown of CoNAn-SNV and SNVMix1 calls from OncoSNP genotyped training data.

Tables S6. Primer specifications for the 200 candidate validation positions.

Figure S1a. Copy number annotations for all chromosomes made by CNAnseq HMM.

Figure S1b. Copy number annotations for all chromosome made by OncoSNP