METHODS FOR TRANSCRIPT VARIANT DISCOVERY AND ALTERNATIVE EXPRESSION ANALYSIS – APPLICATION TO THE STUDY OF FLUOROURACIL RESISTANCE IN COLORECTAL CANCER

by

MALACHI GRIFFITH

B.Sc., The University of Winnipeg, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2010

© Malachi Griffith, 2010

Abstract

RNA transcripts are expressed from tens of thousands of loci across the human genome. Several studies have suggested that many genes are alternatively expressed to produced multiple mRNA isoforms and many of these remain undiscovered. Identifying specific isoforms associated with human diseases such as cancer has potential to lead to improved treatments. The scale and complexity of the transcriptome present significant barriers to (1) identifying isoforms and (2) applying knowledge to human disease research. Recent advances in genome-wide microarray and sequencing platforms have begun to provide the capacity and resolution to address these challenges. The goal of this thesis was to develop novel methods that allow genome-wide identification and quantification of mRNA isoforms. I first approached this problem by creating a microarray design platform for alternative expression analysis called 'ALEXA-array' (www.AlexaPlatform.org). To evaluate the ALEXA-array approach I used it to generate a microarray design that I then used to measure differential expression of mRNA isoforms in 5-fluorouracil (5-FU) sensitive and resistant colorectal cancer cell lines. This approach identified several isoforms potentially involved in 5-FU resistance. While the ALEXA-array approach was successful, I identified several limitations of the method. For example, the approach was insensitive to isoforms with small differences in sequence content and limited by both the transcriptome annotations and the number of microarray features available at design time. I developed a second method, 'ALEXA-seq', to take advantage of advances in massively parallel sequencing. Applying this method to the same cell lines I showed that the approach was able to overcome many limitations of the microarray approach. Several additional candidate 5-FU resistance isoforms were identified. Both the ALEXA-array and ALEXA-seq approaches identified expression of an aberrant isoform of the *uridine monophosphate synthetase* as a top candidate. Interestingly, this gene was suspected to function in the conversion of 5-FU to active anti-cancer metabolites. Additional characterization was performed to elucidate the expression pattern, transcript diversity and sequence variation of this gene in a panel of cell lines and tumours. The methods presented here should help to identify mRNA isoforms with potential utility as therapeutic targets or as prognostic or diagnostic markers.

Table of contents

Abstract ii
Table of contents iii
List of tables vi
List of figures vii
List of equations ix
Acknowledgements x
Co-authorship statement xi
1. Alternative expression analysis: experimental and bioinformatic approaches for the analysis of transcript diversity 1
1.1. Introduction 1
1.2. Thesis overview 1
1.3. Gene expression, alternative expression, and its regulation 2
1.4. Genomic approaches for the study of transcript diversity 5
1.4.1. In silico methods 6
1.4.2. Microarray methods 8
1.4.2.1. 'First generation' expression arrays 9
1.4.2.2. Whole genome and exon tiling arrays 9
1.4.2.3. Splicing arrays 11
1.4.3. Library construction and sequencing methods 15
1.4.3.1. EST sequencing of cDNA libraries 15
1.4.3.2. Full-length sequencing of cDNA libraries 17
1.4.3.3. Sequence-tag based methods 21
1.4.3.4. Massively parallel sequencing methods 23
1.4.4. Limitations of transcriptome analysis methods 26
1.4.4.1. Limitations of microarray approaches 26
1.4.4.2. Limitations of sequencing approaches 27
1.4.5. Functional characterization of mRNA isoforms 28
1.5. Functional significance of alternative expression 30
1.5.1. How much alternative expression is functional? 30
1.5.2. How does alternative expression influence the proteome? 32
1.5.3. Implications of alternative expression for the study of disease 34
1.6. Cancer 38
1.6.1. Colorectal cancer 39
1.6.2. Chemotherapy resistance 40
1.7. Thesis objectives and chapter summaries 41
References 54
2. ALEXA: A microarray design platform for alternative expression analysis 67
2.1. Introduction 67
2.2. Results 69
2.2.1. Pre-computed microarray designs 70
2.2.2. Validation - cross platform analysis 71
2.2.3. Differentially expressed genes and mRNA isoforms associated with 5-FU resistance 73
2.3. Discussion 75
2.4. Methods 78
2.4.1. Probe extraction and filtering for array designs 78
2.4.2. Creation of a validation array design 79

2.4.3. Tissue culture 80
2.4.4. RNA isolation, labeling and hybridizations 80
2.4.5. Data processing 81
2.4.6. Platform comparisons 81
2.4.7. Visualization 82
2.4.8. Identification of significant differential expression events 82
2.4.9. Gene ontology analysis 83
2.4.10. Identification of putative alternative expression events 83
2.4.11. Statistical analysis 84
References 104
3. Alternative expression analysis by RNA sequencing 106
3.1. Introduction 106
3.2. Results 108
3.2.1. Whole transcriptome shotgun sequencing (WTSS) 108
3.2.2. Annotation of features and read mapping 109
3.2.3. Comparison of Illumina WTSS expression data to Affymetrix and ALEXA microarray expression data 110
3.2.4. Expression of canonical and alternative sequence features 112
3.2.5. Differential expression analysis 115
3.2.6. Alternative expression analysis 116
3.2.7. Global disruption of splicing 119
3.2.8. Aberrant expression of candidate 5-FU resistance genes 120
3.3. Discussion 121
3.4. Methods 124
3.4.1. Tissue culture and RNA preparation 124
3.4.2. Illumina library construction and sequencing 124
3.4.3. Data pre-processing 125
3.4.4. Source of gene models 125
3.4.5. Creation and annotation of an alternative expression database 126
3.4.6. Alignment strategy and assignment of reads to features 128
3.4.7. Cross-platform comparison of expression and differential expression estimates 130
3.4.8. Calculation of feature expression values 130
3.4.9. Library depth and feature discovery 132
3.4.10. Determining expression above background 132
3.4.11. Estimating the total copy number of genes expressed in a cell 134
3.4.12. Differential expression analysis 134
3.4.13. Alternative expression analysis 135
3.4.14. Pathway analysis 136
3.4.15. Software implementation and availability 136
3.4.16. Statistics and data visualization 137
References 176
4. Genomic analysis of uridine monophosphate synthetase reveals novel mRNA isoforms and mutations associated with fluorouracil resistance in colorectal cancer 179
4.1. Introduction 179
4.2. Results 181
4.2.1. Differential expression analysis of *UMPS* isoforms in 5-FU sensitive and resistant cell lines. 181
4.2.2. Characterization of *UMPS* transcript structural diversity 182

4.2.3. *UMPS* protein expression 184
4.2.4. Survey of *UMPS* isoform expression in treatment naïve colorectal tumor samples 184
4.2.5. Sequencing of *UMPS* in 5-FU sensitive and resistant cell lines 185
4.2.6. Sequencing of *UMPS* in colorectal cancer samples 187
4.3. Discussion 189
4.4. Methods 192
4.4.1. Cell lines 192
4.4.2. Clinical samples 192
4.4.3. RNA Isolation 193
4.4.4. Genomic DNA isolation 193
4.4.5. Splicing microarray analysis 194
4.4.6. Whole transcriptome shotgun sequencing and analysis 194
4.4.7. RT-PCR and semi-quantitative RT-PCR validation of *UMPS* isoform expression 194
4.4.8. Quantitative real time RT-PCR 194
4.4.9. Cloning & sequence validation of *UMPS* mRNA isoforms 195
4.4.10. Splice site analysis 195
4.4.11. Western analysis 196
4.4.12. PCR and sequencing the *UMPS* locus 197
References 213
5. Conclusions 216
5.1. Summary 216
5.2. Strengths and limitations 217
5.3. Current status, significance and contribution to field of study 219
5.4. Potential applications and future directions 221
References 225
Appendices 227
Appendix A . Description of 5-FU and related drugs (analogs, pro-drugs, 5-FU combination therapies, etc.) 227
References 230
Appendix B . Primer sequences used for *UMPS* analysis 231
Appendix C . Ethics approval certificates 233
Appendix C1. Ethics certificate for samples obtained from the Ontario Institute for Cancer Research (Ontario Tumour Bank) 233
Appendix C2. Ethics certificate for samples obtained from the British Columbia Cancer Agency 234
Appendix C3. Ethics certificate for samples obtained from St. Paul's Hospital 235

List of tables

Table 1.1. Summary of methods for studying transcript diversity 50
Table 1.2. Alternative expression resources 52
Table 2.1. Summary of pre-computed ALEXA designs. 98
Table 2.2. Within platform reproducibility for biological replicates 99
Table 2.3. Summary of differential expression events for genes profiled by the Affymetrix and ALEXA array platforms 99
Table 2.4. Candidate differential gene expression events associated with 5-FU resistance 100
Table 2.5. Candidate differential isoform expression events associated with 5-FU resistance 102
Table 3.1. Summary of alternative expression annotation databases for seven species 164
Table 3.2. Summary of read data, gene model sources and features defined for alternative expression analysis 165
Table 3.3. Top 20 differentially expressed (DE) genes from three gene expression platforms 166
Table 3.4. Comparison of dynamic range, signal-to-noise, sensitivity and specificity for Affymetrix, NimbleGen and Illumina platforms based on an analysis of expression estimates for the exons and introns of 100 housekeeping genes 167
Table 3.5. Comparison of *UMPS* A/B isoform expression ratios from four different platforms capable of measuring alternative isoforms 168
Table 3.6. Summary of feature expression, differential expression, and alternative expression 169
Table 3.7. Summary of novel expressed exon-exon junctions and alternative exon boundaries 170
Table 3.8. Top 50 differential or alternative expression events 171
Table 3.9. Statistically enriched functional categories identified by pathway analysis 175
Table 4.1. Quantification of *UMPS* isoform A and B and the A/B ratio determined using four gene expression platforms 208
Table 4.2. Differential expression values for *UMPS* isoform A and B 209
Table 4.3. Summary of alternative isoforms observed as clones 210
Table 4.4. Summary of putative mutations 212

List of figures

Figure 1.1. Gene expression (transcription and RNA processing) 45
Figure 1.2. Types of alternative expression (AE) 46
Figure 1.3. Splicing acceptor, donor and branch point sequences 47
Figure 1.4. Identification of an alternative exon with application to cancer medicine 47
Figure 1.5. Microarray based method for profiling transcript diversity 48
Figure 1.6. Sequence-based methods for profiling transcript diversity 49
Figure 2.1. Types of alternative expression and corresponding microarray probe design strategies 85
Figure 2.2. ROC curves for ALEXA and Affymetrix control probes 86
Figure 2.3. Correlation of ALEXA and Affymetrix gene differential expression values .. 87
Figure 2.4. Correlation of ALEXA and Affymetrix exon differential expression values .. 88
Figure 2.5. Overlap between Affymetrix and ALEXA gene and exon differential expression events 89
Figure 2.6. Exons identified as differentially expressed by ALEXA but not Affymetrix are biased towards low levels of detected expression in the Affymetrix data 90
Figure 2.7. Absolute gene expression values in ALEXA and Affymetrix data 91
Figure 2.8. Absolute exon expression values in ALEXA and Affymetrix data 92
Figure 2.9. The *OLR1/c12orf59* locus is differentially expressed between sensitive and resistant cells 93
Figure 2.10. A known isoform of *LAMA3* is over-expressed in 5-FU resistant cells 94
Figure 2.11. The last 5 exons of *EPB41L3* are over-expressed in 5-FU resistant cells 95
Figure 2.12. The last 9 exons of the predicted protein *c12orf63* are over-expressed in resistant cells 96
Figure 2.13. Reciprocal DE of *UMPS* isoforms 97
Figure 3.1. Annotation of sequence features 140
Figure 3.2. Illustration of read data generation 141
Figure 3.3. Overview of alternative expression analysis 142
Figure 3.4. Distribution of fragment sizes 143
Figure 3.5. Distribution of average Illumina read qualities 144
Figure 3.6. Distribution of read alignment lengths 145
Figure 3.7. Position bias by transcript size 146
Figure 3.8. Read mapping summary 147
Figure 3.9. Comparison of expression estimates from three expression platforms 148
Figure 3.10. Comparison of differential expression from three expression platforms . 149
Figure 3.11. Comparison of expression estimates for the exons and introns of 100 housekeeping genes derived from three expression platforms 150
Figure 3.12. ROC curves comparing sensitivity and specificity between three expression platforms 151
Figure 3.13. Distribution of percent gene coverage at increasing minimum coverage cutoffs 152
Figure 3.14. Coverage of expressed features as a function of library depth 153
Figure 3.15. Change in percent discovery rate with increasing library depth 154
Figure 3.16. Coverage of exon base positions as a function of increasing library depth at varying minimum depth requirements 155
Figure 3.17. Relationship between gene and intron expression estimates 156
Figure 3.18. Expression distribution for all sequence feature types 157

Figure 3.19. Expression of exon regions contrasted with intronic and intergenic regions 158
Figure 3.20. Example of a transcript, *H19* that is much less abundant in 5-FU resistant cells compared to sensitive cells 159
Figure 3.21. The gene *KRT20* is up-regulated in 5-FU resistant cells compared to sensitive cells 159
Figure 3.22. Percentage of exon-skipping junctions with a particular number of exons skipped for known and observed (i.e. expressed) exon junctions 160
Figure 3.23. The *UMPS* gene exhibits reciprocal differential expression of two isoforms 161
Figure 3.24. Example of gene locus, *OCIAD1* with over-expression of several novel exon-skipping isoforms 162
Figure 3.25. Proportion of expressed features observed in sensitive versus resistant cells 163
Figure 4.1. Simplified 5-FU metabolism pathway 198
Figure 4.2. Differential expression of alternative *UMPS* isoforms in 5-FU sensitive and resistant cells 199
Figure 4.3. Whole transcriptome shotgun sequence data corresponding to the *UMPS* locus 200
Figure 4.4. RT-PCR detection of *UMPS* isoforms in six 5-FU sensitive and resistant colorectal cancer cell lines 201
Figure 4.5. Full-ORF cloning of alternative *UMPS* isoforms 202
Figure 4.6. Sequencing of 96 clones isolated from distinct PCR bands 203
Figure 4.7. Western detection of *UMPS* protein in six 5-FU sensitive or resistant colorectal cancer cell lines 204
Figure 4.8. RT-PCR detection of *UMPS* isoforms A and B in a cohort of fresh frozen colorectal cancer tumours 205
Figure 4.9. Real-time quantitative RT-PCR of *UMPS* isoform A and B expression in a cohort of colorectal cancer tumours 206
Figure 4.10. Overview of SNPs and mutations found by genomic sequencing 207

List of equations

Equation 3.1. Average coverage (AC) 138
Equation 3.2. Normalized average coverage (NAC) 138
Equation 3.3. Splicing index (SI) 138
Equation 3.4. Reciprocity index (RI) 139
Equation 3.5. Percent feature contribution (PFC) 139

Acknowledgements

I would like to thank my thesis graduate supervisor Dr. Marco Marra for his guidance and support. He was a true inspiration and an unfailing advocate of my training, education, and research. The importance of having him in my corner was immeasurable. I would also like to thank Lulu Crisostomo who helped me in more ways than I could ever list here. For scientific guidance and support I am grateful to my supervisory committee of Joseph Connors, Stephane Flibotte, Steven Jones and Gregg Morin. I am also grateful for support from Susan O'reilly and clinical collaborations with Sharlene Gill, David Owens and Carl Brown. I am grateful for salary and travel funding from the Natural Sciences and Engineering Council, the Michael Smith Foundation for Health Research, the National Cancer Institute of Canada, the University of British Columbia (Faculty of Medicine and Department of Medical Genetics), the John Bosdet Memorial Fund, Genome Canada, Genome British Columbia, the British Columbia Clinical Genomics Network, and the British Columbia Cancer Foundation. I have enjoyed the support of many fellow graduate students and other peers including Carri-Lyn Mead, Ben Good, Sorana Morrissy, Kim Wong, Noushin Farnoud, George Yang, Claire Hou, Jaswinder Khattra, Monica Sleumer, Erin Pleasance, George Yang, Dan Fornika, Ryan Morin, Trevor Pugh and Obi Griffith. The laboratory work described in this thesis would not have been possible without the help of co-op and volunteer students, especially, Jessica Paul, Pierre Cheung, Alison Lee, Lisa Miao, and Shaun Drummond. In the lab I am indebted to the production sequencing and microarray groups who make large-scale data generation a reality and to Tesa Severson who kept the Marra lab running like a machine. Thanks also to all the members of the Genome Sciences Centre who I do not mention by name but who helped me by creating an open and exciting atmosphere of scientific collaboration. On a personal level, I would like to thank my Grandma and Grandpa, brothers Obi and Alex, sister Olivia, and aunts and uncles for their support and patience. I have been lucky to have wonderful parent-figures in Werner, Dan and Veronica who were source of calm and perspective along the way. Penultimately, to my father Ron, to whom I owe my ambition and inquisitive spirit. And finally to my mother, Rhéa to whom I owe my passion, motivation and so much more. This work and my commitment to cancer research are dedicated to her memory.

Co-authorship statement

Together with my supervisor Dr. Marco Marra, I was responsible for the conceptualization, design and implementation of the research activities described in this thesis. I was primarily responsible for performing the experimental design, laboratory work and data analysis. **Chapters 1**-**4** correspond to multi-author collaborations. These authors contributed to the laboratory work and analyses. I created all the figures and wrote each manuscript in its entirety with the following exceptions. The Illumina library construction description in **Chapter 3** was provided by Yongjun Zhao and **Figure 4.7** was created in part by Ying-Chen Hou. Furthermore, the co-authors provided valuable scientific and editorial contributions throughout the thesis **Chapters 2**, **3** and **4**. I have included the complete author list for each manuscript as a footnote at the beginning of each chapter. Their specific contributions are summarized briefly here. Marco Marra contributed study designs, concepts, editorial suggestions, funding and supervision for all chapters. Isabella Tai provided cell lines, tissue culture materials, training, concepts and experimental designs (**Chapters 2**-**4**). Steven Jones, Gregg Morin and Stephane Flibotte contributed concepts and experimental designs (**Chapters 2**-**4**). Michelle Tang provided guidance for tissue culturing and other laboratory activities (**Chapters 2**-**4**). Susanna Chan, Jennifer Asano, Adrian Ally, and Agnes Baross generated the Affymetrix exon array data (**Chapter 2**) and assisted with cDNA cloning (**Chapter 4**). Martin Hirst, Richard Moore, Thomas Zeng, Yongjun Zhao and Helen McDonald generated Illumina sequencing libraries and sequence data (**Chapter 3**). Obi Griffith, Ryan Morin, Allen Delaney, Kevin Teague, Rodrigo Goya and Irene Li provided programming, statistical and other advice for bioinformatic analyses (**Chapters 2**-**4**). Greg Taylor assisted in sequence assembly of cDNA clones (**Chapter 4**). Trevor Pugh and Tesa Severson provided advice for various laboratory activities (**Chapters 3**-**4**). Ying-Chen Hou and Grace Cheng assisted with the Western analysis (**Chapter 4**). Jessica Paul, Alison Lee, Pierre Cheung, Shaun Drummond, and Lisa Miao assisted in genomic DNA isolation, total RNA isolation, PCR and sequencing (**Chapter 4**). Karen Novik assisted in sample acquisition and ethics applications (**Chapter 4**). Sharlene Gill and Carl Brown assisted in the identification and retrieval of patient samples (**Chapter 4**). David Owen performed pathology review of archival patient samples.

1. Alternative expression analysis: experimental and bioinformatic approaches for the analysis of transcript diversity[1]

1.1. Introduction

The human genome contains approximately 30,000 genes[1, 2]. These loci generate the functional components of the cell but represent only ~1-2% of the entire genome sequence[3]. Although the human genome sequence itself provides a crucial framework for the study of biology, understanding the function of genes requires analysis of the 'transcriptome' encoded by the genome and the 'proteome' it gives rise to. As our knowledge of the genome has increased so too has the realization that gene expression from most of these loci produces a myriad of distinct alternative isoforms with potentially distinct functions. While the regulation of this process and precise number of functional transcripts generated remains to be determined, it is clear that there may be many times more transcripts than genes contained within the human genome. The proteins encoded by these transcripts represent the building blocks and functional components of the cell. Identifying and categorizing the structure of these transcripts is therefore fundamental to our attempt to explain biological processes and bring genomics data to bear on the study of human diseases such as cancer. Furthermore, specific transcripts may represent targets for the development of novel therapies or act as diagnostic and prognostic markers of disease.

1.2. Thesis overview

A principle aim of this thesis was to develop methods for identifying changes in the expression of mRNA isoforms in human disease and implement these methods to improve our understanding of a specific cancer treatment problem. Recent improvements in genome-wide techniques for detecting and measuring the expression of isoforms, particularly microarray hybridization and massively parallel sequencing platforms, now allow researchers to rapidly create an inventory of the mRNA isoforms present in a sample. Preliminary reports describing the use of these technologies have led to an increased appreciation for the prevalence and diversity of mRNA isoforms

[1] A version of this chapter has been published. Griffith M and Marra MA. *Alternative expression analysis: experimental and bioinformatic approaches for the analysis of transcript diversity.* 2007. Chapter 12. Genes Genomes and Genomics, Volume 2. p201-242. Regency Publications. New Delhi.

expressed in human tissues. The potential usefulness of these mRNA isoforms as novel diagnostic and prognostic disease markers as well as their potential as targets of novel therapies has also become apparent. While improved microarray and sequencing platforms make identifying these markers fundamentally possible, bioinformatic methods to assist the process of biomarker identification are lacking. Interpreting the increasingly large and complex datasets and identifying evidence for important markers in these massive datasets remains a particularly daunting challenge. I approached this challenge by developing novel methods to (1) design and analyze custom microarrays specifically tailored to the identification and quantification of mRNA isoforms in human, (2) create a comparable analytical approach that relied on millions of short RNA sequence reads instead of microarrays, (3) illustrate the utility of both of these methods by applying them to a cell line model of chemotherapy resistance in colorectal cancer and (4) characterize a promising alternative expression event potentially relevant to the problem of cancer treatment resistance that was identified by both methods. This process required the creation of bioinformatic tools, generation of genome-wide datasets to validate each approach and analysis of the output to assess the performance of these approaches by comparison to data from complementary platforms. Finally by developing custom algorithms for processing these novel data I generated lists of candidate markers of drug resistance and performed additional experiments to begin to assess the potential clinical utility of a particularly promising candidate mRNA isoform potentially relevant to 5-FU resistance in colorectal cancer.

1.3. Gene expression, alternative expression, and its regulation

The term ‘gene expression’ broadly encompasses the processes of gene transcription, post-transcriptional processing, translation to a protein product and post-translational modification. This thesis is primarily focused on studying the structure of messenger RNAs (mRNAs) that are the result of gene transcription and post-transcriptional processing. Each gene locus consists of discrete regions of sequence called ‘exons’ that will become part of a transcript, separated by regions called ‘introns’ that must be removed to yield a mature mRNA transcript (**Figure 1.1**). Transcription of protein coding genes occurs in the nucleus followed by capping, RNA splicing, polyadenylation and export of the mature messenger RNA to the cytoplasm where translation of proteins occurs. Transcription thus involves three related processes which collectively define the

exon and intron boundaries of a gene and thus the ultimate sequence content of each transcript. First, an RNA polymerase binds to a transcriptionally competent 'unwound' region of genomic DNA template and results in the synthesis of a pre-mRNA molecule in the 5′-to-3′ direction. RNA polymerase II transcribes most human genes and initiates transcription at specific positions in the genome called transcription initiation sites that are found near promoter elements recognized by transcription factors. The initiation site chosen by the polymerase defines the 5′ end of the resulting transcript (i.e. the beginning of the first exon). Second, RNA splicing results in the removal of most of the nucleotides of the pre-mRNA transcript. Splicing involves the recognition of splice sites, removal of introns from a pre-mRNA transcript and joining of adjacent exons (**Figure 1.1**). The splicing process is mediated by a series of protein-protein, RNA-protein, and RNA-RNA interactions involving a number of sequence motifs in addition to the actual splice sites[4]. The splice sites chosen during this process define the primary structure of the resulting transcript. Finally, the 3′ end of the transcript (i.e. the end of the last exon) is defined by a protein complex consisting of polyA polymerase and cleavage factors that cleaves the transcript and adds a poly-A tail 10 to 30 nucleotides downstream from a recognition site in the RNA transcript. Following the discovery of RNA splicing, these three processes were thought to occur in a single prescribed way for each gene and deviations from the 'one-gene-one-product' model were considered rare[5].

A major challenge in decoding the information content of the human genome is presented by the processes of alternative expression (AE), which can produce from a single locus distinct transcripts with different combinations of exons. More precisely, alternate transcripts may arise from a single locus by the use of alternative transcript initiation (ATI), alternative splicing (AS) and alternative polyadenylation (AP) sites. The mechanisms by which these sites are selected by the transcription machinery are tightly coupled to each other[6-8], involving many of the same protein and RNA factors and will be considered collectively as facets of the same biological phenomenon throughout this chapter. The idea that alternative expression dramatically increases the functional diversity of the proteome has gained general acceptance in recent years[9-12]. Based on an analysis of ~1.4 million sequenced human cDNA clones it was estimated that approximately 52% of human genes utilize alternate transcription initiation sites[13]. Similarly, recent estimates suggest that as many as ~94-95% of human genes undergo alternative splicing, a process which can produce multiple transcripts with different

combinations of exons from a single gene locus[14, 15]. Alternative splicing produces distinct isoforms by several modes including: exon skipping, use of alternate mutually exclusive exons, use of alternate 5' or 3' splice sites and the retention of intronic sequences (**Figure 1.1** and **Figure 1.2**)[16]. Recognition of a particular exon by the splicing machinery is mediated by splicing acceptor and donor sites which define the boundaries of each exon as well as by exonic and intronic splicing enhancers and silencers[4, 17]. Finally, a recent annotation of the transcripts for ~8,000 human genes in the 'AltTrans' database suggests that ~60% of human genes utilize alternate polyadenylation sites[18]. **Figures 1.1–1.3** summarize the types of alternative expression sites and some of the surrounding motifs which influence their selection by the transcriptional machinery. Current challenges of genome research are to catalogue all possible transcriptional outcomes for every gene; to define the pattern of expression of these transcripts associated with development, tissue and disease states; and to determine the regulatory networks which control these patterns. A detailed description of the regulation of these processes is beyond the scope of this thesis, but excellent reviews on the mechanisms of regulation and the experimental methods used to study them are available[4, 19-22].

Based on the apparent prevalence of alternate transcript initiation sites, splice sites, and polyadenylation sites, the number of proteins encoded by the human genome is likely to be much greater than the number of gene loci and has been estimated to be as high as 100,000[23, 24]. The biological consequences of this observation are significant. AE appears to be an important mechanism for encoding a diversity of functions at a single genomic locus and this diversity may be realized in part through alterations in protein-protein interactions and subcellular localization. Mutations or polymorphisms in the genes responsible for transcription initiation, splicing and polyadenylation may affect the transcriptional outcome of many genes and contribute to disease ('trans-acting' effects)[25]. Similarly, inherited or acquired mutations and common polymorphisms within the sequence motifs which regulate these processes for each individual gene could also contribute to disease ('cis-acting' effects). A recent analysis of common genetic variation contributing to gene expression differences in the CEU HapMap population found that only 39% of SNPs associated with gene expression corresponded to changes in whole gene expression compared to 55% that resulted in isoform specific changes[26]. Thus, to effectively characterize the human transcriptome and apply

knowledge of whole gene and transcript expression patterns to problems of medical significance, it is necessary to document the prevalence of AE and consider the biological roles of proteins encoded by alternative transcripts.

Until recently it was not possible to measure the prevalence of AE or detect comprehensively the diverse transcripts produced by it. With the availability of high-density microarrays and the advent of next-generation sequencing technologies, there are now opportunities to study AE on a genome-wide scale. The implications of these technical developments and their application to the study of AE are substantial, for up until this time measurements of gene expression relied largely on the detection of a single transcript for each gene. Microarrays designed to detect differential AE will drive the discovery of transcripts with novel, functionally relevant exon combinations, and such discoveries will inform on the protein coding potential of metazoan genomes. Similarly, ready access to sequence data for multiple transcripts from a single locus will provide invaluable validation of their precise sequence content. In addition to fueling basic research questions, it is easy to imagine how knowledge of the transcripts and proteins produced by AE could lead to medically relevant discoveries. For example, novel exon combinations expressed in disease states might yield excellent candidates for development of new diagnostic tools and therapies (see **Figure 1.4** for an example).

Having introduced what is meant by the term 'alternative expression' and described how a single locus can produce multiple distinct transcripts, the remainder of this chapter will provide background material for **Chapters 2**-**4** by addressing the following areas: (1) the experimental and bioinformatic approaches currently available to profile transcript diversity and what these methods have revealed about the prevalence and nature of AE, (2) the functional significance of AE (3) the implications of AE for the study of disease, and (4) a brief introduction to cancer, colorectal cancer and chemotherapy resistance.

1.4. Genomic approaches for the study of transcript diversity

The prevalence and perceived importance of AE has increased dramatically over the last two decades. For example, early estimates suggested that alternative splicing was a relatively unusual event occurring in approximately 5% of all genes[5]. The advent of genome-wide studies of transcript diversity, involving the analysis of short expressed sequence tags (ESTs) by alignment to the genome and annotation of the exons

revealed by such alignments resulted in predictions that at least 42% of human genes exhibit AS[27]. Such studies have also resulted in the creation of several databases of observed alternative initiation, alternative splicing, and alternative polyadenylation events as well as the identification of AE regulatory motifs for a number of species (**Table 1.1**). More recently, exon-junction microarray experiments used to survey splicing events in 52 human tissues and cell lines found that as many as 74% of all human genes are alternatively spliced[28]. One rationale for identifying the full spectrum of alternative expression is that the determination of gene function and identification of therapeutic targets can be improved by first cataloguing the subset of genes and isoforms which are actually expressed in relevant tissues and disease states. Preliminary experiments suggest that AE occurs most frequently in tissues with diverse cell types such as brain, metabolically active tissues such as testis and liver and cell types with highly diversified functions such as immune cells[29-32]. The following sections will describe the computational and experimental ways in which transcript diversity can be studied by examining genomic DNA sequence, full-length cDNA library sequencing, microarray approaches, tag-based or massively parallel short-read cDNA library sequencing, and finally methods for the visualization and functional validation of alternative transcripts. The advantages and disadvantages of each of these approaches are summarized in **Table 1.1**. Each method is depicted in **Figure 1.5** and **Figure 1.6**.

1.4.1. In silico methods

One starting point for the analysis of a species' transcriptional units and often one of the first large sources of data with relevance to analysis of alternative expression is the genome sequence itself. Perhaps the most important issue faced in analyzing the transcript diversity generated by a particular genome is the problem of accurate and reliable annotation of the genes present. Several algorithms which attempt to annotate the genome by predicting gene structure have been described[33]. Generally these predict a single transcript per gene but some have been adapted to consider the occurrence of multiple alternative transcripts generated from a single locus. A few computational methods have also been recently developed specifically to predict AE directly from genomic sequences without the use of experimentally derived expression data. For example, methods have been developed for the prediction of exon skipping

events by considering only the genomic sequence of an exon in the human genome and its orthologous sequence in another species such as mouse[34-36]. This approach is aided by the fact that the sequence of alternative exons and the flanking intronic sequence exhibit generally higher levels of conservation between related species than the sequence of 'constitutive' exons (those found in every transcript)[37, 38]. Each of these methods generally requires a training set of a few thousand known exon skipping events that are conserved between human and mouse. Although these methods are capable of predicting exon skipping events based solely on the genomic sequence of human and mouse, the data sets used to train them are derived from previously observed expressed sequence tags (ESTs). The training set is used to develop a model by which a 'signature' or classifier is created to enable prediction of skipped exons across the entire genome. Experimental validations of the predictions of these methods have revealed a sensitivity value as high as 73% at 64% specificity[34]. Based on the assumption that alternatively transcribed exons will be highly conserved and surrounded by highly conserved intronic sequences, it is also possible to accurately predict such events based solely on the genomic sequence of related species without use of an EST training set. Philipps et al.[39] used this approach to identify alternative exons representing all of the major classes of AS (**Figure 1.2**) in *Drosophila* by comparing the genomic sequence of *D. melanogaster* and *D. pseudoobscura*. The authors were able to confirm AS in 25% of the predicted alternatively spliced exons generated from this approach by RT-PCR whereas only 3% of randomly selected exons were found to be alternatively spliced. The pool of alternative exons that were confirmed in this experiment was found to be enriched for exons that preserve the reading frame of the predicted protein and the extent of highly conserved intronic sequence surrounding these exons was found to be larger than in constitutive exons. Since these initial reports, more sophisticated methods for distinguishing alternative exons from constitutive exons have emerged. For example, a support vector machine (SVM) learning procedure was used to develop a classifier for identification of alternative exons based on seven major exon attributes (exon size, divisibility of exon size by 3, conservation, splice site strength, etc.) and several additional minor attributes[40]. This approach achieved a sensitivity of 50% with a corresponding specificity of 99.5% for human exons. Methods that are conceptually similar to this approach but use a hidden Markov model (HMM) instead of an SVM to identify

alternative exons have also been described[41, 42]. One of the problems faced by all conservation based AE prediction approaches is that they are difficult to implement for small exons and they are incapable of predicting species-specific events.

An algorithm called 'AUGUSTUS' was proposed as the first purely ab initio method for gene prediction. This method is capable of predicting multiple transcripts for a gene from the sequence features of a single underlying genomic sequence without using conservation between sequences or expression data[43]. Xia et al.[44] also recently described an ab initio method for identifying alternative splice sites which uses a model of predicted competition between neighboring splice sites to classify exons as either constitutive or alternative based on their genomic sequence alone. Although these approaches may be useful for analysis of species where very little expression data or suitable comparative genomes are available, in general such methods perform poorly compared to those that can incorporate comparative genomics and alignment of ESTs or full-length transcripts[43].

1.4.2. Microarray methods

Microarrays consisting of spotted cDNAs or short (25 to 60-mer) oligonucleotides have been used extensively to rapidly and simultaneously determine the overall level of mRNA expression of thousands of genes in a single sample. Briefly, a microarray is a small ordered grid of 'spots' (probes) each consisting of many copies of a single-stranded DNA sequence complementary to a small portion of a target gene. A microarray experiment involves extracting RNA from cells, converting the RNA to cDNA, labeling the cDNA molecules with a fluorescent dye, and hybridizing the labeled sample to an array. Each probe spot forms hybrids with copies of its target sequence and the degree of hybridization is measured by scanning the array and recording fluorescence intensities. The magnitude of the intensity observed at each spot is thus a representation of the amount of probe-target hybridization and therefore an estimate of the number of copies of each target in the sample. Each probe on the array acts as a quantitative detector for a particular RNA sequence. Choosing the size and position of the sequence to target with each probe is an area of active development and is critical to the results of a microarray experiment. The general design and use of microarrays to detect gene expression has been reviewed extensively[45] and each of the following microarray strategies are summarized in **Table 1.1** and depicted in **Figure 1.5**.

1.4.2.1. 'First generation' expression arrays

Despite the heavy use of microarrays for measuring gene expression, the use of these arrays to distinguish alternative transcripts has been limited. Spotted cDNA arrays use probes consisting of copies of entire cDNA transcripts or relatively large portions of them and are therefore unsuitable for the detection of alternative transcripts which have subtle differences involving only a small percentage of their total sequence content. Commercially available oligonucleotide microarrays such as those offered by Affymetrix Inc., NimbleGen Inc., Agilent Inc. and others are composed of sets of 10-20 short probe sequences per gene and therefore have higher resolution for detecting transcription (**Figure 1.5**). However, these designs and corresponding oligo d(T) based labeling procedures have heavily biased detection towards the 3′ end of transcripts (often confined to the UTR), limiting their ability to detect many alternative transcripts. Despite the limitations of these designs, the use of the raw probe values generated from these platforms to predict differential expression of alternate transcripts with variable exons at their 3′ end has been described[46, 47].

1.4.2.2. Whole genome and exon tiling arrays

Whole genome tiling arrays have emerged as a method of profiling transcription across large portions of the genome. These arrays consist of probes representing every non-repetitive base of a genome at 5-35 bp intervals (**Figure 1.5**). Because of this comprehensive approach, these arrays are not limited by the accuracy of gene annotations at the time of array design, but rather the completeness and accuracy of the genome sequence itself. Whole genome tiling arrays are theoretically capable of simultaneously determining the approximate exon-intron boundaries of all genes regardless of their current annotation status and also provide a quantitative measure of expression at every exon of every locus. Due to the size of the human genome, initial experiments focused on the smallest human chromosomes only (20, 21 and 22)[48-50]. Arrays of 25- or 60-mer oligonucleotides were designed to tile across non-repetitive genomic sequence at 30-35 bp intervals and these arrays were hybridized with cytoplasmic polyA+ RNAs isolated from a variety of cell lines and tissues. These and subsequent experiments covering 10 human chromosomes at 5 bp resolution[51] and the entire human genome at ~50 bp resolution[52] have revealed considerable evidence for previously unannotated expression throughout the genome. Despite advances in

microarray technology, the resources required to conduct such experiments are still daunting. For example, achieving ~50 bp resolution on both strands of the entire human genome required ~52 million probes distributed across 134 microarrays each of which was only hybridized with a single polyA+ RNA sample isolated from human liver tissue[52]. In other words, an extremely large number of probes were used to measure transcription of select regions of the genome (those that are actually expressed) from only a single tissue. Furthermore, despite the scale of this approach these arrays were not designed to allow inference of the connectivity of exons. Because of the comprehensive probe design strategy used in these arrays they are adept for detecting novel genes, novel alternative exons within the introns of known genes and novel alternative exon boundaries. However, as the quality of gene annotation improves for the genome of interest, the value of using such a large number of speculative probes is reduced and space on the array can be reclaimed to be used more efficiently. Just as large scale sequencing efforts have revealed an unexpected level of transcript diversity at most loci, whole genome tiling array experiments have challenged accepted notions of the percentage of the genome that is actually transcribed, with indications that the transcribed portion of the genome might be much larger than previously suspected[53]. Whole genome tiling arrays are likely to play an important role in continuing annotation efforts but currently have limited feasibility for profiling transcript diversity.

Affymetrix now offers exon tiling arrays which attempt to use array space more judiciously by designing probes for only those regions which are known to be expressed or predicted to be expressed by gene finding algorithms. Affymetrix's exon tiling arrays are created with a photolithographic in situ oligonucleotide synthesis platform and for the human exome, the design consists of a single array with ~5.5 million features corresponding to ~1.2 million known or predicted exons. This capacity allows each human exon to be covered by an average of 4 probes. This is by far the highest density array currently available but the oligo length is limited to 25-mers and medium to small scale custom designs are costly. The design strategy successfully overcomes some of the limitations of previous Affymetrix gene expression designs (such as the focus on measuring the 3′ end of each gene), but these arrays are still unable to elucidate the connectivity of exons and may yield uninformative results when multiple isoforms are present in the same sample (**Figure 1.5**). Furthermore, Affymetrix currently only offers designs for the human, mouse and rat genomes. For researchers who do not wish to

be limited to probes that only interrogate the exons of each gene or who wish to study AE in additional species, a number of options are available for printed or bead based custom designs of 150 thousand to 1 million features (Agilent, Illumina and others). The maskless photolithography procedure of NimbleGen remains the highest density custom array option allowing 385 thousand to 2.1 million features and additional advantages such as the ability to create probes up to 60 nucleotides in length as well as variable length 'isothermal' designs[54].

1.4.2.3. Splicing arrays

As discussed, 'traditional' microarrays have been designed to measure the expression of only a single canonical transcript of each gene and do not account for the existence of alternate isoforms. The idea of using 'splicing' microarrays consisting of exon-junction and other probe configurations to detect AS events was first suggested by Douglas Black[9]. Since 2002, a number of groups have begun to experiment with measuring expression in the context of AE by using such modifications of existing microarray technology (early efforts were reviewed in[55]). ExonHit Therapeutics offers a commercial service for detection of AS in selected therapeutic targets[56, 57]. Jivan Biologics offers the 'TransExpress™ Whole Spliceome' array which includes probes for ~135,000 alternately spliced sites corresponding to ~23,000 human genes. The splice events selected for this array were identified by bioinformatic analysis of existing EST data. In addition to these commercial options, several groups have described the development of custom splicing arrays using commercially available in situ oligonucleotide synthesis or printing platforms.

A number of studies have specifically addressed the theoretical and practical issues of designing custom splicing microarrays to detect AE events by conducting proof-of-principle experiments in several metazoan species[28, 58-62]. Issues addressed by these experiments include the following. (**1**) Accurately annotating gene models to assist in the selection of oligonucleotides. Annotation involves the identification of all exons for every gene, the precise boundaries of each exon, and the putative connections of these exons to each other. The utility of a splicing microarray is fundamentally limited by the accuracy and comprehensiveness of this annotation process. Defining exon regions as either 'constitutive' or 'alternative' by examining existing expression data is also desirable to distinguish between whole-gene expression and alternative gene

expression. (**2**) Storing gene models and annotations of splicing events in a computer interpretable format such as "splicing graphs"[63]. (**3**) Selecting the number and types of AS events to profile. For example, one may wish to target only sequences within exon boundaries. If the identification of complicated splicing patterns is desired it may be prudent to target exon boundaries, exon junctions, and introns as well (**Figure 1.5**). Each of the array design strategies used to date falls into one of two general categories. In one case, transcript annotations based on existing expression data (ESTs, cDNAs, etc.) are assumed to be an acceptable representation of the transcript diversity in the genome and used to identify known AE events which are then specifically targeted by the array. In the second case, the array design attempts to comprehensively profile all exons and splicing events regardless of existing expression evidence. This approach requires considerably more probes but it has the potential to identify the expression of novel expression events. (**4**) Optimizing the specificity and thermodynamic properties of probes to improve the ability of each probe to accurately and reliably predict the presence of their target during hybridization. A uniformity of probe melting temperature (Tm) and length across the array is desirable. Furthermore, probes that form secondary structures, have low-complexity regions, match repetitive elements, or correspond to expressed sequences from multiple regions of the genome should be avoided. For members of gene families or genes with pseudogenes, it may not be possible to select specific probes. Furthermore, when targeting large exons and introns, choosing an 'optimal' probe is often straightforward, but when the target sequence is constrained to a small exon or a specific exon junction or boundary this may not be possible. (**5**) Reducing 'half-junction crosstalk'[59]. This term refers to a problem related to the use of exon junction probes such that each probe hybridizes over each half of its length to targets containing the same exon sequences in combinations other than those specifically targeted by the junction probe. For example, a probe designed to detect the juxtaposition of exon 1 with exon 3 (e1^e3) will hybridize on each half to RNAs containing e1^e2 and e2^e3. This crosstalk effect increases as the length of a probe is increased or hybridization stringency is reduced. The junction probe length that maximizes sensitivity and specificity has been empirically determined as 35-45 nucleotides in length[58-60]. Crosstalk can theoretically be reduced by offsetting the probe position on the exon junction or allowing the two halves to differ in length such that the difference in Tm between the two halves is minimized. The proof-of-principle

experiments that have helped to resolve these five issues provide invaluable guidance for researchers wishing to create custom splicing arrays without spending considerable time and resources conducting optimization experiments. Furthermore, their results provide general evidence that the splicing microarray approach is feasible. For example, experiments using samples spiked with different mixtures of cloned human and *Drosophila* transcripts showed that an alternate isoform making up as little as 20% of a mixture of two isoforms could be detected by junction probes (observed fold differences were highly correlated with expected values over a range of 0.25 to 12)[59, 64].

The first three large scale experiments with splicing microarrays were conducted in human, mouse and *Drosophila*[28, 61, 62]. Johnson et al.[28] conducted a genome-wide survey of AS in 52 human tissues using a total of 125,000 exon junction probes corresponding to the expected canonical junctions of 10,000 multi-exon genes. The authors observed that similar tissues tend to have similar AS patterns and cell lines have their own distinct patterns, in particular exhibiting the expression of fewer genes but more variants of those genes. By extrapolating from their results and comparing to EST data the authors predicted that 74% of all human genes are alternatively spliced. Pan et al.[62] used a similar approach to analyze ~3,000 previously observed AS events in 10 mouse tissues. Based on RT-PCR validations of the predictions of their splicing microarray the authors determined that the array could predict differential expression of isoforms between tissues with a specificity of approximately 80%. The data described in this initial experiment has recently been analyzed to show that exons that have varying expression levels across mouse tissues are more likely to be a multiple of 3 in length (perhaps indicating a selection for maintenance of reading frame) and are highly conserved relative to constitutively spliced exons[65]. These data have also been used to investigate the potential coupling of AS and nonsense-mediated mRNA decay as a global means of controlling transcript abundance[66].

As the number of published splicing microarray experiments has increased, the variety of analysis methods has also increased[67]. Nevertheless, the availability of suitable analysis methods with open source software implementations remains a challenge to researchers who wish to conduct their own splicing microarray experiments. Standard methods for normalization, background correction and summarizing multiple probe values into a single gene- or exon-level expression estimate may be used [68-70] but methods which specifically address the identification of

differences at the level of alternative transcripts are still required. To date, at least seven distinct analytical methods for identifying differences in isoform expression from splicing microarray data have been proposed: (1) Splicing index values[58, 60, 71], (2) detecting systematic anti-correlation between the log-ratios of two different samples versus a pool containing both samples[72], (3) splice and neighborhood algorithms[46, 47], (4) analysis of splice variation (ANOSVA)[73], (5) sequence based splice variant deconvolution[74], (6) inferring global levels of alternative splicing isoforms using a generative model[75] and (7) microarray detection of alternative splicing (MIDAS)[76]. Although each of these methods uses different mathematical and statistical techniques, the general goal of each is to identify alternative exons, junctions, or whole transcripts that are differentially expressed between two samples. Identifying such events invariably involves some attempt to account for changes in expression at the gene level. For example, the work by Clark et al.[60] was the first to propose the use of a 'splicing index' calculation to identify AS events and several studies since have used it including those described in **Chapter 2** and **Chapter 3** of this thesis. A splicing index is determined by first comparing the expression of each exon to the expression value for the entire gene within a single sample. This results in a 'within-gene' normalized value for each exon which can then be compared across sample pairs to create the splicing index. Statistical methods such as MIDAS also use within-gene normalized values but attempt to identify significant differentially spliced exons by considering the magnitude and variability of exon expression within grouped samples compared to across sample groups (e.g. ten normal versus ten cancer samples).

Using the developments in splicing microarray design and analysis described above, several research groups have applied these arrays to the study of specific biological problems. These include estimating the global prevalence of AE in tissues and throughout development, assessing the implications of AE for protein diversity, studying splicing regulation at the level of trans-acting factors, defining novel cis-acting splicing motifs, and identifying isoforms with disease relevance.

Relogio *et al*.[77] was among the first to publish results obtained using microarray technology to specifically address the role of AS in a cancer model. This group designed a custom array to measure the expression of 86 splicing-related genes and known splicing events in 10 cancer genes and applied their array to RNAs derived from four cell lines representing different stages of Hodgkin lymphoma tumors. Clustering of

the microarray results for 100 splicing events revealed distinct patterns for each of the four tumor stages. Li et al.[71] recently used splicing microarrays to identify differential expression of alternative isoforms between estrogen receptor positive and negative breast cancer cell lines. Zhang et al.[78] predicted that profiling expression at the level of individual exons and AE events with a splicing microarray would improve the accuracy of expression based cancer classification compared to using overall mRNA expression levels. They demonstrated this by conducting a classification of 38 cancer and normal prostate tissues by measuring the expression of 464 isoforms of ~200 genes and concluded that profiling the expression of alternative transcripts increased the information content by at least 30% compared to conventional microarray data. In addition to studying human disease, splicing microarrays have also been shown to have great potential for defining a global 'splicing code' by studying the expression of thousands of exons and identifying novel sequence motifs as well as how the arrangement of these motifs and their interaction with particular trans-acting factors influences the splicing of specific exons in a tissue dependent manor[79-82]. For example, Blanchette et al.[81] used a splicing microarray to study the global effects of RNAi knockdowns of four splicing regulators (two hnRNPs and two SR proteins) in Drosophila. Knocking down each of these four proteins affected a variable number of splicing events, ranging from ~50 to more than 300. Since their array design was limited to only those events that had been previously observed (~8,000 events observed for ~3,000 genes in EST/mRNA data), these are likely to be underestimates. A similar experiment which involved a knock-down of factors involved in nonsense mediated decay (NMD) successfully identified showed that alternative splicing may function as a means of controlling gene regulation via NMD in Drosophila[83]. Perhaps the most comprehensive application of splicing sensitive microarrays used the approach I describe in this thesis (**Chapter 2**) to create an alternative splicing compendium by profiling ~25,000 splicing events across 48 tissues and cell lines[84].

1.4.3. Library construction and sequencing methods

1.4.3.1. EST sequencing of cDNA libraries

The earliest large repositories of data on transcript diversity consisted of expressed sequence tags (ESTs) generated by single sequence reads from systematically selected cDNA clones. Construction of a cDNA library commonly involves extraction of

total RNA from cells, purification of polyA+ mRNAs, RT-PCR with an oligo d(T) primer and cloning into a convenient vector. The rapid generation and sequencing of these libraries from specific human tissues became common in the early 1990's and rapidly accelerated the discovery and annotation of novel genes[85, 86]. EST libraries are generally derived from a single normal or diseased tissue sample or a small pool of tissue samples. Most EST records deposited in public databases contain information on the tissue source and disease status of the sample from which they were derived. Typically each EST represents either the 5′ or 3′ end of a clone and initially the lengths of these reads were 300-500 nucleotides (**Figure 1.6**). Although improvements in Sanger sequencing have approximately doubled this read length, the majority of all ESTs do not represent a complete cDNA sequence and the overall coverage of EST data is heavily biased towards the 3′ end of transcripts[87]. The completion of the human genome, the comprehensive sequencing of EST libraries from a variety of tissues, and the continuing development of algorithms for 'spliced' alignments such as Blat, Spidey and Sim4 has allowed a first comprehensive assessment of the diversity of transcription (refer to **Table 1.2** for a list of spliced alignment algorithms). EST sequences can be rapidly generated and aligned to a reference genome allowing the annotation of exon-intron boundaries and the inference of underlying transcript isoforms[88, 89]. Protein coding information may also be incorporated into predictions by performing 6-frame translations. The size of an EST library has been historically as small as a few hundred sequences or as large as tens of thousands and in rare cases even larger. Since a single cell type is likely to express 10,000 to 30,000 genes with a total of approximately 300k to 500k mRNA molecules per cell, the coverage of these EST libraries is not likely to provide an accurate quantitative measure for the expression of genes in a bulk tissue sample, especially given the fact that a majority of all transcripts will be derived from a minority of loci[90]. The problem of over-representation of highly expressed genes can be addressed by applying normalization techniques during the library construction phase[91, 92]. Such techniques enhance the rate of gene discovery but reduce the quantitative value of the data generated from such libraries. Library normalization techniques can also have the side effect of reducing the presence of transcript variants with subtle but potentially important variations and estimates of AE prevalence in the genome are likely to be underestimates as a result. Approximately 62 million EST sequences have been deposited in the public repository dbEST, of which 8.3 million were generated from

human samples (www.ncbi.nlm.nih.gov/dbEST/)[93]. This collection represents an incredible source of independent transcription observations from a wide variety of tissues and it has been used to identify differentially expressed genes specifically associated with particular tissues or disease states[90]. Perhaps the most prominent examples of the use of EST sequencing are the Cancer Genome Anatomy Project, which has attempted to create a complete catalogue of genes expressed in normal and cancerous tissues, and Unigene, which attempts to group all such sequences into clusters of sequences expressed from a single locus[94].

Analysis of ESTs has proved to be a rich source for discovery of novel genes and transcript diversity and has led to a number of interesting observations about transcription. Early analyses suggested that most AS events affect the 5′ UTR of genes, occur in at least 35% to 42% of all genes[29, 95], and seem to be more prevalent in humans than in other species considered to date[96]. Furthermore, within humans, the prevalence of AE varies dramatically between tissues. Brain and testis have the most exon-skipping events and liver has the most alternate splice site usage but one of the lowest rates of exon skipping[30]. Certain protein domains seem to be preferentially affected by AE and more than 50 domains that are commonly removed by AE have been identified[97]. Analysis of these domains indicates that one of the central roles of AE may be to modulate protein-protein interactions. A number of groups have used ESTs to create databases of annotated AE events and characterize some of the general features of transcript diversity in metazoan species (refer to **Table 1.2** for a complete list). Among the results of these studies were the observations that skipped exons tend to be shorter than constitutively spliced exons, retained introns are generally shorter than those that are constitutively spliced, the introns flanking skipped exons tend to be longer, skipped exons are more likely than constitutively spliced exons to have a length that is a multiple of three, splice sites corresponding to constitutively spliced events tend to more closely resemble the consensus sequence than those involved in AS events, and the average sequence conservation between human and mouse is greater for alternatively spliced exons than constitutively spliced exons.

1.4.3.2. Full-length sequencing of cDNA libraries

As the cost of Sanger sequencing and primer synthesis has gone down it has become more practical to conduct full length sequencing of cDNA clones representing complete

transcripts (**Figure 1.6**). This is conceptually the simplest approach to study transcript diversity because it involves the capture and complete sequencing of single cDNAs. The complete structure of the transcript including the presence of alternative exons is thus determined. Large scale cDNA sequencing projects such as those associated with the Mammalian Gene Collection (MGC) and Full-length Long Japan (FLJ) projects are at various stages of completion for human, mouse and other species[98-100]. The cDNA libraries for these efforts are generated in ways similar to those employed for EST sequencing but additional emphasis is placed on the generation of 'full-ORF' cDNAs. Sequencing of these cDNA clones involves generating EST end reads followed by sequencing of the remainder of the cDNA insert using primer walking, transposon mediated sequencing, or cDNA concatenation[101-103]. The resulting reads are then assembled into a contiguous sequence representing the entire mRNA. Initially, clones were selected for full-length sequencing by first generating EST end reads and identifying a subset of non-redundant clones. Although the primary goal of the MGC was to create a physical resource of cDNA clones for the analysis of gene function, the process of rescuing and sequencing these clones has led to the discovery of considerable transcript diversity. The random clone sequencing approach initially used by the MGC effort was replaced by an RT-PCR targeted approach in which amplicons for a known target gene were generated, cloned and sequenced[104]. The random clone sequencing approach has the potential to identify transcripts that differ in their transcription initiation, polyadenylation, and splicing. Because the targeted approach pre-defines the expected ends of the transcript it is only capable of detecting splice variation that occurs within these boundaries. However, since the cloning and rescue process generates many clones per target sequence, novel transcript variants of this type are routinely observed (**Figure 1.6**). The MGC collection currently contains clones for ~17,500 human genes generated from more than one hundred tissue libraries. A recent study of ~56,000 full-length human clone sequences from the 'H-invitational human transcriptome' annotation meeting[105] found that these clones could be mapped to ~24,000 loci and 41% of these loci were represented by multiple cDNAs[87]. Of these loci, where at least a preliminary assessment of transcript diversity was possible, 68% showed evidence of AE with an average of approximately three unique transcripts per locus. Of these transcripts, 45% exhibited exon skipping events, 52% used at least one alternate 5′ or 3′ splice site, 15% had retained introns, and 3% used one of a series of

mutually exclusive exons. Only 14% of the intron retention events were predicted to result in a transcript possibly subject to nonsense mediated decay (NMD). The majority (73%) of alternate transcripts exhibited a splicing event within the CDS of the predicted protein but 26% had events confined to the 5′ UTR and 6% had events confined to the 3′ UTR. Furthermore, if the rate of each type of event relative to the number of exons in each of these regions is calculated, events affecting the 5′ UTR have the highest frequency. Of all genes with observed AE events, 44% had events which occurred within a known protein motif, 44% were predicted to affect subcellular localization, and 20% were predicted to affect a transmembrane domain. Although many human gene loci are still represented by only a single clone sequence (59% in the study above), this initial data will act as a foundation for future studies of the diversity of transcripts generated from these loci (in EnsEMBL version 53, 48% of protein coding genes still have only one transcript). Analysis of almost 200,000 publicly available full length clone sequences derived from ~200 mouse tissues have resulted in similar findings to those observed in human. At least 40-70% of mouse genes have evidence for AE[87, 99, 106, 107] and an estimated 78,000 distinct proteins are transcribed from only ~20,000 loci[108]. As described for the analysis of large EST datasets, these studies are invaluable for identifying the types of alternative transcripts that occur, revealing patterns in the size distribution, sequence composition and conservation of alternatively transcribed exons themselves and predicting their effect on resulting proteins[107, 109].

The complexity of the mammalian transcriptome generated by AE has been accepted as an outstanding challenge and was specifically discussed at the outset of the Mammalian Gene Collection project which has focused on the goal of acquiring a single ‘representative’ transcript for each known gene[102]. Creating a comprehensive annotation of the complete mammalian transcriptome remains a significant challenge as does obtaining a cDNA clone representing every transcript variant of every gene. Although methods that involve RT-PCR, cloning and sequencing of alternate transcripts can be accurate and reveal much about the structural differences of alternate isoforms, they are costly and difficult to scale. Most analysis of EST and cDNA sequence data has focused on gene annotation and transcript variant discovery rather than quantitative profiling of transcript expression levels. The limited sampling depth generated by EST or full-length cDNA sequencing is insufficient to provide robust identification and quantification of alternatively spliced variants across samples representing comparisons

of large number of tissues, patients or disease states. Bioinformatic analyses of all publicly available EST data are more comprehensive but are limited by the coverage of existing libraries and other problems such as end bias. The EST and cDNA libraries that are publicly available were not specifically intended to provide an accurate and consistent comparison of tissues or the progression of disease states, and often represent pools of individuals or cell types. Furthermore, although the use of EST and cDNA data to study splicing can be effective and has led to significant advances in our knowledge of AE it remains expensive and time consuming to create and sequence libraries of sufficient depth to quantitatively survey the transcripts present in samples representing several conditions.

Experimental approaches have recently been developed to specifically enrich libraries for alternative transcripts and thus increase the discovery of novel transcripts. One approach involves the construction of alternative splicing libraries (ASLs) representing differentially expressed exons from pairs of biological samples[110]. Briefly, this protocol involves creating two cDNA libraries from cytoplasmic RNA, one from each of the samples to be compared. These two libraries are then processed such that single stranded sense DNA molecules are generated from one library and single stranded antisense DNA molecules are generated from the second library. The two libraries are then mixed to allow hybridization and formation of heteroduplex or 'loop' structures. This can occur in the event that a transcript from one library contains exon content not found in the corresponding transcript present in the second library. Hybrid molecules containing these loop structures are then selectively captured with biotin labeled random 25-mers which are purified on streptavidin conjugated magnetic beads and the resulting alternative transcript enriched cDNA population is cloned and sequenced. Use of this approach to compare melanocyte and melanoma cell lines identified 662 AS events representing all of the major categories of AS and differential splicing between the two cell lines was confirmed by RT-PCR for 73% of candidate exons. A comparison of this library construction approach to one without the splicing selection step suggested a ~40-fold enrichment for AS events. Thill et al.[111] recently described a similar method, 'ASEtrap' for the construction of libraries enriched for alternative splicing events from a single RNA sample (rather than from a comparison of two samples). This method also utilizes the formation of loop structures in cDNA heteroduplexes caused by alternative transcripts of a single gene within the sample.

These loops are captured by a recombinant *Escherichia coli* single-stranded DNA binding protein and then cloned and sequenced. Comparison of ~10,000 sequences generated from either an ASEtrap library or a control library revealed a ~10-fold enrichment for AS events in the ASEtrap library. A third approach for enrichment of AS isoforms (EASI) was recently proposed as a simpler version of the ASEtrap method which can be rapidly employed to comprehensively profile all of the isoforms of a single target gene[112].

1.4.3.3. Sequence-tag based methods

The simplest way to overcome the issues of poor representation of rare transcripts and lack of quantitative power in sequence based methods such as EST and full-length cDNA sequencing is to increase the number of sequences available for analysis. Serial analysis of gene expression (SAGE)[113, 114] has been used as an alternative to EST sequencing and libraries as large as several hundred thousand [115, 116] or millions[117] of tags have been reported. SAGE involves double stranded cDNA synthesis with an oligo(dT) primer, followed by digestion of the resulting cDNA with a restriction enzyme predicted to result in at least one cleavage per transcript (typically NlaIII). The resulting fragments are captured at the 3′ end by oligo(dT) primers coupled to streptavidin beads, and a type II restriction enzyme (e.g. MmeI) which cuts outside its recognition sequence, is used to create fragments of a fixed length (up to 21 bp) which are concatenated, cloned into a vector and sequenced. Each sequence read thus produces 30-45 tags corresponding to the 3′ most NlaIII site of transcripts from which they were derived (**Figure 1.6**). By generating sufficiently large numbers of these reads, a quantitative and digital form of expression data is produced with the number of tags mapped to each genomic locus representing the expression level of that gene. This form of data has been shown to have a moderate correlation ($r = 0.5 – 0.8$) of expression values when compared to microarray-based approaches[118, 119]. Two of the largest initiatives to make use of this technology are the Cancer Genome Anatomy Project[115] and the Mouse Atlas of Gene Expression Project[116], each producing several million tags from a wide range of cell types for human and mouse respectively. Analysis of these large datasets has resulted in the identification of differentially expressed genes associated with disease, development or a specific tissue as well as the discovery of novel genes and transcript variants. An analysis of the SAGE tags

mapping to ~13,000 EnsEMBL genes produced a prediction that 64% of genes exhibit AE and many of the variants observed were significantly differentially expressed in specific tissues or developmental stages in mouse[116]. Several bioinformatic tools to assist in the analysis and visualization of SAGE data have been developed[115, 120]. Of these, only 'SAGE2Splice' was specifically designed to identify novel splice junctions in SAGE tags but is limited in its ability to profile exon connections by the fact that only 5-6% of tags span a splice site[121]. The disadvantages of SAGE include the theoretical occurrence of multiple tags per gene from incomplete digestion and the short length of each tag, both of which complicate the process of mapping tags to the gene from which they were expressed. Distinguishing tag artifacts created by mis-priming during library creation from tags derived from the use of alternative polyadenylation sites, alternative splicing or polymorphisms in restriction enzyme sites is also potentially problematic. Finally, because SAGE library construction involves the capture of tags corresponding to restriction enzymes sites closest to the 3′ end of each transcript, any variation observed is heavily biased towards the 3′ end of genes.

A complementary approach to SAGE, cap analysis of gene expression (CAGE), is used in a way similar to SAGE to profile the 5′ end of transcripts and thereby can be used as a means of identifying alternate promoter usage[122]. Briefly, transcripts are captured by their 5′ cap (a modified guanosine nucleotide) and used to generate DNA tags of 20 nucleotides in length which are concatenated, cloned and sequenced. Each sequenced tag corresponds to the 5′ end of a single mRNA transcript and as with SAGE, the short length of each tag allows an increase in throughput and therefore depth of sampling and corresponding reduction in cost. By capturing many tags from a single gene the use of alternate transcription initiation (ATI) sites and their corresponding promoters can be catalogued. Generally 55 - 65% of sequenced tags can be unambiguously mapped to the genome[122]. Analysis of 7.2 and 5.3 million CAGE tags generated from ~200 human and mouse tissues respectively suggests that the use of ATI sites is a common feature of protein coding genes and often results in modified N termini with potentially distinct functions[123]. In both human and mouse, these tags form approximately 200,000 tag clusters which map to ~35,000 loci and ~80% of known protein coding loci are covered by at least one tag cluster. When only protein coding genes were considered, 58% were found to make use of alternative promoters and 93% of these were predicted to result in the use of distinct start codons which for some

genes occurred in a tissue specific manner. Hierarchical clustering of expression levels for all tag clusters revealed distinct global patterns of promoter usage associated with specific tissues, particularly lung, brain and liver.

Experiments that use both SAGE and CAGE have been proposed to allow independent profiling of the 5′ and 3′ ends of transcripts expressed in a single tissue sample[124]. An interesting extension of the 3′ profiling of SAGE and 5′ profiling of CAGE described above has been reported by Ng et al.[125] who developed 'gene identification signature' (GIS) analysis. This approach allows the simultaneous profiling of the 5′ and 3′ end of a transcript by generating paired-end-tags (PETs) from random cDNAs followed by tag concatenation and sequencing. The advantage of this method over combining SAGE and CAGE is that each PET sequence represents a linked transcription start and end position from a single mRNA rather than two independent pools of tags representing start and end positions.

1.4.3.4. Massively parallel sequencing methods

The emergence of 'next generation', massively parallel sequencing technologies[126] has enhanced the potential of sequence based approaches for profiling transcript diversity. The parallel sequencing of many templates on a single compact array was first published in 2000 by a group at Lynx Therapeutics Inc.[127]. This approach, described as massively parallel signature sequencing (MPSS) involved the creation of an array of microbeads, each coupled to a single DNA template, which were used for a ligation-based sequencing protocol involving fluorescently labeled adaptors. Monitoring of fluorescent signals as the sequencing reaction progresses was accomplished by a charge-coupled device (CCD) detector and image analysis, resulting in the simultaneous generation of millions of short sequences. This entire process took place in a flow cell with the array of microbeads remaining in a dense monolayer and reagents flowing past. The accuracy of this platform for profiling gene expression was first assessed by generating ~1.6 million sequences from cDNAs derived from a human cell line and comparing these to EST sequences generated by conventional Sanger sequencing. The resulting qualitative comparison of the most highly expressed genes seemed promising but far from definitive and early MPSS experiments identified strong biases related to the GC content of expressed sequences[128]. Lynx Therapeutics Inc. has since been acquired by Solexa Inc., which in turn has been acquired by Illumina

Inc. The Solexa/Illumina platform has seen rapid and continuous improvements in read quality, length and throughput and has resulted in several publications (some of which are discussed below).

A competing platform that may also be described as a massively parallel sequencing approach has been developed by Roche/454 Life Sciences Inc.[129-131]. This platform, performs sequencing-by-synthesis using a fiber-optic slide with approximately 1.6 million wells (each 44 µm in diameter). The sequencing reaction itself is referred to as 'pyrosequencing', in which fluorescently labeled nucleotides are sequentially washed over the slide and incorporation of each base into a growing complementary strand of a single stranded template DNA is simultaneously observed for all wells by a CCD detector. Homopolymeric sequences in the template DNA result in the incorporation of multiple nucleotides in a single cycle and must be resolved by analyzing the magnitude of fluorescence for each well.

A third platform, referred to as 'SOLiD' offered by Applied Biosystems Inc., became available more recently and initial reports suggest that its ability to profile transcriptomes is comparable to that of the Solexa and 454 platforms[132, 133]. To generate sufficient template for DNA sequencing, this platform uses emulsion PCR amplification of single DNA molecules on beads, which are then deposited on a slide surface. The sequencing reaction then proceeds by repeated ligation of fluorescently labeled di-base probes (e.g. C-A, C-T, etc.) in such a way that each position is interrogated by two independent ligation reactions.

Each of these three 'next generation' sequencing platforms, have achieved dramatic improvements in read length (36-500 bp are now possible depending on the platform) throughput (0.5 - 20 Gb of sequence data per run), and quality. Additional improvements in flexibility have been provided by the adoption of 'paired-end' reads, bar-coding to allow multiplexed analysis of multiple samples within a single sequencing library, and improved library construction protocols to allow analysis of small quantities of nucleic acid.

Several groups have used these sequencing platforms to sequence SAGE-like libraries consisting of tags representing transcript ends or PETs[117, 134-136]. These experiments are conceptually similar to SAGE but are able to produce increased tag counts at reduced cost and have been found to produce gene expression estimates that are similar to longSAGE data ($R^2 = 0.96$)[136]. Gowda et al.[134] used the approach to

profile the 5′ ends of Maize transcripts and described a considerable potential for identifying alternate transcript initiation sites. Ng et al.[135] used a combination of PET library construction and 454 sequencing to generate over 450,000 PETs from the human breast cancer cell line, MCF7. Of these, ~136,000 could be mapped unambiguously to ~21,000 unique loci, and 25% of these represented candidate novel alternative transcript initiation sites or alternative polyadenylation sites. An experiment described by Bainbridge et al.[131] used Roche/454 sequencing to profile full-length transcripts expressed in polyA+ purified RNA from the LnCAP prostate cancer cell line. This direct sequencing of full-length transcripts avoids the artifacts associated with library construction and cloning and does not limit the resulting ESTs to the ends of transcripts. The approach was successful in identifying 25 novel AS events involving known exons but the short read lengths (average of ~100 bp), overrepresentation of a small number of highly expressed genes, and unexpected bias towards transcript ends limited the number of reads which were informative of splice site selection.

More recent reports have described the application of massively parallel RNA sequencing (known as 'RNA-Seq' and whole transcriptome shotgun sequencing or 'WTSS') using the 454, Solexa and SOLID platforms to perform ab initio gene annotation[137, 138] and survey transcript diversity across various diverse tissues in human[14, 15] and mouse[139], human embryonic kidney and B cell lines[140], a prostate cancer cell line[141], a cervical cancer cell line[142], undifferentiated mouse embryonic stem cells and embryoid bodies[132] and mouse blastomeres[133]. Analysis of RNA-seq data revealed that 92-95% of multi-exon genes were alternatively spliced in at least one of 15 tissues and cell lines (where the minor isoform had an expression level of at least 15% of the major isoform)[15]. Variation in splicing was found to be more prevalent between tissues than within the same tissue from different individuals[14, 15]. While many isoforms might represent tissue- or stage- specific markers, transcriptome analysis of a single mouse blastomere still revealed hundreds of genes that expressed at least two isoforms[133]. A comparison of mouse tissues similarly found that the majority (93%) of isoform pairs were found to be co-expressed in the same tissue rather than each being distinct to different tissues[139]. It was also noted that when expression of a particular isoform 'switched' predominantly from one isoform to another between tissues, the result was often the expression of a modified 'full-length' protein, as opposed to a truncated protein or a transcript that would be subject to NMD[15]. This observation, and

the enhanced level of cross-species conservation for exons involved in these tissue-regulated isoforms, provide evidence for the hypothesis that alternative expression increases the cell's repertoire of functionally distinct proteins. While it is likely that many AE events generate functionally distinct proteins, AE is also a mechanism for modifying the level of gene expression. While gene expression is traditionally thought to be controlled by factors that initiate or maintain transcription, AE may provide an additional layer of regulation whereby a gene may be inactivated by switching to a truncated or NMD form without actually decreasing the rate of transcription from the locus.

1.4.4. Limitations of transcriptome analysis methods

1.4.4.1. Limitations of microarray approaches

Many of the technical limitations of microarray analysis stem from the physical limitations of hybridization reactions. For example, the length of oligonucleotide probes used on the array must be optimized in conjunction with the hybridization conditions and must be long enough to ensure sequence specificity (i.e. short sequences are more likely to be redundant to multiple genes). One consequence of this is that short regions, such as those corresponding to small exons, may not be effectively targeted by an oligonucleotide. Another consequence of the physical nature of microarray hybridizations is that they typically require large amounts of RNA (> 5-10 μg), necessitating either large sample inputs or sample amplification. Furthermore, isoforms with low levels of expression may be difficult to distinguish from the levels of background 'noise' common to microarray hybridization. Much of this noise may be attributed to cross-hybridization between the probe and sequences from several genes, particularly members of gene families. Similarly, isoforms with small differences in sequence content such as minor shifts in donor or acceptor site usage or inclusion/exclusion of small exons may be difficult or impossible to detect using microarrays due to the high degree of cross-hybridization expected for sequences with only minor differences. Cross-hybridization is also problematic when profiling exon-exon junctions, as multiple exon junctions representing isoforms from a single locus will by definition share at least ½ of their sequence. Finally, of particular relevance to the study of alternative isoforms is the fact that design of a microarray to detect splicing events is largely dependent on existing gene annotations and the accuracy of exon boundaries present in those gene models.

1.4.4.2. Limitations of sequencing approaches

Although sequencing approaches successfully address many of the challenges encountered with the use of microarrays for AE analysis, they too suffer from limitations. The primary limitation of sequencing approaches for transcriptome analysis relates to their reliance on random sampling of transcriptome space. Only 3-5% of the transcripts in a cell are mRNA molecules, with the remaining transcripts representing a few highly expressed ribosomal RNA (rRNA) species. The majority of rRNA transcripts can be removed by positively selecting for polyA+ sequences or less efficiently by post-sequence computational filtering of rRNA species. Assuming efficient removal of rRNA transcripts, the remaining mRNAs consist of thousands of unique transcripts (at least 10,000) with a large difference in expression level between the least and most abundant mRNA transcripts (at least 10^5) [14, 15, 140]. The high degree of transcript diversity and large dynamic range of expression present a significant challenge for sequencing methods that involve random sampling of a cDNA library. For example, among the mRNA transcripts remaining after rRNA removal, it is estimated that as much as 55% of these are redundant copies of the same mRNAs derived from only 4% of all protein coding loci[143]. Thus, even if a large number of tags can be produced efficiently, sequence based approaches are still faced with the problem of sequencing many transcripts from a few loci at the cost of failing to sample many other loci. For example in a test of Roche/454 Life Sciences GS20 sequencing for the profiling of a cDNA library, we found that ~110,000 reads could be mapped unambiguously to ~8,000 EnsEMBL loci but 39% of these corresponded to only 20 loci[131]. In the analysis described in **Chapter 3**, I found that 50% of all reads correspond to the top 5% of protein coding loci. In addition to this issue of transcript redundancy, because of the complexity of mammalian biology, creating even a snapshot of the human transcriptome remains a daunting challenge. Assuming an average transcript size of ~2000 bp and an average of 300-500k transcripts per cell, sequencing of a single cell type representing just one of hundreds of possible cell types to an average depth of 1X would theoretically require at least ~1 billion bp of sequence[144]. Continued improvements in massively parallel sequencing technologies have largely overcome these sampling limitations and are proving invaluable in characterizing even infrequently expressed transcripts. Furthermore, the combination of sequencing technologies with library normalization strategies such as ‘deep well’ pooling to ensure equal representation of isoforms and

maximize isoform discovery has been successful in avoiding the sampling bias issue, albeit at the cost of complicating library construction[145].

One limitation of both splicing microarrays and massively parallel RNA sequencing is that determining the connection of exons is limited to neighboring exon pairs and the complete connectivity of the exons of a transcript must be inferred from these pairings. Cloning and full-length sequencing remains the best way to unambiguously determine the complete structure of individual transcripts. Improvements in sequencing platforms that allow sequencing of complete cDNA sequences rather than short fragments may overcome this limitation in the future.

An additional limitation of some sequence based approaches is the bias introduced by bacterial cloning constraints in the construction of EST or full-length cDNA libraries. Certain sequences are not well tolerated by bacteria and therefore these sequences are under-represented in sequence libraries. However, massively parallel sequencing platforms do not rely on bacterial cloning for library construction, and thereby avoid such bias.

One approach to overcoming the disadvantages and biases (summarized in **Table 1.1**) inherent to both sequencing and microarray approaches for profiling transcript diversity has been to combine complementary computational and experimental approaches[146]. As a result of these efforts and the continued compilation and synthesis of disparate genome-scale expression data sets in resources such as the UCSC[147] and EnsEMBL[148] genome browsers, researchers now have access to a highly detailed survey of the diversity of transcripts expressed from many loci of many eukaryotic species.

1.4.5. Functional characterization of mRNA isoforms

Due to methodological advances and increases in information as described above, researchers are now increasingly able to identify the complex pattern of alternative transcripts generated by the genes under study in their laboratory. It is therefore becoming increasingly important to have a wide range of tools and protocols to verify expression measurements from high-throughput microarray for sequencing assays and characterize the function of specific isoforms.

Verifying the relative mRNA expression of known or predicted isoforms of a single gene in a tissue of interest is typically accomplished by Northern blot analysis or by

semi-quantitative or quantitative RT-PCR. Similarly, protein-level expression of isoforms with significantly different masses can be confirmed by SDS-PAGE and Western blot analysis with an antibody that recognizes a constitutive portion of the gene. Visualizing the spatial expression of isoforms at the mRNA level can be accomplished by in situ hybridization with digoxigenin labeled riboprobes specific to each isoform[149]. Visualizing spatial expression of isoforms at the protein level by immunohistochemistry is limited by the availability of antibodies specific to the isoforms of interest and the labor-intensive, time-consuming nature of raising novel antibodies to specific isoforms. Although databases of antibodies have been described, considerable effort may still be required to determine which, if any available antibodies will distinguish between the isoforms of interest[150]. *In vivo* methods of visualizing alternate isoforms have been described for model organisms such as *C. elegans*[151] and mouse[152].

Functional characterization of particular isoforms can be performed in a number of ways. Many studies have attempted to infer the function of isoforms by observing differences in expression level, subcellular localization, post-translational modifications and other modifications in cells where the gene of interest is thought to play some role[153, 154]. Examples of direct manipulation of the expression of an isoform are less common. In principle an RNA interference based approach should be able to specifically 'knock down' an isoform of interest in cell culture and considerable resources exist to facilitate these kinds of experiments[155]. Resources to facilitate over-expression of specific isoforms by transfection of open reading frame containing expression vectors into suitable cell lines have also been reported[156, 157]. Creation of transgenic mice expressing a particular isoform has also been widely reported[158, 159]. Altering expression of an isoform can be used in conjunction with studies of particular functions of interest such as apoptosis or cell survival assays. Differences in the protein-protein interactions of alternate isoforms can be studied by methods such as co-immunoprecipitation of expected partners or immunoprecipitation of tagged isoforms followed by HPLC-MS to identify interacting partners[160]. Studying multiple isoforms in these kinds of experiments, although more labor intensive, will become increasingly common as researchers become aware of the transcriptional diversity generated by genes of interest.

1.5. Functional significance of alternative expression

As large scale experimental and bioinformatic approaches have begun to identify the diversity of gene expression across the genomes of several species, parallel efforts to study the functional significance of this diversity have also been reported. One area of intense debate has been the effort to estimate the proportion of AE events that are functional compared to that which represents 'transcription noise'. Other areas which have generated several publications include the effort to identify general themes by which AE influences cellular biology, the study of particular functional classes of genes that are affected by AE and its potential role as a means of globally regulating gene expression. Finally, the implications of AE for the study of human disease has received increasing attention in recent years. For example, the emergence of an 'alternative expression code' has implications for the identification of potential disease mutations; increased knowledge of transcriptome complexity will influence strategies for identifying therapeutic targets; and the mechanisms of RNA processing itself are being considered as a means of directly modulating disease states. The functional significance of alternative expression is discussed in detail in the following sections.

1.5.1. How much alternative expression is functional?

The percentage of alternative transcripts with biologically relevant functions remains a topic of debate. Detailed studies of single genes or pathways have identified differing functions for alternate isoforms. Although these single gene studies hint at the mechanisms by which AE allows a diversity of functions to be encoded from a single locus, they do not confirm the role of AE as a global means of generating biologically relevant diversity in the proteome. To address this outstanding question, a number of studies have attempted to use conservation of AE events between species to infer the fraction of all events that are functionally significant as opposed to transcription 'noise' caused by random splicing errors or observations of immature transcripts derived from the nucleus. The resulting estimates for the percentage of alternative events represented in EST data that are conserved between human and mouse range from 11 to 61%. To estimate the subset of alternatively spliced exons that are functional, one group used ESTs to identify exon skipping events which occur in both humans and mouse[38]. Of a total of 980 exons identified as alternatively skipped in humans, 25% were also skipped in mouse. The characteristics of the conserved subset of alternate

exons were found to be distinct from those of the non-conserved exons and suggested that the majority of non-conserved events are non-functional. Another study observed AS events in 2,603 human genes and their mouse orthologs[62]. The authors found that of all the orthologous exons that are alternatively spliced in human or mouse, 16% are alternatively spliced in both species, and the remaining 84% represent species-specific events. By considering events represented in multiple transcripts from multiple tissues for both human and mouse, the authors estimated that at least 24% of these events represent true examples of species-specific AS. Thanaraj *et al*.[161] argued that studies which utilize EST data will underestimate the conservation of AS between mouse and human because they rely heavily on the level of transcript coverage. In other words, conservation of a splicing event observed in human is often not observed in mouse simply because the EST sampling depth is too low and by chance it has not been observed. These authors conducted a conservation study similar to those previously described but also developed a statistical model to estimate the 'true' level of conservation by extrapolating from existing levels of transcript support. Using this model, they estimated that 61% of alternatively spliced junctions are conserved between mouse and human. In contrast, Yeo *et al*.[35] estimated that only 11% of the alternatively spliced exons in humans are conserved in mouse and suggested that the majority of AS events seen in EST/cDNA data represent aberrant splicing, disease-specific splicing or events that are functionally relevant but specific to humans. One theme that emerges from these works is the considerable disagreement in the literature as to what percentage of AE is truly conserved and indeed what percentage of non-conserved events might be functional but species-specific events that emerged since the divergence of human and mouse 85 million years ago. AE events that are not conserved between human and mouse tend to be expressed at lower levels and may serve as an evolutionary mechanism for testing novel proteins without disrupting the function of the canonical isoform and interfering with the normal functions of the cell[62, 162]. The 'lesser' form is thus unlikely to be detrimental, is relatively free of constraints, can evolve rapidly and in some cases gain a function that is driven by positive selective pressure. It has been suggested that incorporation of novel exons or boundaries in this way represents a major form of gene evolution which is distinct from evolution by gene duplication. This hypothesis is based on the observation that genes which are part of

gene families that have arisen by duplication generally have few alternate transcripts, whereas 'singleton' genes have high rates of AE[163, 164].

It is reasonable to assume that most deeply conserved AE events are functional, that some as yet unknown fraction of non-conserved events are also functional and the remaining fraction are not functional. Although the percentage of events falling into each of these categories remains an area of active debate, any study of AE will certainly be complicated by some level of expression 'noise' with unknown functional relevance.

1.5.2. How does alternative expression influence the proteome?

The number of AE events that result in a protein with a modified biological function is currently a topic of debate. The concept that this subset of AE events could increase the functional diversity of the human genome by generating a combinatorial output of proteins from a genome of perhaps less than 30,000 genes has gained acceptance in recent years[9, 11, 12]. AE of specific genes has been shown to regulate transcript abundance via nonsense mediated decay, alter the subcellular localization of proteins, influence enzymatic activity, modify protein stability, and alter posttranslational modifications (Reviewed in [165]). One of the most striking examples of AE producing diverse products from a single gene locus was observed for the *Drosophila melanogaster DSCAM* gene[166]. When transcribed, this gene's exons are selected from a set of mutually exclusive alternate exons at four positions. Specifically, exons 4, 6, 9 and 17 in each transcript are selected from 12, 48, 33, and 2 possible alternatives respectively. This remarkable arrangement is capable of producing 38,016 possible unique *DSCAM* transcripts. Cloning and sequencing a sample of 50 random cDNAs for this gene yielded 49 unique transcripts which result in distinct proteins with differing abilities to form neuronal connections. A comparably dramatic level of diversity was recently described for the human *basonuclin 2* (*BNC2*) locus, a zinc finger protein which is expressed ubiquitously and thought to function in RNA processing[167]. All 23 exons of this gene are alternatively used and each transcript independently uses one of six promoters and four polyadenylation sites. To date more than 100 distinct *BN2* mRNA isoforms have been produced, but a staggering ~90,000 are possible.

AE may result in the production of protein isoforms that are functionally distinct in a number of ways. It has been suggested that this diversity is realized in part through alterations in protein-protein interactions. Specific examples of genes such as *SMRT*

which produces isoforms differing in their interaction with thyroid hormone receptors have been studied in detail[168]. Furthermore, global analysis of EST data has shown that AE events disproportionately affect domains involved in protein-protein interactions[97]. Although only 10% of AS events can be shown to completely remove or insert a known functional domain, Wang et al.[169] found that many of the remaining 90% of AS events are predicted to effect loop structures in proteins which are thought to mediate protein-protein interactions. Yura et al.[170] also found that the majority of changes observed in isoforms do not affect complete protein domains and based on an analysis of the 3D structures of alternative isoforms concluded that AE modulates the activity of protein networks and associated signaling pathways indirectly by altering the structural core and resulting stability of proteins. For example, replacing a stable domain with an unstable domain in a protein could alter the spatial orientation of other domains resulting in a protein with a distinct conformation and affinity for interaction partners. These observations have led to the general speculation that AE outcomes profoundly influence the protein interaction network of a cell. Supporting this hypothesis is the observation that genes with large numbers of isoforms tend to have many interactions and represent central nodes in protein-protein interaction networks (Hughes and Friedman, 2005). In addition to modifying protein interactions, another common effect of AE is the modification of subcellular localization in which alternative isoforms differ in their signal peptides and/or transmembrane domains[171, 172]. Such modifications can result in post-translational transport to different cellular compartments or the production of a soluble protein rather than a membrane bound one.

As discussed, AE can presumably influence protein interactions, protein stability and subcellular localization and through each of these types of effects has the potential to influence signaling pathways. These observations suggest some of the general modes by which AE influences the function of any protein. Efforts to identify whether genes of particular functional classes are more likely to be modulated by AE have also been reported. For example, Takeda et al.[87] used a comprehensive analysis of 55,000 cDNAs to determine that the gene classes (according to Gene Ontology terms) which are most affected by AE are: nucleic acid binding, transcription factor activity, DNA-binding, protein tyrosine kinase activity, transporter activity, zinc ion binding, insulin-like growth factor-binding, ATP binding, catalytic activity, and oxidoreductase activity. Analysis of cDNA, EST and MPSS data in mouse found that 75% of all kinases and

phosphatases have alternate isoforms and analysis of these variants revealed several tethered, soluble, and secreted isoforms which were predicted to be catalytically inactive and therefore might act as dominant negative forms by competing with other isoforms for ligands and substrates[173]. Similar studies have demonstrated the prevalence of functional isoforms within the G protein coupled receptor family[174], zinc-finger-containing proteins[175] and apoptosis genes[176].

Finally, it is important to note that production of a transcript variant which does not seem to produce a functionally distinct protein may still have functional consequences for the cell by altering the level of gene expression. For example, AS is speculated to act as a gene expression 'switch' whereby genes are effectively turned off by changes in the expression of a splicing factor which disrupts their normal splicing and silences their expression by triggering nonsense mediated decay (NMD). NMD targets transcripts with premature termination codons, which are recognized by the transcription machinery and degraded rather than producing a potentially detrimental protein product. In this system, transcription of a gene may still occur at the same rate but since the mRNA products are quickly degraded the gene's function is essentially silenced. Recent studies have suggested that coupling of NMD and AS is an important but overlooked mechanism of regulating gene expression[177-179]. In addition to NMD, which is triggered by events within the coding region of a transcript, AE within UTRs may also act as a global means of controlling gene expression by altering mRNA stability and translational efficiency in a tissue specific manner[180]. In this case a valid mRNA is produced and would seem to result in production of a normal protein but due to sequence modifications outside the coding region, the stability of the transcript or its rate of translation is modified.

1.5.3. Implications of alternative expression for the study of disease

The role of AE in human disease has received increasing attention in recent years[181-184]. In particular, the apparent existence of a defined 'expression code' has implications for the identification of potential disease-causing genomic variants (e.g. point mutations, insertions, deletions). This code can be considered as the combination of (1) regulatory sequence motifs of a transcribed region and (2) RNA and protein factors which comprise the machinery responsible for correct transcription initiation, splicing and polyadenylation. Genetic changes that have the potential to alter normal expression

and contribute to human disease can thus be classified into two groups, 'cis-acting' variants which affect sequence motifs within each gene locus and 'trans-acting' variants which affect components of the transcriptional machinery itself. Examples of human disease involving both of these classes of variants have been reviewed in the context of neurological disorders and cancer[25, 185].

Disease associated transcripts may arise by the occurrence of cis-acting mutations within the expression regulatory elements of a single gene (**Figure 1.1**) and many examples of heritable diseases have been shown to result from point mutations leading to aberrant splicing of a gene. Such mutations may result in aberrant skipping of a canonical isoform, inclusion of a 'cryptic' exon that is not normally used or simply an alteration of the ratio of alternative isoforms normally expressed[186]. According to the Human Gene Mutation Database, ~10% of all disease associated mutations involve splice sites[187]. In addition to splice site mutations, many other mutations may affect splicing regulatory sequences such as exonic and intronic splicing enhancers and silencers[188]. For example, analysis of the effects of mutations in the well studied human disease genes ATM (ataxia-telengiectasia, OMIM #208900) and NF1 (Neurofibromatosis type I, OMIM #162200) suggests that as many as 50% of all exonic mutations, silent or otherwise, exert their influence by causing splicing defects[189, 190]. Many of these mutations are at splicing regulatory sites, not the actual splice sites. Until recently the only mutations associated with disease that were predicted to affect splicing of a gene product were those associated with the splice acceptor and donor sites specifically. Increasing knowledge of the additional motifs which influence AE has expanded the number of mutations which are predicted to affect transcription. Many non-synonymous mutations may have a more pronounced effect than causing a single amino acid change by influencing the inclusion or exclusion of entire exons. Similarly, many synonymous mutations or mutations outside of the coding sequence may influence exon content. A number of studies have recently begun to investigate the effects of mutations in known disease genes at positions other than the actual splice sites and preliminary attempts to predict and validate the effect of point mutations on AS in splicing regulatory motifs such as exonic splicing enhancers (ESEs) have been reported[191-194]. Some of these studies rely on the observation of mutations and their effect on the splicing of specific genes[191]. Others attempt to computationally predict the effect of mutations occurring within exons or introns on the splicing outcome of a

gene[192-194]. Similar efforts are needed to understand the true implication of mutations on the use of alternate transcription initiation sites and polyadenylation sites. In other words, although it has been accepted that polymorphisms or mutations affecting 'regulatory' sequences may affect the tissue- or developmental-specific expression level of a gene, it is now becoming clear that an entirely additional set of 'regulatory' changes act by influencing AE without necessarily changing the level of expression.

Reports documenting disease associated mutations that occur in trans-acting factors of the splicing machinery and that result in the aberrant processing of several genes are less common than those involving cis-acting mutations but a few examples are well documented. Two forms of the familial disease Retinitis pigmentosa, RP18 and RP13 are caused by mutations in precursor mRNA processing factors 3 and 8 respectively (OMIM #601414 and #600059). For some diseases associated with aberrant splicing such as certain cancers, it is often not known whether a cancer-associated AE event arises because of acquired or inherited mutations in cis-acting transcription regulatory motifs[195] or changes in the expression of trans-acting splicing factors[196, 197]. However, in some cancers such as chronic myeloid leukemia (CML) the evidence for involvement of splicing factors is becoming more convincing. The Bcr-Abl fusion product of CML has been shown to cause changes in the expression of genes involved in pre-mRNA splicing, resulting in the aberrant splicing of a cascade of other genes which in turn contributes to pathogenesis[198]. Bcr-Abl dependent over-expression of the splicing gene *SR Protein Kinase 1* (*SRPK1*) was observed in *CD34*+ blood cells and this over-expression was associated with aberrant splicing of apoptosis and differentiation genes such as *Pyk2*, *SLP65*, *BTK* and *Ikaros*. Both the expression of *Bcr-Abl* and the aberrant splicing of *Pyk2* were partially reversed by treatment with the kinase inhibitor STI571 (Imatinib/Gleevec®).

Regardless of whether the effect is via a cis- or trans-acting effect, the general potential for splice variants to act as diagnostic or prognostic markers or novel therapeutic targets for complex diseases such as cancer seems promising[199]. The observation that the genome is capable of producing a dramatic diversity of products from a relatively small number of loci has already begun to influence strategies for identifying therapeutic targets. For example, a number of studies have used bioinformatic approaches to identify cancer-specific splice variants by analyzing the content of human EST, SAGE and microarray repositories[200-203]. Increasing the

resolution of gene expression screens for therapeutic targets to profile individual exons and AE events has the potential to identify previously unobserved and potentially more definitive events specific to disease states. While 'classic' differential gene expression studies have resulted in useful observations, recent studies have found that differential expression of isoforms may be more prevalent in tissue comparisons than whole gene differential expression[26] and these two groups have limited overlap[84]. Thus, the application of alternative expression microarrays and deep sequencing platforms to the comparison of normal versus diseased tissues, drug responders versus non-responders and other relevant comparisons seems certain to yield novel biomarkers which would have been previously impractical to detect. Since AE can create functionally significant variants, searching for these variants in target discovery efforts should result in the identification of distinct protein isoforms associated with disease which may be more useful targets than proteins that are simply up- or down-regulated in disease. For example, the *Bcl-x* gene is alternatively spliced to form a long isoform which is anti-apoptotic (*Bcl-x*$_L$) and a short isoform which is pro-apoptotic (*Bcl-x*$_S$) and targeting this locus by inactivating one isoform or simply shifting the ratio of isoforms has been proposed as a cancer treatment[57]. Many targets may have evaded detection in previous gene-expression studies of disease because of a technological inability to profile this kind of transcript diversity from each locus. The identification of targets for the development of small molecule drugs and therapeutic antibodies[204] will thus be greatly enhanced by considering alternate isoforms and their subtle differences in amino acid content. In addition to the identification of drug targets, AE also has implications for pharmacogenomics and there is evidence that polymorphisms which alter splicing may underlie differences in drug efficacy and toxicity between patients. For example, the most common polymorphism of *CYP2D6*, a gene which is responsible for the metabolism of at least 40 drugs, results in the aberrant splicing and production of a non-functional protein from this gene[205].

Targeting specific isoforms with small molecule or antibody therapies is a logical extension of current drug design efforts but targeting the transcriptional machinery itself has also been proposed as a means of altering gene expression and treating disease. Proof-of-principle experiments describing the screening of drugs that target splicing factors such as SR-proteins to inhibit aberrant splicing or produce a desired splicing outcome have been reported[206]. Antisense oligonucleotide therapies to directly

manipulate the splicing patterns of specific disease genes have also been described[207]. These molecules can be used to influence splicing in many ways such as preventing the inclusion of an aberrant exon by masking a cryptic splice site, or forcing an exon-skipping event to allow nonsense or frameshift mutations to be by-passed[208]. Current studies have only begun to address the ways in which an understanding of AE can influence the study of human disease by enhancing the identification of therapeutic targets, allowing the design of novel types of therapies and predicting the efficacy and toxicity of drugs for individual patients.

1.6. Cancer

Cancer is a disease in which cells of essentially any tissue become unregulated in their cell division and gain the ability to invade other tissues. The hallmarks of a population of cells representing a cancer include: self-sufficiency in growth signals, insensitivity to anti-growth signals, unlimited potential for replication, the ability to avoid programmed cell death (apoptosis), sustained angiogenesis, and the ability to invade neighboring tissues and metastasize to remote locations[209].

In the late 1990's, cancer overtook heart disease as the number one cause of death in Canada, with 1 of every 4 Canadians dying of the disease (Statistics Canada; 1997). While overall survival has increased and mortality has decreased for most cancers in the past 30 years, it remains a significant cause of morbidity and mortality and therefore an active area of health research. As the occurrence and morbidity associated with cancer increases in proportion relative to other diseases, this has important implications for cost-control and places great emphasis on the need to improve the efficacy of expensive cancer therapies. Improving cancer outcomes is a multifaceted problem but can broadly be summarized as encompassing: prevention, screening to detect cancers early, and treatment involving some combination of surgery, radiation therapy and chemotherapy. While each of these are active areas of research, the development and optimization of cytotoxic chemotherapies and more recently novel biologic agents are most relevant to this thesis. A concerted application of modern genetics, molecular biology, genomics, informatics and other disciplines has lead to the identification of novel molecular targets allowing rational design of biologic agents able to attack cancer cells with an unprecedented level of specificity (see **Figure 1.4** for example).

1.6.1. Colorectal cancer

Colorectal cancer (CRC) is the fourth most common cancer diagnosis but the second most common cause of cancer related death (National Cancer Institute of Canada, 2009). Screening has proven useful in detecting pre-cancerous polyps that can be surgically removed before malignant CRC occurs. If a patient presents with CRC, treatment also typically begins with surgical resection. Surgery alone is often successful in curing patients with stage I and II disease (60-95% five year survival)[210]. Stage I and II CRC is characterized by small minimally invasive to large invasive tumours (T1-4) with no nodal involvement (N0) and no metastasis (M0). Unfortunately, 58% of patients diagnosed with CRC have stage III disease characterized by nodal involvement (T1-4, N1-3, M0) or stage IV characterized by distant metastasis, often to liver (T1-4, NX, M1). For these patients with advanced CRC, the prognosis is comparatively poor (25-60% for stage III and <5% for stage IV)[210]. These patients therefore represent the primary potential beneficiaries of improved therapies. Currently, surgery, including resection of liver metastases is still a common strategy for these patients and is often coupled with adjuvant chemotherapy or chemoradiation. The first drug to be used widely in CRC was fluorouracil (5-FU), a cytotoxic nucleotide analog that results in RNA and DNA damage that triggers apoptosis (discussed in detail in **Chapter 4**). For both chemotherapy and chemoradiation, 5-FU remains the core drug of choice. For example, in my own survey of patient records for 279 CRC cases from the BC Cancer Agency and Ontario Tumour Bank, 223 (80%) received 5-FU, in a neo-adjuvant, adjuvant or palliative context (but mostly adjuvant). The introduction of this drug was successful in doubling the median survival of patients with advanced CRC. During the 1980's, 1990's and 2000's 5-FU biomodulation (5-FU + leucovorin), Irinotecan (a topoisomerase inhibitor), Oxaliplatin (the platinum based DNA cross-linking), Bevacizumab (*VEGF* antibody) and Cetuximab and Panitumumab (*EGFR* antibodies) were added to the oncologist's toolbox. These additions allowed a further doubling of median overall survival[210]. While an overall quadrupling of median survival is certainly encouraging, this corresponds to an improvement in median survival from ~5 months to ~20 months for patients with advanced CRC. In other words, there is considerable need for further research to improve existing therapeutic strategies and develop novel treatments.

1.6.2. Chemotherapy resistance

Chemotherapy resistance is a common challenge in the treatment of many cancers. Resistance simply refers to the failure of a drug to halt tumour growth or kill tumour cells. This resistance may correspond to a sub-population of tumour cells present before treatment (intrinsic resistance) or more commonly arises after exposure to the drug (acquired resistance)[211]. Several mechanisms of drug resistance have been proposed. These mechanisms may apply to all drugs, a class of drugs (e.g. nucleotide antagonists), or a specific drug. In order for any drug to be effective, the drug must be successfully delivered to the tumour site, enter tumour cells, remain extant for a sufficient period of time to be effective and arrest cell growth or preferably induce cell death. Many drugs also need to be converted from a pro-drug to an anti-tumour form. Furthermore, many drugs are toxic to all cells and capable of significant side effects which vary by individual according to the effective dose delivered and drug half-life. A major mechanism that influences both resistance and toxicity is metabolism. A tumour may be resistant simply by virtue of an overly active metabolism of the drug that rapidly removes or inactivates the drug. For example, the cytochrome p450 family of enzymes are involved in the catabolism and clearance of many drugs[211] (often in the liver). Assuming a drug does reach tumour cells, the next requirement is to cross the cellular membrane. Some drugs may enter the cell passively while others utilize active mechanisms or a combination of both. For example, nucleotide analogs may utilize nucleotide transporters to enter the cell and reduced expression or mutation of these genes may confer resistance to the drug. Once inside the cell, a drug is subject to numerous drug efflux pumps. Increased expression of genes such as *MDR1* (*P-gp*) a member of the ABC transporter family can confer multi-drug resistance by pumping out a remarkable array of substrates. Other members of the ABC family are more specific in their recognition of substrates but nevertheless resistance to most if not all drugs can be conferred by abundant expression of one or more of these genes. Assuming the equilibrium between drug entry and efflux allows some exposure of the cell to a drug, the next requirement is often activation of the drug from an inactive- to active- form. For example, drugs belonging to the purine/pyrimidine analog class are converted by nucleotide metabolism enzymes to active anti-tumour metabolites. Reduced expression or mutation of these genes can therefore confer resistance. Assuming entry and activation of sufficient quantities of a drug, the next requirement for many drugs is that

they interact successfully with a target gene product, often an enzyme. Drugs which target an enzyme required by the cell for survival can be rendered ineffective by mutations in the target enzyme that retain enzyme function while preventing binding of the drug thereby conferring drug resistance. In contrast to enzyme targeting drugs, many chemotherapies function by causing DNA damage. Resistance to these drugs may involve modifications to DNA repair pathways. For example, both enhanced DNA repair and increased tolerance of DNA damage by the cell's surveillance systems can confer resistance to these types of drugs. Finally, since the ultimate goal of all chemotherapies is to induce cell death, defects in apoptosis pathways represent a major potential drug resistance mechanism.

Considerable research has begun to elucidate the mechanisms of resistance but this problem remains an active area of study. In particular, determining the dominant mechanisms and associated genes that confer resistance to specific drugs in specific cancers remains a daunting task. Cases where this information has been used successfully to overcome drug resistance are rare but with continued effort remain a promising avenue for future improvements to cancer treatment.

1.7. Thesis objectives and chapter summaries

The human transcriptome is complex and understanding this complexity is critical to our understanding of molecular biology and the application of genomics to improve treatment of human disease. The general aim of this thesis was to develop new computational methods to interpret massive transcriptome profiling datasets. I focused on developing methods that facilitate the identification of novel mRNA isoforms and changes in the expression level of those isoforms associated with cancer progression. Based on the analysis of full-length cDNA sequencing reported in the literature[98, 99, 104] that identified multiple distinct isoforms for many genes, I hypothesized that the current estimate of transcript diversity at the typical gene locus was an underestimate. While yielding high quality data, the methods described in these reports were too expensive and time consuming to apply to the entire genome. To address this challenge I took advantage of emerging developments in microarray and massively parallel sequencing technology to develop novel methods for alternative expression analysis and apply these to a study of 5-FU resistant colorectal cancer. Using these methods my hope was to allow the simultaneous examination of the alternative mRNA isoforms from

thousands or tens of thousands of gene loci. A brief summary of the methods I developed to achieve this goal (**Chapter 2** & **3**) and an example of their application to a model of drug resistance in cancer (**Chapter 4**) is provided below.

The main hypotheses of this thesis were as follows: (1) the number of alternative isoforms represented in current transcriptome annotation efforts (e.g. EnsEMBL) is greatly underestimated, (2) technological advancements in both microarrays and massively parallel sequencing will allow reliable identification and quantification of both known and novel isoforms, and (3) the transition from 5-FU sensitivity to resistance will be associated with differential expression of entire genes as well as specific isoforms for which the overall total change in gene expression might be negligible. In **Chapter 4**, I describe a detailed characterization of one such event that was identified as a top 5-FU resistance candidate in both **Chapter 2** and **3**. Specifically, both ALEXA-array and ALEXA-Seq analysis identified abundant expression of a novel exon-skipping isoform of the gene, *uridine monophosphate synthetase* (*UMPS*) in 5-FU resistant cell lines. A concomitant decrease in the abundance of full-length *UMPS* was also observed in resistant cells. Upon review of a 5-FU pathway from the Pharmacogenomics Knowledge Base (http://www.pharmgkb.org/), I noted that this gene is involved in the conversion of 5-FU from an inactive form to active anti-tumour metabolites. By sequencing of the entire genomic region of *UMPS* near the exon skipping event I determined that the likely cause for the expression of the novel isoform is a heterozygous splice site mutation acquired in 5-FU resistant cells. I then went on to generate 96 full-length sequenced *UMPS* cDNA clones, representing the canonical isoform, the novel isoform identified in **Chapters 2** and **3**, and eight additional isoforms. These isoforms were characterized using bioinformatic techniques and the two most abundant isoforms were profiled in a panel of additional cell lines and colorectal cancer patient samples. Additional mutation analysis was also performed. *UMPS* was found to be recurrently mutated, aberrantly spliced, or under-expressed in 5-FU resistant colorectal cancer cell lines.

In **Chapter 2**, I describe the development of a tool to allow the generation of custom microarrays capable of detecting and measuring the level of specific alternative isoforms without necessarily knowing their identity in advance. This tool, called ‘ALEXA-array’ (ALternative EXpression Analysis by microarrays) used existing transcriptome and genome resources to extract probe sequences corresponding to the

individual exons, exon-exon junctions, exon boundaries, and introns of every human gene. Probes were filtered and scored according to melting Tm, folding potential, sequence specificity and so on. Using the ALEXA-array tool, I created a prototype array and used it to test a series of hybridization conditions using RNA isolated from brain tissue and a single cell line (hybridizations were performed by NimbleGen Inc.). Based on an analysis of these data, I identified the best hybridization parameters offered by NimbleGen and proceeded to create a second custom array to compare 5-FU sensitive and resistant colorectal cancer cell lines. The selection of genes for this array was based on preliminary analysis of Affymetrix exon array data generated for the same cell lines. Extensive validation of this experiment was conducted by comparison of Affymetrix exon arrays and predictions from publicly available expressed sequence databases. I identified a set of candidate alternative mRNA isoforms whose expression level was significantly altered between 5-FU sensitive and resistance cells. I also made the ALEXA tool and related databases publicly available by creating a website (www.AlexaPlatform.org).

In **Chapter 3**, I describe the creation of a method that is conceptually similar to that in **Chapter 2.** While building on my experience with microarrays, the analysis required a novel implementation to accommodate a different data type. Specifically, instead of relying on microarray signal intensities corresponding to hybridization spots on a custom designed array, it relied on randomly sampled reads generated from a fragmented cDNA library. I addressed the challenge of trying to synthesize these massive sequence data sets by developing a pipeline called 'ALEXA-Seq' to obtain expression and differential expression information for specific mRNA isoforms. To assess the output of my pipeline I conducted extensive comparisons to both Affymetrix exon array data, the ALEXA-array data presented in **Chapter 2**, and publicly available expressed sequence and conservation data. I also created novel metrics for identifying interesting alternative expression events and tools for visualizing and interpreting these findings. As in **Chapter 2**, I generated a list of candidate isoforms that were differentially or alternatively expressed between 5-FU sensitive and resistant cells. I made the sequence analysis pipeline, complete analysis and visualization tools available on my website (www.AlexaPlatform.org).

In addition to the work described in this thesis I have been involved in several collaborative projects at the Genome Sciences Centre (GSC) which have resulted in

publications or accepted manuscripts. I modified my microarray design platform (**Chapter 2**) to create a custom array for copy number variant analysis which is currently being applied to the analysis of ~100 mental retardation trios (affected child, unaffected parents) in collaboration with Dr. Jan Freidman at the BC Children & Women's Hospital. More recently, I modified this platform to produce a custom microarray for target array capture experiments similar to those described in [212]. My work at the GSC was initially focused on analysis of full-ORF cDNA sequences generated for targeted reference mRNA sequences as part of the Mammalian Gene Collection effort[98, 104, 213]. This work was partially responsible for my interest in transcript diversity and influential in the development of the sequence based method for transcriptome analysis I describe in **Chapter 3**. The 'ALEXA' database which forms the bases for my microarray design platform was also used for bioinformatic analysis in microRNA profiling studies performed by Ryan Morin[214] and Florian Kuchenbauer[215]. I assisted in the bioinformatic analysis of copy number variation, SAGE, and 454 gene expression datasets performed by Trevor Pugh[216], Asim Siddiqui[128], and Mathew Bainbridge[131] respectively. I contributed bioinformatic support for the integration of publicly available databases into the open-access regulatory annotation project 'ORegAnno'[217]. Finally I have collaborated with Dr. Sharlene Gill, Dr. David Owen, and Dr. Carl Brown in acquiring ~120 patient samples as an ongoing validation of the clinical utility of the findings of **Chapter 4**.

Figure 1.1. Gene expression (transcription and RNA processing)
Expression of typical protein-coding genes involves: gene transcription, pre-mRNA processing and polyadenylation. Each of these processes is regulated by components of the transcription machinery which recognize sequence motifs in the DNA template and pre-mRNA molecule. After pre-mRNA processing, mRNAs are exported to the cytoplasm where ribosomes translate them into proteins. Abbreviations: (UTR) untranslated region; (D) donor site; (A) acceptor site; (SS) splice site; (ESE) exonic splicing enhancer; (ESS) exonic splicing silencer; (ISE) intronic splicing enhancer; (ISS) intronic splicing silencer.

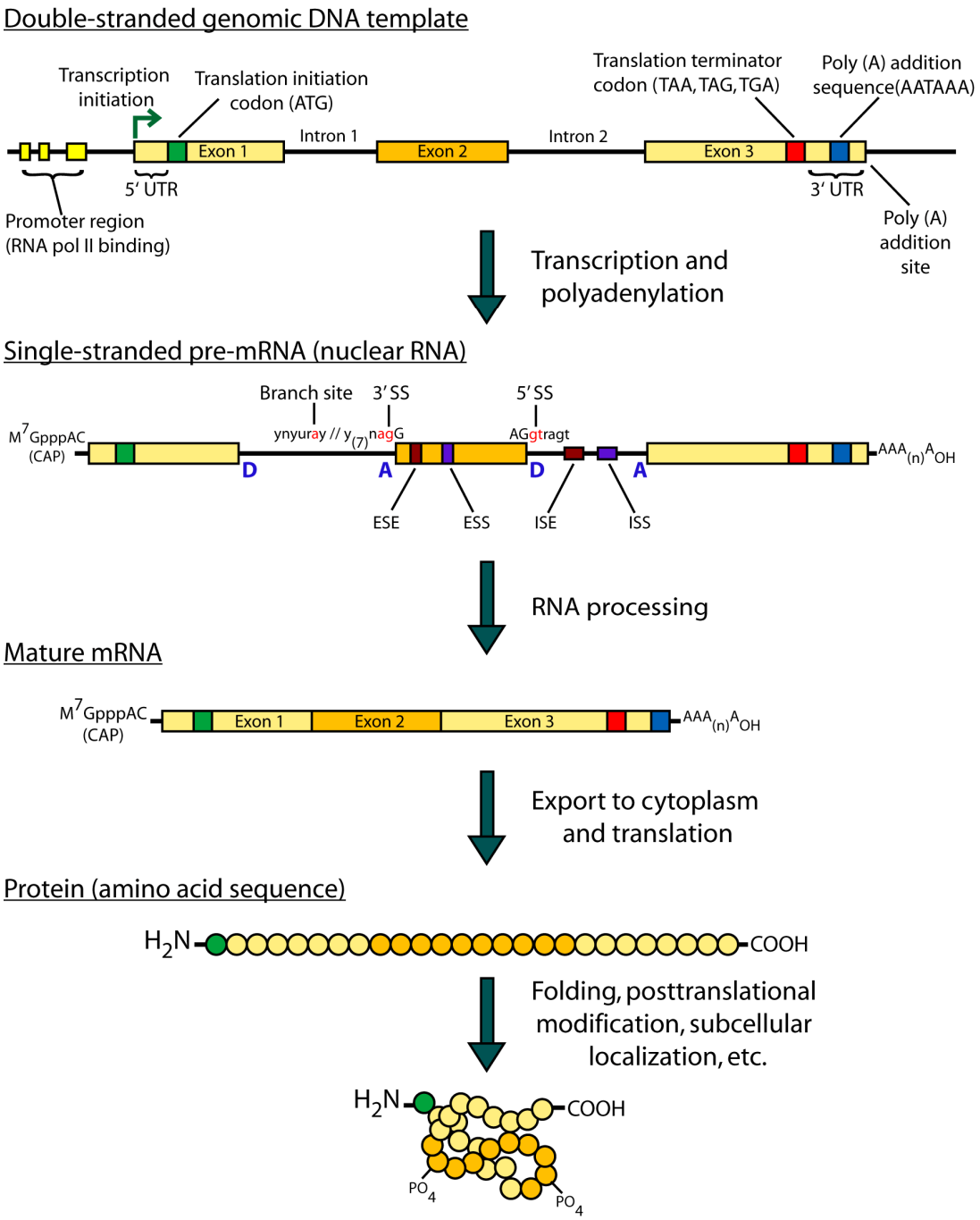


Figure 1.2. Types of alternative expression (AE)
Gene models are depicted as exons (colored rectangles) connected by introns (black lines). Green arrows indicate transcription initiation sites, dotted lines indicate splicing patterns and polyadenylation sites are denoted as 'poly (A)'. The mRNA products generated by each type of AE are shown to the right of each gene model. Simple transcription is contrasted with alternative transcript initiation, the five major classes of alternative splicing, and alternative polyadenylation. In each model, yellow exons are constitutive and blue exons are alternative.

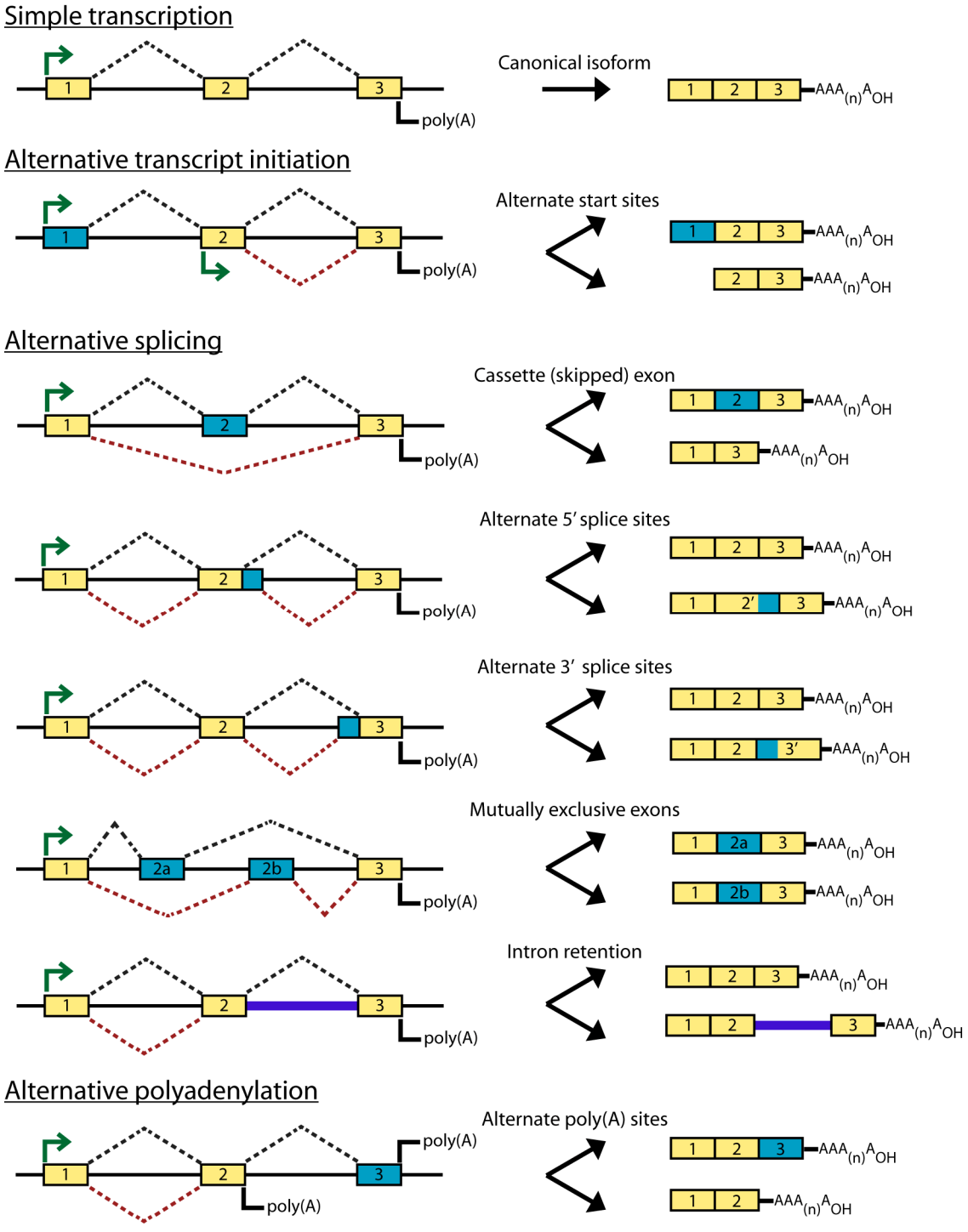


Figure 1.3. Splicing acceptor, donor and branch point sequences
'SeqLogos' showing DNA motifs for human splice acceptor, donor and branch sites [218].

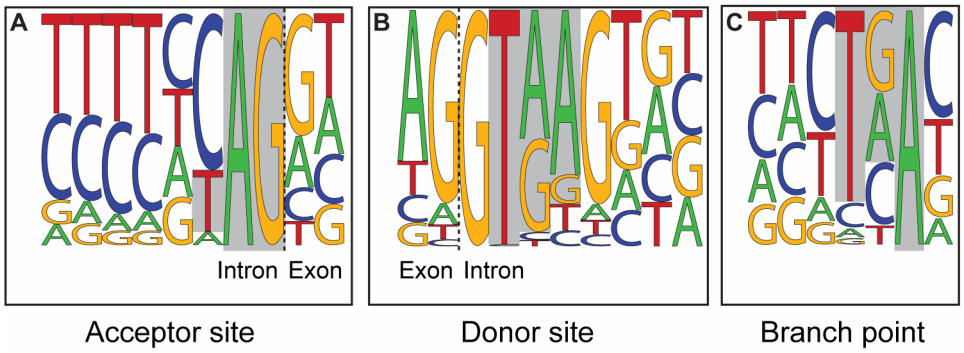


Figure 1.4. Identification of an alternative exon with application to cancer medicine
An alternative exon ('v6') of the hyaluronate receptor *CD44* (colored red) was discovered to be preferentially expressed in head and neck, breast and lung cancers[219]. Based on this observation, an antibody, 'bivatuzumab' was raised against the amino acid sequence encoded by the 'v6' exon. In order to test the efficacy of this antibody in treating cancer, it is was coupled to a radioactive isotope (^{186}Re-bivatuzumab) as well as to the cytotoxic agent known as maytansinoid mertansine or 'DM1' (bivatuzumab-mertansine). Both of these configurations exhibited promising anti-cancer effects in *CD44-v6* expressing tumours, although the latter was withdrawn due to skin toxicity in phase I clinical trials[220]. The image below shows a cartoon depiction of delivery of ^{186}Re-bivatuzumab to a patient with head and next cancer[221].

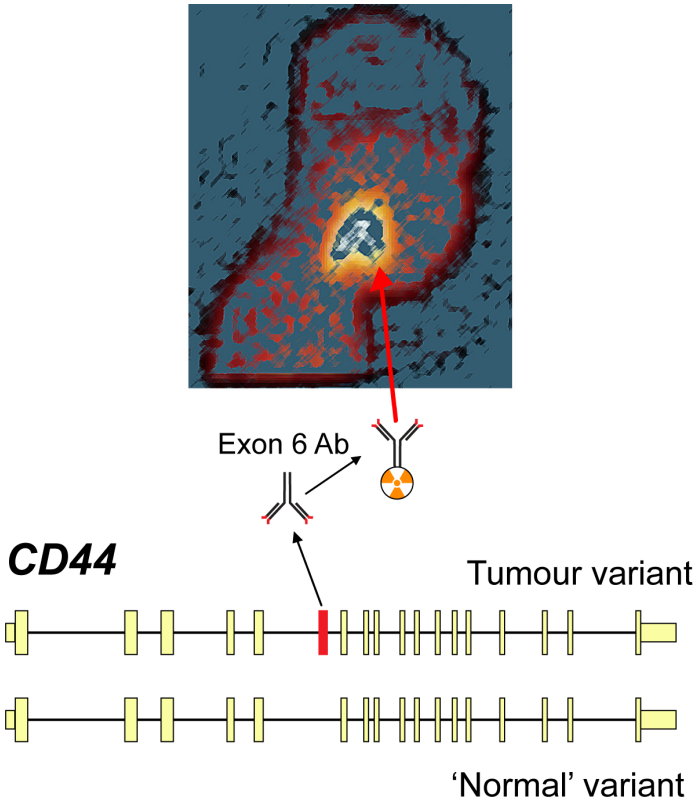


Figure 1.5. Microarray based method for profiling transcript diversity
Gene models are depicted as exons (colored rectangles) connected by introns (black lines). Hypothetical differences in mRNA products which can be detected by each array method are depicted to the right of each gene model. In each model, yellow exons are constitutive and blue exons are alternative. Differences in array design strategy, particularly the position and types of oligonucleotide probes used are shown above each gene model as colored horizontal lines.

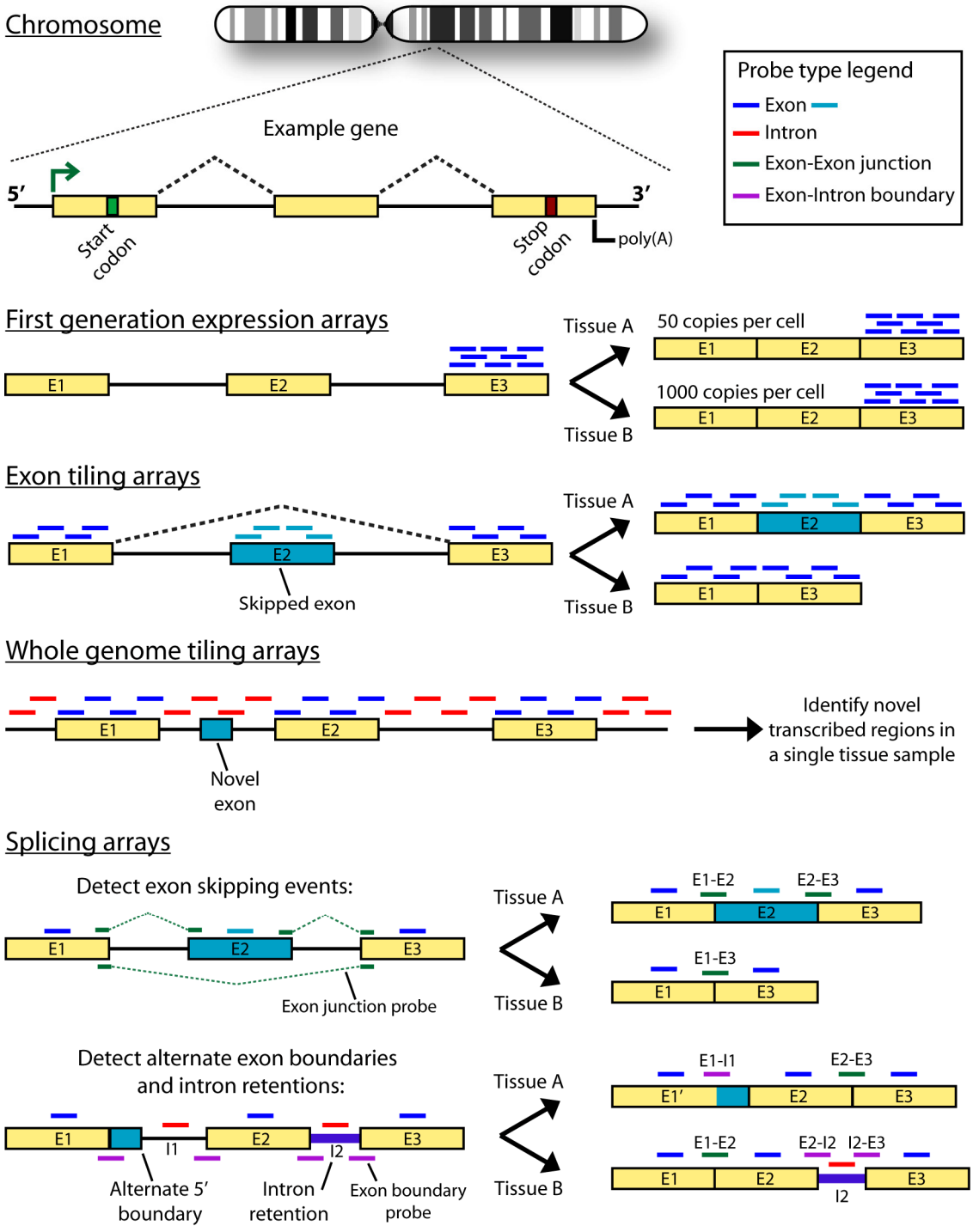


Figure 1.6. Sequence-based methods for profiling transcript diversity
Hypothetical transcript sequences consisting of exons (green rectangles) with intervening introns (black lines) are depicted as gapped alignments to a reference genome. The following tracks represent sequences generated by each sequence-based method. The methods are displayed in order of least to most quantitative. Abbreviations: (EST) expressed sequence tag; (SAGE) serial analysis of gene expression; (CAGE) capped analysis of gene expression; (GIS) gene identification signature.

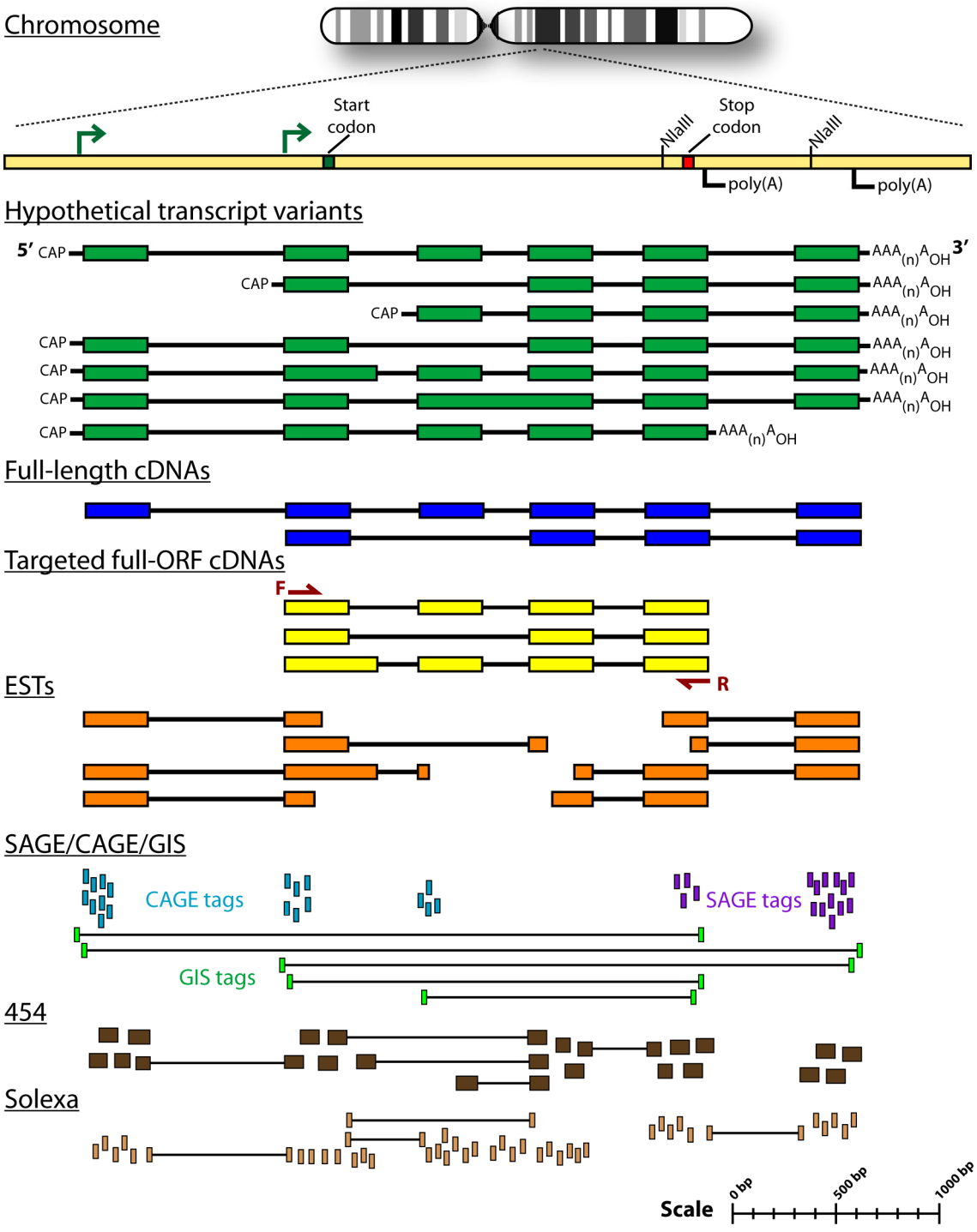


Table 1.1. Summary of methods for studying transcript diversity
Computational methods involve predicting transcription events from genomic sequence without using expression data. Microarray-based methods involve fluorescently labeling RNA and hybridizing it to an array of 'spots' each representing content from a reference genome. All array methods are subject to cross-hybridization between related sequences. Sequence-based methods involve generating expressed sequence data from RNA, aligning it to a reference genome and annotating transcription events. These methods do not rely on pre-existing gene annotations and they are capable of providing exon boundary/connectivity information as well as novel gene discovery.

Abbreviations: (ATI) alternative transcript initiation; (AS) alternative splicing; (AP) alternative polyadenylation; (SBS) sequence by synthesis; (†) limited applicability or supporting evidence.

Method	Events detected	Description (strengths/limitations)
Computational methods		
Ab initio	ATI, AS, AP	Predictions based on a single reference genome. Not quantitative. Low sensitivity/specificity compared to methods that use expression data.
Comparative genomic	ATI, AS, AP	Predictions rely on existence of suitable comparative genomes. Not quantitative. Medium sensitivity/specificity compared to methods that use expression data.
Microarray-based methods		
Spotted cDNA	None	Limited to composition of cDNA library. Not capable of distinguishing transcript variants. Low cost, high throughput. Quantitative.
3′ Expression	AS†, AP†	3′ end bias. Limited by pre-existing gene annotations. Low cost, high throughput. Quantitative.
Whole genome tiling	ATI†, AS†, AP†	Not limited by pre-existing gene annotations. Potential for gene discovery. High cost. Quantitative.
Exon tiling	ATI†, AS†, AP†	Limited by pre-existing gene annotations. Low cost, high throughput. Quantitative.
Splicing arrays	ATI, AS, AP	Limited by pre-existing gene annotations. Provides exon boundary/connectivity information. Medium cost, medium throughput. Quantitative.
Sequence-based methods		
EST cDNA	ATI†, AS†, AP	End bias. Partial transcripts (300-1000 bp reads). High cost, medium throughput. Limited quantitative value.
FL-cDNA	ATI, AS, AP	Complete transcripts. High cost, low throughput. Results in a physical copy of transcript. Not quantitative.

Method	Events detected	Description (strengths/limitations)
Targeted FL-cDNA	AS	Near complete transcripts. High cost, low throughput. Results in a physical copy of transcript. Not quantitative.
SAGE	AS[†], AP	3′ end bias. Short tags (17-21 bp). Medium cost, medium throughput. Quantitative.
CAGE	ATI	5′ end bias. Short tags (20 bp). Medium cost, medium throughput. Quantitative.
GIS	ATI, AP	End bias. Short tags (40 bp paired end tags). Medium cost, medium throughput. Quantitative.
Illumina/Solexa SBS	ATI, AS, AP	Short tags (25-150 bp). Low cost, high throughput. Quantitative.
Roche/454 SBS	ATI, AS, AP	Medium tags (~100-400 bp). Low cost, high throughput. Quantitative.
ABI/SOLID SBS	ATI, AS, AP	Short tags (~100 bp). Low cost, high throughput. Quantitative.

Table 1.2. Alternative expression resources
Abbreviations: (ATI) alternative transcript initiation; (AS) alternative splicing; (AP) alternative polyadenylation; (ESE) exonic splicing enhancer; (Hs) *Homo sapiens*; (Mm) *Mus musculus*; (Rn) *Rattus norvegicus*.

Resource name	Description (applicable species)	Ref.
Ab initio/de novo alternative transcript discovery/prediction/characterization		
ABySS	De novo assembly of short reads	222, 223
AStalavista	Automatic classification and naming of AS events	218
AUGUSTUS	Prediction of ATI, AS, and AP using only human genome sequence	43
EasyCluster	Assembly of gene models from transcriptome data	224
FindPeaks	Identify regions of expression directly from read density data	225
MARS	Human AS transcript prediction from pairwise alignments of mouse, rat, dog, opossum and frog genomes	36
Spliced alignment algorithms [226]		
BLAT, est2genome, Exonerate, SIM4, SPA, SPIDEY, SplicePredictor, Splign, QPALMA, TAP, WebGMAP	Identification of splice sites and gapped alignment of mRNAs to a reference genome	227-236
Databases of transcript diversity derived from EST/mRNA sequences		
AltTrans	Annotation and visualization of AS and AP (Hs, Mm)	18
AS-ALPS	Effect of AS on 3D structure of proteins (Hs, Mm)	237
ASAP II	Annotation and visualization of AS (15 species)	238
ASD	Annotation and visualization of AS (Hs, Mm)	239
AspAlt	Annotation and visualization of ATI, AS and AP (46 species)	240
ASPIC	Annotation and visualization of AS (Hs, Mm, +15 others)	241
ATID	Manual and computational annotation of ATI (Hs, Mm and 32 other species)	242
AVATAR	Annotation of splice sites supported by mRNA and EST data (Hs and 5 other species)	243
BIPASS	Annotation and visualization of AS (Hs and 3 other species)	244
DBTSS	Database of ATI (Hs, Mm, zebrafish, etc.)	13
ECgene	Functional annotation of AS (Hs, Mm, Rn, etc.)	245
G-Mo.R-Se	*De novo* annotation of genomes by analysis of short read sequences	137
H-DBAS	Database of ~40k cDNA clones of representative alternative splicing variants (Hs)	246
Hollywood	Annotation and visualization of AS (Hs, Mm)	247
LSAT	ATI, AS, and AP extracted from literature by text mining	248

Resource name	Description (applicable species)	Ref.
MAASE	Manual annotation of AS (Hs, Mm)	249
PolyA_DB	Annotation and visualization of AP (Hs, Mm)	250
SpliceCenter	Evaluate effect of AS on RT-PCR, RNAi, microarray and peptide studies (9 species)	251
SpliceInfo	Annotation and visualization of AS (Hs)	252
SpliceMiner	Database of splice variant data based on NCBI evidence viewer (Hs)	253
TISA	Annotation of tissue specific transcripts (Hs, Mm)	32
T-STAG	Annotation of tissue specific transcripts (Hs, Mm)	254
Alternative expression regulatory element prediction [255]		
ESEfinder	Identification of ESE sites and predicted effect of mutations within them	256
GRSDB	Identification of G-rich (GRS) processing motifs	257
RegRNA	Identification of transcription and splicing regulatory sequences within RNAs	258
RESCUE-ESE	ESE annotation tool (Hs, Mm, zebrafish, pufferfish)	259
Splicing Factor Finder	Identify splicing factor binding sites by combining splicing motif info with conservation data	260
Splicing Modeler	Uses microarray expression data from multiple tissues to predict splicing regulatory sequences	261
TassDB	Collection of tandem splice sites (human, mouse, etc.)	262
Alternative expression analysis		
ALEXA-array	Design and analysis of splicing alternative expression microarrays	263
easyExon	Processing and visualization of Affymetrix exon array data	264
FIRMAGene	RMA based analysis of Affymetrix exon array data	265
MiDAS	ANOVA based method of detecting alternative expression in microarray data	76
REMAS	Regression base method of identifying alternative expression in microarray data	266
SI-Limma	Identify differential exon splicing in microarray data	267
SPACE	Uses alternative expression array data to predict transcripts expressed in a sample (including novel isoforms)	268
Validation/Visualization Tools		
ASePCR	Electronic PCR utility for validation of alternate isoforms	269
ASGS	Web based tool for AS graphs	270
ASTRA	Visualization and classification of transcription patterns	271
VISTA, UCSC, EnsEMBL	Generic browsers for visualization of expression data and comparative genomics	147, 148, 272

References

1. Venter, J. C. et al. The sequence of the human genome. Science 291, 1304-51 (2001).
2. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).
3. Hubbard, T. J. et al. Ensembl 2009. Nucleic Acids Res 37, D690-7 (2009).
4. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72, 291-336 (2003).
5. Sharp, P. A. Split genes and RNA splicing. Cell 77, 805-15 (1994).
6. Kornblihtt, A. R. Promoter usage and alternative splicing. Curr Opin Cell Biol 17, 262-8 (2005).
7. Matlin, A. J., Clark, F. & Smith, C. W. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6, 386-98 (2005).
8. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. Nature 416, 499-506 (2002).
9. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103, 367-70 (2000).
10. Lareau, L. F., Green, R. E., Bhatnagar, R. S. & Brenner, S. E. The evolving roles of alternative splicing. Curr Opin Struct Biol 14, 273-82 (2004).
11. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418, 236-43 (2002).
12. Roberts, G. C. & Smith, C. W. Alternative splicing: combinatorial output from the genome. Curr Opin Chem Biol 6, 375-83 (2002).
13. Suzuki, Y., Yamashita, R., Sugano, S. & Nakai, K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res 32 Database issue, D78-81 (2004).
14. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413-5 (2008).
15. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-6 (2008).
16. Kalnina, Z., Zayakin, P., Silina, K. & Line, A. Alterations of pre-mRNA splicing in cancer. Genes Chromosomes Cancer 42, 342-57 (2005).
17. Berget, S. M. Exon recognition in vertebrate splicing. J Biol Chem 270, 2411-4 (1995).
18. Le Texier, V. et al. AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC Bioinformatics 7, 169 (2006).
19. Soller, M. Pre-messenger RNA processing and its regulation: a genomic perspective. Cell Mol Life Sci 63, 796-819 (2006).
20. Cooper, T. A. Use of minigene systems to dissect alternative splicing elements. Methods 37, 331-40 (2005).
21. Hicks, M. J., Lam, B. J. & Hertel, K. J. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. Methods 37, 306-13 (2005).
22. Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 37, 376-86 (2005).
23. Goldstrohm, A. C., Greenleaf, A. L. & Garcia-Blanco, M. A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. Gene 277, 31-47 (2001).
24. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res 30, 1083-90 (2002).

25. Srebrow, A. & Kornblihtt, A. R. The connection between splicing and cancer. J Cell Sci 119, 2635-41 (2006).
26. Kwan, T. et al. Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40, 225-31 (2008).
27. Huang, H. D., Horng, J. T., Lee, C. C. & Liu, B. J. ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. Genome Biol 4, R29 (2003).
28. Johnson, J. M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302, 2141-4 (2003).
29. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29, 2850-9 (2001).
30. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. Genome Biol 5, R74 (2004).
31. Watson, F. L. et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science 309, 1874-8 (2005).
32. Noh, S. J., Lee, K., Paik, H. & Hur, C. G. TISA: Tissue-specific Alternative Splicing in Human and Mouse Genes. DNA Res (2006).
33. Jones, S. J. Prediction of Genomic Functional Elements. Annu Rev Genomics Hum Genet (2005).
34. Sorek, R. et al. A non-EST-based method for exon-skipping prediction. Genome Res 14, 1617-23 (2004).
35. Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A 102, 2850-5 (2005).
36. Flicek, P. & Brent, M. R. Using several pair-wise informant sequences for de novo prediction of alternatively spliced transcripts. Genome Biol 7 Suppl 1, S8 1-9 (2006).
37. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34, 177-80 (2003).
38. Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 13, 1631-7 (2003).
39. Philipps, D. L., Park, J. W. & Graveley, B. R. A computational and experimental approach toward a priori identification of alternatively spliced exons. Rna 10, 1838-44 (2004).
40. Dror, G., Sorek, R. & Shamir, R. Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics 21, 897-901 (2005).
41. Cawley, S. L. & Pachter, L. HMM sampling and applications to gene finding and alternative splicing. Bioinformatics 19 Suppl 2, II36-II41 (2003).
42. Ohler, U., Shomron, N. & Burge, C. B. Recognition of unknown conserved alternatively spliced exons. PLoS Comput Biol 1, 113-22 (2005).
43. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34, W435-9 (2006).
44. Xia, H., Bi, J. & Li, Y. Identification of alternative 5'/3' splice sites based on the mechanism of splice site competition. Nucleic Acids Res (2006).
45. Redkar, R., Burzio, L., Haines, D. & Conzone, S. in Genes, Genomes and Genomics (eds. Thangadurai, D., Tang, W. & Pullaiah, T.) 1-39 (2006).
46. Hu, G. K. et al. Predicting splice variant from DNA chip expression data. Genome Res 11, 1237-45 (2001).

47. Fan, W., Khalid, N., Hallahan, A. R., Olson, J. M. & Zhao, L. P. A statistical method for predicting splice variants between two groups of samples using GeneChip expression array data. Theor Biol Med Model 3, 19 (2006).
48. Kampa, D. et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14, 331-42 (2004).
49. Kapranov, P. et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296, 916-9 (2002).
50. Schadt, E. E. et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol 5, R73 (2004).
51. Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149-54 (2005).
52. Bertone, P. et al. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242-6 (2004).
53. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21, 93-102 (2005).
54. Nuwaysir, E. F. et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12, 1749-55 (2002).
55. Lee, C. & Roy, M. Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol 5, 231 (2004).
56. Lyddy, J. ExonHit Therapeutics. Pharmacogenomics 3, 843-6 (2002).
57. Mangasarian, A. Alternative RNA splicing and drug target identification. IDrugs 8, 725-9 (2005).
58. Srinivasan, K. et al. Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods 37, 345-59 (2005).
59. Castle, J. et al. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol 4, R66 (2003).
60. Clark, T. A., Sugnet, C. W. & Ares, M., Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science 296, 907-10 (2002).
61. Stolc, V. et al. A gene expression map for the euchromatic genome of Drosophila melanogaster. Science 306, 655-60 (2004).
62. Pan, Q. et al. Revealing global regulatory features of Mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16, 929-41 (2004).
63. Heber, S., Alekseyev, M., Sze, S. H., Tang, H. & Pevzner, P. A. Splicing graphs and EST assembly problem. Bioinformatics 18 Suppl 1, S181-8 (2002).
64. Fehlbaum, P., Guihal, C., Bracco, L. & Cochet, O. A microarray configuration to quantify expression levels and relative abundance of splice variants. Nucleic Acids Res 33, e47 (2005).
65. Xing, Y. & Lee, C. J. Protein Modularity of Alternatively Spliced Exons Is Associated with Tissue-Specific Regulation of Alternative Splicing. PLoS Genet 1, e34 (2005).
66. Pan, Q. et al. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev 20, 153-8 (2006).
67. Cuperlovic-Culf, M., Belacel, N., Culf, A. S. & Ouellette, R. J. Data analysis of alternative splicing microarrays. Drug Discov Today 11, 983-90 (2006).
68. Butte, A. The use and analysis of microarray data. Nat Rev Drug Discov 1, 951-60 (2002).

69. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185-93 (2003).
70. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249-64 (2003).
71. Li, C. et al. Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays. Cancer Res 66, 1990-9 (2006).
72. Le, K. et al. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. Nucleic Acids Res 32, e180 (2004).
73. Cline, M. S. et al. ANOSVA: a statistical method for detecting splice variation from expression data. Bioinformatics 21 Suppl 1, i107-i115 (2005).
74. Wang, H. et al. Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics 19 Suppl 1, i315-22 (2003).
75. Shai, O., Morris, Q. D., Blencowe, B. J. & Frey, B. J. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. Bioinformatics 22, 606-13 (2006).
76. Della Beffa, C., Cordero, F. & Calogero, R. A. Dissecting an alternative splicing analysis workflow for GeneChip Exon 1.0 ST Affymetrix arrays. BMC Genomics 9, 571 (2008).
77. Relogio, A. et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. J Biol Chem (2004).
78. Zhang, C. et al. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. BMC Bioinformatics 7, 202 (2006).
79. Sugnet, C. W. et al. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol 2, e4 (2006).
80. Ule, J. et al. An RNA map predicting Nova-dependent splicing regulation. Nature (2006).
81. Blanchette, M., Green, R. E., Brenner, S. E. & Rio, D. C. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. Genes Dev 19, 1306-14 (2005).
82. Ule, J. et al. Nova regulates brain-specific splicing to shape the synapse. Nat Genet 37, 844-52 (2005).
83. Hansen, K. D. et al. Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. PLoS Genet 5, e1000525 (2009).
84. Castle, J. C. et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40, 1416-25 (2008).
85. Adams, M. D., Kerlavage, A. R., Fields, C. & Venter, J. C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 4, 256-67 (1993).
86. Hillier, L. D. et al. Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6, 807-28 (1996).
87. Takeda, J. et al. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Res 34, 3917-28 (2006).
88. Xie, H. et al. Computational analysis of alternative splicing using EST tissue information. Genomics 80, 326-30 (2002).
89. Kim, N., Shin, S. & Lee, S. ECgene: Genome-based EST clustering and gene modeling for alternative splicing. Genome Res 15, 566-76 (2005).
90. Schmitt, A. O. et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res 27, 4251-60 (1999).

91. Bonaldo, M. F., Lennon, G. & Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6, 791-806 (1996).
92. Soares, M. B. et al. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci U S A 91, 9228-32 (1994).
93. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST--database for "expressed sequence tags". Nat Genet 4, 332-3 (1993).
94. Strausberg, R. L. The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. J Pathol 195, 31-40 (2001).
95. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. Genome Res 9, 1288-93 (1999).
96. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. Nat Genet 30, 29-30 (2002).
97. Resch, A. et al. Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res 3, 76-83 (2004).
98. Gerhard, D. S. et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14, 2121-7 (2004).
99. Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563-73 (2002).
100. Ota, T. et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36, 40-5 (2004).
101. Strausberg, R. L. et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A 99, 16899-903 (2002).
102. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. Science 286, 455-7 (1999).
103. Butterfield, Y. S. et al. An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. Nucleic Acids Res 30, 2460-8 (2002).
104. Baross, A. et al. Systematic recovery and analysis of full-ORF human cDNA clones. Genome Res 14, 2083-92 (2004).
105. Imanishi, T. et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2, e162 (2004).
106. Hayashizaki, Y. & Carninci, P. Genome Network and FANTOM3: assessing the complexity of the transcriptome. PLoS Genet 2, e63 (2006).
107. Zavolan, M. et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13, 1290-300 (2003).
108. Carninci, P. et al. The transcriptional landscape of the mammalian genome. Science 309, 1559-63 (2005).
109. Zheng, C. L., Fu, X. D. & Gribskov, M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. Rna (2005).
110. Watahiki, A. et al. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. Nat Methods 1, 233-239 (2004).
111. Thill, G. et al. ASEtrap: a biological method for speeding up the exploration of spliceomes. Genome Res 16, 776-86 (2006).
112. Venables, J. P. & Burn, J. EASI--enrichment of alternatively spliced isoforms. Nucleic Acids Res 34, e103 (2006).
113. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. Science 270, 484-7 (1995).

114. Saha, S. et al. Using the transcriptome to annotate the genome. Nat Biotechnol 20, 508-12 (2002).
115. Boon, K. et al. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A 99, 11287-92 (2002).
116. Siddiqui, A. S. et al. A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. Proc Natl Acad Sci U S A 102, 18485-90 (2005).
117. Morrissy, A. S. et al. Next-generation tag sequencing for cancer gene expression profiling. Genome Res (2009).
118. Lu, J., Lal, A., Merriman, B., Nelson, S. & Riggins, G. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. Genomics 84, 631-6 (2004).
119. van Ruissen, F. et al. Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. BMC Genomics 6, 91 (2005).
120. Robertson, N. et al. DiscoverySpace: an interactive data analysis application. Genome Biol 8, R6 (2007).
121. Kuo, B. Y. et al. SAGE2Splice: unmapped SAGE tags reveal novel splice junctions. PLoS Comput Biol 2, e34 (2006).
122. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100, 15776-81 (2003).
123. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38, 626-35 (2006).
124. Wei, C. L. et al. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci U S A 101, 11701-6 (2004).
125. Ng, P. et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2, 105-11 (2005).
126. Metzker, M. L. Emerging technologies in DNA sequencing. Genome Res 15, 1767-76 (2005).
127. Brenner, S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18, 630-4 (2000).
128. Siddiqui, A. S. et al. Sequence biases in large scale gene expression profiling data. Nucleic Acids Res 34, e83 (2006).
129. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature (2005).
130. Leamon, J. H., Braverman, M. S. & Rothberg, J. M. High-throughput, massively parallel DNA sequencing technology for the era of personalized medicine. Gene Therapy and Regulation 3, 15-31 (2007).
131. Bainbridge, M. N. et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:246 (2006).
132. Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5, 613-9 (2008).
133. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377-82 (2009).
134. Gowda, M. et al. Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. Nucleic Acids Res (2006).

135. Ng, P. et al. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res 34, e84 (2006).
136. Nielsen, K. L., Hogh, A. L. & Emmersen, J. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res (2006).
137. Denoeud, F. et al. Annotating genomes with massive-scale RNA sequencing. Genome Biol 9, R175 (2008).
138. Yassour, M. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A 106, 3264-9 (2009).
139. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-8 (2008).
140. Sultan, M. et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956-60 (2008).
141. Li, H. et al. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A 105, 20179-84 (2008).
142. Morin, R. et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45, 81-94 (2008).
143. Alberts, B. et al. Molecular Biology of the Cell (Garland Publishing, Inc., New York & London, 1994).
144. Ruan, Y., Le Ber, P., Ng, H. H. & Liu, E. T. Interrogating the transcriptome. Trends Biotechnol 22, 23-30 (2004).
145. Salehi-Ashtiani, K. et al. Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. Nat Methods 5, 597-600 (2008).
146. Gustincich, S. et al. The complexity of the mammalian transcriptome. J Physiol 575, 321-32 (2006).
147. Kuhn, R. M. et al. The UCSC genome browser database: update 2007. Nucleic Acids Res (2006).
148. Hubbard, T. et al. Ensembl 2005. Nucleic Acids Res 33, D447-53 (2005).
149. David, A. et al. Unusual alternative splicing within the human kallikrein genes KLK2 and KLK3 gives rise to novel prostate-specific proteins. J Biol Chem 277, 18084-90 (2002).
150. Major, S. M. et al. AbMiner: a bioinformatic resource on available monoclonal antibodies and corresponding gene identifiers for genomic, proteomic, and immunologic studies. BMC Bioinformatics 7, 192 (2006).
151. Kuroyanagi, H., Kobayashi, T., Mitani, S. & Hagiwara, M. Transgenic alternative-splicing reporters reveal tissue-specific expression profiles and regulation mechanisms in vivo. Nat Methods 3, 909-15 (2006).
152. Kemp, P. R., Ellis, P. D. & Smith, C. W. Visualization of alternative splicing in vivo. Methods 37, 360-7 (2005).
153. Nanjundan, M., Zhang, F., Schmandt, R., Smith-McCune, K. & Mills, G. B. Identification of a novel splice variant of AML1b in ovarian cancer patients conferring loss of wild-type tumor suppressive functions. Oncogene (2006).
154. Vegran, F. et al. Overexpression of caspase-3s splice variant in locally advanced breast carcinoma is associated with poor response to neoadjuvant chemotherapy. Clin Cancer Res 12, 5794-800 (2006).
155. Paddison, P. J. et al. A resource for large-scale RNA-interference-based screens in mammals. Nature 428, 427-31 (2004).

156. Brasch, M. A., Hartley, J. L. & Vidal, M. ORFeome cloning and systems biology: standardized mass production of the parts from the parts-list. Genome Res 14, 2001-9 (2004).
157. Brizuela, L., Braun, P. & LaBaer, J. FLEXGene repository: from sequenced genomes to gene repositories for high-throughput functional biology and proteomics. Mol Biochem Parasitol 118, 155-65 (2001).
158. Frese, K. K. & Tuveson, D. A. Maximizing mouse cancer models. Nat Rev Cancer 7, 645-58 (2007).
159. Yoshiki, A. et al. The mouse resources at the RIKEN BioResource center. Exp Anim 58, 85-96 (2009).
160. Figeys, D., McBroom, L. D. & Moran, M. F. Mass spectrometry for the study of protein-protein interactions. Methods 24, 230-9 (2001).
161. Thanaraj, T. A., Clark, F. & Muilu, J. Conservation of human alternative splice events in mouse. Nucleic Acids Res 31, 2544-52 (2003).
162. Pan, Q. et al. Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet 21, 73-7 (2005).
163. Kopelman, N. M., Lancet, D. & Yanai, I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet 37, 588-9 (2005).
164. Su, Z., Wang, J., Yu, J., Huang, X. & Gu, X. Evolution of alternative splicing after gene duplication. Genome Res 16, 182-9 (2006).
165. Stamm, S. et al. Function of alternative splicing. Gene 344, 1-20 (2005).
166. Schmucker, D. et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101, 671-84 (2000).
167. Vanhoutteghem, A. & Djian, P. The human basonuclin 2 gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2000 different proteins. Genomics (2006).
168. Goodson, M. L., Jonas, B. A. & Privalsky, M. L. Alternative mRNA splicing of SMRT creates functional diversity by generating corepressor isoforms with different affinities for different nuclear receptors. J Biol Chem 280, 7493-503 (2005).
169. Wang, P., Yan, B., Guo, J. T., Hicks, C. & Xu, Y. Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc Natl Acad Sci U S A 102, 18920-5 (2005).
170. Yura, K. et al. Alternative splicing in human transcriptome: functional and structural influence on proteins. Gene 380, 63-71 (2006).
171. Davis, M. J. et al. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. PLoS Genet 2, e46 (2006).
172. Xing, Y., Xu, Q. & Lee, C. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett 555, 572-8 (2003).
173. Forrest, A. R. et al. Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases. Genome Biol 7, R5 (2006).
174. Bjarnadottir, T. K. et al. Identification of novel splice variants of Adhesion G protein-coupled receptors. Gene (2006).
175. Ravasi, T. et al. Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. Genome Res 13, 1430-42 (2003).
176. Schwerk, C. & Schulze-Osthoff, K. Regulation of apoptosis by alternative pre-mRNA splicing. Mol Cell 19, 1-13 (2005).

177. Hillman, R. T., Green, R. E. & Brenner, S. E. An unappreciated role for RNA surveillance. Genome Biol 5, R8 (2004).
178. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100, 189-92 (2003).
179. Cuccurese, M., Russo, G., Russo, A. & Pietropaolo, C. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. Nucleic Acids Res 33, 5965-77 (2005).
180. Hughes, T. A. Regulation of gene expression by alternative untranslated regions. Trends Genet 22, 119-22 (2006).
181. Caceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet 18, 186-93 (2002).
182. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. Genes Dev 17, 419-37 (2003).
183. Garcia-Blanco, M. A., Baraniak, A. P. & Lasda, E. L. Alternative splicing in disease and therapy. Nat Biotechnol 22, 535-46 (2004).
184. Stoilov, P. et al. Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA Cell Biol 21, 803-18 (2002).
185. Licatalosi, D. D. & Darnell, R. B. Splicing regulation in neurologic disease. Neuron 52, 93-101 (2006).
186. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5, 389-96 (2004).
187. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21, 577-81 (2003).
188. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3, 285-98 (2002).
189. Ars, E. et al. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9, 237-47 (2000).
190. Teraoka, S. N. et al. Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am J Hum Genet 64, 1617-31 (1999).
191. Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nat Genet 30, 377-84 (2002).
192. Wang, Z. et al. Systematic identification and analysis of exonic splicing silencers. Cell 119, 831-45 (2004).
193. Zhang, X. H. & Chasin, L. A. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18, 1241-50 (2004).
194. Smith, P. J. et al. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet 15, 2490-508 (2006).
195. Pajares, M. J. et al. Alternative splicing: an emerging topic in molecular and clinical oncology. Lancet Oncol 8, 349-57 (2007).
196. Hu, A. & Fu, X. D. Splicing oncogenes. Nat Struct Mol Biol 14, 174-5 (2007).
197. Venables, J. P. et al. Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol (2009).
198. Salesse, S., Dylla, S. J. & Verfaillie, C. M. p210BCR/ABL-induced alteration of pre-mRNA splicing in primary human CD34+ hematopoietic progenitor cells. Leukemia 18, 727-33 (2004).
199. Brinkman, B. M. Splice variants as cancer biomarkers. Clin Biochem 37, 584-94 (2004).

200. Hui, L. et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. Oncogene 23, 3013-23 (2004).
201. Kirschbaum-Slager, N., Parmigiani, R. B., Camargo, A. A. & de Souza, S. J. Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data. Physiol Genomics (2005).
202. Xu, Q. & Lee, C. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. Nucleic Acids Res 31, 5635-43 (2003).
203. Kirschbaum-Slager, N., Lopes, G. M., Galante, P. A., Riggins, G. J. & de Souza, S. J. Splicing factors are differentially expressed in tumors. Genet Mol Res 3, 512-20 (2004).
204. Wiles, M. & Andreassen, P. Monoclonals: the billion dollar molecules of the future. Drug Discov World, 17-23 (2006).
205. Bracco, L. & Kearsey, J. The relevance of alternative RNA splicing to pharmacogenomics. Trends Biotechnol 21, 346-53 (2003).
206. Yeo, G. W. Splicing regulators: targets and drugs. Genome Biol 6, 240 (2005).
207. Wilton, S. D. & Fletcher, S. RNA splicing manipulation: strategies to modify gene expression for a variety of therapeutic outcomes. Curr Gene Ther 5, 467-83 (2005).
208. Du, L. & Gatti, R. A. Progress toward therapy with antisense-mediated splicing modulation. Curr Opin Mol Ther 11, 116-23 (2009).
209. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. Cell 100, 57-70 (2000).
210. Carethers, J. M. Review: Systemic treatment of advanced colorectal cancer: Tailoring therapy to the tumor. Therapeutic Advances in Gastroenterology 1, 33-42 (2008).
211. Luqmani, Y. A. Mechanisms of drug resistance in cancer chemotherapy. Med Princ Pract 14 Suppl 1, 35-48 (2005).
212. Albert, T. J. et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 4, 903-5 (2007).
213. Morin, R. D. et al. Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling. Genome Res 16, 796-803 (2006).
214. Morin, R. D. et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18, 610-21 (2008).
215. Kuchenbauer, F. et al. In-depth characterization of the microRNA transcriptome in a leukemia progression model. Genome Res 18, 1787-97 (2008).
216. Pugh, T. J. et al. Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res 36, e80 (2008).
217. Griffith, O. L. et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36, D107-13 (2008).
218. Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 4, e1000147 (2008).
219. Venables, J. P. Unbalanced alternative splicing and its significance in cancer. Bioessays 28, 378-86 (2006).
220. Rupp, U. et al. Safety and pharmacokinetics of bivatuzumab mertansine in patients with CD44v6-positive metastatic breast cancer: final results of a phase I study. Anticancer Drugs 18, 477-85 (2007).
221. Venables, J. P. Aberrant and alternative splicing in cancer. Cancer Res 64, 7647-54 (2004).
222. Simpson, J. T. et al. ABySS: A parallel assembler for short read sequence data. Genome Res 19, 1117-23 (2009).
223. Birol, I. et al. De novo Transcriptome Assembly with ABySS. Bioinformatics (2009).

224. Picardi, E., Mignone, F. & Pesole, G. EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data. BMC Bioinformatics 10 Suppl 6, S10 (2009).
225. Fejes, A. P. et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24, 1729-30 (2008).
226. Churbanov, A., Pauley, M., Quest, D. & Ali, H. A method of precise mRNA/DNA homology-based gene structure prediction. BMC Bioinformatics 6, 261 (2005).
227. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
228. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8, 967-74 (1998).
229. van Nimwegen, E., Paul, N., Sheridan, R. & Zavolan, M. SPA: a probabilistic algorithm for spliced alignment. PLoS Genet 2, e24 (2006).
230. Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. Genome Res 11, 1952-7 (2001).
231. Usuka, J., Zhu, W. & Brendel, V. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16, 203-11 (2000).
232. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res 11, 889-900 (2001).
233. De Bona, F., Ossowski, S., Schneeberger, K. & Ratsch, G. Optimal spliced alignments of short sequence reads. Bioinformatics 24, i174-80 (2008).
234. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477-8 (1997).
235. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct 3, 20 (2008).
236. Liang, C., Liu, L. & Ji, G. WebGMAP: a web service for mapping and aligning cDNA sequences to genomes. Nucleic Acids Res 37, W77-83 (2009).
237. Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K. & Go, M. AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. Nucleic Acids Res 37, D305-9 (2009).
238. Kim, N., Alekseyenko, A. V., Roy, M. & Lee, C. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res (2006).
239. Stamm, S. et al. ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res 34, D46-55 (2006).
240. Bhasi, A., Philip, P., Sreedharan, V. T. & Senapathy, P. AspAlt: A tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. Genomics 94, 48-54 (2009).
241. Castrignano, T. et al. ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization. Nucleic Acids Res 34, W440-3 (2006).
242. Cai, J., Zhang, J., Huang, Y. & Li, Y. ATID: a web-oriented database for collection of publicly available alternative translational initiation events. Bioinformatics 21, 4312-4 (2005).
243. Hsu, F. R. et al. AVATAR: a database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. Bioinformation 1, 16-8 (2005).
244. Lacroix, Z., Legendre, C., Raschid, L. & Snyder, B. BIPASS: BioInformatics Pipeline Alternative Splicing Services. Nucleic Acids Res 35, W292-6 (2007).

245. Kim, P. et al. ECgene: genome annotation for alternative splicing. Nucleic Acids Res 33 Database Issue, D75-9 (2005).
246. Takeda, J. et al. H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. Nucleic Acids Res 35, D104-9 (2007).
247. Holste, D., Huo, G., Tung, V. & Burge, C. B. HOLLYWOOD: a comparative relational database of alternative splicing. Nucleic Acids Res 34, D56-62 (2006).
248. Shah, P. K., Jensen, L. J., Boue, S. & Bork, P. Extraction of transcript diversity from scientific literature. PLoS Comput Biol 1, e10 (2005).
249. Zheng, C. L. et al. MAASE: An alternative splicing database designed for supporting splicing microarray applications. Rna (2005).
250. Zhang, H., Hu, J., Recce, M. & Tian, B. PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res 33, D116-20 (2005).
251. Ryan, M. C. et al. SpliceCenter: a suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. BMC Bioinformatics 9, 313 (2008).
252. Huang, H. D., Horng, J. T., Lin, F. M., Chang, Y. C. & Huang, C. C. SpliceInfo: an information repository for mRNA alternative splicing in human genome. Nucleic Acids Res 33 Database Issue, D80-5 (2005).
253. Kahn, A. B. et al. SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis. BMC Bioinformatics 8, 75 (2007).
254. Gupta, S., Vingron, M. & Haas, S. A. T-STAG: resource and web-interface for tissue-specific transcripts and genes. Nucleic Acids Res 33, W654-8 (2005).
255. Zhang, X. H., Leslie, C. S. & Chasin, L. A. Computational searches for splicing signals. Methods 37, 292-305 (2005).
256. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 31, 3568-71 (2003).
257. Kostadinov, R. et al. GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. Nucleic Acids Res 34, D119-24 (2006).
258. Huang, H. Y., Chien, C. H., Jen, K. H. & Huang, H. D. RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. Nucleic Acids Res 34, W429-34 (2006).
259. Fairbrother, W. G. et al. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32, W187-90 (2004).
260. Akerman, M., David-Eden, H., Pinter, R. Y. & Mandel-Gutfreund, Y. A computational approach for genome-wide mapping of splicing factor binding sites. Genome Biol 10, R30 (2009).
261. Wang, X. et al. Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. BMC Genomics 10 Suppl 1, S4 (2009).
262. Hiller, M. et al. TassDB: a database of alternative tandem splice sites. Nucleic Acids Res (2006).
263. Griffith, M. et al. ALEXA: a microarray design platform for alternative expression analysis. Nat Methods 5, 118 (2008).
264. Chang, T. Y. et al. easyExon--a Java-based GUI tool for processing and visualization of Affymetrix exon array data. BMC Bioinformatics 9, 432 (2008).
265. Robinson, M. D. & Speed, T. P. Differential splicing using whole-transcript microarrays. BMC Bioinformatics 10, 156 (2009).

266. Zheng, H. et al. REMAS: a new regression model to identify alternative splicing events from exon array data. BMC Bioinformatics 10 Suppl 1, S18 (2009).
267. Shah, S. H. & Pallas, J. A. Identifying differential exon splicing using linear models and correlation coefficients. BMC Bioinformatics 10, 26 (2009).
268. Anton, M. A. et al. SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. Genome Biol 9, R46 (2008).
269. Kim, N., Lim, D., Lee, S. & Kim, H. ASePCR: alternative splicing electronic RT-PCR in multiple tissues and organs. Nucleic Acids Res 33, W681-5 (2005).
270. Bollina, D., Lee, B. T., Tan, T. W. & Ranganathan, S. ASGS: an alternative splicing graph web service. Nucleic Acids Res 34, W444-7 (2006).
271. Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M. & Gotoh, O. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics 22, 1211-6 (2006).
272. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. Nucleic Acids Res 32, W273-9 (2004).

2. ALEXA: A microarray design platform for alternative expression analysis[2]

2.1. Introduction

Eukaryotic genomes are predicted to contain ~7,000 to ~30,000 genes[1]. Each of these genes may be alternatively processed to produce multiple distinct mRNAs by alternative transcript initiation, splicing and poly-adenylation (collectively referred to as alternative expression). Although analysis of available transcript resources indicates that up to ~75% of genes are alternatively processed, most microarray expression platforms cannot detect alternative transcripts[2].

Proof-of-principle experiments have described the use of oligonucleotide microarrays to profile transcript isoforms generated by alternative expression but resources to allow the creation of such arrays are lacking[3, 4]. To address this limitation we created a microarray design platform for 'alternative expression analysis' ('ALEXA'), which is capable of designing arrays that can detect all of the major categories of alternative expression. The ALEXA platform facilitates the selection and annotation of oligonucleotide probes representing alternative expression events for any species residing in the EnsEMBL database[1]. For each target gene, probes are selected within every exon, intron, exon junction and exon boundary. This approach allows the detection of constitutive and alternative exons, canonical exon junctions, the junctions of known or novel exon skipping events, alternative exon boundaries and retained introns (**Figure 2.1**). We designed the platform to be flexible to the user's experimental interests and preferred array manufacturer. The user may limit probe selection to known alternative expression events or include all possible exon junctions and boundaries to drive the discovery of novel transcripts. Probes may be designed for an arbitrary subset of genes or for all genes. Most technical parameters of the design can be modified by the user, including: the amount and types of control probes; the use of varying or fixed probe length; and the thresholds for filtering of probe sequences. The probe design process begins with retrieval of genomic sequences from EnsEMBL, removal of pseudogenes, masking of repeat elements and extraction of probe sequences. Random

[2] A version of this chapter has been published. Griffith M, Tang MJ, Griffith OL, Morin RD, Chan SY, Asano JK, Zeng T, Flibotte S, Ally A, Baross A, Hirst M, Jones SJM, Morin GB, Tai IT and Marra MA. *ALEXA – A microarray design platform for alternative expression analysis.* Nature Methods. 2008 Feb. 5(2):118.

probe sequences are generated to uniformly represent the melting temperature and length of all experimental probes. Extracted and randomly generated probes are scored according to their melting temperature, folding potential, complexity, and specificity (**Methods**).

Proof-of-principle experiments have described the use of oligonucleotide microarrays to measure expression of individual exons, exon junctions and exon boundaries[3-7]. The focus of these experiments has been to optimize design parameters such as probe selection and hybridization conditions. Commercial arrays implementing some of these findings have been developed by Affymetrix[8], Exon Hit Therapeutics[9] and Jivan Biologics[10], but these are limited to particular gene families or a subset of alternative expression events and are only available for the human, mouse and rat genomes. Although several publications have described using 'splicing microarrays' to study the genomes of model organisms[5, 7, 11], survey tissues[6] and address specific biological questions[12-17], ours is the first report of a resource that makes such designs readily available. Furthermore, our platform represents the first open-source method for the generation of alternative expression (AE) microarrays for any EnsEMBL annotated species. We used the ALEXA platform to generate AE microarray designs for the human, mouse, rat, fly, and budding yeast genomes as well as the first designs for chimp, dog, chicken, zebrafish and *Caenorhabditis elegans* (**Table 2.1**).

We assessed the ALEXA approach by using a prototype human array to profile the expression of alternative mRNA isoforms in 5-fluourouracil (5-FU) sensitive and resistant colorectal cancer cell lines[18] and comparing the results to those from Affymetrix's 'GeneChip® Human Exon 1.0 ST' array (**Figure 2.2** - **Figure 2.8** and **Table 2.2**). Genes and exons differentially expressed between 5-FU sensitive and resistant cells were identified by both platforms (with significant overlap), but ALEXA arrays provided additional information on the connectivity and boundaries of exons (**Table 2.3**). Furthermore, alternative expression events identified by ALEXA were significantly enriched for known alternative expression events represented in publicly available mRNA and EST databases. Finally, we demonstrated the advantage of the ALEXA approach by identifying several differentially expressed known and predicted isoforms with potential relevance to 5-FU resistance (**Figure 2.9** - **Figure 2.13** and **Table 2.4** - **Table 2.5**). Although we compare the output of analytical approaches involving two microarray platforms, due to differences in the array platforms themselves, this work is

not meant to represent a head-to-head comparison of the oligonucleotide probe performance of the respective microarray platforms.

The approach and resources described in this work have considerable potential to advance studies of gene regulation, transcript processing, human disease and evolutionary biology. The source code, pre-computed array designs and related materials to assist in the creation of custom alternative expression microarrays are provided on the ALEXA website (www.AlexaPlatform.org).

2.2. Results

To illustrate the utility of our design strategy we created a prototype array to profile genes expressed in a 5-fluorouracil (5-FU) sensitive human colorectal cancer cell line, MIP101[19] and in its drug resistant derivative, MIP/5FU[18]. Although the goal of these experiments was to assess the performance of ALEXA arrays we also identified differentially expressed (DE) mRNA isoforms associated with acquired 5-FU resistance. RNA samples isolated in triplicate from each cell line were profiled on Affymetrix's 'GeneChip® Human Exon 1.0 ST' array (hereafter referred to as the 'Affymetrix exon array'), which was designed to measure the expression of ~1.4 million known and predicted exons. A custom ALEXA design consisting of 385,000 features was then created, synthesized by NimbleGen Systems Inc. and used to profile the same RNA samples. Although our prototype arrays were synthesized by NimbleGen, no elements of the ALEXA platform are specific to this manufacturer. Validation of the ALEXA platform consisted of comparison to Affymetrix results as well as to mRNA and EST sequence databases. To our knowledge, this is the first reported comparison of two different microarray expression platforms measuring DE of individual exons for the same set of samples. We describe the level of concordance between these platforms for measuring expression at the level of both genes and exons and highlight the exon connectivity and boundary information provided by ALEXA arrays.

The ALEXA platform was written in Perl and utilizes a MySQL relational database (**Methods**). This database stores information on oligonucleotide probes and associated gene, transcript, exon and protein features. The design platform allows the selection of probes representing AE events for any species with EnsEMBL annotations[1] (35 species as of EnsEMBL version 45). Probes are selected within every exon and intron as well as across every exon junction and boundary. This approach has the potential to detect

expression of constitutive and alternative exons, canonical exon junctions, the junctions of known or novel exon skipping events, alternative exon boundaries and retained introns (**Figure 2.1** compares Affymetrix and ALEXA probe selection strategies). The ALEXA platform was designed to be flexible to the user's design interests. The user may limit probe selection to AE events predicted by ESTs, but to drive the discovery of novel transcripts the option of interrogating all possible exon junctions and boundaries is also available. Probes may be designed for a single gene, all genes, or an arbitrary subset of genes. A variety of technical elements of the design can be controlled by the user. For example, the amount and types of control probes, the use of varying or fixed probe length, and the thresholds used for the filtering of probe sequences according to melting temperature (Tm), sequence complexity, secondary structure and sequence specificity can be modified. The probe design process begins with retrieval of genomic sequences from EnsEMBL, removal of pseudogenes, masking of repeat elements and extraction of probe sequences. Random probe sequences are generated to uniformly represent the Tm and length of all experimental probes. Extracted and randomly generated probes are scored according to their Tm, hairpin or dimerization potential, presence of low complexity elements, specificity of each probe by comparison to all available ESTs, mRNAs and EnsEMBL transcripts, and the specificity of each probe within the total population of probes. All extracted probes are stored but a filtered set is defined to remove probes with sub-optimal specificity and thermodynamic properties (**Methods**). The 'best' exon and intron probes are chosen from probes tiled across these regions at 5 bp intervals.

2.2.1. Pre-computed microarray designs

Using the approach described above, pre-computed designs consisting of ~100 million probe sequences for ten EnsEMBL genomes were generated (**Table 2.1**). All pre-computed array designs, source code, database schemas and user manuals are provided on the ALEXA website (www.AlexaPlatform.org). Source code can be downloaded and installed by the user, and the platform is also available for use on Linux, Mac and Windows operating systems as a preconfigured 'virtual machine appliance'[20].

2.2.2. Validation - cross platform analysis

Total RNA was isolated from biological triplicates of the colorectal cancer cell line MIP101[19] and a derivative 5-FU resistant cell line, MIP/5FU[18]. This RNA was processed and hybridized to Affymetrix exon arrays (**Methods**). The quality of data resulting from these 6 array hybridizations was assessed by comparison to data from 10 Affymetrix exon array experiments also performed in our lab and to publicly available exon array data from 20 assays performed at Affymetrix[21]. A 'receiver operator characteristic - area under the curve' ('ROC AUC') score for all 36 arrays was calculated using Affymetrix's ExACT software[8]. ROC AUC scores are a measure of overall sensitivity and specificity. In this approach a true positive is an exon of a housekeeping gene that is determined to be expressed and a true negative is an intron of a housekeeping gene that is determined to be not expressed. Data from the 6 hybridizations of MIP101 and MIP/5FU samples had the 6 highest scores overall (0.912-0.918 compared to 0.732-0.897 for all other array hybridizations).

Affymetrix exon expression data was used to identify ~2,000 genes with evidence for DE of one or more exons between sensitive and resistant cells. A custom ALEXA array design consisting of 385,000 oligonucleotide probes was then designed and manufactured to represent these genes as well as ~500 additional genes with potential relevance to drug resistance (**Methods**). PolyA+ RNA was isolated from the same total RNA samples used for Affymetrix experiments and processed and hybridized to ALEXA arrays synthesized by NimbleGen. Although all of our arrays were synthesized by NimbleGen, ALEXA oligonucleotides may be synthesized by any custom oligonucleotide array manufacturer.

Sensitivity and specificity of the Affymetrix and ALEXA arrays were compared by examining data for 97 housekeeping genes that were defined by Affymetrix and were also represented by ALEXA oligonucleotides. This allowed a direct comparison of values for genes and exons with a high likelihood of expression. For each of these genes, probes were selected for exons (4,702 in Affymetrix and 1,719 in ALEXA) and introns (14,551 in Affymetrix and 1,738 in ALEXA). For these probesets the overall signal-to-noise ratio (mean exon/intron) on each platform was higher in the ALEXA data (56.0+/-2.3 s.d.) than in Affymetrix (20.9+/-0.42 s.d.) and this difference was significant (Wilcoxon P = 0.0022, n=12). Housekeeping intron and exon probesets were also used to calculate ROC AUC scores to estimate the rate of false positive and false negative

detection of expression for each platform (**Methods**). The resulting AUC scores were 0.952 for ALEXA and 0.913 for Affymetrix (**Figure 2.2**). The ALEXA data achieved a maximum specificity of 94.8% at 87.9% sensitivity and the Affymetrix data achieved a maximum specificity of 85.8% at 84.2% sensitivity. Some of the caveats of these cross-platform comparisons are addressed below (**Discussion**).

The reproducibility of expression and DE estimates for both genes and exons within each set of biological replicates was consistently higher in the ALEXA data (**Table 2.2**). The ability of ALEXA and Affymetrix platforms to detect expression and DE of individual genes and exons was compared directly for all features profiled by both platforms. 2,507 genes and 31,368 EnsEMBL exons were interrogated by at least one probe in both platforms. The 'probesets' corresponding to each EnsEMBL exon generally consisted of 3 probes in ALEXA and 4 probes in Affymetrix. The number of probes for any particular gene depended on the number of exons in the gene. Ranked absolute expression values were compared between platforms and resulted in Spearman correlation coefficients of 0.88 for genes and 0.74 for exons. Mean DE estimates from each platform were also plotted against each other for both genes and exons (**Figure 2.3** - **Figure 2.4**). Cross-platform comparison of DE values resulted in Pearson correlations of 0.87 and 0.67 across all genes and exons respectively. At the gene-level, this is a high level of correlation compared to previously published cross-platform comparisons[22]. To our knowledge a cross-platform correlation of expression at the exon-level has not been previously reported.

The subset of genes and exons with statistically significant DE (and no fold-change cutoff) was identified for each platform and the overlap determined (**Methods**). Of the 667 genes identified as DE by ALEXA, and the 650 by Affymetrix, 482 were identified by both platforms (58%) (**Figure 2.5**). The ALEXA platform identified approximately 3 times as many DE exons as Affymetrix (2,927 compared to 956) with the overlap between platforms at 516 (15%) (**Figure 2.5**). The 2,411 exons detected as differentially expressed by ALEXA but not by Affymetrix were found to have lower overall expression in the Affymetrix data, suggesting a reduced ability to adequately detect these exons compared to ALEXA (**Figure 2.6**). A similar effect was observed for exons detected as DE by Affymetrix but not ALEXA, but the size of this set was considerably smaller (440 exons). The dynamic range of expression values for both genes and exons was larger in the ALEXA data (**Figure 2.7** - **Figure 2.8**).

2.2.3. Differentially expressed genes and mRNA isoforms associated with 5-FU resistance

We summarized the total numbers of each type of expression event profiled by both the Affymetrix exon array and ALEXA array, as well as those events identified as differentially expressed between 5-FU sensitive and resistant states (**Table 2.3**; significant p-value after multiple testing correction and fold-change > 2; see **Methods**). The number of differentially expressed genes was 78 and 233 for Affymetrix and ALEXA, respectively. 46 genes had significant p-values and a fold-change value greater than 4 in one or both expression platforms (**Table 2.4**). Within this list, a number of interesting genes with high expression fold-change values and agreement between the platforms emerged. For example, the top two differentially expressed genes, '*C12orf59*' and '*OLR1*' are expressed from the same locus in a head-to-head arrangement and were down-regulated by ~50 and ~33-fold in 5-FU resistant cells, respectively (**Table 2.4** and **Figure 2.9**).

Although the differentially expressed genes identified in this work have potential relevance to the mechanism(s) of 5-FU resistance, the novelty of Affymetrix exon and ALEXA arrays is their potential to identify discrete expression events corresponding to single exons or introns and in the case of the ALEXA arrays to identify DE of exon junctions and boundaries. For the genes targeted by both platforms, the total number of such events profiled was ~118,000 for the Affymetrix array and ~184,000 for the ALEXA array (**Table 2.3**). The number of these events which were found to be significantly DE between sensitive and resistant cells, the number predicted to affect the open reading frame (ORF) of a protein and the number overlapping a known protein feature are reported in **Table 2.3**. The majority (~98%) of events identified by Affymetrix exon arrays corresponded to exons and only 25 corresponded to introns. Manual inspection of these intronic probesets suggested that these were cryptic or misannotated exons with mRNA and/or EST support that were not correctly identified in EnsEMBL. The ALEXA platform identified almost as many DE events corresponding to canonical exon junctions as those belonging to the exon category. Furthermore, within the DE dataset the ALEXA analysis identified 440 potential alternative expression (AE) events corresponding to 191 exon skips and 253 alternate exon boundaries. Many of these were supported by EST and mRNA data. Specifically, 34.0% of AE events which reached significance were supported by a known mRNA, whereas only 5.9% of all AE

events on the ALEXA array had a corresponding mRNA (Fisher's Exact P = 2.2×10^{-16}). Similarly, 49.8% of AE events which reached significance were supported by an EST, whereas only 12.8% of all AE events profiled had a corresponding EST (Fisher's Exact P = 2.2×10^{-16}).

In the ALEXA data, there is a small but statistically significant enrichment for DE events that occur within ORFs over those occurring in 5′ UTR or 3′ UTRs (Chi-Squared P = 1.012×10^{-14}). The significant DE events were tested for enrichment of probesets within signal peptide motifs, coiled-coil domains, transmembrane domains, and protein family motifs. There was a significant enrichment for events within transmembrane domains in the population of differentially expressed probesets (~20% more than expected by chance, Fisher's Exact P = 9.93×10^{-9}).

The majority of exon and canonical junction probesets with evidence for DE likely reflect the DE of entire genes and only a subset of these are expected to correspond to DE of specific isoforms. A subset of cases were therefore identified by calculating 'splicing index' values[5] in an attempt to identify exons, junctions or boundaries that were differentially expressed relative to changes in expression at the level of the entire gene (**Methods**). A list of the top 25 candidate loci with apparent DE of isoforms and corresponding fold-change values is provided as **Table 2.5**. 19 of these had mRNA and/or EST support, 11 were supported by the Affymetrix data, and 6 could not have been detected by the Affymetrix platform due to a lack of the necessary probes. The data for each of these loci were manually examined by generating custom UCSC genome browser tracks to represent expression data as well as display the position of mRNAs, ESTs, ORFs, and protein features relative to the position of each probe (**Methods**). Using these displays and other bioinformatic resources, each of these genes was manually annotated to describe the DE event, the potential relevance of the gene to 5-FU resistance, the availability of supporting ESTs and mRNAs, the sub-cellular localization of the predicted protein, and the level of agreement between the expression platforms (**Table 2.4** - **Table 2.5**). Candidate DE isoforms ranged from those involving a large number of exons to those with relatively subtle differences involving a single exon skip or alternative exon boundary. For example, alternative transcript initiation or polyadenylation appeared to result in the up-regulation of a known isoform of *LAMA3* and putative novel isoforms of *EPB41L3* and *c12orf63* in resistant cells (**Figure 2.10** - **Figure 2.12**). Another interesting example illustrating the

advantages of the ALEXA approach was the observation of reciprocal expression between 5-FU sensitive and resistant cells of two isoforms of the gene *UMPS* which differ by a single exon (**Figure 2.13**). The expression of *UMPS* isoform A was ~5-fold lower in resistant cells than sensitive cells. Conversely, the expression of isoform B was ~6-fold higher in resistant cells than sensitive cells (**Figure 2.13**). The Affymetrix array reported the DE of *UMPS* exon 2 as ~5-fold but did not identify reciprocal expression of the two isoforms because it lacked the exon-junction probes required to measure both isoforms (**Figure 2.13**). It was not identified as a prominent candidate until the ALEXA data were examined, where it is was ranked 4th among reciprocal isoform expression events.

2.3. Discussion

Although a number of groups have reported the use of microarrays to profile alternative expression events, resources that facilitate the design of such arrays are not readily available. This study describes the first platform for the design of AE microarrays which is open source, flexible to the needs of the user and applicable to any species annotated in EnsEMBL. We determined the effectiveness of this platform by using it to generate a human microarray design and comparing its ability to profile exons to that of the Affymetrix GeneChip® exon array platform. The performance of our ALEXA arrays was comparable or superior to the Affymetrix arrays in every metric examined. A number of differences in array design and experimental protocols complicate the interpretation of comparisons between these two array platforms. For example, the Affymetrix and ALEXA array experiments differed in the following ways: (1) the length of oligonucleotide probes selected (25-mers versus 26- to 46-mers in ALEXA experiments), (2) the preparation of RNA input (total RNA subjected to riboMinus reduction versus total RNA that has been polyA+ purified in ALEXA experiments), (3) the use of a target amplification step (linear IVT amplified target versus unamplified target in ALEXA experiments), (4) the molecular type of the target (cRNA versus ds-DNA in ALEXA experiments), and many other potentially significant differences (see **Methods** for further details). Due to the number and diversity of these experimental differences and the uncertainty of their effect on the resulting hybridization data, it was difficult to identify the source of performance differences between the two approaches. The intention of our cross-platform comparison was therefore not to suggest that our

probe selection strategy is fundamentally superior to that of Affymetrix but rather to use the high level of agreement between these two approaches in the areas where they provide the same information as a validation. Furthermore, it was our hope that this validation exercise would lend credibility to the predictions of the ALEXA approach in cases where was it designed to provide information that the Affymetrix arrays could not provide (such as exon skipping events). We believe that despite the caveats discussed we have shown that Affymetrix exon arrays and ALEXA arrays were similar in their ability to detect the expression of exon features but the ALEXA approach provided additional information on the connectivity and boundaries of those exons. The alternative exon junctions and boundaries which were found to be differentially expressed but did not correspond to known EnsEMBL transcripts were highly enriched for EST supported sequences relative to all junctions and boundaries profiled. This suggests that the ALEXA approach has considerable potential for the identification of novel expressed transcripts generated by alternative expression. Currently the primary limitation of the ALEXA approach is the inability of current custom microarray densities (maximum of 385,000 features) to accommodate genome-wide designs due to the large number of probes required to cover all exons, junctions and boundaries (**Table 2.1** and **Table 2.3**). At least two array manufacturers are currently developing arrays with sufficient density (2-3 million features) to eliminate this limitation[23, 24].

In addition to showing the general utility of the ALEXA approach by comparison to Affymetrix exon expression data, our analysis of colorectal cancer cell lines identified DE of known and putative novel isoforms associated with acquired resistance to the widely used chemotherapy drug 5-FU. This drug is a uracil analog which was designed to exploit the observation that tumor cells preferentially utilize uracil (as opposed to thymidine) for RNA and DNA synthesis[25]. The differential expression events observed in our analysis ranged from those involving entire genes to those involving specific isoforms with subtle differences in exon content. Several of the genes and isoforms we identified have unknown function and represent candidate novel chemotherapy resistance genes. Other examples, such as the differential expression of *UMPS* isoforms, have a clear potential relevance to 5-FU resistance. This gene is known to function in pyrimidine biosynthesis and is involved in the activation of 5-FU[26] by the uridine biosynthesis pathway. *UMPS* encodes two enzymes that facilitate the last two steps in this pathway: EC 2.4.2.10 (Orotate phosphoribosyl-transferase) and EC

4.1.1.23 (Orotidine 5'-phosphate decarboxylase)[27]. *UMPS* isoform A and B differ by the inclusion or exclusion of exon 2, which is predicted to result in use of an alternate start codon in the 3rd exon (**Figure 2.13**). Without the Phosphoribosyltransferase domain of *UMPS* (i.e. isoform B), 5-FU would not be transferred to ribose and would not be incorporated into DNA/RNA where it normally slows tumor growth by inhibiting DNA synthesis and the synthesis, processing, and translation of mRNA[28]. A shift in the expression of *UMPS* isoforms may therefore influence the activation rate and efficacy of 5-FU. Several groups have previously examined the potential role of *UMPS* expression in mediating 5-FU efficacy using clinical specimens or cell lines, but their selection of PCR primers for expression assays and cDNAs for over-expression experiments did not account for the existence of multiple isoforms[25, 28, 29]. Although the existence of an *UMPS* isoform[30] has been known since 1999, to our knowledge no study has considered the functional significance of their relative ratios in uracil or 5-FU metabolism.

We believe that widespread adoption of microarray experiments capable of detecting alternative expression events will advance studies of gene regulation, transcript processing, human disease and evolutionary biology. For example, a recent study based on analysis of mRNAs and ESTs for eight organisms suggested that vertebrates have a higher rate of alternatively spliced exons and genes than invertebrates[31]. One finding was the unusually high frequency of alternative splicing in chicken. Our chicken array design could be used to validate this observation. Similarly, array designs for human, chimp, dog, zebrafish and other genomes might be used to address the evolutionary significance of alternative expression in eukaryotes. Furthermore, we anticipate that the application of microarrays to genome scale studies of alternative expression will help to develop a better understanding of the role alternative expression may play in health and disease states and improve identification of therapeutic and diagnostic targets. We have demonstrated this potential by identifying differential expression of known and novel alternative transcripts associated with 5-FU resistance in a human colorectal cancer cell line.

2.4. Methods

2.4.1. Probe extraction and filtering for array designs

EnsEMBL gene models, their corresponding genomic sequence and related information were downloaded via the EnsEMBL API[1] and imported into a MySQL database for each species (schema available at www.AlexaPlatform.org). Probes were extracted from a repeat masked[32] genomic sequence of each gene model. Exon and intron probes were extracted at 5 bp intervals. Exon-exon junction probes were extracted to represent every possible valid combination of two exons for each gene. A gene with n exons has (n!/(2![n-2]!)) possible junctions. Exon-intron junction probes were extracted to span every unique exon boundary in the gene. Exon-exon and exon-intron probes were extracted such that the sequence was centered on the junction. The length of all probes was varied by up to 10 bp from a desired length of 36 bp to achieve a target Tm of 67ºC (see **Table 2.1** for species specific details). For junction probes, the three lengths which produced the closest Tm to the target were stored. For exon and intron probes only the probe with the closest Tm at each 5 bp position was stored. Finally, 1.5 million random probe sequences (negative controls) were generated to uniformly represent the range of probe Tm and length observed for all exon, intron and junction probes.

Probe sequences were scored according to their thermodynamic properties and specificity. Tm was calculated by a nearest neighbor approach[33, 34] implemented in Perl. The strength of hairpin and dimer formation was determined by 'simfold' and 'pairfold' respectively[35]. Low complexity elements were identified by 'mdust'[36]. Sequence specificity was determined by 'blastn' of probe sequences against databases containing all EnsEMBL transcripts, mRNAs, ESTs, all probe sequences, and the entire genome (for random probes only). The entire probe population was then filtered so that a probe was marked as unsuitable unless it had: a Tm within 3.0ºC of the target Tm; no low complexity regions of 6 nucleotides or longer; a free-energy of hairpin folding greater than -8.0 kcal/mol; a free energy of dimerization greater than -20.0 kcal/mol; no alignments of 67.5% of the length of the probe or greater to any mRNA or EnsEMBL transcript; and no alignments of 80% of the length of the probe or greater to another probe in the design. Multiple probes (comprising a 'probe set') were selected to represent each exon, intron, boundary or junction to increase the accuracy of

expression estimates and statistical tests for each region and to compensate for probes with poor hybridization characteristics that were not successfully removed by the filtering process.

2.4.2. Creation of a validation array design

Before proceeding with the validation experiments described in this study, initial experiments using a prototype design (synthesized by NimbleGen Systems Inc., Madison, WI) and reference RNAs were conducted to optimize probe selection parameters, use of control probes and hybridization conditions (data not shown).

Creation of the validation array first involved identifying genes of interest. Due to the combinatorial nature of the design strategy, the large number of resulting probes per gene and the custom array densities available at the time, a genome-wide design was not possible. Approximately 3 million feature spots would be required to accommodate a genome-wide human design and the maximum density available was 385,000. Potential genes for the array were selected by identifying all genes with 2-fold or greater DE of one or more of their exons according to 'Affymetrix GeneChip Human Exon 1.0 ST' arrays (Affymetrix Inc. Santa Clara, CA). Approximately 100 genes defined as housekeeping controls on the Affymetrix exon array were also selected. Unlike most genes on the array, these were targeted by intron as well as exon probes. An additional ~400 genes were selected for their potential relevance to cancer biology or drug resistance. Specifically, this included genes of the ABC drug transport family, genes with known cancer related isoforms identified in the literature, genes from the cancer gene census[37], and genes associated with the gene ontology[38] terms: 'drug transporter activity', 'response to drug', 'drug metabolism', 'drug catabolism' and 'drug binding'.

The ALEXA validation design was generated by selecting probes corresponding to those genes described above. Only probes which passed all filtering steps were allowed. If less than 67.5% of the probes extracted for a particular gene remained after filtering, the gene was excluded. The final prototype design consisted of 385,000 probes of 26-46 bp in length corresponding to ~2,511 genes. Each exon, intron or junction was represented by 2-4 probes. Exon-exon junction probes were excluded if they represented an event where more than 3 exons would be skipped. The array was composed of probes representing ~31,000 exons, ~93,000 exon-exon junctions, ~50,000 exon-intron junctions, ~500 introns and ~4,500 random sequences. Random

probes were used to estimate false positives and for background correction according to the Tm of each probe. This array design was submitted to NimbleGen Systems Inc. for synthesis. The complete design files and accompanying annotation of features are available online (www.AlexaPlatform.org).

2.4.3. Tissue culture

The colorectal cancer cell line, 'MIP101'[19] and a previously generated 5-FU resistant derivative, 'MIP/5FU'[18] were maintained in Dulbecco's Modified Eagle Medium (Invitrogen Inc., Burlington, ON) supplemented with 10% new born calf serum (Invitrogen), 1% kanamycin (Invitrogen) and 1% penicillin/streptomycin (Invitrogen) at 37ºC and a humidified environment of 5% CO_2.

2.4.4. RNA isolation, labeling and hybridizations

Total RNA was extracted from cultures grown to 75% confluence using Trizol reagent (Invitrogen). Total RNA was DNaseI treated with an RNase free DNase set followed by cleanup on RNeasy columns (Qiagen Inc. Mississauga, ON). RNA was quantified and tested for degradation using an Agilent 2100 Bioanalyzer. Total RNA was used for hybridizations with the Affymetrix exon array platform and processed according to Affymetrix's recommendations as described in the 'GeneChip Whole Transcript Sense Target Labeling Assay Manual' (Affymetrix Inc. Santa Clara, CA). Briefly, this procedure consists of ribosomal RNA reduction, 1st cycle double stranded cDNA synthesis, linear amplification by *in vitro* transcription, 2nd cycle single stranded cDNA synthesis, enzymatic fragmentation, terminal labeling of fragments, hybridization, washing, staining and scanning.

To prepare samples for hybridizations to the validation ALEXA arrays, polyA+ RNA was isolated from total RNA with a μMACS mRNA isolation kit (Miltenyi Biotec, Gladbach, Germany) followed by double stranded cDNA synthesis with a 'Superscript Choice System' for cDNA Synthesis using random hexamers (Invitrogen). 5 μg of each cDNA sample was shipped to NimbleGen. Labeling, hybridization and scanning was conducted by NimbleGen using their 'ChIP-chip' protocol (optimized for 50-mers) and raw data was returned to us.

2.4.5. Data processing

Raw Affymetrix probe intensity values were extracted from .CEL files using Affymetrix's 'Exact-Probe-Intensity-Extraction' tool. Raw probe values for ALEXA arrays were provided directly by NimbleGen. The ALEXA design contained ~4,300 randomly generated probe sequences and the Affymetrix design contained ~17,000 'anti-genomic' probes. In both platforms these probes were selected to uniformly represent the Tm of all experimental probes. For each array hybridization on both platforms, a loess model[39] was fit to a plot of probe intensity versus Tm for all random probes. A Tm-specific estimate of background hybridization was then estimated for every probe on the array by interpolating from the loess model fit. This value was subtracted from the observed intensity and an arbitrary value of 16 was added according to Affymetrix's recommendations for stabilizing variance. The data were then normalized across the arrays within each platform by quantiles normalization[40]. Differential expression (DE) or 'fold-change' values for probesets (each corresponding to an exon, intron or junction) were calculated by taking the mean of individual probe intensities for each probeset, taking the mean of the probeset means across biological triplicates, transforming to a log2 scale and calculating the log2 difference between 5-FU sensitive and resistant cells (sensitive minus resistant). DE values for entire genes were calculated in a similar fashion by combining the probe intensities for all exons of each gene. For the ALEXA platform, both exon and canonical junction probes were considered when estimating expression of the entire gene.

2.4.6. Platform comparisons

The ability of ALEXA and Affymetrix arrays to measure the expression of individual exons and genes was compared and the potential for the ALEXA approach to provide additional information on the connectivity and boundaries of exons was examined.

Receiver operator characteristic (ROC) curves were generated to compare the sensitivity and specificity of each platform for correctly classifying exons and introns corresponding to ~100 housekeeping genes profiled on both platforms. ROC curves were generated by applying an expression level cutoff and determining the proportion of exon probes correctly identified as expressed (sensitivity) and the proportion of intron probes correctly identified as not expressed (specificity). 1000 such cutoffs were chosen across the range of expression observed for each platform and the resulting

sensitivity and specificity values were used to generate ROC curves. The area under each curve was calculated by trapezoid rule integration.

To allow comparisons of expression and DE values for all exons represented on the ALEXA array, common probesets were first identified by mapping probes from both platforms to EnsEMBL exons (version 35). Only those exons with at least one probe in both platforms were compared. Each exon typically had three probes in ALEXA and four in Affymetrix. The resulting common probesets were used to compare absolute and differential expression values. The overlap between platforms for exons and genes identified as significantly DE was also determined.

2.4.7. Visualization

To facilitate manual examination of expression and DE, values for all probes were divided into 20 quantile bins and plotted as individual custom UCSC tracks[41] (See **Figure 2.9** - **Figure 2.13** for examples). Expression or differential expression values are displayed on a log2 scale and positioned and shaded according to their magnitude. The top of each set of tracks displays the gene model which is also shaded according to the expression or DE value of the entire gene. The genomic position of predicted ORFs and protein features such as signal peptides, transmembrane domains, coiled coils, and protein family motifs were also added for reference. Custom tracks representing replicate and mean data were generated for every EnsEMBL gene profiled (www.AlexaPlatform.org).

2.4.8. Identification of significant differential expression events

Significant DE events were identified by comparing the population of individual probes of each probeset (exon, intron or junction) between biological triplicates of 5-FU sensitive and resistant cells (typically 3 probes × 3 replicates or 9 values for each condition). P-values were calculated by a Wilcoxon rank sum test. The probability of an event being DE between conditions was expected to depend on the probe type. For example, many exon-exon junction probes are only predicted sequences which may or may not occur in the transcriptome and therefore may not be expressed in either condition. For this reason, DE events were summarized separately for each probe type (**Table 2.3**). Probes without evidence of expression in one or both conditions were filtered before statistical testing. Specifically, a probeset was required to have a mean

log2 expression value greater than the 97.5th percentile of all negative control probes (~8) in either 5-FU sensitive or resistant cells. Multiple testing problem (MTP) correction was applied to the filtered list. Events with a fold-change of two or greater and a MTP corrected p-value < 0.05 were considered significant (see below for details).

2.4.9. Gene ontology analysis

Tests for statistical enrichment of particular Gene Ontology terms associated with the list of genes identified as DE were conducted with GOstat[42].

2.4.10. Identification of putative alternative expression events

To identify events potentially indicating differential expression of specific isoforms, probesets were filtered as above to eliminate those with low expression. A splicing index value was then calculated to estimate the differential expression of each probeset after normalization to account for DE of the entire gene. The splicing index was calculated as: $SI_i = \log2((eS_i/gS_j)/(eR_i/gR_j))$, for the i-th exon (e) of the j-th gene (g) in 5-FU sensitive (S) and resistant (R) samples. A Wilcoxon test was then applied to test for differences in the SI values for a particular probeset between sensitive and resistant cells (all probes across all replicates). Probesets with a significant Wilcoxon p-value (< 0.05), an SI value > 1 and an absolute difference between their SI and gene level DE > 1.56 (abs(SI - DE)) were selected as putative differential alternative expression events. The resulting list was ranked according to abs(SI-DE) to identify possible cases of reciprocally expressed isoforms. The list of 'top' candidate isoforms was selected from this list by manual examination of data displayed in custom UCSC tracks corresponding to the genomic loci implicated. Each event was classified as 'alternative TSS/polyA', 'alternative exon boundary', 'intron retention', 'exon skipping' and 'complex' (a combination of the other classes). EST and mRNA support was determined by BLAST of all probe sequences to ESTs and mRNAs that map within the target locus of the probe sequence according to UCSC's human 'all_est' and 'all_mrna' SQL tables[41]. Hits of 95% of the length of the probe or greater were considered to be a supporting match. EST and mRNA support was also visually confirmed using custom tracks of expression data in the UCSC browser. Cross-platform validation of an alternative expression event was also determined by manual examination of data from both platforms. In some

cases the necessary probes to allow cross-platform comparison were not available on the Affymetrix array. The fold-change values of putative alternate isoforms (described as fold-change 'isoform A' and 'isoform B' in **Table 2.5**) were determined by manually grouping all probes which correspond to each putative isoform.

2.4.11. Statistical analysis

Correlations and statistical tests were conducted with the programming language R[43] and 'Bioconductor'[44]. All statistical tests were two-tailed. When comparing absolute expression values between platforms or within replicates of a platform, Spearman's rank correlation coefficients were determined. When comparing differential expression values, Pearson correlation coefficients were reported. Comparisons of population means were conducted by a Student's t-test only when the assumption of normality could be satisfied by visual examination of Q-Q plots and the Anderson-Darling[45] test for normality. Otherwise, a non-parametric Wilcoxon rank sum test was used. When identifying significant differential expression events, the sample size for each event often consisted of 6-12 observations for each condition (sensitive and resistant). Reliable demonstration of normality was not possible with samples of this size and therefore the Wilcoxon test was always used for these comparisons. MTP correction was accomplished by Benjamini and Hochberg's step-up false discovery rate controlling procedure[46] using the 'multtest' package of R. An MTP corrected p-value < 0.05 was considered significant. To test for enrichment of particular types of annotations (protein features, etc.) in the group of events identified as significantly DE compared to the total population of events, a Chi-Squared test was used if the assumption of normality could be verified, otherwise a Fisher's Exact test was used.

Figure 2.1. Types of alternative expression and corresponding microarray probe design strategies

(**a**) Alternative expression (alternative transcript initiation, splicing, and polyadenylation). A hypothetical gene locus with four annotated exons (colored rectangles; E1-E4) and three introns (connecting lines; I1-I3) is depicted. Green arrows indicate alternate transcript start sites (TSS). Alternate polyadenylation (polyA) sites are shown in red. Alternative exon boundary usage, exon skipping and intron retention are depicted with black dotted lines. (**b**) Affymetrix array design. Affymetrix exon arrays use multiple sources of gene annotation and prediction in an attempt to measure expression of every known or predicted expressed region of the genome. The resulting design consists of sets of 4 oligonucleotide probes per exon representing most known and predicted exons. (**c**) ALEXA array design. The ALEXA approach attempts to profile exon skipping, alternative exon boundary usage, and intron retentions by selecting probes to represent every exon, intron, exon-exon junction and exon-intron boundary. The positions of exon junctions are depicted over the hypothetical processed mRNAs they represent.

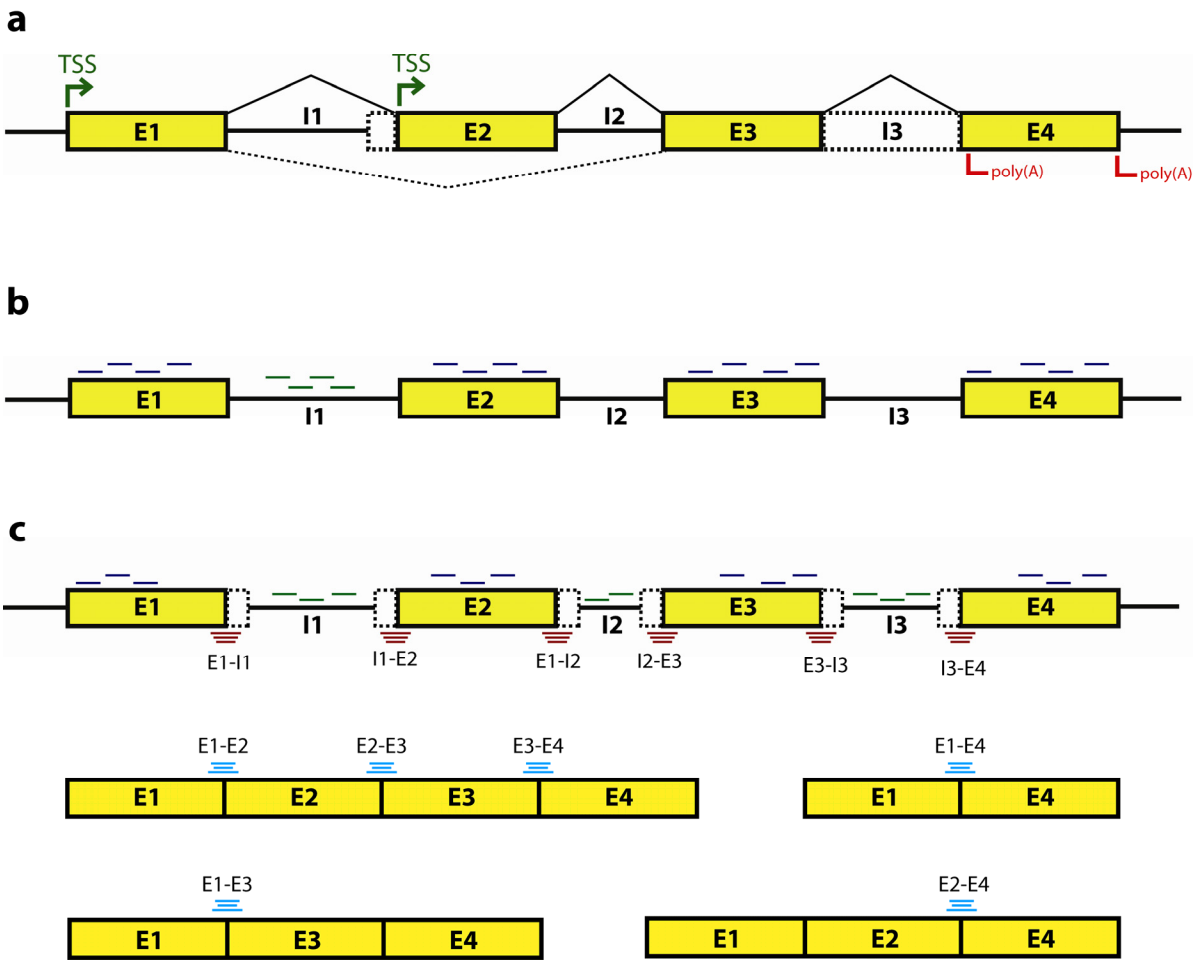


Figure 2.2. ROC curves for ALEXA and Affymetrix control probes
Receiver operator characteristic (ROC) curves describing the sensitivity and specificity of Affymetrix and ALEXA platforms. ROC curves were generated by examining data for ~100 housekeeping genes targeted by both platforms. Probes were designed for all exons (+ve controls) and introns (-ve controls) of these genes. The intensities observed for these control probes were used to calculate sensitivity and specificity scores (see **Methods**). The standard deviation of these scores across six array hybridizations are indicated by dotted lines. The data for each platform were subjected to quantiles normalization before conducting this analysis.

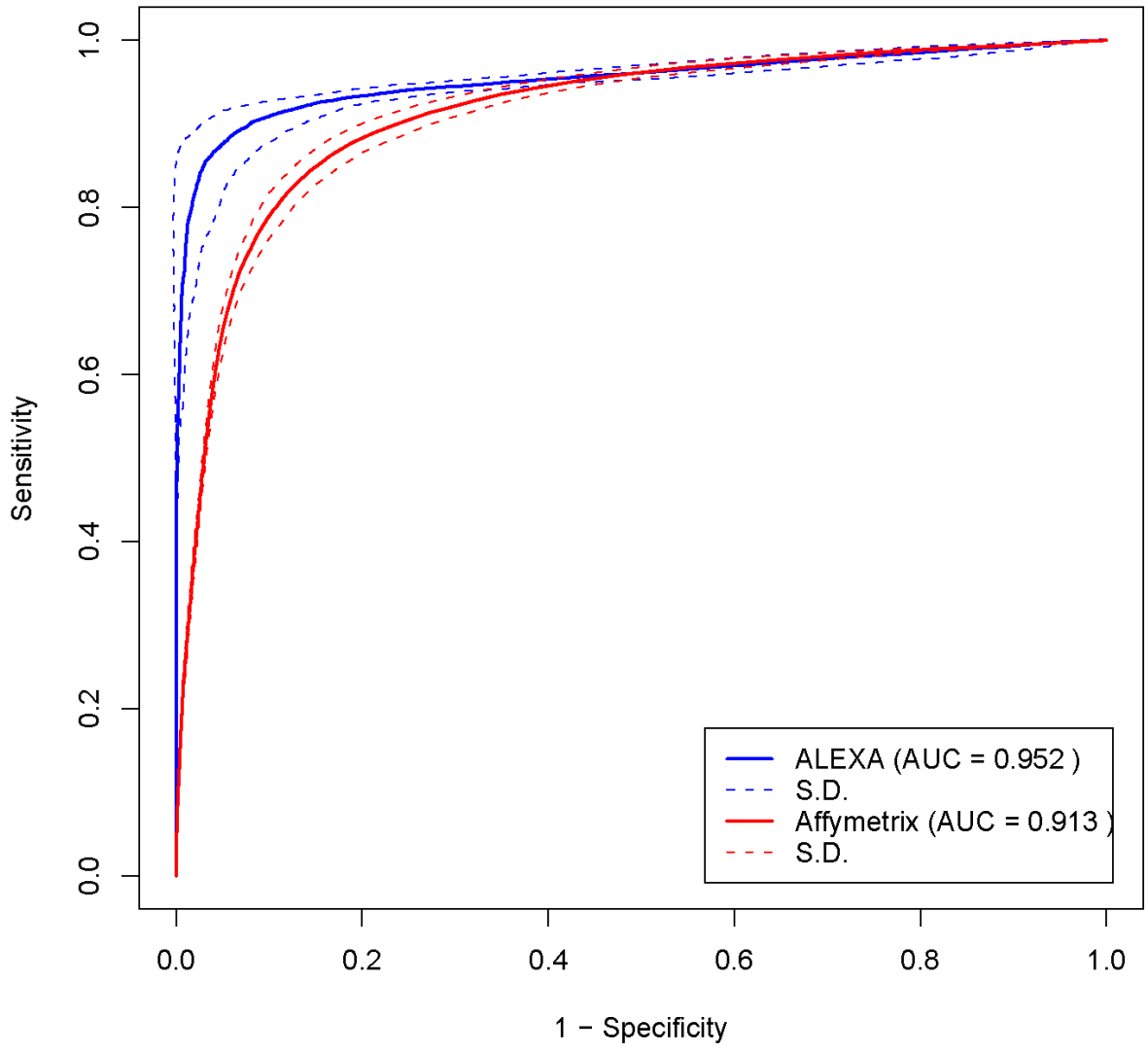


Figure 2.3. Correlation of ALEXA and Affymetrix gene differential expression values

A density plot of differential gene expression values for 2,507 genes from the ALEXA and Affymetrix platforms. Gene expression estimates are derived by calculating the mean expression of all exons of the gene across three biological replicates. Differential expression is the observed log2 gene expression value in 5-FU sensitive cells minus that in 5-FU resistant cells. Positive values indicate over-expression in 5-FU sensitive cells and negative values indicate over-expression in 5-FU resistant cells. The Pearson correlation coefficient between the two platforms is 0.87.

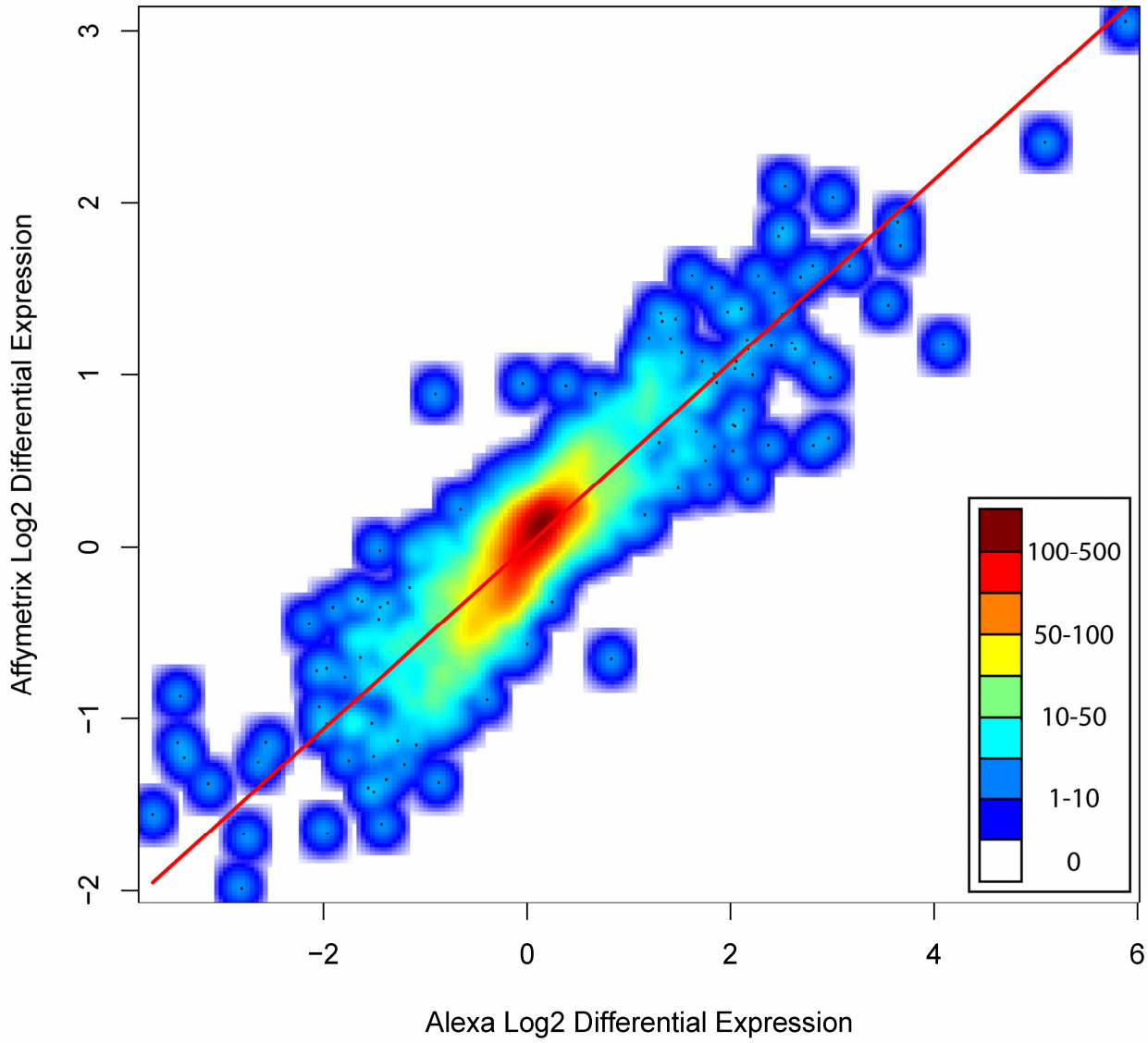


Figure 2.4. Correlation of ALEXA and Affymetrix exon differential expression values

A density plot of differential exon expression values for 31,368 exons from the ALEXA and Affymetrix platforms. The Pearson correlation coefficient for this comparison is 0.67. Positive values indicate over-expression in 5-FU sensitive cells. Negative values indicate over-expression in 5-FU resistant cells.

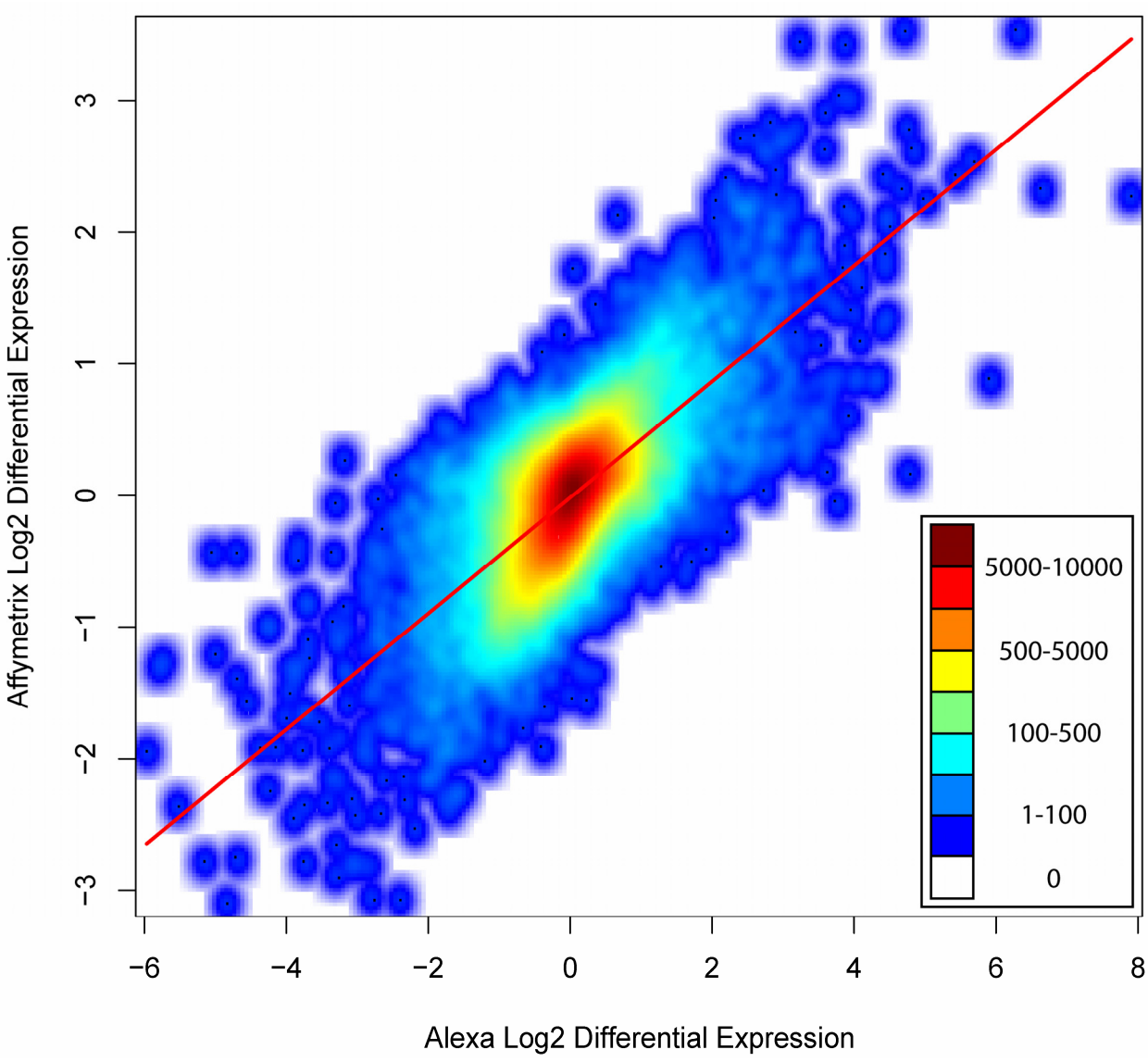


Figure 2.5. Overlap between Affymetrix and ALEXA gene and exon differential expression events

Comparison of ALEXA and Affymetrix expression profiling results. (**a**) Overlap between genes identified as differentially expressed (DE) using the ALEXA and Affymetrix platforms. This analysis was limited to the 2,507 genes covered by both platforms. 835 (33%) of these genes were identified as significantly DE in at least one of the two platforms. The observed overlap of 482 genes is statistically significant ($P < 1\times10^{-5}$ by permutation test; overlap of 175 is expected by chance). (**b**) Overlap between exons identified as DE in ALEXA and Affymetrix platforms. This analysis was limited to only those 31,368 EnsEMBL exons that were represented by at least 1 probe in both platforms. 3,367 (~11%) of these exons were identified as significantly DE in at least one of the two platforms. The observed overlap of 516 exons is statistically significant ($P < 1\times10^{-5}$ by permutation test; an overlap of 90 is expected by chance). These exons correspond to 213 genes.

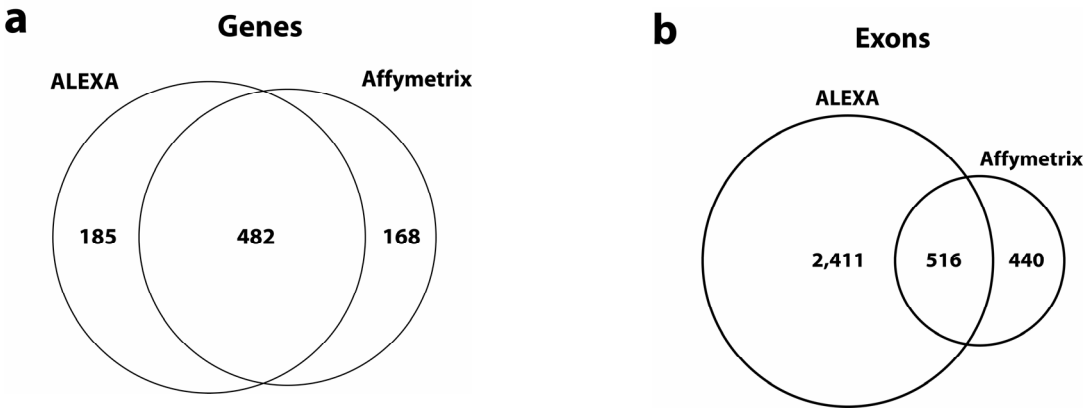


Figure 2.6. Exons identified as differentially expressed by ALEXA but not Affymetrix are biased towards low levels of detected expression in the Affymetrix data

Box plots of the observed expression level in Affymetrix data for exons identified as significant by the ALEXA platform only, by both platforms, or by the Affymetrix platform only. Note that those exons identified as significant by the ALEXA platform only, have low expression values in the Affymetrix data. This poor detection level is likely why these exons were not identified as significantly DE in the Affymetrix data. The expression values of exons identified as DE by ALEXA but not Affymetrix are significantly different from those identified by Affymetrix only (p-value = 1.143×10^{-13} by Wilcoxon rank sum test). It is not surprising that each platform has some difficulty in detecting some exons, but our analysis suggested that the ALEXA data detected ~3 times as many DE exons while also having a higher specificity (i.e. a lower false positive rate, according to the detection of introns for housekeeping genes).

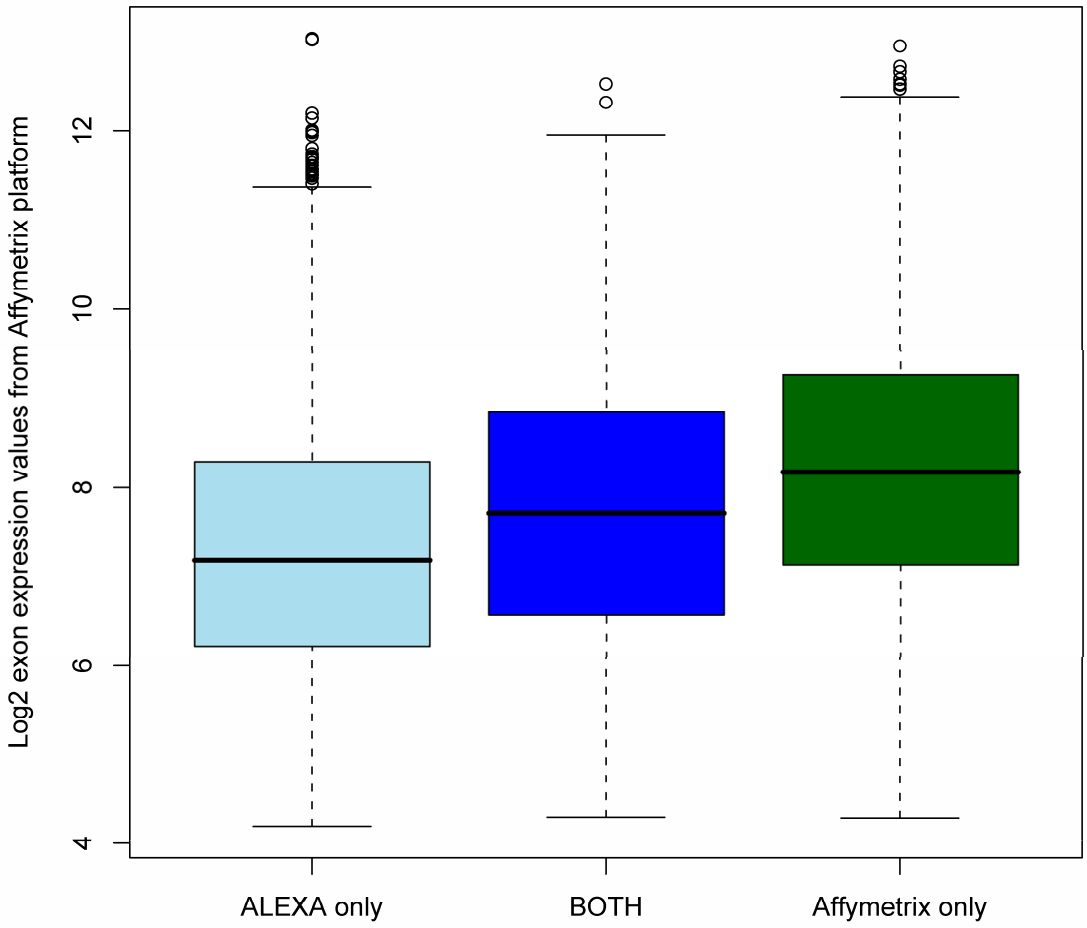


Figure 2.7. Absolute gene expression values in ALEXA and Affymetrix data
A density plot comparing expression estimates between Affymetrix and ALEXA data for 2,507 genes. The mean expression value for most genes is higher in ALEXA compared to Affymetrix. The range of expression values is also wider for the ALEXA data (~4 to ~16 for ALEXA and ~4.5 to ~14 for Affymetrix). The overall trend of the data is represented by a 'lowess' fitted line (red dotted line). Due to the number of technical differences between ALEXA and Affymetrix experiments it is difficult to attribute this difference in dynamic range to any one parameter (see **Methods**). Some of the possible explanations for the observed difference include: the increased length of oligonucleotides used in ALEXA microarrays (36 +/- 10 nucleotides) compared to Affymetrix microarrays (25 nucleotides); the use of poly-A purified RNA in ALEXA hybridizations compared to total RNA in Affymetrix; and the use of unamplified samples in ALEXA hybridizations compared to the use of samples amplified by in vitro transcription for Affymetrix hybridizations.

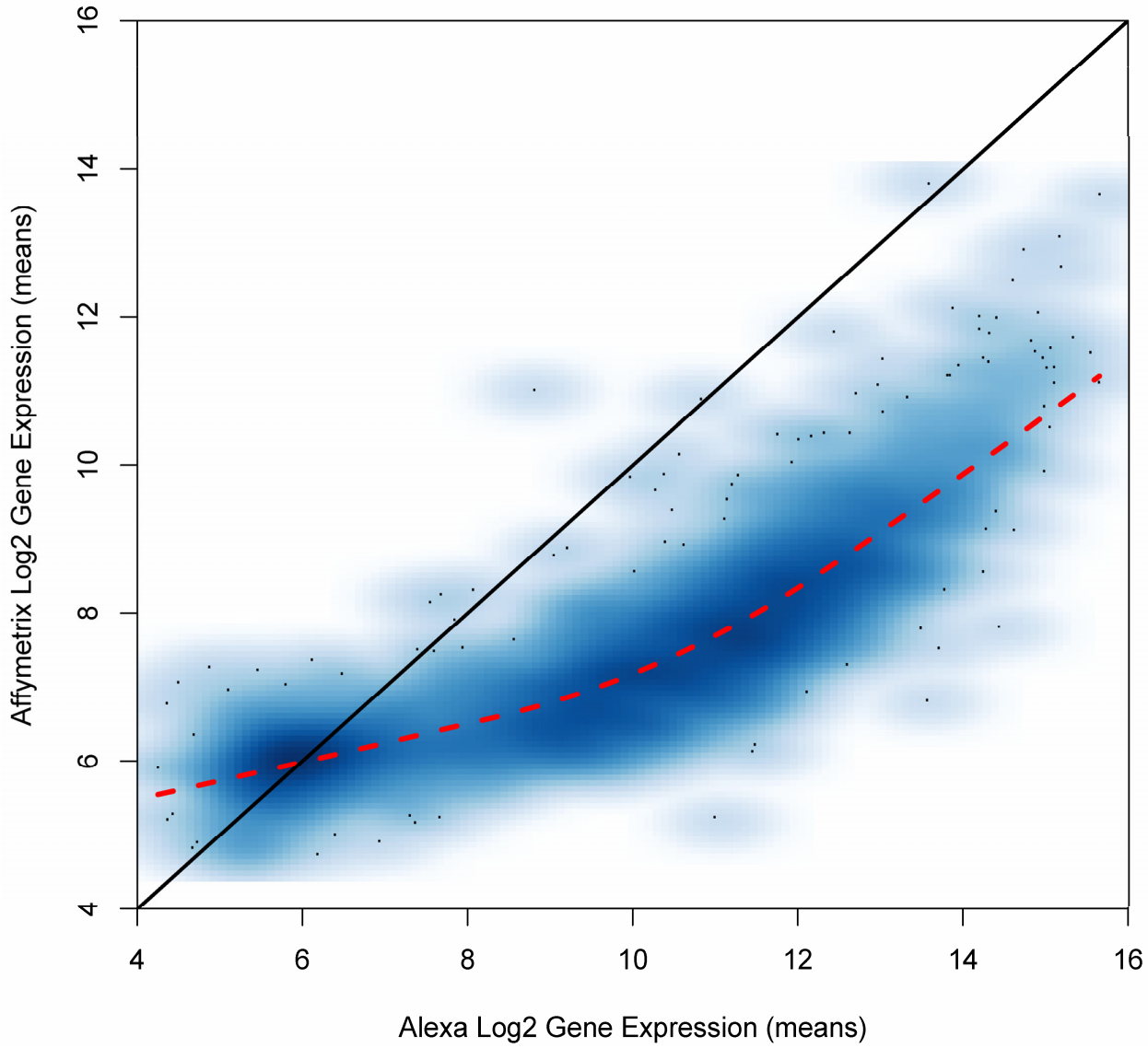


Figure 2.8. Absolute exon expression values in ALEXA and Affymetrix data
A density plot comparing expression estimates between Affymetrix and ALEXA data for 31,368 exons. The mean expression value for most exons is higher in ALEXA compared to Affymetrix. The range of expression values is also wider for the ALEXA data (~4 to ~16 for ALEXA and ~4 to ~14 for Affymetrix). The overall trend of the data is represented by 'lowess' fitted line (red dotted line).

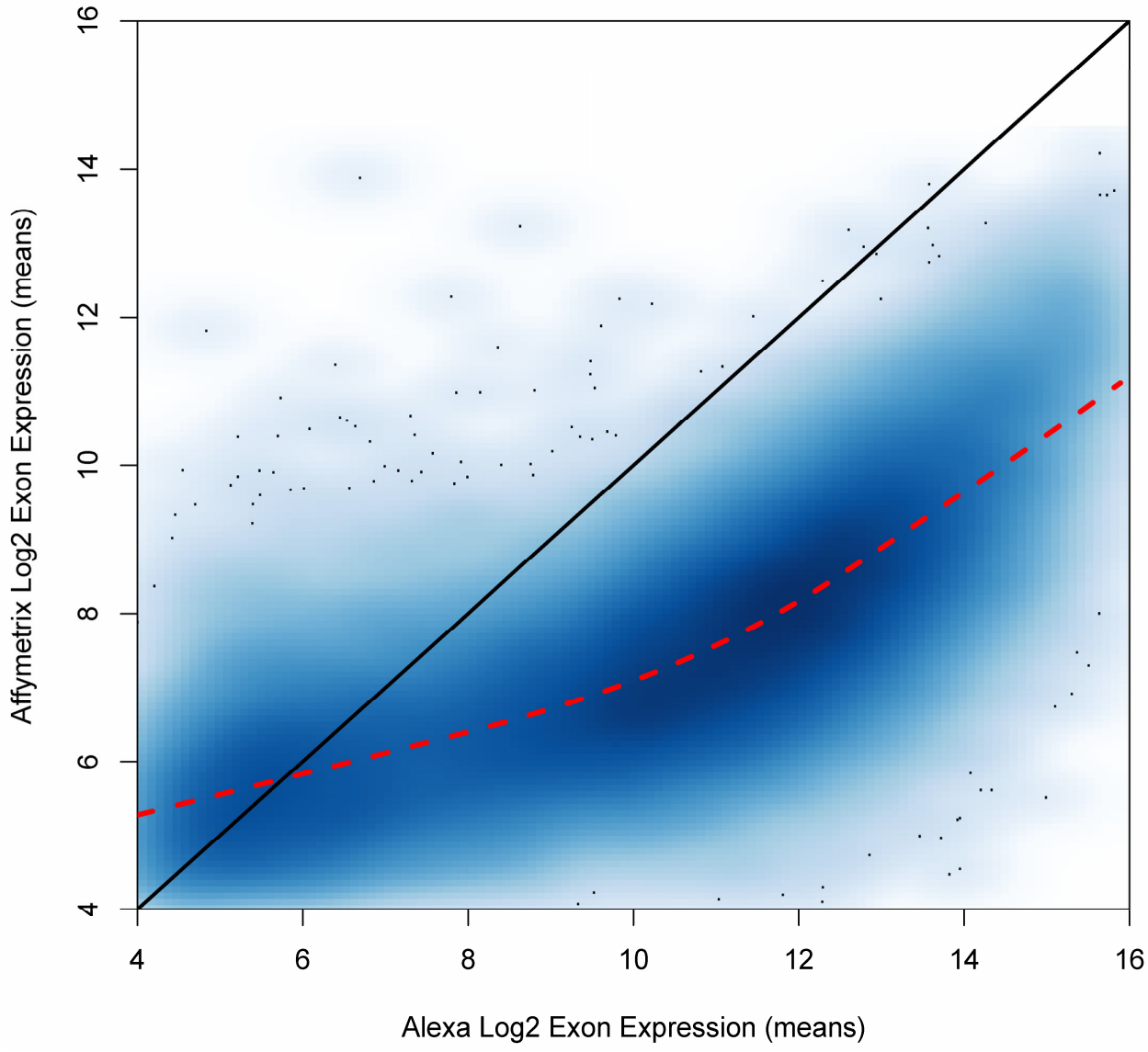


Figure 2.9. The *OLR1/c12orf59* locus is differentially expressed between sensitive and resistant cells

Differential expression† of *OLR1* and *c12orf59*. (**a**) These genes are expressed in a 'head-to-head' fashion and their transcription start sites are separated by ~7,500 bp. *OLR1* is a putative apoptosis gene and *c12orf59* has unknown function (see **Table 2.4** for details). (**b**) DE values for each exon and canonical junction probeset are positioned and colored according to their magnitude on a log2 scale (to simplify this figure other probeset types are not depicted). The red dotted line indicates a fold-change of 0. *OLR1* and *c12orf59* were over-expressed in 5-FU sensitive cells relative to resistant cells with an average fold change of ~40 and fold-changes ranging from ~8 to ~300 for individual probesets. (**c**) The position of predicted open reading frames (ORF), transmembrane domains (TMD), coiled-coil domains (CC) and transcription factor binding sites (GFI1, Oct1 and GATA1) are depicted (position of binding sites was obtained from the UCSC genome browser track displaying data from the Transfac Matrix Database).

† Data shown are from the ALEXA platform only.

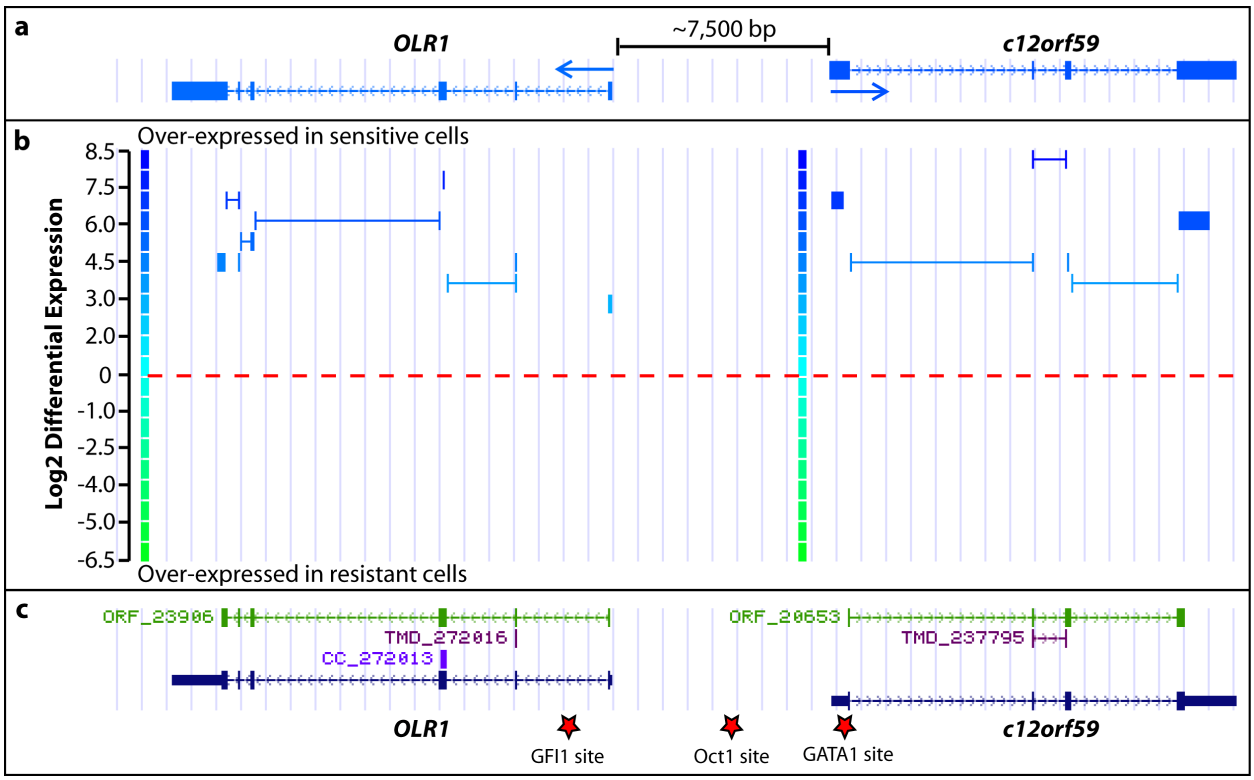


Figure 2.10. A known isoform of *LAMA3* is over-expressed in 5-FU resistant cells
Differential expression† of *LAMA3*. (**a**) *LAMA3* (also known as laminin, alpha 3) is a gene that is thought to be involved in cell adhesion, signal transduction and differentiation (see **Table 2.5** for details). (**b**) DE values for each exon and canonical junction probeset are positioned and colored according to their magnitude on a log2 scale (to simplify this figure other probeset types are not depicted). The red dotted line indicates a fold-change of 0. DE of a known isoform of *LAMA3* was observed. Although exons representing both known isoforms showed evidence of expression in both sensitive and resistant cells, only those exons corresponding to the short isoform were over-expressed in resistant cells (indicated by a dotted blue box). This suggests the possibility that only the transcript initiated at the second known promoter site of *LAMA3* is up-regulated in 5-FU resistant cells. (**c**) The position of known alternative transcript sequences are indicated (RefSeq mRNAs).

† Data shown are from the ALEXA platform only.

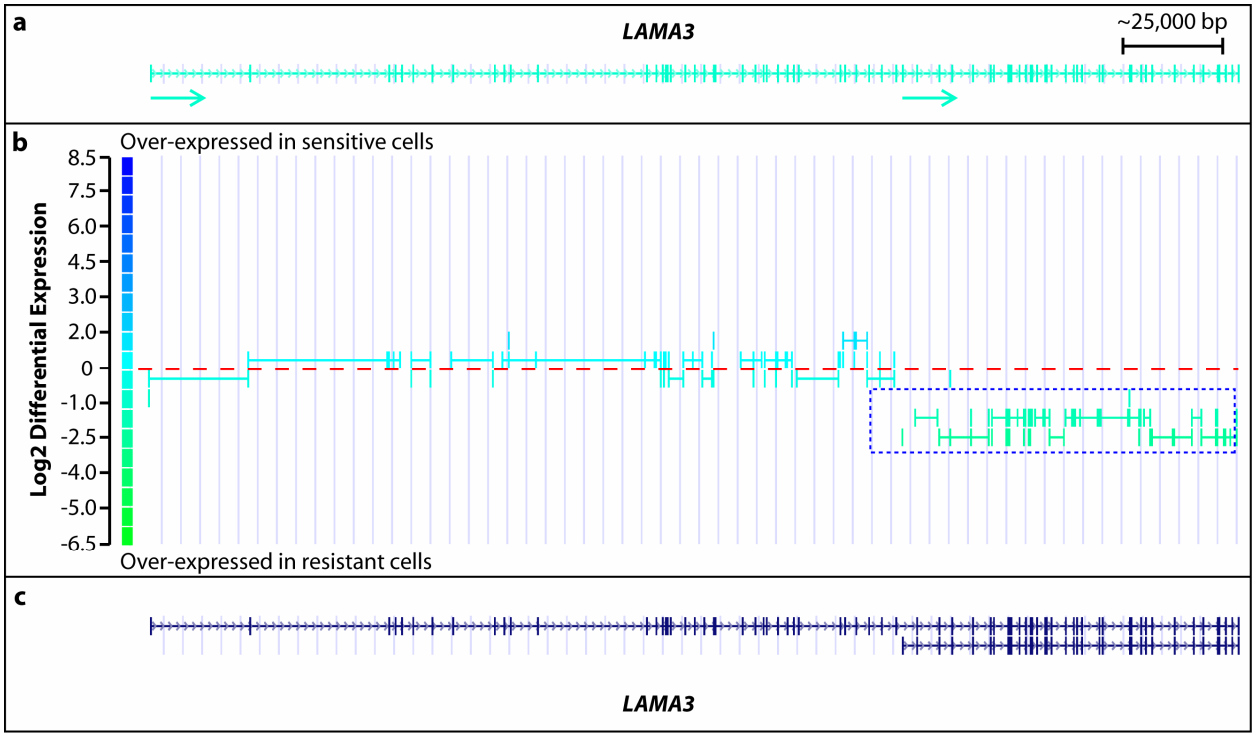


Figure 2.11. The last 5 exons of *EPB41L3* are over-expressed in 5-FU resistant cells

Differential expression† of *EPB41L3*. (**a**) *EPB41L3* (also known as *Dal1*), a gene consisting of 22 exons, is thought to be involved in cell adhesion and cancer progression (see **Table 2.5** for details). (**b**) DE values for each exon and canonical junction probeset are positioned and colored according to their magnitude on a log2 scale (to simplify this figure other probeset types are not depicted). The red dotted line indicates a fold-change of 0. DE of a candidate novel isoform of *EPB41L3* is predicted. Most exons are not DE between sensitive and resistant cells. The last 5 exons (indicated by a dotted blue box) are ~16-fold up-regulated in resistant cells relative to sensitive cells. The remaining exons of this locus are not highly expressed in either cell line. This suggests the possibility that a previously unknown transcript consisting of the last 5 exons of *EPB41L3* is up-regulated in 5-FU resistant cells. Possible mechanisms which could underlie this event include use of an alternate transcription start site or polyadenylation site, deletion, amplification, rearrangement, etc. (**c**) The position of predicted ORFs (green) and known alternative transcripts (red) are indicated.

† Data shown are from the ALEXA platform only.

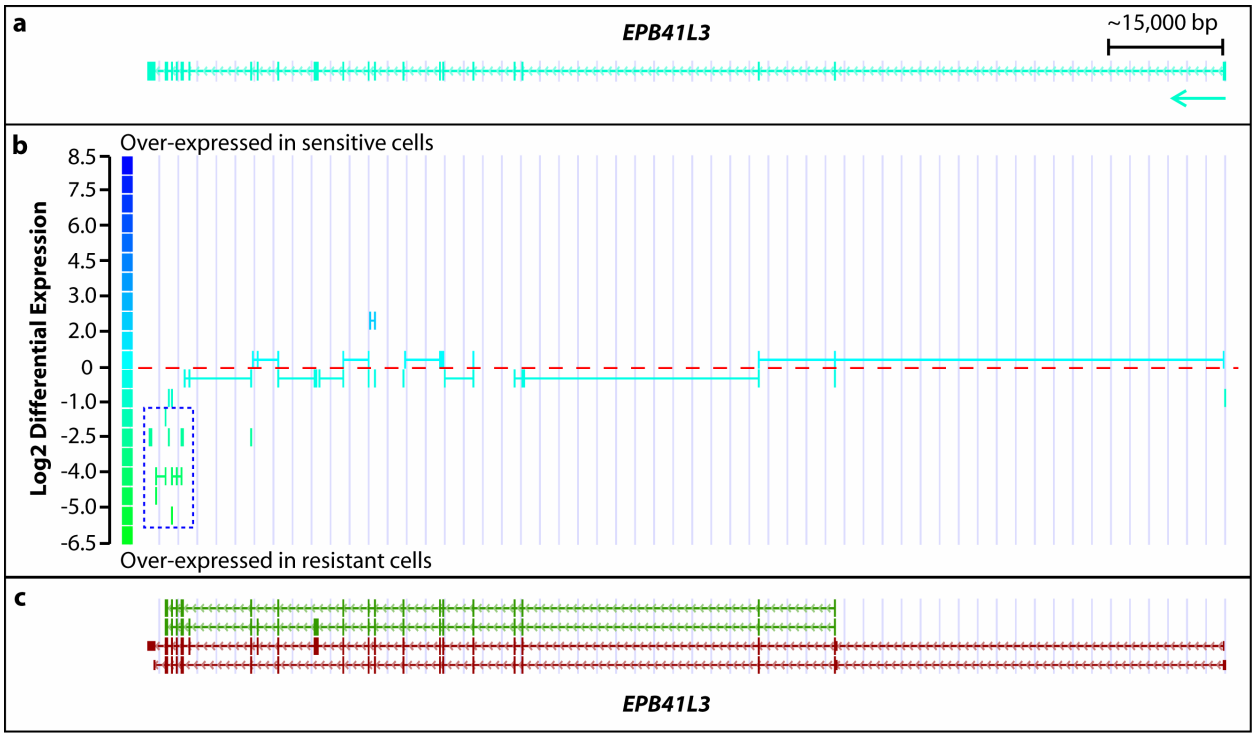


Figure 2.12. The last 9 exons of the predicted protein *c12orf63* are over-expressed in resistant cells

Differential expression† of *c12orf63*. (**a**) *c12orf63*, a gene consisting of 26 exons with unknown function (see **Table 2.5** for details). (**b**) DE values for each exon and canonical junction probeset are positioned and colored according to their magnitude on a log2 scale (to simplify this figure other probeset types are not depicted). The red dotted line indicates a fold-change of 0. DE of a candidate novel isoform of *c12orf63* is predicted. Most exons are not DE between sensitive and resistant cells. The last 9 exons (indicated by a dotted blue box) are ~16-fold up-regulated in resistant cells relative to sensitive cells. The remaining exons of this locus are not highly expressed in either cell line. This suggests the possibility that a previously unknown transcript consisting of the last 9 exons of c12orf63 is up-regulated in 5-FU resistant cells. Possible mechanisms which could underlie this event include use of an alternate transcription start site or polyadenylation site, deletion, amplification, rearrangement, etc. (**c**) The position of the predicted ORF (green) is indicated. Position of a coiled-coil domain is indicated (purple).

† Data shown are from the ALEXA platform only.

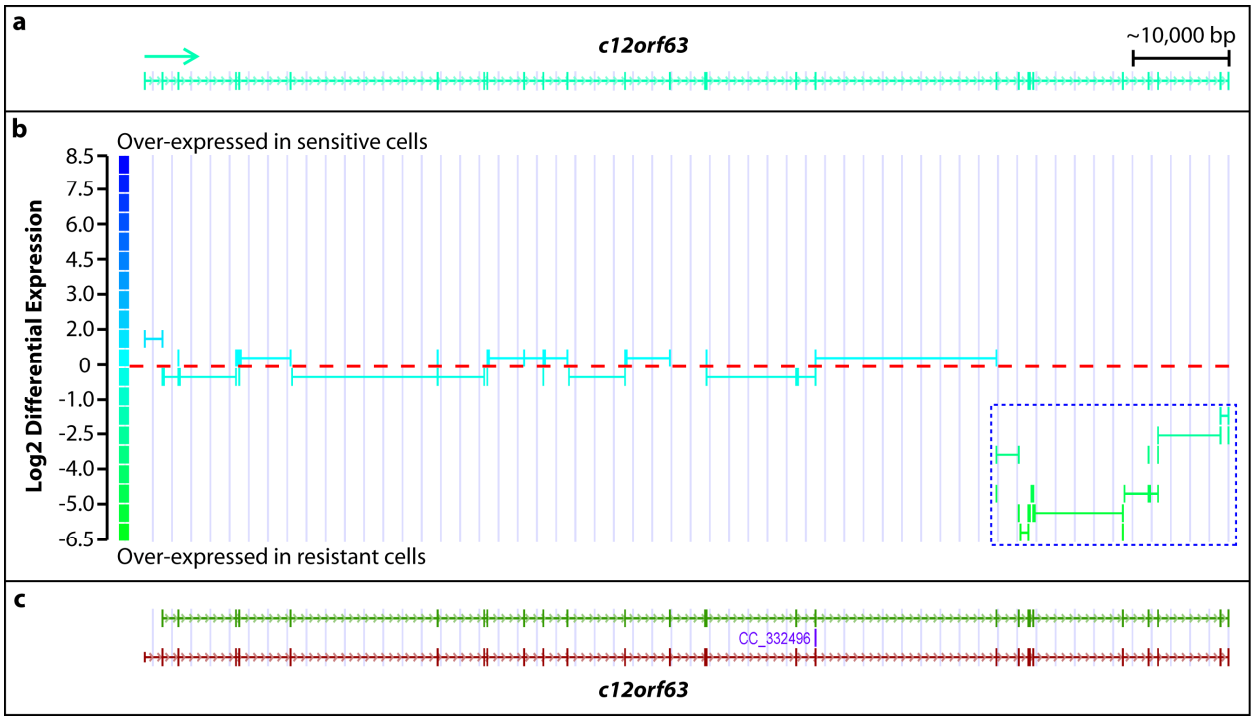


Figure 2.13. Reciprocal DE of *UMPS* isoforms
(**a**) The positions of ALEXA probesets (each consisting of 2-4 oligonucleotide probes) specific to each of the two *UMPS* isoforms are depicted. Probes are labeled according to the exons or junctions they profile (e.g. E1-E3 detects the connection of exon 1 to exon 3). Black arrows indicate the effect of exon skipping on the predicted ORF and the position of known protein domains is indicated. (**b**) ALEXA log2 expression values for the probes specific to each isoform from triplicate samples of each cell line are shown as box plots. The median log2 expression value of all exons (blue dotted line) and all negative controls (red dotted line) on the ALEXA microarray are also shown. Isoform A was ~5-fold over-expressed in 5-FU sensitive cells relative to resistant cells. Isoform B was ~6-fold over-expressed in 5-FU resistant cells relative to sensitive cells. (**c**) Affymetrix log2 expression values for the probes specific to Isoform A from the same triplicate samples. The Affymetrix design did not contain exon junction probes specific to either isoform and therefore could not detect Isoform B.

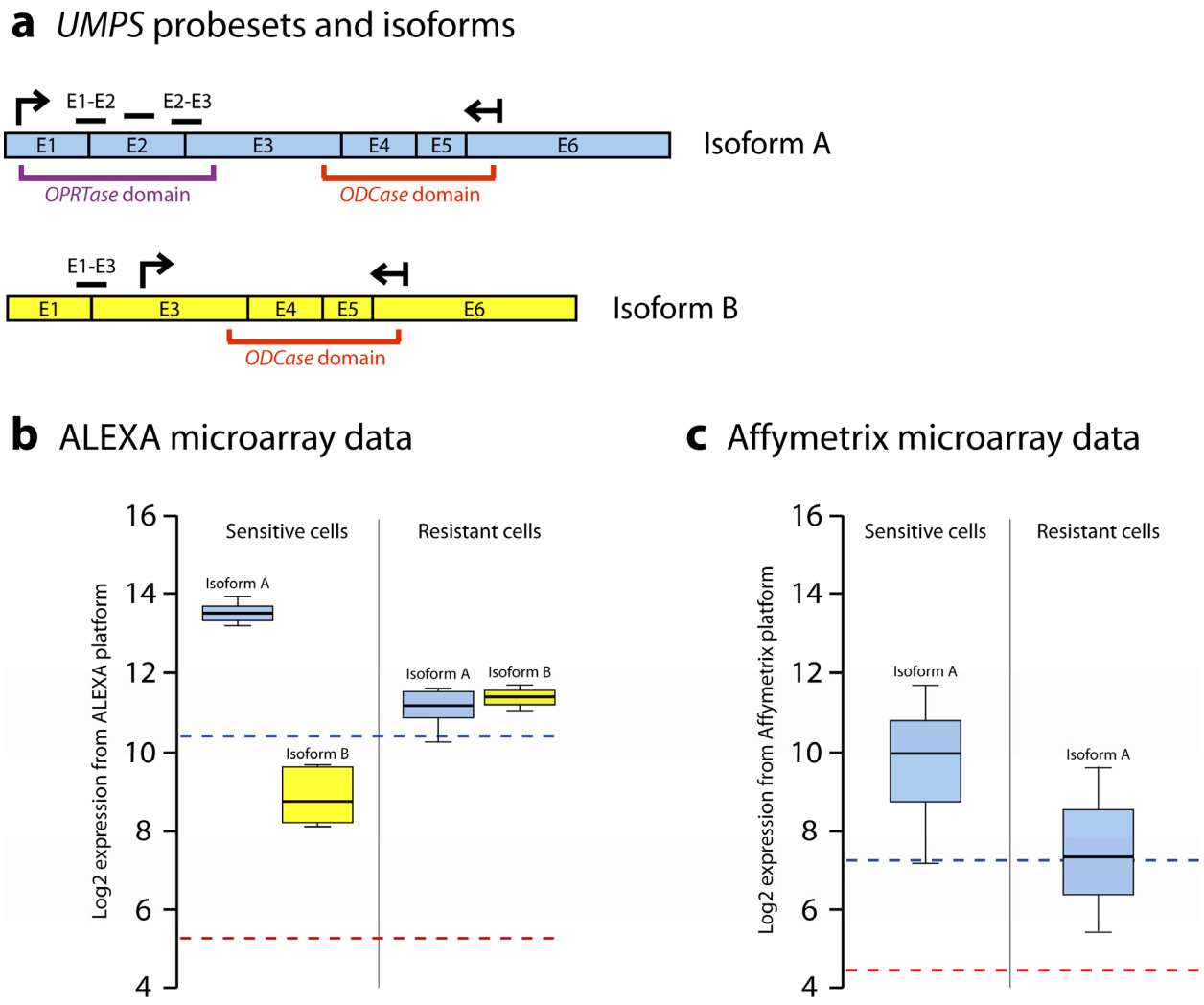


Table 2.1. Summary of pre-computed ALEXA designs.
Summary statistics for ten pre-computed ALEXA designs. Additional information is available for each of these designs on the ALEXA website (www.AlexaPlatform.org). Abbreviations: Exon-Junction (EJ); Exon-Boundary (EB); Exon (E); Intron (I). †Filtering as described in the **Methods** followed by selection of 3 probes per exon and only exon-junction probes representing 3 exons skipped or less. Intron probes were only selected for a set of ~100 housekeeping genes.

Species	# Genes targeted	# Probes	# Probes after filtering / selection†	Probe length (bp)	Target Probe Tm (ºC)	Distribution of Probe types
Canis familiaris	14,121	8,340,667	2,109,103	38 +/- 10	67.9	54.2% EJ, 29.7% EB, 15.9% E, 0.1% I
Caenorhabditis elegans	20,049	6,063,626	1,556,131	42 +/- 10	67.9	48.2% EJ, 29.9% EB, 21.8% E, 0.1% I
Drosophila melanogaster	14,649	5,615,401	700,640	36 +/- 10	67.3	45.8% EJ, 33.2% EB, 20.7% E, 0.3% I
Danio rerio	16,304	8,116,906	1,752,876	38 +/- 10	67.5	52.1% EJ, 31.3% EB, 16.4% E, 0.2% I
Gallus gallus	17,262	10,077,586	2,341,085	38 +/- 10	67.1	54.0% EJ, 30.9% EB, 14.9% E, 0.2% I
Homo sapiens	22,687	15,621,632	3,071,445	38 +/- 10	68.1	52.5% EJ, 29.2% EB, 18.2% E, 0.1% I
Mus musculus	22,283	13,995,763	2,998,157	36 +/- 10	67.3	52.5% EJ, 29.9% EB, 17.5% E, 0.1% I
Pan troglodytes	24,482	13,431,332	2,912,409	38 +/- 10	67.9	53.1% EJ, 28.9% EB, 17.9% E, 0.1% I
Rattus norvegicus	18,799	11,939,250	2,818,987	36 +/- 10	67.2	52.2% EJ, 30.5% EB, 17.2% E, 0.1% I
Saccharomyces cerevisiae	6,678	1,527,330	25,926	42 +/- 10	66.6	1.1% EJ, 2.6% EB, 93.0% E, 3.3% I

Table 2.2. Within platform reproducibility for biological replicates
Mean correlation coefficients representing the reproducibility of biological replicates for expression and DE values are reported for genes and exons. For each platform, the correlation coefficient for pairwise comparisons of three 5-FU sensitive replicates and three 5-FU resistant replicates were calculated. The values reported in the table below are the mean of these coefficients +/- standard deviation. P-values correspond to a two-tailed Wilcoxon rank sum test of the difference between the correlation coefficients determined for ALEXA and Affymetrix platforms. Spearman's rank correlation coefficients are reported for absolute expression values and Pearson correlation coefficients are reported for DE values.

Data Type	ALEXA Platform	Affymetrix Platform	P-value
Gene Expression	0.9904+/-0.0020	0.9732+/-0.0142	0.002165
Gene DE	0.8451+/-0.0058	0.8020+/-0.0519	0.7
Exon Expression	0.9770+/-0.0078	0.9544+/-0.0164	0.004329
Exon DE	0.5936+/-0.0190	0.5554+/-0.0718	0.7

Table 2.3. Summary of differential expression events for genes profiled by the Affymetrix and ALEXA array platforms
Significant events had a multiple testing corrected p-value < 0.05 and a fold-change > 2. 'Canonical junction' refers to the connection of adjacent exons.

	DE event type	Total # events profiled	# Significant DE events	# Within ORF	# Affecting known feature (domain, signal peptide, etc.)
Affymetrix	Gene-level	2,507	78	N/A	N/A
	Exon	49,681	1,117	978	589
	Intron	65,327	25	20	0
	Total	117,515	1,220	998	589
ALEXA	Gene-level	2,507	233	N/A	N/A
	Exon	32,164	2,703	2,537	1,544
	Canonical junction	27,046	2,310	2,260	1,277
	Exon skip	69,761	191	180	103
	Exon boundary	52,402	253	219	100
	Intron	472	0	0	0
	Total	184,354	5,690	5,196	3,024

Table 2.4. Candidate differential gene expression events associated with 5-FU resistance

The top 46 candidate differential gene expression events associated with 5-FU resistance are reported. Genes selected for this table had a significant p-value for differential expression (multiple testing corrected p-value < 0.05) and a fold-change of 4 or greater in either ALEXA or Affymetrix data. Positive values indicate over-expression in 5-FU sensitive cells. Negative values indicate over-expression in 5-FU resistant cells. Additional information for each event listed below is available online (www.AlexaPlatform.org).

Rank	Gene Symbol	Fold Change (ALEXA)	Fold Change (Affy)	Gene name
1	*C12orf59*	50.5	24.8	Chromosome 12 open reading frame 59
2	*OLR1*	33.3	10.1	Oxidized low density lipoprotein (lectin-like) receptor 1
3	*PDZK1*	21.7	4.1	PDZ domain containing 1
4	*ASRGL1*	-18.9	-7.4	Asparaginase like 1
5	*KRT20*	-12.9	-3.4	Keratin 20
6	*IGF2BP3*	-12.3	-4.1	Insulin-like growth factor 2 mRNA binding protein 3
7	*GIPC2*	-12.0	-9.1	GIPC PDZ domain containing family, member 2
8	*ATOH8*	-11.9	-4.4	Atonal homolog 8 (Drosophila)
9	*PRF1*	10.9	6.1	Perforin 1 (pore forming protein)
10	*FUT3*	10.7	11.7	Fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group)
11	*SLAMF6*	-9.5	-4.7	SLAM family member 6
12	*ACSL4*	-8.2	-8.5	Acyl-CoA synthetase long-chain family member 4
13	*ARSE*	7.8	3.7	Arylsulfatase E (chondrodysplasia punctata 1)
14	*LAPTM4B*	7.2	8.1	Lysosomal associated protein transmembrane 4 beta
15	*C1orf25*	-7.1	-5.2	Chromosome 1 open reading frame 25
16	*PON3*	7.1	7.1	Paraoxonase 3
17	*MYEOV*	6.1	1.5	Myeloma overexpressed gene (in a subset of t(11;14) positive multiple myelomas)
18	*FBP1*	5.8	4.2	Fructose-1,6-bisphosphatase 1
19	*MALL*	-5.7	-1.9	Mal, T-cell differentiation protein-like
20	*MR1*	5.4	3.7	Major histocompatibility complex, class I-related
21	*TCF7L1*	5.4	1.3	Transcription factor 7-like 1 (T-cell specific, HMG-box)
22	*EPS8L3*	5.4	1.5	EPS8-like 3
23	*LRIG1*	-5.3	-3.0	Leucine-rich repeats and immunoglobulin-like domains 1
24	*PYGL*	5.2	4.1	Phosphorylase, glycogen; liver (Hers disease, glycogen storage disease type VI)
25	*H19*	5.1	15.5	H19, imprinted maternally expressed untranslated mRNA
26	*PHLDB2*	4.9	3.1	Pleckstrin homology-like domain, family B, member 2
27	*HHIP*	4.9	4.8	Hedgehog interacting protein
28	*SLC7A7*	4.8	6.0	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 7
29	*TNS4*	4.7	2.5	Tensin 4
30	*LXN*	4.6	3.2	Latexin
31	*RTN2*	-4.6	-2.6	Reticulon 2
32	*MLPH*	-4.6	-3.5	Melanophilin
33	*SPTLC3*	4.6	4.1	Serine palmitoyltransferase, long chain base subunit 3
34	*TTC14*	-4.5	-3.8	Tetratricopeptide repeat domain 14
35	*KLK6*	4.4	3.4	Kallikrein-related peptidase 6
36	*PTP4A3*	-4.4	-1.5	Protein tyrosine phosphatase type IVA, member 3

Rank	Gene Symbol	Fold Change (ALEXA)	Fold Change (Affy)	Gene name
37	*IRF8*	-4.3	-3.1	Interferon regulatory factor 8
38	*TMOD2*	-4.3	-2.6	Tropomodulin 2 (neuronal)
39	*PIK3AP1*	4.2	1.3	Phosphoinositide-3-kinase adaptor protein 1
40	*HSD3B1*	4.1	1.7	Hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1
41	*TMEM171*	-4.1	-3.4	Transmembrane protein 171
42	*CACNG4*	-4.1	-1.3	Calcium channel, voltage-dependent, gamma subunit 4
43	*COL4A1*	4.1	1.4	Collagen, type IV, alpha 1
44	*GRIN2B*	4.1	2.5	glutamate receptor, ionotropic, N-methyl D-aspartate 2B
45	*NDP*	-4.0	-2.6	Norrie disease (pseudoglioma)
46	*ABCA3*	-4.0	-1.8	ATP-binding cassette, sub-family A (ABC1), member 3

Table 2.5. Candidate differential isoform expression events associated with 5-FU resistance

The top 25 candidate differential isoform expression events associated with 5-FU resistance are reported (listed alphabetically). These events were selected by applying a series of filters to identify events that were likely to involve differential expression (reciprocal expression in most cases) of specific isoforms as opposed to the entire gene. Fold change 'A' and 'B' refer to values for probes capable of distinguishing expression of putative alternative isoforms (see **Methods**). Positive values indicate over-expression in 5-FU sensitive cells. Negative values indicate over-expression in 5-FU resistant cells. TSS refers to 'transcription start site'. Fold-change values for 'complex' events are listed as 'N/A' because the number of potential isoforms is too large to assign particular probes to particular isoforms. Additional information for each event listed below is available online (www.AlexaPlatform.org).

Gene Symbol	Gene name	Event type	Fold change (Isoform A)	Fold change (Isoform B)	# and type of sequences supporting alternative expression event
AKAP7	A kinase (PRKA) anchor protein 7	Alternative TSS/polyA	-3.07	1.01	Multiple mRNAs
APLP1	Amyloid beta (A4) precursor-like protein 1	Complex	N/A	N/A	mRNA and EST evidence indicates several AS events similar to those observed
ATP6AP1	ATPase, H+ transporting, lysosomal accessory protein 1	Alternative exon boundary	2.85	-1.67	Multiple ESTs
C12orf63	Chromosome 12 open reading frame 63	Alternative TSS/polyA	-21.17	1.23	None
C5	Complement component 5	Alternative TSS/polyA	13.22	-1.75	None
CDC25B	Cell division cycle 25 homolog B	Alternative TSS/polyA	6.31	1.00	Multiple mRNAs and ESTs support the use of alternative 5' exons
COL21A1	Collagen, type XXI, alpha 1	Intron retention	2.72	-3.10	Single EST
DIS3	DIS3 mitotic control homolog (S. cerevisiae)	Exon skipping	-3.44	1.02	Single mRNA and ~50 ESTs
EIF4A2	Eukaryotic translation initiation factor 4A, isoform 2	Alternative exon boundary	3.07	-1.20	Multiple ESTs.
ENO2	Enolase 2 (gamma, neuronal)	Intron retention	2.22	-3.44	Single EST (cloned from Cerebellum)
EPB41L3	Erythrocyte membrane protein band 4.1-like 3	Alternative TSS/polyA	-9.27	1.08	None
FGD5	FYVE, RhoGEF and PH domain containing 5	Alternative TSS/polyA	3.33	1.08	Single mRNA
HHIP	Hedgehog interacting protein	Alternative TSS/polyA	8.08	1.10	Single mRNA
IL10RB	Interleukin 10 receptor, beta	Exon skipping	4.69	-2.55	None
KLK6	Kallikrein-related peptidase 6	Complex	N/A	N/A	Multiple mRNAs which could contribute to observed expression
LAMA3	Laminin, alpha 3	Alternative TSS/polyA	-3.83	1.28	Two mRNAs
MLPH	Melanophilin	Exon skipping	2.02	-4.59	Multiple mRNAs and ESTs indicate skipping of this exon

Gene Symbol	Gene name	Event type	Fold change (Isoform A)	Fold change (Isoform B)	# and type of sequences supporting alternative expression event
MYT1	Myelin transcription factor 1	Complex	N/A	N/A	Multiple mRNAs which could contribute to observed expression
PLCB4	Phospholipase C, beta 4	Exon skipping	-5.03	-1.20	Two mRNAs and two ESTs
PPP2R1B	Protein phosphatase 2, regulatory subunit A, beta isoform	Exon skipping	-3.40	1.14	Single mRNA and ~10 ESTs
RC74	Integrator complex subunit 9	Exon skipping	-12.66	1.06	Single EST (cloned from hepatocellular carcinoma cell line).
RCC1	Melanophilin	Alternative TSS/polyA	-3.72	-1.10	None (some mRNA support for similar alternate TSS usage)
SSBP2	Single-stranded DNA binding protein 2	Alternative exon boundary	3.11	-3.21	None
TPST1	Tyrosylprotein sulfotransferase 1	Exon skipping	5.25	1.10	Multiple ESTs
UMPS	Uridine monophosphate synthetase	Exon skipping	-5.77	5.21	Multiple mRNAs and ESTs in human, mouse and rat
ZNF185	Zinc finger protein 185 (LIM domain)	Exon skipping	2.78	-3.97	None

References

1. Hubbard, T. et al. Ensembl 2005. Nucleic Acids Res 33, D447-53 (2005).
2. Griffith, M. & Marra, M. A. in Genes, Genomes & Genomics (eds. Thangadurai, D., Tang, W. & Pullaiah, T.) 201-242 (Regency Publications, New Delhi, 2007).
3. Castle, J. et al. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol 4, R66 (2003).
4. Srinivasan, K. et al. Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods 37, 345-59 (2005).
5. Clark, T. A., Sugnet, C. W. & Ares, M., Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science 296, 907-10 (2002).
6. Johnson, J. M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302, 2141-4 (2003).
7. Pan, Q. et al. Revealing global regulatory features of Mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16, 929-41 (2004).
8. www.affymetrix.com.
9. www.exonhit.com.
10. www.jivanbio.com.
11. Stolc, V. et al. A gene expression map for the euchromatic genome of Drosophila melanogaster. Science 306, 655-60 (2004).
12. Blanchette, M., Green, R. E., Brenner, S. E. & Rio, D. C. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. Genes Dev 19, 1306-14 (2005).
13. Li, C. et al. Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays. Cancer Res 66, 1990-9 (2006).
14. Relogio, A. et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. J Biol Chem (2004).
15. Sugnet, C. W. et al. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol 2, e4 (2006).
16. Ule, J. et al. Nova regulates brain-specific splicing to shape the synapse. Nat Genet 37, 844-52 (2005).
17. Zhang, C. et al. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. BMC Bioinformatics 7, 202 (2006).
18. Tai, I. T., Dai, M., Owen, D. A. & Chen, L. B. Genome-wide expression analysis of therapy-resistant tumors reveals SPARC as a novel target for cancer therapy. J Clin Invest 115, 1492-502 (2005).
19. Wagner, H. E. et al. Characterization of the tumorigenic and metastatic potential of a poorly differentiated human colon cancer cell line. Invasion Metastasis 10, 253-66 (1990).
20. www.vmware.com.
21. Gardina, P. J. et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 7, 325 (2006).
22. Yauk, C. L. & Berndt, M. L. Review of the literature examining the correlation among DNA microarray technologies. Environ Mol Mutagen (2007).
23. www.agilent.com.
24. www.nimblegen.com.
25. Kidd, E. A. et al. Variance in the expression of 5-Fluorouracil pathway genes in colorectal cancer. Clin Cancer Res 11, 2612-9 (2005).

26. Pinedo, H. M. & Peters, G. F. Fluorouracil: biochemistry and pharmacology. J Clin Oncol 6, 1653-64 (1988).
27. Bairoch, A. The ENZYME database in 2000. Nucleic Acids Res 28, 304-5 (2000).
28. Taomoto, J. et al. Overexpression of the orotate phosphoribosyl-transferase gene enhances the effect of 5-fluorouracil on gastric cancer cell lines. Oncology 70, 458-64 (2006).
29. Kodera, Y. et al. Gene expression of 5-fluorouracil metabolic enzymes in primary gastric cancer: Correlation with drug sensitivity against 5-fluorouracil. Cancer Lett 252, 307-13 (2007).
30. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. Nucleic Acids Res 31, 34-7 (2003).
31. Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res (2006).
32. Smit, A. F. A., Hubley, R. & Green, P. (1999-2004).
33. Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A 83, 3746-50 (1986).
34. Sugimoto, N., Nakano, S., Yoneyama, M. & Honda, K. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res 24, 4501-5 (1996).
35. Andronescu, M., Aguirre-Hernandez, R., Condon, A. & Hoos, H. H. RNAsoft: A suite of RNA secondary structure prediction and design software tools. Nucleic Acids Res 31, 3416-22 (2003).
36. Hancock, J. M. & Armstrong, J. S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput Appl Biosci 10, 67-70 (1994).
37. Futreal, P. A. et al. A census of human cancer genes. Nat Rev Cancer 4, 177-83 (2004).
38. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-9 (2000).
39. Cleveland, W. S., Grosse, E. & Shyu, W. M. in Statistical models in S (eds. Chambers, J. M. & Hastie, T. J.) (Wadsworth & Brooks/Cole, Pacific Grove, 1992).
40. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185-93 (2003).
41. Kuhn, R. M. et al. The UCSC genome browser database: update 2007. Nucleic Acids Res (2006).
42. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20, 1464-5 (2004).
43. R: A language and environment for statistical computing (ed. Team, R. D. C.) (R Foundation for Statistical Computing, Vienna, Austria, 2005).
44. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5, R80 (2004).
45. Thode Jr., H. C. Testing for Normality (Marcel Dekker, New York, 2002).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57, 289-300 (1995).

3. Alternative expression analysis by RNA sequencing[3]

3.1. Introduction

Expression of multiple distinct mRNA transcripts from a single gene locus by alternative transcript initiation, alternative splicing and alternative poly-adenylation (hereafter collectively referred to as 'alternative expression') is widely recognized as a source of proteome diversity in eukaryotic species[1]. PCR, microarray, and sequencing technologies have been applied to the study of transcript diversity generated by alternative expression with increasing success in recent years[1]. As sensitive genome-wide methods have become available[1], the prevalence and diversity of alternative expression has become increasingly apparent. Microarrays sensitive to alternative splicing, such as those developed in our group[2] and by others[3, 4], have made it possible to comprehensively detect and measure the abundance of known and predicted mRNA isoforms. More recently, massively parallel RNA sequencing, known as 'RNA-Seq'[5] or whole transcriptome shotgun sequencing ('WTSS')[6], has been proposed as a method with a number of potential advantages over microarray-based methods. Proof-of-principle experiments conducted in yeast, human and mouse have been reported by several groups. These reports described the use of RNA-Seq to perform *de novo* transcriptome annotation[7], estimate expression of specific isoforms[8], compare gene expression between a pair of contrasting conditions[9-11], comprehensively identify the expression of exons and exon-junctions within a single cell type[6, 12], and catalogue the diversity of known and novel transcripts across a range of tissues and individuals[5, 13, 14]. While these reports illustrated the utility of RNA sequencing for profiling the transcriptome, there remains a lack of methods to identify differences in mRNA isoforms in comparisons of samples.

In this work, we describe a novel method of analyzing WTSS data to allow assessment of the expression, differential expression and alternative expression of known and predicted mRNA isoforms including metrics for identifying reciprocal expression of alternative isoforms. Briefly, the approach consists of the following steps: (1) creation of a database of expression and alternative expression 'features' (**Figure 3.1**), (2) mapping of short paired-end sequence reads (generated by massively parallel

[3] A version of this chapter has been submitted. Griffith M, Griffith OL, Morin RD, Tang MJ, Pugh TJ, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney AD, Hirst M, Morin GB, Jones SJM, Tai IT, Marra MA. *Alternative expression analysis by RNA sequencing*

RNA sequencing) to these features (**Figure 3.2**), (3) identification of features that are expressed above the level of background noise while taking into account locus-by-locus noise levels (**Figure 3.3**), (4) identification of features that are differentially expressed between two disease states (**Figure 3.3**), (5) identification of the subset of differentially expressed features that correspond to alternative expression of mRNA isoforms (**Figure 3.3**), and (6) visualization of the outcome of these analyses. For the first time, we make available databases of sequence features and an associated visualization tool specifically tailored to the analysis of alternative isoforms by WTSS including supporting information from additional sequence resources (e.g. ESTs), cross-species sequence conservation and protein coding effect of each feature.

As proof-of-concept, the approach was applied to a comparison of fluorouracil (5-FU) responsive and non-responsive colorectal cancer cell lines. The drug 5-FU is an anti-metabolite chemotherapy agent commonly used in the treatment of several cancer types including head and neck, pancreatic, breast, stomach, and especially colorectal[15, 16]. We sequenced Illumina paired-end WTSS libraries constructed from polyA+ RNA isolated from 5-FU sensitive and resistant colorectal cancer cell lines (see **Methods**). To derive these lines, a cell line sensitive to 5-FU (MIP101) was passaged in the presence of increasing concentrations of 5-FU resulting in selection of a clone resistant to the drug (MIP/5FU)[17]. The resulting cells were profiled to test the utility of our method in detecting alternative expression events. The resistant cell line exhibited a 10-fold increase in the IC50 for 5-FU and as a clonal derivative of the sensitive cell line was expected to be highly related except for polymorphisms arising during exposure to 5-FU. Based on these characteristics we also sought to study the evolution of drug resistant disease and identify candidate 5-FU resistance genes or specific isoforms. The sequencing libraries consisted of 262 million paired-end reads and approximately 21.5 billion bases of sequence data. We present a detailed analysis of these data with particular emphasis on discriminating expressed features from background noise and identification of events that correspond not just to changes in gene expression but also to changes in the expression of specific mRNA isoforms. Comparison of the 5-FU sensitive state to the 5-FU resistant state revealed a global disruption in splicing characterized by increased expression of alternative mRNA isoforms, many of which were previously unreported in the UCSC and EnsEMBL transcript databases. Several promising candidate events emerged that included differential expression across entire

gene loci as well as loci that were not differentially expressed overall but exhibited a shift in isoform expression patterns. Among the list of candidate resistance loci exhibiting alternative expression, we observed an over-representation of genes with known or suspected roles in 5-FU metabolism and other genes whose disruption is consistent with other drug resistance mechanisms, particularly drug efflux.

The data we used to validate our method relates to a specific area of cancer biology research (5-FU resistance) but we believe that this work represents a general model for the study of diseases such as cancer where comparing the expression of mRNA isoforms between malignant and normal tissue, pre- and post-treatment biopsies, and other contrasting disease states can lead to important biological insights. Sequence data, alternative expression annotation databases for seven species (**Table 3.1**), candidate gene and isoform lists, source code, user manuals, a data viewer and other resources to facilitate alternative expression analysis by WTSS are available at our website: www.AlexaPlatform.org.

3.2. Results

3.2.1. Whole transcriptome shotgun sequencing (WTSS)

Analysis of paired 36- and 42-mer sequencing reads generated from sequencing libraries revealed that the observed median fragment size, inferred by mapping paired reads to known and predicted transcripts was 227 bp (**Figure 3.4**). The data described in this work correspond to a total of 262 million paired-end reads and 21.5 billion bases of sequence, generated on Illumina GAII sequencing devices (see **Methods**). The quality scores for all reads were examined. Quality values decreased substantially from the beginning to the end of the read for both read 1 and read 2 of a read pair (**Figure 3.5**). Despite this decrease in read quality from the beginning of the read to the end, the majority of reads mapping to known or predicted transcripts did so over their complete length (**Figure 3.6**). Lane-by-lane analysis did not identify any lanes with qualities or mapping rates low enough to warrant discarding the lane. Of the ~21 billion bases of sequence data generated, 1.2% were considered ambiguous during base calling (i.e. N's). Reads were removed at the outset of the analysis if they contained more than 1 ambiguous base (1.7% of all reads), or if more than 2/3 of their bases were identified as low complexity by mdust (0.5% of reads) of which the majority (61%) were polyA/T stretches, presumably corresponding to mRNA tails. An additional 0.02% of all reads

were removed because both read 1 and 2 of a pair were identical and might represent artifacts of library construction.

During library construction, polyA+ selection was performed to reduce the presence of highly expressed ribosomal RNAs and enrich for mRNA species. Although we found little evidence of RNA degradation (see **Methods**), even small amounts of degradation could introduce an over-representation of 3′ ends (i.e. 3′ end bias) in the sequence data because the polyA+ selection captures transcripts by their 3′ end. Significant end bias would diminish our ability to detect alternative expression. To investigate the possible presence of end bias in the data we plotted the location of the mapped reads on their cognate transcripts, and expressed this location as a percentage of transcript length with 0% denoting the 5′ end of a transcript and 100% denoting the 3′ end. These percentage values were then divided into bins according to the size of transcripts they were mapped to. The distribution of reads was generally reduced at both the 5′ and 3′ ends of each transcript but enrichment at the 3′ end was only apparent for transcripts larger than 10,000 bp (~1% of all transcripts) (**Figure 3.7**).

3.2.2. Annotation of features and read mapping

To assess differences in alternative expression between drug sensitive and resistant cells, we developed an alternative expression annotation database (see **Methods**). Briefly, this database defines expression 'features' that can be informative of alternative expression events such as exon skipping, alternative exon boundary usage, inclusion of cryptic exons, intron retention, etc. (see **Figure 3.1**). A total of 3.8 million such features were defined. Each feature was annotated with information describing its size, genomic repeat content, protein coding content, cross-species sequence conservation, mRNA/EST sequence support, etc., and assigned a descriptive feature name (see **Methods**). 16% of these features correspond to known EnsEMBL transcripts, 11% are not known to EnsEMBL but have some EST or mRNA support and 73% represent hypothetical events defined in our database but not currently supported by an EnsEMBL transcript, EST or mRNA (**Table 3.2**).

After quality filtering (**Methods**) of 262 million paired reads, the remaining 257 million reads (97.8%) were mapped to the database of sequences encompassing all of the features described above as well as all human and ancestral repeat elements. Briefly, 2.3% of these reads mapped unambiguously to known human or ancestral

repeat elements, 59% to known or predicted transcripts, 0.23% to novel junctions or novel boundaries of known exons, 5.2% to introns and 2.8% to intergenic sequence. 5.2% mapped to more than one sequence feature (i.e. ambiguous origin). 10% had similarity to human sequences but didn't meet our alignment thresholds (too many gaps or mismatches in the alignments) (see **Methods**) and the remaining 13% could not be assigned with high confidence to a human sequence (see **Figure 3.8**). Some of these unmapped reads may represent low-level contamination within our cell cultures or contamination from non-human libraries sequenced at our facility. Based on an analysis of mitochondrial sequences, we estimate that no more than 0.3% of all reads map unambiguously to non-human sequences (a rate within the normal range for libraries generated in our laboratory). The remaining unmapped reads may correspond to library sequencing or data extraction artifacts but they might also correspond to novel exon-exon junctions, not represented in our sequence database. These could theoretically be recovered by mapping directly to the genome with a splicing-aware aligner such as BLAT. Unfortunately, BLAT is not recommended unless both sides of a gapped alignment match perfect over at least 20 bp[18]. Since reads typically consist of two discontinuous 42-mers, either of which may poorly centered on an exon-exon junction, BLAT will often fail to align such reads. An alternative approach would be to use *de novo* assembly or 'peak finding' to identify novel exons and exon-exon junctions. Although *de novo* assemblers such as Velvet[19] and ABySS[20] and peak identification algorithms such as FindPeaks[21] have recently been developed, their use in assembling RNA sequence data remains preliminary[22, 23].

In summary, 203 million reads (77% of all reads) mapped with high confidence to human sequences and for 98.1% of read pairs with similarity to known or predicted transcripts, both reads were mapped as a pair to a transcript of a single gene locus (i.e. both reads of a pair map to the same gene). Additional read statistics summarized individually for each cell line are provided at our website: www.AlexaPlatform.org.

3.2.3. Comparison of Illumina WTSS expression data to Affymetrix and ALEXA microarray expression data

The method of alternative expression analysis we developed relies on accurate expression measurements for whole genes as well as individual features of those genes such as exons and exon junctions. To obtain a preliminary assessment of the utility of

WTSS data for alternative expression analysis we compared expression measurements derived from our Illumina data to measurements from the Affymetrix exon array and ALEXA[2] platforms generated with the same input RNAs (see **Methods**). Expression measurements as well as differential expression measurements were highly concordant between Illumina transcriptome sequencing data, Affymetrix exon arrays and ALEXA arrays fabricated by NimbleGen[2]. For log2 expression estimates the Spearman correlations between these data types were 0.70 to 0.88 (**Figure 3.9**). The sequencing-based method produced comparable correlations to arrays as did the array-based methods to each other. Similarly, differential expression measurements derived from each platform were compared and produced correlation values of 0.72 to 0.79 (**Figure 3.10**). The overlap between genes identified as differentially expressed to a level of 2-fold or greater (and p-value < 0.05) was significant and 56% of all differentially expressed genes were detected by two or three of the platforms. 19 of the top 20 differentially expressed genes identified by WTSS exhibited 2-fold or greater differential expression in the other two platforms (**Table 3.3**).

The signal-to-noise ratio of each platform was estimated by examining the expression estimates for exons (signal) and comparing these to expression estimates for introns, intergenic regions and random sequences (noise). This analysis was limited to a list of 100 housekeeping genes defined by Affymetrix and represented on both the Affymetrix Exon microarrays and custom NimbleGen ALEXA microarrays. Based on these features, the signal-to-noise ratios were 20.8 ± 0.42 (Affymetrix), 56.5 ± 2.5 (NimbleGen) and 381.1 ± 44.7 (Illumina). Thus, the signal-to-noise ratio for WTSS data was ~18 and ~7 times higher than the Affymetrix and NimbleGen arrays, respectively. While the ability to estimate the exon to intron expression ratio in microarray expression data is theoretically limited by physical parameters (such as hybridization stringency, fluorophore performance, detector dynamic range, etc.), the digital expression estimates generated by WTSS are primarily limited by sequencing depth. The greater the sequencing depth available, the greater the ability to detect transcripts with low expression levels and estimate the expression level difference between the least and most abundant transcripts. The observed distribution of exon, intron and random or intergenic sequence expression measurements were compared between Illumina, Affymetrix and NimbleGen data (**Figure 3.11**). The ability of each platform to correctly identify exons as expressed (a measure of sensitivity) and introns as not expressed (a

measure of specificity) was examined by creating receiver operating characteristic (ROC) curves and calculating signal-to-noise, specificity and sensitivity values (**Figure 3.12** & **Table 3.4**). The exon microarray data achieved a maximum specificity of 86.5% at 83.5% sensitivity compared to 95.8% specificity at 86.9% sensitivity for splicing sensitive microarrays and 99.0% specificity at 92.6% sensitivity for WTSS (these values correspond to the 'stationary point' of the ROC curve, equivalent to the global maximum of sensitivity plus specificity values).

We compared the ability of WTSS, splicing sensitive microarrays, semi-quantitative RT-PCR, and-real time quantitative RT-PCR to measure the ratio of two alternatively spliced isoforms of the gene *uridine monophosphate synthetase* (*UMPS*; the canonical isoform 'A' contains exon 2 and isoform 'B' skips exon 2) (**Table 3.5**). While each technology produced similar estimates, the WTSS results predicted the largest difference between the expression of *UMPS* isoforms A and B, suggesting that the dynamic range of this approach may exceed that of microarray and PCR based expression platforms. Indeed since dynamic range is limited primarily by sequencing depth in digital gene expression platforms such as WTSS, improved dynamic range will be obtained as long as sequencing depth is sufficient. In our libraries, more than 20,000 reads mapped to *UMPS* exons and junctions.

3.2.4. Expression of canonical and alternative sequence features

The number of reads corresponding to a single gene in our libraries ranged from 1 to 2.16 million (the gene, *H19*). In each library ~25,000 genes were detected by 1 or more reads and ~20,000 by 10 or more reads. In order to be informative of alternative expression, it was desirable that each gene be sequenced across the majority of exon bases. In each library, ~15,000 genes were sequenced to a depth of 1X or greater over 75% or more of their exonic bases. At a minimum depth of 10X this drops to ~8,000 genes and at a minimum of 100X it drops to ~2,000 genes (**Figure 3.13**). To further assess the degree to which the transcriptome had been comprehensively sequenced (and therefore the potential to measure alternative expression), the identification of transcriptome features within the data was plotted as a function of increasing read depth. The percentage of all possible genes and exon junctions as well as individual exon, intron and intergenic base positions of the EnsEMBL annotated genome detected at a level of 10 or more observations increased rapidly as library depth was increased

from 0% to 20% of all reads (**Figure 3.14**). The shapes of the curves depicted in **Figure 3.14** suggest that for genes, exon junctions and exon base positions, saturation was approached with increasing read depth and that after the first 20-40% of the data were analyzed (60-100 million reads), additional depth resulted in a rapidly diminishing rate of discovery of these features. In contrast, the curves representing intron and intergenic sequence do not show a pattern of saturation. This is not surprising, given that the library represents a polyA+ RNA input and the majority of sequence is therefore expected to correspond to exonic bases of the genome (~2% of the entire genome)[24]. When examined more closely, the intron and intergenic sequence appears to be randomly sampled from across the genome and shows no signs of saturation (i.e. increased depth continues to yield mostly unobserved sequence) in contrast to the reads mapping to exon content. We performed DNaseI treatment and polyA selection of our RNAs, but it is still possible that these sequences represent low-level contamination of the mature mRNA with un-processed RNA (heteronuclear RNA or 'hnRNA') or genomic DNA. They might also represent random transcriptional noise[25] from non-specific transcript initiation events. Expression of these sequences are hereafter referred to as intragenic and intergenic noise for sequences that map within or between the boundaries of known genes respectively. While the saturation curves for exon bases and exon-junctions suggested diminishing returns, a true plateau (slope = 0) was not achieved even at 100% of library depth. To our knowledge the MIP101 library, consisting of ~334 million reads (~167 million paired reads), is the deepest WTSS library produced to date for a single RNA sample (previous experiments reported 8 - 95 million reads[5-7, 9-14]). Despite this, based on extrapolation of these curves we predict that the percent discovery rate of exon bases would not converge with the level of discovery expected from intergenic noise alone until library depth reached 1 billion reads (**Figure 3.15**). We also showed that the degree of 'saturation' was highly dependent on the minimum coverage cutoff threshold required for a feature to be considered detected (**Figure 3.16**).

In order to identify alternatively expressed features, we first identified all those known and predicted features that were detected in our sequence data and expressed above the level of background noise (**Figure 3.3**). To differentiate between true expression and background noise we estimated the level of signal likely to be derived from intergenic and intragenic sources (see **Methods**) and used these estimates to

assess 'expression above background' for every feature. This resulted in a significant difference between the number of genes detected by 1 or more reads (~25,000) and those expressed above the level of intergenic and intragenic noise (~12,500). For the ~12,500 genes detected as transcribed we estimated that if the lowest expressed gene represents a minimum copy number of 1 per cell, then the dynamic range of expression is estimated to be 1 to ~6200. Based on this observed dynamic range we estimated the cumulative mRNA copy number of each MIP101 cell to be ~450,000 transcripts (see **Methods** for details). This estimate is consistent with previously published estimates of 100,000 to 700,000 mRNA transcripts per cell depending on cell type and method of estimation[26-28].

Because of the substantial depth of our libraries, in addition to the inability to completely remove all potential sources of noise such as genomic DNA and random transcription events[25], sequence from any gene may be detected even if it not expressed via a regulated transcription event. Supporting this assertion is the observation that the pool of genes classified as expressed is composed of only 5% pseudogenes, while the remaining genes which are not classified as expressed but have one or more reads are 33% pseudogenes (~7 times enriched for pseudogenes). Similarly, due to the presence of un-processed RNA contamination and random splicing events, the observation of sequences representing a skipped exon, alternative exon boundary or intron retention does not necessarily imply biologically significant alternative expression. Stochastic output from the splicing machinery has been shown to be prevalent, a function of the expression level and intron count of a gene, and result in primarily non-functional products[29, 30]. Consistent with this, the number of novel exon junctions identified as expressed above background did not appear to be correlated with either the expression level or intron count of their corresponding gene (Spearman correlations of 0.006 and 0.140 respectively). Since our evaluation of expression incorporates a gene-by-gene estimate of noise (**Methods**), the majority of spurious stochastic splicing events should be filtered in our analysis (**Figure 3.17**). Finally, since our primary focus is on the change in alternative expression between contrasting states, any isoforms which emerge as significantly different between two conditions are less likely to be a product of random splicing errors.

In the MIP101 library, a total of 12,398 genes were expressed above background. Of these, 11,388 (91.9%) were protein coding genes, 700 (5.6%) were pseudogenes,

167 (1.3%) were snoRNAs, 81 (0.7%) were miRNAs, and the remaining 60 (0.5%) belonged to various miscellaneous classes. Although only 33.5% of all EnsEMBL genes were detected as expressed, this included 54.0% of all protein coding genes.

Another measure of the comprehensiveness of the expression data is the degree to which we detect the expression of the annotated transcript features of EnsEMBL (known exons and exon junctions). For example, we observed expression for 88.3% of the individual exons and 83.0% of the canonical exon-exon junctions of the 12,398 genes detected as expressed in the MIP101 library. While these observations indicate the sensitivity in detecting the exons and junctions of expressed genes, the fact that only 3.2% of the introns and 0.2% of the non-canonical exon-exon junctions of these same genes were detected as expressed suggests that the specificity is also high.

We summarized the total number of genes, transcripts, exon regions, junctions, alternative boundaries, introns and intergenic regions found to be expressed in each library (**Table 3.6**). Junctions and alternative exon boundaries were further classified as either known or novel according to EnsEMBL transcripts (see **Methods**). Introns and intergenic regions were also divided into putative 'active' and 'silent' regions using the coordinates from alignments of all human mRNAs and ESTs. The distribution of expression levels for exon regions and exon-exon junctions were very similar and the expression of alternative exon boundaries, introns and intergenic regions followed the expected pattern of decreasing read support (**Figure 3.18**). The majority of exons were expressed above the 95th percentile of both 'silent' intron and intergenic region estimates (**Figure 3.19**). In the MIP101 library the overall median coverage of all exon regions was 114.5 compared to 0.64 and 0.14 for silent intron and intergenic regions respectively.

3.2.5. Differential expression analysis

Having identified all features expressed above background, we next calculated the differential expression value for each feature (**Figure 3.3**). We identified 7,198 features that were significantly differentially expressed between 5-FU sensitive and resistant cells (**Table 3.6**). These features corresponded to 1,479 distinct genes. Many of these features were differentially expressed exons and canonical exon junctions belonging to loci that were differentially expressed across their entire length (i.e. all canonical exons and junctions) and were not considered to be alternatively expressed. These changes

in gene expression level are presumably caused by regulatory changes (genetic or epigenetic) accrued during the transition from 5-FU sensitivity to resistance. Possible modifications that might affect the expression level of an entire gene include deletions, amplifications, mutation of enhancers, and hyper- or hypo-methylation of promoters sites. The most highly differentially expressed gene, *H19*, exhibited 1/100th the abundance in resistant cells relative to sensitive cells (*H19* was ranked #5 in abundance level in MIP101 and #1 for differential expression between MIP101 and MIP/5FU*).* This gene functions in tandem with its neighbor, *IGF2*, as a regulator of cellular growth. *H19* and *IGF2* are known for being maternally and paternally imprinted[31], respectively, and for their association with Beckwith-Wiedemann syndrome[31]. Interestingly, *IGF2* was not detected above background in either cell line. The expression pattern of all individual features of *H19*, *IGF2* and all other genes can be viewed using the 'ALEXA-Seq' data viewer at our website (www.AlexaPlatform.org). **Figure 3.20** shows a screenshot for *H19* illustrating the loss of expression at all exons and canonical junctions of this locus.

The gene with the greatest increase in abundance in resistant cells relative to sensitive cells was *KRT20* (aka *CD20*), which has been proposed as a prognostic marker for colorectal cancer[32]. The pattern of differential expression was consistent across the 8 exons and 7 canonical junctions of *KRT20* (**Figure 3.21**). No intronic, alternative exon boundary or exon-skipping features for this gene were expressed in either condition. While this gene is not the only example with a simple gene-wide change in expression, there were many genes that exhibited a pattern of alternative expression. These events would be missed by expression analysis that treats each gene model as a single expression unit. Illustrating this point is the observation that only 27 of the top 50 differentially abundant transcripts identified corresponded to differential gene expression events such as those observed for *H19* and *KRT20* (**Table 3.8**).

3.2.6. Alternative expression analysis

Having identified those features expressed above background and differentially expressed between 5-FU sensitive and resistant cells we next sought to identify the subset of expression events representing alternative expression of mRNA isoforms (**Figure 3.3**). Calculating expression of specific known alternative transcripts was possible for 10,151 multi-transcript genes (genes with multiple transcripts where the

sequence features unique to each transcript were identified). Of these multi-transcript genes, 5,389 genes had evidence for at least one isoform being expressed and 1,975 (37%) had evidence for expression of multiple known isoforms within the MIP101 cell line. An additional 5,862 loci had evidence for expression of novel isoforms (1,372 with novel junctions and 4,490 with novel alternative boundaries, novel exons or intron retentions). The known and novel transcripts and other features that were expressed in each library and differentially expressed between libraries are enumerated in **Table 3.6**. The number of expressed novel features enumerated in this table suggests that the annotation of alternative transcripts within EnsEMBL incomplete. Alternative expression is predicted to influence protein coding potential in 90% of alternative expression events involving exons and 98% of events affecting exon junctions (**Table 3.2**). A total of 68% of all multi-exon expressed genes showed evidence for expression of multiple isoforms in our data. While recent studies have predicted that 90-95% of human genes undergo alternative splicing[13, 14] these studies assessed these statistics across a variety of diverse cell types, whereas our estimate corresponds to clonally related cell lines of colorectal cancer origin[17, 33].

The pattern of observed number of exons skipped in expressed exon-exon junctions is revealing. Our exon junction database consisted of ~2.2 million known and hypothetical junctions. 215,743 of these junctions were supported by at least one EnsEMBL transcript, and 203,600 and 194,711 junctions were supported by one or more mRNA and EST sequences respectively. Of 215,743 known junctions, 16,163 correspond to a known exon-skipping event, the remainder representing connections of adjacent exons. Known exon-skipping events involve anywhere from 1 to 151 consecutive exons being skipped but 84% involve only a single skipped exon and 95% involve 5 or fewer skipped exons. In contrast, the number of exons skipped by all possible exon junctions ranges from 1 to 354 and only 8% involve a single exon being skipped. The numbers of exons skipped among the expressed exon junctions matched the percentage of known junctions with these numbers of exons skipped (**Figure 3.22**). If this analysis is limited to only the novel expressed exon junctions, we observe heavy bias towards small numbers of exons skipped (72% have 1 exon skipped and 98% have 5 or less) again following the pattern of known exon skipping events. No such bias is observed if we randomly sample junctions from the database of all possible exon junctions (**Figure 3.22**). These observations support the possibility that these are real

exon-skipping events that simply have not been captured by the EnsEMBL annotation process. Further evidence is supplied by the observation that the pool of putative novel exon-skipping events, although not represented by known EnsEMBL transcripts, are statistically enriched for those with EST support. While 9.7% of all ~2.2 million possible exon junctions have EST support, 57.6% of the 3,802 novel junctions observed in either library have EST support (p-value < 1.0×10^{-300} by two-tailed Fisher's exact test) (**Table 3.7**). The pool of expressed novel alternative boundaries is similarly enriched for those events with EST support. We also found that these novel events tended to exhibit cross-species conservation with at least one mRNA or EST expressed in at least one additional species exhibiting the same exon-exon junction or alternative exon boundary (p-value < 1.0×10^{-159}) (**Table 3.7**). Interestingly, we found that although the number of novel exon-exon junctions predicted to alter the open reading frame of a transcript was high (93.7%), this was less than would be expected by chance (97.0%) (**Table 3.7**). This suggests that alternative splicing may play a significant role in modifying regulatory sequences present in the un-translated regions (UTRs) of transcripts (such as miRNA target sites)

In addition to expressed and differentially expressed features described in **Table 3.6**, the subset of differential expression events that correspond to alternative expression are also summarized in the last column. These correspond to cases where specific features such as exons or exon-skipping junctions are differentially expressed but the level of expression of the gene they belong to is not significantly changed between the two libraries (or the change is in the opposite direction). Such events may indicate differential expression of a single isoform at a locus that expresses multiple isoforms. The top 50 differentially or alternatively expressed genes identified in our pair-wise comparison of the MIP101 and MIP/5FU WTSS libraries were examined in our 'ALEXA-Seq' viewer and summarized (**Table 3.8**). 23 of these events consist of exon-skipping, alternative transcript initiation or poly-adenylation, alternative 5′ or 3′ splice site usage, or intron retentions. While only 23 are shown in this table, a total of 306 genes with these kinds of events were identified by combining differential expression (DE) analysis with use of a 'splicing index' $(SI)^2$ (see **Methods**). We further refined this approach by developing 'reciprocity index' (RI) and 'percent feature contribution' (PFC) calculations to help identify cases where multiple isoforms were reciprocally differentially expressed with respect to overall gene expression or where a single isoform appeared to be

differentially expressed while the overall expression of the gene was constant. The results of scoring and statistical analysis based on DE, SI, RI, and PFC scores were summarized (**Table 3.6**) and can be visualized in the 'ALEXA-Seq' data viewer available at our website (www.AlexaPlatform.org). Using this viewer, we manually examined candidate alternative expression events of many types including differential expression of known alternative isoforms (e.g. *SCL24A1*), novel exon skipping events (e.g. *UMPS*; **Figure 3.23**; and *OCIAD1*, **Figure 3.24**), use of an alternative first exon (e.g. *MAD1L1*, *IGFL2*, *NKIRAS2*), inclusion of a cryptic exon (e.g. *ATP8A1*), alternative exon boundary usage (e.g. *NFIB*), intron retentions (e.g. *IQGAP3, SLC12A7*), and use of alternative transcript initiation sites (e.g. *LAMA3*, *c12orf63*).

3.2.7. Global disruption of splicing

The total number of features expressed above background was determined for the MIP101 (5-FU sensitive) and MIP/5FU (resistant) libraries. Although the MIP101 library was approximately twice the size of the MIP/5FU library (167 million paired reads compared to 95 million), this translated into only a small difference in the number of genes, transcripts, exons, and known exon junctions being identified as expressed (see **Table 3.6**). Specifically, while the MIP/5FU library had ~57% of the depth of the MIP101 library, on average 98.5% of the features were detected (i.e. only 1.5% fewer genes, transcripts, exons and junctions). The same slight drop was also observed for 'silent' intergenic regions (no mRNA/EST evidence for expression). These findings are consistent with the observation that these libraries had effectively reached a point of saturation where additional depth resulted in diminishing discovery of expressed features (**Figure 3.14**). For genes that were differentially expressed between sensitive and resistant cells, there was a trend towards loss of expression in resistant cells with twice as many under-expressed genes in resistance compared to over-expressed (171 under-expressed, 81 over-expressed). A similar bias towards loss of expression in resistant cells was observed for known exons and known exon junctions. Interestingly, for all feature types indicative of possible expression of novel transcripts (novel junctions, alternative exon boundaries, cryptic exons and intron retentions) the number of such events detected was actually higher in the smaller library (**Figure 3.25**). For example, the MIP/5FU library had 31.3% more novel exon junctions, 31.9% more novel exon boundaries, and 37.9% more intron retentions than MIP101. One possible

explanation for this observation is that the level of genomic DNA and/or un-processed RNA contamination was elevated in the MIP/5FU library. However, the comparable detection of intergenic elements in both libraries (**Table 3.6**) suggests that the level of intergenic background noise was not substantially different between the two libraries. Furthermore, increased contamination would not account for the increased expression of novel exon-exon junctions as these sequences occur in mature (spliced) mRNAs but do not generally occur in either the genome or un-processed RNA. If we consider the 3,802 expressed novel exon-skipping junctions observed in either library and eliminate those that were present in both libraries, 1,713 remain and 75% of these were observed only in the MIP/5FU library. In other words, despite fewer sequences we observed expression of approximately three times as many novel exon skipping isoforms in the resistant cells (533 in MIP101 only versus 1,556 in MIP/5FU only). The same trend was observed for alternative exon boundaries and intron retentions (**Table 3.6**).

We hypothesized that the increase in expression of sequence features representing putative novel transcripts might be a consequence of (1) locus specific events acquired in the resistant cell line and (2) changes in the regulation of splicing. In the first scenario, some of the novel isoforms might correspond to mutations acquired at splice sites or within other splicing regulatory sequences. For example, we confirmed by Sanger sequencing the presence of a heterozygous splice site mutation at the acceptor site of exon 2 of *UMPS* which corresponded to increased skipping of this exon in MIP/5FU cells. In the second scenario, the increase in novel isoforms might be caused by differential expression of components of the splicing machinery itself. For example, we observed a loss of expression of U2 snRNA (2-fold down-regulated) in resistant cells as well as a gain of E3 snRNA (2-fold up-regulated). Both of these genes are known to participate in RNA splicing. Pathway analysis (see **Methods**) did not reveal any additional splicing related genes that were differentially or alternatively expressed.

3.2.8. Aberrant expression of candidate 5-FU resistance genes

To identify genes whose disruption might be expected to confer resistance to the drug 5-FU we performed pathway analysis with the Ingenuity Pathway Analysis software package (see **Methods**). The top functional categories identified as significantly enriched in our candidate differentially and alternatively expressed genes are listed in **Table 3.9**. Because this software has limited information on pathways specifically

related to 5-FU resistance we also examined the manually annotated 5-FU pathway from PharmGKB (see **Methods**). Of particular relevance was an observed down-regulation of the canonical isoform of *uridine monophosphate synthetase* (*UMPS)* and up-regulation of a novel exon-skipping isoform (skipping of exon 2) (see **Figure 4.2** - **4.3**). *UMPS* is a member of the pyrimidine metabolism pathway and is thought to be involved in the activation and anti-tumour action of 5-FU[34]. Skipping of exon 2 is predicted to result in use of a second start codon, that would produce a truncated protein missing the orotate phosphoribosyltransferase domain of *UMPS* (considerable details describing this event are presented in **Chapter 4**). Interestingly, *uridine monophosphokinase 2* (*UCK2*), another gene involved in pyrimidine metabolism also shows an up-regulation of a novel exon-skipping event (skipping exon 6). In addition to *UMPS* and *UCK2*, several other candidate genes with known or suspected roles in 5-FU action or multi-drug resistance appeared to be differentially expressed or spliced between 5-FU sensitive and resistant cells. Specifically, 9 of 52 known 5-FU pathway genes (*CDKN1A*, *ERCC2*, *FDXR*, *NFKB1*, *UCK2*, *UGT1A8*, *UMPS*, *TPMT*, and *TYMS*) were in the pool of candidate genes. This represents a statistically significant over-representation compared to expectations by chance (p-value = 2.00×10^{-4} by two-tailed Fisher's exact test). Additional genes with suspected functions that are mechanistically consistent with 5-FU or multi-drug resistance were identified. For example, *ABCA3*, a drug transporter previously reported as having a role in multi-drug resistance to cytostatic chemotherapies[35, 36] was up-regulated by ~4-fold. Similarly, four of six drug transporters thought to be involved in 5-FU efflux or reuptake were over-expressed (*ABCC5*), under-expressed (*ABCG2*) or exhibited over-expression of novel exon-exon skipping isoforms (*ABCC3* and *ABCC4*) in resistant cells[37]. The effect of disruptions to genes involved in 5-FU metabolism and drug efflux, may collectively contribute to reduced exposure to and activity of 5-FU within MIP/5FU cells and thereby confer resistance in these cells.

3.3. Discussion

We have described the creation and use of novel methods and resources for alternative expression analysis that leverages the depth and base-level resolution afforded by massively parallel RNA sequencing. By analyzing these data in the context of known annotations and correlation with existing expression platforms we have shown that they

provide a comprehensive, sensitive and specific digital representation of expressed transcriptomes. We have shown that the quality of these data are primarily limited by sequencing depth and that the expressed features inferred from these data contain a wealth of details on the structure and complexity of mRNA transcripts generated from each gene locus.

A recent publication suggested that "… RT-PCR remains the only sensitive high-throughput method for validation of alternative splicing candidates from genome-wide studies, particularly for low-abundance transcripts" [38]. The authors then described a comprehensive survey of cancer associated AS events by genome-wide application of RT-PCR. For the ~1,600 gene loci examined in their study we examined our own data to assess the degree to which we could profile expression from these loci. In the MIP101 library, we obtained ~881 million bases of sequence for these genes, an average of 13,985 reads per locus, an average coverage of 220X, and on average 64% of the exon bases of each locus were sequenced to a depth of 10X or greater. Thus, we believe that with sufficient depth, the massively parallel RNA sequencing approach described in this work now represents a feasible alternative to this RT-PCR based approach. Furthermore, it has the advantage of not requiring any pre-selection of target genes or transcripts whether they are known or predicted events, thus allowing a relatively un-biased assessment of the transcript isoforms expressed from each locus. In contrast to both splicing microarrays and RT-PCR the data generated are not limited by our knowledge of the transcriptome at the time of data generation. Also in contrast to microarrays, the WTSS data allows base level resolution and mutation detection. Both microarrays and RT-PCR experiments require design of oligonucleotides based on specific known or predicted gene models. As annotations improve or new predictions are made, WTSS data can simply be re-analyzed to accommodate the changing knowledge landscape. Profiling millions of known and predicted exon-exon connections as we describe here is simply not practical by RT-PCR. The sensitivity of WTSS to genes with low expression levels is dependent upon sufficient depth, but with the depth reported in our experiments we achieved detection of ~25,000 genes, half of which were detected above the level of background noise. Included among our expressed genes were well known low copy genes such as telomerase reverse transcriptase (sequenced to an average coverage of 5.5X) for which we detected a novel retention of intron 11. We also detected 72% of a test set of genes recently used in the optimization of long

oligonucleotide arrays that were considered 'low' expression or 'undetectable'[39]. In the MIP101 library, we observed expression of these genes at an average coverage of 95X, well above our estimate of background noise (~3.4X; **Figure 3.19**).

While WTSS data is not fundamentally limited by the quality or completeness of the genome annotation, interpretation of the WTSS data in the context of alternative expression is facilitated by the availability of accurate gene models. For this reason, our analysis placed emphasis on supplementing and tailoring existing gene annotations to allow a more comprehensive characterization of alternative expression. In addition, we believe that an annotation, alignment and expression strategy that combines both genome and transcriptome resources is desirable for RNA sequence data. We propose that one area of future work is to couple annotation from existing databases with entirely *de novo* annotation driven by the sequence data itself. Preliminary work in this area using *de novo* transcript assembly [22] seems promising and it should be possible to incorporate these methods in the near future to further enhance the comprehensiveness of our approach.

By generating WTSS libraries representing contrasting conditions we showed that it is also possible to elucidate changes in the expression of entire transcripts, subtle shifts in the ratio of expressed isoforms, and entirely novel transcripts. We found that the alternative expression analysis described here allowed the identification of potentially important events that would have been missed if we relied on differential gene expression analysis. For example, if we had considered only differential gene expression we would have identified a total of 259 genes (such as *H19* and *KRT20* discussed above) whose expression was altered between chemotherapy sensitive and resistant cells. By defining distinct known and hypothetical expression features such as exon regions, exon junctions, alternative exon boundaries, etc. and using these features to identify genes that are differentially spliced, an additional 306 genes were identified, including relevant 5-FU resistance candidates *UMPS* and *UCK2*. Both the number of differentially expressed genes as well as differentially spliced genes might be increased by increasing library sequence depth (especially for genes that are expressed at low levels).

Alternative expression analysis by WTSS will be useful not just in paired comparisons of disease states, tissue types, etc. but can be easily applied to more global analyses such as disease classification and should be more robust than simple

gene expression estimates typically used for these analyses to date. It is reassuring to note that the analytical approach we describe readily detected a global pattern of aberrant splicing (**Figure 3.25**) as well as specific candidate isoform markers of resistance (**Table 3.8**). Profiling the global regulation of splicing and identifying specific isoforms to be used in a diagnostic, prognostic, or therapeutic context is increasingly cited as an important area of disease research, particularly in cancer [38, 40, 41].

3.4. Methods

3.4.1. Tissue culture and RNA preparation

The colorectal cancer cell line, MIP101[33] and a previously generated 5-FU resistant derivative, MIP/5FU[17] were maintained in Dulbecco's Modified Eagle Medium (Invitrogen Inc., Burlington, ON) supplemented with 10% new born calf serum (Invitrogen), 1% kanamycin (Invitrogen) and 1% penicillin/streptomycin (Invitrogen) at 37ºC in a humidified environment of 5% CO_2. The media for MIP/5FU cultures were supplemented with the chemotherapy drug 5-fluorouracil (5-FU) to a final concentration of 5 µM to maintain resistance. Cultures used for RNA isolation were seeded at a density of 10-30% and grown in media without drug or antibiotics for ~48 hours. Total RNA was isolated using an 'RNeasy' kit (Qiagen, Mississauga, ON.). Total RNA was DNAseI (Qiagen) treated during RNA isolation according to Qiagen's instructions. RNA was quantified and assessed for degradation using an Agilent RNA 6000 Nano assay and was not used for library construction if it had an RNA integrity score less than 9 out of 10. For each sample, polyadenylated RNA was purified from 16.8 µg of DNAseI treated total RNA using the MACS™ mRNA Isolation Kit (Miltenyi Biotec, Germany).

3.4.2. Illumina library construction and sequencing

PolyA+ RNA was purified from total RNA isolated from cell lines (see above) and used for cDNA synthesis followed by fragmentation into 190-210 bp fragments and generation of whole transcriptome shotgun sequencing (WTSS) libraries as follows. Double-stranded cDNA was synthesized from purified polyA+ RNA using a Superscript™ Double-Stranded cDNA Synthesis kit (Invitrogen, Carlsbad, CA) and random hexamer primers (Invitrogen) at a concentration of 5µM. The resulting cDNA was sheared using a Sonic Dismembrator 550 for 5 minutes at amplitude setting '7' (Fisher Scientific, Canada) and size separated by PAGE (8%). The 190-210bp DNA

fraction was excised, eluted overnight at 4°C in 300 µl of elution buffer (5:1, LoTE buffer (3 mM Tris- Cl, pH 7.5, 0.2 mM EDTA)-7.5 M ammonium acetate) and purified using a QIAquick purification kit (Qiagen). The sequencing library was prepared following the Illumina Genome Analyzer paired end library protocol (Illumina Inc., Hayward, CA) with 10 cycles of PCR amplification. PCR products were purified on QIAquick MinElute columns (Qiagen) and assessed and quantified using an Agilent DNA 1000 series II assay and Qubit fluorometer (Invitrogen, Carlsbad, CA) respectively. The resulting libraries were sequenced on an Illumina Genome Analyzer II following the manufacturer's instructions.

3.4.3. Data pre-processing

Image analysis and basecalling was performed by the GA pipeline v1.0 (Illumina Inc., Hayward, CA) using phasing and matrix values calculated from a control phiX174 library run on each flowcell. Raw Quality scores were calibrated by alignment to the reference human genome (NCBI build 36.1, hg18) using ELAND (Illumina Inc., Hayward, CA).

Illumina paired-end sequencing produces paired reads from opposite strands of the ends of double-stranded cDNA fragments (i.e. +/- or -/+ orientations). At the outset of our analysis, the second read of each pair was reverse-complemented (resulting in +/+ or -/- orientations). Since double-stranded cDNA was sequenced, the source strand information of RNAs converted to cDNAs was lost in this approach and we expected an equal representation of both strands in our sequence alignments. In our alignments to known transcripts, 98% of all read paired exhibited the expected strand orientation (+/+ or -/-). There was no apparent bias towards one strand or the other (49.3% of read pairs were +/+ and 49.1% were -/-).

Before further analysis, reads were first filtered on the following basic criteria: (1) poor quality (arbitrarily defined as more than 1 ambiguous base call); (2) low complexity (2/3 or more bases of a read identified as low complexity by mdust[42] using default parameters) and (3) duplicate reads of a pair (read 1 and read 2 of a pair are identical or a reverse complement of each other).

3.4.4. Source of gene models

Annotated transcripts and exons of EnsEMBL version 53 were used for the analysis described in this work (**Table 3.2**). Gene, transcript and exon coordinates as well as

associated descriptive information were retrieved from EnsEMBL by use of the EnsEMBL Application Program Interface [24] (http://www.ensembl.org/info/data/api.html). This dataset consisted of 36,953 genes with 62,371 transcripts and 273,464 non-redundant exons. Genes ranged in size from 8 to ~114,339 bases with a median size of 1,151 bases. 70% of these genes were 'known' and 30% were classified as 'predicted' in the EnsEMBL database. Exons ranged in size from 1 to ~17,546 bases with a median size of 125 bases. The number of exons per gene ranged from 1 to 360 with a median of 3 and an average of 7.5. Gene models from EnsEMBL were supplemented by incorporation of mRNA and EST sequence alignments from the UCSC genome annotation database[43] (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/).

3.4.5. Creation and annotation of an alternative expression database

To facilitate interpretation of WTSS data in the context of alternative expression, we created a tailored annotation database to define and characterize sequences representing expression events for each locus. This database seeks to describe the characteristics of all known and predicted transcripts, exon regions, exon junctions, alternative exon boundaries, introns, and intergenic regions associated with every gene in a genome. These sequences are collectively referred to as sequence 'features' throughout this work (see **Figure 3.1** for an illustration). To create these databases we first retrieved gene, transcript and exon sequences and coordinates from EnsEMBL. We then defined regions of each type and used existing mRNA and EST sequence data to supplement the information in EnsEMBL. Each feature was defined as 'known' or 'novel' relative to EnsEMBL. For each feature, the number of supporting mRNAs and ESTs was also determined. This sequence support was assessed for the target species and then for all other species to assess conservation of each feature. In addition to storing the number of EST and mRNA sequences supporting each feature, the number of species with a supporting sequence was also noted. For all feature types, the size of the feature was noted as well as the number of these bases that correspond to repeat masked positions according to EnsEMBL (all bases identified by mdust, RepeatMasker and tandem repeat finder)[24]. Whenever the number of bases was needed to perform a calculation for a feature, such as determining average read coverage, the number of unmasked bases was used.

Exon regions were defined by identifying clusters of overlapping exons to create a 'block' of exon content. This block was then divided into 'exon regions' using the boundaries of all exons within the cluster. This approach divides overlapping exons into pieces that are often specific to particular isoforms. Exon regions were then labeled from 5′ to 3′ starting at 1 and ending at N (where N is the number of exon clusters for the gene). In the case of overlapping exons resulting in multiple exon regions, these were labeled 'a … z'. For example, a gene with three exons and two known transcripts with an alternate exon 2 boundary might have exon regions defined as: 'ER1', 'ER2a', 'ER2b', 'ER3' (**Figure 3.1**).

Exon-exon junctions were defined by identifying all exon boundaries within a gene and generating all possible pair-wise connections of these exons that could theoretically result from alternative splicing of an exon donor site to an exon acceptor site. A gene with n exons has (n!/(2![n-2]!)) possible junctions. In the case of overlapping exons, multiple acceptor or donor sites within an exon cluster were labeled 'a … z' and the resulting junction name consisted of the exon numbers of two connected exons as well as the donor and acceptor site involved (for example, 'E3a-E4b' represents the connection of exon 3 and 4 using the 3′ most donor site of exon 3 and the second acceptor site of exon 4). The size of exon-exon features was chosen to be 62 bases to accommodate the length of our sequence reads (42-mers). This length was chosen so that a read aligning across its full length would overlap the centre of the junction by at least 10 bases.

Alternative exon boundary features were defined and named in a similar way to that employed for exon-exon junctions. An alternative exon boundary feature was defined for every exon splice acceptor or donor site annotated in EnsEMBL. The size of these features was the same as for exon junctions (62-mers). An alternative exon boundary corresponding to a cluster of overlapping exons was named according to the exon number from 5′ to 3′ and multiple splice donor or acceptor sites were labeled 'a … z'. For example, 'E3_Da' refers to a feature spanning from the 5′ end of exon 3 into the downstream intron at the 3′ most donor site of exon 3.

All introns and intergenic regions were simply defined as those portions of the genome not corresponding to EnsEMBL exons. Introns are those regions between known EnsEMBL exons and were numbered from 5′ to 3′ within each gene. Intergenic regions were defined across each chromosome and labeled as such, and the identity of

genes flanking each intergenic region were stored. For both, introns and intergenic regions, alignments of mRNAs and ESTs were used to supplement the EnsEMBL gene annotations. Specifically, alignments of these sequences to the genome were used to define 'active' portions of each intron or intergenic region ('active intron regions' and 'active intergenic regions'). Messenger RNA or EST alignments were only considered if they had more than 300 matching bases less than 5 mismatches or gaps and less than 2 ambiguous bases. Regions that did not exhibit any such evidence of expression from mRNAs or ESTs were defined as 'silent' portions of each intron or intergenic region ('silent intron regions' and 'silent intergenic regions').

Finally, although transcript features were already defined in EnsEMBL, for each of these transcripts we further identified those exon regions and exon-exon junctions that were specific to each transcript (where possible). The feature collection defining each transcript was used to calculate expression for that transcript.

All features were assigned unique feature IDs and stored in a standardized format including the following characteristics for each feature: source gene id, gene name, feature name, chromosome, strand, coordinates relative to the source gene (in 5′ to 3′ orientation), chromosome coordinates, feature size, number of unmasked bases, number of protein coding bases, number of supporting EnsEMBL transcripts, mRNAs and ESTs (for both the target species and all other species), and if applicable the ID of the transcript this feature was unique to. Exon junction features also indicate the number of exons skipped. Currently, ALEXA annotation databases as described here have been generated for the human genome (used in this analysis) as wells as chimp, mouse, rat, fruit fly and yeast (**Table 3.1**). Additional annotation databases can be generated to accommodate any species annotated by EnsEMBL upon request. Annotation databases, including fasta files representing all sequences can be downloaded from www.AlexaPlatform.org.

3.4.6. Alignment strategy and assignment of reads to features

Our transcriptome analysis strategy involved alignment to a combination of genomic as well as known and predicted transcript sequences. Using the BLAST alignment algorithm[44] we obtained alignments containing both mismatches and gaps. With a word size of 11, it was possible to obtain alignments of 42-mer reads with up to 6 mismatches and 2 gaps, however we pre-filtered alignments to remove those with a bit score less

than 48.1. This resulted in alignments with no more than 4 mismatches and 2 gaps. Using BLAST also allowed us to generate sub-sequence (partial) alignments. For example if the first 25 bases of a read were of high quality and the remainder of the sequence failed to match, an alignment was still returned (some short read aligners such as Eland do not return such alignments). In our data, read alignments for 42-mer reads ranged in size from 20 to 49 bases (**Figure 3.6**). We were also able to learn the identity of an arbitrary number of multiple matches in the genome, allowing us to gauge the degree of ambiguity in the mapping of each read. At the time of analysis, aligners such as Eland and Maq noted mismatches but not the number of ambiguous mapping positions. Similar to the strategy employed by Maq we used read pairing information to resolve mapping ambiguities where possible. Specifically, the combined alignment score was used to assign reads to targets where possible and in cases where one read of a read pair aligned ambiguously, it was often possible to resolve this ambiguity by considering the paired read. In cases where one read of a pair could not be aligned, the other read was still assigned to a feature as a single-end read. The advantages of BLAST over other aligners were accepted at the cost of computation time and disk storage space required for the resulting alignment records. Furthermore, some of the advantages of BLAST as described above have been incorporated into short read aligners such as Maq[45], NovoAlign (www.novocraft.com), Shrimp[46], SOAP[47] and BWA[48]. The approach described in this work should not be fundamentally affected by the use of these or other short read aligners.

Reads that remained after quality filtering (see data pre-processing) were mapped as described above to a database of sequence features of the following types: (1) repeat elements obtained from RepBase[49], including all human and ancestral repeats, (2) known transcripts as defined by EnsEMBL (version 53), (3) all possible 62-mer exon-exon junctions, (4) alternative exon boundaries, (5) introns and (6) intergenic regions (all sequence not covered by the previous categories). Reads were assigned to one of these classes if they had a perfect or near perfect match and the match was non-ambiguous. Ambiguously mapping reads (4.9% of all reads) were noted and their alignments were stored but they were not used for further analysis. Reads that matched repeat elements were flagged at the outset, were not used to calculate expression of any feature, and the effective size of each feature was adjusted to account for the number of repeat-masked (and therefore un-'mappable') bases.

3.4.7. Cross-platform comparison of expression and differential expression estimates

Expression estimates for genes, exons and introns derived from Illumina (WTSS) libraries were compared to previously generated microarray data for the same RNAs. Specifically, we generated Affymetrix GeneChip microarray data (Human Exon 1.0 ST arrays) and custom NimbleGen microarray data (alternative expression analysis arrays)[2]. For direct comparison of expression estimates, gene and exon expression values were calculated for Affymetrix data according to Affymetrix's recommendations using Expression Console and the PLIER algorithm. Custom NimbleGen array data were processed in a similar way. Detailed descriptions of array data processing for both array platforms were previously reported[2]. The WTSS expression estimates were derived as described above. For all three platforms, a value of 1 was added before converting to log2 scale. Scatter plots comparing expression and differential expression data (**Figure 3.9** and **Figure 3.10**) correspond to only those ~2,500 genes profiled by all three platforms (primarily limited by the custom NimbleGen array data). The correlation coefficient for each scatter plot was determined by both the Pearson and Spearman methods. To compare the dynamic range, signal-to-noise ratio, sensitivity and specificity of the three platforms, the analysis was further limited to the introns and exons of ~100 housekeeping genes routinely used by Affymetrix in their array designs. This list of housekeeping genes is available at our website (www.AlexaPlatform.org). For both array platforms, random hybridization signal was estimated by examining a pool of random sequence probe spots included in each array design. Since there is not an analogous data type in WTSS we used the expression estimates for putatively silent intergenic regions as an estimate of background signal. All plots and correlations were performed in R, using the 'geneplotter' package.

3.4.8. Calculation of feature expression values

Expression estimates for genes as well as individual exon, intron and intergenic features were determined by mapping of reads to transcript or genomic sequences and then calculating the observed average coverage (AC) of mapped reads across the base positions of the feature coordinates (see **Equation 3.1**). These estimates were then adjusted to account for varying depth between libraries, resulting in a normalized

average coverage (NAC) (see **Equation 3.2**). The following describes the strategy for specific feature types.

Exon junction and boundary expression estimates were determined by mapping reads to discrete N-mer sequences generated from EnsEMBL annotations. This resulted in a systematic bias towards lower expression estimates for exon junctions and boundaries. This is expected because, while a particular 62-mer junction sequence covered by all theoretically possible 42-mer reads has a coverage potential of 923 sequences bases (cumulative coverage of all possible distinct reads corresponding to the junction sequence), an exon of the same size within the bounds of a transcript can have up to 2,604 bases of potential coverage (i.e. ~2.8 times). This assumes that we require perfect read matching and that the exon region is within a transcript large enough to allow all possible overlapping 42-mers from both sides. Neither of these assumptions were entirely true in our analysis so we used the data to empirically measure the bias between exon and junction expression values. Specifically we used the exon and canonical exon junction expression estimates for genes expressed above background to estimate the observed bias. The median expression estimate for exons was ~2.2 times that of exon junctions. When displaying or summarizing exon junction values, they were corrected by this factor to make them more directly comparable to exon expression estimates.

To estimate the expression for an entire gene locus, all of the non-redundant exon base positions within the bounds of the EnsEMBL gene were considered. In contrast, transcript specific expression was calculated by taking the average of exon region and exon-exon junction expression estimates identified as unique to a particular transcript. In EnsEMBL (version 53), there are 36,953 genes with 62,371 total transcripts. 25,879 of these genes have only a single transcript and for these transcripts, calculating the expression simply utilizes all exon regions and junctions. However, 36,492 transcripts are distributed among the remaining 11,074 multi-transcript genes and expression could be individually assessed for 28,636 (78.5%) of them. Transcript-specific exon regions and exon-exon junctions are close to equally represented in the pool of features used for these measurements (62% of the transcripts had a specific exon region and 59% had a specific exon-exon junction).

3.4.9. Library depth and feature discovery

The relationship between library depth and feature discovery or 'saturation' was examined by sampling reads (without replacement) from the entire data set and after every 100,000 reads the number of events detected above a particular minimum cutoff level was recorded. The cutoffs examined were arbitrarily chosen as 1X, 5X, 10X, 50X, 100X and 500X. The number of events detected at each cutoff was determined for genes and known exon-exon junctions as well as individual exon, intron and intergenic base positions. For genes, and exons, the average coverage value (see **Equation 3.1**) had to exceed the cutoff level. For individual base positions, the position had to be covered by reads in excess of the specified cutoff. In each plot describing library depth and feature discovery, the x-axis corresponds to library depth expressed as a percentage of the total library depth (from 0% of reads sampled to 100%). Feature discovery was expressed as either the cumulative percent of all possible features discovered or the change in this percent relative to the previous sampling iteration. In other words, the percent of features discovered was simply the cumulative number of genes, exons, exon bases, intron bases, and intergenic bases exceeding the cutoff divided by the total number of features of that type (i.e. ~37k genes, ~218k known junctions, ~70 million exon bases, ~594 million intron bases and ~783 million intergenic bases). As discussed previously, only non-repetitive portions of the genome were considered. Plots of percent library depth versus percent of features discovered were generated in R. Power curves of the form $Y = aX^b$ were fit by non-linear least squares estimation. R^2 values were calculated as a Pearson correlation of the fitted Y values from the model fit versus the actual Y values and then raised to the power 2.

3.4.10. Determining expression above background

The goal of evaluating whether a feature is expressed above 'background' was two-fold: (1) to reduce the possibility of false positive expression events that correspond to mapping, PCR amplification, library construction or other artifacts and (2) to distinguish between sequence observations that represent true mRNA expression versus those that correspond to genomic DNA or un-processed RNA contamination and random transcription events. To estimate the level of genomic DNA contamination and random transcription noise[25] contributing to expression estimates, the distribution of expression estimates for all 'silent' intergenic regions was plotted (**Figure 3.18**). While it is likely

that these regions have a low probability of biologically meaningful expression, it is possible that they contain a small number of genes not currently annotated in EnsEMBL and never previously encountered in an mRNA or EST library. For this reason the 95th percentile was chosen as an upper estimate of the level of intergenic noise (hereafter referred to as the intergenic expression cutoff). In addition to intergenic noise, which leads to random low-level background expression signal across the genome, additional noise was expected within the boundaries of known genes from unprocessed RNA contamination (i.e. RNA that has not completed the process of splicing, poly-adenylation and export from the nucleus to the cytoplasm). Unlike the intergenic noise, this source of noise was expected to be positively correlated with gene expression level. To illustrate this, we plotted the expression estimates of all 'silent' intron regions against the expression estimate for the gene locus they corresponded to (**Figure 3.17**). Despite the low expectation of any expression from silent intron regions, the expression estimates for these regions were moderately correlated with gene expression level (Spearman correlation coefficient of 0.52). To derive an estimate for the level of expression signal which can be potentially attributed to the intragenic noise, we used this correlation to derive gene-by-gene expression cutoffs (hereafter referred to as the intragenic expression cutoff). Specifically we generated a scatter plot of gene and intron expression values and fit a linear model corresponding to the 95th percentile (depicted as a dotted line in **Figure 3.17**). The coefficients of the 95th percentile fit were used to estimate the upper limit of intragenic contamination for each gene. The final cutoff used was either the intergenic or intragenic cutoff, whichever was higher. Both the intergenic and intragenic cutoffs were assessed independently for each library to control for differences in contamination between libraries.

For all feature types, in order for the feature to be considered expressed, it had to exceed the expression level cutoffs as described above, but also had to be covered at 1X or greater depth over, 75% or more of the base positions of the feature. The purpose of this criterion is to prevent features that are covered at only a single position (possibly corresponding to a mapping or PCR amplification artifact). This ensures that all expressed features are covered by multiple independent read sequences and never simply a large number of the same sequence resulting in an average coverage value high enough to exceed the expression level cutoff. We believe that the approach described here should largely eliminate sequence artifacts from the expressed feature

lists and result in a false discovery rate of at most 5% among features called as expressed.

3.4.11. Estimating the total copy number of genes expressed in a cell

Traditionally, the number of mRNA transcripts present in a cell has been estimated by comparing the RNA/DNA ratio, the weight of RNA that can be obtained from an estimated number of cells and an estimate of the average molecular mass of an mRNA molecule[28]. We estimated the cumulative copy number of genes expressed in a single cell directly from the expression data for the MIP101 library by making a number of simple assumptions as follows: (1) a cell is a self contained unit, that expresses all genes for its maintenance and growth; (2) the minimum copy number of a gene is 1 copy per cell; (3) each gene expresses only a single transcript (4) our data correctly identified all expressed genes; and (5) our data provides an accurate estimation of the relative difference in expression level between the least and most abundant gene. This last assumption assumes that increased library depth would not increase the observed dynamic range. These are simplifying assumptions that are not necessarily true, but are adequate for first approximations. Starting with these assumptions we obtained the expression values (raw average coverage) for all 12,396 genes classified as expressed above background. This value was 7.6 for the least abundant gene and 47,372 for the most abundant gene. We then set the estimated 'copy number' of the least abundant gene to be 1. All other expression values were then scaled relative to the expression value of this gene. For example, a gene with an expression value of 15.2 would be assigned a copy number of 2 and so on. Estimated copy numbers generated by this approach ranged from 1 to 6,242. We then calculated the cumulative copy number of all genes by adding these scaled copy number values, resulting in a grand total estimate of ~460,000 copies. Due to assumptions 3, 4 and 5, this is likely to be an underestimate.

3.4.12. Differential expression analysis

Differential expression (DE) values were determined as the log2 difference in normalized average coverage values (see **Equation 3.2**) between 5-FU sensitive and resistant cells for all feature types. Before calculating differential expression values, all features that were not expressed above background in at least one library were

removed. A value of 1 was then added to normalized average coverage values before converting to log2 scale (to stabilize variance). The log2 difference was converted to a more intuitive fold-change value by raising the log2 difference to the power of 2. The MIP/5FU library was chosen as a reference for differential expression values such that a negative value implies loss of expression in 5-FU resistant cells relative to sensitive and a positive value implies a gain of expression in resistant cells. A two-sided Fisher's exact test was used to calculate a p-value for the difference in feature expression between two libraries. Since this test accounts for differences in counts between libraries, raw average coverage values were used when calculating p-values. The resulting p-values were adjusted to account for multiple testing by using the Benjamini and Hochberg method[50] implemented in the 'multtest' package of R (www.R-project.org). Features were considered to be differentially expressed if their fold-change exceeded two-fold and their corrected p-value was less than 0.05.

3.4.13. Alternative expression analysis

Alternative expression analysis was performed as an extension to the differential expression analysis described above. The purpose of this analysis was to identify genes exhibiting differential expression of features that might be indicative of a shift in splicing, transcript initiation or poly-adenylation of a specific isoform rather than a shift in expression across the entire locus. Candidate differential splicing events were assessed by calculating a 'splicing index' (SI) value for each feature (see **Equation 3.3** for details). Briefly, this value provides a measure of the change in expression of a feature between two conditions relative to the change in expression of the entire gene locus between the two conditions (5-FU sensitive and resistant cells in this case). SI values were only calculated for a feature if the feature and the gene it corresponded to were expressed above background in at least one of the two conditions being compared. In addition to the SI value, for each feature we also calculated a 'reciprocity index' (RI) and 'percent feature contribution' (PFC) value (see **Equation 3.4** and **Equation 3.5** for details). The purpose of the RI value was to identify those cases where the change in expression of a feature was consistent with reciprocal differential expression of an isoform relative to the gene locus. In other words if a gene was up-regulated overall between 5-FU sensitive and resistant cells while a particular feature such as an exon-skipping junction was down-regulated, a high RI value would be

produced. Similarly, the PFC value gauges the magnitude of differential expression of a feature relative to the differential expression of the locus. For example, if the expression level of a gene is unchanged overall, but an individual feature such as a single exon is changing dramatically, this will result in a large PFC (approaching 100% if the overall gene expression is completely unchanged). Features considered to be candidate differential splicing events were defined as those features having an SI value of at least 1, a significant DE value (p-value < 0.05 after multiple testing correction), and a PFC value of at least 50. The candidate differential expression and alternative expression gene lists were combined and ranked according to the maximum DE or SI value for all features of each gene. The top 50 genes from this list are shown in **Table 3.8** and the complete list is available online (www.AlexaPlatform.org).

3.4.14. Pathway analysis

Analysis of differentially or alternatively expressed genes with respect to pathways was performed using the Ingenuity Pathway Analysis Software (www.ingenuity.com; Ingenuity Systems. Redwood City, CA.). Starting with a list of 1,478 genes with one or more significant differentially expressed features, this software was used to identify statistically significant enrichment for particular gene interaction networks and annotated biological functions or pathways. It was also used to visualize differential expression in the context of specific pathways with known relevance to 5-FU action or general drug resistance mechanisms.

3.4.15. Software implementation and availability

All database annotation, read mapping, expression analysis and alternative expression event discovery was developed in a Linux environment using our previously published ALEXA platform[2] as starting point. The computational platform consists primarily of Perl scripts which interact with mySQL (http://www.mysql.com/) and Berkeley DB (http://freshmeat.net/projects/berkeleydb/) Databases. A Beowulf style cluster consisting of approximately 1500 CPUs was used for database annotation, read mapping and generation of expression estimates. This cluster is managed by the open source cluster application resources (OSCAR) management system (http://oscar.openclustergroup.org). All source code, user manuals, supplementary

methods, data visualizations, etc. are available from the ALEXA website, www.AlexaPlatform.org.

3.4.16. Statistics and data visualization

Generation of all statistics, figures and graphs was performed in R using the 'Base', 'caTools', 'gcrma', 'geneplotter', 'multtest', 'nortest', 'RColorBrewer', and Bioconductor[51] packages.

In each box plot, the box portion displays the median of the distribution flanked by the lower and upper quartiles (i.e. the 25th and 75th percentiles respectively). The whisker portions of each box plot indicate the median plus or minus 1.5 times the interquartile range. If no observed values exceed the interquartile range, then the whisker is set to the largest or smallest observed value. Values which exceed 1.5 times the interquartile range are indicated by dots.

A web accessible data viewer, 'ALEXA-Seq' was created to display expression data for all genes. Summaries of the top expression events as well as all significant differential expression and splicing events are provided as well as search functionality to find specific genes. Indexing of all data to allow searching was accomplished by the open source search engine library, 'Xapian' (http://www.xapian.org/) and the web search package 'Omega'. Plots displaying the expression of each gene and the expression and differential expression of every feature of every gene across libraries were generated in R using the scalable vector graphics (SVG) module of the Cairo package (http://cairographics.org/). Additional visualization of the annotation, expression and differential expression of every feature of every gene is provided by links to the UCSC genome browser which automatically load custom track files containing our data in 'gene feature' (GFF) and 'wiggle' formats [43].

Equation 3.1. Average coverage (AC)
The expression level of an entire gene, transcript, exon region or other feature is determined by considering the cumulative base coverage of that feature by reads that have been mapped to the corresponding segment of the genome or exon-exon junction sequence. Since expression is determined by counting of reads generated by random sampling of cDNA fragments representing RNAs, larger RNAs will tend to produce more reads irrespective of their expression level. To correct for this bias, expression of any feature is normalized to the size of that feature and expressed as an average coverage (AC), where the cumulative base coverage 'N' of the ith feature is divided by its base count 'S'. The cumulative base coverage 'N' refers to the total count of bases from sequence reads aligning within the coordinates of the feature. The base count 'S' refers to the annotated length of the feature (i.e. transcript length, exon length, etc.).

$$AC^i = \frac{N^i}{S^i}$$

Equation 3.2. Normalized average coverage (NAC)
To allow direct comparisons between libraries of different depth an average coverage value normalized to an arbitrary library size of 10 billion bases for the ith feature of each library was calculated for a library with X successfully mapped bases of sequence as follows.

$$NAC^i = AC^i \times \frac{10{,}000{,}000{,}000}{X}$$

Equation 3.3. Splicing index (SI)
The splicing index (SI)[52] for a sequence feature whose expression is compared between two libraries, 'A' and 'B', is calculated for the jth feature of the ith gene as shown below. Essentially, the expression of each feature is normalized to the expression level of the gene to which it belongs before comparison between two libraries. The term 'feature' refers to every transcript, exon, exon-exon junction, alternative exon boundary or intron of a particular gene model.

$$SI^i_j = \frac{\log 2(A_j - A^i)}{\log 2(B_j - B^i)}$$

Equation 3.4. Reciprocity index (RI)
The reciprocity index (RI) is inspired by the SI calculation and is applied when a putative alternative isoform is differentially expressed in the opposite direction to that of the gene locus and assesses the degree to which this occurs in a 'balanced' fashion. The more balanced, the higher the score. For example, if a novel exon-skipping isoform is up-regulated by 4-fold while the exon-containing isoform is down-regulated by close to 4-fold as well, this will result in a high RI score. Similar to the SI score, the RI for a sequence feature, compared between two libraries, 'A' and 'B' is calculated for the jth feature of the ith gene as shown below.

$$RI_j^i = \frac{\left|\log 2(B^i - A^i)\right| + \left|\log 2(B_j - A_j)\right|}{\left|\log 2(B^i - A^i)\right| - \left|\log 2(B_j - A_j)\right|}$$

Equation 3.5. Percent feature contribution (PFC)
The percent feature contribution (PCF) calculation in contrast to the RI seeks to identify cases where the observed change in a feature's expression between two conditions is large relative to the change in expression at the locus overall and the change can be attributed to the expression of the feature itself. An SI value can be large in cases where a sequence feature such as an exon is unchanged between two conditions but the entire gene is differentially expressed. These sometimes correspond to annotation artifacts where the feature in question is not really expressed at all. The PFC can be used to filter a list of events with high SI values to eliminate these spurious cases. Similar to the SI score, the PFC for a feature, compared between two libraries, 'A' and 'B' is calculated for the jth feature of the ith gene as depicted below.

$$PFC_j^i = \left(\frac{\left|\log 2(B_j - A_j)\right|}{\left|\log 2(B_j - A_j)\right| + \left|\log 2(B^i - A^i)\right|} \right) \times 100$$

Figure 3.1. Annotation of sequence features
A hypothetical region of the genome containing two gene loci (green and yellow) illustrates: (1) the source of gene models used for annotation, (2) how these models are supplemented by additional data from mRNA and EST alignments, (3) how features are defined by combining gene models and expressed sequence data, and (4) how features are named as used in this report. The term 'novel', as applied to exon-junctions and alternative exon boundaries indicates a feature not currently represented in an EnsEMBL transcript. The term 'active' as applied to intron and intergenic regions describes a portion of an EnsEMBL intron/intergenic region that has EST or mRNA evidence for the existence of a putative exon not currently annotated in EnsEMBL. A novel exon boundary corresponds to a hypothetical alternative exon donor or acceptor site which results in the incorporation of intronic sequence flanking a known exon. Feature naming conventions: Exon region (ER). 'ER1a' and 'ER1b' describe exon region 1, with two parts corresponding to the boundaries of overlapping exons. 'E1-Da' describes exon 1, donor site 'a'. 'I1' refers to intron 1. 'I1-AR1' and 'I1-SR1' refers to 'active' and 'silent' regions within intron 1 defined by EST alignments. 'IG' refers to an intergenic region. 'ER2-ER3' refers to an exon-exon junction consisting of the connection of the donor site of exon 2 and the acceptor site of exon 3.

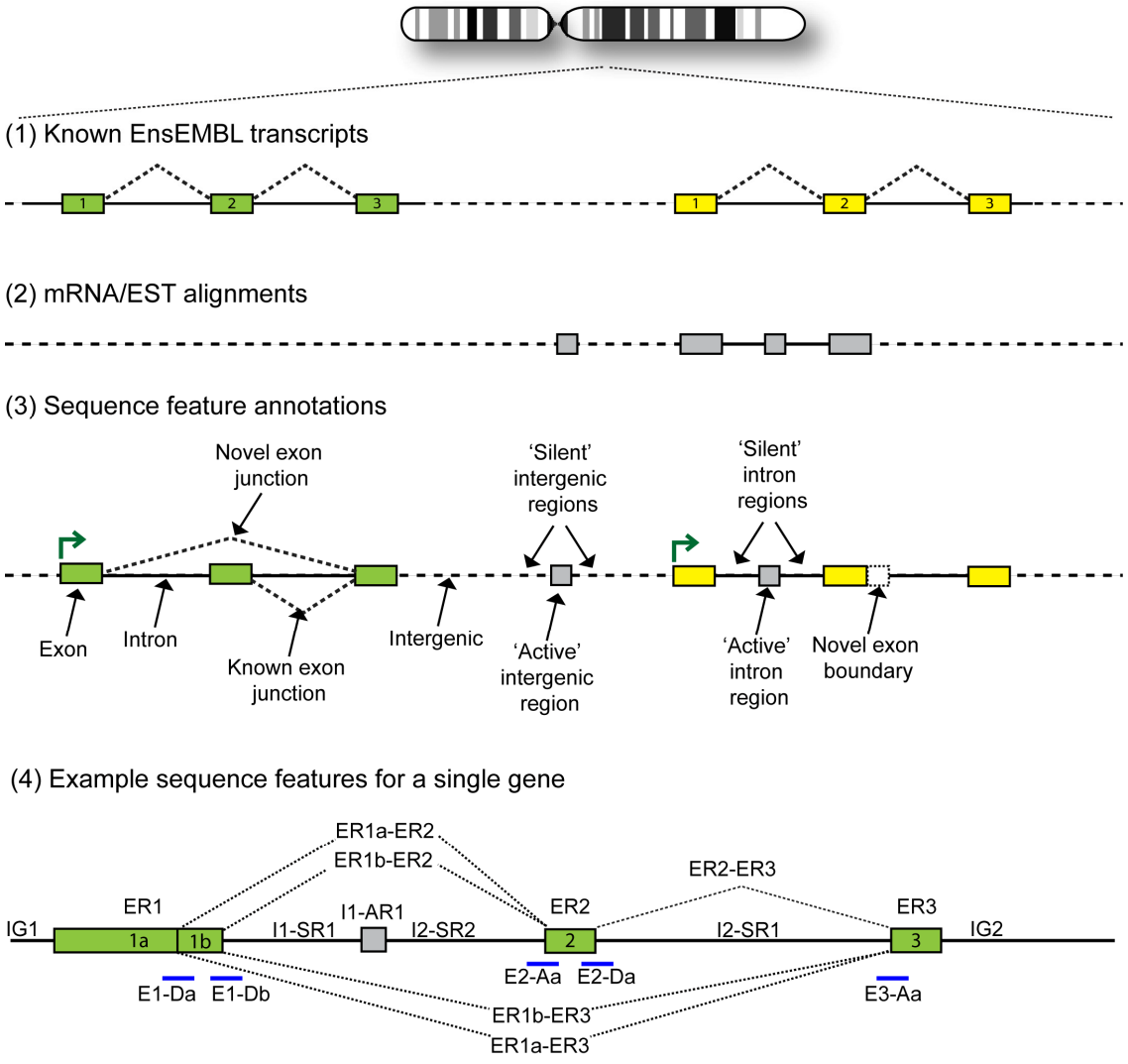


Figure 3.2. Illustration of read data generation
Total RNA was isolated from 5-FU sensitive and resistant colorectal cancer cell lines (images of cultures at 100X magnification). Messenger RNA (depicted in yellow) was purified by polyA+ selection, followed by cDNA generation with random hexamers, fragmentation by sonication, selection of 190-210 bp fragments using gel electrophoresis, ligation of sequencing linkers and sequencing of the ends of 10^6-10^7 such fragments with an Illumina GAII sequencing device. End reads are depicted as two black segments connected by dotted lines. The resulting reads were then mapped to a database consisting of genome, transcript, and exon junction sequences. The coordinates of mapped reads were then used to estimate the expression, differential expression and alternative expression of ~4 million sequence features (**Figure 3.1**). Refer to the **Methods** for additional details.

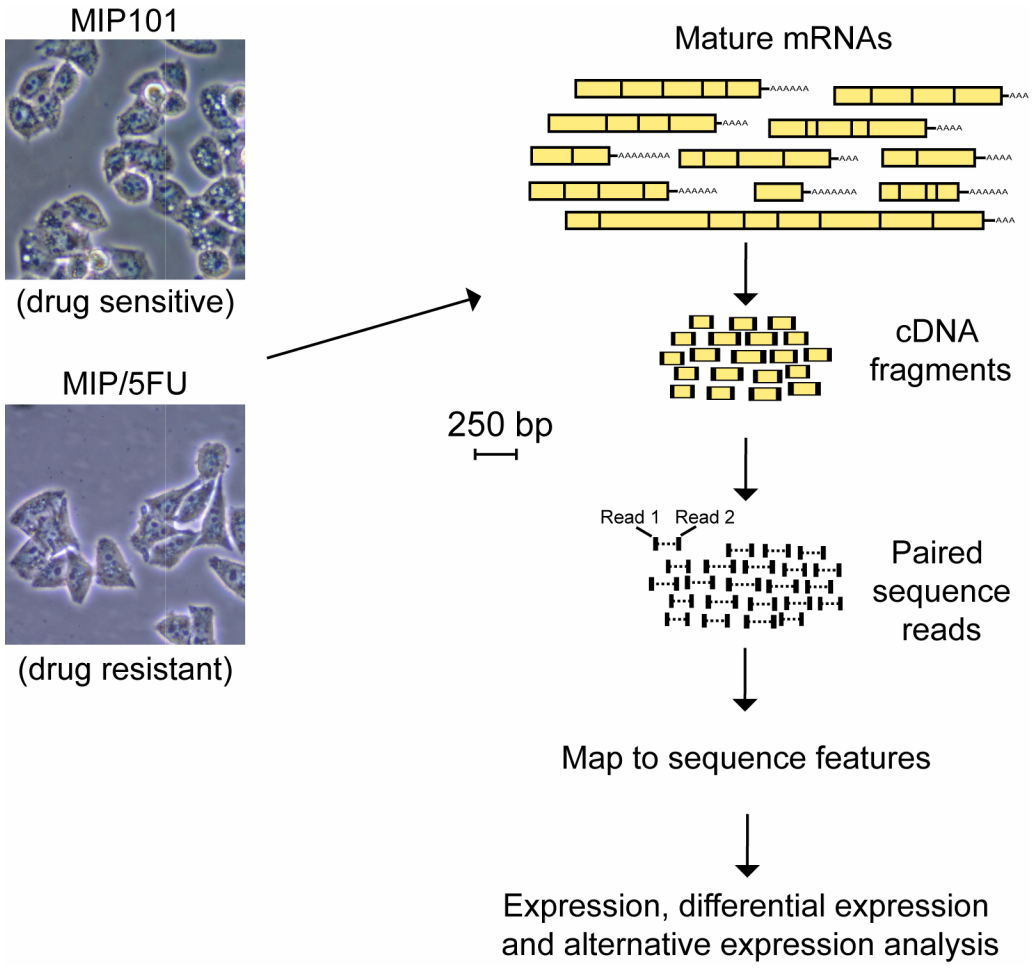


Figure 3.3. Overview of alternative expression analysis

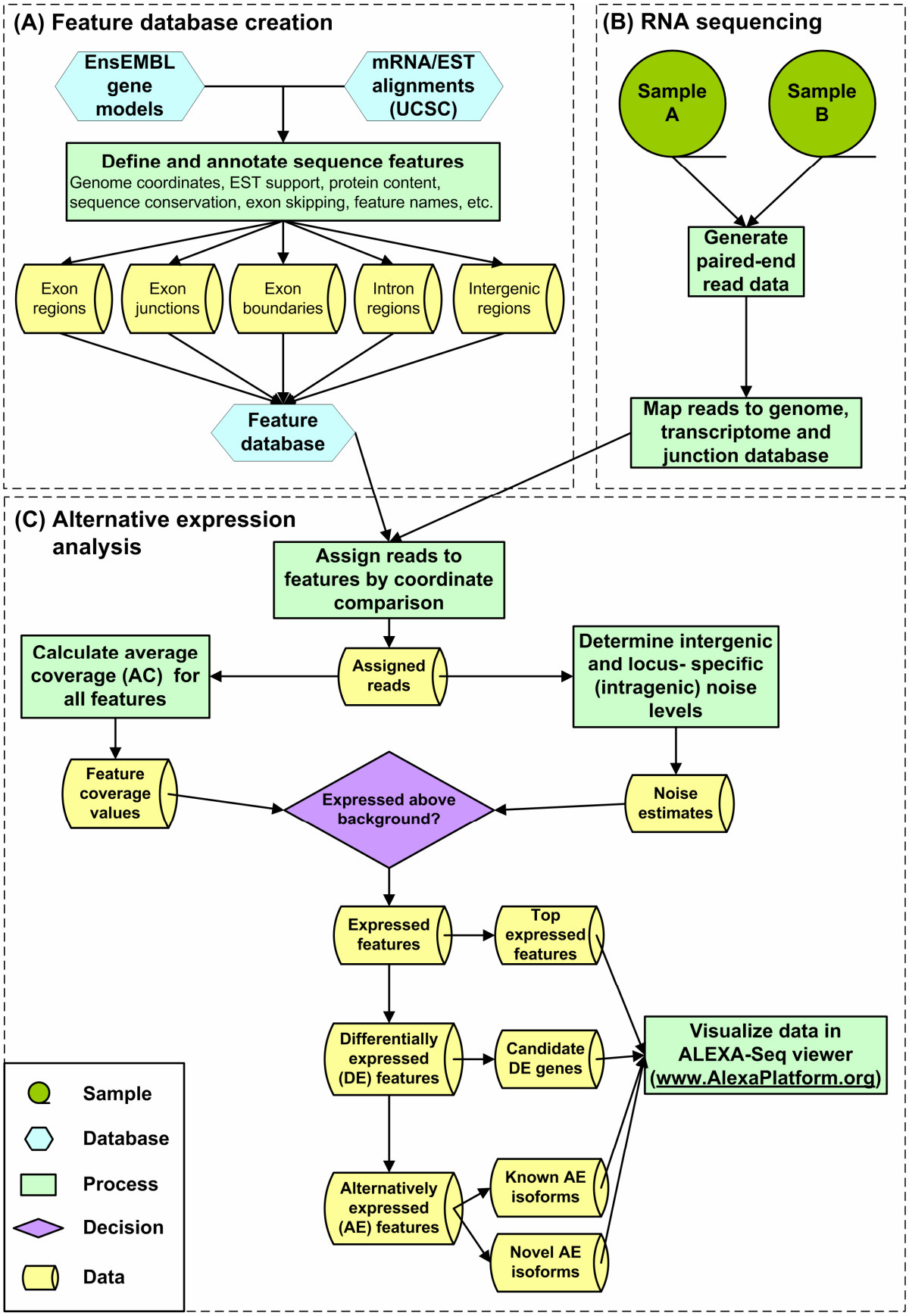


Figure 3.4. Distribution of fragment sizes
A histogram showing the distribution of fragment sizes inferred by mapping paired reads from the MIP101 and MIP/5FU libraries (see **Figure 3.2**) to known transcripts and then calculating the number of bases between the outer coordinates of the mapped reads within the transcript. Only those fragment sizes up to the 95th percentile of all fragment sizes (305 bp) are depicted here. Thus, 5% of all fragment sizes were greater than 305 bp. The minimum, median, mean and maximum fragment sizes of this distribution are provided in the legend.

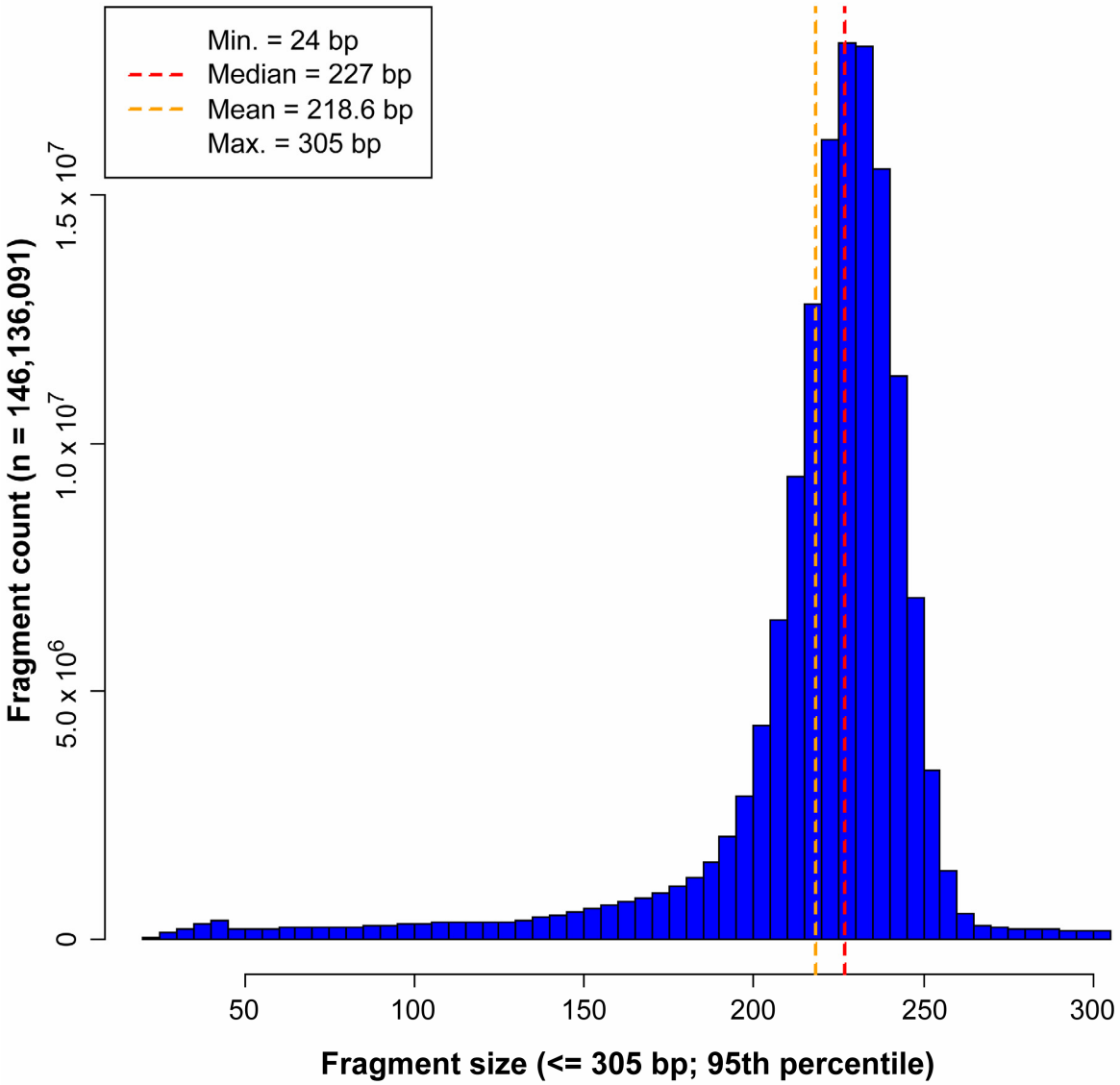


Figure 3.5. Distribution of average Illumina read qualities
The sequence quality values generated by an Illumina GAII sequencing platform (see **Methods** for details) for all positions of paired 36- and 42-mer reads were binned according to their source lane and position within paired reads. The average quality for all qualities of a lane was then calculated for each lane for each read position bin. The averages for 39 flowcell lanes at each read position were then plotted as box plots. The resulting graph shows the distribution of average quality scores across all 39 lanes of data at each read position. Position 1-42 (green) correspond to read 1 of a pair and positions 43-84 (light blue) correspond to read 2 of a pair. Since the data correspond to a pool of 36- and 42-mer reads, positions 37-42 and 79-84 correspond to 42-mers only, and all other position correspond to both 36- and 42-mers.

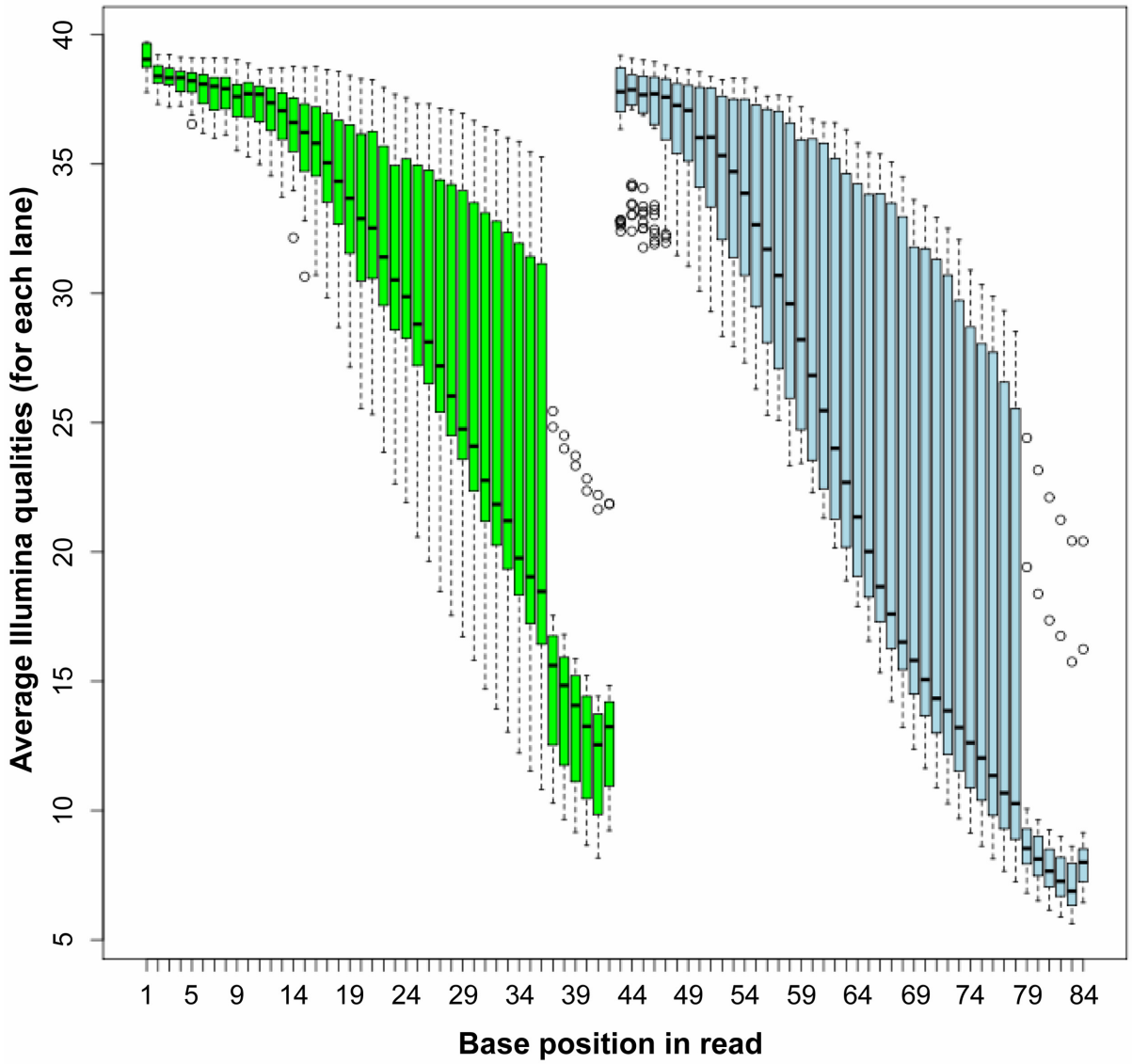


Figure 3.6. Distribution of read alignment lengths
262 million paired 36- and 42-mer reads from the MIP101 and MIP/5FU libraries (i.e. 524 million individual reads) were aligned to a database of known EnsEMBL transcripts. In this analysis read 1 and read 2 of each pair were considered separately. The distribution of alignment lengths for the resulting ~375 million read-to-transcript alignments are displayed as a histogram. The majority of 36- and 42-mer reads aligned over their complete length. The distribution reflects the fact that this library consisted mostly of 42-mers (83%) and a much smaller number of 36-mers (17%). Alignments greater that 42 bases in length suggest the occurrence of gaps relative to the database, while short alignments correspond to sub-sequence alignments (i.e. only a portion of a read aligns).

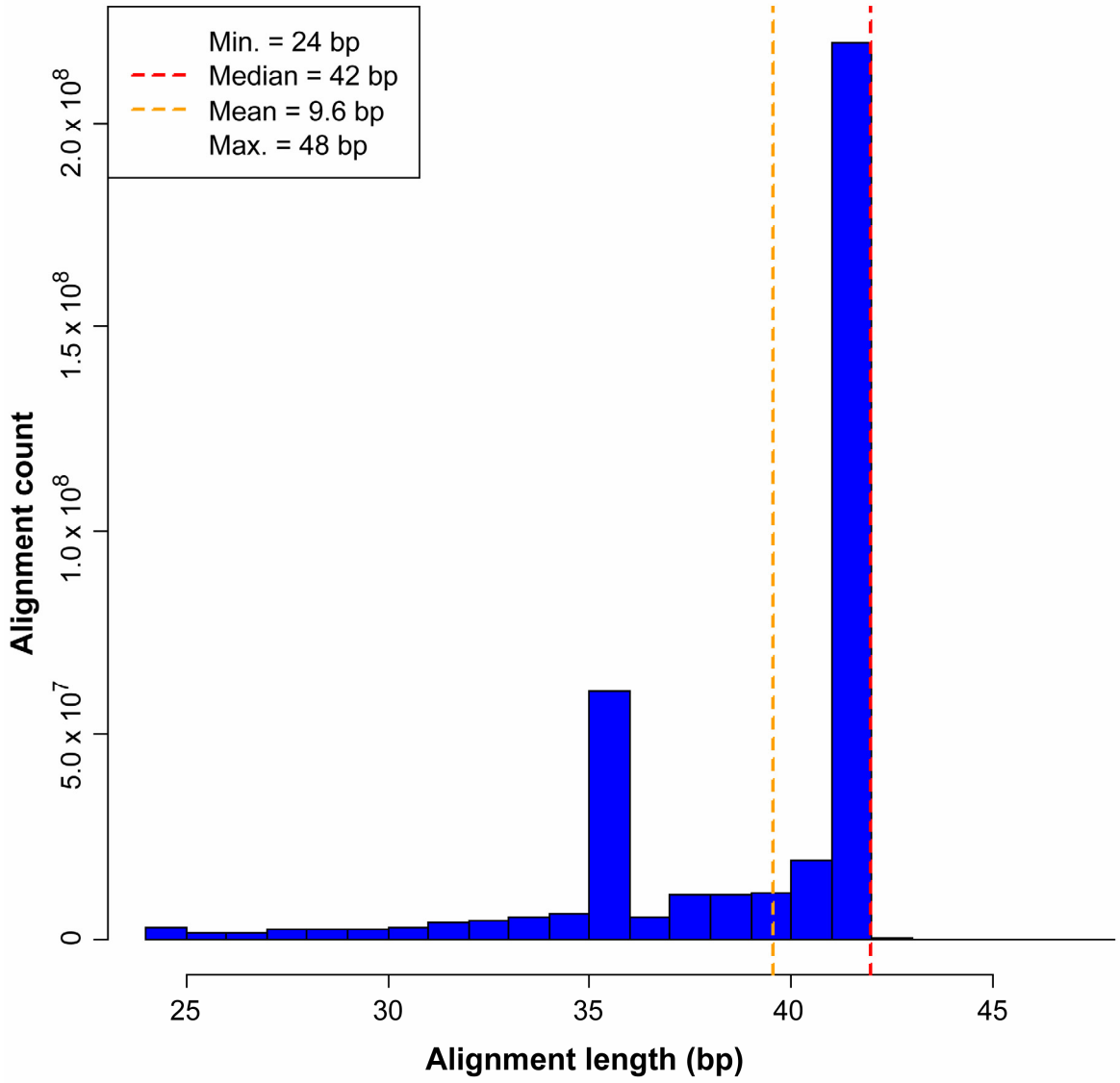


Figure 3.7. Position bias by transcript size
The relative position of a read within a transcript was calculated by determining the centre position of each read mapped to a transcript, dividing this position value by transcript length and multiplying by 100. Thus reads mapping to the 5′ end of a transcript produce values approaching 0, reads mapping to the centre of a transcript produce values near 50, and reads mapping to the 5′ end of a transcript produce values approaching 100. The distribution of these relative position values was plotted as a series of box plots, each corresponding to reads mapping to a particular range of transcript sizes (0 - 500bp, 500 - 1,000bp, etc.). Each box plot graphically displays the median, lower quartile (25th percentile), upper quartile (75th percentile), and the quartiles plus/minus 1.5 times the inter-quartile range. The width of each box plot is drawn proportional to the square root of the number of observations comprising the group (i.e. proportional to the number of mapped reads corresponding to each range of transcript sizes). Hence, a thick box indicates a larger number of reads than a thin box. Each of the 9 distributions below was compared to the distribution for transcripts of 2,500-3,000 (light green box plot, with a median read position of 50.1) by two-sample, two-tailed, Kolmogorov-Smirnov test. P-values (P) are reported for each of these comparisons.

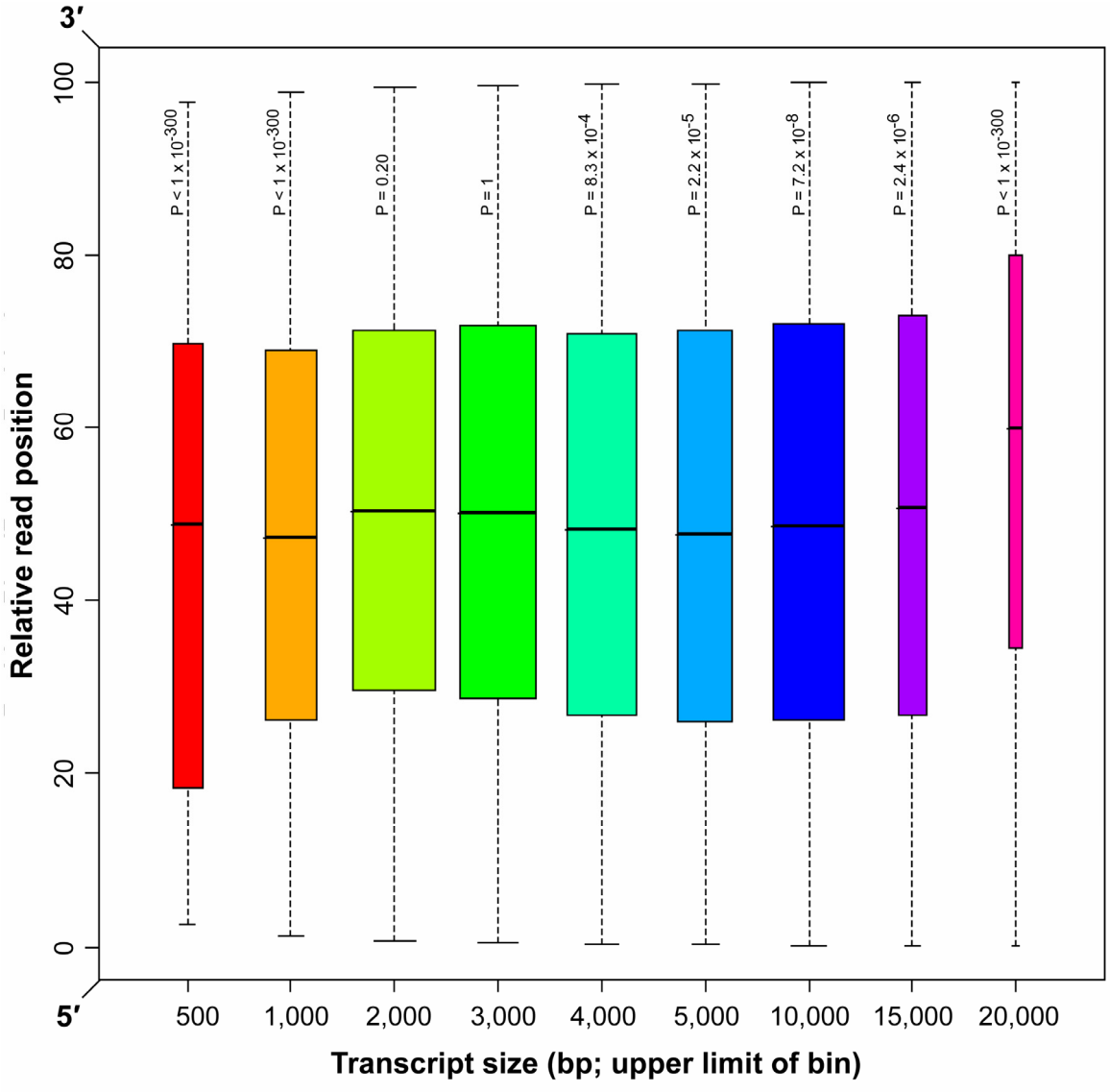


Figure 3.8. Read mapping summary
The pie chart depicts the percentage of ~262 million paired-end reads from the MIP101 and MIP/5FU libraries aligning to each feature type and reads that were filtered due to low complexity sequences, ambiguous bases (i.e. 'N's) and reads where both read 1 and 2 were duplicates. ~26 million reads (10%) had homology to human a transcript, but had an alignment score below our required threshold and an additional ~34 million reads (13%) did not map to any sequence in our database. If both reads of a read pair were identical or a perfect reverse complement of each other, the read was classified as a duplicate. Low complexity reads were identified by mdust (see **Methods** for additional details on read filtering and assignment to feature each type).

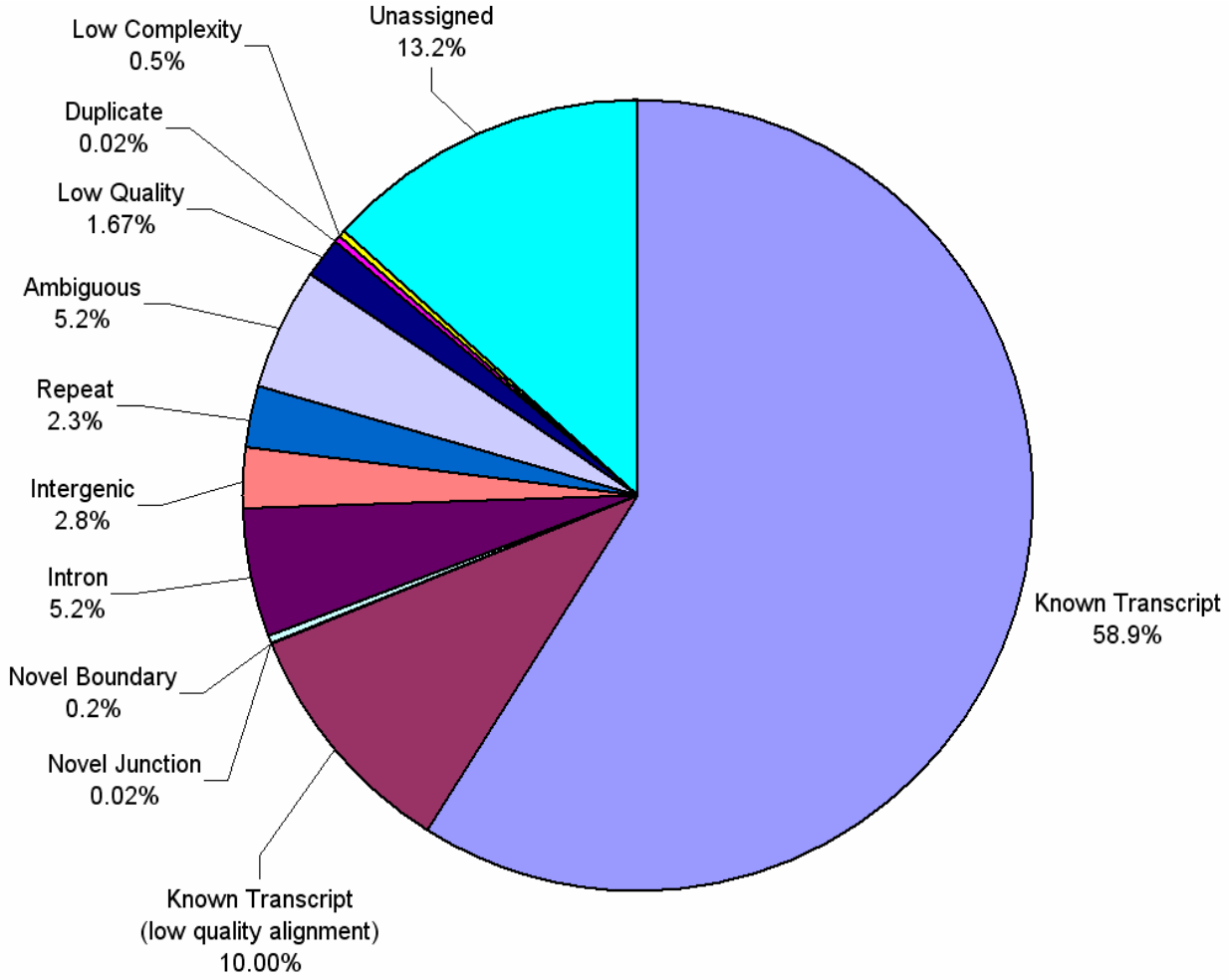
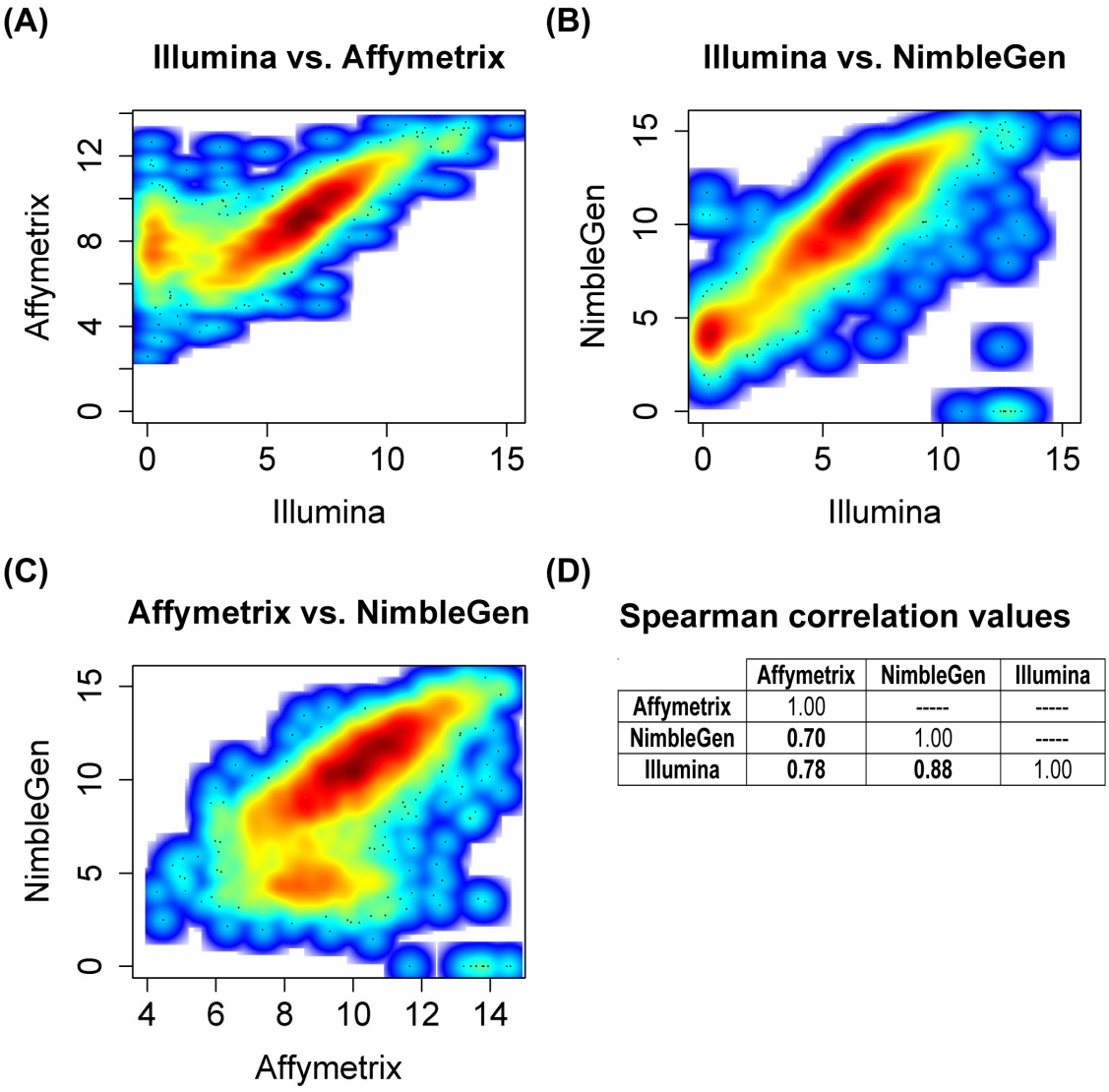


Figure 3.9. Comparison of expression estimates from three expression platforms
Log2 gene expression estimates were generated from the Illumina (WTSS), Affymetrix (exon microarray) and NimbleGen (splicing microarray) expression platforms using the same input RNAs (see **Methods**). The following figures represent data for only the intersection of genes profiled by all three expression platforms: Affymetrix exon arrays, custom NimbleGen splicing arrays[2] and Illumina WTSS (2,434 genes). Two-way comparisons (panels A, B and C) of these data were plotted as density plots using 'R'. The inset table (panel D) shows the Spearman correlation values for all three comparisons.



	Affymetrix	NimbleGen	Illumina
Affymetrix	1.00	-----	-----
NimbleGen	**0.70**	1.00	-----
Illumina	**0.78**	**0.88**	1.00

Figure 3.10. Comparison of differential expression from three expression platforms

Log2 differential gene expression estimates were generated for 2,434 genes profiled by the Illumina (WTSS), Affymetrix (exon microarray) and NimbleGen (splicing microarray) expression platforms (panels A, B and C). Panel D illustrates the overlap in the list of differentially expressed genes (>= 2-fold change) as a Venn diagram created with BioVenn[53].

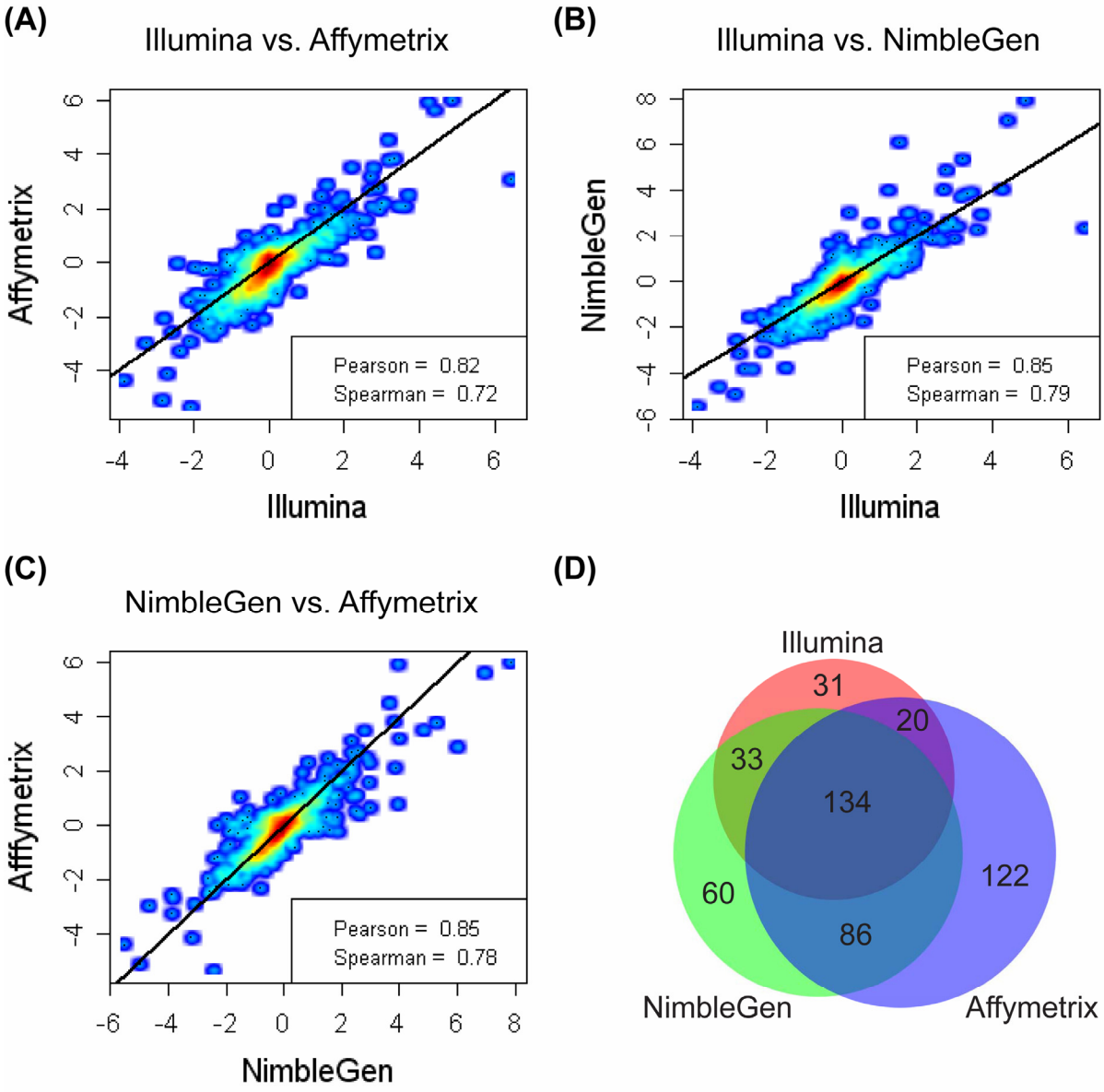


Figure 3.11. Comparison of expression estimates for the exons and introns of 100 housekeeping genes derived from three expression platforms
The log2 expression values of exons and introns for 100 housekeeping genes were plotted as box plots for three expression platforms (see **Methods**). Expression estimates were generated with the Illumina (WTSS), Affymetrix (exon microarray) and NimbleGen (splicing microarray) expression platforms. Note that, for display purposes a value of 1 was added to all expression estimates before converting to log2 scale. This reduces the dynamic range of the Illumina data while having little effect on the array data. See **Table 3.4** for the unadjusted dynamic range. 'Random' refers to random sequence oligonucleotides designed for estimation of background noise in microarrays.

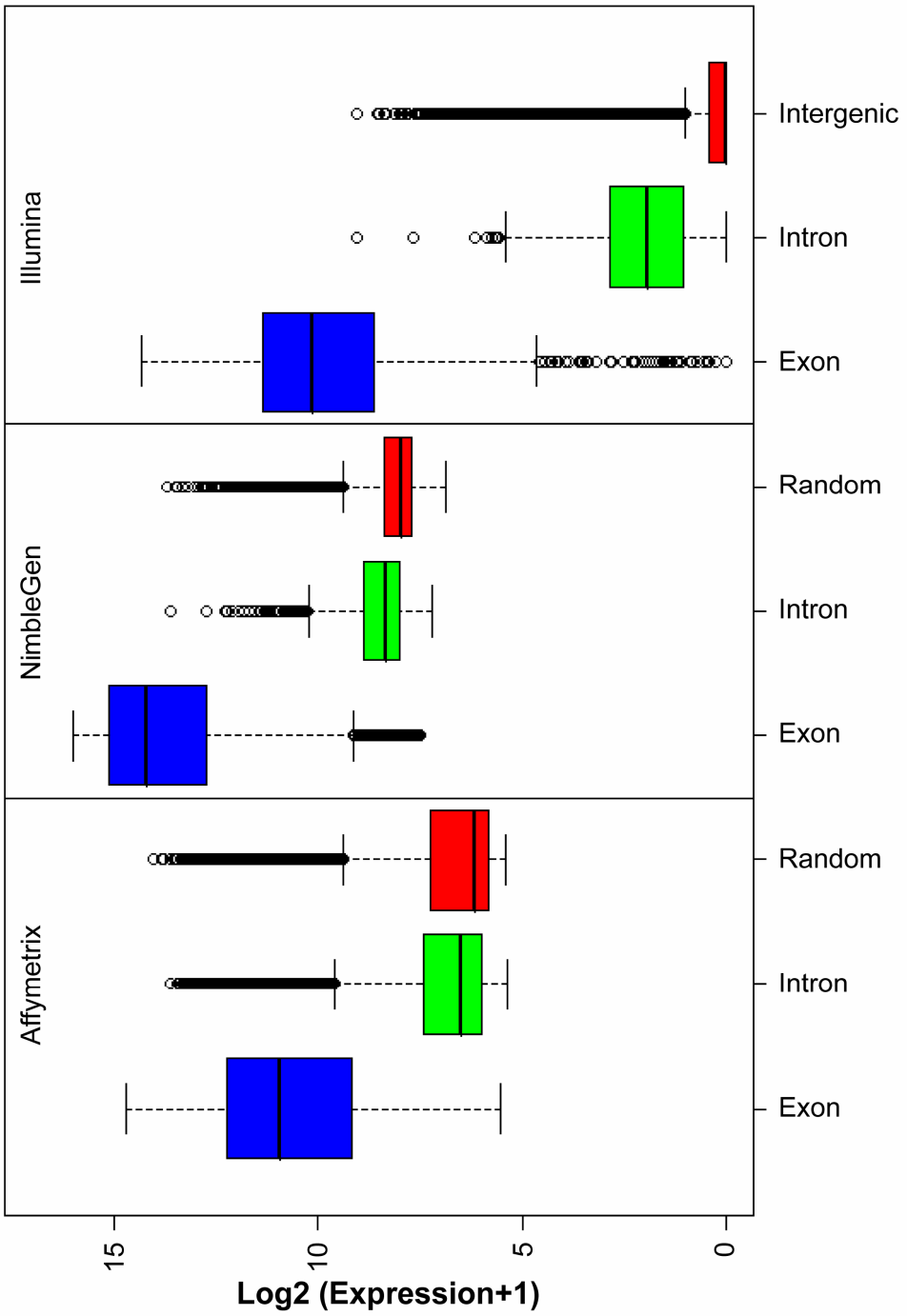


Figure 3.12. ROC curves comparing sensitivity and specificity between three expression platforms

Receiver operator characteristic (ROC) curves were plotted by calculating the sensitivity and specificity for three expression platforms with respect to the ability to correctly identify the exons of 100 housekeeping genes as expressed while at the same time correctly identifying the introns of these genes as not expressed. The three expression platforms consisted of Illumina (WTSS), Affymetrix (exon microarray) and NimbleGen (ALEXA splicing microarray). The Illumina platform achieved the highest ROC area under the curve (an indicator of sensitivity and specificity) followed by NimbleGen and Affymetrix.

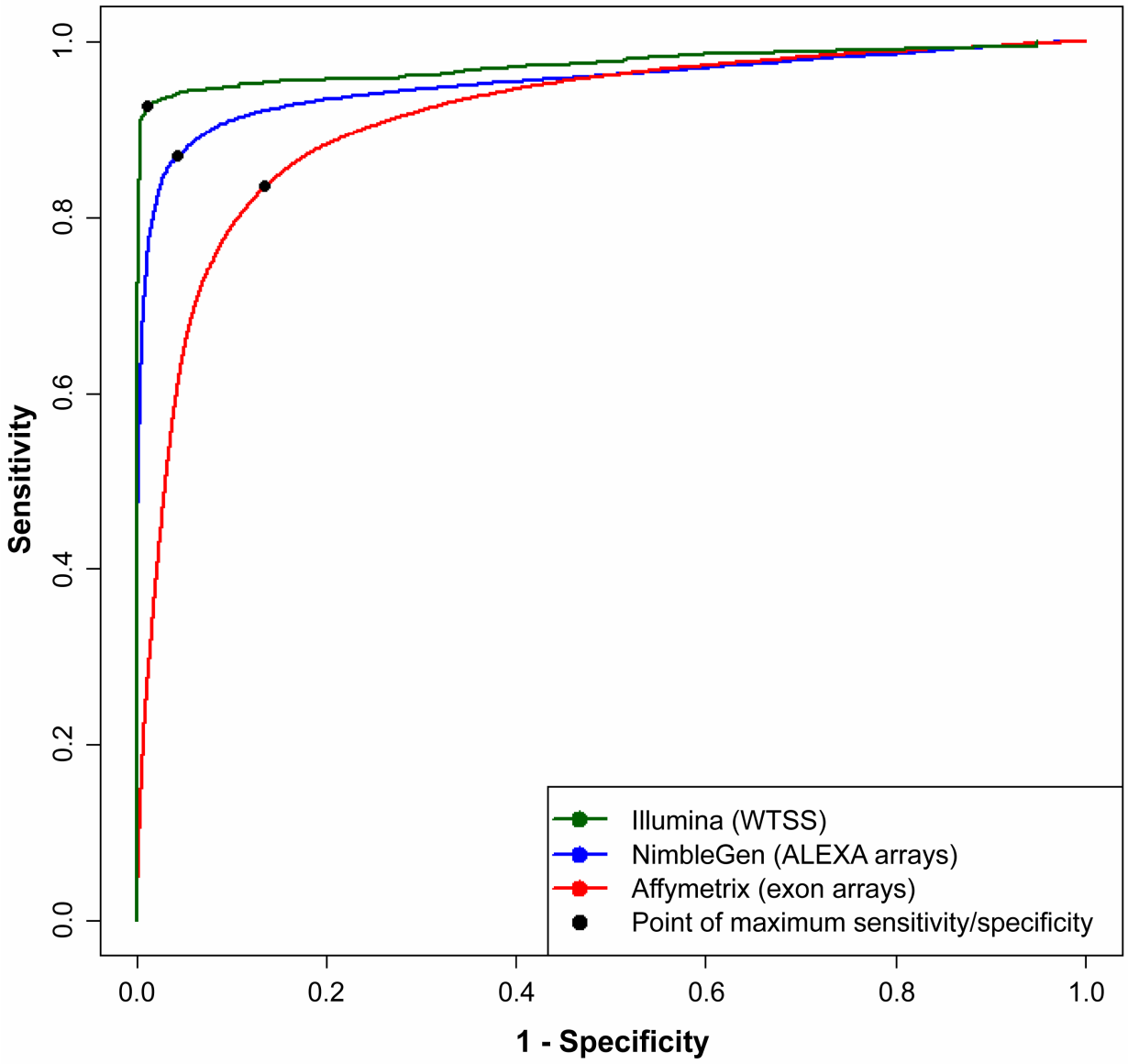


Figure 3.13. Distribution of percent gene coverage at increasing minimum coverage cutoffs

Box plots depict the distribution of percent gene coverage values (i.e. percent of exon bases of a gene covered by reads) at a particular minimum coverage cutoff (X) in the MIP101 library. 25,519 genes were detected by one or more reads in the MIP101 sequence library, but only the 12,396 genes determined to be expressed above background were used to generate the following box plots (see **Methods**). The red box plot shows that the median percent gene coverage for all genes was 99.8% if we required each base position to be covered at 1X depth or greater. When the coverage cutoff requirement was increased to 10X, the median percent gene coverage dropped to ~94.7%. 5,750 genes were covered over 50% or more of their bases at a minimum coverage of 100X.

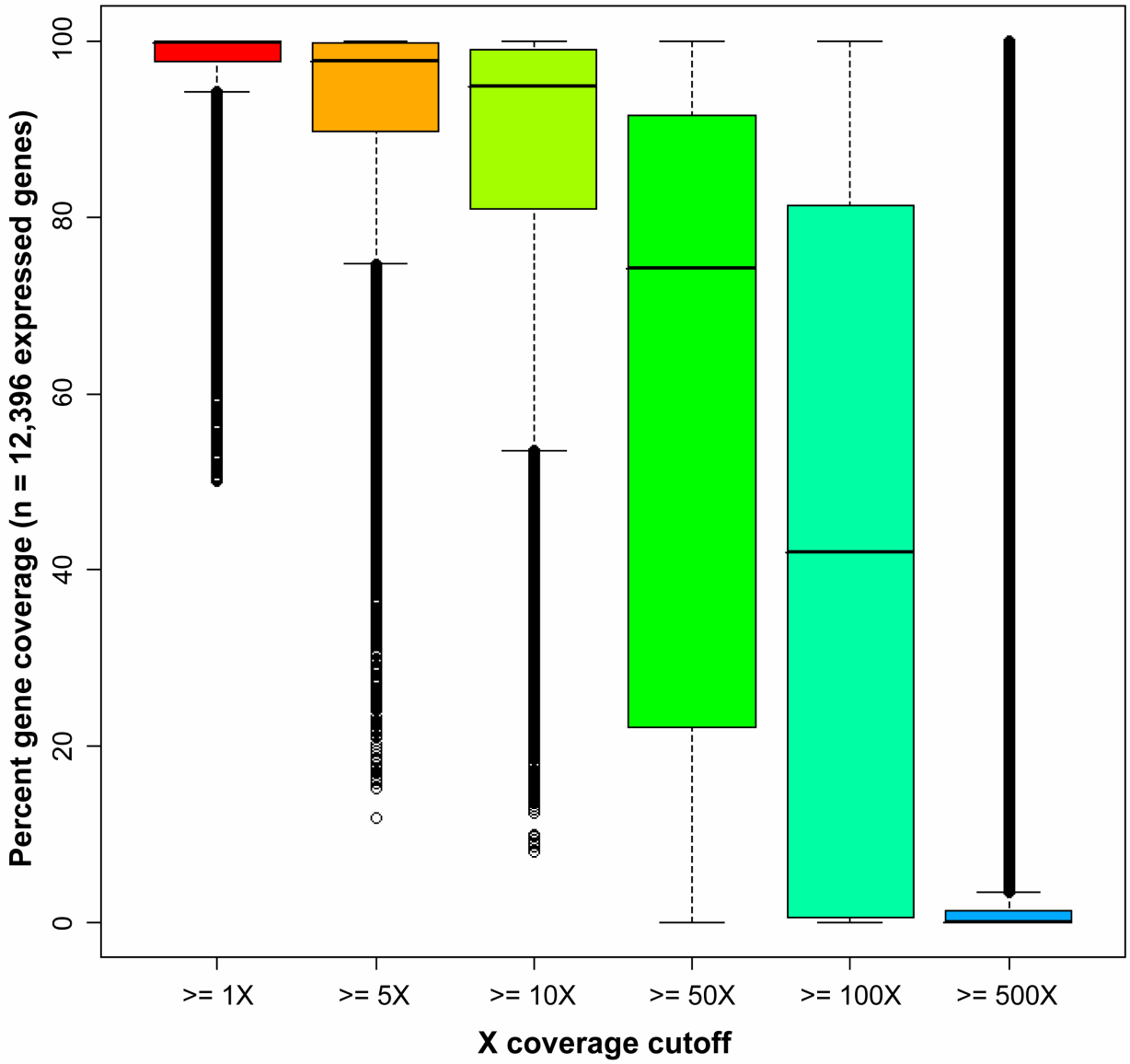


Figure 3.14. Coverage of expressed features as a function of library depth
The number of features (genes, exon junctions, and individual exon, intron and intergenic base positions) detected by 10 or more reads in the MIP101 library was expressed as a percentage of all possible features of each type (e.g. observed genes divided by the ~36,000 genes with annotations in EnsEMBL). Reads were sampled without replacement in blocks of 100,000 mapped reads and after each sampling the percentage of possible features detected was recorded. As library depth increased the number of detected features increased. Sampling was continued in 100,000 read blocks until 100% of the library had been incorporated. The following plot shows the relationship between percent library depth and percent of possible features detected at the 10X threshold. Each colored line represents percentages for a single event type: genes (blue), exon-junctions (green), exon base positions (red), intron base positions (black) and intergenic base positions (grey). Note that, even at 100% library depth, the rate of discovery of exon bases is still greater than that for intergenic bases indicating that exons expressed above background are still being discovered. **Figure 3.15** illustrates the change in discovery rate.

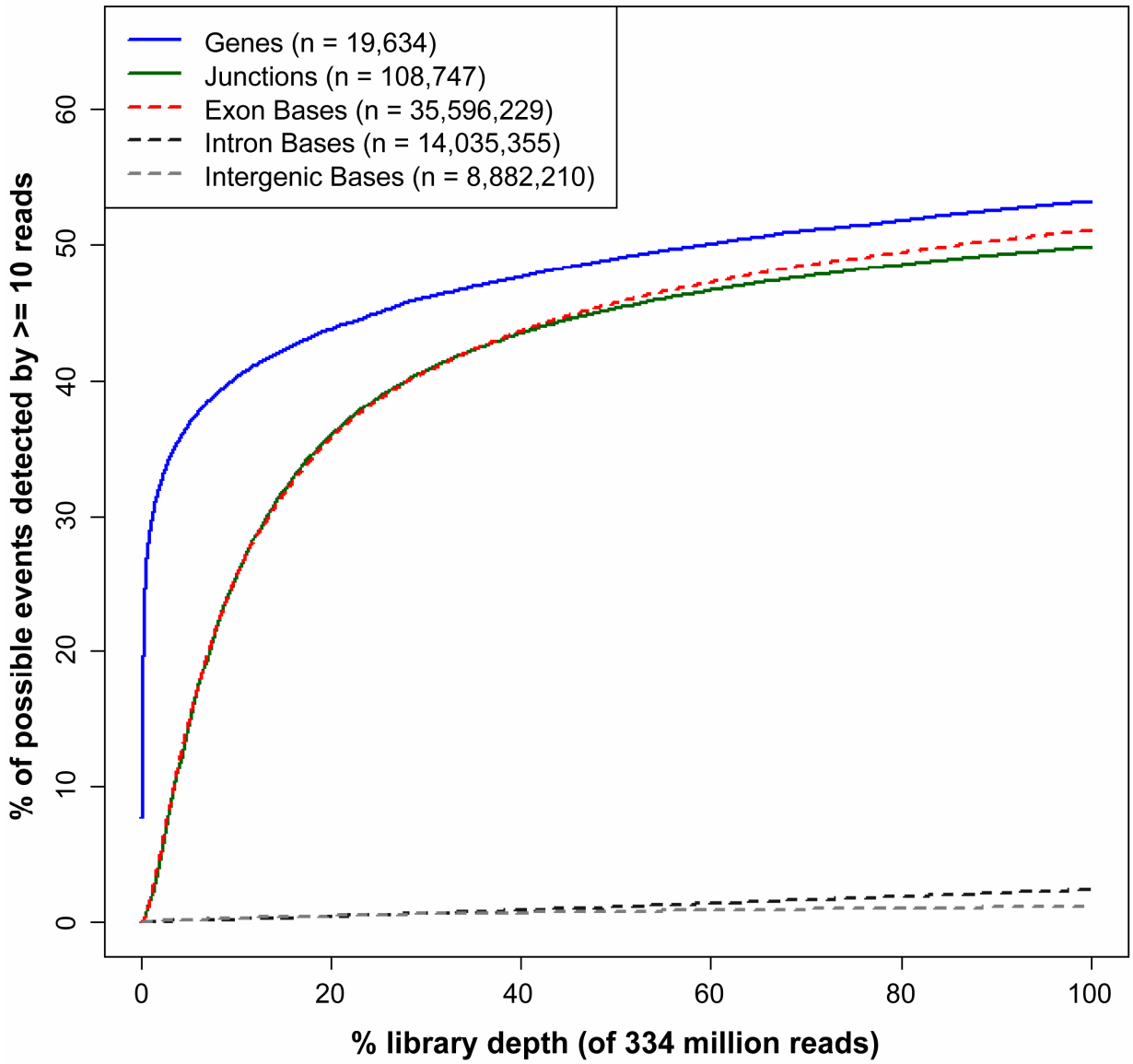


Figure 3.15. Change in percent discovery rate with increasing library depth
The discovery rate of exon and intergenic bases (at 10X or greater depth) in the MIP101 library was calculated after each iteration of read sampling from 0% of library depth to 100% (sampled in blocks of 100,000 reads without replacement). The 'discovery rate' is the change in the percent of all possible exon and intergenic bases discovered at each iteration of sampling. For both data types, a power curve was fit to the data (dotted lines). For this library, 100% (*) of library depth corresponded to ~384 million single-end reads or ~167 paired reads. By extrapolating from the fitted power curves, the rates of discovery for exon and intergenic bases were projected to converge at 308% (**) of our library depth (i.e. ~1 billion single-end reads or ~500 million paired reads). This point of convergence (indicated with a black dot) is not the point at which discovery reaches 0, but rather the point at which the rate of discovery of exon bases reaches the level of discovery of intergenic bases (i.e. the point at which all new observations could be explained by intergenic noise alone). Refer to the **Methods** for details of curve fitting and R^2 calculations.

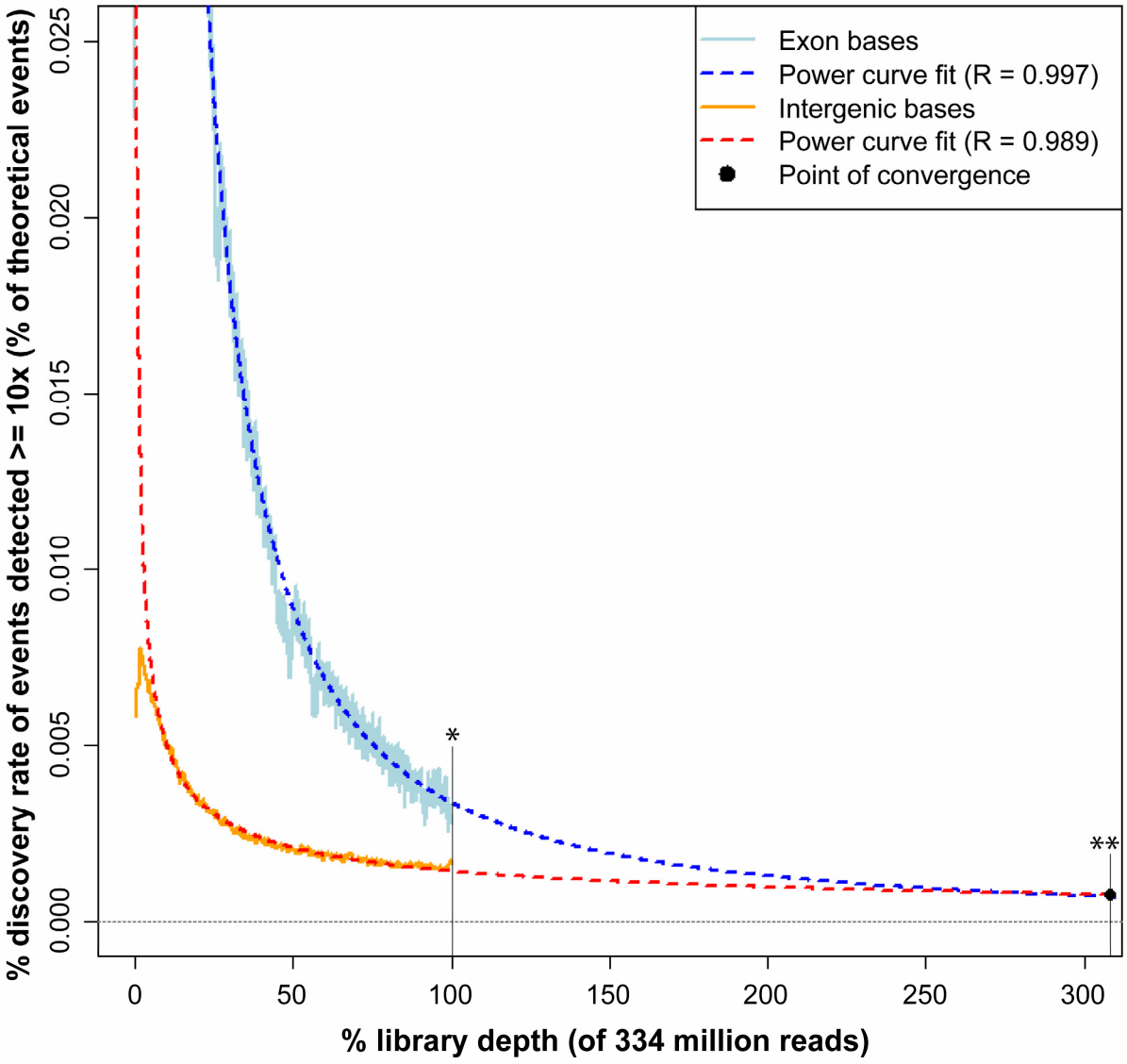


Figure 3.16. Coverage of exon base positions as a function of increasing library depth at varying minimum depth requirements
As in the previous figure, the percentage of all possible exon base positions covered with increasing library depth is reported. In this figure the percentage is shown at increasing expression level cutoffs. The number of bases observed at a minimum cutoff of 1X, 5X, 10X, 50X, 100X and 500X are plotted against the percent of library depth used in the analysis. As library depth increased the number of bases observed increased but the rate of increase was highly dependent on the level of depth required to consider a base observed (i.e. from 1X to 500X). While at 10X depth, the shape of the curve and the slope reported in the legend suggest that saturation is being achieved, a much greater library size would be required to achieve the same level of saturation at 50X or 100X depth. The slope was calculated for the tail of the distribution only (i.e. the data points corresponding to 50-100% library depth)

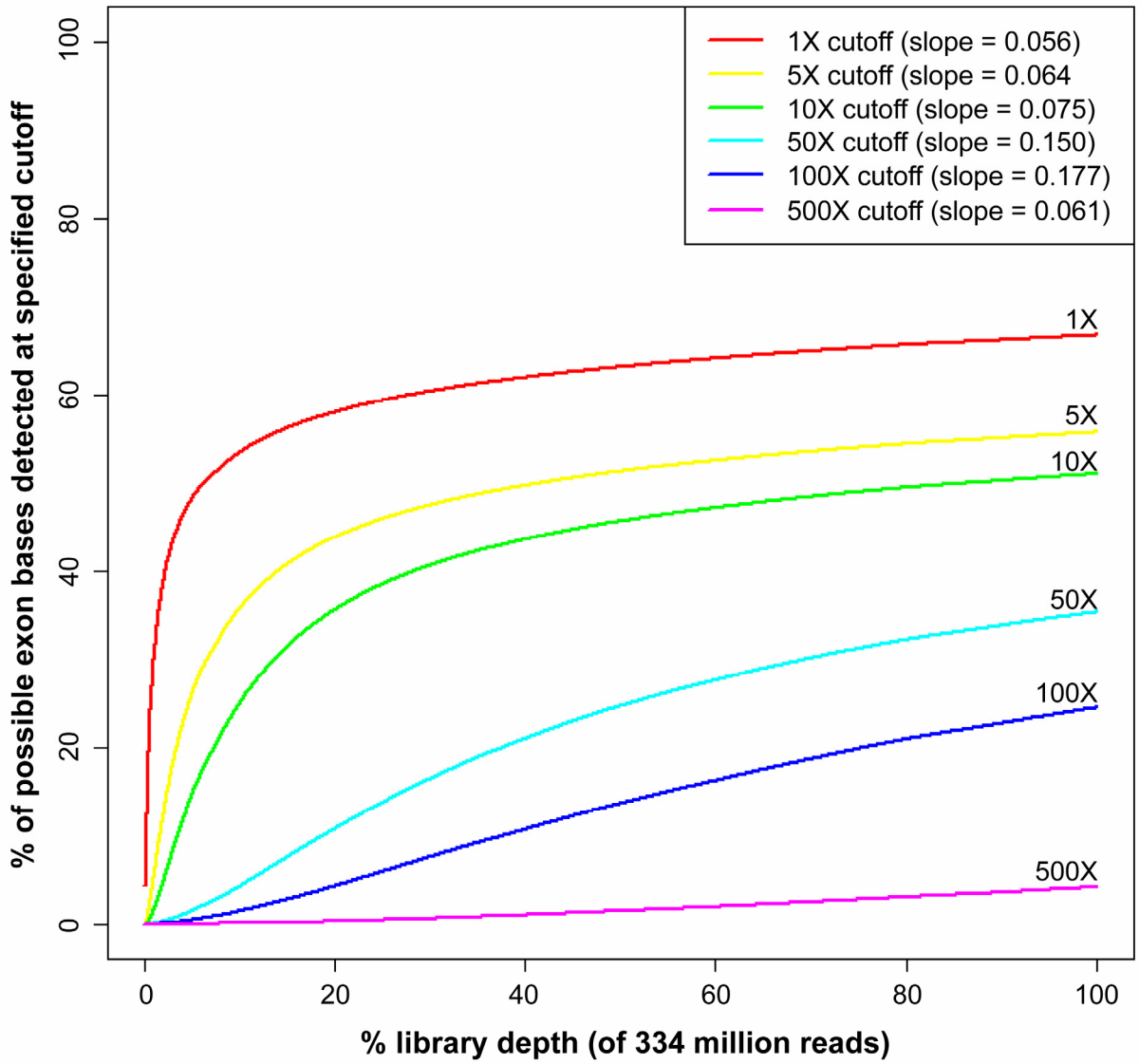


Figure 3.17. Relationship between gene and intron expression estimates
The expression value of every 'silent' intron region was plotted against the expression value for the gene to which they belong. Gene expression values were derived from the observed coverage of the exon base positions of the gene. The expression level of 'silent' intron regions (which generally should not be detected in mRNA expression profiling) was found to be correlated with gene expression level (Spearman correlation = 0.52). This suggests that un-processed RNA contamination (which would be expected to correlate with gene expression level) is present in addition to sources of intergenic sources of noise (such as genomic DNA contamination or random transcript noise). We fit a linear model to the 95th percentile of this correlation and used the coefficients of this fit to derive gene-by-gene estimates of background expression level. Features (exons, exon junctions, etc.) defined within each gene had to exceed this cutoff in order to be considered expressed (see **Methods**).

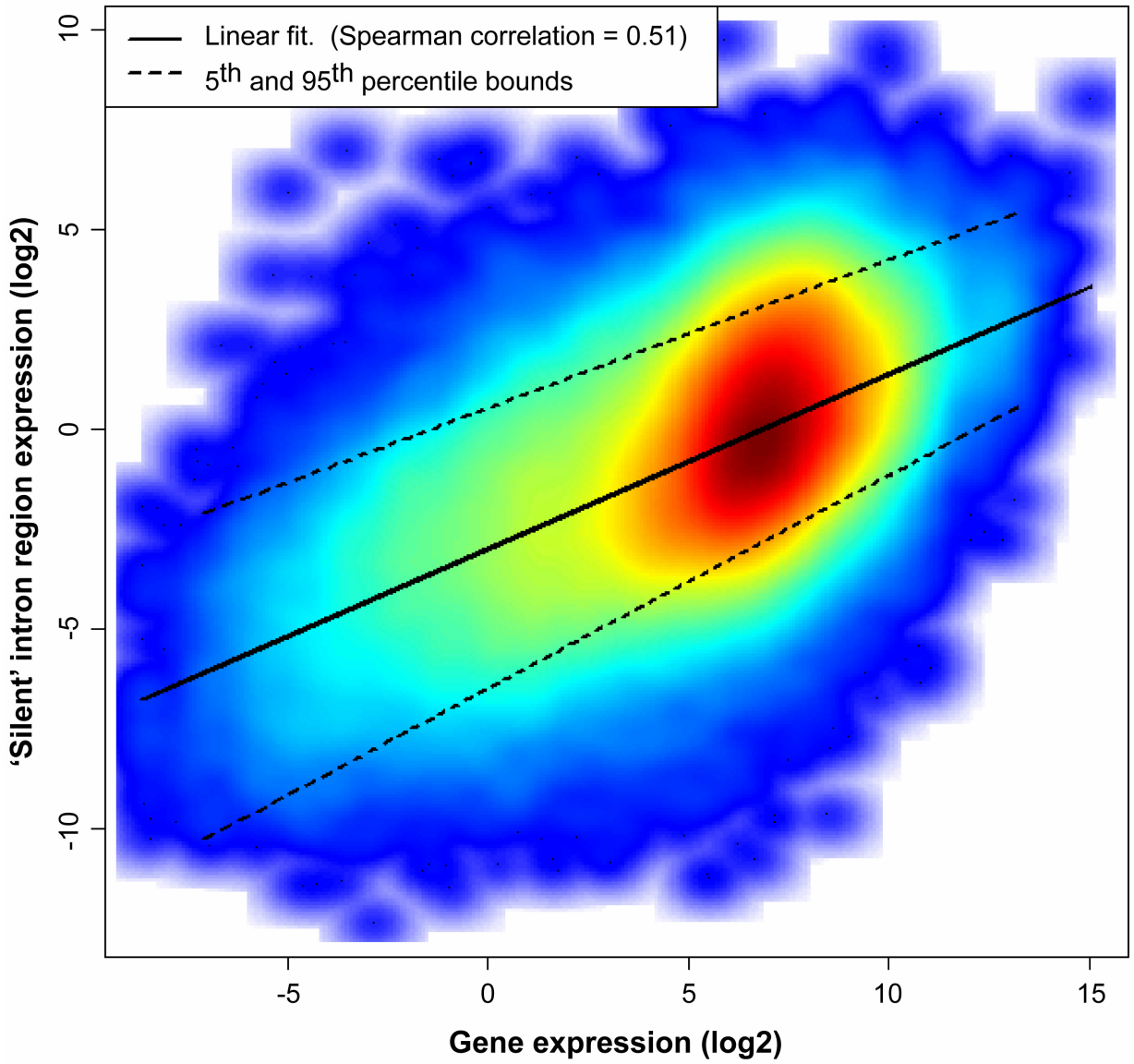


Figure 3.18. Expression distribution for all sequence feature types
The distribution of expression values expressed as average coverage values on a log2 scale are depicted as box plots for 15 feature types (**Methods**). Each feature type is color coded according to its feature type. Multiple box plots of the same color indicate sub-types. The 95th percentile of 'silent' intergenic regions is depicted as a dotted line. This value corresponds to the expression estimates for intergenic regions with no previous evidence for expression in EnsEMBL or sequence databases. We used this value as a conservative estimate of the level of intergenic noise. One consequence of this contamination is that with sufficient sequencing depth, known or hypothetical exons will eventually be observed in the sequence data whether they are expressed or not. However, since the expression level of exons has limited overlap with that of 'silent' intron and intergenic regions we can effectively filter out this noise. There are 14,088 single exon transcripts and genes, of which, 90% are pseudogenes and micro RNAs that were not detected as expressed above background (red and yellow box plots). These non-detected, single exon genes make up ~40% of all genes but only 5% of all exons resulting in the apparent discrepancy between the gene/transcript box plots compared to those for known exons and junctions.

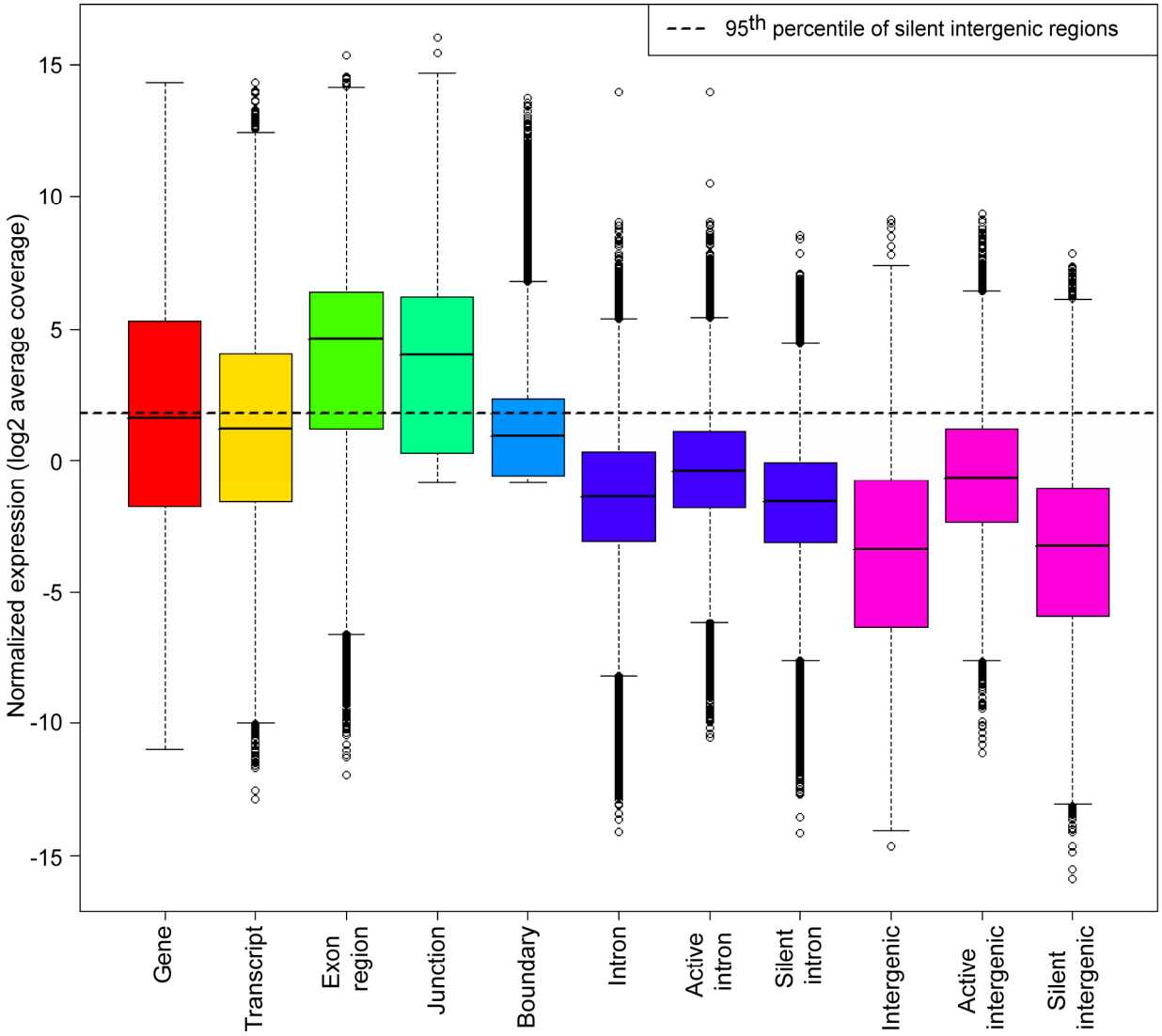


Figure 3.19. Expression of exon regions contrasted with intronic and intergenic regions

The percentiles of expression values for exon regions, silent intron regions and silent intergenic regions were plotted against the corresponding expression values. This shows that approximately 75% of exon expression values exceed the 95th percentile of silent intergenic regions (an expression value of 1.75 on a log2 scale).

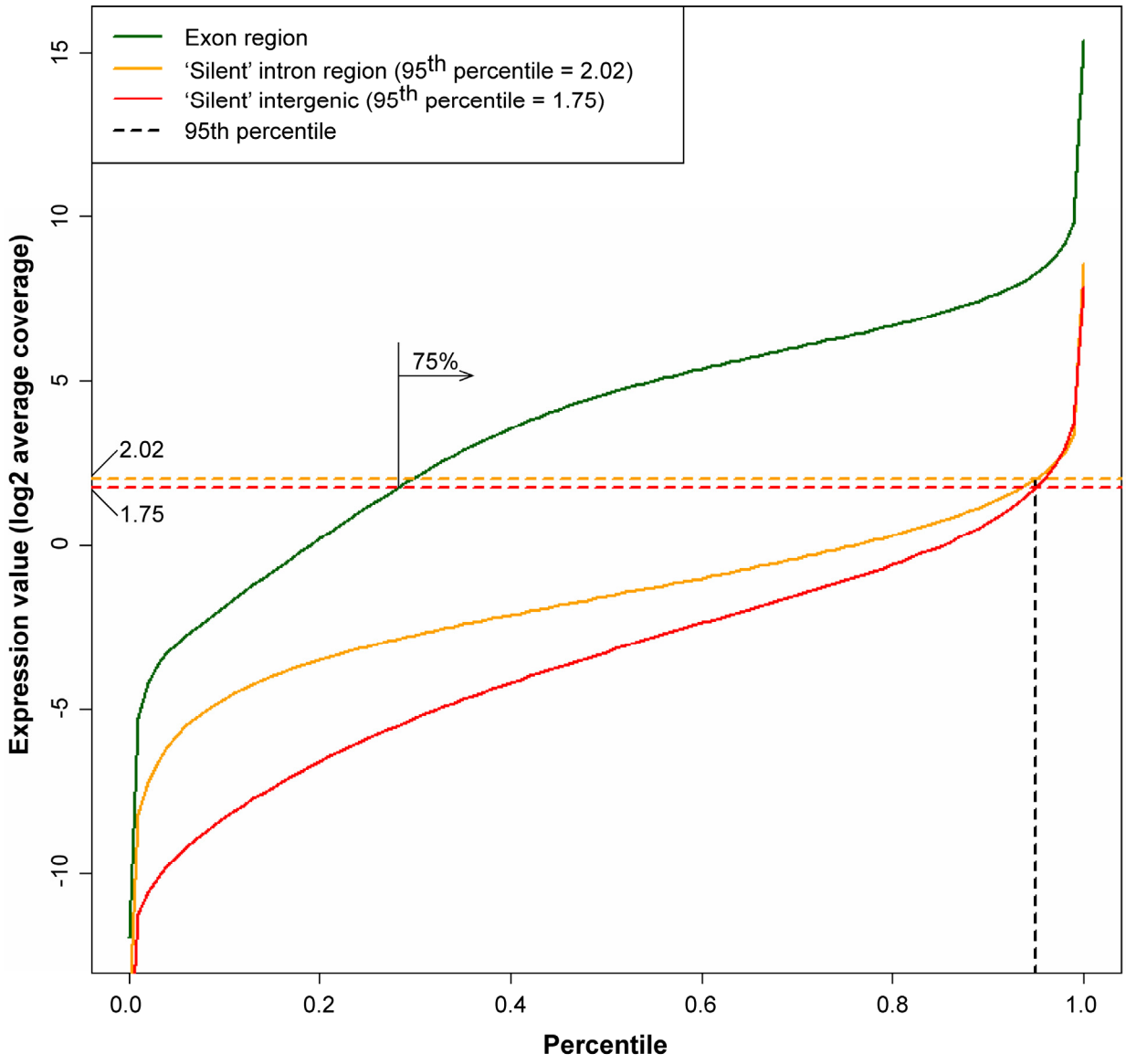


Figure 3.20. Example of a transcript, *H19* that is much less abundant in 5-FU resistant cells compared to sensitive cells
A screen shot taken from the ALEXA-Seq data viewer depicting the expression of all exon and exon junction features at the '*H19*' locus for both MIP101 and MIP/5FU cells. Expression is displayed on a log2 scale as colored lines (see **Methods** for details). Note the difference in expression levels between MIP101 and MIP/5FU cells. Expression values differ by as much as 7 on a log2 scale between the two libraries (i.e. up to 128-fold decrease in expression) (**Table 3.8**).

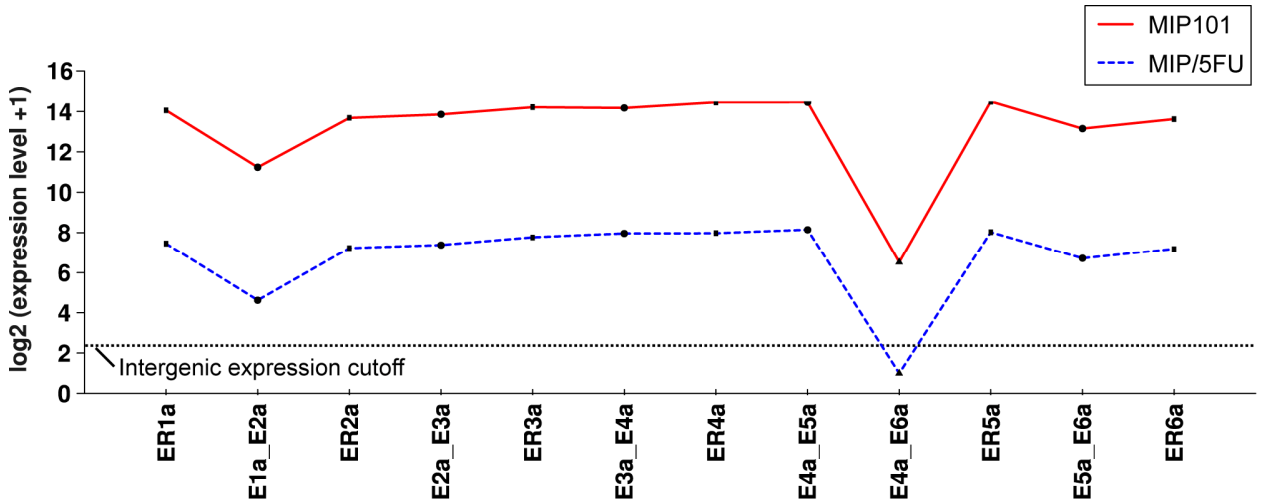


Figure 3.21. The gene *KRT20* is up-regulated in 5-FU resistant cells compared to sensitive cells
A screen shot taken from the ALEXA-Seq data viewer depicting the expression of all features of the locus, '*KRT20*', for both MIP101 and MIP/5FU cells. Refer to legend of **Figure 3.20** for a brief description of the display. In this gene, only the known exons and canonical junctions were detected as expressed in either condition. The expression level of these exons and exon junctions is highly consistent across the gene within each sample but the gene is over-expressed in resistant cells compared to sensitive cells by ~14-fold (**Table 3.8**).

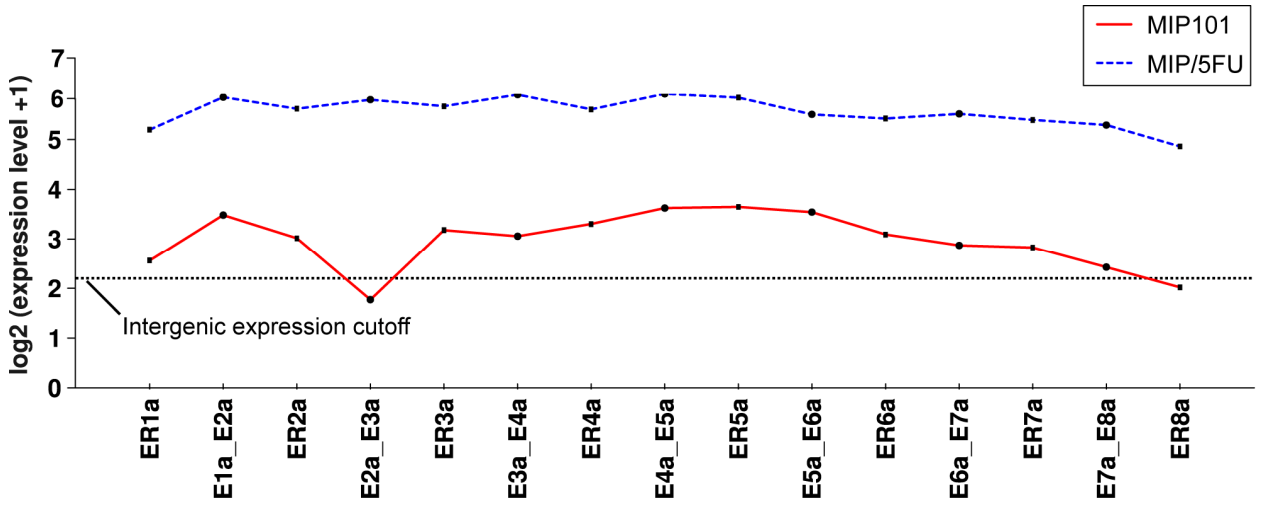


Figure 3.22. Percentage of exon-skipping junctions with a particular number of exons skipped for known and observed (i.e. expressed) exon junctions
The percentage of exon-skipping junctions relative to all exon-skipping junctions was calculated for 1 exon-skips, 2 exon-skips, etc. This calculation was performed for all known exon-skipping junctions (green line; n = 17,287), only those exon-skipping junctions actually detected as expressed (blue line; n = 6,280), only those exon-skipping junctions that were both expressed and novel (magenta line; n = 1,232), and a random sampling of junctions equal in size to the total number of known junctions (red line). In all three cases, the majority (60-70%) of exon-skipping events involved only a single exon being skipped. ~20% involved 2 exons skipped and the occurrence of greater numbers of exons skipped by a splicing event decreased steadily from there. Only 8% of the entire database of ~2.2 million junctions profiled in our analysis involved a single exon being skipped but the expressed junctions were highly biased towards a single exon skipped. The high agreement in this pattern between expressed junctions, novel expressed junctions and all known junctions supports the prediction that these are true novel splicing events and not random false positives. The random data line illustrates that a random selection of junctions does not hold to this pattern.

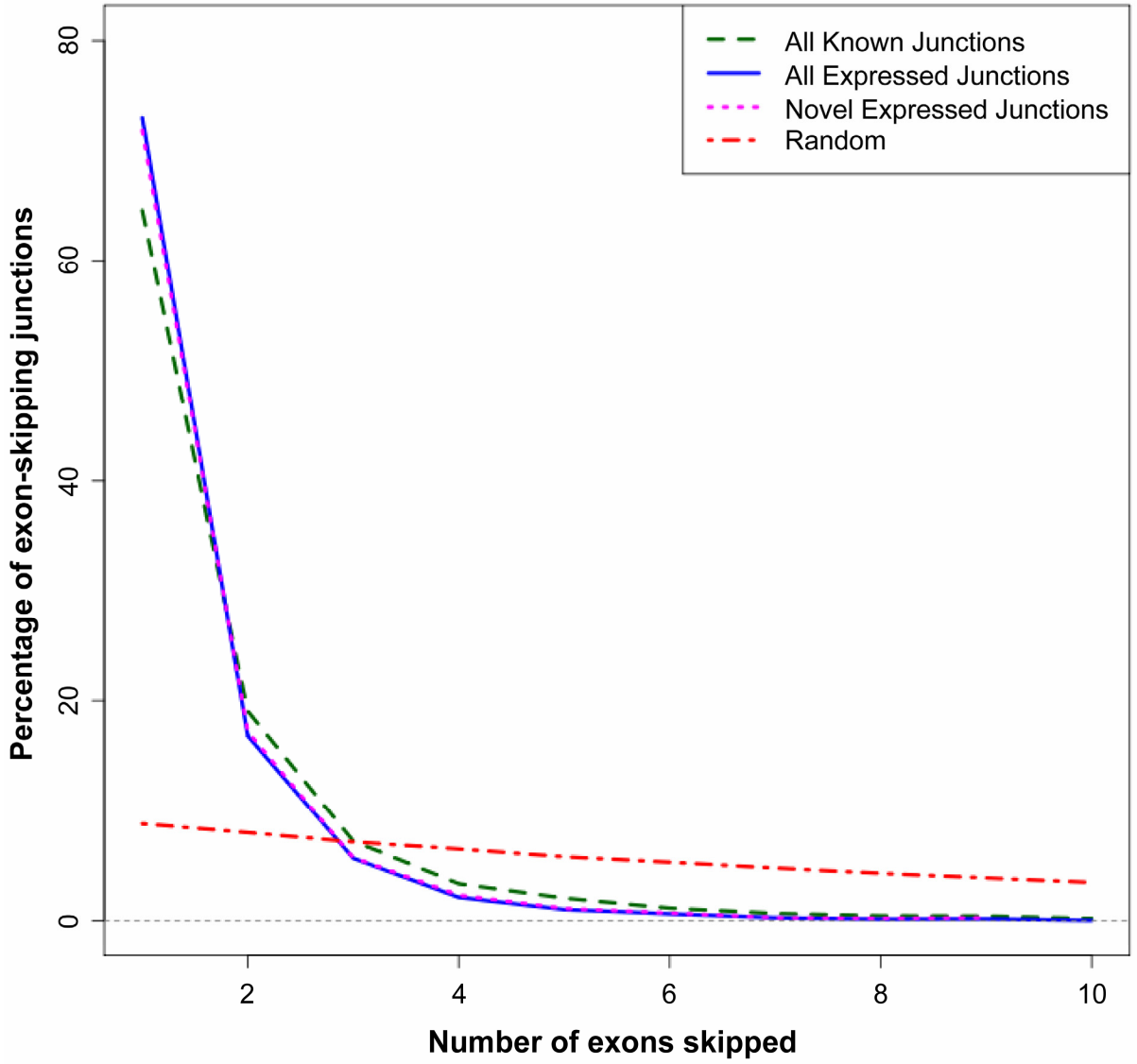


Figure 3.23. The *UMPS* gene exhibits reciprocal differential expression of two isoforms

A screen shot taken from the ALEXA-Seq data viewer depicting the expression of all features of *UMPS*, for both MIP101 and MIP/5FU cells. Refer to legend of **Figure 3.20** for a brief description of the display. Note that while this gene is ~2-fold down-regulated overall from 5-FU sensitive to resistant cells, a novel exon-skipping isoform (skipping exon 2) is up-regulated by ~15-fold (light blue bar labeled 'E1a-E3a' and marked with a star) (**Table 3.8**). The intragenic and intergenic expression cutoff levels are depicted as the upper and lower dashed lines respectively (see **Methods** for details).

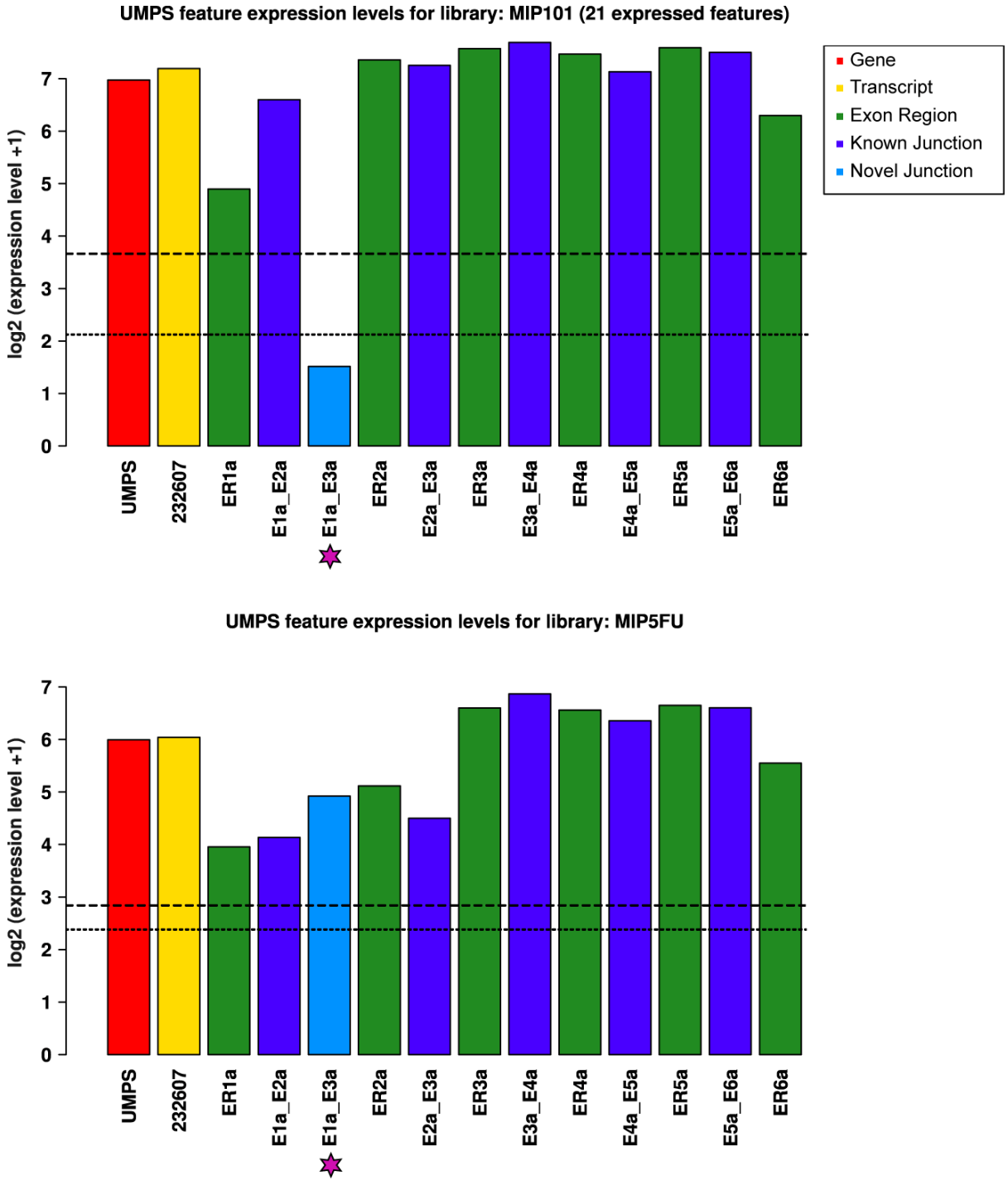


Figure 3.24. Example of gene locus, *OCIAD1* with over-expression of several novel exon-skipping isoforms

A screen shot taken from the ALEXA-Seq data viewer depicting the expression of all features of the locus, '*OCIAD1*', for both MIP101 and MIP/5FU cells. Refer to legend of **Figure 3.20** for a brief description of the display. Note that while this gene is ~2-fold down-regulated overall from 5-FU sensitive to resistant cells, splicing at the donor site of exon 4 of *OCIAD1* appears to be disrupted, resulting in the up-regulation of three novel exon-skipping isoforms (three light blue bars, each marked with a star). In the sensitive cells, exon 4 is only connected to its canonical partner exon 5. In the resistant cells however, exon 4 is connected to exons 5, 6, 8 and 9. The intragenic and intergenic expression cutoff levels are depicted as the upper and lower dashed lines respectively (see **Methods** for details).

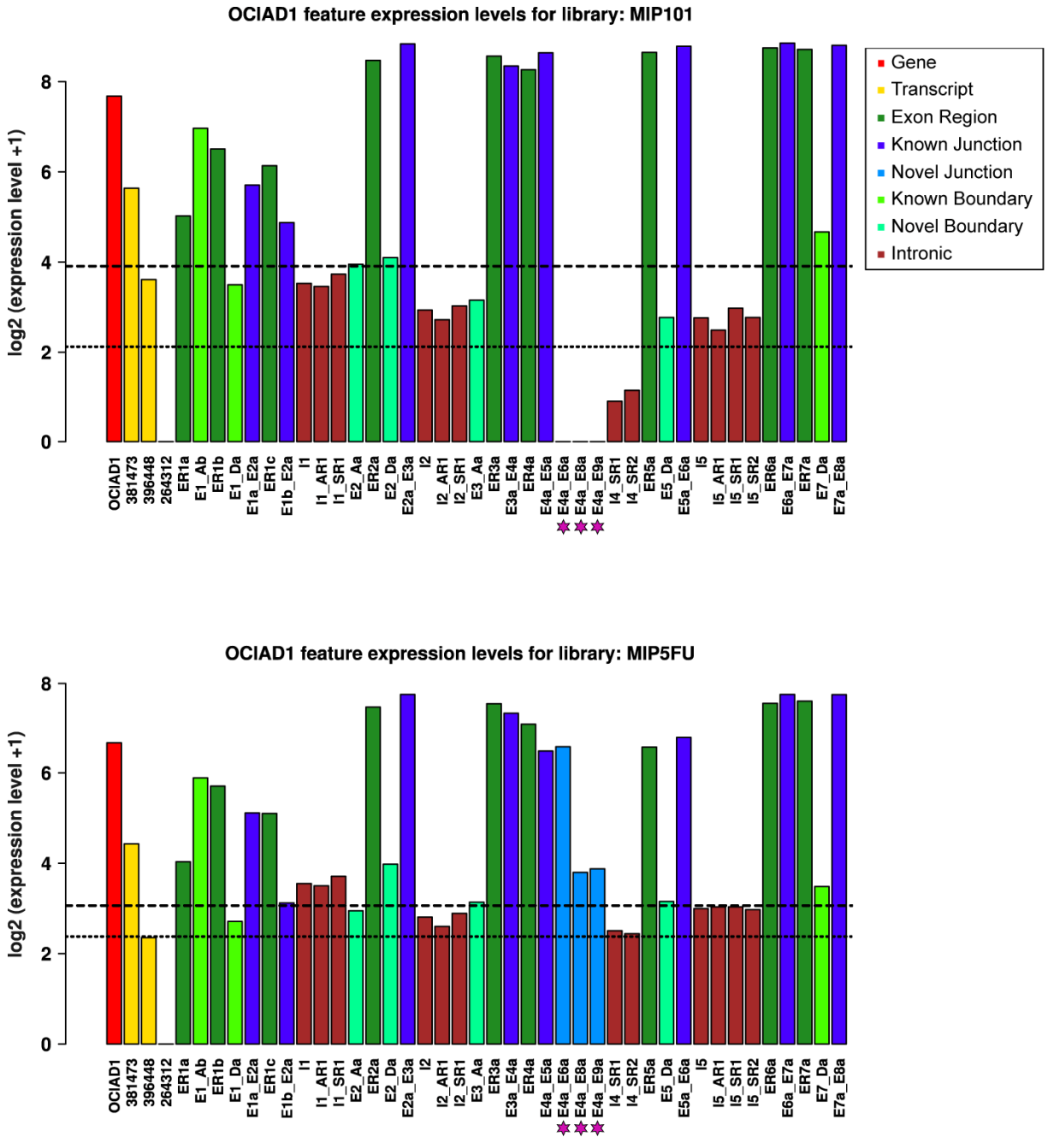


Figure 3.25. Proportion of expressed features observed in sensitive versus resistant cells

The proportion of all expressed features identified as expressed in MIP101 (red bars) and MIP/5FU (blue bars) cells was summarized for feature types corresponding to known isoforms (left side) and predicted novel isoforms (right side). The proportion is the number of expressed features observed in one cell line divided by the number of features expressed in either cell line. For features corresponding to known isoforms (known exons and junctions), the majority were observed in both cell lines. For example, the proportion of expressed known exons is close to 1 for both cell lines indicating that almost all known exons detected were expressed in both cell lines. However, for features corresponding to novel isoforms the proportion of such events was higher in MIP/5FU than MIP101. Specifically, the proportions of exon skipping events (novel exon junctions), alternative exon boundary usage, cryptic exons (active regions with introns), and retained introns (entire introns) were all higher in MIP/5FU than MIP101. The numbers for these events are provided in **Table 3.6**.

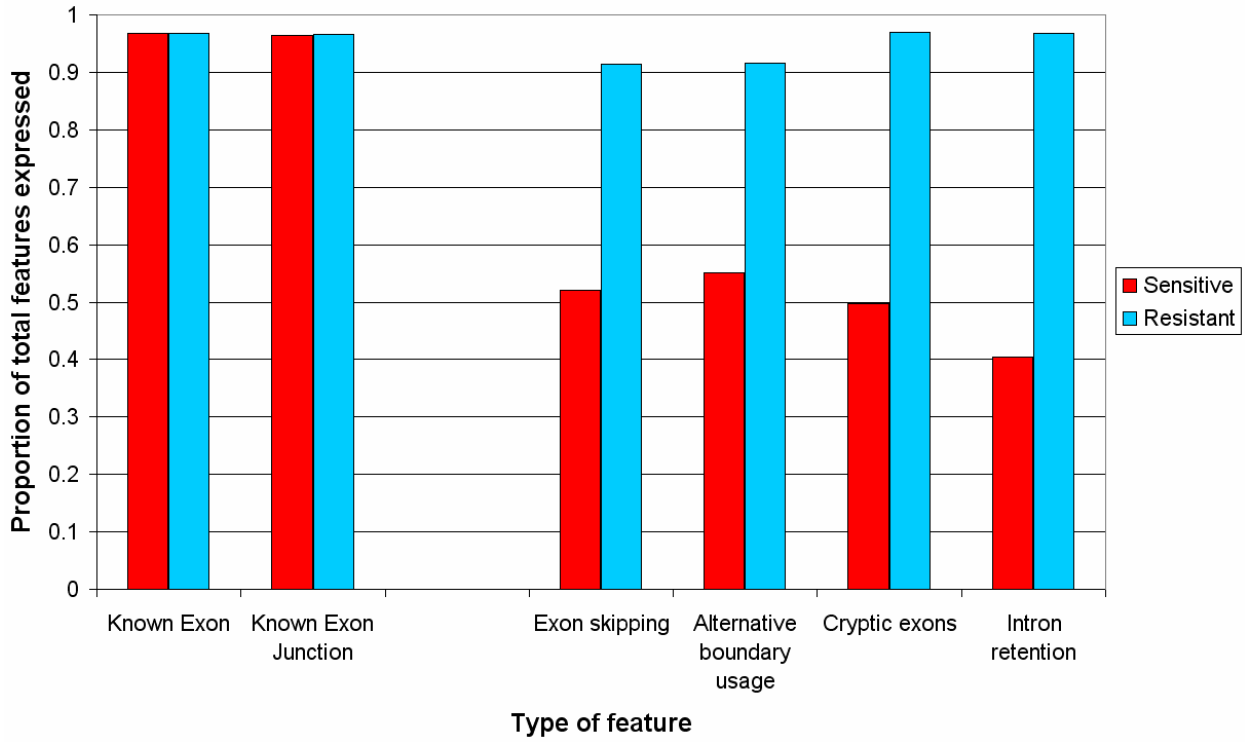


Table 3.1. Summary of alternative expression annotation databases for seven species

A brief summary of databases of alternative expression features we defined for seven species is provided below. Features were defined by analysis of EnsEMBL and mRNA/EST sequence databases (see **Figure 3.1** and **Methods** for details). For further details on the human database used for the data analysis described in the text, refer to **Table 3.2**. A feature was considered to be 'known' if it was represented by one or more EnsEMBL transcripts. All other features were predicted by our annotation pipeline. Abbreviations for features: Gene (G); Transcript (T), Exon region (ER); Exon junction (EJ); alternative exon boundary (EB); Silent intron region (SI); Active intron region (AI); Silent intergenic region (SIG); Active intergenic region (AIG).

Each database can be dowloaded from: www.AlexaPlatform.org

Species (Genome Build)	Features defined and basic statistics	Distribution of feature types
Drosophila melanogaster (EnsEMBL version 54; FlyBase release 5.4)	534,061 (23% known, 77% predicted, 20% EST/mRNA supported, 1% conserved)	2.8% G, 4.1% T, 13.0% ER, 43.3% EJ, 18.7% EB, 9.2% SI, 2.7% AI, 3.4% SIG, 2.7% AIG
Gallus gallus (EnsEMBL version 54; Chicken genome 2.1)	2,392,504 (14% known, 86% predicted, 9% EST/mRNA supported, 6% conserved)	0.8% G, 0.9% T, 7.6% ER, 65.1% EJ, 12.5% EB, 6.8% SI, 2.1% AI, 1.8% SIG, 2.5% AIG
Homo sapiens (EnsEMBL version 53; Human NCBI 36 assembly)	3,814,043 (14% known, 86% predicted, 16% EST/mRNA supported, 9% conserved)	1.0% G, 1.6% T, 7.3% ER, 58.0% EJ, 11.3% EB, 8.2% SI, 6.6% AI, 2.9% SIG, 3.2% AIG
Homo sapiens (EnsEMBL version 55; Human Genome Reference Consortium, Feb. 2009)	4,639,589 (15% known, 85% predicted, 16% EST/mRNA supported, 8% conserved)	1.0% G, 2.2% T, 8.0% ER, 60.5% EJ, 11.6% EB, 7.4% SI, 5.4% AI, 1.9% SIG, 2.0% AIG
Mus musculus (EnsEMBL version 54; Mouse NCBI m37)	3,313,018 (14% known, 86% predicted, 17% EST/mRNA supported, 7.6% conserved)	1.0% G, 1.5% T, 7.6% ER, 61.3% EJ, 12.0% EB, 7.6% SI, 4.5% AI, 2.2% SIG, 2.4% AIG
Pan troglodytes (EnsEMBL version 54; Chimpanzee 2.1)	3,141,903 (14% known, 86% predicted, 15% EST/mRNA supported, 8% conserved)	0.8% G, 1.3% T, 7.3% ER, 61.9% EJ, 11.6% EB, 7.8% SI, 4.3% AI, 2.4% SIG, 2.7% AIG
Rattus norvegicus (EnsEMBL version 54; Rat 3.4)	3,203,902 (14% known, 86% predicted, 10% EST/mRNA supported, 10% conserved)	0.9% G, 1.2% T, 7.7% ER, 64.9% EJ, 12.1% EB, 6.6% SI, 2.5% AI, 2.0% SIG, 2.3% AIG
Saccharomyces cerevisiae (EnsEMBL version 54; Yeast, Saccharomyces Genome Database)	29,361 (26% known, 74% predicted, 0% EST/mRNA supported, 0% conserved)	24.3% G, 24.3% T, 25.7% ER, 1.1% EJ, 2.4% EB, 1.0% SI, 0.0% AI, 21.3% SIG, 0.0% AIG

Table 3.2. Summary of read data, gene model sources and features defined for alternative expression analysis

Features were defined as 'known' if they corresponded to a transcript in the EnsEMBL database. A feature was considered to be supported by an mRNA/EST if the chromosome coordinates of the feature corresponded to an mRNA/EST alignment from UCSC (see **Methods** for details). Conservation was similarly assessed by examining alignments of non-human ESTs and mRNAs (also from UCSC). The total cumulative number of base positions for all of the features of each type are reported as well as the subset of these that were not masked according to EnsEMBL (see Methods for details) and the subset of bases corresponding to coding (translated) position of one or more EnsEMBL transcripts. Note that '% conserved' values were not assessed for genes and transcripts because mRNA and EST alignments from non-human species were generally too short to assess the structure of complete gene models and transcripts.

Sequence data	
Number of reads (paired 36- and 42-mers)	262 million (524 million single-end reads)
Bases of sequence	21.5 billion
Known exon bases covered at >= 1x depth	46.6 million (73.1% of transcriptome)
Bases of the genome covered at >= 1x depth	242.6 million (7.8% of genome)
Gene model sources	
EnsEMBL known transcripts	62,371 (EnsEMBL version 53)
Human mRNA sequence alignments	256,678 (UCSC hg18)
Non-human mRNA sequence alignments	817,042 (UCSC hg18)
Human EST sequence alignments	7,950,883 (UCSC hg18)
Non-human EST sequence alignments	45,745,393 (UCSC hg18)

Feature annotations imported or defined from the gene model sources	**Total**	**% mRNA/EST supported**	**% Conserved**	**Total bases (% unmasked, % coding)**
Gene	36,953	72.3%	-	70.7 Mb (90.2%, 54.7%)
Transcript	62,371	83.6%	-	46.3 Mb (88.4%, 50.7%)
Exon	273,464			
Exon regions	277,805	76.4%	63.5%	70.7 Mb (90.2%, 54.7%)
Exon-exon junctions				
Known	218,463	91.9%	39.4%	13.5 Mb (97.4%, 90.9%)
Novel	1,992,797	0.7%	1.2%	123.5 Mb (97.9%, 95.2%)
Alternative exon boundaries i.e. acceptor/donor sites				
Known	21,431	90.8%	53.0%	1.3 Mb (91.3%, 55.7%)
Novel	407,580	13.1%	8.9%	25.2 Mb (95.6%, 46.0%)
Introns				
'Silent' intron regions	312,429	1.1%	0.9%	1,009.7 Mb (52.3%, 0%)
'Active' intron regions	250,526	28.6%	3.5%	79.5 Mb (81.7%, 0%)
Intergenic regions				
'Silent' intergenic regions	107,912	1.2%	1.0%	1,809.6 Mb (41.0%, 0%)
'Active' intergenic regions	120,066	39.4%	5.6%	43.1Mb (80.5%, 0%)
Total features (non-redundant)	**3,808,333**			**3,013.4 Mb (47.8%, 1.2%)**

Table 3.3. Top 20 differentially expressed (DE) genes from three gene expression platforms

Red values indicate expression was lost in resistant cells compared to sensitive and blue values indicate a gain of expression in 5-FU resistant cell relative to sensitive. A log2 DE value of 1 corresponds to a fold change of 2.

Gene Name	Mean Affymetrix PLIER Log2 DE	Mean ALEXA Log2 DE	Illumina WTSS Log2 DE	Max FOLD CHANGE observed for any platform
C12orf59	-5.968	-7.868	-4.902	233.6
OLR1	-5.598	-6.962	-4.429	124.6
H19	-3.074	-2.376	-6.450	87.4
PDZK1	-2.863	-6.021	-1.548	64.9
FUT3	-5.887	-4.007	-4.290	59.2
ASRGL1	4.414	5.494	3.843	45.1
C12orf63	5.363	2.416	2.098	41.2
PRF1	-3.754	-5.314	-3.211	39.8
GIPC2	5.131	4.975	2.825	35.0
PON3	-3.509	-4.850	-2.773	28.8
ATOH8	2.999	4.635	3.269	24.9
COL4A1	-4.491	-3.676	-3.199	22.5
ACSL4	4.144	3.187	2.707	17.7
FBP1	-3.175	-4.013	-2.700	16.1
MYEOV	-0.767	-3.955	-1.262	15.5
GSPT2	-2.082	-3.890	-3.455	14.8
IGF2BP3	3.251	3.880	2.381	14.7
KRT20	2.581	3.859	2.602	14.5
ARSE	-3.836	-3.824	-3.339	14.3
SLAMF6	2.721	3.821	1.487	14.1

Table 3.4. Comparison of dynamic range, signal-to-noise, sensitivity and specificity for Affymetrix, NimbleGen and Illumina platforms based on an analysis of expression estimates for the exons and introns of 100 housekeeping genes
Since both the Affymetrix and NimbleGen platforms use 16-bit scanners to extract hybridization intensities during scanning of their arrays, the theoretical dynamic range of these platforms is 16 (on a log2 scale). In practice, due to non-specific hybridization and other sources of signal noise, the lower limit is not achieved, although with proper calibration of the scanner, the upper limit can be achieved. Since massively parallel sequencing approaches use random sampling of transcriptome space and produce a digital output (read counts), their dynamic range is limited only by the number of data points generated (i.e. library depth). This allows for improved dynamic range, signal-to-noise ratio, sensitivity and specificity, which we illustrated by examination of a set of 100 housekeeping genes routinely used on Affymetrix microarray designs and mimicked on our own custom NimbleGen array designs. Note that the dynamic range reported in this table differs from that displayed in **Figure 3.11** because raw data were used in this calculation. Instead of adding 1 before converting to log2 scale, features with an expression value of 0 were simply removed from the Illumina data. Zero values do not occur in the microarray data. Also note that the sensitivity and specificity reported in this table correspond to the point at which the sum of sensitivity and specificity was maximized (see **Figure 3.12**).

	Affymetrix exon arrays	**NimbleGen ALEXA arrays**	**Illumina WTSS**
Theoretical dynamic range (log2 scale)	16	16	Unlimited
Observed dynamic range (log2 scale)	8.5	9.2	30.1
Signal-to-Noise Ratio	20.8 ± 0.42	56.5 ± 2.5	381.1 ± 44.7
Specificity	86.5%	95.8%	99.0%
Sensitivity	83.5%	86.9%	92.6%

Table 3.5. Comparison of *UMPS* A/B isoform expression ratios from four different platforms capable of measuring alternative isoforms

Expression values for two isoforms of the gene *UMPS* were derived by four mRNA expression platforms and used to estimate the ratio of *UMPS* isoform A compared to B in both MIP101 and MIP/5FU cells. The Illumina platform estimates the greatest difference between isoform A and B in sensitive cells, while all four platforms indicate a ratio of the two isoforms of nearly 1:1 or lower in MIP/5FU cells (ratio from 0.24 – 1.03).

Platform	UMPS A/B isoform ratio (MIP101 cells)	UMPS A/B isoform ratio (MIP/5FU cells)
ALEXA/NimbleGen splicing microarray	25.62 ± 0.56	0.85 ± 0.13
Semi-quantitative RT-PCR	27.36 ± 9.94	0.24 ± 0.01
Real-time quantitative RT-PCR	22.97 ± 2.64	1.03 ± 1.56
Illumina WTSS	51.4	0.80

Table 3.6. Summary of feature expression, differential expression, and alternative expression

The total number of features of each type in our human sequence annotation database is listed below. This is followed by the number of features expressed above background in each sample (see **Methods**) and the number that were significantly differentially expressed (DE) between the two libraries. Note that the MIP101 library consisted of 167 million paired reads and MIP/5FU consisted of 95 million paired reads. Text is blue if the number of features of a particular type was higher in sensitive than resistant cells and red for the opposite comparison. The DE (differential expression) column represents the number of features of each type that were significantly DE between MIP101 and MIP/5FU (fold-change > 2 and p-value < 0.05 after correction for multiple testing). The 'AE' (alternative expression) column represents the subset of those events in the DE column that are also candidate alternative expression events (see **Methods** for details). Note that for transcripts, only those corresponding to single transcript genes or transcripts that can be measured by unique exon junctions or regions are included (87% of all known transcripts).

Feature Type	# Features	# Expressed (MIP101)	# Expressed (MIP/5FU)	# DE	# AE
Gene	36,953	12,396 (33.54%)	12,004 (32.48%)	259	n/a
Transcript	54,515	12,857 (23.61%)	12,719 (23.33%)	251	15
Exon Region	277,804	144,737 (52.10%)	142,657 (51.35%)	4,040	218
Exon Junction	2,211,260	114,994 (5.20%)	114,613 (5.18%)	2,287	149
Known	218,463	112,748 (51.61%)	111,344 (50.97%)	2,242	114
Novel	1,992,797	2,246 (0.11%)	3,269 (0.16%)	30	21
Alternative Boundary	429,011	34,439 (8.03%)	46,205 (10.77%)	269	63
Known	21,431	9,505 (44.35%)	9,611 (44.84%)	155	26
Novel	407,580	24,934 (6.12%)	36,594 (8.98%)	112	33
Intron	204,177	7,559 (3.70%)	12,178 (5.96%)	47	15
'Silent' region	312,429	5,358 (1.71%)	10,007 (3.20%)	22	11
'Active' region	250,526	12,343 (4.93%)	19,357 (7.73%)	88	28
Intergenic	27,647	1,272 (4.60%)	1,151 (4.16%)	11	n/a
'Silent' region	107,912	2,860 (2.65%)	2,639 (2.45%)	37	n/a
'Active' region	120,066	8,203 (6.83%)	7,617 (6.34%)	116	n/a

Table 3.7. Summary of novel expressed exon-exon junctions and alternative exon boundaries

We summarized the non-redundant, novel (i.e. do not correspond to an EnsEMBL transcript) expressed exon junctions or boundaries observed in either the MIP101 or MIP/5FU library. Expressed exon junctions and alternative boundaries were examined with respect to mRNA/EST support, cross-species conservation, and protein coding affect. Protein coding affect refers to whether the use of each novel exon junction or alternative boundary is predicted to alter the known ORF. Possible over- or under-representation within the novel, expressed features relative to all features was assessed for mRNA/EST support, conservation and protein coding affect (by Fisher's exact test). For example, we found that the percentage of mRNA/EST supported sequences was higher than expected by chance in the expressed, novel exon-exon junctions relative to the total number of exon-exon junctions with EST/mRNA support.

Feature Type	**Count**	**% mRNA/EST Supported**	**% Conserved**	**% Protein Coding**
Exon-exon junctions				
Novel expressed	3,802	57.6%	20.2%	93.7%
All	2,211,260	9.7%	5.0%	97.0%
p-value		$< 1.0\times10^{-300}$	8.7×10^{-159}	5.3×10^{-24}
Alternative Boundary				
Novel expressed	41,076	43.3%	22.2%	87.5%
All	429,011	17.0%	11.1%	87.2%
p-value		$< 1.0\times10^{-300}$	4.2×10^{-282}	0.19

Table 3.8. Top 50 differential or alternative expression events
The following table summarizes the top differential and alternative gene expression events observed between 5-FU sensitive and resistant cells. 'Down-regulated' means that expression of a feature is reduced in resistant cells relative to sensitive (indicated as –ve values in red). 'Up-regulated' means a gain of expression was observed in resistant cells relative to sensitive (indicated as +ve values in blue). In the case of multiple mutually exclusive isoforms being differentially expressed, the largest magnitude fold-change is displayed. Visualization of the expression pattern for these genes (and all other genes) can be accessed with the ALEXA-Seq data viewer at our website (www.AlexaPlatform.org). Gene names marked with an '*' were chosen to illustrate visualizations from this viewer (see **Figures 20, 21, 23 & 24**). Abbreviations. Differential expression (DE). Splicing index (SI). Transcription start site (TSS).

Rank	*Gene Name*	Event Type	Fold Change (by DE or SI)	Description of event
1	*H19**	Gene DE	-113.2	Expression of an intron 1 retaining isoform of *H19* is down-regulated in 5-FU resistant cells. Isoforms exhibiting both retention and splicing of intron 1 are observed.
2	*OCIAD1**	Exon skipping	**103.6**	Normal splicing is disrupted at the donor site of exon 4 resulting in an up-regulation of novel isoforms containing E4-E8, E4-E6 and E4-E9. Overall gene expression is 2-fold down-regulated.
3	*C12orf59*	Gene DE	-78.3	Expression of entire locus is essentially lost, suggesting a possible deletion.
4	*EIF4A2*	Exon skipping	-61.4	Expression of putative novel exon 10 skipping isoform (i.e. E9-E11) is down-regulated.
5	*FUT3*	Gene DE	-44.4	Expression of entire locus is essentially lost, suggesting a possible deletion.
6	*OLR1*	Gene DE	-43.6	Expression of entire locus is essentially lost, suggesting a possible deletion.
7	*UBE2M*	Exon skipping	**39.4**	Expression of a putative novel exon 5 skipping isoform (i.e. E4-E6) is up-regulated. Overall gene expression is unchanged.
8	*C1orf2*	Exon skipping	-35.3	Five known transcripts at the locus can be measured. In sensitive cells 3 are expressed. In resistant cells the isoform containing E7-E8 is lost but the other two remain.

Rank	*Gene Name*	Event Type	Fold Change (by DE or SI)	Description of event
9	*BUD31*	Exon skipping	**33.1**	This gene has two known transcripts with mutually exclusive exon 1 (E1a and E1b). Up-regulation of two putative novel isoforms is observed, both involving skipping of exon 2 (i.e. E1a-E3 and E1b-E3). The canonical junctions are still observed. Could be caused by a heterozygous mutation at the exon 2 acceptor site.
10	*UBE2K*	Exon skipping	-30.9	Expression of a known exon 3 skipping isoform (i.e. E2-E4) is down-regulated
11	*AP2B1*	Exon skipping	**30.7**	Expression of a putative novel exon 21 skipping isoform (i.e. E20-E22) is up-regulated. Overall gene expression is unchanged.
12	*ASRGL1*	Gene DE	**30.6**	Gene is not expressed above background in sensitive cells but is significantly up-regulated in resistant cells
13	*EPS8L3*	Gene DE	-27.6	Expression of entire locus is down-regulated
14	*ALPP*	Gene DE	-24.9	Expression of entire locus is down-regulated
15	*FAU*	Exon skipping	-24.5	Expression of a novel (but mRNA supported) exon 5 skipping isoform (i.e. E4-E6) is down-regulated. Overall gene expression is unchanged.
16	*FOLR1*	Gene DE	-23.8	Expression of entire locus (4 known isoforms) is down-regulated. In sensitive cells, 3 of 4 isoforms are detected. In resistant only a small amount of 1 isoform is expressed.
17	*ZNF702*	Gene DE	-22.3	Expression of entire locus is down-regulated
18	*LAPTM4B*	Gene DE	-21.6	Expression of entire locus is down-regulated
19	*C15orf48*	Gene DE	-20.8	Expression of entire locus is down-regulated. Only 1 of 2 isoforms is expressed in sensitive cells and neither are expressed in resistant
20	*SLC7A7*	Gene DE	-18.8	Expression of entire locus is down-regulated
21	*RAB22A*	Exon skipping	-18.0	Expression of a novel (but mRNA supported) exon 4 skipping isoform (i.e. E3-E5) is down-regulated. Overall gene expression is unchanged.

Rank	*Gene Name*	Event Type	Fold Change (by DE or SI)	Description of event
22	*SPTLC2L*	Gene DE	-17.9	Expression of entire locus is down-regulated
23	*FNDC3A*	Intron retention	-16.0	Expression of the gene is up-regulated overall except for intron 15 which is down-regulated
24	*HHIP*	Alternative TSS	-15.9	In sensitive cells a novel transcript starting at exon 4 is expressed. In resistant cells this isoform is down-regulated.
25	*UMPS**	Exon skipping	**15.4**	A novel exon 2 skipping isoform is up-regulated while the canonical isoform is down-regulated
26	*CST1*	Gene DE	-15.0	Expression of entire locus is down-regulated. Only 1 of 2 known transcripts is expressed
27	*AIM1*	Exon skipping	**14.2**	Normal splicing is disrupted at the donor site of exon 17 resulting in an up-regulation of novel isoforms containing E17-E19, and E17-E21. Overall gene expression is 3-fold down-regulated.
28	*MR1*	Gene DE	-14.1	Expression of entire locus is down-regulated.
29	*KRT20**	Gene DE	**14.0**	Expression of entire locus is up-regulated.
30	*KTN1*	Intron retention	**13.8**	Expression of the gene is 2-fold down-regulated overall except for intron 27 which is up-regulated
31	*PYGL*	Gene DE	-12.9	Expression of entire locus is down-regulated.
32	*ACSL4*	Alternative TSS	**12.9**	Expression of a novel transcript starting at exon 4 is up-regulated
33	*TTC7A*	Intron retention	**12.8**	Expression of the gene is 2-fold down-regulated but retained introns 19 and 24 are up-regulated
34	*PDCD6IP*	Alternative exon boundary	**12.7**	Expression of the gene is 3.5-fold down-regulated but a variant using a novel exon 14 donor site is up-regulated
35	*TNNI2*	Gene DE	-12.7	Expression of entire locus is down-regulated. Retention of introns 3, 4 and 5 (of 6 introns total) is observed.
36	*HLTF*	Gene DE	-12.3	Expression of entire locus is down-regulated.
37	*PRF1*	Gene DE	-12.2	Expression of entire locus is down-regulated.

Rank	*Gene Name*	Event Type	Fold Change (by DE or SI)	Description of event
38	*GSPT2*	Gene DE	-12.2	Expression of entire locus is down-regulated.
39	*ATOH8*	Gene DE	**11.8**	Expression of entire locus is up-regulated.
40	*INTS9*	Exon skipping	**11.6**	Expression of a novel (but EST supported) exon 9 skipping isoform (i.e. E8-E10) is up-regulated. Overall gene expression is unchanged.
41	*KLK6*	Gene DE	-11.1	Expression of entire locus is down-regulated. Expression of novel acceptor sites at exon 2 is also observed.
42	*C4orf27*	Unknown	-10.6	Expression of the gene is 2-fold down-regulated overall while a canonical exon-junction (E1-E2) is 10.6-fold down-regulated.
43	*DNTTIP1*	Exon-skipping	-10.5	Expression of a novel exon 10 skipping isoform (i.e. E9-E11) is down-regulated. Overall gene expression is unchanged.
44	*ZNF185*	Gene DE	**10.4**	Expression of entire locus is up-regulated.
45	*LAMA3*	Alternative TSS	-10.3	Expression of the full-length isoform is down-regulated while expression of a shorter isoform starting at exon region 40 (of 77) is up-regulated. Overall gene expression is unchanged.
46	*CCBL1*	Alternative exon boundary	-10.3	Expression of the gene is 2-fold up-regulated but a variant using a novel exon 6 acceptor site is down-regulated
47	*UGT1A8*	Alternative exon boundary	-10.2	Expression of the gene is 1.4-fold down-regulated but a variant using a novel exon 7 acceptor site is 10.2-fold down-regulated
48	*PHLDB2*	Gene DE	-10.1	Expression of entire locus is down-regulated.
49	*TSPAN12*	Gene DE	-9.9	Expression of entire locus is down-regulated.
50	*TBC1D8B*	Gene DE	-9.8	Expression of entire locus is down-regulated.

Table 3.9. Statistically enriched functional categories identified by pathway analysis

Significant functions and diseases identified by Ingenuity Pathway Analysis are listed below along with the p-value and number of genes found in the specified functional category. P-values were calculated by Fisher's test and corrected by Benjamini and Hochberg. Complete gene lists and sub-categories for each of these functional categories are available at our website (www.AlexaPlatform.org)

Functional Category	P-value	Gene count
Cancer	2.37×10^{-3}	95
Embryonic development	4.11×10^{-2}	26
Tissue development	4.11×10^{-2}	17
Cellular growth and proliferation	4.11×10^{-2}	59
Dermatological diseases and conditions	4.11×10^{-2}	15

References

1. Griffith, M. & Marra, M. A. in Genes, Genomes & Genomics (eds. Thangadurai, D., Tang, W. & Pullaiah, T.) 201-242 (Regency Publications, New Delhi, 2007).
2. Griffith, M. et al. ALEXA: a microarray design platform for alternative expression analysis. Nat Methods 5, 118 (2008).
3. Johnson, J. M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302, 2141-4 (2003).
4. Pan, Q. et al. Revealing global regulatory features of Mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16, 929-41 (2004).
5. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-8 (2008).
6. Morin, R. et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45, 81-94 (2008).
7. Yassour, M. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A 106, 3264-9 (2009).
8. Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25, 1026-32 (2009).
9. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377-82 (2009).
10. Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5, 613-9 (2008).
11. Li, H. et al. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A 105, 20179-84 (2008).
12. Sultan, M. et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956-60 (2008).
13. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413-5 (2008).
14. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-6 (2008).
15. Meisner, N. C. et al. The chemical hunt for the identification of drugable targets. Curr Opin Chem Biol 8, 424-31 (2004).
16. Rich, T. A., Shepard, R. C. & Mosley, S. T. Four decades of continuing innovation with fluorouracil: current and future approaches to fluorouracil chemoradiation therapy. J Clin Oncol 22, 2214-32 (2004).
17. Tai, I. T., Dai, M., Owen, D. A. & Chen, L. B. Genome-wide expression analysis of therapy-resistant tumors reveals SPARC as a novel target for cancer therapy. J Clin Invest 115, 1492-502 (2005).
18. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
19. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-9 (2008).
20. Simpson, J. T. et al. ABySS: A parallel assembler for short read sequence data. Genome Res 19, 1117-23 (2009).
21. Fejes, A. P. et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24, 1729-30 (2008).
22. Birol, I. et al. De novo Transcriptome Assembly with ABySS. Bioinformatics (2009).

23. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-11 (2009).
24. Hubbard, T. J. et al. Ensembl 2009. Nucleic Acids Res 37, D690-7 (2009).
25. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14, 103-5 (2007).
26. Stephenson, F. H. Calculations in molecular biology and biotechnology : a guide to mathematics in the laboratory (Academic Press, Amsterdam ; Boston, 2003).
27. Sul, J. Y. et al. Transcriptome transfer produces a predictable cellular phenotype. Proc Natl Acad Sci U S A 106, 7624-9 (2009).
28. Setlow, J. K. Genetic Engineering: Principles and Methods (ed. Setlow, J. K.) (Plenum Publishers, New York, 2001).
29. Melamud, E. & Moult, J. Stochastic noise in splicing machinery. Nucleic Acids Res (2009).
30. Melamud, E. & Moult, J. Structural implication of splicing stochastics. Nucleic Acids Res (2009).
31. Sparago, A. et al. Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. Nat Genet 36, 958-60 (2004).
32. Rosenberg, R. et al. Prognostic significance of cytokeratin-20 reverse transcriptase polymerase chain reaction in lymph nodes of node-negative colorectal cancer patients. J Clin Oncol 20, 1049-55 (2002).
33. Niles, R. M. et al. Isolation and characterization of an undifferentiated human colon carcinoma cell line (MIP-101). Cancer Invest 5, 545-52 (1987).
34. Sakamoto, E. et al. Orotate phosphoribosyltransferase expression level in tumors is a potential determinant of the efficacy of 5-fluorouracil. Biochem Biophys Res Commun 363, 216-22 (2007).
35. Chapuy, B. et al. Intracellular ABC transporter A3 confers multidrug resistance in leukemia cells by lysosomal drug sequestration. Leukemia 22, 1576-86 (2008).
36. Steinbach, D. et al. ABCA3 as a possible cause of drug resistance in childhood acute myeloid leukemia. Clin Cancer Res 12, 4357-63 (2006).
37. Klein, T. E. et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1, 167-70 (2001).
38. Venables, J. P. et al. Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol (2009).
39. Ramdas, L. et al. Improving signal intensities for genes with low-expression on oligonucleotide microarrays. BMC Genomics 5, 35 (2004).
40. Klinck, R. et al. Multiple alternative splicing markers for ovarian cancer. Cancer Res 68, 657-63 (2008).
41. Venables, J. P. et al. Identification of alternative splicing markers for breast cancer. Cancer Res 68, 9525-31 (2008).
42. Hancock, J. M. & Armstrong, J. S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput Appl Biosci 10, 67-70 (1994).
43. Kuhn, R. M. et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37, D755-61 (2009).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403-10 (1990).

45. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-8 (2008).
46. Rumble, S. M. et al. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5, e1000386 (2009).
47. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713-4 (2008).
48. Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics (2009).
49. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462-7 (2005).
50. Benjamini, Y. & Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Behav. Educ. Statist. 25, 60-83 (2000).
51. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5, R80 (2004).
52. Clark, T. A. et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol 8, R64 (2007).
53. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9, 488 (2008).

4. Genomic analysis of uridine monophosphate synthetase reveals novel mRNA isoforms and mutations associated with fluorouracil resistance in colorectal cancer[4]

4.1. Introduction

The drug fluorouracil (5-FU) is an anti-metabolite chemotherapy commonly used in the treatment of several cancer types including head and neck, pancreatic, breast, stomach and colorectal[1, 2]. Analogs, pro-drugs and oral versions of 5-FU such as Capecitabine and Tegafur are also widely used (**Appendix A**). Intrinsic or acquired resistance to 5-FU is a major limiting factor in the treatment of colorectal and other cancers. Response rates for 5-FU vary from 6% to 53% depending on dose, schedule, combination with modifiers and disease characteristics[3]. While 5-FU is now relatively inexpensive, new versions with improved response rates can cost thousands of dollars per month of therapy. For example, in 2008, the BC Cancer Agency regional cancer centers treated 1,871 cancer patients (primarily gastrointestinal and breast) using chemotherapy protocols involving 5-FU and/or Capecitabine at a total cost of ~1.33 million dollars. Due to the high usage and low-to-moderate response rates for 5-FU, the cost associated with 5-FU resistance is considerable. Developing methods to predict and ultimately overcome this resistance is therefore an important area of research.

5-FU is a pyrimidine (uracil) analog developed in the 1950's that is preferentially utilized by actively dividing tumor cells[4, 5]. Although the precise mechanism of action is still a subject of debate in the literature, 5-FU is thought to act by at least three mechanisms, each requiring metabolic activation of 5-FU (**Figure 4.1**). The most widely cited mechanism involves conversion of 5-FU to 5-fluoro-2'-deoxyuridine 5'-monophosphate (FdUMP) which inhibits thymidylate synthetase (*TYMS* aka *TS*) leading to depletion of thymine, inhibition of DNA synthesis and DNA damage when uracil is subsequently incorporated into DNA[6]. Alternative mechanisms involve conversion of 5-FU into the fluorine-containing cytotoxic metabolites FUTP and FdUTP which are incorporated into RNA and DNA respectively[7]. Varying response to 5-FU is thought to be mediated primarily by differences in the metabolic pathways of 5-FU activation and

[4] A version of this chapter has been submitted. Griffith M, Paul JE, Pugh TJ, Tang MJ, Morin RD, Asano JK, Ally A, Miao L, Cheung P, Lee A, Chan SY, Taylor G, Severson T, Cheng GSW, Novik K, Gill S, Owen D, Brown CJ, Morin GB, Tai IT & Marra MA. *Genomic analysis of uridine monophosphate synthetase reveals novel mRNA isoforms and mutations associated with fluorouracil resistance in colorectal cancer*

degradation[8] but may also be influenced by variation in apoptosis[9, 10], nucleotide transport[11], DNA repair[12] and other mechanisms[13].

At least 13 genes including uridine monophosphate synthetase (*UMPS* aka *OPRT*), dihydropyrimidine dehydrogenase (*DPYD* aka *DPD*), thymidine phosphorylase (*TYMP* aka *TP*), uridine phosphorylase 1 (*UPP1* aka *UP*), and thymidylate synthetase (*TYMS* aka *TS*) are known to be involved in the metabolism of 5-FU (**Figure 4.1**)[14]. Mutations or epigenetic modifications which alter the structure or expression level of these genes and thereby affect the activity of the enzymes they encode may contribute to 5-FU resistance and poor outcome in some cancer patients. One of the most studied genes related to 5-FU action is *DPYD,* for which a clinical test for 5-FU toxicity exists (Myriad Genetics Inc.[15]; DNAVision SA[16]; and GenPath Diagnostics Inc.[17]). *DPYD* deficiency is associated with increased probability of severe adverse response to 5-FU including multi-organ toxicity, especially neurotoxicity[18, 19]. Drugs such as S-1, Tegafur-Uracil (UFT) and Eniluracil have been developed to improve response to 5-FU primarily by inhibiting *DPYD* and thereby reducing catabolism of 5-FU in the liver and increasing the amount of drug reaching the tumor[18] (**Appendix A**). Although *DPYD* is believed to account for a large percentage (as much as 80%) of the catabolism of 5-FU in the liver[18], expression of additional 5-FU metabolism genes (including *UMPS*) within the tumor is required for conversion of 5-FU to active anti-tumour metabolites (**Figure 4.1**)[20, 21].

The *UMPS* protein contains two enzymatic domains, orotate phosphoribosyltransferase (OPRTase; EC 2.4.2.10) and orotidine-5'-phosphate decarboxylase (ODCase; EC 4.1.1.23). In the synthesis of pyrimidines, the OPRTase domain functions by addition of a ribose group to orotate resulting in orotidine monophophate. The ODCase domain removes a carboxyl group from this molecule resulting in uridine monophosphate. In the metabolism of 5-FU, these two domains function by producing fluorinated versions of orotidine monophophate (FOMP) and uridine monophosphate (FUMP) (see **Figure 4.1**). Fluorinated uridine molecules may inhibit RNA synthesis or upon incorporation into RNA may cause RNA damage and interfere with translation[7].

Recent studies have emphasized the potential role for *UMPS* in mediating 5-FU resistance[22] and proposed that this gene may serve as a clinical biomarker of resistance in cancer patients, but there is considerable disagreement in the literature as to the

relative importance of genes predicted to be involved in 5-FU action[23]. Determining those genes that are critical to this process and identifying gene variants that may confer drug resistance can be accomplished by genome-wide analyses that compare drug-sensitive and resistant cell populations.

The use of genomic methods to identify the molecular causes of disease phenotypes is an area of rapid development[24]. We recently developed two novel genome-wide analytical approaches to profile alternatively processed transcripts. The first method uses splicing microarrays for 'alternative expression analysis' (ALEXA)[25] and the second uses an Illumina DNA sequencer for whole transcriptome shotgun sequencing (WTSS)[26]. We applied these methods to the study of 5-FU resistance in colorectal cancer cell lines. Using these approaches we observed differential expression of *UMPS* isoforms in a cell line with acquired 5-FU resistance relative to the sensitive cell line from which it was derived. We used RT-PCR and quantitative real-time RT-PCR to validate these differential expression observations and surveyed the expression of the two most abundant isoforms in additional sensitive/resistant cell lines and a panel of 26 colorectal cancer tumour/normal pairs. We catalogued the diversity of *UMPS* transcript variation in these cell lines by creation, sequencing and analysis of 293 *UMPS* cDNA clones. We then performed genomic DNA sequencing of the *UMPS* locus to identify mutations that might account for differential expression of isoforms or affect *UMPS* function. Finally, we performed mutation analysis of the exons and splice sites of *UMPS* using a panel of 91 pre- and post- treatment colorectal cancer patient cases.

4.2. Results

4.2.1. Differential expression analysis of *UMPS* isoforms in 5-FU sensitive and resistant cell lines.

We previously applied custom designed alternative expression analysis (ALEXA) microarrays to profile polyA+ RNA isolated from MIP101 (5-FU sensitive) and MIP/5FU (5-FU resistant) colorectal cancer cell lines[25]. Analysis of these data revealed an apparent reciprocal differential expression of two isoforms of *UMPS*, one containing six exons and essentially matching the reference sequence (NM_000373) and the other skipping exon 2 (**Figure 4.2**). We subsequently observed differential expression of the same alternative isoforms by WTSS analysis of the same polyA+ RNAs (**Figure 4.3**). In

the MIP101 WTSS library, mapped reads corresponded almost exclusively to the canonical *UMPS* isoform A (A/B ratio of 51.4). In the MIP/5FU WTSS library an approximately equal ratio of reads supporting the exon-exon junctions of *UMPS* isoforms A and B were observed (A/B ratio of 0.8) (**Figure 4.3**). RT-PCR, semi-quantitative RT-PCR and real-time quantitative RT-PCR were used to profile the expression of *UMPS* isoform A and B in three 5-FU sensitive/resistant cell line pairs. Resistant derivatives of MIP101, RKO and HCT116 (MIP/5FU, RKO/5FU and HCT/5FU respectively) were created as previously described[27]. Primers were designed to amplify bands corresponding to both isoform A and B or to specifically amplify one isoform and not the other. Both isoforms could be detected in all six cell lines (**Table 4.1**). All three resistant cell lines (MIP/5FU, RKO/5FU, and HCT/5FU) exhibited a significant down-regulation of *UMPS* isoform A relative to the sensitive cell line from which they were derived (**Table 4.2**). All three resistant cell lines also exhibited a significant up-regulation of *UMPS* isoform B (**Figure 4.4** and **Table 4.2**) although the finding for RKO was less convincing. Isoform A was 5-fold less abundant in MIP/5FU cells relative to MIP101. Isoform B was 20-fold more abundant in MIP/5FU and 5-fold more abundant in HCT/5FU cells relative to MIP101 and HCT116 respectively (**Table 4.2**). The difference in abundance of each isoform between the RKO and RKO/5FU was small but statistically significant in the quantitative real-time RT-PCR results (**Table 4.2**).

4.2.2. Characterization of *UMPS* transcript structural diversity

We created 293 *UMPS* cDNA clones by RT-PCR of polyA+ RNA using an oligo-dT primer for ss-cDNA synthesis followed by amplification with primers designed to flank the *UMPS* ORF. The resulting products were analyzed by gel electrophoresis (**Figure 4.5**). All six cell lines exhibited a band of the expected size for the canonical *UMPS* isoform A (~2,100 bp) and relatively smaller amounts of additional bands ranging in size from ~1,200bp to ~2,400 bp. These additional bands were predicted to correspond to alternatively spliced mRNA isoforms. All visible PCR products were gel purified and cloned by Topo-TA cloning. Clones were verified for size and orientation by restriction enzyme digestion (see **Methods**) and 96 clones were selected for full-length sequencing. Clone sequences were assembled (see **Methods**) and analyzed by BLAT alignment against the human genome (hg18). 95 of 96 clones mapped to the *UMPS* locus and appeared to correspond to either the canonical isoform or a novel isoform.

One clone appeared to represent a contaminating genomic DNA sequence from *Ralstonia pickettii* (99% BLAST identity over 1670 bp).

We successfully generated clones representing both the 'canonical' *UMPS* isoform A, the predicted *UMPS* isoform B as well as 8 other novel isoforms. Represented within our clone collection are a total of 10 distinct spliced *UMPS* isoforms, which we denoted A-J (**Figure 4.6** & **Table 4.3**). Only 4 of 10 alternative isoform sequences were supported by existing mRNAs or ESTs reported by the UCSC genome browser database[28]. The exon sequences, exon junctions, splice sites, and predicted ORFs of all 10 isoforms were examined (**Figure 4.6** & **Table 4.3**). The genome coordinates of exon boundaries from BLAT alignments of each clone to the human genome were used to extract a full length sequence for each apparent alternative isoform using the EnsEMBL API[29]. Each isoform sequence was represented by at least one fully sequenced clone and the sequences of these clones were submitted to GenBank (see **Table 4.3** for GenBank identifiers). These sequences were used to predict open reading frames for each clone using the NCBI ORF finder[30]. Each sequence was examined to determine if it would be targeted by nonsense mediated decay (NMD) (See **Table 4.3**). The predicted coding sequence used for this test was the largest ORF identified by the NCBI ORF finder. A sequence was tagged as a NMD target if the predicted coding sequence terminated more than 50 bp upstream of a splice site (an empirically determined criterion used by the vertebrate genome annotation effort[31] and others[32]). Only isoform E was a candidate for NMD by this criterion. Isoform E appeared to match UCSC known gene 'uc003ehm.1' (hg18) which is classified by UCSC as 'nearCoding' and a NMD target. In addition to scanning for NMD candidates, all exon boundaries were examined to determine whether they appeared to be valid splice sites. The genomic region of *UMPS* (+/- 1 kb of flank) was analyzed using four splice site prediction algorithms: ASSP[33], GeneSplicer[34], NetGene2[35], and NNSPLICE[36]. Of a total of 19 observed splice sites, 14 of these appeared to be valid according to one or more of these algorithms. The 5 non-canonical sites are noted in **Figure 4.6**. 7 of 10 isoforms contained only valid splice sites. The non-canonical splices sites observed in isoform C and isoform I affected only the 3' UTR.

The reading frames identified in all nine alternative isoforms retained the reading frame (at least partially) of the canonical isoform A. Conservation of the protein sequence of isoforms B-J relative to the canonical isoform (A) ranged from 43.1% to

100% (**Table 4.3**). Only isoform I was predicted to result in a protein identical to isoform A. All other isoforms had some loss of amino acid content relative to isoform A and in all cases this affected one or both of the known enzymatic domains of *UMPS*; OPRTase and ODCase (**Figure 4.6**). Furthermore, according to the conserved domain database[37] and published reports examining the structure and function of *UMPS* in yeast, mouse and human, the ODCase domain of *UMPS* contains a dimer interface that results in dimerization of *UMPS in vivo*[38]. 18 residues involved in this dimerization have been annotated within the ODCase domain (in exons 3, 4, 5 and 6). In addition to a loss of portions of the ODCase enzymatic domain, isoforms E, G and J were predicted to differ in their ability to dimerize.

4.2.3. *UMPS* protein expression

Western analysis of total protein lysate isolated from six 5-FU sensitive and resistant cell lines (MIP101, MIP/5FU, RKO, RKO/5FU, HCT116 and HCT/5FU) was performed with three *UMPS* antibodies (**Methods**). A monoclonal mouse antibody (Abnova; H00007372-M06) raised against a partial recombinant protein (amino acids 381-480 of 480 total) detected the presence of a band at the expected molecular weight (52 KDa) for *UMPS* isoform A in all six cell lines. A loss of expression of *UMPS* isoform A from sensitive to resistant cells was apparent in each pair of cell lines (**Figure 4.7a**). Quantitative analysis in which loading amounts were normalized to Actin expression (**Methods**) confirmed the reduced abundance of *UMPS* isoform A protein in MIP/5FU lysate compared to MIP101 lysate (**Figure 4.7b**). A band corresponding to the predicted molecular weight of isoform B (33 KDa) was not detected in any cell line. Two additional antibodies, one a monoclonal mouse antibody raised against a partial recombinant *UMPS* peptide (Abnova; H00007372-M05) and the other a polyclonal mouse antibody raised against a full length *UMPS* (Abnova; H00007372-B01P) were tested. These antibodies detected the purified immunogen purchased from the manufacturer but did not detect *UMPS* expression in any cell line lysate.

4.2.4. Survey of *UMPS* isoform expression in treatment naïve colorectal tumor samples

To determine the normal expression of alternative *UMPS* isoforms A and B in colorectal tumours we measured their expression using PCR (qualitative) and real-time RT-PCR

(quantitative) assays in 26 patient cases obtained from the Ontario Tumour Bank (Toronto, ON) that were mostly treatment naïve (80%) (**Methods**). Each case consisted of a primary colon or rectal tumour and matched normal adjacent tissue that was fresh frozen at the time of surgical resection. Isoform B was easily detectable in the majority of patients and its level was variable from patient to patient, but it represented the minor form compared to isoform A (**Figure 4.8**). Quantitative real-time PCR confirmed that *UMPS* isoform B is a minor variant compared to isoform A in treatment naïve colorectal tumour tissue and matched adjacent normal (**Figure 4.9**). However, the expression of both isoforms, but especially isoform A, was significantly higher in tumour than in normal tissue in this treatment naïve cohort. No significant difference in the expression of either isoform was observed between survivor (n=16) versus non-survivor (n = 10) patient groups, between progressors (n = 9) and non-progressors (n = 13), or between pre-treatment (n = 5) and post-treatment tumours (n = 21) (**Methods**).

Since the samples analyzed in these PCR assays were mostly treatment naïve, they would not be expected to reveal modifications of *UMPS* (such as aberrant splicing) associated with acquired resistance to 5-FU (i.e. that occurs upon exposure the drug). To further investigate the possible role of UMPS in 5-FU resistance we therefore sought to obtain patient samples that had been exposed to 5-FU. The only samples of this type we could obtain were archival materials that had been formalin fixed and paraffin embedded. Unfortunately, we were unable to reliably measure abundance of mRNA isoforms in these archival samples due to the severe RNA degradation resulting from the formalin fixing and long-term storage of the samples (**Methods**). For this reason, we relied on mutation analysis to take advantage of the relative stability of genomic DNA compared to RNA. We started by sequencing the exon and intron boundaries of *UMPS* using genomic DNA isolated from 5-FU sensitive and resistant cell lines and then proceeded to apply similar sequencing assays to a cohort of colorectal cancer patient samples.

4.2.5. Sequencing of *UMPS* in 5-FU sensitive and resistant cell lines

Although the reciprocal differential expression of *UMPS* isoforms A and B that we observed between 5-FU sensitive and resistant cells appeared to be the result of a change in splicing patterns, we sought to elucidate the underlying mechanism. We hypothesized that one or more mutations were acquired at the *UMPS* locus during the

selection of 5-FU resistance and these were responsible for the altered splicing pattern. To investigate this hypothesis, we sequenced genomic DNA extracted from each of the 5-FU sensitive/resistant cell line pairs as well as a commercially available genomic DNA sample. We generated 22 PCR amplicons (with an average size of 710 bp) covering the genomic region of *UMPS* from 1 kb upstream of exon 1 to the end of exon 3 and sequenced each amplicon by direct end sequencing with M13 linkers added to each pair of primers (**Figure 4.10**) (see **Appendix B** for primer sequences). The resulting genomic sequence captured all four splice sites involved in the splicing of exon 2. For the MIP101 and MIP/5FU cell lines we also examined the Illumina WTSS data for evidence of mutations (**Methods**).

Sequence discrepancies relative to the reference human genome within amplicon sequence data or WTSS data were identified and classified as known SNPs or putative novel mutations by comparison to dbSNP[39]. A total of 33 variations relative to the human reference genome were identified. 28 of these were found to be known polymorphisms in dbSNP[39]. Two variants were novel relative to dbSNP but found in both sensitive cells (RKO) and the resistant derivative (RKO/5FU). Thus, three variants appeared to be mutations specific to a 5-FU resistant cell line (i.e. acquired in the resistant line in the process of making it resistant to 5-FU). The first of these was a heterozygous splice site mutation at the acceptor splice site of exon 2 . This mutation was present in the MIP/5FU (resistant) cell line but not the MIP101 (sensitive) cell line (position 5,727 in **Figure 4.10** and **Table 4.4**). This mutation suggests a mechanism for the difference in isoform expression ratios observed between the MIP101 and MIP/5FU cell lines. In the MIP101 cells, both alleles of the *UMPS* locus will produce pre-mRNAs with a canonical splice site at exon 2. The splicing machinery is predicted to recognize this splice site and include exon 2 in the mRNA, resulting in the long isoform (isoform A). However, in MIP/5FU, one allele is wild type but the second allele is mutated at the acceptor splice site for exon 2. For all pre-mRNAs generated from this allele, the splicing machinery is predicted to not recognize the mutated exon 2 acceptor site, resulting in splicing to yield isoform B transcripts. Thus, the two alleles would together be expected to create an approximately equal mixture of the mRNAs for the two isoforms and this is what we observed in our data (**Figure 4.2**, **Figure 4.3** & **Figure 4.4**). We speculated that the mutation might act as either (1) a loss-of-function mutation which confers resistance due to a gene dosage effect because the resistant cells have

one functional copy instead of two or (2) a dominant negative mutation in which the function of an aberrant isoform B actively antagonizes the function of isoform A. The apparent lack of translation of isoform B intro protein as suggested by our Western analysis (**Figure 4.7**) supports the theory that this splice site mutation is a loss-of-function mutation.

A second novel mutation acquired in 5-FU resistant cells was observed in the RKO/5FU line. This mutation is also heterozygous and predicted to result in an amino acid substitution within *UMPS* exon 2 (Pro→Ser) (position 5,809 in **Figure 4.10** and **Table 4.4**).

A third heterozygous coding mutation supported by ~20 reads for each allele in the Illumina data appears to have been acquired in 5-FU resistant MIP/5FU cells and is predicted to result in an amino acid substitution within *UMPS* exon 3 (Arg→Cys) (position 8,211 in **Figure 4.10** and **Table 4.4**). MIP101 had excellent Illumina WTSS sequencing coverage (298X) at this base position and showed no evidence for this mutation. Sanger sequencing confirmed this mutation as heterozygous in MIP/5FU and not present in MIP101 (amplicon 22 in **Figure 4.10**). Several cDNA clones isolated from MIP/5FUR contained this mutation but none from MIP101 did. Those cDNA clones from MIP/5FU containing this mutation also included exon 2 while those with skipping of exon 2 did not have the mutation suggesting that this mutation may exist on the allele without the splice site mutation described above. Both of the amino acid changes observed in the resistant cell lines (Pro→Ser and Arg→Cys) were non-conservative with BLOSUM62 substitution matrix scores[40, 41] of -1 and -3 respectively.

Based on the observation, in 5-FU resistant cell lines, of a splice site mutation (in MIP/5FU) and additional mutations affecting the protein sequence of *UMPS* (in MIP/5FU and RKO/5FU) we next sought to determine whether these or other mutations occurred in colorectal cancer tumours, in particular, those isolated after the patient was exposed to 5-FU.

4.2.6. Sequencing of *UMPS* in colorectal cancer samples

In addition to 26 fresh frozen tumour/normal sample pairs from the Ontario Tumour Bank (discussed above) we obtained 44 archival samples from the BC Cancer Agency (Vancouver, BC) and 20 archival samples from St. Paul's hospital (Vancouver, BC). These 90 cases consisted of 24 primary colon tumours, 22 primary rectal tumours and

44 liver metastases of colorectal cancer. For 50 (55%) of these tumours we able to confirm some amount of exposure to 5-FU prior to resection (**Methods**). We hypothesized that exposure to 5-FU might result in selection for mutations in *UMPS* such as those we observed in cell lines.

To investigate this hypothesis, we sequenced genomic DNA extracted from each sample. We generated 22 PCR amplicons (with an average size of 379 bp) covering the six exons of *UMPS* as well as an alternative exon 2 ('2b') (**Figure 4.10**). Each amplicon was sequenced by direct end sequencing with M13 linkers added to each pair of primers (see **Appendix B** for primer sequences).

A total of 40 variations relative to the human reference genome were identified. 27 of these were found to be known polymorphisms in dbSNP[39]. Of the 13 putative novel mutations identified, 5 were within an intron (but not close to a splice site), 1 was within exon 2b (**Figure 4.6)**, 1 was at the -3 position of the acceptor splice site of exon 2, and the remaining 6 were coding mutations within exons 2, 3, 4 and 6. All 6 putative coding mutations fell within either the OPRTase or ODCase enzymatic domains (2 within the OPRTase domain and 4 within ODCase domain). All but two of the 13 mutations were heterozygous.

Interestingly, the heterozygous splice site mutation (position 5,725 in **Figure 4.10** and **Table 4.4**) occurred very close to the splice site mutation observed in the MIP/5FU cell line. Specifically, it occurred at the -3 position of the acceptor site compared to the -1 position for the splice site mutation identified in MIP/5FU. The patient (#14) harboring the -3 mutation of the exon 2 acceptor site had an unusually aggressive cancer. The patient presented with stage 4 disease with multiple local and distant metastases (including metastases at 13 of 16 lymph nodes) and disease progression resulted in death within 3 months of the primary resection. Patient 14 also had a heterozygous missense (Ala->Thr) mutation within exon 2 (position 5,731 in **Figure 4.10** and **Table 4.4**). Examination of the matched normal sample for patient 14 revealed that both mutations were somatic. Aberrant splicing of exon 2 as observed in the cell line MIP/5FU was not apparent by examination of **Figure 4.8** or the quantitative RT-PCR data. The remaining five putative coding mutations were identified in 5 separate patients (4 of 5 were post-treatment). One of the mutations was silent and the rest were missense mutations although none of these were particularly radical amino acid

substitutions according to BLOSUM62 substitution matrix scores[40, 41] (scores for these substitutions were >= 0; see **Table 4.4** for details).

The number of mutations observed in post-treatment samples compared to pre-treatment samples was not significantly more than expected by chance (taking into account the size of the pre- and post- treatment groups) for either the entire set of 13 mutations (p-value=0.0589; by two-sided Fisher's exact test) or for the 6 mutations predicted to affect amino acid sequence (p-value=1).

4.3. Discussion

Resistance to 5-FU is hypothesized to arise by a number of mechanisms but previous studies have primarily focused on 5-FU metabolism genes and five genes in particular (*DPYD*, *TYMP*, *TYMS*, *UPP1*, and *UMPS*). *UMPS* expression has been described as potentially critical to the response of a tumor to 5-FU and is widely cited as the primary means of activating the pro-drug 5-FU to active anti-tumour metabolites[42, 43] (**Figure 4.1**). A number of previous studies have reported attempts to use *UMPS* expression as a predictor of drug response with varying degrees of success. Some of these studies suggested that measuring *UMPS* expression is useful in predicting 5-FU response[23, 44, 45] while others claimed that *UMPS* expression was not correlated with 5-FU response[46-49]. Many of these studies employed a strategy that involved measuring the expression of *UMPS* by use of probes targeting the 3′ end of the canonical *UMPS* transcript without regard to potential alternative isoforms. Other studies employed biochemical assays of enzymatic activity which are desirable in that they do not rely on inference but are difficult to apply in a clinical setting and require large amounts of fresh tumour tissue. To our knowledge no previous study examining *UMPS* mRNA expression has considered the presence of alternatively spliced or mutated variant isoforms or performed transcript re-sequencing to characterize their diversity.

Using a combination of novel and conventional approaches we observed expression of a mutated or aberrantly spliced isoform in the 5-FU resistant derivatives of three cell lines: (1) significant over-expression of isoform B in MIP/5FU caused by a heterozygous splice site mutation, (2) mild over-expression of isoform B in HCT/5FU (via an unknown molecular mechanism) and (3) expression of a mutant isoform A with a protein coding change in RKO/5FU. We showed that in pre-treatment colorectal tumours, *UMPS* isoform B mRNA was generally present but represented a minor form relative to isoform

A. In addition to isoform B we identified 8 additional novel *UMPS* isoforms. We examined the transcript structure and predicted protein content of these isoforms but further work is required to elucidate the general importance of alternative splicing of *UMPS* in 5-FU resistance and the specific activity of the individual isoforms we report. We also observed novel mutations in human colorectal cancer patient samples. All eight of the coding or splice site point mutations we identified were heterozygous (**Table 4.4**). While these heterozygous mutations might confer a selective advantage in the presence of 5-FU by reducing the activation of 5-FU, homozygous mutations are perhaps unlikely to occur due to the apparent importance of *UMPS* in providing pyrimidines for the normal functioning of the cell. Inherited deficiency of the enzymatic functions of *UMPS* is associated with the rare human disorder 'Orotic Aciduria I' (OMIM: 258900), a disease that is fatal unless treated with large doses of uridine. Inherited heterozygous *UMPS* loss-of-function point mutations in exons 2, 3 and 6 were reported as the likely cause of this disorder in humans[50]. To our knowledge homozygous mutations of *UMPS* have not been reported in humans but homozygous *UMPS* deficiency caused by inheritance of a point mutation that introduced a premature stop codon in exon 5 has been shown to cause early embryonic death in cattle embryos[51, 52]. Based on the deleterious effect of *UMPS* deficiency, germline mutations that might cause intrinsic 5-FU resistance are expected to be rare. However, somatic mutations such as those we report that are acquired during tumorigenesis or chemotherapy treatment may contribute to acquired 5-FU resistance in the treatment of colorectal cancer. Homozygous somatic mutations of *UMPS* are perhaps still unlikely however, given the importance of pyrimidine metabolism in the actively dividing cells of a tumour.

Since the treatment of colorectal cancer generally starts with surgical resection of the primary tumour, few post-treatment primary tumours could be obtained and those we could obtain were of insufficient quality to allow RNA analysis. However, analysis of RNA from frozen post-treatment recurrences or metastases should be possible and will be required to determine whether selection of aberrantly spliced or mutated *UMPS* is a common feature of 5-FU resistant tumours. Analysis of such a cohort with the assays we report should confirm their utility in predicting clinical response to 5-FU. If aberrant splicing in particular proves to be an important mode by which *UMPS* activity is modulated to confer 5-FU resistance then it may be possible to reverse this resistance

by specifically targeting the splicing machinery with anti-sense oligonucleotides designed to shift the balance of expression to more active isoforms[53].

Based on our observations, the inconsistent correlation between *UMPS* expression and 5-FU response or patient outcomes previously reported might be due to the failure to consider transcript structure and mutational status of *UMPS*. For example, a shift in the relative expression levels of *UMPS* isoforms A and B might have an effect on overall *UMPS* activity but would be undetectable by previously reported assays that measured the 3′ end of *UMPS* [45, 54]. Similarly, the expression level might remain unchanged but UMPS activity might be reduced by the presence of point mutations that affect the protein sequence. Future studies of the role *UMPS* in 5-FU resistance should consider not just the level of *UMPS* mRNA abundance but also the expression of alternative *UMPS* isoforms as well as the mutational status of the gene. Increased abundance of aberrant isoforms (such as isoform B) or mutated *UMPS* protein might be useful in predicting 5-FU response. However, future studies will need to develop functional assays to assess the degree to which aberrantly spliced or mutated *UMPS* variants such as those we identified are sufficient to confer resistance (and if so, to what degree). Evaluation of larger numbers of fresh frozen clinical samples with known exposure to 5-FU will also be required to determine to what degree the complexity of *UMPS* isoform expression pattern or presence of mutations within *UMPS* may contribute to 5-FU resistance in the patient population. Information obtained from these studies will increase our understanding of the mechanisms of drug resistance and may be useful in the development of tests to assist in treatment optimization for individual patients. Improved understanding of the mechanisms of 5-FU resistance will also drive the development of new treatments that avoid or overcome resistance. Several examples of modified versions of 5-FU or combinations of 5-FU that take advantage of knowledge of drug resistance mechanisms have already been described (see **Appendix A** for details). For example, the drug ‘S1’ combines 5-FU with 5-chloro-2,4-dihydroxypyridine (CDHP), an inhibitor of *DPYD*, to reduce catabolism in the liver and thereby increase the effective dose that reaches the tumour. A complementary strategy that was capable of restoring or increasing expression of a functional *UMPS* isoform in patients with *UMPS* mutations or aberrant splicing might improve the efficacy of 5-FU treatment in these patients.

4.4. Methods

4.4.1. Cell lines

The 5-FU sensitive cell lines MIP101[55], HCT116 and RKO and 5-FU resistant cell lines MIP/5FU[27], RKO/5FU, and HCT/5FU were maintained in DMEM media supplemented with 1% penicillin-streptomycin, 1% kanamycin (Invitrogen Inc. Burlington, ON., Canada) and 10% newborn calf serum at 37°C and 5% CO_2[27]. For resistant cell lines, media were also supplemented as follows: MIP101 cells resistant to 5-FU (MIP/5FU), 50μM 5-FU; HCT116 cells resistant to 5-FU (HCT/5FU), 10μM 5-FU; RKO cells resistant to 5-FU (RKO/5FU), 25μM 5-FU.

4.4.2. Clinical samples

Samples corresponding to colorectal cancer cases were obtained from the Ontario Tumour Bank (Ontario Institute for Cancer Research, Ontario, Canada), BC Cancer Agency (Vancouver, BC, Canada) and St. Paul's Hospital (Vancouver, BC, Canada). Ethics approval for work with these sample was obtained from the BC Cancer Agency Research Ethics Board (see **Appendix C** for ethics certificates). For all three sample sources, a review of each case was performed by a pathologist and if the sample was deemed to be less that 70% tumour content or found to contain significant signs of necrosis, the sample was excluded. All patients received adjuvant or neo-adjuvant chemotherapy containing 5-FU. No other inclusion or exclusion criteria were applied.

Fresh frozen colorectal tumor samples with matched adjacent normal tissue were obtained for 26 patient cases from the Ontario Tumour Bank. For each case, we received ~250 mg of frozen tissue that had been stored at -80°C for 13 to 46 months. These samples represented resections of primary tumours. 21 of these patients received adjuvant 5-FU and 5 received neo-adjuvant 5-FU (5 cases). 9 'responders' were defined as patients with no progression reported (follow-up was 5 to 22 months). 15 'non-responders' were defined as those patients whose records noted one or more of the following criteria: local or distant recurrence; disease progression resulting in death; adverse drug response (e.g. neutropenia, neuropathy, etc.).

44 samples representing liver metastases of colorectal adenocarcinoma were obtained from the BC Cancer Agency. In these cases the patients presented with primary colorectal cancer which was resected and followed by adjuvant chemotherapy

treatment (including 5-FU). Each of these patients was also diagnosed with a liver metastasis that was also resected. The time between the resection of the primary tumour and metastasis ranged from 0 to 68 months, resulting in a varying degree of exposure to 5-FU prior to resection of the liver metastasis. For 27 out of 44 of these cases we were able to confirm the patient's exposure to 5-FU prior to resection of the metastasis. For each of the 44 liver metastases, we obtained 10 scrolls (of 10 μM thickness each) from formalin fixed paraffin embedded (FFPE) blocks that had been stored for 33-88 months.

20 samples representing primary rectal adenocarcinoma tumours were obtained from St. Paul's Hospital. In these cases, the patients presented with primary rectal cancer and were given neo-adjuvant chemotherapy in conjunction with radiation for six weeks prior to surgery. 18 of 20 of these patients received 5-FU or Capecitabine (oral 5-FU) as their neo-adjuvant chemotherapy. For each of these 20 primary tumours, we obtained 10 scrolls (of 10 μM thickness each) from FFPE blocks that had been stored for 3-38 months.

4.4.3. RNA Isolation

Total RNA was isolated from cells cultured to ~75% confluence using RNeasy columns (Qiagen, Mississauga, ON, Canada). RNA was DNAseI treated using an RNAse free DNAseI kit (Invitrogen). RNA was quantified and tested for degradation using an Agilent 2100 Bioanalyzer and RNA Nano Assay (Agilent, Santa Clara, CA, USA). PolyA+ RNA was purified from total RNA using an oligoTex kit (Qiagen).

4.4.4. Genomic DNA isolation

Homogenization of tissues or cell lines was performed by two 30 second bursts with a hand held homogenizer (VWR. International. Mississauga, Ontario. Cat. No. 47747). Genomic DNA was isolated using a Gentra PureGene kit (Qiagen Inc.). For cell lines, genomic DNA was isolated from cells grown to ~75% confluence. A reference genomic DNA sample was obtained from ClonTech (Mountain View, CA., USA. Cat. #636401). For fresh frozen patient samples (52 total), genomic DNA was isolated from ~20 mg of frozen tissue. For formalin fixed paraffin embedded (FFPE) samples (64 total), genomic DNA was isolated from two 10μM scrolls. Yield of genomic DNA was determined by use of a Nanodrop ND-8000 spectrophotometer (Nanodrop, Wilmington, DE, USA) and

quality of genomic DNA was qualitatively assessed by gel electrophoresis using 0.7% agarose gels.

4.4.5. Splicing microarray analysis

Creation of custom 'ALEXA' splicing microarray designs, sample preparation, microarray hybridization and analysis of microarray data were performed as previously described[25]. Additional details are provided online at www.AlexaPlatform.org.

4.4.6. Whole transcriptome shotgun sequencing and analysis

Library construction, Illumina sequencing, read mapping and analysis were as previously described[26].

4.4.7. RT-PCR and semi-quantitative RT-PCR validation of *UMPS* isoform expression

Single stranded cDNA was generated from 500 ng of polyA+ RNA isolated from each cell line using SuperScript III reverse transcriptase and random hexamer primer (Invitrogen). PCR primers were designed to flank exon 2 (see **Appendix B** for primer sequences). PCR was performed with Invitrogen's Platinum Pfx enzyme. Semi-quantitative detection of PCR products representing alternative *UMPS* isoforms was performed using a 2100 'lab-on-a-chip' Bioanalyzer and DNA 7500 Assay (Agilent. Cat. #5067-1506).

4.4.8. Quantitative real time RT-PCR

Single stranded cDNA was generated from total RNA isolated from cell lines or patient samples using SuperScript III reverse transcriptase and random hexamer primer (Invitrogen). For cell lines and fresh frozen patient samples, 1µg of total RNA was used as input for cDNA synthesis with 50 ng of random hexamers. For formalin fixed paraffin embedded patient samples, 5 µg of total RNA and 250 ng of random hexamers were used. All other conditions for cDNA synthesis were according to the manufacturer's recommendations for SuperScript III. Quantitative PCR was performed on an Applied Biosystems 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) using 20 µL reaction volumes in 384-well plate format. Each reaction consisted of 50 ng of template cDNA, 0.4 µM final primer concentration and Power Sybr Green PCR Master Mix (Applied Biosystems). A relative standard curve was created by

serial dilution of a pool of cDNAs from 25 patient cases. The standard curve was applied to all primer sets. Primers were selected to amplify *UMPS* isoform A, isoform B, both isoforms and a housekeeping gene (*TBP*). Sensitivity and specificity of the *UMPS* A and B specific primers was verified by use of serial dilution of plasmid DNA from previously sequenced clones of each isoform. Analysis was performed using the relative standard curve method and the comparative Ct method (see Applied Biosystems documentation).

4.4.9. Cloning & sequence validation of *UMPS* mRNA isoforms

Single stranded cDNA was generated from polyA+ RNA isolated from each cell line using SuperScript III reverse transcriptase and oligo$(dT)_{20}$ primer (Invitrogen). PCR primers were designed to cover the full *UMPS* open reading frame and most of the UTR (using the *UMPS* reference sequence, NM_000373) (see **Appendix B** for primer sequences). PCR was performed with Invitrogen's Platinum Taq High Fidelity enzyme. PCR products were separated by gel electrophoresis (1.5% agarose gel, 60V for 12 hours at 4°C). Distinct bands were gel purified using a gel purification kit (Qiagen). Clones were generated by TOPO TA cloning into the vector pCR4-TOPO (Invitrogen). 293 clones were screened for correct insert size and forward orientation relative to the M13F site of the cloning vector by restriction enzyme digestion with EcoRI and NotI/XhoI (double digest) enzymes respectively. The EcoRI and NotI restriction enzyme sites do not occur within the *UMPS* reference sequence NM_000373. The XhoI site occurs once near the centre of exon 3 of the *UMPS* reference sequence NM_000373. The inserts of 96 clone fragments were fully sequenced by Sanger sequencing with an ABI 3730 device using M13F and M13R primers as well as custom primers. Clone sequences were assembled by Phred/Phrap and manually checked for quality using Consed as previously described[56]. Briefly, vector sequence was masked except for a short linker sequence at each end of each clone (5'-GAATTCGCCCTT-3'). Each clone consensus sequence was derived from an average of 8.4 reads, had an average overall Phred score[57, 58] of 89.0 and average minimum Phred score of 55.9.

4.4.10. Splice site analysis

The genomic sequence of *UMPS* (exons + introns + 1 kb flank at each end) was extracted from the human genome reference sequence using the UCSC genome

browser[28] (UCSC version hg18). The following splice site prediction algorithms were supplied this genomic sequence as input and using the default options for each: ASSP[33], GeneSplicer[34], NetGene2[35], and NNSPLICE[36]. All *UMPS* clone sequences were aligned to the human genome by BLAT and the exon boundary coordinates from these alignments were compared to the position of splice sites predicted by the four splice site prediction algorithms. In cases where no predicted splice site was found for an exon boundary, the boundary was examined manually for the two most common non-canonical splice sites 'GC … AG' and 'AT … AC'[59].

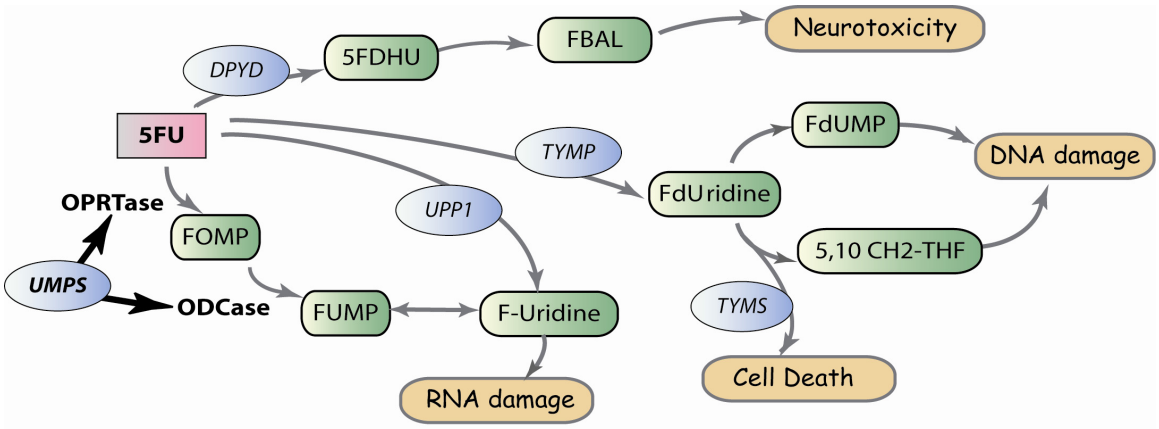
4.4.11. Western analysis

The following *UMPS* antibodies were purchased from Abnova (Walnut, CA, USA): two different clones of a mouse monoclonal antibody raised against a partial recombinant *UMPS* peptide (carboxy terminus; amino acids 381-480 of 480) (Cat. No. H00007372-M05 & H00007372-M06) and a purified mouse polyclonal antibody raised against a full-length *UMPS* protein (Cat. No. H00007372-B01P). Protein lysates were obtained by repeated freezing/thawing in CHAPS lysis buffer (50mM Pipes/HCl, 2mM EDTA, 0.1% CHAPS, 20 ug/mL leupeptin, 10 ug/mL pepstatin A, 10 ug/mL aprotinin, 5mM DTT, 1mM PMSF). Protein concentrations were determined by BioRad Protein Assay reagent (aka BradFord reagent) using a standard curve created with Bovine Serum Albumin (BSA). Each lane was loaded with 75 ug of total protein (determined by Bradford assay). Ladders used were: SeeBlue Plus2 Pre-Stained Standard (Invitrogen; Cat. No. LC5925), Novex Sharp Protein Standard (Invitrogen; Cat. No. LC5800) or PageRuler Pre-Stained Protein Ladder (Fermentas; Cat. No. SM0671). Samples were run on a 4-12% Bis Tris Gradient Gel (Invitrogen) for 90 minutes at 150 volts. Proteins on the gel were transferred overnight to a Nitrocellulose membrane at 100 mA and the membrane was blocked with Odyssey blocking buffer (LI-COR Biosciences, Lincoln, NE, USA) for 2 hours followed by incubation with the antibody at 1:200 dilution (2.5 ug/mL) in Odyssey blocking buffer for 2 hours. The membrane was then washed 3 times for 5 minutes with TBS-Tween (0.1% v/v) and was incubated with secondary antibody (IR700 anti-mouse secondary) for 1 hour. Finally, the membrane was then washed 3 times for 5 minutes with TBS-Tween (0.1% v/v) and scanned with an Odyssey scanner (LI-COR Biosciences).

4.4.12. PCR and sequencing the *UMPS* locus

The *UMPS* locus was sequenced by generating 44 amplicons covering all 6 exons and at least 50 bp of flanking intron sequence (**Figure 4.10**). These amplicons were generated by PCR using genomic DNA template from six cell lines and 118 patient samples. Each primer contained either an M13F or M13R linker that was used for direct sequencing of PCR products (see **Appendix B** for primer sequences). PCR was performed with Platinum Taq High Fidelity enzyme (Invitrogen) and each amplicon was bead purified using Agencourt Ampure Beads from Beckman Coulter (Beverly, MA., USA) and Sanger sequenced with an ABI 3730 device using M13F and M13R primers. Reaction conditions were optimized for each primer pair. The success rate for these PCRs (defined as a single visible band observed by gel electrophoresis) was 86% to 100% for the 44 amplicons (mean of 95%). Sequencing of the target amplicons was carried out by the BC Cancer Agency Genome Sciences Centre production sequencing group using previously published reaction chemistries[60]. The mean number of sequence bases with a Phred quality[57, 58] of 20 or higher ranged from 55% to 89% of the expected amplicon length across the 44 amplicons (overall mean was 79%). Sequence analysis for identification of mutations was conducted with Mutation Surveyor (SoftGenetics).

Figure 4.1. Simplified 5-FU metabolism pathway
Enzymes involved in the metabolism of 5-FU are indicated by blue shading*. The two metabolic steps encoded by *UMPS* are shown separately. Fluorine modified metabolites are indicated by green shading and outcomes are indicated by orange shading.



Gene names used are Entrez Gene IDs (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene)
UMPS (Entrez: 7372) aka: *OPRT*
DPYD (Entrez: 1806) aka: *DPD*, *DHP*
TYMP (Entrez: 1890) aka: *TP*, *ECGF1*, *MNGIE*, *PDECGF*
UPP1 (Entrez: 7378) aka: *UP*, *UPP*, *UPASE*, *UDRPASE*
TYMS (Entrez: 7298) aka: *TS*, *TMS*, *TSase*

UMPS enzymatic domains
OPRTase (EC 2.4.2.10): orotate + 5-phospho-a-D-ribose 1-diphosphate = orotidine 5'-phosphate + diphosphate
ODCase (EC 4.1.1.23): orotidine 5'-phosphate = UMP (uridine 5'-phophate) + CO2

*Relationships used to create this pathway were obtained from the Pharmacogenomics Knowledge Base[14]. http://www.pharmgkb.org/search/pathway/5fu/5fu.jsp

Figure 4.2. Differential expression of alternative *UMPS* isoforms in 5-FU sensitive and resistant cells

(**A**) The uridine monophosphate synthetase (*UMPS*) locus (3q13). Observed alternative splicing of exon 2 is depicted by dotted lines. (**B**) The positions of ALEXA splicing microarray probesets (each consisting of 2-4 oligonucleotide probes) specific to *UMPS* isoforms A and B are depicted. Probes are labeled according to the exons or junctions they profile (e.g., E1-E3 detects the connection of exon 1 to exon 3). Black arrows indicate the start/stop position of the predicted open reading frame of each isoform and the position of protein domains is indicated beneath each isoform. (**C**) Box plots depict expression values for oligonucleotide probes from triplicate samples of MIP101 and MIP/5FU profiled on the ALEXA splicing microarray platform (**Methods**). The median log2 expression value of all exons (blue dotted line) and all negative controls (red dotted line) in the array data are also shown. Isoform A was ~5-fold more abundant in 5-FU sensitive cells than in resistant cells. Isoform B was ~6-fold more abundant in 5-FU resistant cells than in sensitive cells. The ratio of isoform A to B was 25.6 in MIP101 cells and 0.85 in MIP/5FU cells.

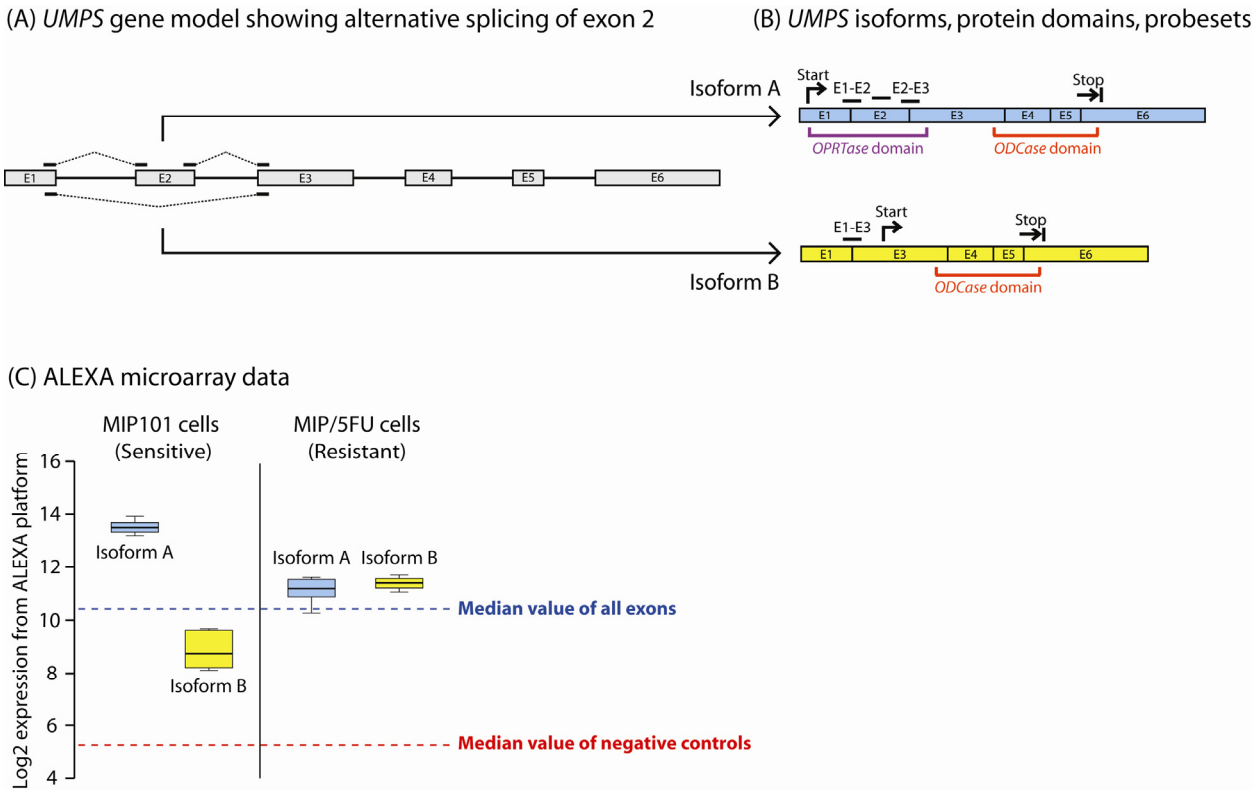


Figure 4.3. Whole transcriptome shotgun sequence data corresponding to the *UMPS* locus

Short read sequencing of polyA+ RNA isolated from MIP101 and MIP/5FU cells was performed with an Illumina GA2 DNA sequencer and the resulting reads (~334 million and ~191 million 42-mer reads respectively) were mapped to known and predicted EnsEMBL transcripts (**Methods**). Sequence read coverage for the *UMPS* locus is shown. The first panel shows the exon/intron structure of *UMPS*. The second and third panel show the base-level coverage and exon junction read counts for Illumina reads mapped to *UMPS* for MIP101 and MIP/5FU respectively. Black arrows indicate the start/stop position of the predicted open reading frame of each isoform. A total of 16,514 MIP101 reads and 4,662 MIP/5FU reads mapped to *UMPS* (**Methods**). For MIP101, 98.2% of known exonic bases were covered by 10 or more reads (average coverage of 293X) and for MIP/5FU, 94.3% of known exonic bases were covered by 10 or more reads (average coverage of 82X). In both MIP101 and MIP/5FU a mixture of exon-exon junctions were observed indicating the presence of both isoform A and B. In MIP101 the predicted ratio of A/B was 51.4 after normalizing for the number of junctions indicating each isoform and differences in library size. In MIP/5FU the predicted ratio of A/B was 0.8. In the figure, coverage values and exon junction read counts have been normalized to the size of the smaller library.

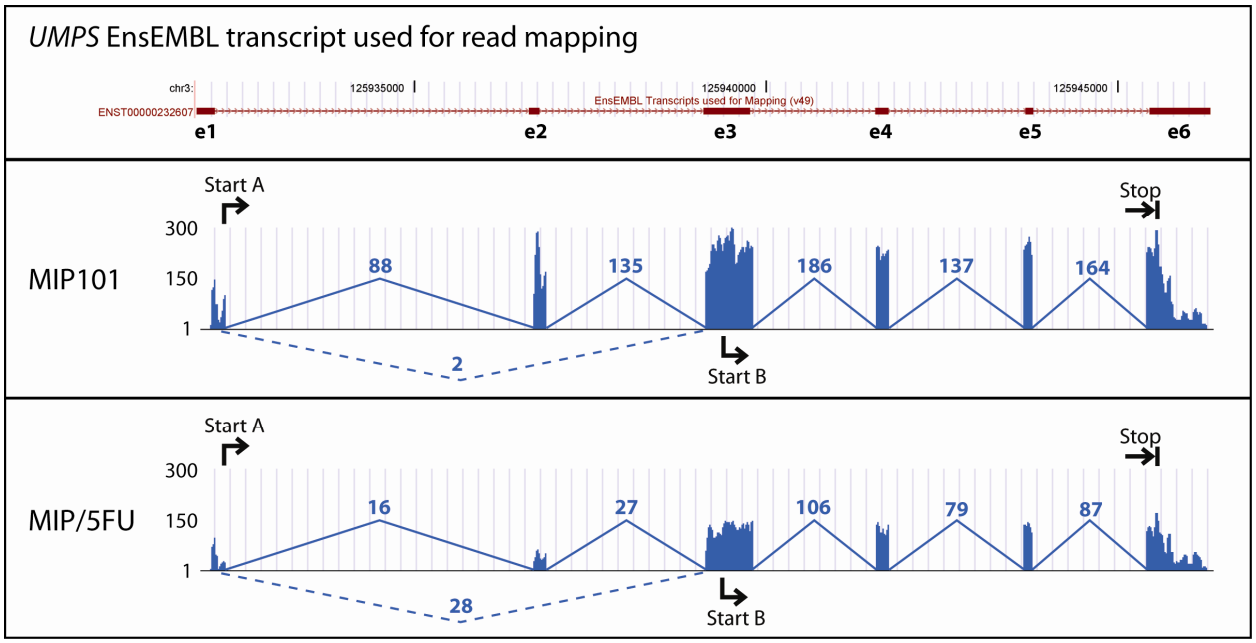


Figure 4.4. RT-PCR detection of *UMPS* isoforms in six 5-FU sensitive and resistant colorectal cancer cell lines

(**A**) Test of RT-PCR primers using mixtures of plasmid DNA clones representing *UMPS* isoforms A & B. These plasmids were mixed in the indicated molar ratios and used as templates for PCR reactions. The expected product sizes for this primer pair are 1,107 bp for isoform A and 953 bp for isoform B (missing exon 2). (**B**) The same PCR primers were used to amplify *UMPS* isoforms from ss-cDNA generated from polyA+ RNA extracted from 5-FU sensitive and resistant colorectal cancer cell lines. Both MIP101 and HCT116 showed an increase in the presence of *UMPS* isoform B in the 5-FU resistant derivative compared to the sensitive line.

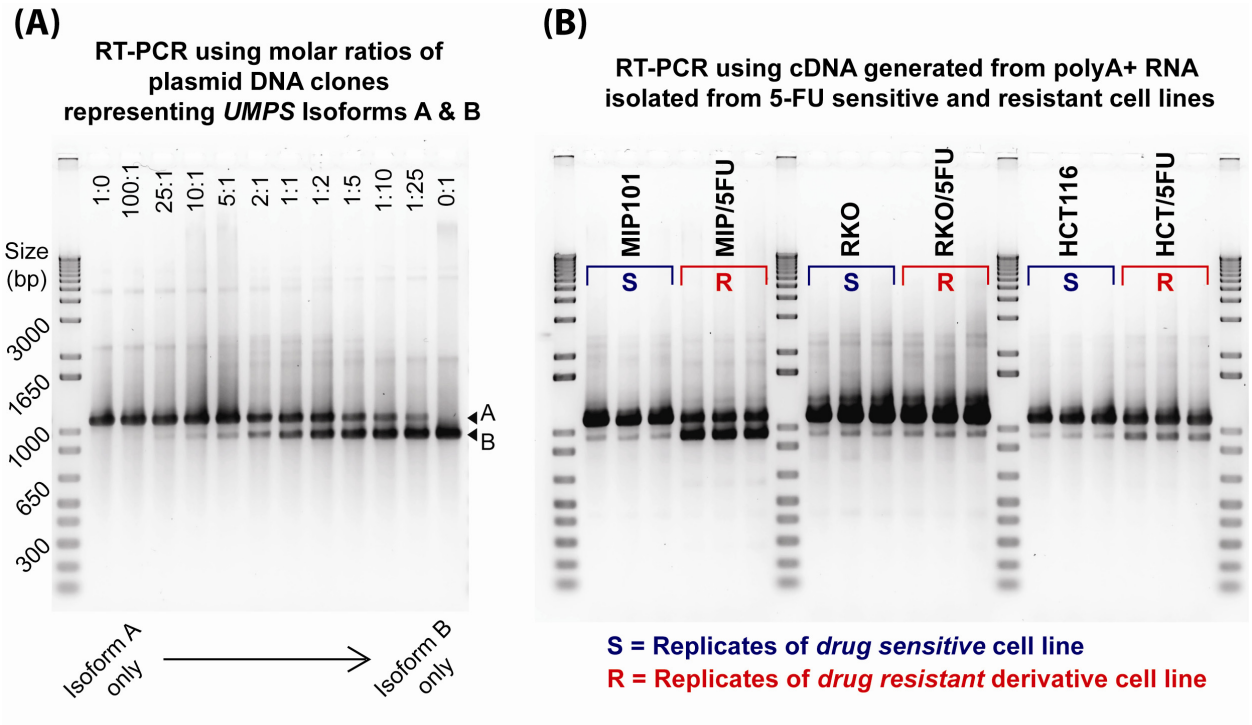


Figure 4.5. Full-ORF cloning of alternative *UMPS* isoforms
Gel electrophoresis of *UMPS* RT-PCR products generated from six cell lines was performed using a 1.5% agarose gel (**Methods**). 21 distinct PCR bands were generated by RT-PCR using polyA+ RNA from six 5-FU sensitive and resistant colorectal cancer cell lines. Each band was gel purified and used as input for TOPO-TA cloning reactions (**Methods**). Bands isolated by gel purification are indicated with colored triangles. Green triangles indicate bands at the expected size for *UMPS* isoform A. Blue triangles indicate additional bands that were successfully cloned and yielded at least one sequenced clone. Orange triangles indicate bands that did not yield sufficient DNA for cloning. The –ve control lane shows the result of RT-PCR with H_2O instead of cDNA template.

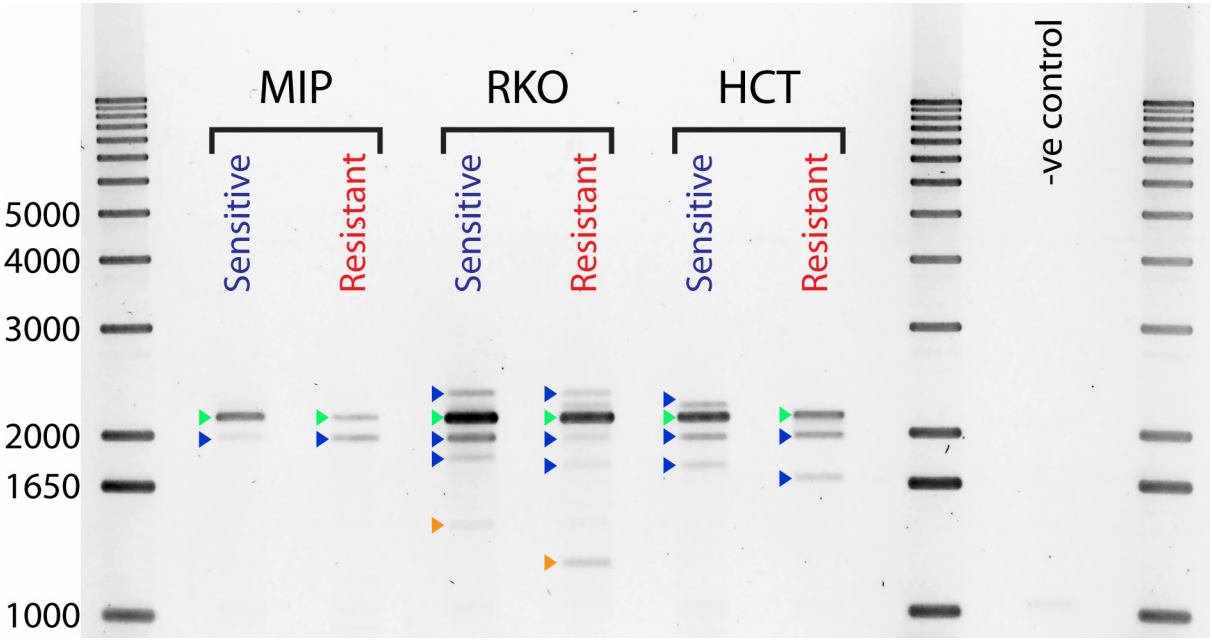


Figure 4.6. Sequencing of 96 clones isolated from distinct PCR bands
Ten unique transcript structures (A-J) observed by sequencing of 96 *UMPS* cDNA clones are depicted (see **Table 3** for further details). Rectangles indicate the boundaries of exons. Introns are indicated as black lines. The predicted coding region of each putative *UMPS* isoform is indicated in green. Unless otherwise indicated, all splice sites were canonical donor and acceptor sites (GT … AG). The position of two known enzymatic domains of *UMPS* (OPRTase and ODCase) is depicted at the bottom of the panel in purple and orange respectively. 'Isoform A' and 'Isoform B' in this figure correspond to isoforms A and B described throughout the text.

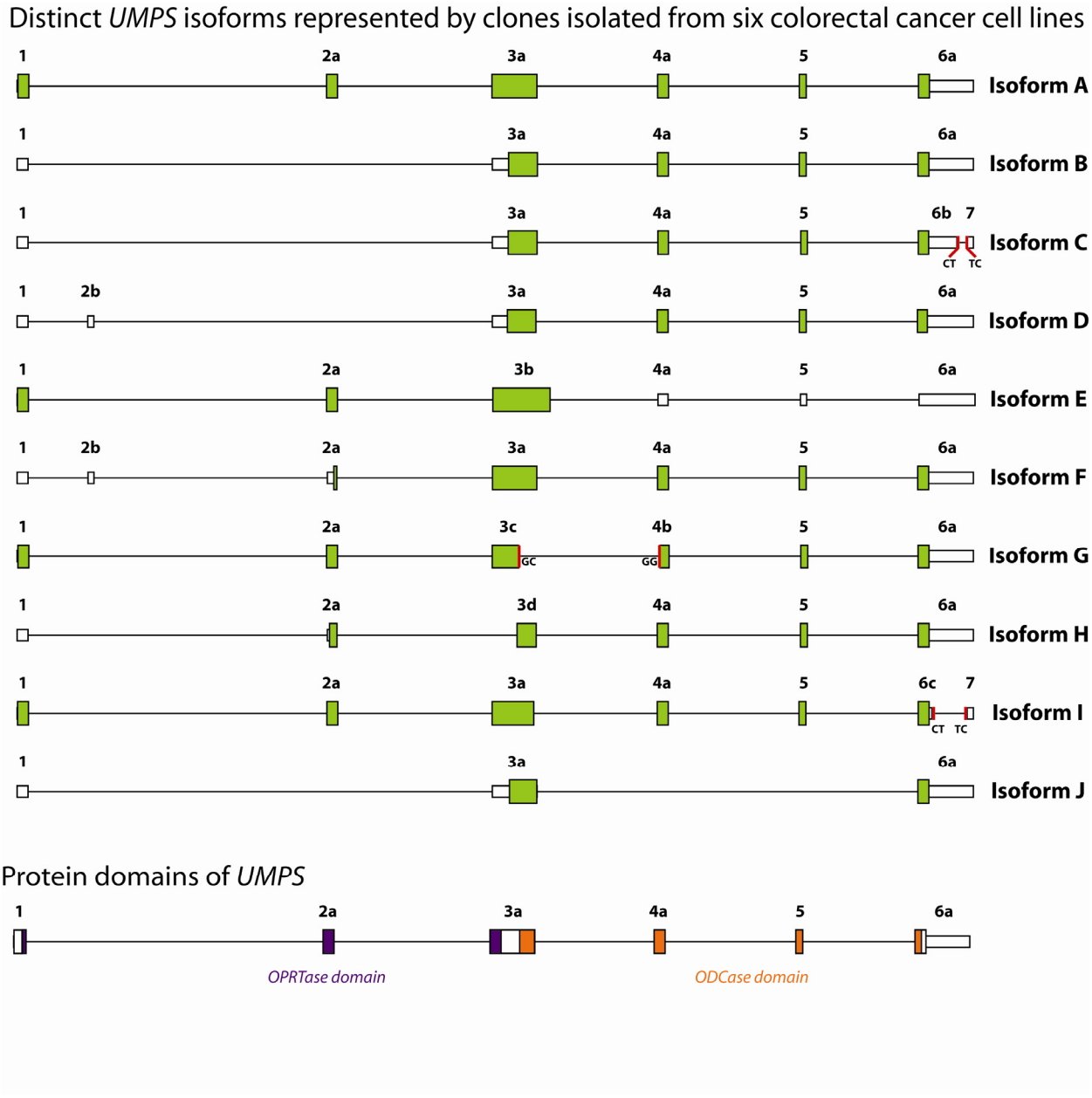


Figure 4.7. Western detection of *UMPS* protein in six 5-FU sensitive or resistant colorectal cancer cell lines

(**A**) Western analysis using an *UMPS* monoclonal antibody raised against partial *UMPS*. The expected size of full length *UMPS* (isoform A) is ~52 kDa and ~33 kDA for isoform B. (**B**) Quantitative western analysis of the MIP101 and MIP/5FU 52 KDa bands. Intensity values are reported beneath the *UMPS* bands. These were determined as previously reported[61] using the Odyssey software (**Methods**). Intensity values were normalized to *Actin* and multiplied by 100. (**C**) The position of amino acids comprising the epitope used to raise the antibody are indicated beneath a diagram of the exons and introns of *UMPS*. The antibody (AbNova Cat. No. H00007372-M06) was raised against a partial recombinant protein of *UMPS* consisting of the carboxy terminal end of the protein (amino acids 381-480 of 480). These amino acids correspond to the end of exon 4 and the complete coding sequence of exons 5 and 6.

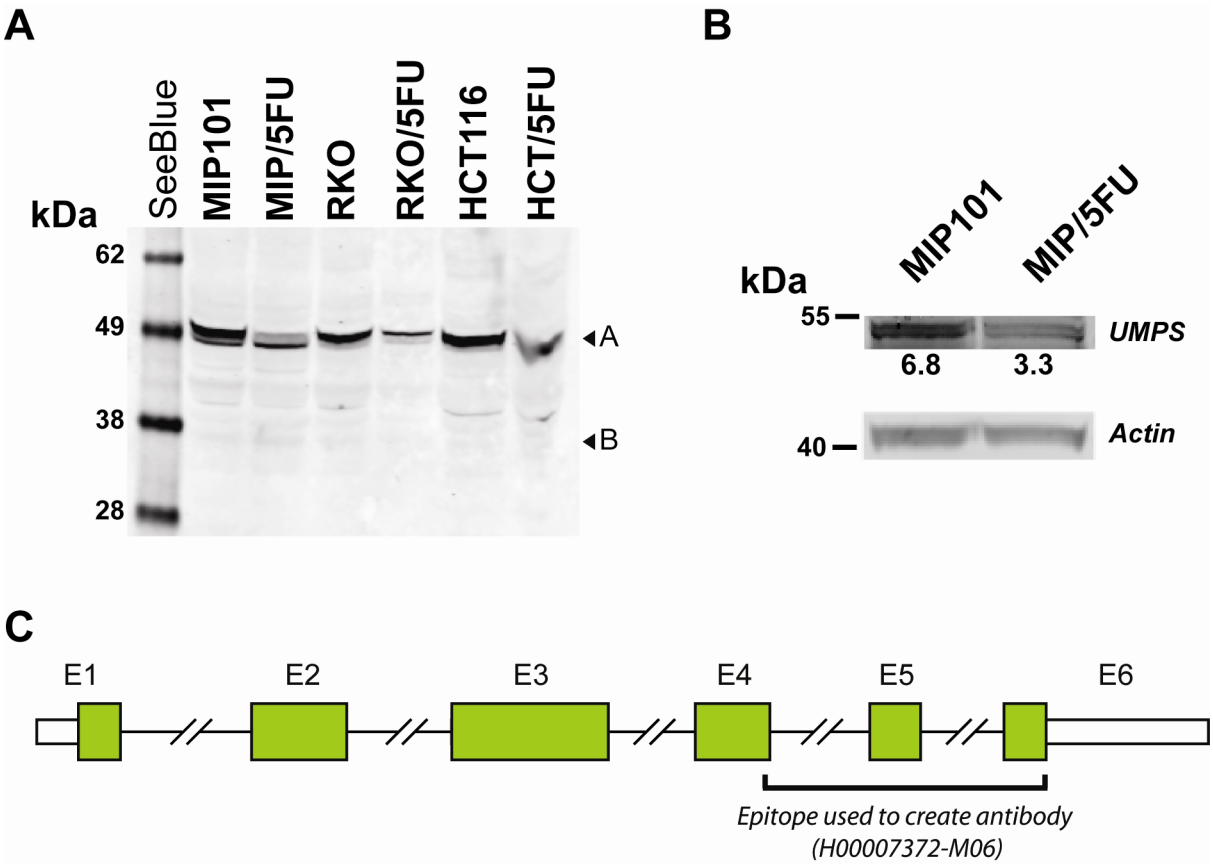
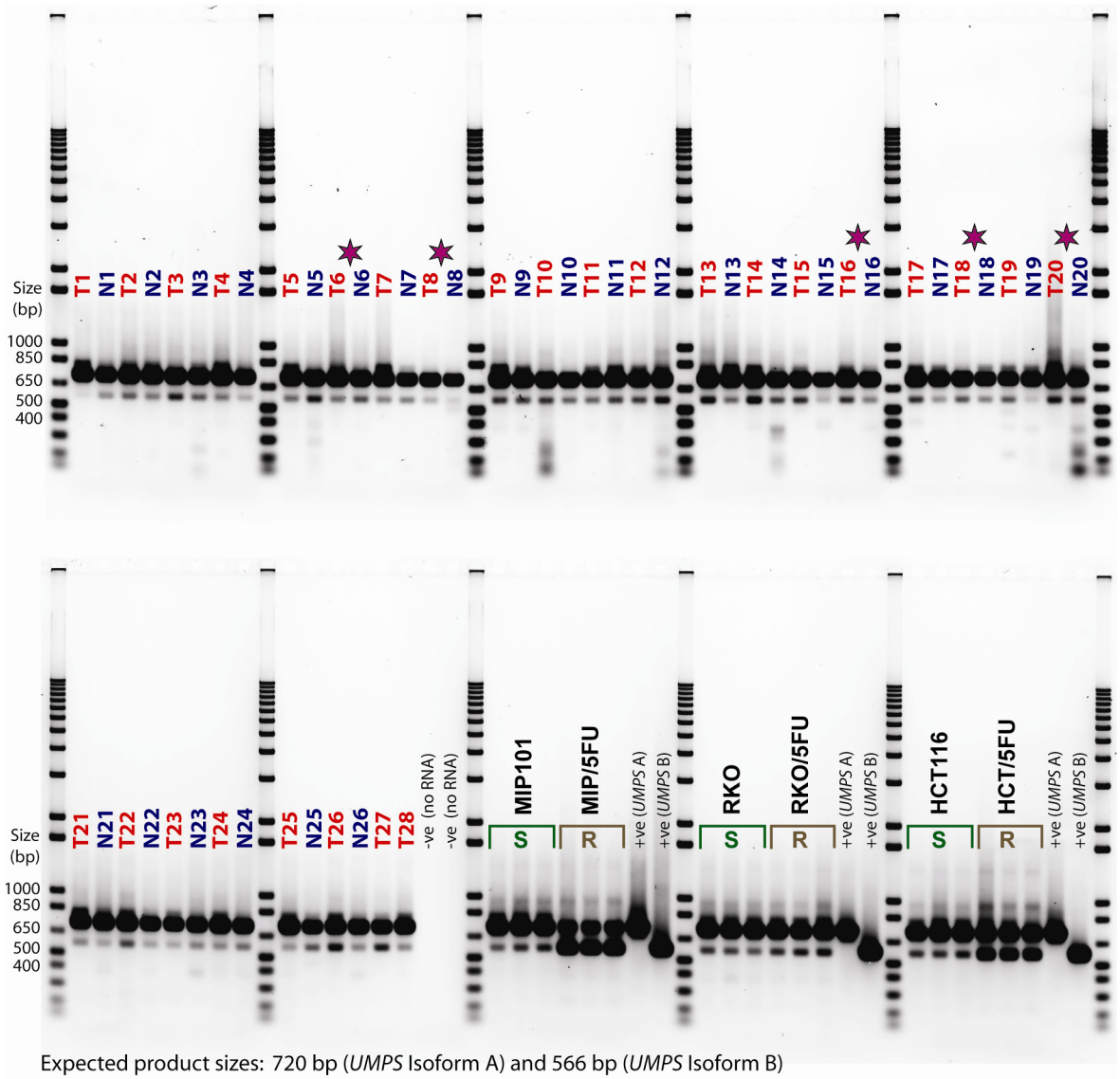


Figure 4.8. RT-PCR detection of *UMPS* isoforms A and B in a cohort of fresh frozen colorectal cancer tumours

The expected product sizes for this PCR reaction are 720 bp for isoform A and 566 bp for isoform B (missing exon 2). The PCR assay was applied to ss-cDNA generated from total RNA extracted from primary colorectal cancer tumour/normal sample pairs obtained from the Ontario Tumour Bank (**Methods**). Total RNAs from the same sensitive (S) and resistant (R) cell lines depicted in **Figure 4.4** were included in the assay as well as plasmid DNA corresponding to clones of *UMPS* isoforms A and B. In all patient samples, *UMPS* isoform A appeared to be the dominant form and isoform B a minor variant. It should be noted that most of these samples represent surgical resections prior to adjuvant treatment with 5-FU (i.e. they are treatment naïve). The five post-treatment cases are indicated with a magenta star.



Expected product sizes: 720 bp (*UMPS* Isoform A) and 566 bp (*UMPS* Isoform B)
S = Replicates of a 5-FU sensitive cell line
R = Replicates of a 5-FU resistant cell line

Figure 4.9. Real-time quantitative RT-PCR of *UMPS* isoform A and B expression in a cohort of colorectal cancer tumours

The expression of *UMPS* isoform A and B was determined by quantitative real-time PCR for treatment naïve primary colorectal tumours and matched adjacent normal tissue in a panel of 26 colorectal cancer cases. The expression level (delta Ct) of each isoform was calculated by normalizing the cycle threshold (Ct) reported by quantitative real time PCR to the Ct of an endogenous control gene (TBP). The inverse of these values (1/delta Ct) were plotted as box plots for both tumour and normal tissue. The statistical significance of the difference in *UMPS* isoform expression between the tumour and normal sample sets was tested using a Student's t-test (two-sided, no assumption of equal variances, $\alpha = 0.05$).

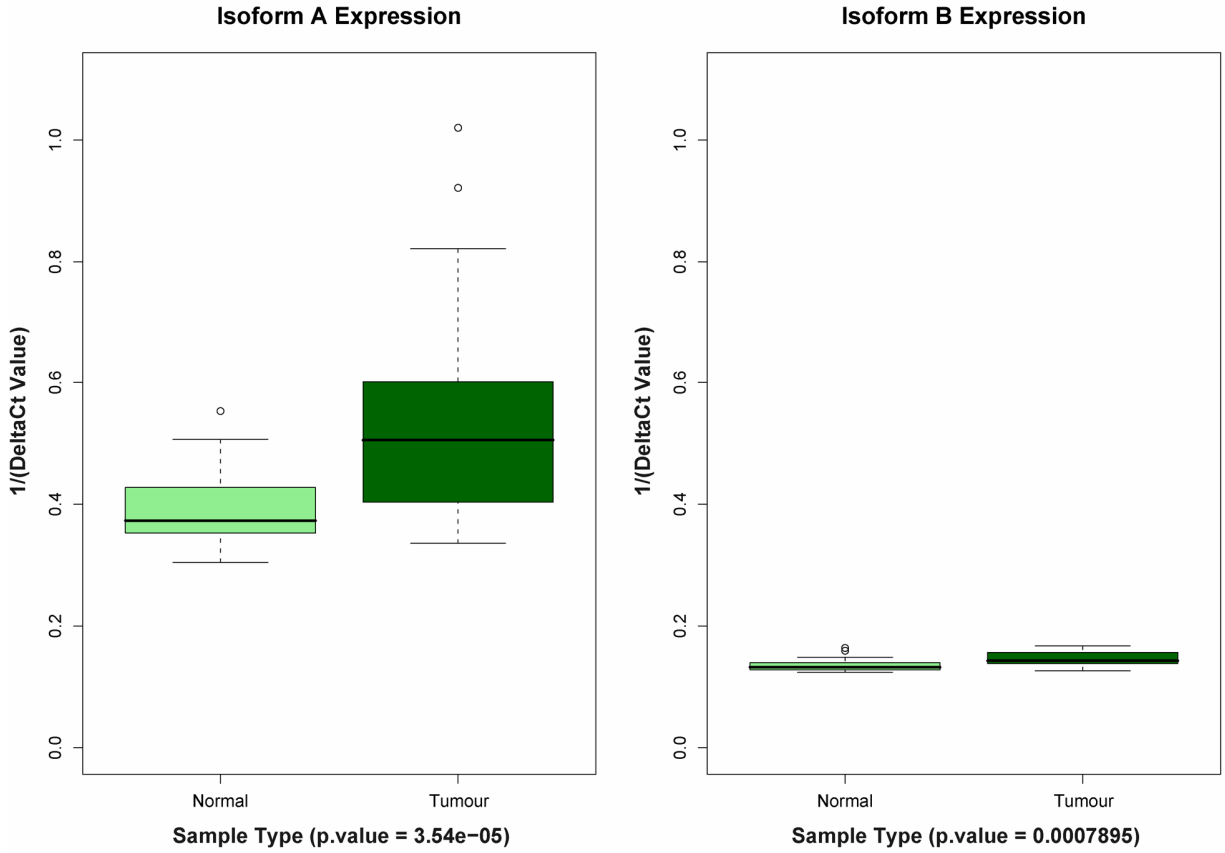


Figure 4.10. Overview of SNPs and mutations found by genomic sequencing
In the following figure, the position of amplicons used for sequencing reactions as well as putative novel mutations are depicted in relation to the coordinates of *UMPS* exons (black boxes) and introns (connecting lines). The position of each amplicon is depicted as a grey rectangle (see **Appendix B** for primer sequences). Note that the coordinates displayed are relative to a fragment of *UMPS* genomic DNA with 1 kb of flank added (see **Table 4.4** for corresponding chromosome coordinates). Intronic mutations are indicated in magenta, splice site mutations in red and protein coding mutations in blue. Additional details on each putative mutation are provided in **Table 4.4**. (**A**) 22 PCR amplicons (labeled A1 to A22) were generated from genomic DNA isolated from 6 cell lines (MIP101, MIP/5FU, RKO, RKO/5FU, HCT116, HCT/5FU) and 1 reference sample (**Methods**). Sequencing of the genomic region of *UMPS* in these cell lines revealed a heterozygous splice site mutation (position 5,727 G/G→G/T; indicated in red) at the splice acceptor site of exon 2 which was present in MIP/5FU (5-FU resistant) but not in MIP101 (sensitive) cells. This splice site mutation, acquired in the resistant cells is predicted to prevent pre-mRNA splicing of exon 2 and favor production of isoform B from the mutated allele. 4 additional mutations were observed (discussed in the text). (**B**) 22 PCR amplicons (labeled A23 to A44) were generated from genomic DNA isolated from 116 colorectal cancer patient samples (91 cases) (**Methods**). 12 putative mutations were identified, of which, 5 were intronic, 1 was close the same acceptor splice site mutation noted in panel A, and 6 were mutations within the open reading frame (discussed in the text).

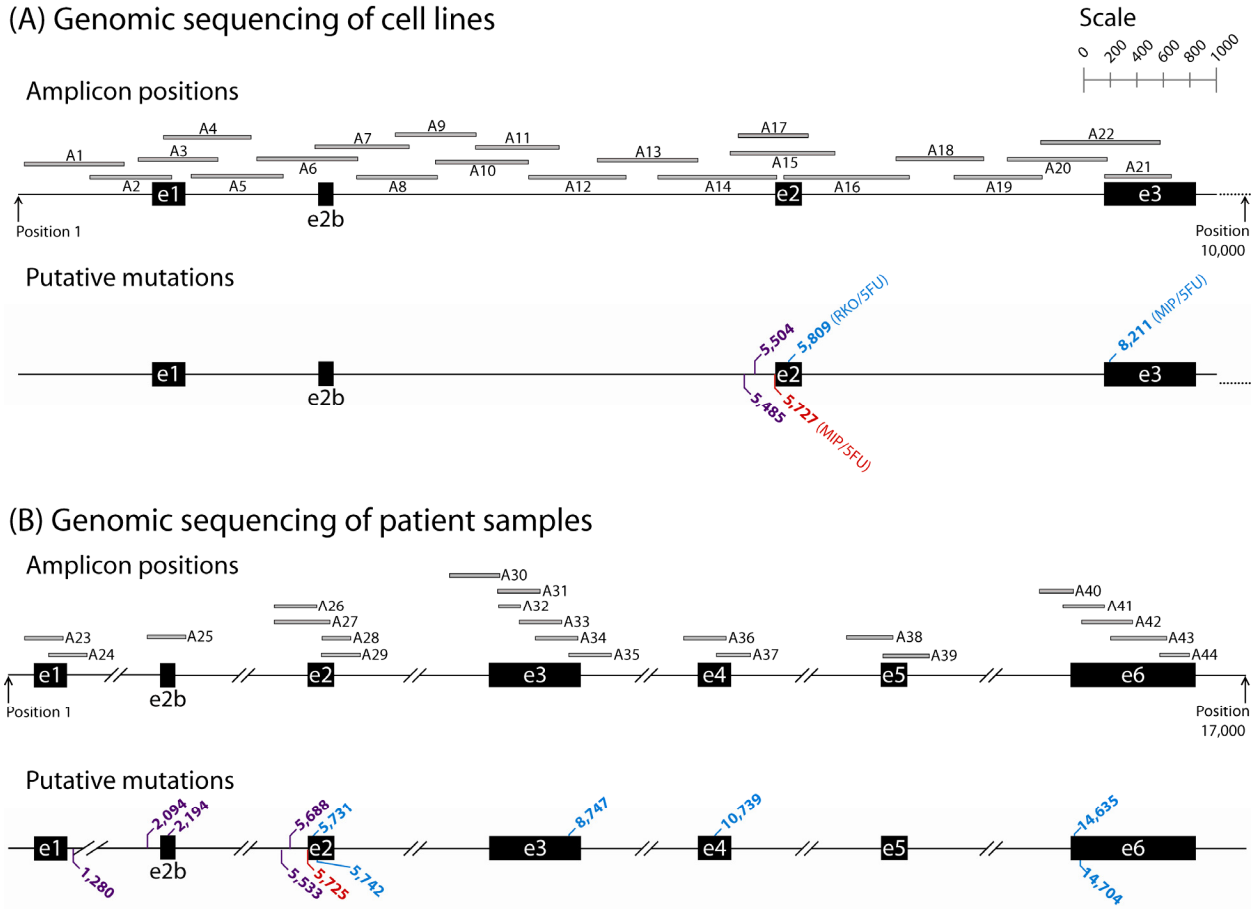


Table 4.1. Quantification of *UMPS* isoform A and B and the A/B ratio determined using four gene expression platforms

The expression of *UMPS* isoforms was determined by four platforms: custom NimbleGen splicing microarrays ('ALEXA'), whole transcriptome shotgun sequencing on an Illumina GA2 sequencer ('WTSS'), semi-quantitative RT-PCR using an Agilent 2100 DNA 7500 assay ('Agilent 2100') and quantitative real-time RT-PCR ('AB 7900HT'). Details of each of these platforms are provided in the **Methods** section. The following log2 expression and isoform A/B ratios were generated by each of these platforms.

Cell Line	Platform	Isoform A (log2)	Isoform B (log2)	Isoform A/B ratio
MIP101 (sensitive)	ALEXA	13.53 ± 0.23	8.85 ± 0.67	25.62 ± 0.56
	WTSS	6.80 ± n/a	1.11 ± n/a	51.38 ± n/a
	Agilent 2100	4.65 ± 0.30	-0.06 ± 0.24	27.36 ± 9.94
	AB 7900HT	5.99 ± 3.37	1.76 ± -0.21	19.22 ± 2.09
MIP/5FU	ALEXA	11.15 ± 0.39	11.38 ± 0.24	0.85 ± 0.13
	WTSS	4.43 ± n/a	4.81 ± n/a	0.77 ± n/a
	Agilent 2100	2.42 ± 0.19	4.49 ± 0.17	0.24 ± 0.01
	AB 7900HT	3.98 ± 0.73	4.93 ± 1.30	0.47 ± 0.18
RKO (sensitive)	Agilent 2100	5.16 ± 0.24	0	35.97 ± 5.75
	AB 7900HT	5.83 ± 2.44	1.40 ± -1.74	21.49 ± 1.34
RKO/5FU (resistant)	Agilent 2100	5.35 ± 0.24	0.42 ± 0.37	31.28 ± 8.99
	AB 7900HT	5.63 ± 2.88	1.87 ± -1.26	13.50 ± 1.18
HCT116 (sensitive)	Agilent 2100	5.26 ± 0.25	0.27 ± 0.36	31.91± 2.47
	AB 7900HT	5.15 ± 1.55	1.07 ± -1.71	17.25 ± 3.08
HCT/5FU (resistant)	Agilent 2100	4.96 ± 0.04	2.92 ± 0.15	4.11 ± 0.35
	AB 7900HT	4.99 ± 1.95	3.25 ± -0.49	3.34 ± 0.41

Table 4.2. Differential expression values for *UMPS* isoform A and B
Log2 differential expression values were derived from each of the gene expression platforms described in the text. Statistical significance of differential expression between sensitive and resistant cells was tested by a Student's t-test (two-sided, no assumption of equal variances, $\alpha = 0.05$). '+/-' values indicate the standard deviation (where replicates were available). The 'ALEXA' platform refers to a custom microarray design synthesized by NimbleGen, 'WTSS' refers to whole transcriptome shotgun sequencing data generated using an Illumina GAII sequencing device, Agilent 2100 refers to analysis of RT-PCR products by quantitative capillary electrophoresis with an Agilent 2100 'lab-on-a-chip' assay, and 'AB 7900 HT' refers to a quantitative real-time RT-PCR assay (see **Methods** for further details on each platform).

Cell Line Comparison	Platform	Isoform A fold change	p-value	Isoform B fold change	p-value
MIP101 (sensitive) versus MIP/5FU (resistant)	ALEXA	-5.21 ± 1.27	2.20×10^{-16}	5.77 ± 0.02	1.07×10^{-4}
	WTSS	-5.17 ± n/a	n/a	12.95 ± n/a	n/a
	Agilent 2100	-4.72 ± 1.40	2.05×10^{-2}	23.45 ± 1.05	4.37×10^{-3}
	AB 7900HT	-3.92 ± 1.10	4.63×10^{-7}	9.20 ± 1.20	3.20×10^{-11}
RKO (sensitive) versus RKO/5FU (resistant)	Agilent 2100	1.14 ± 1.23	0.39	1.34 ± 1.29	0.19
	AB 7900HT	-1.15 ± 1.13	3.11×10^{-2}	1.38 ± 1.11	3.18×10^{-5}
HCT116 (sensitive) versus HCT/5FU (resistant)	Agilent 2100	-1.24 ± 1.20	0.19	6.27 ± 1.40	1.93×10^{-3}
	AB 7900HT	-1.13 ± 1.19	2.97×10^{-2}	4.56 ± 1.16	9.34×10^{-6}

Table 4.3. Summary of alternative isoforms observed as clones
Distinct *UMPS* isoforms were named isoform 'A' to 'J'. A fully sequenced clone representing each isoform was submitted to GenBank (identifiers provided in the first column). The differences present in isoforms B-J relative to the *UMPS* Refseq record (NM_000373) are indicated as well as the number of times each was observed, whether it was supported by existing EST/mRNA sequence data and the effect on the ORF compared to the Refseq ORF. See **Figure 4.6** for a graphical depiction of each isoform.

Isoform (GenBank ID) [WTSS support?]	**Structural differences relative to NM_000373**	**Number of times observed**	**mRNA/EST support**	**Effect on ORF**	**Comments**
Isoform A (EU921886) [WTSS supported]	6 exons. 2136 bp. Identical to NM_000373 within the boundaries of our primers.	37 total 6 MIP101, 5 MIP/5FU, 7 RKO, 7 RKO/5FU, 6 HCT116, 6 HCT/5FU	Multiple mRNAs and ESTs for human and several other species	Contains an OPRTase domain of 123 aa, an ODCase domain of 214 aa with 11 known active sites, and a dimer interface with 18 conserved residues.	The 'canonical' isoform.
Isoform B (EU921887) [WTSS supported]	5 exons. 1982 bp. Skip of exon 2 (154 bases).	30 total 6 MIP101, 6 MIP/5FU, 6 RKO, 6 RKO/5FU, 6 HCT/5FU	At least 8 human mRNAs and ESTs as well as mouse mRNAs and ESTs.	62.9% identity to *UMPS* A. 11 of 11 ODCase active sites and 18 of 18 dimer sites are conserved.	Second most highly expressed and sequence supported alternative isoform
Isoform C (EU921888) [WTSS supported]	6 exons. 1847 bp. Skip of exon2 as well as a 135 base deletion within exon 6.	9 total 6 MIP101, 3 HCT116	None	62.9% identity to *UMPS* A. 11 of 11 ODCase active sites and 18 of 18 dimer sites are conserved.	
Isoform D (EU921889) [No WTSS support]	6 exons. 2081 bp. Use of an alternative exon 2b (of 99 bases) instead of the exon 2 (of 154 bases) used by other isoforms.	1 total 1 MIP/5FU	None	62.9% identity to *UMPS* A. 11 of 11 ODCase active sites and 18 of 18 dimer sites are conserved.	Alternative exon 2b is not well conserved (only to Rhesus)
Isoform E (EU921890) [No WTSS support]	6 exons. 2340 bp. Exon 3 uses an alternate splice donor site downstream of that normally used, resulting in an exon 3 that is 876 instead of 672 bases.	3 total 2 RKO 1 RKO/5FU	One human mRNA and one human EST	Predicted target of Nonsense mediated decay.	

Isoform (GenBank ID) [WTSS support?]	**Structural differences relative to NM_000373**	**Number of times observed**	**mRNA/EST support**	**Effect on ORF**	**Comments**
Isoform F (EU921891) [WTSS supported]	7 exons. 2235 bp. Contains both exon 2 (of 154 bases) and an alternative exon 2b (of 99 bases).	4 total 3 RKO 1 RKO/5FU	Four human ESTs	80.8% identity to *UMPS* A. 11 of 11 ODCase active sites and 18 of 18 dimer sites are conserved.	Alternative exon 2b is not well conserved (only to Rhesus)
Isoform G (EU921892) [WTSS supported]	6 exons, 1818 bp. Exon 3 uses an alternate splice donor site resulting in an exon 3 that is 380 bases instead of 672. Exon 4 also uses an alternate acceptor site changing its size from 176 bases to 150.	3 total 3 RKO	None. Limited EST coverage for this region.	78.5% identity to *UMPS* A. 5 of 11 ODCase active sites and 7 of 18 dimer sites are conserved.	
Isoform H (EU921893) [No WTSS support]	6 exons. 1789 bp. Exon 3 uses an alternate splice acceptor site downstream of that normally used, resulting in an exon 3 that is 325 bases instead of 672 bases.	3 total 3 RKO	None. Limited EST coverage for this region.	54.8% identity to *UMPS* A. 11 of 11 ODCase active sites and 18 of 18 dimer sites are conserved.	
Isoform I (EU921894) [WTSS supported]	7 exons. 1624 bp. Contains a 512 base deletion within the boundaries of exon 6.	3 total 3 RKO	None.	100.0% identity to *UMPS* A.	
Isoform J (EU921895) [WTSS supported]	3 exons. 1691 bp. Skipping of exon 2 (154 bases), exon 4 (176) and exon 5 (115).	2 total 2 HCT/5FU	None.	43.1% identity to *UMPS* A. 6 of 11 ODCase active sites and 12 of 18 dimer sites are conserved.	

Table 4.4. Summary of putative mutations

18 putative mutations identified by sequencing the genomic DNA of colorectal cancer cell lines and patient samples are listed below. All putative mutations identified were single base substitutions. The positions relative to the *UMPS* locus (including 1,000 bp of flank) are listed in the first column. Coordinates relative to chromosome 3 are provided in the second column as well as the location with respect to *UMPS* exon and intron boundaries. The third column describes the mutation (G→G/A indicates a heterozygous mutation, G→A indicates a homozygous mutation). The BLOSUM62 substitution matrix score[40, 41] for each amino acid change is provided in square brackets. The fourth column describes the mutation type. The final column lists the sample(s) that each mutation was observed in, followed by the 5-FU status. '+' indicates that the sample was exposed to 5-FU prior to isolation. '-' indicates that the sample was treatment naïve. Sample names: 'T' for tumour tissue; 'N' for matched normal; 'M' for metastasis. Sample names end with the patient number.

Pos	Chromosome coordinates	Mutation	Mutation Type	Samples (5-FU status)
1,280	chr3:125,932,182 (Intron 1)	G→G/A	Intronic	T85 (+)
2,094	chr3:125,932,996 (Intron 1)	T→T/C	Intronic	T80 (+)
2,194	chr3:125,933,096 (Intron 1)	C→C/T	Intronic (exon 2b)	M67 (+)
5,445	chr3:125,936,347 (Intron 1)	A→A/C	Intronic	T89 (+)
5,485	chr3:125,936,387 (Intron 1)	C→C/T	Intronic	RKO (-), RKO/5FU (+)
5,504	chr3:125,936,406 (Intron 1)	A→A/G	Intronic	RKO (-), RKO/5FU (+)
5,533	chr3:125,936,435 (Intron 1)	G→A	Intronic	M43 (+)
5,688	chr3:125,936,590 (Intron 1)	G→A	Intronic	M43 (+)
5,725	chr3:125,936,627 (splice acceptor site of exon 2)	C→C/T	Splice site (-3 position)	T14 (-)
5,727	chr3:125,936,629 (splice acceptor site of exon 2)	G→G/T	Splice site (-1 position)	MIP/5FU (+)
5,731	chr3:125,936,633 (Exon 2)	G→G/A (A54T) [0]	Missense	T14 (-)
5,742	chr3:125,936,644 (Exon 2)	A→C/T (L57F) [0]	Missense	M57 (+)
5,809	chr3:125,936,711 (Exon 2)	C→C/T (P78S) [-1]	Missense	RKO/5FU (+)
8,211	chr3:125,939,113 (Exon 3)	C→C/T (R107C) [-3]	Missense	MIP/5FU (+)
8,747	chr3:125,939,649 (Exon 3)	T→T/A (D285Q) [0]	Missense	T88 (+)
10,739	chr3:125,941,641 (Exon 4)	C→C/G (Q353E) [2]	Missense	T84 (+)
14,635	chr3:125,945,537 (Exon 6)	A→A/G (I453M) [1]	Missense	N12 (-), T12 (-)
14,704	chr3:125,945,606 (Exon 6)	T→T/C (S476S) [4]	Silent	M66 (+)

References

1. Rich, T. A., Shepard, R. C. & Mosley, S. T. Four decades of continuing innovation with fluorouracil: current and future approaches to fluorouracil chemoradiation therapy. J Clin Oncol 22, 2214-32 (2004).
2. Meisner, N. C. et al. The chemical hunt for the identification of drugable targets. Curr Opin Chem Biol 8, 424-31 (2004).
3. Bleiberg, H., Kemeny, N., Rougier, P. & Wilke, H. Colorectal Cancer: A Clinical Guide to Therapy (Martin Dunitz Ltd., London, 2002).
4. Heidelberger, C. et al. Fluorinated pyrimidines, a new class of tumour-inhibitory compounds. Nature 179, 663-6 (1957).
5. Peters, G. J., van Groeningen, C. J., Laurensse, E. J. & Pinedo, H. M. A comparison of 5-fluorouracil metabolism in human colorectal cancer and colon mucosa. Cancer 68, 1903-9 (1991).
6. Peters, G. J. et al. Induction of thymidylate synthase as a 5-fluorouracil resistance mechanism. Biochim Biophys Acta 1587, 194-205 (2002).
7. Longley, D. B., Harkin, D. P. & Johnston, P. G. 5-fluorouracil: mechanisms of action and clinical strategies. Nat Rev Cancer 3, 330-8 (2003).
8. Pinedo, H. M. & Peters, G. F. Fluorouracil: biochemistry and pharmacology. J Clin Oncol 6, 1653-64 (1988).
9. Cheetham, S. et al. SPARC promoter hypermethylation in colorectal cancers can be reversed by 5-Aza-2'deoxycytidine to increase SPARC expression and improve therapy response. Br J Cancer 98, 1810-9 (2008).
10. Tang, M. J. & Tai, I. T. A novel interaction between procaspase 8 and SPARC enhances apoptosis and potentiates chemotherapy sensitivity in colorectal cancers. J Biol Chem 282, 34457-67 (2007).
11. Oguri, T. et al. MRP8/ABCC11 directly confers resistance to 5-fluorouracil. Mol Cancer Ther 6, 122-7 (2007).
12. An, Q., Robins, P., Lindahl, T. & Barnes, D. E. 5-Fluorouracil incorporated into DNA is excised by the Smug1 DNA glycosylase to reduce drug cytotoxicity. Cancer Res 67, 940-5 (2007).
13. Luqmani, Y. A. Mechanisms of drug resistance in cancer chemotherapy. Med Princ Pract 14 Suppl 1, 35-48 (2005).
14. Klein, T. E. et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1, 167-70 (2001).
15. www.myriad.com.
16. www.dnavision.be.
17. www.genpathdiagnostics.com.
18. Allegra, C. J. Dihydropyrimidine dehydrogenase activity: prognostic partner of 5-fluorouracil? Clin Cancer Res 5, 1947-9 (1999).
19. Han, R. et al. Systemic 5-fluorouracil treatment causes a syndrome of delayed myelin destruction in the central nervous system. J Biol 7, 12 (2008).
20. Maring, J. G., Groen, H. J., Wachters, F. M., Uges, D. R. & de Vries, E. G. Genetic factors influencing pyrimidine-antagonist chemotherapy. Pharmacogenomics J 5, 226-43 (2005).
21. Tokunaga, Y., Sasaki, H. & Saito, T. Clinical role of orotate phosphoribosyl transferase and dihydropyrimidine dehydrogenase in colorectal cancer treated with postoperative fluoropyrimidine. Surgery 141, 346-53 (2007).

22. Sakamoto, E. et al. Orotate phosphoribosyltransferase expression level in tumors is a potential determinant of the efficacy of 5-fluorouracil. Biochem Biophys Res Commun 363, 216-22 (2007).
23. Ichikawa, W. et al. Simple combinations of 5-FU pathway genes predict the outcome of metastatic gastric cancer patients treated by S-1. Int J Cancer 119, 1927-33 (2006).
24. Griffith, M. & Marra, M. A. in Genes, Genomes & Genomics (eds. Thangadurai, D., Tang, W. & Pullaiah, T.) 201-242 (Regency Publications, New Delhi, 2007).
25. Griffith, M. et al. ALEXA: a microarray design platform for alternative expression analysis. Nat Methods 5, 118 (2008).
26. Morin, R. et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45, 81-94 (2008).
27. Tai, I. T., Dai, M., Owen, D. A. & Chen, L. B. Genome-wide expression analysis of therapy-resistant tumors reveals SPARC as a novel target for cancer therapy. J Clin Invest 115, 1492-502 (2005).
28. Kuhn, R. M. et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37, D755-61 (2009).
29. Hubbard, T. et al. Ensembl 2005. Nucleic Acids Res 33, D447-53 (2005).
30. Jenuth, J. P. The NCBI. Publicly available tools and resources on the Web. Methods Mol Biol 132, 301-12 (2000).
31. Wilming, L. G. et al. The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36, D753-60 (2008).
32. Han, A., Kim, W. Y. & Park, S. M. SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23, 397-9 (2007).
33. Wang, M. & Marin, A. Characterization and prediction of alternative splice sites. Gene 366, 219-27 (2006).
34. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29, 1185-90 (2001).
35. Hebsgaard, S. M. et al. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24, 3439-52 (1996).
36. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. J Comput Biol 4, 311-23 (1997).
37. Marchler-Bauer, A. et al. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35, D237-40 (2007).
38. Langdon, S. D. & Jones, M. E. Study of the kinetic and physical properties of the orotidine-5'-monophosphate decarboxylase domain from mouse UMP synthase produced in Saccharomyces cerevisiae. J Biol Chem 262, 13359-65 (1987).
39. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29, 308-11 (2001).
40. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22, 1035-6 (2004).
41. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9 (1992).
42. Ichikawa, W. et al. Both gene expression for orotate phosphoribosyltransferase and its ratio to dihydropyrimidine dehydrogenase influence outcome following fluoropyrimidine-based chemotherapy for metastatic colorectal cancer. Br J Cancer 89, 1486-92 (2003).

43. Peters, G. J. et al. Enhanced therapeutic efficacy of 5'deoxy-5-fluorouridine in 5-fluorouracil resistant head and neck tumours in relation to 5-fluorouracil metabolising enzymes. Br J Cancer 59, 327-34 (1989).
44. Chung, Y. M. et al. Establishment and characterization of 5-fluorouracil-resistant gastric cancer cells. Cancer Lett 159, 95-101 (2000).
45. Matsuyama, R. et al. Predicting 5-fluorouracil chemosensitivity of liver metastases from colorectal cancer using primary tumor specimens: three-gene expression model predicts clinical response. Int J Cancer 119, 406-13 (2006).
46. Ando, T. et al. Relationship between expression of 5-fluorouracil metabolic enzymes and 5-fluorouracil sensitivity in esophageal carcinoma cell lines. Dis Esophagus 21, 15-20 (2008).
47. Harada, K., Kawashima, Y., Yoshida, H. & Sato, M. Thymidylate synthase expression in oral squamous cell carcinoma predicts response to S-1. Oncol Rep 15, 1417-23 (2006).
48. Ijuin, T. et al. Thymidine phosphorylase mRNA level predicts survival of patients with advanced oropharyngeal cancer. Acta Otolaryngol 127, 305-11 (2007).
49. Kodera, Y. et al. Gene expression of 5-fluorouracil metabolic enzymes in primary gastric cancer: Correlation with drug sensitivity against 5-fluorouracil. Cancer Lett 252, 307-13 (2007).
50. Suchi, M. et al. Molecular cloning of the human UMP synthase gene and characterization of point mutations in two hereditary orotic aciduria families. Am J Hum Genet 60, 525-39 (1997).
51. Robinson, J. L., Dombrowski, D. B., Harpestad, G. W. & Shanks, R. D. Detection and prevalence of UMP synthase deficiency among dairy cattle. J Hered 75, 277-80 (1984).
52. Shanks, R. D., Dombrowski, D. B., Harpestad, G. W. & Robinson, J. L. Inheritance of UMP synthase in dairy cattle. J Hered 75, 337-40 (1984).
53. van Ommen, G. J., van Deutekom, J. & Aartsma-Rus, A. The therapeutic potential of antisense-mediated exon skipping. Curr Opin Mol Ther 10, 140-9 (2008).
54. Kidd, E. A. et al. Variance in the expression of 5-Fluorouracil pathway genes in colorectal cancer. Clin Cancer Res 11, 2612-9 (2005).
55. Niles, R. M. et al. Isolation and characterization of an undifferentiated human colon carcinoma cell line (MIP-101). Cancer Invest 5, 545-52 (1987).
56. Baross, A. et al. Systematic recovery and analysis of full-ORF human cDNA clones. Genome Res 14, 2083-92 (2004).
57. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-94 (1998).
58. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175-85 (1998).
59. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res 28, 4364-75 (2000).
60. Pugh, T. J. et al. Correlations of EGFR mutations and increases in EGFR and HER2 copy number to gefitinib response in a retrospective analysis of lung cancer patients. BMC Cancer 7, 128 (2007).
61. Mathew, R. et al. Autophagy suppresses tumorigenesis through elimination of p62. Cell 137, 1062-75 (2009).

5. Conclusions

5.1. Summary

Since the completion of the first reference human genome sequences, the study of the transcriptome, particularly the identification, annotation and quantification of transcripts has remained an active area of research. Efforts to profile the expression of genes across tissues and developmental stages, identify the regulatory elements which control gene expression and characterize the genomic variations between individuals that influence these patterns have been widely reported. Identifying the transcripts expressed in a tissue and measuring differences in expression level between samples can provide valuable insight on the function of genes and may lead to the identification of disease markers and therapeutic targets. The phenomenon of alternative expression described in this thesis has profound implications for these efforts because it can result in the generation of multiple distinct transcripts from each gene locus. A major thrust of this thesis was to investigate the phenomenon of alternative expression, assess its role in generating multiple distinct transcripts from each gene locus and develop methods to characterize these transcripts. In **Chapter 1**, I provided a review of past and present methods used to study the expression of genes with particular emphasis on alternative expression and its implications for the study of the human diseases such as cancer. **Chapters 2** and **3** sought to elucidate the complexity of the transcriptome by the development of novel methods for the identification and quantification of mRNA isoforms. **Chapters 2** and **3** also described the application of these methods to a cell line model of chemotherapy resistance in colorectal cancer and **Chapter 4** described the characterization of a single candidate alternative expression event which emerged from these analyses.

In **Chapter 2**, I described a novel method of microarray design and analysis to allow microarray detection and quantification of both known and novel isoforms. The method proved capable of detecting the expression of known splicing events and predicting the expression of novel isoforms differentially expressed between 5-FU sensitive and resistant colorectal cancer cells. The sensitivity and specificity of the approach were found to be an improvement over Affymetrix exon arrays, the main commercially available microarray platform for alternative expression analysis. In addition, the method was able to identify the connectivity of adjacent exons, a capability not available

in exon tiling microarrays. Although our experiments with the method were largely successful, it remained insensitive to certain types of potentially important alternative expression events such as those involving small exons. Furthermore, the number of alternative expression events that could be simultaneously measured was limited by the number of oligonucleotide features that could be synthesized on a single microarray. Each experiment was also limited by the quality and completeness of transcriptome annotations available at the time each microarray was designed. In **Chapter 3**, I sought to address these limitations by developing a method that relied on recent technological advances in massively parallel sequencing. During the development of this method, several reports were published describing the utility of whole transcriptome shotgun sequencing (WTSS) for profiling the transcriptome. The reports of Wang et al [1] and Pan et al [2] in particular described the preliminary application of WTSS to the study of alternative mRNA isoforms. Based on these reports and the findings presented in **Chapter 3**, it seems that this technology provides a snapshot of the transcriptome that is unprecedented in its resolution and comprehensiveness. In particular it seems well suited to profiling alternative isoforms and offers a number of advantages over microarray approaches. Finally in **Chapter 4**, I described a series of experiments aimed at characterization of an alternative expression event identified in 5-FU resistant cells by both the microarray and sequencing approaches described in **Chapters 2** and **3**. A few studies have used genomics approaches to study 5-FU resistance in colorectal cancer[3, 4] but these studies lacked the ability to measure specific mRNA isoforms or mutation burden.

In the remainder of this chapter I will discuss the strengths, limitations, and significance of the research reported in **Chapters 2-4** and suggest areas for future research.

5.2. Strengths and limitations

The comparative advantages and disadvantages of microarray and sequencing approaches for alternative expression analysis were discussed in some detail in **Chapters 2** and **3**. In general, the primary advantage of these methods compared to previously used methods is that they are more comprehensive. They profile more completely the transcriptionally active regions of the genome and they have the ability to quantify isoforms with low expression or that represent a minor form in a mixture of two

or more isoforms. Some of the limitations and caveats of these methods are enumerated below. (**1**) The increasing amount and complexity of data produced by these methods require increased computational resources. Thus far, informatics and algorithmic developments have shown a remarkable ability to adapt to the changing data landscape. (**2**) Complete exon connectivity for each transcript remains hidden because of the focused nature of the measurements (i.e. short oligonucleotide probes or short sequence reads). The complete structure of an isoform can sometimes be inferred but secondary validations such as RT-PCR, Northern analysis, and cloning and sequencing are required to determine if the inference is correct. This limitation will be difficult to address with microarrays but with sequencing approaches, increased read lengths and use of read-pairing information may be able to at least partially overcome this limitation. (**3**) These methods still rely to some degree on existing gene and transcript annotations. New approaches that utilize *de novo* transcriptome assembly may largely eliminate this limitation in the near future[5, 6]. (**4**) Gene families that consist of groups of genes with high levels of sequence similarity will remain problematic for both microarray and sequence based approaches although longer reads and read-pairing in the sequencing platforms will help to address this challenge. (**5**) As with any use of transcriptome data, inferences regarding gene function based on observations at the transcript level may not be mirrored at the protein level. Since many genes function as proteins, this is a major limitation. Proteomic validation of findings made at the transcript level is therefore of paramount importance. Furthermore, many functionally significant changes related to disease may occur post-translationally (phosphorylation, protein folding, etc.) and while some of these may be predicted computationally, many of these events may be missed by transcriptome analysis. (**6**) The annotation of transcript functions has not kept up with the rate of transcript discovery and this may limit interpretation of the data. Even so, large scale transcriptome profiling experiments often can contribute to our knowledge of function. (**7**) The cost of performing splicing microarray and massively parallel RNA sequencing experiments is not yet low enough to allow use of these technologies to realize the goal of a personalized medicine strategy involving transcriptome profiling but this day is coming and recent experiments have shown the proof-of-principle (Jones et al, unpublished).

A major limitation of the research as an attempt to identify candidate genes involved in 5-FU resistance in colorectal cancer is its reliance on a cell line model which in many

ways is not an ideal representation of drug resistance as it occurs in a patient. For example, while colorectal cancer is a solid tumour, colorectal cancer cell lines are grown in a culture as a mono-layer. It is difficult to image how mechanisms relating to angiogenesis or the tumour micro-environment could be effectively studied in this model. Similarly, cell line models do not account for the role of the liver in mediating drug response. The liver is the primary site of catabolism of many chemotherapies including 5-FU[7]. In other words, cancer cells grown in culture are outside of the environment where drugs work in humans. Furthermore, even if a cell line grown in culture was perfectly analogous to clinical disease, it would still represent only a single instance of drug resistance. The study described in this thesis represents an example of how drug resistance may be studied by transcriptome analysis with microarrays and massively parallel sequencing and not a definitive characterization of the molecular mechanisms of 5-FU resistance. A more comprehensive study of multiple resistant lines derived from a single parental line as well as derived from many different parental lines would have a greater chance of informing on the general mechanisms of acquired 5-FU resistance. Further studies of new cell lines established from primary tumours, metastases and drug refractory tumours, and patient samples isolated from treatment resistant tumours arising after chemotherapy exposure should also be conducted to begin to elucidate the major mechanism(s) of 5-FU resistance. One advantage of this research is that is was not limited to a small handful of genes as many previous studies have been (reviewed in [8, 9]). This allowed us to detect potentially novel genes associated with 5-FU resistance and in fact many of the genes in our candidate lists have not been previously identified as potentially involved in 5-FU resistance. Furthermore, examination of both whole gene differential expression as well as differential expression of specific isoforms allowed us to identify novel candidate isoforms which would have otherwise been missed in typical gene expression analysis approaches. For example, the gene *UMPS* (**Chapter 4**) was ranked 255th in the differential gene expression list, but ranked 3rd in the list of novel differentially spliced isoforms reported in **Chapter 3** (and online at www.AlexaPlatform.org).

5.3. Current status, significance and contribution to field of study

The initial hypothesis of this thesis, that the degree of alternative splicing in the human transcriptome was largely under-represented in existing annotation resources such as

EnsEMBL, was verified by this work and others[1, 2, 10]. In a comparison of 5-FU sensitive and resistant cancer cell lines I identified thousands of previously un-annotated alternative expression events. Our hope that developments in both microarray and sequencing technology would allow quantification of specific alternative isoforms was also realized. Finally, the hypothesis that the transition from 5-FU sensitivity to resistance would be associated with differential expression of both entire genes as well as specific isoforms was also verified. Genes such as *H19* and *KRT20* were found to be highly under- and over-abundant while other genes such *UMPS*, *UCK2* and *LAMA3* exhibited over-abundance of specific alternative isoforms while the canonical isoform remained unchanged or was differentially expressed in the opposite direction. Overall the ranked list of top 5-FU candidates contained almost as many alternative expression events as differential expression events (**Chapter 3**).

In this report I described novel methods for assessing the expression, differential expression and alternative expression of known and predicted mRNA isoforms including metrics for identifying reciprocal expression of alternative isoforms. I described novel methods for determining whether sequence features are expressed above the level of noise emanating from random transcription and possible genomic DNA or heteronuclear RNA contamination for both microarray and RNA sequence data. I also made available to the public databases of sequence features and associated visualization tools tailored to the analysis of alternative isoforms by microarray and RNA sequence analysis along with information on the existence of mRNA or EST sequence support, conservation and the protein coding effect of known and hypothetical alternative expression events. Pre-computed microarray designs and sequence annotation databases for analysis of alternative expression using massively parallel sequencing were generated for ten species. I also made available candidate gene lists, source code and user manuals. To assist in the efficient dissemination and organization of these resources I created the website www.AlexaPlatform.org to supplement the manuscripts (see footnotes for each chapter). Finally, I reported the first application of alternative expression microarray analysis and massively parallel sequencing to a model of chemotherapy resistance in colorectal cancer. While this model system relates to a specific area of cancer biology research, I believe that this work represents a general model for comparison of relevant disease states such as cancer and normal tissue, pre- and post-treatment biopsies, and primary and metastatic tumours. To my knowledge this work was the first genome-wide

effort resulting in the identification of a candidate gene that is aberrantly spliced in 5-FU resistant colorectal cancer cells (*UMPS*). A recent study of the methotrexate metabolism gene, folylpolyglutamate synthetase (*FPGS*) found that acquired resistance to antifolates in leukemia cell lines was caused by aberrant splicing of the gene[11]. Additionally, aberrant splicing of the androgen receptor was reported as a possible cause of hormone refractory prostate cancer[12, 13]. The *UMPS* observation, as well as candidate gene and isoform lists which include genes with suspected involvement in 5-FU metabolism and drug efflux represent a substantial contribution to existing 5-FU resistance candidate lists[3, 4].

The methods I have described have considerable potential to advance studies of gene regulation, transcript processing, human disease and evolutionary biology. For example, analysis of alternative expression by the microarray (**Chapter 2**) and massively parallel RNA sequencing (**Chapter 3**) approaches that I described should assist in the identification of developmental and tissue specific alternative isoforms. This may in turn elucidate some of the mechanisms which control their spatial and temporal expression patterns. The microarray approach I described has since been used to address this area by comprehensive profiling of 48 human tissues and cell lines, resulting in the identification of regulatory sequences associated with alternatively spliced exons[14]. The methods I described should also facilitate improved disease classification. Recent studies have reported the use of candidate gene approaches to identify aberrant splicing of oncogenes and tumour suppressors in cancer[15-17]. The methods I described should help to accelerate this process and lead to the identification of novel biomarkers and therapeutic targets for cancer and other diseases.

5.4. Potential applications and future directions

Integration of gene expression and alternative expression analysis with mutation analysis of both the transcriptome and the whole genome will be instrumental in understanding the mechanisms of alternative expression. By coupling complete genome sequencing to transcriptome sequencing it may be possible to correlate many changes in exon expression and processing with nearby or distant polymorphisms and mutations that reside in splice sites and exonic and intronic splicing enhancers and silencers. Similarly, transcriptome analysis in combination with systematic disruption of members of the transcriptional and splicing machinery (by over-expression and/or

siRNA knockdown for example) will help to unravel the complex code that controls expression and alternative expression.

While the data described in **Chapters 2** and **3** has provided rich information on alternative expression, analysis of these data to identify point mutations, gene fusions, RNA edits and allele specific expression may also be important in identifying novel disease markers. Similarly, when comparing disease states, chromosome copy number, rearrangements, and epigenetic modifications should also be examined. Furthermore, the analysis approach I describe is primarily focused on a pair-wise comparison of samples. This is broadly applicable to comparisons like disease versus matched normal tissue, pre-treated versus post-treatment tissue, unaffected parent versus affected child, and so on. However, for certain efforts such as disease classification, studies of splicing regulation, and others, additional analysis approaches will need to be incorporated to deal with simultaneous analysis of isoform expression patterns across larger groups of samples. These types of approaches have been reported in the analysis of microarray gene expression data sets using statistical approaches such as ANOVA[18], linear and non-linear regression[19, 20], and clustering[14]. One area of future research will be the modification of these approaches to facilitate their application to massively parallel RNA sequencing datasets.

The kinds of analyses advocated in this thesis have potential to be incorporated into a new general approach to disease treatment in the future. Using colorectal cancer as an example, there are already a number of known molecular markers of potential significance for selecting chemotherapy regimes for individual patients. For example, patients that are deficient in the genes that monitor DNA damage are generally less responsive to 5-FU and other DNA damaging cytotoxic drugs[21]. In **Chapter 4**, I describe a candidate gene that when mutated, under-expressed, or aberrantly spliced may result in reduced 5-FU efficacy. Similarly, activating mutations in *KRAS* have been found to predict poor response to *EGFR* targeted antibody therapies[22]. In addition to markers of efficacy, polymorphisms and mutations in *DPYD* and *UGT1A1* are predictive of adverse responses to 5-FU and Irinotecan respectively[22]. These observations could one day form the basis for a battery of clinical tests that would assess mutation status, expression level, isoform composition and other features of individual patients. The output of these tests could serve as the input for an elaborate therapeutic decision tree. Thus, an oncologist's job might increasingly come to involve navigating such decision

trees, customized for cancer type and stage, to create a unique personalized treatment strategy for each patient. In a sense, cancer treatment already operates in this fashion to a limited degree but the current repertoire of molecular predictors and determinants of treatment outcome is likely still largely incomplete. As it becomes more complete, this approach may prove increasingly successful. It is possible that the current markers of 5-FU response represent only a small fraction of those that are relevant to individual patients. Given the number of mechanisms thought to contribute to chemotherapy resistance it is possible that dozens or even hundreds of genes may be involved. Resistance in a single individual may be mediated by one or several small changes drawn from a large set of possible genes. If true, this would place even more emphasis on a personalized approach to predicting resistance using data that is comprehensive.

One of the interesting future applications of this research is not just the potential to identify alternatively spliced isoforms but also to target the splicing of these isoforms directly as a therapy. A number of reports have described the potential for splicing modulation as a disease treatment (reviewed in [23-25]). In this treatment modality, an alternative splicing event is identified and modulated by delivery of antisense oligonucleotides that prevent or encourage the expression of a particular isoform. For example, oligonucleotides complementary to an aberrant exon-skipping junction can reduce the expression of isoforms containing that junction and push the balance of isoforms expressed back towards the normal isoform. Similarly, oligonucleotides can be used to induce skipping of an exon that contains a mutation or to silence a gene entirely by inducing nonsense mediated decay. Preliminary reports of the potential of these treatments have been reported for Duchene muscular dystrophy where exon-skipping of the *DMD* gene was induced by treatment with cocktails of antisense oligonucleotides to effectively reverse frameshift and nonsense mutations in Duchene dystrophy dogs[26]. Another report described shifting the balance of Mcl-1 isoform expression from the long (Mcl-1L) to short (Mcl-1S) isoform by delivery of antisense oligonucleotides in a gastric cancer cell line[27]. Since Mcl-1L is anti-apoptotic and Mcl-1S is pro-apoptotic, the effect of this treatment was to induce apoptosis. Antisense-mediated modification of alternative splicing may also serve as a powerful validation technique for events such as those identified in this thesis in the same way that over-expression assays and siRNA knockdown have proved invaluable in validating findings from differential gene expression studies. In order to develop therapeutic strategies

based on alternative isoforms representing disease markers, these isoforms must first be identified and characterized. In this thesis I have presented a general approach and specific novel methods to help accomplish this goal.

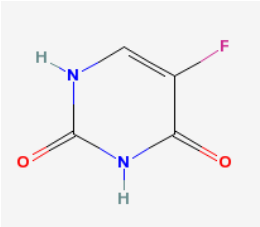
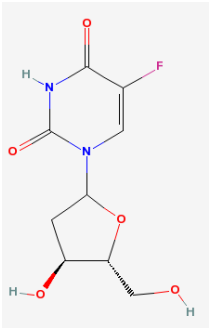
References

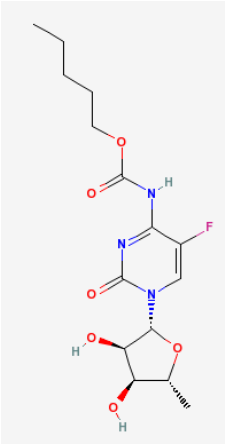
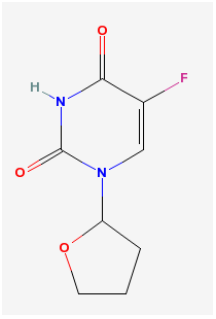
1. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-6 (2008).
2. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413-5 (2008).
3. Tai, I. T., Dai, M., Owen, D. A. & Chen, L. B. Genome-wide expression analysis of therapy-resistant tumors reveals SPARC as a novel target for cancer therapy. J Clin Invest 115, 1492-502 (2005).
4. Shimizu, D. et al. Prediction of chemosensitivity of colorectal cancer to 5-fluorouracil by gene expression profiling with cDNA microarrays. Int J Oncol 27, 371-6 (2005).
5. Birol, I. et al. De novo Transcriptome Assembly with ABySS. Bioinformatics (2009).
6. Simpson, J. T. et al. ABySS: A parallel assembler for short read sequence data. Genome Res 19, 1117-23 (2009).
7. Allegra, C. J. Dihydropyrimidine dehydrogenase activity: prognostic partner of 5-fluorouracil? Clin Cancer Res 5, 1947-9 (1999).
8. Maring, J. G., Groen, H. J., Wachters, F. M., Uges, D. R. & de Vries, E. G. Genetic factors influencing pyrimidine-antagonist chemotherapy. Pharmacogenomics J 5, 226-43 (2005).
9. Imyanitov, E. N. & Moiseyenko, V. M. Molecular-based choice of cancer therapy: realities and expectations. Clin Chim Acta 379, 1-13 (2007).
10. Kwan, T. et al. Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40, 225-31 (2008).
11. Stark, M., Wichman, C., Avivi, I. & Assaraf, Y. G. Aberrant splicing of folylpolyglutamate synthetase as a novel mechanism of antifolate resistance in leukemia. Blood 113, 4362-9 (2009).
12. Hu, R. et al. Ligand-independent androgen receptor variants derived from splicing of cryptic exons signify hormone-refractory prostate cancer. Cancer Res 69, 16-22 (2009).
13. Dehm, S. M., Schmidt, L. J., Heemers, H. V., Vessella, R. L. & Tindall, D. J. Splicing of a novel androgen receptor exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. Cancer Res 68, 5469-77 (2008).
14. Castle, J. C. et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40, 1416-25 (2008).
15. Hu, A. & Fu, X. D. Splicing oncogenes. Nat Struct Mol Biol 14, 174-5 (2007).
16. Klingbeil, P. et al. CD44 variant isoforms promote metastasis formation by a tumor cell-matrix cross-talk that supports adhesion and apoptosis resistance. Mol Cancer Res 7, 168-79 (2009).
17. Pajares, M. J. et al. Alternative splicing: an emerging topic in molecular and clinical oncology. Lancet Oncol 8, 349-57 (2007).
18. Della Beffa, C., Cordero, F. & Calogero, R. A. Dissecting an alternative splicing analysis workflow for GeneChip Exon 1.0 ST Affymetrix arrays. BMC Genomics 9, 571 (2008).

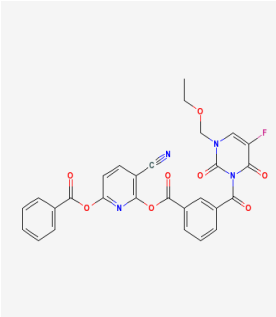
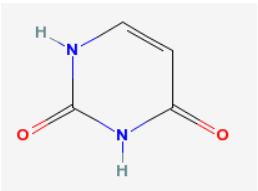
19. Zheng, H. et al. REMAS: a new regression model to identify alternative splicing events from exon array data. BMC Bioinformatics 10 Suppl 1, S18 (2009).
20. Robinson, M. D. & Speed, T. P. Differential splicing using whole-transcript microarrays. BMC Bioinformatics 10, 156 (2009).
21. Luqmani, Y. A. Mechanisms of drug resistance in cancer chemotherapy. Med Princ Pract 14 Suppl 1, 35-48 (2005).
22. Carethers, J. M. Review: Systemic treatment of advanced colorectal cancer: Tailoring therapy to the tumor. Therapeutic Advances in Gastroenterology 1, 33-42 (2008).
23. Bauman, J., Jearawiriyapaisarn, N. & Kole, R. Therapeutic potential of splice-switching oligonucleotides. Oligonucleotides 19, 1-14 (2009).
24. Dery, K. J. et al. Alternative splicing as a therapeutic target for human diseases. Methods Mol Biol 555, 127-44 (2009).
25. Du, L. & Gatti, R. A. Progress toward therapy with antisense-mediated splicing modulation. Curr Opin Mol Ther 11, 116-23 (2009).
26. Yokota, T. et al. Efficacy of systemic morpholino exon-skipping in duchenne dystrophy dogs. Ann Neurol 65, 667-676 (2009).
27. Shieh, J. J., Liu, K. T., Huang, S. W., Chen, Y. J. & Hsieh, T. Y. Modification of Alternative Splicing of Mcl-1 Pre-mRNA Using Antisense Morpholino Oligonucleotides Induces Apoptosis in Basal Cell Carcinoma Cells. J Invest Dermatol (2009).

Appendices

Appendix A. Description of 5-FU and related drugs (analogs, pro-drugs, 5-FU combination therapies, etc.)

Name	Also known as	Description and delivery method	Structure (of 5-FU analog/pro-drug)
Fluorouracil	5-FU; 5FU; Adrucil; Efudex; Carac; Fluoroplex; 5-fluoro-1H-pyrimidine-2,4-dione.	A uracil analog classified as an antimetabolite. Developed in the 1950's[1]. Now administered to approximately 2 million people per year worldwide[2]. Intravenous delivery for many solid tumours or as a topical cream in the treatment of skin cancer.	$C_4H_3FN_2O_2$ 
Fluoro-deoxyuridine	2'-Deoxy-5-fluorouridine; FUDR; Floxuridine; 5-Fluoro-1-[4-hydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl]-1H-pyrimidine-2,4-dione;	Intravenous delivery. An analog of 5-FU[1].	$C_9H_{11}FN_2O_5$ 
S-1	TS-1	Oral delivery. A fluorouracil anti-tumor drug[3-5] that combines three pharmacological agents: tegafur (FT), which is a prodrug of 5-fluorouracil (5-FU); 5-chloro-2,4-dihydroxypyridine (CDHP), which inhibits dihydropyrimidine dehydrogenase (*DPYD*) activity; and potassium oxonate (Oxo), which reduces gastrointestinal toxicity. 5-FU is the neo-plastic agent, CDHP improves efficacy be inhibiting *DPYD* and reducing metabolism of 5-FU in the liver resulting in maintenance of high blood concentrations of 5-FU, and Oxo inhibits *UMPS*/*OPRT* resulting in reduced toxicity by reducing 5-FU activation preferentially in the small intestine while still allowing activation in the bone marrow and tumour regions[6].	Same as Tegafur but delivered along with CDHP and Oxo (see description).

Name	Also known as	Description and delivery method	Structure (of 5-FU analog/pro-drug)
Capecitabine	Xeloda; pentyl[1-(3,4-dihydroxy-5-methyl-tetrahydrofuran-2-yl)- 5-fluoro-2-oxo-1H-pyrimidin- 4-yl] aminomethanoate.	Oral delivery. A pro-drug of 5-FU[7].	$C_{15}H_{22}FN_3O_6$ 
Tegafur-uracil	UFT; Uftoral; UFUR; Florafur; Fluorofur; 5-fluoro-1-(tetrahydro-2-furanyl)-2,4-(1h,3h)-pyrimidinedione.	Oral delivery. A pro-drug of 5-FU. Uracil is included as a competitive inhibitor of *DPYD* to reduce catabolism of 5-FU. UFT is considered a *DPYD* Inhibitory Flouropyrimidine[8, 9].	$C_8H_9FN_2O_3$  (Uracil not shown)
Eniluracil	Ethynyluracil; 5-Ethynyl-2,4(1H,3H)-pyrimidinedione.	Oral delivery. An inactivator of dihydropyrimidine dehydrogenase (*DPYD*) used in combination with 5-FU [10]. Allows an effective dose for oral delivery of 5-FU. The primary purpose of this combination is to reduce the probability of an adverse reaction to 5-FU and improve the effectiveness of 5-FU while lowering the effective dose.	Delivered in combination with 5-FU (see above)

Name	Also known as	Description and delivery method	Structure (of 5-FU analog/pro-drug)
BOF-A2	Emitefur; [6-(benzoyloxy)-3-cyanopyridin-2-yl] 3-[3-(ethoxymethyl)-5-fluoro-2,6-dioxopyrimidine-1-carbonyl]benzoate	Oral delivery. A derivative of 5-FU containing 1-ethoxymethyl-5-FU (EMFU), a masked form of 5-FU, and 3-cyano-2,6-dihydroxypyridine (CNDP), an inhibitor of 5-FU degradation[11].	$C_{28}H_{19}FN_4O_8$ 
Uracil	Pyrimidine-2,4(1*H*,3*H*)-dione; Uracil; 2-oxy-4-oxy pyrimidine; 2,4(1H,3H)-pyrimidinedione; 2,4-dihydroxypryimidine; 2,4-pyrimidinediol.	Oral delivery. The molecule which 5-FU was originally designed to mimic. Sometimes delivered in combination with Tegafur. Thought to reduce the side effects of Tegafur without reducing effectiveness.	$C_4H_4N_2O_2$ 

The structure diagrams displayed above were obtained from the NCBI PubChem (http://pubchem.ncbi.nlm.nih.gov/) via the National Library of Medicine (NLM) webpage.

References

1. Heidelberger, C. et al. Fluorinated pyrimidines, a new class of tumour-inhibitory compounds. Nature 179, 663-6 (1957).
2. Rich, T. A., Shepard, R. C. & Mosley, S. T. Four decades of continuing innovation with fluorouracil: current and future approaches to fluorouracil chemoradiation therapy. J Clin Oncol 22, 2214-32 (2004).
3. Sakata, Y. et al. Late phase II study of novel oral fluoropyrimidine anticancer drug S-1 (1 M tegafur-0.4 M gimestat-1 M otastat potassium) in advanced gastric cancer patients. Eur J Cancer 34, 1715-20 (1998).
4. Sugimachi, K. et al. An early phase II study of oral S-1, a newly developed 5-fluorouracil derivative for advanced and recurrent gastrointestinal cancers. The S-1 Gastrointestinal Cancer Study Group. Oncology 57, 202-10 (1999).
5. Hirata, K. et al. Pharmacokinetic study of S-1, a novel oral fluorouracil antitumor drug. Clin Cancer Res 5, 2000-5 (1999).
6. Shirasaka, T., Shimamoto, Y. & Fukushima, M. Inhibition by oxonic acid of gastrointestinal toxicity of 5-fluorouracil without loss of its antitumor activity in rats. Cancer Res 53, 4004-9 (1993).
7. Miwa, M. et al. Design of a novel oral fluoropyrimidine carbamate, capecitabine, which generates 5-fluorouracil selectively in tumours by enzymes concentrated in human liver and cancer tissue. Eur J Cancer 34, 1274-81 (1998).
8. Fujii, S., Kitano, S., Ikenaka, K. & Shirasaka, T. Effect of coadministration of uracil or cytosine on the anti-tumor activity of clinical doses of 1-(2-tetrahydrofuryl)-5-fluorouracil and level of 5-fluorouracil in rodents. Gann 70, 209-14 (1979).
9. Ho, D. H. et al. Comparison of 5-fluorouracil pharmacokinetics in patients receiving continuous 5-fluorouracil infusion and oral uracil plus N1-(2'-tetrahydrofuryl)-5-fluorouracil. Clin Cancer Res 4, 2085-8 (1998).
10. Baccanari, D. P., Davis, S. T., Knick, V. C. & Spector, T. 5-Ethynyluracil (776C85): a potent modulator of the pharmacokinetics and antitumor efficacy of 5-fluorouracil. Proc Natl Acad Sci U S A 90, 11064-8 (1993).
11. Nakai, Y. et al. Efficacy of a new 5-fluorouracil derivative, BOF-A2, in advanced non-small cell lung cancer. A multi-center phase II study. Acta Oncol 33, 523-6 (1994).

Appendix B. Primer sequences used for *UMPS* analysis

RT-PCR and semi-quantitative RT-PCR primers designed to produce separate products for UMPS isoforms A and B			
Amplicon Name	Forward primer sequence (5′ → 3′)	Reverse primer sequence (5′ → 3′)	Expected amplicon size (s)
F1/R1	TCTGCGGGGCATCGTGTCTC	CGTGCGCCTGCAACTTGTCC	383 / 229
F2/R2	AATGGCGGTCGCTCGTGCAG	GATCCCGTGCGCCTGCAACT	505 / 351
F3/R3	CGACAATGGCGGTCGCTCGT	CTGGGTGGATCCTGGGCAGC	720 / 566
F4/R4	GCATCGTGTCTCGACCGCGT	CCGATGCAAAGGCAGGCCCA	965 / 811
F5/R5	TGGGGCCATTGGTGACGGGT	GGGAGCCGGTGGAGCTCATT	1107 / 953
F6/R6	GGGCATCGTGTCTCGACCGC	TGCTGCTTCCAGACGATCAGC	1267 / 1113

Cloning primers used to generate full-ORF sequences for cloning by Topo-TA cloning			
Amplicon Name	Forward primer sequence (5′ → 3′)	Reverse primer sequence (5′ → 3′)	Expected amplicon size (s)
F3/R5	CAAACAGGCAGCGCGCGACA	GCTGACTTTAGCCTCTTGGTGCCC	2136

Sequencing primers used for clone finishing (designed using reference sequence, NM_000373)	
Primer Name	Primer sequence (5′ → 3′)
M13F	TGTAAAACGACGGCCAGT
M13R	CAGGAAACAGCTATGAC
F1	CATTGGTGACGGGTCTGT
F2	CAAGGACAAGTTGCAGGC
F3	GGGGGTGCCTCCTTATTG
F4	AGCTGATCGTCTGGAAGC
F5	GATGGAGTGCAGTGGTGA
R1	CTCCTTCTGAAGAACCT
R2	CAAGAACTCATGGCATT
R3	CAAGCAGCTTTTCTGTA
R4	TAGTCCCATGGAGAGAT
R5	ACTTTGAGAGACTGAGG

RT-PCR and semi-quantitative RT-PCR primers designed to produce separate products for UMPS isoforms A and B			
Amplicon Name	Forward primer sequence (5′ → 3′)	Reverse primer sequence (5′ → 3′)	Expected amplicon size (s)
TBP	TCAGGGCTTGGCCTCCCCTC	TGTGGGGTCAGTCCAGTGCCA	90
UMPS_A	GCATCGTGTCTCGACCGCGT	GGCACTCCACACACGGTGTCA	100
UMPS_B	CTCCCTGGCCACTGGGGACT	GGTTTCATGCTTACTCGGGAGCCA	105
UMPS_AB	ACCGCGTCTTCTGAGTCAGGAACTA	GATCCCGTGCGCCTGCAACT	194

Genomic sequencing primers (M13F and M13R linker sequence not shown)			
Name	Forward primer sequence (5′ → 3′)	Reverse primer sequence (5′ → 3′)	Size
A1	TGTGTCCTTTAACCACCCCCTACCC	AGCCAATGCCTTCTCGACTAGTTCT	757
A2	ACTCCTCCAGGAAGCAGAGTACAGA	GACCCGTCACCAATGGCCCC	624
A3	CCCCAGCGACCCCATCCAGA	GCGTTCCAAACTGGGTTTCCCGA	630
A4	CGACAATGGCGGTCGCTCGT	CACCTCAGCGTCCCAGGCGT	646
A5	GGTGACAGGAGTTGGGCCGTG	CCGGGGCTGCATAGGGGAGT	677
A6	CGACAAGTGAGACCCTGCCCTC	TTGAGACGAGCCTGGGCAGT	838
A7	TGAGCACAGGGGCTCATGCC	GCTTCTGGCCGCACAGTGCAA	706
A8	CATGTGCCATCATGCCCACCA	CGGCCCTCTACTCAGTGTGGTCT	613
A9	TGCCTCAGCCTCTTCCTGTTAGGT	AGTTAATGGTCCAAAGTACCGCAACC	604
A10	TGCCAGCTGAAACTGCAGACCAC	AGTGAGTTAATGCTACCTTTCCTCTCCA	722
A11	TGGACCATTAACTTCTCACCATCCAGA	TGGGTGCTGCCTCATGACCCT	637
A12	TGGAGAGGAAAGGTAGCATTAACTCACT	TTTCACTCCACTGTGTCCACAGAATTT	751
A13	TCTTGGAAGTAAGGACCTGAGAGAGC	AGGACGATCCCGACTTCCTGGT	730
A14	CTCTTTAGAAGTACTGGTGTAAACCTATGATCT	ATCTGCAACCTGTAAGAAAATGAACACAT	909
A15	ACAATAGGCTGGGCGCAGTGG	TGCTGATGAGGGAAGGGCTGGT	793
A16	TGACACCGTGTGTGGAGTGCC	ACGTTGCCAAAGCTGGTCTCAA	960
A17	GGCAGGTGGATCACAAGGCCA	GTGCACAGAAGCGGAAAAATTCTAAGAG	537
A18	TGAAAATTGGCCAGGCACAGTGG	AGAGTTTTCTTATGGGTAGGGATTCATACAGT	666
A19	TTTTGTGGCGTAAGGGCAGTGT	ACAGCAAGATCACCTAGATCAAGGAAAGTTA	671
A20	CTGAAAATAGTCTAGTGGGATTTGGGTAGAGA	AGTTCCTTCTACAAGACGCTTAGTTCCT	747
A21	AGGAACTAAGCGTCTTGTAGAAGGAACTA	TGCTAGCTGCAACAGCTCTCTGG	502
A22	TCCTTGATCTAGGTGATCTTGCTGT	CTGGGTGGATCCTGGGCAGC	893
A23	CCCCAGCGACCCCATCCAGA	ACGCGGTCGAGACACGATGC	365
A24	CGACAATGGCGGTCGCTCGT	TCGAGGAGCCCAACCCCTGG	342
A25	TACTCCCCTATGCAGCCCCG	GCCTCAGCCTCCCAAAGAAAGCC	319
A26	TGACACCGTGTGTGGAGTGCC	TGCTGATGAGGGAAGGGCTGGT	793
A27	GGCAGGTGGATCACAAGGCCA	GGCACTCCACACACGGTGTCA	372
A28	GGCAGGTGGATCACAAGGCCA	GTGCACAGAAGCGGAAAAATTCTAAGAG	537
A29	TGACACCGTGTGTGGAGTGCC	TCAATACAAAGCATTTCACCCCAAGTCT	301
A30	TAGGCTGGACGTGGTGGCTC	AGTTCCTTCTACAAGACGCTTAGTTCCT	474
A31	TTCTAGGAACTAAGCGTCTTGTAGAAGGAA	GATCCCGTGCGCCTGCAACT	200
A32	AGGAACTAAGCGTCTTGTAGAAGGAACTA	CTGGGTGGATCCTGGGCAGC	407
A33	AAGTTGCAGGCGCACGGGAT	TGCTAGCTGCAACAGCTCTCTGG	327
A34	AGGAGAATGTCTTTGTGGCAGCGA	TCCTATATCTGCAAACTTCCGGTCTTCA	353
A35	TCTGGCAAAATGCCATGAGTTCTTGA	TGCAGAGCAAATTGCCCAGGCT	379
A36	TCACAGATGGAACCATGGGTCTGC	CAGCATAGCAGCCCCTGCCC	396
A37	TGCATCGGGGGTGCCTCCTT	TCTGGGTCTCAAAGTTTCTCTTGGAGT	318
A38	AGCATAAGCCAGGCATGGTGAC	AGAGCCACGACCAGTGACCAGA	366
A39	TCTGGCTCCCGAGTAAGCATGA	GTGTTTACTCTGTGCCAAGACATTGTG	364
A40	ACGCAGTTGCTGTACAAAAGGGGA	GAAGCACTCAAACACCAAGTCTACTCA	318
A41	GGTCGTGGCATAATCTCAGCAGC	CCTAACCCAGACAGGACTGTGGC	363
A42	CCTGCTTGGATTCTTCCACAGGGC	CTGGAAGCTGAGGTGGGAGGAT	359
A43	GGATCCTTCCTATCTCTCCATGGGACT	GCTGACTTTAGCCTCTTGGTGCCC	391
A44	GGCACCAAGAGGCTAAAGTCAGCA	ACTGGTCCATAATTGCTGGCTAGGGG	301

Appendix C. Ethics approval certificates

Appendix C1. Ethics certificate for samples obtained from the Ontario Institute for Cancer Research (Ontario Tumour Bank)

University of British Columbia - British Columbia Cancer Agency Research Ethics Board (UBC BCCA REB)

UBC BCCA Research Ethics Board
Fairmont Medical Building (6th Floor)
614 - 750 West Broadway
Vancouver, BC V5Z 1H5
Tel: (604) 877-6284 Fax: (604) 708-2132
E-mail: reb@bccancer.bc.ca
Website: http://www.bccancer.bc.ca > Research Ethics
RISe: http://rise.ubc.ca

Certificate of Expedited Approval

PRINCIPAL INVESTIGATOR:	INSTITUTION / DEPARTMENT:	REB NUMBER:
Marco A. Marra	BCCA/Administration (BCCA)	H08-00593

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
BC Cancer Agency	Vancouver BCCA

Other locations where the research will be conducted:
N/A

PRINCIPAL INVESTIGATOR FOR EACH ADDITIONAL PARTICIPATING BCCA CENTRE:

Vancouver:	N/A	**Vancouver Island:**	N/A
Fraser Valley:	N/A	**Southern Interior:**	N/A

SPONSORING AGENCIES AND COORDINATING GROUPS:

Alberta Heritage Foundation for Medical Research
Michael Smith Foundation for Health Research
National Cancer Institute of Canada

PROJECT TITLE:
The Use of Novel Genomic Approaches to Identify and Characterize Genes Associated with 5-FU Resistance in Colorectal Cancer (Ontario Tumour Bank)

The UBC BCCA Research Ethics Board Chair, Vice-Chair or second Vice-Chair, has reviewed the above described research project, including associated documentation noted below, and finds the research project acceptable on ethical grounds for research involving human subjects and hereby grants approval.

EXPIRY DATE OF THIS APPROVAL: May 15, 2009

DATE DOCUMENT(S) APPROVED: May 15, 2008

LIST OF DOCUMENTS APPROVED:

Document Name	Version	Date
Consent Forms:		
Ontario Tumour Bank - Letter of Information and Consent Form	1	July 30, 2007
Other Documents:		
Letter from Ontario Tumour Bank - Verification of sample availability and consent processes	1	May 14, 2008

CERTIFICATION:

1. The membership of the UBC BCCA REB complies with the membership requirements for research ethics boards defined in Division 5 of the Food and Drug Regulations of Canada.
2. The UBC BCCA REB carries out its functions in a manner fully consistent with Good Clinical Practices.
3. The UBC BCCA REB has reviewed and approved the research project named on this Certificate of Approval including any associated consent form and taken the action noted above. This research project is to be conducted by the provincial investigator named above. This review and the associated minutes of the UBC BCCA REB have been documented electronically and in writing.

UBC BCCA Ethics Board approval of the above has been verified by one of the following:

Signatures removed

Dr. George Browman, Chair

Dr. Joseph Connors, First Vice-Chair

Dr. Lynne Nakashima Second Vice-Chair

If you have any questions, please call:

Bonnie Shields, Manager, BCCA Research Ethics Board: 604-877-6284 or e-mail: reb@bccancer.bc.ca
Dr. George Browman, Chair: 604-877-6284 or e-mail: gbrowman@bccancer.bc.ca
Dr. Joseph Connors, First Vice-Chair: 604-877-6000-ext. 2746 or e-mail: jconnors@bccancer.bc.ca
Dr. Lynne Nakashima, Second Vice-Chair: 604-707-5989 or e-mail: lnakas@bccancer.bc.ca

Appendix C2. Ethics certificate for samples obtained from the British Columbia Cancer Agency

University of British Columbia - British Columbia Cancer Agency Research Ethics Board (UBC BCCA REB)

UBC BCCA Research Ethics Board
Fairmont Medical Building (6th Floor)
614 - 750 West Broadway
Vancouver, BC V5Z 1H5
Tel: (604) 877-6284 Fax: (604) 708-2132
E-mail: reb@bccancer.bc.ca
Website: http://www.bccancer.bc.ca > Research Ethics
RISe: http://rise.ubc.ca

Certificate of Expedited Approval

PRINCIPAL INVESTIGATOR:	INSTITUTION / DEPARTMENT:	REB NUMBER:
Marco A. Marra	BCCA/BCCA/Administration (BCCA)	H08-03101

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
BC Cancer Agency	Vancouver BCCA

Other locations where the research will be conducted:
N/A

PRINCIPAL INVESTIGATOR FOR EACH ADDITIONAL PARTICIPATING BCCA CENTRE:

Vancouver:	Marco A. Marra	**Vancouver Island:**	N/A
Fraser Valley:	N/A	**Southern Interior:**	N/A

SPONSORING AGENCIES AND COORDINATING GROUPS:

Alberta Heritage Foundation for Medical Research
Michael Smith Foundation for Health Research
National Cancer Institute of Canada

PROJECT TITLE:
Investigation of the clinical utility of UMPS mutation detection in predicting response to 5-FU in metastases of Colorectal Cancer (BCCA)

The UBC BCCA Research Ethics Board Chair, Vice-Chair or second Vice-Chair, has reviewed the above described research project, including associated documentation noted below, and finds the research project acceptable on ethical grounds for research involving human subjects and hereby grants approval.

EXPIRY DATE OF THIS APPROVAL: January 7, 2010

DATE DOCUMENT(S) APPROVED: January 7, 2009

CERTIFICATION:

1. The membership of the UBC BCCA REB complies with the membership requirements for research ethics boards defined in Division 5 of the Food and Drug Regulations of Canada.
2. The UBC BCCA REB carries out its functions in a manner fully consistent with Good Clinical Practices.
3. The UBC BCCA REB has reviewed and approved the research project named on this Certificate of Approval including any associated consent form and taken the action noted above. This research project is to be conducted by the provincial investigator named above. This review and the associated minutes of the UBC BCCA REB have been documented electronically and in writing.

UBC BCCA Ethics Board approval of the above has been verified by one of the following:

Signatures removed

Dr. George Browman, Chair	Dr. Joseph Connors, First Vice-Chair	Dr. Lynne Nakashima Second Vice-Chair

If you have any questions, please call:

Bonnie Shields, Manager, BCCA Research Ethics Board: 604-877-6284 or e-mail: reb@bccancer.bc.ca
Dr. George Browman, Chair: 604-877-6284 or e-mail: gbrowman@bccancer.bc.ca
Dr. Joseph Connors, First Vice-Chair: 604-877-6000-ext. 2746 or e-mail: jconnors@bccancer.bc.ca
Dr. Lynne Nakashima, Second Vice-Chair: 604-707-5989 or e-mail: lnakas@bccancer.bc.ca

Appendix C3. Ethics certificate for samples obtained from St. Paul's Hospital

University of British Columbia - British Columbia Cancer Agency Research Ethics Board (UBC BCCA REB)

UBC BCCA Research Ethics Board
Fairmont Medical Building (6th Floor)
614 - 750 West Broadway
Vancouver, BC V5Z 1H5
Tel: (604) 877-6284 Fax: (604) 708-2132
E-mail: reb@bccancer.bc.ca
Website: http://www.bccancer.bc.ca > Research Ethics
RISe: http://rise.ubc.ca

Certificate of Expedited Approval

PRINCIPAL INVESTIGATOR:	INSTITUTION / DEPARTMENT:	REB NUMBER:
Marco A. Marra	BCCA/BCCA/Administration (BCCA)	H08-01933

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
BC Cancer Agency	Vancouver BCCA
Providence Health Care	St. Paul's Hospital

Other locations where the research will be conducted:
N/A

PRINCIPAL INVESTIGATOR FOR EACH ADDITIONAL PARTICIPATING BCCA CENTRE:

Vancouver:	Marco A. Marra	**Vancouver Island:**	N/A
Fraser Valley:	N/A	**Southern Interior:**	N/A

SPONSORING AGENCIES AND COORDINATING GROUPS:

Alberta Heritage Foundation for Medical Research
Michael Smith Foundation for Health Research
National Cancer Institute of Canada

PROJECT TITLE:
Investigation of the clinical utility of UMPS isoform expression profiling in predicting response to 5-FU in Colorectal Cancer (St. Paul's Hospital)

The UBC BCCA Research Ethics Board Chair, Vice-Chair or second Vice-Chair, has reviewed the above described research project, including associated documentation noted below, and finds the research project acceptable on ethical grounds for research involving human subjects and hereby grants approval.

EXPIRY DATE OF THIS APPROVAL: October 8, 2009

CERTIFICATION:

1. The membership of the UBC BCCA REB complies with the membership requirements for research ethics boards defined in Division 5 of the Food and Drug Regulations of Canada.
2. The UBC BCCA REB carries out its functions in a manner fully consistent with Good Clinical Practices.
3. The UBC BCCA REB has reviewed and approved the research project named on this Certificate of Approval including any associated consent form and taken the action noted above. This research project is to be conducted by the provincial investigator named above. This review and the associated minutes of the UBC BCCA REB have been documented electronically and in writing.

UBC BCCA Ethics Board approval of the above has been verified by one of the following:

Signatures removed

Dr. George Browman, Chair	Dr. Joseph Connors, First Vice-Chair	Dr. Lynne Nakashima Second Vice-Chair

If you have any questions, please call:

Bonnie Shields, Manager, BCCA Research Ethics Board: 604-877-6284 or e-mail: reb@bccancer.bc.ca
Dr. George Browman, Chair: 604-877-6284 or e-mail: gbrowman@bccancer.bc.ca
Dr. Joseph Connors, First Vice-Chair: 604-877-6000-ext. 2746 or e-mail: jconnors@bccancer.bc.ca
Dr. Lynne Nakashima, Second Vice-Chair: 604-707-5989 or e-mail: lnakas@bccancer.bc.ca