Characterization of transcript isoform variations in human and chimpanzee

David Benovoy

Department of Human Genetics
McGill University
Montreal
June 1st, 2009

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Transcript expression and pre-mRNA processing are emerging as important mechanisms that increase the complexity of eukaryotic transcriptomes. These processes allow a genomic locus to produce a number of mRNAs and proteins with distinct properties that affect function, stability, and sub-cellular localization by controlling the rate of transcript expression, by varying the initiation or termination of transcription and by modulating the inclusion of exons (alternative splicing) in mature mRNAs. Thus, it is crucial to determine the extent of these types of variations to better understand their importance in creating organism diversity. The studies described in this thesis provide the first genome-wide estimations of how single nucleotide polymorphisms (SNPs) affect the regulation of transcript expression and pre-mRNA processing in a human population as well as between humans and chimpanzees using a microarray-based approach. We first demonstrated that transcript expression changes at the isoform level are common between two unrelated individuals and that these changes are heritable and therefore have an underlying genetic component. We then investigated what proportion was under genetic control in a normal human population by conducting a genome-wide association analysis between single nucleotide polymorphisms and transcript isoform variants. We found that 50-55% of transcript expression variation is isoform based. We also extended our comparison of human transcript isoform variation to chimpanzee. We showed that genetic substitutions in regulatory sequences are responsible for some of the isoform variations observed between these two closely related species. We ascertained that in our study these isoform variations are responsible for certain phenotypic differences mostly related to immune responses. These results constitute an important change in the way genetic variations are viewed in humans and chimpanzees and they highlight the need for broader investigation into these types of variation and how they affect

gene expression. In the last two chapters of this thesis we also provide solutions for some of the methodological and analysis issues we encountered because they could be of a great benefit to scientist conducting experiments with the Exon Array.

Résumé

Le niveau d'expression d'un transcrit et les processus de maturation de celui-ci en ARN messager (ARNm) se révèlent être des mécanismes augmentant la complexité du transcriptome des eucaryotes. Ces processus permettent au même locus génomique de produire plusieurs ARNm et protéines ayant des propriétés distinctes qui affectent leurs fonctions, leur stabilité et leurs localisations intra cellulaire en contrôlant la vitesse de transcription, en variant le site d'initiation ou de terminaison de la transcription et en modulant l'inclusion d'exons (épissage) dans les ARNm matures. Il est donc primordial de déterminer l'ampleur de ces types de variations afin de mieux comprendre leur impact sur la diversite des oraganismes. Les études décrites dans cette thèse fournissent les premières estimations de la façon dont les variations de polymorphism nucléotidique simple (SNP) peuvent affecter la régulation de l'expression d'un transcrit et ses processus de maturation à l'échelle du génome entier. Ces processus sont examinés dans une population humaine et entre humain et chimpanzé en utilisant une méthode basée sur les puces à ADN. Nous démontrons d'abord l'existence d'un nombre important de variations d'isoformes d'ARNm entre deux individus non apparentés et nous démontrons que ces variations sont héritées ce qui leur révèle une composante génétique. Par la suite, nous avons déterminé quelle proportion et quel type de variation au niveau de l'isoform était sous contrôle génétique dans une population humaine. En réalisant une analyse d'association entre l'expression des transcrits du génome entier et les SNPs présents dans cette population, nous avons observé que 50-55% de la variation était à l'échelle de l'isoforme du transcrit. Nous avons aussi étendu cette comparaison au chimpanzé en utilisant les profils d'expression mesurés lors de l'analyse précédente. Nous avons démontré que des substitutions dans certaines séquences qui régulent l'épissage étaient responsables de variations d'expression au niveau des isoforms

de transcrits entre ces deux espèces apparentées. Nous estimons que ce type de variation est responsable de certaines différences phénotypiques, plus précisément au niveau de certaines réponses immunitaires. Ensemble ces observations amènent un changement important dans notre compréhension du rôle de ces variations dans le contrôle de l'expression des gènes et elles soulignent l'importance de mener des recherches plus étendues sur ces types de variations ainsi que l'impact produit sur l'expression des gènes. De plus, les deux derniers chapitres décrivent diverses solutions que nous avons élaborées afin d'aider la communauté scientifique qui utilise le Exon Array.

Acknowledgments

Since many people have helped me throughout this thesis, I will try to thank everybody…

I would like to first acknowledge my supervisor Jacek Majewski. By giving me the chance to work on so many different and interesting projects you made my Ph.D experience extremely interesting and rewarding. Your enthusiasm for science, approachability and incredible attentiveness really help my projects succeed. Also thank you for your advice and the many hours spent talking about science, editing my papers and listening to my presentations.

I would like to thank my student committee members Ken Dewar, Mathieu Blanchette and Robert Nadon as well as Tomi Pastinen for the great ideas they contributed to my work.

I would also like to thank past and present members of the Majewski lab. Working with all of you made my experience particularly enjoyable and remarkable. I would like to especially thank Tony Kwan with whom I shared an office. You were always there to discuss and help me with the many questions I had on a multitude of subjects.

Big thanks to everybody at the Genome Centre, from the 4 Bases baseball team to our weekly beer, whiskey and/or vodka meetings, all of you made the social aspect of studying there unforgettable.

Finally, many thanks to my family and friends for their constant encouragement throughout these years. To my parents for supporting me and especially my mother for cooking me delicious meals every weekend. To Julie for her unconditional support.

Table of contents

List of Tables

List of Figures

Author contribution

This thesis is written in manuscript format as permitted by the McGill Faculty of Graduate Studies and is comprised of five manuscripts. The contribution of each author is described below.

Chapter 3: Heritability of alternative splicing in the human genome.

Tony Kwan, **David Benovoy**, Christel Dias, Scott Gurd, David Serre, Harry Zuzan, Tyson A. Clark, Anthony Schweitzer, Michelle K. Staples, Hui Wang, John E. Blume, Thomas J. Hudson, Rob Sladek, and Jacek Majewski.

Tony Kwan and I performed the majority of the analysis for this paper. My specific contributions were related to the data analysis and troubleshooting aspects of the data generated with the Exon Array. Tony Kwan also prepared the manuscript. Christel Dias and Scott Gurd performed the array hybridization and validation work (RT-PCR). David Serre and Harry Zuzan contributed ideas related to the statistical aspects of the linkage analysis. Tyson A. Clark, Anthony Schweitzer, Michelle K. Staples, Hui Wang, John E. Blume were collaborators from Affymetrix. Thomas J. Hudson, Rob Sladek contributed ideas and edited the manuscript. Jacek Majewski was the main editor and principal investigator for this study.

Chapter 4: Genome-wide analysis of transcript isoform variation in humans.

Tony Kwan, **David Benovoy**, Christel Dias, Scott Gurd, Cathy Provencher, Patrick Beaulieu, Thomas J Hudson, Rob Sladek & Jacek Majewski

Tony Kwan and I performed all the analyses for this paper. Tony Kwan was the main author for this manuscript and I wrote parts of the Material

and Methods section and made many of the figures. Christel Dias, Scott Gurd, Cathy Provencher performed the wet-lab tasks such as preparing the RNA, hybridizing it to the microarray and performing the validation (real time PCR, and RT-PCR). Patrick Beaulieu posted the data in a genome browser format on the GRID website. Thomas J Hudson, Rob Sladek contributed ideas and edited the manuscript. Jacek Majewski was the main editor and principal investigator for this study.

Chapter 5: Genome-wide comparison of human and chimpanzee exomes.
David Benovoy & Jacek Majewski

I performed the analyses and wrote the manuscript. Jacek Majewksi edited the manuscript.

Chapter 6: Alternative isoform detection using Exon arrays
David Benovoy, Amandine Bemmo, Tony Kwan, Daniel J Gaffney, Roderick V Jensen and Jacek Majewski.

This chapter is an amalgamation of two manuscripts. For the article published in *BMC Genomics* Amandine Bemmo and I performed the majority of the analysis. I included in chapter 6 only the parts I work on because Amandine will use the others in her own thesis. I included the methods I developed in studies described in chapter 3 and 4. The article published as a chapter in the *Handbook of Research on Systems Biology Applications in Medicine* is a description of methods we developed from our experience with the Exon Array.

Chapter7: Effect of polymorphisms within probe–target sequences on olignonucleotide microarray experiments.
David Benovoy, Tony Kwan and Jacek Majewski

I performed the analysis and wrote the manuscript. Tony Kwan and I developed this method for the study described in chapter 4. Jacek Majewski edited the manuscript.

Chapter 1 Introduction

Preface

Over the past two decades, sequencing and comparison of entire genomes from different species have changed our conception of organism complexity and diversity. The surprisingly low number of genes found throughout diverse eukaryotic organisms such as worm, mouse, chimpanzee and human suggests that an increase in biological complexity and diversity is achieved by other means (CHERRY *et al.* 1997; LANDER *et al.* 2001; THE *C. ELEGANS* SEQUENCING CONSORTIM 1998; THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005; WATERSTON *et al.* 2002).

Insight into this paradox was obtained when mRNA was defined as the intermediate between the genetic information contained in sections of DNA (genes) and the protein-synthesizing machinery. Research into its regulation has completely changed our view of how information flows in the cell (CRICK 1970). The concept of one promoter that controls one gene which is transcribed to one mRNA transcript no longer holds. In fact, a single genomic locus can produce multiple mRNA transcript isoforms with the use of alternative transcription initiation and termination as well as alternative pre-mRNA splicing. Processes such as alternative transcription initiation and termination modify the 5' and 3' ends of mRNA transcripts, respectively, while alternative splicing consists of the differential exclusion of exons within mRNA transcripts. Consequently, this can alter mRNA turnover, translation and sub-cellular localization (GRENS and SCHEFFLER 1990; RUSSO *et al.* 2006; WANG *et al.* 2008) or create different protein domain combinations such as in the classical example of the *Dscam* gene (SCHMUCKER *et al.* 2000). Overall, it is estimated that 95% of mammalian genes encode for multiple transcript isoforms (PAN *et al.* 2008; WANG *et al.* 2008). Thus, these processes further diversify eukaryotic transcriptomes and proteomes and have contributed to the evolution of organism

complexity. A cell can adapt to changing environments and states by tightly regulating transcription and pre-mRNA processing. In specialized tissues such as the brain, liver and testis, the frequency of alternative splicing is higher to accommodate their complex functions (JOHNSON *et al.* 2003). Studying transcriptome variation is becoming increasingly important because of its contribution to phenotypic differences among individuals and its regulatory and functional relationship to disease. In fact, splicing defects can result in genetic disorders (FAUSTINO and COOPER 2003) and in some cases confer susceptibility to complex diseases (reviewed in (COOPER *et al.* 2009; LUKONG *et al.* 2008; WANG and COOPER 2007). Consequently, the study of transcriptome variation is important to a broad range of biomedical disciplines from evolutionary biology through development and to medicine.

These transcriptome variations are routinely investigated using DNA microarrays. The typical microarray platform employs a large collection of probes that are designed to hybridize to specific targets, usually a fluorescently labelled nucleic acid sequence from a particular gene. The fluorescence emitted by the bound target to its probe is measured and compared between samples being investigated to identify variation in whole-transcript expression. More recently, advances in microarray design enabled the investigation of mRNA expression at the resolution of a single exon. The Affymetrix GeneChip® Human Exon 1.0 ST Array is the first commercially available microarray product designed for transcriptome-wide exon level analysis. The array relies on targeting multiple probes to individual exons and allows exon-level detection of expression intensity for ~1.4 million exons which theoretically covers the entire set of human exons. The complexity of this array design and the sheer magnitude of data generated per experiment have hindered the use of traditional analysis methods. Therefore, new statistical and data visualization

approaches are needed to adequately analyze expression data derived with the Human Exon Array.

Hypothesis

What is hypothesized in this thesis is that inter- and intra- genetic difference in humans and chimpanzees produce variable expression profiles of mRNA isoforms and that it is possible to adequately measure these types of variations using isoform sensitive microarrays.

Outline

This thesis consists of a literature review, five manuscripts and a discussion that together address the study of gene expression variation at the isoform level in humans and chimpanzees using the Affymetrix GeneChip® Human Exon 1.0 ST Array. The literature review summarizes the basic mechanisms of transcription and pre-mRNA processing, describes how these processes are regulated and explains some of their effects on organism phenotype and diversity. It also describes what tools are used to study gene expression and examines the typical analysis workflow of the Affymetrix GeneChip® Human Exon 1.0 ST Array. The third chapter is a pilot study that was performed to verify the efficacy of the Human Exon Array in detecting transcript isoform variations among two human individuals. This study demonstrated that the Human Exon array was capable of detecting transcript isoform differences that were caused by alternative transcript initiation, alternative splicing and alternative termination. A linkage analysis conducted on single nucleotide polymorphisms also showed that these types of isoform variations were heritable and therefore had an underlying genetic component. This prompted a second study that is described in the fourth chapter of this thesis. A genome-wide association analysis was conducted between single nucleotide polymorphisms and transcript isoform variants. It

demonstrated that expression variation at the isoform level was under genetic control and common in a natural population. It also investigates the relationship between genetic variations associated to certain splicing differences that cause disease phenotypes. The fifth chapter extends the comparison to the chimpanzee. It uses the expression profiles derived from the previous human population study and compares them to the expression profile derived from a chimpanzee lymphoblast cell line. It confirms that genetic substitutions in regulatory sequences are responsible for some of the isoform variations observed between these two closely related species. It also ascertains that these isoform variations are responsible for certain phenotypic differences mostly related to immune responses. The following two chapters (6 and 7) relate to the methodological issues involved in the analysis of the Human Exon Array because substantial time and effort was put into finding solutions to the different technical problems encountered during the analyses described in the last three chapters that could greatly benefit scientist conducting experiments with the Human Exon Array. The sixth chapter outlines problems encountered in the analysis of expression data generated with the Exon array. It also describes some of the statistical and technological problems encountered and proposes solutions to resolve them. The following seventh chapter continues on the technical theme of the preceding one. It describes how polymorphisms present in the probe-target sequence affect hybridization. It shows that this effect is the main source of false positives in Exon Array experiments involving individuals of different genetic backgrounds and a simple solution is proposed to reduce the false positive rate that consists of removing misbehaving probes from the analysis. The last chapter (chapter 8) is a summary of the main results and a discussion of the future work that is needed to better comprehend the role of transcript expression variation in organisms.

Chapter 2 Literature review

Gene expression

The central dogma of molecular biology states that genetic information is transferred in a sequential manner (Figure 2.1) and that each type of molecule (DNA, RNA and protein) is used as a template for the synthesis of another and is entirely dependant on the original molecule (CRICK 1970). The general model (see full lines in Figure 2.1) describes the normal flow of information in cells; (1) the DNA copies itself through DNA replication, (2) genetic information is copied from the DNA to a RNA transcript via transcription and this RNA transcript is then (3) translated into a protein.

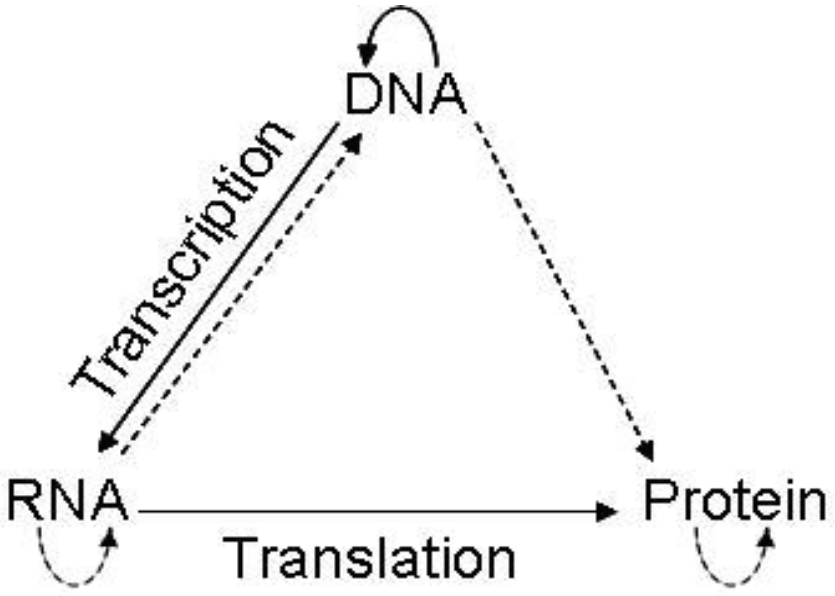


Figure 2.1: The central dogma of molecular biology. Solid arrows show the general transfer of genetic information from DNA to RNA to protein that occurs in most cells. The dashed arrows show the special transfer of genetic information such as RNA to RNA that occurs occasionally in some RNA viruses (LEIS and HURWITZ 1972), DNA to protein transmission has been observed only in-vitro studies (MCCARTHY and HOLLAND 1965; UZAWA *et al.* 2002), protein to protein transmission is taught to occur in

prion replication (WEISSMANN 2004) and there is no evidence of protein to RNA or DNA. Figure modified from (CRICK 1970).

Transcription

Transcription plays a central role in this model (Figure 2.1) because it acts as a messenger between genetic information contained in sections of DNA (genes) and the protein-synthesizing machinery of the cell. In eukaryotes, transcription involves two main phases; the first is the transcription of a gene into a primary RNA transcript (pre-mRNA) that is divided into 5 stages: pre-initiation, initiation, promoter clearance, elongation and termination. The second phase is the processing of this primary transcript (pre-mRNA) into a mature messenger RNA (mRNA) in a 3 step process that consists of 5'-capping, splicing and polyadenylation.

Chromatin remodelling

The first step in gene transcription is called pre-initiation. This is where the gene promoter is exposed by the remodelling of chromatin. Chromatin is formed of proteins that serve as scaffold onto which DNA is packaged. DNA is wrapped around histone proteins (an H3-H4 tetramer flanked by two H2A-H2B dimmers) that make up the nucleosomes and are the primary repeating units of chromatin (KORNBERG 1974; KORNBERG and THOMAS 1974). Transcription is repressed when nucleosomes inhibit the access of the transcription machinery to the promoter (WORKMAN and BUCHMAN 1993; YAGER *et al.* 1989). Therefore, to allow the transcription machinery to gain access to the genomic DNA of the promoter, nucleosomes are modified by histone acetylation (PENNISI 1997; WADE *et al.* 1997) and by chromatin remodelling enzymes (CAIRNS 1998) that together displace the nucleosomes and change the structure of the chromatin in order to expose the promoter.

Transcription initiation

Promoters of genes that encode proteins are usually composed of a core promoter near the transcription start site as well as enhancer elements that can be several kilobases upstream and/or downstream of the transcription start site (BLANCHETTE *et al.* 2006). The DNA segments where these enhancer elements lie can bend back on themselves to allow the placement of regulatory sequences near the core promoter. The core promoter is where the assembly of the transcription initiation complex takes place. This complex is composed of enhancer elements bound by transcription factors that in turn, regulate transcription by promoting or inhibiting the recruitment of the RNA polymerase (KARIN 1990; LATCHMAN 1997). The RNA polymerase, also called DNA-dependent RNA polymerase is responsible for the transcription of DNA into RNA. It uses the complementary nature of DNA and RNA to produce a primary RNA copy based on the segment of DNA it is transcribing (Figure 2.2). In eukaryotes, there are three types of RNA polymerase; RNA polymerase I, II and III (ROEDER and RUTTER 1969). These polymerases consist of 8 to 12 protein subunits and transcribe specific types of genes. For instance, RNA polymerase I and III transcribe RNA genes such as ribosomal, transfer and small nucleolar genes (RUSSELL and ZOMERDIJK 2006; WOLFFE 1991) whereas RNA polymerase II mostly transcribes protein coding genes (BOEGER *et al.* 2005; KORNBERG 1999). Once the transcription initiation complex composed of transcription factors and the RNA polymerase have been assembled on the core promoter, transcription elongation starts.

Transcription elongation

The next step in the transcription of a protein coding gene involves the synthesis of a pre-mRNA transcript by RNA polymerase II (Figure 2.2). The RNA polymerase unwinds the DNA strand using helicase action

(SVEJSTRUP *et al.* 1996), clears the promoter and starts transcription at the transcription start site. The RNA polymerase travels from the $3' \rightarrow 5'$ end of the anti-sense DNA strand and uses it as a template to synthesize the pre-mRNA transcript from the $5' \rightarrow 3'$ end. It assembles ribonucleotides following the rules of base pairing (WATSON and CRICK 1953) and produces an exact copy of the DNA sense-strand although the thymines are replaced by uracils and the nucleotides are composed of ribose sugar instead of a deoxyribose as in DNA. The RNA polymerase continues to transcribe the gene until a transcription termination event.

Transcription termination

The exact mechanism of transcription termination is not well understood in eukaryotes although two scenarios involving the polyadenylation signal have been proposed. The first referred to as the "anti-termination" model suggests that the emergence of the polyadenylation sequence on the RNA transcript and subsequent binding of a polyadenylation factor could displace a positive elongation factor or recruit a negative elongation factor and consequently the RNA polymerase would terminate transcription (LOGAN *et al.* 1987). In the second scenario, the "torpedo" model, the polyadenylation site is cleaved and generates a new uncapped 5' end (CONNELLY and MANLEY 1988). This uncapped end would act as an entry point for an exonuclease or helicase that would track along the RNA and dissociate the RNA polymerase. Other studies have shown factors that induce the pausing of the RNA polymerase such as the transcription of particular RNA sequences that create secondary structures in the RNA or DNA binding proteins that inhibit the forward movement of the RNA polymerase could trigger termination (YONAHA and PROUDFOOT 1999). In general these theories all point to a stochastic process that terminates transcription somewhere downstream of the polyadenylation site (KIM and MARTINSON 2003; TRAN *et al.* 2001) (Figure 2.2).

Pre-mRNA processing

Processing of pre-mRNA usually occurs in a co-transcriptional manner meaning that the pre-mRNA is processed into mRNA while it is synthesized (Figure 2.2). RNA polymerase II contains a unique C-terminal protein domain (CDT) that coupled with processing factors, are responsible for directing the three main post-transcriptional modifications; (1) 5'-end capping, (2) splicing and (3) polyadenylation (CALVO and MANLEY 2003; MCCRACKEN *et al.* 1997a; MCCRACKEN *et al.* 1997b; NEUGEBAUER 2002; REED 2003).

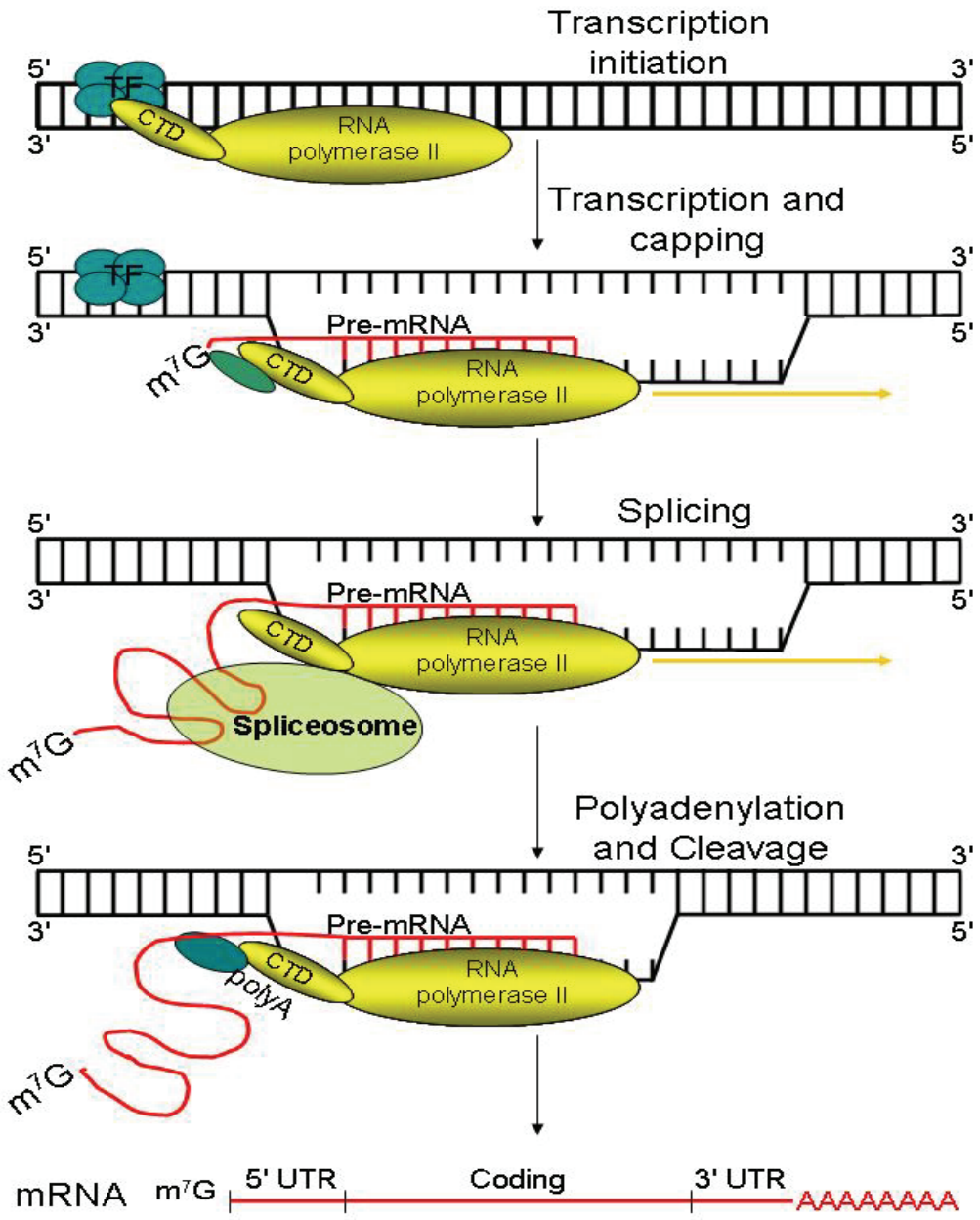


Figure 2.2: Co-transcriptional pre-mRNA processing. Schematic illustrating the principal steps involved in pre-mRNA processing (capping (m7G), splicing and polyadenylation) and their interaction with the C-terminal domain of the RNA polymerase II to form a mature mRNA transcript.

5' capping

Soon after the pre-mRNA has emerged from the RNA polymerase II the 5' end undergoes a chemical modification with the addition of a cap (Figure 2.2). This cap formation involves three enzymatic reactions: a 5'-triphosphatase that removes the γ-phosphate from the first transcribed nucleotide, a guanylyltransferase (GTase) that attaches a guanosine via a 5'–5' triphosphate linkage, and a 7-methyltransferase that modifies the terminal guanine (reviewed in (SHATKIN and MANLEY 2000). Capping the 5'-end mainly stabilizes the mature mRNA against $5' \rightarrow 3'$ exonucleolytic degradation, facilitates mRNA cytoplasmic transport and assists with translation (HOWE 2002).

Constitutive Splicing

Constitutive splicing is the process by which intron sequences are removed from the pre-mRNA and consecutive exons are joined. This process is catalyzed by a complex of small nuclear ribonucleoproteins and associated proteins designated as the spliceosome that assembles on the pre-mRNA in a stepwise manner at the splice sites located at the intron-exon boundaries. The intron-exon boundaries are defined by specific sequences that are recognized by the spliceosome. In addition, both exons and introns contain weak binding sites such as exonic and intronic splicing enhancers and silencers for a multitude of splicing auxiliary and regulatory proteins (MATLIN *et al.* 2005; WANG and COOPER 2007). The donor splice site is located at the 5'-end of the intron and begins with a GU dinucleotide while the acceptor splice site located at the 3'-end of the intron and ends with an AG dinucleotide (Figure 2.3.A). The first steps of spliceosome assembly is the recognition of the donor splice site by the small nuclear ribonucleoprotein (snRNP) U1, the binding of splicing factor SF1 to the branchpoint and the recognition of the acceptor splice site by the U2 snRNP auxiliary factor (U2AF) that together form the E complex.

Following this initiation step, the A complex is formed when the U2 snRNP binds to the branch point dislodging the splicing factor SF1. Subsequently, the A complex is substantially remodelled by the action of 3 other snRNPs (U4, U5, and U6) to form the B complex and leads to the formation of the mature and active spliceosome (C complex) that catalyses both trans-esterification splicing reactions (BLACK 2003; BLAUSTEIN *et al.* 2007; STANCHEV and STANCHEV 1984) (Figure 2.3.B). Some experiments have demonstrated that splicing is tightly coupled to transcription and at least some introns are excised while the nascent transcript is still associated with the polymerase through the action of snRNP and SR proteins associated to the c-terminal domain (CTD) of the RNA polymerase II (CHABOT *et al.* 1995; DAS *et al.* 2007; MORTILLARO *et al.* 1996; VINCENT *et al.* 1996). The majority of introns found in eukaryotes are removed using this U2-dependent process although~700 human introns rely on the U12-dependant spliceosome. The splicing process is very similar to what is described in Figure 2.3 and the major differences between these two types of introns reside in the donor splice site and branch point sequences. The U1, U2, U4 and U6 found in the U2-dependent spliceosome are replaced by four different snRNP proteins U11, U12, U4atac and U6atac in the U12-dependent mechanism (ALIOTO 2007; SHETH *et al.* 2006).

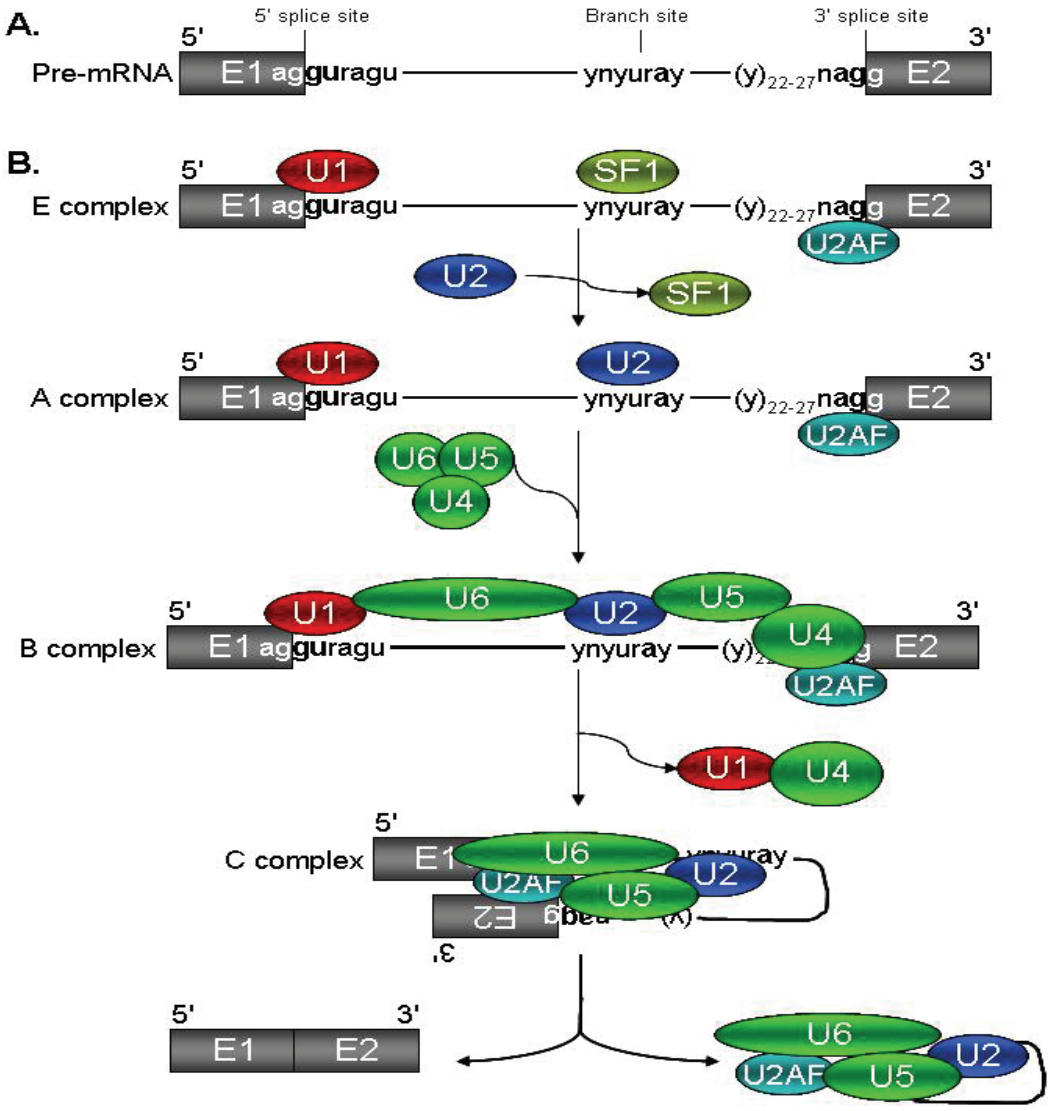


Figure 2.3: Schematic diagram of the spliceosome assembly at the splice site. A. Scheme of a typical intron flanked by exons in pre-mRNA. Cis-acting sequences that are relevant for the splicing reaction are shown for the 5' splice site, branch site and 3' splice site. The grey boxes represent exons and the line represents the intron sequence. B. Steps along spliceosome assembly. Schematic representation of spliceosomal complex E, A, B and C. See text for more details. This figure was modified from (BLAUSTEIN *et al.* 2007).

3'-End processing

Polyadenylation of pre-mRNA at the 3'-end is a vital step in transcription termination and pre-mRNA processing (WAHLE and RUEGSEGGER 1999). Almost all pre-mRNA in eukaryotes are polyadenylated with a few exceptions such as histone genes. (DAVILA LOPEZ and SAMUELSSON 2008). In humans, the pre-mRNA is recognized, cleaved, and then polyadenylated by a complex of enzymes (MANDEL *et al.* 2008) directed by distinct polyadenylation signal sequences present in the pre-mRNA transcript such as the highly conserved upstream AAUAAA sequence and a downstream G/U-rich sequence (BEAUDOING *et al.* 2000; GRABER *et al.* 1999; TIAN *et al.* 2005). PolyA tails have been shown to influence mRNA stability, translation and transport (JACOBSON and PELTZ 1996; LEWIS *et al.* 1995; WICKENS *et al.* 1997). In recent years, studies have shown the interconnection of other transcriptional and post-transcriptional processes (see above), such as splicing and transcriptional termination (MANIATIS and REED 2002).

Gene expression variation

Each cell contains, in its set of genomic loci, all the information required to make many thousand different RNA and protein molecules. However, a typical cell only expresses a subset of these genes because their identity and function, i.e. their phenotypes, is defined by the expression of specific genes in a spatial and temporal manner. To achieve this high level of diversity and precision, the cell regulates each step implicated in gene expression by (1) controlling when and what genes are transcribed (transcriptional control), (2) controlling how the RNA transcript is processed (RNA processing control), (3) selecting which mRNA will be exported and where in the cytoplasm (RNA transport and localization control), (4) controlling the stability of certain mRNA molecules in the cytoplasm (mRNA degradation control) (5) selecting which mRNAs are

translated (translational control), and (6) selectively controlling the activation, degradation and compartmentalization of specific proteins (protein activity control) (reviewed in (ALBERTS 2002). Although all these processes, in addition to others such as environmental signals, interact to form a complex network that coordinates gene expression, the following sections will deal with the regulation of transcription and pre-mRNA processing.

Transcript expression variation

The rate of gene transcription i.e., how many transcripts from a genomic locus are transcribed by the RNA polymerase for a given period, is a central parameter that controls cellular processes. The importance of this process was recognized 40 years ago (BRITTEN and DAVIDSON 1969), however it is only in this last decade that tools needed to study transcript expression variation at the genome-wide level, such as DNA microarrays have become available (see below). Studies using these tools have begun analyzing how environmental and genetic factors contribute to transcript expression variation.

Environmental factors

Organisms can modify the expression of specific genes in order to adapt their physiology to changing environmental conditions. For example, a study of Moroccans living in different environmental conditions (urban, mountain, desert) showed that ~37% of genes expressed in leukocyte samples had significantly different transcript expression levels. The authors of the study tested if this variation was due to genetic or epigenetic factors and found that environmental factors were the most likely cause (IDAGHDOUR *et al.* 2008). This type of environmental influence is well illustrated in another study of goby fish (*Gillichthys mirabilis*) exposed to multiple levels of heat stress (BUCKLEY *et al.* 2006). In this

study, temperature variations were shown to influence the transcript expression of genes responsible (i.e. chaperones) for the physiological adaptation to temperature changes. Interestingly, this expression variation was different between tissues, demonstrating another important aspect of transcript expression regulation, i.e. tissue identity and complexity is strongly defined by specific patterns of transcript expression. In fact, a study comparing the transcript expression profiles of 155 human tissues showed that gene expression was strongly correlated with anatomic locations, cellular compositions and physiologic functions (SHYAMSUNDAR *et al.* 2005).

Genetic factors

It has also been demonstrated that the evolution of an organism is achieved, in part, through changes in transcript expression regulation. The importance of regulatory mutations in the evolution of species was first proposed following the comparison of human and chimpanzee homologue proteins (KING and WILSON 1975). The authors concluded that the modest degree of divergence in homologous protein sequences could not account for the extensive phenotypic differences observed between these two closely related species and postulated that regulatory mutations must play an important role. An interesting example of this process is demonstrated by the comparison of different primate tissues (ENARD *et al.* 2002) where the authors showed that the transcript expression profiles for human brain had significantly diverged from the other primate species. Subsequent studies also found that ~10% of genes showed expression differences between humans and chimpanzees (CACERES *et al.* 2003; KHAITOVICH *et al.* 2005; KHAITOVICH *et al.* 2004). This indicates that some of the complex cognitive abilities found in humans and more generally other species-specific traits (ABZHANOV *et al.* 2004; CLARK *et al.* 2006; STERN 1998), are

the result of transcript expression regulatory variations caused by genetic changes that occurred between species.

Expression quantitative trait loci

Genetic variations present in regulators of transcript expression are also responsible for some of the transcript expression variation observed between individuals of the same population. Expression quantitative trait loci (eQTL) mapping (JANSEN and NAP 2001) is a popular approach to determine the polymorphism(s) or the genomic region containing the polymorphism that is partly responsible for variation of transcript expression regulation (CHEUNG *et al.* 2003; CHEUNG *et al.* 2005; DIXON *et al.* 2007; GORING *et al.* 2007; MORLEY *et al.* 2004; STRANGER *et al.* 2007a; STRANGER *et al.* 2007b). In these studies, gene expression levels are treated as quantitative traits and their genetic basis is studied using well-established linkage and association tools. Linkage mapping uses a study design that is based on tracking the transmission of alleles through families. This approach aims to identify genetic variations that are linked with transcript expression phenotypes (eQTLs) by tracking its transmission patterns through a pedigree. Association analysis uses samples of unrelated individuals to correlate marker genotypes with the eQTL (reviewed in (HIRSCHHORN and DALY 2005). Association analyses are usually more powerful than analyses using a linkage method because they are better at finding eQTLs with a medium to small effect size given a dense enough set of polymorphisms that are in linkage disequilibrium (LD) with the causative polymorphism (GILAD *et al.* 2008). In addition, this technique allows the fine mapping of the region with the causative polymorphisms which depends heavily on the haplotype structure around the eQTL. These types of studies have associated *cis*-acting eQTLs with many disease phenotypes such as resistance to infection with malaria (HAMBLIN and DI RIENZO 2000; TOURNAMILLE *et al.* 1995), risk of heart

disease (BEYZADE *et al.* 2003; YE *et al.* 1996), susceptibility to schizophrenia (HE *et al.* 2006) and many more (see review (WRAY 2007). It is apparent that regulation of transcript expression plays an essential role in gene expression. Moreover, in recent years, biologists have also begun studying how cells regulate the production of alternative transcript structures and the important role this process plays in gene expression.

Transcript structural variations

The ability of the metazoan cell to produce multiple mRNA transcripts from a single genomic locus was a key factor in their evolution. This allowed them to expand their transcriptomes and proteomes without increasing genome complexity, i.e. without increasing the number of genes. This increase in genetic coding potential was achieved by the evolution of specific regulatory processes involved in gene expression such as alternative transcription initiation, alternative splicing and alternative transcription termination.

Alternative transcript initiation

Transcription initiation is one of the first processes involved in regulating transcript expression. Regulation of mRNA synthesis depends heavily on the formation of the pre-initiation complex (see above) at the right time and at the right promoter. This temporal and spatial control relies on the intricate interplay between many transcription factors, *cis*-regulatory DNA elements, core promoter elements as well as chromatin remodelling and modifying factors to properly position the pre-initiation complex near the transcription start site of a genomic locus (LEMON and TJIAN 2000). In the past, genomic loci were thought to contain only one transcriptional start site. However, recent studies suggest that at least 50% of human genes use varying transcription start sites through the use of alternative core promoters (BAEK *et al.* 2007; COOPER *et al.* 2006; KIMURA *et al.* 2006;

TAKEDA *et al.* 2007). These alternative promoters allow a single genomic locus to produce a wide variety mRNA transcript and protein isoforms (Figure 2.4) in response to changing cellular conditions and states (e.g., differentiation, growth and development).

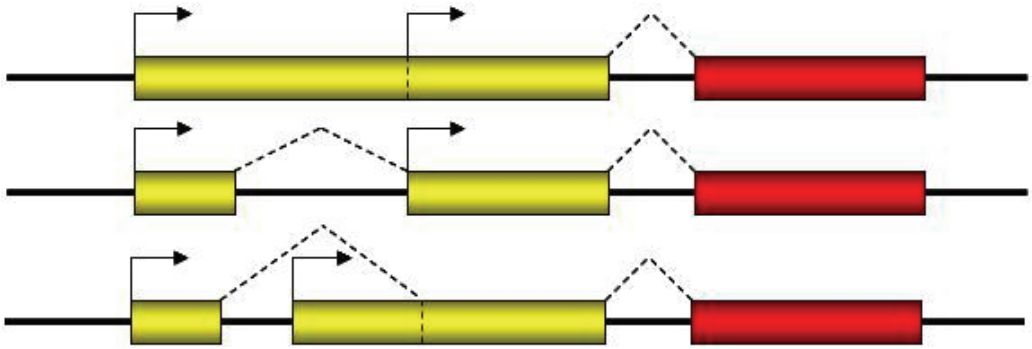


Figure 2.4: Processes that generate alternative transcript initiation. Two promoters on a single exon (top); alternative first exon (middle); and a downstream promoter located within the intron region of another isoform (bottom).

The exact molecular mechanisms responsible for alternative transcription start sites are still not clearly understood. Some mechanisms have been proposed such as the presence of multiple core-promoter structures, variable concentrations of *cis*-regulatory elements and factors, and epigenetic changes in the promoter region (reviewed in (DAVULURI *et al.* 2008). Alternative transcription initiation can result in the production of distinct mRNA isoforms with different 5' untranslated regions (5'-UTR). The 5'-UTRs contain sequences that regulate mRNA stability and translational efficiency such as sequences responsible for mRNA secondary structure and translational initiation sites. Therefore, certain types of alternative transcription initiation can affect these processes without affecting the protein coding potential of the mRNA by only varying the 5'-UTR sequence (DAVULURI *et al.* 2008). Other types of alternative transcription initiation affect the protein coding structure if alternative

translation start sites are included in the pre-mRNA transcript. This might affect protein domains that are important for different biological activities and consequently this diversifies protein functions. Therefore, transcription initiation variation is quite common between different tissue types (BIRNEY *et al.* 2007). Moreover, aberrant alternative transcription initiation has been associated to a number of diseases (LIU *et al.* 2005; MARCU *et al.* 1992; NAKANISHI *et al.* 2006; SUN *et al.* 2007). More recent studies have shown that genetic variations were linked to alternative promoter usage. For example, regional rearrangements (insertions and inversions) in the promoter region of the aromatase (*CYP19A1*) gene increase its expression which is associated with a higher incidence of breast cancer (DEMURA *et al.* 2007). As demonstrated in this example, alternative promoter usage can, in addition to modifying the transcript structure and stability, affect the transcription level of a gene.

Alternative splicing

A typical human gene contains, on average, 8.8 short exonic sequences with a mean size of 145 bp. These exons are usually separated by much larger intron sequences that on average, account for >90% of the pre-mRNA transcript (LANDER *et al.* 2001). The maturation of pre-mRNA into mRNA involves the removal of these intron sequences and the joining of the exon sequences. This process called constitutive splicing is catalyzed by the large ribonucleoprotein complex known as the spliceosome that interacts with the splicing signals (see above). In 1978, Gilbert (GILBERT 1978) proposed that through regulation, splicing could produce multiple mRNA isoforms from the same pre-mRNA transcript by alternatively splicing out specific exons. A few years later, his theory of alternative splicing was validated (EARLY *et al.* 1980; ROSENFELD *et al.* 1982) and more recently it was estimated that almost all mammalian genes (~95%) undergo some form of alternative splicing (PAN *et al.* 2008; WANG *et al.*

2008). This high propensity of alternative splicing is theorized to offset the low level of genome complexity of higher eukaryotes. For example, the *drosophila DSCAM* gene can theoretically produce more than 38,000 different mRNA isoforms (SCHMUCKER *et al.* 2000), which is far superior than the total number of genomic loci in all of its genome (CLARK *et al.* 2007a). Changes in the splicing patterns of different tissues have been proposed as important mechanisms for species evolution. In closely related species such as humans and chimpanzees, the expression profiles for ~1000 orthologous exons from different tissues were compared and this revealed that alternative splicing patterns from brain had the highest level of divergence (CALARCO *et al.* 2007). In addition, comparisons of human and mouse transcripts have revealed that less than 20% of alternative splicing events were conserved between these species (MODREK and LEE 2003; PAN *et al.* 2005; YEO *et al.* 2005). This demonstrates that splicing variation is partially responsible for some of the species specific phenotypes. This observation also holds when comparing organs from the same species. Complex organs comprised of specialized cell types such as brain and liver present more splicing variation than simpler tissues such as kidney and skeletal muscle (JOHNSON *et al.* 2003; XU *et al.* 2002; YEO *et al.* 2004a). The various mechanisms responsible for producing mRNA isoform variation through alternative splicing are illustrated in Figure 2.5.

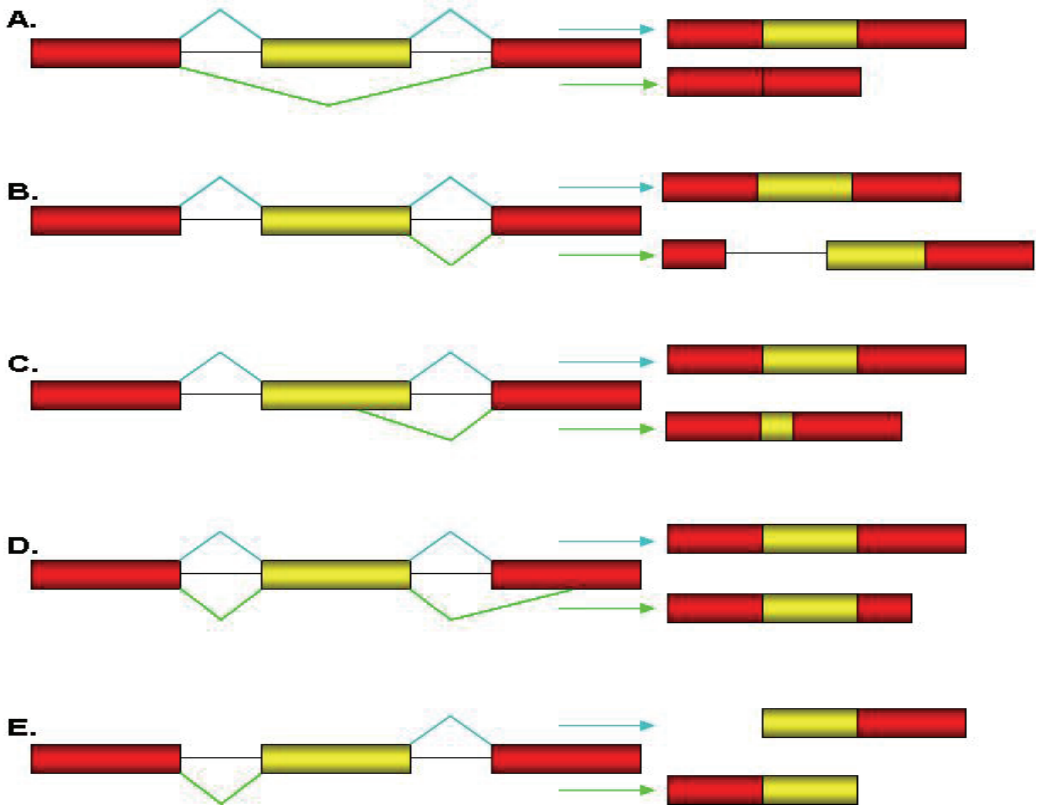


Figure 2.5: Common types of alternative splicing events. A. This represents an alternatively spliced exon where an exon is either included or excluded from the mRNA transcript. B. In this case the intron is not spliced out from the mRNA transcript therefore is annotated as an intron retention event. C. and D. are example of alternative 5' and 3' splice site usage, respectively. Here a second splice site found either in the exon or the intron is used to define the exon boundaries. E. The red exons in this example are mutually exclusive, i.e. when one is included in the mRNA transcript the other is excluded. In many cases, these common mechanisms are combined to generate more complicated alternative splicing events. This figure is modified from (WANG and BURGE 2008).

These events are controlled through the interaction of the spliceosome and specific *cis*-regulatory elements that serve as either splicing enhancers or silencers (Figure 2.6). Elements found in an exon that promote or inhibit its inclusion are respectively classified as exonic splicing

enhancers (ESE) and silencers (ESS). Elements found in an intron that enhance or inhibit the use of adjacent splice sites are known as intronic splicing enhancers (ISEs) and silencers (ISSs), respectively. These splicing regulatory elements promote or inhibit the recruitment of splicing factors by activating or suppressing the recognition of splice sites or by regulating the assembly of the spliceosome (MATLIN *et al.* 2005). Therefore, splicing decisions result from differences in the concentration and/or activity of these proteins. Splicing regulatory elements that enhance splicing are expected to play a predominant role in constitutive splicing while alternative splicing is principally controlled by silencing elements.

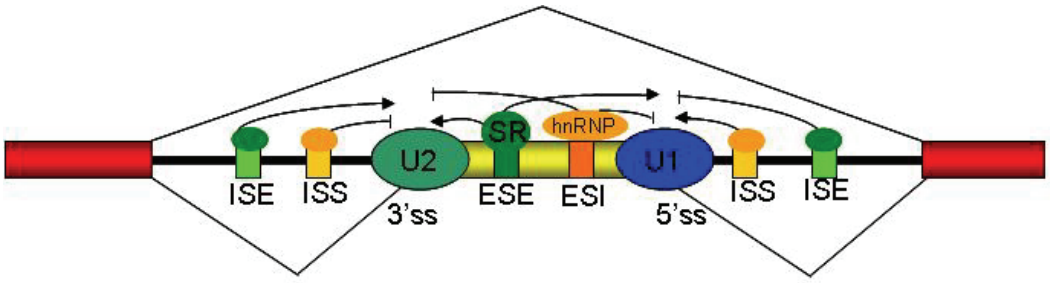


Figure 2.6: A schematic of two alternative splicing pathways for the middle exon. This illustrates the interaction of cis-splicing regulatory elements (ESE, ESS, ISS, and ISE) with trans-splicing factors (hnRNP and SR proteins) that together enhance or inhibit the recruitment of spliceosome proteins (U2 and U1) which leads to the inclusion or exclusion of the middle exon from the mature mRNA transcript. This figure was modified from (WANG and BURGE 2008).

Sequence changes in these splicing regulatory elements can lead to disease phenotypes. A very conservative estimate suggests that at least 15% of point mutations that cause human disease affect splicing (CHEN *et al.* 2003). Spinal muscular atrophy is an example of a recessive disease that is caused by a point mutation in an exonic regulatory element. A C→T

mutation in the *SMN2* gene causes the missplicing of exon 7 which creates a non-functional protein that leads to the disease phenotype (reviewed in (WIRTH *et al.* 2006). Hutchinson–Gilford progeria syndrome is associated with premature aging and is another example of a disease caused by a point mutation but this time in an intron. This mutation activates a cryptic splice site in the Lamin A gene that truncates that last 150 base pairs of exon 11 from the Lamin A gene (DE SANDRE-GIOVANNOLI and LEVY 2006). Mutations can also disrupt proteins belonging to the spliceosome and therefore affect the splicing of multiple exons and consequently create many disease phenotypes. For example, a mutation in the TDP43 splicing factor belonging to the hnRNP family of proteins has been implicated in a number of diseases such as cystic fibrosis (BURATTI *et al.* 2001), frontotemporal lobar degeneration and Lou Gehrig's disease (NEUMANN *et al.* 2006). Splicing variations have also been implicated in different cancers (reviewed in (VENABLES 2006).

Another interesting characteristic of alternative splicing is its ability to regulate transcript expression. It is estimated that approximately ~65% of alternative splicing events occur within the translated regions of mRNA transcripts (GUPTA *et al.* 2004). A splicing event that introduces a premature stop codon in a mRNA transcript is subject to the non-sense mediated decay (NMD) surveillance system (BELGRADER *et al.* 1994). This system recognizes mRNA isoforms containing premature stop codons that are subsequently targeted for degradation. In a study of more than 3000 alternatively spliced human genes, it was shown that 35% of the mRNA isoforms produced contained a premature stop codon and that 75% of these isoforms were degraded by the non-sense mediated decay system (LEWIS *et al.* 2003). Thus, alternative splicing and NMD act together to play an important role in regulating gene expression.

Alternative polyadenylation

Gene expression is also influenced by other types of mRNA structural variation such as alternative polyadenylation. The vast majority of eukaryotic mRNA transcripts are polyadenylated, i.e. they acquire a poly(A) tail at their 3' ends (reviewed in (EDMONDS 2002). Polyadenylation involves a two step process where the pre-mRNA transcript is cleaved and then adenosine (A) residues are added at the 3' end. This process is controlled by core polyadenylation elements as well as auxiliary elements found upstream and downstream of the consensus polyadenylation sequence that interact with the cleavage and polyadenylation machinery (see above). In recent years, studies have demonstrated that genes can contain multiple polyadenylation sites (reviewed in (LUTZ 2008). Recently, it has been estimated that around 50% of human genes are alternatively polyadenylated (TIAN *et al.* 2005). Alternative polyadenylation can create mRNA transcript isoforms that have varying 3' UTR lengths and coupled with alternative splicing can alter the translation region (Figure 2.7).

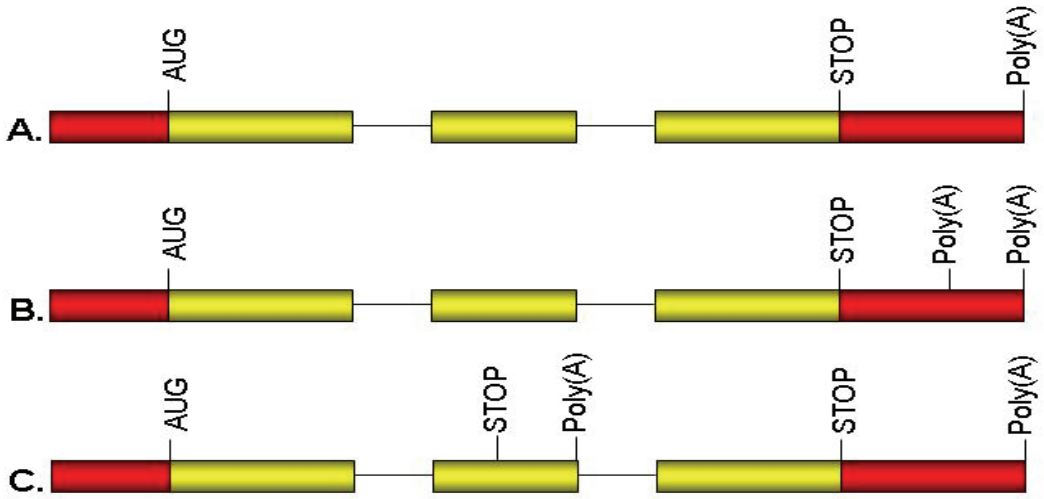


Figure 2.7: Types of alternative polyadenylation. A. This is an example of constitutive polyadenylation because only one polyadenylation site is present in the 3' UTR. B. Example of alternative polyadenylation with multiple polyadenylation sites in the present on the last exon of the 3' UTR. C. This demonstrates an alternative splicing event (last exon is

skipped) coupled with the use of different polyadenylation sites. This figure is modified from (LUTZ 2008).

These types of variation can influence protein coding potential, transcript localization, stability and transport (LEWIS *et al.* 1995; WICKENS *et al.* 1997). Therefore polyadenylation is an important aspect of gene expression. Similar to what was mentioned for transcript expression, alternative transcription initiation and alternative splicing, regulation of alternative polyadenylation varies between tissues (BEAUDOING and GAUTHERET 2001; RIGAULT *et al.* 2006) in response to different developmental or functional cues and has been implicated in evolution (ARA *et al.* 2006) and certain disease phenotypes (DANCKWARDT *et al.* 2008) .

Profiling gene expression

The mRNA population of a cell specifies its identity and helps govern its present and future activities (see above). This has made the efficient analysis of the transcriptome an important aspect in the field of molecular biology. Over the past 30 years, many technological advances have facilitated the study of gene expression. The first technologies developed to study gene expression were the Northern Blot (ALWINE *et al.* 1977) and reverse transcription-polymerase reaction (RT-PCR) (MULLIS *et al.* 1986). These approaches were useful for analysing expression of a small number of genes, however could not be easily scaled up for studies of a large number of genes in many tissues. Thus, higher throughput methods were needed to capture the whole complexity of the transcriptome. High-throughput methods such as expressed sequence tags (EST) (BOGUSKI *et al.* 1994) and serial analysis of gene expression (SAGE) (VELCULESCU *et al.* 1995) were developed to measure gene expression in a multiplex manner. The method relied on sequencing cloned mRNAs and mapping

them back to genomic sequence to identify the genes expressed in cells. These techniques were limited by time and cost constraints as well as by biases that affect coverage and sampling. These limitations led to the development and broad distribution of a technology known as the DNA microarray (AUGENLICHT *et al.* 1987; POUSTKA *et al.* 1986; SCHENA *et al.* 1995).

Microarray applications

The DNA microarray was developed using the concept of the Northern Blots. As with Northern Blots, DNA microarrays are used to measure the abundance of specific nucleic acid sequences in a given sample, except that the DNA microarray does this in a multiplex manner. It uses a collection of probes made of DNA sequences of varying length that are ordered and bound onto the surface of a solid support such as glass. These probes are designed to bind specific targets that consist of fluorescently labelled nucleic acid sequence (cRNA or cDNA). The level of binding between a probe and its target is quantified by measuring the fluorescence emitted by the hybridized target when scanned and corresponds to the abundance of the target. This concept was applied to a variety of DNA microarray designs to study a broad range of nucleic acid variations. They are mainly used for gene expression analysis and screening samples for single nucleotide polymorphisms (genotyping) (HACIA *et al.* 1999). Although in recent years, DNA microarrays have also been used in other application such as ChIP-on-chip experiments (IYER *et al.* 2001; LIEB *et al.* 2001; REN *et al.* 2000), epigenetic studies (ADORJAN *et al.* 2002; HUANG *et al.* 1999; YAN *et al.* 2001) and DNA-mapping (MORAN *et al.* 2004; POLLACK *et al.* 1999). More recently, with advances in manufacturing techniques, DNA microarrays are now used to study gene expression at the sub-transcript level. In fact, expression of individual mRNA isoforms produced by transcriptional and pre-mRNA processing

variations such as alternative transcript initiation and termination as well as alternative splicing can now be assessed with the use of different alternative splicing microarrays. These alternative splicing arrays target their probes to each exon and/or exon junction within a gene to determine mRNA levels at the resolution of a single exon or splice site.

Isoform level detection microarrays

The first attempt at using microarrays to study alternative splicing was explored using the multi-probe design of the Affymetrix Gene Chip (HU *et al.* 2001). In this study, the Affymetrix Gene Chip probes that are usually summarized together into one measure of whole-transcript expression were instead used to measure the expression of individual exons they targeted. This study demonstrated that gene expression at the isoform level could be measured by targeting probes to individual exons within a transcript. This lead to the manufacturing of the first custom microarray designed to measure gene expression variation at the isoform levels by using a mix of exon-body and junction probes (WANG *et al.* 2003). The first high-throughput analyses of alternative splicing (JOHNSON *et al.* 2003; PAN *et al.* 2005; PAN *et al.* 2004) were conducted with custom arrays and measured global alternative splicing patterns in different tissues and species. However, the gene coverage of these custom arrays was insufficient to cover every possible exon in the genome. This prompted the microarray manufacturing company Affymetrix Inc. to design the first truly genome-wide alternative splicing DNA microarray known as the GeneChip® Human Exon 1.0 ST Array.

Human Exon array

The GeneChip® Human Exon 1.0 ST Arrays are constructed using a patented photolithographic process borrowed from the computer chip industry. Probes are synthesized on a wafer slide using photolithographic

masks for selective location activation followed by the addition of the base to the activated site (see affymetrix.com for details). This process produces extremely dense arrays that are composed of 5.5 million 25-mer probes that in turn enable genome-wide analyses of gene expression at the isoform level. Probes target individual exons or portions of an exon when prior evidence of alternative splicing exists. Each exon within a gene is targeted on average by 4 probes (Figure 2.8) which allows the simultaneous exon-level detection of expression intensity for 1.4 million probe sets covering over 1 million known and predicted human exons. Probe sets on the array are divided into 3 levels of annotation: core, extended and full. The core probe sets target ~284,000 exons supported by RefSeq and GenBank. The extended and full annotations are based on less confident annotated exons, with evidence from ESTs and computationally predicted exons. These last two annotation levels are designed to identify novel transcript variants while the core probe set are used for straightforward studies of gene expression variation at the isoform level given the reduced size and high confidence annotation data set they produce (SIEPEL *et al.* 2007).

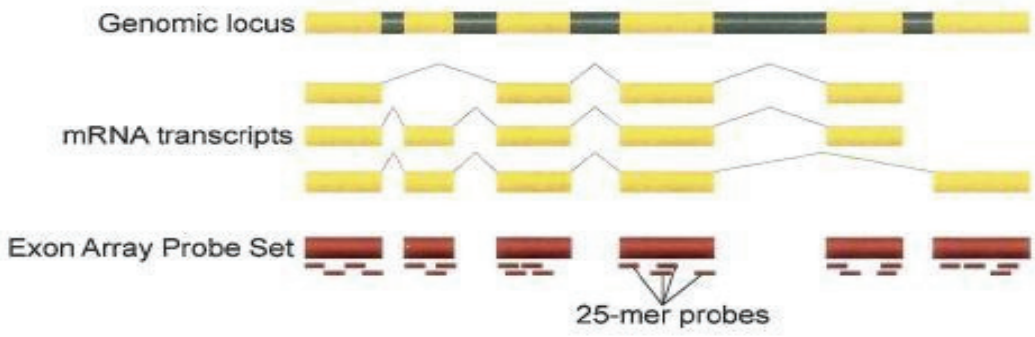


Figure 2.8: Schematic for coverage of probe sets across a gene. Yellow regions are exons and grey regions represent introns. The short dashes below the exon regions (red) indicate individual probes of 25 nucleotides in length and represent a probe set.

Workflow for Exon Array analysis

The biggest challenge of studying transcript isoform variations using microarrays is how to analyse and interpret data generated in these experiments. In the past, whole-gene expression studies using microarrays were based on the dogma that one gene is transcribed to one transcript that is subsequently translated to one protein. Now, the scenario has changed to one gene that produces multiple products. This extra level of complexity has created several problems in microarray analysis that must be solved. For instance, a given probe can represent the sum of intensities from multiple isoforms of a gene and at the same time that probe along with others represent expression for that one gene. Therefore, new analysis method are needed to decouple signals coming from changes in pre-mRNA processing such as variation of alternative splicing from changes in overall gene expression to adequately assess gene expression at the isoform level (CUPERLOVIC-CULF *et al.* 2006). In addition, Exon Array data consists of very noisy signal measurements. The true expression signal is buried by different sources of noise, such as poor sample preparation, labelling, hybridization and many more (ZAKHARKIN *et al.* 2005). Therefore, pre-existing analysis pipelines developed for standard gene expression microarray experiments such as quality assessments, data normalization, detection of differential expression and annotation of differentially expressed isoforms must be adjusted to accommodate this type of data (Figure 2.9).

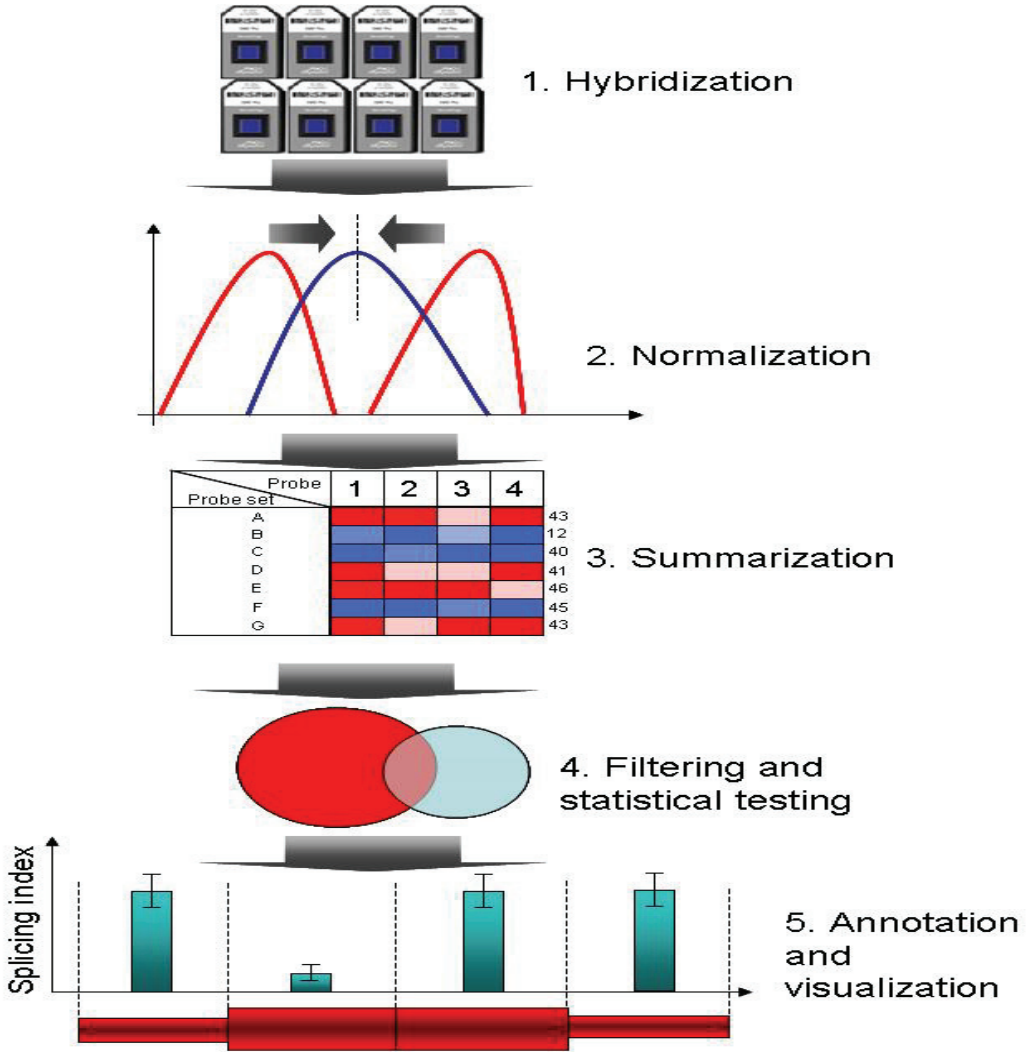


Figure 2.9: Exon array analysis workflow. 1) The first step is preparation and hybridization of cDNA extracted from sample for analysis on the Exon array. 2) Data acquisition and normalization to remove noise and technical biases. 3) Summarization of probe signals into probe set (exon) and meta-probe set (gene) expression scores. 4) Data filtering to remove misbehaving probe set and meta-probe sets folded to statistical testing to identify varying exons or genes within samples. 5) Mapping of interesting probe sets to known and predicted transcript structures overlaid with expression data to determine isoform expression. Image modified from (OKONIEWSKI and MILLER 2008).

Quality control

Quality assessment is an essential first step in the analysis of Exon Array data because it can identify noisy samples due to issues related to RNA quality, probe labelling, hybridization, washing and signal/background detection in the scanning process. During the summarization step (see below) of Exon Array data, a quality report is generated by the Affymetrix Power tool software (affymetrix.com) where summary metrics such as mean probe set intensity for each sample, the number of expressed probe sets per sample (DABG see below) and others are computed to identify outliers samples (see affymetrix.com; Quality Assessment of Exon Arrays). In addition, a principal components analysis (PCA) plot is a tool that is commonly used to identify outliers in a group of samples (DE HAAN *et al.* 2007). The decision regarding which samples is an outlier depends heavily on the experience of the user and varies on a case by case basis. These outlier samples could be flagged and excluded from the analysis or the analysis could be adjusted to account for the outlier by down-weighting it.

Normalization

The next step in the analysis pipeline is normalization. This procedure is essential to reduce noisy microarray data. Many techniques have been developed such as standardization (Z-score), housekeeping gene based normalization and equalized quantile normalization (AUTIO *et al.* 2009). Most of these techniques rely on the assumption that the majority of exon or gene expression is unchanged between samples therefore they attempt to make each sample in a data set have the same probe signal distribution. For example, quantile-normalization is a non-parametric procedure that first consists of constructing quantiles (ranks) for the probe signals on each array individually. The median probe signal in each quantile is then computed across all arrays. That median value now

represents the normalized signal value for each probe of that given quantile (BOLSTAD *et al.* 2003). This type of normalization procedure ensures that all arrays in an experiment have the same median and standard deviation of probe signals and therefore removes some of the high variability and biases introduced by technical artefacts.

Expression summarization

For the Exon array data, expression summarization is the process of combining specific probe signal values into probe set (exon) and meta-probe set (transcript) expression scores. The most popular summarization algorithms are RMA (robust multichip average) (IRIZARRY *et al.* 2003a) and PLIER (see affymetrix.com - Gene signal estimates from exon arrays). Essentially, these algorithms determine the expression level of a probe set or a meta-probe set by performing a type of weighted average and background correction (see below) of probe intensities.

Background correction

The Human Exon Array implements a new system to estimate background noise levels. Instead of using mismatch probes, as was typically the case for earlier Affymetrix designs, they include a collection of probes, called antigenomic probes that have no target in mammalian transcriptomes. The signal intensities of the antigenomic probes mostly originate from non-specific binding which is a function of their GC-content. Therefore, before summarizing probe set and meta-probe set expression scores, probe signals are corrected by subtracting the median non-specific binding signals computed from the distribution of antigenomic probes of the same GC-content. In addition, instead of the classical presence / absent calls used to establish if a gene was expressed in a sample, a new metric called the detected above background (DABG) is computed for each probe set and meta-probe set. This metric represents the probability that the expression of a given probe set or meta-probe set is background noise

and therefore not expressed (www.affymetrix.com; Exon array background Correction). A threshold is usually set (DABG < 0.05) to determine if a probe set or meta-probe set is expressed in a given sample.

Identification of variation

To compare variation of gene expression at the isoform level between a set of samples, the most straightforward method to use is the splicing index (SRINIVASAN *et al.* 2005). The splicing index is a conceptually simple algorithm that aims to identify probe sets (exons) that have different inclusion rates between two sample groups (affymetrix.com - Identifying and Validating Alternative Splicing Events). The Splicing index is first computed as the value of probe set intensity relative to the meta-probe set intensity in a given sample on a $\log_2$ scale ($Intensity_{probe\ set}$ / $Intensity_{meta\text{-}probe\ set}$). Then the normalized intensities (NI) from each group are divided between each other ($NI_{sample1}$ / $NI_{sample2}$) which represent the splicing index. A splicing index of 0 (log2 scale) indicates equal inclusion rates of the exon between both samples, a positive value indicates a skipping of that exon in sample 2 and a negative value indicates skipping of that probe set in sample 1. To identify probe sets that present statistically significant differences between two groups, a statistical test such as the Student's t-test or the analysis of variance (ANOVA) is used on gene-level normalized exon intensities (NI). The splicing index for every exon within a transcript is usually observed in a graphical representation overlaid with the p-value from the statistical test to identify isoform variations between samples (see below).

An issue with statistical testing in microarray experiments is multiple testing. These types of experiment present a challenge because thousands or millions, in the case of the Exon array (1.4 million) of statistical tests are performed and the false positive rate must be

controlled in order minimize the false positive results. Multiple testing corrections aims to address this issue by restricting the stringency threshold ($\alpha = 0.05$) to reduce the false positive rate with as little affect as possible on the number of incorrect rejections of real results, i.e. false negatives. The false discovery rate (FDR) correction (BENJAMINI and HOCHBERG 1995) is a popular strategy that consists of finding a threshold where the number of expected false positives is known. One issue with multiple testing corrections on Exon array data is the violation of certain assumptions such as non-independence of probe sets that makes it difficult to accurately compute the stringency threshold (AICKIN and GENSLER 1996; BENDER and LANGE 2001). However, these techniques can still be used to identify sizable data sets of isoform variation.

Filters

Other strategies, in addition to multiple testing corrections should be used to reduce the false positive rate in Exon array experiments. Filtering of signal data is very important in these types of experiments because it reduces the laborious steps of validating (e.g. by RT-PCR) false results. For analysis of Exon array data at the isoform level, removing all genes that are not expressed in all samples or excluding probe sets that are not expressed in at least one sample are mandatory filtering steps (affymetrix.com; Identifying and Validating Alternative Splicing Events). To date, literature on the filtering criteria is quite poor and new methods need to be developed in order to clean up Exon array data.

Annotation mapping and visualization

Once interesting probe sets have been identified by filtering and statistical testing, the next step is to map the probe sets to their respective exons and genes. The main aim of mapping is to identify genes that are differentially expressed or that present isoform variation in the form of

exon skipping, alternative initiation or termination. This is usually achieved using the standard definition files provided by Affymetrix. These files contain three levels of annotation: core, extended and full (see above). A decision on what level of annotation the analysis will be conducted on can significantly influence its outcome. Studies focusing on core probe sets will deal with high confidence annotated exons whereas the extended and full annotations are used for discovering new genes and exons given the predictive nature of these two levels of annotation. The use of these different annotation levels will also influence multiple testing correction procedures because use of the smaller data sets such as the core will have a beneficial effect on the false discovery rate. Data visualization is the last *in-silico* procedure in the analysis workflow before *in-vitro* validation gene or isoform variations. In-house, open access (XMAP, Integrated Genome Browser, Expression Console) or commercially (Gene sifter) available visualization tools should be used to overlay expression data onto gene structure in order to identify gene expression variation at the isoform level.

Summary of the literature review

This literature review demonstrates the important roles that transcription and pre-mRNA processing play in the regulation of gene expression. These processes work in concert to dictate the quantity and the type of mRNA isoform a gene will produce. The regulatory evolution of these processes has enabled organisms to expand their transcriptomes and proteomes without having to increase the complexity of their genomes. Higher eukaryotes use these processes to create distinct cellular phenotypes that in turn have enabled the development of specialized tissues. Therefore, regulatory disruption of these processes can lead to disease phenotypes. This has prompted the scientific community to

develop methods and tools to explore variation of transcript expression at the isoform level.

Chapter 3: Heritability of alternative splicing in the human genome

Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, David Serre, Harry Zuzan, Tyson A. Clark, Anthony Schweitzer, Michelle K. Staples, Hui Wang, John E. Blume, Thomas J. Hudson, Rob Sladek, and Jacek Majewski.

This chapter is published in *Genome Research* on May 31st, 2007. 17: 1210-1218.

Connecting text

It has recently been shown that variation in whole-transcript expression is under genetic control in human populations and is responsible for phenotypic variation and susceptibility to certain complex diseases (see literature review). However, our understanding of how variable transcript expression is at the isoform level is still poorly understood. Despite a few isolated examples no study has evaluated the prevalence and potential impact of these variations at the genome-wide level. This chapter represents a pilot study that we conducted in order to evaluate the performance of the Human Exon array in detecting transcript isoform differences such as alternative initiation, splicing and termination as well as whole-transcript expression differences among humans.

Abstract

Alternative pre-mRNA splicing increases proteomic diversity and provides a potential mechanism underlying both phenotypic diversity and susceptibility to genetic disorders in human populations. To investigate the variation in splicing among humans on a genome-wide scale, we use a comprehensive exon-targeted microarray to examine alternative splicing in lymphoblastoid cell lines (LCLs) derived from the CEPH HapMap population. We show the identification of transcripts containing sequence verified exon skipping, intron retention, and cryptic splice site usage that are specific between individuals. A number of novel alternative splicing events with no previous annotations in either RefSeq or EST databases were identified, indicating that we are able to discover de novo splicing events. Using family-based linkage analysis, we demonstrate Mendelian inheritance and segregation of specific splice isoforms with regulatory haplotypes for three genes: *OAS1*, *CAST*, and *CRTAP*. Allelic association was further used to identify individual SNPs or regulatory haplotype blocks linked to the alternative splicing event, taking advantage of the high-resolution genotype information from the CEPH HapMap population. In one candidate, we identified a regulatory polymorphism that disrupts a 5′ splice site of an exon in the *CAST* gene, resulting in its exclusion in the mutant allele. This report illustrates that our approach can detect both annotated and novel alternatively spliced variants, and that such variation among individuals is heritable and genetically controlled.

Introduction

The human genome is estimated to contain ~20,000–25,000 genes, and recent studies suggest that ~50%–75% of multi-exon genes undergo alternative splicing (AS), generating multiple mRNA isoforms and greatly increasing human proteomic diversity (LANDER *et al.* 2001; MODREK *et al.* 2001). The splicing of mRNA is a highly regulated process involving the

interactions of *trans*-acting splicing factors and *cis*-acting regulatory motifs. Disruptions of this process through mutations within these factors and regulatory signals may play an important role in phenotypic diversity and genetic disorders (BLACK and GRAVELEY 2006; FAUSTINO and COOPER 2003; NISSIM-RAFINIA and KEREM 2005).

Recent advances in microarray technology hold great promise for the genome-wide detection of AS events (LEE and ROY 2004). Small to large-scale microarrays have been designed using probes spanning predicted exon junctions (JOHNSON *et al.* 2003; MODREK *et al.* 2001; SUGNET *et al.* 2006; ULE *et al.* 2005; ZHANG *et al.* 2006), probes targeted toward individual exons (FREY *et al.* 2005), or a combination thereof (SRINIVASAN *et al.* 2005) and applied to identification of AS events that are tissue-specific, for the most part. However, one caveat of these studies utilizing customized arrays is a bias toward genes with solid EST and cDNA evidence for known AS events and that are therefore limited in their usefulness as a discovery tool for de novo splicing events. Here, we have chosen to use an alternative array design, the Affymetrix GeneChip Human Exon 1.0 ST Array, which is less biased toward known AS events by targeting multiple probes to individual exons and allowing simultaneous, exon-level detection of expression levels for 1.4 million probe sets covering over one million known and predicted human exons (Figure 3.1). Exon-tiling arrays have several advantages over exon-junction arrays: flexibility of probe placement, exact transcript structures do not need to be known a priori, and most AS events can be monitored without designing probes specific to all possible junctions. However, it should be noted that exon arrays do not provide immediate information on transcript structures containing candidate alternative events.

We show that (1) the Exon Array is able to detect AS at a level that is comparable in sensitivity as other microarray methods, and (2) we can identify quantitative and qualitative variations in splicing among individuals. Preliminary analysis estimates that up to 5% of all RefSeq exons are differentially spliced between individuals. Our approach for establishing a genetic basis for the variation in splicing uses lymphoblats derived from individuals of the CEPH population (COHEN *et al.* 1993), where we take advantage of the high resolution HapMap genotype information from these samples (ALTSHULER *et al.* 2005) to perform allelic association studies.

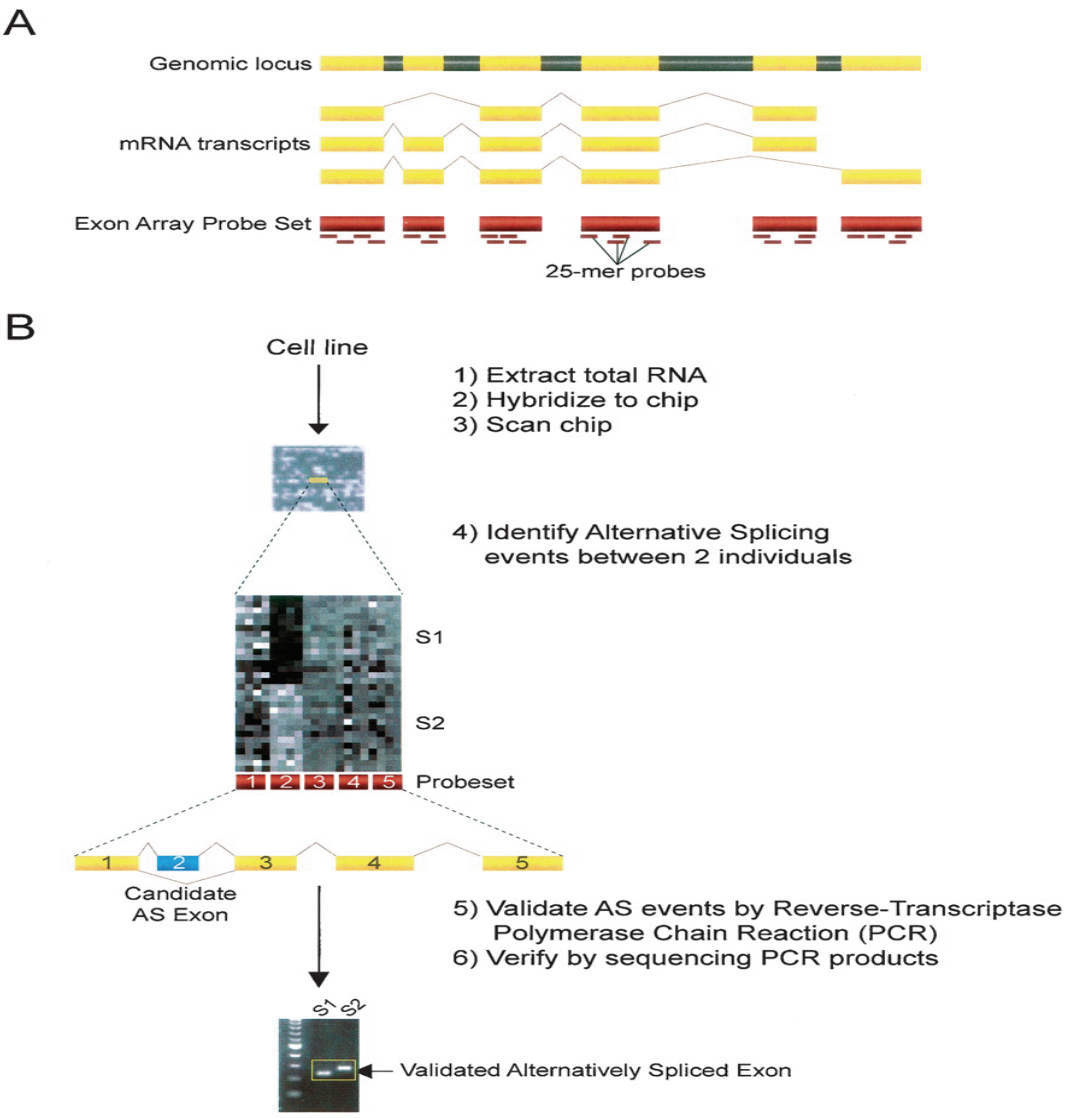


Figure 3.1: (*A*) Schematic for coverage of probe sets across the entire length of the transcript. Yellow regions are exons, whereas grey regions represent introns. The short dashes underneath the exon regions indicate individual probes of 25 nucleotides in length representing the probe set. The Affymetrix GeneChip Human Exon 1.0 ST Array allows for exon-level expression profiling in a single chip, and can interrogate over one million predicted exons within the human genome. (*B*) Flowchart for processing and analysis of chips to validation of alternative splicing events. Total RNA is extracted from the two cell lines (n = 15 replicates per individual) and is transcribed to cDNA and labeled with biotin. The total cDNA is then hybridized to the exon chip, followed by washing and staining with an anti-

streptavidin antibody. Chips are then scanned, and hybridization data are processed and analyzed by the Affymetrix Power Tools (version 1.6) software package. A splicing index is calculated for ~1.4 million probe sets covering one million exons. A subset of 20 alternative splicing events predicted between the two individuals using an unpaired t-test ($P < 8.915 \times 10^{-4}$) on the splicing index and other criteria (see Methods), are then validated by (1) RT-PCR using exon body primers flanking the probe set of interest and (2) sequencing of the RT-PCR products.

Methods

Cell line preparation

RNA samples were obtained from 74 Epstein-Barr virus-transformed LCLs belonging to the CEPH (Center d'étude du polymorphisme humain) reference individuals from the state of Utah in the United States (CEU). For this study, we used DNA samples from 60 unrelated individuals that have been genotyped for approximately four million SNPs by the International HapMap Project (ALTSHULER *et al.* 2005). Additionally, LCLs from CEPH pedigree 1444 (14 samples) were included to examine genetic influences on AS in a three-generation family. Cells were grown at 37°C and 5% CO_2 in RPMI 1640 medium (Invitrogen) supplemented with 15% heat-inactivated fetal bovine serum (Sigma-Aldrich), 2 mM L-glutamine (Invitrogen), and penicillin/streptomycin (Invitrogen). Cell growth was monitored with a hemocytometer, and cells were harvested at a density of 0.8×10^6 to 1.1×10^6 cells/mL. Cells were then resuspended and lysed in TRIzol reagent (Invitrogen). For all LCLs, three successive growths were performed (corresponding to the second, fourth, and sixth passages) after thawing frozen cell aliquots.

Affymetrix exon arrays

RNA was isolated using TRIzol reagent following the manufacturer's instructions (Invitrogen). The RNA quality was assessed using RNA 6000 NanoChips with the Agilent 2100 Bioanalyzer (Agilent). Biotin-labeled target for the microarray experiment were prepared using 1 μg of total RNA. The RNA was subjected to a rRNA removal procedure with the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen), and cDNA was synthesized using the GeneChip WT (Whole Transcript) Sense Target Labeling and Control Reagents kit as described by the manufacturer (Affymetrix). The sense cDNA was then fragmented by UDG (uracil DNA glycosylase) and APE 1 (apurinic/apyrimidic endonuclease 1) and biotin-labeled with TdT (terminal deoxynucleotidyl transferase) using the GeneChip WT Terminal labeling kit (Affymetrix). Hybridization was performed using 5 μg of biotinylated target, which was incubated with the GeneChip Human Exon 1.0 ST array (Affymetrix) at 45°C for 16–20 h. Following hybridization, nonspecifically bound material was removed by washing and detection of specifically bound target was performed using the GeneChip Hybridization, Wash and Stain kit, and the GeneChip Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip Scanner 3000 7G (Affymetrix), and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix).

For the initial study, three separate passages of two unrelated individuals, GM12750 and GM12751, from the CEPH 1444 pedigree were used, with five technical replicates of each growth, for a total of 15 arrays hybridized for each sample. Multiple replicates were used to assess the relative contributions of biological and technical noise to the observed exon and transcript levels. In particular, since this array uses probe cells with a feature size that is only one-quarter of previous expression array designs,

we aimed to determine whether they showed greater technical variability or higher background noise and also to identify a minimum number of biological and technical replicates required for an acceptable signal-to-noise ratio. For the linkage studies of the CEPH 1444 pedigree, three passages for each of GM12739, GM12740, GM12750, and GM12751 were used along with single replicates for the remaining 10 individuals.

Analysis of array hybridization data

The Affymetrix Power Tools software package (Affymetrix) was used to quantile normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta-probe set (representing gene expression) intensities using a probe logarithmic intensity error model (see http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf). Probe sequences that map to SNPs in a particular sample may give rise to altered binding affinities and influence intensity data and the resulting SI scores (data not shown); therefore, probe sets were cross-referenced to the dbSNP database (release 126) for the presence of polymorphisms within the probes, and SNP-containing probes were excluded from this analysis. Probes showing sub-background levels of expression in all samples were also removed to reduce the influence of these probes on total probe set and meta-probe set expression levels. We calculated mean probe intensities for a set of anti-genomic probes, which we designated as background expression. For each probe on the array, if the intensity for all samples was less than the background expression plus two standard deviations for the same GC content, then the probe was excluded from the summary calculations. The SI score was calculated by simply dividing the probe set intensity by the meta-probe set intensity (i.e., exon expression/gene expression) after the addition of a stabilization constant (13) to both the probe set and meta-probe set scores.

PCA was performed on the SI scores from all chips using the Partek Genomics Suite software package (Partek) in order to attribute the variance averaged over all exons to sources of variability, and to determine a confidence level in the consistency of expression profiles from biological and technical replicates. Comparison of expression data from individuals GM12750 and GM12751 identified outliers for three replicates of GM12750 (Figure 3.2) that were excluded from all subsequent analyses.

To analyze splicing differences between the two samples for each probe set, an unpaired Student's *t*-test was performed using the log-transformed SI values for all remaining replicates (12 of GM12750 and 15 of GM12751) of each individual (R statistical package, version 2.3.0). Probe sets showing significantly different SI scores were ranked by *P*-value. Linkage analysis tests of SI scores cosegregating with chromosomal regions for the CEPH 1444 family was carried out using MERLIN (version 1.0.1) with default settings (ABECASIS *et al.* 2002). The scan was performed using a region spanning 20 SNP markers centered on the probe set.

Differentially spliced probe sets were filtered using a number of criteria including: (1) detectable level above background (DABG < 0.05) for both the probe set and the meta-probe set to which it belongs; (2) normalized meta-probe set scores with a minimum intensity score of 50; (3) the transcript defined by a minimum of three exons; and (4) size of the exon corresponding to the probe set is divisible by three. This last criterion was added to ensure that changes resulting from exon inclusion/exclusion would be in frame, which has been observed in a high percentage of conserved and species-specific alternative exons (MAGEN and AST 2005)

comparisons, we also required that transcript expression levels between samples was less than twofold.

RT-PCR and sequence analysis

Total RNA was treated with 4 U of DNase I (Ambion) for 30 min to remove any remaining genomic DNA. First-strand cDNA was synthesized using random hexamers (Invitrogen) and Superscript II reverse transcriptase (Invitrogen). For all candidate probe sets, locus-specific primers within the adjacent, flanking exons were designed using Primer3 software (ROZEN and SKALETSKY 2000). Primers were designed within exons that had the following restrictions: (1) flanking exon expression level above background (DABG < 0.05) and (2) the flanking exon itself was not predicted to be alternatively spliced. Approximately 20ng of total cDNA was then amplified by PCR using Hot Start Taq Polymerase (Qiagen) with an activation step of 15 min at 95°C followed by 35 cycles of 30 sec at 95°C, 30 sec at 58°C, and 40 sec at 72°C and a final extension step of 5 min at 72°C. Amplicons were visualized by electrophoresis on a 2.5% agarose gel. Sequencing of the two products whose sizes corresponded to the predicted larger exon/intron-inclusion and shorter exon-skipped forms confirmed the AS. We performed BLAST analysis of the two splice variants against the non-redundant and EST databases at the National Center for Biotechnology Information (NCBI) to verify if both sequences are known or whether a novel splice isoform has been identified.

Results

Examination of splicing differences between two CEPH HapMap individuals

We investigated differences in exon-level expression in lymphoblastoid cell lines (LCLs; three biological and five technical replicates, for a total of 15 replicates per individual) from two unrelated individuals from the CEPH

HapMap population (GM12750 and GM12751). We defined the splicing index (SI) as the expression level of a given probe set (representing one exon) divided by the expression of the corresponding meta-probe set (representing the gene), to control for differences in gene expression levels between samples (CLARK *et al.* 2002; SRINIVASAN *et al.* 2005). Principal component analysis (PCA) indicates that the majority of the variance in SI is due to individual differences, while the remainder is due to biological and technical factors, suggesting that splicing variation between the two cell lines is frequent (Figure 3.2). Three of the replicates from individual GM12750 appear to be outliers and were removed from all subsequent analyses.

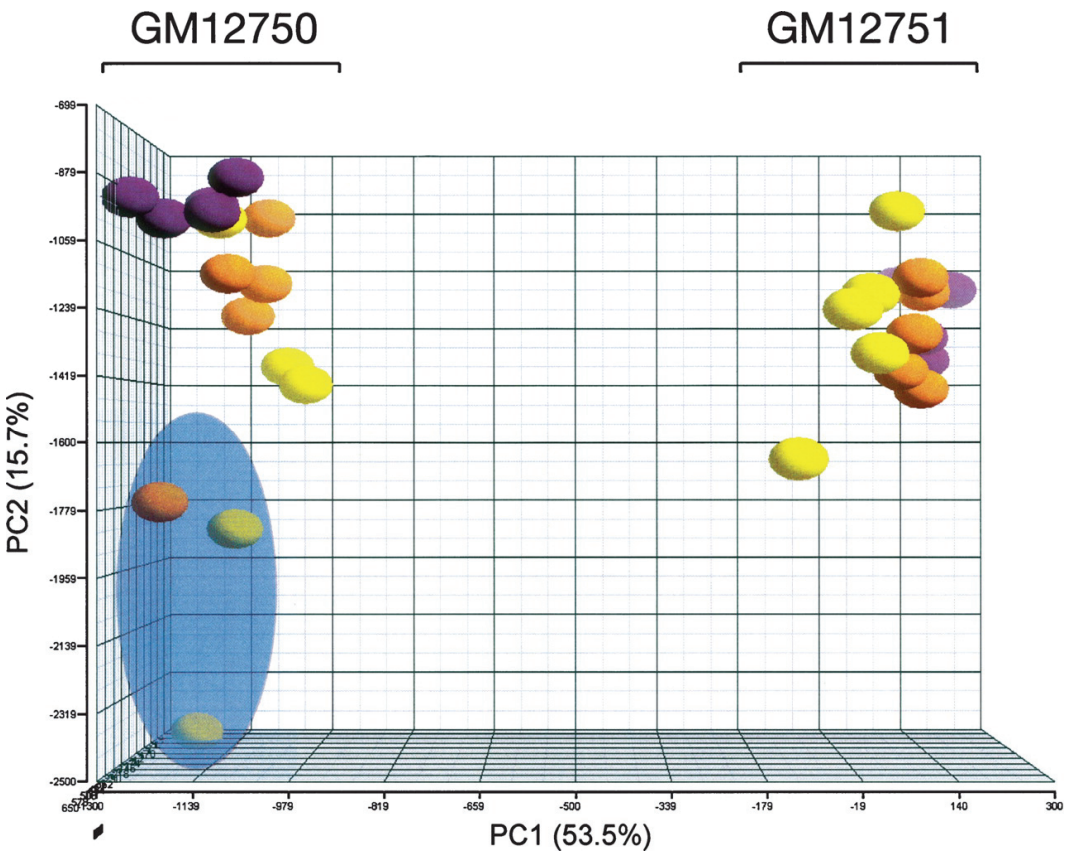


Figure 3.2: Principal component analysis. A three-dimensional plot of the splicing index data showing the three passages of five technical replicates each of individuals GM12750 and GM12751, on the left and right sides,

respectively. The three biological replicates are shown as purple, orange, and yellow spheres, respectively. The three outliers that were removed from all subsequent analyses are shaded in a blue sphere. The percentage of variance attributed to principal components one and two is shown on the X- and Y-axes, respectively. Plots were created using the Partek Genomics Suite software package (Partek).

The array contains sequences from two main sources: high confidence mRNAs from RefSeq and GenBank databases and ESTs from dbEST, and a lower confidence set of speculative gene structures predicted using software such as GENSCAN (BURGE and KARLIN 1997), TWINSCAN (KORF *et al.* 2001), and Exoniphy (SIEPEL and HAUSSLER 2004). For this study, we restricted our analyses to the high confidence set of mRNAs and probe sets. Inclusion of the low confidence theoretical probe sets may contribute expression values that go toward the overall summary and calculations of the meta-probe set score and may adversely affect the SI and all subsequent analyses. In doing so, the number of probe sets has been reduced approximately fivefold, from 1.4 million to 277,000 probe sets belonging to core RefSeq transcripts.

One of the potential issues regarding the use of microarrays, particularly with respect to our study of looking at differences in splicing between individuals, is the effect of polymorphisms within the probes that potentially affect binding affinities. Single nucleotide polymorphisms (SNPs) are very common genetic variations and occur at a frequency of one in 1000bp in the human genome (SACHIDANANDAM *et al.* 2001). Considering such a high frequency of SNPs, we would expect a large number of the probes to contain SNPs and, in some of the cases, to be polymorphic between the individuals that we are examining. In the comparison of two individuals, if a SNP exists within the target sequence

in only one of the individuals, probe binding and intensity will most likely be negatively affected in this sample. This would result in an apparent lower SI relative to the individual with the wild-type allele, potentially leading to a false-positive identification of differential probe set expression. We circumvent this issue by conservatively masking out all probes containing SNPs from the dbSNP database (release 126) and all HapMap SNPs polymorphic between our two samples, from the calculation of probe set and meta-probe set summaries. However, there are most likely unknown SNPs that are not yet annotated that may be present within the probes on the array, and all candidate probe sets will be dealt with on a case by case basis, examining the probe set for any discordant probes within them. Probes showing below background intensities in all samples were also masked out before calculation of probe set summaries in order to avoid potential influences of these low intensity probes on the estimated exon and transcript expression levels. After masking out all of these SNP-containing and background intensity probes, 234K probe sets remain for analysis.

After summarizing probe set scores, ~76K probe sets did not pass the statistical DABG (detected above background) criteria (see Methods) and therefore were not included in subsequent analyses. In order to identify candidates from the remaining 158K probe sets suggestive of differential splicing between the two individuals, we performed a t-test comparing the log-transformed SI scores on replicates of the two groups. Since there is no clear method for optimal determination of statistical cutoffs (THOMAS *et al.* 2005), we applied three different methods for multiple testing correction. The Bonferroni correction provided the most conservative estimate ($P = 3.159 \times 10^{-7}$, significance threshold $P = 0.05$), yielding 1892 potential probe sets (1.2% of expressed “core” probe sets) showing differential splicing. The false discovery rate (FDR) (BENJAMINI and

HOCHBERG 1995; STOREY *et al.* 2007) at a 0.01 significance level provided the least conservative estimate (P = 8.915 × 10^{-4}), with 8771 (5.7%) potential splicing events. We also ascertained the significance values using an empirical null distribution of *P*-values from the observed data, by shuffling the SI scores for all samples of each probe set (CHURCHILL and DOERGE 1994). For each probe set, we calculated an empirical *P*-value by comparing our observed, nonpermuted *P*-value to the distribution of permuted *P*-values, followed by Bonferroni correction on the permuted *P*-values. This method estimates 4020 (2.6%) differentially spliced probe sets between the two individuals. The average fold change in SI of all significant probe sets at the Bonferroni, permuted, and FDR corrected cutoffs are 1.85-fold, 1.48-fold, and 1.45-fold, respectively, showing a positive correlation between significance and fold-change expression.

We applied some additional biological and statistical criteria to the data set (see Methods), reducing the number of candidate probe sets to 1028. From this list, we proceeded to test a random selection of probe sets ranging from the highest significance level to those near the FDR cutoff. A small subset of 20 candidates were subjected to validation by reverse transcriptase–polymerase chain reaction (RT-PCR) using a pair of primers in two distinct exons flanking a third exon containing the predicted probe set. The presence of alternative isoforms for nine transcripts was confirmed by RT-PCR (Table 3.1), which translates into a 45% validation rate. However, our study evaluates the ability of this microarray technology to identify alternative AS events de novo in genetically diverse populations. Restricting our candidates to those showing EST and cDNA evidence of AS in sequence databases reduces the number of cases from 20 to 12, thereby increasing our success rate to 60% (seven out of 12). This is similar to the observed rates in a genome wide junction array study (73/153 = 48%) (JOHNSON *et al.* 2003) and a smaller custom array of both

exon and junction primers (11/20 = 55%) based on a priori knowledge of AS events (LE *et al.* 2004).

Table 3.1: Candidate genes with alternative splicing events

Gene name	RefSeq accession nos.	Function	AS event	Expected sizes of PCR products	PSID	Log2 (ratio)	*P*-value	RefSeq evidence	EST evidence
KIAA0460	NM_015203	Hypothetical protein LOC23248	Exon skipping	226, 304	2358260	−0.55	1.24×10^{-5}	No	Yes
LOC93349	NM_138402	Hypothetical protein LOC93349	Exon skipping	415, 490	2531328	−0.44	2.53×10^{-5}	No	Yes
CRTAP	NM_006371	Cartilage-associated protein precursor	Exon skipping	309, 438	2616180	−0.55	1.80×10^{-5}	No	Yes
SIDT1	NM_017699	SID1 transmembrane family, member 1	Exon skipping	200, 284	2636499	−0.56	1.04×10^{-9}	No	No
CAST	NM_001750 NM_173060 NM_173061 NM_173062 NM_173063	Calpastatin	Exon skipping	234, 273	2821249	−2.82	2.26×10^{-16}	Yes	Yes
PPFIA1	NM_177423 NM_003626	PTPRF interacting protein α1	Exon skipping	283, 436	3338488	−0.73	1.27×10^{-5}	No	No
OAS1	NM_016816 NM_002534 NM_001032409	2′,5′-oligoadenylate synthetase 1	Alternate SS	487, 585	3432462	1.34	5.84×10^{-7}	Yes	Yes
SFRS5	NM_006925 NM_001039465	Splicing factor, arginine/serine-rich 5	Intron retention	155, 439	3542221	1.03	5.46×10^{-9}	No	Yes
HHAT	NM_018194	Hedgehog acyltransferase	Exon skipping	208, 403	2378404	0.51	2.81×10^{-5}	No	Yes

Candidate alternatively spliced (AS) probe sets between two unrelated CEPH HapMap individuals (GM12750 and GM12751) that were validated by RT-PCR. The corresponding Affymetrix probe set ID (PSID), the nature of the observed AS event, and log-transformed fold-change in splicing index ratio between GM12750/GM12751 are indicated, as well as RefSeq and EST-based evidence for AS.

Analysis of validated AS events

Based on EST and RefSeq evidence, seven of the nine probe sets with confirmed AS are predicted to confer exon-skipping events, with the exception of the *OAS1* and *SFRS5* genes. Two *OAS1* splice variants (RefSeq accession nos. NM_016816 and NM_002534) are predicted to encode isoforms with alternative 3′ splice site (ss) usage of the last downstream coding exon. The probe set identified in the *SFRS5* gene is located within an intron between exons 4 and 5 and represents an intron-retention event. In total, seven of the nine probe sets that were identified in this study show annotated evidence in EST and RefSeq databases of AS. Probe sets corresponding to exons from the *PPFIA1* and *SIDT1* genes show no previous evidence of AS, demonstrating that the array can detect novel splicing events.

In three (*CAST*, *PPFIA1*, *OAS1*) of the top four validated splicing events with the highest degree of fold-change in SI between individuals, we observe a clear predominance of one isoform in one individual versus the alternate variant in the second individual. The majority of candidates with lesser fold changes show the presence of both splice variants in each of the individuals. From a biological perspective, the presence or absence of one of the two splice variants between individuals is more likely to have a functional consequence than are cases where two splice variants are expressed in all individuals with subtle differences in relative ratios. Loss of function from one variant without compensatory effects from expression of the alternative splice isoform may have drastic differences in downstream effects. However, until a complete validation of all candidate probe sets is performed, we cannot estimate how many of these "all-or-none" splicing events are present compared with the observation of both isoforms in each individual.

In one of our candidate genes, sequence analysis of the RT-PCR products identified a variant using a cryptic splice site within the predicted exon. Two *OAS1* transcripts show alternative 3′ ss usage in the predicted last exon of the gene, resulting in differential stop codon usage and a longer 3′ UTR in one transcript. In the future, sequence analysis of all validated probe sets will be necessary to accurately determine cryptic splice site usage, especially those in close proximity to the annotated splice site, which may be beyond the resolution of standard gel electrophoresis.

The available EST and mRNA-based evidence of AS in most of our candidate genes provides support and validation for our array-based discovery of known alternatively spliced transcripts. More importantly, the identification of new *PPFIA1* and *SIDT1* splice variants provide confidence

that we may be able to discover novel AS events and increase the catalog of the human transcriptome.

Association of splicing to *cis*-regulatory haplotypes

An important goal of this study was to demonstrate the genetic component of AS, specifically the inheritance of a splicing pattern and its association to a *cis*-regulatory haplotype. Using the SI of an exon as a quantitative trait, we performed regression-based linkage analysis (implemented in Merlin) (ABECASIS *et al.* 2002) within a three-generation family (CEPH 1444) for the nine verified AS events detected in this study. At a nominal level of LOD > 0.59, corresponding to $P < 0.05$, we observed evidence of linkage between SI scores and the corresponding chromosomal region in the *OAS1* (LOD = 0.76), *CRTAP* (LOD = 1.29), and *CAST* (LOD = 1.98) genes. RT-PCR based analysis confirmed segregation of the splicing pattern with the associated haplotype through all three generations of this pedigree (Figure. 3.3).

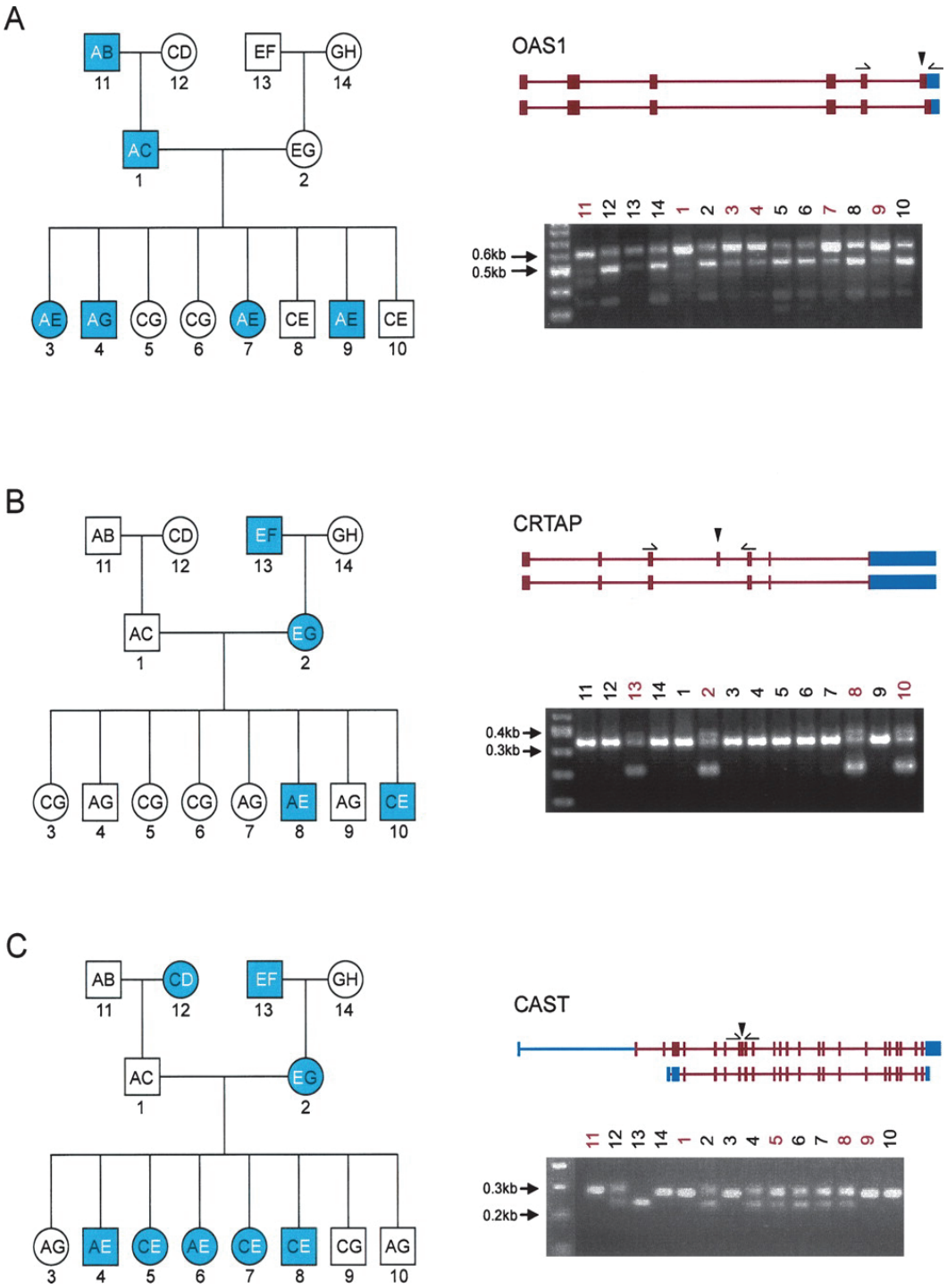


Figure 3.3: Heritability of alternative splicing. Inheritance of alternative splicing for genes (*A*) *OAS1*, (*B*) *CRTAP*, and (*C*) *CAST*. *Left* panel shows pedigree structure of CEPH/UTAH family 1444 with the autosomal dominant inherited splice pattern as blue symbols. Haplotypes for each of the eight founder chromosomes are labeled A, B, C, D, E, F, G, and H,

and the two inherited haplotypes of each family member are indicated within the symbol. The regulatory haplotype is shown as bold white text. Squares represent males, and circles represent females. CEPH/UTAH 1444 pedigree is labeled as follows: 1 (GM12739), 2 (GM12740), 3 (GM12741), 4 (GM12742), 5 (GM12743), 6 (GM12744), 7 (GM12745), 8 (GM12746), 9 (GM12847), 10 (GM12747), 11 (GM12748), 12 (GM12749), 13 (GM12750), and 14 (GM12751). The *right* panel shows the two transcript isoforms of the genes. Exon-body primers are shown above the flanking exons of the predicted alternatively spliced exons. Shown *below* the transcript isoforms are the RT-PCR results. Lanes are numbered from *1–14* according to the pedigree on the *left*.

The association between alternatively spliced isoforms and genetic variation was examined further by testing our nine candidates on a larger panel of 60 unrelated HapMap CEU individuals. In many cases, both splice variants are expressed in different ratios in various individuals, but the RT-PCR approach that was used here was not sensitive enough to quantify the relative isoform levels and establish a statistical association with a regulatory haplotype. Other methods based on the use of fluorescent dyes such as TaqMan PCR (GIBSON *et al.* 1996) may be more sensitive in detecting relative amounts of each isoform, although the cost associated with this technology is prohibitive for large-scale validation of predicted AS events. In clear cases where only one of the isoforms or the other is expressed, classical RT-PCR is a more suitable method. We were able to confirm the previously described association of *OAS1* variants to a candidate regulatory polymorphism (FIELD *et al.* 2005) and establish that the *CRTAP* splicing variant is rare and does not occur outside of members of CEPH family 1444 (data not shown).

The most interesting example of allelic association was identified in the *CAST* gene, which encodes for calpastatin, a calpain protease inhibitor. There are at least 11 known isoforms of calpastatin, all differing in their N-terminal regions (Figure 3.4B) (LEE *et al.* 1992). The predicted alternatively spliced exon of the *CAST* gene is supported by RefSeq and EST evidence of AS and encodes a portion of the first of four repetitive protease-inhibition domains. Consequently, removal or disruption of these calpain-inhibition domains may affect functionality and/or tissue specificity of the protein (TAKANO *et al.* 1993). The splicing pattern in the entire panel was correlated to a single SNP (rs7724759) that is most likely the causative polymorphism resulting in our differentially spliced isoforms. The SNP is located at the 3′ end of the exon and involves a G to A substitution that abates the weak consensus 5′ ss sequence. All individuals genotyped as homozygous GG for rs7724759 have an intact 5′ ss sequence and properly splice the exon, resulting in the larger PCR product. Individuals homozygous for AA at this position have a non-functional 5′ ss on both alleles that is improperly recognized by the splicing machinery; as such, the exon is excluded and accounts for the shorter, lower molecular weight band. When both isoforms are observed, the individual is heterozygous for this SNP and has both wild-type and polymorphic alleles. This exon also demonstrated linkage in the CEPH 1444 family, as previously mentioned, and examination of the pedigree clearly shows the inheritance of the two haplotypes through the three generations (Figure 3.3C).

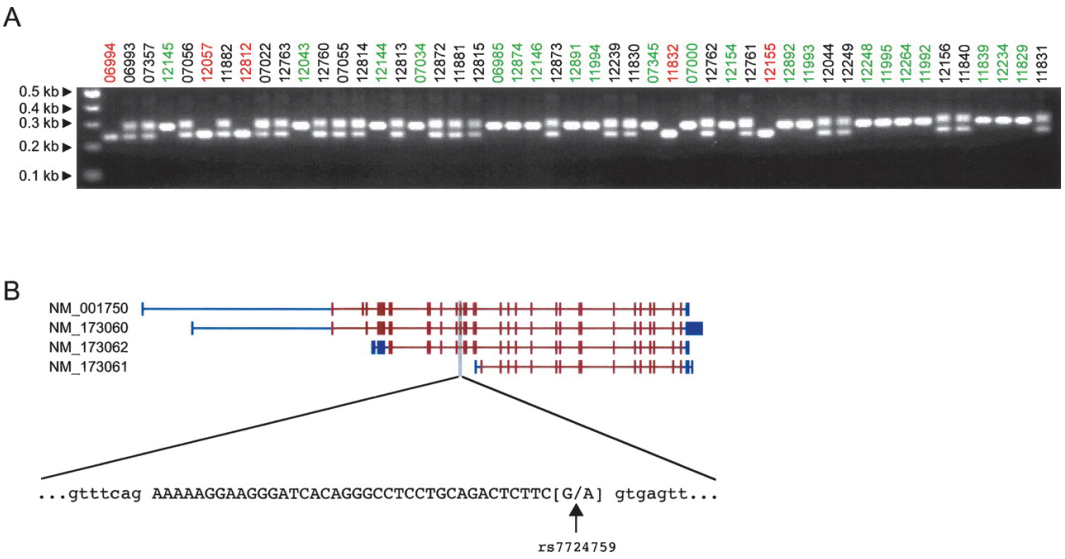


Figure 3.4: Association of alternative splicing and genotypes for the *CAST* gene. (*A*) RT-PCR of *CAST* exon against a panel of unrelated parents from each of the 30 HapMap CEU trios. Sample names are coloured according to their genotype for SNP *rs7724759*: homozygous GG (green), homozygous AA (red), and heterozygous AG (black). (*B*) Four known isoforms of the *CAST* gene are shown with their RefSeq accession numbers on the *left* and the candidate probe set shaded in grey. Shown *below* is the sequence of the exon in capital letters and flanked by the intronic sequence in lower case. The SNP rs7724759 is located at the last position of the exon and is a G to A substitution that disrupts the consensus splice site sequence.

We also examined the remaining eight AS events for both functional domains encoded within the respective exons and also for putative *cis*-acting SNPs that may control the splicing patterns. We did not identify any domains for any of the exons except a putative transmembrane domain within the *HHAT* exon. In most of the cases, the closest polymorphic SNPs between individuals GM12750 and GM12751 were all located either in the 5′ or 3′ flanking introns but at significant distances (>100 bp) from the splice site. We were able to identify SNPs either within or in close proximity (<100 bp) to the putative AS exon for the *SIDT1* and *OAS1*

genes and within the retained *SFRS5* intron. SNP *rs2271494* is located 25 bp upstream of the *SIDT1* exon and is found within the polypyrimidine tract. Mutations within this region may alter binding between the large subunit of the U2 small nuclear ribonucleoprotein particle (snRNP) auxiliary factor, U2AF, to this motif (SINGH *et al.* 1995). The SNP *rs151042* is located within exon 7 of the *OAS1* gene and is part of a haplotype block where another SNP marker, *hCV2567433*, at the exon 7 splice-acceptor site, has been shown to result in the usage of an internal splice site in the mutant allele (BONNEVIE-NIELSEN *et al.* 2005). In the one example of intron retention for the *SFRS5* gene, we identified a SNP (*rs3104*) centrally located within the intron; however, it does not appear to disrupt any known intronic splice enhancer or silencer. These results demonstrate that association studies of alternatively spliced exons with well-genotyped individuals are valuable in identifying the potential polymorphisms linked to the splicing event.

Discussion

Identifying AS events is important to understanding the diversity and complexity of the human genome, and we report on the use of a comprehensive exon-tiling array in our experimental design to discover such events between individuals. The same microarray design has also been recently used for a complete analysis of tissue-specific differences in splicing (GARDINA *et al.* 2006) and is potentially useful for many pairwise comparisons of splicing. Since the design of this array is not biased toward a priori knowledge of AS events, there is more potential for detecting novel splicing events. We demonstrated that novel isoforms can be discovered using this microarray, and others have recently shown the same (CLARK *et al.* 2007b; GARDINA *et al.* 2006). A number of different types of splicing events were identified, including exon exclusion, intron retention, and the use of cryptic splice sites. Exon-tiling arrays provide an

advantage over exon junction arrays in their ability to identify the use of cryptic splice sites due to the design of probes within an exon. Exon-junction probes can detect the joining of two exons at specific, known splice sites and are not as effective at the detection of novel, unannotated cryptic splice site usage. However, one disadvantage of tiling probes only within exons is its inability to provide information on how all the individual exons are linked within the different splice isoforms of a particular gene, a feature more suited to an exon-junction probe array. Proper design of an exon junction array for the entire human genome to interrogate all possible gene structures requires too many probes for every possible joining event. Such a design is more suitable for the examination of a smaller number of events, as demonstrated recently (BEN-ARI *et al.* 2006; VALVERDE *et al.* 2006; ZHANG *et al.* 2006). Each of these array designs possesses advantages and disadvantages, and given comparable false-positive rates obtained in this study and other splicing microarray studies, both are useful and informative in the identification of AS events. A follow-up study using a custom microarray consisting of a combination of exon and exon-junction probes may prove useful for confirming AS events and examining all possible transcript structures for a smaller subset of genes. This study focused on differentially expressed probe sets located within in-frame coding exons. Validation of probe sets corresponding to out-of-frame exons were not looked at, but these may introduce an upstream stop codon through cryptic splice site usage. This may confer differences in post-transcriptional regulation through nonsense-mediated decay. Probe sets located within 5′/3′ UTRs can also have widely varying biological functional consequences, such as changes in promoter regions or polyadenylation and transcript termination differences.

Exactly how much differential splicing is occurring between any two individuals is still unknown. We estimated that up to 2.5% of all RefSeq

exons expressed in lymphoblasts may show differential expression between the two samples tested, after factoring in our current validation rate, although a more accurate determination on the amount of differential splicing events will require a proper ROC-type analysis. However, this study examines splicing in lymphoblasts, and this estimate may change depending on the tissue tested. Alternative splice variants of the same gene can be expressed in multiple cell types to exert different functional and regulatory effects, which may also be individual specific. Neuronal tissues are known to have high levels of splicing (YEO *et al.* 2004a), and it is not unreasonable to assume that the amount of splicing between individuals may be higher in brain tissues than in lymphoblasts. A more complete picture may be ascertained by pairwise comparison of splicing in many tissues between individuals.

The large amount of genotyping information within identified populations from the HapMap project provides a tremendous resource for associating known SNPs or regions of linkage disequilibrium with genetic differences such as copy number variation, allelic imbalance, and AS, or phenotypic traits that may convey an increased risk of disease. Here, we have shown that this approach can be used to identify one or more SNPs associated with some of the splicing events identified. Further examination of the nature of the polymorphisms and their location relative to the spliced exon can give insight as to whether it is part of a larger *cis*-regulatory haplotype or in fact the causative SNP disrupting a splice site consensus sequence, an exonic splicing enhancer (ESE) or silencer (ESS), an intronic splicing enhancer (VIGNAUD *et al.*) or silencer (ISS), or other splice regulatory motifs such as the branch point or the polypyrimidine tract. Assigning a definitive causative effect of the SNP will require further experimental validation in vitro, such as monitoring splicing activity in cells using splice reporter constructs (MAYEDA and KRAINER 1999). However, it is quite

possible that there are unannotated SNPs proximal to the exon that are responsible for the differential splicing, and resequencing of the genomic regions neighboring the exons will be necessary to identify these polymorphisms.

Although we identify a candidate exon from the *CAST* gene showing genetic association with expression level changes, we do not know how often this occurs in a human population on a genome-wide scale. One method of properly assessing how common inherited splicing occurs would be to perform a whole-genome association study with more individuals from the HapMap population, using the SI scores as a quantitative trait. This is very similar to recent whole-genome association studies that suggest that common genetic variation explains much of the gene expression differences among individuals (STRANGER *et al.* 2005; STRANGER *et al.* 2007b). Carrying out a similar analysis at the exon level will yield better estimates of how common this heritability and genetic association is in humans.

The identification of SNPs within specific individuals in a population that affect splicing is an important issue to address when considering its relevance to possible resistance or susceptibility to disease states. An estimated 20%–30% of disease-causing mutations is believed to affect pre-mRNA splicing (FAUSTINO and COOPER 2003), through the disruption of splice sites, exonic and intronic splicing enhancers and silencers, or RNA secondary structure. In this study, the two *OAS1* splice variants identified have been previously associated with a SNP at an exon splice-acceptor site. This polymorphism results in the usage of an internal splice site in the mutant allele, which is thought to confer differences in host susceptibility to viral infection in type I diabetes patients (FIELD *et al.* 2005). A genome-wide analysis with well-genotyped CEPH HapMap

individuals will be an important starting point in identifying many more AS events and the causative polymorphisms involved in human diseases.

Acknowledgments

We thank Eef Harmsen for helpful discussions. This work is supported by Genome Canada and Genome Québec. T.J.H. is the recipient of a Clinician-Scientist Award in Translational Research by the Burroughs Wellcome Fund and an Investigator Award from CIHR. J.M. is a Canada Research Chair holder.

Chapter 4: Genome-wide analysis of transcript isoform variation in humans

Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, Cathy Provencher, Patrick Beaulieu, Thomas J Hudson, Rob Sladek & Jacek Majewski

Published in *Nature Genetics* on January 13th, 2008. **40**:225 – 231

Connecting text

In the previous chapter, we compared the transcript expression patterns derived from lymphoblast cell lines of two unrelated HapMap individual. We established that the Exon Array was capable of detecting different types of isoform differences such as alternative initiation, splicing and termination. We also found by conducting linkage analyses that some of these observed differences were inherited and therefore likely to be under genetic control. The efficacy of this pilot study prompted us to continue this study on a larger scale.

Chapter 4 describes our use of lymphoblast cell lines derived from 60 unrelated HapMap individual of Northern European descent that have been previously genotyped for ~4 millions SNPs by the International HapMap project (THE INTERNATIONAL HAPMAP CONSORTIUM 2003). RNA was isolated from these cell lines for each individual and was hybridized to an Exon Array. The main goal of this study is to combine the genotype information and the transcript expression at the isoform level to carry out genome-wide allelic association analysis.

Abstract

We have performed a genome-wide analysis of common genetic variation controlling differential expression of transcript isoforms in the CEU HapMap population using a comprehensive exon tiling microarray covering 17,897 genes. We detected 324 genes with significant associations between flanking SNPs and transcript levels. Of these, 39% reflected changes in whole gene expression and 55% reflected transcript isoform changes such as splicing variants (exon skipping, alternative splice site use, intron retention), differential 5' UTR (initiation of transcription) use, and differential 3' UTR (alternative polyadenylation) use. These results demonstrate that the regulatory effects of genetic variation in a normal human population are far more complex than previously observed. This extra layer of molecular diversity may account for natural phenotypic variation and disease susceptibility.

Introduction

Alternative pre-mRNA processing increases the complexity of eukaryotic transcriptomes, allowing multiple transcripts and protein isoforms with distinct functions to be produced from a single genomic locus (KIM *et al.* 2004). Within an organism, tissue specific gene isoforms are known to have important functions in development and proper functioning of diverse cell types (BLACK and GRAVELEY 2006). Across individuals, changes in normal isoform structure have phenotypic consequences and have been associated with disease (FAUSTINO and COOPER 2003; NISSIM-RAFINIA and KEREM 2005). Splicing defects in a number of genes, such as the cystic fibrosis transmembrane conductance regulator, *CFTR*, result in several known mendelian disorders (ZIELENSKI 2000). More subtle changes, such as alternative 3' processing and polyadenylation, have recently been associated with complex disorders: *OAS1* in severe acute respiratory syndrome (FIELD *et al.* 2005), *TAP2* in type I diabetes (QU *et al.* 2007),

and *IRF5* in susceptibility to systemic lupus erythematosus (CUNNINGHAME GRAHAM *et al.* 2007; GRAHAM *et al.* 2007).

Several recent studies have suggested that natural variation at the level of whole-gene expression is common in humans and is associated with genetic variants, such as SNPs or copy number variants (CNVs) (CHEUNG *et al.* 2005; SPIELMAN *et al.* 2007; STRANGER *et al.* 2005; STRANGER *et al.* 2007a). Studying variation in gene expression is becoming increasingly important because of its contribution to phenotypic differences among individuals and its possible regulatory and functional relationships to diseases. However, little is known at present about the genetic variation at the sub-transcript level or about differences in multiple transcript isoforms of the same gene. Here, we interrogated transcripts across their entire length, using the Affymetrix GeneChip Human Exon 1.0 ST Array, which can detect splicing differences between various types of samples (CLARK *et al.* 2007b; GARDINA *et al.* 2006; KWAN *et al.* 2007).

Methods

Cell line preparation

We obtained triplicate RNA samples from LCLs derived from the parents of 30 CEPH (CEU) trios (60 individuals) that had been genotyped for approximately 4 million SNPs by the International HapMap Project (THE INTERNATIONAL HAPMAP CONSORTIUM 2005). Cells were grown at 37 °C and 5% CO_2 in RPMI 1640 medium (Invitrogen) supplemented with 15% (vol/vol) heat-inactivated FCS (Sigma-Aldrich), 2 mM L-glutamine (Invitrogen) and penicillin/streptomycin (Invitrogen). Cell growth was monitored with a hemocytometer and cells were collected at a density of 0.8×10^6 to 1.1×10^6 cells/ml. Cells were then resuspended and lysed in TRIzol reagent (Invitrogen). Three successive growths were performed (corresponding to the second, fourth and sixth passages) after thawing

frozen cell aliquots. Three cell lines showed extremely poor growth and were not used in the study, leaving 57 LCLs for subsequent analyses.

Affymetrix exon arrays

We isolated RNA using TRIzol reagent following the manufacturer's instructions (Invitrogen) and assessed the RNA quality using RNA 6000 NanoChips with the Agilent 2100 Bioanalyzer (Agilent). Biotin-labeled targets for the microarray experiment were prepared using 1 μg of total RNA. Ribosomal RNA was removed with the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen) and cDNA was synthesized using the GeneChip WT (Whole Transcript) Sense Target Labeling and Control Reagents kit as described by the manufacturer (Affymetrix). The sense cDNA was then fragmented by uracil DNA glycosylase and apurinic/apyrimidic endonuclease-1 and biotin-labeled with terminal deoxynucleotidyl transferase using the GeneChip WT Terminal labeling kit (Affymetrix). Hybridization was performed using 5 micrograms of biotinylated target, which was incubated with the GeneChip Human Exon 1.0 ST array (Affymetrix) at 45 °C for 16–20 h. After hybridization, non-specifically bound material was removed by washing and specifically bound target was detected using the GeneChip Hybridization, Wash and Stain kit, and the GeneChip Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip Scanner 3000 7G (Affymetrix) and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix).

Preprocessing and analysis of array hybridization data

The Affymetrix Power Tools software package was used to quantile-normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta–probe set (representing

gene expression) intensities using a probe logarithmic-intensity error model (Affymetrix). High false-positive rates are common in microarray studies, and previous studies have suggested that a major factor arises from probes overlapping SNPs that result in changes to hybridization intensity (NAEF and MAGNASCO 2003), potentially influencing the apparent association between the SNP genotype and probe intensities. To reduce potential influences of SNPs on false positives, all probes containing known SNPs (dbSNP release 126) were masked out before summarizing probe set and meta–probe set scores. The presence of unannotated SNPs affecting probe hybridization will remain (see below), but these cannot be detected by any statistical methods except for the impractical solution of resequencing all probes across the panel used in the study. We also filtered probe intensity levels by magnitude of response, removing probes that seemed to be in the background. Probe intensities were extracted for a series of 16,934 antigenomic probes targeted to nonhuman sequences and averaged by their relative G+C content. The threshold for background expression was defined as the average intensity for a given G+C content plus 2 standard deviations. For any given genomic probe on the array, if the intensity across all samples was below the threshold for the same G+C percentage, then it was considered background and masked from the analysis. In total, 670,809 probes corresponding to core annotated probe sets were masked from the analysis, reducing the number of core probe sets in the analysis to 244,027 probe sets.

Association analysis and multiple test correction

We examined probe set expression levels for association with flanking SNPs. For each of the 244,027 core probe sets and 17,653 meta–probe sets, we tested for association of the expression levels to HapMap phase II (release 21) SNPs with a minor allele frequency of at least 5% within a 50-kb region flanking either side of the gene containing the probe set,

using a linear regression model in the R software package. Raw *P*-values were obtained from the regression using the standard asymptotic *t*-statistic.

To correct for testing of associations between multiple probe sets and SNPs, we carried out permutation tests followed by FDR correction. Within each expression-versus-genotype matrix, we randomly permuted the expression values for all probe sets belonging to the same meta–probe set (to preserve the haplotype block structure). For each expression measurement, we computed and retained only the highest asymptotic *P*-value and produced the distribution of maximum *P*-values within the permuted dataset. The maximum asymptotic *P*-values from the experimental data were then converted into empirical *P*-values by mapping onto the permuted distribution. The above procedure corrects for testing multiple SNPs against each expression value. Subsequently, we performed an FDR correction (BENJAMINI and HOCHBERG 1995) on the empirical *P*-values, to control the FDR across multiple expression values. The procedure was applied separately to measurements at the probe set and meta–probe set levels. We used a 0.05 FDR criterion as a significance cutoff in our analysis. For the sake of clarity, all of the values and cutoffs quoted in the results correspond to the raw, uncorrected *P*-values.

Classification of transcript isoforms

We developed an automated method to categorize the transcriptional and isoform changes. The algorithm first classifies transcripts as expression variants if there is an association of the entire meta–probe set significant at the $P < 6.02 \times 10^{-7}$ level (see above for explanation of the cutoffs). Subsequently, the algorithm identifies all individual probe sets significant at the $P < 9.73 \times 10^{-9}$ level that do not belong to the expression variants

detected above. All such significant probe sets are then grouped into blocks corresponding to exons, according to their RefSeq annotation. Each significant block is classified as an initiation, splicing or termination change according to its position within the transcript (3', internal, or 5', respectively). Cases with two or more of the above events occurring in a single transcript are classified as complex. Finally, all results were manually curated. To visualize the potential nature of the isoform changes on a gene level, the probe sets were examined in the context of their transcript, mRNA, and EST information. For each gene predicted to have SNP-associated transcript- or exon-level expression changes, we plotted the *P*-values of all the corresponding probe sets and overlaid the fold change expression levels between the two homozygous genotypes for the significant SNP identified in the association analyses (see Supplementary Figure 2 - www.nature.com/ng/journal/v40/n2/suppinfo). We made minor adjustments (23 of 324 events) to the automated classifications, mostly in cases where the designations were not consistent with annotated alternative isoform structures or where the Affymetrix transcript annotation was incorrect.

Validation of transcript isoform changes

Total RNA was treated with 4 U of DNase I (Ambion) for 30 min to remove any remaining genomic DNA. First-strand complementary DNA was synthesized using random hexamers (Invitrogen) and Superscript II reverse transcriptase (Invitrogen). All primers used for RT-PCR reactions (see Supplementary Table 3 - www.nature.com/ng/journal/v40/n2/suppinfo) were designed using Primer3 software (ROZEN and SKALETSKY 2000). Candidate probe sets showing association were validated in two ways, depending on their location within the gene. For all probe sets located within coding exons and possessing flanking exons in all known RefSeq isoforms, we designed

locus-specific primers within the adjacent flanking exons. Approximately 20ng of total cDNA was then amplified by PCR using Hot Start Taq Polymerase (Qiagen) with an activation step at 95 °C (15 min) followed by 35 cycles at 95 °C (30 s), 58 °C (30 s) and 72 °C (40 s) and a final extension step at 72 °C (5 min). Amplicons were visualized by electrophoresis on a 2.5% agarose gel.

For probe sets located within 5' or 3' untranslated regions or within exons that did not have a flanking exon, we designed a set of primers to amplify the differentially expressed candidate probe set itself. For comparison, other primer pairs were designed to amplify products that corresponded to the adjacent probe sets and were not significantly associated with the same SNP. Total expression measurements were carried out using real-time PCR with Power SYBR Green PCR Master Mix (Applied Biosystems) following the manufacturer's instruction on an ABI 7900HT (Applied Biosystems) instrument. The reaction was set up in 10 µl final volume applying the following conditions: 8 ng of total cDNA and 0.32 µM of gene-specific primers; cycling, 95 °C (15 min) and 95 °C (20 s), 58 °C (30 s), 72 °C (45 s) for 40 cycles. Relative quantification of each amplicon was evaluated on RNA from 57 cell lines in triplicate. For each amplicon, a standard curve was established using dilution series of a mix of cDNA samples with known total cDNA concentration. Human 18S rRNA was also quantified using TaqMan probes as a control for well-to-well normalization (TaqMan Pre-Developed Assay Reagents for Gene Expression – Human 18S rRNA, 4319413E, Applied Biosystems). The cycle threshold (*Ct*) values for each replicate were transformed to relative concentrations using the estimated standard curve function (SDS 2.1, Applied Biosystems) and normalized based on 18S real-time data from the same samples to account for well-to-well variability. The quantitative data was used in regression analyses with the same SNP identified in the

original association to confirm the significance, using a *P*-value threshold of 0.05/*N* where *N* is the number of candidate genes tested using this method. The regression line was required to be in the same direction as the original association. Quantitative RT-PCR of the control probe sets showing no association with the SNP were also required to be nonsignificant at this threshold.

Effect of unannotated SNPs on the analysis

We have previously shown that SNPs located within probes may affect their hybridization to target DNA (KWAN *et al.* 2007), and have therefore conservatively masked out all probes containing SNPs to circumvent this problem. However, probes containing unannotated SNPs are not accounted for; therefore, we wanted to assess the effect of these unknown SNPs on our analysis. We selected 83 genes, each of which contained only a single significant probe set. Many (63) of these probe sets are supported by a single independent, nonoverlapping probe, and such probe sets are the most susceptible to the effect of SNPs, because every probe could potentially be affected by a single SNP. We sequenced the probe sets from the cell lines of six individuals, three from each of the two homozygous genotypes of the associated SNP. We observed that the sequences for 56 probe sets (67.5%) were identical in all samples tested, suggesting that these are more likely to be true events and not an artifact of one or more SNPs located in the individual probes representing the probe set. In the remaining 27 probe sets (32.5%), we identified previously unknown SNPs or indels overlapping one or more of the probes of the probe set, and in most cases, these polymorphisms segregated with one of the two homozygous sample groups, most likely giving rise to the apparent false-positive hit. We excluded these 27 probe sets from our candidate list presented in the manuscript. All of the remaining candidates are supported by two or more independent probes, and are much less

susceptible to the effect of unknown SNPs. Only 2 out of the 32 candidates from the final dataset selected for validation (6%) contained previously unidentified SNPs and hence failed validation, showing that the effect of SNPs on the final results presented here is small.

Results and discussion

Exons within a gene are represented on the microarray by individual probe sets, and were considered discrete units for our analysis of transcript isoform-processing differences. We used triplicate samples of lymphoblastoid cell lines (LCLs) derived from 57 unrelated Centre d'Etudes du Polymorphisme Humain (CEPH) CEU individuals (Utah residents with northern and western European ancestry) genotyped by the HapMap consortium (THE INTERNATIONAL HAPMAP CONSORTIUM 2005), allowing us to establish a possible genetic basis for any observed variations in transcript isoforms with associated SNPs. A linear regression analysis under a codominant model was carried out to associate probe set expression intensities with the genotypes of all SNP markers within a window of 50 kb flanking the boundaries of the transcript cluster (meta–probe set) containing the probe set. We assessed the statistical significance of the variation using the *t*-statistic, and used the regression equation to estimate the fold change in expression between the two homozygous genotypes. We used permutation testing (CHURCHILL and DOERGE 1994) to determine empirical *P*-values corresponding to the asymptotic *P*-values obtained from the regression. Subsequently, we applied the false discovery rate (FDR) correction to establish a cutoff *P*-value of 9.73×10^{-9}, corresponding to the 0.05 FDR level (see Methods). This yielded 757 unique probe sets showing significant SNP associations, belonging to 317 unique meta–probe sets (see Supplementary Table 1 - www.nature.com/ng/journal/v40/n2/suppinfo). Although the most significant SNPs may not be the causative polymorphisms responsible for

these differences in probe set expression, they are very probably in linkage disequilibrium with the causative polymorphism(s). This is reflected in the distance distribution of associated polymorphisms, most of which are in close proximity to the probe sets (see Supplementary Figure 1 - www.nature.com/ng/journal/v40/n2/suppinfo). The association analysis at the transcript (meta–probe set) level resulted in a 0.05 FDR cutoff of 6.02×10^{-7}, yielding 127 unique transcripts with significant genetic association at the gene expression level. Of these 127 transcripts, all but seven were common to the 317 transcripts derived from the regression analysis at the probe-set level; therefore, our final dataset comprised 324 transcripts predicted to have expression changes at the meta–probe set and/or probe set level.

We examined the 324 transcripts in greater detail (Figure 4.1; examples in Figure 4.2) to determine the nature of the isoform changes on a transcript level (summarized in Supplementary Table 2 and Supplementary Figure 2 - www.nature.com/ng/journal/v40/n2/suppinfo). Expression changes were automatically classified on the basis of the positions of the variable probe sets, followed by manual curation based on visualization of the entire transcript (Supplementary Figure 2 - www.nature.com/ng/journal/v40/n2/suppinfo). A large number of genes (127, or 39%) showed whole-gene expression changes. However, an even larger proportion (55%) of genes showed transcript-isoform changes only, without an accompanying change in the expression of the entire locus. Nearly half of these transcript variations were at the splicing level (85, or 26%), with the remaining changes at the level of transcript termination (57, or 18%) and initiation (35, or 11%) (Figure 4.3). It should be noted that some of the genes showing changes in the expression level of the whole gene also showed further changes in splicing, transcript termination and/or transcript initiation, suggesting that transcript isoform

variation constitutes a large part of the genetic variation we have observed. A small number (20, or 6%) of genes showed very complex patterns of isoform variation that were difficult to interpret. Notably, when we compare the proportion (18%) of significant probe sets within the 3' untranslated regions (UTRs) with the proportion of all 3' UTR core probe sets (13%) on the array, we found a significant over-representation (Pearson's chi-squared test, $P = 5.73 \times 10^{-6}$) of probe sets in this region, indicating that transcript termination variations may occur more frequently than expected. Because predicted changes to the 3' UTR may affect mRNA stability and subcellular localization, this type of isoform variation may have important regulatory roles. These findings illustrate a very complex pattern of expression changes associated with genetic variation, encompassing alterations at the whole-gene expression level and/or differences in transcript isoforms.

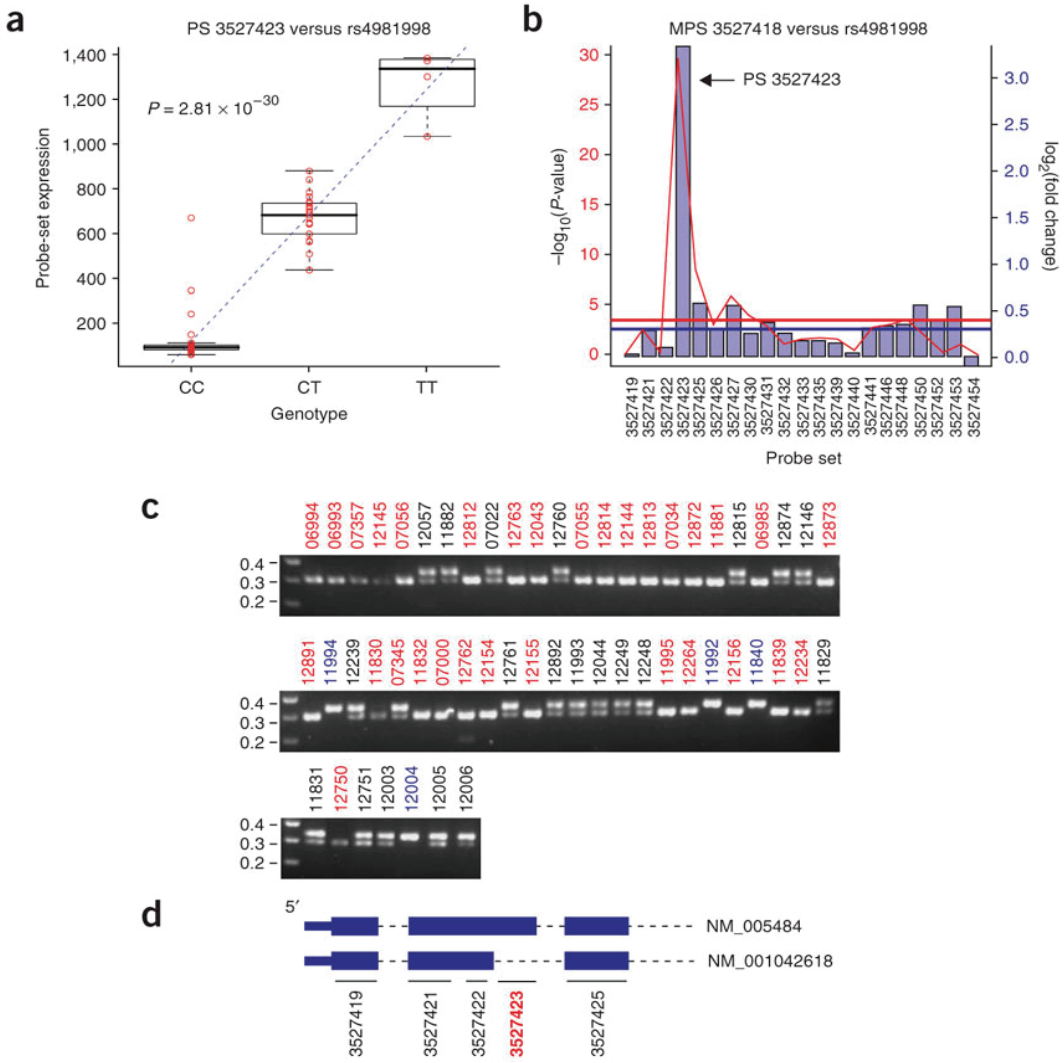


Figure 4.1: Analysis steps from identification of significant probe set in *PARP2* gene to validation. (a) Linear regression analysis of expression scores for probe set (PS) 3527423 with genotypes of SNP rs4981998, giving a P-value of 2.81 x 10^{-30}. Probe set scores for each individual are shown in red and regression line is indicated with blue dashes. (b) Visualization of probe set 3527423 in the context of all other probe sets belonging to the same transcript (meta–probe set 3527418). For each probe set, the significance level (P-value) is graphed (red line), along with fold change expression between the mean scores of the two homozygous genotypes (meanTT / meanCC) (vertical blue bars). The solid horizontal red and blue lines represent the significance and fold change expression

for the regression analysis at the meta–probe set level against SNP rs4981998. Arrow, probe set 3527423. (c) RT-PCR validation of probe set 3527423 using flanking exon-body primers. Individuals are highlighted by color according to their genotype for SNP rs4981998: CC (red), CT (black), TT (blue). (d) Schematic of 5' end of two isoforms of PARP2 with exon array probe sets shown below the exons. The significant probe set 3527423 is highlighted in red and corresponds to alternative 5' splice site use resulting in a larger second exon for NM_005484.

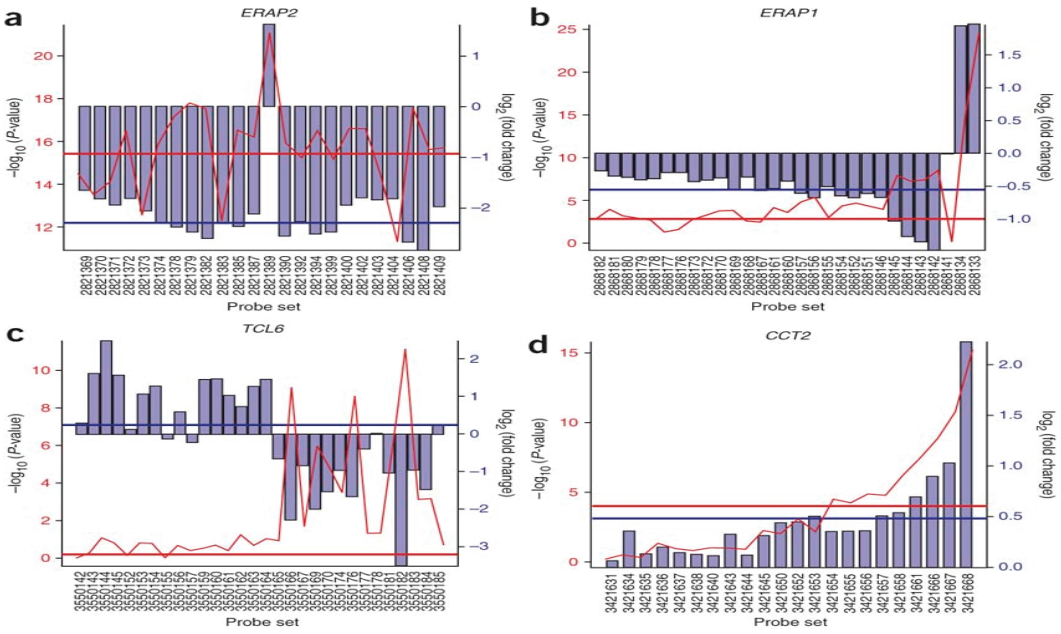


Figure 4.2: Examples of different types of transcript isoform events observed (data is graphed as in Figure 4.1b). (**a**) Gene expression level changes of *ERAP2*, including alternative splicing of a cassette exon. (**b**) Differential 3' UTR change of *ERAP1* resulting in long and short isoforms with alternative stop codon use. (**c**) Expression of two *TCL6* transcript isoforms that contain different 5' and 3' ends. (**d**) Increasing significance and fold change in expression levels toward the 3' end of the *CCT2* gene, suggesting genetic variation associated with mRNA stability.

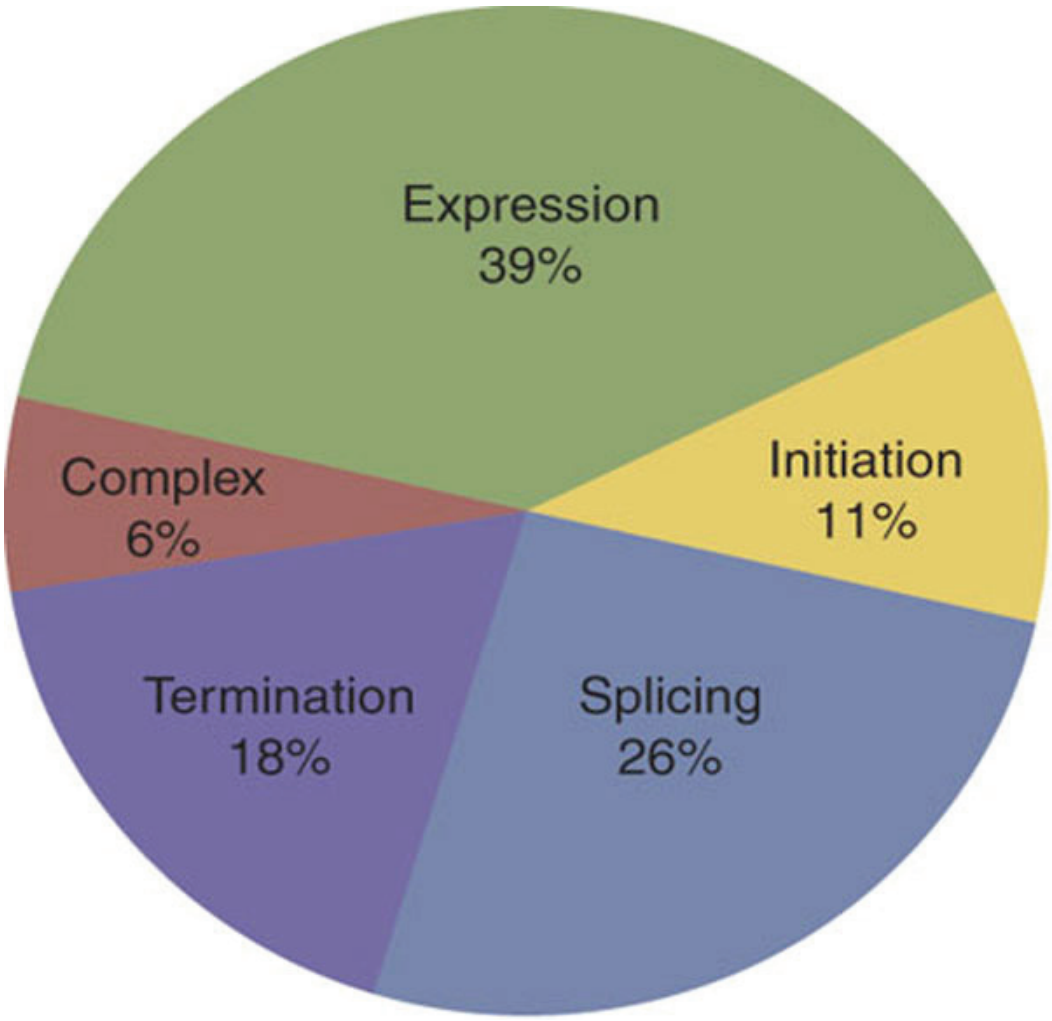


Figure 4.3: Classification of genes showing expression changes at the exon and/or transcript level. The 324 genes were classified into separate categories depending on the nature of the isoform change occurring: expression changes at the whole transcript level (green), transcription initiation changes (yellow), alternative splicing of a cassette exon (blue), transcription termination changes (purple), and complex changes of multiple event types (red). The percentages shown assume a uniform false-positive rate for all results. To obtain a lower bound for the relative frequency of isoform variants, we have also recalculated the frequencies of the isoform changes (but not whole-gene expression and complex changes) based on our current false positive rate estimate of ~20% (from validation experiments). Thus, we obtained the following ranges for each

of the changes: whole gene expression, 39–44%; initiation, 10–11%; splicing, 24–26%; termination, 16–18%; and complex events, 6–7%.

We proceeded, using two different methods, to validate 32 of our top candidate events distributed among the coding (16), 5' UTR (6), and 3' UTR (10) regions. For alternative splicing events of internally located probe sets, we performed RT-PCR on our entire panel of cell lines using exon-body primers in the two exons flanking the candidate probe set (Figure 4.1c). We confirmed 15 probe sets showing SNP association to splicing of a cassette exon or intron (Table 4.1) and classified them as follows: eight probe sets corresponded to splicing of a coding exon, four probe sets were located in the 5' UTR and resulted in the removal of potential promoter sequences or alternative start codon use, two probe sets were found within intronic regions and resulted in intron retention, and the remaining probe set was located in the 3' UTR and altered its length. The second, more sensitive validation method using quantitative real-time RT-PCR was applied to differentially expressed probe sets within the 5' or 3' UTR and to those in which one of the flanking probe sets was missing in one of the alternative isoforms. We designed sets of primers to amplify the differentially expressed probe set itself and compared the resulting PCR products to ones corresponding to adjacent probe sets showing no association to the SNP and also expected to have similar expression levels across all cell lines. Quantitative PCR data was used to perform a linear regression fit with the original associated SNP and confirm the significance and direction of the association analysis with the microarray data at a nominal P-value of $0.05/N$, where N is the number of candidates tested in the real-time RT-PCR. Using this method, we validated six UTR-located probe sets showing SNP association: four in the 3' UTR (alternative polyadenylation) and two in the 5' UTR (differential transcriptional initiation). We also used this method on the candidate

probe sets that failed our initial validation method owing potentially to low sensitivity of endpoint PCR of minor isoforms, and we were able to validate another four probe sets: two within coding regions and two within the 3' UTRs. In total, 25 of 32 candidate probe sets were validated, for a success rate of 78%. The remaining 7 probe sets failed validation, which can be partially accounted for by unannotated SNPs located within the probe sets possibly leading to altered hybridization signals (ALBERTS *et al.* 2007) (see Methods), suboptimal primer design, limited sensitivity of our validation methods, and/or noise from the microarray. We also validated several differentially spliced exons under a more relaxed stringency below our estimated cutoff, indicating that the frequency of genes showing SNP-associated changes is probably greater than what can be estimated from our current analysis. A recent estimate suggests that ~21% of annotated alternatively spliced genes are associated with SNPs that determine the relative abundances of the alternative transcript isoforms (NEMBAWARE *et al.* 2004).

Table 4.1: Validation of probe sets

Gene	Probe set	SNP	*P*-value	Chromosomal location	Probe set location	Type of event	RefSeq/EST evidence
CEP192	3779862	rs482360	3.71×10^{-19}	chr18:13047770–13048132	Coding	Intron retention	Yes
ZNF83	3869658	rs1012531	2.72×10^{-10}	chr19:57808794–57808830	Coding	Intron retention	Yes
C17orf57	3724617	rs3760372	5.54×10^{-12}	chr17:42793744–42793848	Coding	Exon skipping	Yes
CAST	2821249	rs7724759	7.17×10^{-16}	chr5:96102207–96102239	Coding	Exon skipping	Yes
CD46	2377476	rs4844390	1.06×10^{-14}	chr1:204329527–204329556	Coding	Exon skipping	Yes
ATP5SL	3863093	rs1043413	9.38×10^{-11}	chr19:46631033–46631167	Coding	Exon skipping	Yes
ERAP2	2821389	rs2255546	8.37×10^{-22}	chr5:96261677–96261705	Coding	Alternative splice site use	Yes
POMZP3	3057764	rs2005354	3.77×10^{-22}	chr7:75892151–75892256	Coding	Exon skipping	Yes
ULK4	2670619	rs1717020	5.99×10^{-11}	chr3:41932478–41932514	Coding	Exon skipping	No
PARP2	3527423	rs2297616	2.81×10^{-37}	chr14:19883099–19883123	Coding	Alternative splice site use	Yes
ATPIF1	2327383	rs2481974	4.26×10^{-11}	chr1:28248451–28248478	Coding	Alternative splice site use	Yes
MRPL43	3303658	rs12241232	1.24×10^{-11}	chr10:102731257–102731290	Coding	Exon skipping, differential stop codon use and 3′ UTR length	Yes
DKFZp451M2119	2588913	rs10930785	1.93×10^{-28}	chr2:178022380–178022482	5′ UTR	Exon skipping	Yes
RNH1	3358076	rs11821392	4.34×10^{-15}	chr11:494826–494888	5′ UTR	Exon skipping	Yes
SNX11	3725089	rs7224014	4.20×10^{-9}	chr17:43543086–43543116	5′ UTR	Exon skipping	Yes
USMG5	3304753	rs7911488	2.66×10^{-24}	chr10:105143981–105144095	5′ UTR	Exon skipping	Yes
SEP15	2421300	rs1407131	7.57×10^{-13}	chr1:87091818–87092018	5′ UTR	Differential 5′ UTR length	Yes
SLC35B3	2941033	rs3799255	2.12×10^{-10}	chr6:8380460–8380572	5′ UTR	Differential 5′ UTR length	Yes
C17orf81	3708382	rs2521985	2.55×10^{-13}	chr17:7100907–7100934	3′ UTR	Exon skipping, differential 3′ UTR length	Yes
ERAP1	2868133	rs7705827	6.09×10^{-19}	chr5:96123330–96124483	3′ UTR	Differential 3′ UTR length	Yes
TAP2	2950168	rs3763355	1.98×10^{-13}	chr6:32897620–32897880	3′ UTR	Alternative splice site use, differential 3′ UTR length	Yes
IRF5	3023264	rs6969930	8.27×10^{-22}	chr7:128183412–128183723	3′ UTR	Differential 3′ UTR length	Yes
PPIL2	3938301	rs5999098	1.46×10^{-12}	chr22:20374916–20375108	3′ UTR	Differential 3′ UTR length	Yes
PTER	3236819	rs1055340	5.25×10^{-18}	chr10:16595519–16595641	3′ UTR	Differential 3′ UTR length	No
WARS2	2430765	rs1325933	3.53×10^{-8}	chr1:119285989–119286236	3′ UTR	Differential 3′ UTR length	Yes

List of candidate probe sets validated by qualitative or quantitative RT-PCR. The gene name and the significant probe set are indicated along with the SNP and *P*-value from the linear regression analysis. The chromosomal location of the probe set is also shown, including its relative location within the gene. The nature of the isoform change is indicated, as is any existing RefSeq or EST evidence of this change.

A recent study used Illumina arrays to capture gene expression information within the CEU population (STRANGER *et al.* 2007a).The Illumina design, along with many other expression platforms, targets probes to the 3' end of genes and cannot identify specific isoform changes. Our present results demonstrate that the nature of the changes is qualitatively different than previously reported for several genes in that study. For example, our analysis shows that *IRF5*, implicated in susceptibility to systemic lupus erythematosus, shows differences in the 3' UTR (Figure 4.4), where the A allele of rs10954213 creates a functional polyadenylation site, shortening its 3' UTR (CUNNINGHAME GRAHAM *et al.* 2007; GRAHAM *et al.* 2007). This result for *IRF5* contrasts the original predicted change at the gene expression level (CHEUNG *et al.* 2005; SPIELMAN *et al.* 2007; STRANGER *et al.* 2005; STRANGER *et al.* 2007a) and occurs because the Illumina array interrogates *IRF5* with a probe in the 3'

UTR specific to the long isoform. Other examples previously classified as expression changes include *PTER*, which we show to have a variation in the 3' UTR, and *C17orf81* (also known as *DERP6*), which shows alternative splicing of a cassette exon. Another interesting example is *ERAP2*, which has been reported as having an expression change (CHEUNG *et al.* 2005). Our results confirm this variation in expression; however, we additionally detect alternative splice-site use in one of the exons (Figure 4.2a). Many platforms have been used so far in these population-wide expression analyses, and although there is substantial overlap between the studies, significant discordance also exists. A recent paper identified 374 gene-expression phenotypes associated with SNP markers from a study of 3,554 genes (CHEUNG *et al.* 2005). Differences in statistical stringency and false discovery rate most likely explain the higher proportion of SNP associations in their study. However, their set of 3,554 genes was pre-selected for the most variable expression phenotypes among an original set of >8,000 genes. This restricted set of genes may exclude examples of isoform changes without an accompanying change in whole-gene expression, which we observed in our study. In future expression association studies, comparative meta-analyses across different microarray designs may help eliminate platform-specific technical artefacts and allow the elucidation of true isoform and gene-level variations.

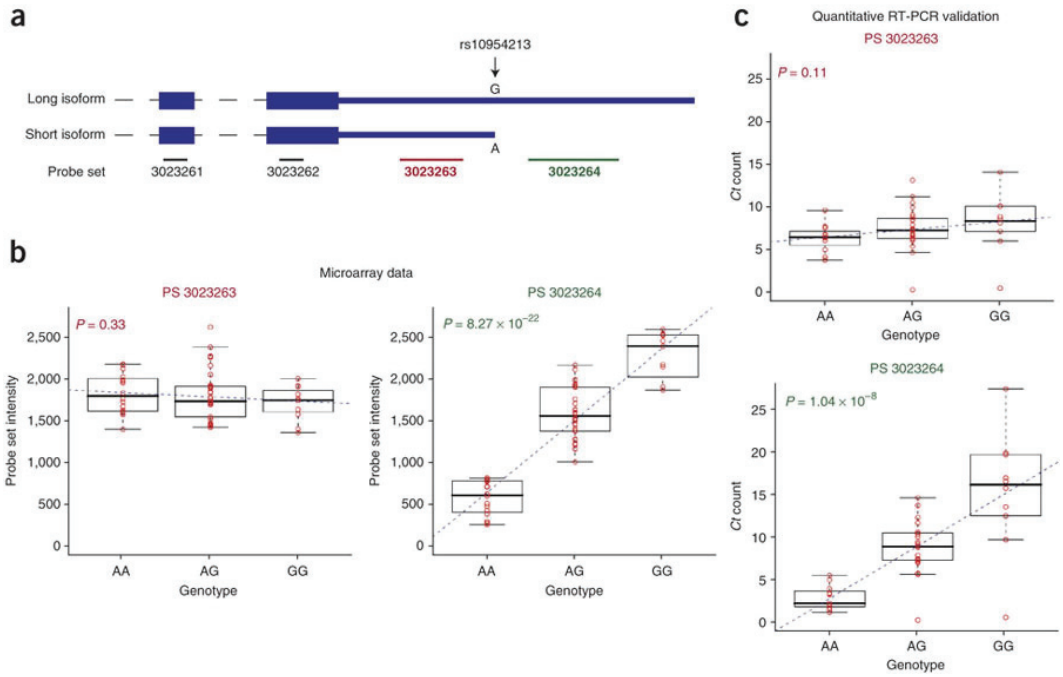


Figure 4.4: Validation of 3' UTR change in IRF5 by quantitative real-rime RT-PCR (a) Schematic of the 3' ends of the long and short isoforms of IRF5. Exons are shown in blue, introns are dashed lines, and solid horizontal lines below the exons indicate probe sets. (b) Regression analyses of probe sets 3023263 and 3023264 against SNP rs10954213. (c) Regression analysis of Ct counts from quantitative real-time RT-PCR against the genotype of SNP rs10954213, to confirm the original microarray data. We used two sets of primers on the panel of individuals, designed to amplify probe sets 3023263 and 3023264, respectively.

We show that tools such as the exon array, targeting probes to many regions of the gene, give a more complete picture of the true complexity of variation in gene expression than previously believed. This variation exists at all levels of transcript processing, beginning with initiation of transcription, through pre-mRNA splicing (HULL *et al.* 2007; KWAN *et al.* 2007; NEMBAWARE *et al.* 2004), to alternative polyadenylation, and it has the potential to exert diverse cellular responses and phenotypic effects. Transcript alterations within coding regions of the gene, such as the

addition or removal of sequences coding for functional domains or the introduction of premature stop codons, may greatly alter the protein sequence, structure and function (LEWIS *et al.* 2003; LIU and ALTMAN 2003). Changes outside the coding regions can also have wide-ranging regulatory consequences. Differential exon selection within the 5' and 3' UTRs may alter mRNA stability and translational efficiency by the addition or removal of regulatory sequences. In some genes (for example, *ATPIF1* and *TAP2*), selection of an alternative splice site for the terminal exon resulted in differential stop codon use and, consequently, changes in the length and composition of the 3' UTR. Alterations in the 3' UTR can also be affected by alternative use of polyadenylation sites and approximately half of human genes are predicted to contain several polyadenylation sites, resulting in transcripts with different 3' UTR lengths (TIAN *et al.* 2005; YAN and MARR 2005). Altering a functional polyadenylation site through a single polymorphism may lead to isoform switching. The 3' UTR is also involved in post-transcriptional regulation through the targeting of specific UTR sequences by microRNAs (miRNA) (VALENCIA-SANCHEZ *et al.* 2006; WU *et al.* 2006). Expression of multiple isoforms may be indirectly controlled through the differential expression of miRNAs or by polymorphisms in these miRNA-specific sequences. The end consequence of many of these alterations in the UTRs affects a cascade of downstream processes such as stability, localization and translation efficiency, and it directly contributes to phenotypic diversity and possible disease states. A systematic characterization of the polymorphisms to determine the true causative SNPs resulting in these changes will lead to the possible identification of new regulatory motifs and is currently being undertaken.

Earlier studies suggested that gene expression constituted an important piece of human variation, and although it remains a significant aspect, the

added complexity of transcript-processing variations and the potential outcome of these differences greatly alter our earlier perceptions. We estimate that between 50 and 55% of gene expression variation is isoform based. Our results constitute an important change in way we view the effects of common genetic variation in humans and highlight the need for broader investigation into the causes of differential gene expression, as well as previously found and new disease associations that lack clear functional variants.

Acknowledgments

The authors would like to thank D. Serre, T. Pastinen, E. Harmsen and H. Zuzan for helpful discussions and D. Sinnett for technical assistance. This work is supported by Genome Canada, Genome Québec and the Canadian Institutes of Health Research (CIHR). T.J.H. is the recipient of a Clinician-Scientist Award in Translational Research by the Burroughs Wellcome Fund and an Investigator Award from CIHR. J.M. is a recipient of a Canada Research Chair.

URLs.

Results from regression analyses at the probe set and meta–probe set levels, including gene-level plots of expression changes, and other relevant information can be found at the GRiD (Genetic Regulators in Disease) website (http://www.regulatorygenomics.org). For the probe logarithmic-intensity error model, see http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

Chapter 5: Exon-level transcriptome comparison of human and chimpanzee

David Benovoy & Jacek Majewski

Submitted to *Genome Biology* on August 12th, 2009

Connecting text

In the last chapter, we demonstrated the existence of common transcript expression variations at the isoform level in a normal human population. We showed that differences such as alternative initiation, splicing and termination were associated to common genetic single nucleotide polymorphisms (SNPs). Our results show that the effects of genetic variants on transcript expression at the isoform level are much more complex than previously believed, and constitute an important step towards understanding the functional consequences of such variations.

Given the extent of isoform variations we observed in a human population, we hypothesized that these types of variation should be prevalent between humans and chimpanzees and that some specie-specific traits evolved through regulatory modifications that control these mechanisms. In this chapter, we describe the first genome-wide comparison of transcript isoform variations between humans and chimpanzees by comparing the isoform variation in from lymphoblast cell lines between the 60 HapMap individuals used in the previous chapter and a single chimpanzee, Clint, for which the chimpanzee genome is derived from (THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM 2005).

Abstract

The sequencing and comparison of the human and chimpanzee genomes has revealed only a small number of genomic variations; yet these closely related species present many different phenotypic traits. Previous studies have begun to identify the mechanisms responsible for these differences. We found that around 58% of the 8,578 genes we defined as expressed in lymphoblast cell lines derived from these two closely related species, presented either whole-gene (34%) or isoform expression changes (24%). The major type of isoform change we observed were represented by differential inclusion of cassette exons but we also observed differences in alternative transcription initiation and polyadenylation sites. We conducted a comparative genomics analysis and showed that the presence of substitutions predicted to alter the strength of splice sites and miRNA binding sites were correlated with isoform and whole transcript expression changes. A functional gene ontology analysis revealed that these genes with expression differences affect many different pathways related to metabolism and immunity. As an example, we described in detail the expression changes that occur in the Nf-κB pathway that is activated following an infection by certain types of viruses, such as HIV-1, and discuss its possible role in conveying different susceptibility of humans and chimpanzees to AIDS. Together our results demonstrate that genomic differences between humans and chimpanzees affect transcription and pre-mRNA processing and may be responsible of certain phenotypic differences observed between these two closely related species.

Introduction

For thousands of years, humans have contemplated their uniqueness. Now with the ushering of the post-genomic era, answers to what makes us human are finally acquiring a molecular perspective. An important challenge in evolutionary biology is to identify the set of molecular characteristics that account for our unique cognitive, behavioral and physiological traits that have emerged since we last shared a common ancestor with chimpanzees, around 6 millions years ago (VIGNAUD *et al.* 2002). At the root of these differences are the molecular changes that stem from genomic variations that in turn have shaped the transcriptomes and proteomes of these species. In recent years, the sequencing and comparison of the human and chimpanzee genomes has revealed the extent of this genomic diversity. These species have accumulated around ~35 million single-nucleotide changes, 5 million insertion/deletion events, and various chromosomal rearrangements (CONSORTIUM 2005). Yet little is known about how these genomic variations translate to variations in the transcriptomes and proteomes and subsequently to overall phenotypic diversity between these two species.

The comparison of human and chimpanzee transcriptomes represents a critical first step toward understanding the evolution of species-specific phenotypes. Researchers have begun to compare gene expression profiles of humans and chimpanzees and have found remarkable diversity, particularly in testis (ENARD *et al.* 2002; KHAITOVICH *et al.* 2005). Another study has highlighted variation between humans and chimpanzees at the sub-transcript level in some genes where differential inclusion of exons produced different mRNA isoforms (CALARCO *et al.* 2007). Other types of processes can generate transcriptome variation such as alternative promoter usage where transcription is initiated at different positions or

alternative termination where the use of different polyadenylation sites marks the end of the transcript.

Here we use an exon-centric expression microarray to compare the human and chimpanzee sets of mRNA molecules from transcribed exons of protein coding gene. To illustrate the variation, we use a model system of lymphoblastoid cell lines (LCLs), which we previously used to study transcriptome diversity in humans (KWAN *et al.* 2008; KWAN *et al.* 2007). Our comparison reveals that around half of the genes expressed in LCLs present either isoform or whole-gene expression changes between humans and chimpanzees. The most common type of isoform variation is caused by alternatively spliced coding exons but we also observed expression differences in the 5' and 3' UTR regions that arise with the use of different transcription start and termination sites, respectively. We also demonstrate an association between these isoform variations and single nucleotide substitutions that occur between the genomes of these two species. We showed that these substitutions can occur in sequences that regulate splicing and gene expression, such as splice site consensus sequences, regulatory motifs, and microRNA binding sites, respectively. An *in-silico* pathway analysis revealed that isoform and whole-expression changes are often targeted immune response genes. As an example of this phenomenon, we describe the changes that occur in the Nf-κB pathway that is activated following an infection by certain types of viruses such as HIV-1 and discuss its possible role in conveying different susceptibility of humans and chimpanzees to AIDS.

Materials and Methods

Microarray data source

Human expression data was obtained from one of our previous studies where we surveyed isoform variation in humans (KWAN *et al.* 2008). This data set comprised of 57 unrelated HapMap individuals of European ancestry (INTERNATIONAL_HAPMAP_CONSORTIUM 2005). Immortalized lymphoblast cells derived from these individuals were grown in triplicate and RNA was extracted from each of these growths and hybridized onto an Affymetrix Human Exon array (n = 171) as described in (KWAN *et al.* 2008).

Four chimpanzee (*Pan troglodytes*) lymphoblast cell lines were obtained at the Coriell Cell Repositories (http://ccr.coriell.org) and processed following the same protocol that was used for the HapMap samples (see above). One of these samples was from Clint (Coriell id: S006006) who was selected for the availability of his genomic sequence (CONSORTIUM 2005) and the other three were from a family trio (Coriell ids: S003657, S003612, S003610). We prepared five (n = 5) successive cell harvest or biological replicates for Clint and one for each of the other three chimpanzees (n =1). Due to issues of probe hybridization (see below) we focused our analysis on samples derived from the chimpanzee Clint.

Noise reduction strategies

We implemented different strategies to reduce the sources of noise that often led to erroneous results. The first strategy we used was to only include probes targeted to the ~260,000 core RefSeq exons because of their high confidence annotation and to reduce the size of our data set. The second was to implement a strategy we described in our previous study (BENOVOY *et al.* 2008). Briefly, we showed that microarray studies conducted on samples with different genetic backgrounds presented high

rates of false positives hits because of mismatches between microarray probes and its intended target resulted in erroneous probe signals and subsequently lead to incorrect estimates of exon (probe set) and gene (meta-probe set) expression. To mitigate this effect, we removed probes targeted to regions that were not identical in chimp and human. The availability of the chimpanzee (Clint) genome sequence (CONSORTIUM 2005) allowed us to identify 297,017 (27%) probes targeted to core exons that contained mismatches. This step removed the majority of misbehaving probes due to inter-species mismatches, however, to remove intra-species difference, we masked out probes that targeted potential polymorphic position in our samples. Based on SNP positions of human and chimpanzee from dbSNP version 128 (http://www.ncbi.nlm.nih.gov/projects/SNP/), we identified and removed 127,087 and 9,515 probes targeted to these polymorphic positions in humans and in chimpanzees, respectively. The lower numbers of probes we identified that were targeted to known heterozygote position in chimpanzee is due to the shallower depth of SNP sampling in chimpanzee when compared to humans. Consequently, this could potentially cause more erroneous expression scores for probe set and meta-probe sets derived from probes targeted to unknown heterozygote position in Clint.

Next, we conducted a principal components analysis (PCA) on the probe expression profiles of all our samples and found that the chimpanzees from the trio were exceedingly variable (results not shown) most likely because of unknown polymorphisms that disrupt probe to target hybridization. Consequently, the chimpanzee trio was excluded from the main analysis.

Cross-hybridization was potentially another source of noise in this study because we used chimpanzee samples and the Affymetrix Human Exon

Array was optimally designed to reduce cross-hybridizing only in humans (Affymetrix). To mitigate this effect, we searched the human and chimpanzee reference genomes using default setting in Blat (KENT 2002) for matches with the probe sequences from the Affymetrix Human Exon array. We found 43,382 and 47,241 probe sequences with more than one significant hit in the human (NCBI Build 35) and chimpanzee (UCSC Build 2.1) genomes, respectively. The larger number of hits for the chimpanzee genome indicates the higher potential for chimpanzee samples to cross-hybridize with probes from the Affymetrix Humans Exon array. To mitigate this effect, we masked out any probe that had more than one significant hit in either genome.

Comparative analysis of array hybridization data

Fluorescent intensities from the remaining 680,676 probes (see above) were quantile-normalized and GC-background corrected using the Power Tools software package from Affymetrix. The normalized probe intensities from each of the arrays (n = 179) were summarized into 212,720 probe sets (representing exons expression) and 15,898 meta-probe set (representing gene expression) scores using a probe logarithm-intensity error (PLIER) model (affymetrix.com). The Exon Array also contains a large number of “antigenomic” probes that do not have a match anywhere in the genome and ideally represent a null signal. The PLIER algorithm groups these antigenomic probes by their GC-content and uses them to produce a Detection Above Background (DABG) p-value (affymetrix.com). We have also established from previous experiments (results not published) that probe sets and meta-probe sets scores with expression score < 15 were generally not expressed therefore we use this threshold along with the DABG metric to ascertain if a probe set or meta-probe set is expressed.

For the gene-level analysis, we compared genes if 50% of their exons showed a detected above background (DABG) probability ≤ 0.05, 95% of the samples had a meta-probe set score that was ≥ 15 and both these criteria were true in 95% of the samples from either the human or chimpanzee groups as suggested in (Affymetrix.com). In addition to this, we restricted our analysis to genes with a clear 1:1 orthologues ratio between human and chimpanzee as defined in (CONSORTIUM 2005) to mitigate any non-specific fluorescence from other orthologous genes. For the exon-level analysis, we defined an exon as expressed if it belonged to an expressed gene (see above) and its DABG value is ≤ 0.05. We only compared exons if their normalized intensities (probe set expression / meta-probe set expression) were between 0.2 and 5 and that the gene they are encoded from is expressed in both chimpanzee and humans but shows no statistically significant difference (see below) at the gene expression level (BEMMO *et al.* 2008). This restricted our analysis to 8,578 meta-probe sets and 51,413 probe sets.

To identify which gene or exon were differentially expressed between the HapMap and Clint samples, and because of our unbalanced experimental design, we first conducted a one-way analysis of variance (ANOVA) by grouping the expression scores of each probe set or meta-probe set into 58 groups; 57 from HapMap samples (humans) with 3 replicates each and 1 from Clint (chimpanzee) with 5 replicates. Following a significant test after false discovery rate (FDR) correction ($\alpha = 0.05$) (BENJAMINI *et al.* 2001), we specifically examined our *a priori* hypothesis by testing for expression differences between the Clint and the HapMap samples using a contrasts analysis. We constructed a contrasts matrix to partition the total variance for a given probe set or meta-probe set into variance derived from Clint and the HapMap samples using a second ANOVA. A significant test indicates that the expression derived from Clint was

significantly different than the expression derived from the 57 HapMap samples.

Classification of Transcript Isoforms

We developed an automated method (perl script available upon request) to categorize isoform changes. The algorithm first classifies probe sets into blocks according to their Refseq annotation. Each significant block is then classified as an initiation, splicing, termination or transcript expression change according to its position within the transcript (5'UTR, coding, 3'UTR or whole-gene, respectively).

Comparative Genomic Analysis

Human exonic and intronic sequences were defined using the using the RefSeq annotation file (September 2008; http://hgdownload.cse.ucsc.edu/goldenPath/hg17/) from the University of California Santa Cruz (UCSC). Orthologous chimpanzee sequences were extracted from UCSC human versus chimpanzee pair-wise alignments (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro2/) and the divergence rate was measured as the number of substitutions in aligned nucleotides divided by the total number of aligned nucleotides.

Splice Site Strength Analysis

We measured the strengths of the donor (5') and acceptor (3') splice site using the MaxEntScan program available at http://genes.mit.edu/burgelab/maxent/. This program defines the acceptor splice site as the last 20 bases from the 5' flanking intron and the first 3 bases from the 5' end of the exon. The donor splice site was defined as the last 3 bases from the 3' end of the exon and the first 6 bases of the flanking 3' intron. We used this program to scan differentially expressed exons from our analysis against a library of known donor and acceptor

splice sites (http://genes.mit.edu/burgelab/maxent/ssdata) and scored each splice site in both species using maximum entropy method (YEO and BURGE 2004). The resulting score was used to compute the difference in splice site strength between human and chimpanzee.

UTR Controlled Gene Expression

We determined the miRNA binding potential in the 3'UTR of human and chimpanzee for each of the 8,578 genes surveyed in this analysis using the MiRanda algorithm (LEWIS *et al.* 2005). The algorithm searched the 3'UTRs (defined by RefSeq, see above) of each gene in each specie against the library of human miRNA targets (version September 2008) (BETEL *et al.* 2008) available at http://www.microrna.org/microrna/getDownloads.do. We expressed the binding potential as the total score from the MiRanda output file for each gene.

Gene Ontology and Pathway Analysis

We conducted gene ontology and pathway analyses with the sets of genes that presented either whole-gene expression changes or isoform differences using the Ingenuity Pathways Analysis (IPA version 6.0) software package (Ingenuity Systems, Mountain View, CA). This software package tests the statistical significance, i.e. assigns a FDR corrected p-value to the biological functions or pathways of genes with expression differences by comparing it to a reference data set. By default, the IPA software package defines the reference data set as all genes represented on the Human Exon array. However, the use of this default reference list may cause erroneous p-value estimates because of the presence of certain experimental biases related to microarray analyses. For instance, genes that are highly expressed are less influenced by background noise compared to genes with low expression levels. This increases the power

to detect an expression change for high expressing and consequently biases our significant hits to highly expressed exons or genes. To reduce the effect of this bias, we constructed reference lists for both levels of analysis (whole-gene and isoform) were we chose genes from a random pool that presented no significant expression difference between HapMap samples and Clint but were expressed in lymphoblasts from both species. More importantly, we chose genes so that the expression distributions for the test list (genes with expression changes) and the reference list were similar. Using the expression-matched reference list, we can more accurately determine (Fisher exact test) to what degree a particular gene ontology term or functional pathway is over-represented for genes with expression changes between species.

Results

Exome comparison

The main objective of our study was to characterize transcript isoform differences between humans and chimpanzees. To asses these differences, we generated isoform expression profiles of lymphoblast cell lines (LCLs) derived from the common chimpanzee (*Pan troglodytes*) Clint (CONSORTIUM 2005) and 57 HapMap individuals (INTERNATIONAL_HAPMAP_CONSORTIUM 2005) using the Affymetrix Human Exon Array (Affymetrix). By comparing these profiles, we found a large number of differentially expressed genes (2,932 or 34.2%) from the 8,578 expressed in both species with an average fold change of 1.79. A similar number of genes (2,095 or 24.3%) with an average fold change of 1.6 showed transcript-isoform changes only without an accompanying whole-transcript expression change. These last differences represent 4,235 (8.2%) differentially expressed probe sets (exons) out of the 51,413 probe sets surveyed where the major type of change is at the splicing level (3,532 or 83.4%) with the remaining changes at the level of transcript initiation (212 or 5,8%) and termination (491 or 14%) (Figure 5.1).

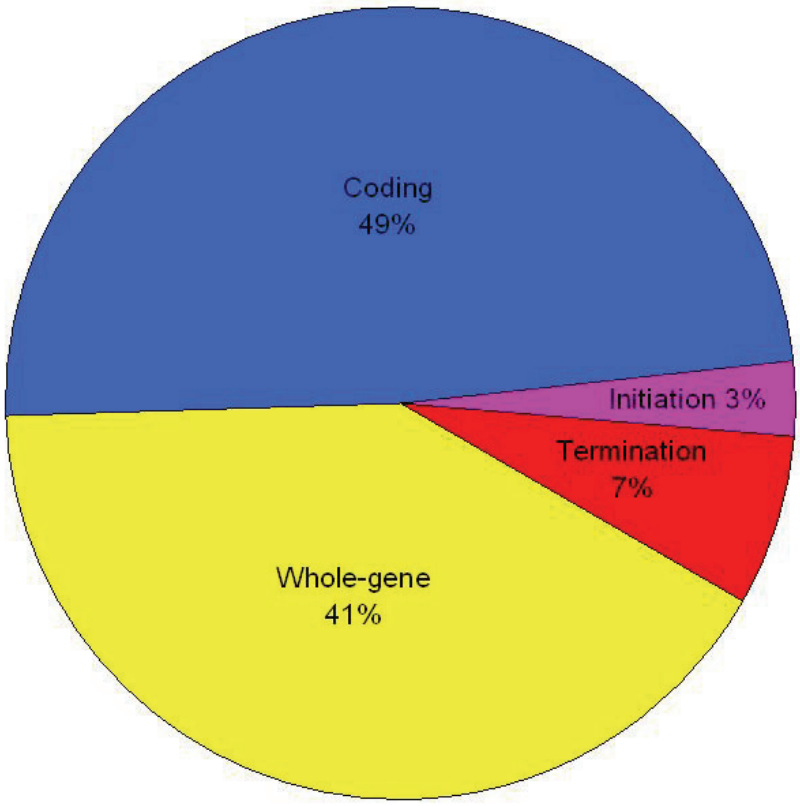


Figure 5.1: Classification of genes showing expression changes at the exon or transcript level. The 5,027 genes were classified into separate categories depending on the nature of the expression change occurring: expression changes at the whole transcript level (yellow), transcription initiation (purple), alternative splicing of a cassette exon (blue) and transcription termination changes (red).

We also compared the number of significant expression difference between the HapMap individuals and Clint to the number of significant difference within the HapMap individuals. This allowed us to estimate the ratio of inter-species divergence to intra-species diversity. We found that this divergence to diversity ratio was ~5.7 at both the whole-transcript and isoform levels. This ratio was compared with the divergence to diversity ratios calculated by (KHAITOVICH *et al.* 2005) at the whole transcript level

for different tissues such as testis (5.6), heart (2.5), kidney (2.1), liver (1.8) and brain (2.3). We find that the ratio we observe in lymphoblast is similar to what was observed in the testis which is an outlier compared to the other tissues (5.7 and 5.6 versus 1.8 to 2.5). This high ratio provides an indication although not proof that strong selection could be operating on transcript expression in lymphoblast.

Graphical visualization of the different types of expression variations mentioned above is presented in Figure 5.2. We represented the expression fold-change (blue bars) on a $\log_2$ scale between chimpanzees (only Clint) and humans (HapMap individuals) and the associated *p*-value (red bars) on a $-\log_{10}$ scale for each probe set (exon) targeted to gene *LCK* (Figure 5.2A). For this gene, each probe set is expressed at a lower level in chimpanzee, which is concordant to the meta-probe set scores ($\log_2$ scale) computed by PLIER (see methods) for humans (9.45) and chimpanzees (5.14) and represents a whole gene expression change. Figure 5.2B illustrates an example of an alternative splicing event in gene *PRKCE,* were the 9th exon exhibits lower inclusion levels in chimpanzee. In Figure 5.2C and 5.2D, we show examples of alternative transcript initiation and termination. In these examples a probe set from a group of probe sets targeted to the same UTR exon is differentially expressed. For gene *TTRAP* (Figure 5.2C) and gene *TMEM63A* (Figure 5.2D) we predict that they produce distinct isoforms in chimpanzees and humans by using different transcription initiation start sites and different polyadenylation sites, respectively.

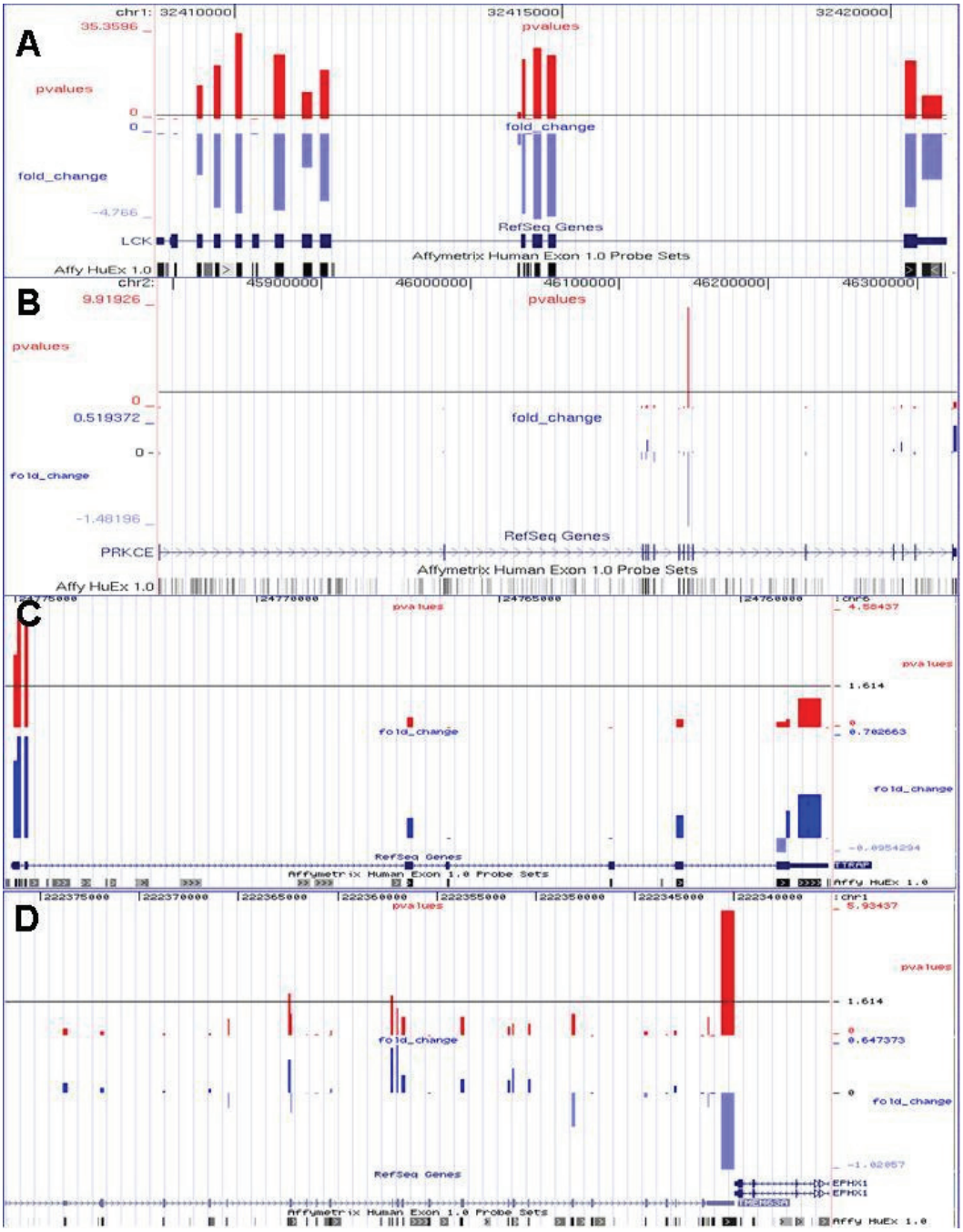


Figure 5.2: Visualization of expression data. Custom view of the UCSC browser with expression data overlaid onto gene structures. Red vertical bars represent the p-value (-log10 scale) derived from the contrast analysis for each probe set for a given gene. Blue vertical bars represent the expression fold-change between Clint and the HapMap samples for each probe set. A. *LCK* is an example of a whole-gene expression change

where all probe sets from the HapMap samples are all expressed at higher levels than in Clint. B. The PRKCE gene is an example of an alternative spliced cassette exon. C. The *TTRAP* gene is an example of alternative initiation where the longer isoform is expressed in Clint. D. The *EPHX1* gene is an example of alternative termination.

Genome-wide microarray analyses, like the one conducted here, are difficult to adequately validate using classical low-throughput experiments such as RT-PCR because of cost and time issues associated to conducting hundreds of these experiments. To circumvent this problem, we used an *in-silico* genome-wide validation method where we compared the 51,413 "expressed" probe sets surveyed in this study (see methods) to a data set of known splicing events derived from EST evidence. We used the "Alt-Splicing" track from the UCSC genome browser that lists known examples of splicing and other transcript isoform events (KAROLCHIK *et al.* 2008) and found that the differentially expressed probe sets from our study were significantly overrepresented as compared to a random expectation (odds ratio = 2.23 (0.11 / 0.049); Chi-square analysis: X^2 = 90.91; *p*-value < 2.2×10^{-16}) in this list. This indicates that our analysis preferentially identifies exons with prior evidence of alternative splicing or alternative inclusion within transcripts.

In addition to this, we examined how the exons and genes that presented the most significantly divergent expression profiles between the chimpanzee Clint and the HapMap individuals behaved in the other 3 chimpanzees that were excluded from the main analysis because of hybridization issues (see methods). Out of the top 10 exons and genes with the most significant (FDR correction at $\alpha = 0.05$) expression differences between the chimpanzee Clint and the HapMap individuals, we found that 80% and 100% of these exons and genes, respectively,

also presented significant expression differences between these 3 chimpanzees and the HapMap individuals. For the top 100 hits the concordance is still good because we found that 75% of the exons and 90% of the genes were also significant between the 3 chimpanzees and the HapMap individuals. These observations are good indications that the expression differences observed between the chimpanzee Clint and the HapMap individuals are not unique to Clint because the majority of the top hits have been validated in 3 other chimpanzees and they potentially represent true inter-species expression variations.

Comparative Genomics Analysis

We hypothesized that the differences in splicing profiles we observed between humans and chimpanzees were in part due to nucleotide substitution that disrupted *cis*-regulatory splicing elements such as splicing enhancers and silencers (BLENCOWE 2006; BRUDNO *et al.* 2001; CALARCO *et al.* 2007; MAJEWSKI and OTT 2002; MATLIN *et al.* 2005; YEO *et al.* 2004b; ZHANG *et al.* 2003). Given that these short, degenerate regulatory elements are over-represented in exonic and intronic regions near the splice sites, we determined the sequence divergence for the entire exon and 150 bp upstream and downstream of the flanking intronic sequences. We found that the sequence divergence was significantly higher (Mann-Whitney; W = 45655458, p-value < 2.2×10^{-16}) for exons that presented significant expression differences (mean sequence divergence = 0.66%) than for exon that were expressed at the same level (mean sequence divergence = 0.46%) between these closely related species. This result indicates that elevated sequence divergence in exonic and intronic regions are correlated with an increased expression divergence and suggests that genetic differences between these species are responsible for some of the differential isoform expression.

Contrary to splicing enhancer and silencers, the donor (5') and acceptor (3') splice site motifs are well characterized in mammals. Therefore, we specifically measured the different splicing potential of these motifs in human and chimpanzee for each differentially expressed exon. We found that these expression differences were significantly correlated to differences in donor splice site strength (Spearman correlation; *rho* = 0.27; *p*-value = 0.010). An example of this phenomenon is illustrated in Figure 5.3 were we show that the donor splice site for the 4th exon of the *C14ORF159* gene is weakened in chimpanzee (MaxEnt score = 0.57) compared to its human orthologue (MaxEnt score = 8.76) by a G to A substitution. Consequently, this substitution is most likely responsible for the lower inclusion of this exon in chimpanzee.

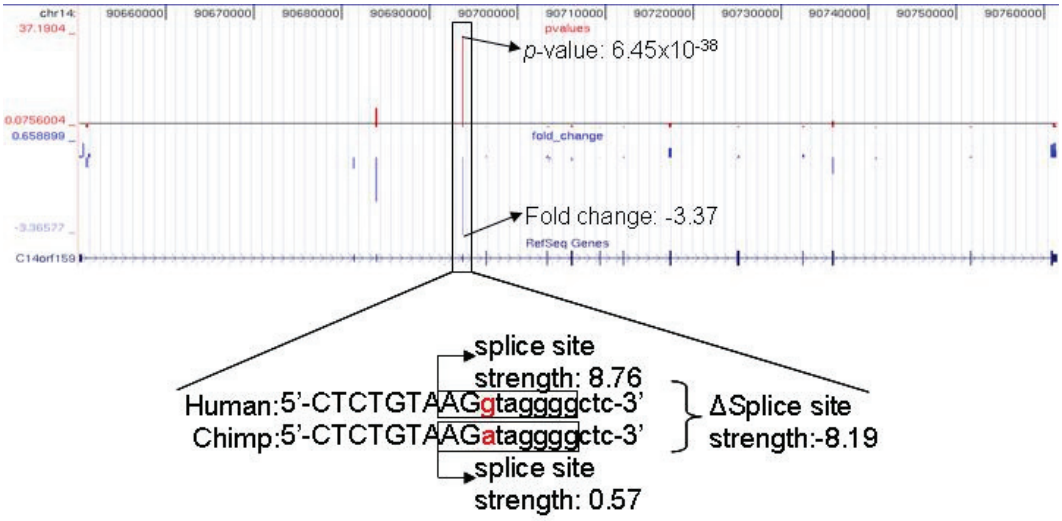


Figure 5.3: Effect of a substitution in the splice site. This example illustrates the effect a G/A substitution in one of the exons of the *C14orf159* gene. The presence of an A in the first base of the intron disrupts the splicing of the exon and consequently lowers the expression of this exon in chimpanzee.

Recent studies have shown that gene expression can be regulated in a post-transcriptional matter by miRNAs (NEILSON and SHARP 2008; SANDBERG *et al.* 2008). We believed that substitutions that disrupt miRNA binding sites in an mRNA transcript would render it less prone to degradation by the dicer pathway and consequently we would detect that transcript to be differentially expressed. This is in fact what we observed when we compared the miRNA binding potential of the human and chimpanzee 3'UTRs (see Materials and Methods). We found that differentially expressed genes had significantly higher differences (Mann-Whitney test; W = 7164348; *p*-value = 2.7×10^{-4}) in binding potential (mean binding potential = 1542.71) compared to genes with no expression differences (mean binding potential = 1348.613).

Gene ontology analysis

We performed a network analysis using the Ingenuity Pathways Analysis (IPA) system on the sets of genes that had either different isoform or whole-gene expression differences between humans and chimpanzees. We observed interesting differences between these two types of expression variation and their related pathways. Many genes with whole transcript expression changes were related to energy metabolism such as carbohydrate synthesis and degradation pathways (fructose, mannose, galactose, starch and sucrose metabolism; Table 5.1) whereas genes with isoform differences were more implicated in signalling pathways (killer cell, B-cell, IL-8, IL-4 and IL-2, NF-κB) related to immunity (OTT *et al.* 1998; SCHRAM and ROTHSTEIN 2003; TRIVEDI *et al.* 2001).

Table 5.1: Top 20 over-represented canonical pathways for genes with isoform differences or whole-transcript expression differences

Type of expression change	Canonical pathway	# of genes with expression changes (Total # of gene in pathway)	p-value
Exon-level analysis	Tight Junction Signaling	29 (160)	2.04E-04
	Estrogen Receptor Signaling	26 (121)	3.80E-03
	Erythropoietin Signaling	18 (75)	1.15E-02
	Role of NFAT in Regulation of the Immune Response	32 (185)	1.15E-02
	Cysteine Metabolism	7 (83)	1.23E-02
	Aminoacyl-tRNA Biosynthesis	15 (83)	1.26E-02
	Protein Ubiquitination Pathway	43 (205)	1.35E-02
	Huntington's Disease Signaling	37 (228)	1.55E-02
	Cell Cycle: G1/S Checkpoint Regulation	15 (57)	1.66E-02
	Butanoate Metabolism	13 (126)	1.78E-02
	FcγRIIB Signaling in B Lymphocytes	11 (52)	1.78E-02
	CCR5 Signaling in Macrophages	11 (85)	1.78E-02
	Nucleotide Excision Repair Pathway	11 (35)	1.78E-02
	Alanine and Aspartate Metabolism	11 (85)	1.78E-02
	Ceramide Signaling	17 (82)	2.00E-02
	PPARα/RXRα Activation	27 (168)	2.29E-02
	IL-8 Signaling	31 (181)	2.82E-02
	Leukocyte Extravasation Signaling	27 (189)	3.24E-02
	IL-15 Production	7 (29)	3.39E-02
	CTLA4 Signaling in Cytotoxic T Lymphocytes	18 (85)	3.63E-02
	Fructose and Mannose Metabolism	17 (131)	2.88E-05
	IL-10 Signaling	21 (71)	1.02E-03
	Xenobiotic Metabolism Signaling	45 (241)	1.66E-03
	Purine Metabolism	72 (412)	1.78E-03

Transcript-level analysis	Arginine and Proline Metabolism	15 (177)	3.09E-03
	Galactose Metabolism	14 (107)	3.31E-03
	N-Glycan Biosynthesis	19 (87)	8.13E-03
	NF-κB Activation by Viruses	23 (80)	9.77E-03
	PXR/RXR Activation	13 (81)	1.12E-02
	Axonal Guidance Signaling	54 (392)	1.23E-02
	Cardiac β-adrenergic Signaling	23 (136)	1.35E-02
	Actin Cytoskeleton Signaling	40 (221)	1.38E-02
	α-Adrenergic Signaling	20 (104)	1.48E-02
	CD27 Signaling in Lymphocytes	17 (49)	1.58E-02
	Starch and Sucrose Metabolism	14 (181)	1.58E-02
	Phototransduction Pathway	8 (62)	1.82E-02
	Urea Cycle and Metabolism of Amino Groups	7 (80)	1.91E-02
	Activation of IRF by Cytosolic Pattern Recognition Receptors	19 (70)	2.34E-02
	Aminosugars Metabolism	14 (103)	2.40E-02
	B Cell Receptor Signaling	45 (153)	2.40E-02

Interestingly some of the pathways mentioned above are involved in HIV-1 infection. We examined in detail one important pathway that is involved in HIV-1 infection; the NF-κB signalling pathway (Figure 5.4). Activation of this pathway can be induced by HIV-1 proteins that interact with the TNF receptor (HERBEIN and KHAN 2008). Once the NF-κB transcription factor is activated it initiates and enhances HIV-1 gene expression in infected cells by binding to the long terminal repeats (LTR) of HIV-1 (HERBEIN and KHAN 2008; TERGAONKAR 2006). Many genes associated with this pathway presented isoform and expression differences or both (highlighted in yellow, blue and red, respectively, in Figure 5.4). Detailed expression profiles of three genes that play important roles (BELTINGER *et al.* 1996; CHAN *et al.* 2000; CHENG *et al.* 1999; RODRIGUES-LIMA *et al.* 2001;

STUMPTNER-CUVELETTE *et al.* 2003) in the activation of this pathway are presented in Figure 5.2A, 5.2B and 5.2C. Together, these figures illustrate the different types of transcript changes as well as the amount of transcriptome diversity that can be present in a pathway.

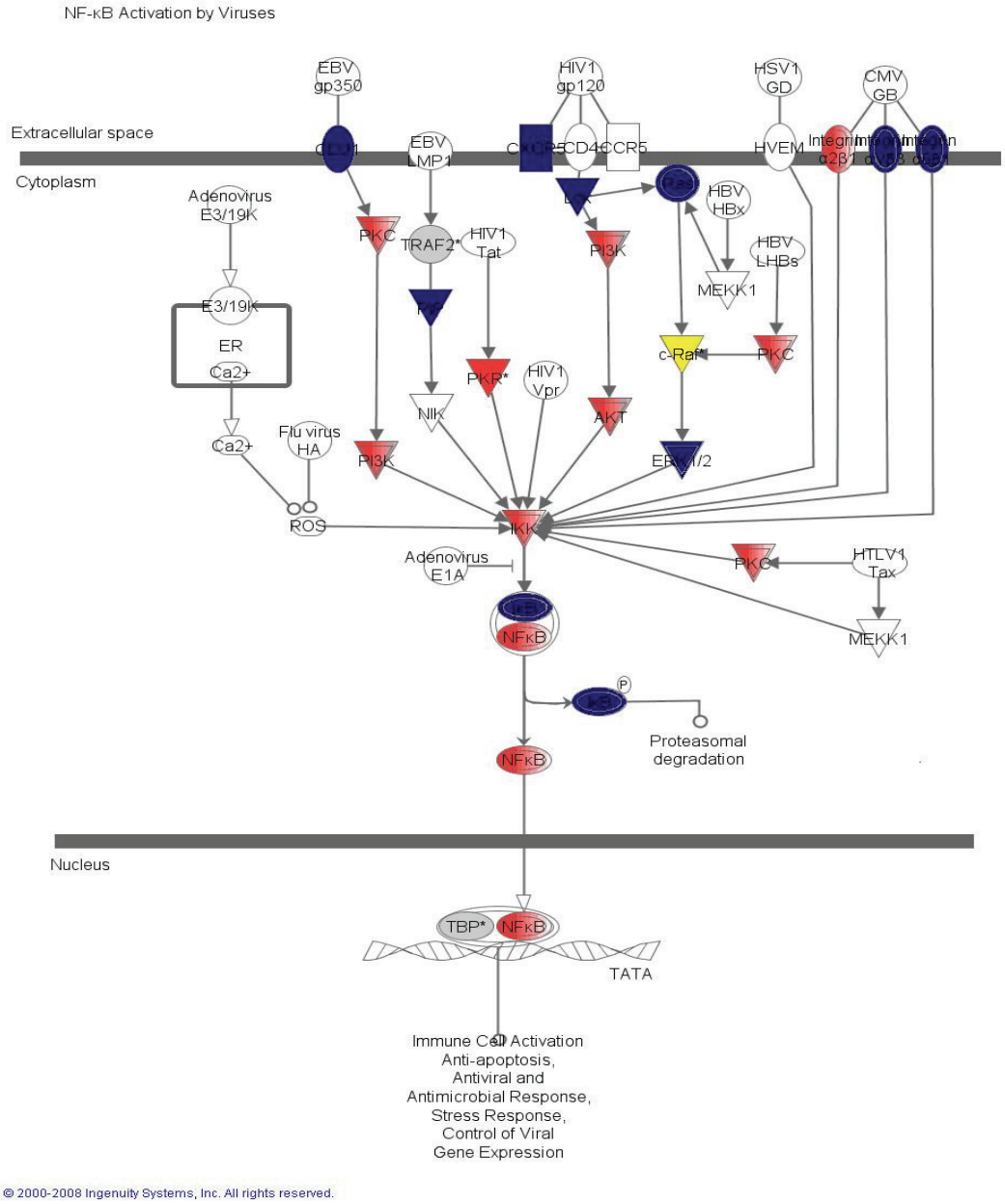


Figure 5.4: Expression changes in the NF-κB pathway. Genes are colored according to the types of expression change. Whole-gene expression changes are colored in blue, isoform changes are colored in yellow and genes with both isoform and whole-gene expression changes are colored in red Colored in red.

Discussion

To our knowledge this study is the most comprehensive analysis of transcriptome variation between humans and chimpanzees. The transcriptome of these two closely related species vary from each other in part because of differences in RNA transcription that stem from differential transcript expression to mRNA processing variations such as pre-mRNA splicing, transcription initiation and alternative termination. These types of variations such as the amount mRNA produced by a genetic locus or the exclusion of a coding exon by alternative splicing can greatly alter the concentration and sequence of protein, respectively, and as a result its function. Also, alternative splicing of a 5' or 3' UTR, in addition to alternative pre-mRNA transcript initiation or termination, may add or remove regulatory sequences that influence mRNA stability, mRNA localization and translational efficiency (KOZAK 1983). Through these processes, evolution has diverged the transcriptome of these two closely related species to a point were we estimate that ~59% of the genes expressed in lymphoblast cells produce transcript with structural or expression variants between humans and chimpanzees.

One goal of the chimpanzee genome sequencing project (CONSORTIUM 2005) was to undertake a comparative analysis with the human genome in order to identify and catalogue human-chimpanzee genomic differences. Hidden among these differences are functional changes that underlie the phenotypic diversity between these two species. The challenge now is to identify how these differences have created the phenotypic diversity observed between humans and chimpanzees. We have shown in previous studies that cis-regulatory single nucleotide polymorphisms were associated to transcript isoform variations in a human population (KWAN *et al.* 2008; KWAN *et al.* 2007) and other studies have shown that these polymorphisms were associated to gene expression variation (CHEUNG *et*

al. 2005; SPIELMAN *et al.* 2007; STRANGER *et al.* 2005). In this study, we demonstrate that disruption of certain regulatory sequences such as splice site motifs and microRNA binding sites by single nucleotide substitutions are correlated with transcript structural and gene expression changes between humans and chimpanzees. From this evidence, future studies should be undertaken to definitively establish associations between genomic and transcriptome variations to further our understanding of human evolution.

Another important challenge is to determine whether changes at the transcriptome level have any phenotypic effect; that is whether they are neutral or under selection. System or network approaches attempt to resolve this issue by placing genes in their functional context and identify networks that have accumulated genes with structural and expression changes more than would be expected by chance. When these genes accumulate in a specific network, we assume that the resulting phenotypic trait encoded by this network is under selection. One interesting observation that emerged from our network analysis is that isoform and whole-gene expression changes tend to affect signalling pathways but more specifically immune response pathways. Given the number of immune related pathways affected by these changes, we propose that the immune systems of humans and chimpanzees have undergone important evolutionary adaptations caused by changes in isoform and whole-gene expression and consequently may respond differently to infectious agents. For example, one striking immunological difference between these closely related species is their response to HIV infection. In fact, once infected with HIV-1 (human immunodeficiency virus) humans will usually develop AIDS (acquired immune deficiency syndrome) whereas chimpanzees infected with a closely related HIV-1 strain, SIV_{chp} (simian immunodeficiency virus) will rarely exhibit symptoms related to AIDS

(HEENEY *et al.* 2006; TEN HAAFT *et al.* 2001). Interestingly, many genes from our study present isoform and whole-gene expression differences that are related to the NF-κB pathway that plays an important role in HIV-1 infection. The exact relationship between these expression differences and varying susceptibility of humans and chimpanzees to the development of AIDS is only speculative at this stage and will require more detailed analyses to establish a clear association, if any. However, we presented this example to illustrate how the exomes have evolved to potentially create distinct phenotypic traits in humans and chimpanzees.

Ultimately, understanding our unique physiological, cognitive and social characteristics, i.e. what makes us human, will require us to connect specific genomic variations to the phenotypes that are most relevant to our evolution. Comparisons like the one conducted here help to reveal the molecular basis for these phenotypic traits as well as the evolutionary forces that have shaped our species. Systematic comparison of other tissues from chimpanzee and other species will likely reveal new important functional pathways that contribute to our uniqueness and help us explain certain variations and abnormalities that lead to diseases.

Acknowlegment

The authors wish to thank E. Eichler for providing us with the chimpanzee cell lines and T. Kwan and C. Murrie for helpful discussions and critical reading of the manuscript. This work is supported by Genome Canada, Genome Quebec and the Canadian Institutes of Health Research (CIHR). J. M is a recipient of a Canada Research Chair. Funding to pay the Open Access publication charges for this article was provided by CIHR.

Chapter 6: Alternative isoform detection using Exon arrays

Exerts taken from

David Benovoy, Amandine Bemmo, Tony Kwan, Daniel J Gaffney, Roderick V Jensen and Jacek Majewski. 2008. Gene Expression and Isoform Variation Analysis using Affymetrix Exon Arrays. *BMC Genomics*. **9**:529.

And

Jacek Majewski, David Benovoy and Tony Kwan. Alternative Isoform Detection Using Exon Arrays. 2009. *Handbook of Research on Systems Biology Applications in Medicine*. Information Science Reference. p.262-277. Hershey, New York.

Connecting text

In the preceding chapter (chapter 5) we showed that transcript isoform variations were common between humans and chimpanzees. We demonstrated that these isoform variations were correlated with genetic differences in certain regulatory motifs. We also showed that these isoform variations were associated with species-specific phenotypic traits and more specifically differences in immune responses. We conclude by proposing that the variation of transcript isoforms regulation is responsible, in part, for the divergence and evolution of these closely related species. These last three chapters end the biological portion of our studies.

The next two chapters present the methodological aspect related to the analysis of data generated with the Affymetrix Human Exon Array. In them, we outline some of the problems we encountered during the analyses presented in the previous three chapters and describe solutions that we developed to overcome them.

Abstract

Eukaryotic genes have the ability to produce several distinct products from a single genomic locus. Recent developments in microarray technology allow monitoring of such isoform variation at a genome-wide scale. These types of experiments generate huge amounts of complex data that in turn create analytical issues that need to be solved. Here, we demonstrates how to analyses data generated with the Exon array using the well studied Quality Control (MAQC) dataset. We outline the analysis involved in detecting alternative mRNA isoforms and point out solutions to problems that may be encountered by researches using this technology.

Introduction

Alternative pre-mRNA splicing is a process allowing the production of several distinct gene isoforms from a single genomic locus. The most common type of alternative splicing events in mammals results in cassette exons, where each such exon can be either included or excluded from the mature mRNA. Other events include alternative use of donor or acceptor splice sites, and intron retention. In addition, processes such as alternative promoter usage and alternative polyadenylation, resulting in differences in initiation and termination of the transcript, respectively, further diversify eukaryotic transcriptomes and proteomes. As researchers are becoming aware of the importance of splicing and mRNA processing in generating transcriptome diversity, isoform-sensitive microarrays are rapidly gaining popularity in gene expression analysis (FREY *et al.* 2005; LEE and ROY 2004).

Splicing sensitive microarrays employ a number of exon body oligonucleotide probes, or exon junction probes, or a combination of the two designs, to determine mRNA levels at the resolution of a single exon or splice site. The Affymetrix GeneChip® Human Exon 1.0 ST Array is the

first commercially available microarray product designed for genome-wide, exon level expression analysis. The array relies on targeting multiple probes to individual exons and allows simultaneous, exon-level detection of expression intensity for 1.4 million probe sets covering over 1 million known and predicted human exons. The Exon Array is a flexible tool, which can be used to perform the function of classical expression arrays and concurrently provide supplementary information on isoform changes. This level of data complexity has introduced the need to develop new statistical and computational tools capable of distinguishing between gene expression differences and isoform differences, and this at the genome wide level.

In this chapter, we will use the example of a well studied system in order to outline the flow of the analysis required to process Exon Arrays, outline problems which may be encountered by potential users of the chips, and describe solutions that we have developed to overcome such problems.. We use the brain and reference human mRNA samples previously studied by the MicroArray Quality Control (MAQC) consortium (CANALES *et al.* 2006; SHI *et al.* 2006). These commercially available samples provide a high quality reference dataset for comparing microarray results across various platforms and laboratories. The human brain has very distinct gene expression signatures, and the comparison with the reference (combined) tissue pool results in detection of numerous genes with differential expression at the isoform level.

Methods

Exon Array Hybridization

The Universal Human Reference RNA (catalogue no. 740000) and Human Brain Reference RNA (catalogue no. 6050) were obtained from Stratagene and ambion, respectively. The RNA quality was assessed

using RNA 6000 nanoChips with the Agilent 2100 Bioanalyzer (Agilent, Palo Alto, USA). Five technical replicates of each sample were hybridized independently at two test sites: McGill University and Genome Quebec Innovation Centre (Montreal, Quebec, Canada) and Virginia Tech (Blacksburg, Virginia, USA). Biotin-labelled target for the microarray experiment were prepared using 1 µg of total RNA. The RNA was subjected to an rRNA removal procedure with the RiboMinus human/Mouse Transcriptome Isolation Kit (Invitrogen) and cDNA was synthesized using the GeneChip® WT (Whole Transcript) Sense Target Labelling and Control Reagents kit as described by the manufacturer (Affymetrix). The sense cDNA was then fragmented by UDG (uracil DNA glycosylase) and APE 1(apurinic/apyrimidic endonuclease 1) and biotin-labelled with TdT (terminal deoxynucleotidyl transferase) using the GeneChip® WT terminal labelling kit (Affymetrix, Santa Clara, USA). Hybridization was performed using 5 micrograms of biotinylated target, which was incubated with the GeneChip® Human Exon 1.0 ST array (Affymetrix) at 45°C for 16–20 hours. Following hybridization, non-specifically bound material was removed by washing and detection of specifically bound target was performed using the GeneChip® Hybridization, Wash and Stain kit, and the GeneChip® Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip® Scanner 3000 7G (Affymetrix) and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix). The microarray data has been deposited in the Gene Expression Omnibus Database (GEO: GSE13072).

Data Pre-processing and Analysis

The Affymetrix Power Tools software package (Affymetrix) was used to quantile normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta-probe set

(representing gene expression) intensities using a probe logarithmic intensity error model (PLIER, www.affymetrix.com) or robust multichip analysis (RMA, (IRIZARRY *et al.* 2003b)). The above procedures were carried out separately for the two test sites (McGill University and Virginia Tech). The raw data (.cel files) was downloaded from the MAQC website for the Illumina and U133 arrays. In order to keep the number of replicates and test sites consistent across platforms, we only used two of the MAQC test sites (a total of 10 technical replicates of each sample). For the probe set-level analysis and alternative isoform detection, we only used the most confident subset of core probe sets from the Exon Array.

Probe set and Gene Mapping

To determine a subset of genes common to the three platforms, we used the mapping provided by the MAQC study (SHI *et al.* 2006) to select 12091 probe sets common Illumina and Affymetrix U133 arrays. Subsequently, we used the Exon Array probe set annotation and retained only the genes where the Exon Array meta-probe set coordinates contained both the Illumina and U133 probe sets. This procedure resulted in 8391 genes with a high confidence concordant mapping across the three platforms.

Results

Variability across labs

Five technical replicates of brain and reference were hybridized in two independent labs: McGill University (MU) and Virginia Tech (VT), for a total of 20 samples. Principal component analysis, which is a commonly used method to visualize sources of variability in the data, is shown in Figure 6.1.

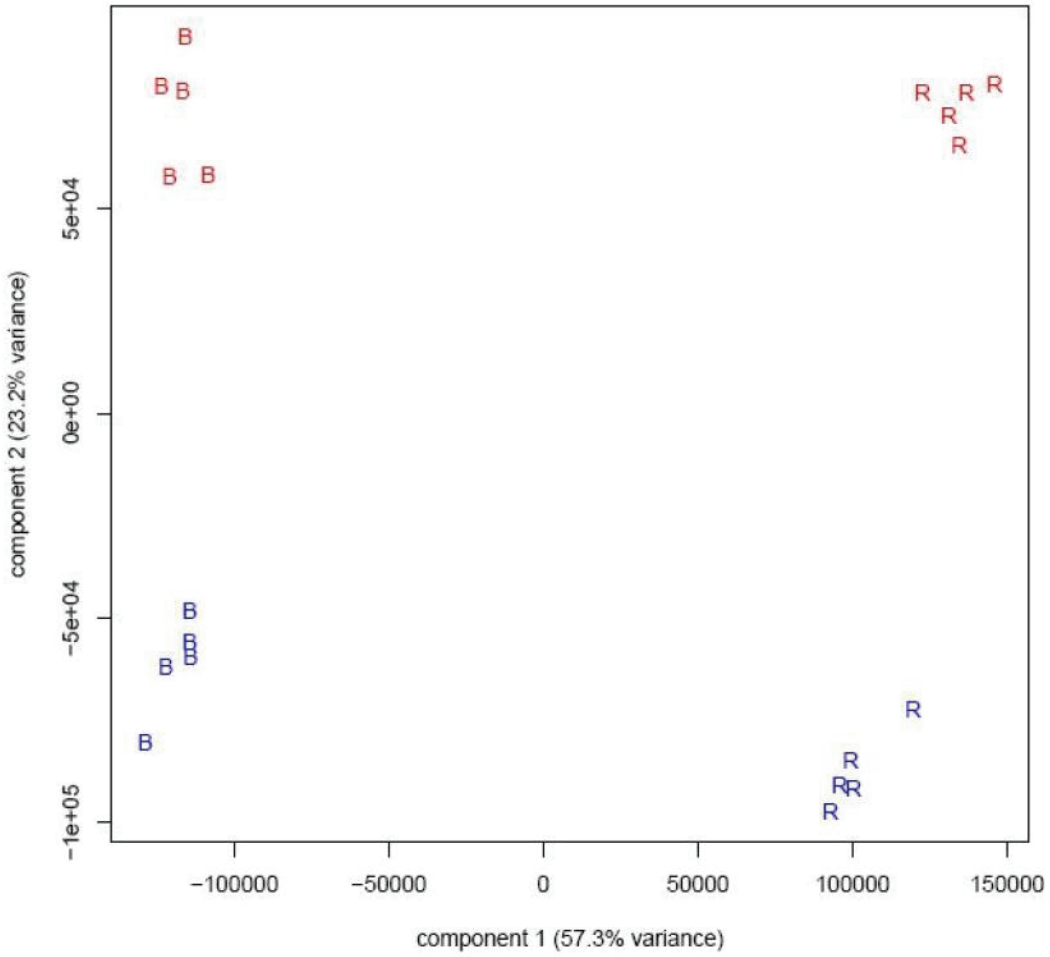


Figure 6.1: PCA plots at the probe set level show two main sources of variation among the 20 samples. The first principal component explains 57% of the variance and corresponds, as expected, to the biological source of the sample: brain (B) vs. reference (R). The second principal component explains 23% of the variance and corresponds to the "lab effect" between VT (blue), and McGill (red) – that is, it illustrates the technical variability across labs.

Our experience with Exon Arrays indicates that in general the ribosomal RNA reduction step is the most inconsistent part of the protocol and is likely to be a major contributor to the differences across labs. Variability in hybridization intensities, background noise, and random errors across labs may contribute to differences in final conclusions resulting from microarray

analyses. In the case of the MAQC data, the final goal was to quantify differences in gene expression levels between the human brain and reference tissues. A relevant metric of such expression difference is the fold change (FC), calculated as FC = $Expression_{Brain}$ / $Expression_{Reference}$. In Figure 6.2, we show a correlation plot comparing the calculated fold changes in genes expression between the two labs. Despite the inter-lab variability in expression levels shown in the PCA plots, the final results (fold changes) are highly consistent for the two labs, with a correlation coefficient of greater than 0.97.

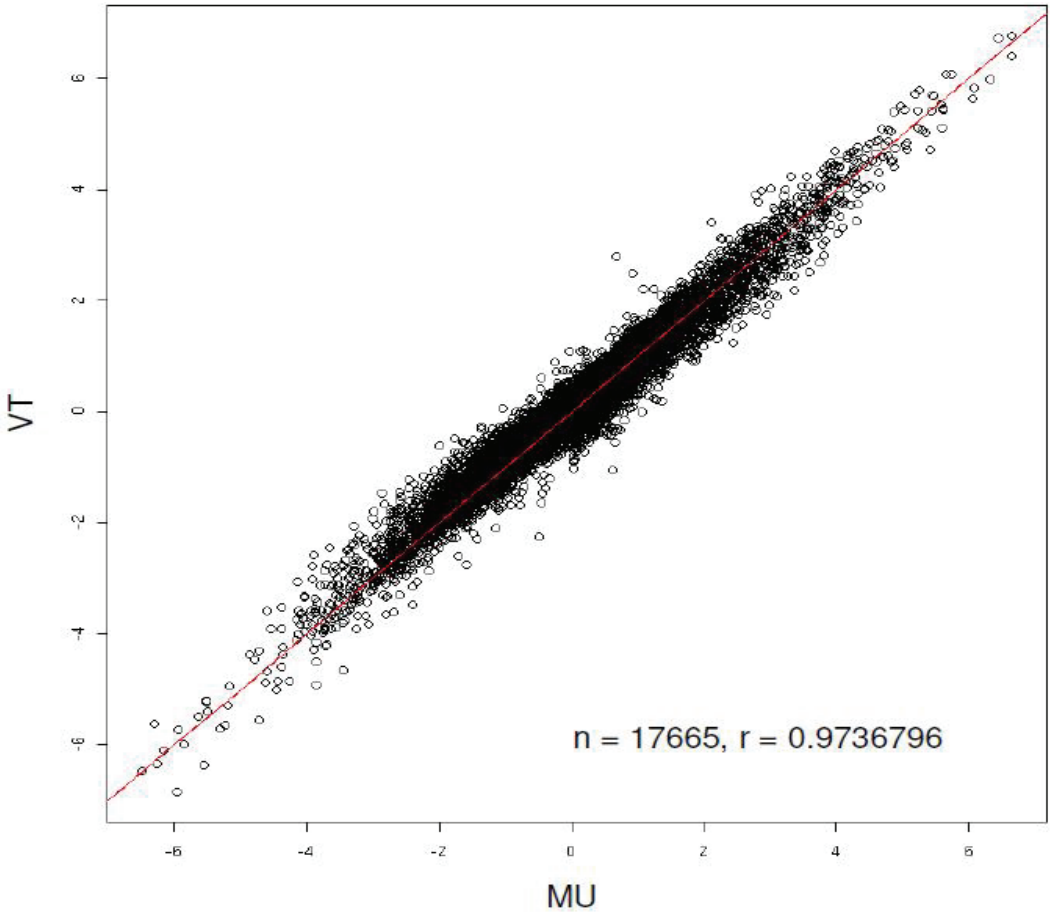


Figure 6.2: Comparison of $\log_2$ (FC) detected between the biological samples for the two labs. Despite significant variation in expression measure across test sites, the fold change estimates are highly correlated.

Variability across summarization methods

The aim of the summarization step in microarray analysis is generally to combine signals from multiple probes, which target the same expression unit, into a single expression index. Most of the popular methods strive for robustness against outlier probes (e.g. cross hybridizing, saturated, or non-responsive probes). We used our fold change results to compare two commonly used summarization methods: PLIER and RMA. We noted that RMA does result in a slight compression of fold changes, as has been observed in prior studies using other microarray platforms (CANALES *et al.* 2006). However, we find that the correlation of fold changes obtained from the two approaches is very high (r = 0.99).

Variability across platforms

The original MAQC studies demonstrated that microarray results are highly consistent across different platforms (CANALES *et al.* 2006). In Figure 6.3, we compare the performance of the Exon Array in determining gene expression levels with two other popular platforms previously used by MAQC: Illumina Bead Array and Affymetrix U133 Gene Chip. In order to facilitate comparison across labs as well as platforms, we selected a number of genes which are reliably annotated and targeted by a common set of probe sets (see Methods).

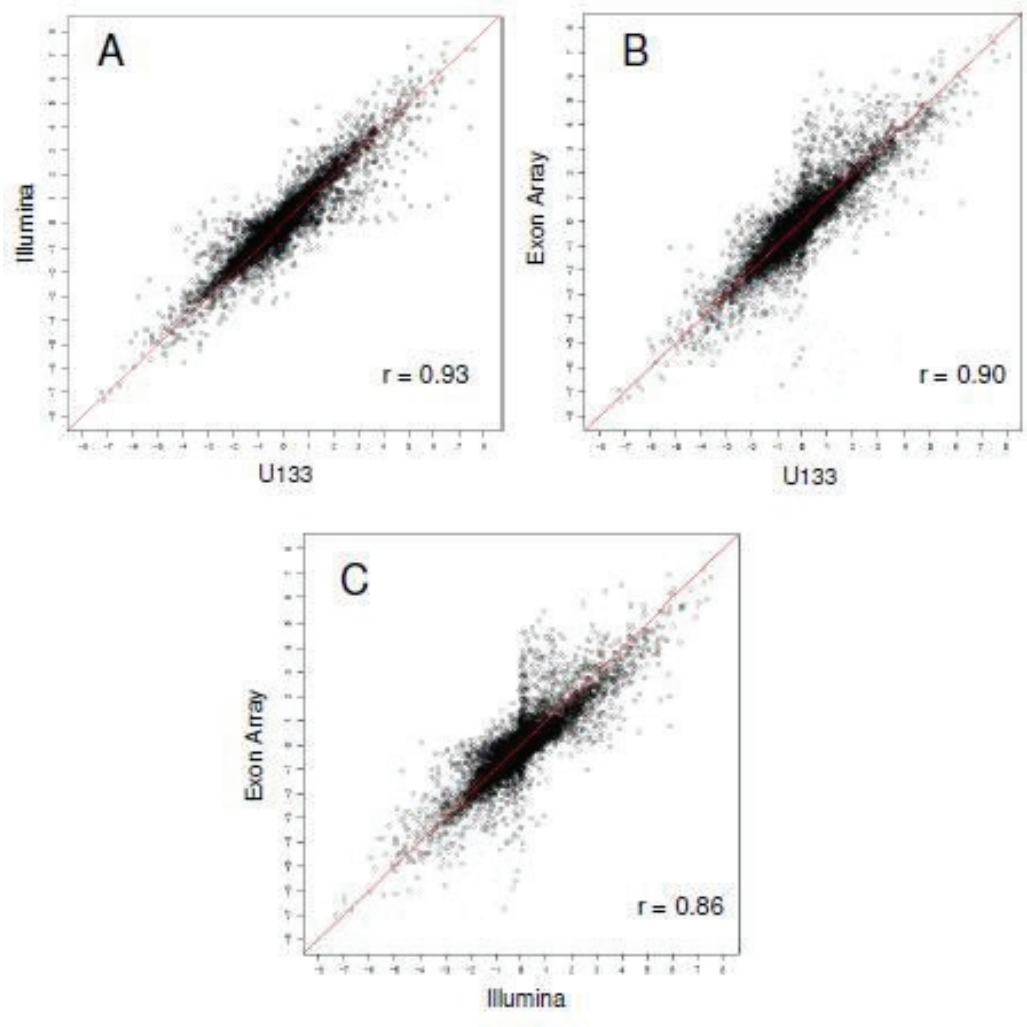


Figure 6.3: Correlation of fold changes between Affymetrix U133, Illumina, and the Affymetrix Exon Array. Fold changes (log2 transformed) between brain and reference expression levels for 8391 genes common to all three platforms: A) Illumina vs. U133. B) Exon Array vs. U133, C) Exon Array vs. Illumina.

For the Exon Arrays, the fold changes were calculated by combining the results from the two labs (MU and VT). For the sake of consistency in the comparison, two test sites were chosen at random and combined for each platform within the MAQC dataset. We find that the 3' targeted platforms, Illumina Human-6 BeadChip and Affymetrix U133, produce the most consistent results (R = 0.92). This is not surprising, since the probe

selection regions for the two platforms largely coincide, and the amplification protocols are poly-A primed and biased towards the 3' ends of genes. The correlation with the Exon Array is slightly lower: R = 0.89 for U133 and 0.85 for Illumina. It has been previously shown (OKONIEWSKI *et al.* 2007a; ROBINSON and SPEED 2007; XING *et al.* 2007), that the Exon Arrays are effective tools for gene expression profiling. Therefore, it is of interest, to examine the main sources of differences between the Exon Arrays and other platforms. Thus, in the analysis below we will concentrate on the genes whose predicted expression patterns are not consistent across platforms. In particular, the Exon Array is able to distinguish between specific isoforms of a given genomic locus, whereas the Illumina and Affymetrix U133 platforms generally target only a single isoform.

Alternative Isoform Detection

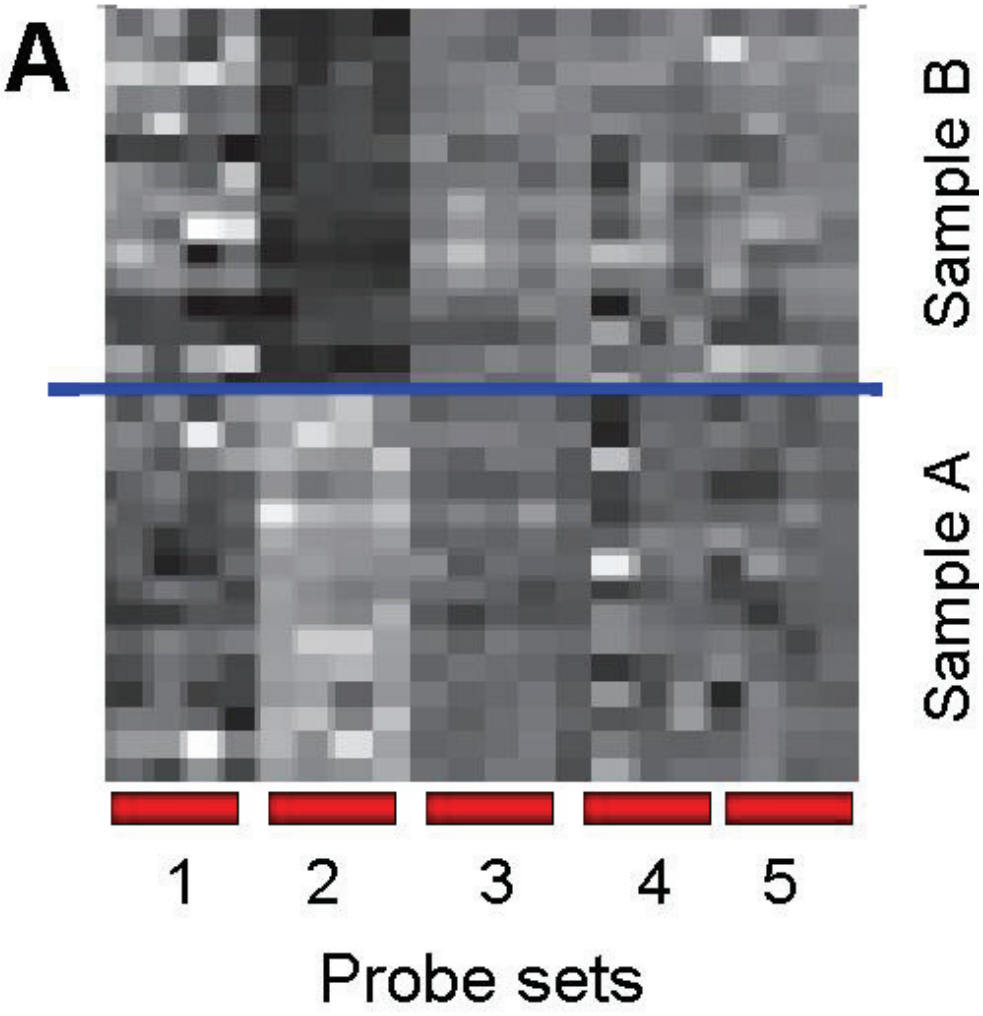
It has previously been pointed out that some discordant results in the original MAQC (CANALES *et al.* 2006) study were caused by differential isoform expression and differences in probe placement across platforms. One particular discordant gene, ELAVL1, was suspected to express two alternative isoforms, differing in the 3' UTR region. In Figure 6.4C, we use the example of ELAVL1 to illustrate the advantages of using the Exon Array for profiling individual isoforms. It is clear that although the Exon Array does not report the entire gene as differentially expressed, individual probe sets within the gene reach high statistical significance levels (*p-value* < 10^{-9}). More interestingly, the gene appears to be composed of two "blocks", with the first block on the 3' end showing elevated expression in the brain, while the second block has elevated expression in the reference sample. In order to understand the more precise nature of this isoform change, it is advantageous to visualize this data in the context of known gene annotation, EST, and mRNA data. Generally, our lab uses the

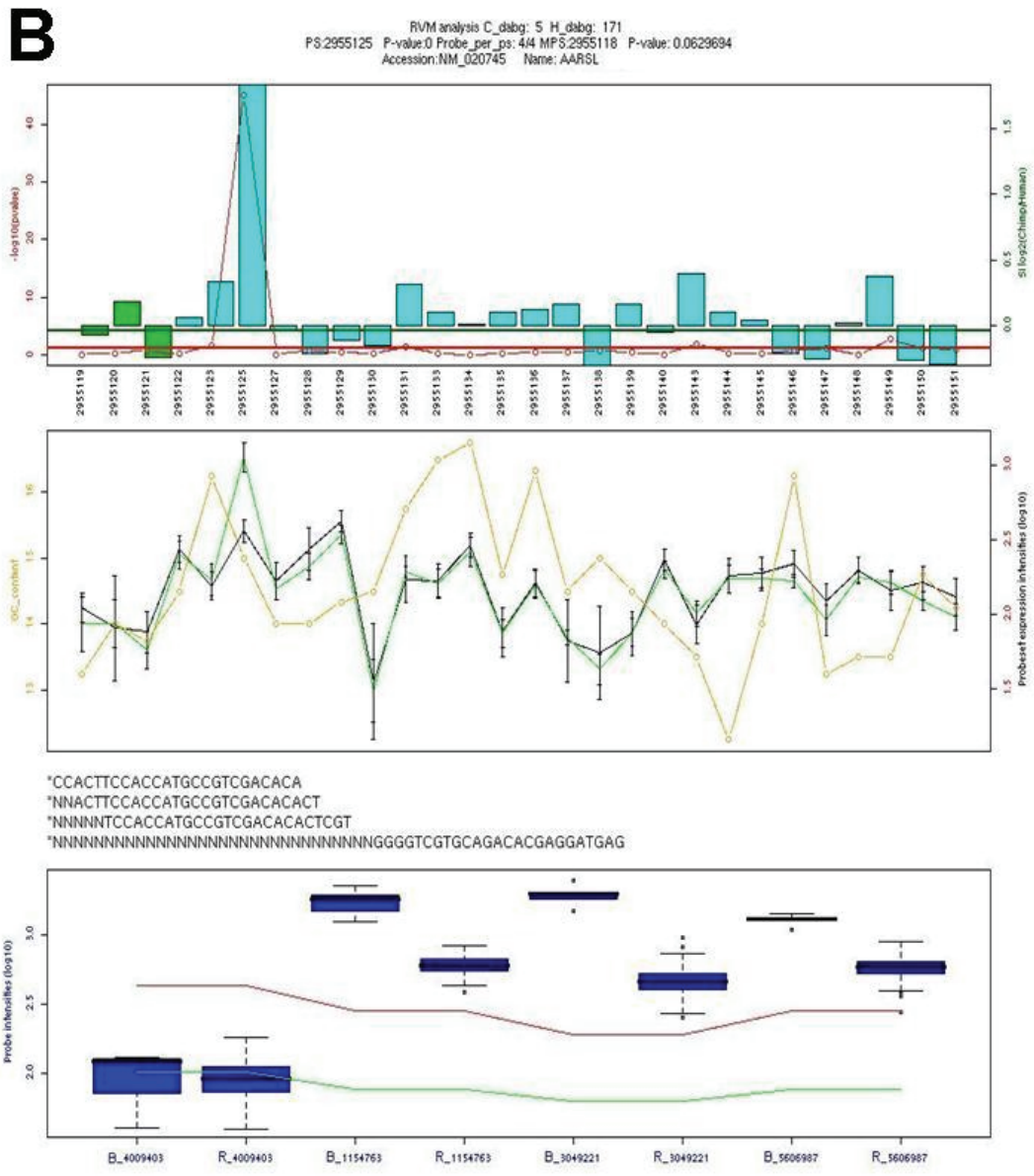
custom track feature of the UCSC genome browser (KAROLCHIK *et al.* 2008), in order to export our own information and combine it with publicly available data. In Additional file 1, we present other examples of discordance between the platforms, further illustrating the value of additional information present on the Exon Array in profiling both "whole transcript" and "isoform-level" changes.

Using the Exon Array to Profile Alternative Isoforms

One of the biggest challenges in profiling alternative isoforms using Exon Arrays is the deconvolution of mRNA processing and transcription. A simple comparison of probe set intensities across samples is not sufficient; if an exon belongs to a transcript that is differentially expressed, the examination of a single exon out of its genomic context will lead to an incorrect conclusion. A very simple and intuitive solution to this problem is the use of the Splicing Index (SI), that is calculated by dividing the probe set intensity by the meta-probe set intensity (i.e. exon expression/gene expression), after the addition of a stabilization constant to both the probe set and meta-probe set scores (www.affymetrix.com).This simple procedure normalizes the expression level of each exon and accounts for any possible gene expression differences between samples. However, we find that the splicing index has some undesirable statistical properties (arising from large errors in the estimates in both the numerator and the denominator) as well as being prone to methodological artifacts and should be used with caution. Thus, we have also used a simpler, but more labor intensive method, of carrying out the entire analysis at the probe set level, and relying on visualization and manual curation of the results in order to distinguish splicing and expression differences between samples. While more robust statistical approaches are being developed, we strongly advocate visualization of results in the context of genome annotation and EST evidence in order to filter out false positive signals. We have relied on

custom scripts and modifications of the UCSC and ENSEMBL genome browsers (Figure 6.4), but increasingly useful and user-friendly commercial packages for the Exon Arrays are available (e.g. Partek Genomics Suite, Biotique XRay) along with academic BioConductor packages (OKONIEWSKI and MILLER 2008; OKONIEWSKI *et al.* 2007b; PURDOM *et al.* 2008). Below, we describe in more detail two approaches to alternative isoform detection. For the case of simplicity, only the core (most confident) subset of Exon Array probe sets was considered in this analysis.




B
RVM analysis C_dabg: 5 H_dabg: 171
PS:2955125 P-value:0 Probe_per_ps: 4/4 MPS:2955118 P-value: 0.0629694
Accession:NM_020745 Name: AARSL
-log10(pvalue)
SI log2(Chimp/Human)
2955119
2955120
2955121
2955122
2955123
2955125
2955127
2955128
2955129
2955130
2955131
2955133
2955134
2955135
2955136
2955137
2955138
2955139
2955140
2955143
2955144
2955145
2955146
2955147
2955148
2955149
2955150
2955151
GC_content
Probeset expression intensities (log10)
*CCACTTCCACCATGCCGTCGACACA
*NNACTTCCACCATGCCGTCGACACACT
*NNNNNTCCACCATGCCGTCGACACACTCGT
*NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGGGTCGTGCAGACACGAGGATGAG
Probe intensities (log10)
B_4009403
R_4009403
B_1154763
R_1154763
B_3049221
R_3049221
B_5606987
R_5606987

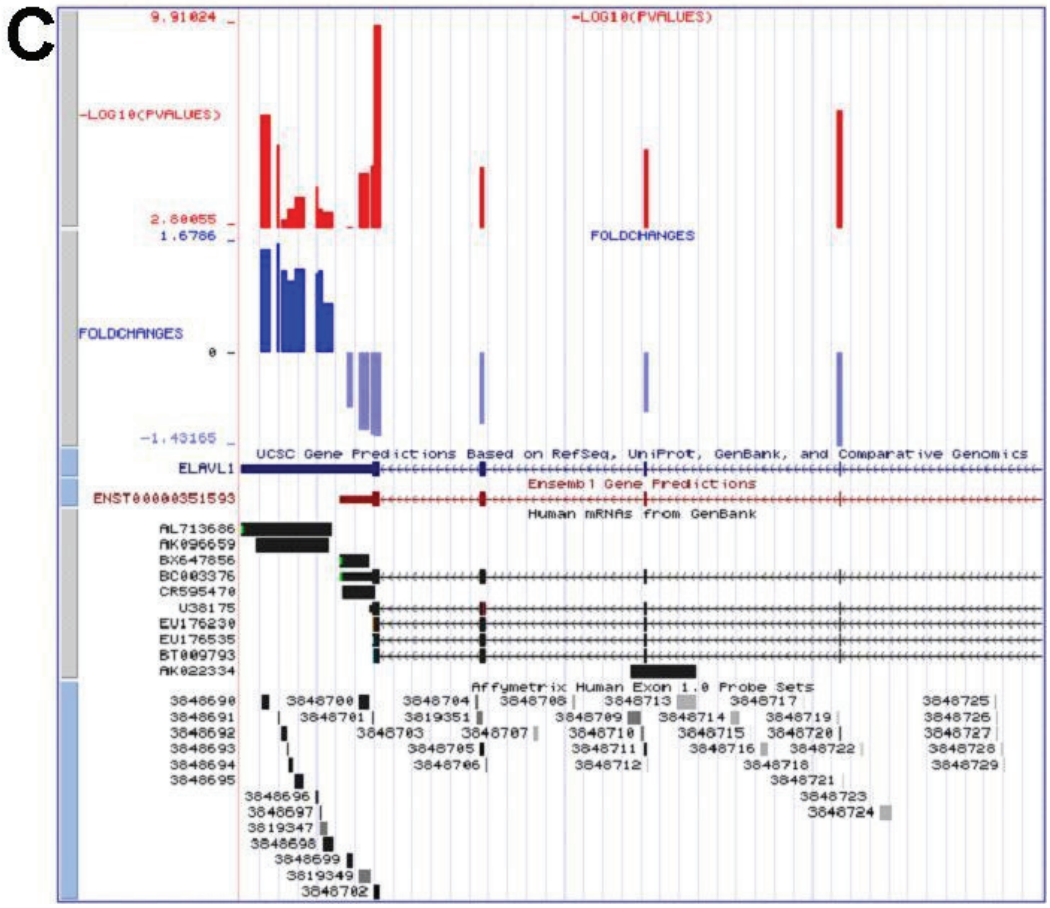


Figure 6.4: Evolution of the different methods developed to visualize expression data. A. First method was developed to visually compare expression data between 2 samples. In this example, each probe (columns) for 5 probe sets is represented as grey-scaled coloured pixels (square) where probe intensity scores increase as the colour whitens. The rows represent 15 technical replicates for each group (A and B). B. This visualization method was developed for assessment of potential transcript isoform variation events. The top panel represents each probe set in their gene context from the 5' to the 3' end (x-axis). The vertical bars represent the splicing index (right y-axis) for each probe set. Their colors represent their position within the gene, i.e. 5' UTR, coding region and 3' UTR. The red line represents the p-value (-$\log_{10}$ scale on the left y-axis) derived from a t-test conducted between the 2 groups using probe set intensities values. The horizontal green and red line represents the fold-change and

p-value, respectively, at the meta-probe set level computed between the two samples. In the middle panel the mean intensity and standard deviation for each probe set (log_2 scale, right y-axis) are represented for both samples. The orange line represents the average GC-content derived from each probe that make up a given probe set (for assessing cross hybridization potential). Bellow the middle panel is an alignment of probe for a probe set that shows significant differential expression between the two groups. In this example, 3 of the 4 probes target the same position indicating that they measure the same region of the exon. This makes the probe set less confident and more prone to be influenced by unknown SNPs (see chapter 7). The bottom panel is a box plot representation of probe intensity from each group under investigation. The horizontal green and red lines represent the mean and 2 standard deviations, respectively, for background expression intensity thresholds derived from the antigenomic probe expression distributions based on a specific GC-content. When boxplots are above these lines the region these probes target are considered expressed. C. Visualization of expression using custom tracks for the UCSC browser to determine the isoform variation event. The p-value and fold-change are represented as the red and blue horizontal bars, respectively. Note that the two probe set "blocks" correspond to the two isoforms of the gene. The long 3'UTR isoform is predominantly expressed in the brain, whereas the short isoform is more abundant in the reference tissues.

Probe set level analysis

At this level of the analysis, each probe set (roughly corresponding to an exon) is used as a unit of expression, instead of a meta-probe set (a transcript) as is done in more traditional gene expression analysis. With appropriate statistical significance cut-offs (e.g. a Benjamini-Hochberg (BENJAMINI *et al.* 2001) False Discovery Rate correction), it is generally

possible to select a highly confident set of probe sets exhibiting significantly altered expression. However, it is not immediately possible to classify the "hits" as results of alternative isoform expression (e.g. alternative splicing), differential gene expression, or both. The easiest way of factoring out of gene expression is to consider only the genes whose expression does not change across samples or treatments. That is, we can select probe sets that are statistically significant, but which belong to genes whose meta-probe set expression does not appear to be significantly altered (nominal $p > 0.05$). For the MAQC samples, we generated a list of the top 100 such genes. The list and links to the UCSC browser are provided in the Additional file 2. The top candidates show evidence for differential promoter usage, polyadenylation, and alternative splicing. A few examples appear to be annotation errors, where the Affymetrix annotation combines two distinct genes into a single transcript cluster. In general, we advocate RT-PCR based validation of alternative isoforms. However, cross validation with existing information is also extremely useful. Extensive EST and mRNA based information on tissue specific splicing is available from many sources, e.g. from the ASAPII (KIM *et al.* 2007) or Hollywood (HOLSTE *et al.* 2006). Most of the source data can be viewed directly in the UCSC genome browser by displaying the mRNA, spliced EST, or AltEvents tracks.

Dataset Reduction

In order to reduce the amount of random noise, and decrease the number of tests being carried out, it is useful to exclude all genes which are either not expressed in all of the samples, or more than one of the samples being compared. Such genes, by definition cannot be alternatively spliced across samples. There is currently no reliable procedure on deciding whether a gene is expressed or not, and Affymetrix recommends using an

ad hoc expression value of 15, and some additional filters using DABG values of individual exons.

Effect of "Dead" Probe sets

A probe set which is not expressed – e.g. an exon which is skipped – in all samples under investigation may produce a false positive signal in the splicing index, in the presence of transcript-level variation. All non-responsive probe sets should be removed from the analysis. A DABG-based criterion may be used here, e.g. DABG p-value < 0.05 in at least 50% of the samples.

Discussion

The recognition of alternative splicing and alternative isoform expression as an important component in gene expression analysis has prompted the introduction of isoform sensitive microarray platforms. By targeting individual exons, exon junctions, and annotated isoform variants, such platforms possess the ability to profile not only the expression levels of the entire transcript, but also variations in the types of expressed isoforms. The Affymetrix Exon Array 1.0 ST is one of such commercially available platforms. To date, it has been shown that the Exon Array produces gene expression measurements that are comparable with the previous generation 3' targeted arrays. However, little is known about the in-depth level of similarities and particularly differences among WT and 3' based technologies. This comparison utilizes the well studied brain and reference samples previously used in the MAQC study to determine sources of variability in profiling gene expression using microarrays. These samples are particularly valuable for the purposes of benchmarking the performance of the Exon Array for two reasons: 1) they allow easy comparison of gene expression level measurements with other platforms that have already been tested, and 2) they allow detection of alternative

splicing and isoform difference, since neural tissues are known to be particularly prone to alternative splicing.

Our first conclusions concern the utility of the Exon Array as an expression profiling tool. We note that although the Exon Array results are very consistent with 3' profiling methods, the level of agreement between the Exon Array and 3' targeted platforms (Illumina and Affymetrix U133) is slightly lower than the agreement between the 3' platforms. Many of the outliers in the correlation plot (Figure 6.3) are due to the presence of real variations in the expression of specific isoforms. This is illustrated using a previously noted example of the ELAVL1 gene, which showed discordance across platforms in the original MAQC study, as well as in additional new examples (Additional file 1). The detected expression differences of transcript variants may have important biological significance. For example the longer 3' UTR in the dominant ELAVL1 transcript in brain has a different set of putative micro RNA binding sites than the shorter 3' UTR in the reference RNA. It should also be noted that discordant results will often be obtained because of differences in the annotation provided by microarray manufacturers. We circumvented most of such problems here by re-mapping the probes and selecting only a subset of genes that we were confident were correctly targeted by all three platforms, but researchers should keep in mind that the annotations and gene assignments provided by manufacturers contain numerous errors (DAI *et al.* 2005). In the case of the Exon Array, we found that the most common annotation error resulted from joining together distinct transcripts into single meta-probe sets, particularly in the case of transcripts that partially overlap. Thus, we recommend that lists of candidates from individual experiments should be carefully curated.

We also outline how the Exon Array can be used to detect alternative splicing and alternative mRNA processing events. Although our analysis methods are not in themselves novel, and most of them have been briefly described elsewhere (KWAN *et al.* 2007), our goal is to convey to the potential users their intuitive appeal and potential pitfalls. The most challenging step remains the decoupling of whole transcript expression, and individual probe set inclusion. The simplest solution to this problem is to consider only the genes that do not change overall expression levels, but contain probe sets that exhibit individual variations. Although this approach produces a highly confident set of alternative events, it can result in a huge reduction of the dataset, particularly in case of comparisons across samples with highly heterogeneous gene expression levels. In the case of MAQC dataset, which has been chosen for the exact reason of it's extreme gene expression variability, imposing the restriction of expression fold change of less than 2 reduces the total number of genes considered by 31% (from 17665 to 12198). A more inclusive approach is to attempt to correct for gene expression differences that may occur concurrently to splicing differences. We discuss two such approaches: 1) the splicing index, which compares probe set inclusion across samples after normalizing by gene expression levels, and 2) two-way ANOVA, where the interaction term between sample type and probe set can be used to indicate differential inclusion of probe sets within transcripts. Both approaches suffer from similar systematic biases; they assume a uniform (linear or log-linear) response of each probe set within a meta-probe set. This assumption is violated in many cases, particularly for probe sets that hybridize at very high levels (saturated response) or probe sets with hybridization levels close to background (poorly or non-responsive). As a result, in the presence of significant gene expression changes, such analyses predominantly indicate three types of events: dead probe sets, saturated probe sets, and probe sets that may be

predominantly skipped (alternative), but not necessarily differentially included across samples.

Many of the above systematic errors can be avoided by filtering out potentially troublesome subsets of the data: probe sets with extremely low variability (saturated), probe set with low inclusion levels (close to background), and genes with extremely high differences in expression levels across samples. However, such filtering decreases the false positive rates at the cost of reduced genomic coverage. In our earlier studies, we have also pointed out that in many experimental designs, particularly when samples originate from different genetic backgrounds (e.g. different individuals), the presence of sequence variants within probe target sequences may be a very significant source of errors (KWAN *et al.* 2008; KWAN *et al.* 2007). This effect can be especially prominent in eQTL association studies, where we have shown that it can be responsible for a false positive rate >80% in alternative splicing analysis (BENOVOY *et al.* 2008). Thus, unless all tested samples are isogenic, we highly recommend additionally "masking".

Conclusion

In summary, the WT profiling provides a wealth of valuable information, which is either not available or misrepresented in traditional 3' gene expression arrays. However, it should be noted that the isoform-level analysis of Exon Arrays is significantly more complicated, suffers from higher false positive rates, and requires more manual intervention than traditional gene expression analysis. We strongly advocate visualization of candidate isoform changes in the context of available genome annotation as a means to both reduce false positive rates and interpret the nature of detected variants.

Acknowledgements

This work was supported by grants from CIHR and Genome Canada/ Genome Quebec (to JM). JM is a Canada Research Chair recipient. DJG is supported by a FRSQ post-doctoral fellowship. We would like to thank the staff of the Genome Quebec Microarray Platform as well as Clive Evans and the technical staff of the Core Laboratory Facility at the Virginia Bioinformatics Institute for expert help and microarray processing.

Chapter 7: Effect of polymorphisms within probe–target sequences on olignonucleotide microarray experiments

David Benovoy, Tony Kwan and Jacek Majewski

Published in *Nucleic Acids Research* on July 2nd, 2008. 36(13):4417-4423

Connecting text

In the last chapter, we evaluated how the Exon arrays behaves when detecting differences at the whole-transcript expression by comparing it to traditional 3' array and found good concordance between these platforms. We also investigated the general sources of noise encountered in experiments using the Affymetrix Human Exon array and discussed ways to reduce the false positive rate. In this chapter, we investigate the main source of false positives when conducting an eQTL experiment using the Affymetrix Human Exon Array. We show that polymorphisms in probe targets are responsible for >80% of the false positives when conducting the analysis at the isoform level. We propose a simple solution to this problem that entails removing probes that target polymorphic regions. This greatly reduces the false positive rate without a significant reduction in exon coverage.

Abstract

Hybridization-based technologies, such as microarrays, rely on precise probe-target interactions to ensure specific and accurate measurement of RNA expression. Polymorphisms present in the probe–target sequences have been shown to alter probe- hybridization affinities, leading to reduced signal intensity measurements and resulting in false-positive results. Here, we characterize this effect on exon and gene expression estimates derived from the Affymetrix Exon Array. We conducted an association analysis between expression levels of probes, exons and transcripts and the genotypes of neighbouring SNPs in 57 CEU HapMap individuals. We quantified the dependence of the effect of genotype on signal intensity with respect to the number of polymorphisms within target sequences, number of affected probes and position of the polymorphism within each probe. The effect of SNPs is quite severe and leads to considerable false-positive rates, particularly when the analysis is performed at the exon level and aimed at detecting alternative splicing events. Finally, we propose simple solutions, based on 'masking' probes, which are putatively affected by polymorphisms and show that such strategy results in a large decrease in false-positive rates, with a very modest reduction in coverage of the transcriptome.

Introduction

Microarray analysis has become an integral part of high-throughput biological research. Microarray-based measurements typically rely on the precise hybridization of a DNA probe to a complementary target DNA or RNA molecule. Advances in technology and miniaturization now allow manufacturers to print up to 10 million probes on a single chip. Such chips are routinely used for truly genome-wide studies of polymorphisms, genomic aberrations (KOMURA *et al.* 2006), gene expression levels (STRANGER *et al.* 2007a) and alternative splicing patterns (KWAN *et al.*

2008). Unfortunately, such massive amounts of data come at the expense of a high potential for false discovery. Common sources of error range from the purely statistical (e.g. multiple testing problems), through experimental techniques, to systematic technical errors (e.g. probe cross-hybridization). As a result, particularly in gene expression analysis, microarray results have often been relegated from the realm of 'proof' to the role of a 'discovery platform' for further validation. In view of their overall popularity and utility, it is of great importance to minimize systematic errors in microarray experiments. In this study, we focus on one particular source of error: the effect of polymorphisms contained within probe target sequences on hybridization levels. Using expression quantitative trait analysis (eQTA) as an example, we show that this effect can be a major source of error, particularly for the latest generation whole-transcript (WT) arrays.

Association of genetic variants to expression phenotypes is becoming a promising strategy to identify sources of phenotypic diversity among individuals. A large number of genome-wide studies have been conducted in recent years, using various microarray platforms to determine gene expression levels (CHEUNG *et al.* 2005; DEUTSCH *et al.* 2005; DIXON *et al.* 2007; EMILSSON *et al.* 2008; GORING *et al.* 2007; MORLEY *et al.* 2004; STRANGER *et al.* 2005; STRANGER *et al.* 2007b). This approach usually treats expression data obtained from microarray experiments as a quantitative trait and tests for association with *cis*-acting polymorphisms. The final goal is to identify regulatory determinants of a particular phenotype, such as a disease state. Once significant associations have been identified, costly and time consuming downstream validations are conducted in order to identify the causative regulatory element. Therefore, it is important to identify candidate *cis*-acting polymorphisms with a high degree of confidence. Recent studies have shown that mismatches

between a microarray probe and its target sequence affect hybridization (SLIWERSKA *et al.* 2007; VALLEE *et al.* 2006; ZHANG *et al.* 2007) that cause erroneous probe signal estimates. This phenomenon leads to an increase in false-positives, particularly in studies across individuals with different genetic backgrounds (WALTER *et al.* 2007). Individuals expressing mRNA that perfectly complements the probes on the microarrays hybridize better than individuals with mRNA sequence diversity in the probe–target region. This results in a difference in probe–signal intensity between individuals, even if both groups express the mRNA at the same level (ALBERTS *et al.* 2007).

Here, we present a detailed analysis of this phenomenon using Affymetrix Human Exon Array data from our previous study of transcript isoform variation in humans (KWAN *et al.* 2008) and describe how it affects association results at the probe, exon and gene levels. In addition, to mitigate the effect of polymorphisms, we propose a simple strategy that consists of removing probes that are targeted to annotated polymorphic regions. We show that this approach greatly reduces false-positive rates, particularly for associations at the exon level, with only a small reduction in exon and gene coverage.

Methods

Microarray data source

In a previous study, we surveyed genetic variation associated with differences in isoform level expression in humans (KWAN *et al.* 2008). We characterized this effect in a sample of 57 unrelated HapMap individuals of European ancestry (ALTSHULER *et al.* 2005) for which ~4 million single nucleotide polymorphism (SNP) genotypes are available. Lymphoblast cells derived from these individuals were grown in triplicates and RNA was extracted from each of these growths and hybridized onto an Affymetrix

Human Exon array (n = 171). The resulting probe-fluorescent intensities were used for the present analysis. We restricted our analysis to probes targeting core exons because of their high confidence annotation.

Effect of mismatches on hybridization

Probe expression signals were quantile-normalized and GC-background corrected using the Affymetrix Power Tools (APT) software package (Affymetrix). To investigate how mismatches affect probe-to-target hybridization on the Affymetrix Human Exon array, we took advantage of the high-resolution genotyping information available from HapMap cell lines and identified 6110 probes that were targeted to a region with only one SNP in at least 1 of the 57 HapMap individuals. These probes were selected because the exon and gene they targeted were considered expressed. Expression of an exon or gene was established using the detected above background (DABG) metric generated by Affymetrix. This metric represents the probability that an exon or gene is expressed below the background. We used false discovery rate (FDR) correction (BENJAMINI *et al.* 2001) to establish the significance threshold for expression above background at DABG ≤0.02 and DABG ≤0.043 for exons and genes, respectively. Next, we categorized each of these probes in 25 bins, depending on the position of the SNP within the target region (from 5' to 3' end). For each of these bins, we determined the fold change between the average probe intensity derived from individuals with a perfect complementary target region and the average probe intensity from individuals with one mismatch (Figure 7.1).

Masking procedure

We have previously shown (KWAN *et al.* 2008; KWAN *et al.* 2007) that SNPs located within probe-targets affect their hybridization to Affymetrix Human Exon array probes and consequently cause erroneous expression

estimates. To mitigate this effect, we devised a simple procedure that consists of removing all probes from the analysis whose target region contains a known SNP. In total, we found 21 843 core probes target regions out of 1 096 799 probes overlapping at least one polymorphic HapMap II SNP (release 21).

Preprocessing and summarization of hybridization data

To study how probe-to-target hybridization is affected by SNPs, we generated two data sets of exon and gene expression estimates. The APT software package was used to quantile-normalize and GC-background correct each data set at the probe level. The average probe set (representing exons) and meta-probe set (representing genes) expression scores (averaged from triplicates) for each data set were computed using the probe logarithmic error intensity model (Affymetrix). The first data set consisted of probe set and meta-probe set expression estimates produced by summarizing all core probes, regardless of polymorphic probe target regions. The second data set was generated by implementing our masking procedure (see above). Thus, probe set and meta-probe set expression scores, for this last data set, were estimated from probes where no HapMap SNP overlapped their target region.

Association analyses

For each of the two data sets, the first generated from the full core probe list and the second from the masked core probe list, we examined probe, exon, and transcript expression estimates (averaged from triplicate samples for each individual) for association with flanking HapMap SNPs (release 21). One of the objectives of our previous analysis (KWAN *et al.* 2008) was to identify possible *cis*-regulatory determinants of differential alternative splicing. The presence of linkage disequilibrium in humans has created haplotype blocks, where SNPs in close proximity to each other

escape rearrangements due to recombination. Therefore, assuming physical proximity of a regulatory variant to the target and to limit the cost of multiple testing, we only tested for SNPs within a 50-kb region flanking either side of the gene containing either the probe or probe set. It should be noted that the SNPs associated with a change in microarray hybridization intensity may either be the actual causative SNPs, or simply be in linkage disequilibrium (part of the same haplotype block) with the causative SNP. We measured the level of association between expression scores (probes, probe sets and meta-probe sets) and the genotypes of a given SNP using linear regression analysis, implemented in the Plink software package (PURCELL *et al.* 2007), under a codominant genetic model. This model considers genotypes AA, AB and BB as the independent discrete variable. The genotypes are encoded as 0, 1 and 2, respectively, whereas expression scores were considered a quantitative trait and treated as the dependent variable in the linear regression. Raw *P*-values were obtained from the linear regression using the standard asymptotic *t*-statistic. To correct for testing multiple SNPs against each probe set and meta-probe set expression values, we carried out permutation tests (CHURCHILL and DOERGE 1994) followed by 5% FDR correction. Permutation analyses were performed using the 'label swapping' and 'adaptive permutation' options implemented in Plink. The 'label swapping' option is used to preserve the haplotype block structure and the 'adaptive permutation' algorithm allows for computationally efficient permutation analyses (PURCELL *et al.* 2007). Subsequently, we performed FDR corrections of 5% on the empirical *P*-values (from permutations) for association of genotype to the expression at the probe set (*P*-value $<9.73 \times 10^{-9}$) and meta-probe set levels (*P*-value $<6.07 \times 10^{-7}$).

Evaluation of SNP mask

To evaluate how SNPs in probe–target regions impacted our association analyses, we estimated the proportion of false-positive and false-negative associations due to polymorphic probe target regions. We treated the association results for the masked data set as the reference (true) data set because they were derived from expression estimates free of influence from known SNPs. This reference data set enables us to evaluate the four scenarios described in Table 7.1.

Table 7.1: Comparison of association analyses with and without a SNP mask

	SNP Mask	
	Positive for Association	Negative for Association
No Mask		
Positive for Association	True Positive	False Positive
Negative for Association	False Negative	True Negative

Associations of probe set or meta-probe set, which were significant (*P*-value below the thresholds) and non-significant (*P*-value above thresholds) in both masked and unmasked data sets, were classified as true positives and true negatives, respectively. We consider a result a false-positive when a significant association is found in the unmasked data set, but become non-significant after masking probes containing SNPs (masked data set). Conversely, associations that were non-significant in the unmasked data set but significant in the masked data set were categorized

as false-negatives. The false-positive and -negative rates are computed by: FPR = FP/ (FP + TP) and FNR = FN/ (FN + TN), respectively. In order to avoid the problem of reduced coverage within the masked data, the above analysis does not include probe sets which were entirely 'masked' due to the presence of SNPs.

Results

Our first objective was to examine the effect of sequence mismatches on probe-to-target hybridization. We selected all probes that contained known SNPs and compared their hybridization intensity between individuals with homozygous match and mismatch genotypes. We illustrated how hybridization intensity changed when a mismatch is present at a given position within a probe in Figure 7.1. We observed that the position of the polymorphism within the probe's target sequence affects its binding affinity. Probe expression scores show a median ~2-fold decrease in expression when a polymorphism is present near the middle of the target area i.e. between positions 6 and 21. This effect decreases linearly towards the edges of the target area and the median fold change in the end is near zero i.e. at positions 1 and 25, which supports the theoretical prediction of Lee *et al.* (LEE *et al.* 2004). It should be noted that the variance in the estimate of the effect is very high and that some mismatches decrease hybridization levels by much more than 2-fold; 7.5% of mismatches cause ≥5-fold decrease in signal intensity. Thus, in some cases the effect of SNPs may be very severe. This corroborates suggestions by earlier studies (HUGHES *et al.* 2001; SLIWERSKA *et al.* 2007; VALLEE *et al.* 2006; WALTER *et al.* 2007; ZHANG *et al.* 2007) that mRNA sequence diversity in probe target regions disrupts hybridization and that polymorphisms in the middle of the probe target regions destabilize hybridization more than those closer to the ends.

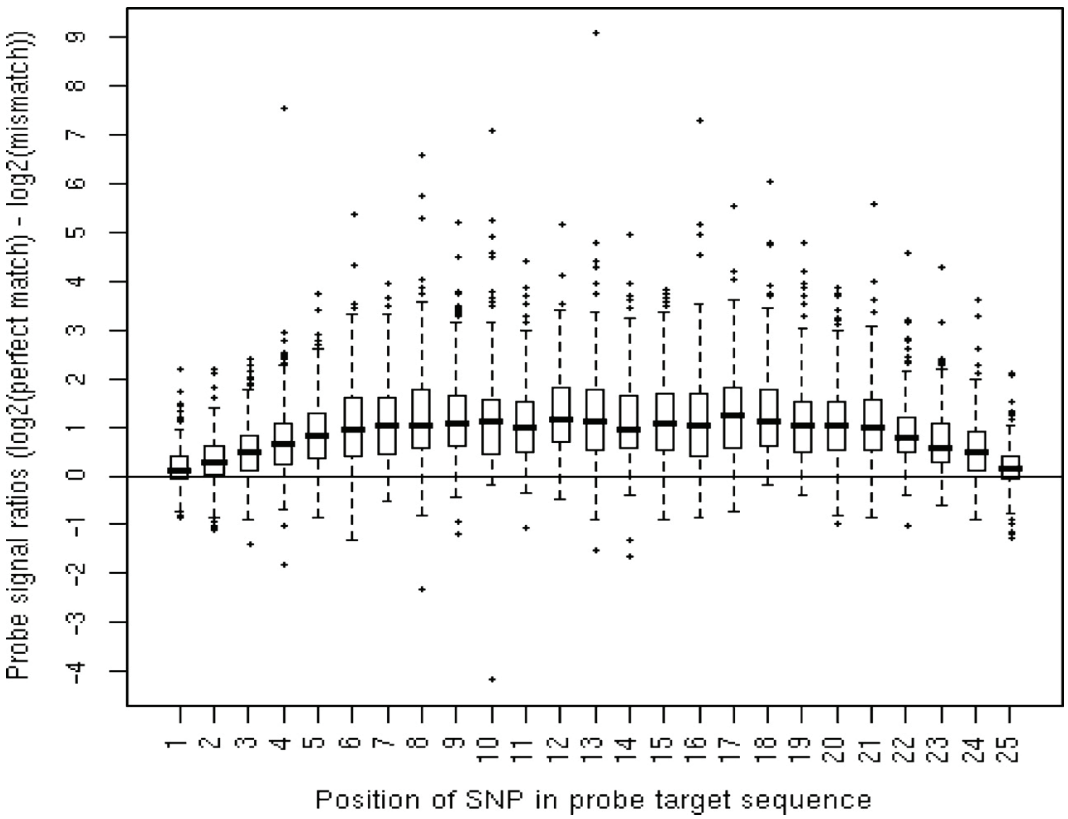


Figure 7.1: Boxplots illustrating the positional effect of SNPs within the probe target region. Probe signal ratios between perfect complementary regions and regions with a single mismatch.

We next investigated how the association of expression phenotypes to neighbouring SNPs, as in our previous analysis (KWAN *et al.* 2008), are distorted by including probes whose target regions were polymorphic. We characterized this by performing an association analysis between expression levels of probes, exons and transcripts, with the genotypes of neighbouring HapMap II SNPs. We compared only the top 1% of significant associations as a way to uniformly correct for multiple testing between the different levels of expression (probe, exon and gene). We observed that probes with polymorphic target regions were highly over-represented in the top 1% of significant association by a factor (odds ratio) of 16.8 (Table 7.2; χ^2 = 33976.74, *P*-value << 10^{-16}) compared to probes with perfectly complementary probe target regions. We also observed this

over-representation at the probe set and meta-probe set levels, although to a lesser degree. In the top 1% of significant associations, we found an enrichment of 6.1-fold (Table 7.2; χ^2 = 1443.88, *P*-value << 10^{-16}) and 2.5-fold (Table 7.2; χ^2 = 19.45, *P*-value = 1.03 x 10^{-5}) for probe sets and meta-probe sets, respectively, whose expression estimates included probes that were targeted to polymorphic regions. In addition, this enrichment is also positively correlated with the number of polymorphisms within probe target region at the probe set (Pearson *r* = 0.956) and meta-probe set (Pearson *r* = 0.967) levels (Table 7.2). This further demonstrated that sequence polymorphisms between an Affymetrix Human Exon array probe and its target sequence resulted in changes to hybridization intensity and influenced the apparent association between the SNP genotypes and expression intensities. Given that probe set and meta-probe set expression estimates are derived by summarizing probe signals, erroneous probe signals due to probe target mismatches are a source of error in comparative expression analyses.

Table 7.2: Enrichment for probes with polymorphic target region in the top 1% of significant association for probes, probe sets and meta-probe sets

Number of SNP overlaps	Enrichment (odds ratio)		
	Probe	Probe set	Meta-probe set
All	16.83	4.30	2.46
1	16.78	1.98	1.94
2	19.39	5.02	2.12
3	NA	10.89	2.40
4	NA	15.64	3.00
≥5	NA	14.84	3.01

To reduce this source of error, we developed a simple masking procedure where we removed all probes targeted to a known polymorphic region (HapMap phase II SNPs). The remaining probes were used to estimate probe set and meta-probe set expression scores. A detailed example of this procedure and how it reduces the false-positive association caused by polymorphic probe target regions for gene *ZNF37A* is illustrated in Figure 7.2. Expression estimates for this gene were derived from four probe sets (Figure7.2a), one of which, probe set 3243183, comprised probes targeting a polymorphic region in the 57 HapMap individuals. The first 3 probes from this probe set (Figure 7.2b) overlapped each other to some degree and targeted a region containing SNP rs176889. Individuals with TT genotypes have higher probe signals than individuals with a TC or CC genotype because the T allele creates a perfectly complementary target to these 3 probes (Figure 7.2c). The fourth probe, probe 496020, targets a region with no known SNP and shows no significant associations with

SNPs rs176889. In addition, we do not find any significant association with neighbouring SNPs that could be in linkage disequilibrium with SNP rs176889. Therefore, by using this single probe to estimate the expression of probe set 3243183, we obtain expression estimates that are not affected by erroneous probe signals and in subsequent association analyses (Figure 7.2d), the same is observed at the gene level (Figure 7.2e). We only used probe set expressions derived from probes unaffected by SNP to estimate meta-probe set expression scores and find no significant association with neighbouring SNPs.

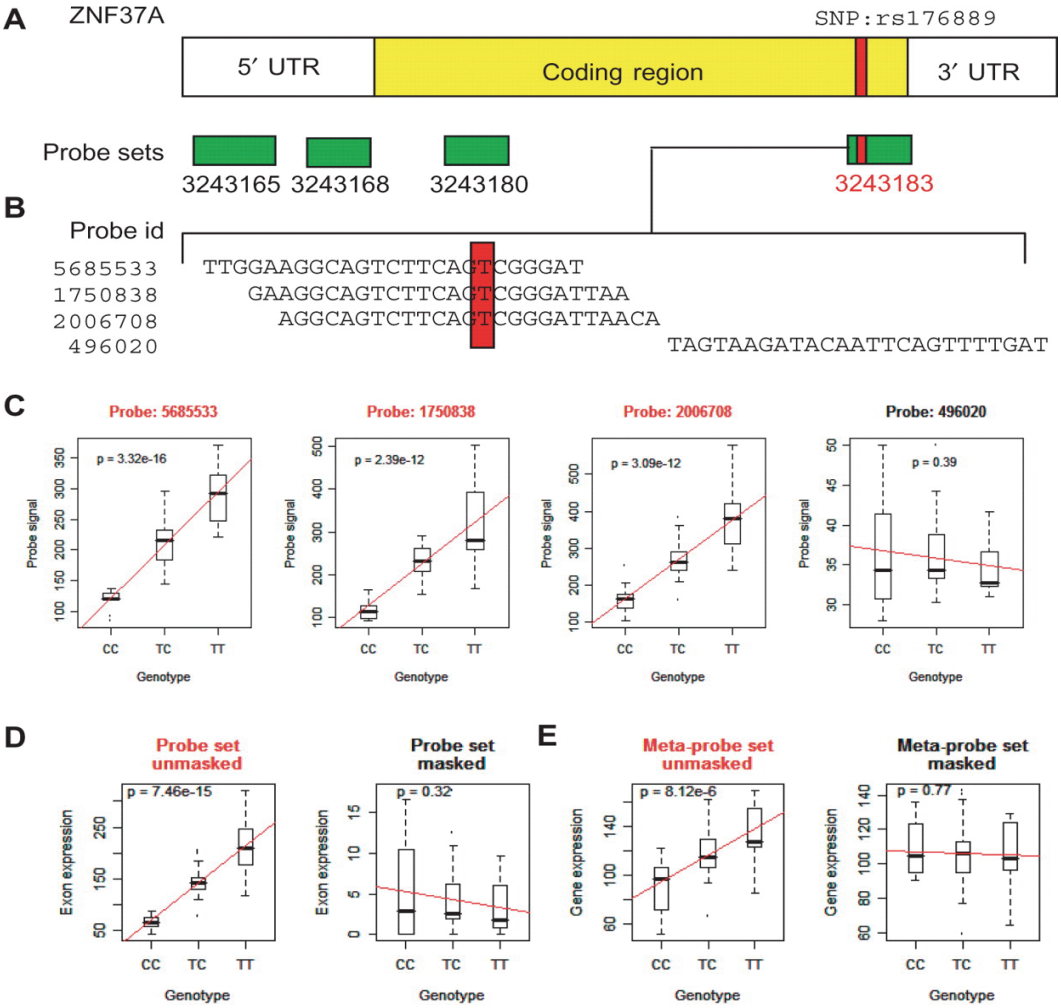


Figure 7.2: ZNF37A is an example of a false-positive induced by a SNP (rs176889). (**A**) The ZNF37A mRNA molecule is illustrated with the coding region in yellow and the 5' and 3' UTRs is represented in white. The

horizontal green rectangles represent the 4 probe sets that target this transcript. The red bars represent the position of SNP rs176889 in the coding sequence of this transcript. (**B**) The alignment of the 4 probe sequences that constitute probe set 3 243 183 and SNP rs176889 falls within each of these probes (red box). (**C**) Plots illustrating the association between each of the 4 probes and the different genotypes for SNP rs176889. Probe 496 020 does not contain any SNP and the association is non-significant. It is the only probe used to estimate probe set 3 243 183 expression scores. (**D**) Probe set 3 243 183 is no longer a false-positive after our masking procedure. (**E**) The same is observed at the meta-probe set level, where this gene is not significantly associated with SNP rs176889 or any other neighbouring SNPs (results not shown).

A potential drawback associated with removing problematic probes, is the reduction of probe set and meta-probe set coverage. For this data set, 21 843 (1.99%) probe target sequences overlapped at least one HapMap SNP and the distribution of affected probes per probe set and meta-probe set is illustrated in Figure 7.3a and b, respectively. We found 1258 (0.47%) probe sets and 99 (0.57%) meta-probe sets where we could not derive any expression estimates because no probes were left after 'masking' which is a very modest amount of lost coverage.

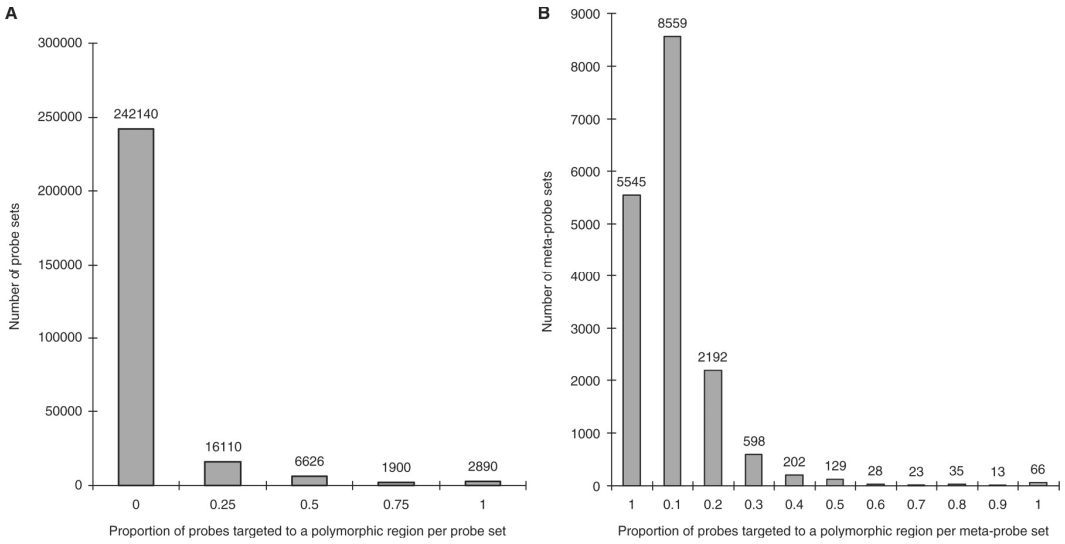


Figure 7.3: Distribution of probe sets and meta-probe sets containing SNPs. (**A**) Proportion of affected probes per exon (**B**) Proportion of probes that contain SNPs per transcript.

Next, we assessed how our masking procedure improved results obtained from our association analyses. For the purpose of the analysis, we assumed that an association is a false-positive when a probe set or meta-probe set is significant in the unmasked data set, and that the same association becomes non-significant after masking probes containing SNPs. This assumption is based on two sources of evidence: (i) the strong over-representation of SNPs in the significant data set and (ii) the fact that in our previous work (KWAN *et al.* 2008; KWAN *et al.* 2007) we were unable to experimentally validate an alternative splicing event supported by an SNP-containing probe. We assumed that the expression data set derived by 'masking' misbehaving probes represents the best estimates of probe set and meta-probe set expression scores. Using this as the reference (true) data set, we evaluated the four scenarios described in Table 7.1 by comparing the *P*-values obtained from the association of the same neighbouring SNPs to the same probe sets or meta-probe sets expression score estimated without 'masking' problematic probes. It should be noted that the reference set itself may not be free of false-positives (due to

sources of errors other than SNPs), but this approach allows us to determine the rates of false-positive results that are induced by the presence of SNPs. We established *P*-value significance thresholds of 9.73 x 10^{-9} and 6.07 x 10^{-7} for probe sets and meta-probe sets, respectively, by permutation testing followed by FDR correction at 5%. We found that the SNP-induced false-positive rate is 86.6 and 8.1% at the probe set and meta-probe set levels, respectively Table 7.3. However, false-negative rates do not seem to be influenced by SNPs because, after masking these potentially misbehaving probes, the false-negative rates were reduced by only 0.3 and 0.05% at the probe set and meta-probe sets Table 7.3, respectively. This demonstrates that the removal of probe signals impacted by SNPs greatly reduces the rate of false-positives particularly for association conducted at the probe set level (e.g. alternative splicing). We concluded that masking probes targeted to known polymorphic regions does not substantially decrease the coverage of the Human Exon array and effectively reduces the SNP-induced false-positives.

Table 7.3: Effect of the masking procedure on results from the association analysis of probe sets and meta-probe sets

	Probe set	Meta-probe set
False positives	446	9
False negatives	41	4
True positives	69	102
True negatives	13,359	8,115
False positive rate	0.866	0.081
False negative rate	0.003	0.0005

Discussion

Our analysis suggests that the presence of SNPs within the target sequence of Affymetrix Human Exon array probes causes false-positives when the analysis is conducted at the exon and transcript levels. Exon expression estimates are affected by misbehaving probes at a higher degree then transcript expression estimates because they are summarized from only 4 probe signals, whereas transcript expression estimates rely, on average, on 30 probes. In addition, we demonstrate that 'masking' a probe targeted to a known polymorphic region is a simple and effective solution for decreasing the rate of false-positives in an association analysis with individuals of different genetic backgrounds.

Alternative filtering approaches have been suggested. Zhang *et al.* (ZHANG *et al.* 2008) proposed to remove from the analysis probe sets with 2 or more probes harboring dbSNPs (release 126). This would result in the removal of 1.96% of probe sets—a much more significant reduction than the 0.47% in the approach outlined here. In addition, we do not advocate leaving probe sets containing single SNPs in the analysis, as we show in Table 7.2, that such probe sets are still ~2-fold over-represented in the significant data set and are likely to produce false-positive results.

Our analysis takes advantage of the HapMap dataset, which has been genotyped at a high resolution. This constitutes an ideal data set for the purpose of illustration and quantification of the effect of SNPs. However, the results and solutions are applicable to most studies, whenever individuals with diverse genetic backgrounds are being compared. This is typically done in cancer studies and should be taken into consideration, particularly since investigation of alternative splicing and the use of WT arrays are quickly gaining popularity in this field (GARDINA *et al.* 2006; THORSEN *et al.* 2008). Generally, when two large groups of patients and

controls are being compared, the effect of SNPs should be minimal in the pooled comparison. However, whenever a single individual or a group of related individuals is being used in a comparison to control samples, the effect of SNPs will be substantial. Similar problems will be encountered in any comparison of alternative splicing across tissues, whenever the tissues do not originate from the same individual. In all such cases, we advocate conservatively masking all probes containing putative SNP sites (from dbSNP). In addition, in our previous study (KWAN *et al.* 2008) we found a non-trivial effect of still unannotated SNPs. While this problem cannot be corrected for *a priori*, we advise investigators to carefully monitor the behavior of individual probes before undertaking further costly functional studies—a single significant outlier probe whose behavior is inconsistent with the rest of the probe set may be an indication of a technical problem.

Finally, while we focus our study on the exon array and the analysis of alternative splicing, we would like to point out that other platforms are not immune to this effect. Examples of similar problems have been identified for the Affymetrix 3' expression arrays (ALBERTS *et al.* 2007; WALTER *et al.* 2007). Other popular expression platforms, such as Agilent and Illumina, use longer probes, which are less sensitive to SNPs, but a slight effect of polymorphisms can be detected in those platforms as well (DOSS *et al.* 2005; STRANGER *et al.* 2005). Therefore, we advocate preventive measures (such as SNP masking) and vigilance (careful scrutiny of final results), and propose that the next generation of microarray designs avoid, when possible, targeting polymorphic sites.

Acknowledgements

The authors wish to thank D. Serre, D. Gaffney and E. Harmsen. This work is supported by Genome Canada, Genome Quebec and the

Canadian Institutes of Health Research (CIHR). J.M. is a recipient of a Canada Research Chair. Funding to pay the Open Access publication charges for this article was provided by CIHR.

Chapter 8: Summary and Conclusion

Biology

The regulation of gene expression is recognized as an important mechanism in numerous biological processes. The study of how processes, such as alternative transcript initiation and termination, alternative splicing and transcript expression, are regulated will provide new insights into organism complexity and diversity. Recent studies have already demonstrated that variation of transcript expression is common among higher eukaryotes and that these types of variation have a genetic basis (CHEUNG *et al.* 2003; CHEUNG *et al.* 2005; STRANGER *et al.* 2005; STRANGER *et al.* 2007b). It is believed that certain regulatory changes that affect transcript expression are responsible for downstream phenotypic differences observed between and within species such as species specific traits and susceptibility to genetic diseases, respectively. This thesis demonstrates that, in addition, to transcript expression differences, a significant amount of expression variation is observed at the transcript isoform level. Moreover, it confirms that this variation also has a strong genetic component, and hence, the effect of common genetic variation in a human population and between humans and chimpanzees is much more complex than previously believed. Single nucleotide polymorphisms affect processes that generate transcript isoforms to an extent that is comparable or even superior to overall transcript expression. Therefore, the downstream phenotypic effects of these variations are likely as important as the ones generated by whole-transcript expression differences. For example, the genome-wide association analysis, described in chapter 4, identified a mutation in the polyadenylation sequence of the *IRF5* gene that is responsible for generating a 3' UTR variant that in turn, is associated with an increased susceptibility to lupus (GRAHAM *et al.* 2007). This example illustrates the type of information needed to better understand the phenotypic effect of isoform variants.

For most of the isoform variations between individuals, identifying the cause is a difficult but essential task to better understand the evolutionary processes that are responsible for generating phenotypic diversity as well as increasing our knowledge of disease mechanisms that in turn can help us develop novel therapeutic applications. Identifying the true causative genetic variant from an association analysis is still a challenge because the majority of polymorphisms are embedded within a haplotype block and consequently in linkage disequilibrium with many other good candidate polymorphisms. This task is even more difficult when conducting inter-species comparisons as the one described in chapter 5 because sequence regulatory elements are poorly defined and therefore any substitution could potentially be the cause of the expression difference. To identify the exact genetic difference responsible for an observed expression variation will require assay systems capable of confirming and further dissecting how genetic differences cause expression variations. Methods such as site-directed mutagenesis or *in-silico* approaches could be used for this purpose which would help the scientific community identify the elements responsible for regulating gene expression and to better understand the processes involved. Another important aspect to consider is how mRNA variations translate to the proteome. It is still not clear if the multiple mRNA isoforms created by alternative splicing and alternative initiation and termination actually produce the predicted protein variants. Moreover, even if these mRNA do produce protein variants the exact phenotypic effect may be hard to ascertain because they may act on the cellular, tissue or organism levels. The answer to these questions will require technological advances in diverse fields.

Technology

In the past, transcript isoform variations were first characterized by very low throughput technologies such as RT-PCR and Northern blots. The advent of EST libraries showed the extent of transcript isoforms variation and motivated research to further develop splicing-sensitive microarrays capable of genome-wide analysis. The Exon Array developed by Affymetrix is the first commercially available array truly capable of genome-wide detection of isoform variations. The studies presented in this thesis demonstrate the capabilities of the Exon array in detecting transcript isoform variations. However, as demonstrated here the analysis of data generated with this technology requires caution. The large amount of data points generated in these experiments can potentially produce a large number of false results. Therefore, many pre- and post- processing steps are necessary to remove systematic artefacts that generate these erroneous results such as unresponsive, cross-hybridizing, unresponsive probes (chapter 6) and SNPs present in probe targets (chapter 7). In genome-wide studies, multiple testing is also an important factor that generates false positives. The development of new statistical methods and microarray designs are essential for improving their analyses.

Newer technological advances will also readily improve our understanding of gene expression. For instance, the next generation microarrays are likely to combine exon body and splice junction probes. This will greatly improve their sensitivity and will allow the detection of other types of splicing events such as alternative splice site junction usage and intron retention events. In the very near future, ultrahigh-throughput parallel sequencing will become very competitive and eventually eclipse microarrays as the preferred transcriptome profiling tool.

Conclusion

In summary, this thesis demonstrated that isoform variations created by processes such as alternative splicing, alternative transcription initiation and termination are common in human and chimpanzee. This thesis also demonstrates an underlying genetic component to these types of variation. Genetic linkage and allelic association analyses confirm that transcript isoform variations are caused, in part, through single nucleotide polymorphisms. These results show that the effects of genetic variants on gene expression are much more complex than previously believed and constitute an important step towards understanding the functional consequences of such variations.

Bibliography

ABECASIS, G. R., S. S. CHERNY, W. O. COOKSON and L. R. CARDON, 2002 Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet **30:** 97-101.

ABZHANOV, A., M. PROTAS, B. R. GRANT, P. R. GRANT and C. J. TABIN, 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science **305:** 1462-1465.

ADORJAN, P., J. DISTLER, E. LIPSCHER, F. MODEL, J. MULLER *et al.*, 2002 Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res **30:** e21.

AICKIN, M., and H. GENSLER, 1996 Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health **86:** 726-728.

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., WALTER, P., 2002 Molecular Biology of the Cell in *Molecular Biology of the Cell*, edited by S. GIBBS. Garland Science, New York.

ALBERTS, R., P. TERPSTRA, Y. LI, R. BREITLING, J. P. NAP *et al.*, 2007 Sequence polymorphisms cause many false cis eQTLs. PLoS ONE **2:** e622.

ALIOTO, T. S., 2007 U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Res **35:** D110-115.

ALTSHULER, D., L. D. BROOKS, A. CHAKRAVARTI, F. S. COLLINS, M. J. DALY *et al.*, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

ALWINE, J. C., D. J. KEMP and G. R. STARK, 1977 Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci U S A **74:** 5350-5354.

ARA, T., F. LOPEZ, W. RITCHIE, P. BENECH and D. GAUTHERET, 2006 Conservation of alternative polyadenylation patterns in mammalian genes. BMC Genomics **7:** 189.

AUGENLICHT, L. H., M. Z. WAHRMAN, H. HALSEY, L. ANDERSON, J. TAYLOR *et al.*, 1987 Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Res **47:** 6017-6021.

AUTIO, R., S. KILPINEN, M. SAARELA, O. KALLIONIEMI, S. HAUTANIEMI *et al.*, 2009 Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. BMC Bioinformatics **10 Suppl 1:** S24.

BAEK, D., C. DAVIS, B. EWING, D. GORDON and P. GREEN, 2007 Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res **17:** 145-155.

BEAUDOING, E., S. FREIER, J. R. WYATT, J. M. CLAVERIE and D. GAUTHERET, 2000 Patterns of variant polyadenylation signal usage in human genes. Genome Res **10:** 1001-1010.

BEAUDOING, E., and D. GAUTHERET, 2001 Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res **11:** 1520-1526.

BELGRADER, P., J. CHENG, X. ZHOU, L. S. STEPHENSON and L. E. MAQUAT, 1994 Mammalian nonsense codons can be cis effectors of nuclear mRNA half-life. Mol Cell Biol **14:** 8219-8228.

BELTINGER, C. P., P. S. WHITE, J. M. MARIS, E. P. SULMAN, S. J. JENSEN *et al.*, 1996 Physical mapping and genomic structure of the human TNFR2 gene. Genomics **35:** 94-100.

BEMMO, A., D. BENOVOY, T. KWAN, D. J. GAFFNEY, R. V. JENSEN *et al.*, 2008 Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics **9:** 529.

BEN-ARI, S., D. TOIBER, A. S. SAS, H. SOREQ and Y. BEN-SHAUL, 2006 Modulated splicing-associated gene expression in P19 cells expressing distinct acetylcholinesterase splice variants. J Neurochem **97 Suppl 1:** 24-34.

BENDER, R., and S. LANGE, 2001 Adjusting for multiple testing--when and how? J Clin Epidemiol **54:** 343-349.

BENJAMINI, Y., D. DRAI, G. ELMER, N. KAFKAFI and I. GOLANI, 2001 Controlling the false discovery rate in behavior genetics research. Behav Brain Res **125:** 279-284.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300.

BENOVOY, D., T. KWAN and J. MAJEWSKI, 2008 Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. Nucleic Acids Res **36:** 4417-4423.

BETEL, D., M. WILSON, A. GABOW, D. S. MARKS and C. SANDER, 2008 The microRNA.org resource: targets and expression. Nucleic Acids Res **36:** D149-153.

BEYZADE, S., S. ZHANG, Y. K. WONG, I. N. DAY, P. ERIKSSON *et al.*, 2003 Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction. J Am Coll Cardiol **41:** 2130-2137.

BIRNEY, E., J. A. STAMATOYANNOPOULOS, A. DUTTA, R. GUIGO, T. R. GINGERAS *et al.*, 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447:** 799-816.

BLACK, D. L., 2003 Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem **72:** 291-336.

BLACK, D. L., and B. R. GRAVELEY, 2006 Splicing bioinformatics to biology. Genome Biol **7:** 317.

BLANCHETTE, M., A. R. BATAILLE, X. CHEN, C. POITRAS, J. LAGANIERE *et al.*, 2006 Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res **16:** 656-668.

BLAUSTEIN, M., F. PELISCH and A. SREBROW, 2007 Signals, pathways and splicing regulation. Int J Biochem Cell Biol **39:** 2031-2048.
BLENCOWE, B. J., 2006 Alternative splicing: new insights from global analyses. Cell **126:** 37-47.
BOEGER, H., D. A. BUSHNELL, R. DAVIS, J. GRIESENBECK, Y. LORCH *et al.*, 2005 Structural basis of eukaryotic gene transcription. FEBS Lett **579:** 899-903.
BOGUSKI, M. S., C. M. TOLSTOSHEV and D. E. BASSETT, JR., 1994 Gene discovery in dbEST. Science **265:** 1993-1994.
BOLSTAD, B. M., R. A. IRIZARRY, M. ASTRAND and T. P. SPEED, 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19:** 185-193.
BONNEVIE-NIELSEN, V., L. L. FIELD, S. LU, D. J. ZHENG, M. LI *et al.*, 2005 Variation in antiviral 2',5'-oligoadenylate synthetase (2'5'AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am J Hum Genet **76:** 623-633.
BRITTEN, R. J., and E. H. DAVIDSON, 1969 Gene regulation for higher cells: a theory. Science **165:** 349-357.
BRUDNO, M., M. S. GELFAND, S. SPENGLER, M. ZORN, I. DUBCHAK *et al.*, 2001 Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res **29:** 2338-2348.
BUCKLEY, B. A., A. Y. GRACEY and G. N. SOMERO, 2006 The cellular response to heat stress in the goby Gillichthys mirabilis: a cDNA microarray and protein-level analysis. J Exp Biol **209:** 2660-2677.
BURATTI, E., T. DORK, E. ZUCCATO, F. PAGANI, M. ROMANO *et al.*, 2001 Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. EMBO J **20:** 1774-1784.
BURGE, C., and S. KARLIN, 1997 Prediction of complete gene structures in human genomic DNA. J Mol Biol **268:** 78-94.

CACERES, M., J. LACHUER, M. A. ZAPALA, J. C. REDMOND, L. KUDO *et al.*, 2003 Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci U S A **100:** 13030-13035.

CAIRNS, B. R., 1998 Chromatin remodeling machines: similar motors, ulterior motives. Trends Biochem Sci **23:** 20-25.

CALARCO, J. A., Y. XING, M. CACERES, J. P. CALARCO, X. XIAO *et al.*, 2007 Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev **21:** 2963-2975.

CALVO, O., and J. L. MANLEY, 2003 Strange bedfellows: polyadenylation factors at the promoter. Genes Dev **17:** 1321-1327.

CANALES, R. D., Y. LUO, J. C. WILLEY, B. AUSTERMILLER, C. C. BARBACIORU *et al.*, 2006 Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol **24:** 1115-1122.

CHABOT, B., S. BISOTTO and M. VINCENT, 1995 The nuclear matrix phosphoprotein p255 associates with splicing complexes as part of the [U4/U6.U5] tri-snRNP particle. Nucleic Acids Res **23:** 3206-3213.

CHAN, F. K., H. J. CHUN, L. ZHENG, R. M. SIEGEL, K. L. BUI *et al.*, 2000 A domain in TNF receptors that mediates ligand-independent receptor assembly and signaling. Science **288:** 2351-2354.

CHEN, Y., W. LIU, L. NAUMOVSKI and R. L. NEVE, 2003 ASPP2 inhibits APP-BP1-mediated NEDD8 conjugation to cullin-1 and decreases APP-BP1-induced cell proliferation and neuronal apoptosis. J Neurochem **85:** 801-809.

CHENG, H., J. P. HOXIE and W. P. PARKS, 1999 The conserved core of human immunodeficiency virus type 1 Nef is essential for association with Lck and for enhanced viral replication in T-lymphocytes. Virology **264:** 5-15.

CHERRY, J. M., C. BALL, S. WENG, G. JUVIK, R. SCHMIDT *et al.*, 1997 Genetic and physical maps of Saccharomyces cerevisiae. Nature **387:** 67-73.

CHEUNG, V. G., L. K. CONLIN, T. M. WEBER, M. ARCARO, K. Y. JEN *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet **33:** 422-425.

CHEUNG, V. G., R. S. SPIELMAN, K. G. EWENS, T. M. WEBER, M. MORLEY *et al.*, 2005 Mapping determinants of human gene expression by regional and genome-wide association. Nature **437:** 1365-1369.

CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963-971.

CLARK, A. G., M. B. EISEN, D. R. SMITH, C. M. BERGMAN, B. OLIVER *et al.*, 2007a Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

CLARK, R. M., T. N. WAGLER, P. QUIJADA and J. DOEBLEY, 2006 A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet **38:** 594-597.

CLARK, T. A., A. C. SCHWEITZER, T. X. CHEN, M. K. STAPLES, G. LU *et al.*, 2007b Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol **8:** R64.

CLARK, T. A., C. W. SUGNET and M. ARES, JR., 2002 Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science **296:** 907-910.

COHEN, D., I. CHUMAKOV and J. WEISSENBACH, 1993 A first-generation physical map of the human genome. Nature **366:** 698-701.

CONNELLY, S., and J. L. MANLEY, 1988 A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev **2:** 440-452.

CONSORTIUM, T. C. S. A. A., 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

COOPER, S. J., N. D. TRINKLEIN, E. D. ANTON, L. NGUYEN and R. M. MYERS, 2006 Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res **16:** 1-10.

COOPER, T. A., L. WAN and G. DREYFUSS, 2009 RNA and disease. Cell **136:** 777-793.
CRICK, F., 1970 Central dogma of molecular biology. Nature **227:** 561-563.
CUNNINGHAME GRAHAM, D. S., H. MANKU, S. WAGNER, J. REID, K. TIMMS *et al.*, 2007 Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. Hum Mol Genet **16:** 579-591.
CUPERLOVIC-CULF, M., N. BELACEL, A. S. CULF and R. J. OUELLETTE, 2006 Microarray analysis of alternative splicing. OMICS **10:** 344-357.
DAI, M., P. WANG, A. D. BOYD, G. KOSTOV, B. ATHEY *et al.*, 2005 Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res **33:** e175.
DANCKWARDT, S., M. W. HENTZE and A. E. KULOZIK, 2008 3' end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J **27:** 482-498.
DAS, R., J. YU, Z. ZHANG, M. P. GYGI, A. R. KRAINER *et al.*, 2007 SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. Mol Cell **26:** 867-881.
DAVILA LOPEZ, M., and T. SAMUELSSON, 2008 Early evolution of histone mRNA 3' end processing. RNA **14:** 1-10.
DAVULURI, R. V., Y. SUZUKI, S. SUGANO, C. PLASS and T. H. HUANG, 2008 The functional consequences of alternative promoter use in mammalian genomes. Trends Genet **24:** 167-177.
DE HAAN, J. R., R. WEHRENS, S. BAUERSCHMIDT, E. PIEK, R. C. VAN SCHAIK *et al.*, 2007 Interpretation of ANOVA models for microarray data using PCA. Bioinformatics **23:** 184-190.
DE SANDRE-GIOVANNOLI, A., and N. LEVY, 2006 Altered splicing in prelamin A-associated premature aging phenotypes. Prog Mol Subcell Biol **44:** 199-232.
DEMURA, M., R. M. MARTIN, M. SHOZU, S. SEBASTIAN, K. TAKAYAMA *et al.*, 2007 Regional rearrangements in chromosome 15q21 cause formation of

cryptic promoters for the CYP19 (aromatase) gene. Hum Mol Genet **16:** 2529-2541.

DEUTSCH, S., R. LYLE, E. T. DERMITZAKIS, H. ATTAR, L. SUBRAHMANYAN *et al.*, 2005 Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet **14:** 3741-3749.

DIXON, A. L., L. LIANG, M. F. MOFFATT, W. CHEN, S. HEATH *et al.*, 2007 A genome-wide association study of global gene expression. Nat Genet **39:** 1202-1207.

DOSS, S., E. E. SCHADT, T. A. DRAKE and A. J. LUSIS, 2005 Cis-acting expression quantitative trait loci in mice. Genome Res **15:** 681-691.

EARLY, P., J. ROGERS, M. DAVIS, K. CALAME, M. BOND *et al.*, 1980 Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell **20:** 313-319.

EDMONDS, M., 2002 A history of poly A sequences: from formation to factors to function. Prog Nucleic Acid Res Mol Biol **71:** 285-389.

EMILSSON, V., G. THORLEIFSSON, B. ZHANG, A. S. LEONARDSON, F. ZINK *et al.*, 2008 Genetics of gene expression and its effect on disease. Nature **452:** 423-428.

ENARD, W., P. KHAITOVICH, J. KLOSE, S. ZOLLNER, F. HEISSIG *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340-343.

FAUSTINO, N. A., and T. A. COOPER, 2003 Pre-mRNA splicing and human disease. Genes Dev **17:** 419-437.

FIELD, L. L., V. BONNEVIE-NIELSEN, F. POCIOT, S. LU, T. B. NIELSEN *et al.*, 2005 OAS1 splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes. Diabetes **54:** 1588-1591.

FREY, B. J., N. MOHAMMAD, Q. D. MORRIS, W. ZHANG, M. D. ROBINSON *et al.*, 2005 Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. Nat Genet **37:** 991-996.

GARDINA, P. J., T. A. CLARK, B. SHIMADA, M. K. STAPLES, Q. YANG *et al.*, 2006 Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics **7:** 325.
GIBSON, U. E., C. A. HEID and P. M. WILLIAMS, 1996 A novel method for real time quantitative RT-PCR. Genome Res **6:** 995-1001.
GILAD, Y., S. A. RIFKIN and J. K. PRITCHARD, 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet **24:** 408-415.
GILBERT, W., 1978 Why genes in pieces? Nature **271:** 501.
GORING, H. H., J. E. CURRAN, M. P. JOHNSON, T. D. DYER, J. CHARLESWORTH *et al.*, 2007 Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet **39:** 1208-1216.
GRABER, J. H., C. R. CANTOR, S. C. MOHR and T. F. SMITH, 1999 In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc Natl Acad Sci U S A **96:** 14055-14060.
GRAHAM, R. R., C. KYOGOKU, S. SIGURDSSON, I. A. VLASOVA, L. R. DAVIES *et al.*, 2007 Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. Proc Natl Acad Sci U S A **104:** 6758-6763.
GRENS, A., and I. E. SCHEFFLER, 1990 The 5'- and 3'-untranslated regions of ornithine decarboxylase mRNA affect the translational efficiency. J Biol Chem **265:** 11810-11816.
GUPTA, S., D. ZINK, B. KORN, M. VINGRON and S. A. HAAS, 2004 Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics **20:** 2579-2585.
HACIA, J. G., J. B. FAN, O. RYDER, L. JIN, K. EDGEMON *et al.*, 1999 Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet **22:** 164-167.

HAMBLIN, M. T., and A. DI RIENZO, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet **66:** 1669-1679.

HE, G., J. ZHANG, X. W. LI, W. Y. CHEN, Y. X. PAN *et al.*, 2006 Interleukin-10 -1082 promoter polymorphism is associated with schizophrenia in a Han Chinese sib-pair study. Neurosci Lett **394:** 1-4.

HEENEY, J. L., A. G. DALGLEISH and R. A. WEISS, 2006 Origins of HIV and the evolution of resistance to AIDS. Science **313:** 462-466.

HERBEIN, G., and K. A. KHAN, 2008 Is HIV infection a TNF receptor signalling-driven disease? Trends Immunol **29:** 61-67.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nat Rev Genet **6:** 95-108.

HOLSTE, D., G. HUO, V. TUNG and C. B. BURGE, 2006 HOLLYWOOD: a comparative relational database of alternative splicing. Nucleic Acids Res **34:** D56-62.

HOWE, K. J., 2002 RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim Biophys Acta **1577:** 308-324.

HU, G. K., S. J. MADORE, B. MOLDOVER, T. JATKOE, D. BALABAN *et al.*, 2001 Predicting splice variant from DNA chip expression data. Genome Res **11:** 1237-1245.

HUANG, T. H., M. R. PERRY and D. E. LAUX, 1999 Methylation profiling of CpG islands in human breast cancer cells. Hum Mol Genet **8:** 459-470.

HUGHES, T. R., M. MAO, A. R. JONES, J. BURCHARD, M. J. MARTON *et al.*, 2001 Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol **19:** 342-347.

HULL, J., S. CAMPINO, K. ROWLANDS, M. S. CHAN, R. R. COPLEY *et al.*, 2007 Identification of common genetic variation that modulates alternative splicing. PLoS Genet **3:** e99.

IDAGHDOUR, Y., J. D. STOREY, S. J. JADALLAH and G. GIBSON, 2008 A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. PLoS Genet **4:** e1000052.

INTERNATIONAL_HAPMAP_CONSORTIUM, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

IRIZARRY, R. A., B. M. BOLSTAD, F. COLLIN, L. M. COPE, B. HOBBS *et al.*, 2003a Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res **31:** e15.

IRIZARRY, R. A., B. HOBBS, F. COLLIN, Y. D. BEAZER-BARCLAY, K. J. ANTONELLIS *et al.*, 2003b Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264.

IYER, V. R., C. E. HORAK, C. S. SCAFE, D. BOTSTEIN, M. SNYDER *et al.*, 2001 Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409:** 533-538.

JACOBSON, A., and S. W. PELTZ, 1996 Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Annu Rev Biochem **65:** 693-739.

JANSEN, R. C., and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends Genet **17:** 388-391.

JOHNSON, J. M., J. CASTLE, P. GARRETT-ENGELE, Z. KAN, P. M. LOERCH *et al.*, 2003 Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302:** 2141-2144.

KARIN, M., 1990 Too many transcription factors: positive and negative interactions. New Biol **2:** 126-131.

KAROLCHIK, D., R. M. KUHN, R. BAERTSCH, G. P. BARBER, H. CLAWSON *et al.*, 2008 The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res **36:** D773-779.

KENT, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res **12:** 656-664.

KHAITOVICH, P., I. HELLMANN, W. ENARD, K. NOWICK, M. LEINWEBER *et al.*, 2005 Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309:** 1850-1854.

KHAITOVICH, P., B. MUETZEL, X. SHE, M. LACHMANN, I. HELLMANN *et al.*, 2004 Regional patterns of gene expression in human and chimpanzee brains. Genome Res **14:** 1462-1473.

KIM, H., R. KLEIN, J. MAJEWSKI and J. OTT, 2004 Estimating rates of alternative splicing in mammals and invertebrates. Nat Genet **36:** 915-916; author reply 916-917.

KIM, N., A. V. ALEKSEYENKO, M. ROY and C. LEE, 2007 The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res **35:** D93-98.

KIM, S. J., and H. G. MARTINSON, 2003 Poly(A)-dependent transcription termination: continued communication of the poly(A) signal with the polymerase is required long after extrusion in vivo. J Biol Chem **278:** 41691-41701.

KIMURA, K., A. WAKAMATSU, Y. SUZUKI, T. OTA, T. NISHIKAWA *et al.*, 2006 Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res **16:** 55-65.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

KOMURA, D., F. SHEN, S. ISHIKAWA, K. R. FITCH, W. CHEN *et al.*, 2006 Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res **16:** 1575-1584.

KORF, I., P. FLICEK, D. DUAN and M. R. BRENT, 2001 Integrating genomic homology into gene structure prediction. Bioinformatics **17 Suppl 1:** S140-148.

KORNBERG, R. D., 1974 Chromatin structure: a repeating unit of histones and DNA. Science **184:** 868-871.

KORNBERG, R. D., 1999 Eukaryotic transcriptional control. Trends Cell Biol **9:** M46-49.

KORNBERG, R. D., and J. O. THOMAS, 1974 Chromatin structure; oligomers of the histones. Science **184:** 865-868.

KOZAK, M., 1983 Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev **47:** 1-45.

KWAN, T., D. BENOVOY, C. DIAS, S. GURD, C. PROVENCHER *et al.*, 2008 Genome-wide analysis of transcript isoform variation in humans. Nat Genet **40:** 225-231.

KWAN, T., D. BENOVOY, C. DIAS, S. GURD, D. SERRE *et al.*, 2007 Heritability of alternative splicing in the human genome. Genome Res **17:** 1210-1218.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LATCHMAN, D. S., 1997 Transcription factors: an overview. Int J Biochem Cell Biol **29:** 1305-1312.

LE, K., K. MITSOURAS, M. ROY, Q. WANG, Q. XU *et al.*, 2004 Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. Nucleic Acids Res **32:** e180.

LEE, C., and M. ROY, 2004 Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol **5:** 231.

LEE, I., A. A. DOMBKOWSKI and B. D. ATHEY, 2004 Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. Nucleic Acids Res **32:** 681-690.

LEE, W. J., H. MA, E. TAKANO, H. Q. YANG, M. HATANAKA *et al.*, 1992 Molecular diversity in amino-terminal domains of human calpastatin by exon skipping. J Biol Chem **267:** 8437-8442.

LEIS, J. P., and J. HURWITZ, 1972 RNA-dependent DNA polymerase activity of RNA tumor viruses. II. Directing influence of RNA in the reaction. J Virol **9:** 130-142.

LEMON, B., and R. TJIAN, 2000 Orchestrated response: a symphony of transcription factors for gene control. Genes Dev **14:** 2551-2569.

LEWIS, B. P., C. B. BURGE and D. P. BARTEL, 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120:** 15-20.

LEWIS, B. P., R. E. GREEN and S. E. BRENNER, 2003 Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A **100:** 189-192.

LEWIS, J. D., S. I. GUNDERSON and I. W. MATTAJ, 1995 The influence of 5' and 3' end structures on pre-mRNA metabolism. J Cell Sci Suppl **19:** 13-19.

LIEB, J. D., X. LIU, D. BOTSTEIN and P. O. BROWN, 2001 Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet **28:** 327-334.

LIU, Q. R., D. WALTHER, T. DRGON, O. POLESSKAYA, T. G. LESNICK *et al.*, 2005 Human brain derived neurotrophic factor (BDNF) genes, splicing patterns, and assessments of associations with substance abuse and Parkinson's Disease. Am J Med Genet B Neuropsychiatr Genet **134B:** 93-103.

LIU, S., and R. B. ALTMAN, 2003 Large scale study of protein domain distribution in the context of alternative splicing. Nucleic Acids Res **31:** 4828-4835.

LOGAN, J., E. FALCK-PEDERSEN, J. E. DARNELL, JR. and T. SHENK, 1987 A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A **84:** 8306-8310.

LUKONG, K. E., K. W. CHANG, E. W. KHANDJIAN and S. RICHARD, 2008 RNA-binding proteins in human genetic disease. Trends Genet **24:** 416-425.
LUTZ, C. S., 2008 Alternative polyadenylation: a twist on mRNA 3' end formation. ACS Chem Biol **3:** 609-617.
MAGEN, A., and G. AST, 2005 The importance of being divisible by three in alternative splicing. Nucleic Acids Res **33:** 5574-5582.
MAJEWSKI, J., and J. OTT, 2002 Distribution and characterization of regulatory elements in the human genome. Genome Res **12:** 1827-1836.
MANDEL, C. R., Y. BAI and L. TONG, 2008 Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci **65:** 1099-1122.
MANIATIS, T., and R. REED, 2002 An extensive network of coupling among gene expression machines. Nature **416:** 499-506.
MARCU, K. B., S. A. BOSSONE and A. J. PATEL, 1992 myc function and regulation. Annu Rev Biochem **61:** 809-860.
MATLIN, A. J., F. CLARK and C. W. SMITH, 2005 Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol **6:** 386-398.
MAYEDA, A., and A. R. KRAINER, 1999 Preparation of HeLa cell nuclear and cytosolic S100 extracts for in vitro splicing. Methods Mol Biol **118:** 309-314.
MCCARTHY, B. J., and J. J. HOLLAND, 1965 Denatured DNA as a direct template for in vitro protein synthesis. Proc Natl Acad Sci U S A **54:** 880-886.
MCCRACKEN, S., N. FONG, E. ROSONINA, K. YANKULOV, G. BROTHERS *et al.*, 1997a 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev **11:** 3306-3318.
MCCRACKEN, S., N. FONG, K. YANKULOV, S. BALLANTYNE, G. PAN *et al.*, 1997b The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature **385:** 357-361.

MODREK, B., and C. J. LEE, 2003 Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet **34:** 177-180.

MODREK, B., A. RESCH, C. GRASSO and C. LEE, 2001 Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res **29:** 2850-2859.

MORAN, G., C. STOKES, S. THEWES, B. HUBE, D. C. COLEMAN *et al.*, 2004 Comparative genomics using Candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in Candida dubliniensis. Microbiology **150:** 3363-3382.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

MORTILLARO, M. J., B. J. BLENCOWE, X. WEI, H. NAKAYASU, L. DU *et al.*, 1996 A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. Proc Natl Acad Sci U S A **93:** 8253-8257.

MULLIS, K., F. FALOONA, S. SCHARF, R. SAIKI, G. HORN *et al.*, 1986 Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol **51 Pt 1:** 263-273.

NAEF, F., and M. O. MAGNASCO, 2003 Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys Rev E Stat Nonlin Soft Matter Phys **68:** 011906.

NAKANISHI, T., K. J. BAILEY-DELL, B. A. HASSEL, K. SHIOZAWA, D. M. SULLIVAN *et al.*, 2006 Novel 5' untranslated region variants of BCRP mRNA are differentially expressed in drug-selected cancer cells and in normal human tissues: implications for drug resistance, tissue-specific expression, and alternative promoter usage. Cancer Res **66:** 5007-5011.

NEILSON, J. R., and P. A. SHARP, 2008 Small RNA regulators of gene expression. Cell **134:** 899-902.

NEMBAWARE, V., K. H. WOLFE, F. BETTONI, J. KELSO and C. SEOIGHE, 2004 Allele-specific transcript isoforms in human. FEBS Lett **577:** 233-238.
NEUGEBAUER, K. M., 2002 On the importance of being co-transcriptional. J Cell Sci **115:** 3865-3871.
NEUMANN, M., D. M. SAMPATHU, L. K. KWONG, A. C. TRUAX, M. C. MICSENYI *et al.*, 2006 Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science **314:** 130-133.
NISSIM-RAFINIA, M., and B. KEREM, 2005 The splicing machinery is a genetic modifier of disease severity. Trends Genet **21:** 480-483.
OKONIEWSKI, M. J., Y. HEY, S. D. PEPPER and C. J. MILLER, 2007a High correspondence between Affymetrix exon and standard expression arrays. Biotechniques **42:** 181-185.
OKONIEWSKI, M. J., and C. J. MILLER, 2008 Comprehensive analysis of affymetrix exon arrays using BioConductor. PLoS Comput Biol **4:** e6.
OKONIEWSKI, M. J., T. YATES, S. DIBBEN and C. J. MILLER, 2007b An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. Genome Biol **8:** R79.
OTT, M., J. L. LOVETT, L. MUELLER and E. VERDIN, 1998 Superinduction of IL-8 in T cells by HIV-1 Tat protein is mediated through NF-kappaB factors. J Immunol **160:** 2872-2880.
PAN, Q., M. A. BAKOWSKI, Q. MORRIS, W. ZHANG, B. J. FREY *et al.*, 2005 Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet **21:** 73-77.
PAN, Q., O. SHAI, L. J. LEE, B. J. FREY and B. J. BLENCOWE, 2008 Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40:** 1413-1415.
PAN, Q., O. SHAI, C. MISQUITTA, W. ZHANG, A. L. SALTZMAN *et al.*, 2004 Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell **16:** 929-941.
PENNISI, E., 1997 Opening the way to gene activity. Science **275:** 155-157.

POLLACK, J. R., C. M. PEROU, A. A. ALIZADEH, M. B. EISEN, A. PERGAMENSCHIKOV *et al.*, 1999 Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet **23:** 41-46.

POUSTKA, A., T. POHL, D. P. BARLOW, G. ZEHETNER, A. CRAIG *et al.*, 1986 Molecular approaches to mammalian genetics. Cold Spring Harb Symp Quant Biol **51 Pt 1:** 131-139.

PURCELL, S., B. NEALE, K. TODD-BROWN, L. THOMAS, M. A. FERREIRA *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81:** 559-575.

PURDOM, E., K. M. SIMPSON, M. D. ROBINSON, J. G. CONBOY, A. V. LAPUK *et al.*, 2008 FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics **24:** 1707-1714.

QU, H. Q., Y. LU, L. MARCHAND, F. BACOT, R. FRECHETTE *et al.*, 2007 Genetic control of alternative splicing in the TAP2 gene: possible implication in the genetics of type 1 diabetes. Diabetes **56:** 270-275.

REED, R., 2003 Coupling transcription, splicing and mRNA export. Curr Opin Cell Biol **15:** 326-331.

REN, B., F. ROBERT, J. J. WYRICK, O. APARICIO, E. G. JENNINGS *et al.*, 2000 Genome-wide location and function of DNA binding proteins. Science **290:** 2306-2309.

RIGAULT, C., F. LE BORGNE and J. DEMARQUOY, 2006 Genomic structure, alternative maturation and tissue expression of the human BBOX1 gene. Biochim Biophys Acta **1761:** 1469-1481.

ROBINSON, M. D., and T. P. SPEED, 2007 A comparison of Affymetrix gene expression arrays. BMC Bioinformatics **8:** 449.

RODRIGUES-LIMA, F., M. JOSEPHS, M. KATAN and B. CASSINAT, 2001 Sequence analysis identifies TTRAP, a protein that associates with CD40 and TNF receptor-associated factors, as a member of a superfamily of divalent cation-dependent phosphodiesterases. Biochem Biophys Res Commun **285:** 1274-1279.

ROEDER, R. G., and W. J. RUTTER, 1969 Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature **224:** 234-237.
ROSENFELD, M. G., C. R. LIN, S. G. AMARA, L. STOLARSKY, B. A. ROOS *et al.*, 1982 Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events. Proc Natl Acad Sci U S A **79:** 1717-1721.
ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386.
RUSSELL, J., and J. C. ZOMERDIJK, 2006 The RNA polymerase I transcription machinery. Biochem Soc Symp**:** 203-216.
RUSSO, A., G. RUSSO, M. CUCCURESE, C. GARBI and C. PIETROPAOLO, 2006 The 3'-untranslated region directs ribosomal protein-encoding mRNAs to specific cytoplasmic regions. Biochim Biophys Acta **1763:** 833-843.
SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928-933.
SANDBERG, R., J. R. NEILSON, A. SARMA, P. A. SHARP and C. B. BURGE, 2008 Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science **320:** 1643-1647.
SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.
SCHMUCKER, D., J. C. CLEMENS, H. SHU, C. A. WORBY, J. XIAO *et al.*, 2000 Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell **101:** 671-684.
SCHRAM, B. R., and T. L. ROTHSTEIN, 2003 NF-kappa B is required for surface Ig-induced Fas resistance in B cells. J Immunol **170:** 3118-3124.
SHATKIN, A. J., and J. L. MANLEY, 2000 The ends of the affair: capping and polyadenylation. Nat Struct Biol **7:** 838-842.

SHETH, N., X. ROCA, M. L. HASTINGS, T. ROEDER, A. R. KRAINER *et al.*, 2006 Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res **34:** 3955-3967.

SHI, L., L. H. REID, W. D. JONES, R. SHIPPY, J. A. WARRINGTON *et al.*, 2006 The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol **24:** 1151-1161.

SHYAMSUNDAR, R., Y. H. KIM, J. P. HIGGINS, K. MONTGOMERY, M. JORDEN *et al.*, 2005 A DNA microarray survey of gene expression in normal human tissues. Genome Biol **6:** R22.

SIEPEL, A., M. DIEKHANS, B. BREJOVA, L. LANGTON, M. STEVENS *et al.*, 2007 Targeted discovery of novel human exons by comparative genomics. Genome Res **17:** 1763-1773.

SIEPEL, A., and D. HAUSSLER, 2004 Computational identification of evolutionarily conserved exons, pp. in *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. ACM Press, San Diego, California, USA.

SINGH, R., J. VALCARCEL and M. R. GREEN, 1995 Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science **268:** 1173-1176.

SLIWERSKA, E., F. MENG, T. P. SPEED, E. G. JONES, W. E. BUNNEY *et al.*, 2007 SNPs on chips: the hidden genetic code in expression arrays. Biol Psychiatry **61:** 13-16.

SPIELMAN, R. S., L. A. BASTONE, J. T. BURDICK, M. MORLEY, W. J. EWENS *et al.*, 2007 Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet **39:** 226-231.

SRINIVASAN, K., L. SHIUE, J. D. HAYES, R. CENTERS, S. FITZWATER *et al.*, 2005 Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods **37:** 345-359.

STANCHEV, B., and V. STANCHEV, 1984 [Split genes and RNA splicing in eukaryotic cells]. Eksp Med Morfol **23:** 166-170.
STERN, D. L., 1998 A role of Ultrabithorax in morphological differences between Drosophila species. Nature **396:** 463-466.
STOREY, J. D., J. MADEOY, J. L. STROUT, M. WURFEL, J. RONALD *et al.*, 2007 Gene-expression variation within and among human populations. Am J Hum Genet **80:** 502-509.
STRANGER, B. E., M. S. FORREST, A. G. CLARK, M. J. MINICHIELLO, S. DEUTSCH *et al.*, 2005 Genome-wide associations of gene expression variation in humans. PLoS Genet **1:** e78.
STRANGER, B. E., M. S. FORREST, M. DUNNING, C. E. INGLE, C. BEAZLEY *et al.*, 2007a Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science **315:** 848-853.
STRANGER, B. E., A. C. NICA, M. S. FORREST, A. DIMAS, C. P. BIRD *et al.*, 2007b Population genomics of human gene expression. Nat Genet **39:** 1217-1224.
STUMPTNER-CUVELETTE, P., M. JOUVE, J. HELFT, M. DUGAST, A. S. GLOUZMAN *et al.*, 2003 Human immunodeficiency virus-1 Nef expression induces intracellular accumulation of multivesicular bodies and major histocompatibility complex class II complexes: potential role of phosphatidylinositol 3-kinase. Mol Biol Cell **14:** 4857-4870.
SUGNET, C. W., K. SRINIVASAN, T. A. CLARK, G. O'BRIEN, M. S. CLINE *et al.*, 2006 Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol **2:** e4.
SUN, T., Y. GAO, W. TAN, S. MA, Y. SHI *et al.*, 2007 A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. Nat Genet **39:** 605-613.
SVEJSTRUP, J. Q., P. VICHI and J. M. EGLY, 1996 The multiple roles of transcription/repair factor TFIIH. Trends Biochem Sci **21:** 346-350.

TAKANO, E., T. NOSAKA, W. J. LEE, K. NAKAMURA, T. TAKAHASHI *et al.*, 1993 Molecular diversity of calpastatin in human erythroid cells. Arch Biochem Biophys **303:** 349-354.

TAKEDA, J., Y. SUZUKI, M. NAKAO, T. KURODA, S. SUGANO *et al.*, 2007 H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. Nucleic Acids Res **35:** D104-109.

TEN HAAFT, P., K. MURTHY, M. SALAS, H. MCCLURE, R. DUBBES *et al.*, 2001 Differences in early virus loads with different phenotypic variants of HIV-1 and SIV(cpz) in chimpanzees. AIDS **15:** 2085-2092.

TERGAONKAR, V., 2006 NFkappaB pathway: a good signaling paradigm and therapeutic target. Int J Biochem Cell Biol **38:** 1647-1653.

THE *C. ELEGANS* SEQUENCING CONSORTIM, 1998 Genome sequence of the nematode C. elegans: a platform for investigating biology. Science **282:** 2012-2018.

THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

THE INTERNATIONAL HAPMAP CONSORTIUM, 2003 The International HapMap Project. Nature **426:** 789-796.

THE INTERNATIONAL HAPMAP CONSORTIUM, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

THOMAS, D. C., R. W. HAILE and D. DUGGAN, 2005 Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet **77:** 337-345.

THORSEN, K., K. D. SORENSEN, A. S. BREMS-ESKILDSEN, C. MODIN, M. GAUSTADNES *et al.*, 2008 Alternative splicing in colon, bladder, and prostate cancer identified by exon-array analysis. Mol Cell Proteomics.

TIAN, B., J. HU, H. ZHANG and C. S. LUTZ, 2005 A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res **33:** 201-212.

TOURNAMILLE, C., Y. COLIN, J. P. CARTRON and C. LE VAN KIM, 1995 Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet **10:** 224-228.

TRAN, D. P., S. J. KIM, N. J. PARK, T. M. JEW and H. G. MARTINSON, 2001 Mechanism of poly(A) signal transduction to RNA polymerase II in vitro. Mol Cell Biol **21:** 7495-7508.

TRIVEDI, H. N., F. A. PLUMMER, A. O. ANZALA, E. NJAGI, J. J. BWAYO *et al.*, 2001 Resistance to HIV-1 infection among African sex workers is associated with global hyporesponsiveness in interleukin 4 production. FASEB J **15:** 1795-1797.

ULE, J., A. ULE, J. SPENCER, A. WILLIAMS, J. S. HU *et al.*, 2005 Nova regulates brain-specific splicing to shape the synapse. Nat Genet **37:** 844-852.

UZAWA, T., A. YAMAGISHI and T. OSHIMA, 2002 Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, Thermus thermophilus HB27 and Sulfolobus tokodaii strain 7. J Biochem **131:** 849-853.

VALENCIA-SANCHEZ, M. A., J. LIU, G. J. HANNON and R. PARKER, 2006 Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev **20:** 515-524.

VALLEE, M., C. ROBERT, S. METHOT, M. F. PALIN and M. A. SIRARD, 2006 Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes. BMC Genomics **7:** 113.

VALVERDE, D., R. RIVEIRO-ALVAREZ, S. BERNAL, K. JAAKSON, M. BAIGET *et al.*, 2006 Microarray-based mutation analysis of the ABCA4 gene in

Spanish patients with Stargardt disease: evidence of a prevalent mutated allele. Mol Vis **12:** 902-908.

VELCULESCU, V. E., L. ZHANG, B. VOGELSTEIN and K. W. KINZLER, 1995 Serial analysis of gene expression. Science **270:** 484-487.

VENABLES, J. P., 2006 Unbalanced alternative splicing and its significance in cancer. Bioessays **28:** 378-386.

VIGNAUD, P., P. DURINGER, H. T. MACKAYE, A. LIKIUS, C. BLONDEL *et al.*, 2002 Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature **418:** 152-155.

VINCENT, M., P. LAURIAULT, M. F. DUBOIS, S. LAVOIE, O. BENSAUDE *et al.*, 1996 The nuclear matrix protein p255 is a highly phosphorylated form of RNA polymerase II largest subunit which associates with spliceosomes. Nucleic Acids Res **24:** 4649-4652.

WADE, P. A., D. PRUSS and A. P. WOLFFE, 1997 Histone acetylation: chromatin in action. Trends Biochem Sci **22:** 128-132.

WAHLE, E., and U. RUEGSEGGER, 1999 3'-End processing of pre-mRNA in eukaryotes. FEMS Microbiol Rev **23:** 277-295.

WALTER, N. A., S. K. MCWEENEY, S. T. PETERS, J. K. BELKNAP, R. HITZEMANN *et al.*, 2007 SNPs matter: impact on detection of differential expression. Nat Methods **4:** 679-680.

WANG, E. T., R. SANDBERG, S. LUO, I. KHREBTUKOVA, L. ZHANG *et al.*, 2008 Alternative isoform regulation in human tissue transcriptomes. Nature **456:** 470-476.

WANG, G. S., and T. A. COOPER, 2007 Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet **8:** 749-761.

WANG, H., E. HUBBELL, J. S. HU, G. MEI, M. CLINE *et al.*, 2003 Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics **19 Suppl 1:** i315-322.

WANG, Z., and C. B. BURGE, 2008 Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA **14:** 802-813.

WATERSTON, R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

WATSON, J. D., and F. H. CRICK, 1953 Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature **171:** 737-738.

WEISSMANN, C., 2004 The state of the prion. Nat Rev Microbiol **2:** 861-871.

WICKENS, M., P. ANDERSON and R. J. JACKSON, 1997 Life and death in the cytoplasm: messages from the 3' end. Curr Opin Genet Dev **7:** 220-232.

WIRTH, B., L. BRICHTA and E. HAHNEN, 2006 Spinal muscular atrophy and therapeutic prospects. Prog Mol Subcell Biol **44:** 109-132.

WOLFFE, A. P., 1991 RNA polymerase III transcription. Curr Opin Cell Biol **3:** 461-466.

WORKMAN, J. L., and A. R. BUCHMAN, 1993 Multiple functions of nucleosomes and regulatory factors in transcription. Trends Biochem Sci **18:** 90-95.

WRAY, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8:** 206-216.

WU, L., J. FAN and J. G. BELASCO, 2006 MicroRNAs direct rapid deadenylation of mRNA. Proc Natl Acad Sci U S A **103:** 4034-4039.

XING, Y., Z. OUYANG, K. KAPUR, M. P. SCOTT and W. H. WONG, 2007 Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol Biol Evol **24:** 1283-1285.

XU, Q., B. MODREK and C. LEE, 2002 Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res **30:** 3754-3766.

YAGER, T. D., C. T. MCMURRAY and K. E. VAN HOLDE, 1989 Salt-induced release of DNA from nucleosome core particles. Biochemistry **28:** 2271-2281.

YAN, J., and T. G. MARR, 2005 Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res **15:** 369-375.
YAN, P. S., C. M. CHEN, H. SHI, F. RAHMATPANAH, S. H. WEI *et al.*, 2001 Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. Cancer Res **61:** 8375-8380.
YE, S., P. ERIKSSON, A. HAMSTEN, M. KURKINEN, S. E. HUMPHRIES *et al.*, 1996 Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. J Biol Chem **271:** 13055-13060.
YEO, G., and C. B. BURGE, 2004 Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol **11:** 377-394.
YEO, G., D. HOLSTE, G. KREIMAN and C. B. BURGE, 2004a Variation in alternative splicing across human tissues. Genome Biol **5:** R74.
YEO, G., S. HOON, B. VENKATESH and C. B. BURGE, 2004b Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci U S A **101:** 15700-15705.
YEO, G. W., E. VAN NOSTRAND, D. HOLSTE, T. POGGIO and C. B. BURGE, 2005 Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A **102:** 2850-2855.
YONAHA, M., and N. J. PROUDFOOT, 1999 Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. Mol Cell **3:** 593-600.
ZAKHARKIN, S. O., K. KIM, T. MEHTA, L. CHEN, S. BARNES *et al.*, 2005 Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics **6:** 214.
ZHANG, C., H. R. LI, J. B. FAN, J. WANG-RODRIGUEZ, T. DOWNS *et al.*, 2006 Profiling alternatively spliced mRNA isoforms for prostate cancer classification. BMC Bioinformatics **7:** 202.

ZHANG, L., C. WU, R. CARTA and H. ZHAO, 2007 Free energy of DNA duplex formation on short oligonucleotide microarrays. Nucleic Acids Res **35:** e18.

ZHANG, W., S. DUAN, E. O. KISTNER, W. K. BLEIBEL, R. S. HUANG *et al.*, 2008 Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet **82:** 631-640.

ZHANG, X. H., K. A. HELLER, I. HEFTER, C. S. LESLIE and L. A. CHASIN, 2003 Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res **13:** 2637-2650.

ZIELENSKI, J., 2000 Genotype and phenotype in cystic fibrosis. Respiration **67:** 117-133.

Annexe

McGill University

APPLICATION TO USE BIOHAZARDOUS MATERIALS*

Projects involving potentially biohazardous materials should not be commenced without approval from the Environmental Safety Office. Submit applications before 1) starting new projects, 2) renewing existing projects, or 3) changing the nature of the biohazardous materials within existing projects.

1. PRINCIPAL INVESTIGATOR: Dr. Robert Sladek PHONE: (514) 398-5458
DEPARTMENT: McGill University and Genome Quebec Innovation Centre FAX: (514) 398-1738
ADDRESS: 740 Ave. Dr. Penfield, Montreal, Quebec H3A 1A4 E-MAIL: rob.sladek@mail.mcgill.ca

PROJECT TITLE: Gene Regulation in Disease / Functional characterization of susceptibility loci for type 2 diabetes mellitus.

2. EMERGENCY: Person(s) designated to handle emergencies
Name: Dr. Robert Sladek Phone No: work: 514-398-5458 home: (514) 842-9406
Name: Huan Chu Pham Dang Phone No: work: ext. 00338 home: (514) 840-0922

3. FUNDING SOURCE OR AGENCY (specify): Genome Canada / CIHR
Grant No.: RI-MUHC 6182 / MOP-84332 Beginning date: April 2006 End date: June 2012

4. Indicate if this is
☐ Renewal: procedures previously approved without alterations.
Approval End Date: _____
☐ New funding source: project previously reviewed and approved under an application to another agency.
Agency: _____ Approval End Date: _____
☐ New project: project not previously reviewed.
☒ Approved project: change in biohazardous materials or procedures.
☐ Work/project involving biohazardous materials in teaching/diagnostics.

CERTIFICATION STATEMENT: The Environmental Safety Office approves the experimental procedures proposed and certifies with the applicant that the experiment will be in accordance with the principles outlined in Health Canada's "Laboratory Biosafety Guidelines" and in the "McGill Laboratory Biosafety Manual".

Containment Level (select one): ☐ 1 ☐ 2 ☒ 2 with additional precautions ☐ 3

Principal Investigator or course director: Robert Sladek (SIGNATURE) date: 20 (day) June (month) 2008 (year)

Approved by: Environmental Safety Office: C. St. Louis (SIGNATURE) date: 20 (day) 06 (month) 08 (year)

Expiry: 30 (day) 06 (month) 12 (year)

*as defined in the "McGill Laboratory Biosafety Manual"

Centre universitaire de santé McGill
McGill University Health Centre
Les meilleurs soins pour la vie
The Best Care for Life

September 16, 2008

Dr. Tomi Pastinen
Montreal Genome Centre
740, avenue du Docteur Penfield
Montreal, Quebec
H3A 1A4

RE: REC. 02-009 entitled *"GRID: Gene Regulators in Disease."*

Dear Dr. Hudson:

We have received an Application for Continuing Review for the research study referenced above, along with an explanation of the delay in submitting this application for continuing review.

It is noted in your correspondence dated 12 September 2008 that since the departure of Dr. Tom Hudson, you have taken over as the leader of the GRiD project, and the reason you overlooked to submit a continuing review application to the Research Ethics Office for the period of July 2007 to September 2008 was of organizational nature.

The study file shows that the study was active, however, no subjects were enrolled in the study since study initiation. We take this opportunity to remind you of the importance in submitting an application for Continuing Review to the REB prior to the expiration of approval on an annual basis to comply with the regulations for continuing review. This review is based on the progress report submitted at the interval set by the REB initial review. Continuing review may establish new conditions based on the study events, and these reviews continue at the same interval until submission of a Termination Report.

At this time, we are pleased to inform you that your explanation was found ethically acceptable and we hereby grant you re-approval for your study via expedited review by the Co-Chairman on September 16, 2008, valid until July 31, 2009. We ask however that in the future a Continuing Review Application be submitted on a yearly basis to prevent the REB from closing the study due to lapsed approval.

At the MUHC sponsored research activities that require US federal assurance are conducted under Federal Wide Assurance (FWA) 00000840.

Should any revision to the research, or other unanticipated development occur prior to the next required review, you are obligated to report in writing promptly to the REB. It is not permitted by regulation to initiate a proposed study modification prior to REB approval.

We trust this will prove satisfactory to you.

Sincerely,

Denis Cournoyer, M. D.
Co-Chairman
Genetics/Population Research/Gen Investigator Initiated Studies
MUHC-Montreal General Hospital

Centre d'innovation Génome Québec et Université McGill
McGill University and Genome Quebec Innovation Centre

September 12, 2008

Denis Coumoyer, MD, Director
Co-Chairman
Genetics/Population Research/Gen Investigator Initiated Studies
MUHC Research Ethics Office
Montreal General Hospital
Research Ethics Office
1650 Cedar Avenue, Room S1.509
Montreal, Quebec, H3G 1A4

Re: REC. 02-009 entitled "GRiD: Gene Regulators in Disease"

Dear Dr. Cournoyer,

My sincere apologies for having delayed to submit the "Continuing review application" (REC. 02-0009) for July 12th 2007. The reasons for not submitting the renewal was not because of the project, but were of an organizational nature.

I acknowledge that we received a renewal note from your office. However, at the time Dr. Eef Harmsen the Project Manager of the GRiD project was on medical leave for three months because of surgery. In addition, after succeeding Tom Hudson as the leader of the GRiD project, my first mandate was to prepare an international midterm review for Genome Canada/Quebec. This review included a mid-term report for July 13, 2007, and a face-to-face meeting in Vancouver on Sept 25, 2007. Gladly, the GRiD project was evaluated as "Excellent to Outstanding", with the ensuing recommendation to proceed as planned. However, during this very hectic phase we did overlook your request to renew the ethical approval for the GRiD project.

The objectives of the GRiD project, and all collaborators, are still the same, which is to identify potential regulatory SNPs in genes which are involved in complex diseases. However, we have implemented two incremental changes which may be of interest for you:

1. In addition to using DNA and RNA from LCL cells from the HapMap project (Coriell), which are grown locally by us, we are now also using human fibroblasts obtained from the McGill repository (Dr Rosenblatt) and from Coriell (USA). The commercially available McGill and Coriell cells are from annonymized donors and are grown by us to obtain DNA and RNA for allelic expression studies.

2. We use novel bead and micro array screenings technology, which allows genome-wide screening and mapping for AE and AS. In addition, some of the regulatory haplotypes in 4 HapMap individuals are sequenced with high-throughput 454 sequencing to identify potential rare regulatory variants.

Thanks for considering our explanation. We hope that this information in addition to the "Continuing review application" form is helpful. If not, I will be happy to provide any other information you might require.

Sincerely,

Tomi Pastinen, MD PhD
Assistant Professor and Canada Research Chair in Human Genomics
Scientific Director GRiD
Departments of Human Genetics and Medical Genetics
McGill University and Genome Quebec Innovation Centre
740 Dr. Penfield Avenue, Rm 6202
Montreal, Quebec, H3A 1A4
tel: 514 398 1777
fax: 514 398 1738
tomi.pastinen@mcgill.ca

Version 11/07

McGILL UNIVERSITY HEALTH CENTRE
RESEARCH ETHICS BOARD

- APPLICATION FOR CONTINUING REVIEW -

1. RESEARCH IDENTIFICATION

Study Title :
GRiD: Gene Regulators in Disease

MUHC Study Code: REC. 02-009 Sponsor's Protocol Number: _____

Previous REB Review : 12 / 06 / 2006 (Day / Month / Year) Approval Expiration : 12 / 06 / 2007 (Day / Month / Year)

Initial REB approval was provided following: ☒ Full Board Review ☐ Expedited Review

Please answer all questions and check all that apply -

➢ *This symbol requires you to take note of the information.*

✓ *This symbol indicates that a supplementary document may be required to complete the review.*

2. RESEARCH PERSONNEL

Principal Investigator: Tomi Pastinen, PhD, MD Study Coordinator: Eef Harmsen, PhD

➢ *A change of Investigator requires a Study Amendment, Sponsor's permission and REB approval for the Amendment. Use Annex A to report changes in the Research Personnel to the REB.*

3. CURRENT STUDY STATUS

☒ Active	*No subjects have been recruited to the study yet.*
☐ Active	*Subjects continue to be recruited to the study.*
☐ On Hold	*Recruitment is interrupted to conduct interim data analysis.*
☐ On Hold	*Discontinuance issued by Health Canada, or Clinical Hold by FDA.*
☐ On Hold	*Status issued by the REB.*
☐ On Hold	*Issued by investigator, sponsor or research cooperative group.*
☐ Closed to Recruitment	*Recruitment is complete and subjects are receiving active treatment.*
☐ Closed to Recruitment	*Recruitment is complete and subjects are in long-term follow-up.*
☐ Closed to Recruitment	*Study is permanently closed to recruitment, data is in final analysis.*

Version 08/07

4. SUBJECT REPORTING

a) Does study approval include legally incompetent adult or child subjects? ☐ Yes ☒ No

b) What number of subjects are projected for enrollment at the MUHC? 0

c) Report on MUHC study enrollment since:

	Previous Review	Study Initiation
Total number of MUHC subjects enrolled:		
Number of minor children enrolled:		
Number of incompetent adults enrolled:		
Number of subjects who completed the study:		
Number of subjects who withdrew voluntarily:		
Number of subjects withdrawn for safety reasons:		
Number of subject deaths "on-study":		

✓ *If a subject(s) was withdrawn, or withdrew voluntarily from the study, please append a summary to provide the explanation for each withdrawl. A subject considered as "lost to follow up" is reported at "Number of subjects who withdrew voluntarily".*

d) Has any breach of confidentiality or intrusion of privacy occurred on the study? ☐ Yes ☒ No

✓ *If Yes, please append new or previous correspondence to REB*

e) Were any complaints about the study brought to investigator's attention? ☐ Yes ☒ No

✓ *If Yes, append a report describing the issue(s) or refer to previous report to REB.*

5. FREE AND INFORMED CONSENT

a) Has new information emerged since the previous review that might influence a subject's willingness to continue as a volunteer in the research? ☐ Yes ☒ No

If Yes, please comment or refer to previous correspondence to REB.

N/A We are using commercially available cells (Coriell and McGill)

b) Has the risk/benefit evaluation changed since the previous review? ☐ Yes ☒ No

If Yes, please comment or refer to previous correspondence to REB.

c) What is the date of the consent document in current use? _____

d) What is the date of the assent document in current use ? _____ ☒ N/A

✓ ***Please append the current consent and assent documents or revised document versions with modifications highlighted for REB review. Revised versions in both English and French must be submitted before the revised consent and assent documents will be approved.***

6. STUDY PROGRESS

a) Please provide the version and date of the research protocol in current use:

Version: _____ *Date:* _____

b) Since previous review was a change made to the study recruitment strategy? ☐ Yes ☒ No

If Yes, please comment or refer to previous correspondence to REB

c) Please summarize any multi-centre trial report, recent literature, interim findings and/or study modifications since the last review. *(Append additional document as necessary)*

N/A

d) Has an unreported protocol revision occurred since previous review? ☐ Yes ☒ No

✓ ***If Yes, please append the information using the "Protocol Violation Report"***

e) Has an unreported protocol violation occurred since previous review? ☐ Yes ☒ No

✓ ***If Yes, please append the information using the "Protocol Violation Report"***

7. FINANCES AND RESOURCES

➢ ***A change in resource utilization or to a condition of financial support in the existing Study Agreement must be approved by REB, and by Office of Clinical Contracts (OCC) prior to implementing the change. Contact OCC concerning every change to the Study Agreemen***

a) Was there any change in a financial arrangement? ☐ Yes ☒ No

✓ ***If Yes, forward the information to the REB and OCC (if not already done)***

b) Did any change occur that could have impact on resource utilization? ☐ Yes ☒ No

✓ ***If Yes, please append Director of Professional Services authorization for the change***

8. SAFETY REPORTING

a) If the study is clinical trial with a therapeutic component was a "Confirmation of Participation in Research" form filed in the Medical Record of each subject? ☐ Yes ☐ No ☐ N/A

b) Were all serious adverse events involving MUHC subjects that occurred since the previous REB review reported to the REB? ☐ Yes ☐ No ☐ N/A

If No, please explain:

N/A this is not a clinical trial

c) Were all "serious and unexpected" adverse events involving non-MUHC subjects received since the previous REB review reported to the REB? ☐ Yes ☐ No ☐ N/A

If No, please explain:

N/A. We use commercially available cells

d) Were any "serious and expected" adverse events reported in greater frequency (trend) than expected by the sponsor? ☐ Yes ☐ No

✓ ***If Yes, please append correspondence with details of the unanticipated frequency.***

e) Does a Safety Committee (DSC or DSMB) oversee the study? ☐ Yes ☒ No

If Yes, what is the date of the most recent summary report? _____

If Yes, was the most recent Summary Report submitted to the REB? ☐ Yes ☐ No

✓ ***Please append all Safety Committee Summary Reports occurring since the previous review.***

f) Did any study amendment require review and authorization by the Radiation Protection Service (RPS) or by the Institutional Biosafety Committee (IBC)? ☒ Yes ☐ No

✓ ***If Yes, submit the corresponding RPS or IBC authorization document(s).***

9. AUTHORIZING SIGNATURES

Signature below certifies that: *(original ink signature is required; no stamps or "as per" signature)*

As Principal Investigator (Qualified Investigator), I will continue to comply with all relevant regulations and guidelines governing the conduct of research involving human subjects and with the requirements of the REB. I understand that the research study or any change to it cannot be carried out without appropriate written REB approval.

- The appropriate REB-approved consent document was signed by each research subject, parent, tutor, curator or mandatary. A copy of the signed consent document was offered to each subject. If the study is research with a therapeutic component, a copy was sent for filling in the medical record of each research subject.
- Each signed original consent document is on file in a secure location.
- If the study is a clinical trial with a therapeutic component a "Confirmation of Participation in Research" form, and the study's "Executive Summary" was forwarded for inclusion in the medical record of each research subject.
- If the study is a clinical trial with a therapeutic component, a "wallet card" with emergency information was given to each research subject. For a legally incompetent adult or child subject, the card was given to the person signing on behalf of the study subject.
- If the study is a clinical trial with a therapeutic component, certain nominal information concerning the subject's participation was included in the Research Subjects Registry held confidentially by the Investigator, and retained for the length of time required by regulations.

I have reviewed the content of this report, and assure the REB of its accuracy.

Signature: _____ *Date:* Sept 15, 2008

Please Note: "Page 6 of 6" of this Application is for Research Ethics Board "optional use" only.

McGill University Health Centre
Research Ethics Board
Application for Continuing Review

Study Title:

_____

_____

_____

_____

MUHC Study Code: _____ Sponsor Protocol Number: _____ *(as applicable)*

Principal Investigator (Qualified Investigator): _____

MUHC/McGill Department or Division: _____

On behalf of the _____ Research Ethics Board (REB), I confirm that the study identified above was reviewed for continuing ethical acceptability on: ___/___/___ *Day Month Year*

- the following decision was authorized by the REB:
  - ☐ Approval
  - ☐ Approval with Conditions
  - ☐ Continuing Review tabled
  - ☐ Approval disallowed
  - ☐ Other: _____
- the decision was provided following: ☐ Full Board Review ☐ Expedited Review
- the duration of the REB approval is provided for (number) : _____ months *or* ☐ N/A

Name: _____ *REB position:* _____

Signature: _____ *Signed on:* _____

REB approval following Continuing Review may be authorized on this page or by REB correspondence signed by the REB Chair.