Abstract

Computational Methodologies for Transcript Analysis in the Age of Next-Generation DNA Sequencing

Lukas Habegger

2012

A cell's transcriptome is characterized by the full repertoire of its condition-specific transcripts and their respective levels. Deciphering the transcriptome is essential for interpreting the functional elements of the genome, unraveling the molecular constituents of cells, and understanding disease. Furthermore, the emergence of next-generation DNA sequencing has significantly reduced costs, thereby revolutionizing the study of genomes and transcriptomes. These technologies have been leveraged for a number of applications, such as the sequencing of personal genomes on a large scale, which has revealed many novel variants and enabled the analysis of their effects on transcripts. In addition, as applied to transcriptome profiling (RNA-Seq), these technologies have allowed the study of transcripts at an unprecedented level. However, new computational methods are required to take advantage of the burgeoning volumes of data. In this thesis we present four computational approaches for transcript analysis in the context of next-generation DNA sequencing, including: (1) the Variant Annotation Tool, a computational framework to functionally annotate variants and assess their effects on the transcript structure of a gene; (2) RSEQtools, a modular approach for analyzing RNA-Seq data using compact anonymized data summaries; (3) FusionSeq, a tool for identifying fusion transcripts using paired-end RNA-Seq data; and (4) DupSeq, a computational approach for assessing the transcriptional activity of highly similar genomic sequences. Finally, as an application of these methods, we investigate the transcriptome dynamics of human embryonic stem cells as they differentiate into neural precursors. Together, these methodologies have been utilized extensively to gain novel insights into the transcriptome in different biological contexts.

Computational Methodologies for Transcript Analysis in the Age of Next-Generation DNA Sequencing

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

By
Lukas Habegger

Dissertation Director: Mark B. Gerstein

May 2012

Table of Contents

List of Figures and Tables

Acknowledgements

This work would not have been possible without the support and encouragement from a number of people. First and foremost, I would like to thank my thesis advisors Mark Gerstein and Michael Snyder for their mentorship, guidance, financial support, as well as the extraordinary opportunity to pursue a number of exciting research projects in their laboratories.

I would also like to thank Hongyu Zhao and Sherman Weissman for serving on my dissertation committee. They have been a tremendous help in guiding me along the way.

In addition, I would like to acknowledge Mark Rubin for providing me with the opportunity to collaborate on a number of projects.

I would also like to extend my gratitude to my fellow colleagues and everyone in the Gerstein lab, especially Joel Rozowsky, Suganthi Balasubramanian, Raymond Auerbach, Pedro Alves, Declan Clarke, Arif Harmanci, David Chen, Ekta Khurana, Alexej Abyzov, Roger Alexander, Baikang Pei, Cristina Sisu, Jing Leng, Rob Kitchen, Rebecca Robilotto, and Mihali Felipe.

I am also deeply indebted to my former lab colleagues, including Andrea Sboner, Jiang Du, Hugo Lam, Ashish Agarwal, David Koppstein, and Nitin Bhardwaj. Tara Gianoulis, although you are not with us today, I know that your spirit lives on, and I am thankful for having had the opportunity to work with you.

I would also like to thank Lisa Sobel and Anne Nicotra, who have been very supportive and helpful throughout the last five years.

Last but not least, I would like to extend my gratitude to my family, especially my parents and siblings, as well as my wife, Sarah. Thank you for your love, support, and encouragement. This work is dedicated to my wife, Sarah, who has been the bedrock of support throughout these years. Without your love, understanding, and support, this would not have been possible.

Chapter 1

Introduction

A cell's transcriptome is defined as its full set of condition-specific transcripts and their corresponding levels. Deciphering the transcriptome is essential for interpreting the functional elements of the genome, unraveling the molecular constituents of cells, as well as understanding development and disease. A comprehensive understanding of the transcriptome requires the identification of different types of transcripts (including mRNAs, non-coding RNAs, small RNAs, and unannotated transcribed regions), the determination of their transcriptional structure and splicing patterns, as well as the quantification of the varying expression levels associated with each transcript under different conditions [1]. Historically, the transcriptome has been investigated by employing either hybridization- or sequence-based approaches. Although hybridization-based approaches [2–5] (in which a fluorescently labeled cDNA sample is hybridized with probes on a customized microarray or a high-density genome tiling array) can be conducted on a large-scale, these methods have several limitations. Such caveats include the relatively high levels of noise (which result from cross-hybridization) [6, 7], limited dynamic range for detecting low-abundance transcripts, poor resolution, the inability to accurately distinguish between different transcript isoforms, as well as the need for complex normalization methods when comparing measured expression levels across different conditions. However, sequence-based methods employ Sanger sequencing to directly provide the underlying sequence of a cDNA molecule. Initially, cDNA and EST libraries [8, 9] were sequenced using this approach, but it was expensive and relatively low throughput. In order to surmount these limitations, tag-based methods were developed, such as serial analysis of gene expression (SAGE) [10, 11] and massively parallel signature sequencing (MPSS) [12, 13]. While these methodologies were high-

throughput and provided precise measurements of gene expression, they were still expensive to conduct.

However, the advent of next-generation DNA sequencing technologies has revolutionized the ways in which genomes and transcriptomes are studied. Specifically, it has significantly reduced the cost of sequencing and has enabled researchers to address many new questions, which had previously been impossible. The advantages afforded by these new technologies have found a number of applications, including whole genome sequencing, targeted resequencing, transcriptome and transcription factor profiling (RNA-Seq [1] and ChIP-Seq [14], respectively), as well as investigating epigenetic marks and chromatin structure. In particular, efforts aimed at sequencing multiple personal genomes, such as those under way as part of The 1000 Genomes Project [15], have revealed many novel variants, thereby enabling researchers to study their effect on transcripts.

Moreover, the application these technologies to transcriptome profiling has provided many novel insights by enabling researchers to study the transcriptome at single-nucleotide resolution [16–20]. In a typical RNA-Seq experiment, RNA from a population of cells is extracted and then converted to a library of cDNA fragments. Adapters are subsequently ligated to each end of a cDNA fragment, and by obtaining sequence reads from one or both ends of a fragment (single- or paired-end, respectively), these molecules are then subjected to high-throughput sequencing, with or without amplification. The size and error profile of a read depends on the sequencing platform utilized. The resulting collection of reads is then either assembled *de novo*, or in conjunction with a splice junction library, they may be aligned to the reference genome. Computational methods are then utilized to define the structure of the transcripts, identify their respective splicing patterns, and calculate their corresponding expression values at the transcript or gene level.

RNA-Seq technology offers a number of advantages over traditional methods. Firstly, RNA-Seq provides single-base resolution, and the signal obtained tends to be less noisy relative to that obtained by hybridization-based approaches [1, 21, 22]. Secondly, RNA-Seq has a larger

dynamic range, and can be used to identify and quantify low-abundance transcripts without prior knowledge of a particular gene. Lastly, the connectivity information provided by paired-end reads can be used to identify alternatively spliced transcript isoforms [20] and fusion transcripts [23, 24].

Although this revolutionary technology offers many advantages, it also presents a number of bioinformatics challenges. Firstly, as the volume of data generated by a typical run continues to grow at a rapidly, there is an increasing need for more efficient data storage and retrieval. Secondly, new computational methods, which are tailored for specific applications, are required to process and analyze this type of data. Moreover, these computational methods must be implemented in an efficient way in order to deal with burgeoning volumes of data being generated.

Therefore, the core of the thesis work described here focuses on four computational methods for transcript analysis in the context of next-generation DNA sequencing. The first is a computer program that describes a computational framework addressing the intersection of variants with annotated elements. Specifically, we investigate the effect of such variants on the transcript structure of a gene. The remaining three are computer programs designed to address specific questions related to applications of next-generation DNA sequencing to transcriptome profiling (RNA-Seq). In particular, we describe RSEQtools [25], which is a modular framework for analyzing RNA-Seq data using compact anonymized data summaries. The second is called FusionSeq [24], which is an extension of RSEQtools and focuses on the identification of fusion transcripts using paired-end RNA-Seq data. Lastly, we describe DupSeq, which is designed to enable the analysis of transcribed regions that share high sequence similarity with other genomic elements.

In Chapter 2, we describe a computational framework to annotate variants in personal genomes in a cloud-computing environment. A number of large-scale studies, including The 1000 Genomes Project [15], aim to sequence and genotype large numbers of individual genomes using next-generation DNA sequencing technologies. These studies have revealed many novel variants, including single nucleotide polymorphisms (SNPs), small insertions and deletions

(indels), and structural variants (SVs). In order to assess the functional impact of identified variants, a key objective is to determine whether those variants intersect with annotated elements, including both coding and non-coding genes. However, the intersection of variants with a gene annotation set is non-trivial; a number of complicating factors make it difficult to assess the overall functional impact on the structure of a gene [26]. To tackle these issues, we have designed and implemented the Variant Annotation Tool (VAT). By implementing VAT, we also addressed the question on how to best position the software relative to the data. Like VAT, other tools have been implemented to assess the functional impact of variants [27–29]. One issue with these tools is that they do not reside in the same space as the data itself. However, as the volumes of data generated continues to grow, the placement of the software relative to the data becomes increasingly important. Thus, the burgeoning volume of data places growing demands on limited bandwidth, making it impractical for software and data to reside in separate spaces. In order to address these issues, we provide VAT as a virtual machine that can be run within a cloud-computing environment (including that operated by Amazon) to take advantage of the scalability and unlimited storage capacity offered by this framework.

In Chapter 3, we present RSEQtools [25]. As noted previously, the application of next-generation DNA sequencing technologies for functional genomics studies has generated large volumes of sequence information. This deluge of data poses many challenges in terms of data storage, processing, and dissemination, thereby necessitating more efficient algorithms and compact data formats. In addition, with the advent of personal genomics, the sequencing data fundamentally stems from individuals, and new mechanisms for protecting confidential information are thus needed. Along these lines, a pivotal challenge is to devise a new data format that enables the dissemination of large amounts of data without revealing the genotypic information of the underlying individual, while still enabling the community to carry out most functional genomics analyses.

In order to address these issues, we have developed the Mapped Read Format (MRF), which not only facilitates the representation of short and long read alignments, but also allows the

anonymization of confidential sequence information. MRF achieves this by separating the read alignments from the actual sequence reads that contain the polymorphisms. In addition to the data format, RSEQtools comprises a suite of tools that use this format for the analysis of RNA-Seq experiments. Specifically, RSEQtools implements several modules using this standardized format for performing common RNA-Seq analyses, such as calculating gene expression values, generating signal tracks of mapped reads, and discovering novel transcribed regions. Moreover, the modules of RSEQtools can easily be used to build customizable RNA-Seq workflows.

In Chapter 4, we describe FusionSeq [24], a downstream analysis pipeline based on RSEQtools. FusionSeq may be used for finding instances of gene fusions by analyzing paired-end RNA-Seq data, and it comprises three main modules. The first aligns each end of a paired-end read to genome and identifies potential fusion candidates, where each end of the read maps to different genes. The second module employs a sophisticated filtration cascade to remove spurious fusion candidates due to misalignment and random pairing of transcript fragments. In addition, the second module classifies the fusion candidates and ranks them according to several statistics. The third module is then utilized to identify the exact junction sequence surrounding the breakpoint between the two genes. By employing this approach we identified several high quality fusion candidates, which were then experimentally validated.

A key advantage of paired-end RNA-Seq data over information obtained from hybridization-based approaches is that they provide connectivity information required to identify fusion transcripts. Although the role of fusion transcripts is still not fully understood, studies have indicated that they play a role in cancer [30, 31]. In addition, gene fusions may reflect an underlying genomic rearrangement between two genes and are thought to drive molecular events. For instance, in chronic myelogenous leukemia, a gene fusion that originates by the reciprocal translocation between chromosome 9 and 22, results in a chimeric fusion oncogene (*BCR-ABL1*) that constitutively activates a tyrosine kinase [32]. Similarly, another gene fusion, called *TMPRSS2-ERG*, has been reported to play a key role in prostate cancer [33]. Thus, in

order to identity gene fusions, we have developed and applied FusionSeq to eight prostate cancer tissue samples [24].

In Chapter 5, we describe DupSeq, which is a computational framework for assessing the transcriptional activity of highly similar genomic sequences. A key objective in modern genomics is to accurately measure the levels of transcription of any given region in the genome. However, this task is non-trivial, especially for genomic elements that share high degrees of sequence similarity, such as pseudogenes and paralogs. Specifically, it is difficult to discriminate between true transcription and potential artifacts when studying these elements. Such artifacts may result from spillover effects from the expression of a highly transcribed region with a similar sequence. This issue has been studied extensively in the context of hybridization-based methods [6, 34, 35], and although the advent of next-generation DNA sequencing applied to transcriptome profiling has led to several improvements, the issue of accurately measuring the levels of transcription for such elements remains unresolved. The reason for this is that reads from an RNA-Seq experiment must be mapped to the reference genome. Occasionally, sequencing errors cause reads from highly expressed genes to be mistakenly aligned to untranscribed elements with highly similar sequences.

In order to surmount these challenges, we have designed DupSeq as a computational framework that employs statistical methods to compare the transcription signal patterns (as obtained from mapped RNA-Seq reads) across multiple samples. The premise is that when comparing the signal of a given sequence across multiple tissues, truly transcribed regions will be characterized by distinctly different (i.e., independent) expression patterns relative to those observed in regions with high sequence similarity, whereas concordant patterns are suggestive of mapping artifacts.

The implementation of DupSeq is based on three main modules. The first identifies all highly similar regions given specific regions of interest. The second processes the various RNA-Seq data sets by mapping the reads and generating the signal tracks associated with those mapped reads. Lastly, the core module utilizes the output of the two modules described above to

statistically evaluate the transcriptional activity of these particular regions of interest. As proof-of-principle, DupSeq has been employed to discriminate between true transcription and artifacts for both human as well as worm [36] pseudogenes.

In Chapter 6, we apply these computational methods to examine the fundamental mechanisms governing neural differentiation by analyzing the transcriptome dynamics that occur during the differentiation of human embryonic stem cells into the neural lineage [37]. By using RNA-Seq data and integrating the sequence information from various next-generation DNA sequencing platforms, we were able to find many previously unannotated transcripts, as well as distinct transcript isoforms that are expressed at each stage of differentiation. Furthermore, a notable finding was that the splicing isoform diversity decreases as human embryonic stem cells differentiate into neural cells.

In Chapter 7, we conclude the thesis with possible future directions.

Chapter 2

VAT: A computational framework to functionally annotate variants in personal genomes within a cloud-computing environment

Abstract

The functional annotation of variants obtained through sequencing projects is generally assumed to be a simple intersection of genomic coordinates with genomic features. However, complexities arise for several reasons, including the differential effects of a variant on alternatively spliced transcripts, as well as the difficulty in assessing the impact of small insertions/deletions and large structural variants. Taking these factors into consideration, we developed the Variant Annotation Tool (VAT; http://vat.gersteinlab.org/) to functionally annotate variants from multiple personal genomes at the transcript level, as well as obtain summary statistics across genes and individuals. VAT also allows visualization of the effects of different variants, integrates allele frequencies and genotype data from the underlying individuals, and facilitates comparative analysis between different groups of individuals. VAT can either be run through a command-line interface or as a web application. Finally, in order to enable on-demand access, and to minimize unnecessary transfers of large data files, VAT can be run as a virtual machine in a cloud-computing environment.

2.1 Introduction

Recent technological advances have significantly reduced the cost of DNA sequencing and have made it possible to sequence complete human genomes on a large scale. Currently, a number of efforts, including the 1000 Genomes Project, aim to sequence and genotype large numbers of individual genomes [15]. These studies have already revealed many novel single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and structural variants

(SVs). In order to assess the functional impact of identified variants, a key objective is to determine whether those variants intersect with annotated elements, including both coding and non-coding genes. However, the intersection of variants with a gene annotation set is non-trivial [26]. First, a variant may affect only a subset of the possible transcript isoforms of a given gene, or it may have different effects on alternatively spliced transcripts. For example, a variant can affect the coding region of one transcript and overlap the canonical splice site of another. Second, depending on their length, indels in coding regions can either preserve the frame or introduce frameshifts. In addition, indels can partially overlap coding exons, thereby not only affecting the coding region of a transcript, but also impairing its splice sites. Assessing the functional impact in such cases is especially challenging. Lastly, large SVs can have drastic effects on the structure of a gene if exons are removed in whole or in part. As a result, it can be difficult to assess the overall functional impact of different types of variants on gene structures without having visual representations. A summary of the various complicating factors is provided in Figure 2.1.

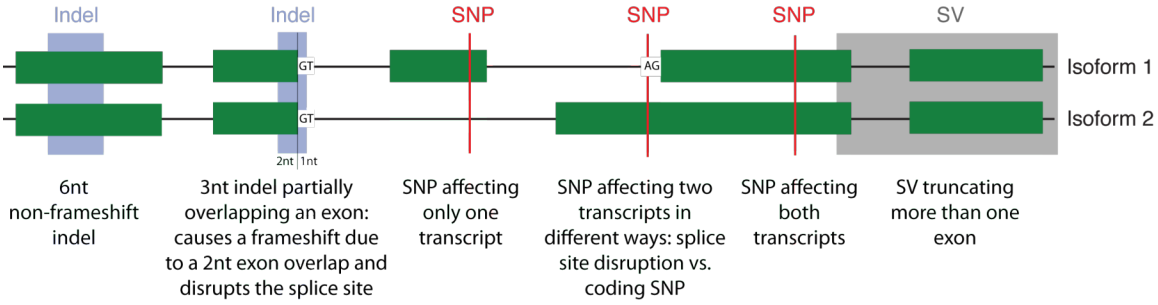


Figure 2.1 Factors that may complicate the functional annotation of variants. As a result of several confounding factors, the functional annotation of variants is non-trivial. A variant may affect all or only a subset of the transcript isoforms of a given gene. Furthermore, a variant may have differential effects on alternatively spliced transcripts. For example, a variant can affect the coding region of one transcript, and overlap with the canonical splice site of another. Depending on their length, indels and SVs can introduce frameshift or non-frameshift variations. In addition, indels and SVs may partially overlap with coding exons, thereby not only affecting the coding region of a transcript, but also impairing the nearby splice site. Lastly, large SVs can have drastic effects on the structure of a gene by potentially removing a number of exons.

To address these issues, we have designed and implemented the Variant Annotation Tool (VAT). Like VAT, other tools have been developed to assess the functional impact of variants [27–29]. One issue with these tools is that they do not reside in the same space as the

data itself. However, as the quantity of data generated by sequencing experiments continues to grow, the positioning of the software relative to the data becomes increasingly important; the burgeoning volume of data places growing demands on limited bandwidth, making it impractical for software and data to be positioned separately, and more advantageous for both to reside in the same space. In order to provide an efficient workflow for annotating variants from large-scale sequencing studies, we have taken great care to position VAT with the data. Specifically, we provide VAT as a virtual machine (VM) that can be run within a cloud-computing environment (including that operated by Amazon) to take advantage of the scalability and unlimited storage capacity offered by this framework. VAT's utility has been demonstrated by its extensive use in annotating the loss-of-function variants obtained as part of the 1,000 Genomes Project [38].

2.2 Features and methods

VAT is implemented in C for efficiency, and consists of a number of modules to pre-process gene annotation sets, intersect variants from multiple individuals with both coding and non-coding genes, generate summary statistics across these individuals and at the single gene level, and provide clear visualization summarizing the functional impact of the annotated variants. The overall workflow is depicted in Figure 2.2A.

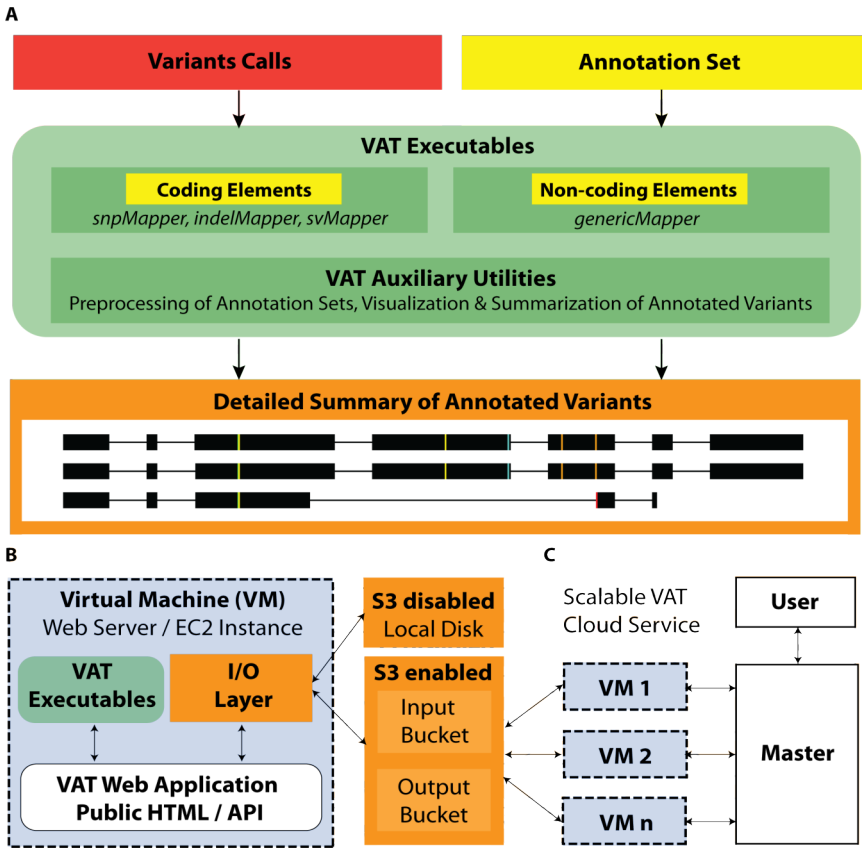


Figure 2.2. Schematic representation of VAT. (A) VAT comprises a number of different modules that relate variants to both protein-coding genes and non-coding elements. These modules use a set of variants and an annotation set as input to generate annotated Variant Call Files (VCFs; [39]). (B) Architecture the VAT web application. The web application may be accessed through the browser or a JSON-based interface. The I/O layer of VAT takes advantage of the Amazon S3 service, and stores all data in S3 buckets or, if S3 support is disabled, simply writes to a local disk. (C) The VAT EC2 cloud service is implemented in a service-oriented architecture consisting of a master node and a number of worker nodes. The master node hosts the user-facing interface and delegates tasks on behalf of the user to the worker nodes.

VAT comprises a number of different modules that relate variants to both protein-coding genes (*snpMapper*, *indelMapper*, and *svMapper*) and non-coding elements (*genericMapper*). These four core modules use an annotation set and a set of variants from multiple individuals as inputs. The variants are typically represented using the Variant Call Format (VCF; [39]). A key feature of VAT is that the annotation task is performed at the transcript level to determine whether all or only a subset of the transcript isoforms of a gene are affected. Therefore, the output of these programs explicitly shows which transcript isoforms are affected by each variant and provides detailed information about the location of a given variant within a transcript, as well as the variant's effect on the coding potential of that transcript. Moreover, VAT can be executed

using various gene annotation sets and genome builds. Preprocessing of gene sets using auxiliary modules in VAT provides the end user with more options and flexibility for variant annotation.

The VAT software package contains a number of utilities for performing downstream analyses on functionally annotated variants. For instance, an auxiliary module generates detailed summaries of the annotated variants across multiple individuals and at the level of single genes. For those variants intersecting protein-coding genes, VAT includes a module for generating an image for each gene in order to provide the user with a clear overview. Specifically, this schematic representation displays the various transcript isoforms of a gene, which are then superimposed with the annotated variants (Figure 2.2A).

As shown in Figure 2.2B, VAT is built to take advantage of the Amazon Web Services (AWS) cloud-computing platform. In addition, VAT can also be deployed as a VM on a private cloud. Each installation consists of the command line executable of the VAT pipeline and a PHP web application. The web application serves as the user interface and driver for the VAT pipeline, and it may be accessed through the browser or a JSON-based RESTful API. The VAT I/O abstraction layer may be customized using the configuration file to take advantage of Amazon's Simple Storage Service (S3), which allows for high availability, reliability, and large storage capacity. With S3 support enabled, VAT reads input files from a bucket storing raw VCF files, and then stores output data sets in a separate bucket. With S3 support disabled, VAT simply reads from and writes to a local disk. Our architecture may also be easily scaled to utilize more sophisticated storage schemes, such as hashing across multiple input and output buckets.

Our hosted VAT cloud service takes advantage of the scalability and reliability of Amazon's Elastic Compute Cloud (EC2) distributed computing platform. Our service is implemented in a service-oriented architecture consisting of a master node and a number of worker nodes. Each node consists of a customized VAT installation running on an EC2 VM (Figure 2.2C). The master node hosts user-facing web components and serves as a load-balancer for the worker nodes. A user action is forwarded by the master node as a request to

one of the worker nodes. Each of the worker nodes communicates with our S3 buckets and reports updates to the master node asynchronously. Furthermore, we take advantage of Amazon's EC2 API to allow the master node to dynamically create additional worker instances. Thus, intensive batch requests and increased traffic may be parallelized and handled efficiently. Finally, our collection of S3 buckets accumulates a growing data set that may be made available to aid in further analyses.

2.3 Conclusions

In summary, VAT offers two primary advantages in variant annotation. First, it addresses the complicating factors frequently involved in variant annotation and visualization. Second, by virtue of operating as a VM in a cloud-computing environment, VAT has direct access to the data on which it operates, and is capable of leveraging unlimited storage capacity and scalability.

2.4 Design and implementation

2.4.1 Pre-processing of the annotation set

The current implementation of VAT uses the GENCODE [40] gene annotation set, although alternative annotation sets can easily be adapted. The goal of the GENCODE project is to accurately annotate all evidence-based features in the human genome, as determined by manual curation, a variety of computational analyses, and targeted experimental approaches [40]. The GENCODE annotation files are updated on a regular basis, and VAT provides the functionality to efficiently parse these files. The raw annotation files, which contain the coordinates of all gene models (including different transcript isoforms) are typically stored in GTF format, and may be downloaded from http://www.gencodegenes.org/.

The variant annotation modules require two inputs obtained from the annotation file. One file provides the underlying coordinates of the coding sequences from all gene models in Interval format (see http://vat.gersteinlab.org/documentation.php for details), whereas the second stores the transcript sequences of all transcripts in FASTA format. The VAT framework includes two

modules to facilitate pre-processing of the raw annotation file. The first module, *gencode2interval*, converts the coordinates of the raw GTF file into Interval format. The second module, *interval2sequences*, extracts the underlying transcript sequences for each entry in the Interval file.

2.4.2 Overview of the annotation modules

VAT comprises a number of different modules that relate variants to both protein-coding genes (*snpMapper*, *indelMapper*, and *svMapper*) and non-coding elements (*genericMapper*). These four core modules use a set of variants from multiple individuals and an annotation set as input. The variants are typically represented using the Variant Call Format (VCF), which was originally developed as part of the 1000 Genomes Project [15, 39], and is also being adapted for cancer variants. A VCF file is a tab-delimited text file that stores information about a variety of variant types, including SNPs, indels, and SVs. Unlike other tools, VAT annotation modules directly capture the annotation information within the VCF file without necessitating a new file format (see http://vat.gersteinlab.org/documentation.php for details). As mentioned, a key feature is that the annotation is performed at the transcript level to determine whether all or only a subset of the transcript isoforms of a given gene are affected. This distinction is important, given that some transcript isoforms are expressed in a tissue-specific manner. Therefore, the output of these core modules explicitly shows which transcript isoforms are affected by each variant. In addition, the output contains detailed information about the location of a given variant within a transcript, and the output from *snpMapper* and *indelMapper* also includes the variant's effect on the coding potential of a transcript. For instance, an annotated SNP may be classified as synonymous, non-synonymous, premature stop, removed stop, or as a splice overlap. On the other hand, indels are grouped into the following categories: frameshift insertion, non-frameshift insertion, frameshift deletion, non-frameshift deletion, start overlap, end overlap, or splice overlap. The term "splice overlap" (as used when describing both SNPs and indels) refers to a variant that overlaps with a canonical splice site (either two nucleotides downstream of an exon or two

nucleotides upstream of an exon). Lastly, for those SVs, which overlap with a transcript, the program calculates the number of overlapping nucleotides.

2.4.3 Generating summaries of functionally annotated variants

The VAT software package contains a number of utilities for performing downstream analyses on functionally annotated variants. One program, *vcfSummary*, creates summary statistics across different individuals and genes. Specifically, it tabulates the different types of variants for each individual and for each gene. The resulting tab-delimited files can then easily be accessed as spreadsheets for further analyses.

2.4.4 Visualization of functionally annotated variants in coding regions

The VAT software package includes *vcf2images*, a program to visualize the impact of different functionally annotated variants on the various transcript isoforms of a protein-coding gene. In particular, this program generates a schematic representation of the exon/intron structure of all transcript isoforms of a gene. The different types of functionally annotated variants are subsequently superimposed on the transcript structures to provide the user with an overview of the various types of variants that affect a particular gene, along with their respective locations along transcripts. In order to clearly visualize coding exon variants, introns are represented as small, fixed-length bars. Thus, introns only account for a small portion of the overall display, while the coding regions are emphasized. The exons are scaled according to their lengths. The main advantage of displaying the variants in this way is that it enables the user to more easily discern the overall impact of the variants on a gene model, in contrast to the displays conventionally provided in genome browsers, in which introns and exons are shown on the same scale (Figure 2.3).

Figure 2.3 Comparison between different visualization methods. This figure shows a gene, *MAST2*, and its associated variants using three different visualization methods: the rendering tool as part of VAT, the UCSC genome browser [41] and the Ensembl genome browser [42]. The top panel (VAT) shows variants which were obtained from the 1000 Genomes Pilot Project [15]. In this representation, introns are rendered at a fixed size, whereas exons are scaled according to their length. The middle panel shows a screenshot of the same gene as represented in the UCSC genome browser [41]. The bottom track in this panel represents SNPs from one individual that was sequenced and genotyped as part of the 1000 Genomes Pilot Project [15]. The lower panel shows a screenshot of the same gene in the Ensembl browser utilizing the Variant Image Viewer. This representation shows the gene models of *MAST2* at the original scale, and also highlights the exonic regions containing variants. This detailed view contains in-depth information about the variants obtained from various sources as well as their effects on coding potential.

2.4.5 Comparative analyses of different data sets

One principal feature of VAT is that it enables users to compare functionally annotated variants across different samples. For instance, samples from three different populations (CEU, CHBJPT, YRI) were sequenced and genotyped as part of the pilot phase of the 1000 Genomes Project [15]. Thus, in this context, a key question is whether the allele frequency of a particular variant differs across the three populations.

In order to address this issue, the VAT package includes *vcfModifyHeader*, a program which modifies the header line in a VCF file to assign each sample to a particular group prior to the annotation step. In the case of the 1000 Genomes Project's pilot phase, each group

represents a different population. This information is subsequently utilized when the results are presented in a web-interface in order to calculate allele frequencies, as well as display the genotype information in a population-specific manner. This concept may further be generalized to clinical re-sequencing projects, in which there is a need to compare cases and controls, as well as to analyze differences between patients.

2.4.6 VAT web application

VAT is designed to take advantage of the Amazon Web Services (AWS; http://aws.amazon.com/) cloud-computing platform. An individual VAT node consists of the C executables comprising the VAT pipeline and the PHP scripts implementing the VAT web application. VAT requires PHP version 5.2 or later. Due to the large sizes of the files being uploaded to be processed by VAT, the user must configure PHP to allow larger upload files by editing `php.ini` and setting `upload_max_filesize` and `post_max_size` to at least 100M. It is also recommended the user turns off output buffering so that `flush()`, which is used by the processing page to update the user, works properly by setting `output_buffering = off`.

The `.vatrc` configuration file, which resides in both the installing user's home directory and the root directory for the web application (as `vat.conf`), may reconfigured without recompilation. The `AWS_USE_S3` configuration directive is set to true to turn on support for Amazon's Simple Storage Service (S3; http://aws.amazon.com/s3/) or false to disable S3 support. To enable S3 support, configure the `AWS_ACCESS_KEY_ID`, `AWS_SECRET_ACCESS_KEY`, and the `AWS_S3_HOSTNAME` directives with the AWS credentials and the S3 hostname, respectively. If S3 support is enabled, file storage is fully backed by S3. Raw VCF files are stored in the S3 bucket specified by the `AWS_S3_RAW_BUCKET` directive, and processed data sets are stored in the bucket specified by `AWS_S3_DATA_BUCKET`. Set `WEB_DATA_URL` to the URL of the data bucket and ensure that the data bucket may be accessed in a browser, as the images generated by VAT and stored in the data bucket are accessed by the URL to be embedded in the results page.

Locally, the data directory under the web root contains the directory tree used by the VAT I/O layer. Regardless of whether S3 support is enabled or disabled, all reference annotation files must be stored in the directory to which `WEB_DATA_REFERENCE_DIR` points (this is set to data/reference by default). The script *get_annotation_sets.sh* may be used to download all the annotation files from our servers using `wget`. Also, the I/O layer uses the directory pointed to by `WEB_DATA_WORKING_DIR`, which is web/working/ to contain the temporary working directories for process instances of VAT. Each instance is given its own unique working directory and hence its own isolated copy of files.

If S3 support is disabled, the S3-specific directives are ignored, and the I/O layer simply stores and retrieves files from local directories. However, the options `WEB_DATA_DIR` and `WEB_DATA_RAW_DIR` must be configured to point to directories used to store processed data sets and raw VCF files respectively. With S3 support disabled, set `WEB_DATA_URL` to the URL of the data directory.

2.4.7 Hosted VAT service and cluster setup

The architecture of the hosted VAT service consists of a master node or load balancer and a set of worker nodes. The web PHP components on the master node are specially adapted for this architecture. Rather than calling local VAT command-line web applications, the master node delegates tasks to worker nodes via requests made through the RESTful API. Load balancing is done by LRU (least-recently used). The information for each server is stored as a JSON string in a local text file on the master that is exclusively locked by a process before reading and updating the server load information. The master also keeps track of information for each data set in a MySQL database. Furthermore, the master implements an atomic fetch-and-increment counter using the ID primary key field, which is specified as `AUTO_INCREMENT`, and assigns the new data set a unique ID using the insertion ID.

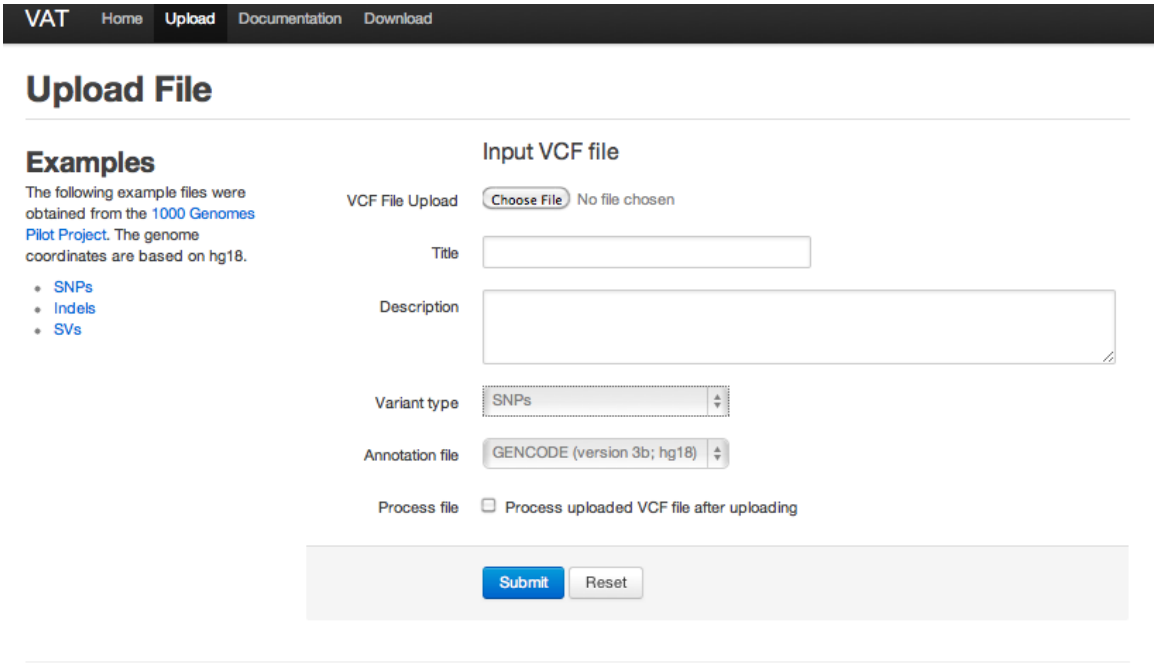


Figure 2.4 Screenshot of the VCF file upload web interface. This web interface enables users to upload their own variants (SNPs, Indels, SVs) for annotation.

As an example of a typical run, a user uploads a VCF file to be processed as shown in Figure 2.4. Upon completion of the upload, the master saves metadata in the database, assigns the data set a unique ID, and transfers the file into a S3 bucket used to store raw VCF files. The master then selects a worker and performs a RESTful API call to the worker to request the worker to download the VCF file to its local cache and forks a background process that performs the six processing steps of the VAT pipeline including intersecting the variant calls in the input file with the reference annotation, compressing and indexing the output, generating summary files for the data set, generating images to visualize the data, and generating subsets of the variants by gene. Because each step takes a few seconds to complete, the background process on the worker also updates the master asynchronously at the end of each processing step by making an API call to the master, which then updates the status for the data set in the database. At the same time, client-side JavaScript code also makes JSON AJAX calls to the master to retrieve the status of the processing in order to update the user interface in real time. Once the processing is

complete, the user may proceed to view the data. Again, rather than calling local command line programs, the master retrieves the necessary information through API calls to workers.

2.4.8 Analysis of the 1000 Genomes pilot data set

As a proof-of-principle, we have applied VAT to annotate the variants identified as part of the 1000 Genomes Pilot Project [15]. We downloaded SNP call sets for three populations (CEU, JPTCHB, and YRI) from their web page. These calls sets were distributed using the VCF format and they contain detailed genotype information on the underlying individuals from these three populations. We subsequently modified the VCF header line in these files to encode the population origin within each VCF file using *vcfModifyHeader*. This step is important because VAT utilizes this information to calculate allele frequencies and to display the genotype information in a population-specific manner. Next, we merged the SNP calls using the *merge-vcf* tool as part of the VCFtools package [39]. We then ran *snpMapper* on the merged SNP set using the GENCODE 3b annotation set. Subsequently, we ran *vcfSummary* to tabulate the functionally annotated variants in a gene- and sample-specific manner. In addition, we generated an image for each gene model containing at least one functionally annotated variant using the *vcf2images* program. Lastly, we visualized and distributed the results using our web service.

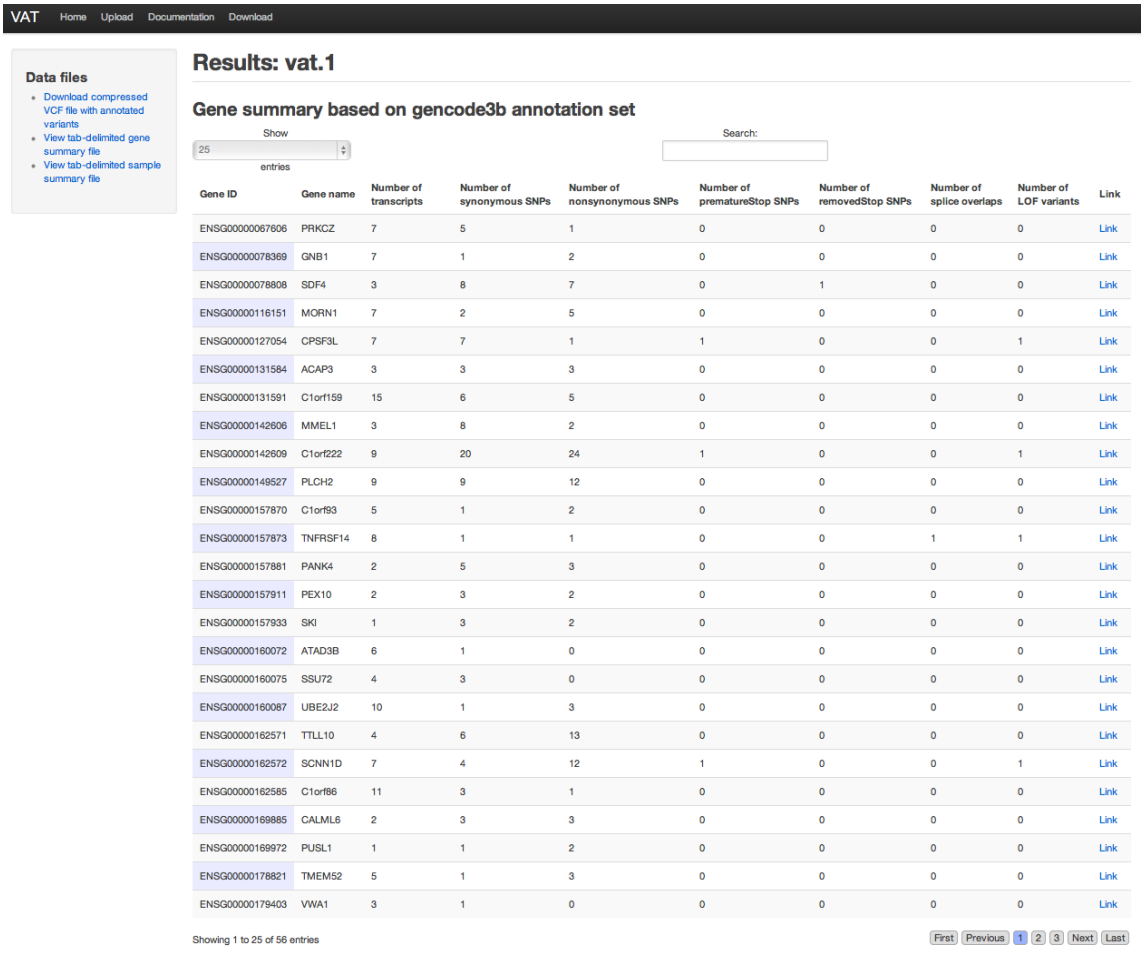


Figure 2.5 Screenshot of the gene summary table. This summary table contains genes that intersect at least one variant.

A sample of the resulting gene summary table is shown in Figure 2.5. Specifically, a gene is listed in this table if it has at least one annotated SNP. Each row, which represents a gene, contains the gene ID, the gene name, and counts for the different types of variants. It should be noted that the columns of this table can be sorted by one or more columns. The upper-right corner of the table provides a search box which enables the user to quickly find a gene of interest. The bottom of the summary web page also contains two links, one to the VCF file with the annotated variants, and the second to a tab-delimited file with the gene summary table. Lastly, VAT provides the user with a unique way to further explore the impact of different variants at the level of individual genes.

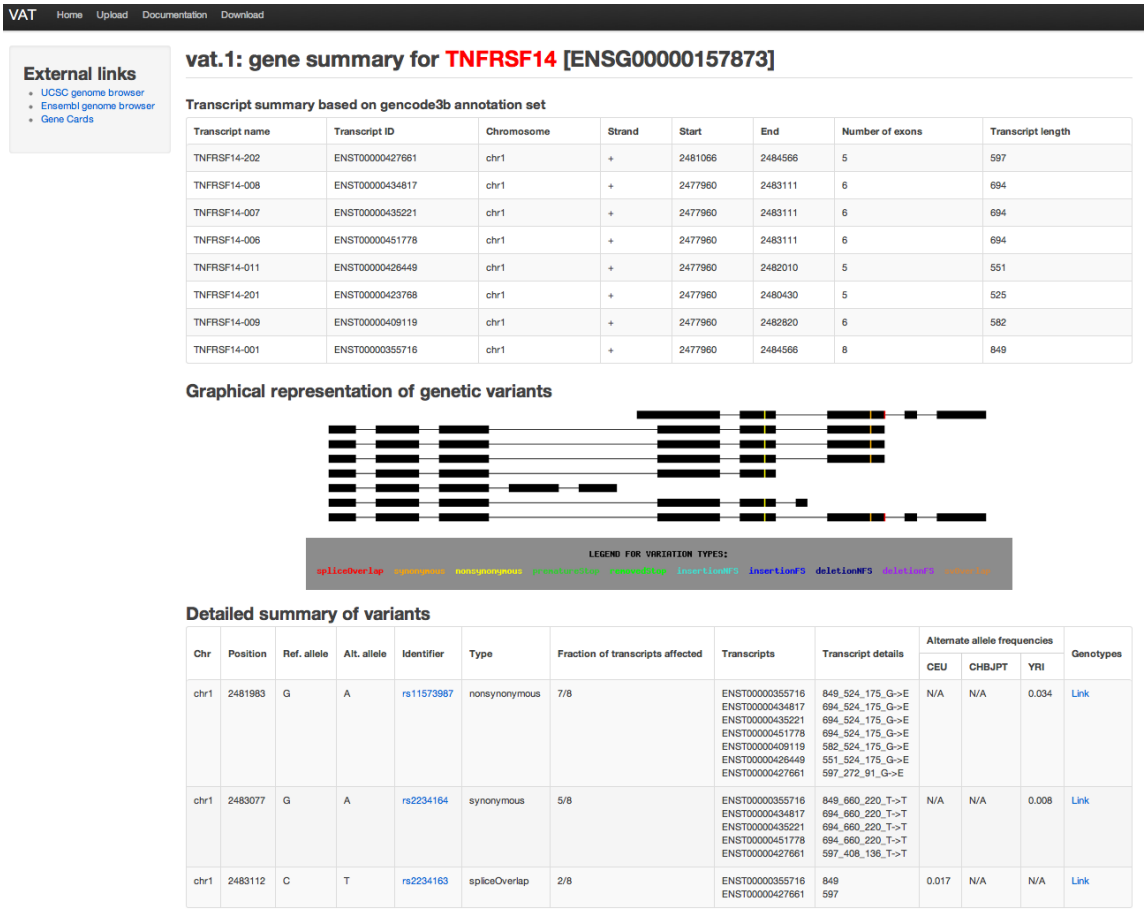


Figure 2.6 Screenshot of the gene-specific view. This page contains detailed information about the variants intersecting various transcript isoforms of a particular gene (*TNFRSF14*).

An example of a gene-specific view is shown in Figure 2.6. This view comprises four parts. The first contains external links to the UCSC genome browser [41], the Ensembl genome browser [42], and the Gene Cards [43] website. The second part provides a detailed summary of the transcript isoforms of the specific gene. The third part provides a graphical representation of the functionally annotated variants, and shows how these variants affect the gene's different transcript isoforms. The fourth part contains a detailed summary table for each variant, including the genomic coordinates of the variant, the reference as well as the alternate alleles, external links to dbSNP [44], and the variant type. This table also contains detailed information on the number of transcripts of the gene that are affected by a particular variant, their transcript IDs, and

the relative position of the variant. The table includes allele frequencies for the different populations (CEU, CHBJPT, and YRI) and a link to the specific genotypes of the underlying individuals.

Chapter 3

RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries

The work described in this chapter was adapted from a manuscript, which was originally published in *Bioinformatics* [25].

Abstract

The advent of next-generation sequencing for functional genomics has given rise to quantities of sequence information that are often so large that they are difficult to handle. Moreover, sequence reads from a specific individual can contain sufficient information to potentially identify and genetically characterize that person, raising privacy concerns. In order to address these issues we have developed the Mapped Read Format (MRF), a compact data summary format for both short and long read alignments that enables the anonymization of confidential sequence information, while allowing one to still carry out many functional genomics studies. We have developed a suite of tools that use this format for the analysis of RNA-Seq experiments. RSEQtools consists of a set of modules that perform common tasks such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. Moreover, these tools can readily be used to build customizable RNA-Seq workflows. In addition to the anonymization afforded by this format it also facilitates the decoupling of the alignment of reads from downstream analyses.

3.1 Introduction

The advent of next-generation sequencing technologies has revolutionized the study of genomes and transcriptomes. In particular, the application of deep sequencing approaches to transcriptome profiling (RNA-Seq) is increasingly becoming the method of choice for studying the transcriptional landscape of cells [1, 18, 45]. Typically, the first step in this analysis is the alignment of the sequence reads to a reference sequence set. Recently, a number of different alignment tools have been developed to map short reads in an efficient manner [46]. While much progress has been made on this front, there is still a great need for a set of software tools that facilitate the downstream analysis of mapped RNA-Seq reads.

Further, two other issues remain to be addressed. First, the immense file size of next generation sequencing data poses many challenges in terms of data processing, storage, and sharing. Secondly, mechanisms to protect personal confidential genetic information need to be established. With the birth of personal genomics, sequencing data stems fundamentally from individuals, and this type of data cannot be distributed as easily because significant privacy concerns arise with sharing all the sequence variations of a particular individual [47, 48]. One critical challenge for genomics, then, is to devise new data summaries that allow the sharing of large amounts of information from sequencing experiments without exposing the genotypic information of the underlying individual.

Although many data formats have been developed such as SAM [49], there is no practical solution yet that addresses the privacy concerns when sharing large sequence alignment files. Addressing this challenge is precisely what we have endeavored to do in putting together the Mapped Read Format (MRF), a format that allows data summaries to be exchanged, enabling many aspects of the RNA-Seq calculation to be performed such as expression measurements, but that also detaches the actual sequence variation in a person into separate files. Further, it provides a very clear way of linking these two pieces of information so that the data summaries can be subsequently conjoined back to the original sequences for more in-depth analyses with potentially confidential data.

Specifically, we developed MRF in the context of RNA-Seq to provide a mechanism to protect the private genotypic information of the underlying individual and to represent the mapped reads in a compact manner. It is important to note that a RNA-Seq experiment with sufficient depth of sequencing is essentially equivalent to an exome sequencing study and thus provides immediate access to the genotypic information of the underlying individual. Thus, in the era of personal genomics it is crucial to protect this private information, while still providing the scientific community a way to access the experimental data without revealing this confidential information. MRF achieves this goal by separating the read alignments from the actual sequence reads that contain the polymorphisms. Thus, the content of the "public" read alignment file is comparable to the information obtained from a standard expression array, which are openly available through multiple data repositories such as Gene Expression Omnibus [50] or ArrayExpress [51]. Also, note that the levels of gene expression are potentially able to identify a person, but they do not genetically characterize someone. Although this approach removes the most obvious genotypic information, other characteristics such as structural variants potentially remain. However, it is not trivial to extract genotypic information from a subset of structural variations. In addition, inferring structural variations from RNA-Seq data as opposed to DNA sequencing would be more complicated due to the presence of alternative splicing and uneven coverage determined by the expression level of the underlying gene.

Here we present an overview of a flexible suite of tools (RSEQtools) that are designed to facilitate easily customizable workflows and efficient pipeline building for the analysis of RNA-Seq experiments using this compact format (Figure 3.1). Briefly, we first convert the aligned reads into MRF and thus decouple the alignment step from the downstream analyses. RSEQtools implements several modules using this standardized format for performing common RNA-Seq analyses, such as expression quantification, discovery of transcribed regions, coverage computations, annotation manipulation, etc.

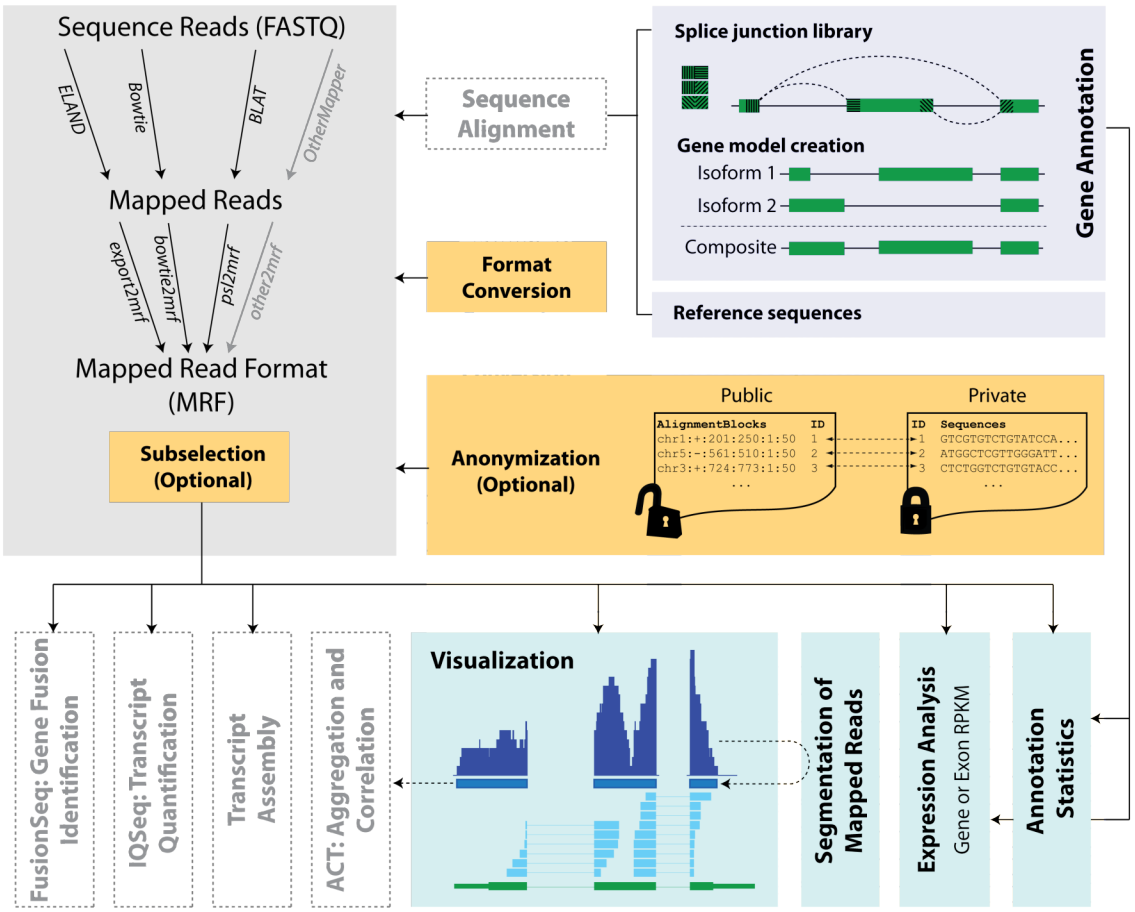


Figure 3.1 Schematic overview of RSEQtools. Mapped reads are first converted into MRF from common alignment tool output formats, including SAM. The resulting MRF files can be divided in two files: one with the alignment only, and another with the corresponding sequence reads. The read identifiers provide a mapping between the two files. Then, several modules perform the downstream analyses independently from the mapping step, such as expression quantification, visualization of the mapped read, and the calculation of annotation statistics, etc. Other tools have been developed based on this framework to perform more sophisticated analyses such as transcript assembly, isoform quantification, fusion transcript identification, as well as aggregation and correlation of signal tracks (described elsewhere).

3.2 Features and Methods

3.2.1 Mapped Read Format (MRF) and converters

MRF only stores a minimal set of information, i.e. information that cannot be derived from the MRF data itself. This has the advantage of keeping the format succinct, while still capturing the relevant information for most analyses. MRF consists of three components: comment lines (optional) denoted by a leading '#' sign, a header line, and the mapped reads. The header line specifies the data type of each column: AlignmentBlocks, Sequences, QualityScores, and

QueryID. The column type AlignmentBlocks is required and represents the mapped reads. Each alignment block contains the coordinates with respect to the reference genome to which the read aligns as well as the read coordinates. A read spanning multiple regions, e.g. multiple exons, is denoted by multiple alignment blocks that are separated by a comma. Paired-end reads can be represented by using a set of alignment blocks for each end, which are separated by the 'l' symbol. By using this format, it is straightforward to specify both gapped and paired-end alignments. The RSEQtools package includes various utilities to convert the output of several mapping tools into MRF. A converter for the commonly used SAM format is included as well. The first example below represents two paired-end reads where one end is spliced, whereas the second example shows two unspliced single-end reads with their associated QueryIDs:

```
# Example 1
AlignmentBlocks
chr2:+:601:630:1:30,chr2:+:921:940:31:50|chr2:+:1401:1450:1:50
chr9:+:451:460:1:10,chr9:+:831:870:11:50|chr9:+:945:994:1:50

# Example 2
AlignmentBlocks          QueryID
chr4:-:1221:1270:1:50    1
chr16:+:511:560:1:50     2
```

The optional types Sequences, QualityScores, and QueryID provide additional information. In particular, the confidentiality issues can be addressed by generating two files: one including the alignments and a second one containing the sequences such as a FASTQ file. The former is useful for most analyses and can be publicly shared because it does not contain confidential information, whereas the latter can be subjected to a higher level of security and control. The two files can be conjoined, if necessary, by using the common QueryID as shown in Figure 3.1.

3.2.2 RNA-Seq analysis with RSEQtools

The RSEQtools suite contains a set of modules to perform a large variety of tasks including the quantification of expression values, manipulation of gene annotation sets,

visualization of the mapped reads, generation of signal tracks, the identification of transcriptional active regions, and several auxiliary utilities as shown in Table 3.1.

Description	Programs

Format conversion utilities, conversion of mapped reads into standardized MRF. The purpose of this conversion step is to decouple the alignment step from the various downstream analyses.	`bowtie2mrf`
	`psl2mrf`
	`export2mrf`
	`sam2mrf`

Genome annotation tools, utilities to manipulate gene annotation sets or to retrieve the genomic/exonic sequences from a given annotation set.	`createSpliceJunctionLibrary`
	`mergeTranscripts`
	`interval2sequences`

Expression analysis, estimation of expression levels (this analysis can be performed at the exon or at the gene level).	`mrfQuantifier` `bgrQuantifier`

Visualization tools, conversion of MRF into common formats for visualization. The WIG and BGR (bedGraph) formats are generally used to represent a signal track of mapped reads whereas the GFF format is utilized to depict splice junction reads.	`mrf2wig`
	`mrf2bgr`
	`mrf2gff`

Segmentation of mapped reads, identification of transcribed active regions (TARs). This analysis is particularly helpful in discovering TARs that are not part of an annotation set.	`wigSegmenter` `bgrSegmenter`
Annotation statistics tools, computation of statistics related to the annotation set used and the mapped reads.	`mrfAnnotationCoverage` `mrfMappingBias`

MRF selection utilities, utilities to sub-select alignment blocks from a MRF flat file based on various criteria.	`mrfSampler`
	`mrfSelectRegion`
	`mrfSelectAnnotated`
	`mrfSelectSpliced`
	`mrfCountRegion`

Auxiliary utilities, additional utilities to convert data into different formats.	`bed2interval`
	`interval2bed`
	`gff2interval`
	`interval2gff`
	`export2fastq`
	`mrf2sam`

Table 3.1 List of RSEQtools modules.

Genome annotation tools. To generate a splice junction library from any annotation set, we extract the genomic sequences of all the exons and synthetically create all splice junctions specified in the annotation set. This splice junction library can be used in combination with the reference sequences. A second tool is particularly useful when estimating expression levels. In order to capture the information of the various transcript isoforms, a “gene model” is required.

The module *mergeTranscripts* collapses the transcript isoforms into a single gene model by either taking the union or intersection of the exonic nucleotides.

Quantification of gene expression. One of the key features of RNA-Seq is the quantification of expression at different levels. Hence, a key module calculates the gene expression values for a given annotation set and a collection of mapped reads in MRF format. The annotation set specifies which "elements" will be quantified. The program *mrfQuantifier* calculates RPKM (reads per kilobase per million mapped reads) values at the nucleotide level [18]. Briefly, for a given entry in the annotation set (typically an exon or gene model) the number of nucleotides from all the reads that overlap with this annotation entry are added up and then this count is normalized by sequence length of the annotation entry (per kb) and by the total number of mapped nucleotides (per million). This calculation is not performed at the transcript level, which requires a more sophisticated analysis [52–54].

Visualization of mapped reads. The RSEQtools package also contains various tools for visualizing the results in genome browsers, by means of wiggle (WIG) and bedGraph files, which are commonly used to represent a signal track of mapped reads. Also, a GFF file can be generated from MRF files to visualize splice junction reads (example in Figure 3.1).

Identification of transcriptionally active regions (TARs). Transcribed regions can be identified de novo by performing a max-Gap/minRun segmentation [35, 55] from the signal files using the *wigSegmenter* program. Briefly, the signal is first thresholded to identify transcribed elements. Contiguous elements whose distance is less than "max-Gap" are joined together and then filtered if the final size is less than "minRun". This type of analysis is particularly useful in discovering novel TARs such as small RNAs, etc.

MRF selection and auxiliary utilities. Lastly, RSEQtools includes a set of utilities to easily manipulate MRF files and a collection of format conversion tools allowing for rapid pipeline development.

Implementation and run time. The modules of the RSEQtools suite were implemented in C and the code was optimized in order to efficiently handle large data sets. The importance of

code scalability cannot be overemphasized in a time where data sets become increasingly large and easily exceed several gigabytes. For example, the conversion of an ELAND export file (uncompressed file size: ~4GB; total number of reads: ~20 million; number of mapped reads: ~12 million) to MRF takes approximately two minutes and the resulting MRF file is significantly smaller (~400MB uncompressed, ~130MB compressed with *gzip*). Converting the same ELAND export file to SAM generates a file of ~3.1GB (uncompressed) and the corresponding BAM file has a size of ~1.2GB. The subsequent quantification of gene expression using *mrfQuantifier* requires 45s to calculate estimates for about 20,000 genes.

In addition, the modularity of RSEQtools also enables the development of additional programs in any programming language and their seamless integration into this framework. Finally, most modules use STDIN and STDOUT to process the data, making them suitable to be integrated into an automated pipeline. For instance, we have developed three downstream analysis pipelines that are based on RSEQtools: FusionSeq [24], a computational framework for identifying fusion transcripts (described in Chapter 4); IQSeq [54], a method for transcript isoform quantification; and DupSeq, an algorithm for analyzing transcribed regions with high sequence similarity to other regions in the genome (described in Chapter 5).

3.3 Conclusions

In summary, RSEQtools contains a number of useful and highly specific modules that can rapidly analyze RNA-Seq data. The MRF format has two major features: it allows the decoupling of downstream analysis from the mapping strategy and addresses the issue of confidentiality that is intrinsic in any sequencing experiments involving human subjects. By separating the actual sequencing reads from the alignments MRF provides a mechanism to protect the private genotypic information of the underlying individual. Although this approach removes the most obvious genotypic features, other distinctive attributes do remain. First of all, the information in a MRF file is at least equivalent to that in traditional expression array, which can potentially identify the underlying individual. Secondly, some information about structural variants may be contained

in the MRF file of an RNA-Seq experiment. However, it is not obvious how to extract genotypic information from a subset of structural variations just affecting genes. In addition, inferring structural variations from RNA-Seq data as opposed to DNA sequencing would be more complicated due to the presence of alternative splicing.

Another advantage of storing the alignments without the underlying sequences is that it saves space, especially as reads become longer. Moreover, a possible future extension is the development of a specific compression schema that could further reduce the size of the files. In addition, this data format could be easily applied to sequence alignments obtained from other high-throughput functional genomic assays such as ChIP-Seq or chromosome conformation capture (3C).

Chapter 4

FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data

The work described in this chapter was adapted from a manuscript, which was originally published in *Genome Biology* [24].

Abstract

We have developed FusionSeq to identify fusion transcripts from paired-end RNA-sequencing. FusionSeq includes filters to remove spurious candidate fusions with artifacts, such as misalignment or random pairing of transcript fragments, and it ranks candidates according to several statistics. It also has a module to identify exact sequences at breakpoint junctions. FusionSeq detected known and novel fusions in a specially sequenced calibration data set, including eight cancers with and without known rearrangements.

4.1 Introduction

Deep sequencing approaches applied to transcriptome profiling (RNA-Seq) are dramatically impacting our understanding of the extent and complexity of eukaryotic transcription [1, 16, 18, 45]. RNA-Seq provides a more accurate measurement of expression levels of genes and more information about alternative splicing of their isoforms as compared to other chip-based methods [16, 20, 21, 45, 56–59].

Large international consortia such as the ENCODE project [60] and the modENCODE project [61] are exploiting this technology to obtain a better picture of the transcriptome. More recently, RNA-Seq was applied to the identification of fusion transcripts, where mRNAs from two

different genes are joined together [23, 62–65]. Although the role of these chimeric transcripts is not fully understood, some studies have shown that they might be implicated in cancer [30, 31]. Also, a fusion transcript may indicate an underlying genomic rearrangement between the two genes. Such gene fusions are thought to be driving molecular events such as in chronic myelogenous leukemia (CML), which is defined by the reciprocal translocation between chromosome 9 and 22 leading to a chimeric fusion oncogene (*BCR-ABL1*) encoding a tyrosine kinase that is constitutively active.

The majority of gene fusions reported in the past have been attributed to hematological cancers [66–68]. Recently, recurrent fusions between the transmembrane protease serine 2 (*TMPRSS2*) gene and members of the ETS family of transcription factors (mainly the v-ets erythroblastosis virus E26 oncogene homolog (avian), *ERG* and the ets variant 1, *ETV1*) were reported in prostate cancer [33]. Other epithelial tumors such as lung and breast cancer also harbor translocations [69–71].

Compared to DNA sequencing, RNA-Seq seems to have less requirements in terms of overall coverage, since it aims at sequencing only the regions of the genome that are transcribed and spliced into mature mRNA, which current estimates set at about 2 to 6%. However, this apparent advantage of RNA-Seq in practice is not so straightforward. Indeed, determining the depth of sequencing needed to completely assess the extent of transcription in complex organisms is complicated by the high dynamic range of gene expression, the presence of alternatively spliced transcripts, and the biological condition of the transcriptome, that is, cell types or environmental conditions [1].

RNA-Seq can be used effectively to detect fusion transcripts. Maher et al. discovered novel fusion transcripts using single-end reads of various lengths [62]. This approach nominated multiple candidates such as *SLC45A3-ELK4*, which was independently confirmed as a common “read-through” transcript identified in prostate cancer (i.e. fusion transcripts resulting by two nearby genes without any genomic rearrangement [31]). This and other non-genomic events of adjacent or neighbouring genes appear to be common. Maher et al. showed in principle how to

use RNA-Seq for discovering fusion transcripts. They used two single-end sequencing platforms, which is rather infeasible in terms of both cost and labor efforts [62]. Since then, Paired-End (PE) RNA-Seq has been introduced and has received broader attention for transcriptome profiling bringing with it a great potential to accelerate fusion discoveries [23, 63].

The concept of sequencing both ends of a fragment, either cDNA or genomic DNA, was introduced in the context of the identification of structural variants (SVs) [72–75]. Such events are among the basic mechanisms generating fusion transcripts. The main advantage of PE reads is that the connectivity information between the sequenced ends is available. PE sequencing is thus the obvious method to employ for identifying fusion transcripts. In a path-breaking study, Maher et al. [23] analyzed PE RNA-Seq data and demonstrated the feasibility of this technology to confirm known gene fusions and identify novel fusion transcripts. Their study also confirmed the need for a systematic analysis accounting for computational complexity and statistical significance. The method proposed, however, relies on the distance between the two ends of a transcript fragment (insert size). This idea, inspired by structural variant analysis, cannot be directly translated to the transcriptome analysis in order to obtain an accurate description of all the occurring events. The main reason is the complexity of the transcription, and in particular the splicing of introns, that can lead to read pairs spanning several exons.

Two more recent studies focus on the identification of novel splice junctions from RNA-Seq data [76, 77]. This problem is related to the discovery of fusion transcripts because, in principle, a “splice junction” can indeed join two different genes and thus suggest a fusion event. Although these methods can in principle be applied for the discovery of fusion transcripts, they mainly focus on the mapping of the reads. They do not analyze the impact of artifacts independent from the mapping procedure on the detection of fusion transcripts, such as the random pairing of transcript fragments during sample preparation (see Materials and methods). These tools also do not provide a means to summarize the results of the detection of potential fusion transcripts. Finally, the experimenter would not have the flexibility of using other mapping

tools that may provide complementary information. Specifically, SplitSeek is currently available only for AB/SOLiD [78].

To address these issues, we developed FusionSeq: a novel computational suite whose aim is to detect candidate fusion transcripts by analyzing PE RNA-Seq data (http://rnaseq.gersteinlab.org/fusionseq). FusionSeq is mapping independent as much as possible, such that it is not bound to a single platform or mapping approach. It accounts for several sources of errors in order to provide a high-confidence list of fusion candidates, which are also scored by using several statistics to prioritize experimental validation. FusionSeq also includes tools to summarize and present its results integrated into a web browser. Furthermore, we sequenced an appropriate data set to calibrate this approach, comprising mostly human prostate cancer tissues with and without known fusion events.

4.2 Results

4.2.1 Mapping the reads

The first step when dealing with next-generation sequencing is the alignment of the reads against known reference sequences. Here the main challenge is how to map millions of reads in a computationally efficient way. Several alignment tools have been developed and, since this research field is quite active, it is likely that improved or new tools will be introduced. In addition, a variety of mapping strategies can be employed. As an example, a splice junction library may be employed along with the reference genome to identify reads bridging exons. Our goal is to develop a method that is independent as much as possible from mapping strategies and alignment tools. As a test, we tried a variety of alignment tools and approaches, all yielding consistent results, thus demonstrating the robustness of FusionSeq. For simplicity, we here report the results obtained by mapping the reads to the genome with ELAND, the standard program supplied with the Illumina platform (see Materials and methods). Table 4.1 reports the results of the mapping step.

Sample ID	Type	Known fusion type	Read size (nt)	Total number of PE reads	Mapped PE reads	Percentage of mapped PE reads
106_T	PCa	TMPRSS2-ERG	51	7,239,733	4,723,941	65.25%
1700_D	PCa	TMPRSS2-ERG	51	12,435,299	7,629,273	61.35%
580_B	PCa	TMPRSS2-ERG	36	18,134,550	7,690,673	42.41%
99_T	PCa	NDRG1-ERG	36	2,844,879	1,515,444	53.27%
2621_D	PCa	SLC45A3-ERG	54	22,079,700	11,899,984	53.90%
1043_D	PCa	No known fusions	51	3,003,305	1,898,332	63.21%
NCI-H660	PCa cell line	TMPRSS2-ERG	51	6,512,688	4,120,365	63.27%
GM12878	Lymphoblastoid cell line	No known fusions	54	44,829,991	20,676,159	46.12%

Table 4.1 Results of the alignment step. Total number of PE reads, number of mapped PE reads and the percentage mapped are reported. Note that the number of single-end reads is double the number of PE reads.

4.2.2 Overall modular framework

The overall schematic of our approach is depicted in Figure 4.1. It consists of three modules.

4.2.3 Module #1: fusion transcript detection

This module only assumes that the PE reads have been aligned and their location is known. It identifies the set of candidate fusions from the mapped sequence reads. Conceptually, it consists of three steps (Figure 4.1A): (1) poor quality reads are removed; (2) PE reads that map to the same gene are considered part of the normal transcriptome; (3) PE reads that map to different genes are selected as potential candidate fusion transcripts; also, reads that do not align anywhere are stored for the computational validation of the candidates and for determining the sequence of the junctions. Note that the mapping of the reads can occur anywhere within a gene: exons, introns or splice junctions.

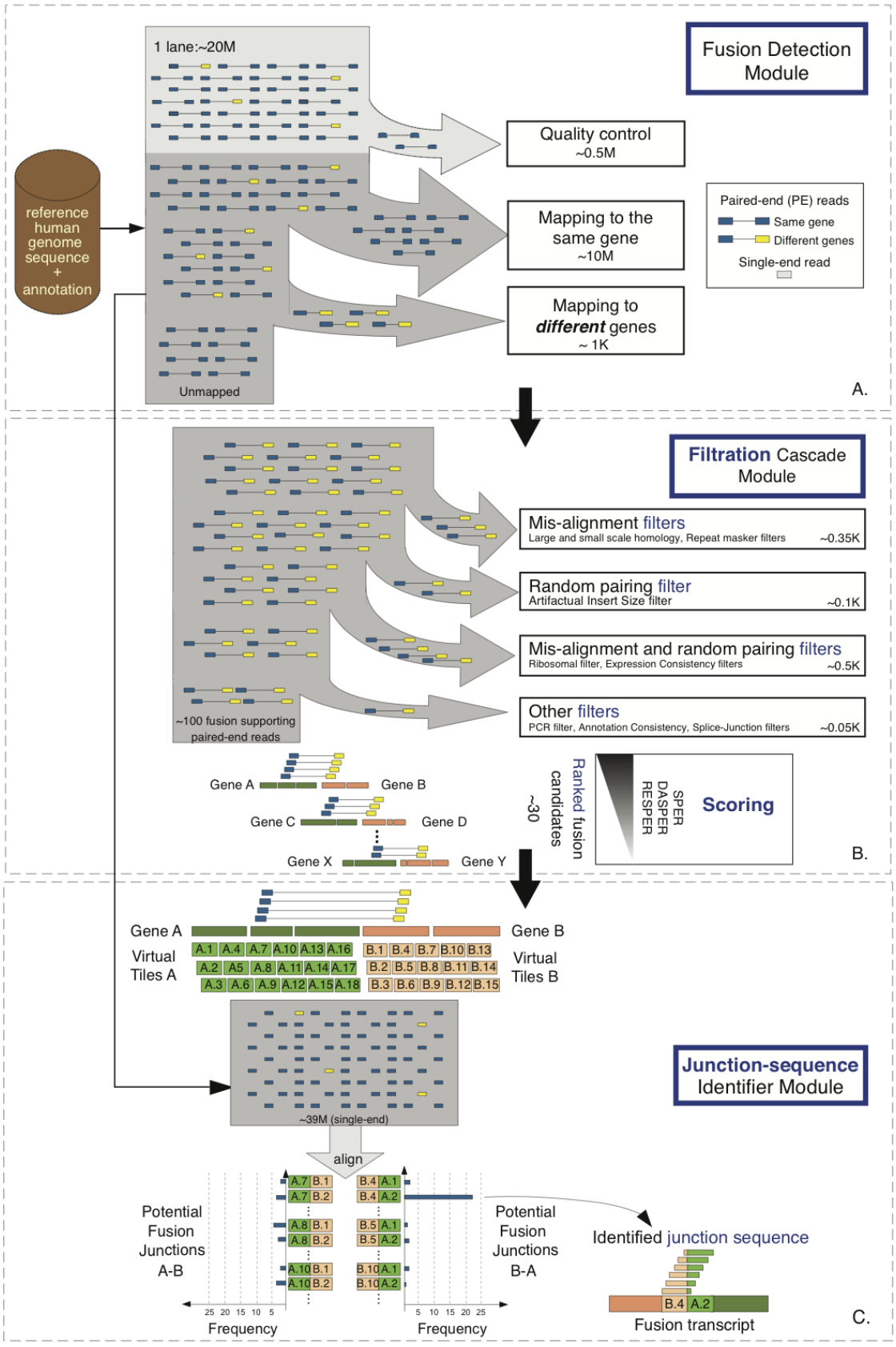

Fusion Detection Module
1 lane:~20M
reference human genome sequence + annotation
Quality control ~0.5M
Mapping to the same gene ~10M
Mapping to different genes ~ 1K
Unmapped
Paired-end (PE) reads
Same gene
Different genes
Single-end read
A.
Filtration Cascade Module
Mis-alignment filters
Large and small scale homology, Repeat masker filters
~0.35K
Random pairing filter
Artifactual Insert Size filter
~0.1K
Mis-alignment and random pairing filters
Ribosomal filter, Expression Consistency filters
~0.5K
Other filters
PCR filter, Annotation Consistency, Splice-Junction filters
~0.05K
~100 fusion supporting paired-end reads
Gene A
Gene B
Gene C
Gene D
Gene X
Gene Y
~30
Ranked fusion candidates
SPER
DASPER
RESPER
Scoring
B.
Gene A
Gene B
Virtual Tiles A
Virtual Tiles B
A.1 A.4 A.7 A.10 A.13 A.16
A.2 A5 A.8 A.11 A.14 A.17
A.3 A.6 A.9 A.12 A.15 A.18
B.1 B.4 B.7 B.10 B.13
B.2 B.5 B.8 B.11 B.14
B.3 B.6 B.9 B.12 B.15
~39M (single-end)
align
Junction-sequence Identifier Module
Potential Fusion Junctions A-B
Potential Fusion Junctions B-A
Frequency
Frequency
Identified junction sequence
Fusion transcript
C.

Figure 4.1 Schematic of FusionSeq. (A) The PE reads are processed to identify potential fusion candidates. Poor quality reads are discarded at first, and the remaining PE reads are aligned to the reference human genome (hg18). The reads are compared to the annotation set (UCSC knownGenes; [79]) in order to classify them as belonging to the same gene or to different genes. Those aligned to two different genes are then selected as potential fusion candidates. All good quality single-end reads are also stored for the identification of the sequence of the junction. (B) The filtration cascade module analyzes the candidates and removes those that have high sequence homology between the two genes or a higher insert-size compared to the transcriptome norm. Additional filters are employed to remove candidates due to random pairing and misalignment as well as PCR artifacts and annotation inconsistencies. The high-confidence list of candidates is then scored and processed to find the sequence of the junction. (C) The junction-sequence identifier detects the actual sequence at the breakpoints by constructing a fusion junction library. It first covers the regions of the potential breakpoint of each gene with "tiles" 1bp apart, and then creates all possible combinations, considering both orientation of the fusion, namely gene A upstream of gene B and vice versa. All single-end reads are then aligned to the fusion junction library and the junction with the highest support is identified as the sequence of the fusion transcript junction.

We employ a reference annotation set (University of California Santa Cruz - UCSC Known Genes [79]) and classify each single-end of a PE read into different categories depending of what parts of the gene is mapped to: exon, intron, splice junction or boundary. The latter case corresponds to reads that might be mapped to the genomic boundary of an exon - for example, in the case of a retained intron or when pre-mRNA is sequenced.

4.2.4 Module #2: filtration cascade

Several types of noise can introduce artifacts at any stage of the sequencing and analysis process. Hence, we developed a number of different filters to reduce the problem of artificial chimeric transcripts (see Figure 4.1B). Additional filters that are specific to the reference annotation set are also employed.

4.2.4.1 Misalignment filters

The reads can be mapped to a different location on the genome compared to where they were generated, mainly because of the sequence similarity of regions in the genome (paralogs, pseudogenes, repetitive elements). Indeed, it is possible that single nucleotide polymorphisms (SNPs), RNA editing, or errors in the base caller can lead to misalignment of one of the ends resulting in artificial chimeric transcripts. This issue is particularly relevant in the intermediate range of sequencing depth (1-100M reads), which FusionSeq has been designed for. We devised three filters to deal with this issue of sequence similarity, briefly described hereafter (see Materials and methods for detail).

4.2.4.1.1 Large scale sequence similarity filter

If the two genes of a candidate fusion transcript are paralogous, they are discarded because of this homology potentially causing a misalignment. We use TreeFam to identify these candidates and remove them from the list [80, 81].

4.2.4.1.2 Small scale sequence similarity filter

The above filter seeks broad similarities between two transcripts. However, it may be possible that there is high similarity between small regions within the two genes where the reads actually map. To identify these cases, for each of the candidate chimeric transcripts, the reads aligned to one gene are searched for sequence similarity against the corresponding partner. If high similarity is found, the pair is removed (Materials and methods).

4.2.4.1.2 Repetitive regions filter

Some reads may be aligned to repetitive regions in the genome, due to the low sequence complexity of those regions and may result in artificial fusion candidates. We thus remove reads mapped to those regions (Materials and methods).

4.2.4.2 Abnormal insert size filter

The filters described so far deal with computationally generated artifacts. However, some artifacts can be intrinsic to the experimental protocol. Library preparation typically requires the fragmentation of the cDNA. This may result in the generation of random chimeric transcripts when inefficient A-tailing may lead to the ligation of random cDNA molecules [82]. This issue affects more highly expressed genes. The Abnormal Insert Size filter addresses this problem by exploiting the fact that the transcript fragments have approximately the same size because a size-selection step is typically part of the experimental protocol. We could filter the set of candidate fusion transcripts by selecting those paired reads having an insert size - that is, the distance between the two mapped reads - comparable to the fragment size and by excluding those with a much higher insert size, somewhat resembling the approach for determining DNA structural variants [72, 83–85]. However, this approach is based on the fact that the alignment of genomic

PE reads to the genome reflects its linearity, where any deviation from this “nominal” insert size will be considered abnormal (Figure 4.2A). These approaches cannot be directly translated to RNA-Seq analysis because of at least three additional layers of complexity: *i)* the splicing mechanism of the transcription; *ii)* the genome of the individual, which contains some differences from the reference genome; and *iii)* the cancer genome of the same individual, which can include additional somatic variations (Figure 4.2B).

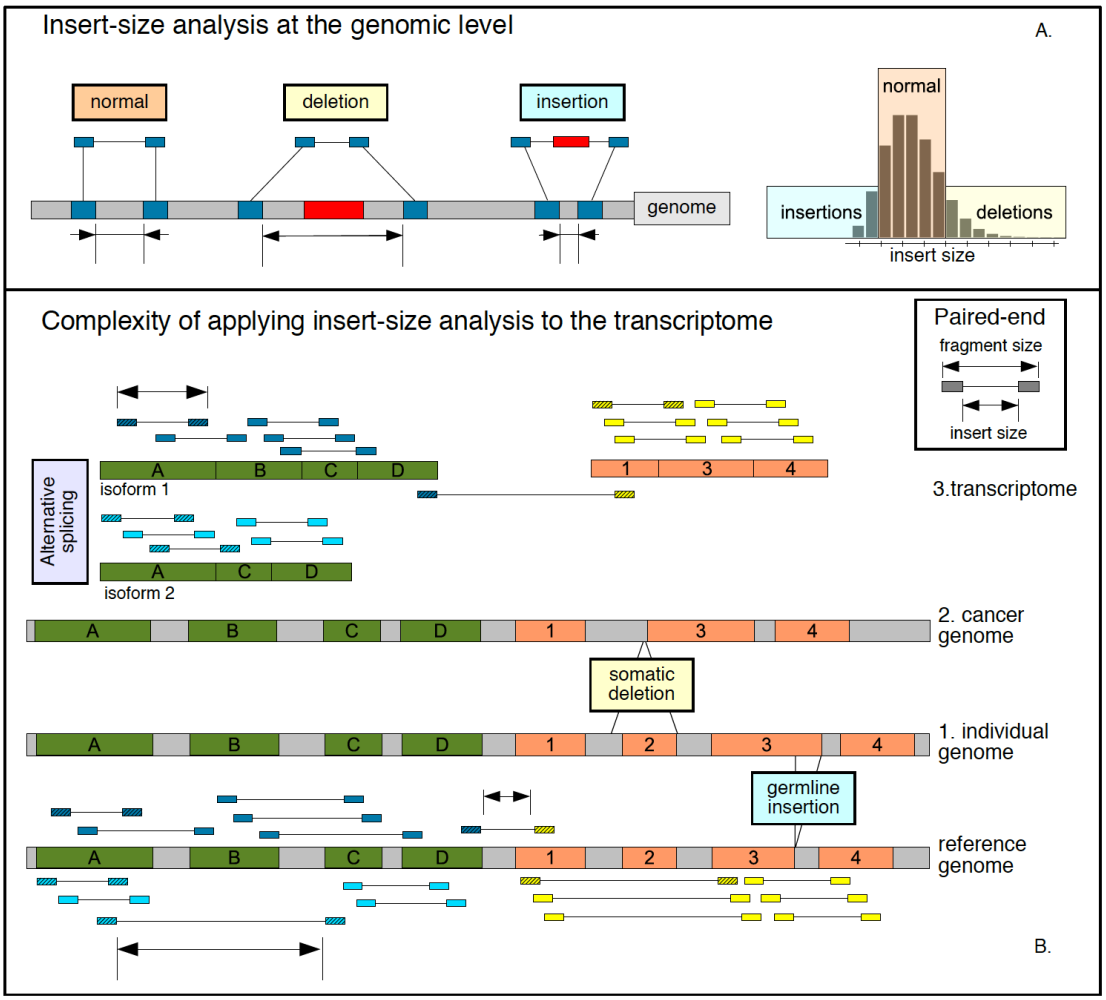


Figure 4.2 Insert-size analysis. (A) The insert-size computation can help identifying structural variations such as insertions and deletions, by comparing to the normal insert-size distribution. Deletions will result in bigger insert sizes, whereas insertions are characterized by smaller insert sizes. (B) The direct application of this principle to the transcriptome is not possible. Three layers of complexity compared to the reference genome can prevent the direct use of the insert-size analysis: 1. germline variations of the individual genome; 2. somatic variations of the cancer genome; and 3. splicing and alternative splicing. If PE reads are mapped to the transcriptome (unknown) the insert-size of normal transcripts will be comparable to the fragment size. However, since PE reads are mapped to the genome, the insert size is not meaningful anymore. PE reads of normal spliced genes may have a bigger insert size (hashed light blue) compared to PE reads of fused genes (hashed blue-yellow). Hashed symbols highlight the correspondence of the reads from the transcriptome to the genome.

We devised a method to address some of these issues and still make use of this concept to identify true chimeric transcripts. We first introduce the concept of the "composite model" of a gene - that is, the union of all exons from all known isoforms of a gene - and then we define the "minimal fusion transcript fragment" (Figure 4.3). This is generated by using all PE reads bridging the two different genes. It is important to note that in the case of a real fusion transcript, we can only identify the region around the fusion junction. Reads generated by a fusion transcript that are distant from the junction will be assigned to one gene or the other. For a real chimeric transcript, the minimal fusion transcript fragment will thus capture the region around the breakpoint and the insert-size distribution computed on it will be similar to the insert size distribution of normal transcripts. Conversely, for an artifactual chimeric transcript, paired reads would randomly join the two genes from all different parts (Figure 4.3B – right). The minimal fusion transcript fragment would be bigger than the expected fragment. Hence, the insert-size distribution computed on this minimal fusion transcript fragment will be higher than that of normal transcripts, i.e. abnormal. The normal insert-size distribution can be estimated from the data by using the composite models of all genes (see Materials and methods).

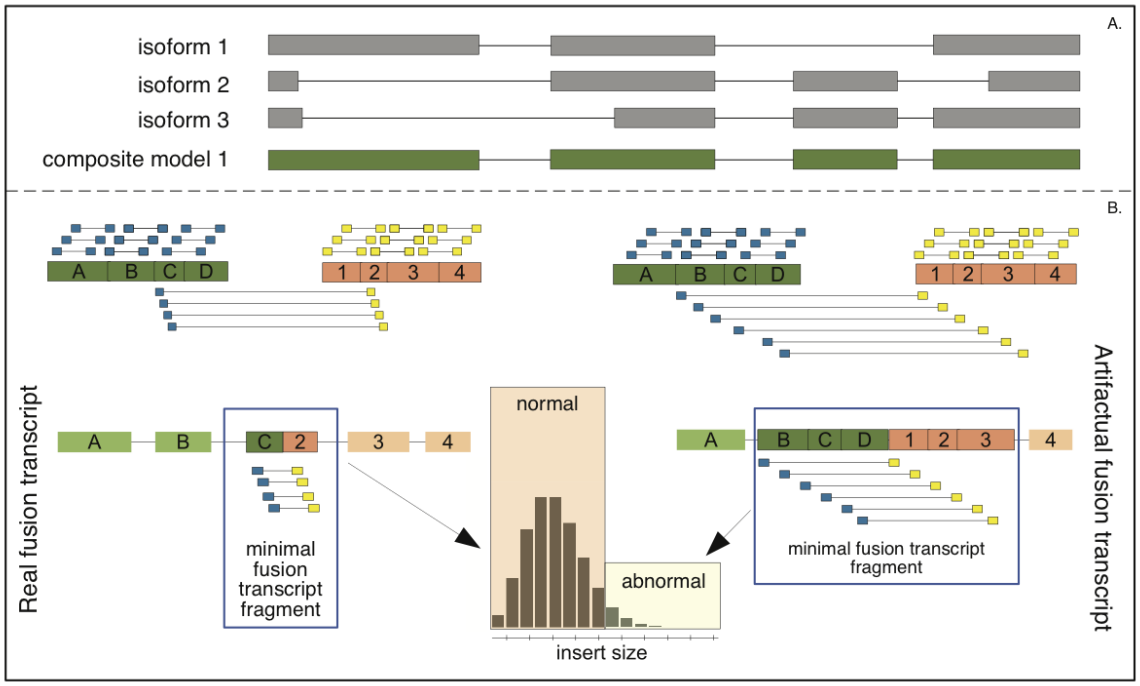


Figure 4.3 Abnormal insert-size principle applied to transcriptome data. (A) The composite model of a gene is created via the union of the exonic nucleotides from all its isoforms. By using the composite model, we can exploit the abnormal insert-size principle. (B) A minimal fusion transcript fragment is created by connecting the regions of the two genes joined by PE reads. Subsequently, the insert-size of these chimeric PE reads is computed and compared to the insert-size distribution of PE reads in the normal transcriptome. The higher insert-size compared to the transcriptome norm would suggest an artifact since it may be due by the random joining of fragments during library generation.

4.2.4.3 Filters for removing misalignments and random pairings

An additional complication is the possibility that random pairing and misalignment occur together. Highly expressed genes may generate transcript fragments that randomly join with another gene. In addition, misalignment can affect the correct identification of the genes involved in this random pairing. This is particularly challenging because only a fraction of the reads from random pairing will be misaligned; specifically, those with high similarity to another region of the genome. This would result in PE reads bridging relatively small regions that can escape the Abnormal Insert Size filter. Hence, we devised two additional filters: one comparing the candidates to the typically highly expressed ribosomal genes, and the other assessing the consistency of the expression levels of the individual genes of a chimeric transcript (see Materials and methods).

4.2.4.4 PCR filter

Most library preparations also require a PCR amplification step. This may lead to potentially artifactual fusion candidates when the same read is over-represented, yielding to a "spike-in-like" signal, i.e. a narrow signal with a high peak. To reduce this effect, we filter candidates that have chimeric reads piling up in a small region (see Materials and methods).

4.2.5 Module #3: junction sequence identifier

After the identification of high-quality candidate fusion transcripts, we can seek the overall support of those candidates taking advantage of the pool of all single-end reads. This process also allows the identification of the exact sequence of the fusion transcript junction. The knowledge of the actual junction sequence has many uses. First, it can help to identify the actual regions that are connected in the fusion transcript. Second, it helps in subsequent experimental validation, for example by RT-PCR. Finally, it can provide additional evidence for the fusion transcript or can be used to rule out artifacts.

In order to identify the junction sequence, we build a "fusion junction library" and align all single-end reads to this library (Figure 4.1C). To be computationally efficient, we first identify the regions where the potential breakpoints are using the information from the PE reads bridging the two genes. The exact size of the regions bears greatly on the resulting complexity of the potential fusion transcript and the computational power (see Materials and methods). Then, we cover these regions with "tiles" that are spaced 1bp apart and, finally, we generate the fusion junction library by creating all pair-wise connections between these tiles. The rationale is that the correct junction sequence will correspond to one of these connected tiles and that there will be full-length single-end reads that will align to that sequence (see Materials and methods).

4.2.6 Scoring the candidates

Although FusionSeq filters out many spurious fusion candidates, some may still be present, especially random chimeric transcripts generated during sample preparation. Hence, candidates are scored based on their likelihood to be real allowing prioritization of validation

experiments. The first obvious measure is simply the number of inter-transcript PE reads (m_i) normalized by the total number of mapped PE reads (N_{mapped}), similarly to RPKM for measuring gene expression [18]. This is expressed per million mapped reads and called SPER for "*S*upportive *PE R*eads". For the *i*-th candidate:

$$SPER_i = \frac{m_i}{N_{mapped}} \cdot 10^6$$

This measure gives an indication of the abundance of the fusion transcript. However, to assess whether a given *SPER* is "high" enough, we compare it with two "expected" values: one is analytically calculated and the other, empirically. The first quantity is *DASPER*, i.e. the *D*ifference between the observed and *A*nalytically calculated expected *SPER*, indicating how many (normalized) inter-transcript PE reads we observe more than expectation. The analytically calculated expected *SPER* (<*SPER*>) is based on the observation that if two ends were randomly joined, the probability that this occurs for gene A and gene B is proportional to the product of the probability that the two single-ends of the pair are mapped to gene A and gene B (see Materials and methods). This scoring method takes into account fusion transcripts that might have been generated during sample preparation from highly expressed genes. Obviously, the higher *DASPER* is, the more likely the fusion candidate is real.

The second measure is *RESPER*: the *R*atio of *E*mpirically computed *SPERs*. The rationale for this measure is the comparison of the observed *SPER* with the *SPERs* of the other candidates. We expect a real fusion transcript to be supported by a higher number of reads compared to the artifactual chimeric transcripts (see Materials and methods). This quantity, contrary to *DASPER*, is independent from the fragment size, thus more suitable for comparisons across samples. While *RESPER* is useful, it suffers in comparison to *DASPER* if a sample has several real fusions.

In summary, by computing these quantities, we can "demote" fusion candidates that may result from random joining of highly expressed genes (*DASPER*), and select those candidates

which “stand out” compared to the others (*RESPER*), thus providing a high-confidence ranked list of candidates.

4.2.7 Classifying the candidates

FusionSeq provides a list of potential fusion candidates which are automatically classified into different categories depending on the genes that are involved [62]: (1) inter-chromosomal - two genes on different chromosomes; (2) intra-chromosomal - two genes on the same chromosomes. The latter can be further subclassified as: (2a) read-through candidates if the two genes are close neighbors on the genomes, that is, if no other gene is present between them; (2b) *cis* candidates - similar to read-through events, but the two genes are on different strands.

Several read-through events have been reported in the literature, although their role remains unclear [86]. This may also be an effect of the pervasive transcription of the genome. Indeed, when considering primary transcripts, more than 90% of the nucleotides of the human genome are transcribed [60]. Although the RNA-Seq protocol requires a poly-A selection step, it may occur that pre-mRNA fragments with stretches of adenosines are still selected and sequenced.

4.2.8 FusionSeq applied to prostate cancer samples

In order to develop and calibrate FusionSeq, we selected a set of prostate cancer tissues harboring the common *TMPRSS2-ERG* fusion, others with less common fusions (*SLC45A3-ERG*, *NDRG1-ERG*) and prostate cancers with no evidence of known ETS fusions. We also sequenced a prostate cancer cell line with the *TMPRSS2-ERG* fusion (NCI-H660) and a lymphoblastoid cell line (GM12878) that was selected for the HapMap project and employed by the ENCODE project as controls. This normal cell line is not expected to have gene fusions (Table 4.1). Overall, FusionSeq takes about two hours to analyze 20M mapped reads. More details about the computational complexity are discussed in Materials and methods.

4.2.8.1 Fusion candidates

The application of FusionSeq to the above samples resulted in the identification of 12 fusion candidates on average per sample with *SPER* greater than 1 (range 0-25). Considering the top candidate for each sample, the average *SPER* is 13.99 for those with known *ERG* rearrangements and 3.09 for those without known fusions (Table 4.2). The vast majority of candidate fusions are intra-chromosomal - they occur between genes that are on the same chromosome - with the majority being read-through events.

The most common fusion, *TMPRSS2-ERG*, is ranked at the top of the list. The other known fusions between *ERG* and other 5' partners, namely *SLC45A3* and *NDRG1*, are also included in the top candidates. The remaining candidates appear to be read-through events, including *ZNF649-ZNF577* (Table 4.2).

Type	ID	Fusion candidate	*SPER*	*DASPER*	*RESPER*
Intra	**580_B**	**TMPRSS2-ERG**	**36.54**	**36.53**	**14.31**
Intra	**1700_D**	**TMPRSS2-ERG**	**19.66**	**19.63**	**8.79**
Intra	**106_T**	**TMPRSS2-ERG**	**10.16**	**10.11**	**3.97**
Inter	**2621_D**	**SLC45A3-ERG**	**4.29**	**4.15**	**3.56**
Inter	**1700_D**	**ERG-GMPR**	**4.59**	**4.59**	**2.05**
Read-through	1700_D	SLC16A8-BAIAP2L2	4.33	4.33	1.93
Read-through	*106_T*	*AK094188-AK311452*	*4.87*	*4.87*	*1.9*
Read-through	*1700_D*	*ZNF473-FLJ26850*	*3.54*	*3.54*	*1.58*
Read-through	*580_B*	*ZNF577-FLJ26850*	*4.03*	*4.03*	*1.58*
Read-through	*1043_D*	*ZNF577-ZNF649*	*5.79*	*5.79*	*1.55*
Read-through	*1700_D*	*CAMTA2-INCA1*	*3.01*	*3.01*	*1.35*
Inter	1700_D	EEF1D-HDAC5	2.88	2.84	1.29
Read-through	*1043_D*	*FLJ00248-LRCH4*	*4.74*	*4.74*	*1.27*
Read-through	*1700_D*	*VMAC-CAPS*	*2.62*	*2.62*	*1.17*
Read-through	*106_T*	*FLJ00248-LRCH4*	*2.96*	*2.96*	*1.16*
Cis	1043_D	AX747861-FLI1	4.21	4.21	1.13
Read-through	*106_T*	*TAGLN-AK126420*	*2.75*	*2.75*	*1.07*
Inter	**580_B**	**PIGU-ALG5**	**2.73**	**2.73**	**1.07**
Inter	**99_T**	**NDRG1-ERG**	**7.26**	**7.15**	**1.02**

Table 4.2 Summary of fusion candidates. *SPER*, *DASPER*, and *RESPER* are provided for the top candidates with *DASPER* > 0 and *RESPER* > 1 across all prostate cancer tissue samples. In bold, the known gene fusions, in italic the confirmed read-through events either experimentally or via additional evidence such as EST or mRNA from GenBank.

Although the candidates are ranked by *RESPER*, it is worth noting that the *TMPRSS2-ERG* fusion has high values for both *SPER* and *DASPER*, as expected. These statistics are almost equivalent for the top candidates; however, they substantially differ in the case of artifacts given by highly expressed genes, suggesting the effectiveness of *DASPER* in identifying those cases. Indicatively, *DASPER* and *RESPER* values greater than 1 seem to conservatively select

for true chimeric events, with 16 out of 19 candidates (84%) being either experimentally confirmed or with EST evidence.

We find a second candidate fusion transcript involving *ERG* and *GMPR* in sample 1700_D in addition to *TMPRSS2-ERG*. By analyzing the regions that are connected, it seems that the exons not involved in the *TMPRSS2-ERG* fusion are linked to *GMPR*, suggesting that *ERG* undergoes a balanced translocation. This novel finding was experimentally validated (as shown in Figure 4.4).

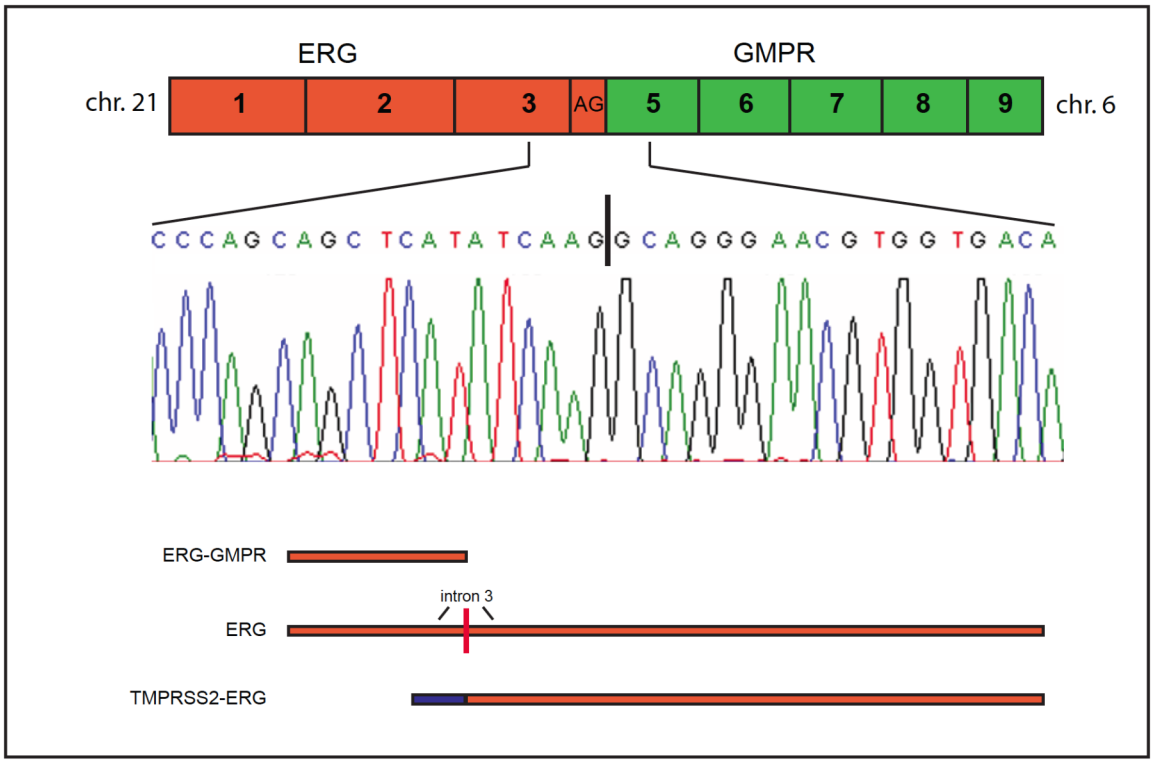


Figure 4.4 Novel fusion candidate: experimental validation. Transcripts of *ERG-GMPR* fusion were amplified by PCR using forward primer in *ERG* exon 2 and using reverse primer in *GMPR* exon 8. The schematic reports the *ERG-GMPR* fusion transcript along with the Sanger sequencing result representing the breakpoint sequence. Also, a schematic of *ERG* fusions by translocating 3' sequences to *TMPRSS2* gene on chr21 and 5' sequences to GMPR gene on chr6 is reported.

Another novel finding is the fusion transcript involving *PIGU* and *ALG5* that was also experimentally confirmed [87]. Finally, there is one *cis* candidate, including *AX747861* and *FLI1*, which may suggest some complex rearrangement (Materials and methods). However, from EST data there is evidence that this may correspond to a single *FLI1* transcript, thus suggesting an

artifact caused by the annotation set. Although FusionSeq can properly handle such cases with the annotation filters, we report it here as an example of how the framework can be employed to refine the search of candidate fusion transcripts and help the experimenter screen this list.

4.2.8.2 Effects of the filters

The application of the filters reduced the number of candidates identified by the fusion detection module. Out of a total of 7342 candidates, only 133 candidates passed all the filters, i.e. a reduction of 98% (average number of identified candidates per sample = 917.75, range [451-1618] – average number of candidates per sample after filtering = 16.63, range [4-41]). In Figure 4.5A, we summarize the effect of the filters. Each filter reduces the number of potential candidates to some extent, indicating that they address these issues. We experimentally verified that some of the candidates filtered out or with negative *DASPER* are artifactual.

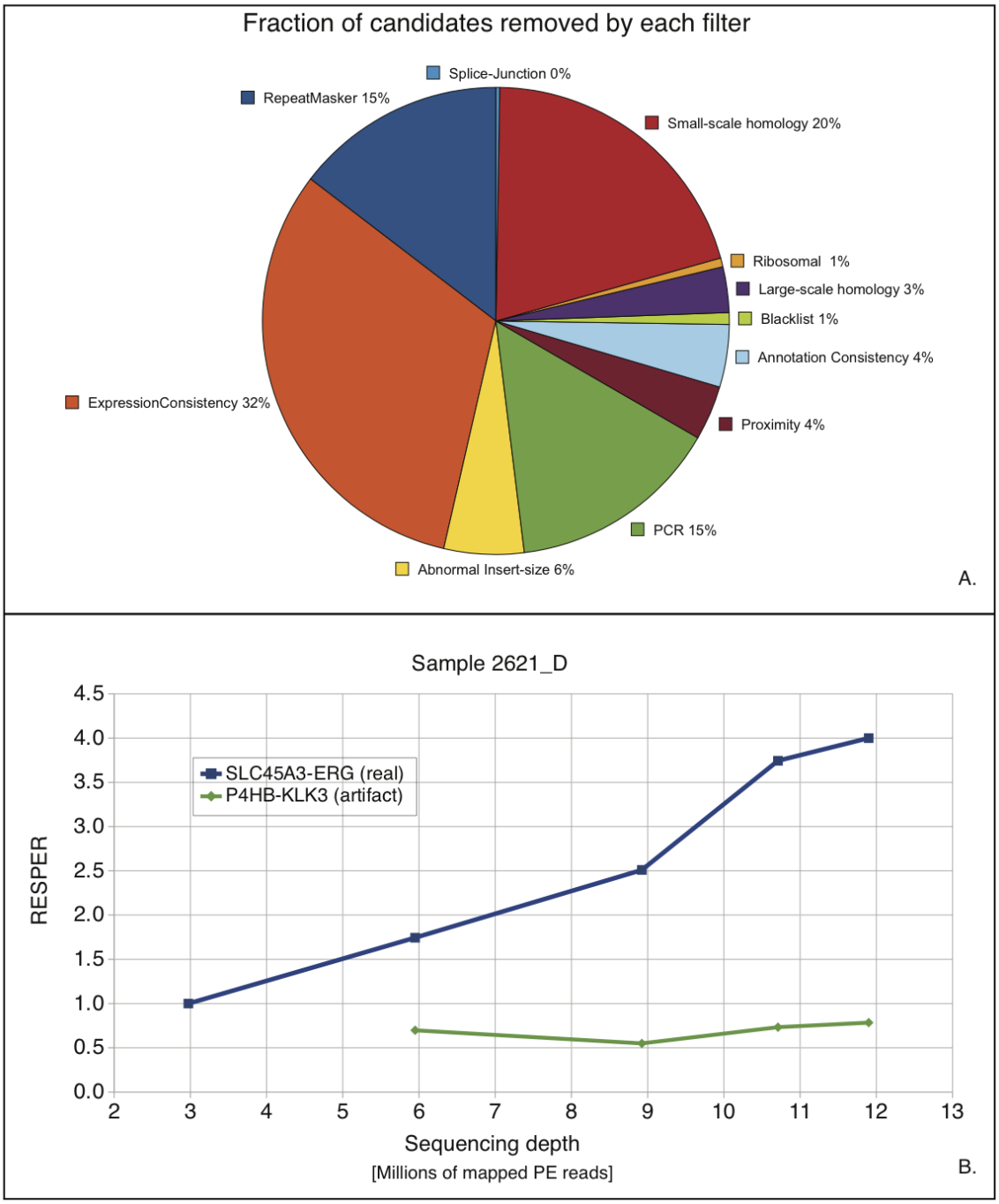


Figure 4.5 Filtration cascade module. (A) The average percentage of candidates identified by the fusion detection module that are removed by each filter is reported. The labels also depict the order the filters have been applied in this case (counter-clockwise starting from the RepeatMasker filter), but it is worth noting that the order of the application of the filters does not affect the final list of candidates. (B) *RESPER* vs. depth of sequencing. The plot shows the *RESPER* values for *SLC45A3-ERG*, a real fusion transcript, and *P4HB-KLK3*, an artifact likely created by the random pairing due to the high expression of *KLK3* at different sequencing depths.

4.2.8.3 Sequencing depth and detection of fusion candidates

We investigated the effect of the number of mapped reads on the detection of fusion transcript. We randomly sampled fractions of mapped reads from sample 2621_D, and applied FusionSeq to the reduced data sets (see Materials and methods). The top candidate is always *SLC45A3-ERG* with an increasing *RESPER*, as expected (Figure 4.5B). That *RESPER* increases with increasing sequencing depth is an indicator that the real fusion transcript stands out compared to the background. Although the number of fusion candidates increases as well, the *DASPER* for the majority of other candidates is negative, suggesting that they are artifacts.

4.2.8.4 *TMPRSS2-ERG* fusion positive prostate cancer tissues

For all the *TMPRSS2-ERG* positive prostate cancer tissues, FusionSeq always detects this fusion transcript at the top of the list. Figure 4.6A shows the PE reads bridging the two genes for the 3 tissue samples and the cell line harboring the fusion for the entire region between *TMPRSS2* and *ERG*. It is worth noting that the regions connected by the PE reads are different across the samples, suggesting the presence of different *TMPRSS2-ERG* isoforms.

4.2.8.5 Exon expression

The expression of a fusion transcript should also be reflected in the intensity of the signal at the exon level. Specifically, if a fusion transcript does not include some exons of the “wild-type” gene, the expression of those excluded exons should be lower compared to that of exons part of the fusion transcript. This observation was originally reported by Tomlins et al. [33] using a standard exon walking experiment and has been confirmed using exon arrays [88].

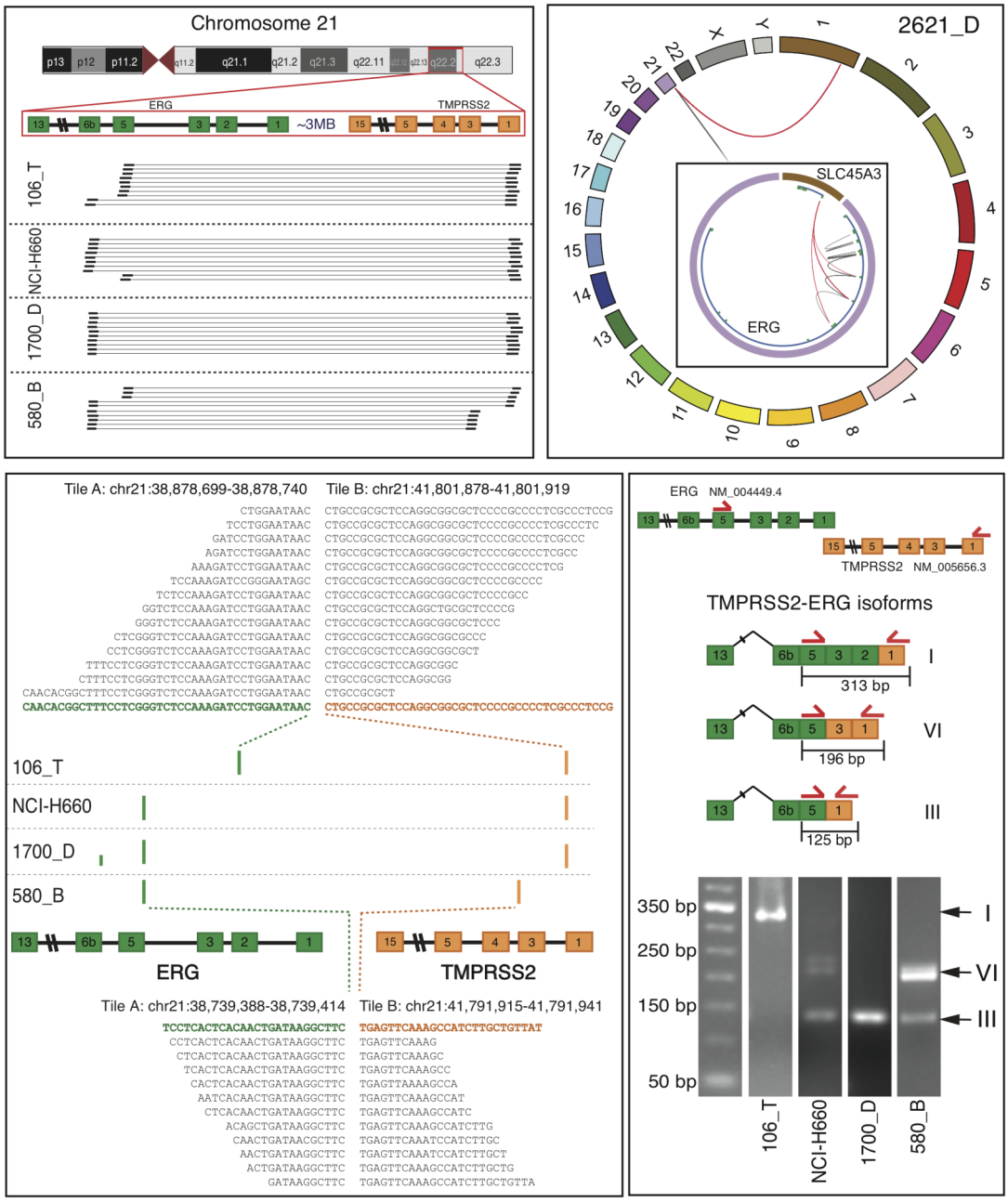


Figure 4.6 Results of FusionSeq. (A) A subset of the PE reads connecting *TMPRSS2* and *ERG* are shown for 4 samples (106_T, NCI-H660, 1700_D, 580_B). (B) PE reads connecting *ERG* and *SLC45A3* for sample 2621_D. The outer circle reports all chromosomes, whereas the inset shows only the region of *ERG* and *SLC45A3*. The gray lines depict the intra-transcript PE reads, whereas the red ones represent the inter-transcript PE reads. Note that for illustration purposes, only the inter-transcript reads are shown for *SLC45A3*. The inset also depicts the composite model (blue line) and its exons (green boxes). (C) Results of the junction-sequence identifier. The location of the breakpoints for the 4 samples with the *TMPRSS2-ERG* fusion are reported as bars (not to scale). Moreover, the sequence of the junctions as well as a subset of the aligned reads for 2 samples is reported (106_T, 580_B). (D) The locations of the PCR primers used for the validation are depicted as red arrows. The isoforms consist of *TMPRSS2* and *ERG* exons fused to form different exon combinations as depicted schematically. For both samples NCI-H660 and 1700_D, isoform III is detected, whereas, for samples 106_T and 580_B, isoforms I and VI are determined, respectively [89, 90]. The transcript isoforms were validated by a PCR assay for each sample separately (gel images). A 50 bp length standard (lane 1) is shown here for the determination of the approximate fragment size. The identity of the PCR products was validated by Sanger sequencing.

For illustration purposes, Figure 4.7 shows the expression values (RPKM) for the exons of *ERG* and *TMPRSS2*. It is common that the expression of *ERG* is driven by its fusion with a 5' partner. Hence, we can expect that the major expression signal is due to the fusion transcript. Indeed, the expression signal of the exons involved in the fusion transcript is higher than that of the region excluded. A similar conclusion is obtained when looking at *TMPRSS2*.

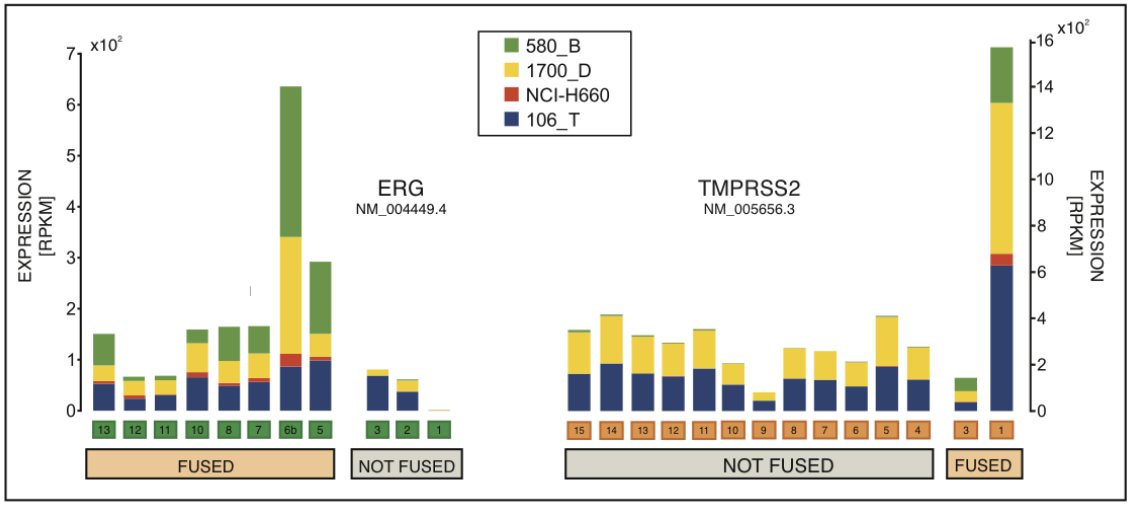


Figure 4.7 Expression values of the exons of *TMPRSS2* and *ERG*. The RPKM values computed on each exons of *ERG* (isoform NM_004449.4) and *TMPRSS2* (isoform NM_005656.3) are shown as stacked bars for the 4 samples with *TMPRSS2-ERG* fusion. For illustration purposes, the exons included in the most common fusion isoforms are labeled as "FUSED".

4.2.8.6 Junction-sequence identification analysis

Figure 4.6C shows the results of the junction-sequence identifier module for the 4 samples with *TMPRSS2-ERG* fusion. The main breakpoints are detected for both *TMPRSS2* and *ERG*. This allows the determination of the correct fusion isoform, which was experimentally validated with RT-PCR (Figure 4.6D). By taking a closer look at the junction-sequence identification results, a second potential breakpoint for sample 1700_D can be detected, albeit with much fewer number of reads (5 compared to 320 for the main breakpoint) (Figure 4.8A). The reads supporting it are uniformly distributed across the junction, suggesting that it is a real breakpoint and that multiple fusion variants are present. This finding has been validated with RT-PCR using a primer specific to this junction (Figure 4.8B).

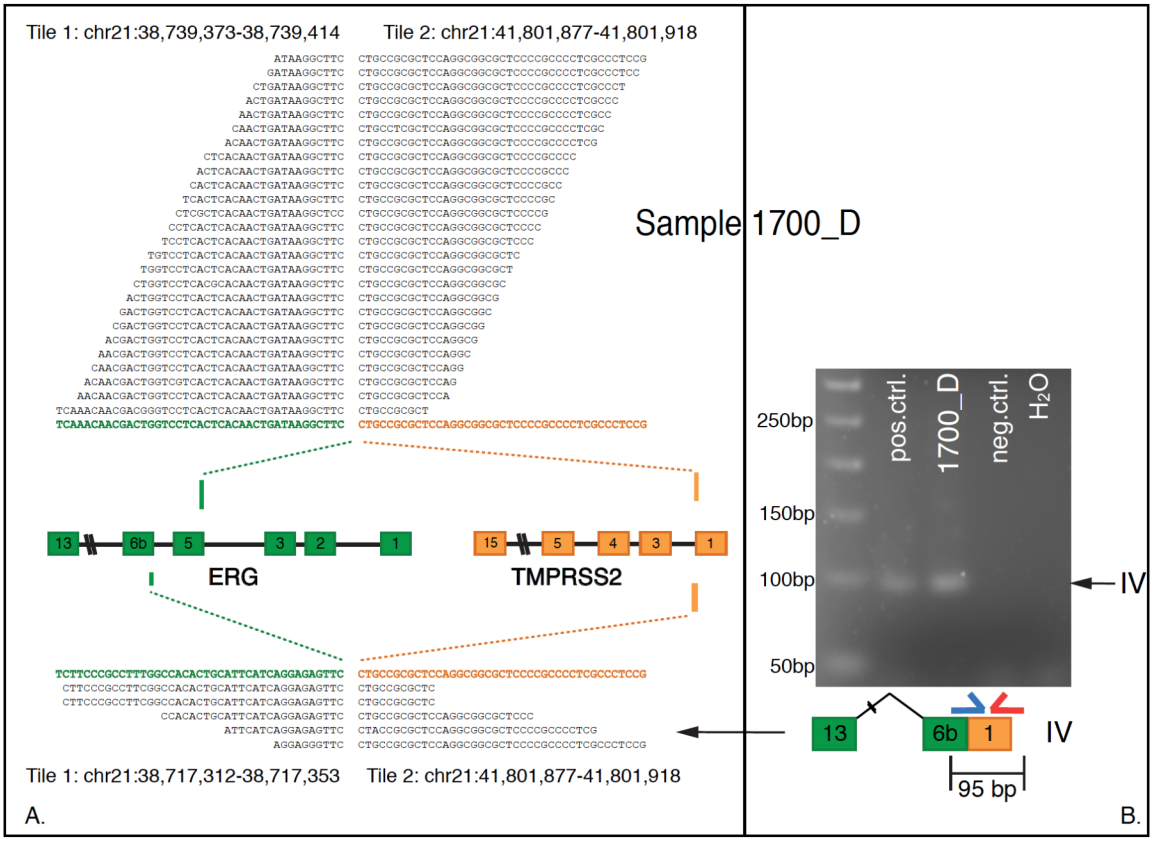


Figure 4.8 Validation of the minor breakpoint. (A) The location of the breakpoints and the sequences of the junctions as well as a subset of the aligned reads is reported. The minor breakpoint between *TMPRSS2* exon 1 and *ERG* exon 6b in sample 1700_D is consistent with the expression of isoform IV. (B) A PCR assay designed specifically for isoform IV, detects a single 95bp PCR product.

4.2.8.7 *ERG* rearranged cases with different 5' partners

We analyzed two other *ERG* rearranged cases where the 5' partner of *ERG* is different from *TMPRSS2*. We previously reported the discovery of a novel rearrangement between *ERG* and *NDRG1* for sample 99_T, resulting from the focused analysis of PE RNA-Seq restricted to the specific region of *ERG* [63]. With the current method that performs a genome-wide analysis, we confirmed the *NDRG1-ERG* fusion transcript as the top candidate (Table 4.2). Furthermore, we applied FusionSeq to another *ERG* rearranged sample, 2621_D, identifying *SLC45A3-ERG* as top candidate (Table 4.2 and Figure 4.6B).

4.2.8.8 *ERG* rearranged negative case and normal cell line

When applied to the sample without known fusion transcripts (1043_D), FusionSeq detected only a few candidates, the top being the read-through event between *ZNF577* and *ZNF649*, which is common in all prostate tissues here analyzed and has been already documented [62]. For the GM12878 cell line, it is noteworthy that, despite having more than 20M mapped PE reads, none of the few candidates (n=4) have a *SPER* higher than 0.3, as expected being a normal cell line. The read-through event with positive *DASPER* appears to be a mis-annotation of the untranslated regions (UTRs) (*BC110369-BC080605*), whereas the inter-chromosomal candidates have a negative *DASPER*, suggesting that they may be due to random chimeric pairing. Indeed, one of the genes involved is a highly expressed gene: *ACTG1*, with an RPKM > 232,000 [18]. Furthermore, the junction-sequence identifier analysis does not yield any result.

4.2.9 Simulation results

In addition to experimental evidence, we also performed a simulation study to assess FusionSeq performance. We employed the GM12878 cell line as an estimate of the background because it is not expected to harbor any fusion transcripts. We randomly generated inter-transcript reads thus simulating the presence of fusion transcripts and added these PE reads to the pool of the actual PE reads of the GM12878 cell line data. The results showed that a *DASPER* score greater than 1 achieves high sensitivity (0.80) even if the fusion transcript is expressed at half the “wild-type” allele (F=0.5) with an Area Under the ROC curve (AUC) higher than 0.95.

4.3 Conclusions

Gene fusions have been considered the key molecular event in leukemias, lymphomas, and some soft tissue tumor (i.e., sarcomas). With the 2005 discovery of common recurrent gene fusions in prostate cancer, there exists a strong likelihood that recurrent gene fusions are present in common epithelial cancers [33]. Numerous studies have now confirmed that approximately

50% of prostate cancers harbor a recurrent fusion between *TMPRSS2* and *ERG* or *ETV1* [91]. In an attempt to identify these fusion events, we employed PE RNA-Seq technology exploiting the connectivity information of the two ends of transcript fragments. As is the case of other applications of deep sequencing, considerations of computational complexity and statistical significance are mandatory.

4.3.1 FusionSeq: a modular framework

In the current study, we describe FusionSeq, a novel computational and statistical framework to identify fusion transcripts by analyzing PE RNA-Seq data. This framework consists of three modules: a fusion transcript detection module; a filtration cascade module, which is composed of a set of filters that remove different types of artifacts and rank the candidates by different scores; and a junction-sequence identifier module, which detects the actual sequence of the fusion junction.

Among the advantages of our method is the decoupling of the alignment approach from the identification of candidate fusion transcripts. Indeed, we developed FusionSeq to be independent from the alignment tool and the mapping strategy as much as possible. Other methods proposed that could potentially identify fusion transcript requires a particular choice of the mapping tool or platform and do not provide any considerations about artifactual fusion transcripts generated by the sequencing protocol [76, 78]. To this end, we develop a set of filters to remove artifactual candidates generated by several sources of errors (see Materials and methods), which are particularly relevant in the intermediate range of sequencing depth (1-100M reads). It is likely that with higher coverage those issues will impact the analysis less since one can use the statistics of the higher coverage to overcome errors.

Of further interest is also the ability of this method to identify the sequence of the junction of the fusion transcript using the full read length. This valuable information allowed us to detect and then experimentally confirm the simultaneous presence of multiple fusion isoforms within a single cancer tissue. Moreover, it enables the experimentalist to narrow the genomic region to look at for the subsequent validation of the fusion candidate. All validated fusions in this data set

have breakpoints lying at the exon boundary. This might indicate that, in case of genomic rearrangement, the splicing machinery is still active and removes the intronic regions harboring the actual genomic breakpoints. Hence, we speculate that insertions or deletions that typically occur at genomic breakpoints might not affect the junction of the fusion transcript.

4.3.2 Scoring the candidates

One of the novel features introduced by FusionSeq is the computation of scores to assign a "confidence value" to the fusion candidates. We propose a classification and scoring approach to prioritize the selection of candidates for experimental validation (see Materials and methods).

We envision that researchers seeking gene fusions can use this tool to focus their efforts on the candidates with the top scores. Validation typically includes seeking confirmation of the putative fusion sequence using standard PCR assays and traditional sequencing as well as exploring for a corresponding genomic rearrangement at the DNA level using such approaches as fluorescence in situ hybridization (FISH).

4.3.3 Sample set

One important aspect of this study is that we tested FusionSeq on data generated from cancer samples derived from human tumors and not only cell lines. Clearly these types of samples are more challenging given their heterogeneity as they may include tumor, stromal, and endothelial cells. We have used a set of prostate cancer samples with and without the *TMPRSS2-ERG* fusion transcript to calibrate FusionSeq. This well-characterized gene fusion was not only detected where present, but the junction sequence identifier also detected the correct junctions, thus enabling the determination of the specific isoform variants. Moreover, we observed that one sample had multiple variants.

Understanding the complexity of isoform splicing in cancer may not only add insight into biology, but may also have useful prognostic information as it has been suggested that some *TMPRSS2-ERG* isoforms play a distinct role in prostate cancer development [89, 92].

Furthermore, FusionSeq identified two novel events (*ERG-GMPR* and *PIGU-ALG5*), demonstrating that our procedure is able to find new fusions in addition to well-characterized ones.

4.3.4 Reporting the results

FusionSeq also includes tools to access and display the results of the analysis through a web-browser by seamlessly integrating the UCSC Genome Browser. Moreover, to display inter-chromosomal events, which is currently not possible in the UCSC Genome Browser, we developed SeqViz, a visualization tool based on Circos [93], an open source software particular suited for this purpose (see Materials and methods). These web-based interface tools enable the user to quickly access the information provided by FusionSeq, an aspect that greatly increase its applicability in comparison to other related tools [76, 78].

4.3.5 Future directions

Although we demonstrated the feasibility of this approach using several cancer tumor samples, there are some limitations to the current approach. The fusion transcript detection module is based on a gene annotation set that provides the information of the genes and their isoforms. Although the framework is flexible and the choice of which annotation to use is left to the user, the identification of the candidate fusion is of course limited to this set. We employed the UCSC knownGenes set, which contains 66,803 isoforms. We believe that this is a reasonable choice and that the use of a different annotation set would not dramatically change our results.

Although FusionSeq is independent from the mapping strategy adopted, it is likely that different mapping approaches would make use of the filtration cascade differently. As an example, if the alignment procedure explicitly excludes repetitive regions, the filter using RepeatMasker will impact on the final list of candidates to a lesser extent. This is why the modularity of FusionSeq allows the users to adapt the framework to their specific goals.

We anticipate that FusionSeq will benefit from the availability of longer sequence reads and deeper sequencing, with an increased ability to identify and score novel fusion events from RNA-Seq data.

4.4 Materials and methods

4.4.1 Prostate cancer selection and RNA extraction

All the prostate cancer samples were collected under an IRB (Institutional Review Board) approved protocol. Hematoxylin and eosin (H&E) slides were prepared from frozen tissue blocks and evaluated for cancer extent and tumor grade by the study pathologist (MAR). To ensure high purity of cancer cells and minimize benign tissue, tumor isolation was performed by first selecting for high-density cancer foci (< 10% stromal or other non-tumor tissue contamination) and then taking 1.5 mm biopsy cores from the frozen tissue block for RNA extraction using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA). The RNA extract was then subjected to DNase treatment using a DNA-*free*™ Kit (Applied Biosystems/Ambion, Austin, TX, USA). The quality of RNA was assessed using the RNA 6000 Nano Kit on a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Up to 10 µg of RNA with RIN (RNA integrity number) ≥7 was determined suitable for sample preparation.

4.4.2 Sample preparation

The samples were prepared in accordance with the Illumina RNA sample preparation protocol (Part # 1004898 Rev. A September 2008). Briefly, mRNAs were fragmented at elevated temperature using divalent cations and transcribed into cDNA thereby generating a library of cDNA fragments. RNA-Seq adapters were then ligated to the blunt ends of the cDNA fragments. The library of cDNA fragments subsequently underwent a size-selection step in which cDNAs were first electrophoresed through a 2.5% agarose gel in TAE buffer. Then, the desired fragment size products (200 bp or 300 bp) were retrieved from the gel and subjected to PCR amplification using universal primer sites present at the end of the ligated adapters. The library was then

subjected to quality control steps such as verification of fragment size and concentration measurements using the DNA 1000 Kit (Agilent Technologies) on an Agilent 2100 Bioanalyzer.

All samples were sequenced using one lane of an Illumina Genome Analyzer II (GAII) flowcell, except for GM12878, which was sequenced using 2 lanes. Since the experiments were performed over several months as Illumina introduced advances to the GAII platform, the total number of reads and the read length vary (see Table 4.1). However, all samples were prepared following the same protocol.

4.4.3 Validation of *TMPRSS2-ERG* fusion isoforms with PCRs

Aliquots from the same RNA stock were used for both RNA-Seq and PCR validation by conventional reverse-transcription PCR. RNA was reverse transcribed using a High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). The *TMPRSS2-ERG* PCR was performed using Platinum Taq DNA Polymerase (Invitrogen) with 1 mM MgCl2, 0.1 µM of each primer (forward, *TMPRSS2* exon 1 - TAGGCGCGAGCTAAGCAGGAG; reverse, *ERG* exon 5 - GTAGGCACACTCAAACAACGACTGG; as published by Tomlins et al. [33]) and 50 ng cDNA at an annealing temperature (Ta) of 63°C for 35 cycles and the PCR products were separated on a 2.5% agarose gel. For *TMPRSS2-ERG* isoform IV, the PCR was performed, using a reverse primer specifically designed for the detection of isoform IV (TGCATTCATCAGGAGAGTTCCTGC), under the same conditions but with Ta 55°C and 40 cycles. The obtained products were isolated from the gel using the MinElute™ Gel Extraction Kit (Qiagen, Valencia, CA, USA) and subsequently sent for Sanger sequencing at the Core facility of Cornell University (Ithaca, NY, USA).

4.4.4 Mapping

We employed ELAND to map the PE reads against the Human Reference Genome (March 2006 Assembly - hg18). We allowed for up to two mismatches of the alignment and selected reads that passed the quality filter from ELAND. In case of pairs mapped to the same chromosome, we selected reads aligned to opposite strands. We also employed bowtie to map

the reads to the human genome sequence [94]. Since bowtie does not allow PE reads to be mapped on different chromosomes, we adopted the following strategy: the two ends were mapped separately to the genome and the best alignment was selected among the top 10 candidates in the case of mapping to multiple locations. Two mismatches were allowed for bowtie too. Then, the two ends were paired together, if both ends were aligned. Moreover, for comparison purposes, we mapped the reads to a splice junction and a ribosomal library in addition to the genome.

4.4.5 Filtration cascade

4.4.5.1 Large scale sequence similarity filter

Two paralogous genes resulting as fusion candidates are discarded because of their homology can potentially cause a misalignment. We use TreeFam to identify and remove these candidates [80, 81]. TreeFam is a database of phylogenetic trees of animal genes with the aim of providing a curated list of orthologs and paralogs.

44.5.2 Small scale sequence similarity filter

The above filter seeks broad similarities between two transcripts. However, it may be possible that there is high similarity between small regions within the two genes where the reads actually map. Hence, to search for similar sequences within the two candidate genes, we employed a two-step strategy. We first perform a fast search of the reads aligned to one gene against the full transcriptome, represented by all composite models, using bowtie [94]. If more than a user-defined threshold (default: 1%) of the reads mapped to one gene “hit” the partner gene, the candidate is discarded. This approach removes candidates where the reads have high similarity, since bowtie allows up to 3 mismatches only. For those candidates not filtered out by this approach, a second, more refined comparison is performed. We align the reads mapped to one gene to its partner's sequence by using BLAT [95]. If the fraction of reads that have similarities to the corresponding partner is higher than a user-defined threshold (default: 1%) then

the pair is discarded. In order to call a hit - that is, similarity to the partner gene - we require that at least 75% of the read has similarity to the corresponding gene.

4.4.5.3 Repetitive regions filter

Some reads may be aligned to repetitive regions in the genome, due to the low sequence complexity of those regions and may result in artificial fusion candidates. We thus remove reads mapped to repetitive regions, using RepeatMasker to identify these regions [96].

4.4.5.4 Abnormal insert size filter

The PE RNA-Seq experimental protocol requires sequencing the ends of cDNA molecules of a determined length: the fragment size. If we mapped those sequenced reads to the transcriptome (which we do not know exactly), we would obtain an insert-size distribution, i.e. the distance between the two reads, similar to the fragment size. However, since the reads are aligned to the reference genome, the insert-size distribution can be rather skewed (Figure 4.2B). Using a splice junction library does not help in this context. Besides having potential biases given by the incomplete knowledge of the junctions, it cannot determine which isoform the two ends belong to. The composite model allows to use the concept of the insert-size also for RNA-Seq data (Figure 4.3A). The composite model is the union of all the exons from all known transcripts of a gene. This ensures that all exonic nucleotides are considered. The insert-size distribution computed using the composite model as reference should thus be comparable to the nominal fragment size selected during sample preparation (if there are more than one isoform, the insert size distribution computed with the composite model will be slightly shifted towards higher values because of the inclusion of all possible exonic regions in the composite model).

We then extend this concept to distinguish potentially real chimeric transcripts from artifactual ones. We generate a minimal fusion transcript fragment by using all the PE reads bridging two different genes (Figure 4.3B). The rationale is that the insert size distribution computed on this minimal fusion transcript fragment of a real chimeric transcript is similar to the insert size distribution of normal transcripts. This is because we expect inter-transcript PE reads

to connect the regions around the junction only. Conversely, a fusion transcript generated by random chimeric pairing would have a rather long minimal fusion transcript because paired reads would randomly join different regions of the two genes. This, in turn, would yield a much higher insert-size distribution compared to that of the real case; that is, it would be abnormal.

Specifically, for each of the candidate chimeras, the insert-size distribution is computed using all paired reads mapping to the composite model of that gene: i.e. the intra-transcript insert-size distribution. For this purpose, only reads that are fully contained within exons are considered. If a candidate has only intronic reads this filter is not applied. Similarly, the "anomalous" reads, i.e. reads that bridge two different genes, are first used to create a minimal fusion transcript (Figure 4.3B). Note that from the PE data we cannot determine the full fusion transcript, but only the region nearby the actual junction of the two genes, i.e. the minimal fusion transcript fragment. Then, the insert-size distribution of the minimal fusion transcript is computed (inter-transcript insert-size distribution) and compared with the intra-transcript insert-size distribution. If the median of the inter-transcript insert-size distribution is much higher than the median of the intra-transcript insert-size distribution, it is likely due by misalignments. A p-value is computed by randomly sampling the intra-transcript insert-size distribution. Candidate fusion transcripts having a p-value lower than a user-defined cut-off are discarded as artifacts. Note that the candidates that are "outliers" with respect to the intra-transcript insert-size distribution are discarded as artifacts, whereas, in the DNA context, those are kept as potential insertions or deletions (Figure 4.2A).

For this analysis, we used a p-value cut-off of 0.01 (corresponding to ~2.5 standard deviations from the transcriptome norm) for all samples, but for 2621_D, where we used a cut-off of 0.0001; the reason being the much tighter intra-transcript insert-size distribution given by the smaller fragment size compared to the other samples.

4.4.5.5 Ribosomal filter

The vast majority of transcripts in the cell are ribosomal RNA. Although the experimental protocol typically requires either selecting for non-ribosomal mRNA with polyA+ selection or

depleting of ribosomal RNAs, this process is imperfect. This translates in a high abundance of ribosomal transcripts with a higher chance of generating random chimera. If misalignment occurs too, this would result in artifactual candidates that appear to not involve ribosomal genes. Hence, this filter compares the reads of the candidates to the ribosomal genes sequence database using a more sensitive alignment tool such as BLAT [95]. If the reads align to ribosomal genes, the candidate is removed.

Specifically, in order to identify reads that bear similarity to ribosomal genes but were mapped to another region, we require a read to have more than 75% of similarity to a ribosomal gene to count it as a hit. If more than 10% of reads map to the ribosomal library the candidate is discarded. Note that this issue, although related, is independent from mapping strategy. Indeed, even if we employ a ribosomal library during the alignment phase, there still may be reads that, due to misalignment, will map best to other regions of the genome.

4.4.5.6 Expression consistency filter

Highly expressed genes give rise to the same issue that occurs with ribosomal genes. This filter compares the expression signal (i.e. number of reads) generated by the chimeric reads to the signal of the individual genes. The rationale is that, in the case of a real fusion transcript, the two genes would be expressed at the same or higher levels than the “chimeric” signal, whereas, in the case of an artifactual candidate, the signal would be generated only from the chimeric reads and the signal of the two individual genes would be much lower.

In more detail, the expression signal of the fusion candidate is computed by counting the number or inter-transcript, i.e. chimeric, reads mapped and normalizing by the length of the region covered by those reads. The expression of the individual genes is computed as the number of reads normalized by the length of the transcripts. If the chimeric reads have a higher signal than that of the individual genes, the candidate is discarded.

4.4.5.7 PCR filter

To avoid that artifacts resulting from the PCR amplification step, we require the reads supporting a candidate fusion transcript to independently cover at least *p* nucleotides (default *p* = 5) in addition to the read size on both end, otherwise the candidate is discarded. This ensures that several instances of the transcripts were expressed in the cells. In the case of sufficient coverage, it is also possible to compute the entropy to identify these cases and remove them.

44.6 Junction-sequence identifier module

The PE reads can identify the genes involved in a fusion transcript, but cannot directly determine the junction sequence, because, typically, short read alignment tools do not allow gapped alignment for the single read. Hence, we developed this module to take advantage of the fast short read alignment tools and identify the sequence of the junction in an efficient way.

Let us assume we have some PE reads joining gene A with gene B, thus suggesting a fusion event between them. Those reads would connect regions around the junction. For each gene, we thus select the region that can include the junction sequence by first considering all exons that can be potentially involved in the junction as well as the intronic regions that are supported by chimeric PE reads. Those regions are extended considering the flanking 150 nucleotides. We then cover them with a set of “tiles” that are spaced 1bp apart and construct a fusion junction library by creating all pair-wise junctions between these tiles. Since we do not know a priori what the specific form of the fusion transcript is, we create two libraries: one assuming gene A is upstream of gene B and the other assuming gene B is upstream of gene A (see Figure 4.1C). This fusion junction library plays the same role as a canonical splice junction library does: it enables the alignment of short reads, thus overcoming the need for a computational expensive gapped alignment for reads bridging two exons or, as in this case, regions of different genes.

All the reads, including the non-mapped ones, are then mapped against this library. In this case we considered the two ends independently. The rationale is that the actual junction sequence will be described by a certain pair of tiles and reads not previously mapped anywhere

in the genome now can be aligned to this fusion junction. Moreover, reads that previously mapped with 1 or 2 mismatches to the reference genome, now may map perfectly to the fusion junction and thus increase the evidence supporting the junction. The size of these tiles depends on the read size as well as the amount of overlap across the two joined tiles required by the user. For example, for reads that are 36bp long and a required overlap of at least 10 nucleotides, each fusion junction element is 52bp long, i.e. each tile is 26bp long. This ensures that every 36bp read, if mapped to this junction element, will have at least 10 nucleotides mapped to the tile of each gene.

To select the true junction sequence, we determine which fusion junction obtains the highest support, i.e. the junction with the highest number of reads aligned to. In addition, we also require the set of single-end reads to be uniformly distributed across the junction to provide further evidence. Provided there is enough coverage overall, we employ a Kolmogorov-Smirnov statistical test, otherwise we apply a simple heuristic by requiring that at least *n* reads align to the junction with at least *m* different starting position on the junction sequence. The latter parameter ensures that no PCR artifacts affect the junction identification. Also, we search for similarity of the identified junction elsewhere in the genome using BLAT [95], in order to eliminate potential spurious junctions.

From a computational viewpoint, let us assume that we have about 1000 virtual tiles for each gene. By creating all pair-wise combinations of these virtual tiles for the two genes and considering both directions, i.e. gene A upstream of gene B or vice-versa, will result in $1000 \times 1000 \times 2 = 2 \times 10^6$ putative junctions. If we have ~30 candidate fusion transcripts, the putative fusion junction library will thus contain $\sim 6 \times 10^7 = 60$ million elements. Using fast alignment tools, this analysis is feasible although it requires large-scale computational resources. Indeed, we use bowtie to first create an index of the fusion library and then map the reads against it [94]. To fully exploit the parallelization of a multi-node computing cluster, each fusion candidate is analyzed independently on different nodes. Moreover, the fusion junction library itself is also split across multiple nodes in order to optimize the generation of the indexes.

4.4.7 Sequencing depth and detection of fusion candidates

To assess the impact of sequencing depth on the detection of the fusion candidates, we randomly selected a fraction of mapped reads from sample 2621_D. Specifically, we extracted 10%, 25%, 50%, 75%, and 90% of all PE mapped reads (1.1M, 3M, 6M, 9M, and 10.8M PE reads, respectively). The number of fusion candidates with more than 5 PE reads clearly correlates with sample size: 0, 1, 3, 4, and 7, respectively. The *SLC45A3-ERG* fusion was detected as the top candidate, starting with 3M mapped PE reads, with a *SPER* of 4.7. The relatively low *SPER* for this candidate is related to the smaller fragment size that has been adopted for this sample (200nt compared to 300-330nt for the others). The smaller fragment size limits the number of PE reads that could span the junction. From this analysis, it appears that 3M reads are sufficient for detecting this fusion in this context. However, this result is difficult to generalize. It might be true only for fusion transcripts that are expressed at a similar level to *SLC45A3-ERG*. We cannot exclude the presence of less abundant fusion transcripts that would have been uncovered by deeper sequencing.

4.4.8 Scoring the candidates

We may take into account different types of information to score the candidates. Potentially we could use the number of PE reads bridging the two genes, the number of reads supporting the main junction, and the "shape" of the coverage as indicators of the reliability of the candidate. Practically, since it may be possible that the true junction is not detected because of lack of coverage, the more general quantities are based on the number of PE reads supporting the fusion candidate. Hence, every fusion transcript candidate is first scored using *SPER*, the normalized number of supportive PE reads, the most intuitive quantitative measure (see Results – Scoring the candidates). One may argue that a "local" score, i.e. a score that takes into account the expression of the genes involved in the fusion might be a reasonable choice. We defined *LSPER* (local *SPER*) as the number of inter-transcript PE reads supporting the fusion divided by the average gene expression value computed as RPKM [18]. However, in many cases, only one allele contributes to the fusion transcript. Hence, the expression of the fusion transcript

(estimated by the number of inter-transcript reads because the structure of the whole fusion is unknown) may not correlate with the expression of the genes generating it and thus this may impair the correct ranking of the candidates. After computing *SPER* for each candidate, we need to assign a "confidence" to this number. We compare it with two expectations. The first one, *DASPER*, i.e. the *D*ifference between the observed and *A*nalytically calculated expected *SPER*, is based on the observation that if two ends were randomly joined, the probability that this occurs for gene A and gene B is proportional to the product of the probability that the two single-ends of the pair are mapped to gene A and gene B:

$$P(A \cap B) = P(A) * P(B)$$

where *P(A)* and *P(B)* are the probabilities that a single-end is mapped to gene A and B, respectively. Note that this is a very conservative estimate because it does not take into account that single ends should also be within a certain distance, based on the fragment size, to be joined in a pair. Nevertheless, as a first approximation, the expected *SPER* can be estimated as the ratio of the number of single-end reads mapped to gene A and gene B and the total number of mapped single-end reads. For the *i*-th candidate, involving gene A and B, we have:

$$\langle SPER_i \rangle = \frac{\langle m_{AB} \rangle}{N_{mapped}} \cdot 10^6 = \frac{1}{N_{mapped}} \cdot \left\{ N_{mapped} \cdot P(A) \cdot P(B) \right\} \cdot 10^6 = \frac{m_A \cdot m_B}{N_{mapped}^2} \cdot 10^6$$

where $\langle m_{AB} \rangle$ is the expected number of inter-transcript PE reads under the null hypothesis, and m_A and m_B are the number of single end reads mapped to gene A and B, respectively. By subtracting this number from the observed *SPER*, we can rank the fusion candidates according to *DASPER* score:

$$DASPER_i = SPER_i - \langle SPER_i \rangle$$

We chose to compute the difference between these two quantities compared to a more traditional ratio or log-ratio because it is more robust in case of low coverage, i.e. low number of reads, than computing a ratio. More accurate estimations of the expected *SPER* can certainly be devised for cases with low coverage, however, they would likely require to take into account the specific characteristics of the sequencing platform and the mapping approach adopted, thus

reducing the broader applicability of this method. Although *DASPER* can reliably rank the candidates within a sample, it may be possible that when comparing candidates from multiple samples *DASPER* may not properly account for different fragment sizes. Indeed, smaller fragment sizes decrease the likelihood of sequencing PE reads bridging two genes, resulting in lower *SPER*, and consequently, lower *DASPER*, affecting the comparison among samples. To address this issue, for each fusion transcript candidate *i*, we compute the ratio of its $SPER_i$ to the average *SPER* of all candidates of a sample: *RESPER*:

$$RESPER_i = \frac{SPER_i}{\frac{1}{M} \cdot \sum_{j=1..M} SPER_j}$$

where *M* is the total number of fusion transcript candidates for a sample. Since this quantity is independent from the fragment size, it is more suitable for comparisons across samples. Also, as long as the sequencing depth increases, *RESPER* is expected to increase for a real fusion transcript compared to an artifactual one (Figure 4.5B).

In the case of sufficient coverage, we can also integrate the information related to the junction-sequence identifier analysis, such as the number of single-end reads supporting a junction as well as how evenly the single-end reads covers it. Ideally, the entire fusion junction should be uniformly covered by the reads. If this does not occur, the chimeric transcript might have been generated during sample preparation and the PCR amplification step resulted in an over-representation of that transcript. However, definitive determination of uniform coverage requires great sequencing depth.

4.4.9 Computational complexity

One of the main issues to address is the computational complexity of processing RNA-Seq data. Computationally, the three modules have different requirements. The fusion transcript detection module depends on the total number of mapped reads. Once the alignment is performed, it takes about 15 minutes to run this module on 20 million mapped PE reads using one core of a dual 2 Intel® Xeon® CPU E5410 @ 2.33GHz (4 cores each, for a total of 8 cores), with

6 MB cache, 32 GB RAM, and 156 GB of local disk. The filtration cascade module takes about 15 to 30 minutes to run on the same architecture. The difference depends on the number of candidates initially identified. A more intensive effort is required for the junction-sequence identifier analysis, the main bottleneck being the indexing of all the virtual tiles. The time complexity also depends on the size of the region being tiled. T he alignment of the reads after the indexing is much less computationally demanding. In fact, the time to complete a junction-sequence identifier analysis for a single candidate in both directions, AB and BA, ranges from about 90 to 180 minutes if run on a single machine. However, by splitting the fusion junction library in different files, it is possible to run the indexing step in parallel, thus substantially decreasing the time complexity. Indeed, by splitting the fusion junction library in files with 2M elements, it is possible to complete the indexing and the mapping in about 15 minutes for both orientations.

4.4.10 Report of the analysis results

We also developed a set of tools to report the analysis results through a web interface and the UCSC Genome Browser (Figure 4.9) [41]. All programs of FusionSeq take as input one of the standard formats we defined, and additional tools convert them in files that can be interpreted by the UCSC Genome Browser such as WIGGLE, BED or GFF. This integration is facilitated by the use of a web interface to interrogate the samples. The user selects the sample and the list of potential candidates is shown with the candidates sorted according to *DASPER* (Figure 4.9A). Information regarding the genes involved, such as gene symbols (including aliases), gene description and genomic coordinates are also reported (Figure 4.9B). By clicking on the genomic coordinates the corresponding UCSC Genome Browser page is displayed. Also, each candidate has a detailed page reporting detailed information, including the junction-sequence identifier analysis results (Figure 4.9C).

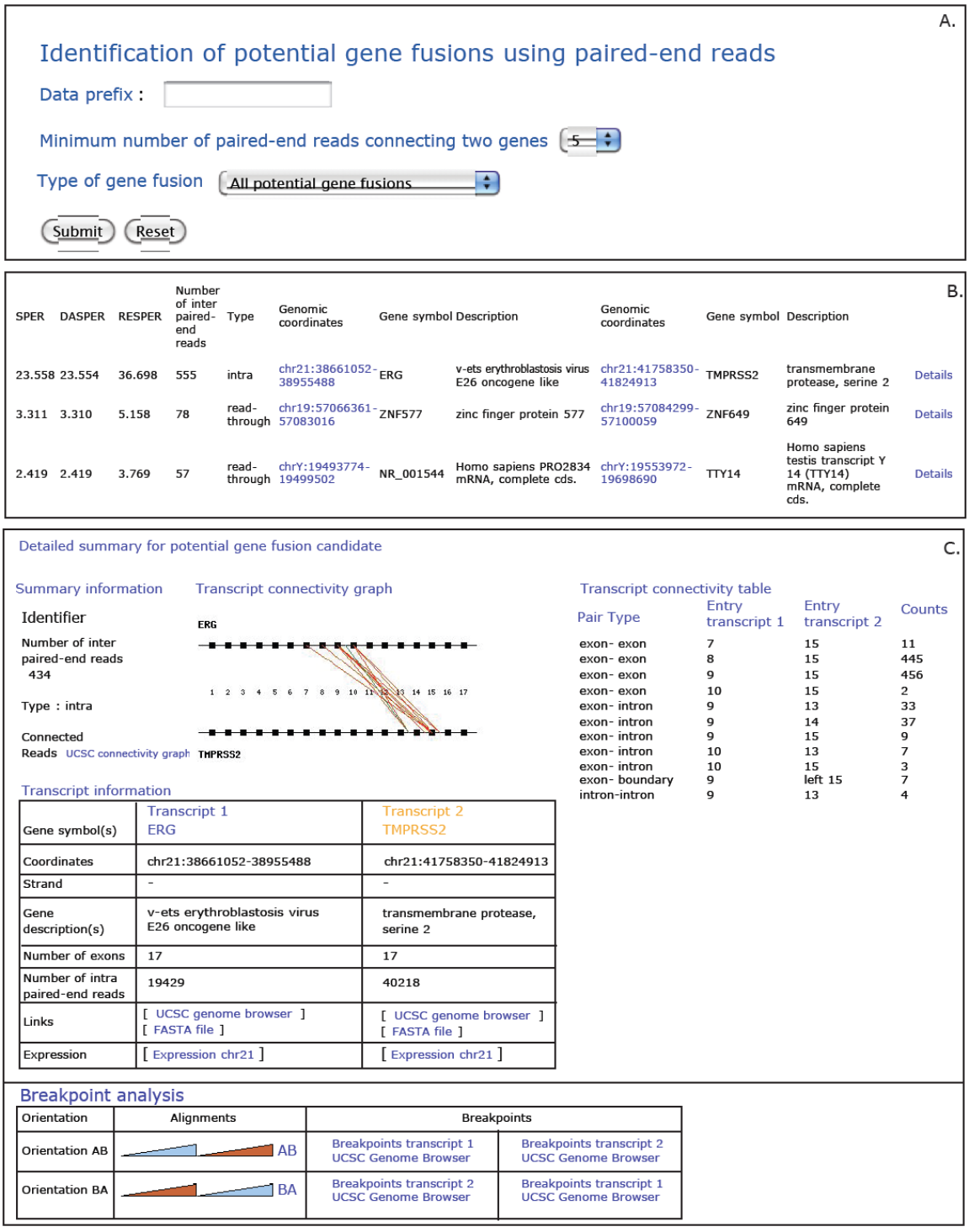


Figure 4.9 Snapshot of the FusionSeq web-interface. (A) Sample name, minimum number of PE reads and type of the fusion candidates can be selected. (B) The list of the candidates is reported along with the statistics we introduce with FusionSeq. The hyperlink provides direct access to the UCSC Genome Browser in order to display the location of the genes. (C) Each candidate has a detailed page with additional details about the fusion, such as the number or inter-transcript reads and the connectivity among exons. If the two genes are located on the same chromosome, a hyperlink points directly to the UCSC Genome Browser and shows all the reads. Moreover, the junction-sequence identifier results can be accessed by clicking on the icons showing the possible combinations, AB or BA. Furthermore, the expression signal of the chromosomes can be loaded as additional track on the Genome Browser.

Although we extensively rely on the data format of the UCSC Genome Browser, it is not possible to show the results for inter-chromosomal fusions (i.e., those between genes on different chromosome) since it can display only one chromosome at the time. In order to address this issue we developed SeqViz, an application that is based on Circos, an open source software that is particularly suited to the display of genomic information by representing the full genomes as a circle [93]. An example of a Circos image can be found in Figure 4.6B. Among the main features of Circos is the high flexibility in adding and showing many types of information: connection between the two ends of a PE read, gene annotation sets, expression values, etc.

Chapter 5

DupSeq: a computational framework for assessing the transcriptional activity of highly similar genomic sequences

Abstract

One of the principal objectives in modern genomics is to accurately measure the levels of transcription from particular regions of the genome. The reliable determination of expression levels from specific genomic elements forms the pillar for many downstream functional analyses, which are of interest from both a basic science and a clinical perspective. Many of these studies have been performed on the genomes of complex organisms, which are abundant in DNA sequences with high mutual similarity, such as that seen between a parent gene and its pseudogene(s), any two members of a gene family, and non-unique sequences which span unannotated regions. Assessing the transcriptional activity of such a genomic element can be challenging because of the difficulty in discriminating between true transcription and potential artifacts, which may result from spillover effects from the expression of a highly transcribed region. To address this issue, we developed DupSeq, a computational framework that performs a set of statistical comparisons between the signal patterns obtained from RNA-Seq reads for similar genomic regions across different tissues. DupSeq has been shown to enable the determination between true transcription and experimental artifact for a number of pseudogenes, and we plan to extend its utility to a broader range of repetitive sequence elements.

5.1 Introduction

5.1.1 Background

It has long been a considerable challenge to distinguish between genuinely transcribed regions and potential artifacts for those genomic elements sharing high degrees of sequence similarity. This problem first emerged in tiling array studies, in which a probe for a specific genomic sequence would cross-hybridize with an off-target element with a similar sequence, and this issue has been studied extensively [6, 34, 35]. These artifacts may lead to highly inaccurate measurements when assessing the transcriptional activity of genomic elements with high degrees of sequence similarity. For instance, the high transcriptional levels of a gene may lead assays to erroneously report that an untranscribed duplicate copy of that gene is being expressed. In other words, the untranscribed duplicate mirrors the transcription signals of the gene as a consequence of their sequence similarity. This problem is especially characteristic in species with very large genomes, which are marked by a great deal of sequence redundancy. For instance, repetitive sequences comprise at least 50% of the human genome [97]. Specifically, roughly 45% of the human genome belongs to the class of transposon-derived repeats, while another 5% represents segmental duplications [97, 98]. Segmental duplications are relatively recent events, in which 1-200kb blocks of genomic sequences are copied from one genomic location to another. As a result, these elements share a high degree of sequence similarity (at least 90%). A recent study has shown that genes, paralogs, as well as pseudogenes are enriched segmental duplications [99], which further highlights the need to assess the transcriptional activity of sequences sharing high sequence similarity.

Although next-generation DNA sequencing technology, as applied to transcriptome profiling (i.e., RNA-Seq), has led to many improvements with respect to resolution and the ability to reliably and accurately detect most transcript isoforms [1, 21, 100], the challenge of accurately assessing the levels of transcription for genomic elements with high sequence similarity remains

unresolved. This is because the reads obtained from RNA-Seq experiments must be aligned to the reference genome, and in some instances, sequencing errors cause reads from highly expressed genes to be mistakenly mapped to untranscribed regions with high sequence similarity, such as pseudogenes or paralogs. Therefore, naïve methods that do not account for these issues often lead to inaccurate results.

In order to address this challenge, we have developed DupSeq, a computational framework designed to statistically analyze and compare the transcription signal patterns (as obtained from mapped RNA-Seq reads) across multiple samples. When comparing signals for a given sequence across different tissues, truly transcribed regions of a particular genomic element are characterized by distinctly different expression patterns relative to those with high sequence similarity, whereas concordant patterns are indicative of mapping artifacts (as shown in Figure 5.1). That is, highly similar regions with discordant expression patterns are transcribed, whereas mapping artifacts are unlikely to give rise to such discordant patterns.

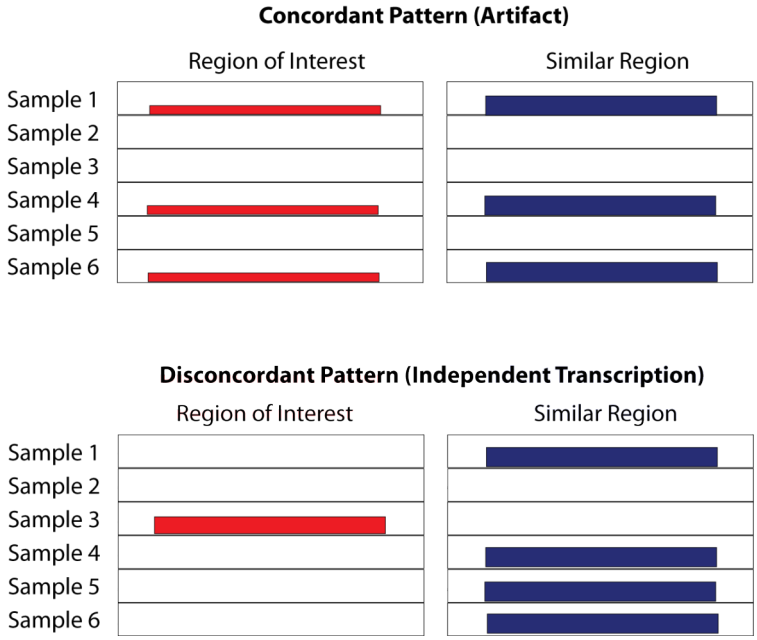


Figure 5.1 Schematic representations of concordant and discordant expression patterns. The top panel shows a region of interest that mirrors the expression pattern of a highly similar genomic region. Such concordance is indicative of a mapping artifact. Conversely, the bottom panel illustrates an example, in which a region of interest exhibits a discordant expression pattern across multiple samples, suggesting independent transcription.

5.1.2 Pseudogenes

Pseudogenes are traditionally defined as defunct genomic elements that share sequence similarity with functional genes, but which lack coding potential as a result of disruptive mutations, such as frame shifts or premature stop codons [101–104]. The conventional thinking surrounding these elements has long dictated that they are entirely devoid of functionality. However, the concept of a "dead" pseudogene may not be an adequate description; there is evidence that some are transcribed, leading to speculation that they may in fact serve functional roles. They may acquire new functions, either as transcribed RNAs or even translated peptides, and some have reported on apparently "revived" pseudogenes. More recently, studies have shown that, in some cases, the mRNA products transcribed from these elements are capable of performing crucial regulatory roles themselves [105–108].

Pseudogenes are usually identified by the rapid accumulation of mutations, such as those which give rise to premature stop codons, and they may be classified as belonging to one of three categories on the basis of how they are formed in the first place: (1) duplicated (or “unprocessed”) pseudogenes are derived from the duplication of functional genes (i.e., the parent genes), (2) processed pseudogenes are created through the retrotransposition of mRNA from functional protein-coding loci back into the genome, and (3) unitary pseudogenes arise through in situ mutations in previously functional protein-coding genes [101, 104, 109, 110]. Depending on their class, pseudogenes exhibit distinct genomic features. As may be expected, duplicated pseudogenes have intron/exon-like genomic structures, and may have inherited upstream regulatory sequences from their parents. In contrast, processed pseudogenes, having lost introns when the mRNAs from which they are derived were spliced, lack both the introns and the upstream regulatory sequences of their parent genes. Processed pseudogenes may preserve evidence of their insertion into the genome, in the form of poly-A features at their 3’ ends.

5.1.3 Paralogs

Like pseudogenes, paralogs constitute a class of genomic elements with high sequence similarity as a consequence of the evolutionary processes from which they are derived. Specifically, paralogs are a subclass of homologs, and they originate when a single sequence is duplicated. This process can occur multiple times, thereby producing a paralog family. Therefore, depending on the length of time which elapsed since duplication, paralogs of the same gene family may share a very high degree of sequence similarity. The initial redundancy in the functionality between the paralog and the gene from which it is derived places less evolutionary pressure on one paralog, thereby better enabling it to assume novel functionality. Thus, although the paralog and the original copy of the gene may share very similar sequences, these genomic elements may not necessarily exhibit similar functionality.

5.1.4 Applying DupSeq to investigate pseudogenes, paralogs, and novel unannotated regions

It should be noted that, in the context of DupSeq, it is significantly easier to discriminate between the transcriptional activity of pseudogenes and parent genes than between related paralogs. This is a consequence of the very nature inherent to the relationship between a pseudogene and its parent. Thus, expression signals from the pseudogene are at first assumed to be artifacts, as only the parent genes is expected to be expressed. Conversely, it is more difficult to define the expected transcription levels of related paralogs. There is no clear definition for which element among paralogs should serve as a benchmark for expression, and which elements should be compared to that benchmark, as it is assumed that all paralog members are equally likely to be expressed, without a priori knowledge of particular members' functions or transcriptional behaviors.

Furthermore, it should be noted that DupSeq can be applied not only to pseudogenes and paralogs, but also to novel unannotated transcribed regions. Though there has been much work devoted to understanding the functions of these elements, they have been difficult to annotate because they often have sequence-similar regions throughout the genome, thereby confounding

measures of their transcriptional activity. For many of these unannotated regions, DupSeq provides a powerful novel means of comparing expression levels, thus providing a first step toward assigning functionality and annotation. Thus, DupSeq can be generalized to process user-defined genomic regions to assess transcriptional activity.

5.2 Methods

5.2.1 Overview of modular implementation

DupSeq is implemented in C for efficiency. As shown in Figure 5.2, DupSeq is a modular framework that comprises three main modules. The first module uses the set of genomic elements under study to identify all the regions with which the member elements of that set share high sequence similarity. The second module processes the various RNA-Seq data sets, including mapping the reads and generating the signal tracks associated with those mapped reads. The third module, which is at the core of DupSeq, utilizes the output of the previous two modules to statistically evaluate the transcriptional activity of these regions of interest, and providing confidence scores for true transcription on the basis of synchronous or discordant signal patterns across the different tissues.

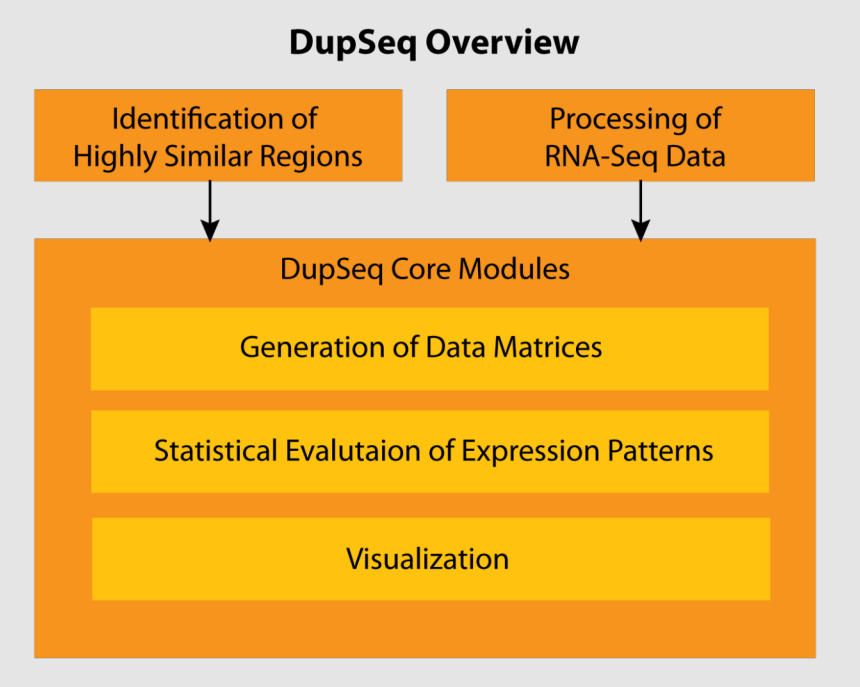


Figure 5.2 Schematic overview of DupSeq. DupSeq is a modular framework comprising three different modules. The first identifies the genomic elements that share a high level of sequence similarity with a given set of specified regions. The second processes the RNA-Seq data by mapping the reads and then generating the signal tracks associated with those mapped reads. Lastly, the core module uses the output of the previous two modules in order to statistically evaluate the transcriptional activity of the regions of interest and to visualize the results.

5.2.2 Module I: Identification of highly similar regions (BLAT alignments)

DupSeq contains a utility to extract the genomic sequences from a set of coordinates representing each element of interest. The current implementation of DupSeq uses BLAT [95] to align each sequence to the reference genome in order to find all genomic regions with similar sequences. It should be noted that BLAT provides a fast way to identify all regions which share at least 90% sequence similarity with the input sequence. Hence, genomic regions that are evolutionarily more distant will not be detected. However, this is inconsequential, as the objective is to identify regions of very high sequence similarity; as discussed, these regions are those which lead to mapping artifacts. Also, the modular design inherent to DupSeq more easily enables the substitution of different mapping programs.

Each alignment pair is then assigned to one of four categories:

1. Entries without any other similar regions in genome
2. Entries giving rise to only one alignment pair
3. Entries with 2 to 5 alignment pairs
4. Entries giving rise to more than 5 sequence alignment pairs

For instance, pseudogenes belonging to the first category have presumably evolved a great deal as a result of their age, thereby conferring them with sequences which are highly divergent from those of their parent genes, and the single alignment in the second category is most likely the parent gene itself.

Elements belonging to the fourth category may be difficult to assess in the framework of DupSeq, as the composite signals from elements with a very high number of sequence-similar regions can be difficult to deconvolute. Alternative methods (including different mapping strategies used to align the RNA-Seq reads) would most likely be required to assess the transcriptional activity of such elements.

5.2.3 Module II: Processing the RNA-Seq data

In order to check for evidence of transcription, DupSeq uses RNA-Seq data sets from multiple tissues to evaluate and compare the expression patterns for each analyzed element and regions with highly similar sequences. Bowtie is used to align the reads from each tissue to both the splice junction library and the reference genome [94]. In the context of this framework the ways in which reads are mapped are essential, and a number of factors must be considered. For instance, when DupSeq is used to analyze pseudogenes, it is important to map these reads to the genome and the splice junction library simultaneously. Otherwise, those reads from the parent gene which span splice junctions would mistakenly be mapped to the pseudogene, thereby falsely providing wrong measurements of pseudogene expression (see Figure 5.3). After the

alignment step, the mapped reads for each tissue are converted to signal tracks using RSEQtools [25]. These signal tracks are then used as input to the third module of DupSeq.

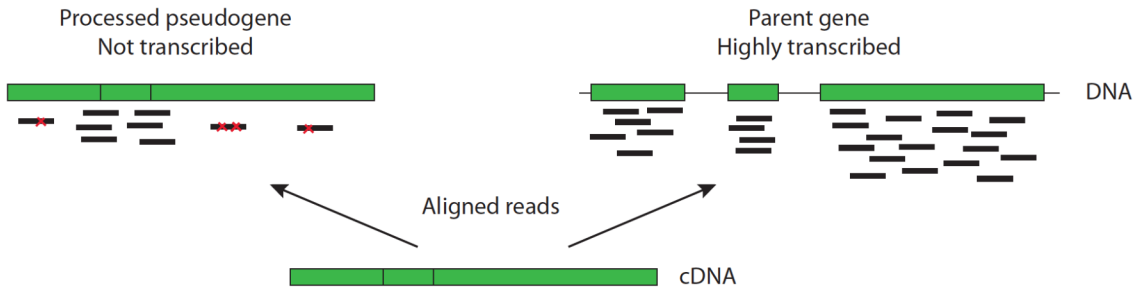


Figure 5.3 Mapping RNA-Seq reads without a splice junction library. In the context of pseudogenes, using a splice junction library during the alignment step is essential. Reads from the parent gene, which span splice junctions, would otherwise mistakenly be aligned to the pseudogene, thereby providing inaccurate measurements of the pseudogene's expression. In addition, sequencing errors, as denoted by red crosses, may cause reads to mistakenly be mapped to the pseudogene.

5.2.4 Module III: Statistical evaluation of expression patterns

The third module of DupSeq includes a utility to merge the information from Modules I and II by extracting the signal track information (representing the mapped RNA-Seq reads) across multiple samples for each alignment pair, as well as a set of programs to perform statistical calculations on these matrices. For example, given an alignment pair (i.e., the region of interest and a matching region of high sequence similarity), one of these programs calculates basic statistics for the signal tracks of each sample. Another program calculates correlation coefficients between the expression values of the region of interest and matching region along every position of the alignment pair. Along with other metrics, these statistics are then used to attribute the original data to independent transcription or mapping artifacts.

It should be noted that we developed a custom data structure to represent the matrices described above, as well as a complementary set of functions to facilitate it's processing. Two key functions include the reader and writer, which enable basic I/O operations, as well as pipelining the multiple programs.

In addition, we have developed and implemented a novel approach for visualizing the expression patterns and statistics associated with the alignment pair. Previously, this step had to be performed manually as there is currently no software available for these tasks. A number of factors complicate this task. First, the region of interest and that to which it matches may be on different chromosomes. Second, these two regions may be located on different strands, and the signal tracks must therefore be inverted for comparison. Also, this visualization tool provides a way of including gene annotations in order to represent those sub-regions of a gene or transcript isoform that were initially aligned.

5.3 Results

5.3.1 Case 1: Application of DupSeq in the context of worm pseudogenes

The modENCODE Project employed DupSeq to investigate the transcriptional activity of 1,198 pseudogenes with identified parent genes [36]. This study used a previously established RNA-Seq read mapping methodology [45], in which MAQ [111] and Crossmatch were used to align all reads to the genome, splice junctions, spliced leaders, and polyA libraries. The best match was selected with a slight bias favoring genome alignments first; reads with equal matches to the genome and other databases were placed against the genome, and the expression values (DCPM) were calculated from these reads [45]. Pseudogenes were considered transcribed if their DCPM exceeded 0.04 in one or more samples. Note that this cutoff is 100-fold higher than the minimum DCPM in the set. Using this approach, we identified 323 pseudogene candidates with evidence of transcriptional activity.

As a first step toward evaluating potential pseudogene transcription, pseudogenes were assigned to one of three categories. The first constitutes pseudogenes exhibiting expression levels which exceed those of their respective parent genes by at least a factor of two. The second consists of those pseudogenes with expression patterns which are asynchronous from those of their respective parent genes across tissue samples (see Figure 5.4). Both of these classes constitute properties suggesting that the pseudogene signal pattern represents

transcription that is truly independent from that of its respective parent gene, rather than mapping artifacts. The last category includes those cases for which the pseudogene signal profile is concordant with that of the parent gene across the multiple tissue samples, suggesting the possibility of a mapping artifact (i.e., such concordance by itself does not exclude the possibility of mapping artifacts). 191 of the 323 candidates were found to belong to the first two categories (87 and 104, respectively). These 191 pseudogenes are thus likely transcribed independently from their parent genes.

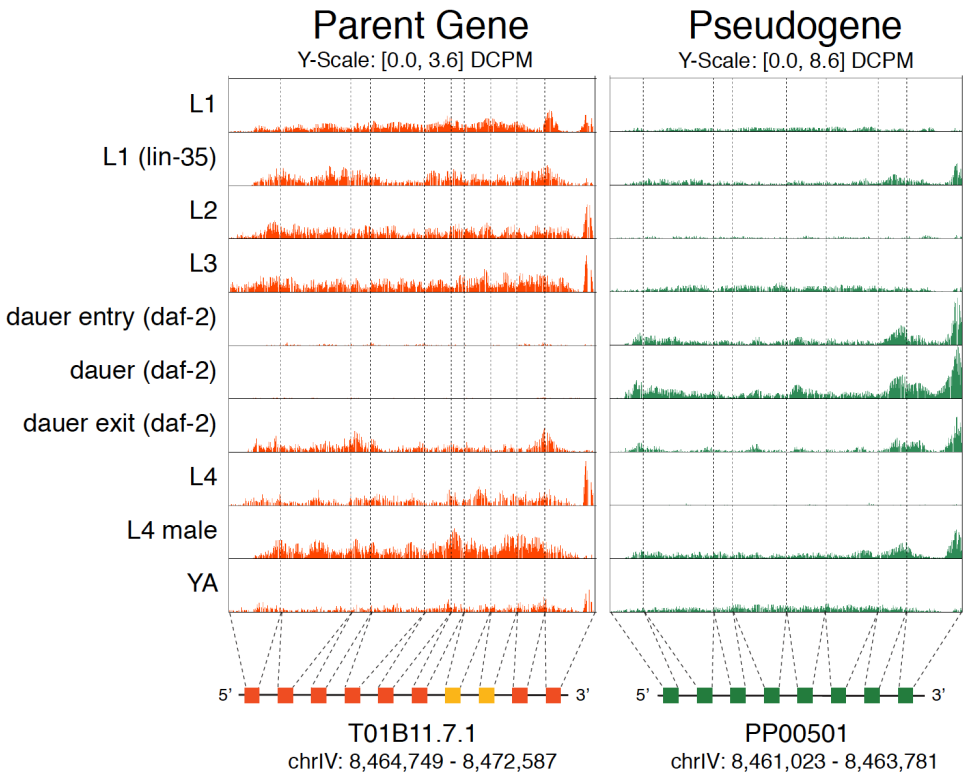


Figure 5.4 Example of a differentially transcribed pseudogene in *C. elegans*. Rows are normalized signal tracks for the various developmental stages, showing the expression pattern of the parent gene (T01B11.7.1; orange) and an associated duplicated pseudogene (PP00501, green).

5.3.2 Case 2: Application of DupSeq in the context of human pseudogenes

Using the GENCODE Pseudogene Resource, we examined the transcriptional activity of the GENCODE pseudogenes across 16 human tissues, as provided by the Illumina Human Body Map RNA-Seq data sets. The genomic coordinates of both the processed and duplicated pseudogenes were extracted and aligned to the human reference genome to identify the regions with high sequence similarity (note that there were 8,107 processed and 1,860 duplicated pseudogenes; see Methods for details). Each pseudogene alignment was assigned to one of four categories (for details, see the category descriptions provided in the section describing Module I under Methods). Out of the total 9,967 pseudogenes, we found that 3,198 belong to the first category, 1,907 belong to the second category (see Figure 5.5A for an example of a transcribed pseudogene), 2,150 belong to the third category, and 2,712 belong to the fourth category.

The set of 3,198 pseudogenes without similar regions was reduced to 344 by requiring that each pseudogene be covered by at least two reads across half of its length in at least one tissue. This filtered subset of potentially transcribed pseudogenes was then selected for validation by RT-PCR, followed by Illumina sequencing.

Finally, we note that artifacts in the signal patterns can sometimes be identified by visual inspection. As shown in Figure 5.5B, artifacts emerge as signal patterns that are concordant with those of the parent gene.

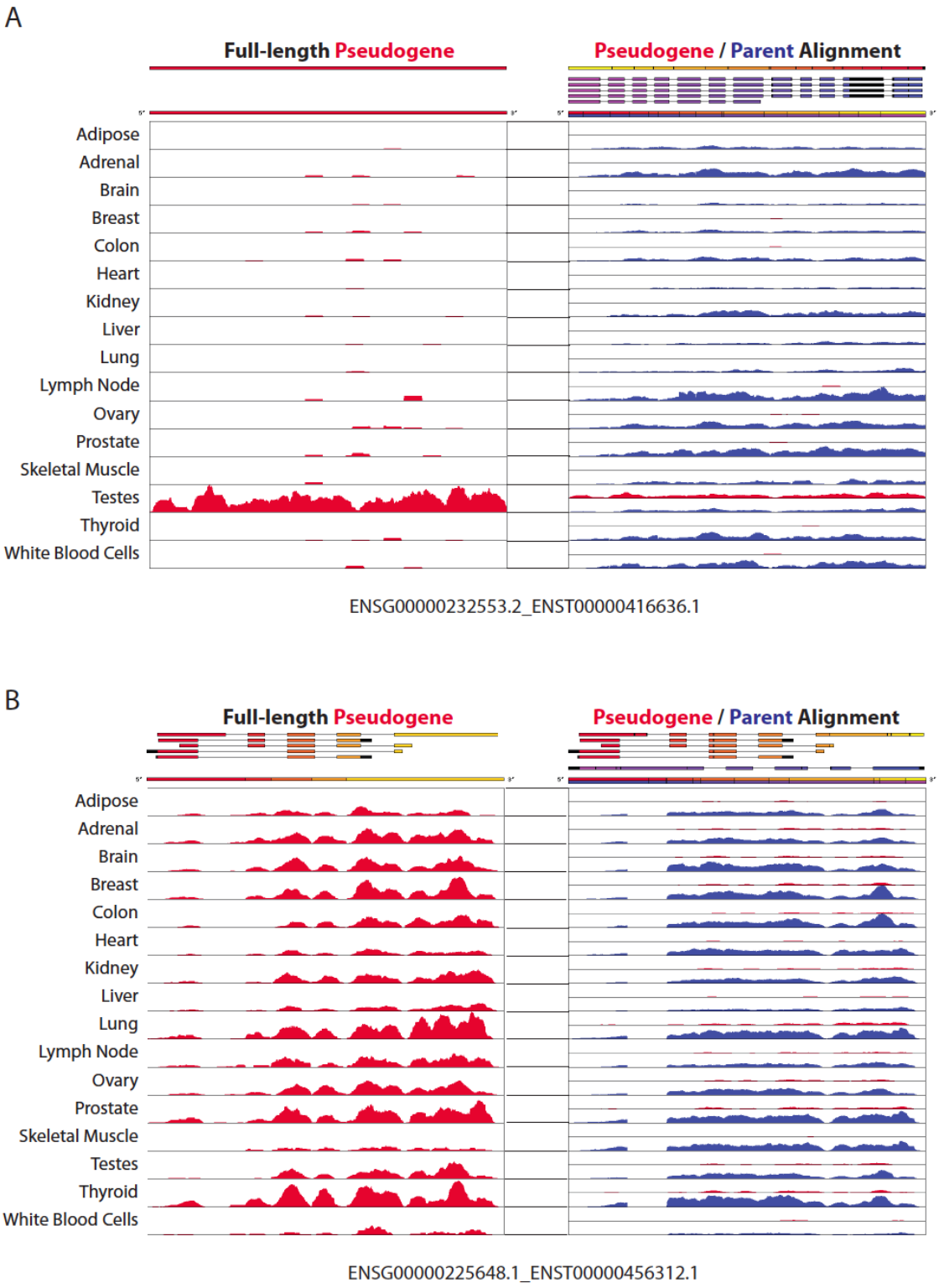


Figure 5.5 Example of a transcribed pseudogene and a mapping artifact. (A) This pseudogene is exclusively expressed in testes, whereas the parent gene is transcribed in a number of different tissues. This constitutes compelling evidence of independent pseudogene transcription. (B) This pseudogene mirrors the expression pattern of the parent gene across multiple tissues, which provides evidence of a likely mapping artifact.

5.4 Conclusions

Accurately measuring the transcription of the specified regions in a given genome has been a canonical pursuit in modern biology, and has led to many important insights. However, measuring the transcription levels associated with certain elements has remained a formidable challenge when those elements share very similar sequences in the genome. In particular, it may be difficult to determine which among these similar elements are truly transcribed or the result of artifacts. Such artifacts arise when silent regions capture some of the signal from truly transcribed elements. This can sometimes lead to the interpretation of signal patterns as evidence that a particular region is transcribed, even though it may in fact be silent. In other cases, with the possibility of artifacts in mind, the investigator sometimes erroneously dismisses apparently expressed regions as false positives when such regions are expected to be silent. Pseudogenes, which are assumed to be silent, serve as an example of the latter; transcription signals observed for a given pseudogene are usually attributed to artifacts, even though some have been shown to perform important regulatory roles [105–108]. Not only has it been difficult to discriminate between true transcription and artifact, but increasingly, this distinction must also be made for data on a large scale. Thus, an automated approach to discriminating between true transcription and artifacts is needed as a first step toward the reliable detection of novel functionality.

To address this need, we designed DupSeq, an efficient and modular piece of software for assessing the transcriptional activity of highly similar sequences. The principle behind DupSeq is to first align sequences of interest to a reference genome in order to identify all regions with highly similar sequences. RNA-Seq data from various samples is then used to generate transcription signal patterns associated with these identified genomic regions.

The novel feature of our approach will be to statistically evaluate the resultant signal patterns across these different samples in order to provide confidence scores associated with true transcription or artifact, which may result when the errors in RNA-Seq cause reads to be aligned to a similar region. The guiding principle behind the statistical framework is to

discriminate between concordant and discordant signal patterns across these various signal tracks. In particular, concordant signal patterns are suggestive of artifacts, whereas discordant signal patterns provide evidence of independent transcription. Figure 5.1 provides a schematic representation of this idea.

As a proof-of-principle, we applied DupSeq to investigate the transcriptional behavior of pseudogenes in two different organisms. Specifically, we identified a number of transcriptionally active pseudogenes in the human genome, which were then validated experimentally by RT-PCR followed by Illumina sequencing. In addition, DupSeq was also used by the modENCODE consortium to examine the pseudogene expression in the *C. elegans* genome [36].

Importantly, as DupSeq operates on any set of elements with similar sequences, it may easily be extended to study not only pseudogenes, but also other types of elements with mutually similar sequences, including paralogs and novel unannotated regions. We first plan to better equip DupSeq to operate on such elements by enhancing its statistical methodologies. We thus anticipate that our resource will serve as a valuable contribution to the genomics community, and will lead to the assignment of novel functionality to many regions that have thus far eluded investigation.

Chapter 6

Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing

The work described in this chapter was adapted from a manuscript, which was originally published in *PNAS* [37].

Abstract

To examine the fundamental mechanisms governing neural differentiation, we analyzed the transcriptome changes that occur during the differentiation of hESCs into the neural lineage. Undifferentiated hESCs as well as cells at three stages of early neural differentiation-N1 (early initiation), N2 (neural progenitor), and N3 (early glial-like)-were analyzed using a combination of single read, paired-end read, and long read RNA sequencing. The results revealed enormous complexity in gene transcription and splicing dynamics during neural cell differentiation. We found previously unannotated transcripts and spliced isoforms specific for each stage of differentiation. Interestingly, splicing isoform diversity is highest in undifferentiated hESCs and decreases upon differentiation, a phenomenon we call isoform specialization. During neural differentiation, we observed differential expression of many types of genes, including those involved in key signaling pathways, and a large number of extracellular receptors exhibit stage-specific regulation. These results provide a valuable resource for studying neural differentiation and reveal insights into the mechanisms underlying in vitro neural differentiation of hESCs, such

as neural fate specification, neural progenitor cell identity maintenance, and the transition from a predominantly neuronal state into one with increased gliogenic potential.

6.1 Introduction

Neural commitment and subsequent differentiation is a complex process. Although the complexity of RNAs expressed in neural tissues is very high [112, 113], a comprehensive analysis of the genes and RNA isoforms that are expressed during the different stages of neural cell differentiation is largely lacking. Such information is expected to be important for understanding mechanisms of neural cell differentiation and ultimately providing therapeutic solutions for neural degenerative diseases, such as Parkinson's and Alzheimer's disease.

Our current knowledge of the mechanisms involved in neural cell formation is derived mostly from studying neurogenesis in the developing embryos of animal models [114, 115]. However, neurogenesis in animals is a complex process involving many different cell types that differentiate asynchronously. This heterogeneity, along with the relatively small number of cells that can be readily obtained, makes the analysis of the temporal differentiation of individual cell types extremely difficult. One solution is to analyze hESCs during in vitro differentiation to different stages of neural development, which can be performed using a relatively large numbers of cells [116–120]. Analysis of the transcriptome in these cells is expected to provide insights into the mechanisms and pathways involved in early cell fate specification, such as the acquisition of neurogenic potential and the transition to gliogenic potential, which may ultimately be extremely useful for pharmacologic screening and neurodegenerative disease therapies.

Many high-throughput methods have been used previously to study global transcription [2, 121–124]. The recent development of massively parallel sequencing of short reads derived from mRNA (RNA-Seq) makes it possible to globally map transcribed regions and quantitatively analyze RNA isoforms at an unprecedented level of sensitivity and accuracy [1, 16–19, 56, 58, 125]. Although the use of short reads enables detection of transcribed regions and spliced

adjacent exons, it has limitations. In particular, the relationship of nonadjacent exons and multiple exons within the same transcript cannot be deduced.

In this study we combined the strengths of several massively parallel sequencing technologies, including short Illumina single and paired-end reads (sequence reads from both ends of cDNA fragments; 35-bp reads) and longer Roche 454 FLX and Titanium sequencing reads (250–450-bp reads) to discern transcript structure and analyze transcriptome complexity at an unprecedented level [125–127]. We applied these technologies to the analysis of early stages of neural differentiation of hESCs. Our results revealed an extraordinary degree of stage-specific transcription and splicing. From more than 150 million uniquely mapped sequence reads, we found thousands of unannotated transcriptionally active regions (TARs) and unannotated isoforms. Some unannotated TARs and splice isoforms are transcribed only at particular stages, implying functional roles in specific steps of neural differentiation. Moreover, we describe a phenomenon we call isoform specialization, whereby splicing isoform diversity is the highest in undifferentiated hESCs and decreases in cells undergoing neural differentiation. Finally, the characterization of dynamic changes of gene transcription levels has provided important insights into the in vitro neural differentiation of hESCs with regard to neural specification, neural progenitor identity maintenance, and the transition from a predominantly neuronal nature to one with increased gliogenic potential.

6.2 Results

6.2.1 RNA-Seq at specific stages of neural differentiation of hESCs

We characterized changes in the transcriptome profiles during early human neural differentiation using H1 hESC cultures. Two differentiation strategies were used (Figure 6.1, Materials and methods).

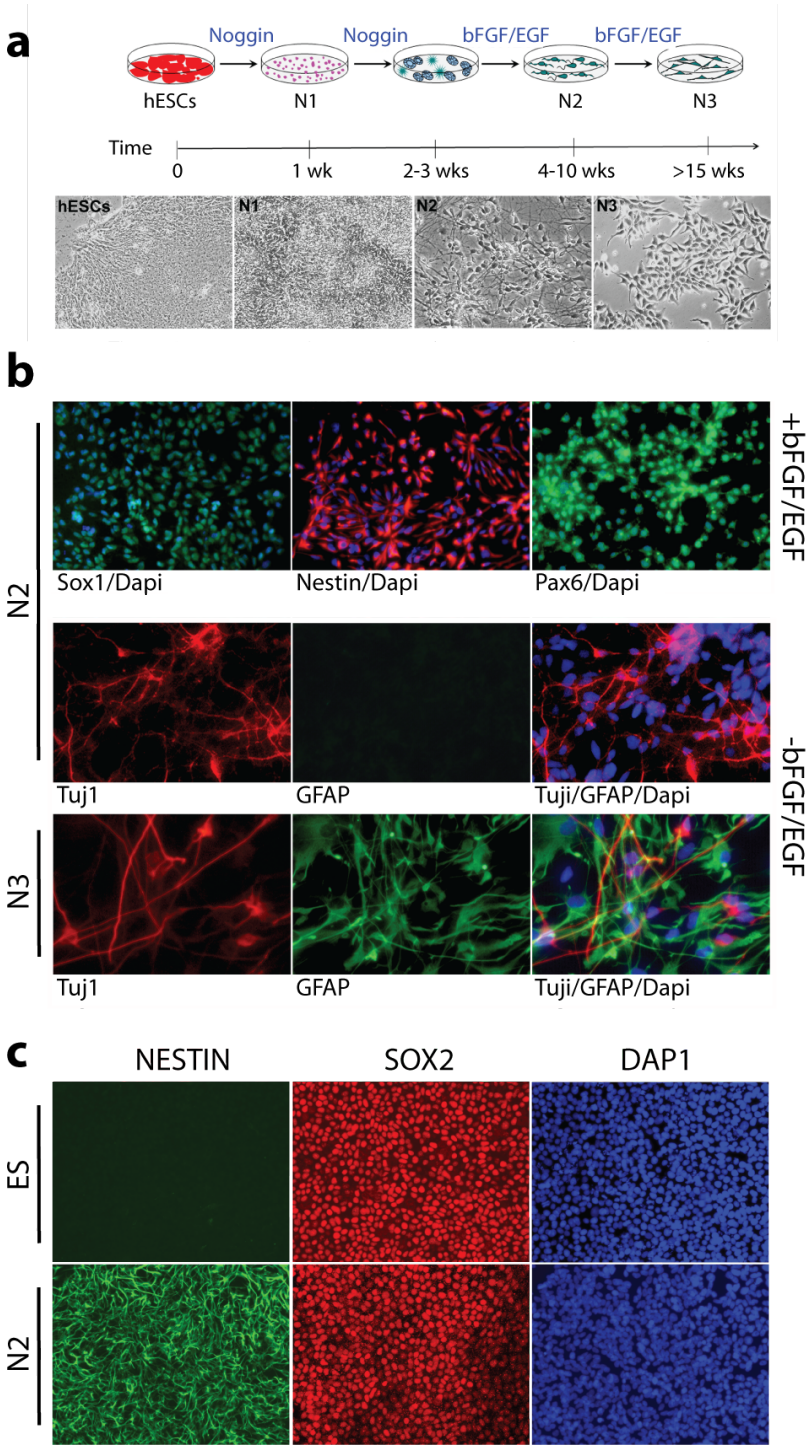


Figure 6.1 Characterization of neural differentiation cell cultures. (A) Schematic representation of the neural differentiation procedure and four stages by approach A: hESCs, N1, N2, and N3. (B) Immunofluorescence labeling of genes specifically expressed in N2 and N3 cells (H1 hESCs) prepared by approach A with or without growth factors. SOX1, NESTIN, and PAX6 are expressed in N2 cells with growth factors bFGF/EGF. TUJ1 is expressed in N2 cells without growth factors; GFAP is not expressed. GFAP is expressed in N3 cells after growth factor withdrawal, whereas TUJ1 expression is still visible. Blue: nuclei are stained by DAPI. (C) Immunostaining characterization of cell cultures (H1 hESCs) by approach B. SOX2 is expressed in hESCs but not NESTIN. Both NESTIN and SOX2 are expressed in N2 cells.

In approach A, the hESC H1 line was differentiated and cultured using feeder-free chemically defined adherent cell culture system through three stages: N1, an initiation stage; N2, a neural progenitor cell (NPC) stage that produces only neurons upon further differentiation; and N3, which produces both neurons and glial cells (Figure 6.1A and B) [118, 119, 128]. In approach B, neural progenitors (N2-B) were generated from undifferentiated H1 hESCs via embryoid body-like neurosphere formation [120]. In each case, we used standard protocols involving bone morphogenic protein signaling antagonists (Noggin) and basic fibroblast growth factor (bFGF) (Materials and methods).

According to qualitative and quantitative analyses, the differentiation protocols were highly reproducible. The derived cell populations from each preparation were characterized by both immunoassays and FACS analysis for a large variety of markers to ensure that the cell cultures were highly homogeneous at the various stages (Figure 6.1). (*i*) Undifferentiated hESCs (present in both approaches A and B) expressed all hESC surface antigens (e.g., TRA-1-60/81 and SSEA4) as well as transcription factors OCT4 and SOX2 (Figure 6.1C). (*ii*) N1 initiation stage cells (present only in approach A) lost TRA-1-60/81 expression but were still positive for SSEA4, although at a lower level, and began to express SSEA1. They also expressed OCT4 at a low level. (*iii*) N2 NPCs generated by approach A lost OCT4 as well as SSEA4 expression but expressed NESTIN, PAX6, and SOX1 (Figure 6.1B). The cells showed a typical morphology of NPCs: bipolar with small soma (Figure 6.1A). Glial fibrillary acidic protein (GFAP) was not expressed (Figure 6.1B). Upon withdrawal of growth factors, the N2 cells predominantly differentiated into neurons rather than glial cells, as indicated by the expression of neuronal marker TUJ1 (Figure 6.1B). Similarly, >95% of N2 neural precursors generated by approach B expressed an array of neural markers (neuroepithelial marker PAX6, neural stem cell markers NESTIN, and SOX2, as well as neuronal stem/precursor markers MUSASHI) (Figure 6.1C). The N2 cell populations were negative for pluripotent hESC markers such as OCT4 (<0.01%). (*iv*) N3 stage cells (only present in approach A) exhibited distinct morphology (Figure 6.1), and GFAP was expressed in these cells (Figure 6.1B). After bFGF/EGF withdrawal, more glial cells than

neurons were generated (Figure 6.1B). These events we observed in vitro are reminiscent of in vivo neurodevelopment, in which neurogenesis occurs before the onset of gliogenesis [128, 129]. This phenomenon is also observed in the H7 hESC line.

6.2.2 Integration of short, long, and paired-end RNA-Seq reads

To characterize the transcription of cells at the specific differentiation stages, we generated a combination of 35-bp single reads, 35-bp paired-end reads, and 250–450-bp long reads. The paired-end reads were from cDNA fragments of different lengths, ≈300 bp, 300–600 bp, and 600–1,000 bp.

Approach A				
Number of uniquely mapped reads	hESC-A	N1-A	N2-A	N3-A
35bp Single-End Reads	6,477,731	11,904,041	17,073,257	10,879,155
35bp Paired-End Reads	724,250	594,053	1,285,140	1,044,559
Grand Total	7,201,981	12,498,094	18,358,397	11,923,714
Approach B				
Number of uniquely mapped reads	hESC-B		N2-B	
250-450bp Reads	1,263,516		258,107	
35bp Single-End Reads	60,967,216		33,668,106	
35bp Paired-End Reads	2,046,128		1,655,169	
Grand Total	64,276,860		35,581,382	

Table 6.1 Summary of sequencing reads by cell type.

A total of 140, 15, and 1.5 million uniquely mapped single, paired-end, and long reads, respectively (summarized in Table 6.1), were generated from two to three biologic replicates from each of the differentiation stages (Spearman correlations: 0.94–0.97). The fraction of genes

detected at 1-fold average coverage approached saturation for the sequencing of hESC-B and N2-B cells (Figure 6.2A).

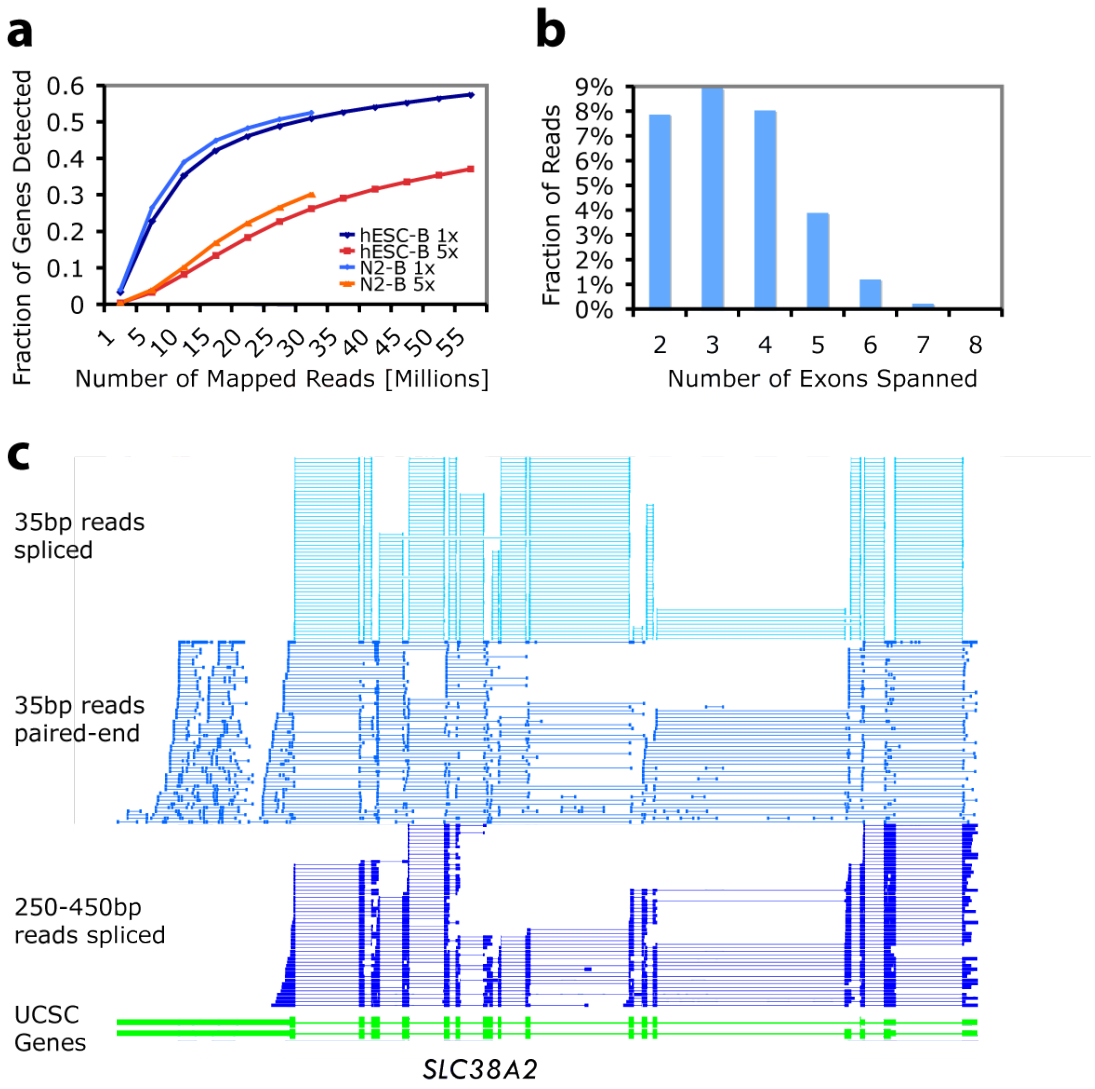


Figure 6.2 Overview of transcript characterization by RNA-Seq. (A) Fraction of genes detected as a function of read depth. (B) Number of exons spanned by 450-bp reads. (C) Transcript complexity revealed by integrating short, long, and pair-end sequence information. Only the spliced reads for single-end and long reads are shown. For paired-end reads the same RNA fragments are shown as two vertical bars connected by a line.

We achieved extensive sequence depth primarily with the 35-bp single reads; the paired-end and 250–450-bp reads provided longer-range exon connectivity information and aided in defining complex splice isoforms. Longer reads, particularly 450-bp reads, can link up to eight

exons (see Figure 6.2B for distribution). Figure 6.2C illustrates the structure of a 16-exon gene that was constructed using a combination of the sequencing technologies.

6.2.3 Identification of unannotated transcribed regions and their connectivity

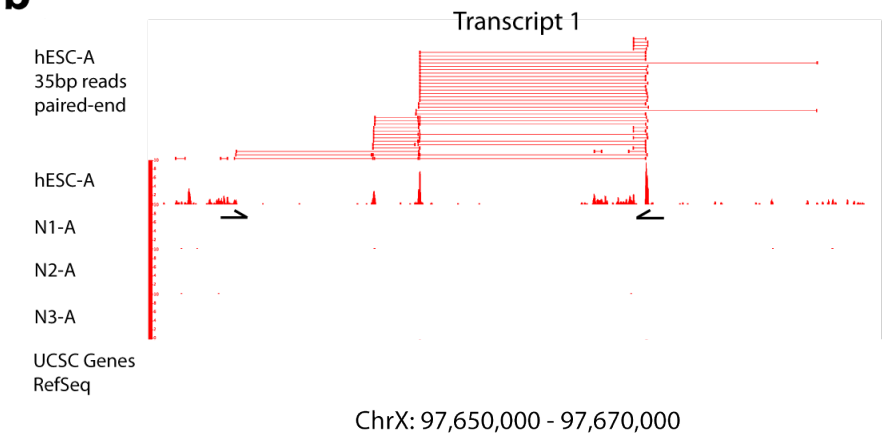
Consistent with our previous studies [2, 124, 130], thousands of unannotated TARs were identified. Specifically, if a TAR overlapped with University of California, Santa Cruz (UCSC) gene annotation it was categorized as "known," and if there was no overlap it was classified as "unannotated." Ninety percent of unannotated TARs were validated by RT-PCR from a random sample of 40 TARs identified from the different stages. We also intersected TARs discovered by our RNA-Seq approach with previously published TARs expressed in the liver that were identified by tiling microarrays [2]. Our study found a large number of hESC RNA that were not identified by the tiling microarray study, consistent with previous observations that RNA-Seq has a higher sensitivity and dynamic range [1]. Interestingly, we also found that a large number of unannotated TARs (35–65%) were specific for each stage; 624, 246, 300, and 353 unique unannotated TARs were found for hESC, N1, N2, and N3 stage cells, respectively (Figure 6.3A – top), raising the possibility that these might have stage-specific functions. Sequences and signal track files can be found in the Gene Expression Omnibus (GEO; accession number GSE20301).

a

Stage	Known TARs	Unannotated TARs	Unique known TARs	Unique unannotated TARs
hESC-A	20,233	1,510	6,497	624
N1-A	14,793	1,182	1,796	246
N2-A	13,538	1,072	2,682	300
N3-A	16,525	1,281	3,299	353

	Known				Unannotated			
Stage	hESC-A	N1-A	N2-A	N3-A	hESC-A	N1-A	N2-A	N3-A
hESC-A	1.00	0.57	0.40	0.53	1.00	0.51	0.37	0.46
N1-A	0.78	1.00	0.56	0.65	0.66	1.00	0.50	0.62
N2-A	0.60	0.61	1.00	0.71	0.52	0.56	1.00	0.63
N3-A	0.64	0.58	0.58	1.00	0.54	0.57	0.53	1.00

b



c

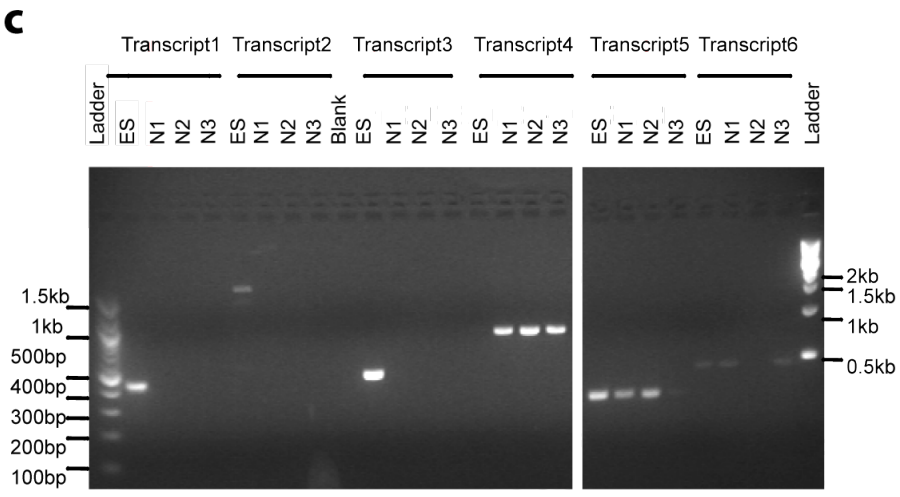


Figure 6.3 Stage-specific unannotated TARs and their connectivity by paired-end reads. (A) (Upper) The number of unannotated, known, and unique TARs found at each stage. TARs that overlap with a UCSC gene are classified as "known." Those that have no overlap with a UCSC gene annotation are called "unannotated." TARs that do not overlap with TARs in other differentiation stages are called "unique" to that particular stage. (Lower) Fractions shared between stages for known TARs and unannotated TARs, respectively. (B) The transcript structure of an unannotated transcript (Transcript 1) that is uniquely transcribed in hESCs is reconstructed using paired-end reads. (C) RT-PCR validation of unannotated transcripts identified by groups of paired-end reads. Transcript 1 and its RT-PCR primer set are shown in B. Transcripts 1–3 are specifically expressed in hESCs, transcript 4 in N1–N3 cells, transcript 5 in ES-N2 cells, and transcript 6 in ES, N1, and N3 cells.

As expected, the majority of the paired-end reads fell within the same known exons. However, a small fraction of paired-end reads linked unannotated TARs to either UCSC-known annotated genes (0.35%, 0.46%, 1.03%, and 0.58% for hESC, N1, N2, and N3, respectively) or to other unannotated TARs (0.36%, 0.38%, 1.50%, and 0.89% for hESC, N1, N2, and N3, respectively). Although the percentage of "linking" reads was low, their large number allowed for unambiguous connection of TARs, and unannotated spliced gene structures could be identified by an overlapping group of paired-end reads. Figure 6.3B shows an unannotated transcript with at least five exons that was uniquely transcribed in hESCs and accurately constructed using a group of overlapping paired-end reads. This transcript and expression pattern was validated by RT-PCR (Figure 6.3C). Twelve such multiexonic unannotated transcripts were further examined using RT-PCR; 11 were validated and 6 were verified to be stage specific (Figure 6.3C).

6.2.4 Alternative splicing during early neural differentiation of hESCs

In addition to cell type–specific gene expression, such as for *OCT4* and *GFAP*, we observed many interesting differentiation-stage-specific alternative splicing isoforms. For instance, an isoform of neural cell adhesion molecule 1 (*NCAM1*) was prevalent at the N3 stage, low level at the N2 stage, but not detectable at the N1 and hESC stages (Figure 6.4A). In addition, an isoform of serine/threonine kinase 2 (*SLK*) was specifically transcribed in hESCs (SI Text), consistent with an independent observation by Gage and colleagues [131]. Although the number of known splice junctions detected in our study was near saturation (Figure 6.4B – top), the number of unannotated splice junctions continued to increase with read depth, indicating that there are many more unannotated isoforms yet to be discovered (Figure 6.4B – bottom).

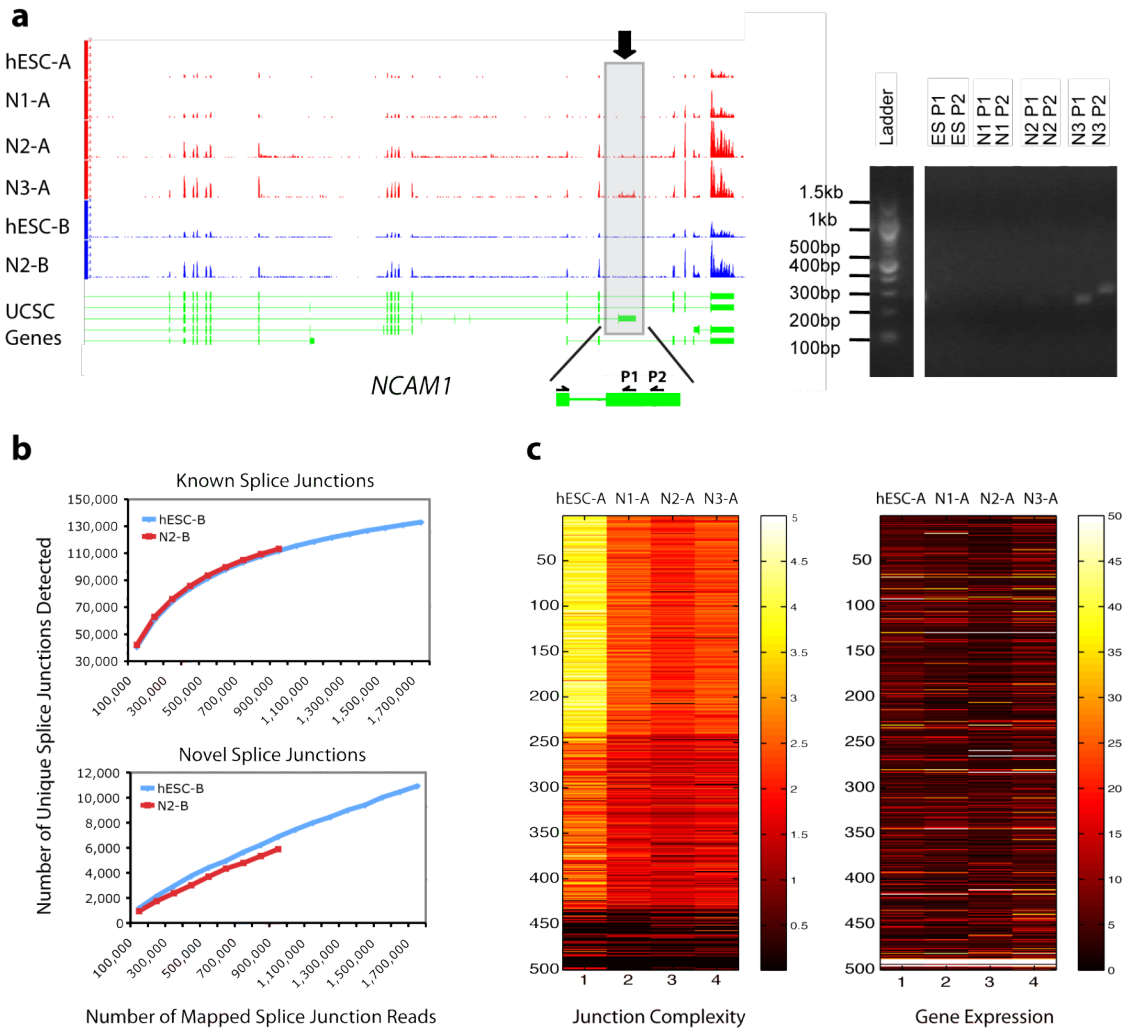


Figure 6.4 Splicing analysis. (A) One of the transcript isoforms of neural cell adhesion molecule 1 (NCAM1) (marked by a rectangle and arrow) is primarily expressed in N3 and very weakly at N2 but not at N1 and hESC stages. The y-axis of the RNA-Seq signal tracks represents the read density normalized by the number of mapped reads per million for each cell type. Two sets of RT-PCR primers were designed on the alternative exon and the adjacent constant exon, which generated two products of slightly different sizes. The DNA ladder and the RT-PCR products were from the same gel. (B) (Upper) The number of known splice junctions detected nears saturation. (Lower) The number of unannotated splice junctions does not saturate at this read depth. (C) Splicing diversity is the highest in hESCs and decreases when cells commit to neural differentiation. The top 500 highly expressed genes shown here were clustered by splice junction diversity (k-means clustering, k = 3). The splice junction diversity value was defined as the number of unique splice junctions detected in the composite gene model given all of the mapped splice junction reads; thus the junction diversity values were normalized for the number of annotated splice junctions in the composite gene model and the number of mapped reads per million. Splice junction diversity is independent of transcript abundance for this set of genes.

Of particularly high interest is how splice isoform diversity changes as a function of cell differentiation, which has not been examined previously. We therefore quantified the number of unique splice junctions per composite gene model for each differentiation stage. To analyze the splice junction diversity, the 500 most highly transcribed genes were selected on the basis of the

sum of their transcription values in the four stages. These abundant transcripts were chosen because they provide large numbers of reads and allow for significant splicing differences to be identified. Our analysis revealed higher isoform diversity in hESCs compared with the neural stages (the median of the junction values for hESC, N1, N2, and N3 are 3.1, 2.2, 1.9, and 2.1, respectively). Interestingly, within the chosen set, this observation is independent of transcript abundance (Figure 6.4C). These data suggest that isoform diversity simplifies during differentiation, a process we have named *isoform specialization*.

6.2.5 Dynamic transcriptome changes during neural differentiation

We next examined the types of genes that exhibited differences in transcription using Gene Ontology analysis. We found that genes involved in nervous system development, neuron differentiation, brain development, regulation of gene expression, and pattern specification were significantly overrepresented among up-regulated genes in N2-B cells compared with hESCs (Figure 6.5A).

Genes were clustered according to the dynamic transcriptome changes between the four stages (ES→N1, N1→N2, and N2→N3) (Figure 6.5B). We found that a group of genes containing *SOX1*, *SOX2*, *PAX6*, *MAP2*, *DCX*, *ZIC1*, *NOTCH2*, *HES1*, and *OLIG2* had the highest transcript levels at the N2 stage and validated the relative transcript levels by quantitative PCR (qPCR) (Figure 6.5C). H7 hESCs showed similar gene expression patterns by qPCR. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed that the neuroactive ligand–receptor interaction pathway was enriched among genes that were up-regulated at the N1 and/or N2 stage but down-regulated at the N3 stage. A wide variety of receptor genes were transcribed in N1 and N2 cells, suggesting that these cells may be capable of differentiating into glutamatergic, GABAergic, dopaminergic, cholinergic, adrenergic, and serotoninergic neuronal subtypes. Although the overall proliferation rate between N2 and N3 cells in cultures with bFGF/EGF supplement is similar, the receptors were diminished at the N3 stage. This is consistent with the fact that the N3 stage cells expressed glial markers and had a propensity to differentiate into glial cells after bFGF/EGF withdrawal [118].

a

Genes enriched in N2-B compared to hESCs

GO Description	P-value
nervous system development	3.84E-11
cellular developmental process	7.90E-09
cell differentiation	3.96E-08
forebrain development	3.80E-06
central nervous system development	1.12E-04
pattern specification process	1.40E-03
axon guidance	3.22E-03
neuron recognition	9.01E-03
regulation of anatomical structure morphogenesis	1.61E-02
neuron differentiation	1.63E-02
regulation of gene expression	1.85E-02

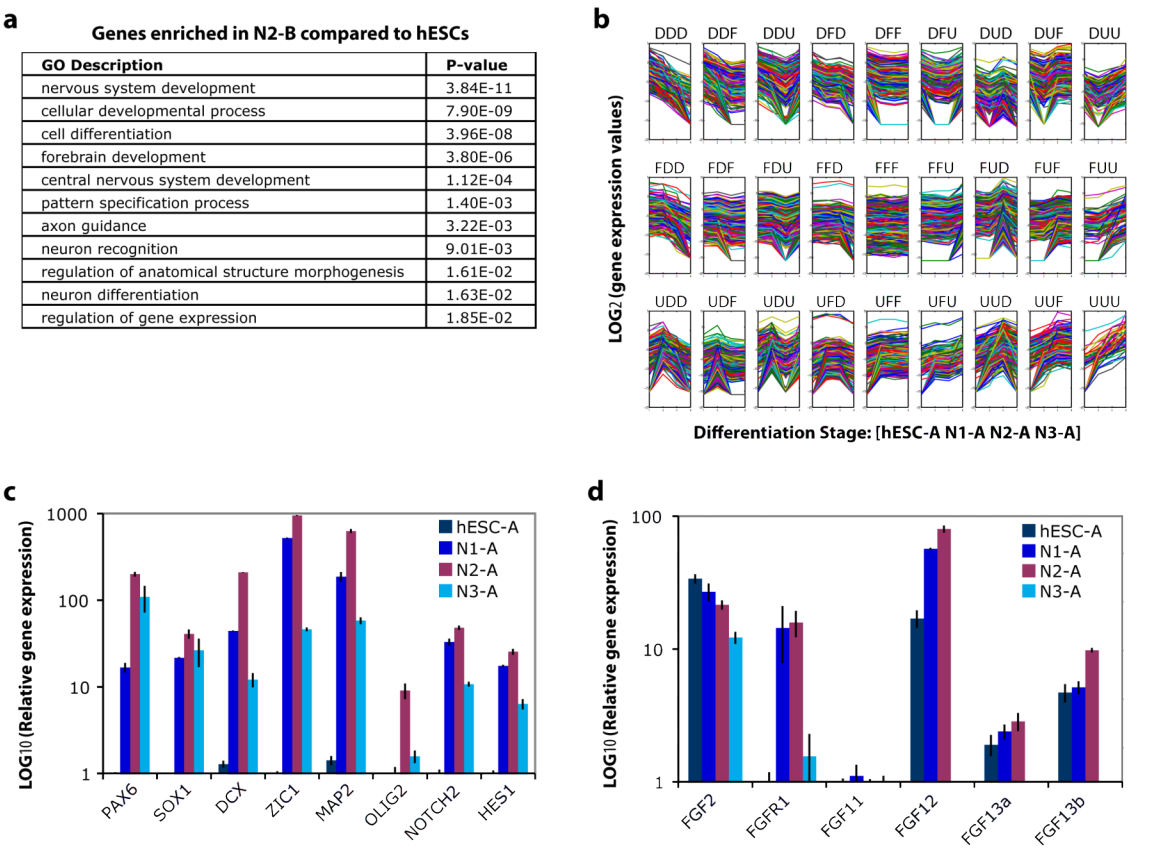


Figure 6.5 Dynamics of gene expression during neural differentiation. (A) The enriched Gene Ontology (GO) categories among up-regulated genes (>2-fold) in N2 cells compared with hESCs. (B) Quantification of dynamic transcriptome changes during neural differentiation process [x axis: differentiation stages hESC, N1, N2, and N3; y axis: log2(gene expression values by RNA-Seq)]. Twenty-seven patterns were identified from clustering the expression changes of the UCSC gene annotation set. U, up-regulation; D, down-regulation; F, flat (<2-fold change). (C) qPCR validation of the genes that show the highest expression at N2. y axis: log2(relative gene expression level for each stage normalized using HPRT). (D) qPCR validation of FGF family gene expression (Note: qPCRs were performed for two isoforms of FGF13.) Expression of FGF12 and FGF13a was not detected in N3 by qPCR.

A coordinated interplay among signaling pathways, such as Wnt and FGF is critical for neural specification [114]. Components of the Wnt pathway previously implicated in noncanonical Wnt signaling [*FZD5* (frizzled homolog 5, Drosophila), *WNT5A*, and *WNT5B*] were found to be down-regulated upon neural differentiation. This is consistent with Wnt pathway function in maintaining the undifferentiated state of stem cells. Interestingly, a group of FGF family genes exhibited increased expression at the N2 stage and then decreased at the N3 stage. This group included *FGF11*, *FGF13*, and *FGF14*, which do not bind to FGF receptor (Figure 6.5D); their roles during neural differentiation are not well understood.

6.3 Discussion

Our study revealed a high level of transcriptome complexity and dynamics during early neural differentiation of hESCs. Alternative splicing has been suggested as a driving force for the evolution of higher eukaryotic phenotypic complexity [132], and it has been shown recently that ≈94% of human genes undergo alternative splicing [20]. Previous studies have shown that massively parallel sequencing technology provides the ability to monitor spliced isoforms at the level of individual splice junctions [20, 58]. Our work incorporates the use of paired-end sequencing and demonstrates that this approach in conjunction with long reads can elucidate multiple exon connections and thereby reconstruct transcribed regions.

Importantly, our study demonstrated that greater splice junction diversity is present in hESCs relative to cells undergoing neural differentiation (Figure 6.4B and 4C). We suggest that this high transcript diversity contributes to the pluripotency of hESCs. Upon differentiation, more specialized transcripts are used, a process that we call isoform specialization. A previous study [133] reported elevated global transcription level in murine ESCs compared with NPCs; increased transcription in ESCs would support a larger number of isoforms.

RNA-Seq also enabled the detection of unknown RNAs (unannotated TARs) (Figure 6.3A). A larger fraction of these unannotated TARs were transcribed in a stage-specific fashion compared with the annotated mRNAs (Figure 6.3A). It is possible that at least some of the unannotated TARs may play an important role in specifying cell differentiation. The largest number of unannotated TARs was observed in hESCs, consistent with our conclusion that the transcriptome complexity is the greatest in pluripotent hESCs.

One important advantage of massively parallel sequencing is that it is more quantitative than other methods [1]. RNA-Seq can accurately measure gene transcription changes at a genome-wide scale, even for low-abundance transcription factors. Thus, these data are a valuable reference data set for researchers studying this process.

Our results provide important insights into the hESCs neural differentiation in vitro. Specifically, we have revealed aspects of neural specification, neural progenitor identity

maintenance, and the transition from neurogenesis to gliogenesis. Our data suggest the temporal order of transcription of the key transcription factors *SOX1* and *PAX6* during human neural specification. *SOX1* is a member of the *SOXB1* transcription factor family that plays important roles in neuroectodermal lineage commitment and maintenance [134, 135]. *PAX6* is a highly conserved transcription factor essential for central nervous system development [136]. The temporal order of their transcription and their roles in human neuroectodermal specification are not fully understood. In mice *SOX1* was found to be the earliest transcribed neural marker, preceding *PAX6*. *PAX6* is first transcribed in radial glial cells during the differentiation of mouse ESCs [137], and it has been reported to be involved in the progression of neuroectoderm toward radial glia [138]. However, in our experiments using hESCs, *PAX6* mRNAs appeared before *SOX1* mRNA, consistent with the immunostaining observations of Gerrard et al. [118]. Thus, *PAX6* may have an earlier role in neural lineage choice in human ESCs than in mouse ESCs.

The transcription of a wide variety of receptor genes at the N1 and N2 stages indicates that if the proper differentiation conditions are applied, these cells could potentially differentiate into glutamatergic, GABAergic, dopaminergic, cholinergic, adrenergic, and serotoninergic neuronal subtypes. Two possibilities can explain why these neuroactive ligand–receptors are not retained in N3 cultures. First, the receptors may be lost in N3 cells owing to cell death and/or less proliferation of proneuronal cells; the proneuronal cells would then be gradually replaced by the proglial cells. However, this cannot explain the complete absence of GFAP when neuronal differentiation is induced at an earlier stage. The second possibility is that a series of gene repression and activation events lead to the transition of the cells from a proneuronal nature to a proglial nature. Our finding that *FGF* family genes, including nonFGF-receptor-binding *FGF11*, *FGF13*, and *FGF14*, increase at the N2 stage and decrease at the N3 stage (Figure 6.5D) raises the possibility that modifying their levels may help to maintain hESC-derived neural cells at the neuronal stage.

Overall, our approach can serve as a template for the investigation of dynamic temporal or spatial transcriptome changes during various developmental processes. Future improvements

of sequencing technologies, including longer reads, higher throughput, and reduced cost will aid in the definition of transcriptomes and alternative splicing in specific temporal and spatial contexts.

6.4 Materials and methods

6.4.1 hESC culture and neural differentiation

6.4.1.1 Approach A

H1 hESCs were cultured in Matrigel-coated plates in mouse embryonic fibroblast conditioned medium supplemented with 8 ng/mL bFGF as previously described [139]. Cells were propagated at a 1:3 ratio by treatment with 200 U/mL collagenase IV and mechanical dissection. Neural differentiation was carried out as previously described [118]. Briefly, hESCs were split with EDTA at 1:5 ratios into culture dishes coated with poly-L-lysine/laminin (Sigma-Aldrich) and cultured in N2B27 medium supplemented with 100 ng/mL mouse recombinant Noggin (R&D Systems). At this stage, cells were defined as passage 1 (P1), and N1 cells were collected at Day 11 of the differentiation. Cells of P1 and P2 were split by collagenase into small clumps, similar to hESC culture, and continuously cultured in N2B27 medium with Noggin. From P3, cells were plated at the density of 5×10^4 cells/cm^2 after disassociation by TrypLE express (Invitrogen) into single-cell suspension, and cultured in N2B27 medium with the addition of 20 ng/mL bFGF and EGF. Cells can be maintained under this culture condition for a long time with a stable proliferative capacity. N2 cells were collected at P9 and N3 at P22. To induce postmitotic cell types, bFGF and EGF were withdrawn, and neural progenitors were continually cultured in N2B27 for 7 days. Cells from the different stages were analyzed for homogeneity by flow cytometry analysis. Cells from N2 and N3 were stained with anti-SOX1 (Abcam), SOX2 (Abcam), and MUSASHI (Chemicon) antibodies; 10–20,000 cells were analyzed.

6.4.1.2 Approach B

All experiments involving hESCs were approved by the Yale Embryonic Stem Cell Oversight Committee. hESC line H1 (WA01, WiCell) was maintained in undifferentiated state by

culturing on Matrigel-coated plates (BD) in feeder-free and serum free, component-defined conditions. Briefly, the cells were cultured in DMEM/F12 medium (Invitrogen) supplemented with 1% MEM-nonessential amino acids (Invitrogen), 1 mM L-glutamine, 1% penicillin-streptomycin, 50 ng/mL bFGF (FGF-2) (Millipore), 1× N2 supplements, and 1× B27 supplements (Invitrogen) [140], with daily media change. H1 cells were passaged every 4–6 days by dissociation with 1 mg/mL collagenase IV (Invitrogen). The hESCs used were between passages 30 and 70 with normal karyotype and expressed conventional hESC markers. hESCs were differentiated by neural sphere formation with some modifications of previously published protocols [120]. Cells were fixed and analyzed by standard fluorescent immunocytochemical techniques.

6.4.1.3 Flow Cytometry

Cells were detached by trypsin, fixed with 4% paraformaldehyde for 15 minutes, permeated with 100% ethanol for 2 minutes and incubated with 10% goat serum (Sigma) for 15 minutes. Cells were then stained with primary antibodies (SOX1 and SOX2 from Abcam, Musashi1 from Chemicon, all 1:100) for 30 minutes on ice followed by secondary antibody (Goat-anti rabbit conjugated fluorescein, Santa Cruz) for 30 minutes. Ten to twenty thousand cells were acquired for each sample using a FACScan (BD Biosciences) and analyzed with CELLQUEST software (BD Biosciences).

6.4.1.4 Immunofluorecent Staining

The cells were either fixed with 4% formaldelhyde or 3% paraformaldehyde for 10 min, followed by standard fluorescent immunocytochemical techniques using the following primary antibodies: monoclonal OCT4 (1:20) and polyclonal SOX2 (1:50) from Santa Cruz, monoclonal NESTIN (1:200), polyclonal MUSASHI (1:100), monoclonal PAX6 (1:50) from Dev.Studies Hybridoma Bank, Iowa, monoclonal TUJ1 (1:1000) and polyclonal GFAP (1: 250) from Chemicon, and polyclonal TUJ1 (1:1000) from Covance. The images were acquired with a Nikon Eclipse E800 fluorescent microscope or LEICA confocal fluorescent microscope.

6.4.2 RNA sequencing

6.4.2.1 Construction of Solexa sequencing library

mRNA samples were extracted and double-polyA purified from cell cultures using Oligotex Direct mRNA Kits followed by Oligotex mRNA Kits, according to the manufacturer's instructions (Qiagen). mRNA (500 ng) was used in each sequencing library. mRNA was fragmented using 10× Fragmentation Buffer (Ambion), and double-stranded cDNA was synthesized using SuperScript II (Invitrogen) RT and random primers. DNA sequencing followed the instructions of the mRNA-Sequencing Sample Prep Kit (Illumina) as previously described [18].

6.4.2.2 454 sequencing library preparation

mRNA was prepared as described above. mRNA samples (200–500 ng) were heat fragmented. Single-stranded cDNA library was synthesized, and adapters were ligated and sequenced using the emPCR II Kit (Amplicon A) and on the 454 Genome Sequencer FLX instrument according to the manufacturer's instructions. GS FLX Titanium cDNA libraries were prepared and sequenced at the 454 Life Sciences Sequencing Centre.

6.4.3 RT-PCR

6.4.3.1 RT-PCR validation experiments

1µg each of polyA RNAs from cell of ES, N1, N2 and N3 stages was separately set up in 200µl Reverse Transcription (RT) reactions (5ng/µl). RT reactions were performed using SuperScript™ III First-Strand Synthesis SuperMix for qRT-PCR (Invitrogen, CA, USA) that contains both oligo(dT)20 and random hexamers. In parallel, reactions without reverse transcriptase (RTase minus) were also performed as the negative control for genomic contamination.

PCR primers were designed using Primer3 or BatchPrimer3. RT was followed by PCR amplification using Advantage™ 2 PCR Enzyme System (Clontech, CA, USA). 1µl RT reaction and 1µl RTase minus negative control from the above were used in 25µl PCR reactions. The PCR program for unannotated singleton TARs was 95°C for 30 seconds, followed by 35 cycles of

95°C for 15 seconds, 68°C 30 seconds, and concluded by an extension cycle of 72°C for 1 minute. The PCR program for multi-exonic unannotated transcripts connected via a group of paired-end reads was: 95°C for 1 minute, followed by either 28 or 35 cycles of 95°C for 15 seconds, 68°C 3 minutes, and concluded by an extension cycle of 72°C for 5 minutes. The PCR products were visualized on a 1% agarose gel.

6.4.3.2 Real-time quantitative RT-PCR

Total RNA was purified with TRI reagent (Sigma) and trace contaminated DNA was removed by DNase treatment (Invitrogen). First-strand cDNA was synthesized from 2 µg total RNA in a 20-µl volume using oligo-dT15 primer and SuperScriptII (Invitrogen). The PCR reaction consisted of 2µl of 1:10-diluted cDNA, 15 µl of SYBR green-Taq mixed solution (Sigma) and 9 pmol each of 5' and 3' primers (see table below) in a total volume of 30 µl and was performed in a Opticon thermal cycler (Biorad) for 40 cycle with denaturation at 95°C for 15 second, annealing at 60°C for 30 second and extension at 72°C for 30 second. RNA without reverse transcriptase treatment was used as negative control.

6.4.4 Bioinformatics analysis

6.4.4.1 Mapping sequence reads to the human genome

The 454 250-450bp long reads were mapped to human genome (hg18) using BLAT [95] with default parameters. Reads were removed in a subsequent post-processing step if less than eighty percent of the read mapped to the genome. A three-step approach was adopted to map the short single-end reads to the genome. First, reads were aligned to the human genome (hg18) with Bowtie [94] allowing up to two mismatches. Only reads that mapped to a unique location in the genome were retained. In a second step, the remaining reads were aligned to a splice junction library consisting of all possible unique pair-wise splice junctions within each transcript of the AceView [141] annotation set. This alignment step was also performed with Bowtie allowing only unique alignments with up to two mismatches. Lastly, the reads that did not align in the previous two steps were aligned to the genome using less stringent parameters. In this step

reads were allowed to map up to five genomic locations. One of these locations was selected according to the read density of the uniquely mapped reads contained within non-overlapping bins (50bp) across the genome. The short paired-end reads were mapped to the genome using ELAND, which is a component of the Illumina software pipeline, operating in the *eland_pair* mode. Each end was aligned separately and then the best-matched pair was selected and reported.

6.4.4.2 Annotation sets and composite gene models

Initially, the AceView annotation set [141], consisting of 258,618 transcripts, was used to create a splice junction library in order to map the short single-end reads. However, for many subsequent analyses the UCSC Genes annotation set [79], comprised of 66,803 transcripts, was utilized because it contains information about the various splice isoforms. The various transcript isoforms of a particular gene were merged into a composite gene model by taking the union of all the exons from the various transcript isoforms.

6.4.4.2 Number of genes detected as a function of read coverage

To assess the number of genes detected as a function of read depth the mapped reads were intersected with the composite gene models of the UCSC Genes annotation set [79]. Reads were sampled randomly at various intervals of five millions and the fraction of genes detected was calculated. The fraction of genes detected was determined at two-fold and five-fold coverage. The coverage is defined as the number of nucleotides obtained form all the reads that overlap with a composite gene model divided by the length of the composite gene model.

6.4.4.3 Quantification of gene transcription

The level of gene transcription (RPKM; [18]) was quantified by intersecting the mapped reads with the composite gene models of the UCSC Genes annotation set [79]. The transcription values were determined by summing the nucleotide overlaps from all the reads that intersect with a composite gene model divided by the length of the composite gene model and the number of mapped reads in millions.

6.4.4.4 Differential gene expression

Gene expression values from various differentiation stages were compared to assess differential gene expression. Genes with at least a two-fold change in their expression values were referred to as differentially expressed. In order to capture the global changes in gene expression across the four differentiation stages each gene was assigned to one of 27 gene expression patterns. Between any two differentiation stages the change in gene expression was assigned to one of three categories: up, down, or flat. The 'up' or 'down' categories represent at least a two-fold change in gene expression between the two differentiation stages while the 'flat' category indicates an unchanged gene expression or a change less than two fold. After assigning each gene to one of the 27 gene expression patterns, the logarithm of the expression values across the four stages were plotted for each pattern type.

6.4.4.5 Splice junction coverage

The splice junction coverage was determined by counting the number of known and unannotated splice junctions obtained from random samples of short single-end reads spanning two exons. Since the splice junction library consisted of all possible unique pair-wise splice junctions within a transcript each splice junction can be categorized as either known or unannotated. Known splice junctions are defined as the junctions that are consistent with annotated transcripts while unannotated splice junctions refer to skipped exons.

6.4.4.6 TAR analyses and connecting TARs using paired-end reads

In order to discover unannotated transcriptionally active regions (TARs) the signal track of the mapped reads was segmented using the common maxGap/minRun algorithm (maxGap = 10, minRun = 50, threshold = 2) [35, 55]. The set of unannotated TARs identified for each differentiation stage were intersected with the set of TARs reported by Bertone et al. [2]. If two TARs from the two different sets overlapped by at least one nucleotide they were counted as overlapping.

6.4.4.7 Splice junction diversity

A subset of genes was selected to analyze the splice junction diversity across the four differentiation stages. Because the sequence coverage for low abundant transcript is lower and it would be difficult to ascertain splicing diversity, the 500 most highly transcribed genes were selected based on the sum of their transcription values in the four stages. For this subset of genes the junction diversity per composite gene model was calculated. The junction diversity is defined as the number of unique splice junctions detected in the composite gene model given all the mapped splice junction reads. In order to facilitate a comparison between the various developmental stages the junction diversity values were normalized for the number of mapped reads per million. In addition, the splice junction diversity values were normalized for the number of annotated splice junctions in the composite gene model. In the next step normalized junction diversity values were clustered using *k*-means clustering ($k = 3$). Lastly, the normalized junction diversity values and the associated gene transcription values were plotted.

6.4.4.8 Enriched Gene Ontology (GO) categories and pathway analysis

An internal software tool was employed to determine statistically significant over-represented GO categories within lists of genes. The hypergeometric distribution was utilized to calculate p-values. These p-values were then corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. Enriched KEGG pathways of differentially expressed genes were identified using Database for Annotation, Visualization and Integrated Discovery (DAVID) [142].

Chapter 7

Conclusion

In this thesis, we presented four different computational methodologies for transcript analysis in the age of next-generation DNA sequencing. First, we described VAT, which is used for functionally annotating variants and analyzing their effects on transcript structure. In this approach, we not only offer efficient software modules to annotate variants, including a novel way for visualizing the results, but also place the software in the same space as the data by providing VAT as a virtual machine that may be run in a cloud-computing environment.

We then described RSEQtools, which introduces a novel data format (MRF) for representing read alignments in a compact manner, thereby enabling the dissemination of large volumes of data. In addition, we provide a mechanism for protecting personal genotypic information, as well as a set of tools that can be assembled to build customizable RNA-Seq workflows for carrying out a number of downstream analyses.

We then described FusionSeq, a downstream analysis pipeline based on RSEQtools. This framework is used for finding instances of gene fusions by analyzing paired-end RNA-Seq data, and comprises three main modules. It exploits the connectivity information provided by paired-end RNA-Seq reads to identify potential fusion candidates. Many of the initially identified gene fusion candidates are then removed using an elaborate filtration cascade. Furthermore, a scoring scheme facilitates the prioritization of these candidates for experimental validation. A third module then determines the exact junction sequence surrounding the breakpoint between two genes. By employing this approach, we were able to identify several high quality fusion candidates in prostate cancer tissue samples.

We next explored new ways for assessing the transcriptional activity of highly similar genomic sequences, which has been non-trivial, especially for those genomic elements that

share high degrees of sequence similarity, such as pseudogenes and paralogs. In order to address these issues, we have designed DupSeq, a computational framework that employs statistical methods to compare the transcription signal patterns (as obtained from mapped RNA-Seq reads) across multiple samples. By comparing the signal of a given sequence across multiple tissues, truly transcribed regions will be characterized by distinctly different expression patterns relative to those observed in regions with high sequence similarity (i.e., the expression patterns are independent), whereas concordant patterns are suggestive of mapping artifacts.

In the last part of this thesis, we applied these computational methods to investigate the fundamental mechanisms governing neural differentiation by analyzing transcriptome dynamics data. This analysis not only revealed many previously unannotated transcripts and differentially expressed transcript isoforms, but also uncovered a reduction in splicing isoform diversity as human embryonic stem cells differentiate into neural cells.

We have identified a number of directions for future research. Specifically, in the context of DupSeq, we intend to implement enhanced statistical methodologies to better discriminate between true transcription and mapping artifacts. In part, we envision using principal component analysis on the matrix, which represents the expression signals across multiple samples, to obtain a set of uncorrelated variables (i.e., the principal components). Thus, if more than one principal component is needed to represent the original matrix, this would indicate that the region of interest is most likely independently transcribed. Conversely, if the original matrix can be collapsed into one principal component, then the region of interest would most likely be a mapping artifact. In addition to improved statistical metrics, we are planning to apply DupSeq not only to pseudogenes, but also to other genomic elements that share high levels of sequence similarity, such as paralogs and unannotated transcribed regions. Furthermore, we aim to set up a web service that would enable researchers to upload a set of specified genomic regions in order to obtain a readout regarding their transcriptional activity.

Bibliography

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.

2. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M: **Global Identification of Human Transcribed Sequences with Genome Tiling Arrays**. *Science* 2004, **306**:2242-2246.

3. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome**. *Proc. Natl. Acad. Sci. U.S.A.* 2006, **103**:5320-5325.

4. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MMH, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, Banh J, Banno F, Bowser L, Brooks S, Carninci P, Chao Q, Choy N, Enju A, Goldsmith AD, Gurjal M, Hansen NF, Hayashizaki Y, Johnson-Hopson C, Hsuan VW, Iida K, Karnes M, Khan S, Koesema E, Ishida J, Jiang PX, Jones T, Kawai J, Kamiya A, Meyers C, Nakajima M, Narusaka M, Seki M, Sakurai T, Satou M, Tamse R, Vaysberg M, Wallender EK, Wong C, Yamamura Y, Yuan S, Shinozaki K, Davis RW, Theologis A, Ecker JR: **Empirical analysis of transcriptional activity in the Arabidopsis genome**. *Science* 2003, **302**:842-846.

5. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**:1149-1154.

6. Royce TE, Rozowsky JS, Gerstein MB: **Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification**. *Nucleic Acids Res* 2007, **35**:e99.

7. Okoniewski MJ, Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations**. *BMC Bioinformatics* 2006, **7**:276.

8. Boguski MS, Tolstoshev CM, Bassett DE Jr: **Gene discovery in dbEST**. *Science* 1994, **265**:1993-1994.

9. Gerhard DS, Wagner L, Feingold EA, et al.: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)**. *Genome Res.* 2004, **14**:2121-2127.

10. Harbers M, Carninci P: **Tag-based approaches for transcriptome research and genome annotation**. *Nat. Methods* 2005, **2**:495-502.

11. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**:484-487.

12. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham

T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays**. *Nat. Biotechnol.* 2000, **18**:630-634.

13. Reinartz J, Bruyns E, Lin J-Z, Burcham T, Brenner S, Bowen B, Kramer M, Woychik R: **Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms**. *Brief Funct Genomic Proteomic* 2002, **1**:95-104.

14. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **advance online publication**.

15. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061-1073.

16. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing**. *Science* 2008, **320**:1344-1349.

17. Cloonan N, Grimmond S: **Transcriptome content and dynamics at single-nucleotide resolution**. *Genome Biology* 2008, **9**:234.

18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Meth* 2008, **5**:621-628.

19. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution**. *Nature* 2008, **453**:1239-1243.

20. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008.

21. Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier L, Sasidharan R, Reinke V, Waterston R, Gerstein M: **Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays**. *BMC Genomics* 2010, **11**:383.

22. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res.* 2008, **18**:1509-1517.

23. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci USA* 2009, **106**:12353 - 12358.

24. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB: **FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data**. *Genome Biol* 2010, **11**:R104.

25. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries**. *Bioinformatics* 2011, **27**:281 -283.

26. Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M: **Gene inactivation and its implications for annotation in the era of personal genomics**. *Genes & Development* 2011, **25**:1 -10.

27. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**:e164.

28. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**:3812-3814.

29. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Research* 2002, **30**:3894 -3900.

30. Li H, Wang J, Mor G, Sklar J: **A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells**. *Science* 2008, **321**:1357-1361.

31. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, Kuefer R, Tewari AK, Setlur SR, Demichelis F, Rubin MA: **SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer.** *Cancer Res* 2009, **69**:2734 - 2738.

32. Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M: **Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics**. *Ann. Intern. Med.* 2003, **138**:819-830.

33. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.** *Science* 2005, **310**:644 - 648.

34. Chen YA, Chou C-C, Lu X, Slate EH, Peck K, Xu W, Voit EO, Almeida JS: **A multivariate prediction model for microarray cross-hybridization**. *BMC Bioinformatics* 2006, **7**:101.

35. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping**. *Trends in Genetics* 2005, **21**:466-475.

36. Gerstein MB, Lu ZJ, Van Nostrand EL, et al.: **Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project**. *Science* 2010, **330**:1775 -1787.

37. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, Hutchison S, Raha D, Egholm M, Lin H, Weissman S, Cui W, Gerstein M, Snyder M: **Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing**. *Proceedings of the National Academy of Sciences* 2010, **107**:5254-5259.

38. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C: **A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes**. *Science* 2012, **335**:823-828.

39. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker R, Lunter G, Marth G, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: **The Variant Call Format and VCFtools**. *Bioinformatics* 2011.

40. Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert J, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis S, Guigo R: **GENCODE: producing a reference annotation for ENCODE**. *Genome Biology* 2006, **7**:S4.

41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**:996-1006.

42. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Vogel J, Searle SMJ: **Ensembl 2011**. *Nucleic Acids Research* 2010, **39**:D800-D806.

43. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: integrating information about genes, proteins and diseases**. *Trends Genet* 1997, **13**:163.

44. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research* 2001, **29**:308 -311.

45. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of C. elegans**. *Genome Res* 2009, **19**:657-666.

46. Trapnell C, Salzberg SL: **How to map billions of short reads onto genomes**. *Nat Biotech* 2009, **27**:455-457.

47. Greenbaum D[1], Du J[2], Gerstein M[2]: **Genomic Anonymity: Have We Already Lost It?** *American Journal of Bioethics* 2008, **8**:71-74.

48. Lowrance WW, Collins FS: **ETHICS: Identifiability in Genomic Research**. *Science* 2007, **317**:600-602.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.

50. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**:207 -210.

51. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A: **ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments**. *Nucleic Acids Res* 2011, **39**:D1002-D1004.

52. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A: **Ab initio reconstruction of cell type-**

specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotech* 2010, **28**:503-510.

53. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotech* 2010, **28**:511-515.

54. Du J, Leng J, Habegger L, Sboner A, McDermott D, Gerstein M: **IQSeq: Integrated Isoform Quantification Analysis Based on Next-Generation Sequencing**. *PLoS ONE* 2012, **7**:e29175.

55. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22**. *Genome Res* 2004, **14**:331-342.

56. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M-L: **A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome**. *Science* 2008, **321**:956-960.

57. Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M, Gendron D, Brosseau J, Thibault P, Lucier J, Tremblay K, Prinos P, Wellinger RJ, Chabot B, Rancourt C, Elela SA: **Identification of alternative splicing markers for breast cancer.** *Cancer Res* 2008, **68**:9525 - 9531.

58. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **advanced online publication**.

59. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** *PLoS Comput Biol* 2009, **5**:e1000598.

60. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**:799-816.

61. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH: **Unlocking the secrets of the genome.** *Nature* 2009, **459**:927 - 930.

62. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**:97-101.

63. Pflueger D, Rickman DS, Sboner A, Perner S, LaFargue CJ, Svensson MA, Moss BJ, Kitabayashi N, Pan Y, de la Taille A, Kuefer R, Tewari AK, Demichelis F, Chee MS, Gerstein MB, Rubin MA: **N-myc downstream regulated gene 1 (NDRG1) is fused to ERG in prostate cancer.** *Neoplasia* 2009, **11**:804 - 811.

64. Gingeras TR: **Implications of chimaeric non-co-linear transcripts.** *Nature* 2009, **461**:206 - 211.

65. Zhao Q, Caballero OL, Levy S, Stevenson BJ, Iseli C, de Souza SJ, Galante PA, Busam D, Leversha MA, Chadalavada K, Rogers Y-H, Venter JC, Simpson AJG, Strausberg RL: **Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line**. *Proceedings of the National Academy of Sciences* 2009, **106**:1886-1891.

66. Mitelman F, Johansson B, Mertens F: **Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer.** *Nat Genet* 2004, **36**:331 - 334.

67. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation.** *Nat Rev Cancer* 2007, **7**:233 - 245.

68. Mitelman F: **Recurrent chromosome aberrations in cancer.** *Mutat Res* 2000, **462**:247 - 253.

69. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H: **Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.** *Nature* 2007, **448**:561 - 566.

70. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM: **Recurrent gene fusions in prostate cancer.** *Nat Rev Cancer* 2008, **8**:497 - 511.

71. Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G, Croft L, Taft R, Rizzi E, Askarian-Amiri M, Bonnal R, Callari M, Mignone F, Pesole G, Bertalot G, Bernardi L, Albertini A, Lee C, Mattick J, Zucchi I, De Bellis G: **A transcriptional sketch of a primary human breast cancer by 454 deep sequencing.** *BMC Genomics* 2009, **10**:163.

72. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318**:420 - 426.

73. The Human Genome Structural Variation Working Group: **Completing the map of human genetic variation.** *Nature* 2007, **447**:161 - 165.

74. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53 - 59.

75. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: **Deciphering the splicing code**. *Nature* 2010, **465**:53-59.

76. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26**:873 - 881.

77. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, Raymond CK: **Digital transcriptome profiling using selective hexamer priming for cDNA synthesis**. *Nat Meth* 2009, **6**:647-649.

78. Ameur A, Wetterbom A, Feuk L, Gyllensten U: **Global and unbiased detection of splice junctions from RNA-seq data.** *Genome Biol* 2010, **11**:R34.

79. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22**:1036 - 1046.

80. Li H, Coghlan A, Ruan J, Coin LJ, Heriche J, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R: **TreeFam: a curated database of phylogenetic trees of animal gene families.** *Nucleic Acids Res* 2006, **34**:D572 - 580.

81. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Heriche J, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R: **TreeFam: 2008 update.** *Nucleic Acids Res* 2008, **36**:D735 - 740.

82. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5**:1005 - 1010.

83. Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo W, Magrane G, De Jong P, Gray JW, Collins C: **End-sequence profiling: sequence-based analysis of aberrant genomes.** *Proc Natl Acad Sci USA* 2003, **100**:7696 - 7701.

84. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37**:727 - 732.

85. Korbel J, Abyzov A, Mu X, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein M: **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.** *Genome Biol* 2009, **10**:R23.

86. Taylor TD, Noguchi H, Totoki Y, Toyoda A, Kuroki Y, Dewar K, Lloyd C, Itoh T, Takeda T, Kim D, She X, Barlow KF, Bloom T, Bruford E, Chang JL, Cuomo CA, Eichler E, FitzGerald MG, Jaffe DB, LaButti K, Nicol R, Park H, Seaman C, Sougnez C, Yang X, Zimmer AR, Zody MC, Birren BW, Nusbaum C, Fujiyama A: **Human chromosome 11 DNA sequence and analysis including novel gene identification.** *Nature* 2006, **440**:497 - 500.

87. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin P-C, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA: **Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing**. *Genome Research* 2011, **21**:56 -67.

88. Jhavar S, Reid A, Clark J, Kote-Jarai Z, Christmas T, Thompson A, Woodhouse C, Ogden C, Fisher C, Corbishley C, De-Bono J, Eeles R, Brewer D, Cooper C: **Detection of TMPRSS2-ERG translocations in human prostate cancer by expression profiling using GeneChip human exon 1.0 ST arrays.** *J Mol Diagn* 2008, **10**:50 - 57.

89. Wang J, Cai Y, Ren C, Ittmann M: **Expression of variant TMPRSS2/ERG fusion messenger RNAs is associated with aggressive prostate cancer.** *Cancer Res* 2006, **66**:8347 - 8351.

90. Clark J, Merson S, Jhavar S, Flohr P, Edwards S, Foster CS, Eeles R, Martin FL, Phillips DH, Crundwell M, Christmas T, Thompson A, Fisher C, Kovacs G, Cooper CS: **Diversity of TMPRSS2-ERG fusion transcripts in the human prostate.** *Oncogene* 2006, **26**:2667 - 2673.

91. Tomlins SA, Bjartell A, Chinnaiyan AM, Jenster G, Nam RK, Rubin MA, Schalken JA: **ETS gene fusions in prostate cancer: from discovery to daily clinical practice.** *Eur Urol* 2009, **56**:275 - 286.

92. Wang J, Cai Y, Yu W, Ren C, Spencer DM, Ittmann M: **Pleiotropic biological activities of alternatively spliced TMPRSS2/ERG fusion gene transcripts.** *Cancer Res* 2008, **68**:8516 - 8524.

93. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19**:1639 - 1645.

94. Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**:R25.

95. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656-664.

96. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16**:418 - 420.

97. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

98. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science* 2002, **297**:1003-1007.

99. Khurana E, Lam HYK, Cheng C, Carriero N, Cayting P, Gerstein MB: **Segmental duplications in the human genome reveal details of pseudogene formation**. *Nucleic Acids Res* 2010, **38**:6997-7007.

100. Metzker ML: **Sequencing technologies [mdash] the next generation**. *Nat Rev Genet* 2010, **11**:31-46.

101. Balakirev ES, Ayala FJ: **Pseudogenes: are they "junk" or functional DNA?** *Annu. Rev. Genet.* 2003, **37**:123-151.

102. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M: **Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes**. *Nucleic Acids Res.* 2002, **30**:2515-2523.

103. Harrison PM, Echols N, Gerstein MB: **Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome**. *Nucleic Acids Res.* 2001, **29**:818-830.

104. Mighell AJ, Smith NR, Robinson PA, Markham AF: **Vertebrate pseudogenes**. *FEBS Lett.* 2000, **468**:109-114.

105. Piehler AP, Hellum M, Wenzel JJ, Kaminski E, Haug KBF, Kierulf P, Kaminski WE: **The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference**. *BMC Genomics* 2008, **9**:165.

106. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP: **A coding-independent function of gene and pseudogene mRNAs regulates tumour biology**. *Nature* 2010, **465**:1033-1038.

107. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, Hannon GJ: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes**. *Nature* 2008, **453**:534-538.

108. Sasidharan R, Gerstein M: **Genomics: Protein fossils live on as RNA**. *Nature* 2008, **453**:729-731.

109. Harrison PM, Gerstein M: **Studying genomes through the aeons: protein families, pseudogenes and proteome evolution**. *J. Mol. Biol.* 2002, **318**:1155-1174.

110. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M: **Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates**. *Genome Biol.* 2010, **11**:R26.

111. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res* 2008, **18**:1851-1858.

112. Nelson SB, Sugino K, Hempel CM: **The problem of neuronal cell types: a physiological genomics approach**. *Trends Neurosci.* 2006, **29**:339-345.

113. Watakabe A, Komatsu Y, Nawa H, Yamamori T: **Gene expression profiling of primate neocortex: molecular neuroanatomy of cortical areas**. *Genes Brain Behav.* 2006, **5 Suppl 1**:38-43.

114. Suter DM, Krause K-H: **Neural commitment of embryonic stem cells: molecules, pathways and potential for cell therapy**. *J. Pathol.* 2008, **215**:355-368.

115. Murry CE, Keller G: **Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development**. *Cell* 2008, **132**:661-680.

116. Ying Q-L, Stavridis M, Griffiths D, Li M, Smith A: **Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture**. *Nat. Biotechnol.* 2003, **21**:183-186.

117. Reubinoff BE, Itsykson P, Turetsky T, Pera MF, Reinhartz E, Itzik A, Ben-Hur T: **Neural progenitors from human embryonic stem cells**. *Nat. Biotechnol.* 2001, **19**:1134-1140.

118. Gerrard L, Rodgers L, Cui W: **Differentiation of human embryonic stem cells to neural lineages in adherent culture by blocking bone morphogenetic protein signaling**. *Stem Cells* 2005, **23**:1234-1241.

119. Schwartz PH, Brick DJ, Stover AE, Loring JF, Müller F-J: **Differentiation of neural lineage cells from human pluripotent stem cells**. *Methods* 2008, **45**:142-158.

120. Cohen MA, Itsykson P, Reubinoff BE: **Neural differentiation of human ES cells**. *Curr Protoc Cell Biol* 2007, **Chapter 23**:Unit 23.7.

121. Carninci P, Kasukawa T, Katayama S, et al.: **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**:1559-1563.

122. The MGC Project Team, Temple G, Gerhard DS, et al.: **The completion of the Mammalian Gene Collection (MGC)**. *Genome Research* 2009, **19**:2324 -2333.

123. Wu JQ, Garcia AM, Hulyk S, Sneed A, Kowis C, Yuan Y, Steffen D, McPherson JD, Gunaratne PH, Gibbs RA: **Large-scale RT-PCR recovery of full-length cDNA clones**. *BioTechniques* 2004, **36**:690-696, 698-700.

124. Wu J, Du J, Rozowsky J, Zhang Z, Urban A, Euskirchen G, Weissman S, Gerstein M, Snyder M: **Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome**. *Genome Biology* 2008, **9**:R3.

125. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**:872-876.

126. Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman M-L, Taillon BE, Du L, Bouffard P, Kingsmore SF, Miller NA, Farmer AD, Jensen RV, Gullans SR, Bueno R: **Transcriptome sequencing of malignant pleural mesothelioma tumors**. *Proc Natl Acad Sci U S A* 2008, **105**:3521-6.

127. Emrich SJ, Barbazuk WB, Li L, Schnable PS: **Gene discovery and annotation using LCM-454 transcriptome sequencing**. *Genome Res.* 2007, **17**:69-73.

128. Elkabetz Y, Studer L: **Human ESC-derived neural rosettes and neural stem cell progression**. *Cold Spring Harb. Symp. Quant. Biol.* 2008, **73**:377-387.

129. Temple S: **The development of neural stem cells**. *Nature* 2001, **414**:112-117.

130. Rozowsky J, Wu J, Lian Z, Nagalakshmi U, Korbel J o., Kapranov P, Zheng D, Dyke S, Newburger P, Miller P, Gingeras T r., Weissman S, Gerstein M, Snyder M: **Novel Transcribed Regions in the Human Genome**. *Cold Spring Harbor Symposia on Quantitative Biology* 2006, **71**:111 -116.

131. Yeo GW, Xu X, Liang TY, Muotri AR, Carson CT, Coufal NG, Gage FH: **Alternative splicing events identified in human embryonic stem cells and neural progenitors**. *PLoS Comput. Biol.* 2007, **3**:1951-1967.

132. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays**. *Science* 2003, **302**:2141-2144.

133. Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RDG, Buetow KH, Gingeras TR, Misteli T, Meshorer E: **Global transcription in pluripotent embryonic stem cells**. *Cell Stem Cell* 2008, **2**:437-447.

134. Kan L, Israsena N, Zhang Z, Hu M, Zhao L-R, Jalali A, Sahni V, Kessler JA: **Sox1 acts through multiple independent pathways to promote neurogenesis**. *Dev. Biol.* 2004, **269**:580-594.

135. Zhao S, Nichols J, Smith AG, Li M: **SoxB transcription factors specify neuroectodermal lineage choice in ES cells**. *Mol. Cell. Neurosci.* 2004, **27**:332-342.

136. Osumi N, Shinohara H, Numayama-Tsuruta K, Maekawa M: **Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator**. *Stem Cells* 2008, **26**:1663-1672.

137. Mori T, Buffo A, Götz M: **The novel roles of glial cells revisited: the contribution of radial glia and astrocytes to neurogenesis**. *Curr. Top. Dev. Biol.* 2005, **69**:67-99.

138. Suter DM, Tirefort D, Julien S, Krause K-H: **A Sox1 to Pax6 switch drives neuroectoderm to radial glia progression during differentiation of mouse embryonic stem cells**. *Stem Cells* 2009, **27**:49-58.

139. Xu C, Inokuma MS, Denham J, Golds K, Kundu P, Gold JD, Carpenter MK: **Feeder-free growth of undifferentiated human embryonic stem cells**. *Nat. Biotechnol.* 2001, **19**:971-974.

140. Yao S, Chen S, Clark J, Hao E, Beattie GM, Hayek A, Ding S: **Long-term self-renewal and directed differentiation of human embryonic stem cells in chemically defined conditions**. *Proc. Natl. Acad. Sci. U.S.A.* 2006, **103**:6907-6912.

141. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation**. *Genome Biol.* 2006, **7 Suppl 1**:S12.1-14.

142. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists**. *Nucleic Acids Res.* 2007, **35**:W169-175.