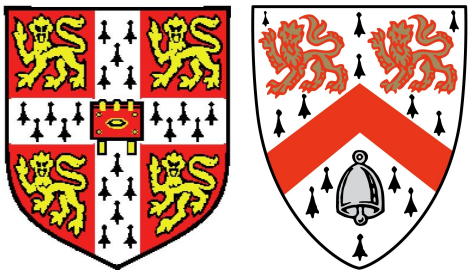
Transcriptome sequencing analysis with application to embryonic stem cell self-renewal



Tamara Steijger

European Bioinformatics Institute

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

4th October 2013

To family and friends.

Declaration

Transcriptome sequencing analysis with application to embryonic stem cell self-renewal

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This thesis does not exceed the specified length limit of 60.000 words as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using LaTeX according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Acknowledgements

First, I would like to thank my supervisor Paul Bertone for giving me the opportunity to conduct my PhD project in his research group and in the great environment of the European Bioinformatics Institute. Thank you for your support and for bringing me in contact with the people at the Cambridge Stem Cell Institute.

Furthermore, I would like to thank all the members of the Bertone group. Thanks to Remco, Myrto, Maria and Ewan for keeping me sane, especially during the last year of my PhD. But my gratitude also goes to previous members of the lab: Pär Engström for all his patience and his ability to answer all of my questions, Mali Salmon-Divon who helped me during my first steps of the RGASP project and Heidi Dvinge for moral support even from afar.

I also would like to thank my collaborators, especially Graziano Martello and Tuzer Kalkan, from the Stem Cell Institute for introducing me to the fascinating world of stem cell biology.

This thesis would not have been possible, without all the patient proof-readers. Here, I would like to thank in particular Remco Loos for all his patience and helpful discussions, but also Heidi Dvinge, Myrto Kostadima and Ewan Johnstone for proof-reading parts of my thesis.

My time in Cambridge would not have been the same without the Wolfson College Boat Club. Rowing proved to be the perfect to free my mind and release the daily stress of PhD life. I would in particular like to thank Christine Seeliger, Filipa Azevedo, Rasha Rezk, Evelyn Tichy, Carmen Jack, Zhen Sun, and our coach Chris Parkhouse for their constant support throughout the years. Thanks, you guys are amazing!

A personal thanks goes to Thomas Zichner for convincing me to apply to the EMBL PhD program. He made me reach out for the starts and without him, I would never have ended up doing a PhD at Cambridge University.

I am especially grateful to Pascal Maas who has been there for me in the last 1.5 years. Thank you for being so understanding when I had a bad day and for your incredible ability to cheer me up within seconds.

I am enormously grateful to my family, my parents Wilfridus and Anita Steijger and my lovely sister Wanda Steijger, for their support over all these years. Thank you for believing in me and for letting me go my way. Without your continuous support I would not be who I am today.

Abstract

Cell identity is defined by the set of active genes in each cell. Studying changes in gene expression between conditions or developmental stages provides insights into gene function. For many years microarrays have been the technology of choice for the study of gene expression. However, they come with several technical limitations and can only query genes with relevant probe sets on the array. In recent years, sequencing of RNA (RNA-seq) has become a widely accessible method for the genome-wide quantification of transcripts, overcoming several of these limitations.

The advent of this new technology has led to the development of a wide variety of RNA-seq specific software. The two main software categories for RNA-seq are: 1) spliced aligners, which can align reads across intron-exon boundaries, and 2) transcript reconstruction algorithms, based either on aligned reads to a reference genome or de-novo assembly. These methods provide the basis for subsequent transcript quantification and differential expression analysis. In the first part of my thesis I benchmarked a comprehensive set of spliced aligners and transcript reconstruction tools for RNA-seq. Results indicated that choice of alignment software is critical for accurate interpretation of RNA-seq data. I also demonstrated that exons can be accurately identified from RNA-seq data, but that most methods fail to connect them into valid transcript isoforms, especially when transcriptome complexity increases.

In the second part of my thesis I studied the response to leukemia inhibitory factor (LIF) signalling in mouse embryonic stem (ES) cells. LIF-dependent Stat3 activation plays an important role in the maintenance of mouse ES cells in culture. Combined RNA-seq and ChIP-seq analyses identified a set of direct Stat3 targets induced following LIF exposure. Genes involved in transcriptional regulation were selected for functional assays, revealing that ES cells can be maintained in the absence of LIF through overexpression of either Klf4 or Tfcp2l1. Notably, most LIF induced genes not bound by Stat3 were bound by either Klf4 or Tfcp2l1. Integrating ChIP-seq data of known pluripotency factors, such as Oct4, Sox2, Nanog, and Klf4, demonstrated that proximity of their binding sites and Stat3 sites increased significantly for the subset of induced genes. Notably, of the Klf4 binding sites associated with Stat3-bound upregulated genes, 50% overlap directly with Stat3. These results suggest that Stat3 physically interacts with core pluripotency regulators to induce expression of its target genes.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

cDNA complementary DNA

CDS coding sequence

ChIP-seq chromatin immunoprecipitation followed by sequencing

DNA deoxyribonucleic acid

EG cells embryonic germ cells

ENCODE encyclopaedia of human DNA elements

ES cells embryonic stem cells

FCS Fetal calf serum

iPS cells induced pluripotent stem cells

kb kilo base

LIF leukemia inhibitory factor

lncRNA long non-coding RNA

ncRNA non-coding RNA

NGS next-generation sequencing

RGASP RNA-seq Genome Annotation Assessment Project

nt nucleotide

PCR polymerase chain reaction

qPCR quantitative PCR

RNA ribonucleic acid

RNA-seq RNA sequencing

rRNA ribosomal RNA

Stat3 Signal transducer and activator of transcription 3

TSS transcription start site

UTR untranslated region

Chapter 1

Introduction

The identity of a cell is determined by the set of actively transcribed genes and their post-transcriptional regulation. Comparing gene expression between conditions or developmental stages can identify genes involved in certain pathways or specifically expressed in certain tissues. Technologies to study gene expression and regulation have drastically improved during the last few decades. In particular, the advances in sequencing technologies allow the study of genetics and epigenetics at unprecedented resolution. Gene expression can be measured using RNA-sequencing (RNA-seq), which provides a read-out of a complete transcriptome and enables the quantification of expressed transcripts as well as the detection of novel genes and isoform composition. Combining gene expression measures with chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) can provide further insights into the mechanisms by which protein-DNA interactions, chromatin modifications and nucleosome positioning regulate gene expression.

The advent of RNA-seq has triggered the development of a whole range of methods tailored for this kind of data, such as spliced aligners, which can map reads across exon-intron boundaries, and transcript reconstruction methods, which assemble short fragments into full transcript isoforms. These methods form the foundation of any RNA-seq based transcriptome study and all subsequent analyses depend on the initial alignment and transcript reconstruction. Therefore, the right method choice is crucial for a correct interpretation of RNA-seq data. As no comprehensive comparison of these methods exists to date, there is no consensus which methods are appropriate for which

application. Depending whether the main aim is to measure gene expression of annotated genes or to identify novel transcripts, different methods might be required.

The aim of this thesis is to first assess the performance of RNA-seq aligners (Chapter 2) and transcriptome reconstruction algorithms (Chapter 3). The insights gained from this evaluation then informed the analysis of RNA-seq data from mouse embryonic stem cells upon leukemia inhibitory factor stimulation. These data were combined with the results from ChIP-seq experiments to study the mechanisms of LIF-mediated activation of the Jak/Stat3 signalling pathway and its role in the regulation of self-renewal and pluripotency (Chapter 4).

1.1 High-throughput sequencing

The field of genomics experienced a major breakthrough when Watson and Crick solved the structure of DNA (Watson and Crick, 1953) and a few years later the genetic code – sets of three nucleotides encoding an amino acid (Crick et al., 1961) – was deciphered. The development of DNA sequencing methods (Section 1.1.1) enabled the identification of genes encoded within the DNA. Soon researchers sought to sequence whole genomes and to identify all coding genes. The first complete genome to be sequenced was the genome of the ΦX174 bacteriophage (Sanger et al., 1977b). The first larger genomes to be completed were of *Escherichia coli* (Blattner et al., 1997) and *Saccharomyces cerevisiae* (Goffeau et al., 1996). In the 1980s the human genome project was launched aiming at decoding the human genome sequence, and in 2004 the first draft was completed (Lander et al., 2001; Venter et al., 2001; International Human Genome Sequencing Consortium, 2004). The high demand for sequencing information has lead to the development of new sequencing methods, resulting in a major speed-up enabling simultaneously sequencing of millions of sequences (Section 1.1.2).

1.1.1 First generation sequencing

In 1975 Frederick Sanger and and Alan Coulson published a method to sequence DNA by primed synthesis with DNA polymerase (Sanger and Coulson, 1975). In 1977 they improved this method by using chain-terminating inhibitors (Sanger et al., 1977a). In the same year Allan Maxam and Walter Gilbert published a DNA sequencing method based

chemical modification of DNA and subsequent cleavage at specific bases (Maxam and Gilbert, 1977). However, the chain-termination method developed by Sanger (therefore also often referred to as Sanger sequencing) was easier to use and required fewer toxic chemicals and thus became soon the method of choice. It formed the foundation for the first generation of DNA sequencers. Adding fluorescent tags to the chain terminators enabled the automated readout of the sequence (Smith et al., 1986). Using capillary electrophoresis instead of electrophoresis on acrylamide gels resulted in another speed increase (Hutchison, 2007). Reads of length up to 700 bp could be sequenced with this technique, but throughput was limited.

1.1.2 Second generation sequencing

Automated Sanger sequencing is now often referred to as *first-generation* sequencing and newer methods as *next-generation sequencing* (NGS) or also *second generation* sequencing (Mardis, 2008; Metzker, 2010). These methods are capable of generating a large amount of data in a short time frame, but at the cost of shorter read length.

There are several different approaches to next-generation sequencing, but the most popular one is the one of the Illumina platform. The approach implemented in the Illumina Genome Analyzer is based on the concept of *sequencing by synthesis* (Bentley et al., 2008). Instead of using beads, fragments are randomly bound to a solid surface, the flow cell, and clusters are build by bridge amplification (Figure 1.1a). All four nucleotides are added simultaneously onto the flow cell during each cycle of the sequencing process. Each nucleotide is labelled with a different dye and has a chemically blocked 3′-OH group, enabling the incorporation of only one nucleotide per sequencing cycle. After washing away unbound nucleotides, the incorporated nucleotides are identified in an imaging step. The blocking group is removed and the next sequencing cycle can begin. Two sequencing cycles are illustrated in Figure 1.1b, whereas Figure 1.1c illustrates a sequence of images determining the nucleotide sequence of each cluster. In *paired-end* sequencing both ends of a fragment are sequenced. After the first read has been sequenced the template strand is used to generate a bridge and re-synthesise the second strand, which is then used as template strand for the second read of the read pair. Depending on the fragment length and the read length, read pairs will map to the genome with similar distances. Anomalous paired reads can indicate structural varia-

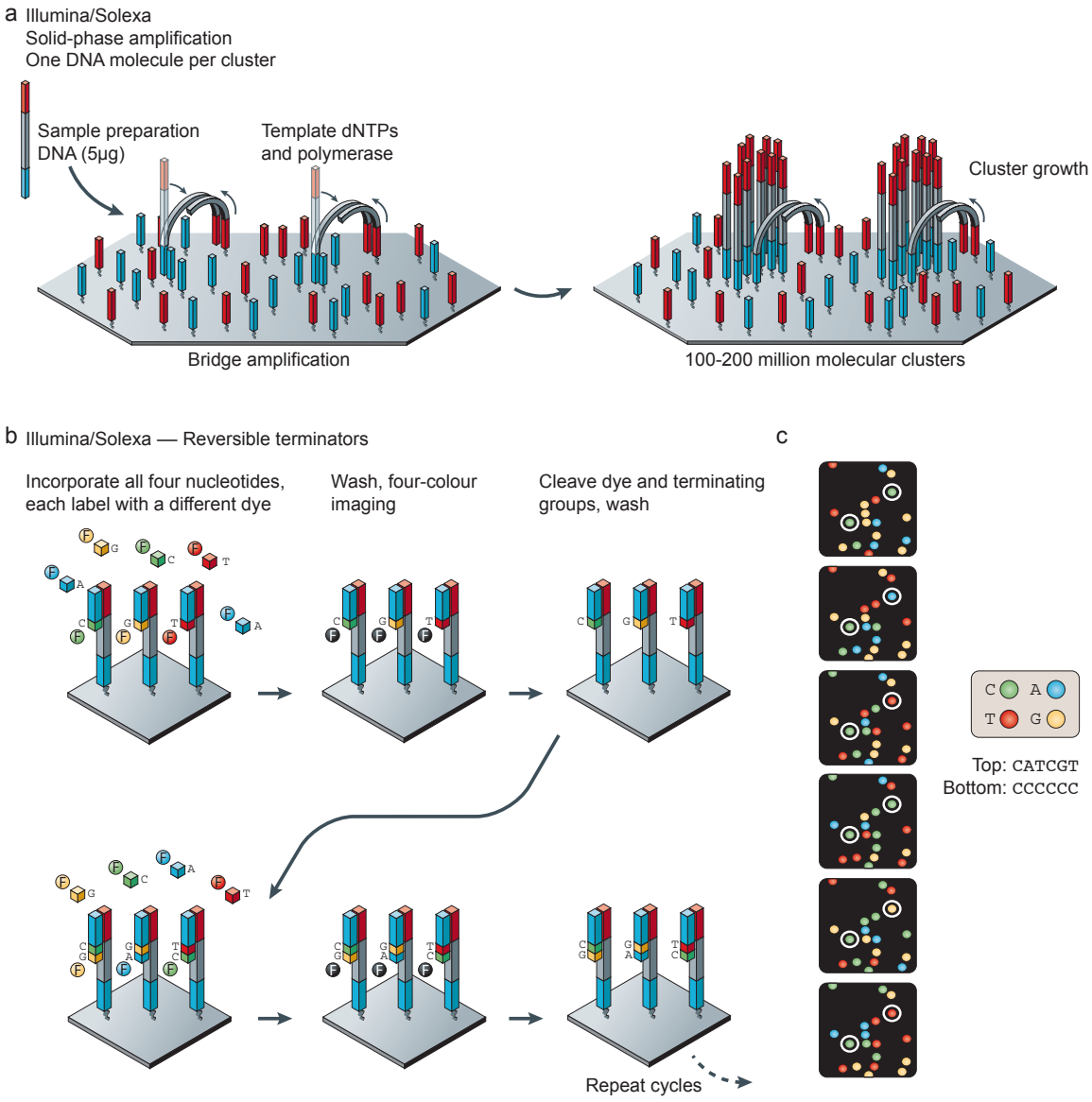


Figure 1.1: Overview of the Illumina sequencing technology. (a) Solid-phase amplification is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. (b) The four-colour cyclic reversible termination (CRT) method uses Illumina/Solexa's reversible terminator chemistry using solid-phase-amplified template clusters (shown as single templates for illustrative purposes). Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3′-OH group. (c) The four-colour images highlight the sequencing data from two clonally amplified templates. (Modified from Metzker, 2010)

tions (Bentley et al., 2008). To derive longe-range paired-end fragments longer than a 1kb, fragment ends are marked by the incorporation of biotinylated nucleotides. The fragment is then circularised, randomly fragmented and the biotinylated fragments are recovered for paired-end sequencing as described before.

There are other sequencing platforms with conceptually similar strategies. The Helicos Genetic Analysis System platform is also based on the concept of sequencing by synthesis while the SOLiD sequencing technology developed by Life Technologies is based on the concept of sequencing by ligation.

1.1.3 Applications

Population-scale sequencing, such as the 1000 genomes project, enables the study of genetic variation and their corresponding phenotypes within a population (1000 Genomes Project Consortium et al., 2010). In addition to single nucleotide polymorphisms (SNPs) and small indels (insertions or deletions), larger structural variations can be identified from paired-end genome sequencing data (Korbel et al., 2009), and differences in read coverage can be used to detect copy number variations (Mills et al., 2011). Converting RNA to cDNA and subsequent sequencing enables the analysis of gene expression in unprecedented detail (Morin et al., 2008; Mortazavi et al., 2008; Nookaew et al., 2012).

Within the field of epigenetics, chromatin modifications or binding profiles of DNA-binding proteins, such as transcription factors, can be analysed using chromatin immunoprecipitation sequencing (ChIP-seq, Johnson et al., 2007; Barski et al., 2007). DNA methylation patterns can be analysed using bisulfide sequencing (Bock, 2012; Cokus et al., 2008), and the three-dimensional structure of the genome can be studied using chromosome conformation capture methods, enabling the identification of interactions between distal genomic regions (Zhao et al., 2006; Lieberman-Aiden et al., 2009).

1.2 Gene expression profiling

The process of gene expression, protein synthesis and post-translational regulations, is controlled at various levels. Chromatin structure and DNA-binding proteins influence transcription rate, while RNA-binding proteins regulate RNA-stability and translation rate. Post-translational modification and protein localisation further control the activity

of proteins. Studying changes in gene expression between conditions, different tissues and developmental stages provides insights into gene function as some genes display tissue-specific expression profiles. Therefore, methods measuring gene expression of all genes present in a sample are required to obtain an unbiased read-out of the entire transcriptome.

1.2.1 Gene expression profiling using microarrays

For many years DNA microarrays have been the technology of choice for gene expression profiling. This technology is based on the principle of Southern Blotting where fragmented DNA is attached to a surface and then probed with a known labelled DNA sequence (Southern, 1975). If the DNA sequence of interest is within the sample, the probe will hybridise with it and can be detected after washing away remaining probes.

Microarrays enable the probing of many DNA sequences at the same time. Known probe sequences are attached or directly synthesised on defined spots on a solid surface, such as glass or a silicon chip (Schena et al., 1995; Lockhart et al., 1996; Shalon et al., 1996). Fluorescently labelled fragmented cDNA or cRNA (antisense RNA), that is the target sequences, are then added to hybridise with the probes on the array. The strength of the bond depends on the level of complementarity between the probe and the target sequence. Fully complementary targets bind more strongly than partially complementary targets. After hybridisation the remaining target sequences are washed away. During this process also weak bonds are broken off. Finally the fluorescent signal of each spot is measured in an imaging step. The strength of the signal provides information about the amount of target sequence present in the sample.

Microarrays have been adapted to various applications. In addition to gene expression profiling and detection of differentially expressed genes (Gautier et al., 2004), microarrays can be used to study binding profiles of DNA binding proteins (Ren et al., 2000), novel transcripts using tiling arrays (Bertone et al., 2004; Huber et al., 2006; Xu et al., 2011) and alternative splicing using exon-arrays (French et al., 2007; Gardina et al., 2006).

Microarrays, however, come with certain limitations. As hybridisation efficiency differs between probes, microarrays do not provide absolute gene expression estimates (Gautier et al., 2004). Nevertheless, over the years methods for the detection of differentially

expressed genes between samples have been optimised (Allison et al., 2006). Another drawback of microarrays is a high background signal caused by unspecific hybridisation. The dynamic range of gene expression profiling is further limited by the saturation of the fluorescent signal, which precludes proper quantification of highly expressed genes. Finally, the set of interrogated transcripts is limited by the probe set present on a microarray.

1.2.2 Expression profiling using RNA sequencing

Sequencing of mRNA (RNA-seq) overcomes several of the limitations of microarrays. A single experiment can be used to quantify known genes, identify novel transcripts and analyse isoform composition expressed within a sample (Mortazavi et al., 2008). Due to sinking sequencing costs RNA-seq is becoming more and more accessible and is slowly replacing gene expression analysis using microarrays.

Several studies have compared the performance of gene expression microarrays and RNA-seq (Marioni et al., 2008; Hoen et al., 2008; Malone and Oliver, 2011; Nookaew et al., 2012). All studies found RNA-seq to correlate well with results from microarrays, however, RNA-seq provides a wider range of expression levels and thus a more detailed insight into gene expression. RNA-seq was further found to be more robust and show a higher inter-lab portability (Hoen et al., 2008) and to be superior to microarrays in the context of transcript discovery (Marioni et al., 2008; Nookaew et al., 2012; Malone and Oliver, 2011).

Further comparisons between RNA-seq and exon arrays have found high correlations on the level of exon expression estimated between the two methodologies (Bradford et al., 2010; Raghavachari et al., 2012), but RNA-seq was superior for the detection of novel features and isoforms.

1.2.2.1 RNA-seq experimental workflow

A typical RNA-seq experiment workflow consists of several steps (Figure 1.2). First, ribosomal RNA (rRNA) has to be removed from the sample since it is the predominant RNA species in a cell. Next RNA molecules get fragmented either before or after they are reverse transcribed into double-stranded cDNA. Often a size selection step is performed following adapter ligation and PCR amplification. The cDNA library can then

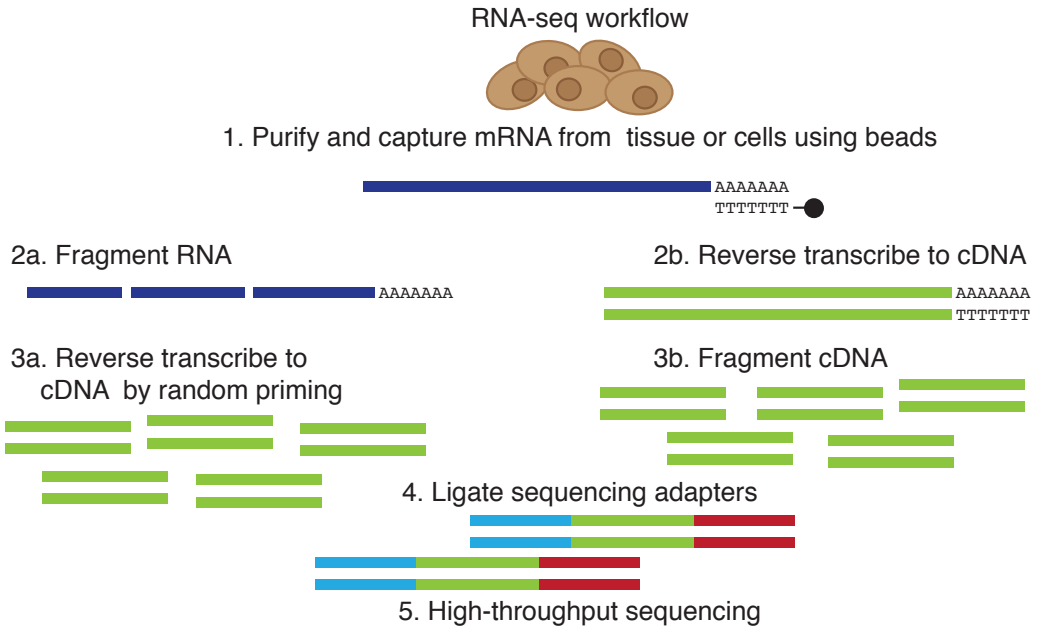


Figure 1.2: RNA-seq experimental workflow. The initial step of RNA-seq is the purification of mRNA. The RNA is either fragmented prior to reverse transcription (workflow a) or after reverse transcription (workflow b). The double stranded DNA fragments are then ligated to sequencing adapters for the subsequent sequencing on a next generation sequencer. Modified from Cullum et al., 2011.

be sequenced using for example the Illumina GenomeAnalyze or ABI SOLiD platforms. In the following, the single steps of RNA-seq library preparation are described in more detail.

Purification of mRNA Ribosomal RNA is the most abundant RNA species in the cell (Karpinets et al., 2006) and it must be depleted during library preparation to enrich for the RNA sequences of interest. Common methods are the enrichment for mRNA using poly(A) selection or targeted removal of ribosomal RNA (O'Neil et al., 2013). Two commonly used ribosomal depletion strategies are subtractive hybridisation with rRNA specific probes and the digestion with exonuclease that preferentially acts on rRNA (see He et al., 2010 for a comparison of these two strategies).

RNA fragmentation Current sequencers are not able to sequence full length transcripts. Therefore, RNA is fragmented to obtain fragments of 200 to 500 bp length. Fragmentation can be achieved using RNA hydrolysis, nebulisation or sonication.

It can also be performed after cDNA synthesis using DNAse I treatment or sonication. However, this increases the likelyhood of 3′ bias within the data, as reverse transcription using oligo-dT primers is often incomplete (Wang et al., 2009).

cDNA synthesis Most sequencers do not process RNA, but only double-stranded DNA. Therefore, RNA fragments need to be reverse transcribed into DNA molecules, which can be achieved through hybridisation of primers to the RNA sequence. In the case of cDNA synthesis following RNA fragmentation, sequences of six random bases (random hexamers) that bind to random positions within the RNA fragment are commonly used. They can be combined with short sequences of Ts (oligo-dTs) complementary to polyA tails. The sole use of oligo-dTs is only applicable if cDNA synthesis is performed prior to fragmentation.

Adapter ligation and PCR amplification Before adapter ligation the cDNA is treated to generate blunt ends, followed by ligation of adapter sequences to both ends of the cDNA. PCR amplification is performed to create enough material to sequence.

Several adaptations of this general workflow exist. For example, fragment circulation can be used to generate large-range paired-end data (Bentley et al., 2008) while dUTP second strand marking can be used to preserve strand orientation. In a comparison of different strand-specific RNA-seq protocols, the dUTP protocol performed most favourable (Levin et al., 2010).

1.2.2.2 RNA-seq alignment

Challenges of RNA-seq mapping include dealing with reads that map to multiple places in the genome and reads that span exon-intron junctions (Ozsolak and Milos, 2011). In the early days of RNA-seq reads were often aligned against transcript sequences derived from annotated genomes using a conventional, not splice-aware, aligner as for example BWA or Bowtie (Li and Durbin, 2009; Langmead and Salzberg, 2012; Langmead et al., 2009). This approach avoids the problem of having to split up reads if they span an exon-intron junction. However, this simplification comes with a drawback: only annotated transcripts can be detected. The potential of RNA-seq to detect novel isoforms of annotated genes as well as unkown genes is then lost.

In the last few years several alignment strategies tailored for the application of RNA-seq data have been developed. One of the first methods was TopHat (Trapnell et al., 2009), where initial read alignment against the genome using Bowtie (Langmead and Salzberg, 2012; Langmead et al., 2009) is used to identify islands of read coverage (representing potential exons). As a next step canonical splice signals (GT-AG, GC-AG and AT-AC) are identified between islands to determine potential intron boundaries. Previously unmapped reads are then re-aligned against these. In a later version of TopHat, reads longer than 76 bp are split into smaller segments, which are aligned independently. Unmapped segments can then be placed by extending alignments of adjacent segments (see supplement of Trapnell et al., 2010). To date, several spliced aligners are available, including PALMapper (Jean et al., 2010), GSNAP (Wu and Nacu, 2010), MapSplice (Wang et al., 2010a), and STAR (Dobin et al., 2013).

1.2.2.3 The Sequence Alignment/Map format

The Sequence Alignment/Map (SAM) format, a tab-delimited text format, has become the standard way of storing sequence alignments (Li et al., 2009). It consists of an optional header section and an alignment section. The header section contains basic information about the sequencing platform, which genome assembly the reads were aligned against and which tool was used to obtain the alignment. The alignment section contains one line per read mapped to a given reference sequence. Each alignment line has eleven mandatory fields, as shown in Table 1.1. Besides the actual read position, the line also contains information about other segments of a sequenced fragment such as the other half of the read pair in paired-end sequencing. In the following three fields of the SAM format are described that were used to calculate most of the alignment metrics in Chapter 2.

The flag field indicates whether the data is single- or paired-end, along with other information about the read mapping. In case of paired-end data the flag records whether the current read is the first or second mate of a read pair; whether the other mate is mapped as well; and to which strand the mate maps to. The flag can also be used to flag an alignment as secondary in the case of multiple possible mappings for a read (see Table 1.1 for a full list).

The **cigar string** specifies whether the full read was mapped to the reference; whether the alignment contains indels; and whether sequences at the start or end of the read were clipped. Each of these possibilities are known as *operations*, and are abbreviated using a single upper-case letter, such as M, X or N (see Table 1.1 for a full list).

In addition to the mandatory fields, each alignment line can contain an arbitrary number of optional tags following the TAG:TYPE:VALUE format. The **MD tag** contains more detailed information about mismatch positions and has to be consistent with the cigar string. This tag enables SNP analysis without looking up the reference sequence. Its format is described with the following regular expression: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*. Not many aligners provide this optional tag. It can be added to a SAM file using the SAMTools calmd functionality (Li et al., 2009).

1.2.2.4 Gene expression quantification

Read alignments are used to quantify gene expression levels. The raw gene read count is defined as the number of reads that map to exons of known genes. These raw counts, however, cannot be used to compare gene expression levels between genes within the same sequencing run or the same gene between different sequencing runs. A long gene will have a higher read count compared to a short gene expressed at the same level. Equally, a gene will have a higher read count if the sequencing run resulted in 40M reads instead of 20M reads.

Therefore, it is necessary to normalise the raw counts both for gene length and total read number. One popular way of normalising raw counts for length and total number of reads are RPKM values, which stands for reads per kilobase of exon model per million mapped reads (Mortazavi et al., 2008). FPKM (fragments per kilobase of exon model per million mapped fragments) values are the RPKM equivalent for paired-end data where the two reads coming from one fragment are counted as one (Trapnell et al., 2010). Several slight adaptations to this normalisation have been proposed, e.g., only using uniquely mappable regions of genes for length normalisation (Lee et al., 2011).

Gene expression quantification is impaired by different biases present in RNA-seq data. There are two main kind of biases: sequence dependent bias and positional bias. The latter one can often be observed when cDNA synthesis is performed prior to frag-

Table 1.1: The SAM format. The mandatory fields of the alignment line and description of the flag and the cigar string (field 2 and 6 of the alignment line).

Alignment line:		
Column	**Field**	**Description**
1	QNAME	read name
2	FLAG	bitwise flag
3	RNAME	chromosome name
4	POS	1-based leftmost mapping position
5	MAPQ	mapping quality
6	CIGAR	cigar string
7	RNEXT	chromosome name for next segment
8	PNEXT	mapping position of next segment
9	TLEN	template length
12	SEQ	read sequence
11	QUAL	read base quality
FLAG, bitwise flag:		
	Bit	**Description**
	0x1	template having multiple seqments in sequencing
	0x2	each segment properly aligned according to the aligner
	0x4	segment unmapped
	0x8	next segment in the template unmapped
	0x10	SEQ being reverse complemented
	0x20	SEQ of the next segment in the template being reversed
	0x40	the first segment in the template
	0x80	the last segment in the template
	0x100	secondary alignment
	0x200	not passing quality control
	0x400	PCR or optical duplicate
CIGAR, cigar string:		
	Operation	**Description**
	M	alignment match (can be a sequence match or mismatch)
	I	insertion to the reference
	D	deletion from the reference
	N	skipped region from the reference
	S	soft clipping (clipped sequences present in SEQ)
	H	hard clipping (clipped sequences NOT present in SEQ)
	P	padding (silent deletion from padded reference)
	=	sequence match
	X	sequence mismatch

mentation and priming is performed exclusively with oligo-dT primers, leading to a 3′-bias (Wang et al., 2009). Estimating the global bias and accounting for it can improve gene expression estimates (Wu et al., 2011).

Sequence-dependent bias is when reads tend to have certain nucleotides overrepresented at the start of their sequence. Using random hexamers for reverse transcription seems to be the cause of this bias as the hexamers differ in their priming efficiency. Priming is therefore not entirely random, but some primers are more likely to hybridise than others. There are several methods available to correct for this kind of bias when generating gene counts. These methods reweigh the count of each read alignment depending on their sequence instead of counting each read alignments exactly once (Hansen et al., 2010; Roberts et al., 2011b; Jones et al., 2012).

1.2.2.5 Differential gene expression analysis

Many tools for differential expression analysis work on raw counts. For differential expression analysis it is sufficient to normalise for total read count, whereas for absolute expression estimates it is necessary to take both read count and gene length into account. The expression of the same gene is compared within different condition or tissues and therefore normalising for gene length is not necessary. Ideally the samples will have been processed together, so any bias present in the data should affect all samples to a similar degree and therefore not affect differential gene expression. Differential expression analysis between data sets that have been generated using different library construction protocols is more complicated and will require bias correction. Commonly used methods for differential expression analysis are the Bioconductor (Gentleman et al., 2004) packages DESeq (Anders and Huber, 2010), DEGSeq (Wang et al., 2010b), edgeR (Robinson and Smyth, 2007; Robinson et al., 2010), and baySeq (Hardcastle and Kelly, 2010).

The biological variance of a gene, that is the natural variance of expression levels of a gene within the same conditions, has to be estimated to identify differentially expressed genes. Accordingly, the usage of biological replicates is crucial for the estimation of biological variance. Originally, differential expression between conditions was tested using a Poisson model for read counts. This model provides a good fit for technical replicates (Marioni et al., 2008). However, samples from biological replicates show higher variance than predicted by the Poisson model (Robinson et al., 2010; Anders and Hu-

ber, 2010). The Negative Binomial (BN) distribution has been suggested to model read counts across samples to capture this over-dispersion (Robinson et al., 2010). Most methods require biological replicates to estimate biological variation. DESeq, however, can estimate variation by sharing information across genes. This will typically overestimate the real variation in the data and lead to very conservative differential expression calls, and is therefore not recommended whenever biological replicates are available (Anders and Huber, 2010).

The aforementioned methods work on count data and do not consider ambiguously mapped reads. They also do not consider gene structure and cannot identify isoform switching, where genes are expressed at the same level in two or more conditions, but where the major isoform is different. Another approach is to use a probabilistic method to quantify transcript isoforms (Bohnert and Rätsch, 2010; Li and Dewey, 2011; Du et al., 2012; Li et al., 2010) and subsequently use these isoform quantifications rather than gene counts for differential expression analysis. Finally, there are methods that are designated for the identification of differential expressed isoforms and isoform switching from alignments given a gene model as rDiff (Drewe et al., 2013), cuffdiff (Trapnell et al., 2012), MISO (Katz et al., 2010), DSGseq (Wang et al., 2013) and DEXSeq (Anders et al., 2012; Reyes et al., 2013).

1.3 Studying protein-DNA interactions with ChIP sequencing

The identification of chromatin modifications and protein-DNA interactions is essential for a full understanding of transcriptional regulation. Mapping transcription factor binding sites, chromatin modifications or components of the core transcriptional machinery, provides insights into the gene regulatory networks that control transcription.

Chromatin immunoprecipitation (ChIP) followed by either hybridisation to a microarray (ChIP-chip) or deep sequencing (ChIP-seq) enables the genome-wide detection of protein-DNA binding events (Park, 2009; Pepke et al., 2009; Furey, 2012).

1.3.1 ChIP-seq experimental workflow

A ChIP-seq workflow differs depending on the type of interaction of interest. To detect binding sites of DNA-binding proteins, formaldehyde treatment crosslinks proteins to the DNA onto which they are bound. The DNA is then fragmented using sonication or digestion by exonucleases. Antibodies specific to the protein of interest are used to enrich for the DNA-protein complex. After reversing the crosslinking the bound DNA can be analysed using microarray technologies or next-generation sequencing.

To identify nucleosome positions or histone modifications, micrococcal nuclease (MNase) digestion without prior crosslinking is more commonly used to fragment the DNA. It removes linker DNA more efficiently than sonication and therefore allows more precise mapping of single nucleosomes.

In order to identify significant binding sites, an additional control experiment is performed. This can either use input DNA from the ChIP experiment or be derived using an unspecific antibody. Only regions that show a statistically significant enrichment of signal in the treatment experiment compared to the control experiment are regarded as peaks.

1.3.2 Mapping strategy and peak detection

ChIP-seq produces DNA sequences that represent the protein-bound DNA. Therefore, there is no need for spliced alignment and traditional DNA sequencer as Bowtie or MAQ can be used to align the reads to the reference sequence (Langmead et al., 2009; Langmead and Salzberg, 2012; Li et al., 2008a).

After obtaining an alignment, the next step is to find regions that are significantly enriched in the ChIP sample compared to the control sample. Several so called 'peak callers' are available, as for example MACS (Zhang et al., 2008), PeakSeq (Rozowsky et al., 2009), FindPeaks (Fejes et al., 2008) and PeakFinder (Glynn et al., 2004). The simplest strategy for peak detection is to calculate the number of reads within a window and the enrichment relative to the number of reads in the control. More advanced methods make use of the directionality of the reads. As the fragments are sequenced from the $5'$ end, the positions of the aligned reads should form two distributions, one on each strand, with a consistent distance between the two peaks of the distributions. A

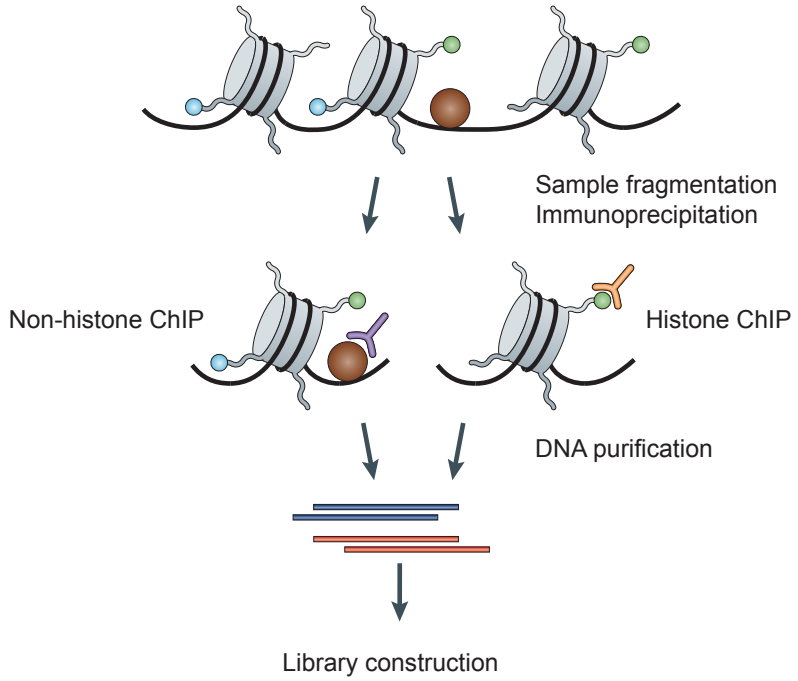


Figure 1.3: ChIP-seq experimental workflow. The specific DNA sites that interact with transcription factors or other chromatin-associated proteins (non-histone ChIP) and sites that correspond to modified nucleosomes (histone ChIP) can be profiled, using chromatin immunoprecipitation (ChIP) followed by massive parallel sequencing. The ChIP process enriches the crosslinked proteins or modified nucleosomes of interest using an antibody specific to the protein or the histone modification. Purified DNA can then be sequenced on any of the next-generation platforms. (Modified from Park, 2009)

combined profile is then calculated by shifting each distribution towards the center or by extending all reads to an estimated fragment size and adding the fragments together.

This approach is mostly applicable to sharp peaks as for example transcription factor binding sites. ChIP-seq analysis of histone modifications typically results in much broader peaks, which initially posed an additional challenge to peak detectors. Today several specialised methods are available for the detection of broader peak domains (Xu et al., 2008; Zang et al., 2009).

Statistical significance of a detected peak needs to account for enrichment relative to the control sample as well as for the absolute number of reads. Both a Poisson or a binomial model can be used for this purpose. Furthermore, peaks can be scored by considering how similar the read distributions on the two strands are, and whether the distance between the two peaks is close to the distance expected based on the fragment length.

1.3.3 Downstream analysis

Following peak calling the downstream analysis depends on the biological process under investigation. For transcription factor binding, a common follow-up analysis is the detection of enriched sequence motifs. These motifs can indicate sequence-specific binding of transcription factors. To identify such motifs, the sequences of top-scoring peaks is used by motif-finding algorithms such as Meme (Machanick and Bailey, 2011), MDScan (Liu et al., 2002) and Weeder (Pavesi et al., 2004). To get more accurate results it can be advantageous to select only the region surrounding the peak summits for motif discovery. This can be done using PeakAnalyzer (Salmon-Divon et al., 2010), which furthermore identifies individual peaks consisting of several merged peaks and splits them into separate peaks. Once a motif has been identified, methods like TOMTOM can be used to find similar known motifs (Gupta et al., 2007).

Several methods have been developed for more specialised tasks, as for example identifying slight differences in motifs depending whether a certain cofactor is present or not (Mason et al., 2010), detecting several binding modes within the same set of sequences (Narlikar, 2013); or the co-occurrence of transcription factor binding motifs (Ha et al., 2012).

Another common downstream analysis is the annotation of peaks to known genomic features (Salmon-Divon et al., 2010; Shin et al., 2009). This can reveal whether a factor predominantly maps to promoter regions, to exons or other features. More interestingly, peaks can be annotated with potential target genes (Salmon-Divon et al., 2010; Zhu et al., 2010; Shin et al., 2009; Cheng et al., 2011). However, close proximity does not always indicate that a binding site has a functional role related to the most proximal gene. A combinatorial approach integrating expression data as well can help to distinguish between functional and non-functional binding events (Ouyang et al., 2009; Vallania et al., 2009).

1.4 The biology of embryonic stem cells

The fertilised oocyte has the unique potential to generate all cell types of the developing embryo as well as of the *trophoblast*, which generates the extra-embryonic tissue that will build the placenta. The oocyte is therefore referred to as *totipotent*. Cells of the inner cell mass (ICM, see Figure 1.4), at a later stage in embryonic development, lose the ability to contribute to extra-embryonic tissues. However, these cells can still give rise to all cell types of the growing embryo and are therefore referred to as *pluripotent*.

Embryonic stem (ES) cells are the pluripotent *in vitro* counterpart of cells in the ICM (Nichols and Smith, 2012). These cells can be cultured in an undifferentiated state indefinitely and can be differentiated *in vitro* into various cell types of the three primary germ layers: the *endoderm*, *mesoderm* and *ectoderm*. These germ layers give rise to all somatic cells in the adult organism, such as the stomach, colon, liver and pancreas (endoderm), muscles and the hematopoietic system (mesoderm), and neurons and skin cells (ectoderm). Upon injection into a blastocyst, ES cells contribute to all three germ layers including the germ line of the embryo resulting in the formation of a chimeric animal containing both cells derived from the original zygote and cells originating from the injected stem cells.

The potential of pluripotent stem cells to generate all cell types of an adult body makes them perfect candidates for cell therapy, where cells are introduced into a tissue in order to replace defect or degenerated cells. However, this requires protocols to efficiently differentiate stem cells into the desired cell type and to test their ability to integrate and execute their function. For some cell types, it might not be possible to

differentiate these cells outside of their normal environment and it will be necessary to model their biological environment. Also it is extremely important to obtain highly pure cell populations. Contamination with remaining undifferentiated stem cells might lead to tumour formation following cell implantation.

There are ethical implications of deriving human ES cells as their derivation requires the destruction of a human embryo (Ramalho-Santos, 2011). Therefore, initially research focused on the use of adult stem cells for clinical purposes as they can be derived for example from bone marrow and blood and also decrease the risk of cells being rejected after implantation. Adult stem cells have been used to treat spinal cord injuries, liver cirrhosis and peripheral vascular disease (Dedeepiya et al., 2012; Terai et al., 2006; Subrammaniyan et al., 2011). The success of these initial cell therapies has lead to the creation of stem cell banks, where people can store their adult stem cells for later use. However, adult stem cells often have limited differentiation potential and cannot generate any required cell type.

More recently, the discovery of induced pluripotent stem (iPS, Takahashi and Yamanaka, 2006; Takahashi et al., 2007) cells has revolutionised the field as they enable the generation of patient-specific stem cells from for example the patient's skin cells. However, before iPS cells can be routinely used for cell therapy more efficient reprogramming protocols are needed and the differentiation potential of iPS cells has to be studied. It is still unclear to what extent iPS cells are identical to truly pluripotent stem cells and whether they can efficiently be differentiated into any cell type, or whether they retain epigenetic memory of their origin, which may bias subsequent differentiation.

Therefore, there remains a need to study how pluripotency is established and maintained in the embryo and their *in vitro* counterpart embryonic stem cells. Gaining a deeper understanding of the pluripotent state will also improve our understanding of the medical potential of induced pluripotent stem cells and enable the development of more efficient reprogramming protocols.

To study the pluripotent state, it is important to study the set of active genes and their regulation in embryonic stem cells. Gene expression is regulated on several levels. Epigenetic modifications, such as DNA methylation and chromatin modification, regulate the accessibility of genes for the transcriptional machinery, and DNA-binding proteins can further activate or repress transcription. The translation and degeneration rate of the mRNA can be modified by RNA-binding proteins post-transcriptionally, thereby

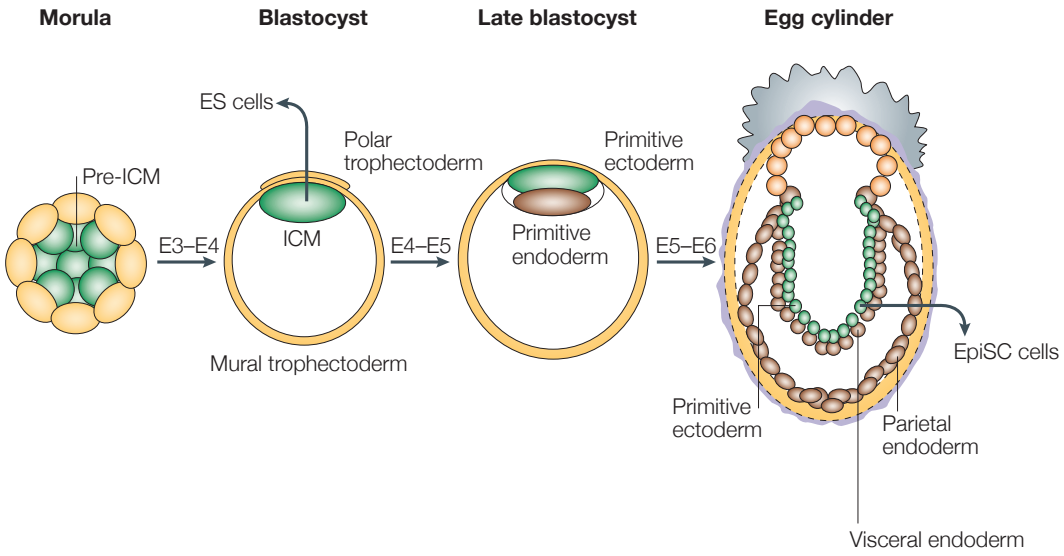


Figure 1.4: Embryonic stem cells are derived from the pre-implantation ICM. Depicted is the development of the mouse embryo during embryonic days 2.5 to 6 (E2.5-E6). Pluripotent cells are marked in green. Embryonic stem cells are derived from the pre-implantation blastocyst between time point E3-E4. At this stage *in vivo* implantation occurs. EpiSCs can be derived from the post-implantation epiblast (E5-E6). Modified from Boiani and Schöler, 2005.

regulating the amount of protein synthesised from a given mRNA. Finally, proteins can be activated, inactivated or marked for degradation by post-translational modifications, such as phosphorylation. This results in a very complex interconnected regulatory network stabilising the pluripotent state. Many recent studies have focused on gene regulation through DNA methylation, histone modifications and transcription factor binding in embryonic stem cells to gain insights into their regulatory network.

1.4.1 Types of pluripotent stem cells

Embryonic stem cells can be derived from the ICM of the pre-implantation blastocyst (see Figure 1.4, Evans and Kaufman, 1981; Martin, 1981). These cells can be maintained indefinitely in an undifferentiated state when grown in co-culture with fibroblastic feeder cells and in the presence of calf serum.

Feeder cells secrete leukaemia inhibitory factor (LIF) and supplementing the culture media with LIF enables the maintenance of ES cells in the absence of feeders (Smith and Hooper, 1987; Smith et al., 1988; Williams et al., 1988). LIF activates the transcription

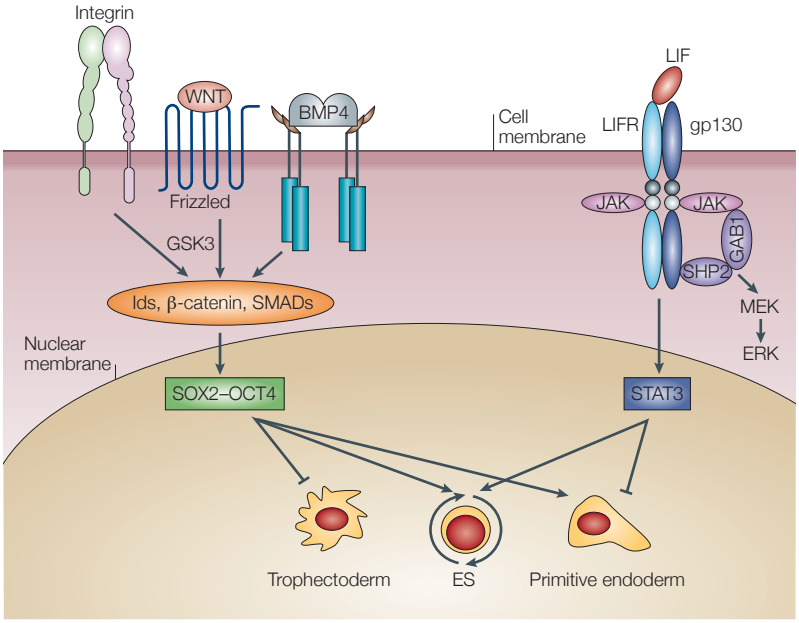


Figure 1.5: Signalling pathways involved in ES cell maintenance. LIF and BMP4 signalling maintain ES cell pluripotency by repressing differentiation along the primitive endoderm and the trophectoderm lineage, respectively. Modified from Boiani and Schöler, 2005.

factor Stat3 and causes its translocation into the nucleus where it activates expression of its target genes (Niwa et al., 1998; Matsuda et al., 1999). However, LIF signalling alone is insufficient to maintain ES cells in the absence of fetal calf serum, but BMP4 (bone morphogenic protein 4) together with LIF-induced Stat3 activation can sustain ES cells self-renewal in the absence of serum (Ying et al., 2003). In LIF alone, ES cells will eventually differentiate into neuroectoderm, while in BMP4 alone, they differentiate into non-neural lineages (Ying et al., 2003). This suggests that LIF and BMP4 support self-renewal by repressing differentiation along different lineages (Figure 1.5).

Another way of preventing the differentiation of ES cells is the inhibition of the mitogen-activated protein kinase (MAPK) and the glycogen synthase kinase 3 (Gsk3) using two small chemical inhibitors PD03 and Chiron (Ying et al., 2008; Wray et al., 2010). These culture conditions, commonly referred to as 2i, lead to major advances in the field. For the very first time it enabled the derivation of ES cells from hitherto recalcitrant strains, such as non-obese diabetic (NOD) mice embryos (Nichols et al., 2009), as well as the derivation of the first germline-competent rat ES cells (Li et al., 2008b; Buehr et al., 2008). It has been reported that the 2i system can also be used

Table 1.2: Characteristics of different pluripotent stem cells. Comparing the main characteristics of mouse ES cells and EpiSCs as well as human ES cells. Xa - active X chromosome; Xi - inactivated X chromosome

Properties	Mouse ES cells	Mouse EpiSCs	Human ES/iPS cells
Origin	ICM	Epiblast	ICM
Morphology	Dome shaped	Flat	Flat
Clonogenicity	High efficiency	Low efficiency	Low efficiency
Passaging	Insensitive	Sensitive	Sensitive
Contribution to chimaera	High efficiency	Very low efficiency	N/A
Female X Inactivation	XaXa	XaXi	XaXi
Culture condition	LIF/Stat3	TGFβ/Activin	TGFβ/Activin

to derive ES cells morphologically similar to mouse and rat ES cells from chicken and gecko (Nakanoh et al., 2013), suggesting that the mechanisms regulating pluripotency and the exit from the pluripotent state date back to the common ancestor of mammals, birds and reptiles.

Another type of pluripotent cells can be derived from the post-implantation epiblast (see Figure 1.4, Tesar et al., 2007; Brons et al., 2007). These post-implantation epiblast-derived stem cells (EpiSC) can robustly differentiate into the major somatic cell types as well as germ cells. However, EpiSC differ from mouse ES cells in their epigenetic state and culture requirements; they are dependent on FGF2 and activin. Furthermore, they exhibit lower clonogenicity efficiency than mouse ES cells and do not contribute effectively to chimeric animals unless they are injected into postimplantation embryos (Huang et al., 2012). Therefore they are referred to as *primed* pluripotent cells in contrast to the *naive* pluripotent mouse ES cells (Nichols and Smith, 2009; Nichols and Smith, 2012). However, culturing EpiSCs in high density on feeder cells and in the presence of serum/LIF, or in 2i/LIF, can reprogram EpiSCs to ES cells at low frequency (Hayashi and Surani, 2009; Bernemann et al., 2011).

The first primate ES stem cells were obtained from the rhesus macaque blastocyst (Thomson et al., 1995) and the common marmoset (Thomson et al., 1996). Using embryos from *in vitro* fertility treatment, the first human ES cell lines were derived in 1998 (Thomson et al., 1998). However, these cell lines exhibit several differences to mouse ES cells. They are not LIF responsive (Dahéron et al., 2004; Sumi et al., 2004; Humphrey et al., 2004) and are reliant on FGF and activin. These observations were

commonly attributed to species-specific differences. However, Austin Smith suggested that the human pluripotent cells might represent a later stage of embryonic development, rather than the human equivalent of mouse ES cells (Smith, 2001). The later discovery of EpiSCs supports this position as they share several characteristics with primate ES stem cells, see Table 1.2.

Mouse fibroblasts can be reprogrammed into a pluripotent like state through forced expression of four transcription factors: Oct4, Sox2, Klf4 and c-Myc, often referred to as the Yamanaka reprogramming factors (Takahashi and Yamanaka, 2006). These induced pluripotent stem cells (iPS cells) are able to contribute to chimeric animals. Human iPS cells can be obtained using the same four factors (Takahashi et al., 2007). However, reprogramming occurs at low efficiency and often via a partial reprogrammed state (pre-iPS cells), where the second X-chromosome (in female cells) has not been inactivated yet and endogenous Oct4 and Nanog are not stably expressed (Silva et al., 2008; Theunissen et al., 2011). Culturing pre-iPS cells in 2i/LIF supplemented media leads to fully reprogrammed iPS cells (Silva et al., 2008; Theunissen et al., 2011). It remains unclear how closely iPS cells resemble ES cells and whether they have the same developmental potential. If iPS cells retain some epigenetic memory of their origin, this can bias subsequent differentiation (Bar-Nur et al., 2011; Kim et al., 2010; Hu et al., 2010).

Human iPS cells share several features with EpiSC (see Table 1.2). Adding Nanog to the Yamanaka reprogramming factors (Takahashi and Yamanaka, 2006) during reprogramming can generate iPS cells that display morphological, molecular and functional properties of murine ES cells (Buecker et al., 2010). Similarly, ectopic expression of Oct4, Klf4 and Klf2 can convert human ES cells into cells which are more similar to mouse ES cells and can be maintained in 2i media supplemented with LIF (Hanna et al., 2010). Wang et al. show that a combination of the four Yamanaka reprogramming factors (Takahashi and Yamanaka, 2006) together with expression of RAR-γ and Lrh-1 accelerates reprogramming and generates human iPS cells which resemble mouse ES cells in gene expression and signalling dependency (Wang et al., 2011). However, all these 'naive' human ES cells are dependent on the expression of the transgene and become unstable in the absence of transgene expression.

Transgene independent promotion of human naive pluripotent stem cells can be established using defined conditions called naive human stel cell medium (Gafni et al.,

2013). It consists of 2i/LIF, p38i, JNKi together with FGF2 and TGF-β1 cytokine supplementation and Rho-associated coiled-coil kinases and protein kinase C inhibitors and enables the derivation of human pluripotent stem cells with similar characteristics to naive mouse ES cells in the absence of transgene expression.

1.4.2 Epigenetic landscape of embryonic stem cells

DNA methylation is defined as the addition of a methyl group to cytosine or adenine nucleotides of the DNA. This epigenetic mark is sustained through cell division and one of the main mechanisms for the stable maintenance of gene expression patterns through mitotic cell division (Holliday and Pugh, 1975). It is associated with gene silencing, imprinting and X-chromosome inactivation (Urnov and Wolffe, 2001; Wolffe and Matzke, 1999). Several studies have compared the methylation patterns in ES cells to those in differentiated cells and showed that they change during differentiation (Shiota et al., 2002; Meissner et al., 2008; Farthing et al., 2008; Laurent et al., 2010; Hawkins et al., 2010). For example, promoter regions of pluripotency factors are hypomethylated in ES cells and hypermethylated in differentiated cells (Farthing et al., 2008). Already EpiSC exhibit a different methylation pattern from ES cells (Senner et al., 2012). The 2i culture condition maintains the ES methylation pattern via Prdm14-mediated repression of de novo DNA methyltransferases (Leitch et al., 2013; Yamaji et al., 2013; Ficz et al., 2013).

DNA methylation is associated with chromatin modifications (Meissner et al., 2008; Hawkins et al., 2010). Chromatin structure and its modifications provide another level of gene expression regulation by modulating the accessibility of the DNA to the transcriptional machinery (Kouzarides, 2007). Embryonic stem cells exhibit significantly different epigenetic profiles compared to differentiated cells (Serrano et al., 2013; Mikkelsen et al., 2007). They are generally known to have a more open chromatin structure and more dynamic change-over rate of chromatin proteins (Meshorer et al., 2006). Activating chromatin marks as AcH3 and AcH4 and H3K36me2 and H3K4me3 are increased in ES cells and iPS cells compared to somatic cells (Meshorer et al., 2006), while the repressing mark H3K9me3 is reduced (Wen et al., 2009; Krejcí et al., 2009; Hawkins et al., 2010). Consistent with these observations, undifferentiated cells display a more global transcription pattern than differentiated cells (Efroni et al., 2008; Guenther et al., 2007; Golan-Mashiach et al., 2005).

Also non-expressed lineage-specific genes are associated with high levels of the activating chromatin marks H3K4me3 and AcH3 in ES cells. However, these markers of open chromatin are often found in combination of the repressing mark H3K27me3 (Azuara et al., 2006; Bernstein et al., 2006; Rugg-Gunn et al., 2010). These so called bivalent domains are thought to poise lineage-specific genes for expression in ES cells to respond quickly to differentiation cues. Under self-renewing conditions they are held in check by the opposing chromatin mark. Another study that compared the epigenetic pattern in mouse ES cells cultured in LIF and serum with ES cells maintained in LIF and 2i, showed that bivalent domains are more frequent in LIF and serum culture conditions (Marks et al., 2012). Instead of repressing chromatin marks, in 2i lineage specific genes were held in check by polymerase pausing (Marks et al., 2012).

Chromatin modifications have been directly linked to the pluripotency network. Oct4 positively regulates the H3K9me2 and H3K9me3 demethylase genes Jmjd1a and Jmjd2c (Loh et al., 2007). Oct4 also binds together with Stat3 to the promoter of Eed, inducing its expression (Ura et al., 2008). Eed is part of the polycomb repressive complex 2 (PRC2), which targets developmental genes for silencing by H3K27me3 (Boyer et al., 2005). While polycomb is required for differentiation it is dispensable for ES cell self-renewal (Chamberlain et al., 2008). Stat3 also directly interacts with the chromatin remodelling ATPase Brg1 (Ni and Bremner, 2007).

1.4.3 The transcription factor network of pluripotency

The undifferentiated state of ES cells is governed by a regulatory network of transcription factors (Boiani and Schöler, 2005). Oct4, Sox2 and Nanog are thought to be at the center of this network (Boyer et al., 2005; Loh et al., 2006; Zhou et al., 2007). Chromatin immunoprecipitation coupled with DNA microarrays showed that these three transcript factors co-occupy a substantial proportion of their target genes, including themselves (Boyer et al., 2005). Many of the targets encode developmentally important transcription factors. Notably, Oct4 and Sox2 are also two of the four reprogramming factors (Takahashi and Yamanaka, 2006) and are thus critical for the establishment of the pluripotent state.

Oct4 (also known as Pou5f1) is a POU domain-containing transcription factor essential to ES cells and early embryonic development (Nichols et al., 1998). Oct4 knock-out

in ES cells induces differentiation towards the trophectodermal lineage, while 2-fold overexpression leads to differentiation into primitive endoderm-like cells (Nichols et al., 1998; Niwa et al., 2000). Recently it has been shown that a low expression level of Oct4 stabilises pluripotency while a high expression level enables effective differentiation (Karwacki-Neisius et al., 2013; Radzisheuskaya et al., 2013).

Oct4 interacts with many components of the pluripotency network (Berg et al., 2010; Pardo et al., 2010). However, its main interaction partner is Sox2 and together they co-bind many target genes (Boyer et al., 2005). A joint Oct4-Sox2 binding motif is enriched within the promoter sequences of shared target genes (Loh et al., 2006). Deletion of Sox2 causes ES cells to differentiate along the trophectodermal lineage, similar to Oct4 deletion (Masui et al., 2007). Surprisingly, expression of many Oct4 and Sox2 target genes is not affected by Sox2 deletion, but Sox2 is required for the maintenance of Oct4 expression. Consistent with this, enforced Oct4 expression can rescue ES cells from differentiation induced by Sox2 deletion (Masui et al., 2007). Overexpression of Sox2 will induce differentiation of ES cells even in self-renewing conditions (Kopp et al., 2008).

While Oct4 and Sox2 need to be expressed at a defined level and both deletion and overexpression will cause differentiation of ES cells, overexpression of Nanog can maintain mouse ES cells in a pluripotent state even in the absence of LIF (Chambers et al., 2003; Mitsui et al., 2003). Nanog-null ES cells on the other hand differentiate into parietal endoderm-like cells (Mitsui et al., 2003).

In recent years this core regulatory network has been extended with additional transcription factors. Rex1 (also known as Zfp42) is a transcription factor expressed in ground state pluripotency, but is rapidly downregulated once the cells become primed to differentiate. Therefore, it is often used to monitor the pluripotent state of ES cells (Marks et al., 2012; Martello et al., 2012). The Krueppel factors Klf2, Klf4 and Klf5 enhance ES cell self-renewal (Li et al., 2005; Hall et al., 2009; Jiang et al., 2008). Notably, Klf4 is one of the Yamanaka reprogramming factors (Takahashi and Yamanaka, 2006) and forced Klf4 expression can revert EpiSC to ES cells (Guo et al., 2009). The transcription factor Esrrb is required to mediate the self-renewal response to Gsk3 inhibition and its forced expression can mimic Gsk3 inhibition (Martello et al., 2012). Prdm14 regulates the DNA methylation pattern in ES cells by repressing the expression of de novo DNA methyltransferases (Leitch et al., 2013; Chia et al., 2010; Yamaji et al., 2013;

Ficz et al., 2013). Many of these transcription factors have highly overlapping target genes (Kim et al., 2008; Chen et al., 2008; Kidder et al., 2008; Marson et al., 2008; Ang et al., 2011; Whyte et al., 2013). In addition, miRNAs play a major role in this regulatory network by fine-tuning the expression levels of several pluripotency transcription factors (Mallanna and Rizzino, 2010; Marson et al., 2008; Tay et al., 2008).

1.4.4 Pathways involved in stem cell maintenance and differentiation

Signalling pathways communicate extrinsic and intrinsic signals to the gene regulatory machinery enabling the cells to respond to outer cues. The activation of different pathways leads to the maintenance of the pluripotent state or to the initiation of differentiation. The most important signalling pathways involved in stem cell maintenance are the Jak/Stat pathway, the Wnt-pathway and the PI(3)K pathway (Hirai et al., 2011). All three pathways have been reported to enhance ES cell self-renewal. However, the detailed mechanisms and the overlap between these three pathways are not yet fully understood.

1.4.4.1 Jak/Stat3-pathway

LIF signalling is mediated via heterodimerisation of the low affinity LIF receptor and gp130, a common receptor subunit of the interleukin (IL)-6 cytokine family (Yoshida et al., 1994). Heterodimerisation of the LIF receptor and gp130 leads to phosphorylation and activation of associated JAK tyrosine kinases (Figure 1.6, Stahl et al., 1994). This in turn causes phosphorylation of tyrosine residues of gp130, leading to the activation of MAP ERK signalling cascade and the phosphorylation of Stat3 (Boeuf et al., 1997; Akira et al., 1994; Burdon et al., 2002; Hirai et al., 2011). Phophorylated Stat3 dimerises, enters the nucleus where it activates the expression of its target genes (Boeuf et al., 1997; Zhong et al., 1994; Burdon et al., 2002; Bourillot et al., 2009).

Mutational analyses of the receptors and the use of a dominant-negative Stat3 construct revealed Stat3 activation as the main mediator of enhanced stem cell self-renewal upon LIF stimulation (Niwa et al., 1998). Further mutational analysis of gp130 showed that gp130-mediated SHP2 and MAP kinase activation is dispensable for the self-renewal of ES cells (Matsuda et al., 1999). Artificial activation of Stat3 using a hormone inducible

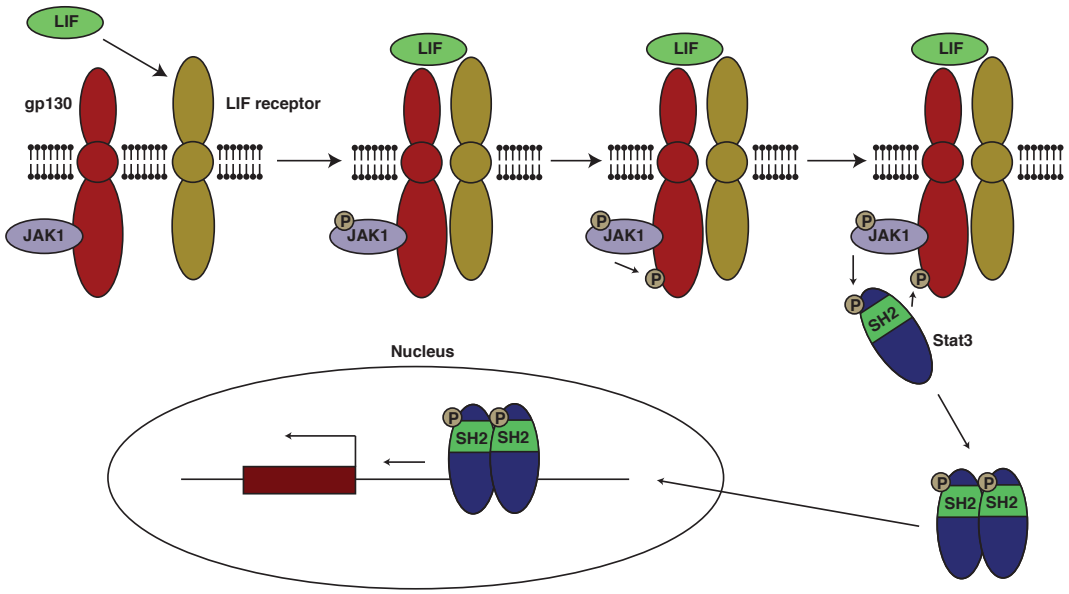


Figure 1.6: The Jak/Stat3 signalling pathway. The LIF receptor and gp130 receptor heterodimerise after LIF binding. This results in the phsophorylation of associated JAL tyrosine kinases which in return leads to phosphorylation of gp130. This causes activation of Stat3 which dimerises and translocates to the nucleus where it activates its target genes. Modified from Hirai et al., 2011.

Stat3 construct (Matsuda et al., 1999) or mutation of tyrosine 118 of gp130 (Burdon et al., 1999) demonstrated that Stat3 activation is sufficient to maintain an undifferentiated state of ES cells.

The main pathway downstream of LIF is the Jak/Stat3 pathway. However, LIF also activates the MAPK and the PI(3)K pathway via JAK (Niwa et al., 2009; Griffiths et al., 2011; Welham et al., 2007). Activation of the MAPK pathway typically drives differentiation. Blocking differentiation by MAPK inhibition is one main mechanism of the 2i system (as described in Section 1.4.1, Ying et al., 2008). The PI(3)K pathway on the other hand is promoting self-renewal (Niwa et al., 2009; Welham et al., 2007).

The JAK/Stat3 pathway plays an important role in both the induction and maintenance of the pluripotent state, and LIF is routinely used to convert primordial germ cells into pluripotent embryonic germ (EG) cells (Matsui et al., 1992; Resnick et al., 1992). Yang et al. showed that EpiSC (Brons et al., 2007; Tesar et al., 2007) can be reprogrammed to ES cells by transient activation of Stat3 even in the presence of FGF and Activin, which instruct and maintain the primed state (Yang et al., 2010; Oosten et al., 2012). Reprogramming using the canonical reprogramming factors (Oct4, Klf4, Sox2,

c-Myc, Takahashi and Yamanaka, 2006) is greatly enhanced by Stat3 activation even in the presence of antagonistic cues like FGF-Erk signalling (Oosten et al., 2012). On the other hand, inhibition of the LIF/Stat3 pathway abolishes iPS cell generation (Tang et al., 2012).

1.4.4.2 Canonical Wnt-pathway

The canonical Wnt signalling pathway is activated by the binding of Wnt ligands to the Frizzled and low-density lipoprotein receptor-related protein. Its activation can promote self-renewal even in conditions that induce differentiation (Sato et al., 2004). Although Wnt signalling activates Stat3 (Hao et al., 2006), the self-renewal effect of Wnt activation was demonstrated to be LIF/Stat3 independent (Sato et al., 2004). Wnt enhances self-renewal by stabilising β-catenin and protecting it from Gsk3 mediated phosphorylation. Phosporylation of β-catenin leads to its ubiquitylation and subsequent degradation. Wnt signalling further causes β-catenin to translocate to the nucleus where it interacts with the transcription factor Tcf3 (Molenaar et al., 1996). Tcf3 co-localises with pluripotency factors and and holds the expression levels of their target genes in check, thereby balancing between pluripotency and differentiation (Cole et al., 2008). Consequently, Tcf3-null ES cells show increased resistance to differentiation (Pereira et al., 2006; Guo et al., 2011). For some time, the main mechanisms of β-catenin was believed to be the conversion of Tcf3 from a repressor to an activator. However, it has been shown that a truncated β-catenin lacking the transactivation domain is sufficient to promote ES cell self-renewal, indicating that β-catenin stabilises ES cell self-renewal by de-repression of Tcf3 targets rather than activation (Wray et al., 2011).

Gsk3 inhibition, as used in the 2i culture condition, thus activates the canonical Wnt/β-catenin signalling pathway to promote stem cell self-renewal (Sato et al., 2004; Wray et al., 2011). While β-catenin is not necessary for ES cell self -renewal, its absence eliminates the self-renewal response to chiron (Wray et al., 2011). Esrrb is one of the key targets of Tcf3 mediated repression and its expression is essential for the self-renewal response to chiron (Martello et al., 2012).

1.4.4.3 PI(3)K-pathway

Beside the Jak/Stat3 pathway, LIF also induces the PI(3)K pathway which leads to the activation of the AKT serine/threonine kinases (Paling et al., 2004). These stabilise ES cell self-renewal by Gsk3β inhibition (Storm et al., 2009; Storm et al., 2007) and upregulation of pluripotency factor Tbx3 (Niwa et al., 2009). The PI(3)K pathway also feeds back into the Jak/Stat3 pathway (Ohbayashi et al., 2007). Watanabe et al. report that PI(3)K activation is sufficient to maintain pluripotency (Watanabe et al., 2006). As LIF-induced activation of the PI(3)K pathway is, however, not sufficient to maintain $Stat3^{-/-}$ cells self-renewing, this indicates that self-renewal mediated by PI(3)K activation is dependent on Stat3. $Stat3^{-/-}$ cells are strictly dependent on 2i and cannot be maintained with LIF (Martello et al., 2013).

1.5 Aims of the analyses

Since the advent of the RNA-seq technology a wide variety of software packages have been developed for the analysis of this kind of data. The first step of most RNA-seq analysis pipelines is the alignment of the data to a reference genome or transcriptome. The alignment of reads spanning exon-intron boundaries poses an extra challenge to the aligner when aligning against a genome. Splitting up reads to map individual segments across splice junctions increases the risk of spurious alignments. In the first part of my thesis I used RNA-seq data from *H. sapiens* and *M. musculus* to compare the performance of a comprehensive set of spliced aligners (including unpublished methods such as GSTRUCT or an unreleased version of BAGET) on different aspects, from simple statistics as alignment yield to the influence of alignments on subsequent transcript reconstruction. The human data sets included simulated data, enabling accurate sensitivity and precision calculations for read placement, junction discovery and the identification of small indels.

The ultimate goal of RNA-seq analysis is the identification of the set of expressed transcript isoforms and their quantification. This requires that reads are assembled into transcripts. Most methods make use of read alignments to the reference genome. Some *de-novo* methods, however, assemble the reads directly into transcripts without the use of a reference genome. In the second part of my thesis, I used RNA-seq data

from *H. sapiens*, *D. melanogaster* and *C. elegans* to assess the performance of several transcript reconstruction methods, including alignment-based methods as well as *de-novo* methods. The reconstructed transcript isoforms were compared to high-quality annotations (filtered to include only expressed isoforms) to assess the ability of methods to detect exons, genes and their isoforms. As these three organisms differ in transcript diversity, the influence of transcriptome complexity on the transcript assembly can be analysed. Transcript quantification was compared between methods and quantification accuracy was assessed using NanoString data.

In the last part of my thesis, I studied the response of mouse ES cells to LIF signalling using a combination of RNA-seq and ChIP-seq data. LIF-dependent Stat3 activation plays an important role in the maintenance of mouse ES cells in culture. Several Stat3 downstream targets have been identified, but none are essential for LIF/Stat3-mediated ES cell self-renewal. This suggests that Stat3 activates a set of target genes with cross-compensatory functionality. Another possibility is the presence of a master regulator downstream of Stat3 that has yet to have been revealed. Through integrated analysis of RNA-seq and ChIP-seq data, I aimed to identify Stat3 targets involved in ES cell self-renewal. Furthermore, I incorporated public available ChIP-seq data for other pluripotency factors to gain insight in to the mechanisms of Stat3 mediated gene activation.

Chapter 2

Comparing Alignment methods for RNA-seq data

2.1 Introduction

The recent development of RNA sequencing (RNA-seq) has generated the need for software tailored to the requirements of partial transcript sequencing. Over the last few years a large number of aligners specialised for the application of RNA-seq data have been developed as have methods for further downstream analysis, as for example methods for transcript reconstruction and quantification which are discussed in Chapter 3.

2.1.1 Basics of RNA-seq alignment

There are two fundamentally different approaches to RNA-seq analysis. Fragmented transcript reads can be assembled into transcripts using a de-novo transcriptome assembler, such as TransAbyss, Trinity or Oases (Schulz et al., 2012; Grabherr et al., 2011; Robertson et al., 2010). This method is the only available approach when analysing data for species for which the reference genome has not yet been determined. However, for species with a sequenced reference genome, such as *H. sapiens* and most model organisms, the more common approach is to map reads to the reference genome or annotation based transcriptome. The latter approach was commonly used in some of the earliest mammalian RNA-seq studies as it enabled the alignment with a contiguous DNA aligner as Bowtie, BWA or SSAHA2 (Langmead et al., 2009; Li and Durbin, 2009; Ning et al.,

2001). This approach. however, prevented the identification of novel transcribed features which is one of the main advantages of RNA-seq over microarray analysis.

Therefore the advent of RNA-seq resulted in the development of a new generation of spliced alignment methods improving on earlier spliced alignment programs such as BLAT (Kent, 2002). The two main challenges of RNA-seq read alignment are the mapping of reads across exon-intron boundaries and resolving ambiguity if reads match to multiple genomic locations. To improve the mapping across introns a two-step approach can be implemented, where initial read alignments are analysed to discover exon boundaries, which are then used to guide final alignment. Similarly, many methods can be provided with known splice sites from existing annotations. To determine the origin of reads that map to multiple genomic locations, many aligners make use of base call quality scores as well as the density of uniquely mapped read surrounding ambiguously mapped reads.

2.1.2 The benchmark

The performance of 26 RNA-seq alignment protocols was assessed on real and simulated data for human and mouse transcriptomes. To ensure that programs were executed in the intended manner and with appropriate settings, developers were invited to run their aligners and submit results for evaluation as part of the RNA-seq Genome Annotation Assessment Project (RGASP). The set of evaluated aligners includes eight spliced alignment programs (GEM (Marco-Sola et al., 2012), GSNAP (Wu and Nacu, 2010), MapSplice (Wang et al., 2010a), PALMapper (Jean et al., 2010), PASS (Campagna et al., 2009), ReadsMap, STAR (Dobin et al., 2013) and TopHat (Trapnell et al., 2009)) and two alignment pipelines (GSTRUCT and BAGET). GSTRUCT is based on GSNAP, while BAGET is an implementation of the strategy used in the earliest mammalian RNA-seq studies (Sultan et al., 2008; Cloonan et al., 2008; Mortazavi et al., 2008) where reads are mapped with a contiguous DNA aligner to the genome as well as to a library of exon junction sequences based on genome annotation. To compare the spliced alignment programs with a contiguous aligner SMALT was included in this study. SMALT can map reads in a split manner, but cannot determine precise exon-intron boundaries. This study demonstrates how important the choice of alignment software is for the interpre-

Table 2.1: RNA-seq data sets used for the evaluation. Repl.=replicate

Name	ID	Organism	Fragments	Lanes
K562 whole cell repl. 1	LID16627	Human cell line	113588758	3
K562 whole cell repl. 2	LID16628	Human cell line	119053315	3
K562 cytoplasmic fraction repl. 1	LID8465	Human cell line	124826068	3
K562 cytoplasmic fraction repl. 2	LID8466	Human cell line	88445339	3
K562 nuclear fraction repl. 1	LID8556	Human cell line	117113622	3
K562 nuclear fraction repl. 2	LID8557	Human cell line	105769104	3
Mouse brain	mouse	Mouse strain C57B6	57187342	2
Simulation 1	sim1	Human	40000000	n.a
Simulation 2	sim2	Human	40000000	n.a

tation of RNA-seq data and identifies aspects of the spliced alignment problem that still need to be addressed.

2.1.3 Outline

In order to assess the quality of several commonly used RNA-seq aligners, their performance was compared using different metrics. This project was a highly collaborative effort and a manuscript has been accepted for publication by *Nature Methods*[1]. The artificial data was generated by Botond Sipos, European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge UK, and Gregory Grant, Department of Genetics and the Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia USA. TopHat alignments were created by Pär Engström, European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge UK. The post-processing of bam files was performed by Pär Engström and me. Pär Engström calculated alignment yields and alignment accuracy for the simulated data sets. Botond Sipos calculated the fraction of reads falling within annotated features as exons and introns. I calculated positioning and numbers of indels, mismatches and introns and performed the Cufflinks analysis. Supplemental information for this project can be found in Appendix A.

[1]Engström P, Steijger T, Sipos B, Grant G, RGASP Consortium, Goldman N, Hubbard T, Harrow J, Guigo R and Bertone P. **Systematic evaluation of spliced aligners for RNA-seq**, *Nature Methods* (in Press)

Table 2.2: Results on key metrics. Data sets: Mean over K562 samples (K562), K562 whole cell replicate 1 (K562 w), mouse brain (M), simulation 1 (S1) and 2 (S2). Metrics: [a]percentage of sequenced or simulated reads mapped by each protocol; [b]percentage of all simulated bases that were correctly/incorrectly aligned; [c]number of splices in primary alignments divided by the number of sequenced reads; [d]junction discovery accuracy when requiring at least two supporting mappings per junction. All values are given as percentages. Bold indicates the highest or lowest value in each column. The PALMapper protocols were not applied to all data sets, as indicated (n.a.). The lower splice frequencies on mouse data are expected as a result of a more pronounced 3′ bias in this data set (not shown).

	Mapped reads[a]				Correctly mapped bases[b]		Incorrectly mapped bases[b]		Splice frequency[c]				Junction recall (≥2 mappings)		Junction precision (≥2 mappings)[d]	
	K562	**M**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**K562 w**	**M**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
BAGET ann	92.94	95.71	98.58	96.77	90.61	87.49	5.23	4.83	8.38	4.95	9.05	9.17	63.03	61.89	95.56	94.91
GEM ann	**93.87**	**98.33**	**99.9**	**99.4**	96.54	94.33	3.29	4.76	16.23	6.91	15.55	14.62	95.34	90.8	95.6	89.6
GEM cons	93.85	98.31	99.88	99.36	96.49	94.25	3.3	4.8	16.01	6.7	15.35	14.26	84.08	77.39	96.56	91.57
GEM cons ann	93.86	**98.33**	**99.9**	99.39	96.53	94.32	3.29	4.77	16.07	6.81	15.5	14.53	90.14	86.22	96.15	91.51
GSNAP	93.8	96.71	99.24	97.95	96.84	94.55	1.75	2.01	16.55	6.19	13.66	13.79	95.61	95.34	95.58	93.58
GSNAP ann	93.82	96.72	99.25	97.97	97.52	95.27	1.35	1.7	23.21	8.2	18.01	18.78	**98.12**	**97.9**	93.28	91.04
GSTRUCT	**93.87**	97.44	99.26	98.11	96.95	94.85	1.95	2.34	21.35	8.63	17.87	18.65	96.79	96.42	96.95	95.16
GSTRUCT ann	**93.87**	97.43	99.26	98.11	**97.59**	**95.43**	1.31	1.76	22.37	8.77	18.12	18.89	97.24	97.02	97.24	95.51
MapSplice	90.02	93.95	98.61	94.61	96.83	91.46	1.35	1.62	18.65	7.32	16.98	15.09	95.94	90.35	**98.26**	**95.86**
MapSplice ann	90.01	93.98	98.68	94.79	96.95	91.67	1.34	1.64	18.51	7.41	17.2	15.57	97	93.54	94.54	90.78
PALMapper	91.15	n.a.	98.35	96.78	95.2	93.03	3.05	3.74	21.62	n.a.	17.09	17.79	94.89	93.14	61.49	58.58
PALMapper ann	n.a.	n.a.	98.42	96.99	94.96	92.99	3.37	4	n.a.	n.a.	17.82	19.1	96.27	95.18	58.66	52.07
PALMapper cons	52.14	n.a.	80.81	84.77	78.54	81.91	1.7	2.86	3.82	n.a.	8.31	8.88	87.97	86.59	95.74	91.85
PALMapper cons ann	n.a.	n.a.	97.74	94.32	94.85	90.92	2.78	3.4	n.a.	n.a.	15.44	15.94	92.65	89.47	78.79	71.63
PASS	89.86	92.78	96.97	90.15	90.83	80.52	3.46	3.38	11.2	5.9	12.48	10.72	91.18	85.1	86.33	76.3
PASS cons	87.62	90.29	95.99	87.48	90.47	79.28	3.01	2.8	11.02	5.77	12.42	10.49	91.1	84.94	89.41	80.37
ReadsMap	77.18	72.82	88	86.49	77.15	72.65	9.87	13.83	22.84	**10.57**	**22.94**	**20.24**	94.63	89.53	20.68	20.25
SMALT	91.45	92.25	96.73	96.34	91.62	90.13	1.92	2.1	2.8	1.51	3.32	3.15	35.34	34.88	30.69	28.43
STAR 1-pass	91.52	89.23	98.77	96.23	96.2	92.21	1.7	1.96	14.02	5.55	12.07	10.39	93.01	87.24	97.68	95.79
STAR 1-pass ann	91.69	89.26	98.85	96.71	97.19	93.73	1.27	1.6	22.64	7.1	17.32	16.49	96	93.23	91.72	89.8
STAR 2-pass	91.68	89.31	98.86	96.77	97.26	93.85	**1.23**	1.58	24.24	8.47	17.55	16.92	96.53	92.38	95.66	92.59
STAR 2-pass ann	91.67	89.34	98.85	96.77	97.26	93.9	1.25	1.59	**24.33**	8.67	17.74	17.25	97.71	95.02	91.66	88.81
TopHat v1	84.22	84.92	95.44	86.09	92.79	83.82	2.44	2.27	15.12	6.58	15.31	14.21	91.01	83.85	94.97	92.33
TopHat v1 ann	84.25	84.96	95.58	86.53	92.94	84.26	2.45	2.27	15.15	6.65	15.48	14.7	93.59	88.99	94.62	92.15
TopHat v2	83.47	85.1	93.96	77.93	91.96	76.18	1.85	1.74	17.23	7.32	16.41	13.31	91.78	86.23	95.04	93.36
TopHat v2 ann	84.52	85.41	93.84	79.64	93.16	78.1	1.46	**1.55**	22.11	8.33	17.76	15.54	95.76	92.61	88.4	86.87

2.2 Results

The data sets used for the evaluation were Illumina 76 nt paired-end RNA-seq data from the human leukemia cell line K562 ($1.3x10^9$ reads), mouse brain ($1.1x10^8$ reads) and two simulated human transcriptome data sets ($8.0x10^7$ reads). For more details see Table 2.1. Nine development teams participated in the evaluation. They run their own software and provided their alignment results for benchmarking. Most of them submitted alignments using several protocols, corresponding to different ways of running their software, that is different parameter choices. Additionally, two versions of the widely used RNA-seq aligner TopHat were included in the analysis.

A comprehensive set of metrics, aiming to assess a wide range of attributes was calculated. These metrics range from very basic characteristics such as the percentage of aligned reads, to more advanced ones such as splice junction detection. Some of the metrics used in this study have been described before as for example basewise accuracy (Dobin et al., 2013; Grant et al., 2011; Langmead and Salzberg, 2012). Results on key metrics are summarised in Table 2.2.

2.2.1 Alignment yield

One of the most basic attribute of an alignment is the percentage of mapped reads. Methods differed substantially in their alignment yield. The proportion of read pairs with reported alignments ranged from 42.4% to 96.7% for the K562 data. For detailed numbers refer to Table A.1. When considering whether both mates of a read pair were mapped and whether they were mapped uniquely or ambiguously, protocols showed even greater variation (Figure 2.1).

TopHat, ReadsMap, PASS and the conservative PALMapper protocol showed the highest fraction of read pairs with only one mate mapped. PALMapper output contained a high fraction of read pairs where one or both reads of a pair were ambiguously mapped. GEM also reported a high proportion of ambiguously mapped reads, however, typically both mates were mapped ambiguously (37% of sequenced reads per data set on average). The conservative PALMapper protocol showed reduced mapping ambiguity compare to the other PALMapper protocols, but still higher levels of ambiguity compare to most other methods. These trends were consistent across all data sets (Figure 2.1). To

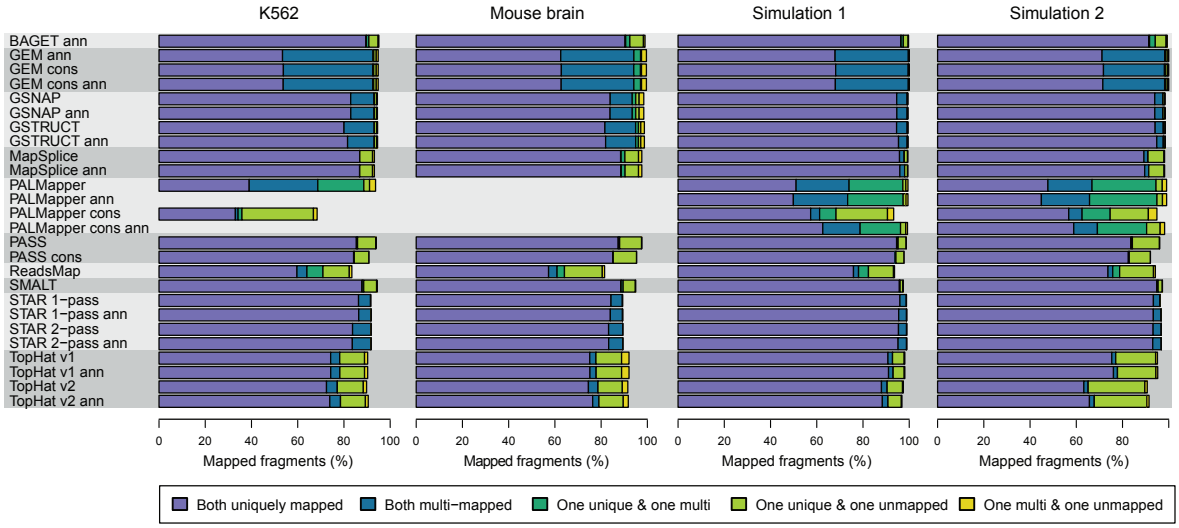


Figure 2.1: Alignment yield. Percentage of sequenced or simulated read pairs (fragments) mapped by each protocol, for the four data sets used in this study. Read pairs are classified by the number of alignments reported per read. Read pairs with both mates unmapped were excluded. Protocols are groups by the underlying alignment program (grey shading). Protocol names contain the suffix "ann" if annotation was used. The suffix "cons" distinguishes the more conservative protocols from others based on the same aligner. The K562 data set comprises six samples, and the metrics presented here were averaged over them. The PALMapper protocols were not applied to all data sets (as indicated), and some K562 samples were not processed by PALMapper and ReadsMap. Exact percentages for each data set are also listed in Table A.1.

avoid introducing bias at further evaluation stages due to differences in the number of alignments per read, we instructed developer teams to assign a preferred (primary) alignment for each read mapped in their program output. Unless further specified the following analyses are focusing on primary read alignments only.

2.2.2 Mismatches and basewise accuracy

Another important attribute of aligners is the number of mismatches allowed per read alignment. The methods GSNAP, GSTRUCT, MapSplice, PASS, SMALT and STAR reported more primary alignments devoid of mismatches compared to the other methods (Figure 2.2a). This can partly be explained by their tendency to trim the end of reads and to output partial alignments (Figure 2.2b). Especially, PASS and SMALT performed extensive read truncation, suggesting that these programs often report alignments shorter than the optimal. While most methods allowed more mismatches for reads with lower quality scores, MapSplice, PASS and TopHat 2 displayed a low tolerance for mismatches independent of quality scores (Figure 2.3). Consequently, they were not able to map a large proportion of reads with low base call quality scores.

The K562 data set differs from the other data sets in respect to the degree of basewise similarity to their respective reference genome. Due to polymorphisms and the accumulation of mutations the K562 genome is more distant to the human reference assembly, which itself is the consensus sequence based on samples from different individuals (International Human Genome Sequencing Consortium, 2004). On the other hand, the mouse RNA samples were obtained from the mouse strain C57BL/6NJ, which genome has high similarity to the mouse reference assembly (Keane et al., 2011). This explains why high-quality reads from the mouse data set were mapped at a greater rate and with fewer mismatches compared to reads from the K562 data set (Figure 2.2). However, the overall trends in the aligner's mismatch and truncation frequencies were consistent across all data sets. This suggests that mapping properties are largely dependent on software algorithms even when the genome and transcriptome are virtually identical.

On the simulated data sets GSNAP, GSTRUCT, MapSplice and STAR outperformed the other methods for basewise accuracy (see Table 2.2 and Tables A.2- A.4). The error rate was significantly lower for uniquely mapped reads compared to primary alignments

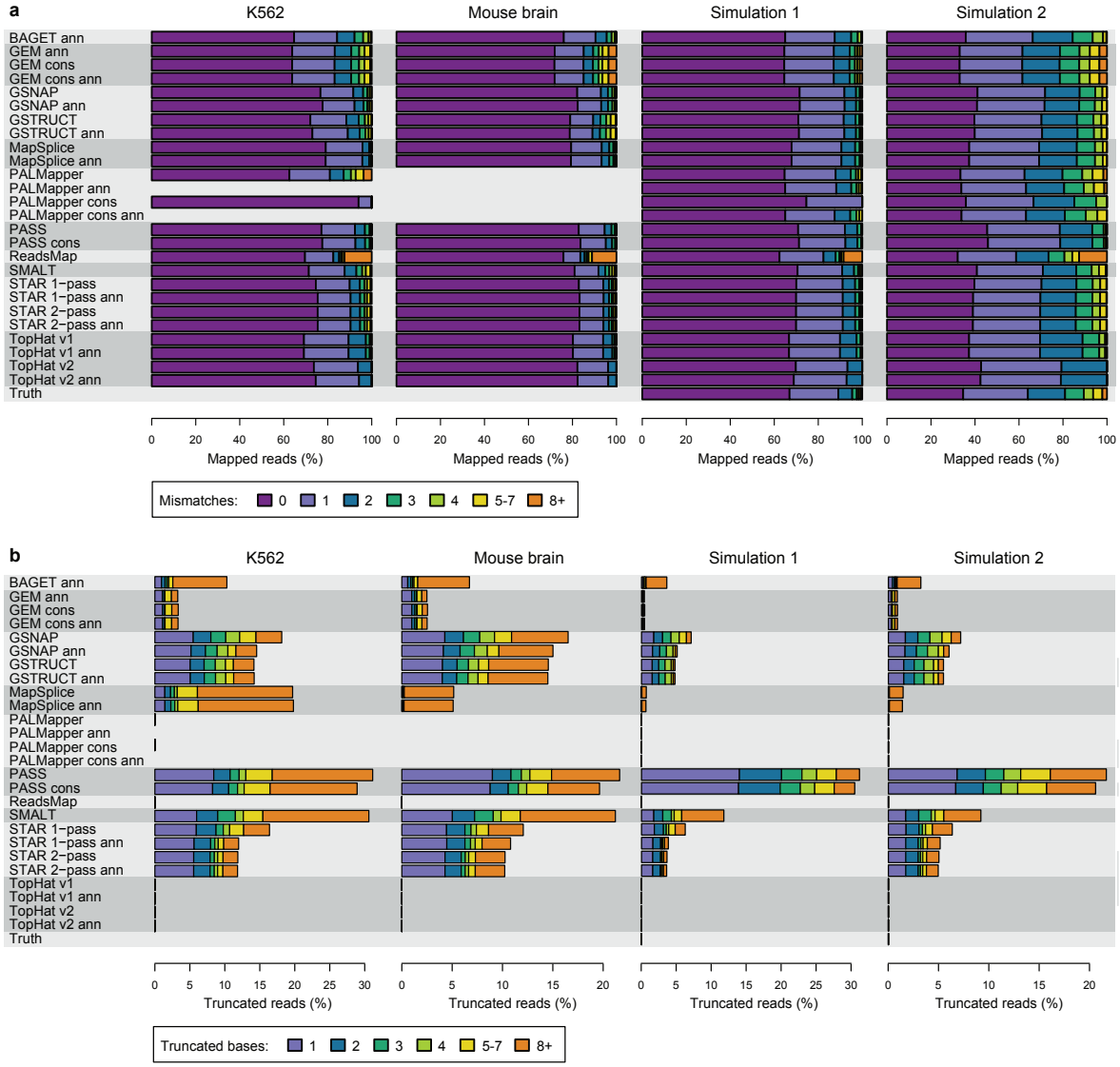


Figure 2.2: Mismatches and read trimming. (a) Percentage of sequenced reads mapped with the indicated number of mismatches. (b) Percentage of sequenced reads that were trimmed at either or both ends. Bar colours indicate the number of bases removed.

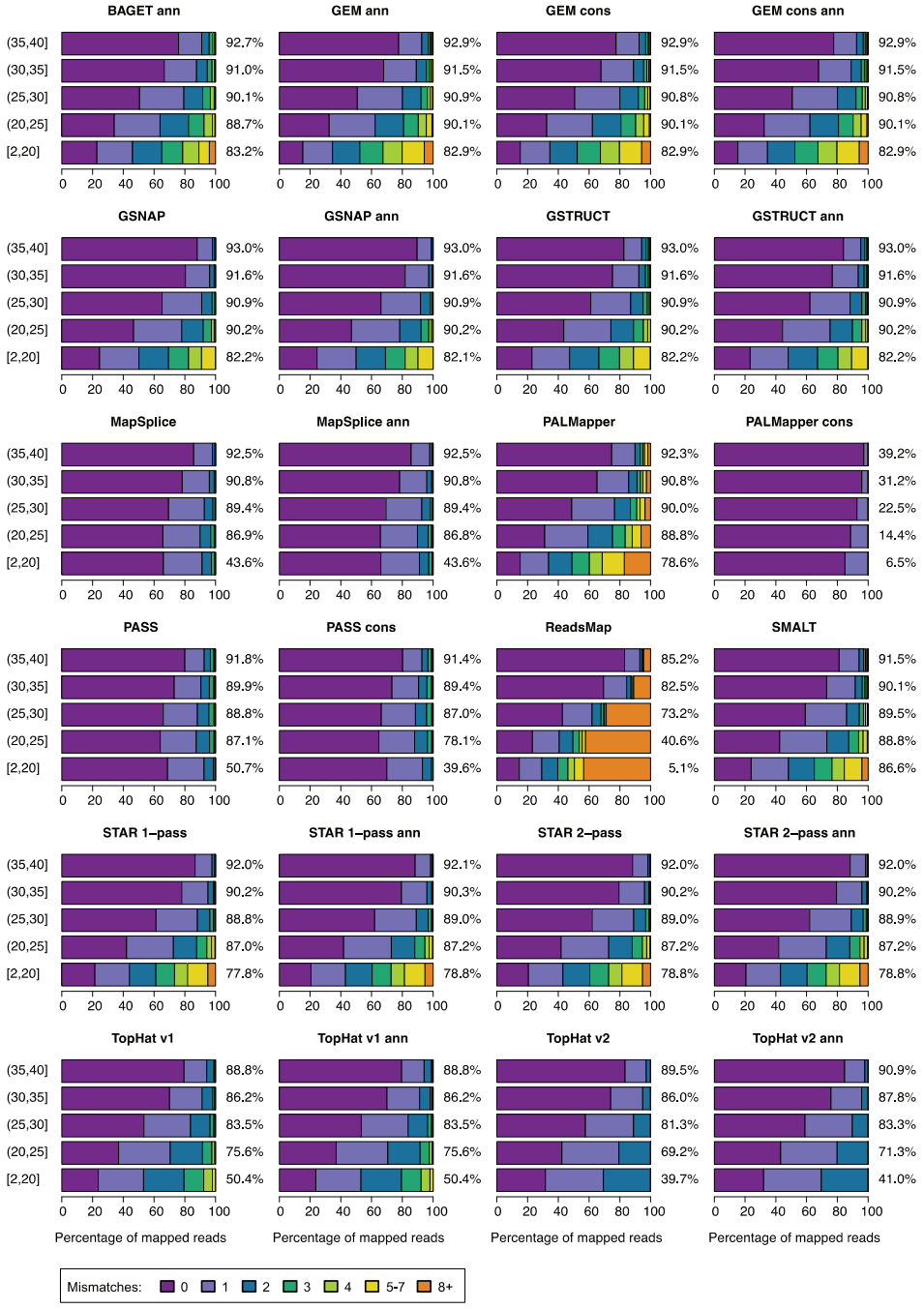


Figure 2.3: Mismatch frequencies stratified by base caller quality scores. Results for K562 whole cell replicate 1 are shown. Reads were divided into five categories by mean quality score. Quality scores ranges from 2 to 40, with lower scores corresponding to less confident base calls. Bars show distribution of mismatches per alignment, demonstrating that most methods tend to align low-quality reads with more mismatches. Percentages of aligned reads are tabulated for each protocol and quality score category, showing that protocols differ in the extent to which alignability depends on quality scores.

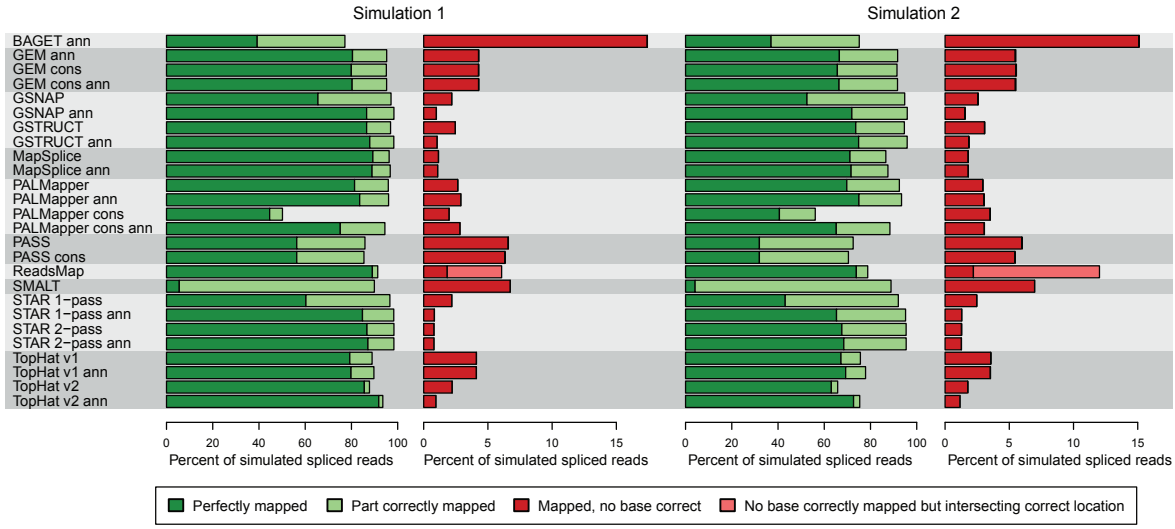


Figure 2.4: Read placement accuracy for simulated spliced reads. Reads with all bases positioned correctly (dark green) are distinguishes from those with a subsequence (1-75 bases) accurately placed (light green). Red bars indicate the percentage of reads that are mapped, but have no base correctly placed. A subset of these mappings overlap the correct alignment (light red); this may occur in repetitive regions or indicate an error in the alignment program.

of multi-mapped reads. Notably, GEM and PALMapper were able to identify the correct primary alignment for a high fraction of their ambiguous mappings.

The greatest differences between methods was observed for the correct placement of spliced reads. For the first simulated data set, GSNAP, GSTRUCT, MapSplice and STAR mapped 96.3-98.4% of spliced reads to the correct genomic location and only misaligned 0.9-2.9% of spliced reads (Figure 2.4 and Table A.3). Thus, they nearly identify the correct locus for all spliced reads. However, for a significant part of the spliced reads, they only map part of the read (indicated with light green bars in Figure 2.4). ReadsMap and the annotation-based TopHat 2 protocol identified the correct location for a slightly lower fraction of spliced reads, however, due to the lack of trimming, overall they identified a higher fraction of perfectly mapped spliced reads compared to the other methods. ReadsMap placed an exceptionally high proportion of bases to the wrong genomic location. Further investigation indicated, that due to a programmatic error, reads got placed a few bases from their correct locations, as indicated by light red bars in Figure 2.4.

2.2.3 Coverage of annotated genes

For the real data it is not possible to calculate exact recall and precision metrics. However, one can analyse the read placement in relation to annotated gene structures from the Ensembl database (Flicek et al., 2013). Both the genome annotation for human and mouse are of high quality and show a high level of completeness. Therefore, the majority of RNA-seq reads would be expected to originate from known exons.

Figure 2.5 shows the results exemplarily for K562 whole cell replicate 1 and simulated data set 1 as the trends were similar for the remaining data sets. More importantly, trends were very similar between real and simulated data indicating that the simulation results reflect alignment performance in real RNA-seq experiments. GSNAP and GSTRUCT mapped the highest number of reads to annotated exons. For the simulated data set the number of reads mapped to exons was close to the true number (Figure 2.5a).

Relative to the frequency of exonic alignments, BAGET and SMALT mapped a high proportion of reads to intronic sequence (Figure 2.5b). This indicates that BAGET and SMALT give priority to unspliced read alignments. ReadsMap and also to some extend annotation based TopHat 2 showed very little mapping to introns relative to the frequency of exonic alignments. The annotation-based TopHat 2 protocol aligns reads to the known transcriptome first. Accordingly, intronic mappings might be under represented. The lower number of inronic alignments for ReadsMap can be explained by ReadsMap avoiding alignments to genomic repeat elements, which are prevalent in introns (data not shown).

A trend that was observable for all methods was the dispersion of reads across too many genes. For example, for simulation 1 the expression of 16554 Ensembl genes was simulated. However, all protocols report primary alignments for more than 17800 genes. Figure 2.5c shows that this effect was largely due to the placement of reads at pseudogenes. SMALT, BAGET and GEM displayed the highest number of read alignments at pseudogenes.

2.2.4 Indel frequency and accuracy

Methods also differed substantially in the detection of indels (Figure 2.6). GEM and PALMapper alignments included more indels than any other method, however, differed

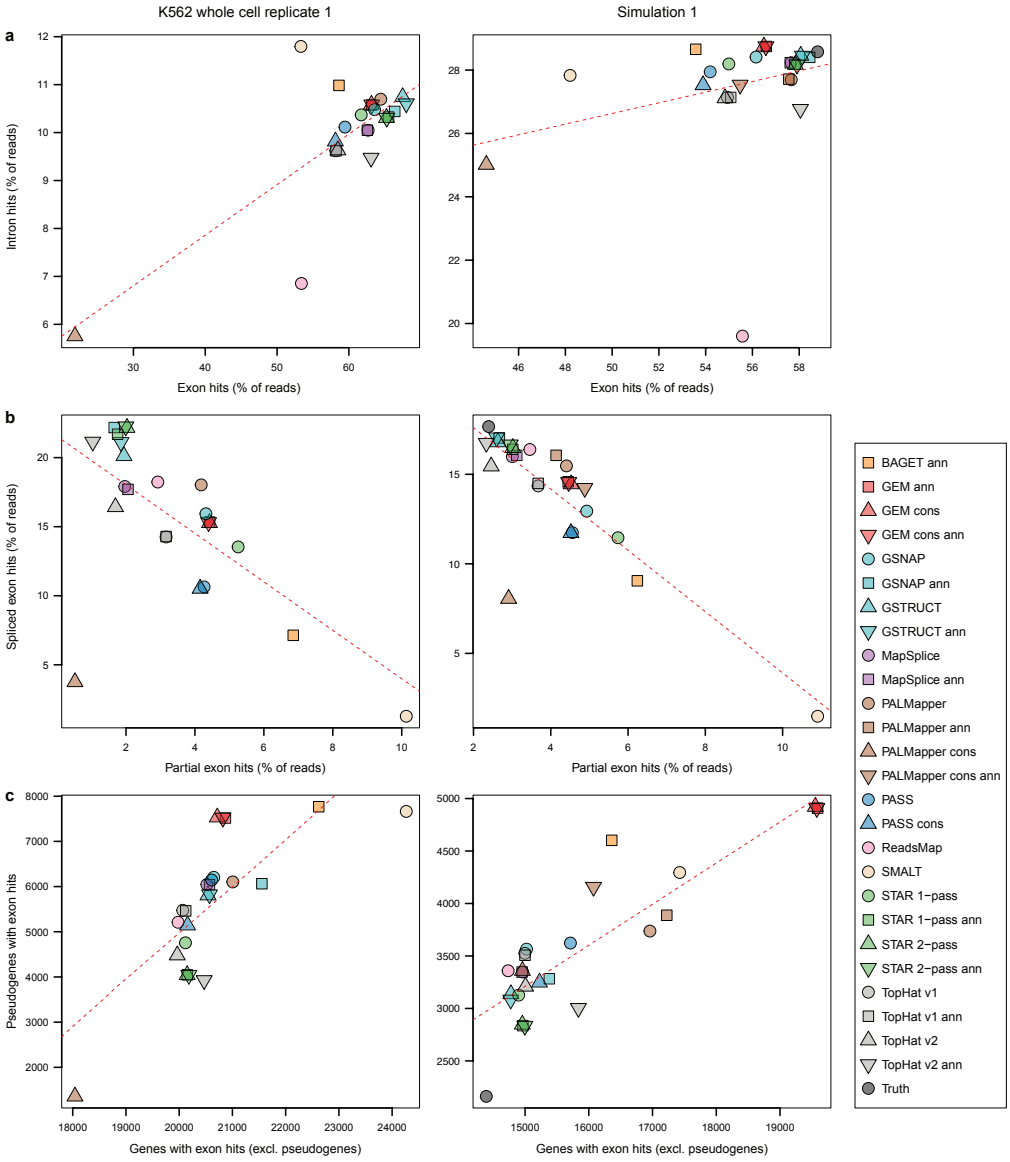


Figure 2.5: Coverage of annotated genes for K562 whole cell and simulation 1. Scatter plots show a range of metric reflecting coverage of Ensembl genes by RNA-seq read alignments, for K562 whole cell replicate 1 (left) and simulated data set 1 (right). (a) Percentage of sequenced and simulated reads for which all mapped bases fall within exon sequence versus those with all mapped bases confined to intron sequence. (b) Percentage of reads for which mappings partially overlap exons (that is alignments where a subset of the genomic positions are annotated as exonic) versus those aligned in a spliced manner with all mapped bases in exon sequence. Note the negative correlation, suggesting that partial exon hits often result from failure to identify splice junctions. (c) Number of genes (including non-coding genes) with fully exonic mappings versus number of pseudogenes with such mappings. For simulated data, "Truth" corresponds to the results expected for a perfect aligner.

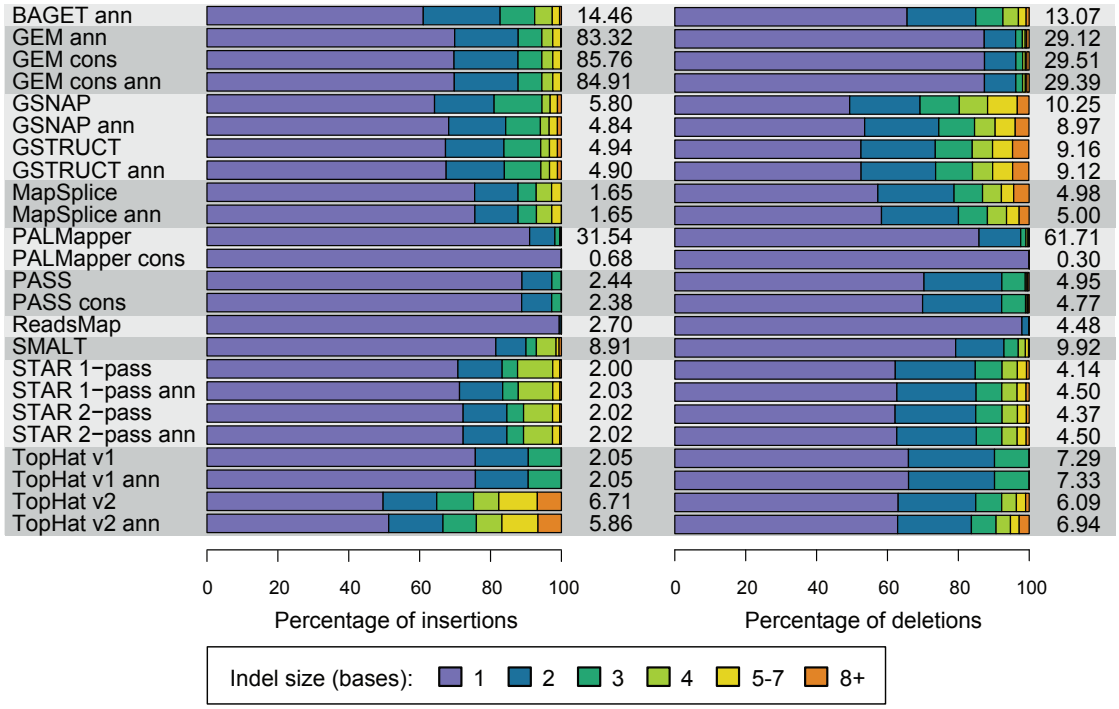


Figure 2.6: Indel frequency. Bars show size distribution of insertions (a) and deletions (b) for the four data sets. Indel frequencies are tabulated (number of indels per thousand sequenced reads).

in their preference for insertions or deletions. While GEM mostly reported insertions, PALMapper reported mainly deletions. All methods reported mostly short indels, however, TopHat 2 reported a few long insertions and GSNAP and GSTRUCT a few long deletions. In contrast, PASS, ReadsMap and TopHat 1 reported mostly short indels and the conservative PALMapper protocol allowed single-nucleotide indels only.

These trends were also observed when analysing the simulated data sets. GEM and PALMapper reported a high number of indels including many false positives leading to an indel precision of $< 37\%$ for all protocols except for the conservative PALMapper protocol for Simulation 1 data set. GSNAP and GSTRUCT showed high sensitivity in the detection of deletions largely independent on the size of the deletion, detecting $> 69\%$ indels for each length interval (Figure 2.7). TopHat 2 was most sensitive for long insertions, reporting $> 87\%$ of insertions ≥ 5 bp correctly for Simulation 1. However, the high sensitivity to detect long indels resulted in a high false discovery rate for GSNAP, GSTRUCT and TopHat 2. MapSplice had a lower recall, however, achieved a better balance between recall and precision for long deletions compared to GSNAP

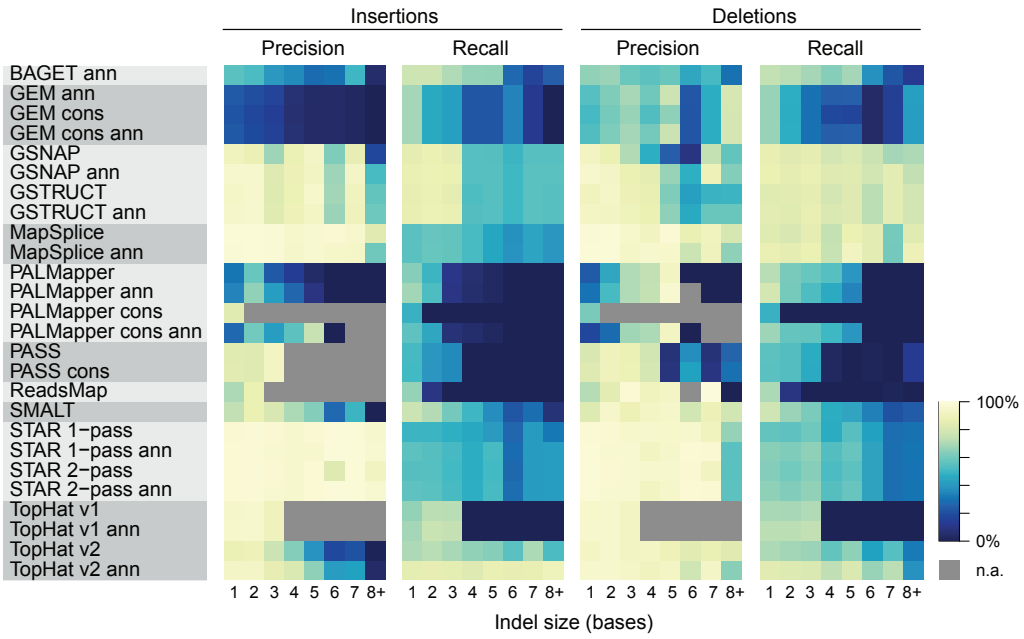


Figure 2.7: Indel accuracy. Precision and recall, stratified by the indel size, for human simulated data set 1. Precision is not applicable (n.a.) if no indels were called.

(Figure 2.7). Balance can be quantified using the F-score, which is the harmonic mean between precision and recall.

$$F - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

For deletions $\geq$ 5 bp in Simulation 1 MapSplice reached an F-score of 87% while GSNAP reached 36% (protocols without annotation input).

2.2.5 Positioning of mismatches and gaps in reads

To assess whether the algorithms showed any bias in positioning certain operations within the read sequence, the spatial distributions of mismatches, indels and introns over the reads sequences were determined. The distributions for these operations are illustrated in Figure 2.8.

All methods except MapSplice, PASS and ReadsMap consistently reported an increasing frequency of mismatches along the reads, reflecting the decrease in base-call quality score along the reads (Figure A.2). As discussed in Section 2.2.2, MapSplice and

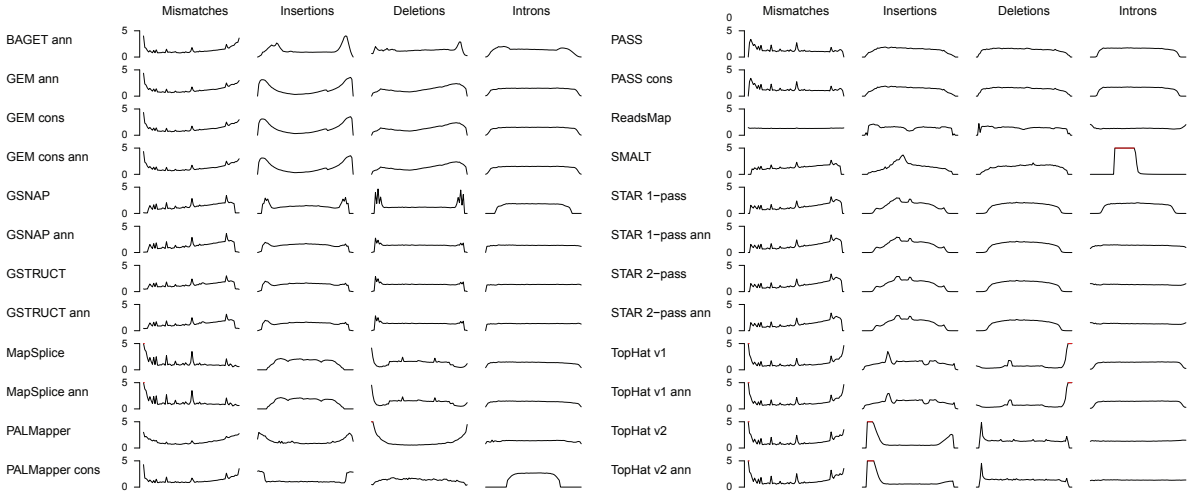


Figure 2.8: Positional distribution of mismatches and gaps over read sequences. Curves show the distribution (percentage) of the indicated operations along the 76 nt read sequences, computed over the primary alignments for K562 whole cell replicate 1. Red lines indicate positions exceeds 5%.

PASS did not report alignments with many mismatches independent of the read quality scores and because of that aligned only a small proportion of reads with low average base quality scores. This might explain why these two methods did not show an increase in mismatch frequency along the reads. ReadsMap however, did allow more mismatches for reads of lower mean base quality and still displayed an even distribution of mismatch placements. This might be due to the reads that were placed a few bases from their correct location (as discussed in Section 2.2.2) which will cause mismatches across the whole read.

BAGET, GEM, MapSplice, PALMapper and TopHat placed many mismatches at read termini (Figure 2.8). Methods not showing this behaviour typically performed a higher degree of read trimming of a few bases (Figure 2.2). For indels two contrary behaviours were observed. Some methods placed indels preferentially near ends of reads, such as PALMapper and TopHat, while others, such as MapSplice and STAR, tended to place them internally. GSTRUCT produced the most uniform distribution of indel frequency over the K562 data (coefficient of variation, $CV = 0.32$) and TopHat the most variable ($CV = 1.5$ and 1.1 for TopHat 1 and 2, respectively). Splice junctions were positioned more evenly by all methods. Annotation based methods and methods implementing a two-step approach placed introns across the whole read sequence. The

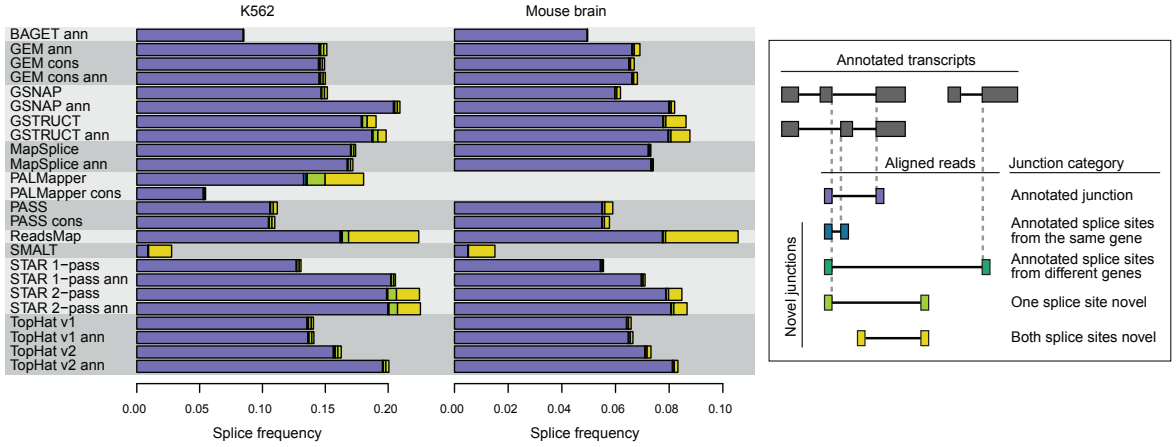


Figure 2.9: Classification of reported splices. Splice frequencies for the K562 (left) and mouse brain (right) data sets. Frequencies were computed as the total number of splices (as opposed to unique junctions) in primary alignments divided by the number of sequenced reads. Splices were classified by overlap with junctions annotated in the Ensembl database (see pictogram).

remaining methods called less junction sites near read termini, especially visible for GSNAP without annotation, the conservative PALMapper, PASS and STAR 1-pass in Figure 2.8.

2.2.6 Spliced alignment

In this section, detection of *splices* in individual reads are distinguished from unique *splice junctions* on the genomic sequence. Splices that overlap known splice junctions are called *supporting* splices.

In general, GSNAP, GSTRUCT, ReadsMap, STAR 2-pass and the annotation based TopHat 2 protocol reported more splices than other aligners, exemplarily shown for the K562 and mouse data sets in Figure 2.9. GSNAP, STAR 1-pass and TopHat 2 reported significantly fewer spliced mappings if no annotation data was used to guide spliced read alignment. In contrast to that, GEM, GSTRUCT, MapSplice, STAR 2-pass and TopHat 1, reported similar number of splices whether they were run with annotation or not. SMALT, BAGET, PASS and the conservative PALMapper protocols inferred the fewest splice sites from the data. ReadsMap and PALMapper, and to a lesser extend SMALT, GSTRUCT and STAR 2-pass reported a large number of splice sites not corresponding to known introns (indicated by yellow bars in Figure 2.9). Figure 2.11a shows that

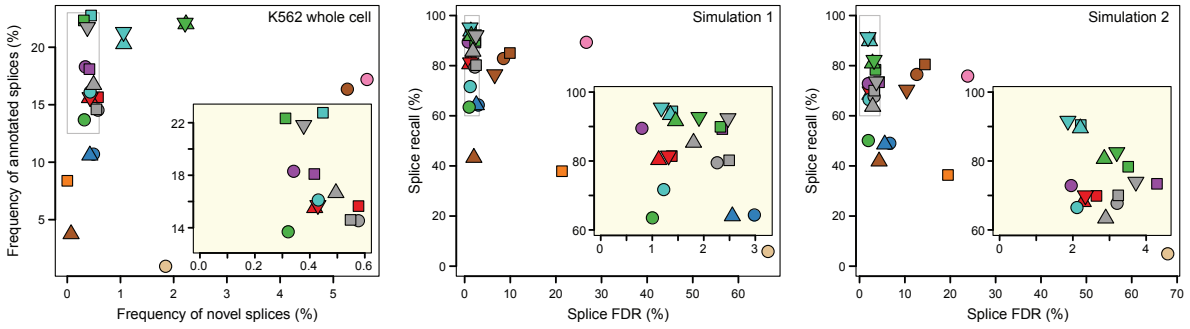


Figure 2.10: Accuracy of splices. For real data (K562 whole cell RNA replicate 1, left), splices were classified as annotated or novel by comparison to transcripts annotated in the Ensembl database, and frequencies computed as the number of reported splices divided by the number of sequenced reads. For simulated data (middle, right) splice recall and false discovery rate (FDR) is presented. Insets show details of the dense upper-left areas (grey rectangles). Note that simulation 2 was designed to be more challenging (see Methods).

novel splice junctions (all junction categories except annotated junctions) were typically supported by fewer alignments in contrast to known splice junctions, suggesting that these might be spurious alignments (see also Table A.5).

Highest accuracy in splice detection for the simulated data was reached by protocols based on GSNAP, GSTRUCT, MapSplice and STAR (Figure 2.10). ReadsMap, PALMapper and SMALT predicted a large fraction of novel splices. The analysis of simulated data confirmed that many of those novel predicted splices were false positives (Figure 2.10), but also all other methods called a considerable fraction of false splice junctions. These, however, could be greatly reduced when junctions were filtered by the number of supporting splices (Figure 2.11). At a threshold of two supporting splices, GSTRUCT outperformed most other methods on both simulated data sets, when assessed by numbers of true and false junction calls (Figure 2.11b, Table 2.2 and Table A.5). MapSplice displayed similar performance on the first simulated data set, but only if used without annotation.

The simulated data sets were based on the Ensembl annotation that was provided to the aligners. To make the data set more realistic, junctions from other gene catalogues were included additionally as well as junctions created by simulating alternate isoforms of known genes. This mimics a realistic scenario where a subset of known transcripts are expressed in the assayed sample and knowledge of the transcriptome is incomplete. As expected, most known junctions were recovered by protocols that were provided

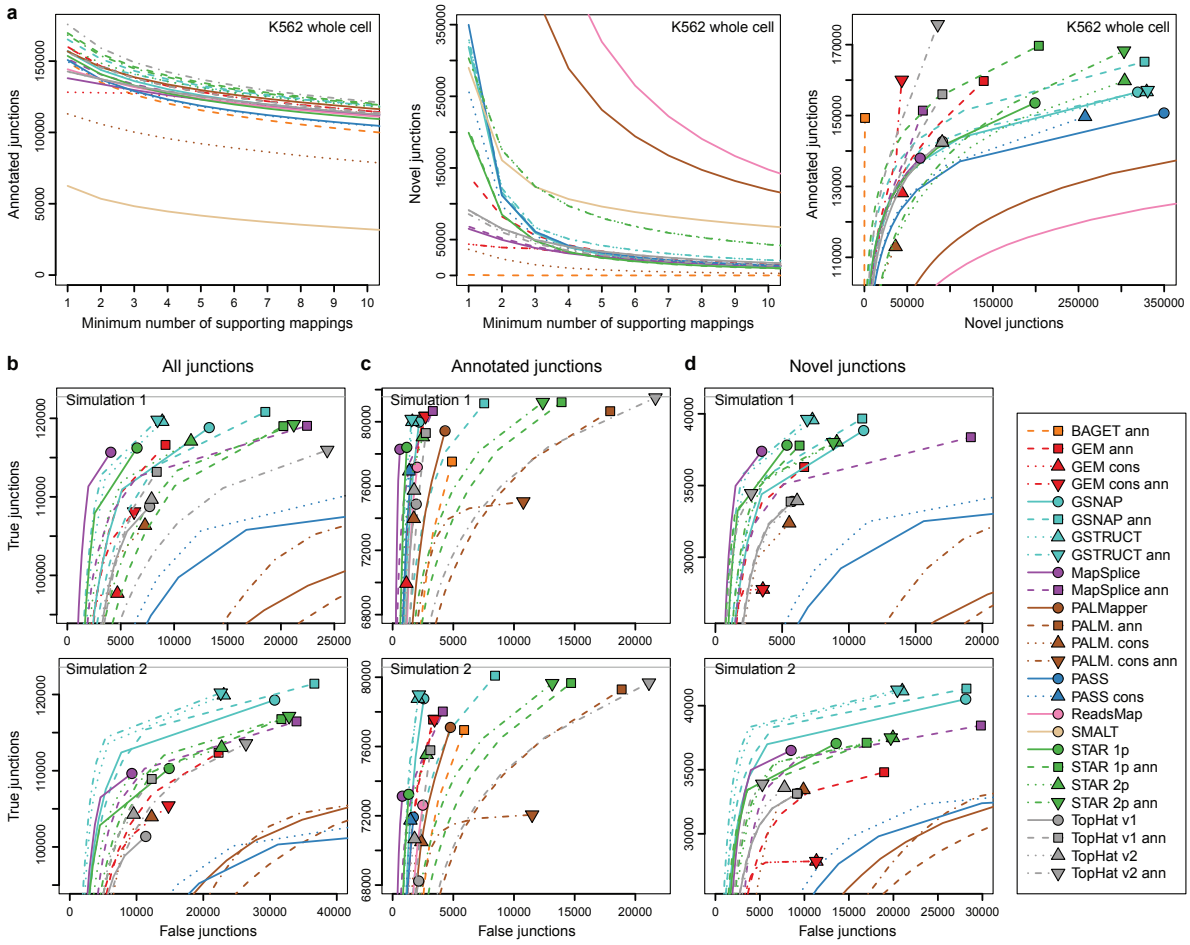


Figure 2.11: Accuracy of splices stratified by read support. (a) Number of annotated and novel junctions reported at different thresholds for the number of supporting mappings. In the rightmost plot, filled symbols depict the number of junctions with at least one supporting mapping, and lines demonstrate the result of thresholding. (b) Junction discovery accuracy for simulated data set 1 (top) and 2 (bottom). Counts of true and false junctions were computed at increasing thresholds for the number of supporting mappings, and results depicted as in panel b to obtain receiver operator characteristic-like curves. Grey vertical lines indicate the number of junctions supported by true simulated alignments. (c) Accuracy for the subset of junctions contained in the Ensembl annotation (used by protocols with the suffix ann). (d) Accuracy for junctions absent from the Ensembl annotation. See also Table A.5.

with annotation. However, most of them also reported many annotated junctions that were not expressed in the simulated transcriptome (Figure 2.11c). Especially TopHat 2, PALMapper and STAR reported annotated splice junctions that were not supported by the simulated data. For novel junction discovery, GSTRUCT and MapSplice outperformed other methods (Figure 2.11d). The high performance of GSTRUCT can be explained by the generation of a splice index that includes junctions of possible alternate isoforms as well, which was one of the strategies to create novel splices in the simulated data sets.

In general, precision was higher for splices identified in the middle of reads compared to splices placed at read termini (data not shown). Also most reported splices corresponded to the main splice-signal GT-AG as indicated in blue in Figure 2.12. However, some splices with GC-AG and AT-AC splice signals were also reported, especially PALMapper reported novel splices with GC-AG splice signals. For novel splices, GSNAP, GSTRUCT, PALMapper reported a small fraction of splices with noncanonical splice signals while for ReadsMap and SMALT the majority of reported novel splice sites were corresponding to noncanonical splice signals.

Over the last few years, read length that can be produced by commonly used sequencers has increased continually. While early RNA-seq data sets had typically read lengths of 36 bp, 100 bp reads and longer are not unusual these days. Therefore, it is more and more important that aligners can detect the presence of multiple splice junctions within a single read. BAGET and SMALT never reported more than one intron per read; and PASS and PALMapper mostly reported reads with one or two introns, but rarely more. All other protocols were able to detect three of more introns as indicated in Table A.6. ReadsMap, STAR and the annotation based TopHat 2 protocol reported the greatest number of primary alignments with at least three introns. Annotation based TopHat2 reached the highest sensitivity for multi-intron alignments with a recall of 79.3% for simulation 1, see Table A.7. Among the protocols run without annotation, ReadsMap was the most sensitive for detecting alignments with three or more introns (recall of 72.1%), but at the expense of a low precision of 7.0%. For all aligners except ReadsMap and PALMapper, most multi-intron alignments tended to be correct.

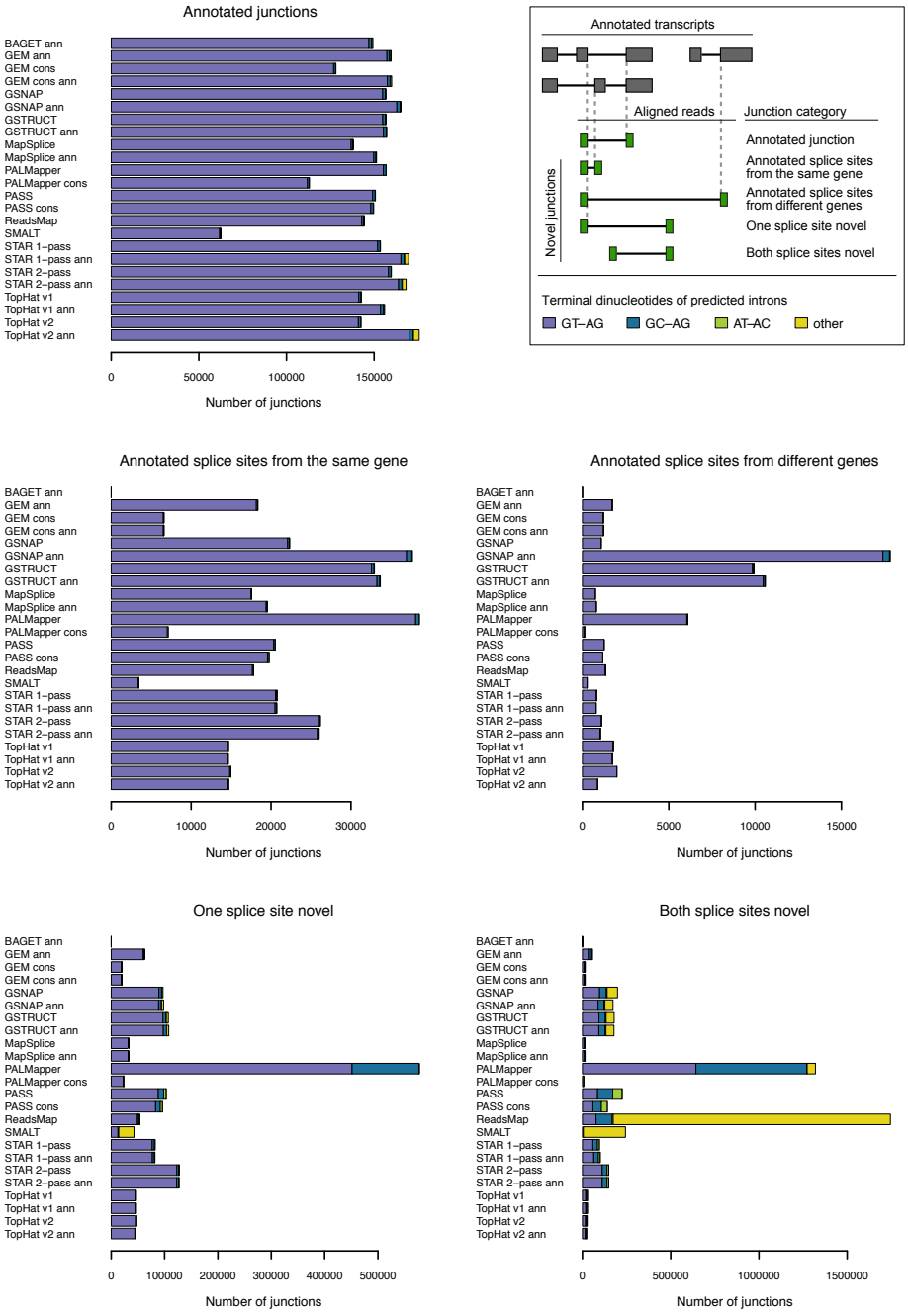


Figure 2.12: Classification of reported splice with respect to splice signals. Splice signal frequencies for the K562 whole cell replicate 1 data set. Frequencies were calculated as the total number of splices (as opposed to unique junctions) in primary alignments divided by the number of sequenced reads.

2.2.7 Influence of aligners on transcript reconstruction

The placement of spliced reads has a big influence on the down-stream analysis task of transcript reconstruction. As protocols displayed significant differences in their accuracy of splice junction detection, they might not all be equally suited for transcript reconstruction. In order to analyse the influence of the aligner choice on exon discovery and transcript reconstruction, Cufflinks was applied to all submitted alignments. GEM, GSNAP, GSTRUCT, MapSplice, STAR and TopHat alignments resulted in similar exon detection rates (Figure 2.13a). Up to 69% of the exons reported by Cufflinks matched exons present in the Ensembl annotation, and up to 51% of all annotated exons from protein-coding genes were recovered when using primary alignments only (K562 whole cell replicate 1). Exon detection by Cufflinks performed significantly lower for the other alignment programs (Figure 2.13a). Including secondary alignments negatively affected Cufflinks performance of GEM and PALMapper which report numerous such alignments, however, had little effect on other methods (Figure 2.13b).

Consistent with the results on real data, the six aligners noted above enabled highly accurate exon detection for the simulated data, with recall reaching 85% and precision 84% (Simulation 1, Figure 2.13a). For the second, more challenging simulated data set, the annotation based TopHat 2 protocol outperformed the other methods, closely followed by GSNAP (with annotation) and GSTRUCT (with or without annotation, Figure 2.13a). The same alignment methods gave the best Cufflinks accuracy for the more complex task of transcripts reconstruction (Figure 2.13b).

Surprisingly the TopHat 2 protocol using annotation, despite exhibiting relatively poor precision for spliced alignment (Figure 2.11b-d), produces the best input for transcript reconstruction. One underlying cause is that Cufflinks seems to be able to discard erroneous exon junctions in the input data at a high rate. For example, on the data from the first simulation, 71% of true junctions identified by the annotation-based TopHat 2 protocol were incorporated into transcript by Cufflinks, compared to 5% of false junctions as indicated in Table A.9. Another reason for the superior performance of annotation based TopHat 2 is its high detection rate for multi-intron alignments which significantly improves transcript reconstruction (see Table A.7).

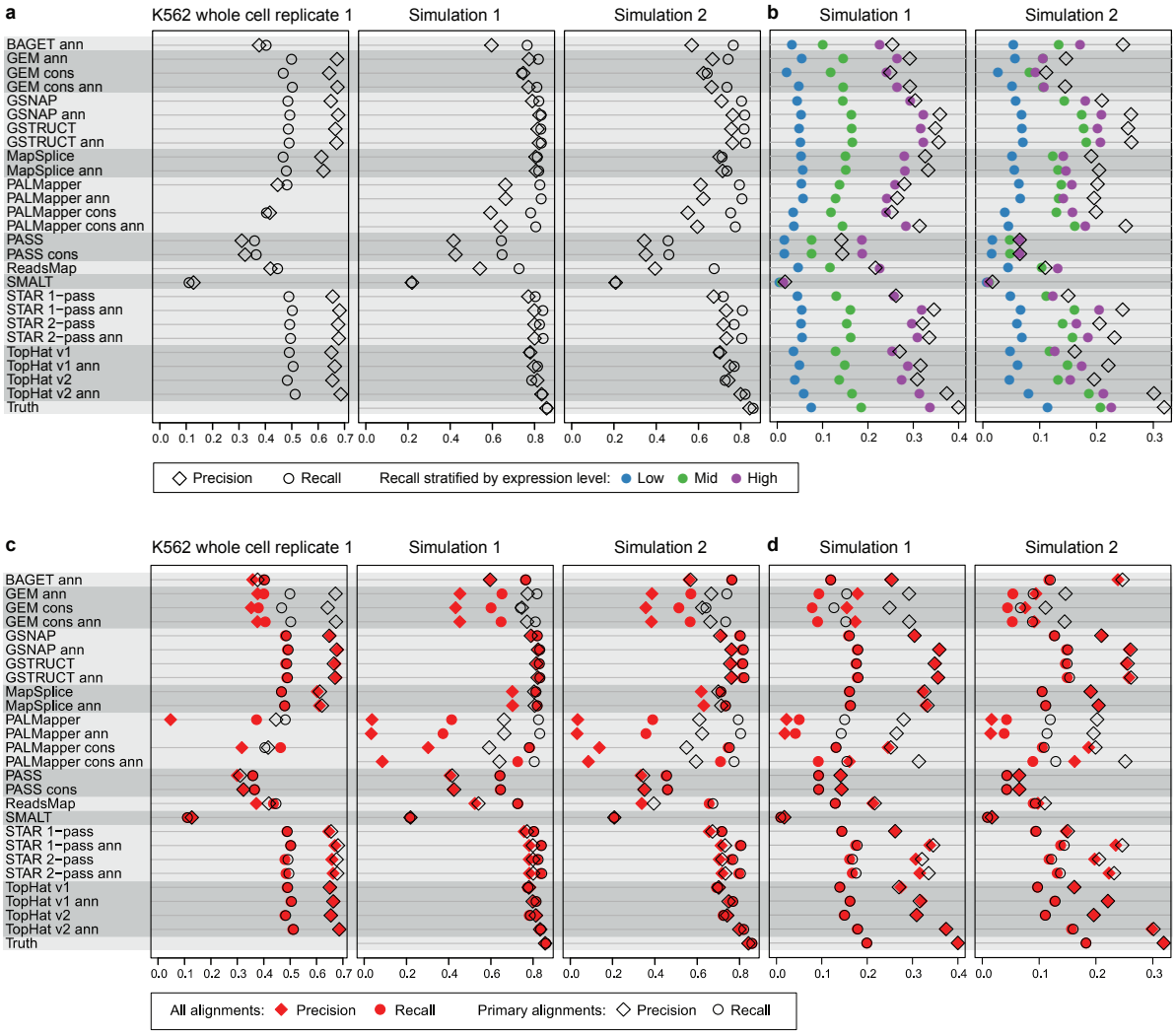


Figure 2.13: Aligner influence on transcript assembly. Cufflinks performance was assessed by measuring precision and recall for individual exons (a) and spliced transcripts (b). For K562 data, precision was defined as the fraction of predicted exons matching Ensembl annotation, and recall as the fraction of annotated exons that were predicted. Only exons from protein-coding genes were considered. Results on simulated data were benchmarked against simulated gene models, using analogous definitions of precision and recall. Simulated transcripts were divided into three groups of equal size according to expression level, where the "mid" groups comprise those with expression corresponding to an RNA-seq coverage of 121-646 (simulation 1) and 182-855 (simulation 2) reads per kb. (c-d) Effect of secondary alignments on transcript assembly by Cufflinks. Performance was assessed by measuring precision and recall for individual exons (c) and spliced transcripts (d), using all alignments from each protocol (red symbols) or the subset of primary alignments (open symbols). For K562 data, precision was defined as the fraction of predicted exons matching Ensembl annotation, and recall as the fraction of annotated exons that were predicted. Only exons from protein-coding genes were considered. Results on simulated data were benchmarked against simulated gene models, using analogous definitions of precision and recall. The last row shows the results obtained when using the perfect alignment produced by the simulator (Truth).

2.3 Conclusion

In this chapter, I presented a systematic evaluation of spliced alignment software for RNA-seq data, including the most widely used programs and several less well-known alternatives. To assess their performance, a comprehensive set of metrics was analysed. These ranged from basic alignment statistics, such as the percentage of mapped reads, to suitability for higher-order analyses such as transcript reconstruction. Large performance differences between aligners were apparent for all evaluated metrics.

The top four performing aligners in this study were MapSplice, GSNAP, GSTRUCT and STAR. A previous study by Grant et al. also reported good performance for MapSplice and GSNAP, but did not include GSTRUCT and STAR (Grant et al., 2011). Nevertheless, also these four methods displayed certain weaknesses. GSNAP, GSTRUCT and STAR reported many false exon junctions in their output and it might be advisable to filter junctions based on the number of supporting alignments. MapSplice on the other hand appeared to be a conservative aligner, both with respect to reporting mismatches and indels as well as exon junction calls.

The use of a dedicated spliced alignment program is crucial for the proper interpretation of RNA-seq data, as the results for BAGET and SMALT show. As expected, SMALT displayed the lowest junction recall on the simulated data as this method is able to split reads, but lacks the ability to identify exact exon-intron boundaries. Accordingly, transcript reconstruction based on SMALT performed very poorly. BAGET, which maps reads against exon junction sequences derived from reference genome annotation, can identify exact exon-intron boundaries. Still it showed a significantly reduced performance in junction detection compared to real spliced aligners as it is unable to detect new junctions that are not present in the reference annotation. Therefore, transcript reconstruction based on BAGET alignments also showed reduced performance even though much less prominent compared to SMALT.

Several methods included in this evaluation feature algorithmic innovations the incorporation of which could improve other RNA-seq aligners as well. As an example, PALMapper creates a map of nucleotide variations within the RNA-seq data compared to the reference genome enabling faster and more accurate alignment of reads containing such variants. This approach, however, is mostly applicable for mismatches and not directly transferable to the mapping of indels for which PALMapper shows quite poor

detection performance. The mapping of indels (but also mismatches) could be further improved by implementing a realignment strategy. This is commonly used in DNA re-sequencing studies, where variants are called after targeted multiple sequence alignment of reads independently mapping to the same locus.

GEM and PALMapper differed largely from the other methods included in this evaluation with respect to the fraction of multiple alignments reported. Reporting all possible positions for a read might be useful when the aim is to identify all genes including pseudogenes. However, if the aim is to identify transcripts that are actually expressed in a given sample, the excessive reporting of multiple mapping positions will cause the identification of transcripts that are not actually expressed. Accordingly, including all alignments decreased the performance of subsequent transcript reconstruction drastically for GEM and PALMapper while for all other methods performance was hardly influenced.

Surprisingly TopHat version 2 showed several deficits compared with version 1, exhibiting decreased junction precision when provided with an annotation and mapping a smaller fraction of reads with lower average base call quality. However, it outperformed all other methods with respect to transcript reconstruction with Cufflinks, facilitating a performance close to that observed for perfect alignments of simulated data. The main reason for this is most likely the higher accuracy for multi-intron alignments for TopHat version 2 compared with any other method. The accuracy for multi-intron alignments was significantly lower when running TopHat version 2 without annotation. This suggests that the direct mapping to spliced sequences performed by TopHat version 2 when provided with an annotation is the main reason for the higher multi-intron accuracy and subsequently the better transcript reconstruction with Cufflinks.

Similarly to TopHat also other methods produced better input for transcript reconstruction when provided with gene annotations, with the exception of MapSplice and GSTRUCT which only showed marginal improvements upon annotation usage. While provision of annotation improved transcript reconstruction it showed little effect on fundamental metrics, such as alignment yield and basewise accuracy. The use of annotation improved recovery of known exon junctions for GSNAP, STAR and TopHat 2, but also resulted in an increase of false positive junction calls, and hence lower accuracy overall. Only GSTRUCT and TopHat 1 consistently benefited from gene annotation, though

GSTRUCT performance was only marginal affected. Making optimal use of annotation information thus leaves a lot of space for improvements.

One more aspect that is important for the usability of a tool is its runtime. If a tool performs very well, but takes weeks to compute, the user will probably switch to a tool with similar performance and lower run time. As developers computed the alignments themselves, it was not possible to collect comparable runtime information. However, a recent study reported that GSNAP and MapSplice require similar runtimes while TopHat 2 and STAR run about 3 and 180 times faster, respectively (Dobin et al., 2013).

The advent of RNA-seq has prompted the development of sophisticated alignment tools for shotgun transcriptome sequencing. However, several algorithmic challenges remain, such as exploiting gene annotation without introducing bias, correctly placing multi-mapped reads, achieving optimal yet fast alignment around gaps and mismatches, and reducing the number of false exon junctions reported. Over the last few years, read length has been continuously increasing and this trend is likely to continue. Therefore, methods will require to deal with longer reads with higher error rates. Since longer reads are more likely to span multiple exon-intron boundaries, methods will need to perform more extensive spliced alignment. Differential treatment of these issues will enhance and expand the range of RNA-seq aligners suited to varied computational setups and analysis aims.

2.4 Methods

2.4.1 RNA-seq data

The human K562 data used here correspond to the K562 poly(A)+ samples produced at Cold Spring Harbor Laboratory for the ENCODE project (Djebali et al., 2012) and can be accessed at http://www.encodeproject.org. RNA-seq libraries were sequenced using a strand-specific protocol and comprise two biological replicates each of whole cell, cytoplasmic and nuclear RNA.

The mouse RNA-seq data set was produced at the Wellcome Trust Sanger Institute as part of the Mouse Genomes Project using brain tissue from adult mice of strain C57BL/6NJ. The library was sequenced using the standard Illumina protocol that does not retain strand information. These data have been previously described and are available from the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accessions ERR033015 and ERR033016.

Simulated RNA-seq data were generated as previously described (Grant et al., 2011), additionally modelling base-call errors and quality scores with simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS). It should be noted that alignment protocols making use of gene annotation were provided with annotation from Ensembl only, while the simulated transcriptomes were based on Ensembl as well as several additional gene catalogues. In addition, novel transcript isoforms were simulated. This corresponds to a realistic scenario, as knowledge of the transcriptome is incomplete even for well-studied organisms such as human and mouse. Retained introns were also simulated, to reflect the fact that a proportion of transcripts captured by RNA-seq correspond to pre-spliced mRNAs. Transcriptome simulation parameters were set as previously described (Grant et al., 2011). Briefly, substitution variants were introduced in exons at rates of 0.001 (simulation 1) and 0.0005 (simulation 2) events per bp, and indel polymorphisms at rates of 0.0005 (simulation 1) and 0.0025 (simulation 2). The proportion of signal originating from novel simulated transcript isoforms was 20% and 35% for simulations 1 and 2, respectively.

The program simNGS recreates observations from Illumina sequencing machines using the statistical models underlying the AYB base-calling software (Massingham and Goldman, 2012). To simulate base-call errors and quality scores, simNGS version 1.5 was

applied using a paired-end simulation model. The model was trained on intensity data released by Illumina from a sequencing run on the HiSeq instrument using the TruSeq chemistry.

All sequencing data (real and simulated) as well as the alignment files used in this study have been consolidated as a single experimental record in the ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress) under excession E-MTAB-1728.

2.4.2 Evaluation of alignments

Developer teams provided alignments in BAM format. These files were processed to ensure compliance with the SAM specification (Li et al., 2009) and eliminate formatting discrepancies that otherwise could have impacted the evaluation. Mismatch information (NM and MD tags) was stripped from the files and recomputed using the samtools command `calmd` to ensure that mismatches were counted in the same manner for all protocols (Li et al., 2009).

In the BAM format, alignment gaps in read sequences can be described either as deletions or introns. Small gaps are typically labelled deletions and longer gaps considered introns, but the exact criteria differ among aligners. To prevent the introduction of bias from such differences, we reclassified deletions and introns were appropriate. Specifically, for the results presented in Figures 2.6 and 2.7 an alignment gap in the read sequences was considered a deletion if shorter than 19 bp and otherwise counted as an intron. The adjustment only noticeably affected the results for GSNAP and GSTRUCT, which reported numerous long deletions (up 2.0% of the deletions in the output from these tools exceeded 18 bp).

For alignments of simulated RNA-seq data, accuracy metrics were computed by comparison with the alignments produced by the simulator. When computing basewise and indel accurracy, ambiguity in indel placement was accounted for (Grant et al., 2011). For example, in an alignment of the sequences `ATTTA` and `ATTA`, there are three equivalent gap placements in the latter sequence (`A-TTA`,`AT-TA` and `ATT-A`), all of which were considered correct. A general strategy was implemented to handle positional ambiguity for indels on any size.

Unless otherwise noted, evaluation metrics for alignments of K562 RNA-seq data were averaged over the six K562 data sets (Table 2.1). Comparisons with the gene

annotation were preformed using the Ensembl annotation that was provided to aligners. Comparisons to the simulated transcriptomes were carried ot as described in Results.

2.4.3 Transcript reconstruction

Transcript assembly was conducted with Cufflinks version 2.0.2. with the parameter `library-type` set to `fr-firststrand` for the K562 data, which are strand-specific, and to `tr-unstranded` for the simulated data, which are not. Default values were used for other parameters.

Cufflinks requires spliced alignments to have a SAM format tag (XS) indicating the genomic strand (plus or minus) on which the transcript represented by the read is likely to be encoded. Alignment programs such as TopHat can set the XS tag based on information about the library construction protocol (for strand-specific libraries) or by inspecting sequences at exon-intron boundaries. Five of the methods evaluated here (BAGET, GEM, ReadsMap, SMALT and STAR) did not provide XS tags; we therefore post-processed the alignment output from these methods to add them. For the strand-specific K562 data, XS tags were set based on alignment orientation and mate number, as done by TopHat. For alignments of simulated reads, we set XS tags according to the initial and terminal dinucleotides of the inferred introns, which are expected to be GT/AG, GC/AG or AT/AC for plus-strand transcripts and CT/AC, CT/GC or GT/AT for minus-strand transcripts (Iwata and Gotoh, 2011). For the XS tag to be added to an alignment, at least one exon junction was required to have these signals and conflicting signals among junctions were not allowed.

We noted that the annotation-based on TopHat 2 protocol uses the annotation provided to set the XS tag for unspliced alignments that overlap annotation exons. As this is a unique feature of TopHat 2 that might confer and advantage in the evaluation of transcript reconstruction, we investigated the effect of removing the XS tag from unspliced alignments in the TopHat 2 output prior to running Cufflinks. This modification had negligible effect on the Cufflinks accuracy metrics presented here (data not shown), demonstrating that provision of XS tags for unspliced alignments cannot explain why the annotation-based TopHat 2 protocol resulted in better Cufflinks performance than other protocols.

Precision and recall were computed as described in Chapter 3. Specifically, internal exons were required to be recovered with exact boundaries, first and terminal exons were required to have correctly predicted internal borders only, and exons constituting unspliced transcripts were scored as correct if covered to at least 90% by a predicted unspliced transcript. For the simulated data, only exons of spliced transcripts were required to be placed on the correct strand, as orientation of single-exon transcripts cannot be reliably predicted unless RNA-seq libraries are strand-specific. Spliced transcripts were considered to be correctly assembled if the strand and all exon junctions matched.

Chapter 3

Automated transcript reconstruction from RNA-seq data

3.1 Introduction

RNA-seq provides a genome-wide readout of expressed transcript isoforms. Current sequencing technologies, however, cannot sequence full transcripts yet and short reads need to be computationally assembled into gene structures. Transcript reconstruction from RNA-seq data enables the identification of expressed isoforms, their quantification and the subsequent identification of differentially expressed isoforms.

3.1.1 The early days of gene prediction

Sequencing of whole genomes generated the need for automated annotation of these sequences. Automated gene prediction algorithms were designed to identify protein-coding genes within these genomes. The first gene predictors could be divided into two main categories: *de novo* or alignment based predictors (see Brent, 2008 for a review). De novo predictors make use of genomic sequences only. Single-genome de novo gene predictors, identify protein coding genes solely from the presence of coding sequences, splice signals, translational start and stop signals and polyadenylation signs. Dual-genome de novo gene predictors can furthermore incorporate sequence conservation between species as evidence for functional sequences. Alignment based predictors use the alignment of sequences derived from cDNA libraries (*cis* alignment) or from cDNAs

from homologous genes in the same or another species (*trans* alignment). Alternatively, protein sequences are aligned back to the genome instead of the cDNAs directly.

A first attempt to evaluate the quality of automated gene prediction algorithms on a defined test data set was the "genome annotation assessment project" (GASP) which was launched in 2000 (Reese et al., 2000). It assessed the performance of gene prediction algorithms on 2.9 Mb of the well-characterised ADH region of the drosophila genome. This was followed up by the EGASP project in 2006 (Guigo et al., 2006), which again assessed the status of automated gene prediction algorithms, but in the 44 regions selected by the Encode pilot project totalling 30 Mb (1% of the human genome, Harrow et al., 2012; ENCODE Project Consortium et al., 2007). At that time, the best methods were predicting at least one transcript correctly for 70% of genes. However, the accuracy of predicting all alternatively spliced transcripts was approximately 40-50%. NGASP analysing gene predictions in nematode genomes followed in 2008 (Coghlan et al., 2008) with comparable prediction accuracy to EGASP.

3.1.2 Transcript reconstruction from RNA-seq data

During the last decade more high-throughput technologies for analysing expressed genes have been developed. Strand specific tiling arrays allowed for the identification of unannoated transcripts including non-coding genes and antisense transcripts (Bertone et al., 2004; Huber et al., 2006; David et al., 2006; Xu et al., 2011). Although, tiling arrays can provide some information about alternative splicing between samples (Eichner, 2013), they have limited support for the study of alternatively splice isoforms within the same sample. RNA-seq provides the same power when it comes to the detection of unannotated transcripts at a lower cost. One single experiment enables the quantification of known genes as well as the discovery of unannotated gene products and novel splicing events. With adequate sequencing depth the dynamic range of RNA-seq data substantially exceeds that of microarray technologies providing a more detailed picture of the transcriptional landscape of a given system (Sultan et al., 2008; Nagalakshmi et al., 2008; Cloonan et al., 2008; Wang et al., 2009). However, with the current technology it is not yet possible to sequence full transcripts in a high-throughput fashion, but only short reads obtained from fragmented cDNA. To obtain full transcript sequences these short reads need to be computationally assembled into gene structures. Transcript reconstruc-

tion can be done by either mapping the reads to the genome and inferring transcript models from the read alignments (Trapnell et al., 2010; Roberts et al., 2011a; Li et al., 2011; Mezlini et al., 2013) or by de novo assembly of the reads, where the transcripts are reconstructed without the use of a reference genome, but solely by utilizing sequence overlaps between reads (Schulz et al., 2012; Grabherr et al., 2011; Robertson et al., 2010).

Both approaches are computationally heavy. The most limiting factor is sequencing coverage. Lowly expressed genes, therefore, pose a difficult challenge for transcript reconstruction and quantification. Some methods exploit features in the underlying genomic sequence, such as splice signals and translational start and stop signals, as further evidence for coding regions (Stanke et al., 2006; Schweikert et al., 2009; Blanco et al., 2007), however, this approach can not be applied to the discovery of non-coding features. Another challenge is the presence of multiple alternative isoforms of a gene (Wang et al., 2008; Nilsen and Graveley, 2010). It is not always possible to unambiguously quantify the isoforms as there may be more than one set of isoform quantifications that fit the observed gene coverage. As read coverage tends to be lower at gene termini, the exact borders of untranslated regions (UTRs) are hard to determine from RNA-seq data. Also these may be ill-defined in the reference annotation or vary accross biological samples (Lenhard et al., 2012; Di Giammartino et al., 2011; Tian et al., 2005). Additional challenges arise from non-uniform exon coverage because of sequencing bias (Wu et al., 2011; Roberts et al., 2011b; Jones et al., 2012), and read mapping to homologous sequences elsewhere in the genome (Lee and Schatz, 2012; Derrien et al., 2012). All together, these factors impair the identification of transcript isoforms and their accurate quantification.

3.1.3 The benchmark

RNA-seq data were generated as part of the ENCODE (Harrow et al., 2012) and modENCODE projects (modENCODE Consortium et al., 2010), along with a third data set of compatible sequencing format and read depth, and represent three widely-studied species: *Homo sapiens* (liver hepatocellular carcinoma cell line HepG2) (Djebali et al., 2012), *Drosophila melanogaster* (L3 stage larvae) (Graveley et al., 2011), and *Caenorhabditis elegans* (L3 stage larvae) (Mortazavi et al., 2010). For all three species high quality

annotated reference genomes are available, which is crucial to minimize the impact of alignment problems due to low quality genome assemblies, and to use as a gold standard for the predictions. As not all genes are expressed in a sample, the annotations were filtered to represent expressed transcript isoforms only (see Methods). Another motive for the three selected species was their varying transcriptome complexity (see Table B.1. In order to analyse the impact of transcriptome complexity on transcript reconstruction, the data sets were derived with similar library construction methods and the same read length. All three data sets were sequenced on the Illumina platform in 76 nt paired-end format to obtain approximately 100 million read pairs per sample (see Methods).

These three data sets were used to assess the performance of 25 transcript reconstruction protocols based on 14 different software packages. Similar as in the EGASP project (Guigo et al., 2006), developers of leading software programs were invited to participate in the RNA-seq Genome Annotation Assessment Project (RGASP), to benchmark methods to predict and quantify expressed transcripts from RNA-seq data. This ensured that programs were executed in the intended manner and with appropriate settings. Most evaluated methods are based on genome alignments (AUGUSTUS (Stanke et al., 2006), Cufflinks (Roberts et al., 2011a), Exonerate (Slater and Birney, 2005), GSTRUCT, iReckon (Mezlini et al., 2013), mGene (Schweikert et al., 2009), mTim, NextGeneid (Blanco et al., 2007), SLIDE (Li et al., 2011), Transomics, Trembly, and Tromer (Sperisen et al., 2004)), but also two de novo assembly methods participated (Oases (Schulz et al., 2012) and Velvet (Zerbino and Birney, 2008)).

3.1.4 Outline

In order to assess the quality of these transcript reconstruction tools and to identify aspects that impede or promote transcript assembly, performance was compared using a set of increasingly stringent metrics. All predictions were compared against a filtered annotation containing only transcripts that show evidence of being expressed. A manuscript has been accepted for publication by *Nature Methods*[1]. NanoString quantification was performed in Ali Mortazavi's lab, Biological Sciences III, University of California Irvine, Irvine, USA. Pär Engström, European Molecular Biology Laboratory,

[1]Steijger T, Abril J F, Engström P G, Kokocinski F, RGASP Consortium, Hubbard T J, Guigo R, Harrow J and Bertone P. **Assessment of transcript reconstruction methods for RNA-seq**, *Nature Methods* (in Press)

European Bioinformatics Institute, Cambridge UK, provided the code for the nucleotide level analysis and filtered the reference annotation. All other metrics described in this chapter have been implemented and calculated by me. Supplementary information is available in Appendix B.

3.2 Results

A total of 25 transcript reconstruction protocols were evaluated, based on alternate parameter usage of 14 different software packages (Table 3.1). Programs were run by the original developers, with the exception of SLIDE and iReckon. These became available during the analytical phase of the project and were subsequently included for further comparison. As the aim of this project was to evaluated the performance at reconstructing transcriptomes from RNA-seq data without prior knowledge of gene content, programs were run without genome annotation to guide transcript assembly. SLIDE and iReckon, however, require some annotation information and were run with minimum requirements (see Methods).

3.2.1 Nucleotide level evaluation

As a first measure of performance, basewise agreement between the annotation and predictions was calculated. Nucleotide level sensitivity denotes the fraction of bases in annotated features that were covered by assembled transcripts, irrespective of the strand. Precision consequently was calculated as the proportion of reported exonic sequences that matched known exons. Most protocols exceeded a sensitivity of 75% for both model organisms, but sensitivity was generally higher for *C. elegans* (Figure 3.1 and Table B.2). Notably, the methods AUGUSTUS, mGene and Transomics reached almost 100% sensitivity while their precision was not noticeably lower than for other methods. Precision, however, was generally lower for *C. elegans*, which can be explained by the lack of extensive UTR information in the *C. elegans* annotation. Performance decreased for human data, where tradeoffs between between precision and sensitivity were more apparent. Seven unguided methods (AUGUSTUS, Exonerate, GSTRUCT, NextGeneid, Oases, Trembly and Velvet) attained both precision and sensitivity above 60% on the human data. However, the greatest sensitivity for methods without annota-

Table 3.1: Developer team submission details. CDS - coding sequence, n.a. - protocol was run by me.

Developer team	Protocol designation	Underlying alignment programs	CDS	Quantified features	Multiple transcripts reported per gene
H. sapiens					
Iseli	Tromer	fetchGWI, megablast SIBsim4	yes	transcript	yes
Gerstein	Trembly all	TopHat	no	transcript	yes
	Trembly high	TopHat	no	transcript	yes
Rtsch	mGene	PALMapper	yes	transcript	yes
	mGene graph	PALMapper	yes	transcript	yes
	mTim	PALMapper	yes	transcript	yes
Richard	Oases	BLAT	no	no	yes
Pachter	Cufflinks	TopHat	no	transcript	yes
Stanke	AUGUSTUS high	BLAT	yes	transcript	yes
	AUGUSTUS all	BLAT	yes	transcript	yes
	AUGUSTUS de-novo	n.a	yes	transcript	no
Searle	Exonerate SM all	Exonerate	yes	no	no
	Exonerate SM high	Exonerate	yes	no	no
Wu	GSTRUCT	GSNAP	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no
	NextgeneidAS	GEM	yes	no	no
	NextgeneidAS de-novo	GEM	yes	no	no
Solovyev	Transomics all		yes	transcript	no
	Transomics high		yes	transcript	no
Wold	Velvet	BLAT	no	exon	yes
	Velvet AUGUSTUS	BLAT	no	exon	yes
n.a	iReckon full	TopHat	no	transcript	yes
	iReckon ends	TopHat	no	transcript	yes
n.a	SLIDE all	TopHat	no	transcript	yes
	SLIDE high	TopHat	no	transcript	yes
D. melanogaster					
Iseli	Tromer	fetchGWI, megablast SIBsim4	yes	transcript	yes
Rtsch	mGene	PALMapper	yes	transcript	yes
	mGene graph	PALMapper	yes	transcript	yes
	mTim	PALMapper	yes	transcript	yes
Richard	Oases	BLAT	no	no	yes
Pachter	Cufflinks	TopHat	no	transcript	yes
Stanke	AUGUSTUS all	BLAT	yes	transcript	yes
	AUGUSTUS de-novo	n.a.	yes	transcript	no
Wu	GSTRUCT	GSNAP	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no
	NextgeneidAS	GEM	yes	no	no
	NextgeneidAS de-novo	GEM	yes	no	no
Solovyev	Transomics all		yes	transcript	no
	Transomics high		yes	transcript	no
Wold	Velvet	BLAT	no	exon	yes
n.a.	iReckon full	TopHat	no	transcript	yes
	iReckon ends	TopHat	no	transcript	yes
n.a.	SLIDE all	TopHat	no	transcript	yes
	SLIDE high	TopHat	no	transcript	yes
C. elegans					
Iseli	Tromer	fetchGWI, megablast SIBsim4	yes	transcript	yes
Rtsch	mGene	PALMapper	yes	transcript	yes
	mGene graph	PALMapper	yes	transcript	yes
	mTim	PALMapper	yes	transcript	yes
Richard	Oases	BLAT	no	no	yes
Pachter	Cufflinks	TopHat	no	transcript	yes
Stanke	AUGUSTUS high	BLAT	yes	transcript	yes
	AUGUSTUS all	BLAT	yes	transcript	yes
	AUGUSTUS de-novo	n.a.	yes	transcript	no
Searle	Exonerate SM all	Exonerate	yes	no	no
	Exonerate SM high	Exonerate	yes	no	no
Wu	GSTRUCT	GSNAP	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no
	NextgeneidAS	GEM	yes	no	no
	NextgeneidAS de-novo	GEM	yes	no	no
Solovyev	Transomics all		yes	transcript	no
	Transomics high		yes	transcript	no
Wold	Velvet	BLAT	no	exon	yes
	Velvet AUGUSTUS	BLAT	no	exon	yes
n.a.	iReckon full	TopHat	no	transcript	yes
	iReckon ends	TopHat	no	transcript	yes
n.a.	SLIDE all	TopHat	no	transcript	yes
	SLIDE high	TopHat	no	transcript	yes

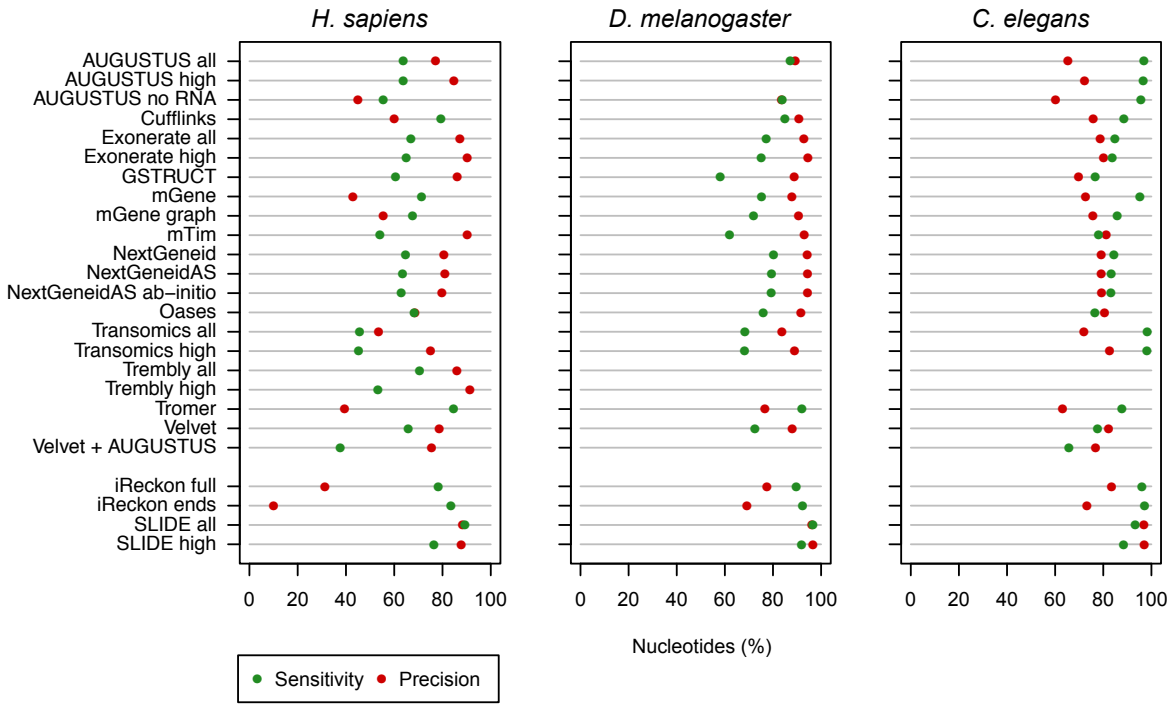


Figure 3.1: Summary of nucleotide-level performance for the methods evaluated. Performance at detecting exonic Nucleotides. Sensitivity (green) indicates the proportion of known exon sequence in each genome that were covered by assembled transcripts, and precision (red) the proportion of reported expressed sequence that corresponded to known exon sequences. Programs run with gene annotation are grouped separately.

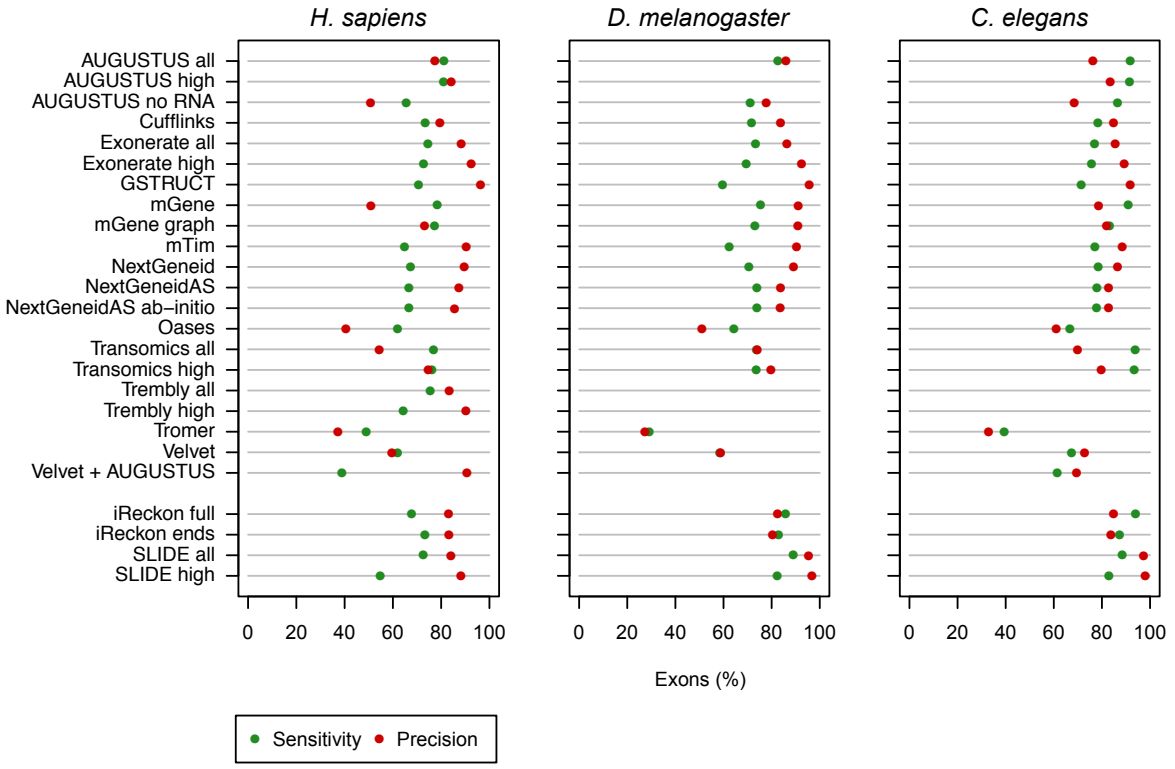


Figure 3.2: Summary of exon-level performance for the methods evaluated. Performance at detecting individual exons, shown as the percentage of reference exons with a matching feature in the submission (sensitivity, green), and the proportion of reported exons that agree with annotation (precision, red). To account for biological variation in transcript start and end sites, external boundaries of first and last exons were allowed to differ from the reference annotation (see Methods).

tion was observed for Tromer and Cufflinks at the cost of low precision. These programs consistently displayed high sensitivity across the three species, but the low precision of Tromer in particular indicates a tendency for overprediction. SLIDE and iReckon must be provided with gene annotation, and therefore outperform most other methods. iReckon, however, suffers from low precision at the nucleotide level due to the prediction of transcript isoforms with retained introns. Notably, the AUGUSTUS de novo protocol, which predicts transcripts using the genomic sequence alone, reached nearly the same level of sensitivity as the corresponding protocol that also integrates RNA-seq data, but with significantly lower precision.

3.2.2 Exon identification from RNA-seq data

Next the ability of each method to identify individual exons from RNA-seq data relative to the reference annotation was assessed (Figure 3.2). Exact agreement in translation start site and stop site positioning between predicted and annotated coding exons was extremely rare, while internal exons, which can be inferred from spliced alignments, were identified at high rate (Figure 3.3, left). This trend was even stronger when including non-coding exons of protein coding genes, that is UTR exons where prediction accuracy for transcription start and polyadenylation sites was almost zero (Figure 3.3, center). As these discrepancies may result from biological variation in precise transcript boundaries (Lenhard et al., 2012; Di Giammartino et al., 2011; Tian et al., 2005), 5′-ends of first exons and 3′-ends of terminal exons were allowed to differ from the reference coordinates (see Methods). This led to significant improvements (Figure 3.3, right). Most methods exhibited the lowest exon detection rates for the human RNA-seq data (Figure 3.2). Note that for all three species the performance of most methods approached that of iReckon and SLIDE, despite the latter two benefiting from the use of high-quality gene annotation.

Coding exons can be identified directly from genomic sequence by the presence of translation start/stop sites and splice acceptors and donors. Programs such as AUGUSTUS, Exonerate, mGene, mTim, NextGeneid, Tromer and Transomics exploit these features to improve exon discovery. Of these, AUGUSTUS, mGene, and Transomics identified a greater proportion of annotated coding exons than mTim, Exonerate, NextGeneid, and Tromer (Table B.3). These methods augment data-driven transcript reconstruction with *ab initio* gene prediction, suggesting that higher sensitivity measures are due to more extensive utilization of the underlying genomic sequence, thereby reducing the need for support from RNA-seq data.

RNA-seq read coverage had a high impact on exon detection rates (Figure 3.4). Through the use of *ab initio* prediction, AUGUSTUS, mGene and Transomics were able to detect exons from protein-coding transcripts present at very low abundance. All other methods required a minimum average read coverage to detect exons. Exon detection increased with read coverage at a roughly linear rate until reaching a plateau. One exception was Tromer, which often reported short exon fragments of 50-75 bp flanking introns and failed to extend them to full exons, exemplarily shown in Figure 3.5c.

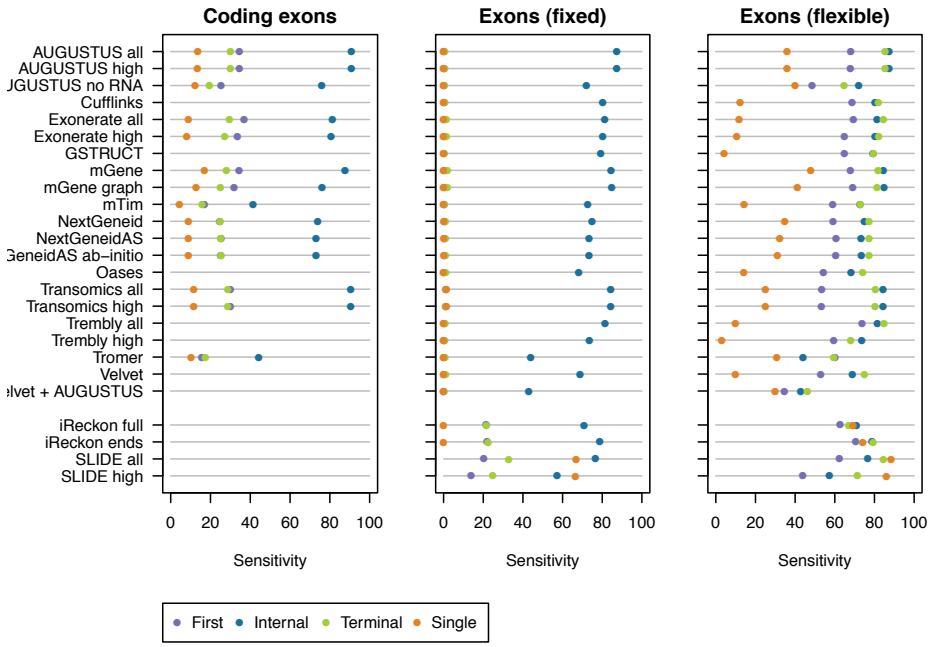


Figure 3.3: Influence of exon rank on exon detection performance. Detection sensitivity for annotated human exons classified as first, internal, terminal or single (that is those comprising an entire transcript). Exon boundaries were required to be predicted exactly as annotated (left, center) or by employing relaxed criteria for exons at transcript boundaries (right).

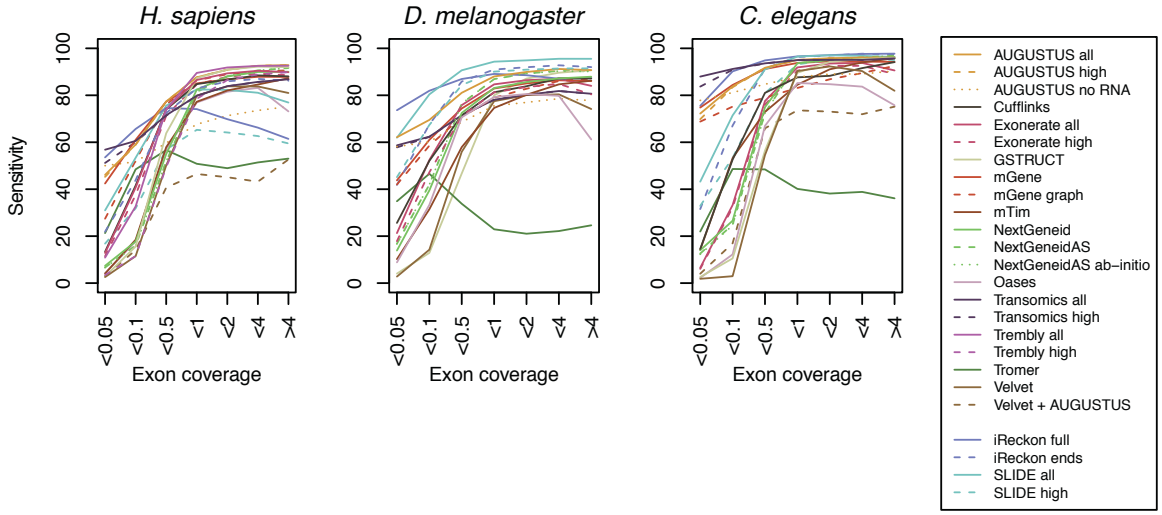


Figure 3.4: Influence of sequencing coverage on exon detection performance. Sensitivity for detection of annotated exons stratified by read depth.

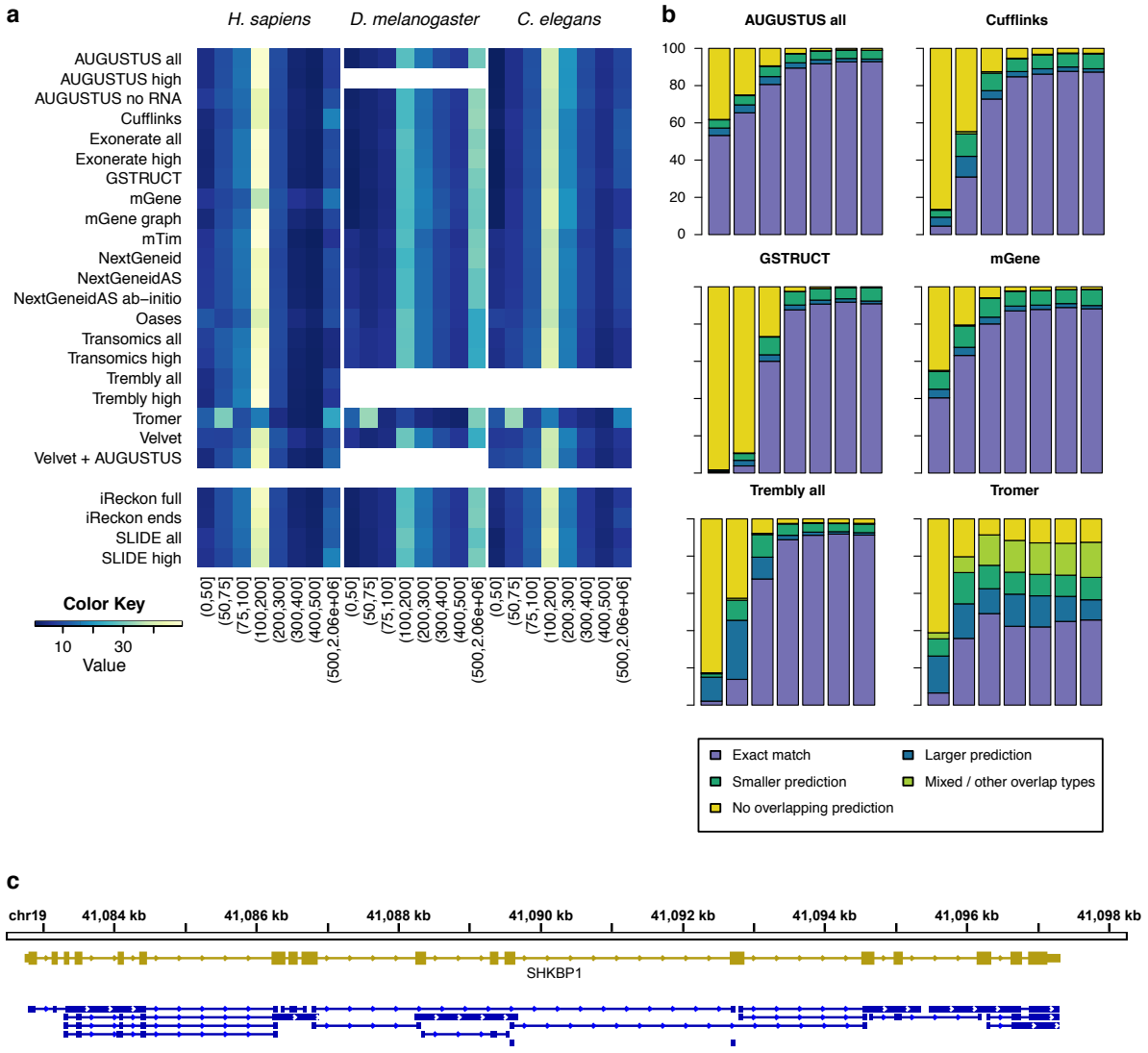


Figure 3.5: Exon length distribution and internal exon detection rate stratified by read coverage. (a) Exon length distributions in transcriptome assembly results. Colours indicate percentage of exons within the indicated length intervals. (b) Exon detection rate stratified by read coverage exemplarily for AUGUSTUS, Cufflinks, GSTRUCT, mGene, Trembly and Tromer. Bars indicate the percentage of annotated internal exons (of human protein-coding genes) that overlap with reported exons. Reference exons were binned by read coverage (x-axis) and further classified based on overlap with predicted exons (inset legend). Specifically, the classes represent exons with a perfectly matching prediction (green); exons for which all overlapping predictions span a larger region, including the entire reference exon (dark blue); exons for which all overlapping predictions are contained within the reference exon (light blue); and exons with other or multiple overlap types (pink). Note the high frequency of imperfect overlaps for Tromer. See also Figure 3.4. (c) Isoforms predicted by Tromer (blue) at the SHKBP1 locus (yellow). The predicted isoforms contain several examples of Tromer failing to extend exons to full exons as well as predicting exons spanning several annotated exons.

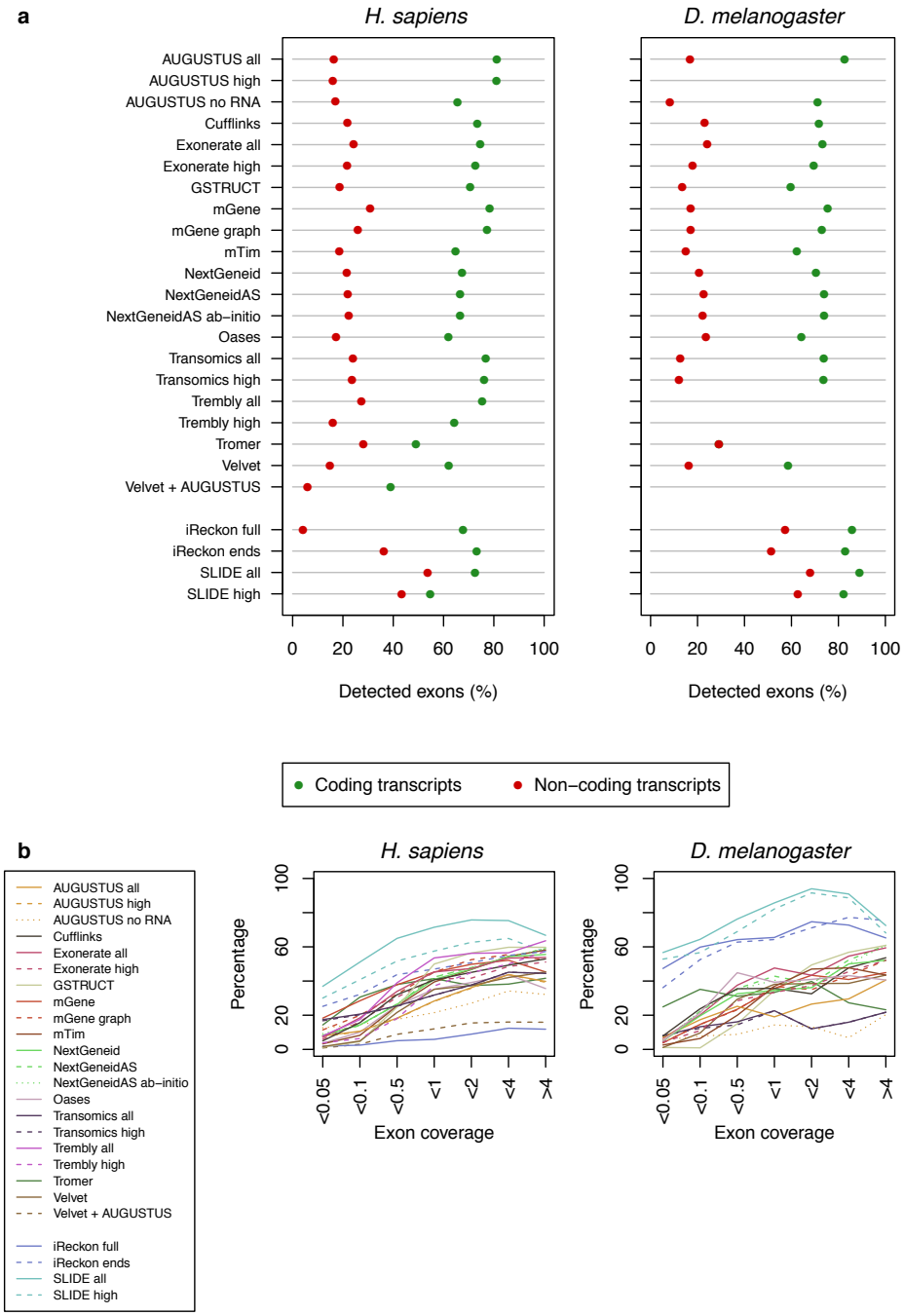


Figure 3.6: Exon-level performance for non-coding exons. (a) Exon detection sensitivity relative to coding potential. Percentage of detected exons belonging to coding (green) and non-coding (red) transcripts in *H. sapiens* and *D. melanogaster*. (b) Influence of sequencing coverage on non-coding exon-level sensitivity. Annotated exons of non-coding transcripts were binned according to RNA-seq read coverage and method sensitivities were calculated for each bin separately.

Therefore, Tromer predicted a higher fraction of very short exons compared to any other method (Figure 3.5a), while also showing a tendency to predict very long exons spanning multiple annotated exons (Figure 3.5b and c). This trend increased with read coverage (Figure 3.5b). To a lesser extent Oases and Velvet also showed reduced performance for high coverage exons (Figure 3.4).

The advantage of AUGUSTUS, mGene and Transomics was lost for non-coding transcripts, which lack the signals exploited by these methods for coding sequence prediction. Hence, these methods needed a certain read coverage in order to identify non-coding transcripts (Figure 3.6b). Only annotation-based methods like SLIDE and iReckon were able to identify lowly expressed non-coding transcripts at high rate. All unguided methods detected exons of non-coding transcripts with lower accuracy (Figure 3.6a). Non-coding RNAs tend to be expressed at lower levels than protein-coding genes (data not shown), but even when controlling for differences in read coverage they were detected with lower sensitivity (Figure 3.4 versus Figure 3.6b).

As the RNA-seq data used in this study did not retain strand information, strand orientation had to be inferred from sequence features such as translation start and stop signal or from spicing signals at spliced reads. Methods that did not exploit the genomic sequence could therefore not identify the strand orientation for unspliced transcripts (Figure 3.3). Accordingly, AUGUSTUS, mGene, and NextGeneid outperformed other methods in the assignment of single-exon transcripts to the correct strand (Figure 3.3).

3.2.3 Intron detection from RNA-seq data

A distinguishing feature of the three species used for this study is intron composition, where the relative number and size of introns differ markedly between the two model organisms and the human genome. Whereas greater than 73% of all introns in *C. elegans* and *D. melanogaster* are under 500 bp in length, this is the case for only 25% of introns in *H. sapiens* (Table B.1).

Figure 3.7a shows the degree to which intron length influenced the intron detection capabilities of various methods. Overall AUGUSTUS, mGene and Transomics showed the highest intron detection rates. However, Transomics exhibited a sharper decline with increased intron length. This trend was apparent for all methods except Tromer, which exhibited a markedly lower detection rate for introns shorter than 300 bp. For the *C.*

Figure 3

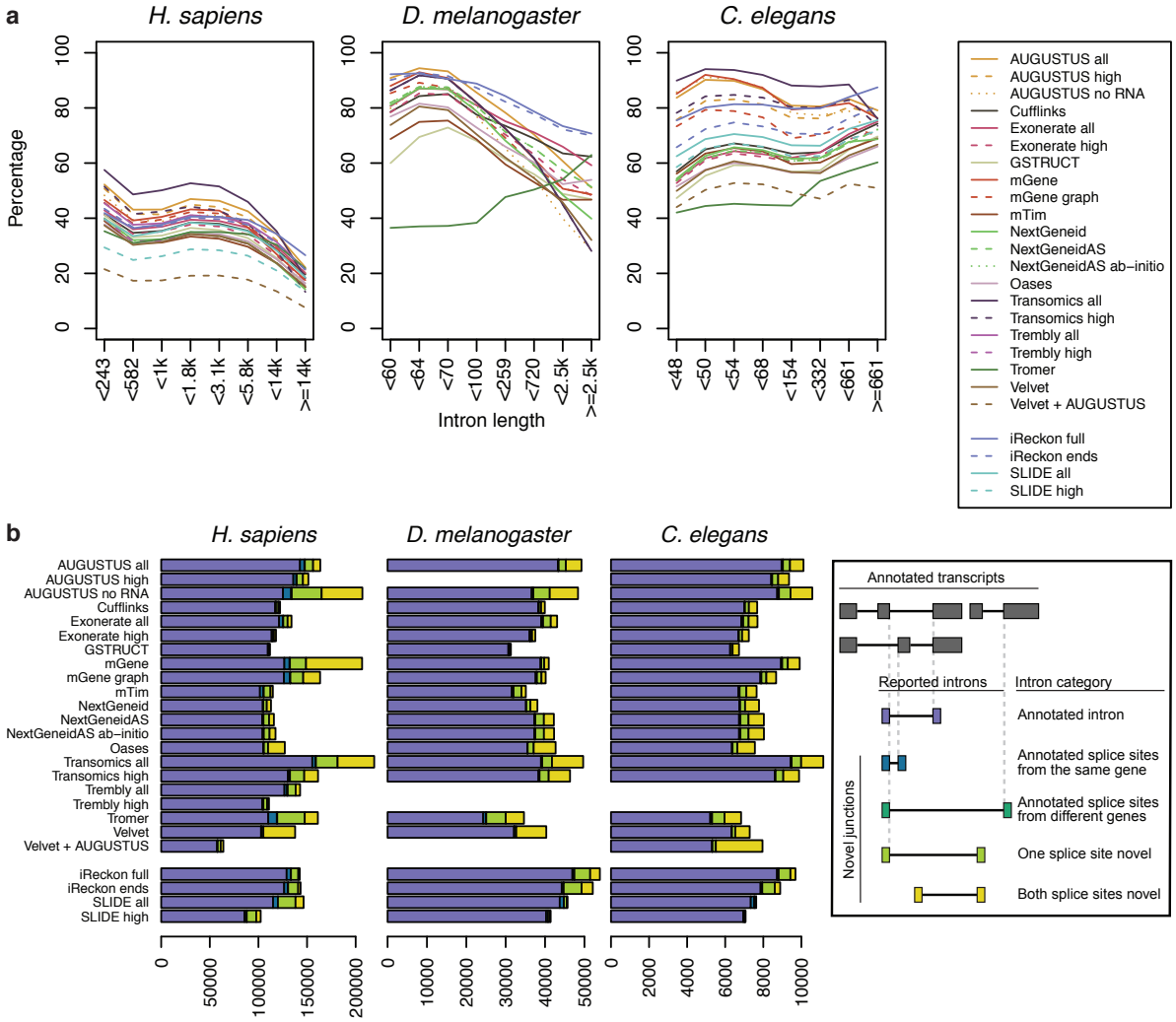


Figure 3.7: Intron detection performance. (a) Annotated introns were binned on length and sensitivity was calculated separately for each bin. (b) Reported introns were classified by overlap with splice sites annotated in the reference gene sets.

elegans data, the original Cufflinks submission failed to identify any intron under 50 bp in length (data not shown). This was due to the standard behaviour of the software: a parameter specifies the minimum intron length, with a default setting of 50 bp. This is not appropriate for species with a large fraction of short introns, such as *C. elegans* where this population comprises 25% (Table B.1). To more accurately capture the behaviour of the program, Cufflinks was re-run on the same TopHat alignments used as input to iReckon and SLIDE, for which minimum intron lengths were specified according to the intron size distributions in each genome (see Methods).

To better characterise the differences in intron detection between methods, reported introns were classified based on overlap with known splice sites (Figure 3.7b). Most protocols predominantly detected known introns; several, however, also predicted a substantial number of introns with one or two novel splice sites. The highest frequencies of novel junctions were predicted by mGene, Transomics, Tromer, Velvet, and the AUGUSTUS protocol that only used genomic sequence.

Intron detection is highly dependent on the underlying read alignments. For example, PALMapper (Jean et al., 2010) was used as the alignment component in the mGene and mTim protocols. As shown in Chapter 2, this aligner places more reads across unannotated splice sites than does GEM (Marco-Sola et al., 2012), GSNAP (Wu and Nacu, 2010), and TopHat (Trapnell et al., 2009; Kim et al., 2013); the latter programs form part of the NextGeneid, GSTRUCT, and Cufflinks protocols, respectively. These differences are reflected in the numbers of reported introns, where the behavior of the alignment software influences the likelihood of a given method to produce novel junction calls.

3.2.4 Assembly of exons into transcript isoforms

In this section, the ability of each method to link exons into defined splice products based on RNA-seq data is analysed. Initially the gene loci for which any expression was reported were determined, regardless of whether a valid transcript was identified, followed by those consistent with at least one annotated isoform (Figure 3.8). Although for most methods expression was detected at more than 80% of all genes, performance decreased markedly when considering only genes for which at least one annotated transcript had been identified. For unguided transcript reconstruction, valid isoforms are

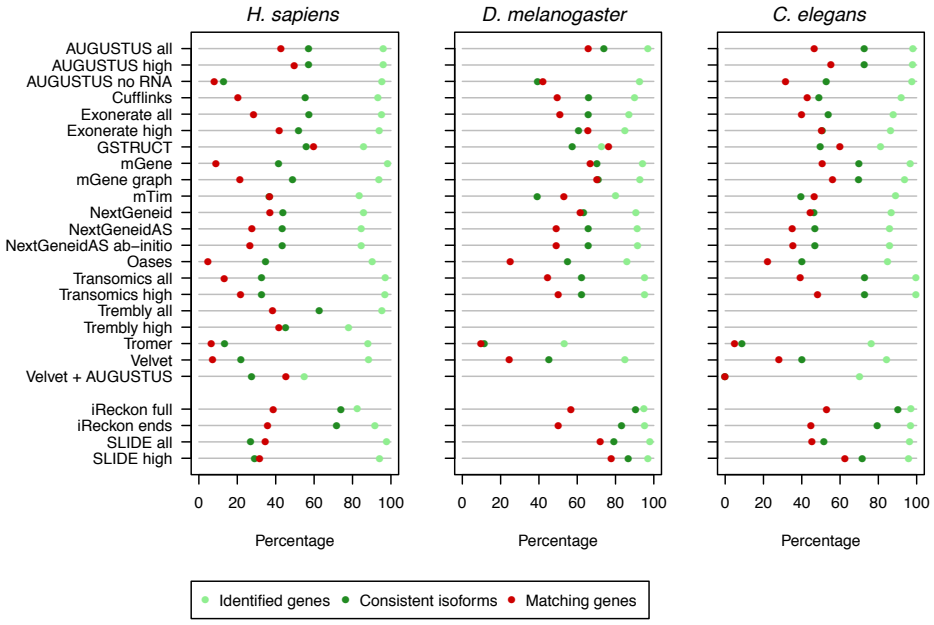


Figure 3.8: Gene detection performance. Light green dots indicate the percentage of genes for which some overlapping transcript was predicted. Dark green points indicate the percentage of reference genes with matching assembled transcripts (sensitivity) and red ones the percentage of reported genes with at least one transcript matching the reference (precision).

assembled for roughly half of expressed genes on average (*H. sapiens* mean 41%, max 61%; *D. melanogaster* mean 55%, max 73%; *C. elegans* mean 50%, max 73%), and for those only one isoform was typically identified (Figure 3.9).

A significant reduction in sensitivity was also observed from the gene to transcript level, even when using the flexible evaluation mode for first and terminal exons described above (Figure 3.10a and Tables B.3- B.5). The best performing methods identified at most 56-59% of spliced protein-coding transcripts from *C. elegans* (AUGUSTUS, mGene and Transomics), 43% from *D. melanogaster* (AUGUSTUS) and merely 21% from *H. sapiens* (Trembly). Sensitivity increased by roughly 10% when partial isoform matches were considered, as did precision when including partial predictions consistent with annotated isoforms (Figure 3.10).

Greater sequencing depth improved transcript assembly for *D. melanogaster* and *C. elegans* (Figure 3.11a), whereas in *H. sapiens* transcript detection remained low despite sequencing coverage in excess of 4.000 reads per kb of exon sequence. Generally, at least one consistent isoform was identified for highly expressed genes: $>$ 50% in *D.*

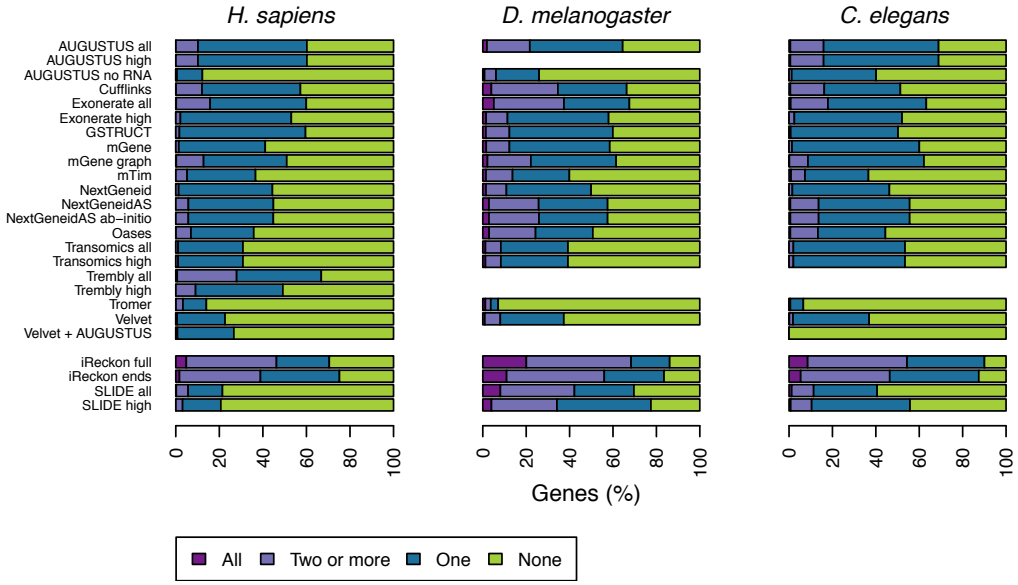


Figure 3.9: Number of isoforms detected per gene. Genes with at least three annotated splice products for which various subsets have been reported.

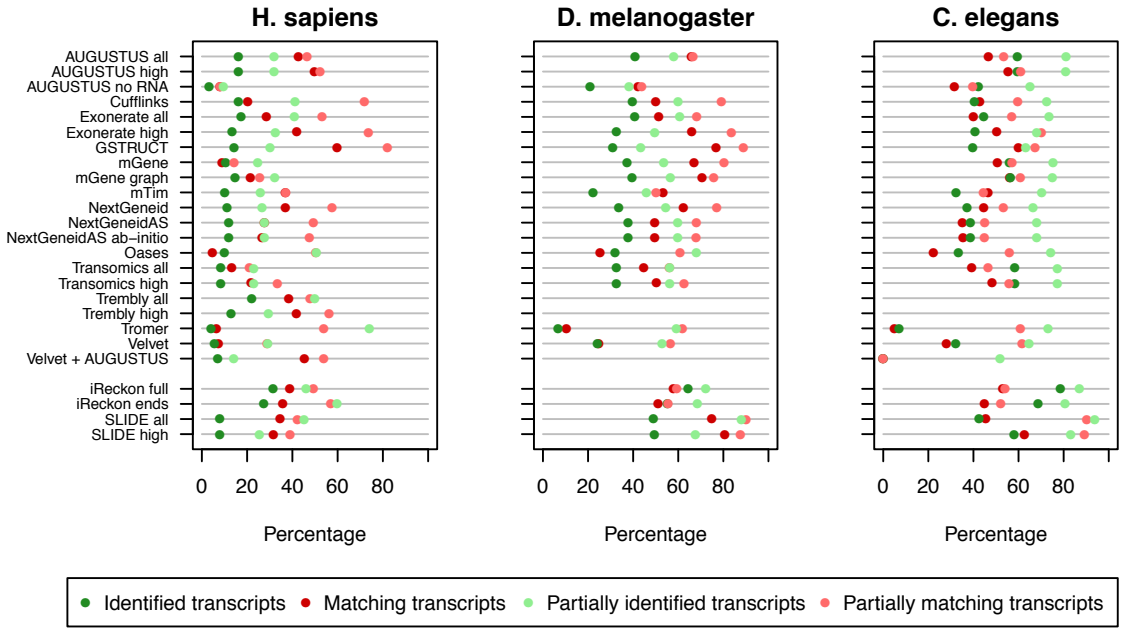


Figure 3.10: Transcript level performance. Reference transcripts with a matching submission entry (transcript-level sensitivity, dark green) and reported transcripts that match the reference (transcript-level precision, dark red). Reference transcripts with partial matching submission entry (transcript-level sensitivity (partial), light green) and reported transcripts that are consistent with the reference (transcript level precision (partial), light red).

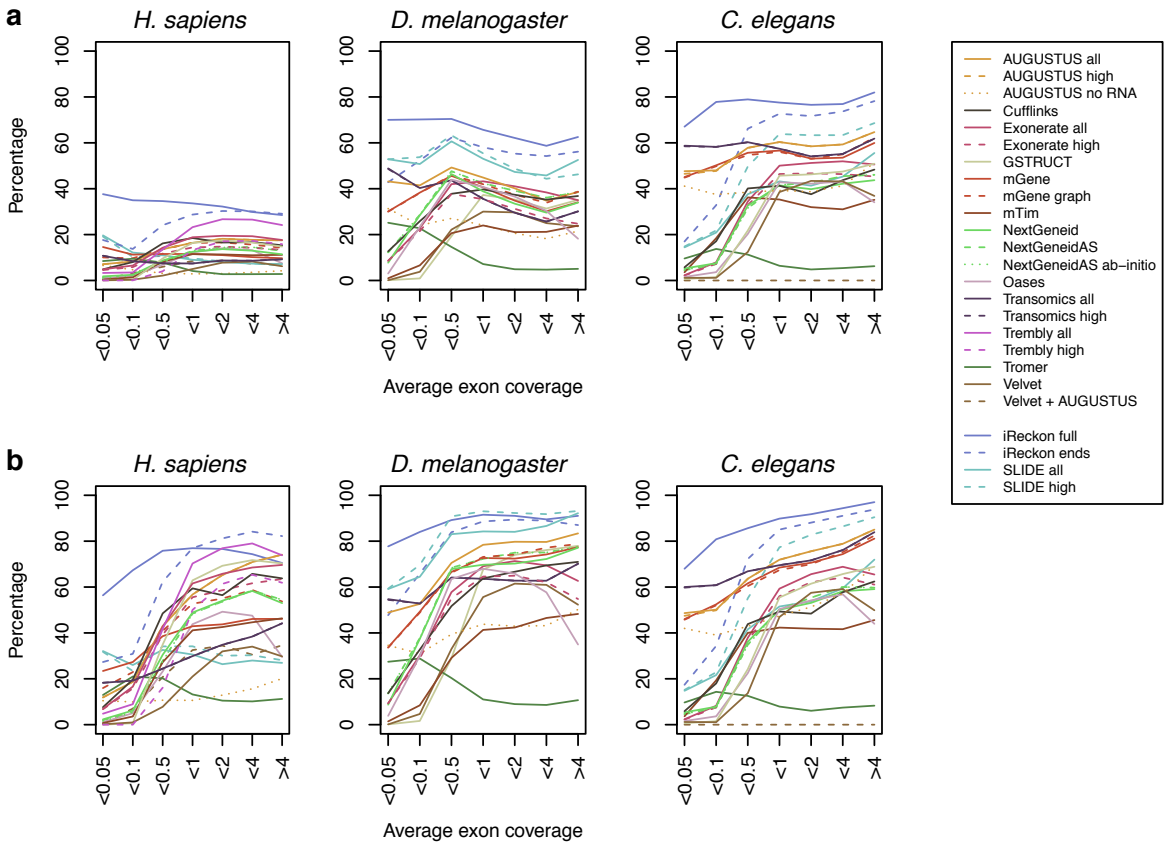


Figure 3.11: Influence of coverage on (a) transcript- and (b) gene level performance. Annotated transcripts and genes were binned according to RNA-seq read coverage and method sensitivity was calculated for each bin separately.

melanogaster and *C. elegans*, and $> 35\%$ in *H. sapiens* (Figure 3.11b). Detection rates were even lower for non-coding RNAs (data not shown). Pseudogenes were reported with similar frequency to protein-coding genes by AUGUSTUS, mGene, NextGeneid and Transomics, as pseudogenes retain partially intact coding sequences that can be identified by these methods (data not shown).

One explanation for the dramatic differences between species at the transcript level is the tendency of methods to assign one splice product per gene (Table 3.1). Whereas it is rare for genes in *C. elegans* and *D. melanogaster* to give rise to more than one or two transcript isoforms, human genes are annotated with an average of three to four, and it is unclear how many are simultaneously expressed. Assigning a single transcript

model per gene may therefore impede the detection of multiple isoforms expressed in a given sample.

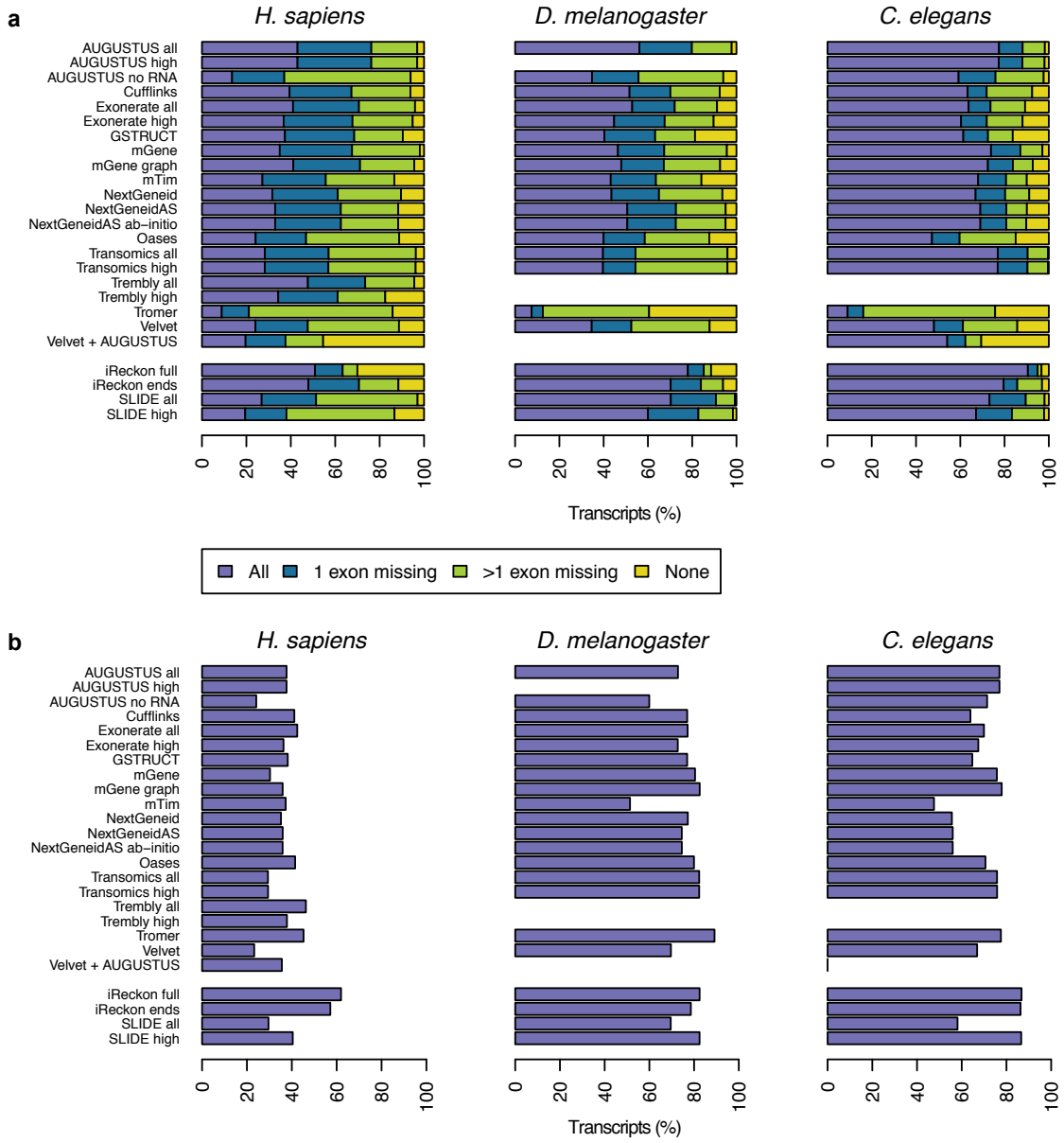


Figure 3.12: Transcript assembly performance. (a) Transcripts for which various subsets of constituent exons have been reported. (b) Percentage of transcripts, for which all exons have been identified, that were correctly assembled to a full-length annotated splice variant.

To identify the limiting factors in this process, for each method the number of known transcripts for which 1) all exons were identified, 2) exactly one exon was missing, 3) more than one exon was missing and 4) no exons were detected at all, was calculated (Figure 3.12a). The results clearly show that missing exons severely compromise transcript identification. For a significant fraction of transcripts not all exons are identified, ranging from 30% in *C. elegans* to greater than 60% in *H. sapiens*. Interestingly, while Trembly did not perform as well as AUGUSTUS, mGene, and Transomics at the exon level, this method reported the highest number of transcripts for which all exons were identified from *H. sapiens* data. In contrast, AUGUSTUS, mGene and Transomics identified at least one exon for most transcripts. The remaining methods failed to identify any exons for nearly 10% of all transcripts.

The topology of transcript structures was examined to determine how well each method was able to link detected exons into complete isoforms. Even in cases where all exons of an annotated transcript had been identified, the full isoform was often not assembled (Figure 3.12b). For *C. elegans* and *D. melanogaster* most methods were able to reconstruct 60% of the transcripts from the RNA-seq data. However, from the *H. sapiens* data less than 40% of known transcripts were assembled. Tromer stands out as an exception: the program identified all exons for relatively few genes, but once accounted for these were frequently linked into annotated transcript structures. Further inspection showed that these tended to be short isoforms comprising two to three exons on average, and thus represent a more tractable subset of the transcriptome. Provision of transcript start and end sites gives iReckon an advantage for the more complex human transcriptome, as evidenced by increased accuracy in assembling partial transcripts. In contrast, SLIDE consults exon coordinates, but ignores their connectivity, performing at a level similar to methods without any prior transcript-level information.

3.2.5 Agreement between methods

Transcript isoforms predicted by different methods often differed substantially as for example for RPF2 in Figure 3.13a, and only few examples where consistent isoforms were predicted by all methods were found (exemplarily shown for COX5B in Figure 3.13b). Therefore the agreement between methods was determined on the intron and isoform level.

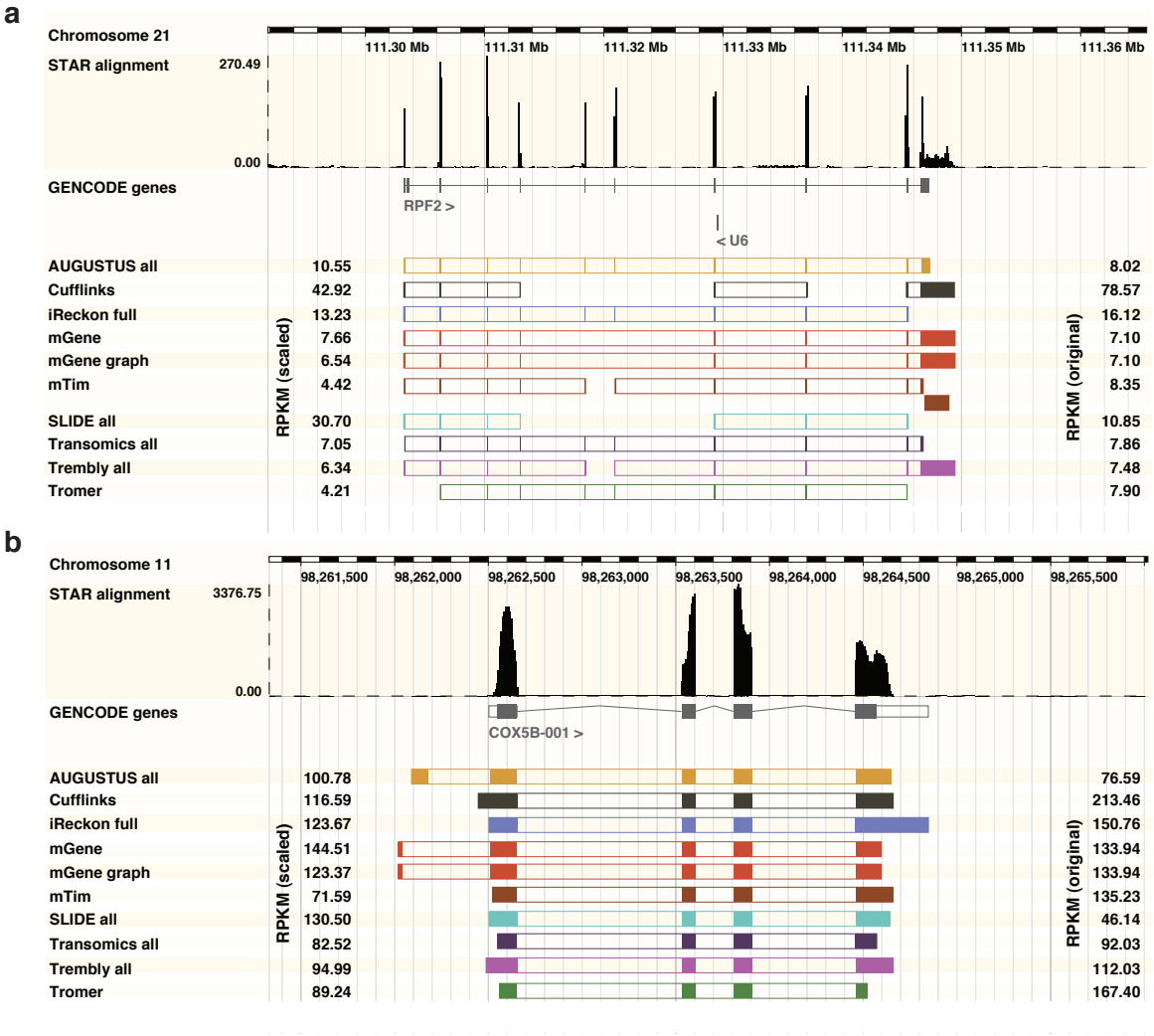


Figure 3.13: Examples of assembled and quantified transcripts. The upper tracks show RNA-seq read coverage (from STAR alignments; see Methods) and annotated genes. Exon predictions from the 10 methods that quantified transcripts are illustrated below the annotated gene by colored boxes. Exons predicted to belong to the same transcript isoform are connected. Original and median-scaled RPKM values are presented to the right and left, respectively, of the transcript models. (a) For RPF2, all methods reported different isoforms and expression levels. (b) For the gene COX5B, all methods reported similar transcript isoforms, but RPKM values differ substantially. Where multiple overlapping isoforms were reported, that with the higher RPKM was chosen for visualization, and spliced isoforms were prioritized over unspliced ones.

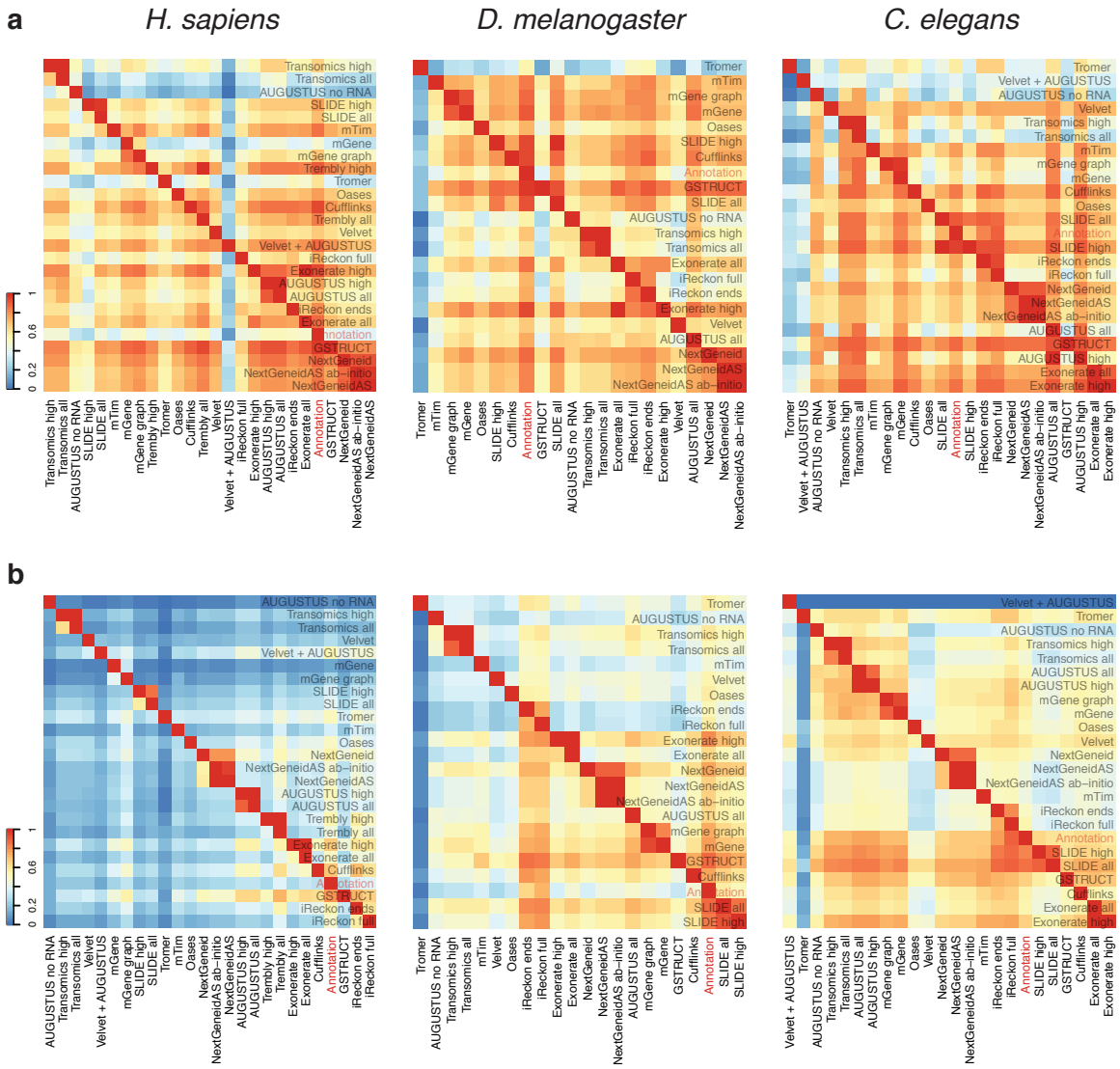


Figure 3.14: Intron and transcript consistency between methods. (a) Heatmaps illustrate pairwise intron agreement between methods. The agreement score $A_i[i,j]$ denotes the fraction of introns predicted by method i that were also predicted by method j. Methods were ordered by hierarchical clustering using $1-(A_i[i,j]+A_i[j,i])/2$ as the distance metric. (b) Heatmaps illustrate pairwise isoform agreement between methods. The agreement score $A_j[i,j]$ denotes the fraction of isoforms predicted by method i that are consistent with isoforms predicted by method j. Methods were ordered by hierarchical clustering using $1-(A_j[i,j]+A_j[j,i])/2$ as the distance metric.

First the intron agreement $A_j[i,j]$ between methods i and j was defined as the fraction of introns reported by method i that was also reported by method j. The pairwise agreement was used as input for hierarchical clustering, illustrated as heatmaps in Figure 3.14a. The level of agreement between methods was similar for all three organisms. Tromer showed little agreement with other methods on the two model organisms. For some methods there are notable differences between their column and row in the heatmap. A method that predicts only very few introns, but these are usually consistent with other methods, will show good agreement across its row but little agreement across its column as most introns predicted by other methods will be missed.

Similarly the isoform agreement $A_i[i,j]$ between methods i and j was defined as the fraction of isoforms predicted by method i that were equal or a subsequence of at least one isoform reported by method j. Generally, the level of agreement was much lower for isoforms compared to intron agreement. The agreement between methods was similar for the model organisms (Figure 3.14b), but isoforms predicted for human showed little agreement between methods (median agreement 25%) suggesting that methods are predicting different subparts of transcripts as in the example of RPF2 in Figure 3.13b.

3.2.6 Quantification of expression levels from RNA-seq data

A core feature of transcript reconstruction software is the estimation of expression levels from transcribed genes. These are given as digital read counts normalized by transcript length and sequencing depth (reads per kilobase of exon model per million mapped reads, RPKM) (Mortazavi et al., 2008). RPKM values were reported at the transcript level from a subset of methods. A range of expression level distributions was evident (Figure 3.15a). Generally, expression values show strong agreement between AUGUSTUS, Trembly, iReckon and mGene for all three data sets (Figures B.1- B.3). The greatest variation arises from gene loci where divergent or incomplete transcript models have been computed (Figure 3.13a). However, expression level estimates can vary considerably even where concordant transcript structures are reported as shown exemplarily in Figure 3.13b. Such differences were apparent also after scaling the RPKM distributions to equalize medians (Figure 3.13a and b).

To establish independent expression level quantification a set of human genes was assayed using the NanoString nCounter amplification-free detection system (Kulkarni,

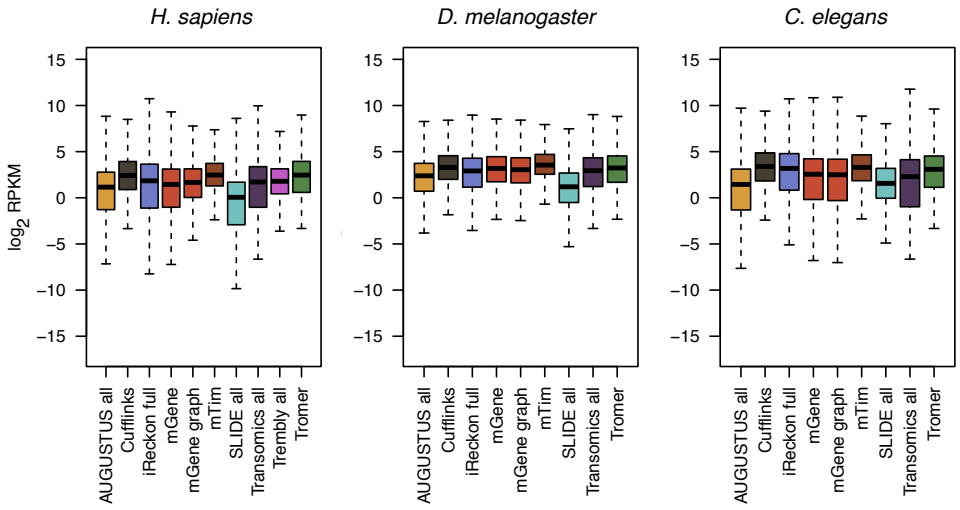


Figure 3.15: Comparison of transcript quantification. Distribution of gene expression values (RPKM) for each method. Results are shown for annotated genes only. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method.

2011). The NanoString probe sequences can be found in Table B.6. Correlation between NanoString counts and RNA-seq RPKMs ranged from 0.34 for Transomics to 0.68 for Cufflinks, iReckon and Trembly (Figure 3.16). NanoString counts showed high correlation with the number of reads mapped to the corresponding exon or junction using STAR or TopHat2 (Figure B.4). Many methods failed to report numerous exons or junctions targeted by probes that were expressed according to NanoString counts. Read support at those loci was typically sparse, with 19 probes having no corresponding alignments from the RNA-seq data (Figure B.4). These were, however, represented by low NanoString counts, indicating that the nCounter assay exhibits higher sensitivity for low-abundance transcripts than RNA-seq. For 10 of the unsupported NanoString probes, consistent isoforms were still reported by either AUGUSTUS, iReckon, mGene, SLIDE or Transomics. Thus, although the expression levels of these genes reflect the lower limits of detection for both technologies, sequencing reads dispersed over the gene body can allow for adequate transcript identification where ab initio methods or gene annotation were applied.

In general, all methods displayed higher identification rates for exons and junctions with higher NanoString counts (Figure B.5), and reliable detection from RNA-seq data

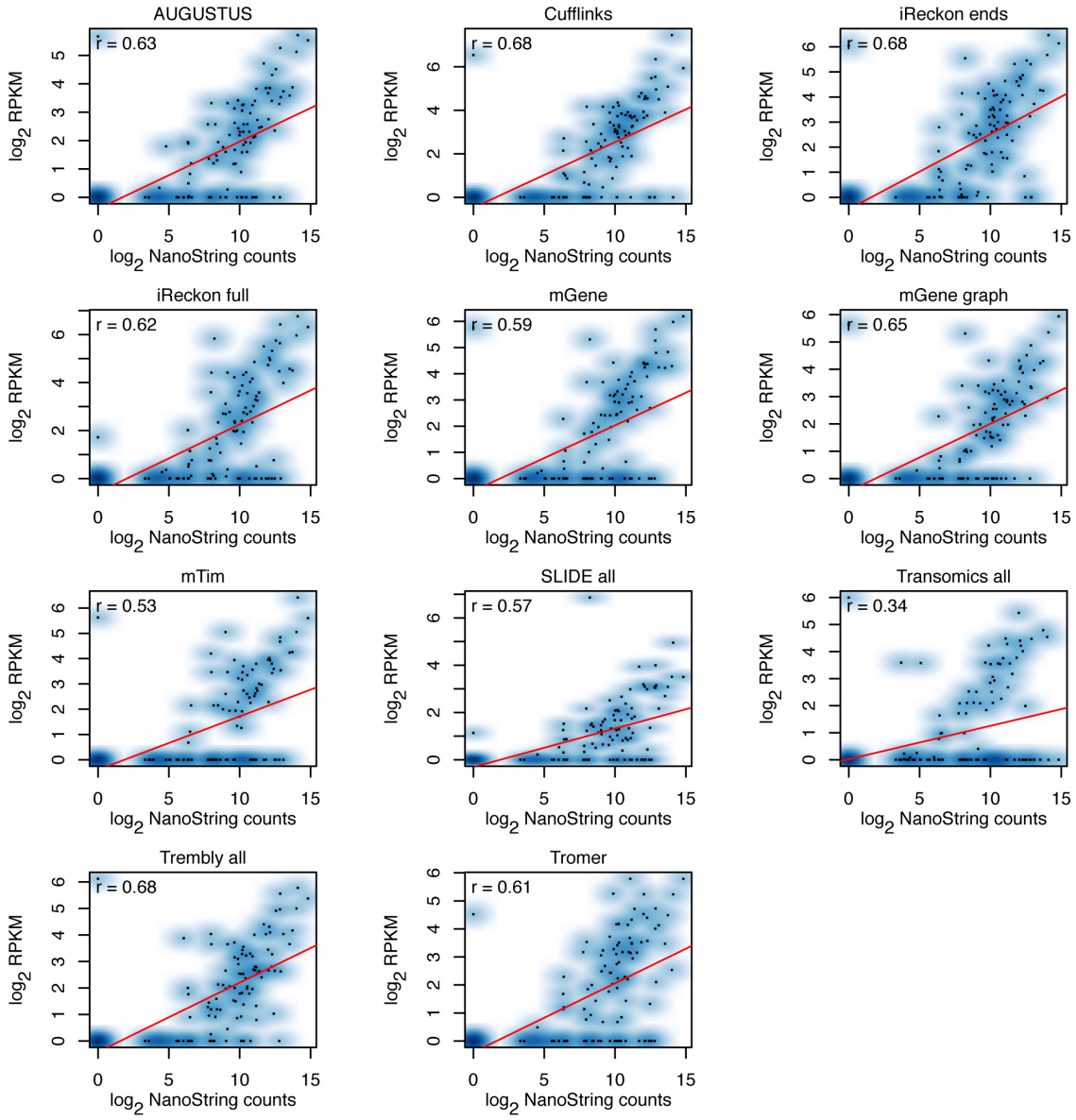


Figure 3.16: Correlation between NanoString counts and transcript RPKMs. Scatter plots show individual data points in black, with color intensity indicating the density of data points. Predicted transcript were required to contain the exon or junction targeted by the NanoString probe. Where multiple transcripts were reported for the same gene, the highest RPKM value was used. Where no such transcript was reported, an RPKM of zero was assigned. Correlation coefficients (Pearson r) are given for each comparison. Expression values were incremented by 1 prior to log transformation to avoid infinite numbers.

is dependent on read depth. Nonetheless, each failed to report a subset of exons and junctions despite the availability of adequate RNA-seq alignments (Figure B.5b). Comparing NanoString counts with RPKM values of the predominant isoform reported for each gene (irrespective of whether the targeted exon or junction was identified) improved correlation for most methods, and significantly for mTim and Transomics (Figure B.6).

3.3 Conclusion

With the advent of high-throughput RNA sequencing a variety of approaches for the assembly of short reads into complete transcripts have been proposed. The performance of 25 transcript reconstruction protocols from 14 developer teams was evaluated, in realistic RNA-seq applications against high-quality reference genomes. The assessments spanned a range of increasingly stringent metrics, from nucleotide-level correspondence with basic transcript components to the assembly of full-length, multi-exonic splice variants.

AUGUSTUS, mGene and Transomics performed well at nucleotide- and exon-level detection for protein-coding genes. These methods predict transcript features from translation and splicing signals in the genomic sequence, then report the model best supported by RNA-seq data. Combining *ab initio* prediction with experimental data provides an advantage in detecting genes expressed at low abundance, or from samples with low sequencing coverage. Even so, the benefits of this approach were most evident for *C. elegans* and lessened with increased transcriptome complexity.

The more challenging task of isoform reconstruction was reflected by lower performance in transcript-level evaluations. These results underscore the difficulty of transcript assembly, which relies on two outcomes: all exons comprising a given transcript must be identified, and these must then be connected to form the correct isoform structure. For most transcripts, automated methods failed to identify all constituent exons, and in cases where all exons were reported the protocols tested often failed to assemble them into complete transcript structures. Whereas methods using *ab initio* prediction retain an advantage in detecting individual exons, others performed better at linking them together. For example, Trembly detected fewer exons than AUGUSTUS, mGene and Transomics, but identified the greatest number of valid transcript isoforms in *H. sapiens*. GSTRUCT did not reach the same detection rate as Trembly, but featured the highest

overall precision of all methods. In *H. sapiens* and *C. elegans*, 57% and 60% of transcripts reported by GSTRUCT matched known isoforms and 76% in *D. melanogaster*.

AUGUSTUS, GSTRUCT, mGene, Transomics, and Trembly thus outperformed other methods evaluated here, but no single protocol excelled at all metrics. Comparing the performance of AUGUSTUS with and without RNA-seq data as input revealed that using experimental evidence only slightly improved exon-level detection, but increased transcript-level precision. Transomics featured enhanced precision for high-abundance transcripts, but expression level differences had little impact on detection sensitivity. Precision was a consistent strength of GSTRUCT. Tailoring the analysis protocol to suit more general attributes of a particular species can also be an important criterion. For example, the default threshold for minimum intron size impaired performance of the original Cufflinks submission on the *C. elegans* data significantly.

Accurate identification of single-exon transcripts proved difficult for many of the methods tested. In these cases the detection of discrete transcription products may benefit from sequencing protocols that preserve strand orientation, such that exon coverage and read depth better reflect the expression of contiguous RNAs. For spliced transcripts, the correct strand can typically be inferred from the genomic sequence at exon-intron boundaries.

Significant variation was observed in the range of expression level estimates reported for transcripts arising from the same gene loci. This was exacerbated by non-uniform exon detection and linkage between methods, but was also apparent when similar or identical transcript structures were reported. Thus, it may be unreliable to directly compare gene-based RPKM values from sample data processed independently with different software tools. RNA-seq data to be compared from disparate sources should be treated in an identical manner from the initial processing steps. Where this is not possible, care should be taken to ensure that similar gene models have been identified, and RPKM distributions should be inspected before applying expression level thresholds in downstream analyses.

The potential for non-coding RNA discovery and characterization is a significant advantage of RNA-seq. However this remains a challenging area for automated analysis methods. Performance is often impaired by lower expression levels of non-coding transcripts relative to many protein-coding genes, coupled with the inherent lack of translational features at the sequence level. The presence of open reading frames and

translation start/stop signals allowed some methods to identify protein-coding transcripts even at very low expression levels, whereas the detection of non-coding RNAs at high confidence required much greater read depth. Sequencing coverage thus appears to be crucial for accurate non-coding RNA profiling.

Two of the methods evaluated here, Velvet and Oases, carry out direct transcriptome assembly rather than aligning sequencing reads to a reference genome, and can therefore be applied to RNA-seq data from species where a genome build is not available. Velvet was originally developed for genome assembly from short read data. Oases is an adaption of that software, taking into account the unique challenges posed by transcriptome assembly from RNA-seq data, such as uneven coverage across transcript exons due to amplification bias or where multiple isoforms are expressed. As expected, both methods reached similar performance at the nucleotide and exon levels. At the transcript level, however, Oases achieved higher detection rates for the more complex transcriptomes of *D. melanogaster* and *H. sapiens*. Although performance is generally enhanced by genome alignment strategies, studies on organisms with a reference genome assembly of lower quality may benefit from a combined approach exploiting both genome alignments and *de novo* strategies. An ongoing project focused entirely on *de novo* transcriptome assembly methods is the DREAM6 challenge (`www.the-dream-project.org`).

The methods evaluated here can be applied to a range of analysis strategies, largely dependent on the state of the reference genome assembly and associated gene annotation for the target species. To improve the accuracy of existing annotation using RNA-seq, both Cufflinks and iReckon consult known gene structures during the transcript assembly process and may be useful in refining the coordinates of exon and transcript boundaries. Where a finished genome and high-quality annotation are available, Cufflinks and rQuant (part of the mGene protocol) can be applied solely for transcript quantification, which can further be improved by correcting for fragment bias. Gene prediction algorithms such as AUGUSTUS and mGene can be used to automate the annotation of novel genomes, whereas RNA-seq experiments based on partial or low-quality genome builds can be approached with a de novo assembler like Oases. This last application is expected to receive increasingly wider attention with the continued sequencing of new genomes.

RNA-seq offers the potential for precise refinement of existing gene annotation through the discovery of novel exons and junction sites. However, unannotated transcript isoforms assembled from RNA-seq data should be interpreted with care, and those critical

to an experimental study subjected to independent validation. The expression of multiple transcript isoforms and novel splice variants presents a major obstacle to accurate transcriptome reconstruction. Both exon identification and novel RNA discovery can improve with increased read depth, but the benefits of additional sampling to transcript assembly are inherently limited by the library construction requirements of current high-throughput sequencing instruments. Ultimately, the evolution of RNA-seq will move toward the single-pass determination of intact transcripts. Third-generation instruments will realize that potential and inspire new computing approaches to meet the next wave of innovation in transcriptome analysis.

3.4 Methods

3.4.1 Gene expression data

The *H. sapiens* RNA-seq data corresponds to the ENCODE (Harrow et al., 2012) data set HepG2 whole cell long polyA+ RNA CALTECH replicate 2, available from http://www.encodeproject.org. The *D. melanogaster* data set comprised a total of five sequencing runs from the modENCODE project (Graveley et al., 2011) for three L3 stage larval samples, and can be obtained from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRR023546, SRR023608, SRR023505, SRR027108 and SRR026433. The *C. elegans* data has previously been described (Mortazavi et al., 2010) and is available under accesssion SRR065719. All three data sets consist of 75-76 bp paired-end reads sequenced on the Illumina Genome Analyzer II. All of the data used in this study have been consolidated as a single experimental record in the ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress) under accession E-MTAB-1730.

3.4.2 Establishment of reference genome annotations

As not all genes are expressed in the samples used in the study, benchmarking methods against the entire set of annotated genes would underestimate transcript detection sensitivity. Therefore, the genome annotations (*H. sapiens*: GENCODE (Harrow et al., 2012)/Ensembl version 70, *D. melanogaster*: FB2013_01, *C. elegans*: WS200) were preprocessed to include only exons and transcripts with sufficient support in the RNA-seq data. Reads were mapped to the reference genomes using STAR version 2.2.0c, an independent RNA-seq aligner that is not a component in any of the evaluated transcript assembly methods (Dobin et al., 2013). To improve spliced alignment, STAR was provided with exon junction coordinates from the reference annotations. Default alignment parameters were used for the human data. For D. melanogaster and C. elegans, the intron size limit was reduced to 100000 and 15000 respectively (using options `--alignIntronMax` and `--alignMatesGapMax`). For each annotated exon, the number of mapped read pairs per base was computed and exons with a value below 0.01 were excluded from further analysis. Only transcripts for which all exons satisfied this crite-

rion were included in transcript-level assessments. Changing the underlying aligner did not change the results (Figure 3.17).

3.4.3 Data processing for Cufflinks, iReckon and SLIDE

RNA-seq reads were aligned with TopHat version 2.0.3 using parameters suited to each species. The genomes of *D. melanogaster* and *C. elegans* contain a high percentage of small introns; examining their size distributions led us to set the parameters `-i`, `--min-coverage-intron` and `--min-segment-intron` to 30 for *C. elegans*, 40 for *D. melanogaster* and 50 for *H. sapiens*.

Cufflinks was re-run with default settings except for the `--min-intron-length` parameter which was set to 30 for *C. elegans*, 40 for *D. melanogaster* and 50 for *H. sapiens*, consistent with the TopHat alignments. To maintain the greatest compatibility with submitted results that were computed without annotation, iReckon was run with the minimum annotation requirements, that is start and end sites of all annotated transcripts (not filtered by read coverage). SLIDE was run in discovery mode and provided with the full unfiltered annotation for each genome.

3.4.4 Evaluation of prediction sets

Feature predictions were evaluated against the filtered reference annotation sets at four structural levels: nucleotide, exon, transcript and gene. The nucleotide-level metrics measure the ability of methods to identify exonic regions, ignoring the strand and exact boundaries of features. Nucleotide-level precision was computed as the number of genomic base pairs within both annotated and predicted exons, divided by the number of genomic base pairs within predicted exons. Similarly, nucleotide sensitivity was computed as the number of genomic base pairs shared between annotated and predicted exons, divided by the number of genomic base pairs within annotated exons.

The exon-level metrics measure the ability of the different algorithms to identify the correct strand and boundaries of exons. Precision was calculated as the percentage of reported exons with an annotated counterpart, and sensitivity denotes the percentage of annotated exons that were correctly assembled. Annotated exons were classified as first, internal, terminal and those comprising unspliced transcripts (single exons). Unless stated otherwise, a flexible evaluation mode was employed for first, terminal

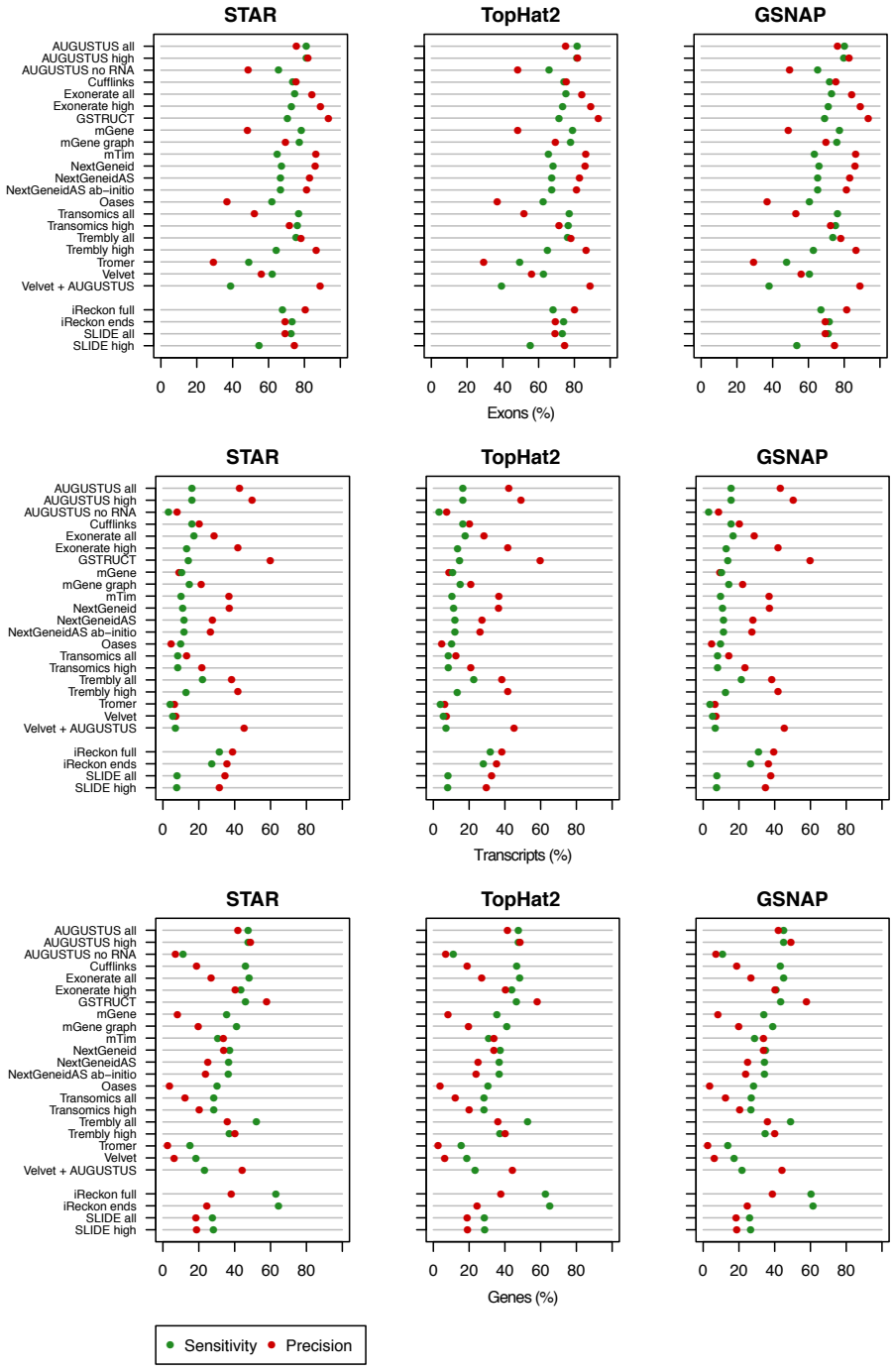


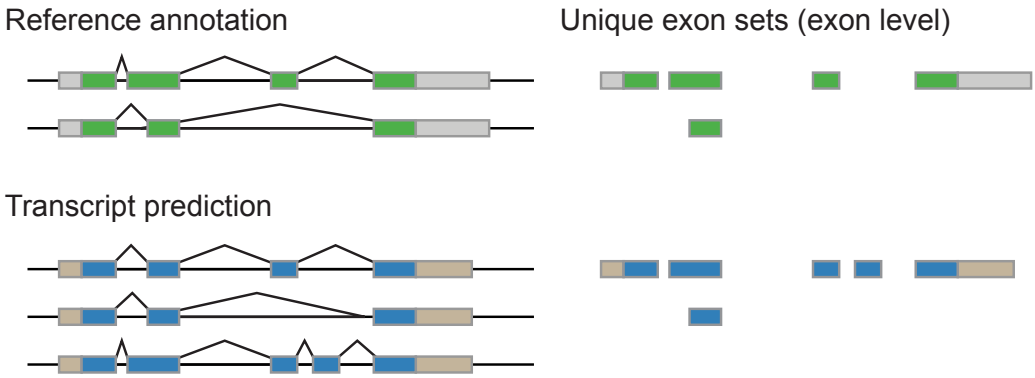
Figure 3.17: Influence of aligner choice on the assessment. Exon-, transcript-, and gene level performance based on filtered annotation derived from either a STAR, TopHat2 or GSNAP alignment.

and single exons. Specifically, first and terminal exons were required to have correctly predicted internal borders only, and exons constituting unspliced transcripts were scored as correct if covered to at least 60% by a predicted transcript. Exons shared between different transcript isoforms were counted once. For comparison, certain analyses were also carried out using a fixed evaluation mode, where annotated and predicted exons were required to match exactly.

Transcript-level precision was computed as as the percentage of reported spliced transcripts matching an annotated transcript, and sensitivity as the percentage of annotated spliced transcripts with a counterpart in the transcript reconstruction output. Consistent with the flexible evaluation mode for exons (see above), transcript start and end sites were allowed to differ between reference and prediction, but splice sites were required to match exactly. Genes were scored as correctly predicted if at least one annotated transcript isoform in a given gene locus was correct.

3.4.5 Evaluation of transcript quantification

In order to compare transcript quantification results among methods, for each annotated gene the corresponding predominant transcript reported by each method was identified. The predominant transcript was defined as the transcript with the highest reported RPKM value, among those isoforms intersecting annotated exons of the gene. A subset of human transcripts was quantified independently by NanoString assays in Ali Mortazavi's lab, Biological Sciences III, University of California Irvine, Irvine, USA. Genes of at least 1 kb in length, for which annotated exon-intron structures have been manually curated, and having at least two transcripts satisfying these criteria were selected. A total of 109 genes were targeted by 141 distinct probes, designed against specific exons or splice junctions. NanoString counts were compared to the highest RPKM value reported for transcript isoforms consistent with the probe design (correlation r_c) or for any isoform from the locus (correlation r_a). Pearsons r was calculated based on the log-transformed NanoString counts and RNA-seq RPKM values. Expression values were incremented by 1 prior to log2 transformation to avoid infinite numbers.



Level		fixed mode		flexible mode	
	Feature	Sensitivity	Precision	Sensitivity	Precision
Exon	CDS	1.00	0.83	NA	NA
	Exon	0.80	0.67	1.00	0.83
Transcript	CDS	0.50	0.33	NA	NA
	Exon	0.00	0.00	0.50	0.33
Gene	CDS	1.00	1.00	NA	NA
	Exon	0.00	0.00	1.00	1.00

Figure 3.18: Structural validation strategy. Transcript models were validated against annotated isoforms. For exon-level evaluation, transcripts were collapsed into unique exon sets, that is exons shared between transcript isoforms are counted once. Sensitivity (a.k.a. sensitivity) was calculated as the proportion of reference features (exons, transcripts, or genes) matched by a reported feature. Precision was calculated as the proportion of reported features matching a reference feature. In this study primarily the flexible evaluation strategy, where exact agreement between transcript boundaries was not required, was used. For comparison, certain analyses were also carried out using a fixed evaluation mode, where annotated and predicted exons were required to match exactly.

Chapter 4

Exploring the Stat3-dependent Transcriptome in ES Cells

4.1 Introduction

As described in the introduction, embryonic stem (ES) cells are derived from the pre-implantation mouse embryo (Evans and Kaufman, 1981; Martin, 1981; Nichols et al., 1990; Brook and Gardner, 1997) and can be maintained in culture indefinitely when grown in the right conditions. The LIF-dependent activation of Stat3 is known to play an important role in the promotion of ES cell self-renewal and ES cell derivation (Niwa et al., 1998). It also is a limiting factor of somatic cell reprogramming (Yang et al., 2010; Tang et al., 2012). The fact that EpiSCs and primate ES cells are cultured in the absence of LIF supports the notion that Stat3 responsiveness is a feature of naive pluripotency (Dahéron et al., 2004; Sumi et al., 2004; Humphrey et al., 2004). Therefore, understanding LIF-dependent Stat3 signalling is crucial for a full understanding of ground state pluripotency. However, little is known about the mechanism of Stat3 dependent induction of gene expression. The identification of novel Stat3 targets (both coding and non-coding) and the analysis of interactions between Stat3 and the pluripotency network are the main focus of this chapter.

4.1.1 Stat3 signalling in ES cells

Despite the important role of Stat3 in mouse stem cell maintenance and induction of pluripotency, little is know about its downstream effectors. The core pluripotency factors Oct4, Sox2, Nanog or Esrrb are not directly regulated by LIF. Although Stat3 binds to the promoter region of Oct4, Sox2 and Nanog as well as to an intronic region of Esrrb the expression of these factors is not induced upon LIF exposure. Several efforts to identify direct targets of Stat3 have identified transcription factors Klf4, Pim1 and Gbx2 (Li et al., 2005; Niwa et al., 2009; Hall et al., 2009; Tai and Ying, 2013). However, none of these factors are indispensable for LIF-responsiveness of ES cells and their forced expression cannot substitute for LIF signalling. Notably, Klf4 is one of the canonical Yamanaka reprogramming factors (Takahashi and Yamanaka, 2006) found to enable reprogramming of somatic cells, but as discussed in the introduction LIF is required in addition to the four factors (Tang et al., 2012). Since none of the known Stat3 targets can substitute for LIF signalling, the downstream effect of Stat3 is accomplished either via additive effects of multiple target genes or via a crucial unidentified target.

Several attempts have been made to identify Stat3 target genes involved in mouse ES cell self-renewal. Sekkai et al. used microarrays to identify transcriptional differences between ES cells cultured in LIF and serum supplemented media and ES cells where Stat3 activity has been shut down, either through LIF deprivation or expression of a tetracycline regulable dominant-negative Stat3F mutant (Sekkaï et al., 2005). Changes in gene expression were analysed at 16, 24 and 48 hours of LIF withdrawal in the presence of tetracycline to suppress Stat3F expression. The same time points were analysed to identify variations in gene expression induced by Stat3F expression. Genes commonly downregulated in both experiments and thus potential Stat3 targets were Socs3, Inhbb and Aes. Stat3 binding to the Aes promoter was confirmed by ChIP experiments. No functional analysis of Aes or other potential targets was performed. Notably, known Stat3 targets as Junb, Klf4 and Klf5 were not identified in this study.

Bourillot et al. made an attempt to identify genome-wide Stat3 targets using E14-S3ER cells expressing a hormone-dependent STAT3-ER from an episomal vector (Bourillot et al., 2009). These cells can be maintained in serum/LIF activating endogenous Stat3 or in serum and 4′-hydroxytamoxifen (4′-OHT) activating the ligand-dependent STAT3-ER. E14/T ES cells expressing hormone-dependent Cre-ER recombinase (E14-CER)

were used as a control. These cells are strictly dependent on LIF. In order to detect Stat3 downstream targets both E14-S3ER ES and E14-CER cells were deprived of LIF and 4′-OHT for 24 hours. E14-S3ER cells were then stimulated by either LIF for one hour or 4′-OHT for two hours. Additionally E14-CER were stimulated by 4′-OHT for two hours. Genes that were both activated in E14-S3ER ES cells stimulated by LIF and 4′-OHT and not in E14-CER cells upon 4′-OHT stimulation, were selected for further investigation. Selection for fold-change > 1.3 and excluding targets that required de-novo protein synthesis (blocked by cycloheximide treatment) resulted in 58 Stat3 target genes. Knock-down of 23 of these target genes increased the frequency of differentiated colonies in self-renewal assays, indicating that they are involved in ES cell maintenance.

All studies described above are based on mouse ES cells cultured in media supplemented with serum and LIF (serum/LIF). Target genes are identified after LIF withdrawal which causes mouse ES cells to differentiate. As discussed in the introduction, media supplemented with two selective small molecule inhibitors (2i) of Gsk3 and Mek kinases eliminate mouse ES cell differentiation and can sustain self-renewal in the absence of LIF (Ying et al., 2008; Wray et al., 2010). This allows the derivation and expansion of $Stat3^{-/-}$ ES cells. Comparing the transcriptome of Stat3 positive and negative cells in the absence of LIF as well as after LIF exposure enables exploration of the Stat3-dependent transcriptome.

Stat3 plays an important role in various cell types. Additional to its role as maintainer of pluripotency, Stat3 induces acute-phase response in hepatoma cells, stimulates proliferation in B lymphocytes and activates terminal differentiation and growth arrest in monocytes (see Levy and Lee, 2002 for a review). Depending on the cellular context Stat3 induces distinct sets of target genes. Therefore, it is desirable to elucidate the mechanisms how Stat3 regulates gene expression of its target genes. For several cell types Stat3 is known to cooperate with other factors to regulate gene expression. In T-cells, for example, Stat3 cooperates with IRF4 transcription factors to induce gene expression. Also in AtT-20 cells, a cancer cell line, Stat3 regulates gene expression together with glucocorticoid receptor (GR). In ES cells Stat3 has been proposed to control gene expression in a c-myc dependant manner (Kidder et al., 2008). However, the overlap of Stat3 and c-myc is quite small and more recent studies show a more prominent overlap between binding sites of Stat3 and other pluripotency factors.

4.1.2 Outline

In order to explore the Stat3-dependent transcriptome in ES cells, gene expression was analysed in the presence and absence of LIF by RNA sequencing. In order to distinguish between Stat3 targets and other LIF downstream effects, ES cells devoid of Stat3 were used as a control. Information on Stat3 binding sites from in-house ChIP-seq experiments was combined with publicly available data to study the cooperation of Stat3 with other transcription factors in ES cells. The experimental part of the work was performed by Graziano Martello and Jason Wray from Austin Smith's laboratory at the Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, Cambridge UK. Library preparation was performed by Paul Bertone, European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge UK, for the RNA-seq experiment and by Maike Paramor, Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, Cambridge UK, for the ChIP-seq experiments. Parts of the analysis have been used and acknowledged in Martello et al. (Martello et al., 2013). Supplementary information is available in Appendix C.

4.2 Results

RNA was sequenced for wild-type ES ($Stat3^{+/+}$) cells as well as Stat3 null ($Stat3^{-/-}$) cells in three different conditions each (Figure 4.1a): cultured in 2i only, cultured in 2i and one hour LIF, cultured in 2i and 24 hours LIF. For each cell line and condition combination two biological replicates were sequenced. Each library yielded roughly 36M reads which were aligned to the mouse genome using GSNAP (Wu and Nacu, 2010). Gene expression showed high correlation between biological replicates (Figure C.1). ChIP-seq was performed for total Stat3 and phosphorylated Stat3 in $Stat3^{+/+}$ cells exposed to LIF for one hour.

4.2.1 Using $Stat3^{-/-}$ ES cells to identify genuine Stat3 targets

As described above and depicted in Figure 4.1b, LIF activates more than the Jak/Stat3 pathway. $Stat3^{-/-}$ ES cells, which can be derived and expanded in 2i media, enable the identification of LIF downstream targets that are independent of Stat3 (Figure 4.1c). A set of 3964 genes was identified as differentially expressed between $Stat3^{-/-}$ and $Stat3^{+/+}$

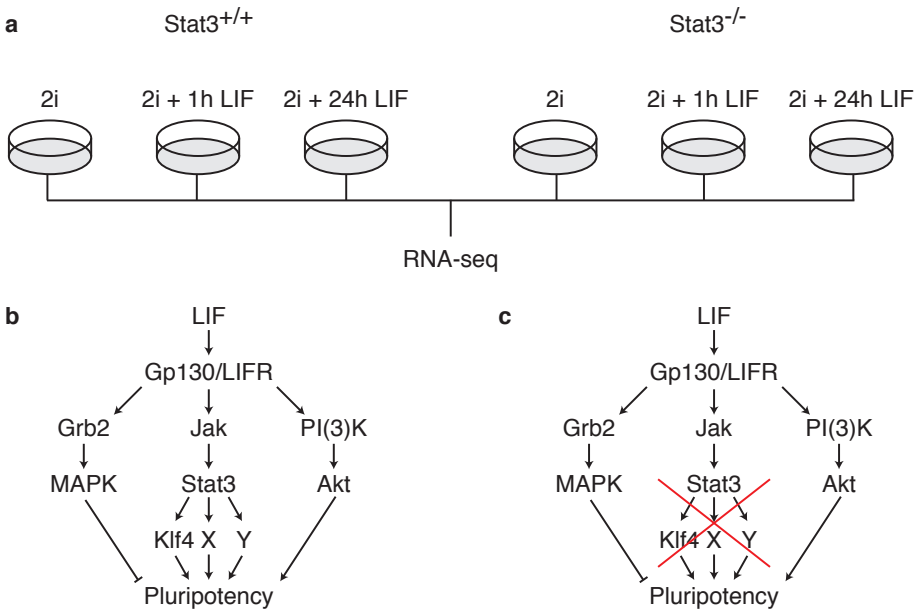


Figure 4.1: Experimental setup and pathways induced by LIF in $Stat3^{+/+}$ and $^{-/-}$ cells. (a) RNA-seq data was derived for $Stat3^{+/+}$ and $Stat3^{-/-}$ cells for cultured in 2i, 2i + 1h LIF and 2i + 24h LIF. Two biological replicates were generated for each cell line and condition combination. (b) Pathways induced by LIF in wild-type ES cells. (c) Pathways induced by LIF in $Stat3^{-/-}$ cells. Genes whose expression is induced by Stat3 can be identified by subtracting the set of upregulated genes in $Stat3^{-/-}$ cells from the set of upregulated genes in $Stat3^{+/+}$ cells.

Table 4.1: Enriched GO terms. The top ten enriched GO terms for genes differentially expressed between Stat3$^{-/-}$ and Stat3$^{+/+}$.

GO term	***P*-value**
Translation	$1.96E-13$
Transcription	$2.68E-10$
RNA processing	$4.50E-10$
Cell cycle	$6.26E-10$
Regulation of transcription	$5.42E-7$
Cell cycle process	$4.23E-6$
Cell cycle phase	$4.34E-5$
Mitotic cell cycle	$4.17E-5$
Chromosome organization	$1.88E-4$
ncRNA processing	$1.70E-4$

ES cells, with roughly half of them being up- and half of them being downregulated. GO term analysis identified biological processes that were enriched among the differentially expressed genes. Four out of the top ten enriched GO terms were related to the cell cycle (Table 4.1). Indeed the Stat3$^{-/-}$ ES cells showed a slightly lower proliferation rate than Stat3$^{+/+}$ ES cells (Graziano Martello, personal communication). Examination of expression levels of genes associated with either pluripotency or germ layer specification showed no major differences between Stat3$^{-/-}$ and Stat3$^{+/+}$ cells maintained in 2i (Figure 4.2).

Stat3$^{-/-}$ cells do not exhibit any overt sign of spontaneous differentiation or appreciable cell death and can form high-contribution chimeric embryos upon blastocyst injection (Martello et al., 2013). Colony formation efficiency is unchanged compared to wild-type ES cells. Deletion of Stat3 therefore does not impair ES cell self-renewal efficiency in 2i. However, these cells are strictly dependant on 2i and are not able to self-renew in other culture conditions since they are not responsive to LIF. This indicates that activation of Stat3 cannot be substituted by other LIF-induced signals.

Besides JAK/Stat3, LIF activates PI3 kinase and Erk signalling as well (Burdon et al., 1999). LIF enhances self-renewal efficiency in all ES cell culture conditions, including 2i (Wray et al., 2010). However, LIF shows no increase in colony-forming efficiency of Stat3$^{-/-}$ cells in 2i while efficiency for wild-type cells was increased (Martello et al., 2013). Clustering the RNA-seq samples based on the Euclidean distance between their gene counts showed that the gene expression profile of Stat3$^{-/-}$ cells is very similar in

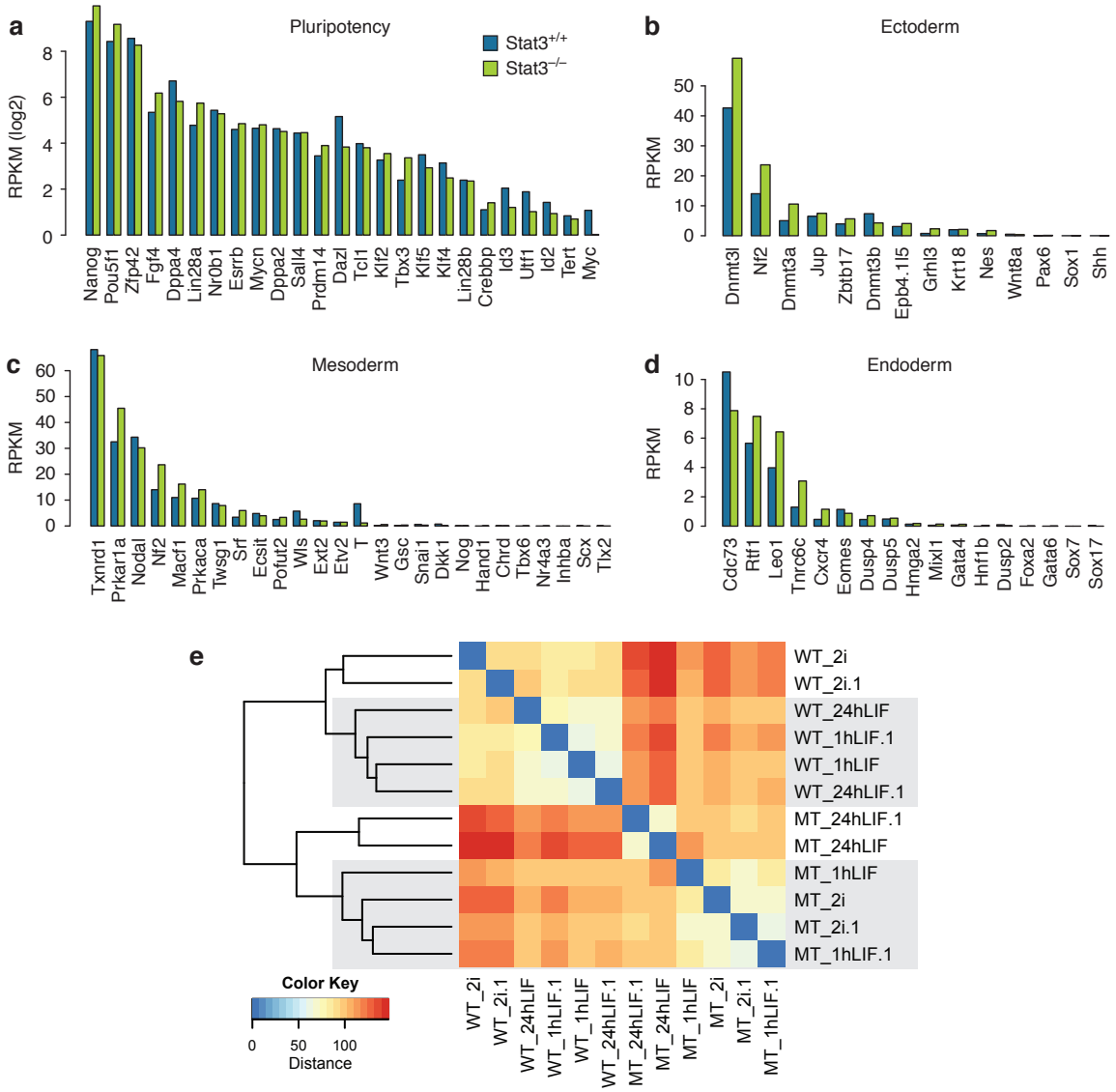


Figure 4.2: Expression levels of genes associated with pluripotency or lineage priming. (a) Log transformed RPKM values for pluripotency factors. (b) RPKM values for genes associated with ectoderm lineage priming. (c) RPKM values for genes associated with mesoderm development. (d) RPKM values for endoderm marker genes. (e) Hierarchical clustering of all samples based on gene counts.

Table 4.2: Number of differentially expressed genes. The table lists the number of up- and downregulated genes for five different comparisons. Complete lists of differentially expressed genes can be found in Appendix C.

Comparison		Upregulated	Downregulated
$Stat3^{-/-}$ in 2i	$Stat3^{+/+}$ in 2i	1737	2227
$Stat3^{-/-}$ in 2i	$Stat3^{-/-}$ in 2i + 1h LIF	43	2
$Stat3^{-/-}$ in 2i	$Stat3^{-/-}$ in 2i + 24h LIF	1196	1083
$Stat3^{+/+}$ in 2i	$Stat3^{+/+}$ in 2i + 1h LIF	188	360
$Stat3^{+/+}$ in 2i	$Stat3^{+/+}$ in 2i + 24h LIF	295	652

2i and after one hour LIF exposure (Figure 4.2e). In contrast, in $Stat3^{+/+}$ the samples exposed to LIF cluster together, indicating that one hour LIF exposure is enough to alter the gene expression profile in $Stat3^{+/+}$ cells. Together these findings verify the primary role of Stat3 mediating the LIF signal to ES cell self-renewal.

4.2.2 Response of wild-typ ES cells to LIF

Potential LIF targets were identified by analysing genes that were differentially expressed in $Stat3^{+/+}$ cells following LIF treatment, but not LIF-responsive in $Stat3^{-/-}$ cells. Both annotated and novel genes were analysed (Section 4.2.2.1 and Section 4.2.2.2 respectively). $Stat3^{+/+}$ cells responded within one hour to LIF treatment (Figure 4.2e) enabling the study of primary targets. However, since Stat3 binds to its target genes with different affinities, it is likely that not all direct targets respond within one hour. Therefore, a later time point was included, which also enabled the study of secondary targets.

4.2.2.1 Differential expression analysis

Differentially expressed genes were identified based on gene read counts using DESeq (see Methods). The pairwise comparisons and the resulting number of differentially expressed genes are shown in Table 4.2. After one hour of LIF exposure, 548 genes are differentially expressed of which 188 genes are upregulated (Figure 4.3 and Table 4.2). Socs3 (suppressor of cytokine signalling 3) showed the highest induction of gene expression after LIF exposure. Socs3 is a known Stat3 target and forms a negative feedback loop of the Jak/Stat3 pathway to regulate Stat3 activation (Boyle et al., 2009). The LIF-induced

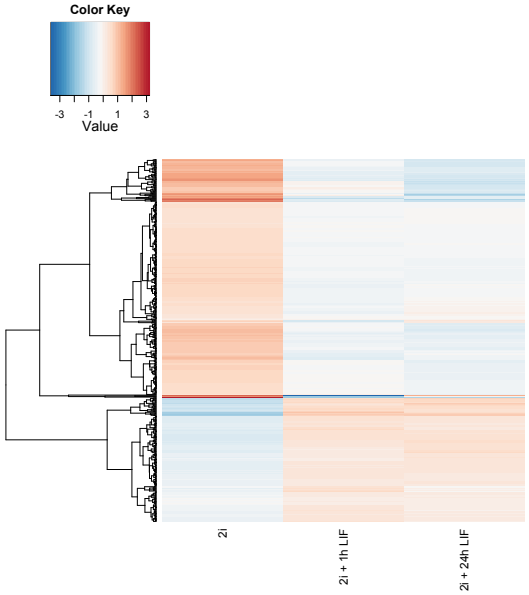


Figure 4.3: Differentially expressed genes in Stat3$^{+/+}$ cells. Displayed are mean-centred RPKM values of differentially expressed genes.

genes also include the known Stat3 targets Klf4, Pim1, Gbx2, Zfp36l1, Spry2 and Icam1 as well as Stat3 itself. Additionally, the known Stat3 targets Klf5, Sbno2, Zfp36 and Junb are upregulated after 24 hours of LIF stimulation. Stat3 is commonly regarded as a transcriptional activator (Boeuf et al., 1997; Darnell, 1997). Nonetheless, more genes were repressed than activated following LIF exposure (Figure 4.3 and Table 4.2). This could be either a direct effect of Stat3 or a secondary effect. The LIF-induced genes are discussed more in detail in Section 4.2.3, where Stat3 binding data is used to distinguish between direct and secondary effects and to study the integration of Stat3 within the pluripotency network.

4.2.2.2 Detection of novel transcripts

One main advantage of RNA-seq experiments is the detection of unannotated transcripts, including non-coding ones. Non-coding transcripts have been shown to play a role in ES cell maintenance although the mechanisms remains unknown for most of them (Loewer et al., 2010; Guttman et al., 2011; Livyatan et al., 2013). In order to identify novel LIF-dependent transcripts, Cufflinks (version 2.1.1) was used to reconstruct transcript isoforms guided by the Ensembl annotation. The Cuffcompare utility classifies assem-

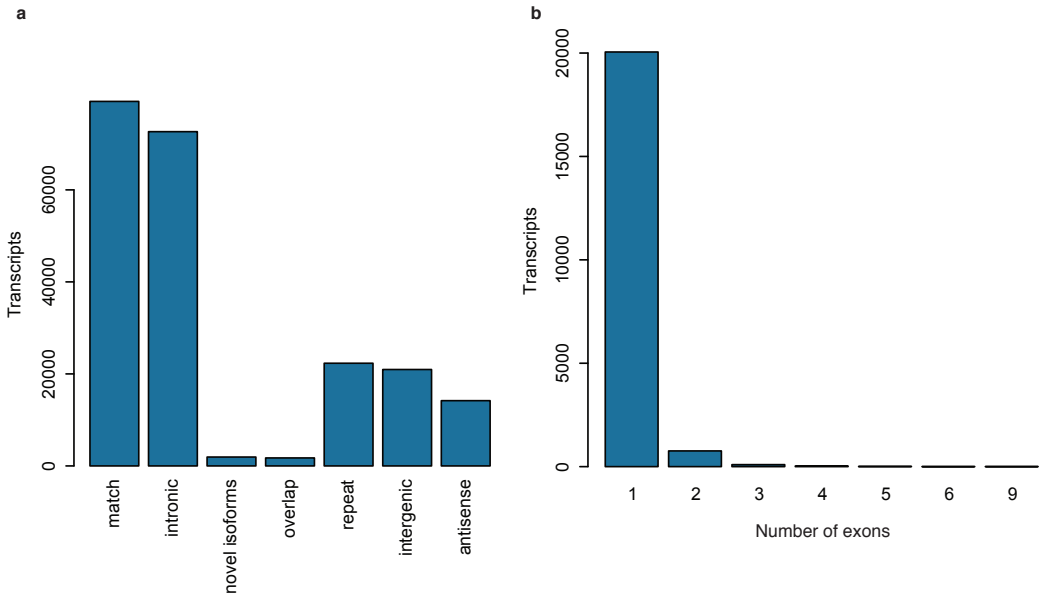


Figure 4.4: Classification of isoforms assembled by Cufflinks. (a) Classification of assembled transcripts in relation to annotated genes from Ensembl. (b) Number of exons of novel intergenic transcripts identified by Cufflinks.

bled transcript isoforms relative to annotated transcripts and was used to distinguish between known and novel transcripts (Figure 4.4a). The majority of novel isoforms predicted by Cufflinks were intronic which may be due to a higher level of pre-mRNAs in libraries derived from ribosomal depletion. About 20000 novel intergenic transcripts were identified. While most of these were single-transcript isoforms (95%), about 900 spliced isoforms were detected (Figure 4.4b).

To determine whether any of these novel intergenic transcripts responded to LIF, differential expression analysis was performed. A set of 10 novel transcripts was differentially expressed following LIF exposure, of which three were activated by LIF. None of the LIF-induced transcripts were spliced. However, visual inspection of the data revealed that several transcripts, whose expression was induced by LIF, were not identified by Cufflinks. To compensate for this, a simple sliding window approach was implemented to assembly mapped reads into novel transcripts (see Methods). These were classified according to their position relative to their closest annotated gene (Figure 4.5a).

Again intronic transcripts were most abundant (Figure 4.5b). To determine whether these might be due to the presence of pre-mRNAs in the sample, the relation between

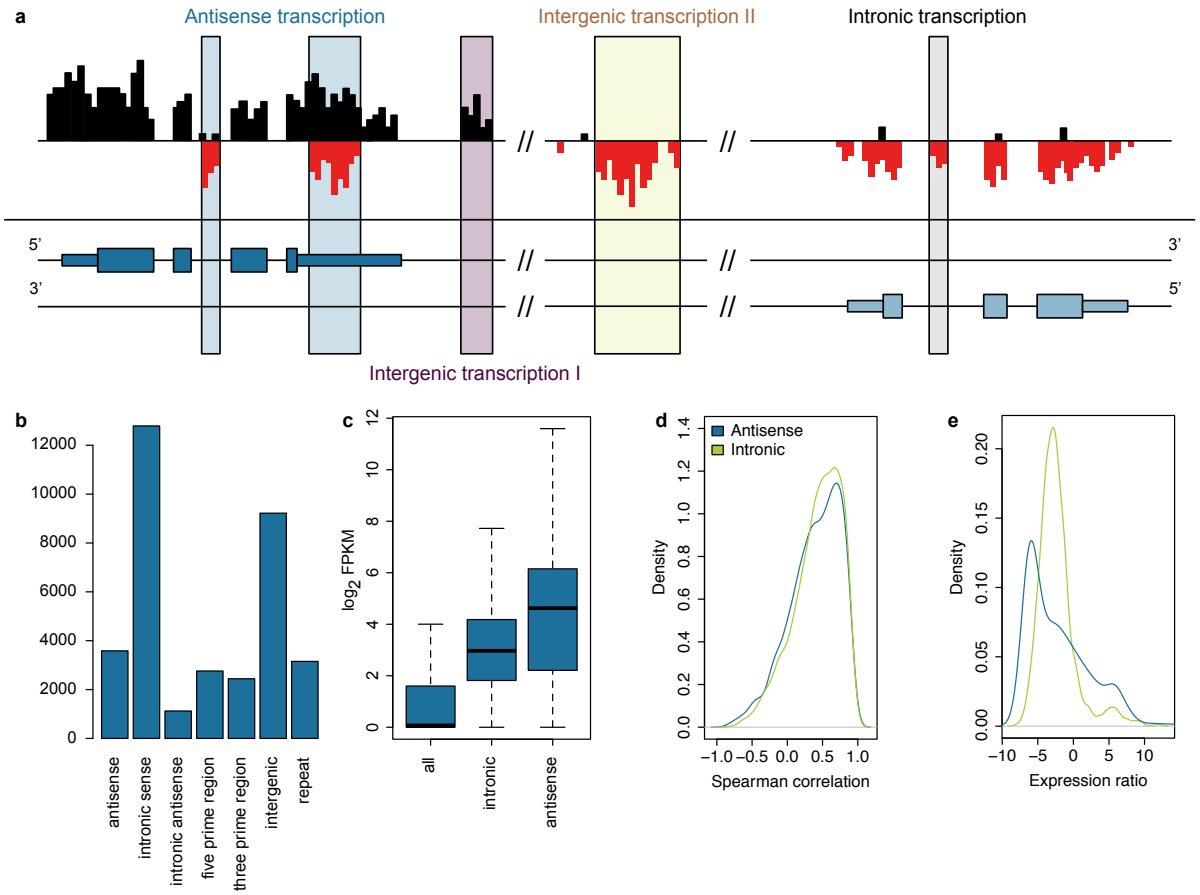


Figure 4.5: Novel region classification. (a) Regions were classified according to their relative position to annotated genes and whether they are marked as repeats: Antisense, intronic, intergenic and repeat. Antisense transcripts were further divided into exonic and intronic and intergenic transcripts within 50 kb of an annotated gene were classified as five and three prime transcripts. (b) Number of identified transcripts for each category. (c) FPKM distribution for all genes compared to genes associated with intronic and antisense transcripts. A value of one was added to the FPKM values prior to log transformation to avoid infinite values. (d) Spearman correlation between the expression of intronic and antisense transcripts and their associated annotated gene. (e) Expression ratio between antisense and intronic transcripts and their corresponding annotated genes.

the expression of intronic transcripts and their associated annotated transcripts was analysed. Genes associated with intronic transcription were on average expressed at a higher level (t-test P-value $< 1.0e-10$, see Methods, Figure 4.5c) and their expression was highly correlated with the expression of intronic transcripts (Figure 4.5c). To identify real intergenic transcripts, the expression ratio between the intronic transcript and the corresponding gene was determined. For 290 intronic transcripts, the unannotated transcript was expressed at least at 50% of the expression level of the associated gene and supported by 50 reads or more. Of these, 16 were activated by LIF. An example of an LIF-induced intronic transcript is shown in the Appendix in Figure C.2.

Using a strand-specific library preparation protocol enabled the analysis of antisense transcription. Antisense transcripts play an important role in controlling gene regulation (Katayama et al., 2005; Xu et al., 2011; Modarresi et al., 2012). Most antisense transcripts overlapped annotated exons (Figure 4.5b). Strand-specific RNA-seq methods are known to have a certain error rate (0.5-1% for the best methods such as the dUTP protocol used for this study, Levin et al., 2010) and therefore some of the observed exonic antisense transcripts may be due to this artefact. The expression of most antisense transcripts correlated with the expression of their associated gene (Figure 4.5d), which tended to be highly expressed (P-value $< 1.0e-10$, Figure 4.5c). This indicates that a large proportion of detected antisense transcription may be due to strand specificity not being entirely preserved during the library preparation. In order to identify real antisense transcripts, the ratio between the expression of the antisense transcript and of the associated annotated gene was determined. For the majority of antisense transcripts the expression level was around 3% of the expression level of the associated gene and thus likely due to a small error rate of the strand-specific protocol (Figure 4.5e). Nevertheless, 249 antisense transcripts were supported by at least 50 reads and were expressed at a level of at least 50% of the expression level of the associated gene. The expression of 18 of these antisense transcripts was upregulated upon LIF exposure. However, none of the associated annotated genes were differentially expressed after LIF induction.

Differential expression analysis for the intergenic transcripts identified 368 differential expressed transcripts of which 174 were upregulated in Stat3$^{+/+}$ cells following one hour of LIF stimulation. The three novel transcripts identified by Cufflinks were also identified in this analysis. For six of the novel transcripts, qPCR was used to validate that their expression was upregulated following LIF exposure. All six transcripts were detected

using qPCR, but only the expression of five of them was upregulated with a fold change > 1.3 (Figure 4.6). The fold changes observed using qPCR were generally lower than fold changes derived from the RNA-seq data, most likely due to the wider range of expression levels detectable by RNA-seq.

Correlation of expression levels between the twelve samples was used to identify novel regions that potentially belong to the same spliced transcript (see Methods). In particular, for lowly expressed transcripts which may not have support of spliced reads, expression level correlation can provide some indication whether some transcribed regions are expressed together. A total of 2239 potentially spliced transcripts were assembled this way. One of these was selected for validation by qPCR. For this purpose primers were designed to target the predicted the splice sites. All three junctions were confirmed by qPCR and also the whole spliced transcript could be confirmed by qPCR (Figure 4.6). Notably, Cufflinks identified only the first two exons of this transcript. This shows that expression correlation between samples can be exploited for transcript assembly. To my knowledge, none of the current available transcript assemblers make use of this kind of information. None of these potentially spliced transcripts were LIF-responsive and were therefore not analysed further.

These results indicate that Stat3 induces the expression of unannotated transcripts. Knock-down experiments using shRNAs may determine whether these transcripts have an effect on the self-renewal capacity of ES cells.

4.2.3 Integration of Stat3 binding data

To gain functional insights into the role of Stat3 in the maintenance of ES cells, information about Stat3 binding sites is essential. Stat3 binding profiles were generated using ChIP-seq experiments in mouse ES cells cultured in 2i/LIF supplemented media (see Methods). Two experiments were performed targeting both total Stat3 and phosphorylated Stat3 (pStat3). A total of 1523 binding sites were identified for Stat3 and 1026 for pStat3 (see Methods). More than 550 sites overlapped between the two data sets. For the following analyses binding sites from both experiments were combined.

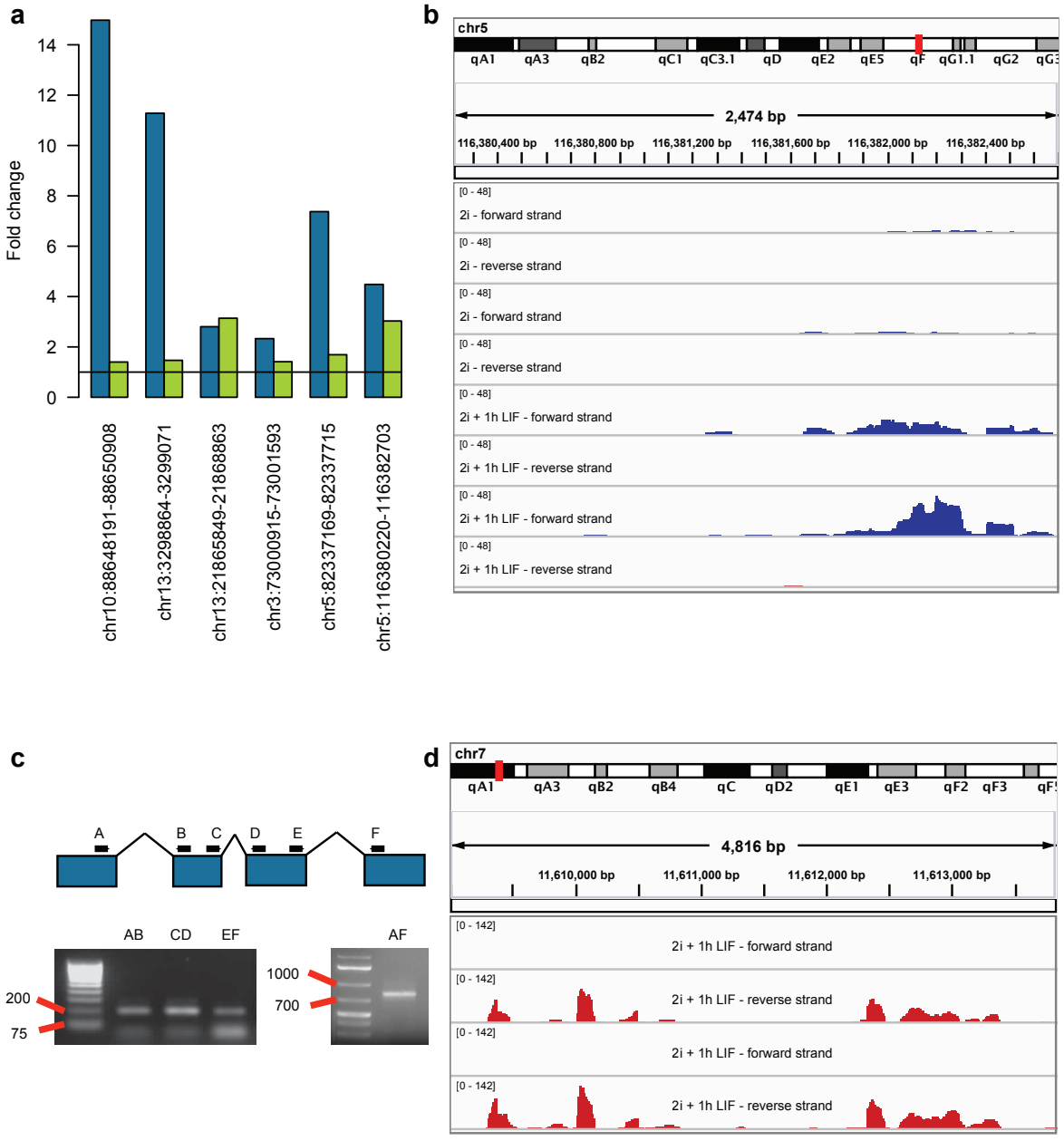


Figure 4.6: Novel region validation. (a) Fold-changes observed by RNA-seq (blue) and qPCR (green) for six unannotated transcripts. (b) Genome brower image showing expression of the unannotated transcript at locus chr5:116380220-116382703 after LIF exposure. (c) Validation of 3 predicted junctions and the full transcript using different sets of primers. A qPCR product was present for each tested junction. Amplifying the whole transcript resulted in a PCR product with a length between 700 and 1000 nt (expected length was around 800 nt). (d) Genome browser image showing the expression of the unannotated multi-exonic transcript.

4.2.3.1 Identification of Stat3 targets involved in ES cell self-renewal

Direct Stat3 downstream targets must be distinguished from Stat3-independent LIF targets, since LIF activates other pathways besides the JAK/Stat3 pathway. Exploiting $Stat3^{-/-}$ cells enables the identification of genes that are directly regulated by Stat3 activation rather than other signals downstream of the LIF receptor. Both $Stat3^{-/-}$ and $Stat3^{+/+}$ ES cells were exposed to LIF for either 1 hour or 24 hours. The first time point of LIF stimulation is expected to enrich for primary transcriptional targets. Differential expression analysis identified 188 genes upregulated in $Stat3^{+/+}$ cells (see Methods). 43 genes were activated by LIF in $Stat3^{-/-}$ cells. The expression of Irf9, which is involved in immune response, was induced both in $Stat3^{-/-}$ and $Stat3^{+/+}$ cells suggesting that it is a Stat3 independent target of LIF.

To further enrich for direct Stat3 target genes, Stat3 binding sites were annotated with known genes and compared to the set of LIF-induced genes (see Methods, for a list of Stat3 bound genes see Appendix C). This resulted in 1851 Stat3 bound genes, representing about 8% of all annotated genes. A high proportion of LIF-responsive genes in $Stat3^{+/+}$ cells was bound by Stat3 (25.5%, Pearson's χ^2 test P-value $\leq 1.0e-10$, see Methods, Figure 4.7a). For genes down-regulated after LIF stimulation, only 5.6% were bound by Stat3, which could be expected by chance (P-value = 0.08, Figure 4.7a). These results are consistent with the common notion of Stat3 as a transcriptional activator (Boeuf et al., 1997; Darnell, 1997). Only 2.3% of the LIF-induced genes in $Stat3^{-/-}$ cells were bound by Stat3 (P-value = 0.26).

After filtering out genes induced in $Stat3^{-/-}$ cells and selecting genes with Stat3 binding sites, LIF-responsive genes in $Stat3^{+/+}$ cells were further narrowed down by selecting genes involved in the regulation of transcription, such as transcription factors or chromatin modifiers, using GO term annotation (Figure 4.7b, filtering for genes associated with the following GO terms: GO:0005667, transcription factor complex; GO:0003700, sequence-specific DNA binding transcription factor activity; GO:0008134, transcription factor binding; or GO:0006355, regulation of transcription, DNA-dependent). This strategy resulted in a list of nine genes: Bcl3, Erf, Gbx2, Irf1, Klf4, Stat3, Tfap2c, Tfeb and Tfcp2l1 (Figure 4.7c). Except for Irf1 all of these genes were downregulated upon Stat3 knock-down indicating that their expression is dependent on Stat3 (Nishiyama et al., 2013).

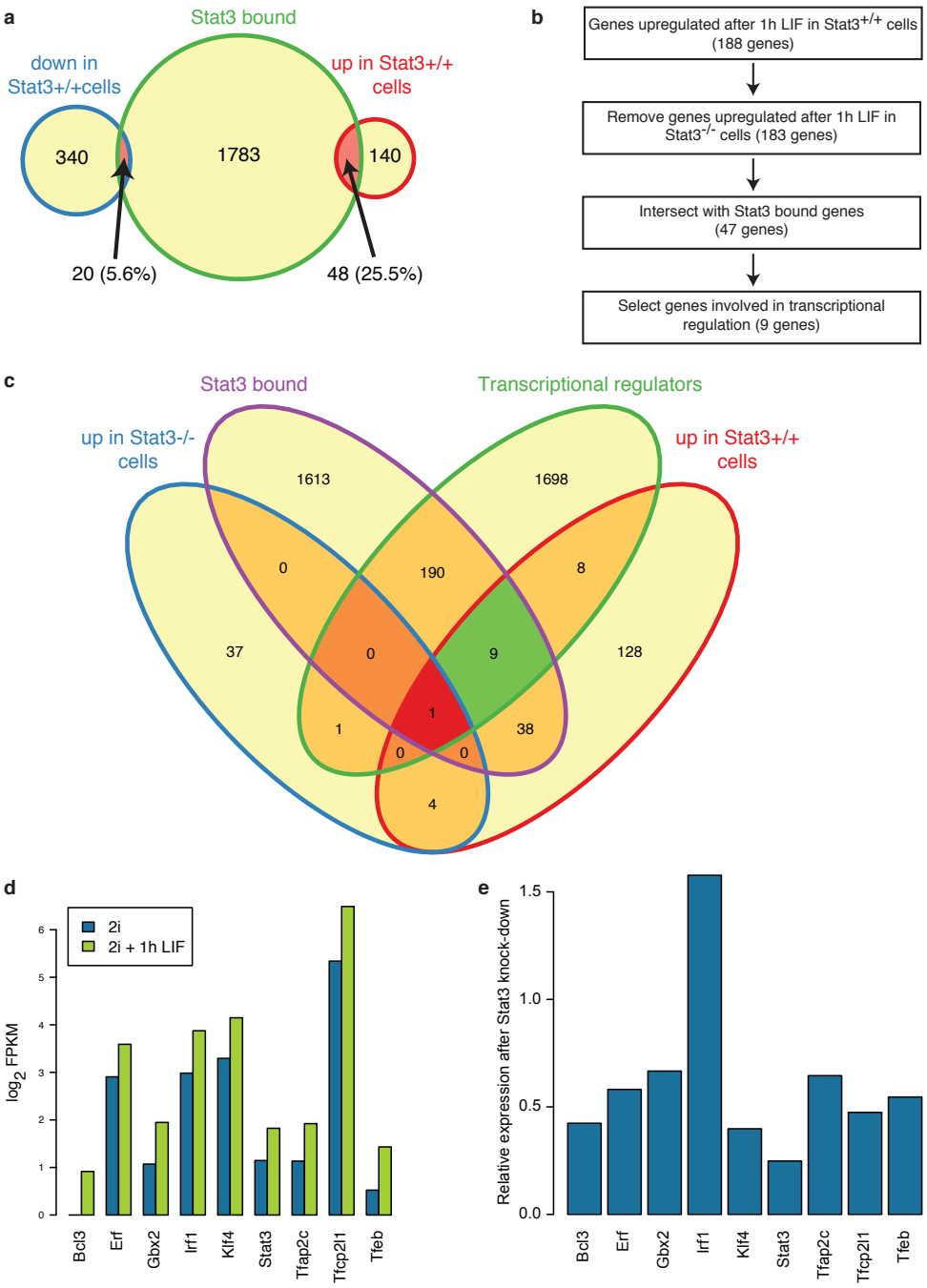


Figure 4.7: Ranking Stat3 targets. (a) Differentially expressed genes bound by Stat3. (b) Strategy to rank Stat3 target genes to find targets involved in ES cell self-renewal. (c) This strategy leads to the selection of 9 candidate target genes (green area in Venn diagram). (d) Log-transformed FPKM values for the selected candidate genes. FPKM values were incremented by 1 prior to log transformation to avoid infinite values. (e) Relative expression of candidate genes after Stat3 knock-down (data from Nishiyama et al., 2013).

Among the candidate genes are Stat3 itself and its known downstream targets Gbx2 and Klf4 (Bourillot et al., 2009). Interestingly, however, the most highly expressed transcription factor in this group, Tfcp2l1 (also known as Crtr-1) has not previously been linked to LIF signalling (Figure 4.7d). Tfcp2l1 induction was unaffected when cells were stimulated with LIF in the presence of the protein synthesis inhibitor cycloheximide (Martello et al., 2013). This confirmed Tfcp2l1 as a direct response to LIF stimulation rather than a secondary effect requiring protein synthesis.

Forced expression of Gbx2 allows long-term expansion of mouse ES cells in the absence of LIF and can also reprogram EpiSCs into mouse ES cells (Tai and Ying, 2013; Martello et al., 2013). Similarly, Klf4 can maintain mouse ES cells in a pluripotent state in the absence of LIF (Martello et al., 2012). However, for both factors a high level of spontaneous differentiation in the absence of LIF can be observed indicating that they are not able to fully substitute for LIF signalling. Also their knock-down does not affect LIF signalling, showing that they are dispensable for LIF signalling. In contrast to these results, Tfcp2l1 can maintain ES cells in the absence of LIF with a similar level of spontaneous differentiation as observed in serum/LIF conditions (Martello et al., 2013). Expression of Tfcp2l1 at endogenous levels can maintain mouse ES cells in combination with PD03. These cells can give rise to chimeric animals after transgene excision (Martello et al., 2013). These results indicate that Tfcp2l1 can phenocopy Stat3 signalling. Furthermore, Tfcp2l1 knock-down significantly reduced the number of undifferentiated colonies in clonogenicity assay and caused the down-regulation of the pluripotency factors Nanog, Oct4, Sox2, Esrrb and Tbx3 (Martello et al., 2013). These results suggest that Tfcp2l1 is not only able to substitute for LIF, but is required for LIF signalling.

Of the 174 novel transcripts identified in Section 4.2.2.2, 71 had a Stat3 binding site within 50 kb. Six novel transcripts were even overlapping with Stat3 binding sites. These results indicate that several of the novel transcripts identified in this analysis are indeed direct targets of Stat3.

4.2.3.2 Secondary Stat3 targets

Over the time course studied in this project, the expression of 327 genes was significantly increased upon LIF exposure. Of these LIF-induced genes 77 were bound by Stat3,

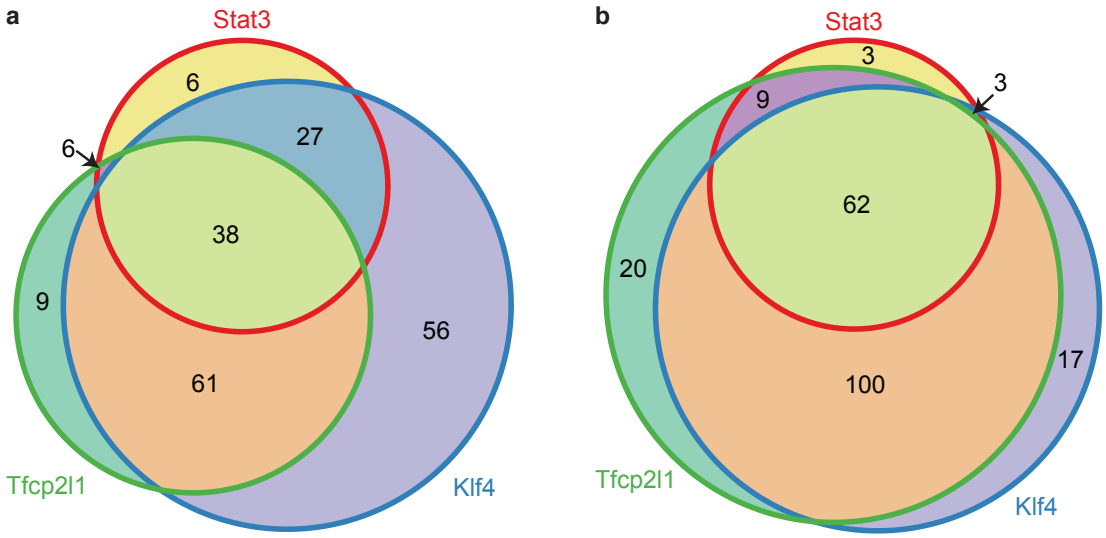


Figure 4.8: LIF-induced genes bound by Stat3, Klf4 or Tfcp2l1. (a) Shared target genes of Stat3, Tfcp2l1 and Klf4 based on inhouse Stat3 and Tfcp2l1 binding data and publicly available Klf4 binding data (Chen et al., 2008). (b) Shared target genes of Stat3, Tfcp2l1 and Klf4 based on inhouse Stat3 and publicly available Tfcp2l1 and Klf4 binding data (Chen et al., 2008).

suggesting that they are primary targets of Stat3, while the unbound genes are potential secondary targets. To determine whether the expression of the LIF-induced genes that are not bound by Stat3 might be activated by primary targets of Stat3, ChIP-seq for the transcriptional factors Klf4 and Tfcp2l1 was performed. These two transcription factors were confirmed Stat3 targets and had the most prominent effect on ES cell self-renewal. The Klf4 ChIP-seq data set showed little enrichment and was therefore replaced by the Klf4 data set from Chen et al. (Chen et al., 2008) for further analysis. The binding sites of Tfcp2l1 and Klf4 were annotated with known genes (for detailed lists of Klf4 and Tfcp2l1 bound genes see Appendix C).

Most LIF-responsive genes bound by Stat3 were also bound by Tfcp2l1 and Klf4. Furthermore, a large number of LIF-induced genes that were not bound by Stat3 were bound by Tfcp2l1 and Klf4, often by both. Including the Tfcp2l1 data set from Chen et al. (Chen et al., 2008) furthermore increased the number of genes bound by both Tfcp2l1 and Klf4, but not Stat3. This suggests that a large proportion of genes whose expression is upregulated following LIF exposure and that are not bound by Stat3 are secondary targets and regulated by Tfcp2l1 and Klf4 downstream of Stat3. These secondary targets include the genes Vegfa and Nkx6-2, which are both involved in cell differentiation, were

activated by LIF and bound by Klf4 and Tfcp2l1, but not Stat3, as was Dtx1, a negative regulator of neural differentiation. Furthermore, Leprot a known negative regulator of the Jak/Stat3 pathway was activated upon LIF exposure. The Wnt-targets Sfrp1 and Aes were also identified as secondary targets, indicating that LIF-signalling is feeding back into the canonical Wnt-pathway. Finally, Smarcd3, which is part of the chromatin-remodelling complex SNF/SWI, was found to be regulated by Klf4 and Tfcp2l1.

This is further supported by the observation that 107 of the 117 LIF-induced genes bound by Klf4, but not Stat3, are downregulated more than 2-fold following Klf4 knock-down (Nishiyama et al., 2013). Similarly, 111 of the 120 LIF-responsive genes bound by Tfcp2l1, but not Stat3, are downregulated more than 2-fold following Tfcp2l1 knock-down (Nishiyama et al., 2013).

4.2.3.3 Correlating Stat3 binding and expression changes

To gain insights into the mechanism of Stat3 dependent gene regulation, the influence of certain binding characteristics on gene expression was analysed, such as peak location, peak size and distance to associated genes.

Peaks were annotated with known genes (see Methods). Peaks were furthermore classified according to their relative location to their potential target genes: overlapping introns, coding exons, UTRs, promoters or downstream regions (3 kb upstream of the transcriptional start site and 3 kb downstream of the transcriptional termination site), or intergenic (Figure 4.9a). Stat3 binding was enriched at promoter regions accounting for around 20% of all binding sites despite promoter regions only representing 3% of the genomic sequence. Comparing binding sites with RNA-seq data of mouse ES cells in 2i/LIF showed that genes bound by Stat3 were generally expressed at higher levels than genes not bound (P-value $< 1.0e-10$, Figure 4.9b). Furthermore, genes bound in their promoter region were higher expressed than genes bound outside their promoter region (P-value $1.197e-06$). For genes transcriptionally upregulated upon LIF exposure Stat3 binding sites were enriched even more at promoter regions (Figure 4.9c and d). Genes whose expression was only induced after 24 hours of LIF exposure were not enriched in Stat3 binding (Figure 4.9d). This supports the idea that direct Stat3 targets respond within one hour and that genes responding later are mostly secondary targets. The expression of Stat3 bound genes was induced significantly more often than of unbound

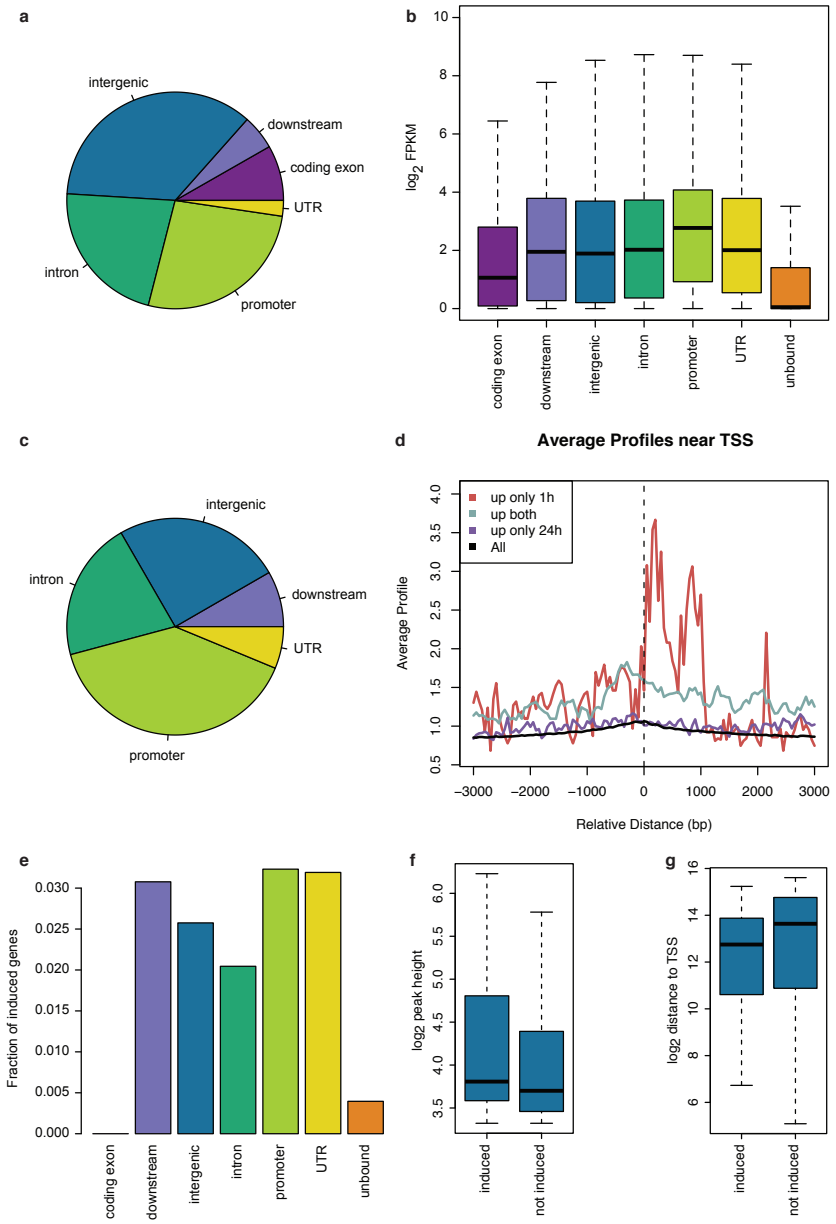


Figure 4.9: Location of Stat3 peaks and expression levels of associated genes. (a) Pie chart illustrating the location of Stat3 binding sites relative to their associated gene. (b) Log-transformed FPKM values for genes classified by Stat3 binding sites. FPKM values were log-transformed after adding one to avoid infinite values. (c) Pie chart illustrating the location of Stat3 binding sites associated with LIF-responsive genes. (d) Stat3 binding site density at promoters, classified by the expression changes of the corresponding genes upon LIF exposure. (e) Genes were classified according to Stat3 binding sites and the fraction of induced genes was calculated for each class separately. (f) Distribution of log-transformed peak heights for peaks associated with induced genes versus peaks associated with genes that do not respond to LIF signalling. (g) Distribution of log-transformed distances between peak and TSS for peaks associated with induced genes versus peaks associated with genes that are not activated by LIF.

genes (Figure 4.9e). LIF-induced genes were associated with significantly higher peaks than not induced genes (P-value 0.03888). However, the average distance to transcription start sites was not significantly different between induced and non-induced genes (P-value 0.1473).

4.2.3.4 Stat3 and the pluripotency network

As the previous results indicate, the expression of only around 3% of Stat3 bound genes was significantly upregulated upon LIF exposure. As Stat3 is known to collaborate with other factors to promote expression of its target genes in other cell types (Kwon et al., 2009), the binding profile of Stat3 was compared to binding data of other pluripotency factors and transcriptional regulators in ES cells (Martello et al., 2012). Clustering these factors according to shared binding sites showed that Stat3 was part of the cluster containing pluripotency factors Oct4, Sox2, Nanog, Klf4, Esrrb and Tfcp2l1 (Figure 4.10). Additionally, markers for active enhancer such as p300, Med12 and Med1 clustered together with Stat3 (Figure 4.10).

Around 50% of Stat3 binding positions were also bound by other pluripotency factors, such as Oct4, Sox2, Nanog, Esrrb, Klf4, Prdm14, Tcf3 and Tfcp2l1 (Figure 4.11a). For Stat3 peaks associated with LIF-induced genes, this fraction increased to almost 70% (Figure 4.11b). Considering associated genes rather than exact binding positions, only 5% of genes bound by Stat3 were bound by Stat3 only (Figure 4.11c). Of the 48 LIF-induced genes bound by Stat3, none were bound by Stat3 alone and 29 were bound by 7 or all 8 of the aforementioned pluripotency factors (Figure 4.11d). For example, the LIF-induced genes Klf4, Klf5 and Tfcp2l1 were bound by all eight factors (Figures C.4-C.6).

To get a better understanding of the cooperation between Stat3 and other factors, a distance matrix was calculated indicating for each Stat3 how far the next peak of a given factor was. The distance matrix was then used to cluster Stat3 peaks (Figure C.3). LIF-responsive genes were enriched in Stat3 peaks that were in close proximity to other pluripotency factors (Figure 4.11e).

Closer inspection of the distance between the Stat3 peaks and the peaks of the other factors showed that for upregulated genes the median distance between Stat3 and the pluripotency factors was significantly shorter (Figure 4.12). While the overall median

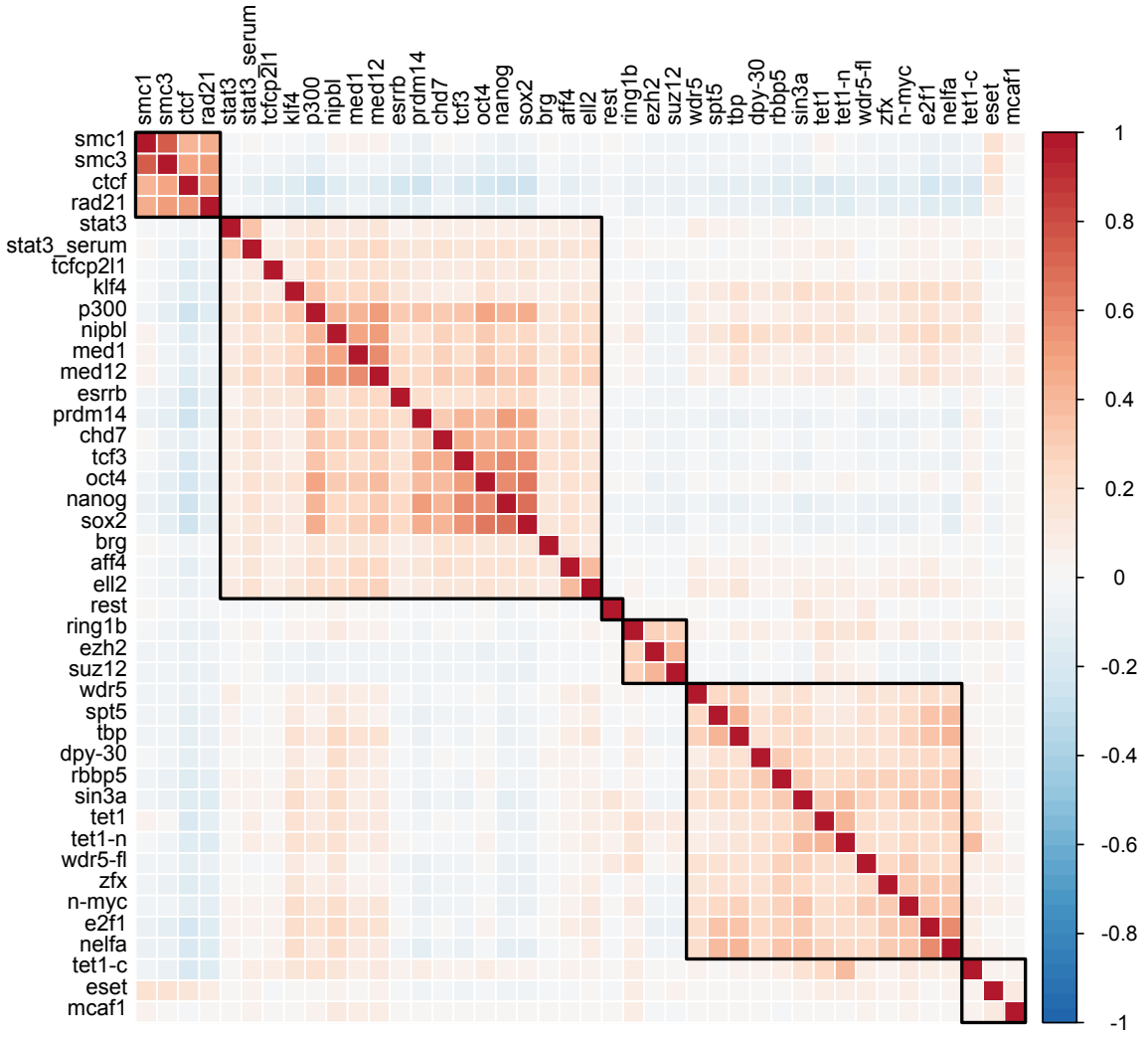


Figure 4.10: Hierarchical clustering of transcriptional regulators in ES cells based on binding sites. For all factors from the compendium the summits +- 100 bp was taken as peak positions and a list of total binding sites from all factors was generated. A binary table $T[i, j]$ was calculated storing for each binding site i whether factor j was bound to this location. This binary table was used for hierarchical clustering. The black boxes indicate the largest 6 clusters.

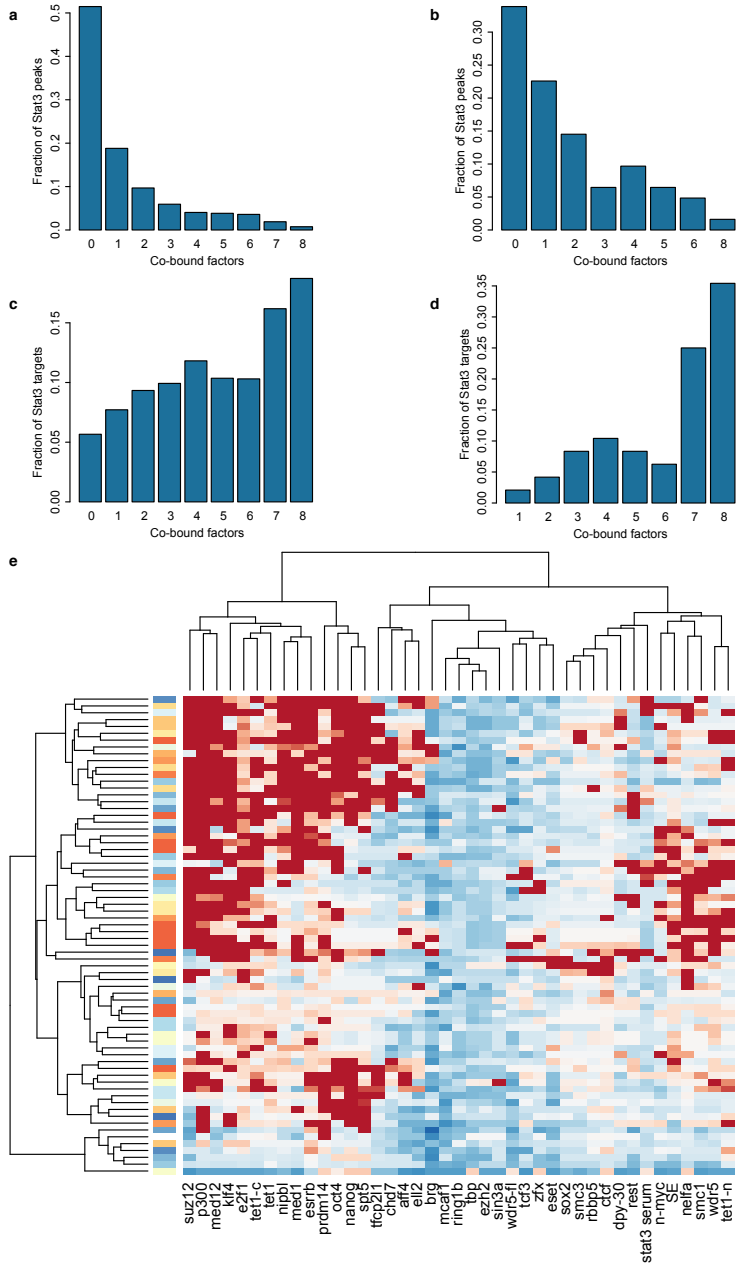


Figure 4.11: Cooperative binding of Stat3 and other pluripotency factors. (a) For each Stat3 binding sites the number of co-bound factors was determined (Oct4, Sox2, Nanog, Esrrb, Klf4, Prdm14, Tcf3 and Tfcp2l1). Shown are the fractions of Stat3 binding sites co-bound by a certain number of other factors. (b) as in (a) but only for Stat3 peaks that were associated with genes activated after LIF exposure. (c) As (a) but based on Stat3 target genes rather than binding sites. (d) as (c) but only for LIF-induced Stat3 target genes. (e) For each Stat3 peak the distance to the closest binding site of a certain other factor was determined. The distance matrix was used as input for hierarchical clustering. Red colours indicate close proximity while blue colours indicate larger distances. The column on the left indicates the fold change of the associated gene.

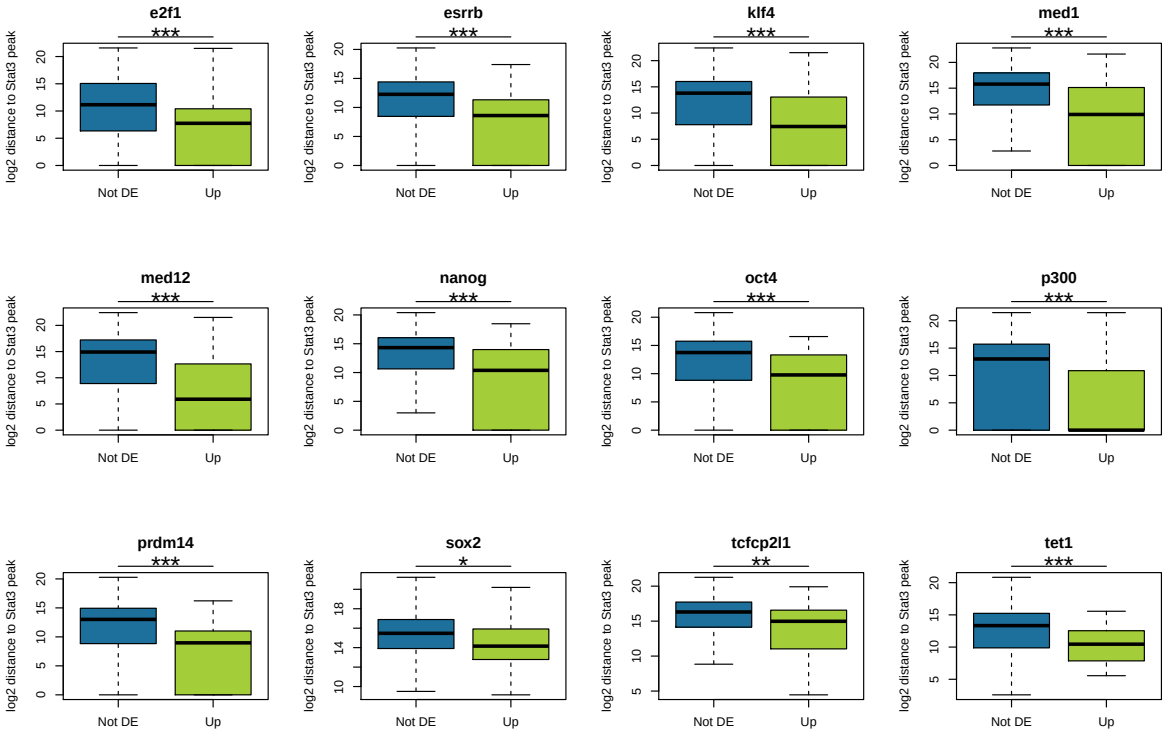


Figure 4.12: Distance to Stat3 binding sites. Displayed is the distance between Stat3 peaks and peaks of other transcriptional regulators for genes induced by LIF compared to genes unresponsive to LIF. Distance was log2 transformed after adding 1 to avoid infinite values. Proximity increases for upregulated genes associated with Stat3 binding. Two-sided t-tests were used to determine whether the increase in proximity was significant: * P-value ≤ 0.05, ** P-value ≤ 0.005, *** P-value ≤ 0.0005

distance between Stat3 peaks and the closest peak of the other factors was typically more than 15 kb, the distance to Stat3 peaks associated with upregulated genes was less than 1 kb. Notably, in the case of Klf4 50% of the Klf4 peaks associated with Stat3 bound upregulated genes, overlapped directly with the Stat3 peaks. Also p300, a marker for active enhancers, was present at the majority of Stat3 binding sites that were associated with genes whose expression was positively regulated by Stat3 activation.

These results suggests that Stat3 is part of the core pluripotency network and controls gene expression of its targets in close collaboration of other pluripotency factors, such as Klf4 and Esrrb. The close proximity of these factors and Stat3 at LIF-induced genes further suggest that Stat3 might physically interact with core pluripotency factors and regulators as p300 to induce expression of its target genes.

4.2.4 Comparing LIF signalling in 2i versus serum

The most comprehensive study on Stat3 in mouse ES cells to date has been performed in serum-containing media supplemented with LIF (Bourillot et al., 2009). A major drawback of this studies is the fact that ES cells cannot be maintained in serum without LIF. In order to study LIF signalling LIF was withdrawn and then re-added after 24 hours. In response to LIF withdrawal the majority of cells will have started differentiating and generating a heterogeneous cell population.

Culturing ES cells in 2i enables the study of Stat3 signalling in a homogeneous environment, where the cells remain pluripotent at any time (Ying et al., 2008; Wray et al., 2010). This is the first comprehensive data set examining LIF signalling and Stat3 binding in ground state pluripotency conditions. Despite these experimental differences, gene expression levels measured by microarray and RNA-seq correlated with a Pearson's $r = 0.74$ (Figure 4.13a).

Re-analysing the microarray data resulted in the identification of 176 genes (FDR 5%), out of which 69 were upregulated following LIF exposure. Despite the different experimental setups, both experiments identified the most important Stat3 targets which have been linked to stem cell self-renewal (Figure 4.13b): Gbx2, Icam1, Klf4, Pim1, Socs3, Spry2, Stat3 and Zfp36l1 (Tai and Ying, 2013; Yang et al., 2005; Kim et al., 2006; Hall et al., 2009; Aksoy et al., 2007; Cinelli et al., 2008; Wegmüller et al., 2007).

A set of 61 genes were found to be induced in serum only, of which 3 genes (Klf5, Junb and Zfp36) were found upregulated in 2i after 24 hours of LIF exposure. The later induction in our setup might be due to the different LIF concentrations used in the two studies. While Bourillot et al. reactivated endogenous Stat3 using 10.000 U/ml of LIF (Bourillot et al., 2009), for the RNA-seq experiment Stat3 was activated using 100 U/ml of LIF. A total of 20 genes out of the 58 genes not found in the RNA-seq data are involved in the regulation of transcription. Furthermore the set of genes was enriched for genes involved in the MAPK pathway (KEGG pathway analysis with DAVID, *P*-value 0.022): Fos, Jun, Dusp1, Dusp6, Gadd45a and Map2k5. Activation of the MAPK pathway is known to drive differentiation and is blocked in 2i (Ying et al., 2008; Wray et al., 2010). Higher expression of components of the MAPK pathway in serum is thus consistent with the observation of spontaneous differentiation in ES cells cultured in serum (Wray et al., 2010).

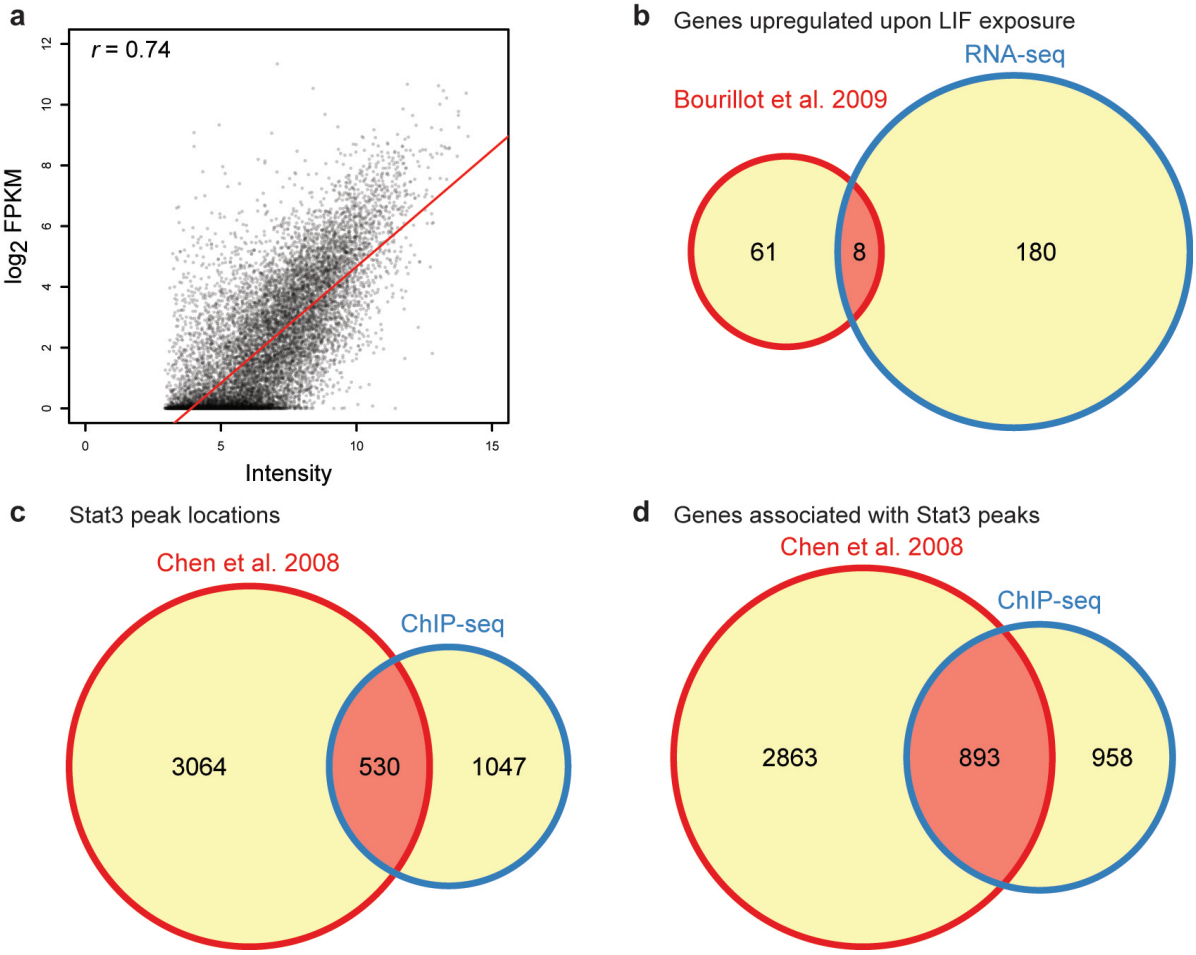


Figure 4.13: Comparing Stat3 targets in 2i and serum. (a) Scatterplot showing the correlation of gene expression from Bourillot et al., 2009 and the RNA-seq data. Pearson's r is shown in the upper left corner. (b) Showing the overlap of Stat3-regulated genes in Bourillot et al., 2009 and the RNA-seq data. (c) Showing the overlap of binding sites detected by Chen et al., 2008 and the ChIP-seq data. (d) Overlap of associated genes.

In order to test whether genes found upregulated in the microarray study were true Stat3 targets or due to prior LIF withdrawal, the set of LIF-induced genes in serum or 2i were compared with the genes downregulated upon Stat3 knock-down (Nishiyama et al., 2013). Knock-down of Stat3 leads to the downregulation of 509 genes and upregulation of 1206 genes with an absolute fold change ≥ 1.5 (Nishiyama et al., 2013). Only two of the 61 (2%) genes identified in the microarray data, were downregulated upon Stat3 knock-down. In contrast, 39 of the 180 (22%) genes induced only in 2i overlapped with downregulated genes. Notably, five genes that were upregulated both in serum and in 2i were downregulated upon Stat3 knock-down. These results indicate that our experimental setup was more suitable to identify Stat3-regulated genes.

Stat3 regulatory activity was compared between 2i/LIF and serum/LIF culture condition, comparing Stat3 binding sites of ES cells cultured in 2i/LIF versus serum/LIF. For this purpose, Stat3 ChIP-seq data from Chen et al. was re-analysed, identifying 5730 peaks associated with 3756 genes (Chen et al., 2008). The greater number of peaks may indicate a higher Stat3 activity in serum/LIF conditions. A total of 530 peaks and 893 associated genes were shared between the two data sets (Figure 4.13c and d). While 31 of the 61 genes whose expression was induced only in serum were bound by Stat3 in serum (of which 18 are only bound in serum), 41 of the 180 genes whose expression was induced in 2i only were bound by Stat3 in 2i (of which 9 are only bound in 2i).

Marks et al. identified genes that were differentially expressed between mouse ES cell cultured in 2i/LIF or serum/LIF (Marks et al., 2012). However, they did not make use of the presence of biological replicates, but rather used average expression levels across replicates and a simple Student's t-test to identify differentially expressed genes without adjusting for multiple testing. This data set was re-analysed to gain further insights into the transcriptional differences of ES cells in 2i/LIF and ES cells in serum/LIF. Differential expression analysis lead to the identification of 230 genes with higher expression in 2i/LIF and 197 genes with higher expression in serum. A significant fraction of the genes more highly expressed in serum were Stat3 bound only in serum (45 genes, P-value $3.054e-05$). In contrast, the fraction of genes more highly expressed in 2i/LIF that were Stat3 bound only in 2i/LIF was not significant (6 genes, P-value 0.9461). These results indicate that the larger number of Stat3 binding sites in serum/LIF are indeed associated with the activation of more genes under these culture conditions.

The gene, whose expression was most robustly induced in serum/LIF, was Dnmt3l (20-fold upregulated). The DNA methylase is responsible for the establishment of DNA methylation and maternal imprinting (Hata et al., 2002; Deplus et al., 2002; Jia et al., 2007). This is consistent with the finding that 2i induces global DNA hypomethylation (Leitch et al., 2013; Yamaji et al., 2013; Ficz et al., 2013). Furthermore, several genes of the MAPK pathway were expressed at a higher level in serum (Igf2, Spry4, Kitl, Dusp6) as were genes involved in cell differentiation (Igf2, Dusp6, Efnb2, Otx2, Cyr61, Zfp423, Peg10). Also more highly expressed in serum were the Id proteins Id1 and Id3. This is in line with the observation that BMP can substitute for serum by induction of Id proteins in collaboration with Stat3 (Ying et al., 2003).

These results suggest that the differences observed between ES cell cultured in 2i/LIF and ES cells kept in serum/LIF are partly due to differences in LIF/Stat3 signalling between the two conditions. Stat3 is more active in media supplemented with LIF and serum and promotes the expression of more target genes, including genes involved in differentiation.

4.3 Conclusion

The aim of this project was to gain a better understanding of Stat3 signalling in mouse ES cells. The 2i culture condition was used to maintain undifferentiated ES cells and RNA-seq at several time points was used to follow changes in expression following LIF exposure. Stat3 independent LIF targets could be identified by performing RNA-seq of the same time points for $Stat3^{-/-}$ cells.

The RNA-seq data was used to identify novel transcripts regulated by Stat3, including antisense transcripts. A number of antisense transcripts showed changes in their expression level following LIF induction. However, none of these resulted in changes of the corresponding annotated gene. Furthermore, several intergenic transcripts that were upregulated by LIF were identified. A total of five out of six tested transcripts were confirmed using qPCR. Fold changes observed in the RNA-seq analysis were higher than fold changes measured by qPCR, most likely due to the wider range of expression levels detectable with RNA-seq. A large proportion of the LIF-induced novel transcripts showed Stat3 binding sites, suggesting that the expression of these transcripts is Stat3

dependent. Further investigation is needed to analyse whether these transcripts are functional and whether they play a role in ES cell self-renewal.

Differential expression analysis of known genes identified several known Stat3 targets. Most highly induced was the expression of Socs3 which is a known negative regulator of LIF signalling (Boyle et al., 2009). Stat3 binding data was used to prioritise the set of LIF-induced genes to further enrich for direct Stat3 targets, leading to a small set of candidate targets. This gene set contained several genes with known roles in ES cell self-renewal, such as Klf4, Gbx2 and Pim1 (Hall et al., 2009; Tai and Ying, 2013; Aksoy et al., 2007). However, the most highly expressed gene upregulated upon LIF exposure, Tfcp2l1, had not been linked to Stat3 signalling before. While the known Stat3 targets are not sufficient to replace LIF and are dispensable for LIF signalling, expression of a Tfcp2l1 transgene at an endogenous level can substitute for LIF signalling. Tfcp2l1 is further required for LIF dependent LIF signalling, as knock-down of Tfcp2l1 reduces the number of undifferentiated colonies even the presence of LIF (Martello et al., 2013). Furthermore, Tfcp2l1 can reprogram EpiSCs to ES cells (Martello et al., 2013; Ye et al., 2013) and has been identified as a downstream target of the 2i culture condition as well, emphasizing the important role of Tfcp2l1 in the regulatory circuitry that governs ES cell self-renewal (Ye et al., 2013).

Genes upregulated by LIF without evidence of Stat3 binding were compared to genes with Tfcp2l1 and Klf4 binding sites to determine whether these genes might be secondary Stat3 targets regulated by these two factors. More than 50% of LIF-induced genes without Stat3 binding sites were bound by Tfcp2l1 or Klf4, and most often co-bound by both factors. These results indicate that Klf4 and Tfcp2l1 are the two main downstream targets of Stat3 and regulate the expression of most of its secondary targets.

The expression of only a small percentage of Stat3 bound genes was activated following LIF exposure. Since Stat3 has been shown to act in collaboration with other factors to regulate the expression of its target genes in other cell types (Kwon et al., 2009; Langlais et al., 2012; Kidder et al., 2008), Stat3 binding sites were compared to binding sites of other pluripotency factors and transcriptional regulators (Martello et al., 2012). LIF-responsive genes were typically bound by Stat3 and several other of the aforementioned factors. Furthermore, the proximity between the binding sites of Stat3 to other pluripotency factors was significantly increased for the set of LIF-induced genes,

indicating that Stat3 regulates gene expression in cooperation with other pluripotency factors, such as Klf4 and Esrrb.

As this is the first comprehensive data set of LIF signalling in ground state pluripotency, the results of the transcriptome analysis as well as the Stat3 binding sites were compared to data sets previously derived in serum/LIF conditions (Bourillot et al., 2009; Chen et al., 2008). Despite the very different exprimental setups, both differential expression analyses identified the most prominent Stat3 targets Socs3, Klf4, Gbx2 and Pim1. Notably, Tfcp2l1 was not identified in the microarray data set. The majority of genes identified only in the microarray data set were not responsive to Stat3 knockdown, indicating that their induction was not due to Stat3 activation, but rather to cells differentiating during the 24 hours of LIF withdrawal.

Stat3 showed more binding sites in serum/LIF. Another RNA-seq data set comparing gene expression between serum/LIF and 2i/LIF culture conditions was used to test whether the changes in Stat3 binding were reflected in the expression of the corresponding genes (Marks et al., 2012). A significant fraction of genes more highly expressed in serum/LIF showed Stat3 binding in serum/LIF conditions only, including a number of genes involved in differentiation. Therefore, changes between the two conditions are partly due to differences in LIF/Stat3 signalling. Stat3 is more active in serum/LIF condition and regulates the expression of more target genes, including genes involved in differentiation which can explain the spontaneous differentiation which is commonly observed for ES cells cultured in serum/LIF (Wray et al., 2010).

The mechanism leading to higher Stat3 activity in serum/LIF conditions remains unclear. Stat3 was not differentially expressed between ES cells in 2i/LIF and ES cells in serum/LIF, indicating that the changes must be due to differences in activation through phosphorylation, changes in protein degradation rate or different localisation of the Stat3 protein within the cell.

4.4 Methods

4.4.1 Experimental methods

4.4.1.1 Embryonic Stem Cell Culture

Embryonic stem cells were cultured without feeders on plastic coated with 0.1% gelatine (Sigma, cat. G1890) and replated every three days at a split ratio of 1 in 10 following dissociation with Accutase (PAA, cat. L11-007). Cells were cultured in the serum-free media N2B27 (NDiff N2B27 base medium, Stem Cell Sciences Ltd, cat. SCS-SF-NB-02) supplemented with the small-molecule inhibitors PD03 (1μM, PD0325901) and Chiron (3μM CHIR99021) and LIF prepared in-house. Colony forming assays were carried out by plating 60 ES cells/cm^2 on plates coated with laminin (Sigma, cat. L2020). Plates were fixed and stained for alkaline phosphatase (Sigma, cat. 86R-1KT) according to the manufacturer's protocol. Plates were scanned using a CellCelector (Aviso) and scored manually.

4.4.1.2 RNA-seq Library Construction

RNA was extracted using the TRIzol method (Invitrogen) followed by treatment with TURBO DNase (Ambion). Ribosomal RNA was depleted using RiboMinus (Invitrogen), and the remaining RNA was sheared by ultrasonication on a Covaris S2 for 90s with the following parameter settings: Duty cycle = 10, Cycles per Burst = 200, Intensity = 5. Fragmented RNA was reverse-transcribed with SuperScript III (Invitrogen) at 50°C for 2h and random hexamer primers in the presence of 6μg/ml actinomycin D to inhibit the generation of second-strand products. Second-strand cDNA was synthesized by DNA Polymerase I for 2h at 16°C in the presence of RNase H and dUTPs instead of dTTPs. End repair of double-strand cDNAs was carried out with T4 DNA polymerase and T4 polynucleotide kinase (New England Biolabs). Blunt-end, 3′-phosphorylated products were 3′-adenylated by exo- Klenow fragment in the presence of dATPs and ligated to sequencing adapters (Illumina) by T4 DNA ligase (New Englad Biolabs) at 20°C for 30 minutes. Following adapter ligation, the second strand of the library constructs was digested with uracil DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE 1) for 30 min at 37°C. PCR amplification of first-strand library constructs was carried out with Phusion DNA polymerase (Finnzymes) for 15 cycles. Purification

of reaction products between each step was performed with Ampure XP paramagnetic beads (Beckman Coulter). The molarity and size of the libraries was assessed by DNA 1000 microfluidic chips on the Agilent 2100 Bioanalyzer. Sequencing was performed on the Illumina GAIIx yielding $35 - 40$M 105bp reads per sample.

4.4.1.3 ChIP-seq Library Construction

ES cells were fixed by formaldehyde crosslinking for 10 min and lysed. DNA-protein complexes were sonicated on a Diagenode Bioruptor to produce fragments in the range of 200-400 bp. Bound DNA was immunoprecipitated using Santa Cruz Biotech SC-7179 (total Stat3), Cell Signalling Technologies 9131 (phosphorylated Stat3), Abcam ab123354 (Tfcp2l1) and R&D Systems AF3158 (Klf4) and crosslinks were reversed by heating at 65°C for 1h. End repair of purified DNA fragments was carried out with T4 DNA polymerase and T4 polynucleotide kinase (New England Biolabs). Blunt-end, 3′-phosphorylated products were 3′-adenylated by exo- Klenow fragment in the presence of dATPs and ligated to sequencing adapters (Illumina) by T4 DNA ligase (New England Biolabs) at 20°C for 30 minutes. PCR amplification was performed with Phusion DNA polymerase (Finnzymes) for 15 cycles. Purification of reaction products between each step was performed with Ampure XP paramagnetic beads (Beckman Coulter). The molarity and size distribution of the libraries was assessed by DNA 1000 microfluidic chips on the Agilent 2100 Bioanalyzer. Sequencing was performed on the Illumina HiSeq 2000 yielding 36-43M 50bp reads per sample.

4.4.2 Computational methods

4.4.2.1 Mapping of the RNA-seq Data

Before mapping, reads were filtered for ribosomal content. Bowtie (Langmead et al., 2009) was used to align the reads against an rRNA reference consisting of annotated rRNA sequences from Ensembl and RefSeq. Between 9 and 16% of the reads were identified as of ribosomal origin. Remaining reads were aligned to the mouse reference genome (build mm10) using GSNAP (version of 2012-05-24,Wu and Nacu, 2010). GSNAP was provided with known splice sites based on the Ensembl annotation (version 68, Flicek et al., 2011), but enabled to identify novel splice sites ($N = 1$). Maximum 10 mismatches

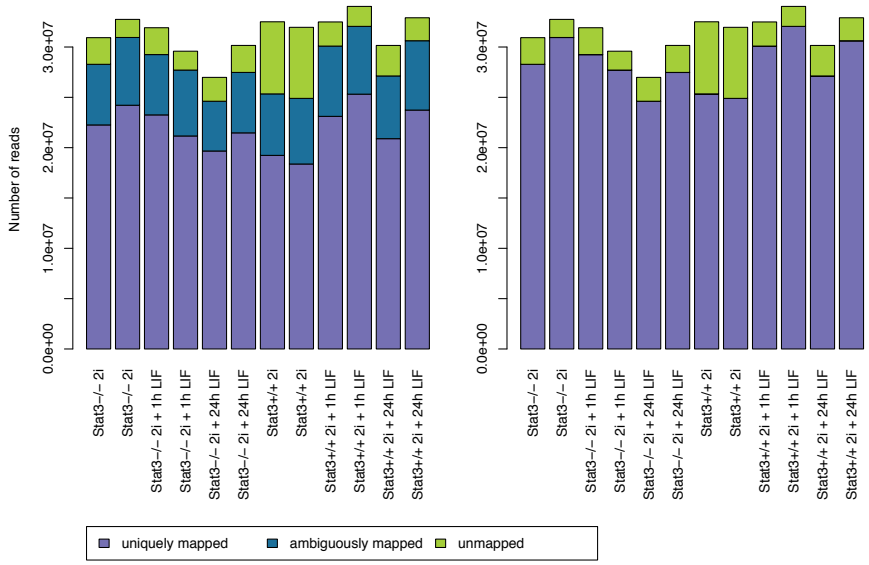


Figure 4.14: Number of aligned reads. Showing the number of reads that were mapped uniquely, ambiguously or unmapped for the full data set (left) and after filtering secondary alignments (right).

were allowed as well as maximum 10 alignments per read. As discussed in Chapter 2, GNSAP is able to place ambiguously mapped reads by analysing informative reads from the surrounding region. Ambiguously mapped reads marked as secondary alignments using the 0x100 flag were excluded from further analysis. This notably reduced the amount of ambiguously mapped reads ($<<$ 1%, Figure 4.14).

4.4.2.2 Genome annotation

Gene counts for annotated genes were calculated using the htseq-count utility (version 0.5.3p3, http://www.huber.embl.de/users/anders/HTSeq/doc/count.html) and the ensembl genome annotation for mouse version 69. The following parameters were used: mode = intersection_nonempty; s = reverse; type = exon; idattr = gene_id.

4.4.2.3 Detection of unannotated transcripts

Regions with RNA-seq support, but not present in the Ensembl annotation (mouse version 69), were detected using a sliding-window approach. As the data was strand-

specific novel region detection was performed for each strand independently. A region had to have a minimum length of 80 bp and support of at least 10 reads and an average coverage of 2. If two regions were less than 50 bp apart, they were merged together. Novel transcribed regions fulfilling these requirements were classified according to their position relative to the next annotated gene: **antisense**, novel regions overlapping exons of known genes in antisense orientation; **intronic sense**, novel regions within introns of annotated genes; **intronic antisense** novel regions within introns of annotated genes on the other strand; **intergenic** novel regions that are more than 50 kb away from any annotated gene; **five prime region** novel region that is five prime of its closest annotated gene. **three prime region** novel region that is three prime of its closest annotated gene; **repeat** Novel region that is marked as repeat region by the RepeatMasker of the UCSC genome browser (Kuhn et al., 2013; Fujita et al., 2011).

Intergenic regions were assembled into potential spliced isoforms based on their expression level correlation. This was performed independently for each chromosome. For this purpose a correlation matrix with a column and row for each region was initialised with zeros. For regions on the same strand and within 50 kb of each other the corresponding matrix entry was replaced by their spearman correlation. Regions with spearman correlation ≥ 0.8 were assembled together.

4.4.2.4 Calculation of differentially expressed genes

Gene counts were used as input for differential gene expression analysis with DESeq (version 1.10.1, Anders and Huber, 2010). Genes with a P-value of < 0.05 after Benjamini-Hochberg correction to account for multiple hypothesis testing were selected for further analysis (Benjamini and Hochberg, 1995). GO term analysis for differentially expressed genes was performed using the online functional annotation tool DAVID (version 6.7, Huang et al., 2009). GO terms were sorted according to their P-value after Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

4.4.2.5 Analysis of published microarray data

The data from Bourillot et al. (E-TABM-562) was re-analysed using the ArrayExpress Bioconductor package (Bourillot et al., 2009; Kauffmann et al., 2009). The arrays were normalised using robust multi-array average (RMA) normalisation (Irizarry et al., 2003).

Differentially expressed genes were identified between mouse ES cells deprived of LIF for 24 hours and after 1 hour LIF exposure (Smyth, 2004). Probe identifiers were mapped to Ensembl gene identifiers using the Bioconductor package biomaRt (Durinck et al., 2005).

4.4.2.6 Mapping of the ChIP-seq Data

We performed ChIP sequencing for total Stat3 and phosphorylated Stat3, as well as for Klf4 and Tfcp2l1. We also obtained ChIP-seq data for Stat3 in ES cells cultured in LIF and serum and a GFP control from Chen et al. (Chen et al., 2008). Reads were mapped to the mouse genome (build mm10) using bowtie (version 0.12.8, Langmead et al., 2009) discarding reads with more than one possible alignment.

4.4.2.7 Peak calling

Peak detection was performed using MACS (version 1.4.1, Zhang et al., 2008) with the following parameters: tsize = 36; mfold = 10,30; gsize = 2644093988. Resulting peaks were processed with PeakSplitter (Salmon-Divon et al., 2010) to resolve situations where several peaks have been merged into one peak region.

4.4.2.8 Integrating mouse ES cell ChIP-seq compendium

Martello et al. collected ChIP-seq data sets that are available on mouse ES cells and provide processed data files (coverage and peak files, see Martello et al., 2012). Peak files were obtained from the compendium website (http://lila.results.cscr.cam.ac.uk/ES_Cell_ChIP-seq_compendium_UPDATED.html). Peak summits were detected using PeakAnalyzer (Salmon-Divon et al., 2010). As the peak files are for mouse genome build mm9, the UCSC utility liftOver was used to convert the peak summits to mouse genome build mm10 (Kuhn et al., 2013).

4.4.2.9 Peak annotation

Peaks were annotated with genes using a slight variation of the method implemented in Martello et al. (Martello et al., 2012). Each peak was mapped to up to two genes as follows: the two closest genes, one for each strand, were identified by examining

50 kb up- and downstream of the peak. If a gene is overlapping a peak, the peak is associated with this gene only unless there is another overlapping gene on the other strand. Annotated peaks were further classified according to their relative position to the annotated gene. The promoter region was defined as 3 kb upstream of the transcription start site. Similarly, the downstream region was defined as the 3 kb downstream of the transcription termination site. Other possible classifications are bound within the UTR, coding exon, intron or intergenic region (that is within 50 kb of the transcription start site, but not within the promoter region).

4.4.2.10 Statistical testing

The significance of overlapping gene or peak sets have been calculated using a Pearson's Chi-squared test. Significant differences in continuous features, such as distances or expression levels have been detected using two-sided t-tests after log-transformation.

Chapter 5

Discussion

RNA sequencing provides a genome-wide readout of actively transcribed genes in a cell. The analysis of this kind of data requires sophisticated methods tuned to the specific needs and biases of RNA-seq data. In the first part of my thesis I have assessed the performance of a comprehensive set of RNA-seq aligners and transcript reconstruction methods. In the second part of my thesis, I used RNA-seq data in combination with ChIP-seq data to study LIF-dependent Stat3 signalling in mouse ES cells.

5.1 Methods for the analysis of RNA-seq data

The analysis of RNA-seq data requires splice-aware aligners and methods to infer transcript isoforms from short reads. Two of the main challenges of RNA-seq alignment are placing spliced reads and resolving ambiguously mapped reads. Over the last few years, several spliced aligners have been developed, such as TopHat, GSNAP and MapSplice (Trapnell et al., 2009; Kim et al., 2013; Wu and Nacu, 2010; Wang et al., 2010a). Some developers have compared their aligner to a small subset of other methods (Dobin et al., 2013; Grant et al., 2011). However, these studies were inherently susceptible to bias in the selection of tools and data sets and limited in scope. For example, aspects like the influence of annotation and of base call qualities were not considered for these comparisons.

The RNA-seq aligner analysis presented in this thesis is the first comprehensive study of spliced aligners to date where the analysing team was not involved in the development of any of the methods under investigation. The quality of the simulated data exceeds

the quality of previously used simulated data sets as the simulation process included simulating base call qualities. This enabled the analysis of how methods incorporate this kind of information and how it influences the placement of mismatches. Furthermore, this study is the first one to analyse the influence of guided alignment which make use of genome annotations for the alignment of spliced reads. Our study has demonstrated that the use of a dedicated spliced aligner is crucial for the interpretation of RNA-seq data. It also revealed that some methods make more extensive use of the base call qualities and allow for more mismatches when aligning reads of poor quality, allowing them to align more reads. Annotation guided alignment of spliced reads improved sensitivity, but increased the false discovery rate at the same time.

Also the transcriptome assembly study presented in this thesis is the first comprehensive study of this kind. As for RNA-seq aligners, some developers have compared their software against a small set of other methods (Mezlini et al., 2013; Li et al., 2011), but no comparison of a more inclusive set of transcriptome reconstruction methods has been performed yet. The results of this benchmark make clear that the problem of transcriptome reconstruction based on RNA-seq data is far from being solved. While most methods accurately detect expressed exons, the identification of their connectivity proves problematic for all methods. Being able to sequence whole transcripts would solve that problem and allow the direct identification of transcript isoforms. While current sequencing machines are not able to sequence whole transcripts, the development of *third-generation* sequencing technologies might enable this in the near future (Morey et al., 2013).

Also for current sequencing technologies read length is continuously increasing, for example the Illumina platform can now generate 150nt reads. However, these advances only marginally improve transcript reconstruction, which benefits from reads spanning multiple introns. Since the average human exon length is about 200nt, even with 150nt long reads, the chances of spanning several introns is very small. Read length would have to increase much more in order to substantially improve transcript assembly. There are, however, limits to read length in RNA-seq data. As described in the introduction the library preparation protocol includes a fragmentation and subsequent size selection step. Typically, transcripts are fragmented into sequences of roughly 250nt length. Sequencing longer reads will cause overlapping mate pairs in the case of paired-end sequencing which defeats the benefits of paired-end sequencing. Increasing fragment size could solve that

problem but will also involve new challenges. Transcripts shorter than a given fragment size will get lost if a size selection step is performed. If short transcripts are included the insert size distribution will become broader. Many methods make assumptions about the insert size distribution and use these to assemble transcript isoforms and thus show decreased performance for RNA-seq datasets based on longer fragment lengths.

A better way to improve transcriptome reconstructions may be the incorporation of other data sources. Providing additional information about transcript start sites, for example from CAGE sequencing (Hoon and Hayashizaki, 2008), and transcript termination sites, for example using poly(A) site mapping protocols (Wilkening et al., 2013), may increase the ability of methods to link exons into valid transcript isoforms.

5.1.1 Future work

Since RNA-seq is an emerging technology, new methods for the analysis of this kind of data are still being developed. During the write-up of this thesis several new transcript reconstruction methods (Song and Florea, 2013; Hiller and Wong, 2013; Legault and Dewey, 2013; Bernard et al., 2013; Behr et al., 2013) and a new RNA-seq aligner (Hu et al., 2012) have been published. It would be desirable to include these methods in the benchmarks. A possible solution for this would be the generation of a webservice where developers can upload their alignments or transcript assemblies based on the data used in the benchmarks. The webservice would then run the analysis and test the performance of their software. The result could be graphically displayed together with the results presented in this thesis. Storing these results would also provide an useful resource for the community and enable interested readers of these studies to find up to date results on new methods online.

The transcriptome reconstruction study focused mainly on the influence of transcriptome complexity on transcript assembly. It might also be interesting to study the influence of other factors, such as read length, single-end versus paired-end protocols and strand-specific protocols. In order to assess these factors, data from the same organism has to be generated that only differs in these aspects. If available, simulated data should be included as it provides a proper gold standard to compare against. Especially, the benefit of paired-end data in RNA-seq data is a controversial topic. While the use of paired-end data in genome sequencing and the detection of structural variations is un-

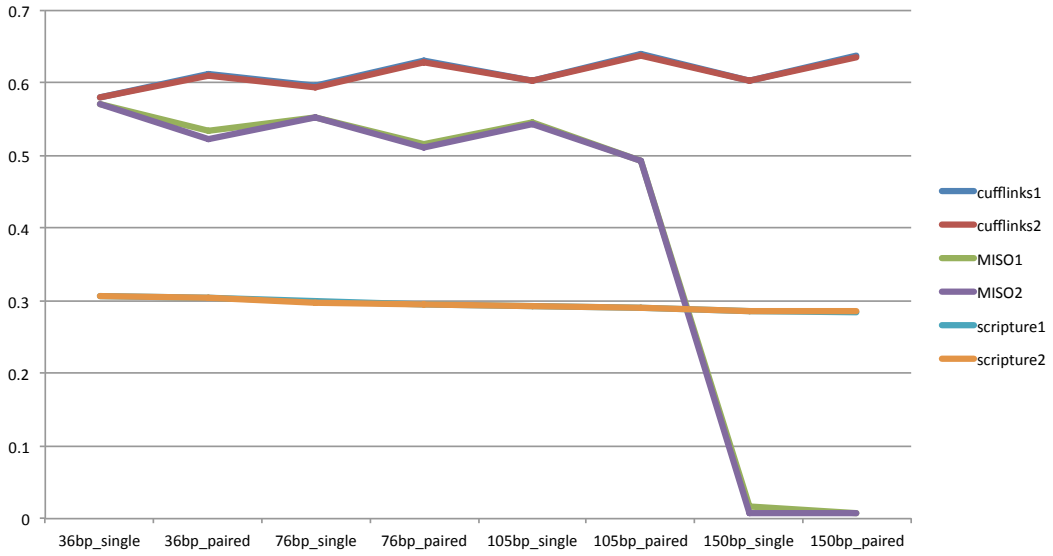


Figure 5.1: Quantification correlation dependent on different read characteristics. Shown is the correlation between real expression level and reported RPKM values for two simulated RNA-seq data sets.

questionable, it is not really known whether transcript reconstruction and quantification benefits more from paired-end sequencing or from longer reads which are more likely to span multiple introns. In an initial analysis using simulated RNA-seq data of differing read lengths and single- and paired-end, the influence of these aspects on transcript quantification was evaluated. Preliminary results for transcript quantification using Cufflinks, Scripture and MISO showed that only Cufflinks consistently benefited from the paired-end information while both MISO and Scripture methods were not able to make proper use of this information (Figure 5.1). Similarly, quantification with Cufflinks improved with increasing read length, while the other two methods did not show benefit from longer reads. MISO (version 0.4.1) had problems with read length greater than 100bp. These preliminary results on just a small set of quantification methods indicates that a closer study of the influence of these aspects of RNA-seq data on quantification and assembly methods might provide the community with valuable information which methods are better suited for which kind of data.

A more advanced aspect that could be included in this study would be the use of multiple insert sizes in RNA-seq data. It has been suggested that using parallel libraries with different insert sizes can provide valuable information for transcript reconstruc-

tion for complex transcriptomes (Smith et al., 2012). But it is questionable how many currently available methods are actually able to make use of this kind of data.

In the aligner study, the influence of providing annotation data was analysed. Similarly, it would be interesting to assess how much transcript reconstruction methods benefit from annotation data. This would require artificial data that simulates the expression of both annotated transcripts as well as novel splice variants and novel transcripts.

5.2 LIF/Stat3-signalling in mouse ES cells

LIF-mediated Stat3 signalling has long been known for its important role in maintaining mouse ES cell self-renewal. The discovery that primed stem cells such as EpiSCs and primate ES cells are not responsive to LIF suggests that LIF signalling is intrinsically linked to naive pluripotency. More detailed insights into the mechanisms of LIF signalling in mouse ES cells will lead to a better understanding of ground state pluripotency.

Before the development of the 2i culture system, an unbiased study of LIF signalling was not possible as LIF was required at all times to maintain ES cells in an undifferentiated state. The 2i culture system renders LIF unnecessary for the maintenance of undifferentiated mouse ES cells. However, adding LIF has positive effect on ES cells self-renewal even in 2i indicating that 2i and LIF act through independent pathways. The 2i system furthermore enable the generation and maintenance of Stat3 null embryonic stem cells which are vital to distinguish between Stat3-dependent signalling and other LIF downstream effects. Taken together, this makes the 2i culture conditions a superior system for the study of LIF-mediated Stat3 signalling. Making use of the RNA-seq technology is another improvement compared to previous studies as it enables the study of known genes as well as the identification of novel transcripts. Intersecting the RNA-seq data with genome wide Stat3-binding data provided further insights into Stat3-dependent gene regulation.

The study lead to the identification of the hitherto unknown Stat3 target Tfcp2l1. This transcription factor was known to be expressed in ES cells and to be downregulated upon differentiation but nothing was known about its function. Martello et al. demonstrated that Tfcp2l1 is required for the propagation of LIF signalling and that forced expression of Tfcp2l1 can substitute for Stat3 activation (Martello et al., 2013).

Tfcp2l1 is also required for the LIF-dependent reprogramming of EpiSCs (Martello et al., 2013; Ye et al., 2013), indicating that Tfcp2l1 plays an important role not only in the maintenance of pluripotency but also during the induction of it.

In addition to the identification of Stat3 targets the study aimed to gain insights into the mechanisms of Stat3-dependent gene regulation. It has been suggested that Stat3 and c-myc, a target of the Wnt-pathway and one of the Yamanaka reprogramming factors, co-occupy a significant number of pluripotency-related genes and that these two factors might cooperate to regulate gene expression (Kidder et al., 2008). However, only 22% of Stat3 bound genes are bound by c-myc as well (Kidder et al., 2008). The availability of binding data for a large number of pluripotency factors and transcriptional regulators in mouse embryonic stem cells, enabled the study of Stat3 gene regulation in the context of the pluripotency factor network.

A large fraction of Stat3 bound genes are bound by several pluripotency factors (Oct4, Sox2, Nanog, Esrrb, Klf4, Prdm14, Tcf3 and Tfcp2l1). LIF-responsive genes were usually bound by most of these factors (60% were bound by at least 7 of these 8 factors). Furthermore, the proximity between Stat3 binding sites and binding sites of these factors significantly increased for LIF-responsive genes, in particular for Klf4 and Esrrb. Furthermore, Stat3 bound at LIF-induced genes typically co-localised with markers of active enhancers, such as Med1, Med12 and p300. This shows that Stat3 is tightly integrated in the core pluripotency network and regulates gene expression in cooperation with these factors.

5.2.1 Future work

The analysis of the RNA-seq data identified several novel transcripts whose expression was significantly increased by LIF signalling. Several of these were bound by Stat3 suggesting that they are indeed regulated by Stat3. Non-coding RNAs have been shown to play an important role in ES cell self-renewal as well as reprogramming (Loewer et al., 2010; Guttman et al., 2011; Livyatan et al., 2013).

Therefore, it would be of interest to test whether the novel LIF-induced transcripts identified in this study have an effect on self-renewal or reprogramming. This could be achieved by shRNA induced knock-down of the transcripts. In case that these transcripts

affect the self-renewing capacity of ES cells or reprogramming efficiency, it will need to be investigated through what mechanism these transcripts execute their function.

Comparing LIF/Stat3 signalling in serum/LIF versus 2i/LIF based media suggested different levels of Stat3 activation within these two different conditions. However, the Stat3 mRNA was not found differentially expressed between mouse ES cells cultured in the two different conditions, suggesting that the changes have to be post-transcriptional. For example Stat3 mRNA stability may be different in the two conditions. The activity and stability of the Stat3 protein is further controlled by different protein modifications, such as phosphorylation, acetylation and ubiquitination (Hatakeyama, 2012). Studying the levels of these modifications between the two conditions might provide an explanation for the higher number of Stat3 binding sites and activated genes in the serum/LIF conditions.

The binding sites of Stat3, Klf4 and Tfcp2l1 showed a high degree of overlap. Most LIF-induced genes bound by Stat3 were also bound by Klf4 and Tfcp2l1. The same was true for LIF-responsive genes that lacked evidence for Stat3 binding. This suggests that Klf4 and Tfcp2l1 are two of the main Stat3 targets and responsible for the activation of a large proportion of Stat3 secondary targets. However, only forced Tfcp2l1 expression can substitute for Stat3 activation while forced expression of Klf4 cannot. Therefore, it would be interesting to study the differences between Klf4 and Tfcp2l1 in more detail to study the reason for Tfcp2l1 effect on self-renewal potential. of mouse ES cells. One first observation is that Tfcp2l1 binds more genes than Klf4 and that most genes bound by Klf4 are also bound by Tfcp2l1.

Interestingly, in a study overexpressing various transcription factors in ES cells, overexpression of Klf4 cause a large number of differentially expressed genes while both overexpression of Stat3 and Tfcp2l1 did not result in large expression changes (Correa-Cerro et al., 2011). However, these cells were cultured in serum/LIF and therefore all three factors should already be expressed at a high level, which might explain the lack of expression changes. In the complementary experiment, where transcription factors were knocked down using shRNAs, Klf4 knock down resulted in the downregulation of less genes than knock down of Tfcp2l1 (Nishiyama et al., 2013). Finally, the overlap with genes downregulated after Stat3 knock down was higher for genes affected by Tfcp2l1 knock down.

Appendix A

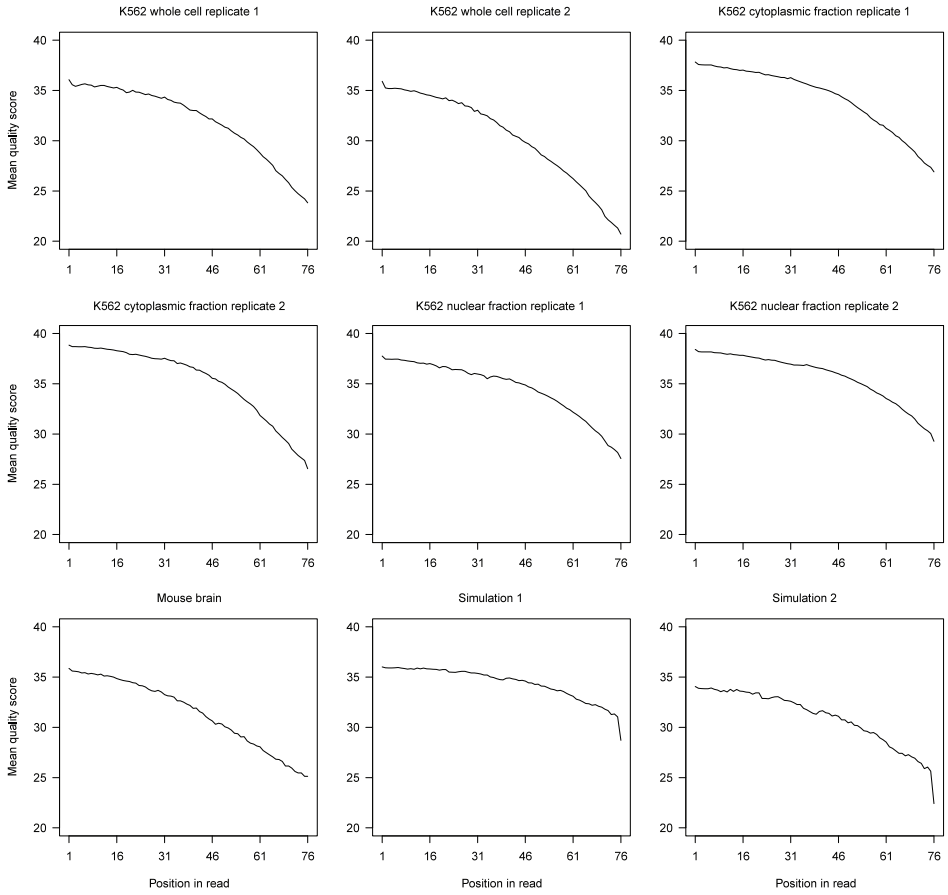


Figure A.2: Quality scores for RNA-seq datasets in this study. Mean base caller quality scores for each read position.

Table A.1: Alignment yield. Percentage of sequenced or simulated read pairs mapped by each protocol, for the data sets used in this study. Read pairs are classified by the number of alignments reported per mate.

	Both mates uniquely mapped	Both mates multi-mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
A. K562 whole cell replicate 1							
BAGET ann	87.78	0.13	0.98	3.43	0.24	92.57	90.73
GEM ann	47.13	42.92	0.37	0.77	0.72	91.91	91.17
GEM cons	47.45	42.57	0.37	0.79	0.73	91.91	91.15
GEM cons ann	47.38	42.65	0.37	0.78	0.73	91.91	91.16
GSNAP	79.50	10.98	0.04	0.90	0.35	91.77	91.14
GSNAP ann	79.61	10.86	0.04	0.92	0.35	91.78	91.15
GSTRUCT	74.48	16.01	0.04	0.88	0.39	91.80	91.17
GSTRUCT ann	77.86	12.63	0.04	0.92	0.35	91.80	91.16
MapSplice	83.31	0.01	0.05	5.81	0.88	90.07	86.72
MapSplice ann	83.30	0.01	0.05	5.81	0.89	90.07	86.71
PALMapper	32.84	36.97	18.72	1.50	1.67	91.69	90.11
PALMapper cons	18.12	0.00	0.00	24.23	0.00	42.36	30.24
PASS	82.13	0.33	0.18	8.17	0.05	90.86	86.75
PASS cons	80.95	0.32	0.00	6.65	0.00	87.93	84.60
ReadsMap	55.49	4.42	6.46	11.17	1.17	78.70	72.54
SMALT	85.76	0.03	1.02	6.49	0.38	93.68	90.24
STAR 1-pass	83.76	5.68	0.00	0.00	0.00	89.45	89.45
STAR 1-pass ann	84.17	5.45	0.00	0.00	0.00	89.61	89.61
STAR 2-pass	81.75	7.85	0.00	0.00	0.00	89.60	89.60
STAR 2-pass ann	81.66	7.93	0.00	0.00	0.00	89.59	89.59
TopHat v1	73.35	4.09	0.00	9.65	1.39	88.48	82.96
TopHat v1 ann	73.39	4.11	0.00	9.60	1.39	88.49	82.99
TopHat v2	70.58	4.49	0.00	11.29	1.58	87.95	81.51
TopHat v2 ann	72.57	4.59	0.00	10.63	1.33	89.12	83.14
B. K562 whole cell replicate 2							
BAGET ann	84.16	0.14	1.54	8.79	0.51	95.14	90.49
GEM ann	47.12	41.72	0.49	3.09	2.43	94.85	92.09
GEM cons	47.46	41.34	0.49	3.14	2.42	94.85	92.07
GEM cons ann	47.40	41.42	0.49	3.12	2.42	94.85	92.08
GSNAP	78.74	11.68	0.05	2.35	0.93	93.75	92.11
GSNAP ann	78.86	11.61	0.05	2.32	0.93	93.77	92.14
GSTRUCT	73.07	17.45	0.05	2.22	1.03	93.82	92.19
GSTRUCT ann	74.60	15.91	0.05	2.26	0.99	93.81	92.19
MapSplice	76.45	0.01	0.06	12.23	2.06	90.82	83.68
MapSplice ann	76.43	0.01	0.06	12.23	2.07	90.81	83.66
PALMapper	31.12	35.15	18.23	4.72	4.80	94.03	89.27

Continued on next page

Table A.1 – continued from previous page

	Both mates uniquely mapped	Both mates multi -mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
PALMapper cons	34.52	0.00	0.00	34.39	0.00	68.92	51.72
PASS	74.45	0.32	0.17	17.64	0.13	92.72	83.83
PASS cons	73.19	0.32	0.00	10.36	0.00	83.87	78.69
SMALT	86.08	0.02	0.75	6.91	0.21	93.96	90.40
STAR 1-pass	82.90	5.99	0.00	0.00	0.00	88.89	88.89
STAR 1-pass ann	83.68	5.68	0.00	0.00	0.00	89.36	89.36
STAR 2-pass	81.25	8.10	0.00	0.00	0.00	89.36	89.36
STAR 2-pass ann	81.15	8.20	0.00	0.00	0.00	89.35	89.35
TopHat v1	62.54	3.61	0.00	16.54	2.34	85.03	75.59
TopHat v1 ann	62.56	3.63	0.00	16.50	2.35	85.04	75.62
TopHat v2	59.15	3.97	0.00	17.63	2.37	83.12	73.12
TopHat v2 ann	60.74	3.85	0.00	17.39	2.16	84.15	74.37
C. K562 cytoplasmic fraction replicate 1							
BAGET ann	91.83	0.11	1.00	3.51	0.29	96.74	94.84
GEM ann	52.24	42.33	0.63	0.72	0.72	96.63	95.91
GEM cons	52.66	41.87	0.63	0.75	0.72	96.63	95.90
GEM cons ann	52.57	41.98	0.63	0.74	0.72	96.63	95.90
GSNAP	82.59	12.69	0.12	0.68	0.31	96.39	95.89
GSNAP ann	82.53	12.75	0.12	0.69	0.31	96.40	95.90
GSTRUCT	77.97	17.34	0.12	0.79	0.31	96.53	95.98
GSTRUCT ann	79.34	15.97	0.12	0.81	0.30	96.53	95.98
MapSplice	90.31	0.01	0.09	4.29	0.63	95.33	92.87
MapSplice ann	90.29	0.01	0.09	4.31	0.63	95.32	92.86
PASS	89.47	0.19	0.19	5.90	0.03	95.78	92.82
PASS cons	88.33	0.18	0.00	5.51	0.00	94.03	91.27
ReadsMap	61.37	5.60	9.37	10.00	1.02	87.37	81.86
SMALT	88.12	0.00	0.49	5.63	0.14	94.39	91.50
STAR 1-pass	87.75	5.96	0.00	0.00	0.00	93.71	93.71
STAR 1-pass ann	87.72	6.12	0.00	0.00	0.00	93.84	93.84
STAR 2-pass	83.73	10.08	0.00	0.00	0.00	93.81	93.81
STAR 2-pass ann	83.60	10.20	0.00	0.00	0.00	93.80	93.80
TopHat v1	77.44	4.61	0.00	9.24	1.18	92.47	87.26
TopHat v1 ann	77.46	4.65	0.00	9.18	1.18	92.48	87.29
TopHat v2	75.66	5.96	0.00	9.56	1.25	92.43	87.03
TopHat v2 ann	77.35	6.01	0.00	8.96	1.08	93.39	88.37
D. K562 cytoplasmic fraction replicate 2							
BAGET ann	90.78	0.12	1.03	3.12	0.25	95.30	93.61
GEM ann	44.72	49.11	0.46	0.51	0.40	95.20	94.74
GEM cons	45.14	48.67	0.46	0.54	0.40	95.20	94.73

Continued on next page

Table A.1 – continued from previous page

	Both mates uniquely mapped	Both mates multi -mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
GEM cons ann	45.05	48.78	0.46	0.52	0.40	95.20	94.74
GSNAP	83.12	11.16	0.11	0.56	0.17	95.12	94.75
GSNAP ann	83.11	11.18	0.11	0.57	0.17	95.13	94.76
GSTRUCT	79.62	14.68	0.11	0.66	0.17	95.25	94.83
GSTRUCT ann	81.25	13.06	0.10	0.67	0.16	95.25	94.83
MapSplice	90.66	0.01	0.08	3.31	0.34	94.40	92.58
MapSplice ann	90.60	0.01	0.08	3.36	0.34	94.40	92.55
PASS	88.31	0.20	0.18	5.90	0.03	94.63	91.66
PASS cons	87.05	0.20	0.00	5.63	0.00	92.87	90.05
SMALT	87.24	0.00	0.55	5.99	0.17	93.94	90.86
STAR 1-pass	86.92	5.63	0.00	0.00	0.00	92.55	92.55
STAR 1-pass ann	86.65	6.02	0.00	0.00	0.00	92.67	92.67
STAR 2-pass	82.97	9.67	0.00	0.00	0.00	92.64	92.64
STAR 2-pass ann	82.83	9.79	0.00	0.00	0.00	92.63	92.63
TopHat v1	75.71	4.51	0.00	9.96	1.11	91.29	85.75
TopHat v1 ann	75.79	4.52	0.00	9.88	1.11	91.30	85.80
TopHat v2	73.38	5.91	0.00	10.80	1.20	91.30	85.29
TopHat v2 ann	75.16	6.32	0.00	10.00	1.07	92.55	87.02
E. K562 nuclear fraction replicate 1							
BAGET ann	92.05	0.25	1.02	2.65	0.40	96.36	94.84
GEM ann	64.76	29.89	0.40	0.72	0.45	96.22	95.63
GEM cons	65.17	29.45	0.40	0.74	0.45	96.22	95.62
GEM cons ann	65.11	29.52	0.40	0.73	0.45	96.22	95.62
GSNAP	87.22	7.68	0.06	0.65	0.26	95.87	95.42
GSNAP ann	87.25	7.66	0.06	0.65	0.26	95.88	95.43
GSTRUCT	88.12	6.84	0.06	0.71	0.21	95.93	95.47
GSTRUCT ann	88.70	6.25	0.06	0.71	0.21	95.93	95.47
MapSplice	90.43	0.01	0.08	4.10	0.55	95.16	92.84
MapSplice ann	90.43	0.01	0.08	4.09	0.55	95.16	92.84
PALMapper	46.25	24.11	21.33	1.87	2.00	95.57	93.64
PALMapper cons	37.19	2.26	3.25	33.81	3.40	79.90	61.29
PASS	89.41	0.39	0.26	5.43	0.04	95.53	92.79
PASS cons	88.22	0.38	0.00	5.23	0.00	93.83	91.22
ReadsMap	62.17	2.93	4.98	12.99	1.13	84.20	77.14
SMALT	90.82	0.01	0.54	3.88	0.17	95.42	93.40
STAR 1-pass	88.94	4.06	0.00	0.00	0.00	93.00	93.00
STAR 1-pass ann	88.77	4.31	0.00	0.00	0.00	93.08	93.08
STAR 2-pass	87.00	6.08	0.00	0.00	0.00	93.08	93.08
STAR 2-pass ann	86.95	6.13	0.00	0.00	0.00	93.07	93.07
TopHat v1	78.26	3.74	0.00	9.99	1.19	93.19	87.59

Continued on next page

Table A.1 – continued from previous page

	Both mates uniquely mapped	Both mates multi -mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
TopHat v1 ann	78.29	3.75	0.00	9.96	1.19	93.19	87.62
TopHat v2	77.30	4.10	0.00	10.23	1.26	92.88	87.14
TopHat v2 ann	78.12	3.71	0.00	9.97	1.07	92.87	87.35
F. K562 nuclear fraction replicate 2							
BAGET ann	90.76	0.19	0.80	2.47	0.34	94.55	93.15
GEM ann	64.61	28.22	0.32	0.66	0.36	94.17	93.66
GEM cons	64.95	27.85	0.32	0.68	0.37	94.17	93.64
GEM cons ann	64.89	27.92	0.32	0.67	0.36	94.17	93.65
GSNAP	86.38	6.72	0.05	0.55	0.18	93.88	93.52
GSNAP ann	86.39	6.71	0.05	0.55	0.18	93.89	93.52
GSTRUCT	86.95	6.19	0.04	0.62	0.15	93.96	93.57
GSTRUCT ann	87.68	5.46	0.04	0.62	0.15	93.96	93.57
MapSplice	89.59	0.01	0.07	3.22	0.35	93.24	91.46
MapSplice ann	89.59	0.01	0.07	3.22	0.35	93.24	91.46
PALMapper	45.82	22.50	21.39	1.80	1.92	93.43	91.57
PALMapper cons	42.29	2.43	3.46	31.26	2.97	82.42	65.30
PASS	88.47	0.43	0.26	4.24	0.04	93.44	91.30
PASS cons	87.31	0.42	0.00	4.37	0.00	92.09	89.91
SMALT	89.36	0.01	0.55	4.46	0.23	94.61	92.26
STAR 1-pass	87.62	3.89	0.00	0.00	0.00	91.50	91.50
STAR 1-pass ann	87.34	4.23	0.00	0.00	0.00	91.57	91.57
STAR 2-pass	85.67	5.91	0.00	0.00	0.00	91.57	91.57
STAR 2-pass ann	85.61	5.96	0.00	0.00	0.00	91.57	91.57
TopHat v1	78.17	3.07	0.00	8.95	0.91	91.10	86.17
TopHat v1 ann	78.20	3.08	0.00	8.92	0.91	91.10	86.19
TopHat v2	78.66	3.68	0.00	7.91	0.87	91.12	86.73
TopHat v2 ann	79.32	3.34	0.00	7.64	0.72	91.03	86.85
G. Mouse brain							
BAGET ann	90.34	0.28	1.81	5.87	0.67	98.98	95.71
GEM ann	62.53	31.64	2.89	0.42	2.12	99.60	98.33
GEM cons	62.80	31.33	2.89	0.45	2.12	99.60	98.31
GEM cons ann	62.72	31.44	2.89	0.43	2.11	99.60	98.33
GSNAP	83.92	9.54	1.51	1.46	2.01	98.45	96.71
GSNAP ann	83.88	9.59	1.51	1.46	2.01	98.45	96.72
GSTRUCT	81.63	13.29	1.23	1.00	1.56	98.71	97.44
GSTRUCT ann	81.94	13.00	1.20	1.01	1.56	98.71	97.43
MapSplice	88.42	0.24	1.63	5.89	1.42	97.60	93.95
MapSplice ann	88.49	0.24	1.63	5.81	1.43	97.60	93.98
PASS	87.38	0.31	0.33	9.48	0.04	97.54	92.78

Continued on next page

Table A.1 – continued from previous page

	Both mates uniquely mapped	Both mates multi -mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
PASS cons	84.99	0.27	0.00	10.07	0.00	95.33	90.29
ReadsMap	57.26	3.68	3.20	16.36	0.99	81.50	72.82
SMALT	88.66	0.01	0.86	5.27	0.18	94.97	92.25
STAR 1-pass	84.28	4.95	0.00	0.00	0.00	89.23	89.23
STAR 1-pass ann	83.98	5.28	0.00	0.00	0.00	89.26	89.26
STAR 2-pass	83.23	6.08	0.00	0.00	0.00	89.31	89.31
STAR 2-pass ann	83.26	6.07	0.00	0.00	0.00	89.34	89.34
TopHat v1	75.09	2.68	0.00	11.08	3.21	92.06	84.92
TopHat v1 ann	75.16	2.70	0.00	11.00	3.21	92.07	84.96
TopHat v2	74.51	4.14	0.00	10.51	2.38	91.54	85.10
TopHat v2 ann	76.35	2.71	0.00	10.52	2.18	91.75	85.41
H. Simulation 1							
BAGET ann	96.37	0.12	0.95	2.08	0.18	99.71	98.58
GEM ann	67.92	31.68	0.20	0.11	0.08	100.00	99.90
GEM cons	68.18	31.38	0.20	0.15	0.10	100.00	99.88
GEM cons ann	68.04	31.55	0.20	0.12	0.08	100.00	99.90
GSNAP	94.59	4.54	0.00	0.18	0.05	99.35	99.24
GSNAP ann	94.65	4.49	0.00	0.19	0.05	99.37	99.25
GSTRUCT	94.54	4.60	0.00	0.20	0.04	99.38	99.26
GSTRUCT ann	95.37	3.77	0.00	0.20	0.04	99.38	99.26
MapSplice	95.80	2.06	0.01	1.38	0.08	99.34	98.61
MapSplice ann	95.95	2.06	0.01	1.24	0.08	99.34	98.68
PALMapper	51.06	22.92	23.26	1.30	0.91	99.46	98.35
PALMapper ann	49.88	23.48	24.02	1.21	0.88	99.46	98.42
PALMapper cons	57.35	3.89	7.10	22.26	2.67	93.27	80.81
PALMapper cons ann	62.61	16.14	17.49	2.22	0.78	99.25	97.74
PASS	94.53	0.44	0.23	3.52	0.02	98.73	96.97
PASS cons	93.82	0.44	0.00	3.46	0.00	97.72	95.99
ReadsMap	75.90	2.17	4.29	10.83	0.45	93.64	88.00
SMALT	95.79	0.01	0.25	1.30	0.04	97.39	96.73
STAR 1-pass	95.97	2.80	0.00	0.00	0.00	98.77	98.77
STAR 1-pass ann	95.44	3.41	0.00	0.00	0.00	98.85	98.85
STAR 2-pass	95.36	3.50	0.00	0.00	0.00	98.86	98.86
STAR 2-pass ann	95.18	3.67	0.00	0.00	0.00	98.85	98.85
TopHat v1	90.80	1.98	0.00	5.04	0.27	98.10	95.44
TopHat v1 ann	91.05	2.00	0.00	4.78	0.27	98.10	95.58
TopHat v2	88.00	2.46	0.00	6.64	0.36	97.46	93.96
TopHat v2 ann	88.38	2.45	0.00	5.77	0.26	96.85	93.84
I. Simulation 2							

Continued on next page

Table A.1 – continued from previous page

	Both mates uniquely mapped	Both mates multi -mapped	One mate uniquely and one multi-mapped	One mate uniquely mapped and one unaligned	One mate multi-mapped and one unaligned	Total mapped read pairs	Total mapped reads
BAGET ann	91.36	0.35	2.47	4.66	0.51	99.36	96.77
GEM ann	71.15	27.08	0.58	0.74	0.44	99.99	99.40
GEM cons	71.76	26.38	0.59	0.81	0.45	99.99	99.36
GEM cons ann	71.50	26.72	0.59	0.75	0.44	99.99	99.39
GSNAP	93.95	3.60	0.01	0.65	0.14	98.35	97.95
GSNAP ann	93.97	3.58	0.01	0.68	0.14	98.39	97.97
GSTRUCT	94.11	3.57	0.01	0.71	0.12	98.52	98.11
GSTRUCT ann	94.82	2.87	0.01	0.72	0.11	98.52	98.11
MapSplice	89.26	1.75	0.02	6.88	0.26	98.19	94.61
MapSplice ann	89.59	1.74	0.03	6.61	0.25	98.21	94.79
PALMapper	47.73	19.09	27.68	2.70	1.87	99.06	96.78
PALMapper ann	44.90	20.84	29.17	2.37	1.80	99.08	96.99
PALMapper cons	56.73	5.78	12.15	16.52	3.70	94.88	84.77
PALMapper cons ann	58.91	10.22	21.30	5.85	1.92	98.21	94.32
PASS	83.60	0.39	0.29	11.70	0.05	96.03	90.15
PASS cons	82.52	0.38	0.00	9.15	0.00	92.06	87.48
ReadsMap	73.71	2.06	2.99	14.63	0.81	94.21	86.49
SMALT	94.92	0.01	0.48	1.82	0.04	97.27	96.34
STAR 1-pass	93.36	2.87	0.00	0.00	0.00	96.23	96.23
STAR 1-pass ann	93.33	3.38	0.00	0.00	0.00	96.71	96.71
STAR 2-pass	93.24	3.53	0.00	0.00	0.00	96.77	96.77
STAR 2-pass ann	93.09	3.69	0.00	0.00	0.00	96.77	96.77
TopHat v1	75.36	1.71	0.00	17.25	0.80	95.11	86.09
TopHat v1 ann	76.09	1.74	0.00	16.62	0.79	95.24	86.53
TopHat v2	63.27	1.88	0.00	24.52	1.03	90.70	77.93
TopHat v2 ann	65.70	2.11	0.00	22.70	0.98	91.48	79.64

Table A.2: Mapping accuracy for simulated data (all reads). Results are shown for simulated reads from the nuclear genome, and percentages are relative to the total number of such reads. Perfectly mapped reads have all 76 bases correctly placed (accounting for ambiguity in indel placement as described in Methods). Part correctly mapped reads have at least one base correctly placed, but not all 76. Reads mapped near the correct location are those for which no base is correctly placed, but the mapping overlaps with the correct mapping (this may occur in repetitive regions or indicate a bug in the aligner, as for ReadsMap).

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
A. Simulation 1												
BAGET ann	97.75	85.04	7.63	0.02	90.41	5.18	98.49	85.04	8.24	0.02	90.61	5.23
GEM ann	70.79	68.66	1.98	0.00	70.45	0.31	99.90	93.38	3.53	0.01	96.54	3.29
GEM cons	71.08	68.84	2.09	0.00	70.71	0.33	99.87	93.28	3.61	0.01	96.49	3.30
GEM cons ann	70.93	68.76	2.01	0.00	70.57	0.31	99.89	93.35	3.56	0.01	96.53	3.29
GSNAP	95.65	82.99	12.39	0.01	94.68	0.37	99.23	84.90	12.66	0.01	96.84	1.75
GSNAP ann	95.72	87.04	8.62	0.01	95.30	0.07	99.24	89.07	8.82	0.01	97.52	1.35
GSTRUCT	95.70	87.43	8.08	0.01	95.18	0.19	99.24	89.05	8.24	0.01	96.95	1.95
GSTRUCT ann	96.59	88.27	8.14	0.01	96.08	0.18	99.24	89.65	8.28	0.01	97.59	1.31
MapSplice	96.70	94.34	1.98	0.01	95.94	0.40	98.55	95.22	1.99	0.02	96.83	1.35
MapSplice ann	96.79	94.28	2.15	0.01	96.07	0.39	98.63	95.16	2.16	0.02	96.95	1.34
PALMapper	68.50	67.84	0.63	0.00	68.41	0.09	98.25	91.30	4.20	0.02	95.20	3.05
PALMapper ann	67.57	67.18	0.37	0.00	67.51	0.06	98.33	91.26	3.94	0.02	94.96	3.37
PALMapper cons	73.43	72.39	0.92	0.00	73.28	0.15	80.24	77.34	1.24	0.01	78.54	1.70
PALMapper cons ann	77.83	76.28	1.51	0.00	77.72	0.11	97.62	90.05	5.10	0.01	94.85	2.78
PASS	96.24	48.11	45.10	0.02	90.62	3.08	96.85	48.21	45.21	0.02	90.83	3.46
PASS cons	95.35	48.08	44.60	0.02	90.27	2.75	95.83	48.18	44.71	0.02	90.47	3.01
ReadsMap	83.60	75.22	0.82	3.90	75.95	7.65	87.02	76.40	0.86	3.95	77.15	9.87
SMALT	96.72	73.17	21.80	0.00	91.56	1.89	96.89	73.17	21.90	0.00	91.62	1.92
STAR 1-pass	96.14	84.61	11.20	0.00	94.87	0.47	98.72	85.81	11.35	0.01	96.20	1.70
STAR 1-pass ann	95.56	88.83	6.60	0.00	95.06	0.16	98.81	90.64	6.94	0.01	97.19	1.27
STAR 2-pass	95.48	89.11	6.24	0.00	95.01	0.16	98.82	91.05	6.57	0.01	97.26	1.23
STAR 2-pass ann	95.29	89.00	6.16	0.00	94.84	0.15	98.81	91.08	6.52	0.01	97.26	1.25
TopHat v1	93.37	90.12	1.96	0.02	92.00	1.37	95.23	90.87	2.00	0.02	92.79	2.44
TopHat v1 ann	93.51	90.21	2.00	0.02	92.13	1.37	95.39	90.97	2.05	0.02	92.94	2.45
TopHat v2	91.38	90.41	0.46	0.01	90.84	0.54	93.81	91.45	0.56	0.02	91.96	1.85
TopHat v2 ann	92.00	91.35	0.36	0.01	91.69	0.31	94.62	92.64	0.54	0.02	93.16	1.46
B. Simulation 2												
BAGET ann	94.93	80.35	9.77	0.01	86.98	4.73	96.77	80.35	11.31	0.01	87.49	4.83
GEM ann	71.85	66.36	5.21	0.01	70.84	0.84	99.40	87.23	8.29	0.02	94.33	4.76
GEM cons	72.50	66.77	5.45	0.01	71.44	0.88	99.36	87.06	8.40	0.02	94.25	4.80

Continued on next page

Table A.2 – continued from previous page

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
GEM cons ann	72.21	66.64	5.28	0.01	71.19	0.85	99.39	87.21	8.30	0.02	94.32	4.77
GSNAP	94.28	70.94	22.95	0.01	92.48	0.47	97.95	72.54	23.45	0.01	94.55	2.01
GSNAP ann	94.31	74.66	19.48	0.01	93.18	0.19	97.97	76.35	19.91	0.01	95.27	1.70
GSTRUCT	94.47	76.06	18.01	0.01	93.18	0.42	98.11	77.42	18.33	0.01	94.85	2.34
GSTRUCT ann	95.18	76.75	18.08	0.01	93.95	0.36	98.11	77.96	18.37	0.01	95.43	1.76
MapSplice	92.72	86.05	6.00	0.01	90.55	0.68	94.61	86.94	6.05	0.02	91.46	1.62
MapSplice ann	92.90	86.13	6.11	0.01	90.77	0.70	94.78	87.00	6.16	0.01	91.67	1.64
PALMapper	62.96	61.43	1.43	0.00	62.70	0.26	96.78	85.21	8.47	0.02	93.03	3.74
PALMapper ann	60.71	59.87	0.77	0.00	60.55	0.16	96.99	85.65	7.83	0.02	92.99	4.00
PALMapper cons	71.11	68.47	2.41	0.00	70.77	0.34	84.76	76.60	5.55	0.02	81.91	2.86
PALMapper cons ann	72.53	70.46	1.96	0.00	72.34	0.19	94.32	82.50	8.88	0.02	90.92	3.40
PASS	89.59	25.96	60.51	0.02	80.35	3.04	90.15	26.01	60.64	0.02	80.52	3.38
PASS cons	87.10	25.94	58.58	0.02	79.12	2.60	87.48	25.99	58.71	0.02	79.28	2.80
ReadsMap	82.51	69.79	2.02	7.49	71.54	10.97	86.48	70.87	2.06	7.59	72.65	13.83
SMALT	96.07	64.55	29.38	0.00	90.04	2.07	96.34	64.55	29.55	0.00	90.13	2.10
STAR 1-pass	93.36	72.55	20.39	0.00	90.75	0.62	96.23	73.72	20.74	0.01	92.21	1.96
STAR 1-pass ann	93.33	76.53	16.55	0.00	91.66	0.36	96.71	78.10	17.11	0.01	93.73	1.60
STAR 2-pass	93.24	76.80	16.14	0.00	91.58	0.39	96.77	78.54	16.74	0.01	93.85	1.58
STAR 2-pass ann	93.08	76.85	15.98	0.00	91.51	0.35	96.77	78.67	16.61	0.01	93.90	1.59
TopHat v1	83.98	80.94	2.04	0.01	82.90	1.08	86.09	81.76	2.14	0.01	83.82	2.27
TopHat v1 ann	84.40	81.32	2.08	0.01	83.32	1.07	86.53	82.16	2.19	0.01	84.26	2.27
TopHat v2	75.53	74.31	0.87	0.01	75.13	0.40	77.92	75.29	0.97	0.01	76.18	1.74
TopHat v2 ann	77.05	75.94	0.83	0.01	76.73	0.32	79.65	77.14	1.02	0.01	78.10	1.55

Table A.3: Mapping accuracy for simulated data (spliced reads). Results are shown for simulated spliced reads from the nuclear genome, and percentages are relative to the total number of such reads. Perfectly mapped reads have all 76 bases correctly placed (accounting for ambiguity in indel placement as described in Methods). Part correctly mapped reads have at least one base correctly placed, but not all 76. Reads mapped near the correct location are those for which no base is correctly placed, but the mapping overlaps with the correct mapping (this may occur in repetitive regions or indicate a bug in the aligner, as for ReadsMap).

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
A. Simulation 1												
BAGET ann	91.73	39.23	35.43	0.01	64.14	17.83	94.59	39.23	37.95	0.01	64.96	17.96
GEM ann	21.58	13.33	7.71	0.01	20.19	1.25	99.52	80.46	14.78	0.01	93.57	5.64
GEM cons	22.39	13.58	8.23	0.01	20.84	1.37	99.38	79.94	15.16	0.01	93.31	5.72
GEM cons ann	21.93	13.54	7.84	0.01	20.50	1.27	99.49	80.30	14.91	0.01	93.50	5.67
GSNAP	96.44	64.51	31.00	0.00	93.21	1.44	99.31	65.51	31.61	0.00	94.78	2.68
GSNAP ann	96.82	85.14	11.58	0.00	96.20	0.18	99.36	86.57	11.81	0.00	97.84	1.07
GSTRUCT	95.09	84.77	10.13	0.00	94.45	0.26	99.38	86.57	10.35	0.00	96.46	2.52
GSTRUCT ann	97.33	86.94	10.23	0.00	96.73	0.23	99.37	87.93	10.39	0.00	97.88	1.11
MapSplice	97.09	89.22	7.01	0.00	95.09	0.96	97.46	89.27	7.04	0.00	95.17	1.24
MapSplice ann	97.51	88.89	7.82	0.00	95.69	0.93	97.86	88.91	7.86	0.00	95.75	1.21
PALMapper	35.11	32.21	2.85	0.00	34.77	0.33	98.58	81.36	14.58	0.00	94.77	3.81
PALMapper ann	33.21	31.60	1.59	0.00	33.05	0.16	98.95	83.57	12.49	0.00	95.22	3.74
PALMapper cons	41.10	35.99	4.56	0.00	40.42	0.68	52.14	44.62	5.54	0.00	49.96	2.18
PALMapper cons ann	63.26	55.90	7.30	0.00	62.88	0.39	97.29	75.16	19.30	0.00	93.27	4.02
PASS	92.26	56.31	29.43	0.01	82.78	7.01	92.44	56.35	29.50	0.01	82.86	7.09
PASS cons	91.61	56.31	29.01	0.01	82.52	6.81	91.75	56.34	29.07	0.01	82.59	6.87
ReadsMap	94.52	87.94	2.26	4.19	89.99	4.53	97.44	89.05	2.32	4.24	91.14	6.29
SMALT	96.10	5.52	83.88	0.00	72.96	8.06	96.65	5.52	84.39	0.00	73.27	8.09
STAR 1-pass	96.68	59.73	35.86	0.00	91.57	1.79	98.81	60.31	36.32	0.00	92.53	2.88
STAR 1-pass ann	94.73	82.28	12.26	0.00	93.43	0.35	99.14	84.77	13.53	0.00	97.11	1.03
STAR 2-pass	94.46	83.70	10.47	0.00	93.22	0.41	99.18	86.72	11.66	0.00	97.34	0.95
STAR 2-pass ann	93.82	83.73	9.90	0.00	92.80	0.29	99.16	87.13	11.23	0.00	97.46	0.93
TopHat v1	91.77	78.88	9.43	0.00	87.93	3.84	93.03	79.29	9.64	0.00	88.53	4.50
TopHat v1 ann	92.48	79.36	9.66	0.00	88.63	3.85	93.81	79.82	9.89	0.00	89.30	4.51
TopHat v2	88.01	84.78	1.78	0.00	86.42	1.59	90.02	85.57	2.23	0.00	87.56	2.46
TopHat v2 ann	91.24	90.04	1.08	0.00	91.06	0.18	94.51	91.82	1.75	0.00	93.47	1.04
B. Simulation 2												
BAGET ann	85.25	36.99	33.78	0.01	59.31	14.73	90.25	36.99	38.15	0.01	60.74	14.96
GEM ann	27.43	12.78	13.67	0.03	23.77	2.95	97.25	66.53	25.21	0.05	87.09	8.79
GEM cons	28.68	13.04	14.63	0.04	24.78	3.14	97.02	65.67	25.79	0.05	86.66	8.98

Continued on next page

Table A.3 – continued from previous page

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
GEM cons ann	27.85	13.00	13.87	0.03	24.14	2.99	97.20	66.39	25.29	0.05	87.02	8.82
GSNAP	94.22	51.61	41.43	0.00	89.48	1.66	97.36	52.53	42.26	0.00	91.15	3.02
GSNAP ann	94.43	70.60	23.51	0.00	92.82	0.44	97.45	71.93	23.97	0.00	94.59	1.65
GSTRUCT	93.53	72.18	20.61	0.00	91.67	0.85	97.73	73.62	21.03	0.00	93.50	3.16
GSTRUCT ann	95.04	73.83	20.68	0.00	93.40	0.63	97.72	74.86	21.01	0.00	94.74	1.93
MapSplice	88.03	71.09	15.41	0.00	82.84	1.63	88.44	71.12	15.51	0.00	82.95	1.88
MapSplice ann	88.97	71.60	15.83	0.00	84.03	1.69	89.35	71.61	15.93	0.00	84.12	1.93
PALMapper	30.30	24.79	5.36	0.00	29.51	0.79	95.42	69.75	22.73	0.00	90.24	5.18
PALMapper ann	26.37	23.94	2.37	0.00	26.05	0.32	96.47	75.00	18.43	0.00	92.08	4.39
PALMapper cons	41.40	30.85	9.70	0.00	40.07	1.34	59.63	40.60	15.50	0.00	55.29	4.34
PALMapper cons ann	58.04	50.08	7.82	0.00	57.55	0.49	91.45	65.20	23.19	0.00	86.99	4.47
PASS	78.31	31.92	40.48	0.02	66.53	6.17	78.50	31.94	40.57	0.02	66.61	6.25
PASS cons	75.77	31.92	38.45	0.02	65.27	5.76	75.90	31.94	38.52	0.02	65.34	5.81
ReadsMap	87.63	72.81	4.81	9.70	77.02	10.61	90.82	73.88	4.93	9.82	78.17	12.65
SMALT	94.88	4.13	83.91	0.00	70.66	7.84	95.85	4.13	84.75	0.00	71.16	7.91
STAR 1-pass	91.80	42.50	48.05	0.00	82.96	2.14	94.50	43.11	48.93	0.00	84.26	3.30
STAR 1-pass ann	91.98	63.42	28.13	0.00	87.96	0.80	96.47	65.30	29.87	0.00	91.37	1.69
STAR 2-pass	91.89	65.27	25.99	0.00	87.96	0.95	96.70	67.61	27.80	0.00	91.90	1.62
STAR 2-pass ann	91.34	65.81	25.10	0.00	87.89	0.73	96.71	68.46	26.98	0.00	92.22	1.59
TopHat v1	77.46	66.62	8.04	0.00	74.35	3.11	79.17	67.23	8.38	0.00	75.26	3.91
TopHat v1 ann	79.59	68.59	8.23	0.00	76.49	3.10	81.39	69.26	8.60	0.00	77.50	3.90
TopHat v2	65.76	62.38	2.35	0.00	64.57	1.19	67.56	63.07	2.73	0.00	65.54	2.02
TopHat v2 ann	73.10	70.93	1.99	0.00	72.81	0.29	76.50	72.68	2.67	0.00	75.18	1.32

Table A.4: Mapping accuracy for simulated data (unspliced reads). Results are shown for simulated unspliced reads from the nuclear genome, and percentages are relative to the total number of such reads. Perfectly mapped reads have all 76 bases correctly placed (accounting for ambiguity in indel placement as described in Methods). Part correctly mapped reads have at least one base correctly placed, but not all 76. Reads mapped near the correct location are those for which no base is correctly placed, but the mapping overlaps with the correct mapping (this may occur in repetitive regions or indicate a bug in the aligner, as for ReadsMap).

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
A. Simulation 1												
BAGET ann	99.22	96.23	0.84	0.02	96.83	2.09	99.45	96.23	0.98	0.02	96.88	2.11
GEM ann	82.81	82.18	0.58	0.00	82.73	0.08	99.99	96.53	0.79	0.02	97.27	2.71
GEM cons	82.98	82.34	0.59	0.00	82.89	0.08	99.99	96.53	0.78	0.02	97.27	2.71
GEM cons ann	82.90	82.26	0.59	0.00	82.81	0.08	99.99	96.53	0.78	0.02	97.27	2.71
GSNAP	95.46	87.50	7.85	0.01	95.03	0.11	99.21	89.64	8.03	0.02	97.35	1.52
GSNAP ann	95.45	87.51	7.89	0.01	95.08	0.05	99.21	89.68	8.09	0.02	97.44	1.42
GSTRUCT	95.84	88.08	7.59	0.01	95.36	0.17	99.21	89.66	7.72	0.02	97.07	1.81
GSTRUCT ann	96.40	88.60	7.63	0.01	95.92	0.17	99.21	90.07	7.76	0.02	97.52	1.36
MapSplice	96.61	95.60	0.75	0.01	96.15	0.26	98.82	96.68	0.75	0.02	97.24	1.38
MapSplice ann	96.61	95.60	0.76	0.01	96.16	0.25	98.82	96.68	0.77	0.02	97.25	1.37
PALMapper	76.66	76.55	0.09	0.00	76.62	0.04	98.17	93.73	1.66	0.03	95.31	2.87
PALMapper ann	75.97	75.87	0.07	0.00	75.93	0.04	98.17	93.14	1.85	0.03	94.89	3.28
PALMapper cons	81.33	81.28	0.03	0.00	81.31	0.02	87.11	85.34	0.19	0.01	85.52	1.59
PALMapper cons ann	81.39	81.26	0.09	0.00	81.35	0.04	97.71	93.69	1.62	0.02	95.23	2.47
PASS	97.21	46.10	48.92	0.02	92.53	2.12	97.93	46.23	49.05	0.02	92.78	2.57
PASS cons	96.26	46.06	48.41	0.02	92.16	1.76	96.82	46.18	48.54	0.02	92.40	2.07
ReadsMap	80.93	72.11	0.47	3.83	72.52	8.41	84.47	73.31	0.50	3.88	73.73	10.74
SMALT	96.87	89.70	6.63	0.00	96.11	0.38	96.95	89.70	6.63	0.00	96.11	0.41
STAR 1-pass	96.01	90.69	5.17	0.00	95.68	0.15	98.70	92.04	5.25	0.01	97.10	1.41
STAR 1-pass ann	95.76	90.43	5.22	0.00	95.46	0.12	98.73	92.08	5.33	0.01	97.21	1.32
STAR 2-pass	95.73	90.42	5.21	0.00	95.45	0.10	98.73	92.11	5.33	0.01	97.24	1.30
STAR 2-pass ann	95.65	90.29	5.24	0.00	95.34	0.12	98.73	92.04	5.37	0.01	97.21	1.33
TopHat v1	93.76	92.87	0.13	0.02	93.00	0.77	95.77	93.70	0.14	0.02	93.83	1.94
TopHat v1 ann	93.76	92.86	0.13	0.02	92.99	0.77	95.77	93.70	0.14	0.02	93.83	1.94
TopHat v2	92.20	91.78	0.14	0.01	91.92	0.29	94.73	92.89	0.15	0.02	93.03	1.70
TopHat v2 ann	92.19	91.67	0.19	0.01	91.85	0.34	94.65	92.85	0.25	0.02	93.08	1.57
B. Simulation 2												
BAGET ann	97.27	90.86	3.94	0.01	93.68	2.31	98.35	90.86	4.81	0.01	93.97	2.38
GEM ann	82.62	79.35	3.16	0.00	82.25	0.33	99.92	92.25	4.19	0.02	96.09	3.78
GEM cons	83.12	79.79	3.22	0.00	82.75	0.33	99.92	92.25	4.18	0.02	96.08	3.79

Continued on next page

Table A.4 – continued from previous page

	Uniquely mapped reads						All reads (primary alignment counted)					
	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases	Mapped reads	Perfectly mapped reads	Part correctly mapped reads	Reads mapped near correct location	Correctly mapped bases	Incorrectly mapped bases
GEM cons ann	82.96	79.65	3.20	0.00	82.59	0.33	99.92	92.25	4.18	0.02	96.09	3.78
GSNAP	94.29	75.62	18.47	0.01	93.21	0.18	98.10	77.39	18.89	0.02	95.38	1.77
GSNAP ann	94.29	75.64	18.51	0.01	93.26	0.13	98.10	77.42	18.93	0.02	95.44	1.71
GSTRUCT	94.70	77.00	17.38	0.01	93.55	0.31	98.20	78.34	17.68	0.02	95.17	2.14
GSTRUCT ann	95.21	77.46	17.46	0.01	94.08	0.29	98.20	78.71	17.74	0.02	95.60	1.72
MapSplice	93.85	89.68	3.72	0.01	92.41	0.44	96.11	90.77	3.76	0.02	93.53	1.56
MapSplice ann	93.86	89.65	3.76	0.01	92.40	0.46	96.10	90.73	3.79	0.02	93.51	1.57
PALMapper	70.87	70.31	0.48	0.00	70.75	0.13	97.11	88.95	5.02	0.02	93.71	3.40
PALMapper ann	69.03	68.58	0.38	0.00	68.91	0.12	97.12	88.23	5.27	0.03	93.21	3.91
PALMapper cons	78.31	77.59	0.64	0.00	78.21	0.10	90.86	85.32	3.13	0.02	88.36	2.50
PALMapper cons ann	76.04	75.40	0.55	0.00	75.92	0.12	95.01	86.69	5.41	0.02	91.87	3.14
PASS	92.32	24.52	65.36	0.02	83.71	2.29	92.97	24.58	65.51	0.02	83.89	2.69
PASS cons	89.84	24.50	63.46	0.02	82.47	1.84	90.29	24.55	63.60	0.02	82.65	2.08
ReadsMap	81.27	69.06	1.35	6.95	70.22	11.06	85.43	70.14	1.37	7.05	71.32	14.11
SMALT	96.36	79.19	16.17	0.00	94.73	0.67	96.45	79.19	16.17	0.00	94.73	0.69
STAR 1-pass	93.74	79.84	13.68	0.00	92.64	0.26	96.65	81.13	13.90	0.01	94.14	1.64
STAR 1-pass ann	93.66	79.71	13.74	0.00	92.55	0.25	96.77	81.20	14.02	0.01	94.30	1.58
STAR 2-pass	93.56	79.59	13.76	0.00	92.45	0.26	96.79	81.18	14.07	0.01	94.33	1.57
STAR 2-pass ann	93.51	79.52	13.77	0.00	92.39	0.26	96.79	81.14	14.09	0.01	94.31	1.59
TopHat v1	85.55	84.41	0.59	0.02	84.97	0.58	87.77	85.29	0.63	0.02	85.89	1.88
TopHat v1 ann	85.56	84.41	0.59	0.02	84.98	0.58	87.78	85.29	0.63	0.02	85.90	1.87
TopHat v2	77.90	77.20	0.51	0.01	77.69	0.21	80.44	78.25	0.54	0.02	78.76	1.68
TopHat v2 ann	78.01	77.15	0.55	0.01	77.68	0.33	80.41	78.22	0.62	0.02	78.80	1.61

Table A.5: Accuracy of junction discovery on simulated data. Number of unique junctions reported for the two simulated data sets, at a range of thresholds (1-7) for the number of primary alignments supporting a junction. Higher thresholds correspond to a more conservative interpretation of alignment results. Junctions were classified as true and false by comparison to the true simulated alignments. The row labeled Truth shows the result expected for a perfect aligner.

	True junctions							False junctions						
	1	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A. Simulation 1														
BAGET ann	77534	73135	69946	67551	65549	63739	62133	5141	3395	2751	2371	2148	1970	1809
GEM ann	116615	110636	105531	101265	97581	94269	91300	9191	5090	3887	3238	2774	2488	2266
GEM cons	97700	97566	97408	96169	94150	91740	89266	4689	3479	2986	2638	2330	2125	1956
GEM cons ann	108107	104597	101922	99273	96423	93548	90804	6250	4192	3456	2999	2593	2345	2142
GSNAP	118833	110948	104977	100268	96218	92800	89724	13284	5127	3813	3102	2612	2320	2096
GSNAP ann	120819	113862	108590	104301	100617	97363	94504	18537	8207	5821	4545	3783	3275	2862
GSTRUCT	119587	112317	106795	103806	100704	97789	95044	8887	3529	2830	2432	2183	2016	1869
GSTRUCT ann	119781	112837	108495	104822	101316	98144	95322	8445	3206	2521	2171	1924	1743	1599
MapSplice	115690	111331	106663	102584	99019	95922	93075	4070	1970	1595	1348	1190	1072	991
MapSplice ann	119040	112564	107469	103222	99589	96460	93553	22445	6504	3917	2911	2369	2066	1862
PALMapper	117215	110115	105936	102172	98686	95554	92752	283031	68953	41034	29074	22520	18426	15638
PALMapper ann	118661	111717	107420	103688	100221	97047	94200	325926	78720	49656	37115	29989	25269	21977
PALMapper cons	106357	102086	95731	90320	85874	81982	78553	7268	4538	3554	3032	2691	2421	2240
PALMapper cons ann	108259	107511	105181	101961	98241	94968	91998	43228	28942	23058	19389	16954	15107	13702
PASS	114022	105801	99746	94887	90901	87486	84487	62597	16756	10398	7824	6304	5291	4682
PASS cons	113836	105711	99698	94842	90869	87454	84451	37285	12524	8435	6605	5485	4721	4220
ReadsMap	114152	109814	105454	101663	98322	95362	92695	898709	421113	272287	199815	156863	128490	108553
SMALT	50497	41008	35546	31504	28431	25900	23692	140685	92591	77404	67930	60578	54614	49584
STAR 1-pass	116239	107930	101800	96987	92897	89403	86335	6525	2562	2081	1795	1603	1456	1356
STAR 1-pass ann	119013	111395	105573	100954	97036	93665	90791	20220	10055	7322	5831	4870	4181	3677
STAR 2-pass	117085	112014	107278	103202	99619	96518	93727	11575	5088	3789	3105	2640	2327	2092
STAR 2-pass ann	119227	113384	108426	104254	100620	97498	94669	21198	10323	7304	5775	4764	4065	3569
TopHat v1	108787	105603	101945	98448	95190	92247	89519	7701	5590	4512	3857	3424	3089	2827
TopHat v1 ann	113187	108603	104173	100272	96756	93679	90818	8366	6175	5041	4304	3820	3429	3180
TopHat v2	109676	106506	102743	99095	95858	93023	90461	7888	5563	4469	3845	3404	3041	2788
TopHat v2 ann	115957	111121	106711	102839	99398	96426	93769	24324	14579	10352	8035	6570	5572	4816
Truth	122745	116040	110976	106744	103132	99965	97158	0	0	0	0	0	0	0

Continued on next page

Table A.5 – continued from previous page

	True junctions							False junctions						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
B. Simulation 2														
BAGET ann	76962	72959	70253	68045	66147	64426	62808	6354	3910	3085	2658	2421	2242	2063
GEM ann	112360	107048	102924	99412	96309	93360	90819	22292	12427	9024	7213	6154	5368	4820
GEM cons	91374	91235	91121	90396	89195	87536	85811	12621	8403	6631	5539	4854	4320	3908
GEM cons ann	105417	101642	99134	96816	94557	92150	89914	14780	9433	7314	6059	5271	4663	4214
GSNAP	119283	112397	107561	103605	100140	97025	94184	30687	7713	5394	4324	3651	3238	2904
GSNAP ann	121426	115409	111207	107760	104783	102064	99575	36633	11356	7860	6122	5071	4369	3834
GSTRUCT	119921	113669	109265	106978	104639	102231	99928	23066	5776	4305	3671	3242	2983	2746
GSTRUCT ann	120219	114373	111154	108143	105365	102726	100331	22609	5372	3989	3377	2942	2674	2467
MapSplice	109651	106509	102957	99736	96811	94145	91687	9306	4601	3771	3314	2914	2682	2473
MapSplice ann	116473	110276	105742	101987	98721	95808	93193	33960	11202	7265	5771	5000	4495	4121
PALMapper	115695	109803	106569	103578	100885	98240	95805	383907	77630	47548	34827	27831	23501	20511
PALMapper ann	118153	112207	108634	105591	102760	100085	97664	528205	103293	63774	48093	39406	33876	29784
PALMapper cons	103942	102085	98624	94689	90864	87283	84025	12256	9061	7605	6709	6087	5578	5170
PALMapper cons ann	105895	105479	104427	102706	100487	97938	95346	59112	41785	34466	29870	26572	24055	21986
PASS	107839	100326	95229	91086	87382	84197	81314	125286	31170	19365	14722	12252	10594	9386
PASS cons	107564	100131	95032	90855	87144	83990	81059	77357	24451	16072	12485	10534	9203	8256
ReadsMap	109056	105550	102275	99313	96644	94126	91780	942675	415584	259001	184536	141592	113821	94463
SMALT	50726	41116	35239	30947	27516	24694	22277	181841	103528	82418	70436	61738	54887	49302
STAR 1-pass	110302	102852	97694	93452	89795	86566	83641	14892	4525	3471	2883	2563	2322	2130
STAR 1-pass ann	116775	109904	105141	101334	97941	94936	92329	31692	12483	9021	7239	6063	5289	4655
STAR 2-pass	113032	108903	105346	102261	99503	96811	94492	22749	8721	6366	5282	4581	4082	3690
STAR 2-pass ann	117148	112023	108148	104872	101968	99188	96774	32851	14116	9978	8051	6782	5913	5236
TopHat v1	101398	98845	96123	93387	90865	88453	86205	11376	8217	6551	5535	4871	4347	3947
TopHat v1 ann	108927	104905	101273	98039	95126	92462	89997	12267	8942	7192	6072	5316	4732	4300
TopHat v2	104281	101660	98663	95569	92708	90091	87729	9560	7236	6024	5278	4720	4265	3920
TopHat v2 ann	113572	109180	105497	102151	99131	96423	93939	26381	16509	12193	9784	8177	7057	6225
Truth	123581	117890	113826	110530	107667	105088	102713	0	0	0	0	0	0	0

Table A.6: Number of introns reported per alignment. There were no alignments with more than five introns.

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
												A. K562 whole cell replicate 1
BAGET ann	187077254	19044438	0	0	0	0	190657737	19044438	0	0	0	0
GEM ann	171310851	34730775	1064776	1509	0	0	499923501	76562905	13247099	429798	3108	0
GEM cons	171601899	34567687	898840	209	0	0	494153538	69729534	8070043	378798	2443	0
GEM cons ann	171519716	34628760	938481	1012	0	0	495520648	72100675	8478289	385549	2445	0
GSNAP	169869232	36780143	403327	2387	14	0	222746918	39531697	411451	2541	14	0
GSNAP ann	155959666	49495787	1595726	11463	35	0	207664707	53698921	1707068	12197	36	0
GSTRUCT	160220367	45290527	1586733	9674	38	0	222315512	54330580	1848701	10801	39	0
GSTRUCT ann	157984101	47426882	1683304	10286	37	0	209374754	54242555	1855777	11163	39	0
MapSplice	155715733	40229115	1062125	1353	0	0	158023362	40229125	1062125	1353	0	0
MapSplice ann	156017042	39908235	1068363	1869	0	0	158330301	39908238	1068363	1869	0	0
PALMapper	158313084	43659317	2733642	0	0	0	1615733787	98720318	9598992	0	0	0
PALMapper cons	60036227	8633289	25053	0	0	0	145387747	19789755	82347	0	0	0
PASS	171743999	25222924	115456	6	0	0	173860480	26029419	115739	6	0	0
PASS cons	167272372	24802361	114869	6	0	0	168590407	25496149	115053	6	0	0
ReadsMap	115308436	47132457	2282147	61828	135	1	142468047	54354636	2811288	104747	158	1
SMALT	198636832	6370637	0	0	0	0	200297027	6370637	0	0	0	0
STAR 1-pass	171619910	31320692	260829	357	0	0	189704784	33803026	272426	374	0	0
STAR 1-pass ann	153632518	48484160	1426139	35255	16	0	168619810	52590167	1583953	37982	20	0
STAR 2-pass	150202827	51651315	1653897	39505	68	0	168220716	62174408	2662407	52111	164	0
STAR 2-pass ann	150004846	51809144	1671503	40203	70	0	168011972	62573092	2707290	53504	168	0
TopHat v1	155458487	31685988	1321699	4249	2	0	169288810	32281213	1508011	7658	6	0
TopHat v1 ann	155458440	31765329	1311500	7347	9	0	169300835	32399210	1499216	12984	14	0
TopHat v2	147253729	36710072	1204310	6029	3	0	163857967	38962028	1249600	6498	4	0
TopHat v2 ann	140335591	46878285	1637296	27952	23	0	154155848	51536690	1841780	55839	46	0
												B. K562 whole cell replicate 2
BAGET ann	197805101	17665445	0	0	0	0	204566639	17665445	0	0	0	0
GEM ann	185275025	33132187	851569	15660	0	0	539610202	69130436	8433960	388289	215	0
GEM cons	185615881	32821350	795067	209	0	0	533876729	61377573	4873485	204158	2170	0
GEM cons ann	185536089	32882879	833009	861	0	0	535319029	63471618	5318079	228545	2170	0
GSNAP	182769289	36179016	364673	2077	8	0	246051100	38691591	373129	2257	10	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
GSNAP ann	167694005	50170089	1526399	11917	25	0	229895554	53925349	1628445	13450	27	0
GSTRUCT	173861658	44234621	1412799	9059	22	0	262536766	54202009	1640658	10504	22	0
GSTRUCT ann	172181493	45847374	1466966	10111	28	0	254420149	54043823	1657727	11527	28	0
MapSplice	161505723	36819919	910651	1197	0	0	167168530	36819931	910651	1197	0	0
MapSplice ann	161906407	36398227	886835	1465	0	0	167579746	36398233	886835	1465	0	0
PALMapper	164821984	44270272	3460350	0	0	0	1692507126	102137645	11501769	0	0	0
PALMapper cons	107664915	15415151	67570	0	0	0	107664915	15415151	67570	0	0	0
PASS	176544216	22968986	102988	3	0	0	178806957	23480856	103265	6	0	0
PASS cons	165095387	22162161	102235	4	0	0	166432987	22527135	102409	6	0	0
SMALT	209130532	6127506	0	0	0	0	210307069	6127506	0	0	0	0
STAR 1-pass	181284498	30131002	236178	252	0	0	201282547	32495226	246466	265	0	0
STAR 1-pass ann	163348660	48075196	1316387	37029	6	0	180032268	52109009	1454382	40190	13	0
STAR 2-pass	159706127	51458084	1556531	42043	41	0	179508185	62397070	2526198	55231	116	0
STAR 2-pass ann	159514745	51624217	1571320	42549	43	0	179324495	62854431	2567410	56279	121	0
TopHat v1	149782758	29146862	1049421	4511	28	0	164950835	29832399	1242963	19291	56	1
TopHat v1 ann	149783322	29183515	1072472	6439	24	0	164962013	29912688	1282587	24737	42	0
TopHat v2	139741709	33321108	1028619	5325	0	0	156890641	35541225	1062079	5479	0	0
TopHat v2 ann	132677412	42947041	1435715	29132	6	0	146628377	46898096	1615166	56043	7	0
												C. K562 cytoplasmic fraction replicate 1
BAGET ann	208521101	28241977	0	0	0	0	212070658	28241977	0	0	0	0
GEM ann	197702900	40606845	1134440	1230	0	0	493631577	77253508	9906030	216131	1004	0
GEM cons	197967587	40352093	1087093	190	0	0	488624837	69430179	7101212	101765	426	0
GEM cons ann	197887670	40413389	1125713	1137	0	0	489553614	71590585	7396418	94242	162	0
GSNAP	195321629	43621344	451596	2673	15	0	250621888	46533423	460979	2806	17	0
GSNAP ann	179534877	58044452	1822973	9814	24	0	233740631	62778133	1962993	10388	24	0
GSTRUCT	188613775	49280508	1717291	8100	26	0	257506322	64135677	2357715	9061	26	0
GSTRUCT ann	185601418	52112656	1894798	9080	28	0	249941538	64060885	2380522	9518	28	0
MapSplice	180858050	49673933	1318983	1094	0	0	182639049	49673956	1318983	1094	0	0
MapSplice ann	181889965	48648848	1276719	2309	0	0	183671968	48648855	1276719	2309	0	0
PASS	200240345	31338049	142379	7	0	0	201821496	31749762	142891	7	0	0
PASS cons	196768072	30949027	141723	6	0	0	197592326	31222304	142046	6	0	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
ReadsMap	140851097	60522677	2872481	107427	28	0	178414125	69875721	3461753	234391	39	0
SMALT	220887722	7544213	0	0	0	0	221684991	7544213	0	0	0	0
STAR 1-pass	195727486	37897252	324449	373	0	0	215758379	40394998	335875	386	0	0
STAR 1-pass ann	173885738	58529123	1736876	127424	11	0	191298592	63538256	1915457	129617	12	0
STAR 2-pass	166816809	65075051	2176309	136258	54	1	190222297	81106244	3686507	156631	159	2
STAR 2-pass ann	166571761	65289207	2186183	137206	58	1	190002411	81790333	3749852	158565	163	2
TopHat v1	179661623	36831248	1344808	6148	3	0	195310467	37339628	1491492	13720	9	0
TopHat v1 ann	179660581	36913991	1351375	6896	4	0	195359319	37471945	1492452	12005	9	0
TopHat v2	170911303	44935030	1410915	5034	0	0	195605668	49344971	1519378	6374	0	0
TopHat v2 ann	161001926	57456987	2047066	116223	34	0	181974554	64861853	2429409	212221	51	0
												D. K562 cytoplasmic fraction replicate 2
BAGET ann	146679343	18913128	0	0	0	0	149493992	18913128	0	0	0	0
GEM ann	133664041	33050924	878073	1183	0	0	364341755	69827925	9762875	283774	227	0
GEM cons	133880600	32852609	832767	227	0	0	360205170	63087139	7306424	116213	60	0
GEM cons ann	133794150	32919330	871043	968	0	0	361093746	65076885	7579619	117473	75	0
GSNAP	132905278	34363515	337927	1960	3	0	165065576	36381786	343986	2063	3	0
GSNAP ann	121020777	45309882	1287171	6066	16	0	152256757	48498217	1357665	6492	16	0
GSTRUCT	126275712	40116144	1346547	5748	18	0	163482441	49159012	1687419	6252	18	0
GSTRUCT ann	124224907	42031798	1480673	5994	22	0	157225878	49052558	1718690	6442	22	0
MapSplice	122865216	39932973	961603	642	0	0	123668821	39932980	961603	642	0	0
MapSplice ann	123621200	39143336	939761	1171	0	0	124425977	39143339	939761	1171	0	0
PASS	136440767	25599042	103327	2	0	0	137655948	25954539	103689	2	0	0
PASS cons	133861132	25334401	102917	2	0	0	134571977	25602161	103163	2	0	0
SMALT	154486033	6245646	0	0	0	0	155123972	6245646	0	0	0	0
STAR 1-pass	133548716	29927431	240562	341	0	0	146845730	32174114	249305	345	0	0
STAR 1-pass ann	117478919	45195504	1209188	49421	8	0	129335370	49198277	1335727	50591	9	0
STAR 2-pass	113072761	49193805	1555212	54778	50	0	128635253	60047230	2680991	67904	86	0
STAR 2-pass ann	112880538	49349458	1563581	55596	53	0	128413460	60494713	2731292	69343	90	0
TopHat v1	120392290	30257696	1035799	4436	184	0	130952871	30719274	1157794	10051	2135	0
TopHat v1 ann	120391265	30338687	1041674	5541	142	0	130960638	30820610	1166744	11262	1586	0
TopHat v2	113472066	36323162	1078664	3395	1	0	130552639	39703856	1187823	3773	6	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
TopHat v2 ann	108225194	44237331	1446875	15471	18	0	124134286	49790480	1737874	22532	18	0
												E. K562 nuclear fraction replicate 1
BAGET ann	206705594	15431582	0	0	0	0	210513852	15431582	0	0	0	0
GEM ann	197660121	25731659	603640	457	1	0	538778273	55458548	6666902	142307	564	0
GEM cons	198011964	25376373	571676	143	0	0	534612925	48611961	3671708	58066	18	0
GEM cons ann	197935331	25446958	591336	397	0	0	535168811	49677122	3944727	62174	18	0
GSNAP	198639580	24590060	262861	1334	16	0	256546541	26582431	274731	1408	16	0
GSNAP ann	190841927	31745811	918286	6212	26	0	248218509	34499476	1018447	6398	26	0
GSTRUCT	191751353	30929967	925084	4376	24	0	234821855	34917124	1039863	5648	25	0
GSTRUCT ann	191009154	31640474	955681	4636	23	0	232797536	34747221	1045346	5741	23	0
MapSplice	188915608	27849966	693371	1118	2	0	190811556	27849975	693371	1118	2	0
MapSplice ann	188993965	27766670	696752	975	8	0	190883732	27766675	696752	975	8	0
PALMapper	186629755	30952119	1738721	0	0	0	2554195311	69199335	4910535	0	0	0
PALMapper cons	130671178	12853961	42233	0	0	0	173449918	15341457	51658	0	0	0
PASS	195909426	21349613	88759	11	0	0	198364701	21972613	89174	11	0	0
PASS cons	192644938	20923625	88109	7	0	0	194283800	21322859	88327	7	0	0
ReadsMap	141145652	37445251	2033113	54916	516	0	167139782	43692129	2500076	74332	699	0
SMALT	213355159	5407107	0	0	0	0	214209294	5407107	0	0	0	0
STAR 1-pass	195898559	21759750	167482	119	0	0	212012843	23263830	184064	133	0	0
STAR 1-pass ann	186225918	30943289	827329	13566	2	0	201653603	33228051	945262	16152	2	0
STAR 2-pass	183252391	33771914	974031	17624	25	1	200881258	40044714	1515066	30013	62	1
STAR 2-pass ann	183132331	33871096	981924	17919	25	1	200789609	40384352	1543146	30642	64	1
TopHat v1	180403610	23968166	795467	4044	18	0	194534527	24543719	907738	7399	51	0
TopHat v1 ann	180403717	24002859	809383	4819	18	0	194547983	24605666	924258	8167	57	0
TopHat v2	176258365	27079714	765656	3311	1	0	195355523	29138021	812726	4028	1	0
TopHat v2 ann	173129155	30494610	966483	12655	15	0	188174689	33348476	1105695	21125	22	0
												F. K562 nuclear fraction replicate 2
BAGET ann	183216474	13830210	0	0	0	0	186375573	13830210	0	0	0	0
GEM ann	174331871	23282063	510180	487	0	0	495481257	49632919	5445987	144699	723	0
GEM cons	174618785	22998748	475199	115	0	0	492567840	44678365	2919699	68430	7	0
GEM cons ann	174535471	23073017	497136	453	0	0	493060432	45622893	3173625	71260	30	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
GSNAP	175703134	21897666	218107	1175	18	0	228881337	23886945	228125	1219	21	0
GSNAP ann	169165577	27931661	732314	4283	47	0	221805293	30627969	813255	4383	50	0
GSTRUCT	169585880	27600513	745440	3448	45	0	209168236	30808839	819232	4025	46	0
GSTRUCT ann	169136046	28030982	763899	3681	46	0	207369719	30707614	815426	4146	46	0
MapSplice	167648664	25235898	579628	696	0	0	169001827	25235908	579628	696	0	0
MapSplice ann	167680834	25198102	583636	875	0	0	169027659	25198107	583636	875	0	0
PALMapper	164784505	27454461	1464692	0	0	0	2648436542	59977471	3825122	0	0	0
PALMapper cons	125464945	12642822	36162	0	0	0	169630360	15090565	44023	0	0	0
PASS	172713017	20332388	82128	7	0	0	175158312	20940060	82626	9	0	0
PASS cons	170121237	19986490	81579	6	0	0	171768050	20398941	81940	7	0	0
SMALT	190114114	5057768	0	0	0	0	190958188	5057768	0	0	0	0
STAR 1-pass	173679550	19741734	144171	139	0	0	188503852	21223846	159341	143	0	0
STAR 1-pass ann	165699327	27329037	664747	6623	16	0	180111098	29433467	761984	8159	28	0
STAR 2-pass	163151433	29763532	782416	10236	39	0	179388811	35235658	1224137	17880	70	0
STAR 2-pass ann	163046387	29852674	788567	10554	40	0	179313087	35576486	1246049	18437	77	0
TopHat v1	159567534	22051237	652260	3266	8	0	170842949	22579102	740729	6198	39	0
TopHat v1 ann	159567334	22085062	668529	4089	11	0	170850087	22632500	760666	6961	44	0
TopHat v2	157907882	24910089	648982	2180	8	0	173381001	26705557	690659	2610	8	0
TopHat v2 ann	155678376	27241854	788145	6765	30	0	167766127	29760695	899188	9923	32	0
												G. Mouse brain
BAGET ann	103801100	5664502	0	0	0	0	106626404	5664502	0	0	0	0
GEM ann	104783501	7457537	221379	1310	0	0	1680600227	13138441	651413	4024	2	0
GEM cons	104968184	7288614	186078	2	0	0	1680148684	11536810	433152	47	0	0
GEM cons ann	104887829	7350261	219901	1306	0	0	1680274464	11796821	525613	3358	2	0
GSNAP	103644953	6862434	104655	1095	12	0	184900539	7565613	105529	1102	12	0
GSNAP ann	101540035	8778287	296948	3678	72	0	182680583	9705444	311914	3688	72	0
GSTRUCT	101873894	9269701	294801	3204	54	0	172915091	10623869	323401	3443	54	0
GSTRUCT ann	101706267	9423372	300023	3347	58	0	173240656	10703268	325579	3589	58	0
MapSplice	99314039	7907439	230820	261	0	0	103798423	7907441	230820	261	0	0
MapSplice ann	99269151	7976595	245448	1465	1	0	103777468	7976598	245448	1465	1	0
PASS	99405349	6668953	38963	8	0	0	102761515	6822469	39213	8	0	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
PASS cons	96704058	6527373	38602	8	0	0	97577969	6625437	38765	8	0	0
ReadsMap	71638776	11207219	435295	3756	59	0	105324775	12711034	558163	6808	59	0
SMALT	103786857	1723772	0	0	0	0	104394448	1723772	0	0	0	0
STAR 1-pass	95776345	6210490	69044	209	0	0	103446176	6367339	69751	210	0	0
STAR 1-pass ann	94157652	7746192	185813	1914	47	0	101968585	8188551	192088	1961	53	0
STAR 2-pass	92751597	9109186	283715	4850	72	0	100940213	10100321	309283	6020	75	0
STAR 2-pass ann	92573028	9304825	295568	5523	82	0	100633165	10421178	325112	6749	94	0
TopHat v1	89822037	7075370	224883	1230	0	0	97517105	7113106	233819	1359	0	0
TopHat v1 ann	89822004	7109187	244436	2428	62	0	97520713	7159649	253887	2588	64	0
TopHat v2	89216210	7853672	258284	1668	0	0	101874048	8062729	262267	1711	0	0
TopHat v2 ann	88490497	8865297	321994	4119	88	0	98587259	9357618	341839	4300	88	0
												H. Simulation 1
BAGET ann	71619289	7243380	0	0	0	0	72438310	7243380	0	0	0	0
GEM ann	67895614	11620668	404930	2038	0	0	165877216	21162135	1610192	11556	22	0
GEM cons	67993410	11540114	369445	285	0	0	165099634	19616501	1116746	2305	0	0
GEM cons ann	67923682	11592880	400762	1902	0	0	165500190	20423970	1355142	7442	19	0
GSNAP	68655070	10552570	183396	2029	19	0	80840006	10999735	187525	2031	19	0
GSNAP ann	65523836	13357541	511492	8496	193	0	77578022	13982407	536732	8498	193	0
GSTRUCT	65645586	13235700	517317	8182	173	0	72687872	13849148	531732	8207	173	0
GSTRUCT ann	65455071	13417975	524927	8385	173	0	71751504	13774977	531934	8388	173	0
MapSplice	65759487	12673956	450754	2255	7	0	71522223	12682474	450844	2255	7	0
MapSplice ann	65654189	12822124	466648	2542	8	0	71412112	12829010	466776	2542	8	0
PALMapper	65223152	13245519	211792	74	8	0	650778760	34807250	296256	276	21	0
PALMapper ann	64762109	13687402	285014	60	8	0	649459058	37952667	447335	276	21	0
PALMapper cons	58013224	6615783	17596	0	0	0	86278432	8411815	19808	0	0	0
PALMapper cons ann	66026622	11986230	182774	10	0	0	308967867	19314514	225714	32	2	0
PASS	67661818	9840645	71539	27	0	0	68505088	9924704	71705	27	0	0
PASS cons	66930406	9789895	71387	27	0	0	67528822	9830463	71431	27	0	0
ReadsMap	53769449	15043852	1461515	121373	4855	30	60544097	15956506	1583285	132148	5132	30
SMALT	74725141	2655012	0	0	0	0	74854021	2655012	0	0	0	0
STAR 1-pass	69477331	9422613	116704	346	0	0	73248740	9807635	121345	346	0	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
STAR 1-pass ann	65688823	12927586	452392	8511	228	0	69505024	13655950	495790	9510	233	1
STAR 2-pass	65533283	13070836	472720	8758	209	0	69397907	13798019	530437	10221	226	0
STAR 2-pass ann	65396794	13190256	486595	9129	220	0	69287811	14134611	559707	10894	238	1
TopHat v1	64563528	11337574	443740	6547	41	0	66733540	11483082	465724	6915	41	0
TopHat v1 ann	64560841	11429576	465337	7337	142	0	66734561	11583780	488154	7702	153	0
TopHat v2	62535958	12139147	483319	7601	45	0	66201271	12550205	500929	7705	45	0
TopHat v2 ann	61436183	13067748	555565	10513	292	0	64401859	13755211	606975	11567	295	0
												I. Simulation 2
BAGET ann	70075321	7339683	0	0	0	0	72248546	7339683	0	0	0	0
GEM ann	68175403	11000317	343047	2378	5	0	168057919	21101019	1893735	14365	32	0
GEM cons	68381701	10800863	302917	436	0	0	167084140	19138153	1371492	5152	0	0
GEM cons ann	68230123	10945100	337369	2119	5	0	167467722	19973685	1596525	9704	16	0
GSNAP	67502298	10688746	170716	1275	13	0	80649602	11221314	176598	1283	13	0
GSNAP ann	63884518	13975304	512648	7390	102	2	76939310	14761371	554446	7409	102	2
GSTRUCT	64091183	13880674	509423	6581	62	1	71586151	14577724	535411	6624	62	1
GSTRUCT ann	63903567	14056596	518470	6965	95	1	70899214	14590806	538133	6975	95	1
MapSplice	63972131	11366972	350883	1924	6	0	69298637	11373139	351091	1924	6	0
MapSplice ann	63736012	11730852	358910	2314	14	0	69046142	11734861	359129	2318	14	0
PALMapper	63372043	13871065	180109	48	2	0	615694813	43911626	306150	335	14	0
PALMapper ann	62590372	14732160	272967	45	2	0	615958105	51083521	562036	335	14	0
PALMapper cons	60732770	7062992	19571	0	0	0	168635025	9846617	22824	0	0	0
PALMapper cons ann	62858130	12438785	157020	3	0	0	276166832	21541293	230330	5	0	0
PASS	63588965	8487179	45659	11	0	0	64555002	8590501	45820	11	0	0
PASS cons	61639107	8302610	45296	11	0	0	62249886	8347417	45362	11	0	0
ReadsMap	53673659	14854263	649065	12091	143	0	61876899	16658795	776727	13173	159	0
SMALT	74549890	2520940	0	0	0	0	74761393	2520940	0	0	0	0
STAR 1-pass	68751704	8161114	74265	83	0	0	73570492	8673633	80486	85	0	0
STAR 1-pass ann	64558922	12434337	369404	5882	83	0	69221948	13312964	417012	6571	85	0
STAR 2-pass	64293445	12714005	401107	6073	70	0	69034173	13549823	460189	7024	79	0
STAR 2-pass ann	64051280	12942089	419355	6527	79	0	68779769	14066697	497348	7696	88	0
TopHat v1	57899907	10586748	382874	4434	1	0	60653984	10893767	425260	5195	2	0

Continued on next page

Table A.6 – continued from previous page

	Primary alignments						All alignments					
	0 introns	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**0 introns**	**1 intron**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
TopHat v1 ann	57893669	10914044	412454	5855	44	0	60648703	11232934	458045	6815	47	0
TopHat v2	52071715	9893020	370824	4710	8	0	56467249	10487685	397883	4826	8	0
TopHat v2 ann	51775131	11460488	470507	8668	155	0	55456172	12423060	552738	9462	180	0

Table A.7: Accuracy of multi-intron alignments. For each intron count n, the tabulated percentages were computed as follows: recall = number of primary alignments with at least n correctly identified introns / number of simulated reads with at least n introns; precision = number of primary alignments with at least n correctly identified introns / number of primary alignments with at least n reported introns. The number of simulated reads with n introns is given on the last row of each table. Precision is n.a. (not applicable) where no aligments were reported.

	Recall					Precision				
	1 introns	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**1 introns**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
A. Simulation 1										
BAGET ann	39.50	0.00	0.00	0.00	0.00	78.70	n.a.	n.a.	n.a.	n.a.
GEM ann	82.20	65.20	15.80	0.00	0.00	98.60	98.30	97.60	n.a.	n.a.
GEM cons	81.60	59.70	2.30	0.00	0.00	98.90	98.70	97.50	n.a.	n.a.
GEM cons ann	82.00	64.60	14.80	0.00	0.00	98.70	98.50	98.80	n.a.	n.a.
GSNAP	73.60	29.10	11.40	3.10	0.00	98.80	95.70	68.50	89.50	n.a.
GSNAP ann	94.90	82.20	67.60	34.20	0.00	98.70	96.50	95.90	96.90	n.a.
GSTRUCT	94.10	83.90	65.80	30.70	0.00	98.70	97.50	97.10	97.10	n.a.
GSTRUCT ann	95.60	85.20	66.20	30.70	0.00	98.90	97.50	95.30	97.10	n.a.
MapSplice	90.30	72.50	17.90	1.30	0.00	99.30	97.80	97.70	100.00	n.a.
MapSplice ann	90.00	74.10	19.20	1.30	0.00	97.70	96.40	93.00	87.50	n.a.
PALMapper	85.40	24.40	0.10	0.00	0.00	91.80	70.40	19.50	0.00	n.a.
PALMapper ann	87.20	36.50	0.10	0.00	0.00	90.30	78.30	20.60	0.00	n.a.
PALMapper cons	45.00	2.80	0.00	0.00	0.00	98.00	98.20	n.a.	n.a.	n.a.
PALMapper cons ann	78.60	28.70	0.10	0.00	0.00	93.40	95.70	80.00	n.a.	n.a.
PASS	66.70	11.30	0.20	0.00	0.00	97.00	96.70	100.00	n.a.	n.a.
PASS cons	66.60	11.20	0.20	0.00	0.00	97.50	96.80	100.00	n.a.	n.a.
ReadsMap	89.70	82.70	72.10	40.80	0.00	77.80	31.80	7.00	4.60	0.00
SMALT	6.20	0.00	0.00	0.00	0.00	33.50	n.a.	n.a.	n.a.	n.a.
STAR 1-pass	65.40	19.00	2.80	0.00	0.00	99.00	99.00	99.40	n.a.	n.a.
STAR 1-pass ann	90.70	71.90	64.70	40.20	0.00	97.80	95.40	91.30	96.50	n.a.
STAR 2-pass	92.60	76.80	70.70	38.20	0.00	98.60	97.30	97.20	100.00	n.a.
STAR 2-pass ann	93.00	78.40	72.40	40.00	0.00	98.10	96.70	95.60	99.50	n.a.
TopHat v1	80.00	67.40	47.60	7.50	0.00	98.00	91.40	89.20	100.00	n.a.
TopHat v1 ann	80.60	70.20	54.30	21.60	0.00	97.80	90.70	89.40	83.10	n.a.
TopHat v2	86.00	78.30	60.90	8.20	0.00	98.20	97.40	98.20	100.00	n.a.
TopHat v2 ann	92.30	87.40	79.30	52.70	0.00	97.70	94.50	92.40	98.60	n.a.

Continued on next page

Table A.7 – continued from previous page

	Recall					Precision				
	1 introns	**2 introns**	**3 introns**	**4 introns**	**5 introns**	**1 introns**	**2 introns**	**3 introns**	**4 introns**	**5 introns**
Number of simulated reads	13808336	598297	11781	493	54					
B. Simulation 2										
BAGET ann	37.90	0.00	0.00	0.00	0.00	80.50	n.a.	n.a.	n.a.	n.a.
GEM ann	70.70	50.80	13.80	0.00	0.00	97.40	94.50	86.40	0.00	n.a.
GEM cons	69.40	45.20	2.90	0.00	0.00	97.80	95.40	82.20	n.a.	n.a.
GEM cons ann	70.50	50.20	13.20	0.00	0.00	97.70	95.10	94.30	0.00	n.a.
GSNAP	68.30	24.20	9.10	4.00	0.00	98.00	89.10	84.90	84.60	n.a.
GSNAP ann	91.00	76.40	56.10	33.00	0.00	98.00	93.20	89.50	85.70	0.00
GSTRUCT	90.30	78.40	53.10	18.30	0.00	97.80	96.50	95.70	78.10	0.00
GSTRUCT ann	91.80	80.00	56.60	29.70	0.00	98.20	96.60	95.90	83.50	0.00
MapSplice	73.80	51.10	15.20	2.20	0.00	98.20	92.00	94.10	100.00	n.a.
MapSplice ann	74.30	51.20	17.60	0.70	0.00	95.90	90.00	90.50	14.30	n.a.
PALMapper	79.00	17.10	0.00	0.00	0.00	87.70	60.20	10.00	0.00	n.a.
PALMapper ann	82.70	28.30	0.00	0.00	0.00	85.90	65.90	10.60	0.00	n.a.
PALMapper cons	43.50	3.00	0.00	0.00	0.00	95.70	97.10	n.a.	n.a.	n.a.
PALMapper cons ann	72.40	22.00	0.00	0.00	0.00	89.60	88.80	66.70	n.a.	n.a.
PASS	50.80	6.60	0.10	0.00	0.00	93.30	93.90	100.00	n.a.	n.a.
PASS cons	50.40	6.60	0.10	0.00	0.00	94.50	94.20	100.00	n.a.	n.a.
ReadsMap	76.20	67.80	56.30	30.80	0.00	76.60	65.10	55.10	58.70	n.a.
SMALT	5.20	0.00	0.00	0.00	0.00	32.10	n.a.	n.a.	n.a.	n.a.
STAR 1-pass	51.80	11.30	0.70	0.00	0.00	98.10	96.30	94.00	n.a.	n.a.
STAR 1-pass ann	79.30	54.10	44.30	28.20	0.00	96.60	91.80	91.30	92.80	n.a.
STAR 2-pass	81.80	59.90	47.70	25.60	0.00	97.30	93.50	93.00	100.00	n.a.
STAR 2-pass ann	83.10	61.80	48.50	28.90	0.00	96.90	92.40	90.10	100.00	n.a.
TopHat v1	68.30	54.90	32.80	0.00	0.00	97.00	89.90	88.60	0.00	n.a.
TopHat v1 ann	70.50	58.80	44.00	14.70	0.00	97.00	89.30	89.30	90.90	n.a.
TopHat v2	64.00	56.60	37.70	0.00	0.00	97.20	95.70	95.70	0.00	n.a.
TopHat v2 ann	73.80	68.70	64.00	40.30	0.00	96.50	91.40	88.80	71.00	n.a.
Number of simulated reads	14962090	622980	11701	270	3					

Table A.8: Transcript reconstruction accuracy. The exons and transcripts constituting the simulated transcriptomes were classified as known or novel, depending whether they were included in the annotation provided to aligners. Note that lower accuracy for novel transcripts is expected even for protocols not using annotation, as the expression levels are lower for novel transcripts on average. The precision estimates for known and novel features serve to assess the effect on precision when excluding a defined subset of matches. Precision for known features was computed as $TP_{known}/(TP_{known} + FP)$, i.e. by excluding predictions matching novel transcripts. Similarly, precision for novel features was computed as $TP_{novel}/(TP_{novel} + FP)$. These values should not be interpreted as absolute precision estimates, but in a relative manner, for comparison among methods.

	Exon recall			Exon precision			Spliced transcript recall			Spliced transcript precision		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
A. Simulation 1												
BAGET ann	76.50	81.60	6.50	59.60	59.50	0.80	12.00	38.60	3.80	25.40	20.50	7.70
GEM ann	81.80	84.10	50.50	77.40	76.70	12.30	15.50	39.60	8.10	29.20	19.80	14.20
GEM cons	74.00	76.20	44.10	74.60	73.90	10.40	12.70	31.80	6.80	24.90	16.40	12.00
GEM cons ann	81.10	83.80	44.50	77.10	76.50	10.90	15.30	40.30	7.60	29.30	20.40	13.60
GSNAP	82.00	84.10	54.20	78.90	78.20	14.10	16.10	40.00	8.70	30.40	20.30	15.40
GSNAP ann	83.10	85.30	53.30	82.30	81.70	16.60	18.00	45.60	9.50	35.90	25.00	18.50
GSTRUCT	83.00	85.00	55.70	81.50	80.80	16.40	17.70	43.70	9.70	34.90	23.70	18.40
GSTRUCT ann	83.20	85.30	55.20	82.20	81.50	16.90	18.00	44.80	9.80	35.60	24.40	18.70
MapSplice	81.30	83.30	54.00	80.50	79.80	15.40	16.10	40.00	8.80	32.60	22.00	16.80
MapSplice ann	82.00	84.50	47.60	80.70	80.10	13.90	16.30	42.50	8.30	33.30	23.40	16.30
PALMapper	82.50	84.60	54.00	66.20	65.20	7.80	15.00	37.10	8.30	28.00	18.40	14.10
PALMapper ann	83.10	85.30	53.70	66.30	65.30	7.70	14.30	35.50	7.80	26.50	17.40	13.10
PALMapper cons	78.20	80.40	47.20	59.20	58.20	5.50	13.20	32.20	7.30	25.20	16.20	12.60
PALMapper cons ann	80.60	82.70	50.70	64.00	63.00	6.90	15.50	38.30	8.50	31.40	20.90	16.20
PASS	64.30	66.30	36.50	41.60	40.70	2.60	9.30	23.50	4.90	14.10	8.90	6.30
PASS cons	64.60	66.60	37.00	42.50	41.60	2.80	9.30	23.40	5.00	14.30	9.00	6.40
ReadsMap	72.60	74.50	46.70	54.10	53.00	4.80	13.00	31.40	7.30	21.70	13.60	10.70
SMALT	21.60	22.20	14.30	21.90	21.20	1.10	1.00	2.00	0.60	1.70	0.90	0.90
STAR 1-pass	80.30	82.30	53.30	77.10	76.30	12.90	14.40	36.20	7.80	26.20	17.30	12.70
STAR 1-pass ann	83.90	86.20	52.80	79.90	79.20	14.30	17.80	46.40	9.00	34.60	24.40	17.00
STAR 2-pass	82.40	84.30	55.70	80.00	79.20	15.30	16.80	41.50	9.20	32.10	21.50	16.60
STAR 2-pass ann	84.10	86.20	55.10	80.00	79.30	14.90	17.60	44.80	9.20	33.50	23.20	16.90
TopHat v1	77.60	79.80	46.90	78.00	77.30	12.40	14.00	35.40	7.40	27.00	18.00	13.00
TopHat v1 ann	81.40	83.90	47.00	79.80	79.20	13.20	16.20	42.40	8.20	31.60	22.10	15.20

Continued on next page

Table A.8 – continued from previous page

	Exon recall			Exon precision			Spliced transcript recall			Spliced transcript precision		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
TopHat v2	78.70	80.90	47.70	81.30	80.70	14.90	15.00	38.00	8.00	30.90	20.90	15.40
TopHat v2 ann	83.60	86.30	46.50	83.40	82.80	15.70	17.90	48.10	8.70	37.40	27.30	18.10
Truth	86.00	87.60	65.20	85.70	85.10	23.10	19.90	48.40	11.20	40.00	27.50	22.30
B. Simulation 2												
BAGET ann	76.50	81.70	7.30	56.80	56.70	0.70	12.00	38.20	3.70	24.70	20.00	7.20
GEM ann	74.00	76.40	42.40	66.60	65.80	7.00	9.00	21.00	5.20	14.60	8.80	7.00
GEM cons	64.10	66.20	36.40	62.40	61.50	5.80	6.70	14.30	4.30	11.10	6.00	5.80
GEM cons ann	73.50	76.20	36.70	66.20	65.50	6.10	8.80	21.70	4.80	14.50	9.00	6.60
GSNAP	80.40	82.60	50.40	70.90	70.00	9.20	12.70	30.10	7.30	21.00	13.00	10.40
GSNAP ann	81.90	84.30	49.70	76.20	75.50	11.50	15.00	36.50	8.30	26.10	16.90	12.90
GSTRUCT	81.60	83.80	52.20	75.60	74.80	11.70	14.90	35.10	8.60	25.60	16.10	13.10
GSTRUCT ann	82.20	84.50	52.10	76.10	75.30	11.90	15.30	35.90	8.90	26.20	16.50	13.50
MapSplice	71.20	73.30	43.50	70.00	69.10	8.70	10.50	24.10	6.30	19.10	11.40	9.70
MapSplice ann	73.50	76.10	37.80	71.40	70.70	7.80	11.20	27.10	6.20	20.50	13.00	9.80
PALMapper	79.40	81.70	49.20	61.00	60.00	6.00	12.00	28.30	6.90	20.20	12.50	10.00
PALMapper ann	80.60	82.90	49.10	62.30	61.40	6.20	11.40	27.80	6.20	19.60	12.40	9.20
PALMapper cons	75.20	77.70	43.00	55.00	54.00	4.40	10.90	25.00	6.50	19.90	12.00	10.10
PALMapper cons ann	77.40	79.90	45.10	59.40	58.50	5.40	12.90	30.90	7.30	25.10	16.10	12.60
PASS	45.60	47.10	25.80	34.40	33.50	2.00	4.30	9.70	2.60	6.50	3.60	3.10
PASS cons	46.00	47.50	26.30	35.10	34.20	2.10	4.30	9.70	2.60	6.50	3.60	3.10
ReadsMap	67.40	69.30	42.70	39.50	38.40	2.70	9.30	21.10	5.70	11.00	6.20	5.40
SMALT	20.70	21.20	14.00	20.80	20.20	1.00	0.90	1.60	0.70	1.70	0.70	1.00
STAR 1-pass	71.70	73.80	44.50	67.20	66.30	7.70	9.50	21.10	5.80	15.00	8.60	7.70
STAR 1-pass ann	80.80	83.40	45.30	73.10	72.40	9.10	14.40	37.10	7.30	24.60	16.70	11.20
STAR 2-pass	76.90	79.10	48.10	71.90	71.00	9.60	12.20	28.30	7.10	20.50	12.50	10.30
STAR 2-pass ann	80.50	82.90	48.40	73.30	72.50	9.90	13.70	33.40	7.60	23.20	14.90	11.20
TopHat v1	69.60	71.70	41.40	70.10	69.30	8.50	9.70	21.50	6.10	16.20	9.20	8.40
TopHat v1 ann	76.90	79.50	42.10	74.80	74.10	9.80	12.80	31.50	7.00	22.10	14.20	10.50
TopHat v2	72.60	74.90	42.30	74.10	73.40	10.00	11.10	25.20	6.70	19.60	11.60	10.10
TopHat v2 ann	82.10	85.10	41.90	79.90	79.40	12.00	16.00	41.70	8.00	30.10	21.10	14.10
Truth	85.90	87.50	63.60	84.20	83.60	20.80	18.20	41.40	11.00	31.90	20.20	17.70

Table A.9: Cufflinks incorporation rates for exon junctions in alignments of simulated RNA-seq data. Number and percentage of exon junctions incorporated into transcript isoforms by Cufflinks. The junctions counted are those present in primary alignments, which were used as input to Cufflinks. Junctions are further classified as true and false by comparison to the simulated gene models. n.a., not applicable.

	Type	Incorporated	Discarded	Percent incorporated	Percent incorporated, stratified by number of mappings supporting junction									
					1	2	3	4	5	6	7	8	9	10
														A. Simulation 1
BAGET ann	TRUE	71461	6073	92.20%	60.10%	79.10%	87.10%	89.20%	90.50%	90.50%	91.30%	92.00%	92.70%	95.80%
	FALSE	2644	2497	51.40%	28.60%	40.50%	48.70%	57.80%	56.70%	59.60%	63.20%	58.30%	67.00%	78.70%
GEM ann	TRUE	81780	34835	70.10%	22.20%	34.50%	41.90%	47.50%	47.20%	49.80%	52.30%	53.70%	52.60%	81.20%
	FALSE	1681	7510	18.30%	7.80%	10.20%	12.90%	19.20%	22.40%	24.80%	25.70%	25.40%	39.40%	44.50%
GEM cons	TRUE	73935	23765	75.70%	23.90%	32.90%	29.00%	36.40%	39.90%	45.90%	50.90%	50.50%	52.20%	81.40%
	FALSE	1398	3291	29.80%	19.20%	18.70%	13.50%	20.80%	25.40%	26.60%	22.40%	25.70%	42.40%	47.40%
GEM cons ann	TRUE	80343	27764	74.30%	29.70%	52.20%	55.90%	54.90%	50.30%	51.70%	53.60%	52.90%	53.40%	81.40%
	FALSE	1635	4615	26.20%	14.70%	15.90%	15.80%	20.40%	23.80%	25.10%	24.80%	25.60%	44.10%	47.40%
GSNAP	TRUE	81905	36928	68.90%	19.60%	31.90%	39.80%	44.30%	45.80%	49.30%	49.90%	54.20%	55.40%	82.40%
	FALSE	879	12405	6.60%	1.70%	6.30%	7.20%	9.00%	14.00%	14.70%	15.40%	12.70%	17.00%	25.90%
GSNAP ann	TRUE	82283	38536	68.10%	16.30%	28.50%	36.70%	42.00%	42.70%	45.10%	46.60%	49.60%	49.60%	80.50%
	FALSE	697	17840	3.80%	1.50%	3.50%	5.30%	4.90%	3.90%	6.50%	4.30%	5.60%	7.70%	12.70%
GSTRUCT	TRUE	82639	36948	69.10%	20.60%	33.40%	26.40%	38.80%	41.60%	46.00%	47.00%	52.90%	50.30%	81.00%
	FALSE	624	8263	7.00%	1.50%	5.40%	9.00%	9.20%	11.40%	13.60%	11.10%	15.20%	15.70%	23.50%
GSTRUCT ann	TRUE	82815	36966	69.10%	18.00%	24.20%	36.80%	42.70%	43.40%	47.00%	47.90%	52.50%	50.90%	80.90%
	FALSE	667	7778	7.90%	1.50%	5.40%	9.10%	10.50%	11.60%	16.70%	14.50%	17.50%	20.30%	30.90%
MapSplice	TRUE	80694	34996	69.80%	17.00%	30.10%	38.70%	43.10%	44.70%	47.80%	49.40%	49.90%	52.40%	80.30%
	FALSE	613	3457	15.10%	3.60%	8.80%	18.60%	15.20%	16.90%	19.80%	19.00%	12.20%	26.50%	43.90%
MapSplice ann	TRUE	81525	37515	68.50%	17.90%	30.80%	39.90%	43.00%	44.20%	47.00%	49.00%	49.80%	51.40%	80.10%
	FALSE	943	21502	4.20%	0.90%	1.20%	3.70%	4.60%	4.60%	5.90%	5.00%	3.50%	8.50%	43.60%
PALMapper	TRUE	81809	35406	69.80%	23.60%	35.50%	41.60%	45.20%	45.30%	48.10%	49.30%	51.20%	50.50%	80.70%
	FALSE	6232	276799	2.20%	1.40%	1.80%	2.30%	3.30%	3.80%	5.00%	4.70%	5.40%	6.90%	16.20%
PALMapper ann	TRUE	82172	36489	69.20%	23.00%	42.30%	47.40%	49.20%	45.10%	47.40%	49.70%	49.60%	50.40%	79.10%
	FALSE	8091	317835	2.50%	1.60%	2.20%	2.90%	3.60%	3.80%	4.20%	4.30%	6.10%	6.00%	13.80%
PALMapper cons	TRUE	79281	27076	74.50%	35.00%	42.80%	49.90%	51.10%	55.50%	59.00%	61.50%	62.50%	64.20%	86.40%
	FALSE	2072	5196	28.50%	7.80%	15.40%	27.80%	35.80%	33.30%	43.60%	47.20%	49.30%	46.10%	58.80%
PALMapper cons ann	TRUE	79825	28434	73.70%	25.00%	28.50%	40.30%	48.60%	48.50%	50.80%	52.90%	54.00%	53.60%	81.30%
	FALSE	3472	39756	8.00%	3.00%	4.40%	6.20%	7.10%	8.40%	10.10%	9.00%	10.80%	11.60%	16.60%

Continued on next page

Table A.9 – continued from previous page

	Type	Incorporated	Discarded	Percent incorporated	Percent incorporated, stratified by number of mappings supporting junction									
					1	2	3	4	5	6	7	8	9	10
PASS	TRUE	70212	43810	61.60%	9.50%	19.60%	28.30%	34.50%	37.50%	40.20%	41.30%	45.80%	45.50%	77.60%
	FALSE	1599	60998	2.60%	0.50%	1.70%	3.60%	4.00%	6.50%	4.40%	8.20%	8.50%	14.10%	25.40%
PASS cons	TRUE	70270	43566	61.70%	9.30%	19.90%	28.10%	35.30%	37.50%	40.70%	41.90%	46.90%	46.70%	77.60%
	FALSE	1424	35861	3.80%	0.60%	2.20%	4.00%	5.70%	7.70%	5.40%	9.40%	11.70%	14.80%	26.50%
ReadsMap	TRUE	74212	39940	65.00%	19.10%	33.50%	39.60%	41.60%	42.70%	44.30%	47.10%	47.60%	46.70%	73.80%
	FALSE	10530	888179	1.20%	0.10%	0.20%	0.30%	0.50%	0.70%	1.10%	1.40%	1.70%	2.00%	11.60%
SMALT	TRUE	26213	24284	51.90%	28.50%	39.60%	46.50%	51.90%	54.60%	54.20%	57.30%	57.90%	59.90%	66.20%
	FALSE	55687	84998	39.60%	14.60%	32.20%	42.70%	48.60%	51.00%	54.60%	57.10%	57.20%	57.90%	62.30%
STAR 1-pass	TRUE	80602	35637	69.30%	22.00%	34.60%	42.50%	45.70%	48.50%	52.00%	52.00%	55.30%	57.40%	83.00%
	FALSE	1217	5308	18.70%	3.40%	13.30%	18.90%	32.30%	31.30%	36.00%	35.30%	41.60%	37.70%	65.40%
STAR 1-pass ann	TRUE	81625	37388	68.60%	15.70%	27.80%	36.00%	40.30%	42.50%	44.90%	45.20%	48.20%	51.10%	83.00%
	FALSE	2008	18212	9.90%	3.30%	7.00%	7.90%	9.10%	12.20%	14.70%	13.50%	15.00%	17.30%	35.90%
STAR 2-pass	TRUE	82230	34855	70.20%	16.80%	29.50%	38.50%	43.60%	45.50%	48.00%	48.30%	51.60%	52.60%	81.30%
	FALSE	1104	10471	9.50%	2.50%	6.10%	7.90%	9.50%	15.30%	15.70%	16.50%	17.60%	17.70%	36.80%
STAR 2-pass ann	TRUE	83682	35545	70.20%	18.10%	31.30%	39.40%	44.90%	45.10%	47.90%	48.70%	51.50%	53.10%	81.50%
	FALSE	1818	19380	8.60%	2.70%	5.80%	7.20%	8.00%	8.70%	11.50%	14.40%	12.00%	15.30%	34.20%
TopHat v1	TRUE	77951	30836	71.70%	18.30%	31.00%	39.70%	45.20%	46.70%	49.10%	53.20%	52.60%	54.50%	81.00%
	FALSE	1470	6231	19.10%	3.80%	7.80%	10.70%	12.70%	15.80%	18.30%	25.00%	21.90%	24.30%	41.30%
TopHat v1 ann	TRUE	81346	31841	71.90%	24.60%	38.90%	45.70%	50.20%	50.10%	51.20%	54.90%	54.00%	55.40%	81.30%
	FALSE	1569	6797	18.80%	4.50%	8.90%	10.60%	14.30%	14.80%	15.30%	20.50%	22.40%	26.10%	38.30%
TopHat v2	TRUE	78219	31457	71.30%	16.20%	28.80%	37.20%	43.10%	44.80%	46.50%	49.20%	51.50%	52.90%	81.20%
	FALSE	587	7301	7.40%	2.30%	3.80%	5.40%	8.20%	5.50%	9.90%	11.70%	11.80%	11.60%	13.90%
TopHat v2 ann	TRUE	82301	33656	71.00%	21.40%	35.90%	42.40%	46.10%	46.90%	48.40%	49.20%	51.20%	50.50%	80.90%
	FALSE	1276	23048	5.20%	3.10%	4.50%	4.30%	4.60%	4.20%	5.30%	5.30%	7.30%	6.60%	13.00%
Truth	TRUE	85827	36918	69.90%	17.80%	32.70%	40.60%	46.30%	46.90%	48.00%	48.40%	52.80%	53.10%	81.10%
	FALSE	0	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
														B. Simulation 2
BAGET ann	TRUE	70768	6194	92.00%	58.10%	78.50%	86.10%	87.70%	89.30%	90.60%	90.20%	91.80%	91.90%	95.40%
	FALSE	2748	3606	43.20%	20.90%	29.80%	42.20%	52.30%	53.60%	56.40%	51.60%	67.20%	71.00%	74.10%
GEM ann	TRUE	78127	34233	69.50%	23.10%	38.20%	47.20%	51.30%	55.10%	57.90%	58.50%	59.70%	60.00%	77.30%
	FALSE	3878	18414	17.40%	12.80%	15.50%	16.70%	17.10%	16.30%	17.30%	21.70%	17.30%	18.70%	31.30%
GEM cons	TRUE	68279	23095	74.70%	20.10%	32.50%	35.90%	42.70%	50.80%	55.90%	56.90%	59.60%	58.80%	77.70%
	FALSE	3623	8998	28.70%	28.70%	27.90%	23.90%	23.50%	21.30%	21.60%	25.90%	22.80%	19.60%	35.70%

Continued on next page

Table A.9 – continued from previous page

	Type	Incorporated	Discarded	Percent incorporated	Percent incorporated, stratified by number of mappings supporting junction									
					1	2	3	4	5	6	7	8	9	10
GEM cons ann	TRUE	76677	28740	72.70%	26.30%	47.40%	57.80%	60.00%	61.20%	62.30%	61.20%	60.50%	62.30%	77.80%
	FALSE	3878	10902	26.20%	23.20%	24.30%	24.30%	22.30%	21.90%	20.70%	26.00%	22.10%	24.30%	36.10%
GSNAP	TRUE	85566	33717	71.70%	21.10%	36.40%	46.00%	53.10%	56.80%	58.60%	59.80%	62.20%	63.10%	81.50%
	FALSE	1508	29179	4.90%	1.30%	4.50%	9.20%	12.90%	13.80%	16.20%	19.50%	19.60%	20.20%	30.50%
GSNAP ann	TRUE	86561	34865	71.30%	17.00%	32.30%	42.70%	50.20%	53.10%	54.80%	57.50%	59.10%	57.80%	80.20%
	FALSE	1099	35534	3.00%	1.00%	3.50%	4.50%	5.80%	7.10%	7.10%	10.50%	8.50%	7.10%	14.10%
GSTRUCT	TRUE	87072	32849	72.60%	20.70%	37.80%	32.90%	43.30%	51.00%	54.60%	58.70%	61.20%	59.30%	81.30%
	FALSE	1223	21843	5.30%	1.00%	4.80%	6.80%	11.70%	15.40%	15.60%	17.70%	18.70%	18.40%	31.90%
GSTRUCT ann	TRUE	87729	32490	73.00%	18.90%	29.00%	41.30%	48.90%	53.10%	56.00%	59.00%	61.90%	60.80%	81.40%
	FALSE	1156	21453	5.10%	1.10%	5.10%	7.50%	13.80%	16.40%	20.30%	20.90%	21.60%	17.70%	31.10%
MapSplice	TRUE	73923	35728	67.40%	15.50%	30.60%	40.80%	46.80%	51.30%	53.50%	54.30%	56.80%	57.20%	74.30%
	FALSE	894	8412	9.60%	2.00%	7.20%	12.00%	14.50%	13.40%	9.10%	12.60%	16.90%	11.20%	25.40%
MapSplice ann	TRUE	76680	39793	65.80%	17.90%	32.20%	42.00%	46.10%	50.80%	51.50%	53.20%	54.50%	56.10%	74.60%
	FALSE	1901	32059	5.60%	0.90%	1.80%	3.30%	5.20%	6.10%	4.50%	4.80%	5.20%	6.30%	40.70%
PALMapper	TRUE	82117	33578	71.00%	25.30%	39.20%	46.60%	51.30%	51.30%	54.30%	54.40%	59.00%	56.50%	78.60%
	FALSE	8866	375041	2.30%	1.40%	1.70%	2.70%	3.30%	3.70%	4.80%	6.40%	6.10%	8.40%	18.00%
PALMapper ann	TRUE	81977	36176	69.40%	25.70%	44.20%	49.30%	50.30%	50.00%	53.20%	55.10%	55.50%	53.70%	76.50%
	FALSE	11698	516507	2.20%	1.50%	2.10%	2.90%	3.90%	3.90%	4.00%	4.80%	6.00%	6.00%	13.80%
PALMapper cons	TRUE	79185	24757	76.20%	31.10%	44.00%	53.60%	57.80%	59.70%	63.10%	63.80%	66.60%	66.80%	83.30%
	FALSE	3809	8447	31.10%	9.90%	17.00%	23.20%	27.20%	29.70%	36.30%	39.10%	39.80%	37.50%	52.00%
PALMapper cons ann	TRUE	79610	26285	75.20%	21.20%	25.80%	35.90%	47.30%	55.00%	59.60%	60.70%	61.10%	64.80%	79.50%
	FALSE	5162	53950	8.70%	3.00%	4.80%	6.50%	7.40%	9.30%	9.70%	9.70%	9.90%	11.10%	16.30%
PASS	TRUE	55961	51878	51.90%	9.60%	19.00%	26.60%	30.90%	34.90%	36.40%	39.00%	40.30%	43.00%	63.40%
	FALSE	2383	122903	1.90%	0.40%	1.20%	2.50%	3.60%	4.90%	5.60%	6.90%	8.20%	7.70%	18.10%
PASS cons	TRUE	56849	50715	52.90%	9.90%	19.30%	26.20%	31.40%	35.70%	38.00%	39.80%	41.70%	43.70%	64.60%
	FALSE	2186	75171	2.80%	0.50%	1.60%	2.90%	4.90%	6.00%	6.00%	8.50%	8.90%	9.30%	20.00%
ReadsMap	TRUE	72491	36565	66.50%	16.10%	31.30%	42.30%	44.50%	48.90%	48.50%	50.40%	51.60%	53.40%	73.50%
	FALSE	16563	926112	1.80%	0.10%	0.40%	0.70%	1.10%	1.70%	2.50%	3.30%	4.30%	5.90%	19.60%
SMALT	TRUE	23924	26802	47.20%	27.90%	37.70%	41.70%	45.90%	47.60%	48.00%	50.60%	51.60%	50.40%	61.60%
	FALSE	53865	127976	29.60%	10.90%	25.00%	34.50%	39.10%	43.30%	45.80%	47.50%	49.20%	49.60%	56.80%
STAR 1-pass	TRUE	75728	34574	68.70%	23.40%	39.20%	48.40%	51.60%	55.10%	59.10%	60.00%	63.20%	62.00%	78.60%
	FALSE	1880	13012	12.60%	3.00%	10.60%	15.30%	24.70%	22.80%	29.70%	37.10%	32.10%	44.20%	59.10%
STAR 1-pass ann	TRUE	81935	34840	70.20%	17.20%	30.00%	39.20%	44.90%	49.70%	51.00%	54.70%	55.70%	57.00%	81.50%

Continued on next page

Table A.9 – continued from previous page

	Type	Incorporated	Discarded	Percent incorporated	Percent incorporated, stratified by number of mappings supporting junction									
					1	2	3	4	5	6	7	8	9	10
	FALSE	2504	29188	7.90%	2.40%	6.10%	8.70%	11.30%	9.40%	12.50%	11.50%	16.80%	13.50%	35.10%
STAR 2-pass	TRUE	81369	31663	72.00%	17.20%	31.90%	41.80%	49.30%	54.60%	55.20%	59.20%	58.90%	61.10%	79.80%
	FALSE	1687	21062	7.40%	1.80%	5.60%	8.10%	11.10%	13.00%	13.50%	14.70%	16.90%	11.50%	31.10%
STAR 2-pass ann	TRUE	84855	32293	72.40%	19.40%	34.00%	45.00%	51.50%	56.00%	56.50%	59.20%	59.60%	63.10%	80.50%
	FALSE	2254	30597	6.90%	1.80%	4.70%	7.00%	9.70%	8.60%	10.20%	10.70%	16.70%	12.30%	28.90%
TopHat v1	TRUE	74198	27200	73.20%	19.10%	36.70%	46.10%	53.20%	57.30%	61.40%	61.30%	62.00%	63.40%	79.30%
	FALSE	1647	9729	14.50%	3.00%	5.90%	8.90%	11.60%	13.70%	16.00%	17.30%	15.90%	17.70%	32.20%
TopHat v1 ann	TRUE	81144	27783	74.50%	25.60%	45.20%	53.60%	59.00%	60.90%	63.00%	64.20%	64.70%	65.20%	81.00%
	FALSE	1722	10545	14.00%	3.50%	6.50%	7.60%	11.20%	14.00%	12.70%	18.40%	15.80%	17.60%	30.20%
TopHat v2	TRUE	76695	27586	73.50%	21.20%	35.80%	48.90%	53.70%	57.10%	59.40%	60.70%	61.30%	62.30%	80.20%
	FALSE	550	9010	5.80%	2.10%	2.90%	4.70%	4.80%	4.60%	5.20%	10.10%	6.60%	9.50%	9.50%
TopHat v2 ann	TRUE	84921	28651	74.80%	25.60%	43.00%	52.20%	57.40%	59.90%	59.50%	61.90%	61.00%	65.20%	81.90%
	FALSE	1492	24889	5.70%	4.10%	4.30%	5.80%	6.50%	4.90%	5.60%	7.50%	6.90%	5.30%	9.60%
Truth	TRUE	92247	31334	74.60%	21.10%	39.20%	52.10%	55.10%	59.40%	61.20%	62.00%	64.40%	64.00%	82.10%
	FALSE	0	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Appendix B

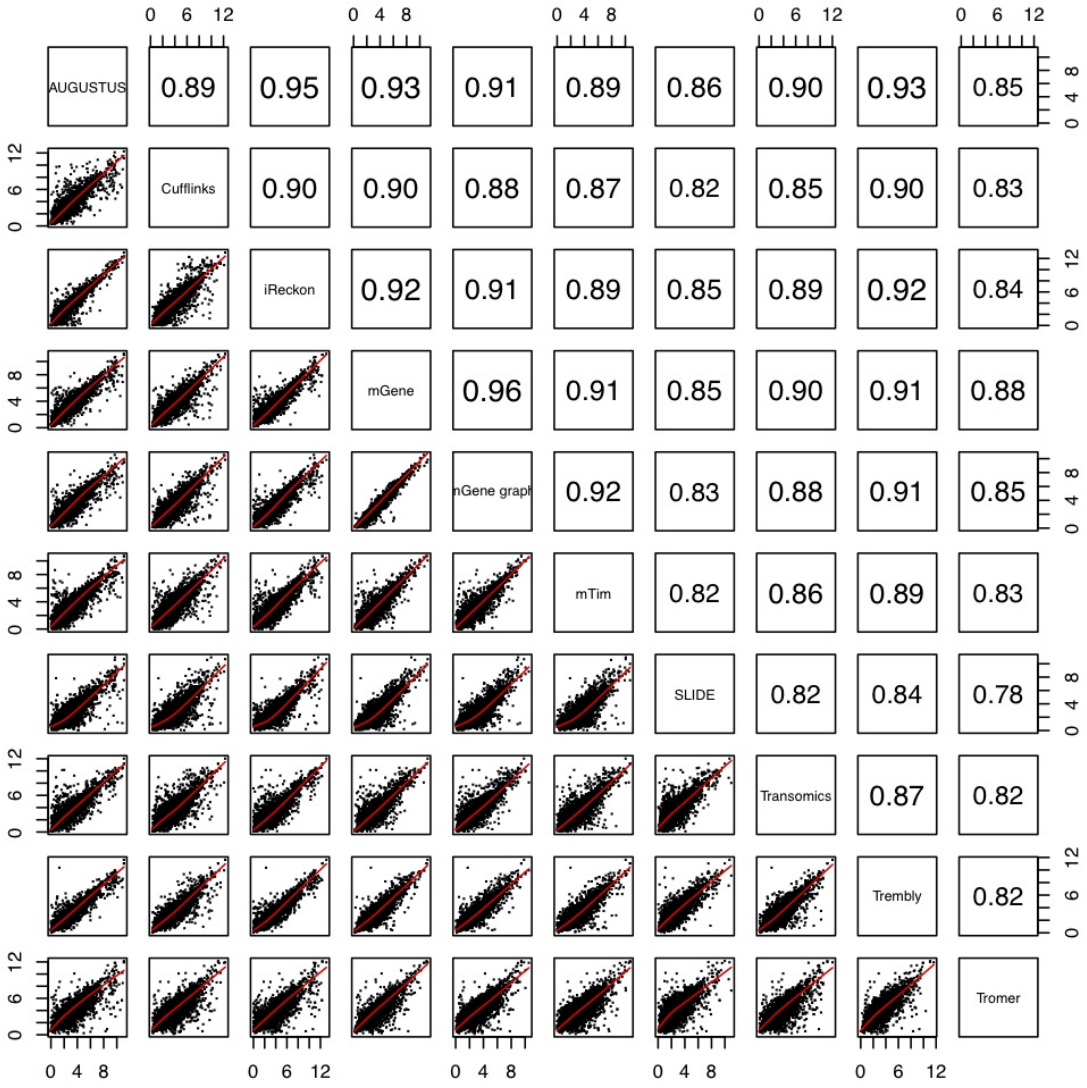


Figure B.1: Comparison of quantification methods for *H. sapiens*. For each pair of methods, scatter plots relate $\log_2$ RPKM values for the genes identified by all methods. The corresponding correlation coefficients (Pearson's r) are shown opposite. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method. RPKM values for AUGUSTUS, iReckon, SLIDE, Transomics and Trembly correspond to the values reported by their *all* and *full* protocols.

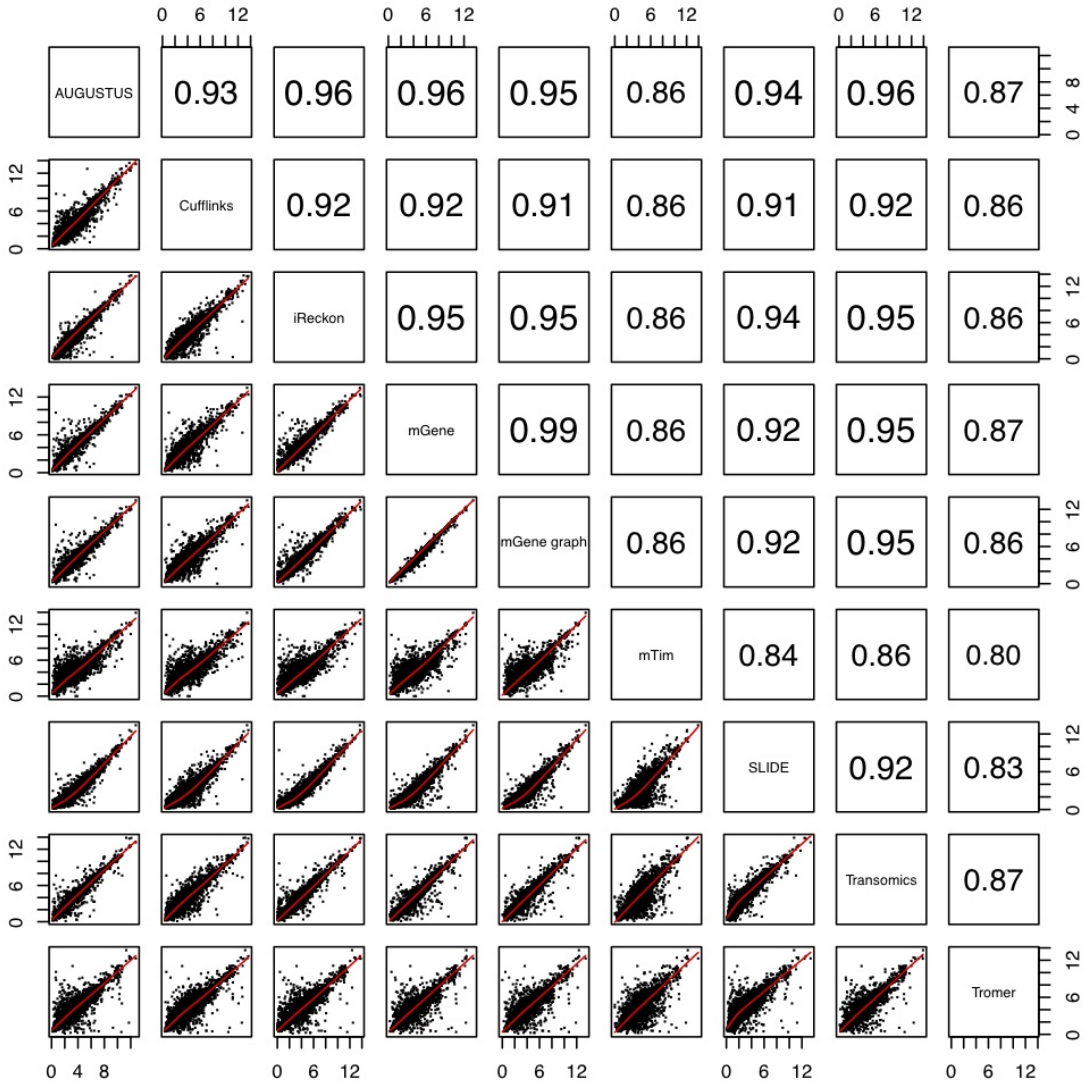


Figure B.2: Comparison of quantification methods for *D. melanogaster*. For each pair of methods, scatter plots relate $\log_2$ RPKM values for the genes identified by all methods. The corresponding correlation coefficients (Pearson's r) are shown opposite. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method. RPKM values for AUGUSTUS, iReckon, SLIDE, and Transomics correspond to the values reported by their *all* and *full* protocols.

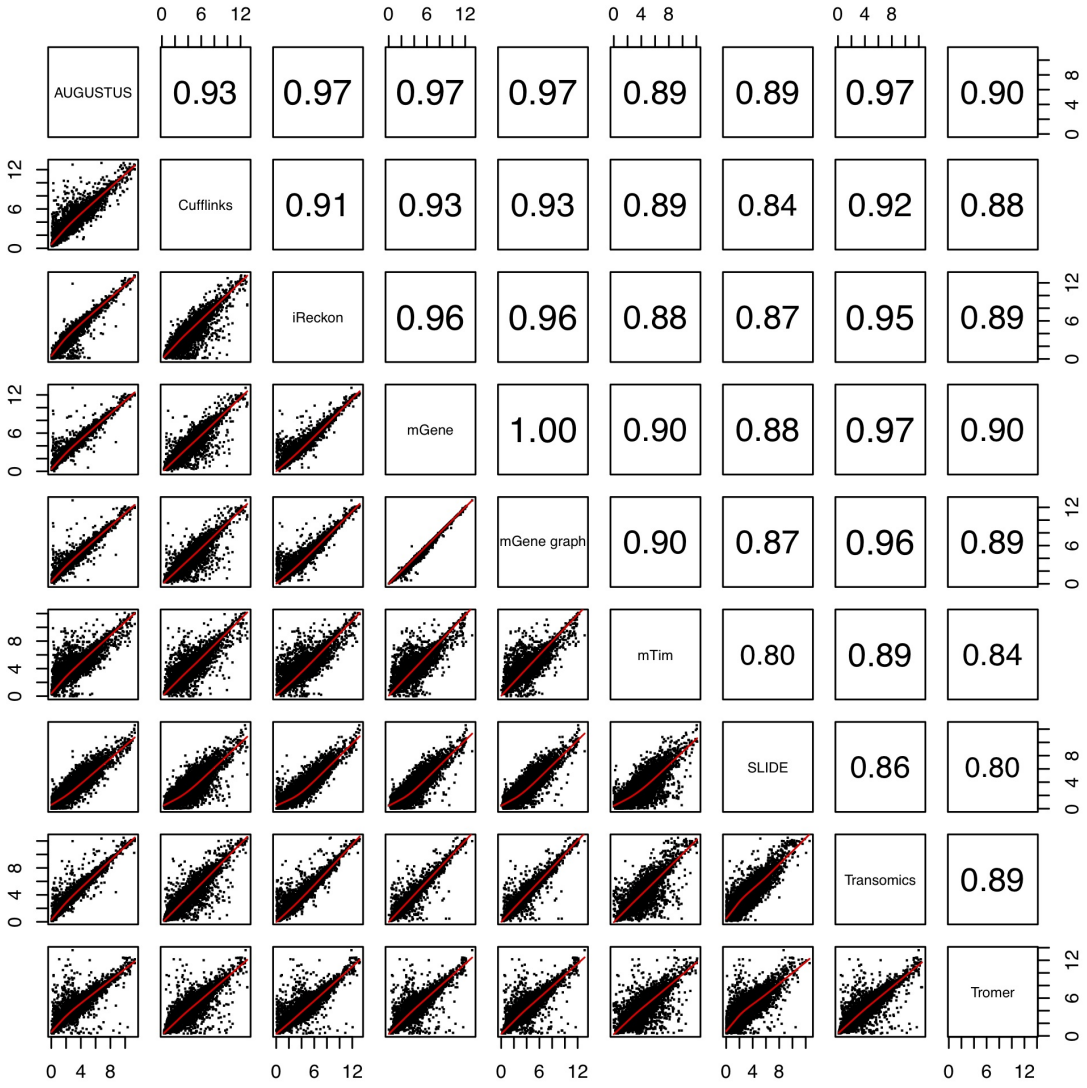


Figure B.3: Comparison of quantification methods for *C. elegans*. For each pair of methods, scatter plots relate $\log_2$ RPKM values for the genes identified by all methods. The corresponding correlation coefficients (Pearson's r) are shown opposite. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method. RPKM values for AUGUSTUS, iReckon, SLIDE, and Transomics correspond to the values reported by their *all* and *full* protocols.

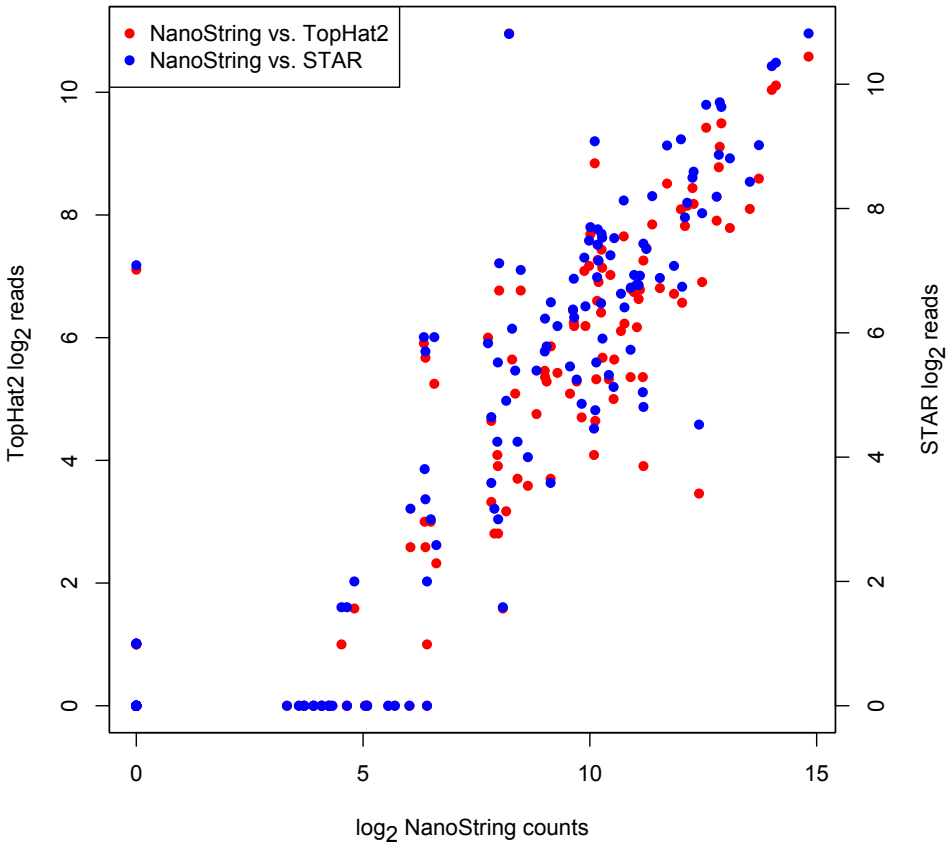


Figure B.4: Correlation between NanoString counts and numbers of mapped reads for targeted exons and junctions. Scatter plots show individual data points in red (TopHat2) and blue (STAR). Count values were incremented by 1 prior to log transformation to avoid infinite numbers.

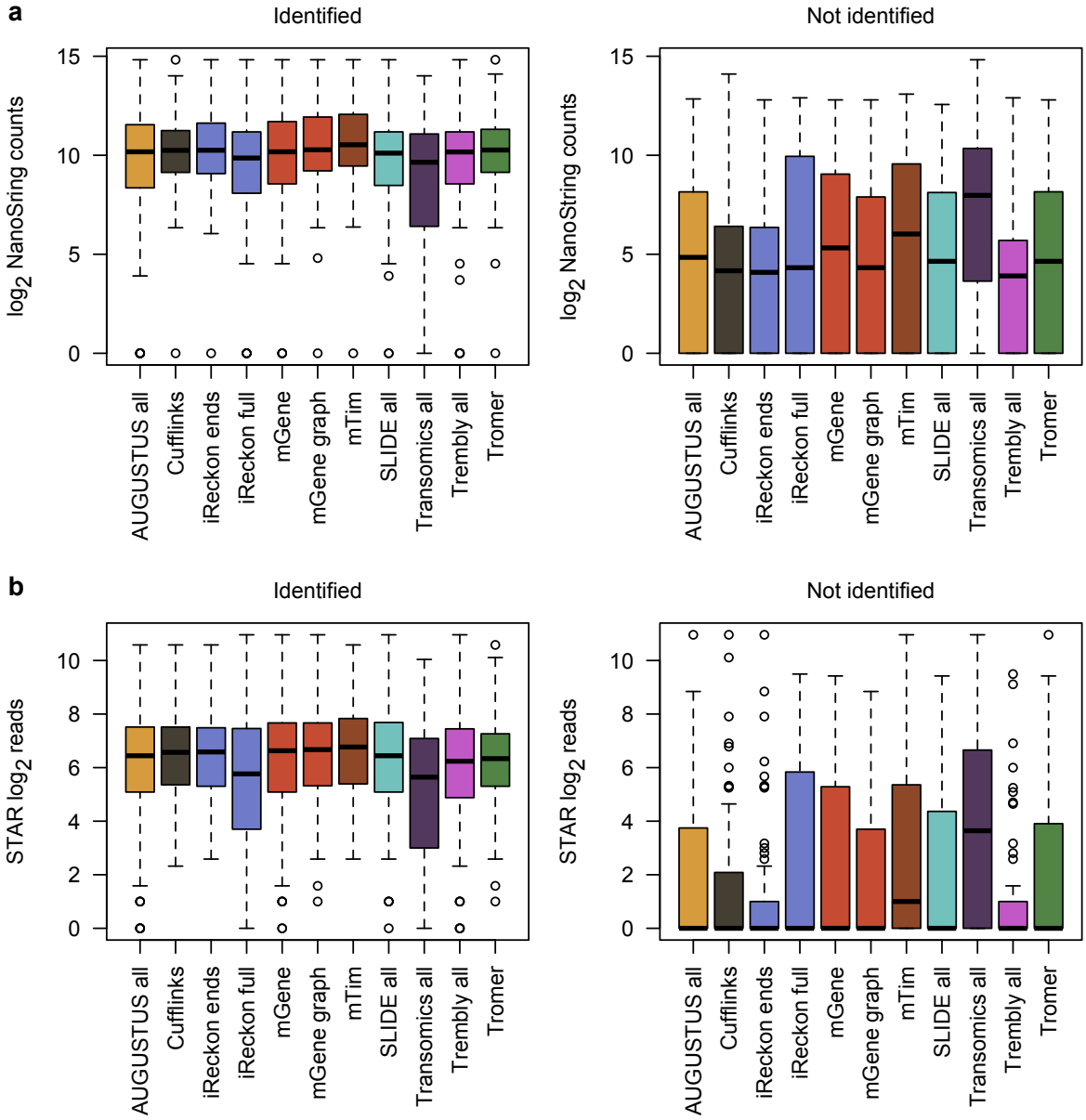


Figure B.5: Distribution of NanoString counts (a) and mapped reads by the STAR aligner (b) for probes depending on whether a method identified an isoform consistent with a probe (left) or not (right). Both mTim and Transomics failed to identify many exons or junctions with RNA-seq read support. Count values were incremented by 1 prior to log transformation to avoid infinite values.

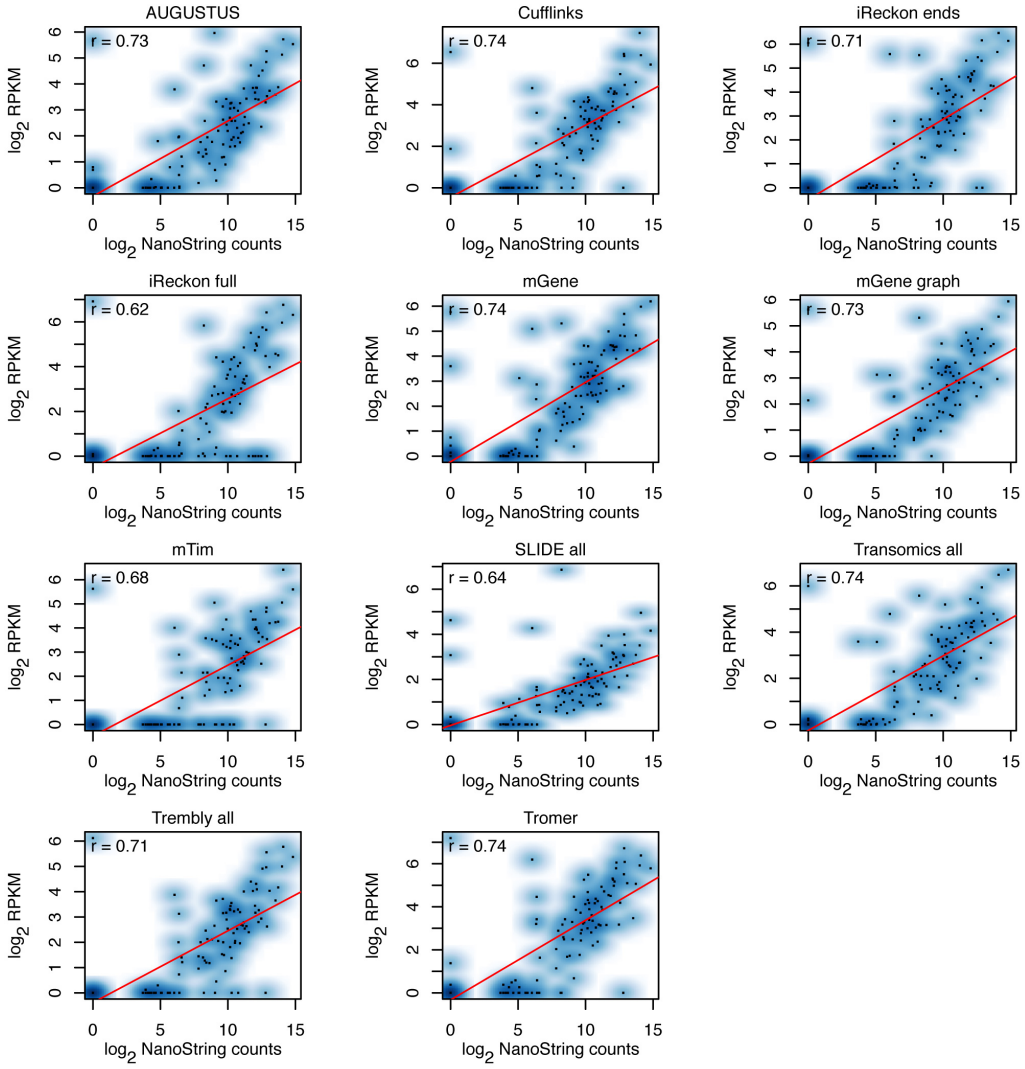


Figure B.6: Correlation between NanoString counts and gene RPKMs. Scatter plots show individual data points in black, with colour intensity indicating the density of data points. Where multiple transcript were reported for the same gene, the highest RPKM value was used (irrespective of whether that transcript contained the exon or junction targeted by the NanoString probe). Correlation coefficients (Pearson's r) are given for each comparison.

Table B.1: Alternative splicing and transcript diversity

	Minimum	First quartile	Median	Mean	Third quartile	Maximum
H. sapiens						
CDS length (bp)	1	79	115	143.7	159	17330
Exon length (bp)	1	87	126	224.9	186	91670
Intron length (bp)	3	498	1569	6410	4481	4251000
Exons per transcript	1	3	5	6.646	8	118
Transcripts per gene	1	1	1	5.225	7	80
D. melanogaster						
CDS length (bp)	1	124	197	372.6	402	27710
Exon length (bp)	1	144	246	476.2	544	28070
Intron length (bp)	4	65	104	1540	733	139300
Exons per transcript	1	2	4	5.496	7	78
Transcripts per gene	1	1	1	1.943	2	31
C. elegans						
CDS length (bp)	1	99	146	205.3	234	14980
Exon length (bp)	1	99	146	205.3	234	14980
Intron length (bp)	3	50	71	330.6	345	21230
Exons per transcript	1	4	6	6.819	9	66
Transcripts per gene	1	1	1	1.255	1	15

Table B.2: Nucleotide-level performance

	H. sapiens		*D. melanogaster*		*C. elegans*	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
AUGUSTUS all	63.77%	77.14%	87.17%	89.26%	96.88%	65.28%
AUGUSTUS high	63.70%	84.75%			96.65%	72.32%
AUGUSTUS no RNA	55.46%	44.91%	83.93%	83.63%	95.68%	60.17%
Cufflinks	79.31%	59.98%	84.98%	90.87%	88.61%	75.81%
Exonerate all	66.89%	87.21%	77.19%	92.87%	84.82%	78.73%
Exonerate high	65.01%	90.33%	75.22%	94.54%	83.78%	80.12%
GSTRUCT	60.59%	86.08%	58.13%	88.91%	76.67%	69.75%
iReckon full	78.33%	31.29%	89.72%	77.56%	96.08%	83.42%
iReckon ends	83.44%	10.01%	92.40%	69.20%	97.17%	73.29%
mGene	71.38%	42.86%	75.33%	87.87%	95.28%	72.73%
mGene graph	67.56%	55.51%	71.98%	90.68%	85.81%	75.70%
mTim	54.06%	90.24%	61.97%	93.06%	78.08%	81.31%
NextGeneid	64.73%	80.59%	80.36%	94.22%	84.51%	79.16%
NextGeneidAS	63.48%	81.01%	79.51%	94.44%	83.32%	79.13%
NextGeneidAS ab-initio	62.95%	79.72%	79.32%	94.37%	83.26%	79.32%
Oases	68.28%	68.53%	75.98%	91.65%	76.58%	80.52%
SLIDE all	89.28%	88.33%	96.67%	96.15%	93.36%	96.90%
SLIDE high	76.48%	87.87%	91.94%	96.64%	88.41%	97.07%
Transomics all	45.65%	53.57%	68.35%	83.70%	98.36%	72.00%
Transomics high	45.29%	75.06%	68.30%	88.99%	98.12%	82.61%
Trembly all	70.54%	85.99%				
Trembly high	53.24%	91.41%				
Tromer	84.67%	39.50%	92.06%	76.70%	87.74%	63.15%
Velvet	65.79%	78.69%	72.54%	88.07%	77.73%	82.18%
Velvet + AUGUSTUS	37.66%	75.51%			65.72%	76.78%

Table B.3: Exon-, transcript- and gene-level performance for CDS reconstruction

	Exon		Transcript		Gene	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
H. sapiens						
AUGUSTUS all	66.18%	75.03%	19.51%	43.70%	61.47%	45.64%
AUGUSTUS high	66.09%	81.46%	19.50%	49.45%	61.46%	53.23%
AUGUSTUS no RNA	54.96%	48.88%	5.34%	9.28%	17.61%	9.28%
Exonerate all	57.36%	85.11%	19.77%	31.88%	58.12%	31.88%
Exonerate high	56.04%	89.39%	16.24%	42.65%	54.29%	42.65%
mGene	63.13%	50.32%	14.62%	10.01%	50.01%	10.02%
mGene graph	53.49%	82.44%	16.03%	34.44%	49.33%	46.01%
mTim	28.76%	92.55%	8.82%	46.66%	27.52%	52.53%
NextGeneid	50.47%	85.22%	11.29%	38.01%	40.96%	38.01%
NextGeneidAS	50.11%	82.48%	11.77%	31.47%	39.84%	31.47%
NextGeneidAS ab-initio	50.14%	80.49%	11.76%	29.20%	39.82%	29.20%
Transomics all	66.23%	50.68%	11.10%	14.59%	39.52%	14.59%
Transomics high	65.58%	69.73%	11.10%	23.89%	39.51%	23.89%
Tromer	31.58%	29.65%	2.23%	0.93%	6.30%	1.66%
D. melanogaster						
AUGUSTUS all	73.74%	77.11%	24.47%	39.36%	48.53%	44.03%
AUGUSTUS no RNA	64.97%	70.15%	16.60%	34.09%	33.18%	34.09%
Exonerate all	62.39%	77.96%	17.31%	28.01%	33.88%	28.01%
Exonerate high	60.48%	81.96%	16.34%	39.36%	32.63%	39.36%
mGene	70.73%	81.30%	22.00%	44.02%	43.99%	44.02%
mGene graph	62.54%	82.58%	19.34%	41.26%	38.43%	47.14%
mTim	35.55%	82.59%	8.90%	34.06%	17.66%	40.08%
NextGeneid	59.77%	76.11%	18.69%	38.84%	37.37%	38.84%
NextGeneidAS	61.43%	73.08%	19.20%	32.29%	37.92%	32.29%
NextGeneidAS ab-initio	61.44%	72.99%	19.24%	32.21%	37.99%	32.21%
Transomics all	73.62%	66.12%	23.48%	33.54%	46.95%	33.54%
Transomics high	73.56%	71.22%	23.48%	37.72%	46.93%	37.72%
Tromer	13.85%	18.64%	3.26%	2.81%	6.46%	5.75%
C. elegans						
AUGUSTUS all	84.38%	72.19%	48.20%	36.02%	60.15%	38.76%
AUGUSTUS high	84.21%	79.22%	48.13%	42.44%	60.06%	45.98%
AUGUSTUS no RNA	78.94%	64.66%	36.28%	27.20%	45.52%	27.20%
Exonerate all	67.67%	82.06%	34.18%	32.66%	42.46%	32.66%
Exonerate high	66.55%	84.82%	32.55%	41.25%	40.84%	41.25%
mGene	83.62%	74.15%	45.48%	41.94%	57.05%	41.94%
mGene graph	72.62%	77.61%	45.02%	45.46%	56.38%	47.01%
mTim	45.34%	85.51%	20.24%	36.66%	25.24%	43.02%
NextGeneid	70.15%	81.18%	30.28%	39.80%	37.98%	39.80%
NextGeneidAS	69.78%	79.51%	30.39%	33.37%	37.96%	33.37%
NextGeneidAS ab-initio	69.78%	79.43%	30.39%	33.26%	37.97%	33.26%
Transomics all	86.37%	65.43%	48.30%	32.82%	60.56%	32.82%
Transomics high	86.10%	74.75%	48.28%	40.33%	60.54%	40.33%
Tromer	20.85%	26.55%	1.20%	0.51%	1.50%	1.11%

Table B.4: Exon-, transcript-, and gene-level performance (fixed evaluation mode)

	Exon		Transcript		Gene	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
H. sapiens						
AUGUSTUS all	49.07%	64.29%	0.00%	0.00%	0.01%	0.01%
AUGUSTUS high	49.00%	69.85%	0.00%	0.01%	0.01%	0.01%
AUGUSTUS no RNA	41.27%	42.41%	0.00%	0.00%	0.00%	0.00%
Cufflinks	42.63%	62.90%	0.00%	0.00%	0.01%	0.00%
Exonerate all	43.32%	71.18%	0.00%	0.00%	0.01%	0.00%
Exonerate high	42.75%	75.85%	0.00%	0.01%	0.01%	0.01%
GSTRUCT	41.58%	79.67%	0.00%	0.00%	0.00%	0.00%
iReckon full	48.48%	68.40%	0.06%	0.06%	0.24%	0.06%
iReckon ends	50.26%	64.32%	0.05%	0.03%	0.18%	0.03%
mGene	47.14%	41.63%	0.00%	0.00%	0.01%	0.00%
mGene graph	46.52%	59.98%	0.00%	0.01%	0.02%	0.01%
mTim	38.28%	73.72%	0.00%	0.00%	0.01%	0.01%
NextGeneid	39.79%	73.26%	0.01%	0.12%	0.02%	0.12%
NextGeneidAS	38.94%	69.02%	0.00%	0.09%	0.02%	0.09%
NextGeneidAS ab-initio	38.97%	67.58%	0.00%	0.14%	0.02%	0.14%
Oases	36.00%	29.91%	0.00%	0.00%	0.01%	0.00%
SLIDE all	52.62%	79.66%	3.13%	4.75%	12.03%	18.18%
SLIDE high	38.56%	84.06%	3.82%	11.53%	14.75%	18.76%
Transomics all	48.52%	45.72%	0.05%	0.63%	0.19%	0.63%
Transomics high	48.01%	62.85%	0.05%	1.02%	0.19%	1.02%
Trembly all	43.14%	65.35%	0.01%	0.01%	0.02%	0.01%
Trembly high	38.76%	74.96%	0.00%	0.00%	0.01%	0.01%
Tromer	23.87%	21.04%	0.00%	0.00%	0.00%	0.00%
Velvet	36.28%	46.45%	0.00%	0.00%	0.00%	0.00%
Velvet + AUGUSTUS	22.57%	73.75%	0.00%	0.00%	0.00%	0.00%
D. melanogaster						
AUGUSTUS all	45.45%	51.16%	0.01%	0.02%	0.03%	0.02%
AUGUSTUS no RNA	40.06%	47.22%	0.00%	0.01%	0.01%	0.01%
Cufflinks	39.06%	49.65%	0.00%	0.01%	0.01%	0.01%
Exonerate all	40.32%	51.70%	0.02%	0.03%	0.04%	0.03%
Exonerate high	39.52%	57.66%	0.02%	0.04%	0.03%	0.04%
GSTRUCT	33.34%	59.03%	0.00%	0.01%	0.01%	0.01%
iReckon full	61.82%	61.71%	0.03%	0.03%	0.05%	0.03%
iReckon ends	58.56%	59.42%	0.02%	0.02%	0.04%	0.02%
mGene	42.61%	55.60%	0.03%	0.05%	0.05%	0.05%
mGene graph	40.36%	54.37%	0.03%	0.05%	0.05%	0.05%
mTim	33.87%	55.67%	0.01%	0.02%	0.02%	0.03%
NextGeneid	37.49%	51.53%	0.01%	0.02%	0.03%	0.02%
NextGeneidAS	38.88%	47.37%	0.02%	0.02%	0.03%	0.02%
NextGeneidAS ab-initio	38.88%	47.33%	0.03%	0.03%	0.04%	0.03%
Oases	33.61%	28.54%	0.01%	0.01%	0.03%	0.01%
SLIDE all	84.78%	94.14%	44.19%	25.52%	76.99%	70.36%
SLIDE high	77.01%	95.78%	44.75%	50.81%	82.73%	75.38%
Transomics all	41.16%	44.40%	0.44%	0.65%	0.88%	0.65%
Transomics high	41.09%	47.77%	0.44%	0.73%	0.88%	0.73%
Tromer	10.46%	10.56%	0.00%	0.00%	0.00%	0.00%
Velvet	30.27%	32.91%	0.02%	0.02%	0.03%	0.02%
C. elegans						
AUGUSTUS all	62.92%	53.19%	0.00%	0.00%	0.00%	0.00%
AUGUSTUS high	62.80%	58.34%	0.00%	0.00%	0.00%	0.00%
AUGUSTUS no RNA	59.53%	47.94%	0.00%	0.00%	0.00%	0.00%
Cufflinks	53.55%	59.41%	0.00%	0.00%	0.00%	0.00%
Exonerate all	53.51%	60.84%	0.00%	0.00%	0.00%	0.00%
Exonerate high	53.01%	63.98%	0.00%	0.00%	0.00%	0.00%
GSTRUCT	49.19%	64.98%	0.00%	0.00%	0.00%	0.00%
iReckon full	78.70%	71.81%	0.00%	0.00%	0.00%	0.00%
iReckon ends	71.85%	69.63%	0.00%	0.00%	0.00%	0.00%
mGene	62.93%	55.35%	0.00%	0.00%	0.00%	0.00%
mGene graph	56.22%	56.33%	0.00%	0.00%	0.00%	0.00%
mTim	52.80%	62.77%	0.00%	0.00%	0.00%	0.00%
NextGeneid	59.65%	67.18%	4.16%	5.03%	5.22%	5.03%
NextGeneidAS	59.67%	64.82%	4.58%	4.21%	5.66%	4.21%
NextGeneidAS ab-initio	59.74%	64.92%	4.63%	4.28%	5.73%	4.28%
Oases	45.33%	42.17%	0.00%	0.00%	0.00%	0.00%
SLIDE all	87.50%	97.24%	40.85%	12.66%	50.26%	43.98%
SLIDE high	81.65%	97.94%	56.51%	50.52%	70.19%	61.42%
Transomics all	88.87%	67.32%	54.36%	36.94%	68.16%	36.94%
Transomics high	88.45%	76.78%	54.35%	45.40%	68.14%	45.40%
Tromer	22.72%	19.22%	0.00%	0.00%	0.00%	0.00%
Velvet	45.87%	50.48%	0.00%	0.00%	0.00%	0.00%
Velvet + AUGUSTUS	58.49%	67.24%	0.00%	0.00%	0.00%	0.00%

Table B.5: Exon-, transcript-, and gene-level performance (flexible evaluation mode)

	Exon		Transcript		Gene	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
H. sapiens						
AUGUSTUS all	81.16%	77.47%	16.27%	39.26%	56.16%	42.71%
AUGUSTUS high	81.05%	84.15%	16.27%	44.41%	56.16%	49.78%
AUGUSTUS no RNA	65.55%	50.73%	3.52%	8.03%	13.38%	8.03%
Cufflinks	73.45%	79.57%	16.03%	19.29%	53.55%	20.36%
Exonerate all	74.59%	88.34%	17.22%	28.60%	55.53%	28.60%
Exonerate high	72.72%	92.46%	13.23%	41.92%	50.28%	41.92%
GSTRUCT	70.63%	96.42%	14.10%	59.85%	54.02%	59.85%
iReckon full	67.78%	83.05%	31.80%	38.87%	73.63%	38.87%
iReckon ends	73.24%	83.26%	27.78%	35.77%	71.61%	35.77%
mGene	78.40%	50.96%	10.83%	8.98%	41.35%	9.00%
mGene graph	77.27%	73.14%	14.90%	14.89%	48.17%	21.42%
mTim	64.86%	90.42%	10.14%	22.66%	35.78%	36.95%
NextGeneid	67.36%	89.58%	11.30%	36.99%	43.27%	36.99%
NextGeneidAS	66.65%	87.32%	11.96%	27.73%	42.73%	27.73%
NextGeneidAS ab-initio	66.68%	85.56%	11.94%	26.60%	42.63%	26.60%
Oases	62.02%	40.55%	10.04%	5.11%	34.05%	4.74%
SLIDE all	72.58%	84.10%	8.78%	12.68%	29.08%	34.60%
SLIDE high	54.75%	88.22%	8.65%	22.66%	31.09%	31.62%
Transomics all	76.84%	54.35%	8.41%	13.28%	32.16%	13.28%
Transomics high	76.14%	74.64%	8.41%	21.74%	32.16%	21.74%
Trembly all	75.45%	83.44%	21.82%	22.73%	60.53%	38.37%
Trembly high	64.30%	90.24%	12.84%	30.28%	43.60%	41.80%
Tromer	49.02%	37.22%	4.27%	3.70%	13.92%	6.48%
Velvet	62.05%	59.56%	5.57%	6.81%	21.35%	7.26%
Velvet + AUGUSTUS	38.97%	90.68%	7.17%	45.22%	27.44%	45.35%
D. melanogaster						
AUGUSTUS all	82.61%	86.03%	42.66%	61.14%	73.03%	65.92%
AUGUSTUS no RNA	71.12%	77.83%	23.30%	42.22%	41.21%	42.22%
Cufflinks	71.73%	83.70%	36.18%	46.32%	56.37%	50.06%
Exonerate all	73.32%	86.35%	38.51%	51.44%	58.83%	51.44%
Exonerate high	69.54%	92.50%	30.98%	66.02%	54.54%	66.02%
GSTRUCT	59.66%	95.60%	28.21%	76.77%	49.10%	76.77%
iReckon full	85.85%	82.43%	65.22%	57.94%	89.07%	57.94%
iReckon ends	82.93%	80.44%	57.05%	51.13%	83.09%	51.13%
mGene	75.49%	91.06%	38.02%	67.02%	67.17%	67.05%
mGene graph	73.03%	90.92%	40.00%	64.82%	67.78%	70.56%
mTim	62.35%	90.40%	20.59%	40.70%	34.09%	53.27%
NextGeneid	70.53%	89.19%	36.30%	62.23%	64.27%	62.23%
NextGeneidAS	73.97%	83.76%	40.05%	49.64%	66.30%	49.64%
NextGeneidAS ab-initio	73.95%	83.66%	40.05%	49.61%	66.25%	49.61%
Oases	64.31%	51.05%	33.99%	18.61%	55.47%	25.39%
SLIDE all	89.08%	95.36%	52.80%	31.26%	82.02%	74.88%
SLIDE high	82.31%	96.83%	53.22%	58.60%	88.50%	80.58%
Transomics all	73.74%	74.03%	35.11%	44.78%	62.71%	44.78%
Transomics high	73.63%	79.67%	35.11%	50.38%	62.69%	50.38%
Tromer	29.15%	27.41%	9.37%	4.79%	15.69%	10.47%
Velvet	58.59%	58.75%	24.09%	22.63%	42.39%	24.81%
C. elegans						
AUGUSTUS all	91.76%	76.26%	59.28%	44.09%	72.35%	46.62%
AUGUSTUS high	91.48%	83.55%	59.21%	51.96%	72.26%	55.33%
AUGUSTUS no RNA	86.51%	68.56%	42.25%	31.53%	52.77%	31.53%
Cufflinks	78.38%	84.84%	39.75%	37.53%	48.04%	42.88%
Exonerate all	76.99%	85.55%	43.82%	40.02%	52.90%	40.02%
Exonerate high	75.75%	89.27%	39.98%	50.42%	49.83%	50.42%
GSTRUCT	71.48%	91.85%	39.03%	59.92%	48.72%	59.92%
iReckon full	94.09%	84.88%	78.10%	52.98%	89.87%	52.98%
iReckon ends	87.39%	83.71%	68.45%	44.85%	79.25%	44.85%
mGene	90.99%	78.64%	55.27%	50.73%	69.01%	50.73%
mGene graph	83.26%	81.96%	55.78%	54.68%	68.88%	56.18%
mTim	77.16%	88.48%	31.86%	37.72%	38.96%	46.54%
NextGeneid	78.45%	86.60%	37.04%	44.55%	46.19%	44.55%
NextGeneidAS	77.91%	82.80%	38.45%	35.17%	46.71%	35.17%
NextGeneidAS ab-initio	77.85%	82.76%	38.45%	35.41%	46.71%	35.41%
Oases	66.77%	61.04%	32.98%	17.39%	39.70%	22.24%
SLIDE all	88.52%	97.34%	42.85%	13.34%	51.92%	45.44%
SLIDE high	82.97%	98.04%	58.13%	51.78%	71.61%	62.66%
Transomics all	93.90%	69.92%	58.10%	39.32%	72.52%	39.32%
Transomics high	93.43%	79.73%	58.08%	48.32%	72.50%	48.32%
Tromer	39.43%	32.92%	7.24%	1.96%	8.97%	4.96%
Velvet	67.44%	72.86%	31.77%	27.15%	39.54%	28.09%
Velvet + AUGUSTUS	61.49%	69.46%	0.00%	0.00%	0.00%	0.00%

Table B.6: NanoString probes. Genes and their transcript isoforms targeted by NanoString probes.

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
adar1_sp1	ENST00000292205, ENST00000494866, ENST00000368471	ACTGGCAGTCTCCGGGTGTCCGGCCGTGTCCCGAGGAAGTGCAAGACCCGGGGTATTCCCTCAGCGGATACTACACCCATCCATTTCAAGGCTATGAGCA
adar1_sp2	ENST00000368474	CGGCGGGTCGGGCCGGGCAATGCCTCGCGGGCGCAATGAATCCGCGGCAGGGGTATTCCCTCAGCGGATACTACACCCATCCATTTCAAGGCTATGAGCA
atf2_common	ENST00000264110, ENST00000345739, ENST00000392543, ENST00000392544, ENST00000409499, ENST00000409833, ENST00000413123, ENST00000415955, ENST00000417080, ENST00000421438, ENST00000426833, ENST00000428760, ENST00000429579, ENST00000435231, ENST00000437522, ENST00000445349, ENST00000456655, ENST00000487334, ENST00000538946, ENST00000542046, ENST00000409635	GGTACTAGATGGAAACTTGAGAAAGGACTGCTTATTGATAACAGCTAAGGTATTCCTGGAAGCAGAGTAAATAAAGCTCATGGCCCACCAGCTAGAAAG
atf2_sp1	ENST00000345739, ENST00000437522, ENST00000435231, ENST00000415955, ENST00000409635, ENST00000456655	ATATGAGTGATGACAAACCCTTTCTATGTACTGCGCCTGGATGTGGCCAGATCAGACCCCAACACCAACAAGATTCTTGAAAAACTGTGAAGAAGTGGGT
atf2_sp2	ENST00000429579, ENST00000538946, ENST00000428760, ENST00000417080, ENST00000421438, ENST00000409437, ENST00000392544, ENST00000264110, ENST00000435004, ENST00000542046, ENST00000426833, ENST00000409833, ENST00000487334	ATATGAGTGATGACAAACCCTTTCTATGTACTGCGCCTGGATGTGGCCAGCGTTTTACCAACGAGGATCATTTGGCTGTCCATAAACATAAACATGAGAT
ATP5J_common	ENST00000284971, ENST00000400087, ENST00000400090, ENST00000400093, ENST00000400094, ENST00000457143, ENST00000486002, ENST00000400099	CAGAGTATCAGCAAGAGCTGGAGAGGGAGCTTTTTAAGCTCAAGCAAATGTTTGGTAATGCAGACATGAATACATTTCCCACCTTCAAATTTGAA
Bcl11a_sp1	ENST00000335712, ENST00000356842, ENST00000359629, ENST00000489516	TTTATCAACGTCATCTAGAGGAATTTGCCCCAAACAGGAACACATAGCAGATAAACTTCTGCACTGGAGGGGCCTCTCCTCCCCTCGTTCTGCACATGGA
Bcl11a_sp2	ENST00000409351	TATCAACGTCATCTAGAGGAATTTGCCCCAAACAGGAACACATAGCAGCTCAGACTGAACTGGAGGATGTATTTGTGTACCTTATGGTGT
BCL3	ENST00000164227, ENST00000403534	CGGAGCCTTACTGCCTTTGTACCCCACTCGGGCCATGGGCTCCCCGTTTCCTCTGGTGAACCTGCCTACACCCCTATACCCCATGATGTGCCCCATGGAA
BHLHB2	ENST00000256495	AAAAGCTTCAAAGTCTTGGTCTGTGAGTCACTCTTCAGTTTGGGAGCTGGGTCTGTGGCTTTGATCAGAAGGTACTTTCAAAAGAGGGCTTTCCAGGGCT

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
Blnk_sp1_T	ENST00000371176	TGAAAACTATATTCATCCCACAGAAAGCAGTTCACCTCCA-CCTGAAAAAGGTCGAAACAGTGGGGCCTGGGAAACCAAGT-CACCTCCACCAGCTGCACCA
Blnk_sp2	ENST00000413476, ENST00000224337, ENST00000427367, ENST00000467799	TGAAAACTATATTCATCCCACAGAAAGCAGTTCACCTCCA-CCTGAAAAAGCTCCCATGGTGAATAGATCAACCAAGCCAA-ATTCCTCAACGCCCGCCTCT
CARM1_sp1_T	ENST00000592516, ENST00000344150	TGTTATTGCCAGTGGCTCCAGCGTGGGCCACAACAACCTG-ATTCCTTTAGGGTCCTCCGGCGCCCAGGGCAGTGGTGGTG-GCAGCACGAGTGCCCACTAT
CD19	ENST00000324662, ENST00000538922, ENST00000565089, ENST00000567541	GAAGGTCTCAGCTGTGACTTTGGCTTATCTGATCTTCTGC-CTGTGTTCCCTTGTGGGCATTCTTCATCTTCAAAGAGCCC-TGGTCCTGAGGAGGAAAAGA
Cd79b_sp1_T	ENST00000349817	TGAGCCAGTACCAGCAGCCAGATCGGAGGACCGGTACCGG-AATCCCAAAGGATTCAGCACCTTGGCACAGCTGAAGCAGA-GGAACACGCTGAAGGATGGT
Cd79b_sp2_T	ENST00000559358, ENST00000006750, ENST00000392795	CAACACCTCGGAGGTCTACCAGGGCTGCGGCACAGAGCTG-CGAGTCATGGGATTCAGCACCTTGGCACAGCTGAAGCAGA-GGAACACGCTGAAGGATGGT
cdkn1a_sp1_T	ENST00000244741	GAGCCGGAGCTGGGCGCGGATTCGCCGAGGCACCGAGGCA-CTCAGAGGAGGCGCCATGTCAGAACCGGCTGGGGATGTCC-GTCAGAACCCATGCGGCAGC
cdkn1a_sp2_T	ENST00000405375, ENST00000478800	GGATGCGTGTTCGCGGGTGTGTGCTGCGTTCACAGGTGTT-TCTGCGGCAGGCGCCATGTCAGAACCGGCTGGGGATGTCC-GTCAGAACCCATGCGGCAGC
CEBPA	ENST00000425420, ENST00000498907	CTAGTATTTAGGATAACCTTGTGCCTTGGAAATGCAAACT-CACCGCTCCAATGCCTACTGAGTAGGGGGAGCAAATCGTG-CCTTGTCATTTTATTTGGAG
CTCF_common	ENST00000264010, ENST00000401394	CCCAACGGAGACCTCACGCCCGAGATGATCCTCAGCATGA-TGGACCGGTGATGGCGGAGCCTTGTGCGTCGCCAGGACTT-CTCTGGGCTGTGTTTAAACG
CTCF_sp1	ENST00000566078, ENST00000264010	GAGCTGGGTTCTATTTTCCCTCCTCAAACTGACTTTGCA-GCCACGGAGAGGCAGGGGAAATGGAAGGTGATGCAGTCGA-AGCCATTGTGGAGGAGTCCGA
CTCFL_sp1	ENST00000433949, ENST00000243914, ENST00000539382, ENST00000371196, ENST00000426658, ENST00000423479, ENST00000422869, ENST00000502686	TGCCAGCAGAGATACCTACAAGCTGAAACGCCACATGAGA-ACGCACTCAGGTGAGAAGCCTTACGAATGCCACATCTGCC-ACACCCGCTTCACCCAGAGC
CTCFL_sp2	ENST00000429804	TGCCAGCAGAGATACCTACAAGCTGAAACGCCACATGAGA-ACGCACTCAGGTGTGCATATGCGCAACTTGCATGCTTACA-GCGCTGCAGAGCTGAAATGC
CTDSL_sp2_T	ENST00000273179, ENST00000443503, ENST00000486978	GCCCCAGTGTGCTTCCGCCACTGGTGGAGGAGAATGGTGG-GCTTCAGAAGGGTGACCAGAGGCAGGTCATTCCCATACCA-AGTCCACCAGCTAAGTACCT

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
CTDSP1_common_T	ENST00000273062, ENST00000428361, ENST00000443891, ENST00000452977, ENST00000464255, ENST00000473420, ENST00000482272, ENST00000488627, ENST00000491064, ENST00000497677, ENST00000498160	AGCTGACCTGCTGGACAAATGGGGGGCCTTCCGGGCCCGGCTGTTTCGAGAGTCCTGCGTCTTCCACCGGGGGAACTACGTGAAGGACCTGAGCCGGTT
CTDSP2	ENST00000398073	CCTGTCCTGTACCGAGCTCTGTCTGTTCCAGCCTTCATCCTTCCTGGCTGTTGCTTTTCCTCTTAAGGGCCTCAGAACTCTTGCTCTTCCTGGGCTGAGG
CTDSPL_sp1	ENST00000443503	GCCCCAGTGTGCTTCCGCCACTGGTGGAGGAGAATGGTGGGCTTCAGAAGCCACCAGCTAAGTACCTTCTTCCAGAGGTGACGGTGCTTGACTATGGAAA
DES	ENST00000492726, ENST00000477226, ENST00000373960	GAGAACAATTTGGCTGCCTTCCGAGCGGACGTGGATGCAGCTACTCTAGCTCGCATTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGT
DNMT1_common_T	ENST00000340748, ENST00000359526, ENST00000540357, ENST00000586588, ENST00000587197, ENST00000588913, ENST00000589294, ENST00000592705	CCTTTCCGGACCATCACGGTGCGAGACACGATGTCCGACCTGCCGGAGGTGCGGAATGGAGCCTCGGCACTGGAGATCTCCTACAACGGGGAGCCTCAGT
E2F4	ENST00000379378, ENST00000567007	GTCAGAAATCTTTGATCCCACACGAGAGTGCATGAGCTCGGAGCTGCTGGAGGAGTTGATGTCCTCAGAAGTGTTTGCCCCTCTGCTTCGTCTTTCTCCA
E2F6_common	ENST00000307236, ENST00000362009, ENST00000381525, ENST00000421117, ENST00000428221, ENST00000437573, ENST00000444832, ENST00000455198, ENST00000468775, ENST00000542100, ENST00000546212	AGGTTGCAACGAAACTGGGAGTCCGAAAGCGGAGAGTGTATGACATCACCAATGTCTTAGATGGAATCGACCTCGTTGAAAAGAAATCCAAGAACCATAT
ebf1_sp1	ENST00000519890, ENST00000380654, ENST00000518836, ENST00000519739, ENST00000313708, ENST00000522192	TGGACAACTGGCCGTGAATGTCTCCGAGGCATCACAAGCCACCAATCAGGGTTTCACCCGCAACTCAAGCAGCGTATCACCACACGGGTACGTGCCGAGC
EGR1	ENST00000239938	CTTCAATGCTAGAAAATCGAGTTGGCAAAATGGGGTTTGGGCCCCTCAGAGCCCTGCCCTGCACCCTTGTACAGTGTCTGTGCCATGGATTTCGTTTTTC
EOMES_common	ENST00000295743, ENST00000449599, ENST00000537516	CAACAAACTAGACATCAGTTCCTATGAATCTGAATATACTTCTAGCACATTGCTCCCATATGGCATTAAATCCTTGCCCCTTCAGACATCCCATGCCCTG
EP300	ENST00000263253	ACAAATATCCCTTTGGCTCCGTCCAGCGGTCAAGCTCCAGTGTCTCAAGCACAAATGTCTAGTTCTTCCTGCCCGGTGAACTCTCCTATAATGCCTCCAG
esr1_common	ENST00000206249, ENST00000338799, ENST00000406599, ENST00000427531, ENST00000440973, ENST00000443427, ENST00000456483, ENST00000544394	GTAGAGGGCATGGTGGAGATCTTCGACATGCTGCTGGCTACATCATCTCGGTTCCGCATGATGAATCTGCAGGGAGAGGAGTTTGTGTGCCTCAAATCTA

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
esr2_sp1	ENST00000353772, ENST00000554572, ENST00000344288, ENST00000358599	ATGCGCCTGGCTAACCTCCTGATGCTCCTGTCCCACGTCAGGCATGCGAGGGCAGAAAAGGCCTCTCAAACACTCACCTCATTTGGAATGAAGATGGAGA
esr2_sp2	ENST00000554520, ENST00000341099, ENST00000267525, ENST00000555483	ATGCGCCTGGCTAACCTCCTGATGCTCCTGTCCCACGTCAGGCATGCGAGTAACAAGGGCATGGAACATCTGCTCAACATGAAGTGCAAAAATGTGGTCC
ets1_common	ENST00000319397, ENST00000345075, ENST00000526145, ENST00000530924, ENST00000531611, ENST00000535549, ENST00000392668	CAGGAGATGGGGAAAGAGGAAAAACAAACCTAAGATGAATTATGAGAAACTGAGCCGTGGCCTACGCTACTATTACGACAAAAACATCATCCACAAGACA
FBXO15	ENST00000269500, ENST00000581214, ENST00000419743, ENST00000585174, ENST00000583443	CTAGCTGACATTCTCAAACCTGTCAACCCTTACACAGGCCTTCCAGTTAAGACCAAAGAGGCCCTCAGAATATTTGGTTTAGGTTGGGCAATTATACTGA
FOS	ENST00000303562	TCAAGTCCTTACCTCTTCCGGAGATGTAGCAAAACGCATGGAGTGTGTATTGTTCCCAGTGACACTTCAGAGAGCTGGTAGTTAGTAGCATGTTGAGCCA
foxa2_body	ENST00000319993, ENST00000377115, ENST00000419308	CGTTCCGTCCCAAACAGAGGGCCACACAGATACCCCACGTTCTATATAAGGAGGAAAACGGGAAAGAATATAAAGTTAAAAAAAAGCCTCCGGTTTCCAC
foxa2_sp2_T	ENST00000377115	CGGGTCCCTGGCGGCCGGTGTCTGAGGAGTCGGAGAGCCGAGGCGGCCAGACCGTGCGCCCCGCGCTTCTCCCGAGGCCGTTCCGGGTCTGAACTGTAAC
FOXA3	ENST00000302177	CCCCGTGTTGGCCATGTCGTCACCATTCTCTCTGGCATGGGTTGGGTAGGGGATGGAGGTGAGAATACTCCTTGGTTTTCTCTGAAGCCCACCCTTTCCC
gabpa_sp1_T	ENST00000354828	CCGGACGGGTCTAGGTGAGACAGAAGCCAAACAGGAGGAGGAAGTGGAGGGACTGATCCTTTGAAATACTCCAGCCATGACTAAAAGAGAAGCAGAGGAG
gabpa_sp2_T	ENST00000400075	CAGCCGGCTCTGGAGTGCGGGCGGGGGCGACAGGGCCGATTCCGGAGTGGGACTGATCCTTTGAAATACTCCAGCCATGACTAAAAGAGAAGCAGAGGAG
GATA1_T	ENST00000376665, ENST00000376670	GTGTCCCACCCGCGAGGACTCTCCTCCCCAGGCCGTGGAAGATCTGGATGGAAAAGGCAGCACCAGCTTCCTGGAGACTTTGAAGACAGAGCGGCTGAGC
HDAC1	ENST00000373548, ENST00000476391	CTGTTTTCGTACCTTCCCACTGGCCTCAAGTGAGCCAAGAAACACTGCCTGCCCTCTGTCTGTCTTCTCCTAATTCTGCAGGTGGAGGTTGCTAGTCTAG
HDAC3_sp1	ENST00000305264, ENST00000495485, ENST00000523353	CATTGACCCATAGCCTGGTCCTGCATTACGGTCTCTATAAGAAGATGATCGTCTTCAAGCCATACCAGGCCTCCCAACATGACATGTGCCGCTTCCACTC
HDAC4	ENST00000345617	TGTCAGCTCACTCCAGCTTCACAAATGTGCTGAGAGCATTACTGTGTAGCCTTTTCTTTGAAGACACACTCGGCTCTTCTCCACAGCAAGCGTCCAGGGC
HDAC5	ENST00000225983, ENST00000336057, ENST00000393622	CAGGGGAGGATCTGGAGGATCCACTACTGTCTTTAAGATGCAGAGTGGAGGGGAGGTGGGCACCCACCCTGCGATTCTCCACCCTTTCCCCTTCTTTCGT

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
Hif1a_common	ENST00000323441, ENST00000337138, ENST00000394997, ENST00000555014, ENST00000539097, ENST00000557538	TCCAGCAGACTCAAATACAAGAACCTACTGCTAATGCCAC-CACTACCACTGCCACCACTGATGAATTAAAAACAGTGACA-AAAGACCGTATGGAAGACAT
HNF1A	ENST00000257555, ENST00000400024, ENST00000402929, ENST00000538646, ENST00000540108, ENST00000541395, ENST00000541924, ENST00000543427, ENST00000544413	GTGCGCTATGGACAGCCTGCGACCAGTGAGACTGCAGAAG-TACCCTCAAGCAGCGGCGGTCCCTTAGTGACAGTGTCTAC-ACCCCTCCACCAAGTGTCCC
HNF1B	ENST00000561193, ENST00000225893	GTTTCCATCTGCAATGGTGGTCACAGATACCAGCAGCATC-AGTACACTCACCAACATGTCTTCAAGTAAACAGTGTCCTC-TACAAGCCTGGTGATGCCCA
Hnf4g_common_T	ENST00000396423	AGAAAATAGTTATCCATTGACTAGAAATTAGTACATGCCC-ACAGCTGGCTCCCACGGTAGCCAGGAGAATTATCTATAGG-TGGAAAGTCTGTGTCAGCCA
HSF1	ENST00000400780, ENST00000528838, ENST00000528988, ENST00000533240	TGTTCGACCAGGGCCAGTTTGCCAAGGAGGTGCTGCCCAA-GTACTTCAAGCACAACAACATGGCCAGCTTCGTGCGGCAG-CTCAACATGTATGGCTTCCG
IGF1R_common_T	ENST00000268035, ENST00000558762	GCGATTGCTGGGTGTGGTGTCCCAAGGCCAGCCAACACTG-GTCATCATGGAACTGATGACACGGGGCGATCTCAAAAGTT-ATCTCCGGTCTCTGAGGCCA
IKZF1_common	ENST00000331340, ENST00000343574, ENST00000346667, ENST00000349824, ENST00000357364, ENST00000359197, ENST00000438033, ENST00000439701, ENST00000440768, ENST00000471793	CCTGCTGCGCGCCGCCTCCGAGAACTCGCAGGACGCGCTC-CGCGTGGTCAGCACCAGCGGGGAGCAGATGAAGGTGTACA-AGTGCGAACACTGCCGGGTG
IKZF1_sp1_T	ENST00000357364, ENST00000331340, ENST00000343574, ENST00000413698, ENST00000346667, ENST00000349824, ENST00000440768, ENST00000359197	CGAGGATCAGTCTTGGCCCCAAAGCGCGACGCACAAATCC-ACATAACCTGAGGACCATGGATGCTGATGAGGGTCAAGAC-ATGTCCCAAGTT
IKZF1_sp2	ENST00000438033, ENST00000492782, ENST00000462201, ENST00000439701	GTGTGGAAAAGGCAGCTCTCACTTGGCCTTGGCGAGGCCT-CGGTTGGTTGATAACCTGAGGACCATGGATGCTGATGAGG-GTCAAGACATGTCCCAAGTT
IKZF3_sp1_T	ENST00000377958, ENST00000377944, ENST00000346872, ENST00000535189, ENST00000467757	CAAGGAGCGCTGCCGTACATTTCTTCAGAGCACTGACCCA-GGGGACACTGCAAGTGCGGAGGCAAGACACATCAAAGCAG-AGATGGGAAGTGAAAGAGCT
IKZF3_sp2	ENST00000583368, ENST00000293068, ENST00000348427	TCACTGACCACAGCAGGTACCCCAGGCAAGAATCTGAGCA-GTTATAACAGCAAGTGCGGAGGCAAGACACATCAAAGCAG-AGATGGGAAGTGAAAGAGCT
IL6	ENST00000258743, ENST00000401630, ENST00000404625, ENST00000407492, ENST00000485300	GCTCTTCGGCAAATGTAGCATGGGCACCTCAGATTGTTGT-TGTTAATGGGCATTCCTTCTTCTGGTCAGAAACCTGTCCA-CTGGGCACAGAACTTATGTT
IL6receptor_common	ENST00000344086, ENST00000368485	TCCAATATTCGCTGTGTCAGCATAGAAGTAACTTACTTAG-GTGTGGGGGAAGCACCATAACTTTGTTTAGCCCAAAACCA-AGTCAAGTGAAAAAGGAGGA

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
IL8	ENST00000307407, ENST00000483500, ENST00000401931	GGAAGGAACCATCTCACTGTGTGTAAACATGACTTCCAAGCTGGCCGTGGCTCTCTTGGCAGCCTTCCTGATTTCTGCAGCTCTGTGTGAAGGTGCAGTT
IL8RA	ENST00000295683	GTCCATTGGGCAGGCAGATGTTCCTAATAAAGCTTCTGTTCCGTGCTTGTCCCTGTGGAAGTATCTTGGTTGTGACAGAGTCAAGGGTGTGTGCAGCATT
IL8RB	ENST00000318507	GATAGACAAATCTCCACCTTCAGACTGGTAGGCTCCTCCAGAAGCCATCAGACAGGAAGATGTGAAAATCCCCAGCACTCATCCCAGAATCACTAAGTGG
IRF8	ENST00000268638, ENST00000566369	CCCTCTGTCTGGGGTGGGATGCCTTACTTTGCACTTAATTTAATAAGGGCATTCTCGGAGGAGTAGACGTTTAATACGAAGTGGCGGCATAGCCCTGCCG
JUND	ENST00000252818	TGCTACGAGTCCACATTCCTGTTTGTAATCCTTGGTTCGCCCGGTTTTCTGTTTTCAGTAAAGTCTCGTTACGCCAGCTCGGCAAAAAAAAAAAAAAAAA
KAISO_sp1_T	ENST00000326624	CCAGCCTTCCGCGCGTCCGGAGGAGGAGAAGCGGCGGCGCCGGGAAGCAGGCATGGAGAGTAGAAAACTGATTTCTGCTACAGACATTCAGTACTCTGGC
KAP1_common_T	ENST00000253024, ENST00000341753	CAGGCCGAGTGCAAACAGGGCAGCAGGCGGGGCTCCCTCTCGGGTCGCAGGCGCTCTCTGCACACGCCGCAGTGCTCCAGCAGCTCCAGCGCCTCGGCGC
KLF4	ENST00000374672, ENST00000493306, ENST00000497048	CCGAGCATTTTCCAGGTCGGACCACCTCGCCTTACACATGAAGAGGCATTTTTAAATCCCAGACAGTGGATATGACCCACACTGCCAGAAGAGAATTCAG
LEF1_sp1	ENST00000509428, ENST00000438313, ENST00000505379, ENST00000510624, ENST00000504775, ENST00000510135, ENST00000379951	TATCCCTTGTCTCCGGGTGGTGTTGGACAGATCACCCCACCTCTTGGCTGGTTTTCCCATCATATGATTCCCGGTCCTCCTGGTCCCCACACAACTGGCA
LEF1_sp2	ENST00000265165, ENST00000506680, ENST00000510717, ENST00000504950, ENST00000515500	TCCCTTGTCTCCGGGTGGTGTTGGACAGATCACCCCACCTCTTGGCTGGCAAGGTCAGCCTGTATATCCCATCACGGGTGGATTCAGGCAACCCTACC
LIN28_common	ENST00000254231, ENST00000326279	GTCTGGAATCCATCCGTGTCACCGGACCTGGTGGAGTATTCTGTATTGGGAGTGAGAGGCGGCCAAAAGGAAAGAGCATGCAGAAGCGCAGATCAAAAGG
MAX_sp1	ENST00000556443, ENST00000358664, ENST00000358402, ENST00000553928, ENST00000555667, ENST00000394606, ENST00000284165, ENST00000553951, ENST00000556979, ENST00000556892, ENST00000557746, ENST00000557277	ACAGCTTTCACAGTTTGCGGGACTCAGTCCCATCACTCCAAGGAGAGAAGGCATCCCGGGCCCAAATCCTAGACAAAGCCACAGAATATATCCAGTATAT
mef2a_common_T	ENST00000354410, ENST00000449277, ENST00000557785, ENST00000557942, ENST00000558812, ENST00000561125, ENST00000338042, ENST00000453228	CCCGCAGCCCCAGCCCCGACAGGAAATGGGGCGCTCCCCTGTGGACAGTCTGAGCAGCTCTAGTAGCTCCTATGATGGCAGTGATCGGGAGGATCCACGG

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
mef2a_sp3	ENST00000558812, ENST00000557785, ENST00000449277, ENST00000557942, ENST00000453228, ENST00000338042	GATTCAGCAAACTAAATGAAGATAGTGATTTTATTTTCAA-ACGAGGCCCTCCTGGTCTGCCACCTCAGAACTTTTCAATG-TCTGTCACAGTTCCAGTGAC
mef2a_sp4	ENST00000354410	AAAAAATTAATGAGGAATTTGATAATATGATGCGGAATCA-TAAAATCGCACCTGGTCTGCCACCTCAGAACTTTTCAATG-TCTGTCACAGTTCCAGTGAC
mef2b_sp10_T	ENST00000477565	CCCCACTGCCACTCCCAGCTGCAAGGACCGTCTCTCAGCT-GCGCTGGGAACCGCTGCTTCTCGCTTATTAGAAAACTGTC-TCTTTCCTTTTGTCCTGGT
mef2b_sp11_T	ENST00000585679, ENST00000514819, ENST00000462498, ENST00000444486	GTCGCTATGGAGGAGCCGGAGATGCAGCTCAAGGGGAAGA-AAGCGCCGTGAAGAACCTGGTGGACAGCAGCGTCTACTTC-CGCAGCGTGGAGG
mef2b_sp12	ENST00000591398, ENST00000162023, ENST00000494489, ENST00000462790, ENST00000588208, ENST00000488252, ENST00000354191, ENST00000477565	TGGGAGGAGCAGAGCCAGGGAGCCATCTACACTGTGGAGT-ACGCCTGCAGCGCCGTGAAGAACCTGGTGGACAGCAGCGT-CTACTTCCGCAGCGTGGAGG
mef2b_sp7_T	ENST00000410050, ENST00000424583, ENST00000409224	CAGCCGCCGCGGGTCCGTGCGCCCAGCGTCCCAGGGCCCA-GGCCGAGCAGACAAAGATCATTCCACTCAGCCTGGGACGA-TGGGGAGGAAAAAAATCCAG
mef2c_sp1	ENST00000508569, ENST00000514015, ENST00000510942, ENST00000514028, ENST00000437473, ENST00000504921, ENST00000503554, ENST00000506554	GGAAAATTAACGAAGATATTGATCTAATGATCAGCAGGCA-AAGATTGTGTGCTGTTCCACCTCCCAACTTCGAGATGCCA-GTCTCCATCCCAGTGTCCAG
mef2c_sp3	ENST00000424173, ENST00000340208	GAAGTAAAGAACGGAAGGCAAATGATTGTGGCAGTAAAGA-AGTGTATGTGCAGGAACGAATGCAGGAATTTGGGAACTGA-GCTGTGCAAGTGCTGAAGAA
mef2c_sp4	ENST00000514015, ENST00000510942, ENST00000514028, ENST00000504921, ENST00000513252, ENST00000506554	TTAAGAAAGGAAAATATCCCAAGGACTAATCTGATCGGGT-CTTCCTTCATCAGGAACGAATGCAGGAATTTGGGAACTGA-GCTGTGCAAGTGCTGAAGAA
mef2d_sp4	ENST00000454816, ENST00000360595, ENST00000368240	CCTGCGAGTCATCACTTCCCAGGCAGGAAAGGGGTTAATG-CATCACTTGACTGAGGACCATTTAGATCTGAACAATGCCC-AGCGCCTTGGGGTCTCCCA
MYC	ENST00000377970, ENST00000524013	AGGAGCAAAAGCTCATTTCTGAAGAGGACTTGTTGCGGAA-ACGACGAGAACAGTTGAAACACAAACTTGAACAGCTACGG-AACTCTTGTGCGTAAGGAAA
MYF5	ENST00000228644	TGGATTGCTTATCCAACATAGTGGACCGGATCACCTCCTC-AGAGCAACCTGGGTTGCCTCTCCAGGATCTGGCTTCTCTC-TCTCCAGTTGCCAGCACCGA
MYF6	ENST00000228641	GGAGGAGCAAGTATTGATTCGTCAGCCTCGAGTAGCCTTC-GATGCCTTTCTTCCATCGTGGACAGTATTTCCTCGGAGGA-ACGCAAACTCCCCTGCGTGG
MYOD1	ENST00000250003	GCATGGTGTGTGGTGCTACAGGGAATTTGTACGTTTATAC-CGCAGGCGGGCGAGCCGCGGGCGCTCGCTCAGGTGATCAA-AATAAAGGCGCTAATTTATA

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
NANOG	ENST00000229307, ENST00000526286, ENST00000526434, ENST00000541267	CTTCACCTATGCCTGTGATTTGTGGGCCTGAAGAAAACTATCCATCCTTGCAAATGTCTTCTGCTGAGATGCCTCACACGGAGACTGTCTCTCCTCTTCC
ncor2_sp1_T	ENST00000404621, ENST00000429285, ENST00000397355, ENST00000404121, ENST00000448614	GGCGTCGGGCGTGAGCGGAAATGAGGAGGAGATGGTGGAGGAGGCTGAAGCCACTGTCAACAACAGCTCAGACACCGAGAGCATCCCCTCTCCTCACACT
NFKB1_common	ENST00000226574, ENST00000394820, ENST00000504044, ENST00000505458, ENST00000600343	ACTGAATCTAAAAAGGACCCTGAAGGTTGTGACAAAAGTGATGACAAAAACACTGTAAACCTCTTTGGGAAAGTTATTGAAACCACAGAGCAAGATCAGG
NOTCH1	ENST00000277541	GAGGGCTTCAGCGGTCCCAACTGCCAGACCAACATCAACGAGTGTGCGTCCAACCCATGTCTGAACCAGGGCACGTGTATTGACGACGTTGCCGGGTACA
NR2F2_sp2_T	ENST00000394166	CCAGTACTGCCGCCTCAAAAAGTGCCTCAAAGTGGGCATGAGACGGGAAGCGGTGCAGAGGGGCAGGATGCCGCCGACCCAGCCGACCCACGGGCAGTTC
OCT4_common	ENST00000259915, ENST00000441888, ENST00000471529, ENST00000512818, ENST00000513407	ATTCAGCCAAACGACCATCTGCCGCTTTGAGGCTCTGCAGCTTAGCTTCAAGAACATGTGTAAGCTGCGGCCCTTGCTGCAGAAGTGGGTGGAGGAAGCT
ONECUT1	ENST00000305901	CTTGGCAAGACAAATGATGAGCAGGAAAACACCACTGGATCTCACACCTTCAATCCATGACCATCCTCGCTGTGCTTGGCTGTTTAGTGGTTTGGAGCAT
ONECUT2	ENST00000262095, ENST00000491143	CCCAAACCAAAATGCTTGACATAAAGCCAAATCAACTGCCAAGCACACTTTATTTTGCATAGGAGTATGCAGCCTAGGGAACCTTGGTTGAAAAGCAGCA
pbx1_common	ENST00000367897, ENST00000420696, ENST00000465089, ENST00000468104, ENST00000496120, ENST00000560469, ENST00000560641	TTTCTCTCCCAACGCTGAAGCGGTCAGACTGGAGGTCGAAGCAATCAGCAAACACAATAAGAGTCTCCTTCTCTTCTCTTCTTTGGGATGCTATTTCAGC
pbx3_sp1	ENST00000373483, ENST00000373482, ENST00000373492	AATGAAACCAGCGCTCTTCAGCGTCCTGTGTGAGATCAAAGAGAAAACAGGCATGTAATGAATTTACTACACATGTGATGAACCTTCTCCGAGAACAGAG
pbx3_sp2_T	ENST00000491787, ENST00000447726, ENST00000342287, ENST00000373489, ENST00000373487	AATGAAACCAGCGCTCTTCAGCGTCCTGTGTGAGATCAAAGAGAAAACAGGTCTCAGCATCAGAGGAGCCCAGGAGGAGGACCCTCCCGATCCCCAGCTA
PER1	ENST00000581395, ENST00000579065, ENST00000354903, ENST00000582719, ENST00000317276	AGCACATCACGTCTGAGTACACACTTCAGAACCAGGATACCTTCTCAGTGGCTGTCTCCTTCCTGACGGGCCGAATCGTCTACATTTCGGAGCAGGCAGC
POLR2A	ENST00000322644	TTCTACTCCAACATTCAGACTGTCATTAACAACTGGCTCCTCATCGAGGGTCATACTATTGGCATTGGGGACTCCATTGCTGATTCTAAGACTTACCAGG
POU5F1_T	ENST00000259915, ENST00000441888, ENST00000471529, ENST00000512818, ENST00000513407	GAGGCTGCTGGGTCTCCTTTCTCAGGGGGACCAGTGTCCTTTCCTCTGGCCCCAGGGCCCCATTTTGGTACCCCAGGCTATGGGAGCCCTCACTTCACTG
PTEN	ENST00000371953	TTGGATGTGCAGCAGCTTACATGTCTGAAGTTACTTGAAGGCATCACTTTTAAGAAAGCTTACAGTTGGGCCCTGTACCATCCCAAGTCCTTTGTAGCTC

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
rbpj_sp2	ENST00000506956, ENST00000514807, ENST00000355476, ENST00000504907, ENST00000511546, ENST00000342320, ENST00000509158, ENST00000511451, ENST00000505958, ENST00000511401, ENST00000514730, ENST00000512351, ENST00000514675	CTGTGACTTACCTTAACATGTTCTTGAAGTACCATGGCGTGGATTAAAAGGAAATTTGGTGAGCGGCCTCCACCTAAACGACTTACTAGGGAAGCTATGC
RCOR1	ENST00000262241, ENST00000570597	GAAGGAACCACACCCCAGTTGTGCCGATTACATTAGTGTTGGCACACAGTCGGGTGCTAGTGTAACACAAATGCCGCGTTGTCTGGGTGTACAGTGTTTG
RELL2_common_T	ENST00000297164, ENST00000521367, ENST00000518856, ENST00000517794, ENST00000444782	GAATGAGGACACAGTAGAGAGGATTGTTCGCTGCATCATCCAGAATGAAGCCAATGCTGAGGCCTTGAAGGAGATGCTGGGGGACAGTGAAGGAGAAGGG
rrad_common	ENST00000299759, ENST00000420652, ENST00000566577, ENST00000568915	ACTCAGACGAGAGCGTTTACAAGGTGCTGCTGCTGGGGGCGCCCGGCGTGGGCAAGAGCGCCCTGGCGCGCATCTTCGGCGGTGTGGAGGACGGGCCTGA
Runx1_sp1	ENST00000300305, ENST00000437180, ENST00000475045, ENST00000416754, ENST00000486278	AGACAGCATATTTGAGTCATTTCCTTCGTACCCACAGTGCTTCATGAGAGAATGCATACTTGGAATGAATCCTTCTAGAGACGTCCACGATGCCAGCACG
Runx1_sp2	ENST00000344691, ENST00000358356, ENST00000399240	CCCTGTCGCCGTCTGGTAGGAGCTGTTTGCAGGGTCCTAACTCAATCGGCTTGTTGTGATGCGTATCCCCGTAGATGCCAGCACGAGCCGCC
SIN3A	ENST00000394949, ENST00000360439, ENST00000394947	CTTCTATGGCAGATGCCAGCAAACATGGTGGTGGAACAGAATCGTTATTTTTTGATAAGGTCCGAAAGGCTCTTCGGAGTGCAGAAGCCTACGAAAATTT
SOX2	ENST00000325404, ENST00000431565	GCCTTTCCAAAAAATAATAATAACAATCATCGGCGGCGGCAGGATCGGCCAGAGGAGGAGGGAAGCGCTTTTTTTTGATCCTGATTCCAGTTTGCCTCTCT
SOX4	ENST00000244745, ENST00000543472	GCATGCAGGCTTTTTGGCTTCCTACCTTGCAACAAAATAATTGCACCAACTCCTTAGTGCCGATTCCGCCCACAGAGAGTCCTGGAGCCACAGTCTTTTT
SP1	ENST00000426431	AGCCCTGGTGCTACTTGCTTGAAGTTTTCAGTGTAAGTACCCTGATGCCTTTTGGACCTTGGGATCAGATCAAGAGTTTTGGAGATCAGGTACCAAGGAA
SREBF2	ENST00000361204, ENST00000424354, ENST00000491541	CTGAGTTGCTGTAGCGTCTTGATTCTCTCCCTGGGTCTGCGTTCCCTCCCCTGGGCCTGACTGAGCCTGCTCATTGTTTTTCCCTTTATTACACAGGACA
SRF	ENST00000265354	AGAGCCTACCTTCACCACCTATATCCAGAAGGGGAGCTTTTTCAGAAACAGGGCAGCAGTGGGGTGAAATTTTCTTAACCCCTAAGACTGCCTTCAGTAG
stat1_common	ENST00000361099, ENST00000392322, ENST00000409465, ENST00000452281, ENST00000540176, ENST00000392323	TGCTGAATGTCACTGAACTTACCCAGAATGCCCTGATTAATGATGAACTAGTGGAGTGGAAGCGGAGACAGCAGAGCGCCTGTATTGGGGGGCCGCCCAA
STAT2	ENST00000314128, ENST00000555665, ENST00000556539, ENST00000557235	CAACATTTTAATAGTTGGTTAGGCTAAACTGGTGCATACTGGCATTGGCCCTTGGTGGGGAGCACAGACACAGGATAGGACTCCATTTCTTTCTTCCATT

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
stat3_common	ENST00000264657, ENST00000389272, ENST00000404395, ENST00000585517, ENST00000588969	GCTGAAATCATCATGGGCTATAAGATCATGGATGCTACCAATATCCTGGTGTCTCCACTGGTCTATCTCTATCCTGACATTCCCAAGGAGGAGGCATTCG
STAT5A	ENST00000588868, ENST00000345506, ENST00000452307	GCCGCCGGTTTGAGTGAGGGTTTCTGAGCTGCTCTGAATTAGTCCTTGCTTGGCTGCTTGGCCTTGGGCTTCATTCAAGTCTATGATGCTGTTGCCCACG
STAT5B	ENST00000293328	GCCTAGAGAGTGGAGATTTTTGATGAAAGGTGTGCTCGCTCTCTGCGTTCTATCTTCTCTCTCCTCCTTGTTCCTGCAAACCACAAGATAAAGGTAGTGG
TAF1_sp1	ENST00000449580, ENST00000373790	CCAATGAAGAAGGATAAGGACCAGGATTCTATTACTGGTGAGAAAGTGGACTTCAGTAGTTCCTCTGACTCAGAATCTGAGATGGGACCT
TAF1_sp2	ENST00000276072, ENST00000423759	CCAATGAAGAAGGATAAGGACCAGGATTCTATTACTGGTGTGTCTGAAAATGGAGAAGGCATCATCTTGCCCTCCATCATTGCCCCTTCC
TCF12_common_T	ENST00000267811, ENST00000333725, ENST00000438423, ENST00000452095, ENST00000557843, ENST00000559609, ENST00000560190, ENST00000561449, ENST00000343827, ENST00000537840, ENST00000543579, ENST00000559703, ENST00000559710, ENST00000561420	CCTGAACAGAAGATAGAAAGGGAGAAGGAGAGGCGGATGGCTAACAATGCCAGAGAACGCTTACGCGTGCGGGATATTAATGAAGCATTCAAAGAGCTTG
TCF3_sp2_T	ENST00000593064, ENST00000588136	CCCACCCAGGCCTGAGCGAAGCCCACAACCCCGCCGGGCACATGTGAAAGTAAACAAAACCTGAAAGCAAGCAACAAAACATACACTTTGTCAGAGAAGA
TCF3_sp3_T	ENST00000262965, ENST00000344749, ENST00000395423, ENST00000585731, ENST00000590684	CCCACCCAGGCCTGAGCGAAGCCCACAACCCCGCCGGGCACATGTGAAAGGTATGCCTCCGTGGGACGAGCCACCCGCTTTCAGCCCTGTGCTCTGGCCC
TCF3_sp4	ENST00000586164, ENST00000593064, ENST00000587425, ENST00000395423, ENST00000592628, ENST00000262965	CACCAGGCTGTCTCGGTCATCCTGAACTTGGAGCAGCAAGTGCGAGAGCGGAACCTGAATCCCAAAGCAGCCTGTTTGAAACGGCGAGAAGAGGAA
TCF3_sp5_T	ENST00000588136, ENST00000453954, ENST00000344749, ENST00000585731, ENST00000585855, ENST00000590684, ENST00000592395	CCTGCAGCAGGCCGTGCAGGTCATCCTGGGGCTGGAGCAGCAGGTGCGAGAGCGGAACCTGAATCCCAAAGCAGCCTGTTTGAAACGGCGAGAAGAGGAA
TCF3_sp6_T	ENST00000395423	CGGGAAGGGCCGGCCCGCCTCCCTGGCCGGGGCGCAGTTCGGAGGTTCAGGCAAGAGCGGTGAGCGGGGCGCCTATGCCTCCTTCGGGAGAGACGCAGGC
TCF3_sp7_T	ENST00000588136, ENST00000344749, ENST00000262965	GTCGCACAGCAGCCTCTCTTCATCCACATTCCTGGGACCGGGACTCGGAGGCAAGAGCGGTGAGCGGGGCGCCTATGCCTCCTTCGGGAGAGACGCAGGC
TNFRSF13B	ENST00000261652, ENST00000437538, ENST00000579315, ENST00000581616, ENST00000582931, ENST00000583789, ENST00000584950	AGGAGCAAGGCAAGTTCTATGACCATCTCCTGAGGGACTGCATCAGCTGTGCCTCCATCTGTGGACAGCACCCTAAGCAATGTGCATACTTCTGTGAGAA

Continued on next page

Table B.6 – continued from previous page

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
TNFRSF13C	ENST00000291232	GTTTGGTGTGCTTGCCTTTGGCTTCAGACCTCACCATCTT-TGACAGCCCTTGAAGGTGGTAGCCCAGCTCCTGTTCCTGT-GCCTTCAAAAGGCTGGGGCA
TNFRSF17_sp1	ENST00000053243	GCTAAGGAAGATAAACTCTGAACCATTAAAGGACGAGTTT-AAAAACACAGGATCAGGTCTCCTGGGCATGGCTAACATTG-ACCTGGAAAAGAGCAGGACT
TNFSF13B	ENST00000375887, ENST00000430559	AACAGGAAATGATCCATTCCCTGTGGTCACTTATTCTAAA-GGCCCCAACCTTCAAAGTTCAAGTAGTGATATGGATGACT-CCACAGAAAGGGAGCAGTCA
USF1_common_T	ENST00000368019, ENST00000368021, ENST00000435396, ENST00000472217, ENST00000473969, ENST00000528768, ENST00000531842, ENST00000368020	AAGCTTGTGATTATATCCAGGAGCTTCGGCAGAGTAACCA-CCGCTTGTCTGAAGAACTGCAGGGACTTGACCAACTGCAG-CTGGACAATGACGTGCTTCG
YY1	ENST00000262238, ENST00000554579, ENST00000554804, ENST00000555735	ACATGCTAAGGCCAAAAACAACCAGTGAAAAGAAGAGAGA-AGACCCTTCTCGACCACGGGAAGCATCTTCCAGAAGTGTG-ATTGGGAATAAATATGCCTC
ZIC3	ENST00000287538	TCAGTTAGTGGCCATGACATCTCAATCTTGTACTTCAAAG-ACTGAGAAGCTGGATTTAATCATCCCTGCCCTACATATAT-AAACATAAGGTAACCTACTG
ZNF217_sp3_T	ENST00000371471	AGTCCCGGCCGCCGCCGCGCGAGGAAATGGCCGAGGAGC-CGGAGCCGCAGGGTTTGGAAATCCCTTGTCTCCAGGTTGC-TGGGATTGACTTCTTGCTCAA

Table B.7: NanoString counts and RPKM values for predominant compatible isoforms.

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
adar1_sp1	2101	0.00	12.26	23.98	27.99	0.00	8.02	4.74	0.00	0.00	15.10	11.80
adar1_sp2	2321	13.18	11.64	1.95	0.42	15.64	6.34	0.00	3.53	18.42	0.73	10.10
atf2_common	4911	9.55	12.50	15.86	0.00	13.87	6.02	12.91	6.90	0.00	5.97	20.10
atf2_sp1	2299	0.00	2.71	4.64	0.00	0.00	6.02	12.91	0.00	0.00	1.50	19.10
atf2_sp2	2156	9.55	12.50	15.86	0.00	13.87	4.21	0.00	6.90	21.38	5.97	54.30
ATP5J_common	16504	33.92	174.53	50.08	61.53	18.58	6.76	32.22	10.38	22.28	30.98	4.80
Bcl11a_sp1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bcl11a_sp2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BCL3	1737	2.79	6.19	2.47	5.93	0.56	0.00	3.77	0.88	3.78	0.00	1.70
BHLHB2	17578	51.58	0.00	87.48	107.52	62.18	39.79	84.31	29.95	0.00	53.85	36.60
Blnk_sp1_T	33	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	10.93	0.00	0.00
Blnk_sp2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CARM1_sp1_T	1168	8.52	19.49	18.04	0.00	0.00	4.85	9.98	0.00	0.00	9.81	3.70
CD19	0	0.00	0.00	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cd79b_sp1_T	24	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
Cd79b_sp2_T	0	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
cdkn1a_sp1_T	6057	21.89	43.57	42.89	53.02	0.00	21.88	0.00	0.00	0.00	29.10	0.00
cdkn1a_sp2_T	5429	0.00	0.00	0.79	0.70	0.00	5.58	11.08	0.00	0.00	1.04	0.00
CEBPA	7109	8.78	0.00	8.37	21.21	0.00	0.00	0.00	7.46	0.00	0.00	0.00
CTCF_common	255	8.98	16.96	16.73	20.36	11.82	9.57	10.03	1.34	0.00	11.56	10.60
CTCF_sp1	563	8.98	16.96	16.73	20.36	11.82	9.57	10.03	2.04	0.00	11.56	8.40
CTCFL_sp1	19	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
CTCFL_sp2	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CTDSL_sp2_T	940	9.70	16.73	22.16	0.00	20.48	18.99	3.87	0.00	17.71	7.57	37.30
CTDSP1_common_T	4117	11.71	14.31	20.84	25.55	20.14	10.70	14.56	8.13	42.12	9.89	22.10
CTDSP2	3330	25.31	23.74	38.76	44.53	0.00	0.00	0.00	14.32	0.00	20.10	0.00
CTDSPL_sp1	4193	5.74	11.59	6.03	0.00	20.48	18.99	3.87	0.00	0.00	5.33	36.70
DES	0	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DNMT1_common_T	1011	3.57	7.06	6.46	11.40	6.62	2.21	2.27	2.19	6.19	4.82	0.00
E2F4	4533	13.40	23.89	25.41	29.67	19.53	16.23	14.25	0.00	17.58	16.07	0.00
E2F6_common	5693	4.10	0.00	0.00	0.00	5.45	3.92	0.00	1.56	2.96	5.26	1.50
ebf1_sp1	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGR1	956	4.39	8.91	8.18	10.04	7.90	7.68	8.31	2.69	0.00	6.46	8.70
EOMES_common	14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.17	0.00	0.00
EP300	1914	8.55	14.06	13.80	16.92	11.11	11.00	10.82	2.78	12.51	8.14	16.70
esr1_common	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
esr2_sp1	33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
esr2_sp2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
ets1_common	561	0.21	0.38	0.15	0.30	0.31	0.00	0.00	0.57	0.32	0.37	0.60
FBXO15	14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
FOS	755	1.26	2.59	2.40	2.96	1.78	1.78	0.00	0.54	0.00	1.89	2.40

Continued on next page

Table B.7 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
foxa2_body	8700	12.33	21.70	12.24	14.81	17.95	14.69	0.00	3.49	0.00	15.41	16.50
foxa2_sp2_T	250	0.00	4.92	8.89	11.10	0.00	0.00	17.36	3.49	0.00	15.41	0.00
FOXA3	1223	8.65	17.66	15.65	19.05	12.66	9.92	12.00	3.16	0.00	10.75	17.10
gabpa_sp1_T	251	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	1.29	0.00
gabpa_sp2_T	237	0.00	0.00	0.26	1.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GATA1_T	12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.04	0.00	0.00
HDAC1	7349	0.00	45.48	38.68	48.94	37.45	9.32	24.48	7.23	0.00	30.27	15.10
HDAC3_sp1	3703	13.39	26.60	23.59	0.00	7.76	7.76	0.00	0.00	12.67	15.36	10.50
HDAC4	902	0.00	0.83	0.09	3.02	0.00	0.00	1.53	0.79	0.00	0.81	3.10
HDAC5	65	0.00	0.00	1.19	0.00	0.00	0.00	0.00	0.45	0.00	13.69	0.00
Hif1a_common	13561	13.67	33.05	18.02	21.75	27.50	18.76	18.17	7.49	26.69	17.02	21.50
HNF1A	1485	4.97	2.70	7.03	10.80	8.25	8.25	9.18	0.75	10.81	3.16	3.60
HNF1B	836	1.28	2.44	2.30	2.90	2.90	2.29	2.84	1.49	2.57	1.82	4.40
Hnf4g_common_T	1906	0.00	3.80	1.98	4.06	3.55	2.90	0.00	0.79	0.00	2.90	3.60
HSF1	4997	18.88	22.36	28.28	39.01	19.78	13.67	9.78	0.35	15.14	8.65	10.90
IGF1R_common_T	2654	4.48	12.12	13.00	11.00	12.10	5.52	7.03	2.55	10.78	5.22	25.30
IKZF1_common	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF1_sp1_T	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF1_sp2	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF3_sp1_T	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
IKZF3_sp2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IL6	0	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IL6receptor_common	2319	0.00	8.48	8.53	9.79	6.58	6.04	6.12	2.34	0.00	5.92	10.30
IL8	84	0.03	0.00	0.23	0.52	0.30	0.30	0.00	0.82	0.00	0.00	0.00
IL8RA	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IL8RB	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IRF8	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
JUND	297	0.00	0.00	45.64	56.16	38.72	38.72	0.00	115.20	0.00	0.00	0.00
KAISO_sp1_T	1368	2.85	4.74	5.59	6.76	4.85	4.85	0.00	2.19	0.00	4.15	0.00
KAP1_common_T	0	49.95	91.94	65.31	2.30	53.39	48.40	48.37	1.20	62.88	68.57	22.00
KLF4	226	0.00	0.60	0.47	0.69	0.54	0.54	0.00	0.50	0.00	0.94	1.20
LEF1_sp1	0	0.00	0.00	0.00	0.05	0.10	0.00	0.00	0.00	0.00	0.00	0.00
LEF1_sp2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LIN28_common	0	0.01	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.01	0.00	0.00
MAX_sp1	2984	4.59	10.64	3.79	0.00	5.16	6.12	6.80	0.00	8.04	4.28	7.90
mef2a_common_T	515	2.03	4.04	4.21	5.45	4.39	2.00	5.08	1.39	4.70	2.93	0.00
mef2a_sp3	1470	2.03	4.04	4.21	5.45	0.00	2.00	5.08	0.31	0.00	2.93	8.00
mef2a_sp4	1090	0.85	1.63	1.47	1.74	4.39	1.27	0.00	1.39	4.70	0.89	7.60
mef2b_sp10_T	81	4.95	0.00	1.18	0.00	0.00	0.00	0.00	0.00	2.28	0.00	8.00
mef2b_sp11_T	248	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00
mef2b_sp12	311	0.00	0.00	1.18	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00

Continued on next page

Table B.7 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
mef2b_sp7_T	215	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.42	0.00
mef2c_sp1	9	2.84	3.43	0.00	0.00	3.85	3.85	0.00	1.40	0.00	0.00	2.20
mef2c_sp3	18	45.25	60.13	0.00	0.00	72.16	60.37	47.51	10.33	0.00	0.00	54.20
mef2c_sp4	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mef2d_sp4	80	0.00	0.00	2.58	3.05	0.00	0.00	0.00	0.00	0.00	3.00	0.00
MYC	28966	0.00	0.00	69.47	78.57	0.00	0.00	0.00	0.00	0.00	40.51	0.00
MYF5	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
MYF6	51	2.00	6.60	0.00	0.00	0.00	3.42	2.77	1.05	7.36	0.00	9.70
MYOD1	32	2.85	9.30	0.00	0.00	3.33	5.07	3.44	1.64	4.76	0.00	0.00
NANOG	64	2.48	0.00	0.01	0.00	0.23	0.23	0.00	0.00	0.19	0.00	0.00
ncor2_sp1_T	1143	5.16	9.87	1.03	0.00	8.05	8.05	0.00	0.00	11.05	2.51	0.00
NFKB1_common	355	1.30	0.50	4.90	4.06	1.01	1.01	3.43	0.00	0.98	3.38	0.00
NOTCH1	27	3.31	6.75	0.03	0.28	6.43	6.43	0.00	1.52	0.00	0.00	8.00
NR2F2_sp2_T	800	2.18	2.16	3.78	4.22	4.32	4.32	0.00	0.96	0.00	7.90	4.40
OCT4_common	284	0.00	0.00	1.80	0.08	0.00	0.00	0.00	0.02	0.00	0.00	0.00
ONECUT1	2013	0.00	0.82	5.99	7.12	0.00	0.00	0.00	0.00	0.00	5.49	0.00
ONECUT2	805	0.77	1.01	3.81	4.05	1.09	0.75	1.15	0.00	0.96	3.01	0.00
pbx1_common	0	2.40	4.57	0.00	0.00	2.91	2.91	2.99	1.23	3.32	0.00	0.00
pbx3_sp1	97	13.50	22.03	0.59	0.70	17.97	15.25	15.28	7.54	20.14	0.00	0.00
pbx3_sp2_T	89	1.30	0.00	1.03	1.22	0.00	0.00	3.43	0.83	0.98	0.86	0.00
PER1	452	0.00	0.00	4.59	5.53	4.50	3.61	0.00	0.00	0.00	3.32	2.70
POLR2A	4377	2.73	5.56	26.52	31.67	0.00	0.00	0.00	0.00	0.00	18.84	3.90
POU5F1_T	94	3.44	0.00	1.80	0.08	0.00	0.00	5.16	0.89	0.00	0.00	7.80
PTEN	1128	1.57	1.23	1.04	4.29	2.28	0.63	0.00	0.26	3.24	0.00	0.00
rbpj_sp2	82	0.41	1.14	5.92	0.00	0.58	0.58	0.60	1.88	0.70	2.40	1.30
RCOR1	2199	0.00	0.00	6.89	8.78	0.00	0.00	0.00	0.00	0.00	5.34	0.00
RELL2_common_T	226	1.47	3.15	0.00	0.00	0.00	0.00	0.00	0.60	0.00	1.73	0.00
rrad_common	82	4.97	8.38	0.74	0.41	6.49	5.61	5.68	1.45	6.62	0.00	10.40
Runx1_sp1	24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Runx1_sp2	338	0.00	0.00	1.44	1.73	0.00	0.00	0.00	1.11	0.00	1.29	4.60
SIN3A	1241	0.00	6.52	8.16	5.74	0.00	0.00	32.10	1.76	0.00	0.00	7.30
SOX2	0	37.41	80.46	0.03	0.00	50.74	28.33	27.72	14.94	0.00	0.00	25.50
SOX4	528	7.76	13.85	2.14	0.06	0.00	0.00	14.38	4.70	0.00	0.00	12.20
SP1	513	7.30	12.28	6.25	7.67	9.05	8.43	7.99	3.29	3.29	5.30	0.00
SREBF2	7451	2.61	8.96	3.23	84.86	6.43	3.72	0.00	1.52	0.00	46.20	6.00
SRF	790	10.56	30.04	14.30	17.67	26.86	8.99	10.63	7.88	21.14	12.14	3.30
stat1_common	1154	0.00	1.74	10.89	13.72	1.49	1.01	0.00	0.42	0.00	8.73	1.70
STAT2	1235	6.81	12.71	6.20	2.85	9.71	7.42	0.00	4.18	0.00	1.60	0.00
stat3_common	7640	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.17	2.60
STAT5A	396	1.78	2.44	1.39	1.11	2.42	0.96	0.00	1.89	3.29	1.26	3.00
STAT5B	1717	5.62	8.75	12.87	15.30	7.71	2.93	5.65	1.19	5.40	8.82	6.70

Continued on next page

Table B.7 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
TAF1_sp1	270	3.95	5.75	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TAF1_sp2	326	0.00	7.68	0.00	2.18	0.00	0.00	0.00	3.25	0.00	2.01	16.90
TCF12_common_T	2427	0.00	7.18	6.72	8.49	0.00	1.82	0.00	0.00	0.00	4.82	0.60
TCF3_sp2_T	1214	3.95	7.68	10.25	0.00	8.29	1.90	0.00	3.25	10.60	4.78	0.00
TCF3_sp3_T	1102	0.00	0.00	5.13	0.00	0.00	1.82	1.39	0.00	0.00	0.00	5.70
TCF3_sp4	1153	3.95	7.18	8.66	0.00	0.00	1.65	4.49	0.00	10.60	4.78	3.90
TCF3_sp5_T	1031	0.00	0.00	10.25	0.00	0.00	0.00	0.00	0.00	0.09	2.45	0.00
TCF3_sp6_T	1111	0.00	0.00	2.46	0.00	0.16	0.00	0.00	0.16	0.00	4.78	0.40
TCF3_sp7_T	1646	0.00	0.00	10.25	0.00	0.00	0.00	0.00	0.00	0.00	2.45	0.00
TNFRSF13B	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.11	0.00	0.00
TNFRSF13C	22	3.58	5.18	0.13	0.20	4.31	2.95	2.83	2.27	3.25	0.00	5.40
TNFRSF17_sp1	0	11.01	13.89	0.00	0.00	17.69	17.69	17.87	5.46	0.00	0.00	10.10
TNFSF13B	84	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
USF1_common_T	623	0.00	13.41	4.73	5.82	0.00	0.00	0.00	5.35	0.00	3.15	0.80
YY1	11801	0.00	12.26	18.36	22.80	0.00	8.02	4.74	0.00	0.00	11.60	11.80
ZIC3	46	13.18	11.64	0.00	0.00	15.64	6.34	0.00	3.53	18.42	0.00	10.10
ZNF217_sp3_T	1401	9.55	12.50	15.95	20.48	13.87	6.02	12.91	6.90	0.00	8.34	20.10

Table B.8: NanoString counts and RPKM values for predominant isoforms.

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
adar1_sp1	2101.00	13.18	12.26	27.99	23.98	15.64	8.02	4.74	10.28	18.42	18.42	29.50
adar1_sp2	2321.00	13.18	12.26	27.99	23.98	15.64	8.02	4.74	10.28	18.42	18.42	29.50
atf2_common	4911.00	9.55	12.50	0.00	15.86	13.87	6.02	12.91	6.90	21.38	21.38	54.30
atf2_sp1	2299.00	9.55	12.50	0.00	15.86	13.87	6.02	12.91	6.90	21.38	21.38	54.30
atf2_sp2	2156.00	9.55	12.50	0.00	15.86	13.87	6.02	12.91	6.90	21.38	21.38	54.30
ATP5J_common	16504.00	33.92	174.53	61.53	50.08	18.58	6.76	32.22	10.38	22.28	22.28	59.50
Bcl11a_sp1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.14	0.00
Bcl11a_sp2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.14	0.00
BCL3	1737.00	2.79	6.19	5.93	3.57	4.49	4.49	3.77	1.56	3.78	3.78	7.20
BHLHB2	17578.00	51.58	82.71	107.52	87.48	62.18	39.79	84.31	29.95	88.55	88.55	82.60
Blnk_sp1_T	33.00	0.01	0.00	0.01	0.00	7.72	7.72	0.00	0.01	10.93	10.93	0.00
Blnk_sp2	0.00	0.01	0.00	0.01	0.00	7.72	7.72	0.00	0.01	10.93	10.93	0.00
CARM1_sp1_T	1168.00	8.52	19.49	0.00	18.04	20.11	9.54	9.98	3.74	9.54	9.54	43.90
CD19	0.00	0.61	2.67	0.07	0.07	0.68	0.00	0.00	7.42	0.19	0.19	1.60
Cd79b_sp1_T	24.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
Cd79b_sp2_T	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
cdkn1a_sp1_T	6057.00	21.89	43.57	53.02	42.89	27.36	21.88	11.36	5.72	37.47	37.47	17.20
cdkn1a_sp2_T	5429.00	21.89	43.57	53.02	42.89	27.36	21.88	11.36	5.72	37.47	37.47	17.20
CEBPA	7109.00	8.78	0.00	21.21	8.37	5.98	5.27	0.00	7.46	6.38	6.38	0.00
CTCF_common	255.00	8.98	16.96	20.36	16.73	11.82	9.57	10.03	3.17	13.75	13.75	21.10
CTCF_sp1	563.00	8.98	16.96	20.36	16.73	11.82	9.57	10.03	3.17	13.75	13.75	21.10
CTCFL_sp1	19.00	0.26	0.00	0.01	0.04	0.10	0.00	0.00	0.72	0.06	0.06	0.20
CTCFL_sp2	11.00	0.26	0.00	0.01	0.04	0.10	0.00	0.00	0.72	0.06	0.06	0.20
CTDSL_sp2_T	940.00	9.70	16.73	0.00	22.16	20.48	18.99	19.44	6.40	17.71	17.71	37.30
CTDSP1_common_T	4117.00	11.71	14.31	25.55	20.84	20.14	10.70	14.56	8.13	42.12	42.12	38.70
CTDSP2	3330.00	25.31	23.74	44.53	38.76	30.80	25.75	24.82	14.32	20.10	20.10	22.30
CTDSPL_sp1	4193.00	9.70	16.73	0.00	22.16	20.48	18.99	19.44	6.40	17.71	17.71	37.30
DES	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DNMT1_common_T	1011.00	3.57	12.12	11.40	6.46	6.62	2.21	2.27	2.89	6.19	6.19	16.40
E2F4	4533.00	13.40	23.89	29.67	25.41	19.53	16.23	14.25	3.34	17.58	17.58	35.10
E2F6_common	5693.00	4.10	7.95	0.00	0.00	5.45	3.92	4.73	2.14	2.96	2.96	8.00
ebf1_sp1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EGR1	956.00	4.39	8.91	10.04	8.18	7.90	7.68	8.31	2.69	9.34	9.34	8.70
EOMES_common	14.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.17	0.17	0.00
EP300	1914.00	8.55	14.06	16.92	13.80	11.11	11.00	10.82	2.78	12.51	12.51	18.20
esr1_common	12.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
esr2_sp1	33.00	0.00	0.00	0.00	0.08	0.09	0.00	0.00	1.20	0.02	0.02	0.20
esr2_sp2	0.00	0.00	0.00	0.00	0.08	0.09	0.00	0.00	1.20	0.02	0.02	0.20
ets1_common	561.00	0.21	0.38	0.30	0.15	0.31	0.00	0.00	0.57	0.32	0.32	0.60
FBXO15	14.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.94	0.01	0.01	0.30
FOS	755.00	1.26	2.59	2.96	2.40	1.78	1.78	0.00	1.00	2.29	2.29	2.40

Continued on next page

Table B.8 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
foxa2_body	8700.00	12.33	21.70	14.81	12.24	17.95	14.69	17.36	3.49	18.41	18.41	33.20
foxa2_sp2_T	250.00	12.33	21.70	14.81	12.24	17.95	14.69	17.36	3.49	18.41	18.41	33.20
FOXA3	1223.00	8.65	17.66	19.05	15.65	12.66	9.92	12.00	3.16	11.93	11.93	17.10
gabpa_sp1_T	251.00	2.39	6.09	2.23	0.26	2.72	1.77	0.00	1.39	3.48	3.48	6.40
gabpa_sp2_T	237.00	2.39	6.09	2.23	0.26	2.72	1.77	0.00	1.39	3.48	3.48	6.40
GATA1_T	12.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.04	11.04	0.00
HDAC1	7349.00	27.70	86.13	48.94	38.68	37.45	9.32	24.48	7.23	27.57	27.57	44.60
HDAC3_sp1	3703.00	13.39	26.60	0.00	23.59	20.75	20.75	19.40	4.83	12.67	12.67	21.70
HDAC4	902.00	0.86	1.52	3.02	2.40	2.22	0.94	1.53	0.79	1.72	1.72	8.30
HDAC5	65.00	12.84	26.93	0.00	47.10	33.16	7.62	15.08	18.30	26.61	26.61	72.30
Hif1a_common	13561.00	13.67	33.05	21.75	18.02	27.50	18.76	18.17	7.49	26.69	26.69	32.50
HNF1A	1485.00	4.97	2.70	10.80	7.03	8.25	8.25	9.18	2.09	10.81	10.81	3.60
HNF1B	836.00	1.28	2.44	2.90	2.30	2.90	2.29	2.84	1.49	2.57	2.57	6.60
Hnf4g_common_T	1906.00	2.36	3.80	4.06	1.98	3.55	2.90	2.04	0.82	1.22	1.22	4.20
HSF1	4997.00	18.88	22.36	39.01	28.28	19.78	13.67	9.78	5.11	15.14	15.14	65.50
IGF1R_common_T	2654.00	9.78	12.12	11.00	13.00	17.78	17.78	7.03	2.55	14.32	14.32	25.30
IKZF1_common	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF1_sp1_T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF1_sp2	16.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IKZF3_sp1_T	0.00	0.00	0.00	0.00	0.02	11.14	3.41	0.00	0.00	0.06	0.06	0.00
IKZF3_sp2	0.00	0.00	0.00	0.00	0.02	11.14	3.41	0.00	0.00	0.06	0.06	0.00
IL6	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
IL6receptor_common	2319.00	4.37	8.48	9.79	8.53	6.58	6.04	6.12	3.05	5.37	5.37	14.90
IL8	84.00	0.03	0.48	0.52	0.23	0.30	0.30	0.00	0.82	0.16	0.16	0.50
IL8RA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00
IL8RB	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.26	0.01	0.01	0.00
IRF8	18.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
JUND	297.00	25.22	0.00	56.16	45.64	38.72	38.72	0.00	115.20	46.83	46.83	0.00
KAISO_sp1_T	1368.00	2.85	4.74	6.76	5.59	4.85	4.85	0.00	2.19	4.92	4.92	6.00
KAP1_common_T	0.00	49.95	91.94	119.82	65.31	53.39	48.40	48.37	23.63	62.88	62.88	145.10
KLF4	226.00	0.14	0.60	0.69	0.47	0.54	0.54	0.00	0.50	0.37	0.37	1.20
LEF1_sp1	0.00	0.03	0.00	0.05	0.01	0.10	0.00	0.00	0.00	0.04	0.04	0.30
LEF1_sp2	0.00	0.03	0.00	0.05	0.01	0.10	0.00	0.00	0.00	0.04	0.04	0.30
LIN28_common	0.00	0.01	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.01	0.01	0.00
MAX_sp1	2984.00	4.59	10.64	0.00	3.79	5.16	6.12	6.80	2.23	8.04	8.04	7.90
mef2a_common_T	515.00	2.03	4.04	5.45	4.21	4.39	2.00	5.08	1.39	4.70	4.70	9.20
mef2a_sp3	1470.00	2.03	4.04	5.45	4.21	4.39	2.00	5.08	1.39	4.70	4.70	9.20
mef2a_sp4	1090.00	2.03	4.04	5.45	4.21	4.39	2.00	5.08	1.39	4.70	4.70	9.20
mef2b_sp10_T	81.00	4.95	3.46	0.00	1.18	1.38	1.09	1.29	1.76	2.28	2.28	8.00
mef2b_sp11_T	248.00	0.00	0.00	0.00	3.17	0.00	0.00	0.00	0.00	0.00	2.28	0.00
mef2b_sp12	311.00	0.00	0.00	0.00	3.17	0.00	0.00	0.00	0.00	0.00	2.28	0.00

Continued on next page

Table B.8 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
mef2b_sp7_T	215.00	0.00	0.00	0.00	3.17	0.00	0.00	0.00	0.00	0.00	2.28	0.00
mef2c_sp1	9.00	2.84	3.43	0.00	0.00	3.85	3.85	0.00	1.40	3.90	0.00	8.20
mef2c_sp3	18.00	45.25	60.13	0.00	0.00	72.16	60.37	47.51	16.87	103.10	0.00	54.20
mef2c_sp4	16.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mef2d_sp4	80.00	0.73	0.00	3.05	2.58	0.00	0.00	0.00	0.00	0.00	3.90	0.00
MYC	28966.00	0.00	0.00	78.57	69.47	0.00	0.00	0.00	0.00	0.00	103.10	0.00
MYF5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
MYF6	51.00	2.00	6.60	0.00	0.00	8.09	3.42	2.77	5.11	7.36	0.00	24.70
MYOD1	32.00	2.85	9.30	0.00	0.00	3.33	5.07	3.44	1.64	4.76	0.00	9.70
NANOG	64.00	2.48	0.50	0.00	0.01	0.23	0.23	0.00	0.31	0.19	0.07	0.50
ncor2_sp1_T	1143.00	5.16	9.87	0.00	7.65	8.05	8.05	9.02	2.37	11.05	7.36	5.80
NFKB1_common	355.00	1.30	0.50	4.06	4.90	1.01	1.01	3.43	0.83	0.98	4.76	0.00
NOTCH1	27.00	3.31	6.75	0.28	0.06	6.43	6.43	4.97	1.52	4.42	0.19	8.00
NR2F2_sp2_T	800.00	2.18	2.16	6.08	4.06	4.32	4.32	0.00	0.96	1.90	11.05	4.40
OCT4_common	284.00	0.00	0.00	0.08	1.80	0.00	0.00	0.00	0.02	0.01	0.98	0.00
ONECUT1	2013.00	0.77	1.01	7.12	5.99	1.09	0.75	1.15	0.48	0.96	4.42	2.10
ONECUT2	805.00	0.77	1.01	4.05	3.81	1.09	0.75	1.15	0.48	0.96	1.90	2.10
pbx1_common	0.00	2.40	4.57	0.00	0.00	2.91	2.91	10.53	1.60	3.32	0.01	2.30
pbx3_sp1	97.00	13.50	22.03	1.22	1.03	17.97	15.25	15.28	7.81	20.14	0.96	30.50
pbx3_sp2_T	89.00	1.30	0.50	1.22	1.03	1.01	1.01	3.43	0.83	0.98	0.96	0.00
PER1	452.00	5.49	7.90	5.53	4.59	4.50	3.61	0.00	3.31	35.66	3.32	7.00
POLR2A	4377.00	2.93	11.23	31.67	26.52	6.30	3.90	6.44	2.16	1.79	20.14	9.70
POU5F1_T	94.00	3.44	6.24	0.08	1.80	4.72	4.72	5.16	0.89	2.20	0.98	7.80
PTEN	1128.00	1.57	1.23	4.29	4.05	2.28	0.63	0.00	0.70	3.24	35.66	3.10
rbpj_sp2	82.00	0.41	1.14	0.00	5.92	0.58	0.58	0.60	1.88	0.70	1.79	21.00
RCOR1	2199.00	1.47	3.15	8.78	6.89	1.99	1.74	2.39	0.99	1.86	2.20	4.60
RELL2_common_T	226.00	1.47	3.15	0.00	1.06	1.99	1.74	2.39	0.99	1.86	3.24	4.60
rrad_common	82.00	4.97	8.38	0.41	0.74	6.49	5.61	5.68	1.45	6.62	0.70	13.80
Runx1_sp1	24.00	0.01	0.00	1.73	1.44	0.02	0.02	0.00	0.01	0.02	1.86	0.00
Runx1_sp2	338.00	0.61	0.00	1.73	1.44	1.57	1.57	0.00	1.11	1.95	1.86	4.60
SIN3A	1241.00	61.27	6.52	5.74	8.16	5.82	5.80	32.10	1.76	6.87	6.62	8.80
SOX2	0.00	37.41	80.46	0.00	0.03	50.74	28.33	27.72	14.94	60.26	0.02	104.40
SOX4	528.00	7.76	13.85	0.06	2.14	12.00	12.00	14.38	5.38	11.91	1.95	12.20
SP1	513.00	7.30	16.60	7.67	6.25	9.05	8.43	7.99	3.29	3.29	6.87	18.90
SREBF2	7451.00	2.61	8.96	84.86	70.53	6.43	3.72	1.66	2.55	5.45	60.26	9.50
SRF	790.00	10.56	30.04	17.67	14.30	26.86	8.99	10.63	10.48	21.14	11.91	66.80
stat1_common	1154.00	0.91	1.74	13.72	10.89	1.49	1.01	1.71	0.42	5.16	3.29	2.20
STAT2	1235.00	6.81	12.71	2.85	6.20	9.71	7.42	8.86	4.18	8.57	5.45	19.70
stat3_common	7640.00	1.78	5.08	0.00	0.00	2.42	0.96	10.92	1.89	3.29	21.14	5.00
STAT5A	396.00	1.78	5.08	1.11	1.39	2.42	0.96	10.92	1.89	3.29	5.16	5.00
STAT5B	1717.00	5.62	9.24	15.30	12.87	7.71	2.93	5.65	3.92	5.40	8.57	13.60

Continued on next page

Table B.8 – continued from previous page

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon ends	iReckon full	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
TAF1_sp1	270.00	3.95	7.68	3.53	6.24	8.29	1.90	4.49	3.25	10.60	3.29	16.90
TAF1_sp2	326.00	3.95	7.68	3.53	6.24	8.29	1.90	4.49	3.25	10.60	3.29	16.90
TCF12_common_T	2427.00	3.95	7.68	8.49	6.72	8.29	1.90	4.49	3.25	10.60	5.40	16.90
TCF3_sp2_T	1214.00	3.95	7.68	0.00	10.25	8.29	1.90	4.49	3.25	10.60	10.60	16.90
TCF3_sp3_T	1102.00	3.95	7.68	0.00	10.25	8.29	1.90	4.49	3.25	10.60	10.60	16.90
TCF3_sp4	1153.00	3.95	7.68	0.00	10.25	8.29	1.90	4.49	3.25	10.60	10.60	16.90
TCF3_sp5_T	1031.00	0.00	0.00	0.00	10.25	0.34	0.00	0.00	0.03	0.09	10.60	0.00
TCF3_sp6_T	1111.00	0.00	0.00	0.00	10.25	0.16	0.00	0.00	0.16	0.11	10.60	0.40
TCF3_sp7_T	1646.00	0.00	0.00	0.00	10.25	0.00	0.00	0.00	0.00	0.00	10.60	0.00
TNFRSF13B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.11	0.09	0.00
TNFRSF13C	22.00	3.58	5.18	0.20	0.13	4.31	2.95	2.83	2.27	3.25	0.11	13.50
TNFRSF17_sp1	0.00	11.01	13.89	0.00	0.00	17.69	17.69	17.87	5.46	10.59	0.00	10.10
TNFSF13B	84.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.17	2.11	0.00
USF1_common_T	623.00	7.43	13.41	5.82	4.73	12.48	9.75	0.00	5.35	17.10	3.25	11.20
YY1	11801.00	13.18	12.26	22.80	18.36	15.64	8.02	4.74	10.28	18.42	10.59	29.50
ZIC3	46.00	13.18	12.26	0.00	0.00	15.64	8.02	4.74	10.28	18.42	0.17	29.50
ZNF217_sp3_T	1401.00	9.55	12.50	20.48	15.95	13.87	6.02	12.91	6.90	21.38	17.10	54.30

Appendix C

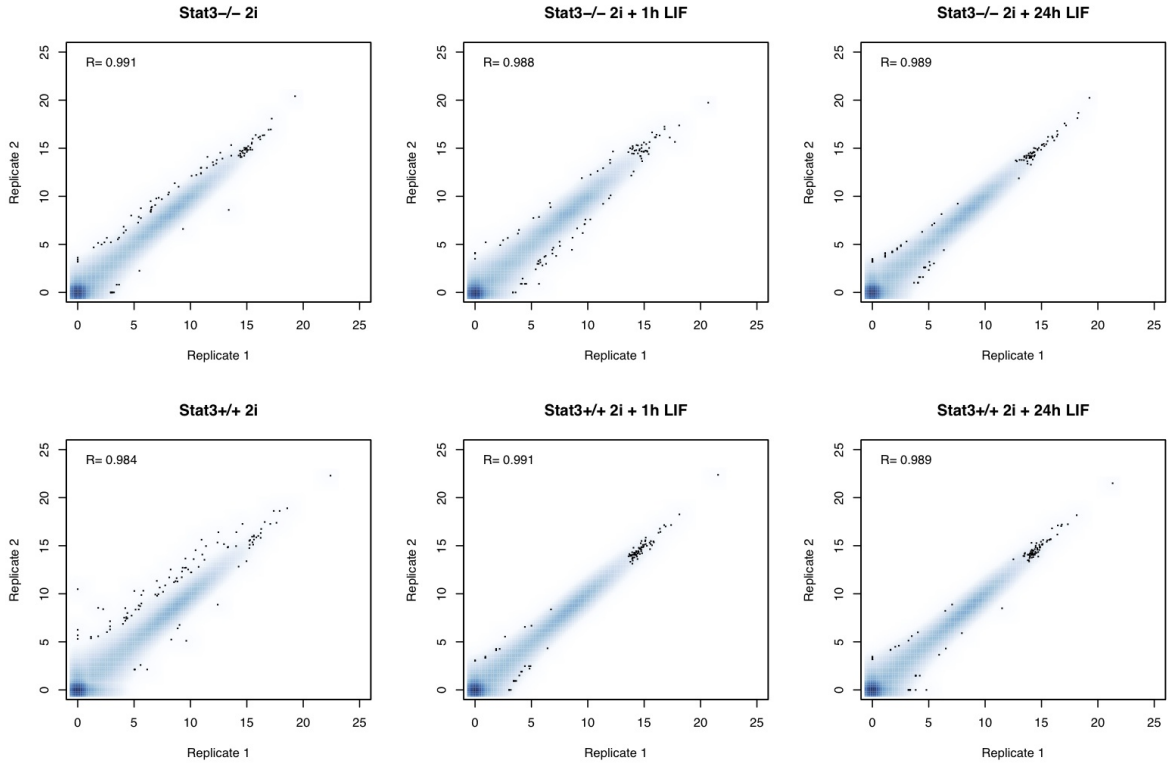


Figure C.1: Comparing gene expression between replicates. Gene counts were log2 transformed after adding the count of one to avoid infinite numbers. Pearson's R was computed on the transformed read counts and is displayed in the top left of each correlation plot.

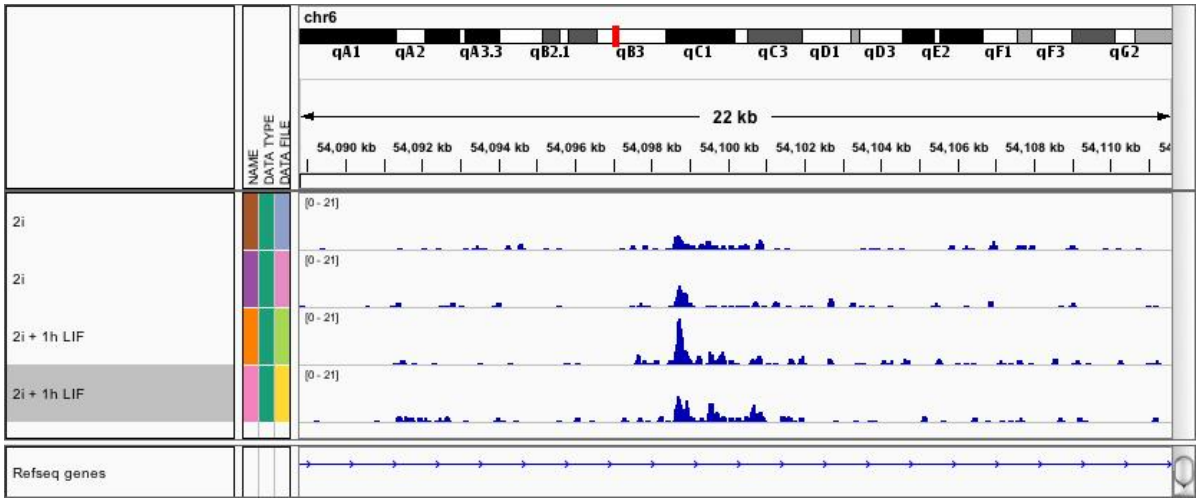


Figure C.2: Intronic transcript at the *Chn2* locus.

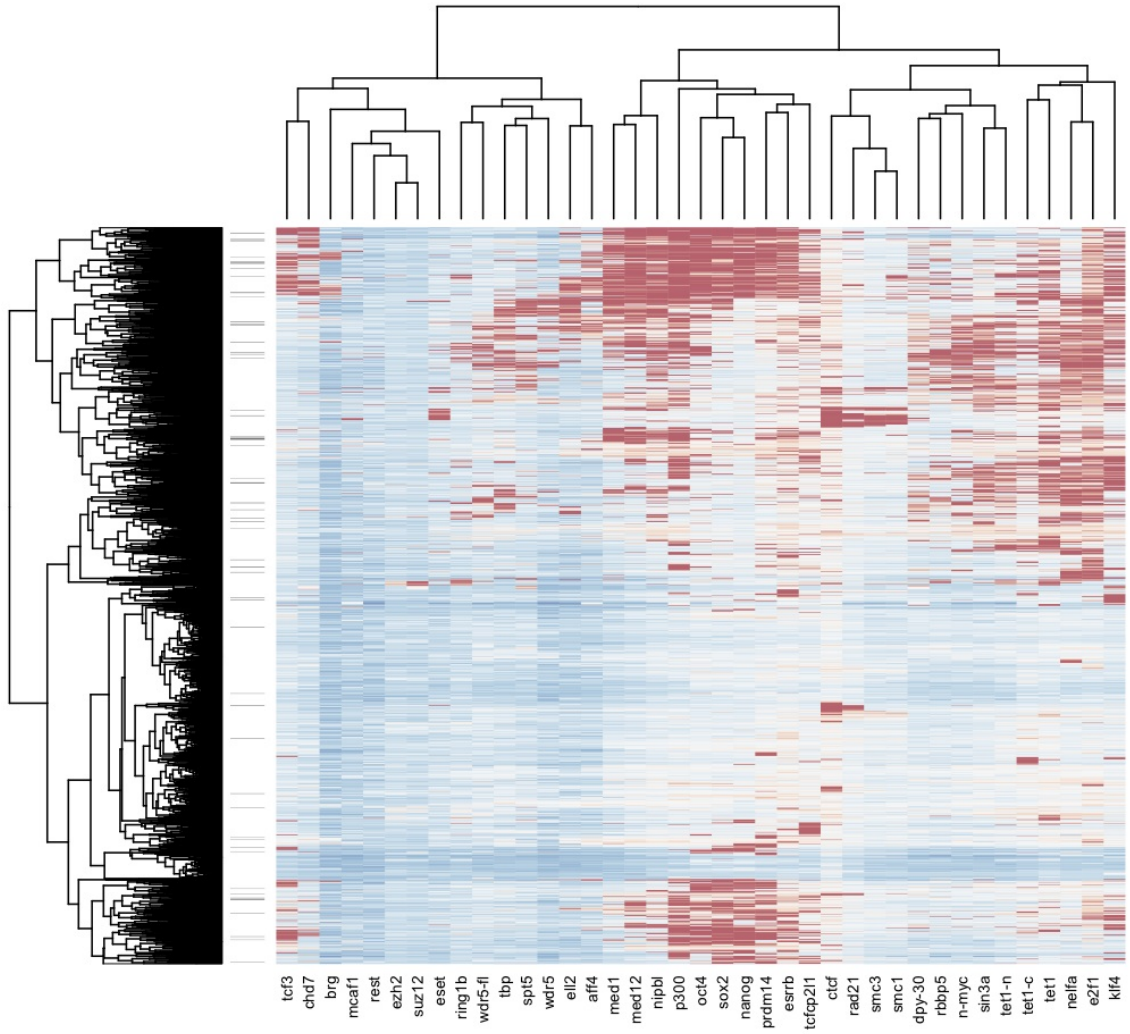


Figure C.3: Hierarchical clustering of transcriptional regulators in ES cells based on distance to Stat3 peaks. For each Stat3 peak the distance to the closest binding site of other factors was derived. The distance matrix was used as input for hierarchical clustering. Red colours indicate close proximity while blue colours indicate larger distances. The column on the left indicates peaks associated with upregulated genes.

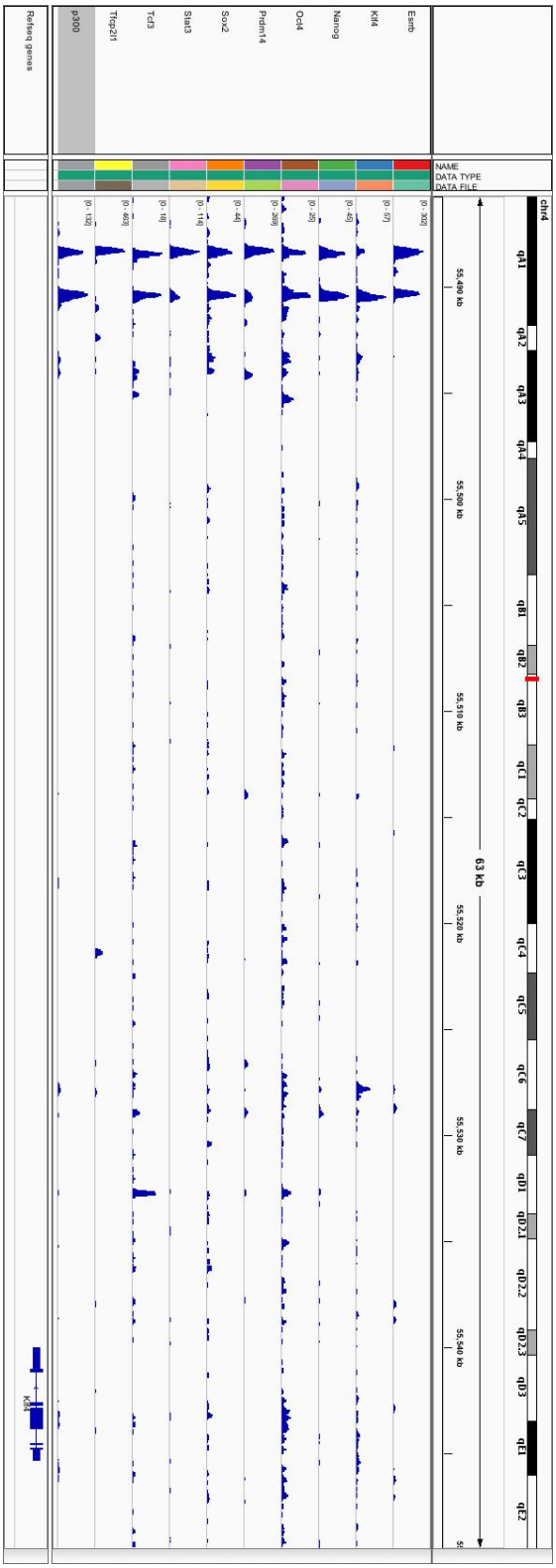


Figure C.4: Cooperative binding at the *Klf4* locus

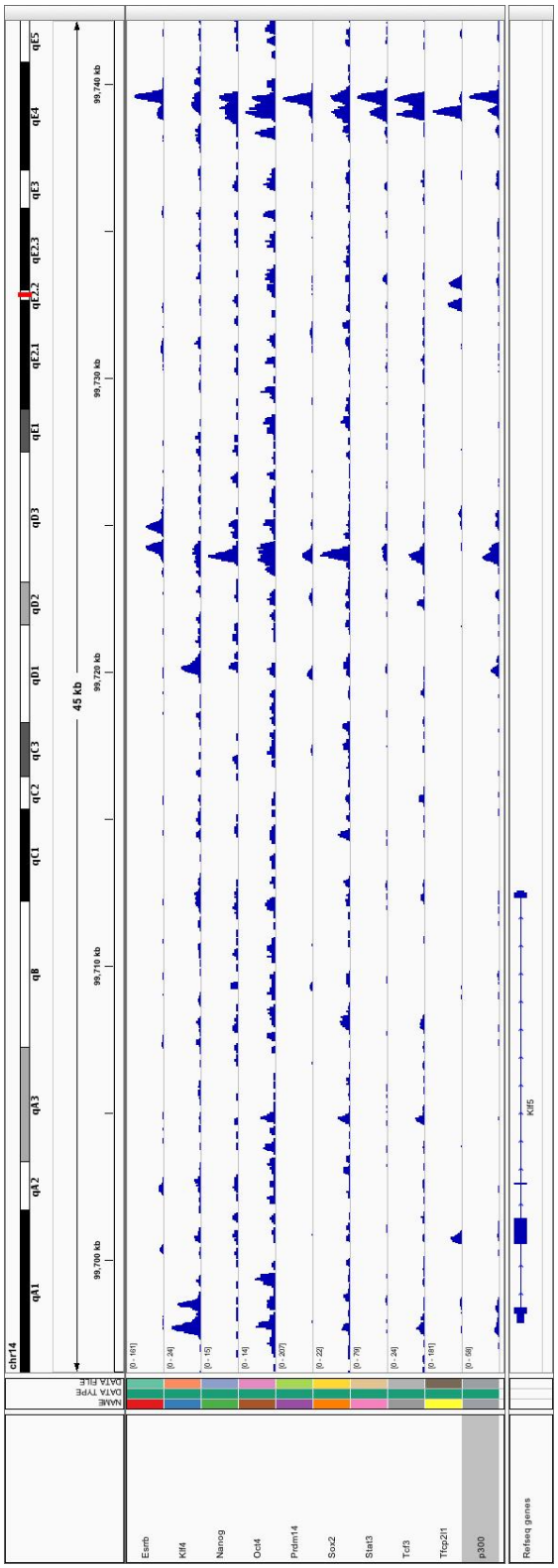


Figure C.5: Cooperative binding at the *Klf5* locus

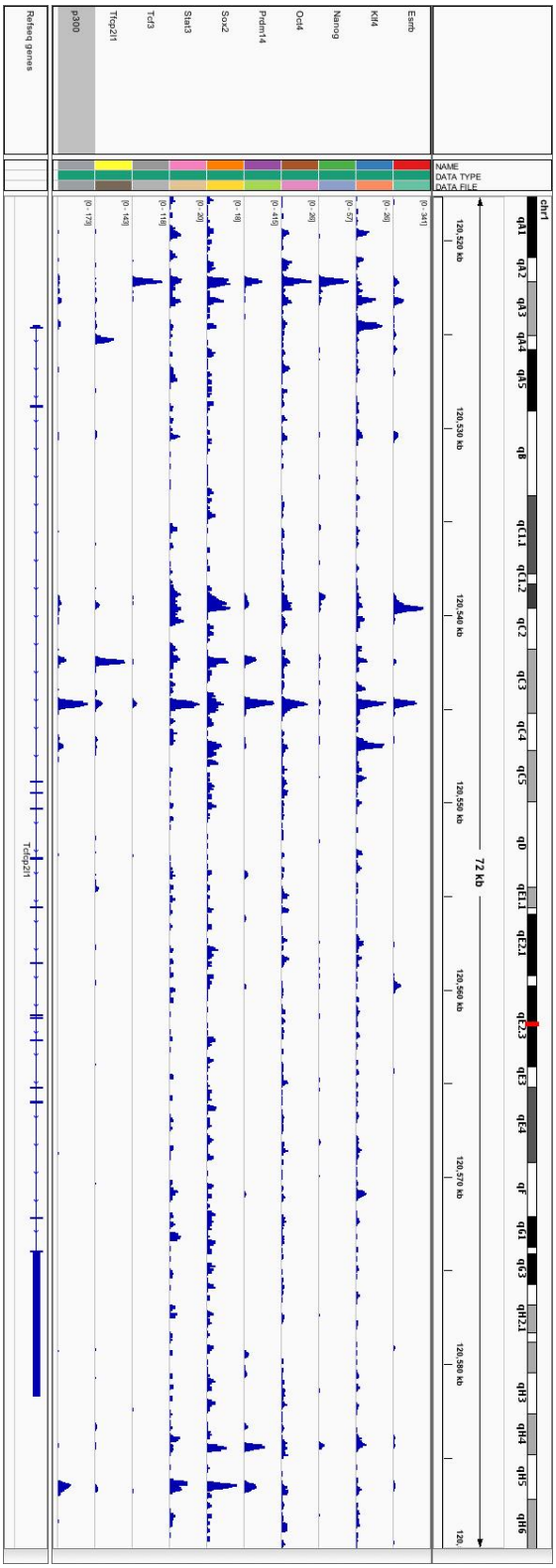


Figure C.6: Cooperative binding at the *Tfcp2l1* locus

Table C.1: Information on enclosed CD.

File name	**Description**
mm10.ens69.deseq_table.[xls/txt]	Raw gene counts, FPKM values average across biological replicates, fold-changes and *P*-values for comparisons disussed in Chapter 4.
stat3_annot.[xls/txt]	Stat3 peaks annotated with Ensembl genes
pstat3_annot.[xls/txt]	pStat3 peaks annotated with Ensembl genes
tfcp2l1_annot.[xls/txt]	Tfcp2l1 peaks annotated with Ensembl genes
tfcp2l1_chen_annot.[xls/txt]	Tfcp2l1 peaks annotated with Ensembl genes (Chen et al., 2008)
klf4_chen_annot.[xls/txt]	Klf4 peaks annotated with Ensembl genes (Chen et al., 2008)

Full information about differentially expressed genes and Stat3, Klf4 and Tfcp2l1 bound genes can be found on the enclosed CD. Table C.1 lists the files that can be found on the CD.

References

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean (2010). "A map of human genome variation from population-scale sequencing." In: *Nature* 467.7319 (Oct. 2010), pp. 1061–1073.

Akira, S, Y Nishio, M Inoue, X. J. Wang, S Wei, T Matsusaka, K Yoshida, T Sudo, M Naruto, and T Kishimoto (1994). "Molecular cloning of APRF, a novel IFN-stimulated gene factor 3 p91-related transcription factor involved in the gp130-mediated signaling pathway." In: *Cell* 77.1 (Apr. 1994), pp. 63–71.

Aksoy, I., C. Sakabedoyan, P.-Y. Bourillot, A. B. Malashicheva, J. Mancip, K. Knoblauch, M. Afanassieff, and P. Savatier (2007). "Self-renewal of murine embryonic stem cells is supported by the serine/threonine kinases Pim-1 and Pim-3." In: *Stem cells (Dayton, Ohio)* 25.12 (Dec. 2007), pp. 2996–3004.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour (2006). "Microarray data analysis: from disarray to consolidation and consensus." In: *Nature reviews. Genetics* 7.1 (Jan. 2006), pp. 55–65.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." In: *Genome biology* 11.10 (2010), R106.

Anders, S., A. Reyes, and W. Huber (2012). "Detecting differential usage of exons from RNA-seq data." In: *Genome research* 22.10 (Oct. 2012), pp. 2008–2017.

Ang, Y.-S., S.-Y. Tsai, D.-F. Lee, J. Monk, J. Su, K. Ratnakumar, J. Ding, Y. Ge, H. Darr, B. Chang, J. Wang, M. Rendl, E. Bernstein, C. Schaniel, and I. R. Lemischka (2011). "Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network." In: *Cell* 145.2 (Apr. 2011), pp. 183–197.

Azuara, V., P. Perry, S. Sauer, M. Spivakov, H. F. Jørgensen, R. M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, and A. G. Fisher (2006). "Chromatin signatures of pluripotent cell lines." In: *Nature cell biology* 8.5 (May 2006), pp. 532–538.

Bar-Nur, O., H. A. Russ, S. Efrat, and N. Benvenisty (2011). "Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells." In: *Cell stem cell* 9.1 (July 2011), pp. 17–23.

Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao (2007). "High-resolution profiling of histone methylations in the human genome." In: *Cell* 129.4 (May 2007), pp. 823–837.

Behr, J., A. Kahles, Y. Zhong, V. T. Sreedharan, P. Drewe, and G. Rätsch (2013). "MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples." In: *Bioinformatics (Oxford, England)* (Sept. 2013).

Benjamini, Y and Y Hochberg (1995). "JSTOR: Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289-300". In: *Journal of the Royal Statistical Society Series B (...* (1995).

Bentley, D. R. et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." In: *Nature* 456.7218 (Nov. 2008), pp. 53–59.

Berg, D. L. C. van den, T. Snoek, N. P. Mullin, A. Yates, K. Bezstarosti, J. Demmers, I. Chambers, and R. A. Poot (2010). "An Oct4-centered protein interaction network in embryonic stem cells." In: *Cell stem cell* 6.4 (Apr. 2010), pp. 369–381.

Bernard, E., L. Jacob, J. Mairal, and J.-P. Vert (2013). "Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows". In: (Mar. 2013).

Bernemann, C., B. Greber, K. Ko, J. Sterneckert, D. W. Han, M. J. Araúzo-Bravo, and H. R. Schöler (2011). "Distinct developmental ground states of epiblast stem cell lines determine different pluripotency features." In: *Stem cells (Dayton, Ohio)* 29.10 (Oct. 2011), pp. 1496–1503.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." In: *Cell* 125.2 (Apr. 2006), pp. 315–326.

Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder (2004). "Global identification of human transcribed sequences with genome tiling arrays." In: *Science (New York, N.Y.)* 306.5705 (Dec. 2004), pp. 2242–2246.

Blanco, E., G. Parra, and R. Guigo (2007). "Using geneid to identify genes." In: *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 4 (June 2007), Unit 4.3.

Blattner, F. R., G Plunkett, C. A. Bloch, N. T. Perna, V Burland, M Riley, J Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B Mau, and Y Shao (1997). "The complete genome sequence of Escherichia coli K-12." In: *Science (New York, N.Y.)* 277.5331 (Sept. 1997), pp. 1453–1462.

Bock, C. (2012). "Analysing and interpreting DNA methylation data." In: *Nature reviews. Genetics* 13.10 (Oct. 2012), pp. 705–719.

Boeuf, H, C Hauss, F. D. Graeve, N Baran, and C Kedinger (1997). "Leukemia inhibitory factor-dependent transcriptional activation in embryonic stem cells." In: *The Journal of cell biology* 138.6 (Sept. 1997), pp. 1207–1217.

Bohnert, R. and G. Rätsch (2010). "rQuant.web: a tool for RNA-Seq-based transcript quantitation." In: *Nucleic acids research* 38.Web Server issue (July 2010), W348–51.

Boiani, M. and H. R. Schöler (2005). "Regulatory networks in embryo-derived pluripotent stem cells." In: *Nature reviews. Molecular cell biology* 6.11 (Nov. 2005), pp. 872–884.

Bourillot, P.-Y., I. Aksoy, V. Schreiber, F. Wianny, H. Schulz, O. Hummel, N. Hubner, and P. Savatier (2009). "Novel STAT3 target genes exert distinct roles in the inhibition of mesoderm and endoderm differentiation in cooperation with Nanog." In: *Stem cells (Dayton, Ohio)* 27.8 (Aug. 2009), pp. 1760–1771.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." In: *Cell* 122.6 (Sept. 2005), pp. 947–956.

Boyle, K., J.-G. Zhang, S. E. Nicholson, E. Trounson, J. J. Babon, E. J. McManus, N. A. Nicola, and L. Robb (2009). "ScienceDirect.com - Cellular Signalling - Deletion of the SOCS box of suppressor of cytokine signaling 3 (SOCS3) in embryonic stem cells reveals SOCS box-dependent regulation of JAK but not STAT phosphorylation". In: *Cellular signalling* 21.3 (Mar. 2009), pp. 394–404.

Bradford, J. R., Y. Hey, T. Yates, Y. Li, S. D. Pepper, and C. J. Miller (2010). "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling." In: *BMC genomics* 11 (2010), p. 282.

Brent, M. R. (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation." In: *Nature reviews. Genetics* 9.1 (Jan. 2008), pp. 62–73.

Brons, I. G. M., L. E. Smithers, M. W. B. Trotter, P. Rugg-Gunn, B. Sun, S. M. Chuva de Sousa Lopes, S. K. Howlett, A. Clarkson, L. Ahrlund-Richter, R. A. Pedersen, and L. Vallier (2007). "Derivation of pluripotent epiblast stem cells from mammalian embryos." In: *Nature* 448.7150 (July 2007), pp. 191–195.

Brook, F. A. and R. L. Gardner (1997). "The origin and efficient derivation of embryonic stem cells in the mouse." In: *Proceedings of the National Academy of Sciences of the United States of America* 94.11 (May 1997), pp. 5709–5712.

Buecker, C., H.-H. Chen, J. M. Polo, L. Dahéron, L. Bu, T. S. Barakat, P. Okwieka, A. Porter, J. Gribnau, K. Hochedlinger, and N. Geijsen (2010). "A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells." In: *Cell stem cell* 6.6 (June 2010), pp. 535–546.

Buehr, M., S. Meek, K. Blair, J. Yang, J. Ure, J. Silva, R. McLay, J. Hall, Q.-L. Ying, and A. Smith (2008). "Capture of authentic embryonic stem cells from rat blastocysts." In: *Cell* 135.7 (Dec. 2008), pp. 1287–1298.

Burdon, T, C Stracey, I Chambers, J Nichols, and A Smith (1999). "Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells." In: *Developmental biology* 210.1 (June 1999), pp. 30–43.

Burdon, T., A. Smith, and P. Savatier (2002). "Signalling, cell cycle and pluripotency in embryonic stem cells." In: *Trends in cell biology* 12.9 (Sept. 2002), pp. 432–438.

Campagna, D., A. Albiero, A. Bilardi, E. Caniato, C. Forcato, S. Manavski, N. Vitulo, and G. Valle (2009). "PASS: a program to align short sequences." In: *Bioinformatics (Oxford, England)* 25.7 (Apr. 2009), pp. 967–968.

Chamberlain, S. J., D. Yee, and T. Magnuson (2008). "Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency." In: *Stem cells (Dayton, Ohio)* 26.6 (June 2008), pp. 1496–1505.

Chambers, I., D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie, and A. Smith (2003). "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." In: *Cell* 113.5 (May 2003), pp. 643–655.

Chen, X. et al. (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." In: *Cell* 133.6 (June 2008), pp. 1106–1117.

Cheng, C., R. Min, and M. Gerstein (2011). "TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles." In: *Bioinformatics (Oxford, England)* 27.23 (Dec. 2011), pp. 3221–3227.

Chia, N.-Y., Y.-S. Chan, B. Feng, X. Lu, Y. L. Orlov, D. Moreau, P. Kumar, L. Yang, J. Jiang, M.-S. Lau, M. Huss, B. S. Soh, P. Kraus, P. Li, T. Lufkin, B. Lim, N. D. Clarke, F. Bard, and H.-H. Ng (2010). "A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity." In: *Nature* 468.7321 (Nov. 2010), pp. 316–320.

Cinelli, P., E. A. Casanova, S. Uhlig, P. Lochmatter, T. Matsuda, T. Yokota, T. Rülicke, B. Ledermann, and K. Bürki (2008). "Expression profiling in transgenic FVB/N embryonic stem cells overexpressing STAT3." In: *BMC developmental biology* 8 (2008), p. 57.

Cloonan, N., A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing." In: *Nature methods* 5.7 (July 2008), pp. 613–619.

Coghlan, A., T. J. Fiedler, S. J. McKay, P. Flicek, T. W. Harris, D. Blasiar, nGASP Consortium, and L. D. Stein (2008). "nGASP–the nematode genome annotation assessment project." In: *BMC bioinformatics* 9 (2008), p. 549.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." In: *Nature* 452.7184 (Mar. 2008), pp. 215–219.

Cole, M. F., S. E. Johnstone, J. J. Newman, M. H. Kagey, and R. A. Young (2008). "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells." In: *Genes & development* 22.6 (Mar. 2008), pp. 746–755.

Correa-Cerro, L. S., Y. Piao, A. A. Sharov, A. Nishiyama, J. S. Cadet, H. Yu, L. V. Sharova, L. Xin, H. G. Hoang, M. Thomas, Y. Qian, D. B. Dudekula, E. Meyers, B. Y. Binder, G. Mowrer, U. Bassey, D. L. Longo, D. Schlessinger, and M. S. H. Ko (2011). "Generation of mouse ES cell lines engineered for the forced induction of transcription factors." In: *Scientific reports* 1 (2011), p. 167.

Crick, F. H., L Barnett, S Brenner, and R. J. Watts-tobin (1961). "General nature of the genetic code for proteins." In: *Nature* 192 (Dec. 1961), pp. 1227–1232.

Cullum, R., O. Alder, and P. A. Hoodless (2011). "The next generation: using new sequencing technologies to analyse gene regulation." In: *Respirology (Carlton, Vic.)* 16.2 (Feb. 2011), pp. 210–222.

Dahéron, L., S. L. Opitz, H. Zaehres, M. W. Lensch, W. M. Lensch, P. W. Andrews, J. Itskovitz-Eldor, and G. Q. Daley (2004). "LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells." In: *Stem cells (Dayton, Ohio)* 22.5 (2004), pp. 770–778.

Darnell, J. E. (1997). "STATs and gene regulation." In: *Science (New York, N.Y.)* 277.5332 (Sept. 1997), pp. 1630–1635.

David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz (2006). "A high-resolution map of transcription in the yeast genome." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.14 (Apr. 2006), pp. 5320–5325.

Dedeepiya, V. D., Y. Y. Rao, G. A. Jayakrishnan, J. K. B. C. Parthiban, S. Baskar, S. R. Manjunath, R. Senthilkumar, and S. J. K. Abraham (2012). "Index of CD34+ Cells and Mononuclear Cells in the Bone Marrow of Spinal Cord Injury Patients of Different Age Groups: A Comparative Analysis." In: *Bone marrow research* 2012 (2012), p. 787414.

Deplus, R., C. Brenner, W. A. Burgers, P. Putmans, T. Kouzarides, Y. de Launoit, and F. Fuks (2002). "Dnmt3L is a transcriptional repressor that recruits histone deacetylase." In: *Nucleic acids research* 30.17 (Sept. 2002), pp. 3831–3838.

Derrien, T., J. Estellé, S. Marco-Sola, D. G. Knowles, E. Raineri, R. Guigo, and P. Ribeca (2012). "Fast computation and applications of genome mappability." In: *PloS one* 7.1 (2012), e30377.

Di Giammartino, D. C., K. Nishida, and J. L. Manley (2011). "Mechanisms and consequences of alternative polyadenylation." In: *Molecular cell* 43.6 (Sept. 2011), pp. 853–866.

Djebali, S. et al. (2012). "Landscape of transcription in human cells." In: *Nature* 489.7414 (Sept. 2012), pp. 101–108.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics (Oxford, England)* 29.1 (Jan. 2013), pp. 15–21.

Drewe, P., O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, and G. Rätsch (2013). "Accurate detection of differential RNA processing." In: *Nucleic acids research* 41.10 (May 2013), pp. 5189–5198.

Du, J., J. Leng, L. Habegger, A. Sboner, D. McDermott, and M. Gerstein (2012). "IQSeq: integrated isoform quantification analysis based on next-generation sequencing." In: *PloS one* 7.1 (2012), e29175.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." In: *Bioinformatics (Oxford, England)* 21.16 (Aug. 2005), pp. 3439–3440.

Efroni, S., R. Duttagupta, J. Cheng, H. Dehghani, D. J. Hoeppner, C. Dash, D. P. Bazett-Jones, S. Le Grice, R. D. G. McKay, K. H. Buetow, T. R. Gingeras, T. Misteli, and E. Meshorer (2008). "Global transcription in pluripotent embryonic stem cells." In: *Cell stem cell* 2.5 (May 2008), pp. 437–447.

Eichner, J. (2013). "Inference of alternative splicing from tiling array data." In: *Methods in molecular biology (Clifton, N.J.)* 1067 (2013), pp. 143–164.

ENCODE Project Consortium et al. (2007). "Identification and analysis of functional elements in 1the human genome by the ENCODE pilot project." In: *Nature* 447.7146 (June 2007), pp. 799–816.

Evans, M. J. and M. H. Kaufman (1981). "Establishment in culture of pluripotential cells from mouse embryos." In: *Nature* 292.5819 (July 1981), pp. 154–156.

Farthing, C. R., G. Ficz, R. K. Ng, C.-F. Chan, S. Andrews, W. Dean, M. Hemberger, and W. Reik (2008). "Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes." In: *PLoS genetics* 4.6 (June 2008), e1000116.

Fejes, A. P., G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. M. Jones (2008). "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology." In: *Bioinformatics (Oxford, England)* 24.15 (Aug. 2008), pp. 1729–1730.

Ficz, G., T. A. Hore, F. Santos, H. J. Lee, W. Dean, J. Arand, F. Krueger, D. Oxley, Y.-L. Paul, J. Walter, S. J. Cook, S. Andrews, M. R. Branco, and W. Reik (2013).

"FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency." In: *Cell stem cell* (July 2013).

Flicek, P. et al. (2011). "Ensembl 2011." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D800–6.

Flicek, P. et al. (2013). "Ensembl 2013." In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. D48–55.

French, P. J., J. Peeters, S. Horsman, E. Duijm, I. Siccama, M. J. van den Bent, T. M. Luider, J. M. Kros, P. van der Spek, and P. A. Sillevis Smitt (2007). "Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays." In: *Cancer research* 67.12 (June 2007), pp. 5635–5642.

Fujita, P. A. et al. (2011). "The UCSC Genome Browser database: update 2011." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D876–82.

Furey, T. S. (2012). "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." In: *Nature reviews. Genetics* 13.12 (Dec. 2012), pp. 840–852.

Gafni, O. et al. (2013). "Derivation of novel human ground state naive pluripotent stem cells." In: *Nature* 504.7479 (Dec. 2013), pp. 282–286.

Gardina, P. J., T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, and Y. Turpaz (2006). "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array." In: *BMC genomics* 7 (2006), p. 325.

Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry (2004). "affy–analysis of Affymetrix GeneChip data at the probe level." In: *Bioinformatics (Oxford, England)* 20.3 (Feb. 2004), pp. 307–315.

Gentleman, R. C. et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." In: *Genome biology* 5.10 (2004), R80.

Glynn, E. F., P. C. Megee, H.-G. Yu, C. Mistrot, E. Unal, D. E. Koshland, J. L. DeRisi, and J. L. Gerton (2004). "Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae." In: *PLoS biology* 2.9 (Sept. 2004), E259.

Goffeau, A, B. G. Barrell, H Bussey, R. W. Davis, B Dujon, H Feldmann, F Galibert, J. D. Hoheisel, C Jacq, M Johnston, E. J. Louis, H. W. Mewes, Y Murakami, P Philippsen, H Tettelin, and S. G. Oliver (1996). "Life with 6000 genes." In: *Science (New York, N.Y.)* 274.5287 (Oct. 1996), pp. 546–563–7.

Golan-Mashiach, M., J.-E. Dazard, S. Gerecht-Nir, N. Amariglio, T. Fisher, J. Jacob-Hirsch, B. Bielorai, S. Osenberg, O. Barad, G. Getz, A. Toren, G. Rechavi, J. Itskovitz-Eldor, E. Domany, and D. Givol (2005). "Design principle of gene expression used by human stem cells: implication for pluripotency." In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 19.1 (Jan. 2005), pp. 147–149.

Grabherr, M. G. et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." In: *Nature biotechnology* 29.7 (2011), pp. 644–652.

Grant, G. R., M. H. Farkas, A. Pizarro, N. Lahens, J. Schug, B. Brunk, C. J. Stoeckert, J. B. Hogenesch, and E. A. Pierce (2011). "Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)." In: *Bioinformatics (Oxford, England)* (July 2011).

Graveley, B. R. et al. (2011). "The developmental transcriptome of Drosophila melanogaster." In: *Nature* 471.7339 (Mar. 2011), pp. 473–479.

Griffiths, D. S., J. Li, M. A. Dawson, M. W. B. Trotter, Y.-H. Cheng, A. M. Smith, W. Mansfield, P. Liu, T. Kouzarides, J. Nichols, A. J. Bannister, A. R. Green, and B. Göttgens (2011). "LIF-independent JAK signalling to chromatin in embryonic stem cells uncovered from an adult stem cell disease." In: *Nature cell biology* 13.1 (Jan. 2011), pp. 13–21.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." In: *Cell* 130.1 (July 2007), pp. 77–88.

Guigo, R., P. Flicek, J. F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T. R. Gingeras, J. Harrow, T. Hubbard, S. E. Lewis, and M. G. Reese (2006). "EGASP: the human ENCODE Genome Annotation Assessment Project." In: *Genome biology* 7 Suppl 1 (2006), S2.1–31.

Guo, G., J. Yang, J. Nichols, J. S. Hall, I. Eyres, W. Mansfield, and A. Smith (2009). "Klf4 reverts developmentally programmed restriction of ground state pluripotency." In: *Development (Cambridge, England)* 136.7 (Apr. 2009), pp. 1063–1069.

Guo, G., Y. Huang, P. Humphreys, X. Wang, and A. Smith (2011). "A PiggyBac-based recessive screening method to identify pluripotency regulators." In: *PloS one* 6.4 (2011), e18189.

Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble (2007). "Quantifying similarity between motifs." In: *Genome biology* 8.2 (2007), R24.

Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." In: *Nature* (Aug. 2011).

Ha, N., M. Polychronidou, and I. Lohmann (2012). "COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets." In: *PloS one* 7.12 (2012), e52055.

Hall, J., G. Guo, J. Wray, I. Eyres, J. Nichols, L. Grotewold, S. Morfopoulou, P. Humphreys, W. Mansfield, R. Walker, S. Tomlinson, and A. Smith (2009). "Oct4 and LIF/Stat3 additively induce Krüppel factors to sustain embryonic stem cell self-renewal." In: *Cell stem cell* 5.6 (Dec. 2009), pp. 597–609.

Hanna, J., A. W. Cheng, K. Saha, J. Kim, C. J. Lengner, F. Soldner, J. P. Cassady, J. Muffat, B. W. Carey, and R. Jaenisch (2010). "Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.20 (May 2010), pp. 9222–9227.

Hansen, K. D., S. E. Brenner, and S. Dudoit (2010). "Biases in Illumina transcriptome sequencing caused by random hexamer priming." In: *Nucleic acids research* 38.12 (July 2010), e131.

Hao, J., T.-G. Li, X. Qi, D.-F. Zhao, and G.-Q. Zhao (2006). "WNT/beta-catenin pathway up-regulates Stat3 and converges on LIF to prevent differentiation of mouse embryonic stem cells." In: *Developmental biology* 290.1 (Feb. 2006), pp. 81–91.

Hardcastle, T. J. and K. A. Kelly (2010). "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." In: *BMC bioinformatics* 11 (2010), p. 422.

Harrow, J. et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." In: *Genome research* 22.9 (Sept. 2012), pp. 1760–1774.

Hata, K., M. Okano, H. Lei, and E. Li (2002). "Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice." In: *Development (Cambridge, England)* 129.8 (Apr. 2002), pp. 1983–1993.

Hatakeyama, S. (2012). "Ubiquitin-mediated regulation of JAK-STAT signaling in embryonic stem cells." In: *JAK-STAT* 1.3 (July 2012), pp. 168–175.

Hawkins, R. D. et al. (2010). "Distinct epigenomic landscapes of pluripotent and lineage-committed human cells." In: *Cell stem cell* 6.5 (May 2010), pp. 479–491.

Hayashi, K. and M. A. Surani (2009). "Self-renewing epiblast stem cells exhibit continual delineation of germ cells with epigenetic reprogramming in vitro." In: *Development (Cambridge, England)* 136.21 (Nov. 2009), pp. 3549–3556.

He, S., O. Wurtzel, K. Singh, J. L. Froula, S. Yilmaz, S. G. Tringe, Z. Wang, F. Chen, E. A. Lindquist, R. Sorek, and P. Hugenholtz (2010). "Validation of two ribosomal RNA removal methods for microbial metatranscriptomics." In: *Nature methods* 7.10 (Oct. 2010), pp. 807–812.

Hiller, D. and W. H. Wong (2013). "Simultaneous isoform discovery and quantification from RNA-seq." In: *Statistics in biosciences* 5.1 (May 2013), pp. 100–118.

Hirai, H., P. Karian, and N. Kikyo (2011). "Regulation of embryonic stem cell self-renewal and pluripotency by leukaemia inhibitory factor." In: *The Biochemical journal* 438.1 (Aug. 2011), pp. 11–23.

Hoen, P. A. C. t, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen (2008). "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms." In: *Nucleic acids research* 36.21 (Dec. 2008), e141.

Holliday, R and J. E. Pugh (1975). "DNA modification mechanisms and gene activity during development." In: *Science (New York, N.Y.)* 187.4173 (Jan. 1975), pp. 226–232.

Hoon, M. de and Y. Hayashizaki (2008). "Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference." In: *BioTechniques* 44.5 (Apr. 2008), pp. 627–8–630–632.

Hu, J., H. Ge, M. Newman, and K. Liu (2012). "OSA: a fast and accurate alignment tool for RNA-Seq." In: *Bioinformatics (Oxford, England)* 28.14 (July 2012), pp. 1933–1934.

Hu, Q., A. M. Friedrich, L. V. Johnson, and D. O. Clegg (2010). "Memory in induced pluripotent stem cells: reprogrammed human retinal-pigmented epithelial cells show tendency for spontaneous redifferentiation." In: *Stem cells (Dayton, Ohio)* 28.11 (Nov. 2010), pp. 1981–1991.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." In: *Nature protocols* 4.1 (2009), pp. 44–57.

Huang, Y., R. Osorno, A. Tsakiridis, and V. Wilson (2012). "In Vivo differentiation potential of epiblast stem cells revealed by chimeric embryo formation." In: *Cell reports* 2.6 (Dec. 2012), pp. 1571–1578.

Huber, W., J. Toedling, and L. M. Steinmetz (2006). "Transcript mapping with high-density oligonucleotide tiling arrays." In: *Bioinformatics (Oxford, England)* 22.16 (Aug. 2006), pp. 1963–1970.

Humphrey, R. K., G. M. Beattie, A. D. Lopez, N. Bucay, C. C. King, M. T. Firpo, S. Rose-John, and A. Hayek (2004). "Maintenance of pluripotency in human embryonic stem cells is STAT3 independent." In: *Stem cells (Dayton, Ohio)* 22.4 (2004), pp. 522–530.

Hutchison, C. A. (2007). "DNA sequencing: bench to bedside and beyond." In: *Nucleic acids research* 35.18 (2007), pp. 6227–6237.

International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." In: *Nature* 431.7011 (Oct. 2004), pp. 931–945.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." In: *Biostatistics (Oxford, England)* 4.2 (Apr. 2003), pp. 249–264.

Iwata, H. and O. Gotoh (2011). "Comparative analysis of information contents relevant to recognition of introns in many species." In: *BMC genomics* 12 (2011), p. 45.

Jean, G., A. Kahles, V. T. Sreedharan, F. De Bona, and G. Rätsch (2010). "RNA-Seq read alignments with PALMapper." In: *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 11 (Dec. 2010), Unit 11.6.

Jia, D., R. Z. Jurkowska, X. Zhang, A. Jeltsch, and X. Cheng (2007). "Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation." In: *Nature* 449.7159 (Sept. 2007), pp. 248–251.

Jiang, J., Y.-S. Chan, Y.-H. Loh, J. Cai, G.-Q. Tong, C.-A. Lim, P. Robson, S. Zhong, and H.-H. Ng (2008). "A core Klf circuitry regulates self-renewal of embryonic stem cells." In: *Nature cell biology* 10.3 (Mar. 2008), pp. 353–360.

Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions." In: *Science (New York, N.Y.)* 316.5830 (June 2007), pp. 1497–1502.

Jones, D. C., W. L. Ruzzo, X. Peng, and M. G. Katze (2012). "A new approach to bias correction in RNA-Seq." In: *Bioinformatics (Oxford, England)* 28.7 (Apr. 2012), pp. 921–928.

Karpinets, T. V., D. J. Greenwood, C. E. Sams, and J. T. Ammons (2006). "RNA:protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis." In: *BMC biology* 4 (2006), p. 30.

Karwacki-Neisius, V., J. Göke, R. Osorno, F. Halbritter, J. H. Ng, A. Y. Weiße, F. C. K. Wong, A. Gagliardi, N. P. Mullin, N. Festuccia, D. Colby, S. R. Tomlinson, H.-H. Ng, and I. Chambers (2013). "Reduced oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by oct4 and nanog." In: *Cell stem cell* 12.5 (May 2013), pp. 531–545.

Katayama, S et al. (2005). "Antisense transcription in the mammalian transcriptome." In: *Science (New York, N.Y.)* 309.5740 (Sept. 2005), pp. 1564–1566.

Katz, Y., E. T. Wang, E. M. Airoldi, and C. B. Burge (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation." In: *Nature methods* 7.12 (Dec. 2010), pp. 1009–1015.

Kauffmann, A., T. F. Rayner, H. Parkinson, M. Kapushesky, M. Lukk, A. Brazma, and W. Huber (2009). "Importing ArrayExpress datasets into R/Bioconductor." In: *Bioinformatics (Oxford, England)* 25.16 (Aug. 2009), pp. 2092–2094.

Keane, T. M. et al. (2011). "Mouse genomic variation and its effect on phenotypes and gene regulation." In: *Nature* 477.7364 (Sept. 2011), pp. 289–294.

Kent, W. J. (2002). "BLAT–the BLAST-like alignment tool." In: *Genome research* 12.4 (Apr. 2002), pp. 656–664.

Kidder, B. L., J. Yang, and S. Palmer (2008). "Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells." In: *PloS one* 3.12 (2008), e3932.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." In: *Genome biology* 14.4 (Apr. 2013), R36.

Kim, D. H., I. H. Cho, H. S. Kim, J. E. Jung, J. E. Kim, K. H. Lee, T. Park, Y. M. Yang, S.-Y. Seong, S.-K. Ye, and M.-H. Chung (2006). "Anti-inflammatory effects of 8-

hydroxydeoxyguanosine in LPS-induced microglia activation: suppression of STAT3-mediated intercellular adhesion molecule-1 expression." In: *Experimental & molecular medicine* 38.4 (Aug. 2006), pp. 417–427.

Kim, J., J. Chu, X. Shen, J. Wang, and S. H. Orkin (2008). "An extended transcriptional network for pluripotency of embryonic stem cells." In: *Cell* 132.6 (Mar. 2008), pp. 1049–1061.

Kim, K et al. (2010). "Epigenetic memory in induced pluripotent stem cells". In: *Nature* 467.7313 (July 2010), pp. 285–290.

Kopp, J. L., B. D. Ormsbee, M. Desler, and A. Rizzino (2008). "Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells." In: *Stem cells (Dayton, Ohio)* 26.4 (Apr. 2008), pp. 903–911.

Korbel, J. O., A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein (2009). "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data." In: *Genome biology* 10.2 (2009), R23.

Kouzarides, T. (2007). "Chromatin modifications and their function." In: *Cell* 128.4 (Feb. 2007), pp. 693–705.

Krejcí, J., R. Uhlírová, G. Galiová, S. Kozubek, J. Smigová, and E. Bártová (2009). "Genome-wide reduction in H3K9 acetylation during human embryonic stem cell differentiation." In: *Journal of cellular physiology* 219.3 (June 2009), pp. 677–687.

Kuhn, R. M., D. Haussler, and W. J. Kent (2013). "The UCSC genome browser and associated tools." In: *Briefings in bioinformatics* 14.2 (Mar. 2013), pp. 144–161.

Kulkarni, M. M. (2011). "Digital multiplexed gene expression analysis using the NanoString nCounter system." In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* Chapter 25 (Apr. 2011), Unit25B.10.

Kwon, H., D. Thierry-Mieg, J. Thierry-Mieg, H.-P. Kim, J. Oh, C. Tunyaplin, S. Carotta, C. E. Donovan, M. L. Goldman, P. Tailor, K. Ozato, D. E. Levy, S. L. Nutt, K. Calame, and W. J. Leonard (2009). "Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors." In: *Immunity* 31.6 (Dec. 2009), pp. 941–952.

Lander, E. S. et al. (2001). "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822 (Feb. 2001), pp. 860–921.

Langlais, D., C. Couture, A. Balsalobre, and J. Drouin (2012). "The Stat3/GR interaction code: predictive value of direct/indirect DNA recruitment for transcription outcome." In: *Molecular cell* 47.1 (July 2012), pp. 38–49.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." In: *Nature methods* 9.4 (Apr. 2012), pp. 357–359.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome biology* 10.3 (2009), R25.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring, and C.-L. Wei (2010). "Dynamic changes in the human methylome during differentiation." In: *Genome research* 20.3 (Mar. 2010), pp. 320–331.

Lee, H. and M. C. Schatz (2012). "Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score." In: *Bioinformatics (Oxford, England)* 28.16 (Aug. 2012), pp. 2097–2105.

Lee, S., C. H. Seo, B. Lim, J. O. Yang, J. Oh, M. Kim, S. Lee, B. Lee, C. Kang, and S. Lee (2011). "Accurate quantification of transcriptome from RNA-Seq data by effective length normalization." In: *Nucleic acids research* 39.2 (Jan. 2011), e9.

Legault, L. H. and C. N. Dewey (2013). "Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs." In: *Bioinformatics (Oxford, England)* 29.18 (Sept. 2013), pp. 2300–2310.

Leitch, H. G., K. R. McEwen, A. Turp, V. Encheva, T. Carroll, N. Grabole, W. Mansfield, B. Nashun, J. G. Knezovich, A. Smith, M. A. Surani, and P. Hajkova (2013). "Naive pluripotency is associated with global DNA hypomethylation." In: *Nature structural & molecular biology* 20.3 (Mar. 2013), pp. 311–316.

Lenhard, B., A. Sandelin, and P. Carninci (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." In: *Nature reviews. Genetics* 13.4 (Apr. 2012), pp. 233–245.

Levin, J. Z., M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev (2010). "Comprehensive comparative analysis of strand-specific RNA sequencing methods." In: *Nature methods* 7.9 (Sept. 2010), pp. 709–715.

Levy, D. E. and C.-k. Lee (2002). "What does Stat3 do?" In: *The Journal of clinical investigation* 109.9 (May 2002), pp. 1143–1148.

Li, B. and C. N. Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." In: *BMC bioinformatics* 12 (2011), p. 323.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey (2010). "RNA-Seq gene expression estimation with read mapping uncertainty." In: *Bioinformatics (Oxford, England)* 26.4 (Feb. 2010), pp. 493–500.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." In: *Bioinformatics (Oxford, England)* 25.14 (July 2009), pp. 1754–1760.

Li, H., J. Ruan, and R. Durbin (2008a). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." In: *Genome research* 18.11 (Nov. 2008), pp. 1851–1858.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup (2009). "The

Sequence Alignment/Map format and SAMtools." In: *Bioinformatics (Oxford, England)* 25.16 (Aug. 2009), pp. 2078–2079.

Li, J. J., C.-R. Jiang, J. B. Brown, H. Huang, and P. J. Bickel (2011). "Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.50 (Dec. 2011), pp. 19867–19872.

Li, P., C. Tong, R. Mehrian-Shai, L. Jia, N. Wu, Y. Yan, R. E. Maxson, E. N. Schulze, H. Song, C.-L. Hsieh, M. F. Pera, and Q.-L. Ying (2008b). "Germline competent embryonic stem cells derived from rat blastocysts." In: *Cell* 135.7 (Dec. 2008), pp. 1299–1310.

Li, Y., J. McClintick, L. Zhong, H. J. Edenberg, M. C. Yoder, and R. J. Chan (2005). "Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4." In: *Blood* 105.2 (Jan. 2005), pp. 635–637.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science (New York, N.Y.)* 326.5950 (Oct. 2009), pp. 289–293.

Liu, X. S., D. L. Brutlag, and J. S. Liu (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." In: *Nature biotechnology* 20.8 (Aug. 2002), pp. 835–839.

Livyatan, I., A. Harikumar, M. Nissim-Rafinia, R. Duttagupta, T. R. Gingeras, and E. Meshorer (2013). "Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation." In: *Nucleic acids research* (Apr. 2013).

Lockhart, D. J., H Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M Mittmann, C Wang, M Kobayashi, H Horton, and E. L. Brown (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." In: *Nature biotechnology* 14.13 (Dec. 1996), pp. 1675–1680.

Loewer, S., M. N. Cabili, M. Guttman, Y.-H. Loh, K. Thomas, I.-H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P. D. Manos, S. Datta, E. S. Lander, T. M. Schlaeger, G. Q. Daley, and J. L. Rinn (2010). "Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells." In: *Nature genetics* 42.12 (Dec. 2010), pp. 1113–1117.

Loh, Y.-H. et al. (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." In: *Nature genetics* 38.4 (Apr. 2006), pp. 431–440.

Loh, Y.-H., W. Zhang, X. Chen, J. George, and H.-H. Ng (2007). "Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells." In: *Genes & development* 21.20 (Oct. 2007), pp. 2545–2557.

Machanick, P. and T. L. Bailey (2011). "MEME-ChIP: motif analysis of large DNA datasets." In: *Bioinformatics (Oxford, England)* 27.12 (June 2011), pp. 1696–1697.

Mallanna, S. K. and A. Rizzino (2010). "Emerging roles of microRNAs in the control of embryonic stem cells and the generation of induced pluripotent stem cells." In: *Developmental biology* 344.1 (Aug. 2010), pp. 16–25.

Malone, J. H. and B. Oliver (2011). "Microarrays, deep sequencing and the true measure of the transcriptome." In: *BMC biology* 9 (2011), p. 34.

Marco-Sola, S., M. Sammeth, R. Guigo, and P. Ribeca (2012). "The GEM mapper: fast, accurate and versatile alignment by filtration." In: *Nature methods* 9.12 (Dec. 2012), pp. 1185–1188.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." In: *Annual review of genomics and human genetics* 9 (2008), pp. 387–402.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." In: *Genome research* 18.9 (Sept. 2008), pp. 1509–1517.

Marks, H., T. Kalkan, R. Menafra, S. Denissov, K. Jones, H. Hofemeister, J. Nichols, A. Kranz, A Francis Stewart, A. Smith, and H. G. Stunnenberg (2012). "The transcriptional and epigenomic foundations of ground state pluripotency." In: *Cell* 149.3 (Apr. 2012), pp. 590–604.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young (2008). "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." In: *Cell* 134.3 (Aug. 2008), pp. 521–533.

Martello, G., T. Sugimoto, E. Diamanti, A. Joshi, R. Hannah, S. Ohtsuka, B. Göttgens, H. Niwa, and A. Smith (2012). "Esrrb is a pivotal target of the gsk3/tcf3 axis regulating embryonic stem cell self-renewal." In: *Cell stem cell* 11.4 (Oct. 2012), pp. 491–504.

Martello, G., P. Bertone, and A. Smith (2013). "Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor." In: *The EMBO journal* (Aug. 2013).

Martin, G. R. (1981). "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells." In: *Proceedings of the National Academy of Sciences of the United States of America* 78.12 (Dec. 1981), pp. 7634–7638.

Mason, M. J., K. Plath, and Q. Zhou (2010). "Identification of context-dependent motifs by contrasting ChIP binding data." In: *Bioinformatics (Oxford, England)* 26.22 (Nov. 2010), pp. 2826–2832.

Massingham, T. and N. Goldman (2012). "All Your Base: a fast and accurate probabilistic approach to base calling." In: *Genome biology* 13.2 (2012), R13.

Masui, S., Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. H. Ko, and H. Niwa (2007). "Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells." In: *Nature cell biology* 9.6 (June 2007), pp. 625–635.

Matsuda, T, T Nakamura, K Nakao, T Arai, M Katsuki, T Heike, and T Yokota (1999). "STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells." In: *The EMBO journal* 18.15 (Aug. 1999), pp. 4261–4269.

Matsui, Y, K Zsebo, and B. L. Hogan (1992). "Derivation of pluripotential embryonic stem cells from murine primordial germ cells in culture." In: *Cell* 70.5 (Sept. 1992), pp. 841–847.

Maxam, A. M. and W Gilbert (1977). "A new method for sequencing DNA." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.2 (Feb. 1977), pp. 560–564.

Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." In: *Nature* 454.7205 (Aug. 2008), pp. 766–770.

Meshorer, E., D. Yellajoshula, E. George, P. J. Scambler, D. T. Brown, and T. Misteli (2006). "Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells." In: *Developmental cell* 10.1 (Jan. 2006), pp. 105–116.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." In: *Nature reviews. Genetics* 11.1 (Jan. 2010), pp. 31–46.

Mezlini, A. M., E. J. M. Smith, M. Fiume, O. Buske, G. L. Savich, S. Shah, S. Aparicio, D. Y. Chiang, A. Goldenberg, and M. Brudno (2013). "iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data." In: *Genome research* 23.3 (Mar. 2013), pp. 519–529.

Mikkelsen, T. S. et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." In: *Nature* 448.7153 (Aug. 2007), pp. 553–560.

Mills, R. E. et al. (2011). "Mapping copy number variation by population-scale genome sequencing." In: *Nature* 470.7332 (Feb. 2011), pp. 59–65.

Mitsui, K., Y. Tokuzawa, H. Itoh, K. Segawa, M. Murakami, K. Takahashi, M. Maruyama, M. Maeda, and S. Yamanaka (2003). "The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells." In: *Cell* 113.5 (May 2003), pp. 631–642.

Modarresi, F., M. A. Faghihi, M. A. Lopez-Toledano, R. P. Fatemi, M. Magistri, S. P. Brothers, M. P. van der Brug, and C. Wahlestedt (2012). "Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation." In: *Nature biotechnology* (Mar. 2012).

modENCODE Consortium et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." In: *Science (New York, N.Y.)* 330.6012 (Dec. 2010), pp. 1787–1797.

Molenaar, M, M van de Wetering, M Oosterwegel, J Peterson-Maduro, S Godsave, V Korinek, J Roose, O Destrée, and H Clevers (1996). "XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos." In: *Cell* 86.3 (Aug. 1996), pp. 391–399.

Morey, M., A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, and J. A. Cocho (2013). "A glimpse into past, present, and future DNA sequencing." In: *Molecular genetics and metabolism* 110.1-2 (Sept. 2013), pp. 3–24.

Morin, R., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." In: *BioTechniques* 45.1 (July 2008), pp. 81–94.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nature methods* 5.7 (July 2008), pp. 621–628.

Mortazavi, A., E. M. Schwarz, B. Williams, L. Schaeffer, I. Antoshechkin, B. J. Wold, and P. W. Sternberg (2010). "Scaffolding a Caenorhabditis nematode genome with RNA-seq." In: *Genome research* 20.12 (Dec. 2010), pp. 1740–1747.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." In: *Science (New York, N.Y.)* 320.5881 (June 2008), pp. 1344–1349.

Nakanoh, S., K. Okazaki, and K. Agata (2013). "Inhibition of MEK and GSK3 Supports ES Cell-like Domed Colony Formation from Avian and Reptile Embryos." In: *Zoological science* 30.7 (July 2013), pp. 543–552.

Narlikar, L. (2013). "MuMoD: a Bayesian approach to detect multiple modes of protein-DNA binding from genome-wide ChIP data." In: *Nucleic acids research* 41.1 (Jan. 2013), pp. 21–32.

Ni, Z. and R. Bremner (2007). "Brahma-related gene 1-dependent STAT3 recruitment at IL-6-inducible genes." In: *Journal of immunology (Baltimore, Md. : 1950)* 178.1 (Jan. 2007), pp. 345–351.

Nichols, J, E. P. Evans, and A. G. Smith (1990). "Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity." In: *Development (Cambridge, England)* 110.4 (Dec. 1990), pp. 1341–1348.

Nichols, J, B Zevnik, K Anastassiadis, H Niwa, D Klewe-Nebenius, I Chambers, H Schöler, and A Smith (1998). "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4." In: *Cell* 95.3 (Oct. 1998), pp. 379–391.

Nichols, J. and A. Smith (2009). "Naive and primed pluripotent states." In: *Cell stem cell* 4.6 (June 2009), pp. 487–492.

Nichols, J. and A. Smith (2012). "Pluripotency in the embryo and in culture." In: *Cold Spring Harbor perspectives in biology* 4.8 (Aug. 2012), a008128.

Nichols, J., K. Jones, J. M. Phillips, S. A. Newland, M. Roode, W. Mansfield, A. Smith, and A. Cooke (2009). "Validated germline-competent embryonic stem cell lines from nonobese diabetic mice." In: *Nature medicine* 15.7 (July 2009), pp. 814–818.

Nilsen, T. W. and B. R. Graveley (2010). "Expansion of the eukaryotic proteome by alternative splicing." In: *Nature* 463.7280 (Jan. 2010), pp. 457–463.

Ning, Z, A. J. Cox, and J. C. Mullikin (2001). "SSAHA: a fast search method for large DNA databases." In: *Genome research* 11.10 (Oct. 2001), pp. 1725–1729.

Nishiyama, A. et al. (2013). "Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells." In: *Scientific reports* 3 (Mar. 2013), p. 1390.

Niwa, H, T Burdon, I Chambers, and A Smith (1998). "Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3." In: *Genes & development* 12.13 (July 1998), pp. 2048–2060.

Niwa, H, J Miyazaki, and A. G. Smith (2000). "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." In: *Nature genetics* 24.4 (Apr. 2000), pp. 372–376.

Niwa, H., K. Ogawa, D. Shimosato, and K. Adachi (2009). "A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells." In: *Nature* 460.7251 (July 2009), pp. 118–122.

Nookaew, I., M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen (2012). "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae." In: *Nucleic acids research* (Sept. 2012).

Ohbayashi, N., O. Ikeda, N. Taira, Y. Yamamoto, R. Muromoto, Y. Sekine, K. Sugiyama, T. Honjoh, and T. Matsuda (2007). "LIF- and IL-6-induced acetylation of STAT3 at Lys-685 through PI3K/Akt activation." In: *Biological & pharmaceutical bulletin* 30.10 (Oct. 2007), pp. 1860–1864.

O'Neil, D., H. Glowatz, and M. Schlumpberger (2013). "Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity." In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* Chapter 4 (July 2013), Unit4.19.

Oosten, A. L. van, Y. Costa, A. Smith, and J. C. R. Silva (2012). "JAK/STAT3 signalling is sufficient and dominant over antagonistic cues for the establishment of naive pluripotency." In: *Nature communications* 3 (2012), p. 817.

Ouyang, Z., Q. Zhou, and W. H. Wong (2009). "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.51 (Dec. 2009), pp. 21521–21526.

Ozsolak, F. and P. M. Milos (2011). "RNA sequencing: advances, challenges and opportunities." In: *Nature reviews. Genetics* 12.2 (Feb. 2011), pp. 87–98.

Paling, N. R. D., H. Wheadon, H. K. Bone, and M. J. Welham (2004). "Regulation of embryonic stem cell self-renewal by phosphoinositide 3-kinase-dependent signaling." In: *The Journal of biological chemistry* 279.46 (Nov. 2004), pp. 48063–48070.

Pardo, M., B. Lang, L. Yu, H. Prosser, A. Bradley, M. M. Babu, and J. Choudhary (2010). "An expanded Oct4 interaction network: implications for stem cell biology, development, and disease." In: *Cell stem cell* 6.4 (Apr. 2010), pp. 382–395.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." In: *Nature reviews. Genetics* 10.10 (Oct. 2009), pp. 669–680.

Pavesi, G., P. Mereghetti, G. Mauri, and G. Pesole (2004). "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." In: *Nucleic acids research* 32.Web Server issue (July 2004), W199–203.

Pepke, S., B. Wold, and A. Mortazavi (2009). "Computation for ChIP-seq and RNA-seq studies." In: *Nature methods* 6.11 Suppl (Nov. 2009), S22–32.

Pereira, L., F. Yi, and B. J. Merrill (2006). "Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal." In: *Molecular and cellular biology* 26.20 (Oct. 2006), pp. 7479–7491.

Radzisheuskaya, A., G. Le Bin Chia, R. L. Dos Santos, T. W. Theunissen, L. F. C. Castro, J. Nichols, and J. C. R. Silva (2013). "A defined Oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages." In: *Nature cell biology* (Apr. 2013).

Raghavachari, N., J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, C. J. O'Donnell, P. J. Munson, and G. J. Kato (2012). "A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease." In: *BMC medical genomics* 5 (2012), p. 28.

Ramalho-Santos, J. (2011). "Human procreation in unchartered territory: new twists in ethical discussions." In: *Human reproduction (Oxford, England)* 26.6 (June 2011), pp. 1284–1287.

Reese, M. G., G Hartzell, N. L. Harris, U Ohler, J. F. Abril, and S. E. Lewis (2000). "Genome annotation assessment in Drosophila melanogaster." In: *Genome research* 10.4 (Apr. 2000), pp. 483–501.

Ren, B, F Robert, J. J. Wyrick, O Aparicio, E. G. Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young (2000). "Genome-wide location and function of DNA binding proteins." In: *Science (New York, N.Y.)* 290.5500 (Dec. 2000), pp. 2306–2309.

Resnick, J. L., L. S. Bixler, L Cheng, and P. J. Donovan (1992). "Long-term proliferation of mouse primordial germ cells in culture." In: *Nature* 359.6395 (Oct. 1992), pp. 550–551.

Reyes, A., S. Anders, R. J. Weatheritt, T. J. Gibson, L. M. Steinmetz, and W. Huber (2013). "Drift and conservation of differential exon usage across tissues in primate species." In: *Proceedings of the National Academy of Sciences of the United States of America* (Sept. 2013).

Roberts, A., H. Pimentel, C. Trapnell, and L. Pachter (2011a). "Identification of novel transcripts in annotated genomes using RNA-Seq." In: *Bioinformatics (Oxford, England)* 27.17 (Sept. 2011), pp. 2325–2329.

Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter (2011b). "Improving RNA-Seq expression estimates by correcting for fragment bias." In: *Genome biology* 12.3 (Mar. 2011), R22.

Robertson, G. et al. (2010). "De novo assembly and analysis of RNA-seq data." In: *Nature methods* 7.11 (Nov. 2010), pp. 909–912.

Robinson, M. D. and G. K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance." In: *Bioinformatics (Oxford, England)* 23.21 (Nov. 2007), pp. 2881–2887.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 2010), pp. 139–140.

Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." In: *Nature biotechnology* 27.1 (Jan. 2009), pp. 66–75.

Rugg-Gunn, P. J., B. J. Cox, A. Ralston, and J. Rossant (2010). "Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.24 (June 2010), pp. 10783–10790.

Salmon-Divon, M., H. Dvinge, K. Tammoja, and P. Bertone (2010). "PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci." In: *BMC bioinformatics* 11 (2010), p. 415.

Sanger, F and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." In: *Journal of molecular biology* 94.3 (May 1975), pp. 441–448.

Sanger, F, S Nicklen, and A. R. Coulson (1977a). "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (Dec. 1977), pp. 5463–5467.

Sanger, F, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M Smith (1977b). "Nucleotide sequence of bacteriophage phi X174 DNA." In: *Nature* 265.5596 (Feb. 1977), pp. 687–695.

Sato, N., L. Meijer, L. Skaltsounis, P. Greengard, and A. H. Brivanlou (2004). "Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor." In: *Nature medicine* 10.1 (Jan. 2004), pp. 55–63.

Schena, M, D Shalon, R. W. Davis, and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." In: *Science (New York, N.Y.)* 270.5235 (Oct. 1995), pp. 467–470.

Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney (2012). "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels." In: *Bioinformatics (Oxford, England)* 28.8 (Apr. 2012), pp. 1086–1092.

Schweikert, G., A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, and G. Rätsch (2009). "mGene: accurate SVM-based gene finding with an application to nematode genomes." In: *Genome research* 19.11 (Nov. 2009), pp. 2133–2143.

Sekkaï, D., G. Gruel, M. Herry, V. Moucadel, S. N. Constantinescu, O. Albagli, D. Tronik-Le Roux, W. Vainchenker, and A. Bennaceur-Griscelli (2005). "Microarray analysis of LIF/Stat3 transcriptional targets in embryonic stem cells." In: *Stem cells (Dayton, Ohio)* 23.10 (Oct. 2005), pp. 1634–1642.

Senner, C. E., F. Krueger, D. Oxley, S. Andrews, and M. Hemberger (2012). "DNA Methylation Profiles Define Stem Cell Identity and Reveal a Tight Embryonic-Extraembryonic Lineage Boundary." In: *Stem cells (Dayton, Ohio)* (Oct. 2012).

Serrano, L., B. N. Vazquez, and J. Tischfield (2013). "Chromatin structure, pluripotency and differentiation." In: *Experimental biology and medicine (Maywood, N.J.)* 238.3 (Mar. 2013), pp. 259–270.

Shalon, D, S. J. Smith, and P. O. Brown (1996). "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization." In: *Genome research* 6.7 (July 1996), pp. 639–645.

Shin, H., T. Liu, A. K. Manrai, and X. S. Liu (2009). "CEAS: cis-regulatory element annotation system." In: *Bioinformatics (Oxford, England)* 25.19 (Oct. 2009), pp. 2605–2606.

Shiota, K., Y. Kogo, J. Ohgane, T. Imamura, A. Urano, K. Nishino, S. Tanaka, and N. Hattori (2002). "Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice." In: *Genes to cells : devoted to molecular & cellular mechanisms* 7.9 (Sept. 2002), pp. 961–969.

Silva, J., O. Barrandon, J. Nichols, J. Kawaguchi, T. W. Theunissen, and A. Smith (2008). "Promotion of reprogramming to ground state pluripotency by signal inhibition." In: *PLoS biology* 6.10 (Oct. 2008), e253.

Slater, G. S. C. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." In: *BMC bioinformatics* 6 (2005), p. 31.

Smith, A. G. (2001). "Embryo-derived stem cells: of mice and men." In: *Annual review of cell and developmental biology* 17 (2001), pp. 435–462.

Smith, A. G. and M. L. Hooper (1987). "Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells." In: *Developmental biology* 121.1 (May 1987), pp. 1–9.

Smith, A. G., J. K. Heath, D. D. Donaldson, G. G. Wong, J Moreau, M Stahl, and D Rogers (1988). "Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides." In: *Nature* 336.6200 (Dec. 1988), pp. 688–690.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P Hughes, C Dodd, C. R. Connell, C Heiner, S. B. Kent, and L. E. Hood (1986). "Fluorescence detection in automated DNA sequence analysis." In: *Nature* 321.6071 (June 1986), pp. 674–679.

Smith, L. M., L. Hartmann, P. Drewe, R. Bohnert, A. Kahles, C. Lanz, and G. Rätsch (2012). "Multiple insert size paired-end sequencing for deconvolution of complex transcriptomes." In: *RNA biology* 9.5 (May 2012), pp. 596–609.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." In: *Statistical applications in genetics and molecular biology* 3 (2004), Article3.

Song, L. and L. Florea (2013). "CLASS: constrained transcript assembly of RNA-seq reads." In: *BMC bioinformatics* 14 Suppl 5 (2013), S14.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." In: *Journal of molecular biology* 98.3 (Nov. 1975), pp. 503–517.

Sperisen, P., C. Iseli, M. Pagni, B. J. Stevenson, P. Bucher, and C. V. Jongeneel (2004). "trome, trEST and trGEN: databases of predicted protein sequences." In: *Nucleic acids research* 32.Database issue (Jan. 2004), pp. D509–11.

Stahl, N, T. G. Boulton, T Farruggella, N. Y. Ip, S Davis, B. A. Witthuhn, F. W. Quelle, O Silvennoinen, G Barbieri, and S Pellegrini (1994). "Association and activation of Jak-Tyk kinases by CNTF-LIF-OSM-IL-6 beta receptor components." In: *Science (New York, N.Y.)* 263.5143 (Jan. 1994), pp. 92–95.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern (2006). "AUGUSTUS: ab initio prediction of alternative transcripts." In: *Nucleic acids research* 34.Web Server issue (July 2006), W435–9.

Storm, M. P., B. Kumpfmueller, B. Thompson, R. Kolde, J. Vilo, O. Hummel, H. Schulz, and M. J. Welham (2009). "Characterization of the phosphoinositide 3-kinase-dependent transcriptome in murine embryonic stem cells: identification of novel reg-

ulators of pluripotency." In: *Stem cells (Dayton, Ohio)* 27.4 (Apr. 2009), pp. 764–775.

Storm, M. P., H. K. Bone, C. G. Beck, P.-Y. Bourillot, V. Schreiber, T. Damiano, A. Nelson, P. Savatier, and M. J. Welham (2007). "Regulation of Nanog expression by phosphoinositide 3-kinase-dependent signaling in murine embryonic stem cells." In: *The Journal of biological chemistry* 282.9 (Mar. 2007), pp. 6265–6273.

Subrammaniyan, R., J. Amalorpavanathan, R. Shankar, M. Rajkumar, S. Baskar, S. R. Manjunath, R. Senthilkumar, P. Murugan, V. R. Srinivasan, and S. Abraham (2011). "Application of autologous bone marrow mononuclear cells in six patients with advanced chronic critical limb ischemia as a result of diabetes: our experience." In: *Cytotherapy* 13.8 (Sept. 2011), pp. 993–999.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo (2008). "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome." In: *Science (New York, N.Y.)* 321.5891 (Aug. 2008), pp. 956–960.

Sumi, T., Y. Fujimoto, N. Nakatsuji, and H. Suemori (2004). "STAT3 is dispensable for maintenance of self-renewal in nonhuman primate embryonic stem cells." In: *Stem cells (Dayton, Ohio)* 22.5 (2004), pp. 861–872.

Tai, C.-I. and Q.-L. Ying (2013). "Gbx2, a LIF/Stat3 target, promotes reprogramming to and retention of the pluripotent ground state." In: *Journal of cell science* 126.Pt 5 (Mar. 2013), pp. 1093–1098.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." In: *Cell* 126.4 (Aug. 2006), pp. 663–676.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka (2007). "Induction of pluripotent stem cells from adult human fibroblasts by defined factors." In: *Cell* 131.5 (Nov. 2007), pp. 861–872.

Tang, Y., Y. Luo, Z. Jiang, Y. Ma, C.-J. Lin, C. Kim, M. G. Carter, T. Amano, J. Park, S. Kish, and X. C. Tian (2012). "Jak/Stat3 Signaling Promotes Somatic Cell Reprogramming by Epigenetic Regulation." In: *Stem cells (Dayton, Ohio)* (Sept. 2012).

Tay, Y., J. Zhang, A. M. Thomson, B. Lim, and I. Rigoutsos (2008). "MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation." In: *Nature* 455.7216 (Oct. 2008), pp. 1124–1128.

Terai, S., T. Ishikawa, K. Omori, K. Aoyama, Y. Marumoto, Y. Urata, Y. Yokoyama, K. Uchida, T. Yamasaki, Y. Fujii, K. Okita, and I. Sakaida (2006). "Improved liver function in patients with liver cirrhosis after autologous bone marrow cell infusion therapy." In: *Stem cells (Dayton, Ohio)* 24.10 (Oct. 2006), pp. 2292–2298.

Tesar, P. J., J. G. Chenoweth, F. A. Brook, T. J. Davies, E. P. Evans, D. L. Mack, R. L. Gardner, and R. D. G. McKay (2007). "New cell lines from mouse epiblast share defining features with human embryonic stem cells." In: *Nature* 448.7150 (July 2007), pp. 196–199.

Theunissen, T. W., A. L. van Oosten, G. Castelo-Branco, J. Hall, A. Smith, and J. C. R. Silva (2011). "Nanog overcomes reprogramming barriers and induces pluripotency in minimal conditions." In: *Current biology : CB* 21.1 (Jan. 2011), pp. 65–71.

Thomson, J. A., J Kalishman, T. G. Golos, M Durning, C. P. Harris, R. A. Becker, and J. P. Hearn (1995). "Isolation of a primate embryonic stem cell line." In: *Proceedings of the National Academy of Sciences of the United States of America* 92.17 (Aug. 1995), pp. 7844–7848.

Thomson, J. A., J Kalishman, T. G. Golos, M Durning, C. P. Harris, and J. P. Hearn (1996). "Pluripotent cell lines derived from common marmoset (Callithrix jacchus) blastocysts." In: *Biology of reproduction* 55.2 (Aug. 1996), pp. 254–259.

Thomson, J. A., J Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones (1998). "Embryonic stem cell lines derived from human blastocysts." In: *Science (New York, N.Y.)* 282.5391 (Nov. 1998), pp. 1145–1147.

Tian, B., J. Hu, H. Zhang, and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." In: *Nucleic acids research* 33.1 (2005), pp. 201–212.

Trapnell, C., L. Pachter, and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." In: *Bioinformatics (Oxford, England)* 25.9 (May 2009), pp. 1105–1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." In: *Nature biotechnology* 28.5 (May 2010), pp. 511–515.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." In: *Nature protocols* 7.3 (2012), pp. 562–578.

Ura, H., M. Usuda, K. Kinoshita, C. Sun, K. Mori, T. Akagi, T. Matsuda, H. Koide, and T. Yokota (2008). "STAT3 and Oct-3/4 control histone modification through induction of Eed in embryonic stem cells." In: *The Journal of biological chemistry* 283.15 (Apr. 2008), pp. 9713–9723.

Urnov, F. D. and A. P. Wolffe (2001). "Above and within the genome: epigenetics past and present." In: *Journal of mammary gland biology and neoplasia* 6.2 (Apr. 2001), pp. 153–167.

Vallania, F., D. Schiavone, S. Dewilde, E. Pupo, S. Garbay, R. Calogero, M. Pontoglio, P. Provero, and V. Poli (2009). "Genome-wide discovery of functional transcription

factor binding sites by comparative genomics: the case of Stat3." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.13 (Mar. 2009), pp. 5117–5122.

Venter, J. C. et al. (2001). "The sequence of the human genome." In: *Science (New York, N.Y.)* 291.5507 (Feb. 2001), pp. 1304–1351.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge (2008). "Alternative isoform regulation in human tissue transcriptomes." In: *Nature* 456.7221 (Nov. 2008), pp. 470–476.

Wang, K., D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu (2010a). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." In: *Nucleic acids research* 38.18 (Oct. 2010), e178.

Wang, L., Z. Feng, X. Wang, X. Wang, and X. Zhang (2010b). "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data." In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 2010), pp. 136–138.

Wang, W., J. Yang, H. Liu, D. Lu, X. Chen, Z. Zenonos, L. S. Campos, R. Rad, G. Guo, S. Zhang, A. Bradley, and P. Liu (2011). "Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.45 (Nov. 2011), pp. 18283–18288.

Wang, W., Z. Qin, Z. Feng, X. Wang, and X. Zhang (2013). "Identifying differentially spliced genes from two groups of RNA-seq samples." In: *Gene* 518.1 (Apr. 2013), pp. 164–170.

Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews. Genetics* 10.1 (Jan. 2009), pp. 57–63.

Watanabe, S, H Umehara, K Murayama, M Okabe, T Kimura, and T Nakano (2006). "Activation of Akt signaling is sufficient to maintain pluripotency in mouse and primate embryonic stem cells." In: *Oncogene* 25.19 (May 2006), pp. 2697–2707.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." In: *Nature* 171.4356 (Apr. 1953), pp. 737–738.

Wegmüller, D., I. Raineri, B. Gross, E. J. Oakeley, and C. Moroni (2007). "A cassette system to study embryonic stem cell differentiation by inducible RNA interference." In: *Stem cells (Dayton, Ohio)* 25.5 (May 2007), pp. 1178–1185.

Welham, M. J., M. P. Storm, E Kingham, and H. K. Bone (2007). "Phosphoinositide 3-kinases and regulation of embryonic stem cell fate." In: *Biochemical Society transactions* 35.Pt 2 (Apr. 2007), pp. 225–228.

Wen, B., H. Wu, Y. Shinkai, R. A. Irizarry, and A. P. Feinberg (2009). "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." In: *Nature genetics* 41.2 (Feb. 2009), pp. 246–250.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young (2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes." In: *Cell* 153.2 (Apr. 2013), pp. 307–319.

Wilkening, S., V. Pelechano, A. I. Järvelin, M. M. Tekkedil, S. Anders, V. Benes, and L. M. Steinmetz (2013). "An efficient method for genome-wide polyadenylation site mapping and RNA quantification." In: *Nucleic acids research* 41.5 (Mar. 2013), e65.

Williams, R. L., D. J. Hilton, S Pease, T. A. Willson, C. L. Stewart, D. P. Gearing, E. F. Wagner, D Metcalf, N. A. Nicola, and N. M. Gough (1988). "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells." In: *Nature* 336.6200 (Dec. 1988), pp. 684–687.

Wolffe, A. P. and M. A. Matzke (1999). "Epigenetics: regulation through repression." In: *Science (New York, N.Y.)* 286.5439 (Oct. 1999), pp. 481–486.

Wray, J., T. Kalkan, and A. G. Smith (2010). "The ground state of pluripotency." In: *Biochemical Society transactions* 38.4 (Aug. 2010), pp. 1027–1032.

Wray, J., T. Kalkan, S. Gomez-Lopez, D. Eckardt, A. Cook, R. Kemler, and A. Smith (2011). "Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation." In: *Nature cell biology* 13.7 (July 2011), pp. 838–845.

Wu, T. D. and S. Nacu (2010). "Fast and SNP-tolerant detection of complex variants and splicing in short reads." In: *Bioinformatics (Oxford, England)* 26.7 (Apr. 2010), pp. 873–881.

Wu, Z., X. Wang, and X. Zhang (2011). "Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq." In: *Bioinformatics (Oxford, England)* 27.4 (Feb. 2011), pp. 502–508.

Xu, H., C.-L. Wei, F. Lin, and W.-K. Sung (2008). "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data." In: *Bioinformatics (Oxford, England)* 24.20 (Oct. 2008), pp. 2344–2349.

Xu, Z., W. Wei, J. Gagneur, S. Clauder-Münster, M. Smolik, W. Huber, and L. M. Steinmetz (2011). "Antisense expression increases gene expression variability and locus interdependency." In: *Molecular systems biology* 7 (Feb. 2011), p. 468.

Yamaji, M., J. Ueda, K. Hayashi, H. Ohta, Y. Yabuta, K. Kurimoto, R. Nakato, Y. Yamada, K. Shirahige, and M. Saitou (2013). "PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells." In: *Cell stem cell* 12.3 (Mar. 2013), pp. 368–382.

Yang, J., A. L. van Oosten, T. W. Theunissen, G. Guo, J. C. R. Silva, and A. Smith (2010). "Stat3 activation is limiting for reprogramming to ground state pluripotency." In: *Cell stem cell* 7.3 (Sept. 2010), pp. 319–328.

Yang, X. P., K. Irani, S. Mattagajasingh, A. Dipaula, F. Khanday, M. Ozaki, K. Fox-Talbot, W. M. Baldwin, and L. C. Becker (2005). "Signal transducer and activator

of transcription 3alpha and specificity protein 1 interact to upregulate intercellular adhesion molecule-1 in ischemic-reperfused myocardium and vascular endothelium." In: *Arteriosclerosis, thrombosis, and vascular biology* 25.7 (July 2005), pp. 1395–1400.

Ye, S., P. Li, C. Tong, and Q.-L. Ying (2013). "Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1." In: *The EMBO journal* (Aug. 2013).

Ying, Q.-L., J. Nichols, I. Chambers, and A. Smith (2003). "BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3." In: *Cell* 115.3 (Oct. 2003), pp. 281–292.

Ying, Q.-L., J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen, and A. Smith (2008). "The ground state of embryonic stem cell self-renewal." In: *Nature* 453.7194 (May 2008), pp. 519–523.

Yoshida, K, I Chambers, J Nichols, A Smith, M Saito, K Yasukawa, M Shoyab, T Taga, and T Kishimoto (1994). "Maintenance of the pluripotential phenotype of embryonic stem cells through direct activation of gp130 signalling pathways." In: *Mechanisms of development* 45.2 (Feb. 1994), pp. 163–171.

Zang, C., D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng (2009). "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data." In: *Bioinformatics (Oxford, England)* 25.15 (Aug. 2009), pp. 1952–1958.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." In: *Genome research* 18.5 (May 2008), pp. 821–829.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." In: *Genome biology* 9.9 (2008), R137.

Zhao, Z., G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." In: *Nature genetics* 38.11 (Nov. 2006), pp. 1341–1347.

Zhong, Z, Z Wen, and J. E. Darnell (1994). "Stat3: a STAT family member activated by tyrosine phosphorylation in response to epidermal growth factor and interleukin-6." In: *Science (New York, N.Y.)* 264.5155 (Apr. 1994), pp. 95–98.

Zhou, Q., H. Chipperfield, D. A. Melton, and W. H. Wong (2007). "A gene regulatory network in mouse embryonic stem cells." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.42 (Oct. 2007), pp. 16438–16443.

Zhu, L. J., C. Gazin, N. D. Lawson, H. Pagès, S. M. Lin, D. S. Lapointe, and M. R. Green (2010). "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data." In: *BMC bioinformatics* 11 (2010), p. 237.