Chapter 1
The Analyses of Global Gene Expression and Transcription Factor Regulation

Raquel Cuevas Diaz Duran, Sudheer Menon and Jiaqian Wu

Abstract A major challenge in molecular cell biology lies in understanding how the same genome can give rise to different cell types and how gene expression is regulated. Gene expression and regulation studies focus on the abundance and structure of transcripts as well as how RNA production is controlled. High-throughput sequencing technologies such as RNA sequencing have allowed more accurate profiling of the transcriptome and the rapid identification of differentially expressed genes among samples. The regulation of gene expression is orchestrated by transcription factors. The development of ChIP sequencing assay has made it possible to comprehensively identify transcription factor-binding sites in vivo, allowing rapid unraveling of signaling pathways. The following chapter described the common methods used in studying global gene expression and transcription factor regulation with a special emphasis on bioinformatic analyses. The final section illustrates an example of an integrated gene expression and regulation study for identifying key factors regulating self-renewal and differentiation in hematopoietic precursor cells.

Keywords RNA sequencing · ChIP sequencing · Transcription factors · Transcriptome

R. Cuevas Diaz Duran · S. Menon · J. Wu (✉)
The Vivian L. Smith Department of Neurosurgery, University of Texas Medical School at Houston, Houston, TX 77030, USA
e-mail: jiaqian.wu@uth.tmc.edu

R. Cuevas Diaz Duran · S. Menon · J. Wu
Center for Stem Cell and Regenerative Medicine, UT Brown Institution of Molecular Medicine, Houston, TX 77030, USA

J. Wu, *Transcriptomics and Gene Regulation*,
Translational Bioinformatics 9, DOI 10.1007/978-94-017-7450-5_1

1.1 Introduction

Gene transcription and regulation are important areas of study because they underlie many biological processes and phenotypic variations in living organisms. Aberrant gene expression and regulation lead to diseases. The transcriptome consists of all transcripts synthesized in an organism including protein-coding, noncoding, alternatively spliced, polymorphic, sense, antisense, and edited RNAs. Transcriptome data analyses, namely the analyses of gene expression levels and structures, are essential for interpreting the functional elements of the genome and understanding the molecular constituents of cells and tissues. The regulation of gene expression is a basic mechanism through which RNA production is coordinated, and it controls important events such as development, homeostasis, and responses to environmental stimuli. Transcription factors, a type of DNA-binding proteins which recognize specific sequences, and other proteins work together through a variety of mechanisms to regulate gene transcription.

In this volume, different aspects regarding the analyses of gene transcription and regulation are described in individual chapters. In this chapter, we focus on gene expression level analyses by RNA sequencing (RNA-Seq) and transcription factor regulation by chromatin immunoprecipitation coupled with sequencing (ChIP-Seq). First, we review some useful methods developed in the past for characterizing global gene transcription.

1.2 Methods for Characterizing Global Gene Transcription

1.2.1 EST Sequencing

It is widely recognized that expressed sequence tag (EST) sequencing has provided an invaluable resource for identification of novel human genes [1, 2]. EST clustering methods allow EST to be systematically mapped, so that the information is readily integrated into the positional cloning project UniGene database (http://www.ncbi.nlm.nih.gov/UniGene). Because ESTs are from single-pass sequencing, they have to be carefully analyzed to remove genomic DNA and other contaminating sequences, such as mitochondrial, ribosomal, vector, and bacterial sequences. However, EST databases still contain a significant portion of (estimated 5–10 %) artifact sequences such as intronic or intergenic DNA [3, 4]. This is likely due to the presence of heterogeneous nuclear RNA (hnRNA) in RNA samples where EST libraries were generated. Moreover, it is difficult to understand the relationships among short EST sequences. EST clustering may confuse genes sharing similarities and alternatively spliced transcripts. Additionally, because of their short length and generally low quality, ESTs only provide limited information about gene structure and function [1]. Since EST sequencing is biased toward genes with high expression levels, the transcripts which are tissue-specific or low-abundant are less likely to be disclosed by EST sequencing. Therefore, methods that are less biased, more accurate, and sensitive are needed [5].

1.2.2 SAGE

The serial analysis of gene expression, or SAGE [6], is another technique used to quantify gene expression levels. SAGE method is designed to add a 9–14 bp tag adjacent to an *Nla*III restriction site at the gene's 3′ end. It measures transcript levels by automatically sequencing and counting each SAGE tag. The expression level of SAGE tags is analyzed and accessioned through the GEO repository. Additionally, an anatomic viewer "SAGE Genie" makes it easy to search and visualize the transcription level in different tissues and cell types of the human body (http://cgap.nci.nih.gov/SAGE). However, SAGE is a high-throughput technology which measures not the expression level of a gene, but a "tag" that represents a transcript. Due to alternative transcription, sequencing errors, and other potential effectors, sometimes two or more genes share the same tag or one gene may have more than one tag. Thus, the potential loss of fidelity should be taken into consideration.

Long serial analysis of gene expression (LongSAGE) is an adaptation of the original SAGE approach that can be used to rapidly identify novel genes and exons [7]. Instead of using an *Nla*III restriction site, LongSAGE uses a modification of longer tags (21 bp) added to a different restriction site (*Mme*I). The 21 bp tags include a constant 4 bp restriction site sequence where the transcript was cleaved and a unique 17 bp adjacent sequence of each transcript. The advantage of LongSAGE is the uniqueness of each tag in the human genome, which is not guaranteed by 14 bp tags. Sequencing tag concatamers and searching for the localization of tags in the genome help to verify predicted genes and to identity novel transcripts. LongSAGE has been reported to be at least an order of magnitude more efficient than EST sequencing [8].

1.2.3 Full-Length CDNA Sequencing Projects

In order to better access the biological information of genes including location of open reading frames, 5′ and 3′ untranslated regions, and splicing patterns, full-length and high-quality sequences of cDNAs are needed. cDNA sequencing is a valuable resource not only for characterizing the structure and function of known genes, but also for discovering novel genes. Especially with the completion of the human genome, the comparisons of the full-length and high-quality cDNA sequences with genomes are especially useful in identifying alternative gene structure and better understanding transcriptome composition during physiological and disease processes. Moreover, full-length cDNA sequencing projects paved the way for proteomic study by identifying new enzymes and proteins, generating physical clones for expression profiling, testing protein interactions, and generating hypotheses for biochemical studies.

There have been multiple efforts that aim at capturing the sequence of full-length clones which can be directly obtained from cDNA libraries generated from mammals and other selected organisms, such as *zebra fish, drosophila,* and *Caenorhabditis elegans*, mouse, and human [9–15]. In particular, NIH Mammalian Gene Collection project, utilizing large-scale RT-PCR-based cloning methods, provided thousands of clones of full-length human and mouse open reading frames [16–18].

1.2.4 Microarrays

DNA microarray is a hybridization-based technology which enables researchers to analyze the expression of large number of genes in a single reaction. DNA microarray technology was developed in the early 1990s. The technical advancement of this methodology is to manufacture slides or chips with thousands of nucleic acid probes immobilized on a small surface area. DNA probes are conformed of several specific DNA sequences of genes to which cDNA samples are hybridized. Samples, also referred to as targets, may be obtained from cells in different biological or experimental conditions, tissues, organisms, or developmental stages. Probe–target hybridization is quantified by fluorescent labeling. The signal intensities captured as images after scanning are converted into a data matrix and processed using software specific to the application of the array to determine relative abundance of specific cDNA sequences from the samples. The DNA microarray is an effective tool to investigate the structure and activity of genes at a genome-wide scale, and it helps to elucidate the molecular mechanisms underlying normal and dysfunctional biological processes [19–24].

1.2.5 RNA-Seq

Even though microarray technology has provided valuable insights into gene function throughout the last decade, it suffers from limitations in resolution, dynamic range, and accuracy. The recently developed RNA-Seq methodology uses next-generation sequencing (NGS) to sequence cDNAs generated from RNA samples producing millions of short reads. The number of reads mapped within a genomic feature of interest (such as a gene or an exon) can be used as a measurement of the feature's abundance in the analyzed sample. Typical RNA-Seq procedure is depicted in Fig. 1.1. Briefly, RNAs are converted to a library of cDNA fragments with adaptors attached to one or both ends. The molecules, with or without amplification, are then sequenced with high-throughput technology, and

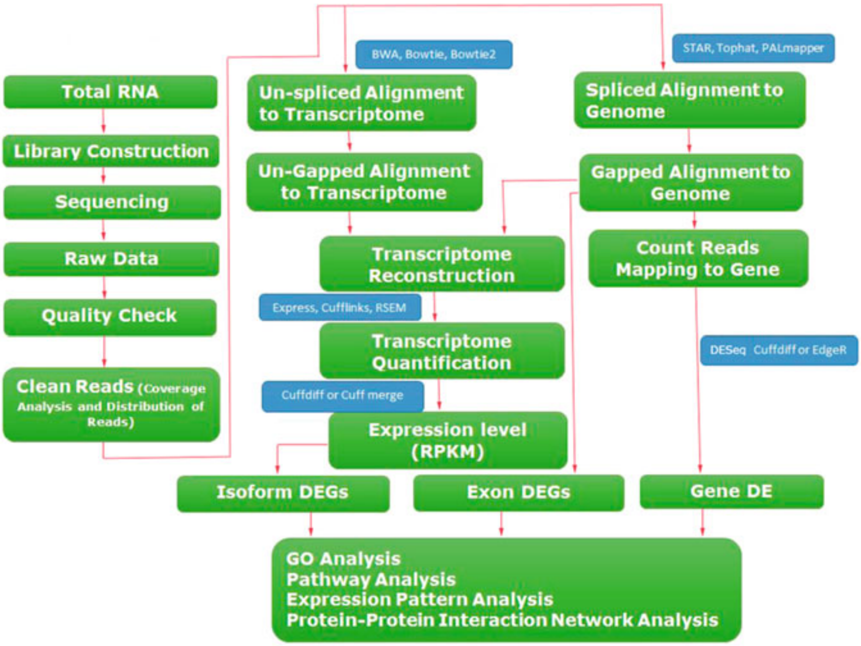

Total RNA
Library Construction
Sequencing
Raw Data
Quality Check
Clean Reads (Coverage Analysis and Distribution of Reads)
BWA, Bowtie, Bowtie2
Un-spliced Alignment to Transcriptome
Un-Gapped Alignment to Transcriptome
Transcriptome Reconstruction
Express, Cufflinks, RSEM
Transcriptome Quantification
Cuffdiff or Cuff merge
Expression level (RPKM)
STAR, Tophat, PALmapper
Spliced Alignment to Genome
Gapped Alignment to Genome
Count Reads Mapping to Gene
DESeq Cuffdiff or EdgeR
Isoform DEGs
Exon DEGs
Gene DE
GO Analysis
Pathway Analysis
Expression Pattern Analysis
Protein-Protein Interaction Network Analysis

Fig. 1.1 RNA-Seq workflow. RNA-Seq begins with isolation of high-quality total RNA followed by conversion into cDNA, fragmentation, and adaptor ligation. Fragmented cDNA is used to construct a library for sequencing. Raw data, consisting of reads of a defined length, are preprocessed according to a set of quality control metrics, such as base quality, minimum read length, untrimmed adaptors, and sequence contamination. After filtering and trimming, reads are aligned to a reference genome or transcriptome depending on the objective of the experiment and the nature of the samples. Subsequently aligned reads are assembled. RNA-Seq assembly involves merging reads into larger contiguous sequences based on similarity. After assembly, reads are quantified in order to measure transcriptional activity. Read counts are generally computed in RPKM of FPKM in order to perform further downstream analysis, such as differential expression, pathway and gene set overrepresentation analysis, and interaction networks

short sequences from one end or both ends are obtained. The reads are typically 30–400 bp, depending on the DNA sequencing technology used. There are various high-throughput sequencing platforms available for RNA-Seq such as Illumina, Applied Biosystems SOLiD, and Roche 454 Life Science systems. The total reads obtained after sequencing are either aligned to a reference genome or transcriptome, or assembled de novo without genomic sequence guidance to create a genome-scale transcription map providing both transcriptional structure and expression level for each gene.

RNA-Seq has several advantages over microarrays [25–27]. First, sequencing technology is much more sensitive and quantitative than microarrays and it can provide a large dynamic range of detection (>9000-fold) [28]. Additionally, sequencing data are more specific and have less background. Moreover, sequencing experiments do not depend on the limited features of tiled microarrays and can therefore be used to interrogate any location in the genome and to detect and quantify the expression of previously unknown transcripts and splicing isoforms. Finally, sequencing is not limited by array hybridization chemistry, such as melting temperature, cross-hybridization, and secondary structure concerns.

1.2.5.1 Data Analysis General Workflow

RNA-Seq experiments result in millions of short sequence reads which require computational methods for comprehensive transcriptome characterization and quantification. Steps for data analysis vary according to the desired biological problem to be assessed and to the availability of reference genome or transcriptome data. A generic overview of the routine analyses performed is included in Fig. 1.1. The main tasks of data analysis are read mapping also referred to as alignment, transcriptome assembly, expression quantification, and downstream applications. Steps for data analysis, although it is sequential, may be performed with different computational tools and algorithms which require specific data formats and external files. It is desirable to automate the multiple data analysis steps in a pipeline. A pipeline is a reusable script with defined inputs, outputs, and parameters for each processing step. Several software platforms which collect different RNA-Seq analysis tools for each task have been developed such as PRADA [29], Tuxedo [30], MAP-R-SEQ [31], and GENE-Counter [32]. However, pipelines may be custom-built by users, selecting the most appropriate tools according to experimental data and desired downstream analysis [33]. The following subsections describe the different RNA-Seq data analysis tasks.

1.2.5.2 Quality Control and Preprocessing

The first step in data analysis is quality control. Accuracy in library preparation and sequencing steps contribute to the quality of reads, which if overseen may lead to erroneous mappings, misassembles, and false expression estimates. The quality control should include the assessment of read length, GC content, sequence complexity, sequence duplication, polymerase chain reaction artifacts, untrimmed adapters, low-confidence bases, 3′/5′ positional bias, sequence contamination, and fragment biases [34]. Quality control metrics are obtained directly from raw reads. Raw reads are typically in FASTQ format, text-based files which contain a sequence identifier, a nucleotide sequence, and its corresponding quality score [35]. A brief overview of the most important quality control processes to be performed is described next.

(a) Base Quality: Since RNA-Seq technologies rely on complex interactions between chemistry, hardware, and optical sensors, sequencing platforms provide metrics for assessing error probabilities. Base quality is measured by computing the confidence on base calling, the process by which the sequencer analyzes colorimetric sensor signals to predict individual bases. Base calling quality values are expressed in *Phred* scale, an error probability log transformation which has the advantage of converting low error probabilities to high-quality values and vice versa [36]. Quality values q are calculated for each base as:

$$q = -10 \log_{10}(p) \tag{4.1}$$

where p is the estimated error probability of a base call. Base quality values are encoded with ASCII characters with the sequence data in FASTQ format [35]. Typically, good reads should have base qualities greater than 20; however, this threshold depends on the platform used. It is important to inspect the reads' base quality distribution to detect regions of poor base quality, which may be filtered or trimmed preserving the order of reads, thus increasing mapping efficiency. This process is also referred to as quality trimming and will be discussed next.

(b) Filtering and Trimming: Reads should be inspected for the presence of sequencing adapters, tags, and contaminating sequences, which should be removed before quality control processing. Adapter sequences and tags used during library construction should be removed prior to mapping. Contaminating sequences such as DNA, rRNA, or sequences from other organisms or vectors should be filtered out.

Additionally, reads should be filtered according to mean base quality or to the proportion of bases whose quality is below a user-defined threshold. There is no consensus on the optimal base quality threshold for trimming, and it is rather a trade-off between mapping efficiency and coverage. Software for filtering generally outputs synchronized filtered reads. When low-quality bases are located at the ends of reads, trimming is a better option than filtering. The basic principle of read trimming is to assess base quality keeping the longest possible high-quality read segments. Trimming with respect to base quality may be performed using running sum algorithms or sliding window-based algorithms. Running sum algorithms basically find the summation of the differences between all base quality values against a quality threshold, and sequences are trimmed at the base that makes the running sum minimum. When analyzing reads with a sliding window, the user defines a window size and a mean base quality threshold. Depending on the tool used, the window may slide from the 5′ or the 3′ ends of reads. Sliding the window from the 5′ end will trim the read until the window's quality lies below a threshold, maintaining the beginning of the sequence, whereas sliding from 3′ end will trim until a passing quality window is encountered. An excellent evaluation on the performance of several trimming tools was published by Del Fabbro et al. [37]. Some common tools used for trimming are Trimmomatic [38], Cutadapt [39], PRINSEQ [34], and ConDeTri [40].

Reads with a high frequency of ambiguous bases, bases not identified during sequencing, should be filtered out since they can lead to erroneous mapping and misassemblies. Low-complexity sequences (homopolymers, di-trinucleotides) will also result in mapping errors and should therefore be trimmed.

(c) GC Content Determination: Another important metric that should be considered is reads' mean GC content, which if plotted should follow a normal distribution centered on the organism's normal content. Variations on the GC content might be due to the PCR amplification process and therefore may be sample specific. An approach to reduce GC content systematic bias is conditional quantile normalization, a technique in which the distribution of read counts is modified by estimating quantiles obtained with a median regression on a subset of genes [41].

(d) Minimum Read Length: The distribution of read lengths should be verified after trimming since reads may have ended up as very short fragments, which become difficult for mapping. The minimum read length is a user-defined variable, and its value depends on desired downstream applications.

(e) Fragment Biases: RNA fragmentation creates segments whose starting points were assumed to be located uniformly at random within a transcript. However, it has been demonstrated that there are both positional [42] and sequence-specific [43, 44] biases derived from fragmentation and reverse transcription. Positional bias describes the fact that reads' starting positions are non-uniformly distributed since they are preferentially located toward the transcripts' boundaries. Sequence-specific bias refers to the phenomenon by which the sequences at the reads' boundaries, such as the random hexamer primers used for reverse transcription priming, introduce biases in nucleotide composition and influence the likelihood of being sequenced. Furthermore, fragment length also generates bias since long transcripts result in more reads mapping to them than smaller transcripts. Thereby, for genes with equal levels of expression, the long genes will be overrepresented, distorting the relative expression among genes [45]. Since RNA-Seq read counts are proportional to transcript abundances, expression estimates should be made after fragment bias correction. An effective approach for fragment bias correction has been implemented in the Cufflinks [46] transcriptome assembly and differential expression RNA-Seq analysis tool. The fragment bias correction was based on an algorithm which learns the read sequences and models them as a likelihood function involving abundance and bias parameters such as the probability of finding a fragment with a specific length in a given position [47]. In this manner, bias and expression estimation are performed simultaneously.

1.2.5.3 Read Alignment

In order to determine transcript abundance from reads, it is necessary to align or map reads to a previously assembled reference genome or transcriptome to determine the read's origin. Mapping to a reference genome is more common since it increases the potential information which may be obtained (e.g., identification of novel transcripts and genes) and because many transcriptomes are incomplete. Mapping is a challenging task since RNA-Seq reads are relatively short and they may match non-contiguous regions of the genome due to splice junctions. Furthermore, alignment tools must cope with mismatches and indels caused by genomic variation and sequencing errors. Many alignment tools have been developed. A comprehensive list of alignment tools and their properties was initially published by Fonseca et al. [48] and is kept updated on the Web [49]. A list of some common aligners and their main properties is included in Table 1.1.

The main consideration to be addressed when selecting an aligner is whether RNA-Seq reads span splice junctions. As depicted in Fig. 1.1, unspliced or contiguous alignment tools such as BWA [50], Bowtie [51], and Bowtie2 [52] are useful when mapping reads to a transcriptome, when sequencing microRNAs, or when the organism under study has no intronic regions. Unspliced aligners are thus limited to identifying known exons and do not allow for new splicing event identification. Spliced alignment tools are used when mapping to reference genomes without relying on previously known splice sites. Some of the most commonly used tools for spliced alignment are TopHat [53], TopHat2 [54], Palmapper [55], and STAR [56].

Alignment to a reference genome starts with indexing, the process with which auxiliary structures called indices are created for either the reference sequence or the sequenced reads to allow for faster queries. Indexing the reference genome is more time efficient and thus is used by most alignment tools. Alignment algorithms used for sequencing data analysis are mainly classified into hash tables and suffix trees according to the property of the index used. Hash table indexing was first introduced as an alignment tool by BLAST [67], using a seed and extend approach. In hash table indexing, reads are divided into short *k-mer* subsequences called "seeds" and stored in a hash table. The algorithm assumes that at least one "seed" in a read will match the reference. Once a "seed" is aligned, it is extended using more sensitive algorithms such as Smith–Waterman [68] or Needleman–Wunsch [69]. Modifications to hash table indexing algorithms have been performed, and they have been implemented in Novoalign [59], MAQ [65], SHRiMP2 [70], and BFAST [57], among other alignment tools. Suffix trees, on the other hand, are based on the premise that an inexact matching problem may be converted into an exact matching task by constructing a tree (an ordered tree data structure) with all the possible substrings that make up a sequence. The suffix tree data structure enables fast substring searches regardless of sequence size [71]. Among different suffix tree algorithms, one of the most efficient is the FM-index [72] which is based on the Burrows–Wheeler transform (BWT) [73]. BWT is a reversible permutation of characters in a string, and FM-indexing addresses permutations (nodes in a tree) constantly using a backward search. FM-index and BWT, both originally designed for data compression, have been successfully implemented for storing reference genomes and performing rapid queries.

Table 1.1 Overview of common alignment tools

ALIGNERS	Operating system	Language	Alignment algorithm	Input	Output	Paired-end mapping	Splice junction	Read length range	Ref.
BOWTIE	Unix-based, windows	C++	FM-index based on BWT	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–1 k	[51]
BOWTIE2	Unix-based, windows	C++	FM-index based on BWT, dynamic programming	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–5000 k	[52]
PALMapper	Unix-based, web interface	C++	Reference indexing	FAST(A/Q)	SAM, BED (x), SHORE	Yes	Yes	12 bp–12 k	[55]
STAR	Unix-based	C++	Reference indexing	FAST(A/Q)	SAM	Yes	Yes	15 bp–10 k	[56]
BFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, TSV	Yes	No	25–100 bp	[57]
GENOME-MAPPER	Unix-based	C	Reference indexing	FAST (A/Q), SHORE	BED, SHORE	No	No	12 bp–2 k	[58]
NOVAALIGN	Unix-based	C++	Reference indexing	FAST (A/Q), CSFASTA	SAM	Yes	Yes	1–250 bp	[59]
SHRiMP2	Unix-based	Python	Reference indexing	FAST(A/Q)	SAM	Yes	No	30 bp–1 k	[60]
SOAP2	Unix-based	C++	BWT + reference indexing	FAST(A/Q)	SAM/BAM	Yes	No	27 bp–1 k	[61]
MrFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, DIVET	Yes	No	25 bp–1 k	[62]
GNUMAP	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, TSV	No	No	16 bp–1 k	[63]
RMAP	Unix-based	C++	Read indexing	FAST(A/Q)	BED	Yes	No	11 bp–10 k	[64]

(continued)

Table 1.1 (continued)

ALIGNERS	Operating system	Language	Alignment algorithm	Input	Output	Paired-end mapping	Splice junction	Read length range	Ref.
MAQ	Unix-based	C, C++, Perl	Hashing reads	FAST(A/Q)	TSV	Yes	No	8–63 bp	[65]
Mosaik	Unix-based, windows	C++	Reference indexing	FAST(A/Q)	BAM	Yes	No	15 bp–1 k	[66]
BWA	Unix-based, windows	C, C++	FM-index based on BWT	FAST(A/Q)	SAM	Yes	No	4–200 bp	[50]

Reference genome indexes may be built or downloaded as GTF/GFF annotation files most commonly from Ensembl [74] and Illumina iGenomes Web sites [75]. GTF files must be selected carefully using a standard assembly so that chromosome names, gene identifiers, transcription starting sites, and all genomic annotations match between experiments.

Bowtie and Bowtie2 are two of the most efficient unspliced alignment tools because of their low memory requirements and high speed; they both implement an FM-indexing algorithm for achieving ultra-fast alignments. However, neither of these tools are suitable for performing spliced alignments since they cannot align reads when there are large gaps (introns). TopHat addresses spliced alignment limitations by performing a multistep alignment process and using Bowtie as an alignment engine. In the first step, reads are mapped to a reference genome, setting aside reads which were not aligned. In the next step, reads that could not be mapped are broken down into segments and remapped. Finally, reads whose segments were mapped into the same user-defined intronic region are assembled and mapped to that genomic region in an attempt to find splice sites. With this approach, TopHat identifies splice sites without previous splice site annotations and finds novel splicing events [53].

RNA-Seq alignment results are output as SAM/BAM files, and they generally need some further processing such as conversion, sorting, indexing, or merging. SAMTools, implemented in C and Java, is a library for parsing and manipulating alignments prior to downstream analysis [76]. Visualizing aligned reads in a genomic context is recommended for assessing exon coverage, spotting indels and SNPs, displaying splice junctions, identifying novel transcripts, etc. Some available tools for visualization of alignment files are Integrative Genomics Viewer (IGV) [77], Savant [78], and Integrated Genome Browser (IGB) [79].

1.2.5.4 Transcriptome Assembly

In order to quantify gene expression levels from aligned reads, it is necessary to identify which gene isoform generated each read. Therefore, the main aim of transcript assembly is to reconstruct complete transcripts from small overlapping fragments. There are several methods for transcriptome reconstruction, and they can be categorized into genome-guided and genome-independent methods. In genome-guided methods, reads are first mapped to a reference genome and a splicing or exon graph is then constructed for each gene to identify all possible isoforms according to exon combinations. In the splicing graph, each node represents an exon and each connection is an exon junction. Paths that are not evidenced by RNA-Seq reads are eliminated. There are different graph topologies which are implemented to best describe exon combinations for building transcript isoforms. One of the most commonly used tools for genome-guided transcript assembly is Cufflinks, which connects aligned reads based on the location of their spliced alignments [46].

Genome-independent transcriptome reconstruction aims at finding as many long contiguous segments as possible from an assembly graph. The most common strategy is to build a de Bruijn graph, which models overlapped sequence data as a set of *k-mers* (*k* consecutive nucleotides) and their connections [80–82]. Sequences are represented as paths, and branches not supported by reads are eliminated; remaining paths are considered transcript variants. The length of the *k-mer* has an effect on the complexity of the graph, and, although it is conceptually simple, de Bruijn reconstruction approaches have complications such as finding the balance between sensitivity and graph complexity [83]. The value of *k* must be smaller than the read length. However, if *k* is too small, the graph will have excessive connections and will be very sensitive to sequencing errors. If *k* is too large, there must be enough data to make the graph connected. To resolve such issues, several assemblies should be performed with variable values of *k*. Some common de novo assemblers based on de Bruijn graphs are ABySS [84], Trinity [85], Velvet [82], and Oases [86].

1.2.5.5 Expression Quantification

Expression quantification may be performed with respect to transcripts or to genes. Gene expression, the sum of the expression of all its isoforms, is computed by counting reads per gene according to the reference genome's annotation used for mapping. Read counts need to be normalized due to variability introduced by read length bias [45, 87] and due to fluctuations in the number of reads per run [88]. Quantification tools generally output read counts in raw counts, reads per kilobase of transcript per million mapped reads (RPKM), or fragments per kilobase of transcript per million mapped reads (FPKM). RPKM measure normalizes read counts according to the length and to the number of mapped reads per sample [88]. FPKM is used for normalization of paired-end reads since it incorporates dependency estimation [46]. In statistical dependence between two variables (paired-end reads), the levels of one of the variables vary in an exactly determined way with respect to levels of the other variable. All quantification tools are taken as input read alignments in SAM/BAM formats and their reference genome annotation files in GTF/GFF or BED format. They differ in how they handle multimapping reads, which affects expression quantification accuracy [46, 89, 90]. To deal with mapping uncertainty, tools such as Cufflinks use a maximum likelihood function which works by dividing multimapping reads probabilistically according to the abundance of genes they were mapped to [91].

1.2.5.6 Differential Expression Analysis

Often, it is necessary to compare the expression levels of genes or other genomic features between different samples or biological conditions; this is referred to as differential expression analysis. Comparisons are typically performed in a univariate way since it is not possible to fit a multivariate statistical model due to the number of samples being much less than the number of genes.

The ability of detecting differential expression in RNA-Seq experiments depends on the sequencing depth, gene expression, and even on the gene's length, as previously mentioned. A difference in gene expression between two groups is significant only if it is greater than the variability within the group. For estimating variability, biological replicates should be considered. The number of replicates to be used depends on the experiment and the statistical power desired. The purpose of replication is to estimate the variability between and within groups, which is important for hypothesis testing. The set of standards, guidelines, and best practices for RNA-Seq published by the ENCODE Consortium [92] states that two or more biological replicates are sufficient as long as the Pearson correlation of gene expression between them lies between 0.92 and 0.98.

Since RNA-Seq experiments are based on read counts, the initial methods for differential analysis modeled reads as Poisson distributions [46, 87]. However, due to biological variability and the limited number of samples, methods that model count variability as a nonlinear function of mean counts with parametric approaches (e.g., normal, negative binomial distributions) have become popular. Commonly used tools such as DESeq [93], edgeR [94], and Cuffdiff [46] use a negative binomial distribution for modeling RNA-Seq counts. Recently, it has been suggested that RNA-Seq count data may be transformed to apply normal-based microarray-like statistical methods as in the case of Limma software [95]. RNA-Seq data must be normalized transforming counts to have similar empirical distributions across all samples in order to enable comparisons between samples and genes. This step is executed internally by differential expression analysis tools. Table 1.2 makes a comparison of some commonly used differential expression analysis tools. A differential expression analysis should produce a ranked list of differentially expressed genes to be used in downstream applications.

1.2.5.7 Downstream Analysis

Interpretation, visualization, and summarization of differential expression results are important for downstream interpretations. Heat maps and PCA plots are common for finding clusters of differentially expressed genes.

It is of interest to correlate differentially expressed genes to gene sets representing functions, categories, pathways, and others incorporating existing biological knowledge into the analysis. An overrepresentation analysis requires a list of differentially expressed genes which are tested statistically for enrichment in gene sets such as gene ontology (GO) categories, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome pathways, and many other databases [96, 97].

Table 1.2 Overview of common tools for differential expression analysis

Properties	Cuffdiff	DESeq	edgeR	baySeq	Limma
Language	C++	R	R	R	R
Operating system	Unix-based OS	Unix-based OS, Windows	Unix-based OS, Windows	Unix-based OS, Windows	Unix-based OS, Windows
Data normalization	Geometric mean, quartile, FPKM	Scaling	Model-based global scaling (TMM)	Scaling, quantile, TMM	Quantile normalization, loess regression, TMM
Read count distribution	Beta negative binomial distribution	Negative binomial distribution	Negative binomial distribution	Negative binomial distribution	Voom transformation of counts into a log distribution
Differential expression test	t-test	Fisher's exact test	Fisher's exact test	Empirical Bayes method to obtain posterior probabilities	Empirical Bayes method
False discovery rate (FDR) estimation method	Benjamini–Hochberg	Benjamini–Hochberg	Benjamini–Hochberg	Bayesian	Benjamini–Hochberg
Reference	[30]	[93]	[94]	[98]	[95]

TMM Trimmed mean of M values

1.3 Characterizing Transcription Factor Regulation by ChIP-chip and ChIP-Seq Methods

The mapping of binding sites for transcription factors (TF), the core transcriptional machinery, and other DNA-binding proteins is essential for understanding gene regulation. Regulatory networks formed by transcription factors and the coordinated activation of their specific target genes play a major role in controlling many cellular processes. The traditional way of constructing gene regulatory networks, via sequential analysis of one or a few genes, is time consuming and labor intensive. Recently, the development of ChIP-chip or ChIP-Seq technology has made it possible to comprehensively identify most in vivo target genes of a given TF at a genome-wide scale, allowing rapid unraveling of signaling pathways [99–101].

In ChIP experiments, TFs are first cross-linked to DNA by treatment with formaldehyde, and chromatin is fragmented to ~300–500 bp fragments. TF-bound chromatins are then immunoprecipitated with specific antibody. Next, the cross-link is reversed by heating to release the precipitated DNA. Immunoprecipitated DNA fragments are hybridized to a microarray (ChIP-chip) or sequenced to generate millions of short sequence tags (ChIP-Seq). Various arrays have been used for ChIP-chip analysis, for example, proximal promoter arrays where about ~1 kb PCR products encompassing transcription start sites are used as probes; arrays composed of CpG islands amplified by PCR; large promoter arrays which consist of tiling oligonucleotides of promoter sequences extending up to several kb upstream of the transcription start site; and genome tiling arrays in which a non-repetitive sequence from entire chromosomes is reconstituted using oligonucleotides. As chromosomal sequence is densely covered, higher resolution can be achieved with genome tiling microarrays.

As described previously, sequencing offers various advantages over microarray methods; thus, it has become the predominant technique for profiling genome-wide protein–DNA interactions, chromosomal proteins, and histone marks in vivo [102–104]. For example, the ChIP-Seq assays have higher resolution, lower noise, and better genomic coverage when compared to ChIP-chip assays. Therefore, ChIP-Seq provides more precise mapping of protein-binding sites and sequence motif identification [103, 105]. Several factors influencing ChIP-Seq fidelity need to be addressed.

1.3.1 Analysis of ChIP-Seq Data

The typical output of a ChIP-Seq experiment is a list of millions of short sequence reads. Processing such reads requires filtering and cleaning, mapping to a reference genome, and identification of peak regions. Once significant peaks have been identified, they must be examined, annotated, and associated to a genomic region. The final result is the identification of a transcription factor’s motif and binding sites. A general ChIP-Seq workflow is shown in Fig. 1.2. The main issues to consider when analyzing ChIP-Seq data are the following:

(a) Control Sample: ChIP-Seq experiments are prone to artifacts due to effects of DNA shearing and repetitive DNA sequences. DNA shearing during sonication is not uniform because open chromatin regions are fragmented more easily, thus resulting in an uneven distribution of reads. Repetitive DNA sequences may seem enriched when the number of repeats is not considered in calculations. Therefore, the use of a control sample is recommended for peak comparison and significance assessment. Three commonly used control samples are as follows: DNA prior to immunoprecipitation, immunoprecipitated DNA without an antibody, and immunoprecipitated DNA using a non-DNA-binding antibody (e.g., anti-IgG antibody). There is no consensus on which is the most appropriate control.

(b) Sequencing Depth: For a ChIP-Seq analysis to be effective, sufficient genomic coverage, referred to as sequencing depth, is important. However, sequencing

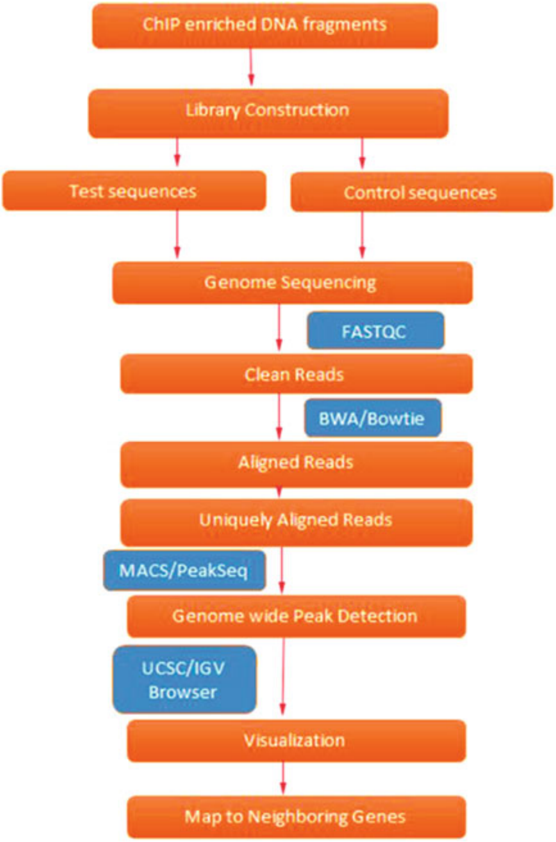


Fig. 1.2 ChIP-Seq workflow. ChIP-Seq experiments allow in vivo determination of where proteins, such as transcription factors, bind to the genome. Bound proteins are cross-linked to chromatin, then fragmented, and immunoprecipitated. ChIP-enriched DNA fragments are used for library construction and sequencing. Reads are filtered according to base quality. Test and control sequences are used for computational mapping to identify genomic locations of bound DNA transcription factors, unveiling potential protein–DNA interactions. Mapped reads are converted into an integer count of "tags." As illustrated, different tools may be used for finding statistically relevant peaks. Finally, the peaks can be visualized and mapped to nearby genes

depth is a potential source of bias since at high sequencing depths, open chromatin regions generate redundant reads which represent false positives [106]. The choice of sequencing depth depends on the genome's size and on the expected number and size of the transcription factor-binding sites. Transcription factors generate highly localized ChIP-Seq signals, and for mammalian genomes, there are thousands of binding sites. For mammalian transcription factors, at least 20 million uniquely mapped reads are currently used for most experiments [107]. For histone marks or proteins with more binding sites such as RNA polymerase II, a higher sequencing depth (e.g., 60 million reads) is needed. To verify whether the sequencing depth was appropriate, a saturation analysis is recommended. Saturation analysis consists of increasing the number of randomly selected reads during read mapping and peak calling for verifying the consistency of peaks called. Saturation analyses are included in some peak caller tools such as SPP [108].

(c) Quality Control Filtering and Read Mapping: Like RNA-Seq data, ChIP-Seq reads must be preprocessed before mapping in order to identify possible sequencing errors and biases. The first filtering criterion is the base calling confidence, computed with the *Phred* quality score for each sequence tag. Low-quality reads should be filtered out and low base quality read ends trimmed. Tools for filtering, trimming, and mapping ChIP-Seq data are the same as for RNA-Seq. After filtering and trimming, the reads are aligned. Alignment/mapping of ChIP-Seq reads is less complex than RNA-Seq reads since large gaps corresponding to splice junctions will not be present. ChIP-Seq aligners generally consider mismatches due to sequencing errors, single nucleotide polymorphisms, and indels. Commonly used mappers are Bowtie [51], BWA [50], SOAP [109], and MAQ [65]. The percentage of uniquely mapped reads must be calculated, and values above 70 % are generally considered normal [110]. However, these values are organism, platform, or protocol dependent. Multimapped reads are most likely due to regions of repeated DNA, and most peak-calling algorithms will filter them out.

Library complexity is the fraction of mapped DNA fragments which are non-redundant, and it may be addressed using the PCR bottleneck coefficient (PBC) from the ENCODE project [111]. PBC computes the fraction of genomic locations with only one unique read mapped against the ones with at least one mapped read. Low-complexity libraries might be due to not enough recovered DNA, resulting in the same PCR-amplified products being sequenced repeatedly. Generally, library complexity is related to antibody quality, over cross-linking, sonication, or over PCR amplification.

(d) Background Signal: Another metric to be considered after mapping is the signal to noise ratio (SNR) of the experiment. During the ChIP-Seq experiment, most of the unbound DNA fragments are washed in the immunoprecipitation step and the library is built with protein–DNA-bound fragments. However, due to nonspecific binding of molecules, non-useful fragments may remain in the library and be sequenced. Such reads become generally spread in the genome and are referred to as background noise and may result in false positives. Noise may be computed from the control sample or modeled with a Poisson or negative binomial distribution.

(e) Peak Calling: One of the most important aims of ChIP-Seq experiments is finding enriched regions in the genome in which more transcription factors (ChIP-Seq tags) were bound to DNA through the number of mapped reads, referred to as "peaks." Several peak callers have been developed, and they mostly differ from each other in the algorithms for signal smoothing and background modeling. Models implemented for statistical assessment of peaks range from Poisson (CSAR [112]), local Poisson (MACS [113]), negative binomial (CisGenome [114]), and even some machine learning techniques, such as hidden Markov models (HPeak [115]). Peak callers report a *p* value or false discovery rate (FDR) as an enrichment metric which is greatly affected by variables, such as sequencing depth, real number of binding sites, and the statistical model used. There is no consensus on how to best estimate the best FDR value for ChIP-Seq experiments. Table 1.3 lists the main characteristics of some commonly used peak callers.

(f) Reproducibility: It is recommended to develop experiments with at least two biological replicates for verifying reproducibility of reads and identified peaks [107]. The reproducibility of reads can be computed with metrics such as Pearson correlation coefficient of mapped read counts at each genomic site [116].

(g) Downstream Analysis: Once significant and reproducible peaks have been found, it is necessary to associate them with relevant genomic regions, such as transcription start sites, gene promoters, and intergenic regions. It is common to view the identified peaks and reads in a genome browser to examine regions of interest. Generally, peaks are uploaded as BED or GFF file formats and reads with WIG file format. HOMER or BEDTools may be used to calculate distances from peaks to landmark regions (e.g., genes). The most common downstream analysis of a ChIP-seq experiment is the discovery of binding sequence motifs [117]. The read sequences of the top-scoring peaks can be entered in FASTA format into motif discovery programs such as MEME [118] resulting in motif discovery, enrichment, and location analysis.

The in vivo binding targets of TFs identified above can be further correlated with the differentially expressed genes using Gene Set Enrichment Analysis (GSEA) software [101, 123]. The factors that show enriched binding to the differentially expressed genes can be selected for further genetic testing. Finally, to understand the intricate relationship of the TFs that are differentially expressed, one can construct a network among coregulated TFs and incorporate ChIP-Seq result into the network. Thus, the underlying regulatory mechanism can be revealed, such as autoregulation (where a factor interacts with its own promoter region), cross-factor control (where pairs of factors directly bind each other's promoter regions), and positive/negative feedback loop.

Table 1.3 Overview of common peak callers used for ChIP-Seq data analysis

Software	Background model	Signal profile creation method	False discovery rate (FDR) estimation method	Statistical model for peak identification	Ref.
SPP	Poisson	Window scan	Ratio of significant scores between sample and control	Poisson model	[108]
CisGenome	Negative binomial/control sample	Window scan	Ratio between expected to observed peak number	Conditional/negative binomial model	[114]
PeakSeq	Local Poisson/control sample	Extended tag aggregation	Benjamini–Hochberg	Conditional binomial model	[119]
MACS	Local Poisson/control sample	Shift tags + window scan	Ratio between number of peaks in control and in ChIP	Local Poisson model	[113]
QuEST	Control sample	Kernel density estimation	Based on control sample	Threshold-based model	[120]
FindPeaks	Uniform	Overlap-based	Monte Carlo simulation	Peak height threshold	[121]
F-Seq	Kernel density estimation	Kernel density estimation	None	Peak height threshold	[122]

1.4 Integrated Study of Gene Expression and Transcriptional Regulation—An Example: Identification of Key Factors Regulating Self-renewal and Differentiation in EML Hematopoietic Precursor Cells by RNA-Seq and ChIP-Seq Analyses

1.4.1 The Multipotential EML Cell Line Is a Favorable System to Study the Control of Early Hematopoietic Self-renewal and Differentiation

The hematopoietic system has provided a leading model for stem cell studies, and there is great interest in elucidating the mechanisms that control the decision of HSC self-renewal and differentiation [124–130]. This switch is important for understanding hematopoietic diseases and manipulating HSCs for therapeutic purposes. However, because HSCs are currently unable to proliferate extensively in vitro, this severely limits the types of biochemical analyses that can be performed, and consequently, the mechanisms that control the decision between early-stage HSC self-renewal and differentiation remain unclear [131].

The mouse (*Mus musculus*) EML (erythroid, myeloid, and lymphocytic) multipotential hematopoietic precursor cell is an ideal system for studying the molecular control of early hematopoietic differentiation events. EML cells are derived from mouse bone marrow cells and are cultured in the presence of stem cell factor (SCF). These cells can be rederived or repeatedly cloned, and still retain their multipotentiality [132–134]. The ability of EML cells to propagate extensively in medium containing SCF makes them ideal for biochemical and genetic assays, as well as for high-throughput functional screens [126, 135]. Phenotypically, EML cells express many of the cell surface markers' characteristic of hematopoietic progenitor cells, including SCA1, CD34, and c-KIT. Functionally, when treated with different growth factors, such as SCF, IL-3, GM-CSF, and EPO, EML cells can differentiate into distinct cell lineages including B-lymphocyte, erythrocyte, neutrophil, macrophage, mast cell, and megakaryocyte lineages [132].

Interestingly, in culture, the Lin-SCA+ CD34+ subpopulation of EML cells gives rise to a mixed population containing similar numbers of self-renewing Lin-SCA+ CD34+ precursor cells and partially differentiated Lin-SCA-CD34− cells (henceforth referred to as CD34+ and CD34− cells, respectively) [136]. Although the two populations resemble each other morphologically, only the CD34+ population propagates in SCF-containing media, while the growth of CD34− cells requires the cytokine IL-3 [136]. The closest normal analogs of CD34+ cells are short-term (ST) HSC or multipotent progenitors (MPP). Similar to short-term (ST) HSC, CD34+ cells are capable of self-renewal; like MPP, when treated with cytokines such as IL-3, CD34+ cells can give rise to CD34− cells with more restricted potential. A number of erythroid genes, such as α- and β-hemoglobin, Gata1, Epor (erythropoietin receptor), and Eraf (erythroid associated factor), as well as mast cell proteases are expressed at a significantly higher level in the CD34− cell population than in CD34+ cells [136, 137]. This indicates that the CD34− cells were, at minimum, differentiated into a state with prominent erythroid potential.

The ability of CD34+ cells to both differentiate and self-renew in suspension culture, in the absence of any anatomical niche or other cell types, suggests that CD34+ cells are regulated by a tightly controlled endogenous mechanism that guides the generation of the variety and relative abundance of the cell types in culture. Understanding the molecular events that regulate the transition between the two types of putative precursors in the EML multipotent hematopoietic cell line will give insights into the fundamental mechanisms of autonomous and balanced cell fate choice available to stem cells and intermediate-stage cancer precursor cells [126].

1.4.2 Mapping Transcription Regulatory Networks and Identifying Master Regulators

The regulatory inputs and functional outputs of various downstream genes constitute network-like architectures [138]. The linkage relationships in a complex network provide causal clues about how a specific eukaryotic process is regulated at the molecular level. Using these methods, regulatory networks have been constructed for the yeast cell cycle [139–141], yeast development [141, 142], and human embryonic stem cell self-renewal [99]. For example, in the study of yeast pseudohyphal development, the binding targets of six key transcription factors (Ste12, Tec1, Sok2, Phd1, Mga1, and Flo8) were identified. The binding network formed by these factors and their target genes were analyzed, and Mga1 and Phd1 were found to be the targets of many factors in the network. These factors were called target hubs [142].

It appears that target hubs are especially likely to be “master regulators.” Master regulators have been identified as transcription factors whose ectopic expression alone is capable of activating a biological pathway. For example, MyoD is capable of activating a terminal muscle differentiation program in primary cells and in differentiated cell lines [143]. The target hubs Mga1 and Phd1 also appear to be such “master regulators,” serving as key nodal points that orchestrate a large number of regulatory inputs into a complex response [144–146]. Overexpression of either of these target hub proteins under conditions that do not normally activate the pseudohyphal response specifically induces this process. The distinct nature of the master regulators allows us to use them as a switch to control cellular processes, which has important therapeutic applications.

1.4.3 *Identifying Key Factors Regulating Self-renewal and Differentiation*

In order to identify the "switch" in cell self-renewal and differentiation, we constructed regulatory circuits controlling early hematopoietic differentiation by using the gene expression and ChIP-Seq data. We examined transcription factors that were significantly upregulated in CD34+ cells relative to CD34− cells using RNA-Seq and found *Tcf7* (also referred to by the symbol *Tcf1*) to be the most strongly upregulated transcription factor (Fig. 1.3) [27].

The binding motifs of the TCF family of transcription factors are significantly enriched among genes that are expressed at a higher level in CD34+ than in CD34−

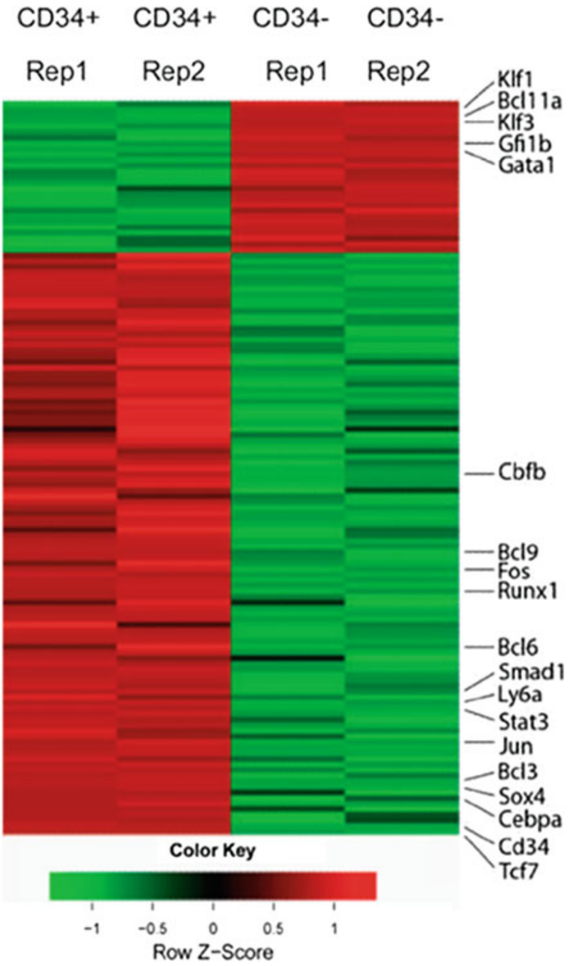
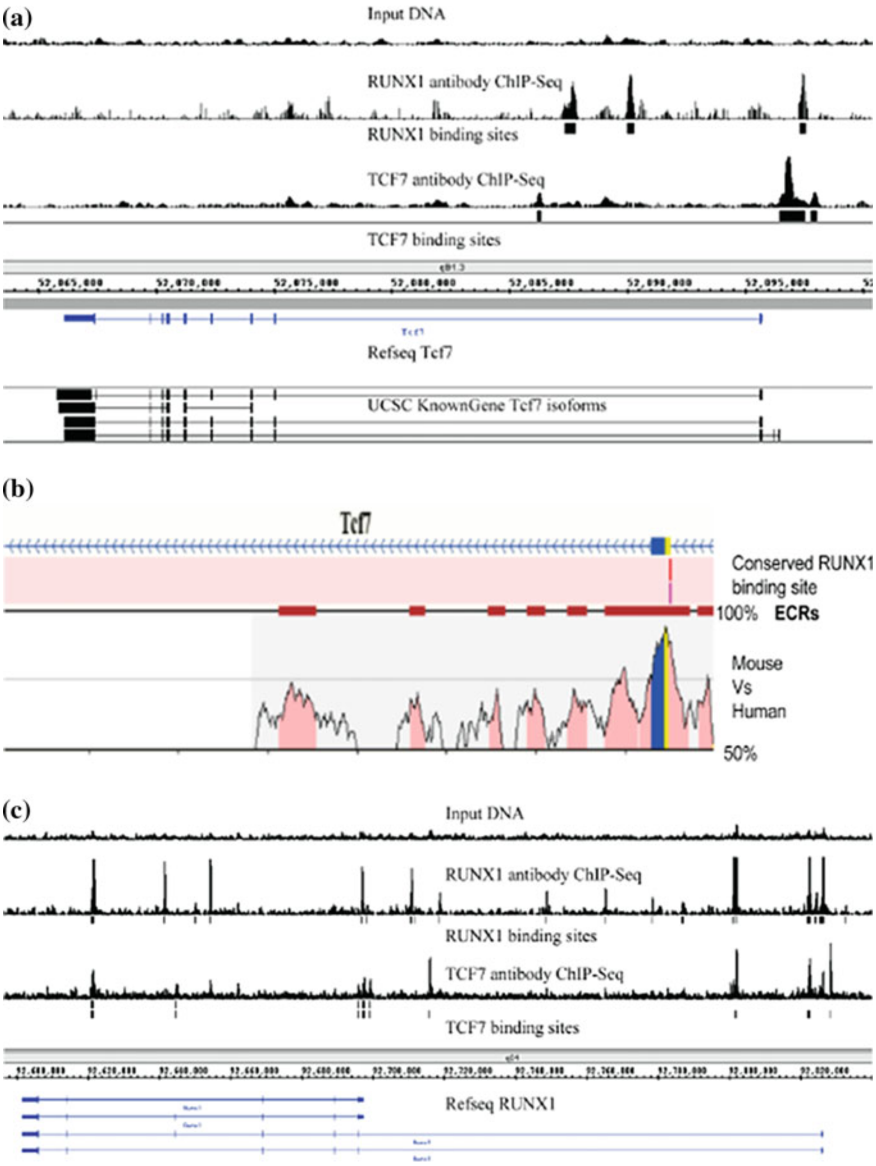


Fig. 1.3 Heat map of differentially expressed transcription factors (>1.5-fold) between Lin-CD34+ cells and Lin-CD34− cells. Two replicates were shown for each cell type. *Red color* represents upregulated genes and *green color* represents downregulated genes. Genes mentioned in the text are labeled. CD34 and Ly6a (Sca1) are cell surface markers. Adapted from Wu et al. [27]


(a)
Input DNA
RUNX1 antibody ChIP-Seq
RUNX1 binding sites
TCF7 antibody ChIP-Seq
TCF7 binding sites
qB1.3
52,065,000
52,070,000
52,075,000
52,080,000
52,085,000
52,090,000
52,095,000
Tcf7
Refseq Tcf7
UCSC KnownGene Tcf7 isoforms
(b)
Tcf7
Conserved RUNX1 binding site
100%
ECRs
Mouse Vs Human
50%
(c)
Input DNA
RUNX1 antibody ChIP-Seq
RUNX1 binding sites
TCF7 antibody ChIP-Seq
TCF7 binding sites
qC4
92,400,000
92,420,000
92,440,000
92,460,000
92,480,000
92,700,000
92,720,000
92,740,000
92,760,000
92,780,000
92,800,000
92,820,000
Refseq RUNX1
Runx1

◀ **Fig. 1.4** Identification of transcription factor-binding targets using ChIP sequencing. **a** Tcf7 is bound by both itself and by RUNX1 (AML1). Peaks indicate ChIP sequencing signal. Input genomic DNA serves as the negative control. The “binding sites” tracks (*black vertical bars*) show the transcription factor-binding loci determined using the PeakSeq program (normalized against genomic input DNA; q-value 0.001). Data are visualized in Integrated Genome Browser. **b** Identification of evolutionarily conserved RUNX1-binding sites at Tcf7 promoter region using REGULATORY VISTA. The graph shows conserved and aligned AML1/RUNX1 transcription factor-binding sites between mouse and human genomes using a matrix similarity score of 1 (the mt stringent). Two versions of the AML1-binding sites were found (AML1 and AML_Q6). The *ECRs: Evolutionarily conserved regions are indicated by deep *red blocks*. The degree of conservation (50–100 %) is indicated by the height of the peaks. Coding region is shown in *blue,* and UTR is shown in *yellow*. **c** Runx1 promoter is bound by both TCF7 and itself. Adapted from Wu et al. [27]

cells [27]. Therefore, we hypothesize that there are key regulators in transcriptional regulatory networks that determine the choice between EML cell self-renewal and differentiation, and TCF7 is one of the key transcription factors.

Subsequently, we identified in vivo binding targets of TCF7 using ChIP-Seq [27]. We found that TCF7 binds to its own promoter and the promoter of *Runx1 (Aml1),* a developmental determinant in hematopoietic cells that is best known for its critical role in haematological malignancies [147, 148] (Fig. 1.4). We showed that TCF7 and RUNX1 (AML1) bind to each other’s promoter regions, and a large number of common target genes are bound by RUNX1 and TCF7. TCF7 is necessary for the production of the short isoforms, but not the long isoforms of RUNX1, suggesting that TCF7 and the short isoforms of RUNX1 function coordinately in regulation. TCF7 knockdown experiments and Gene Set Enrichment Analyses suggest that TCF7 plays a dual role in promoting the expression of genes characteristic of self-renewing CD34+ cells while repressing genes activated in the partially differentiated CD34− state. Finally, through network analysis, we found that TCF7 and RUNX1 bind and regulate a network of upregulated transcription factors in the CD34+ cells which characterize the self-renewal property of the CD34+ cells, including Stat3, Sox4, F, Scl/Tal1, Etv6/Tel, Ppard, Smads, Cebpa, Gfi1, and Fli-1 (Fig. 1.5).

In summary, our results elucidated novel components and mechanisms that control the renewal and differentiation of hematopoietic precursor cells. The elucidation of the networks suggested potential master regulators that control early hematopoietic differentiation. Genetic manipulation of the master regulators may reveal how to induce hematopoietic precursor cell self-renewal in vitro or reprogram partially differentiated hematopoietic precursor cells back to a self-renewing state. Increasing the long-term ability of human hematopoietic precursor cells to reconstitute bone marrow is highly relevant for the therapy of leukemia and regenerative medicine.

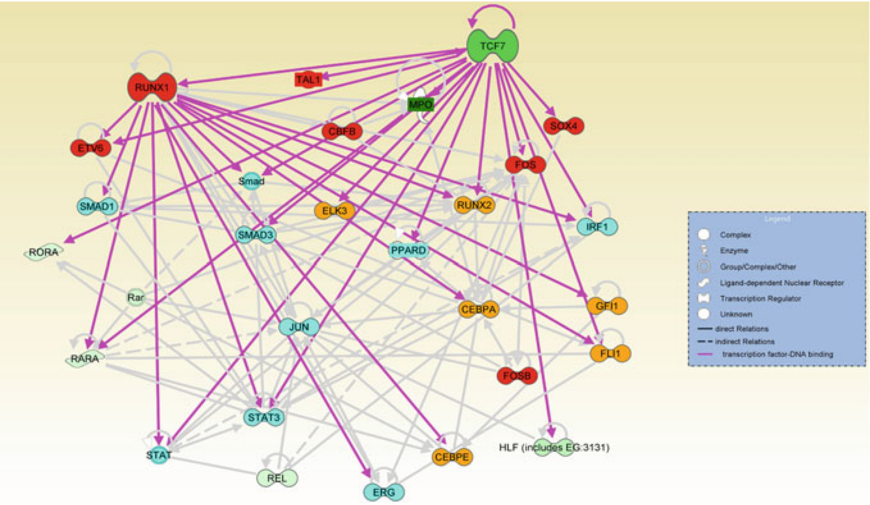


Fig. 1.5 Transcription factors TCF7 together with RUNX1 regulate a transcriptional regulatory network. The network involved in HSC establishment and development (*red nodes*), cell growth control (*blue nodes*), and multipotency (*orange nodes*) was identified among upregulated genes in CD34+ cells (twofold) and displayed by Ingenuity Pathway Analysis software (IPA). *Gray lines* are IPA-annotated relations based on the literature. *Pink lines* indicate TCF7 or RUNX1 binding to gene targets that were identified by our ChIP-Seq experiments. The shades of *green color* of the nodes in the network indicate the level of upregulation in CD34+ cells. Sox4, Mpo, Tal1, and Ppard were TCF7-binding targets that were added to the network manually because of their obvious interesting function in hematopoiesis and self-renewal. All other nodes were from default IPA analysis. Direct relations were indicated by *solid line* or *arrows*. Indirect relations were indicated by *dotted line*. Please see Ingenuity Pathway Analysis software (IPA, https://analysis.ingenuity.com/) Online Help section for detailed definitions. Adapted from Wu et al [27]

1.5 Future Prospective

The advancement of sequencing technology and computational analyses has greatly increased our knowledge of gene transcription and regulation. However, many challenges still remain. Difficulties in deciphering the anatomy of mammalian genes exist at multiple levels, including topics discussed in later chapters, such as complicated RNA, large amounts of intervening (noncoding) sequences, and the imperfection of computational algorithms. Additional issues include overlapping reading frames of protein-coding genes, antisense transcriptional units, the situation where the exon of one gene is encoded within the intron of another, and pseudogenes [149–151]. It will be impossible to find all genes and regulatory elements solely by analyzing genomic nucleotide sequences. Therefore, the eventual solution of annotation lies in large-scale systematic functional genomics experiments and conservation information from cross-genome comparisons [152].

Additionally, we should always take caution when interpreting data from a single kind of "omic" approach. For example, we cannot immediately conclude that a protein is expressed at a higher level from an upregulated signal by using microarray or RNA-Seq alone. Integrating data obtained from multiple distinct approaches will make conclusions more reliable. Theoretically, as multiple "omic" functional maps are overlaid, genes involved in the same process will cocluster in various maps. There are many challenges ahead in developing statistical and computational strategies for integrating these data, for improving annotation, and for making them available to the scientific community. The long-term goal is to understand the intricate and dynamic functional relationships between all components involved in particular biological processes as a whole, in order to be able to predict the potential behaviors of these systems in response to perturbations and thus be able to restore. This approach will provide answers for treating diseases.

Acknowledgments We thank Dr. Eva Zsigmond for reading and editing our manuscript. JQW, RCDD, and SM are supported by grants from the National Institutes of Health R01 NS088353 and R00 HL093213, the Staman Ogilvie Fund—Memorial Hermann Foundation, Mission Connect—a program of the TIRR Foundation, the Senator Lloyd & B.A. Bentsen Center for Stroke Research, UTHealth BRAIN Initiative and CTSA UL1 TR000371, and a grant from the University of Texas System Neuroscience and Neurotechnology Research Institute (Grant #362469).

References

1. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet. 2000;25(2): 239–40.
2. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252(5013):1651–6.
3. Wolfsberg TG, Landsman D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. 1997;25(8):1626–32.
4. Bailey LC Jr, Searls DB, Overton GC. Analysis of EST-driven gene annotation in human genomic sequence. Genome Res. 1998;8(4):362–76.
5. Das M, Burge CB, Park E, Colinas J, Pelletier J. Assessment of the total number of human transcription units. Genomics. 2001;77(1–2):71–8.
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270(5235):484–7.
7. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, et al. Using the transcriptome to annotate the genome. Nat Biotechnol. 2002;20(5):508–12.
8. Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, et al. 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci USA. 2004;101(32):11701–6 (Epub 2004/07/24).
9. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional annotation of a full-length mouse cDNA collection. Nature. 2001;409(6821):685–90.
10. Clark MD, Hennig S, Herwig R, Clifton SW, Marra MA, Lehrach H, et al. An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. Genome Res. 2001;11(9):1594–602.

11. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002;420(6915):563–73.
12. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, et al. A Drosophila full-length cDNA resource. Genome Biol. 2002;3(12):RESEARCH0080.
13. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, et al. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet. 2003;34(1):35–41.
14. Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, Ansorge W, et al. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Res. 2001;11(3):422–35.
15. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 2004;36(1):40–5.
16. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. 2004;14(10B):2121–7.
17. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci USA. 2002;99(26):16899–903.
18. Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, Robinson C, et al. The completion of the Mammalian Gene Collection (MGC). Genome Res. 2009;19(12):2324–33.
19. Bareyre FM, Schwab ME. Inflammation, degeneration and regeneration in the injured spinal cord: insights from DNA microarrays. Trends Neurosci. 2003;26:555–63.
20. Carmel JB, Galante a, Soteropoulos P, Tolias P, Recce M, Young W, et al. Gene expression profiling of acute spinal cord injury reveals spreading inflammatory signals and neuron loss. Physiol Genomics 2001;7:201–13.
21. Velardo MJ, Burger C, Williams PR, Baker HV, López MC, Mareci TH, et al. Patterns of gene expression reveal a temporally orchestrated wound healing response in the injured spinal cord. J Neurosci. 2004;24:8562–76.
22. Liu CL, Jin AM, Tong BH. Detection of gene expression pattern in the early stage after spinal cord injury by gene chip. Chin J Traumatol. 2003;6(1):18–22 (Epub 2003/01/25).
23. Tachibana T, Noguchi K, Ruda MA. Analysis of gene expression following spinal cord injury in rat using complementary DNA microarray. Neurosci Lett. 2002;327(2):133–7 (Epub 2002/07/06).
24. Jaerve A, Kruse F, Malik K, Hartung HP, Muller HW. Age-dependent modulation of cortical transcriptomes in spinal cord injury and repair. PLoS One. 2012;7(12):e49812 (Epub 2012/12/14).
25. Kahvejian A, Quackenbush J, Thompson JF. What would you do if you could sequence everything? Nat Biotechnol. 2008;26:1125–33.
26. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, Hutchison S, et al. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc Natl Acad Sci USA. 2010;107(11):5254–9 (Epub 2010/03/03).
27. Wu JQ, Seay M, Schulz, V., Hariharan, M., Tuck, D., Lian, J., Du, J., Shi, M., Ye, Z.J, Gerstein M, Snyder M, Weissman S. TCF7 is a key regulator of the self-renewal and differentiation switch in a multipotential hematopoietic cell line. PLoS Genet. 2012;8(3): e1002565 (Epub 2012).
28. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63 (Epub 2008/11/19).
29. Torres-Garcia W, Zheng S, Sivachenko A, Vegesna R, Wang Q, Yao R, et al. PRADA: pipeline for RNA sequencing data analysis. Bioinformatics. 2014;30(15):2224–6 (Epub 2014/04/04).
30. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78 (Epub 2012/03/03).

31. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, et al. MAP-RSeq: mayo analysis pipeline for RNA sequencing. BMC Bioinf. 2014;15:224 (Epub 2014/06/29).
32. Cumbie JS, Kimbrel JA, Di Y, Schafer DW, Wilhelm LJ, Fox SE, et al. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. PLoS ONE. 2011;6(10):e25279 (Epub 2011/10/15).
33. Fonseca NA, Marioni J, Brazma A. RNA-Seq gene profiling–a systematic empirical comparison. PLoS ONE. 2014;9(9):e107026 (Epub 2014/10/01).
34. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4 (Epub 2011/02/01).
35. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71 (Epub 2009/12/18).
36. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8(3):186–94 (Epub 1998/05/16).
37. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS ONE. 2013;8(12):e85024 (Epub 2014/01/01).
38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20 (Epub 2014/04/04).
39. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2.
40. Smeds L, Kunstner A. ConDeTri–a content dependent read trimmer for Illumina data. PLoS ONE. 2011;6(10):e26314 (Epub 2011/11/01).
41. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204–16 (Epub 2012/01/31).
42. Bohnert R, Ratsch G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. Nucleic Acids Res. 2010;38(Web Server issue):W348-51 (Epub 2010/06/17).
43. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. 2010;38(17):e170 (Epub 2010/07/31).
44. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12):e131 (Epub 2010/04/17).
45. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4:14 (Epub 2009/04/18).
46. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5 (Epub 2010/05/04).
47. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22 (Epub 2011/03/18).
48. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012;28(24):3169–77 (Epub 2012/10/13).
49. Updated listing of mappers. Available from: http://wwwdev.ebi.ac.uk/fg/hts_mappers/.
50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60 (Epub 2009/05/20).
51. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25 (Epub 2009/03/06).
52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9 (4):357–9 (Epub 2012/03/06).
53. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11 (Epub 2009/03/18).
54. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36 (Epub 2013/04/27).

55. Jean G, Kahles A, Sreedharan VT, De Bona F, Ratsch G. RNA-Seq read alignments with PALMapper. Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al]. 2010;Chapter 11:Unit 11 6 (Epub 2010/12/15).
56. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21 (Epub 2012/10/30).
57. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS ONE. 2009;4(11):e7767 (Epub 2009/11/13).
58. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. Genome Biol. 2009;10(9): R98 (Epub 2009/09/19).
59. Novocraft. 2010. Available from: http://www.novocraft.com/.
60. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical SHort read mapping. Bioinformatics. 2011;27(7):1011–2 (Epub 2011/02/01).
61. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7 (Epub 2009/06/06).
62. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7 (Epub 2009/09/01).
63. Clement NL, Clement MJ, Snell Q, Johnson WE. Parallel mapping approaches for GNUMAP. IPDPS. 2011;435–43 (Epub 2011/01/01).
64. Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, et al. Updates to the RMAP short-read mapping software. Bioinformatics. 2009;25(21):2841–2 (Epub 2009/09/09).
65. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8 (Epub 2008/08/21).
66. Maher MC, Hernandez RD. Rock, paper, scissors: harnessing complementarity in ortholog detection methods improves comparative genomic inference. G3 (Bethesda). 2015;5(4): 629–38 (Epub 2015/02/26).
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10 (Epub 1990/10/05).
68. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7 (Epub 1981/03/25).
69. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53 (Epub 1970/03/01).
70. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. 2009;5(5):e1000386 (Epub 2009/05/23).
71. Barsky M, Stege U, Thomo A, Upton C, editors. Suffix trees for very large genomic sequences. CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management; 2009; New York, NY, USA.
72. Ferragina P, Manzini G, editors. Opportunistic data structures with applications. Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000); 2000; Redondo Beach, CA.
73. Burrows M, Wheeler D. A block sorting lossless data compression algorithm. Palo Alto, CA: Digital Equipment Corporation; 1994.
74. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662-9 (Epub 2014/10/30).
75. iGenomes. Available from: https://support.illumina.com/sequencing/sequencing_software/igenome.html.
76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9 (Epub 2009/06/10).

77. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14 (2):178–92 (Epub 2012/04/21).
78. Fiume M, Smith EJ, Brook A, Strbenac D, Turner B, Mezlini AM, et al. Savant Genome Browser 2: visualization and analysis for population-scale genomics. Nucleic Acids Res. 2012;40(Web Server issue):W615-21 (Epub 2012/05/29).
79. Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 2009;25(20):2730–1 (Epub 2009/08/06).
80. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 2010;20(10):1432–40 (Epub 2010/08/10).
81. De Bruijn NG. A combinatorial problem. Koninklijke Nederlandse Akademie v Wetenschappen. 1946;46(6).
82. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9 (Epub 2008/03/20).
83. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909–12 (Epub 2010/10/12).
84. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23 (Epub 2009/03/03).
85. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52 (Epub 2011/05/17).
86. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92 (Epub 2012/03/01).
87. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18 (9):1509–17 (Epub 2008/06/14).
88. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8 (Epub 2008/06/03).
89. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25(8):1026–32 (Epub 2009/02/27).
90. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;26(4):493–500 (Epub 2009/12/22).
91. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010;7(12):1009–15 (Epub 2010/11/09).
92. Consortium TE. Standards, Guideline and Best Practices for RNA-Seq. 2011; V1.0. Available from: https://www.encodeproject.org/.
93. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106 (Epub 2010/10/29).
94. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40 (Epub 2009/11/17).
95. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47 (Epub 2015/01/22).
96. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57 (Epub 2009/01/10).
97. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27 (12):1739–40 (Epub 2011/05/07).

98. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422 (Epub 2010/08/12).
99. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947–56.
100. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 2001;409(6819):533–8.
101. Wu JQ, Seay M, Schulz, V., Hariharan, M., Tuck, D., Lian, J., Du, J., Shi, M., Ye, Z. J.,, Gerstein M, Snyder M, Weissman S. TCF7 is a key regulator of the self-renewal and differentiation switch in a multipotential hematopoietic cell line. PLoS Genetics. 2012;In Press.
102. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823–37 (Epub 2007/05/22).
103. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–502 (Epub 2007/06/02).
104. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80 (Epub 2009/09/09).
105. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457(7231):854–8 (Epub 2009/02/13).
106. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, et al. Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods. 2012;9(6):609–14 (Epub 2012/04/24).
107. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813–31 (Epub 2012/09/08).
108. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008;26(12):1351–9 (Epub 2008/11/26).
109. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24(5):713–4 (Epub 2008/01/30).
110. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS computational biology. 2013;9(11): e1003326 (Epub 2013/11/19).
111. Jung LY, Kharchenko P, Wold B, Sidow A, Batzoglou S, Park P. Assessment of ChIP-seq data quality using cross-correlation analysis.
112. Muino JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P. ChIP-seq analysis in R (CSAR): an R package for the statistical detection of protein-bound genomic regions. Plant Methods. 2011;7:11 (Epub 2011/05/11).
113. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137 (Epub 2008/09/19).
114. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 2008;26(11):1293–300 (Epub 2008/11/04).
115. Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, et al. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinf. 2010;11:369 (Epub 2010/07/06).
116. Bardet AF, He Q, Zeitlinger J, Stark A. A computational pipeline for comparative ChIP-seq analyses. Nat Protoc. 2012;7(1):45–61 (Epub 2011/12/20).
117. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005;23(1):137–44 (Epub 2005/01/08).
118. Ma W, Noble WS, Bailey TL. Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nat Protoc. 2014;9(6):1428–50 (Epub 2014/05/24).

119. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66–75.
120. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008;5 (9):829–34 (Epub 2009/01/23).
121. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008;24(15):1729–30 (Epub 2008/07/05).
122. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008;24(21):2537–8 (Epub 2008/09/12).
123. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
124. Bryder D, Rossi DJ, Weissman IL. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. Am J Pathol. 2006;169(2):338–46.
125. Shizuru JA, Negrin RS, Weissman IL. Hematopoietic stem and progenitor cells: clinical and preclinical regeneration of the hematolymphoid system. Annu Rev Med. 2005;56:509–38.
126. Faubert A, Lessard J, Sauvageau G. Are genetic determinants of asymmetric stem cell division active in hematopoietic stem cells? Oncogene. 2004;23(43):7247–55.
127. Zhou JX, Huang S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends Genet. 2011;27(2):55–62.
128. Waltzer L, Gobert V, Osman D, Haenlin M. Transcription factor interplay during Drosophila haematopoiesis. Int J Dev Biol. 2010;54(6–7):1107–15.
129. Bertrand V, Hobert O. Lineage programming: navigating through transient regulatory states via binary decisions. Curr Opin Genet Dev. 2010;20(4):362–8.
130. Jukam D, Desplan C. Binary fate decisions in differentiating neurons. Curr Opin Neurobiol. 2010;20(1):6–13.
131. Moore KA, Lemischka IR. “Tie-ing” down the hematopoietic niche. Cell. 2004;118 (2):139–40.
132. Tsai S, Bartelmez S, Sitnicka E, Collins S. Lymphohematopoietic progenitors immortalized by a retroviral vector harboring a dominant-negative retinoic acid receptor can recapitulate lymphoid, myeloid, and erythroid development. Genes Dev. 1994;8(23):2831–41.
133. Pinto do OP. Kolterud A, Carlsson L. Expression of the LIM-homeobox gene LH2 generates immortalized steel factor-dependent multipotent hematopoietic precursors. EMBO J. 1998;17 (19):5744–56.
134. Yu WM, Hawley TS, Hawley RG, Qu CK. Immortalization of yolk sac-derived precursor cells. Blood. 2002;100(10):3828–31.
135. Sauvageau G, Iscove NN, Humphries RK. In vitro and in vivo expansion of hematopoietic stem cells. Oncogene. 2004;23(43):7223–32.
136. Ye ZJ, Kluger Y, Lian Z, Weissman SM. Two types of precursor cells in a multipotential hematopoietic cell line. Proc Natl Acad Sci USA. 2005;102(51):18461–6.
137. Raich N, Clegg CH, Grofti J, Romeo PH, Stamatoyannopoulos G. GATA1 and YY1 are developmental repressors of the human epsilon-globin gene. EMBO J. 1995;14(4):801–9.
138. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. Genome Biol. 2003;4(3):R22.
139. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, et al. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. Genes Dev. 2002;16(23):3017–33.
140. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science. 2002;298(5594):799–804.

141. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431(7004):99–104.
142. Borneman AR, H. Yu, P. Bertone, M. Gerstein and M. Snyder. The transcription factors Mga1 and Phd1 are master regulators of a complex transcriptional network controlling pseudohyphal growth. Cell, submitted. 2005.
143. Weintraub H, Tapscott SJ, Davis RL, Thayer MJ, Adam MA, Lassar AB, et al. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. Proc Natl Acad Sci USA. 1989;86(14):5434–8.
144. Tapscott SJ. The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development. 2005;132(12):2685–95.
145. Asakura A, Lyons GE, Tapscott SJ. The regulation of MyoD gene expression: conserved elements mediate expression in embryonic axial muscle. Dev Biol. 1995;171(2):386–98.
146. Goldhamer DJ, Brunk BP, Faerman A, King A, Shani M, Emerson CP Jr. Embryonic activation of the myoD gene is regulated by a highly conserved distal control element. Development. 1995;121(3):637–49.
147. Kurokawa M, Hirai H. Role of AML1/Runx1 in the pathogenesis of hematological malignancies. Cancer Sci. 2003;94(10):841–6.
148. Friedman AD. Cell cycle and developmental control of hematopoiesis by Runx1. J Cell Physiol. 2009;219(3):520–4.
149. Coelho PS, Bryan AC, Kumar A, Shadel GS, Snyder M. A novel mitochondrial protein, Tar1p, is encoded on the antisense strand of the nuclear 25S rDNA. Genes Dev. 2002;16 (21):2755–60.
150. Tycowski KT, Shu MD, Steitz JA. A mammalian gene with introns instead of exons generating stable RNA products. Nature. 1996;379(6564):464–6.
151. Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res. 2002;12(10):1466–82.
152. Snyder M, Gerstein M. Genomics. Defining genes in the genomics era. Science. 2003;300 (5617):258–60.

Author Biographies

Dr. Raquel Cuevas Diaz Duran obtained her PhD degree in biotechnology from Tecnologico de Monterrey, Mexico, in 2014 working in the Cell Therapy Research Group. Her thesis involved the genomic characterization of adipose-derived stem cells and the identification of regulatory modules driving differentiation using time-series gene expression. In 2015, she joined Prof. Jiaqian Wu's laboratory in the Vivian L. Smith Department of Neurosurgery at the University of Texas as a postdoctoral research fellow. Her research is focused on the implementation of systems-based approaches for the integration of epigenetic, transcriptomic, and proteomic data obtained from time-series induction experiments of stem cells in order to identify molecular mechanisms regulating fate commitment and differentiation.

Dr. Sudheer Menon obtained his PhD degree in bioinformatics in the Department of Bioinformatics, Bharathiar University and DBEB, IIT Delhi, India, in 2010. His thesis was on "Computational identification of promoter regions in fungal genome." He worked as a postdoctoral fellow in Bioinformatics at Umea University, Sweden, and University of Witwatersrand, Johannesburg, in 2011–2012. Subsequently, he joined as a postdoctoral visiting fellow at National Institutes of Health, Bethesda, Maryland, in 2013 where his study mainly focused on "Characterization of transcriptional end sites and gene looping in mouse and human genome." In 2015, he worked in Prof. Jiaqian Wu's laboratory at the University of Texas as a postdoctoral research fellow. He has presented his research at several national and international conferences.

Prof. Jiaqian Wu An assistant professor in the Vivian L. Smith Department of Neurosurgery and Center for Stem Cell and Regenerative Medicine at the University of Texas Medical School at Houston, and Dr. Wu earned her doctorate in molecular and human genetics at Baylor College of Medicine and did her postdoctoral work at Yale and Stanford University. The Wu laboratory combines stem cell biology and systems-based approaches involving functional genomics, bioinformatics, and NGS technologies to unravel gene transcription and regulatory mechanisms governing neural and blood development and differentiation. Dr. Wu's work has been recognized with prestigious honors and awards, including the National Institute of Health Pathway to Independence (PI) Award (K99/R00), R01 and the Senator Lloyd & B.A. Bentsen Investigator Award which she currently holds; the National Institutes of Health Ruth L. Kirschstein National Research Service Award for Individual Postdoctoral Fellows; and the International Society for Stem Cell Research (ISSCR) Annual Meeting Travel Award. A reviewer for NIH, MRC, the journals Nucleic Acids Research, Genome Research, and Genome Biology, Dr. Wu has presented invited talks and lectures on stem cell biology, functional genomics, and proteomics at international conferences, the Multiple Sclerosis Research Center of New York, Lawrence Livermore National Laboratory, the University of Florida, etc. She has developed a patent, authored a book, and wrote many articles that have appeared in PNAS, Genome Biology, Plos Genetics, Genome Research, the Journal of Neuroscience, and Nature.