Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI)

Sebastian Schafer,[1,2] Kui Miao,[3] Craig C. Benson,[4] Matthias Heinig,[1,5,9] Stuart A. Cook,[2,3,6] and Norbert Hubner[1,7,8]

[1] Cardiovascular and Metabolic Sciences, Max-Delbrück-Center for Molecular Medicine, Berlin, Germany
[2] National Heart Center Singapore, Singapore
[3] Duke-National University of Singapore, Singapore
[4] Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts
[5] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany
[6] National Heart and Lung Institute, Imperial College London, London, United Kingdom
[7] German Center for Cardiovascular Research (partner site), Berlin, Germany
[8] Charité-Universitätsmedizin, Berlin, Germany
[9] Present address: Institute of Computational Biology, Helmholtz Zentrum München, Neuerberg, Germany

Thousands of alternative exons are spliced out of messenger RNA to increase protein diversity. High-throughput sequencing of short cDNA fragments (RNA-seq) generates a genome-wide snapshot of these post-transcriptional processes. RNA-seq reads yield insights into the regulation of alternative splicing by revealing the usage of known or unknown splice sites as well as the expression level of exons. Constitutive exons are never covered by split alignments, whereas alternative exonic parts are located within highly expressed splicing junctions. The ratio between reads including or excluding exons, also known as percent spliced in index (PSI), indicates how efficiently sequences of interest are spliced into transcripts. This protocol describes a method to calculate the PSI without prior knowledge of splicing patterns. It provides a quantitative, global assessment of exon usage that can be integrated with other tools that identify differential isoform processing. Novel, complex splicing events along a genetic locus can be visualized in an exon-centric manner and compared across conditions. © 2015 by John Wiley & Sons, Inc.

Keywords: alternative splicing • RNA-seq • percent spliced in • PSI • transcript processing • isoform expression

How to cite this article:
Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. 2015. Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). *Curr. Protoc. Hum. Genet.* 87:11.16.1-11.16.14.
doi: 10.1002/0471142905.hg1116s87

INTRODUCTION

Alternative splicing generates multiple gene products with diverse functions and increases protein diversity. Exon skipping and mutually exclusive splicing, as well as the usage of alternative 3′ and 5′ donor sites, are tightly regulated processes that result in a multitude of RNA isoforms transcribed from the identical genetic locus. High-throughput sequencing of cDNA fragments (RNA-seq; Lister et al., 2008) has become a popular tool to investigate

Current Protocols in Human Genetics 11.16.1-11.16.14, October 2015
Published online October 2015 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/0471142905.hg1116s87

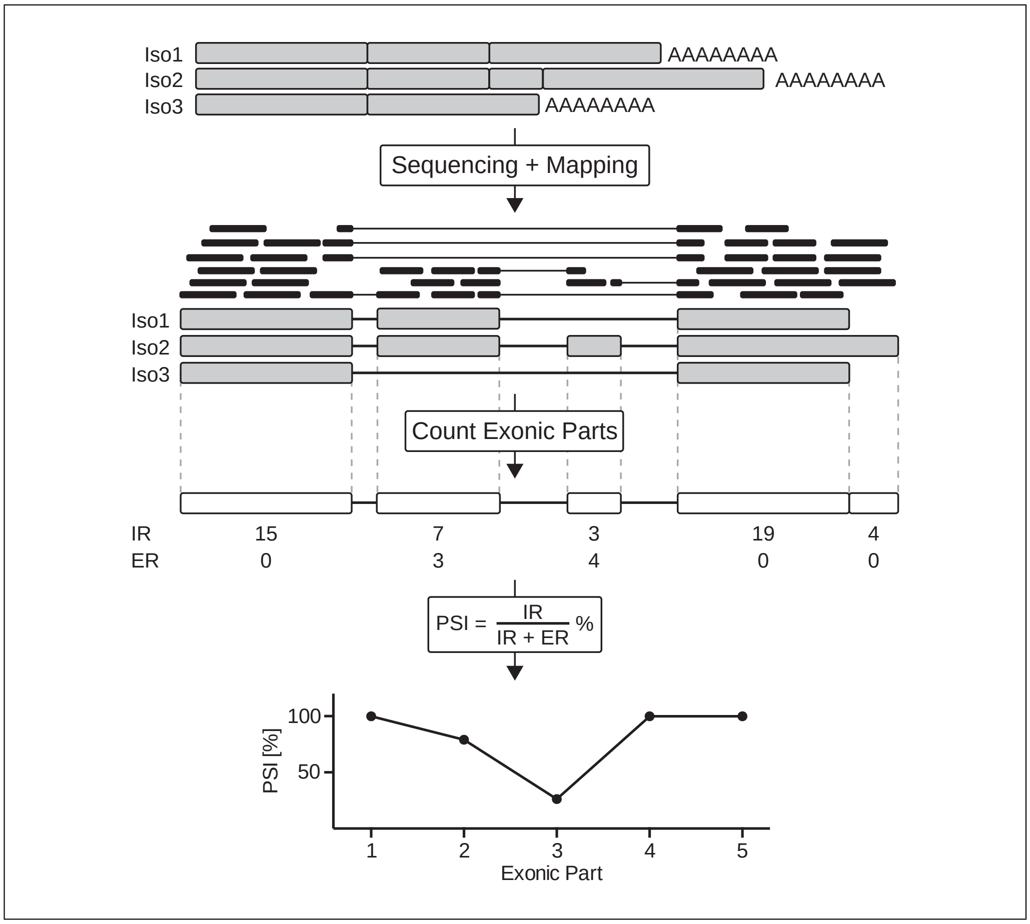

Figure 11.16.1 RNA-seq data reveals alternative splicing of exonic parts. RNA transcripts are fragmented and sequenced on a high-throughput sequencing platform. The reads must be aligned to the genome to reveal coverage of exons and splicing junctions. Split alignments across exonic parts indicate the removal of alternative exons from processed isoforms. The ratio of inclusion reads (i.e., overlapping reads; IR) and exclusion reads (i.e., overlapping split alignments; ER) indicates how efficient an exonic part is spliced into the isoform population.

RNA expression and post-transcriptional regulation. Sequencing of the transcriptome reveals tens of thousands of transcripts expressed simultaneously (Mortazavi et al., 2008). Despite providing a deep and genome-wide view on post-transcriptional processing, current technologies cannot reveal transcripts in full-length due to limitations in read length.

The percent spliced in index (PSI) indicates the efficiency of splicing a specific exon into the transcript population of a gene. The score summarizes alternative splicing events across individual exons without the need to know the underlying composition of full-length transcripts. A PSI of 100% indicates constitutive exons that are included in all transcripts and never removed from expressed isoforms. PSI values below 100% imply reduced inclusion of alternative exons and denote the percentage of isoforms that contain the exon compared to the total transcript population (Fig. 11.16.1). The splicing pattern observed along a genetic locus eventually results in the final isoform composition expressed in a sample.

The PSI is based on two read populations within an RNA-seq dataset (Wang et al., 2008): inclusion reads (IR) and exclusion reads (ER). IRs overlap with the exonic features and originate from transcripts that contain the exon of interest. These short sequence tags are evidence for the presence of the exon within the transcript population. ERs derive from isoforms that do not contain the exon. They are split alignments that map to

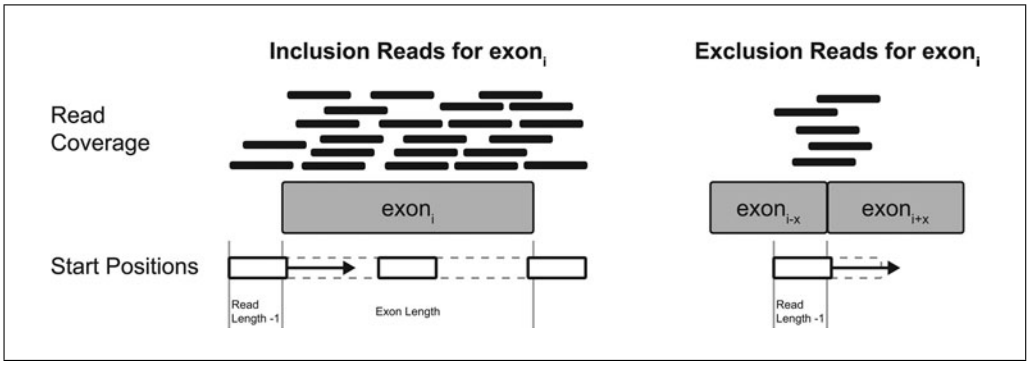

Figure 11.16.2 Length normalization required for PSI calculation. Inclusion reads (IR) can map to positions upstream of exon$_i$ (dependent on read length) and cover the whole body of exon$_i$. Exclusion reads (ER) for exon$_i$ overlap with splice sites between upstream and downstream exons and only intersect with a feature of 0 length. IR and ER counts must be normalized according to possible read start positions to calculate the PSI.

positions upstream of the exon as well as to positions downstream of the exon but never into the exon itself. These ERs do not have to map to adjacent exons or cover known splice junctions to be considered in this protocol. RNA-seq mapping strategies such as TopHat (Trapnell et al., 2009; Kim et al., 2013) or STAR (Dobin et al., 2013) detect split alignments across both known and unknown splice junctions.

The PSI is a ratio based on two read counts within the same dataset, and is thus independent of library size. PSI values can be compared between conditions without the need to adjust for sequencing depth. However, the read counts must be normalized for exon and read length to obtain meaningful results (Fig. 11.16.2). The ERs only cover a very small genomic feature, the splice site, compared to IRs, which cover the whole exon. Without normalization, the PSI of long exons would always approximate 100%.

We consider the possible number of start positions for each read population to normalize IRs and ERs. IRs can map upstream of the exon feature and cover the entire exon body. Thus, at equal expression levels, long exons have higher read counts than short exons. ERs cover splicing junctions of 0 base pair length. Their starting positions are only dependent on the read length. To compare IRs and ERs, both read counts must first be normalized:

$$\mathrm{IR}_{i,n} = \frac{\mathrm{IR}_i}{\text{length exon}_i + \text{read length} - 1}$$

$$\mathrm{ER}_{i,n} = \frac{\mathrm{ER}_i}{\text{read length} - 1}$$

where the subscript *i* is the exon number, and the subscript *n* indicates normalized read counts. The PSI$_i$ of exon$_i$ can then be calculated based on normalized counts as follows:

$$\mathrm{PSI}_i = \frac{\mathrm{IR}_{i,n}}{\mathrm{IR}_{i,n} + \mathrm{ER}_{i,n}}\%$$

Integration of PSI results with differential splicing analysis pipelines such as Cufflinks (Trapnell et al., 2010) or DEXSeq (Anders et al., 2012) will improve splicing predictions. Differences in PSI values (dPSI) between conditions can be used to filter differential exon or isoform expression (see Anticipated Results). The splicing index also allows visualization of complex isoform transitions across conditions (see Anticipated Results). While this method has been tested with RNA-seq data, it is also possible to use other RNA-based sequencing datasets to calculate PSI scores. The PSI based on ribosome

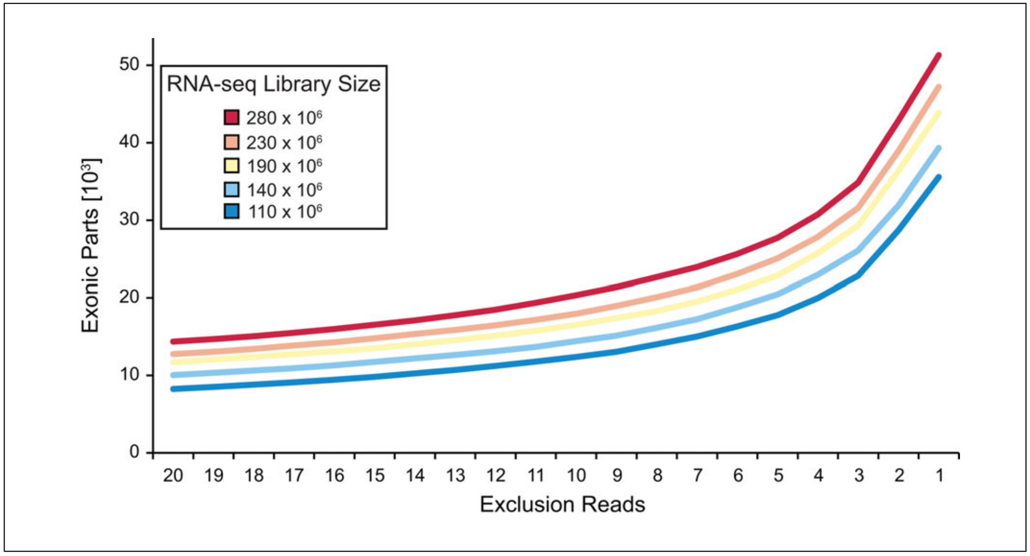

Figure 11.16.3 Cumulative sum of exonic parts with at least *x* exclusion reads in the human cardiac transcriptome (100 bp poly(A)+ RNA-seq). ER counts reveal thousands of alternative exonic parts in the cardiac transcriptome. The splicing of more than 20,000 alternative exonic parts can be investigated very accurately (>10 ER) when 280 million reads are generated. Smaller library sizes reduce the number of alternative exons that can be examined on the posttranscriptional level.

profiling data (Ribo-seq; Ingolia et al., 2009) reveals sequences that are removed from protein translation.

STRATEGIC PLANNING

To identify and quantify alternative splicing sufficiently using RNA-seq, methods for RNA extraction and library preparation must ensure sufficient read coverage across full-length mRNA transcripts. The quality of the RNA should have an RNA integrity number (RIN) of >8 according to the Agilent Bioanalyzer platform to guarantee that intact transcripts are present in the sample. To sequence fully processed transcripts, a poly(A) selection step should be part of the library preparation (Sultan et al., 2014).

The library size (number of reads per sample) is crucial for all RNA-seq-based analyses. The sequencing depth is especially important for the calculation of a splicing index (Fig. 11.16.3). ERs covering splice sites are direct evidence for the removal of an exon from a transcript and are critical for PSI calculation. However, splicing junctions are very small genomic features (0 base pair length) compared to full-length genes or exons and are thus not covered by a large fraction of reads.

Given the large dynamic range of transcript abundances, splicing analyses also depend heavily on the RNA expression level of the gene of interest. Alternative splicing can be estimated precisely in highly expressed genes even if the sequencing depth is rather low.

NOTE: For all protocols, command-line instructions begin with `$`, which denotes a bash shell prompt. Do not type the first `$` symbol for each command.

BASIC PROTOCOL

CALCULATION OF PERCENT SPLICED IN FOR ANNOTATED EXONS

Once RNA-seq data has been generated and mapped to the genome, the usage of exons can be estimated from the data by calculating the percent spliced in index (PSI). In the following steps, we quantify the number of reads supporting the inclusion or exclusion of any given exon in processed transcripts. The PSI is the ratio between both read populations. This index, ranging from 0% to 100%, is an estimation of the fraction of isoforms that include the exon. The protocol does not rely on previous knowledge of alternative splicing processes and does not require information regarding the expressed isoforms. This protocol details each step to calculate the PSI using the command line. The PSI can also be calculated with a shell script that combines all steps (see Alternate Protocol).

Materials

Hardware

Computer running Unix, Linux, or Mac OS X (see Critical Parameters for memory requirements)
Protocol testing environment: CentOS release 6.3; CPU architecture x86_64

Software

`dexseq_prepare_annotation.py` of the DEXSeq package available at *http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html*
HTSeq and Python 2.5 or later, not including Python 3, available at *http://www-huber.embl.de/HTSeq/* and *https://www.python.org/*
BEDTools toolset available at *https://github.com/arq5x/bedtools2*

Files

`accepted_hits.bam` (see Critical Parameters)
`junctions.bed` (see Critical Parameters)

1. Obtain a gene annotation file in the GTF format for the organism of interest by going to the homepage of the Ensembl database (Cunningham et al., 2015) and viewing all available gene sets (*http://www.ensembl.org/info/data/ftp/index.html*). The annotation must be based on the same genome version that was used to map the RNA-seq data. You can download the file directly from the homepage or type this on the command line to download the human gene set v78 based on GRCh38:

   ```
   $ wget ftp://ftp.ensembl.org/pub/release-78/gtf/
     homo_sapiens/Homo_sapiens.GRCh38.78.gtf.gz
   ```

   You can also download this file by accessing the Ensembl ftp site directly (ftp://ftp.ensembl.org/pub/release-78/gtf/homo_sapiens/).

   Annotation files are available in `.gtf.gz` *format to reduce file size and facilitate fast file transfers. You can browse previous Ensembl releases (ftp://ftp.ensembl.org/pub/) if you require a gene annotation file for older genome versions such as GRCh37. If you want to use a UCSC genome browser annotation, please refer to Support Protocol 1.*

2. Extract the GTF file:

   ```
   $ gunzip Homo_sapiens.GRCh38.78.gtf.gz
   ```

3. Create an exonic part annotation based on the Ensembl gene models. This command will create the file `GRCh38.78_reduce.gtf` that contains an exonic part annotation based on the gene models proposed by `Homo_sapiens.GRCh38.78.gtf`.

   ```
   $ python dexseq_prepare_annotation.py Homo_sapiens.
     GRCh38.78.gtf GRCh38.78_reduce.gtf
   ```

   Download the DEXseq suite (Anders et al., 2012) and copy the script to the home directory by typing `cp ./DEXSeq/inst/python_scripts/dexseq_prepare_annotation.py ~/`*. The script* `dexseq_prepare_annotation.py` *is part of the DEXSeq package (Anders et al., 2012). It creates a model transcript for each gene that combines all known exons (or parts of exons) into one artificial transcript (see Background Information).*

4. Extract all exonic parts from the annotation and assign a unique identifier:

```
$ awk '{OFS="\t"}{if ($3 == "exonic_part") print
  $1,$2,$3,$4,$5,$6,$7,$8,$14":"$12}' GRCh38.78_
  reduce.gtf | sed 's=[";]==g' > GRCh38.78_exonic_
  parts.gff
$ rm GRCh38.78_reduce.gtf
```

The annotation file GRCh38.78_reduce.gtf *is not required for subsequent steps of this protocol and can be deleted. However, it contains valuable information such as the genomic location of exonic parts that can be useful for downstream analyses beyond this protocol. One unique identifier is assigned to each exonic part in the format* GeneID:ExonicPart#*. This information can be used to identify and sort the exonic parts and later also to combine various sample files. If the annotation is not created by the* dexseq_prepare_annotation.py *script,* $14 *and* $12 *might have to be altered to point to fields that contain unique gene and exon identifiers.*

5. Determine the IR count for each exonic part. Count the number of reads in accepted_hits.bam that overlap with features in GRCh38.78.exonic_parts to create a file that reports the location, length, IR count, and identification for each exonic part.

```
$ coverageBed -split -abam accepted_hits.bam -b
  GRCh38.78_exonic_parts.gff | awk 'BEGIN{OFS="\t"}
  {print $1,$4,$5,$5-$4+1,$9,$10}' | sort -k 5 >
  exonic_parts.inclusion
```

A protocol for basic usage and installation of BEDTools is available (see installation option 1; Quinlan, 2014). Add the binary files to the system path by typing: export PATH=/your_path_to_bedtools/bin/:$PATH.

The coverageBed *command of the BEDTools suite counts the number of alignments overlapping with exonic parts. Split alignments are considered individually. If a read, or a fraction of a read, overlaps with an exonic part, it is considered an indication for the inclusion of an exon.*

6. Filter junctions based on the annotation of exons.

```
$ sed 's/,/\t/g' junctions.bed | awk 'BEGIN{OFS="\t"}
  {print $1,$2,$2+$13,$4,$5,$6}' > left.bed
$ sed 's/,/\t/g' junctions.bed | awk 'BEGIN{OFS="\t"}
  {print $1,$3-$14,$3,$4,$5,$6}' > right.bed
$ intersectBed -u -s -a left.bed -b GRCh38.78_
  exonic_parts.gff > left.overlap
$ intersectBed -u -s -a right.bed -b GRCh38.78_
  exonic_parts.gff > right.overlap
$ cat left.overlap right.overlap | cut -f4 | sort |
  uniq -c | awk '{ if($1 == 2) print $2 }' >
  filtered_junctions.txt
$ grep -F -f filtered_junctions.txt junctions.bed >
  filtered_junctions.bed
$ rm left.bed right.bed left.overlap right.overlap
  filtered_junctions.txt
```

TopHat does not only report read alignments but also detects, quantifies, and reports splicing junctions present in the RNA-seq data in the file junctions.bed*. This filtering step ensures that only splicing between known exons is considered for further analyses and reduces the effect of misalignments on the calculation of PSI indices. This is a very conservative approach that can be skipped to consider all split alignments detected in the RNA-seq data.*

Filtering the junctions based on the location of known exons can be too stringent in cases where the genome annotation is incomplete. To skip filtering, use the file `junctions.bed` *instead of* `filtered_junctions.bed` *in step 7. The filtering of junctions can be particularly slow on Mac OS X systems after version 10.7 (Lion). Install GNU grep (http://rudix.org/packages/grep.html) to improve running time.*

7. Define intronic locations within the junction file. Transform the junction file to determine the genomic location of introns covered by exclusion reads. It is the inner part of the splicing junctions that corresponds to the sequence that is removed from processed transcripts.

```
$ sed 's/,/\t/g' filtered_junctions.bed | grep -v
  description | awk '{OFS="\t"}{print $1,$2+$13,
  $3-$14,$4,$5,$6}' > intron.bed
$ rm filtered_junctions.bed
```

The inner segment of each junction designates the genomic region that is spliced out of a transcript. Each split alignment indicates an intron (i.e., the absence of transcribed sequence within an RNA molecule). These introns, due to the presence of conserved sequences at splice sites, are assigned to a strand even though the RNA-seq data itself might not be stranded. Retaining the strand information ensures that splicing of antisense RNA or overlapping genes does not interfere with the calculation of the PSI.

8. To count the ER, intersect introns with exonic parts, and count the total number of split alignments overlapping each exonic part. Finally, sort the results according to the unique exonic part identification, and save a file reporting the genome-wide exclusion information.

```
$ intersectBed -wao -f 1.0 -s -a GRCh38.78_exonic_
  parts.gff -b intron.bed | awk 'BEGIN{OFS="\t"}{$16
  == 0? s[$9] += 0:s[$9] += $14}END{for (i in s)
  {print i,s[i]}}' | sort -k 1 > exonic_parts.
  exclusion
$ rm intron.bed
```

The exonic part must be located completely within the intron on the same strand to be considered spliced out.

9. Calculate PSI values and generate a file that contains the unique exonic part identifier, the exon length, and number of inclusion and exclusion reads as well as the PSI. A header line denotes each column of the results written to `exonic_parts.psi`. The read length of the RNA-seq dataset must be supplied to the pipeline via `readLength` to accurately normalize the read counts.

```
$ readLength=100
$ paste exonic_parts.inclusion exonic_parts.exclusion
  | awk -v "len=$readLength" 'BEGIN{OFS="\t"; print
  "exon_ID" , "length" , "inclusion" , "exclusion" ,
  "PSI"}{NIR=$6/($4+len-1) ; NER=$8/(len-1)}{print
  $5,$4,$6,$8,(NIR+NER<=0)? "NA":NIR / (NIR + NER)}'
  > exonic_parts.psi
$ rm exonic_parts.inclusion
$ rm exonic_parts.exclusion
```

The inclusion and exclusion reads must be normalized first (`NIR` *and* `NER`*) before the PSI can be calculated. We normalize the read counts by the possible number of starting positions for each read population across the genomic feature (see also Fig. 11.16.2). If an exonic part is supported neither by inclusion nor exclusion reads, the PSI is set to* `NA`*. It is important to consider the number of reads the PSI value is based on. The* *sum of inclusion and exclusion reads for an exonic part can be used to filter the results. To ensure correct results, each line in* `exonic_parts.inclusion` *and* `exonic_parts.exclusion` *must correspond to the identical exonic part.*

ALTERNATE PROTOCOL

CALCULATION OF PERCENT SPLICED IN FOR ANNOTATED EXONS WITH THE SCRIPT `PSI.sh`

This protocol performs the same task as the Basic Protocol, using a shell script. This script can execute individual steps of the protocol as well as the entire protocol.

Materials

Hardware

Computer running Unix, Linux, or Mac OS X (see Critical Parameters for memory requirements)
Protocol testing environment (i.e., CentOS release 6.3; CPU architecture x86_64)

Software

BEDTools toolset available at *https://github.com/arq5x/bedtools2*

Files

Script `PSI.sh` (see *http://www.currentprotocols.com/protocol/hg1116*)
`accepted_hits.bam` (see Critical Parameters)
`junctions.bed` (see Critical Parameters)
Exonic parts annotation (e.g., `GRCh38.78_reduce.gtf`; see Basic Protocol, step 3)

1. Download the script `PSI.sh` to the home folder.

2. Set the executable bit on `PSI.sh` on the command line:

```
$ chmod +x ./PSI.sh
```

3. Calculate the PSI index by executing the script:

```
$ ./PSI.sh StartPSI GRCh38.78_reduce.gtf 100
  accepted_hits.bam junctions.bed sample1
```

The first argument specifies the mode of PSI calculation; `StartPSI` *will start the whole pipeline and calculate the PSI index from mapped RNA-seq data. The second argument must indicate the annotation of exonic parts. The third argument specifies the read length of the RNA-seq library. The fourth and fifth arguments refer to the mapped reads in BAM format and the TopHat junction file, respectively. The sixth argument is a sample prefix that can be chosen freely to suit the project. Generated files will be named after this prefix. A unique prefix for each sample in the experiment is required. If the mode* `StartPSINoFilter` *is specified as first argument, the junctions will not be filtered.*

This command will create several output files. The files `sample1_exonic_parts.inclusion` *and* `sample1_exonic_parts.exclusion` *are intermediate files that can be deleted if the analysis is completed. The PSI values are reported in the file* `sample1_exonic_parts.psi`.

4. Individual steps of the analysis can be repeated. Running the script with the argument `getPSI` and supplying the files containing IR (`sample1_exonic_parts.inclusion`) and ER (`sample1_exonic_parts.exclusion`) data will recalculate the PSI index and save the results to a new files based on the prefix chosen (`sample1`).

```
$ ./PSI.sh getPSI 100 sample1_exonic_parts.inclusion
  sample1_exonic_parts.exclusion sample1
```

This recalculation of the PSI index will be much faster than running the entire pipeline, since inclusion and exclusion reads do not have to be counted again.

5. It is also possible to only count inclusion or exclusion reads with the script if needed. This will create the files `sample1_exonic_parts.inclusion` and `sample1_exonic_parts.exclusion`, respectively.

```
$ ./PSI.sh CountInclusion GRCh38.78_reduce.gtf
  accepted_hits.bam sample1
$ ./PSI.sh CountExclusion GRCh38.78_reduce.gtf
  junctions.bed sample1
```

Starting the script with `./PSI.sh StartPSI` *will run all three steps sequentially:* `CountInclusion`*,* `CountExclusion`*, and finally* `getPSI`*. The most computing intense step is* `CountInclusion` *due to the intersection of reads from the file* `accepted_hits.bam` *with the exonic part annotation. The number of reads in the library of the RNA-seq experiment will thus determine the running time of the script.*

SUPPORT PROTOCOL 1

CREATE AN EXONIC PART ANNOTATION BASED ON UCSC GENE ANNOTATION FILES

We create an exonic part annotation based on Ensembl gene models in the Basic Protocol, step 3, using the script `dexseq_prepare_annotation.py`. Annotations supplied by the UCSC genome browser (Kent et al., 2002) do not strictly adhere to the GTF specifications. To generate an exonic matrix annotation based on the UCSC gene predictions, use the customized `dexseq_prepare_annotation_UCSC.py` script.

Materials

Hardware

Computer running Unix, Linux, or Mac OS X

Software

`dexseq_prepare_annotation_UCSC.py` script (see *http://www.currentprotocols.com/protocol/hg1116*) based on `dexseq_prepare_annotation.py` of the DEXSeq package

HTSeq and Python 2.5 or later, not including Python 3, available at *http://www-huber.embl.de/HTSeq/* and *https://www.python.org/*

Files

Gene annotation file for the organism of interest (see step 1)

1. Obtain a gene annotation file for the organism of interest.

   a. Go to the homepage of the UCSC genome browser (Kent et al., 2002) at *http://genome.ucsc.edu/cgi-bin/hgTables* to view all available gene sets.
   b. Select Genes and Gene Prediction, Refseq, or UCSC genes for the organism of interest. The annotation must be based on the same assembly version that was used to map the RNA-seq data.
   c. Choose GTF as output format, and download the gene annotation for the entire genome.

2. Download the script `dexseq_prepare_annotation_UCSC.py` to the home folder.

3. Create an exonic part annotation based on the UCSC-derived gene models. This command will create the file `UCSC_gene_models_reduce.gtf` that

contains an exonic part annotation based on the gene models proposed by `UCSC_gene_models.gtf`.

```
$ python dexseq_prepare_annotation_UCSC.py UCSC_gene_
  models.gtf UCSC_gene_models_reduce.gtf
```

Use the file `UCSC_gene_models_reduce.gtf` *instead of* `GRCh38.78_reduce.gtf` *in Basic Protocol, step 4, to run the PSI pipeline based on UCSC-derived gene models.*

SUPPORT PROTOCOL 2

REFORMAT STAR ALIGNER OUTPUT DATA TO CALCULATE THE PSI

The STAR aligner (Dobin et al., 2013) is an alternative to TopHat for mapping RNA-seq data to the genome and detecting known and novel splicing junctions. However, the STAR aligner does not report junctions in a format that is supported by either the Basic Protocol or the Alternate Protocol. Reformat the junctions files created by STAR as follows for use with the protocols in this unit.

Materials

Hardware

Computer running Unix, Linux, or Mac OS X

Software

Samtools available at *http://www.htslib.org/*

Files

`SJ.out.tab`
`aligned.out.sam`

1. The STAR aligner reports split alignments in a file called `SJ.out.tab`. Run the following command to transform this file to a format resembling the output of TopHat:

```
$ awk 'BEGIN{OFS="\t"}{print $1, $2-20-1, $3+20,
  "JUNCBJ"NR, $7, ($4 == 1)? "+":"-",$2-20-1, $3+20,
  "255,0,0", 2, "20,20", "0,300" }' SJ.out.tab >
  junctions.bed
```

The file `SJ.out.tab` *is a junctions file created by STAR aligner (Dobin et al., 2013).*

The reformatted junctions are saved in the file `junctions.bed`*. Use this file in Basic Protocol, step 6, to calculate PSI values based on STAR alignments. This command will create a junction file containing split alignments supported by uniquely aligning reads. To consider multi-mapping reads only, select field* `$8` *instead of* `$7`*. To count both read populations choose* `$7+$8`*.*

2. Convert the read alignments from SAM to BAM format:

```
$ samtools view --hb aligned.out.sam > accepted_hits.
  bam
```

Transforming the STAR output from SAM to BAM will reduce file size and allow for processing the alignment files with BEDTools in the Basic Protocol, step 5.

COMMENTARY

Background Information

The calculation of percent spliced in index (PSI) scores in this protocol is based on an exonic part annotation that can be generated from any gene annotation. One artificial model transcript combines all known exons, or parts of exons, that are known for each gene (Fig. 11.16.4). By quantifying each exonic part individually using the PSI methodology, all possible splicing processes that result

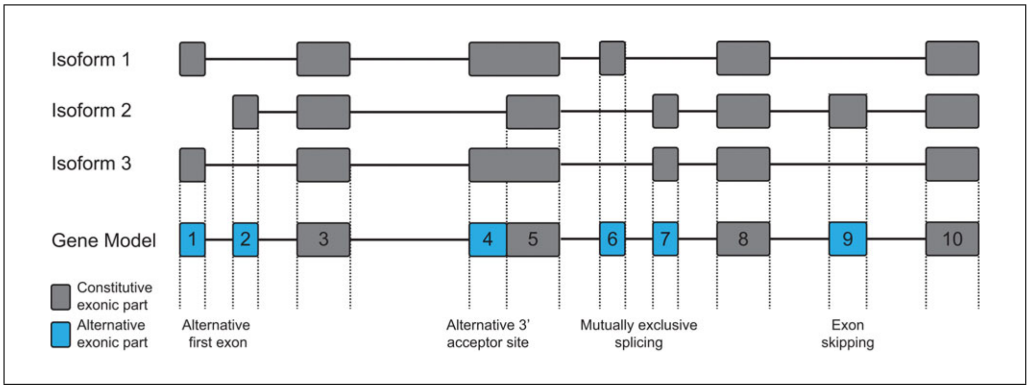

Figure 11.16.4 Isoform population represented as artificial transcript of exonic parts. Knowledge of the PSI for all exonic parts along the artificial transcript yields insights into the splicing processes. The PSI of exonic part 2 indicates the usage of the alternative first exon between both transcript populations. A decrease in PSI can be explained by an upregulation of isoforms 1 and 3 compared to isoform 2. The PSI value of exonic part 4 indicates the efficiency of the 3′ acceptor site of the second intron. Higher inclusion of this exonic part indicates an upregulation of the long form of exon 3 within the transcript population. Exonic parts 6 and 7 represent mutually exclusive exons. Isoforms can contain either exon, but not both together. Changes in the PSI of exonic part 6 between conditions should therefore always be in the opposite direction to exonic part 7.

in a complex variety of isoforms can be discussed referring to only one model transcript per gene.

Knowledge of the exact splicing processes is not required to calculate the PSI. All reads indicating splicing between known exons are counted and contribute to exclusion counts. The PSI score summarizes all splicing events that result in the removal of an exonic part. Previously considered constitutive exons will be revealed as alternative exons if their PSI is below 100%.

Critical Parameters

Computer memory requirements are dependent on the library size of the sequencing experiment as well as the genome annotation. A 2 Gb BAM alignment file in combination with the rn5 Ensembl (Cunningham et al., 2015) gene annotation requires 1 Gb of RAM. Deeper datasets will require more computational resources. Large alignment files can be split by chromosomes to lower the requirements. This does not affect the calculation of the PSI score.

The protocols presented in this unit were tested using BEDTools version 2.23 (Quinlan and Hall, 2010) available at *https://github.com/arq5x/bedtools2*. A protocol for basic usage and installation of BEDTools is available (see installation option 1; Quinlan, 2014). Add the binary files to the system path by typing `export PATH=/your_path_to_bedtools/bin/:$PATH`.

The `accepted_hits.bam` file contains RNA-seq alignments reported by either TopHat (Trapnell et al., 2009) or TopHat2 (Kim et al., 2013). Since these alignments are the basis for all subsequent calculations, the mapping parameters should be considered carefully beforehand. Setting a threshold for maximum intron (splice junction) length and filtering for uniquely mapping reads can improve the results. Detailed protocols for the alignment procedures are available (Trapnell et al., 2012). To use output created by the STAR aligner (Dobin et al., 2013) to calculate the PSI, please refer to Support Protocol 2. The `junctions.bed` file contains splice junctions covered by RNA-seq alignments reported by TopHat (Trapnell et al., 2009) or TopHat2 (Kim et al., 2013).

The initial mapping of split reads is a crucial step toward obtaining PSI values. It is recommended to map all reads against the genome, transcriptome, and novel splice junctions to identify reads mapping uniquely to only one position in the genome.

The minimum number of reads to calculate a PSI value is also an important parameter and should be at least ten. This will limit the view on splicing of lowly expressed genes, but also filter out noisy splicing and misalignments in the data.

Troubleshooting

The IR and ER files (`exonic_parts.inclusion` and `exonic_parts.exclusion`, respectively) must contain data of the identical exonic parts in the same

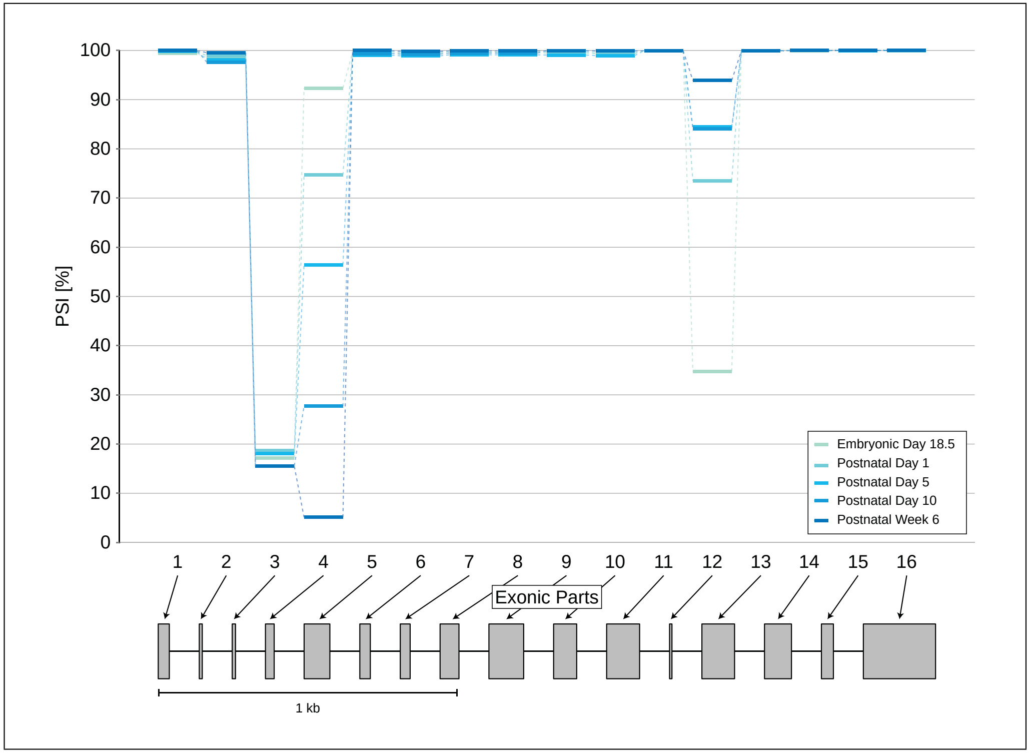

Figure 11.16.5 *Cardiac Troponin T* isoform transition in cardiac development. The hearts of embryonic (day 18.5), 1-day, 5-day, 10-day, and adult rats were sequenced on the RNA level. DEXSeq analyses identified exons 4 and 12 of *Cardiac Troponin T* to be differentially expressed (FDR ≤0.05) between stages of cardiac development. These isoform transitions are also reflected on the PSI level (average of 3 biological replicates for each time point). The third exonic part is alternatively spliced and present in only ~15% of transcripts. It is not differentially spliced across development; the PSI remains constant across time points. Exon 4 is present in the majority of embryonic isoforms but gradually removed across development and rarely spliced into adult *Cardiac Troponin T* transcripts. The opposite is true for exon 12; it is only present in a small fraction of transcripts in embryonic heart muscle but spliced into most adult isoforms of *Cardiac Troponin T* (unpublished data, together with Martin Liss and Michael Gotthardt).

order to calculate the PSI. The shell script introduced in Alternate Protocol tests for this requirement and will exit stating `Unsorted exonID exit` if this criterion is not met. If the same format that is created by `dexseq_prepare_annotation.py` for the input annotation is not used, the script may have to be altered to generate unique exonic part identifiers (also Basic Protocol, step 4).

If the PSI of all exonic parts is `NA`, the reads and the exonic part annotation are likely not based on the same nomenclature or genome version. Reads might be mapped to contigs called "chr1," whereas exonic parts could be located on chromosome "1." The location of reads, junctions, and exonic parts must be compatible to determine and count possible overlaps.

Anticipated Results

Popular methods to detect differential splicing include the tool cuffdiff from the Cufflinks suite (Trapnell et al., 2010). The cuffdiff tool assigns reads to transcripts and then performs differential expression analysis. This approach reveals isoform expression differences due to differential splicing but not the exons that undergo splicing transitions directly. It is also heavily dependent on the knowledge, or correct prediction, of isoforms. DEXSeq is another powerful, exon-centered approach to detect differential expression of exons due to posttranscriptional regulation (Anders et al., 2012). Unlike cuffdiff, it is not dependent on the previous knowledge or prediction of isoforms. However, it relies solely on exon coverage and does not take exon exclusion information into account.

Genome-wide PSI values quantify complex splicing processes and can be compared across conditions to complement methods that test for differential splicing. Combining PSI data with previously mentioned approaches can help to identify differentially spliced exons. The PSI index integrates both inclusion and exclusion information for each exon, thus showing direct evidence of alternative splicing processes in the sequencing data. The presence of exclusion reads is a qualitative indication for alternatively spliced exons. A difference in the PSI value (dPSI), in addition to differential exon or isoform expression, is an additional indicator for differential splicing between conditions. The dPSI value can therefore be used to filter significant exon or isoform expression results similar to a fold-change cutoff for gene expression analyses.

The PSI scores of genes across conditions can also be used to visualize complex splicing transitions (Fig. 11.16.5). Alternative exons can be identified, and the directionality and extent of splicing can be plotted.

The exon-centric view of splicing based on PSI values explores complex transcript populations with unknown composition. *Titin* is the largest gene expressed in humans with more than 360 exons, and the exact isoform composition remains unknown. This protocol can identify splicing transitions of *Titin* exons that underlie cardiac disease (Guo et al., 2012). The knowledge of exons that are used less (low PSI) in cardiac transcripts of *Titin* has shed light on the pathogenicity of rare variants (Roberts et al., 2015). The genome-wide integration of PSI values with protein-RNA interaction datasets of the splicing factor *Rbm20* has also elucidated the mechanisms underlying the regulation of alternative splicing of *Titin* and other *Rbm20* target genes (Maatz et al., 2014).

Time Considerations

The library size of the RNA-seq dataset determines the duration for calculating the PSI scores. The calculation of the inclusion reads can become a lengthy process and is dependent on the size of the files `accepted_hits.bam` and the annotation file. It can take up to 2 hr to analyze a 2 Gb BAM alignment file in combination with the most recent rn5 Ensembl (Cunningham et al., 2015) gene annotation. Deeper datasets will require a longer running time. The data can be split according to chromosomes to calculate PSI scores in parallel and reduce overall running time. The pipeline requires some software tools to be installed on the system (Python and BEDTools), and the setup might take up to 1 to 2 hr.

Acknowledgments

This work was supported by funding from the European Union EURATRANS award (HEALTH-F4-2010-241504 to N.H. and S.A.C.), the Helmholtz Alliance ICEMED, and the Deutsche Forschungsgemeinschaft (Forschergruppe 1054, HU 1522/1-1) to N.H. S.A.C was supported by the NIHR CV BRU of Royal Brompton and Harefield NHS Foundation Trust.

Literature Cited

Anders, S., Reyes, A., and Huber, W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res.* 22:2008-2017. doi: 10.1101/gr.133744.111.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., García, G.C., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., and Flicek, P. 2015. Ensembl 2015. *Nucl. Acids Res.* 43:D662-D669. doi: 10.1093/nar/gku1010.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21. doi: 10.1093/bioinformatics/bts635.

Guo, W., Schafer, S., Greaser, M.L., Radke, M.H., Liss, M., Govindarajan, T., Maatz, H., Schulz, H., Li, S., Parrish, A.M., Dauksaite, V., Vakeel, P., Klaassen, S., Gerull, B., Thierfelder, L., Regitz-Zagrosek, V., Hacker, T.A., Saupe, K.W., Dec, G.W., Ellinor, P.T., MacRae, C.A., Spallek, B., Fischer, R., Perrot, A., Özcelik, C., Saar, K., Hubner, N., and Gotthardt, M. 2012. RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat. Med.* 18:766-773. doi: 10.1038/nm.2693.

Ingolia, N., Ghaemmaghami, S., Newman, J., and Weissman, J. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218-223. doi: 10.1126/science.1168978.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* 12:996-1006. doi: 10.1101/gr.229102.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079. doi: 10.1093/bioinformatics/btp352.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133:523-536. doi: 10.1016/j.cell.2008.03.029.

Maatz, H., Jens, M., Liss, M., Schafer, S., Heinig, M., Kirchner, M., Adami, E., Rintisch, C., Dauksaite, V., Radke, M.H., Selbach, M., Barton, P.J., Cook, S.A., Rajewsky, N., Gotthardt, M., Landthaler, M., and Hubner, N. 2014. RNA-binding protein RBM20 represses splicing to orchestrate cardiac pre-mRNA processing. *J. Clin. Investig.* 124:3419-3430. doi: 10.1172/JCI74523.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621-628. doi: 10.1038/nmeth.1226.

Quinlan, A.R. 2014. BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinform.* 47:11.12.1-11.12.34.

Quinlan, A.R. and Hall, I.M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842. doi: 10.1093/bioinformatics/btq033.

Roberts, A., Ware, J., Herman, D., Schafer, S., Baksi, J., Bick, A., Buchan, R., Walsh, R., John, S., Wilkinson, S., Mazzarotto, F., Felkin, L.E., Gong, S., MacArthur, J.A., Cunningham, F., Flannick, J., Gabriel, S.B., Altshuler, D.M., Macdonald, P.S., Heinig, M., Keogh, A.M., Hayward, C.S., Banner, N.R., Pennell, D.J., O'Regan, D.P., San, T.R., de Marvao, A., Dawes, T.J., Gulati, A., Birks, E.J., Yacoub M.H., Radke M., Gotthardt M., Wilson J.G., O'Donnell C.J., Prasad S.K., Barton P.J., Fatkin D., Hubner N., Seidman J.G., Seidman C.E., and Cook S.A. 2015. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Science Transl. Med.* 7:270ra6.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., and Yaspo, M.-L. 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* 15:675. doi: 10.1186/1471-2164-15-675.

Trapnell, C., Pachter, L., and Salzberg, S.L. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105-1111. doi: 10.1093/bioinformatics/btp120.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511-515. doi: 10.1038/nbt.1621.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:562-578. doi: 10.1038/nprot.2012.016.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470-476. doi: 10.1038/nature07509.

Internet Resources

https://github.com/arq5x/bedtools2

BEDTools Suite Web site, which allows for downloading and installing the BEDTools applications. This toolset can analyze genome-wide datasets and work with genomic intervals.

http://genome.ucsc.edu/

UCSC Genome browser Web site to download gene annotations and view genomic interval files.

http://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html

Download the DEXSeq package from this site to obtain the dexseq_prepare_annotation.py script.

http://www-huber.embl.de/HTSeq/

Download, install, and read about the Python package HTSeq to analyze high-throughput sequencing data.

https://www.python.org/

Python Web site to download and install the Python software to run code that was written in the Python programming language.

http://www.ensembl.org/

Ensembl creates and provides gene annotations for several genomes. Gene annotations downloaded from this site can be used to create exonic part annotations.

http://www.htslib.org/

Samtools is a software suite that enables users to read/write and edit high-throughput sequencing data.