Introduction to RNA-Seq

Dhivya Arasappan

(With some slides borrowed from Scott Hunicke-Smith and Jeff Barrick)

Goals of the Class

- When considering an RNA-Seq experiment
  - What kind of options are available for library prep?
- When you have an RNA-Seq dataset
  - What kind of options are available for analysis?

Logistics

- Commands that I will run today also on BIoIteam wiki:
  - https://wikis.utexas.edu/display/bioiteam/Introduction+to+RNA+Seq+Short+Course+Commands

Resources

- BioIteam Wiki- Bookmark it!
  https://wikis.utexas.edu/display/bioiteam

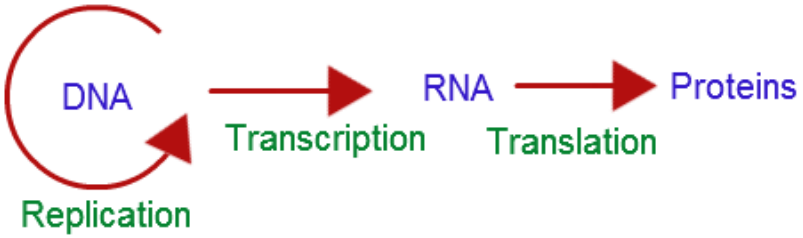
- Summer School course materials:
  https://wikis.utexas.edu/display/bioiteam/Introduction+to+RNA+Seq+Course+2015

- Other CCBB Short courses:
  http://ccbb.biosci.utexas.edu/shortcourses.html

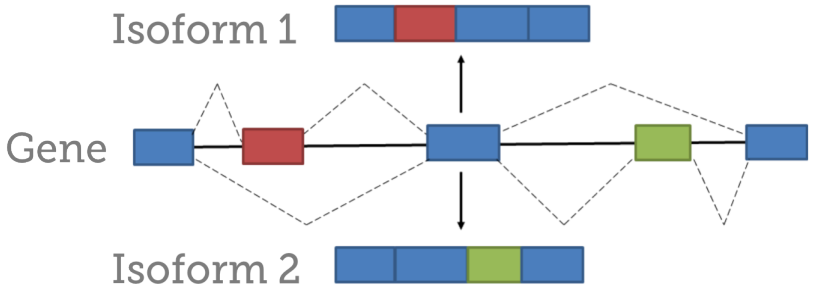
- CCBB Bioinformatics consultants

The Purpose of RNA-Seq

- Examine the state of the transcriptome.



- Genes expression patterns vary in:
  - Tissue types
  - Cell types
  - Development stages
  - Disease conditions
  - Time points



- RNA-Seq measures these expression variations using high-throughput sequencing technologies.

- Additionally, RNA-Seq allows detection of novel isoforms of genes.

Other Uses of RNA-Seq

- Assembling and annotating a transcriptome
- Characterization of alternative splicing patterns
- Gene fusion detection
- Small RNA profiling
- Targeted approaches using RNA-Seq
- Direct RNA sequencing

Advantages of RNA-Seq

Technology	Tiling microarray	RNA-Seq
Technology specifications		
Principle	Hybridization	High-throughput sequencing
Resolution	From several to 100 bp	Single base
Throughput	High	High
Reliance on genomic sequence	Yes	In some cases
Background noise	High	Low
Application		
Simultaneously map transcribed regions and gene expression	Yes	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes
Ability to distinguish allelic expression	Limited	Yes
Practical issues		
Required amount of RNA	High	Low
Cost for mapping transcriptomes of large genomes	High	Relatively low

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, **Mark Gerstein**, and **Michael Snyder**

Nat Rev Genet. 2009 January ; 10(1): 57–63. doi:10.1038/nrg2484.

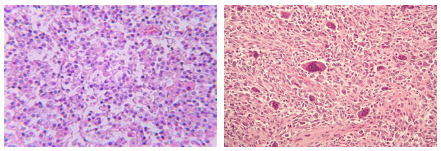
What are your questions ?

- This determines how you set up your experiment and how you analyze the data.
- What are you looking for?
  - Annotating a transcriptome?
  - Differential expression?
    - Novel transcripts, junctions?
    - Differential gene expression?
    - Differential exon level counts?
    - Differential regulation?
  - Small RNA?

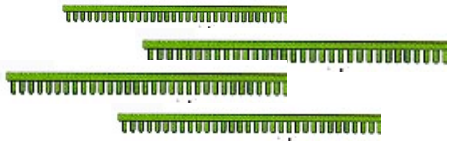
Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important Actually important!

From RNA-seqlopedia

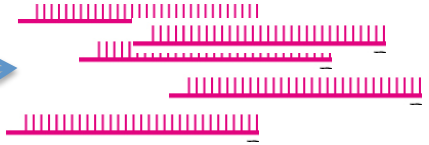
RNA-Seq… at it's Most Basic Form



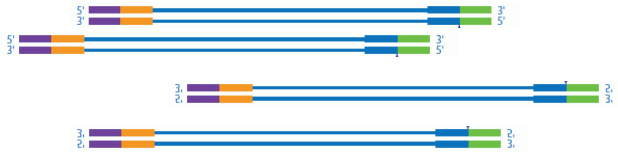
Samples from two conditions



Isolate RNA

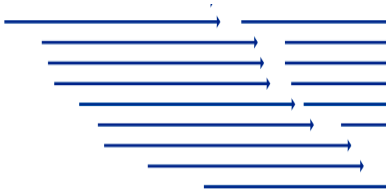


Generate cDNA



Create sequencing library by size selection and adding adaptors

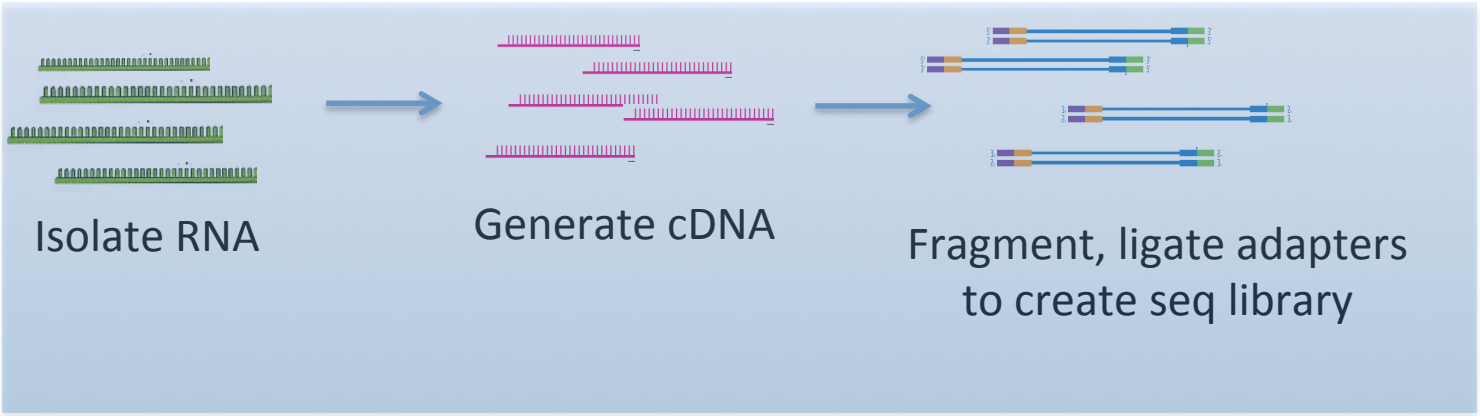
Run sequencer



Generate short reads

Identify differentially expressed genes

RNA-Seq Libraries… with More Details



A. rRNA Depletion

Ribominus kit

B. Normalized library

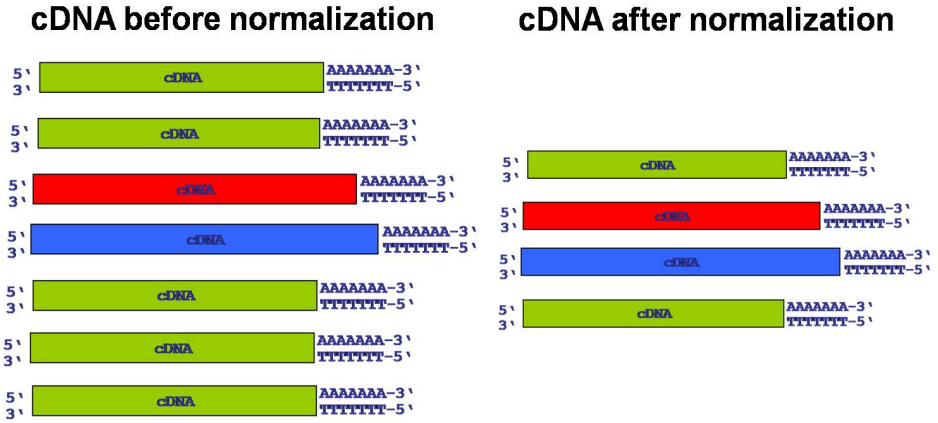
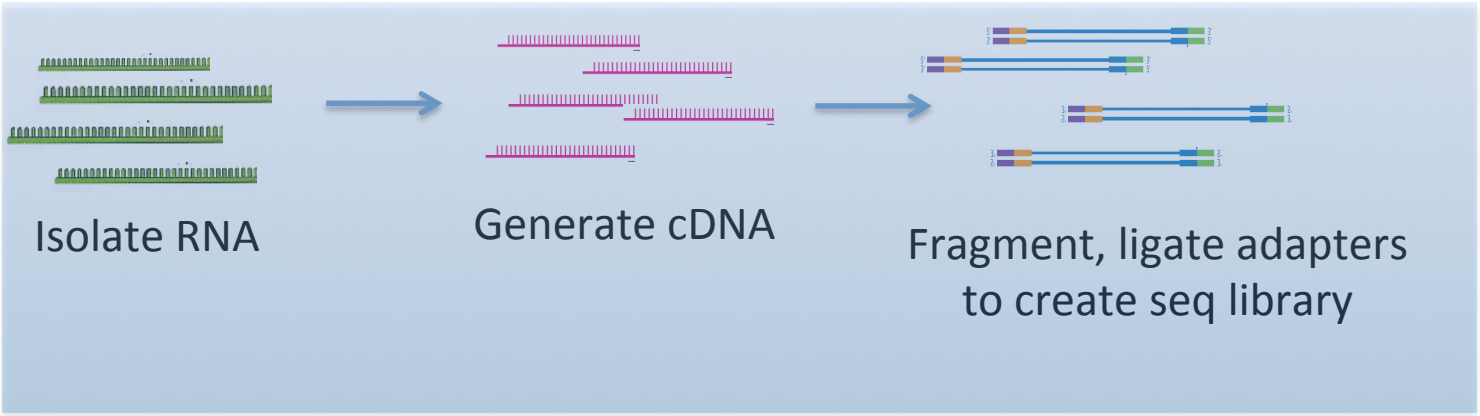


Image from :www.genxpro.info

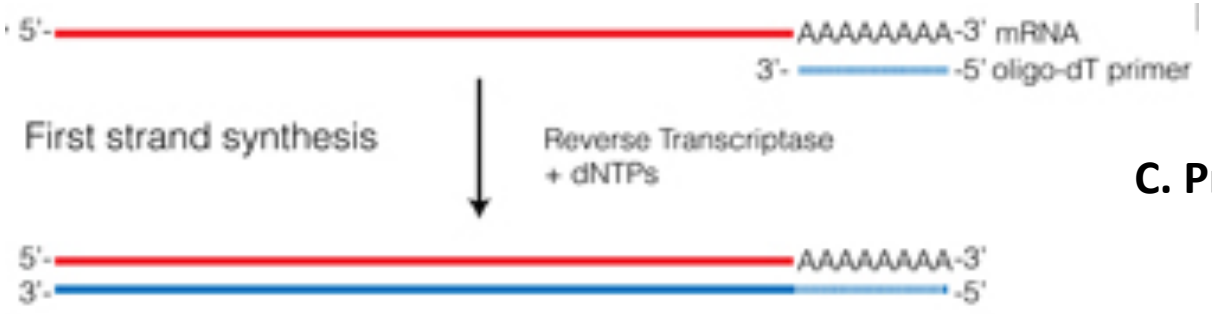
C. Size selection

Reserved for miRNA, siRNA profiling

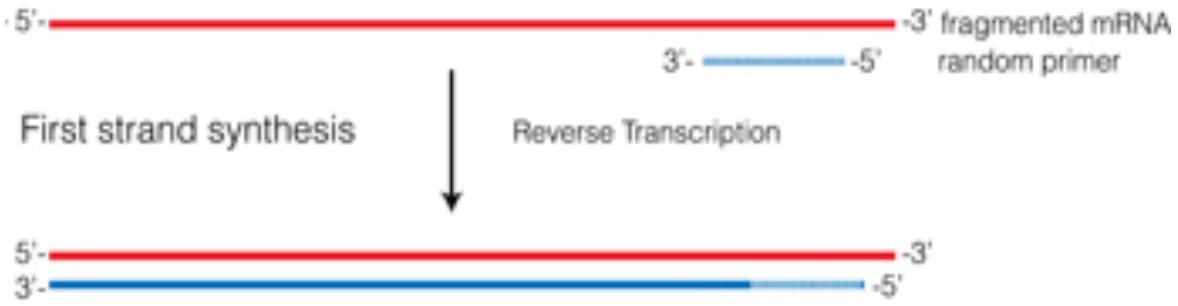
RNA-Seq Libraries… with More Details



A. Poly A Priming



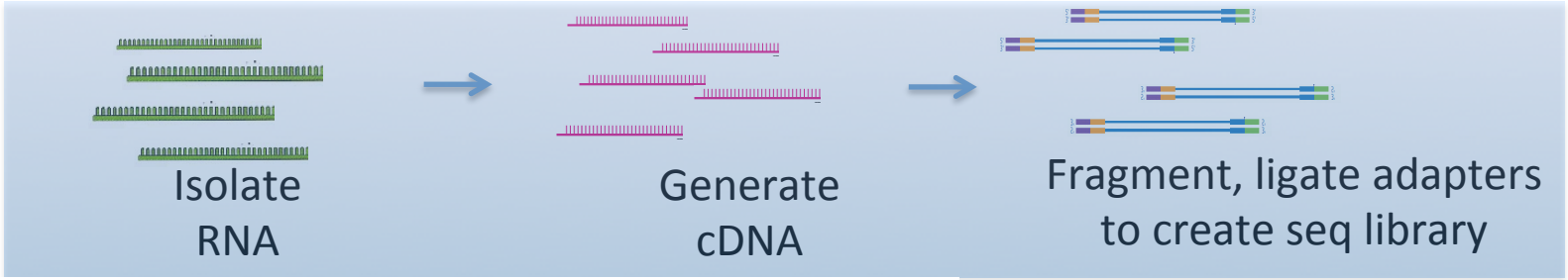
B. Random Priming

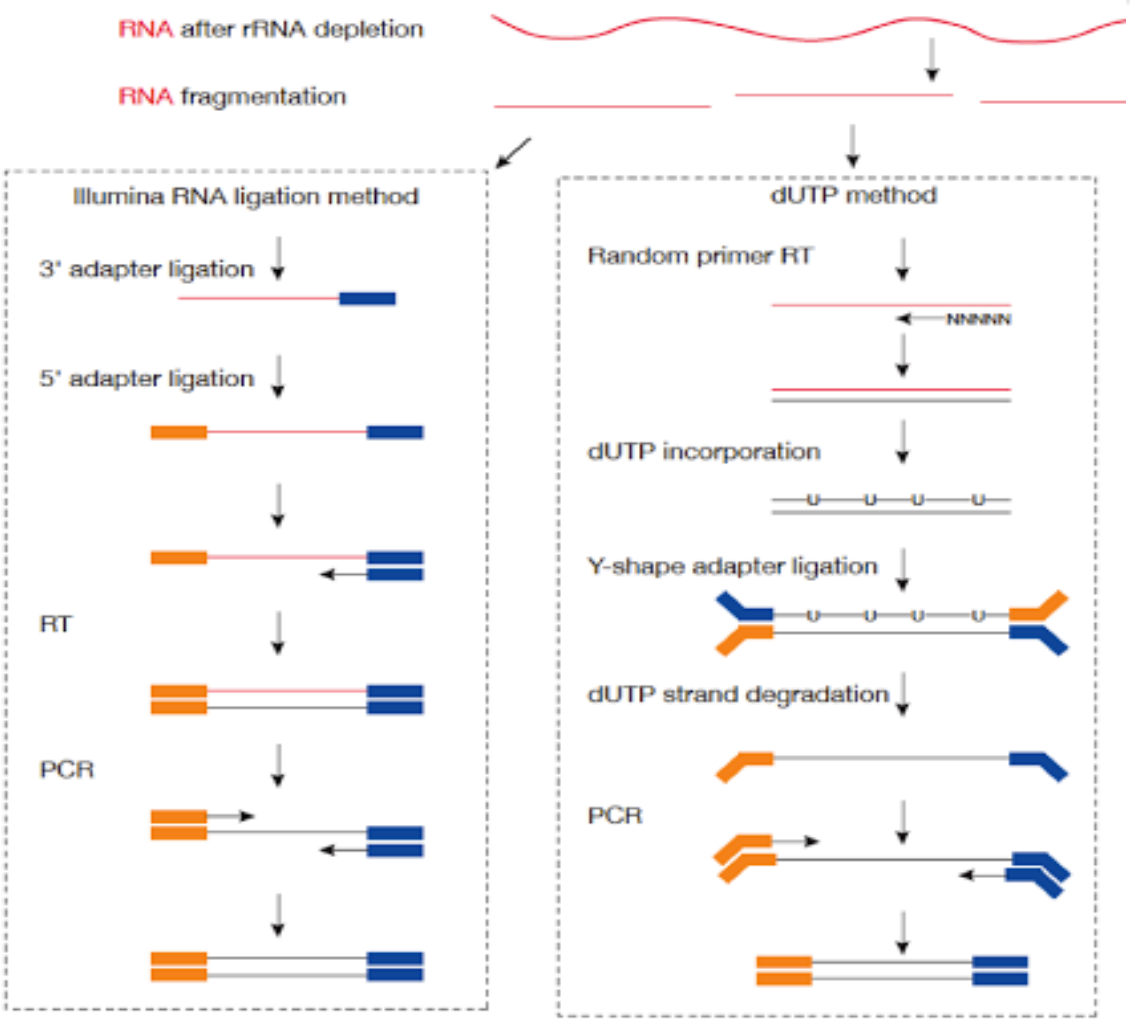


C. Priming using pre-ligated oligo

Illumina Small RNA kit
SOLiD RNA kits

RNA-Seq Libraries… with More Details



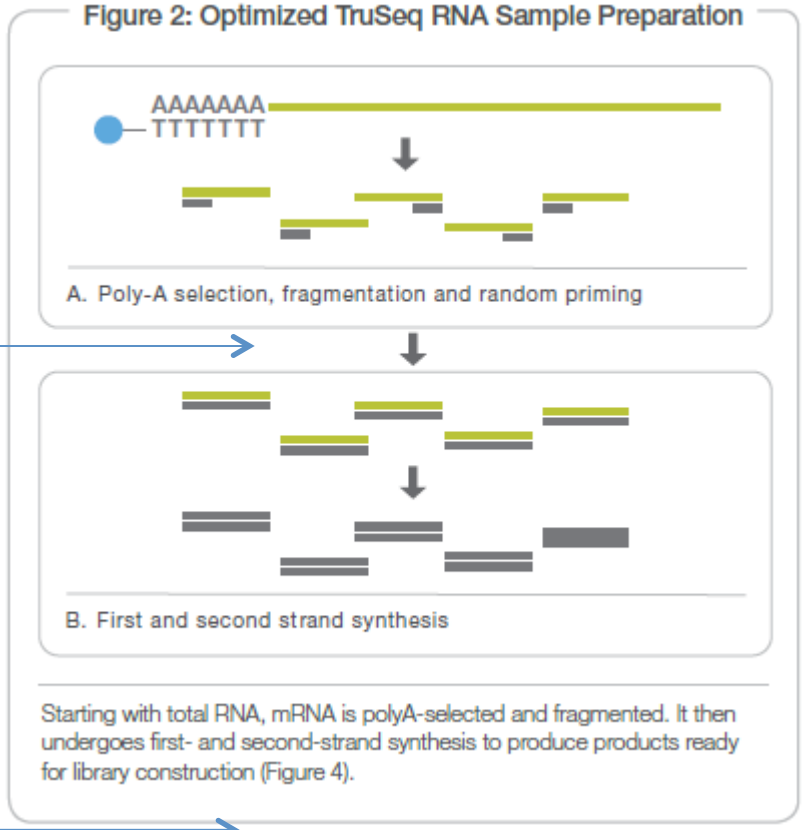


Second Strand Synthesis- Many Strand Specific Methods.

Strand-specific libraries for high throughput RNA sequencing prepared without poly(A) selection, Zhang et al.

RNA Illumina Tru-Seq library prep

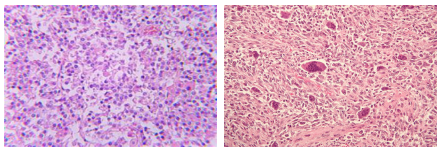
Size selection step



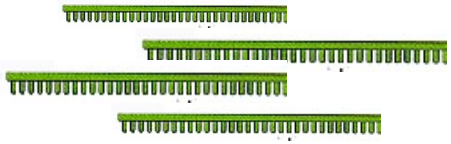
Adaptor ligation and standard library preparation

2 days for 8 samples

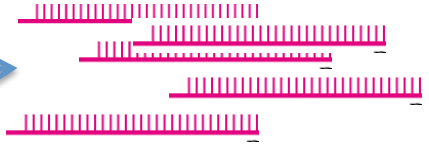
RNA-Seq… at it's Most Basic Form



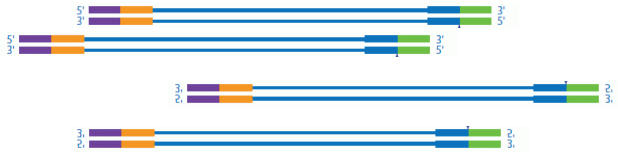
Samples from two conditions



Isolate RNA

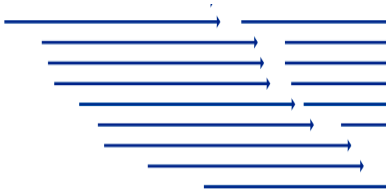


Generate cDNA



Create sequencing library by size selection and adding adaptors

Run sequencer



Generate short reads

Identify differentially expressed genes

Types of Illumina Fragment Libraries

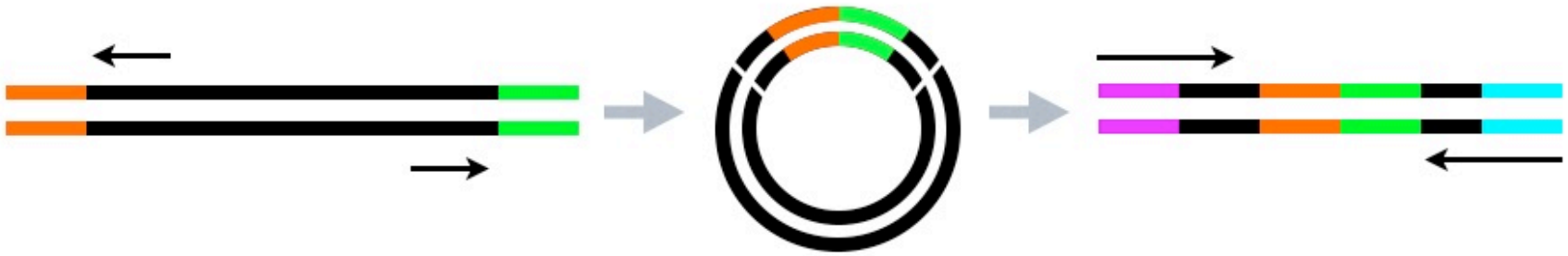
single-end

independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired



two outwardly oriented reads separated by ~3000 nt

Comparing Stranded RNA-Seq Library Protocols

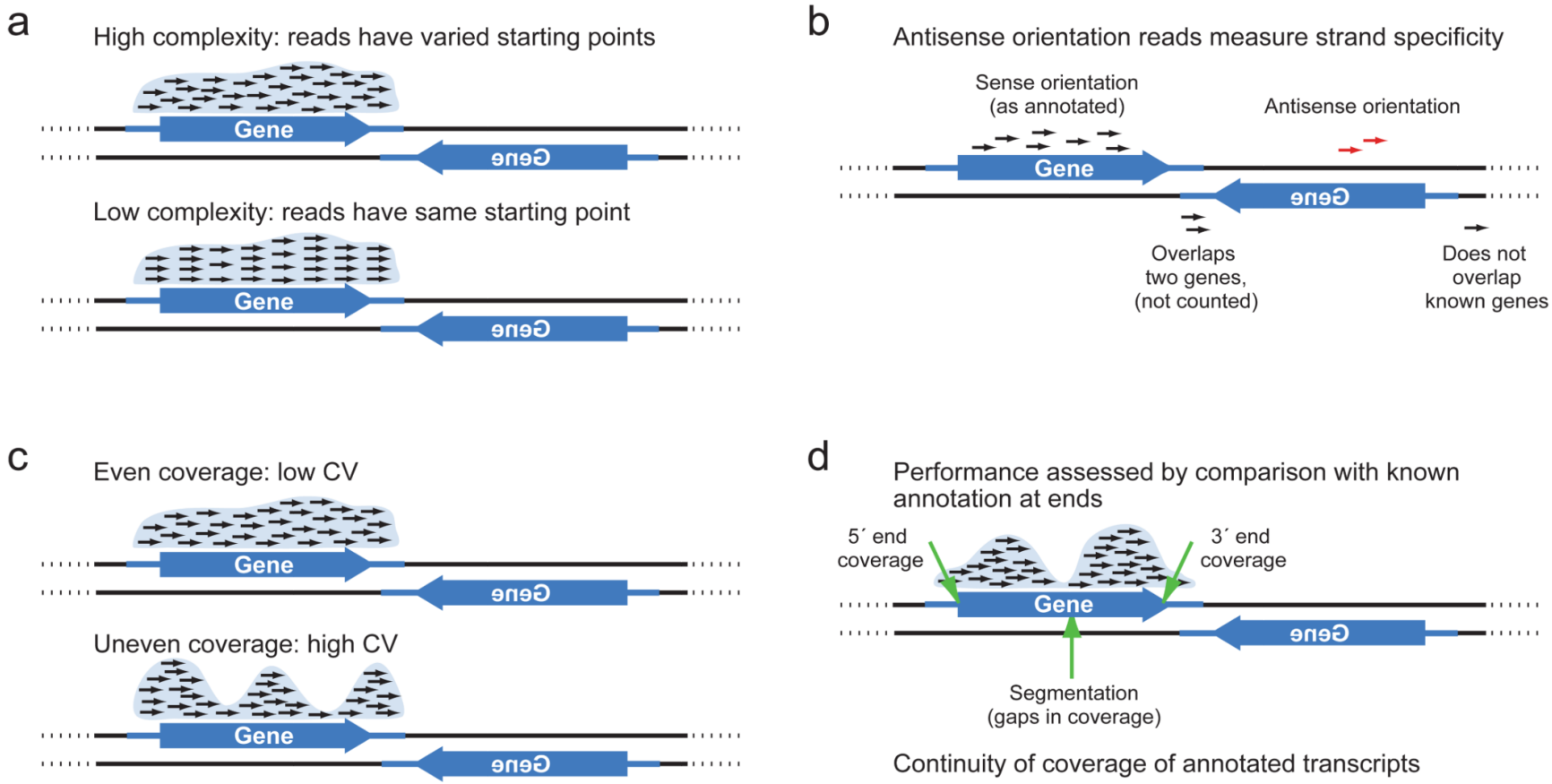
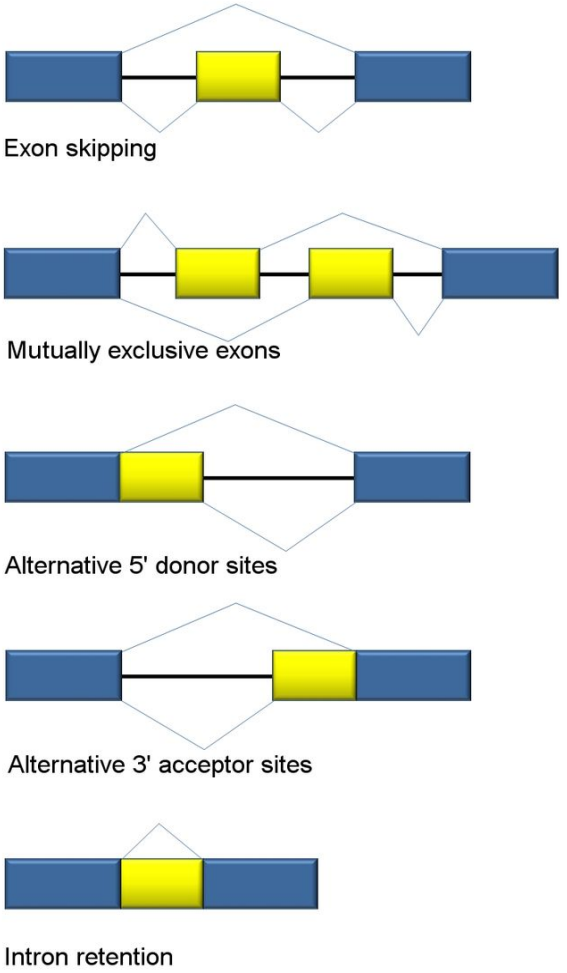


Figure 2. Key criteria for evaluation of strand-specific RNAseq libraries
Four categories of quality assessment. Double stranded genome (black parallel lines), with Gene ORF orientation (thick blue arrow) and UTRs (thin blue line), along with mapped reads (short black arrows – reads mapped to sense strand; red – reads mapped to antisense strand). (**a**) Complexity. (**b**) Strand Specificity. (**c**) Evenness of coverage. (**d**) Comparison to known transcript structure..

Comprehensive comparative analysis of strand-specific RNA sequencing methods, Levin et al, 2010

Why is RNA-Seq Difficult?

- Biases may mean what we are seeing is not reflective of true state of the transcriptome.
- Ugh, splicing!
- Gene level, exon level?
- Multimapping, partial mapping, not mapping.
- Normalization issues
  - some datasets are larger than others, some genes are larger than others



From Wikipedia- alternative splicing

How do we analyze RNA-Seq data?

- **STEP 1**: EVALUATE AND MANIPULATE RAW DATA
- **STEP 2**: MAP TO REFERENCE, ASSESS RESULTS
- **STEP 3**: ASSEMBLE TRANSCRIPTS
- **STEP 4**: QUANTIFY TRANSCRIPTS
- **STEP 5**: TEST FOR DIFFERENTIAL EXPRESSION
- **STEP 6**: VISUALIZE AND PERFORM OTHER DOWNSTREAM ANALYSIS

STEP 1 - Evaluate Raw Data

FASTQ FORMAT

```
@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBB@BBBBAAAA>@AABA?BBBAAB??>A?
```

Line 1: @read name

Line 2: called base sequence

Line 3: +read name (optional after +)

Line 4: base quality scores

STEP 1 - Evaluate Raw Data

Illumina Base Quality Scores

http://www.asciitable.com/

```
Quality character   !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                    |          |          |          |          |
ASCII Value         33         43         53         63         73
Base Quality (Q)    0          10         20         30         40
```

$$\text{Probability of Error} = 10^{-Q/10}$$

(This is a **Phred** score, also used for other types of qualities.)

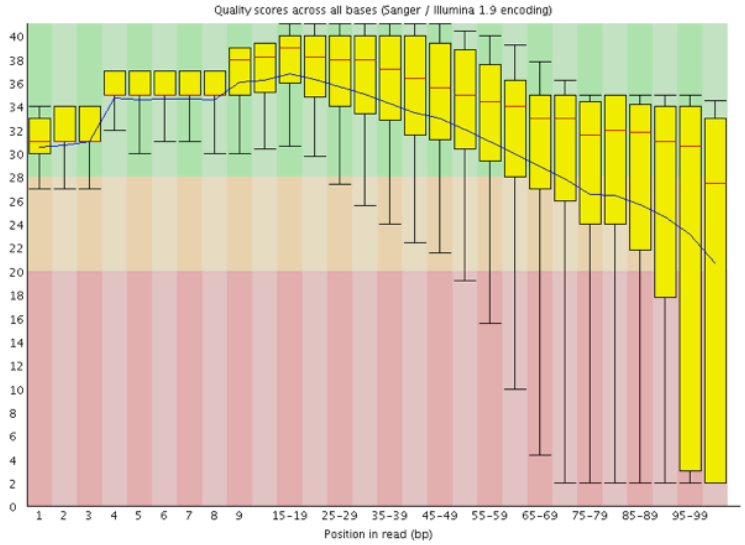
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

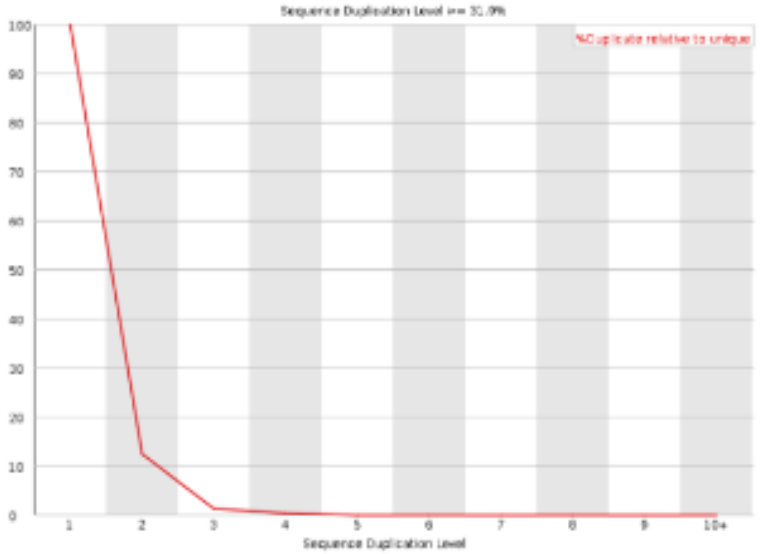
Quality scores are ASCII encoded in fastq files. Different platforms/older sequencing data can have different encoding! Illumina HiSeq 2500 produces Sanger encoded data.

Phred +33 =ASCII

STEP 1 - Evaluate Raw Data

- Count your reads!
- Assess quality using FastQC reports



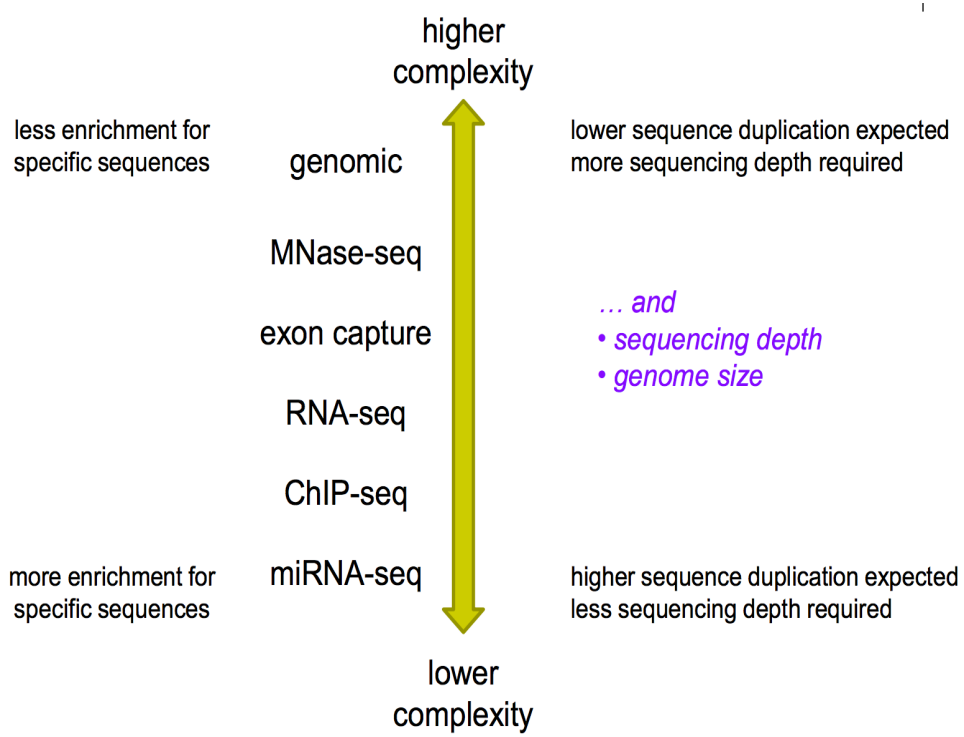


Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATG	60030	5.01369306977828	TruSeq Adapter, Index 1 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGC	42955	3.5875926338884896	TruSeq Adapter, Index 1 (97% over 37bp)
CACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGCCGTCTTCTGCT	3574	0.29849973398946483	RNA PCR Primer, Index 40 (100% over 41bp)
CAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	2519	0.2103863542024236	TruSeq Adapter, Index 1 (97% over 37bp)
GAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	1251	0.10448325887543942	TruSeq Adapter, Index 1 (97% over 37bp)

STEP 1 - Evaluate Raw Data

- Sequence duplication levels does not always indicate PCR amplication issues.

Library complexity is a function of experiment type…



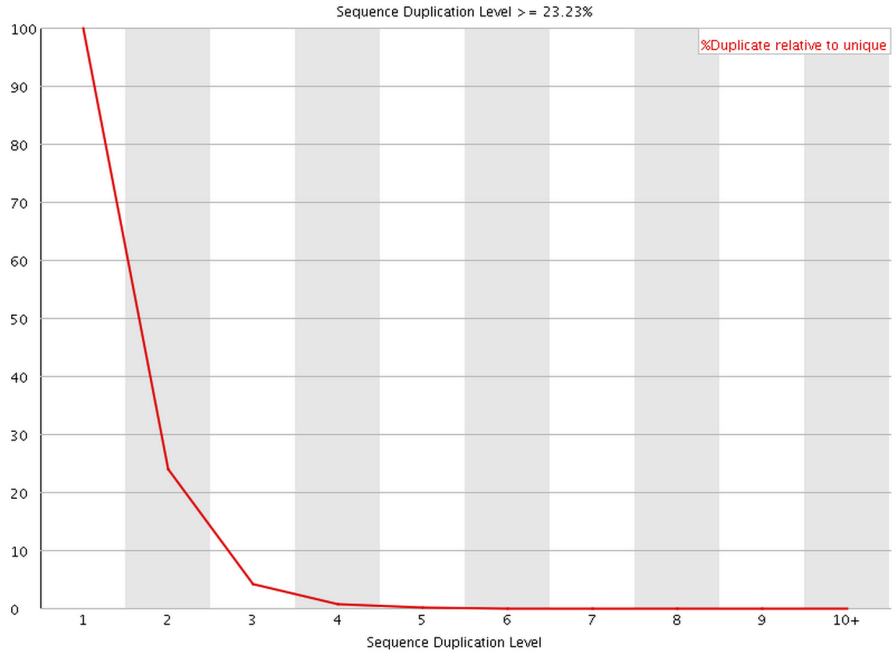
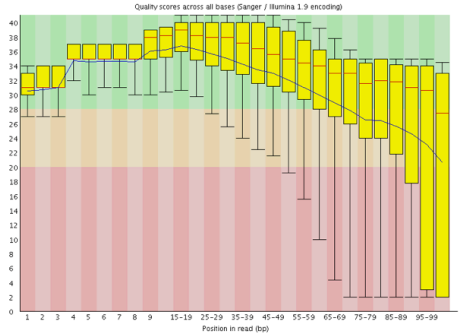
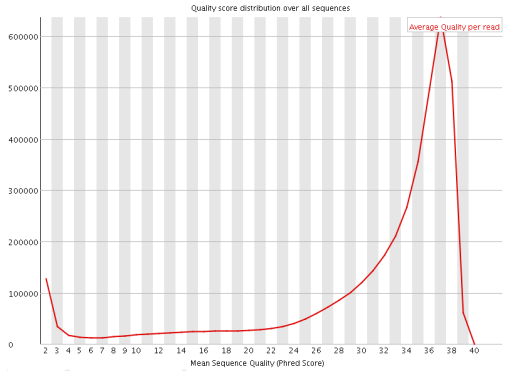


Image created by Anna Battenhouse

STEP 1 – Manipulate Raw Data





- Trim low quality bases
  - Fastx toolkit- **fastx_trimmer**
    - Trim X number of low quality bases from each read.
- Filter out low quality reads
  - Fastx toolkit- **fastq_quality_filter**
    - Filter out reads with more than X percent of low quality bases.

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATG	60030	5.01369306977828	TruSeq Adapter, Index 1 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGC	42955	3.5875926338884896	TruSeq Adapter, Index 1 (97% over 37bp)
CACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGCCGTCTTCTGCT	3574	0.29849973398946403	RNA PCR Primer, Index 40 (100% over 41bp)
CAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	2519	0.2103863542024236	TruSeq Adapter, Index 1 (97% over 37bp)
GAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTAT	1251	0.10448325887543942	TruSeq Adapter, Index 1 (97% over 37bp)

- Trim Adaptor
  - Fastx toolkit- **fastx_clipper**
    - Look for and clip a given sequence from the end of reads
  - **Cutadapt**
    - Allows for mismatches
    - Paired -end support

RNA-Seq Analysis Pipelines

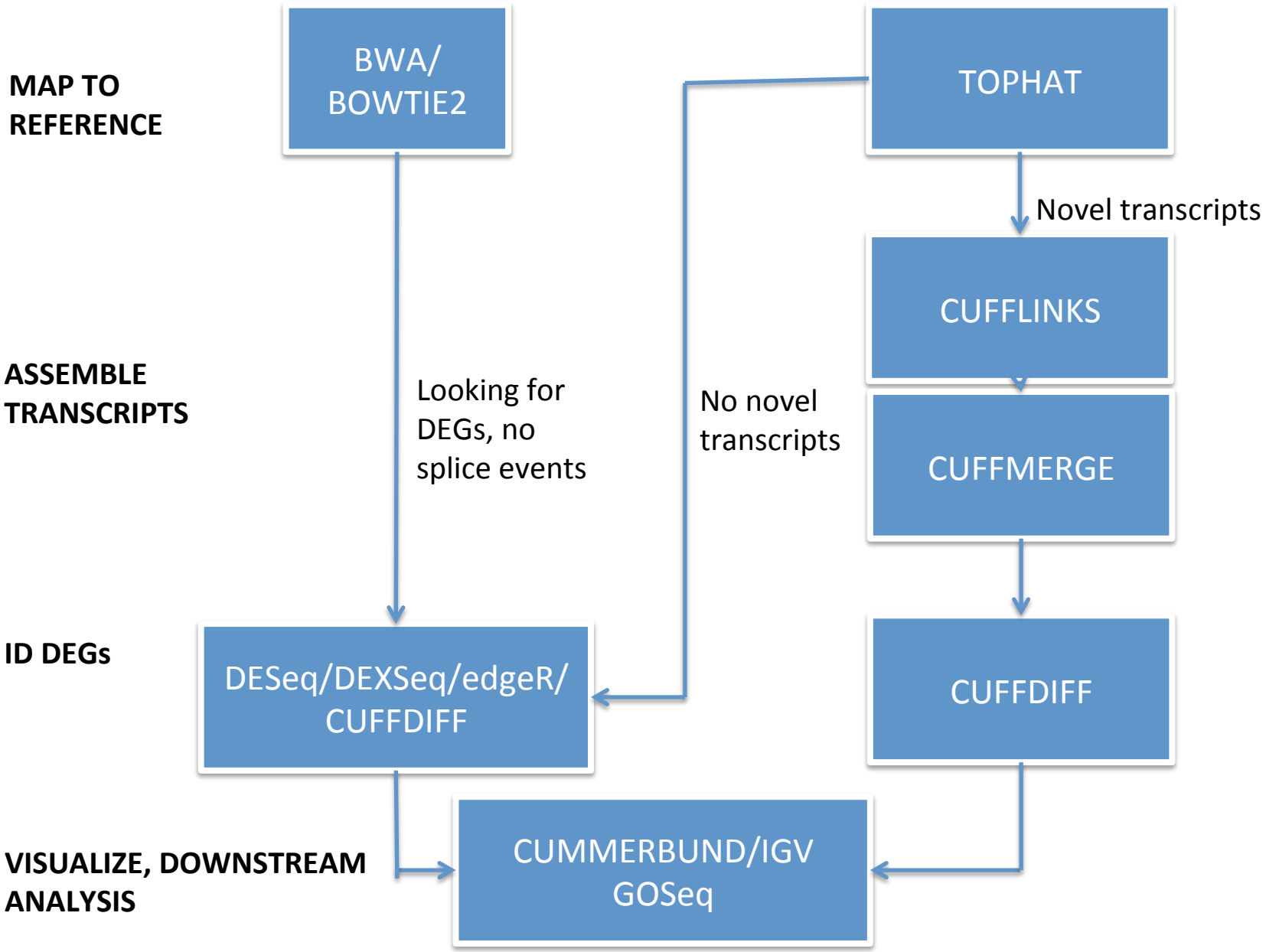
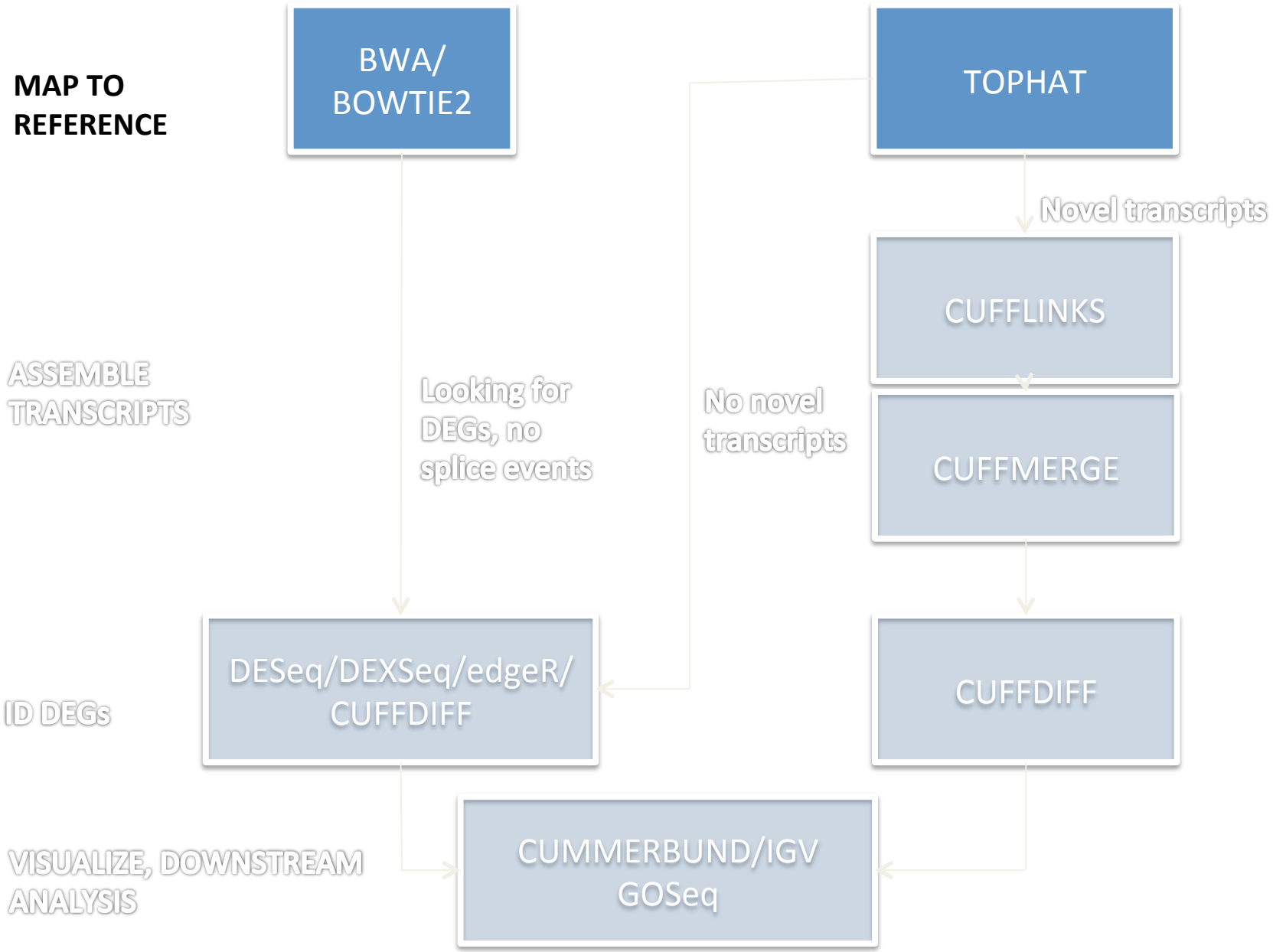


Table 1 | Selected list of RNA-seq analysis programs

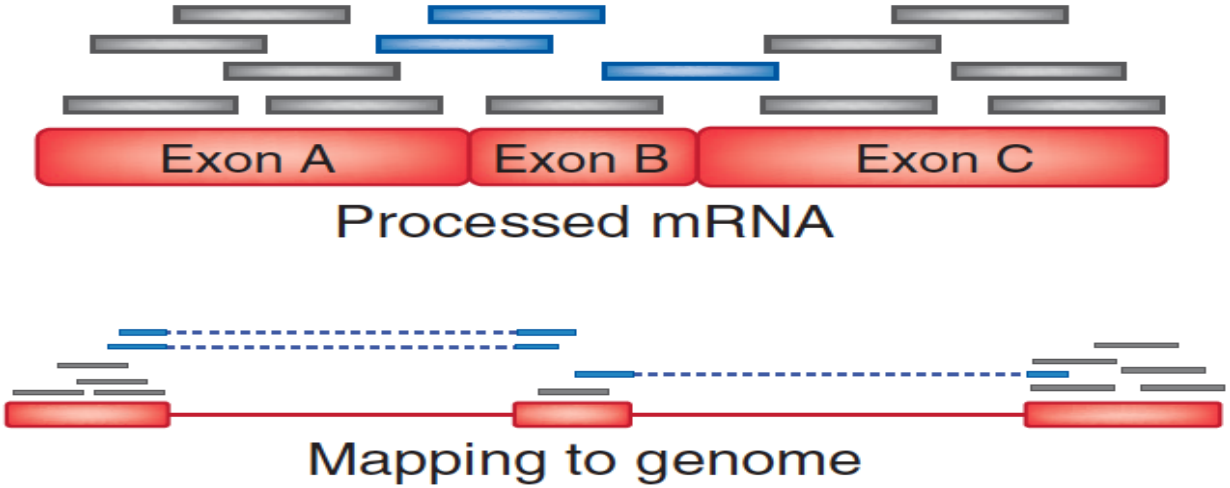
Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

Figure:
Garber et al, Nature Methods, 2011

STEP 2 - Map to Reference



Unspliced Mapping

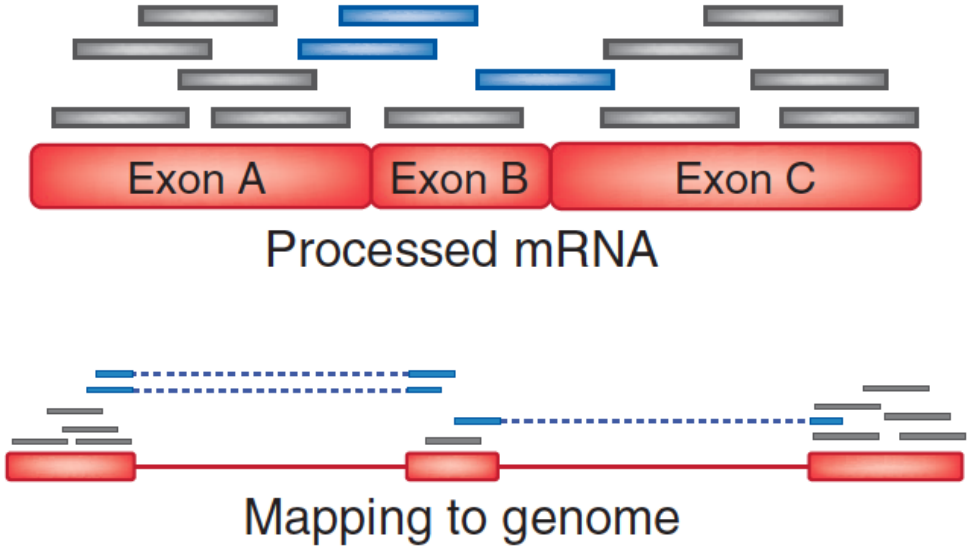


Class	Category	Package	Notes
Read mapping			
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension
		Stampy[39]	Probabilistic model
	Burrows-Wheeler transform methods	Bowtie[43]	
		BWA[44]	Incorporates quality scores

Garber et al, Nature Methods, 2011

Spliced mapping

- Needed for identifying and quantifying splice variants from RNA Seq data.
- Tools:
  - Tophat
  - SpliceMap
  - MapSplice
  - STAR
  - RUM



Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

Spliced mapping

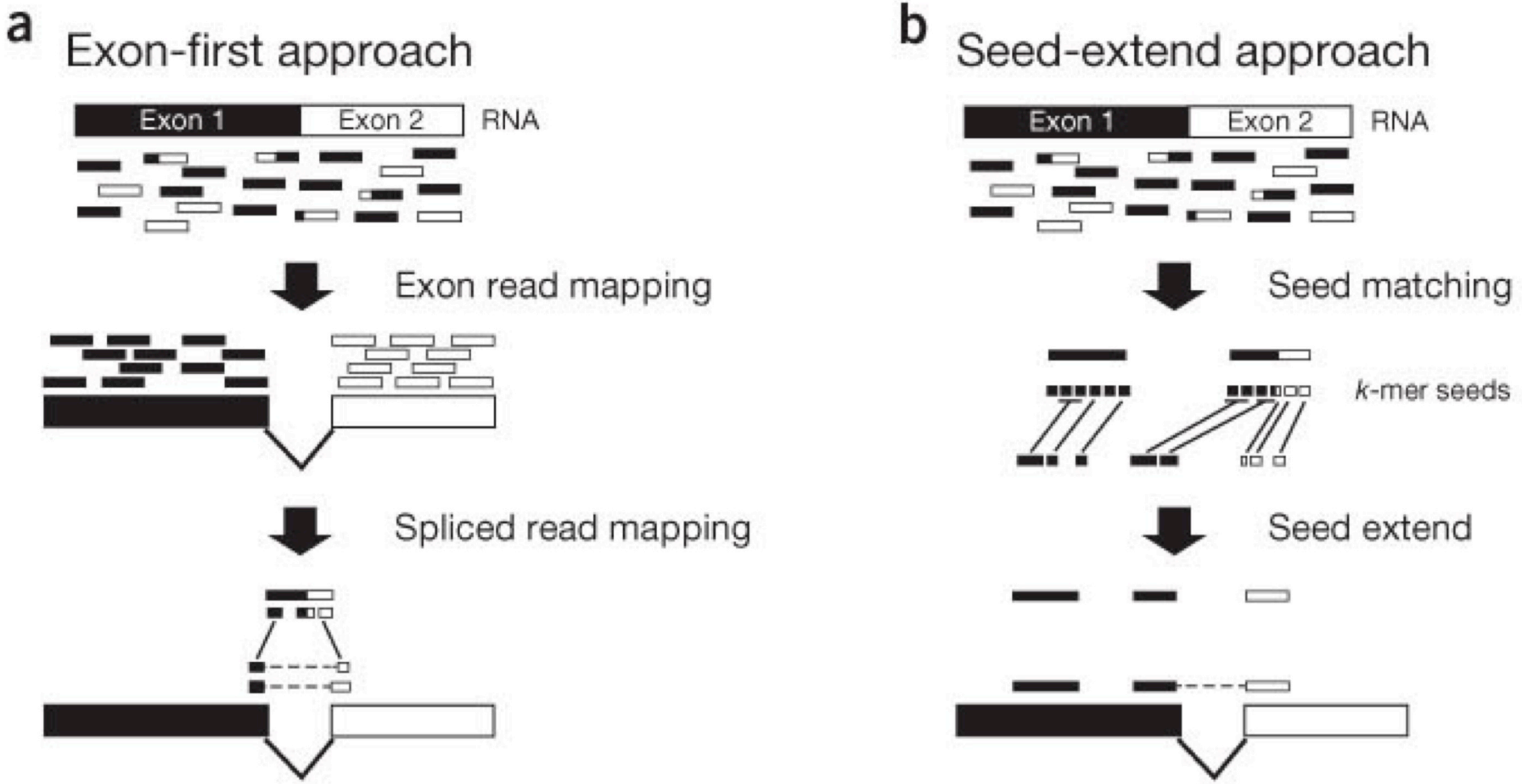


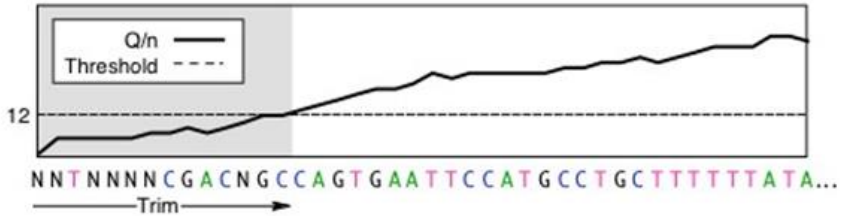
Figure :
Garber et al, Nature Methods, 2011

What to know about your data before mapping?

KNOW YOUR DATA!

- Paired end? Single end?
- Traditional RNA-Seq? 3’ tag ?
- Insert size estimate?

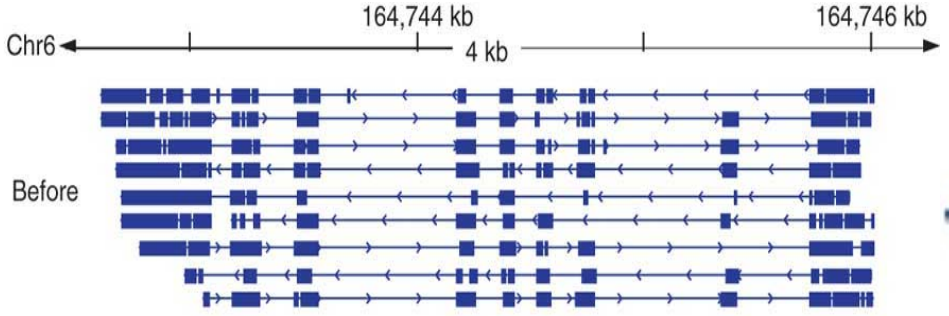
PREPROCESSING



- Adaptor sequences trimmed?
- Primer sequences/barcodes removed?
- Poor quality regions trimmed?

What to know about your reference before mapping?

- Mapping to genome vs transcriptome?



- Is your reference the right version?
- Does your annotation match your reference?

What will your reference look like?

- FASTA Format

  >gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
  agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
  tgatagcagcttctgaactggttacctgccgtgagtaaattaaaattttattgacttagg
  tcactaaatactttaaccaatataggcatagcgcacagacagataaaaattacagagtac
  acaacatccatgaaacgcattagcaccaccattaccaccaccatcaccattaccacaggt
  ....

- Using complex reference sequence names is a common problem during analysis. Might rename:

  >REL606
  agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
  tgatagcagcttctgaactggttacctgccgtgagtaaattaaaattttattgacttagg

What will your annotation look like?

- GFF3 Format
  - seqname - The name of the sequence.
  - source - The program that generated this feature.
  - feature – Examples: "CDS", "start_codon", "stop_codon", and "exon".
  - start - The starting position of the feature in the sequence.
  - end - The ending position of the feature (inclusive).
  - score - A score between 0 and 1000.
  - strand - Valid entries include '+', '-', or '.' (for don't know/don't care).
  - Frame – reading frame
  - group – ID and other information about the entry

  Example:

  Rel606 refseq cds 1450 1540 500 + . Gene_id=« test_gene »

- Make sure the GFF3 file matches your reference fasta file.

Mapping with BWA

- BWA is a fast short read aligner that uses the burrows-wheeler transform to perform alignment in a time and memory efficient manner.

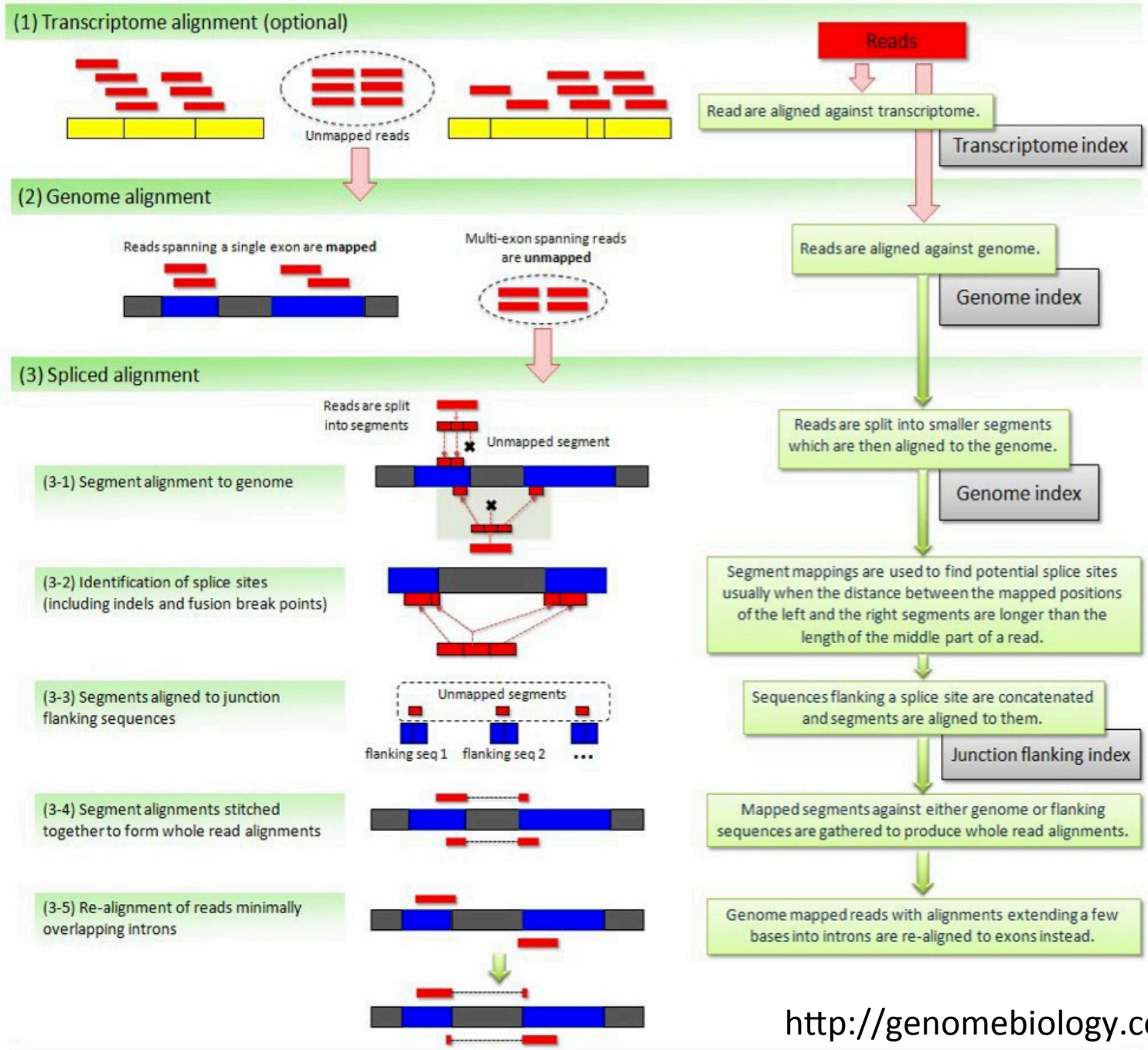
- BWA Variants
  - For reads upto 100 bp long
    - BWA-backtrack: BWA aln/samse/sampe
  - For reads upto 1 Mbp long
    - BWA-SW
    - **BWA-MEM**: Newer! Typically faster!

Mapping with BWA

- Create an index of your reference – bwa index
- Run mapping - bwa mem

- Help! I have a large number of reads. Make BWA go faster!
  - Use threading option (bwa –t <threads>)
  - Split one data file into smaller chunks, run multiple, parallel BWA instances, concatenate results.
    - Wait! We have a pipeline for that on lonestar – **runBWA_mem.sh** in $BI/bin

Mapping with Tophat



http://genomebiology.com/2013/14/4/R36

Mapping with Tophat

Steps:

1. Index the genome using bowtie
2. Map using tophat

Let's look at the command.

Help! I have a large number of reads. Make tophat go faster!

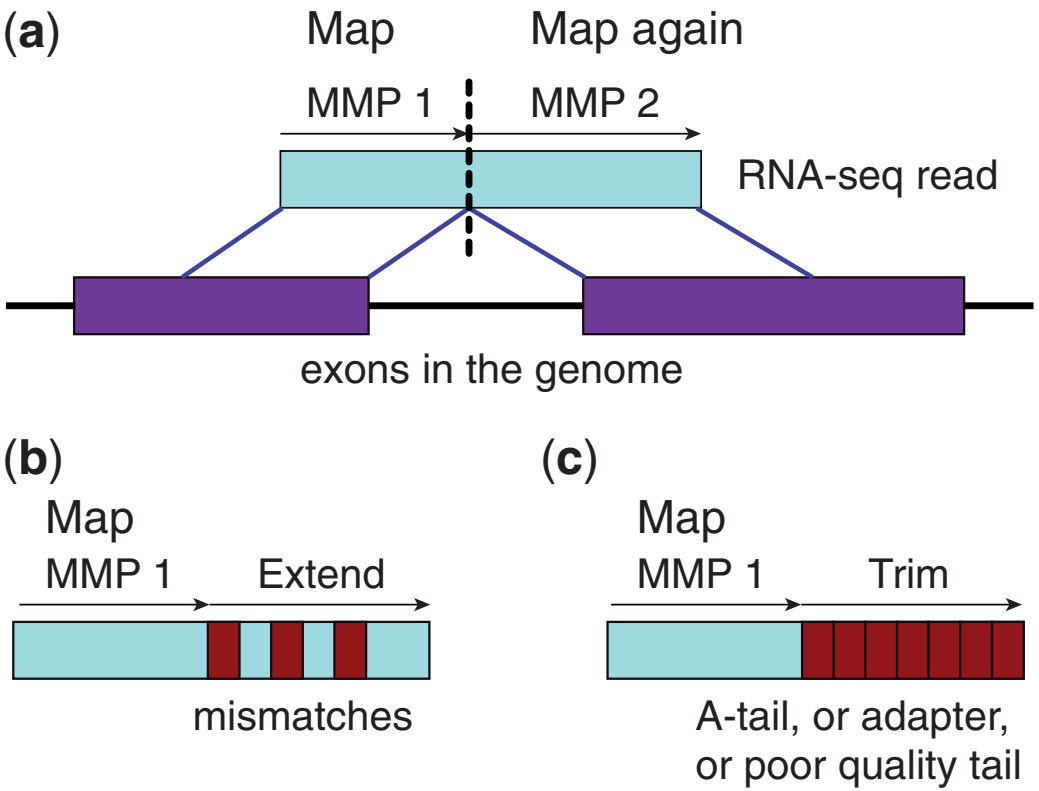
Use threading option (tophat -p <threads>)

Split one data file into smaller chunks, run multiple, parallel tophat instances, concatenate results.

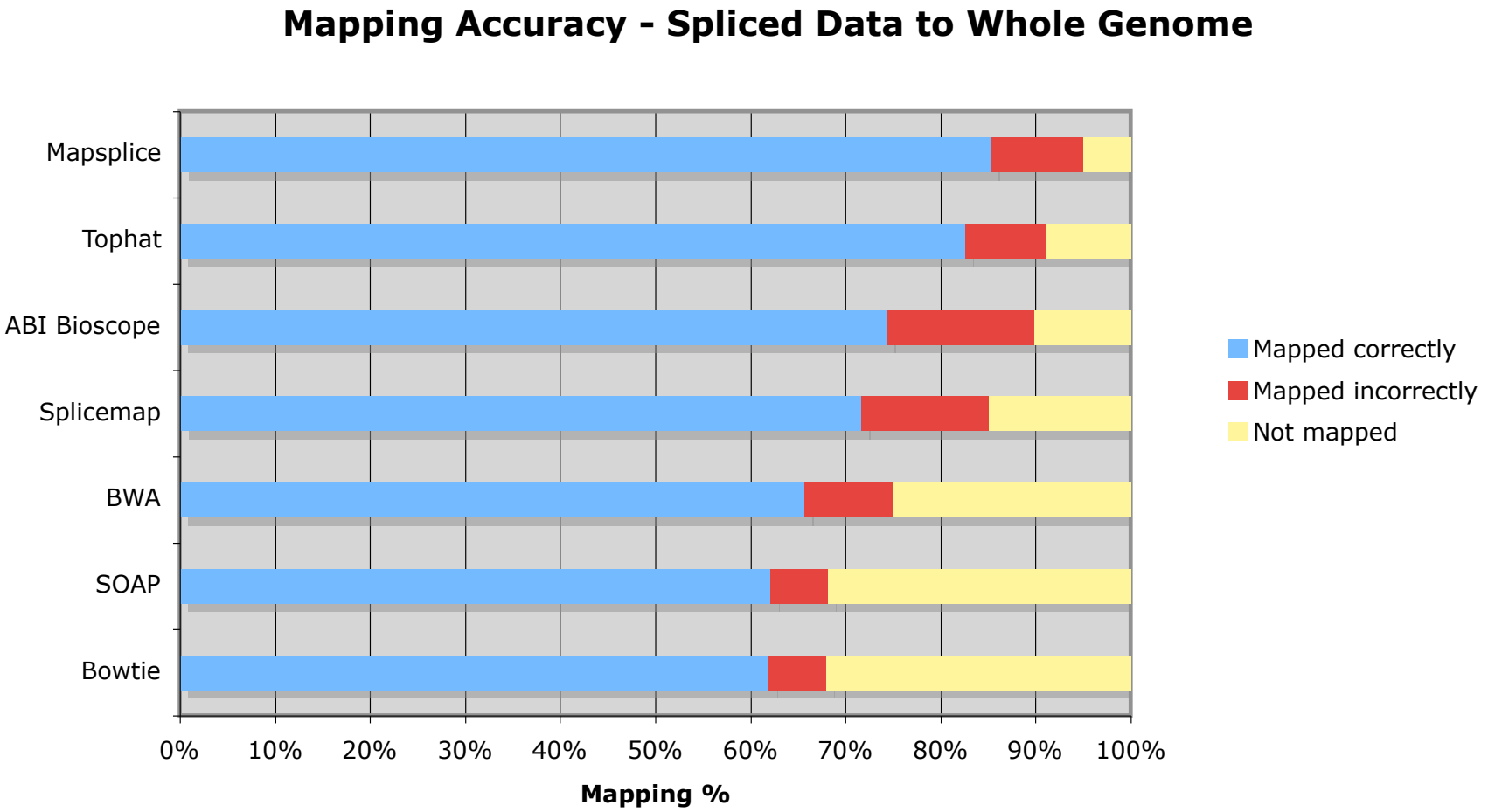
Wait! We have a pipeline for that on lonestar – **fastTophat.sh** in $BI/bin

Mapping with STAR

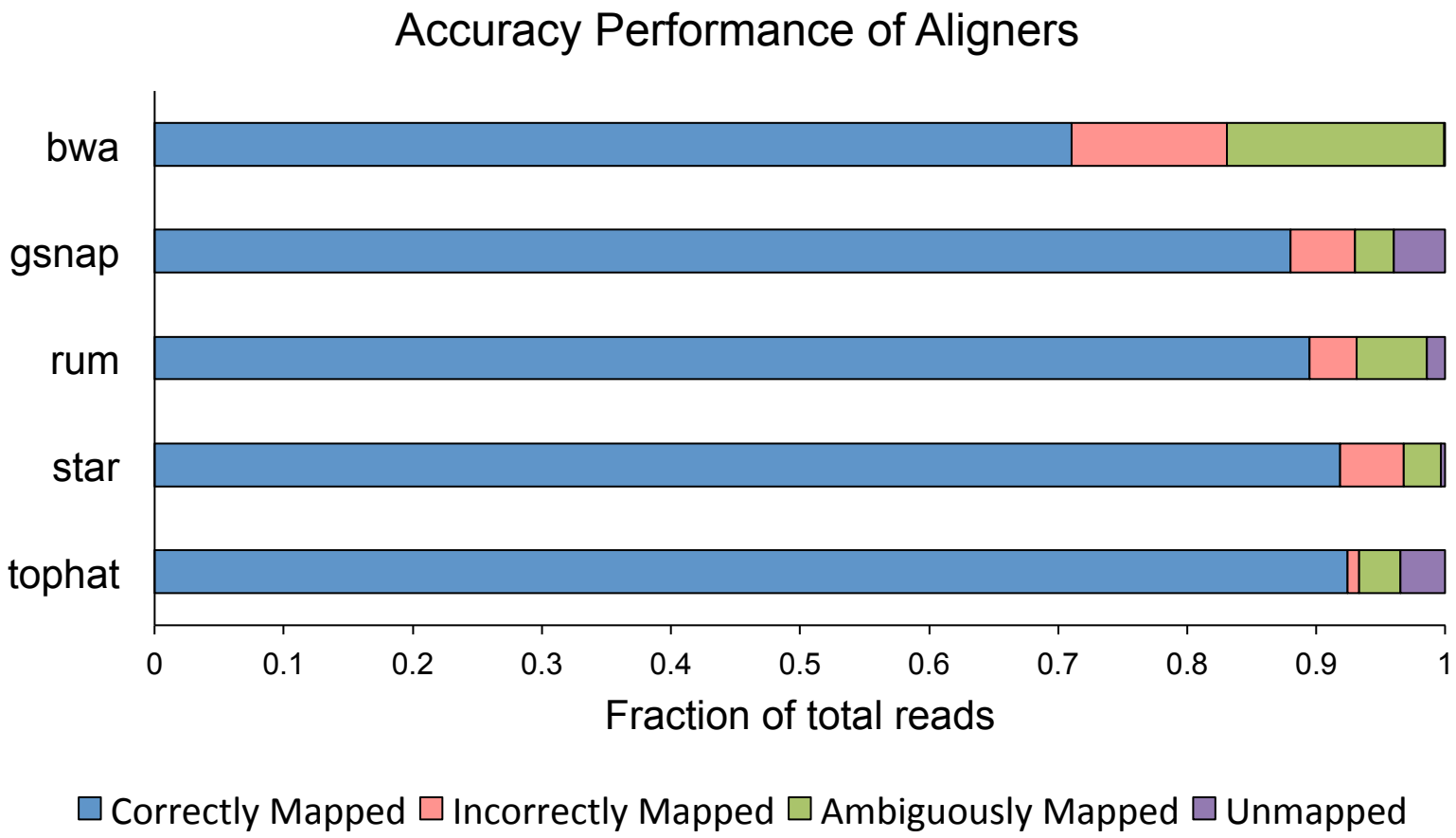
- “Spliced Transcripts Alignment to a Reference”
- Faster splice-aware mapper



Mappers comparisons

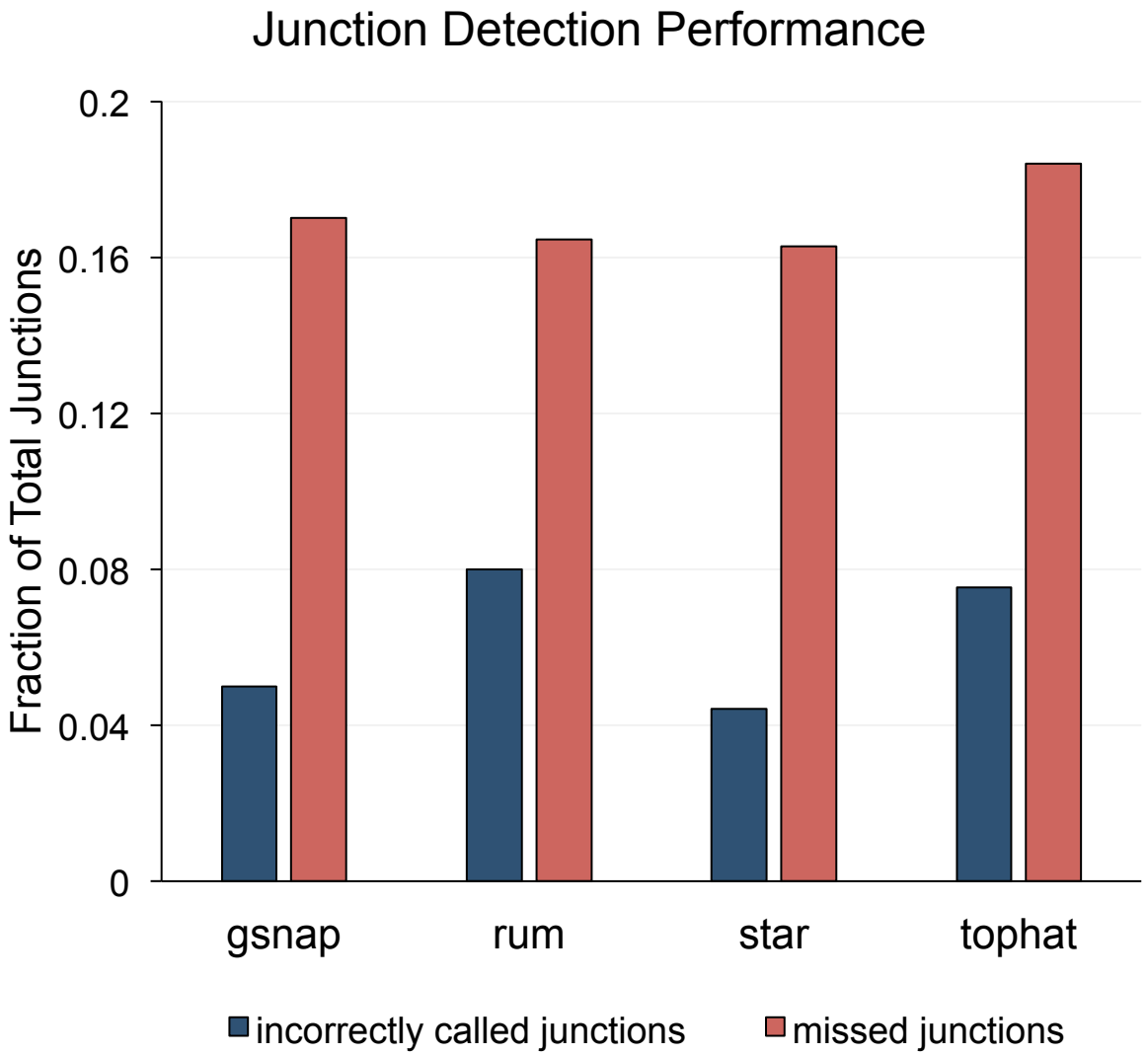


Mappers comparisons



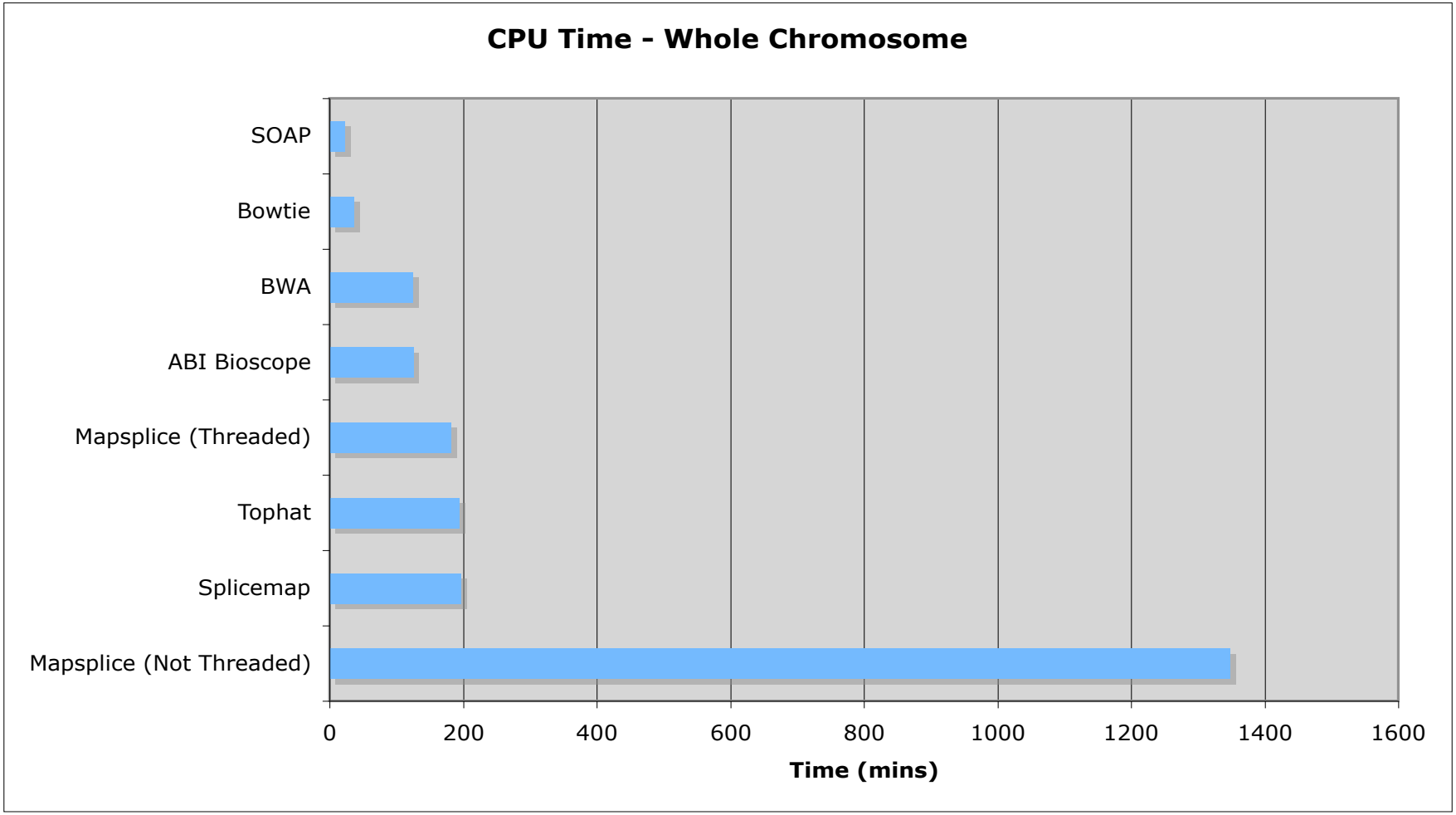
New benchmarking analysis performed by **Raghav Shroff**

Mappers comparison

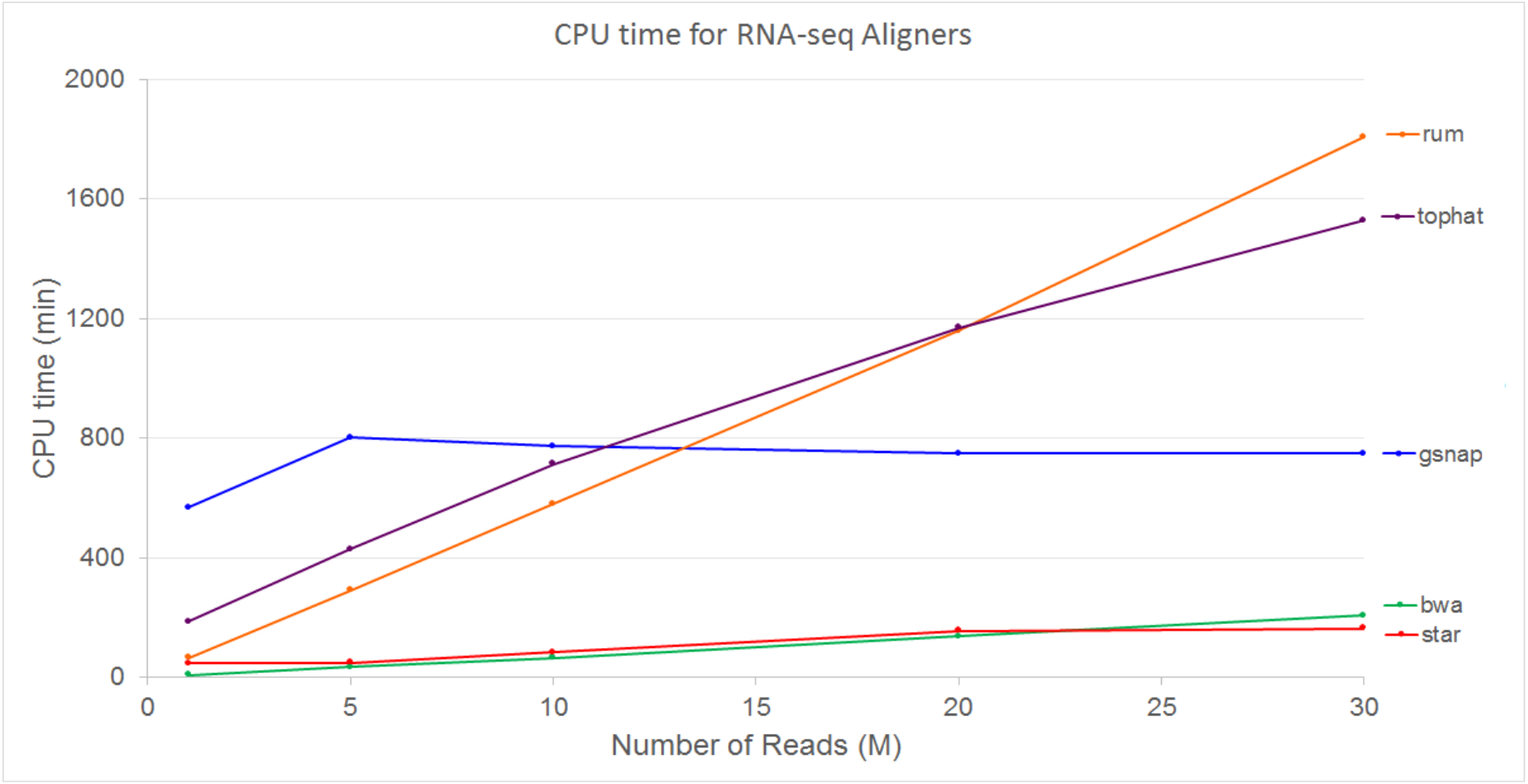


New benchmarking analysis performed by **Raghav Shroff**

Mappers comparison



Mappers comparison



New benchmarking analysis performed by **Raghav Shroff**

Mapping Output: SAM file format

- Alignment results generated in Sequence Alignment/ Map format
- Tab delimited, with fixed columns followed by user-extendable key:data values.
- Most mappers also output unmapped reads in SAM file.
- SAMTOOLS – toolkit to manipulate, parse SAM files.

Mapping Output: SAM File Format

SAM fixed fields: http://samtools.sourceforge.net/

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	`[0,2^16-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2^29-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,2^8-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	`[0,2^29-1]`	Position of the mate/next segment
9	TLEN	Int	`[-2^29+1,2^29-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SRR030257.264529 99 NC_012967 1521 29 34M2S = 1564
79 CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
XT:A:M NM:i:3 SM:i:29 AM:i:29 XM:i:3 XO:i:0 XG:i:0 MD:Z:23T0G4T4

Mapping Output: Mapping Quality

- Mapping quality is the probability that a read is aligned to the wrong place.

$$p = 10^{(-q/10)}$$

- BWA mapping quality calculated by considering:
  - Repeat structure of reference
  - Read base quality
  - Read alignment quality (mismatches etc)
  - Number of mappings

Mapping Output: CIGAR score

```
Ref   CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC
Read        aaATGTCGCGGTG.TAGGAggatcc
```

2S5M2I4M1D4M6S

	Op	BAM	Description
	M	0	alignment match (can be a sequence match or mismatch)
	I	1	insertion to the reference
	D	2	deletion from the reference
*	N	3	skipped region from the reference
	S	4	soft clipping (clipped sequences present in SEQ)
*	H	5	hard clipping (clipped sequences NOT present in SEQ)
*	P	6	padding (silent deletion from padded reference)
*	=	7	sequence match
*	X	8	sequence mismatch

*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

Mapping Output: BAM format

- SAM files are converted to BAM format through SAMTOOLS command:
    - samtools view –b –S samfile > bamfile
- BAM file is binary format.
- BAM file is compressed.
- BAM files are usually what you need for post mapping analysis and visualization.

Assess Mapping Results - Samtools

- For parsing and manipulating mapping output files in SAM and BAM formats.
  - Sorting mapping output files
  - Merging multiple mapping output files
  - Converting from SAM to BAM and vice versa
  - Retrieving reads based on different criteria: reads mapping to a particular region, unmapped reads etc
  - Collecting statistics about your mapping results

Assess Mapping Results - Samtools

1. Convert SAM file to BAM format

   samtools view

2. Sort and index newly created BAM file

   samtools sort

   samtools index

3. Mapping Statistics

   samtools flagstat

   samtools idxstats

Assess Mapping Results - RNASEQC

Transcript-associated Reads

Sample	Note	Intragenic Rate	Exonic Rate	Intronic Rate	Intergenic Rate	Expression Profiling Efficiency	Transcripts Detected	Genes Detected
K-562	v1.0 dUTP|Cell Line	0.897	0.538	0.359	0.103	0.411	79,585	18,663
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	0.888	0.446	0.442	0.111	0.327	87,101	20,970
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	0.907	0.464	0.443	0.092	0.276	90,362	21,217

Coverage Metrics for Bottom 1000 Expressed Transcripts

The metrics in this table are calculated across the transcripts that were determined to have the highest expression levels.

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'100Base Norm	No. Covered 3'	3' 100Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
K-562	v1.0 dUTP|Fibroblast	7.17	0.84	739	0.90	791	0.833	2204	230166	15.6
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	5.35	0.75	742	0.68	836	0.954	2403	207728	13.8
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	4.60	0.77	713	0.69	788	0.843	2792	227526	14.7

It is important to note that these values are restricted to the bottom 1000 expressed transcripts. **5'** and **3'** values are per-base coverage averaged across all top transcripts. 5' and 3' ends are 100 base pairs. Gap % is the total cumulative gap length divided by the total cumulative transcript lengths.

Coverage Metrics for Middle 1000 Expressed Transcripts

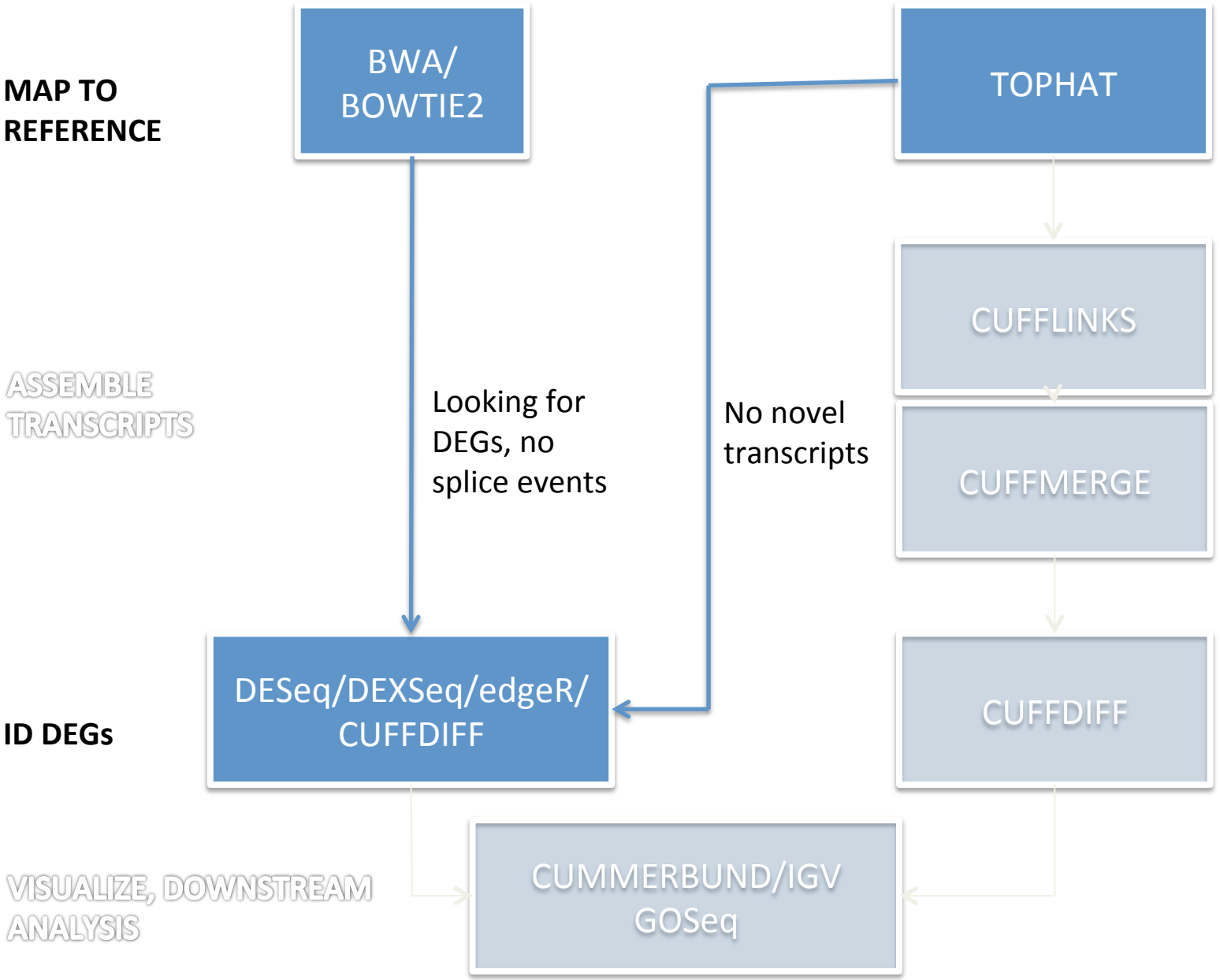
The metrics in this table are calculated across the transcripts that were determined to have the highest expression levels.

Sample	Note	Mean Per Base Cov.	Mean CV	No. Covered 5'	5'100Base Norm	No. Covered 3'	3' 100Base Norm	Num. Gaps	Cumul. Gap Length	Gap %
K-562	v1.0 dUTP|Fibroblast	24.42	0.62	863	0.79	890	0.787	1045	83828	4.3
GTEX-N7MS-2526	v1.0 dUTP|Brain|9.638445	14.61	0.61	854	0.59	943	0.949	972	69905	3.5
GTEX-N7MT-0126	v1.0 dUTP|Lung|9.074045	11.90	0.63	852	0.63	877	0.841	1316	90803	4.5

Mapping Summary

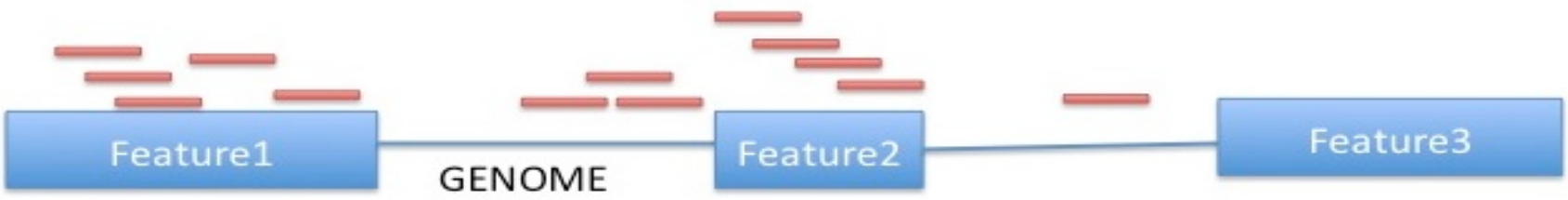
- Unspliced mappers (BWA, bowtie2) ok when mapping to the transcriptome.
- Spliced mappers (tophat, STAR) are good for mapping to the genome.
- Samtools can be used to gather basic mapping statistics, RNASEQC for RNA specific statistics

STEP 4 and 5: Quantify Expression and ID DEGs

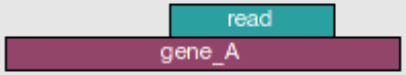
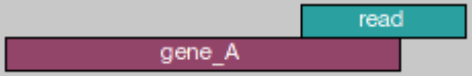
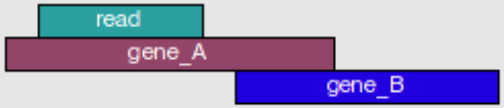
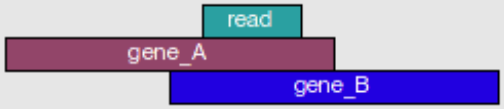


STEP 4: Quantify Expression

- Bedtools
  - **Bedtools multicov :** Takes a feature file (GFF) and counts how many reads in the mapped output file (BAM) overlap the features.
  - Remember that the chromosome names in your gff file should match the chromosome names in the reference fasta file used in the mapping step.



STEP 4 : Quantify Expression

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTSeq –

– Gives you fine grained control over how to count genes, especially when a read overlaps more than one gene/feature.

STEP 5: ID Differentially Expressed Genes

- Normalize gene counts
- Represent the gene counts by a distribution that defines the relation between mean and variance.
- Perform statistical test to compare this distribution between conditions.
- Provide fold change, P-value information, false discovery rate for each gene.

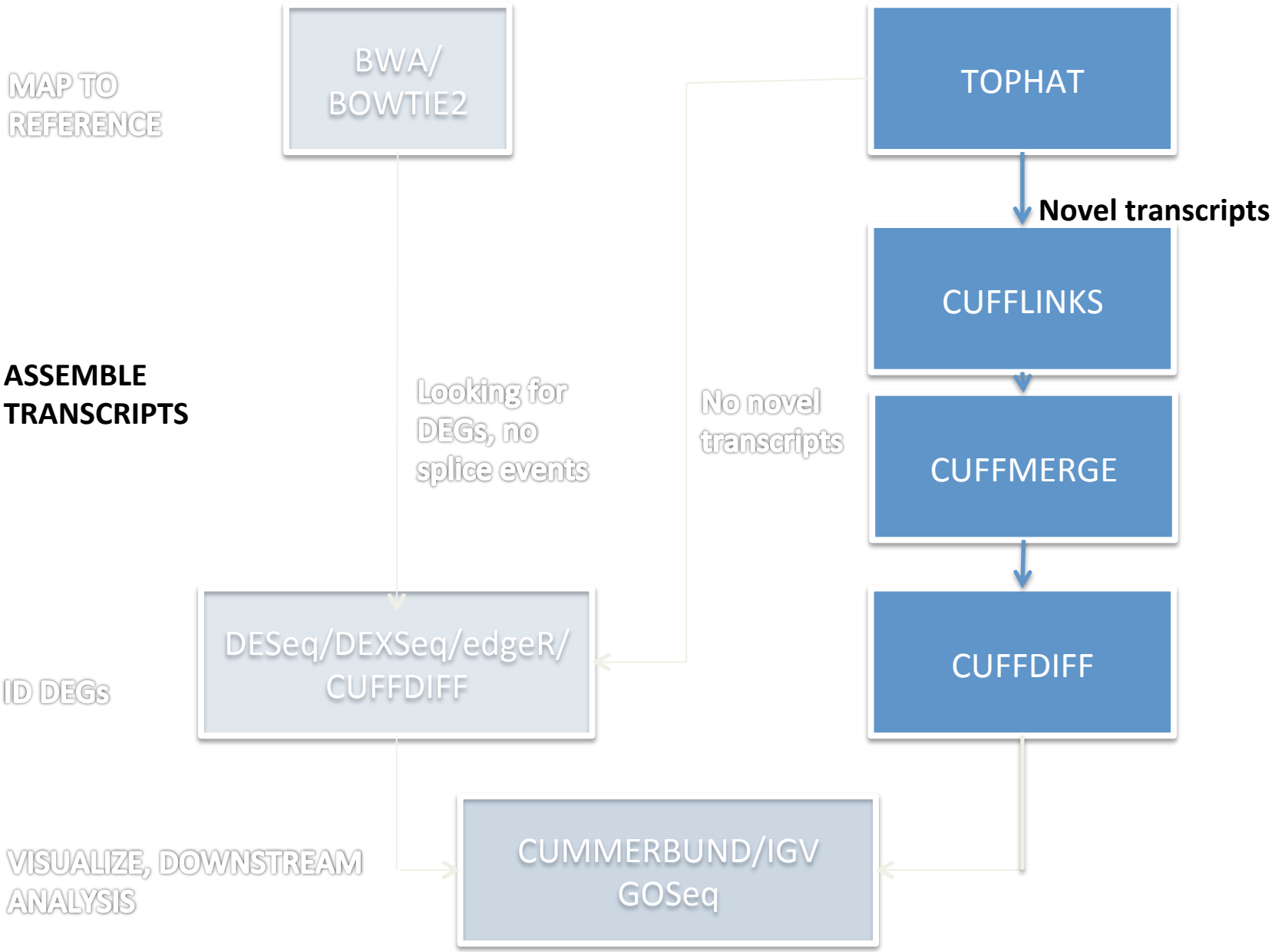
STEP 5: ID Differentially Expressed Genes

	DESeq2	edgeR	DEXSeq	Cuffdiff
Normalization	Median scaling size factor	Median scaling size factor/TMM	Median scaling size factor	FPKM, but also has provisions for others
Distribution	Negative binomial	Negative binomial	Negative binomial	Negative binomial
Statistical Test	Negative binomial test	Fisher exact test	Modified T test	T test
Advantages	Straightforward, fast, good with small number of replicates. Allows for complicated study designs, with multiple factors.	Straightforward, fast, good with small number of replicates.	Good for identifying exon-usage changes	Good for identifying isoform-level changes, splicing changes, promotor changes. Not as straightforward

STEP 5: ID Differentially Expressed Genes

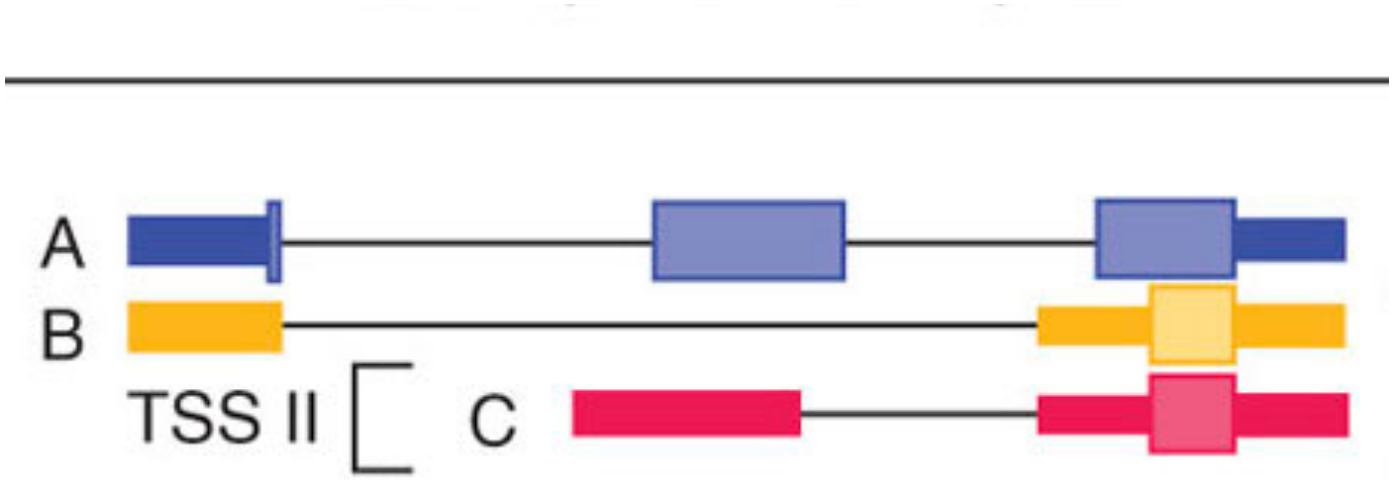
- **DESeq2** Input: **RAW** count data , with each column representing a biological replicate/condition.
- DESeq2 R commands available at: https://wikis.utexas.edu/display/bioiteam/Testing+for+differential+expression
- Let's look at bedtools, htseq, DESeq2 results for now.
- **Cuffdiff** covered further down the line.

STEP 3: Assemble Transcripts



What is a gene? What is a transcript?

A gene can have multiple transcripts!



- We want to identify all these transcripts, whether annotated or not.

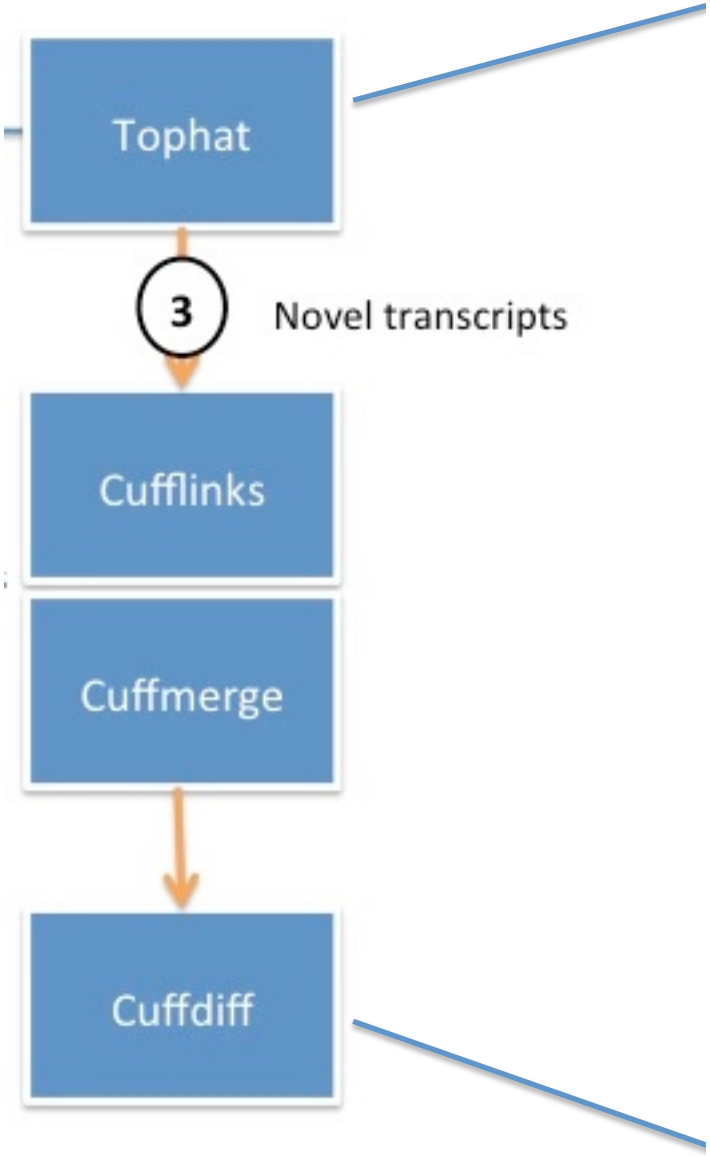
Why transcript assembly?

Transcript assembly = assembly of mapped reads into transcriptional units.

Why?

- Define a precise map of all transcripts expressed in a sample.
- How does our transcriptome look in comparison to the known transcriptome?
- Look for novel transcripts between conditions/ samples.
- Look for differences in expression for these novel transcripts between conditions/samples.

TUXEDO PIPELINE



TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

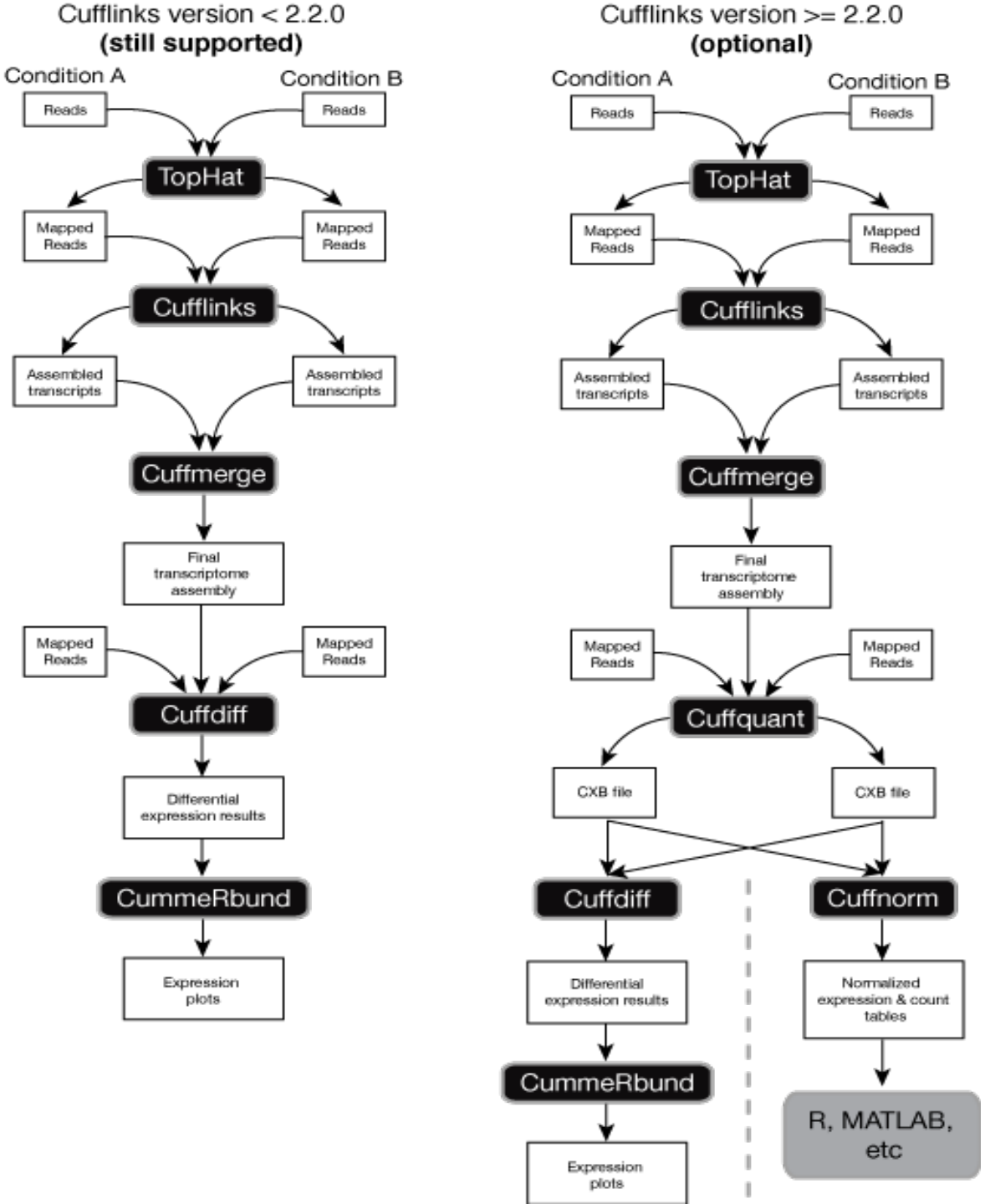
Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use



The pipeline is sequential.

Output of one step becomes input of next step.

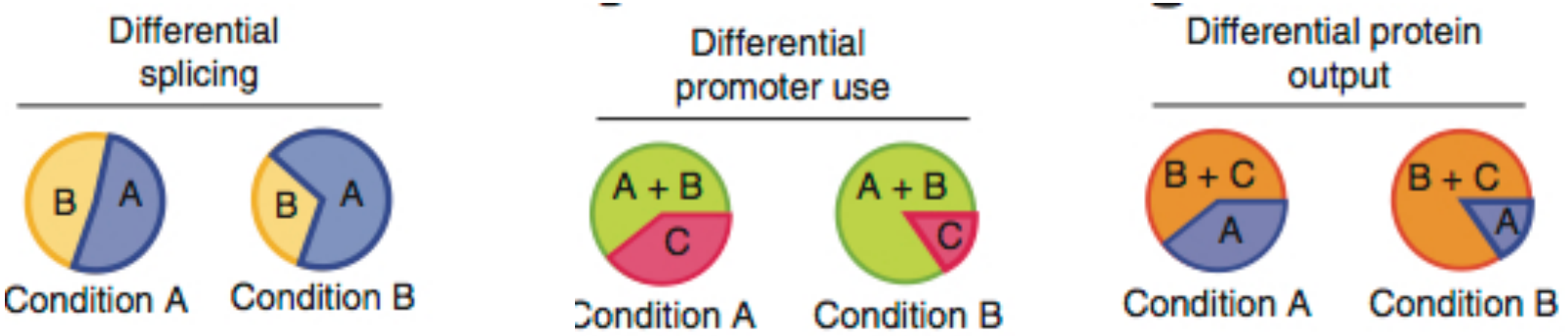
Figure from:
Trapnell et al, Nature protocols, 2012.

What do we get at the end of running this pipeline?

A view of how the transcriptome is different between condition C1 and condition C2

- Both in terms of annotated genes and transcripts.
- And novel genes and transcripts

Differential gene expression and so much more…



A. TOPHAT

Tophat maps your data to your reference in a splice-aware manner, that will also identify junctions. We've already looked at to run it.

Output: Mapped output in bam format

B. CUFFLINKS

Reconstructs/assembles transcript for each sample.

Why is transcript assembly hard?

Difficult to tell which read came from which transcript

- Many short reads, many transcripts!
- Transcripts are expressed in different amounts. So, coverage of reads can be vastly different.
- Reads can come from mature mRNA (exons only) and precursor RNA (containing partial introns).

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

Figure :
Garber et al, Nature Methods, 2011

Most commonly used, if you have a genome.

Less resource-intensive

We'll call this coverage islands method

Transcriptome reconstruction				
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms	
		Cufflinks[29]	Reports a minimal set of isoforms	
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome
		TransABySS[56]		

If you don't have a genome.
If you believe your sample has major rearrangements

More CPU and RAM intensive

Figure :
Garber et al, Nature Methods, 2011

Genome guided transcript assesmbly

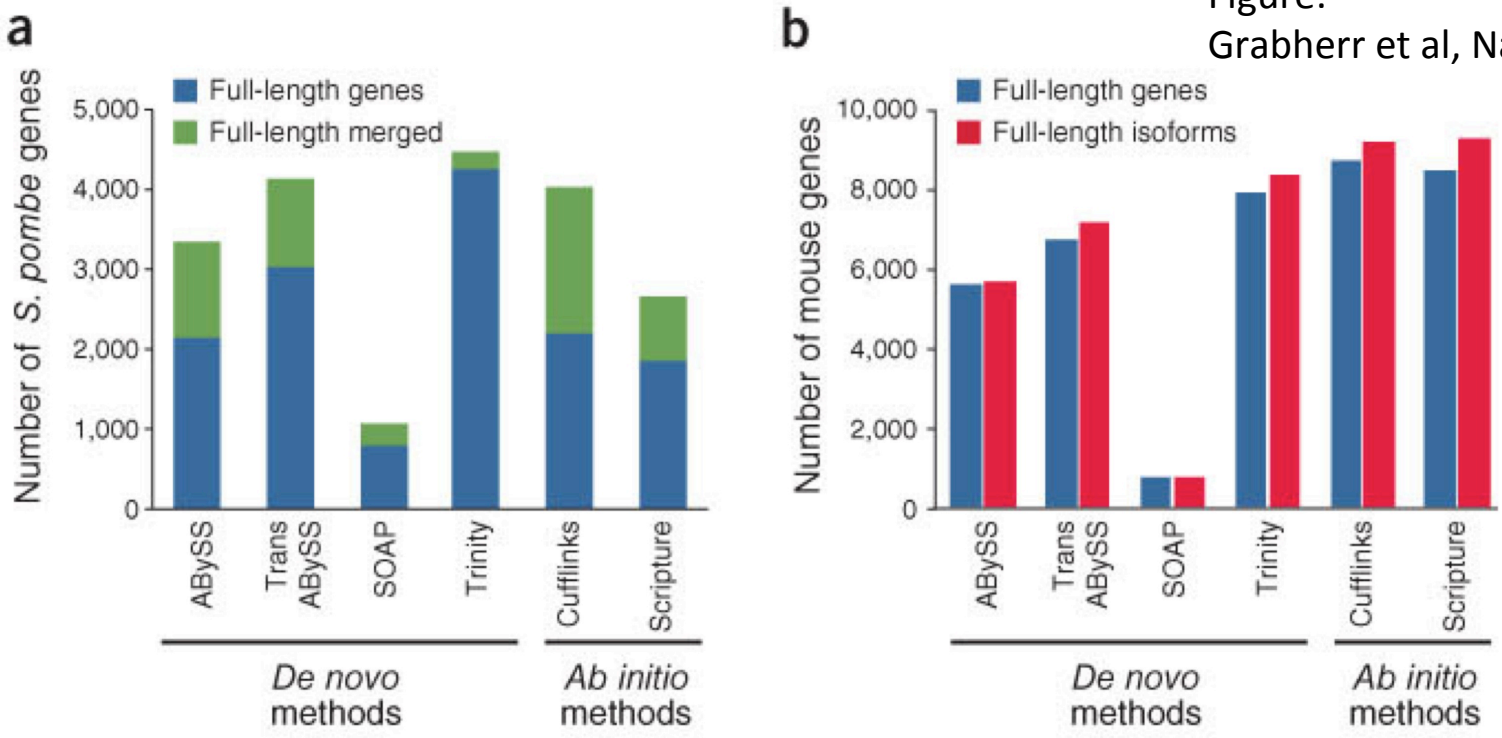
Different assembly methods

- **Coverage islands**
  - ID putative exons by looking for coverage islands.
  - Older method, were meant for shorter read lengths.
  - G.MorSe

- **Exon first approach**
  - Directly uses mappings of spliced reads to reconstruct transcriptome.
  - Uses graph topology.
  - **Cufflinks (part of tuxedo suite),** scripture

How do these tools compare?

Figure:
Grabherr et al, Nature Biotech, 2011



Program combination	*vs.* orthology annotation		*vs.* EST annotation	
	Base-level accuracy (%)[1]	Confirmed junctions (%)[1]	Base-level accuracy (%)[1]	Confirmed junctions (%)[1]
TopHat + Cufflinks	83.9	75.8	68.9	63.0
GSNAP + Cufflinks	79.4	71.2	65.7	58.4
GSNAP + Cufflinks (subsample[2])	80.3	72.7	60.2	66.3
TopHat + Scripture	70.3	67.9	60.8	62.5

[1]Base level accuracy and percentage of confirmed junctions with different combinations of mapper and assembler on the sample ps94 males compared to the orthology annotation and the EST annotation ([2]based on 48 M reads).
doi:10.1371/journal.pone.0046415.t001

Figure:
Palmieri et al, PLOS One, 2012

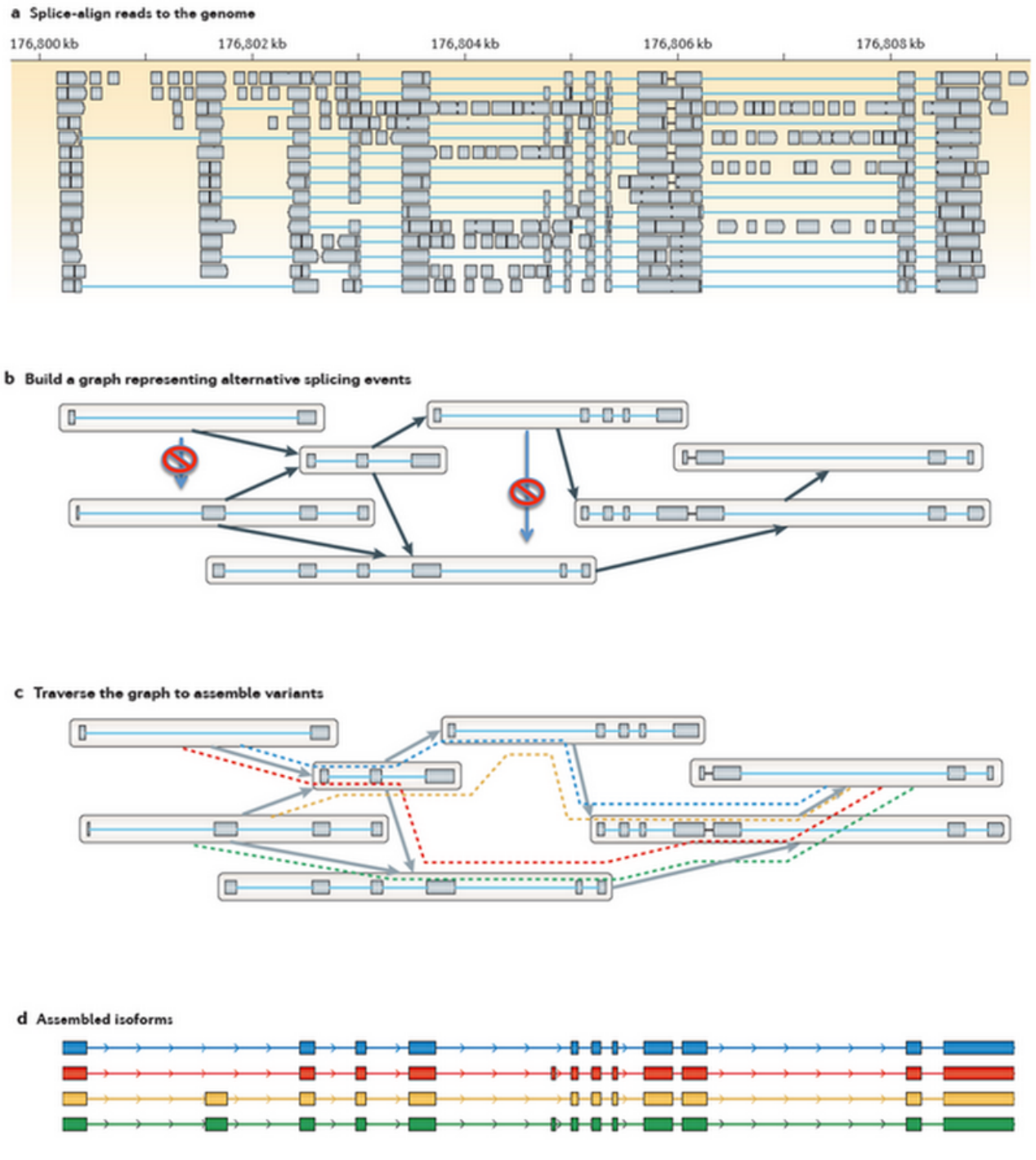
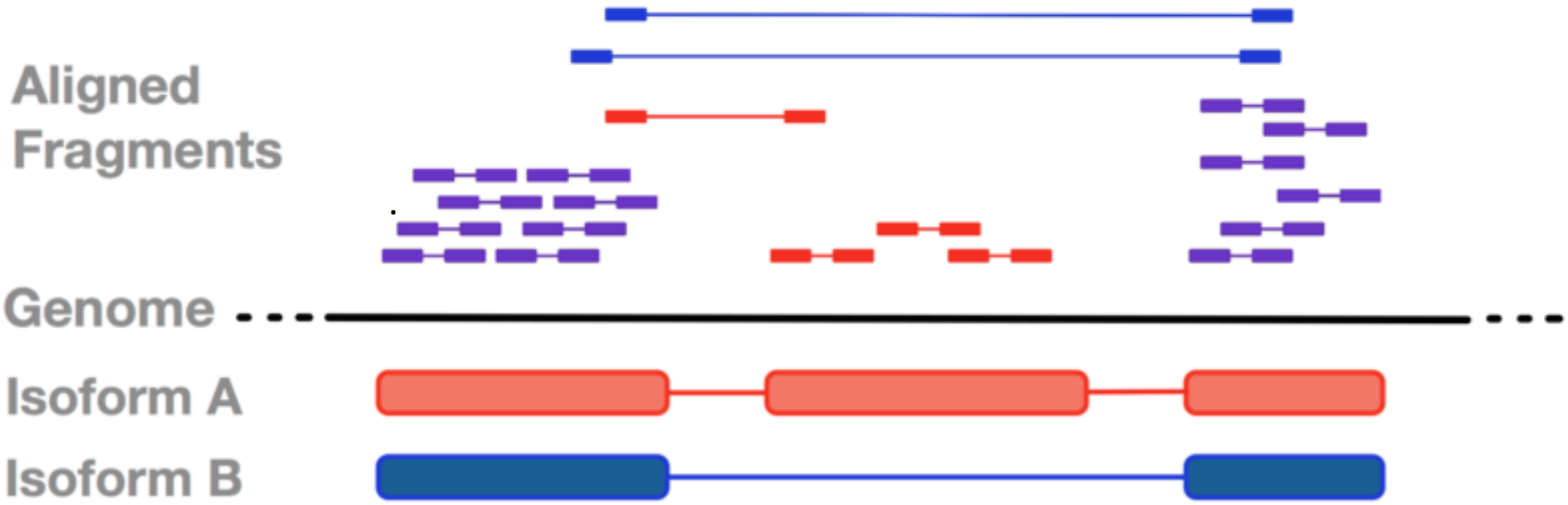


Figure : http://sourceforge.net/projects/trinityrnaseq/files/misc/RNASEQ_WORKSHOP/rnaseq_workshop_slides.pdf

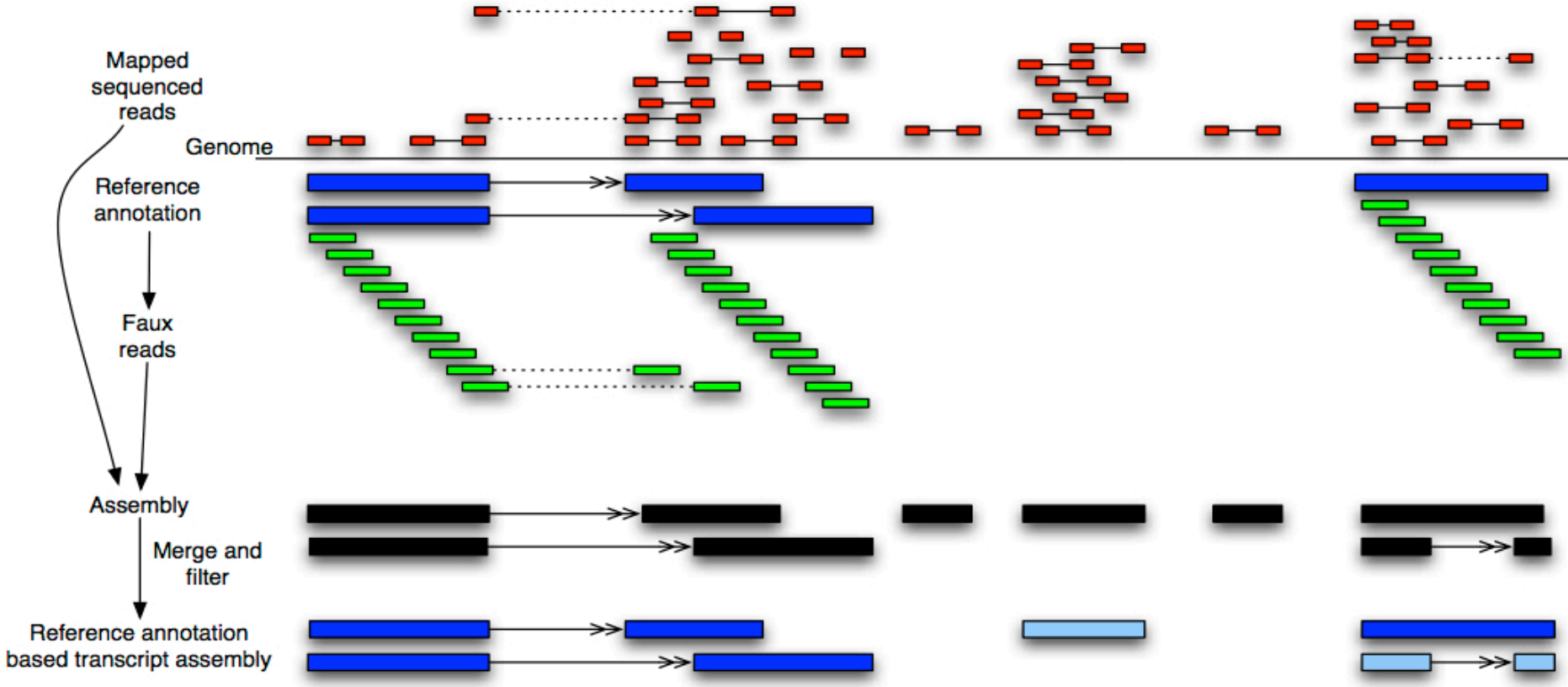
How does Cufflinks do transcript assembly

Exon first method!



RABT

- Reference annotation based transcript assembly (RABT)
  - Uses existing annotation to guide assembly of transcripts.



After assembly

- Calculates abundance for these assembled transcripts.

- Normalized using FPKM (Fragments Per Kilobase of Exon Per Million) (variation of RPKM)
  - RPKM normalizes for **transcript length variations** and **sequencing depth**.
  - RPKM= (No.of Mapped reads*10^9)/ (**length of transcript *total no.of reads**)
  - FPKM just exchanges reads with fragments.

General syntax for cufflinks command

cufflinks [options] <accepted_hits.bam>

Some of the important options:

-p/--num-threads

-g/--GTF-guide (both annotated and novel transcripts)

-b/--frag-bias-correct

-u/--multi-read-correct

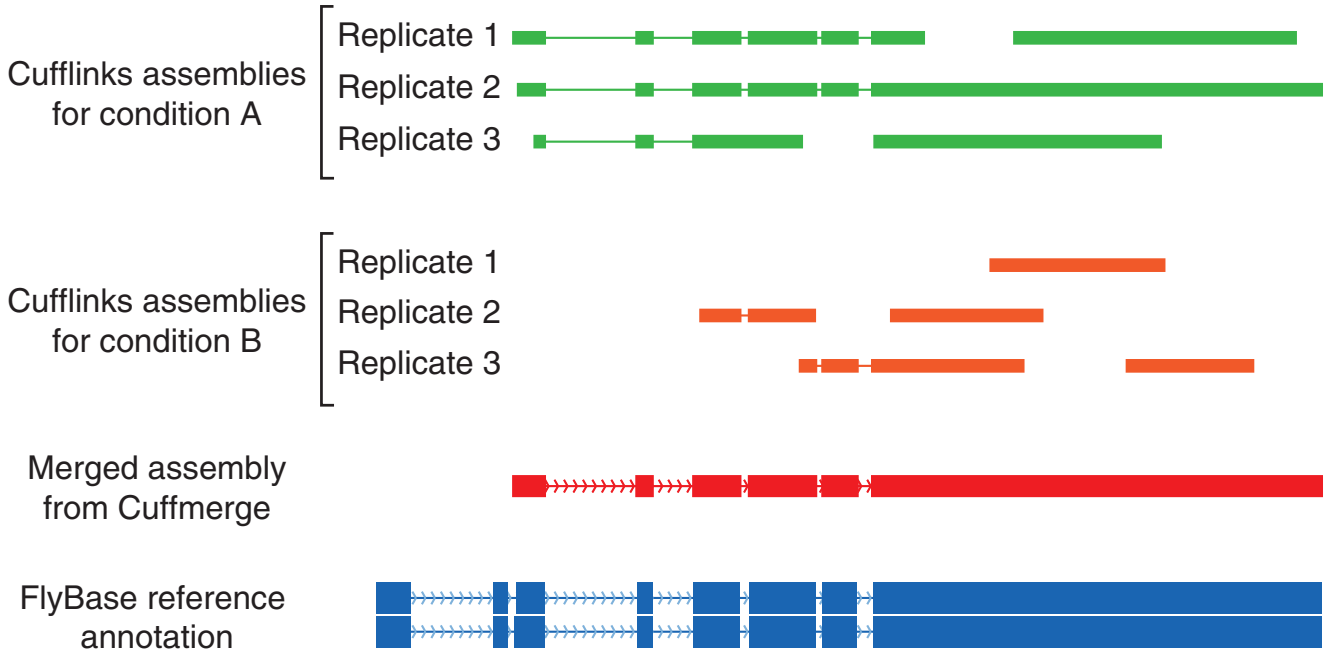
Let's look at some results from a cufflinks transcript assembly

- Input:
  - Tophat mapped results (bam files)
  - Transcriptome annotation (genes.gtf)

- Let's look at the wiki and the output files.

C. CUFFMERGE

- Cuffmerge is used to merge all the transcripts that cufflinks assembled into one file.



C. CUFFMERGE

- Input: All cufflinks assembly files (in gtf format)
- Output: merged.gtf
  - Your very own gtf file, containing all the transcripts found in your samples (both novel and otherwise).
  - Also information about how the novel transcripts relate to the known transcripts
- SWITCH TO THE WIKI for instructions on viewing these results

D. CUFFQUANT

- Given alignment output files and an annotation file, quantifies isoform expression.
  - Don't care about novels? Just provide an existing annotation (gtf) file
  - Care about novels? Just provide a cufflinks/ cuffmerge assembled (gtf) file

E. CUFFDIFF

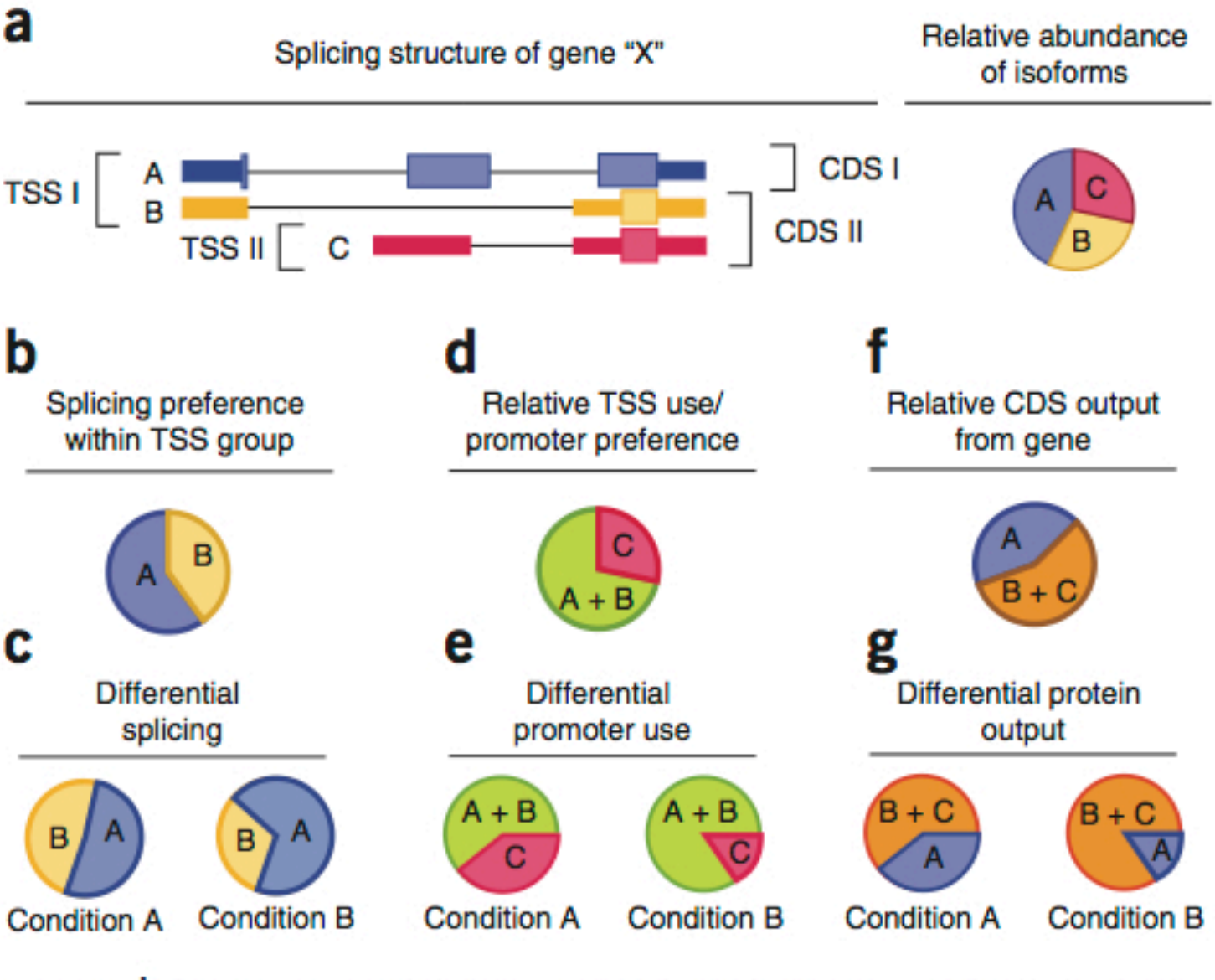
- Calculates differential expression!

- Input:
  - Our newly created merged.gtf file or A gtf file we downloaded (genes.gtf)
  - Mapped bam files/cuffquant quantification output

- Calculates difference in isoform-level expression among conditions.
- If the chance of seeing this difference is small enough under the chosen statistical model, it is deemed signficantly differentially expressed.

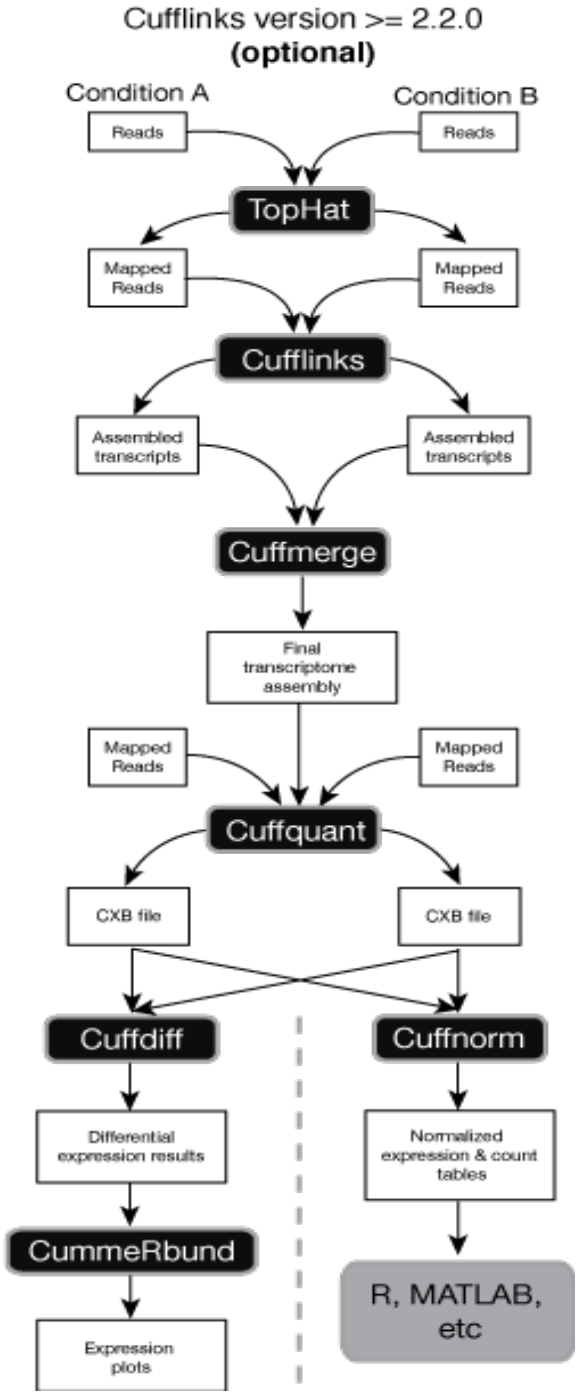
E. CUFFDIFF

Figure from: Differential analysis of gene regulation at transcript resolution with rNA-seq, Trapnell et al, Nature Biotechnology, 2013



E. CUFFDIFF

- SWITCH TO THE WIKI for instructions on viewing these results



OPTIONAL STEP
D2. CUFFNORM

- Generates normalized tables of expression values.
- Ready to take outside the tuxedo pipeline for any further analysis

DESeq/edgeR output vs Tuxedo pipeline output

- We generated differential expressed genes using DESeq too. So, why the big fuss?
  - They were all from annotated genes. So, they all has flybase ids.
  - Now our output has genes with ids 'CUFF…' - they are novel.
  - In addition to differential gene expression, we also have results for differential regulation.
  - We also have results telling us where our novel transcripts are with respect to the annotated ones.

Limitations of the Tuxedo Pipeline

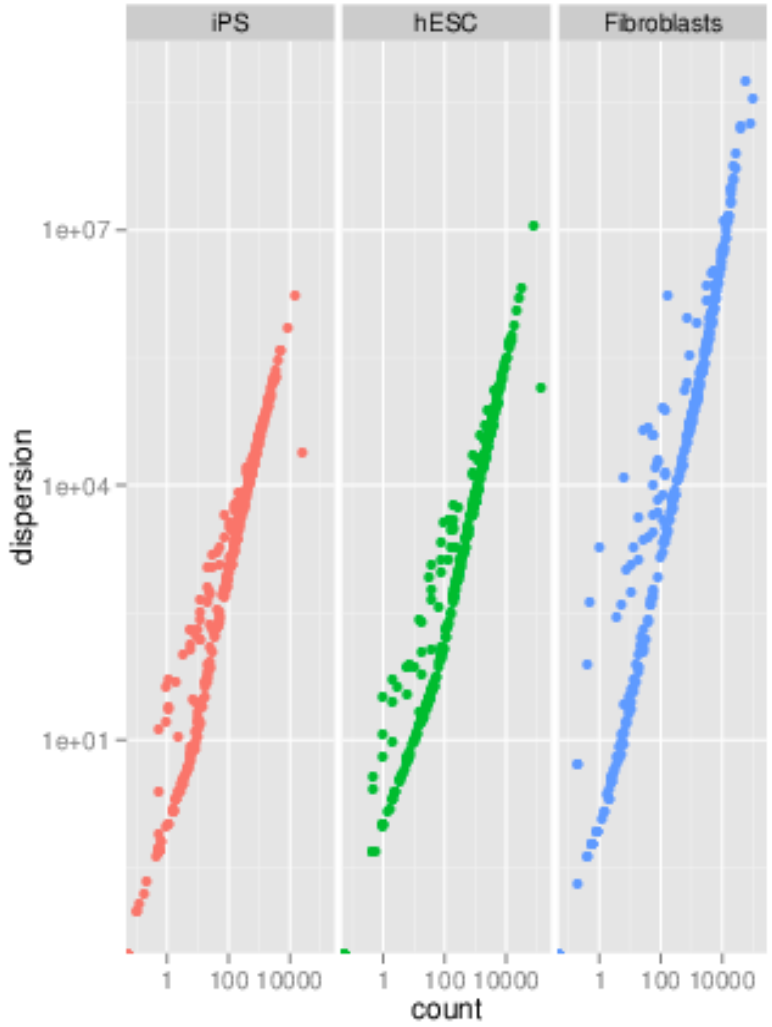
- A Reference is needed.
- Kind of a black box.
- Not quick.
  - Step 1, align the RNA-seq reads to the genome: ~6 h
  - Steps 2–4, assemble expressed genes and transcripts: ~6 h
  - Step 5, identify differentially expressed genes and transcripts:~6 h

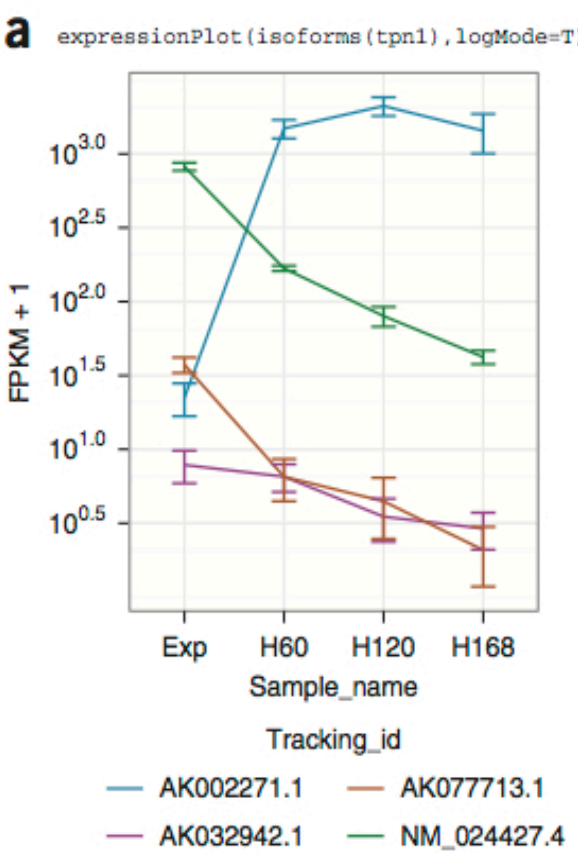
If you don't have a genome:

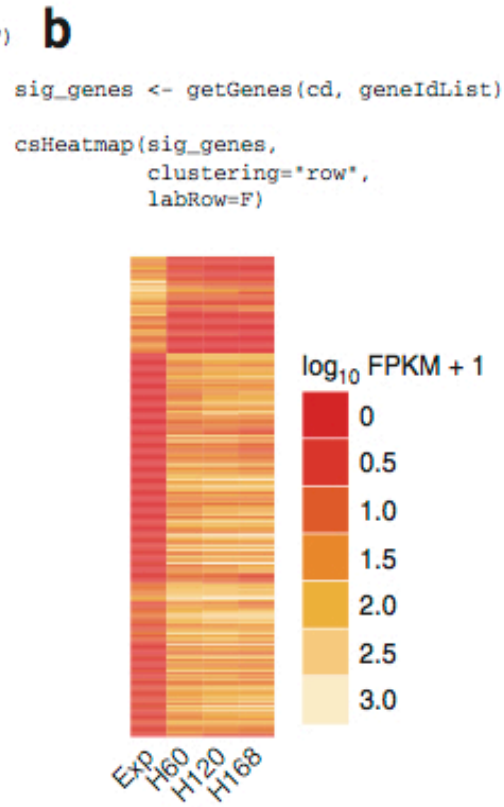
- De novo transcriptome assembly using trinity.
- Map your data to this to calculate gene expression changes.

STEP 6: Visualize and Perfom Other Downstream Analysis

- Visualize using Cummerbund

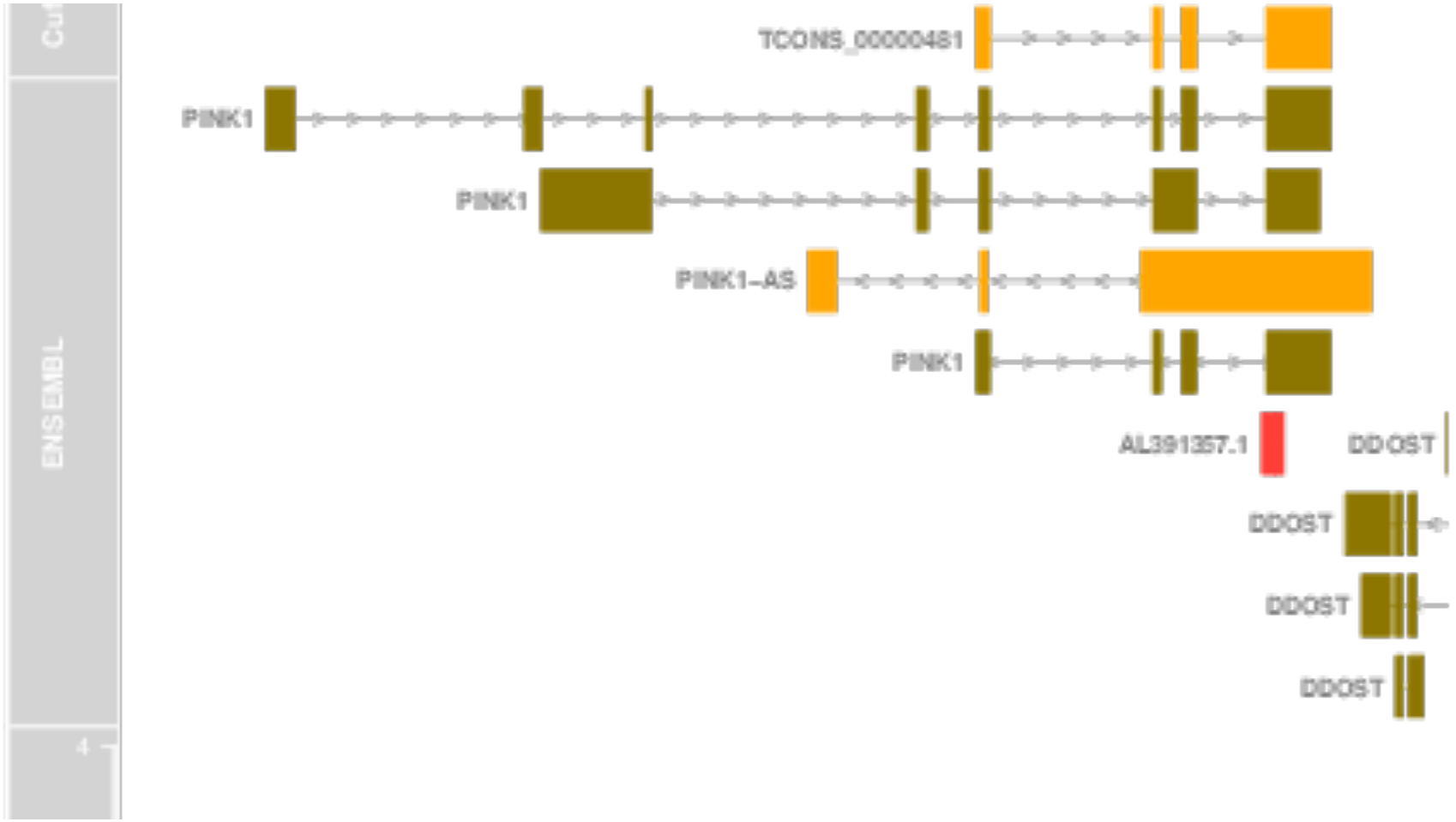






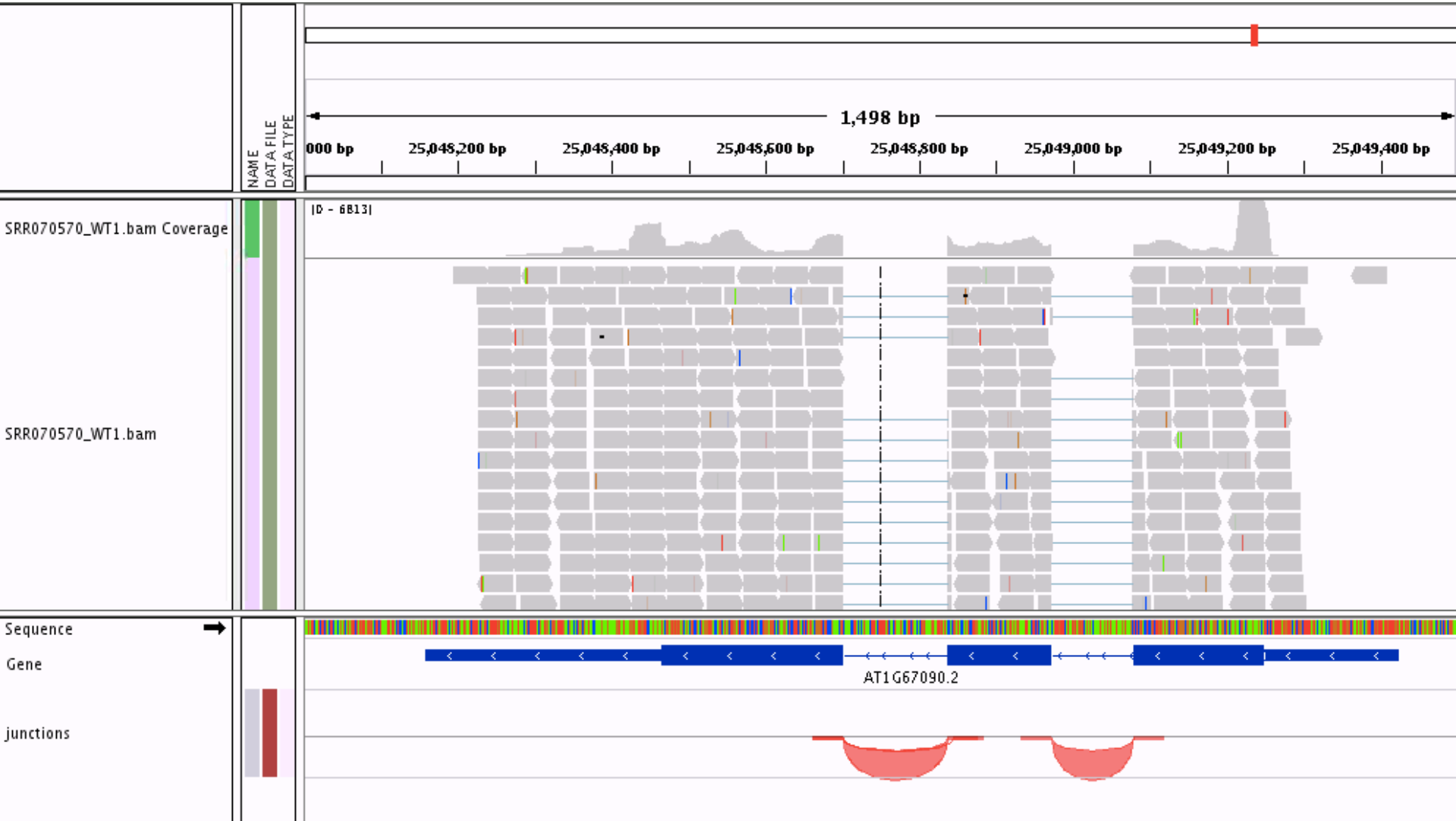
STEP 6: Visualize and Perfom Other Downstream Analysis

Visualize using Cummerbund



STEP 6: Visualize and Perfom Other Downstream Analysis

- Visualization using IGV



STEP 6: Perfom Other Downstream Analysis

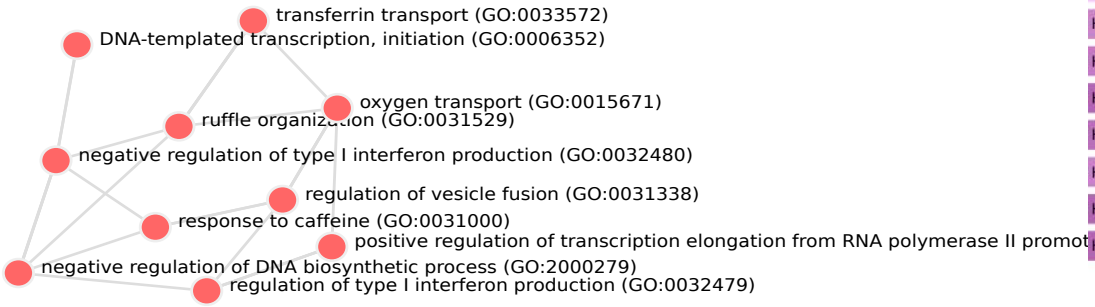
- ID enriched gene ontology (GO) terms in our DEGs using GOSeq
- Commands and the examples on the wiki.
- For GO enrichment, we take the following things into account:
  A. Total number of genes we are looking at.
  B. Number of genes of interest, that is, in our DEG list.
  C. Total number of genes in the GO term
  D. Number of genes from our genes of interest that are also in the GO term.

If the number of genes from our list that belong to GO term (D) is significant compared to the total number of genes in that GO term (C) and the total number of genes in our experiment (A), we consider that GO term to be enriched in our data.

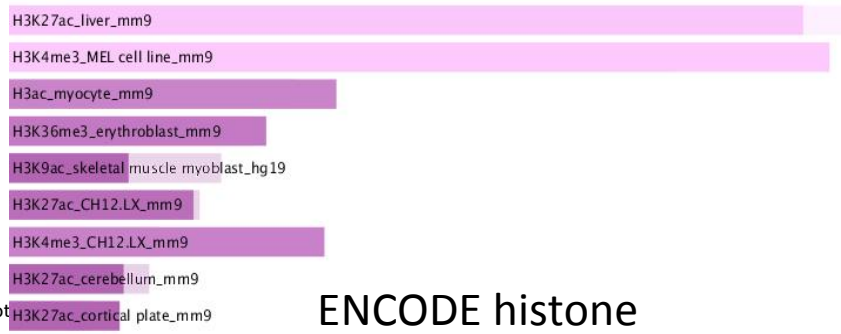
STEP 6: Perfom Other Downstream Analysis

- Enrichr- GUI for GO/pathway enrichment analysis

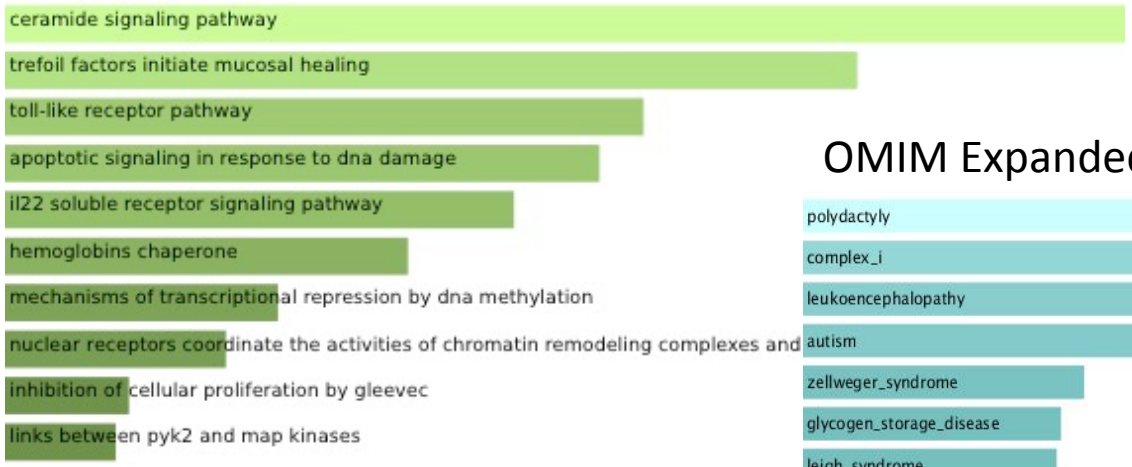
  amp.pharm.mssm.edu/Enrichr



GO terms network graph

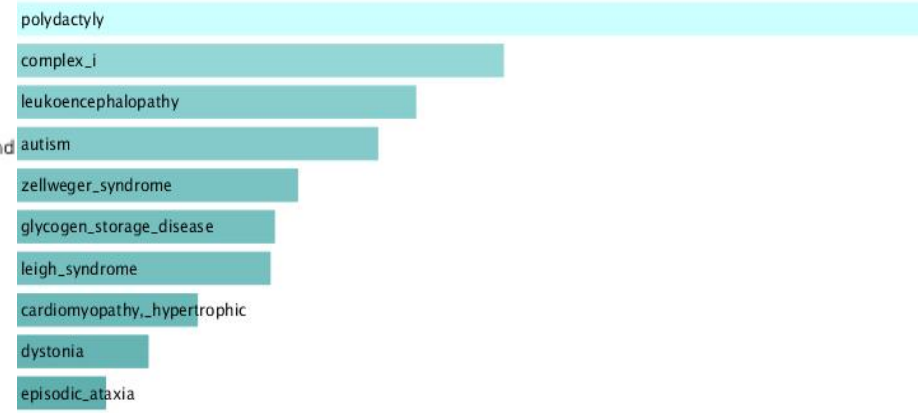


ENCODE histone modifications



Biocarta pathways bar chart

OMIM Expanded terms



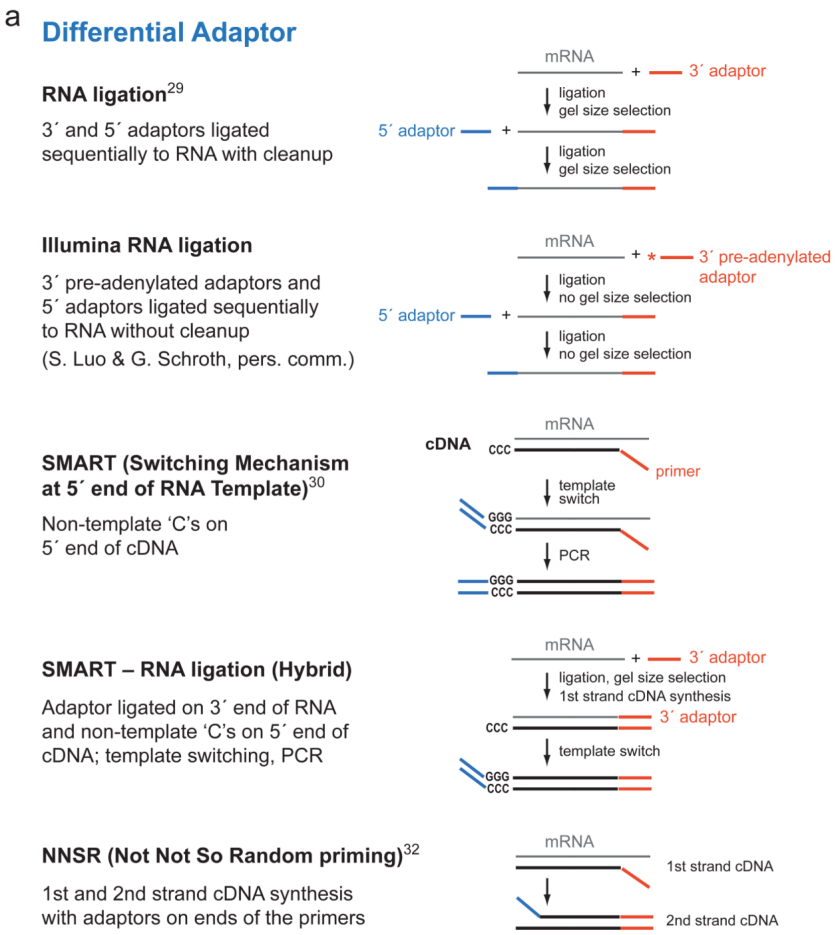
Thank you!

- Visit the Bioinformatics Consultants at GDC
- Come to Byte Club meetings
  - Join UT Lists-bioiteam

APPENDIX: Submitting Jobs to Lonestar

- https://wikis.utexas.edu/display/bioiteam/Submitting+Jobs+to+Lonestar

Appendix



Levin et al.
Page 10

Figure 1. Methods for strand-specific RNA-Seq

Salient details for seven protocols for strand-specific RNA-Seq, differential adaptor methods (a) and differential marking methods (b). mRNA is shown in grey, and cDNA in black. For differential adaptor methods, 5' adaptors are shown in blue, and 3' adaptors in red.

QNAME SRR035022.262
FLAG 163

APPENDIX SAM FILE FLAGS EXPLAINED

The QNAME is the query name. For the FLAG of 163 we transform this into a binary string: 10100011. So accordingly to the flag table:

Flag	Description
0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment) [1]
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped [1]
0x0010	strand of the query (0 for forward; 1 for reverse strand)
0x0020	strand of the mate [1]
0x0040	the read is the first read in a pair [1,2]
0x0080	the read is the second read in a pair [1,2]
0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
0x0200	the read fails platform/vendor quality checks
0x0400	the read is either a PCR duplicate or an optical duplicate

1 the read is paired in sequencing, no matter whether it is mapped in a pair
1 the read is mapped in a proper pair
0 not unmapped

0 mate is not unmapped
0 forward strand
1 mate strand is negative
0 the read is not the first read in a pair
1 the read is the second read in a pair

APPENDIX
Of course, Tuxedo Pipeline can be run without looking for novel events

- NO NOVEL JUNCTIONS: Simple differential gene expression analysis against a set of known transcripts.
  - User provides a gff/gtf file containing annotated features. Quantify only the annotated features and id DEGs.

- NOVEL JUNCTIONS ALSO: In addition to known transcripts, novel transcripts should be explored.
  - User provides a gff/gtf file containing annotated features. But you also allow the search for novel variants as well. Both annotated and novel variants are quantified and DEGs are identified.

- ONLY NOVEL/*DE NOVO* JUNCTIONS: No gff/gtf file is provided. Using just the read data and the genome reference, construct *de novo* transcripts, quantify them and id DEGs.

APPENDIX
Other differential expression tools vs cuffdiff

Others	Cuffdiff
Raw count method for assigning counts to genes	**Isoform deconvolution method** for assigning counts to genes
Count the reads mapping to **exons** of each gene/normalization factor = expression for gene	Count the reads that map to each **isoform** of the gene/normalization factor = expression for gene
If all isoforms of the gene are up/down, works fine	If all isoforms of the gene are up/down, works fine
If some isoforms of the gene are up and some are down, **inaccurate results**	If some isoforms of the gene are up and some are down, **works fine**

APPENDIX

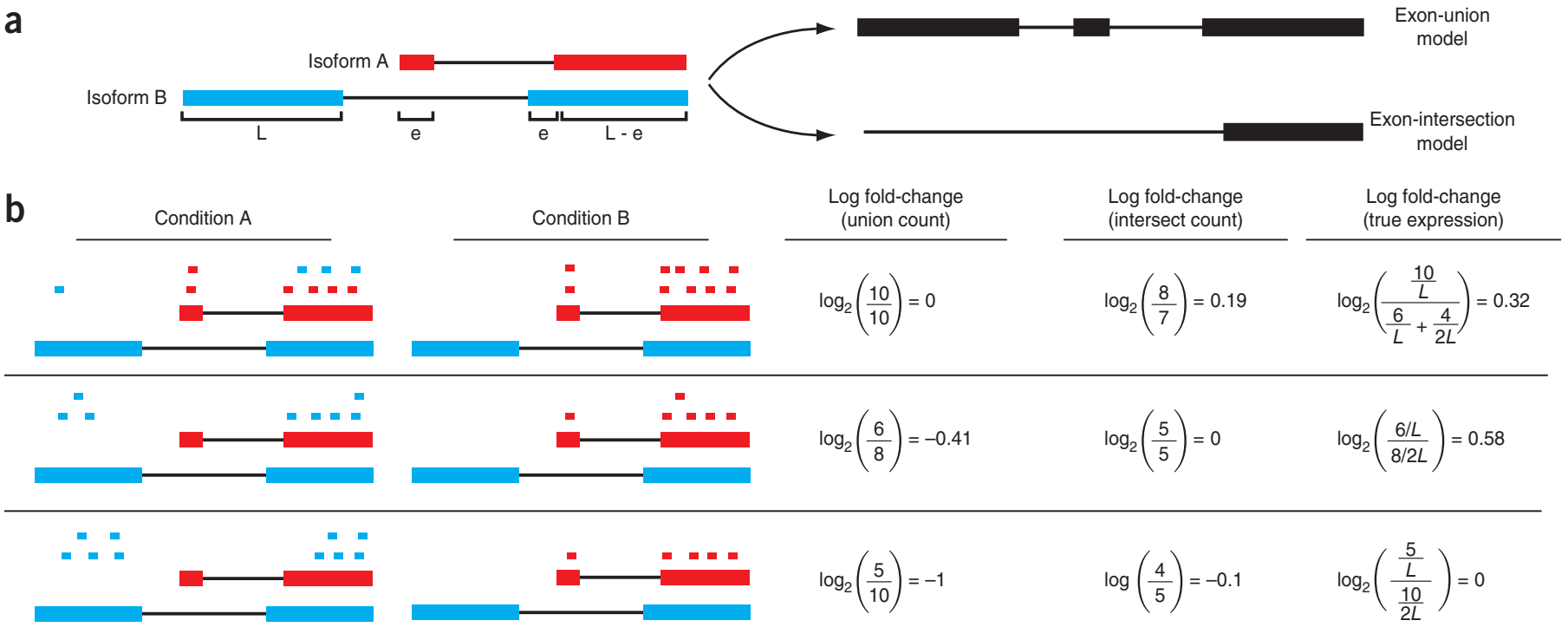


Figure 1 Changes in fragment count for a gene does not necessarily equal a change in expression. (**a**) Simple read-counting schemes sum the fragments incident on a gene's exons. The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads on constitutive exons. (**b**) Both of the exon-union and exon-intersection counting schemes may incorrectly estimate a change in expression in genes with multiple isoforms. The true expression is estimated by the sum of the length-normalized isoform read counts. The discrepancy between a change in the union or intersection count and a change in gene expression is driven by a change in the abundance of the isoforms with respect to one another. In the top row, the gene generates the same number of reads in conditions A and B, but in condition B, all of the reads come from the shorter of the two isoforms, and thus the true expression for the gene is higher in condition B. The intersection count scheme underestimates the true change in gene expression, and the union scheme fails to detect the change entirely. In the middle row, the intersection count fails to detect a change driven by a shift in the dominant isoform for the gene. The union scheme detects a shift in the wrong direction. In the bottom row, the gene's expression is constant, but the isoforms undergo a complete

Figure from: Differential analysis of gene regulation at transcript resolution with rNA-seq, Trapnell et al, Nature Biotechnology, 2013