Introduction to RNA-Seq

Dhivya Arasappan

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

Goals of the Class

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

Logistics

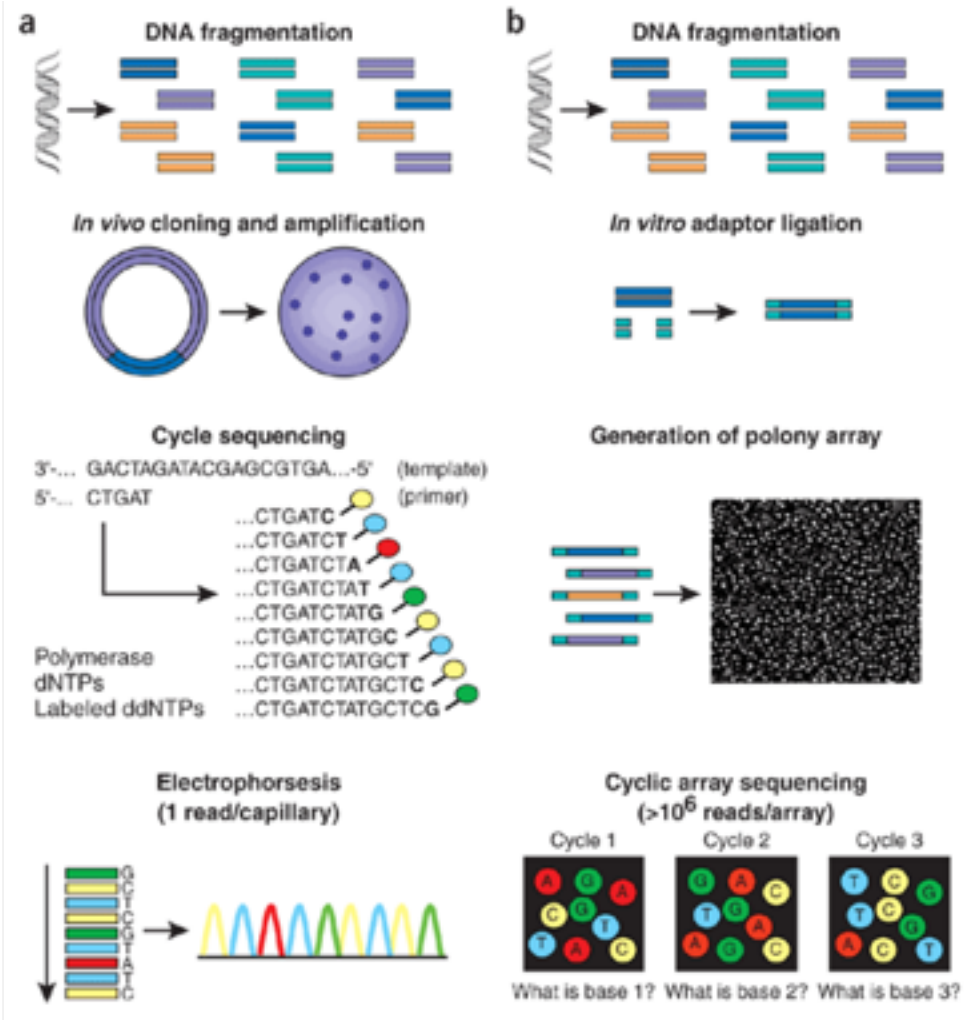
- Commands that I will run today also on BioIteam wiki:
  - https://wikis.utexas.edu/display/bioiteam/Introduction+to+RNA+Seq+Short+Course+Commands

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+2017
- Other Short courses: http://ccbb.biosci.utexas.edu/shortcourses.html
- Bioinformatics consultants

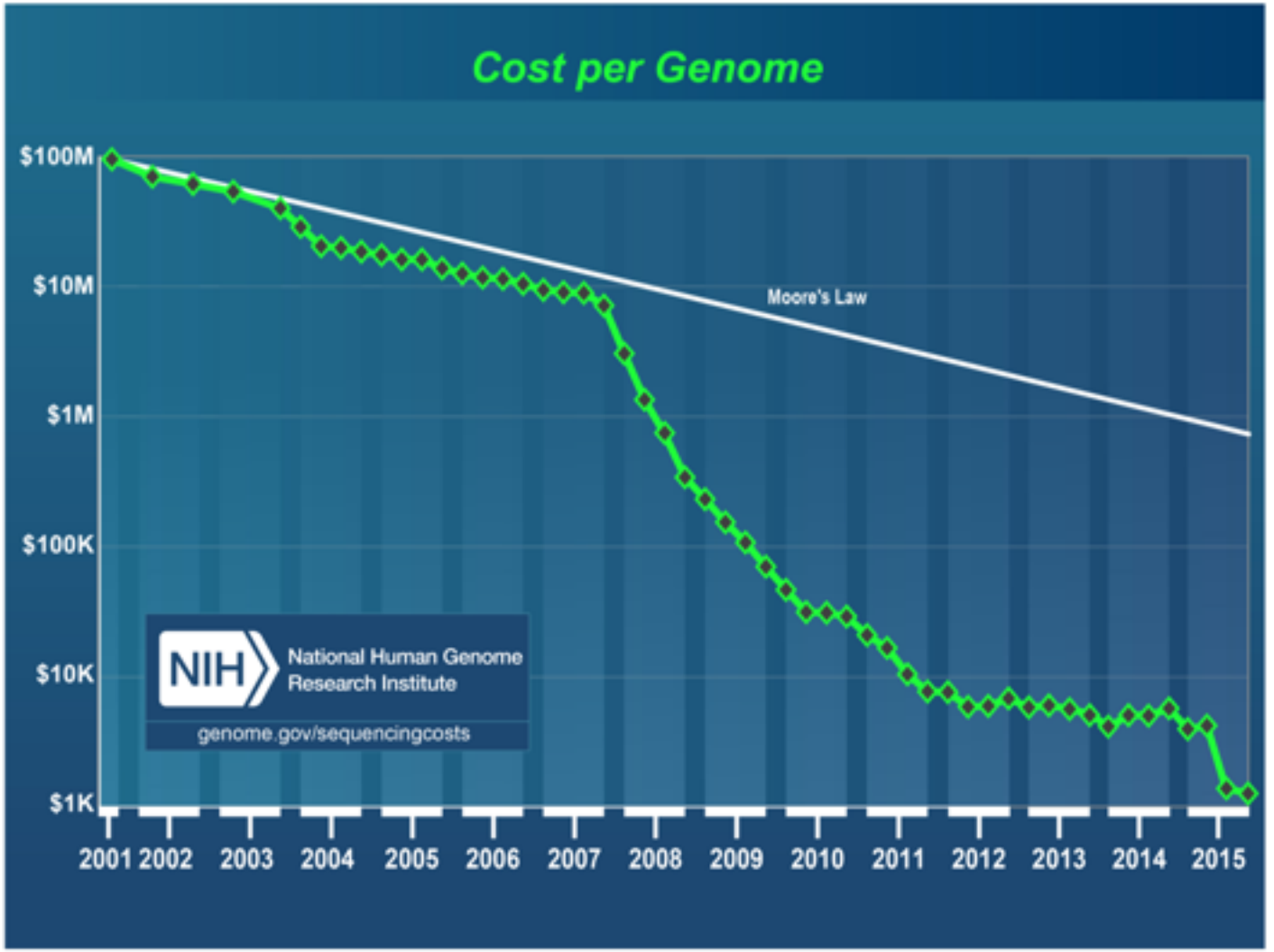
What is Next Generation(or) Second Generation Sequencing?

- Massively parallel sequencing
- The template DNA is attached to a cluster.
- Billions of clusters sequenced in parallel.
- 3-10 billion independent DNA fragments sequenced in one run.



So, what's so great about second generation sequencing?

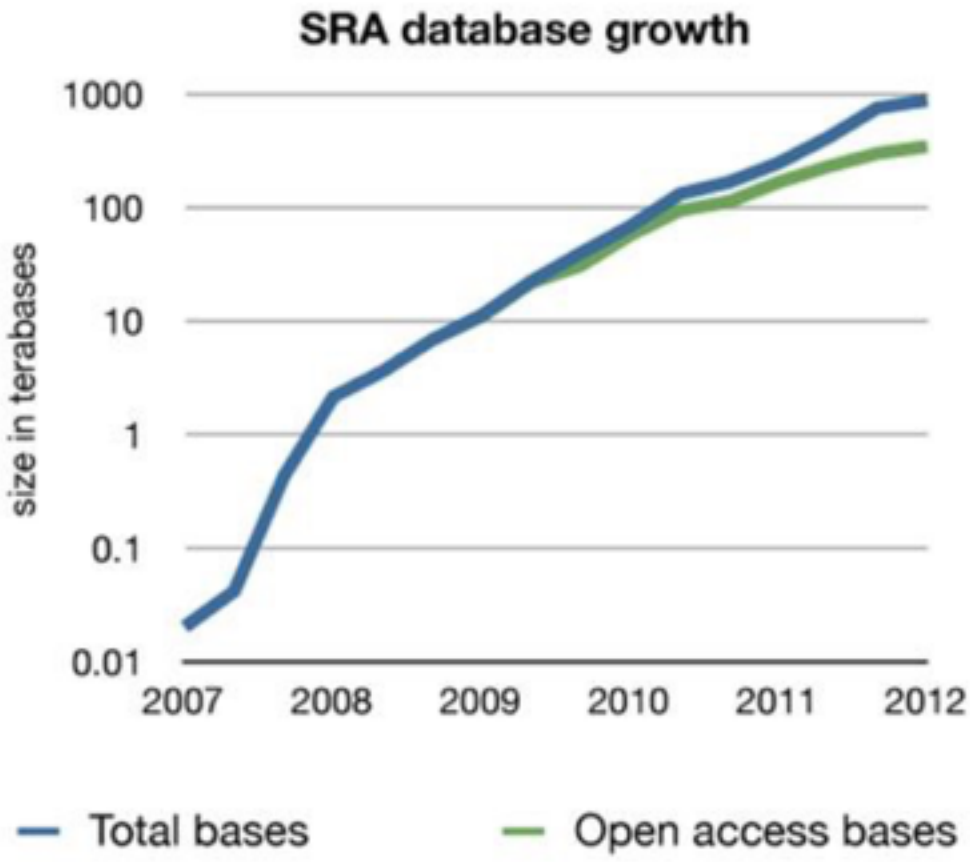
- + Sequence lots more, faster!
- + More cost effective.



So, what's NOT so great about second generation sequencing?

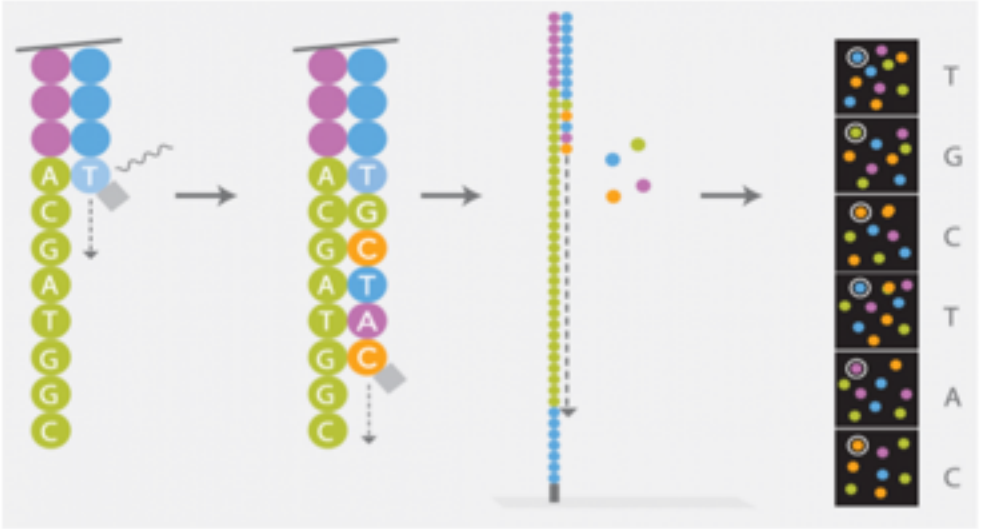
- Data deluge

Data analysis skills are required to make sense of all the data



How does the sequencer work?

- Library prep
- Cluster generation/ amplification
- Sequencing by synthesis
- Done in parallel for billions clusters at once.



http://www.cegat.de/

Different Types of Illumina Sequencers

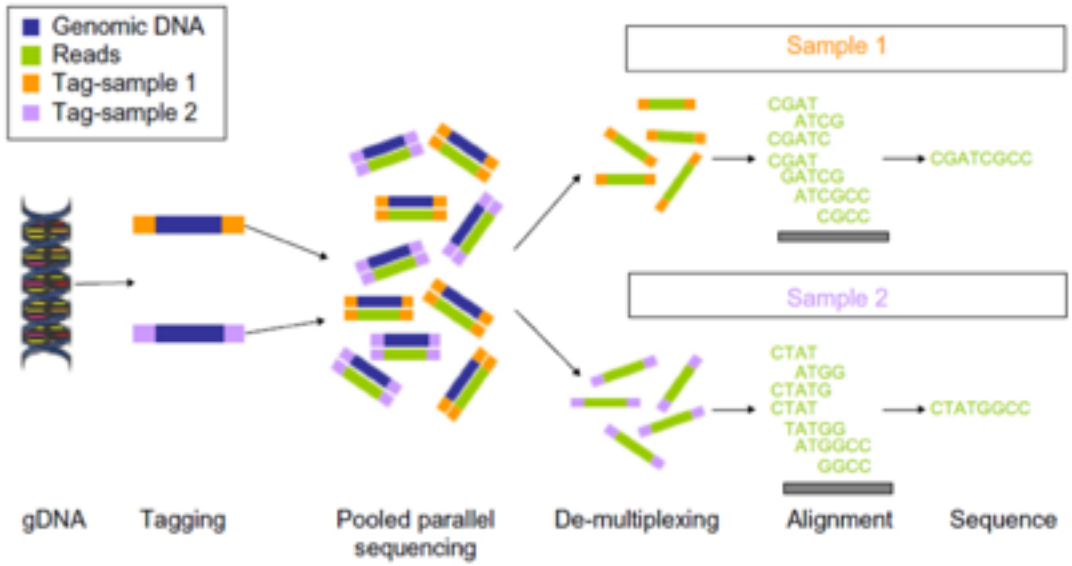
Illumina Specifications Table

	HiSeq X Ten*	Hi Seq 2500			NextSeq 500		MiSeq
		HT v4	HT v3	Rapid	High	Mid	
Total output	1.8 Tb	1 Tb	600 Gb	180 Gb	129 Gb	39 Gb	15 Gb
Run time	3 days	6 days	11 days	40 hrs	29 hrs	26 hrs	~65 hrs
Output/day	600 Gb	167 Gb	55 Gb	~110 gb	~100 Gb	~36 Gb	~5.5 Gb
Read length	2 X 150	2 X 125	2 X 100	2 X 150	2 X 150	2 X 150	2 X 300
# of single reads	6B	4B	3B	600M	400M	130M	25M
Instrument price	$1M*	$740K	$740K	$740K	$250K	$250K	$125K
Run price	~$12k	~$29k	~$26k	~$8k	$4k	?	~$1.4k
$/Gb	$7	$29	$43	$44	$33	?	$93

allseq.com/knowledgebank/sequencing-platfor ms/illumina

Multiplexing

- Sample specific Indexes/ Barcodes are attached to the DNA template.
- 6-8bp indexes/barcodes
- Data off the sequencer must first be demultiplexed to identify which reads belong to which sample.

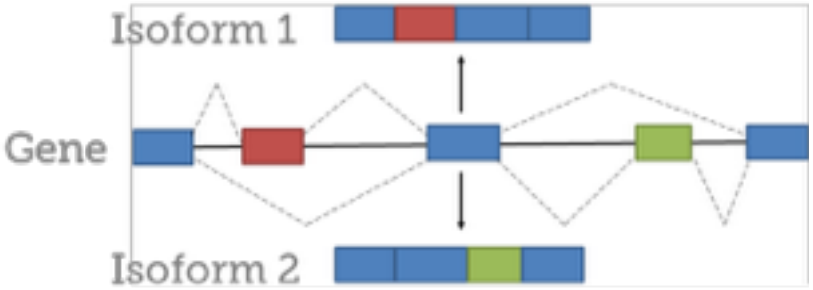


https://doi.org/10.2147/BLCTT.S51503

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

Advantages of RNA-Seq

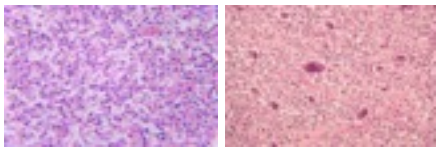
Technology	Tiling microarray	RNA-Seq
Technology specifications		
Principle	Hybridization	High-throughput sequencing
Resolution	From several to 100 bp	Single base
Throughput	High	High
Reliance on genomic sequence	Yes	In some cases
Background noise	High	Low
Application		
Simultaneously map transcribed 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

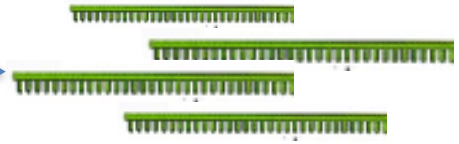
Zhong Wang, **Mark Gerstein**, and **Michael Snyder**

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

RNA-Seq… at it's Most Basic Form

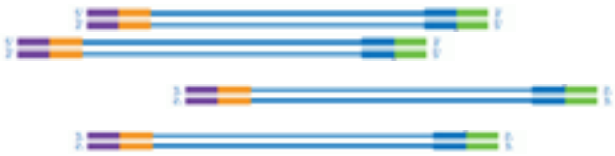


Samples from two conditions



Isolate RNA

Generate cDNA



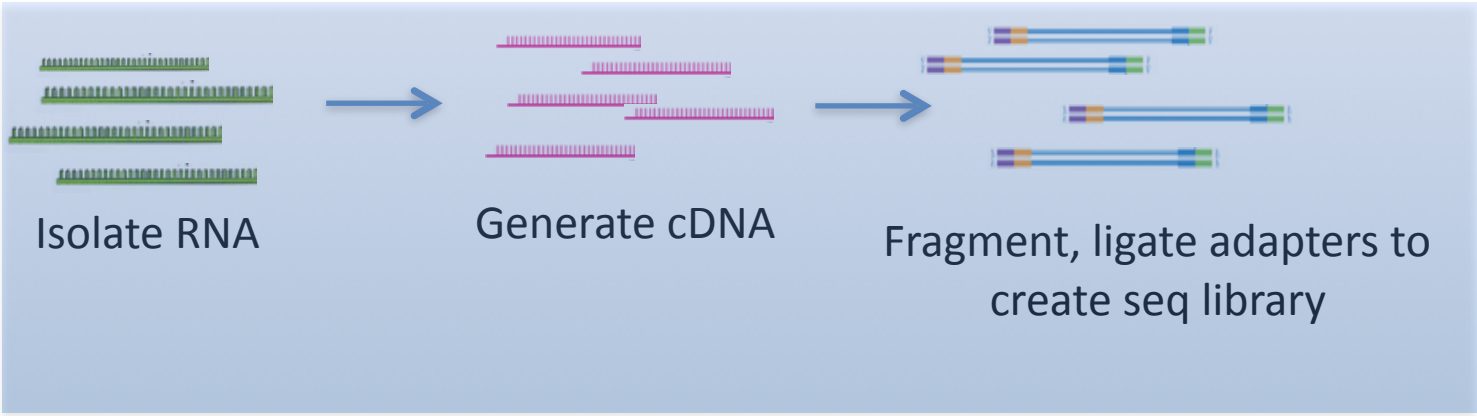
Create sequencing library by size selection and adding adaptors

Run sequencer

Generate short reads

Identify differentially expressed genes

RNA-Seq Libraries… with More Details



A. rRNA Depletion

Ribominus kit

B. Normalized library

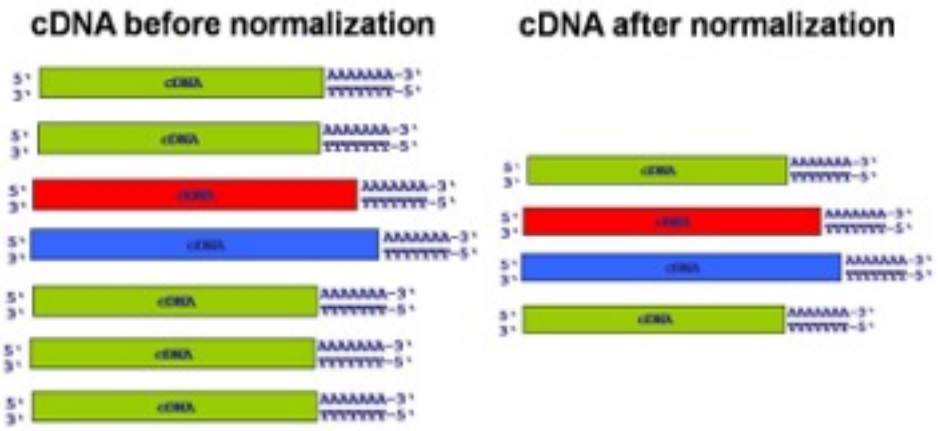
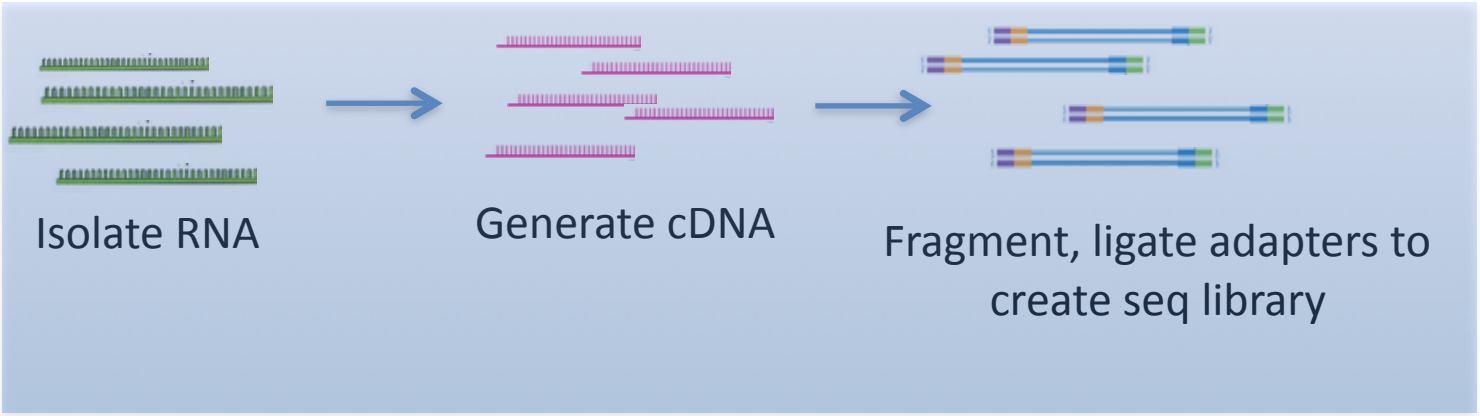


Image from :www.genxpro.info

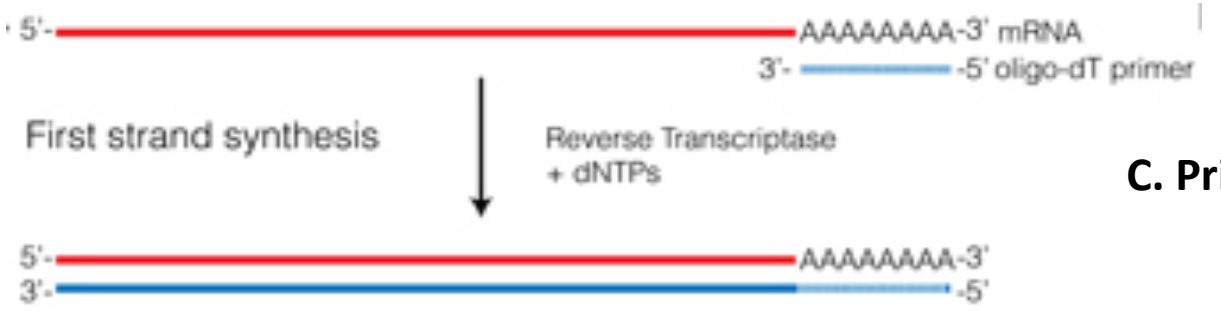
C. Size selection

Reserved for miRNA, siRNA profiling

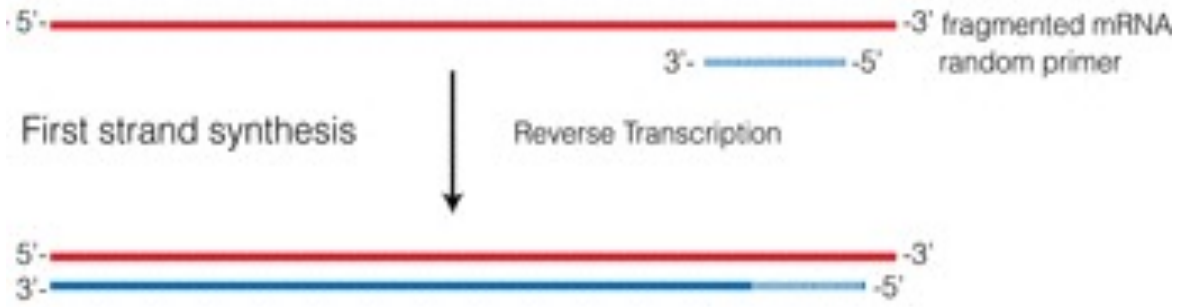
RNA-Seq Libraries… with More Details



A. Poly A Priming



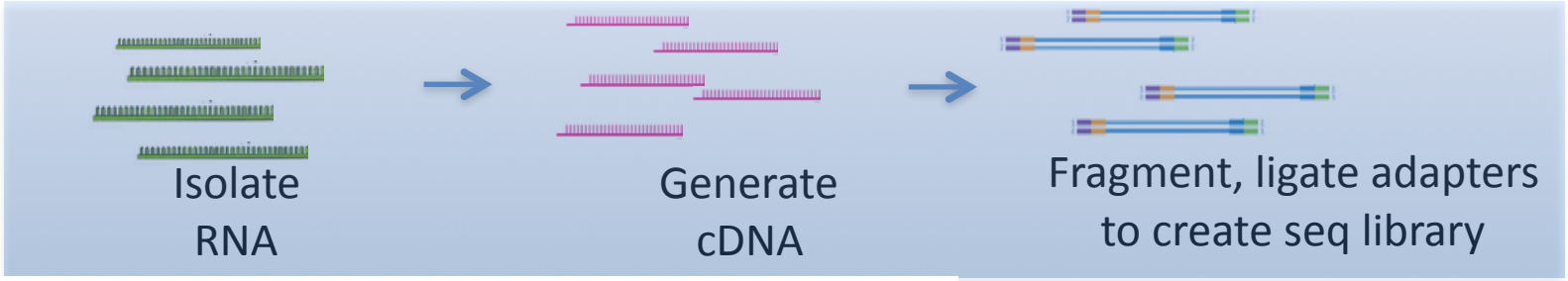
B. Random Priming

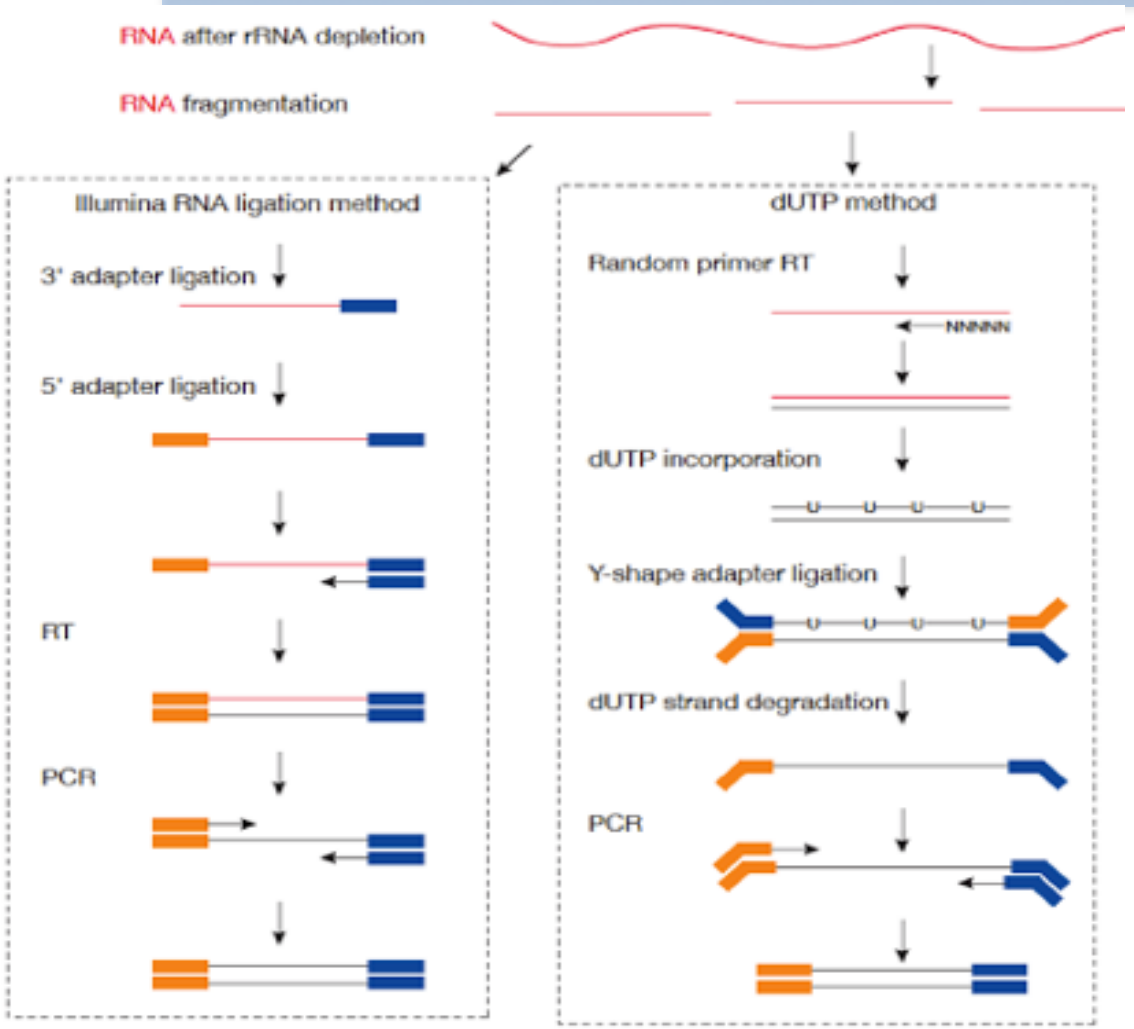


C. Priming using pre-ligated oligo

Illumina Small RNA kit
SOLiD RNA kits

RNA-Seq Libraries… with More Details



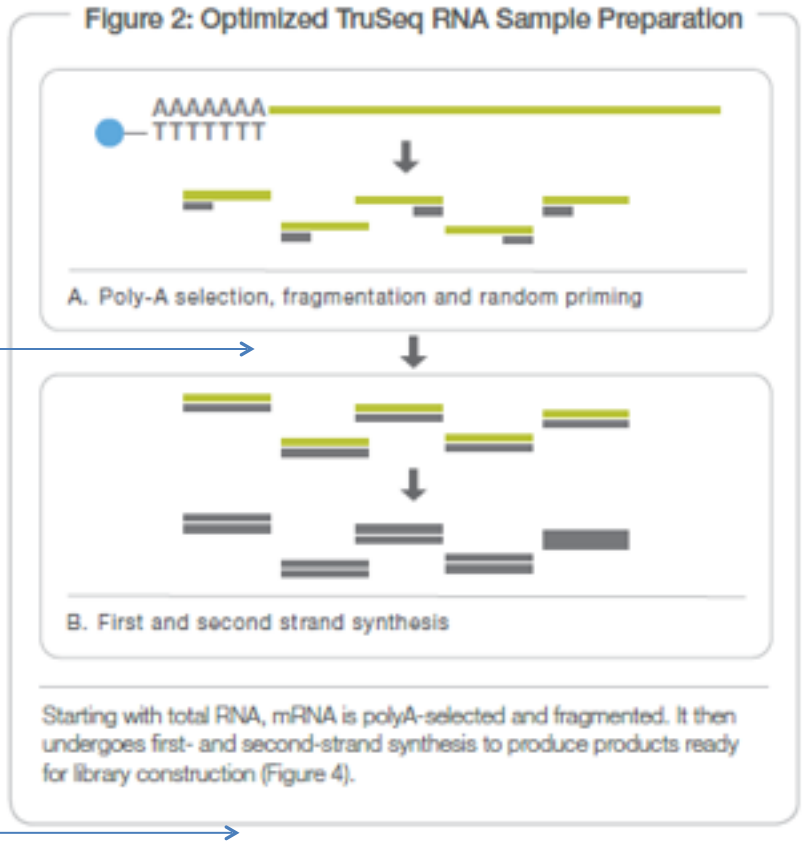


Second Strand Synthesis- Many Strand Specific Methods.

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

RNA Illumina Tru-Seq library prep

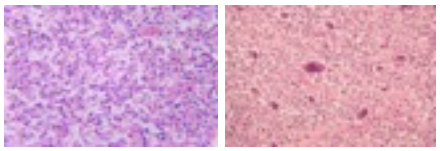
Size selection step



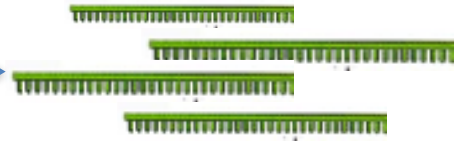
Adaptor ligation and standard library preparation

2 days for 8 samples

RNA-Seq… at it's Most Basic Form

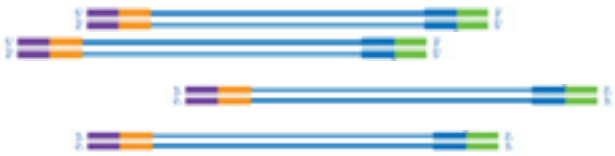


Samples from two conditions



Isolate RNA

Generate cDNA



Create sequencing library by size selection and adding adaptors

Run sequencer

Generate short reads

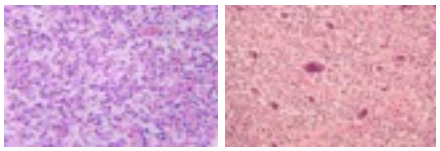
Identify differentially expressed genes

What is the adaptor?

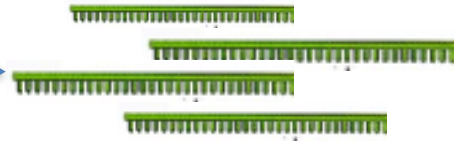
Adaptor :

- Allows template to attach to cluster in flowcell.
- Has a primer to start synthesis from.
- Has barcodes for multiplexing

RNA-Seq… at it's Most Basic Form

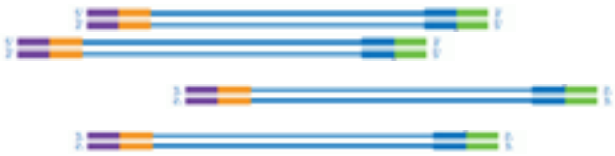


Samples from two conditions



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

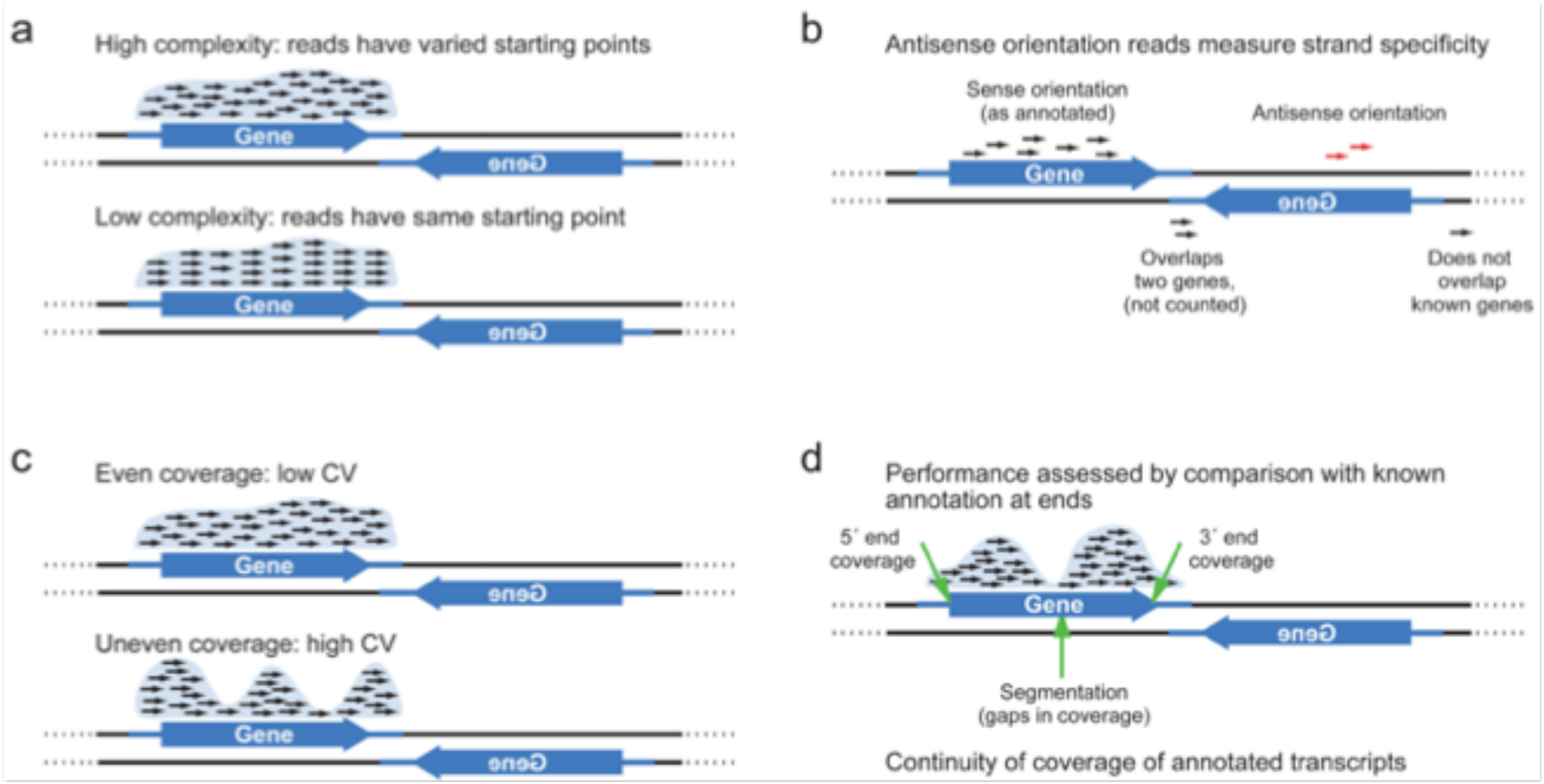
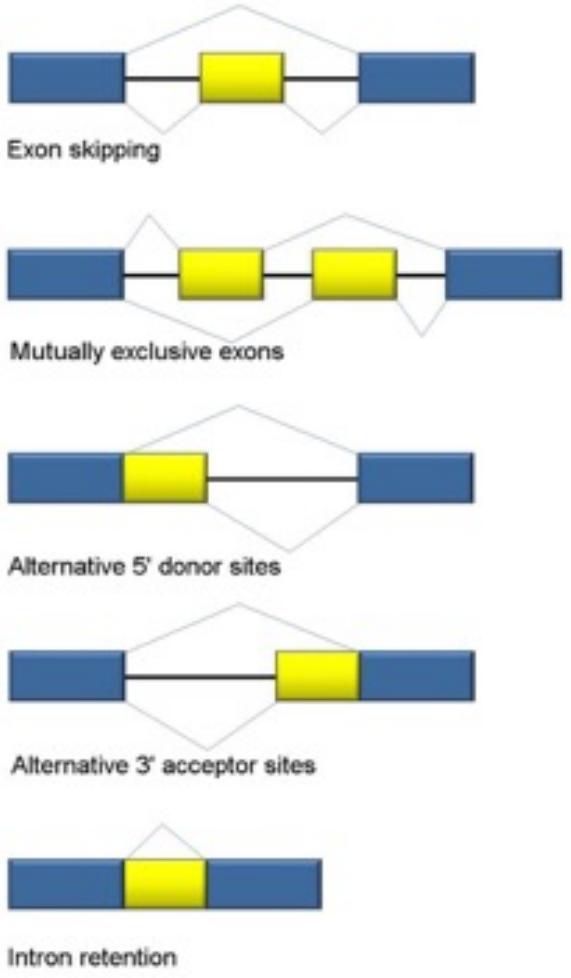


Figure 2. Key criteria for evaluation of strand-specific RNAseq libraries
Four categories of quality assessment. Double stranded genome (black parallel lines), with Gene ORF orientation (thick blue arrow) and UTRs (thin blue line), along with mapped reads (short black arrows – reads mapped to sense strand; 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

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!

Actually important!

From RNA-seqlopedia

3’ TAGSEQ- An Alternative to Whole RNA-Seq

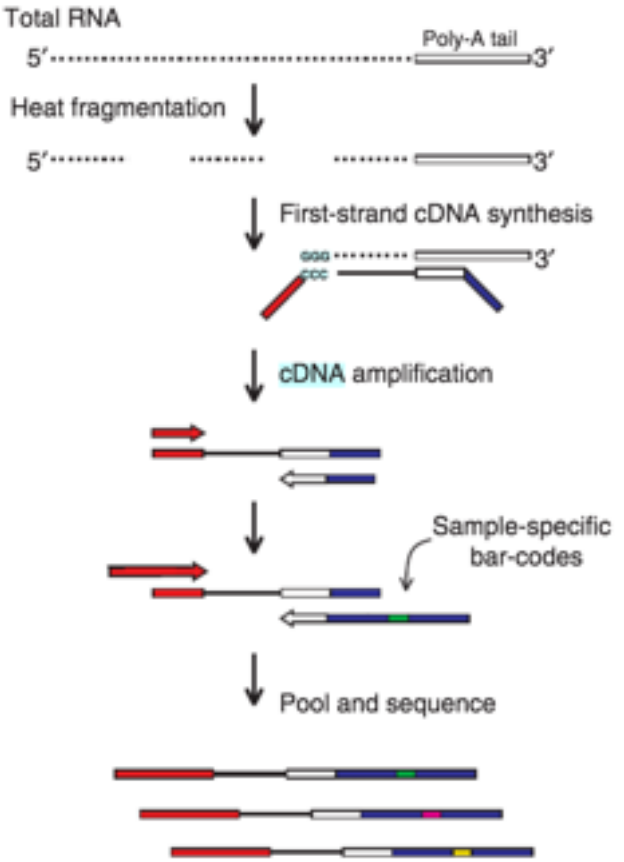
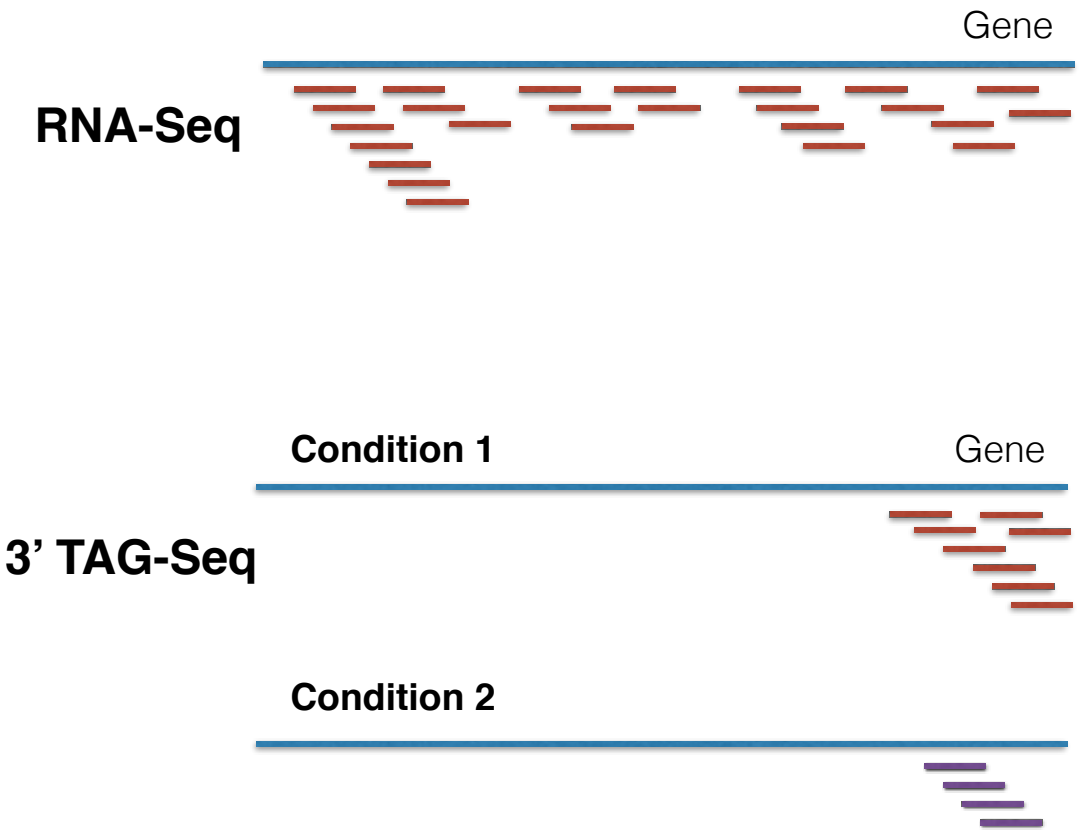


Fig. 1 Overview of the protocol used to prepare 3′ cDNA tag libraries from total RNA. RNA was fragmented at the beginning to eliminate biases resulting from differences in transcript lengths. First-strand cDNA was primed with a modified oligo-dT containing primer to target 3′ ends. Each sample was prepared with a sample-specific oligonucleotide barcode, then quantified and pooled prior to sequencing.

Targeting the 3’ prime end of RNA

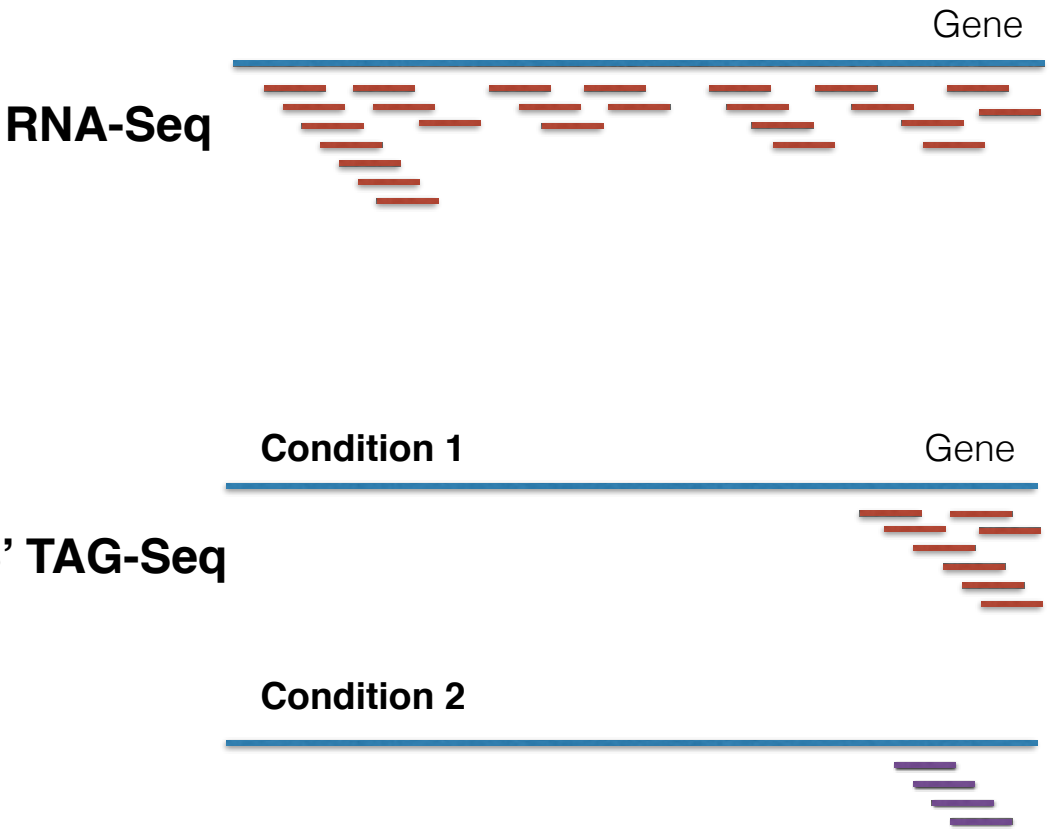
WHY TAGSEQ?

- Cheaper to sequence 3' end instead of the entire RNA.
- Amount of input RNA required is less.
- You can still identify differential expression.

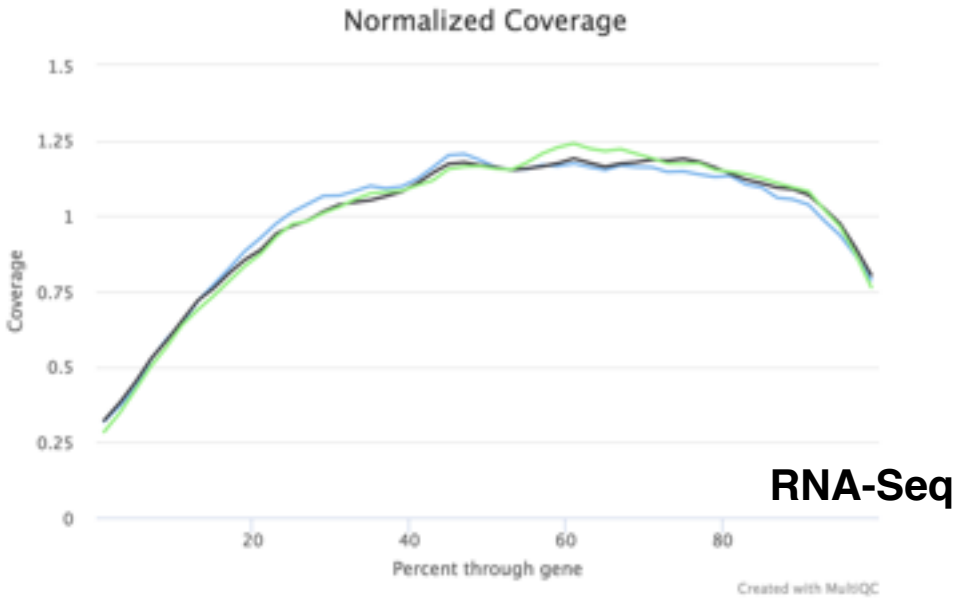


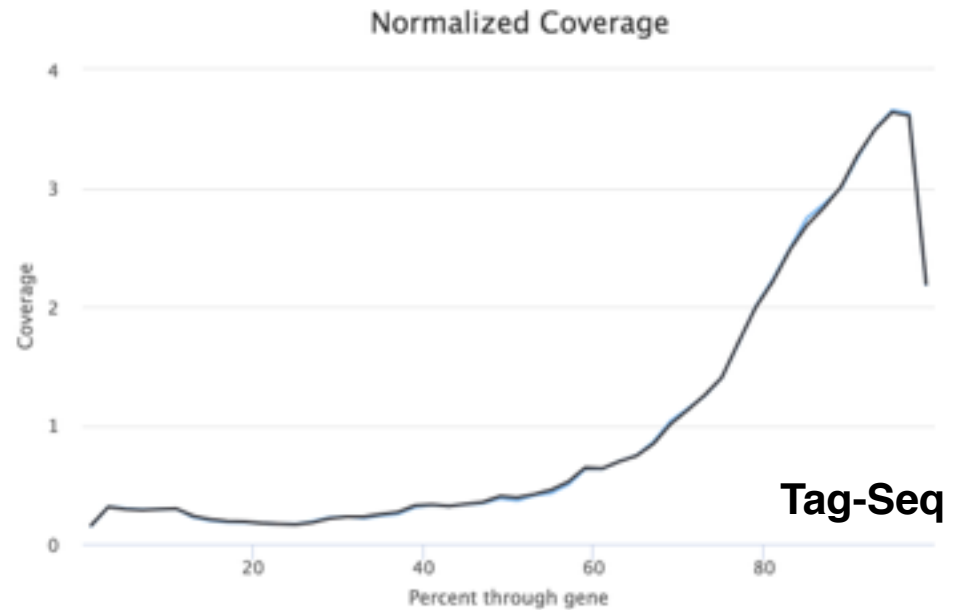
WHY NOT TAGSEQ?

- If you want to look at differential splicing
- If you want to identify polymorphisms in gene sequences



Whole RNA-Seq vs TagSeq





Whole RNA-Seq vs TagSeq

TagSeq recovers known concentrations of mRNA (ERCC controls) with more accuracy than whole mRNASeq

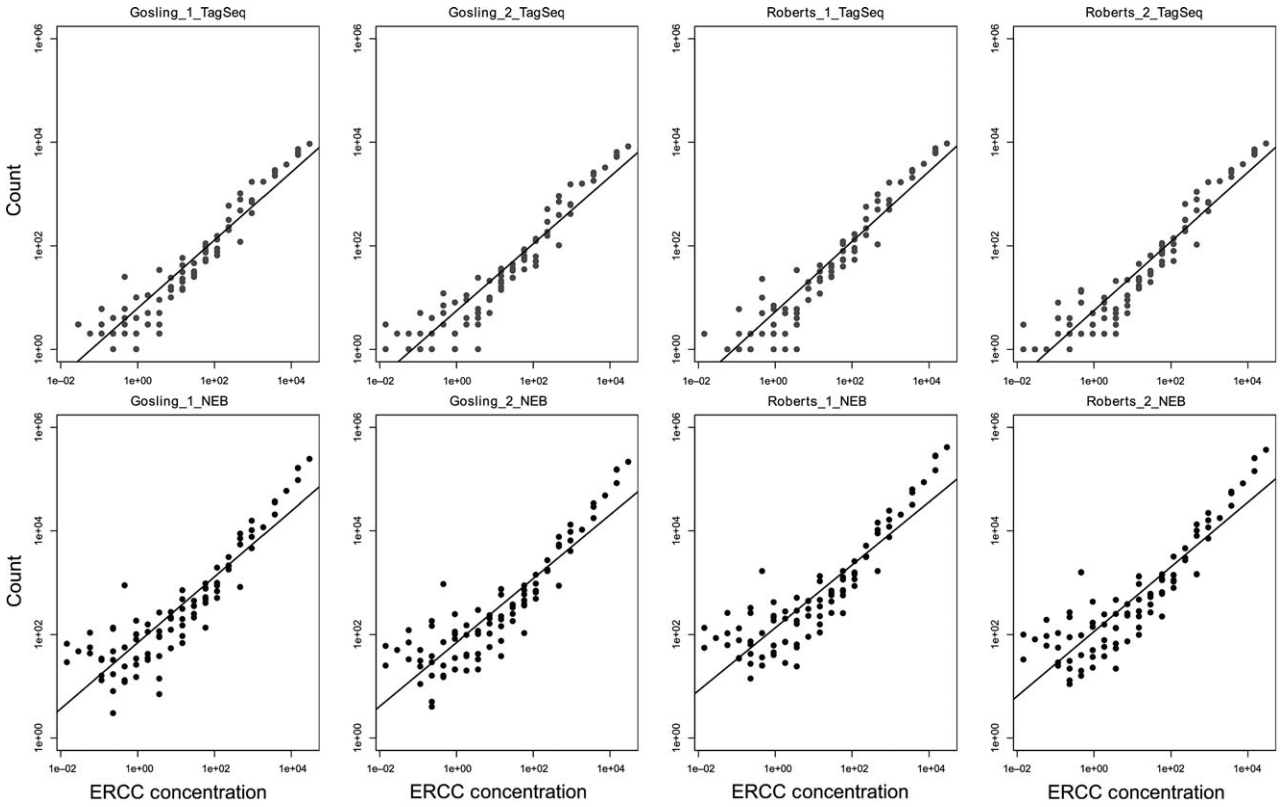


Fig. 1 Regression of observed vs. expected ERCC transcripts shows TagSeq has higher adjusted R^2 values for four different biological samples prepared with both methods (paired t-test, $t = 18.63$, d.f. = 3, $P < 0.001$).

Lohman et al, Molecular Ecology Resources, 2016

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 **OPTIONAL**
- **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

h" p://www.asciitable.com/2

```
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.)

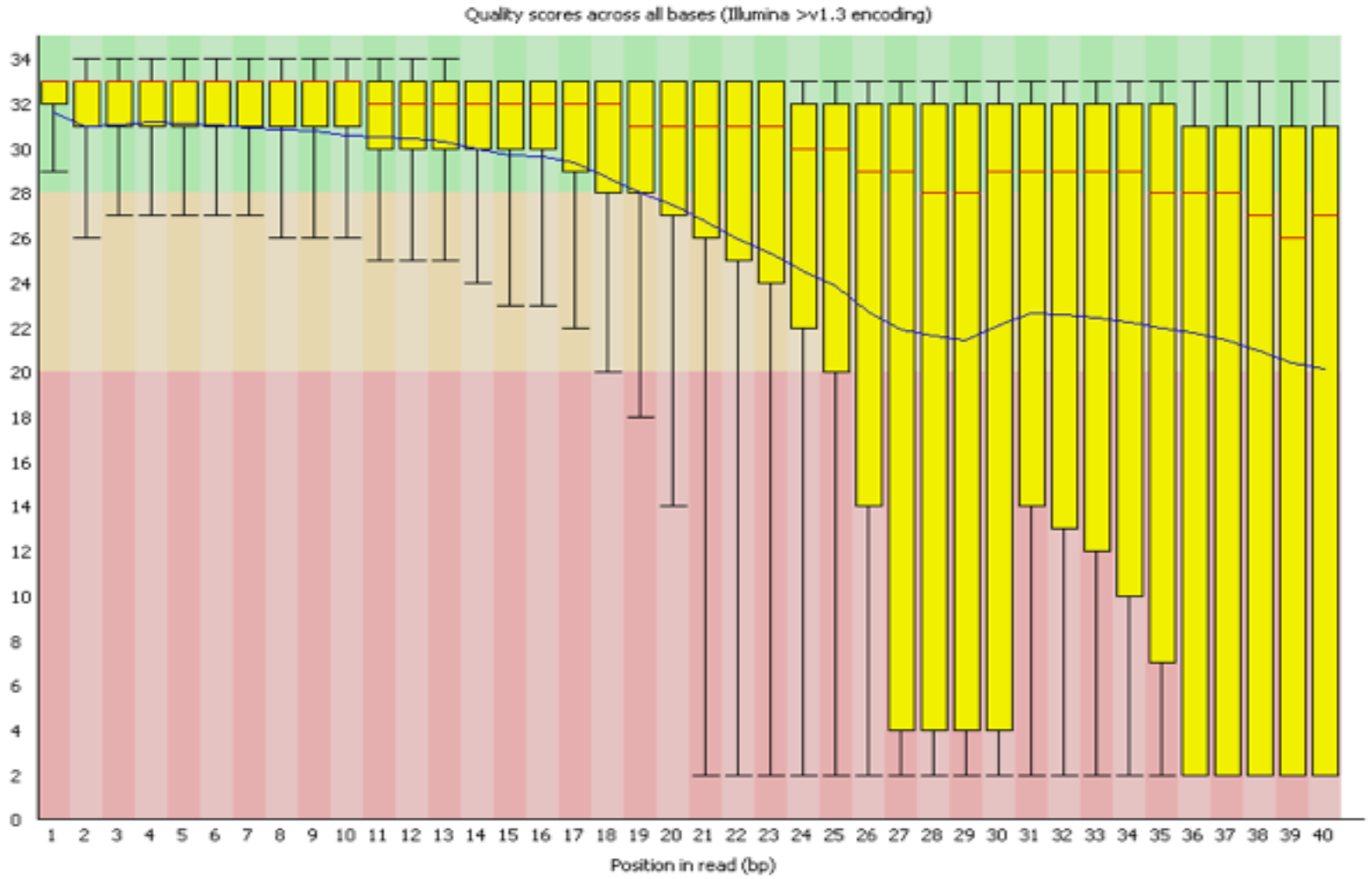
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

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

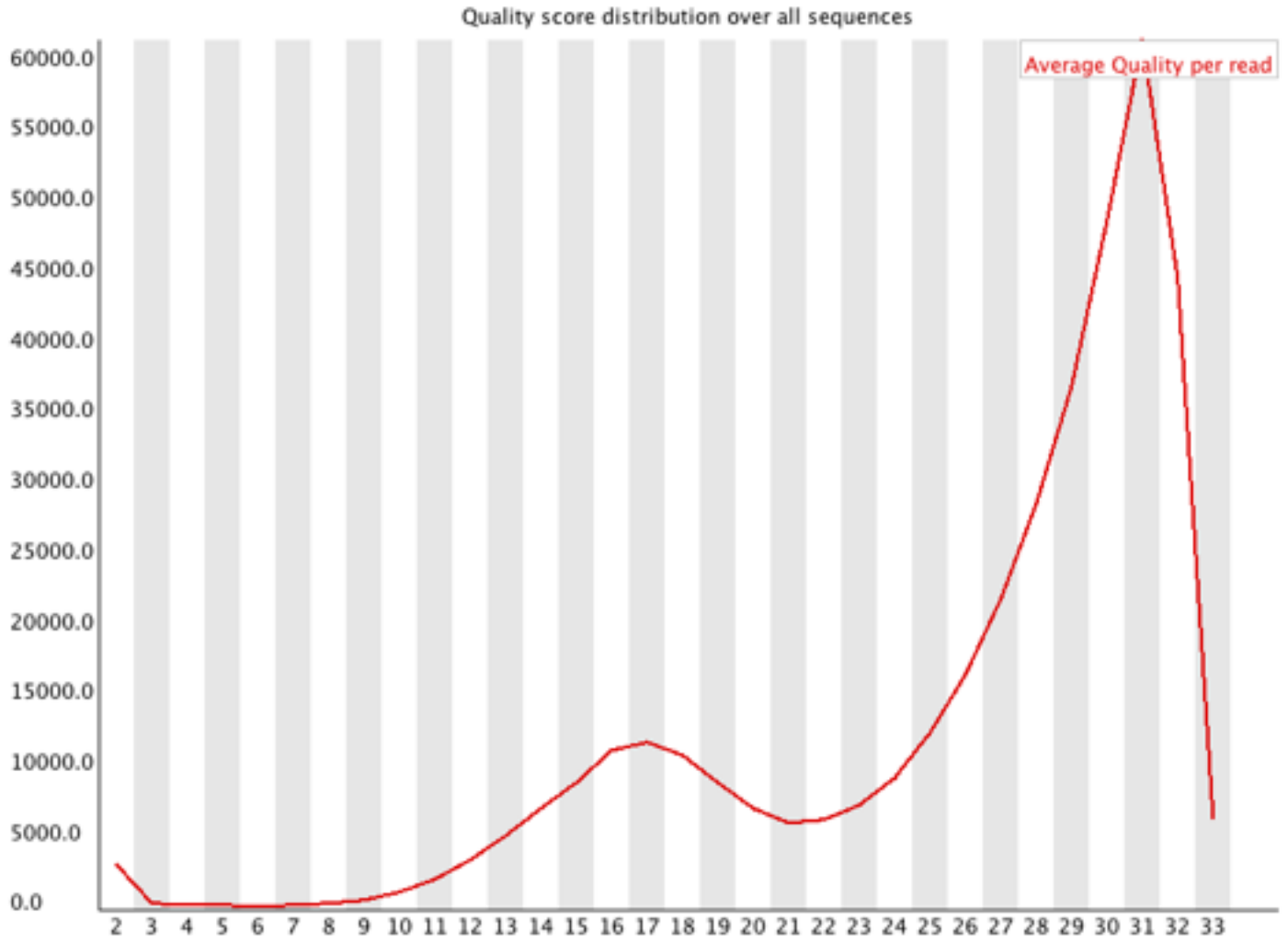
STEP 1 - Evaluate Raw Data

- Count your reads!
- Assess quality using FastQC reports
  - Median base quality across the length of the read



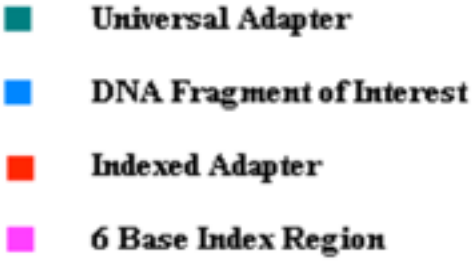
STEP 1 - Evaluate Raw Data

- Assess quality using FastQC reports
  - Quality score distribution across all reads



STEP 1 - Evaluate Raw Data

- Assess quality using FastQC reports
  - Adapter content

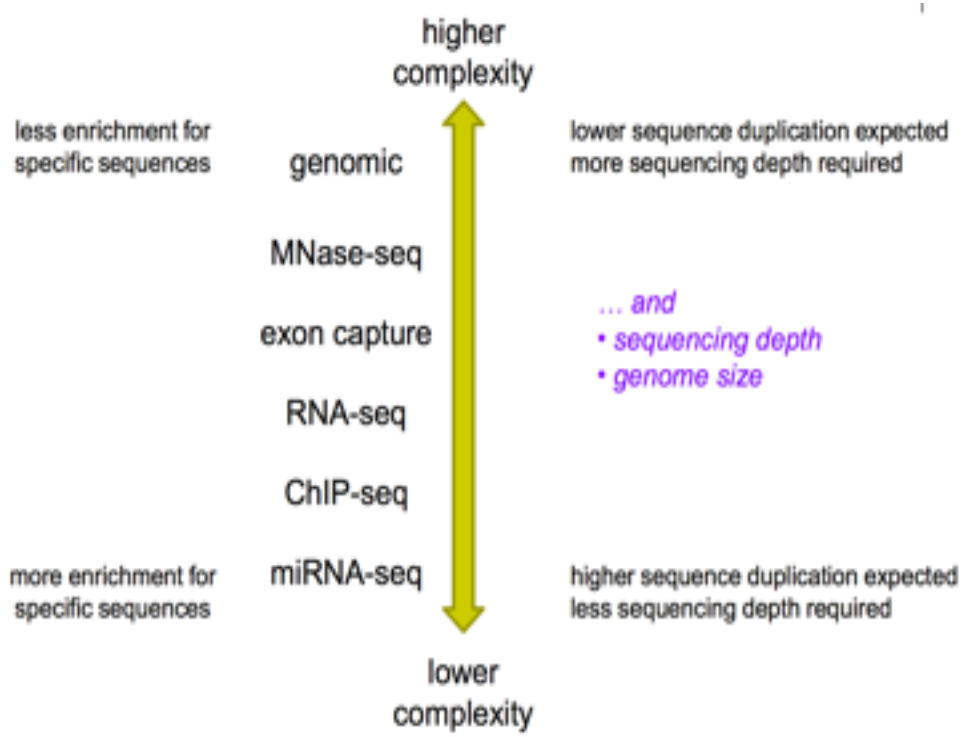


Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATG	60030	5.01369306977828	TruSeq Adapter, Index 1 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGC	42955	3.5875926338884896	TruSeq Adapter, Index 1 (97% over 37bp)
CACACGTCTGAACTCCAGTCACCTCAGAATCTCGTATGCCGTCTTCTGCT	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…



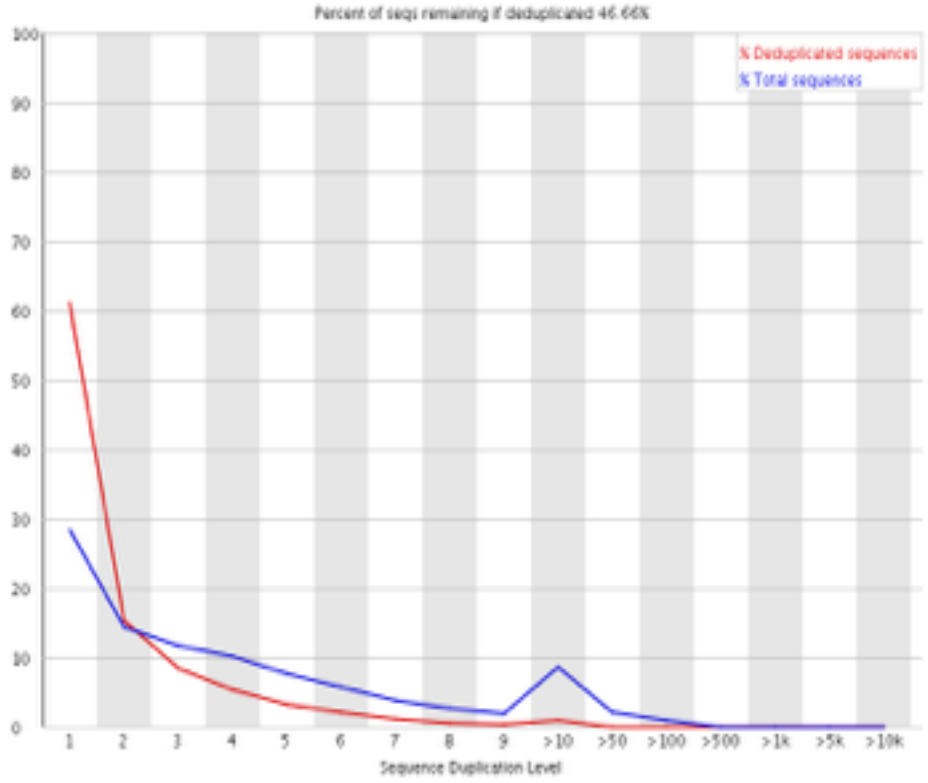
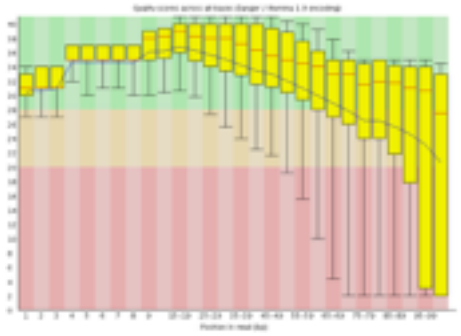
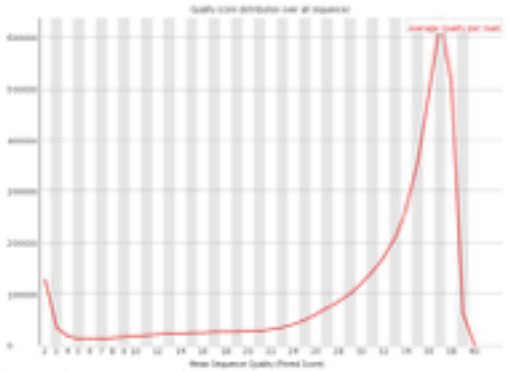


Image created by Anna Battenhouse

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

STEP 1 – Manipulate Raw Data

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

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

Table 1 | Selected list of RNA-seq analysis programs

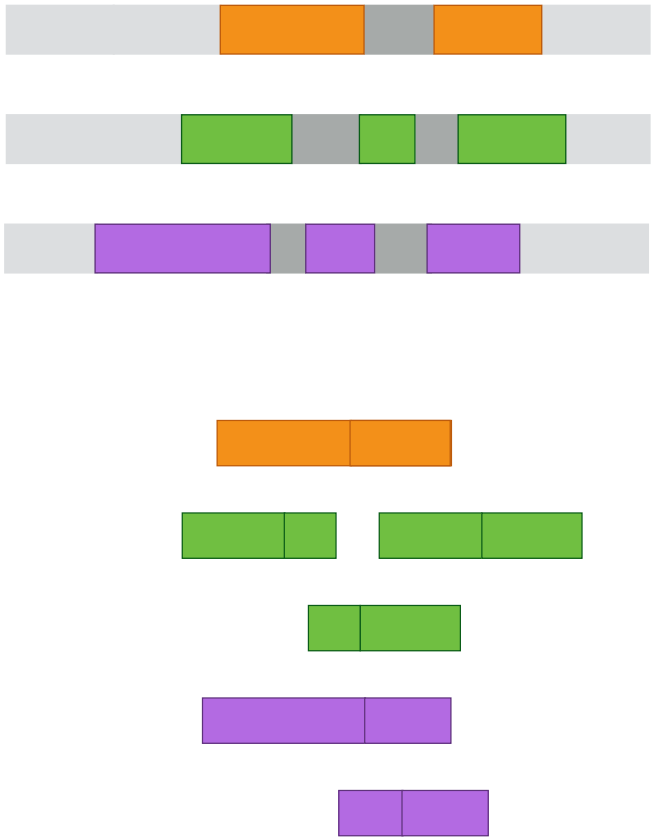
Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	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

What does the reference look like?

- **Genome:** All the DNA of an individual, organized by chromosome, containing non-coding and coding regions.
- **Transcriptome**: All the gene isoforms. No non-coding sequences.



What does an alignment look like?

Ref=TAGATCAGATTCGATACCAGACCATGATCATACGATCCA

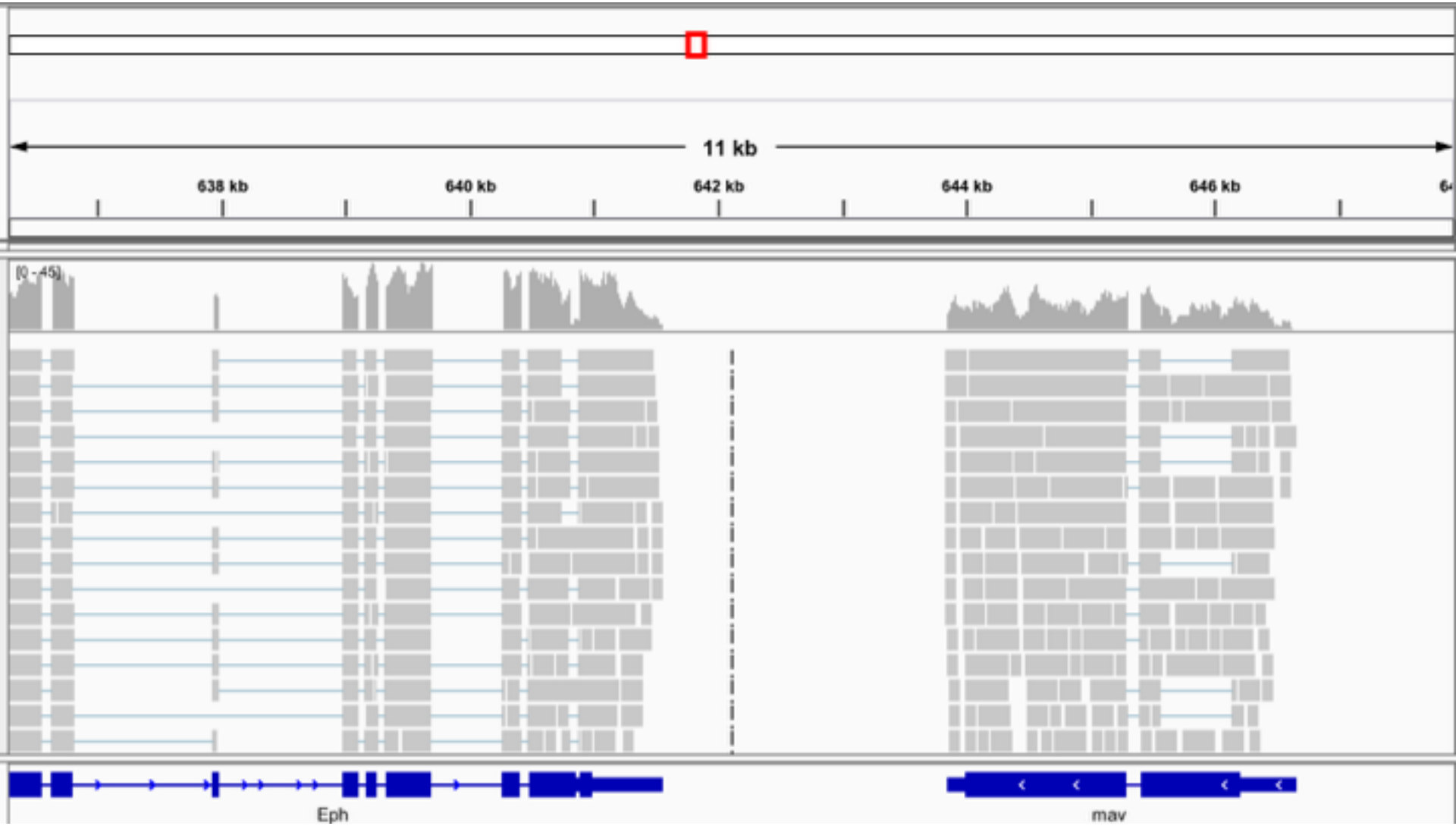
Read=AGACCATG

Found at offset 18!
TAGATCAGATTCGATACC**AGACCATG**ATCATACGATCCA

What does an alignment look like?

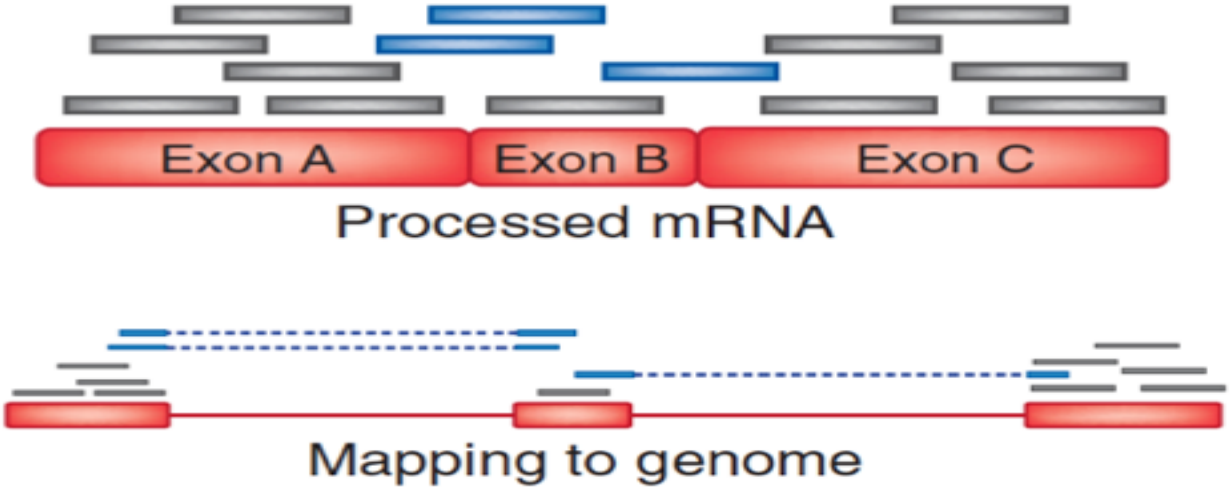
gatkforums.broadinstitute.org

What does an alignment look like?



https://vlsci.github.io/lscc_docs/tutorials/rna_seq_dge_basic/rna_seq_basic_tutorial/

Unspliced Mapping

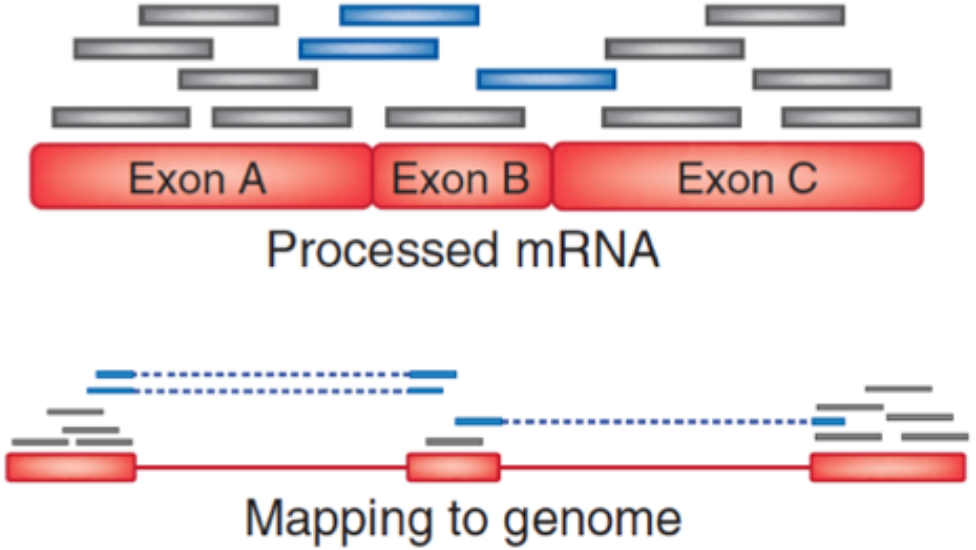


Class	Category	Package	Notes
Read mapping			
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension
		Stampy[39]	Probabilistic model
	Burrows-Wheeler transform methods	Bowtie[43]	
		BWA[44]	Incorporates quality scores

Garber et al, Nature Methods, 2011

Spliced mapping

- Needed for identifying and quantifying splice variants from RNA Seq data.
- Tools:
  - **HiSat2**
  - **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

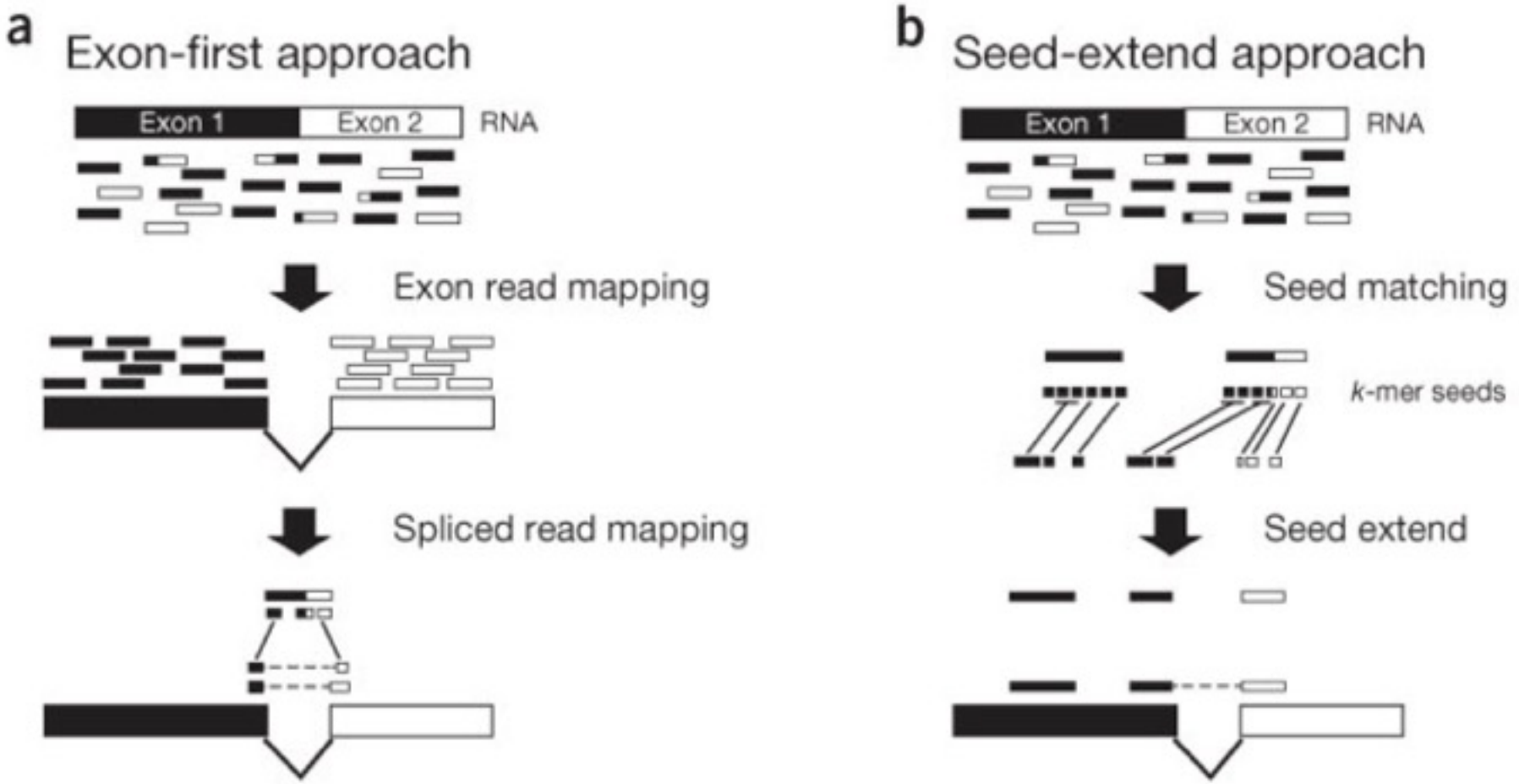


Figure :
Garber et al, Nature Methods, 2011

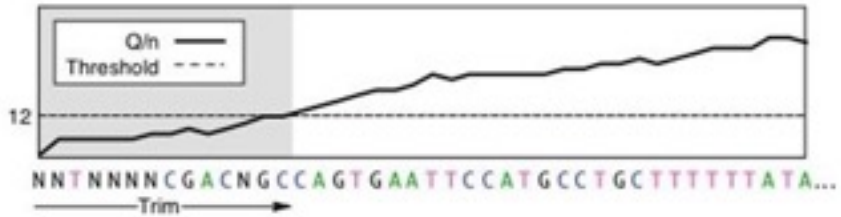
What to know about your data before mapping?

KNOW YOUR DATA!

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

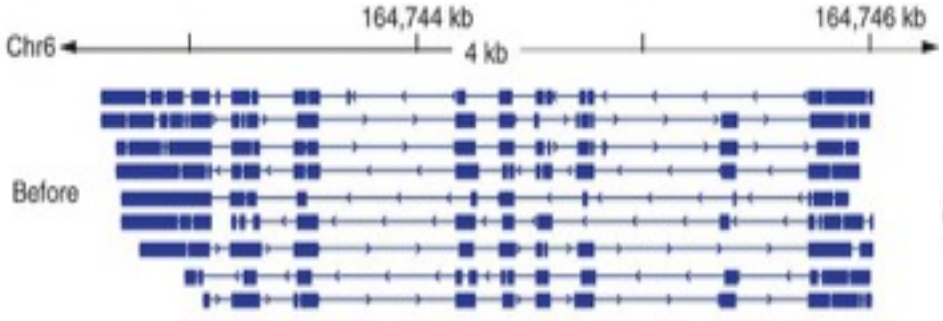
PREPROCESSING

- 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 14501540500 + . 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!

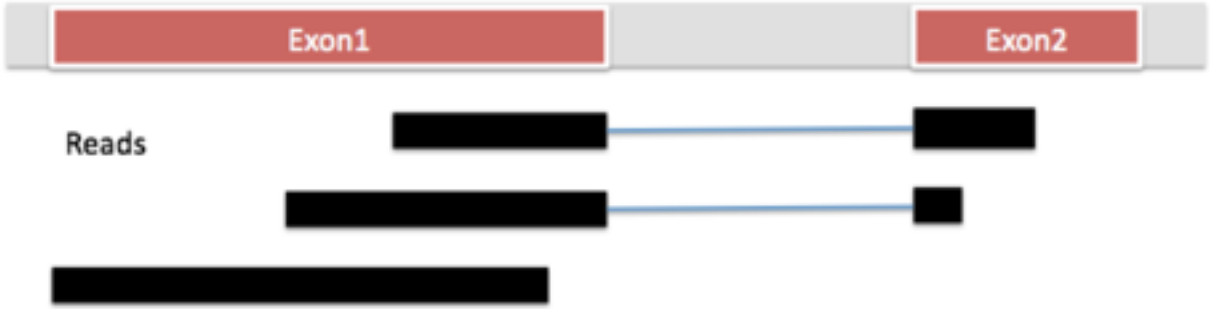
Mapping with BWA

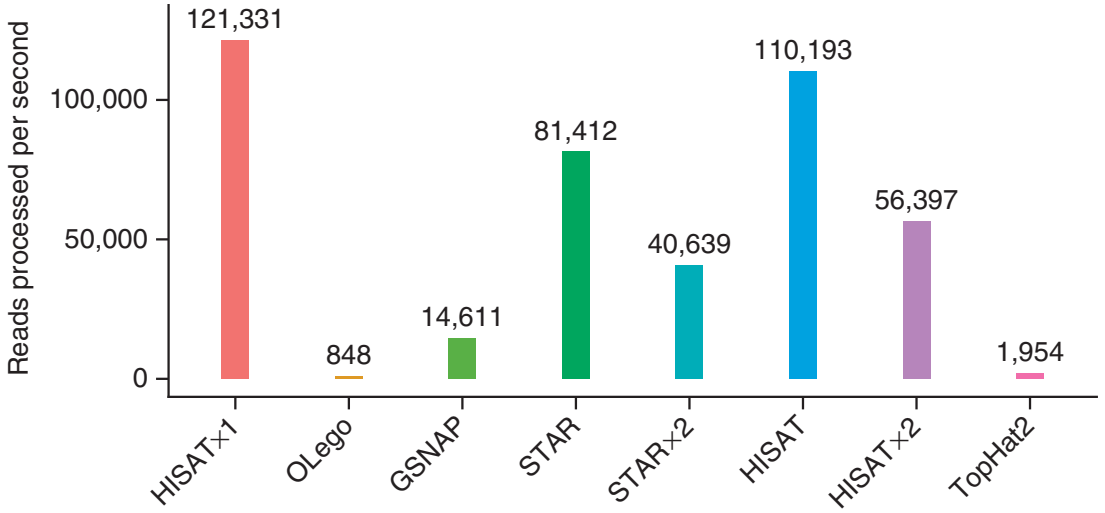
- Create an index of your reference – bwa index
- Run mapping - bwa mem

- 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 TACC – **runBWA_mem.sh** in $BI/bin

Mapping with Hisat2

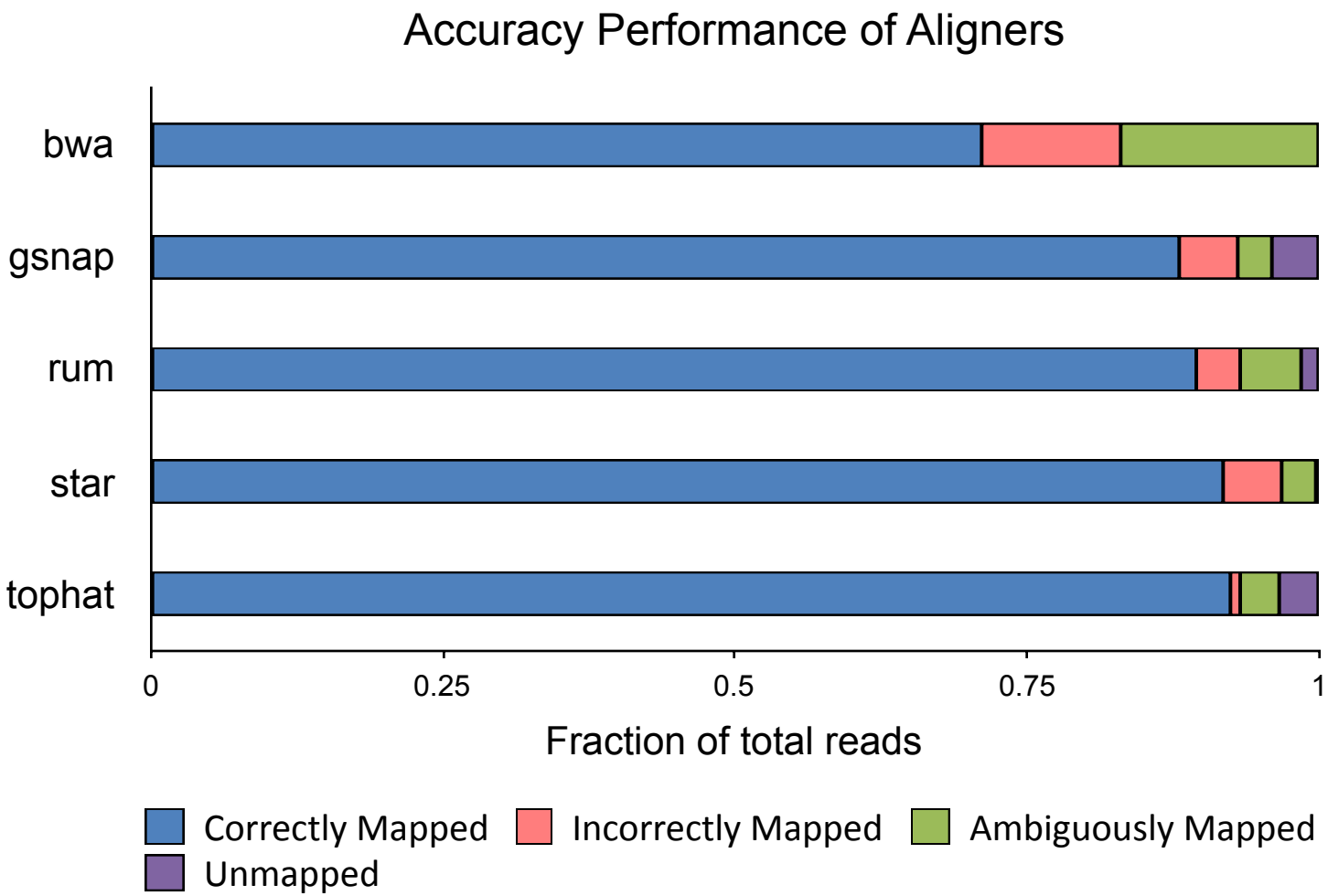
- **“Heirarchical indexing for spliced alignment of transcripts”**
- Faster splice-aware mapper
- Global indexes + Many small overlapping local indexes





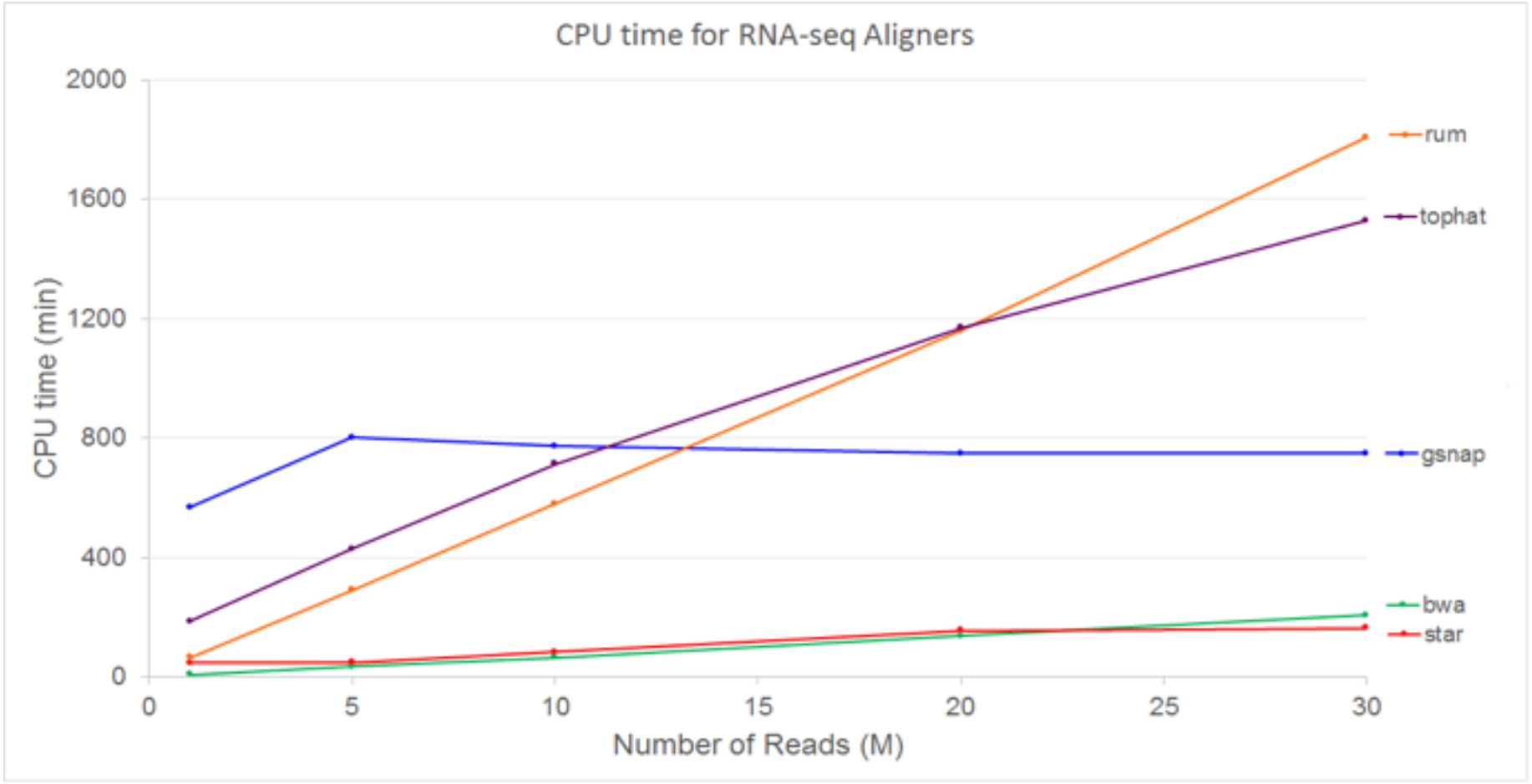
Daehwan et al. Nature Methods. 2015

Mappers comparisons



New benchmarking analysis performed by **Raghav Shroff**

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!
```

2S5M2I4M1D4M6S2

	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
2. Sort and index newly created BAM file
3. Mapping Statistics

   samtools flagstat/samtools idxstats

flagstat output

```
37144063 + 0 in total (QC-passed reads + QC-failed
reads)
0 + 0 duplicates
31291926 + 0 mapped (84.24%:-nan%)
37144063 + 0 paired in sequencing
18565040 + 0 read1
18579023 + 0 read2
28963894 + 0 properly paired (77.98%:-nan%)
31255233 + 0 with itself and mate mapped
36693 + 0 singletons (0.10%:-nan%)
2485389 + 0 with mate mapped to a different chr
```

idxstats output

```
GEET01005312.66.2790    2725    0   0
GASX01002166.56.2741    2686    4   0
GBXQ01000253.110.2765   2656    0   0
GBZK01006874.2968.6337  3370    50  0
GCVF01002796.135.2107   1973    0   0
```

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

STEP 2– Mapping to Reference

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

Mapping Summary

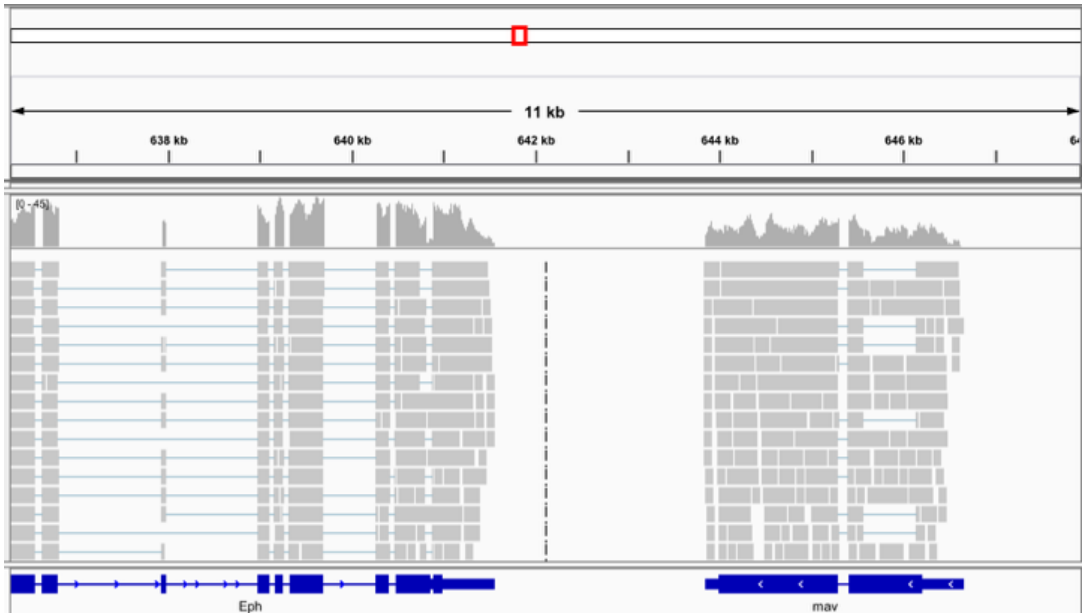
- Unspliced mappers (BWA, bowtie2) ok when mapping to the transcriptome.
- Spliced mappers (Hisat2, STAR) are good for mapping to the genome.
- Samtools can be used to gather basic mapping statistics, RNASEQC for RNA specific statistics

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

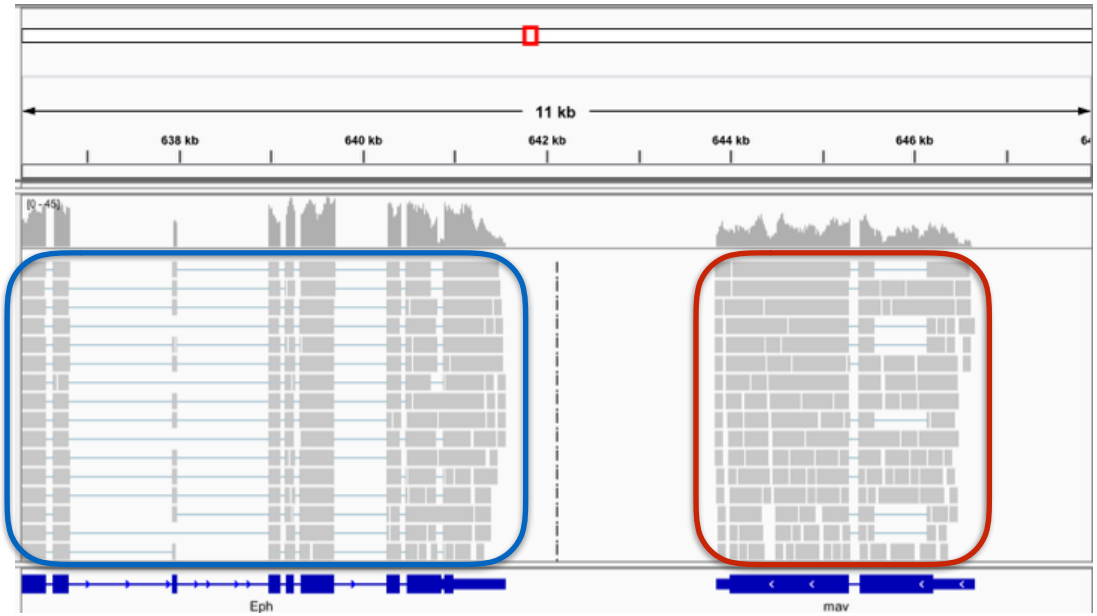
STEP 4- Quantify Expression

- Quantify expression=gene counting=transcript counting
- Mapping tells us where every read came from.
- How do we go from that to gene expression?
  - What genes are expressed?
  - What is the expression level for each gene/gene isoform?



STEP 4- Quantify Expression

- What is gene expression?
  - A gene is expressed when it's corresponding DNA sequence is transcribed into mRNA (for translation into protein).
- What is gene expression level?
  - The amount of mRNA detected in a sample.



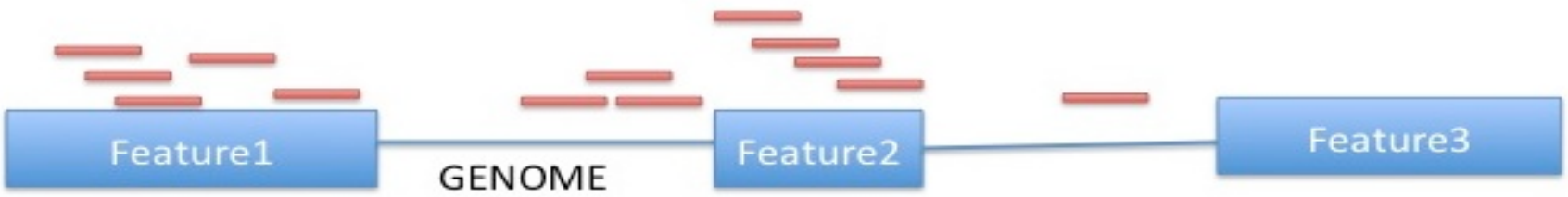
Count for gene Eph

Count for gene mav

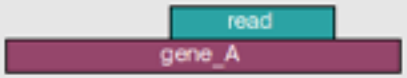
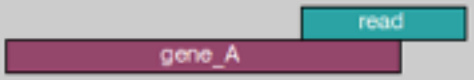
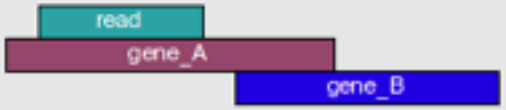
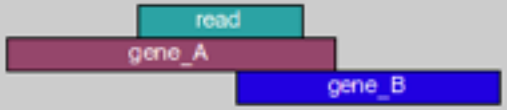
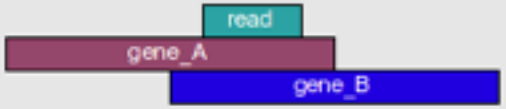
- **Read depth= mRNA amount= expression level of gene**

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

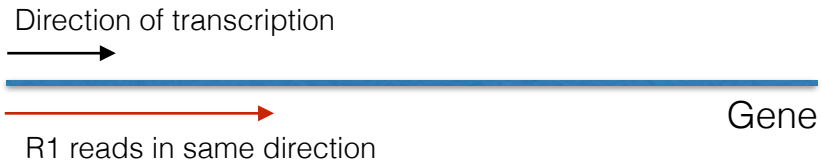
HTSeq –

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

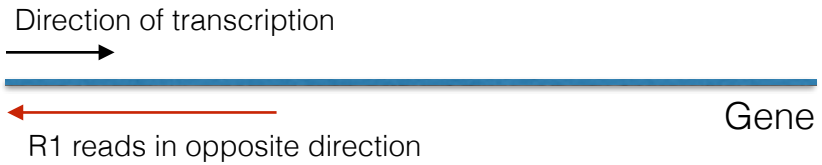
STEP 4 : Quantify Expression

Ligation vs dUTP stranded library

Ligation



dUTP

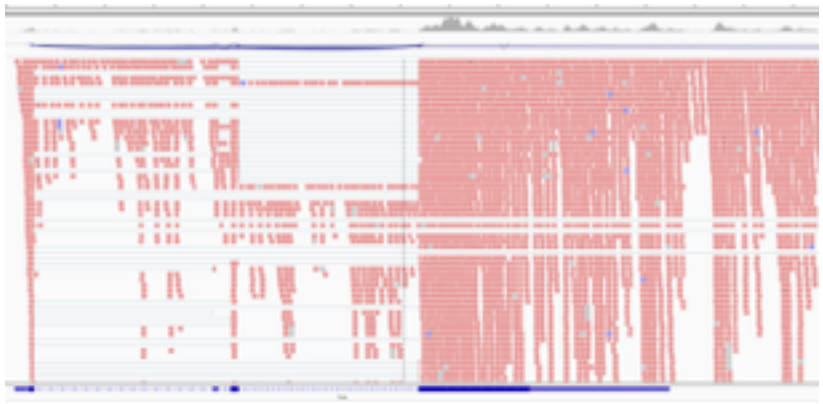


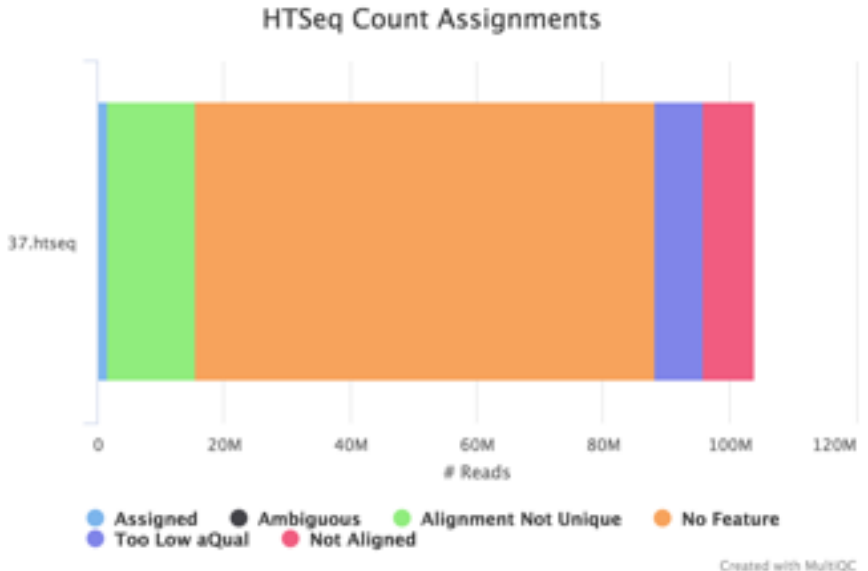
HT-SEQ —reverse tag applicable for dUTP library prepped data

HT-Seq —reverse

TLR4 = 0

TLR4 Pileup



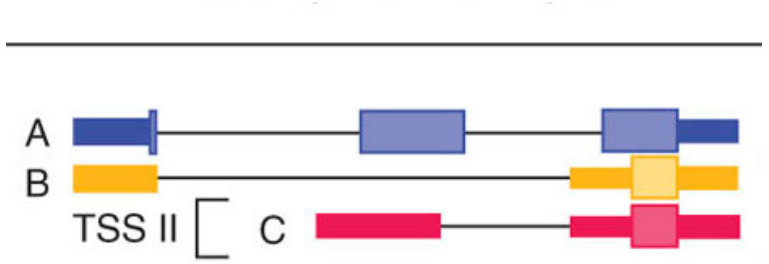


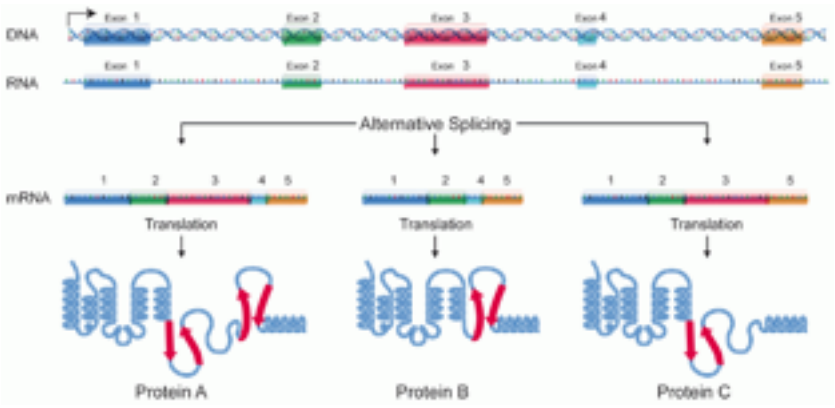
STEP 4- Quantify Expression

- Quantifying a gene is simpler than quantifying its different isoforms/ transcripts.
- Tools: kallisto, stringtie, and cufflinks

What is a gene? What is a transcript?

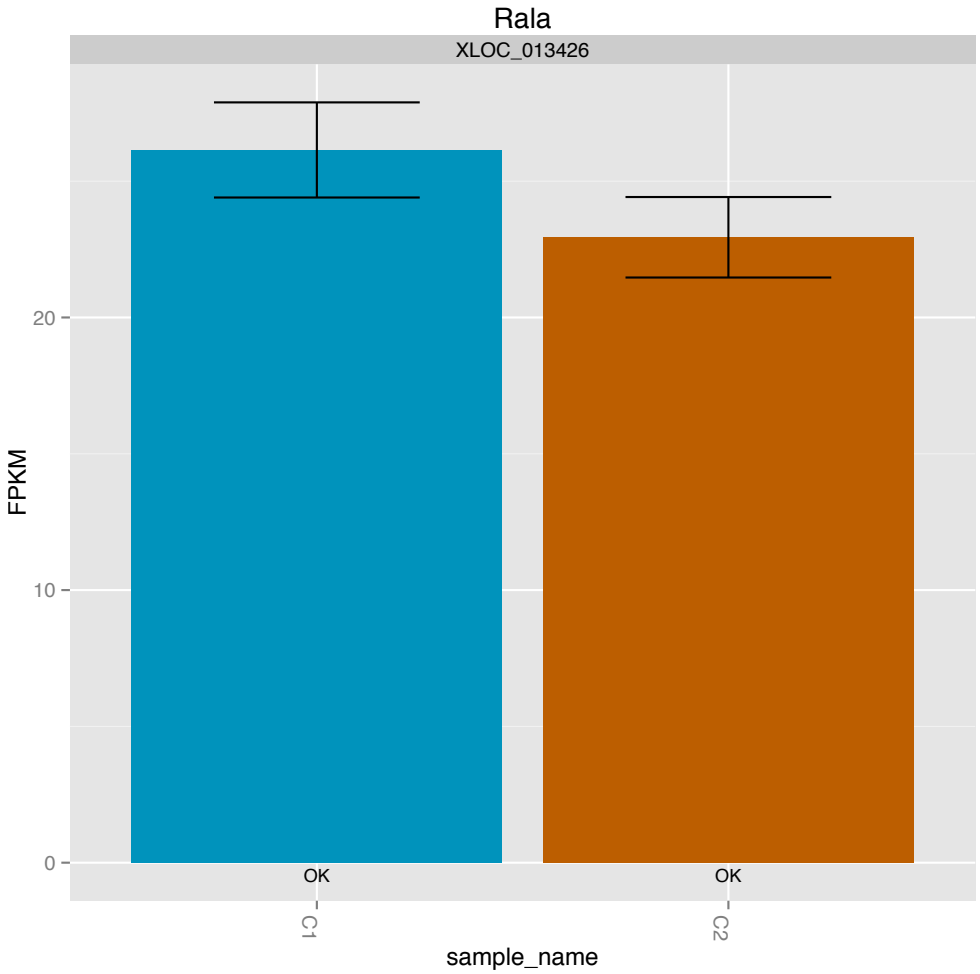
A gene can have multiple transcripts!

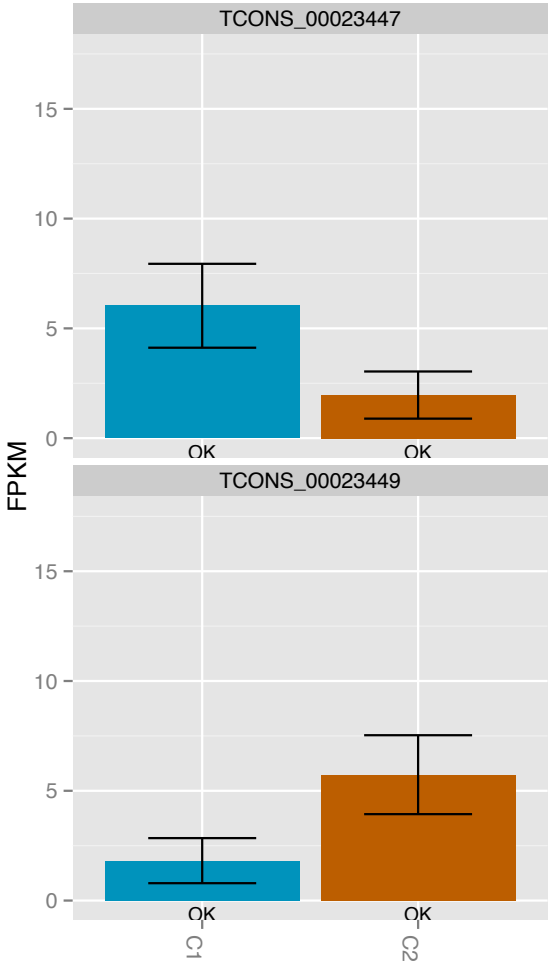




STEP 4- Quantify Expression

Why quantifying all transcripts of the gene may be important?

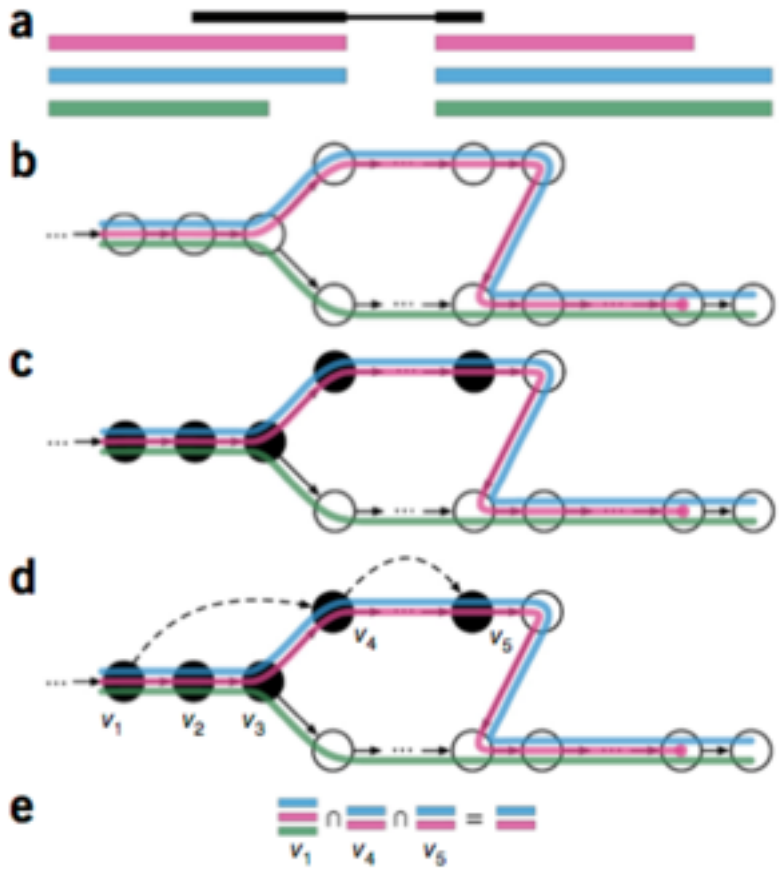




Pseudomapping and Quantification with Kallisto

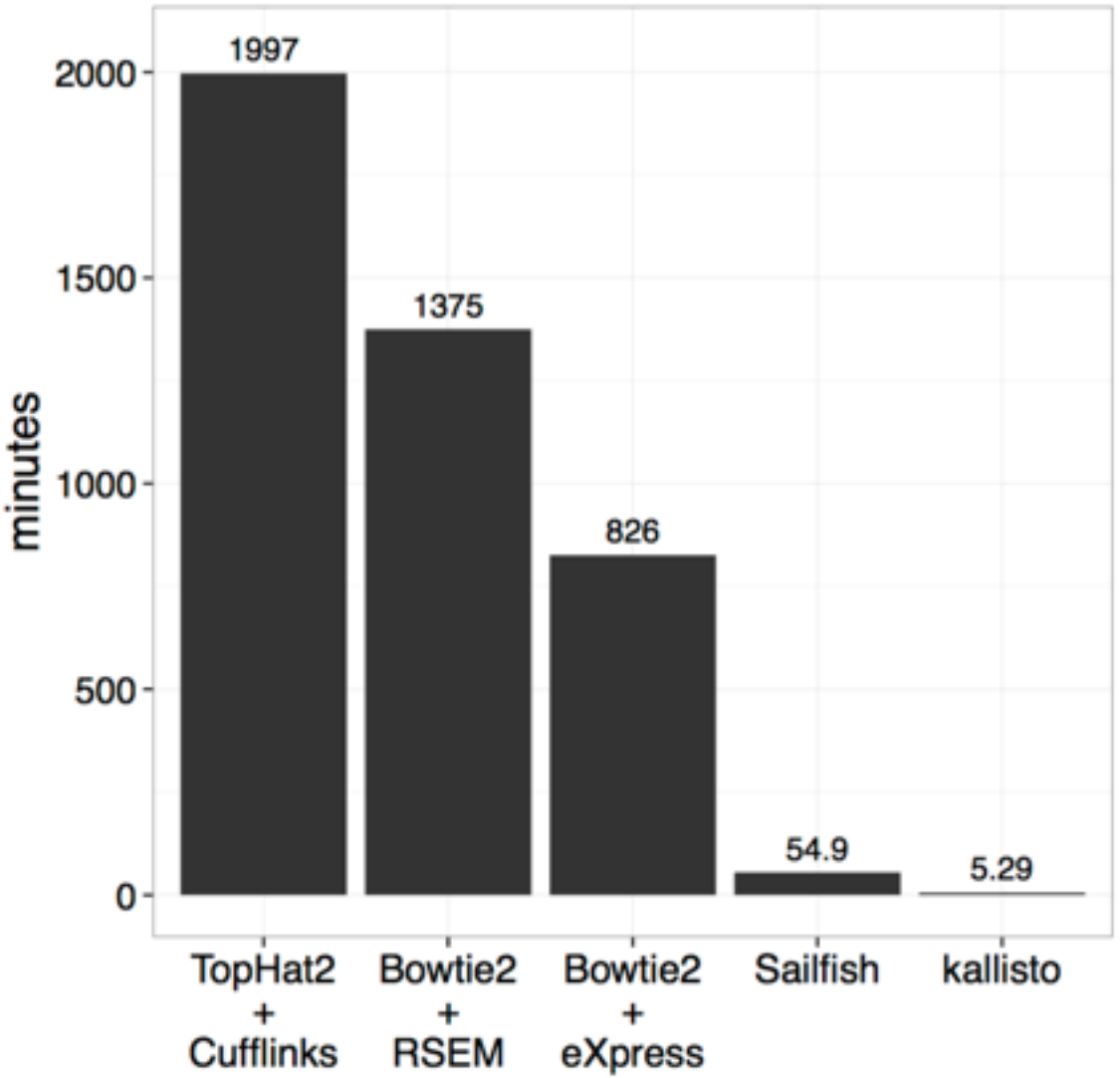
- **Kallisto** - super fast tool for transcript quantification.
- Skips the alignment step.
- Identifies transcripts that a read is compatible with in order to quantify the transcript.

- ONLY FOR MAPPING TO TRANSCRIPTOME!



Pseudomapping and Quantification with Kallisto

- Quantifies 30 million reads in 3 minutes on a mac laptop.



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 5- Test for Differential Expression

- Input: Gene Expression Matrix

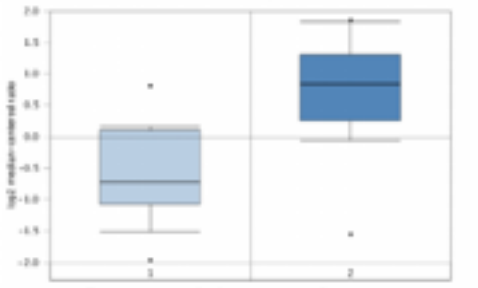
Gene

Sample

Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Plh1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

Image from babelomics

- Outputs like:



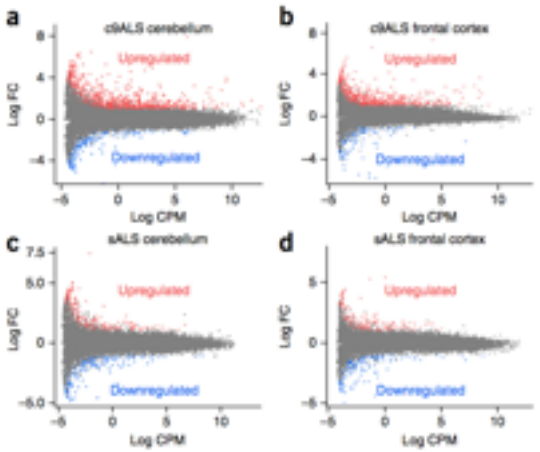


Figure: doi:10.1038/nn.4065

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- Test for Differential Expression

- Even before normalization, you may want to filter out genes with low counts.
- Remove genes with less than 1 count in most samples.
- Remove genes with very low variance across samples.

Gene

Sample

Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

STEP 5- Test for Differential Expression

- After mapping and quantifying the genes for each sample:
  - compare gene counts across samples/conditions.
- But first, **normalize!**
  - Normalization evens out the technical variations so that any variation you see between samples is “hopefully” due to real biological reasons.
  - Normalize for **read depth** differences
  - Normalize for **gene/transcript length** differences
    - RPKM = Reads Per **Kilobase of transcript** per **Million mapped reads**
    - **RPK= No.of Mapped reads/ length of transcript in kb (transcript length/1000)**
    - **RPKM = RPK/total no.of reads in million (total no of reads/ 1000000)**
  - Other normalization methods: upper quartile, median read count and more complicated scaling factors (DESeq2 R package)

Gene	Read Count
ABAR1	1200
ATXN1	1345
ATXN2	2
BRAT2	0
GABA	24
GABRA2	456
GABRA4	45345

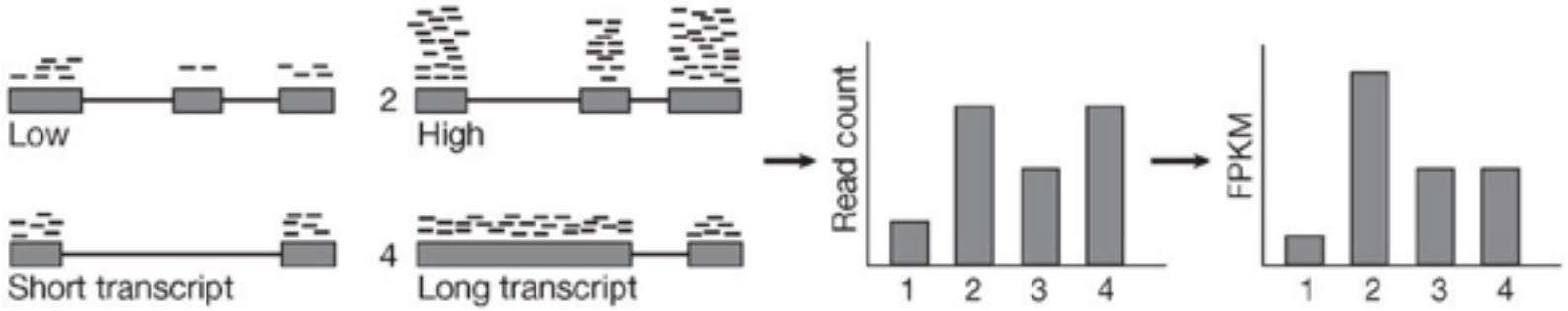
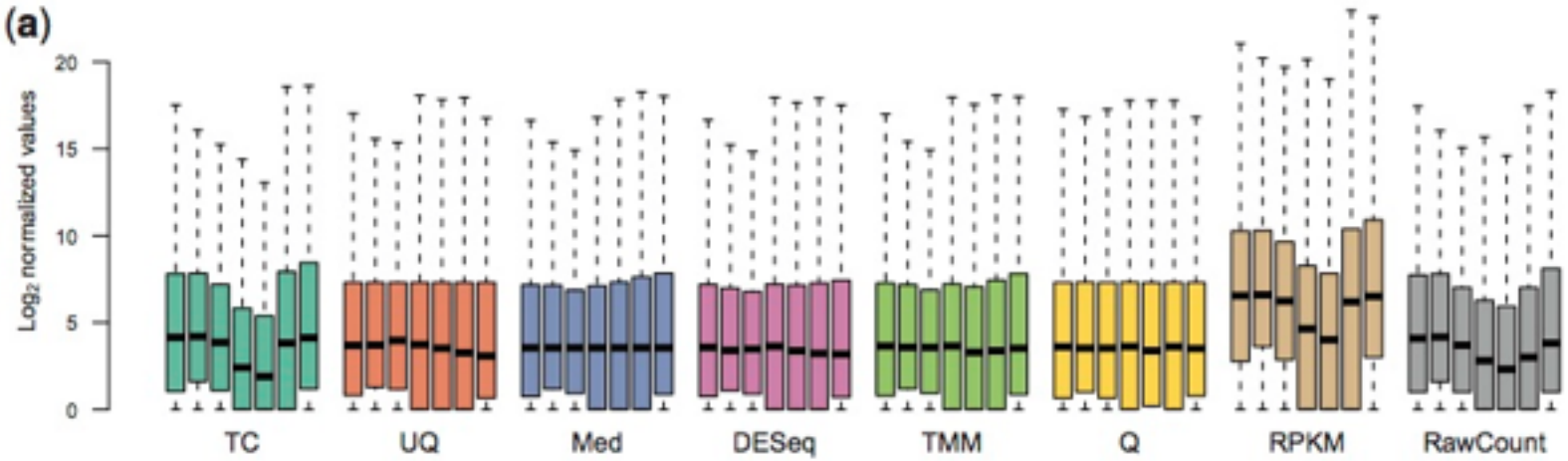
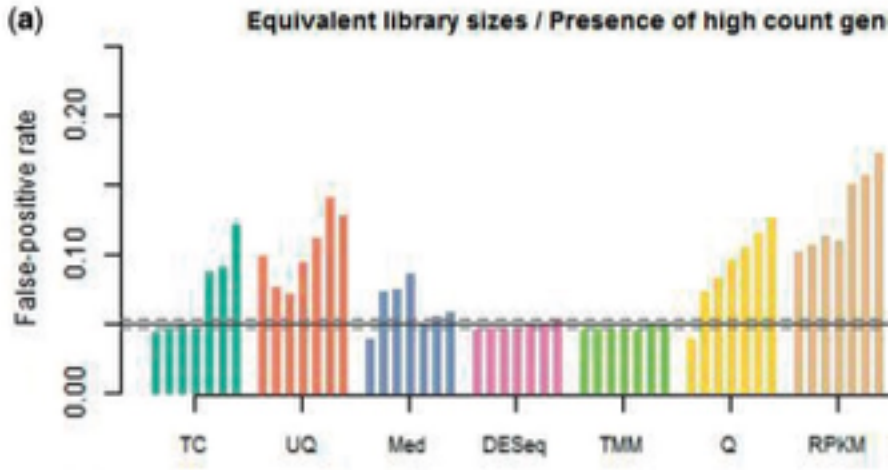


Figure: doi:10.1038/nmeth.1613

STEP 5- Test for Differential Expression

- Comparing different normalization methods





From: Dillies A et al, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, doi:10.1093/bib/bbs046 .

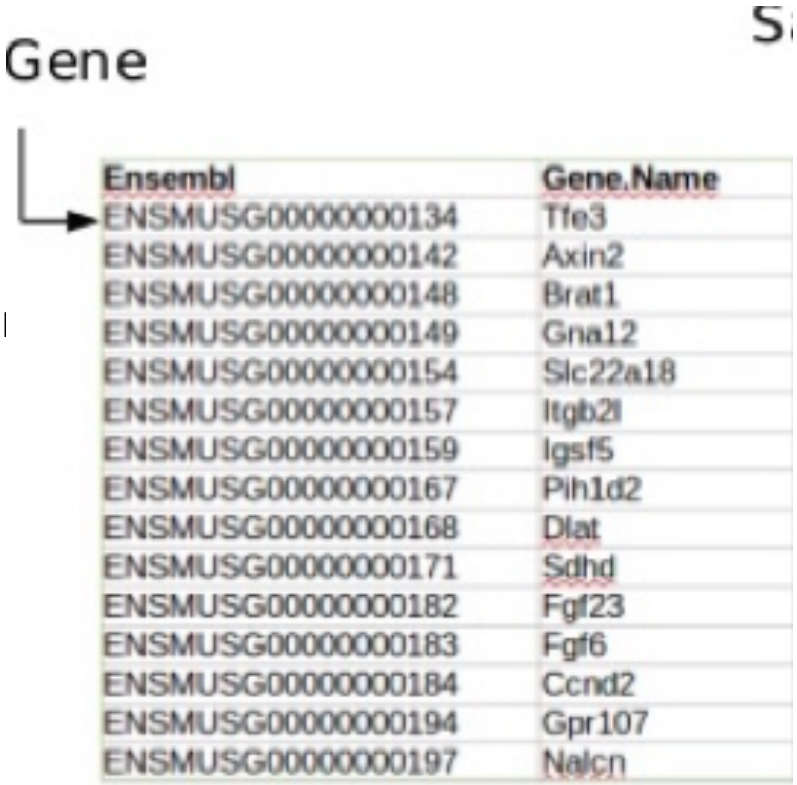
STEP 5- Test for Differential Expression

- Methods differ in how they normalize, what statistical test they use etc.

	DESeq2	edgeR	DEXSeq	Cuffdiff
Normalization	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
DE Test	Negative binomial test	Fisher exact test	Modified T test	T test
Advantages	Straightforward, fast, DESeq2 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, somewhat of a black box

STEP 5- Test for Differential Expression

- Output from differential expression testing is usually a table with the following values for every gene:
  - Log2 Fold change
  - P value
  - Corrected P value/FDR



STEP 5: ID Differentially Expressed Genes

- **DESeq2** Input: **RAW** count data , with each column representing a biological replicate/condition.

- DESeq2 Output: Normalized count data, but mainly a table with fold change, pvalue and adjusted pvalue for every gene/transcript.

- DESeq2 R commands available at: https://wikis.utexas.edu/display/bioiteam/Testing+for+differential+expression

STEP 4&5 – Quantification and ID Differentially Expressed Genes

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

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

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.

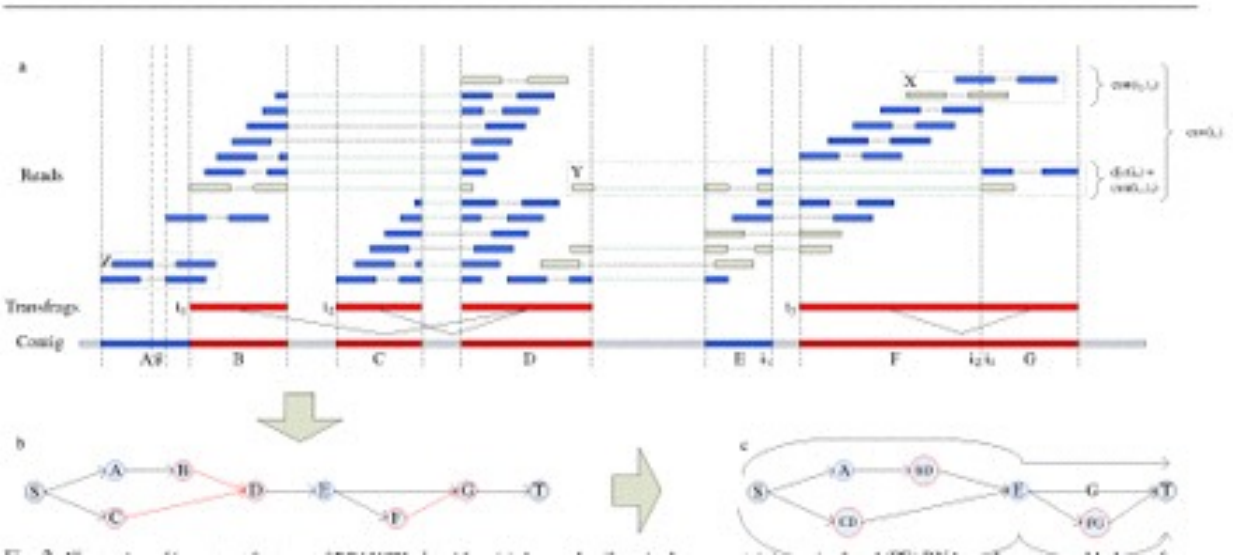
Optional: Assemble Transcripts

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).

Optional: Assemble Transcripts



doi: 10.1093/bioinformatics/btt127

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?

IDENTIFY NOVEL TRANSCRIPTS!!

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

Optional: Assemble Transcripts

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

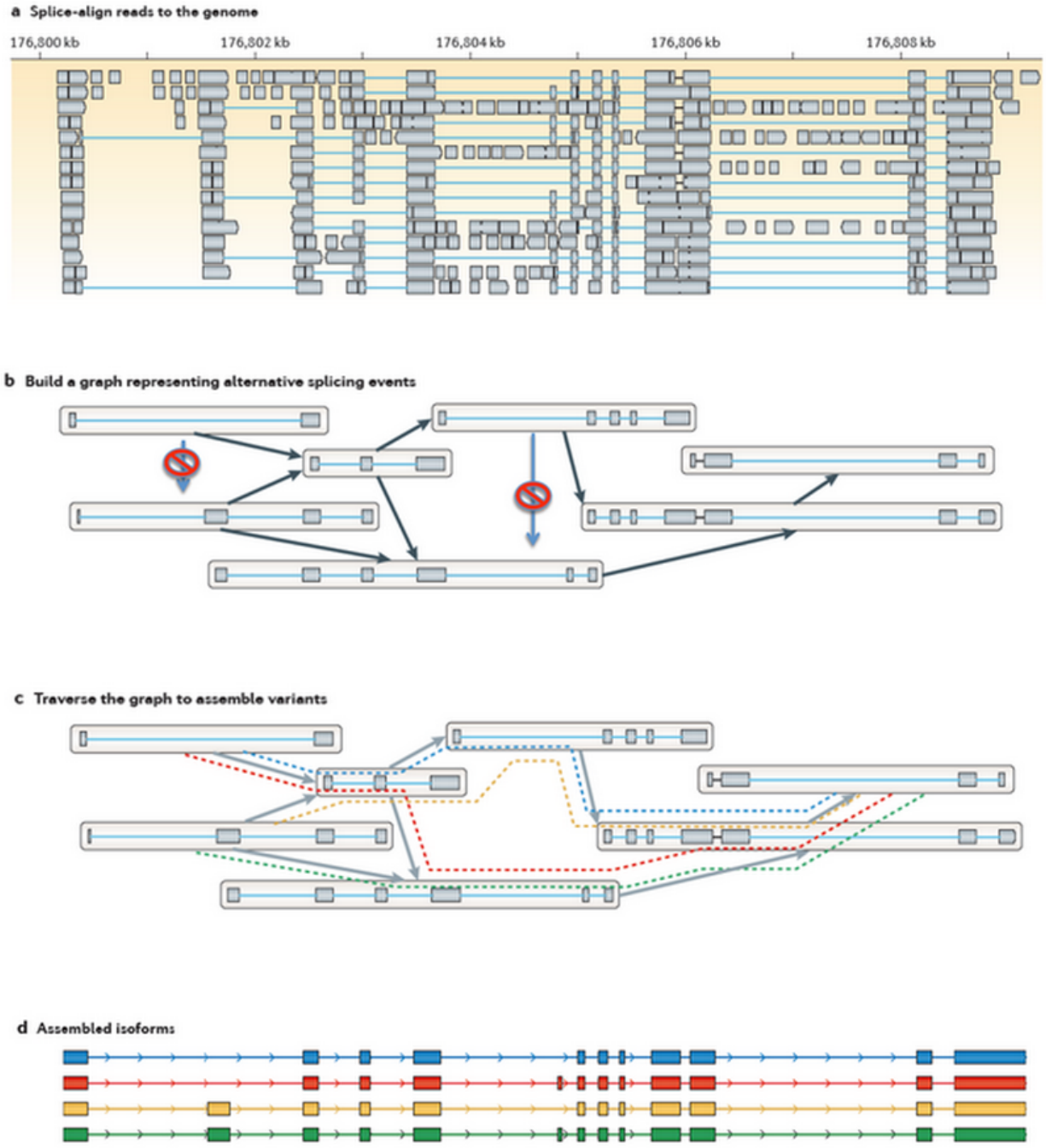
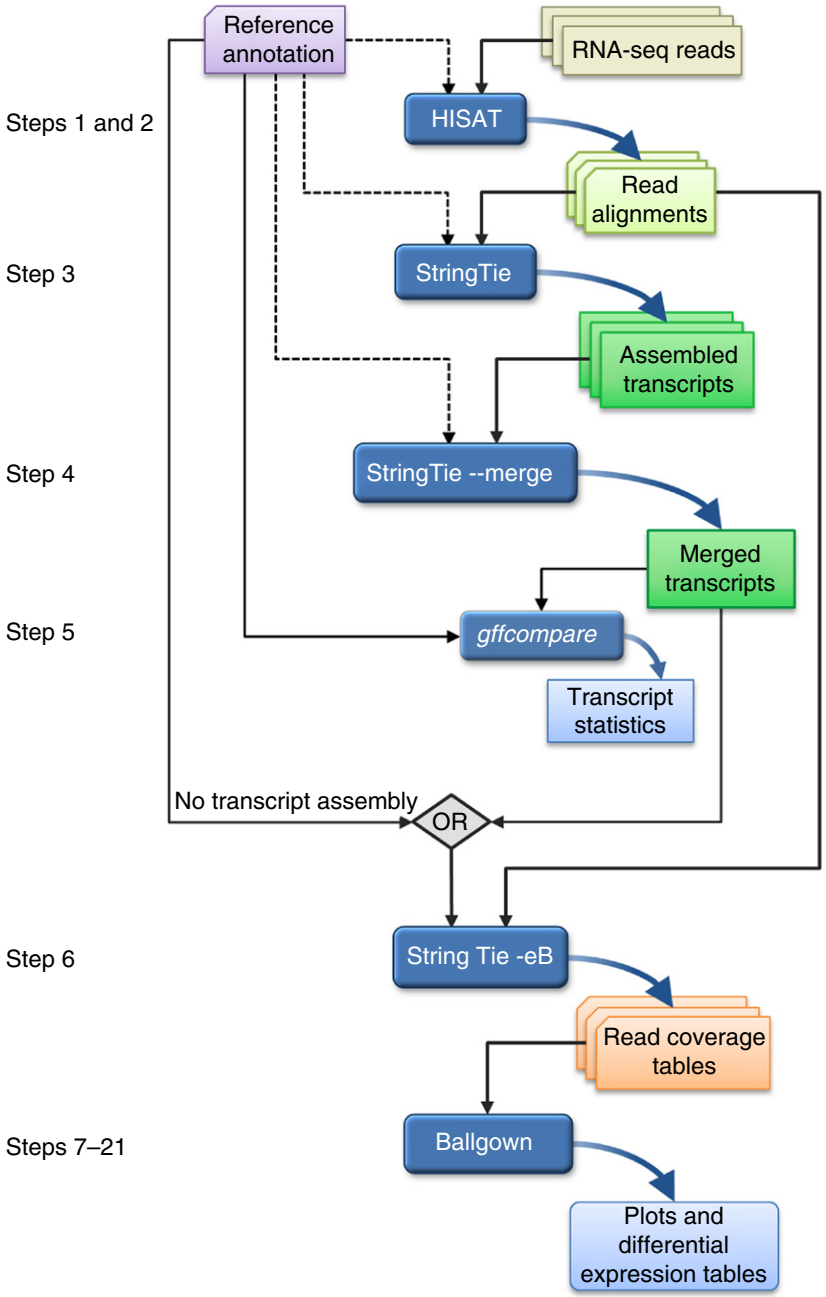


Figure :
http://sourceforge.net/projects/trinityrnaseq/files/misc/RNASEQ_WORKSHOP/rnaseq_workshop_slides.pdf

NEW TUXEDO PIPELINE

The pipeline is sequential.

Output of one step becomes input of next step.



DESeq/edgeR output vs New Tuxedo pipeline output

- DESeq2 can be used to generate differentially expressed genes too. So, why the big fuss?
  - From DESeq2, all genes are annotated genes. So, they all have flybase ids.
  - With this pipeline, our output will also have genes with ids ‘MSTR…’ - they are novel. (**stringtie, ballgown**)
  - It also gives us results telling us where our novel transcripts are with respect to the annotated ones. (**gffcompare**)

STEP 6: Visualize and Perfom Other Downstream Analysis

- Visualizations are useful for:
  - Identifying patterns/issues in the data
  - Summarizing results
  - Illustrating relationships between variables

STEP 6: Visualize

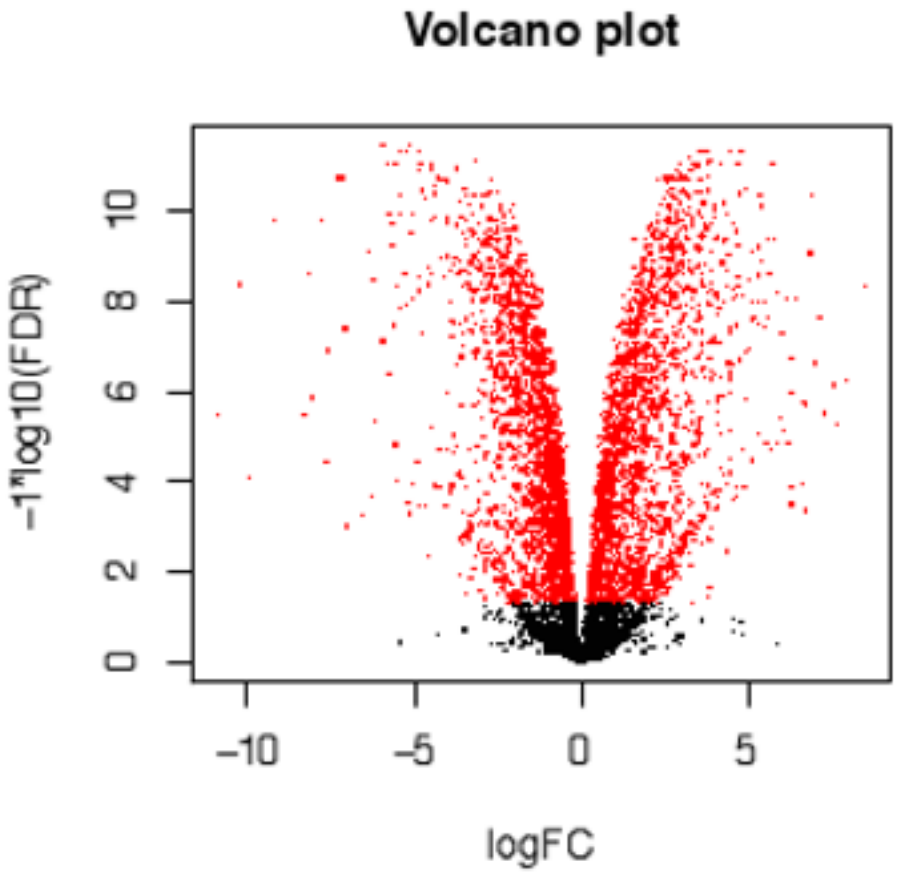
MA Plots

- For visualizing differences in measurements (in this case, gene expression) between two groups.
- M- log fold change (differences between two groups)
- A - mean gene expression (average value across samples)

STEP 6: Visualize

Volcano Plots

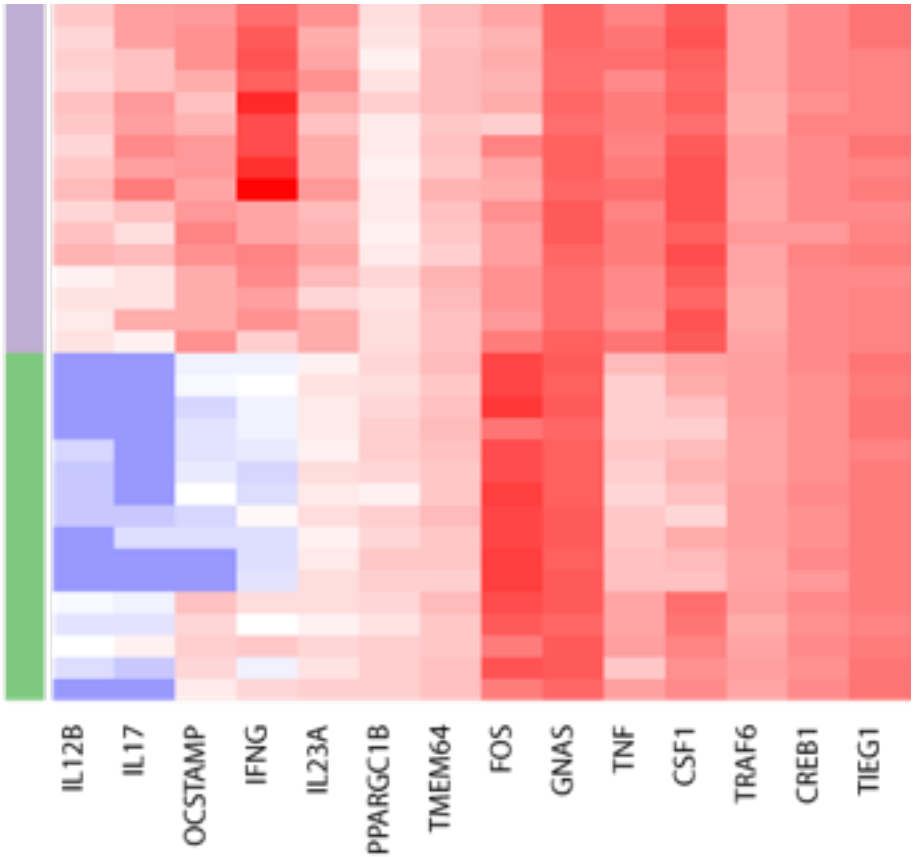
- Plots fold change vs significance value for all genes.
- Helps quickly see how many significantly differentially expressed genes are present.



STEP 6: Visualize

Heatmaps

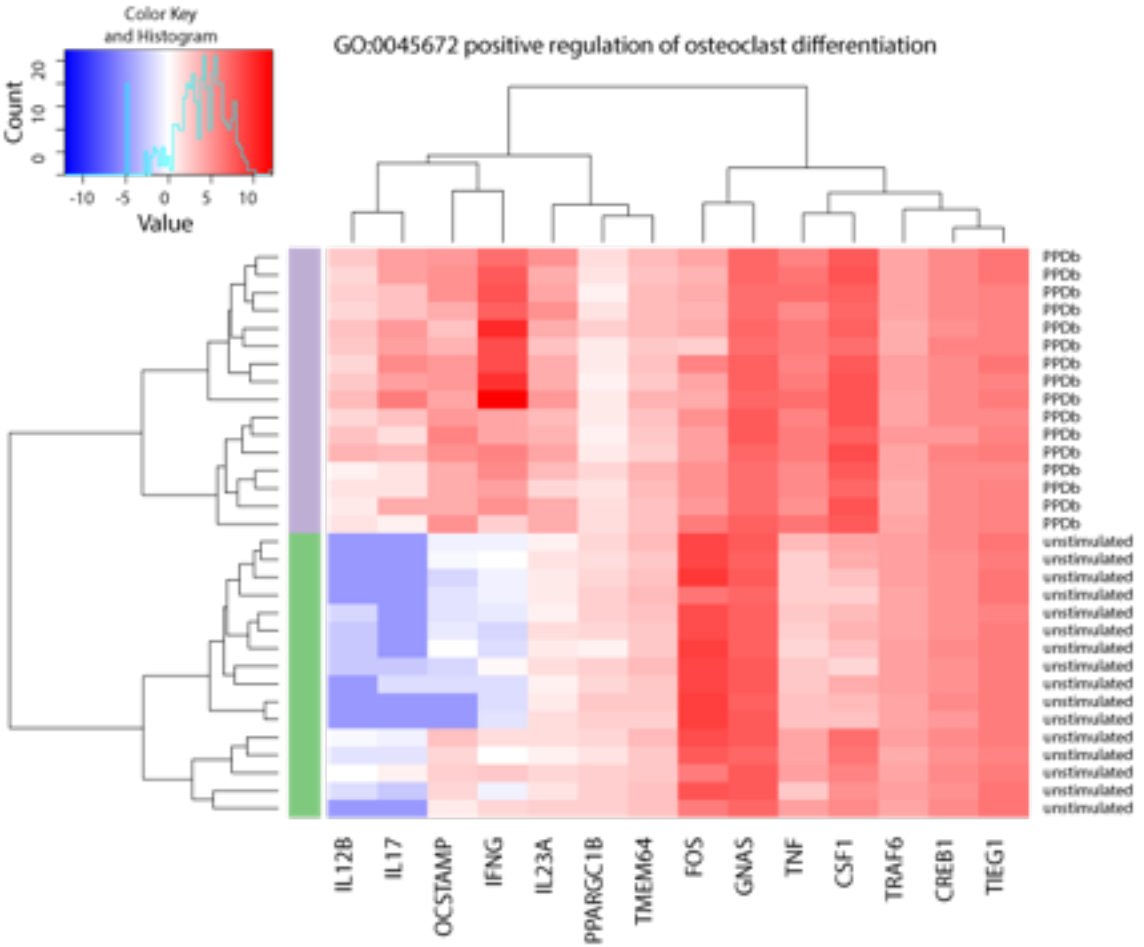
- Heat Maps represent gene expression by colors.
  - For visualizing how gene expression changes in different samples.
  - Columns are genes
  - Rows are Samples



STEP 6: Visualize

Heatmaps/Clustering

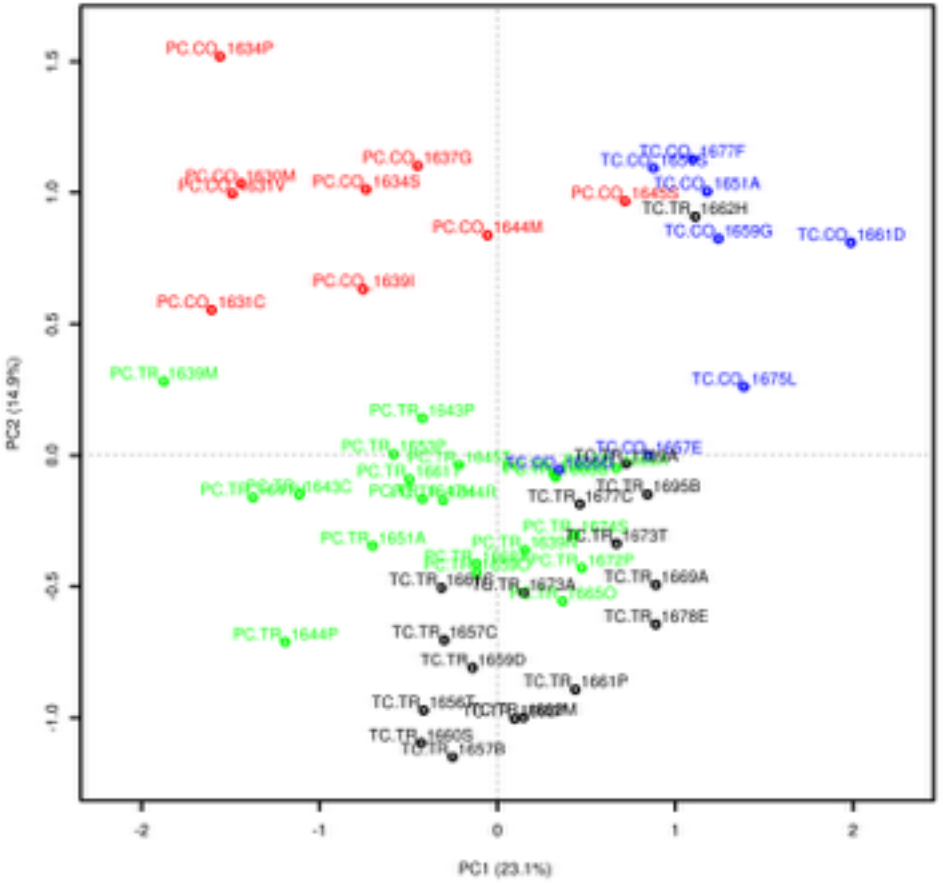
- Dendograms can be added to heat maps
- Samples can be clustered by gene expression
- Genes can be clustered by gene expression
    - time consuming for large number of genes



STEP 6: Visualize

Principal Component Analysis

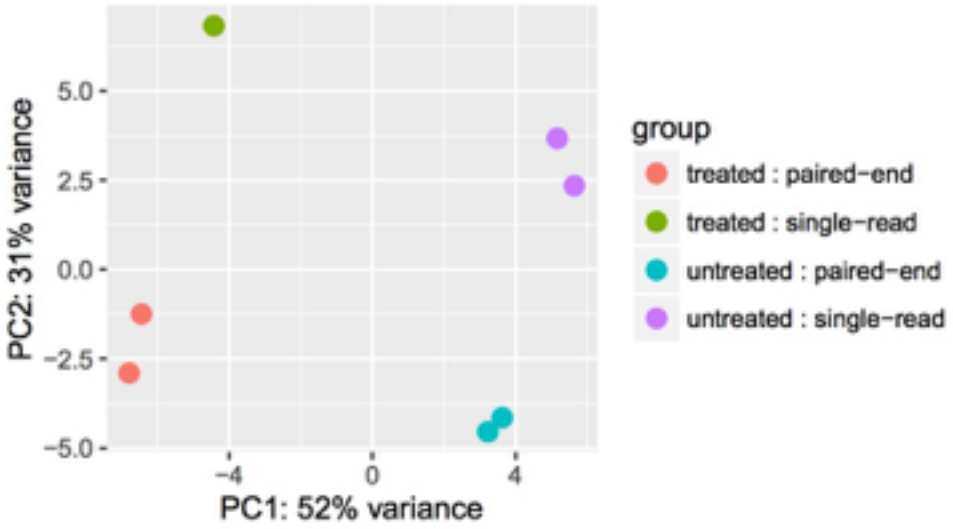
- Each principal component is one dimension in the data.
- Illustrates how the data groups based on the dimensions that represent the highest variability.



STEP 6: Visualize

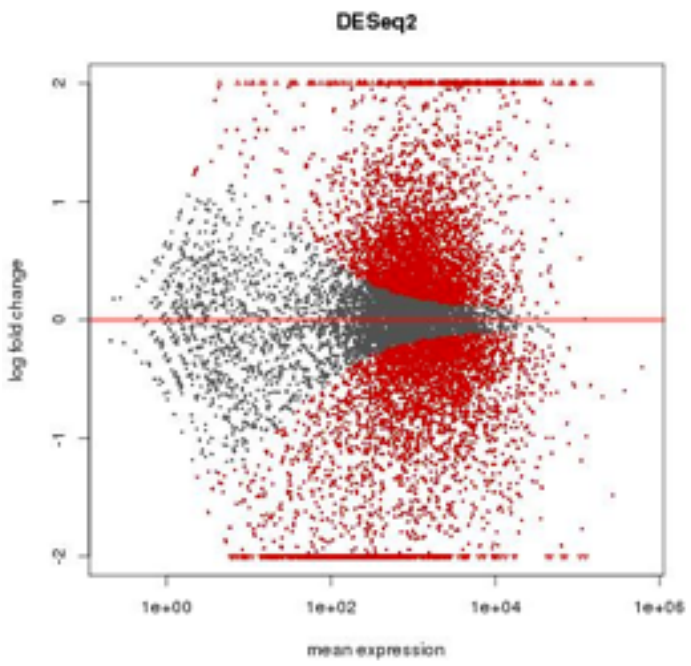
Principal Component Analysis

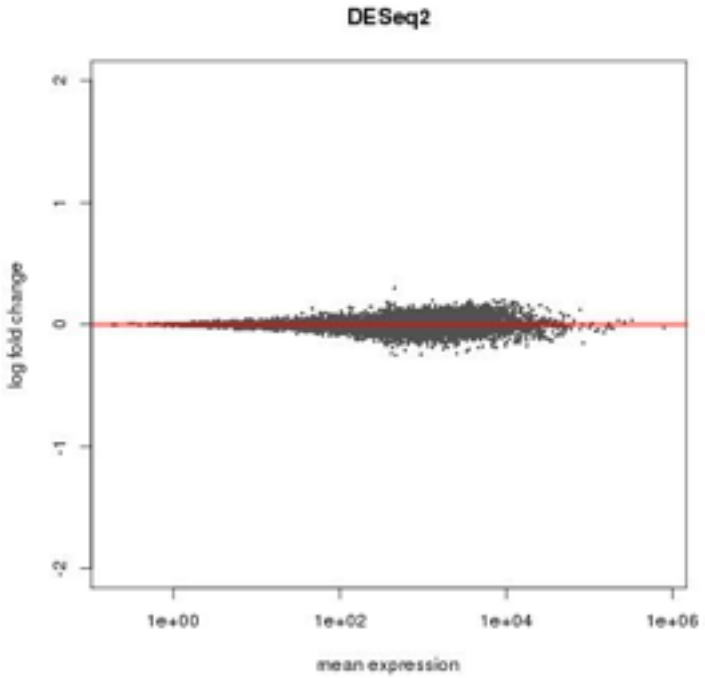
- Each principal component is one dimension in the data.
- Illustrates how the data groups based on the dimensions that represent the highest variability.



Looking at Some Real Data

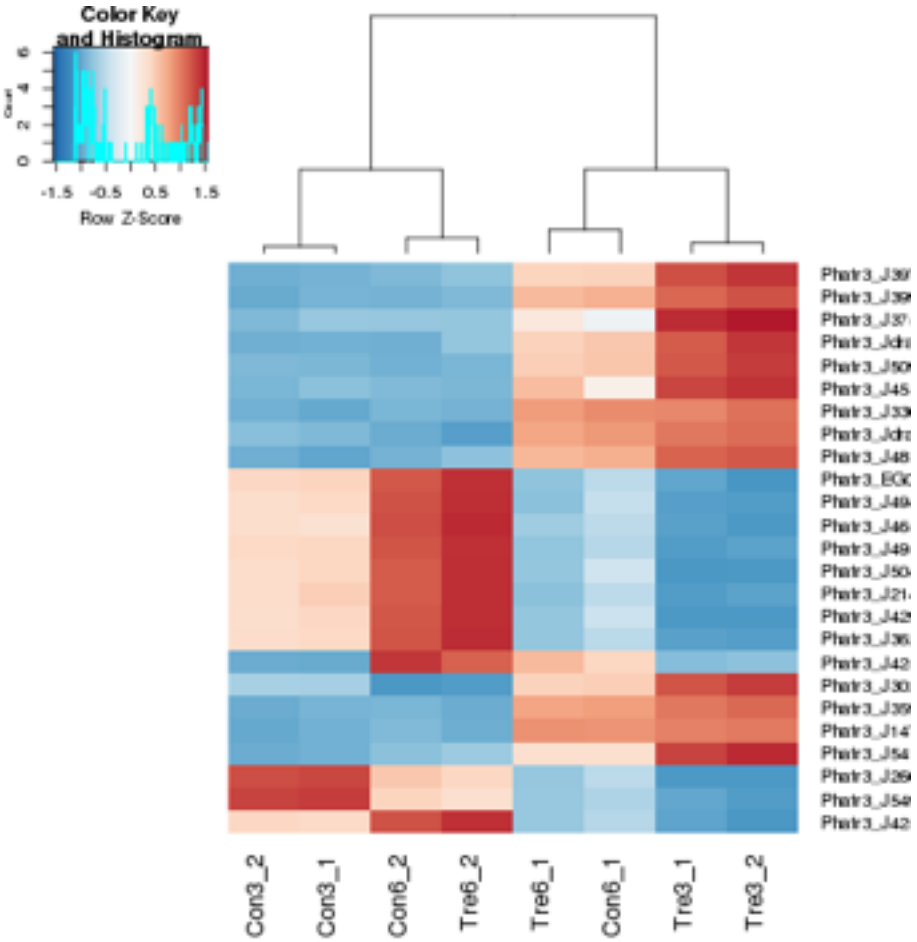
- **Mysterious results for an experiment with 6 samples across:**
  - 2 different time points, 2 different conditions: control vs treated. 3replicates each.

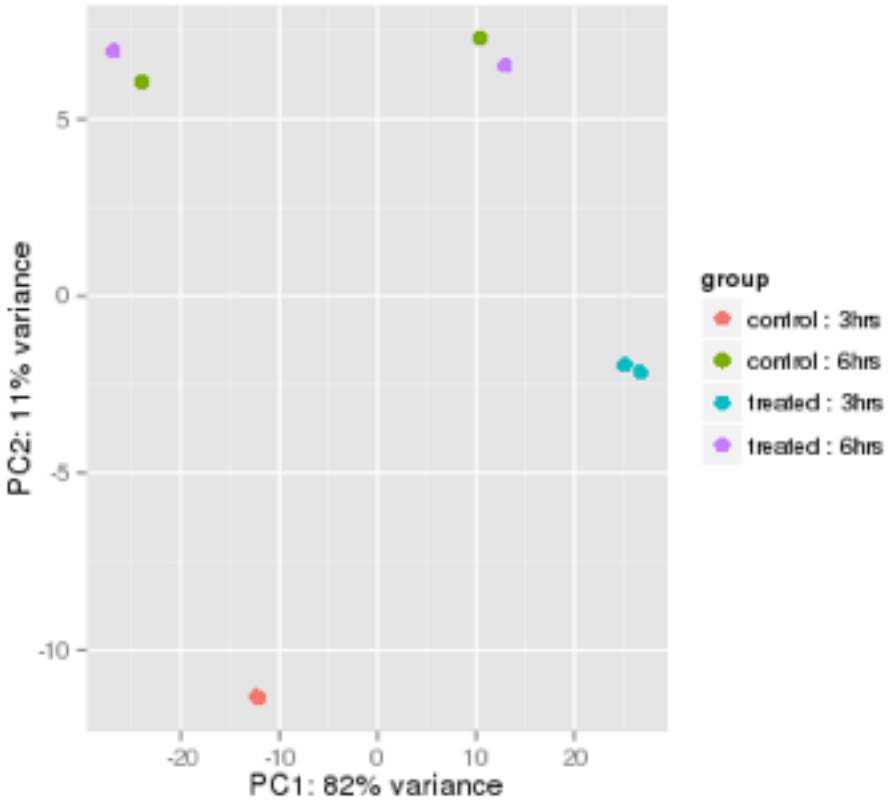




Looking at Some Real Data

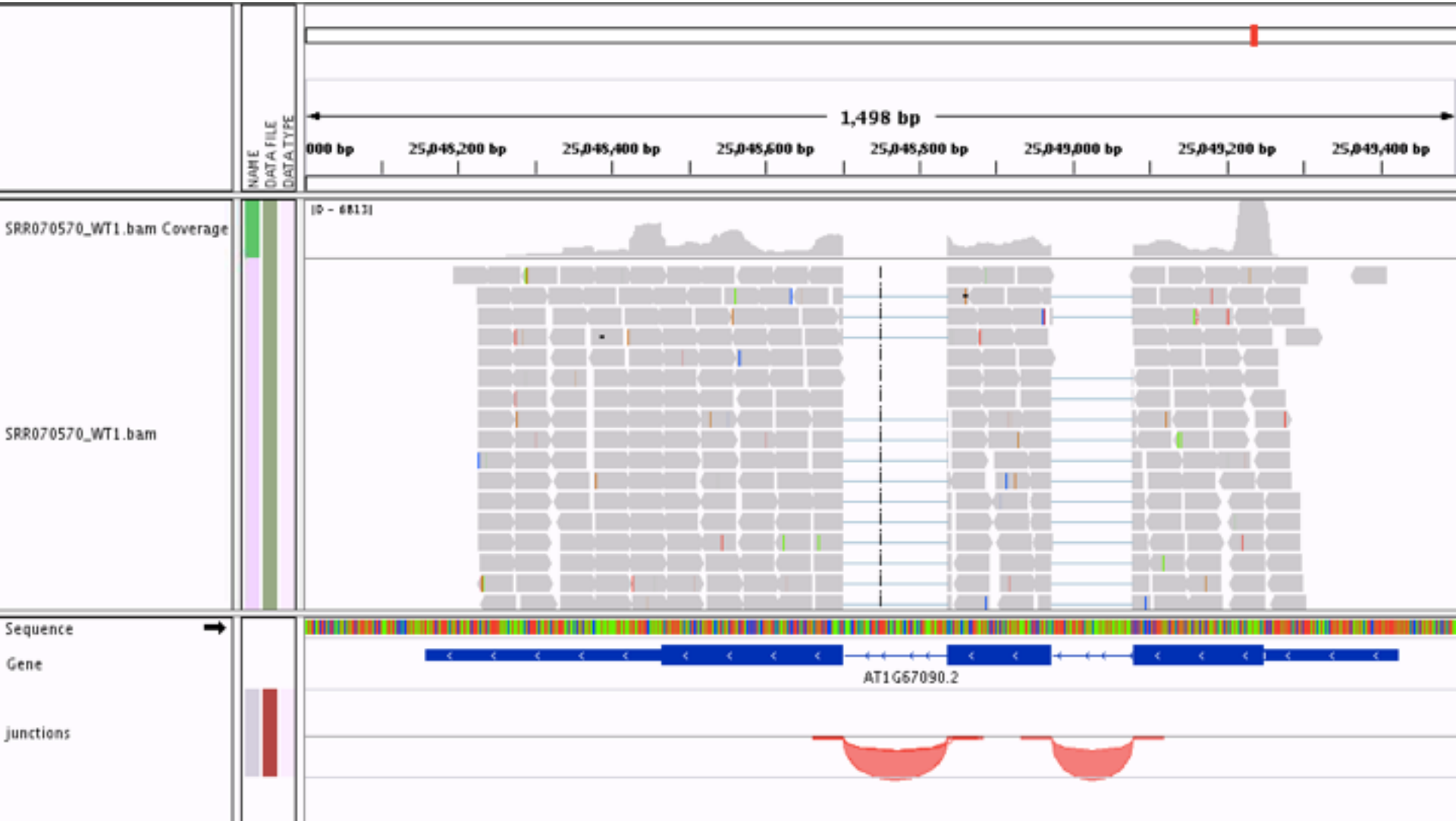
- **Can these plots inform us about what might be going on?**





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 **GORILLA**

- WHAT ARE GO TERMS?
- Standardized vocabulary to describe genes and gene products from different species. GO terms allow us to assign functionality to genes.

    - **cellular component**, describes where in a cell a gene acts, what cellular unit the gene is part of
    - **molecular function**, describes the function carried out by the gene, such as binding or catalysis;
    - **biological process**, a set of molecular functions, with a defined beginning and end, makes up a biological process.
    - All GO terms have an ID that looks like GO:0006260.
    - All GO terms have a list of genes that belong to that particular term.

STEP 6: Perform Other Downstream Analysis

GORILLA

Gene Ontology enRIchment anaLysis and visuaLizAtion tool

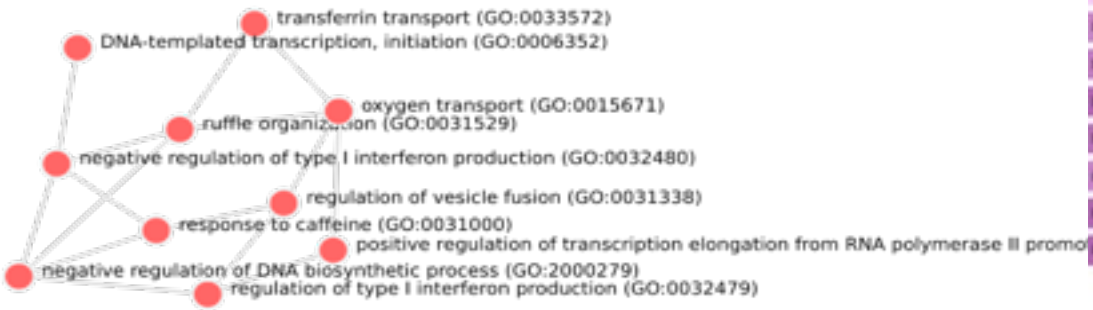
- For GO enrichment, we take the following things into account:
  A. Total number of genes we are looking at.
  B. Number of genes of interest, that is, in our DEG list.
  C. Total number of genes in the GO term
  D. Number of genes from our genes of interest that are also in the GO term.

- 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: Perform Other Downstream Analysis

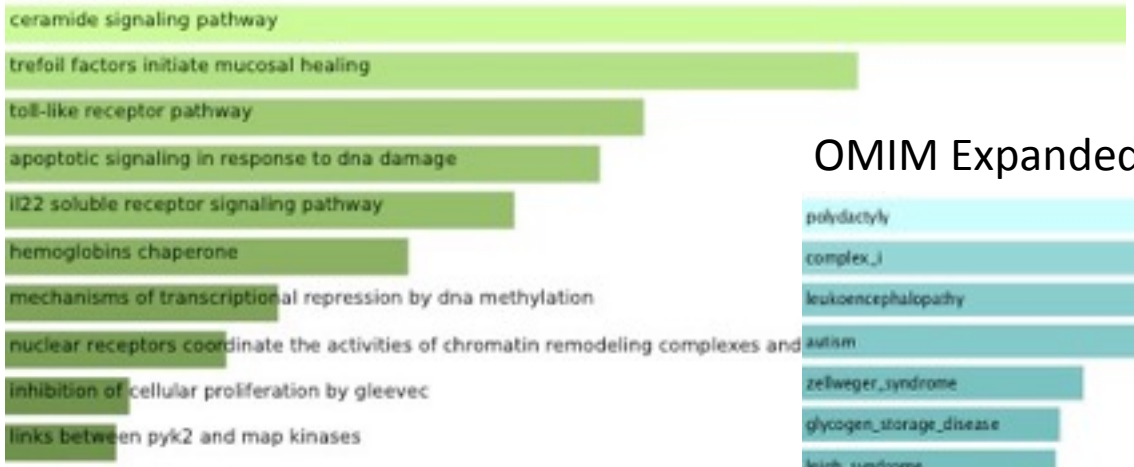
- Enrichr- GUI for GO/pathway enrichment analysis

  amp.pharm.mssm.edu/Enrichr



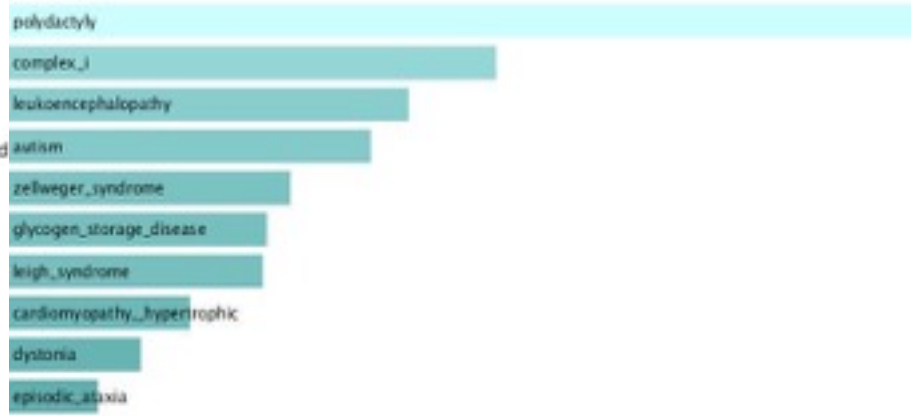
GO terms network graph

ENCODE histone modifications

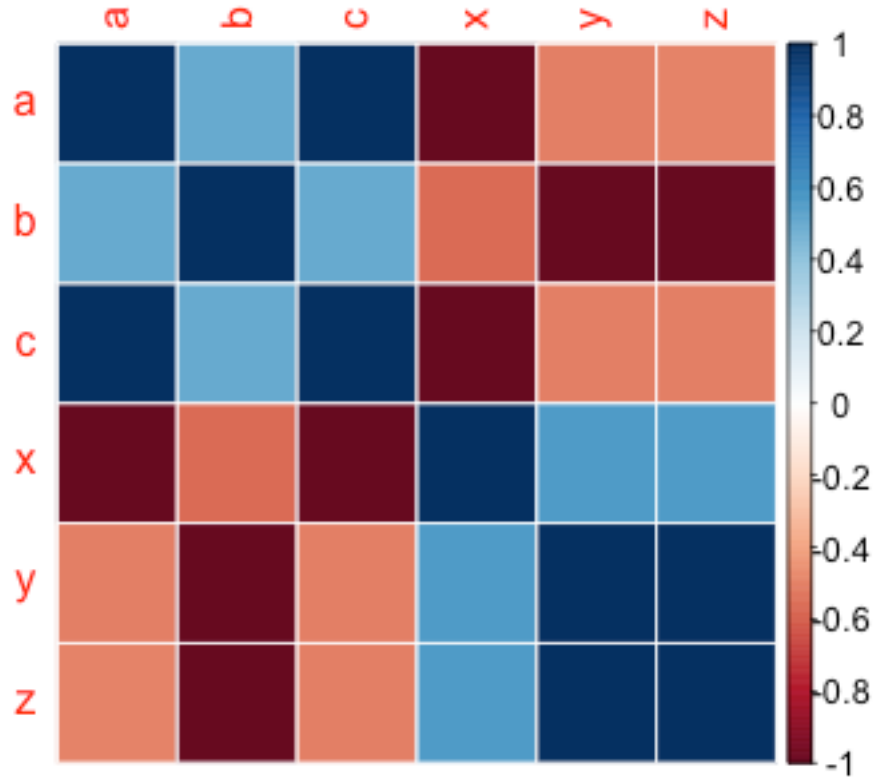


Biocarta pathways bar chart

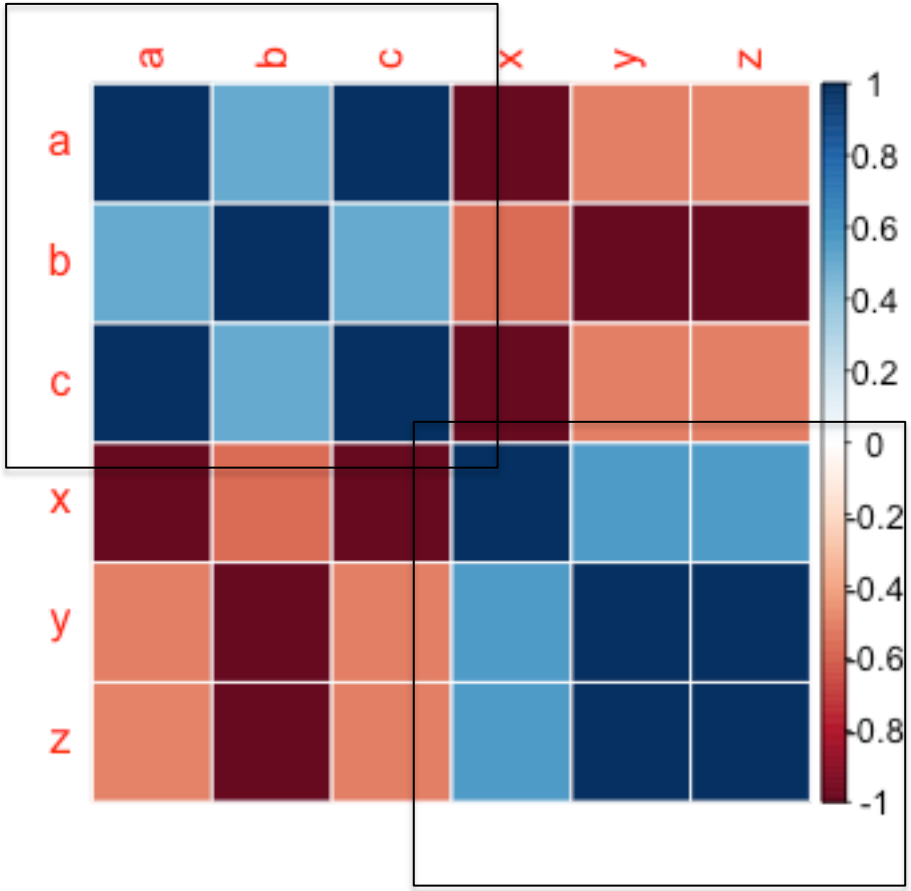
OMIM Expanded terms



Co-regulated genes have correlated expression



Co-regulated genes have correlated expression

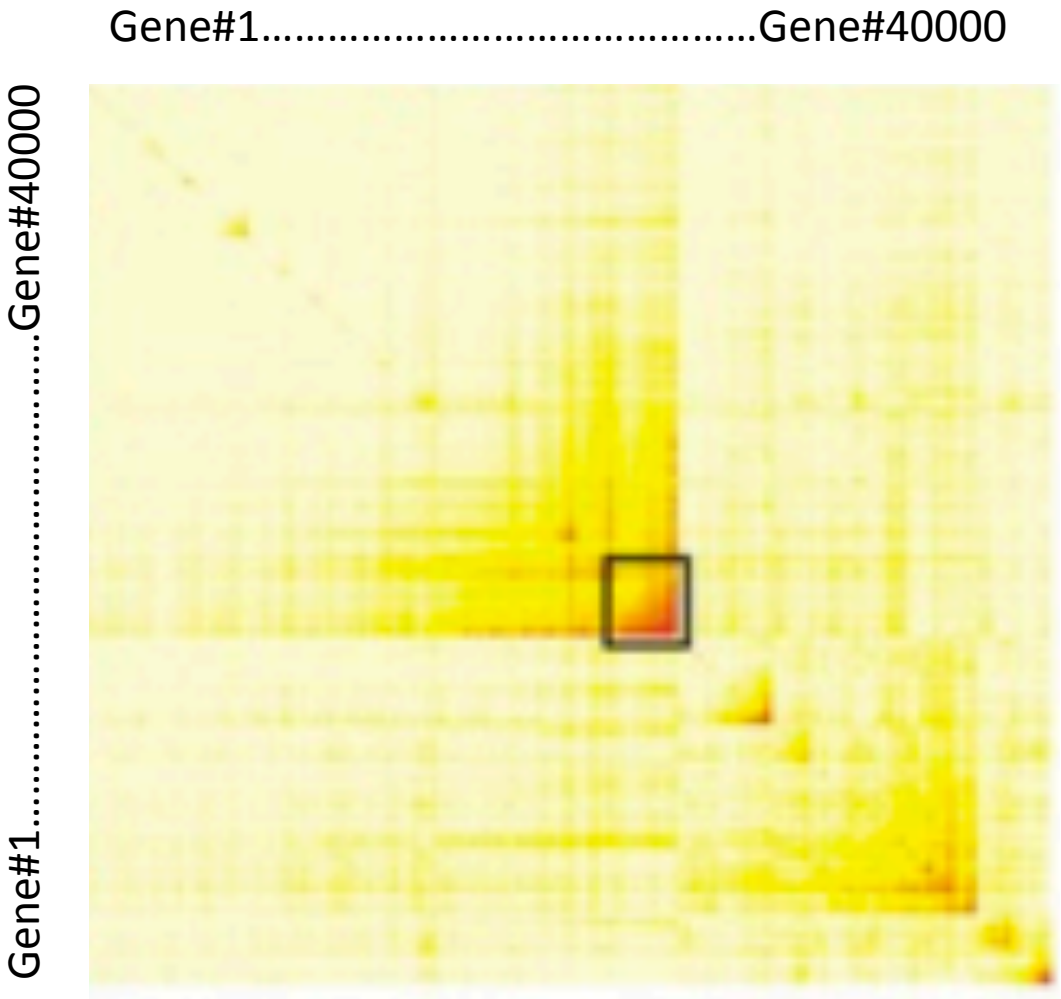


A <-> B <-> C

and

X <-> Y <-> Z

Now with 40,000 genes…

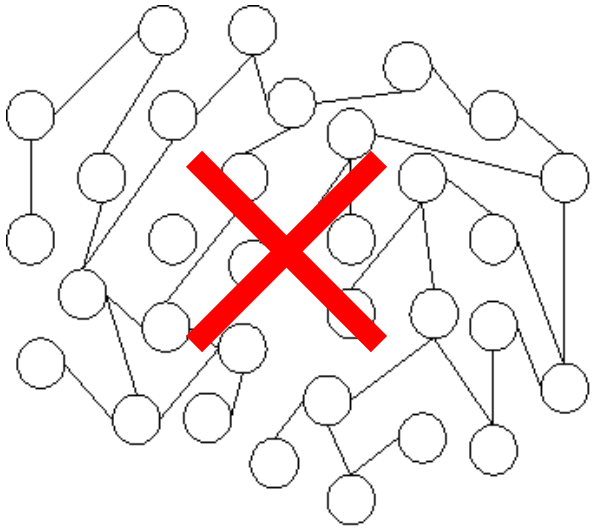


Pro tip: use a supercomputer

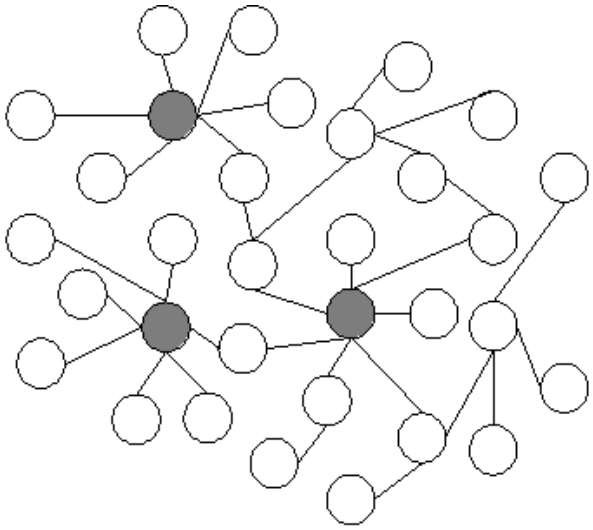
STEP 6: Perform Other Downstream Analysis

WGCNA- Weighted Gene Co-expression Network Analysis

We assume a scale-free topology



(a) Random network

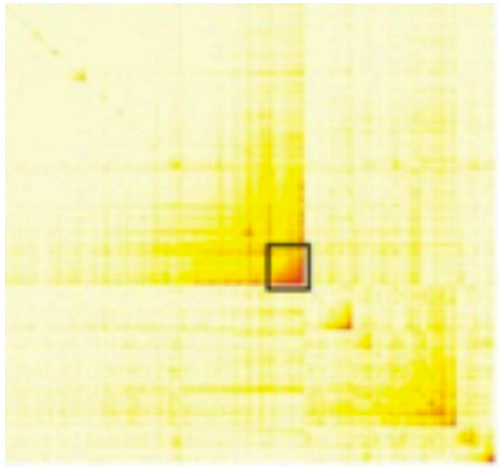


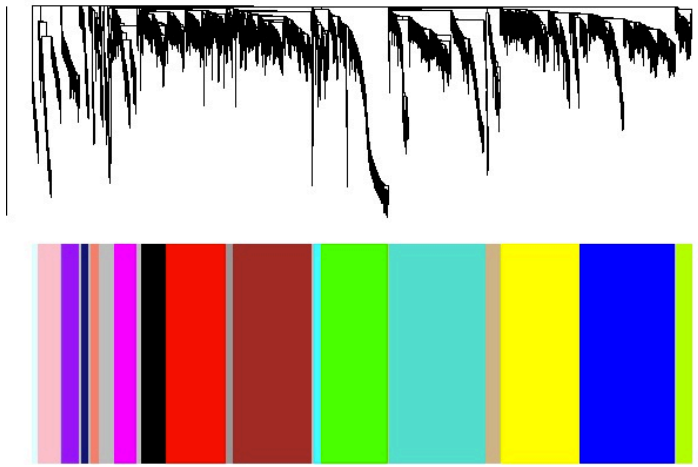
(b) Scale-free network

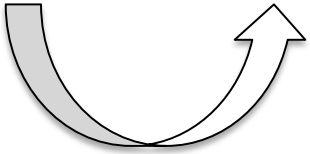
STEP 6: Perform Other Downstream Analysis

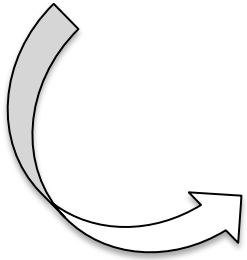
WGCNA- Weighted Gene Co-expression Network Analysis

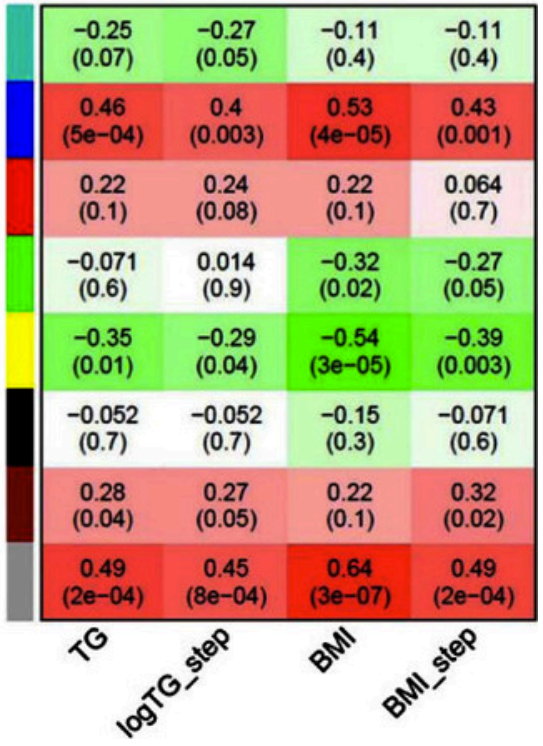
- Identify groups (modules) of co-regulated genes
- Correlate module expression values to trait data











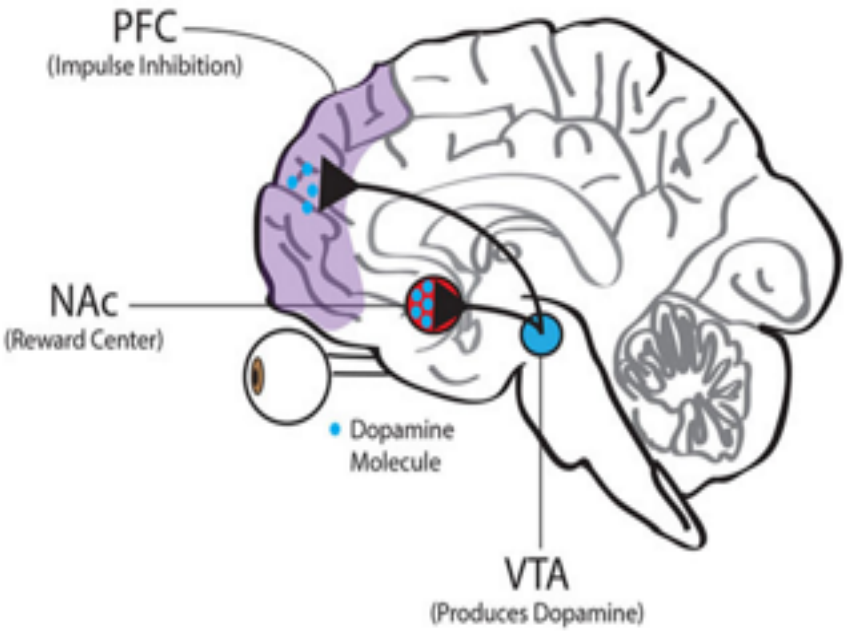
Thank you!

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

Apendix: RNA-SEQ Example Study

What Does Our Data Look Like?

- 60 matched brain samples
- 30 Alc, 30 Controls
- 4 different brain regions
- Total RNA for samples with RIN >5
- ~60-100x coverage of the transcriptome

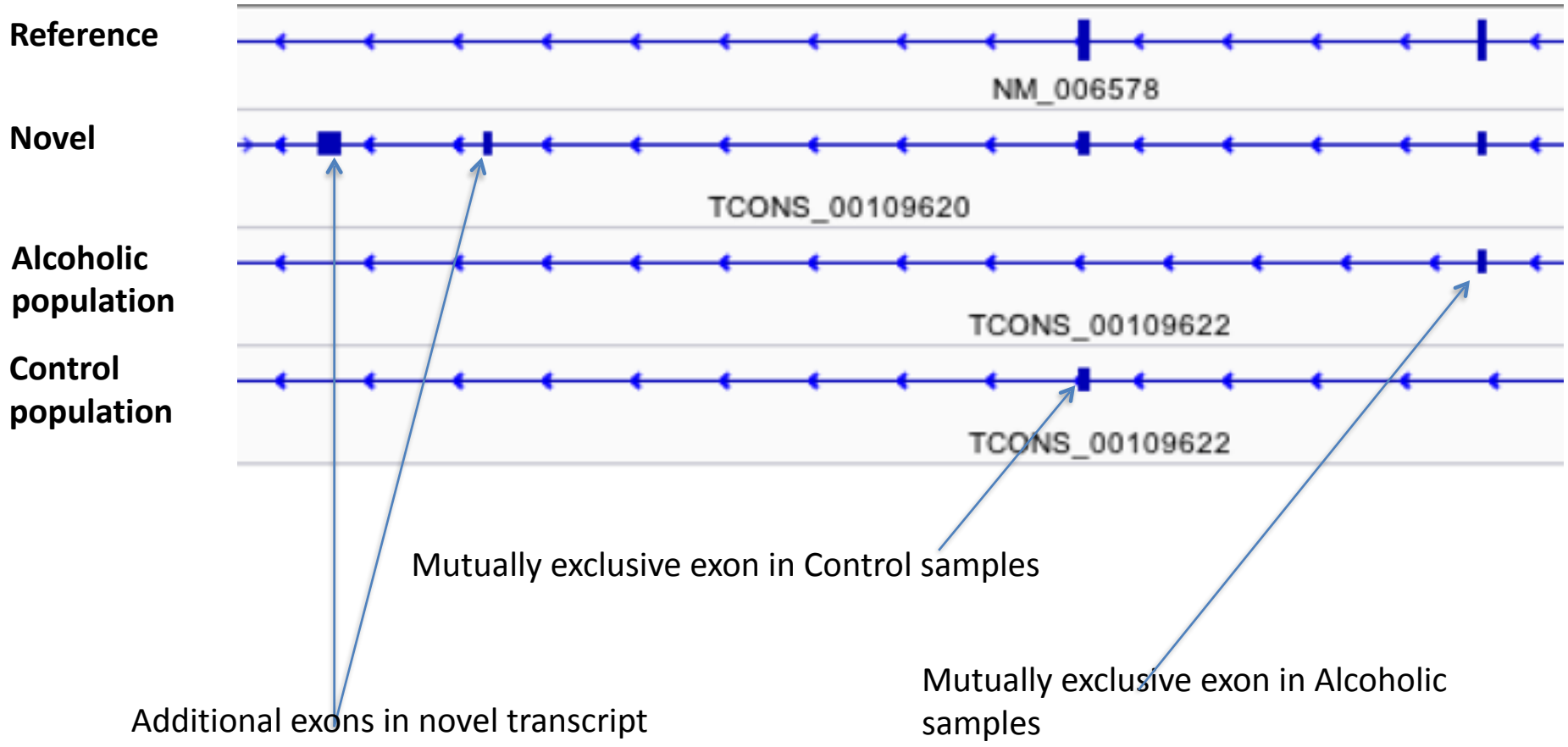


- Covariates: Age, Gender, Ethnicity, Cause of Death, Age of drinking onset, Duration, Daily Quantity, Number of drinks/day, Lifetime Consumption

What do we do with it?

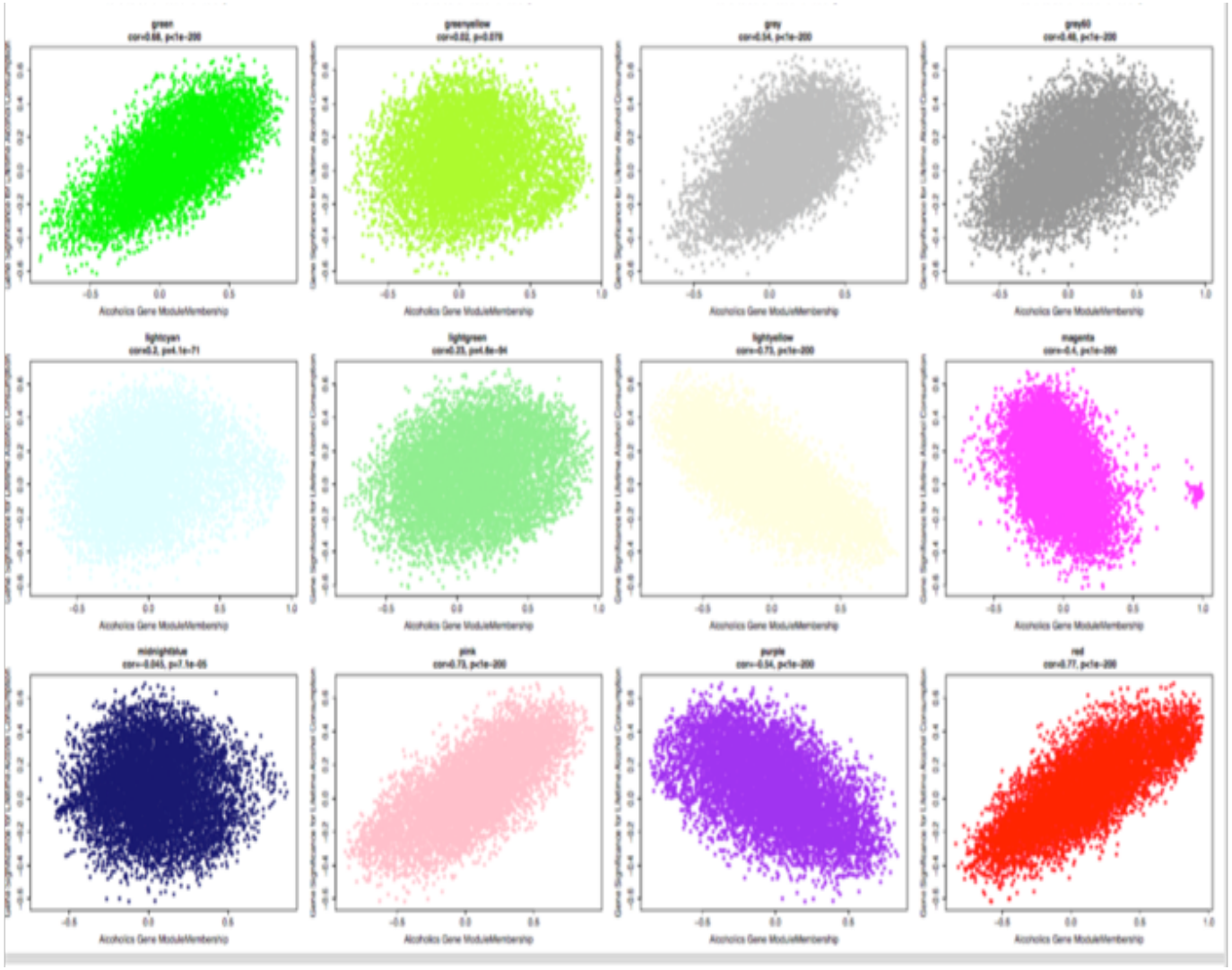
What kind of results?

NOVEL TRANSCRIPTS,
DIFFERENCES IN SPLICING



What kind of results?

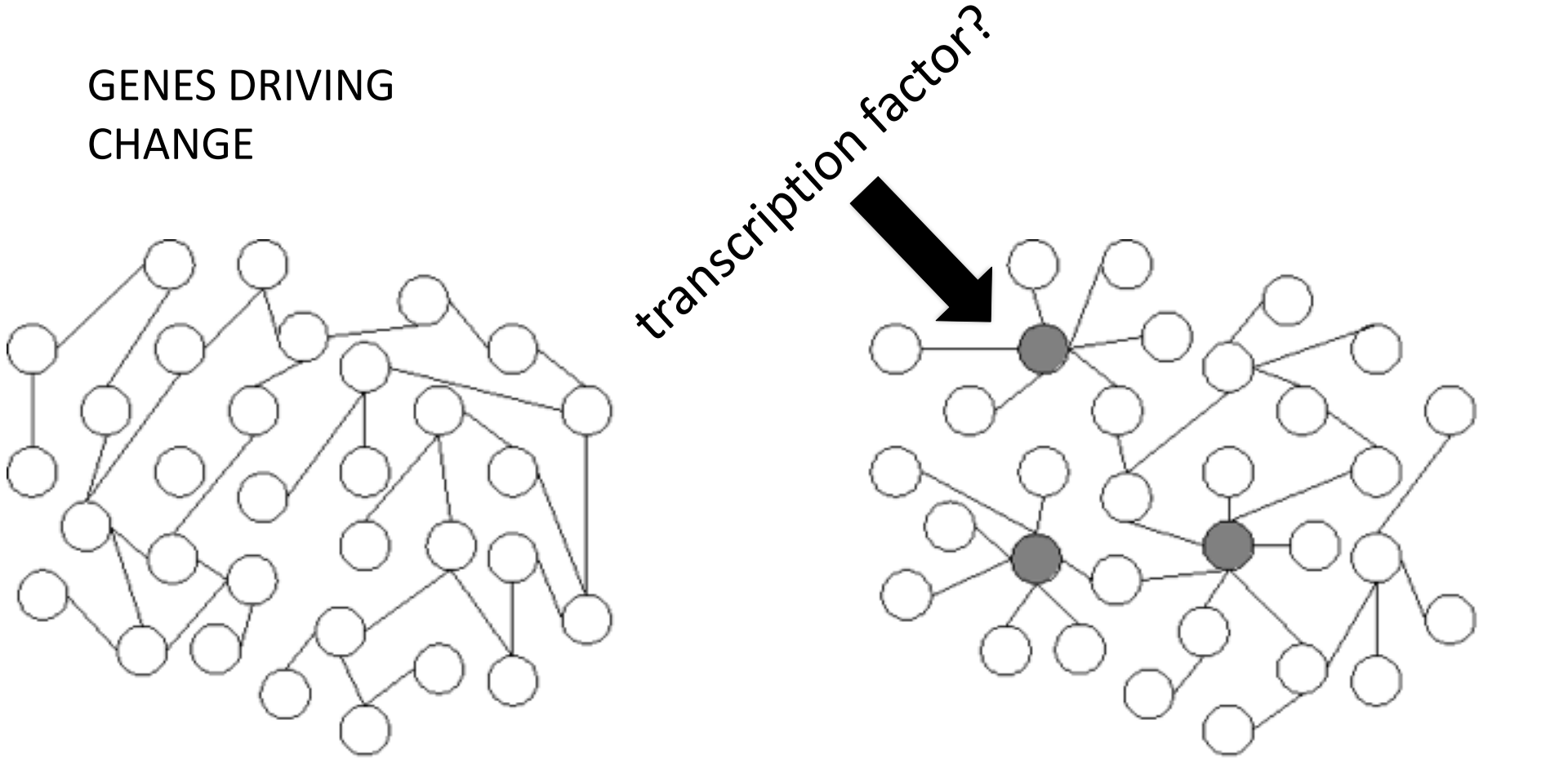
Module Correlation to Lifetime Alcohol Consumption in CNA Brain Region



COEXPRESSED
GENE MODULES

What kind of results?

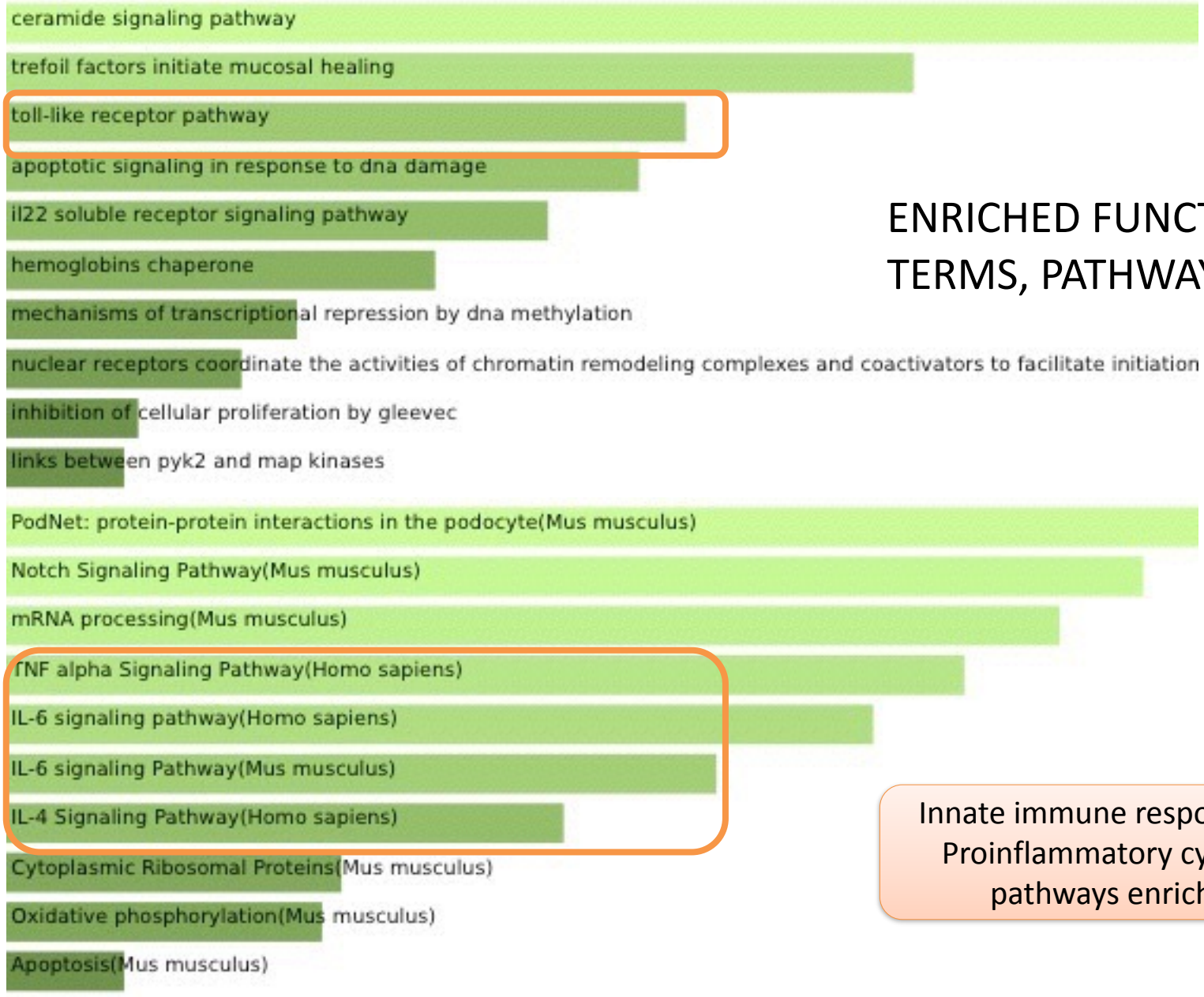
GENES DRIVING
CHANGE



(a) Random network

(b) Scale-free network

What kind of results?

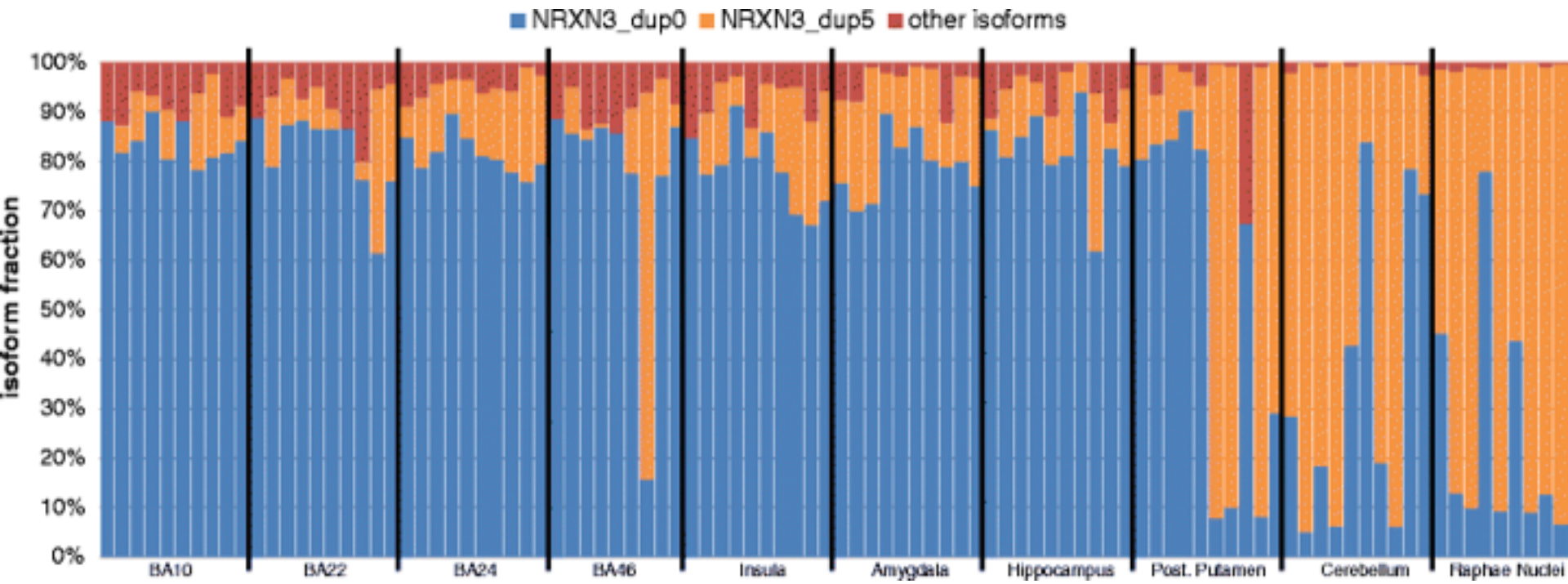


ENRICHED FUNCTIONAL TERMS, PATHWAYS

Innate immune response and Proinflammatory cytokine pathways enriched

What kind of results?

COMPARISON ACROSS BRAIN REGIONS



RNA sequencing of transcriptomes in human brain regions: protein-coding and non-coding RNAs, isoforms and alleles

Webb et al, 2015

What kind of results would we like?

DRUG COMPOUNDS WHICH CAN REGULATE GENE CHANGES

FDA Approved drugs in NAC Module

Contributed by Boone Miller

Drug Name	FDA Status	Connectivity Score
formestane	Approved, Investigational, Withdrawn	-95.11
metolazone	Approved	-96.17
Erlotinib	Approved	-96.38
timolol	Approved	-95.3
finasteride	Approved	-95.65
flunarizine	Approved	-98.77
alprazolam	Approved, Illicit, Investigational	-95.95
ipratropium bromide	Approved	-96.71
labetalol	Approved	-95.56
etodolac	Approved, Investigational, Vet Approved	-95.18
clofibrate	Approved	-95.91

What kind of results would we like?

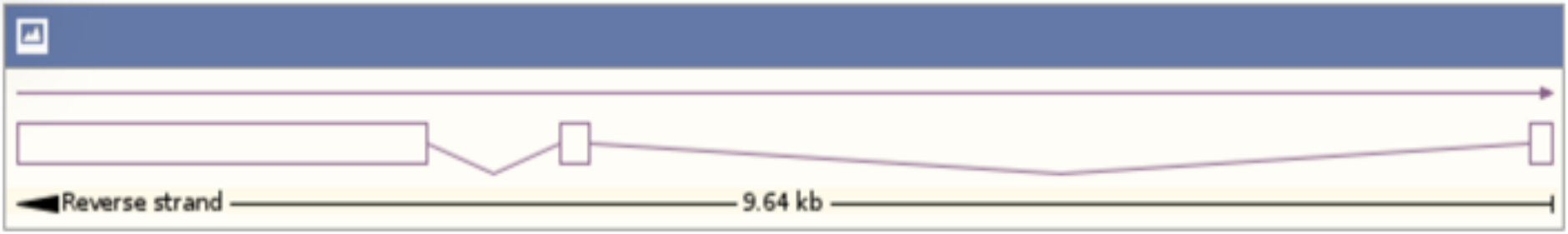
DIFFERENCES IN NON-CODING RNA

Contributed by: Edgar Marroquin

Transcript: LINC01567-001 ENST00000414816.1

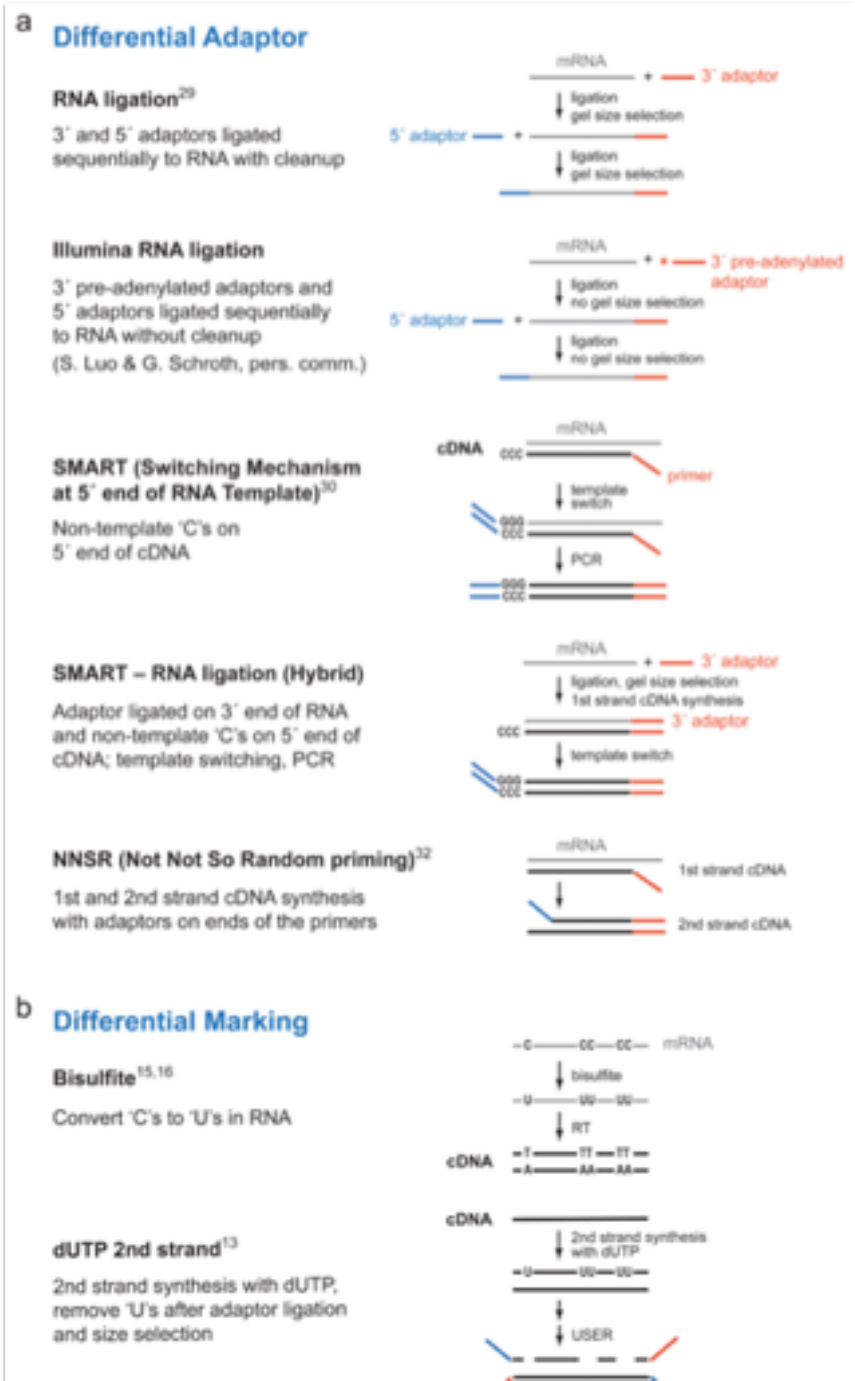
Description	long intergenic non-protein coding RNA 1567 [Source:HGNC Symbol;Acc:HGNC:51367]
Location	Chromosome 16: 24,661,422-24,671,062 reverse strand.
About this transcript	This transcript has 3 exons and maps to 13 oligo probes.
Gene	This transcript is a product of gene ENSG00000224310 Show transcript table

Summary



THANK YOU!

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

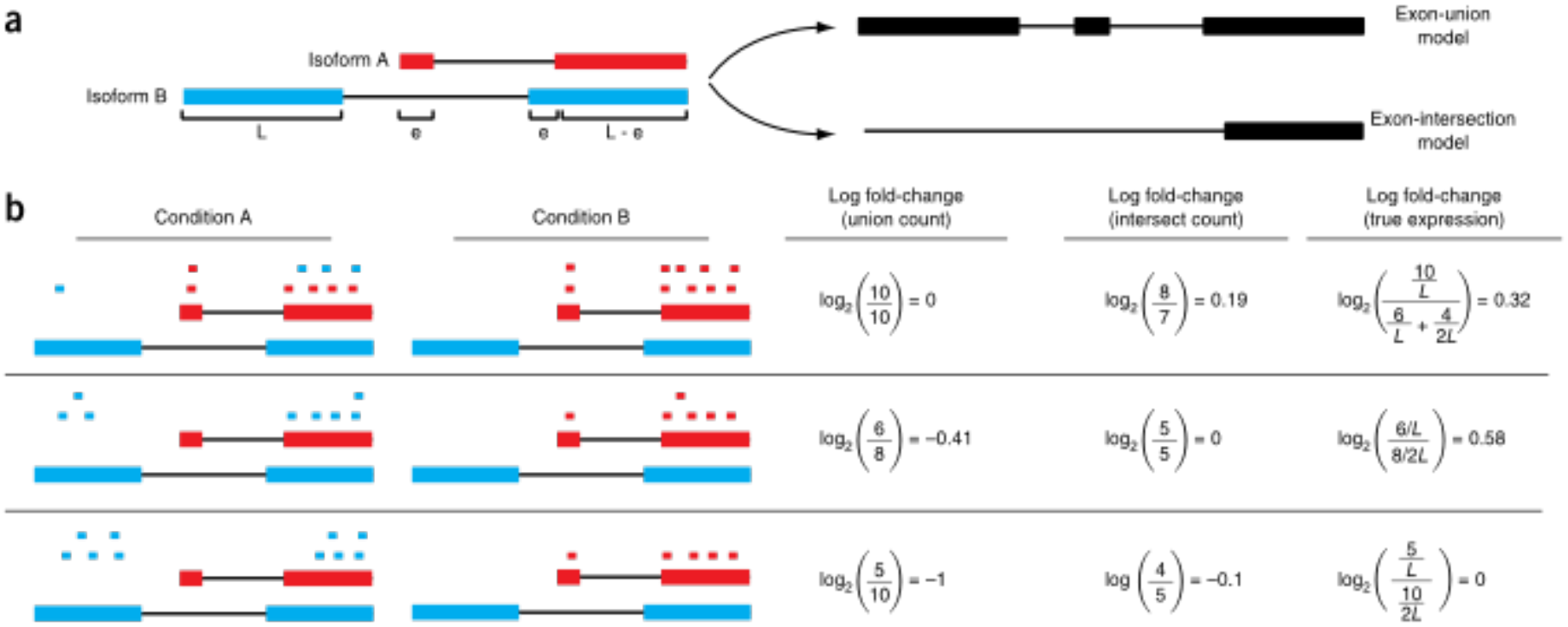


Figure 1 Changes in fragment count for a gene does not necessarily equal a change in expression. (**a**) Simple read-counting schemes sum the fragments incident on a gene's exons. The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads 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