Introduction to NGS and RNA-Seq

Dhivya Arasappan

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

Some background

- Research scientist- bioinformatician at CBRS.
  - RNA-Seq
  - Genome Assembly
  - Exome data analysis
  - Benchmarking of tools

Center for Biomedical Research Support
Enabling Cutting-Edge Research that Changes the World
THE UNIVERSITY OF TEXAS AT AUSTIN
102 E 24th Street - C4500 • Austin, TX 78712 • Tel (512) 471-5261 biomedsupport.utexas.edu

FRI

BIG DATA IN BIOLOGY

- Assistant Professor of Practice
  - Training grad students, post-docs.
  - Undergraduate- FRI

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?
- Hands-on experience running typical RNA-Seq workflows on TACC
  - Some unix, R, TACC skills
- Learn the terminology
- Brief introductions to 3' targeted RNA-Seq (tag-seq) and Single Cell RNA-Seq.

Setting General Expectations

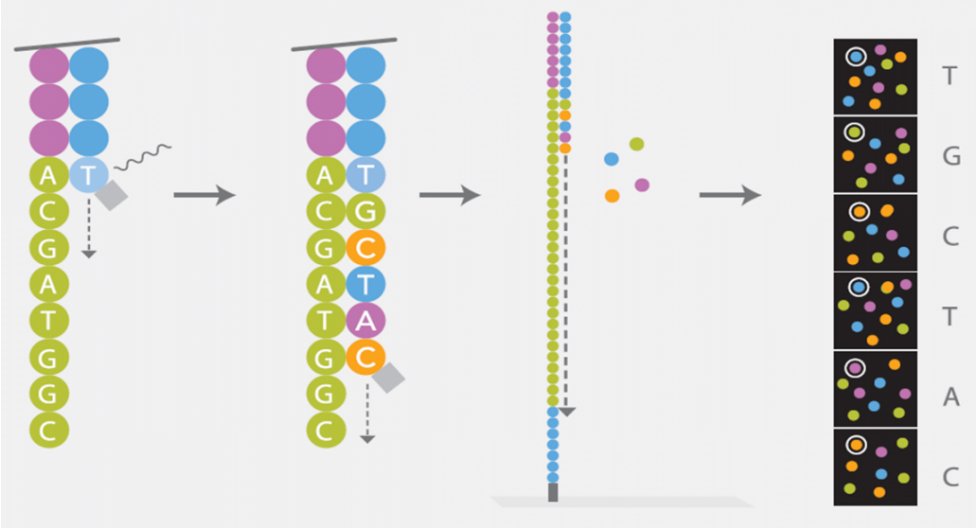
- Lots of background and basics to provide comfort with terminology and key concepts.
- Exposure to commands and typically used analysis tools using an example RNA-Seq dataset.
  - No one ‘best’ or ‘standard’ tool.
- A starting point for you to design your RNA-Seq study or analyze your dataset.
- First time teaching it virtually

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
- CBRS Bioinformatics consultants: https://research.utexas.edu/cbrs/cores/cbb/bioinformatics-services/

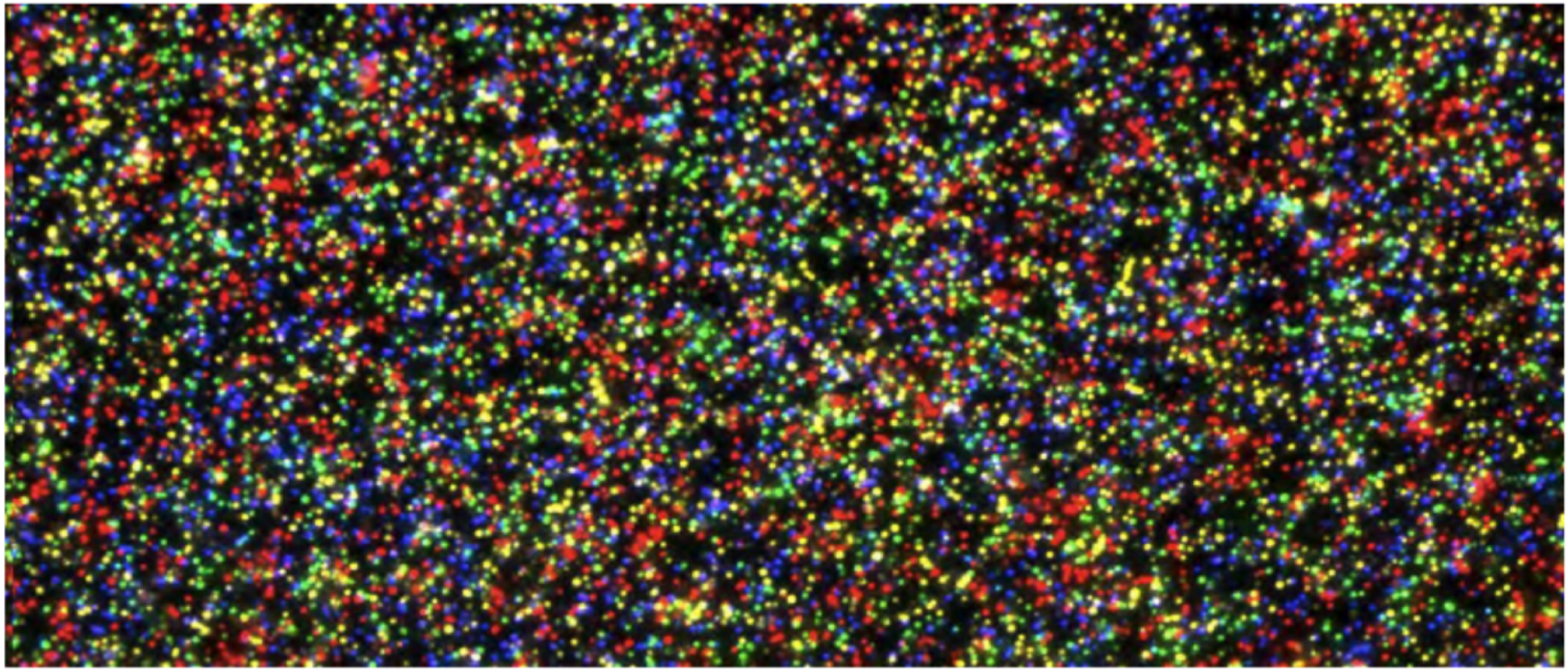
Second Generation Sequencing (or) Next Generation Sequencing

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



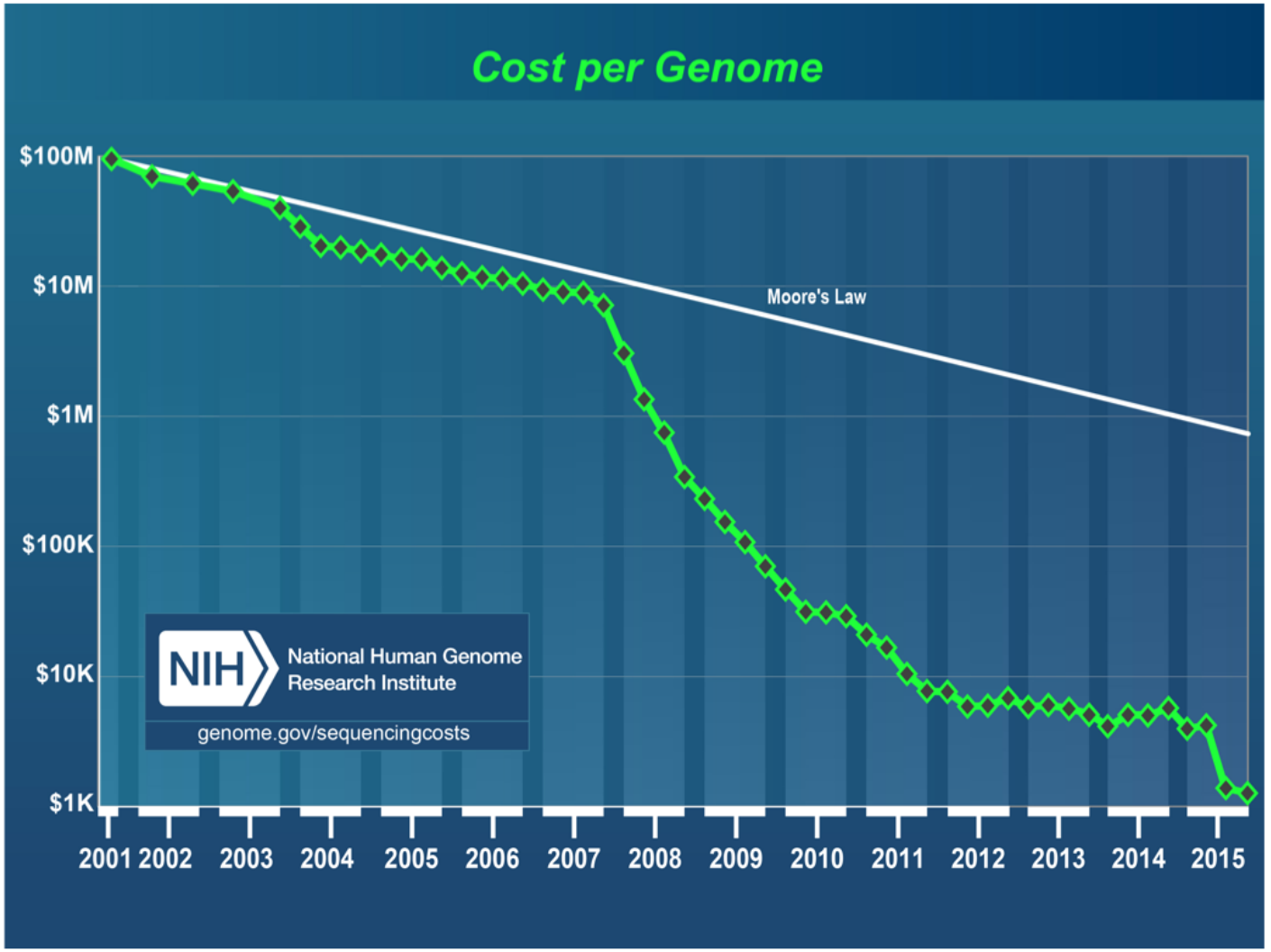
http://www.cegat.de/

How do next generation sequencers work?!



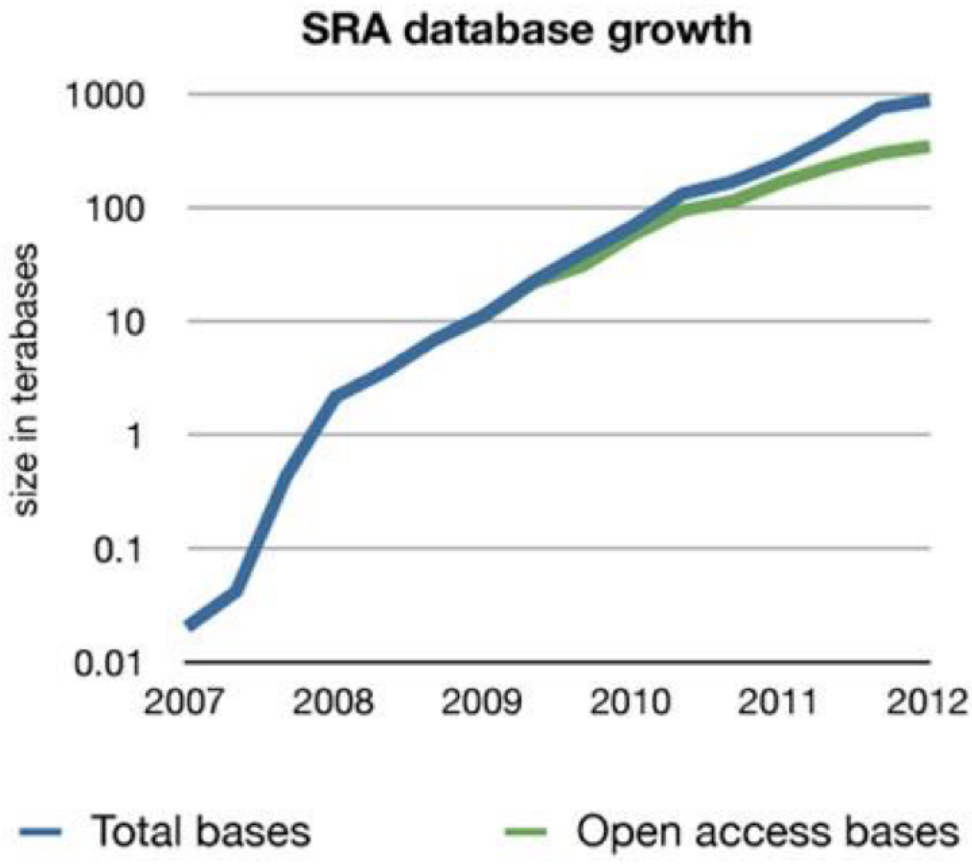
So, what's so great about second generation sequencing?

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



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

- Data deluge
- Bioinformaticians and computational biologists to the rescue!



Illumina Sequencing Platforms

	NextSeq System	HiSeq System	NovaSeq Series††	
	NextSeq 500*	HiSeq 4000*	NovaSeq 5000*	NovaSeq 6000*
Output Range	20–120 Gb	125–1500 Gb	167—2000 Gb	167—6000 Gb
Run Time	11–29 hr	<1–3.5 days	TBA	19—40 hr
Reads per Run	130–400 million	2.5–5 billion	1.4–6.6 billion	1.4–20 billion
Max Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Samples per Run†	1	6–12	4—16	4—48
Relative Price per Sample†	Higher Cost	Mid Cost	Lower Cost	Lower Cost
Relative Instrument Price†	Lower Cost	Mid Cost	Higher Cost	Higher Cost
Downloads	Spec Sheet	Spec Sheet	Spec Sheet	Spec Sheet

enseqlopedia.com

Illumina Sequencing Platforms

NextSeq Series

HiSeq Series

NovaSeq Series

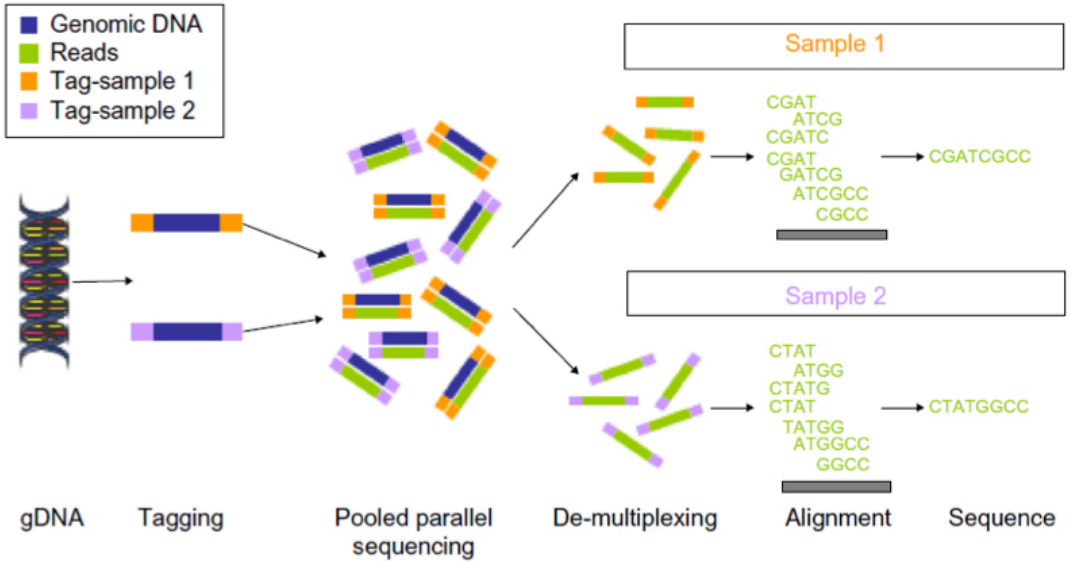
HiSeq X Series†

Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	
Exome Sequencing	●	●	●	
Targeted Gene Sequencing (amplicon, gene panel)	●	●	●	
Whole-Transcriptome Sequencing	●	●	●	
Gene Expression Profiling with mRNA-Seq	●	●	●	
miRNA & Small RNA Analysis	●	●	●	
DNA-Protein Interaction Analysis	●	●	●	
Methylation Sequencing	●	●	●	
Shotgun Metagenomics	●	●	●	

enseqlopedia.com

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.



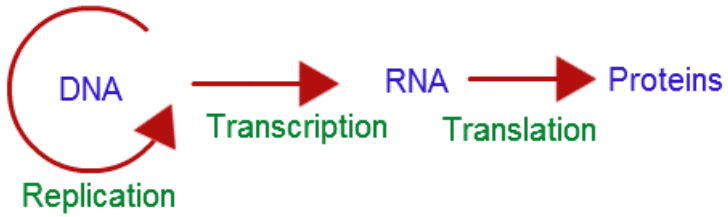
https://doi.org/10.2147/BLCTT.S51503

What are the Limations/Challenges?

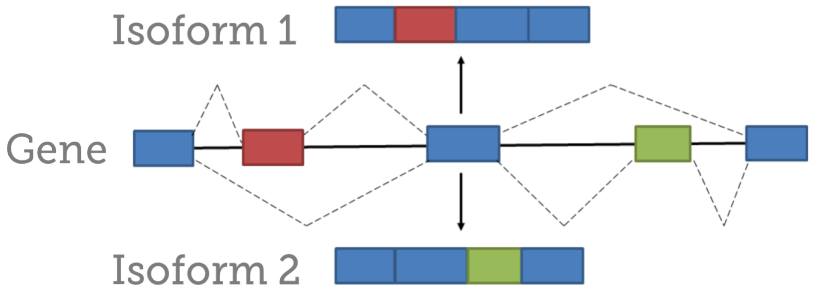
- Amplification can cause problems.
  - Clusters are made by using PCR amplification.
- Reads are short
  - difficult to align, assemble.
  - too short to span long repeat regions.
  - Difficult to detect large structural variations like inversions.

What is RNA-Seq?

- Examine the state of the transcriptome.



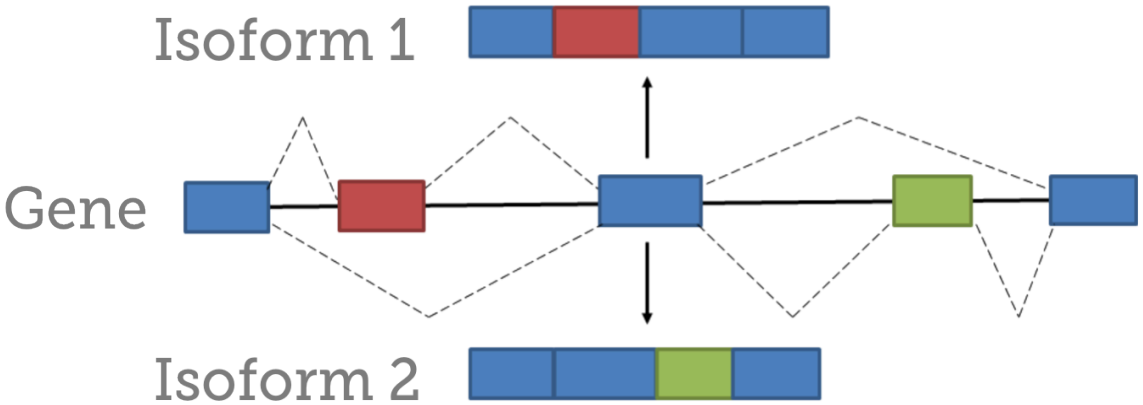
- Genes expression patterns vary in:
  - Tissue types
  - Cell types
  - Development stages
  - Disease conditions
  - Time points



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

What is RNA-Seq?

- RNA-Seq measures these expression variations
  - At gene level
  - At isoform level



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
- Profiling gene expression at single cell level

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

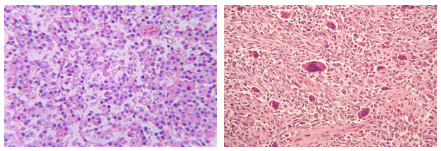
Zhong Wang, **Mark Gerstein**, and **Michael Snyder**

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

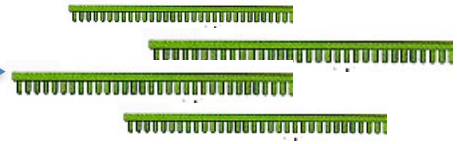
What are your questions ?

- 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/isoforms, junctions?
    - Differential gene expression?
    - Differential exon level counts?
    - Differential regulation?
  - Small RNA?
  - Identifying cell-types using gene expression?

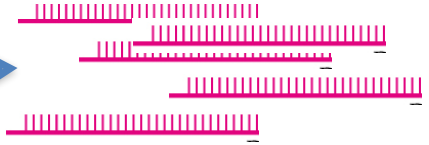
RNA-Seq… at it's Most Basic Form



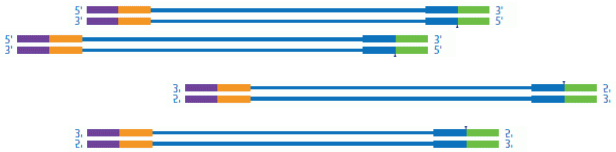
Samples from two conditions



Isolate RNA

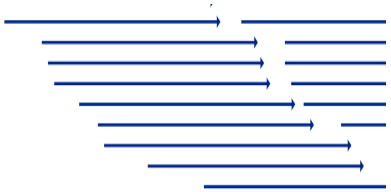


Generate cDNA



Create sequencing library by fragmenting, size selection and adding adapters

Run sequencer

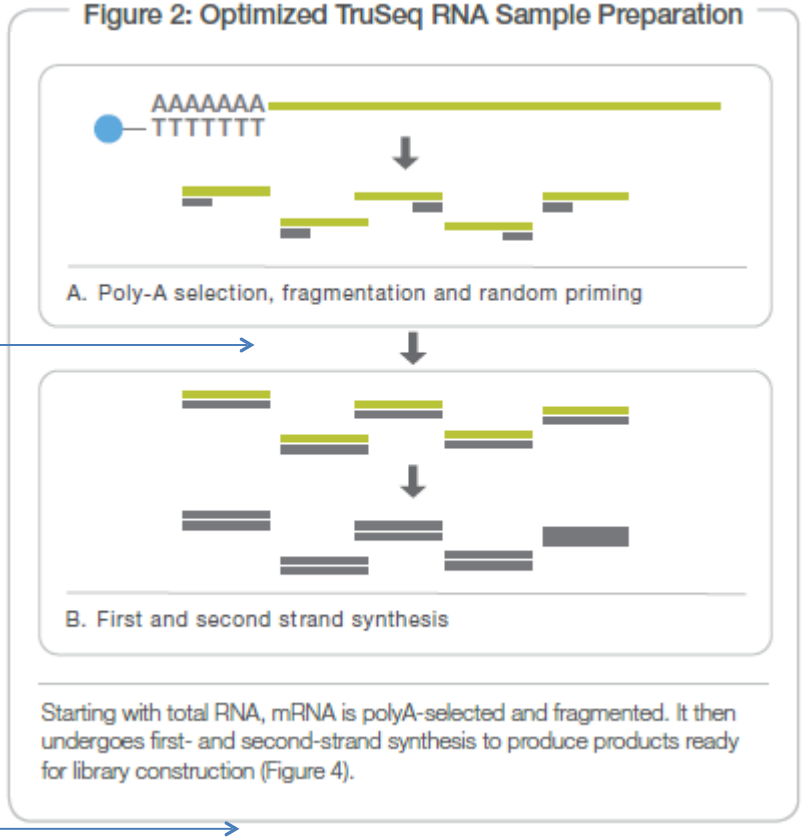


Generate short reads

Identify differentially expressed genes

RNA Illumina Tru-Seq library prep

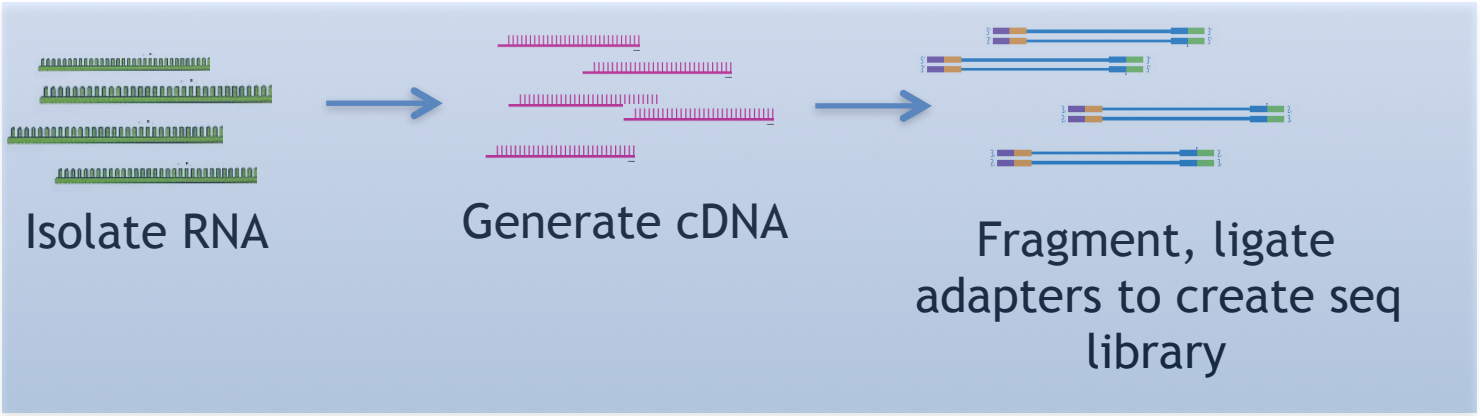
Size selection step



Adaptor ligation and standard library preparation

2 days for 8 samples

RNA-Seq Libraries… with More Details



A. rRNA Depletion

Ribominus kit

B. Normalized library

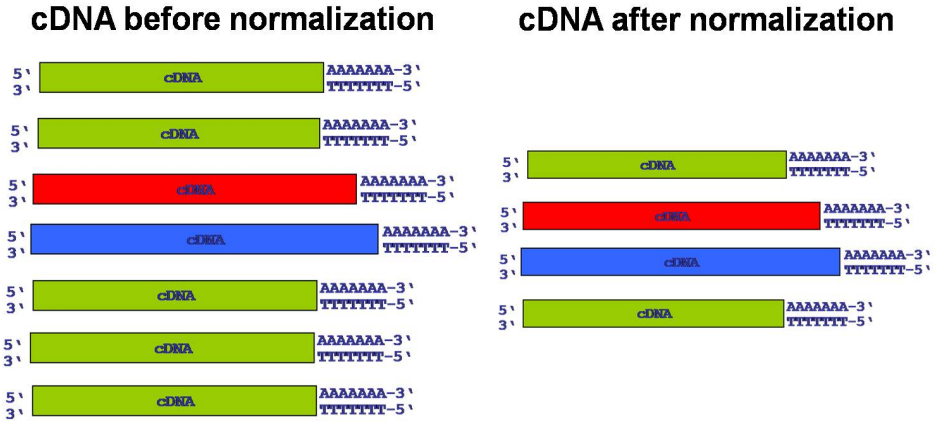
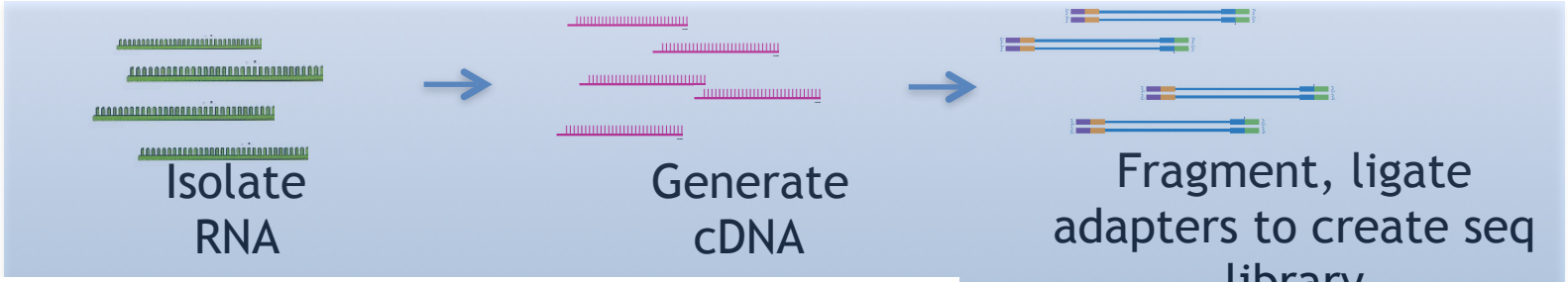


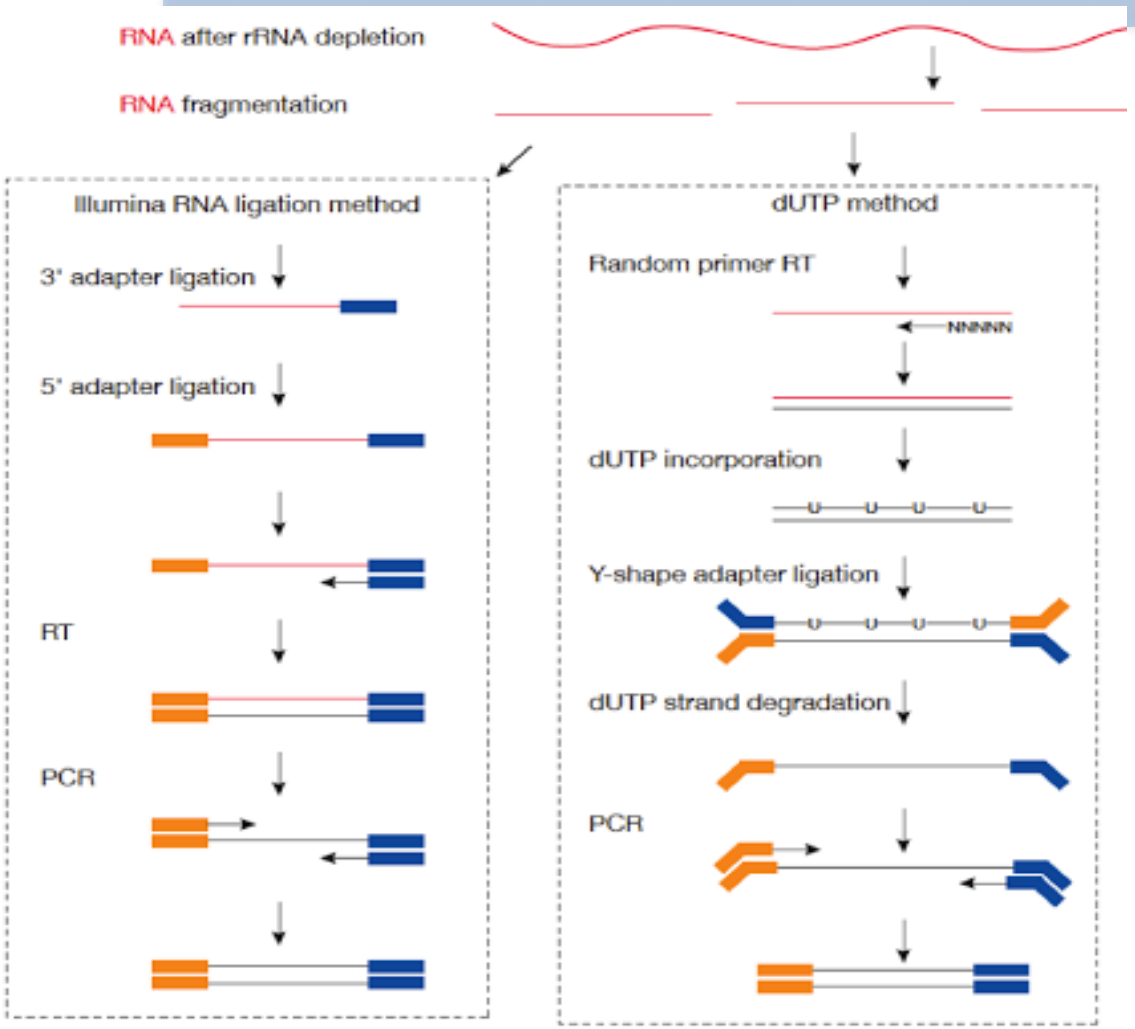
Image from :www.genxpro.info

C. Size selection

Reserved for miRNA, siRNA profiling

RNA-Seq Libraries… with More Details

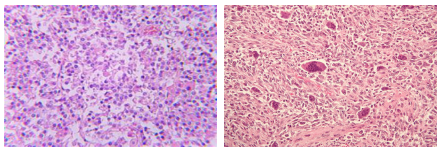




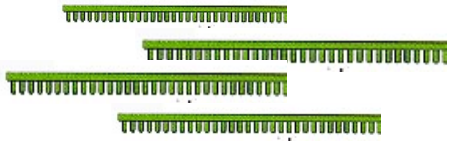
Second Strand Synthesis- Many Strand Specific Methods.

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

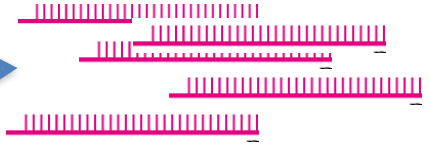
RNA-Seq… at it's Most Basic Form



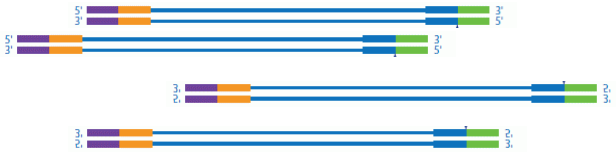
Samples from two conditions



Isolate RNA

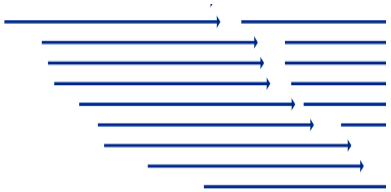


Generate cDNA



Create sequencing library by size selection and adding adaptors

Run sequencer



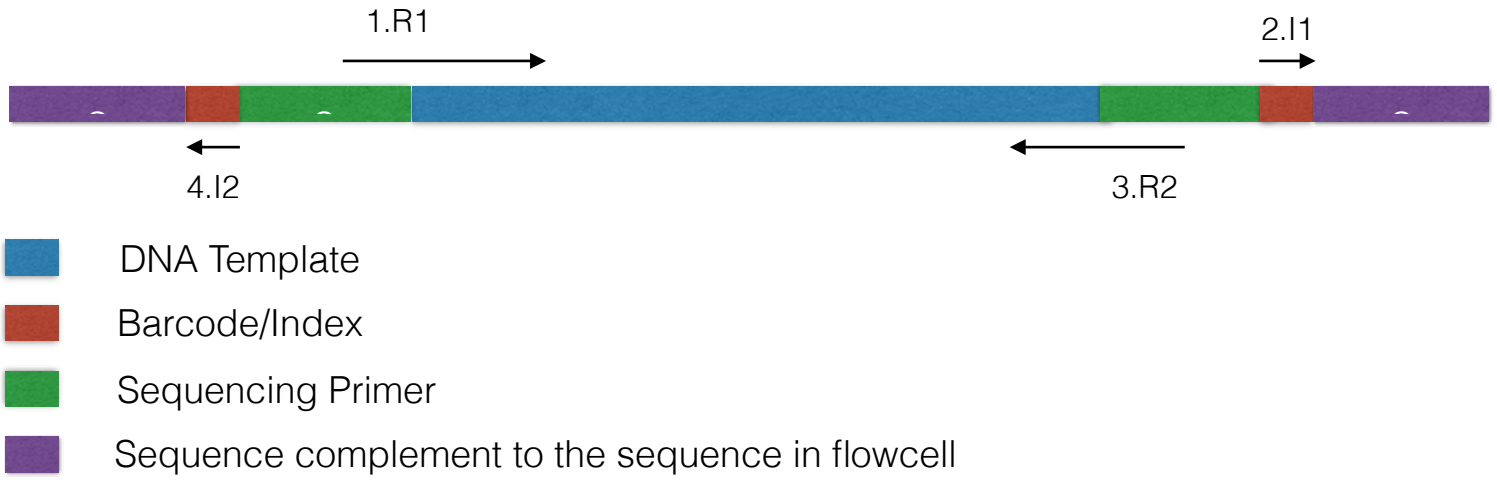
Generate short reads

Identify differentially expressed genes

What is an adapter?

Adapter:

- Allows the template DNA to attach to the flowcell/cluster
- Has primer sequences to start synthesis off of.
- Has barcodes/indexes for multiplexing



Types of Illumina Fragment Libraries

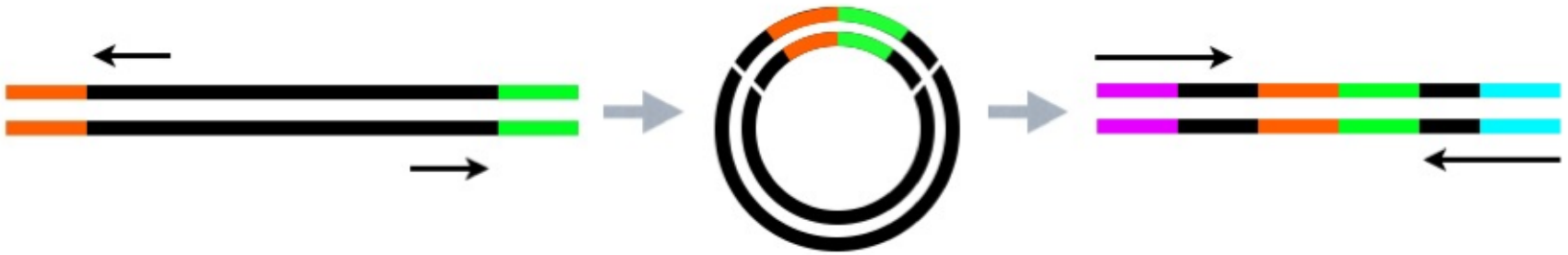
single-end

independent reads

paired-end

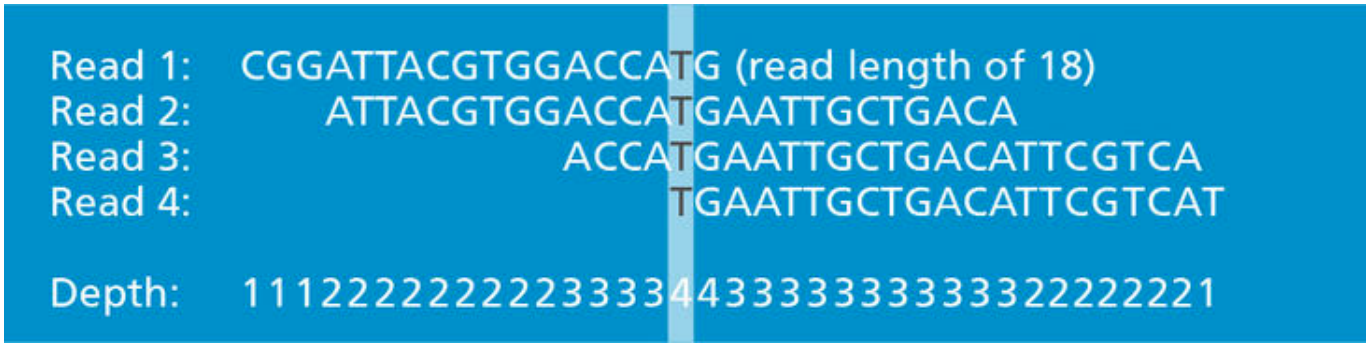
two inwardly oriented reads separated by ~200 nt

mate-paired



two outwardly oriented reads separated by ~3000 nt

What is Depth of Coverage?



Number of reads 'covering' each position in the genome/ transcriptome.

coverage = (read count * read length) / total genome size

- **Example:**
  - read count: 1000000
  - read length: 2x150bp = 300bp
  - genome size: 2MB = 2000000bp
  - **Coverage= (1000000*300)/2000000= 150x coverage**

Encode : recommendations for how much data you need **encodeproject.org/data-standards/rna-seq/long-rnas/**

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

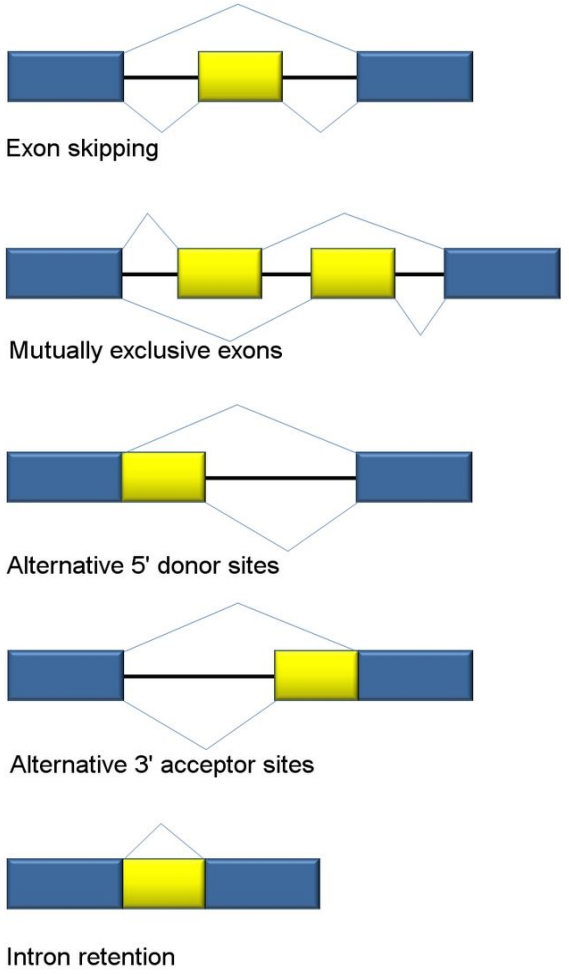
Actually important!

Actually important!

From RNA-seqlopedia

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

Illumina Fastq file

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

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%

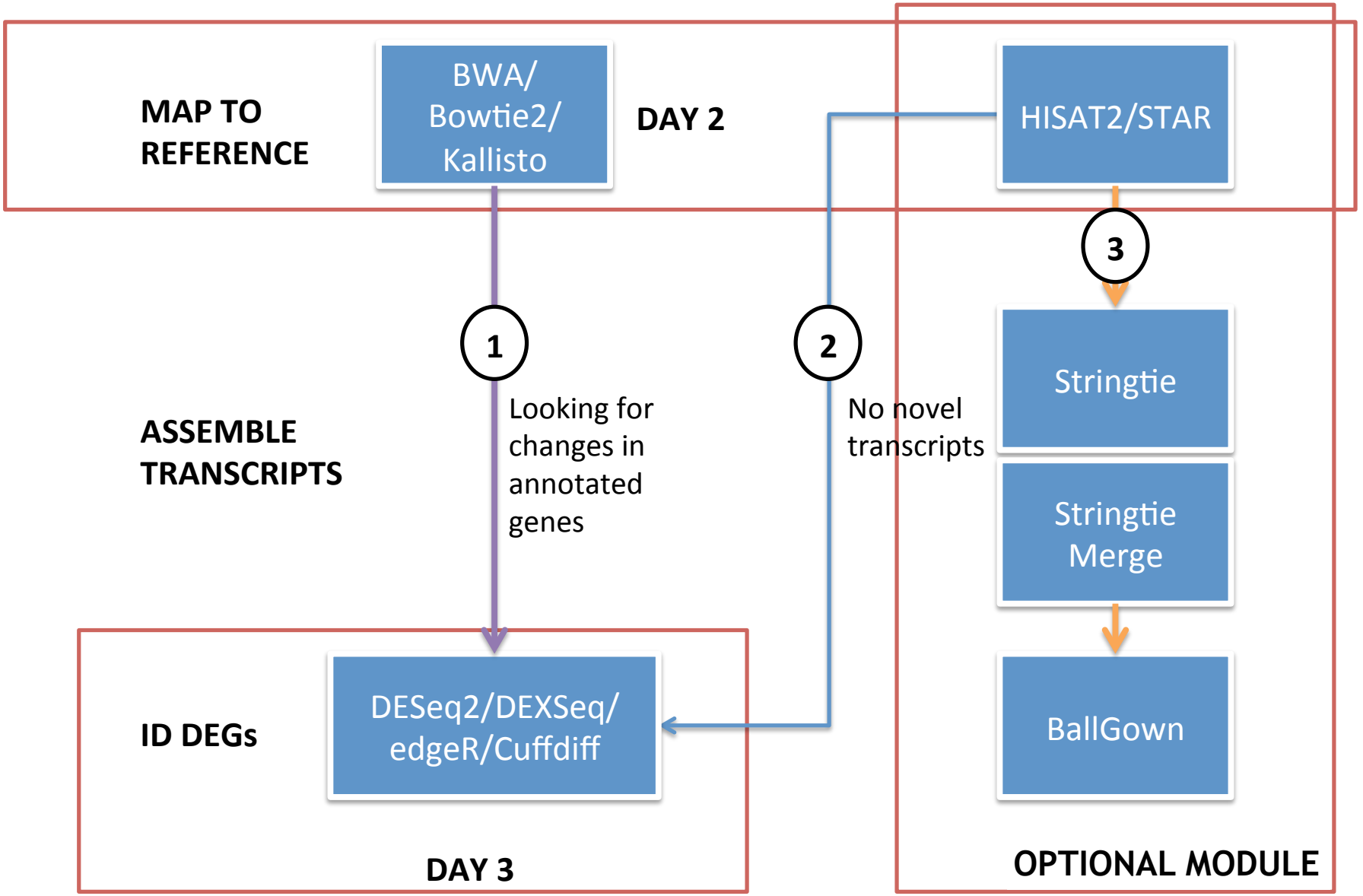
Quality scores are ASCII encoded in fastq files. Different platforms/older sequencing data can have different encoding! Illumina HiSeq 2500 produces Sanger encoded data.

Phred +33 =ASCII

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

The Big Picture

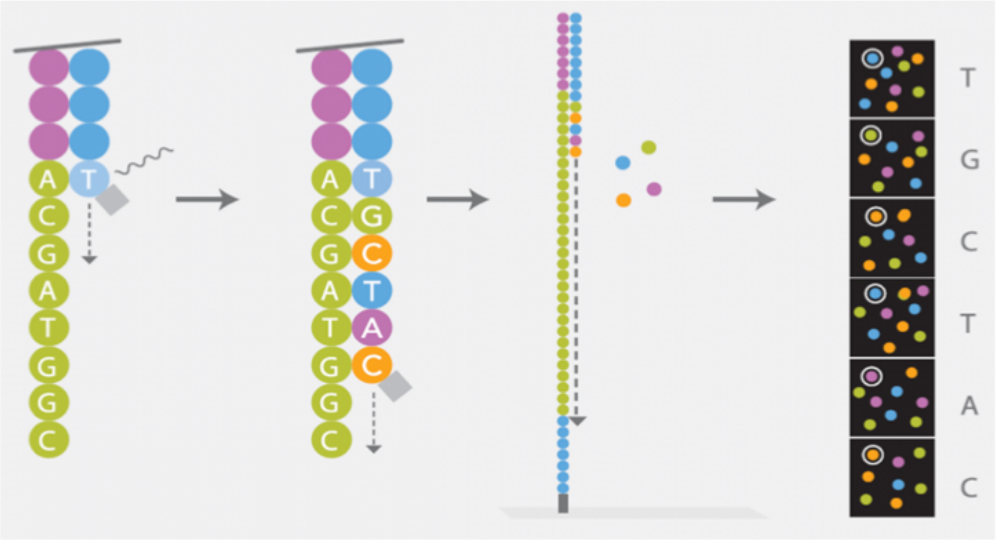


APPENDIX

How do next generation sequencers work?!

- Attach a short DNA template on a chip. !
- Flood with polymerase, fluorescent labeled nucleobases.!
- When a complementary base is generated, take a picture of the fluorescence.!
- Do this for millions/billions of templates at the same time.!

Sequencers simply observe DNA Replication!



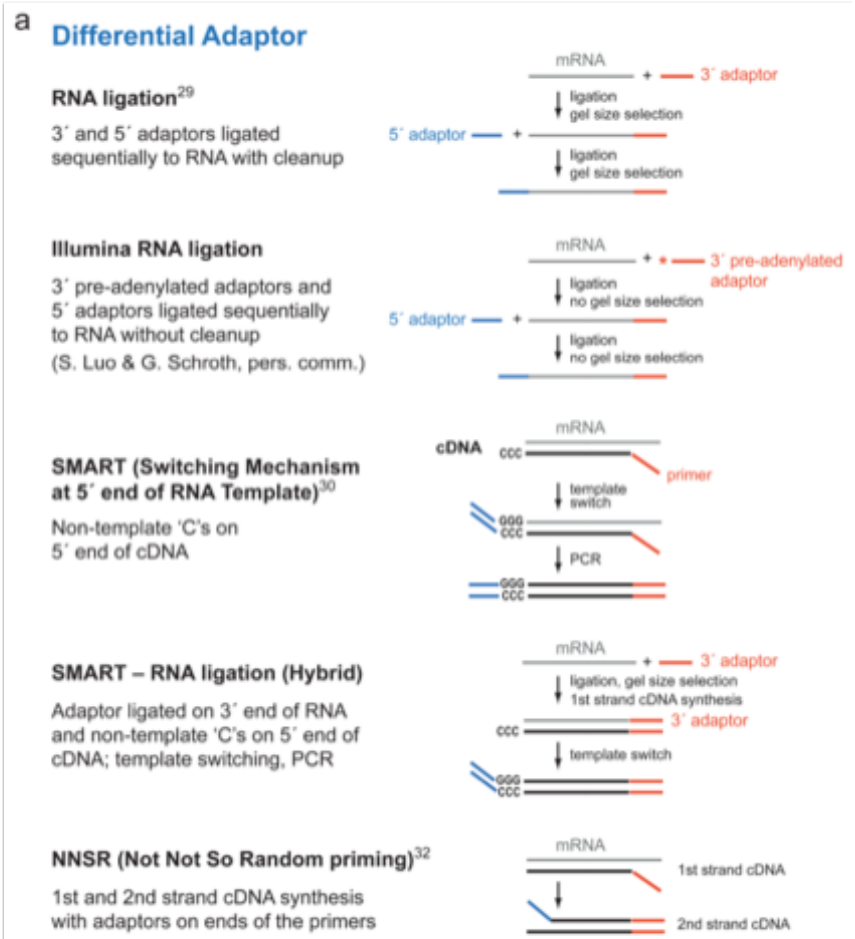
http://www.cegat.de/!

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

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

Third generation sequencing

- Next, next generation sequencing?
- Single molecule sequencing- takes care of all above mentioned issues.
- Much longer reads (1-100kb)
- Many issues- high error rate and expensive
- Two categories:
  - Sequencing by synthesis (pacbio)
    - WATCH DNA as it is sequenced in realtime
    - ZMW technology lets smallest amount of light to be detected.
  - Direct sequencing
    - Oxford nanopore
    - Hydrogen ion changes ph in well. Change in ph indicates base has been incorporated.