

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 CCBB.
 - RNA-Seq
 - Genome Assembly
 - Exome data analysis
 - Benchmarking of tools
- Training
 - Grad students, post-docs.
 - Undergraduate- FRI



CCBB
Center for Computational Biology
and Bioinformatics



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

Setting General Expectations

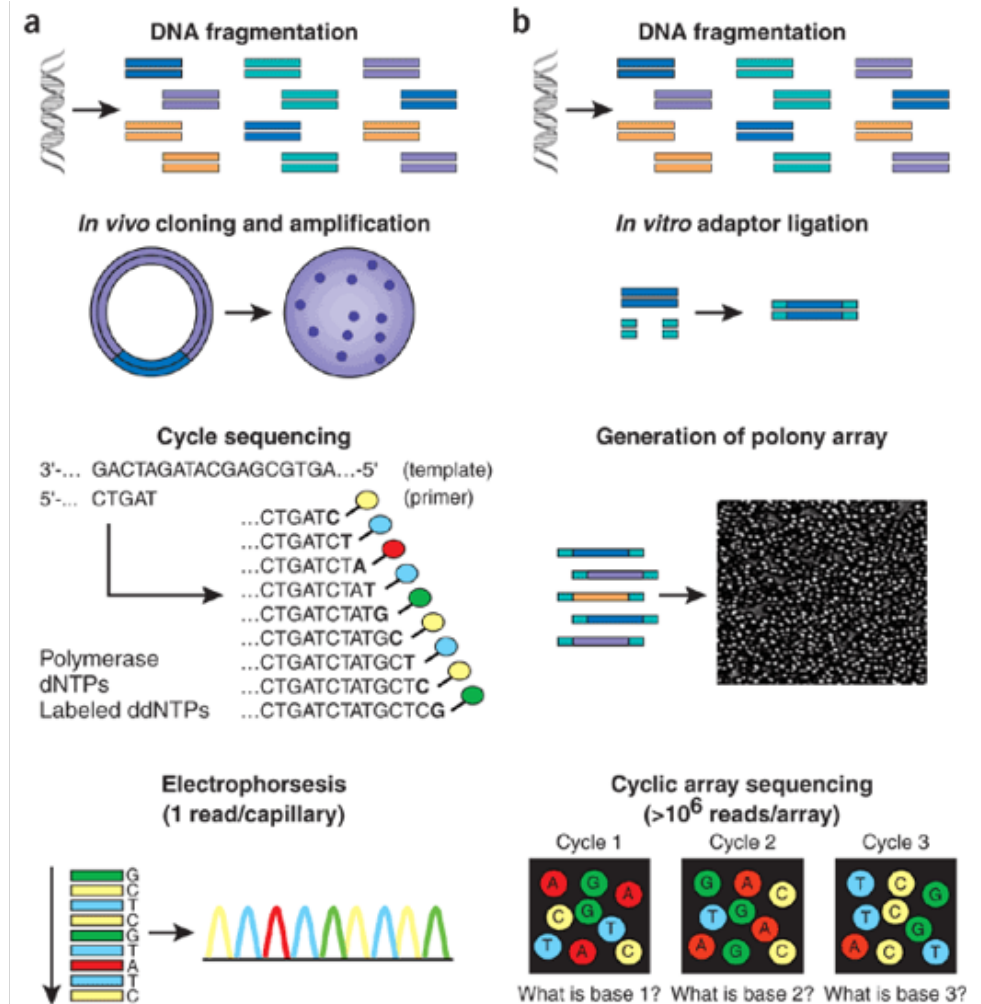
- Lots of background and basics to provide comfort with terminology and key concepts.
- Exposure to commands and typically used analysis tools using an example RNA-Seq dataset.
 - No one ‘best’ or ‘standard’ tool.
- A starting point for you to design your RNA-Seq study or analyze your dataset.

Resources

- Biolteam 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>
- Byte Club: Meets Third Wednesday of every month
- <https://wikis.utexas.edu/display/bioiteam/Byte+Club>
- CCBB Bioinformatics consultants

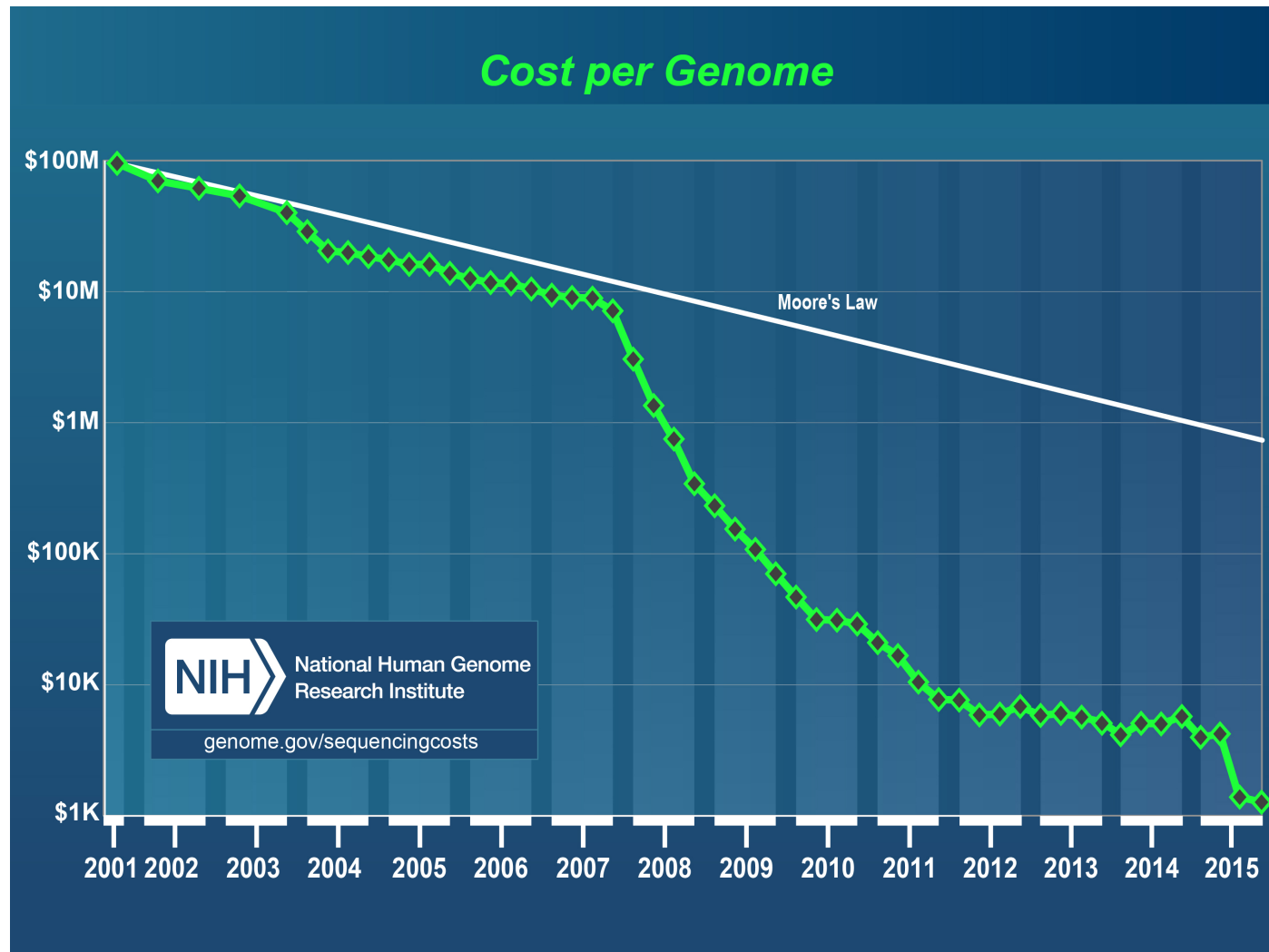
What is Next Generation(or) Second Generation Sequencing?

- 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.



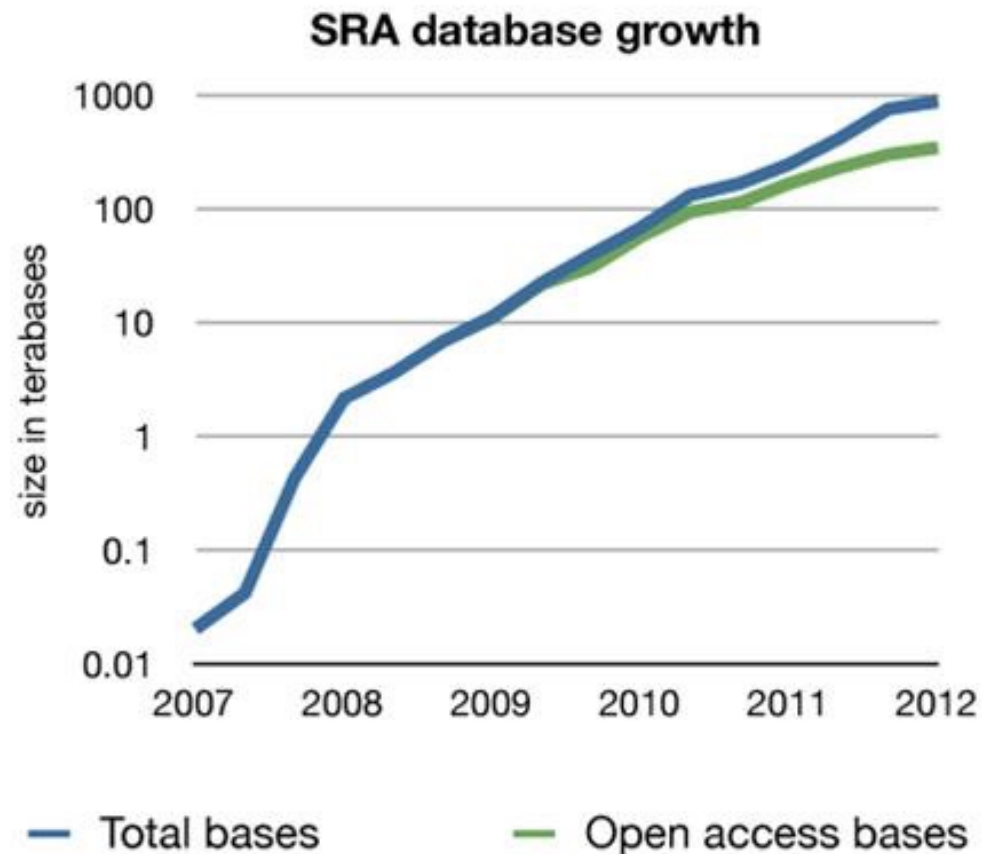
So, what's so great about second generation sequencing?

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



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

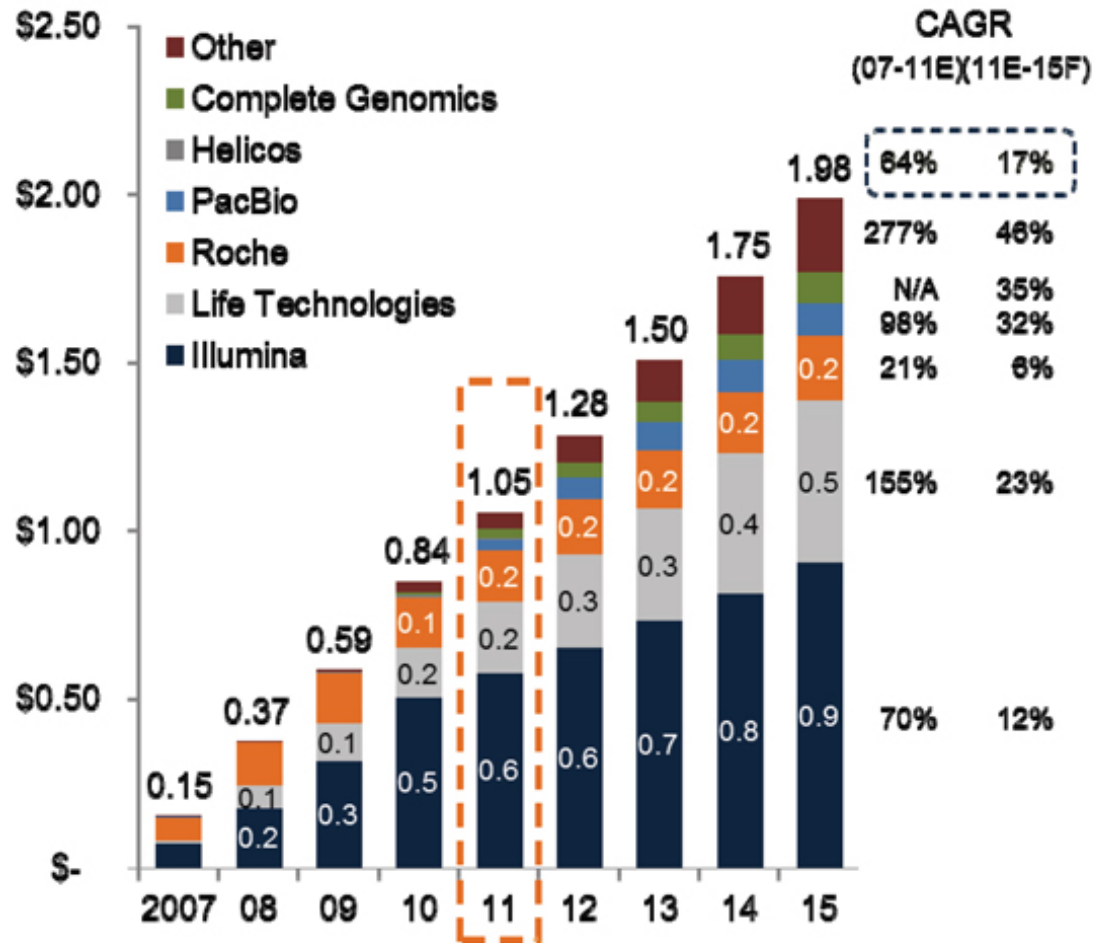
- Data deluge
- Bioinformaticians and computational biologists to the rescue!



Who are the players?

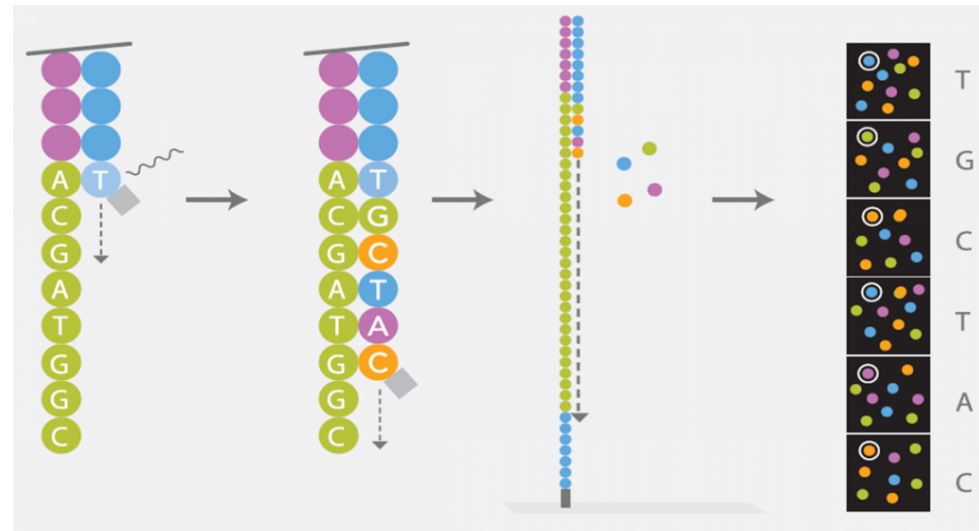
WWNGS market by competitor (2007-15F)*

Billions of dollars



How does the sequencer work?

- Library prep
- Cluster generation/
amplification
- Sequencing by synthesis
- Done in parallel for billions
clusters at once.
- Let's watch the official Illumina
video.



<http://www.cebcat.de/>

Different Types of Illumina Sequencers

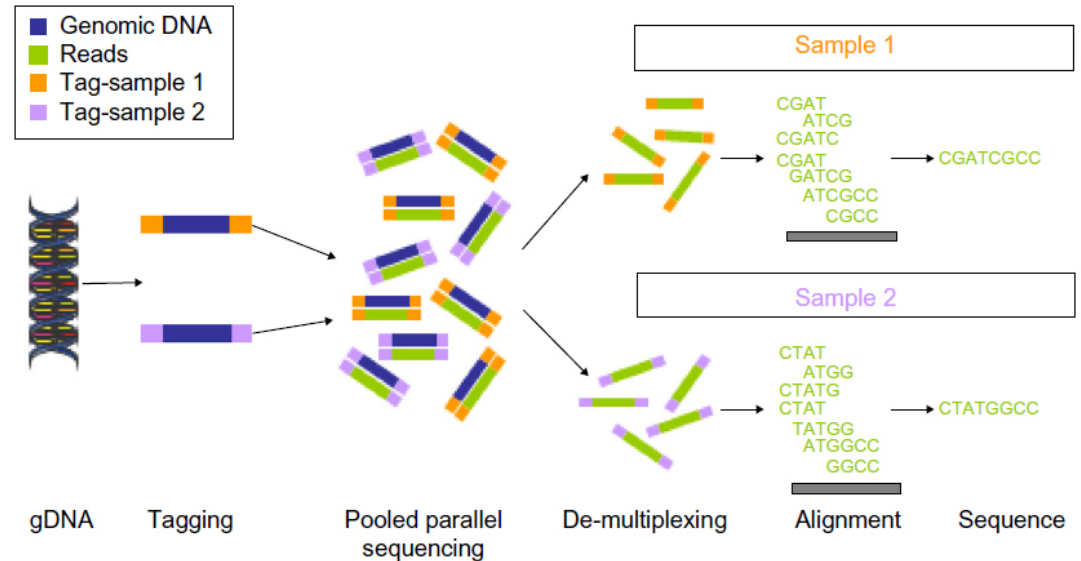


Illumina Specifications Table

	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

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.



What are the Limitations/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.



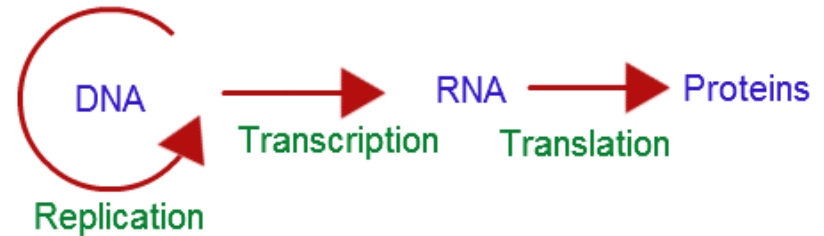
Third generation sequencing

- 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.



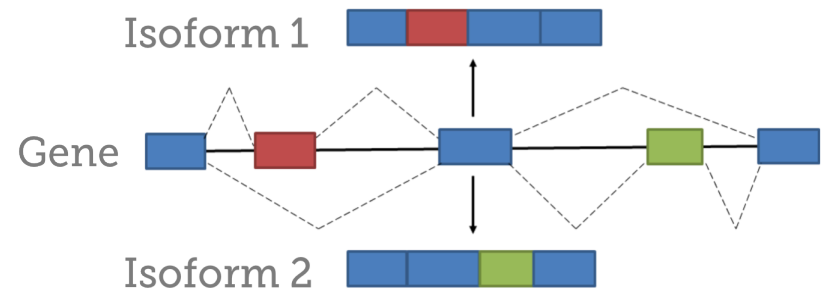
What is RNA-Seq?

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

Other Uses of RNA-Seq

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

Advantages of RNA-Seq

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

RNA-Seq: a revolutionary tool for transcriptomics

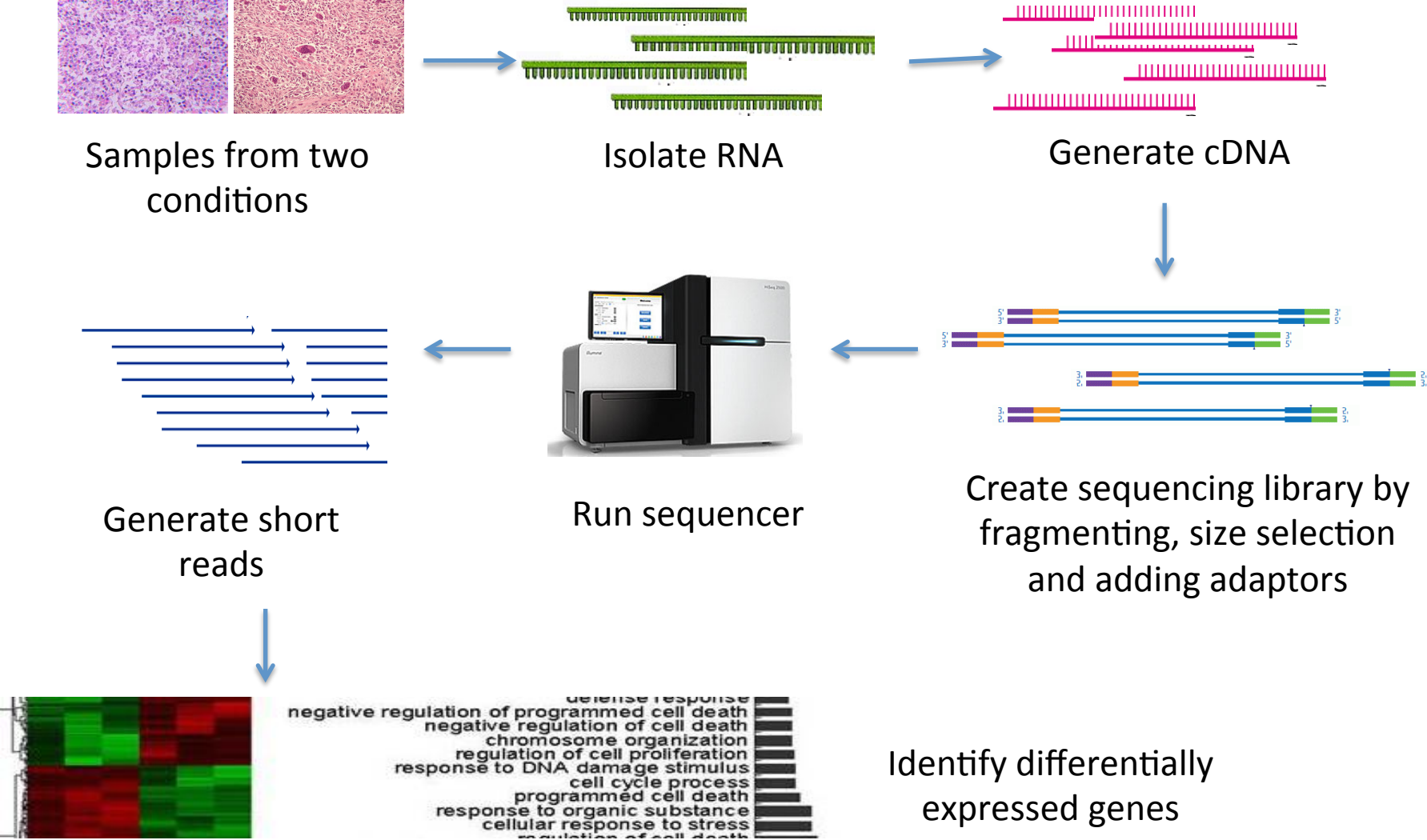
Zhong Wang, Mark Gerstein, and Michael Snyder

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

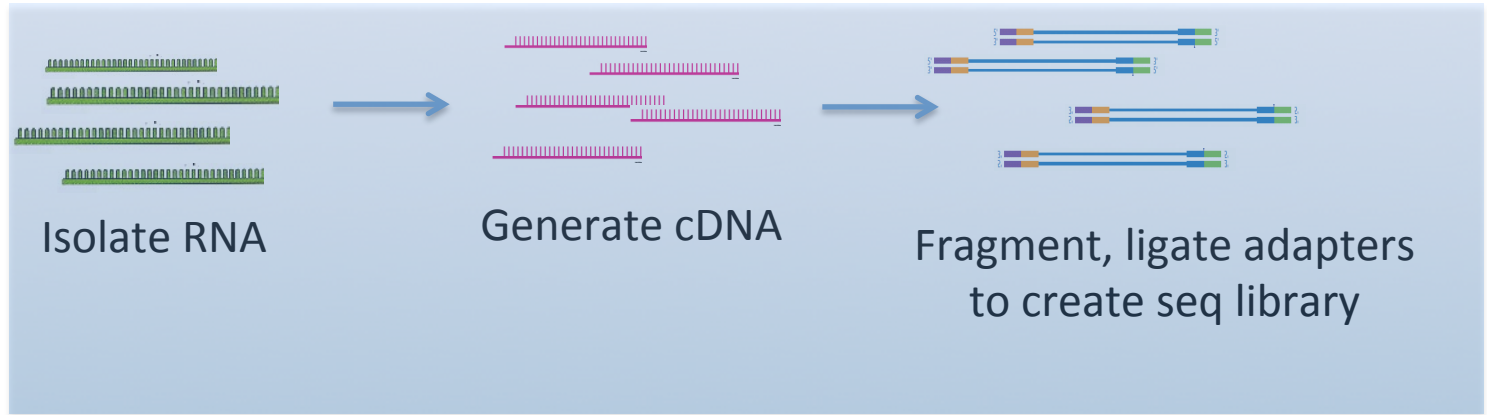
What are your questions ?

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

RNA-Seq... at it's Most Basic Form



RNA-Seq Libraries... with More Details



B. Normalized library

cDNA before normalization



cDNA after normalization

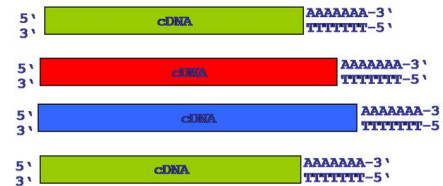


Image from :www.genxpro.info

A. rRNA Depletion

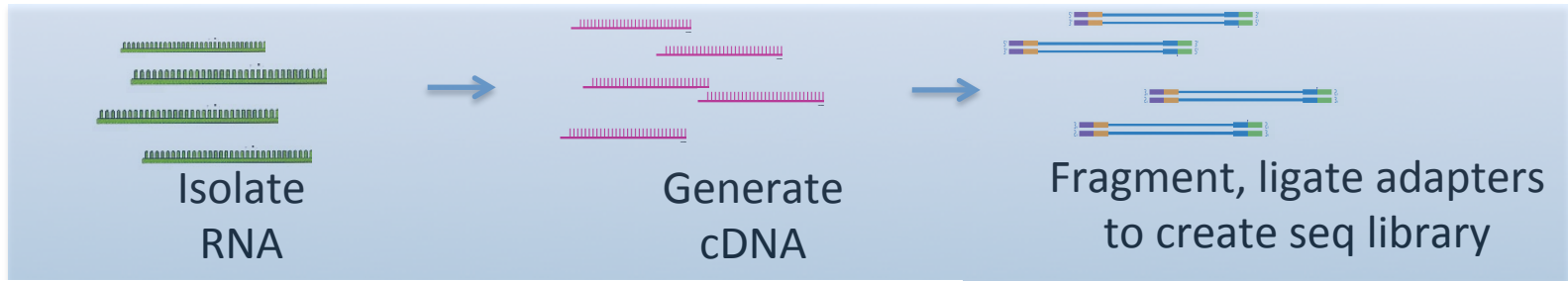


Ribominus kit

C. Size selection

Reserved for
miRNA,
siRNA
profiling

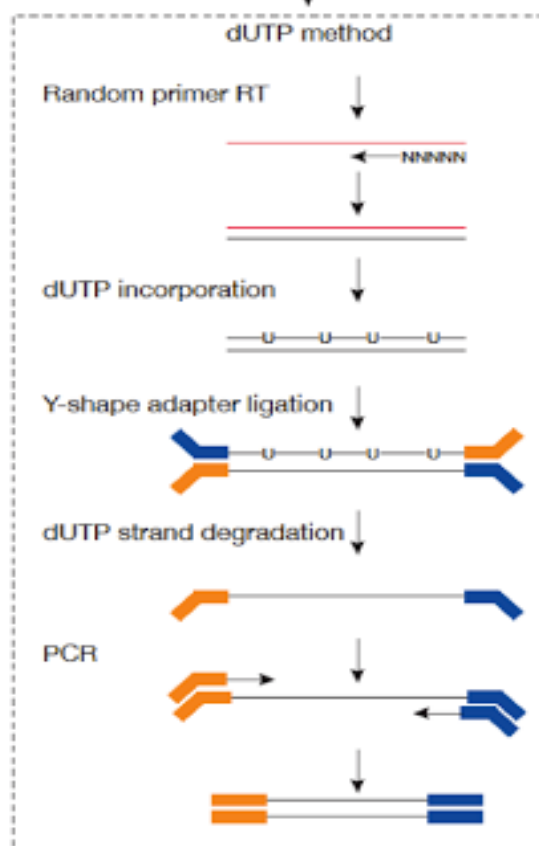
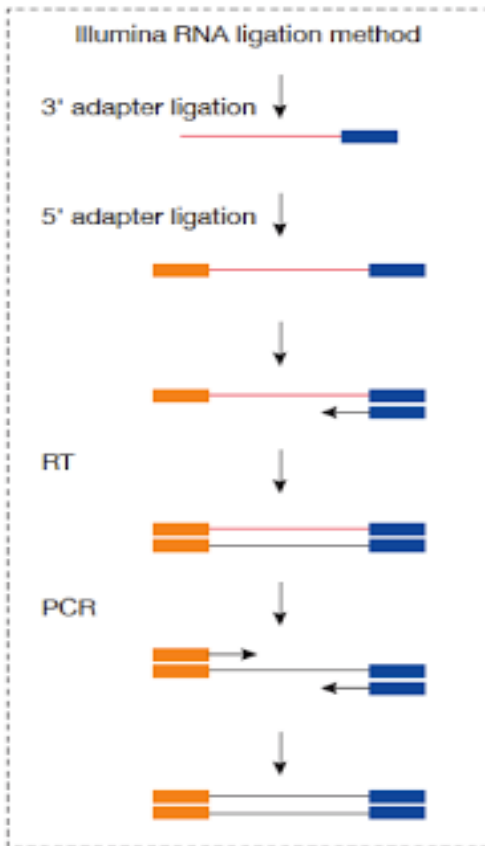
RNA-Seq Libraries... with More Details



RNA after rRNA depletion

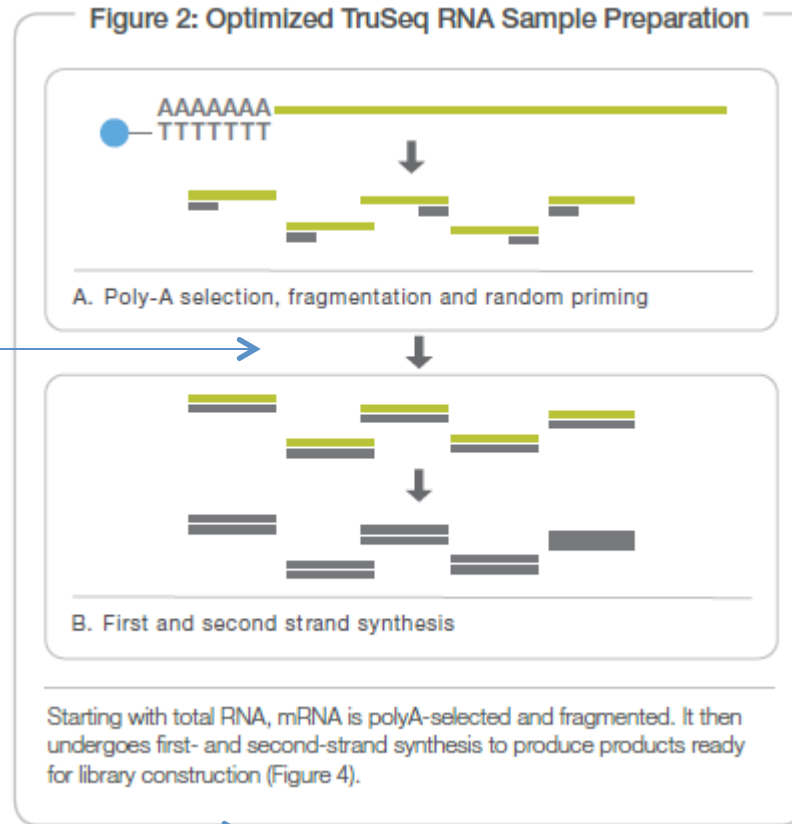
RNA fragmentation

**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

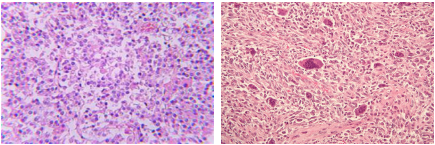


Size selection step

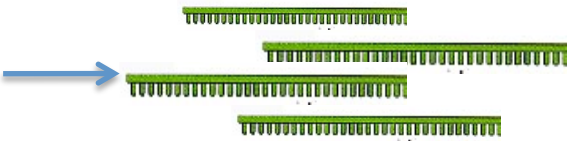
Adaptor ligation and standard library preparation

2 days for 8 samples

RNA-Seq... at it's Most Basic Form



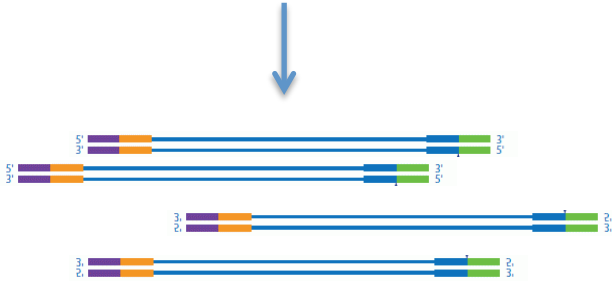
Samples from two conditions



Isolate RNA



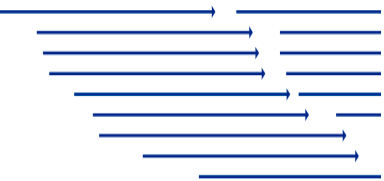
Generate cDNA



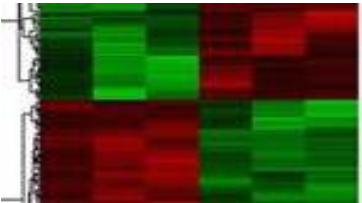
Create sequencing library by size selection and adding adaptors



Run sequencer



Generate short reads



- defense response
- negative regulation of programmed cell death
- negative regulation of cell death
- chromosome organization
- regulation of cell proliferation
- response to DNA damage stimulus
- cell cycle process
- programmed cell death
- response to organic substance
- cellular response to stress
- regulation of cell death

Identify differentially expressed genes

Types of Illumina Fragment Libraries

single-end



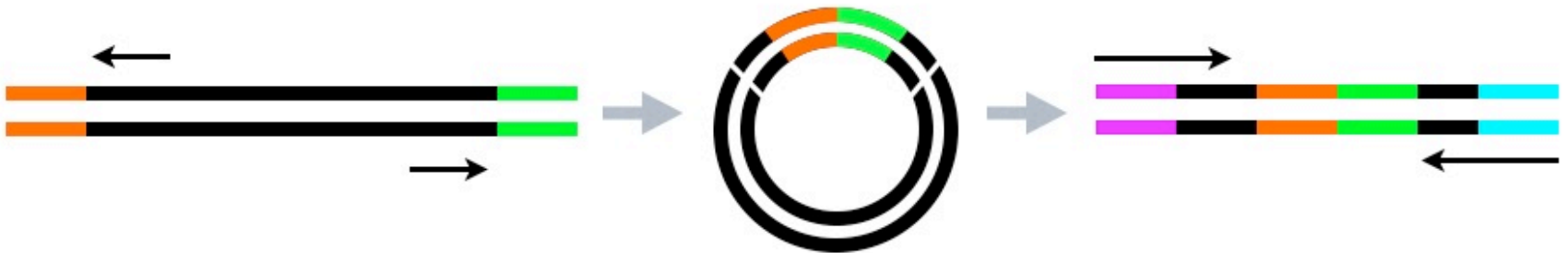
independent reads

paired-end



two inwardly oriented reads separated by ~200 nt

mate-paired



two outwardly oriented reads separated by ~3000 nt

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 accurate 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 transcripts	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!

Comparing Stranded RNA-Seq Library Protocols

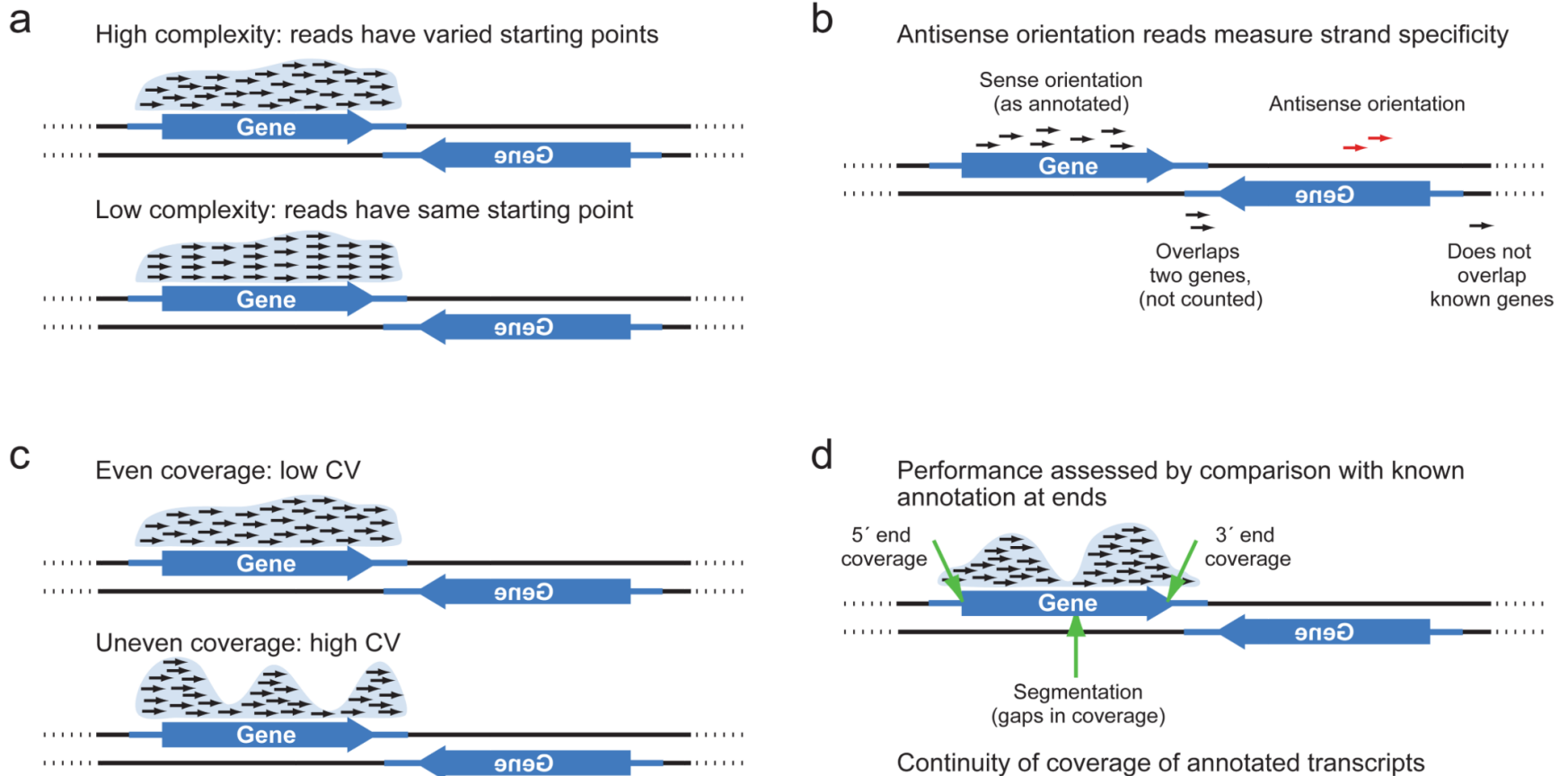
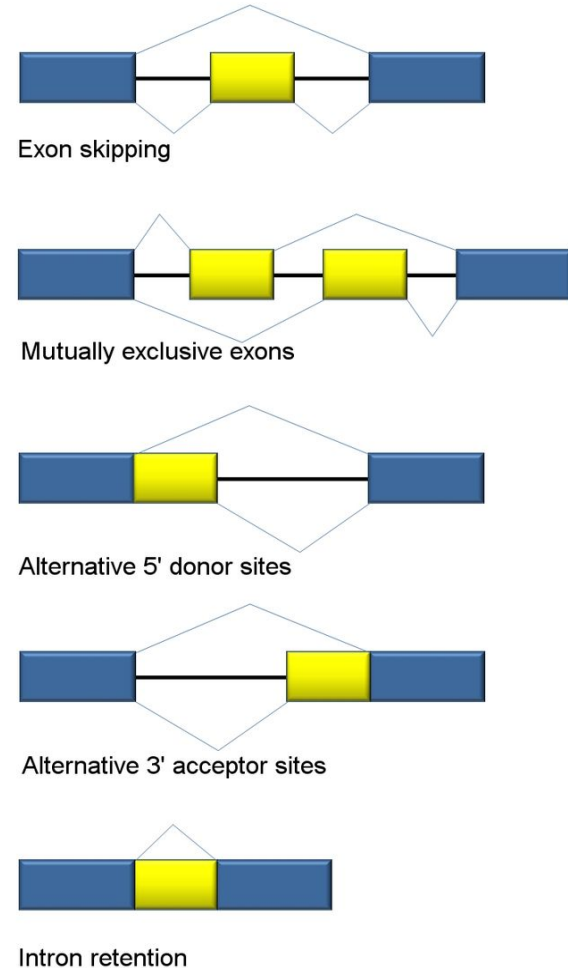


Figure 2. Key criteria for evaluation of strand-specific RNAseq libraries

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..

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

<http://www.asciitable.com/>

Quality character	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I																	
ASCII Value	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
Base Quality (Q)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50							

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

(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

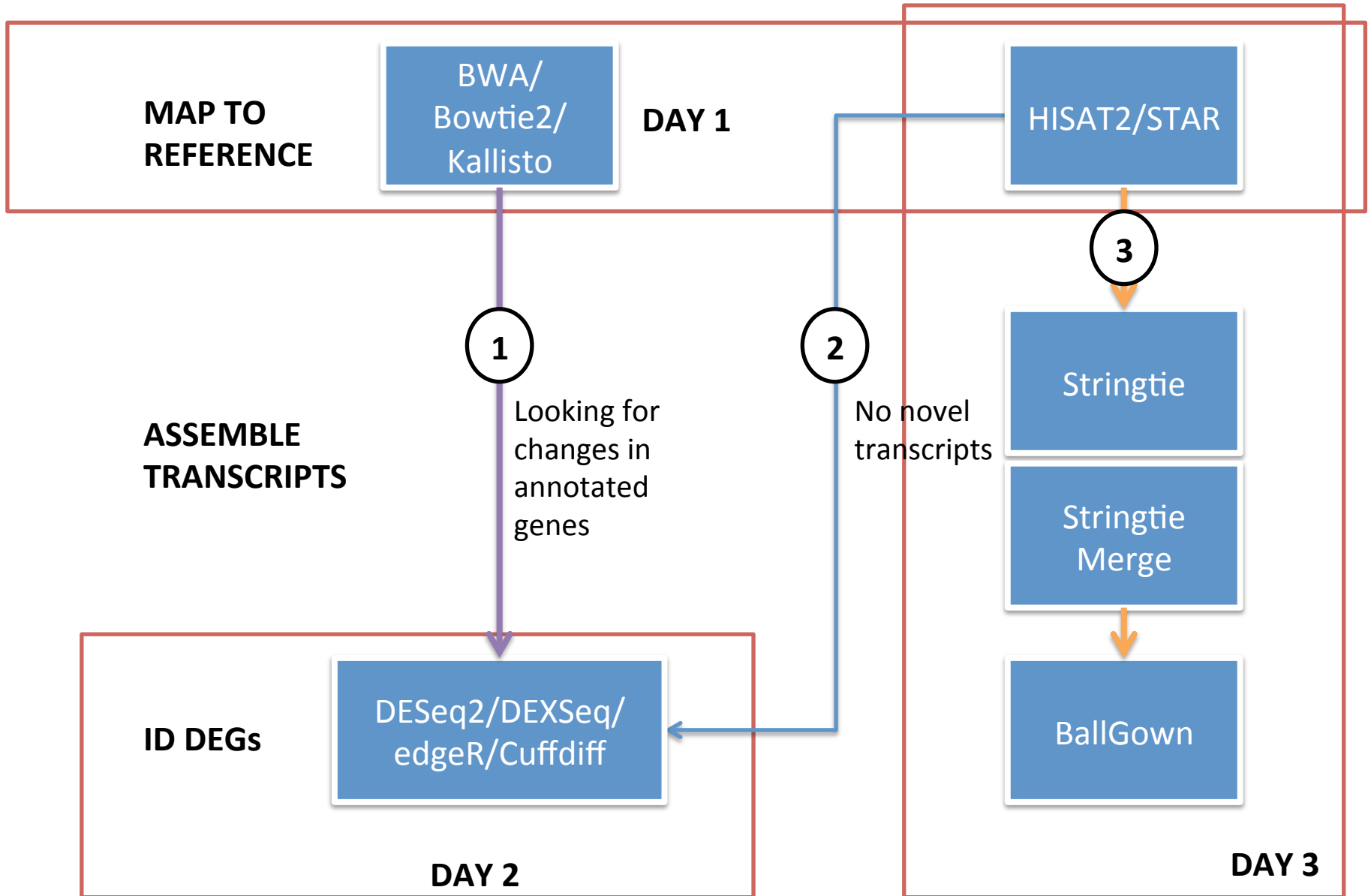


Table 1 | Selected list of RNA-seq analysis programs

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		
Spliced aligners	Burrows-Wheeler transform methods	Bowtie ⁴³ BWA ⁴⁴	Incorporates quality scores	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
	Exon-first methods	MapSplice ⁵² SpliceMap ⁵⁰ TopHat ⁵¹	Works with multiple unspliced aligners Uses Bowtie alignments		
	Seed-extend methods	GSNAP ⁵³ QPALMA ⁵⁴	Can use SNP databases 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 Cufflinks ²⁹	Reports all isoforms Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet ⁶¹ TransABySS ⁵⁶	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
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		
	Isoform quantification	Normalization by expected uniquely mappable area (NEUMA) ⁸² Cufflinks ²⁹ MISO ³³ RNA-seq by expectation maximization (RSEM) ⁶⁹	Quantifies using unique reads Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
Differential expression		Cuffdiff ²⁹	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq ⁷⁹ EdgeR ⁷⁷	Uses a normal distribution		
		Differential Expression analysis of count data (DESeq) ⁷⁸			
		Myrna ⁷⁵	Cloud-based permutation method		

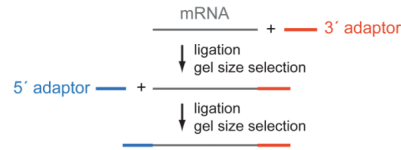
Figure:
Garber et al, Nature Methods, 2011

Appendix

a Differential Adaptor

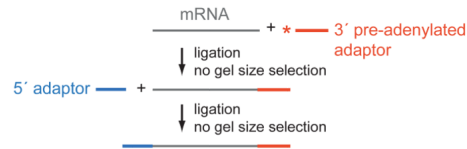
RNA ligation²⁹

3' and 5' adaptors ligated sequentially to RNA with cleanup



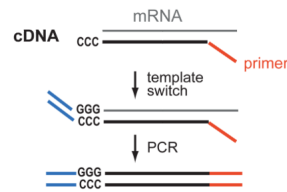
Illumina RNA ligation

3' pre-adenylated adaptors and 5' adaptors ligated sequentially to RNA without cleanup (S. Luo & G. Schroth, pers. comm.)



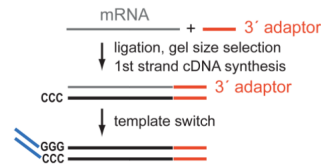
SMART (Switching Mechanism at 5' end of RNA Template)³⁰

Non-template 'C's on 5' end of cDNA



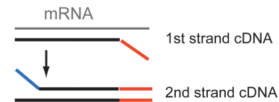
SMART – RNA ligation (Hybrid)

Adaptor ligated on 3' end of RNA and non-template 'C's on 5' end of cDNA; template switching, PCR



NNSR (Not Not So Random priming)³²

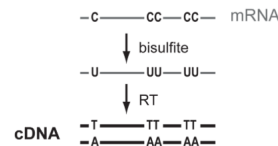
1st and 2nd strand cDNA synthesis with adaptors on ends of the primers



b Differential Marking

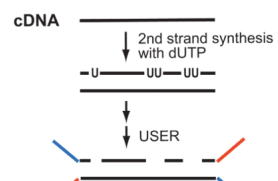
Bisulfite^{15,16}

Convert 'C's to 'U's in RNA



dUTP 2nd strand¹³

2nd strand synthesis with dUTP, remove 'U's after adaptor ligation and size selection



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.