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

CCBB
Center for Compuational Biology
and Bioinformatics

FRI

BIG DATA IN BIOLOGY

- Training
  - Grad students, post-docs.
  - Undergraduate- FRI

Setting General Expectations

- Lots of background and basics to provide comfort with terminology and key concepts.

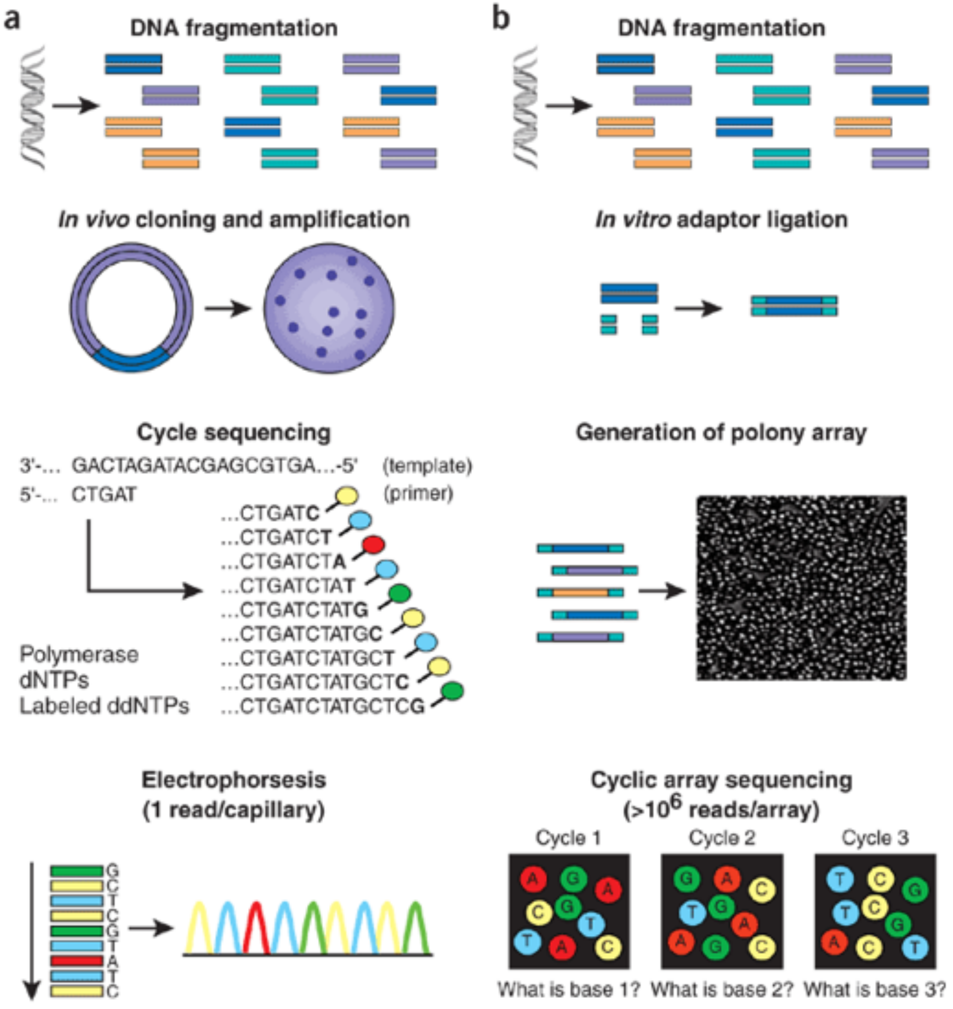
- Exposure to commands and typically used analysis tools using an example RNA-Seq dataset.
  - No one ‘best’ or ‘standard’ tool.

- A starting point for you to design your RNA-Seq study or analyze your dataset.

- Slides for lectures, wiki for tutorials.

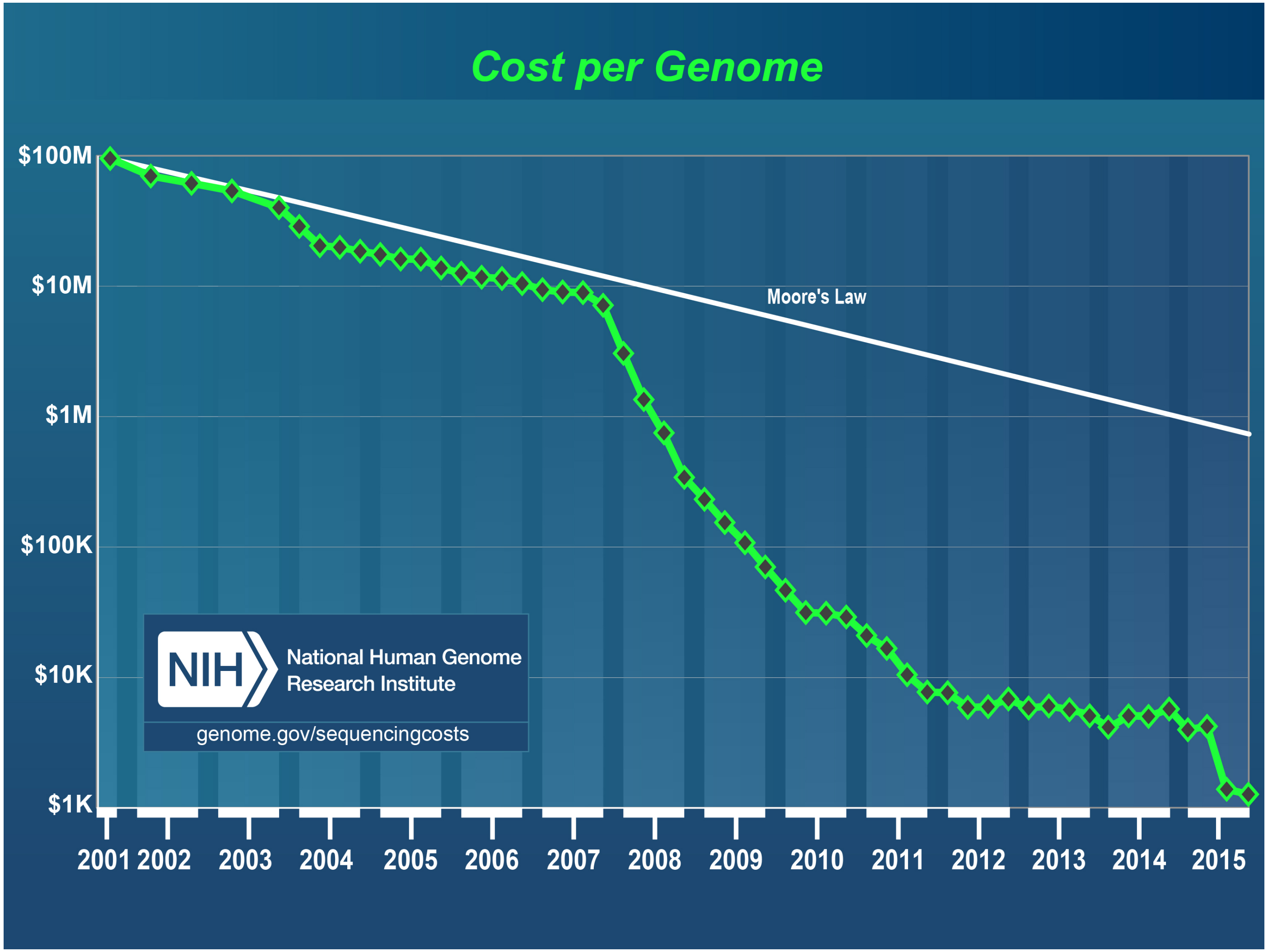
What is Next Generation(or) Second Generation Sequencing?

- 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.



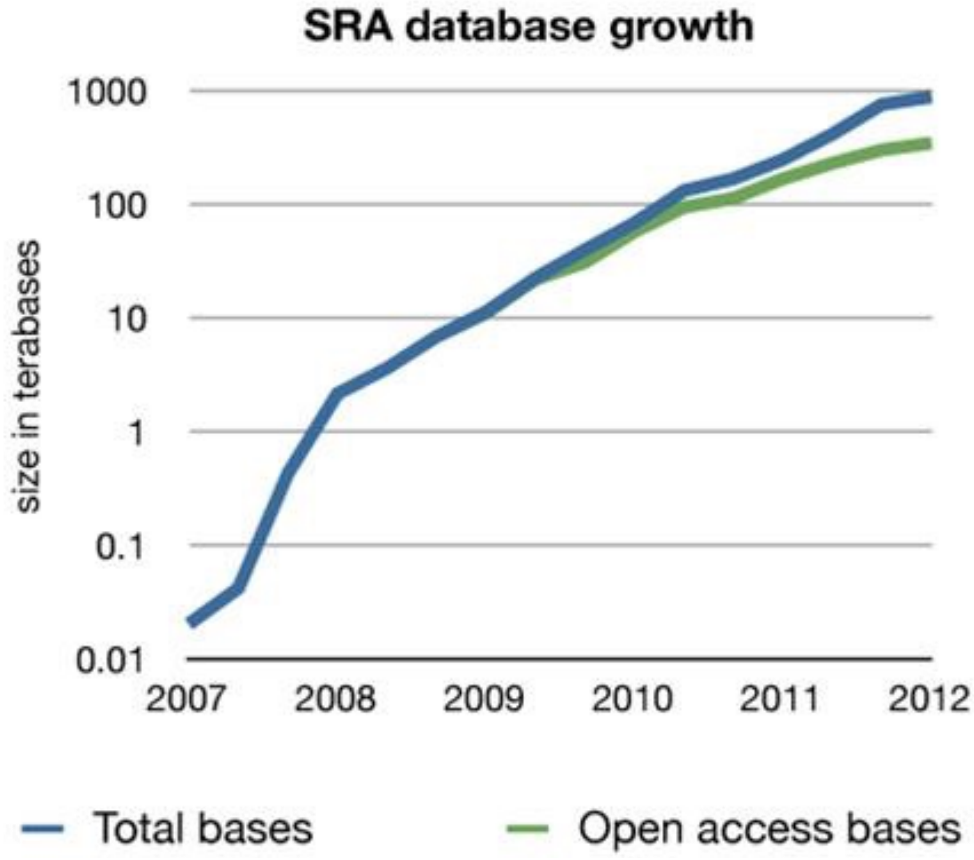
So, what's so great about second generation sequencing?

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

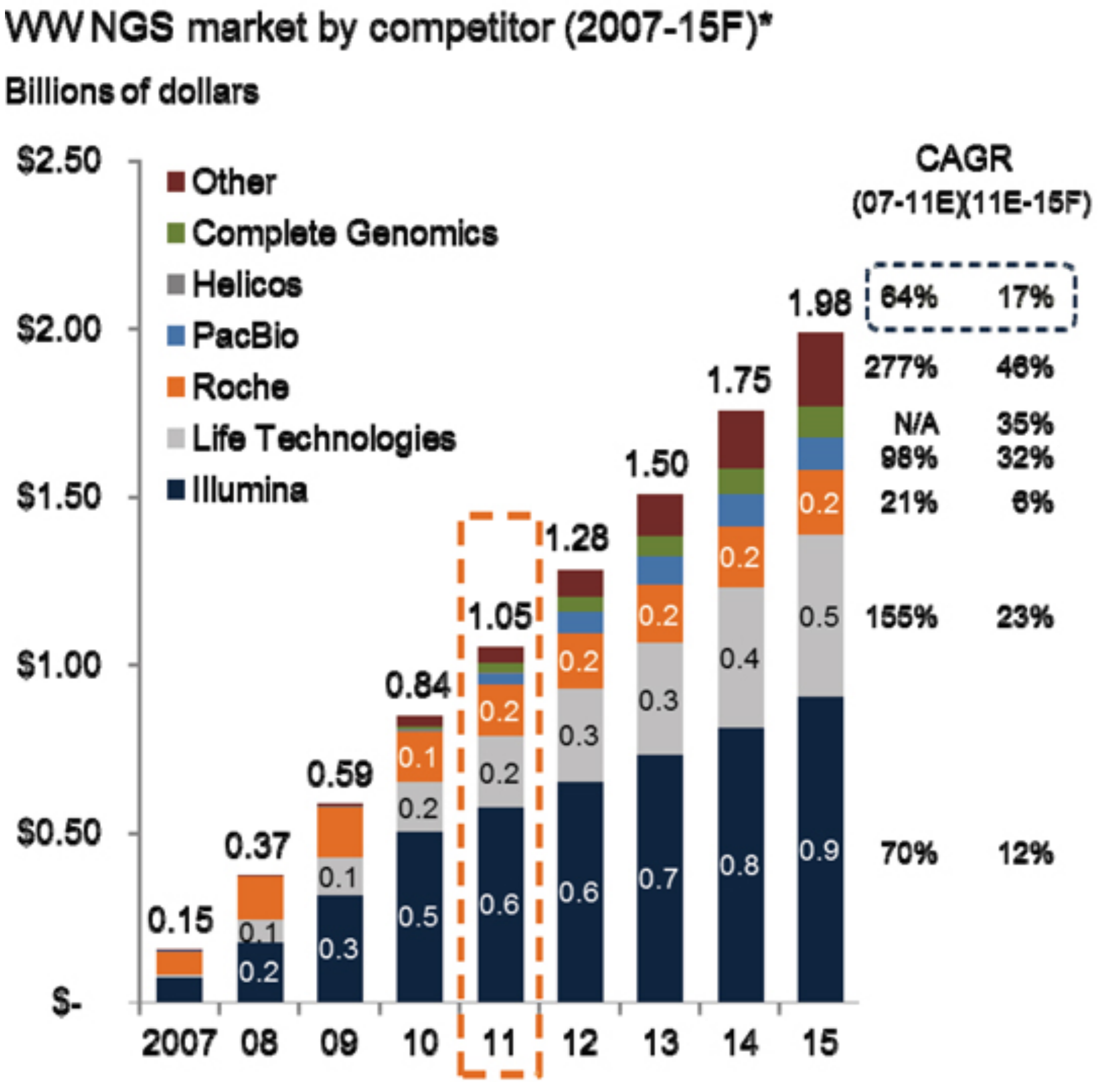


So, what's NOT so great about second generation sequencing?

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

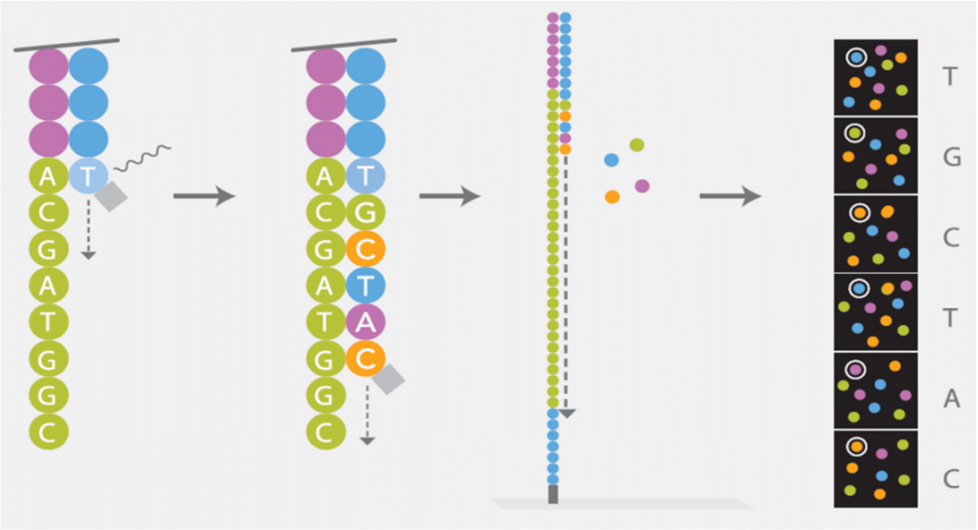


Who are the players?



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/

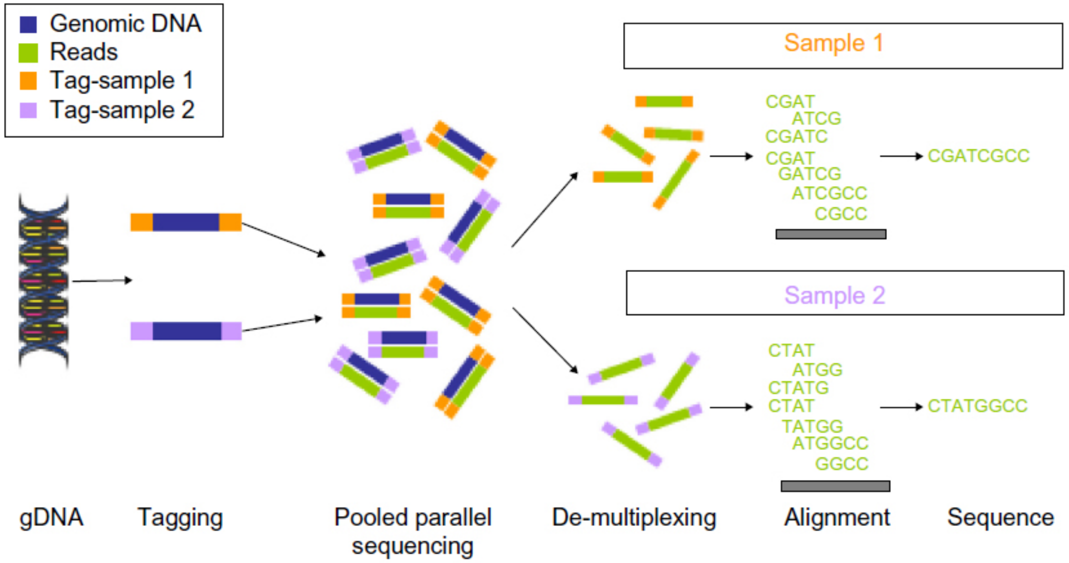
- Let’s watch the official Illumina video.

Multiplexing

- Sample specific Indexes/Barcodes are attached to the DNA template.

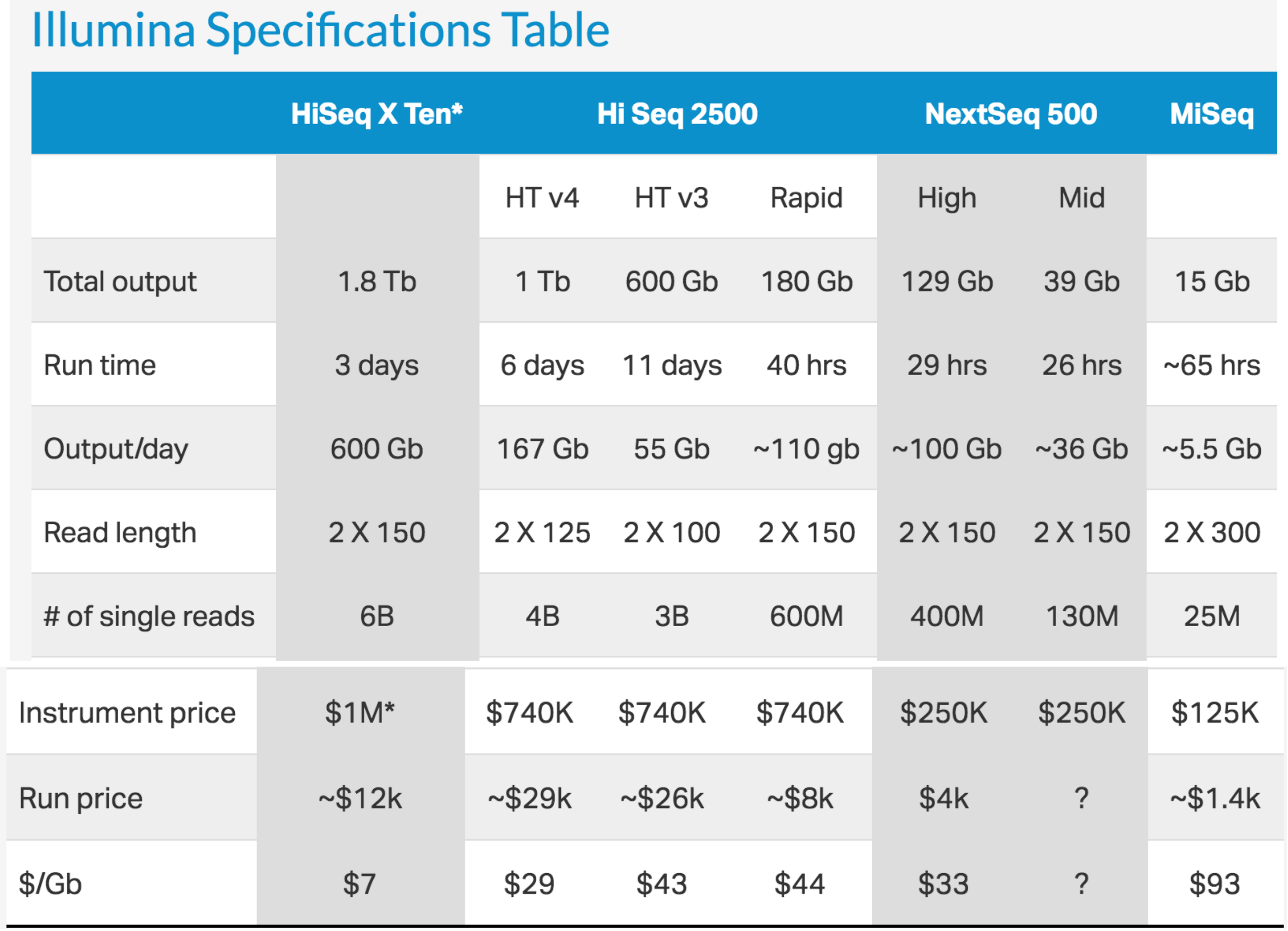
- 6-8bp indexes/barcodes

- Data off the sequencer must first be demultiplexed to identify which reads belong to which sample.



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

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-platforms/illumina

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 structura variations like inversions.

Third generation sequencing

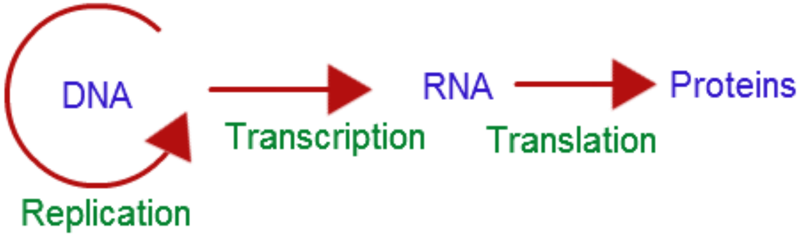
- Next, next generation sequencing?
- Single molecule sequencing- takes care of all above mentioned issues.
- Much longer reads (1-100kb)
- Many issues- high error rate and expensive
- Two categories:
  - Sequencing by synthesis (pacbio)
    - WATCH DNA as it is sequenced in realtime
    - ZMW technology 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.

PACBIO®

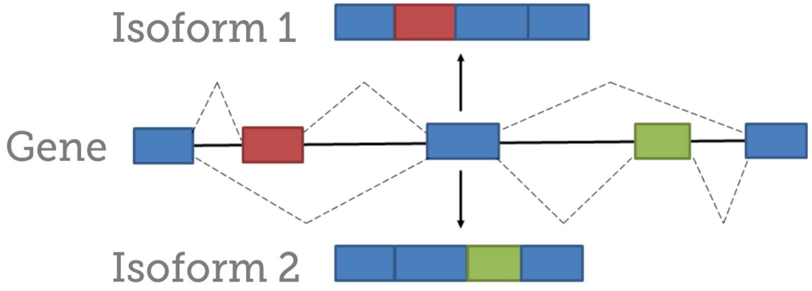
ion torrent

by life technologies™

What is RNA-Seq?



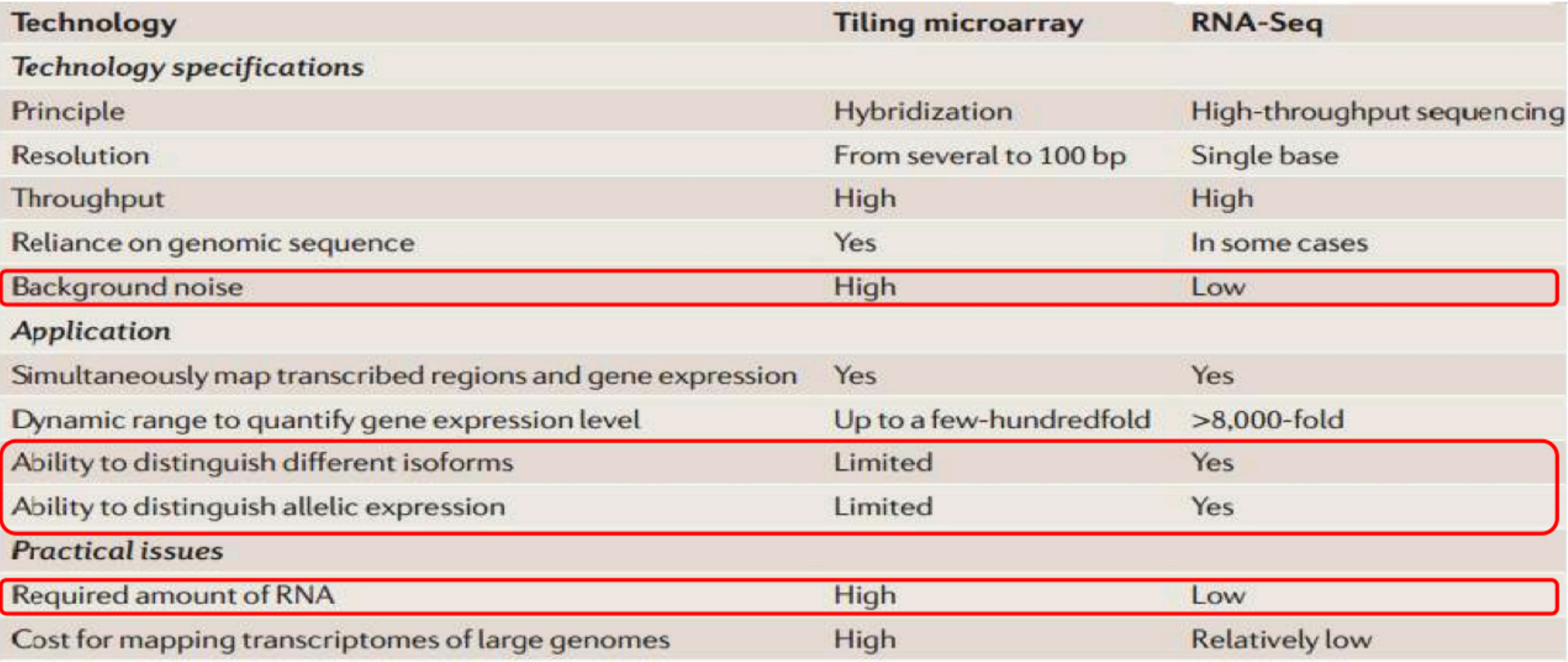
- Genes expression patterns vary in:
  - Tissue types
  - Cell types
  - Development stages
  - Disease conditions
  - Time points



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

- Additionally, RNA-Seq allows detection of novel transcripts.

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

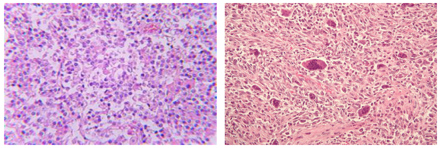
Zhong Wang, **Mark Gerstein**, and **Michael Snyder**

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

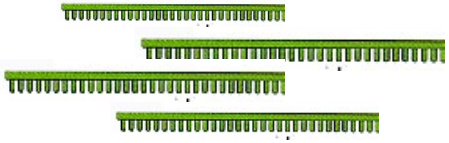
What are your questions ?

- This determines how you analyze the data.
- What are you looking for?
  - Novel transcripts, junctions?
  - Differential Gene expression?
  - Differential exon level counts?
  - Differential regulation?
  - Differential splicing?

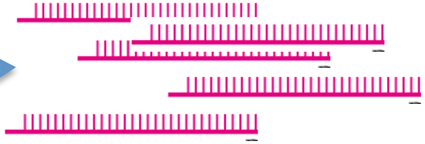
RNA-Seq… at it’s Most Basic Form



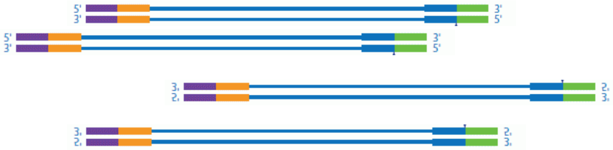
Samples from two conditions



Isolate RNA

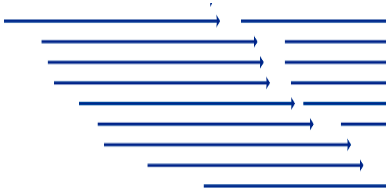


Generate cDNA

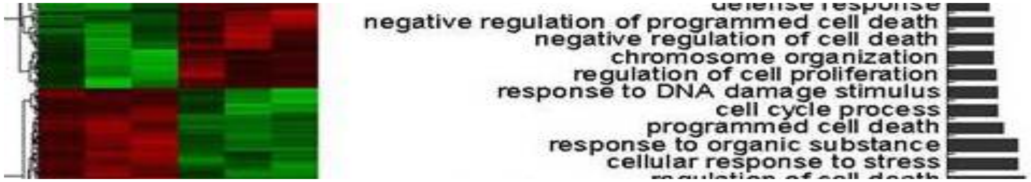


Create sequencing library by fragmenting, size selection and adding adaptors

Run sequencer

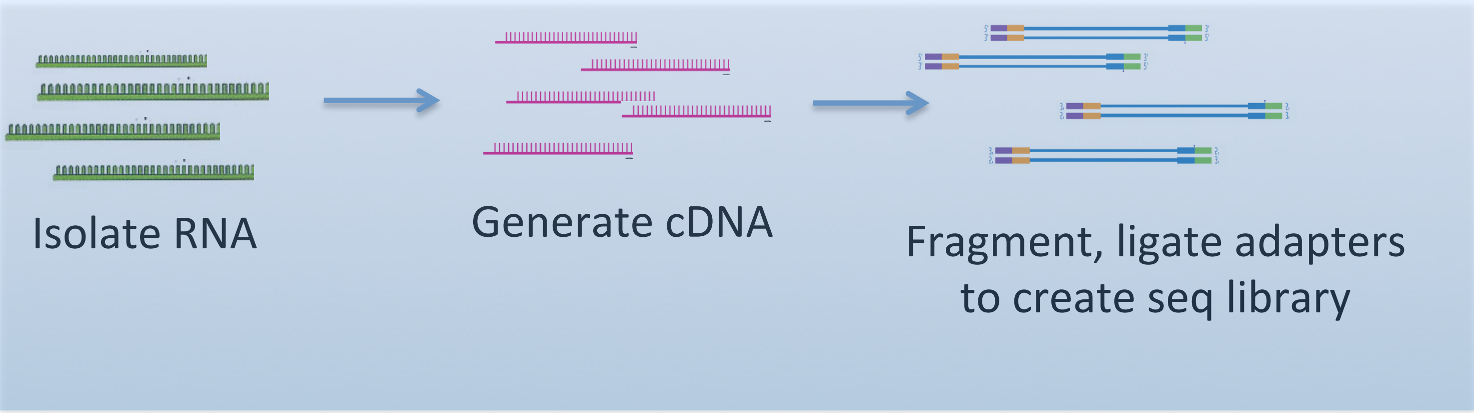


Generate short reads



Identify differentially expressed genes

RNA-Seq Libraries… with More Details

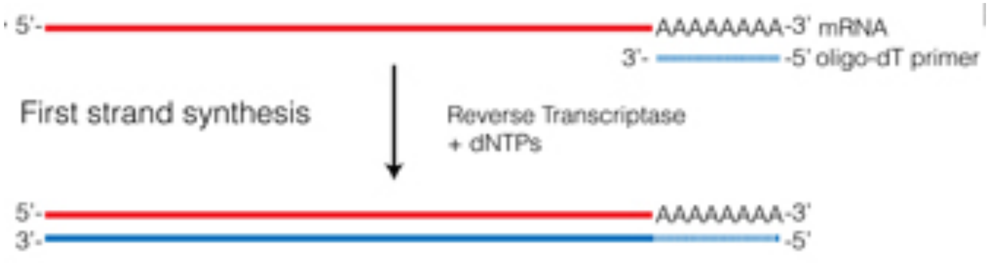


Isolate RNA

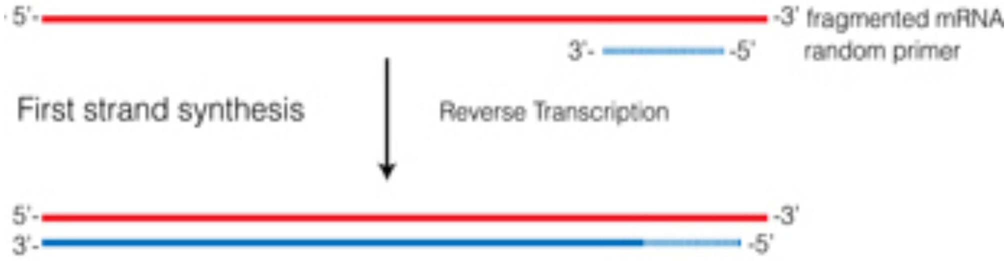
Generate cDNA

Fragment, ligate adapters to create seq library

A. Poly A Priming



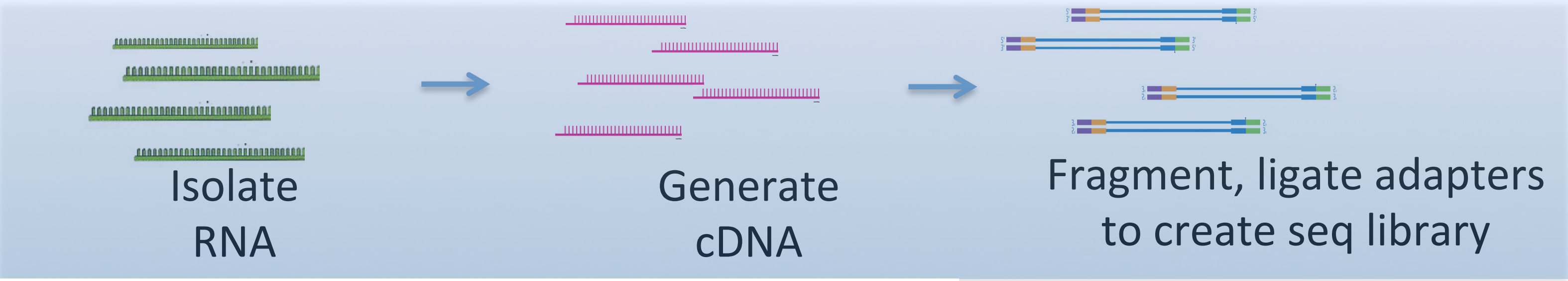
B. Random Priming



C. rRNA Depletion

Ribominus kit

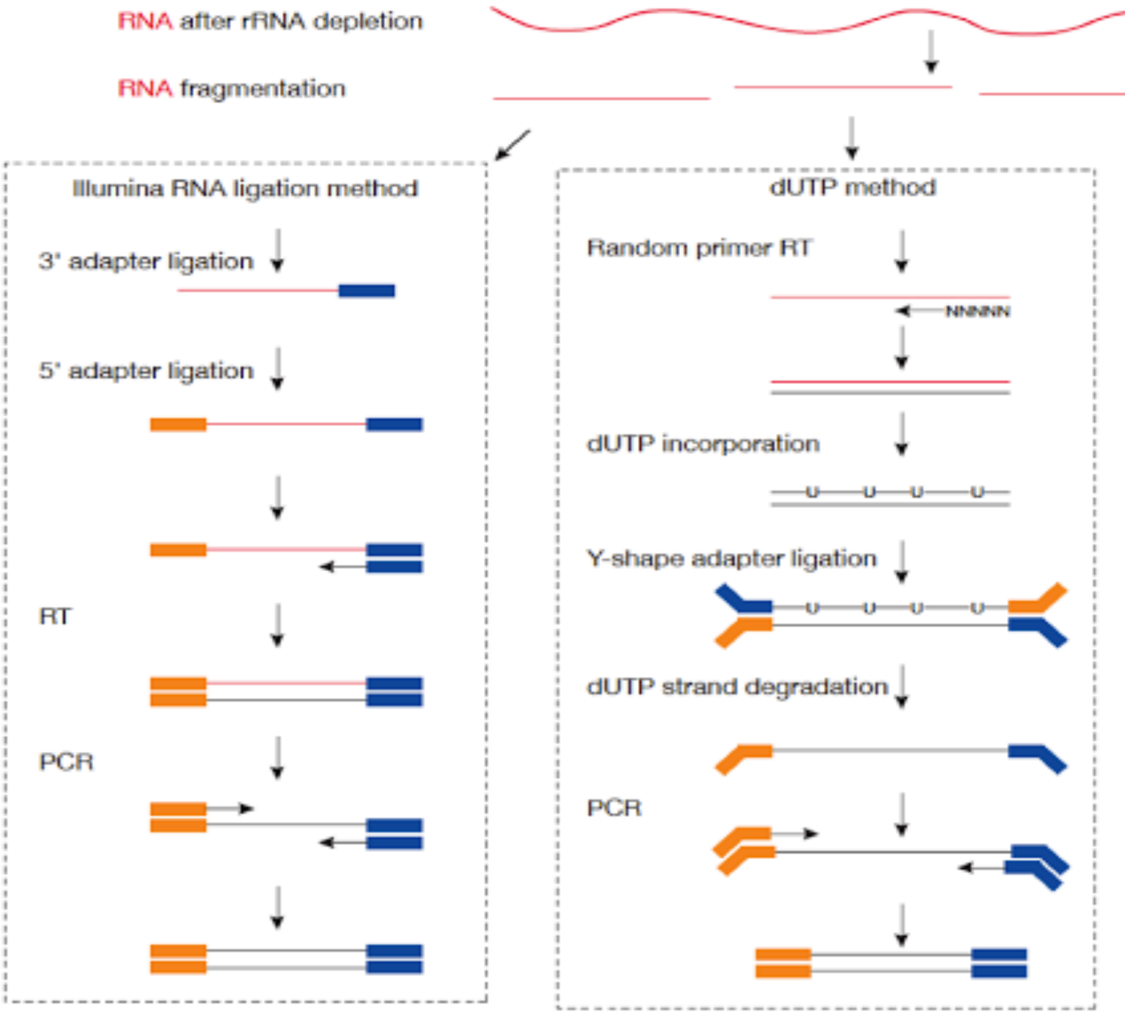
RNA-Seq Libraries… with More Details



Isolate RNA

Generate cDNA

Fragment, ligate adapters to create seq library



Second Strand Synthesis-

Many Strand Specific Methods.

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

Comparing Stranded RNA-Seq Library Protocols

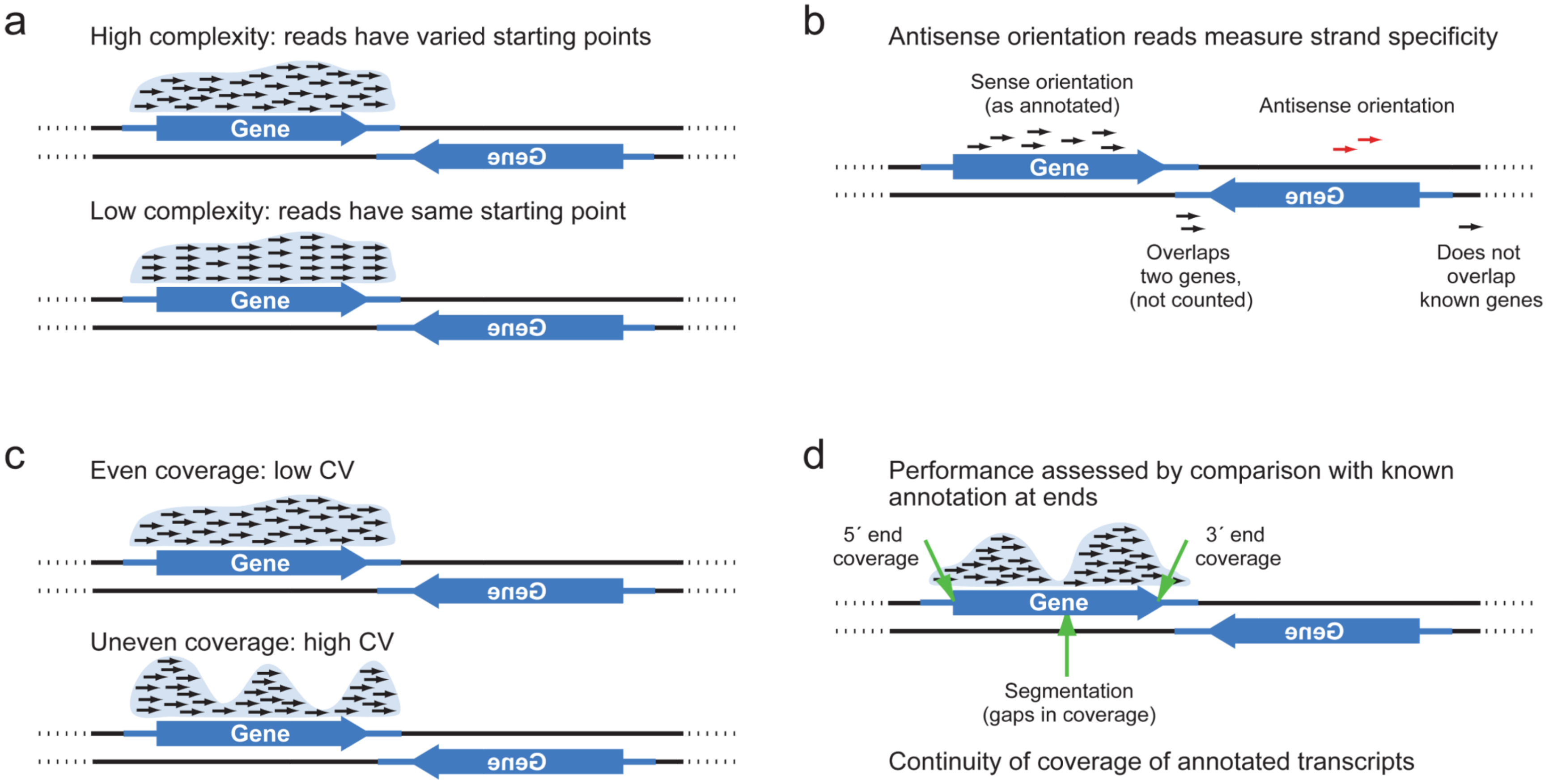
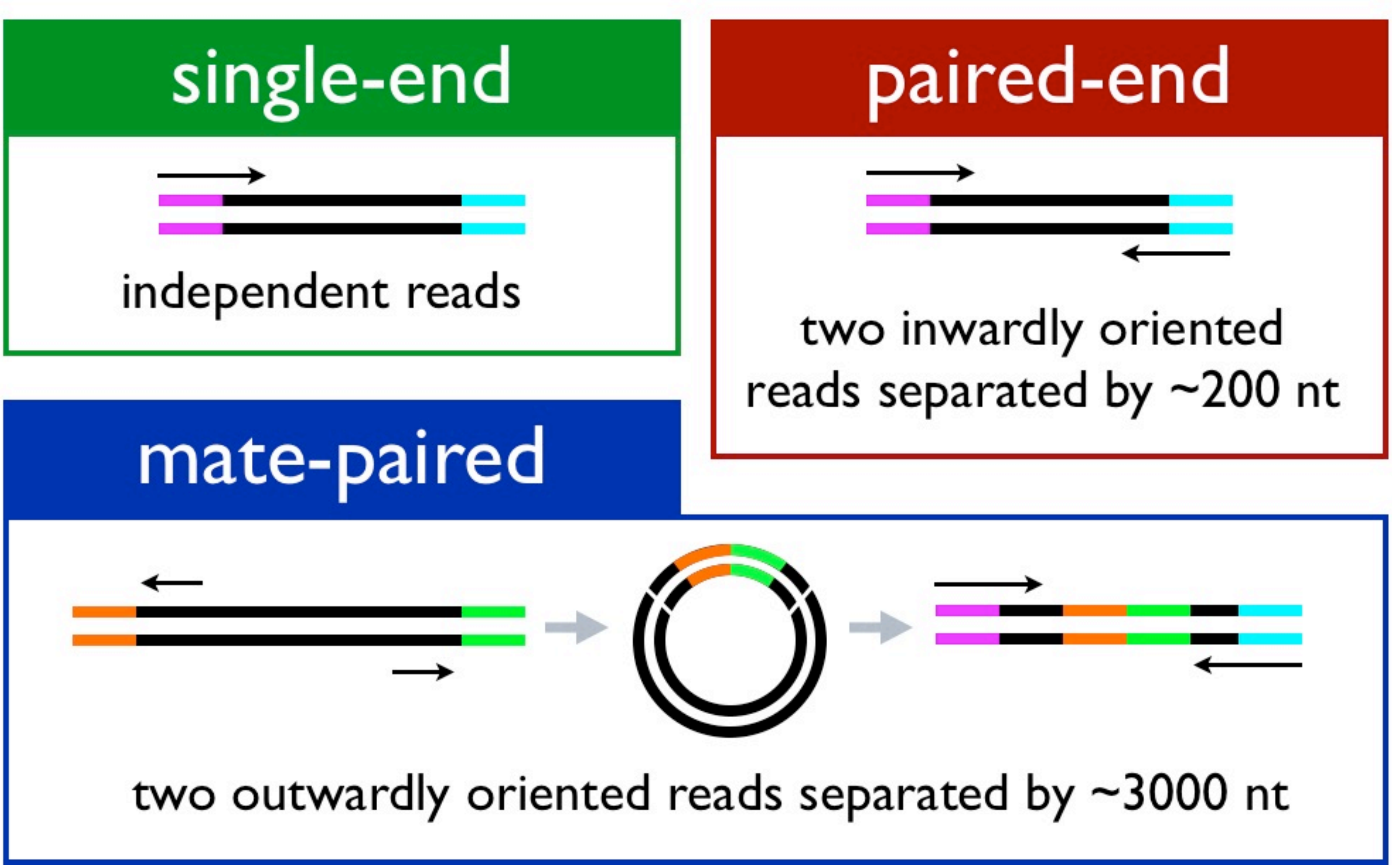

a
High complexity: reads have varied starting points
Gene
Gene
Low complexity: reads have same starting point
Gene
Gene
b
Antisense orientation reads measure strand specificity
Sense orientation (as annotated)
Antisense orientation
Gene
Gene
Overlaps two genes, (not counted)
Does not overlap known genes
c
Even coverage: low CV
Gene
Gene
Uneven coverage: high CV
Gene
Gene
d
Performance assessed by comparison with known annotation at ends
5´ end coverage
3´ end coverage
Gene
Gene
Segmentation (gaps in coverage)
Continuity of coverage of annotated transcripts

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

Types of Illumina Fragment Libraries



single-end

independent reads

paired-end

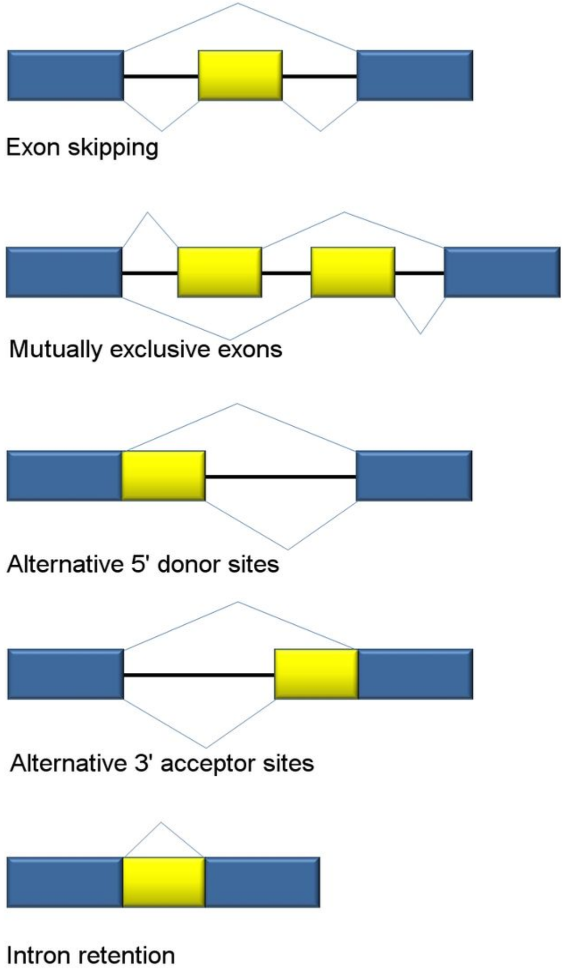
two inwardly oriented reads separated by ~200 nt

mate-paired

two outwardly oriented reads separated by ~3000 nt

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

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%

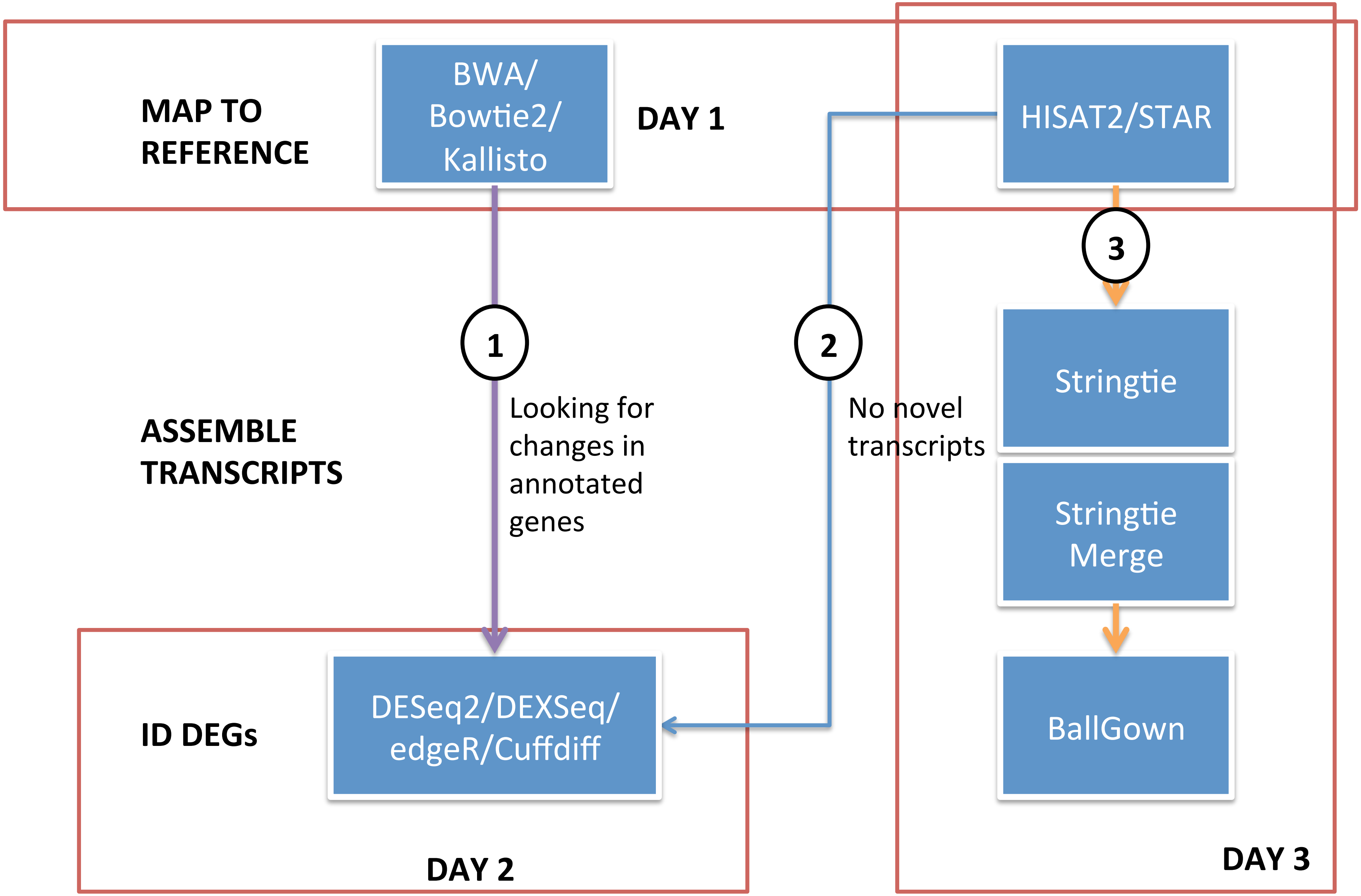
Quality scores are ASCII encoded in fastq files. Different platforms/older sequencing data can have different encoding! Illumina HiSeq 2500 produces Sanger encoded data.

Phred +33 =ASCII

How do we analyze RNA-Seq data?

- EVALUATE QUALITY AND MANIPULATE RAW DATA
- ALIGN READS
- ASSEMBLE TRANSCRIPTS
- QUANTIFY TRANSCRIPTS
- TEST FOR DIFFERENTIAL EXPRESSION
- VISUALIZE
- DOWNSTREAM ANALYSIS

The Big Picture



MAP TO REFERENCE

BWA/ Bowtie2/ Kallisto

DAY 1

HISAT2/STAR

3

1

2

ASSEMBLE TRANSCRIPTS

Looking for changes in annotated genes

No novel transcripts

Stringtie

Stringtie Merge

ID DEGs

DESeq2/DEXSeq/ edgeR/Cuffdiff

BallGown

DAY 2

DAY 3

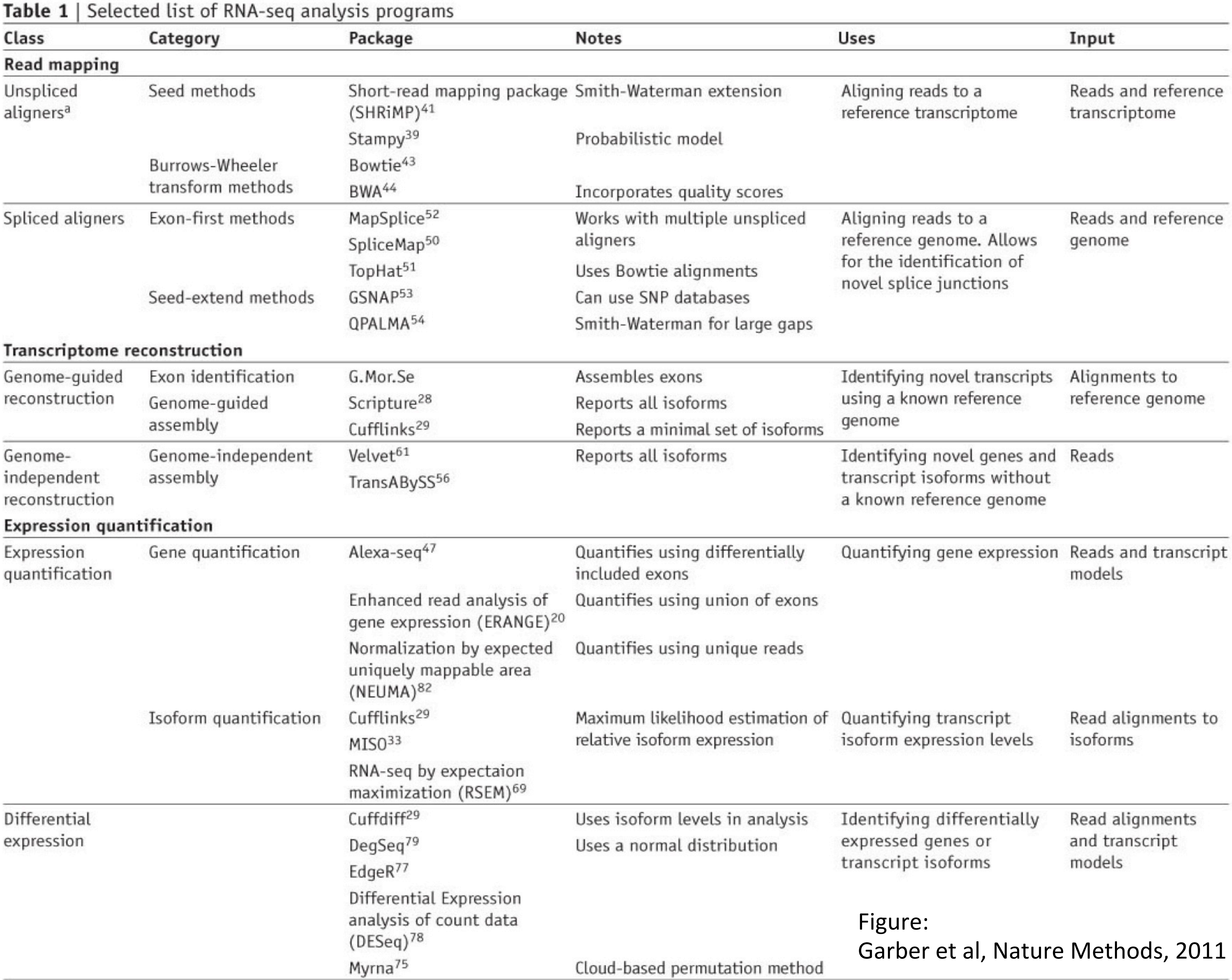
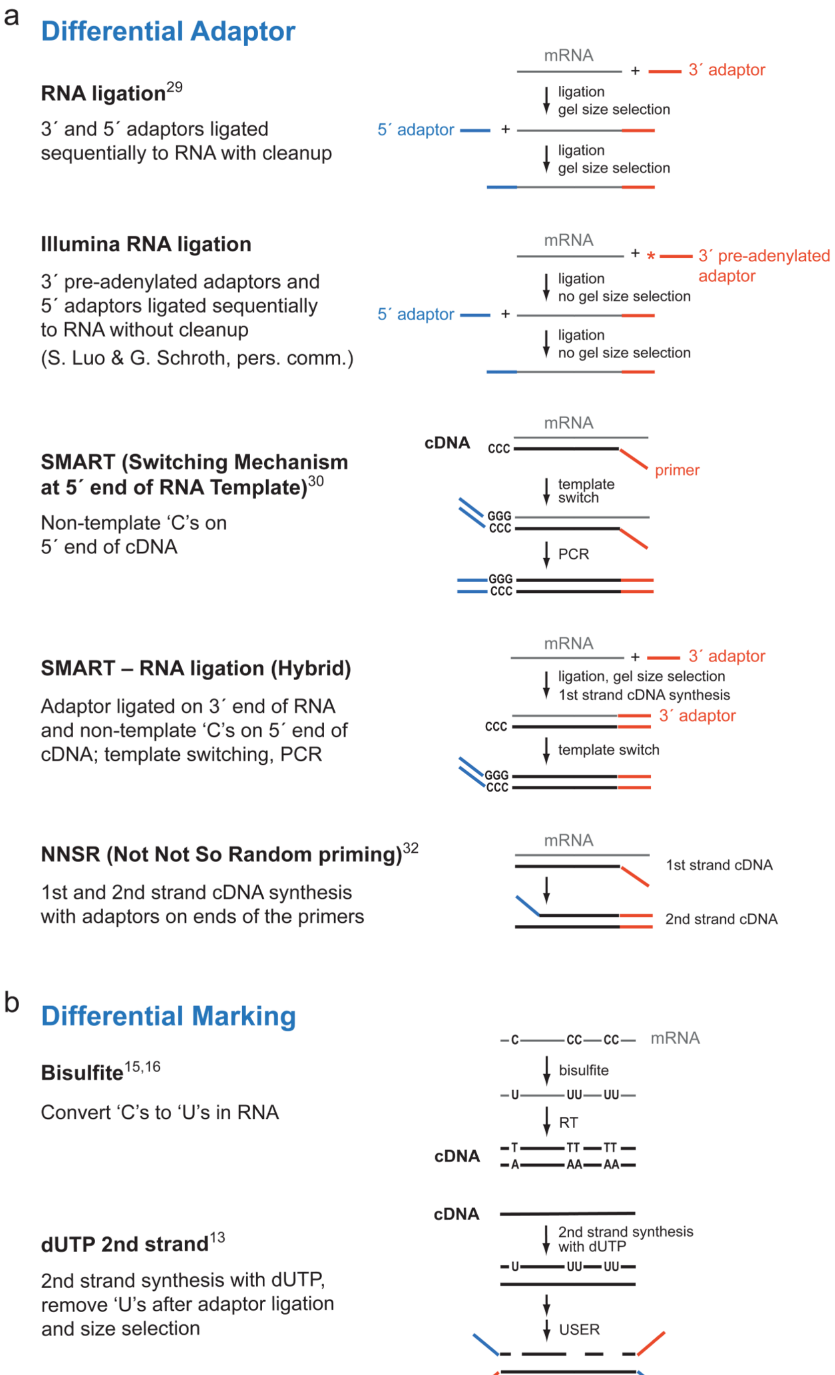


Table 1 | Selected list of RNA-seq analysis programs

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



a **Differential Adaptor**

RNA ligation[29]

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 strand[13]

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.