BCB 5200 Introduction to Bioinformatics I

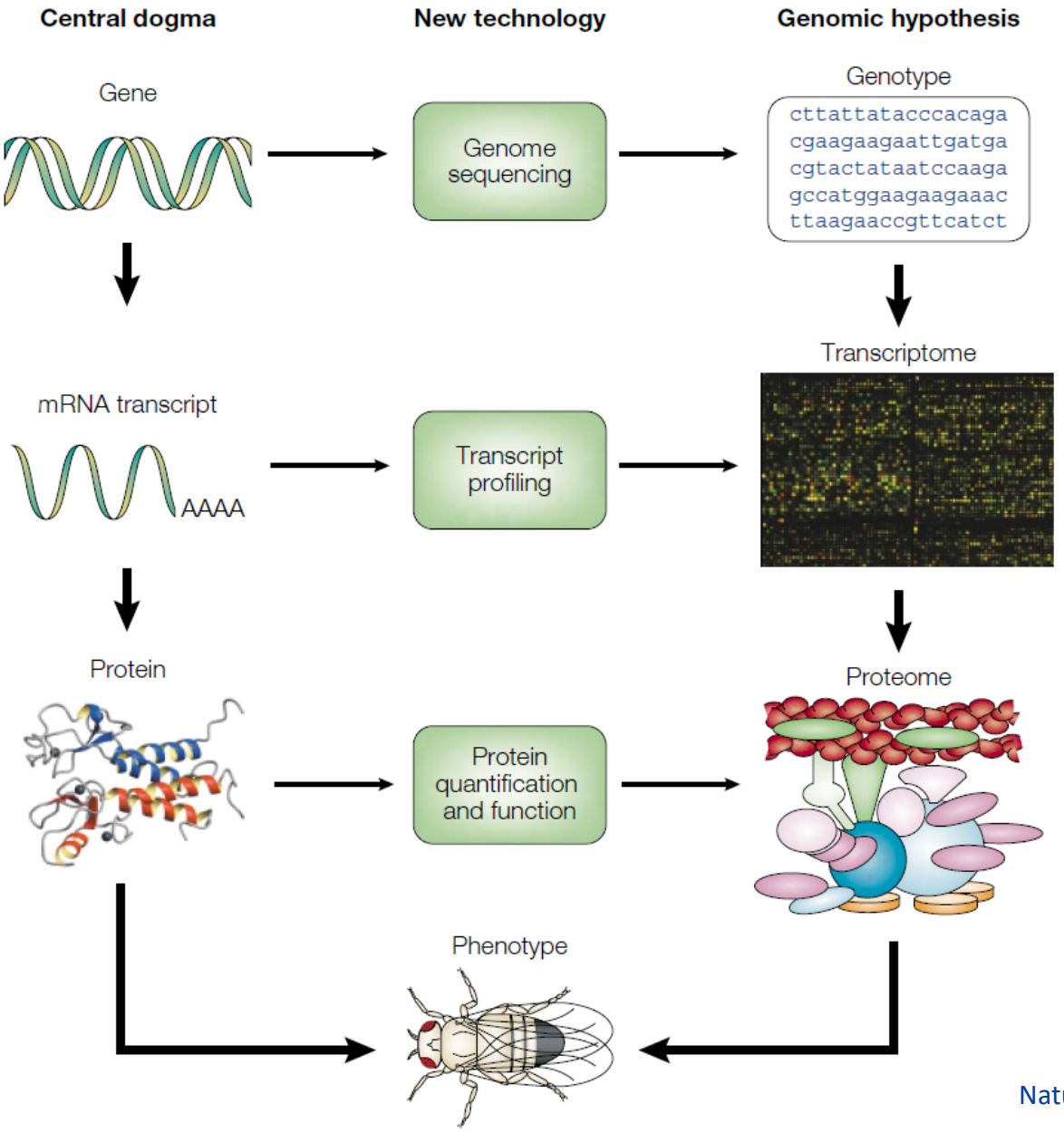
Transcriptome and RNA-seq: an overview

Zhenguo Lin, PhD
Department of Biology
Fall 2017

Today's Topic

- Transcriptome and RNA-seq: an overview
  - Background information
  - RNA-seq workflow
    - Experimental design
    - RNA preparation
    - Library preparation
    - Sequencing
    - Data analysis

Evolution of the central dogma



Each step of the central dogma is accompanied by recent technological innovations that allow genome-wide analysis

Nature Reviews Genetics 3: 43-53 (2002).

What is transcriptome?

- Transcriptome
  - A transcriptome is a collection of all the transcripts present in a given cell, and their quantity, for a specific developmental stage or physiological condition
- Is a transcriptome the same as a genome?
  - No, a transcriptome represents the very small percentage of the genome

http://www.genome.gov/13014330

What can a transcriptome tell us?

- Can determine when and where each gene is turned on or off in the cells and tissues of an organism
- Interpreting the functional elements of the genome
- Revealing the molecular constituents of cells, tissues
- Understanding development and disease

RNA-Seq: a revolutionary tool for transcriptomics Nat Rev Genet. 2009 Jan;10(1):57-63.

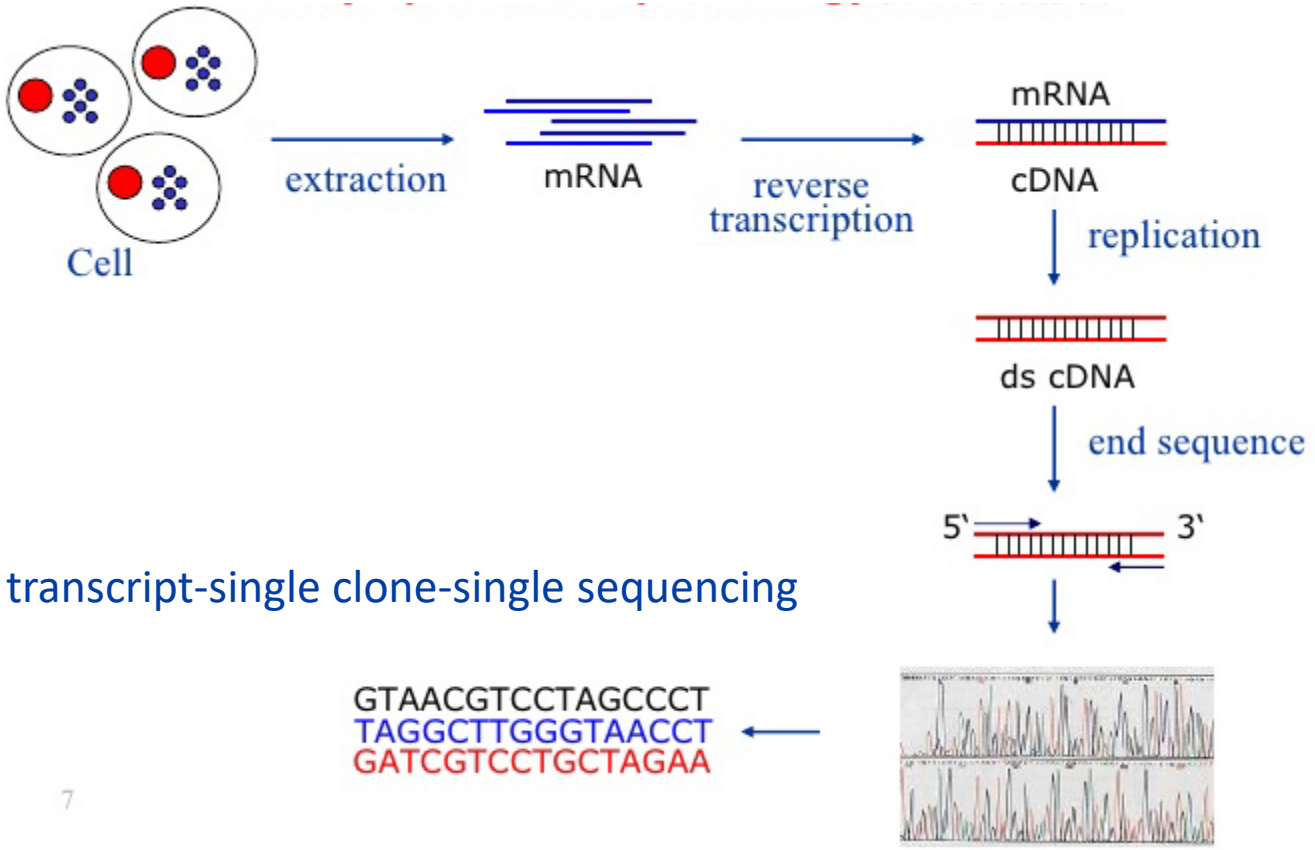
How to obtain a transcriptome?

- Sequencing of expressed sequence tags (ESTs):
  - high cost, ideal for discovery of new gene
- Serial analysis of gene expression (SAGE):
  - lowered costs by minimizing the amount of information collected per transcript, but still relatively high
- Microarray
- RNA-sequencing

Nature Methods 5, 585 - 587 (2008)

EST

Ideal for discovery of new gene, but high cost

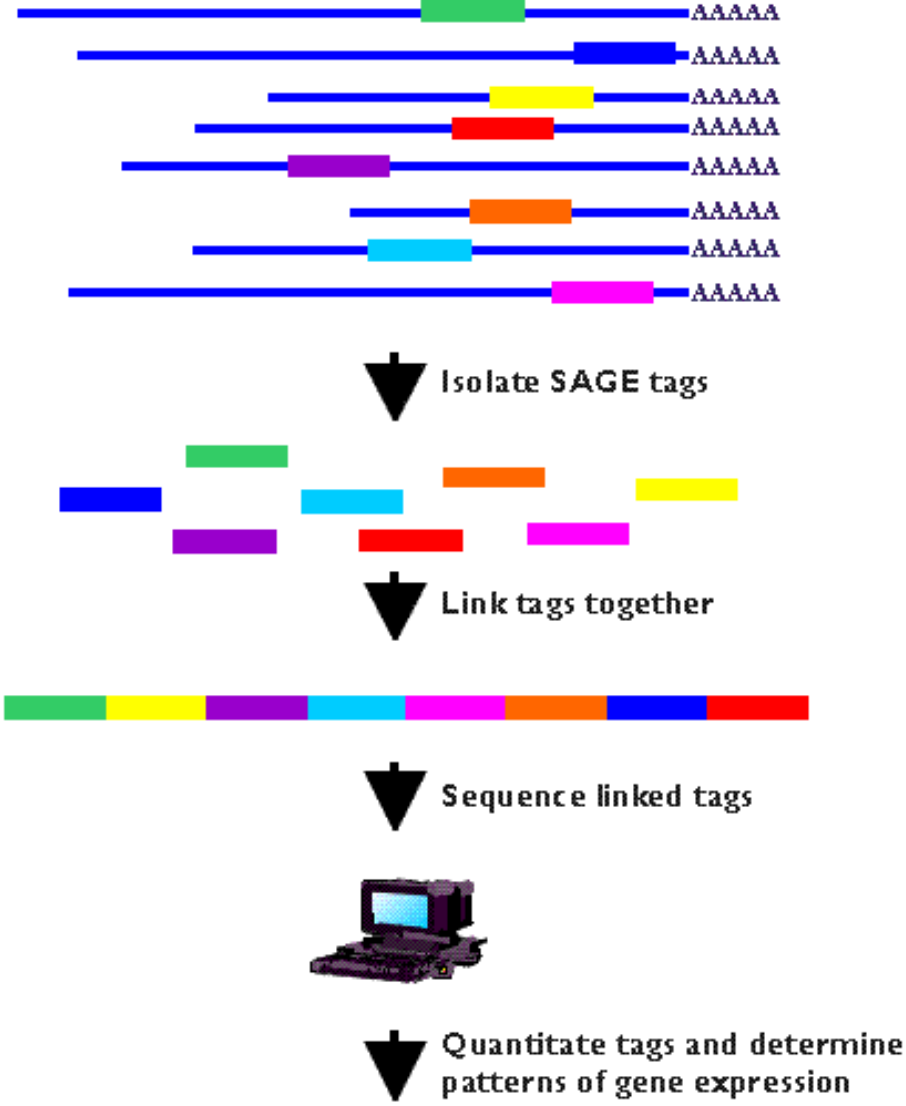


Single transcript-single clone-single sequencing

Adams, M. D., et al. (1991). Science 252(5013): 1651-1656

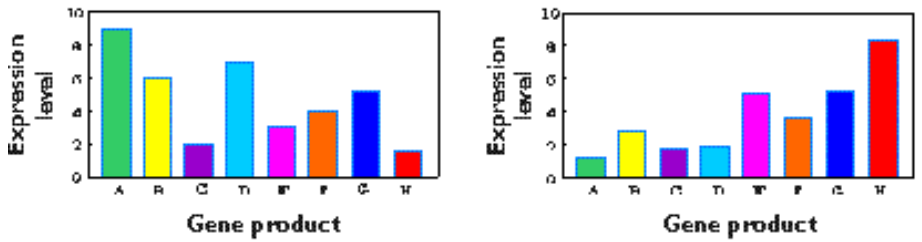
SAGE

- lowered costs by minimizing the amount of information collected per transcript, but still relatively high
- multiple transcripts-multiple tags-single clone-single sequencing

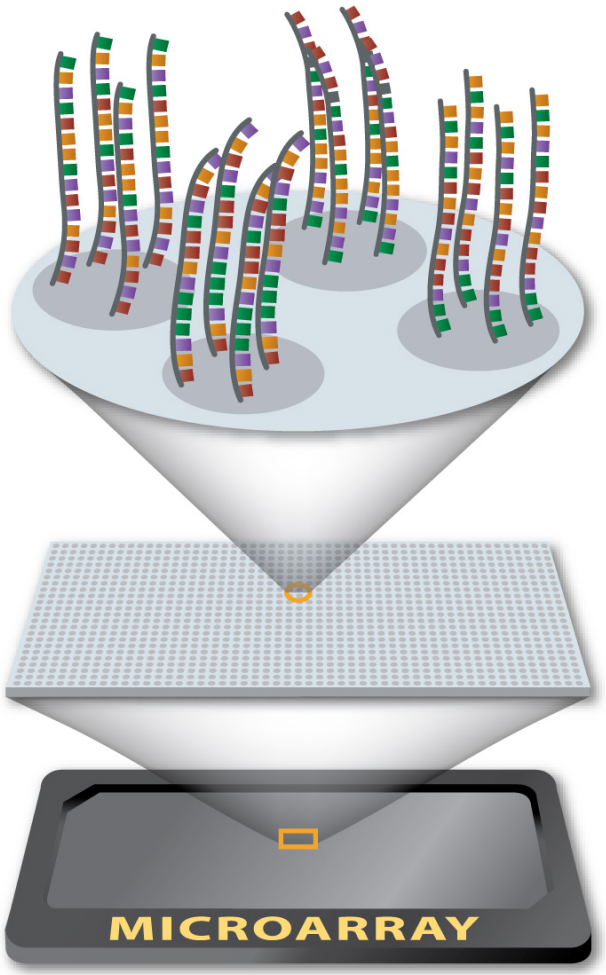


Velculescu, V. E., et al. (1995). "Serial analysis of gene expression." Science 270(5235): 484-487

http://www.sagenet.org/findings/index.html



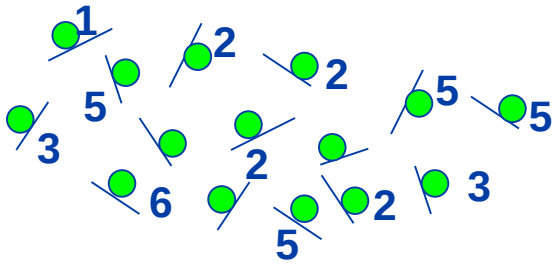
Microarray



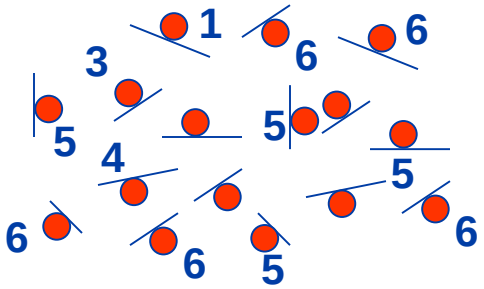
- A DNA micorarray allows scientists to perform an experiment on thousands of genes at the same time.
  - Each spot on a microarray contains multiple identical strands of DNA.
  - The DNA sequence on each spot is unique.
  - Each spot represents one gene.
  - Thousands of spots are arrayed in orderly rows and columns on a solid surface (usually glass).
  - The precise location and sequence of each spot is recorded in a computer database.
  - Microarrays can be the size of a microscope slide, or even smaller.

Microarray procedures: Illumina platform

cDNA "control"
Cy3 labeled



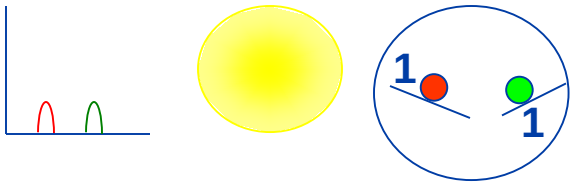
cDNA "treament"
Cy5 labeled



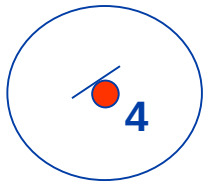
HYBRIDIZATION

Relative intensity **Output**

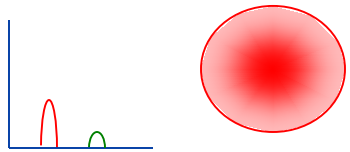
Relative intensity **Output**

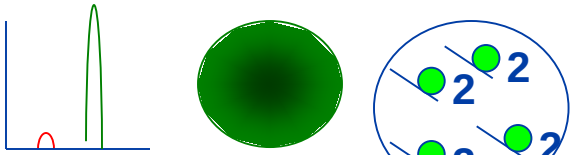


Spot #1

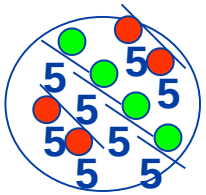


Spot #4

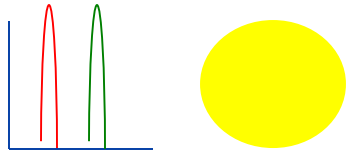


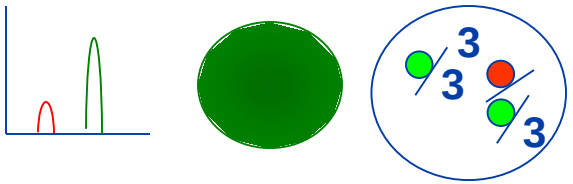


Spot #2

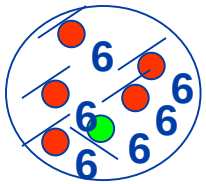


Spot #5

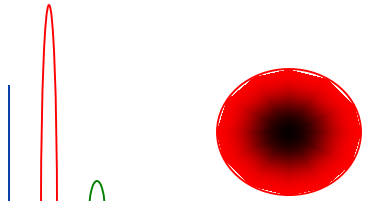




Spot #3



Spot #6



Microarray

- Limitations
  - knowledge of the sequences being interrogated is a prerequisite for array design
  - analysis of highly related sequences is problematic because of cross-hybridization
  - the analog nature of the signal makes it difficult to confidently detect and quantify low-abundance species
  - Low reproducibility of results between laboratories and across platforms

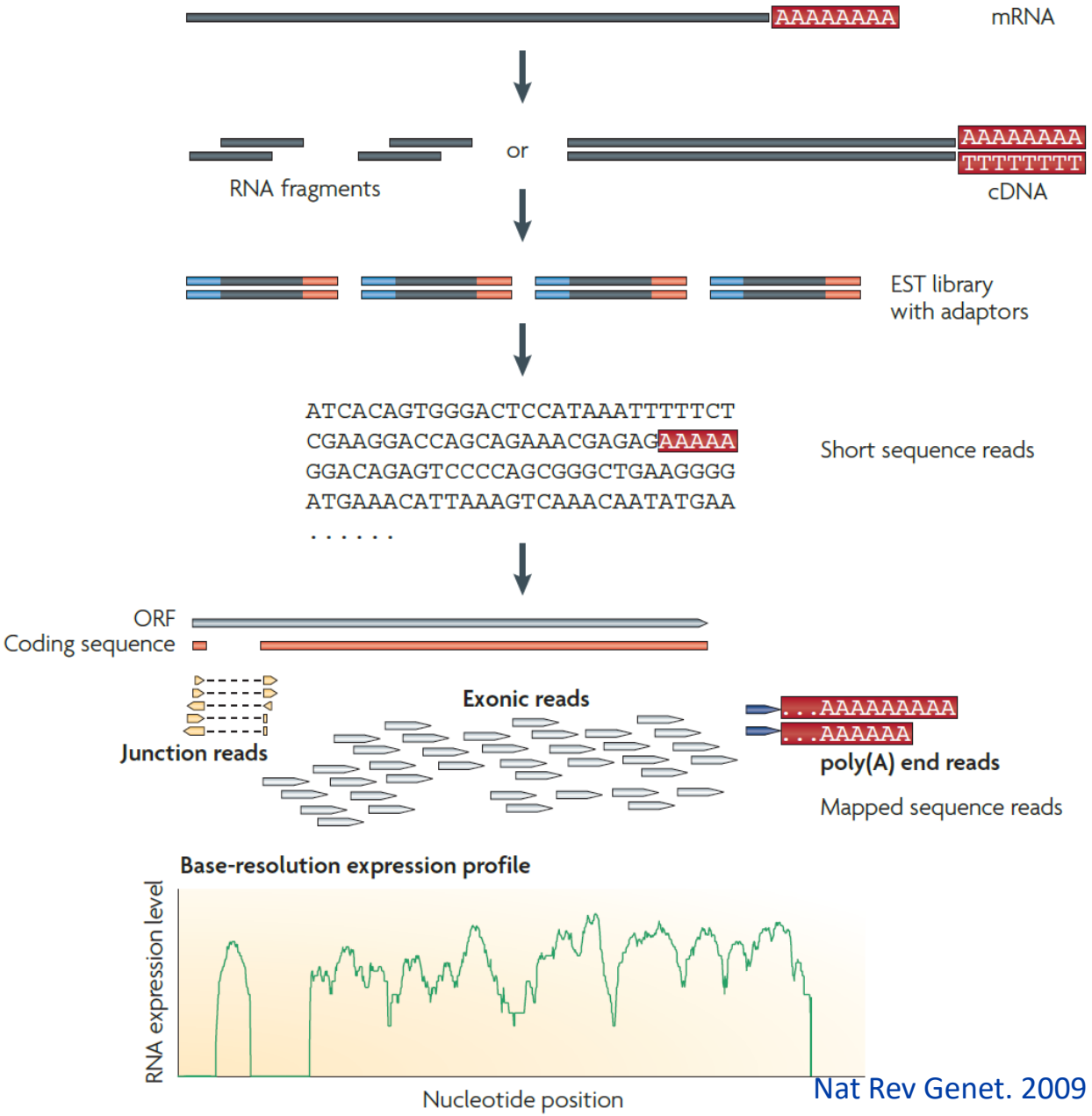
Nature Methods 5, 585 - 587 (2008)

RNA sequencing

- RNA sequencing is experimental procedures that generate DNA sequence reads derived from the entire RNA molecules
- In theory, RNA-seq can be used to build a complete map of the transcriptome across all cell types, perturbations and states

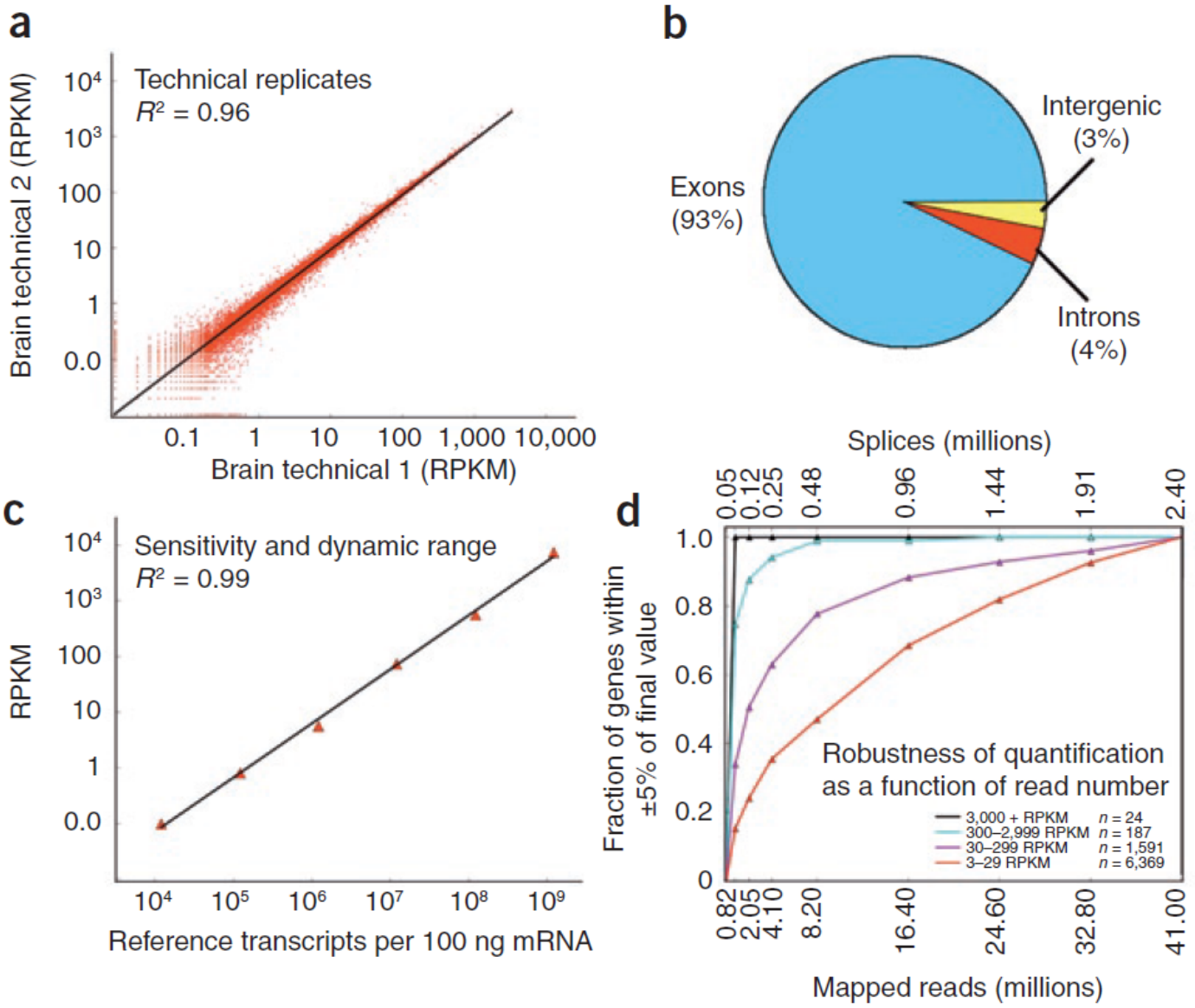
DOI:10.1038/NMETH.1613

A typical RNA-seq experiment



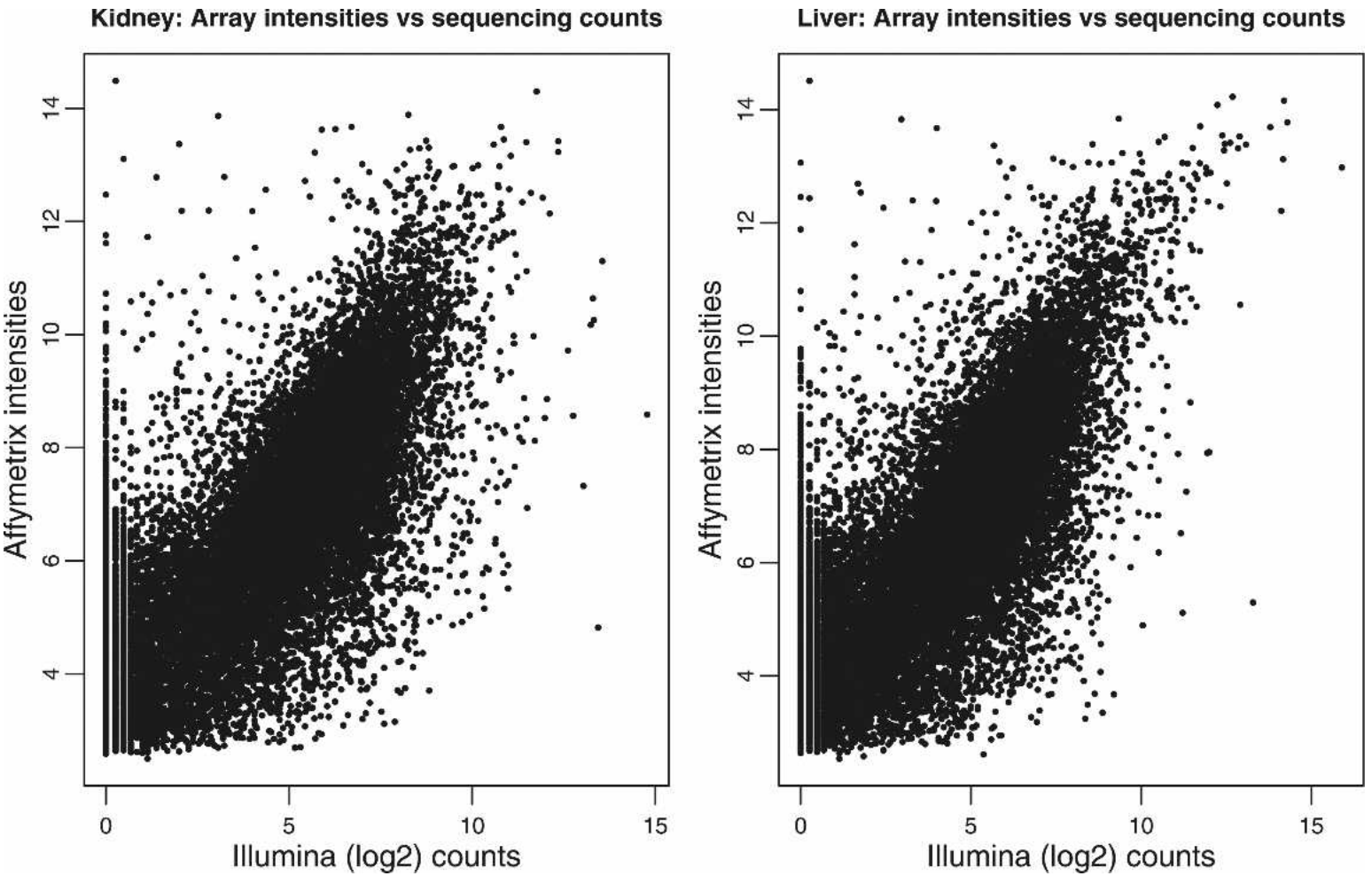
Nat Rev Genet. 2009 Jan;10(1):57-63.

RNA-seq: Reproducibility, linearity and sensitivity



Ali Mortazavi et al., Nature Methods - 5, 621 - 628 (2008)

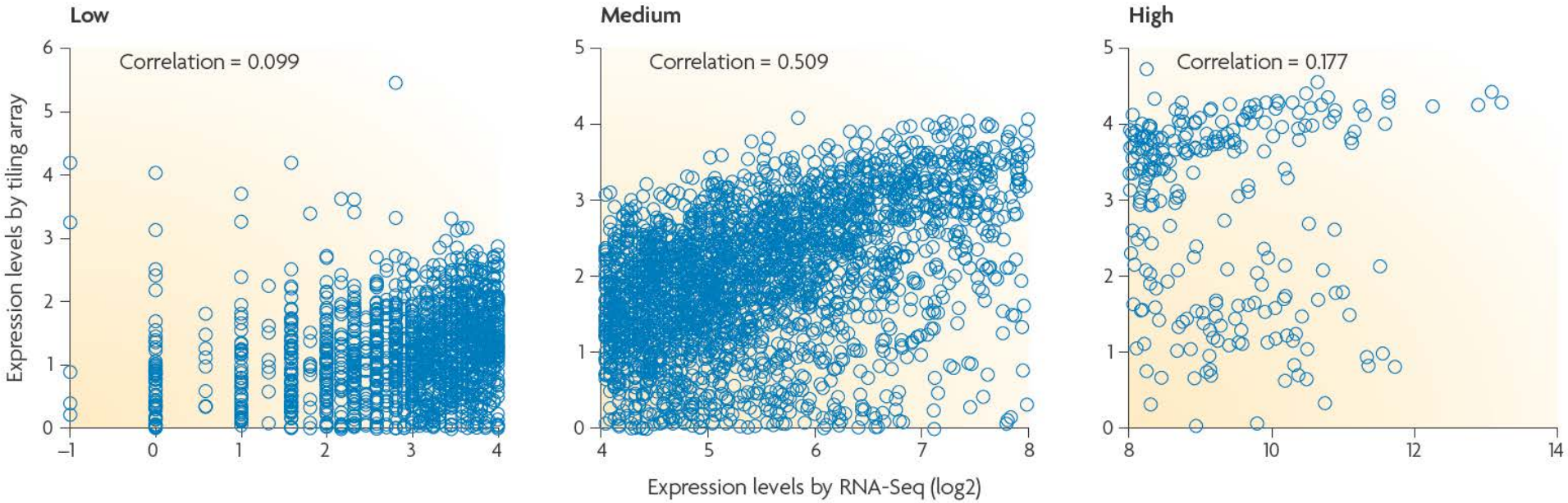
Illumina counts vs microarray intensities



SAINT LOUIS UNIVERSITY.

Genome Res. 2008. 18: 1509-1517

Quantifying expression levels: RNA-seq vs. microarray



Saccharomyces cerevisiae cells grown in nutrient rich media

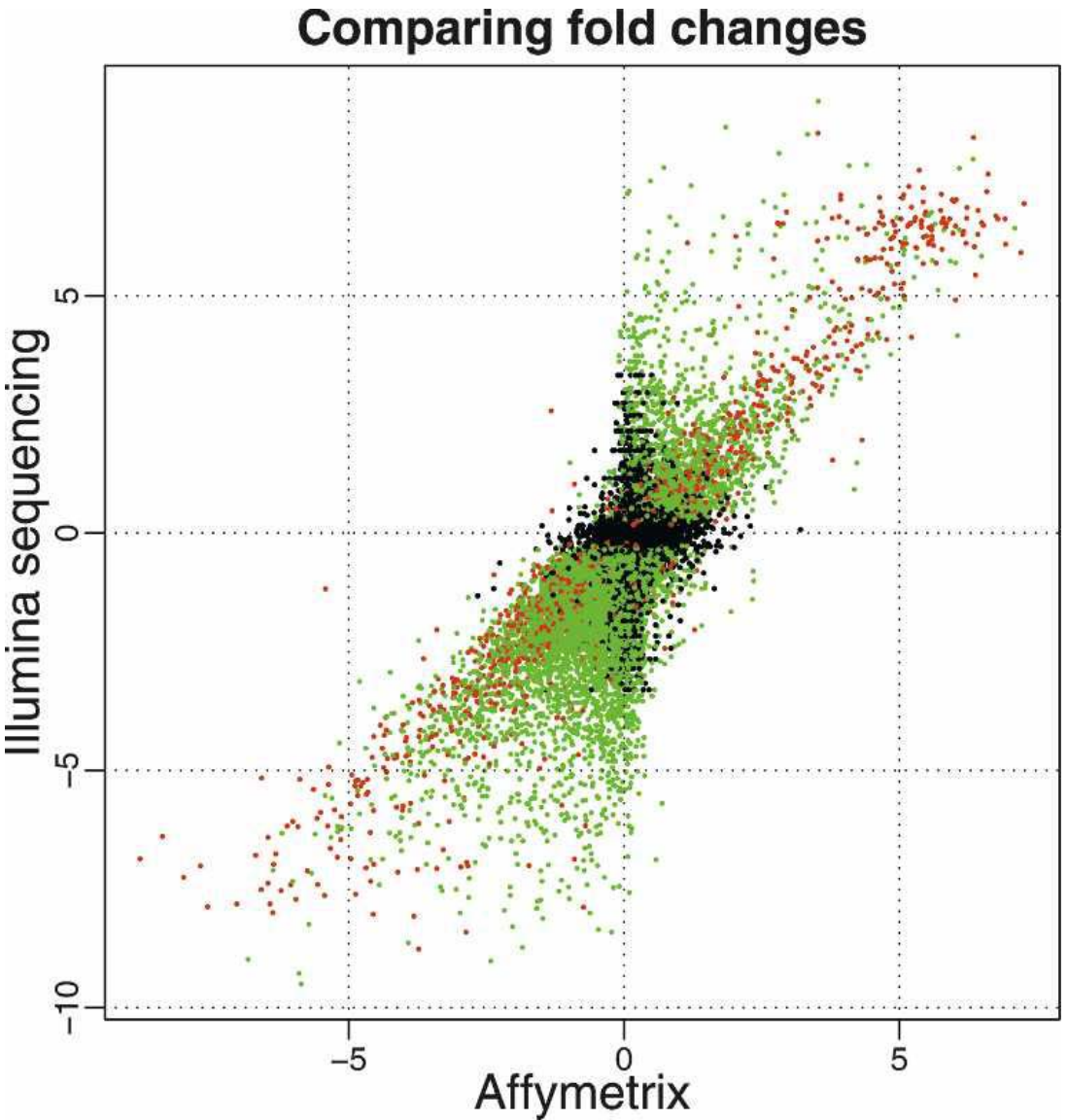
SAINT LOUIS UNIVERSITY.

Wang et al 2009, Nature Reviews Genetics 10:57-63

log2 fold changes from Illumina and Affymetrix

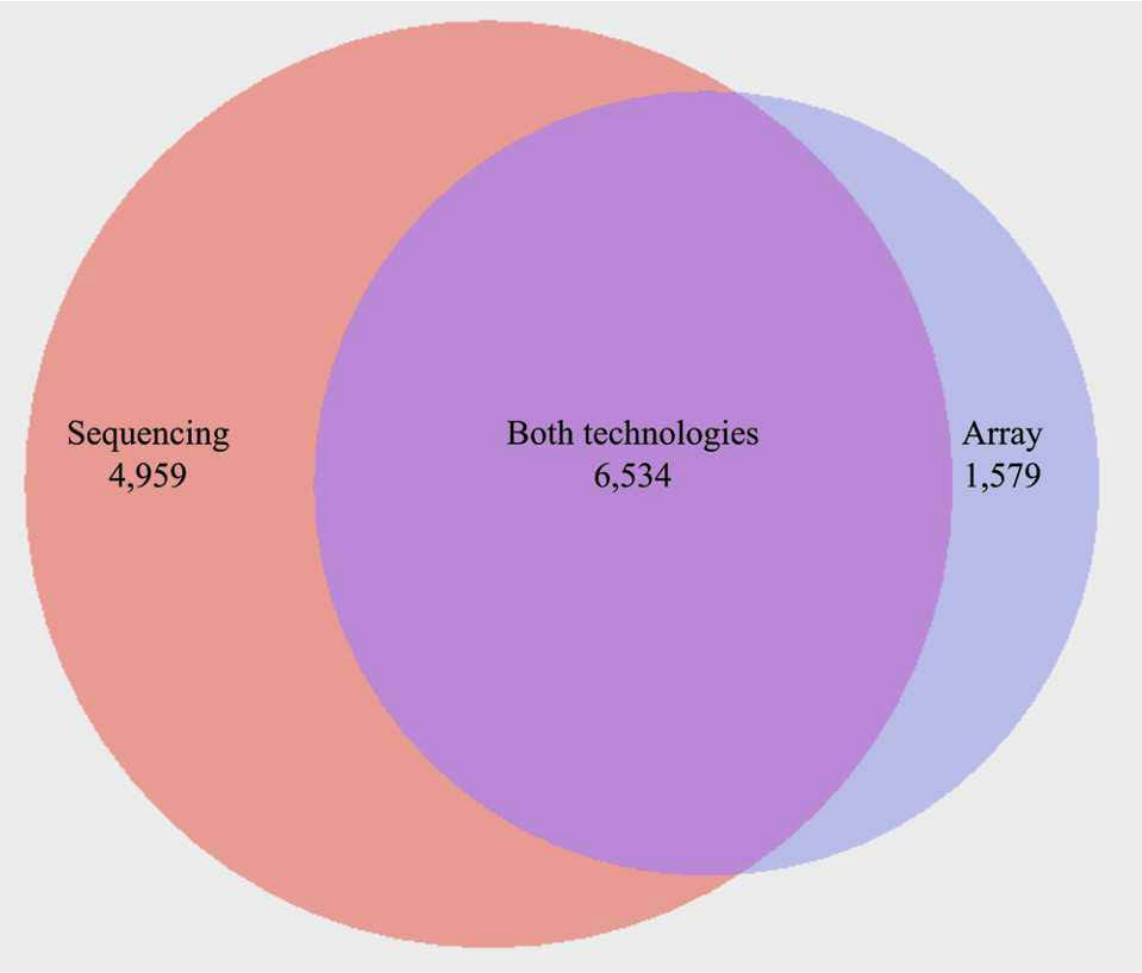
mean number of counts greater than (red) or less than (green) 250 reads in both tissues
Black dots: Genes not called as differentially expressed based on the Illumina sequencing data

The strongest correlation between the two technologies seems to be those that are mapped to by many reads (red), while the correlation is weaker for differentially expressed genes mapped to by fewer reads (green).



Genome Res. 2008. 18: 1509-1517

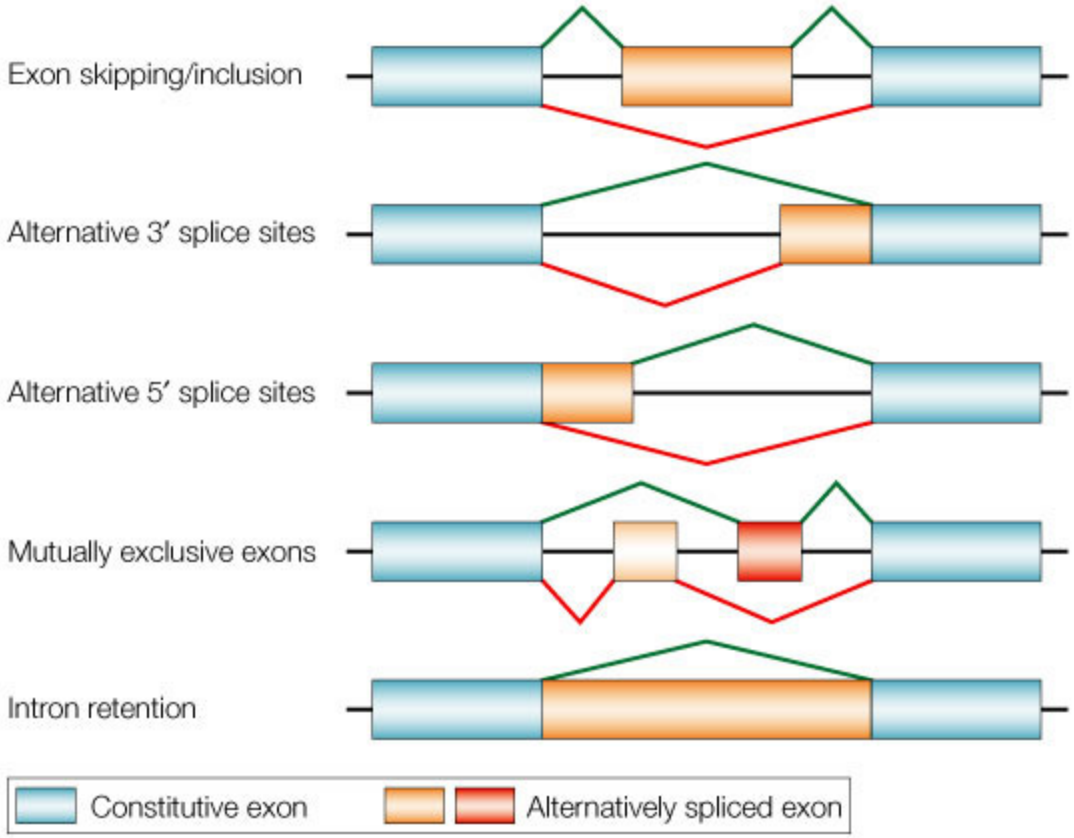
Illumina vs Affymetrix: differentially expressed genes



SAINT LOUIS UNIVERSITY.

Genome Res. 2008. 18: 1509-1517

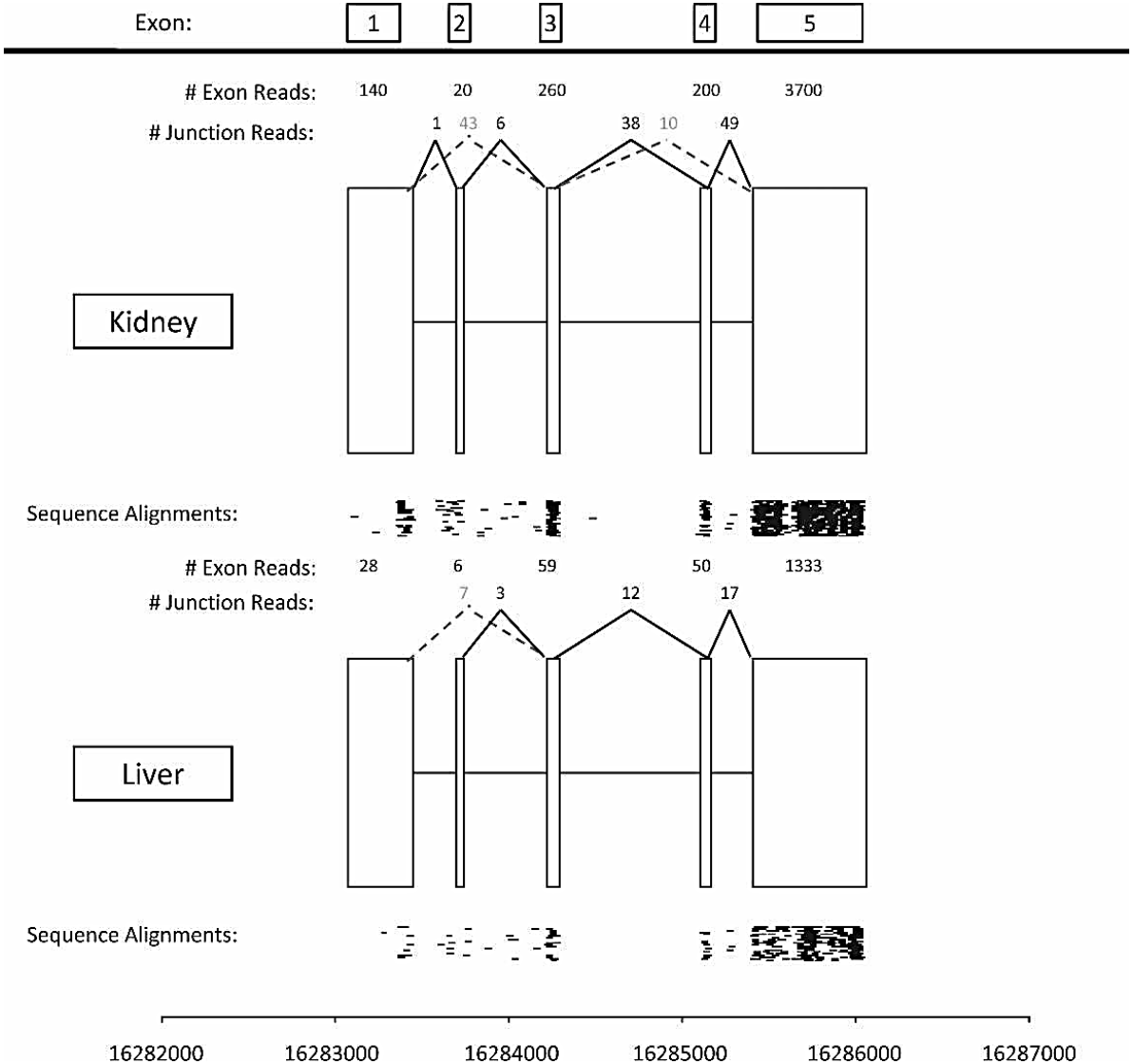
A schematic representation of alternative splicing



SAINT LOUIS UNIVERSITY.

Nature Reviews Genetics 3, 285–298 (2002).

An example of alternative splicing



SAINT LOUIS UNIVERSITY.

Genome Res. 2008. 18: 1509-1517

Microarrays, EST, RNA-Seq

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang et al 2009, Nature Reviews Genetics 10:57-63

RNA-seq resolve Microarray issues

- Microarray Limitations
  - knowledge of the sequences being interrogated is a prerequisite for array design
    - RNA-seq: Prior sequence knowledge is not required
  - analysis of highly related sequences is problematic because of cross-hybridization
    - RNA-seq: paralogous sequences can be distinguished
  - the analog nature of the signal makes it difficult to confidently detect and quantify low-abundance species
    - RNA-seq: quantitation is 'digital' rather than 'analog', with ensuing benefits for dynamic range and sample comparison
  - Low reproducibility of results between laboratories and across platforms
    - RNA-seq: the digital nature of the quantitation will also lead to superior inter-platform reproducibility

Nature Methods 5, 585 - 587 (2008)

Other Advantages of RNA-seq

- Cost: Lower than traditional sequencing
- Can discover new alternative splice junctions and transcriptional units simultaneously with gene expression measurements
- Quantitative linearity over a broad dynamic range
- Can reveal sequence variations (SNPs)

Nature Methods 5, 585 - 587 (2008)

Applications of RNA-Seq

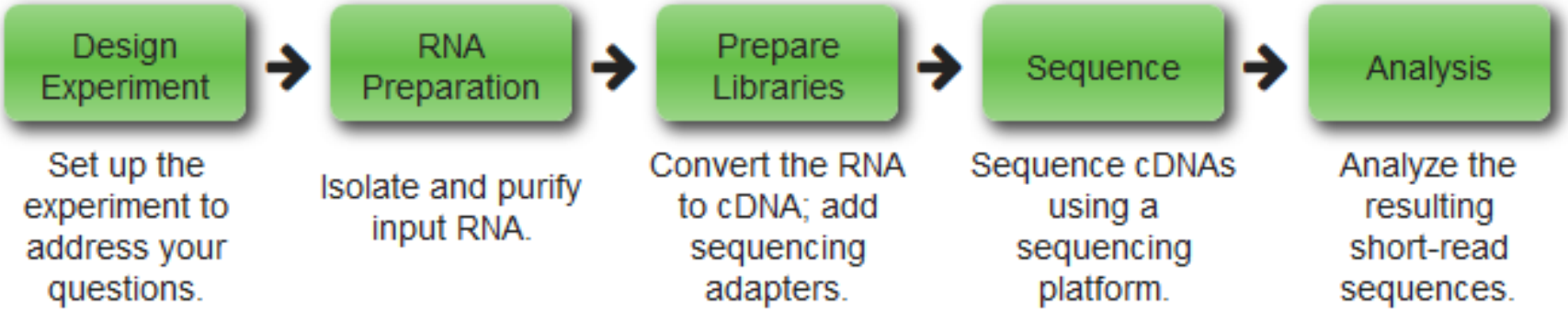
- **Annotation (Qualitative): identifying**
  - expressed transcripts
  - exon/intron boundaries
  - alternative splicing
  - poly-A sites or alternative poly-A sites (if use polyA enrichment to prepare library)
- **Measure differential gene expression (DGE)** (Quantitative)**:** measuring between two or more treatments or groups

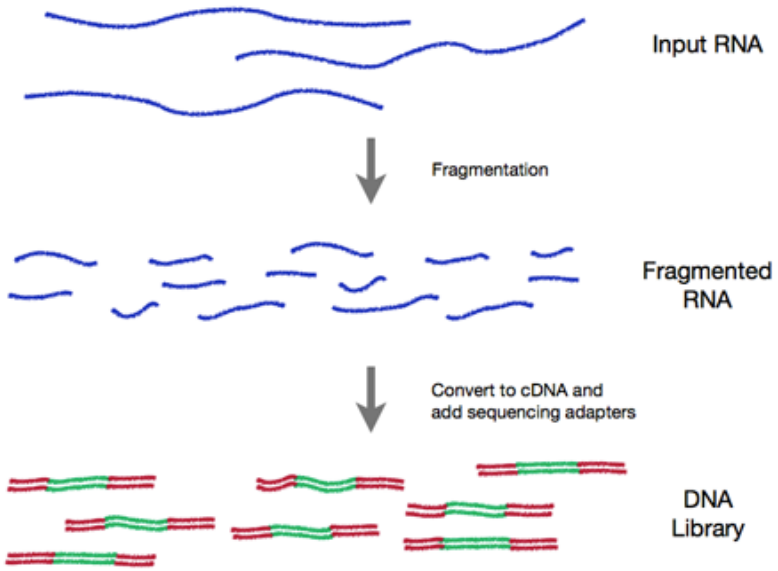
Common challenges

- Mapping of short sequence reads to the genome
- Appropriate assignment of ‘multi-mapping’ reads
- Identification of new alternative splice junctions
- Classification of reads mapping outside annotated boundaries
  - i.e., distinguishing genomic DNA contamination vs. pre-mRNA vs. new transcriptional units vs. belonging to adjacent transcriptional units;
- Comparison of samples to identify differentially expressed genes
- Improving efficiency and reducing bias in library construction

Nature Methods 5, 585 - 587 (2008)

RNA-seq Workflow





http://rnaseq.uoregon.edu/#introduction

Step 1: Experimental design

- What are your goals?
  - Transcriptome assembly?
  - Differential expression analysis?
  - Identify rare transcripts?
- What are the characteristics of your system?
  - Large, complex genome?
  - Introns and high degree of alternative splicing?
  - No reference genome or transcriptome?

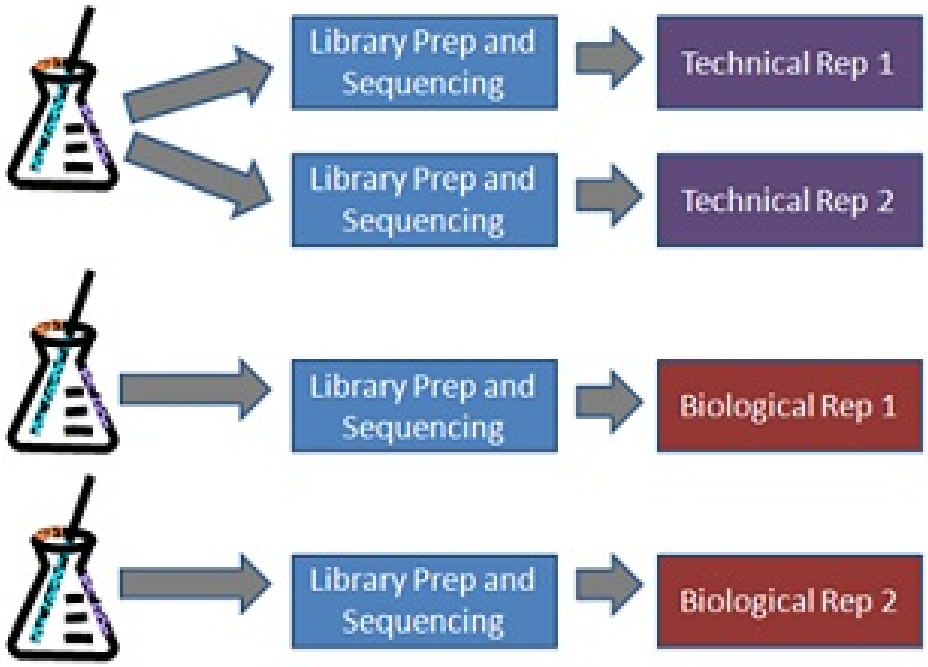
Recommendations for RNA-seq options based upon

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 normalization	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

http://rnaseq.uoregon.edu/#introduction

Technical vs. Biological Replications

- Biological replication: the same type of organism is grown/treated under the same conditions
- Technical replication: when the exact same sample is analyzed multiple times
- Biological variation is large relative to technical variation
- Recommend greater resource allocation to biological replication

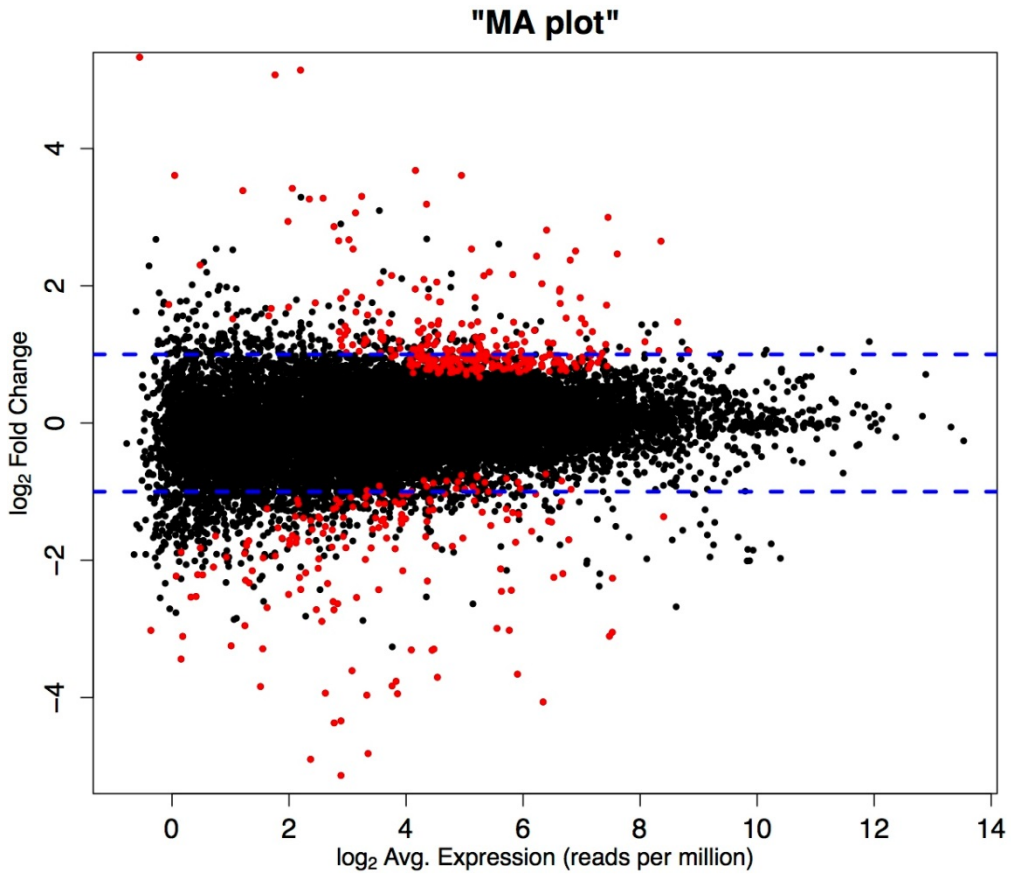


http://rnaseq.uoregon.edu/#exp-design-depth-of-sequencing

Technical vs. Biological Replications

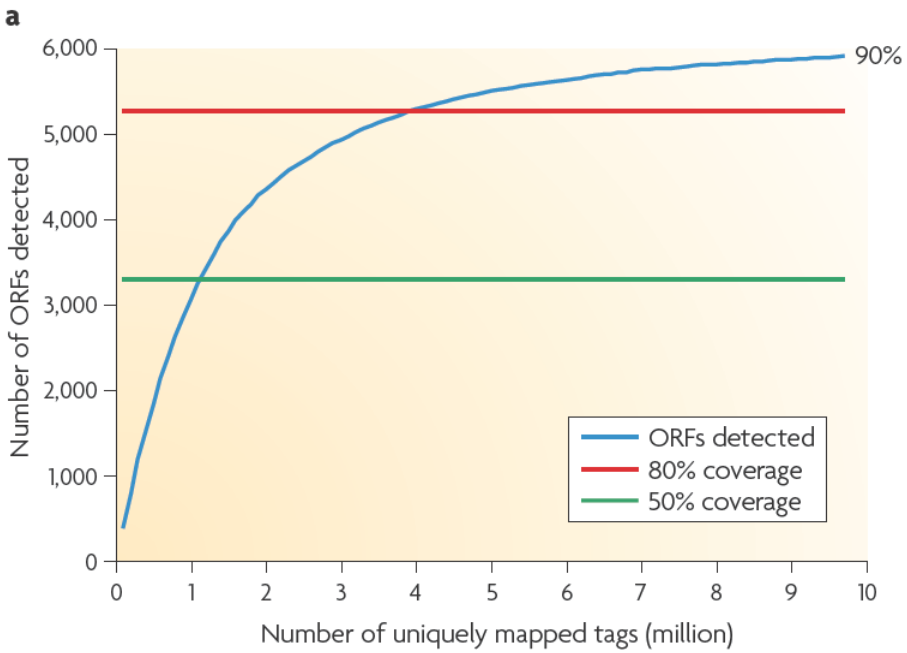
- Technical replicates
  - Not necessary, low technical variation
- • Biological replicates
  - Not necessary for transcriptome assembly
  - Essential for differential expression analysis
  - Difficult to estimate
    - 3+ for cell lines
    - 5+ for inbred lines
    - 20+ for human samples

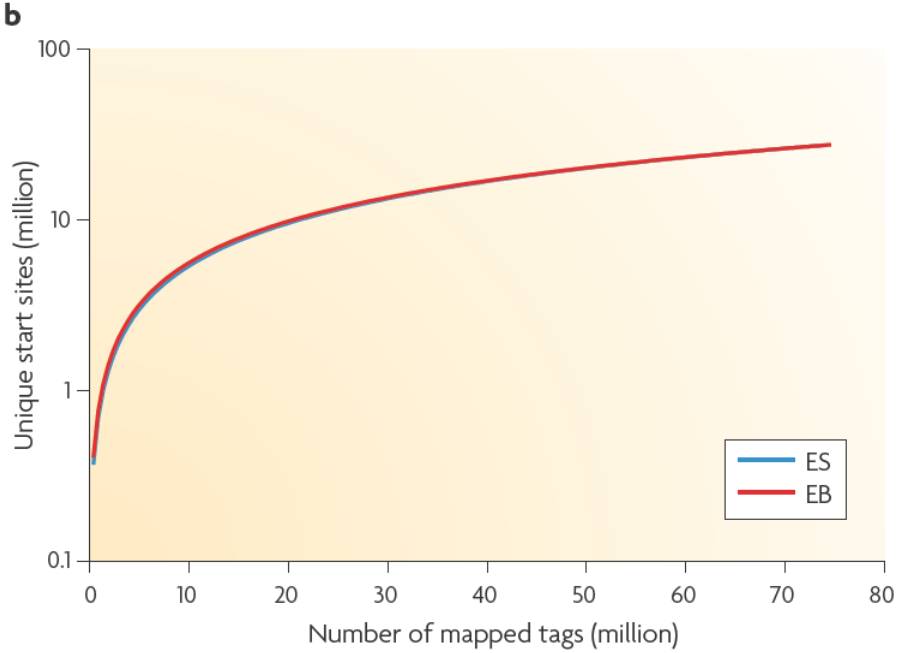
Depth of Sequencing



Variation due to the sampling process makes an especially large contribution to the total variance among individuals for transcripts represented by few reads

How many reads are sufficient?





In yeast, 30 million 35nt reads from poly(A) mRNA libraries are sufficient to observe transcription from most (>90%) genes.

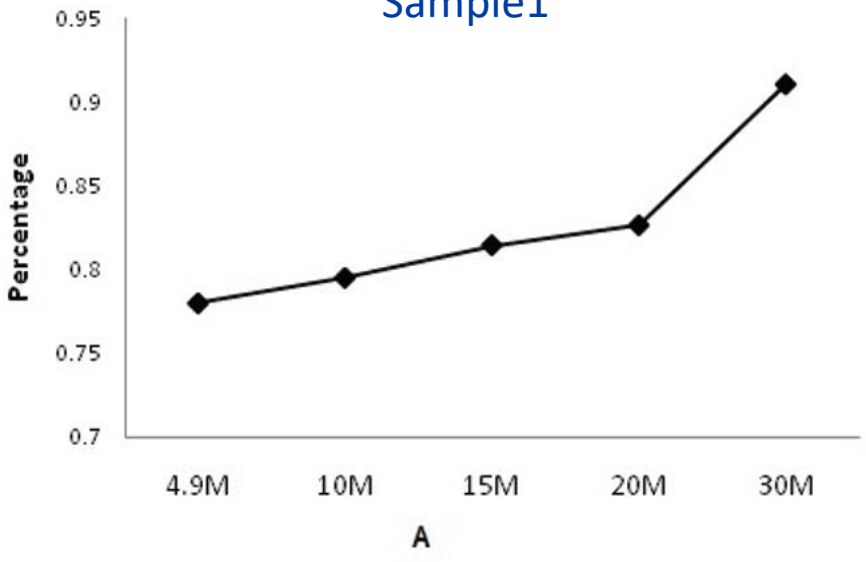
In mouse embryonic cells, at 80 million reads (25 nt), the number of start sites reached a plateau

SAINT LOUIS UNIVERSITY.

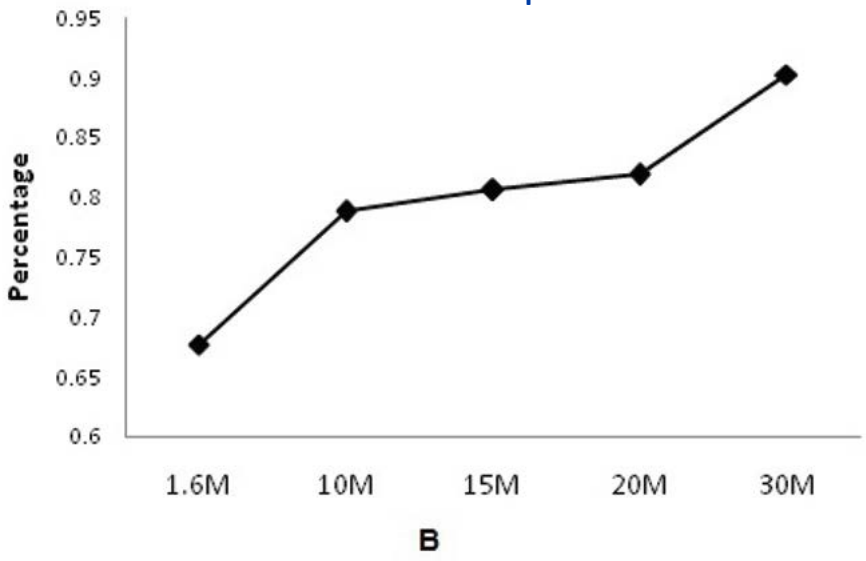
Nat Rev Genet. 2009 Jan;10(1):57-63.

How many reads are sufficient?

Sample1



Sample2

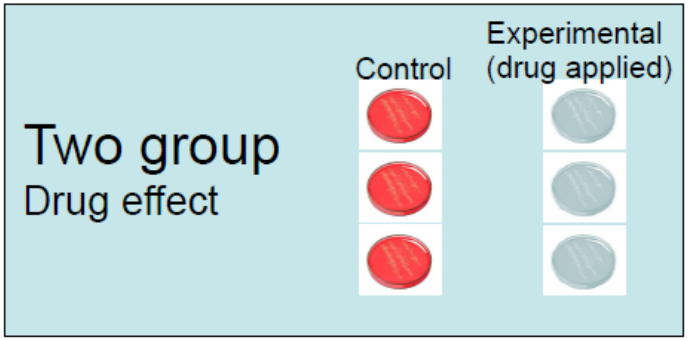


Percentages of detected chicken genes at different levels of sequence depth across all annotated chicken genes

30 M (75 bp) reads is sufficient to detect all annotated genes in chicken lungs. Ten million (75 bp) reads could detect about 80% of annotated chicken genes, and RNA-Seq at this depth can serve as a replacement of microarray technology

Wang et al. BMC Bioinformatics 2011, 12(Suppl 10):S5

Experimental Complexity



Drug 1 with 2 treatments (+ or -)

+ N = 6	- N = 6

Total # ind. = 12

Drug 1 with 2 treatments (+ or -) by Drug 2 with 2 treatments (+ or -) by Drug 3 with 2 treatments (+ or -)

- / - / - N = 6	- / - / + N = 6	- / + / - N = 6	- / + / + N = 6
+ / - / - N = 6	+ / - / + N = 6	+ / + / - N = 6	+ / + / + N = 6

Total # ind. = 48

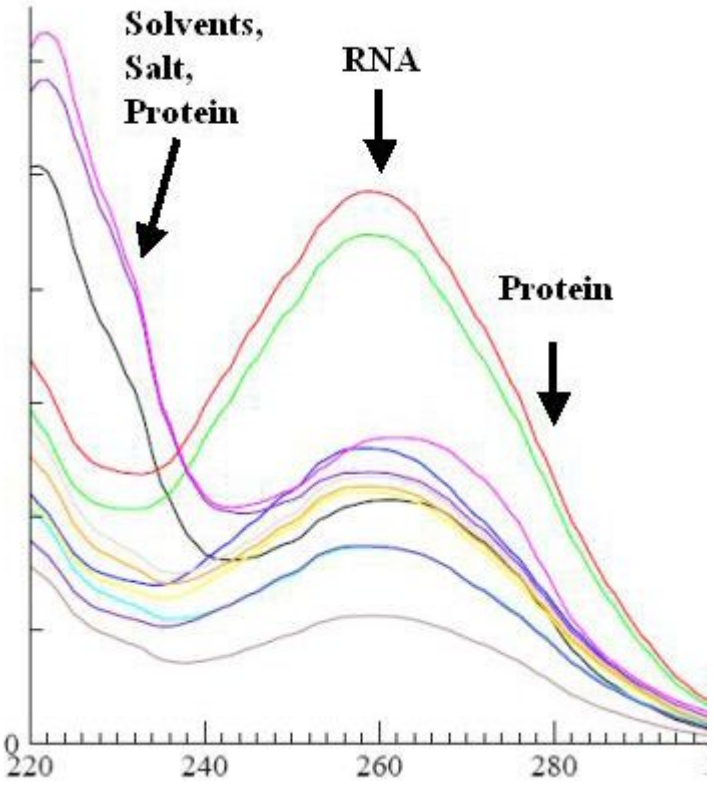
any individual receives one of multiple possible treatments at each factor.

Step 2: RNA preparation

- The success of RNA-seq experiments is highly dependent upon recovering pure and intact RNA.
  - Isolate and purify RNA
  - rRNA removal (Target enrichment):
    - Total RNA consists of >80% ribosomal RNA (rRNA). Remove rRNA
  - RNA fragmentation
    - ~40-400bp depending upon the platform

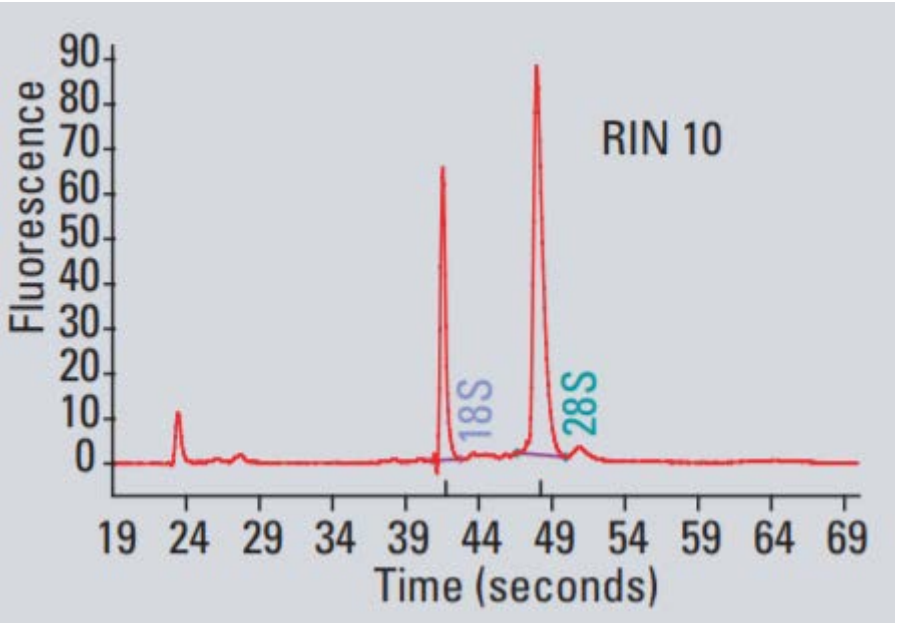
RNA purity

- RNA purity is determined by measuring the 260/280 and 260/230 ratios
- Nucleic acids have a peak absorbance at approximately 250 - 260 nm, this includes RNA, DNA, and free nucleotides
- Good purity of RNA sample
  - 260/280 ≈ 2
  - 260/230 ≈ 2.0-2.2
- Excessive absorbance at 280 indicates the presence of protein in your sample
- Excessive absorbance at 230 may indicate the presence of residual phenol in your sample



RNA Integrity

- To assess the integrity of the ribosomal subunits, a Bioanalyzer profile is run and an RNA integrity number (RIN)
- The RIN value is a measurement of the intactness of the two ribosomal bands



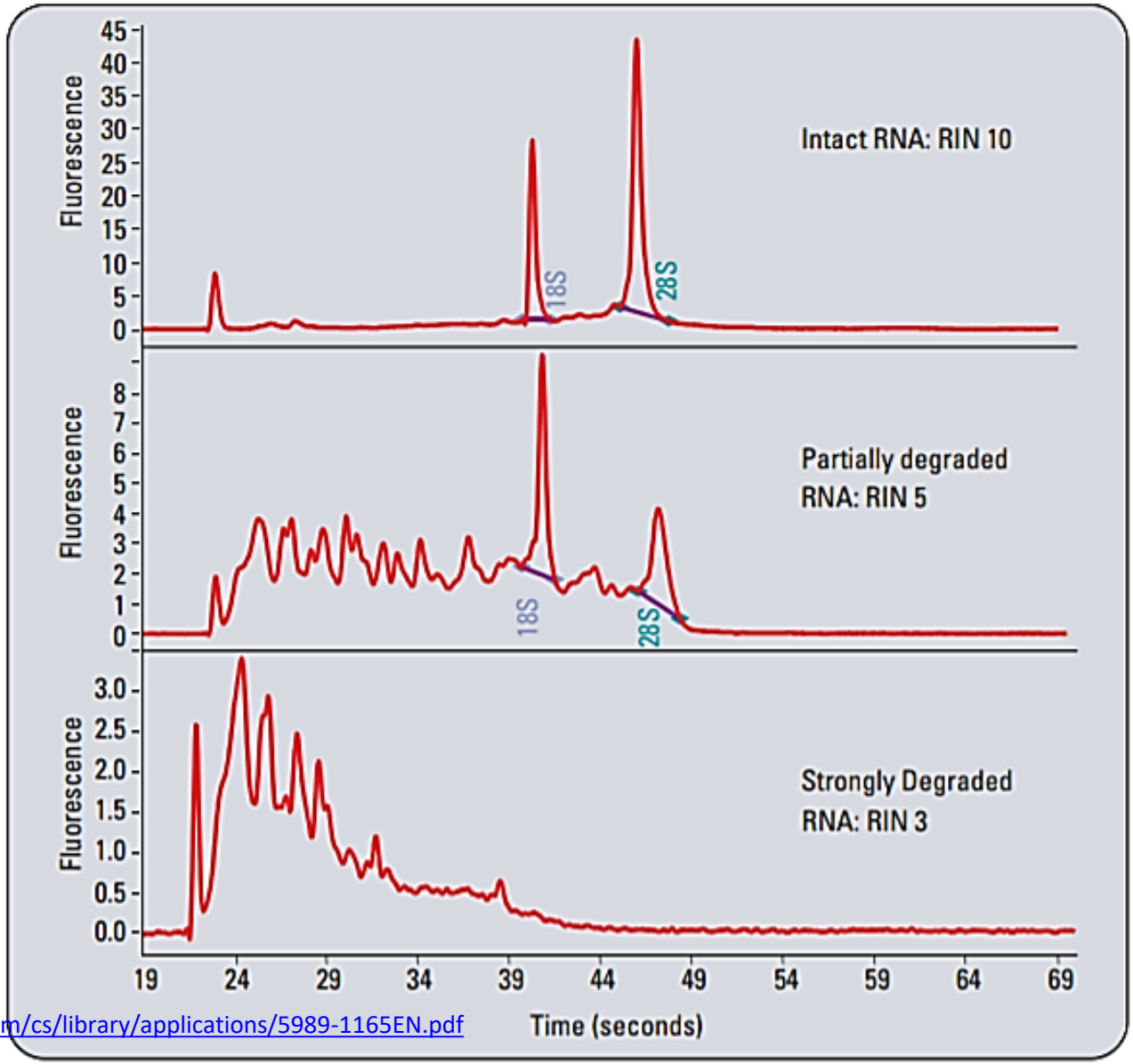
http://www.exiqon.com/ls/Documents/Scientific/Guidelines-NGS-whole-transcriptome-mRNA.pdf

RNA Integrity

- In general, for sequencing:
  - RIN 7-10: Little to no degradation; whole transcriptome or mRNA RNA sequencing possible.
  - RIN value 5-7: Partial degradation; whole transcriptome and mRNA sequencing possible. However, mRNA sequencing will be poor since mRNAs will start to lose their 3´ poly-A tails.
  - RIN value <5: Degraded RNA; partial sequencing of degraded rRNA material unavoidable

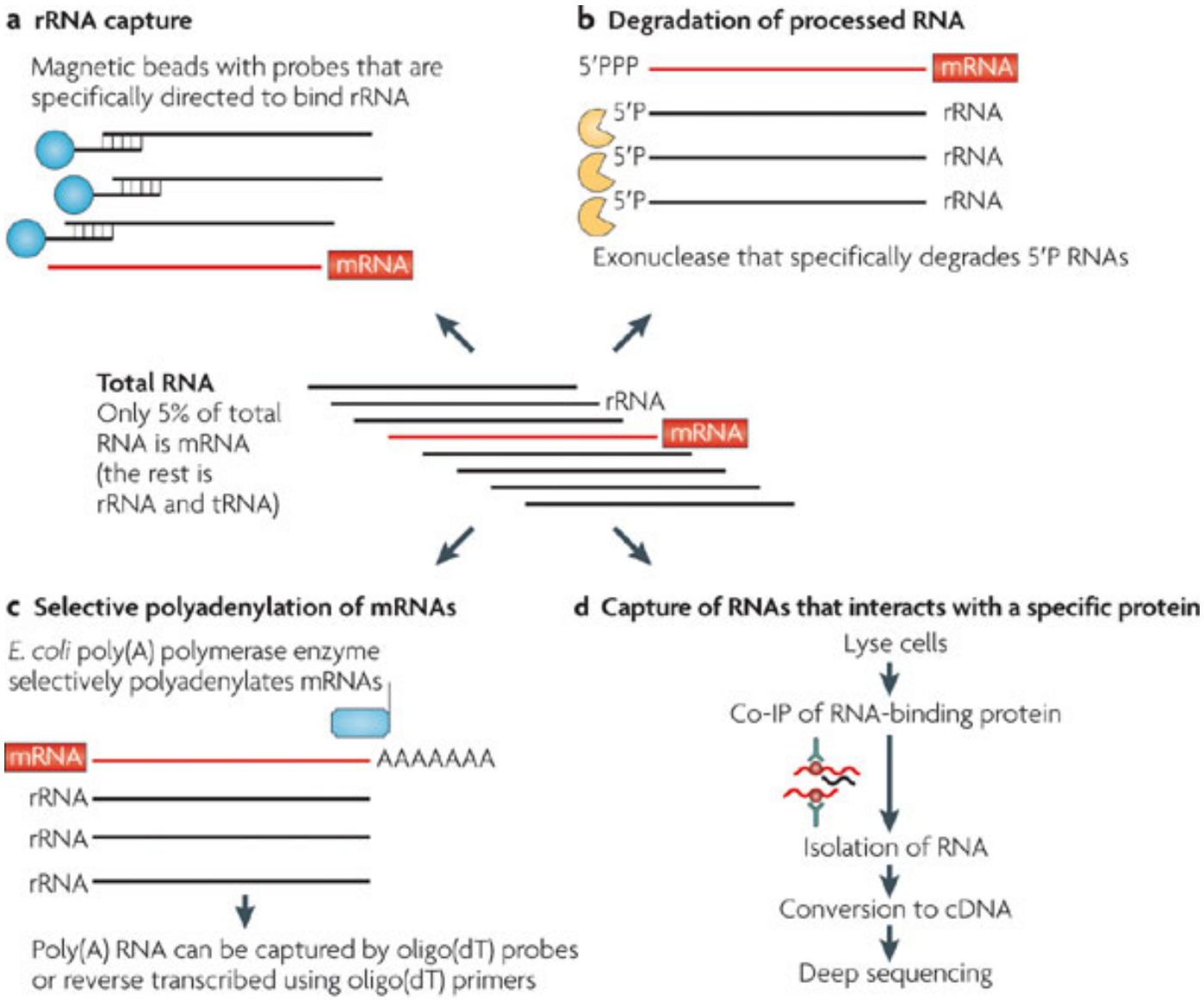
http://www.exiqon.com/ls/Documents/Scientific/Guidelines-NGS-whole-transcriptome-mRNA.pdf

RNA Integrity



https://www.agilent.com/cs/library/applications/5989-1165EN.pdf

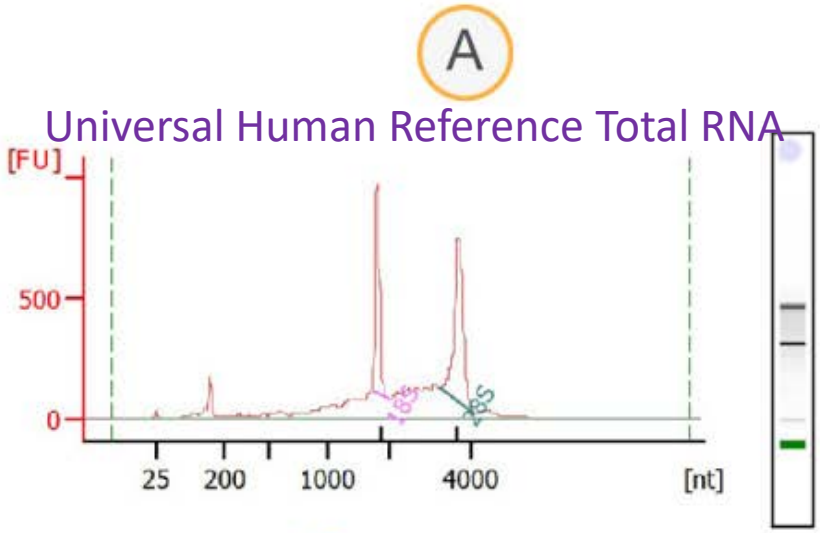
rRNA removal (Target enrichment)



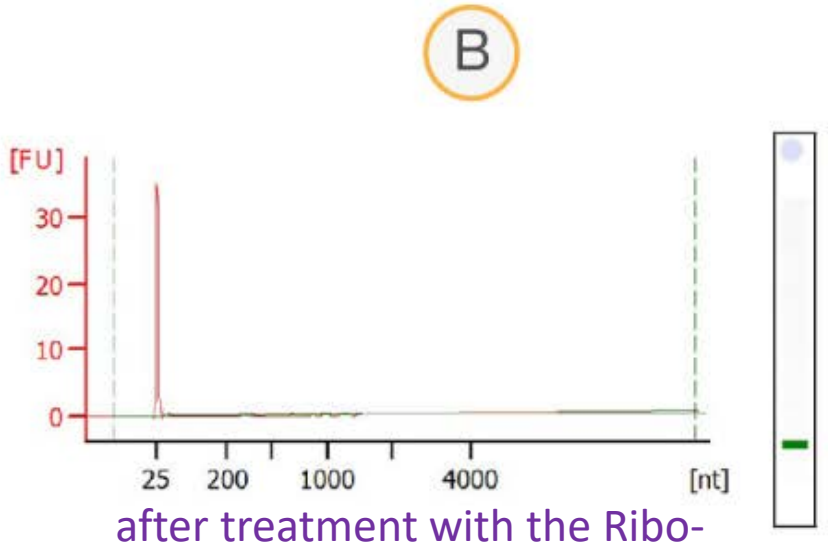
Nature Reviews Genetics 11, 9-16 | doi:10.1038/nrg2695

Nature Reviews | Genetics

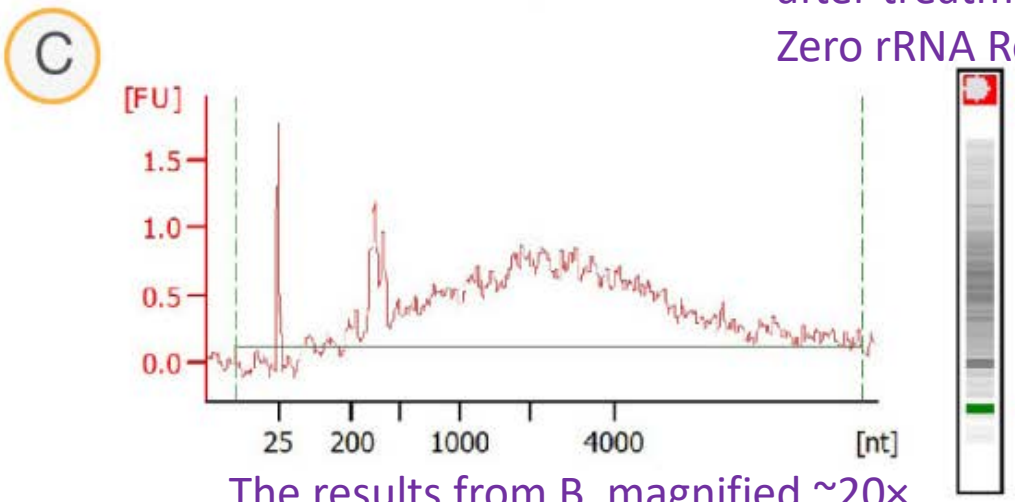
rRNA removal: Assess Quality



Universal Human Reference Total RNA



after treatment with the Ribo-Zero rRNA Removal Kit

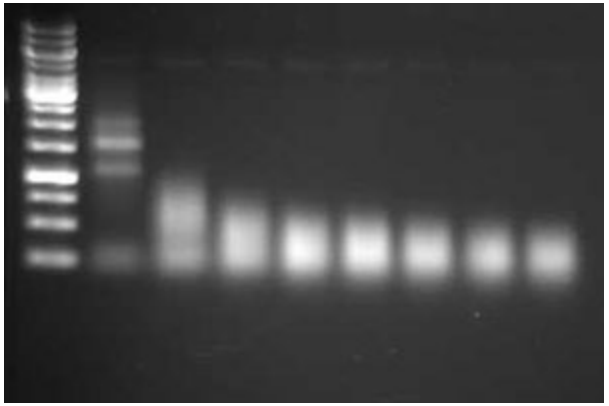


The results from B, magnified ~20×

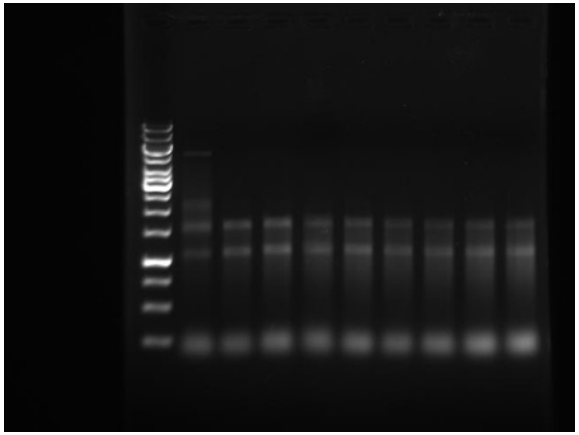
https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/ribosomal-depletion/ribo-zero/ribo-zero-reference-guide-15066012-02.pdf

RNA fragmentation

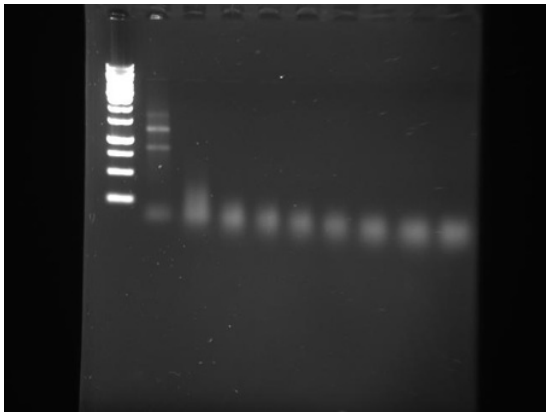
- fragmentation of the RNA (most)
  - enzymatic, metal ion, heat, and sonication
- fragmentation of the DNA
  - mechanical (*e.g.*nebulization or sonication) or enzymatic methods



Gel analysis of 0, 2, 4, 6, 5, 8, 9, and 10 minutes RNA fragmentation. Fragmentation begins at 2 minutes and fragments converge to between 500 and 250 bp.



This RNA is not fragmented at all (across time)

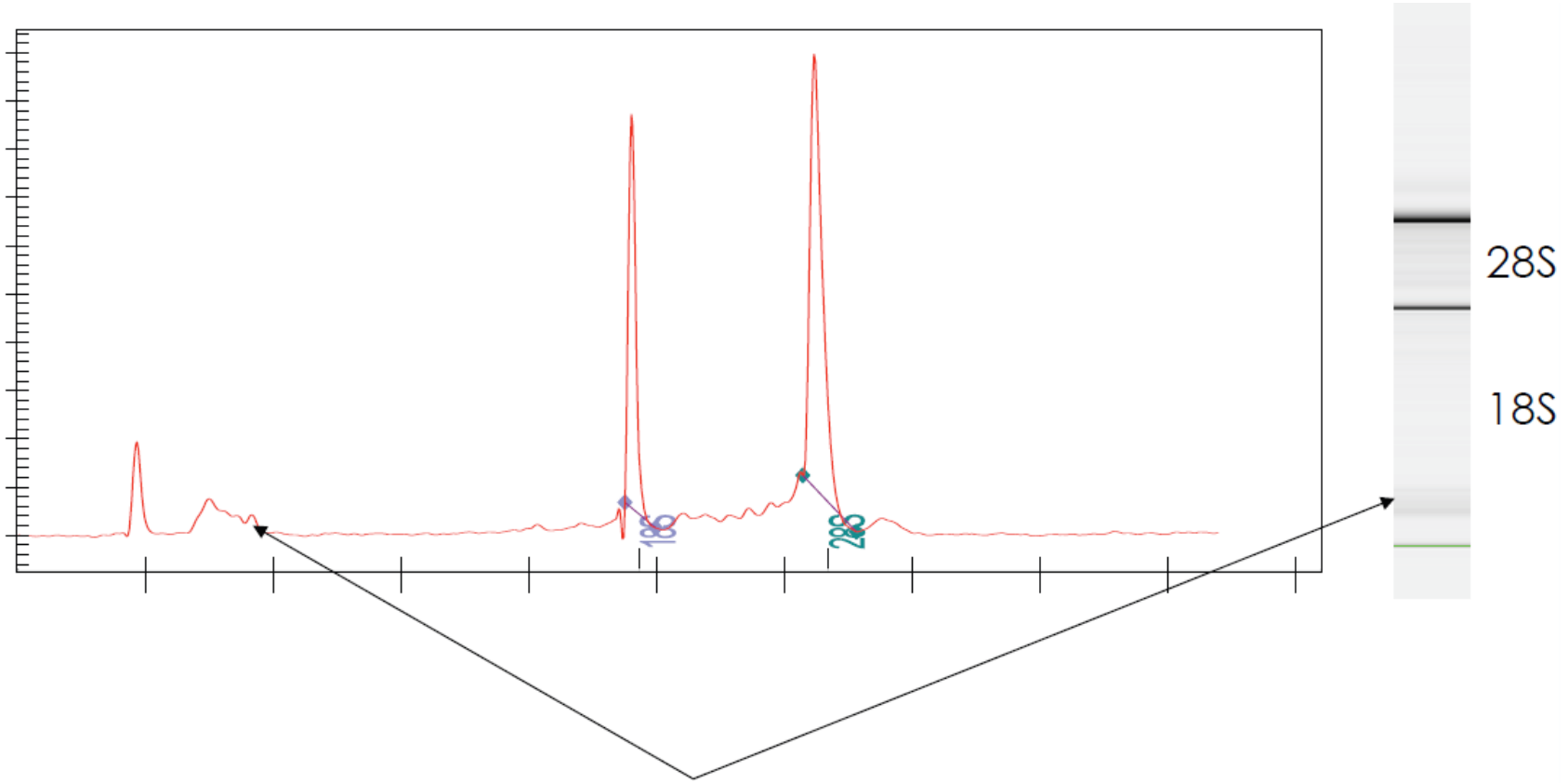


An example of over fragmentation

an **optimal** result of **RNA fragmentation**

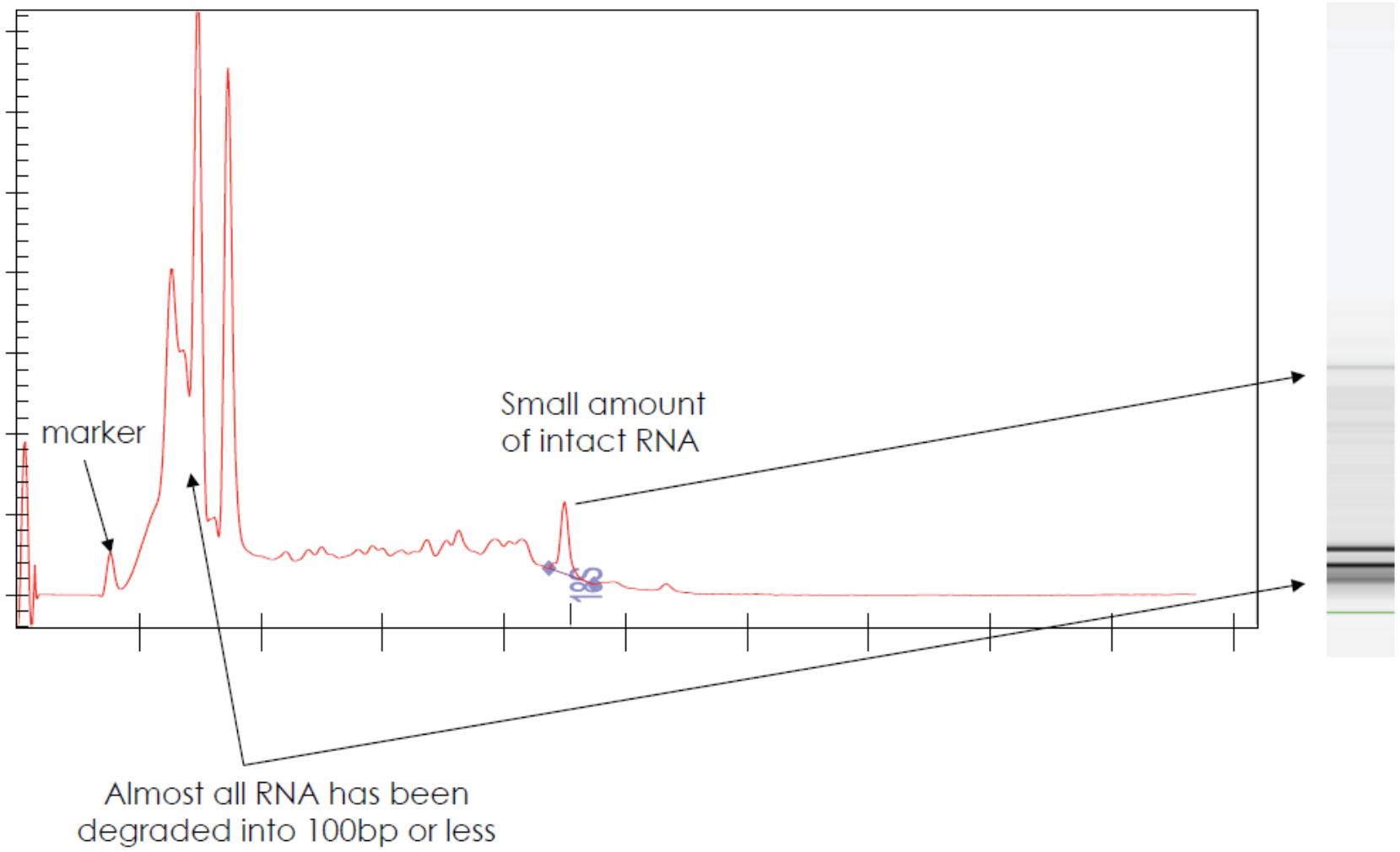
http://fg.cns.utexas.edu/fg/protocol__RNA-seq_Library_Preparation.html

Intact Total RNA: Bioanalyzer results

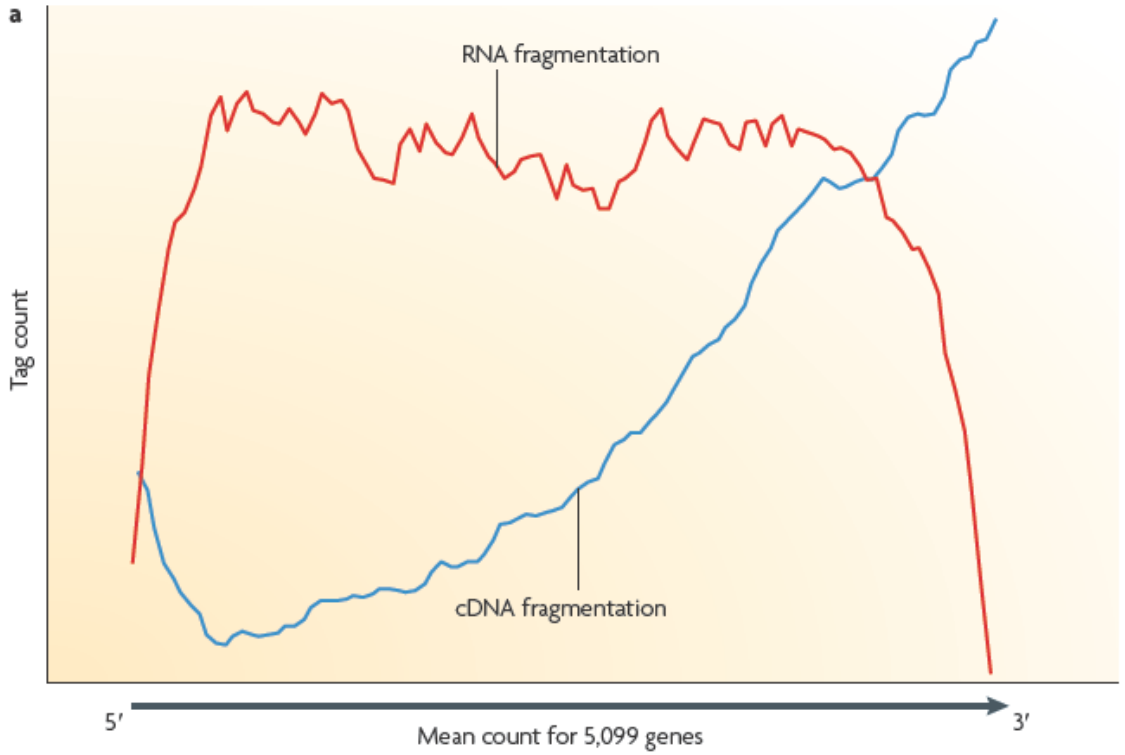


Small peaks: 5S and 5.8S subunits, tRNAs, and small RNA fragments about 100bp.

Completely Digested RNA: Bioanalyzer results



RNA fragmentation

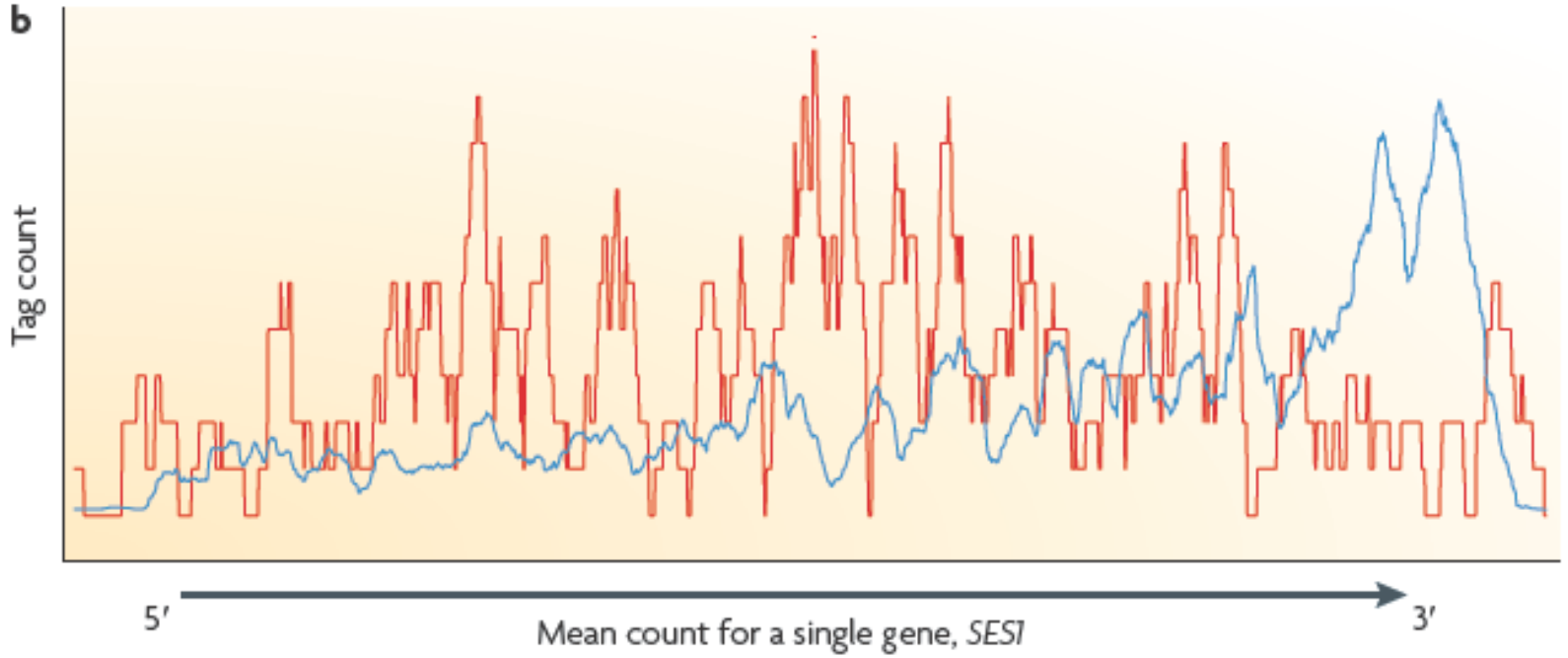


Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3′ end of the transcript

RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5′ and 3′ ends

Nat Rev Genet. 2009 Jan;10(1):57-63.

RNA fragmentation



A specific yeast gene, *SES1*

SAINT LOUIS UNIVERSITY.

Nat Rev Genet. 2009 Jan;10(1):57-63.

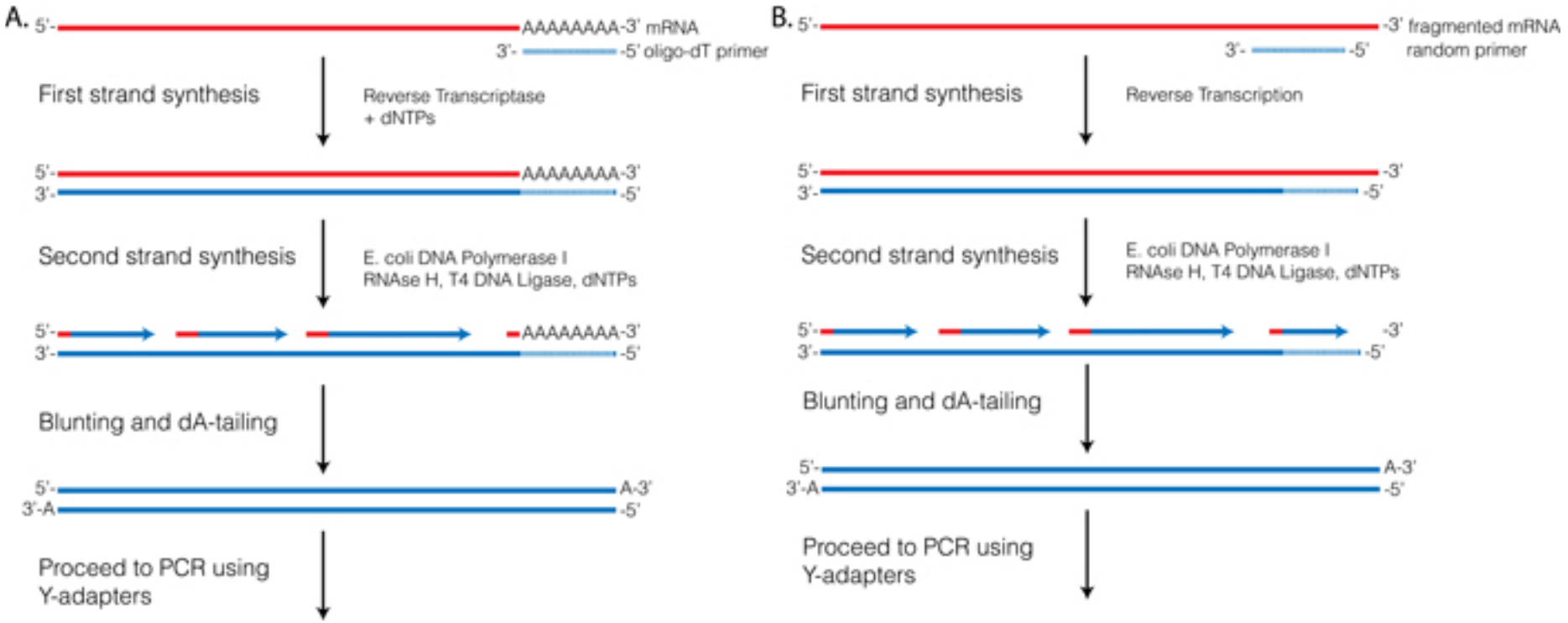
Step 3: Library Preparation

- Currently available sequencing technologies require a DNA template with platform-specific "adaptor" sequences at either end of each molecule.
- process of “library preparation”
  - Generating the cDNA
  - Adding the adaptors
  - Amplifying the DNA

Generating the cDNA

- First-strand synthesis
- Second-strand synthesis
- Optional: Fragmentation of cDNA

Generating the cDNA



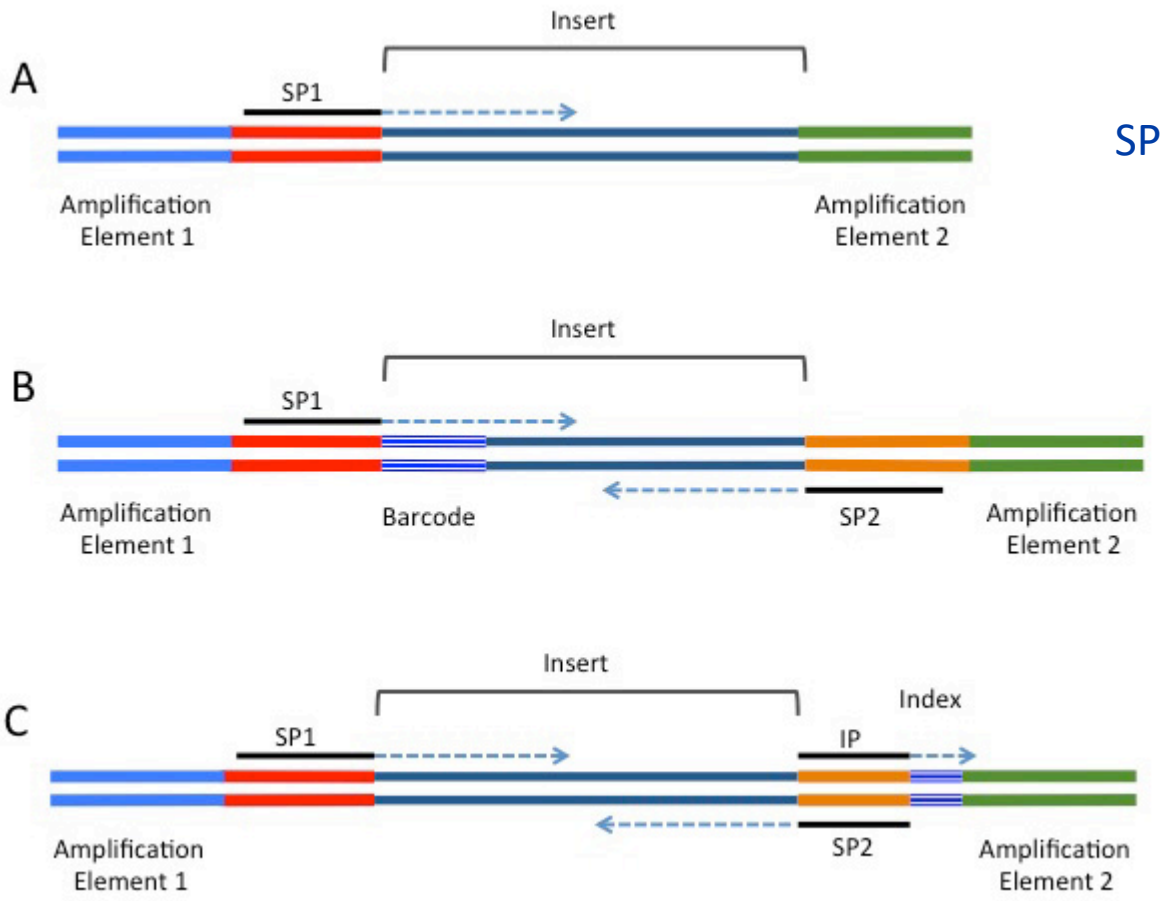
Schematic overview of cDNA synthesis using (A) oligo-dT priming or (B) random oligo priming for first-strand synthesis. In both cases second-strand synthesis is via-RNA priming/displacement.

https://rnaseq.uoregon.edu/

Adding the adaptors

- Adapter sequences must be present at the ends of the fragments.
- Two types of sequence elements in adapters are required
  - (1) sequences for clonal amplification and attachment to the sequencing support
  - (2) primer for sequences for priming the sequencing reaction.
- Optional elements
  - barcodes or indices for multiplexing
  - a second primer for paired-end sequencing

Adding the adaptors



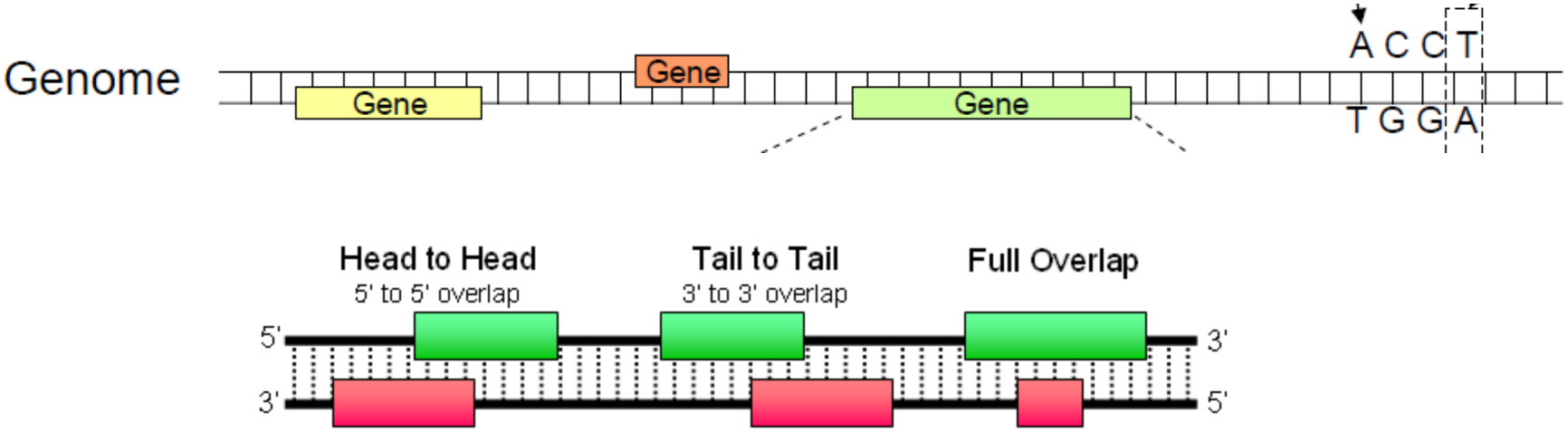
SP1 (primary sequencing primer)

SP2 (paired-end primer)

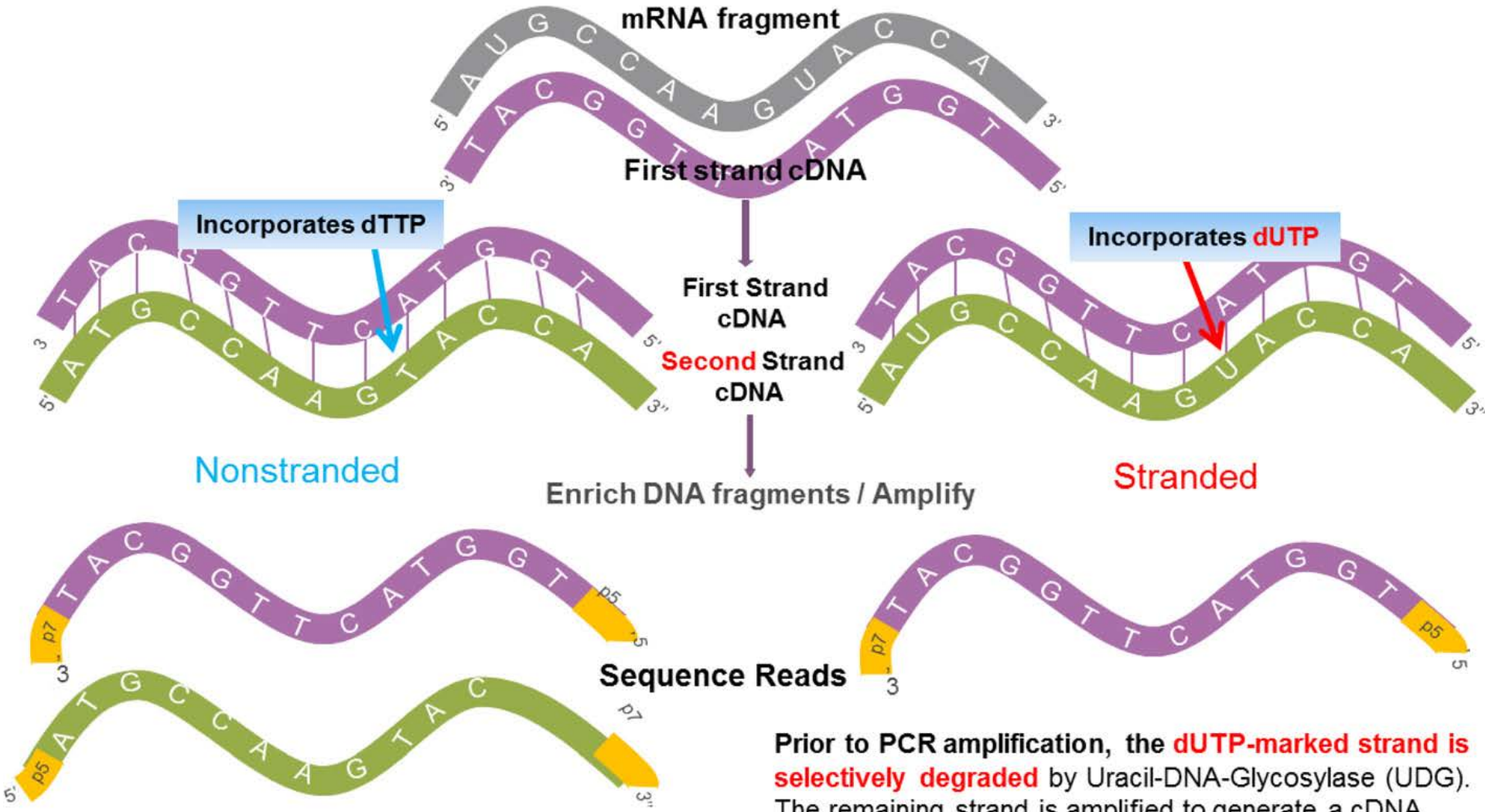
http://rnaseq.uoregon.edu/#library-prep-sequencing-adapters

Preparation of stranded libraries

- Gene coding sequence can be found on both strands
- Conventional RNA-seq library protocols do not keep strand information
- Need to distinguish overlapping genes, or overlapping anti-sense transcripts



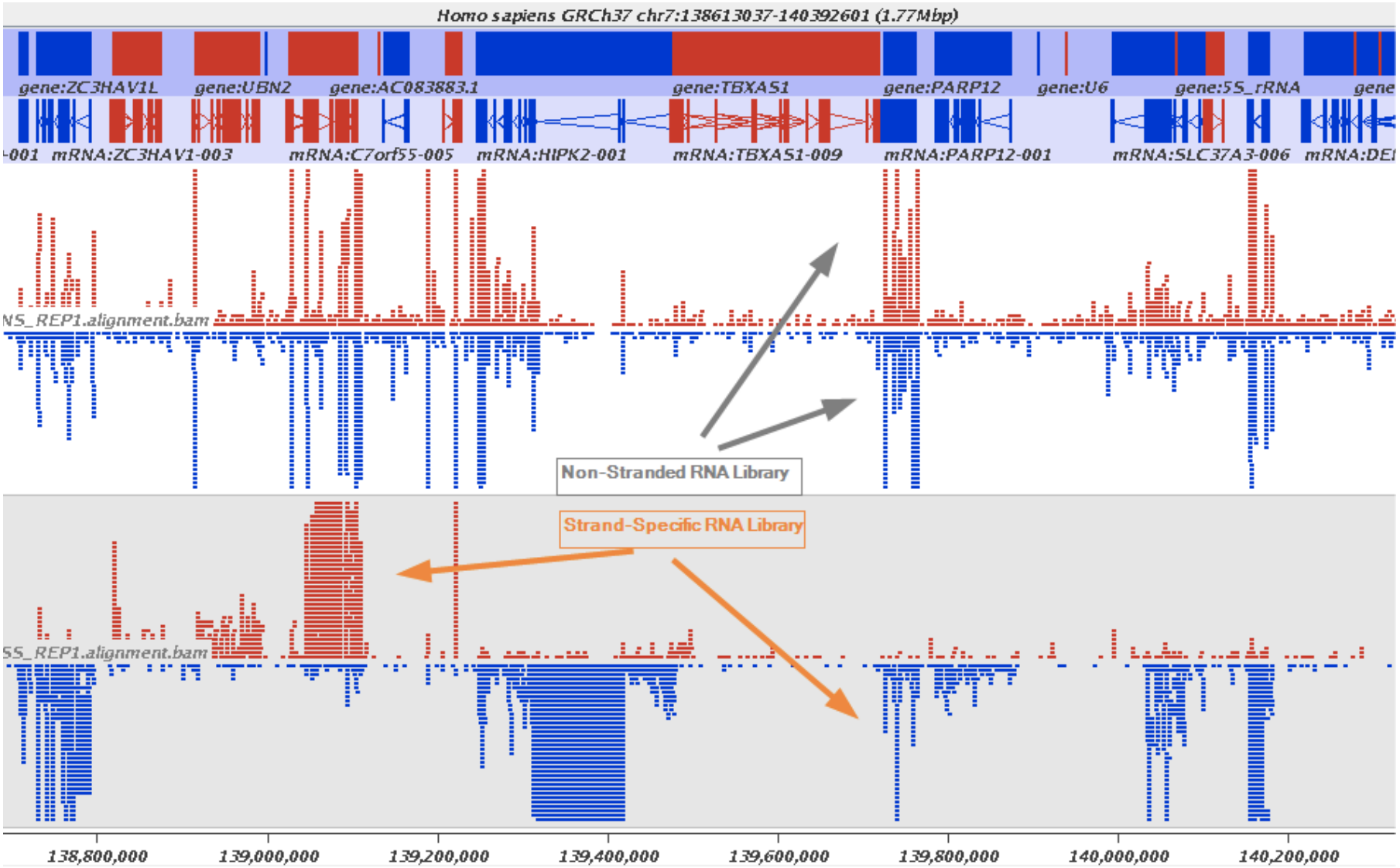
Non-stranded versus stranded RNA-seq protocol



BMC Genomics. 2015; 16(1): 675.

SAINT LOUIS UNIVERSITY.

Stranded RNAseq data look like this



Red transcripts are from + strand and blue are from - strand. In stranded example reads are clearly stratified between the two strands. A small number of reads from opposite strand may represent anti-sense transcription.

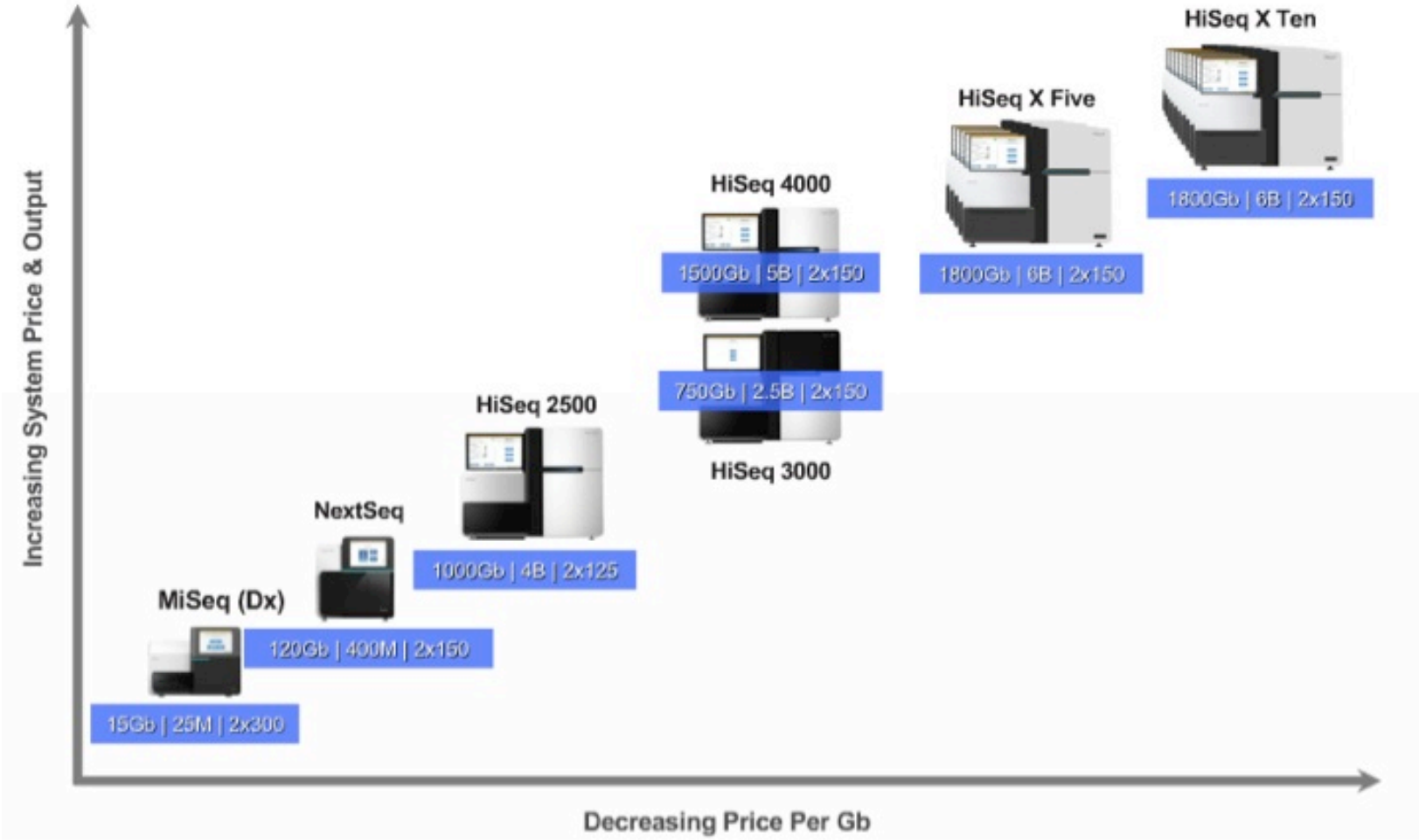
https://galaxyproject.org/tutorials/rb_rnaseq/

Step 4: Sequencing

- Platforms:
  - Illumina, Ion Torrent, and PacBio, Nanopore, etc
  - Each is based upon different proprietary chemistries and technologies and each has unique strengths and weaknesses
- The current leading platform for RNA-seq is Illumina

M.L. Metzker M.L. (2009) Sequencing technologies - the next generation, Nature reviews Genetics. 11, 31-46.

Sequencing Power For Every Scale.



Technical specifications of Sequencing platforms

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	$128 K	$80 K**	$695 K	$256 K	$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-341

Detailed comparisons of the available platforms

- M.L. Metzker M.L. (2009) Sequencing technologies - the next generation, Nature reviews Genetics. 11, 31-46.
- Liu L., Li Y., Li S., et al. (2012) Comparison of next-generation sequencing systems, *Journal of biomedicine & biotechnology.* 2012, 251364.
- Glenn T.C. (2011) Field guide to next-generation DNA sequencers, *Molecular ecology resources.* 11, 759-769.

Sequence Data Format: FASTQ

- An example

```
@SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
NGAGAATAACAACAATCTCAAATGAATTCATGCAGAAAAATATTAAAAGGCAAAAACACTATATCATCCATGGATGAACACTTTTCACAAAACAATCATT
+SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
#1=DDDEDHFFHGJJJJJJJJIGJJJIJJJJGIJHIJIJIJJJJJJIJIJJFJJIIJJJJJJJGGIJIJIHIIGFHFFHEFFFFFECCEEDDDDDDDDD
@SRR647683.2 HWI-ST444:154:D0M2UACXX:3:1101:1507:2144 length=100
NGCATAACTTTCTAACAAGTTACAGGGTAAAATATTGGTCTGAGTGCCCCAAGAGTGGCTGCTCTTAAAAGTTAAACCTTCCAAAACCAAATTCTTTTTC
+SRR647683.2 HWI-ST444:154:D0M2UACXX:3:1101:1507:2144 length=100
#1:BDADA>F>CDFGHIIBHHHCHHGEEF@@CFGIIGHHGGEHE<?DFHBDDGII??DFFHC.).887=@)=C77))7=?);.7?;>??BB(((-;@CA#
@SRR647683.3 HWI-ST444:154:D0M2UACXX:3:1101:1587:2164 length=100
NCTACTCCTACAACTGCTGCTCGAAAAAAGAAAGACTTGAAGATATTTAGAAACGTTGATAGTAATCTTGCTAATCTAATTCTGAATGAAATTGTTGACA
+SRR647683.3 HWI-ST444:154:D0M2UACXX:3:1101:1587:2164 length=100
#1=DDFFFHHHHHIGHJIJIGIJJGJJJJJIJJJIIJJJJJI@HIJIJIJIJJIIFHIEHIJHHHHHHGHEFFFFFCCEEEEEDEDDDDDCCCCCDDCCC
@SRR647683.4 HWI-ST444:154:D0M2UACXX:3:1101:1648:2172 length=100
NGTATGCTGGCCCTGAGGTGGACATCTGGAGTTGTGGTGTTATTCTCTATGCTCTTCTCTGCGGGACGCTGCCATTCGATGATGAACATGTTCCCACGCT
+SRR647683.4 HWI-ST444:154:D0M2UACXX:3:1101:1648:2172 length=100
#11=ADDDDAAD>AEI@E+AE>EFEIIIIEEEEI@DDDDDDCEDDDIEICEICEEIC<CCEIIAA@@@????AAA>BDABAAAADAAAAAD#########
@SRR647683.5 HWI-ST444:154:D0M2UACXX:3:1101:1709:2175 length=100
NAGGCTGCAAAGAGGAAGCTTTGCTAGATCAGCCCATCTCCATGCCATCTTGGCCATGCTTTTCCCCCCCAACAAATGCCTCACACTGTTCAAGCTTCACT
+SRR647683.5 HWI-ST444:154:D0M2UACXX:3:1101:1709:2175 length=100
#11ADDDFBFDHFIGGIGHGDEGGGIIGIIGDGIIICHIIHBEB@GHGIIGGDBDHGGEIIIIIGEABBDFCCCCCACCCCCCCCCCCC>A>@CCCCACC
```

Sequence Data Format: FASTQ

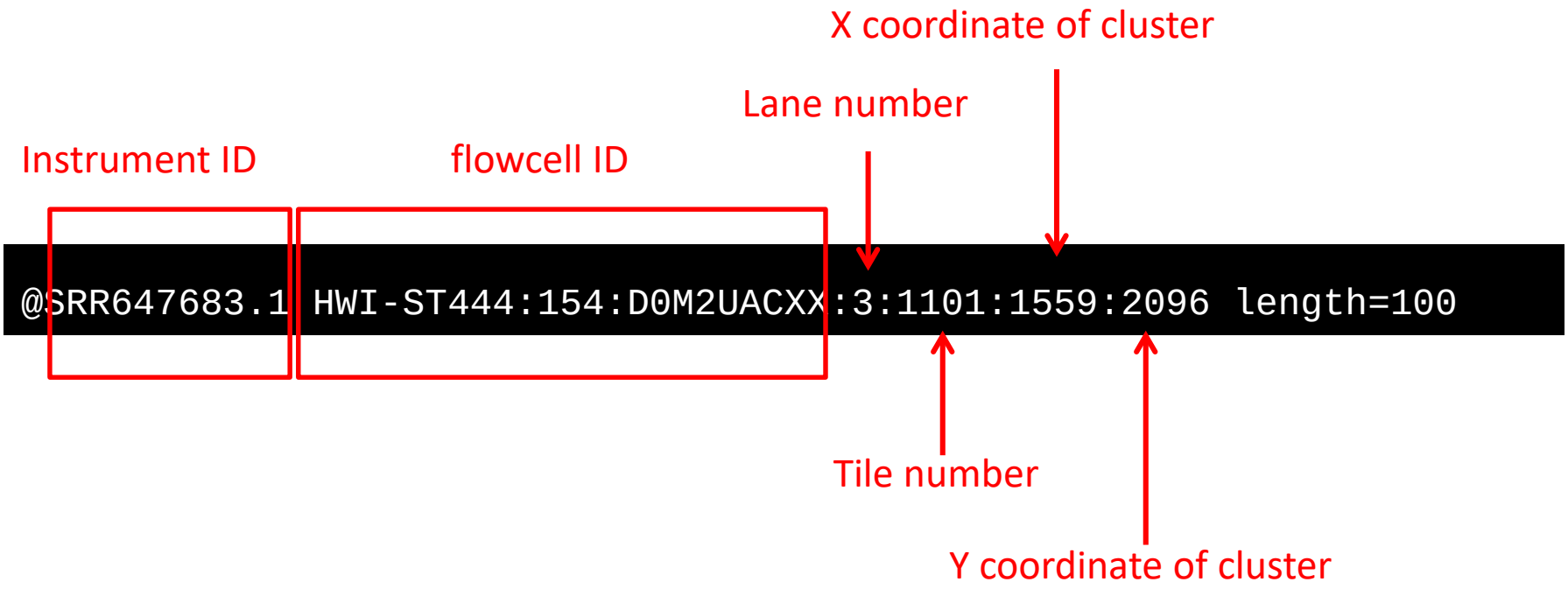
Each entry in a FASTQ file consists of four lines:

- Sequence identifier
- Sequence
- Quality score identifier line (consisting of a +)
- Quality score

An example of a valid entry is as follows; note the space preceding the read number element:

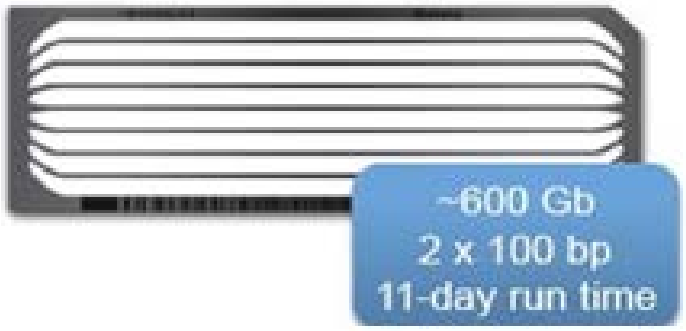
```
@SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
NGAGAATAACAACAATCTCAAATGAATTCATGCAGAAAAATATTAAAAGGCAAAAACACTATATCATCCATGGATGAACACTTTTCACAAAACAATCATT
+SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
#1=DDDEDHFFHGJJJJJJJJIGJJJJIJJJJGIJHIJIJIJJJJJJIJIJJFJJIIJJJJJJJGGIJIJIHIIGFHFFHEFFFFFECCEEDDDDDDDDD
```

Sequence identifier: Illumina



https://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm

Illumina flowcell



high output mode

each flow cell has eight lanes

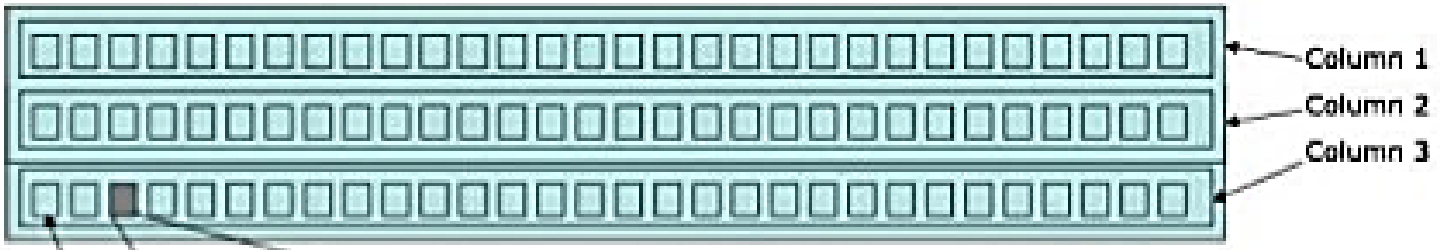
rapid run mode

each flow cell has two lanes

Tile

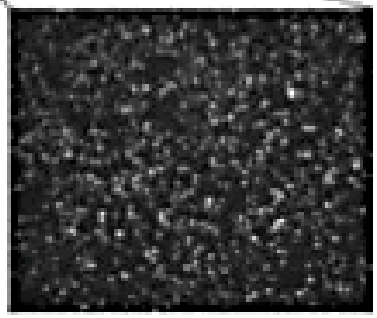
A **flow cell** contains eight lanes

Each **lane/channel** contains **three columns** of tiles



Each **column** contains **100 tiles**

Tile



20K-30K Clusters

350 X 350 µm

Each tile is imaged four times per cycle – one image per base.

345,600 images for a 36-cycle run

Quality score

Table 1 ASCII Characters Encoding Q-scores 0–40

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

PHRED quality score (Q_{PHRED}) of a base call, defined in terms of the estimated probability of error (P_e):

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

Q_{PHRED}	P_e	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

```
@SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
NGAGAATAACAACAATCTCAAATGAATTCATGCAGAAAAATATTAAAAGGCAAAAACACTATATCATCCATGGATGAACACTTTTCACAAAACAATCATT
+SRR647683.1 HWI-ST444:154:D0M2UACXX:3:1101:1559:2096 length=100
#1=DDDEDHFFHGJJJJJJJJIGJJJJIJJJJGIJHIJIJIJJJJJJIJIJJFJJIIJJJJJJJGGIJIJIHIIGFHFFHEFFFFFECCEEDDDDDDDDD
```

Step 5: RNA-seq data analysis

- Two scenarios
  - Reference genome sequence available
  - No Reference genome sequence available
    - De novo assembly of the reads (Trinity)
    - Map the reads to the assembly (RSEM mapper, Bowtie)
    - Extract count table

Scenario 1

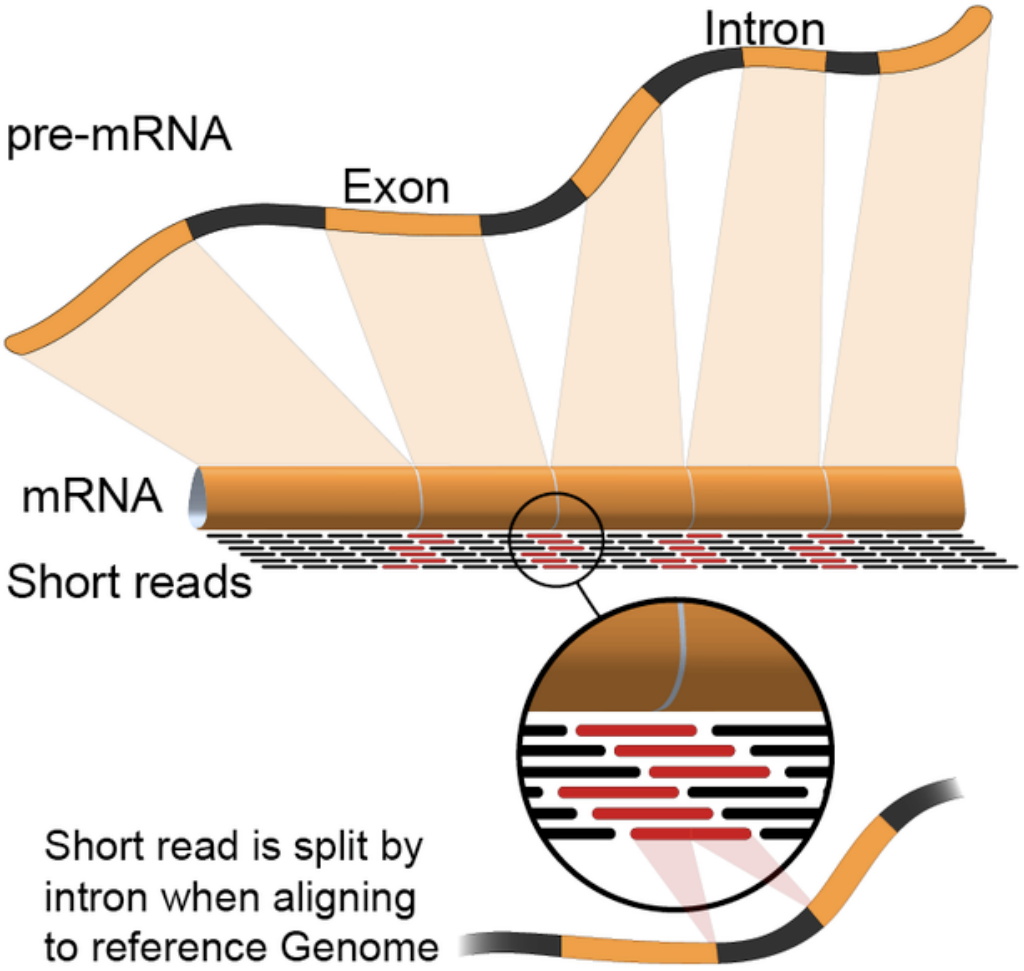
- With reference genome or transcriptome
  - Step 1: Read alignment:
    - align reads directly to a reference genome (spliced) or transcriptome (unpliced)
  - Step 2: Transcriptome reconstruction
    - identify expressed genes and isoforms
  - Step 3: Expression qualification and differentiation expression analysis
    - Expression quantification
      - Gene quantification
      - Isoform quantification
    - Differential expression
  - Step 4: Functional Interpretation:

Scenario 2

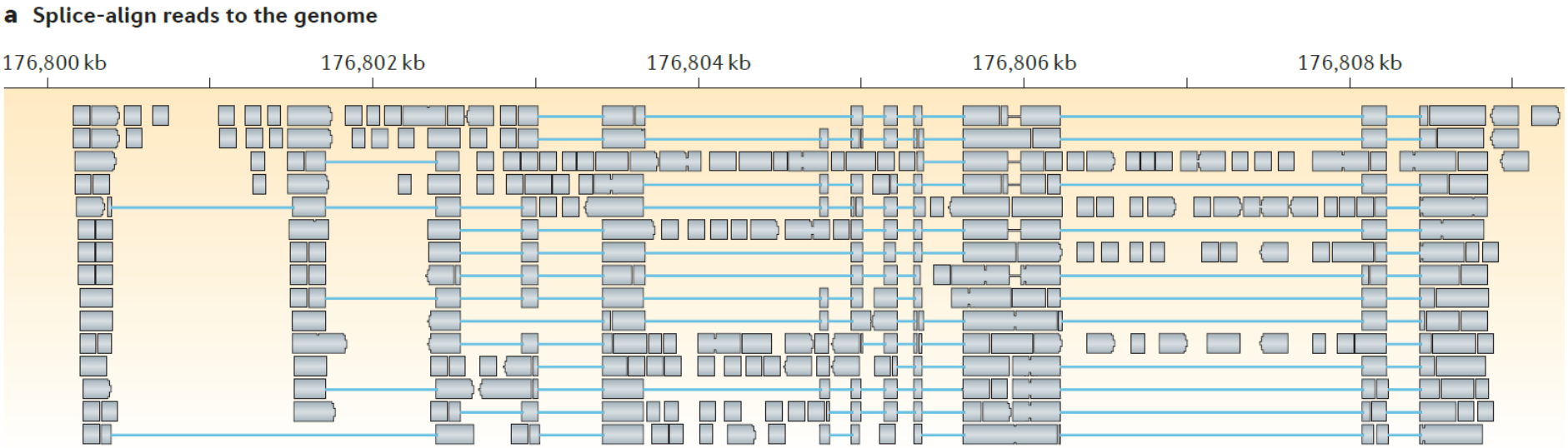
- Without reference genome or transcriptome
  - Step 1: De novo transcriptome assembly
    - align reads directly to a reference transcriptome or genome
  - Step 2: Map reads back to assembled transcripts
  - Step 3: Expression qualification and differentiation expression analysis
    - Expression quantification
      - Gene quantification
      - Isoform quantification
    - Differential expression
  - Step 4: Functional Interpretation:

Scenario 1: mapping to reference

- Potential issues by mapping to genome:
  - Exon Boundaries
  - Multiple matches



Splice-align reads to the genome



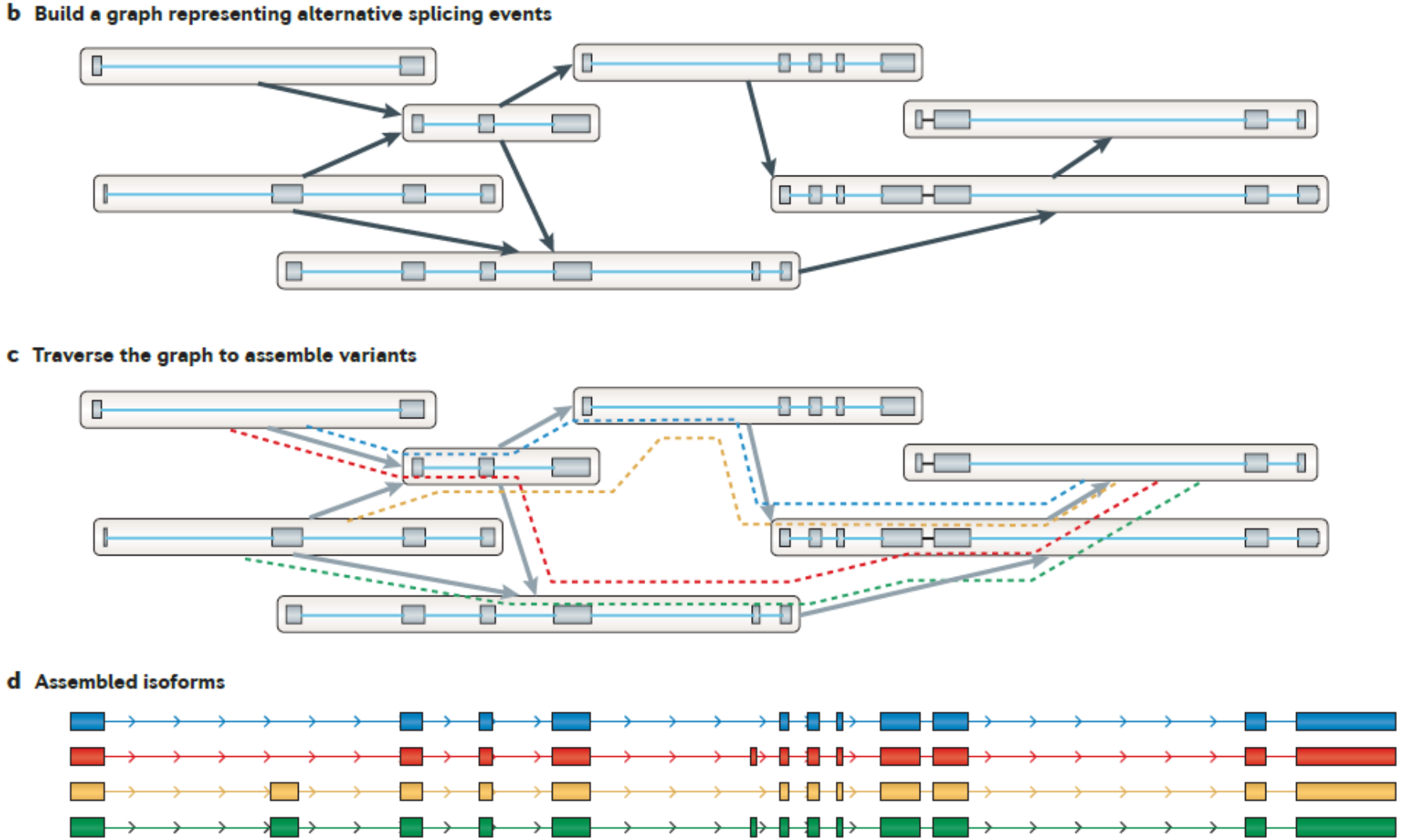
RNA-seq reads are aligned to a reference genome using a splice-aware aligner

an example of a maize gene (GRMZM2G060216)

SAINT LOUIS UNIVERSITY.

Nature Reviews Genetics **12**, 671-682

Transcripts reconstruction and isoform identification



SAINT LOUIS UNIVERSITY.

Nature Reviews Genetics **12**, 671-682

Advantage of reference-based assembly

- Fast and high sensitivity: can assemble transcripts of low abundance
- Requires less RAM
- Contamination or sequencing artefacts are not a major concern
- reads that span exon-exon junctions are aligned correctly
- it allows users to discover novel transcripts that are not present in the current annotation

Nature Reviews Genetics **12**, 671-682

Disadvantages of reference-based assembly

- depends on the quality of the reference genome being used
- Spliced reads that span large introns can be missed
- How to deal with reads that align equally well to multiple places in the genome

Nature Reviews Genetics **12**, 671-682

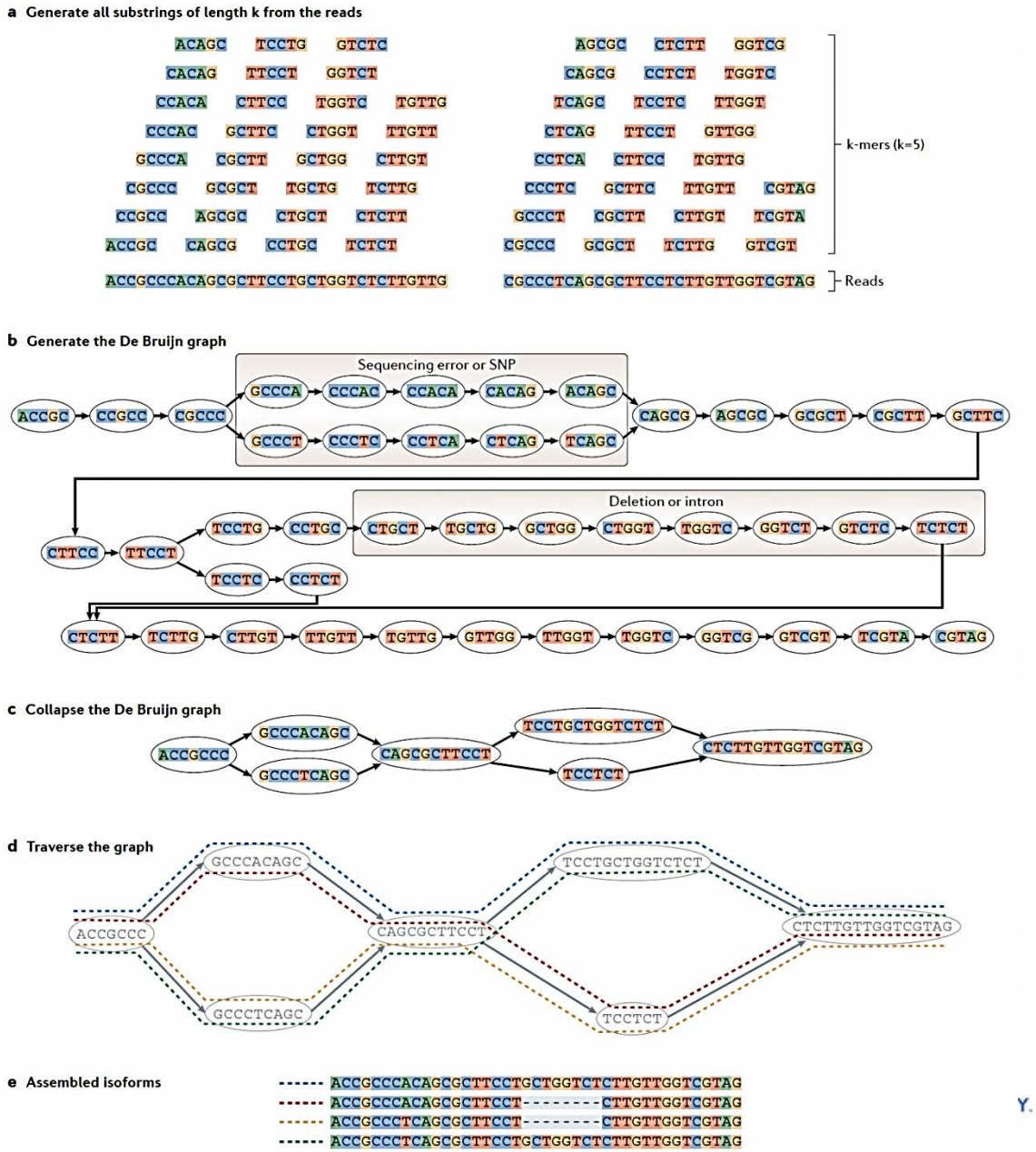
Reference-based Alignment

- Summary
  - Preferable where a high-quality reference exists.
  - Can assemble full-length transcripts at depth of 10x.
  - Can include longer reads (e.g . 454) to capture connectivity between more exons.

Scenario 2: De-novo assembly

- Doesn’t use a reference sequence.
- Finds overlaps between reads and assembles them into contigs/transcripts.
- Constructs De Bruijn graph which breaks reads into k-mers and connects overlapping nodes.

the *de novo* transcriptome assembly strategy



All substrings of length k (k-mers) are generated from each read.

De Bruijn graph created by kmers that overlap by k–1.

Single-nucleotide differences cause 'bubbles' of length k in the De Brujin graph

Insertions or deletions introduce a shorter path in the graph.

Collapse adjacent nodes.

Calculate paths through graph.

Isoforms.

Nature Reviews Genetics **12**, 671-682

De-novo Assembly

- Applications:
  - Microbes and lower eukaryotic organisms.
  - Overlapping genes from opposite strands can be detected by not allowing reverse complements in De Bruijn graph and using odd k-mers (for stranded library).
  - Higher eukaryotes more challenging due to larger datasets and difficulties in identifying alternative splice sites.

De-novo Assembly

- Advantages
  - Doesn't need a reference sequence.
  - Sometimes better than reference-based assembly when:
    - reference is of low quality (e.g. missing bits).
    - Unknown exogenous transcripts want to be detected.
    - Where long introns are expected.
  - Doesn't depend on the correct alignment of reads to splice sites.

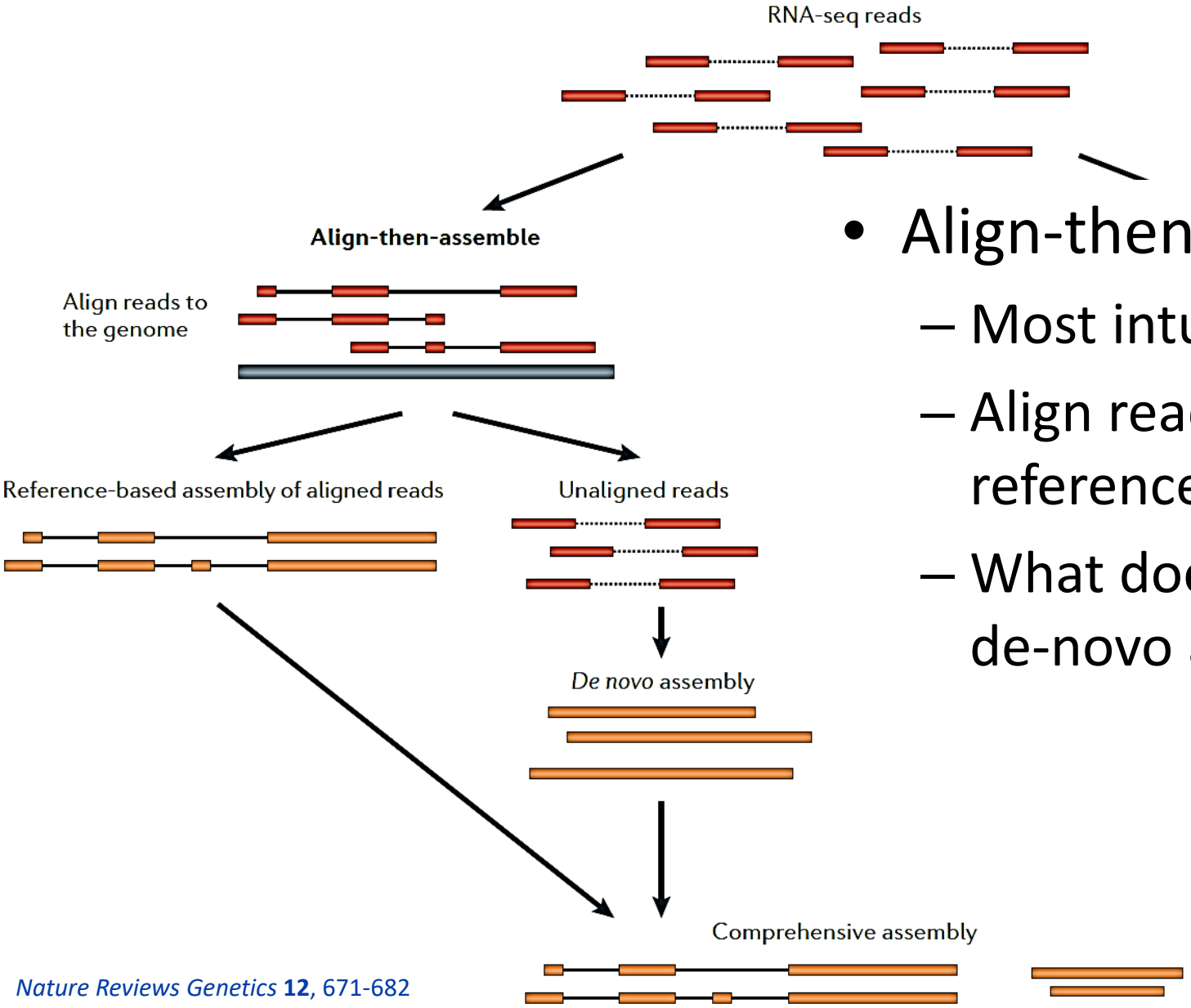
De-novo Assembly

- Disadvantages:
  - With higher eukaryotic datasets needs lots of RAM
  - Requires higher sequencing depth than reference-based assembly (30x vs.10x).
  - Highly similar transcripts are likely to be assembled into single transcripts.
  - Sensitive to read-errors. Hard to tell errors from low-abundance transcripts.

Combined strategy

- Use both de-novo assembly and reference-based alignment methods to get the best results.
- Two techniques:
  - Align-then-assemble
  - Assemble-then-align
- Make use of sensitivity of reference-based aligners and use de-novo assembly for novel sequences.

Combined strategy

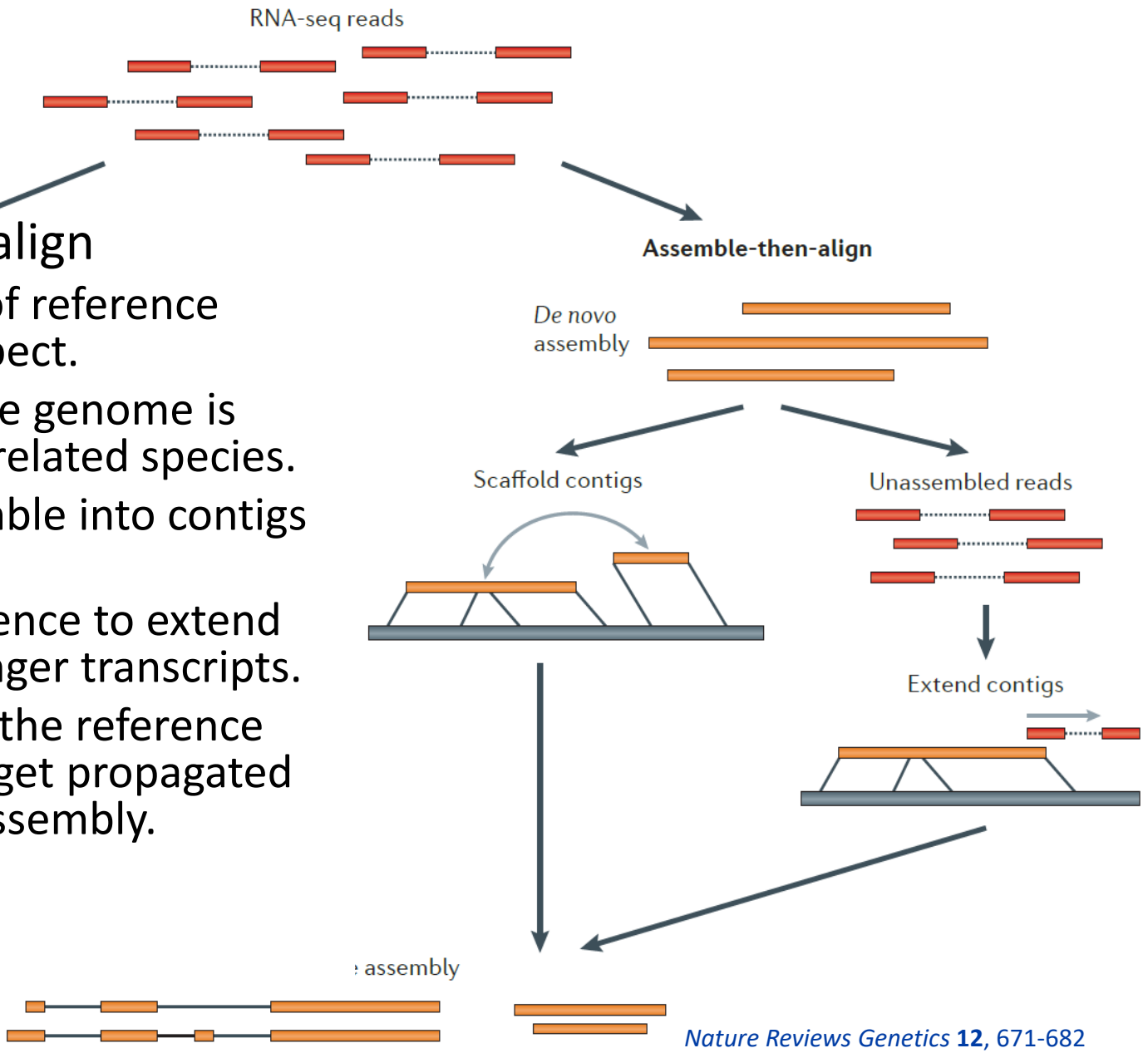


- Align-then-assemble
  - Most intuitive.
  - Align reads to a reference.
  - What doesn't align – de-novo assemble.

Nature Reviews Genetics **12**, 671-682

Combined strategy

- Assemble-then-align
  - When quality of reference genome is suspect.
  - When reference genome is from distantly related species.
  - De-novo assemble into contigs first.
  - Then use reference to extend contigs into longer transcripts.
  - Small errors in the reference genome don't get propagated into the new assembly.

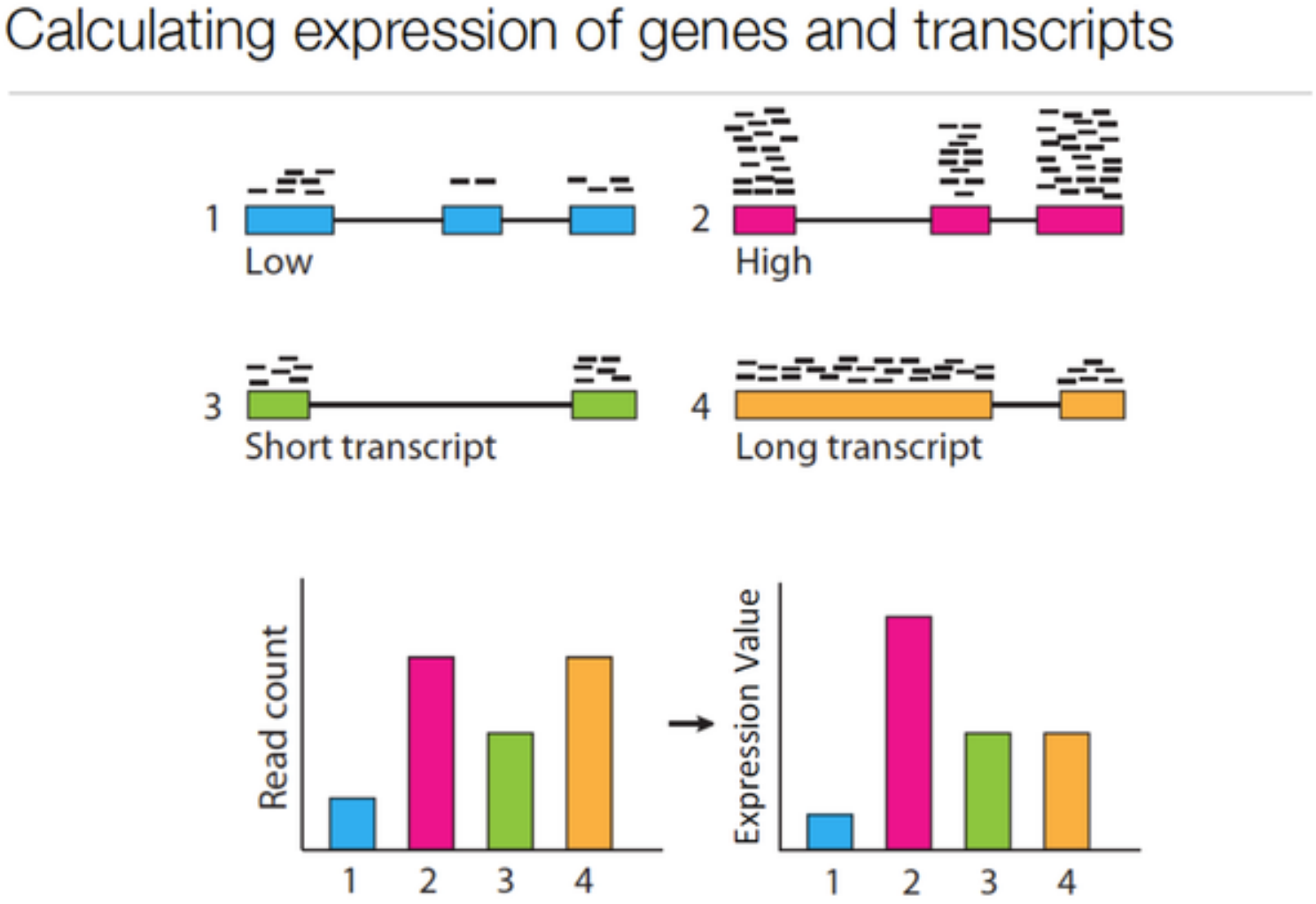


Nature Reviews Genetics **12**, 671-682

Choosing a strategy

- Factors to consider
  - Reference genome available?
    - Good quality?
    - Closely-related species?
  - Aim of project
    - Annotation
    - Identify novel transcripts
    - Expression analysis

Normalize expression value



Units of measurement

- RPKM : Reads per kilobase per million mapped reads

$$\text{RPKM} = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

  - 1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have RPKM = 1000/(1 * 8) = 125
- FPKM : for paired-end sequencing
  - A pair of reads constitute one fragment

$$\text{FPKM} = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

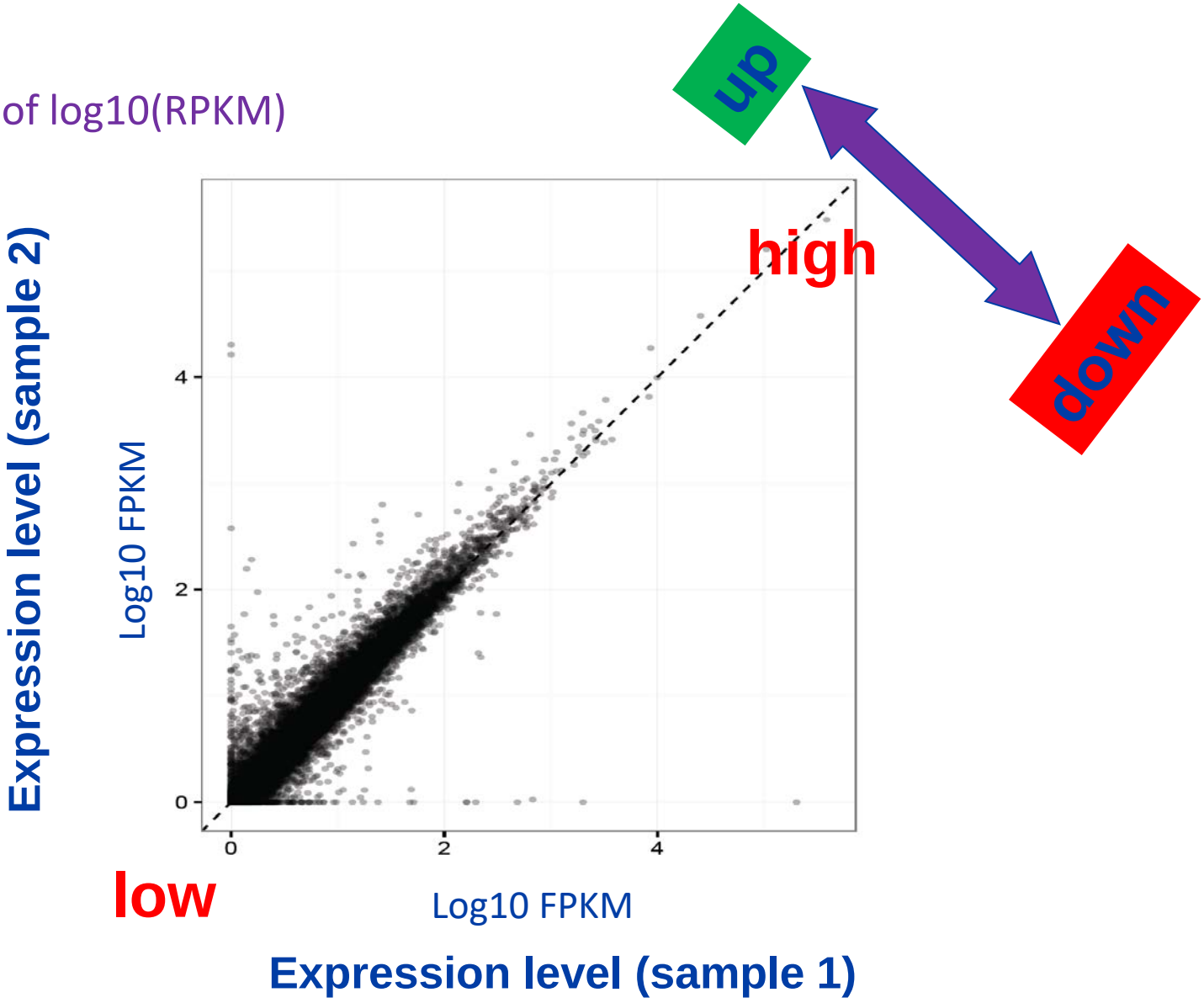
RPKM vs FPKM
What is the difference?

- They're the same thing is the reads are not paired-end.
- If the reads are paired-end
  - Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable.
  - For example, the second read is of poor quality. If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value. Thus, FPKM is calculated by counting fragments, not reads.

Use FPKM for paired-end reads

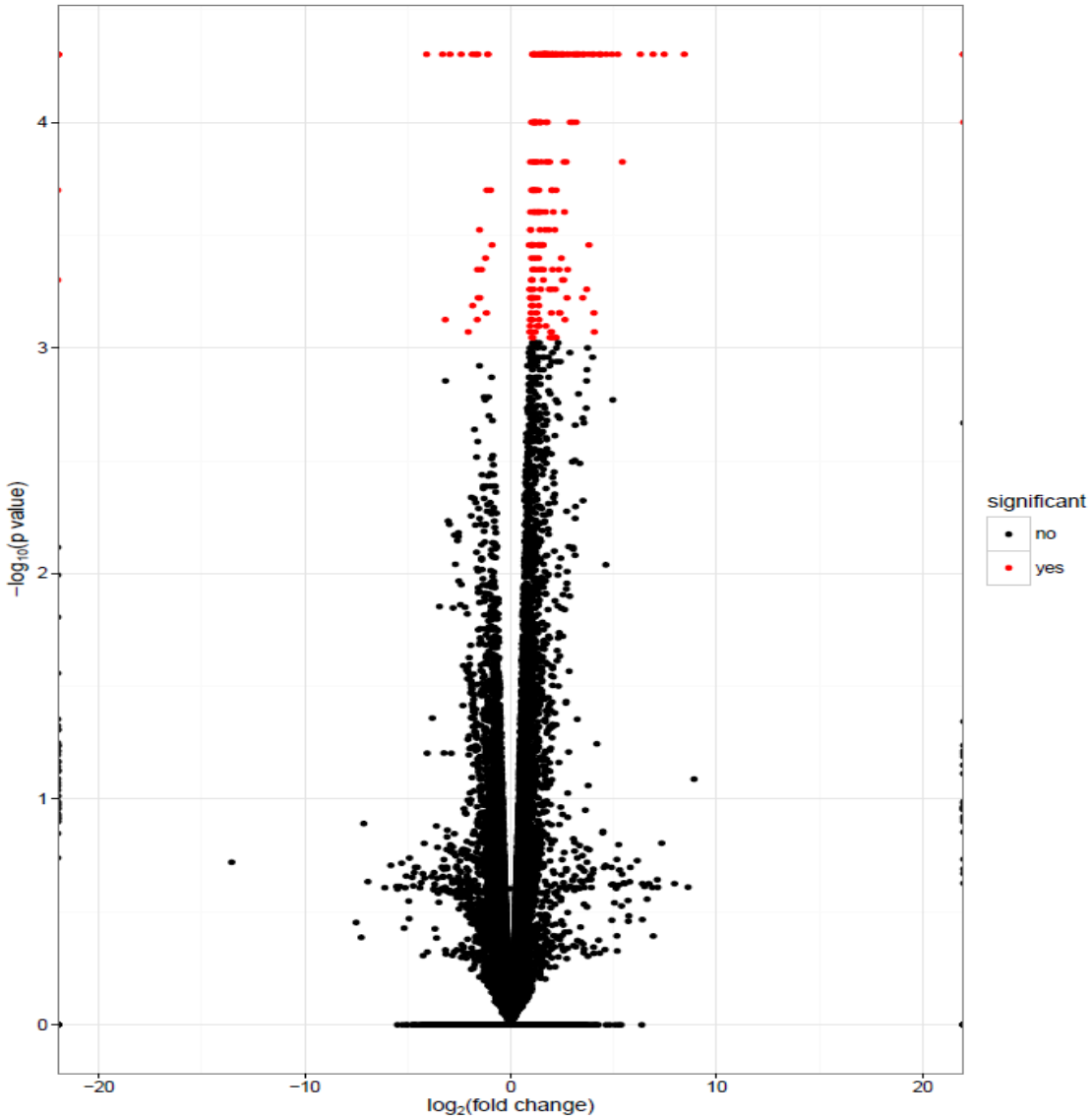
Differential expression analysis

Scatter plots of log10(RPKM)

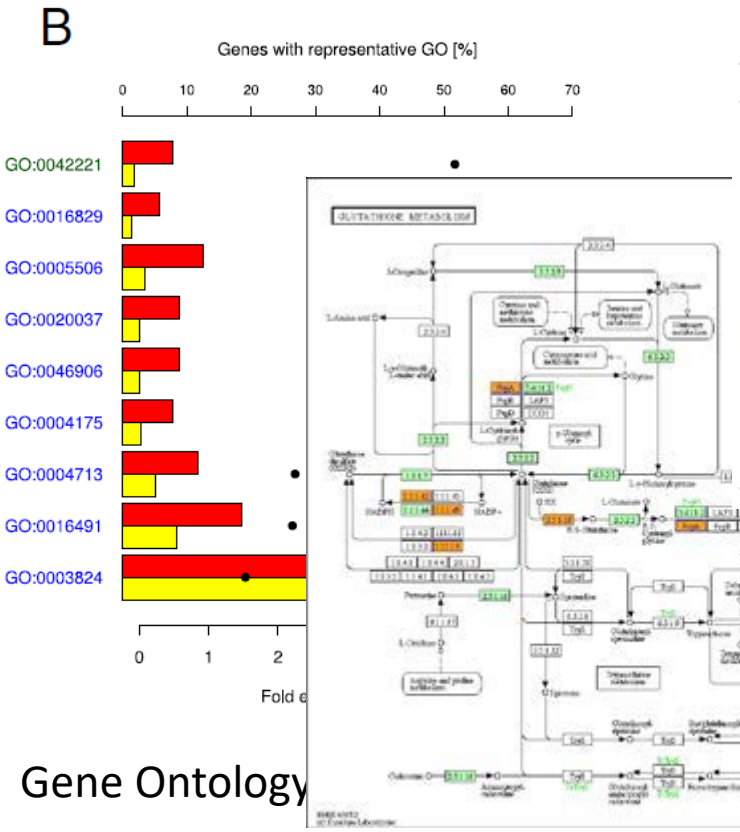


Differential expression analysis

A volcano plot displays both p values and fold change



Functional annotation

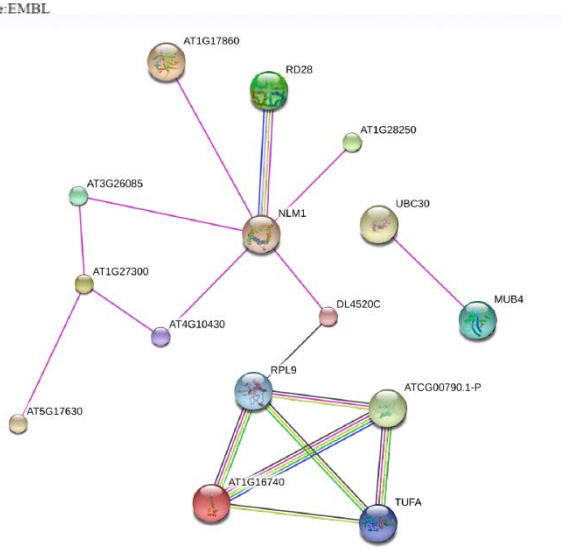


Gene Ontology

Pathway analysis

Family name	Gene NO.	Mutant vs. WT	Gene name	Description
bHLH	AT1G09530	up	POC1	transcription factor PIF3 [Source:EMBL
bHLH	AT1G10610	up	AT1G10610	transcription factor bHLH90 [Source:EMBL
bHLH	AT5G04150	down	BHLH101	transcription factor bHLH101 [Source:EMBL
C3H	AT1G29560	up	AT1G29560	Zinc finger C-x8-C-x5-C-x3-H type family protein [Source:EMBL
C3H	AT2G40140	up	SZF2	zinc finger CCCH domain-containing protein 29 [Source:EMBL
C3H	AT4G29190	up	AT4G29190	zinc finger CCCH domain-containing protein 49 [Source:EMBL
CO-like	AT3G02380	down	COL2	zinc finger protein CONSTANS-LIKE 2 [Source:EMBL
CO-like	AT5G15850	down	COL1	zinc finger protein CONSTANS-LIKE 1 [Source:EMBL
ERF	AT1G21910	down	DREB26	ethylene-responsive transcription factor ERF012 [Source:EMBL
FAR1	AT1G80010	down	FRS8	FAR1-related sequence 8 [Source:EMBL
GRF	AT5G53660	up	GRF7	growth-regulating factor 7 [Source:EMBL
MIKC	AT5G10140	down	FLF	MADS-box protein FLOWERING LOCUS C [Source:EMBL
MYB_related	AT1G01060	down	LHY	protein late elongated hypocotyl [Source:EMBL
MYB_related	AT3G09600	down	AT3G09600	myb family transcription factor [Source:EMBL
MYB_related	AT5G17300	down	RVE1	myb family transcription factor [Source:EMBL
MYB_related	AT5G37260	down	RVE2	protein REVEILLE 2 / DNA binding / transcription factor [Source:EMBL
RAV	AT1G13260	down	RAV1	AP2/ERF and B3 domain-containing transcription factor RAV1 [Source:EMBL
WRKY	AT5G49520	up	WRKY48	

DE transcı



Protein-protein interaction

BMC Genomics. 2015; 16(1): 123.

Suggested readings

- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews 10, 57-63.
- Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews 12, 87-98.
- Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nature reviews 12, 671-682.

Next lecture

- RNA-seq pipeline and Tuxedo tools