Next-generation Sequencing

Lecture 9

Alignment

- GCACTTCACAAATTAATGACCATGAGCTCGTTTTTGATAAACTCCAACTACATCGAGCCC
- ||||||||||| |||||||||
- ACCATGAGCTCGATTTTGATAAA

GOAL: to efficiently find the true location of each read from a potentially large quantity of reference data while distinguishing between technical sequencing errors and true genetic variation within the sample.

1. Efficient
2. True location
3. Distinguishing between technical sequencing errors and true genetic variation

Mapping Algorithms

- Hash Table (Lookup table)
  - FAST, but requires perfect matches. [O(m, n + N)]
- Array Scanning
  - Can handle mismatches, but not gaps. [O(m, N)]
- Dynamic Programming (Smith-Waterman)
  - Indels
  - Mathematically optimal solution
  - Slow (most programs use Hash Mapping as a prefilter) [O(mnN)]
- Merge Sorting

  - i.e. Slider [Malhis et al 2009]
- Indexing method, i.e. Burrows-Wheeler Transform (BW Transform)
  - FAST. [O(m + N)] (without mismatch/gap)
  - Memory efficient.
  - But for gaps/mismatches, it lacks sensitivity

Short-Read Alignment Tools with indexing

- Indexing Reads with Hash Tables
  - ZOOM: uses spaced seeds algorithm [Lin et al 2008]
  - RMAP: simpler spaced seeds algorithm [Smith et al 2008]
  - SHRiMP: employs a combination of spaced seeds and the Smith-Waterman
  - MAQ [Li et al 2008b]
  - Eland (commercial Solexa Pipeline)
- Indexing Reference with Hash Tables
  - SOAPv1 [Li et al 2008]
- Indexing Reference with Sux Array/Burrows-Wheeler
  - Bowtie [Langmead et al 2009]
  - BWA
  - SOAPv2

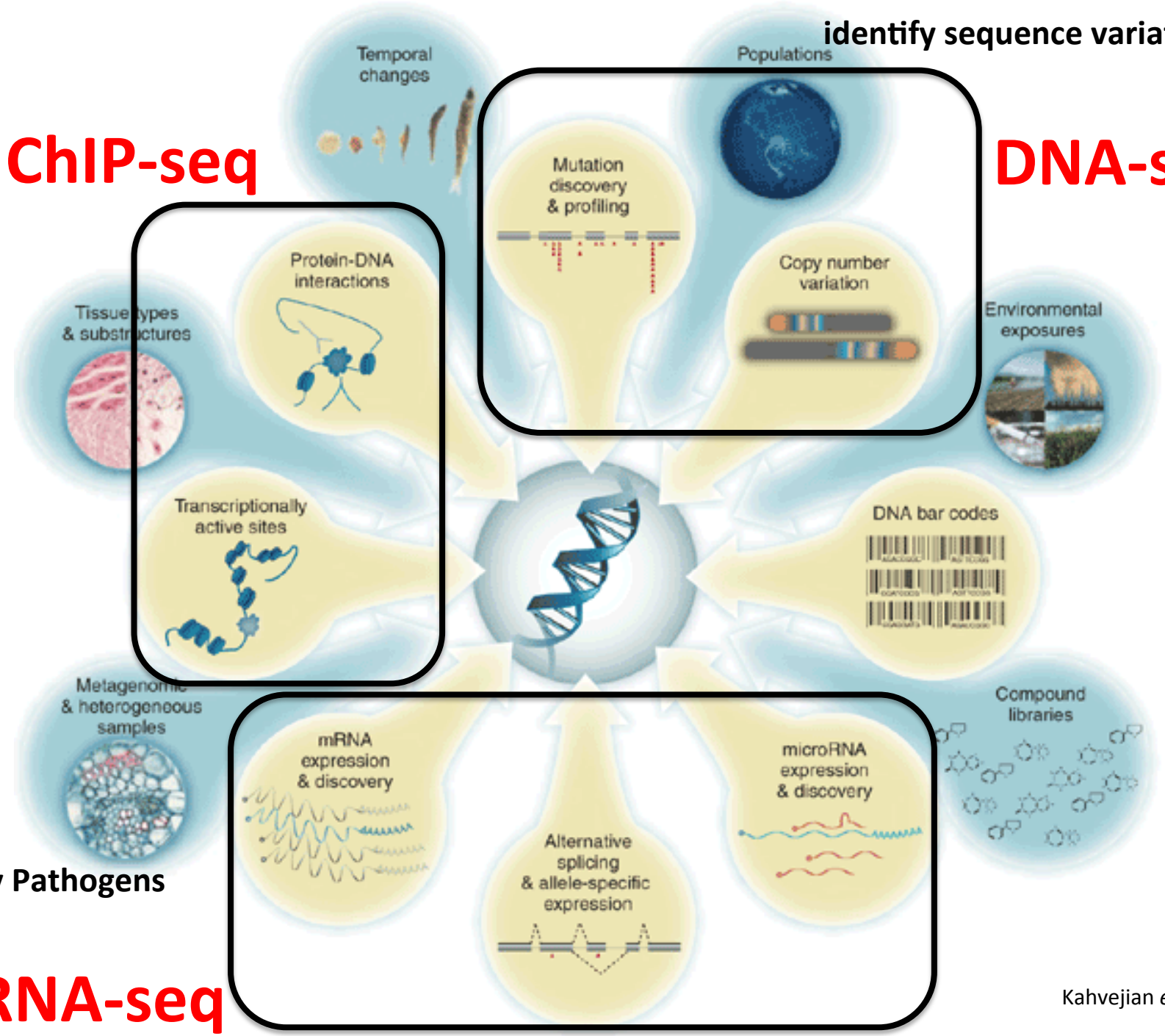
Output: SAM format

A SAM file consists of two parts:

- Header
  - contains meta data (source of the reads, reference genome, aligner, etc.)
  - All header lines start with "@".
  - Header fields have standardized two-letter codes for easy parsing of the information.
  - Most current tools omit and/or ignore the header.
- Alignment section
  - A tab-separated table with at least 11 columns
  - Each line describes one alignment

http://samtools.sourceforge.net/SAM1.pdf

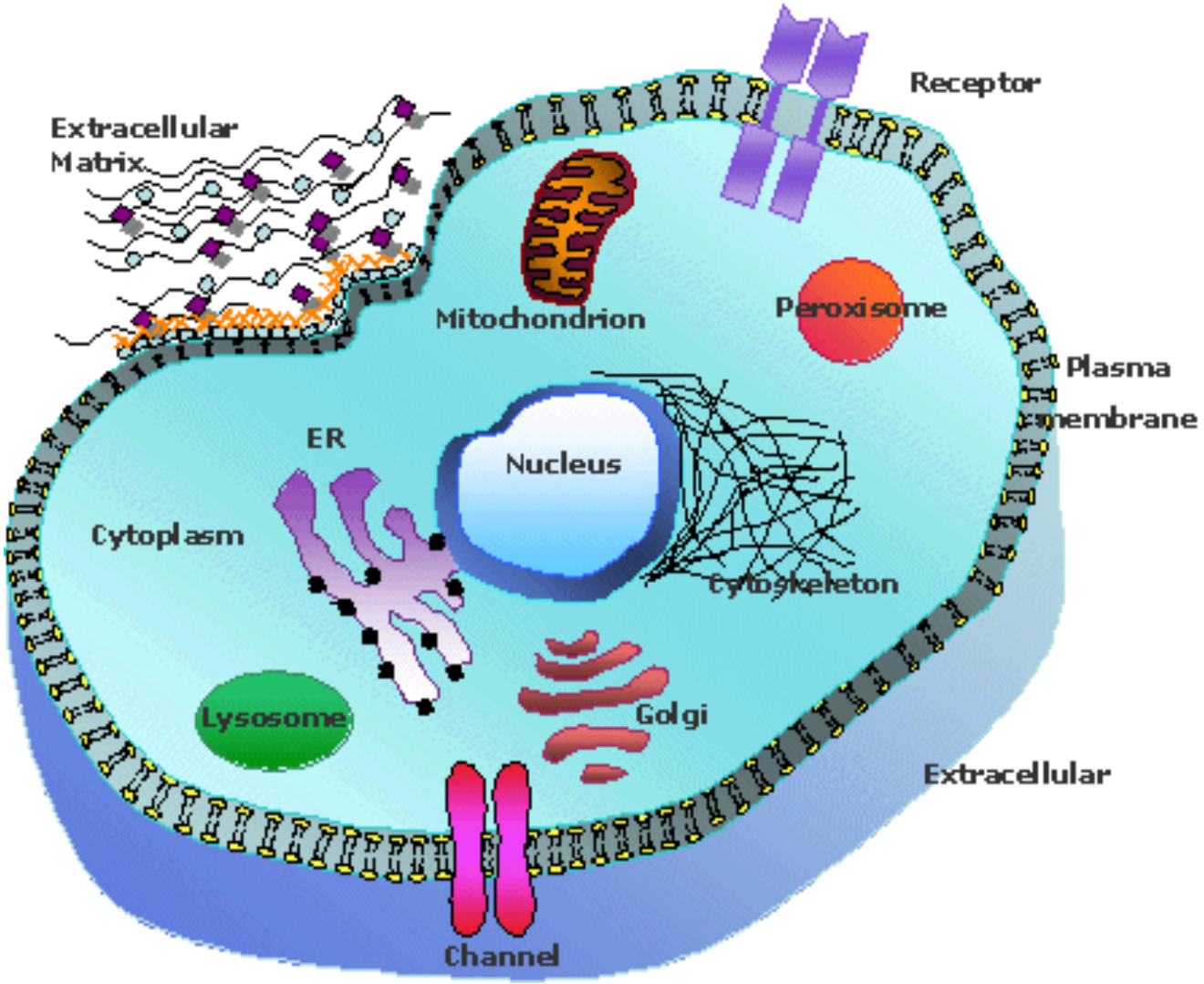
identify sequence variations



Identify Pathogens

Kahvejian *et al*, 2008

Cells: Building Blocks of Life

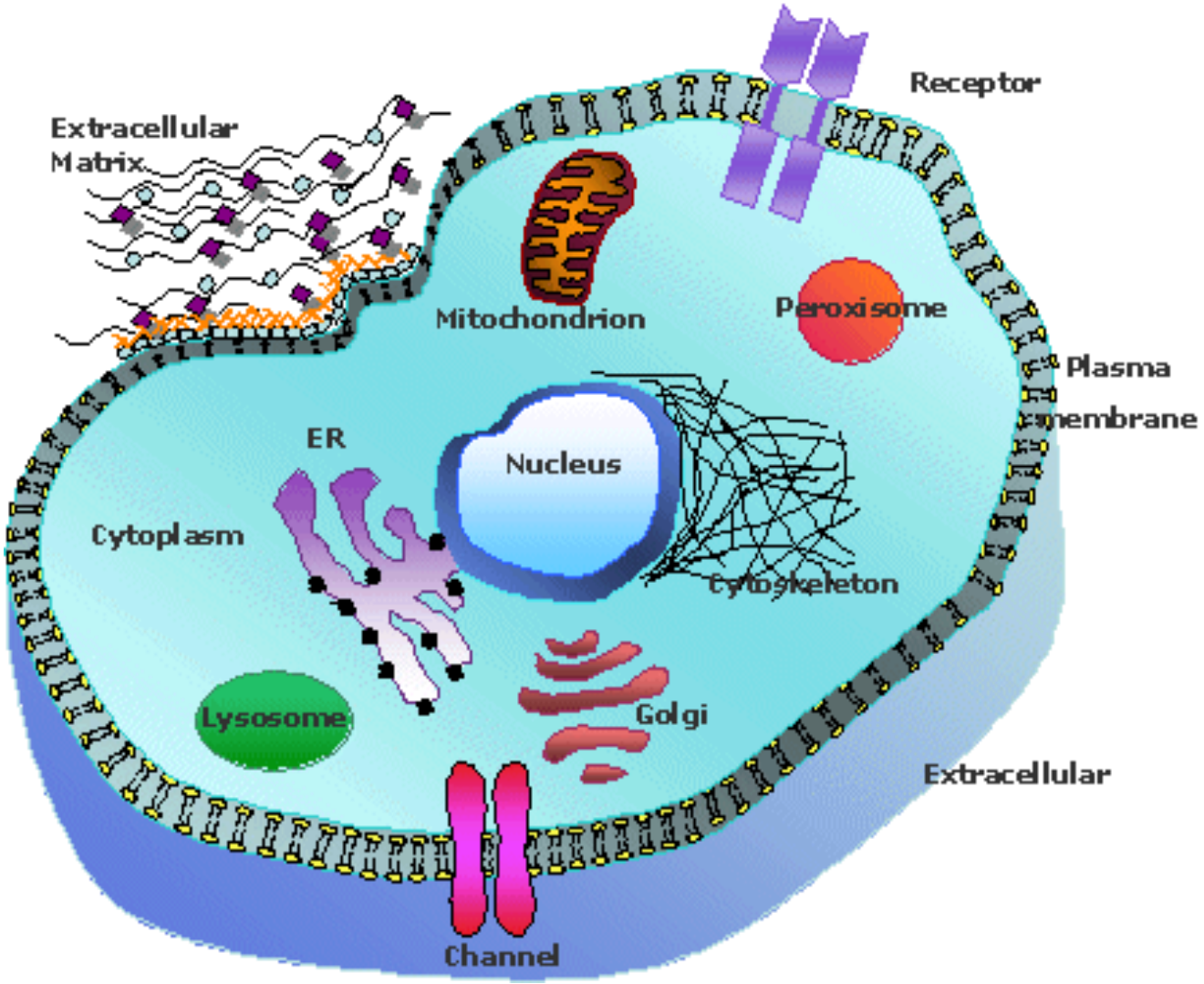


- Cells are the smallest form of life—the functional and structural units of all living things.

Cells: Building Blocks of Life

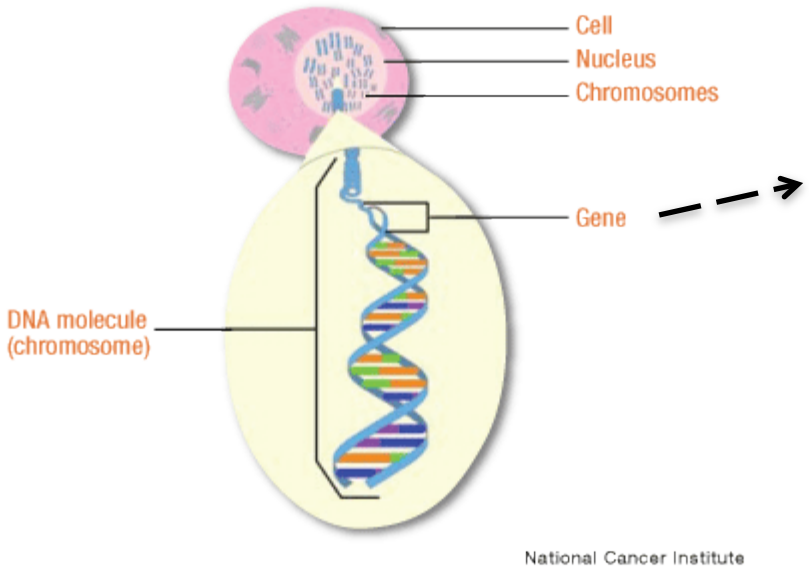
- Approximately how many cells make up human body?
  - ❑ 1
  - ❑ 100
  - ❑ 1000
  - ❑ 100,000
  - ❑ 1,000,000 (1 million)
  - ❑ 1,000,000,000,000,000 (1 trillion, 10^{12})
  - ☒ 100,000,000,000,000,000 (100 trillion, 10^{14})

Cells: Building Blocks of Life



- Each cell has a nucleus which contains genetic material, that is, DNA molecules.

DNA: "Blueprints" for a cell

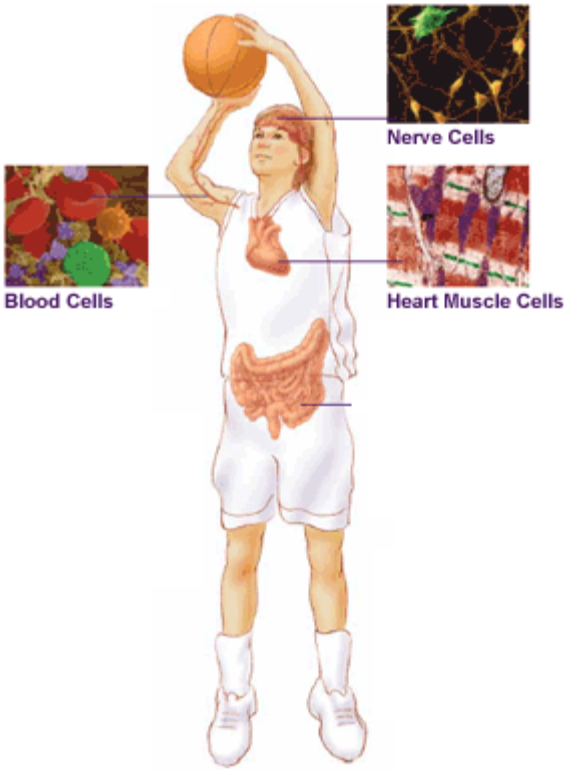


GAATTCTCTTTGGTATCCAATGAAGAAATCGAATCCATACCCATAGCTATAAAAAACAT
TTCAGGAGAAAATAAGACCGAAGCTGCTCAATTAGGCGCAATTGATTCGTTTCAAAAAT
GTGAAACTTGCCAGCTTACTTCGGCATGTCCTGGTCATTTTGGAAAATTTCATCTTACT
CAACCATTATTTAAAGTCGCATTTAAAAAACTTGTTGAAAATATTTTTAAATATACTTG
TTCTTTCTGTGGTGCTTTACAAAATCTTGAACTTCTGGAATTGATCAAGCAGATAGACG
AACGAAATACTGGAATAACAGTTAAAGATCGTGCTGCTTTTAAAAAAATTTTAGAAGCT
ACCAAACAAAGCAAATTCAAGTGTATTGCACCTAATTGCCAAAAACAAGTCTCTCCTTT
ACAATATTCGAAAAATAATAACTTTATATATAATTCGGGTACTACAAAGGGTATAGTTT
TGGATAACAGGCATGTGTTTAATATCTTACAAAATCTTCCACAAACGTTTAAATTATTG
TTAACCCCTTCGAATGCTCATCAAATCGTATCTCCCGAAAATGTCTTTTATGCTAATAG
TATCTTACTTCCACCACATAATCTACGAACTATCAATGTTTATGATGGTCAGGTTACGA
GTTTGTTAACAAGTGATTTGAATCTGATAATGCGAAGAGTTGCTAATAATGAGACAAAT
GCAAAAATACAAAAAATCTTGGATTCTATCGATAACAGCCGAGGTGCCAATCCATATGC
TACAAATAAAAAGCTTACTTTGGATACTTTGACAGGTGGACACTCAAAAGAATCTTATT
TGCGAAGTTATATTAATGGCAAACGTATTCCTGAGACTGCCAGAGCTGTAATCGAACCC
TCTATGAATAAAACTGGCTTTATTGAAGTACCATCTTACATTTTAAACAAGTTAAGAGA
TGTTGTCTTTTTATAATCACGTTACGAAAGATAACATACTCAAAAGTCTTCAAAACGAAC
AAGCTTTTCTAACATATATCAAAAGTGATCATAATTCTGAAAATCCTTATATGGTTTAT
GATTTAGCACAGAAGAATGGATATTTAACCTTGGCTCCTAATTTCGGTGATATTTTCGA
AAAAAGGAAAGAGGAAGGTGGTTTTGTAACTATTTGCAGACATCCATCTATCTGGTTAA
CTAATATCCAATCTGGTATAATAAAAAGATCAGAAGGGTTTACTATTAACATCCCAACC
ACAATTTGCACATCTTTTAATGCTGATTTTGATGGAGATGAGATGACAATATATTCTTT
CAAATCCCCATGTGCCAATCTCGAACAAGCTTTGATTATGAACTCACGAAATCTCTTCA
AAAATTCTATAACAAGCAATCCAATGTTCGGCTTGGTCCAAGATCAAATACCAGCCTTG
AATAAGTTATATAGACGACAAAATTATACATATAACGATGCGTTGGTGATTTTAGGACA
ATTCGGATTTCTGTTAACACCTGGAAAAGATAATTATACCGGAAAAGATATACTTTCTT
GTGTATTCCCAAAACATTATACACTCAAAGGAATTGTTGAAAATGGCGAACTTATTTTG
GAGAATTTTACAAATAAACTCGTTTCCGCAAATTCCTCAAAGTCCATCTTTGGGCATCT
TGTTTTATTTTATGGACAAGAGTATGGTTTGACTATATTGGATACAATGCGAGATATTG
TTCAAAATTTTATTACACATTTTGGTTTCAGTGTAAAAATCCGAGATATGATCCCAAGC
CCAAAAATTTTGGATATTCTAGAAAAGATCGTAGACCAAGAAGTGGATAAAATTGATAA
ACAAACAAAACTTCTATATGACGATATCGAACAAGGTAAGGTTATAATCAACTCTTATG
ATGATATTTCTGAGTTCAGATTAAAAAATGTGGCTATTATGAAAAAGAAACTAGAAAGC
AAACTTTTGGAACTTTTGGATGAATATTATGATGAAGACAATAATTTCCTAGAGATGTA
TAGAACGGGATATAAGGTCAACATTAACGAACTTCTCTCTATTATGTGTTTCTCGGGTT
TTAAAAATTATGGAAATATCGAAATGATTACACCGGGTCTTAATGGTAAAACATCTTTG
TTTAGCTTACCAGATTCTATAAACTTACAAGATTATGGGTTCATCAAAAGCTCTATTGC
CAAAGGGTTAACGTTTGAAGAATATGCTACAATCGTAAAACAAGAAGCTTTTCCACAAA
TTGTTAATGTTACAACTGGTACTTCACAAACAGGATTTTTGGGGAAAAAAATGGTTAAA
ATGGCTTCTGAATTC

- Each human cell has **identical** genetic information – a total of 3 billion DNA base pairs, including 25,000 genes

Why are cells different?

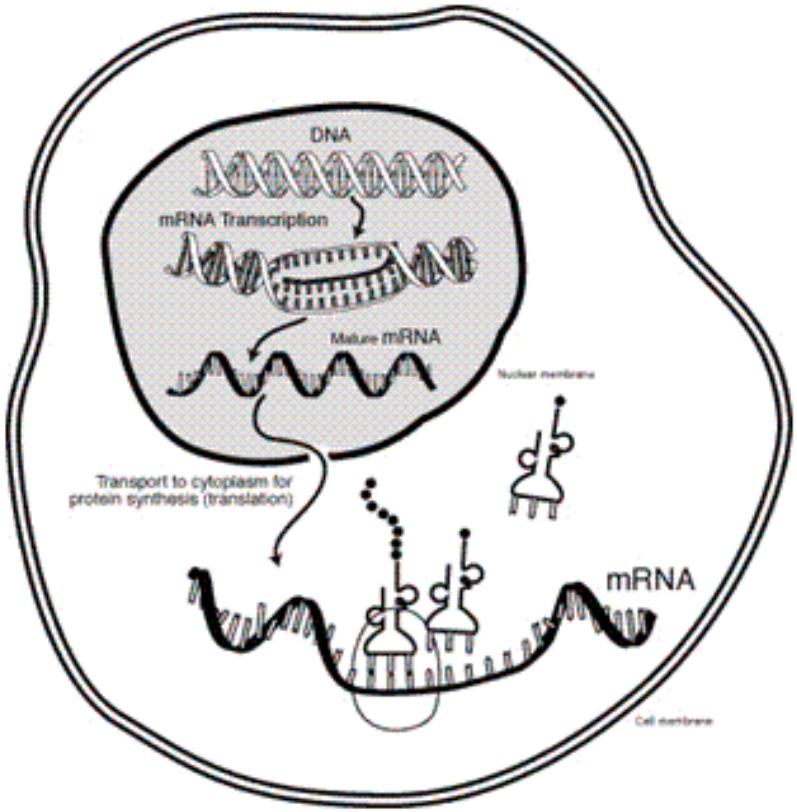
- The trillions of cells in human body are organized into >200 major tissue types, each customized for a particular role, for example
  - Red blood cells carry life-giving oxygen to every corner of your body.
  - Nerve cells sling chemical and electrical messages that allow you to think and move.
  - Heart cells constantly pump blood, enabling life itself.

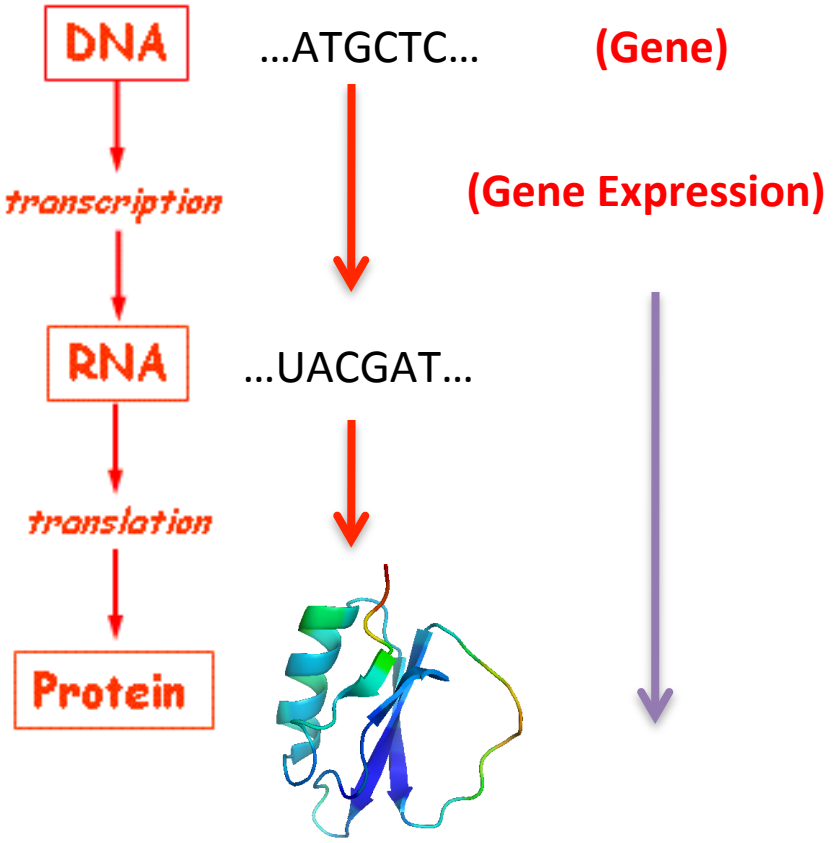


Question

- Q: What make those cells different?
- Cells contain the same genetic information (3 billion DNA base pairs, 25,000)

Flow of Genetic Information

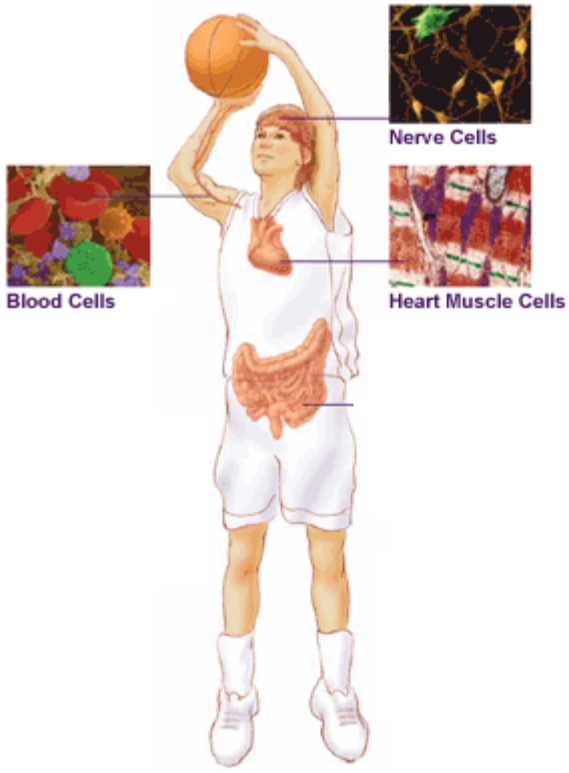




Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product.

Why are cells different?

- Q: Since the cells contain the same genetic information (3 billion DNA based pairs), what make them different?
- A: The ~25,000 genes in our DNA are like a tool kit, are used (i.e., expressed) by different cells in different ways at different time.
- Gene expression is regulated by different cells.



Studying the Expression of Groups of Genes

- A major goal of biologists is to learn how genes act together to produce and maintain a functioning organism.
- Large groups of genes are studied by a systems approach.
- Such approaches allow networks of expression across a genome to be identified.
- Genome-wide expression studies can be carried out using **RNA-seq** or microarray assay.

Transcriptome

- Transcriptome: How to genome-wide measure the expression of those genes? How to get the gene expression profiles.
- **gene expression profiling** is the measurement of the expression of thousands of genes at once, to create a global picture of cellular functions.
- These profiles can distinguish between cells that actively dividing, or show how the cells react to a particular treatment.
- Gene regulation network: who regulates those genes expression.

What is RNA-seq?

- RNA-seq refers to the method of using Next-Generation Sequencing technology to measure RNA levels.

Applications of RNA-seq

- Gene expression
  - Expression of individual genes/loci
  - Quantitatively discriminate isoforms using junction reads and coverage of individual exons, introns, etc.
- Annotation
  - New features of the transcriptome: genes, exons, splicing, ncRNAs
- SNP
- Fusion gene detection

Comparison between different technologies

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

Some Advantages of RNA-seq over Microarrays

- Microarrays measure only genes corresponding to predetermined probes on a microarray while RNA-seq measures any transcripts in a sample.

- With RNA-seq, there is no need to identify probes prior to measurement or to build a microarray.

- RNA-seq provides count data which may be closer, at least in principle, to the amount of mRNA produced by a gene than the fluorescence measures produced with microarray technology.

Some Advantages of RNA-seq over Microarrays

- RNA-seq provides information about transcript sequence in addition to information about transcript abundance.

- Thus, with RNA-seq, it is possible to separately measure the expression of different transcripts that would be difficult to separately measure with microarray technology due to cross hybridization.

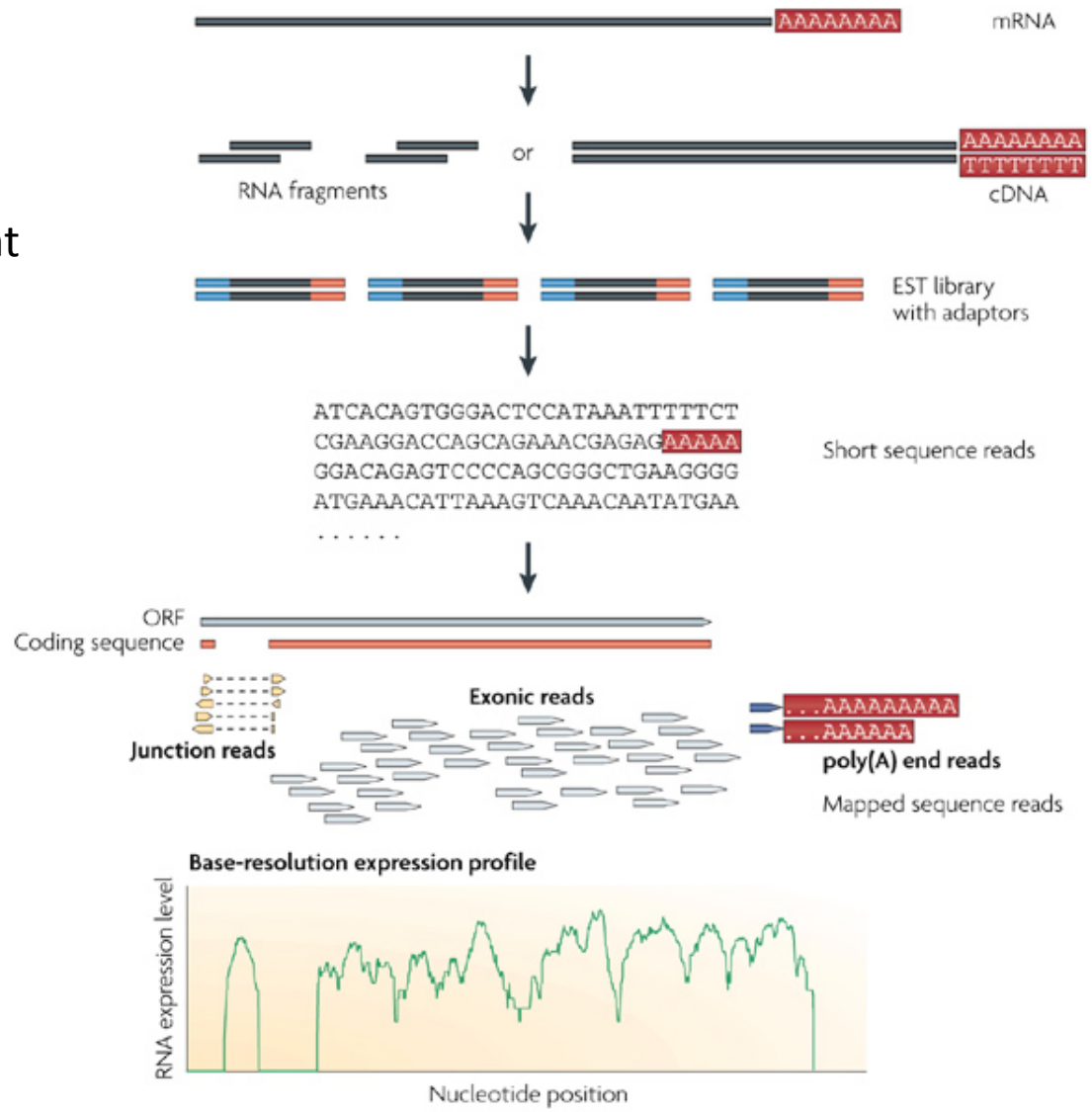
- Sequence information also permits the identification of alternative splicing, allele specific expression, single nucleotide polymorphisms (SNPs), and other forms of sequence variation.

RNA-seq Exp.

Extract sufficient mRNA from total using either poly-A selection or depletion of rRNA (RiboMinus).

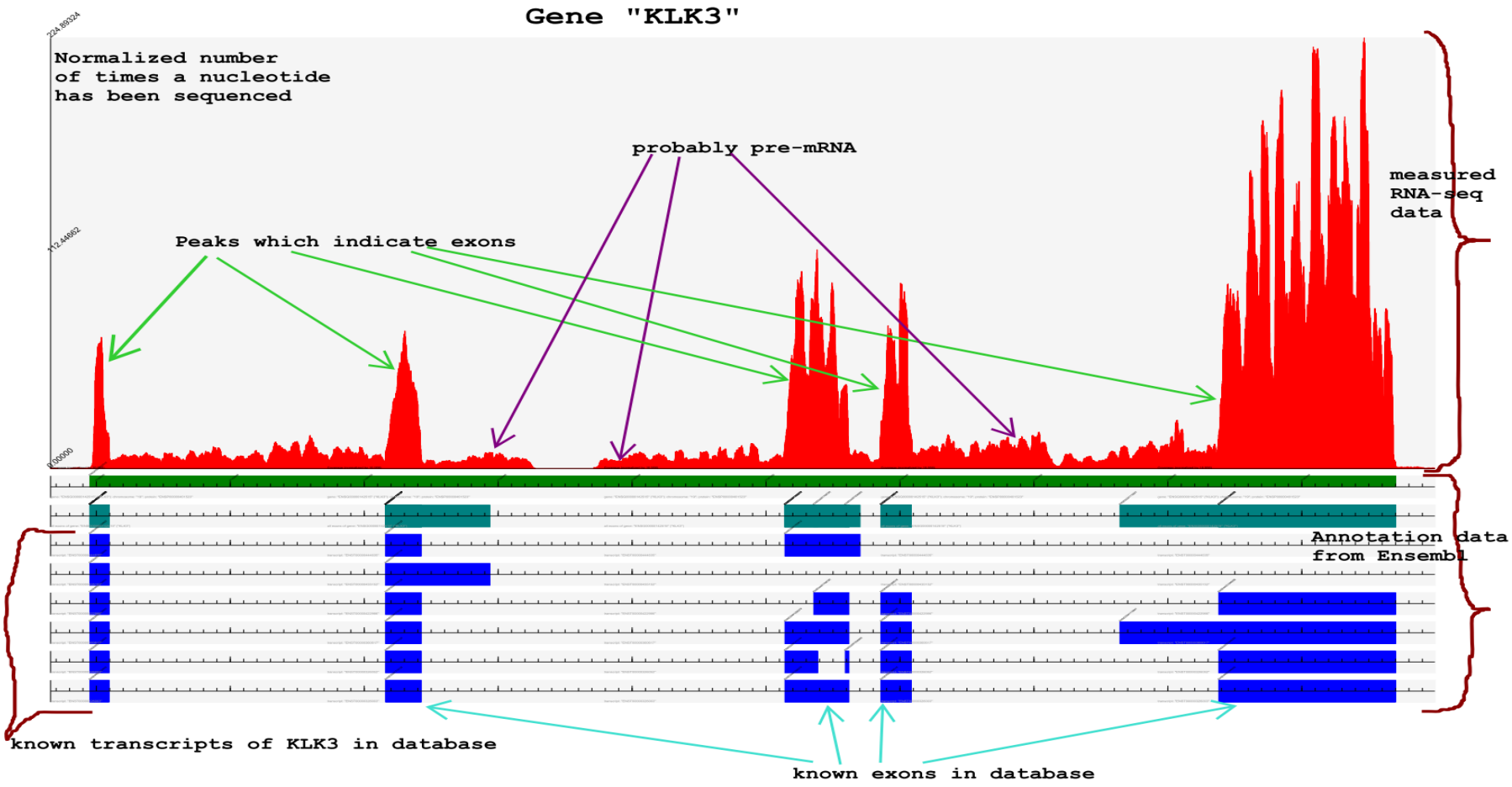
Non-poly(A) RNA) can yield important noncoding RNA gene discovery

reads are aligned with the reference genome

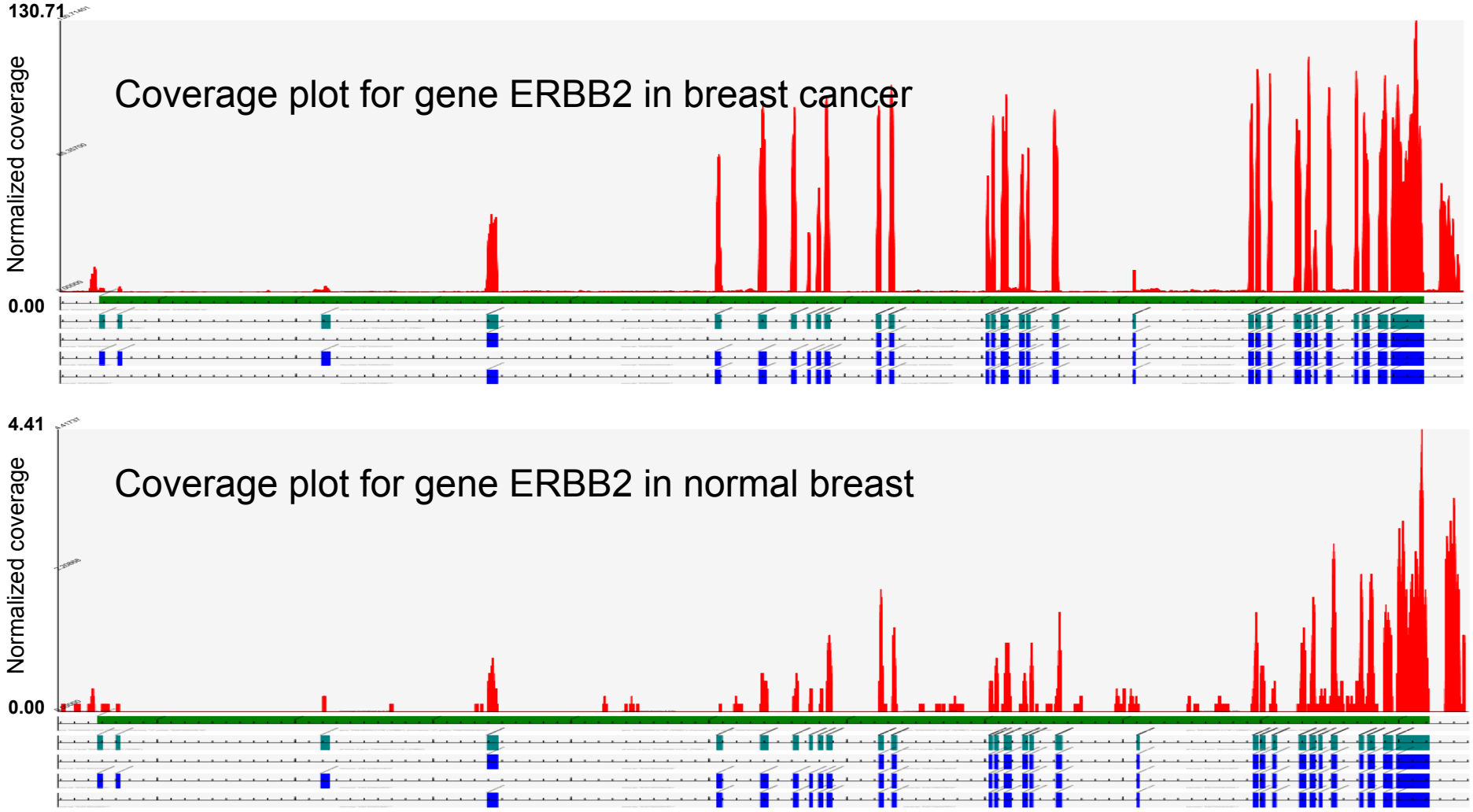


Wang, Gerstein, Snyser, 2009

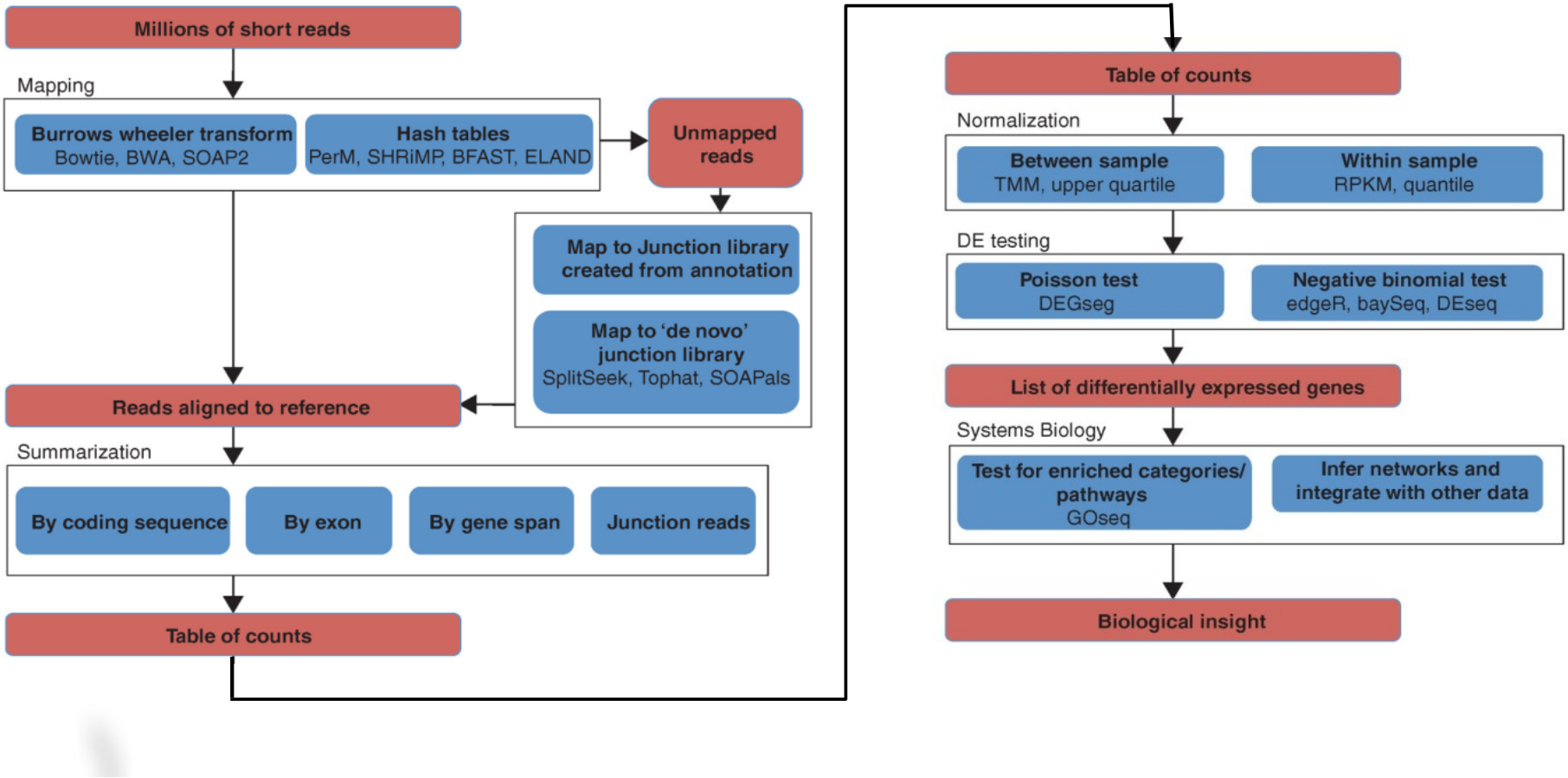
Examples of RNA-seq



Examples of RNA-seq



RNA-seq analysis pipeline



Steps involved on RNA-seq analysis

- Experimental design
- Preprocess
  - Split by barcodes
  - Quality control and removal of poor-quality reads
  - Remove adapters and linkers
- Map the reads
- Count how many reads fall within each feature of interest (gene, transcript, exon etc).
- Remove absent genes and add offset (such as 1)
  - Prevent dividing by 0
  - Moderate fold change of low-count genes
- Normalization and Identify differentially expressed genes.

Experimental design

- Include replicates in your experiment.
  - drawn from a single RNA-seq experiment can be misleading.
- Estimate the number of reads needed for an experiment.
  - Depends on the organism and the level of the differences you want to detect.

Coverage Requirements: How many lanes/plates/wells?

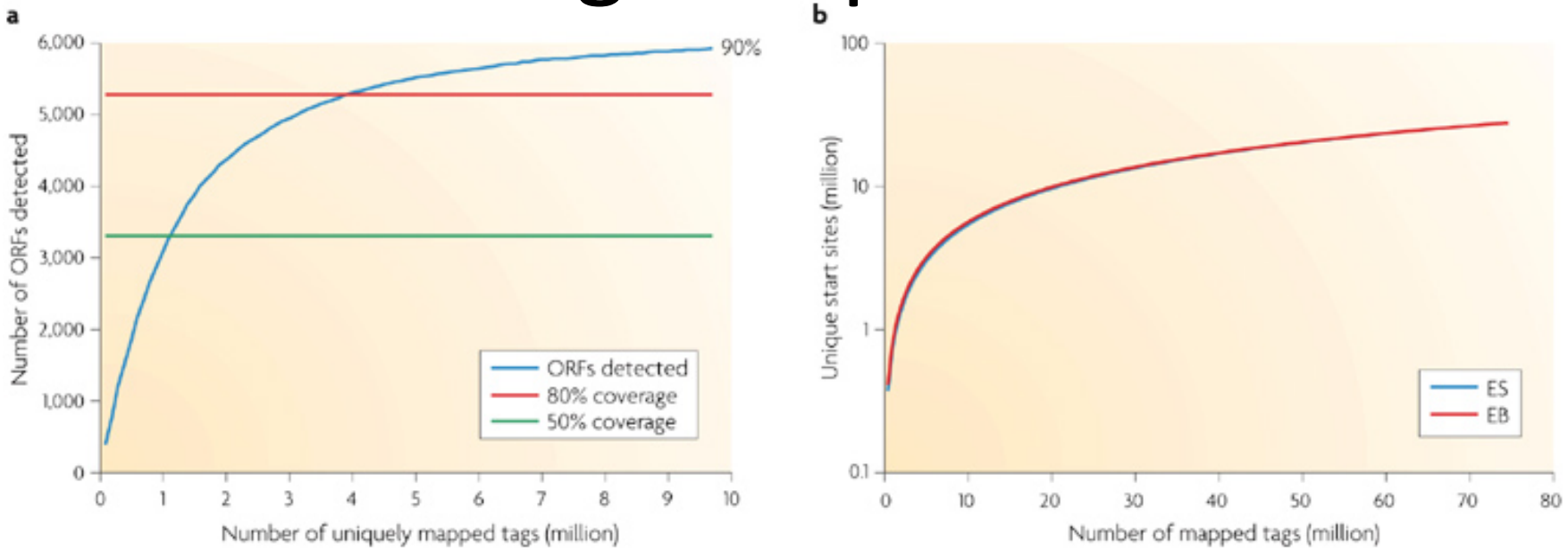
- Depends on
- Read length
- Size of transcriptome
- Complexity of tissue
- Biological variance
- System errors

How many lanes do we need?

Table 1. Power to detect differentially expressed genes depends on the number of lanes used for each sample

No. of lanes compared	Differentially expressed genes	Overlap with genes called from the array	Correlation of fold changes between the sequence data and the array
One vs. one	5670	4208	0.67
Two vs. two	7994	5340	0.70
Three vs. three	9482	5909	0.71
Four vs. four	10,580	6278	0.72
Five vs. five	11,493	6534	0.73

Coverage Requirements



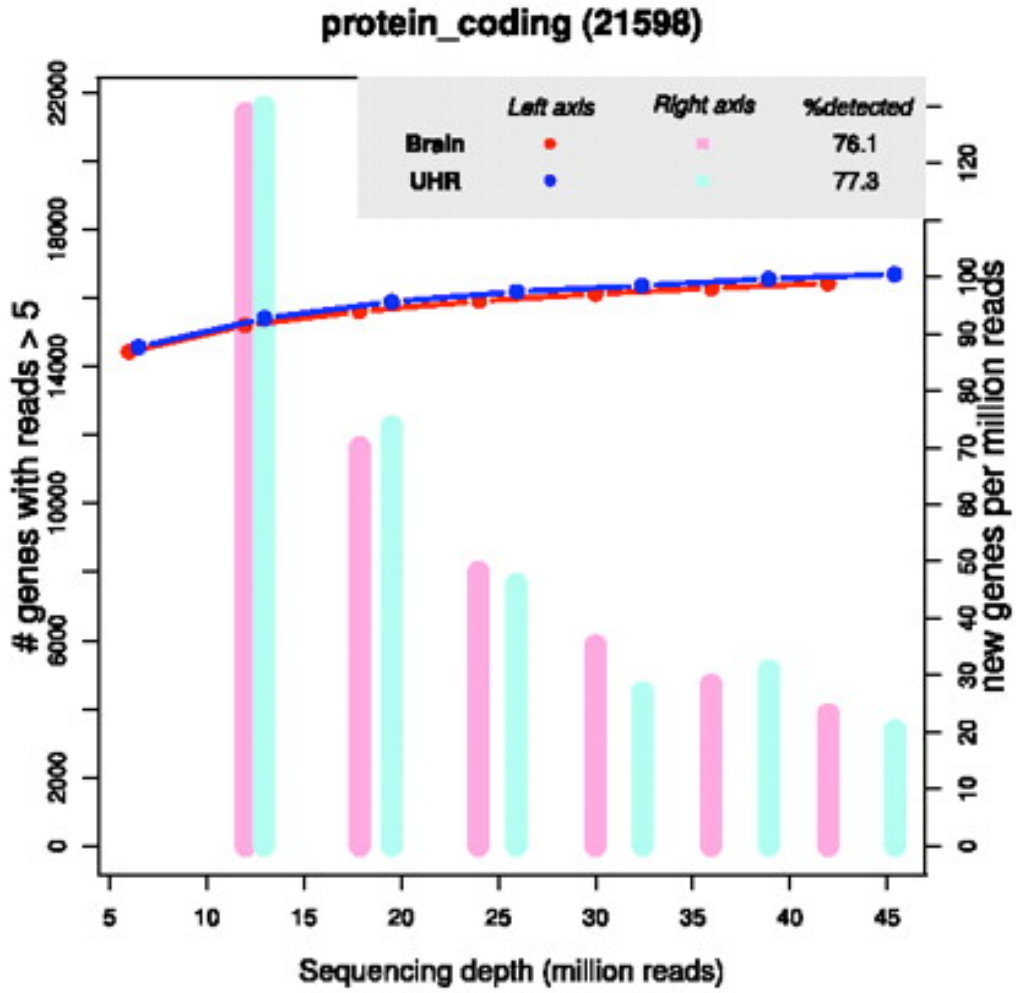
Nature Reviews | Genetics

A: 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads

B: The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes.

Wang, Gerstein, Snyser, 2009

Coverage Requirements



Differential expression in RNA- seq: A matter of depth. Genome Res. 2011.

Summary of Example Illumina RNA-Seq Data

- 40% of reads mapped uniquely to a genomic location
- Of these, 65% mapped to autosomal or sex chromosomes

	Lane 1	Lane 2	Lane 3	Lane 4	Lane 6	Lane 7	Lane 8
	Solexa Run 1						
	kidney	**liver**	**kidney**	**liver**	**liver**	**kidney**	**liver**
Concentration (pM)	3	3	3	3	3	3	3
# Reads	13,017,169	14,003,322	13,401,343	14,230,879	13,525,355	12,848,201	13,096,715
Total Sequence (Mb)	417	448	429	455	433	411	419
# Mapped Reads	5,025,044	5,142,214	5,199,295	5,167,290	4,997,324	4,901,266	4,822,319
Mapped to chr1-22,X,Y	3,261,380	3,460,175	3,369,521	3,480,325	3,363,455	3,179,248	3,249,417
Mapped in Genes	2,706,150	2,847,704	2,792,026	2,861,877	2,761,468	2,630,987	2,668,148
Mapped in Exons	1,926,217	1,815,816	1,981,182	1,821,860	1,752,042	1,861,126	1,692,041
	Solexa Run 2						
	liver	**kidney**	**liver**	**kidney**	**kidney**	**liver**	**kidney**
Concentration (pM)	1.5	3	3	1.5	3	1.5	1.5
# Reads	9,096,595	13,687,929	14,761,931	8,843,158	13,449,864	9,341,101	8,449,276
Total Sequence (Mb)	291	438	472	283	430	299	270
# Mapped Reads	4,138,533	5,293,547	5,320,141	4,394,988	5,422,895	4,437,111	4,266,893
Mapped to chr1-22,X,Y	2,794,909	3,456,114	3,591,760	2,885,222	3,533,100	2,989,819	2,799,046
Mapped in Genes	2,328,896	2,875,214	2,959,436	2,416,834	2,938,079	2,488,832	2,345,160
Mapped in Exons	1,532,142	2,055,876	1,896,001	1,751,854	2,096,458	1,634,684	1,701,056

- Of these, 83% were located in genic regions
- Of those outside…

Marioni and Mason et al, 2008

Coverage Requirements: How many lanes/plates/wells?

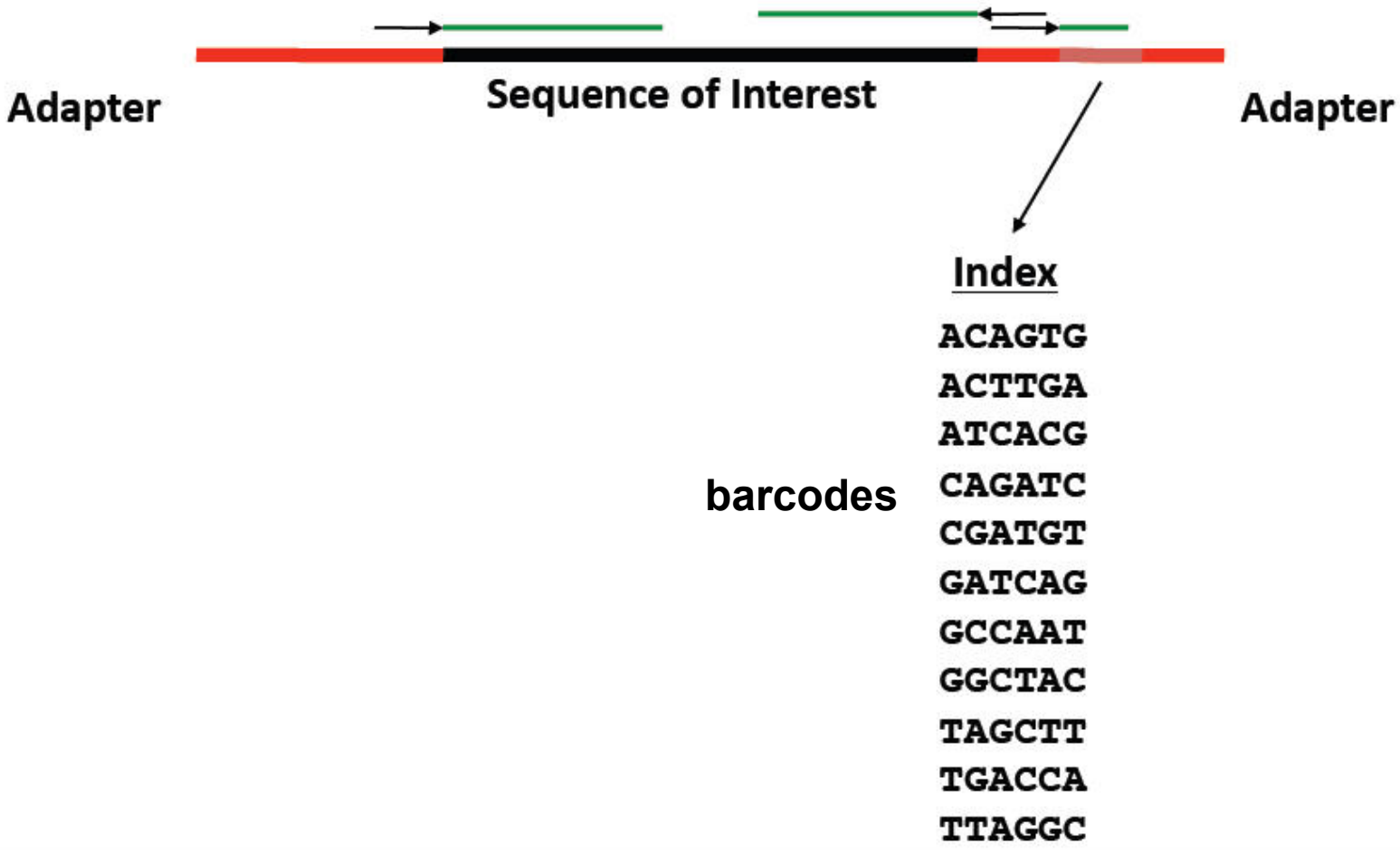
- Depends on
- Read length
- Size of transcriptome
- Complexity of tissue
- Biological variance
- System errors

HiSeq 2000
180-240 million reads/lane
10-20 million reads/sample
10-18 samples / lane

Preprocess

- Split by barcodes
- Conduct quality control and removal of poor-quality reads
- Remove adapters and linkers

Adaptor and barcode



RNA-Seq Error Sources

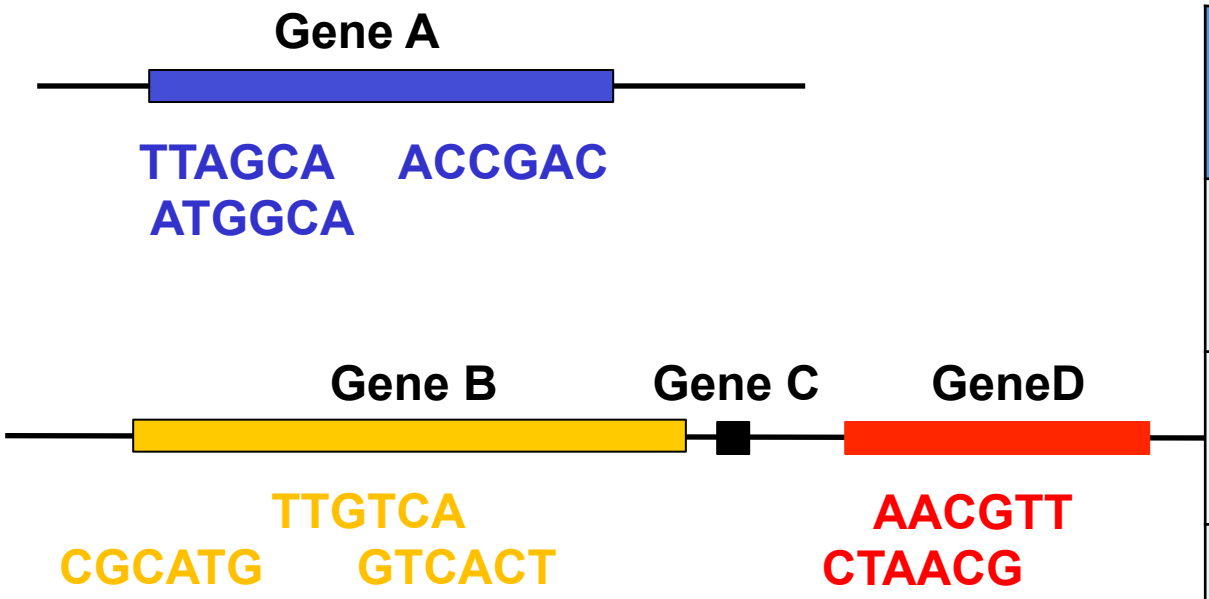
- read length
- Sample preparation
- Quality scores
- All errors from sequencing
- Number of genes expressed in tissue

Useful tools for preprocessing

- Fastx Toolki
  http://hannonlab.cshl.edu/fastx_toolkit
  - Split by barcodes: fastx_barcode_splitter.pl
  - Remove adapters: fastx_clipper
  - removal of poor-quality reads: fastq_quality_filter
- FastQC:

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Align reads to Genome and count



Gene ID	Sample1
A	3
B	3
C	0
D	2

For a given gene, the number of reads aligned to the gene measures its expression level.

Determine Abundance

- Count reads in gene, coding area, or exons.
- Need gene annotation files in GFF (General Feature Format) format, which gives complete gene, RNA transcript or protein structures
- Tools:
  - Cufflinks (http://cufflinks.cbcb.umd.edu/)
  - Sam2counts (https://github.com/vsbuffalo/sam2counts)
  - HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

Example Dataset after Aligning Reads

Gene	Control			Treatment 1		
1	14	18	10	47	13	24
2	10	3	15	1	11	5
3	1	0	10	80	21	34
4	0	0	0	0	2	0
5	4	3	3	5	33	29
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
53256	47	29	11	71	278	339
Total						
	22910173	30701031	18897029	20546299	28491272	27082148

Differential Expression (DE) Analysis

- To determine if gene-1 is DE, we would like to know whether the proportion of reads aligning to gene-1 tends to be different for experimental units that is for control than for experimental units that received a treatment.

14 out of 22910173		47 out of 20546299
18 out of 30701031	vs.	13 out of 28491272
10 out of 18897029		24 out of 27082148

Need Normalization

- More reads mapped to a transcript if it is
  - i) long
  - ii) at higher depth of coverage
- Normalize data such that i) features of different lengths and ii) total sequence from different conditions can be comparable.

Normalization

- Total Count (TC): Gene counts are divided by the total number of mapped reads
- Median (Med): the total counts are replaced by the median counts different from 0
- Upper Quartile (UQ): the total counts are replaced by the upper quartile of counts different from 0 Bullard et al., 2010)
- Quantile (Q): was for microarray, Hansen et al., 2012
- RPKM (Reads Per Kilobase of exon model per Million mapped reads) (Mortazavi et al., 2008)
- Trimmed Mean of M-values (TMM): used by edgeR Robinson and Oshlack, 2010
- DEseq normalization: Anders and Huber, 2010

Comparison between different normalization methods

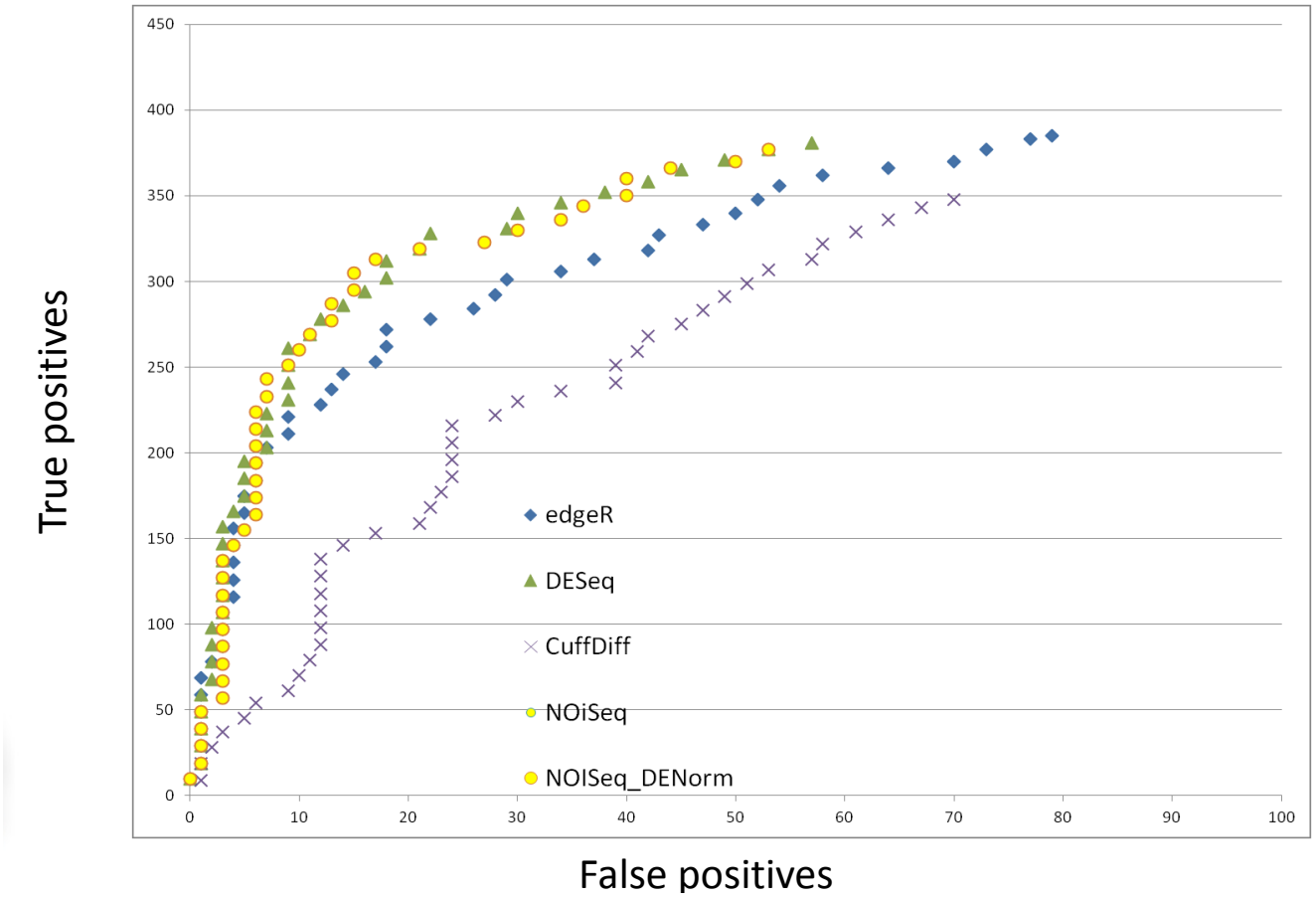
Method	Distribution	Intra-Variance	Housekeeping	clustering	False-positive rate
TC	-	+	+	-	-
UQ	++	++	+	++	-
Med	++	++	-	++	-
DEseq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	-	+	++	-
RPKM	-	+	+	-	-

Dillies et al. Briefings in Bioinformatics, 2012

Differentially expressed gene Analysis Tools

Tools	Statistics			speed
edgeR	Empirical Bayes estimation and exact tests based on the negative binomial distribution	Robinson et al., 2010	High TPR	media
DEseq	Negative binomial distribution.	Anders and Huber, 2010	Low TPR	media
NOISeq	Compares replicates within the same condition to estimate noise distribution of M (log-ratio) and D (absolute value of the difference).	Tarazona et al., 2011	High TPR	Data size
baySeq	Empirical Bayesian methods using the negative binomial distribution.	Hardcastle and Kelly, 2010		slow
TSPM		Auer and Doerge, 2011	Data size	media
BitSeq	a hierarchical log-normal model and determines the probability of differential expression by Bayesian model averaging	Glaus et al., 2012		
POME	Poisson mixed-effects model	Hu et al., 2012		

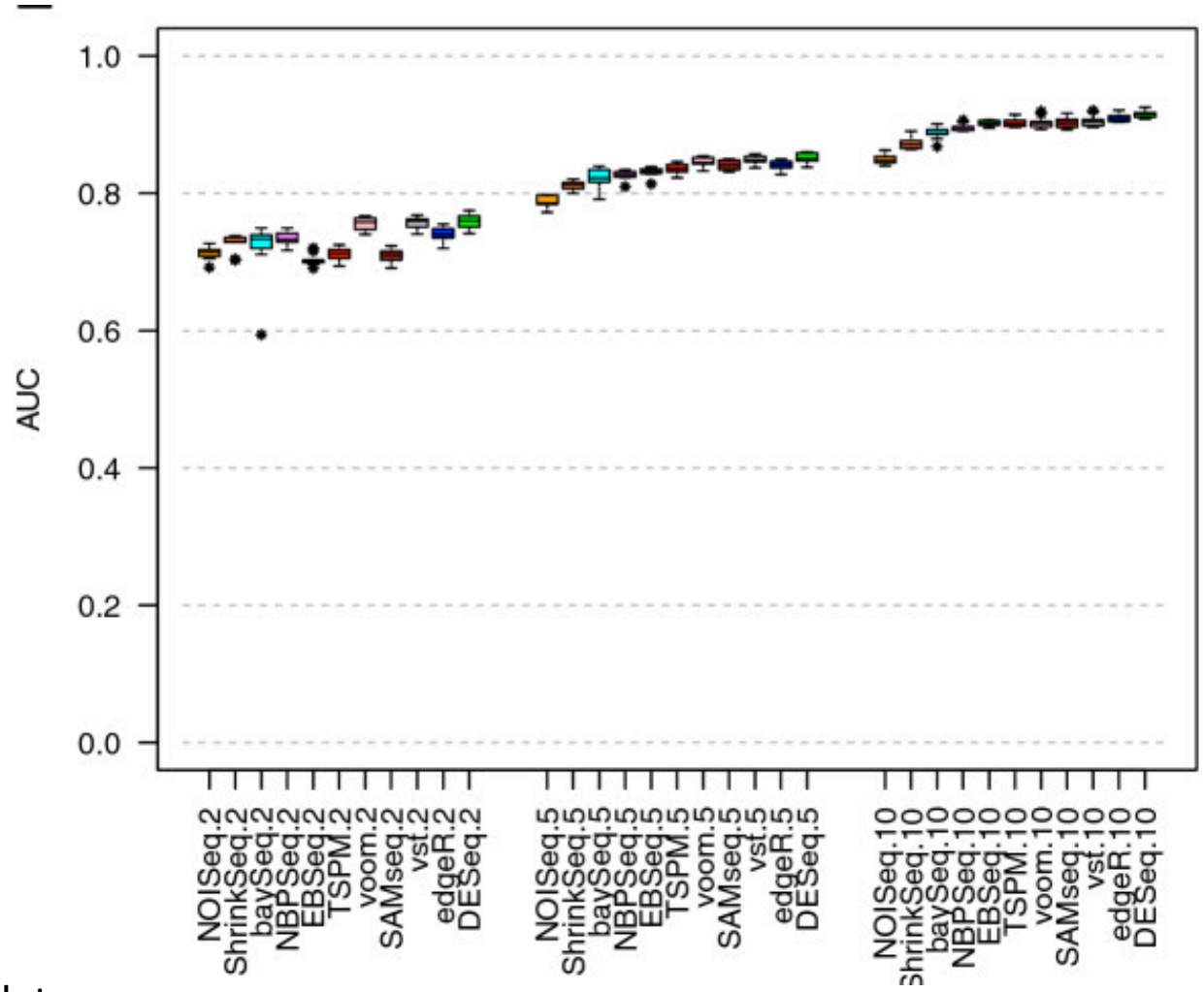
Performance of different tools



Data from MicroArray Quality Control (MAQC) Project

Bullard et al. BMC Bioinformatics, 2010)

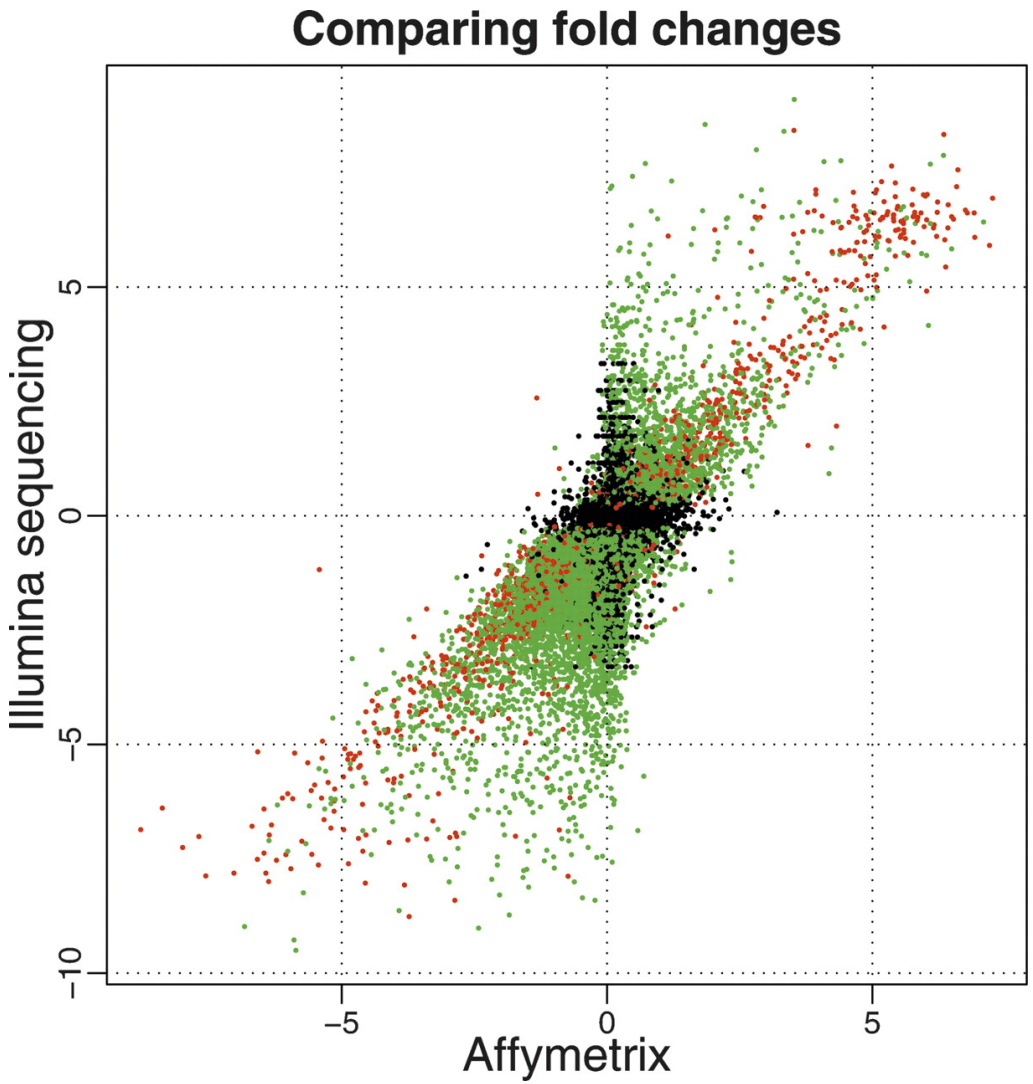
Performance of different tools



Simulation data

Soneson and Delorenzi. BMC Bioinformatics, 2013

Comparison with Microarray



$\log_2$ fold changes (liver/kidney)

Red: number of reads > 250 /gene
Green: number of reads < 250 / gene
Black: Genes not called as differentially expressed

The set of differentially expressed genes that show 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).

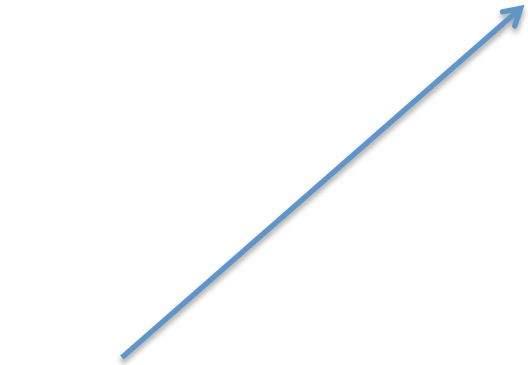
Using edgeR

Installation of edgeR

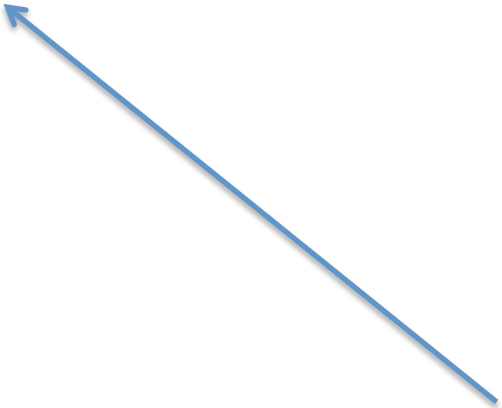
```
> source("http://bioconductor.org/biocLite.R")
> biocLite("edgeR")
```

Data file

66	98
38	7
128	238
91	64
4	3
302	508
37	71
517	309
490	1531
….	
0	1
328	236
16	16
0	0
4	2
780	1133
469	900
121901	298290
44	46



Control



Treatment

Using edgeR

```
> library(edgeR)
> library(stats)        #edgeR needs this lib
> set.seed(133)
> y <- as.matrix(read.table("your_data_file"))
> # need more lines to associate gene names with data
> # use sum() function to calculate the total number of reads
> lib.sizes <- c(13041385, 16428242 ) #the total number of reads
> d<- DGEList(counts=y, group=c("WT","Tr"), remove.zeros = TRUE)
> d <- estimateCommonDisp(d)
> ms <-  exactTest(d)
> result=topTags(ms, n=1000, adjust.method= "fdr")
```

Absent genes

- Remove absent genes (zero counts in all samples). It reduces the number of tests and the false discovery rate correction.
- Add 1 pseudocount (prevent dividing by 0).

Results

Comparison of groups: WT - Tr

	logConc	logFC	PValue	FDR
GRMZM2G304548	-3.040312	-0.8187952	0.000000e+00	0.000000e+00
GRMZM2G337229	-5.020442	0.6845438	0.000000e+00	0.000000e+00
GRMZM2G085260	-7.547953	-1.0193668	0.000000e+00	0.000000e+00
GRMZM2G060429	-2.328527	0.2535723	0.000000e+00	0.000000e+00
GRMZM2G092125	-8.148289	1.2663013	0.000000e+00	0.000000e+00
GRMZM2G123212	-10.028034	-2.1860488	0.000000e+00	0.000000e+00
GRMZM2G102356	-7.632246	1.5565755	0.000000e+00	0.000000e+00
GRMZM2G007256	-10.076032	-2.2233622	0.000000e+00	0.000000e+00
GRMZM2G168651	-4.621981	1.4587438	0.000000e+00	0.000000e+00
GRMZM2G322819	-30.122305	-39.7874987	0.000000e+00	0.000000e+00
GRMZM2G331701	-11.805408	-3.7628796	7.313298e-313	2.022459e-310
GRMZM2G068202	-11.807386	3.4861947	9.541348e-297	2.418732e-294
GRMZM2G162284	-9.951298	-1.7878318	3.360447e-254	7.863447e-252
GRMZM2G010783	-8.502862	-1.0699938	9.945640e-251	2.161045e-248

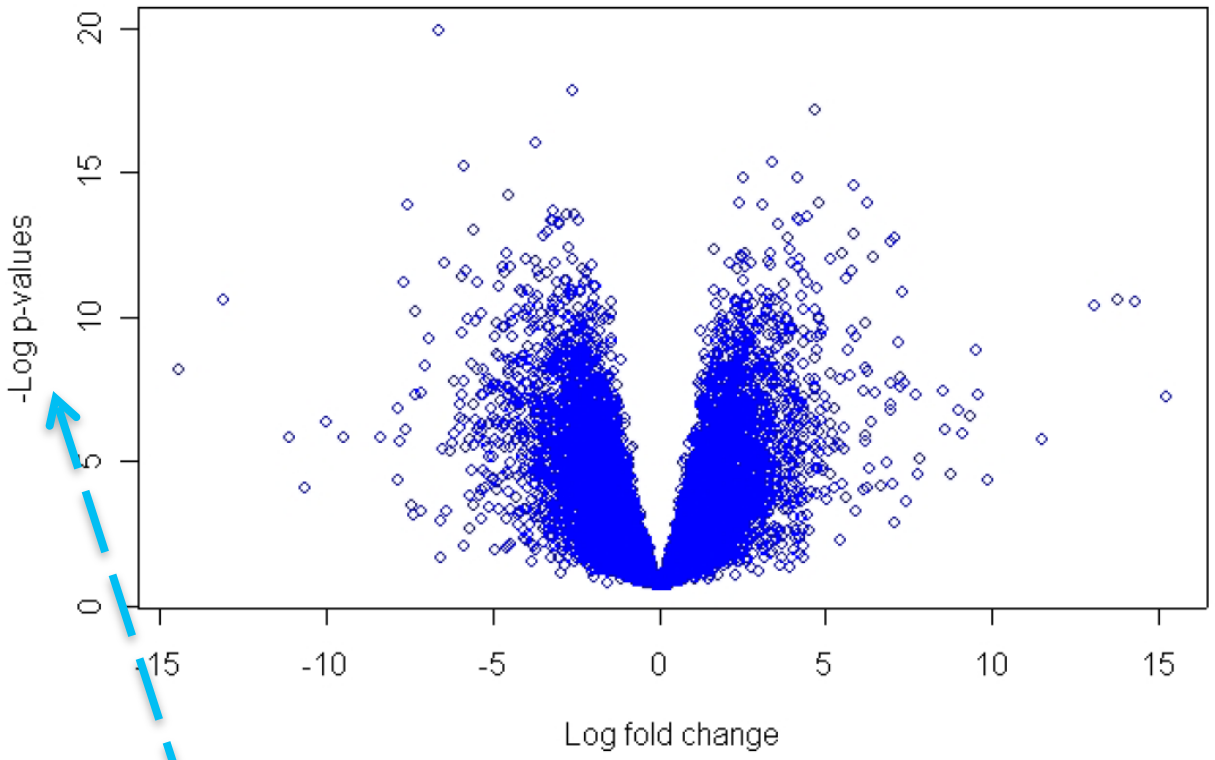
Default: it is sort by P-value

Volcano plot

- In statistics, a **volcano plot** is a type of scatter-plot that is used to quickly identify changes in large datasets composed of replicate data.

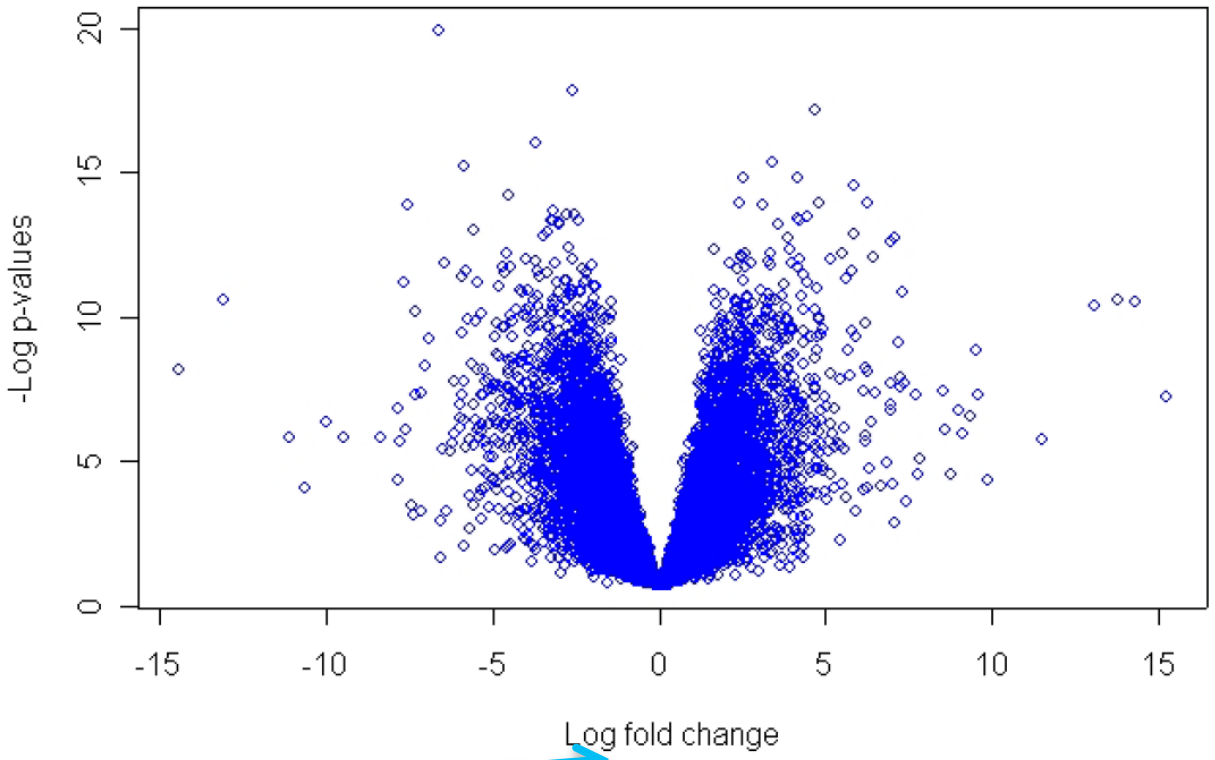
- It plots significance versus fold-change on the y- and x-axes, respectively.

Volcano plot



- A volcano plot is constructed by plotting the negative log of the p-value on the y-axis (usually base 10). This results in data points with low p-values (highly significant) appearing towards the top of the plot.

Volcano plot



- The x-axis is the log of the fold change between the two conditions. The log of the fold-change is used so that changes in both directions (up and down) appear equidistant from the center.

Volcano plot: R

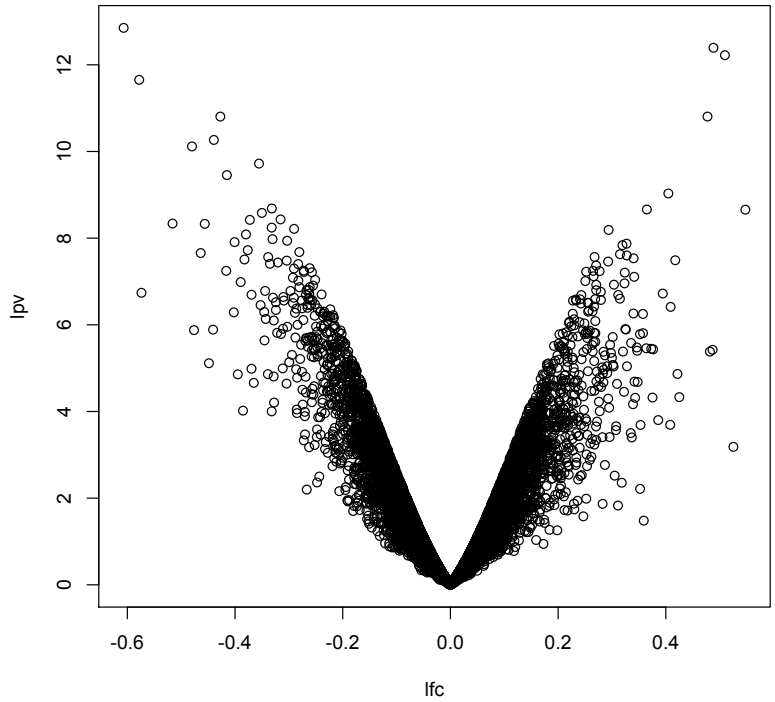
topTags

> results=topTagsms, n=20000, adjust.method="fdr", lfc=0)

> logfc=results[,2]

> logpvalue=-log2(results[,3])

> plot(logfc,logpvalue)



HW5

- Using edgeR and volcano plot
- Due on Feb. 24th