BIO306: Bioinformatics

Lecture 10

Gene expression profiling and RNA-seq

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Vocabulary – Review

Gene: a sequence of DNA/RNA which codes for a molecule that has a function. Broad, heritable DNA/ RNA sequence which affect an organism's traits.

Genetics: study of heredity & variation in organisms

Genome: an organism’s total genetic content (full DNA sequence)

Genomics: study of organisms in terms of their genome

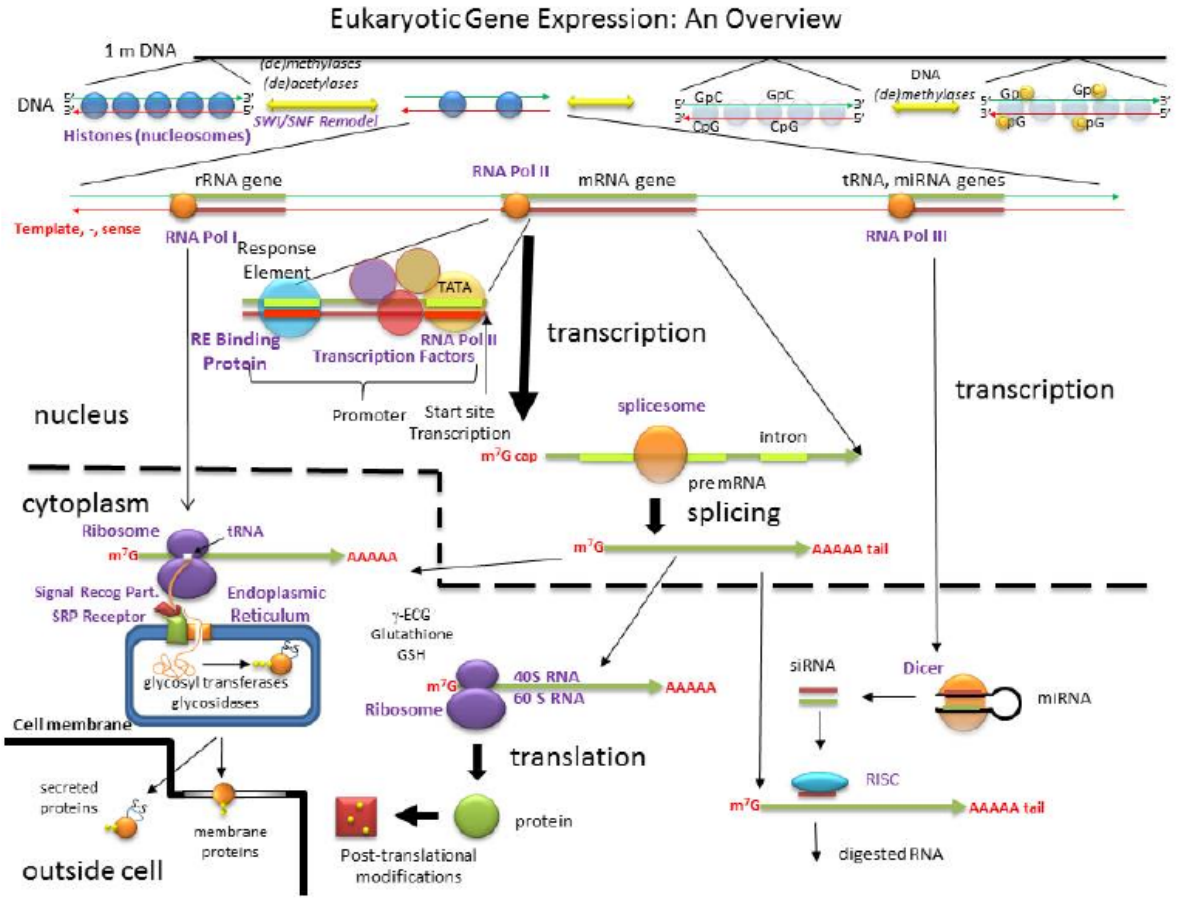
Review last lecture

- Gene mapping/Genetics association
- Common disease-common variant (CD-CV)
- Methods for identifying disease associated variants
- Challenges in sequencing-based genome-wide association study

What is gene expression profiling?

Gene expression profiling is the measurement of expression of thousands of genes at once, to create a global picture of cellular function.

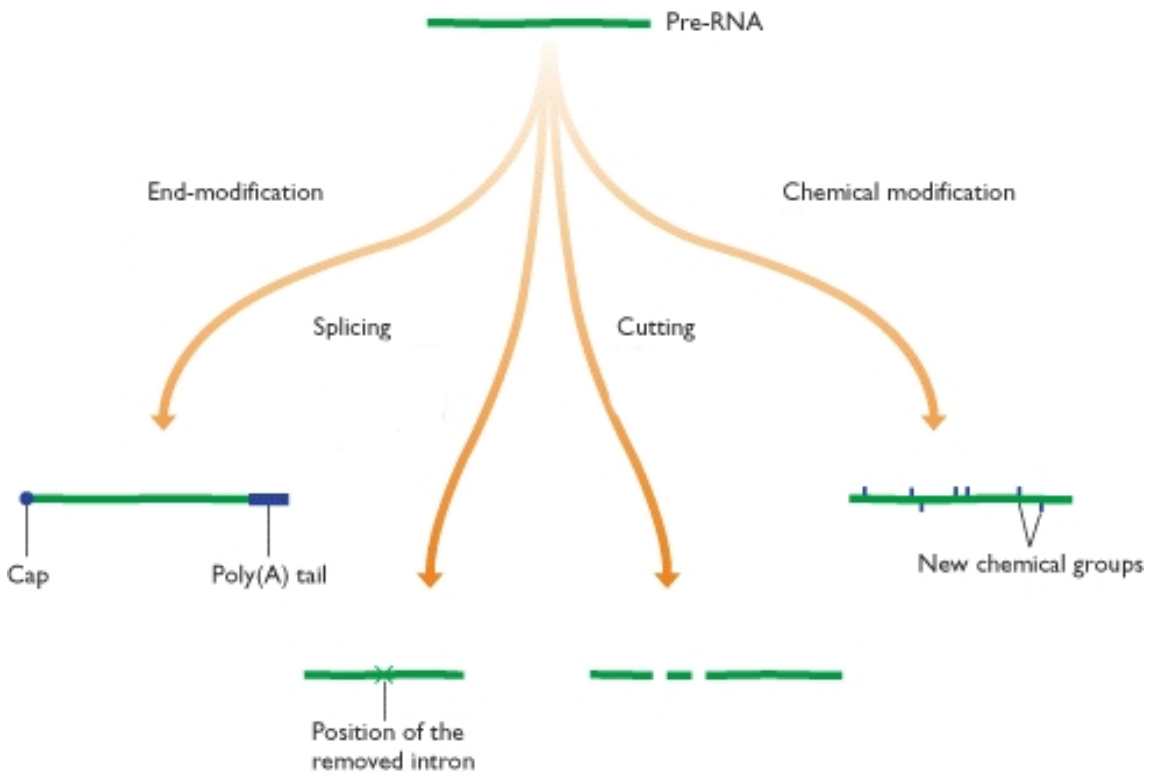
Eukaryotic Gene Expression: Overview



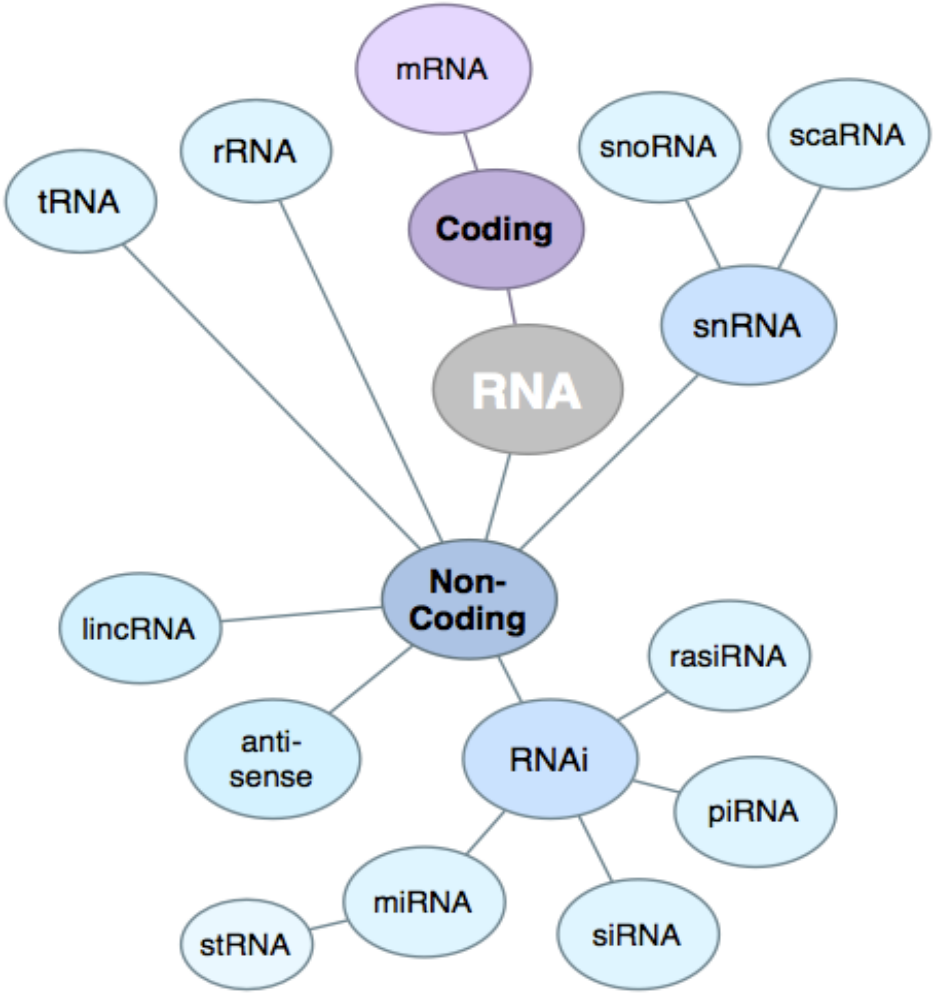
Graphics credit: CSBCJU; Biochemistry, Dr Jakubowski

http://employees.csbsju.edu/hjakubowski/classes/ch331/bind/olbindtransciption.html

Transcriptome summary



Transcriptome: RNA World



Significance of gene expression profiling

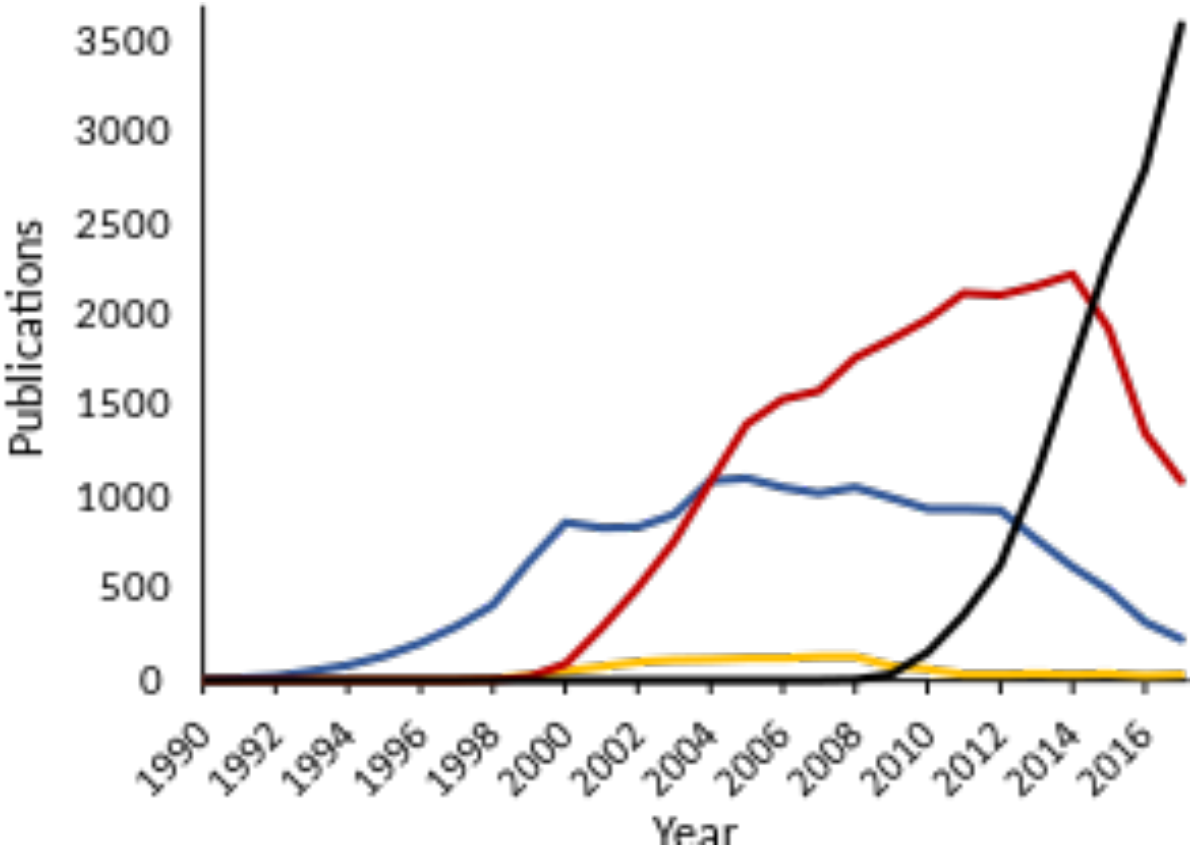
- Measuring the expression of an organism's genes in different tissues or conditions, or at different times, gives information on how genes are regulated and reveal details of an organism's biology. It can also be used to infer the functions of previously unannotated genes.

- Enable the study of how gene expression changes in different organisms and has been instrumental in the understanding of human disease. An analysis of gene expression in its entirety allows detection of broad coordinated trends which cannot be discerned by more targeted assays.

Methods for gene expression profiling

- Low throughput
  - Expressed sequence tag (EST)
    - a short sub-sequence of a cDNA sequence
- Early genome-wide technique
  - Serial analysis of gene expression (SAGE)
- Contemporary Techniques
  - Microarrays
  - RNA-seq

Methods usage over time



EST(blue)

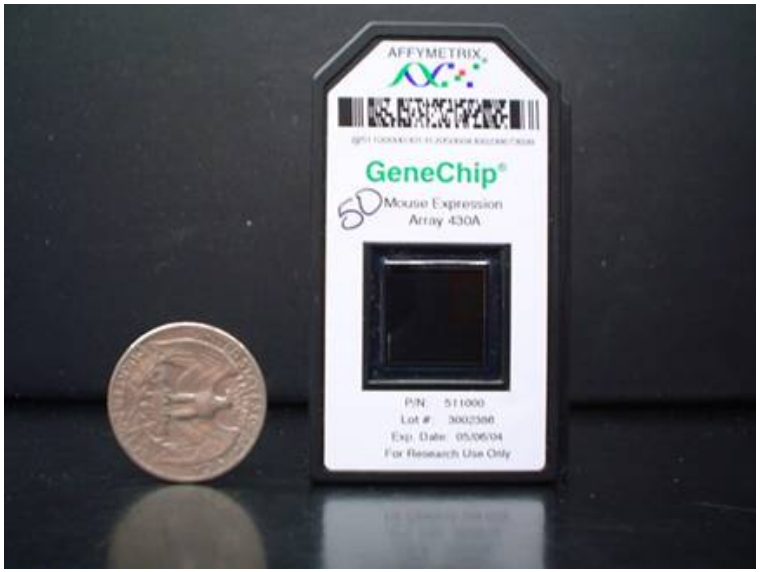
SAGE (yellow)

Microarray (red)

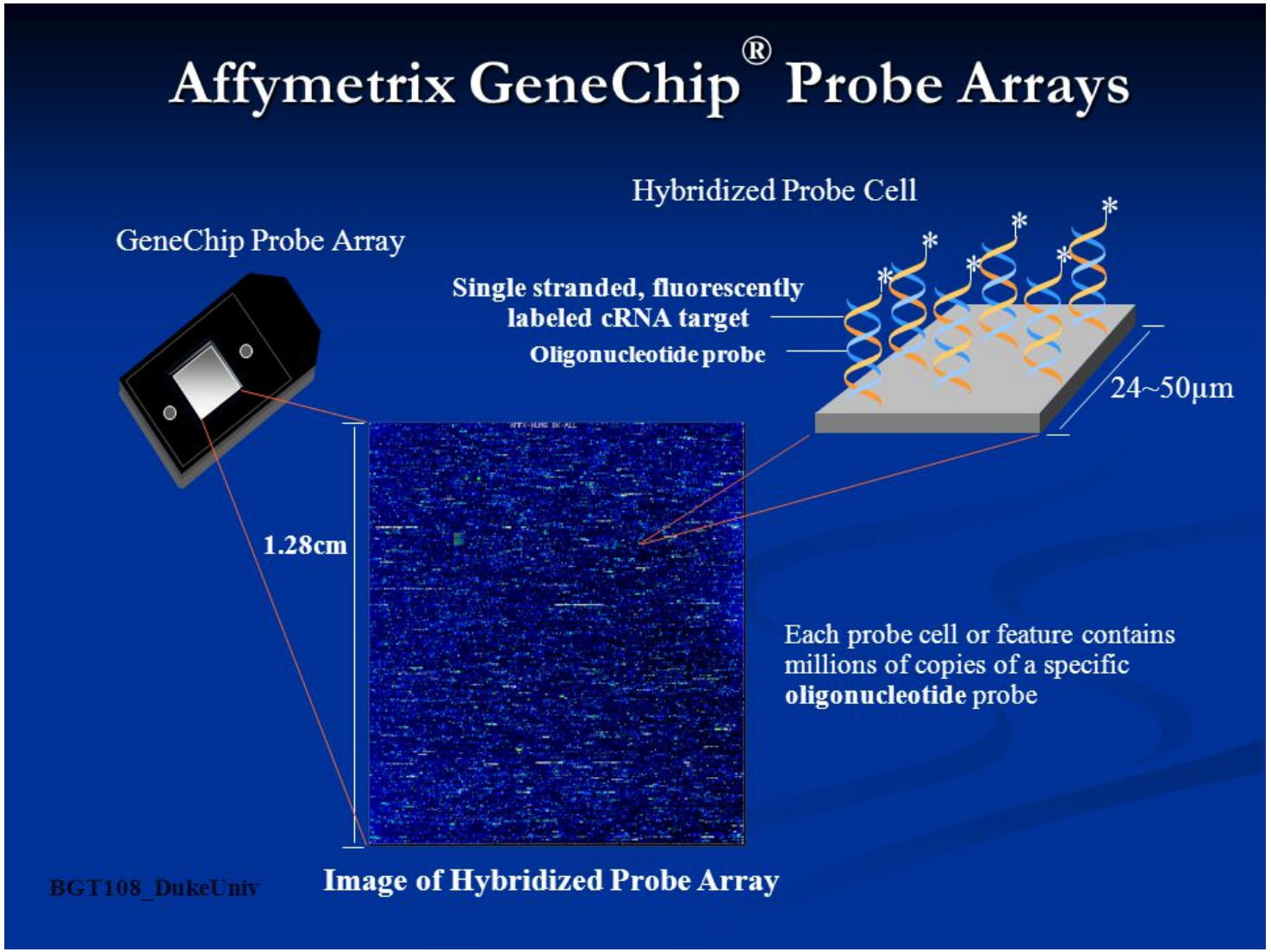
RNA-Seq (black)

What is a DNA Microarray?

- Also known as DNA Chip
- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression)



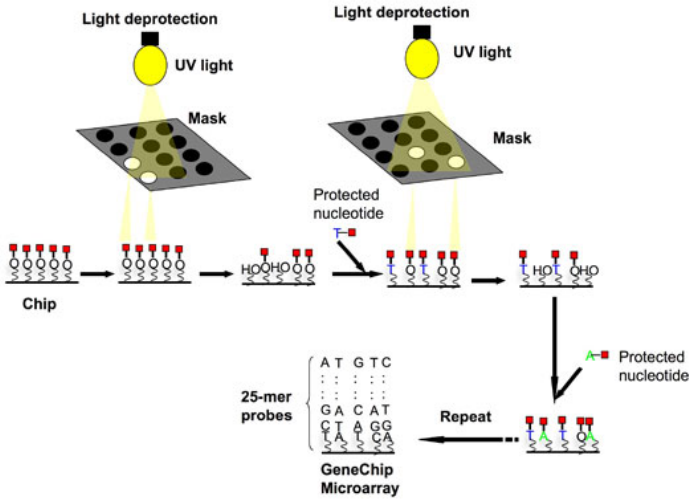
Zoom in microarray



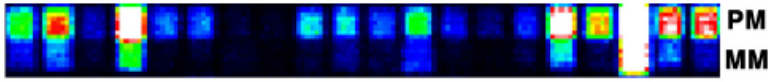
Applications

- Transcriptional profiling (search for DE genes)
- SNP genotyping
- DNA protein interaction (chip-Chip)
- Copy number variations (CNVs) detection
- Various gene detection

Affymetrix Technology

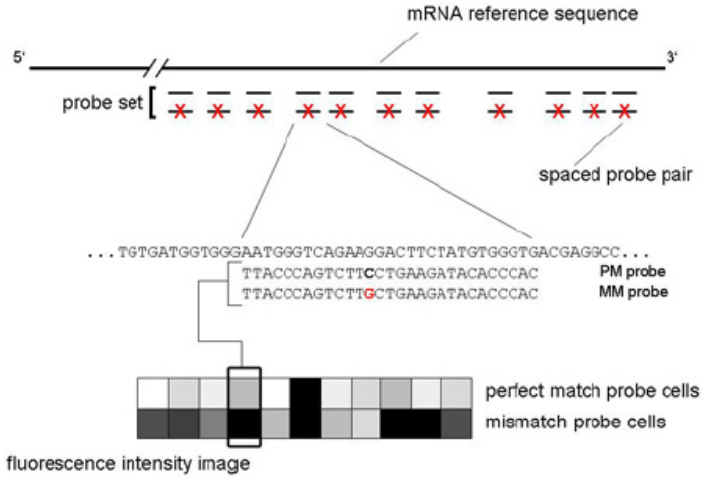


- The Affymetrix technology uses photolithographic synthesis of oligonucleotides on microarrays.
- The chip can hold up to 1.6 million features



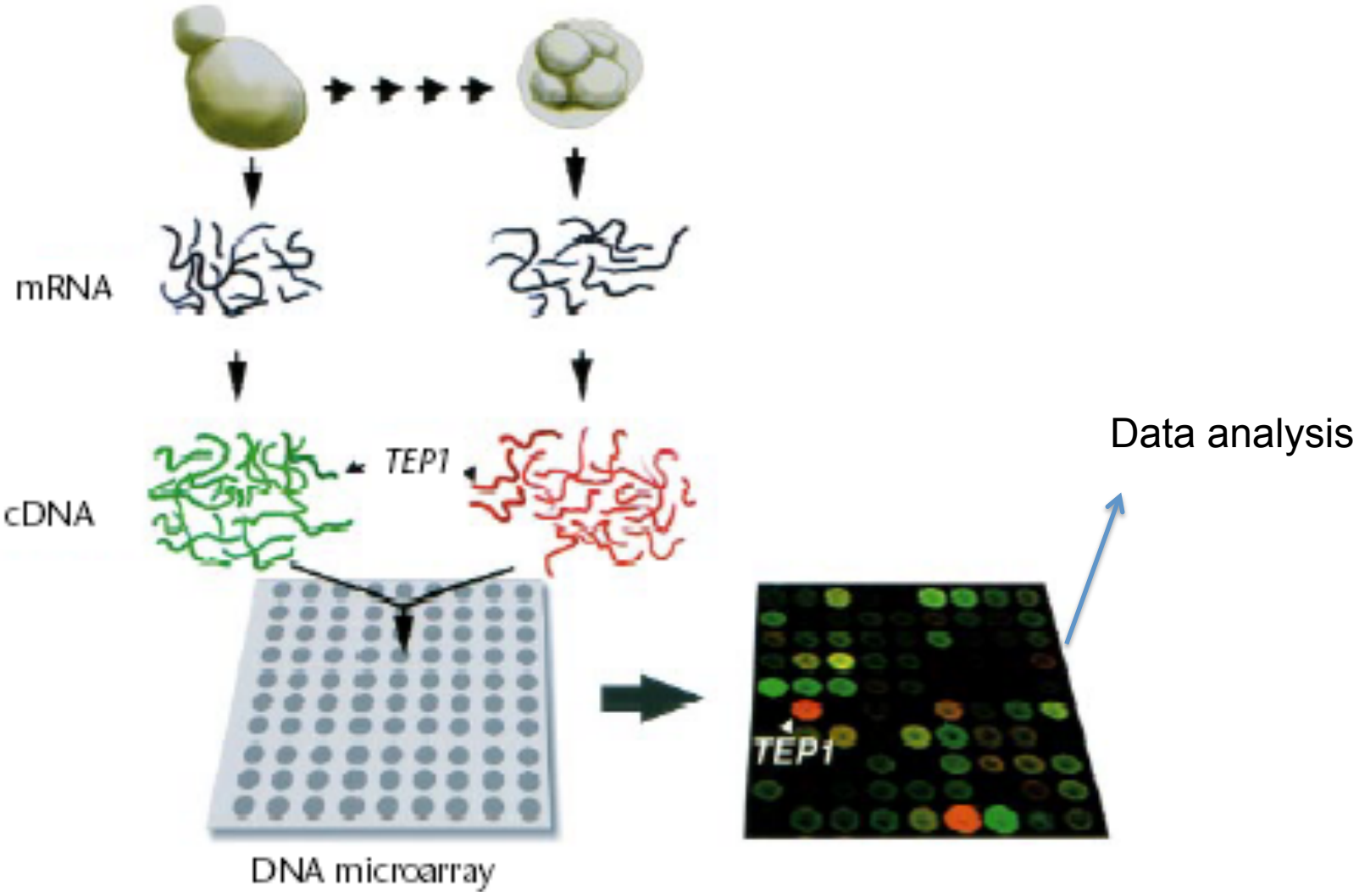
- Two 25-mer oligonucleotides make up one probe pair of a perfect match (PM) oligo and a corresponding mismatch (MM) oligo (mismatch at base 13)
- The probe pairs allow the quantization and subtraction of signals caused by non-specific cross-hybridization.
- PM - MM $\Rightarrow$ indicators of specific target abundance.

Affymetrix Technology



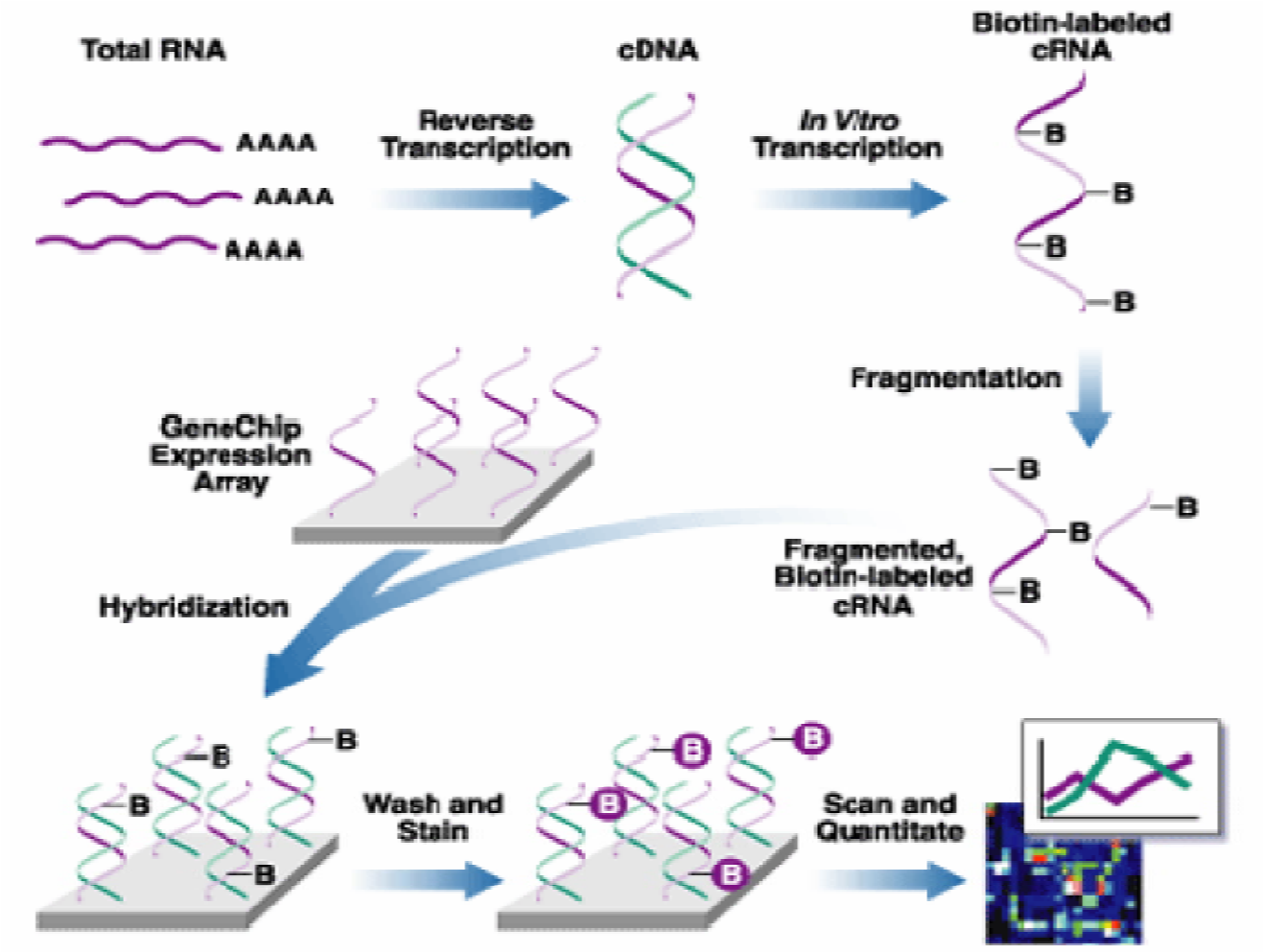
- The presence of messenger RNA (mRNA) is detected by a series of probes that differ in only one nucleotide.
- Hybridization of fluorescent mRNA to these probes on the chip is detected by laser scanning of the chip surface.
- A probe set consists 11 PM, MM pairs – the expression level is calculated by synthesizing information from all such PM/MM probes

Summary of expression array

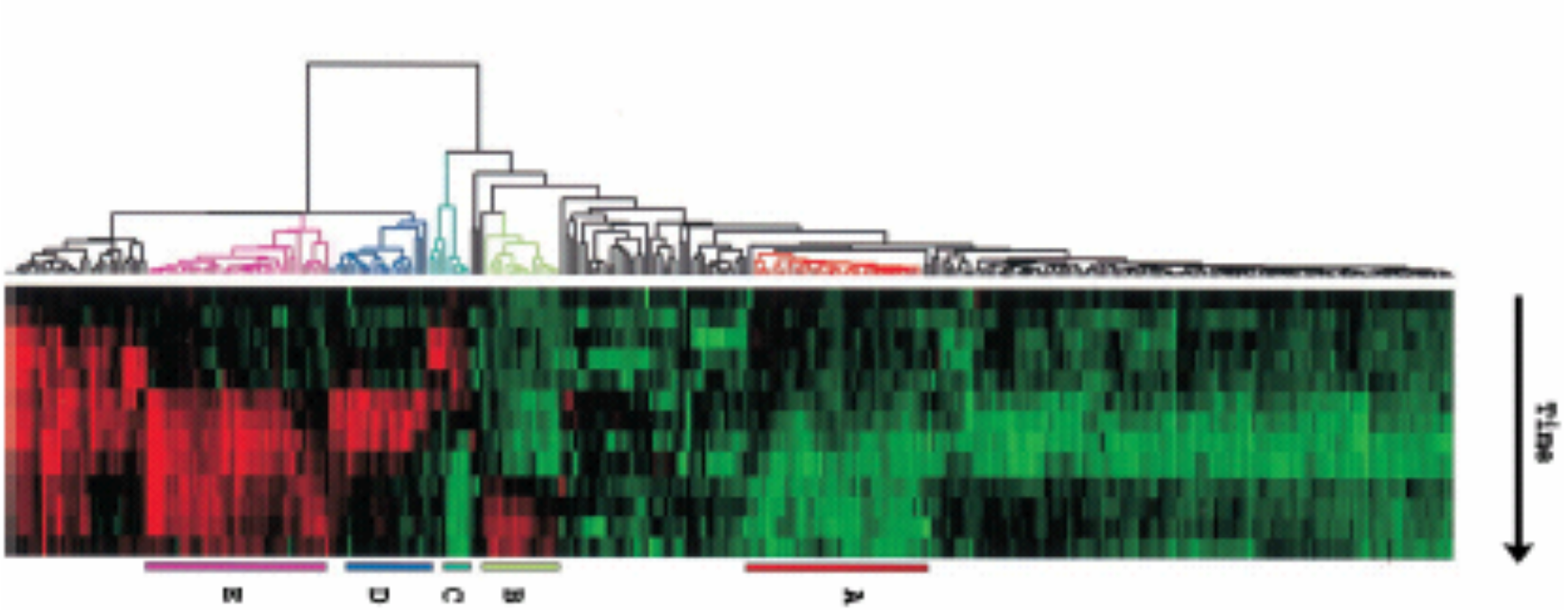


Brown & Botstein, 1999

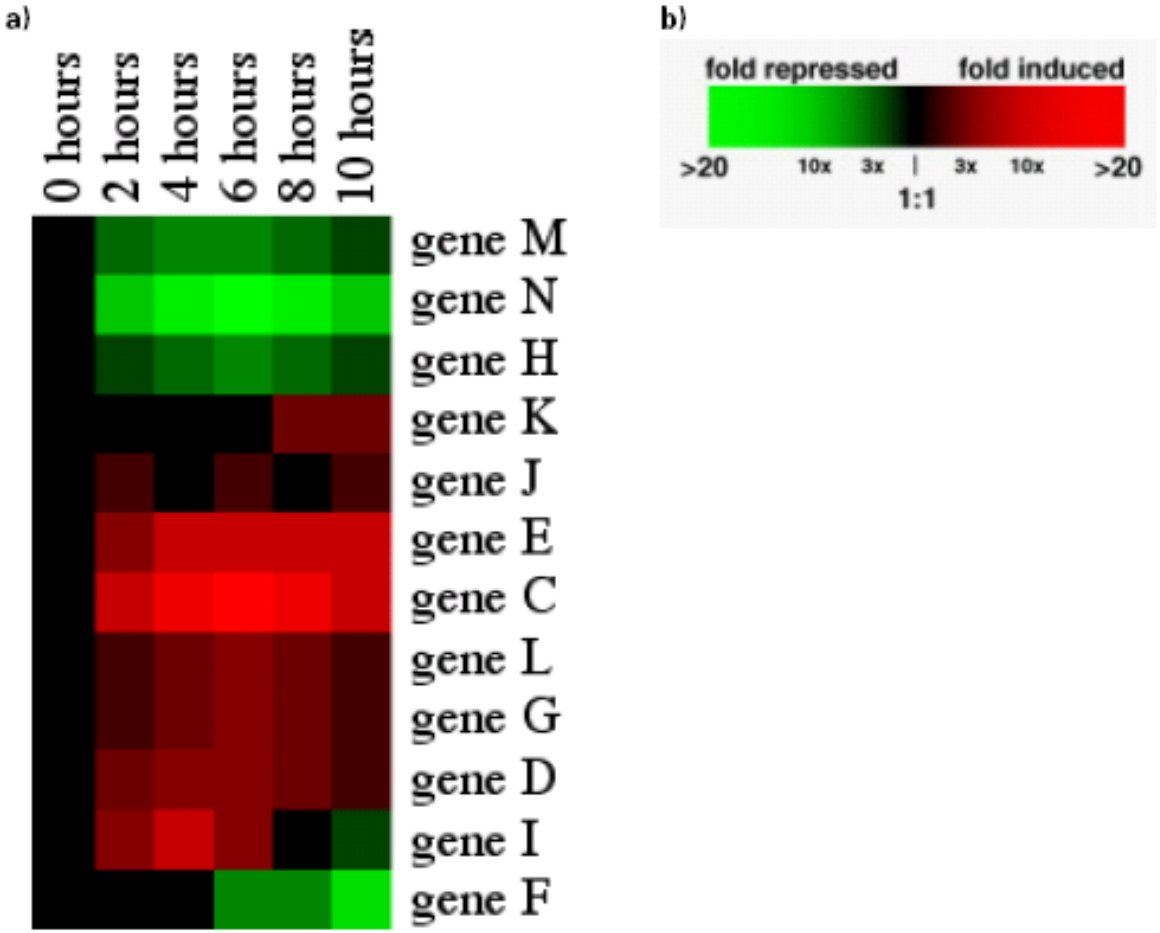
Steps of the Expression Assay



Sample clustering

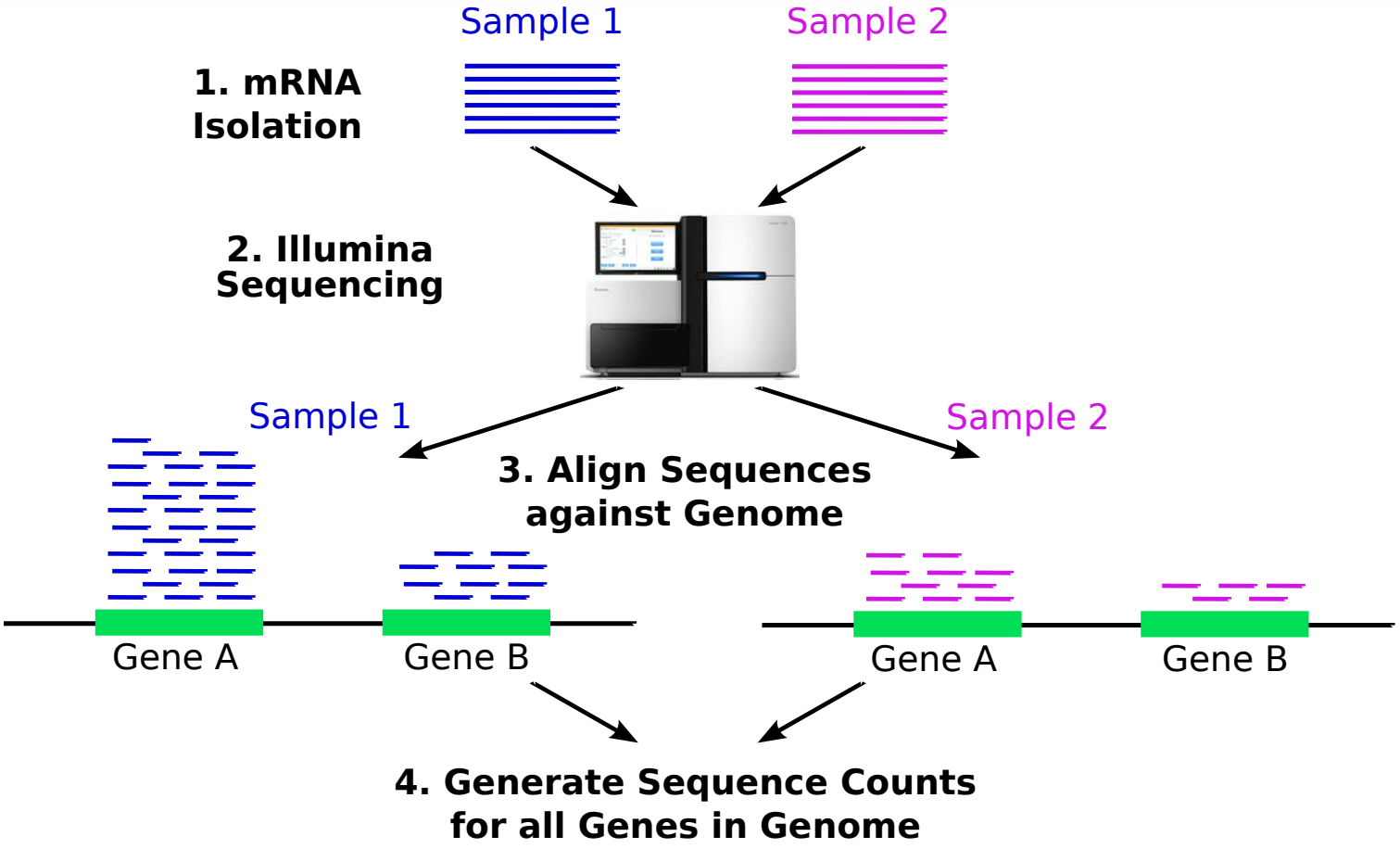


Change of gene expression



Campbell & Heyer, 2003

RNA-seq technology



RNA–seq

RNA-seq can be used for many different types of experiment

- Measuring gene expression
- Differential expression
- Detecting novel transcripts
- Splice junction analysis
- De novo assembly
- SNP analysis
- Allele specific expression
- RNA–editing
- Studying small/microRNAs

blue: (Nearly) impossible with microarrays
green: Requires special chip

Analysis Workflow of RNA-Seq Gene Expression Data

1. Alignment of RNA reads to reference
   - Reference can be genome or transcriptome.
2. Count reads overlapping with annotation features of interest
   - Most common: counts for exonic gene regions, but many viable alternatives exist here: counts per exons, genes, introns, etc.
3. Normalization
   - Main adjustment for sequencing depth and compositional bias.
4. Identification of Differentially Expressed Genes (DEGs)
   - Identification of genes with significant expression differences.
   - Identification of expressed genes possible for strongly expressed ones.
5. Specialty applications
   - Splice variant discovery (semi-quantitative), gene discovery, antisense expressions, etc.
6. Cluster Analysis
   - Identification of genes with similar expression profiles across many samples.
7. Enrichment Analysis of Functional Annotations
   - Gene ontology analysis of obtained gene sets from steps 5-6.

Major topics for RNA-seq analysis

- Mapping
- Reads count and Normalization
- Differential expression
- Alternative splicing
- Single cell RNA-seq

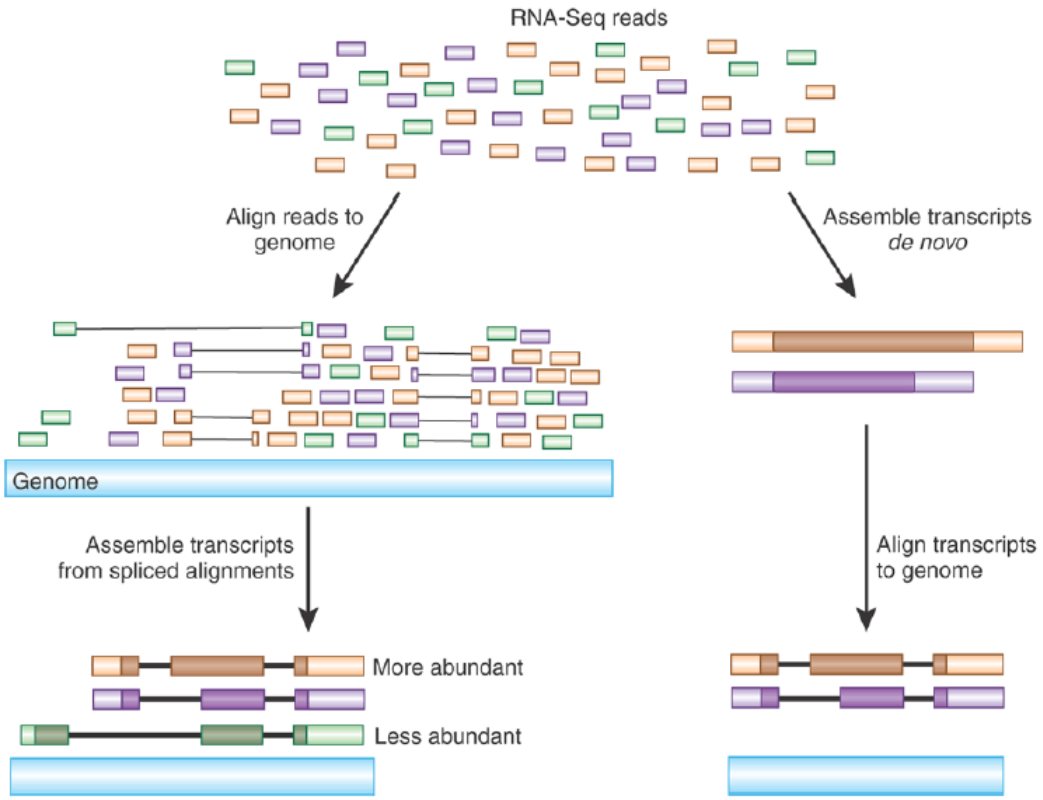
Mapping Reads to Transcriptome

One of the critical steps in an RNA-Seq experiment is that of mapping the NGS reads to the reference transcriptome. However, we still do not know all transcripts even for well studied species such as our own.

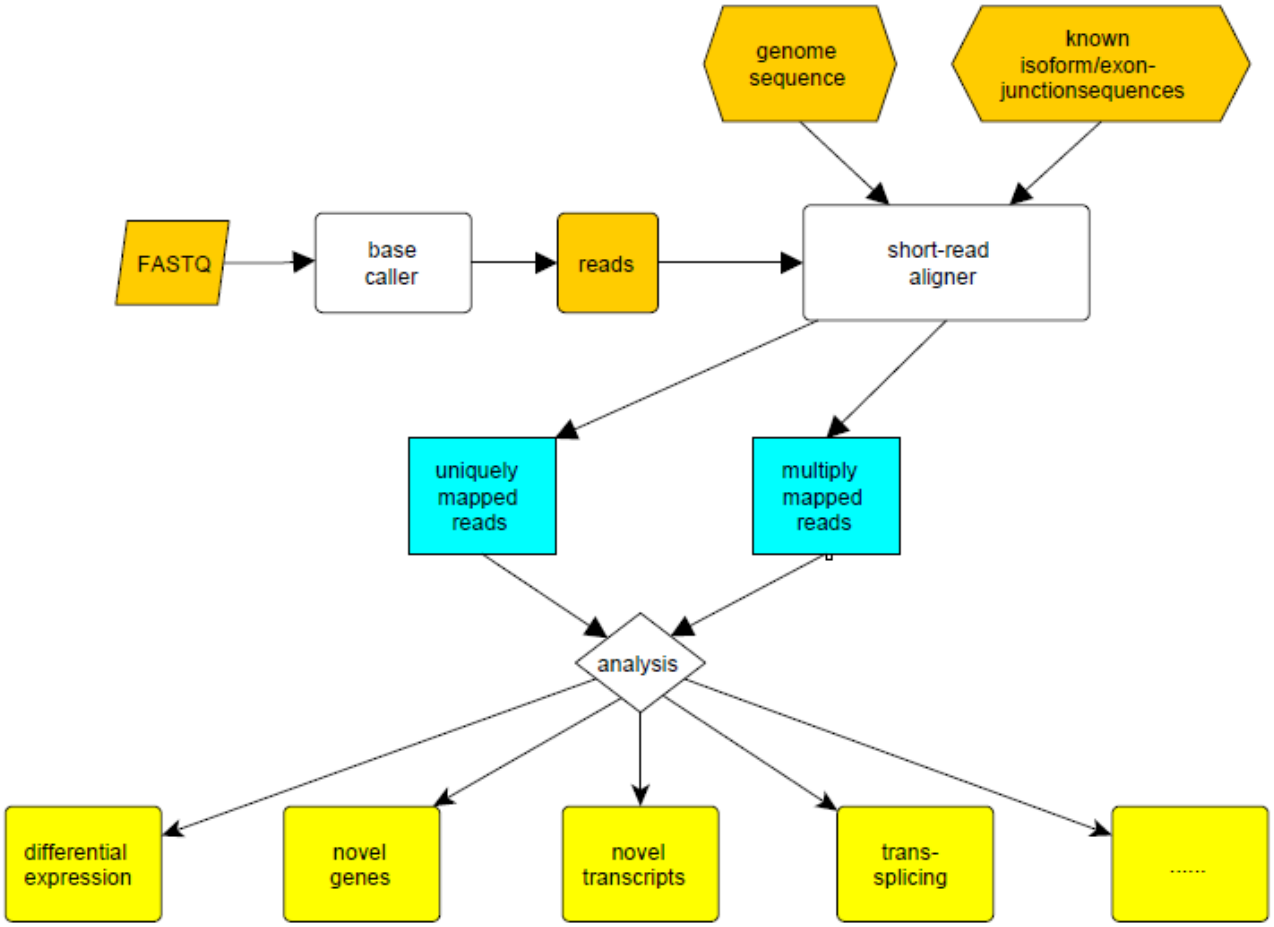
- RNA-Seq analyses are thus forced to map to the reference genome as a proxy for the transcriptome.
- Mapping to the genome achieves two major objectives of RNA-Seq experiments:
  1. Identification of novel transcripts from the locations of regions covered in the mapping.
  2. Estimation of the abundance of the transcripts from their depth of coverage in the mapping.

Two methods for mapping

General Bioinformatics Workflow to map transcripts from RNA-seq data



Multiple downstream applications...

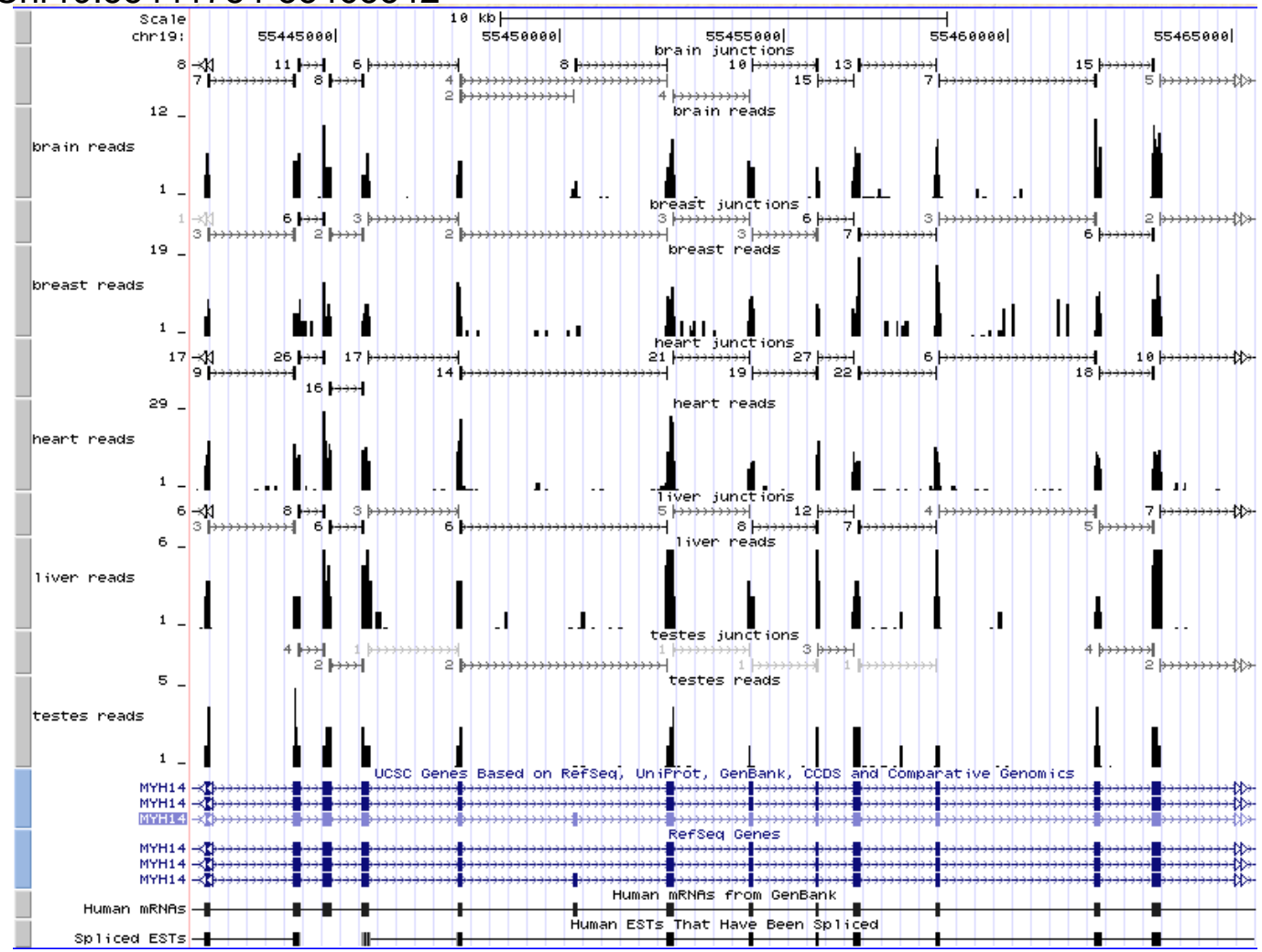


Major topics for RNA-seq analysis

- Mapping
- Reads count and Normalization
- Differential expression
- Alternative splicing
- Single cell RNA-seq

RNA−Seq on UCSC genome Browser

Chr19:55441784-55465542



RNA–Seq on genome Browser

Need for Normalization

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

Quantifying Expression

- RPM: Reads Per Million

- RPKM: Reads Per Kilobase per Million reads
- FPKM: Fragments Per Kilobase per Million fragments

- TPM: Transcripts Per Kilobase per Million

RPM/CPM

Reads per million/Counts per million

$$RPM = C/N$$

- C : Number of mappable reads on a feature (eg. transcript, exon, etc.)
- N: Total number of mappable reads (in millions)

CPM/RPM = NumReads /(totalNumReads/1,000,000)

RPKM

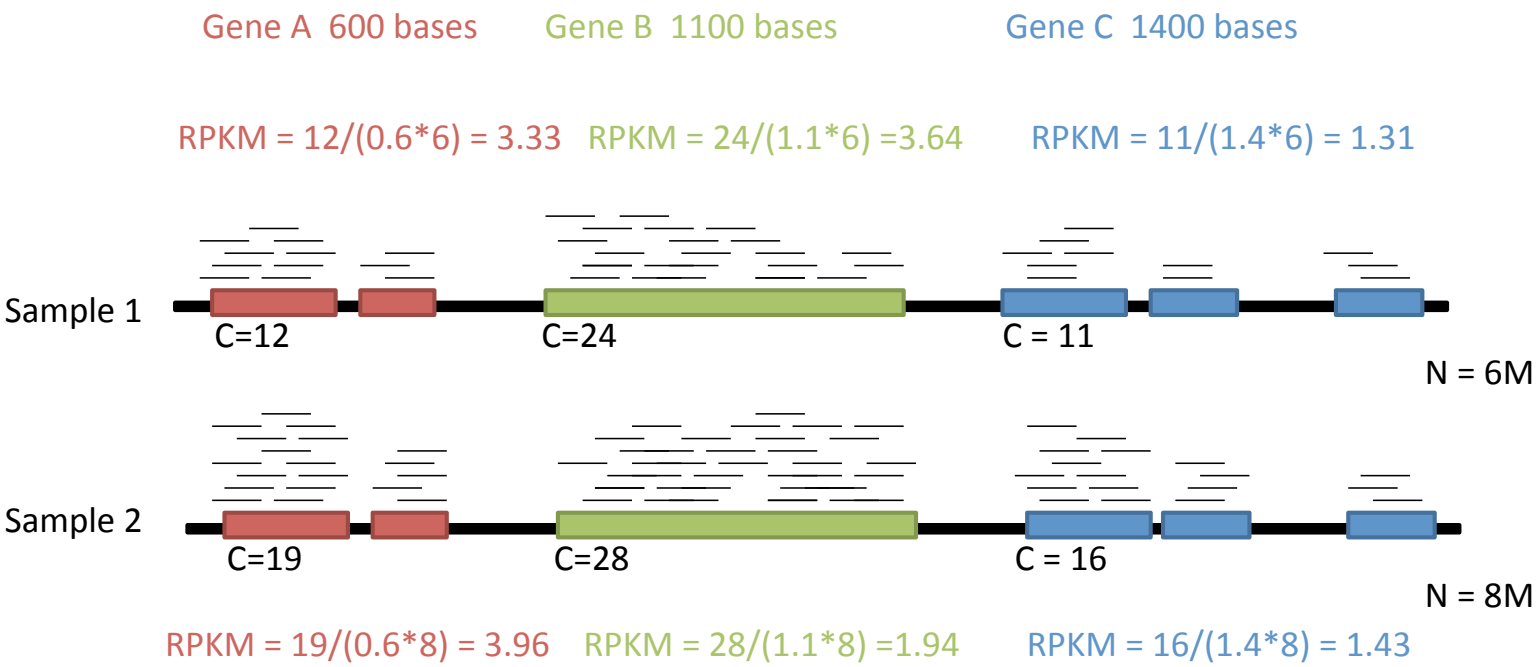
RPKM: Reads Per Kilobase per Million mapped reads

$$RPKM=C/(LN)$$

- C : Number of mappable reads on a feature (eg. transcript, exon, etc.)
- L: Length of feature (in kb)
- N: Total number of mappable reads (in millions)

RPKM = numReads / (Length/1000 * totalNumReads/1,000,000)

RPKM Example



Rewrite the formula

- $$RPKM = \#MappedReads * \frac{1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- Example 1:
  - 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):

Examples

- $$RPKM = \#MappedReads * \frac{1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- Example 1:
  - 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):
    - RPKM=$900 * \frac{1000 * 10^6}{2500 * 8.10^6} = 45$

FPKM

RPKM: Fragments Per Kilobase per Million mapped fragments

- FPKM is very similar to RPKM. RPKM was made for single-end RNA-seq.FPKM was made for paired-end RNA-seq.

- The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn’t count this fragment twice).

TPM

TPM: Transcripts Per Million

$$TPM = (C*L_r*10^6) / (L* T)$$

$$T = \Sigma C*L_r/L$$

1. C: Number of mappable reads on a feature (eg. transcript, exon, etc.)
2. L_r: Read length
3. L: Length of feature (in kb)
4. T: Total number of transcripts sampled in a sequencing run

TPM considers the gene length for normalization
TPM does not count for total number of mapped reads
TPM proposed as an alternative to RPKM due to inaccuracy in RPKM measurement (Wagner et al., 2012)
TPM is suitable for sequencing protocols where reads sequencing depends on gene length

The sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample.

TPM and RPKM/FPKM

$$RPKM = TPM * (T * 10^3) / (N * L_r)$$

1. T: Total number of transcripts sampled in genome
2. N: Total number of mappable reads
3. L_r: Read length

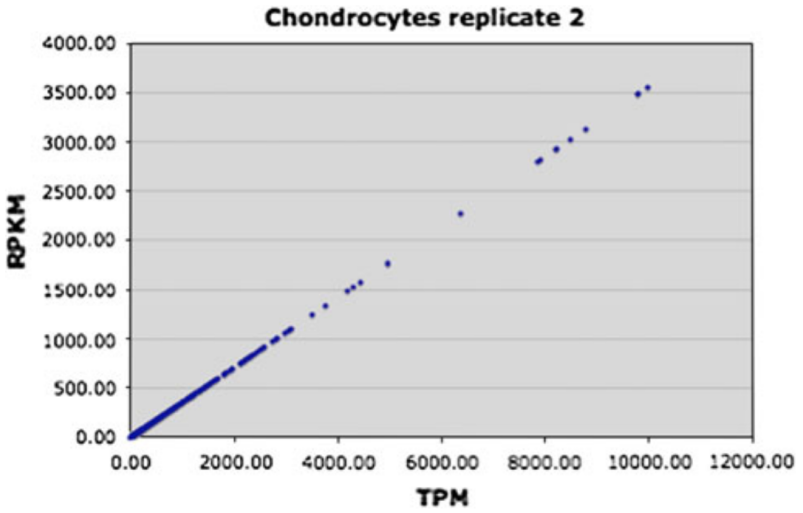


Fig. 1 Relationship between RPKM and TPM in data from RNA abundance in cultured human chondrocytes (ATCC, Cat. No. CRL-2847). RPKM and TPM are proportional to each other within a given sample, but see Table 1 for variation between samples

Wagner et al., 2012

TPM reduce false positive

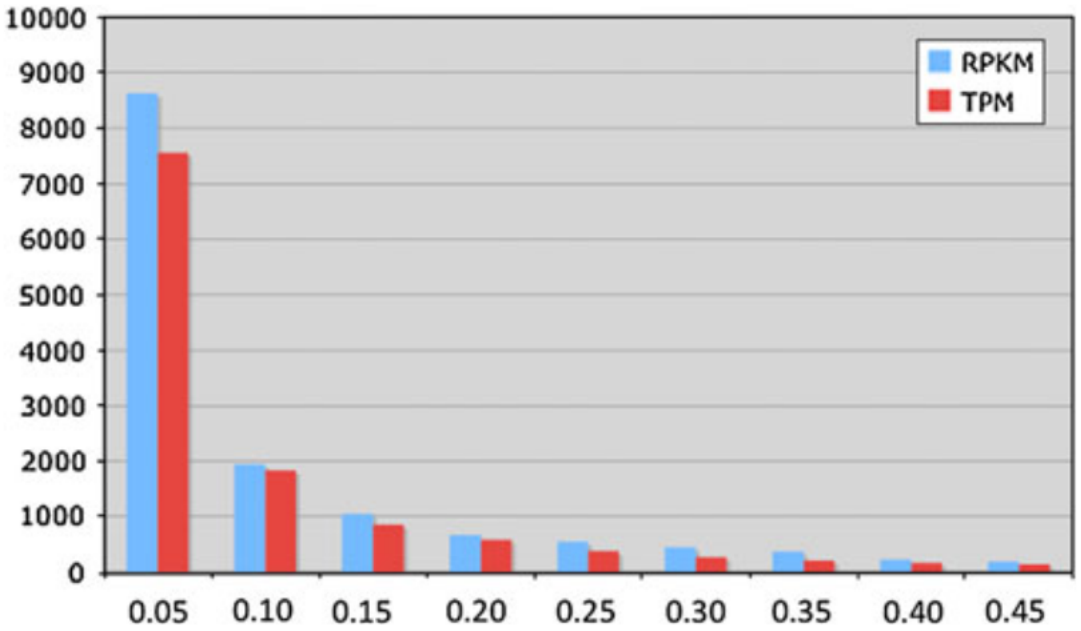


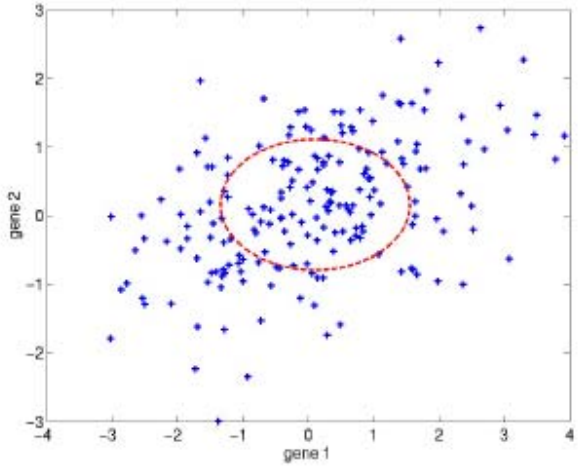
Fig. 2 *p* value distribution of RNA abundance data from human chondrocytes and myometrial cells for data expressed in RPKM and TPM. The *p* values were calculating from two-tailed *t* test assuming different variances. The *p* values are binned in 0.05 bins. Note that *t* tests using RPKM lead to higher number of low *p* values as expected if RPKM introduces artifactual differences in RNA abundance measures between samples

Major topics for RNA-seq analysis

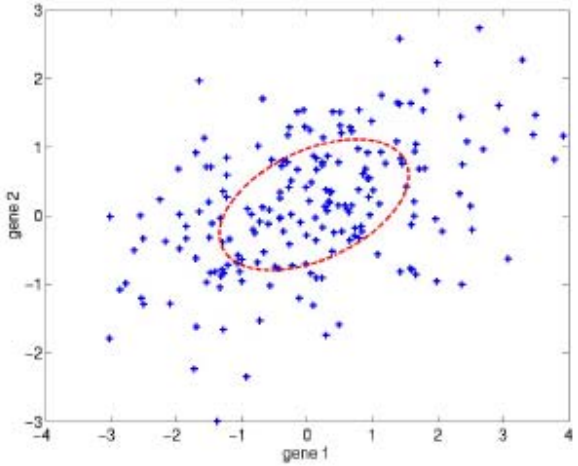
- Mapping
- Reads count and Normalization
- Differential expression
- Alternative splicing
- Single cell RNA-seq

Statistical tests: example

- The alternative hypothesis H_1 is more expressive in terms of explaining the observed data



null hypothesis



alternative hypothesis

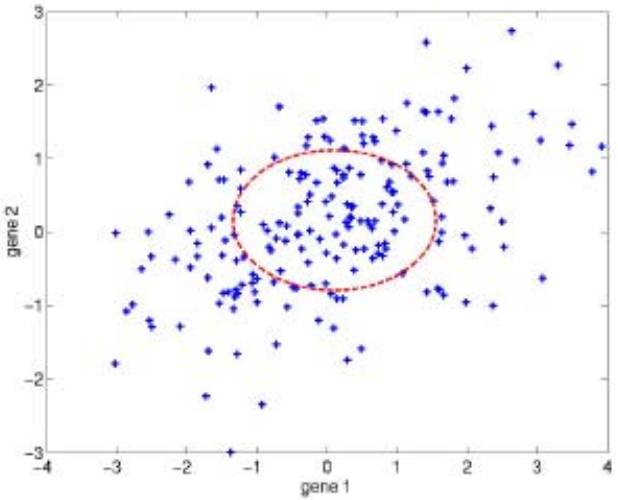
- We need to find a way of testing whether this difference is significant

Degrees of freedom

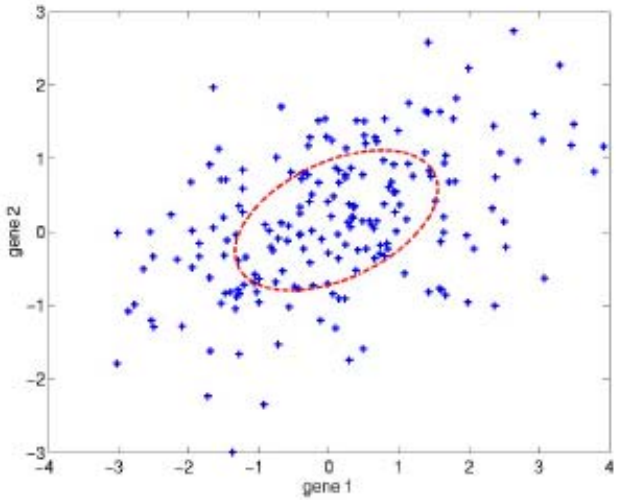
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



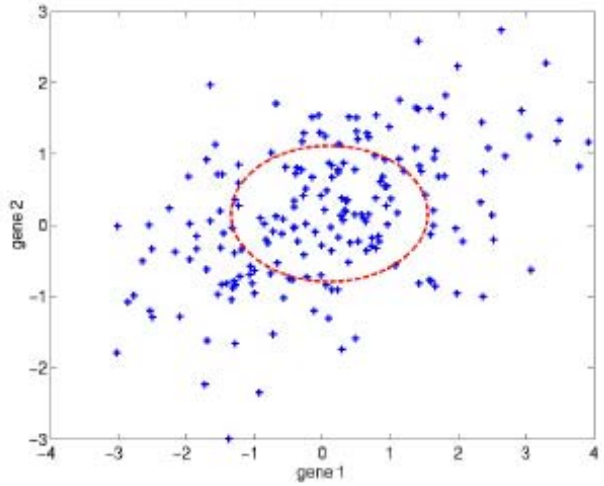
H_1

Degrees of freedom

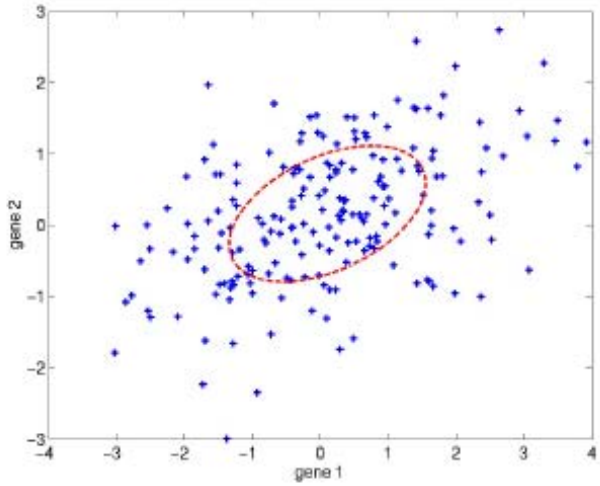
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \qquad (1)$$

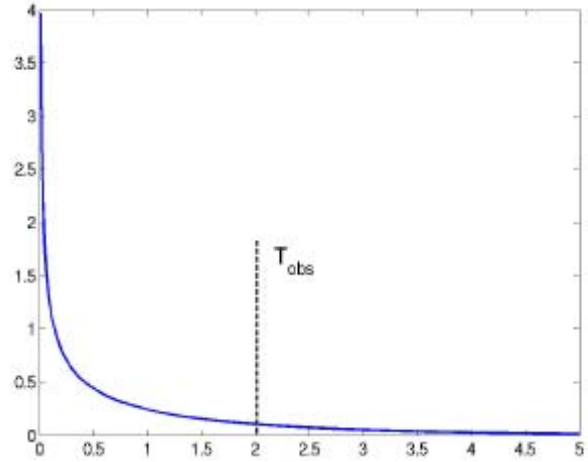
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis H_0 is correct

Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



- Once we know the sampling distribution, we can compute the P-value

$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \qquad (2)$$

Scaling RNA-seq data (DESeq)

- i gene or isoform
- j sample (experiment)
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- s_j sampling depth for experiment j (scale factor)

$$s_j = \underset{i}{median} \frac{K_{ij}}{\left(\prod_{v=1}^{m} K_{iv}\right)^{1/m}}$$

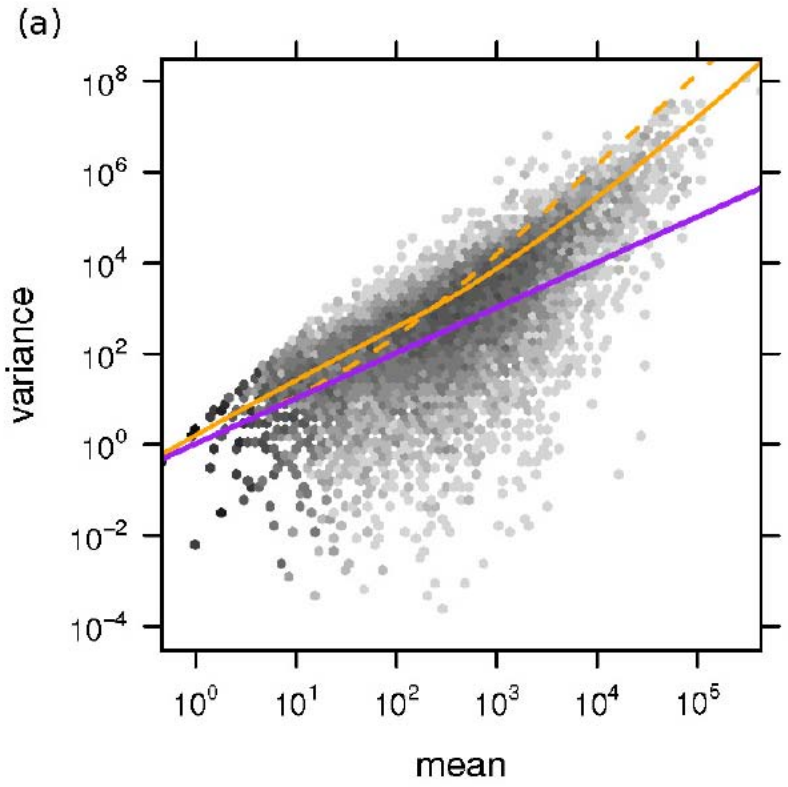
Model for RNA-seq data (DESeq)

- i gene or isoform p condition
- j sample (experiment) p(j) condition of sample j
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- q_{ip} Average scaled expression for gene i condition p

$$q_{ip} = \frac{1}{\#\text{ of replicates}} \sum_{\text{j in replicates}} \frac{K_{ij}}{s_j}$$

$$\mu_{ij} = q_{ip(j)} s_j \qquad \sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

Source: Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11, no. 10 (2010): R106.

Significance of differential expression using test statistics

- Hypothesis H0 (null) – Condition A and B identically express isoform i with random noise added
- Hypothesis H1 – Condition A and B differentially express isoform
- Degrees of freedom (dof) is the number of free parameters in H1 minus the number of free parameters in H0; in this case degrees of freedom is 4 – 2 = 2 (H1 has an extra mean and variance).
- Likelihood ratio test defines a test statistic that follows the Chi Squared distribution

$$T_i = 2\log\frac{P(K_{iA} \mid H1)P(K_{iB} \mid H1)}{P\left(K_{iA}, K_{iB} \mid H0\right)}$$

$$P\left(H0\right) \approx 1 - ChiSquaredCDF\left(T_i \mid dof\right)$$

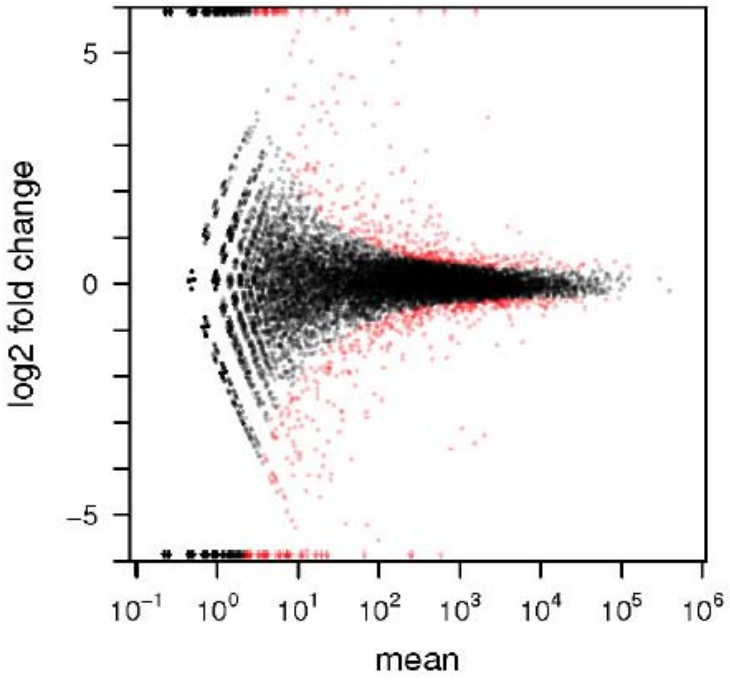


Figure 3 Testing for differential expression between conditions *A* and *B*: Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at the upper and lower plot border indicate genes with very large or infinite log fold change. The corresponding volcano plot is shown in Supplementary Figure S8 in Additional file 2.

Source: Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11, no. 10 (2010): R106.

Hypergeometric test for overlap significance

N – total # of genes	1000
n1 - # of genes in set A	20
n2 - # of genes in set B	30
k - # of genes in both A and B	3

$$P(k) = \frac{\binom{n1}{k}\binom{N-n1}{n2-k}}{\binom{N}{n2}}$$

0.017

$$P(x \geq k) = \sum_{i=k}^{\min(n1,n2)} P(i)$$

0.020

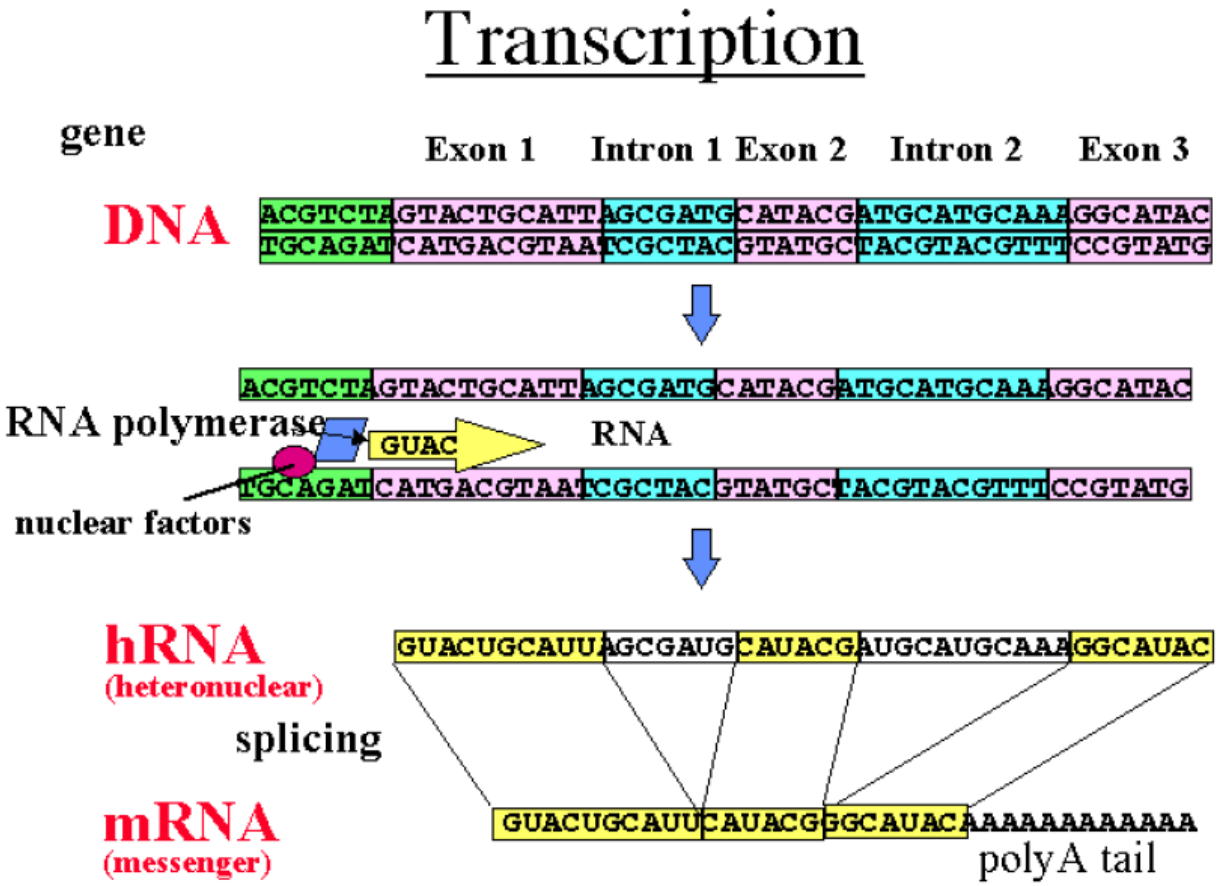
Software for RNA-Seq DEG Analysis

- edgeR (Robinson et al., 2010)
- DESeq/DESeq2 (Anders and Huber, 2010)
- DEXSeq (Anders et al., 2012)
- limmaVoom
- Cuffdiff/Cuffdiff2 (Trapnell et al., 2013)
- PoissonSeq
- baySeq
- ...

Major topics for RNA-seq analysis

- Mapping
- Reads count and Normalization
- Differential expression
- Alternative splicing
- Single cell RNA-seq

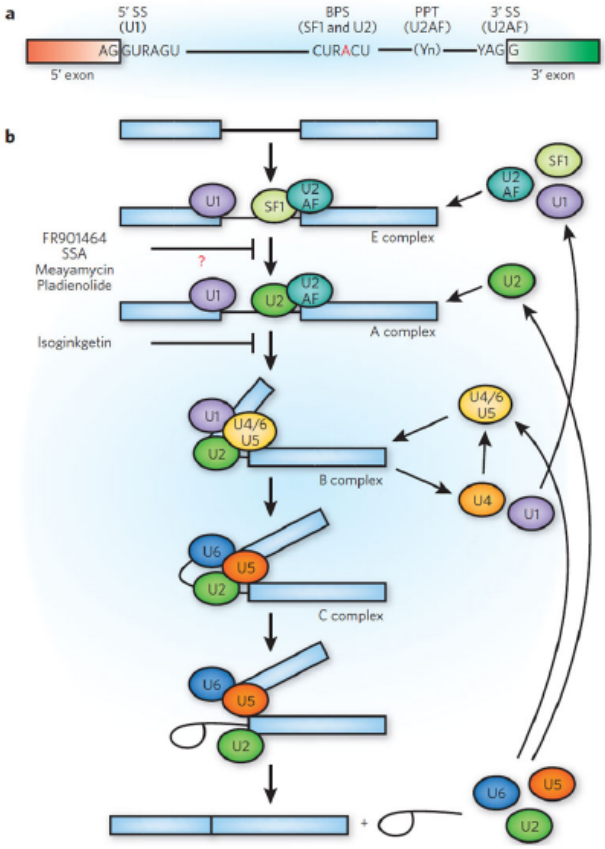
Splicing (review)



You should know (or review) general concepts of transcription, pre-RNA (near synonym to "heteronuclear RNA"), spliceosome, splicing

Splicing (review)

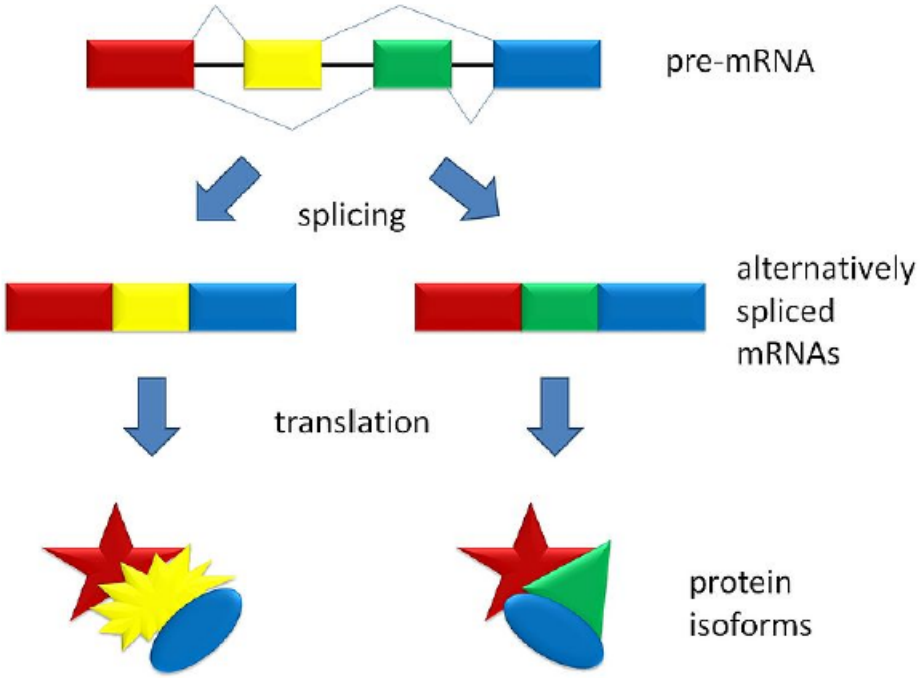
- A spliceosome is a complex of snRNA and protein subunits
- A spliceosome removes introns from a transcribed pre-mRNA (hnRNA) segment.



Schneider-Poetsch et al (2010) *Nature Chemical Biology* **6**:189–198

Alternative splicing

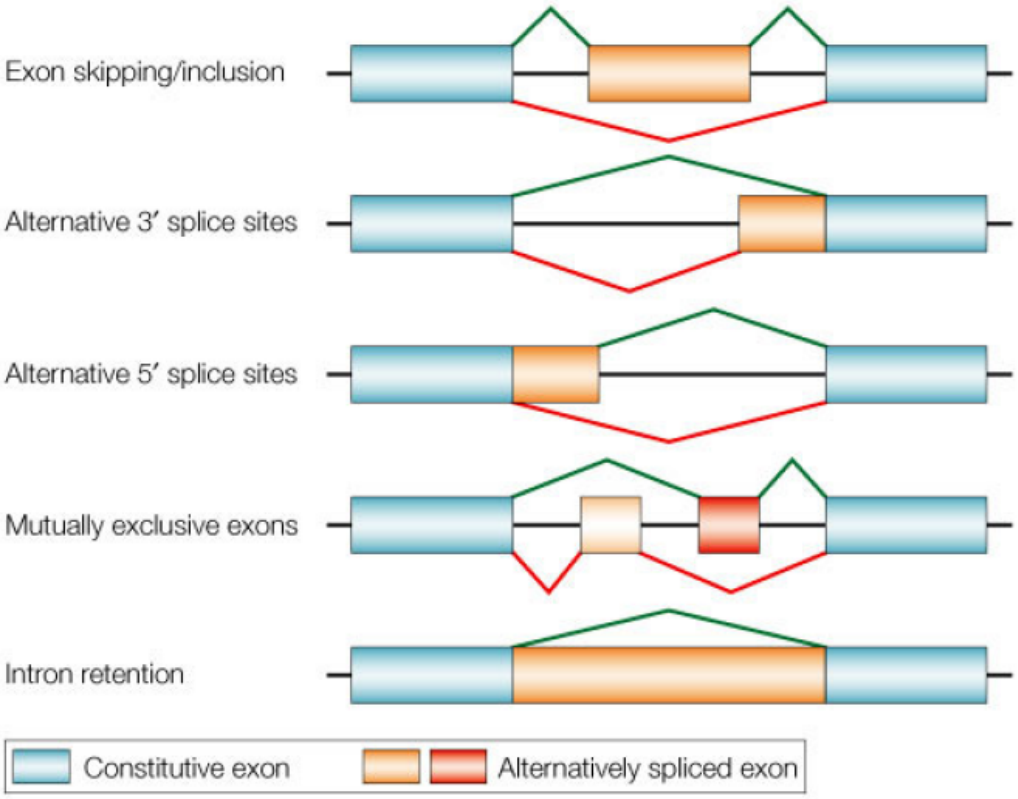
Single gene coding for multiple proteins. Each distinct splicing is known as an `isoform` or `transcript` of the gene.



graphic credit: wikipedia

Alternative splicing

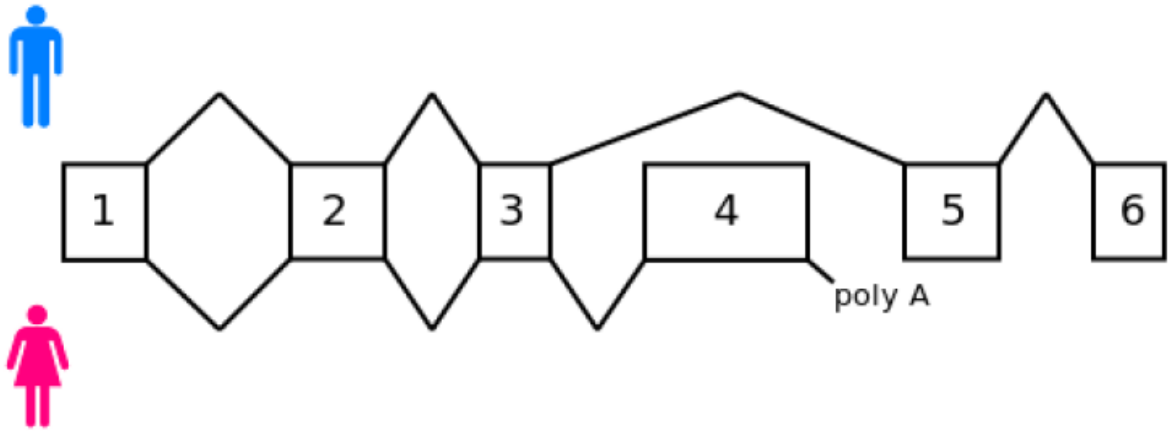
Several different classes of alternative splicing events



Cartegni L et al. (2002) *Nature Reviews Genetics* **3**:285–298

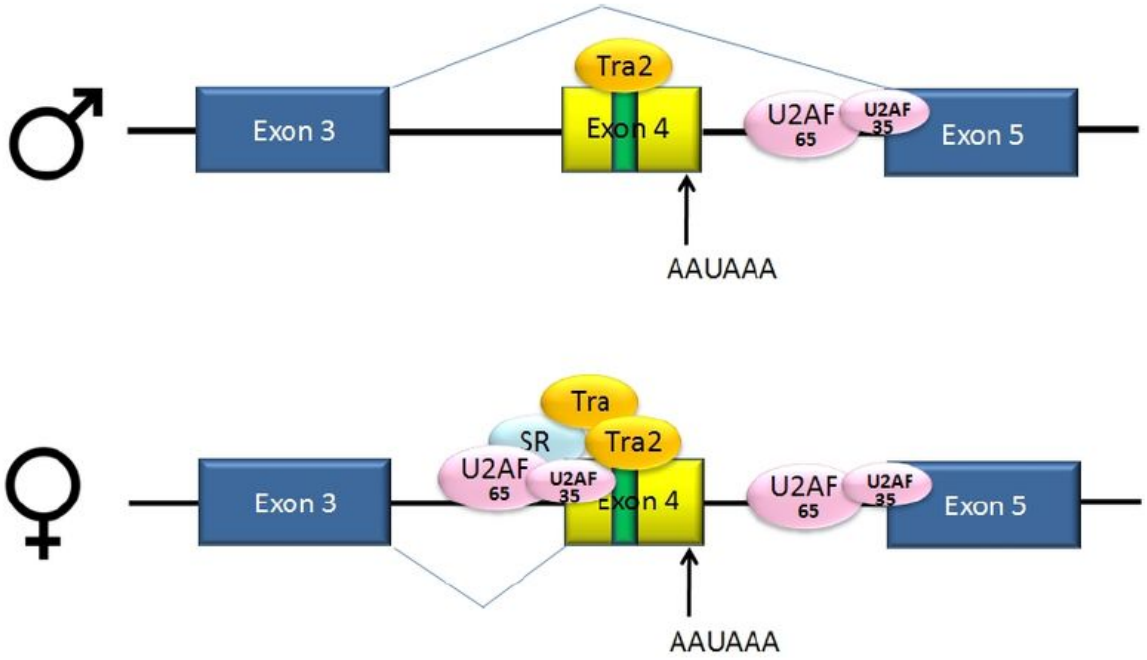
Alternative splicing: Biological roles

The different isoforms of a gene can have quite distinct functional roles. Here we see the *Drosophila dsx* gene.



- Males: exons 1–3,5–6 $\Rightarrow$ transcriptional regulatory protein required for male development.
- Females: exons 1–4 $\Rightarrow$ transcriptional regulatory protein required for female development

Alternative splicing: Regulation

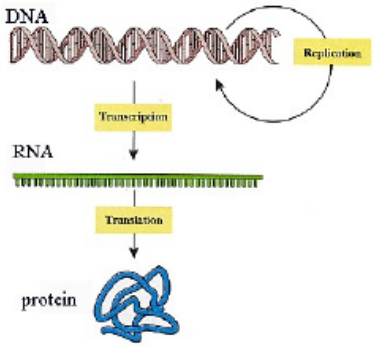


graphics credit: wikipedia

- The intron upstream from exon 4 has a polypyrimidine tract that doesn't match the consensus sequence well, so that U2AF proteins bind poorly to it without assistance from splicing activators. This 3' splice acceptor site is therefore not used in males.
- In general, we are just beginning to understand the regulatory mechanisms responsible for alternative splicing

Alternative splicing: Regulation

The **central dogma** of molecular biology...is thus slightly dodgy

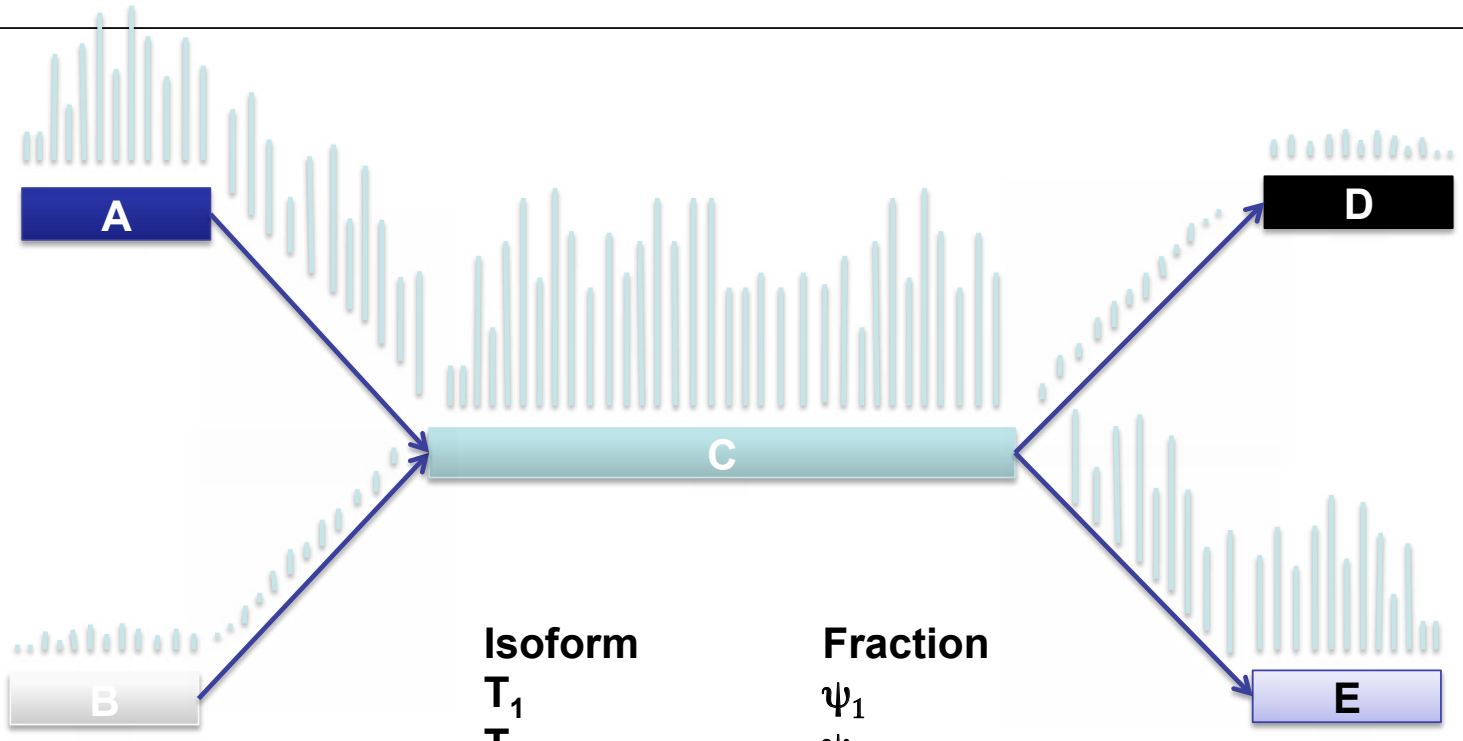


- Instead: One gene – many polypeptides
- Several proteins can be encoded by a single gene, rather than requiring a separate gene for each, and thus allowing a more varied proteome from a genome of limited size.
- Evolutionary flexibility. ("change just one isoform at a time")

Alternative splicing and RNA-seq

- In the rest of this lecture, we will therefore discuss how one might investigate alternative splicing with RNA-seq
- There are by now a multitude of methods and algorithms, each with particular focuses, strengths, and weaknesses.

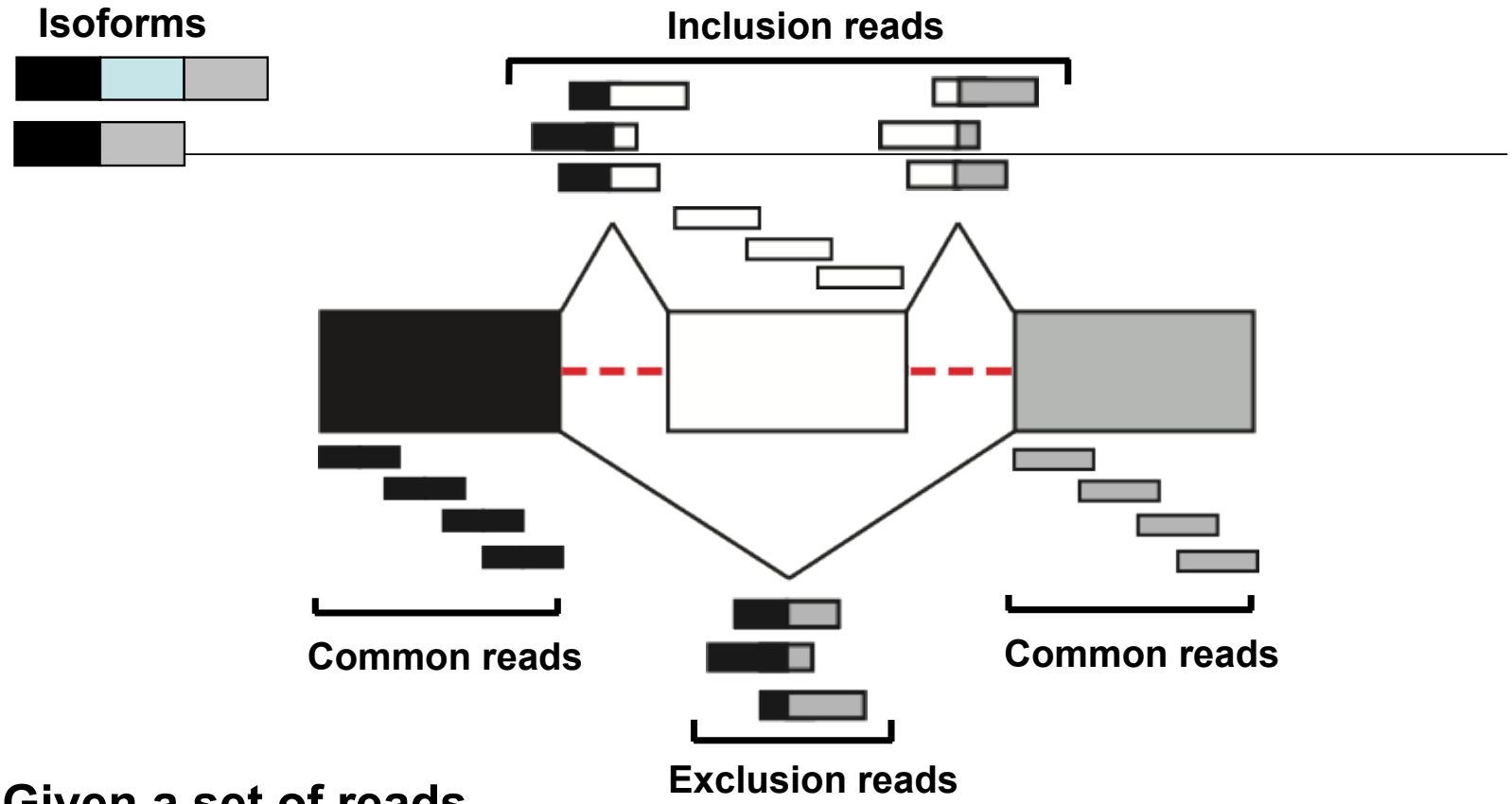
We can use mapped reads to learn the isoform mixture ψ



Isoform	Fraction
T_1	ψ_1
T_2	ψ_2
T_3	ψ_3
T_4	ψ_4

Slide courtesy Cole Trapnell

Detecting alternative splicing from mRNA-Seq data



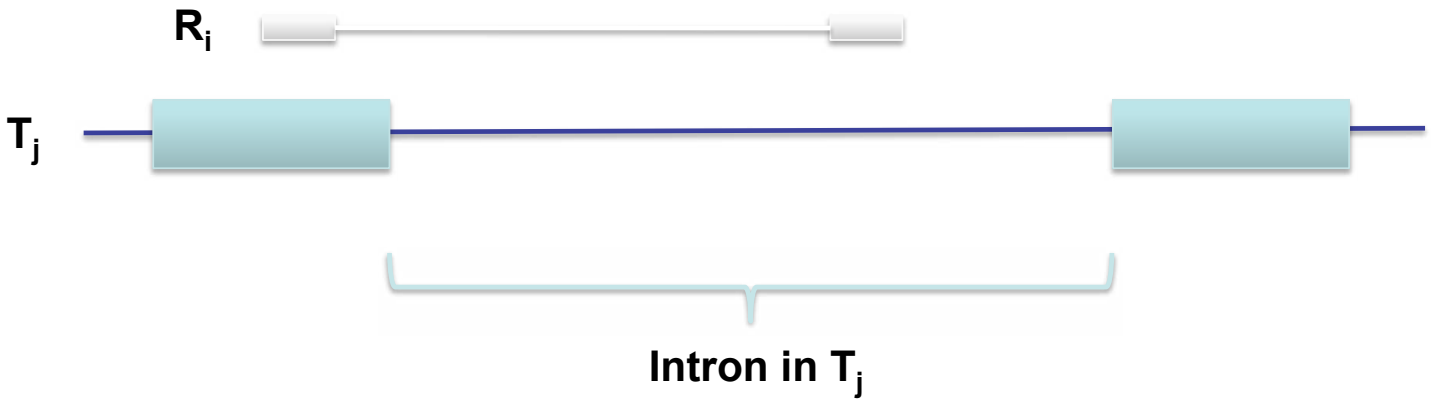
Given a set of reads, estimate:

Ψ = **Distribution of isoforms**

$P(R_i \mid T=T_j)$ – Excluded reads

If a single ended read or read pair R_i is structurally incompatible with transcript T_j, then

$$P(R = R_i \mid T = T_j) = 0$$



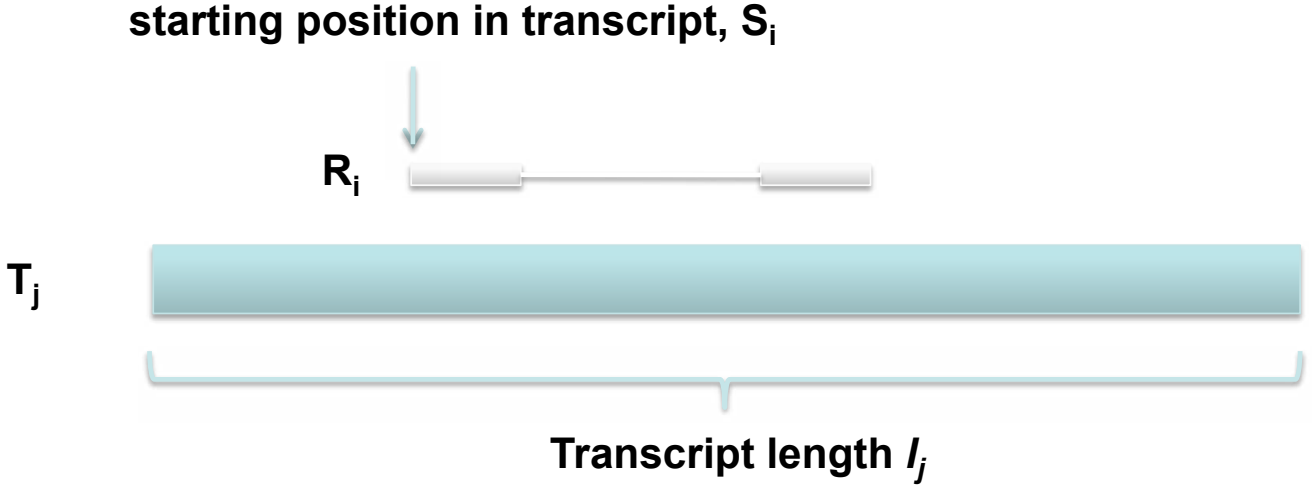
Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – Single end reads

Cufflinks assumes that fragmentation is roughly uniform. The probability of observing a fragment starting at a specific position S_i in a transcript of length l_j is:

$$P(S = S_i \mid T = T_j) = \frac{1}{l_j}$$



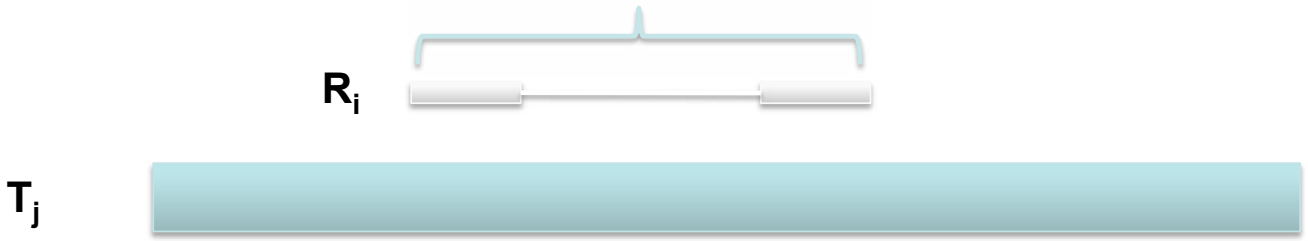
Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – Paired end reads

Assume our library fragments have a length distribution described by a probability density F. Thus, the probability of observing a particular paired alignment to a transcript:

$$P(R = R_i \mid T = T_j) = \frac{F(l_j(R_j))}{l_j}$$

Implied fragment length $l_j(R_i)$



Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnell

Estimating Isoform Expression

- Find expression abundances $\psi_1,\ldots,\psi_n$ for a set of isoforms $T_1,\ldots,T_n$
- Observations are the set of reads $R_1,\ldots,R_m$

$$P(R \mid \Psi) = \prod_{i=0}^{m} \sum_{j=0}^{n} \Psi_j P(R = R_i \mid T = T_j)$$

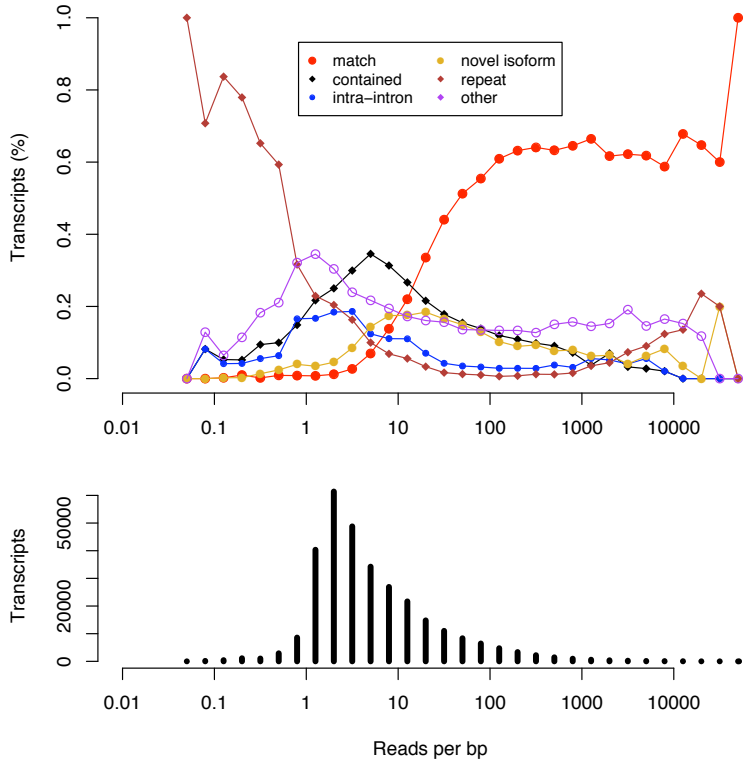
$$L(\Psi \mid R) \propto P(R \mid \Psi) P(\Psi)$$

$$\Psi = \underset{\Psi}{\operatorname{argmax}} L(\Psi \mid R)$$

- Can estimate mRNA expression of each isoform using total number of reads that map to a gene and ψ

Case study: myogenesis

Transcript categories, by coverage

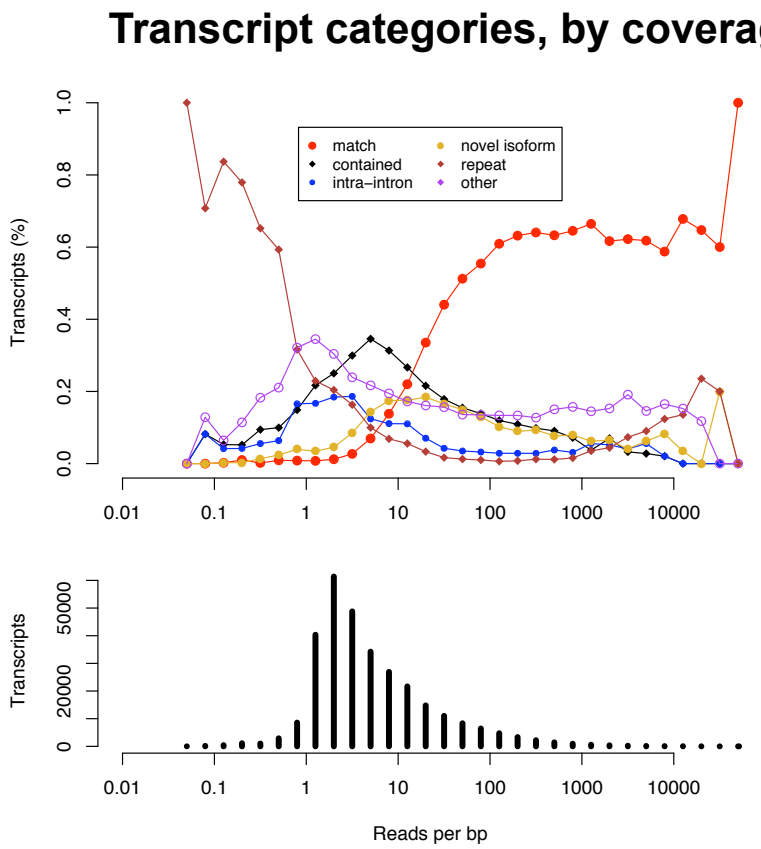


Courtesy of Cole Trapnell. Used with permission.

- Cufflinks identified 116,839 distinct transcribed fragments (transfrags)
- Nearly **70%** of the reads in 14,241 matching transcripts
- Tracked 8,134 transfrags across all time points, **5,845 complete matches** to UCSC/ Ensembl/VEGA
- Tracked **643** new isoforms of known genes across all points

Slide courtesy Cole Trapnell

Case study: myogenesis

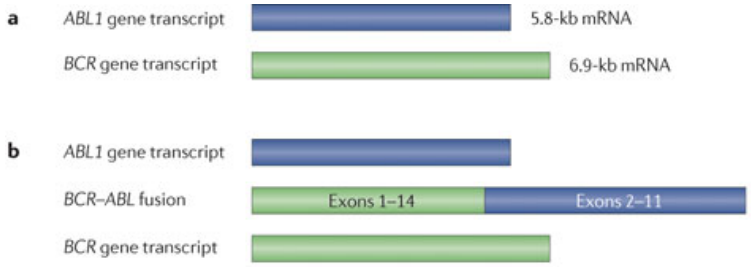


Courtesy of Cole Trapnell. Used with permission.

- ~25% of transcripts have light sequence coverage, and are fragments of full transcripts
- Intronic reads, repeats, and other artifacts are numerous, but account for less than 5% of the assembled reads.

Slide courtesy Cole Trapnell

Gene Fusion



c

```
GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGGAGGTGGATTCCTTTGGGTATTTT      BCR

AGGCATGGGGGTCCACACTGCAATGTTTTTGTGGAACATGAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAG      ABL1

GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAG      BCR–ABL fusion transcript


GTCAT-GTCCACTCAGC-ACTGGATT-AAGCAGAGTTCAAAAGC
 TCAT-GTCCACTCAGCCACTGGATTTAA-CAGAGTTCAAAAGC
       TCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCC
        CCACTCAGCTACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGC
             CAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGC
             CAGCCA-TGGATTTAAGC-GAGTTCAAAAGCCCTTCAG
              AGCCACTGGATTTAAGCAGAGTTCAAAAG
               GCCACTGGATTTAAGCAGAGTTCAAAAGCCCT
                 CACTGGATTTAAGCAGAGTTCAAAAGCCCTT
                    CTGGATTTAAGCAGAGTTCAAAAGC--TTCAGCGGC-AGTAG
                     TG-ATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGC
                        ATTTAAGCAGAGTTCAAAAGCCCTTCAGCG-CCAGTAGCA
                          TTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAG
                             AGCAGAGT-CAAAAGCCCTTC-GCGGCCAGTAGCATCTGACTTTGA-C
                             AT-A-AGTTCAAAAGCC-TTCAGCGGCCA-TAGCATCTG
                               CAGAGTTCAAAAGCCCTTCAGCGGCCAG
                               CAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTG
                                  AGTTCAAAAGCCCTTCAGCG-CCAGT-GCATCT
                                   GTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACT
                                     TCAAA-GCCCT-C-GCGGCCAGTAGCATCTGAC
                                     TCAAA-GCCCTTCAGCGGCCAGTAGCATCTGACTTTGAG
                                      CAAA-GCCCTTCAG-GGCCAGTAGCATCTGACTTTGAGCCTCAG
                                       AAAAGCCCTTCAG-GGCCAGTAGCATCTGACTTTGAG
```

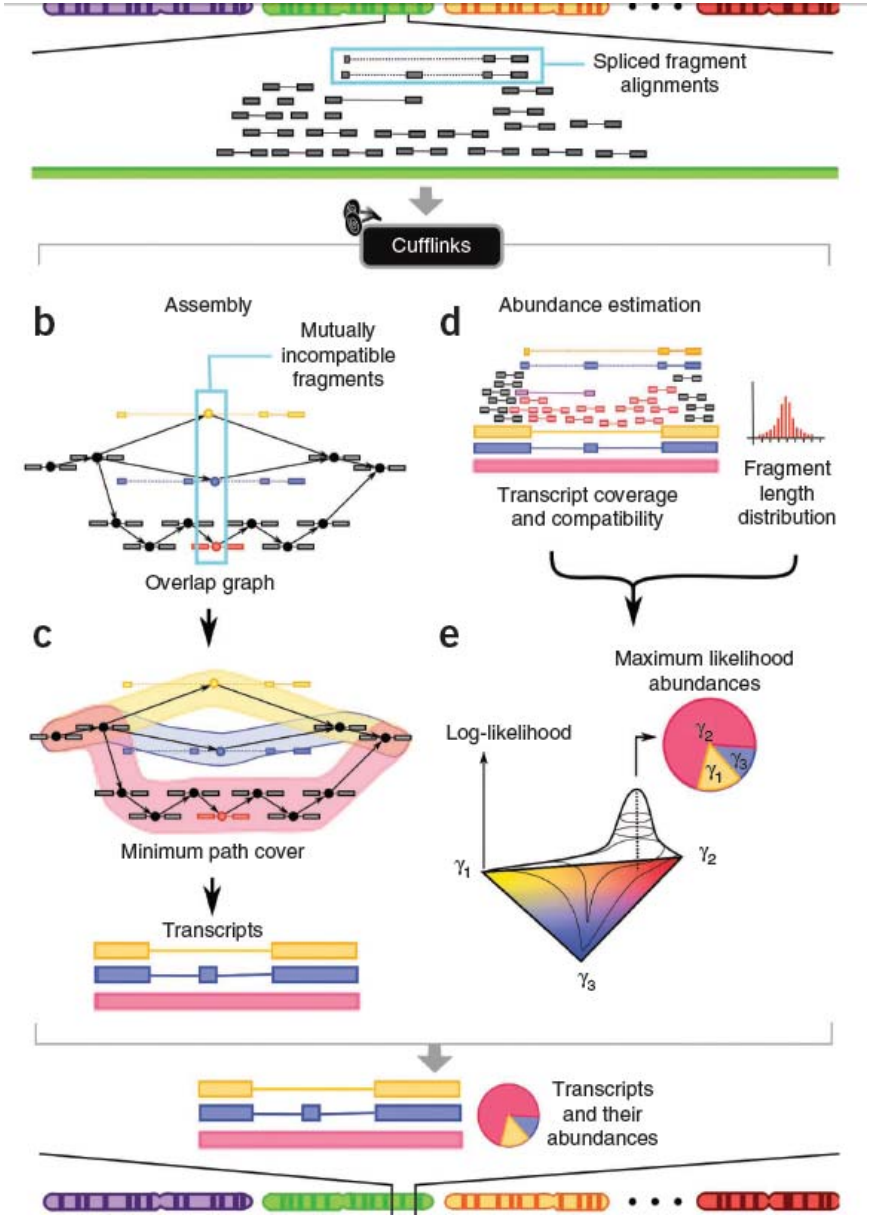
Cufflinks Software

Trapnell, *Nature Biotech*, 2010

Identify all compatible pairs of reads, connect them with an edge

Find a minimal set of paths that cover all the fragments in the overlap graph.

Regard the the paths as isoforms

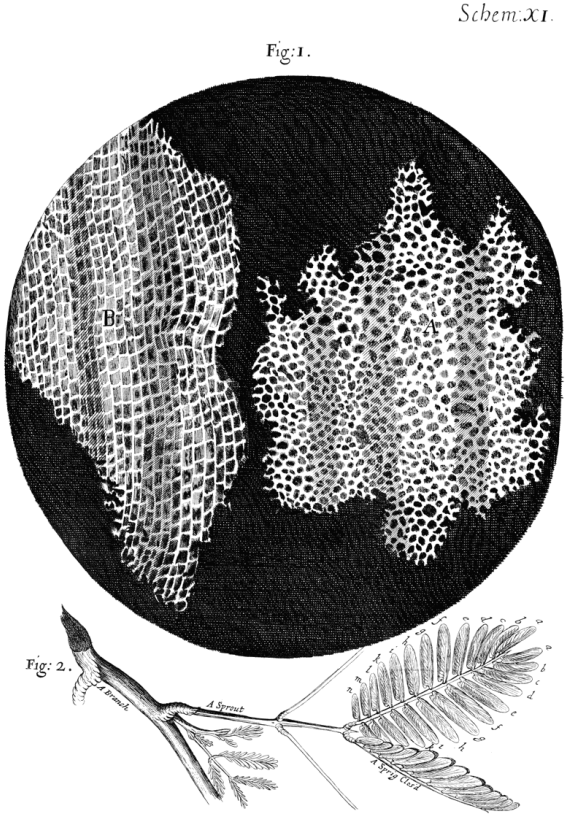


Major topics for RNA-seq analysis

- Mapping
- Reads count and Normalization
- Differential expression
- Alternative splicing
- Single cell RNA-seq

Cell -- Basic unit of life

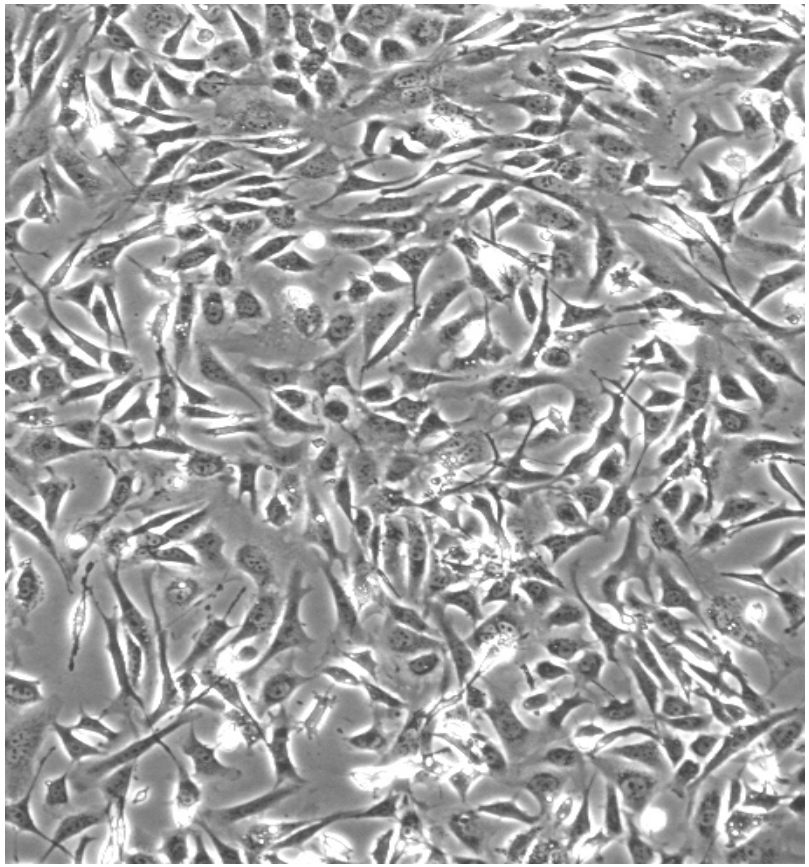
Discovery of cell by Hooke



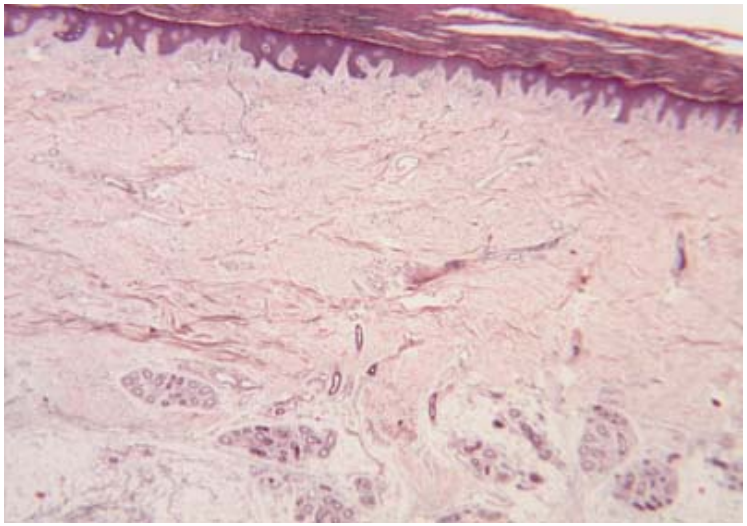
Micrographia, 1665

- All living organisms are composed of one or more cells.
- The cell is the basic unit of structure and organization in organisms.

Population-averaged assays are dominant

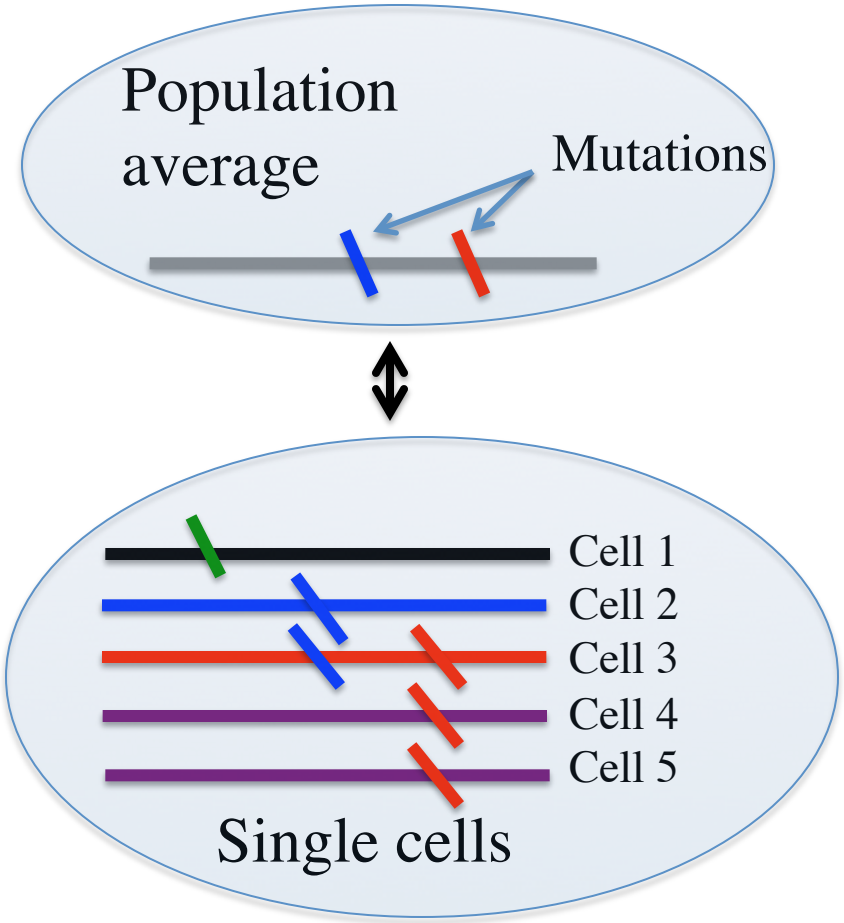


Mouse embryonic fibroblasts (MEF) cells

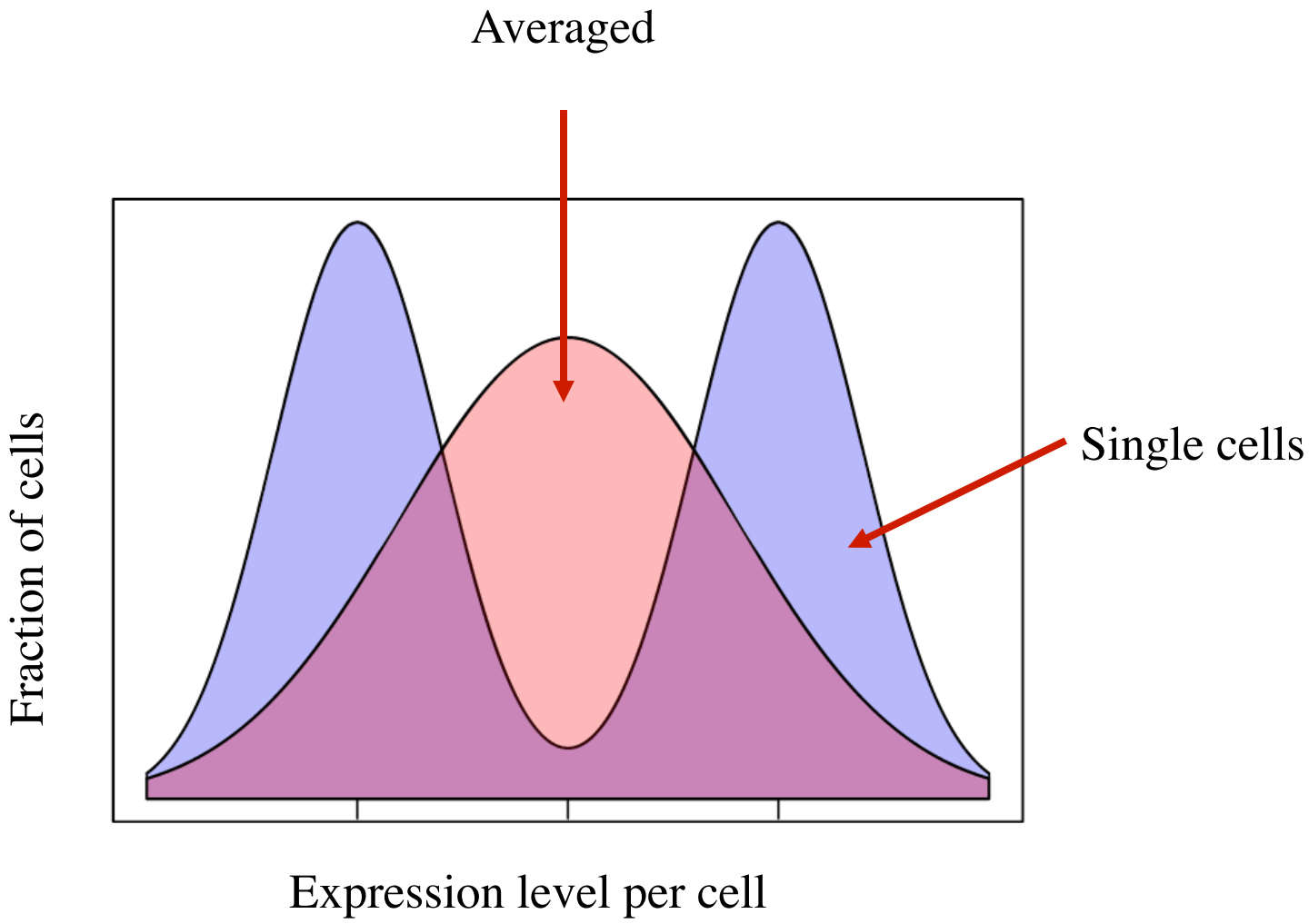


Skin(Tissue)

Population average obliterate uniqueness



Population average may be misleading



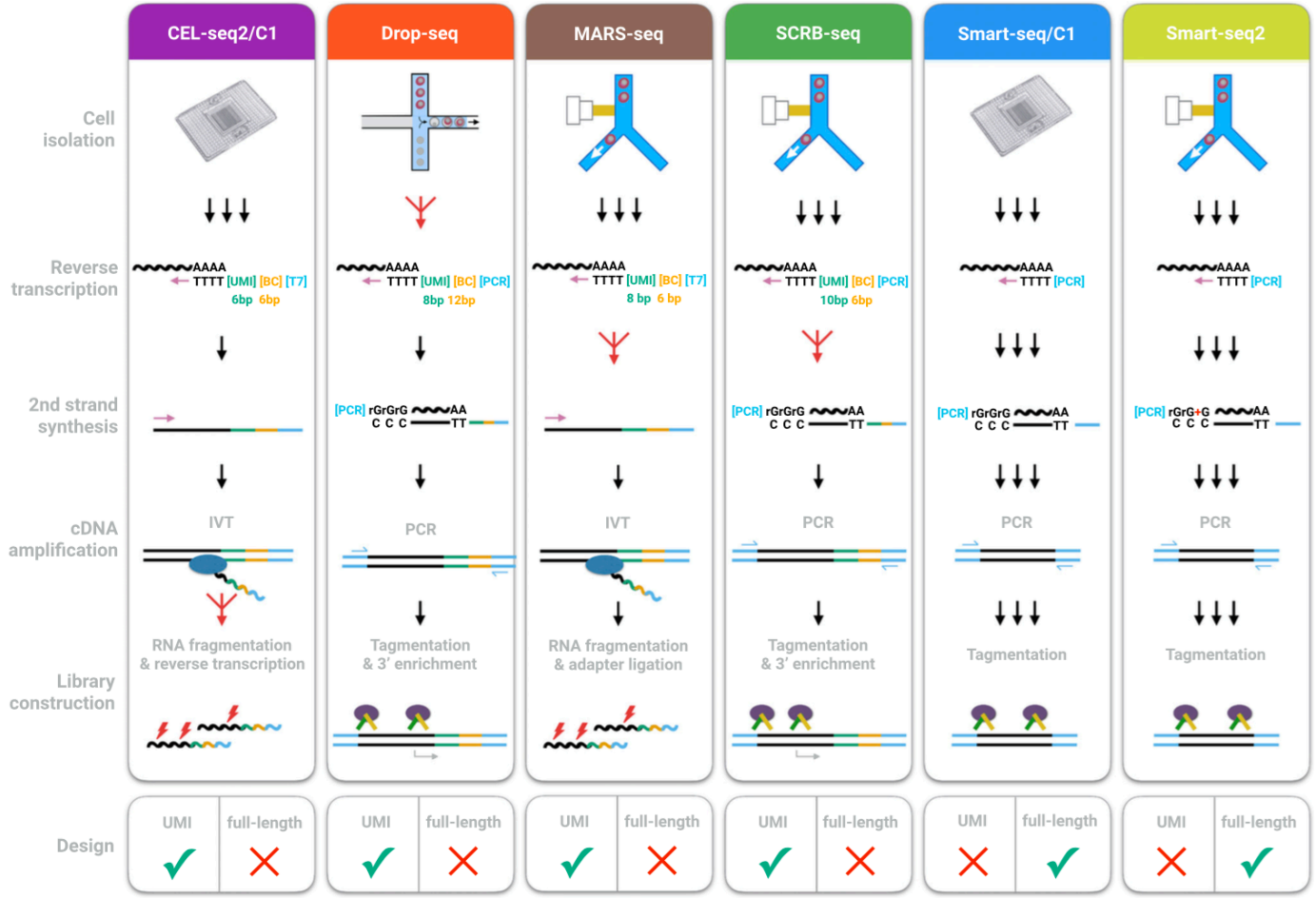
Important steps in single cell sequencing

- Cell isolation
- Library preparation (Amplification)
- Data analysis

Cell isolation

- Classic
  - Micromanipulation
  - Laser-capture microdissection
  - Fluorescence activated cell sorting (FACS)
  - Magnetic activated cell sorting (MACS)
- High throughput (recent)
  - Microfluidics (C1)
  - Droplet (10Xgenomics, Dolomite)

Library preparation(RNA-seq)



Ziegenhain et al., 2017, Molecular Cell

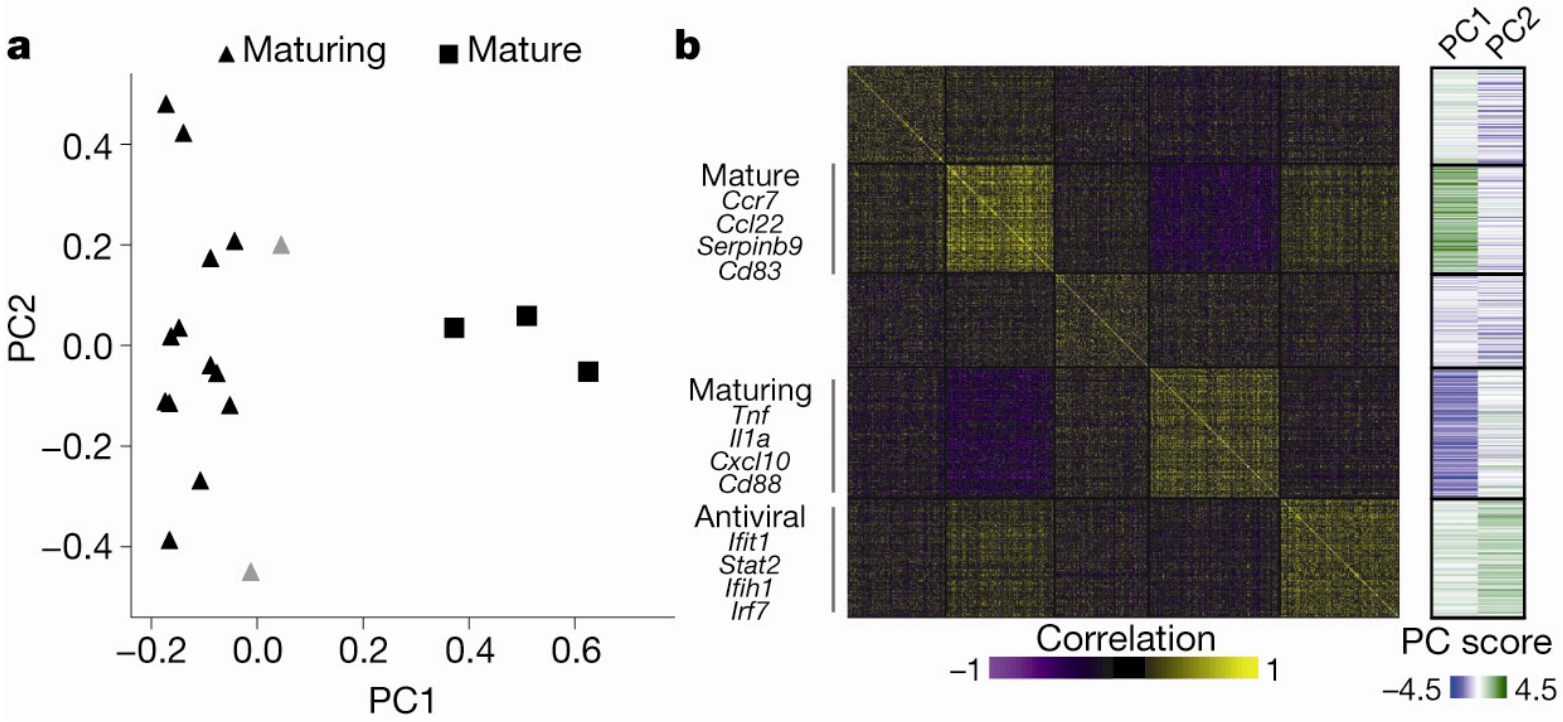
Features of single cell data

- Challenges
  - High noisy and high missing data
  - High amplification bias
  - Low genome coverage
- New feature
  - Low cost per library/cell
  - Big cell population, makes single cell sequencing data become very similar to classic population genetics data

Application of single cell sequencing

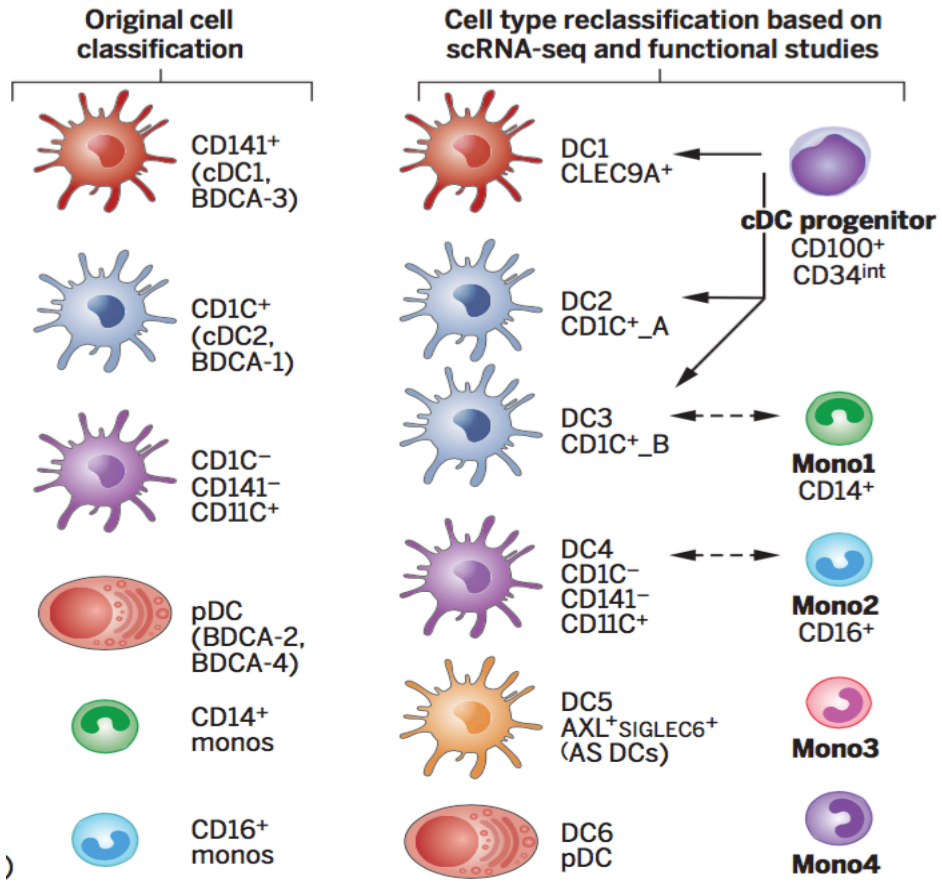
- Cell population genetics, especially cell lineage relationship
- Development biology and cell fate decisions
- Identifying new cell types
- Precision gene regulation at single cell level
- Cancer diagnosis and treatment
- Metagenomics
- Other clinical studies, e.g. immunity, neuroscience

Analysis of co-variation in single-cell mRNA expression levels reveals distinct maturity states and an antiviral cell circuit.



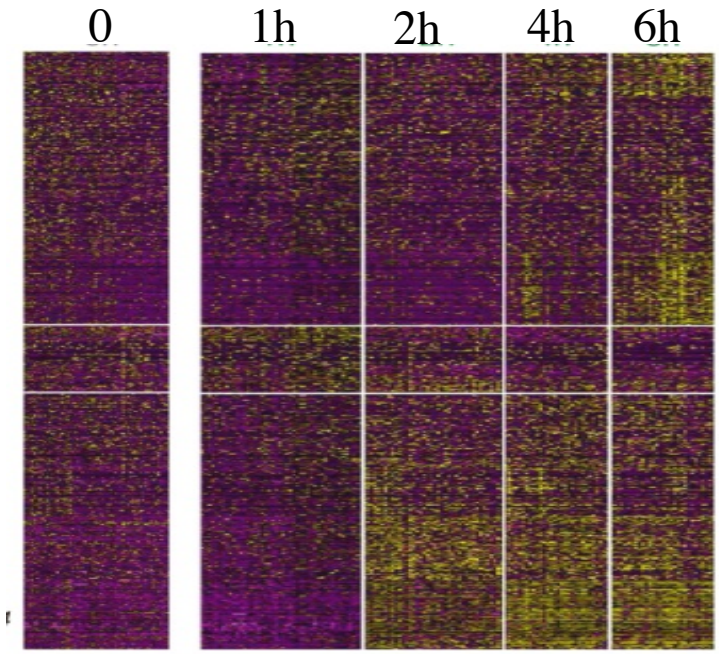
AK Shalek *et al. Nature* 000, 1-5 (2012) doi:10.1038/nature12172

Discovery of new cell types



Villani *et al*. 2017 Science

Heterogeneity is common

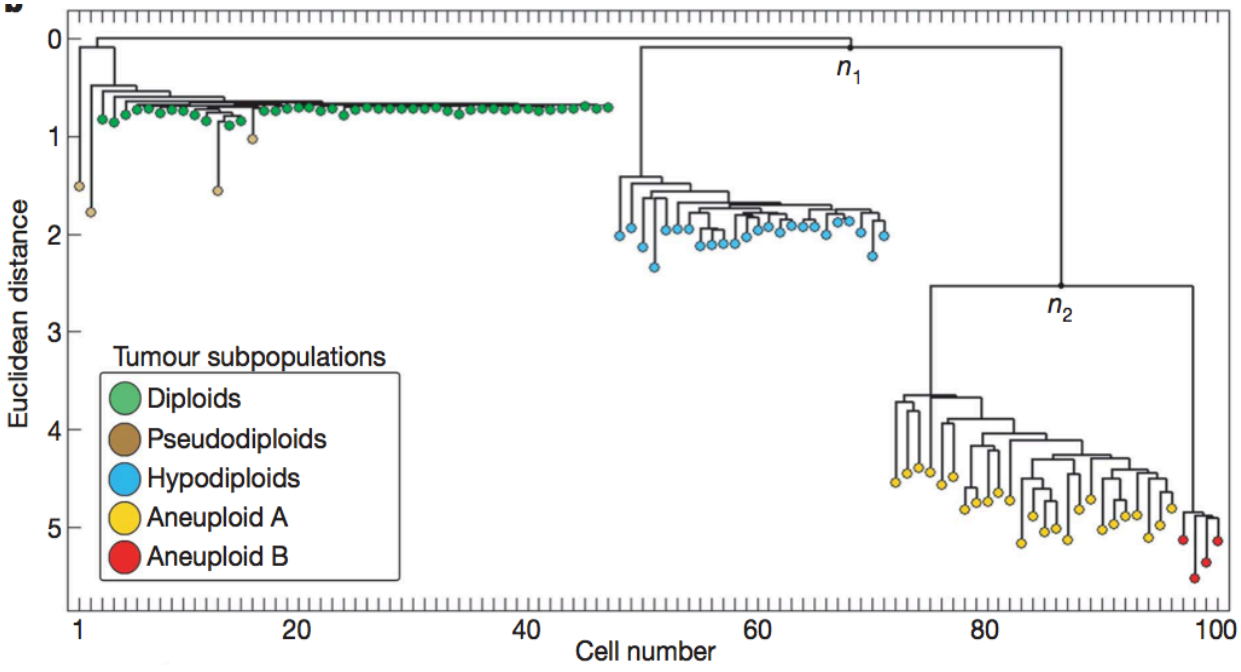


Even a seemly homogeneous cell population displays heterogeneity in gene expression and response to environmental stimulation.

(Shalek *et al. Nature* 2014)

Cancer cell subpopulations

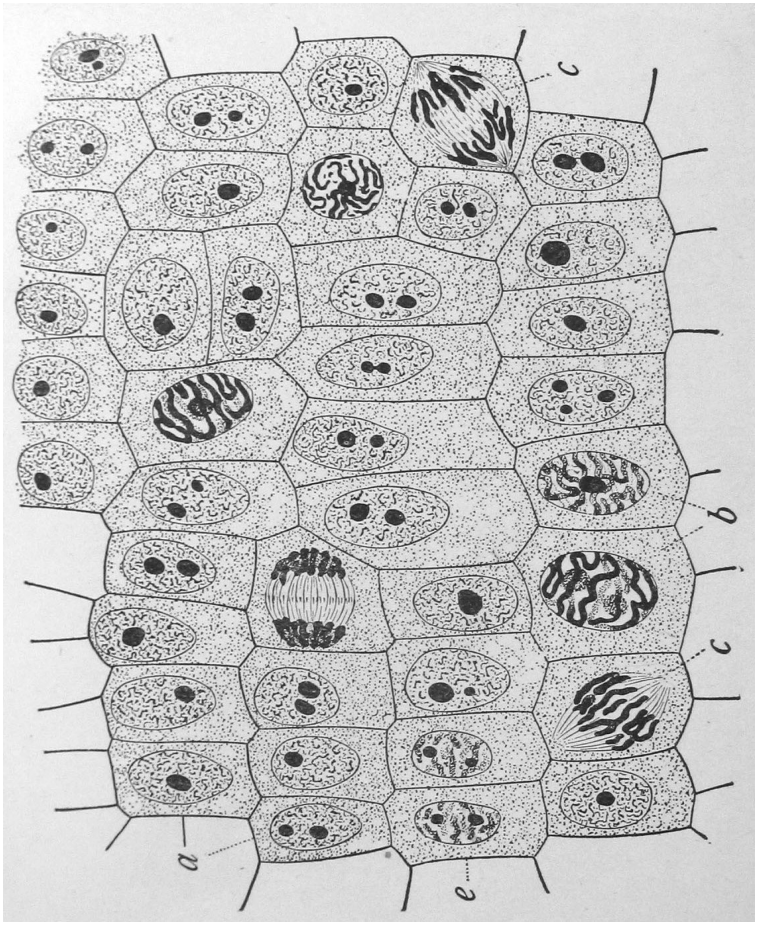
Single cell genome-seq revealed cancer cell subpopulations



Navin *et al*. Nature 2011

Each cell is unique

Cells in growing root-tip of onion



Wilson, 1900

Every cell is unique—it occupies an exclusive position in space, carries distinct errors in its copied genome and is subject to programmed and induced changes in gene expression.

Comparison of Methods for Studying the Transcriptome

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, Z. et al. *RNA-Seq: a revolutionary tool for transcriptomics* Nature Reviews Genetics (2009)

Thank you for your attention