Preliminaries

- Introduction
- Survey
- Etherpad
- Technologies
- Applications
- Experimental Design
- Exploratory data analysis: dimensionality reduction + clustering
- limma: what we learned from microarray data

Introduction

Mark D Robinson

Twitter @markrobinsonca
GitHub markrobinsonuzh
Google Scholar

My research interests are diverse, but more-or-less encompass the general application of statistical methods and data science to experimental data with biological applications. Often, this is within the context of genomics data types, but we are interested in methodological challenges and robust solutions in data, generally. We also try to be modern scientists, with a focus on reproducibility (repos for code) and open science (preprints).

PhD in Medical Biology (2008), University of Melbourne (Walter and Eliza Hall Institute)
MSc in Statistics (2001), University of British Columbia
BSc in Applied Mathematics and Statistics (1999), University of Guelph

https://robinsonlabuzh.github.io/people.html

Introduction

Simone Tiberi

Website
Twitter @tiberi_simone
GitHub SimoneTiberi
Google Scholar

I am a Postdoc working on the development of cutting-edge statistical methods in bioinformatics, mostly for bulk and single-cell RNA-seq data.In particular, I am currently focusing on two methodological projects. The first one consists in developing a Bayesian hierarchical model to identify differentially spliced genes, via differential transcript usage (DTU), from bulk RNA-seq data. The second project (in its early stages) aims at creating a methodology, based on permutation tests, to perform differential state analyses from single-cell RNA-seq data. Previously, I was a PhD student at the Department of Statistics at the University of Warwick, where I worked on Bayesian hierarchical models to investigate stochastic systems in single cells. In general terms, my interests are broad and lie in the development of statistical methods for applications in biology.

PhD in Statistics (2017), The University of Warwick
MSc in Statistics (2012), The University of Padua
BSc in Statistics (2010), Sapienza University of Rome

https://robinsonlabuzh.github.io/people.html

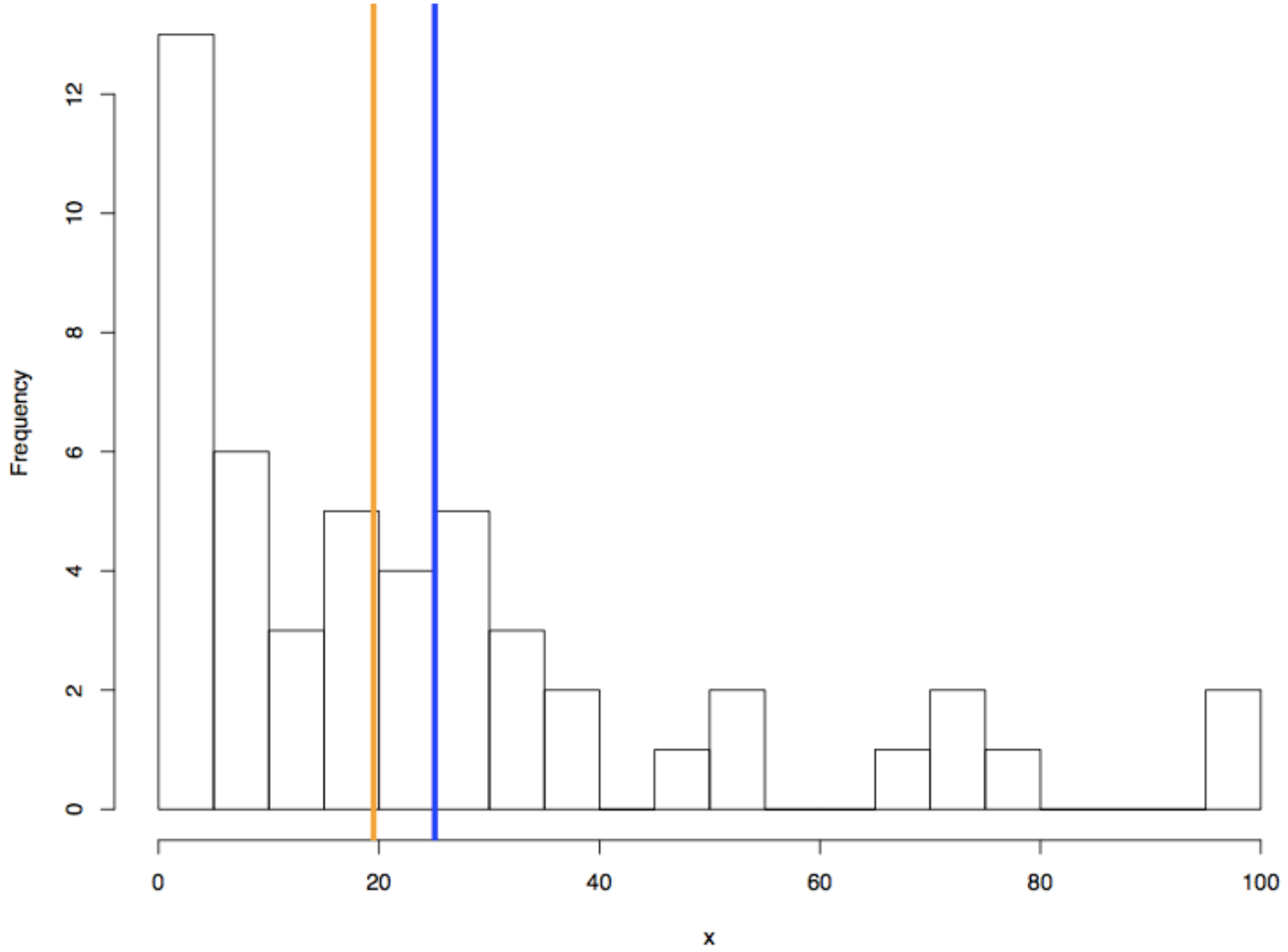
Survey: Statistical Insight

movo.ch

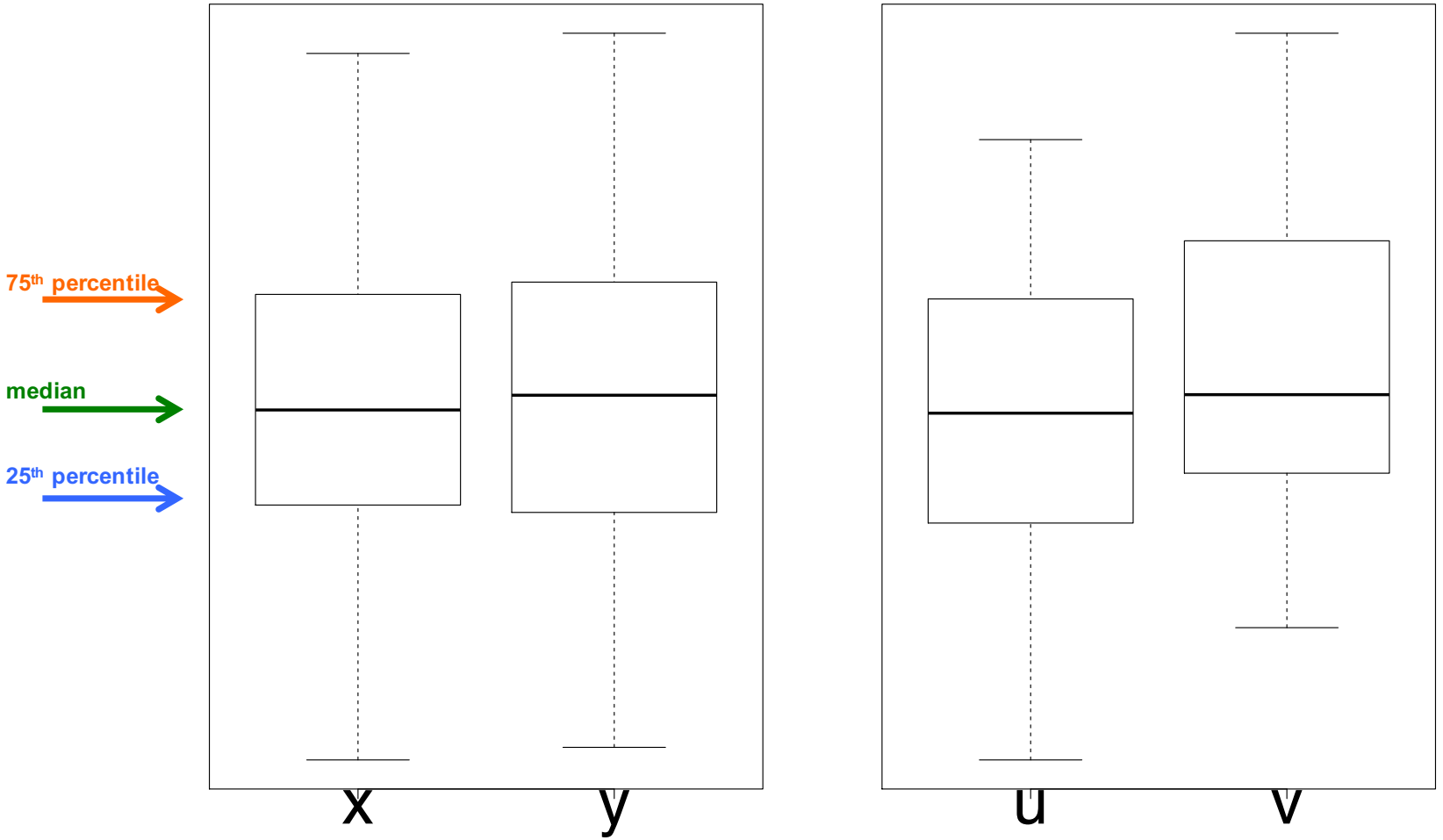
Token:

KI KY BY LU

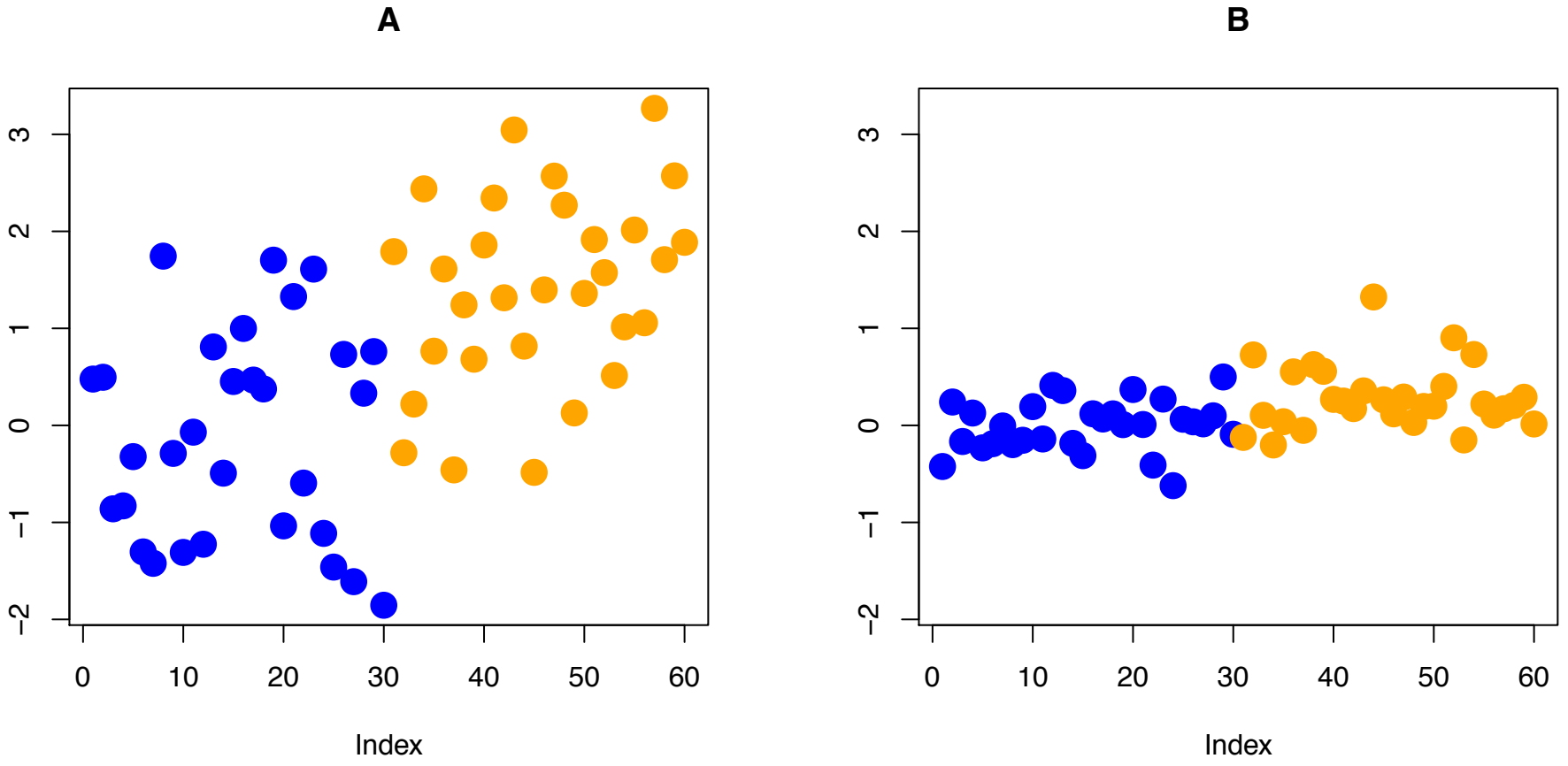
Question 1: From the histogram, decide whether blue or orange represents the mean/median



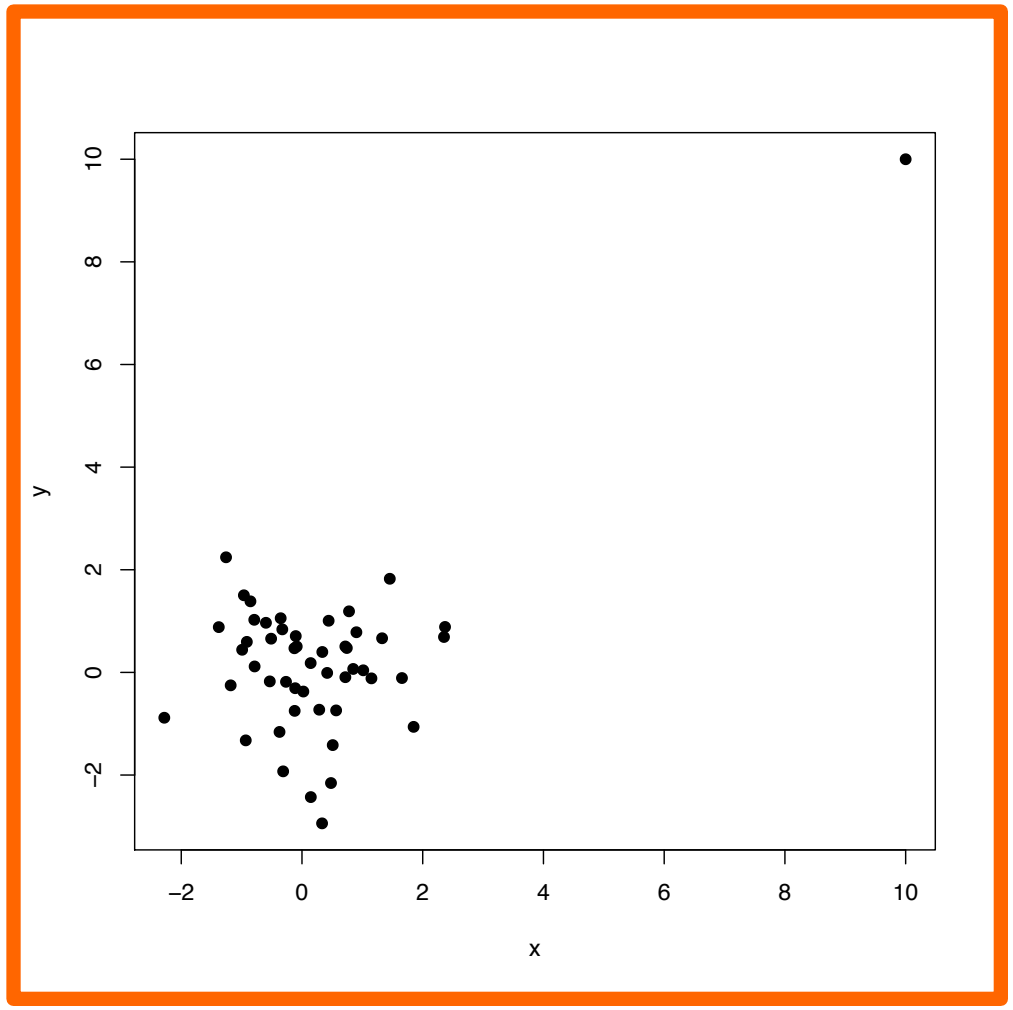
Question 2: Given these boxplots, which of two underlying distributions are more similar?



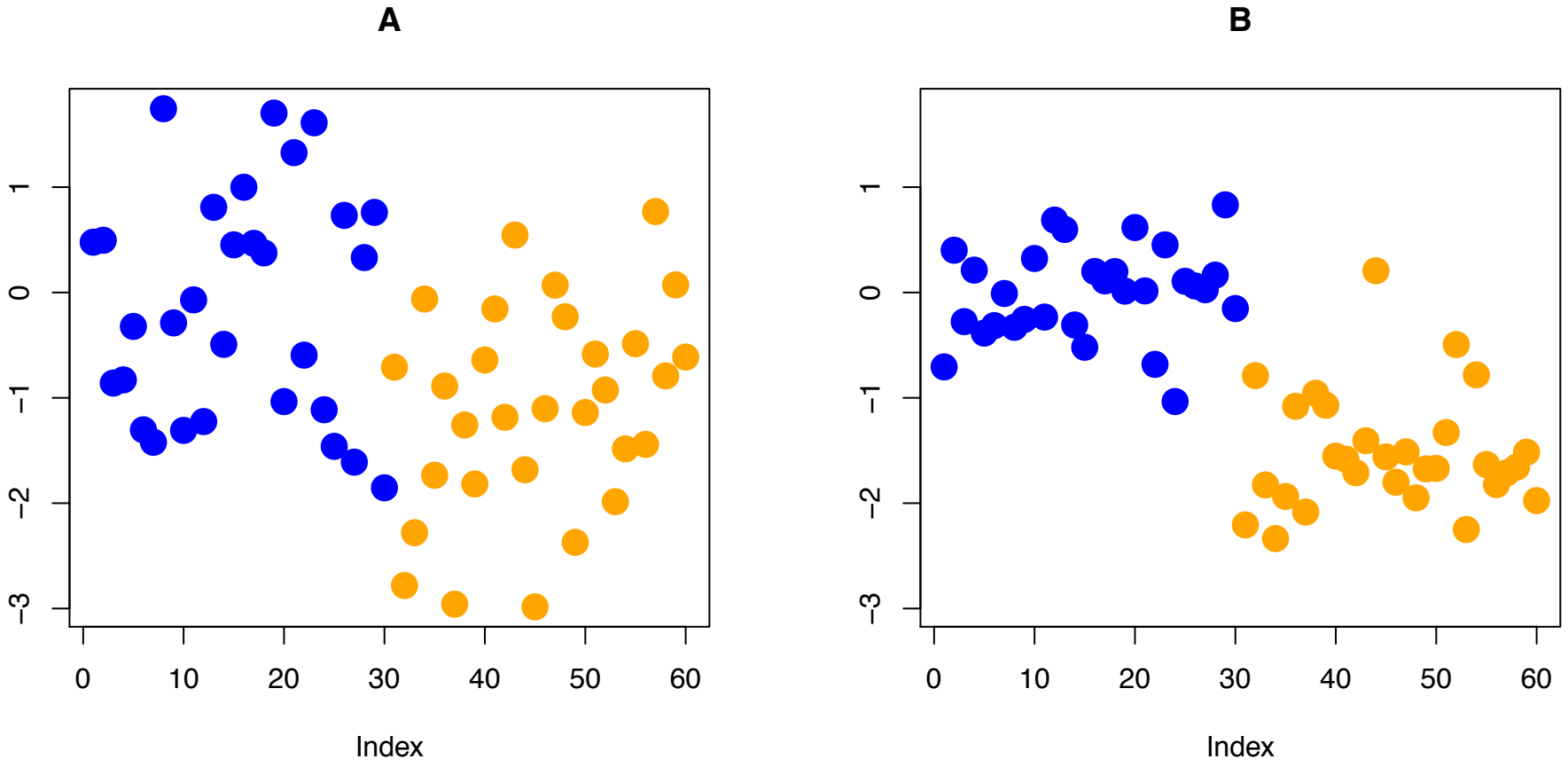
Question 3: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



Question 4: In your view, what best describes the associations shown in the plot of 'x' and 'y' ?



Question 5: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



Question 6: Of these equations, which one resembles the standard two sample t-test ?

1

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

2

$$\sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

3

$$\frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

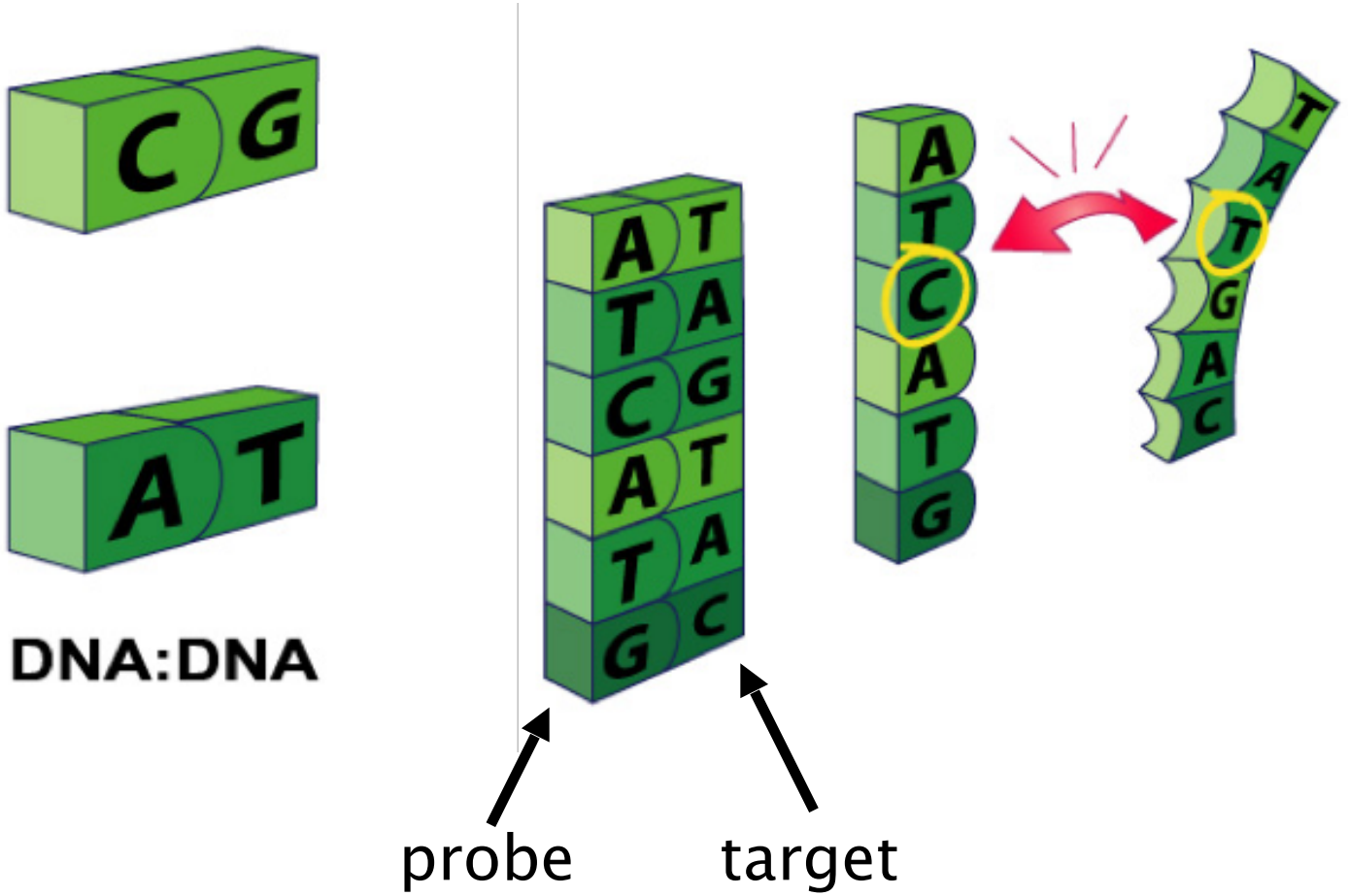
Etherpad

- http://bit.ly/2UoeA5c
- https://public.etherpad-mozilla.org/p/Pretoria_RNAseq_Feb2019

- Task: what would you like to get out of this RNA-seq course? —> write 1 or 2 sentences in the Etherpad

Technologies

Microarray fundamentals: Nature gives a complementary pairing



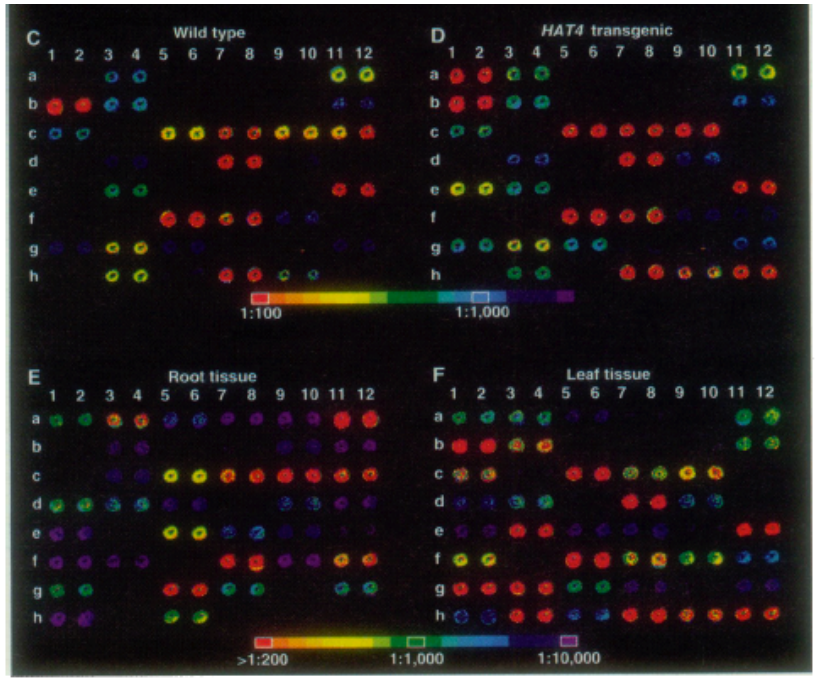
Microarrays: Where it all began …

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization v… of 0… microliters could be used that enabled detection of rare transcripts in pro… derived from 2 micrograms of total cellular messenger RNA. Differential… measurements of 45 *Arabidopsis* genes were made by means of simultaneou… fluorescence hybridization.

Science, 1995



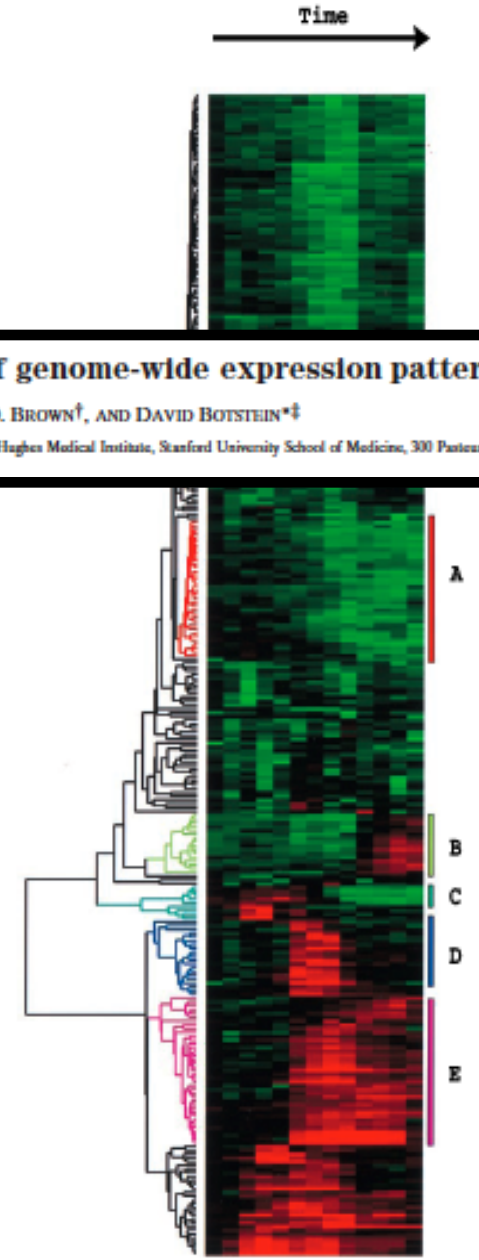
45 Arabidopsis genes
2 probes each
printed on glass

Cluster analysis and display of genome-wide expression patterns

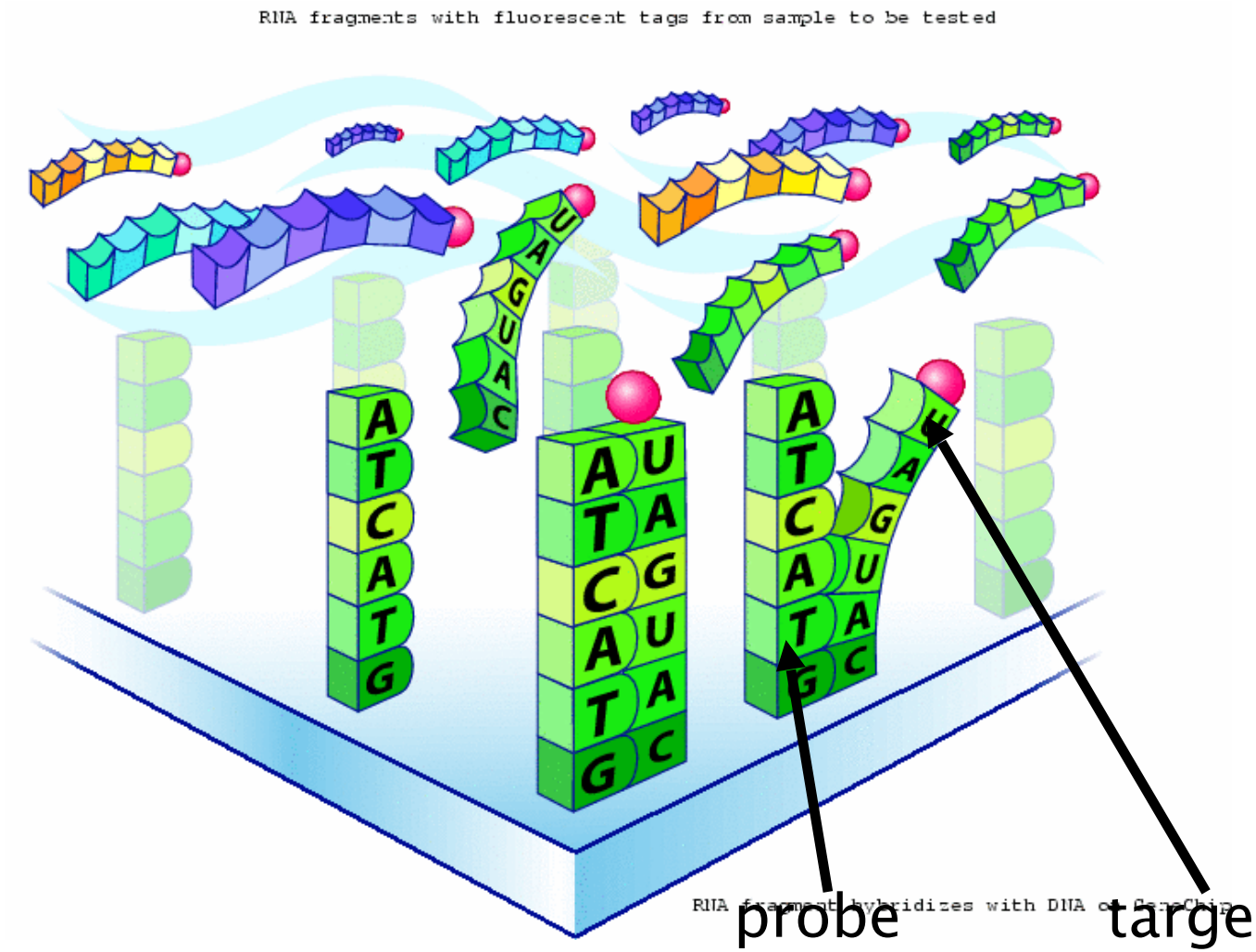
MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

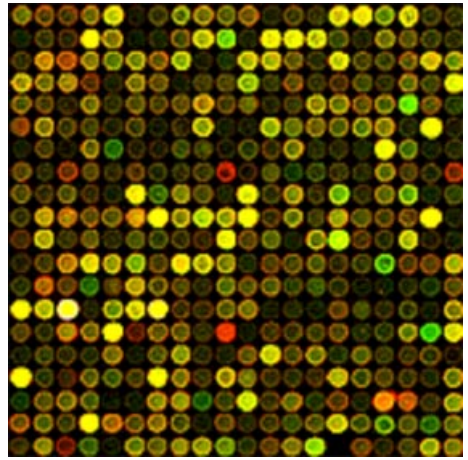
PNAS, 1998



DNA microarray: **parallel** northern blots



Abundance (of complementary DNA species) measured by flouresence intensity

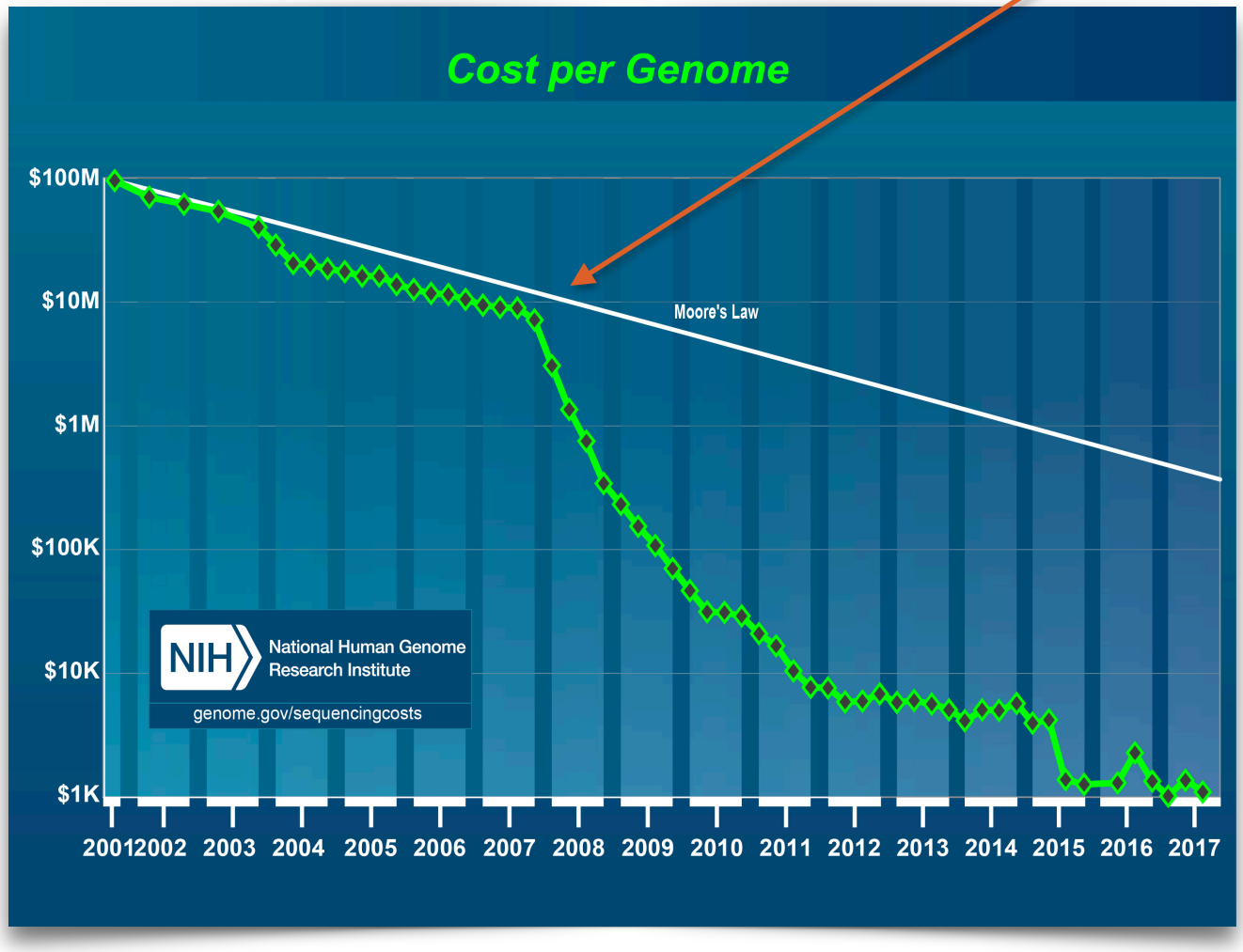


Gene Expression Profiling: questions of interest

- What genes have changed in expression? (e.g. between disease/normal, affected by treatment)
  Gene discovery, differential expression
- Is a specified group of genes all up-regulated in a particular condition?
  Gene **set** differential expression
- Can the expression profile predict outcome?
  Class prediction, classification
- Are there tumour sub-types not previously identified? Do my genes group into previously undiscovered pathways?
  Class discovery, clustering

High-throughput sequencing

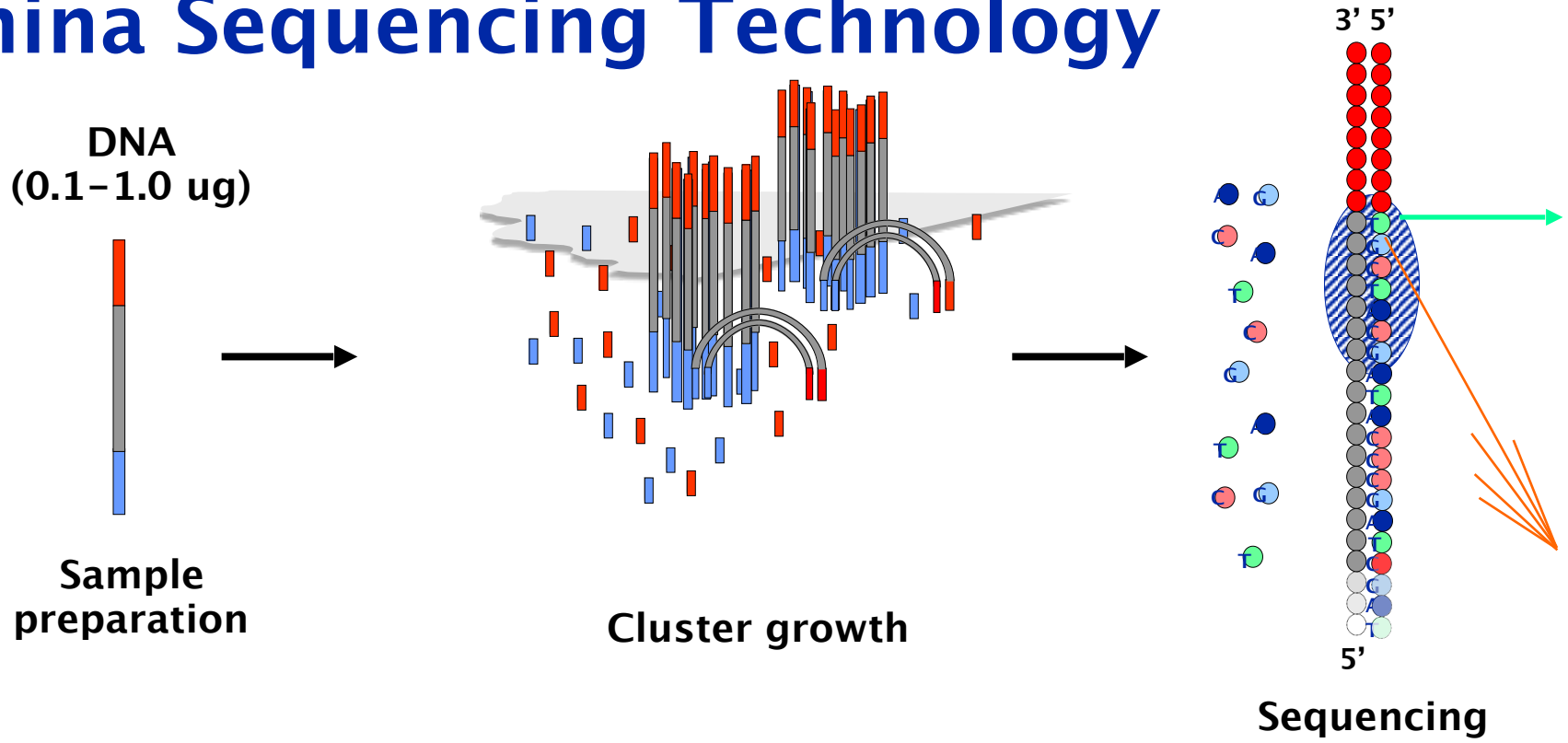
(Solexa) Illumina



https://www.statnews.com/2017/01/09/illumina-ushering-in-the-100-genome/

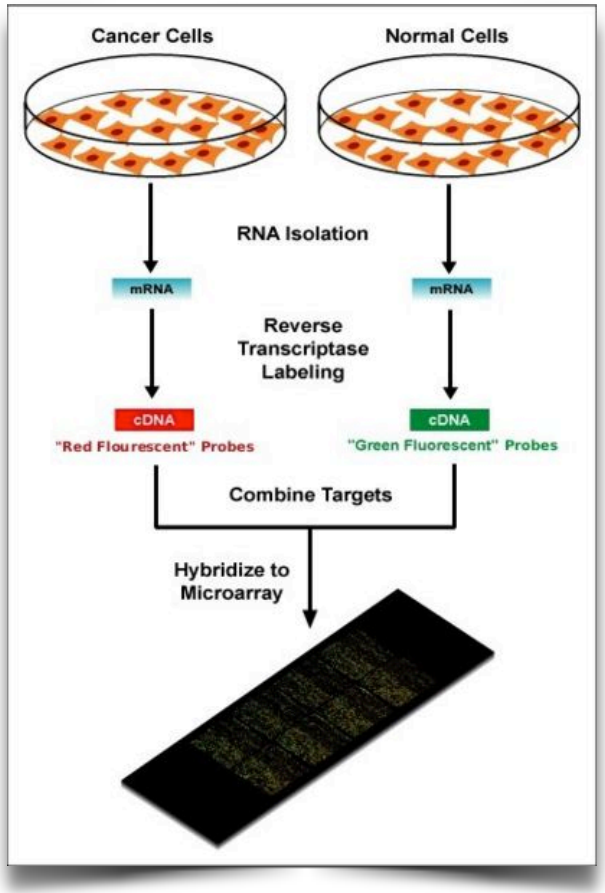
Many assays based on sequencing DNA:
https://liorpachter.wordpress.com/seq/

Illumina Sequencing Technology



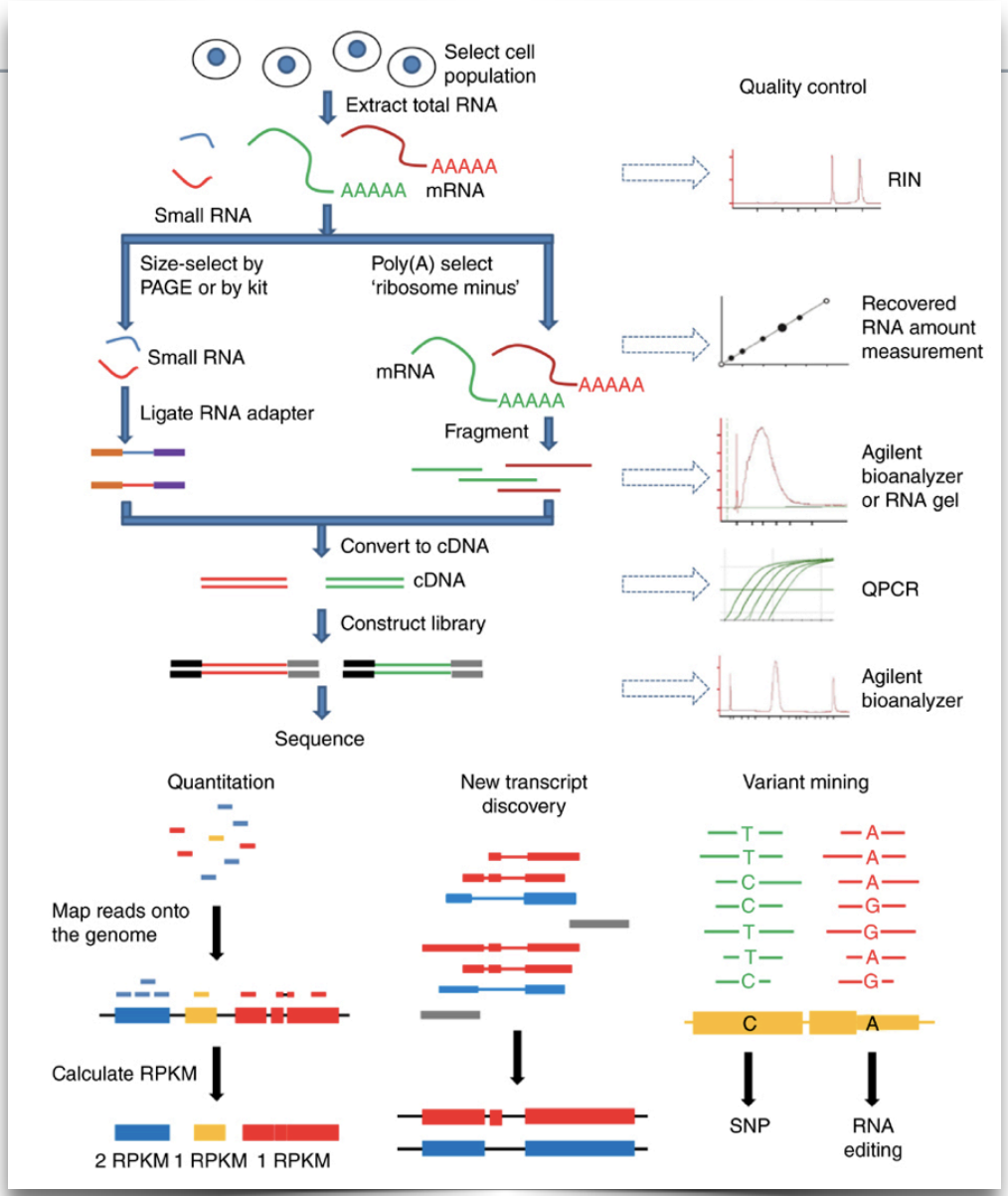
This slide courtesy of Gary Schroth, Illumina

Abundance by Fluorescence Intensity (DNA microarray)



http://en.wikipedia.org/wiki/DNA_microarray

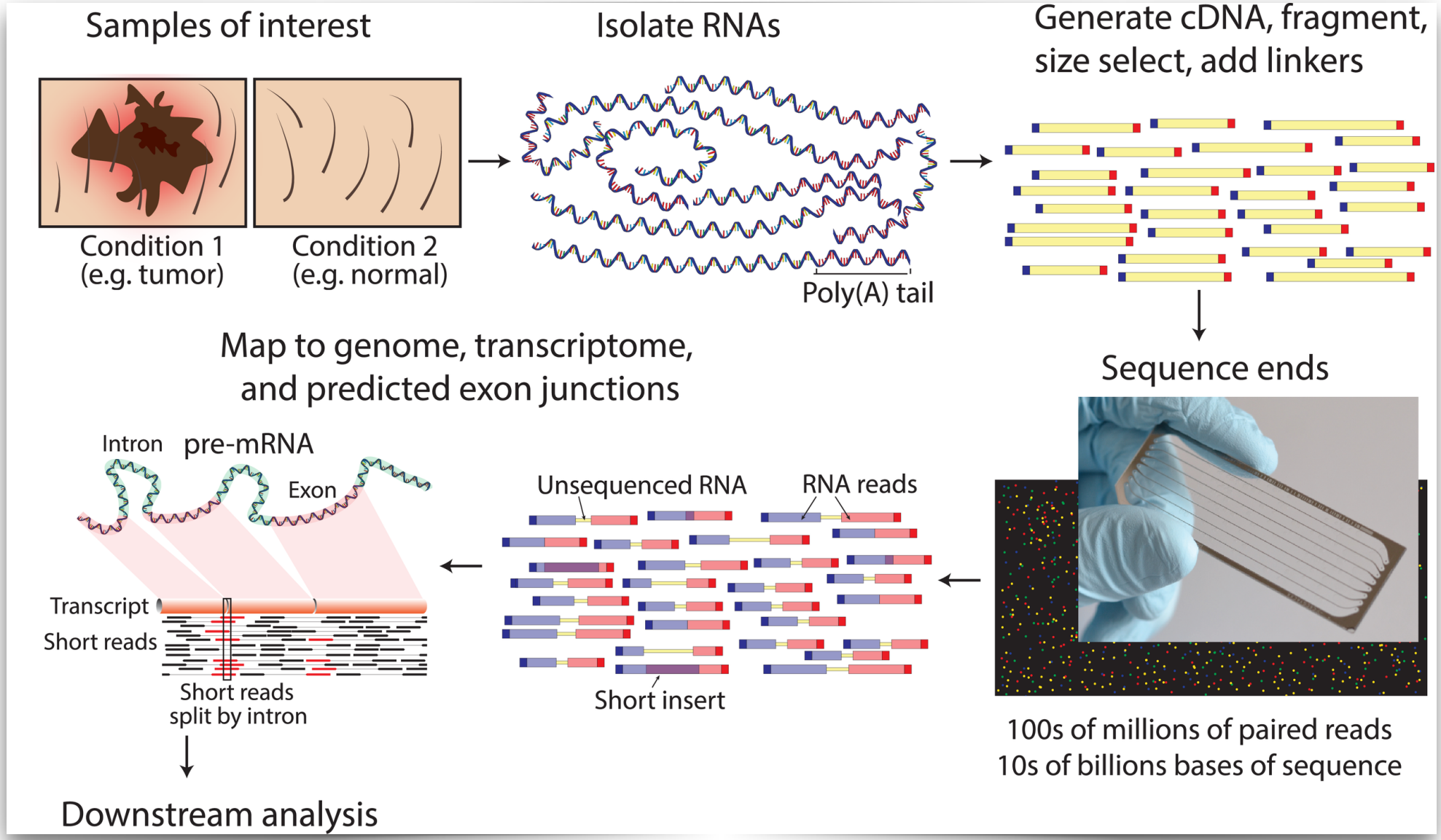
Abundance by Counting (RNA-seq)



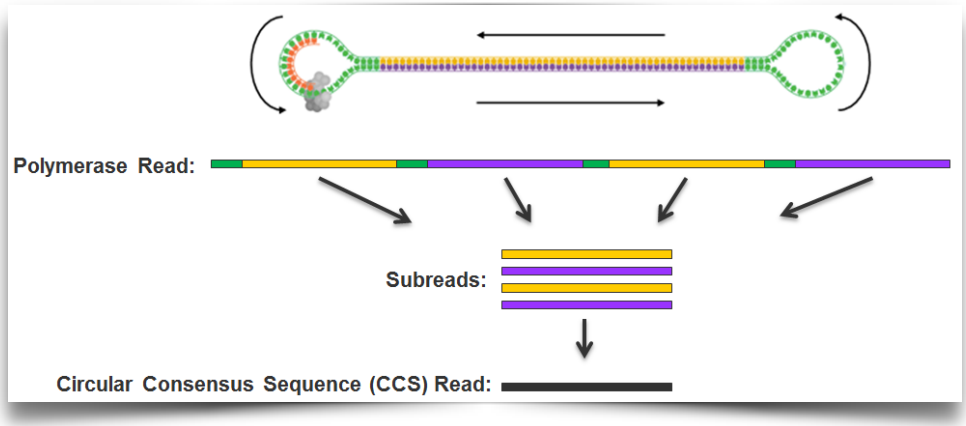
Zeng & Mortazavi, Nature Immunology, 2012

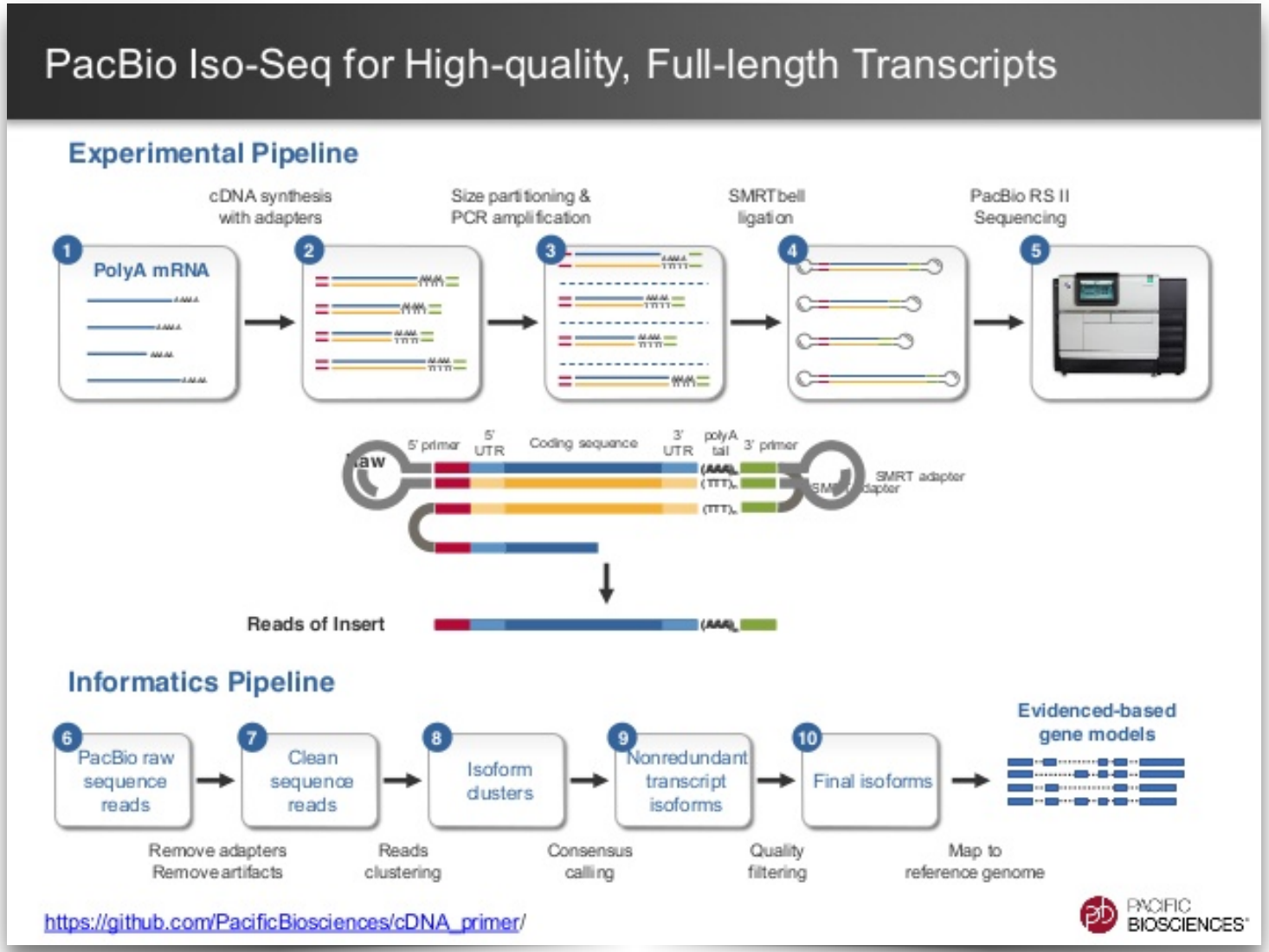
https://en.wikipedia.org/wiki/RNA-Seq

Illumina



PacBio

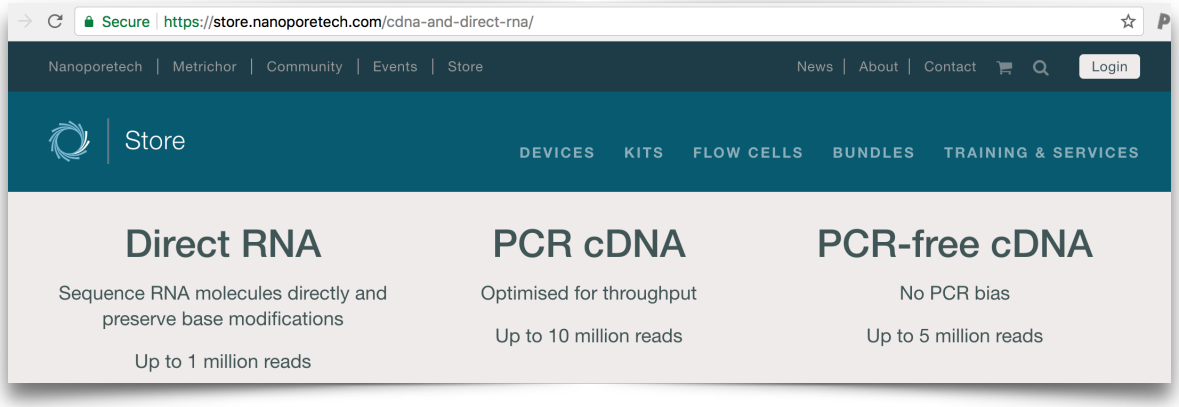




https://www.slideshare.net/ERCC-Workshop/20140711-3-tclarkercc20workshop

https://nanoporetech.com/applications/dna-nanopore-sequencing

ONT (Oxford Nanopore)



—> attachment of processive enzyme, leads RNA/DNA fragment to pore, combination of nucleotides going through pore creates a “characteristic disruption of the electrical current” —> order of signals can be used to determine the sequence of bases on that single strand.

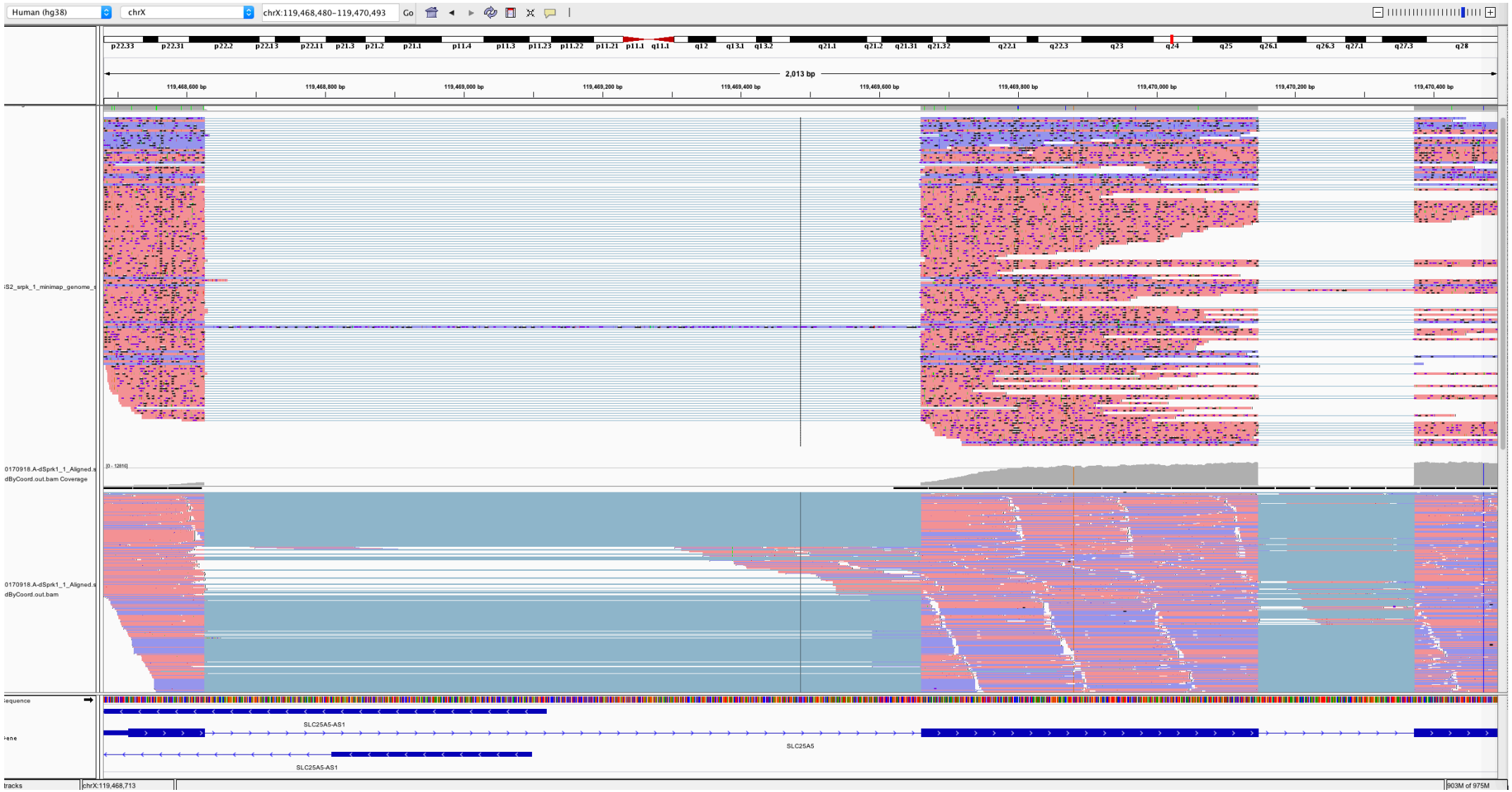
Illumina - PacBio - Nanopore

Technology	*Advantages*	*Disadvantages*
Illumina	cheap, quantitative	short reads, assembly/transcript
PacBio	full length reads	(error rate), non-quantitative, cost
ONT	full length reads, quantitative	error rate, cost

Quick look at reads in a browser

ONT

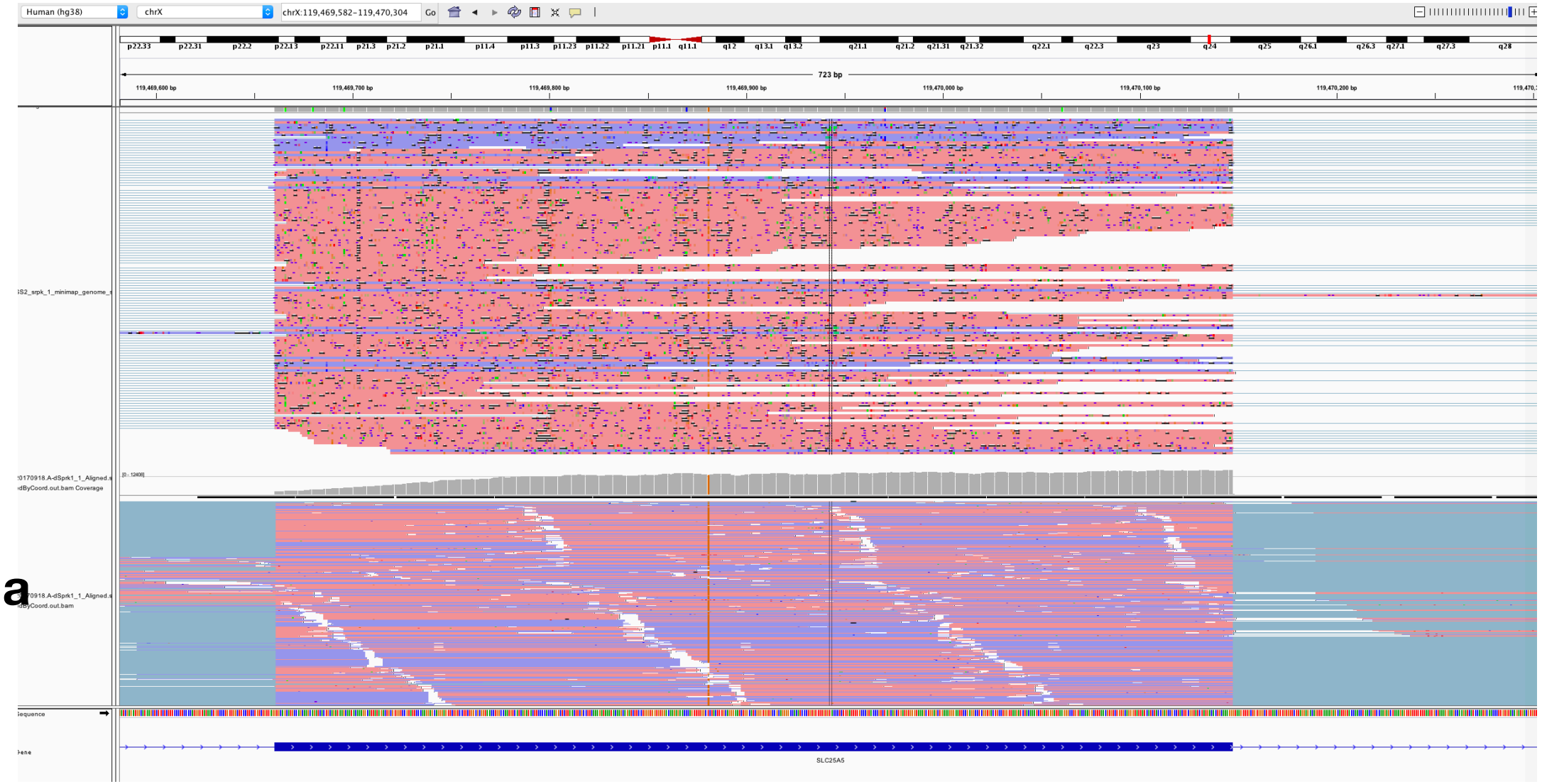
Illumina



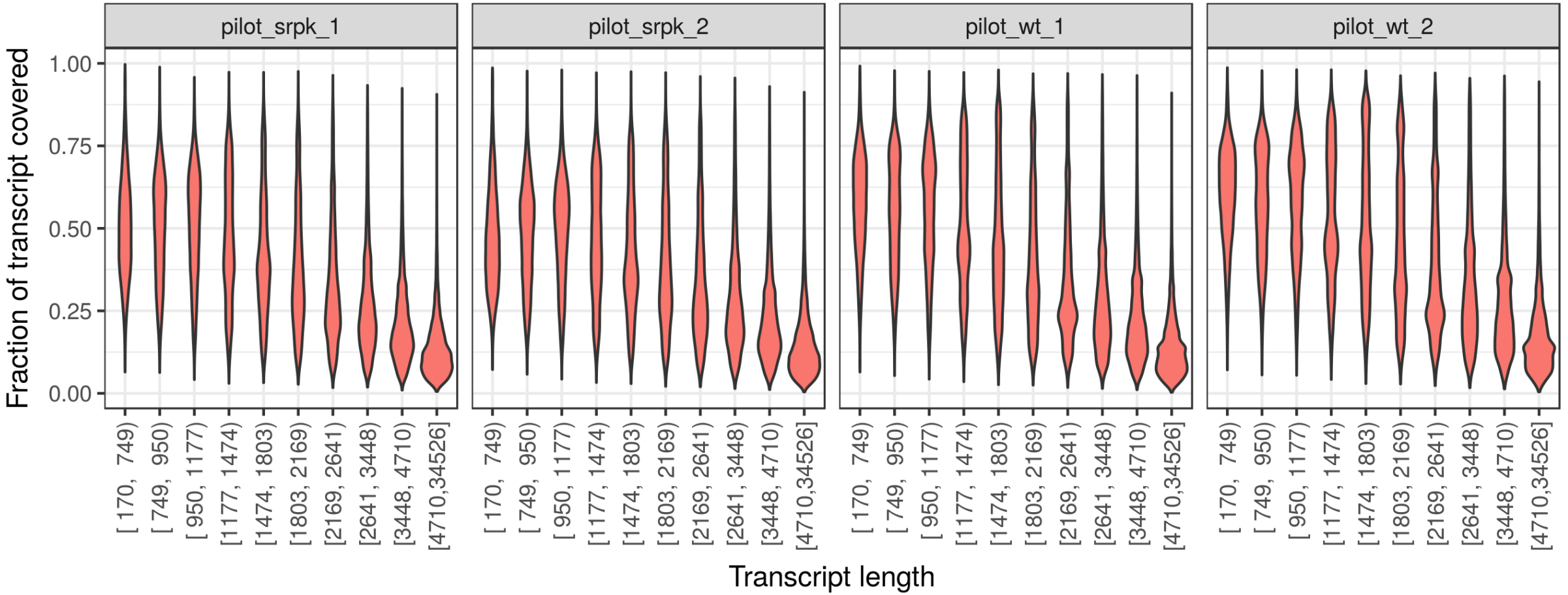
Quick look at reads in a browser

ONT

Illumina



Basic properties: do we get full length transcripts?



https://peerj.com/preprints/27283/

Many more details here (228 cited papers!)

RNA sequencing data: hitchhiker's guide to expression analysis

Literature review | Bioinformatics | Computational Biology | Genomics | Data Science

Koen Van Den Berge *[1], Katharina Hembach *[2], Charlotte Soneson *[2,3], Simone Tiberi *[2], Lieven Clement [1], Michael I Love [4], Rob Patro [5], Mark Robinson ✉[2]

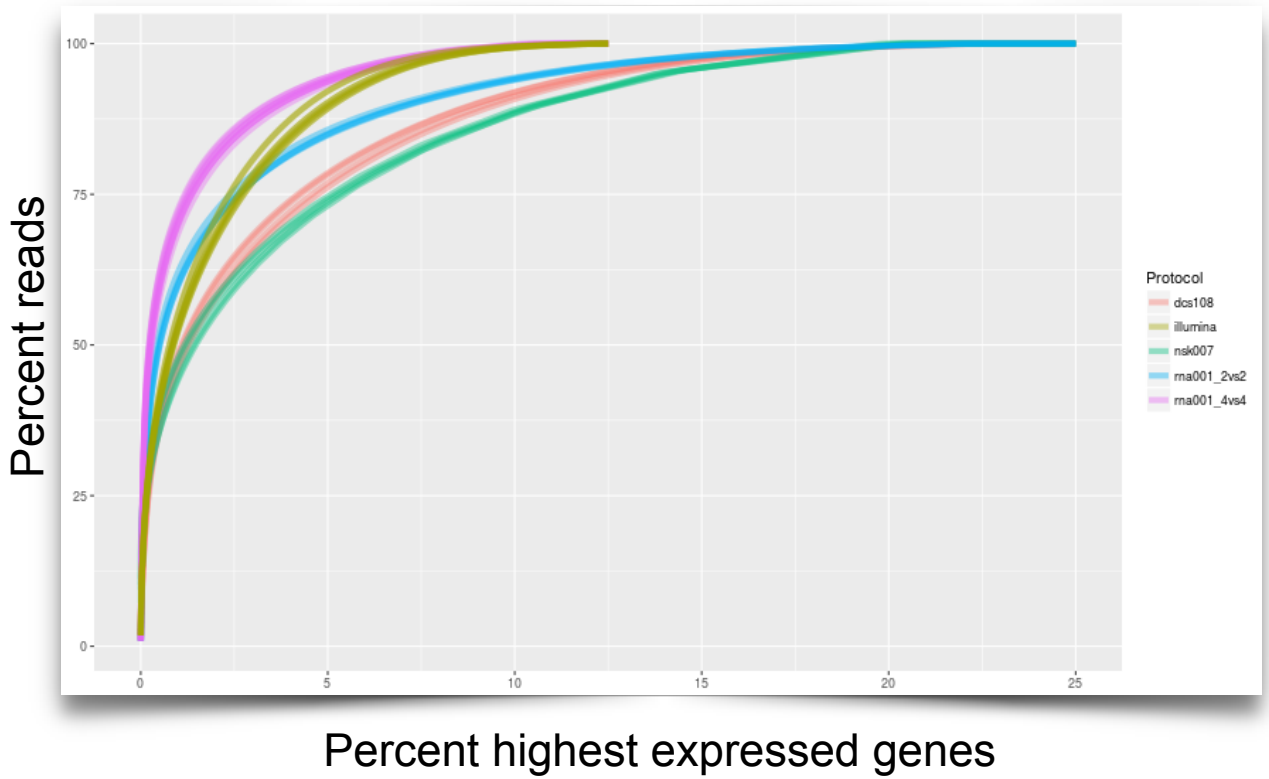
November 24, 2018

Applications

- popularity of RNAseq is driven by its large number of applications
- genome annotation: exon skipping, alternative 3’ acceptor or 5’ donor sites or intron retention .. microexons, cryptic exons, “skiptic” exons, circular RNAs, enhancer RNAs, fusion genes, epi-transcriptomics (RNA base modifications)
- gene regulation: comparison of gene / transcript / exon expression between different tissues, cell types, genotypes, stimulation conditions, time points, disease states, growth conditions and so on, understanding of the molecular pathways that are used or altered or the regulatory components that are utilized.
- molecular subclassification e.g., cancer; future clinical applications
- spatial transcriptomics, where cellular positional information, host-pathogen interactions via “dual RNAseq”, where the transcriptomes of both host and pathogen are simultaneously assayed, the analysis of genetic variation among expressed genes, RNA editing events, characterization of long noncoding RNAs and metatranscriptomics.
- sometimes, single cell resolution is desired, e.g., when studying heterogeneous tissues that consist of more than one cell type. Deconvolution of bulk RNAseq possible, but cannot discover new cell types or perform cell-type-specific analyses

Experimental Design

- basics of experimental design apply, e.g., randomize experimental units to treatments
- avoid confounding factors (e.g., via blocking over batches): represent every experimental condition in each batch
- number of replicates versus depth of sequencing; first driver of sample size is budget.
- In most cases, the budget is better spent on replicates. For example, Schurch *et al*. show that a higher number of replicates is required to identify DE of genes with low fold change and ideally at least 6 replicates per condition should be used.
- sample size calculators
- targeted RNAseq (RNA CaptureSeq): specific regions are first captured by probes



Schurch *et al*.: https://rnajournal.cshlp.org/content/22/6/839

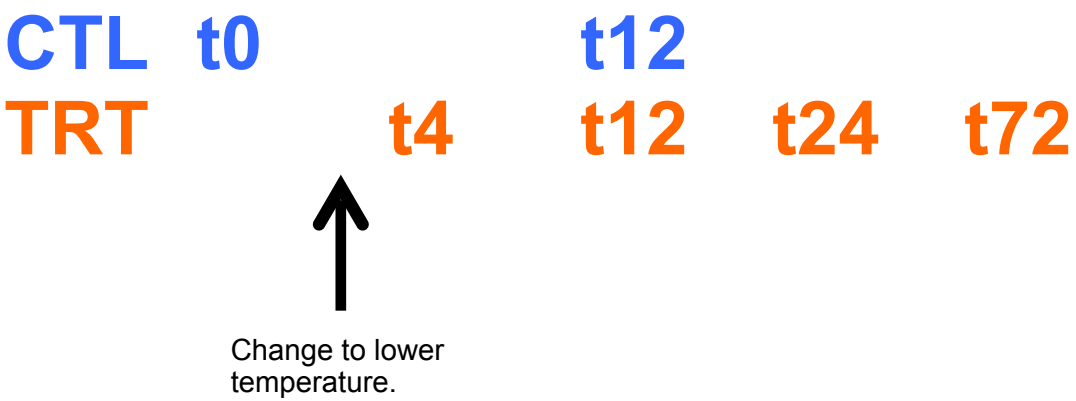
"To consult the statistician after an experiment is finished is often merely to ask [them] to conduct a post mortem examination. [They] can perhaps say what the experiment died of." R. A. Fisher

Motivating example: exploratory data analysis

(from Stefano, a former M.Sc. student in my Institute)

He is studying gene expression in fruitfly and is interested in transcriptional responses following "heat shock".

Basic schematic of experiment:

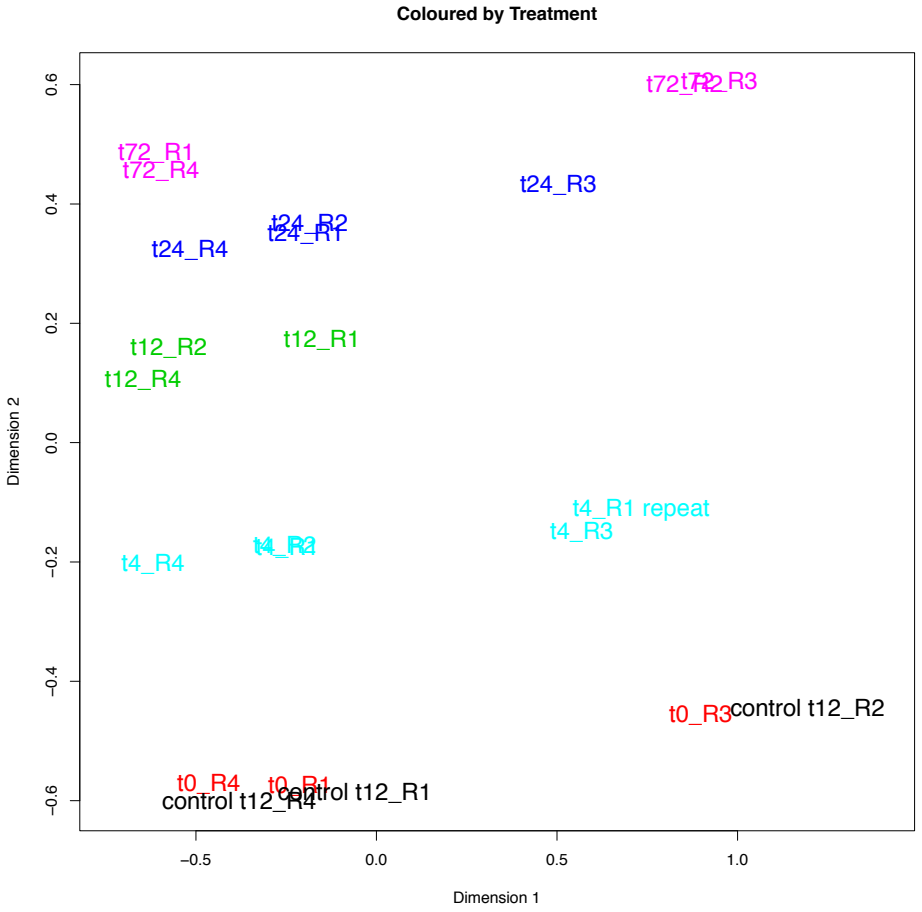


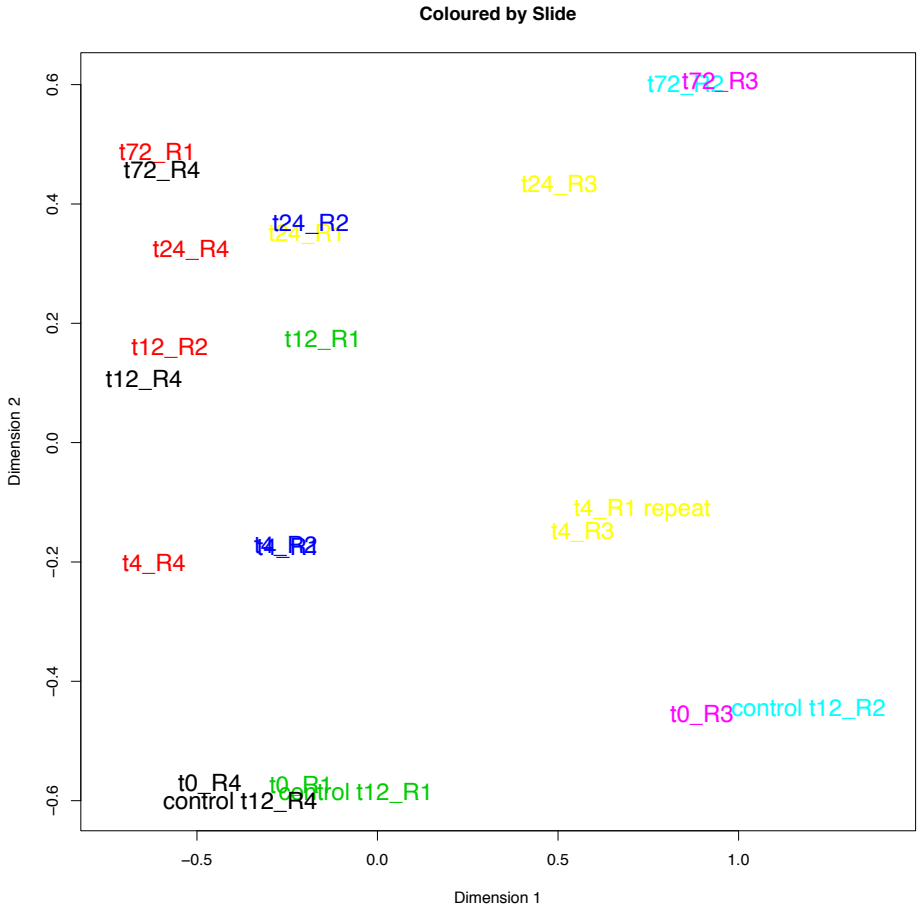
~4 replicates for each condition

http://en.wikipedia.org/wiki/Multidimensional_scaling

```
library(limma)
plotMDS(d)  # 'd' is a matrix
"Plot samples on a two-dimensional scatterplot so that
distances on the plot approximate the typical log2 fold
changes between the samples."
```

Take a close look at where the 24 samples are to each other relative to the X- and Y-axes





22 samples x ~20,000 genes

reduced to 22 samples x 2 dimensions

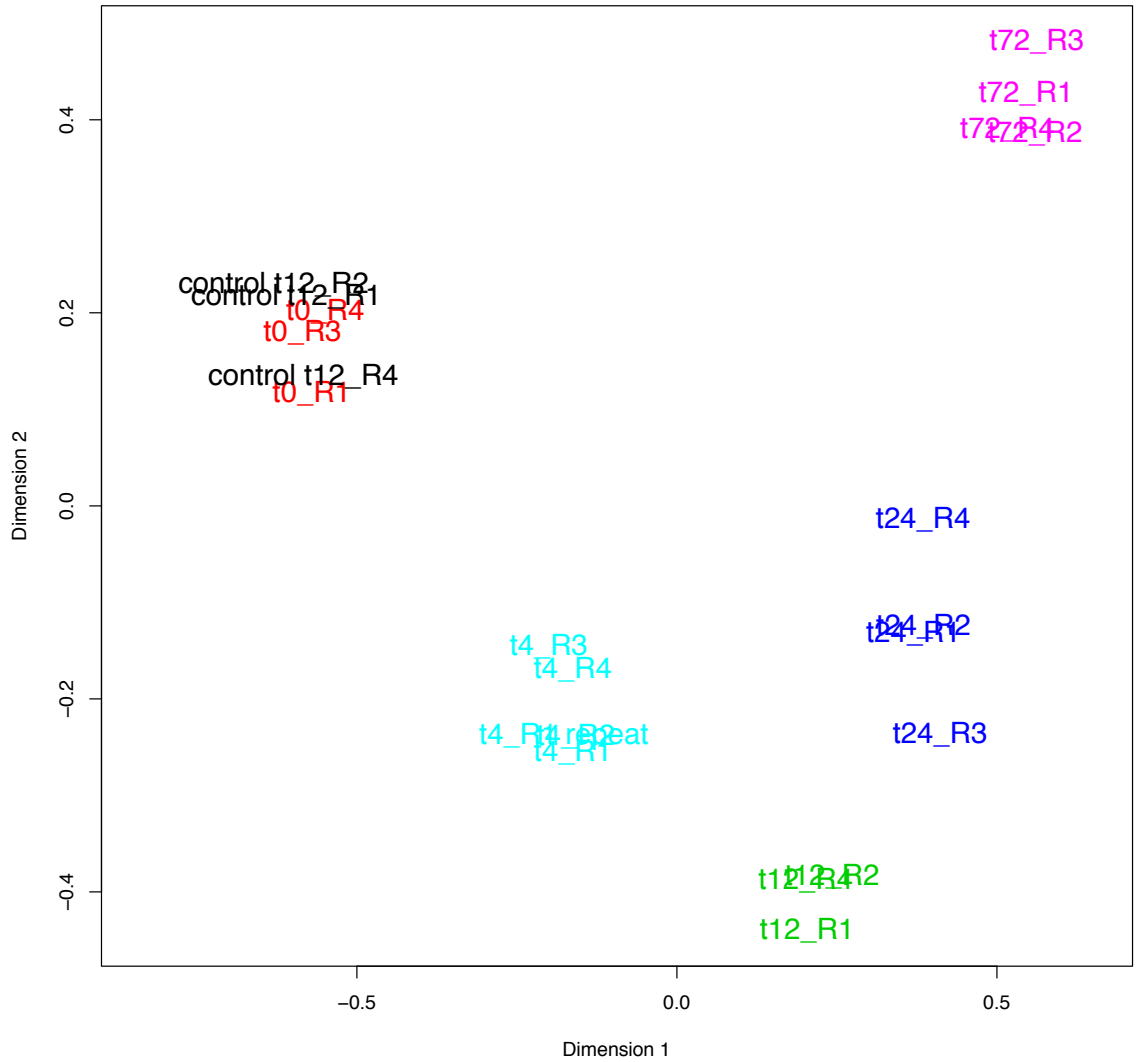
Magic: Surrogate variable analysis to detect and "remove" batch effects

OPEN ACCESS Freely available online PLOS GENETICS

Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

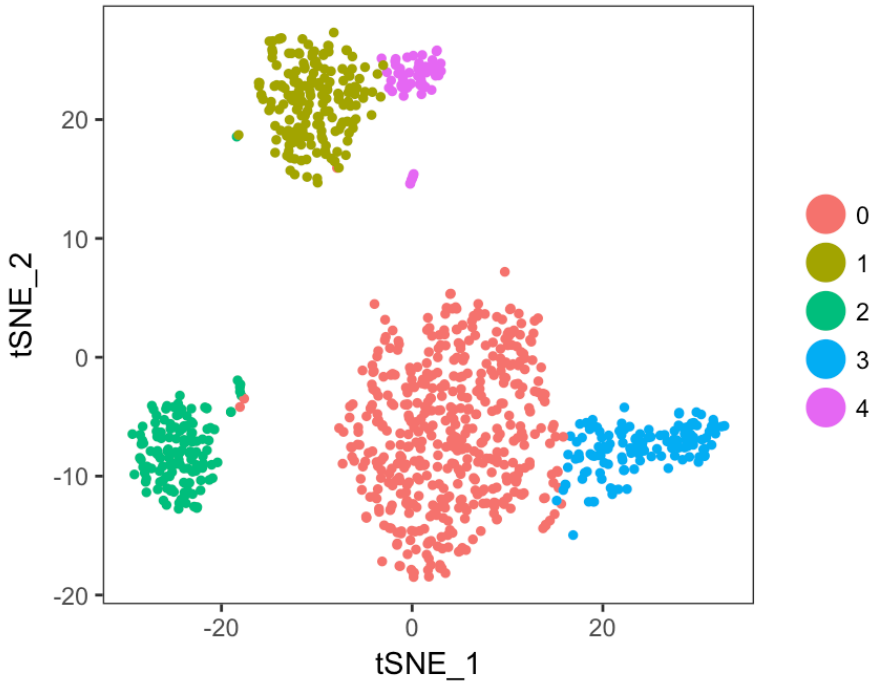
Jeffrey T. Leek[1], **John D. Storey**[1,2*]

1 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America



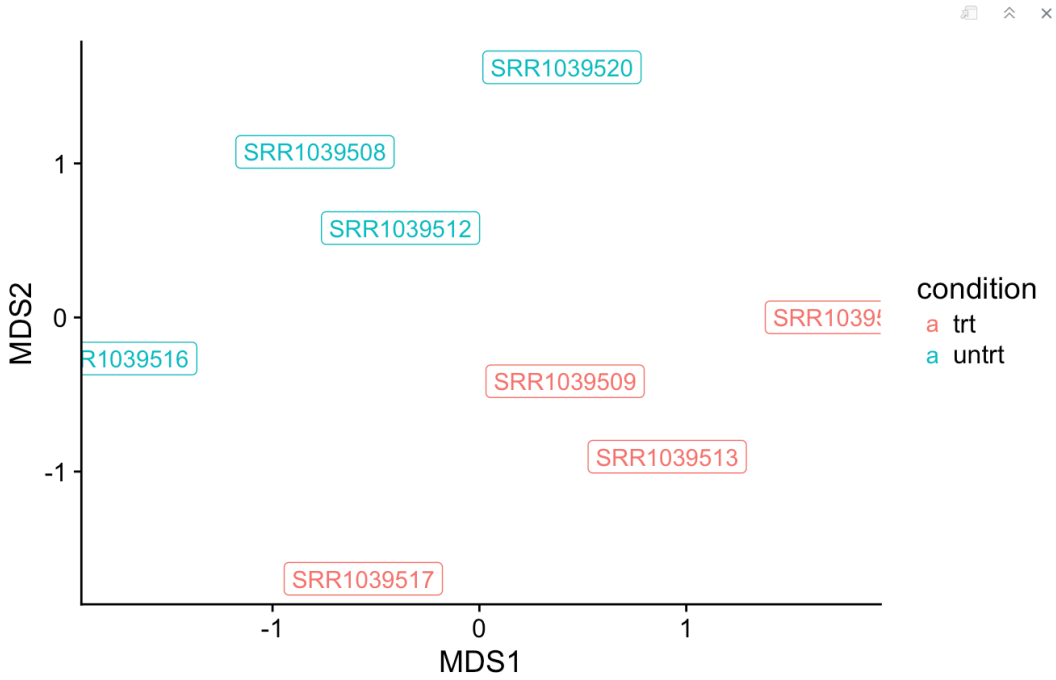
Dimension reduction: to visualize high dimensional datasets

N cells x 5000 genes markers —> N cells x 2 dimensions



Each point = **single cell**
(10x PBMC)

P samples x 10000 genes —> P samples x 2 dimensions

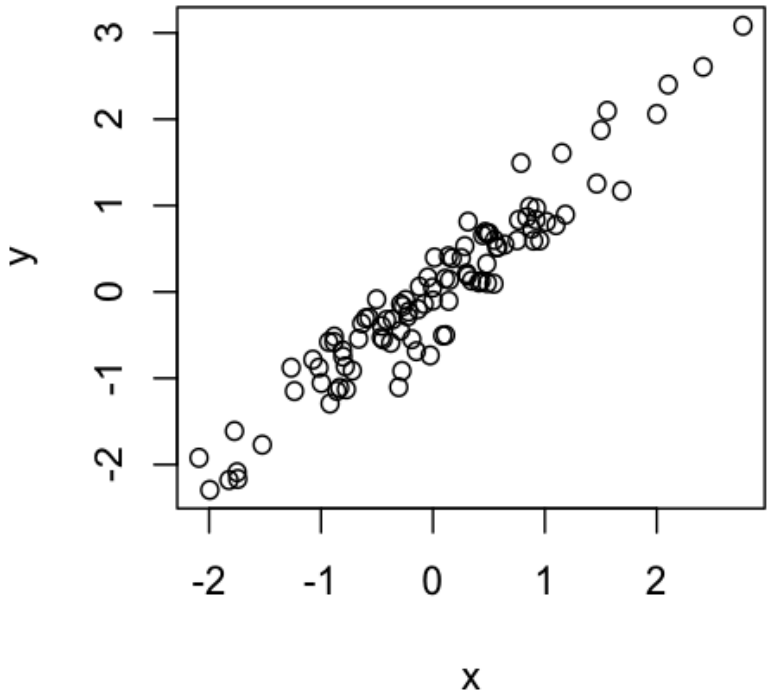


Each point = **sample**
(airway)

Dimensionality reduction (generally)

Many techniques exist to *project* high-dimensional data (our situation: 100s-1000s of cells each with measurements of 5'000 genes) into a small number of dimensions (2 or 3, for humans)

Many techniques: **linear PCA**, **multidimensional scaling**, t-distributed stochastic neighbor embedding (**tSNE**), **diffusion map**, **UMAP**, **SIMLR**, **…**

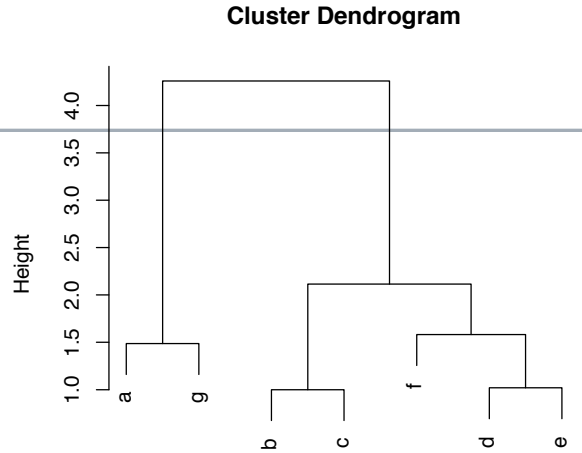


Linear PCA: uses a linear combination of original variables such that the components decrease in variability (highest variance first) and are orthogonal to previous dimensions. Often, first 2 or 3 are used for visualization.

Visual explanation:
http://setosa.io/ev/principal-component-analysis/

http://en.wikipedia.org/wiki/Hierarchical_clustering

Hierarchical (Agglomerative) Clustering

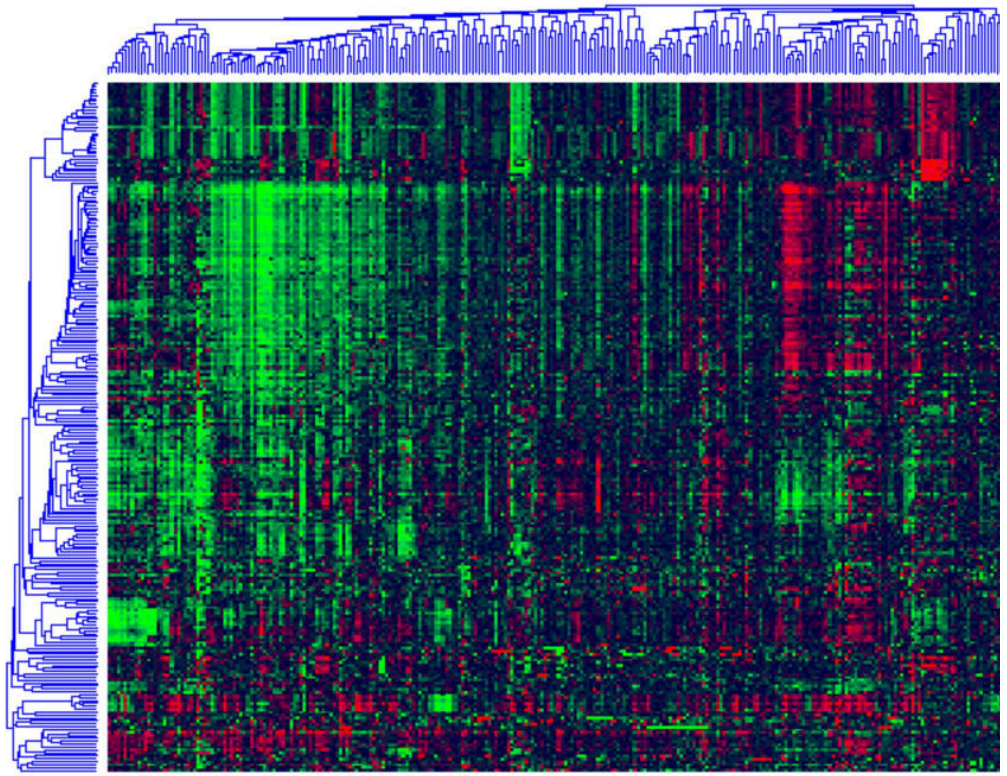


Divisive: all features start as 1 cluster, then subsequently split

Agglomerative: every feature is it's own cluster, then subsequently merged

Metric: to define how similar any two vectors are.

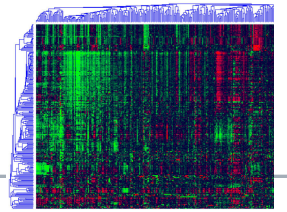
Linkage: determines how clusters are merged into a tree



http://en.wikipedia.org/wiki/Euclidean_distance

Euclidean distance:

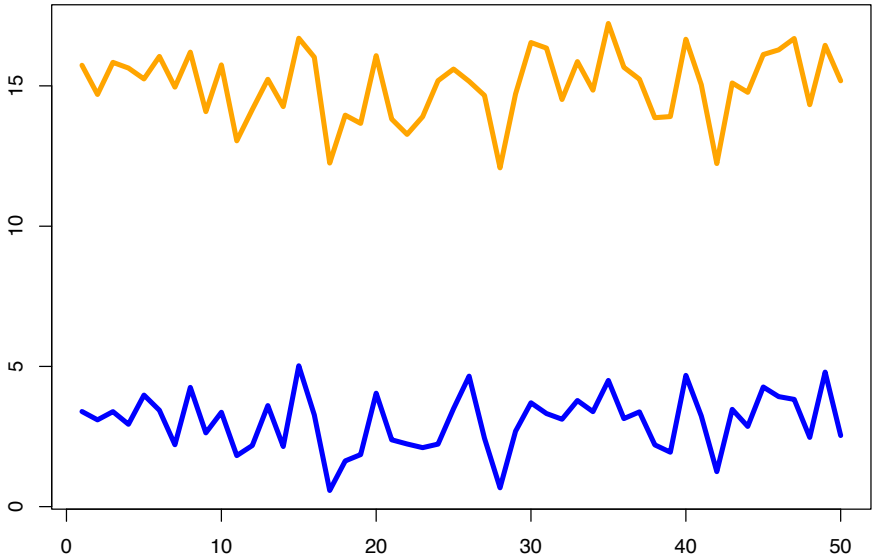
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

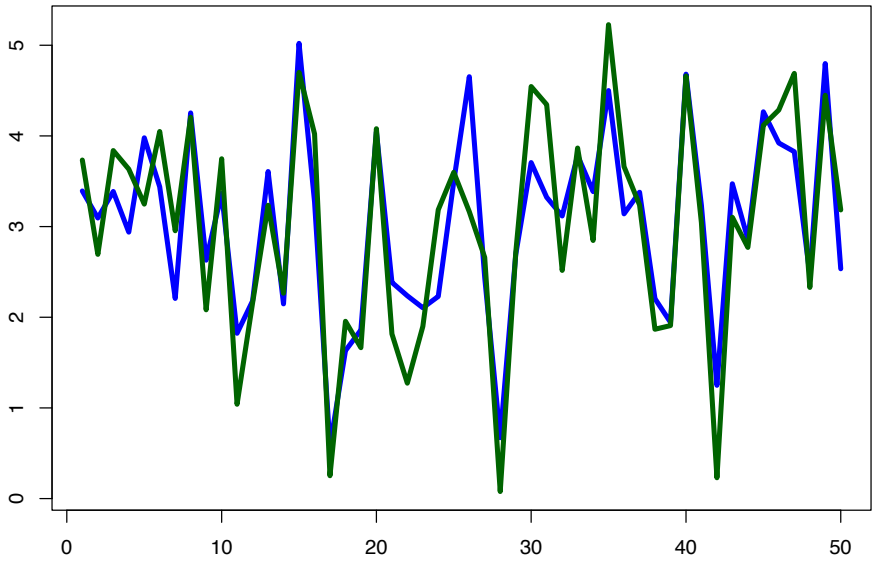


Are these “vectors” similar ?

```
> sqrt(sum((x-(y-12))^2))
[1] 3.926007
> sqrt(sum((x-y)^2))
[1] 84.84028
```

It depends how you define similar.

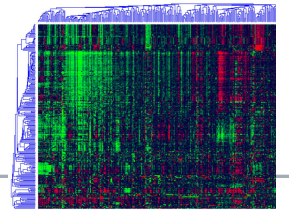




Euclidean distance: 84.84 3.92

http://en.wikipedia.org/wiki/Correlation_and_dependence

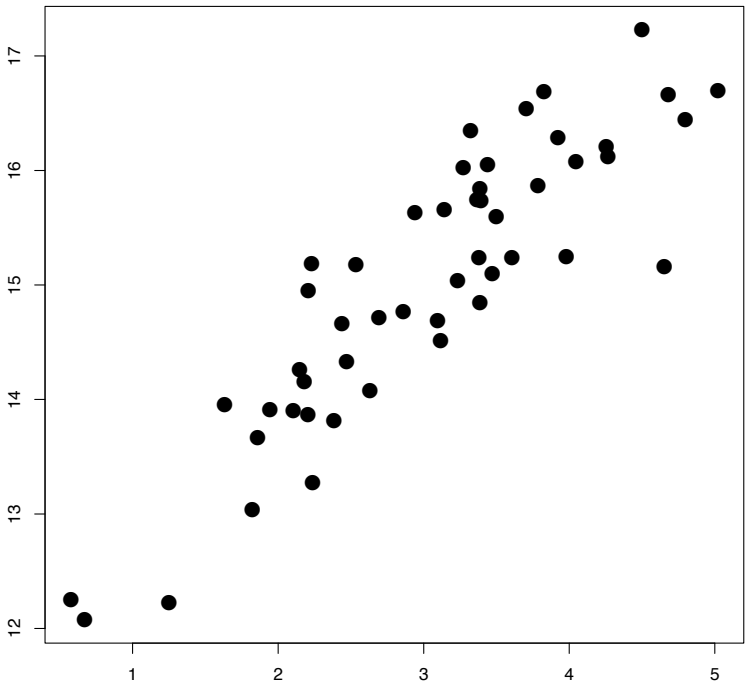
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

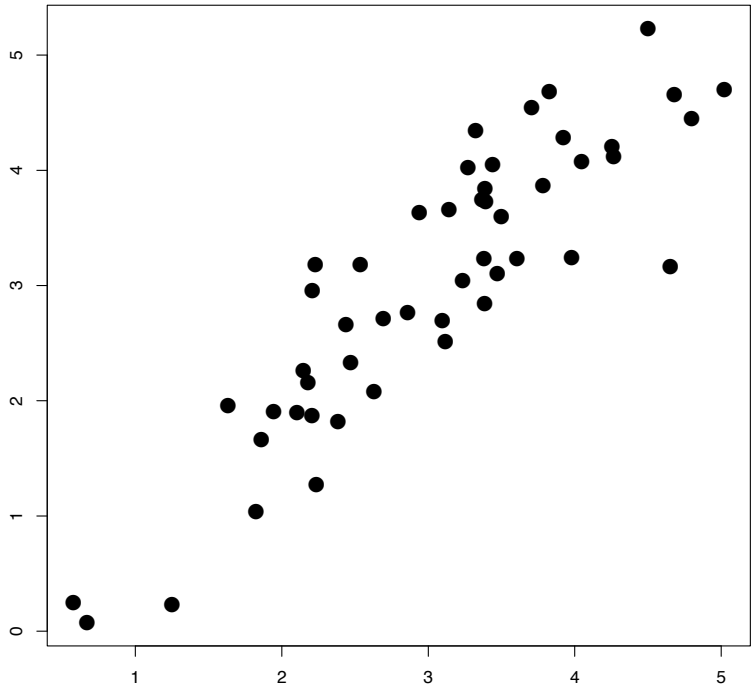


Are these “vectors” similar ?

```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

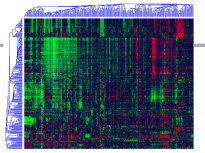
It depends how you define similar.





Correlation: 0.89 0.89

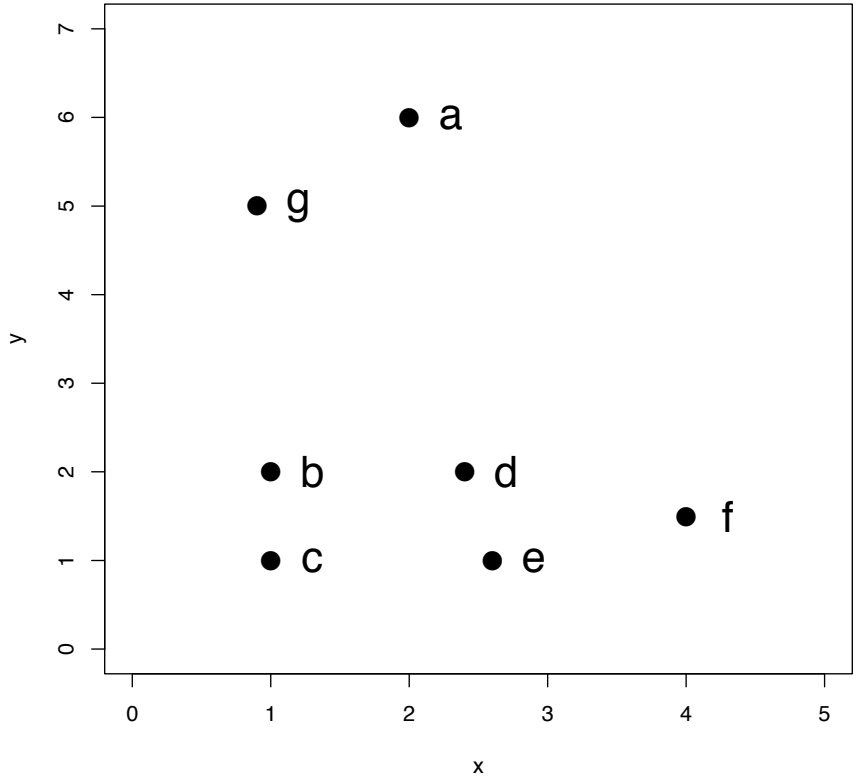
http://en.wikipedia.org/wiki/Hierarchical_clustering



Hierarchical (Agglomerative) Clustering

Start with distances.

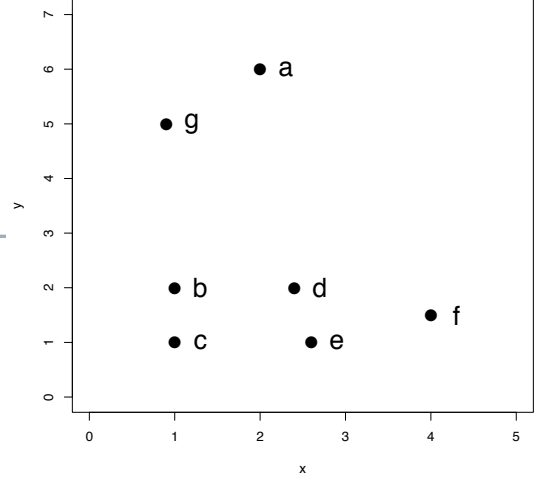
Linkage: determines how clusters are merged into a tree.

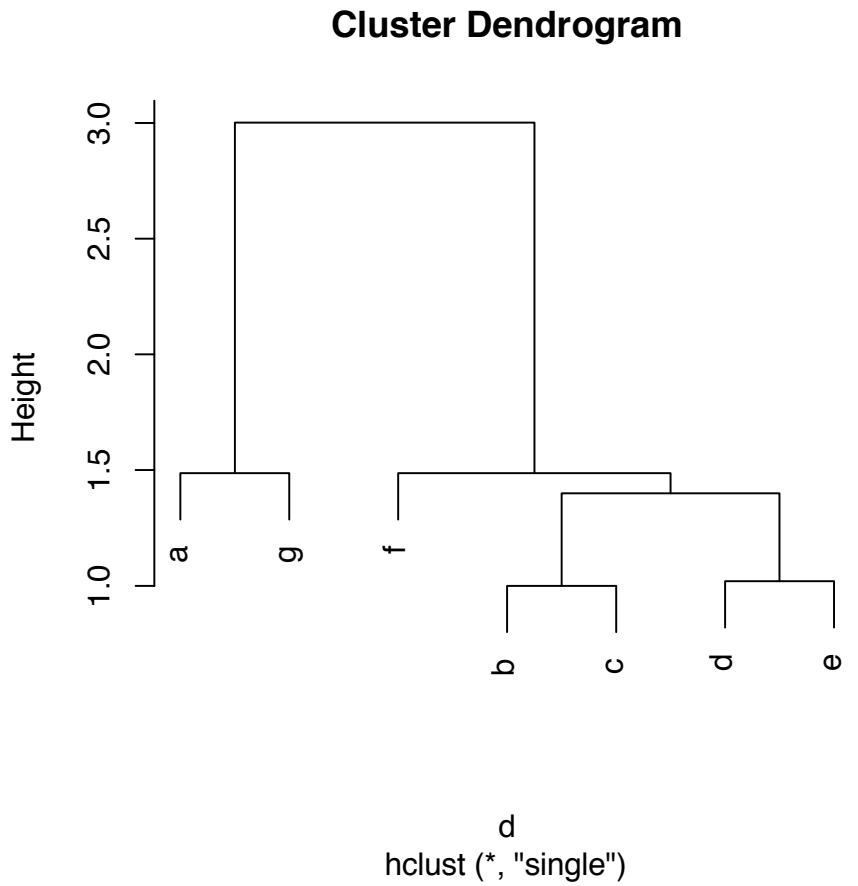


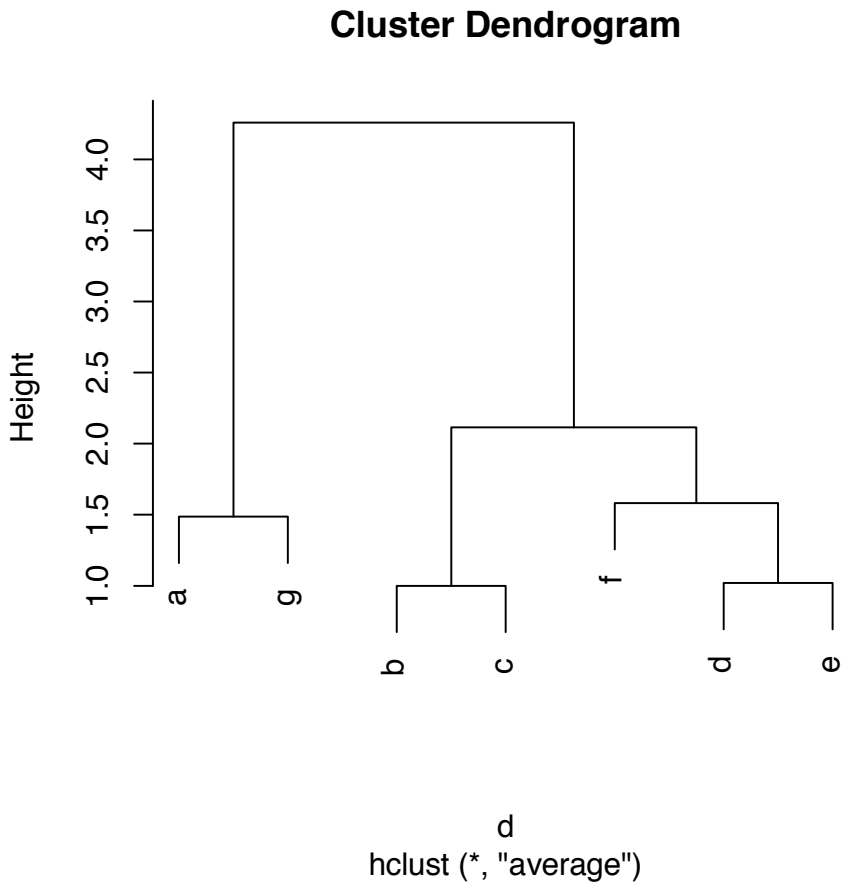
From eyeballing, here is a likely set of merges:

b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL

Different linkages



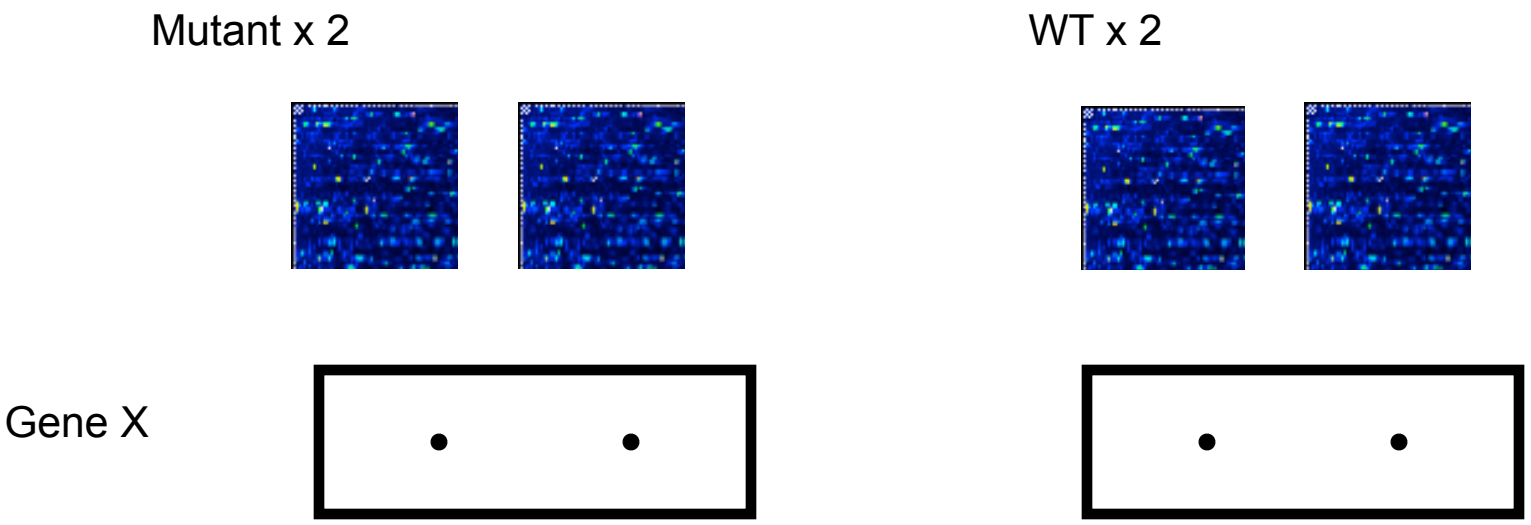




Limma concept: borrowing information across genes

- Small data sets: few samples, generally under-powered for 1 gene
- Curse of dimensionality: many tests, need to adjust for multiple testing (= loss of power)
- Benefit of parallelism: same model is fit for every gene. Can borrow information from one gene to another
  - Hard: assume parameters are constant across genes
  - Soft: smooth genewise parameters towards a common value in a graduated way, e.g., Bayes, empirical Bayes, Stein shrinkage …

A very common experiment (1-colour)



Which genes are differentially expressed?

$n_1 = n_2 = 2$ Affymetrix arrays

~30,000 probe-sets

Ordinary t-tests (1-colour)

$$t_g = \frac{\overline{y}_{\mathrm{mu}} - \overline{y}_{\mathrm{wt}}}{s_g\, c}$$

give very high false discovery rates

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Residual df = 2

t-tests with common variance

$$t_{g,\text{pooled}} = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{s_0\, c}$$

with residual standard deviation s_0 pooled across genes

More stable, but ignores gene-specific variability

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

A better compromise

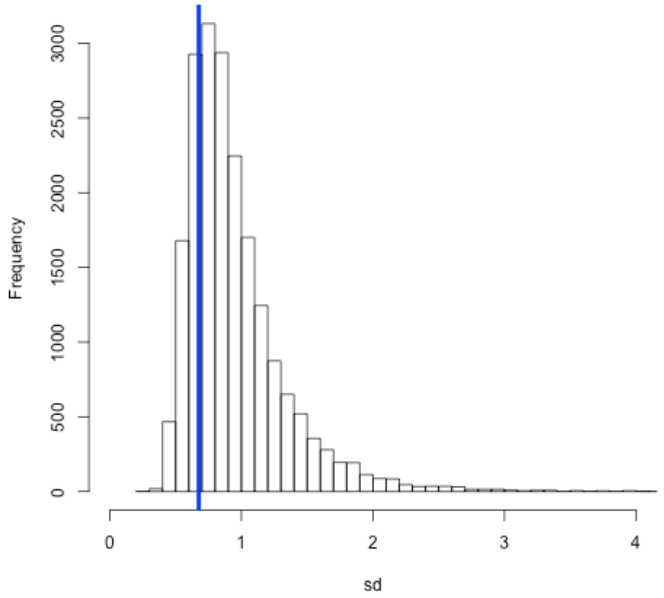
Shrink standard deviations towards common value

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

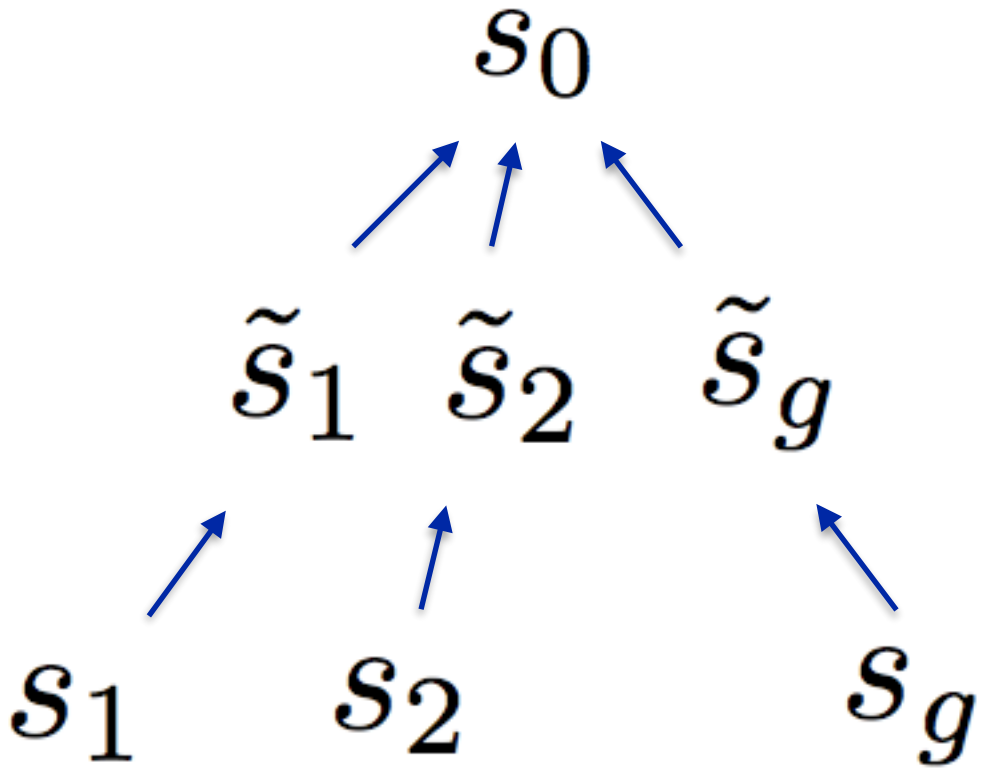
d = degrees of freedom

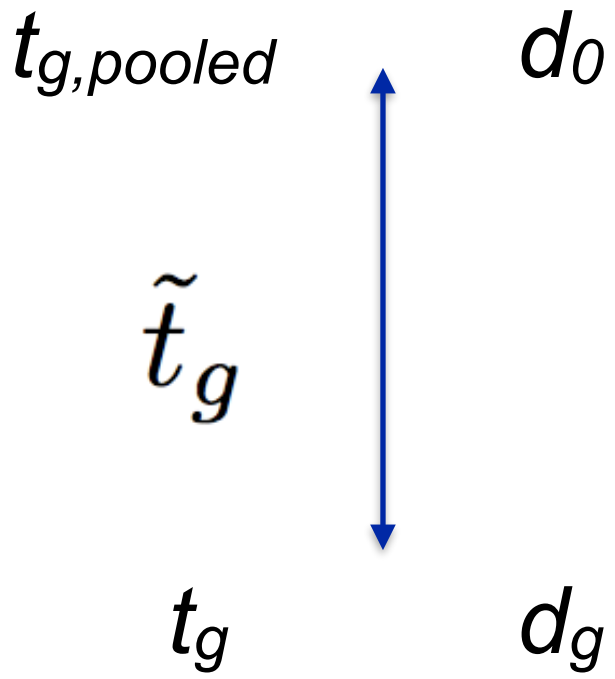
Moderated t-statistics

$$\tilde{t}_g = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{\tilde{s}_g \, u}$$



Shrinkage of standard deviations





The **data decides** whether $\tilde{t}_g$ should be closer to $t_{g,pooled}$ or t_g

Why does it work?

- We learn what is the typical variability level by looking at all genes, but allow some flexibility from this for individual genes
- Adaptive – data (through hyperparameter estimates, d_0 and s_0) suggests how much to “squeeze” toward common value

Hierarchical model for variances

Data $$s_g^2 \sim \sigma_g^2 \frac{\chi^2_{d_g}}{d_g}$$

Prior $$\frac{1}{\sigma_g^2} \sim s_0^2 \frac{\chi^2_{d_0}}{d_0}$$

Posterior $$E\left(\frac{1}{\sigma_g^2} \mid s_g^2\right) = \frac{d_0 + d_g}{s_0^2 d_0 + s_g^2 d_g}$$

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_0^2 d_0 + s_g^2 d_g}{d_0 + d_g}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Baldi & Long 2001, Wright & Simon 2003, Smyth 2004

Exact distribution for moderated t

An unexpected piece of mathematics shows that, under the null hypothesis,

$$\tilde{t}_g \sim t_{d_0+d_g}$$

The degrees of freedom add!

The Bayes prior in effect adds d_0 extra arrays for estimating the variance.

Wright and Simon 2003, Smyth 2004

Multiple testing and adjusted p-values

- Each statistical test has an associated false positive rate
- Traditional method in statistics is to control family wise error rate, e.g., by Bonferroni.
- Controlling the false discovery rate (FDR) is more **appropriate** in microarray studies
- Benjamini and Hochberg method controls expected FDR for independent or weakly dependent test statistics. Simulation studies support use for genomic data.
- All methods can be implemented in terms of adjusted p-values.