



BIO390 Introduction to Bioinformatics

Statistical Bioinformatics: motivation via data examples
(1st hour), some fundamental concepts (2nd hour)

Upcoming

October 04, 2022
Statistical Bioinformatics
BIO390 UZH HS22 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85
Mark Robinson
[→ Continue reading](#)

October 11, 2022
Machine Learning for Biological Use Cases
BIO390 UZH HS22 - INTRODUCTION TO BIOINFORMATICS
08:00-09:45 @ UZH IRCHEL Y03-G-85
Valentina Boeva (ETHZ)
Brief note: In this lecture V. Boeva will cover the standard machine learning methods used in the analysis of biological data: dimensionality reduction, clustering, classification and regression.

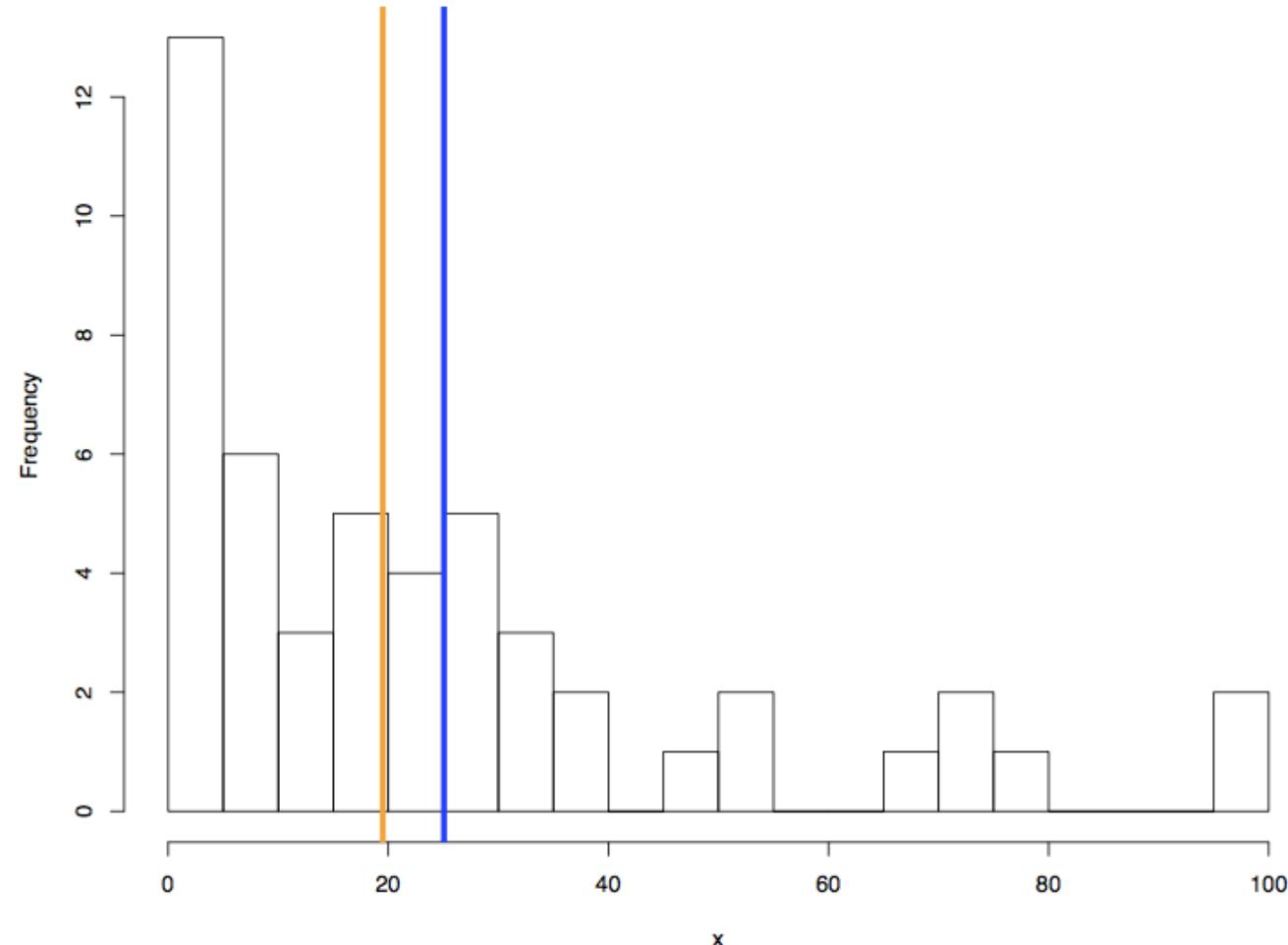


Survey: Statistical Insight

klicker

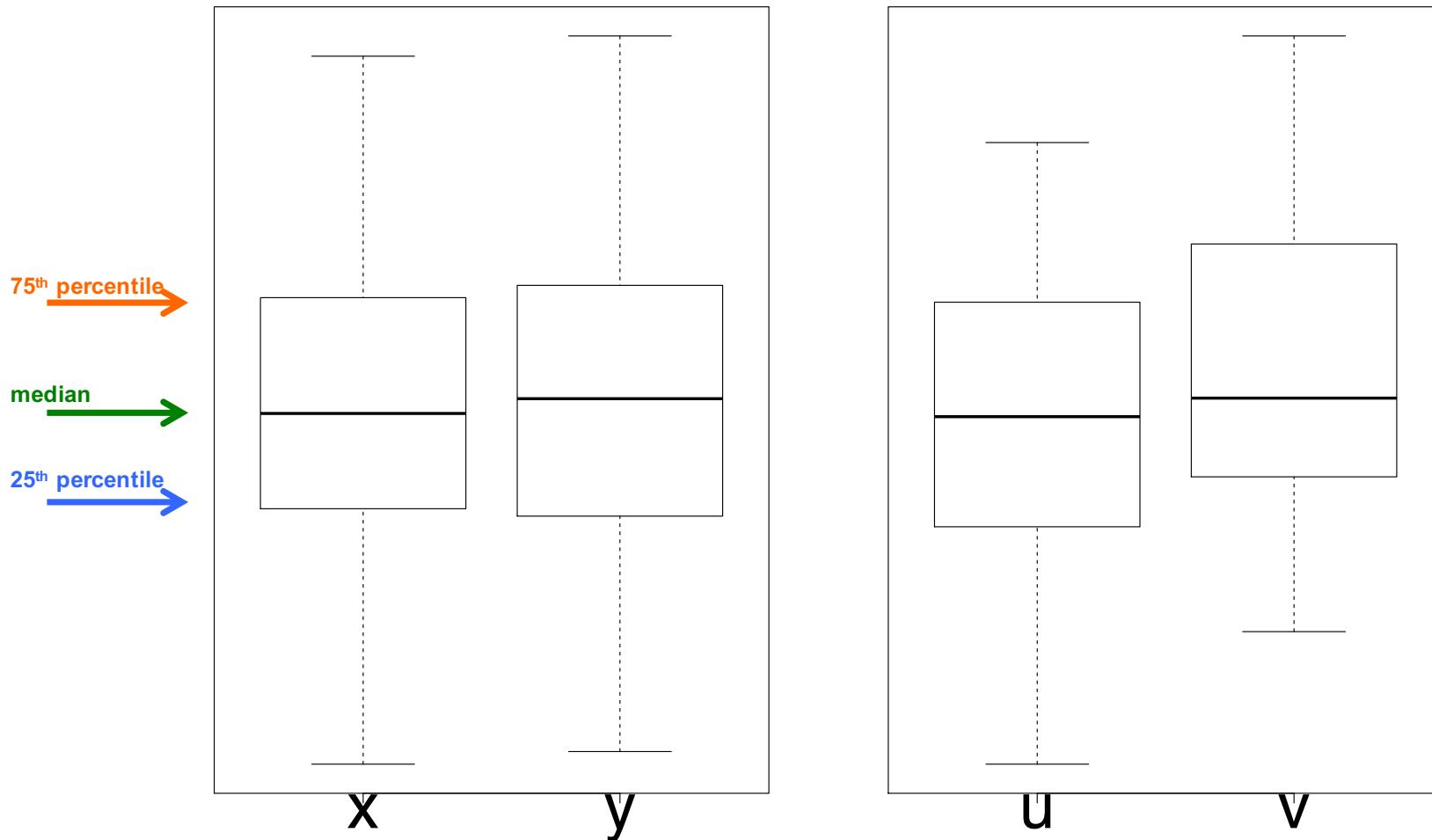


From the histogram, determine whether blue or orange represents the mean/median



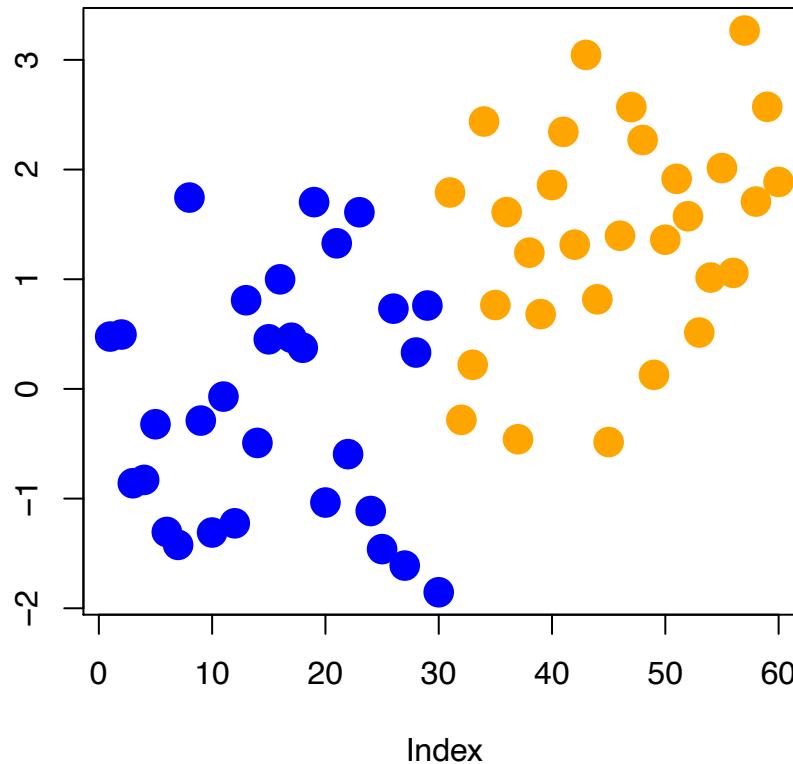


Given these boxplots, which of two underlying distributions are more similar?

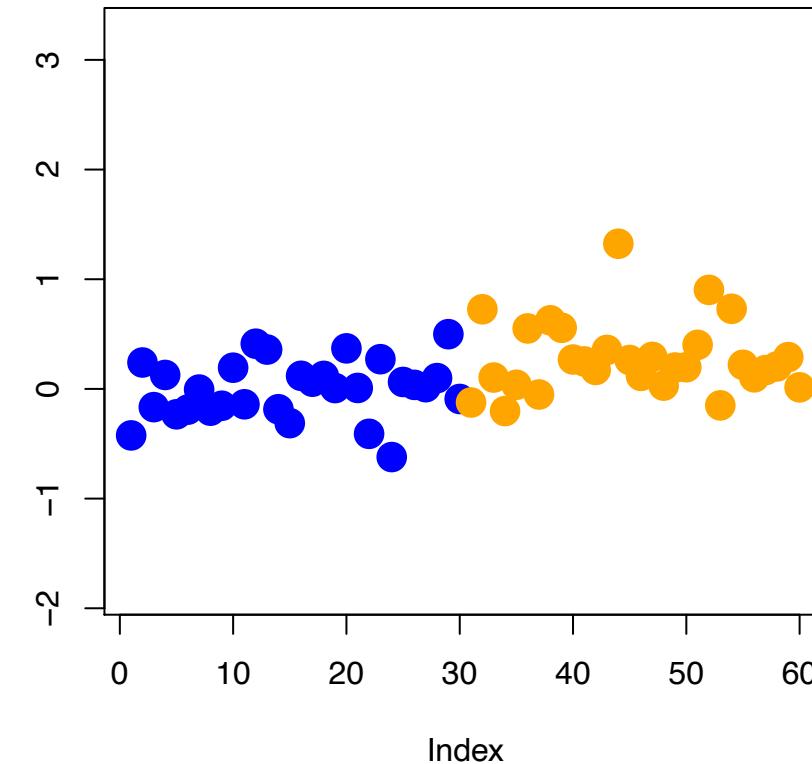


Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?

A

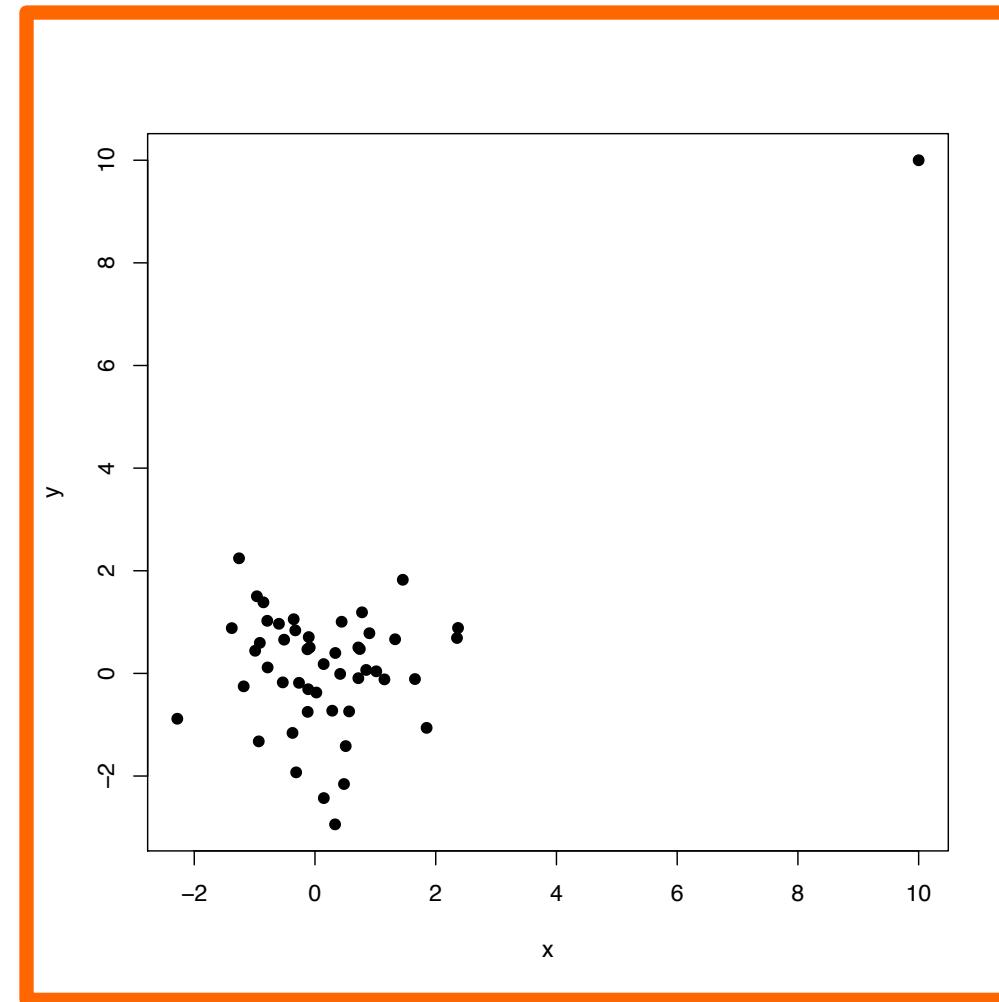


B



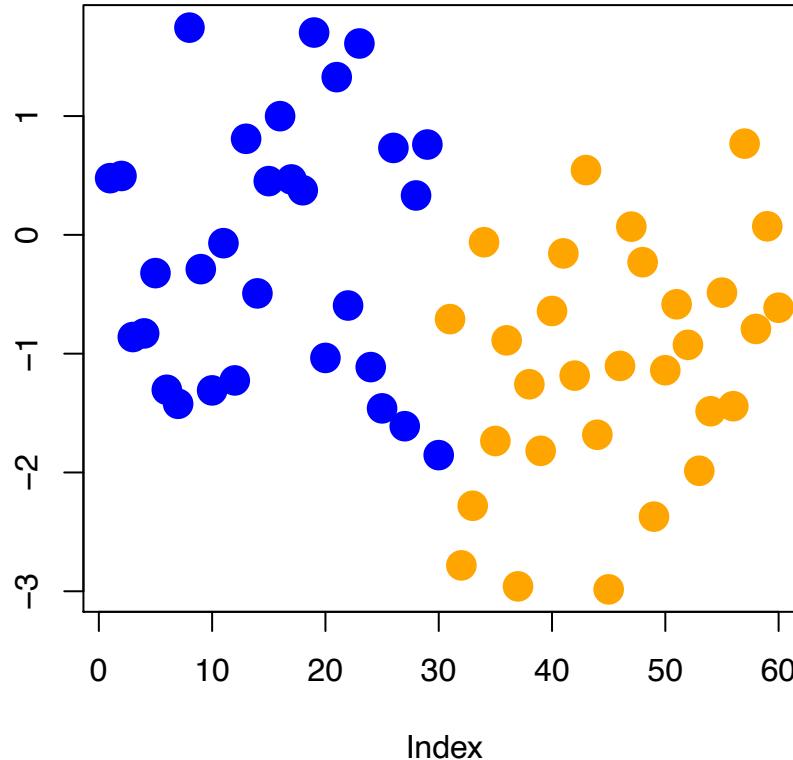


In your view, what best describes the associations shown in the plot of 'x' and 'y' ?

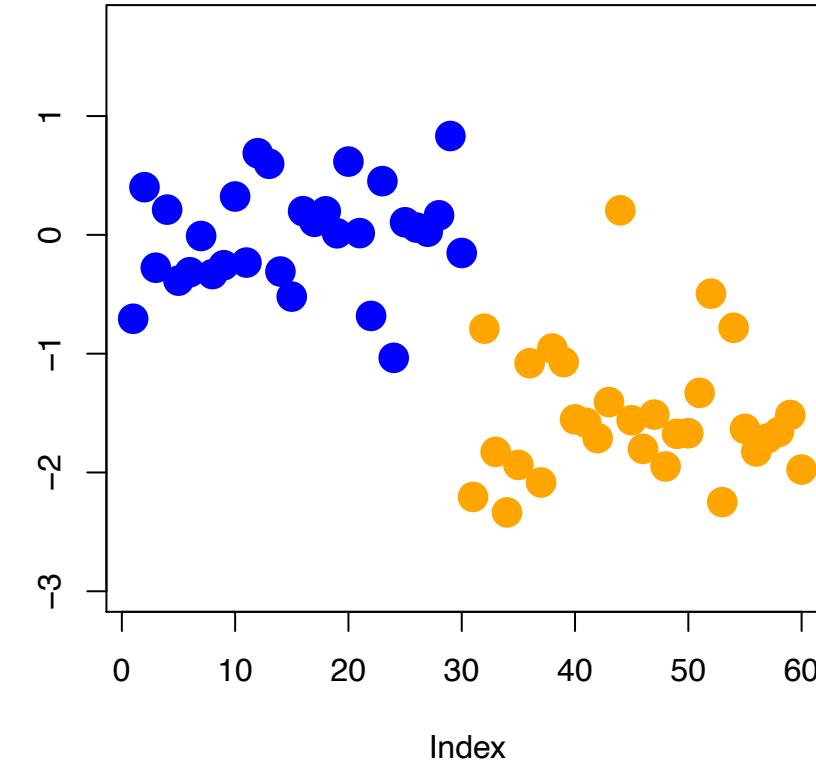


Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?

A



B





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

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



Outline

- Motivation: nowadays we are inundated with data, such as microarrays, sequencing, cytometry, imaging —> modern biologists need to be data-savvy (data science, statistics, computation)
- Fundamental statistical concepts: central limit theorem, false positives / false negatives, P-values, multiple testing, exploratory data analysis, regression, clustering, dimension reduction, reproducibility, ...
- Data science / programming: BIO 134, BIO 144, (BIO 334, STA 426)



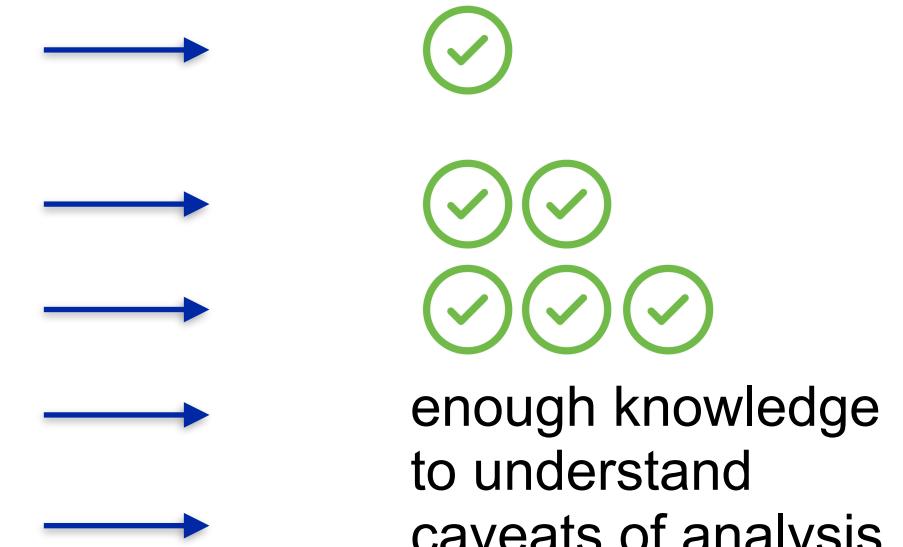
Critical skills needed by statisticians (Jeffrey Leek's words):

With all the excitement going on around statistics, there is also increasing diversity. It is increasingly hard to define “statistician” since the definition ranges from very mathematical to very applied. An obvious question is: what are the most critical skills needed by statisticians?

So just for fun, I made up my list of the top 5 most critical skills for a statistician by my own definition. They are by necessity very general (I only gave myself 5).

1. **The ability to manipulate/organize/work with data on computers** - whether it is with excel, R, SAS, or Stata, to be a statistician you have to be able to work with data.
2. **A knowledge of exploratory data analysis** - how to make plots, how to discover patterns with visualizations, how to explore assumptions
3. **Scientific/contextual knowledge** - at least enough to be able to abstract and formulate problems. This is what separates statisticians from mathematicians.
4. **Skills to distinguish true from false patterns** - whether with p-values, posterior probabilities, meaningful summary statistics, cross-validation or any other means.
5. **The ability to communicate results to people without math skills** - a key component of being a statistician is knowing how to explain math/plots/analyses.

Modern biologists



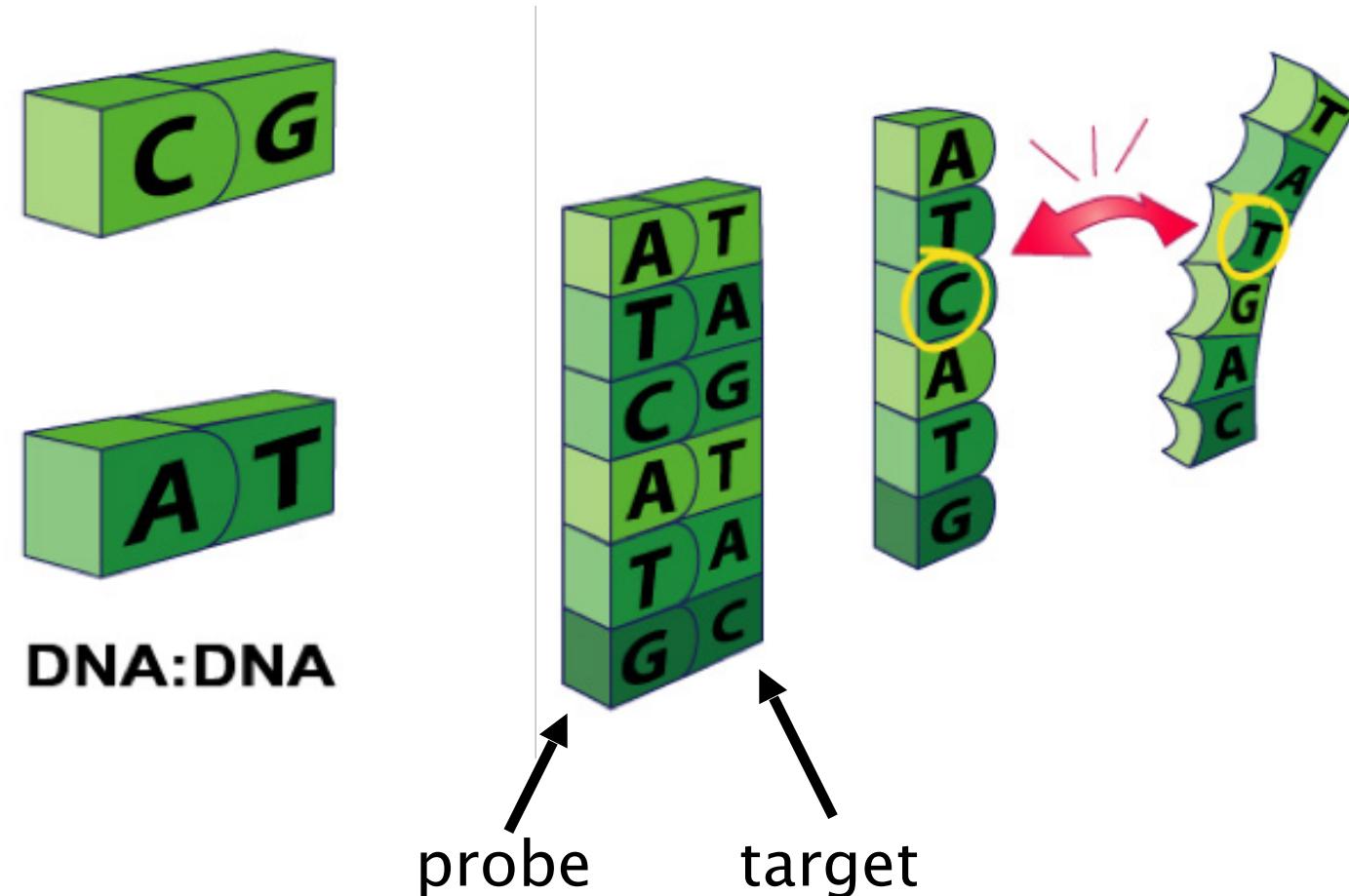
Technologies in my research area

microarray, high-throughput sequencing, single cell, cytometry, etc.

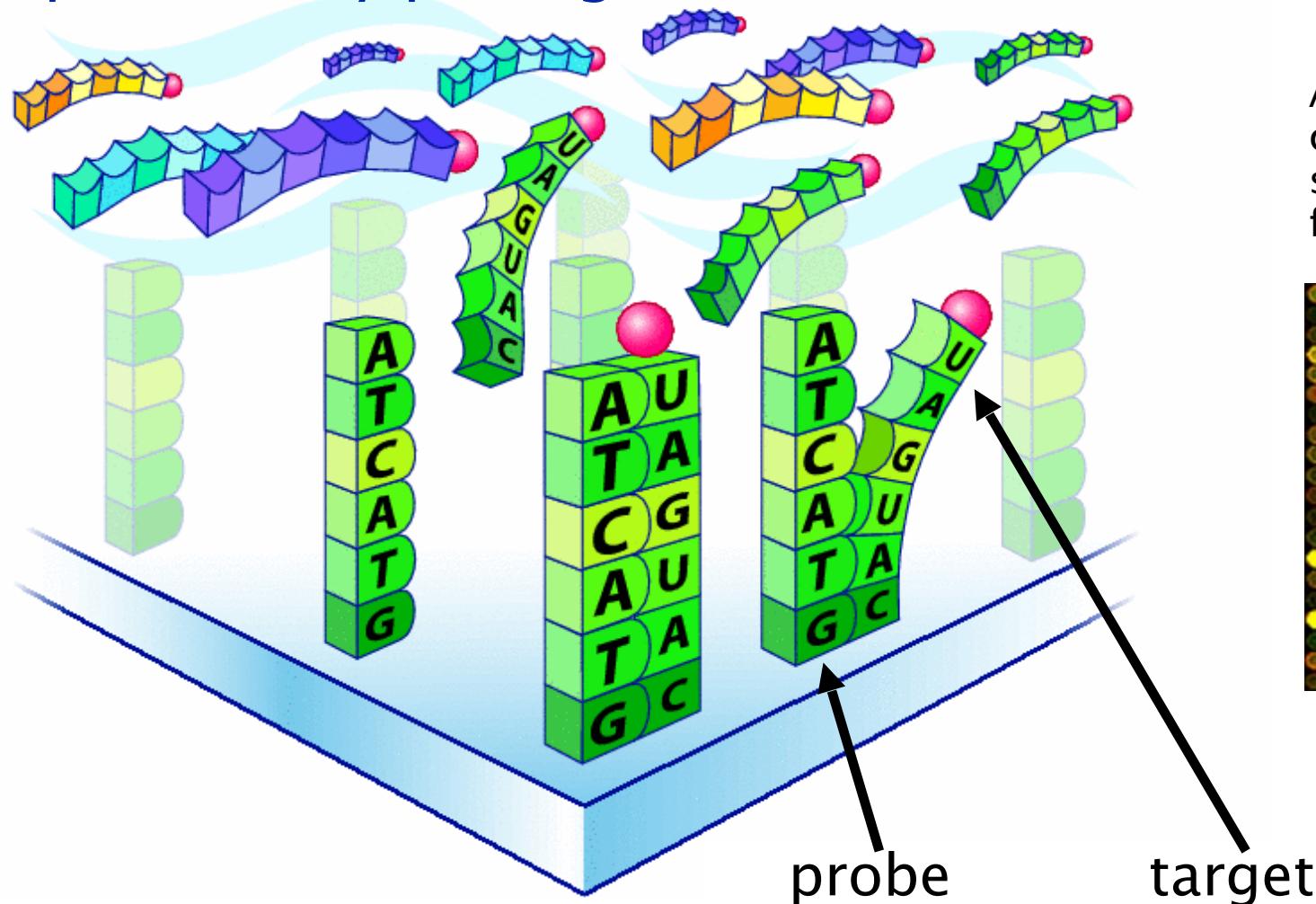
“it's just data”



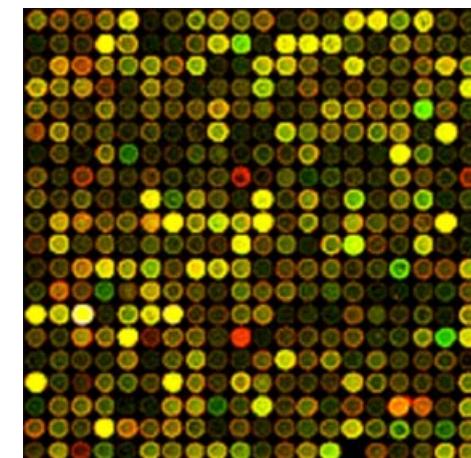
Microarray fundamentals: Nature gives a complementary pairing



DNA microarray: parallel northern blots; Nature gives a complementary pairing



Abundance (of complementary DNA species) measured by fluorescence 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



“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

Motivation for exploratory data analysis: Case Study

(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:

CTL	t0		t12		
TRT		t4	t12	t24	t72



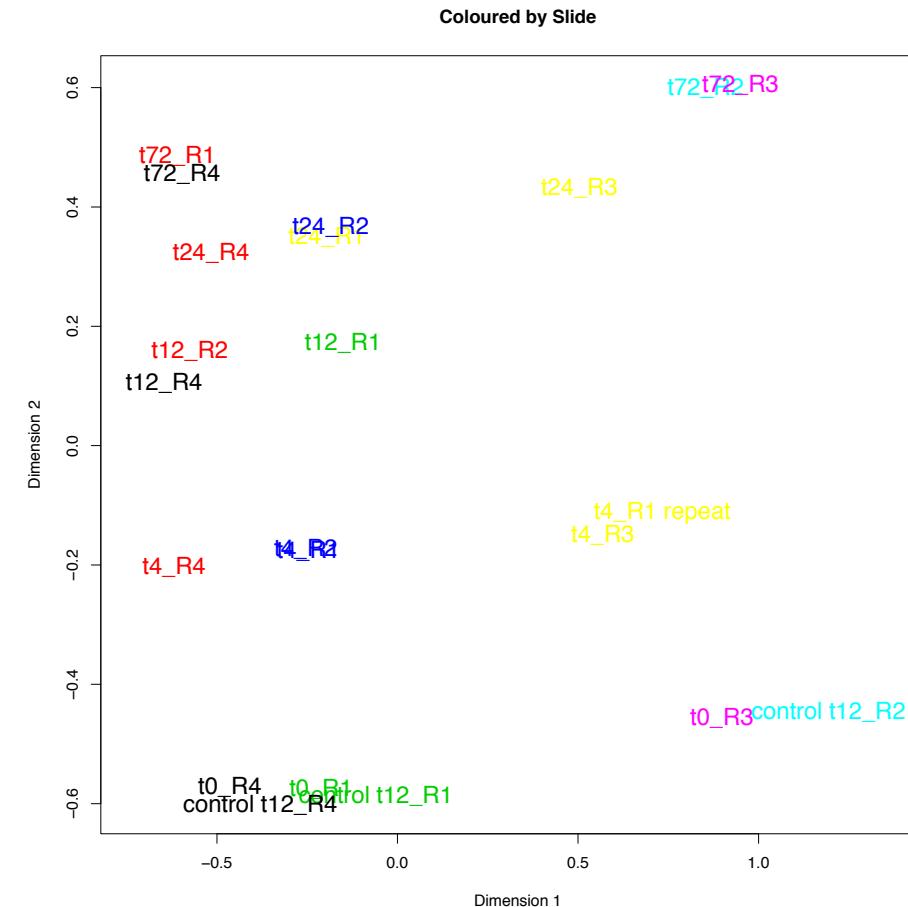
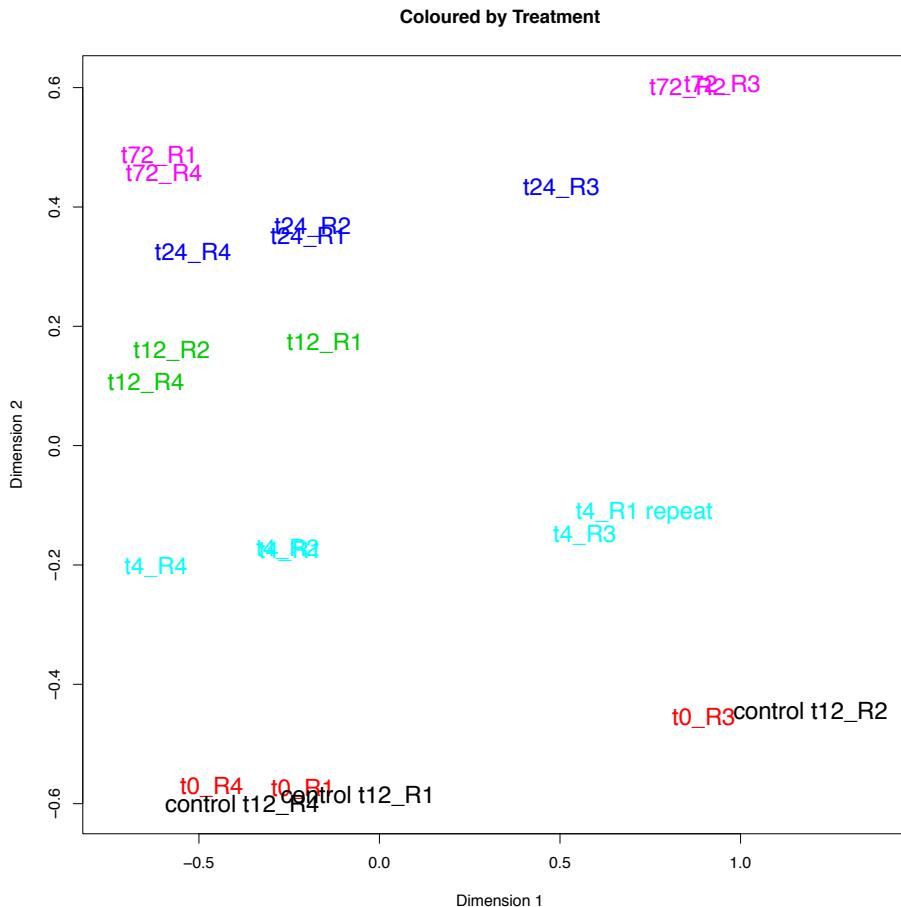
Change to lower
temperature.

~4 replicates for each condition



```
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

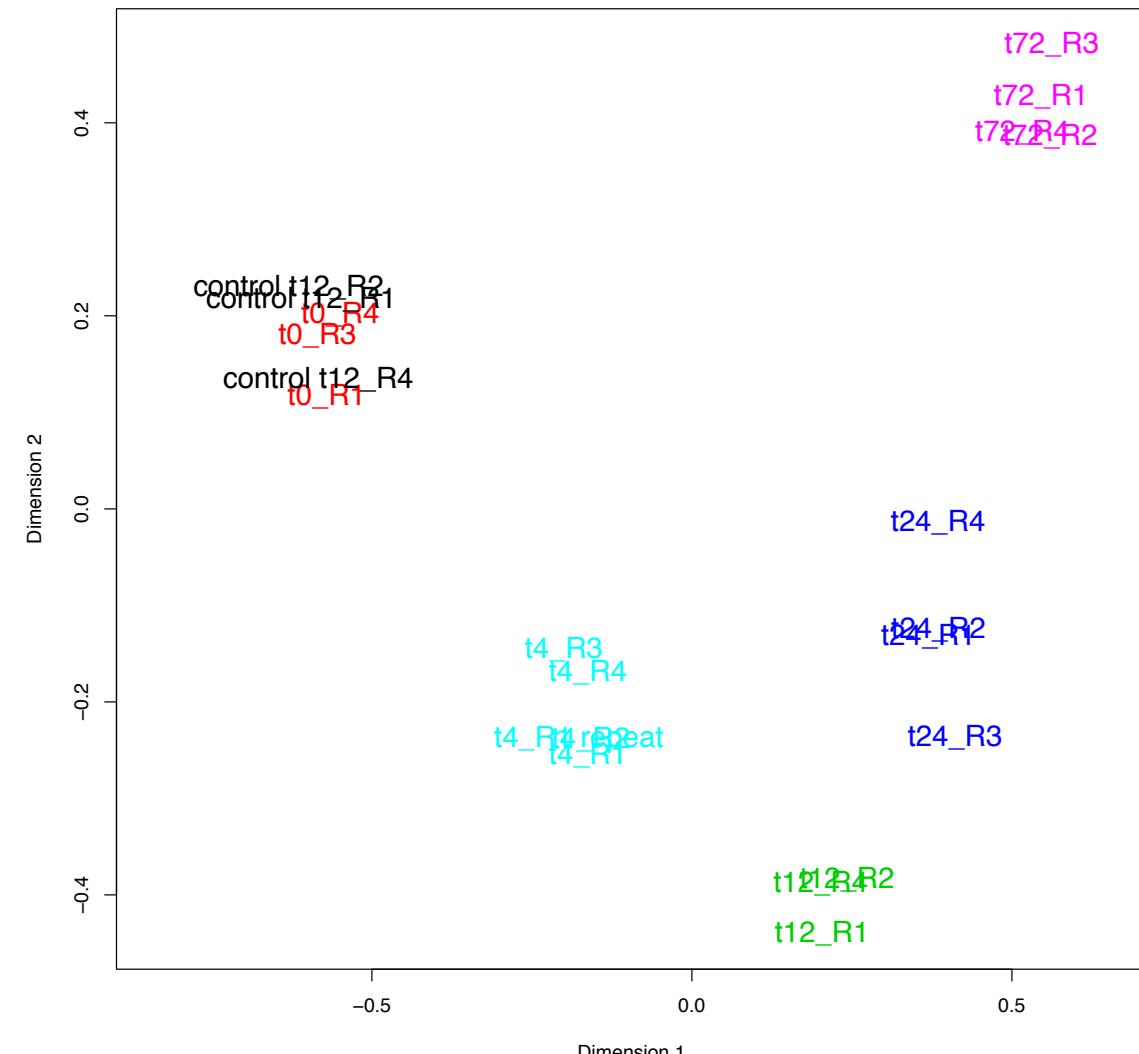
OPEN ACCESS Freely available online

PLOS GENETICS

Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

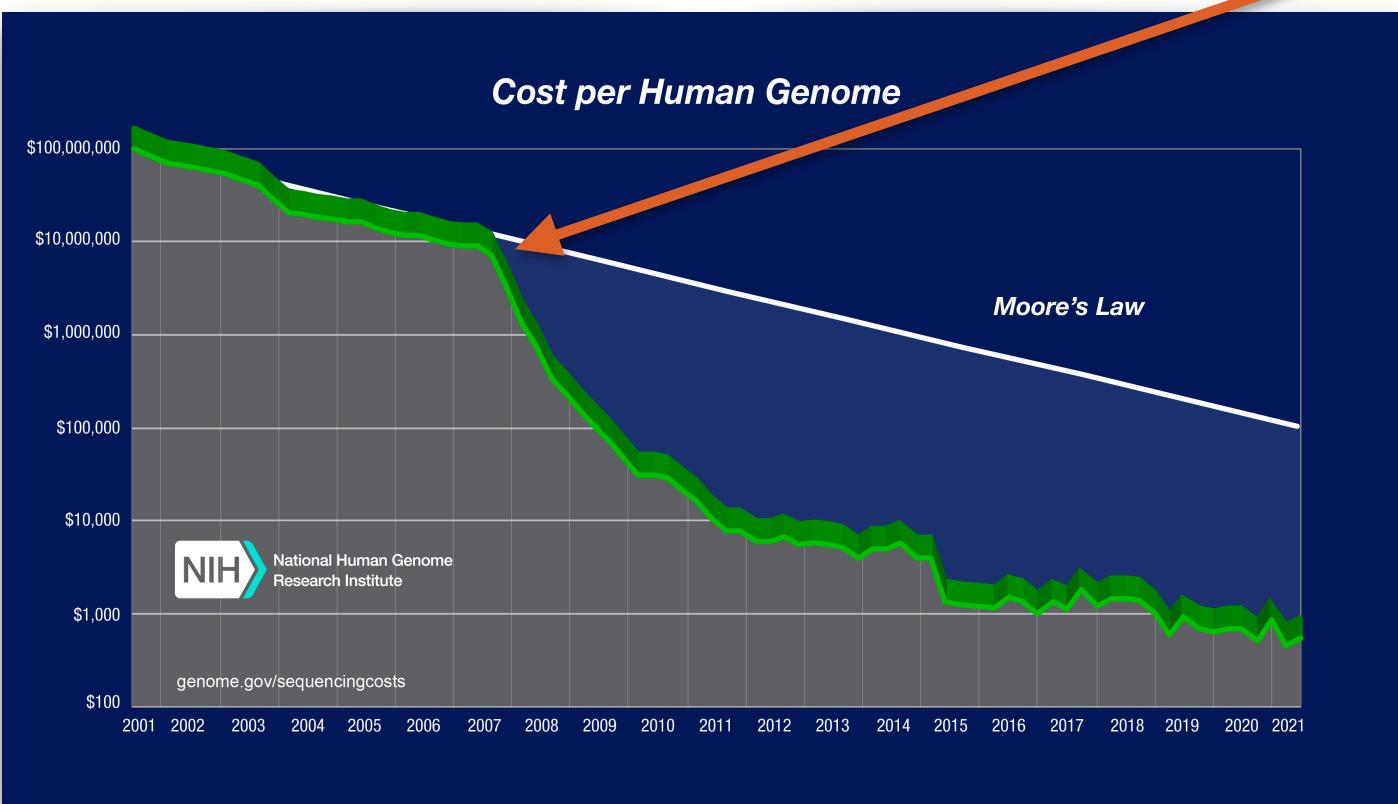
Jeffrey T. Leek¹, John D. Storey^{1,2*}

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





High-throughput sequencing



(Solexa) Illumina

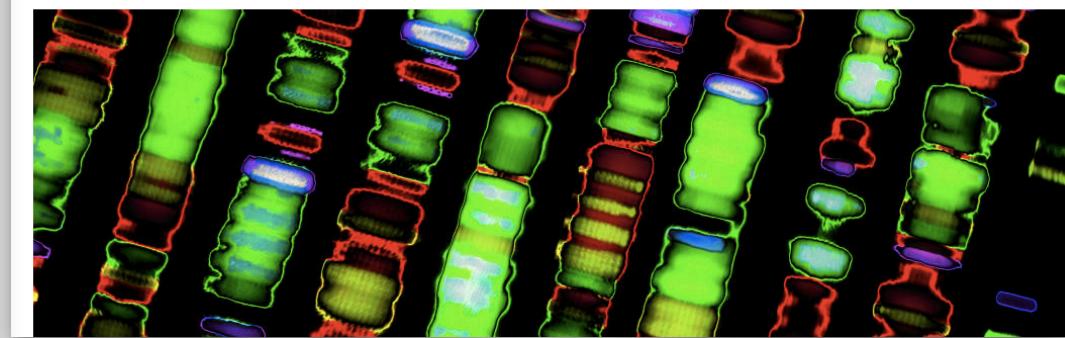
<https://www.statnews.com/2017/01/09/illumina-ushering-in-the-100-genome/>

BUSINESS

STAT+

Illumina says it can deliver a \$100 genome — soon

By MEGHANA KESHAVAN @megkesh / JANUARY 9, 2017



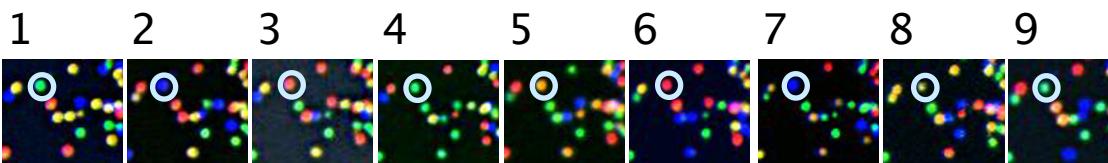
Sept 29th 2022 - **NovaSeq X Series**, unveiled earlier today, ushers in the era of the genome with revolutionary new production-scale sequencers .. can generate more than 20,000 whole genomes per year – 2.5 times the throughput of prior sequencers – greatly accelerating genomic discovery and clinical insights, to understand disease and ultimately transform patient lives.¹⁸



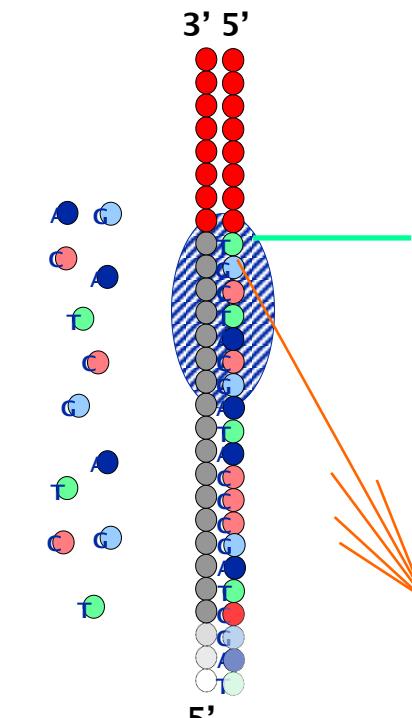
Illumina Sequencing Technology

DNA
(0.1–1.0 ug)

Sample
preparation



Cluster growth



Sequencing

T G C T A C G A T ...

Image acquisition

Base calling

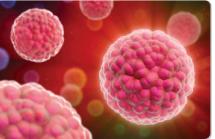


Applications of high-throughput sequencing

Common Sequencing Applications

Cancer Research

NGS-based sequencing enables cancer researchers to detect rare somatic variants, tumor subclones, and circulating DNA fragments. [Learn more about sequencing for cancer research.](#)



Microbiology Research

From environmental metagenomics studies to infectious disease surveillance and more, NGS-based sequencing can help researchers gain genetic insight into bacteria and viruses. [Learn more about microbial genomics.](#)



Complex Disease Research

Illumina sequencing is introducing new avenues for understanding immunological, neurological, and other complex disorders on a molecular level. [Learn more about complex disease genomics.](#)



Reproductive and Genetic Health

Illumina sequencing and array technologies deliver fast, accurate information that can guide choices along the reproductive and genetic health journey. [Find reproductive and genetic health solutions.](#)



ETH zürich  University of
Zurich^{UZH}

Functional Genomics Center Zurich

About us [Working with us](#) OMICS areas Education Research & Publications FAQ News & Events

ETH Zurich > UZH > FGCZ

Services

Proteomics/Protein analysis services
[Genomics/Transcriptomics services](#)
Metabolomics/Biophysics services

User Lab Access
Collaboration
FGCZ Policies
Job Offers
FGCZ Terms and Conditions

Genomics/Transcriptomics services

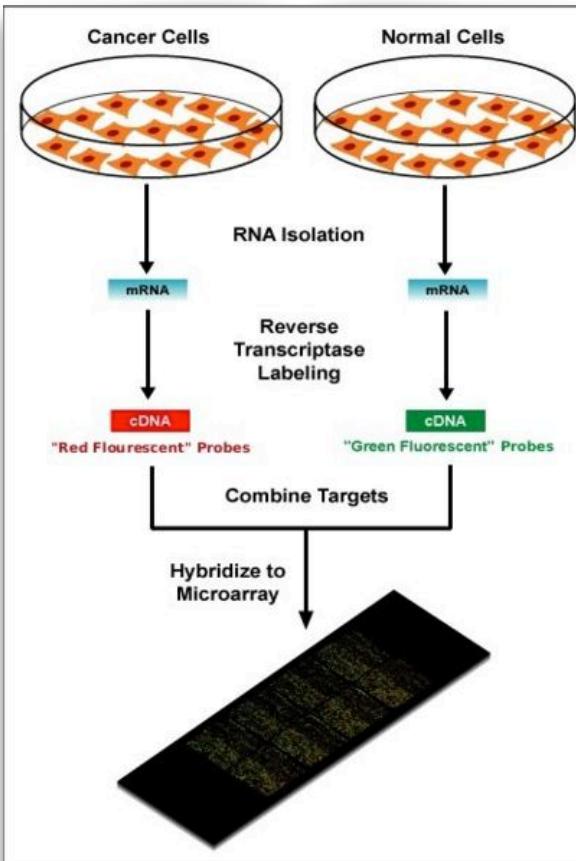
All services in Genomics/Transcriptomics require a project submission via B-Fabric, our project management system.

If you have specific questions about our Genomics/Transcriptomics services please refer to our [FAQ](#) section; alternatively, or in case you would like to request a quote, please do not hesitate to get in touch with our sequencing team at sequencing@fgc.z.ethz.ch

Application Group	Application	Order via B-Fabric
DNA sequencing	Whole Exome Sequencing	Project
DNA sequencing	Methylation Profiling	Project
DNA sequencing	ChIP-Seq	Project
DNA sequencing	Targeted Sequencing and Metagenomics	Project
DNA sequencing	De novo Genome Assembly	Project
DNA sequencing	Whole Genome Resequencing	Project
RNA sequencing	Transcriptome Profiling	Project
RNA sequencing	Small RNA Profiling	Project
RNA sequencing	De novo Transcriptome Assembly	Project

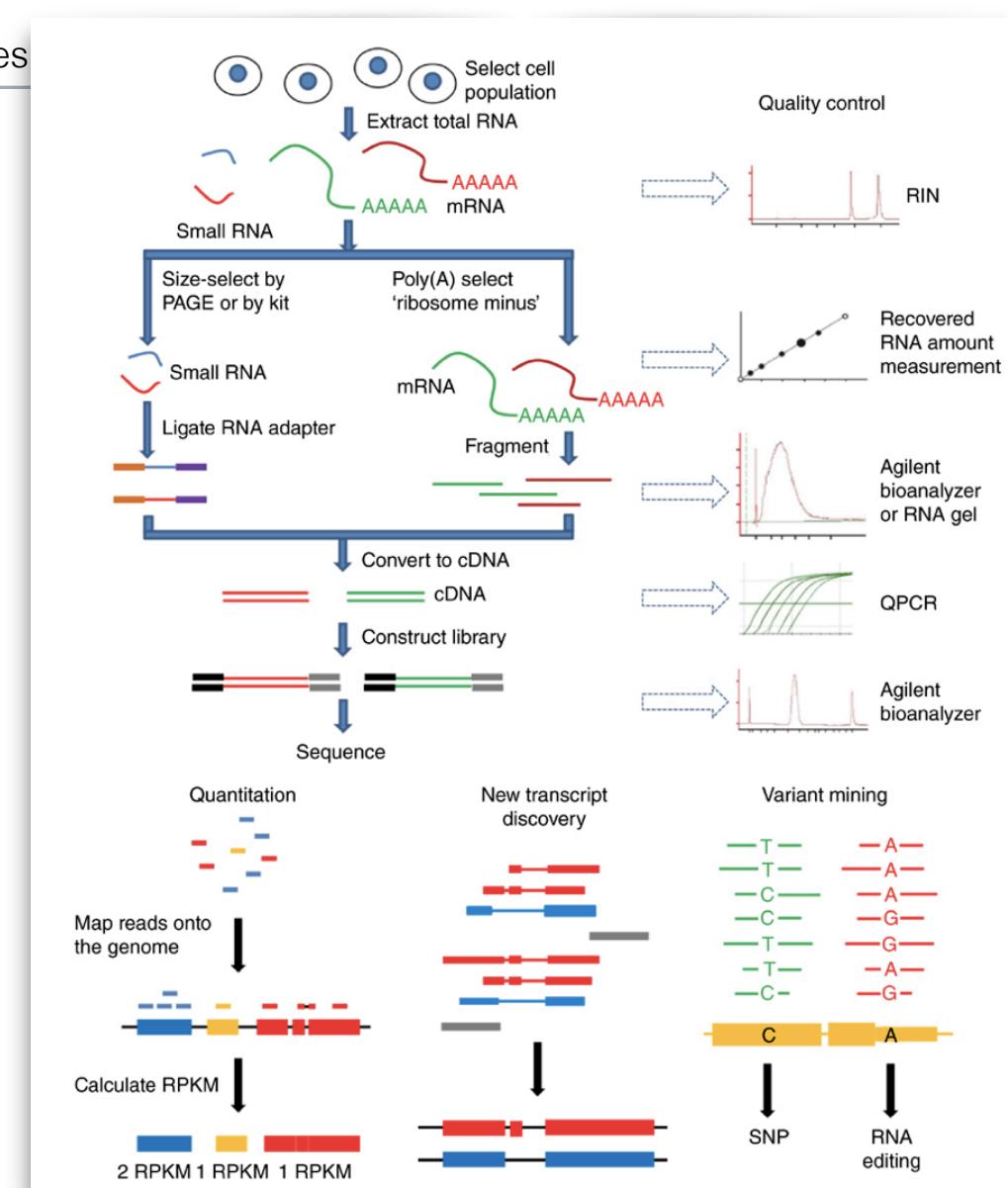


Abundance by Fluorescence Intensity (DNA microarray)



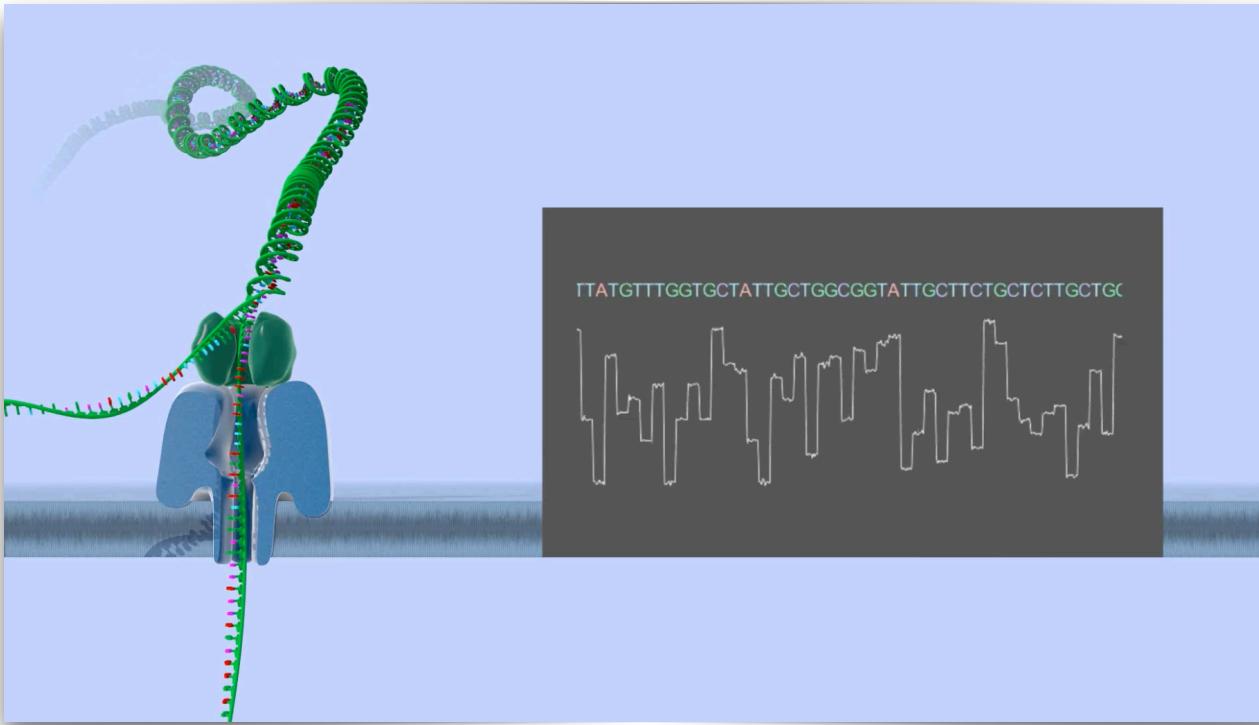
http://en.wikipedia.org/wiki/DNA_microarray

Abundance by Counting (RNA-seq)



Zeng & Mortazavi, Nature Immunology, 2012

ONT (Oxford Nanopore)



Secure | https://store.nanoporetech.com/cdna-and-direct-rna/

Nanoporetech | Metrichor | Community | Events | Store

Store

DEVICES KITS FLOW CELLS BUNDLES TRAINING & SERVICES

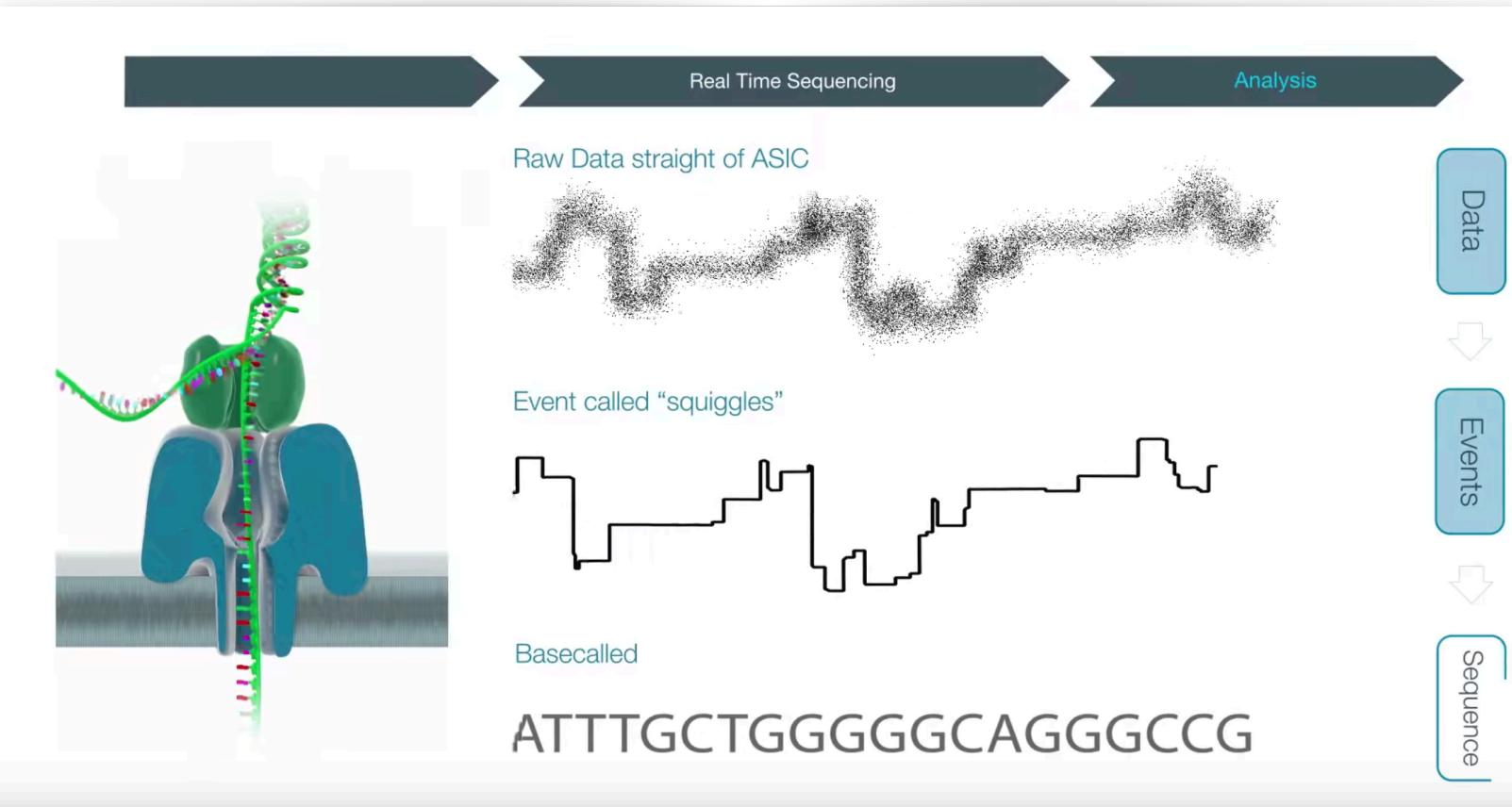
Direct RNA
Sequence RNA molecules directly and preserve base modifications
Up to 1 million reads

PCR cDNA
Optimised for throughput
Up to 10 million reads

PCR-free cDNA
No PCR bias
Up to 5 million reads

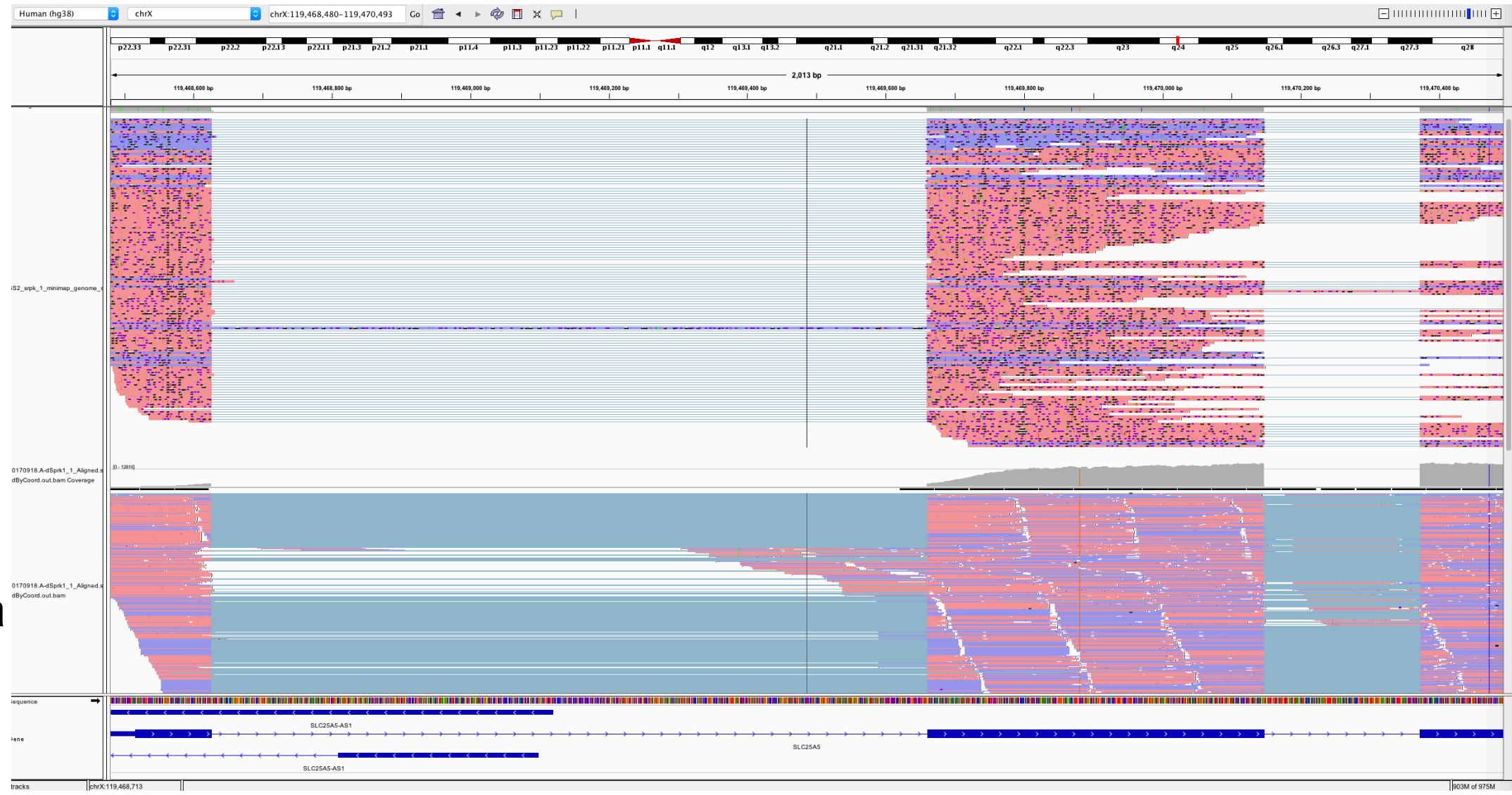
—> attachment of processive enzyme, leads RNA/DNA fragment to pore, combination of nucleotides going through pore creates a “characteristic disruption of the electrical current” —> order of signals can be used to determine the sequence of bases on that single strand.

ONT (Oxford Nanopore)



Quick look at reads in a browser

ONT



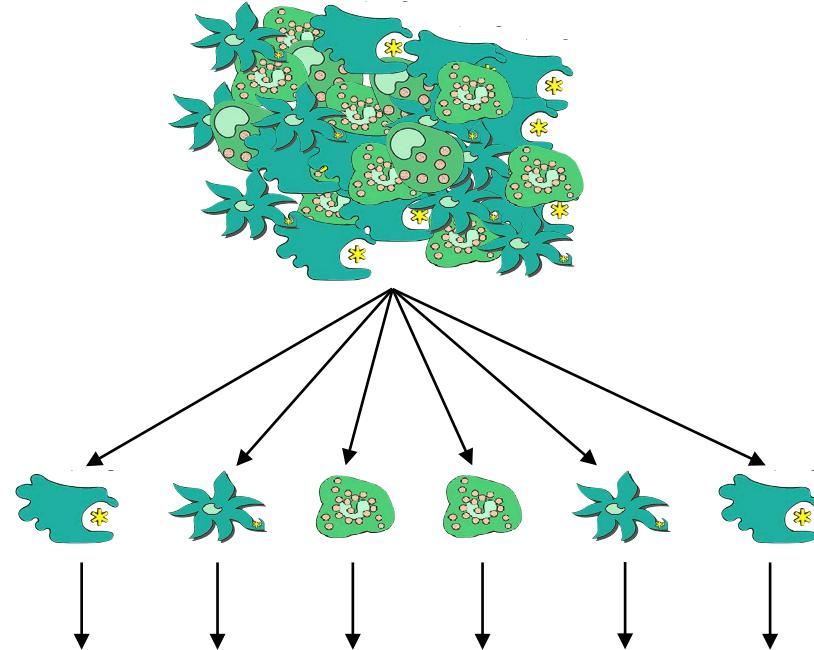
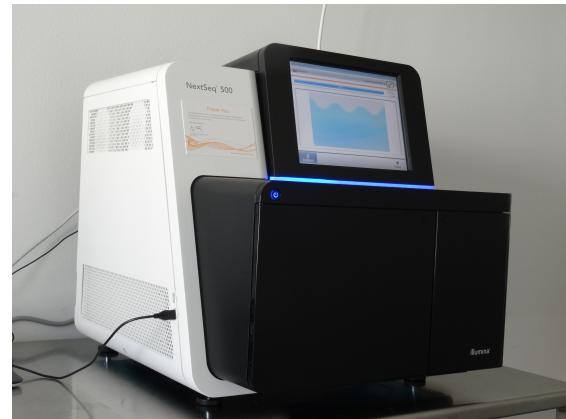
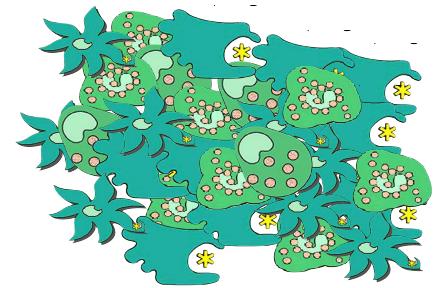
Illumina

Bulk vs single-cell RNA-sequencing

Cell sorting, tissue dissociation

RNA extraction,
preparation of cDNA,
cell barcoding, UMLs
(scRNA-seq only)

sequencing





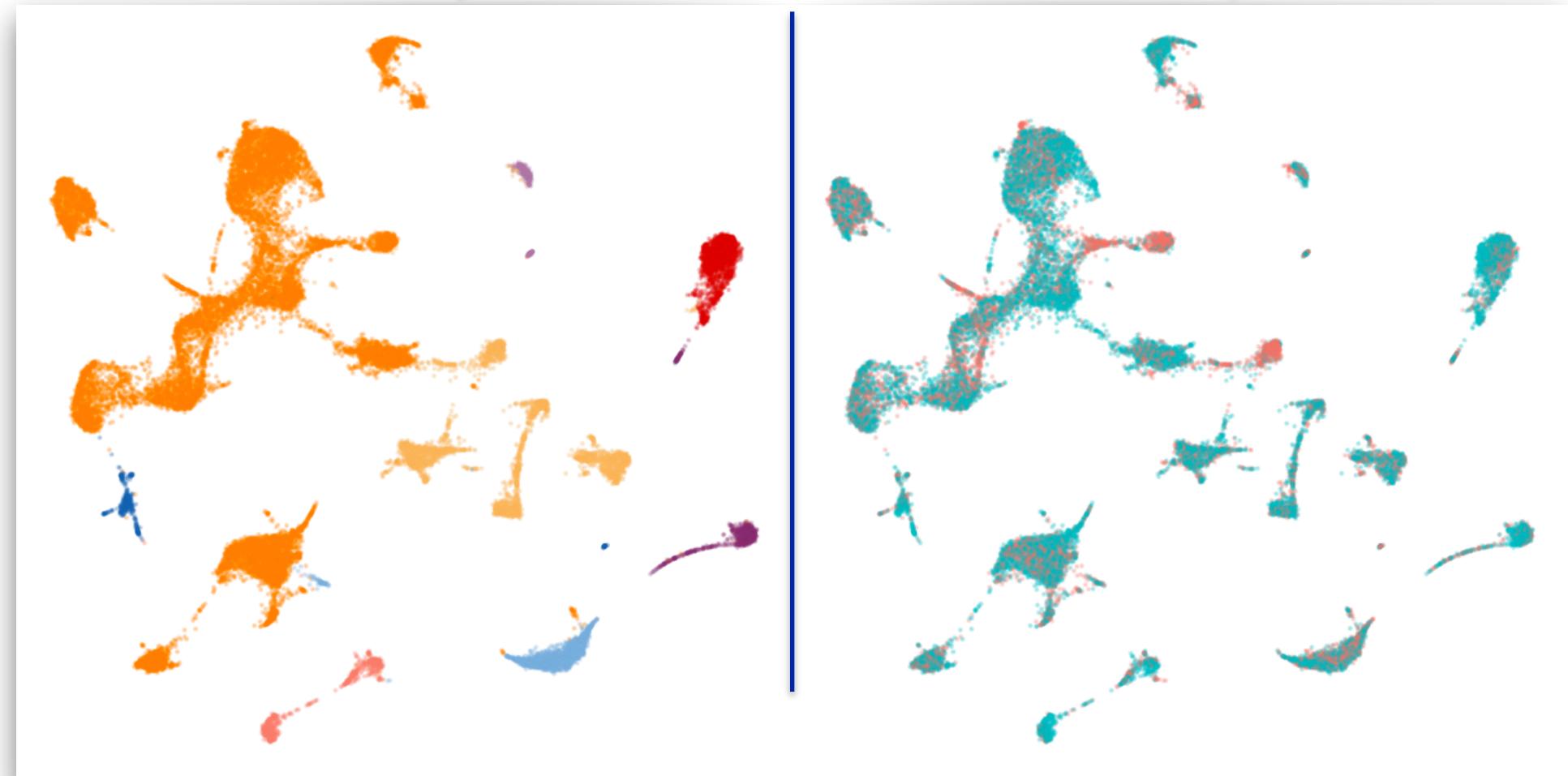
Motivation: Single-cell RNA-seq: finding cell subpopulation-specific changes in state

frontal cortex

single nuclei RNA-seq
(10x)

Data from:
4 mice vehicle treated
4 mice LPS treated

Each dot is one cell



Flow cytometry

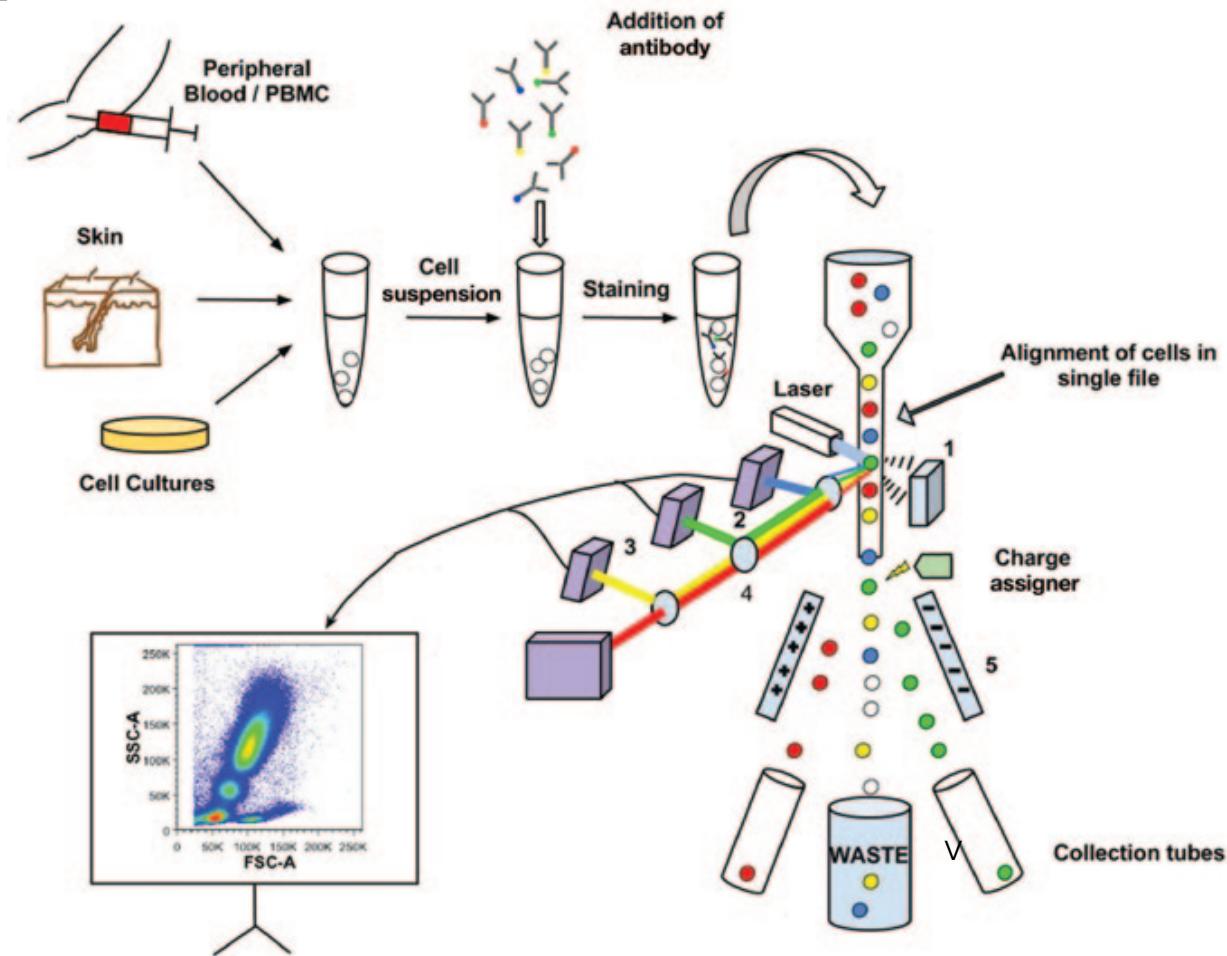
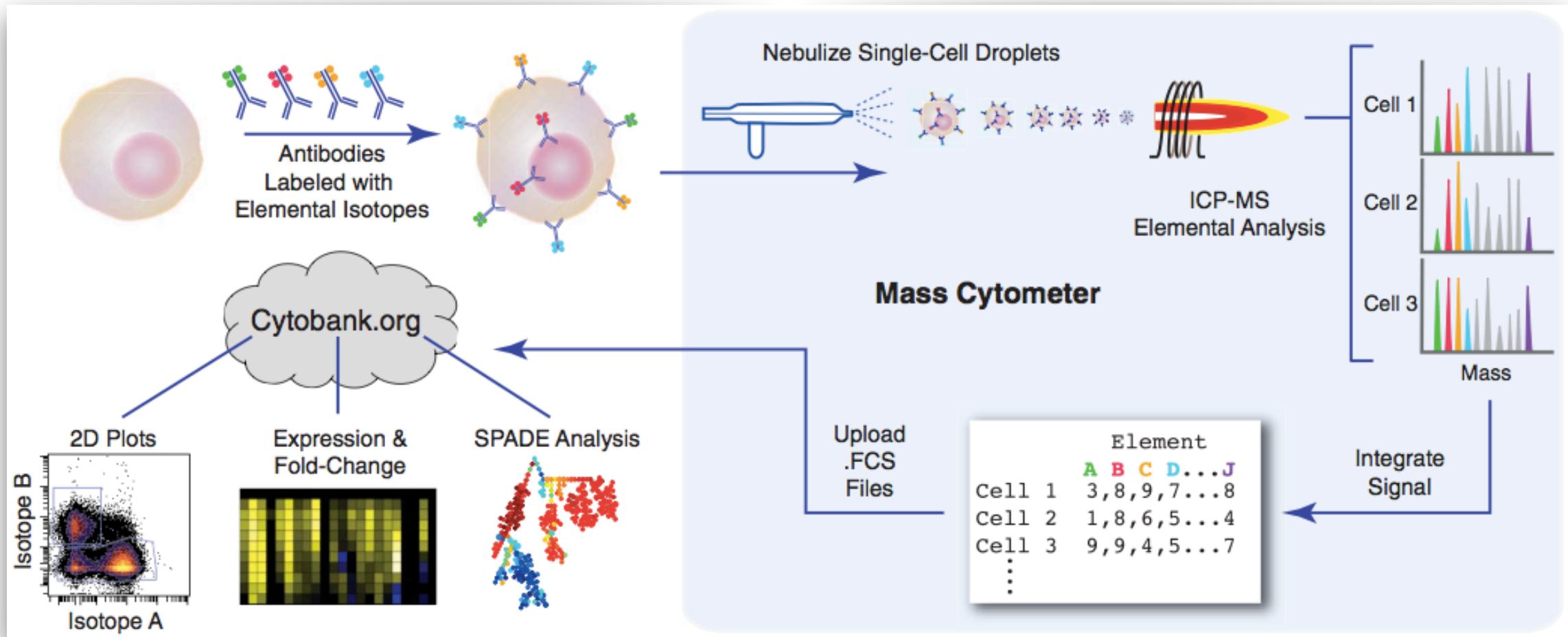


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.



Mass cytometry

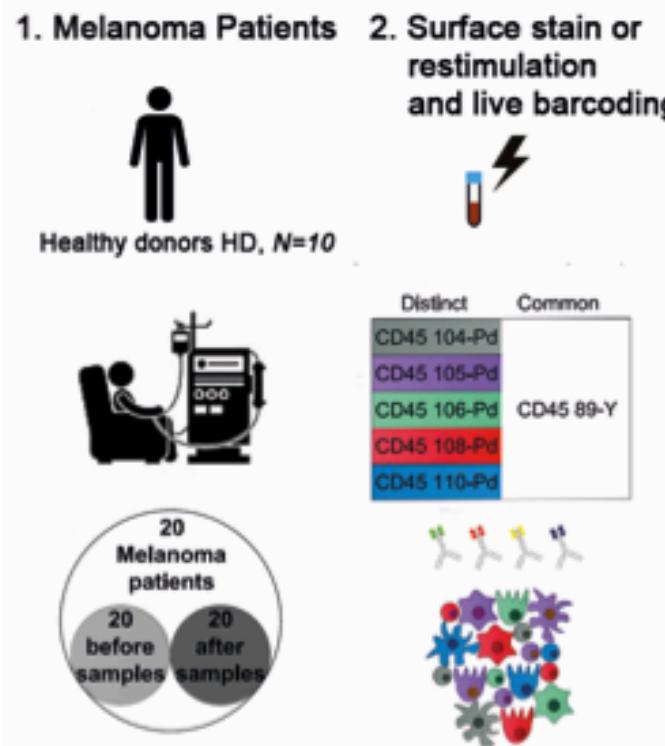


Bendall et al. (2011), Fig. 1A

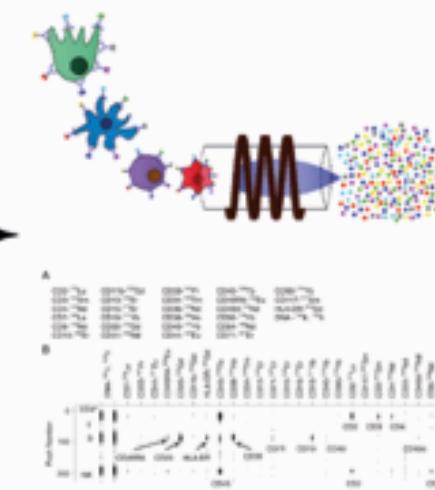


Finding molecular biomarkers associated with drug response

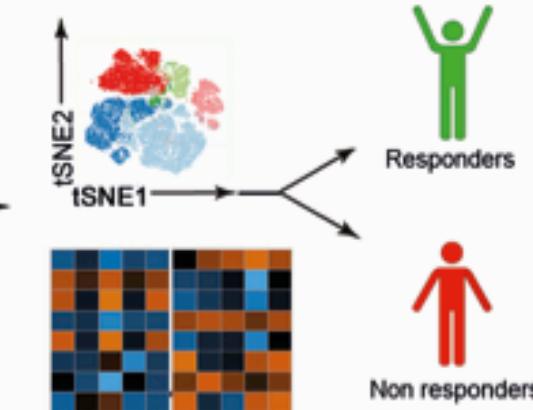
A Workflow



3. Single cell mass cytometry



4. Algorithm guided analysis



5. Biomarker discovery



High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy

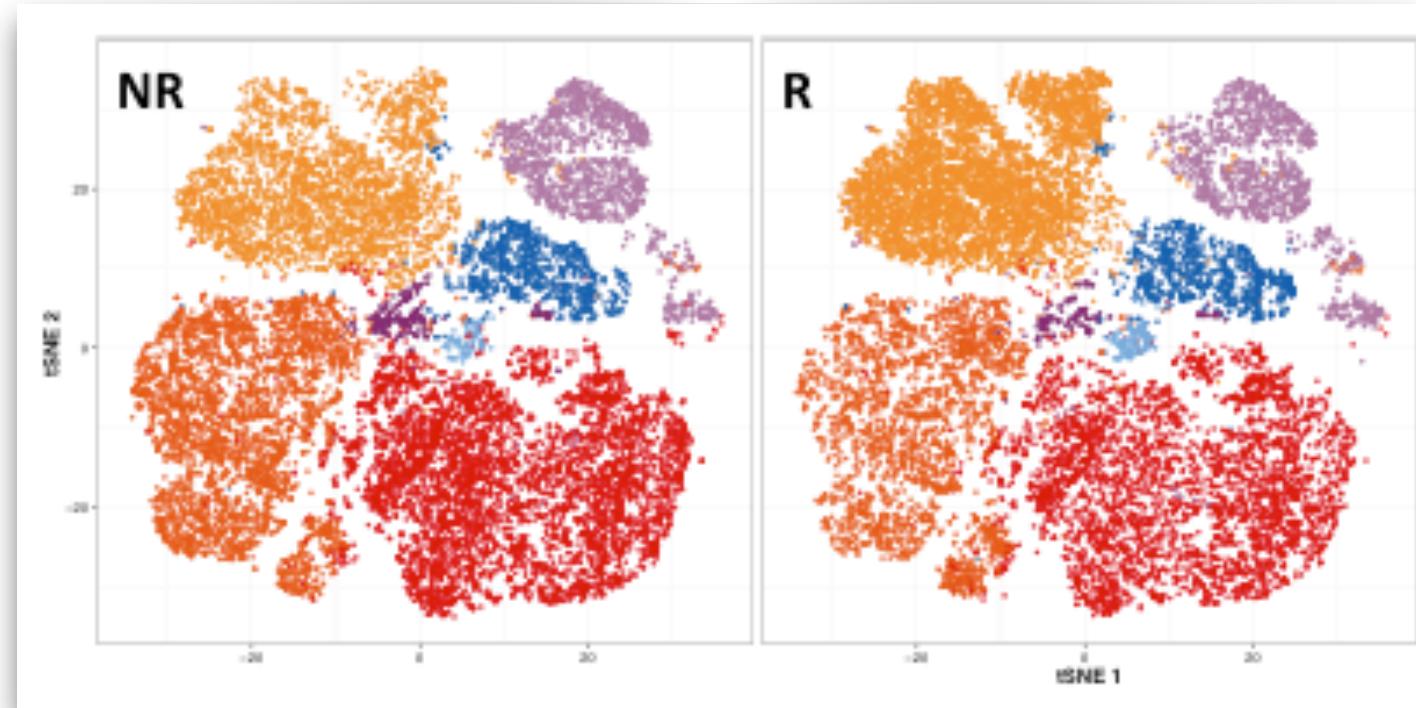
Carsten Krieg^{1,6} , Malgorzata Nowicka^{2,3}, Silvia Guglietta⁴, Sabrina Schindler⁵, Felix J Hartmann¹ , Lukas M Weber^{2,3} , Reinhard Dummer⁵, Mark D Robinson^{2,3} , Mitchell P Levesque^{5,7}  & Burkhard Becher^{1,7} 



Differential abundance of cell populations

tSNE projection
(each dot = cell,
cells from multiple
patients)

NR: non-responders
R: responders



Under the hood: Generalized linear mixed model to assess the change in relative abundance of subpopulations.



Some of the statistical fundamentals that underpin much of our research .. and our discoveries (.. but also underpin analyses that you may do in the future)

- central limit theorem
- false positives / false negatives (error control)
- statistical tests, multiple testing, P-values
- sharing information (limma)
- clustering
- exploratory data analysis, e.g., dimensionality reduction

Central limit theorem

Central limit theorem

The short non-technical version: once you start taking sums (averages), **sampling distributions** of the mean converge to the Gaussian (normal) bell shaped curve as the sample size increases.

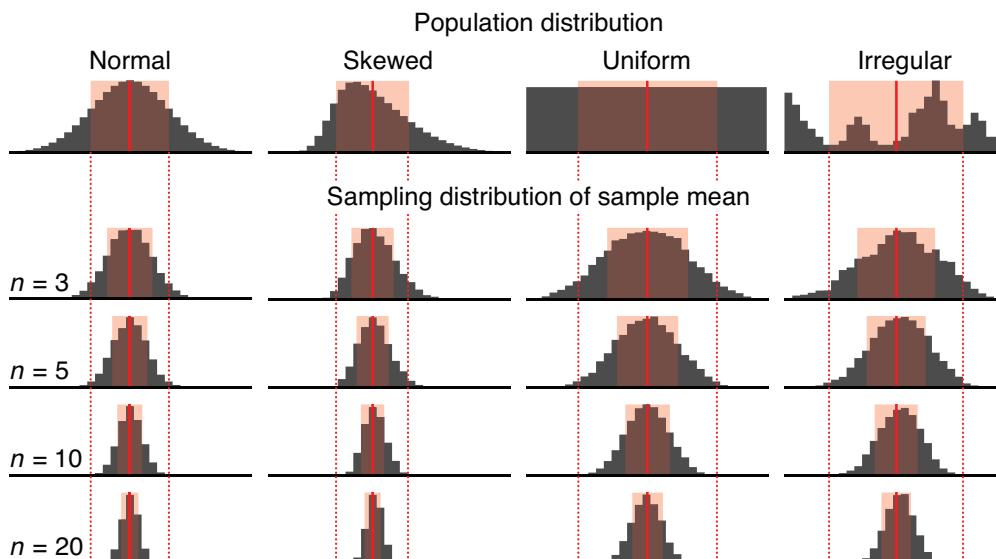


Figure 3 | The distribution of sample means from most distributions will be approximately normally distributed. Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in **Figure 1**.

If time, demonstrate this in R.

false positives, false negatives,
multiple testing, P-values



Hypothesis testing

- Method of making a decision
- Is this result "statistically significant"? ("Is my finding likely to occur by chance?")
- (Controversial) 
- Statistical significance != Biological significance

Operationally, it works (something) like:

- Define "null hypothesis" (usually some kind of baseline setting)
- Define alternative: non-null
- Calculate test statistics (e.g. where the sampling distribution under the null is known) and/or P-value
- If P-value < some (magic) cutoff, decide to reject the null hypothesis in favour of the alternative; otherwise, accept the null hypothesis

EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.

"Researchers should seek to analyse data in multiple ways to see whether different analyses converge on the same answer."



NHST (Null hypothesis statistical testing): Hypothetical example

Say we wanted to know whether ETHZ students are scoring better or worse in a particular course than UZH students. First, we take a random sample from each population.

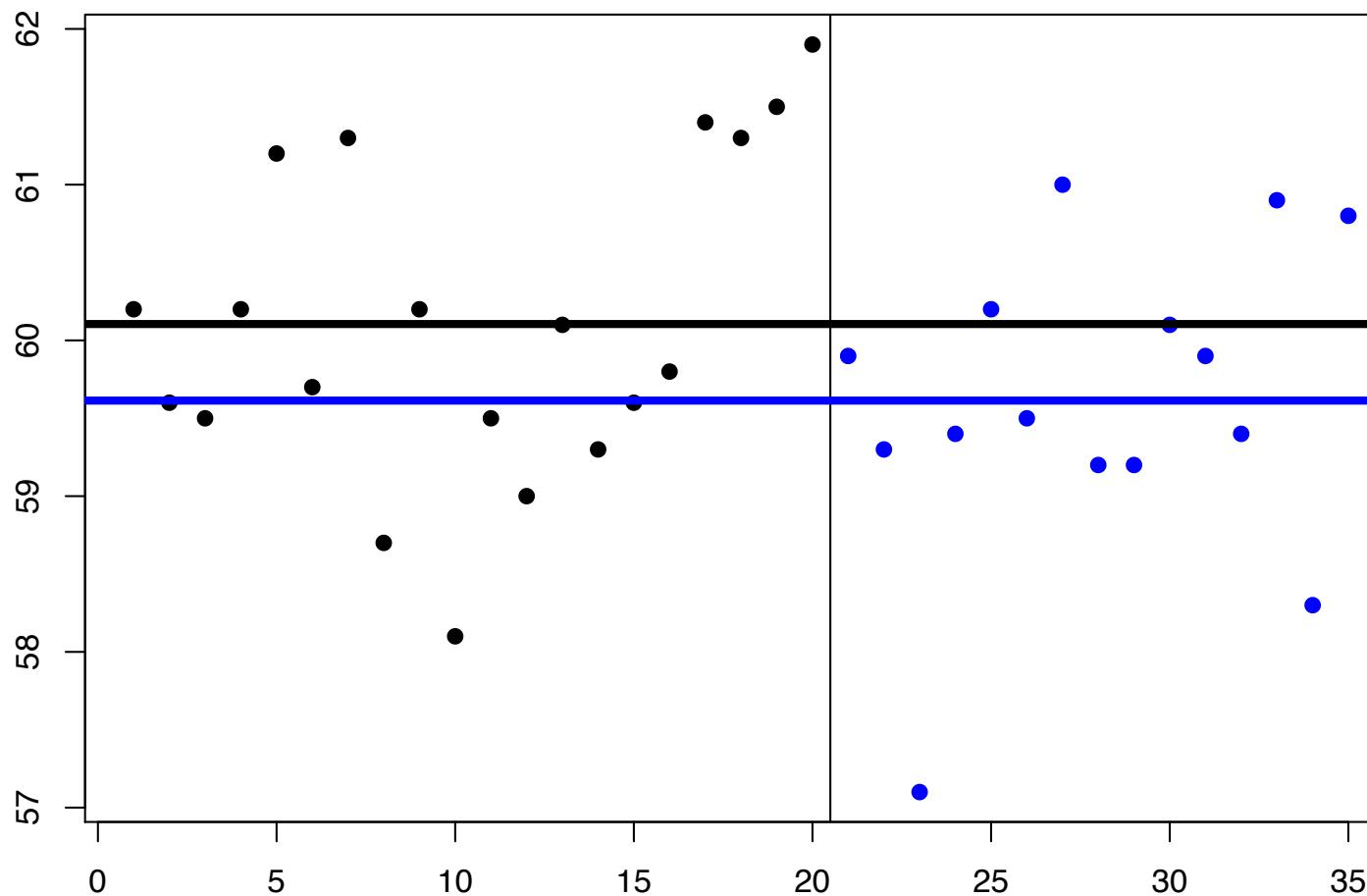
Null hypothesis: population mean of ETHZ scores = population mean of UZH scores

Alternative: means are different

Critical point: Assume that null hypothesis is true (i.e., means are equal), calculate a test statistic that we know the distribution of (under the null). Calculate the probability of observing something as or more extreme than our test statistic.

We'll use a t-statistic.

There are some variations of the t-test, but let us assume that the variances are equal



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$s_{X_1 X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}.$$



Where does the t-test come from?

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the “error of random sampling” the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabulated, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.



OK, but mathematically, where does the t-distribution come from?

$$Z = (\bar{X}_n - \mu) \frac{\sqrt{n}}{\sigma}$$

$$T \equiv \frac{Z}{\sqrt{V/\nu}} = (\bar{X}_n - \mu) \frac{\sqrt{n}}{S_n},$$

$$V = (n - 1) \frac{S_n^2}{\sigma^2}$$

Clever discovery by William Gosset (i.e. “Student”)

The variance parameter cancels out —> straightforward extension to the 2-sample problem.



False positives / false negatives

Most statistical testing
regimes set an error rate (5%)

Type I error = false positive
Type II error = false negative

Arthur Charpentier
@freakonometrics [Follow](#)

Statistical Errors

	$\hat{Y} = 0$ NEGATIVE	$\hat{Y} = 1$ POSITIVE
$Y = 0$ NOT PREGNANT	TRUE NEGATIVE You're not pregnant	FALSE POSITIVE You're pregnant TYPE 1 ERROR
$Y = 1$ PREGNANT	FALSE NEGATIVE You're not pregnant TYPE 2 ERROR	TRUE POSITIVE You're pregnant

limma (sharing information)



Differential expression, small sample inference

- Table of data (e.g., microarray gene expression data with replicates of each of condition A, condition B)
 - *rows* = features (e.g., genes), *columns* = experimental units (samples)
- Most common problem in statistical bioinformatics: want to infer whether there is a **change in the response** → a statistical test for each row of the table.

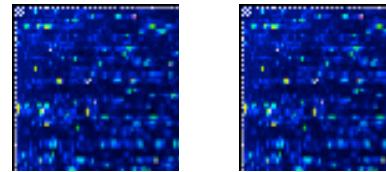
```
> head(y)
      group0     group0     group0     group1     group1     group1
gene1 -0.1874854  0.2584037 -0.05550717 -0.4617966 -0.3563024 -0.03271432
gene2 -3.5418798 -2.4540999  0.11750996 -4.3270442 -5.3462622 -5.54049106
gene3 -0.1226303  0.9354707 -1.10537767 -0.1037990  0.5221678 -1.72360854
gene4 -2.3394536 -0.3495697 -3.47742610 -3.2287093  6.1376670 -2.23871974
gene5 -3.7978820  1.4545702 -7.14796503 -4.0500796  4.7235714 10.00033769
gene6  1.4627078 -0.3096070 -0.26230124 -0.7903434  0.8398769 -0.96822312
```

[1] <http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>

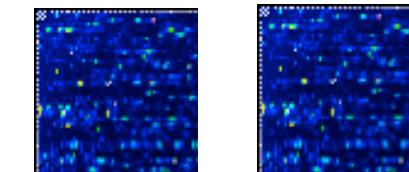


A very common experiment: microarray or RNA sequencing

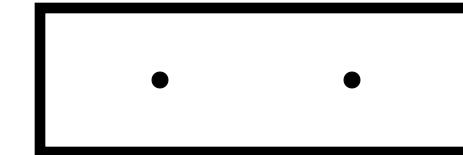
Mutant x 2



WT x 2



Gene X



Which genes are differentially expressed?

$n_1 = n_2 = 2$ Affymetrix arrays

~30,000 probe-sets



In genomics, there is often a **multiple testing** problem

- You often make multiple tests (e.g., for every gene). Say, you set your cutoff such that you had a 5% false positive rate.
- In doing 20,000 tests (for 20,000 genes), ~1000 would be rejected just by chance.
- There are various ways to "correct" for multiple testing. Two popular ones include:
 1. False discovery rate (weak)
 2. Bonferroni correction (strong)



Classical 2-sample t-tests

$$t_g = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{s_g c}$$

give very high false discovery rates

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

Residual df = 2



t-tests with **common** variance (pooled over all genes measured)

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

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: moderate between

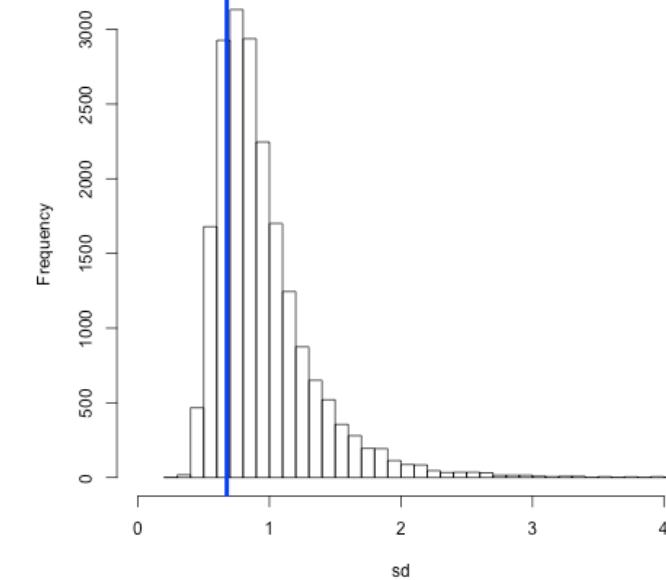
Shrink standard deviations towards common value

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

d = degrees of freedom

Moderated t-statistics

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





Exact distribution for moderated t

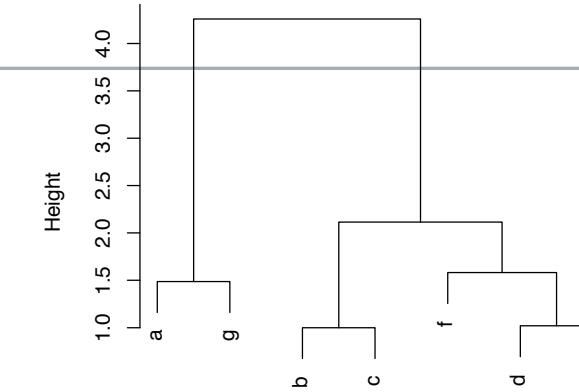
An unexpected piece of mathematics shows that, under the null hypothesis,

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

The degrees of freedom add.

In effect, the moderated variance adds d_0 extra samples to the analysis, thus increasing the statistical power.

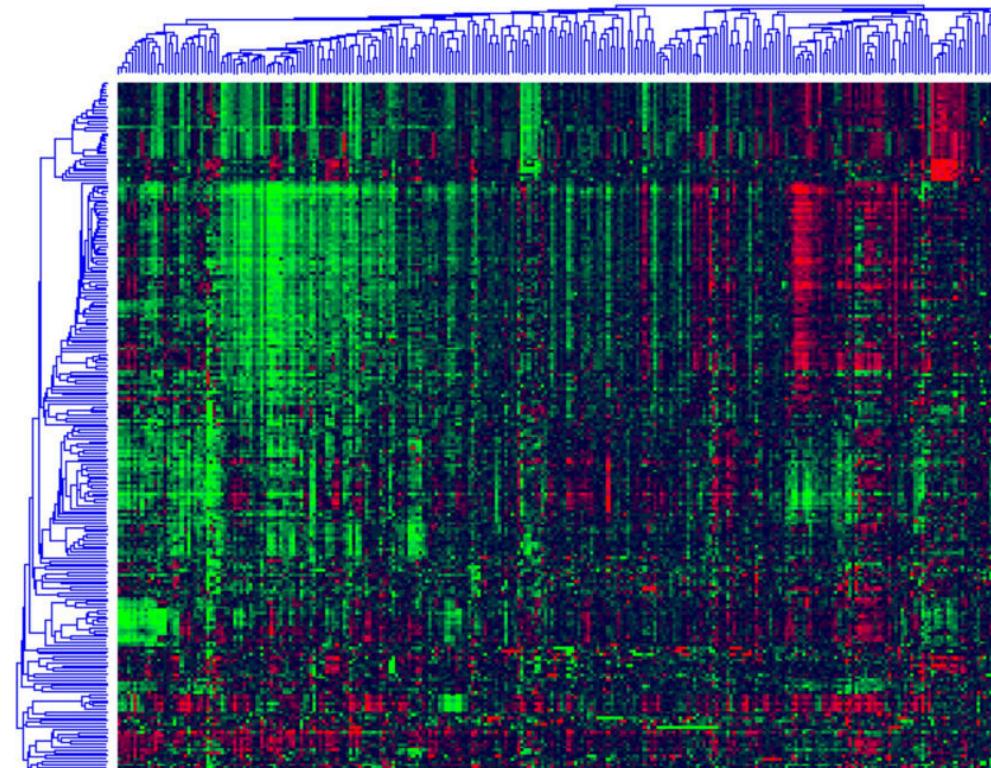
clustering (hierarchical)



Divisive (all features start as 1 cluster, then subsequently split) versus Agglomerative (every feature is it's own cluster, then subsequently merged)

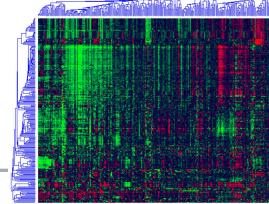
Metric: to define how similar any two vectors are.

Linkage: determines how clusters are merged into a tree





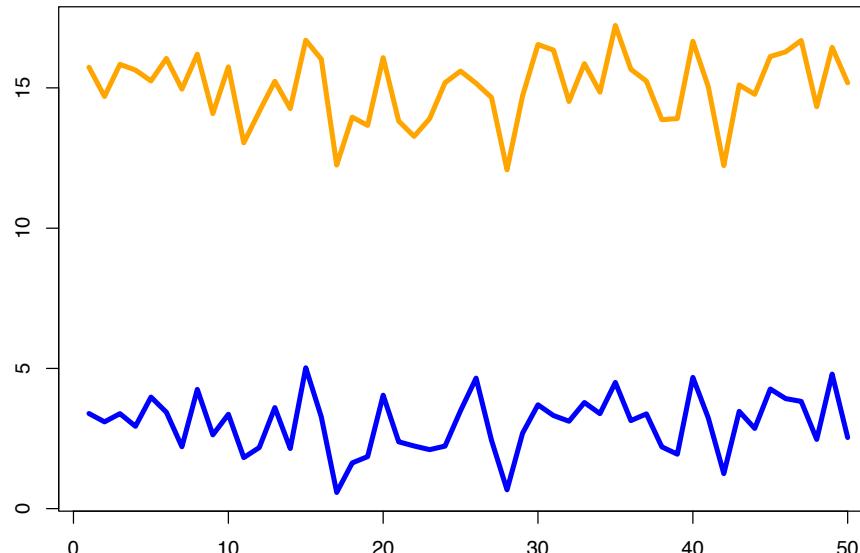
$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$



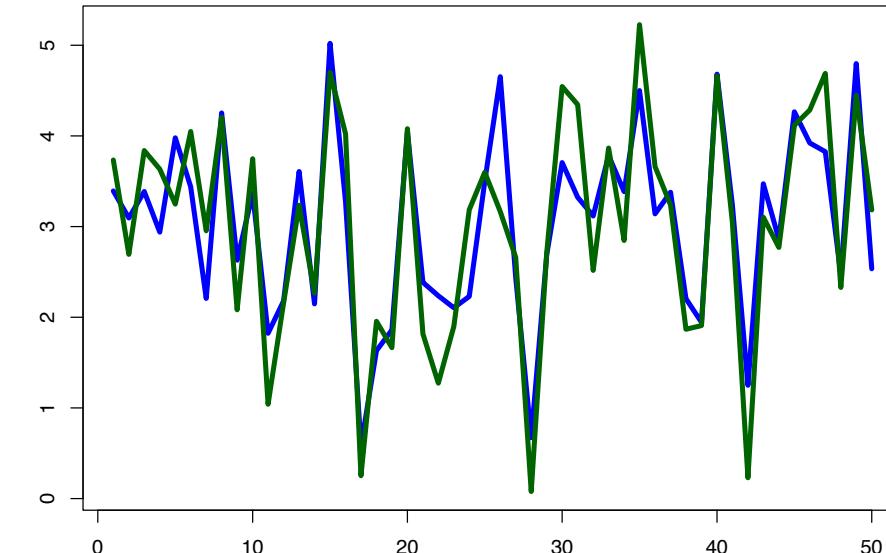
Are these “vectors” similar ?

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

It depends how you define similar.



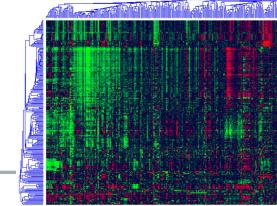
Euclidean distance: 84.84



3.92



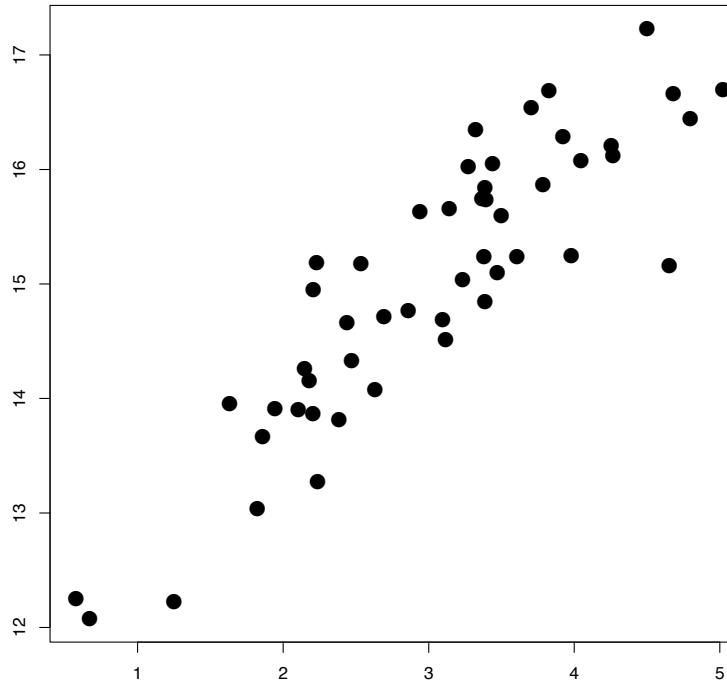
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$



Are these “vectors” similar ?

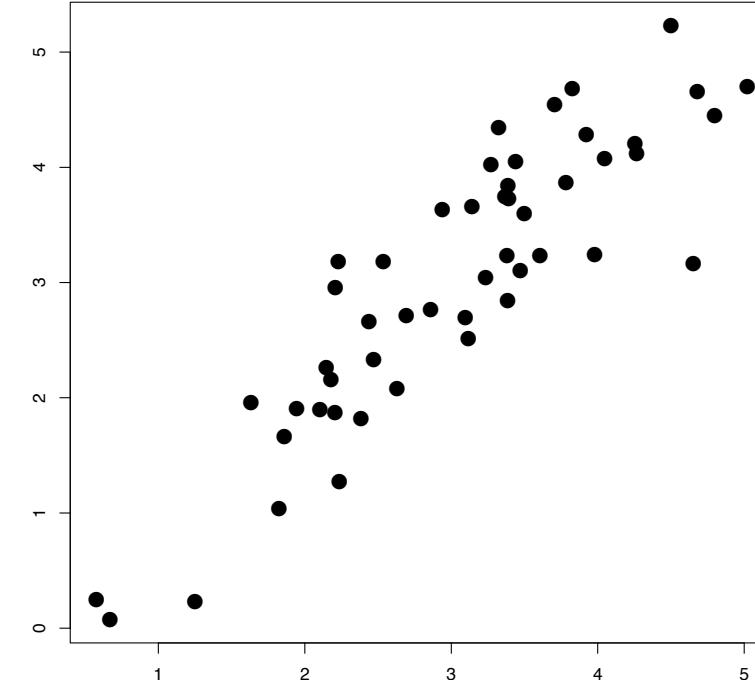
```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

It depends how you define similar.

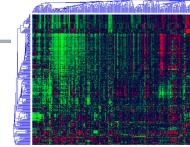


Correlation:

0.89



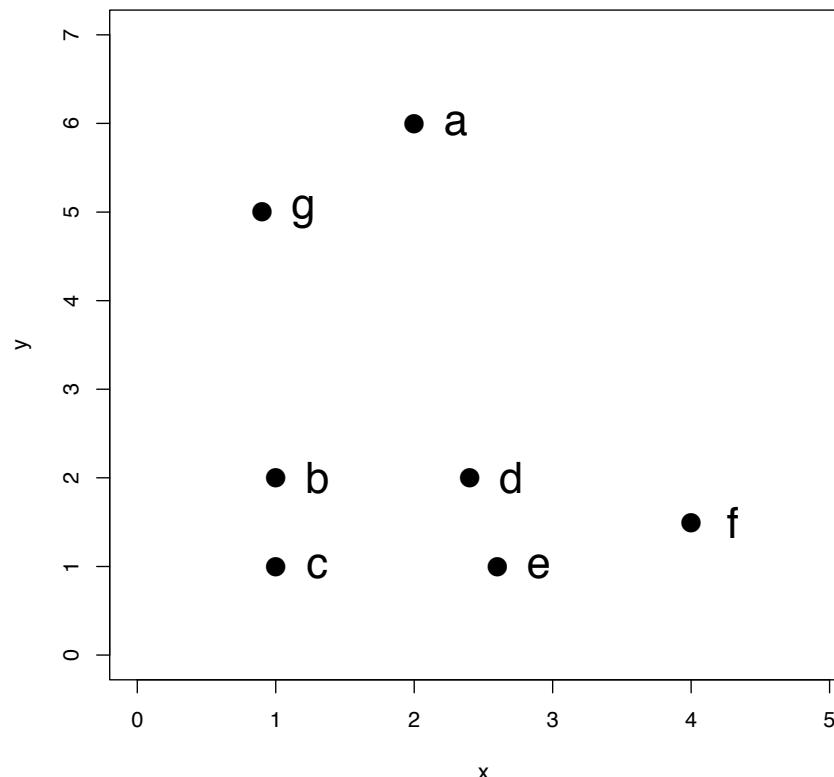
0.89



Hierarchical (Agglomerative) Clustering

Start with distances.

Linkage: determines how clusters are merged into a tree.

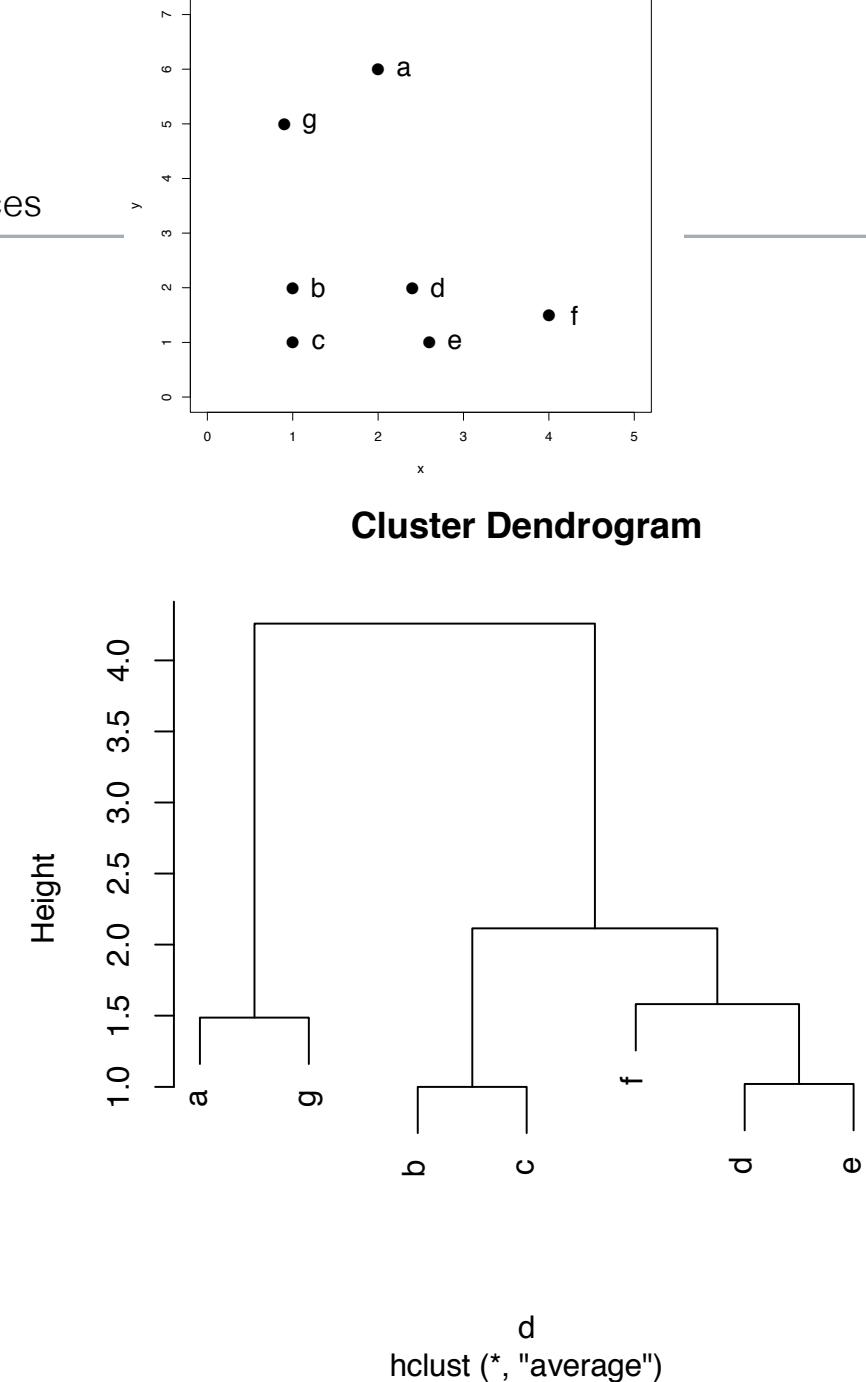
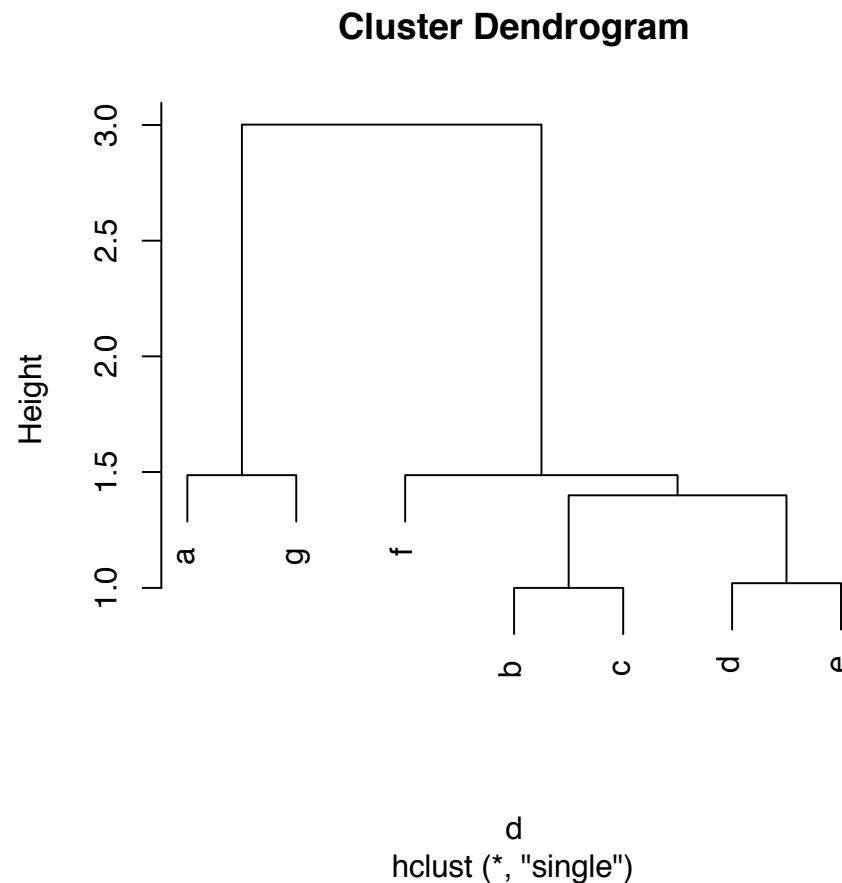


From eyeballing, here is a likely set of merges:

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



Different linkages



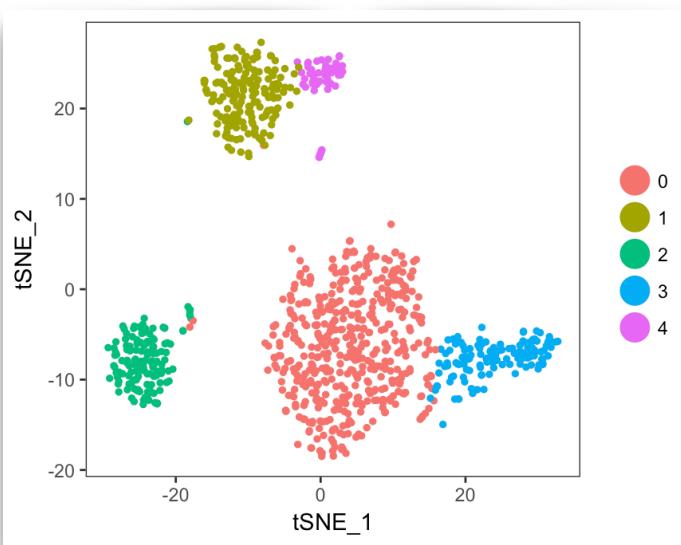
dimension reduction
(exploratory data analysis)

Dimension reduction: general introduction

- Many types of data come as a matrix of N samples (e.g., cells, patients) x G features (e.g., genes, proteins)
- Each sample is a point in G-dimensional space
- Goal: represent the data in 2-3 dimensions, but preserve **structure** as best as possible (i.e., points that are **close** in G dimensions should be close in 2 or 3 dimensions)

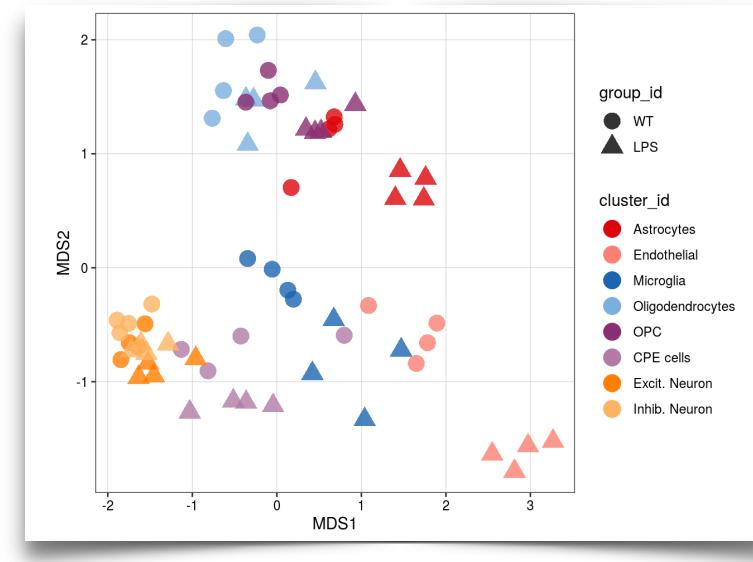
Dimension reduction is versatile

$K \text{ features} \times N \text{ cells} \rightarrow$
 $2 \text{ dimensions} \times N \text{ cells}$



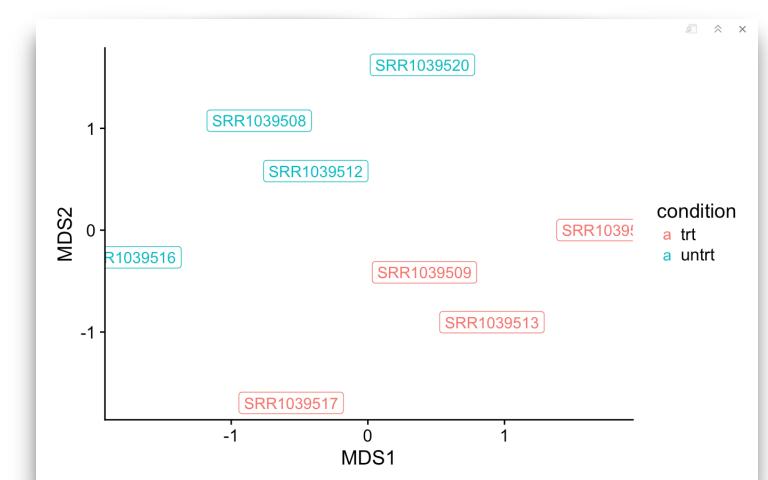
Each point =
single cell
(10x PBMC)

$N \text{ cells} \times K \text{ features} \rightarrow N \text{ cell}$
subpopulations $\times 2 \text{ dimensions}$



Each point =
subpopulation from a
single sample (LPS mouse cortex)

$P \text{ samples} \times K \text{ features} \rightarrow$
 $P \text{ samples} \times 2 \text{ dimensions}$



Each point =
sample (airway)

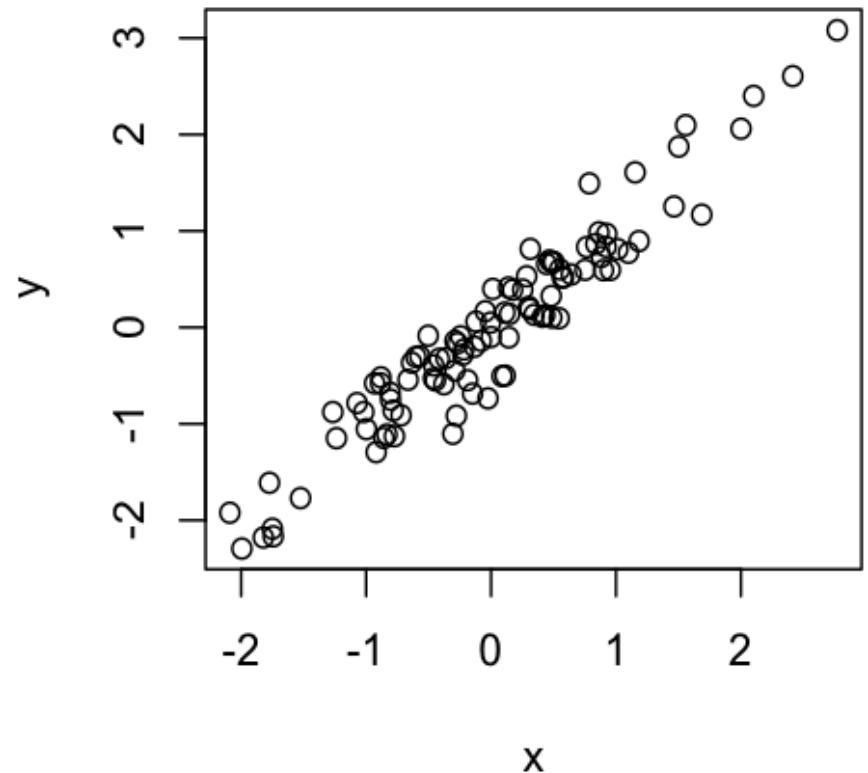
Introduction to dimension reduction: PCA (principal components analysis)

- Form successive *linear* combinations of the features that are: orthogonal, ordered by variance

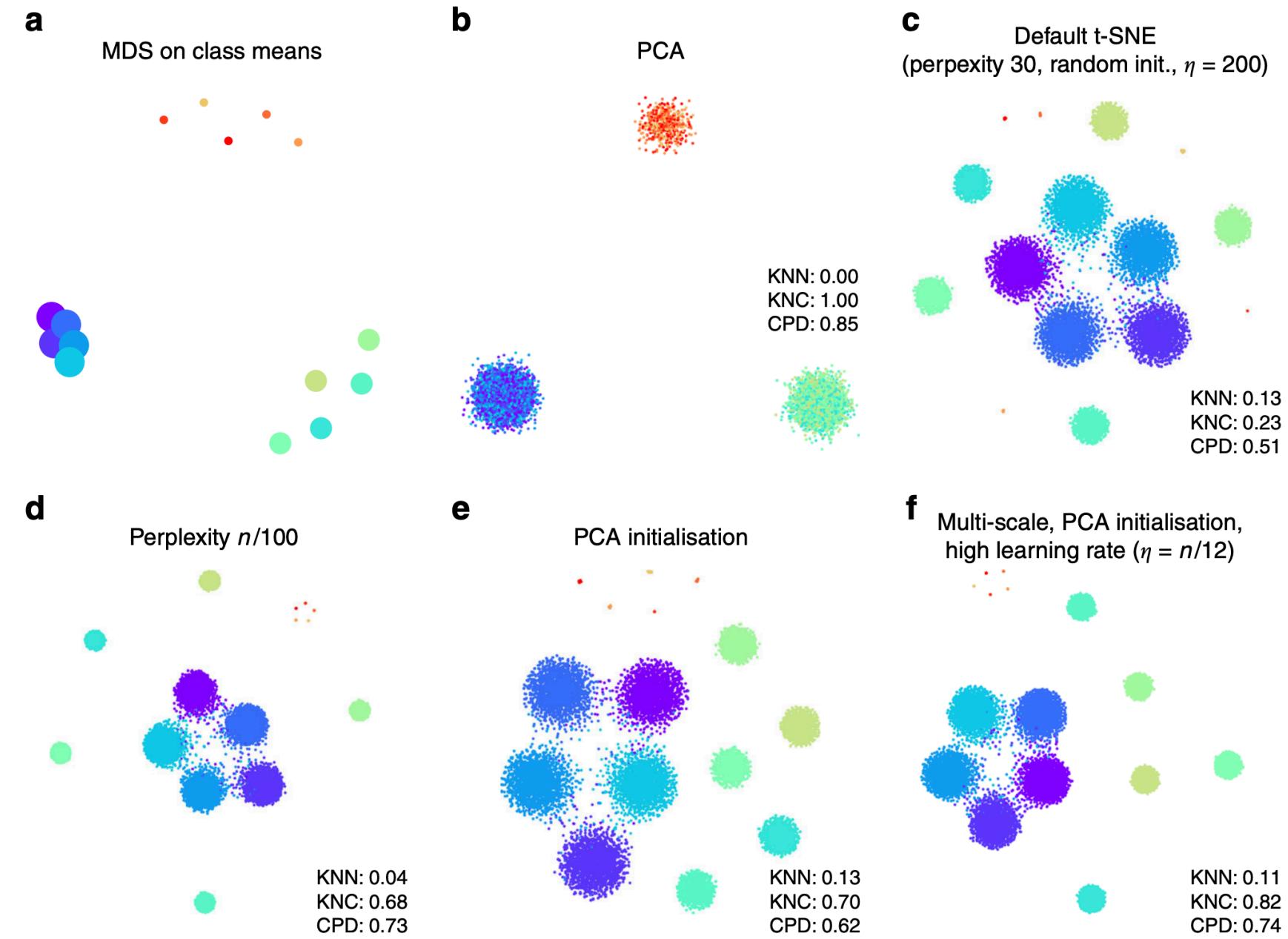
$$Y = XA$$

$$Y_{rk} = a_{1k}x_{r1} + a_{2k}x_{r2} + \cdots + a_{pk}x_{rp}$$

- A is the loadings matrix
- Typically, first 2-3 columns ('principal components') of Y are retained for visualisation; often top P PCs are retained for other analyses (e.g., clustering)



Many variations
(linear/non-
linear), many
notions of
distance, many
ways to
“compress”

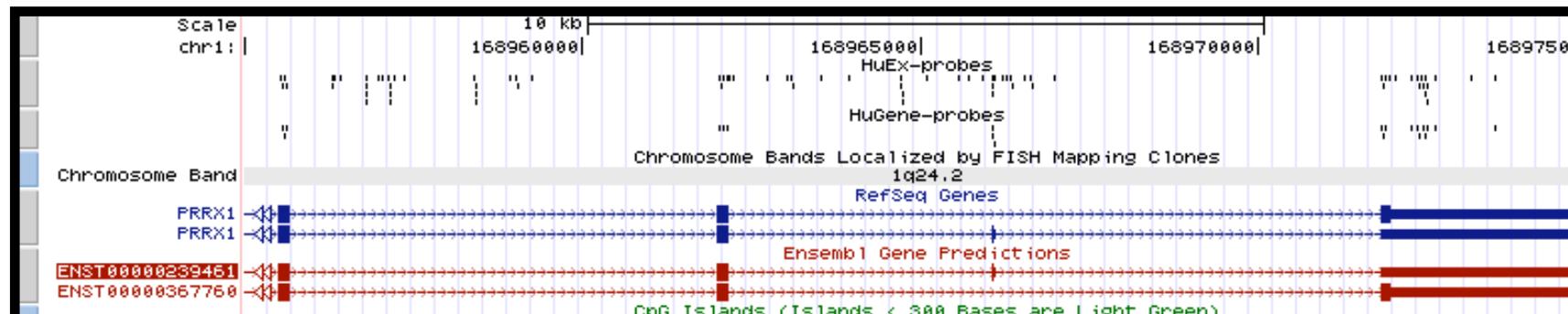




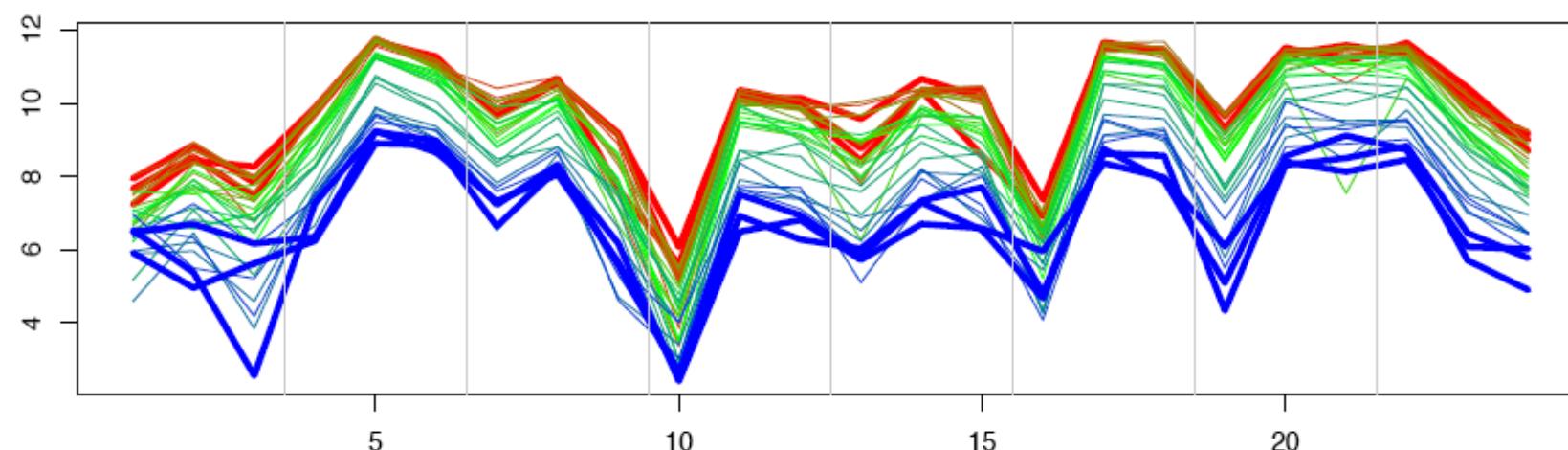
Another data example .. a regression model to separate interesting signal (gene expression) from technical effects (probes)

The nature of Affymetrix Probe Level Data

Statistical Bioinformatics // Department of Molecular Life Sciences

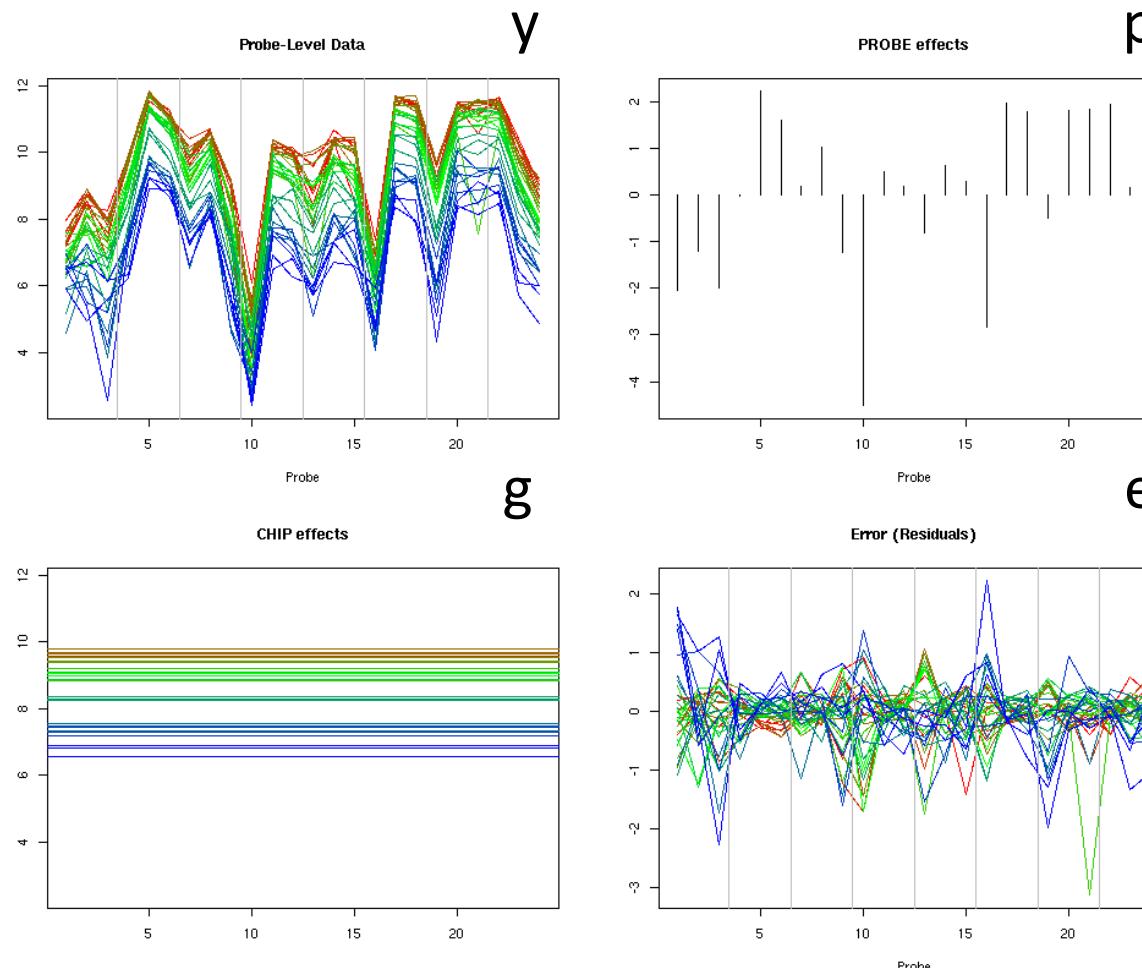


HuGene data [red-heart,blue-brain,mixtures] 10 ENSG00000116132



- Data for one gene that is differentially expressed between heart (red is 100% heart) and brain (blue is 100% brain).
- 11 mixtures x 3 replicates = 33 samples (33 lines)
- Note the parallelism: probes have different **affinities**

Linear model decomposes the probe-level data into **PROBE** effects and **CHIP** effects



Linear model:

$$y_{ik} = g_i + p_k + e_{ik}$$

Robust Multichip Analysis (RMA)
uses this model.
Irizarry et al. 2003,
Biostatistics

Parameters are estimated **robustly**, meaning a small number of outliers have minimal effect