BORDEAUX POPULATION HEALTH | Centre de Recherche - U1219

ISPED SCHOOL OF PUBLIC HEALTH / université de BORDEAUX

Digital Public Health Graduate Program

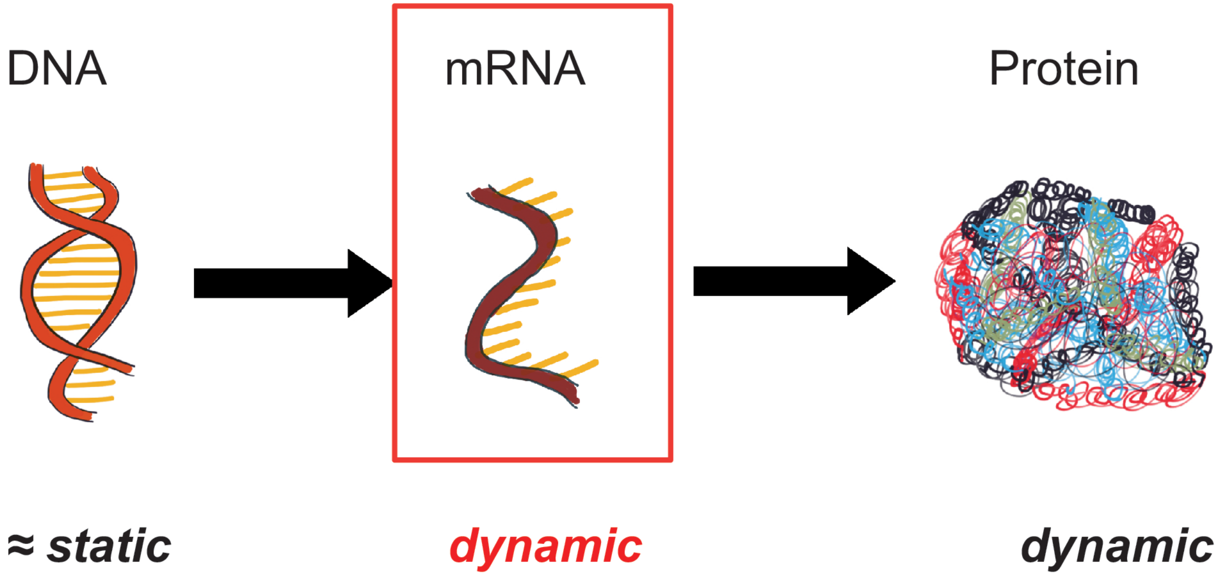
Transcriptomics Data:
Generation, Management & Analysis

T.U. OMICS

Boris Hejblum, PhD
Assistant Professor in Biostatistics
Université de Bordeaux – ISPED
SISTM team, BPH Inserm U1219 / Inria BSO

INTRODUCTION

Central dogma of molecular biology



DNA

mRNA

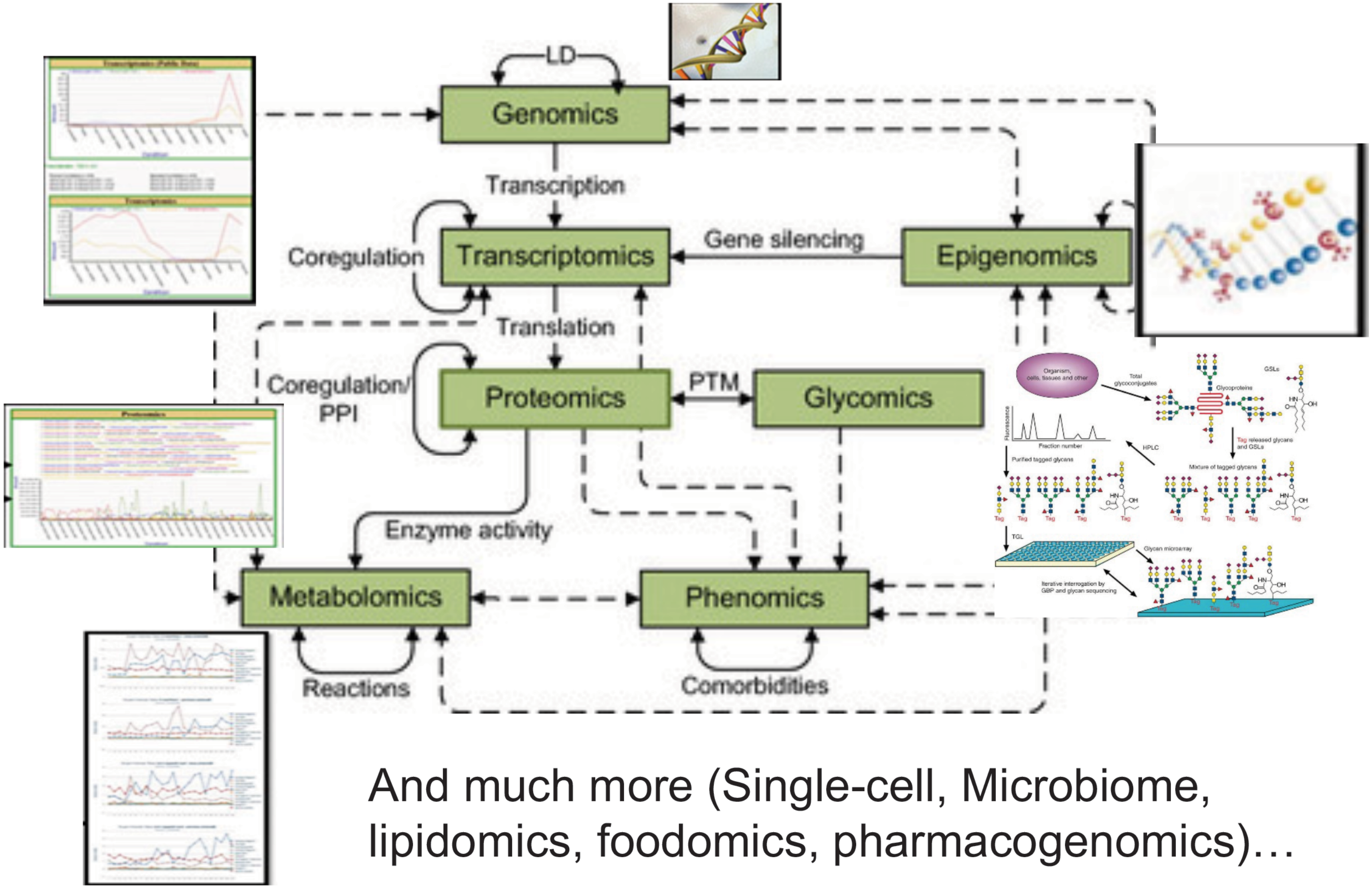
Protein

≈ ***static***

dynamic

dynamic

Omics Data



And much more (Single-cell, Microbiome, lipidomics, foodomics, pharmacogenomics)…

Adapted from Zierer, Aging Cell 2015

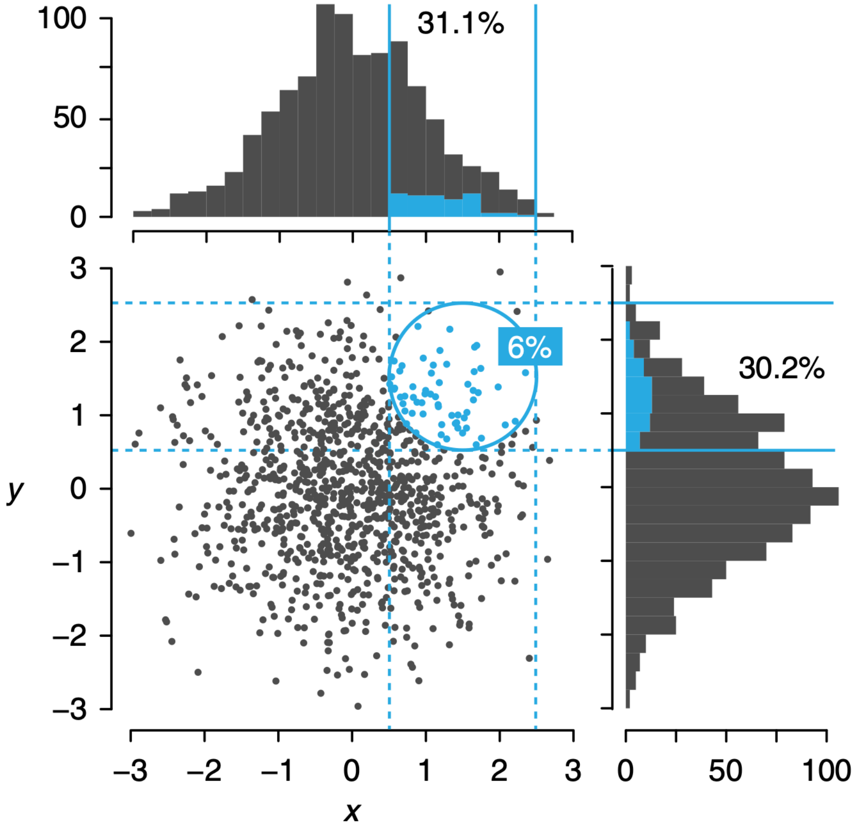
Many (gen)omics data

- **Genomics**
  - Single Nucleotide Polymorphisms
  - eQTL
  - …
- **Transcriptomics**
  - qPCR
  - microarrays
  - RNAseq
- **Proteomics**
- **Metabolomics**
  - Targeted
  - Untargeted
- **Microbiomics**
- **Cytomics**
  - Flow cytometry
  - Mass cytometry
- **…**

Measuring gene expression

- PCR
- Microarrays
- (bulk) RNA-seq
- single-cell RNA-seq

- ➔ A blessing
  - › *Lots of information*
- ➔ A curse
  - › *Signal “drowned” in dimension*



Altman & Krzywinski, Nat. Meth., 2018

HIERARCHICAL CLUSTERING

BORDEAUX POPULATION HEALTH | Centre de Recherche - U1219

ISPED SCHOOL OF PUBLIC HEALTH

université de BORDEAUX

Digital Public Health Graduate Program

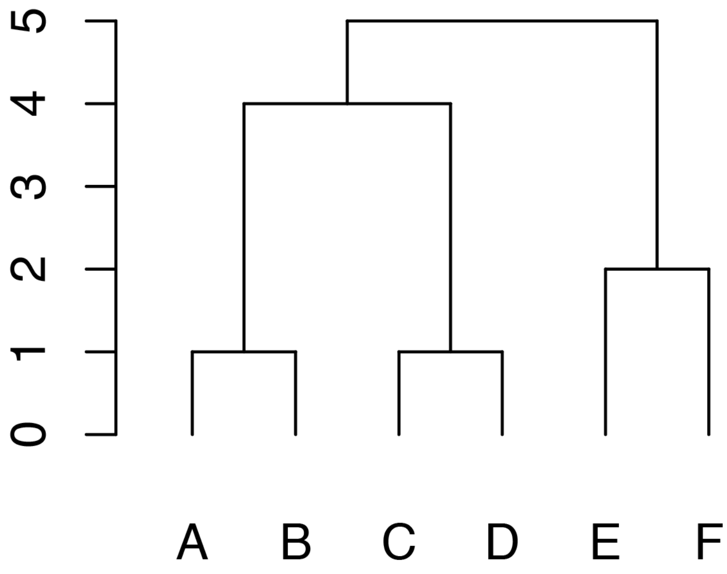
Clustering

Clustering:

Synthetic representation of a collection of objects (samples, variables, …) **highlighting resemblance**

→ *hierarchical clustering*

Hierarchical tree



NB: lateral proximities are not interpretable (several ways to display the same tree)

Formalizing “resemblance”

➔ **Aim**: build a hierarchical tree which branches are gather individuals, in a group defined by a set of properties (on the variables)

1. **Each individual** within the **group** possesses a **large fraction** of those **properties**
2. **Each properties** is possessed by a **large fraction** of the **individuals in the group**

Usually rely on a ***distance*** *between 2 individuals* i *and* i’ :

1. **Euclidean** distance

$$d(i, i')^2 = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

2. **Manhattan** (or city-block) distance

$$d(i, i') = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$$

Distance choice can influence the results

Data:

$$\begin{array}{cccc} & V1 & V2 & V3 \\ A & 1 & 1 & 3 \\ B & 1 & 1 & 1 \\ C & 2 & 2 & 2 \end{array}$$

Euclidean distance:

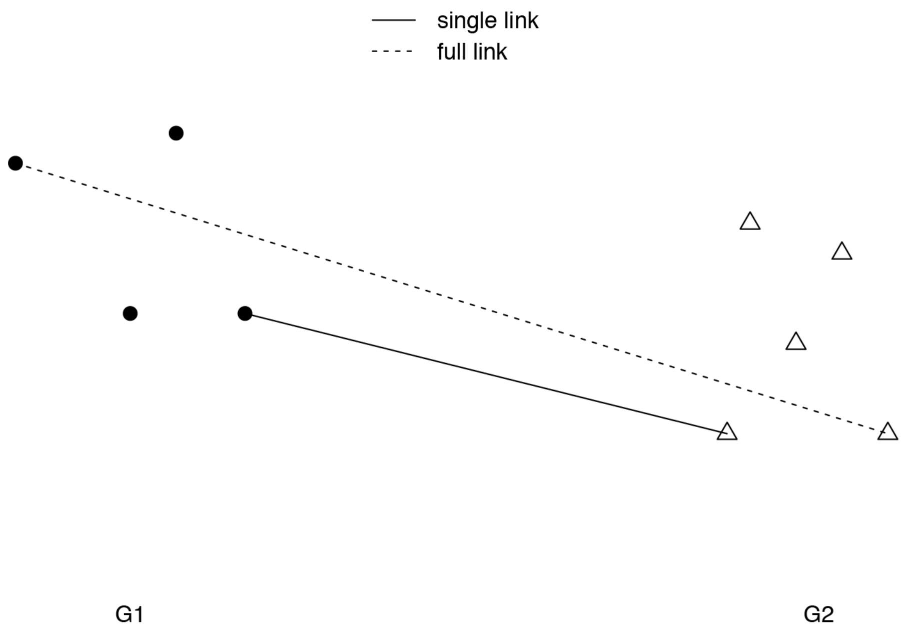
$$\begin{array}{cccc} & A & B & C \\ A & 0 & & \\ B & 2 & 0 & \\ C & \sqrt{3} & \sqrt{3} & 0 \end{array}$$

Manhattan distance:

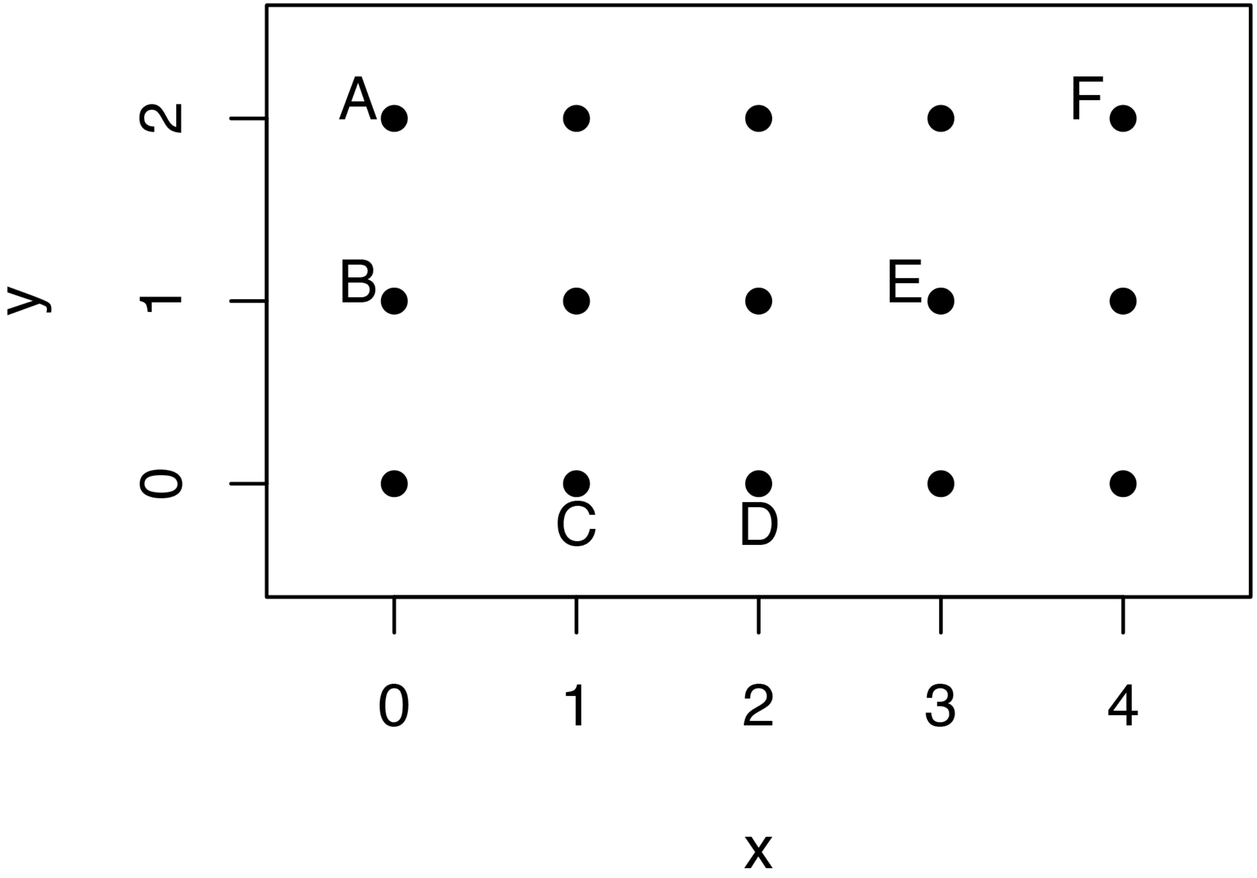
$$\begin{array}{cccc} & A & B & C \\ A & 0 & & \\ B & 2 & 0 & \\ C & 3 & 3 & 0 \end{array}$$

Resemblance between groups

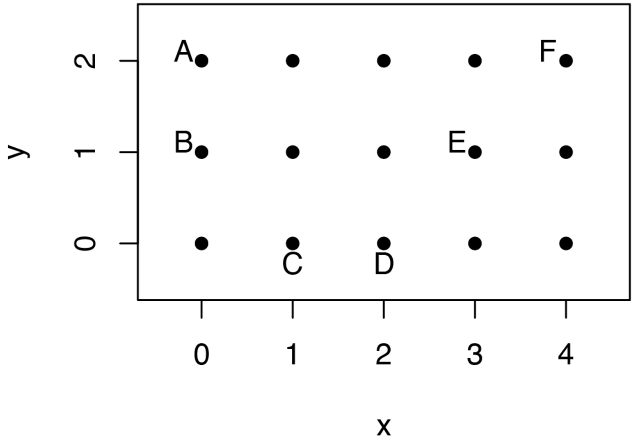
- ➔ Minimum distance
  - › *single link*
- ➔ Maximum distance
  - › *complete link*
- ➔ Average distance (barycenter)
  - › *Average*
- ➔ Within group variability



A toy example: cluster 5 points in 2 dimensions



A toy example: cluster 5 points in 2 dimensions



Manhattan distance matrix:

$$
\begin{array}{c|cccccc}
 & A & B & C & D & E & F \\
A & 0 & & & & & \\
B & \mathbf{1} & 0 & & & & \\
C & 3 & 2 & 0 & & & \\
D & 4 & 3 & 1 & 0 & & \\
E & 4 & 3 & 3 & 2 & 0 & \\
F & 4 & 5 & 5 & 4 & 2 & 0
\end{array}
$$

A toy example: cluster 5 points in 2 dimensions

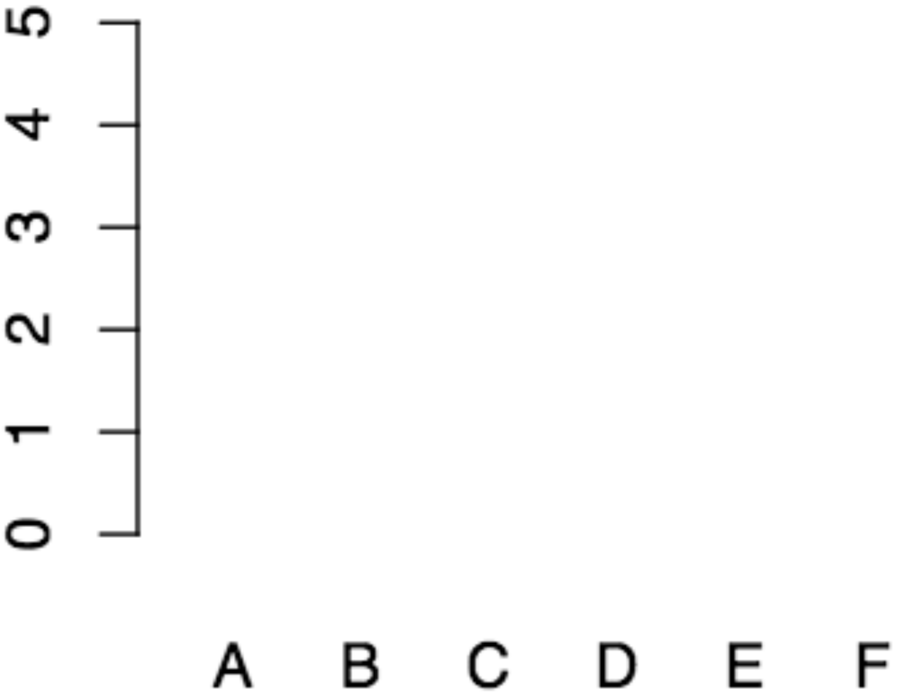
➔ Building the dendrogram (complete link)

$$
\begin{array}{ccccccc}
 & A & B & C & D & E & F \\
A & 0 & & & & & \\
B & \mathbf{1} & 0 & & & & \\
C & 3 & 2 & 0 & & & \\
D & 4 & 3 & 1 & 0 & & \\
E & 4 & 3 & 3 & 2 & 0 & \\
F & 4 & 5 & 5 & 4 & 2 & 0
\end{array}
$$

A toy example: cluster 5 points in 2 dimensions

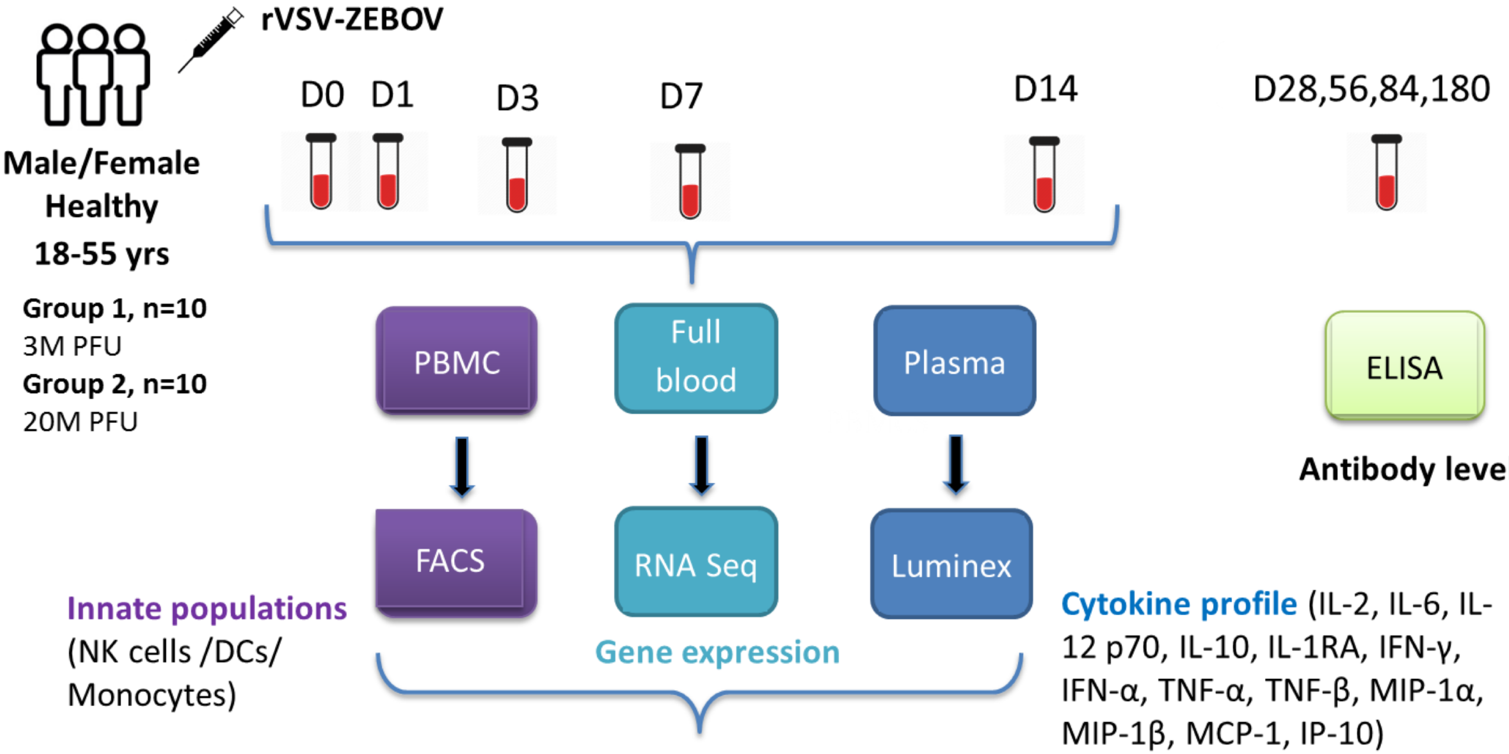
- Building the dendrogram (complete link)

$$
\begin{array}{c|cccccc}
 & A & B & C & D & E & F \\
A & 0 & & & & & \\
B & \mathbf{1} & 0 & & & & \\
C & 3 & 2 & 0 & & & \\
D & 4 & 3 & 1 & 0 & & \\
E & 4 & 3 & 3 & 2 & 0 & \\
F & 4 & 5 & 5 & 4 & 2 & 0
\end{array}
$$



rVSV-ZEBOV RNA-seq data example

→ Clinical trial against Ebola virus



rVSV-ZEBOV

D0 D1 D3 D7 D14 D28,56,84,180

Male/Female
Healthy
18-55 yrs

Group 1, n=10
3M PFU
Group 2, n=10
20M PFU

PBMC → FACS

Full blood → RNA Seq

Plasma → Luminex

ELISA

Antibody level

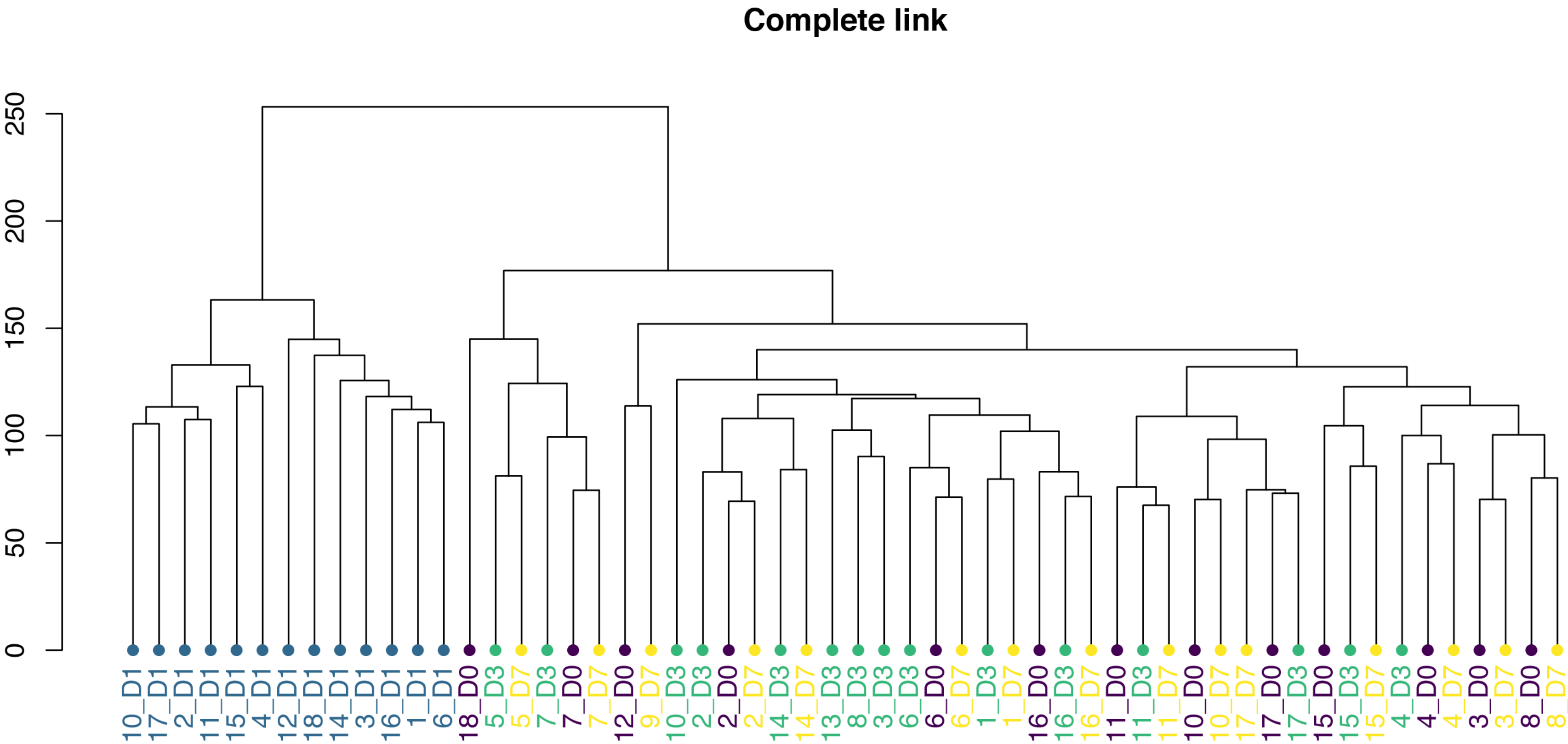
Innate populations (NK cells /DCs/ Monocytes)

Gene expression

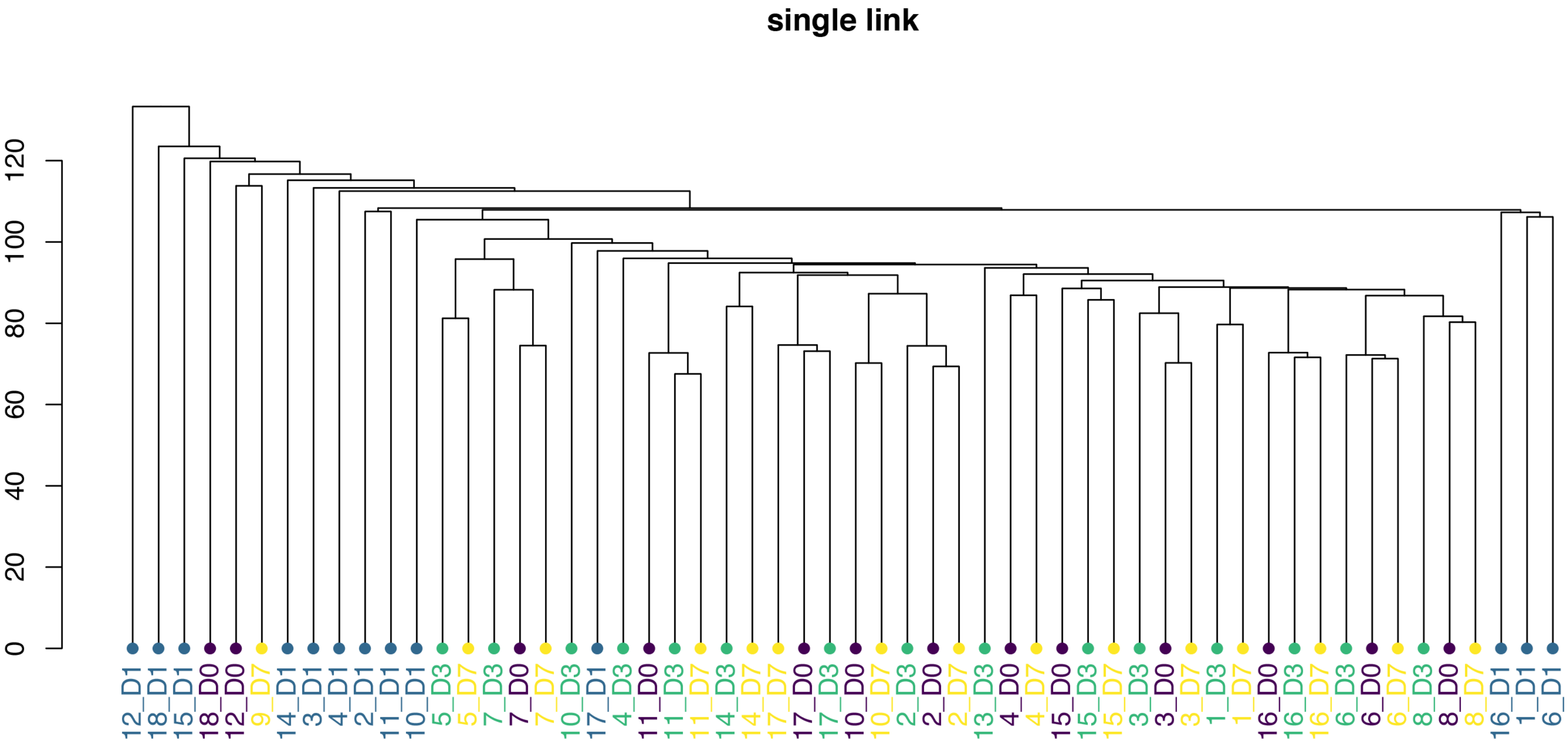
Cytokine profile (IL-2, IL-6, IL-12 p70, IL-10, IL-1RA, IFN-γ, IFN-α, TNF-α, TNF-β, MIP-1α, MIP-1β, MCP-1, IP-10)

Systems vaccinology:
D1, D3, D7 integrative analyses and statistical modelling to identify early signature correlating with antibody response

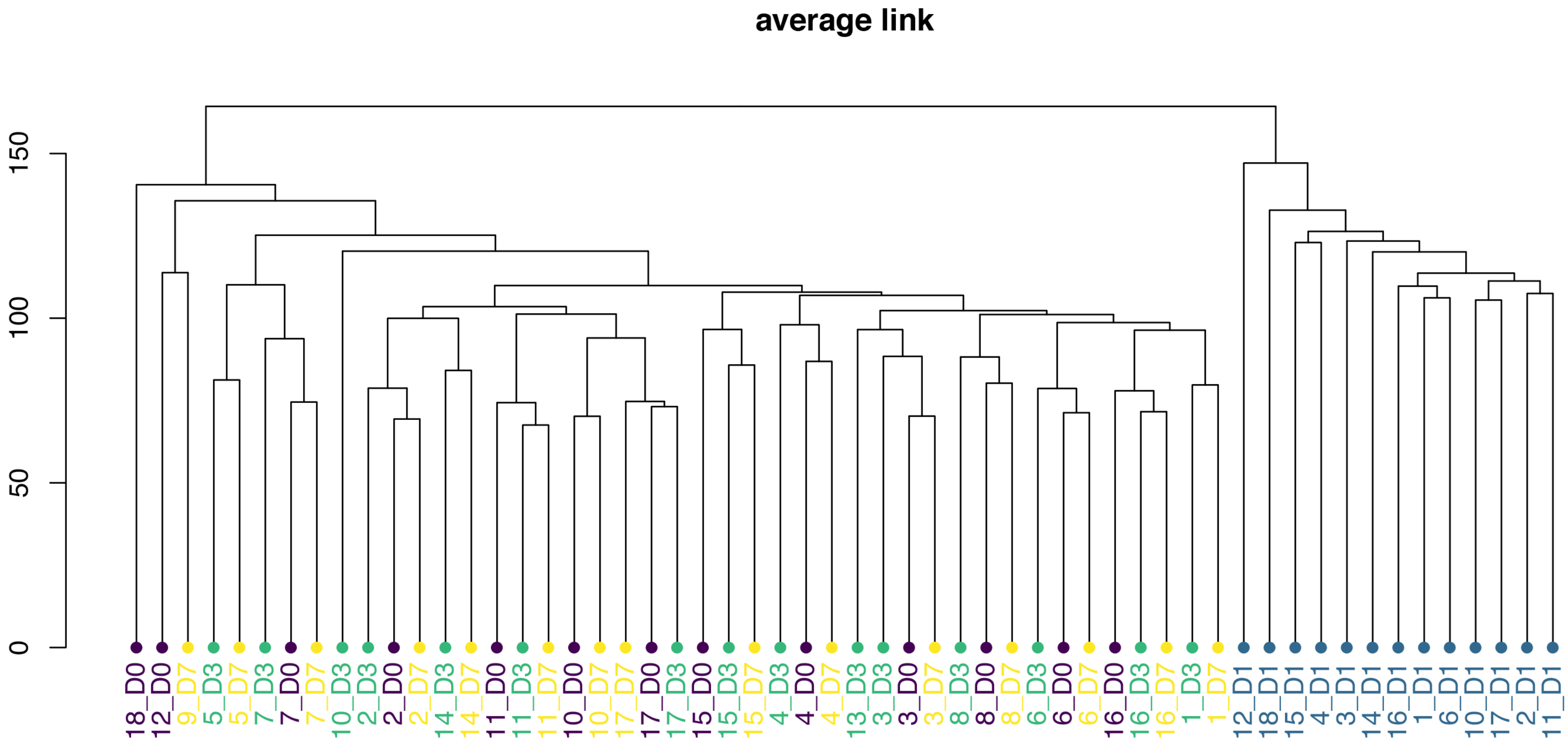
rVSV-ZEBOV RNA-seq: complete link results



rVSV-ZEBOV RNA-seq: single link results



rVSV-ZEBOV RNA-seq: average link results



Ward method and choosing the number of clusters

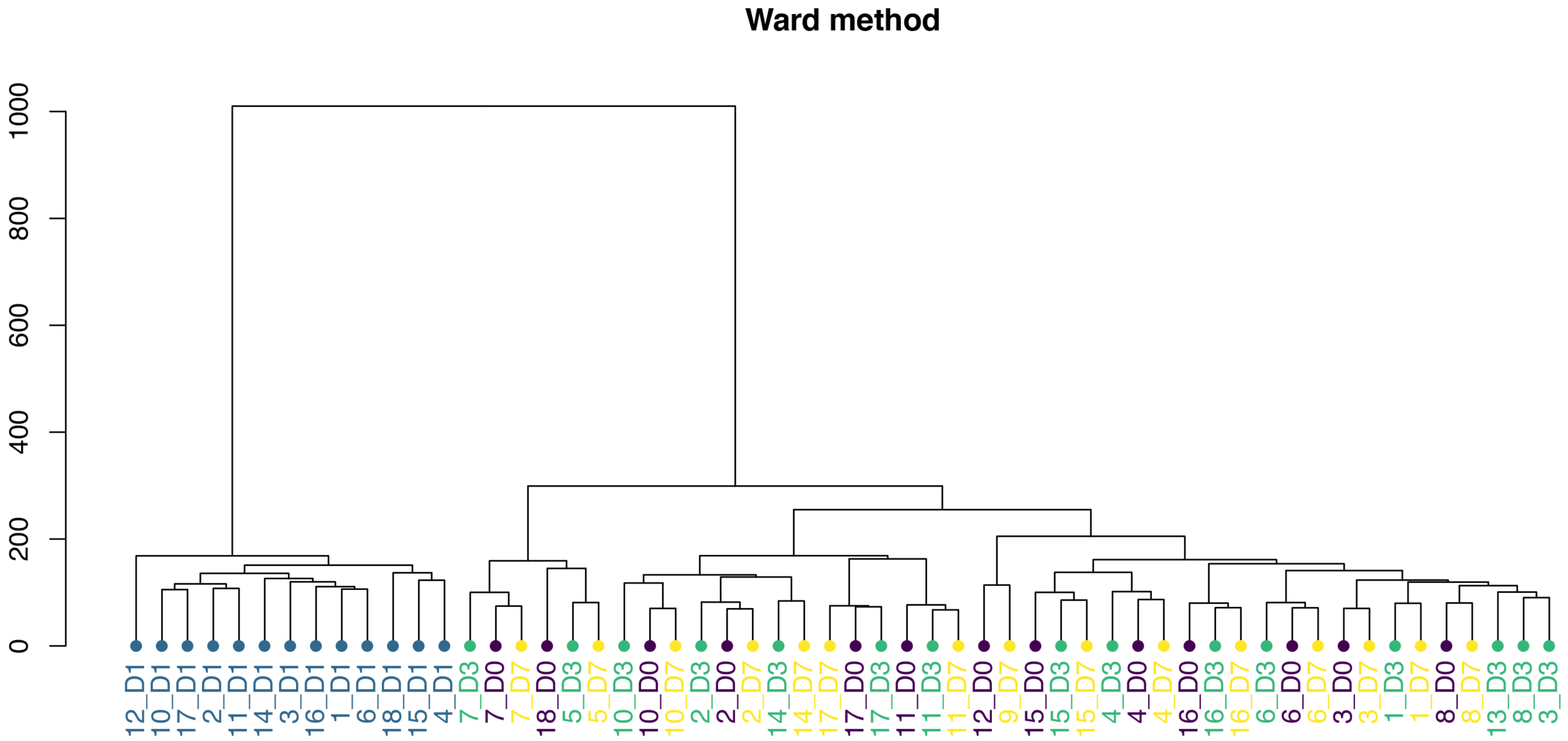
Ward method: Ascending hierarchical method aggregating individuals one at a time by minimizing the intra-group variability increase

The number of clusters K is chosen by maximizing the following criterion

$$\frac{\text{Inter-group variability at } K \text{ clusters vs } K\text{-}1}{\text{Inter-group variability at } K+1 \text{ clusters vs } K}$$

i.e. we pick the number of clusters for which the next aggregation step represents the biggest jump in the tree

rVSV-ZEBOV RNA-seq: Ward method results



Important things to consider for **Hierarchical Clustering**

- What distance/similarity measure was used ?
  - *usually Euclidean*

- What grouping rule was used
  - *Ward is recommended*

- How was the number of clusters chosen?

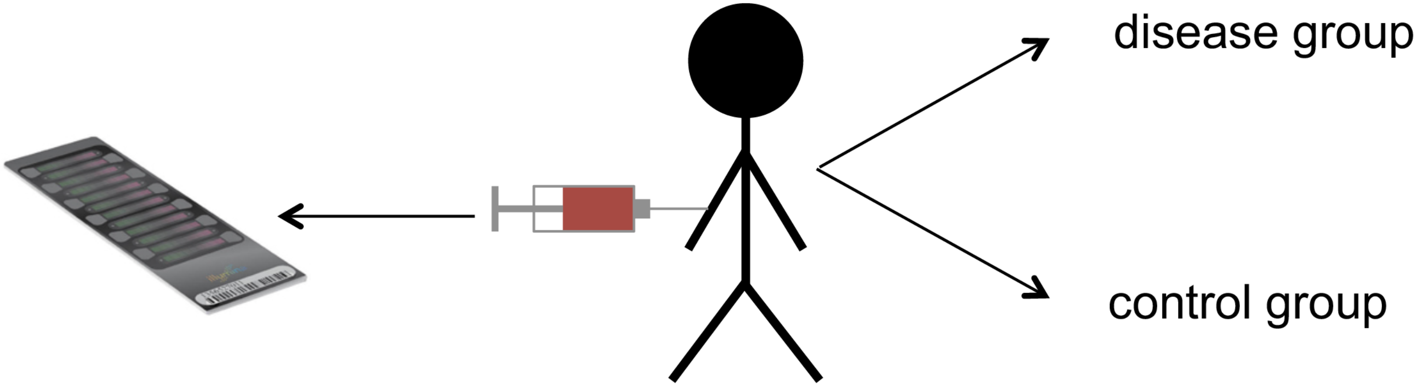
NB: can be slow to compute on large data

Other unsupervised clustering algorithms

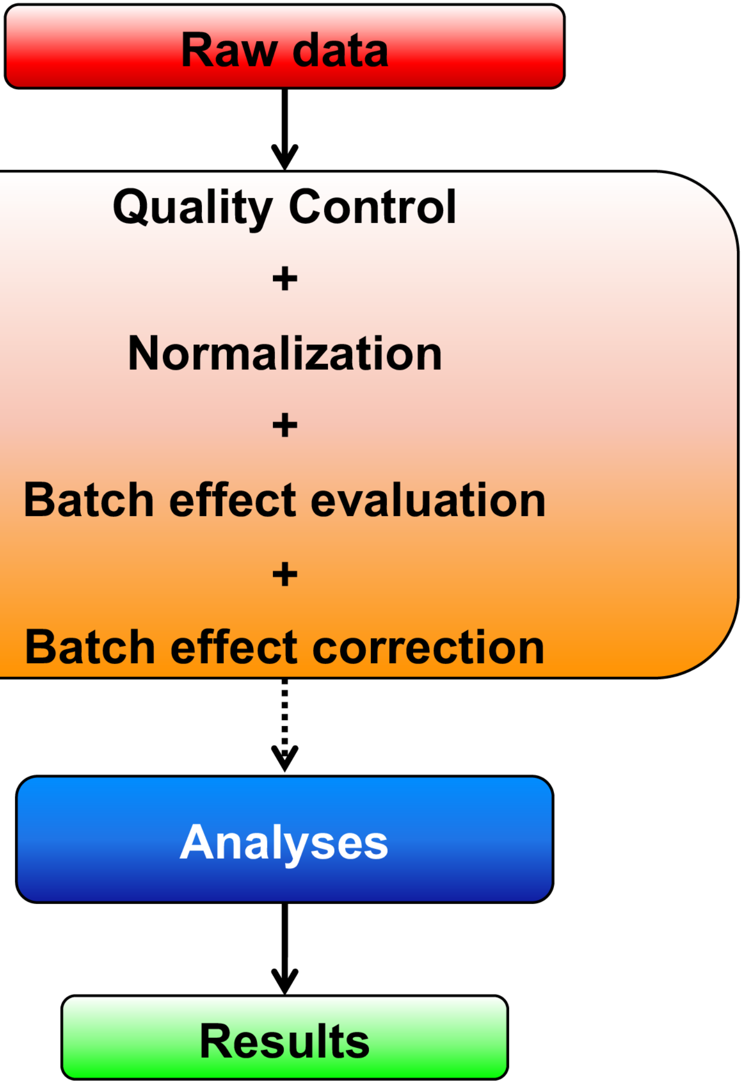
- K means
- Mixture models
- …

TRANSCRIPTOMICS

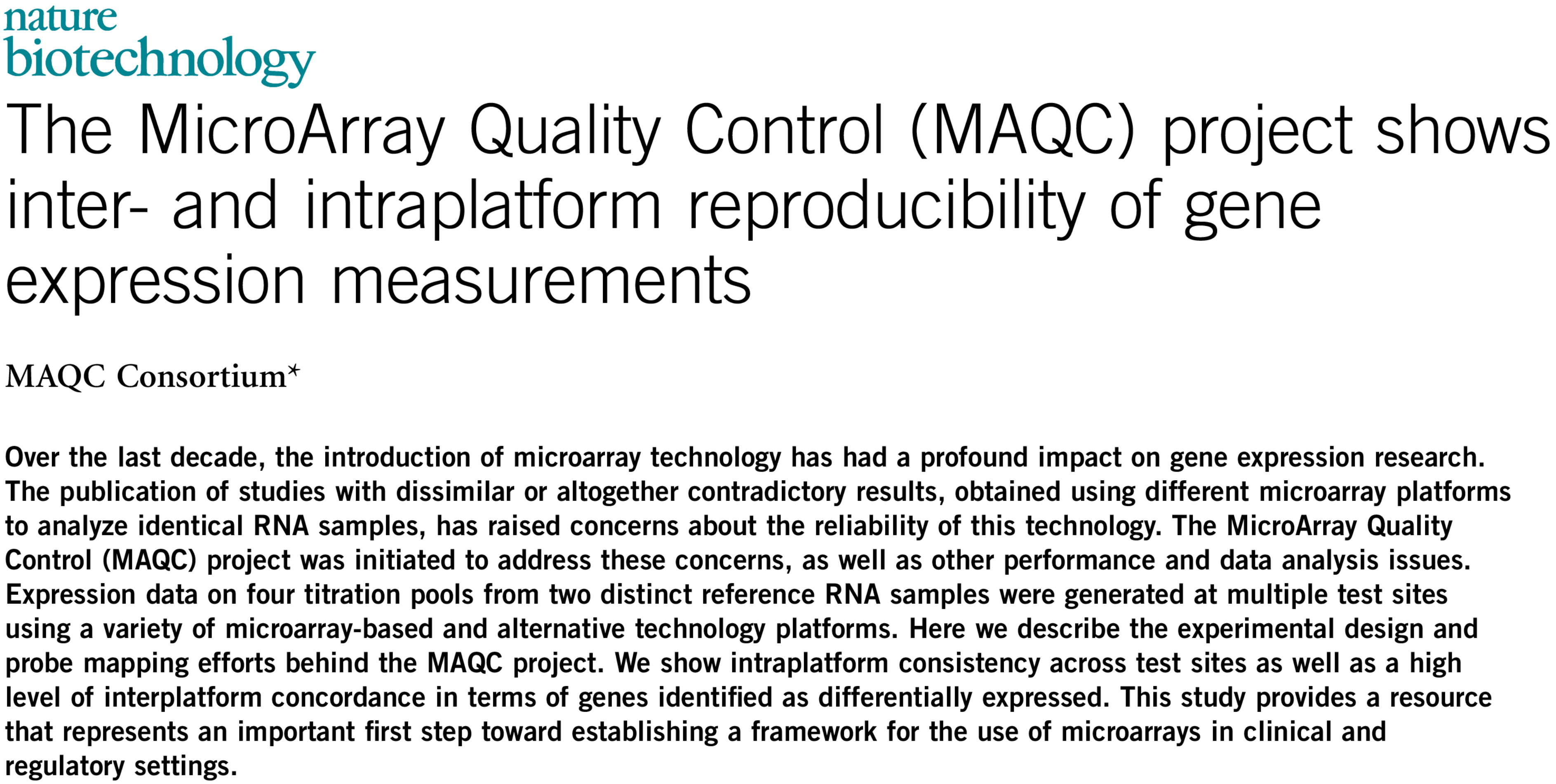
Differential gene expression



Gene expression analysis pipeline



Microarray



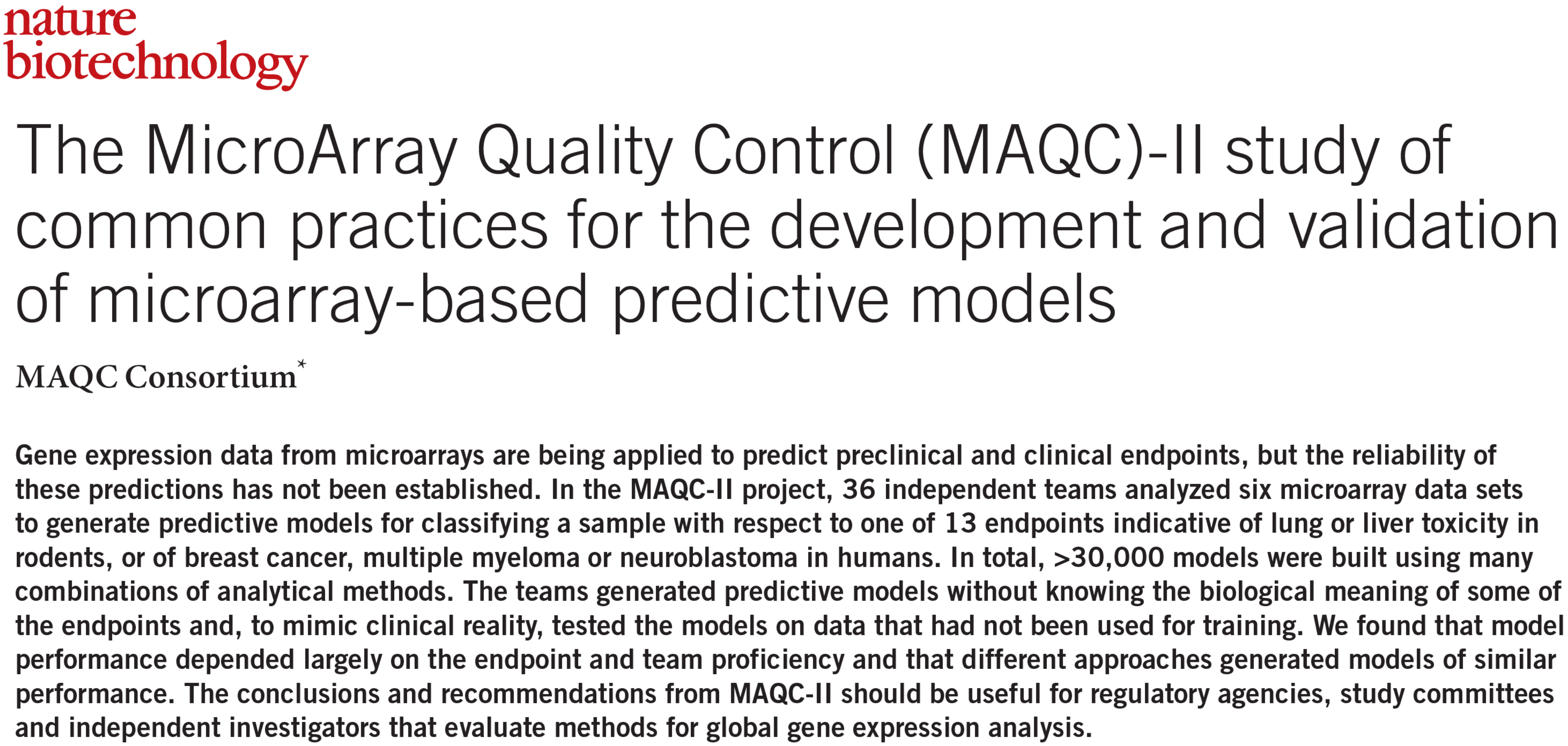
nature biotechnology

The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements

MAQC Consortium*

Over the last decade, the introduction of microarray technology has had a profound impact on gene expression research. The publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms to analyze identical RNA samples, has raised concerns about the reliability of this technology. The MicroArray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues. Expression data on four titration pools from two distinct reference RNA samples were generated at multiple test sites using a variety of microarray-based and alternative technology platforms. Here we describe the experimental design and probe mapping efforts behind the MAQC project. We show intraplatform consistency across test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed. This study provides a resource that represents an important first step toward establishing a framework for the use of microarrays in clinical and regulatory settings.

Microarray



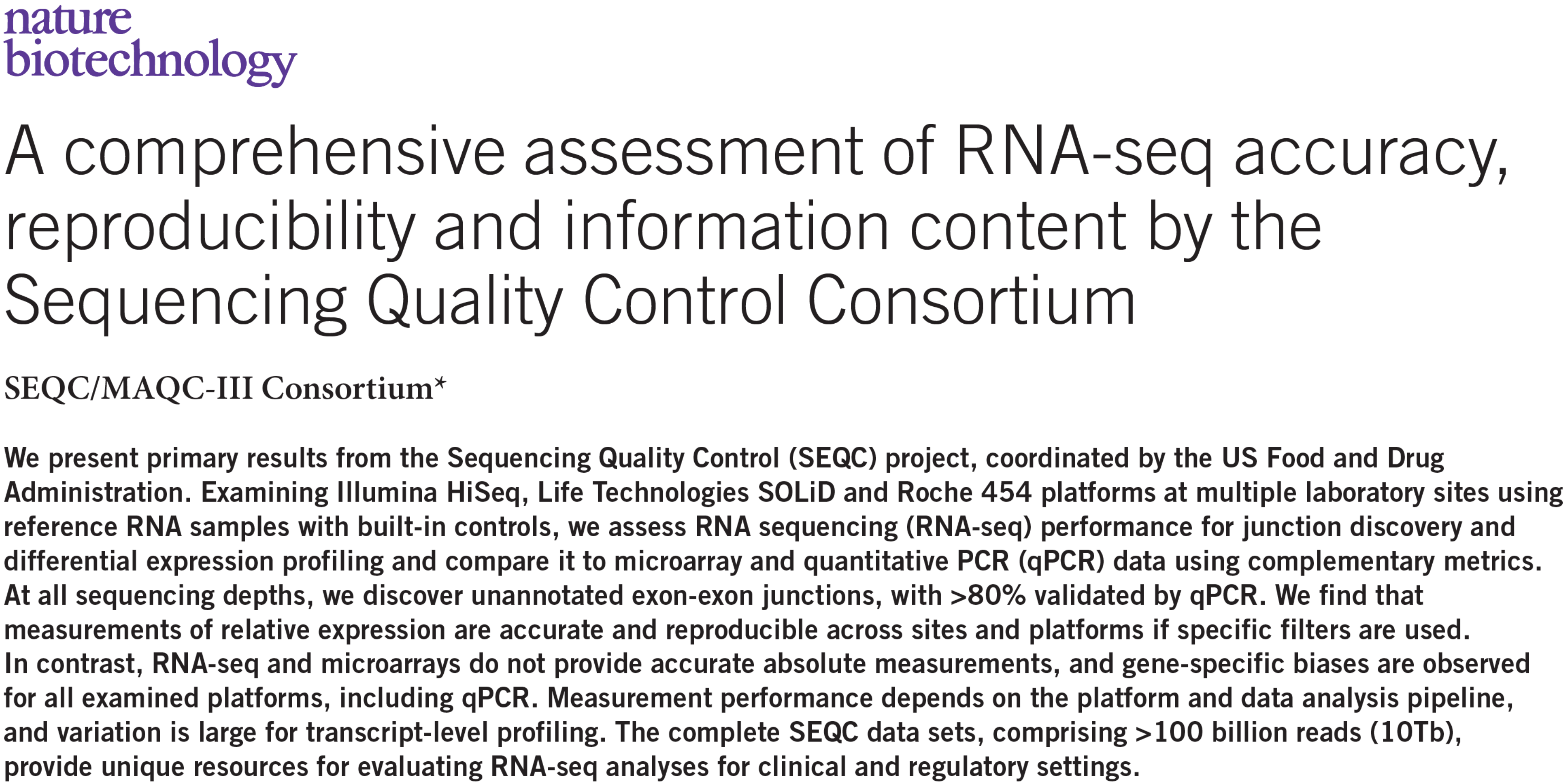
nature
biotechnology

The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

MAQC Consortium*

Gene expression data from microarrays are being applied to predict preclinical and clinical endpoints, but the reliability of these predictions has not been established. In the MAQC-II project, 36 independent teams analyzed six microarray data sets to generate predictive models for classifying a sample with respect to one of 13 endpoints indicative of lung or liver toxicity in rodents, or of breast cancer, multiple myeloma or neuroblastoma in humans. In total, >30,000 models were built using many combinations of analytical methods. The teams generated predictive models without knowing the biological meaning of some of the endpoints and, to mimic clinical reality, tested the models on data that had not been used for training. We found that model performance depended largely on the endpoint and team proficiency and that different approaches generated models of similar performance. The conclusions and recommendations from MAQC-II should be useful for regulatory agencies, study committees and independent investigators that evaluate methods for global gene expression analysis.

RNA-seq technology



nature biotechnology

A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium

SEQC/MAQC-III Consortium*

We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the US Food and Drug Administration. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, we assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics. At all sequencing depths, we discover unannotated exon-exon junctions, with >80% validated by qPCR. We find that measurements of relative expression are accurate and reproducible across sites and platforms if specific filters are used. In contrast, RNA-seq and microarrays do not provide accurate absolute measurements, and gene-specific biases are observed for all examined platforms, including qPCR. Measurement performance depends on the platform and data analysis pipeline, and variation is large for transcript-level profiling. The complete SEQC data sets, comprising >100 billion reads (10Tb), provide unique resources for evaluating RNA-seq analyses for clinical and regulatory settings.

Data generation

RNA-seq data generation

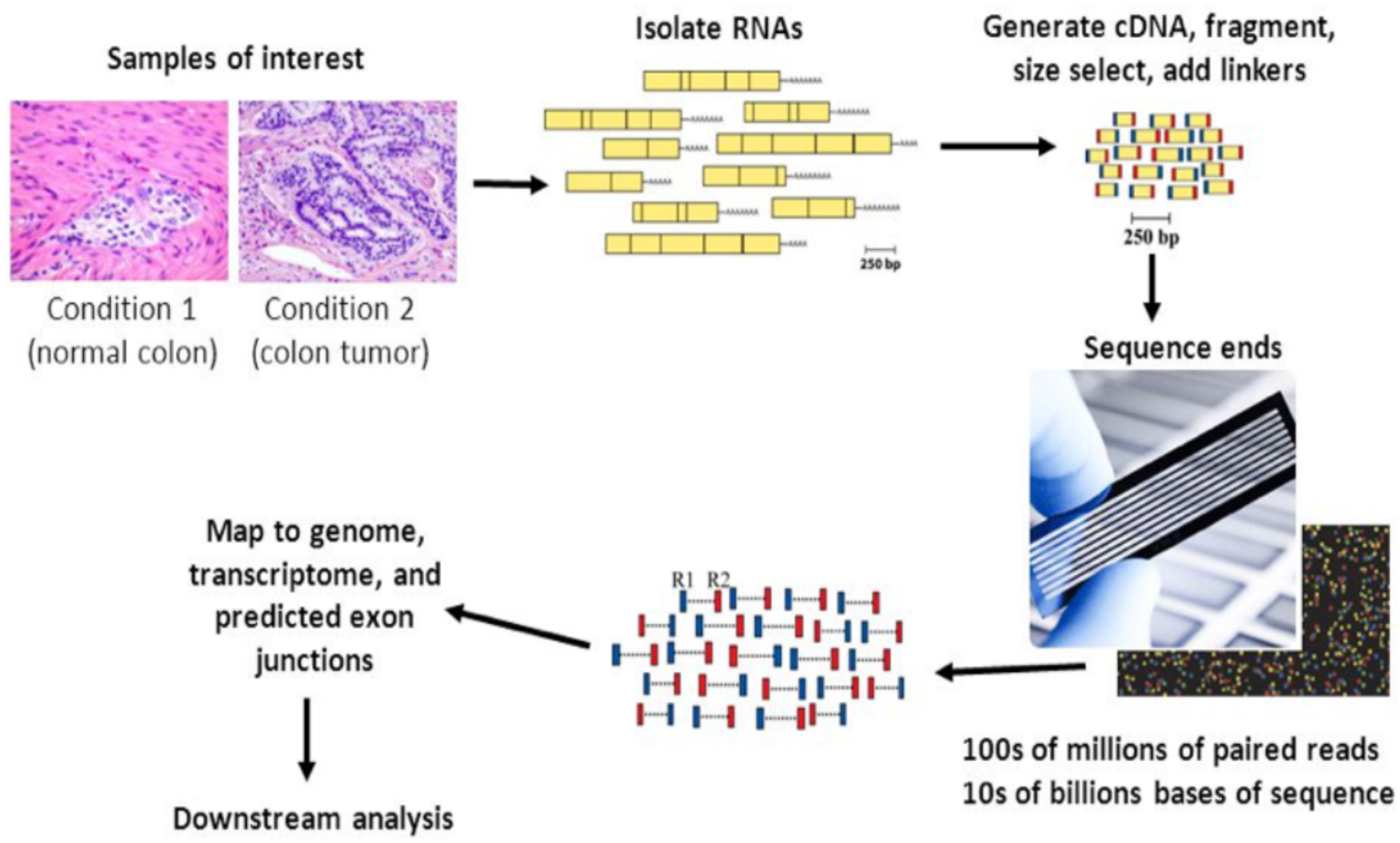
- RNA-Seq is used to analyze the continually changing cellular transcriptome

- **Data acquisition:**
  - Illumina HiSeq (2000, 2500, 4000)
  - SOLiD
  - Ion Torrent
  - …

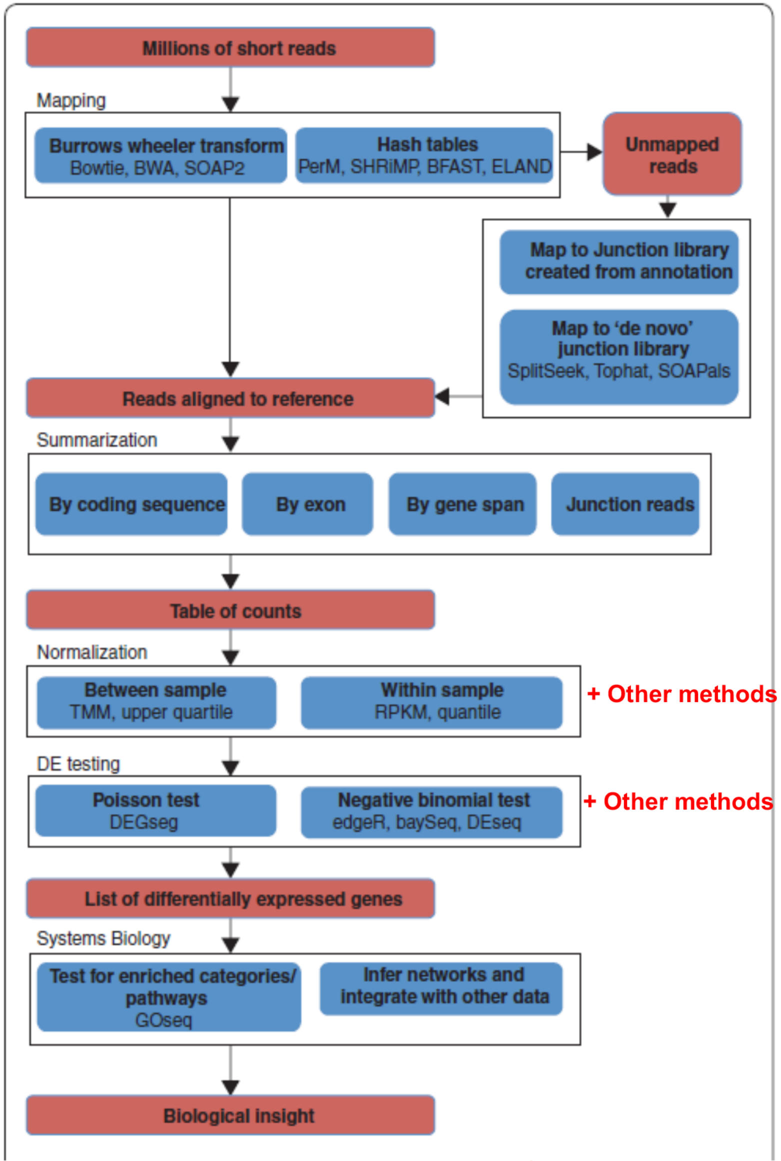
- **Ilumina technologics characteristics**
  - Illumina tutorial
  - Single/paired end
  - Length reads : ~20-500 bp
  - ~ 60 000 reads per sample -> Size per sample: ~6 Go

- **Fastq files**

RNA-seq data process



RNA-seq analysis pipeline



Millions of short reads

Mapping

Burrows wheeler transform
Bowtie, BWA, SOAP2

Hash tables
PerM, SHRiMP, BFAST, ELAND

Unmapped reads

Map to Junction library created from annotation

Map to 'de novo' junction library
SplitSeek, Tophat, SOAPals

Reads aligned to reference

Summarization

By coding sequence

By exon

By gene span

Junction reads

Table of counts

Normalization

Between sample
TMM, upper quartile

Within sample
RPKM, quantile

+ Other methods

DE testing

Poisson test
DEGseq

Negative binomial test
edgeR, baySeq, DEseq

+ Other methods

List of differentially expressed genes

Systems Biology

Test for enriched categories/ pathways
GOseq

Infer networks and integrate with other data

Biological insight

Quality Control

Mapping/Alignment

Normalization

Differential gene expression
Integrative analysis

Adapted from Oshlack, *Genom. Biol.*, 2010

Quality control: definitions

→ **RIN**: **RNA Integrity Number**

between 1 to 10, with 10 least degraded RNA: evaluated after RNA isolation

→ **Base composition**: CG content

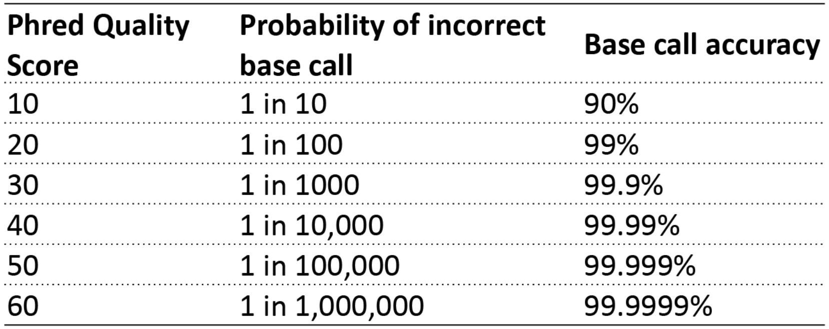
(**NB:** Rich CG content reads are under represented)

→ **Library concentration/Sequencing depth**:

Deep sequencing improves quantification/identification **BUT** can result in increased noise

→ **Quality Score**: Sequencing quality score (probability P) that a base is called incorrectly

Phred/Q score: $P = 10^{-Q/10}$



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality control: criteria

- **RIN ≥ 8** (polyA selection -> biais towards 3' ends of sequences if degraded RNA)

- Rich CG content reads are underrepresented in the sequencing results in term of abundance transcript

- **Sequencing depth > 1 million reads**

The higher the sequencing depth, the higher the sensitivity to capture weakly expressed genes

Read mapping/alignment

→ **Alignment of the sequenced reads**:

- *Reference genome*, GRCh37 or 38 (ensembl.org) for human
- *De novo*

Alignment tools *(Engström, Nature Methods, 2013)*

- Unspliced Aligners: align continuous reads which not contain gaps results by splicing
  - Burrows-Wheeler transform method (**Bowtie2/BWA**…) FAST
  - Needleman-Wunsch or Smith-Waterman algorithms, *seed-extended method* (BFAST, NovoAlign…) MORE SENSITIVE
- Spliced Aligners: (**STAR**, TopHat)
- Pseudo-count (**salmon**, kalisto) FAST, probabilistic

→ **Paired-end (PE) information** improves alignment precision in genome assembly

Can be computationly costly (time)…

Normalization & Batch effect

BORDEAUX POPULATION HEALTH Centre de Recherche - U1219

ISPED SCHOOL OF PUBLIC HEALTH

université de BORDEAUX

Digital Public Health Graduate Program

Why normalize ?

Objective: ***Remove noise and calibrate sample in observed gene counts***

➔ **Between-sample normalization**: *sequencing depth normalization*
(+/- normalization for relative gene abundance – important for generalization)

➔ **Within-sample normalization**: *transcript length normalization*
(if expression of several genes within the same sample are to be compared)

- Differences in library size (sequencing depth) is the most obvious source of variation between lanes !

➔ **Methods** *(Dillies, Brief Bioinfo, 2013)*

- Distribution adjustment of read counts *(assume similarities between distribution)*
  - Total Counts (**TC**): Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the datasets
  - Reads Per Kilobase per Million (**RPKM**):
  - …

Batch effects

➔ **Data correlated for non-biological reason**

- Date of experiments
- Operating technicians
- Laboratory effect
- Atmospheric condition
- Instruments
- …

Unwanted sources of heterogeneity could dramatically reduce the accuracy of statistical inference in genomic data analysis

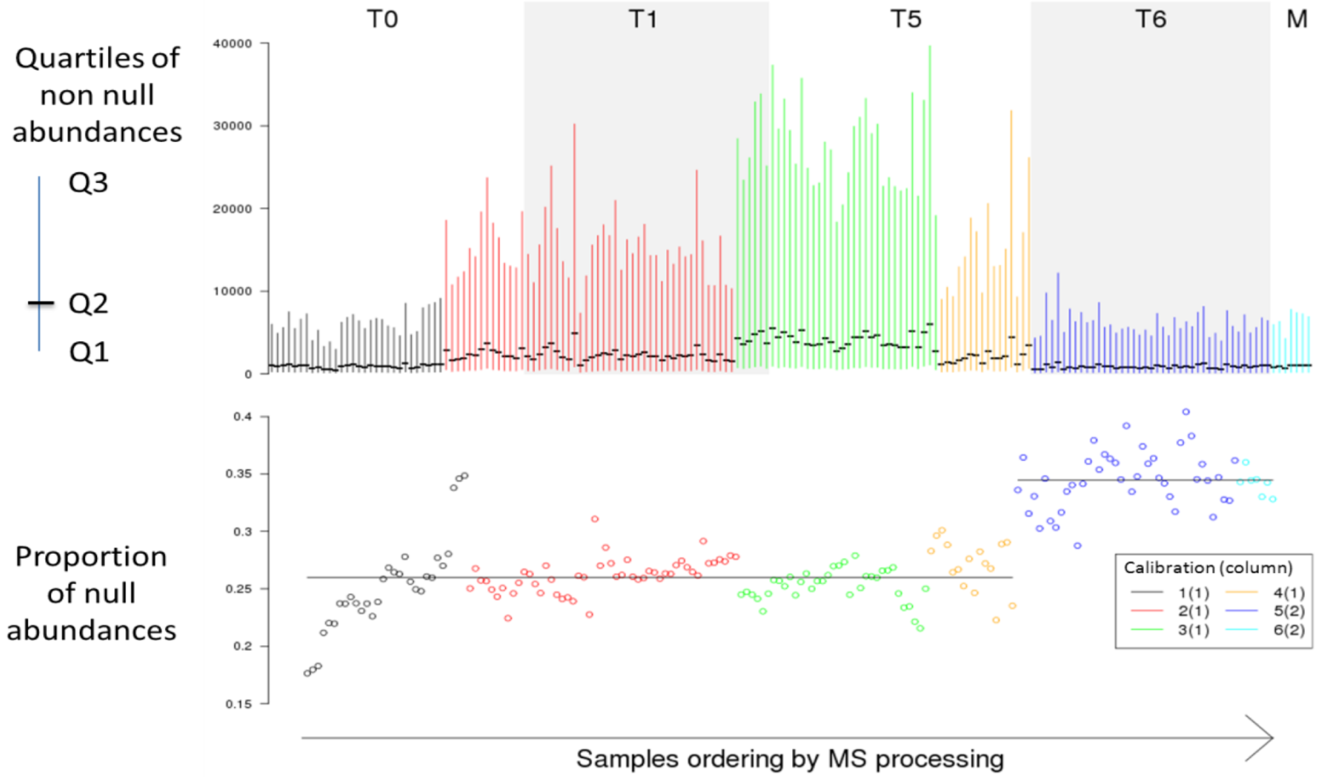
➔ **Methods to deal with batch effects in data**: developed for microarray data

- Adjustment for known batch effects in multivariable DE model
- « removeBatchEffect » function in limma package
- ComBat
- SVA and RUVSeq packages for estimating unknown batch effects

An example of batch effect

Proteomics:

- ➔ Association between responder status after radiotherapy and peptides in serum
- ➔ Mass spectrometry data at 4 time points



Time effect confounded with calibration effect !

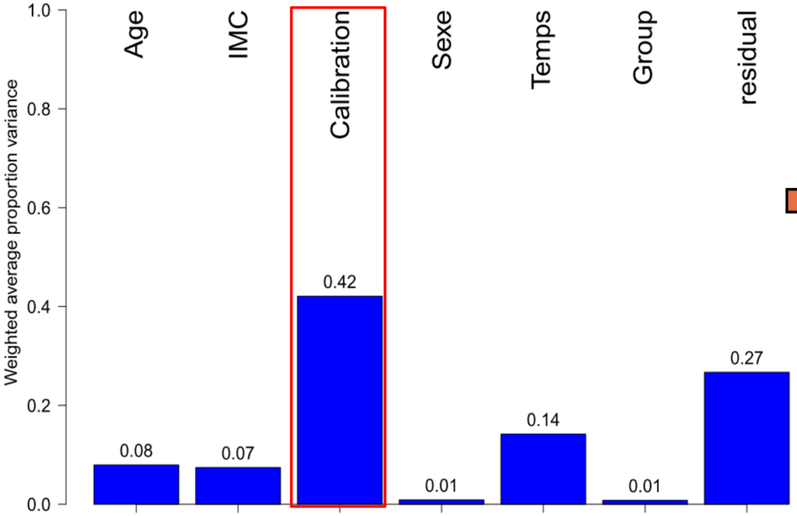
Proportion zeros also associated with calibration…

Measuring batch effects and correction impact

PVCA (Principal Variance Component Analysis)

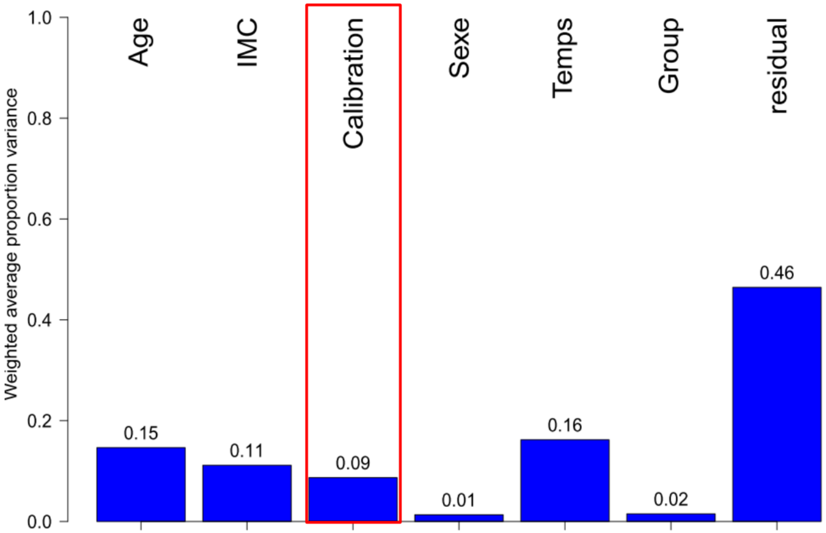
➔ Estimate the source of variability of experimental effects/batch by combining two popular approaches :

➔ **PCA** (dimension reduction) + **VCA** (*mixed effect regression* on PCs)



PVCA on batch effects before ComBat correction

ComBat correction



PVCA on batch effects after ComBat correction

ComBat model

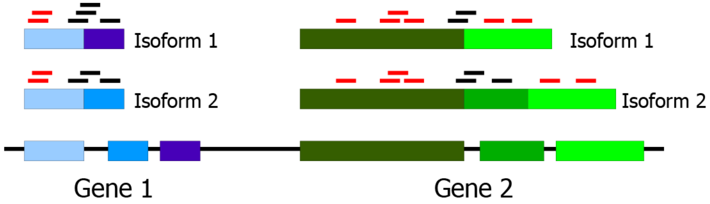
- ➔ The Batch has 2 effects on the expression value:
  - › Additive γ_{ig}
  - › Multiplicative δ_{ig}

- ➔ Bayesian estimation gives a corrected expression:

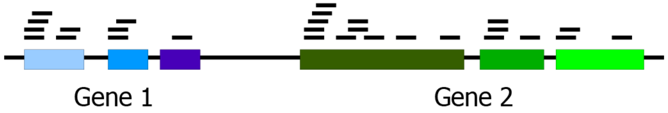
$$Y_{ijg}^{*} = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g - \widehat{\gamma}_{ig}}{\widehat{\delta}_{ig}} + \widehat{\alpha}_g + X\widehat{\beta}_g$$

Quantifying gene expression

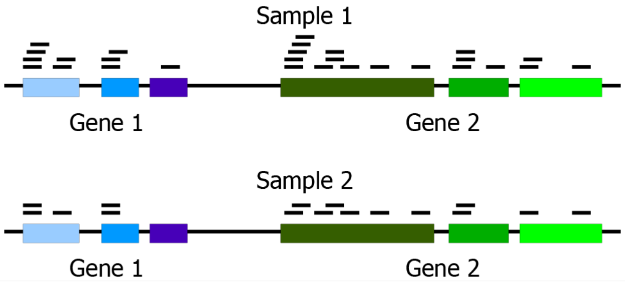
- Ambiguity in reads (multireads, align to more than one isoform)



- Longer genes yield more reads (as they have a higher sampling rate)



- Gene counts depend on total number of sequences (= "**library size**")

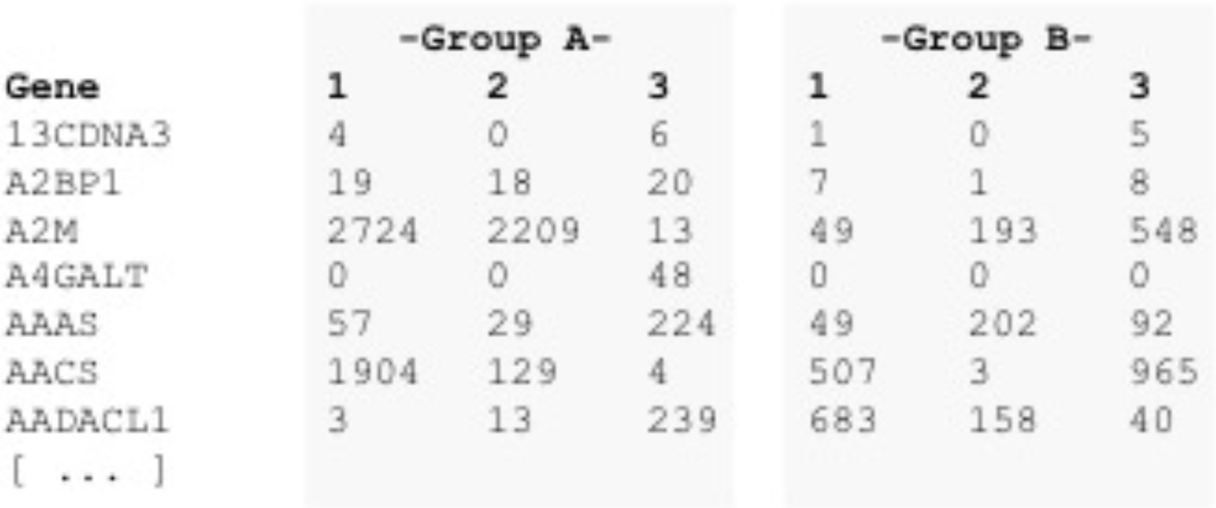


Adapted from A. Rau

Raw counts

Quantification of gene expression is still an **open and active** area of research: isoform-specific expression, strand-specific expression, ambiguity in mapping, ...

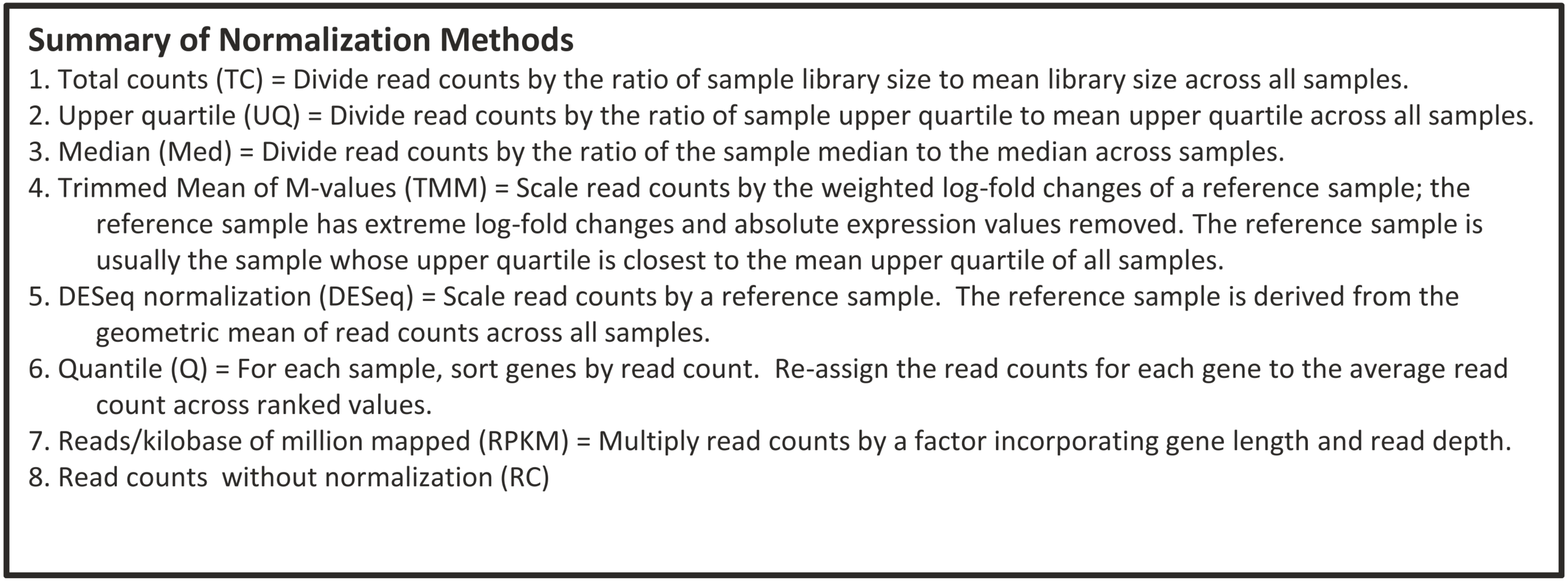
→ **Generally, focus on analysis of count-based measures of gene expression**



	-Group A-			-Group B-		
Gene	**1**	**2**	**3**	**1**	**2**	**3**
13CDNA3	4	0	6	1	0	5
A2BP1	19	18	20	7	1	8
A2M	2724	2209	13	49	193	548
A4GALT	0	0	48	0	0	0
AAAS	57	29	224	49	202	92
AACS	1904	129	4	507	3	965
AADACL1	3	13	239	683	158	40
[...]						

Adapted from A. Rau

Normalization methods



Summary of Normalization Methods

1. Total counts (TC) = Divide read counts by the ratio of sample library size to mean library size across all samples.
2. Upper quartile (UQ) = Divide read counts by the ratio of sample upper quartile to mean upper quartile across all samples.
3. Median (Med) = Divide read counts by the ratio of the sample median to the median across samples.
4. Trimmed Mean of M-values (TMM) = Scale read counts by the weighted log-fold changes of a reference sample; the reference sample has extreme log-fold changes and absolute expression values removed. The reference sample is usually the sample whose upper quartile is closest to the mean upper quartile of all samples.
5. DESeq normalization (DESeq) = Scale read counts by a reference sample. The reference sample is derived from the geometric mean of read counts across all samples.
6. Quantile (Q) = For each sample, sort genes by read count. Re-assign the read counts for each gene to the average read count across ranked values.
7. Reads/kilobase of million mapped (RPKM) = Multiply read counts by a factor incorporating gene length and read depth.
8. Read counts without normalization (RC)

Lin *et al.*, *BMC Genomics*, 2016

Which normalization to chose ?

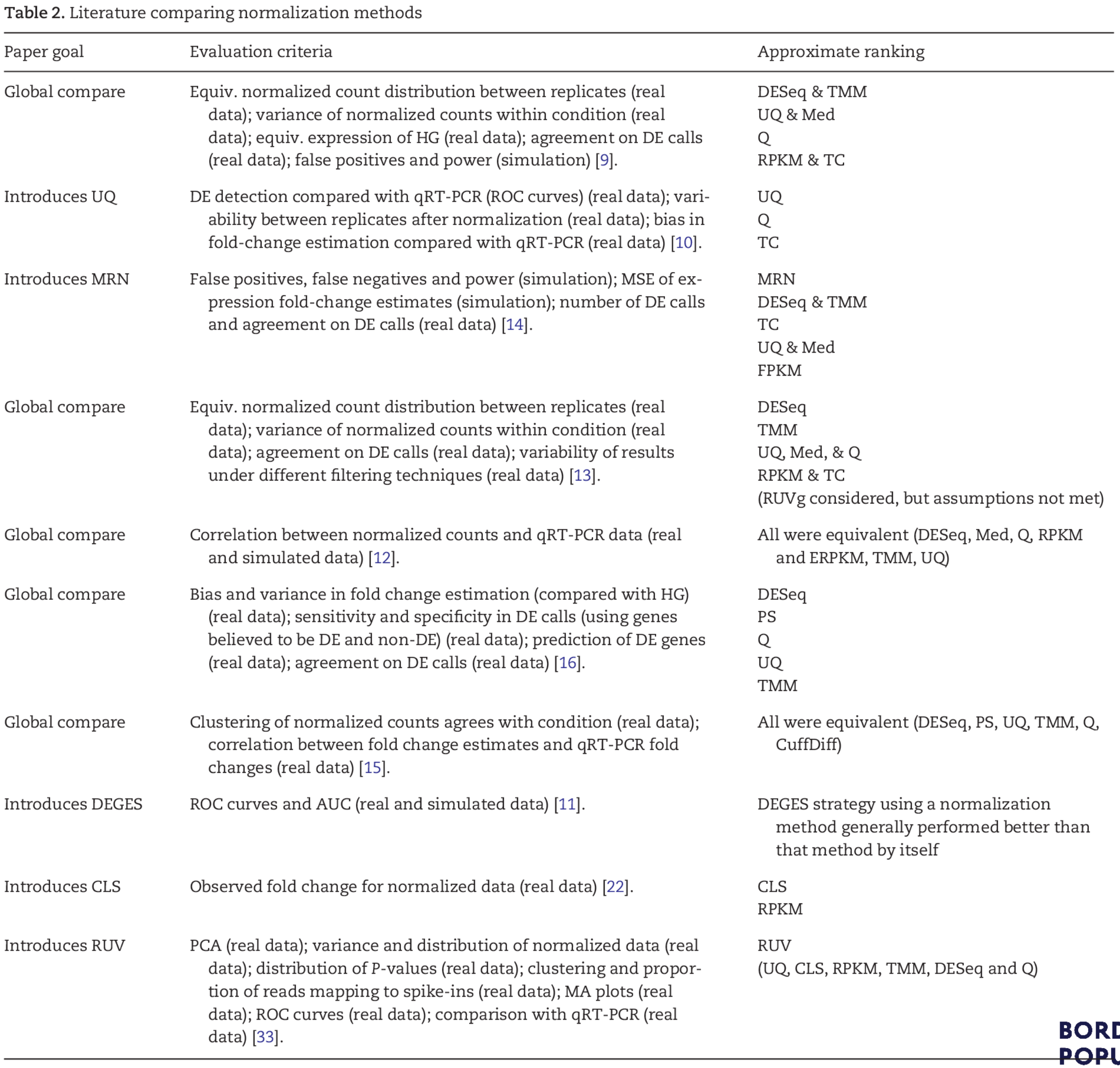


Table 2. Literature comparing normalization methods

Paper goal	Evaluation criteria	Approximate ranking
Global compare	Equiv. normalized count distribution between replicates (real data); variance of normalized counts within condition (real data); equiv. expression of HG (real data); agreement on DE calls (real data); false positives and power (simulation) [9].	DESeq & TMM UQ & Med Q RPKM & TC
Introduces UQ	DE detection compared with qRT-PCR (ROC curves) (real data); variability between replicates after normalization (real data); bias in fold-change estimation compared with qRT-PCR (real data) [10].	UQ Q TC
Introduces MRN	False positives, false negatives and power (simulation); MSE of expression fold-change estimates (simulation); number of DE calls and agreement on DE calls (real data) [14].	MRN DESeq & TMM TC UQ & Med FPKM
Global compare	Equiv. normalized count distribution between replicates (real data); variance of normalized counts within condition (real data); agreement on DE calls (real data); variability of results under different filtering techniques (real data) [13].	DESeq TMM UQ, Med, & Q RPKM & TC (RUVg considered, but assumptions not met)
Global compare	Correlation between normalized counts and qRT-PCR data (real and simulated data) [12].	All were equivalent (DESeq, Med, Q, RPKM and ERPKM, TMM, UQ)
Global compare	Bias and variance in fold change estimation (compared with HG) (real data); sensitivity and specificity in DE calls (using genes believed to be DE and non-DE) (real data); prediction of DE genes (real data); agreement on DE calls (real data) [16].	DESeq PS Q UQ TMM
Global compare	Clustering of normalized counts agrees with condition (real data); correlation between fold change estimates and qRT-PCR fold changes (real data) [15].	All were equivalent (DESeq, PS, UQ, TMM, Q, CuffDiff)
Introduces DEGES	ROC curves and AUC (real and simulated data) [11].	DEGES strategy using a normalization method generally performed better than that method by itself
Introduces CLS	Observed fold change for normalized data (real data) [22].	CLS RPKM
Introduces RUV	PCA (real data); variance and distribution of normalized data (real data); distribution of P-values (real data); clustering and proportion of reads mapping to spike-ins (real data); MA plots (real data); ROC curves (real data); comparison with qRT-PCR (real data) [33].	RUV (UQ, CLS, RPKM, TMM, DESeq and Q)

Evans *et al.*, *Brief. Bioinform.*, 2017.

Normalization methods: take home

- → **Assumptions** allow **normalization** to **translate raw read counts into meaningful measures** of expression.

- → **Suitable normalization** method to use **depends on which assumptions are valid** for a **given** biological experiment.

- → **Incorrect normalization leads to problems** in downstream analysis, such as *inflated false positives*, that mean results cannot be trusted.

- → **No normalization method is perfect**, and for every method there exists cases for which the assumptions are violated. There are examples of global shifts in expression that violate assumptions of conventional normalization methods, requiring controls.

- → **Understanding of assumptions can help** pick the most suitable normalization method for a given experiment.

DIFFERENTIAL EXPRESSION ANALYSIS OF RNA-SEQ DATA

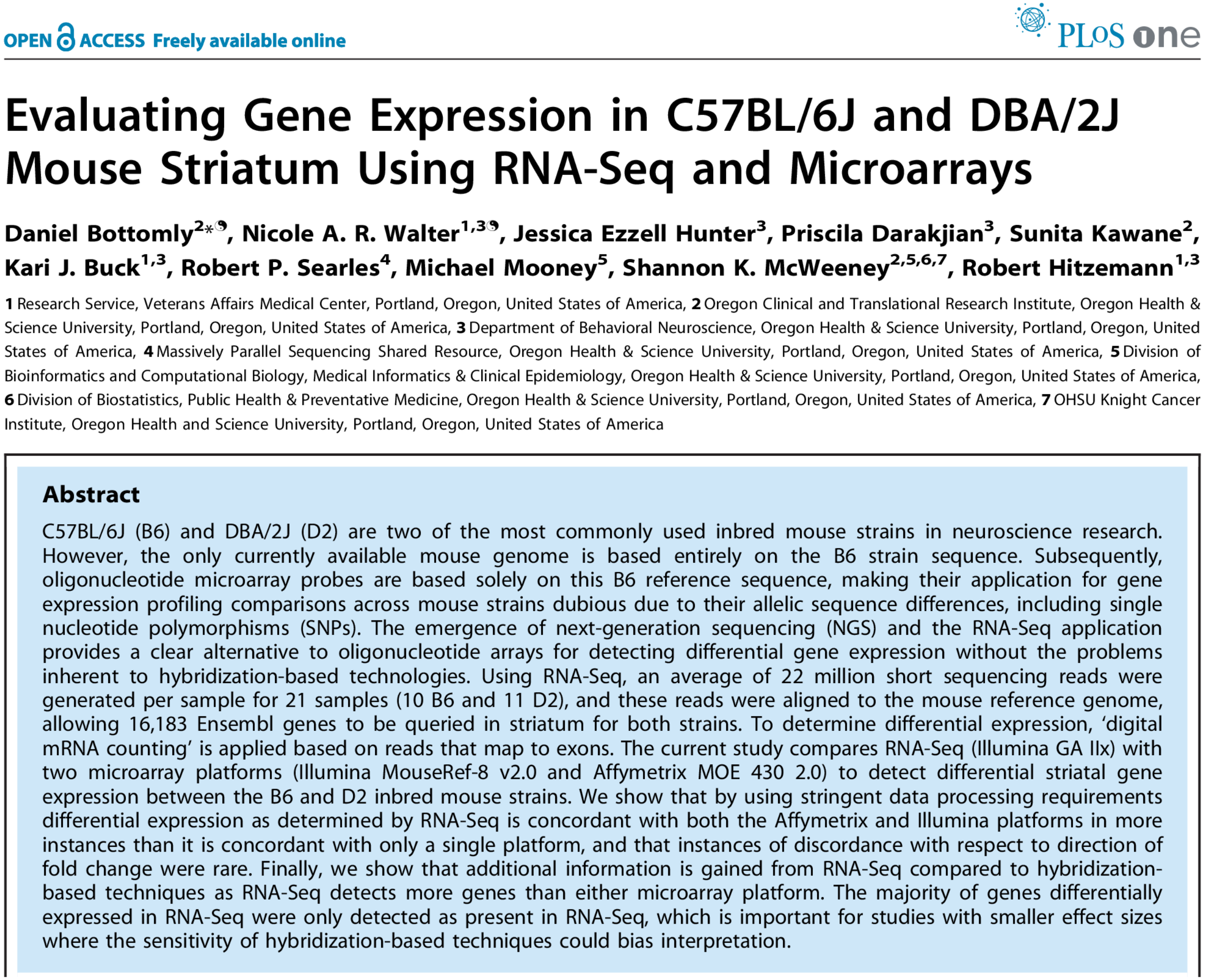
BORDEAUX POPULATION HEALTH Centre de Recherche - U1219

ISPED SCHOOL OF PUBLIC HEALTH

université de BORDEAUX

Digital Public Health Graduate Program

Example dataset



OPEN ACCESS Freely available online

PLoS one

Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays

Daniel Bottomly[2*¶], Nicole A. R. Walter[1,3¶], Jessica Ezzell Hunter[3], Priscila Darakjian[3], Sunita Kawane[2], Kari J. Buck[1,3], Robert P. Searles[4], Michael Mooney[5], Shannon K. McWeeney[2,5,6,7], Robert Hitzemann[1,3]

1 Research Service, Veterans Affairs Medical Center, Portland, Oregon, United States of America, **2** Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, Oregon, United States of America, **3** Department of Behavioral Neuroscience, Oregon Health & Science University, Portland, Oregon, United States of America, **4** Massively Parallel Sequencing Shared Resource, Oregon Health & Science University, Portland, Oregon, United States of America, **5** Division of Bioinformatics and Computational Biology, Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, United States of America, **6** Division of Biostatistics, Public Health & Preventative Medicine, Oregon Health & Science University, Portland, Oregon, United States of America, **7** OHSU Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, United States of America

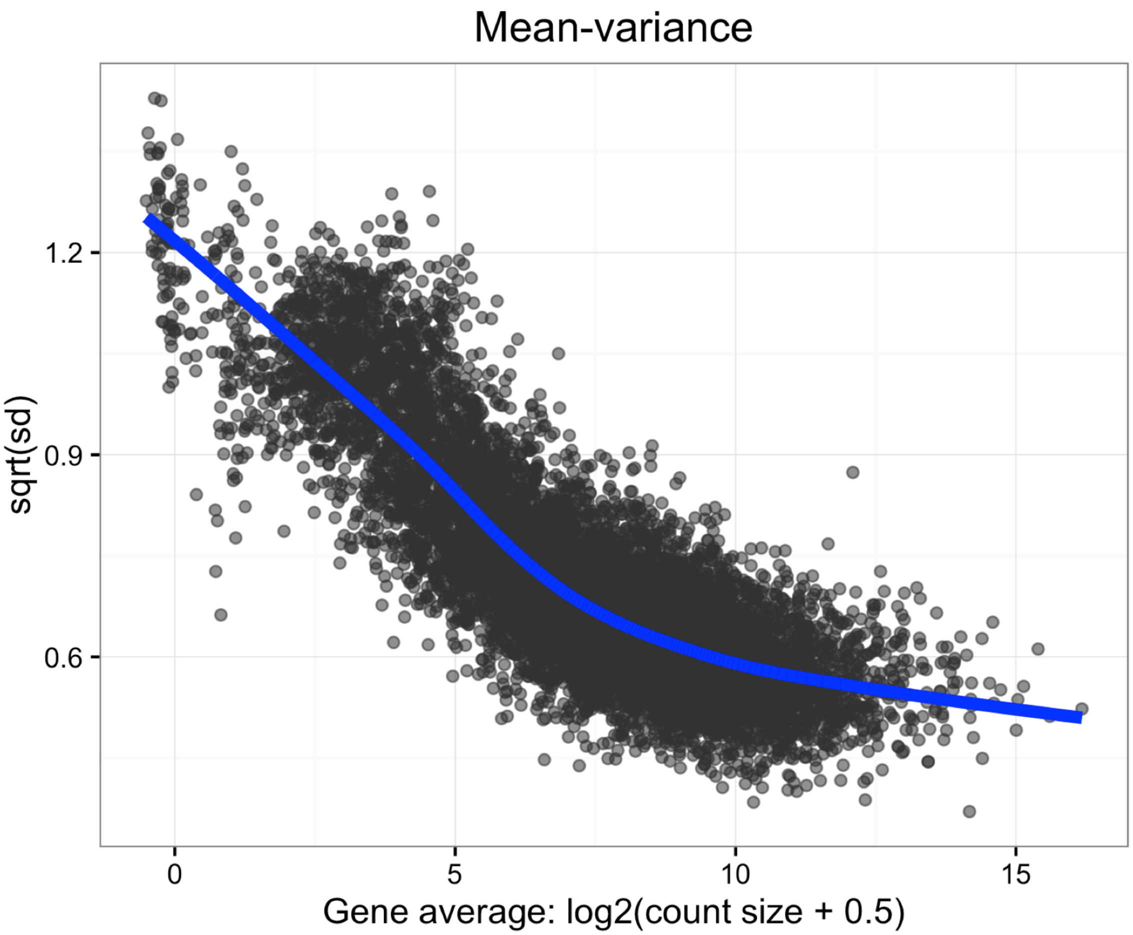
Abstract

C57BL/6J (B6) and DBA/2J (D2) are two of the most commonly used inbred mouse strains in neuroscience research. However, the only currently available mouse genome is based entirely on the B6 strain sequence. Subsequently, oligonucleotide microarray probes are based solely on this B6 reference sequence, making their application for gene expression profiling comparisons across mouse strains dubious due to their allelic sequence differences, including single nucleotide polymorphisms (SNPs). The emergence of next-generation sequencing (NGS) and the RNA-Seq application provides a clear alternative to oligonucleotide arrays for detecting differential gene expression without the problems inherent to hybridization-based technologies. Using RNA-Seq, an average of 22 million short sequencing reads were generated per sample for 21 samples (10 B6 and 11 D2), and these reads were aligned to the mouse reference genome, allowing 16,183 Ensembl genes to be queried in striatum for both strains. To determine differential expression, 'digital mRNA counting' is applied based on reads that map to exons. The current study compares RNA-Seq (Illumina GA IIx) with two microarray platforms (Illumina MouseRef-8 v2.0 and Affymetrix MOE 430 2.0) to detect differential striatal gene expression between the B6 and D2 inbred mouse strains. We show that by using stringent data processing requirements differential expression as determined by RNA-Seq is concordant with both the Affymetrix and Illumina platforms in more instances than it is concordant with only a single platform, and that instances of discordance with respect to direction of fold change were rare. Finally, we show that additional information is gained from RNA-Seq compared to hybridization-based techniques as RNA-Seq detects more genes than either microarray platform. The majority of genes differentially expressed in RNA-Seq were only detected as present in RNA-Seq, which is important for studies with smaller effect sizes where the sensitivity of hybridization-based techniques could bias interpretation.

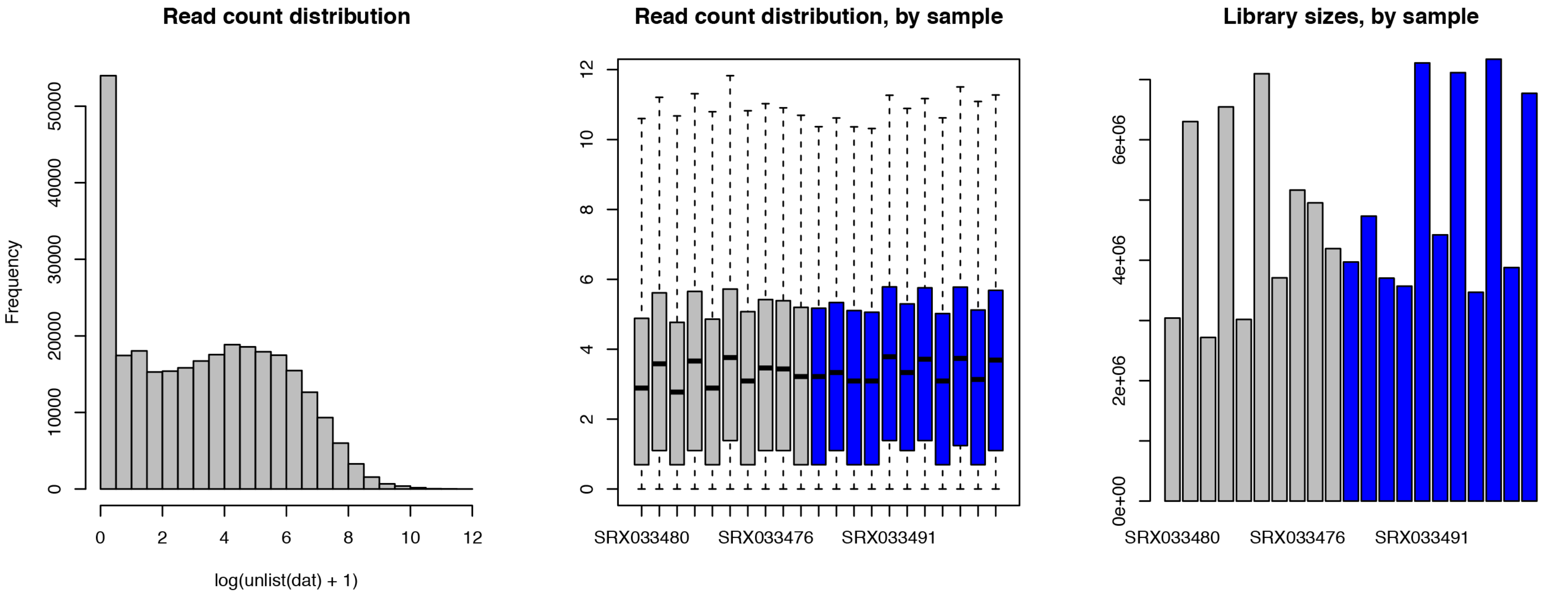
Statistical challenges – 1/2

- **High dimensionality** (large number of genes, few replicates)
- **Discrete**, **positive**, and **skewed** data (heteroscedastic)
- **Large dynamic range** among genes (106 orders of magnitude), presence of 0 counts
  - Typically remove absent genes (those with 0 counts for all samples)
- Sequencing depth (= “**library size**”) varies among samples

Statistical challenges – 2/2



First step: Descriptive exploratory analysis

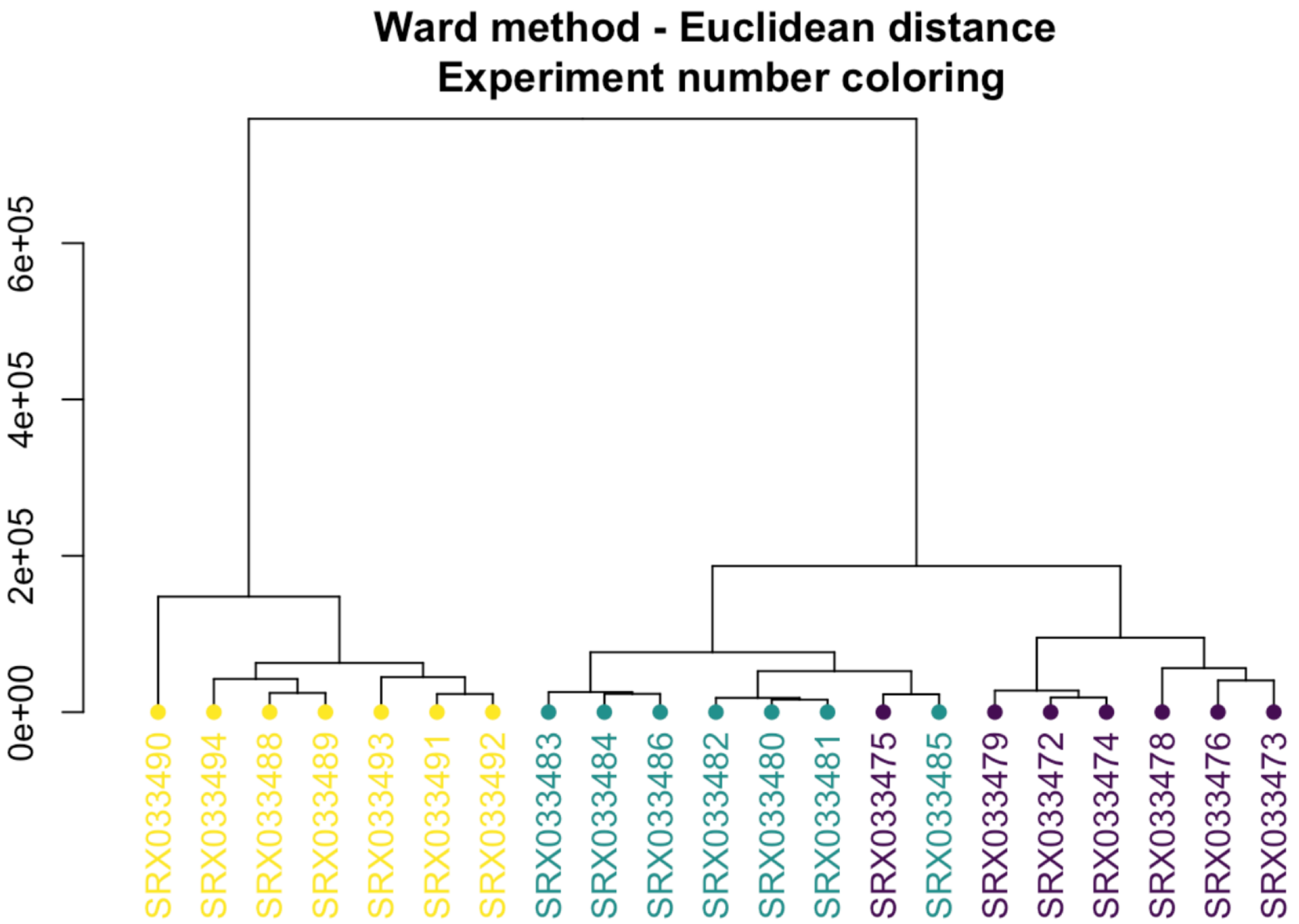


Read count distribution

Read count distribution, by sample

Library sizes, by sample

Hierarchical clustering of samples



Parametric model for RNA-seq data

Poisson modeling

Poisson model

$$\Pr(Y_{ijk} = y_{ijk}) = f(y_{ijk}; \mu_{ijk}) = \frac{e^{-\mu_{ijk}}(\mu_{ijk}^{y_{ijk}})}{y_{ijk}!}$$

Loglinear representation of Poisson model:

$$Y_{ijk} \sim \mathcal{P}(\mu_{ijk})$$

$$\frac{\mu_{ijk}}{m_{jk}} = \exp(\alpha_i + \beta_{ij}) \Leftrightarrow \log\left(\frac{\mu_{ijk}}{m_{jk}}\right) = \alpha_i + \beta_{ij}$$

where m_{jk} is a normalization factor included as an offset in the model

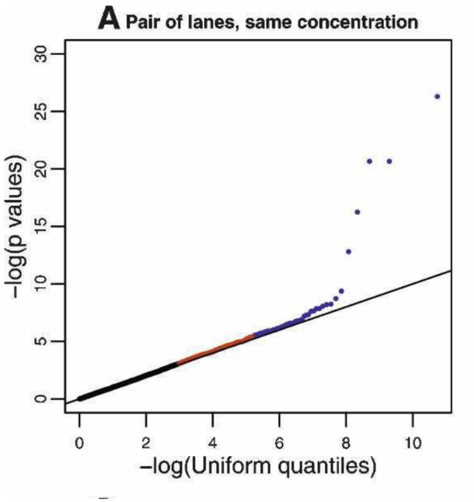
- $\mathrm{E}(Y) = \mathrm{Var}(Y)$
- If $Y_1 \sim \mathcal{P}(\mu_1)$ and $Y_2 \sim \mathcal{P}(\mu_2)$, then $Y_1 + Y_2 \sim \mathcal{P}(\mu_1 + \mu_2)$

Adapted from A. Rau

Poisson modeling: reasonable ?

Is the Poisson model really appropriate for RNA-seq data?

- Nagalakshmi et al. (2008) and Marioni et al. (2008) found that genes from different **technical** replicates have a variance equal to the mean (= Poisson)
- Generally, technical replicates are summed (as the sum of two Poisson random variables is also Poisson)



Marioni et al. (2008), Fig 1 (based on hypergeometric test statistic to compare tech reps)

Adapted from A. Rau

Poisson modeling: reasonable ?

Counts from **biological** replicates tend to have variance exceeding the mean (= **overdispersion**)...

What causes this overdispersion?

- Correlated gene counts
- Clustering of subjects
- Within-group heterogeneity
- Within-group variation in transcription levels
- Different types of noise present...

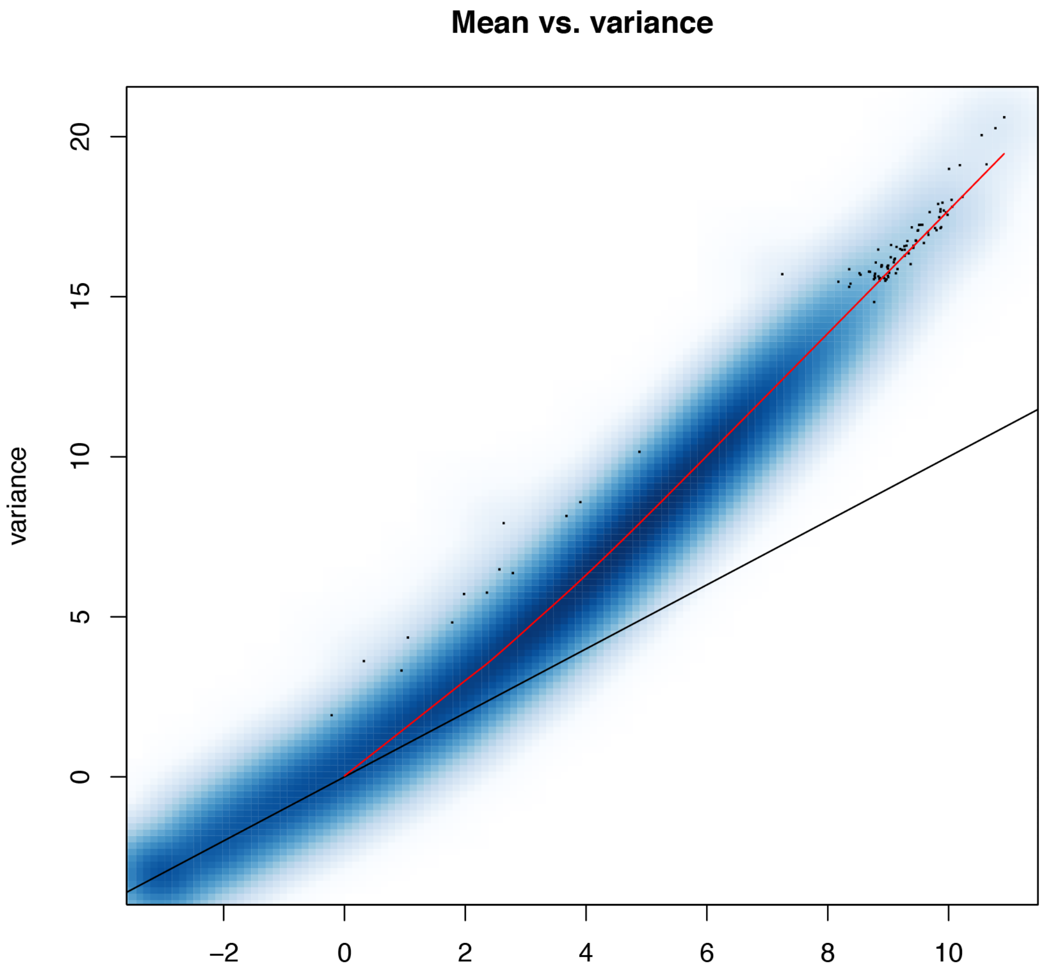
Adapted from A. Rau

Noise sources in RNA-seq data

1. **Shot noise**: unavoidable noise inherent in counting process (dominant for weakly expressed genes)
2. **Technical noise**: from sample preparation and sequencing, hopefully negligable
3. **Biological noise**: unaccounted for differences between samples (dominant for strongly expressed genes)

Adapted from A. Rau

Overdispersion in Bottomly *et al.*



Adapted from A. Rau

Negative Binomial (NB) modeling

Negative binomial model (I)

- Generalization of Poisson with two parameters
- Number of successes in a sequence of Bernoulli trials (with probability p of success) before a specified number of failures (r) occurs:

$$\Pr(Y_{ijk} = y_{ijk}) = f(y_{ijk}; r, p) = \binom{y_{ijk} + r - 1}{y_{ijk}} (1-p)^r p^{y_{ijk}}$$

- $\mathrm{E}(Y_{ijk}) = \frac{pr}{1-p}$
- $\mathrm{Var}(Y_{ijk}) = \frac{pr}{(1-p)^2} = \frac{1}{1-p}\mathrm{E}(Y_{ijk})$

Adapted from A. Rau

Negative Binomial (NB) modeling

Negative binomial model (II)

$$\Pr(Y_{ijk} = y_{ijk}) = f(y_{ijk}; \mu_{ijk}, \phi) =$$

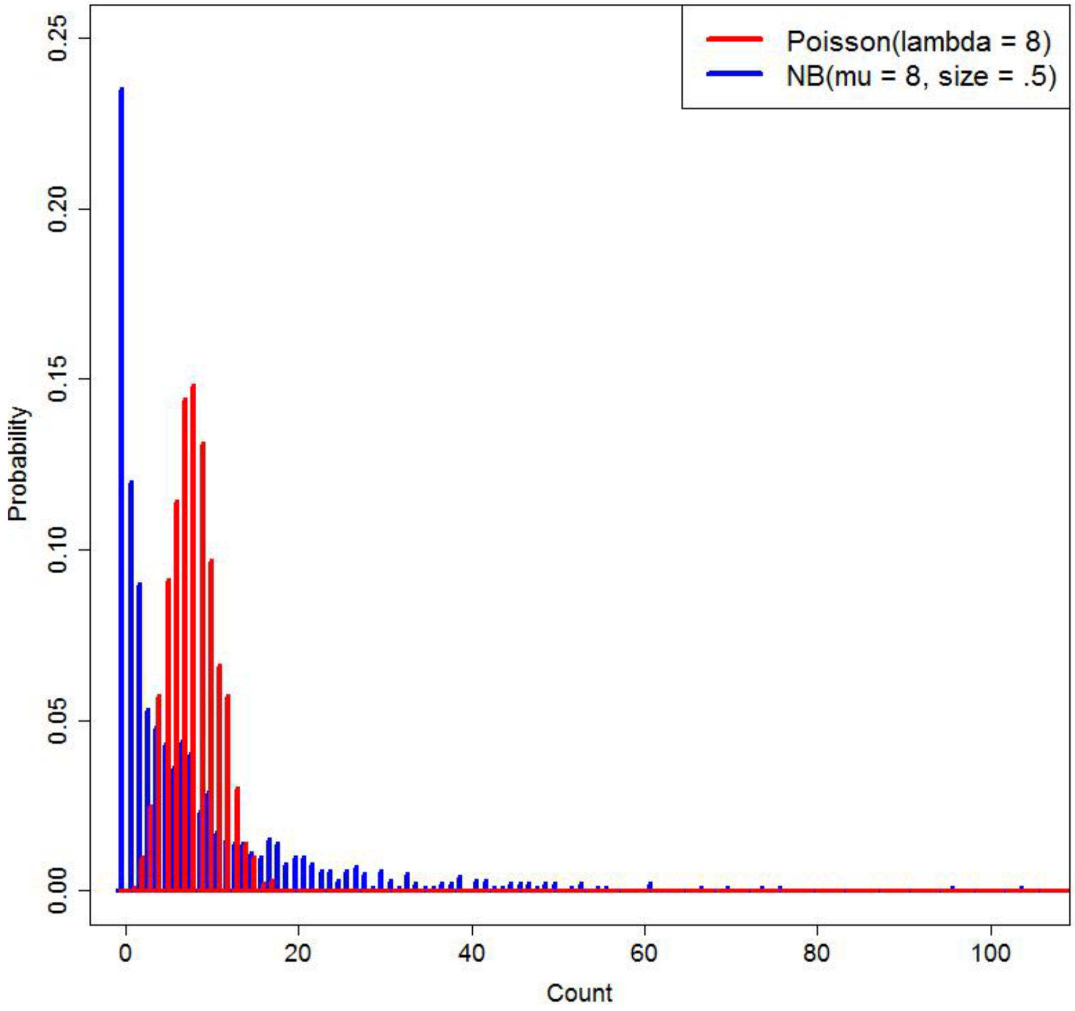
$$= \frac{\Gamma(y_{ijk} + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_{ijk}+1)} \left(\frac{1}{1+\mu_{ijk}\phi}\right)^{\phi^{-1}} \left(\frac{\mu_{ijk}}{\phi^{-1}+\mu_{ijk}}\right)^{y_{ijk}}$$

- $E(Y_{ijk}) = \mu_{ijk}$
- $\mathrm{Var}(Y_{ijk}) = \mu_{ijk} + \phi\mu_{ijk}^2$

- We may consider ϕ (common dispersion parameter) or ϕ_i (per-gene dispersion parameter)

Adapted from A. Rau

Negative Binomial vs Poisson distribution



Adapted from A. Rau

Negative Binomial estimation

- Many genes, very few biological samples – difficult to estimate overdispertion for each gene

- **Borrow information accross genes** (Empirical Bayes) !

edgeR modeling

Preprocessing: none – working on the gene counts matrix

GLM for each gene: $Y_{ij} \sim NB(\mu_{ij}, \alpha_i)$

- $\mu_{ij} = M_j p_{ij}$

  ⇒ M_j: library size

  ⇒ glm link: $\log_2\left(p_{ij}\right) = \sum_r x_{jr}\beta_{ir}$

- dispersion parameter α_i

  ⇒ adjusted profile likelihood (penalized likelihood) + Empirical Bayes

edgeR Likelihood Ratio Test

$\widehat{\beta}_{ir}$ estimated through ML

edgeR test:

- Fisher exact test adapted for overdispersion in pairwise comparisons for one factor
- Likelihood Ratio Test for more general designs

edgeR bibliography

- Robinson, Smyth. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881-7

- Robinson, Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008;9:321-32

- Robinson, McCarthy, Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-40

- McCarthy, Chen, Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40:4288-97

- `http://www.bioconductor.org/packages/release/bioc/html/edgeR.html`

voom-limma method

General idea slightly different than edgeR and DESeq2 :

`voom`:

1. normalize to log-counts per millions (remove library size effect)
2. estimate heteroscedasticity weights

`limma`:

3. Gaussian linear model
   (accounting for RNA-seq data heteroscedasticity through weighting)
4. Empirical Bayes t-test

voom weights computation — details

- y_{ij}: read count from sample j for gene i
- $L_j = \sum_{j=1}^{G} y_{ij}$: library size
- $y^*_{ij} = \log_2\left(10^6 \frac{0.5 + y_{ij}}{1 + L_j}\right)$: log-count per million
- X: experiment design matrix

1. $\widehat{\beta}_i \Rightarrow$ OLS estimate from the linear model $y^*_{ij} = \beta_i x_j + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i)$
2. $s_i = \sqrt{\sum_{i=1}^{n} \left(y^*_{ij} - \widehat{\beta}_i x_j\right)^2}$
3. average log-count value:
   $\tilde{y}_i = \frac{1}{n}\sum_{j=1}^{n} y^*_{ij} + \log_2\left(\prod_{j=1}^{n}(1 + \sum_{i=1}^{G} y_{ij})\right)^{1/n} - \log_2(10^6)$
4. $\widehat{f}(\cdot)$: predictor obtained from the LOWESS regression of $\sqrt{s_i}$ over $\tilde{y}_i$
5. $\widehat{w}_{ij} = \left[\widehat{f}\left(\widehat{\beta}_i x_j + \log_2(1 + \sum_{i=1}^{G} y_{ij}) - \log_2(10^6)\right)\right]^{-4}$

Voom-limma bibliography

- Law, Chen, Shi, Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29
- `http://www.bioconductor.org/packages/release/bioc/html/limma.html`

boris.hejblum@u-bordeaux.fr

BORDEAUX POPULATION HEALTH Centre de Recherche - U1219

ISPED SCHOOL OF PUBLIC HEALTH

université de BORDEAUX

Digital Public Health Graduate Program