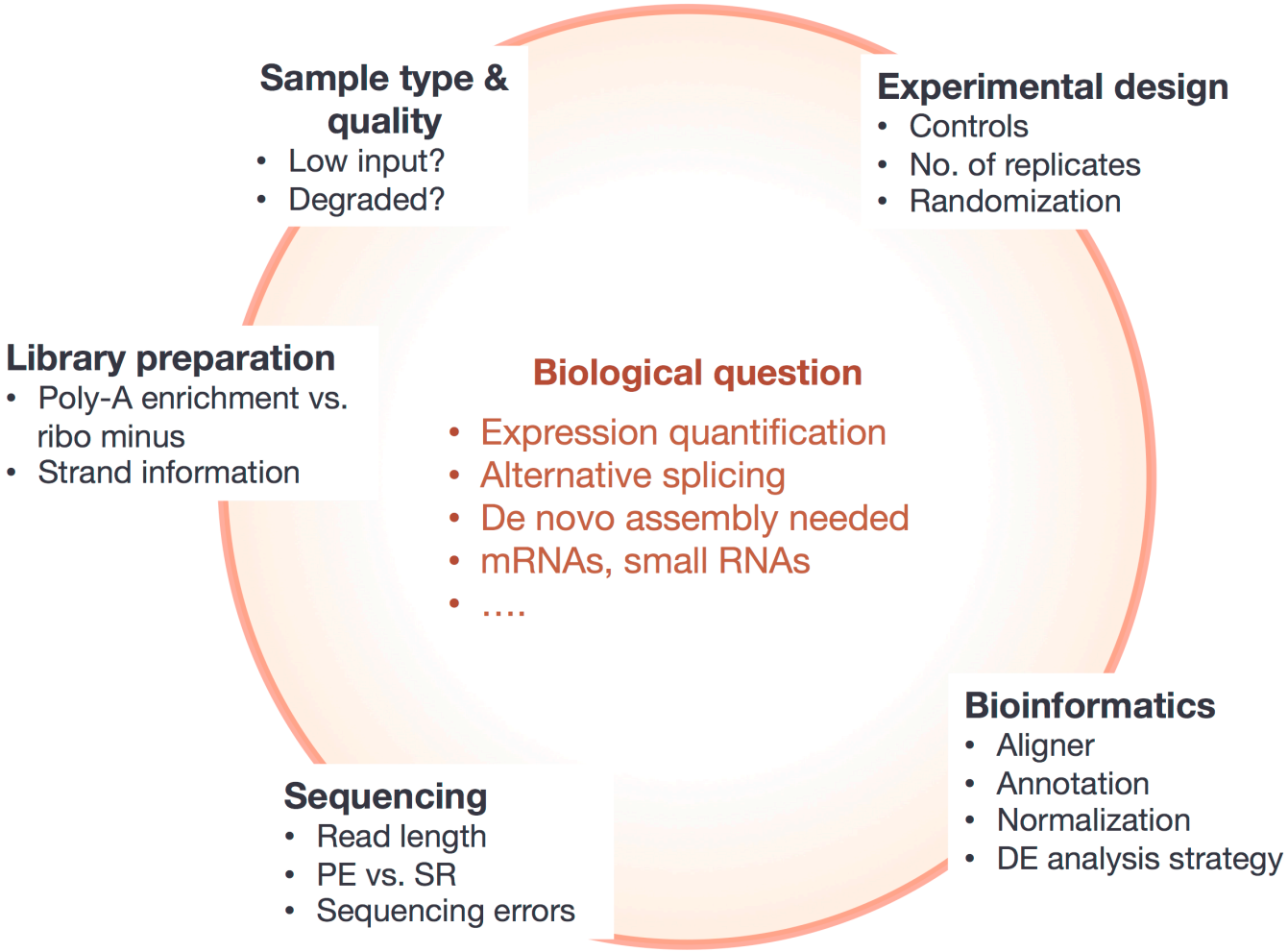
RNA-Seq analysis:
Differential expression and transcriptome assembly

Beibei Chen Ph.D
BICF
8/3/2016

Agenda

- Brief about RNA-seq
- Gene oriented analysis
  - Gene quantification
  - Gene differential analysis
- Transcript oriented analysis
  - Transcripts assembly and quantification
  - Transcripts differential expression
- Brief about experiment design

Everything's connected…



Everything's connected slide by Dündar et al. (2015)

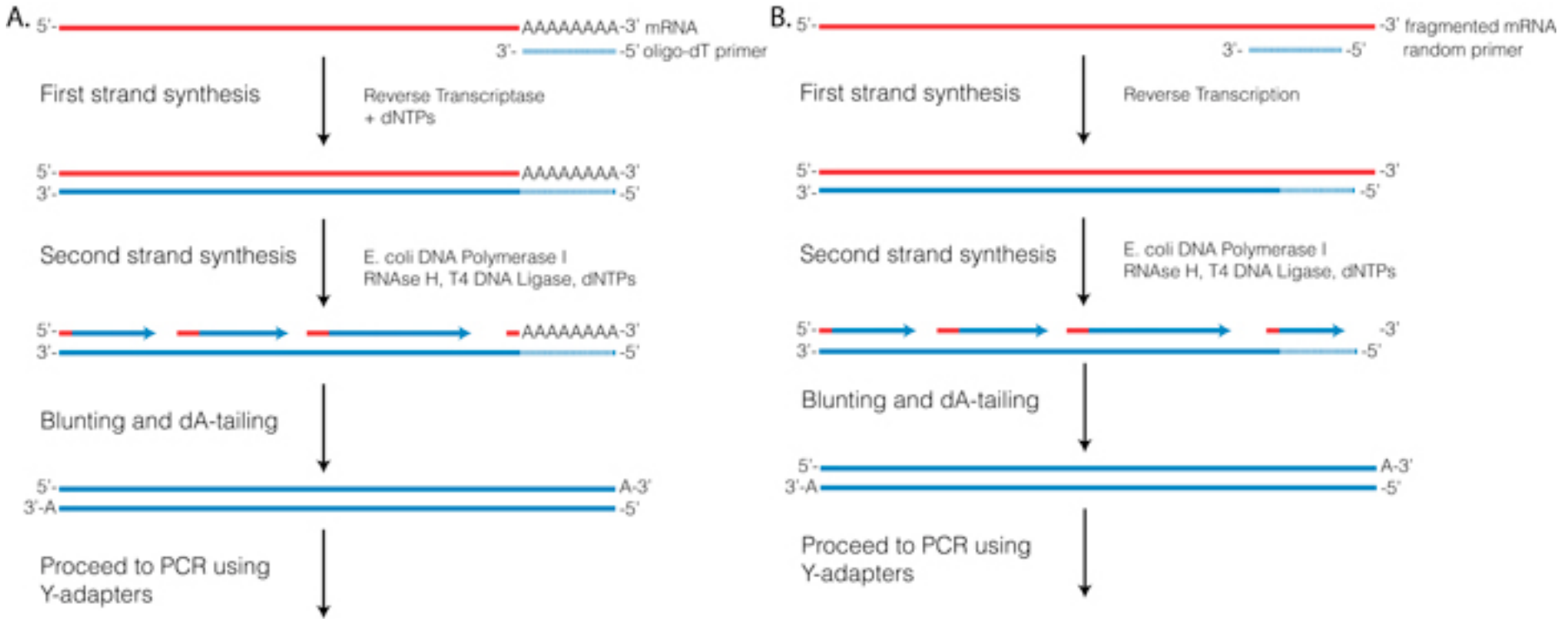
General RNA-seq workflow

- RNA purification
- Reverse transcription using Reverse Transcriptase (RT), which produces the first strand of cDNA
- Second strand synthesis using DNA polymerase
- Library preparation for sequencing

Two considerations

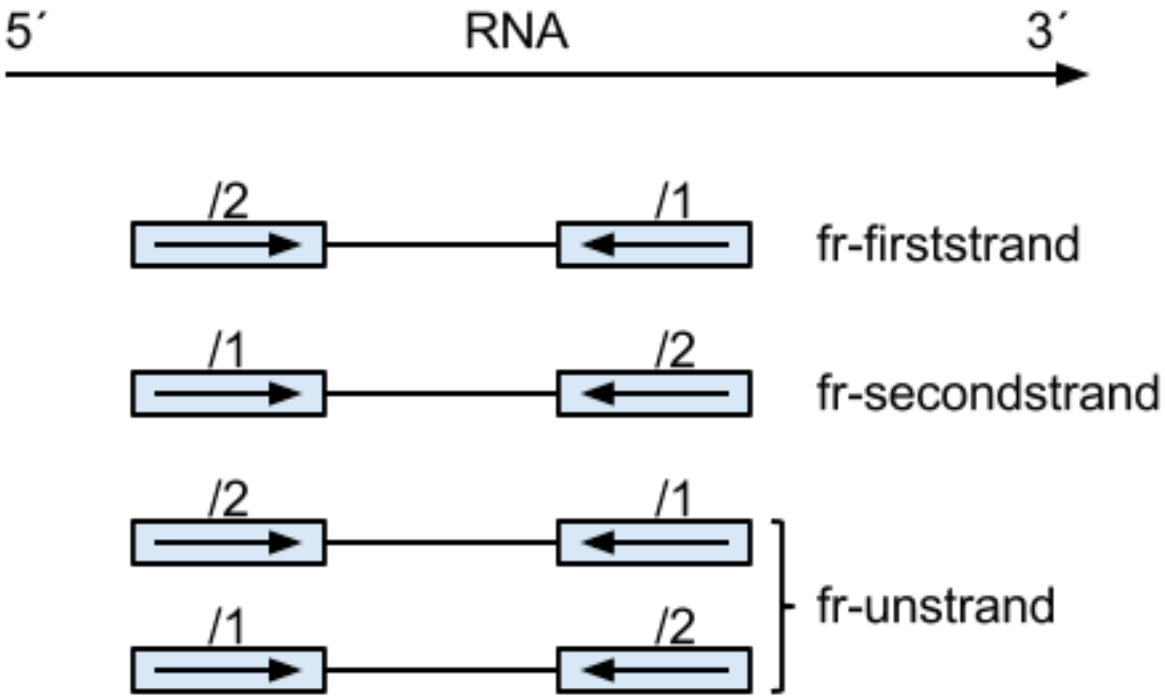
- Priming for the first cDNA strand synthesis
- Stranded versus Non-stranded libraries

Priming for the first strand synthesis



A: Use oligo-dT primer to (mostly) restrict cDNA synthesis to fully processed mRNAs
B: Use a mix of random oligonucleotides to prime RT at a multitude of internal sites irrespective of RNA type and maturation status:

Strand-specific RNA-seq



0 (fr-unstrand), 1 (fr-secondstrand) and 2 (fr-firststrand)

Strand-specific RNA-seq

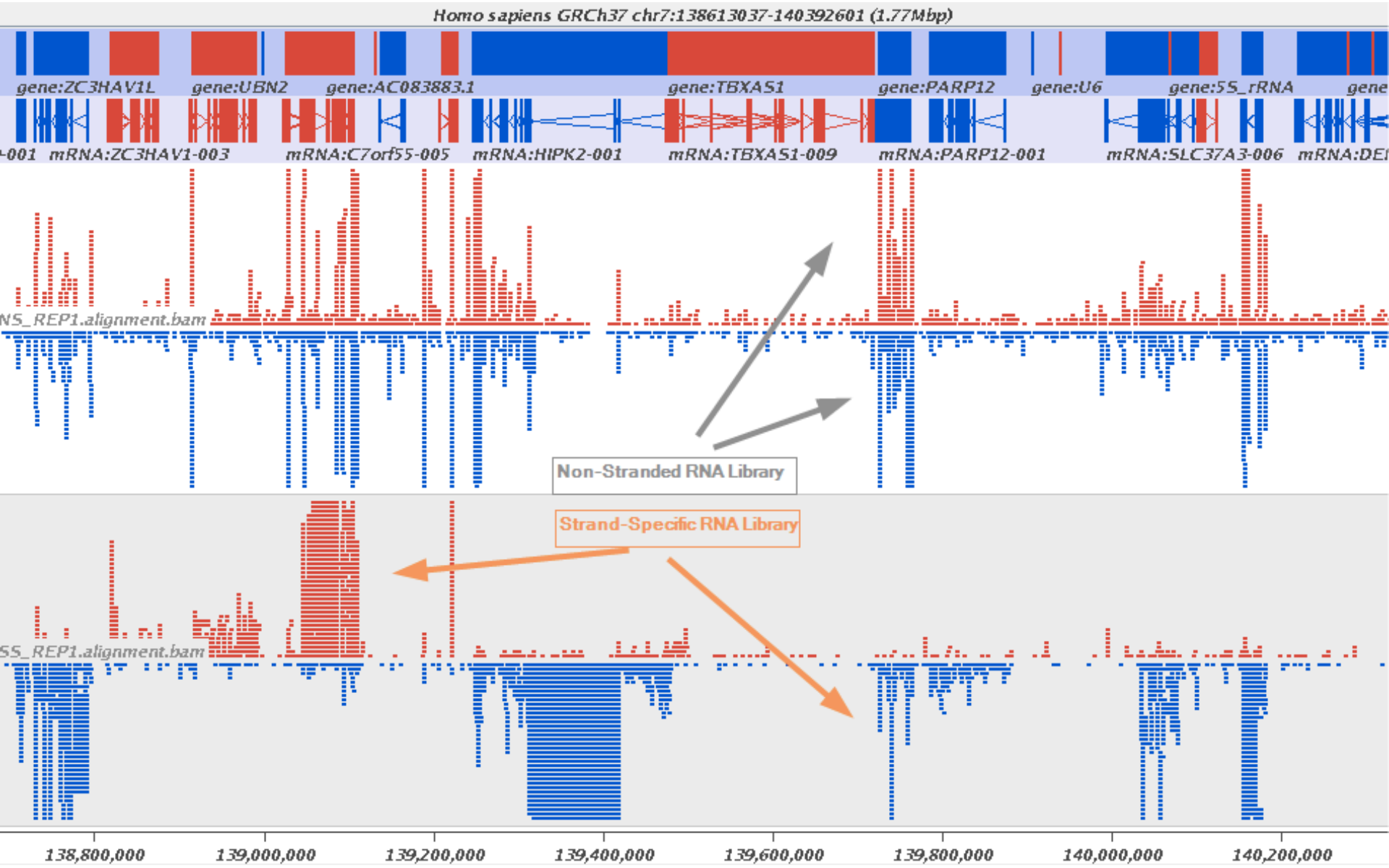
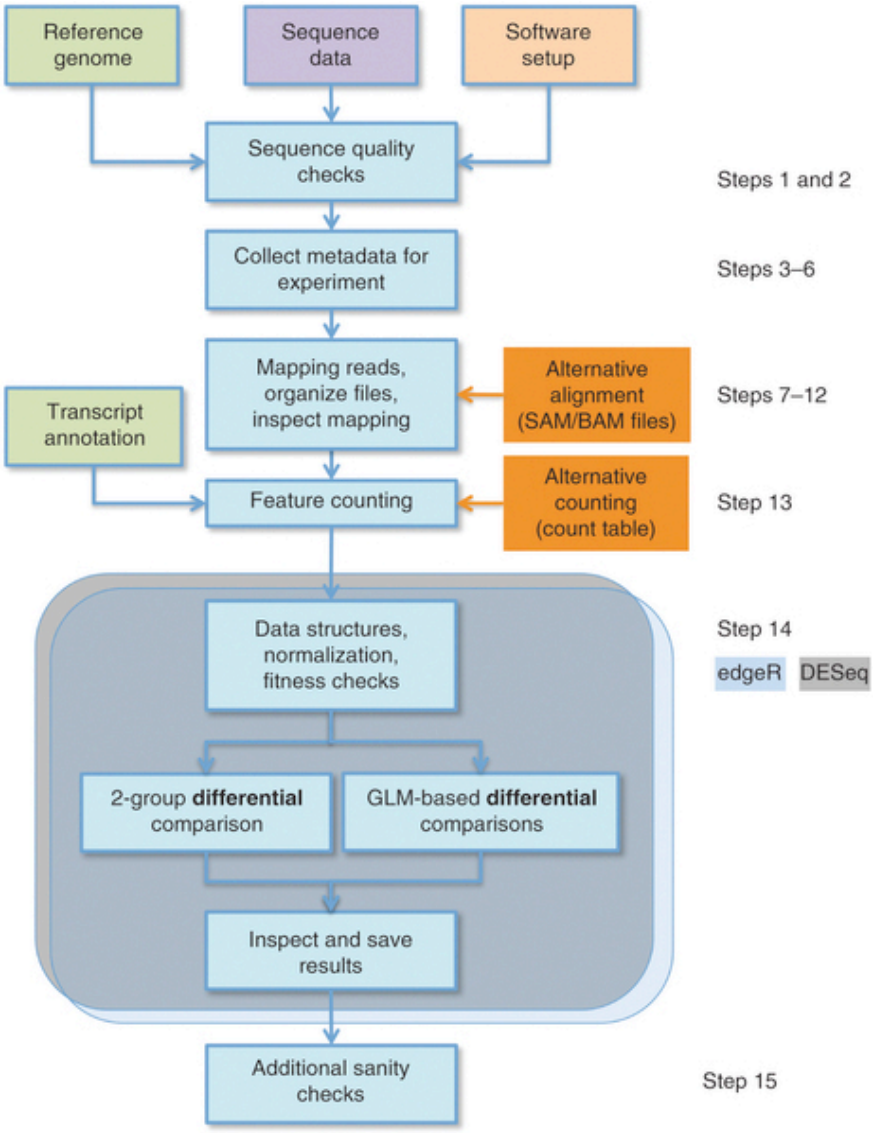


image from GATC Biotech

General RNA-seq analysis workflow



Nature Protocols 8,1765–1786 (2013)

Gene oriented analysis

- Gene quantification
- Gene differential analysis
  - Exploratory analysis
  - Differential expression analysis

Gene quantification

- In RNA-Seq, the abundance level of a gene is measured by the number of reads that map to that gene.

Things to prepare for counting

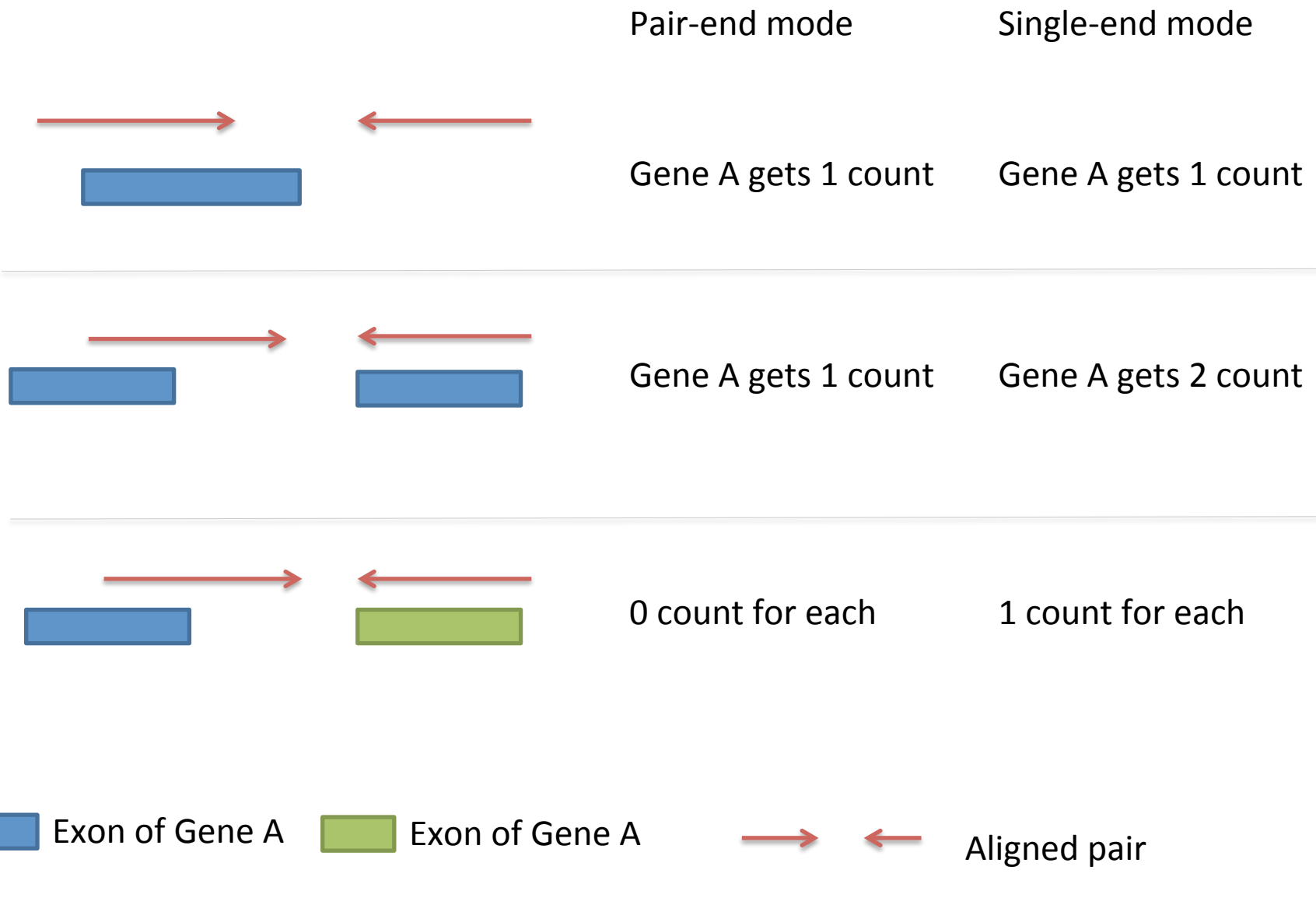
- featureCounts program
- Aligned BAM file
- GTF (*Gene Transfer Format*) file
  - Gencode: comprehensive annotation for human and mouse
  - igenome: refSeq annotation for many species
  - Ensembl: comprehensive annotation for many species (caution on the chromosome name!)

GTF format

column-number	content	values/format
1	chromosome name	chr{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,M} or GRC accession [a]
2	annotation source	{ENSEMBL,HAVANA}
3	feature type	{gene,transcript,exon,CDS,UTR,start_codon,stop_codon,Selenocysteine}
4	genomic start location	integer-value (1-based)
5	genomic end location	integer-value
6	score (not used)	.
7	genomic strand	{+,-}
8	genomic phase (for CDS features)	{0,1,2,.}
9	additional information as key-value pairs	see below

key name	value format
gene_id	ENSGXXXXXXXXXX.X [b,c]
transcript_id [d]	ENSTXXXXXXXXXX.X [b,c]
gene_type	**list of biotypes**
gene_status	{KNOWN, NOVEL, PUTATIVE}
gene_name	string
transcript_type [d]	**list of biotypes**
transcript_status [d]	{KNOWN, NOVEL, PUTATIVE}
transcript_name [d]	string
exon_number [e]	indicates the biological position of the exon in the transcript
exon_id [e]	ENSEXXXXXXXXXX [b]
level	1 (verified loci), 2 (manually annotated loci), 3 (automatically annotated loci)

How does featureCounts work



featureCounts Parameters

- Required:
  - Input file
  - -a file name: annotation file
  - -o string : output file
- Basic optional parameters for RNA-seq
  - -p: indicates the alignment is pair-end
  - -s int: strand-specific. 0 (unstranded), 1 (stranded) and 2 (reversely stranded)
  - -g gene_name (useful if you are using Ensembl GTF)

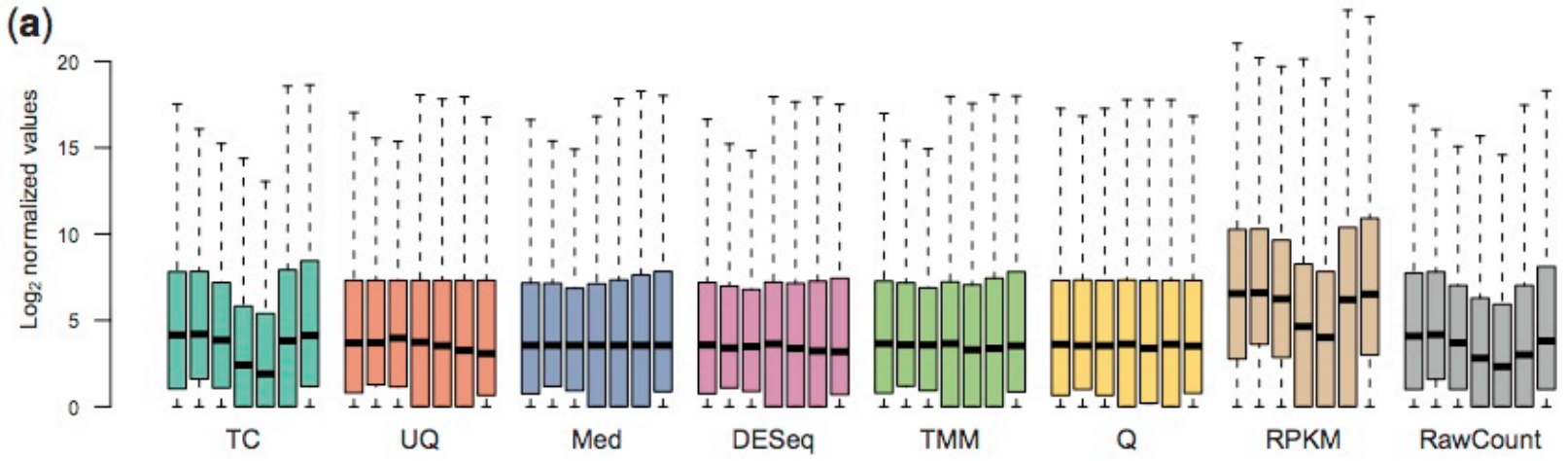
Differential expressed gene detection

- Normalization
- Explore your data

Why normalize

- To smooth out technical variations among the samples
  - Sequencing depth: genes have more reads in a deeper sequenced library
  - Gene length: longer genes are likely to have more reads than the shorter genes

Effects of different normalization methods



Assuming reads count distribution should be the same

- Total count (TC): Gene counts are divided by the total number of mapped reads
- Upper Quartile (UQ):Gene counts are divided by the upper quartile of counts
- Median (Med): Gene counts are divided by the median counts
- Quantile (Q): Matching distributions of gene counts across samples (limma)
- Reads Per Kilobase per Million mapped reads (RPKM): Re-scales gene counts
- to correct for differences in both library sizes and gene length

Assuming most gene are not differentially expressed

- DESeq
- Trimmed Mean of *M*-values (TMM): edgeR

doi:10.1093/bib/bbs046

Trimmed Mean M values (TMM)

- Applied in edgeR package
- Rationale:
  - TMM is the weighted mean of log ratios between this test and the reference.
  - TMM should be close to 0 according to the hypothesis of low DE. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis
- Reference sample can be assigned or the sample whose upper quartile is closest to the mean upper quartile is used

Trimmed Mean M values (TMM)

Gene-wise log-fold-changes $$M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$$

Absolute expression levels

$$A_g = \frac{1}{2}\log_2\left(Y_{gk}/N_k \bullet Y_{gk'}/N_{k'}\right) \text{ for } Y_{g\bullet} \neq 0$$

- By default, trim the M_g values by 30% and the A_g values by 5% (can be tailored in program)
- Weights are from the delta method on Binomial data
- Normalization factor for sample *k* using reference sample *r* is calculated as:

$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r} \text{ where } M_{gk}^r = \frac{\log_2\left(Y_{gk}/N_k\right)}{\log_2\left(Y_{gr}/N_r\right)} \text{ and } w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$Y_{gk}, Y_{gr} > 0.$

Median-of-ratios normalization

- Applied in DESeq and DESeq2
- Rationale:
  - Calculate the ratio of between a test and a pseudosample (For each gene, the geometric mean of all samples)
  - Non-DE genes should have similar read counts across samples, leading to a ratio of 1.
  - Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis

Median-of-ratios normalization

```
> log(raw_data)
        sample_1 sample_2 sample_3 sample_4
gene_1 2.564949 2.197225 2.772589 2.833213
gene_2 2.890372 2.639057 3.091042 3.637586
gene_3 4.605170 4.852030 4.905275 5.187386
gene_4 6.214608 6.445720 6.641182 6.917706
gene_5 6.919684 7.071573 7.328437 7.606885
gene_6 8.493105 8.696510 8.923458 9.210440
```

```
> loggeomeans <- rowMeans(log(raw_data))
> loggeomeans
  gene_1   gene_2   gene_3   gene_4   gene_5   gene_6
2.591994 3.064514 4.887465 6.554804 7.231645 8.830878
```

Pseudo sample

Get the median of log ratio of test comparing to pseudo sample:

```
> a <- apply(raw_data, 2, function(cnts) exp(median((log(cnts) - loggeomeans)[is.finite(loggeomeans)])))
> a
 sample_1  sample_2  sample_3  sample_4
0.7429489 0.8631042 1.0936042 1.4463899
```

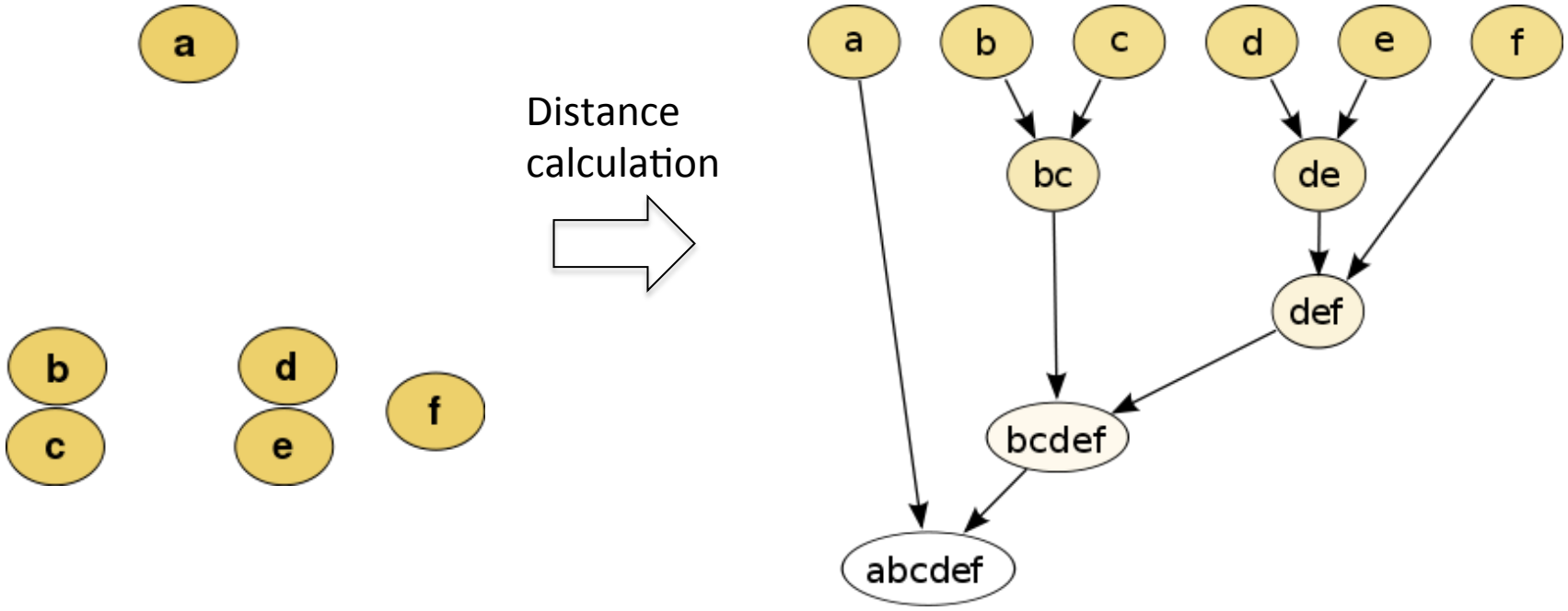
Data exploration

- Use log transformed normalized gene reads count
- Check if replicates from the same group are well concordance and grouped together
  - Hierarchy clustering
  - PCA plot

Hierarchy clustering

Raw data

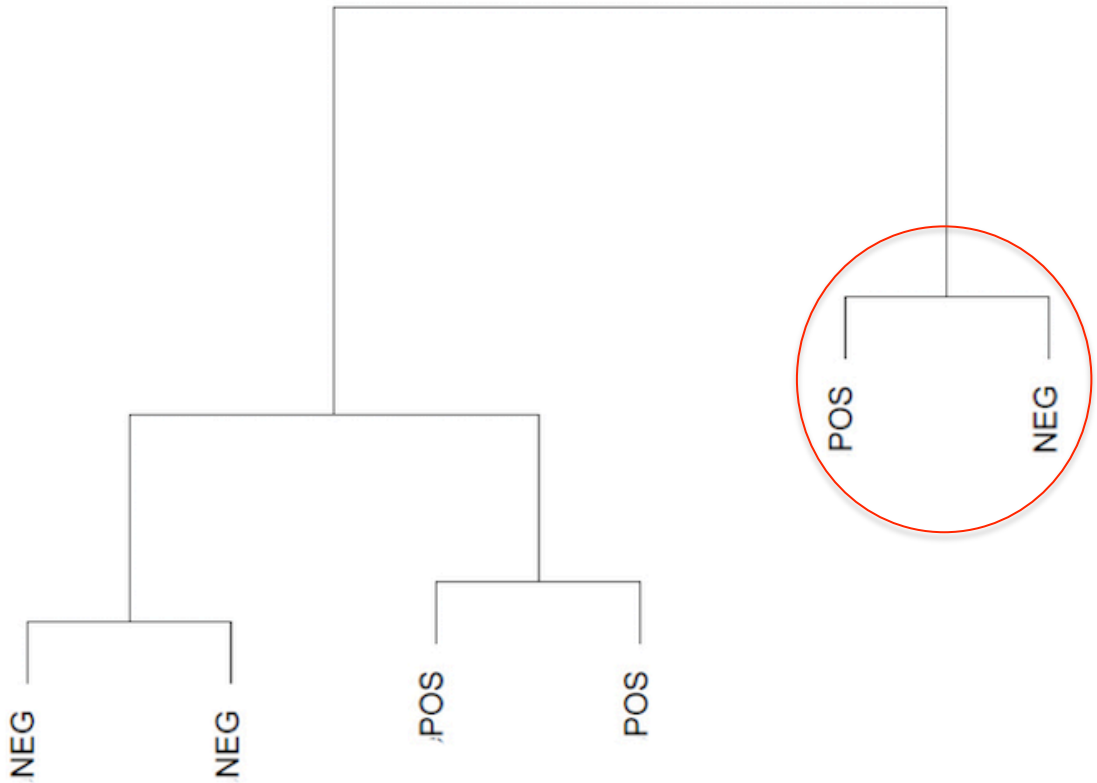
hierarchical clustering dendrogram



Euclidean distance: $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$

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

Hierarchy plot example

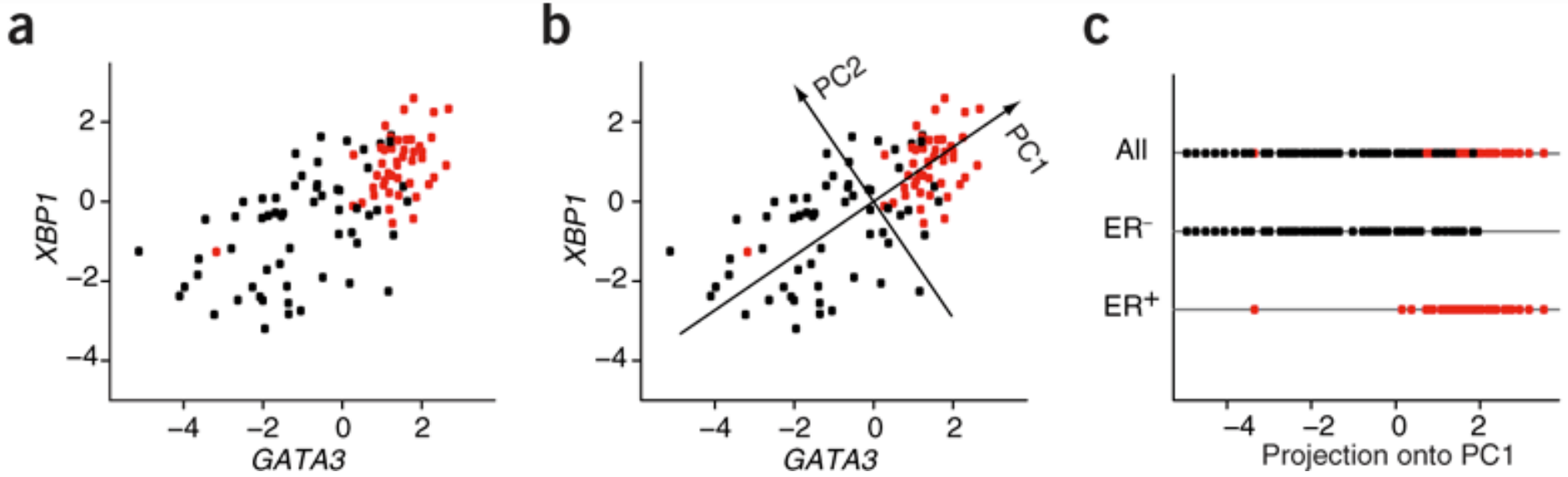


Samples prepared a year ago

Principal component analysis (PCA)

- A mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set
- It identifies directions, called principal components, along which the variation in the data is maximal
- By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables.

Simple example of PCA



Separate breast cancer ER+ from ER- : get profiles with only two genes

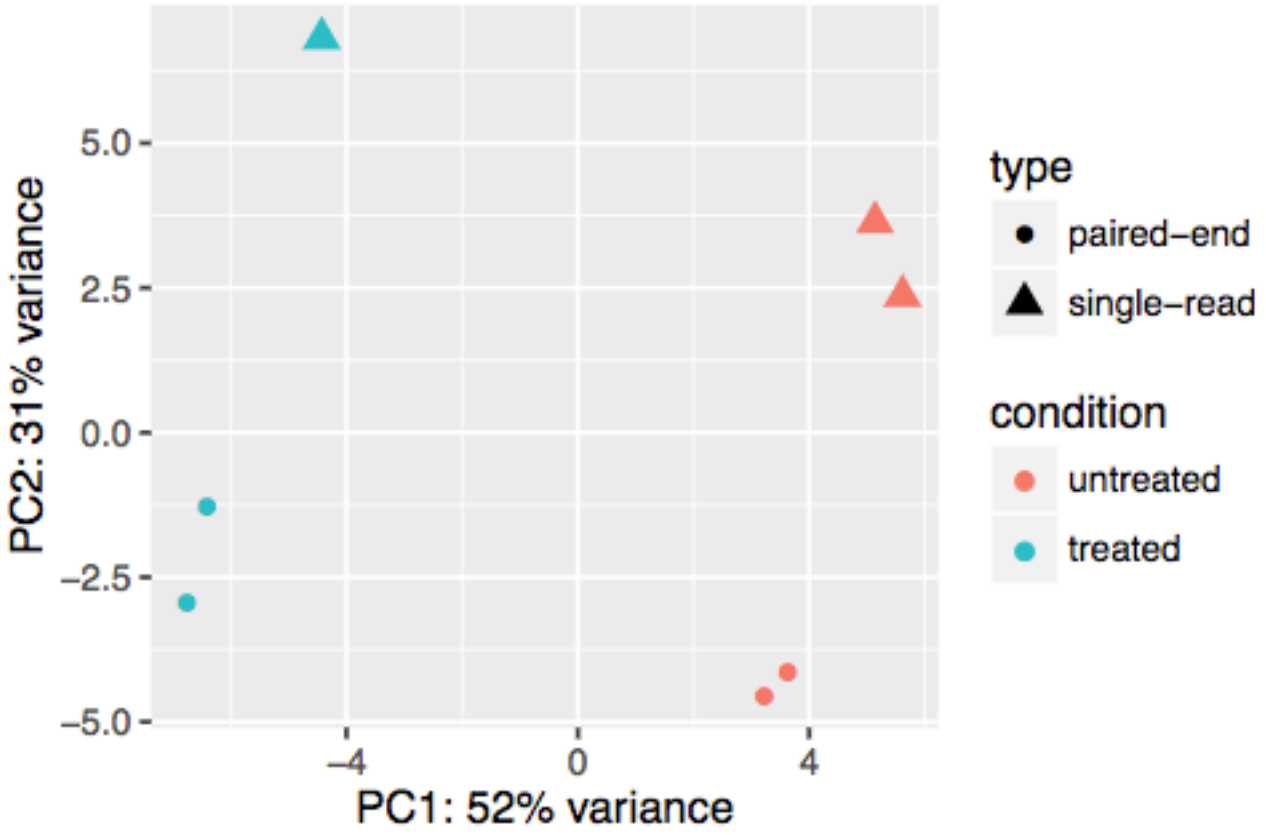
(a) Each dot represents a breast cancer sample plotted against its expression levels for two genes. (ER^+, red; ER^-, black).

(b) (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.

(c) (**c**) Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER^+, ER^- and all samples separately.

Nature Biotechnology 26, 303 - 304 (2008)

Real example



Test for differential expressed genes

- General liner model: negative binomial distribution
- Design matrix to define the linear model
  - ~condition
  - More complicated cases: edgeR user manual

Test for differential expressed genes

Two sample groups, treatment and control.

Assumption:

- Count value for a gene in sample j is generated by NB distribution with mean $s_j \mu_j$ and dispersion α.

Null hypothesis:

- All samples have the same μ_j.

Alternative hypothesis:

- Mean is the same only within groups:
  $\log \mu_j = \beta_0 + x_j \beta_T$

$x_j = 0$ for if j is control sample
$x_j = 1$ for if j is treatment sample

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/2_Tuesday/lectures/DESeq2-Anders.pdf

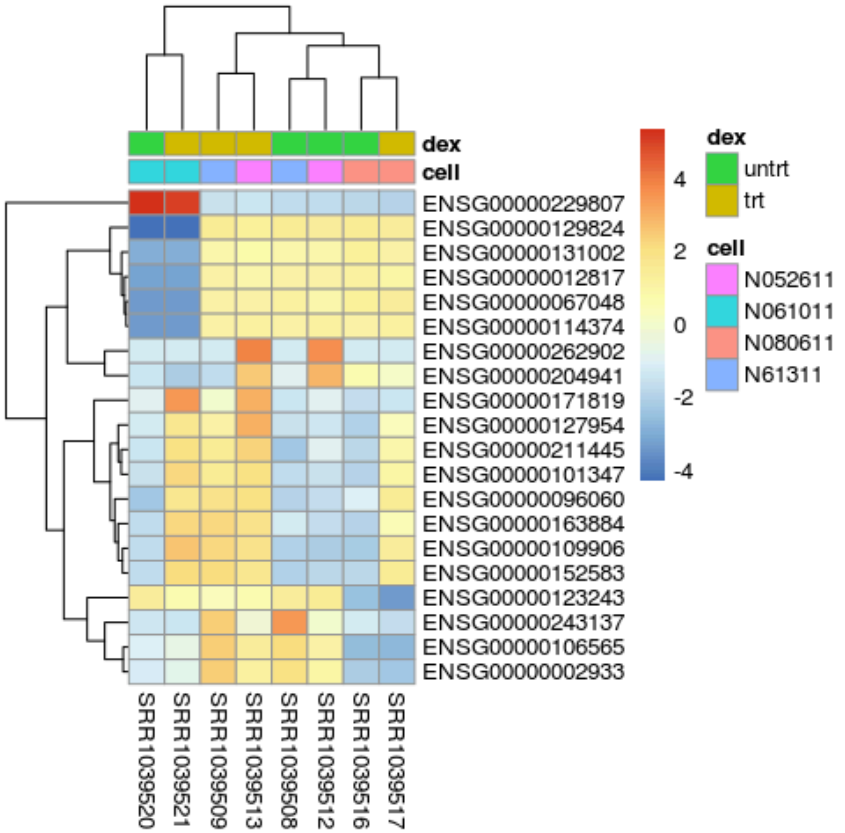
Test for differential expressed genes

- After GLMs are fit for each gene
- Wald test: whether each model treatment coefficient differs significantly from zero
- Multiple testing adjust
  - For a genome with 10,000 gene, using p<=0.05 as cutoff, there are 500 genes are significant by chance
  - BH method

https://bioconductor.org/help/course-materials/2010/BioC2010/2010-07-29_DESeq_Bioc-Meeting_Seattle.pdf

Define differential expressed genes

FDR and/or logFC cutoff



INGENUITY PATHWAY ANALYSIS

GSEA
Gene Set Enrichment Analysis

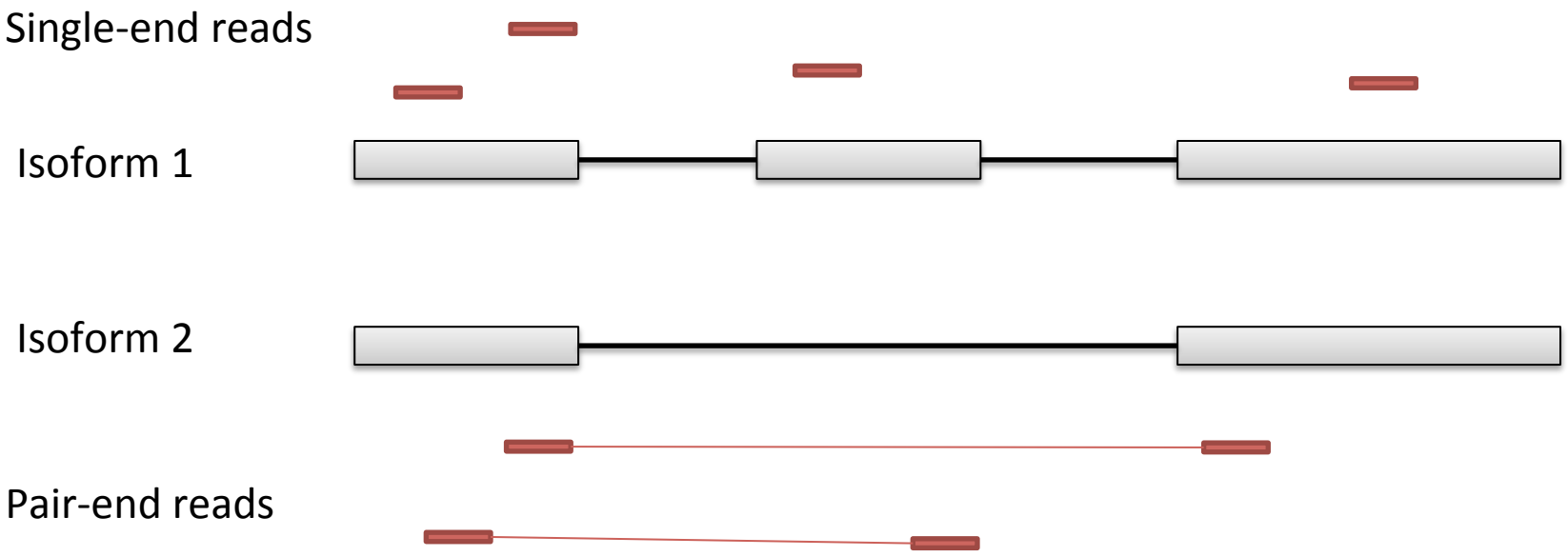
PANTHER
Classification System

DAVID BIOINFORMATICS DATABASE

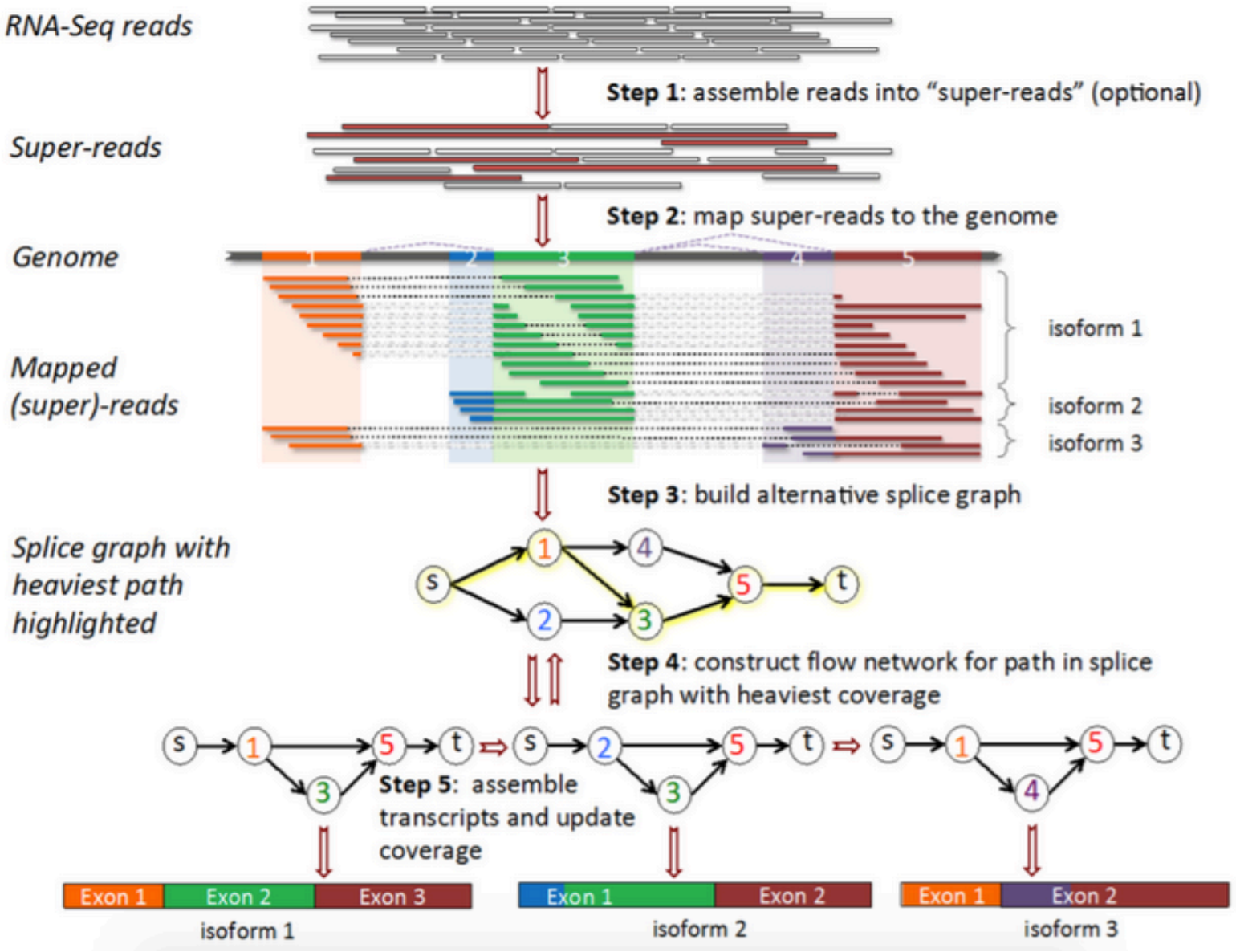
Transcript oriented analysis

- Transcripts assembly and quantification
  - Stringtie
- Transcripts differential expression
  - Ballgown

Pair-end and single-endsequencing

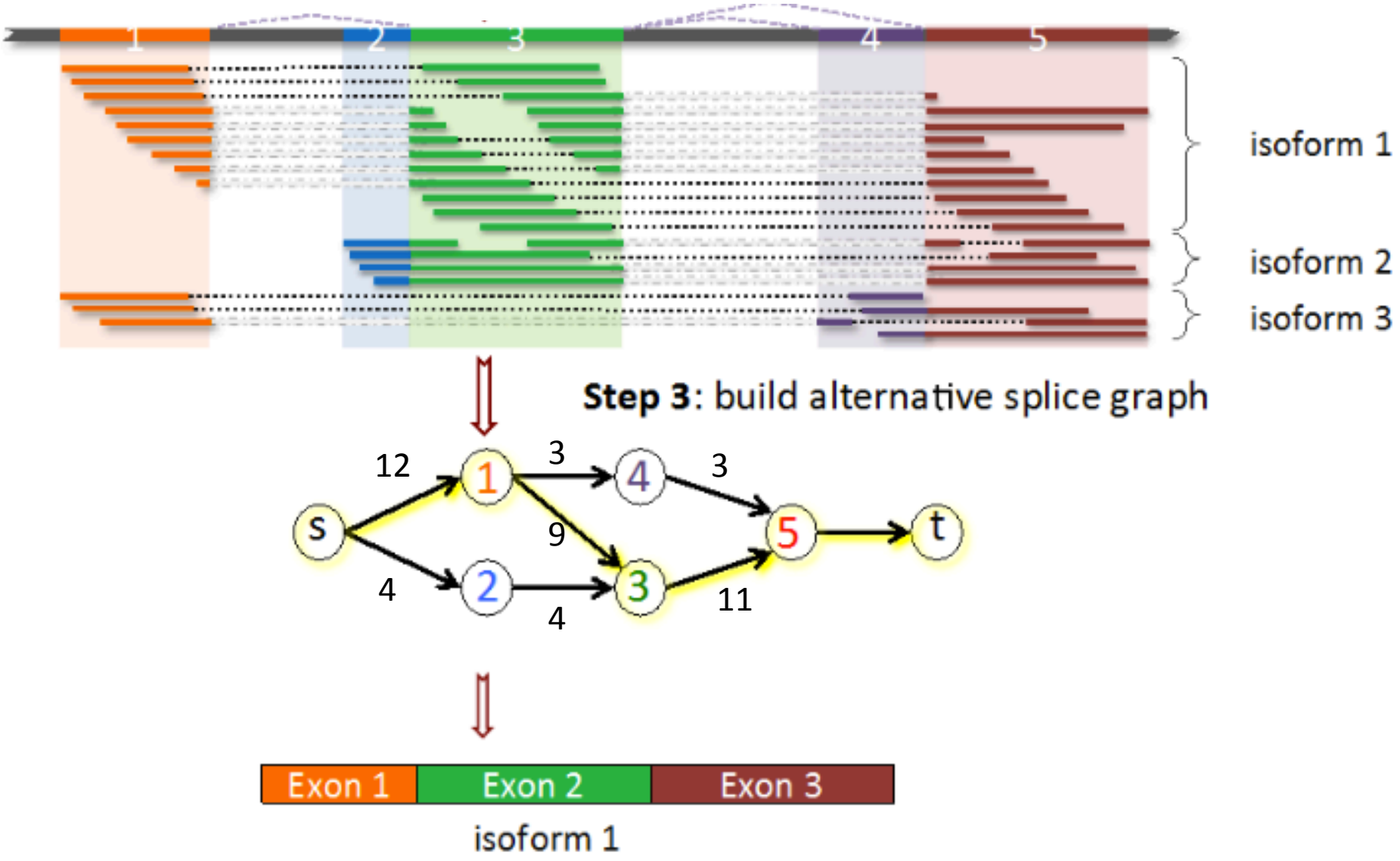


StringTie workflow

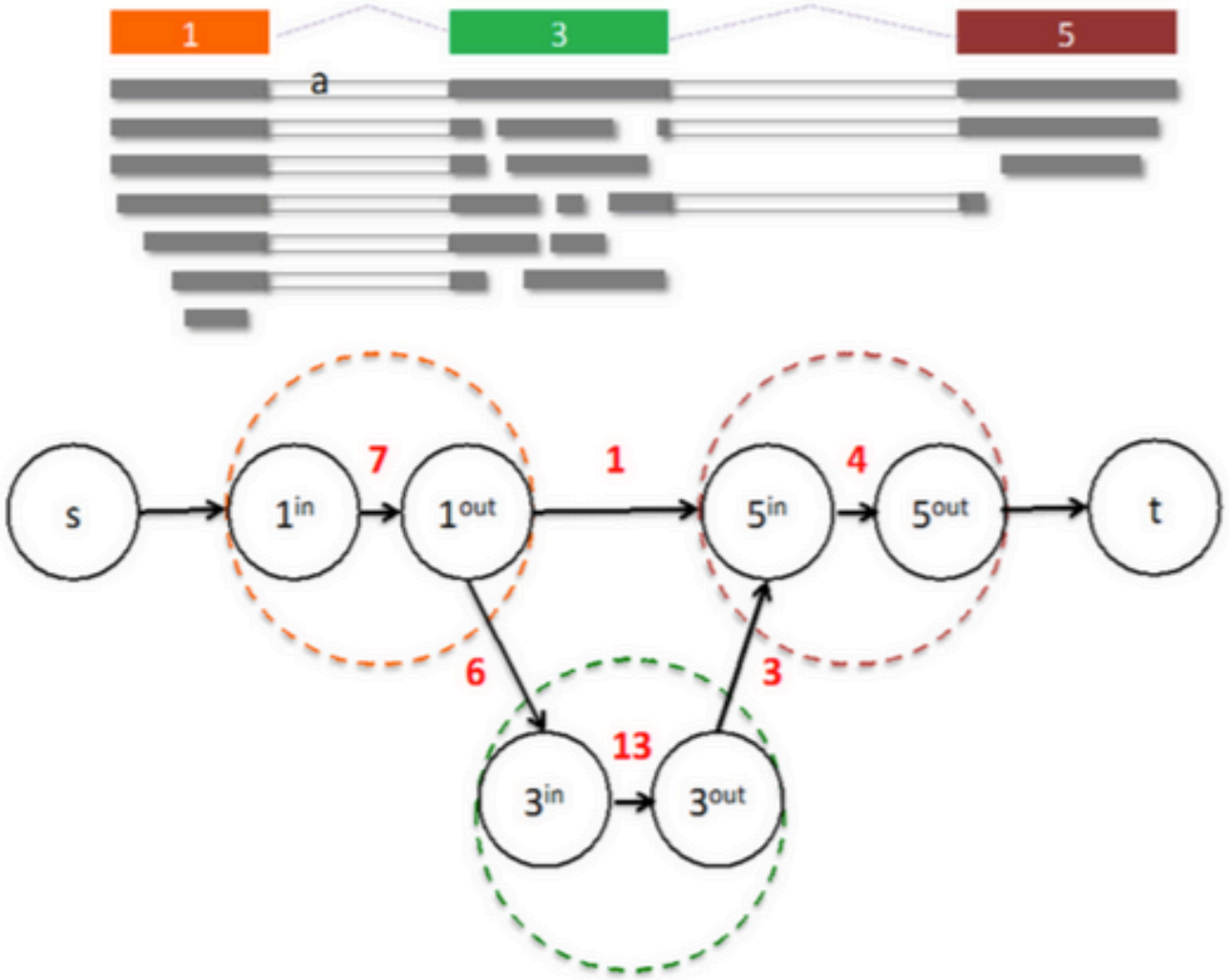


Nature Biotechnology 33, 290–295 (2015)

Build alternative splicing graph



StringTie estimates transcript levels



Nature Biotechnology 33, 290–295 (2015)

Ballgown

- Bridged the gap of transcripts assembly and differential expression analysis
  - RSEM + edgeR
- Statistical methods are conceptual similar to limma
- Super fast

Ballgown algorithm

$$h(FPKM(\hat{t}_k, z)) = \alpha_k + \sum_{p=1}^{P} \beta_{pk} X_{zp} + \varepsilon_{zk} \quad (1)$$

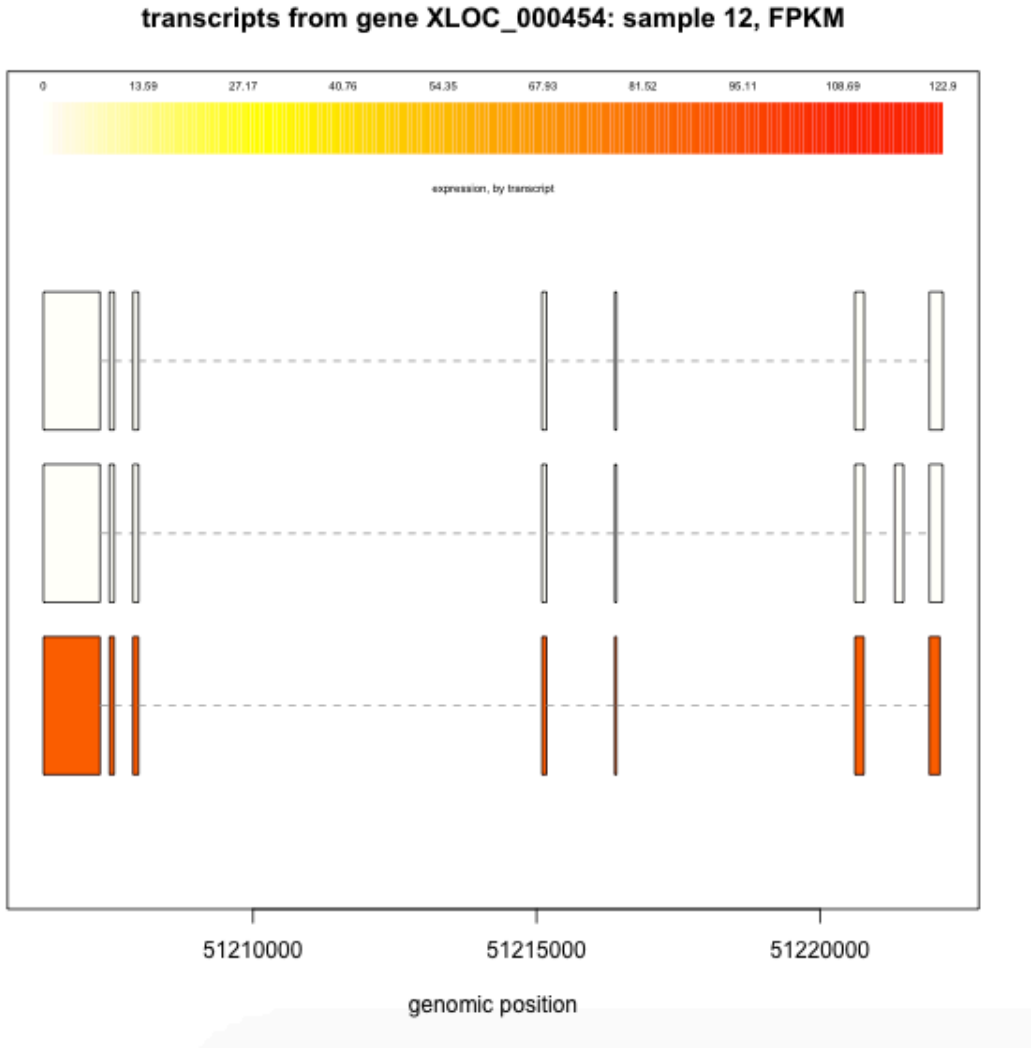
where:

- $FPKM(\hat{t}_k, z)$ is the FPKM expression measurement for transcript k for sample z
- h is a transformation [2] to reduce the impact of mean-variance relationships observed in the counts [1]. For example, the transformation $h(\cdot) = \log_2(\cdot + 1)$ is commonly applied in the analysis of sequence-count data [8].
- α_k represents the baseline expression for transcript k
- X_{zp} represents covariate p for sample z. These covariates differ by experiment type. X_{z1} generally represents a library size adjustment for sample z; `ballgown`'s default for this value is $X_{1z} = \text{median}_k\{FPKM(\hat{t}_k, z)\}$
- β_{pk} quantifies the association of covariate p on te expression of transcript k
- ε represents residual measurement error

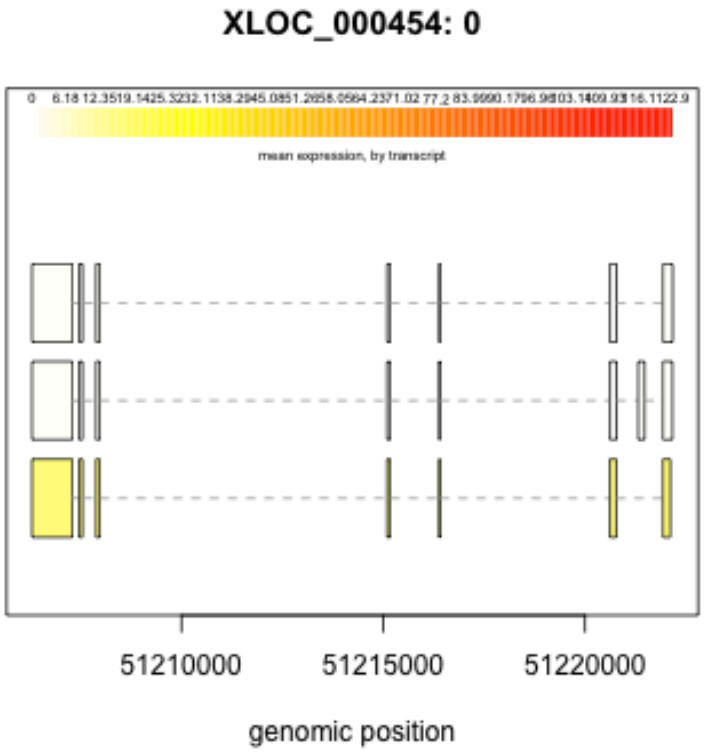
After model fitting, F test is used to test for differential expressed transcripts.

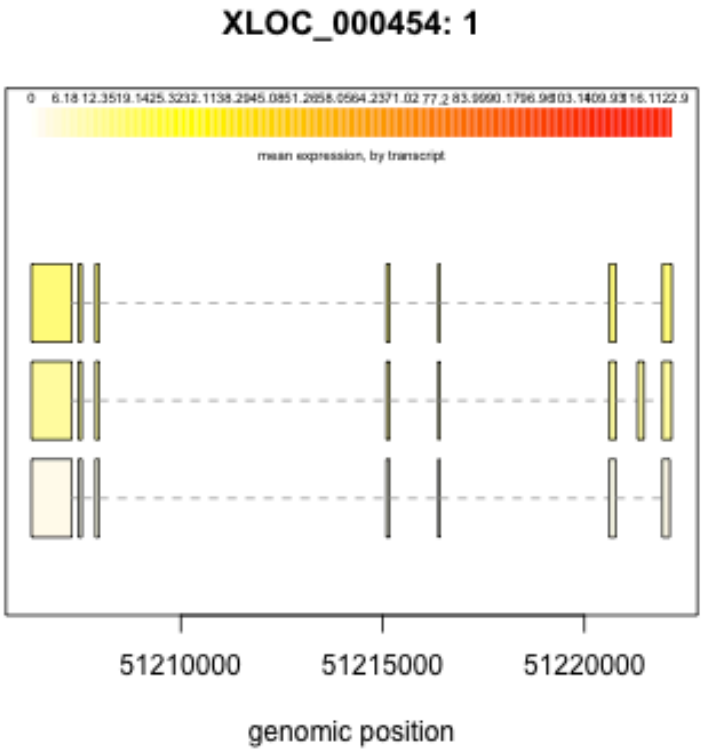
http://dx.doi.org/10.1101/003665

Ballgown visualization



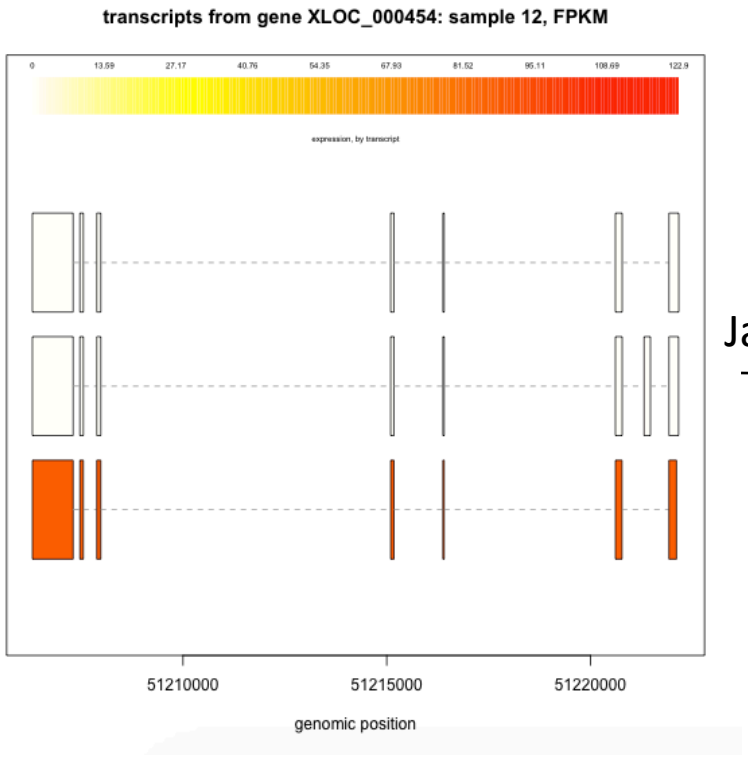
Ballgown visualization



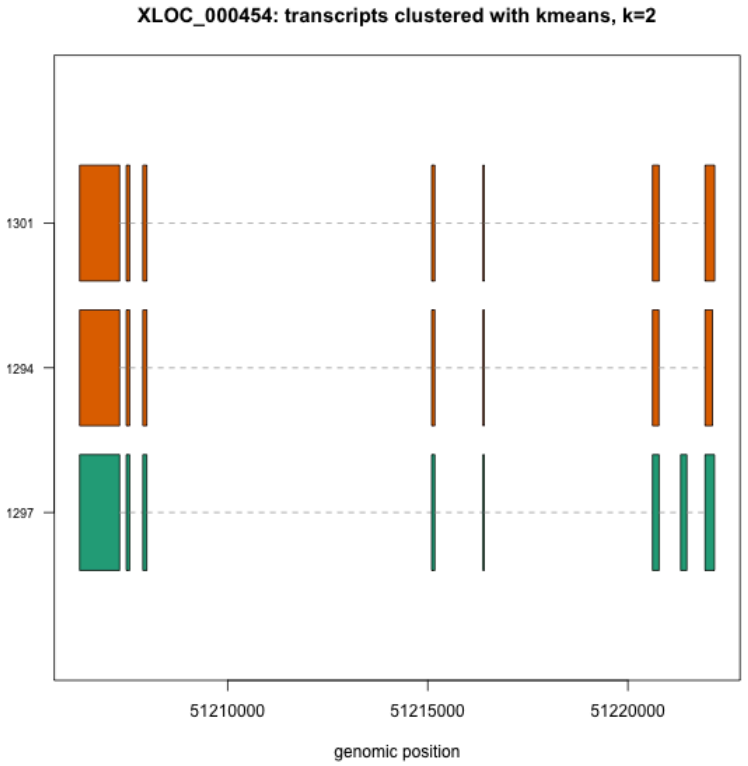


Ballgown: transcripts clustering

- Expression estimates are unreliable for very similar transcripts of a same gene



Jaccard distance

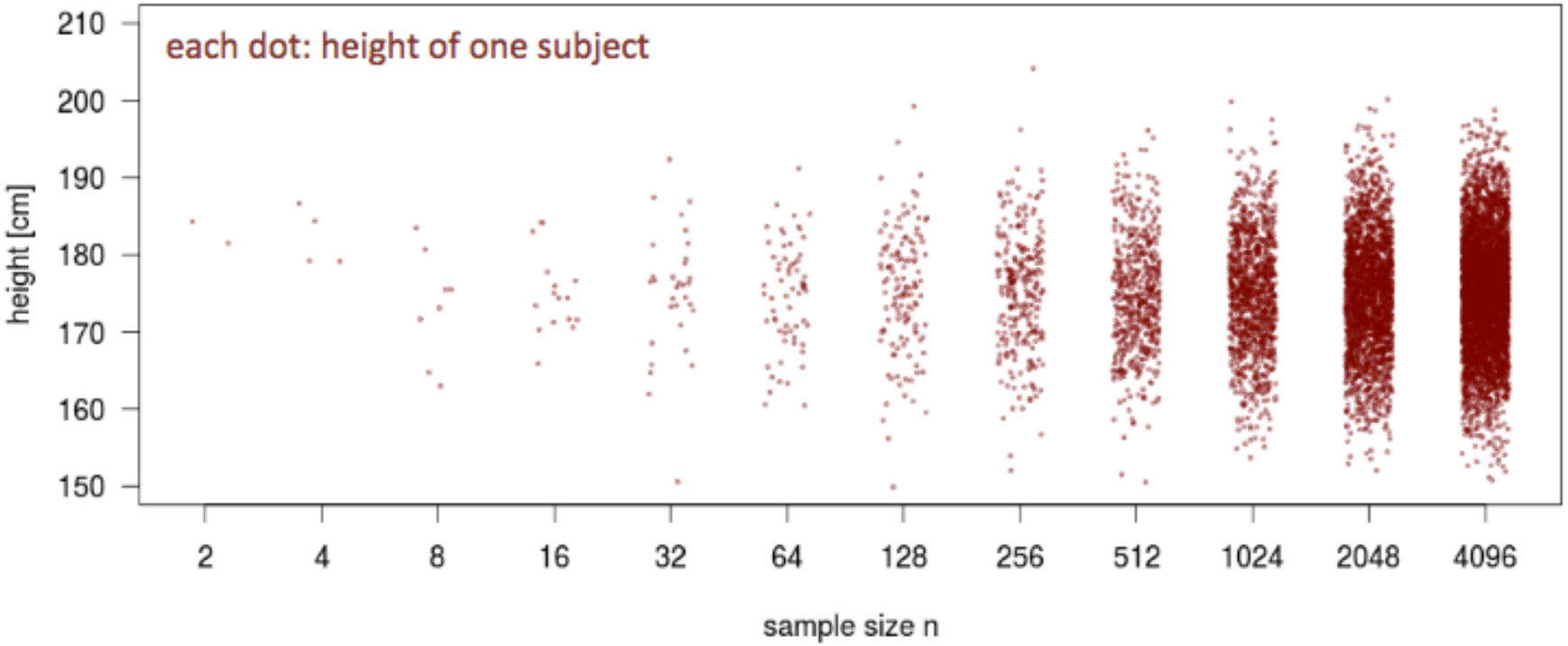


Brief about experiment design

- Why and how many replicates
  - In this section, contents are from Dr. Simon Anders's talk in CSAMA2014
- Pair-end or single-end
- Sequence depth

Why replicates

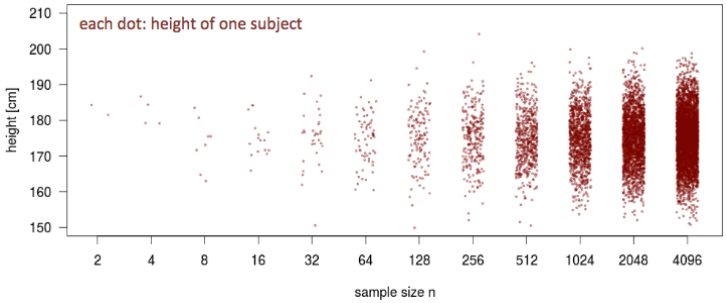
What is the average height of a man in Germany?

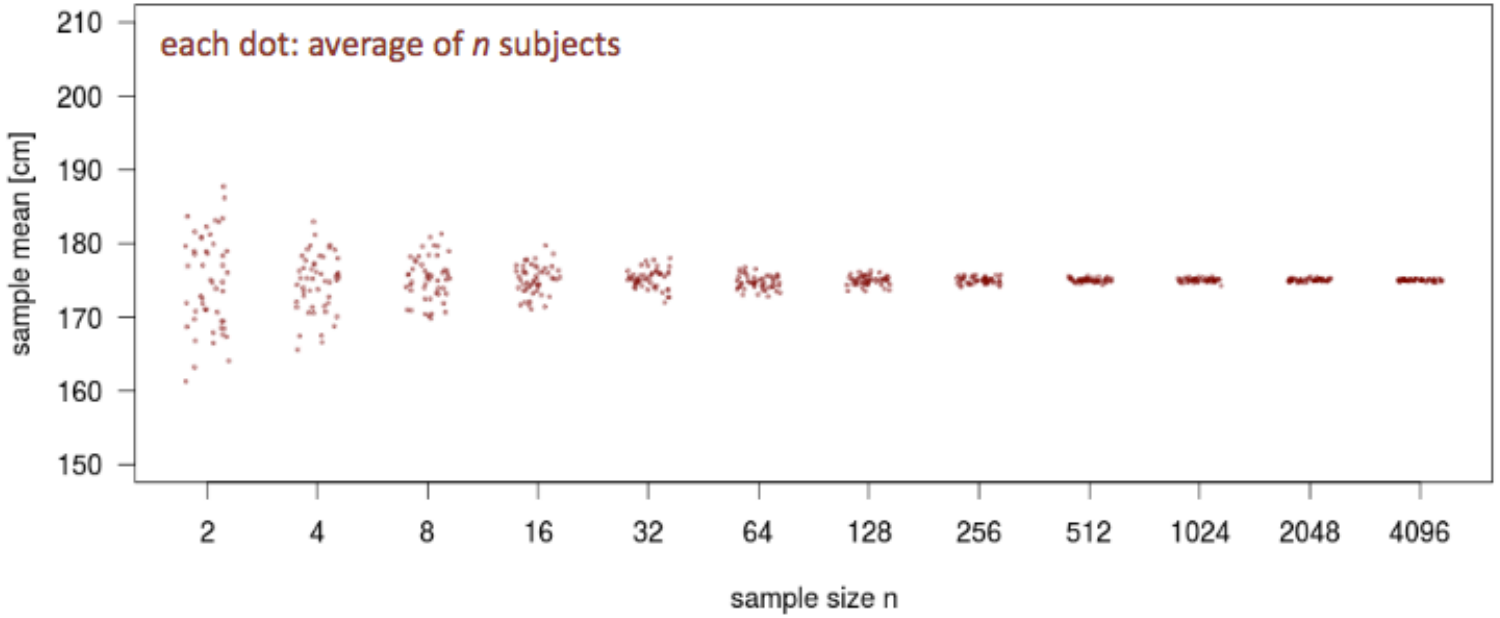


https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Why replicates

What is the average height of a man in Germany?



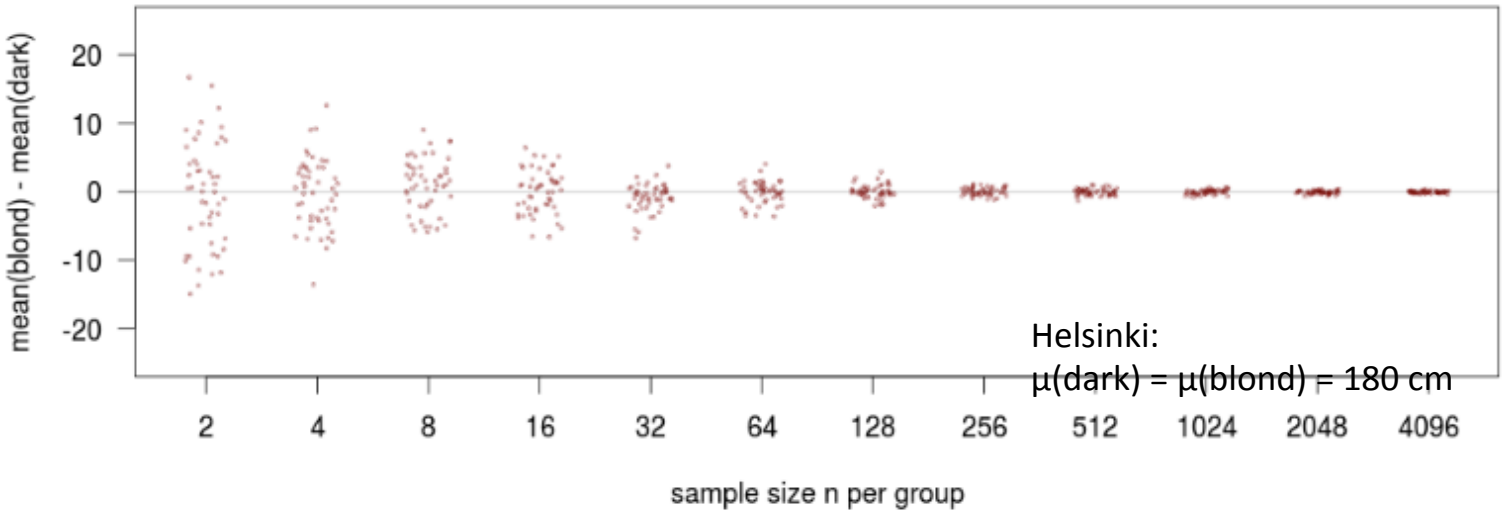


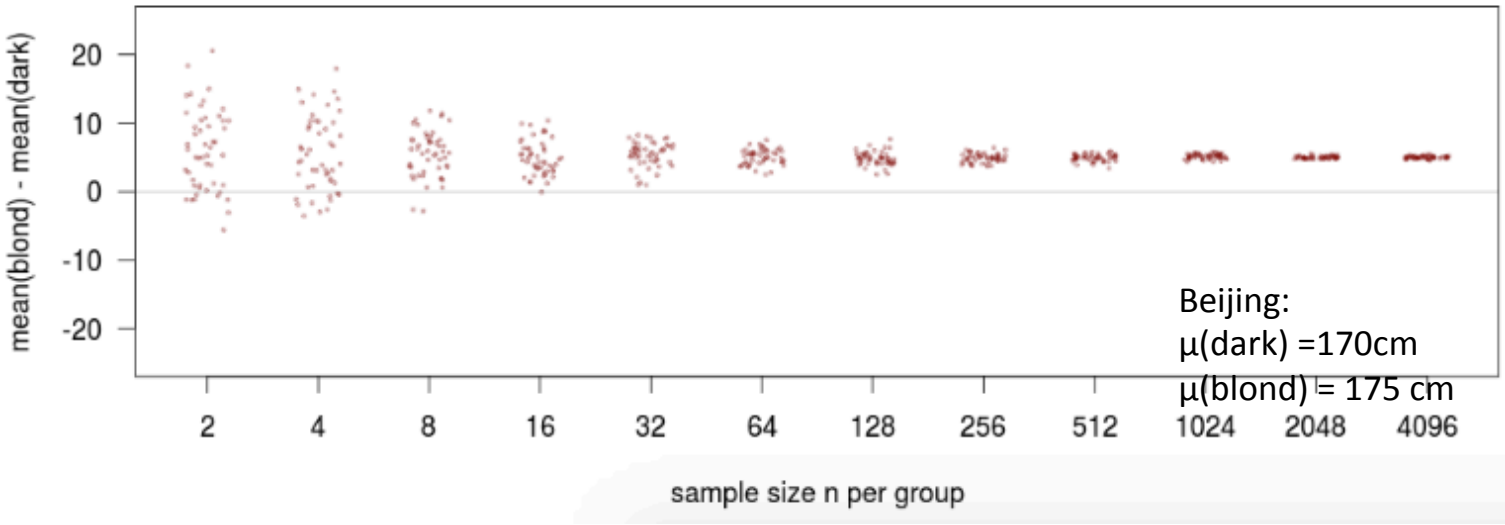
$$\text{standard error of mean} = \frac{\text{standard deviation of observations}}{\sqrt{\text{sample size}}}$$

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Why replicates

Is height correlated with hair color?





https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

How many replicates

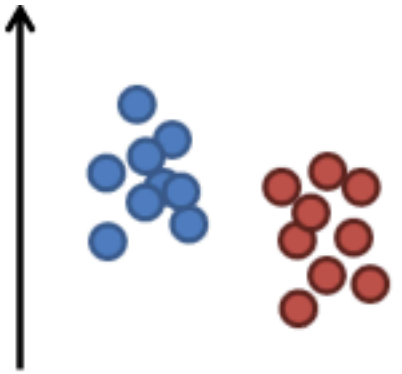
- Your power to detect an effect depends on
  - Effect size (difference between group means)
  - Within group variance
  - Sample size

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Two extremes

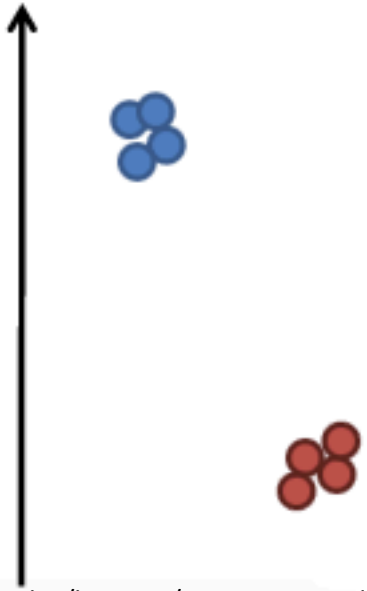
- effect size ≪ withing-group SD

  many replicates to get precise mean estimates



- effect size ≫ within-group SD

  few replicates sufficient, just to verify that SD is small



https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Purpose of replicates

- More replicates allow for more precise estimates of effect sizes
- Replicates allow for estimating how precise our effect size estimates are
- Enough replicates allow for proper randomization. (For causation)
- Enough replicates reduce the need for assumptions on distribution
- Replicates allows to spot outliers

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

How many replicates?

- "Estimating a SD from just two subjects is pointless"
- Controlled experiment: variation ≪ effect size
  - Single measure: 5~15 per group
  - RNA-seq etc: 3 per group
- Study with Study with strong inter subject variation
  - Dozens to hundreds of subjects!

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Respond to common objections by Dr. Andres

- "I cannot afford replicates"
  - Use multiplexing: 5 samples sequenced to 20M reads each offer more power than 2 samples sequenced to 50M reads
- "I know I need at least 50 samples, but I cannot get hold of more than 10. So I use what I have."
  - Performing underpowered experiments is a waste of time
- "I know that my within group variability is much smaller than the effect size."
  - Then prove it.

https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/5_Friday/lectures/ExpDesign_Anders.pdf

Pair-end or single-end and depth

- Gene/transcript quantification
  - pair-end
  - 20M is minimum. For transcript oriented experiment, try to get deeper
- CLIP/RIP/smallRNA sequence
  - You can use stranded single-end but pair-end improves the unique mapping ratio and reduces duplication rate
- ChIP-seq
  - ENCODE's protocol is single-end, so…

Questions?