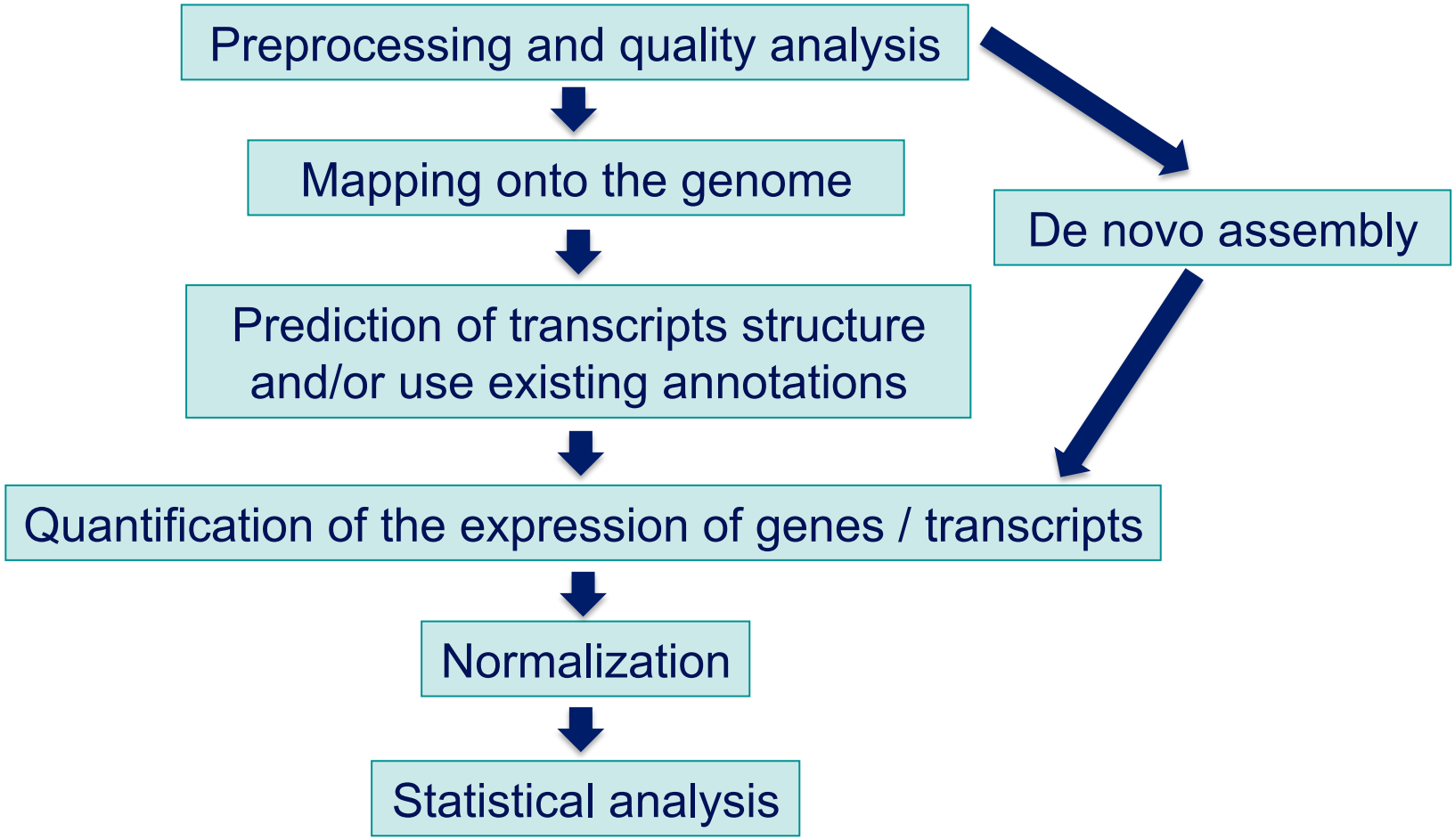
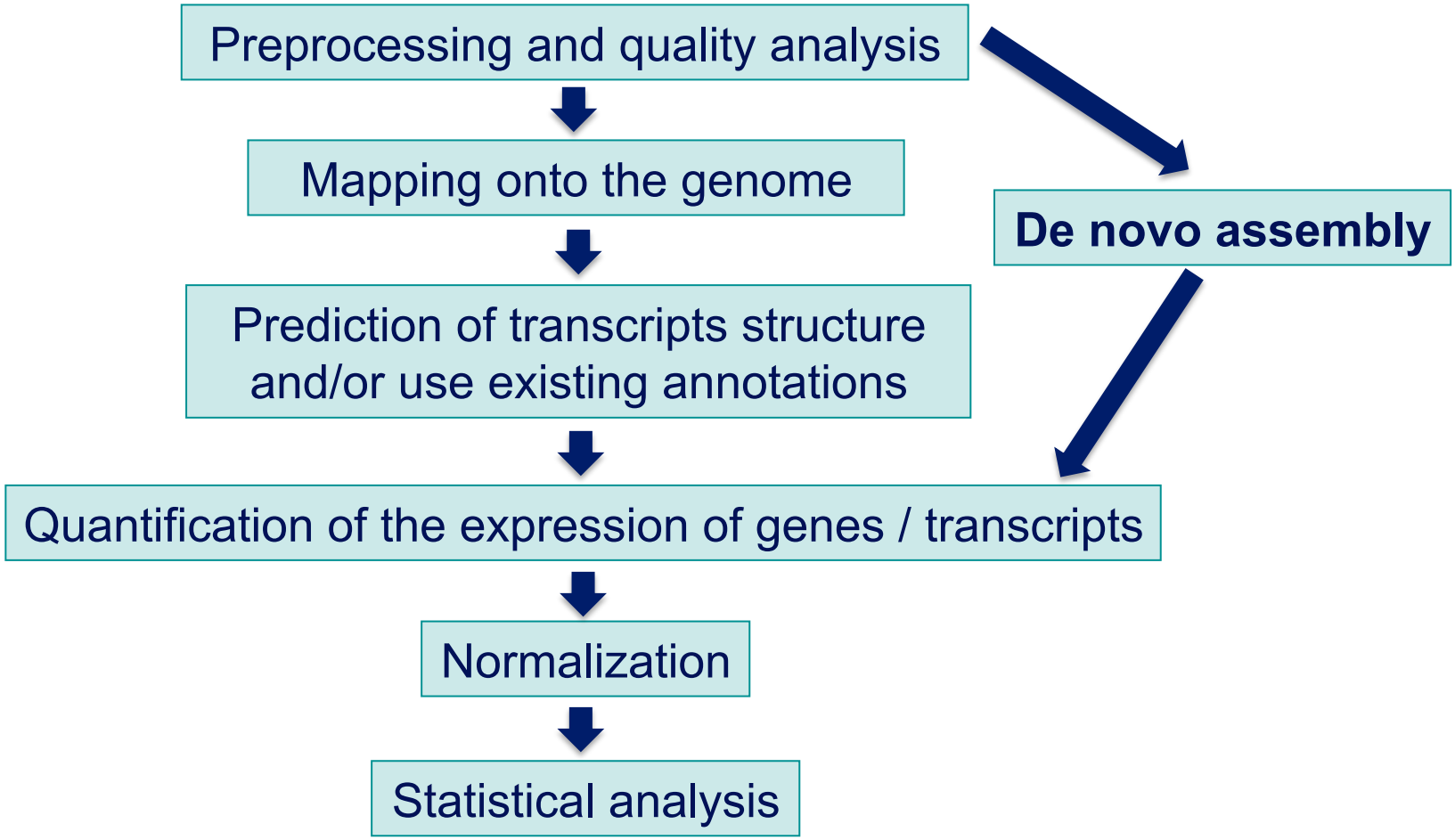
Analysis of RNAseq data

Céline Keime
keime@igbmc.fr

Analysis of RNA-seq data



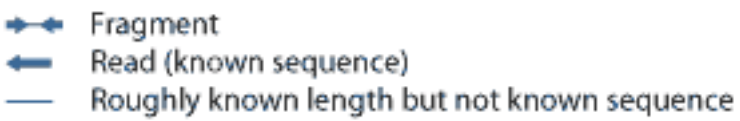
Analysis of RNA-seq data

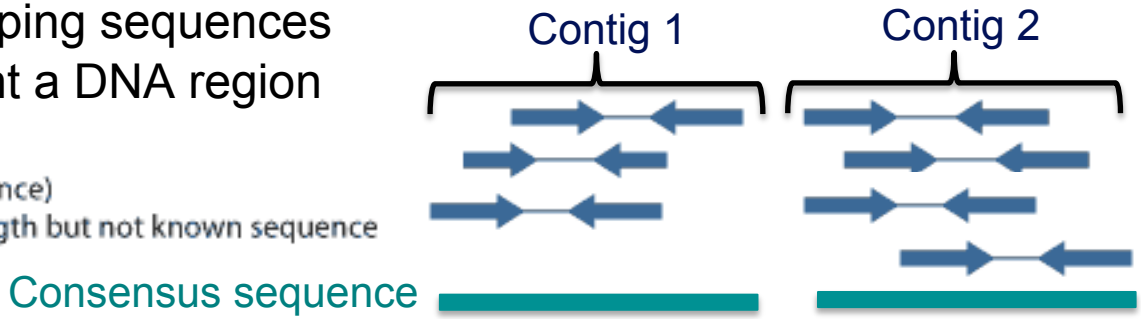


De novo transcriptome assembly

- Purpose
  - Analyse transcriptome on organisms without reference genome
  - Detect chimeric transcripts from chromosomal rearrangements
- Read coverage need to be high enough to build contigs

Contig : set of overlapping sequences that together represent a DNA region





- Challenges (as for genome assembly)
  - Repetitive regions, sequencing errors
- And more challenges specific to transcriptome assembly
  - Transcriptome coverage highly dependent on gene expression
  - Ambiguities in transcriptome assembly due to alternative splicing, alternative promoter usage, alternative polyA, overlapping transcripts

Programs for *de novo* transcriptome assembly

- Different programs
  - Velvet/Oases (Shulz et al. Bioinformatics 2012;28(8):1086-1092)
  - Trans-ABySS (http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss)
  - Trinity (Haas et al. Nature Protocols 2013; 8:1494–1512)
  - SOAPdenovo-Trans (http://soap.genomics.org.cn/SOAPdenovo-Trans.html)
  - Commercial software : CLC cell, Newbler

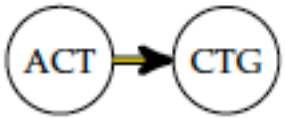
- Comparisons
  - On 454 data : Mundry et al. (Plos One 2012;7(2):e31410)
  - On Illumina data : Zhao et al. (BMC Bioinformatics 2011; 12(14):S2)
  - Which method will perform best is a function of read length, sequencing coverage and transcriptome complexity

De novo transcriptome assembly : general method

- Breaks reads into k-mers (short sub-sequences of length k)

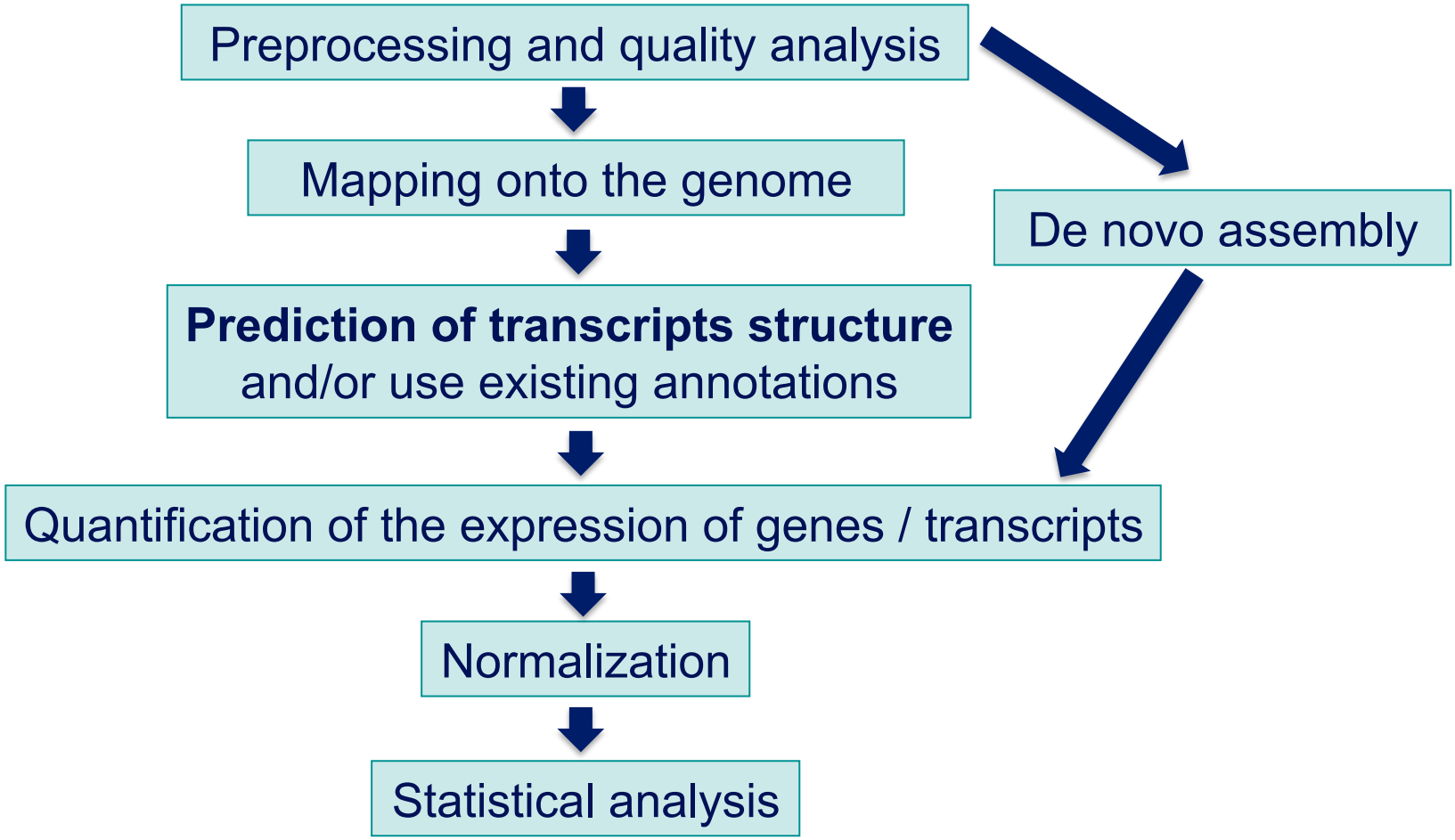
  e.g. 1 read = ACTG, k=3 ➔ k-mers = ACT, CTG

- Arranges k-mers into a graph structure (De Brujn graph)
  - Nodes : all sub-sequences of length k present in the sample
  - Arcs : link nodes to represent all sequences present in the sample



- Parse graph in order to create contigs
  - Look at the coverage to decide to follow a path or to remove it in order to avoid sequencing errors
- Choice of k-mer length greatly influence result of the assembly
- Functional annotation of contigs (with Gene Ontology e.g. Blast2GO, screen for Open Reading Frames, for known protein domains, ..)

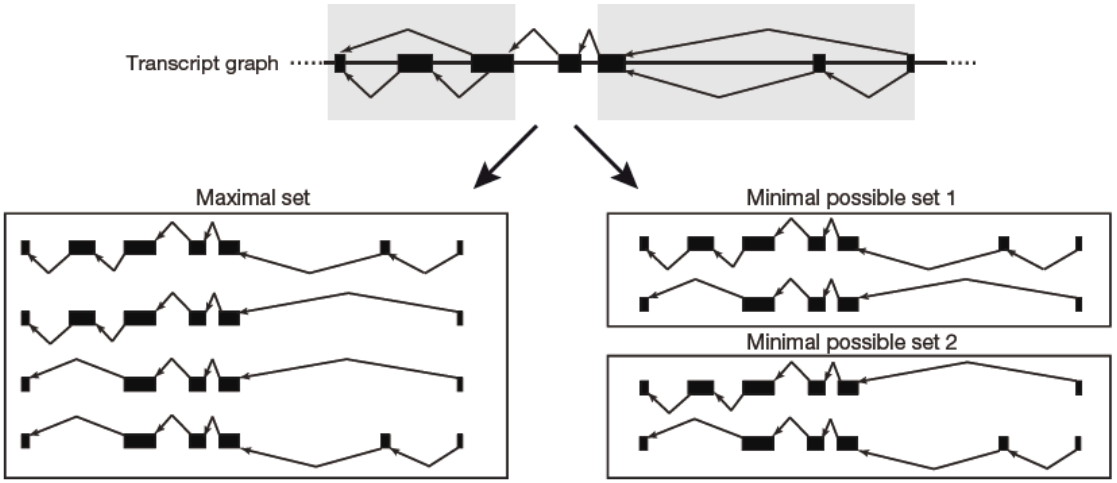
Analysis of RNA-seq data



Genome-guided assembly methods

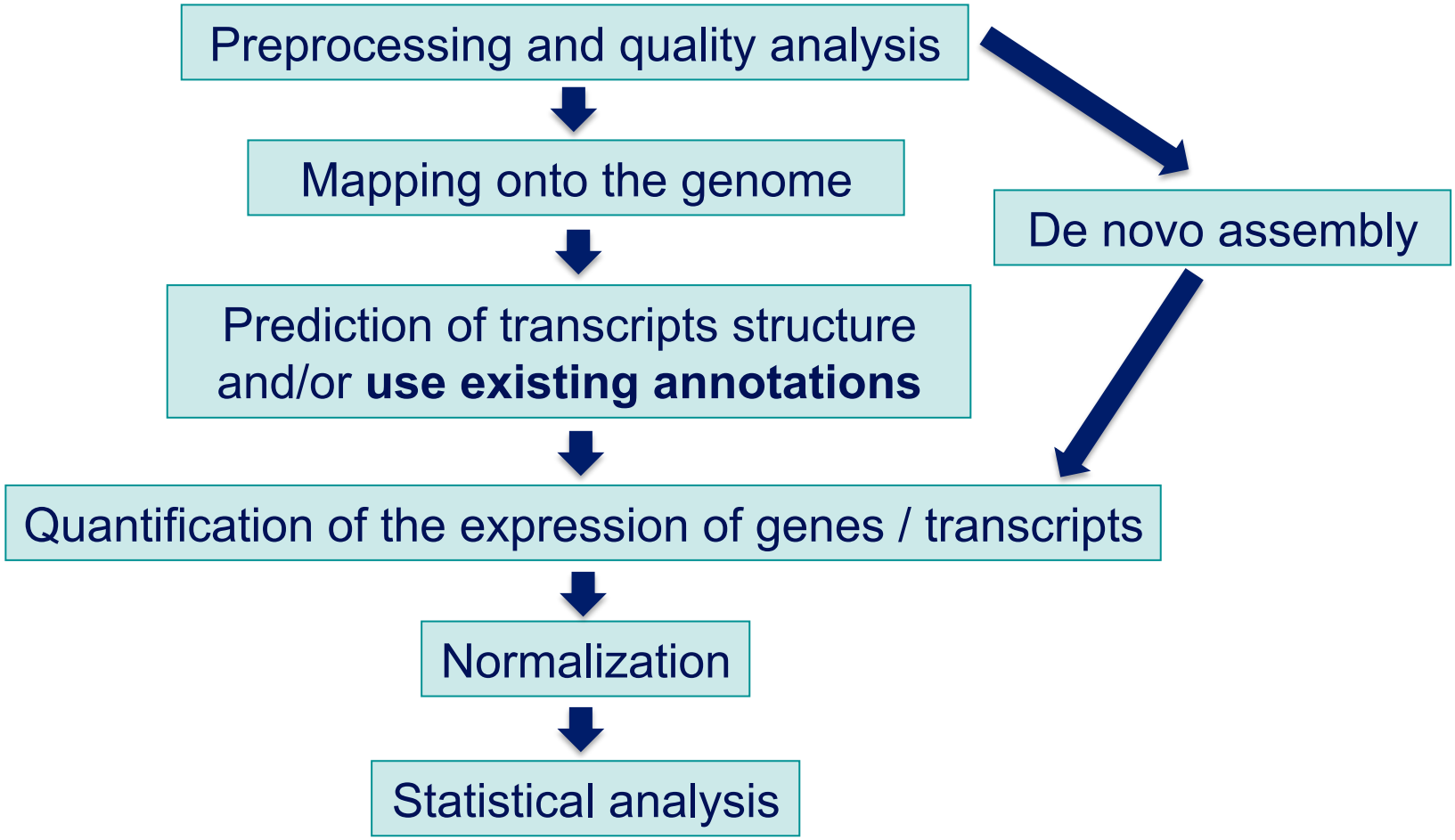
- Use spliced reads to reconstruct the transcriptome

1. Build a transcriptome assembly graph
2. Parse the graph into transcripts (1 path = 1 isoform)
   - ➔ Scripture report all isoforms that are compatible with the reads
   - ➔ Cufflinks reports the minimal number of compatible isoforms
     i.e. a minimal number of isoforms such that all reads are included in at least one path → use read coverage to decide which combination of isoforms is most likely to originate from the same RNA



Scripture (Guttman et al. Nature Biotechnology 2010;28(5):503-10)
Cufflinks (Trapnell et al. Nature Biotechnology 2010;28(5):511-5)

Analysis of RNA-seq data

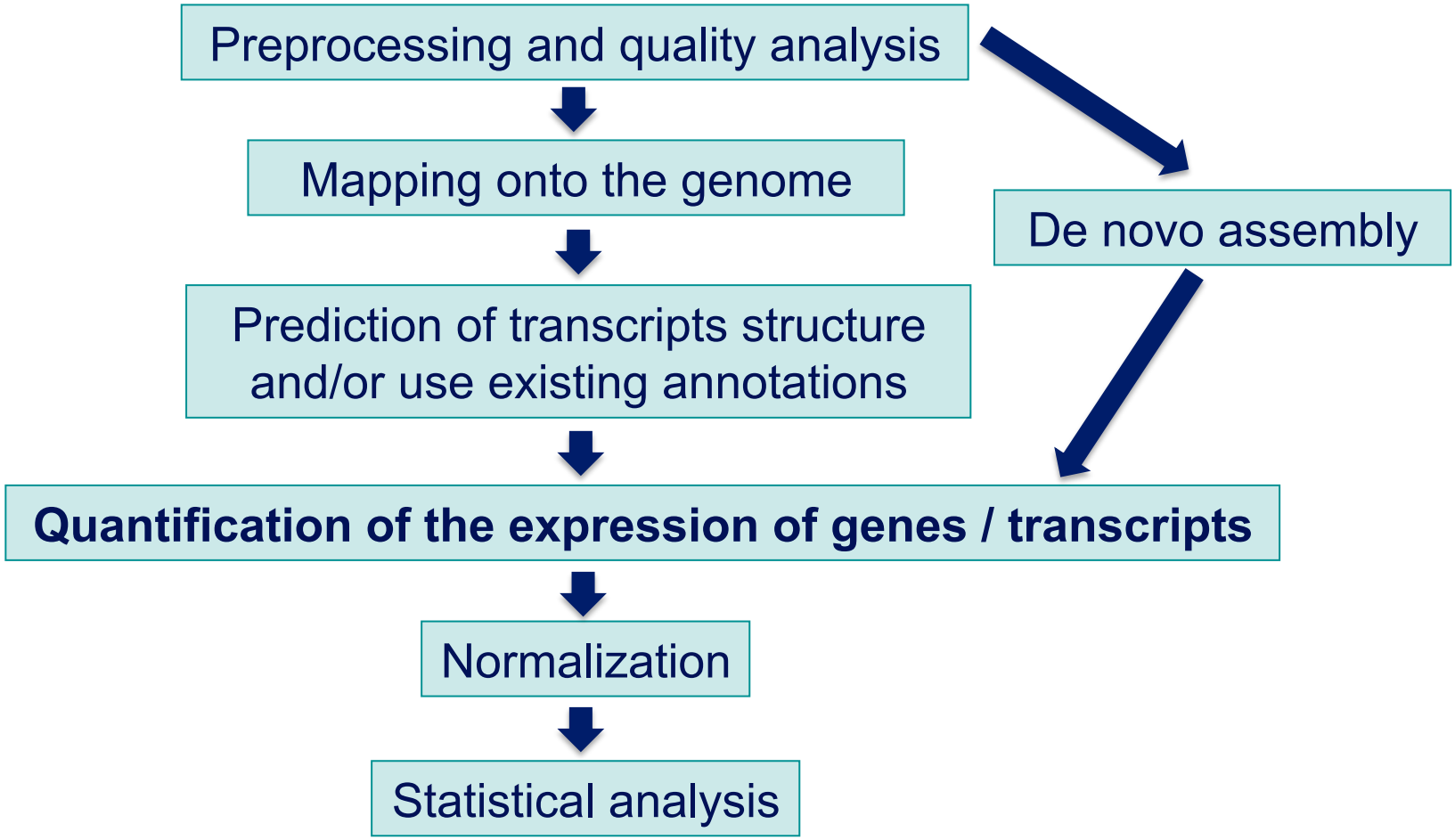


Genome annotations

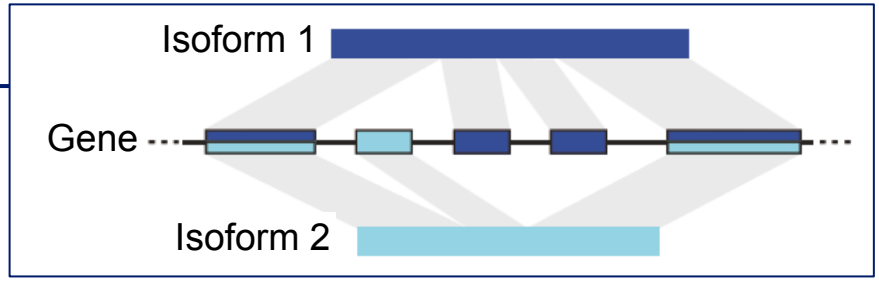
- Generally provided in a GFF/GTF file
  - GFF: General Feature Format / GTF : General Transfert Format
  - Text file format to describe genes and other features associated to DNA, RNA and protein sequences
  - Specifications : http://www.sanger.ac.uk/resources/software/gff/spec.html
  - eg human Ensembl 75 GTF file

```
#!genome-build GRCh37.p13
#!genome-version GRCh37
#!genome-date 2009-02
#!genome-build-accession NCBI:GCA_000001405.14
#!genebuild-last-updated 2013-09
1	pseudogene	gene	11869	14412	.	+	.	gene_id
"ENSG00000223972"; gene_name "DDX11L1"; gene_source "ensembl_havana"; gene_bioty
pe "pseudogene";
1	processed_transcript	transcript	11869	14409	.	+
.	gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "D
DX11L1"; gene_source "ensembl_havana"; gene_biotype "pseudogene"; transcript_nam
e "DDX11L1-002"; transcript_source "havana";
1	processed_transcript	exon	11869	12227	.	+	.
gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; exon_number "1"; gen
e_name "DDX11L1"; gene_source "ensembl_havana"; gene_biotype "pseudogene"; trans
cript_name "DDX11L1-002"; transcript_source "havana"; exon_id "ENSE00002234944";
```

Analysis of RNAseq data

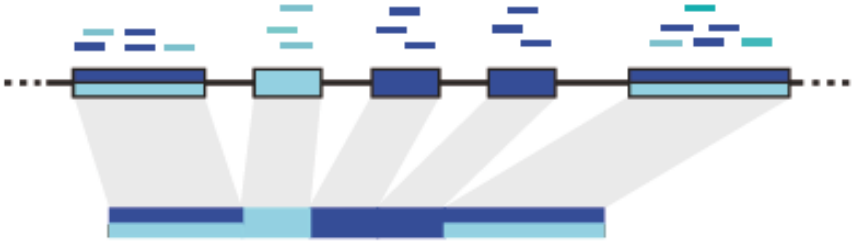


Gene-level quantification



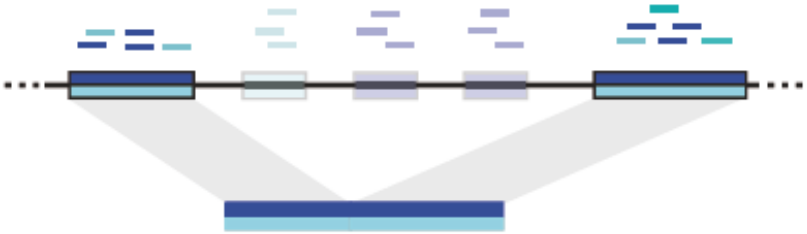
- Exon-union method

  Count reads mapped to all exons from all isoforms of the gene



- Exon-intersection method

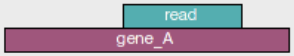
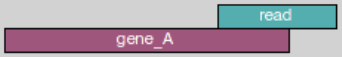
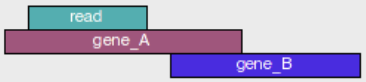
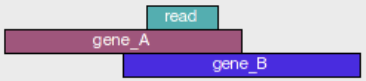
  Count only reads mapped to its constitutive exons



➔ reduce power for differential expression analysis

Gene-level quantification

- Exon-union method
  - HTSeq (Anders et al., Bioinformatics 2015;31(2):166-9)

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Application

- htseq-count has been used on the 4 RNAseq samples from MITF dataset to quantify gene expression, using
  - BAM alignment files
    - Only reads with one reported alignment are considered
  - intersection_nonempty method
  - Annotations from Ensembl v75
    - ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz

Application : results

- One tabulated text file per sample
  - Number of reads for each Ensembl gene

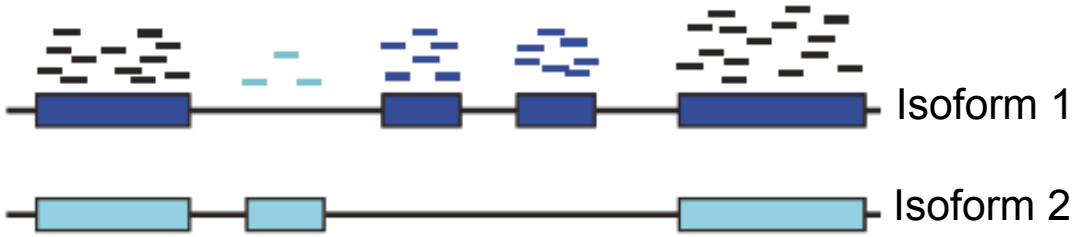
```
ENSG00000000003 1056
ENSG00000000005 0
ENSG00000000419 2661
ENSG00000000457 602
ENSG00000000460 2077
ENSG00000000938 2
ENSG00000000971 75
```

- Summary of quantification results

Sample ID	Sample name	% of assigned reads	% of no feature reads	% of ambiguous reads
TSB-11_5_S1	siLuc2	87.42	8.52	4.05
TSB-12_6_S1	siLuc3	87.13	8.88	3.99
TSB-13_19_S	siMitf3	87.06	8.91	4.03
TSB-14_12_S2	siMitf4	88.07	7.86	4.07

Transcript-level quantification

- Some reads cannot be assigned unequivocally to a transcript



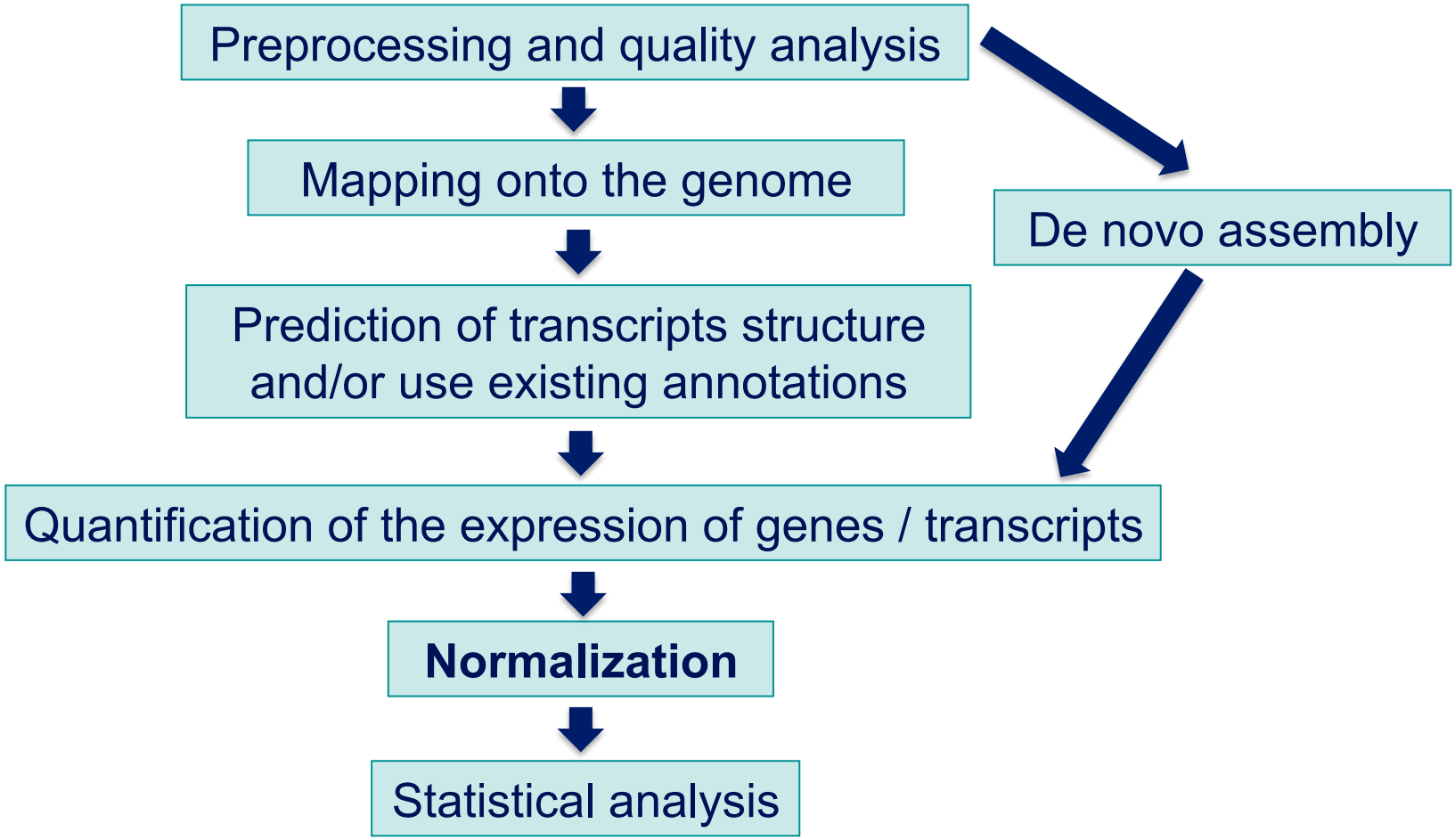
- Alexa-seq (Griffith et al. Nature methods 2010;7(10):843-7)

  Count only reads that map uniquely to a single isoform

  ➔ Fails for genes that do not contain unique exons from which to estimate isoform expression

- Cufflinks (Trapnell et al. Nature Biotechnology 2010;28(5):511-5)
  MISO (Nature Mathods 2010 Dec;7(12):1009-15)
  - Construct a likelihood function that models the sequencing process
  - Calculate isoforms abundance estimates that best explain reads observed in the experiment

Analysis of RNA-seq data



Normalization : why ?

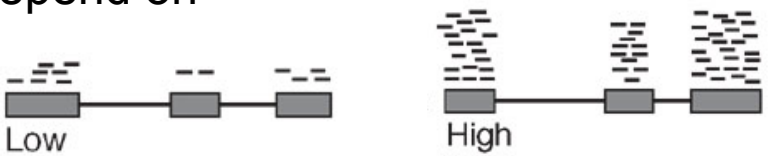
- To compare RNA-seq libraries
  - with different sizes, eg :

Sample ID	Sample name	Total number of reads
TSB-11_5_S1	siLuc2	44,340,015
TSB-12_6_S1	siLuc3	49,763,265
TSB-13_19_S2	siMitf3	42,595,950
TSB-14_12_S2	siMitf4	39,065,527

- To compare the expression level of several genes within a library

Indeed read counts depend on

- Expression level



- Gene length

- Library size

Different normalization methods

- Based on distribution adjustment
  - Total read count
    - Motivation

    Higher library size ➔ higher counts
    - Method

    Divide counts by total number of reads
  - Upper quartile (Bullard et al. BMC Bioinformatics 2010;11,94) / Median
    - Motivation

    Total read count is strongly dependent on a few highly expressed transcripts
    - Method

    Divide counts by the upper quartile/median of the counts different from 0
  - Quantile (Bolstad et al. Bioinformatics 2003; 19:185–93)
    - Assumption

      Read counts have identical distribution across libraries
    - Method

      Count distributions are matched between libraries

Different normalization methods

- Take into account gene/transcript length
  - RPKM (Mortazavi et al. Nat Methods 2008;5:621–8), FPKM
  - **R**eads (**F**ragments) per **K**ilobase per **M**illion mapped reads
  - Assumption
    - Read counts =f(expression level, gene length, library size)
  - Method
    - Divide counts by gene length (kb) and total nb of reads (million)
  - Allows to compare expression levels between genes

Different normalization methods

- Based on the “effective library size” concept
  - Assumption
    - Most genes are not differentially expressed
  - 2 methods
    - Trimmed Mean of M values (Robinson et al. Genome Biol. 2010;11:R25)
    - DESeq normalization (Anders et al. Genome Biol. 2010;11:R106)

Which normalization method to choose ?

- Comparison on 4 real and 1 simulated dataset
- Summary of comparison results

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	–	+	++	–
RPKM	–	+	+	–	–

- : the method provided unsatisfactory results for the given criterion
+ : satisfactory results
++ : very satisfactory results

(Dillies et al. Brief. Bioinformatics 2013 Nov;14(6):671-83)

DESeq normalization method

	lib1	lib2	lib3	…	lib **j**	lib **n**
gene1	468	475	501			
gene2	45	56	76			
gene3	2576	560	578			
gene4	1678	1798	1867			
…						
gene **i**					x_{ij}	

n : number of samples to compare

xij : number of reads
for gene i in sample j

DESeq normalization method

	lib1	lib2	lib3	…	lib **j**	lib **n**
gene1	468	475	501			
gene2	45	56	76			
gene3	2576	560	578			
gene4	1678	1798	1867			
…						
gene **i**					x_{ij}	

n : number of samples to compare

xij : number of reads for gene i in sample j

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

➔ Each value is divided by the geometric mean of its row

➔ Normalization factor = median of all these ratios

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468	475	501	m1=481.1263
gene2	45	56	76	m2=57.64187
gene3	2576	560	578	m3=941.2115
gene4	1678	1798	1867	m4=1779.271

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468 / m1	475 / m1	501 / m1	m1=481.1263
gene2	45 / m2	56 / m2	76 / m2	m2=57.64187
gene3	2576 / m3	560 / m3	578 / m3	m3=941.2115
gene4	1678 / m4	1798 / m4	1867 / m4	m4=1779.271

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{v=1}^{n} x_{iv})^{1/n}}$$

DESeq normalization method

	lib1	lib2	lib3	mean
gene1	468 / m1	475 / m1	501 / m1	m1=481.1263
gene2	45 / m2	56 / m2	76 / m2	m2=57.64187
gene3	2576 / m3	560 / m3	578 / m3	m3=941.2115
gene4	1678 / m4	1798 / m4	1867 / m4	m4=1779.271
median	0.9577858	0.9793598	1.0452989	

normalization factors

Normalization factor for library j :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

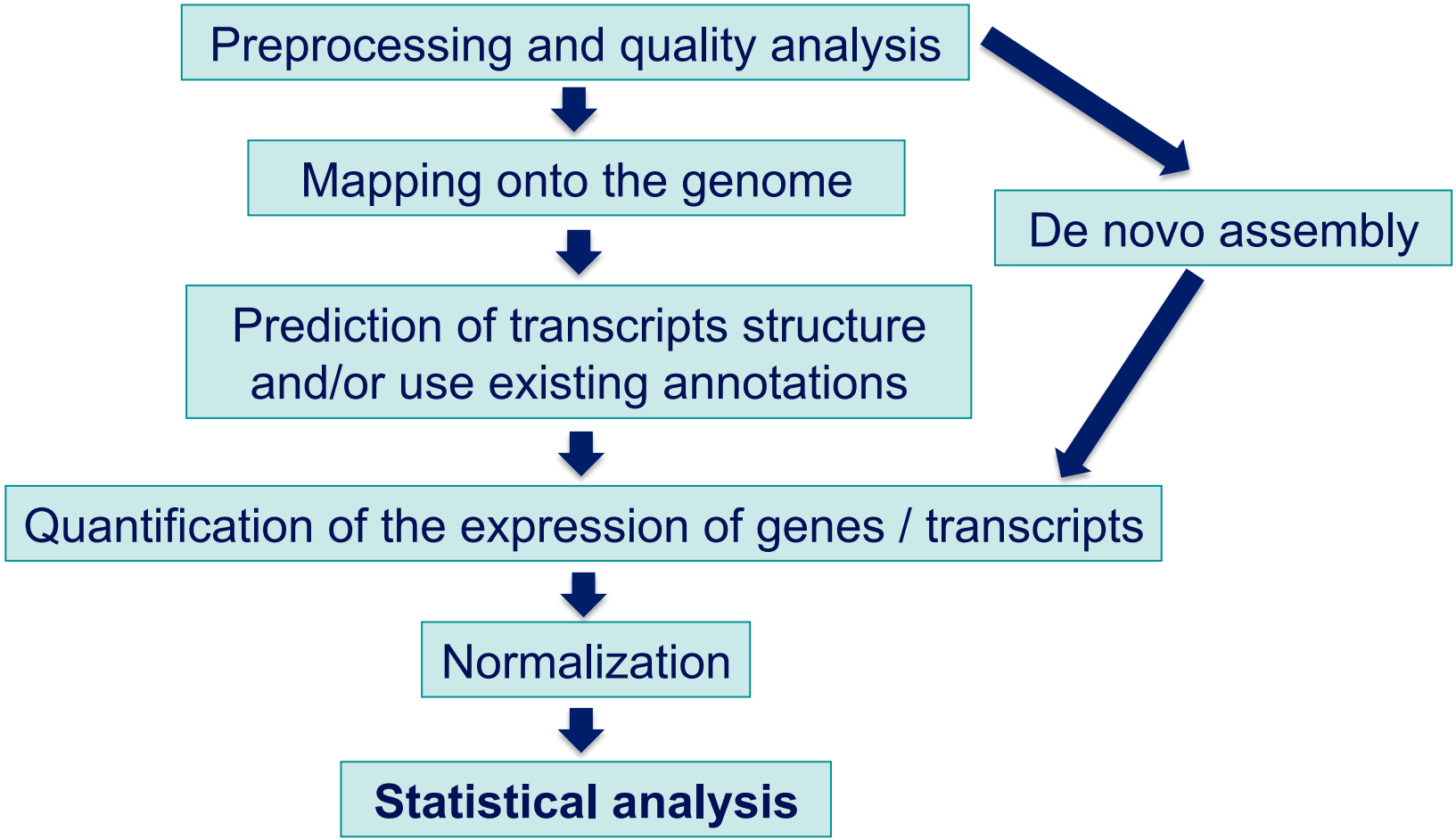
Application

- The DESeq normalization method has been used to normalize the 4 RNA-seq samples from MITF dataset, using R and DESeq2 Bioconductor package available in

  http://bioconductor.org/packages/release/bioc/html/DESeq2.html
- Resulting normalization factors

Sample ID	Sample name	Total number of reads	Normalization factors
TSB-11_5_S1	siLuc2	44,340,015	1.0141592
TSB-12_6_S1	siLuc3	49,763,265	1.1547005
TSB-13_19_S2	siMitf3	42,595,950	0.9725945
TSB-14_12_S2	siMitf4	39,065,527	0.8927402

Analysis of RNA-seq data



Data exploration

- Exploration and visualisation of data
  - Essential step before any analysis
  - Allows data quality assessment and control
  - Eventually leads to remove data with insufficient quality

Data exploration

- Samples clustering
  - Distance that could be used
    - d=1-ρ (ρ=Spearman correlation coefficient)
    - SERE coefficient (Schulze et al. BMC Genomics 2012;13:524)

      Simple Error Ratio Estimate

      $$\text{SERE} = \frac{\text{Observed standard deviation between two samples}}{\text{Value that would be expected from an ideal experiment}}$$

      SERE = 0 ➔ data duplication

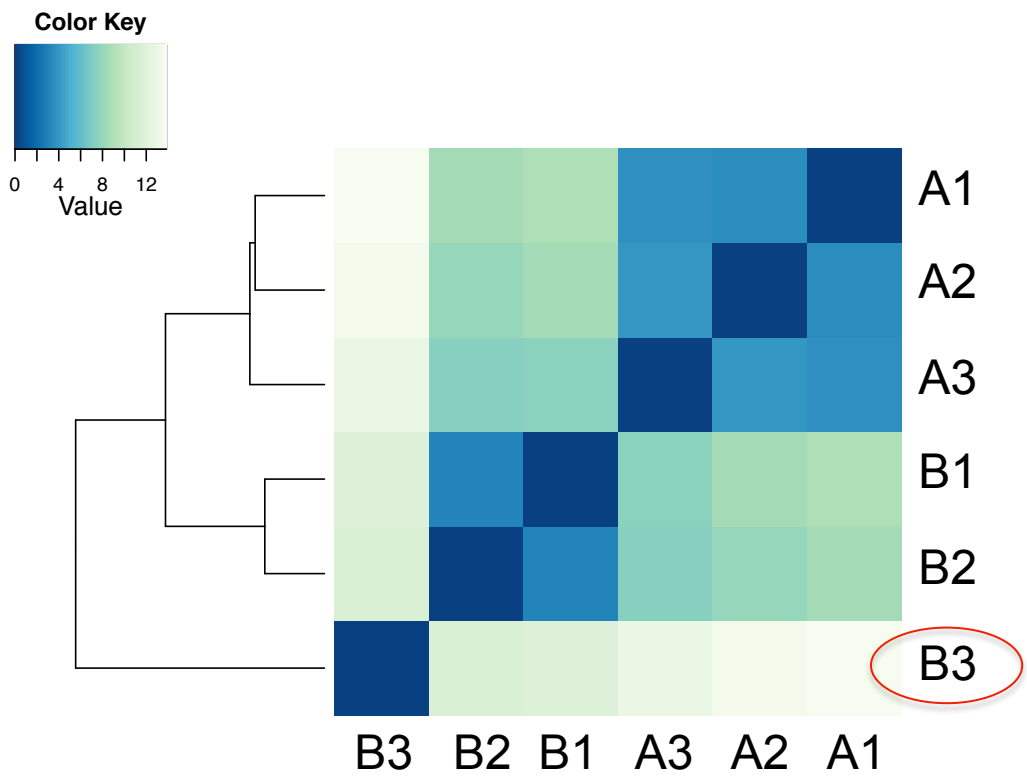
      SERE = 1 ➔ technical replication

      SERE > 1 ➔ biological variation

- Multivariate analyses
  - Useful for visualizing the overall effect of experimental covariates and batch effects
  - e.g. Principal Component Analysis ➔ Anders et al. proposed data transformation methods that can be used before performing PCA

Data exploration

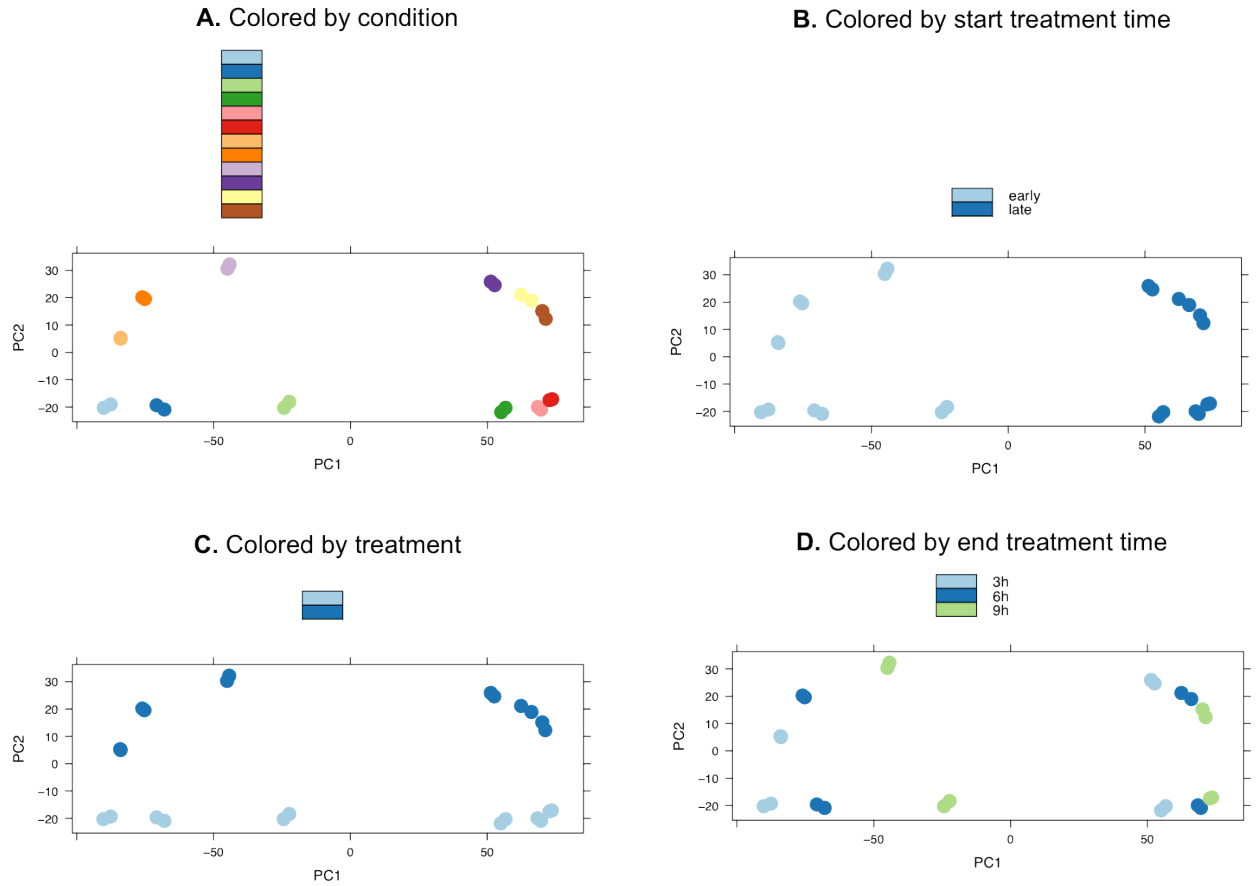
- Example of data clustering with SERE coefficient
  - A-B : 2 different conditions, 1-2-3 : replicate samples



Data exploration

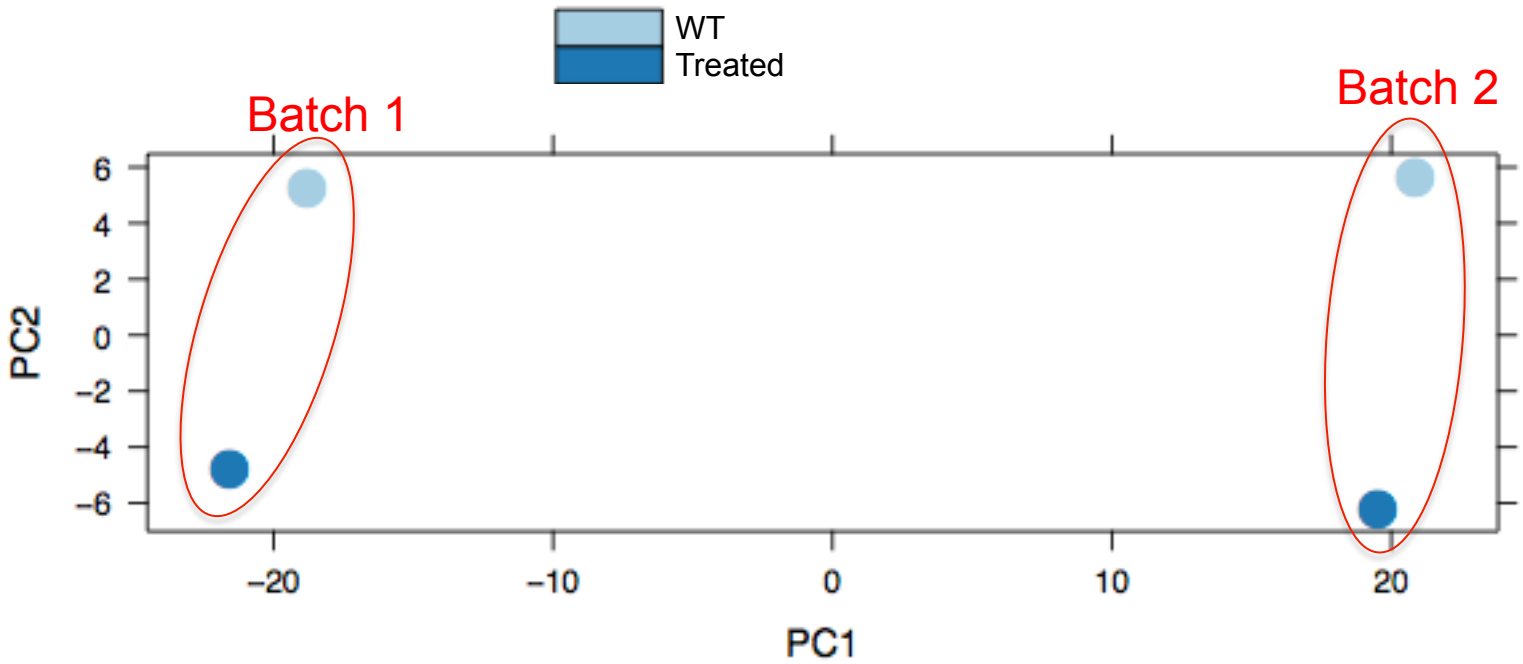
- Example of PCA calculated on variance stabilized data from 24 RNA-seq libraries : first factorial plan

  the 1st axis explains 81% and the 2nd 8% of the variability



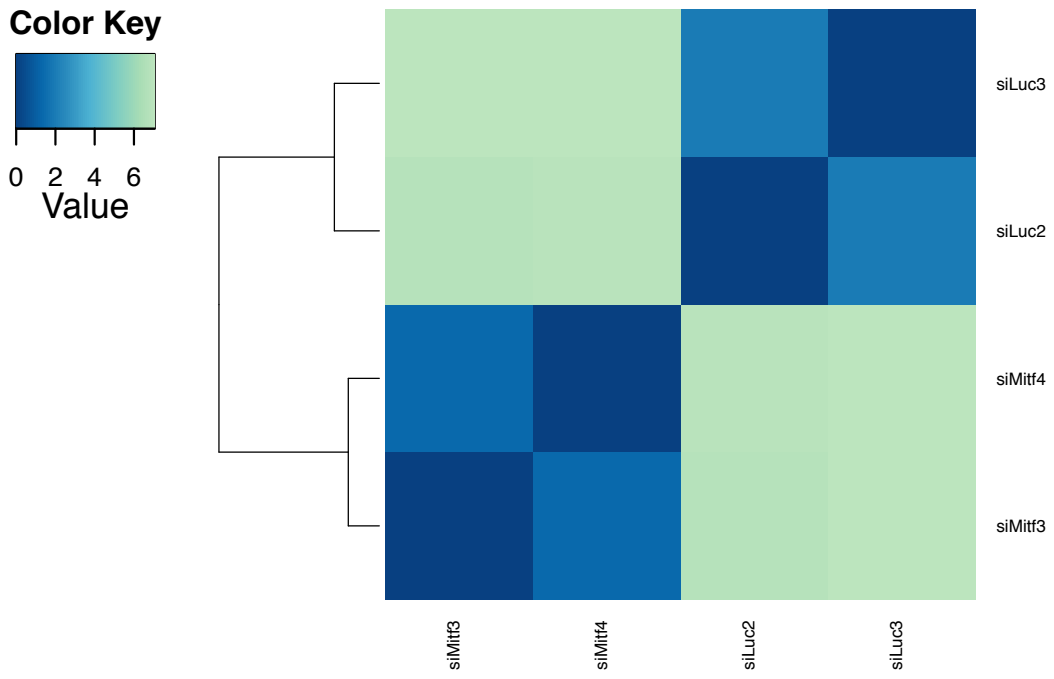
Data exploration

- Example with batch effect
  - PC1 : 30% of the variability = batch effect
  - PC2 : 22% of the variability = WT vs treatment
  - ➔ This batch effect has to be taken into account in statistical analysis



Application : data exploration

- Heatmap and clustering - 4 RNA-seq libraries from MITF project
  - On SERE coefficients calculated between all pairs of libraries
  - Clustering calculated with UPGMA
  - Performed using R software



Search for significantly differentially expressed genes

- What is significant differential expression ?
  - The observed difference between conditions is statistically significant i.e. greater than expected just due to random variation

- Microarray vs RNA-seq
  - Microarray

    Fluorescence proportional to expression ➔ continuous data
  - RNA-seq

    Number of reads assigned to a feature (gene, transcript) proportional to expression ➔ count data

- Here we focus on count-based measures of **gene** expression

Search for significantly differentially expressed genes

- Use only a fold-change ranking ?
  - Do not take variability into account
  - Do not take level of expression into account
  - No control of the false positive rate
- Hypothesis testing
  - For each gene
    - H0 : No gene expression difference between the compared conditions
    - H1 : There is a gene expression difference between the compared conditions
- Steps
  - Choose a statistic
  - Define a decision rule
    - Define a threshold below which we will reject H0

Statistic to search for significantly differentially expressed genes

- Sequencing a library = randomly and independently choose N sequences from the library

  → read counts = multinomial distribution
- High number of reads, probability of a read assigned to a given gene small → Poisson approximation
  - Distribution of counts across **technical** replicates for the majority of genes fit well to a Poisson distribution

    Marioni et al. Genome Research 2008;18(9):1509-17

    Bullard et al. BMC Bioinformatics 2010;11,94

- But Poisson distribution : variance = mean

  ➔ Across **biological** replicates variance > mean for many genes

  (Anders et al. Genome Biology 2010;11:R106) : **overdispersion**

  ➔ **Negative binomial distribution** : a good alternative to Poisson in the case of overdispersion

Negative binomial models

- How to estimate the overdispersion parameter ?
  - Very few replicates ➔ challenging issue
  - A common dispersion for all genes (Robinson et al. Biostatistics 2008;9(2):321-32)
    ➔ rarely appropriate assumption
  - edgeR (Robinson et al. Bioinformatics. 2010;26(1):139-40)
    DESeq (Anders et al. Genome Biology 2010;11:R106)
    DESeq2 (Anders et al. Genome Biol. 2014;15(12):550)

    Allow the different genes to have different dispersion parameters but improve the estimation of these parameters by borrowing information across genes
- Generalized linear models : edgeR, DESeq2
  - Generalization of a linear model that allows response variables to have other than normal distribution, e.g. negative binomial
  - Allow to analyse multifactor designs

Definition of a decision rule

- p-value
  - Probability of obtaining a statistic at least as extreme as the one that was actually observed, assuming that H0 is true

- Reject H0 if p-value < threshold
  - Common threshold = 0.05

  ➔ the observed result would be highly unlikely under H0

  But be careful : you perform multiple testing !

Multiple testing problem

- To identify significantly differentially expressed genes
  ➔ as many tests as the number of genes (G)
- With a type I error α for each gene
  - we expect to find $G\alpha$ false positives
  - i.e. $G\alpha$ genes declared to be differentially expressed even though there are not
  - e.g. G=30,000 genes α=0.05
    ➔ We expect to find 1,500 false positives
  ➔ Important to control the false positive rate when we make a lot of tests
- 2 points of views
  - Individually consider the differentially expressed genes sorted according to a statistic
  - Consider a list of differentially expressed genes, in which we would like to control the false positive rate
    ➔ Use a multiple testing correction

Multiple testing correction methods

- Control the Family-Wise Error Rate (FWER)
  - Definition
    - FWER : Probability to have at least one false positive
    - e.g. FWER = 0.05 ➔ 5% chances of having at least one false positive
  - Methods to control the FWER
    - Bonferroni

    $p_{g\_adjusted} = \min(Gp_g, 1)$

    ➔ Each test is performed with a type I error α/G
    - Westfall et Young (1993)
  - Very conservative methods (Ge et al. TEST 2003;12(1):1-77)

Multiple testing correction methods

- Control the False Discovery Rate (FDR)
  - Definition
    - Expected proportion of false positives among genes declared as differentially expressed
    - e.g. FDR = 0.05 ➔ We expect to find 5% of false positives among genes declared as significantly differentially expressed
  - Methods to control the FDR
    - Benjamini and Hochberg (Journal of the R. Stat. Soc., Series B 57 (1): 125–133)
      - Hypothesis : independence the of tests performed
    - Benjamini and Yekutieli (Ann Stat 2001; 29:1165-1188)
      - Hypothesis : dependency of the tests performed (e.g. due to genes co-regulations)
      - Very conservative method (Ge et al. TEST 2003;12(1):1-77)

→ Less stringent than controlling the FWER

Application :
Statistical analysis results

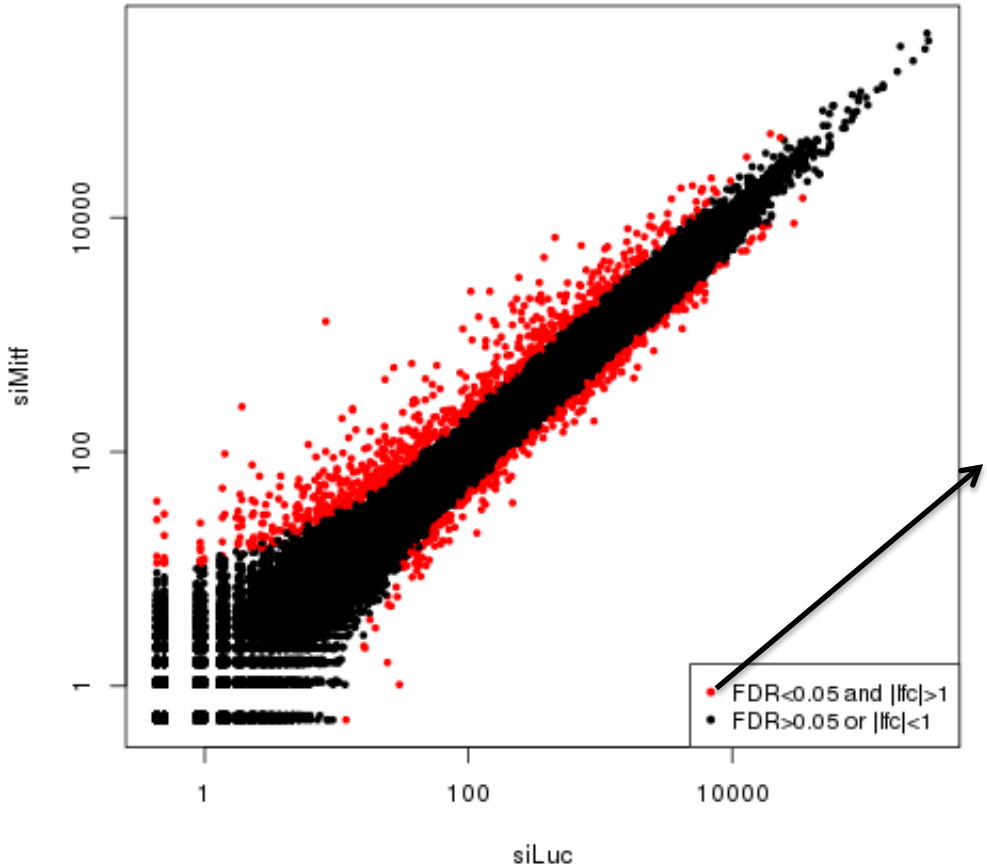
- Test to search for significantly differentially expressed genes performed using R and DESeq2 Bioconductor package available in

  http://bioconductor.org/packages/release/bioc/html/DESeq2.html
- Adjustment for multiple testing performed using the Benjamini and Hochberg method
- Annotations performed with the biomaRt Bioconductor package available in

  http://bioconductor.org/packages/release/bioc/html/biomaRt.html

Application :
Statistical analysis results

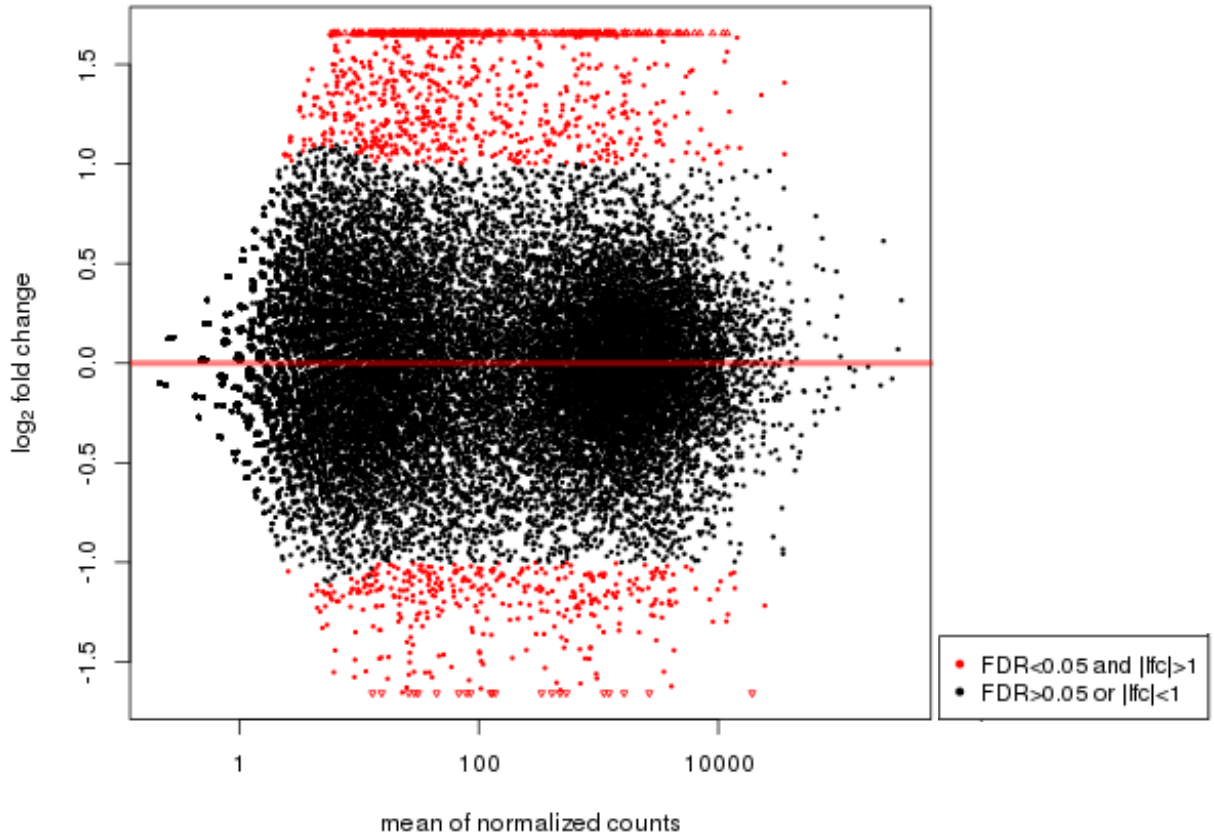
- Scatter plot



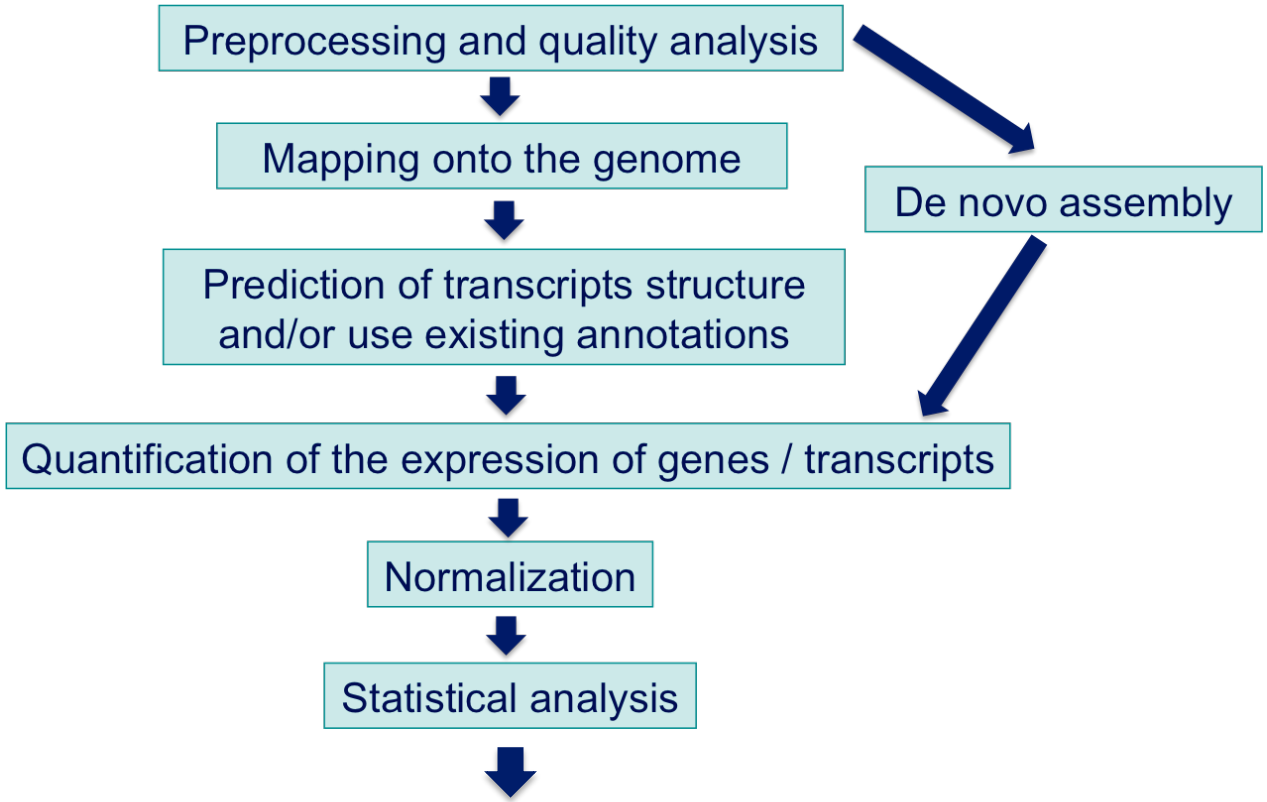
Significantly differentially expressed genes
742 over-expressed genes
(siMitf > siLuc)
272 under-expressed genes
(siMitf < siLuc)

Application :
Statistical analysis results

- MA-plot



Analysis of RNA-seq data

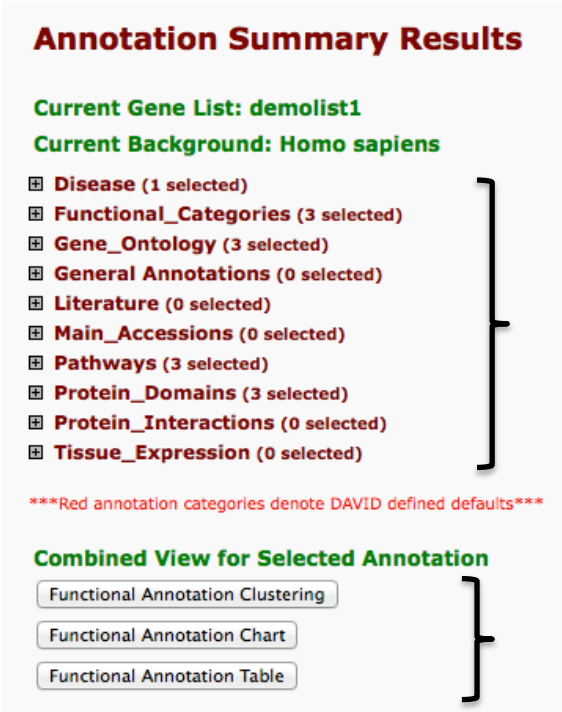


Functional enrichment analysis, pathway analysis, integration with other data, …

Functional analysis

- A lot of functional analysis tools available
  - Initially developed for microarray data
  - e.g. GO tools listed in
    http://omictools.com/gene-ontologies-c25-p1.html
  - Methods specific to RNA-seq data
    - goseq (Young et al., Genome Biology 2010;11:R14)
    - SeqGSEA (Wang et al. BMC Bioinformatics 2013, 14(Sup5):S16)
    - GSAASeqSP (Xiong et al Scientific Reports 2014; 4:6347)
- DAVID will be used for this practical session because
  - Graphical interface & free software
- DAVID
  - **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery
  - http://david.abcc.ncifcrf.gov/
  - A very interested article describing how to use DAVID :
    Huang et al. Nature Protocols 2009;4(1):44-57.

DAVID



Different sources of annotation

- Disease (OMIM)
- Gene Ontology
- Pathways (KEGG, Biocarta)
- Protein Domains (InterPro, SMART)
- Protein Interaction (BIND)
- …

Different tools

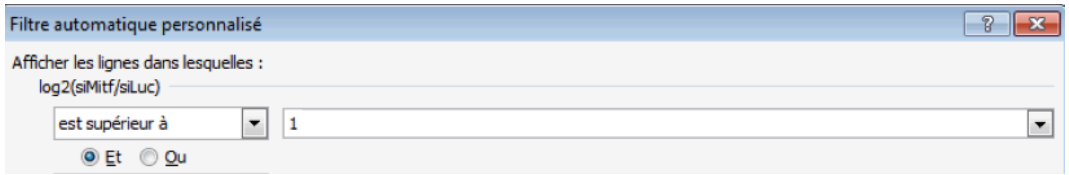
- Functional Annotation Clustering
  - Cluster functionally similar terms associated with a gene list into groups
- Functional Annotation Chart
  - Identify enriched annotation terms associated with a gene list
- Functional Annotation Table
  - Query associated annotations for all genes from a list

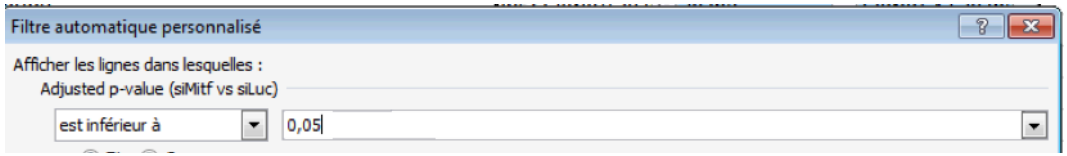
Exercise : functional analysis

- Use DAVID to perform functional analysis of genes significantly over-expressed in siMitf vs control samples
  - Proposed thresholds to select significantly differentially expressed genes : Adjusted p-value < 0.05 and log2FoldChange > 1
  - Go to http://david.abcc.ncifcrf.gov
  - Click on Start Analysis

Exercise : functional analysis

- Excel
  - Select the 2 columns containing log2FC and Adjusted-pvalue
  - Données -> Filtrer
  - Click on the filter icon
  - Filtres numériques :

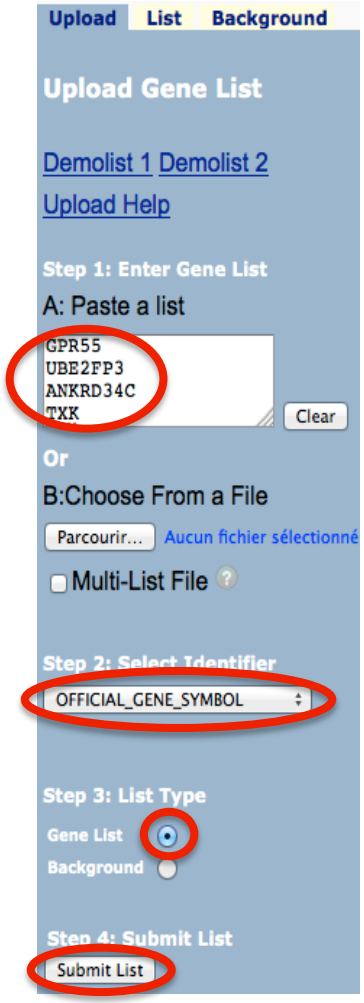




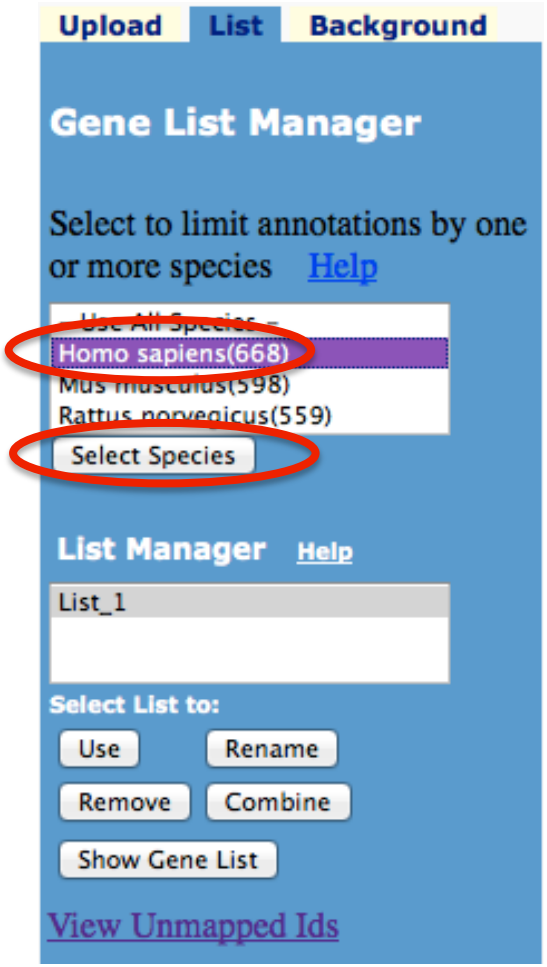
742 enregistrement(s) trouvé(s) sur 24956

Exercise : functional analysis

- Enter your gene list



- Select species



Exercise : functional analysis

1. What are the 5 most enriched functional annotation terms among annotations of the genes from your list ?

   How many genes are annotated with each of these terms ?

   What are the genes annotated with the most enriched term ?
2. As you see redundancy in previous results, it could be interesting to cluster functionally similar terms into groups.

   Perform this clustering.

   What is the first identified cluster ? Visualize members of this cluster (genes and annotation terms) by clicking on 
3. claudin 15 gene is a member of this cluster.

   What are all associated annotations for this gene ?

   Among these annotations you will find the KEGG pathway “Cell adhesion molecules”.
   Are other genes from your list member of this pathway ?