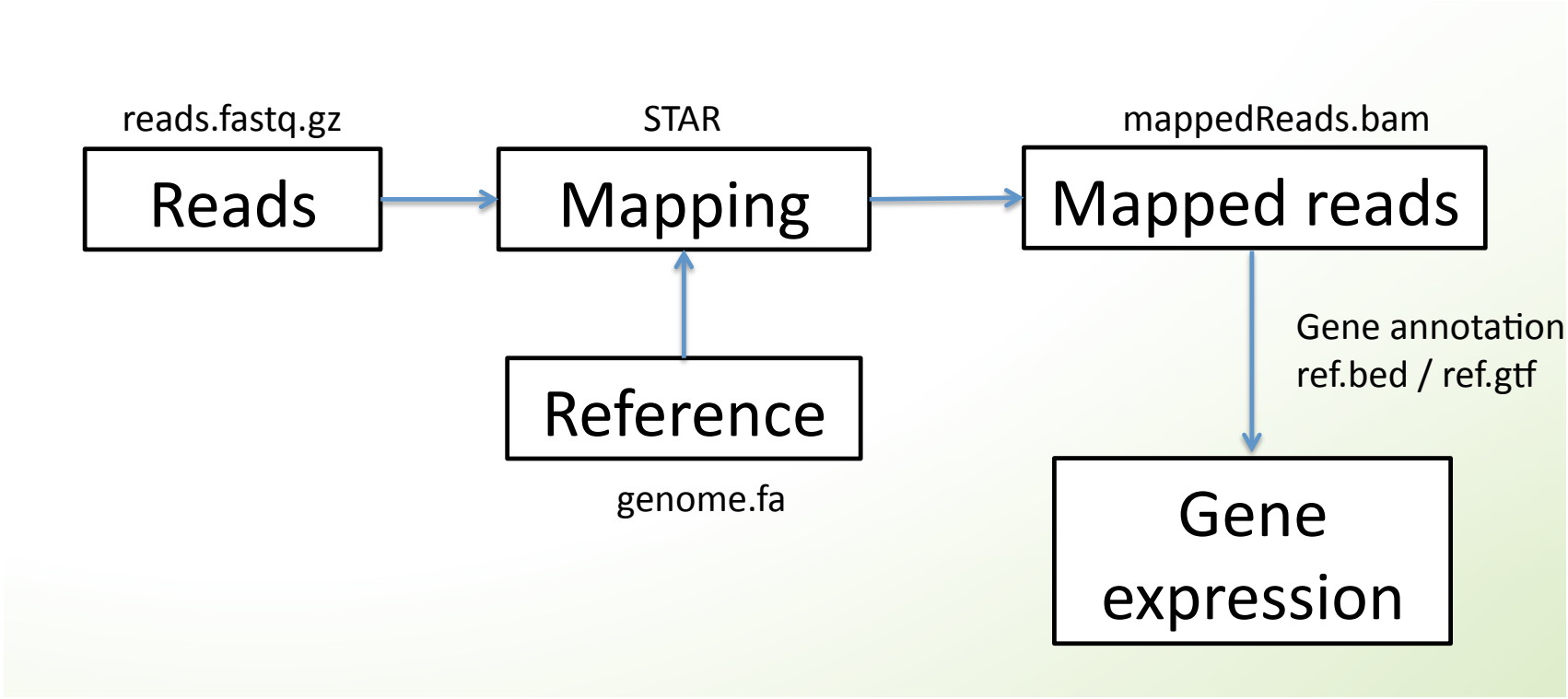
RNAseq course

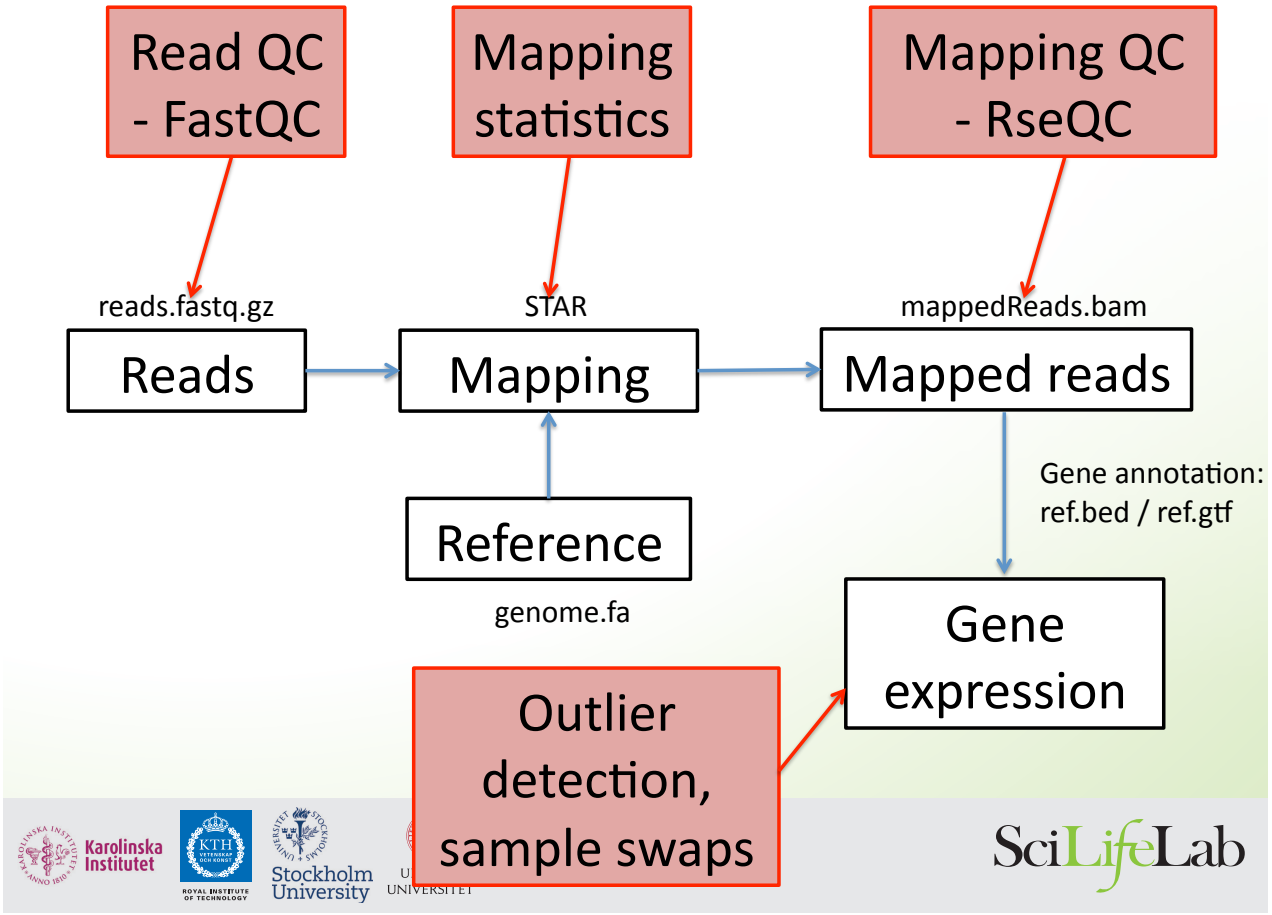
November 2015

23th – 25th

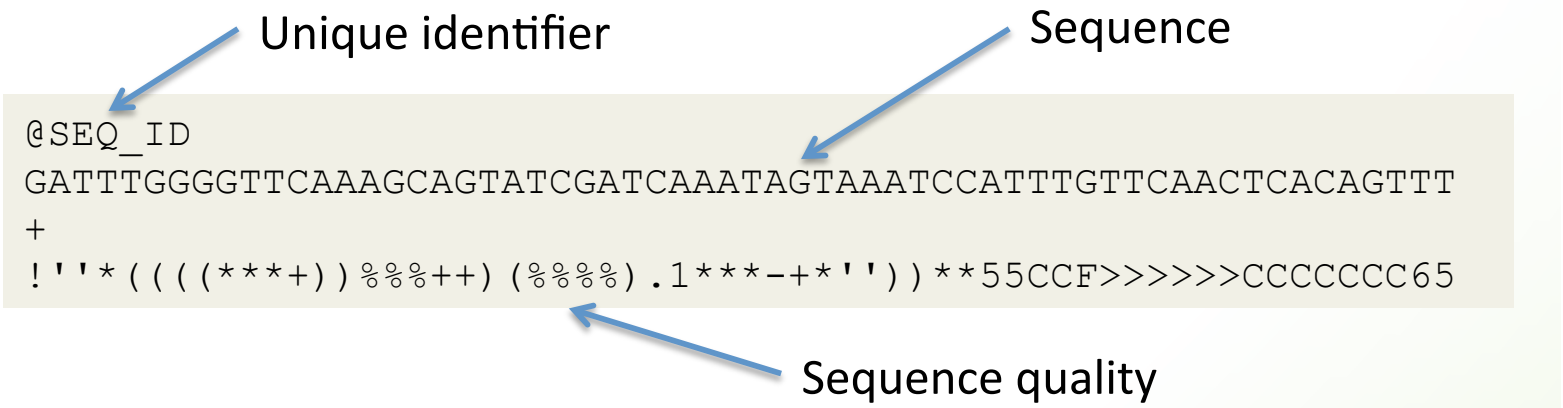
RNA-seq analysis workflow



Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

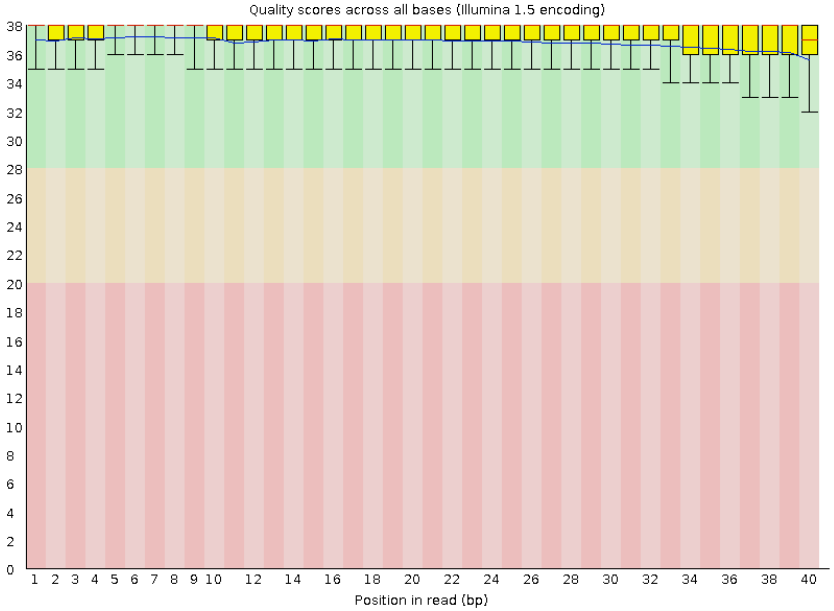
Fastq – read file format

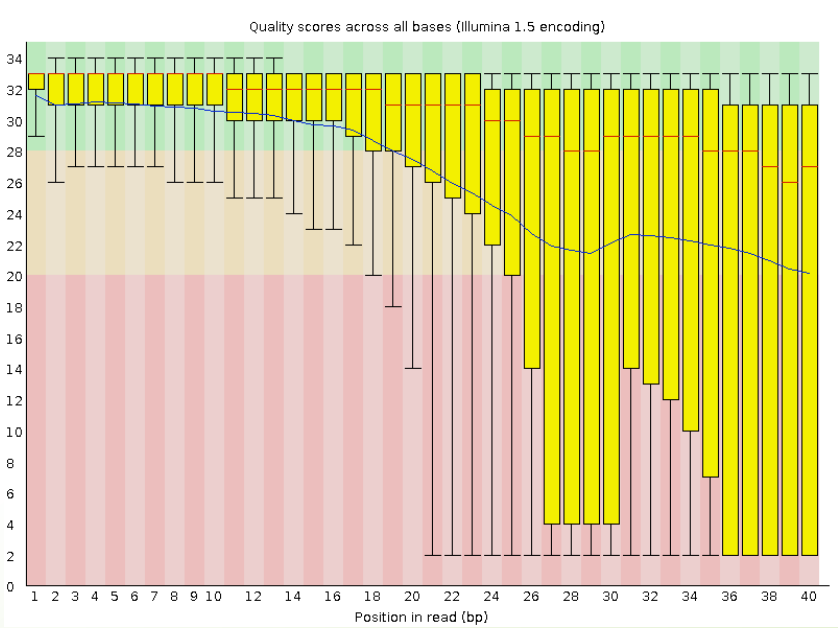


Sequence quality (Phred score) = Value between 0 and 40

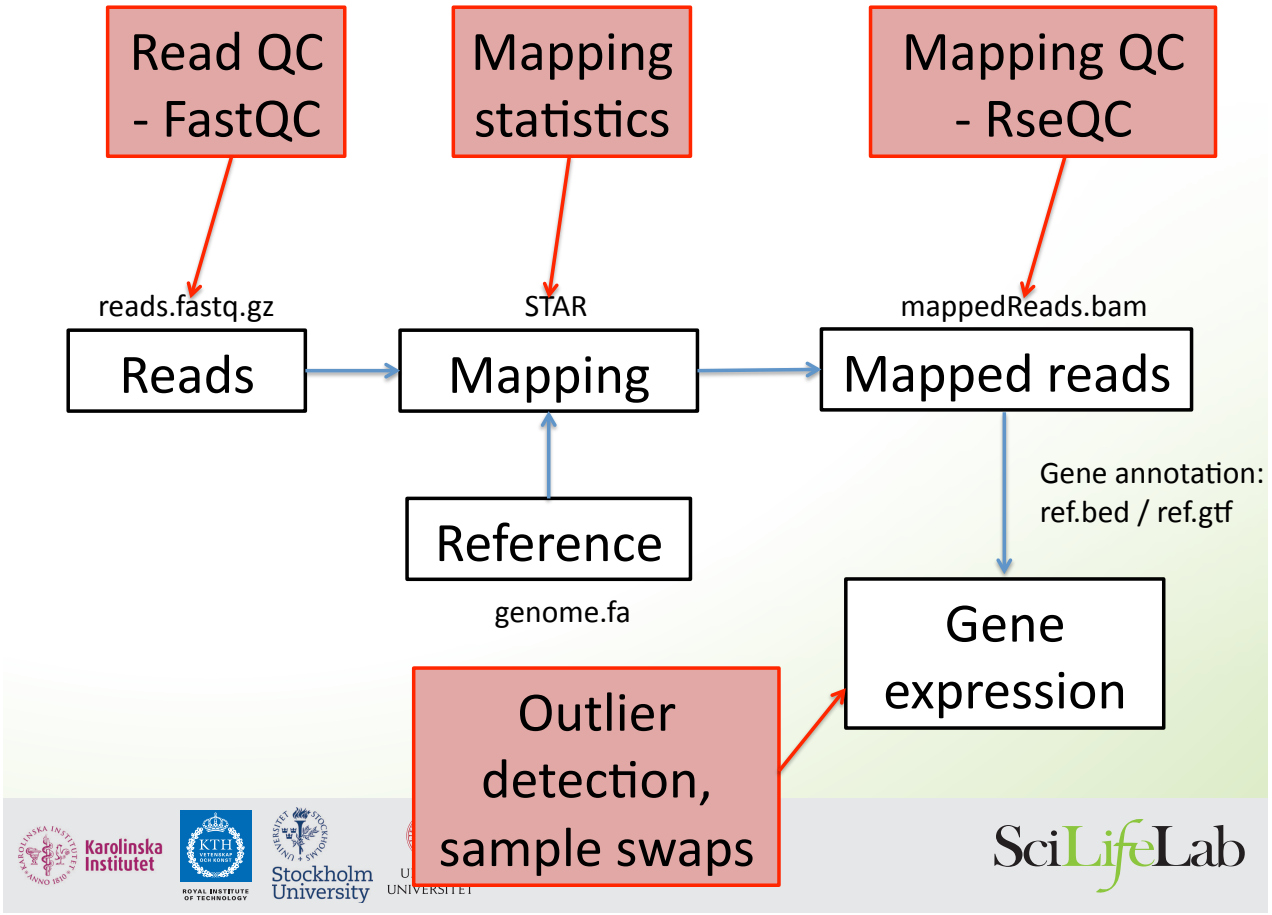
Probability that the nt call is wrong = $10^{-Phred/10}$

Per base sequence quality





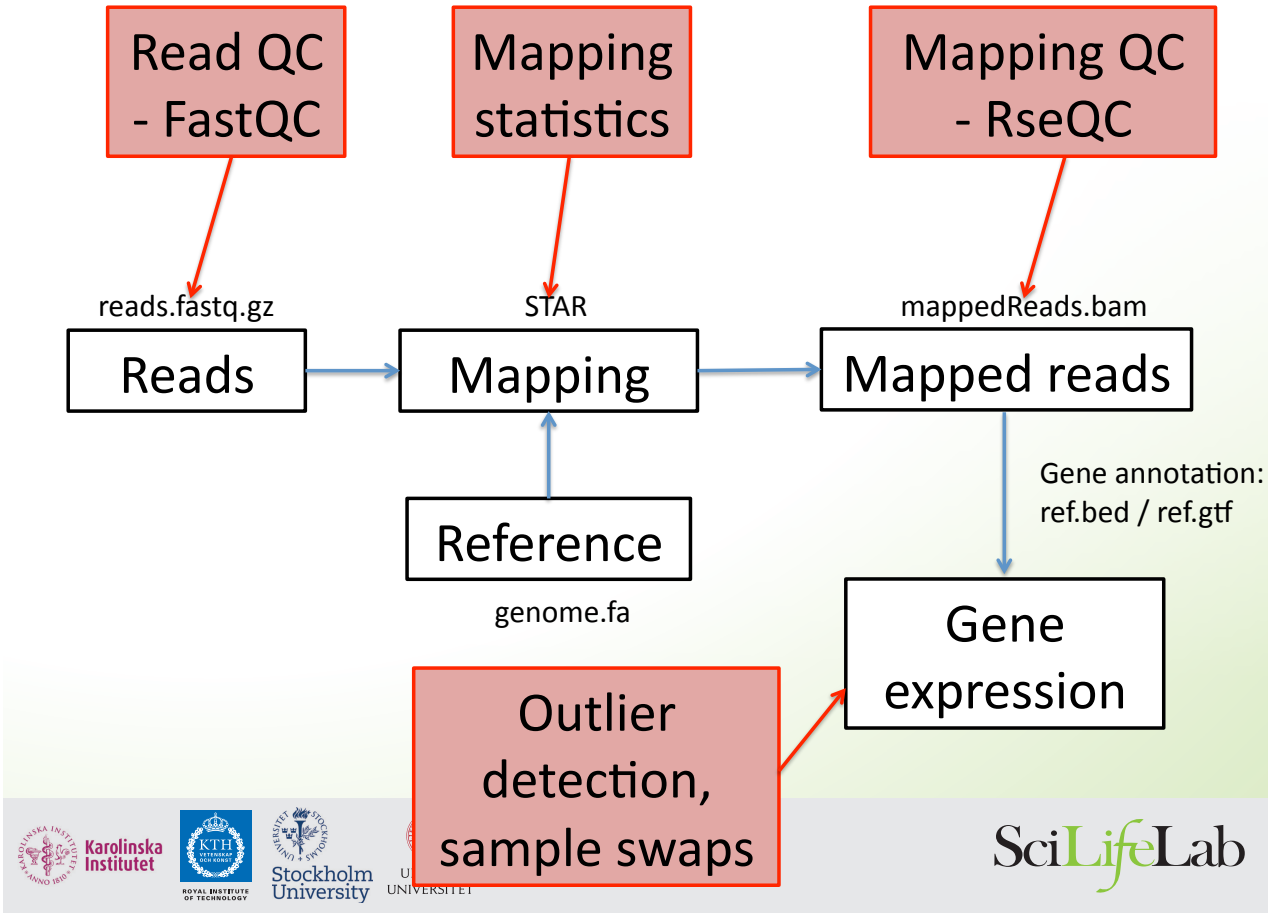
Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

Mapping (Johan Reimegård)

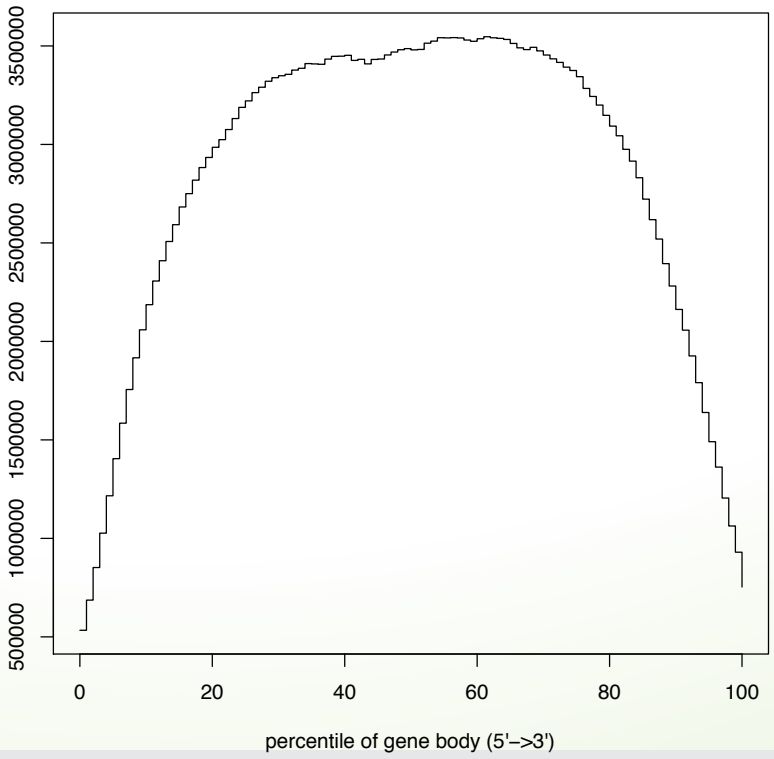
- Use a two-pass workflow
- STAR or HISAT
- If you want to run Cufflinks, use TopHat or HISAT
- For long (PacBio) reads, STAR, BLAT or GMAP can be used

Do a lot of QC

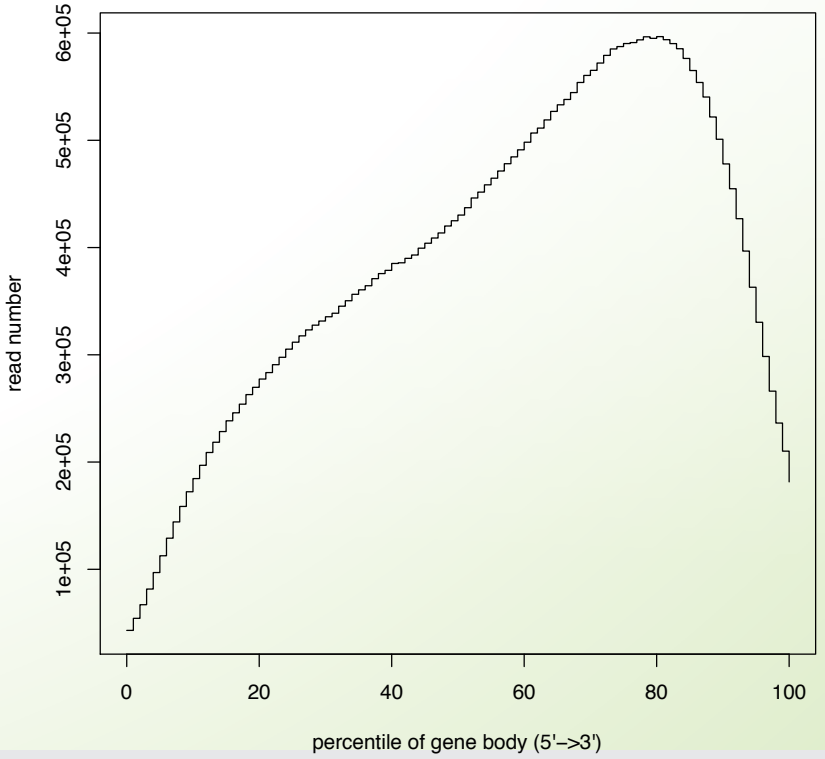

Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

RNA QC Åsa Björklund

Not degraded



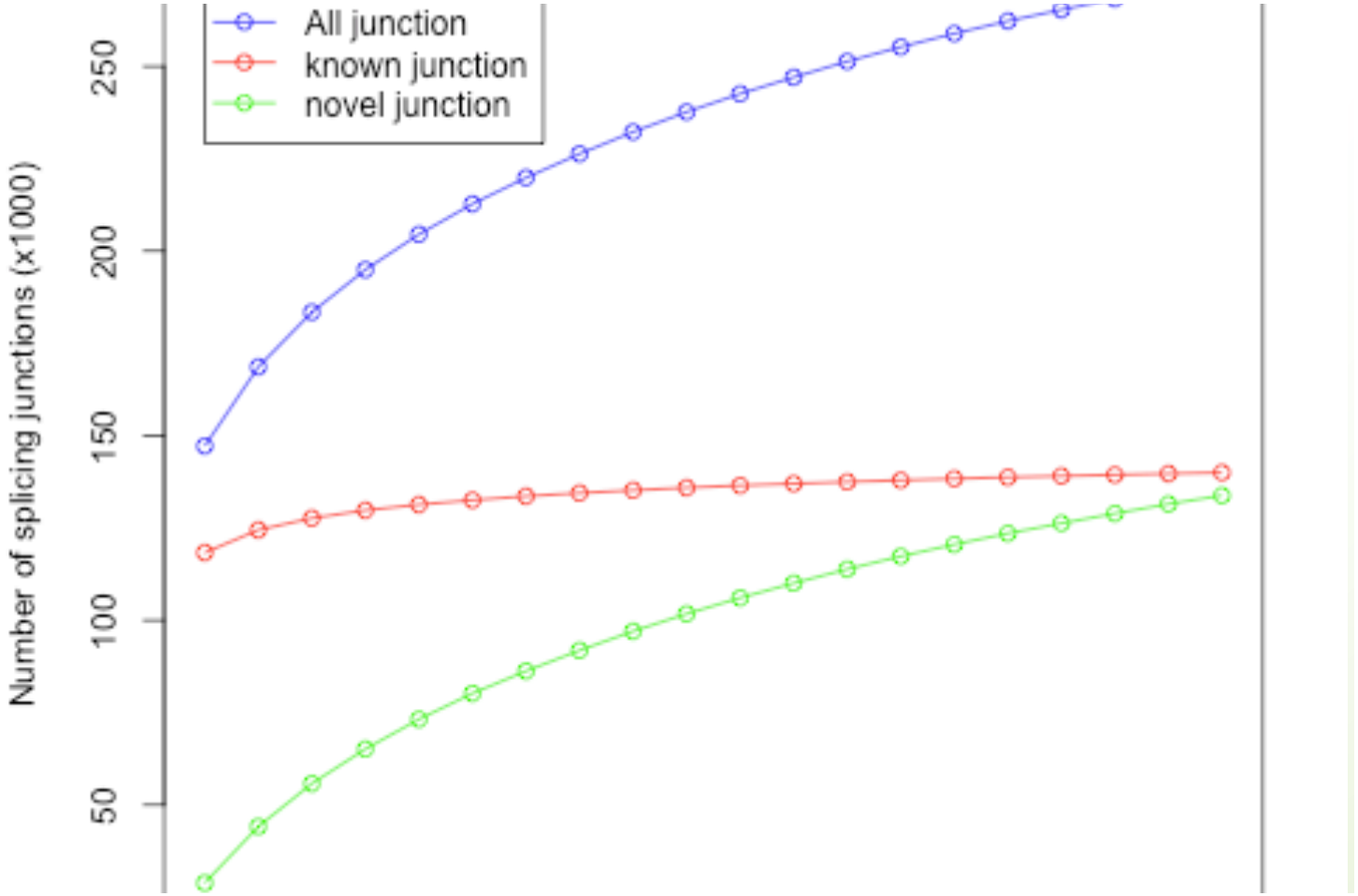
Degraded



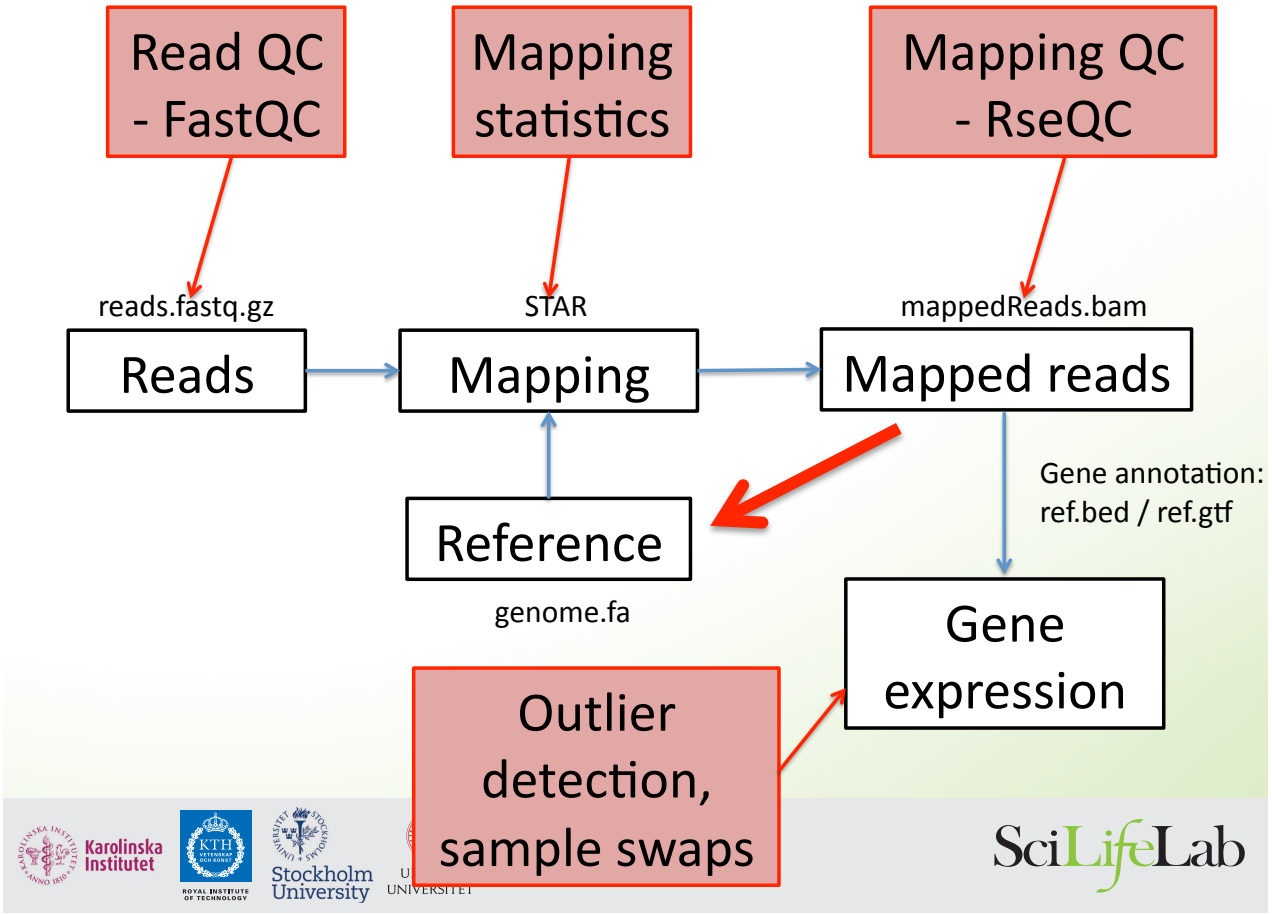
RNA QC Åsa Björklund

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

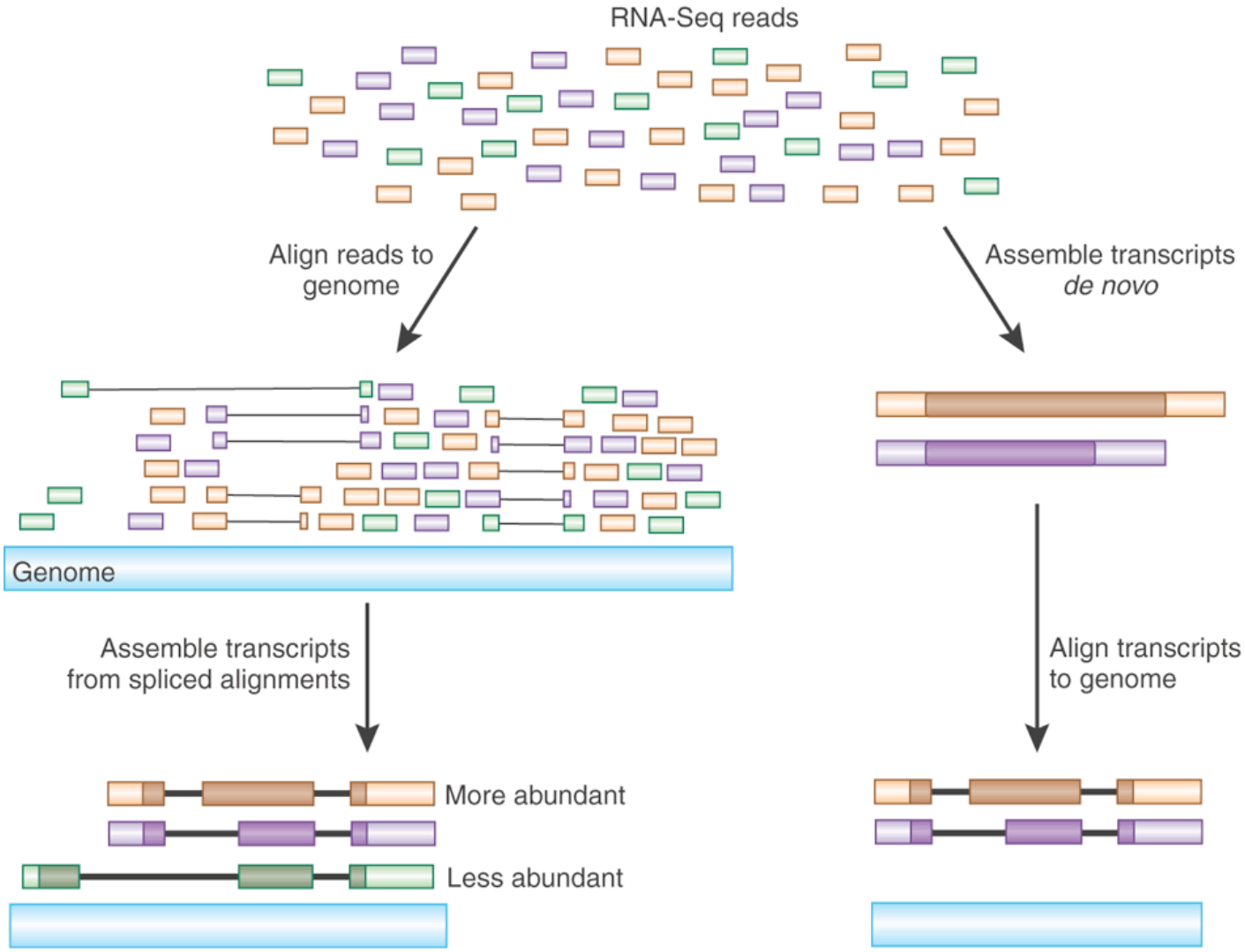
RNA QC Åsa Björklund



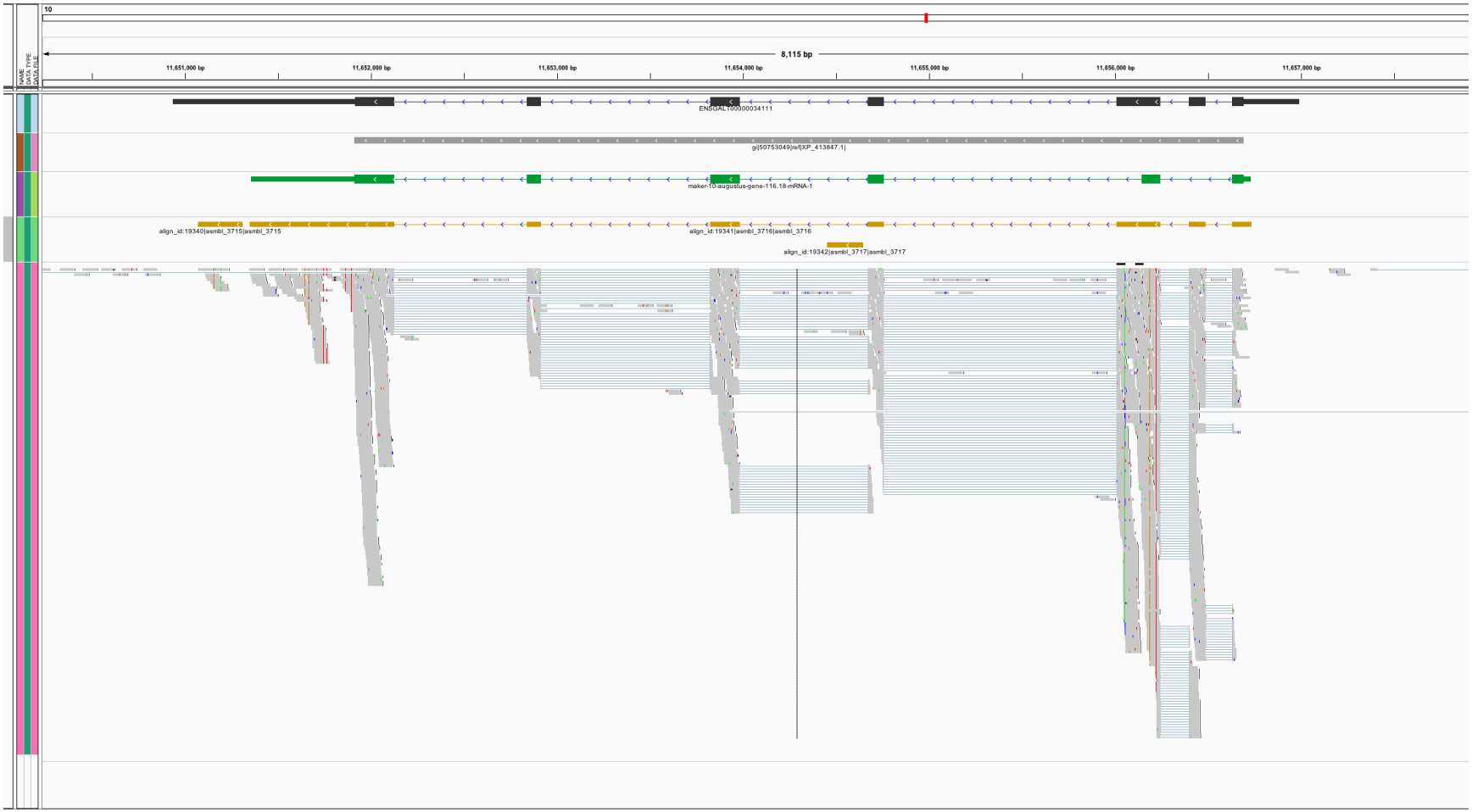
Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

Gene and Isoform detection and visualisation (Estelle)



Crappy annotation?

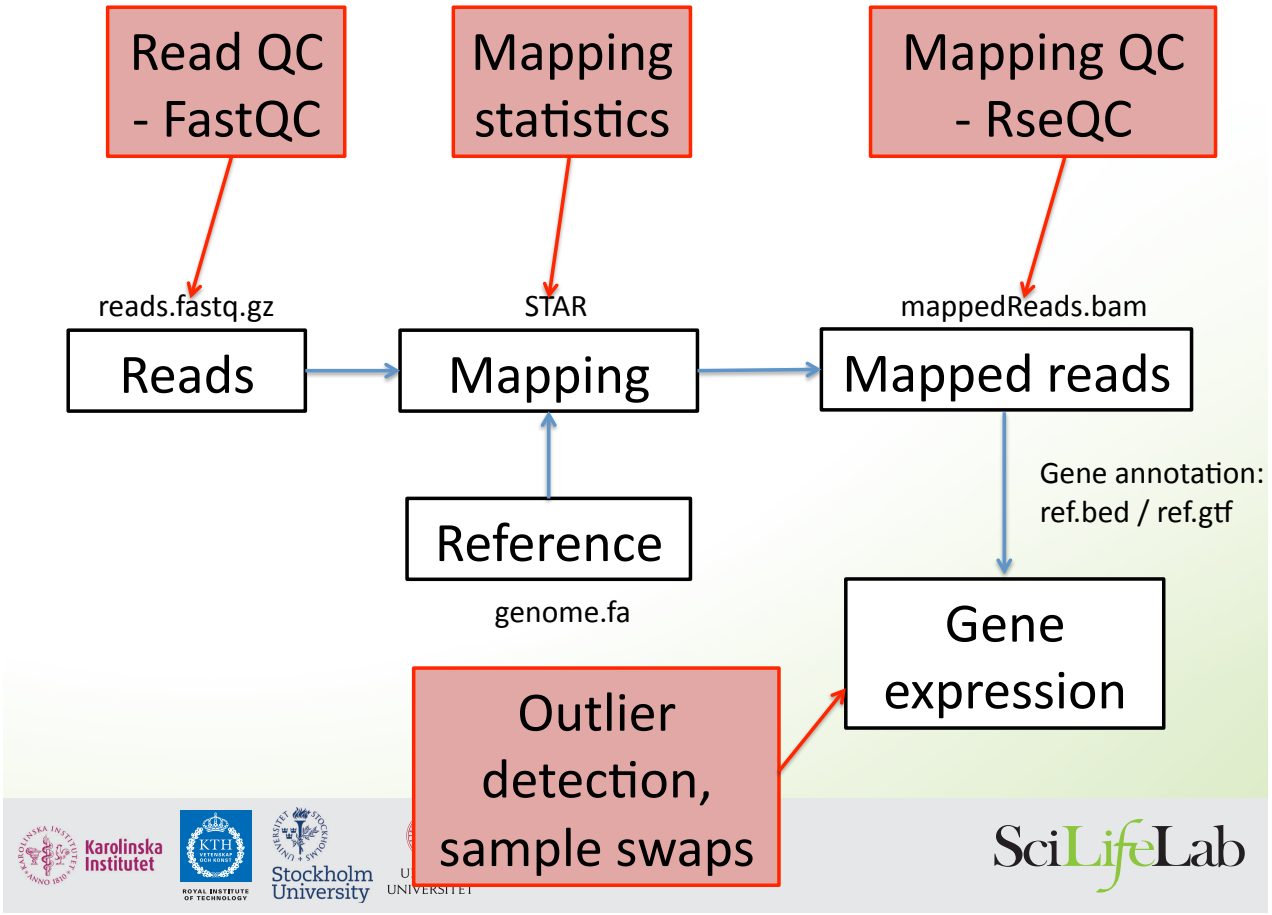


Make new annotation

- String tie or cufflinks

Combined annotation

Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

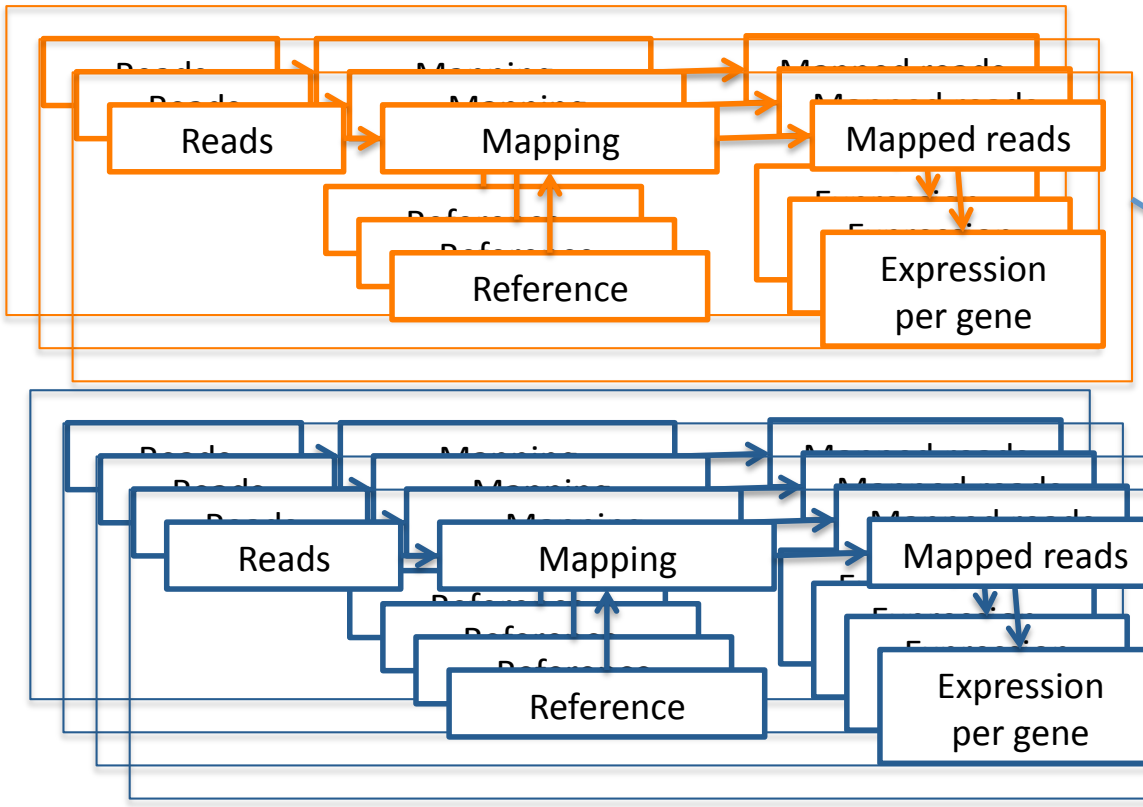
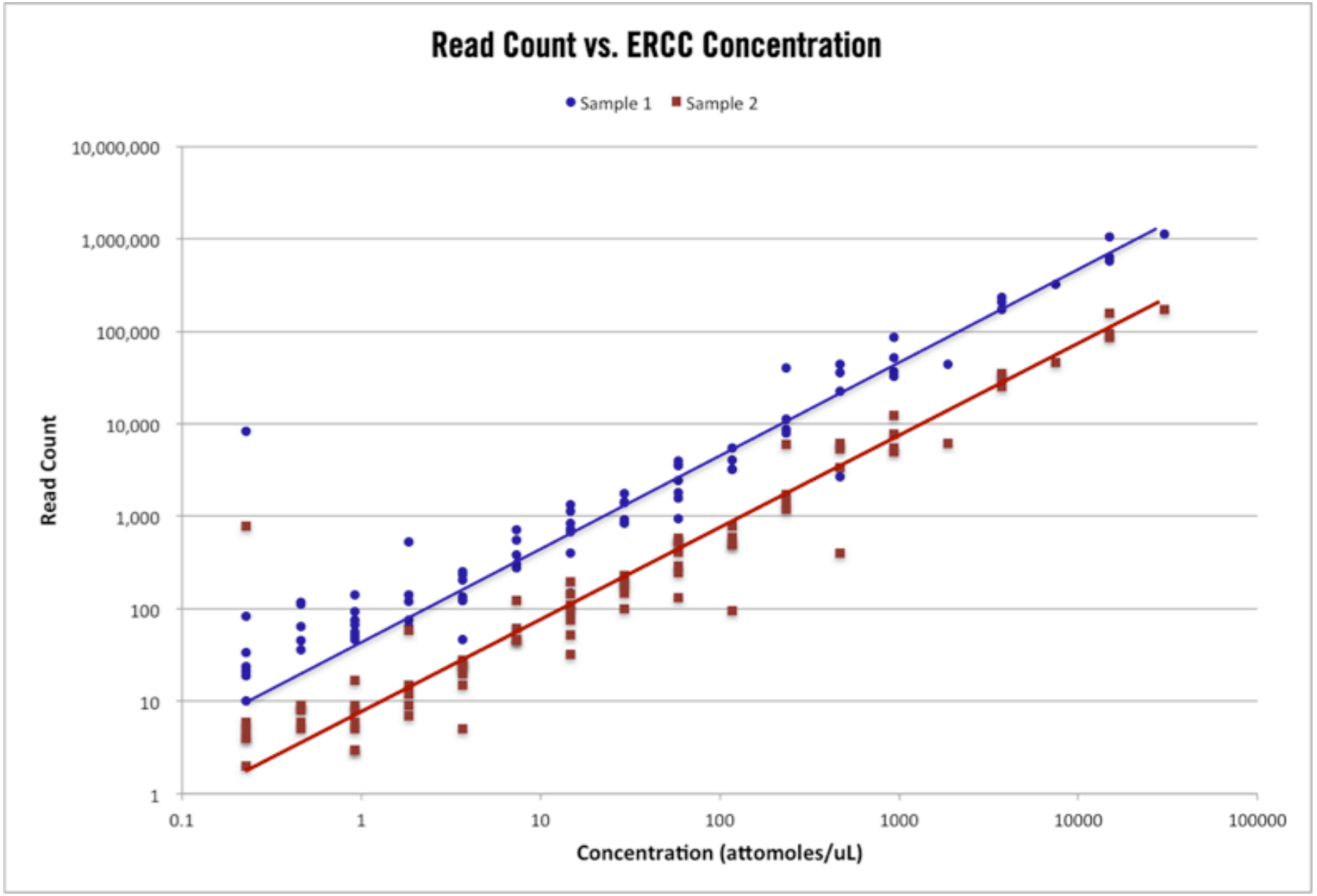


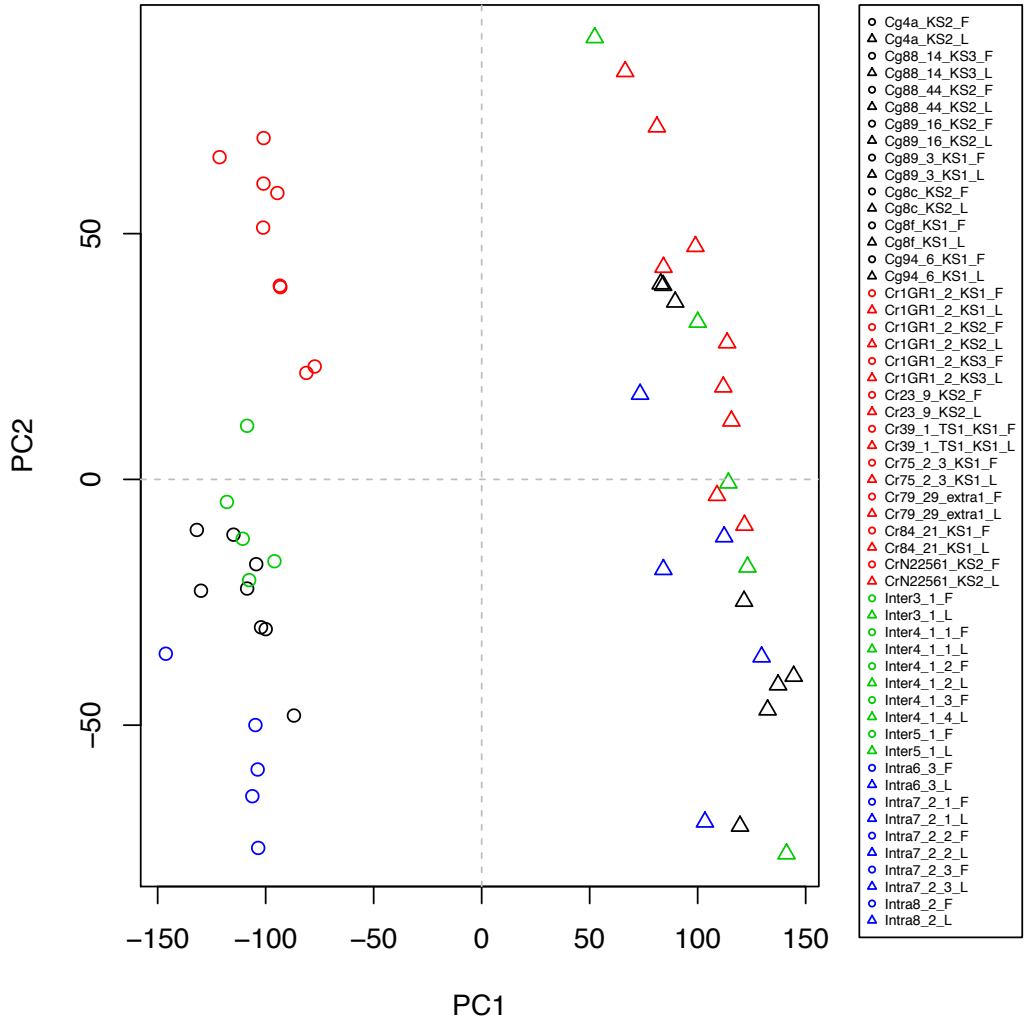
Table with counts, rpkms, fpkms or similar

Sample swaps and outliers can be identified using PCA



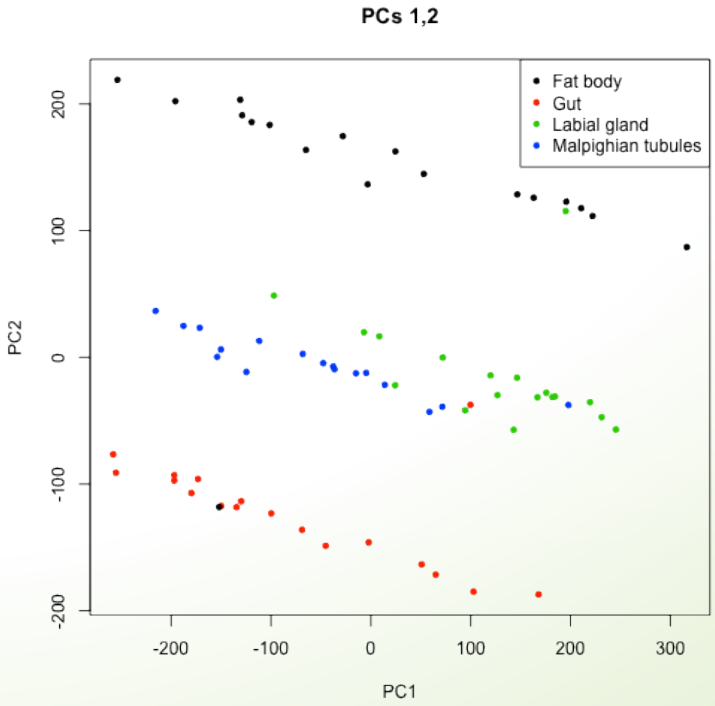
(https://cofactorgenomics.com)

Principal component 1 separates samples from flowers and leaves

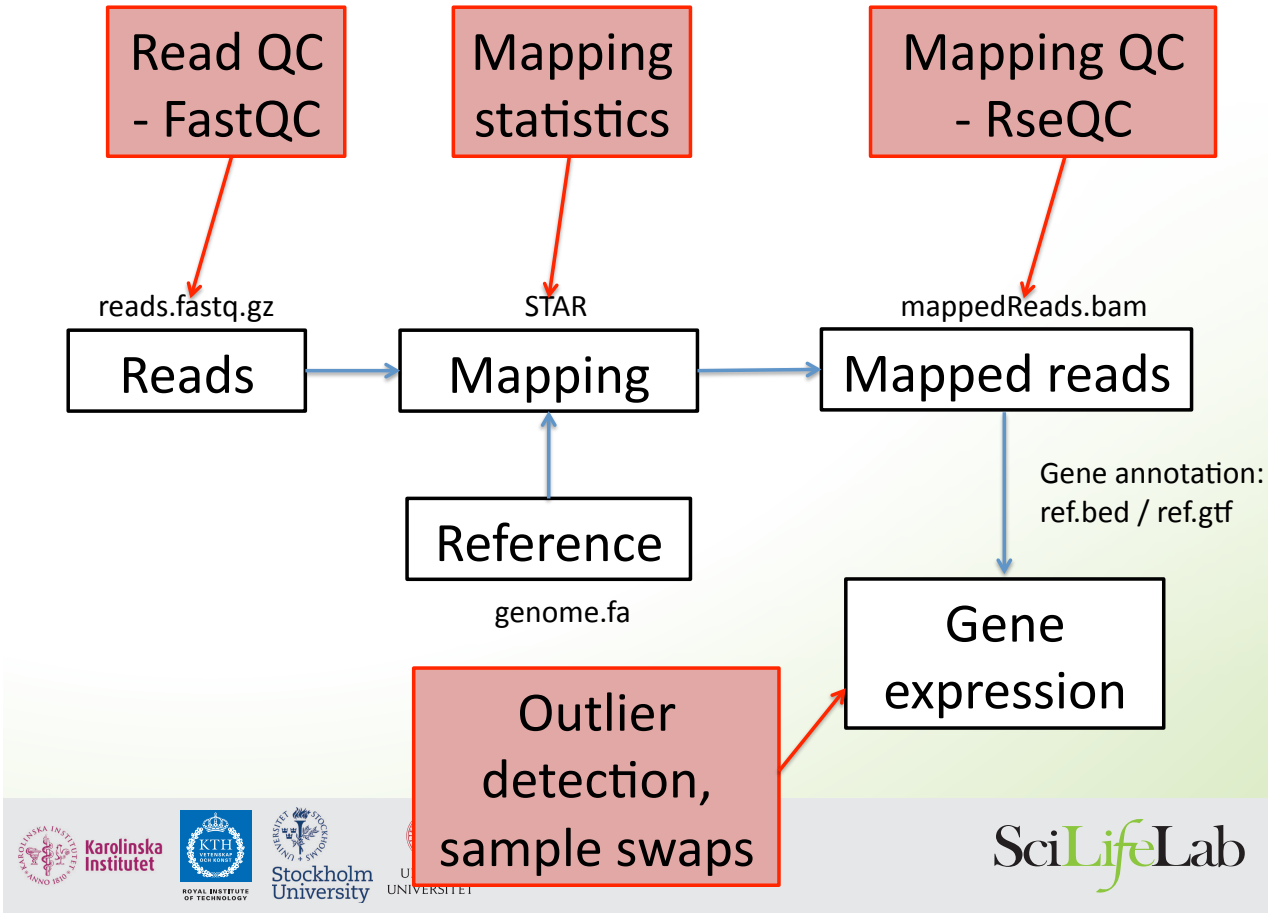


RNA QC Åsa Björklund

PCA analysis detected potential sample swaps



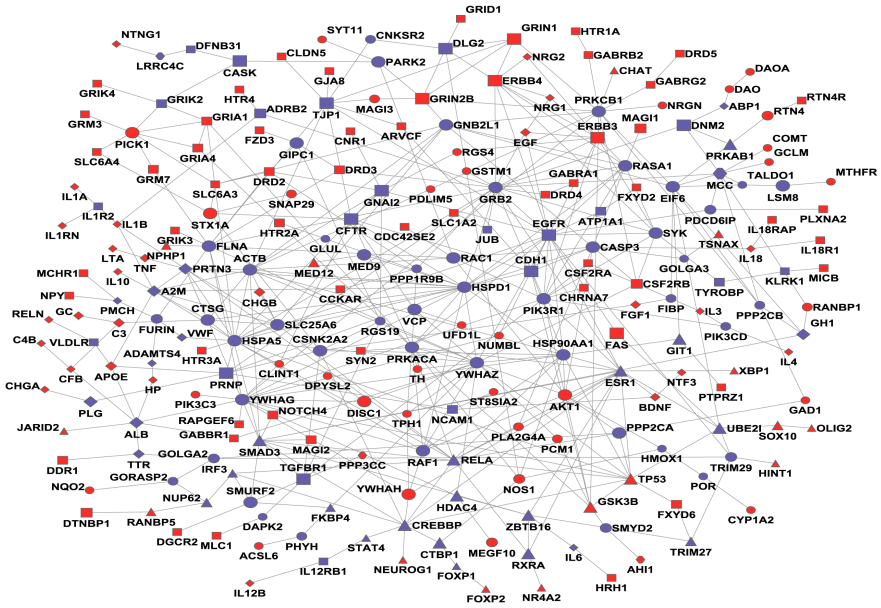
Do a lot of QC


Read QC
- FastQC
Mapping
statistics
Mapping QC
- RseQC
reads.fastq.gz
STAR
mappedReads.bam
Reads
Mapping
Mapped reads
Gene annotation:
ref.bed / ref.gtf
Reference
genome.fa
Gene
expression
Outlier
detection,
sample swaps
Karolinska
Institutet
Stockholm
University
SciLifeLab

Differential expression analysis
Mikael Huss

The identification of genes (or other types of genomic features, such as transcripts or exons) that are expressed in significantly different quantities in distinct groups of samples, be it biological conditions (drug-treated vs. controls), diseased vs. healthy individuals, different tissues, different stages of development, or something else.

Typically **univariate** analysis (one gene at a time) – even though we know that genes are not independent



Decision tree for software selection
Åsa Björklund

Differentially expressed **exons** => *DEXSeq*

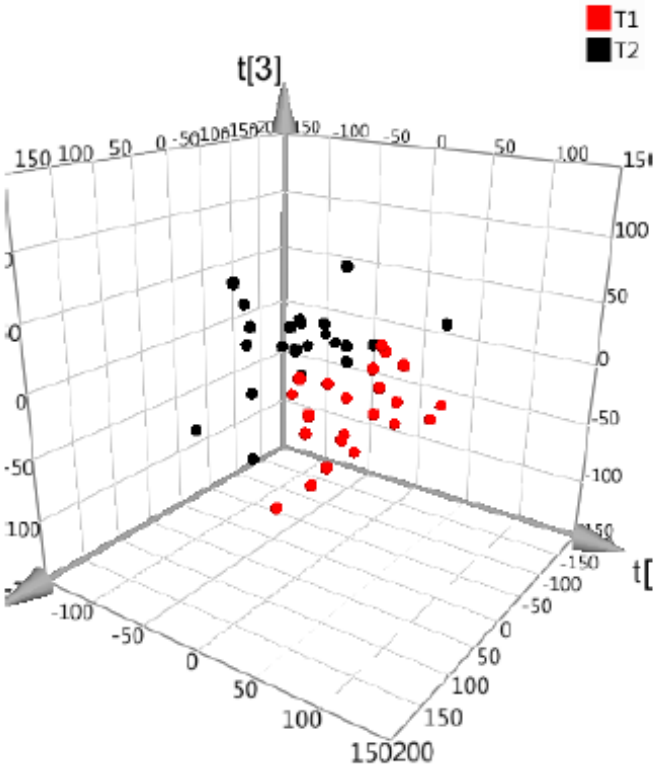
Differentially expressed **isoforms** => *BitSeq, Cuffdiff or ebSeq*

Differentially expressed genes => **Select type of experimental design**

- Complex design (more than one varying factor) => *DESeq, edgeR, limma*
- Simple comparison of groups => **How many biological replicates?**
  - More than about 5 biological replicates per group => *SAMSeq*
  - Less than 5 biological replicates per group => *DESeq, edgeR, limma*

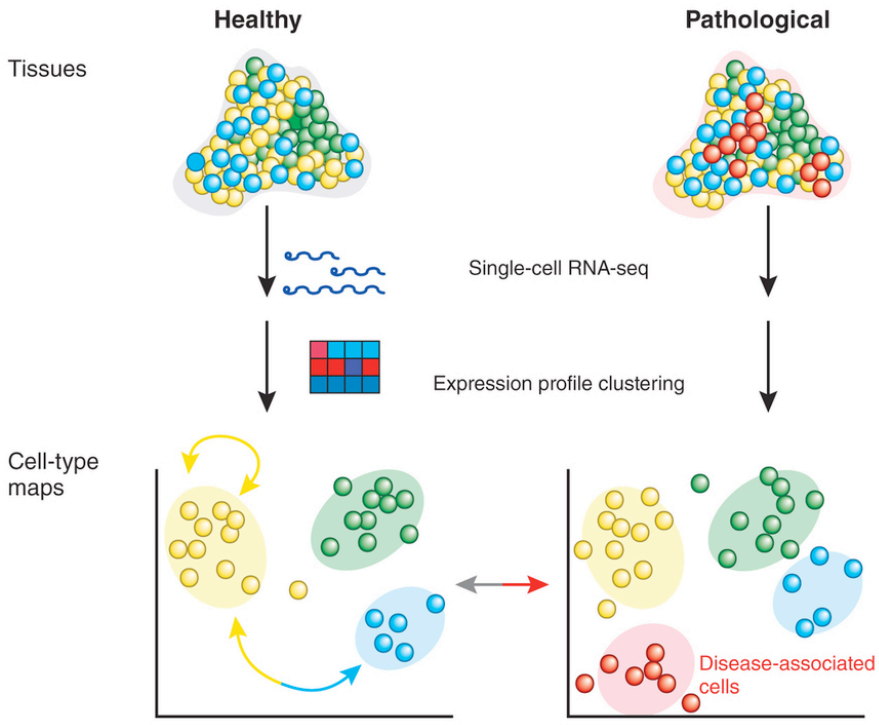
Beyond univariate differential expression
Sanela Kjellqvist

Multivariate methods such as PCA (unsupervised) or PLS (supervised) can be used to obtain loadings for features (genes/transcripts/…) that contribute to separation of groups



The loading scores can be used as a different kind of measure of which genes are interesting

Single cell sequencing Åsa Björklund



Types of analyses

Within cell type
- Stochasticity, variability of transcription
- Regulatory network inference
- Allelic expression patterns
- Scaling laws of transcription

Between cell types
- Identify biomarkers
- (Post)-transcriptional differences

Between tissues
- Cell-type compositions
- Altered transcription in matched cell types

(Sandberg, Nature Methods 2014)

Need help??

- We are here for you. Apply for help.