MSIB32500 Advanced Bioinformatics Spring 2019

RNAseq Data Analysis and Clinical Applications, Part II

Riyue Bao, Ph.D.

Research Assistant Professor (Bioinformatics)

Associate Director of Cancer Immunology (Bioinformatics)

Center for Research Informatics & Department of Pediatrics

The University of Chicago

Outline

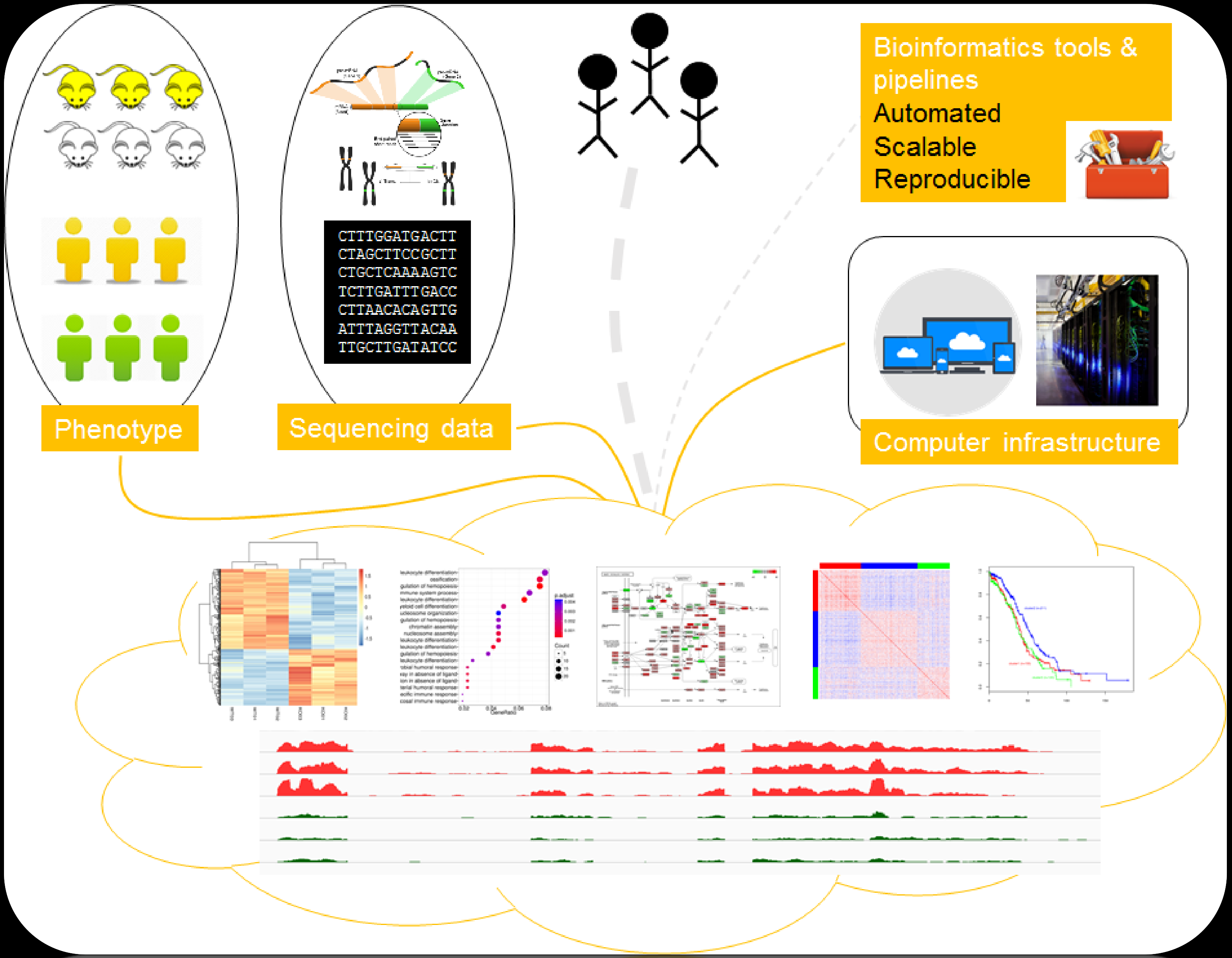
- Part I (05/25/2019)
  - Introduction to RNAseq technology and clinical applications
  - Hands on: From raw data to gene expression quantification
- Part II (06/01/2019)
  - Differential gene expression analysis and data visualization
  - Hands on: Identification of genes and pathways significantly changed under condition
  - ***Homework assignment (DUE 06/08/2019 11:59 PM; and optional tasks)***
- Part III (06/08/2019)
  - How to associate gene expression data with clinical outcome
  - Hands on: Use gene expression data to discover tumor subtypes and survival analysis

Class materials

- GitHub
  - https://github.com/MScBiomedicalInformatics/MSIB32500
  - This lecture note contains the same contents as the notebook. In addition, the notebook also contains hands-on materials
  - **lecture9.pdf**
  - **Handson9.Rmd**
- Rstudio (or R console) on personal computers (hands on practice)

Objective

- *(Recap from last class)*
- Detect genes differentially expressed between conditions
- Identify pathways / network enriched in genes of interest
- Generate high-quality figures for publication (PCA, heatmap, sample/gene cluster, GO/pathways, etc.)
- Become familiar with running commands in R / Rstudio



Phenotype

Sequencing data

```
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
```

Bioinformatics tools & pipelines

Automated

Scalable

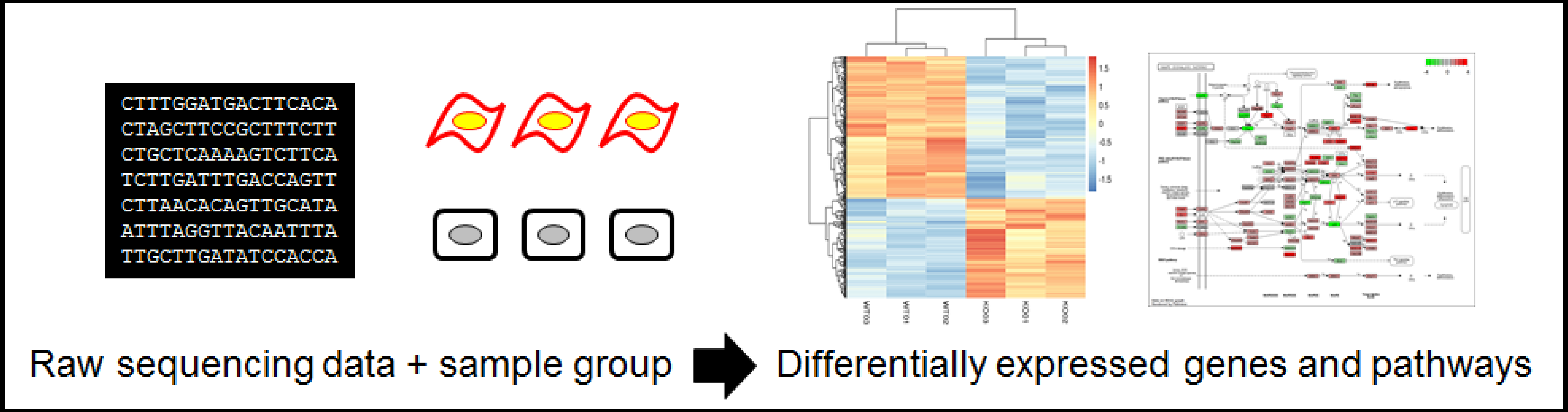
Reproducible

Computer infrastructure

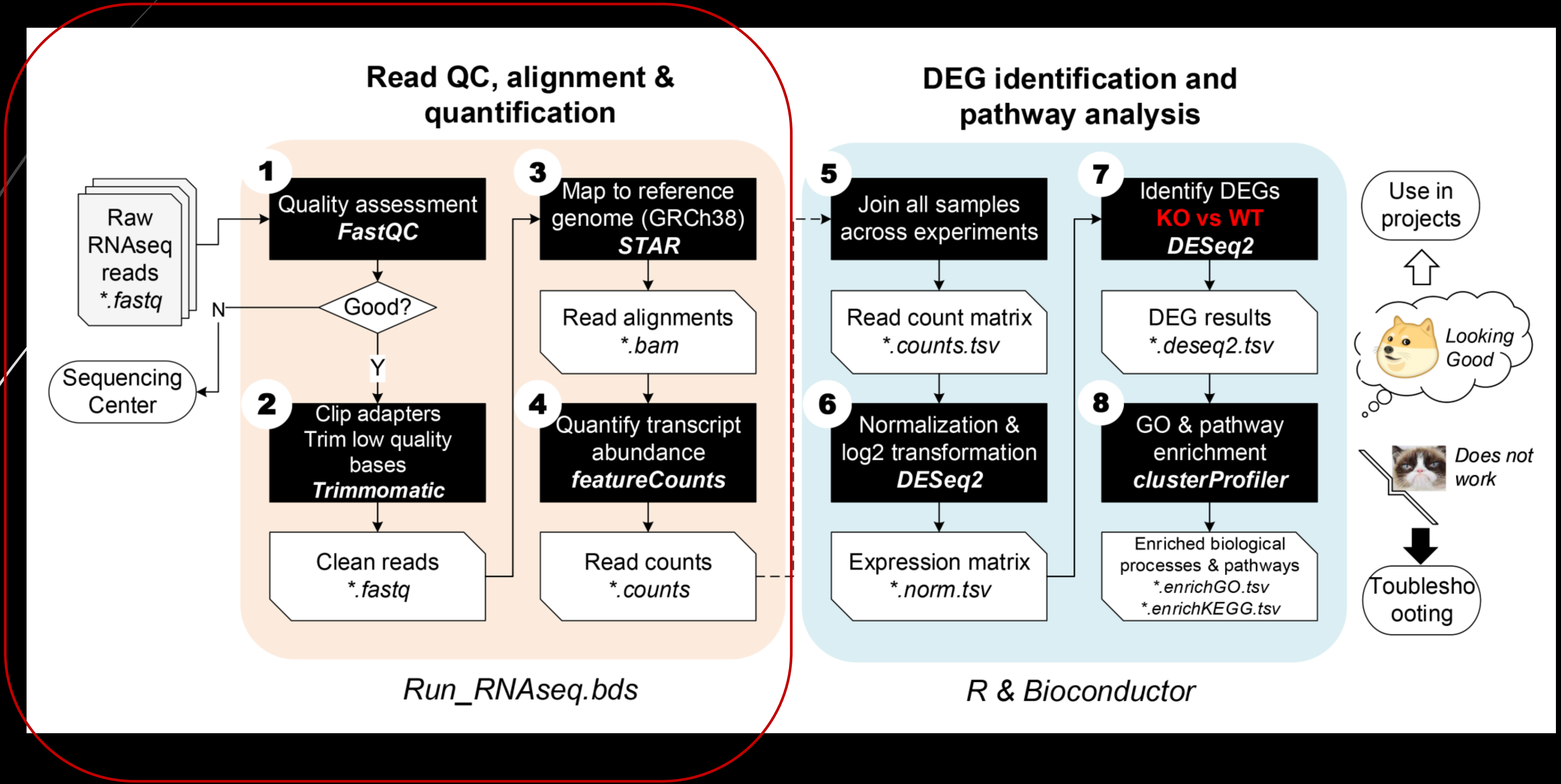
How to perform RNAseq analysis

The good-practice analysis protocol takes 8 major steps.

- **01-04**: From raw sequencing to transcript quantification
- **05-08**: DEG and pathway analysis (06/01, part II)



How to perform RNAseq analysis



Read QC, alignment & quantification

Raw RNAseq reads **.fastq*

1 Quality assessment ***FastQC***

Good?

N

Sequencing Center

Y

2 Clip adapters Trim low quality bases ***Trimmomatic***

Clean reads **.fastq*

3 Map to reference genome (GRCh38) ***STAR***

Read alignments **.bam*

4 Quantify transcript abundance ***featureCounts***

Read counts **.counts*

Run_RNAseq.bds

DEG identification and pathway analysis

5 Join all samples across experiments

Read count matrix **.counts.tsv*

6 Normalization & log2 transformation ***DESeq2***

Expression matrix **.norm.tsv*

7 Identify DEGs **KO vs WT** ***DESeq2***

DEG results **.deseq2.tsv*

8 GO & pathway enrichment ***clusterProfiler***

Enriched biological processes & pathways **.enrichGO.tsv* **.enrichKEGG.tsv*

R & Bioconductor

Use in projects

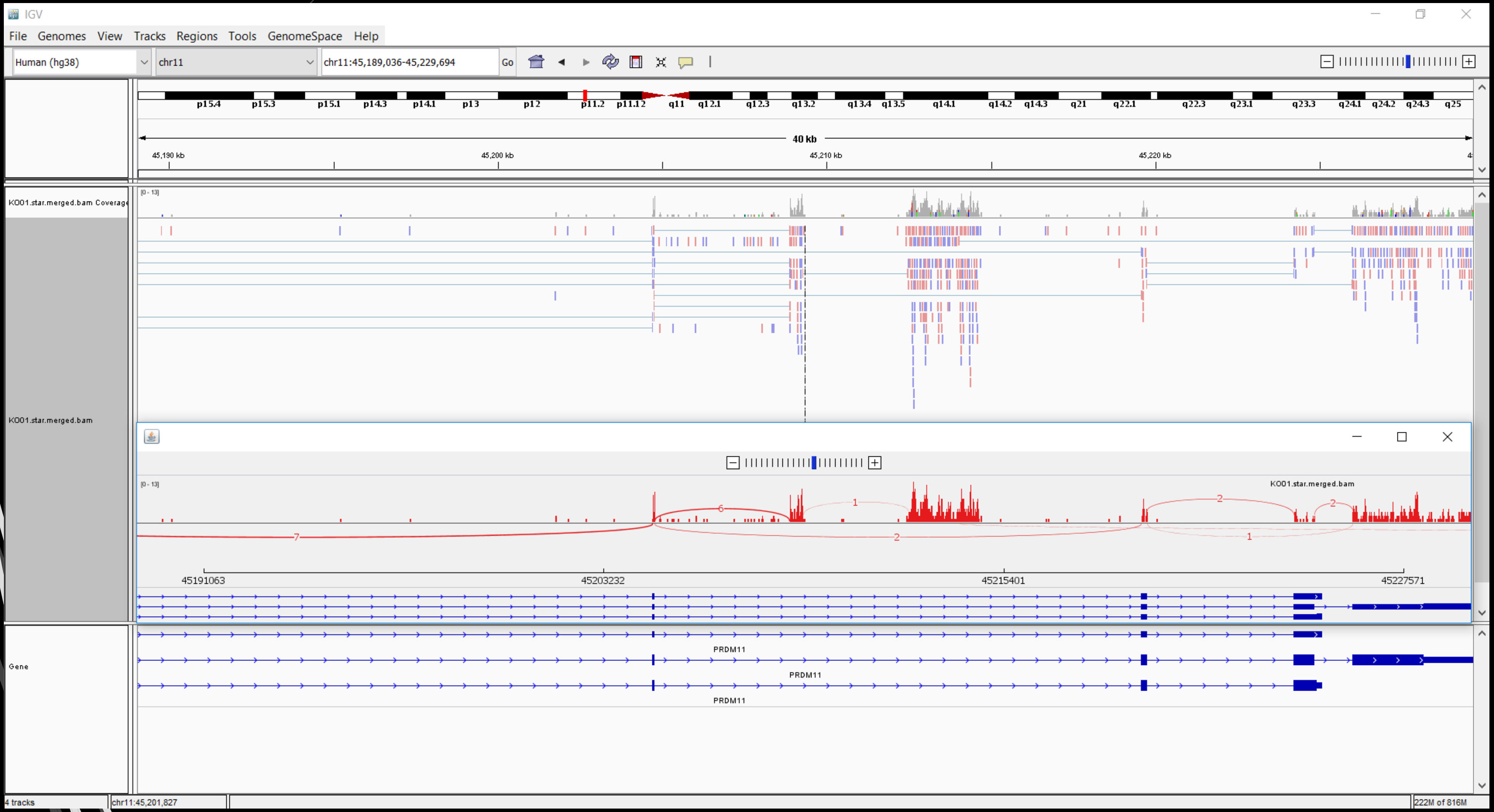
Looking Good

Does not work

Toubleshooting

IGV (Integrative Genome Viewer)

http://software.broadinstitute.org/software/igv/home


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Go
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.12 q11 q12.1 q12.3 q13.2 q13.4 q13.5 q14.1 q14.2 q14.3 q21 q22.1 q22.3 q23.1 q23.3 q24.1 q24.2 q24.3 q25
40 kb
45,190 kb
45,200 kb
45,210 kb
45,220 kb
[0 - 13]
KO01.star.merged.bam Coverage
KO01.star.merged.bam
45191063
45203232
45215401
45227571
Gene
PRDM11
PRDM11
PRDM11
4 tracks
222M of 816M

Human (hg38)

chr11

chr11:45,189,036-45,229,694

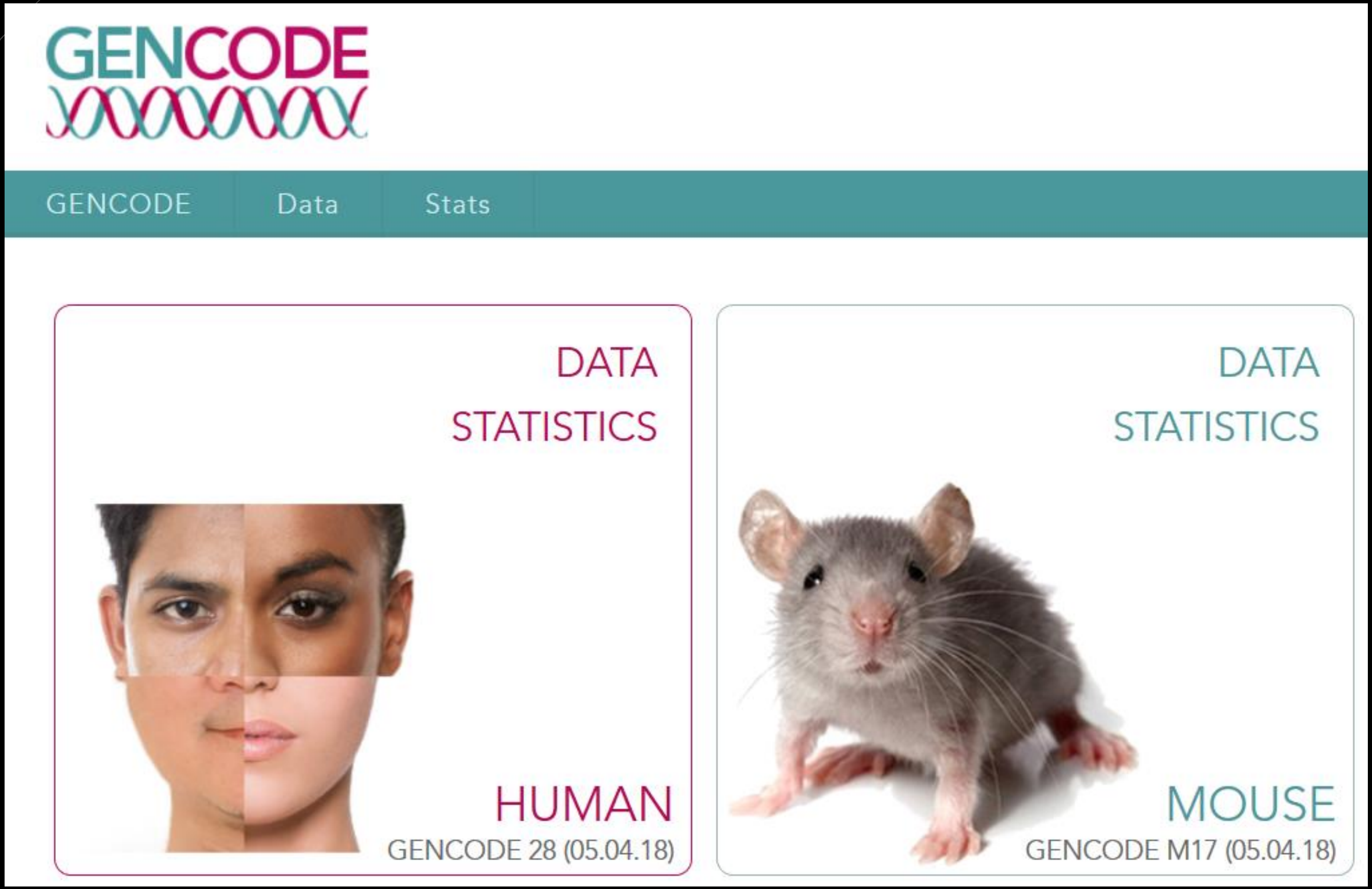
chr11:45,201,827

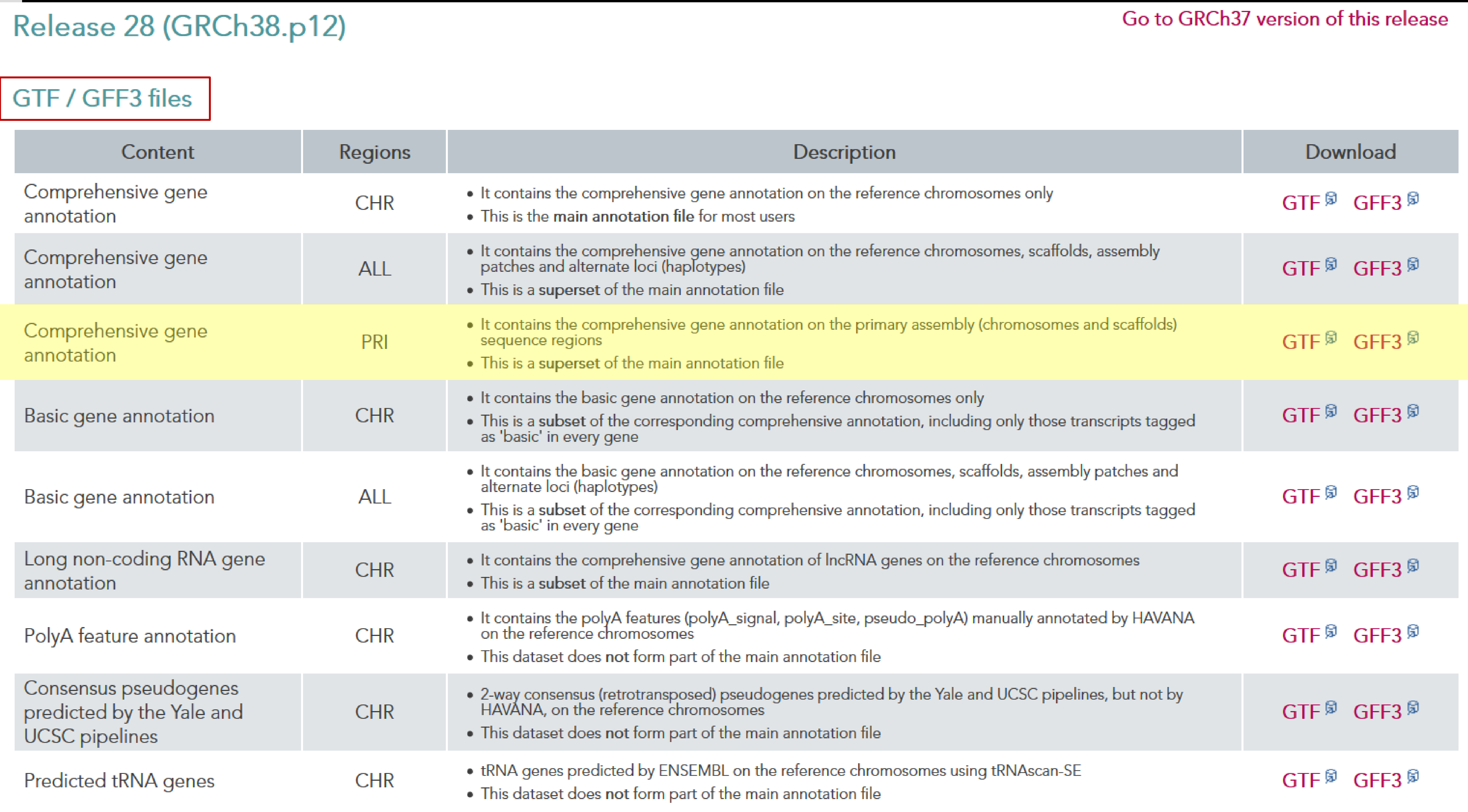
- Load existing genomes, or generate custom genomes
- Visualize standard file formats
  - BAM
  - BED
  - GTF
  - … and more!

Reference databases

- **Gene annotation database: GENCODE**
  - https://www.gencodegenes.org/
- **Ensembl database**
  - https://www.ensembl.org/index.html
- UCSC Genome Browser
  - https://genome.ucsc.edu/
- NCBI databases
  - https://www.ncbi.nlm.nih.gov/guide/genomes-maps/
- Genomic databases
  - GDC: https://portal.gdc.cancer.gov/
  - GTEx: https://gtexportal.org/home/
  - Single-cell RNAseq: https://portals.broadinstitute.org/single_cell

https://www.gencodegenes.org/ model organisms



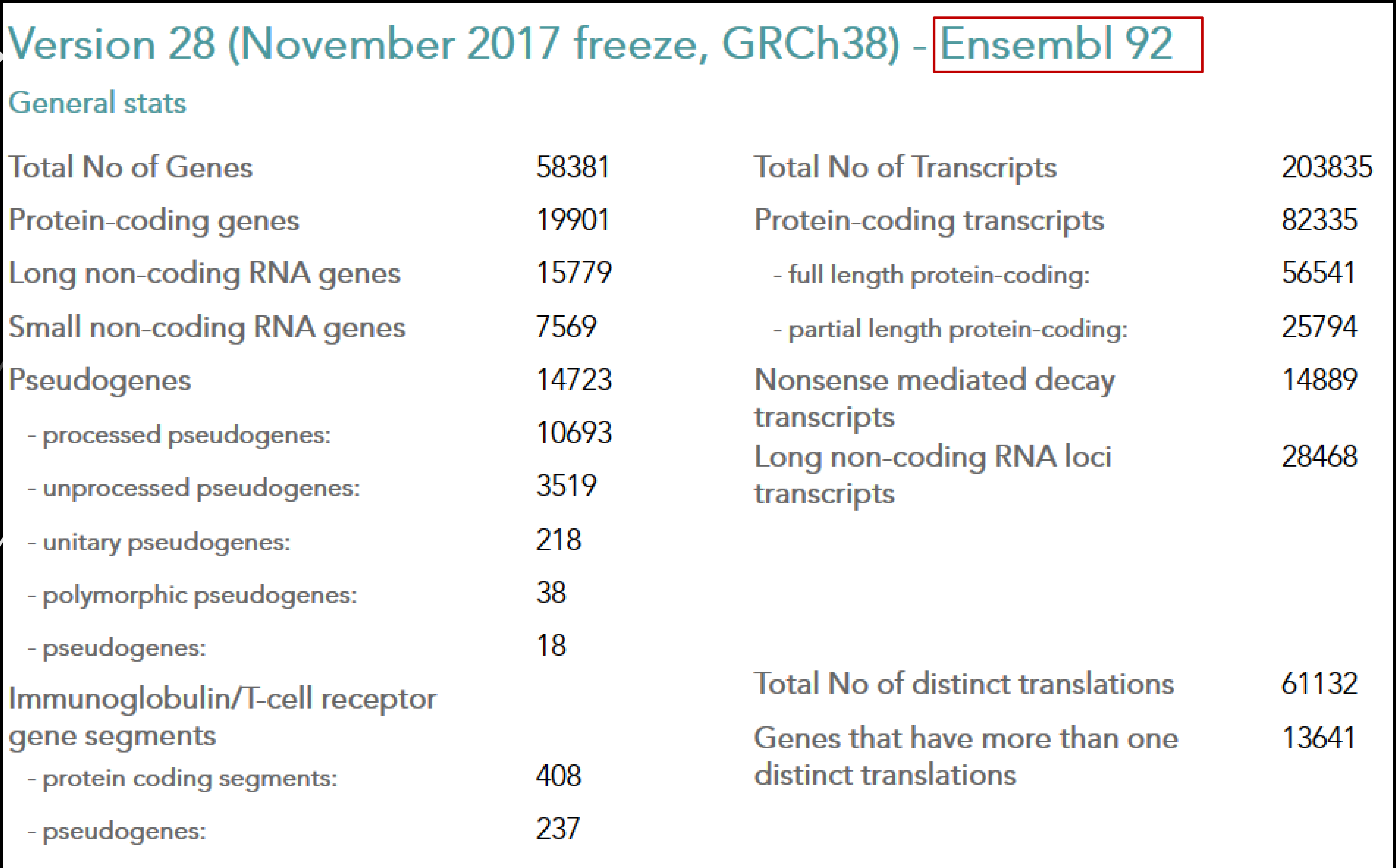


Release 28 (GRCh38.p12)

Go to GRCh37 version of this release

GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only • This is the **main annotation file** for most users	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **superset** of the main annotation file	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Long non-coding RNA gene annotation	CHR	• It contains the comprehensive gene annotation of lncRNA genes on the reference chromosomes • This is a **subset** of the main annotation file	GTF GFF3
PolyA feature annotation	CHR	• It contains the polyA features (polyA_signal, polyA_site, pseudo_polyA) manually annotated by HAVANA on the reference chromosomes • This dataset does **not** form part of the main annotation file	GTF GFF3
Consensus pseudogenes predicted by the Yale and UCSC pipelines	CHR	• 2-way consensus (retrotransposed) pseudogenes predicted by the Yale and UCSC pipelines, but not by HAVANA, on the reference chromosomes • This dataset does **not** form part of the main annotation file	GTF GFF3
Predicted tRNA genes	CHR	• tRNA genes predicted by ENSEMBL on the reference chromosomes using tRNAscan-SE • This dataset does **not** form part of the main annotation file	GTF GFF3

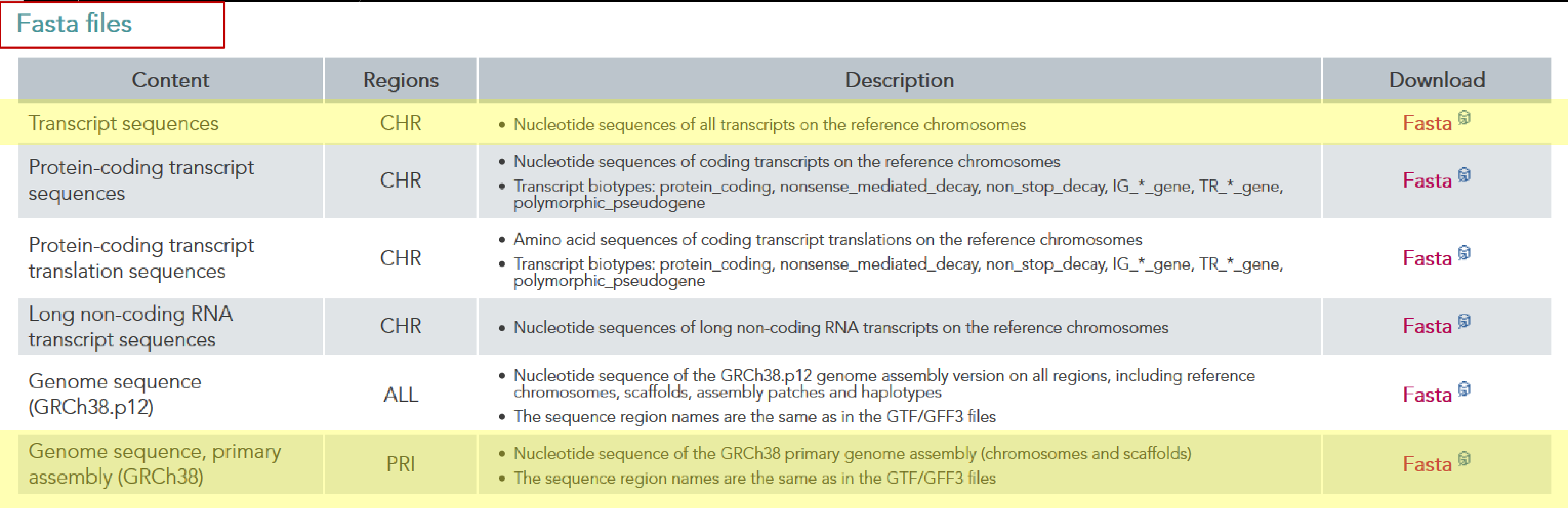


Version 28 (November 2017 freeze, GRCh38) - Ensembl 92

General stats

Total No of Genes	58381
Protein-coding genes	19901
Long non-coding RNA genes	15779
Small non-coding RNA genes	7569
Pseudogenes	14723
- processed pseudogenes:	10693
- unprocessed pseudogenes:	3519
- unitary pseudogenes:	218
- polymorphic pseudogenes:	38
- pseudogenes:	18
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	408
- pseudogenes:	237

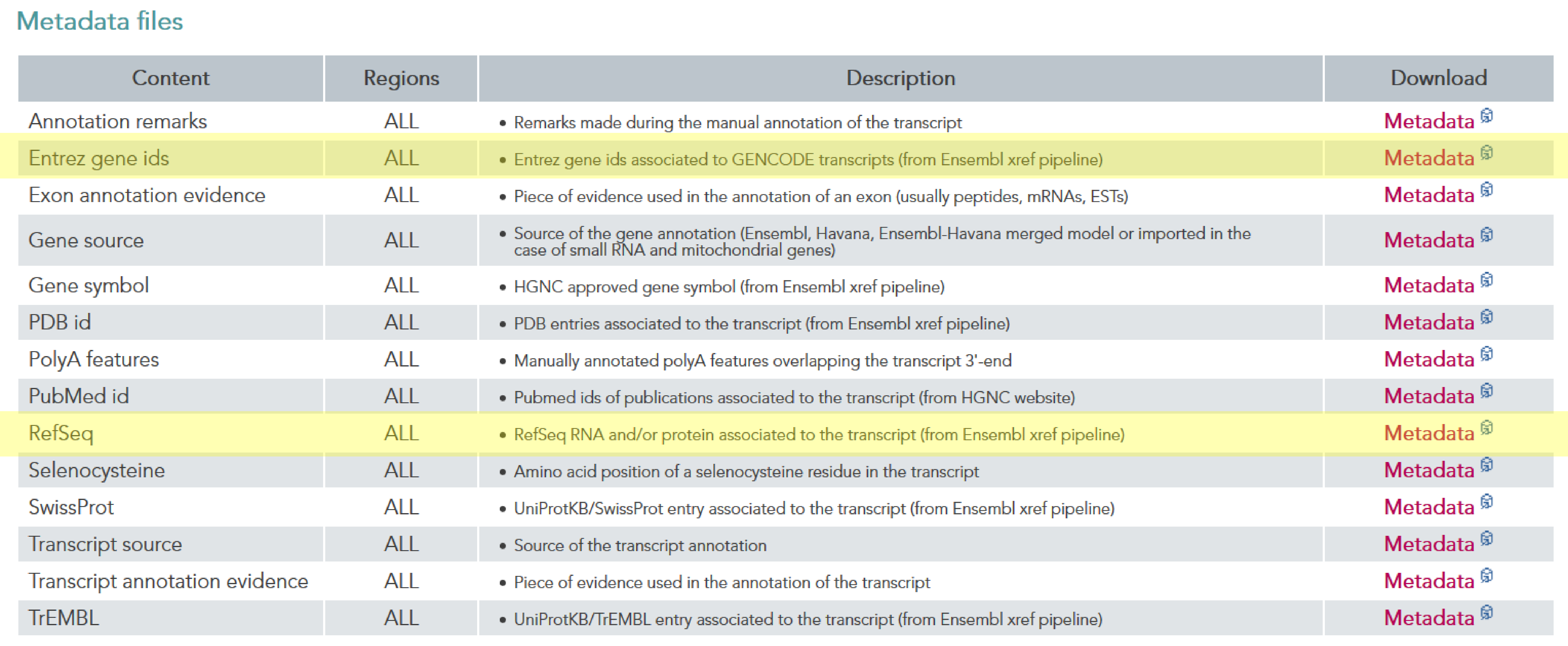
Total No of Transcripts	203835
Protein-coding transcripts	82335
- full length protein-coding:	56541
- partial length protein-coding:	25794
Nonsense mediated decay transcripts	14889
Long non-coding RNA loci transcripts	28468
Total No of distinct translations	61132
Genes that have more than one distinct translations	13641



Fasta files

Content	Regions	Description	Download
Transcript sequences	CHR	• Nucleotide sequences of all transcripts on the reference chromosomes	Fasta
Protein-coding transcript sequences	CHR	• Nucleotide sequences of coding transcripts on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Protein-coding transcript translation sequences	CHR	• Amino acid sequences of coding transcript translations on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Long non-coding RNA transcript sequences	CHR	• Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes	Fasta
Genome sequence (GRCh38.p12)	ALL	• Nucleotide sequence of the GRCh38.p12 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes • The sequence region names are the same as in the GTF/GFF3 files	Fasta
Genome sequence, primary assembly (GRCh38)	PRI	• Nucleotide sequence of the GRCh38 primary genome assembly (chromosomes and scaffolds) • The sequence region names are the same as in the GTF/GFF3 files	Fasta

- Genome-based alignment: STAR
- Transcriptome-based (pseudo)alignment: Kallisto, Salmon

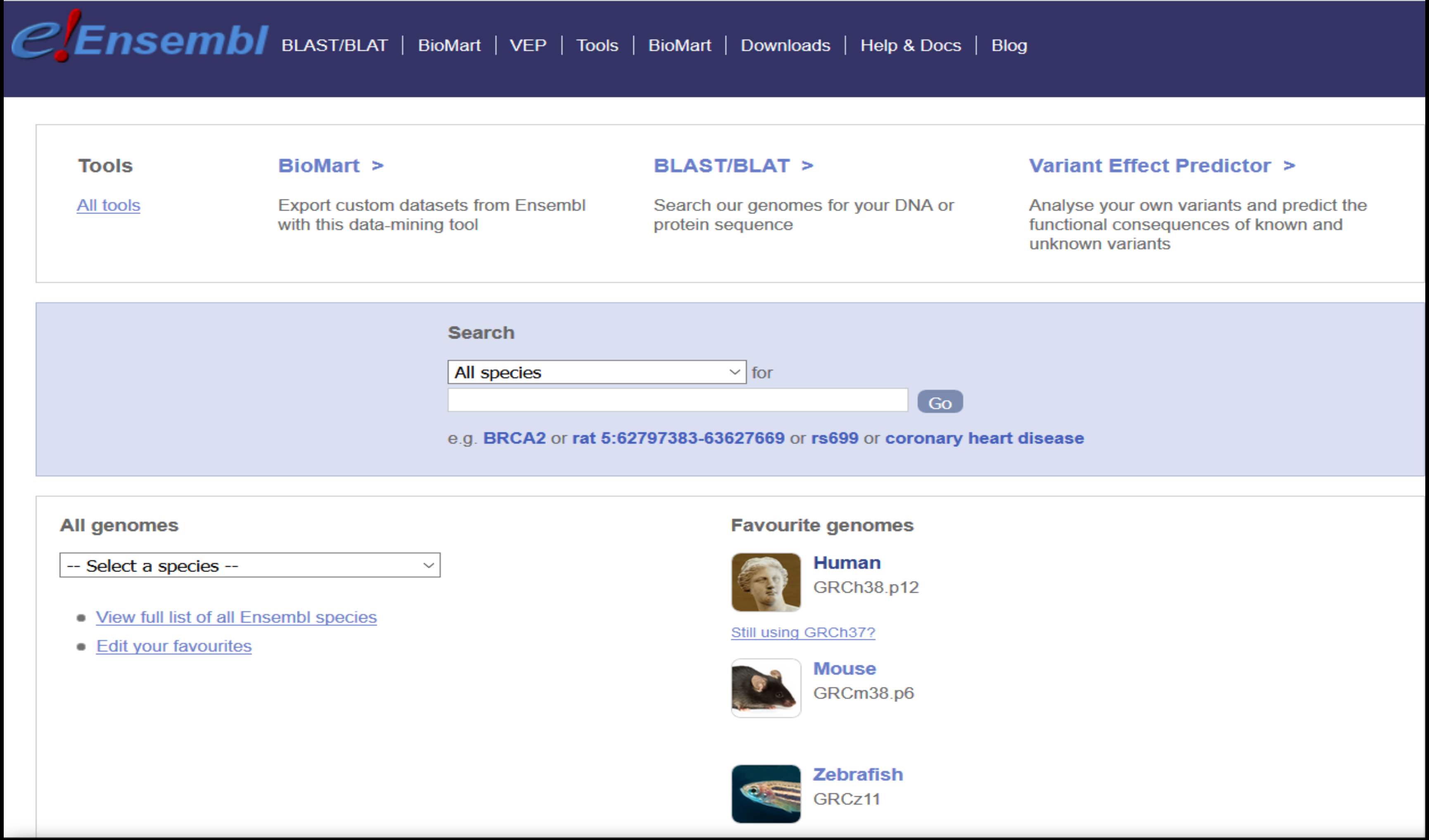


Metadata files

Content	Regions	Description	Download
Annotation remarks	ALL	• Remarks made during the manual annotation of the transcript	Metadata
Entrez gene ids	ALL	• Entrez gene ids associated to GENCODE transcripts (from Ensembl xref pipeline)	Metadata
Exon annotation evidence	ALL	• Piece of evidence used in the annotation of an exon (usually peptides, mRNAs, ESTs)	Metadata
Gene source	ALL	• Source of the gene annotation (Ensembl, Havana, Ensembl-Havana merged model or imported in the case of small RNA and mitochondrial genes)	Metadata
Gene symbol	ALL	• HGNC approved gene symbol (from Ensembl xref pipeline)	Metadata
PDB id	ALL	• PDB entries associated to the transcript (from Ensembl xref pipeline)	Metadata
PolyA features	ALL	• Manually annotated polyA features overlapping the transcript 3'-end	Metadata
PubMed id	ALL	• Pubmed ids of publications associated to the transcript (from HGNC website)	Metadata
RefSeq	ALL	• RefSeq RNA and/or protein associated to the transcript (from Ensembl xref pipeline)	Metadata
Selenocysteine	ALL	• Amino acid position of a selenocysteine residue in the transcript	Metadata
SwissProt	ALL	• UniProtKB/SwissProt entry associated to the transcript (from Ensembl xref pipeline)	Metadata
Transcript source	ALL	• Source of the transcript annotation	Metadata
Transcript annotation evidence	ALL	• Piece of evidence used in the annotation of the transcript	Metadata
TrEMBL	ALL	• UniProtKB/TrEMBL entry associated to the transcript (from Ensembl xref pipeline)	Metadata

ID conversion between different annotation databases (e.g. NCBI/RefSeq, Ensembl)

https://www.ensembl.org/index.html all organisms


e!Ensembl
BLAST/BLAT | BioMart | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
All species
for
Go
-- Select a species --

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

All genomes

- View full list of all Ensembl species
- Edit your favourites

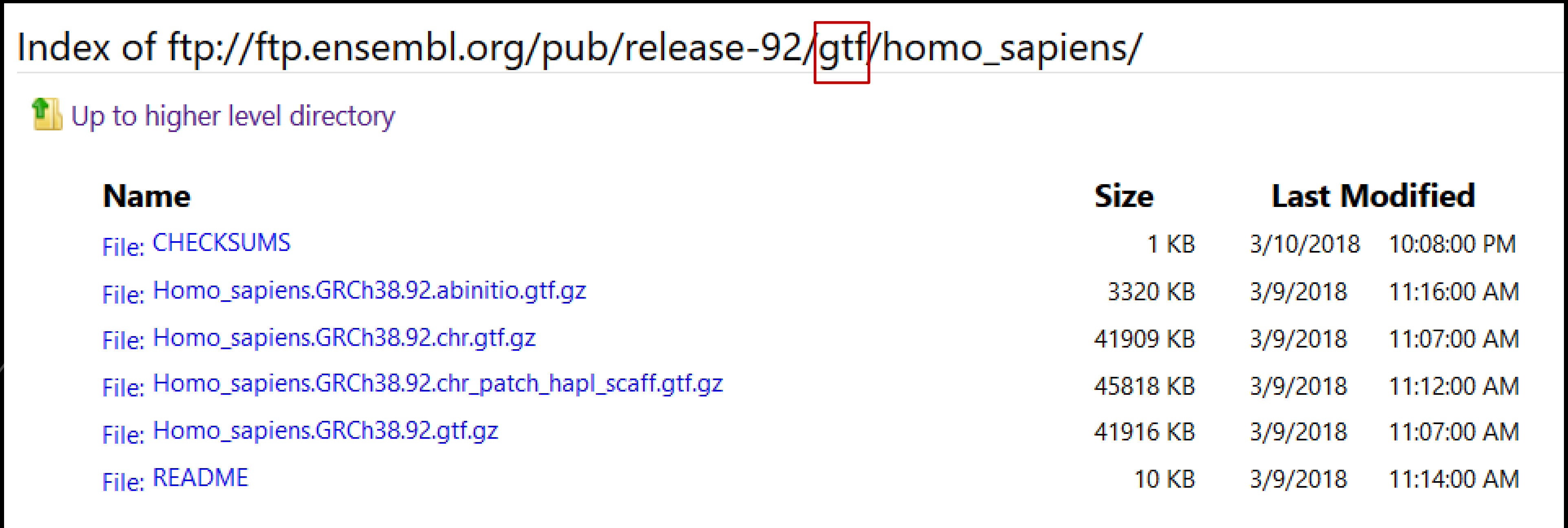
Favourite genomes

Human
GRCh38.p12

Still using GRCh37?

Mouse
GRCm38.p6

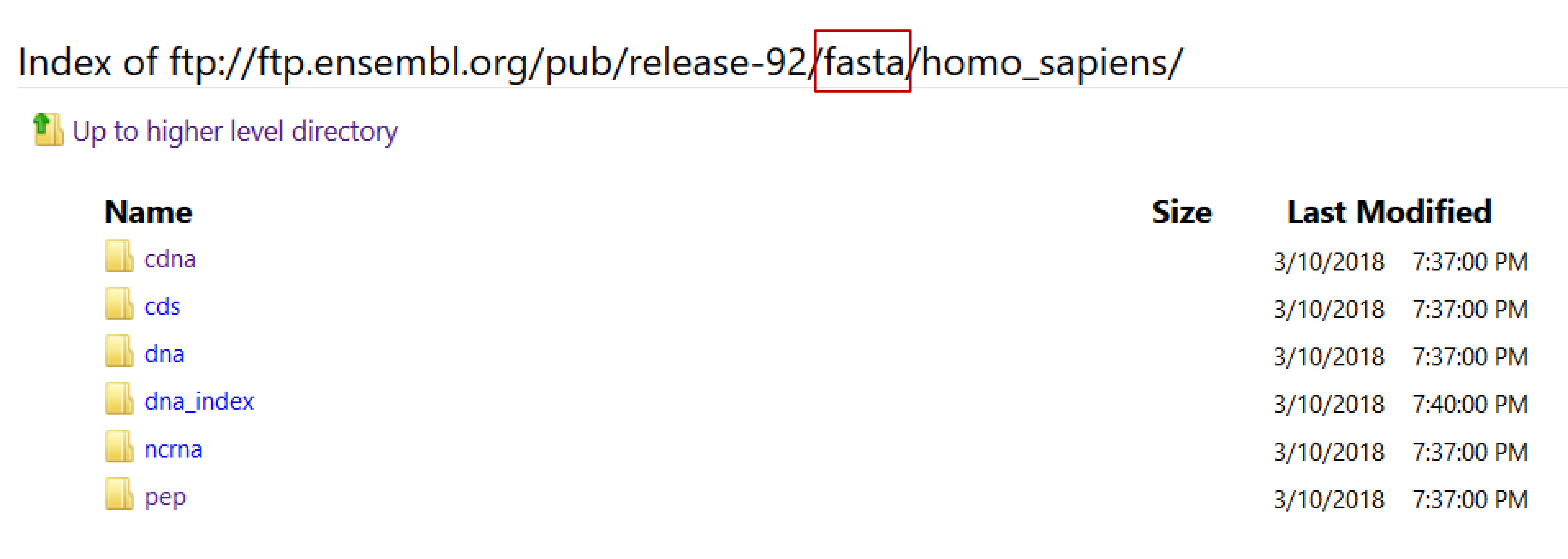
Zebrafish
GRCz11



Index of ftp://ftp.ensembl.org/pub/release-92/gtf/homo_sapiens/

Up to higher level directory

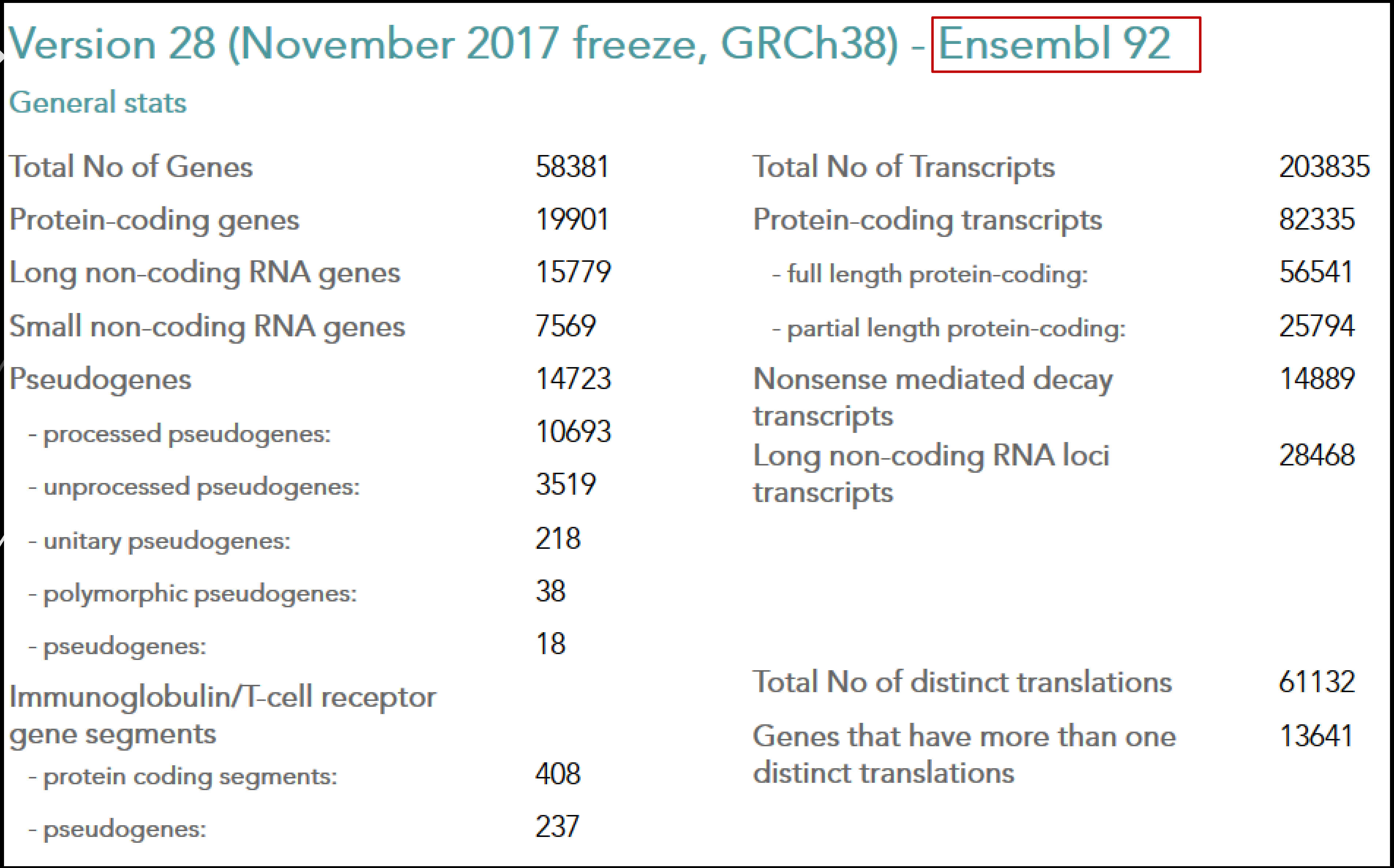
Name	Size	Last Modified	
File: CHECKSUMS	1 KB	3/10/2018	10:08:00 PM
File: Homo_sapiens.GRCh38.92.abinitio.gtf.gz	3320 KB	3/9/2018	11:16:00 AM
File: Homo_sapiens.GRCh38.92.chr.gtf.gz	41909 KB	3/9/2018	11:07:00 AM
File: Homo_sapiens.GRCh38.92.chr_patch_hapl_scaff.gtf.gz	45818 KB	3/9/2018	11:12:00 AM
File: Homo_sapiens.GRCh38.92.gtf.gz	41916 KB	3/9/2018	11:07:00 AM
File: README	10 KB	3/9/2018	11:14:00 AM



Index of ftp://ftp.ensembl.org/pub/release-92/fasta/homo_sapiens/

Up to higher level directory

Name	Size	Last Modified	
cdna		3/10/2018	7:37:00 PM
cds		3/10/2018	7:37:00 PM
dna		3/10/2018	7:37:00 PM
dna_index		3/10/2018	7:40:00 PM
ncrna		3/10/2018	7:37:00 PM
pep		3/10/2018	7:37:00 PM



Version 28 (November 2017 freeze, GRCh38) - Ensembl 92

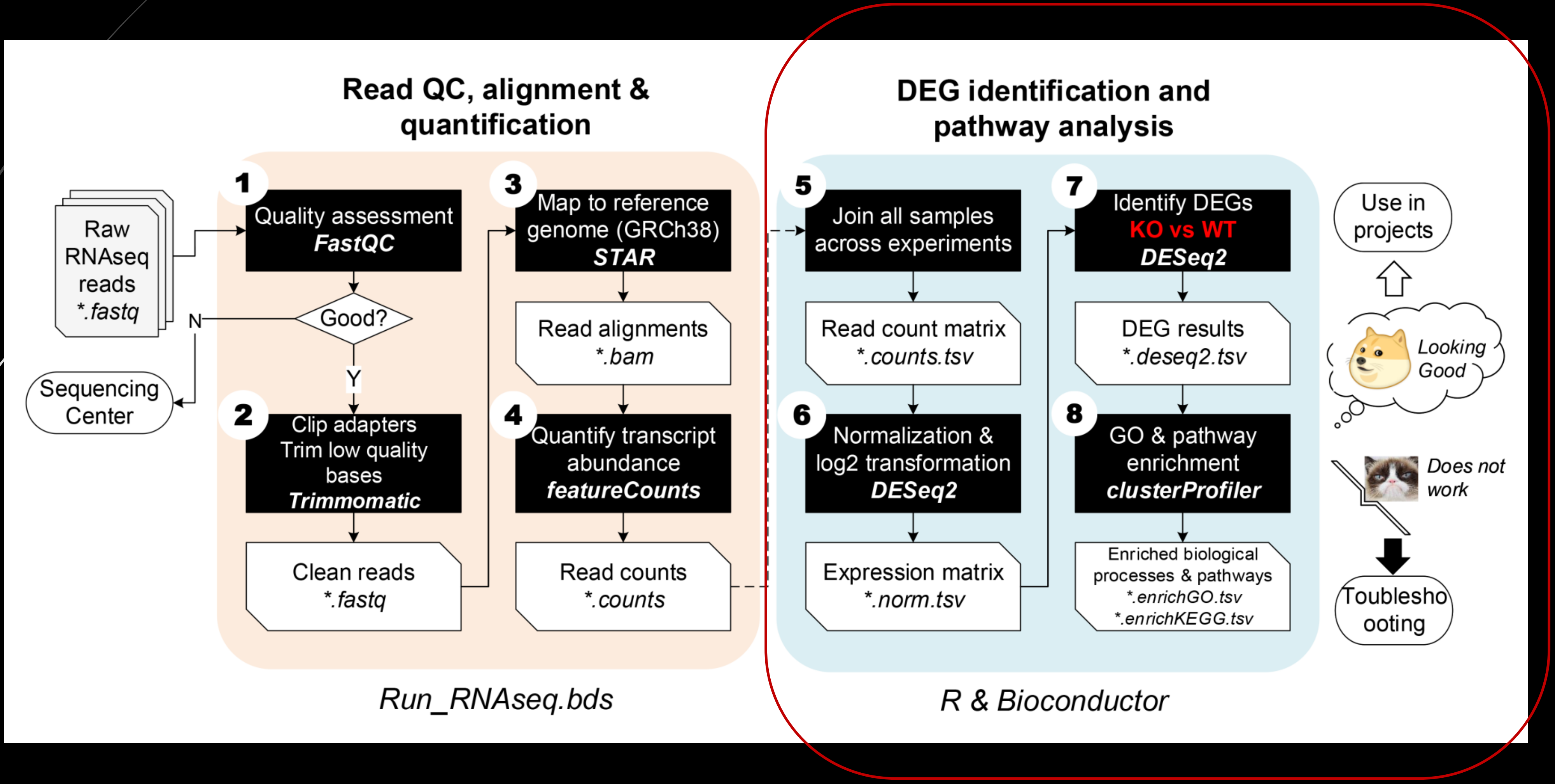
General stats

Total No of Genes	58381
Protein-coding genes	19901
Long non-coding RNA genes	15779
Small non-coding RNA genes	7569
Pseudogenes	14723
- processed pseudogenes:	10693
- unprocessed pseudogenes:	3519
- unitary pseudogenes:	218
- polymorphic pseudogenes:	38
- pseudogenes:	18
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	408
- pseudogenes:	237

Total No of Transcripts	203835
Protein-coding transcripts	82335
- full length protein-coding:	56541
- partial length protein-coding:	25794
Nonsense mediated decay transcripts	14889
Long non-coding RNA loci transcripts	28468
Total No of distinct translations	61132
Genes that have more than one distinct translations	13641

Gencode and Ensembl are generally in sync

How to perform RNAseq analysis



Read QC, alignment & quantification

Raw RNAseq reads *.fastq

1 Quality assessment *FastQC*

Good?

N

Sequencing Center

Y

2 Clip adapters Trim low quality bases *Trimmomatic*

Clean reads *.fastq*

3 Map to reference genome (GRCh38) *STAR*

Read alignments *.bam*

4 Quantify transcript abundance *featureCounts*

Read counts *.counts*

Run_RNAseq.bds

DEG identification and pathway analysis

5 Join all samples across experiments

Read count matrix *.counts.tsv*

6 Normalization & log2 transformation *DESeq2*

Expression matrix *.norm.tsv*

7 Identify DEGs KO vs WT *DESeq2*

DEG results *.deseq2.tsv*

8 GO & pathway enrichment *clusterProfiler*

Enriched biological processes & pathways *.enrichGO.tsv *.enrichKEGG.tsv

R & Bioconductor

Use in projects

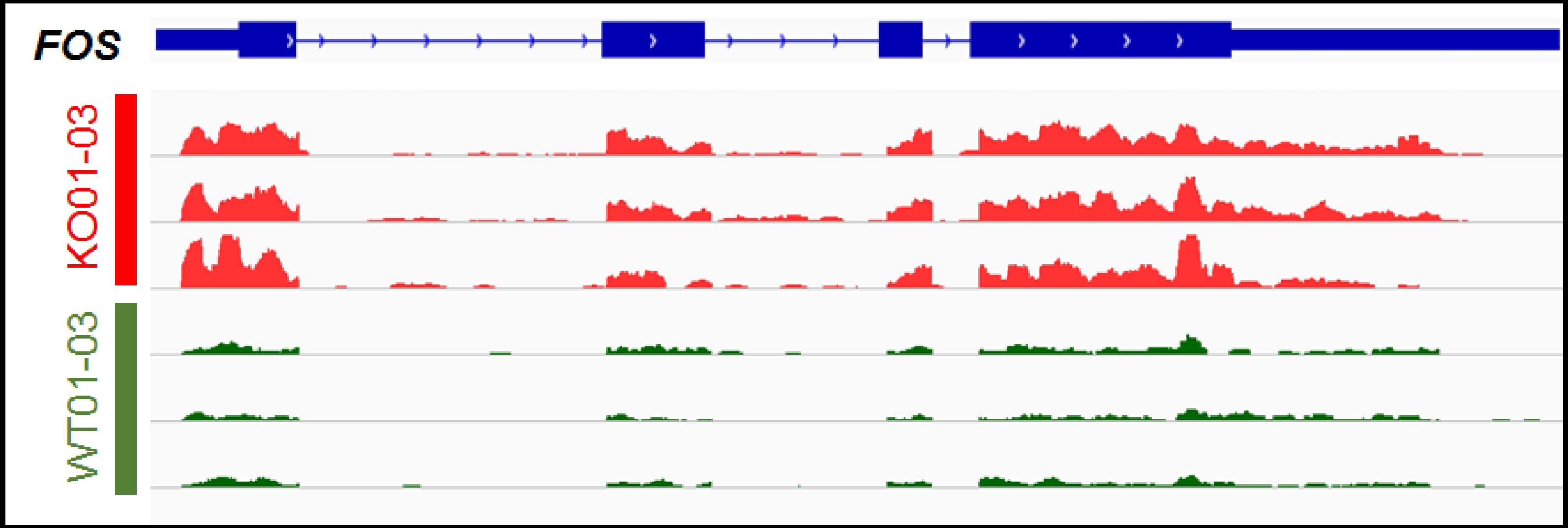
Looking Good

Does not work

Toubleshooting

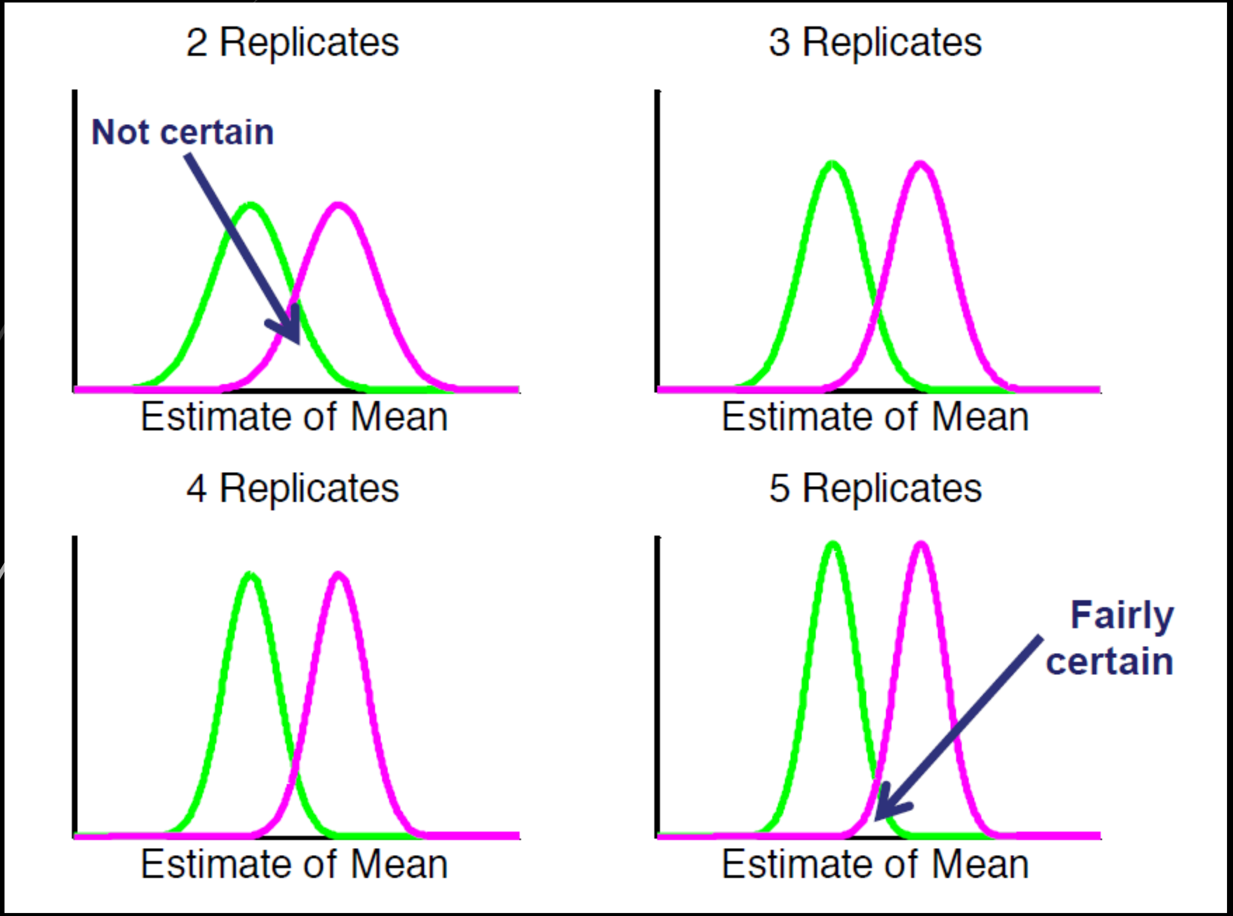
05-08: Identify differentially expressed genes and pathways: DESeq2, clusterProfiler

- After steps 01-04, we have generated read alignment and counts for every annotated gene on the genome
- The next step is to utilize the read counts data to detect DEGs
- For example, if we visualize *FOS* gene across 6 samples in genome browser



FOS = Fos proto-oncogene, AP-1 transcription factor subunit

DEG detection



$$SE = \frac{\sigma}{\sqrt{n}}$$

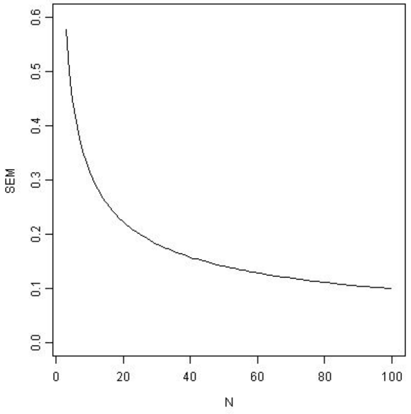
← Standard deviation

← Number of samples

Standard Error of the Mean

$$SE_{M_x} = \frac{\sigma}{\sqrt{N}}$$

- This equation implies that sampling error decreases as sample size increases.
- This is important because it suggests that if we want to make sampling error as small as possible, we need to use as large of a sample size as we can manage.

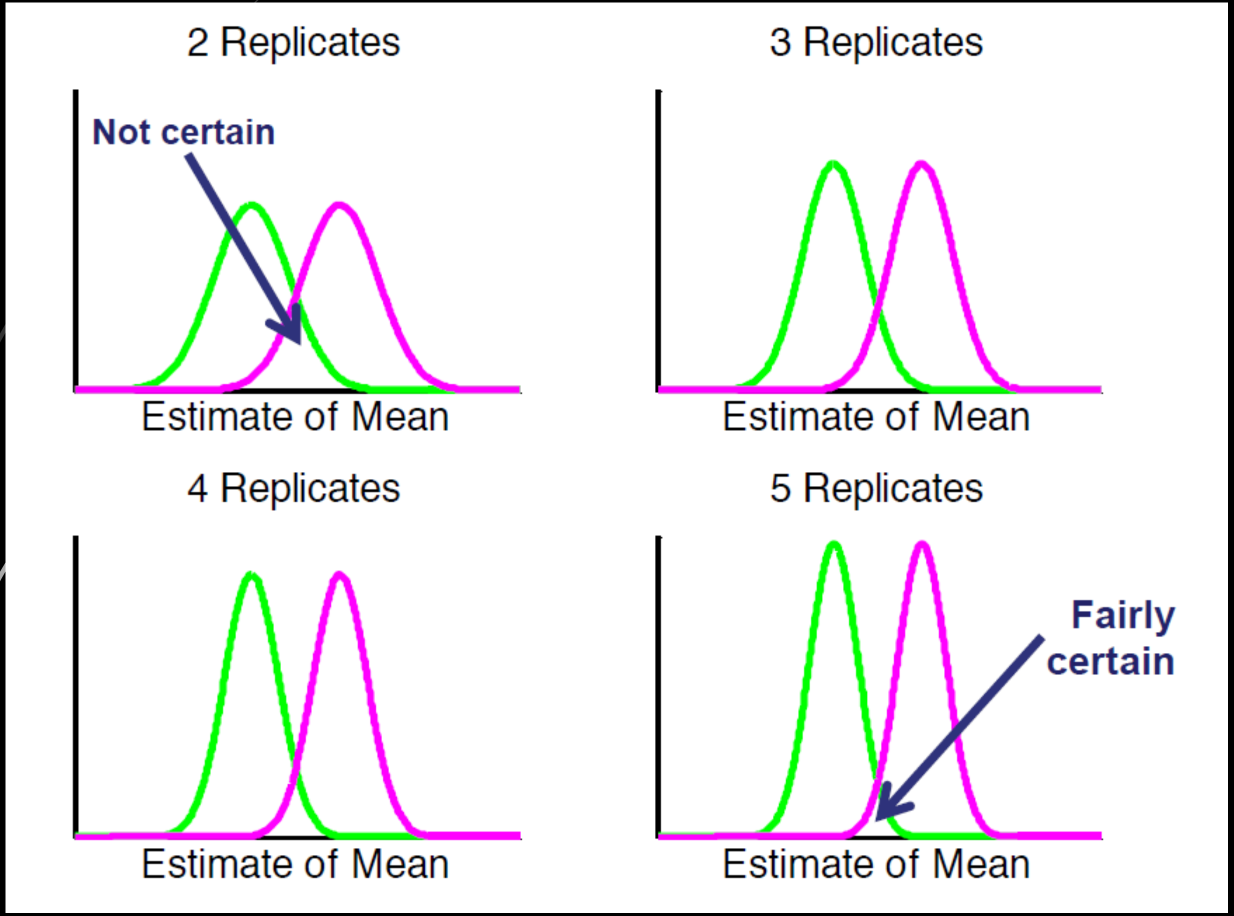


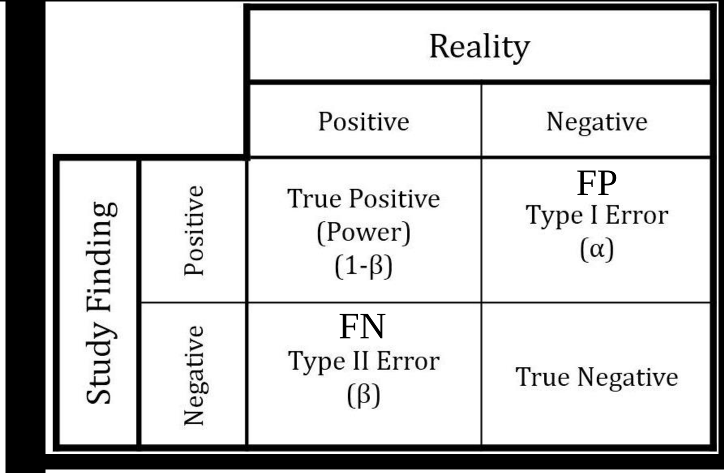
More biological replicates per group lead to higher discovery power, sensitivity and specificity.

DEG detection

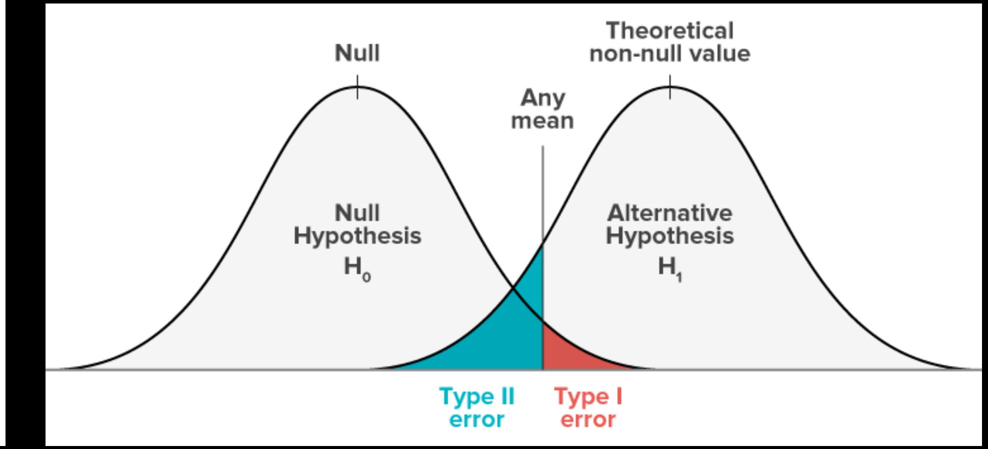
Sensitivity: TP/(TP+FN)

Specificity: TN/(FP+TN)



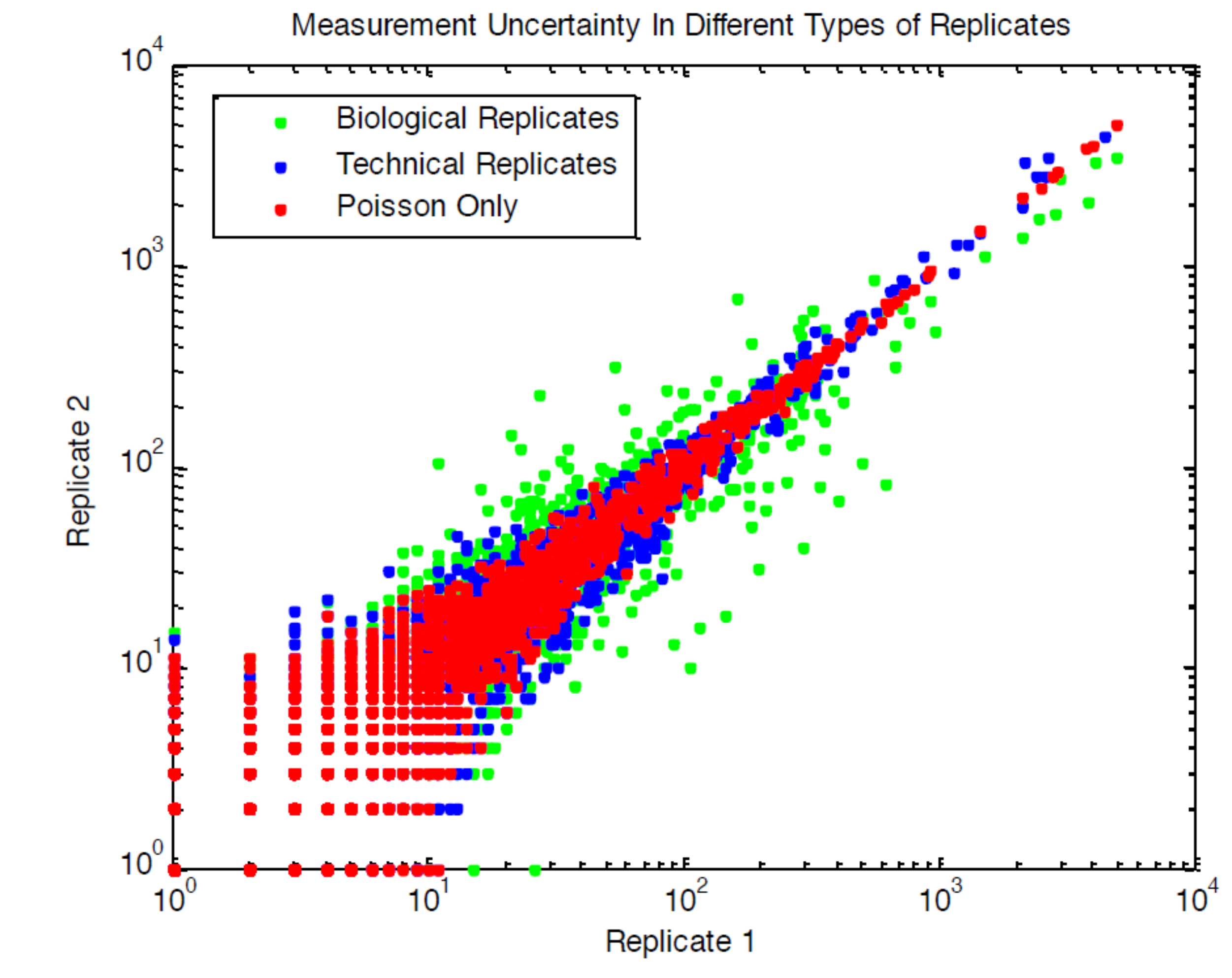


		Reality	
		Positive	Negative
Study Finding	Positive	True Positive (Power) (1-β)	FP Type I Error (α)
	Negative	FN Type II Error (β)	True Negative



More biological replicates per group lead to higher discovery power, sensitivity and specificity.

DEG detection



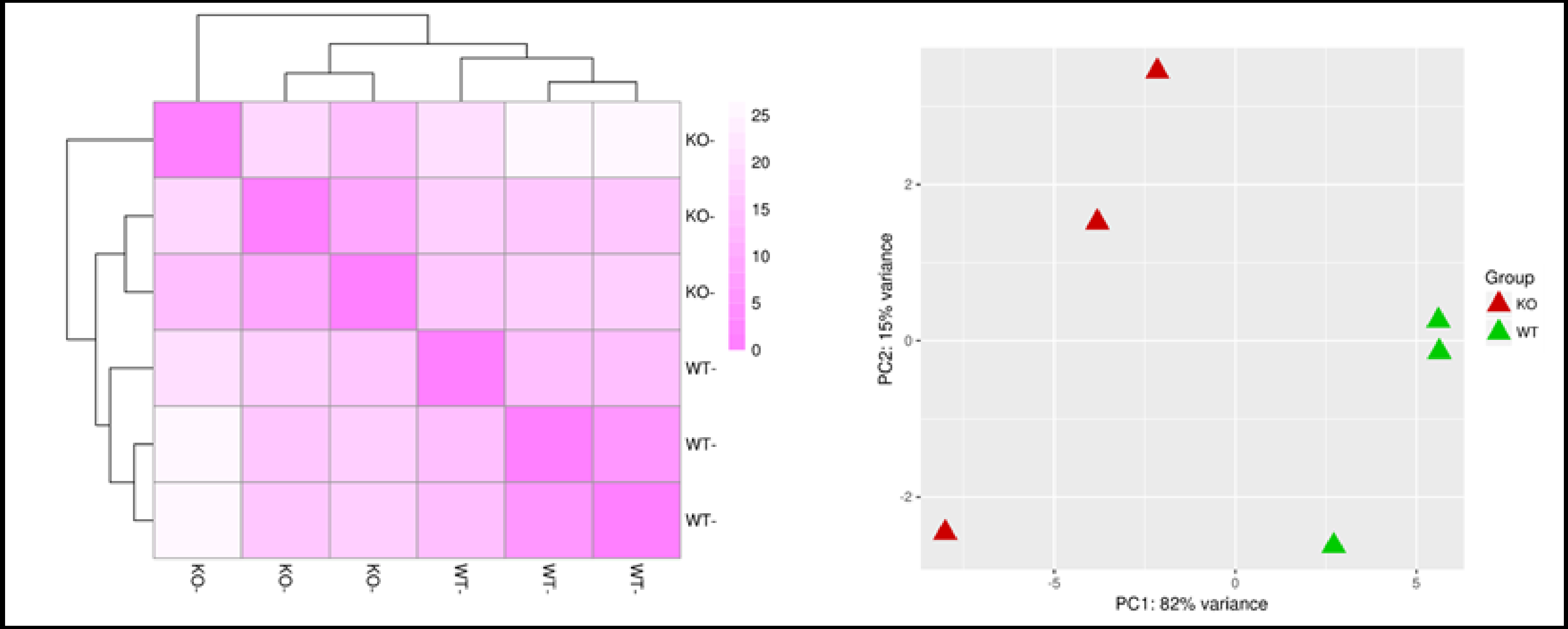
http://www.broadinstitute.org

High correlation is expected between biological replicates.

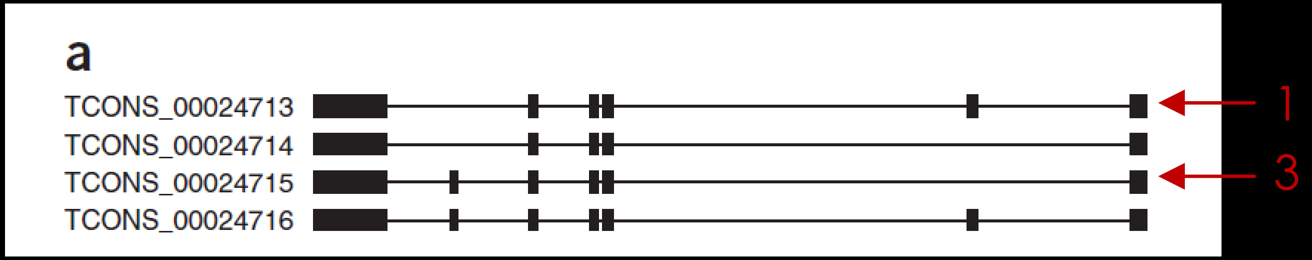
If one sample is an outlier, it can be identified if multiple replicates are included in an experiment.

How to identify an outlier?

- PCA plot (visualization)
- Unsupervised sample clustering based on all genes or *top variable genes (e.g. 1500)*

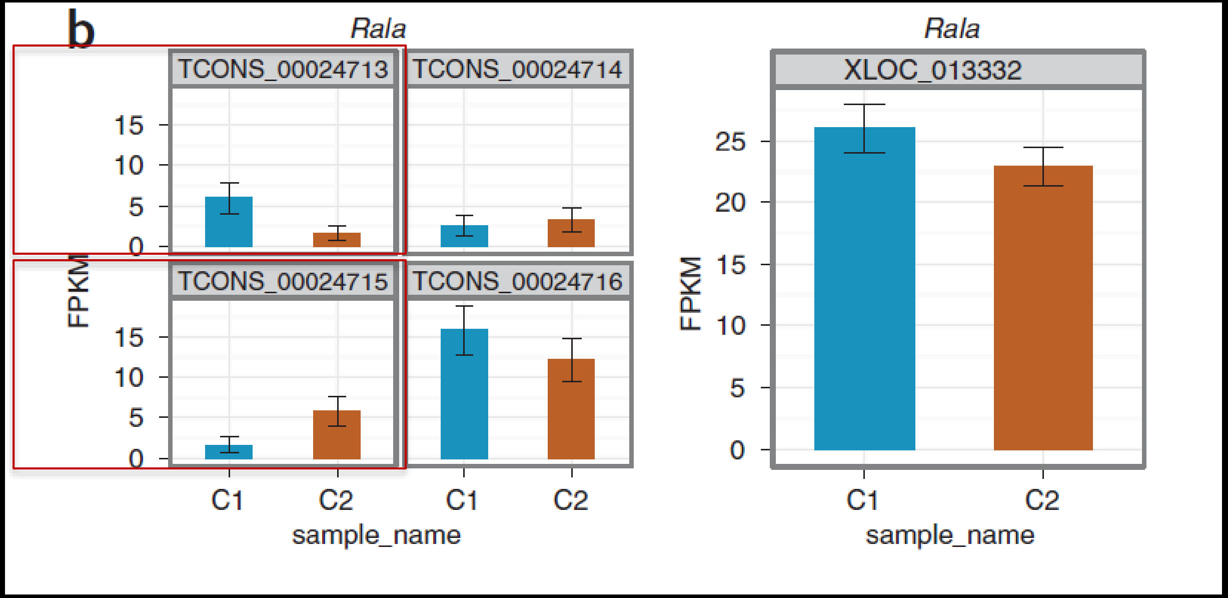


Transcript vs gene level quantification



Isoform 1 *

Isoform 3 *



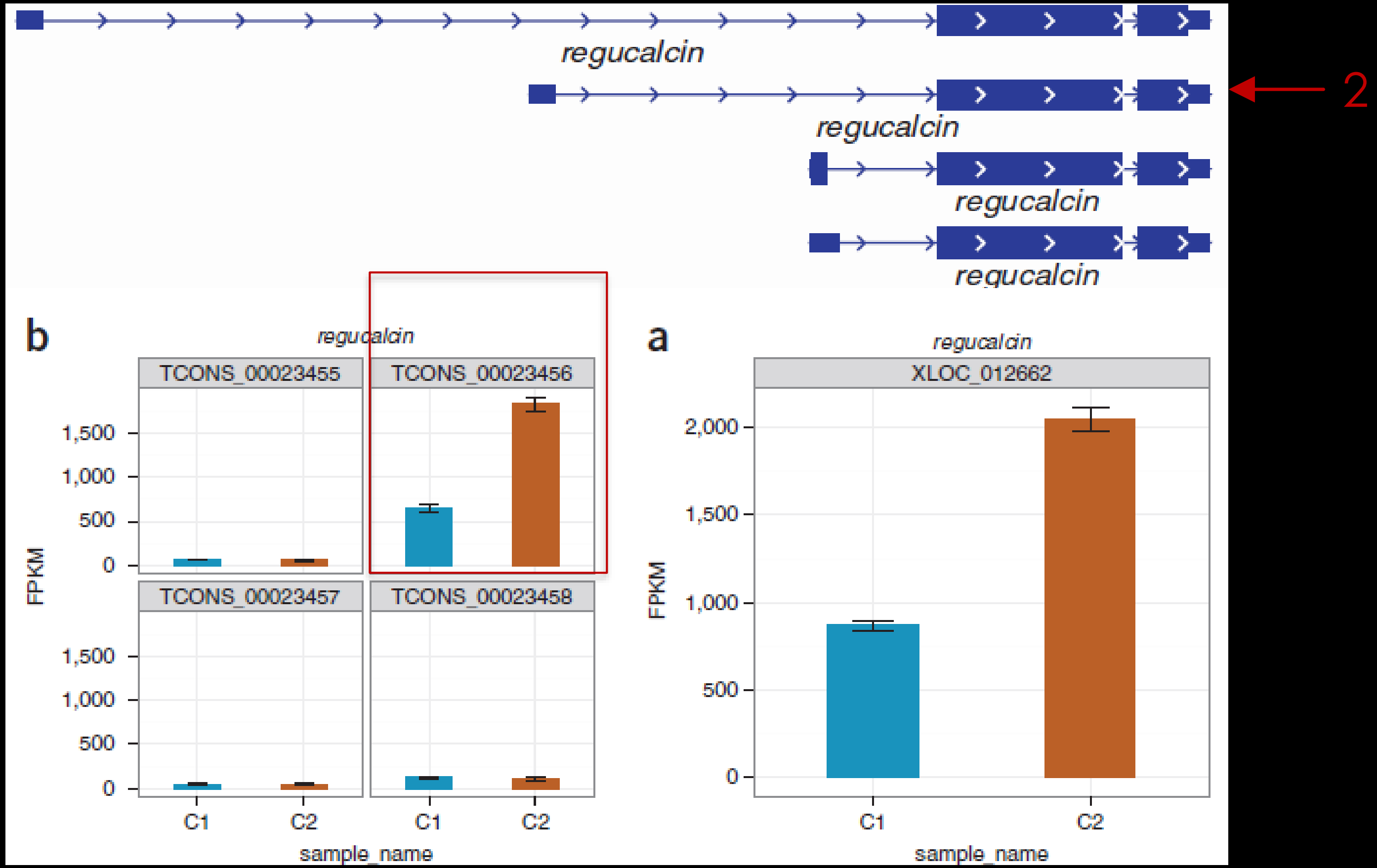
Isoforms 1 – 4

Gene

Difference in gene-level expression is not significant due to variability of isoforms

Transcript vs gene level quantification

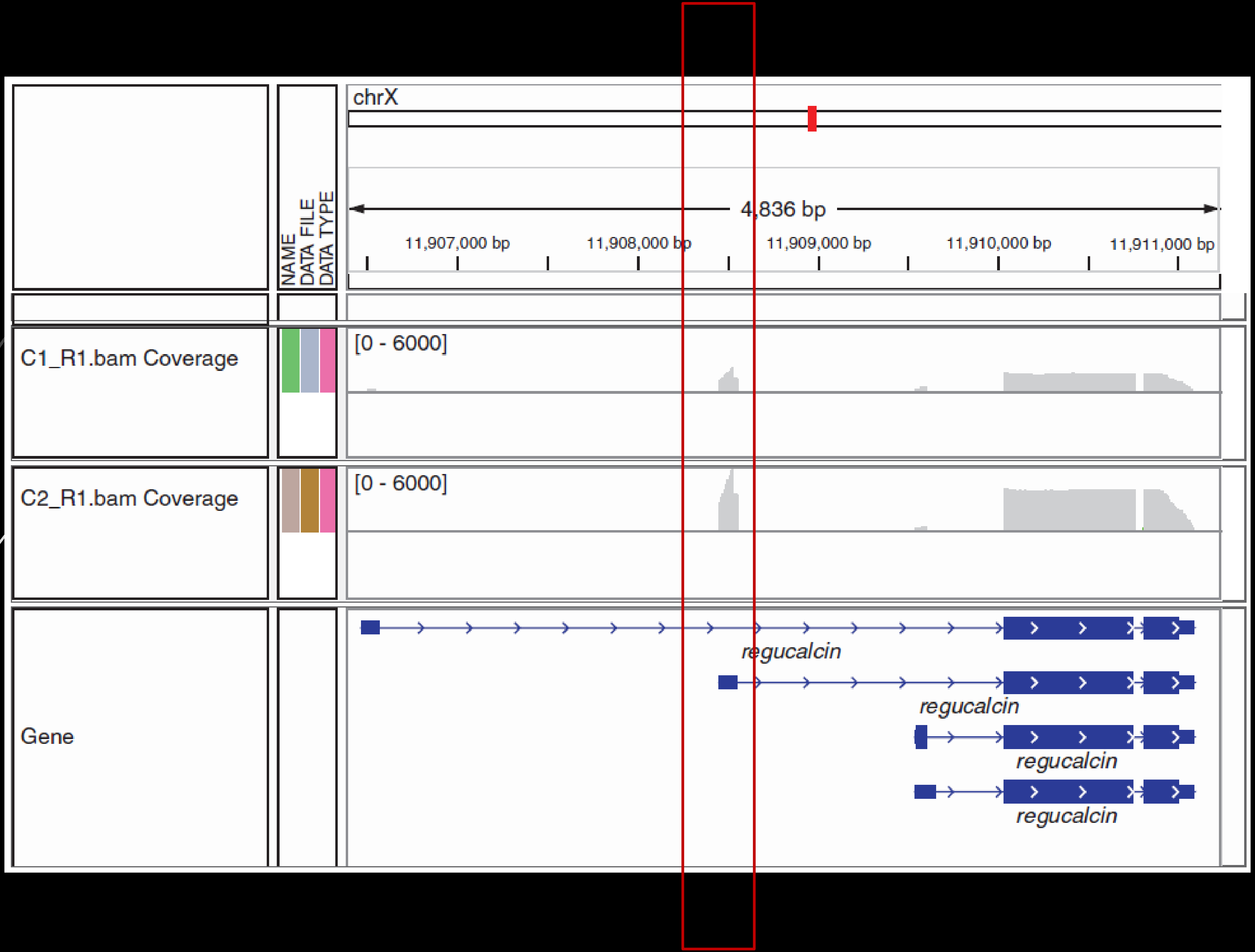
Isoform 2 *



Isoforms 1 – 4

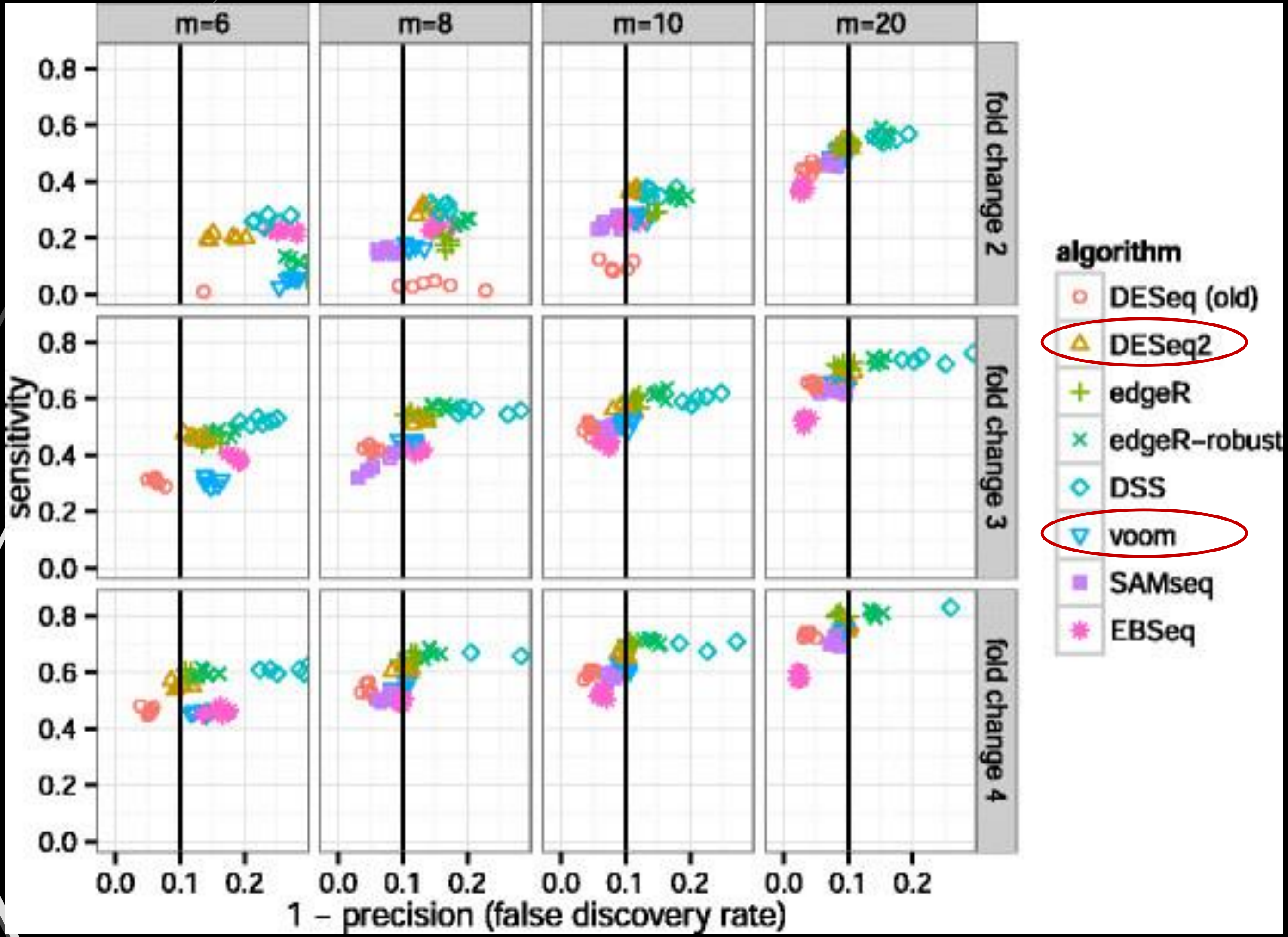
Gene

Difference in gene-level expression is significant, which is largely due to a great increase in the expression of isoform 2



Difference in gene-level expression is significant, which is largely due to a great increase in the expression of isoform 2

Comparison of different DEG identification methods



Sensitivity and precision of algorithms across combinations of sample size and effect size.

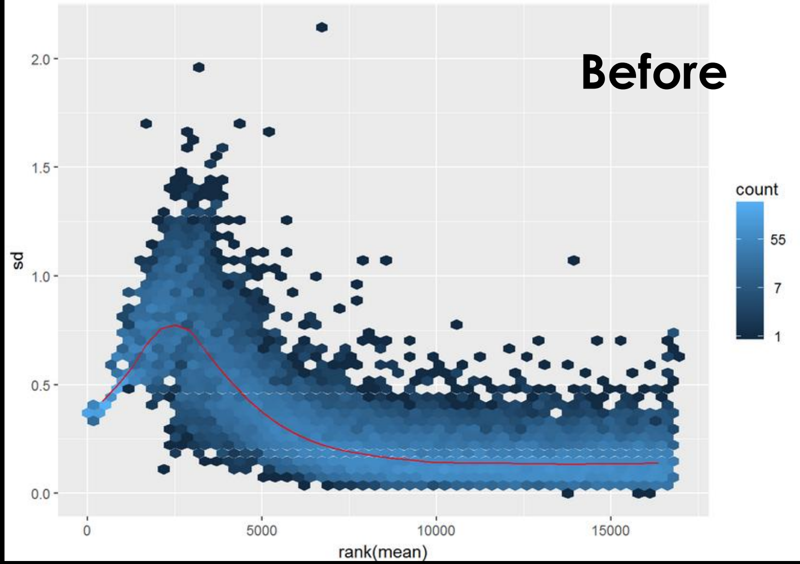
DESeq2 and edgeR often had the highest sensitivity of those algorithms that controlled the FDR, i.e., those algorithms which fall on or to the left of the vertical black line.

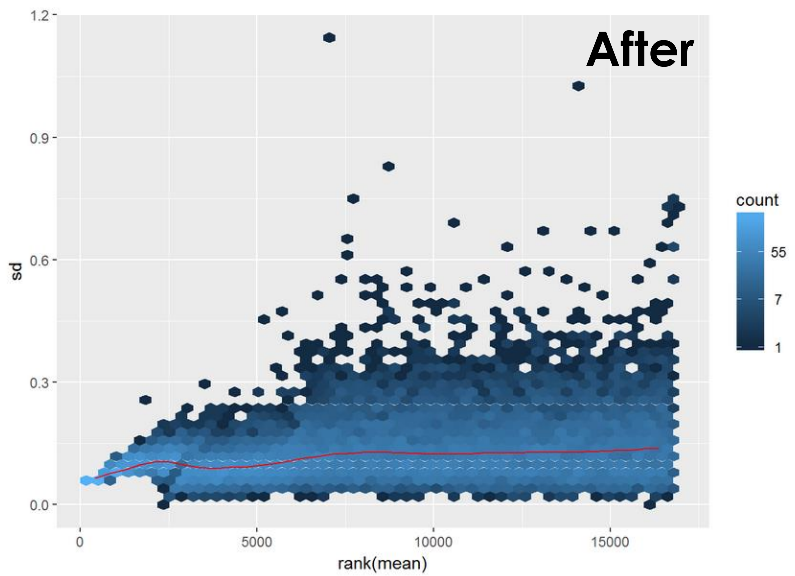
m: total sample size; split into two even-sized groups for comparison

Love et al., Genome Biology 2014

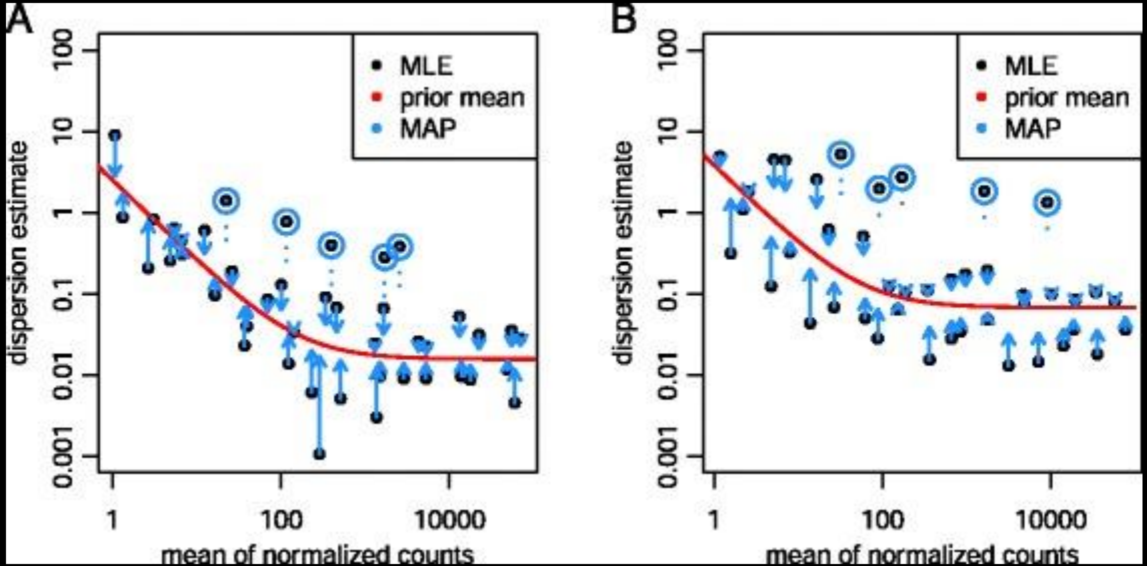
DESeq2

- Count matrix data
- Assume data follow negative binomial distribution (sometimes also called a gamma-Poisson distribution) with mean (μ) and dispersion (α) parameters
- Within-group variability, i.e., the variability between replicates, is modeled by the dispersion parameter alpha, which describes the *variance* of counts
- Empirical Bayes shrinkage for dispersion estimation





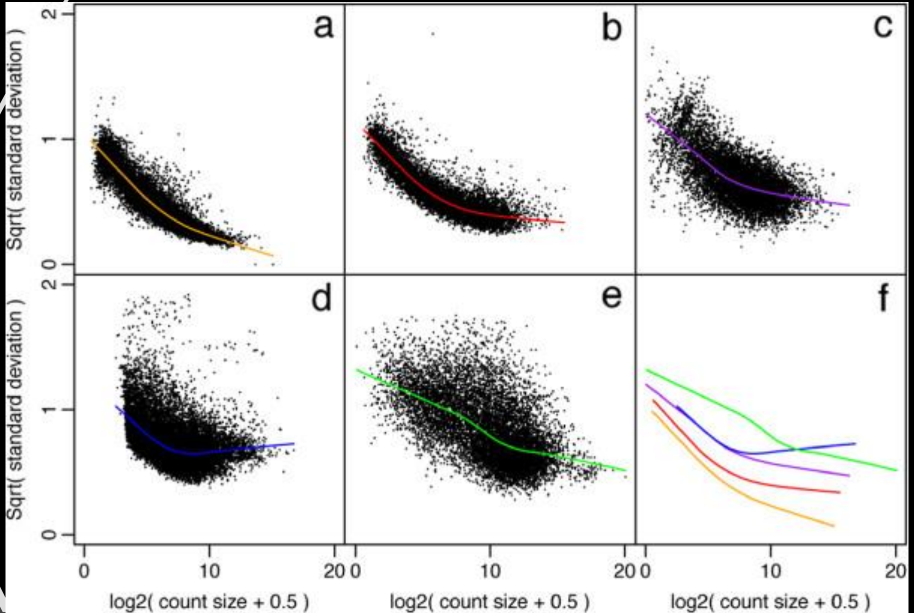
MAP, maximum a posteriori; MLE, maximum-likelihood estimate



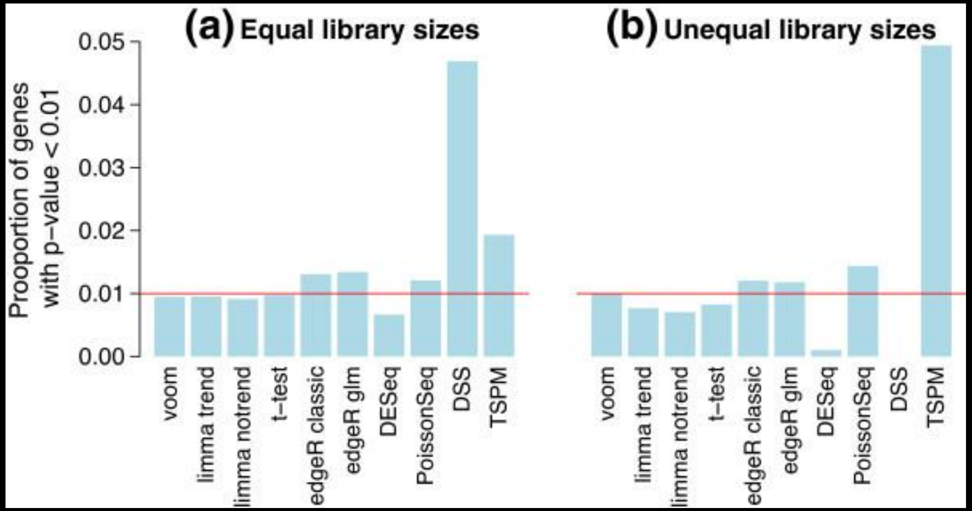
Limma voom (*weighted* algorithm)

- To model the mean-variance relationship than to specify the exact probabilistic distribution of the counts (e.g. NB or Poisson)
- Provide accurate Type I (alpha) and Type II error (beta) control compared to other methods, especially when sample size is small
- *Voom with sample quality weights*

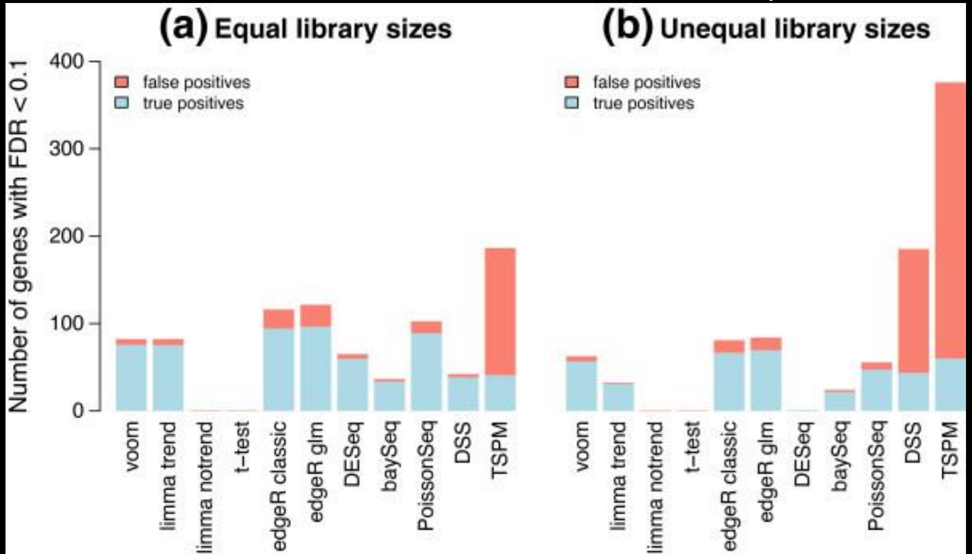
Law et al., Genome Biology 2014



Type I error rates in the absence of true differential expression

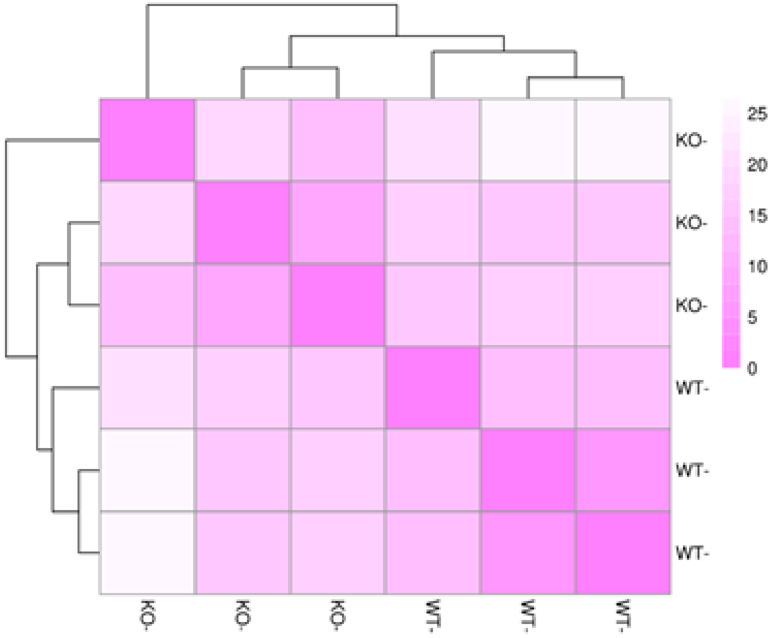


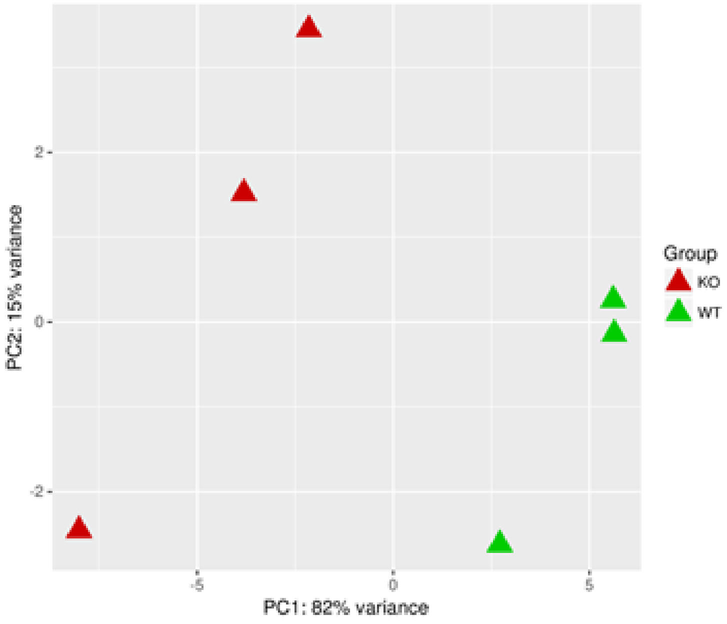
Power to detect true differential expression



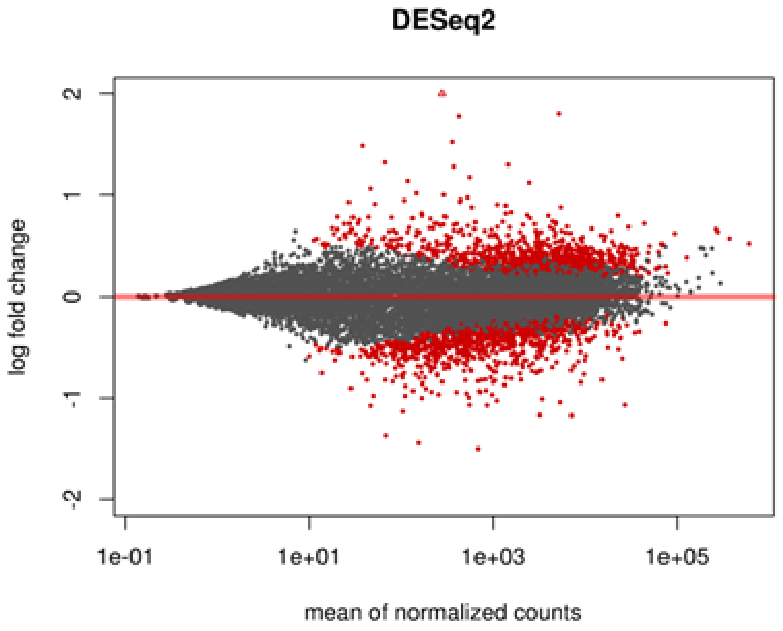
More databases!

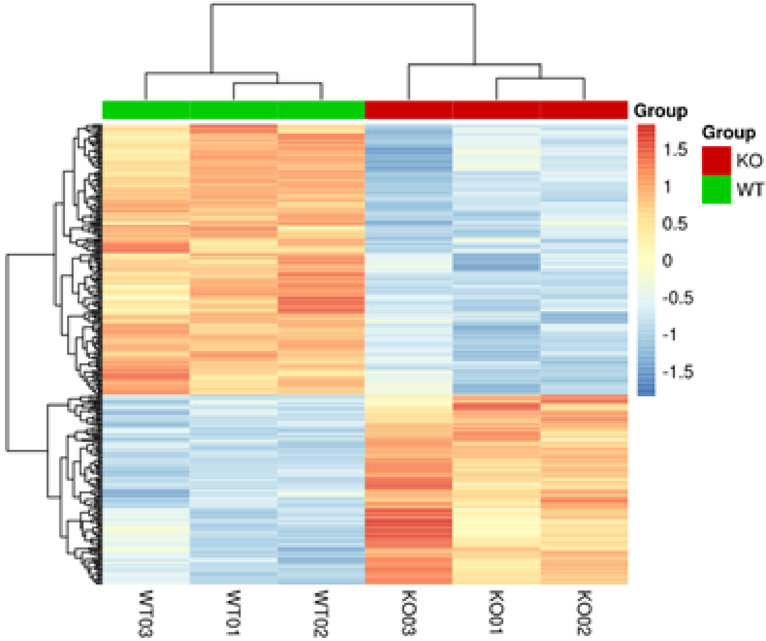
- Gene annotation database: GENCODE
  - https://www.gencodegenes.org/
- Gene Ontology (GO) database: Gene Ontology Consortium
  - http://www.geneontology.org/
- Pathway database: KEGG
  - http://www.genome.jp/kegg/
- Predefined gene sets: MSigDB
  - http://software.broadinstitute.org/gsea/msigdb/

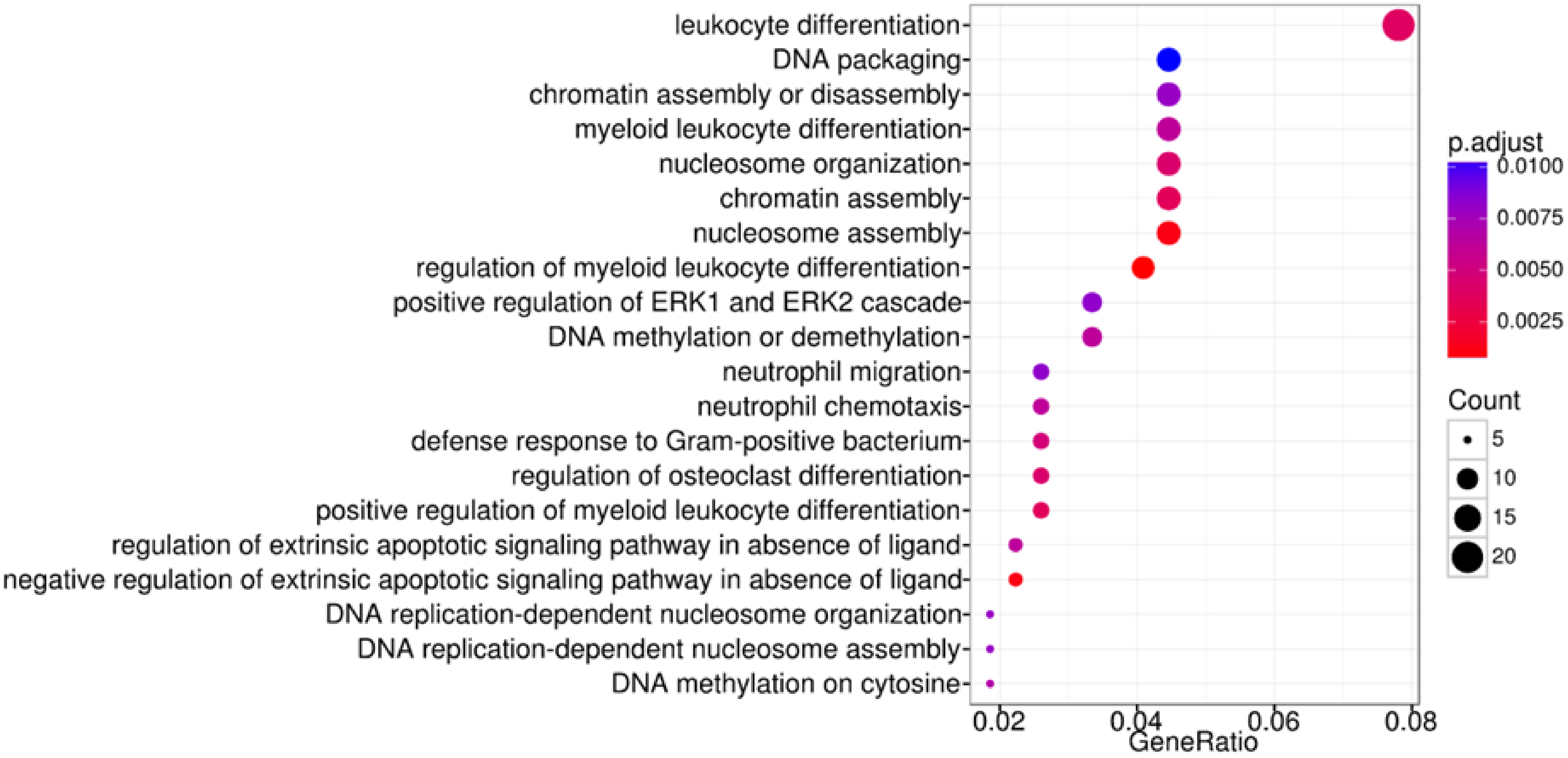


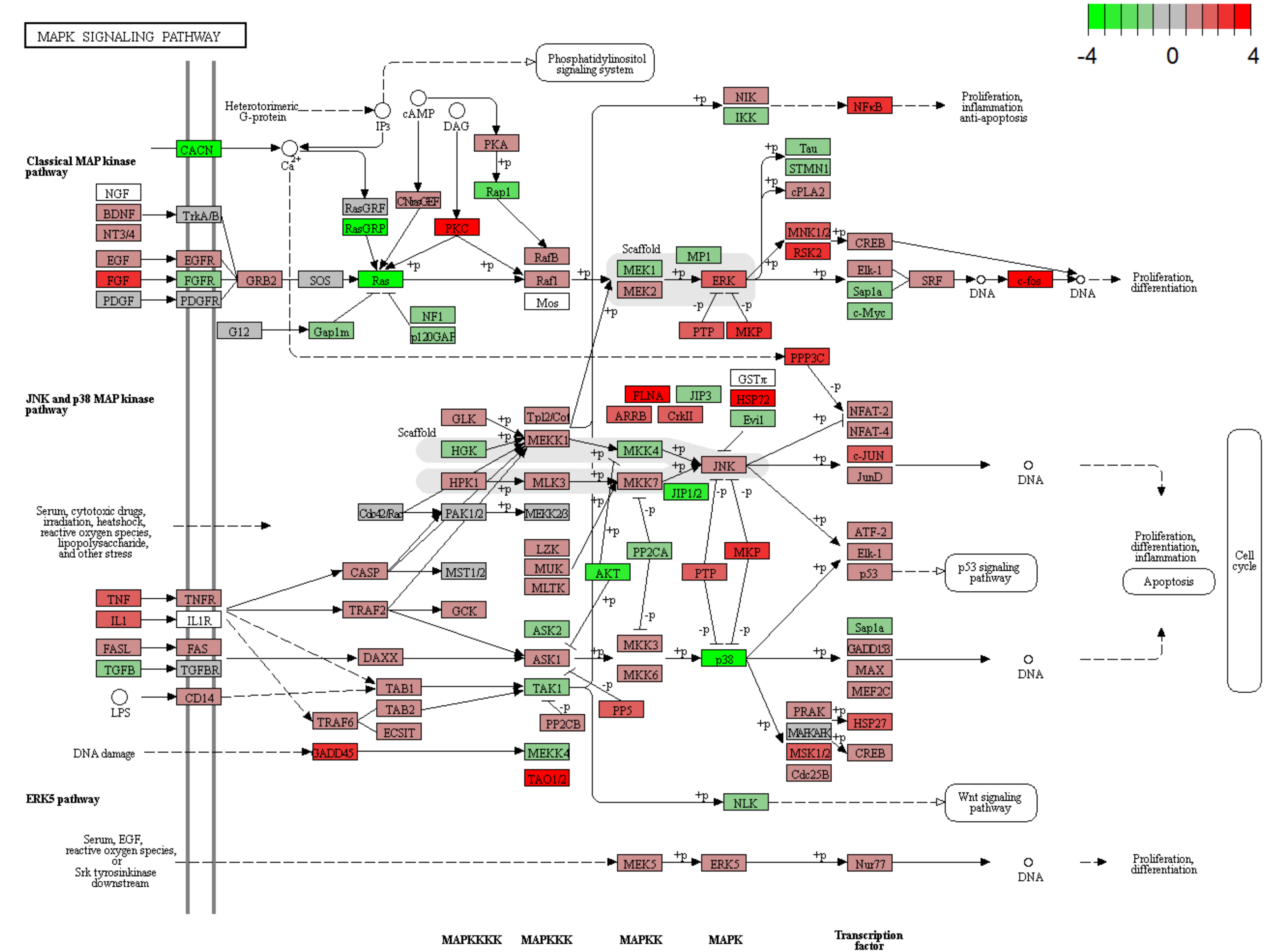


```
[1] "Genes significant = 296 (fc, 1.5, fdr 0.05)"
[1] "Heatmap = 296 genes on the row, 6 samples on the column"
```






leukocyte differentiation
DNA packaging
chromatin assembly or disassembly
myeloid leukocyte differentiation
nucleosome organization
chromatin assembly
nucleosome assembly
regulation of myeloid leukocyte differentiation
positive regulation of ERK1 and ERK2 cascade
DNA methylation or demethylation
neutrophil migration
neutrophil chemotaxis
defense response to Gram-positive bacterium
regulation of osteoclast differentiation
positive regulation of myeloid leukocyte differentiation
regulation of extrinsic apoptotic signaling pathway in absence of ligand
negative regulation of extrinsic apoptotic signaling pathway in absence of ligand
DNA replication-dependent nucleosome organization
DNA replication-dependent nucleosome assembly
DNA methylation on cytosine
0.02
0.04
0.06
0.08
GeneRatio
p.adjust
0.0100
0.0075
0.0050
0.0025
Count
5
10
15
20



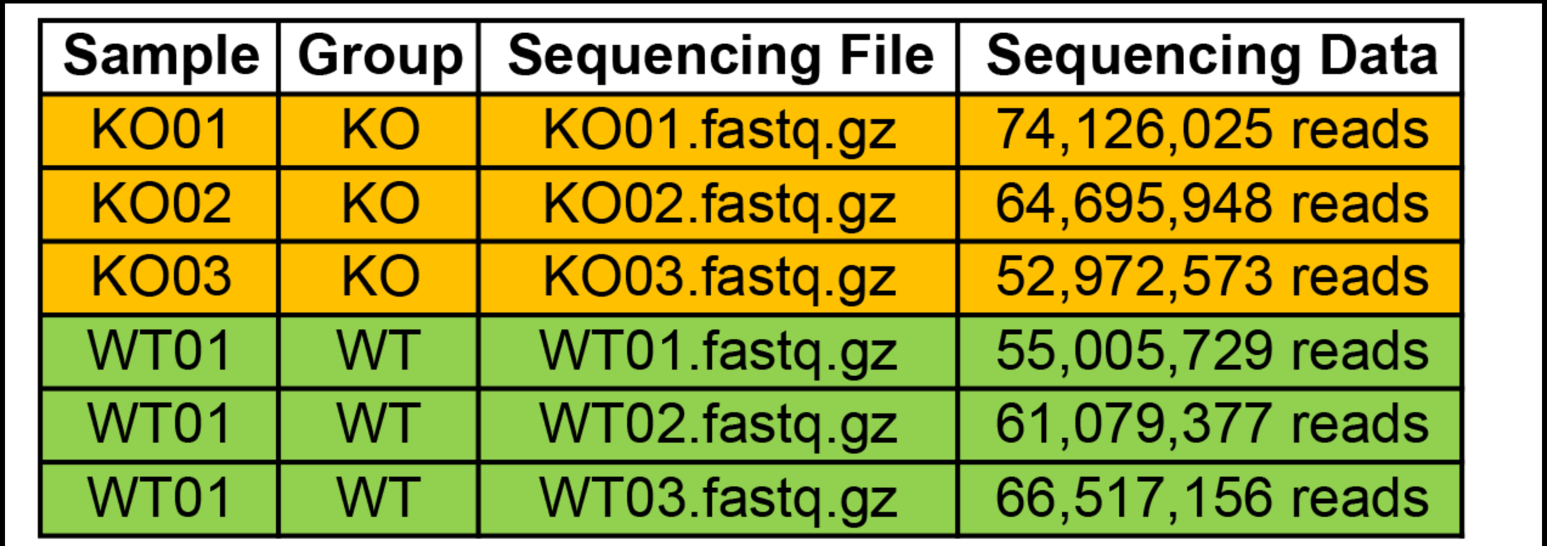
Data on KEGG graph
Rendered by Pathview

Thank you!

Questions

Hands-on practice START

- Open your handson.Rmd on the Github or download to local computer
- https://github.com/MScBiomedicalInformatics/MSIB32500/blob/master/lectures/handson9.html
- *Dataset: two groups (PRDM11 KO vs WT, human U2932 cells), 6 samples*
- *Single-end reads, unstranded libraries*



Sample	Group	Sequencing File	Sequencing Data
KO01	KO	KO01.fastq.gz	74,126,025 reads
KO02	KO	KO02.fastq.gz	64,695,948 reads
KO03	KO	KO03.fastq.gz	52,972,573 reads
WT01	WT	WT01.fastq.gz	55,005,729 reads
WT01	WT	WT02.fastq.gz	61,079,377 reads
WT01	WT	WT03.fastq.gz	66,517,156 reads

Fog. et al. 2015. Loss of *PRDM11* promotes MYC-driven lymphomagenesis. Blood 125(8):1272-81

PRDM11 = PR/SET domain 11

INDEX
EXTRA QUALITY
UK90