

# Transcriptome Data Analysis

Federico Marini ([marinif@uni-mainz.de](mailto:marinif@uni-mainz.de))



IMBEI@



2023/07/05-06

 @FedeBioinfo

# Let's do a quick introduction round!

I'd like to know

- Your name
- The link between you and RNA-seq data
- A fun fact about you

I'll start :)

Hi, I am Federico, I am the head of the Bioinformatics group at the IMBEI in Mainz.  
People told me I am somehow well versed with RNA-seq data.  
I like to cook and talk about food.

# Questions

**What are the steps to process RNA-Seq data?**

- How to convert RNA-seq reads into counts?
- How to perform quality control (QC) of RNA-seq reads?

**How to identify differentially expressed genes across multiple experimental conditions?**

- How to properly analyze RNA count data using DESeq2?
- How to perform quality control (QC) and exploratory data analysis (EDA) of RNA-seq count data?

**What are the biological functions impacted by the differential expression of genes?**

- How can I perform a gene ontology enrichment analysis?

**How can I create neat visualizations of the data?**

- How can I visualize the results for my enrichment analysis?

**How can I generate interactive reports to summarise my analyses?**

# What you will learn

- the basics of RNA-seq data
- the basics of RNA-seq data analysis
- to get familiar with the concepts of gene expression, high-dimensional data, expression quantification, differential expression analysis
- the importance to pose the right question, in order to get the right answer :)

This material has been developed together with Charlotte Soneson in the scope of the GTIPI Summer School  
(<https://imbeimainz.github.io/GTIPI2022>)

# Setup for practical sessions

Got R/RStudio?

Latest versions highly recommended!

See <https://imbeimainz.github.io/GTIPI2022/material.html> for details!

+

This repo should contain it all: [https://github.com/imbeimainz/MSE\\_GenEpi\\_2023](https://github.com/imbeimainz/MSE_GenEpi_2023)

# Decomposing the title

RNA

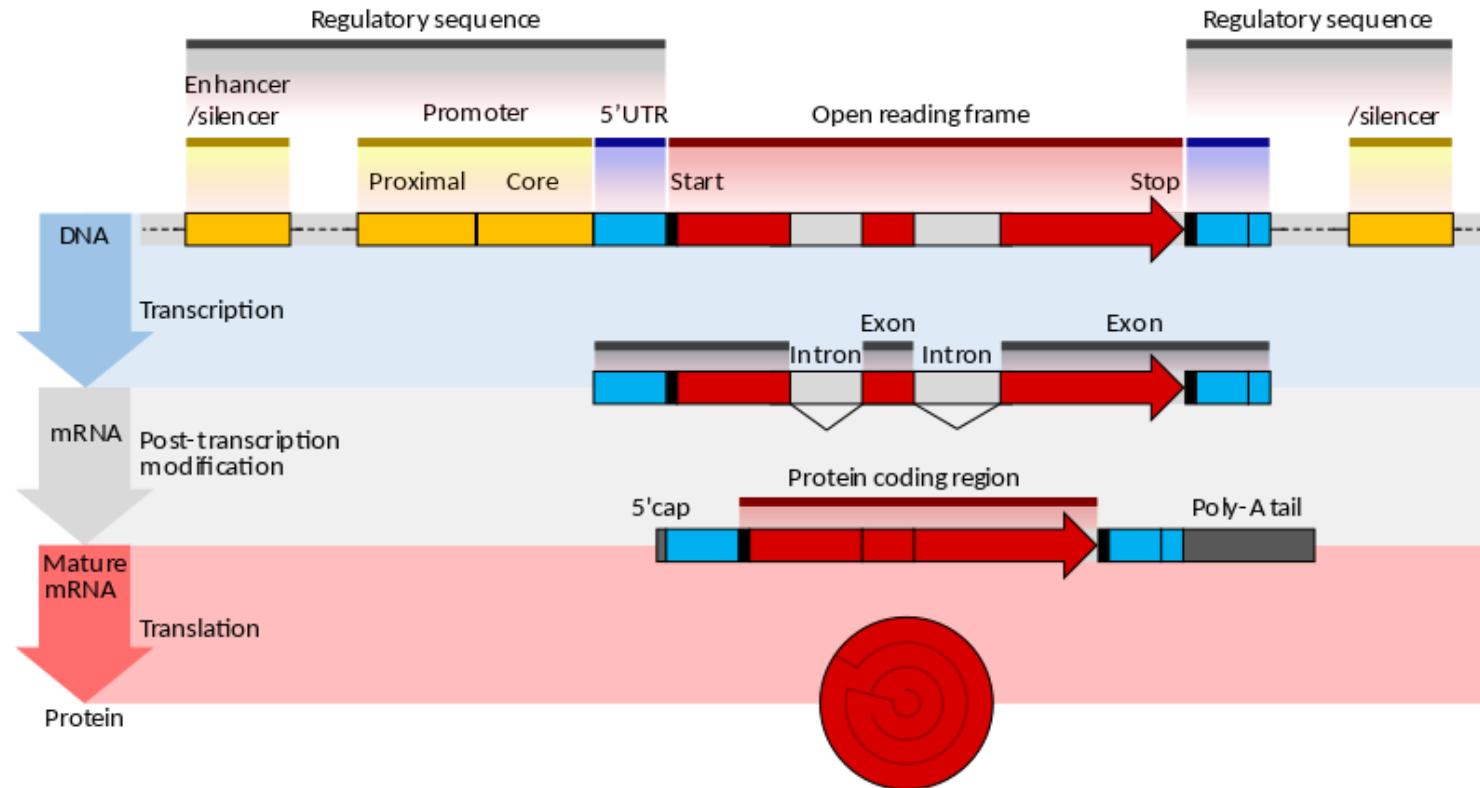
Sequencing

Bioinformatics

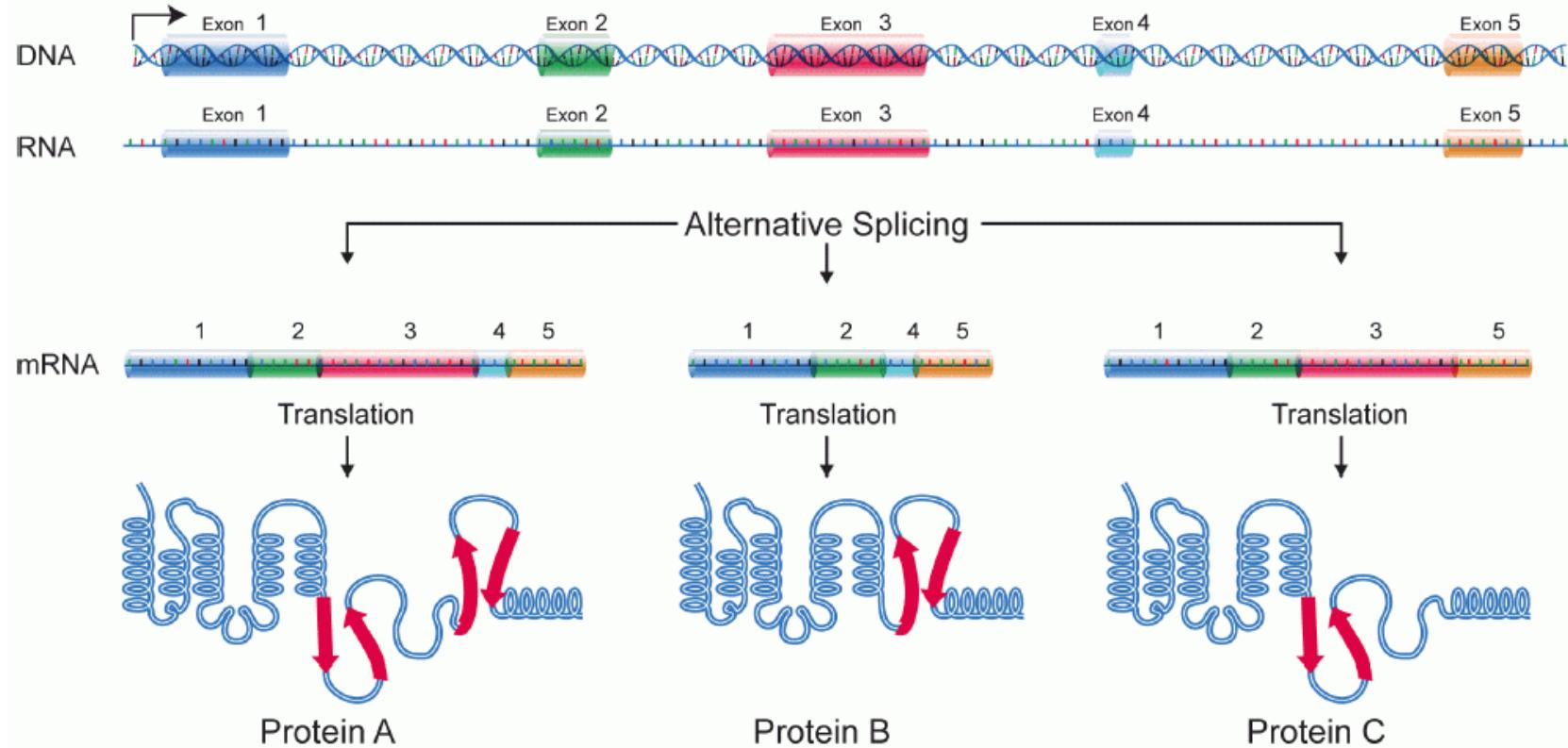
Transcriptome

Analysis

# (messenger) RNA

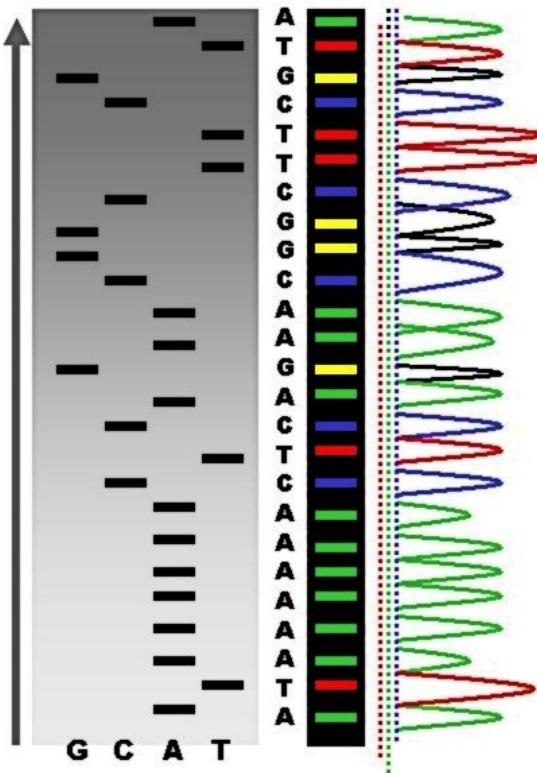


# (messenger) RNA



Exons, introns, transcripts, isoforms

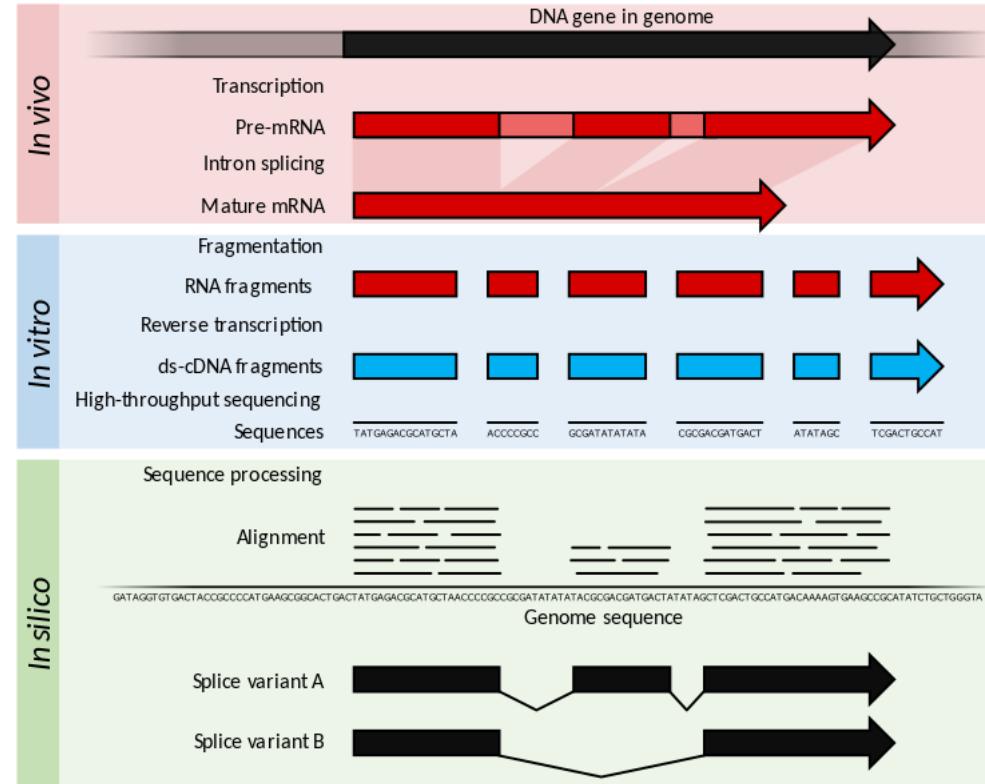
# Sequencing



4 base pairs, 21 aminoacids

Excellent review on next-generation sequencing: <https://www.nature.com/articles/nrg.2016.49>

# RNA-sequencing



- RNA quantification at single base resolution
- Cost efficient analysis of the whole transcriptome in a high-throughput manner

# Challenges in RNA-seq

- Different origin for the sample RNA and the reference genome
- Presence of incompletely processed RNAs or transcriptional background noise
- Sequencing biases (e.g. PCR library preparation)

## Benefits

- sensitive
- specific
- high-throughput
- cost-efficient
- basepair resolution

You can see: transcripts, splicing, lncRNA, circRNA, gene fusions

# Bioinformatics

"an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex."

A combination of

- biology
- computer science
- information engineering
- mathematics
- statistics

... to analyze and interpret the biological data

# Transcriptome

Gene expression is a fundamental level at which the results of various genetic and regulatory programs are observable.

RNA sequencing (RNA-seq) provides a quantitative and open system for profiling transcriptional outcomes on a large scale

Much has been learned about the characteristics of the RNA-seq data sets, as well as the performance of the myriad of methods developed

"RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis" ->

<https://www.annualreviews.org/doi/abs/10.1146/annurev-biodatasci-072018-021255>

# Analysis

There's data involved!

and these datasets have particular properties (how they are generated, ...)

How to make sense out of it?

There is a large diversity of applications to deal with

Among the most widely adopted workflows:

- Transcript discovery

*Which RNA molecules are in my sample?*

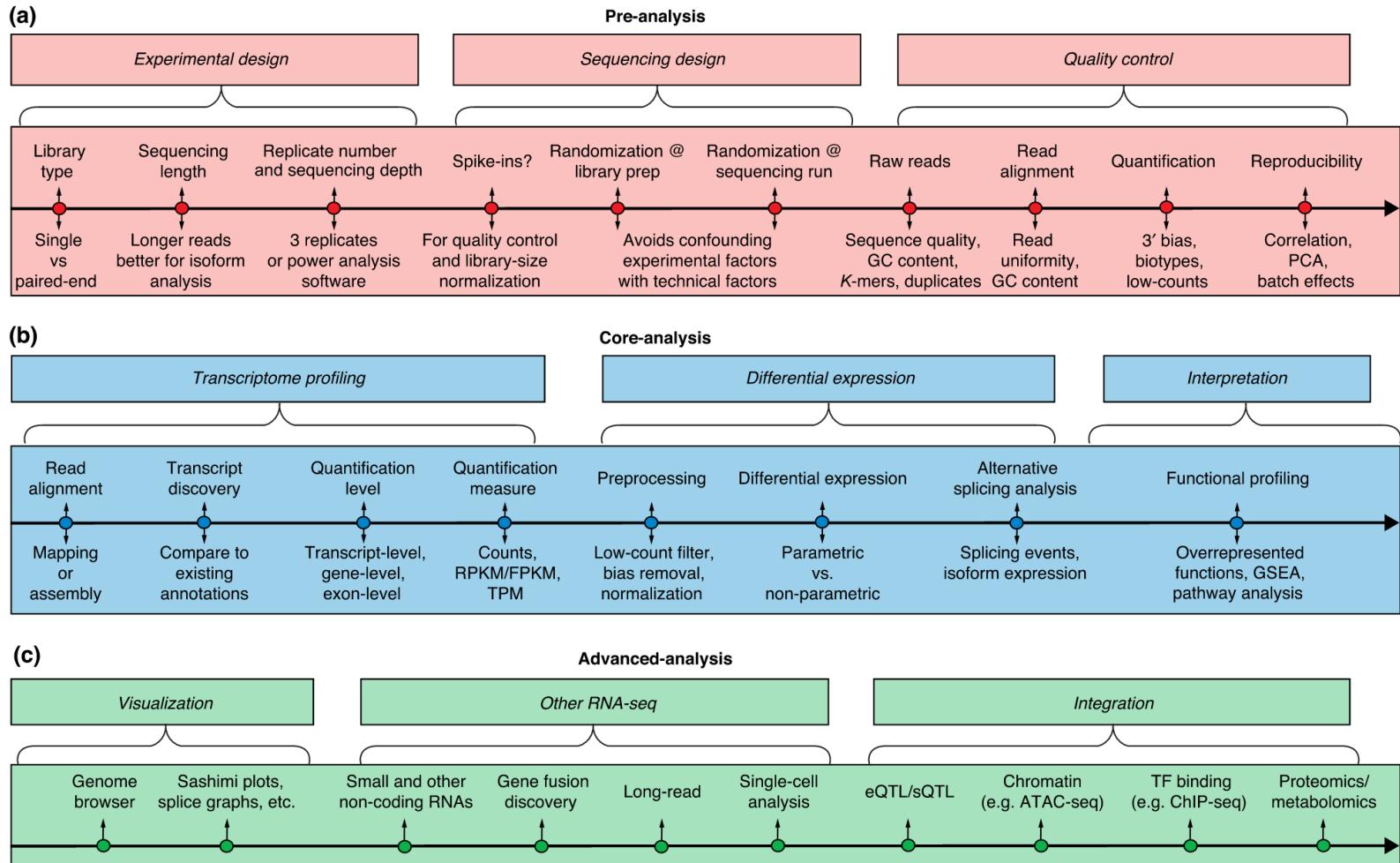
Novel isoforms and alternative splicing, Non-coding RNAs, Single nucleotide variations, Fusion genes

- RNA quantification

*What is the concentration of RNAs?*

Absolute gene expression (within sample), Differential expression (between biological samples)

# Analysis - a bird's eye view



# The essential

- Expression quantification
- Data exploration: principal components analysis, gene plots
- Differential analysis: DE modeling, design, effect size, variability, significance
- Functional interpretation: gene sets, pathways, biological themes

# Differential analysis types for RNA-seq

No single available standardized workflow

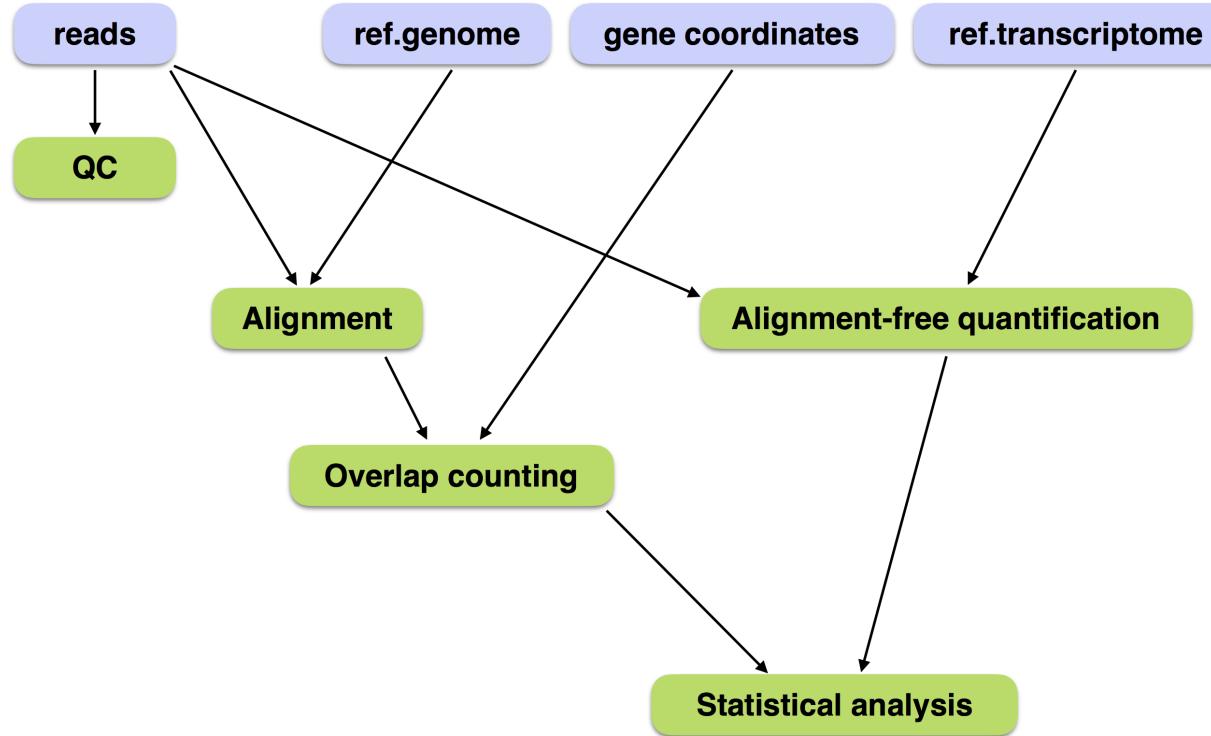
Multiple possible best practices for every dataset

To get the right answer, you have to pose the right question.

- Does the total output of a gene change between conditions? Differential Gene Expression (DGE)
- Does the expression of individual transcripts change? Differential Transcript Expression (DTE)
- Does any isoform of a given gene change? DTE+G
- Does the isoform composition for a given gene change? Differential Transcript Usage/Differential Exon Usage (DTU/DEU)

Each needs different computational approaches (quantifications + tests)

# Overview of the processing workflow



Ingredients + operations

# The raw data: sequencing reads

FASTQ files: sequence + base quality (phred score)

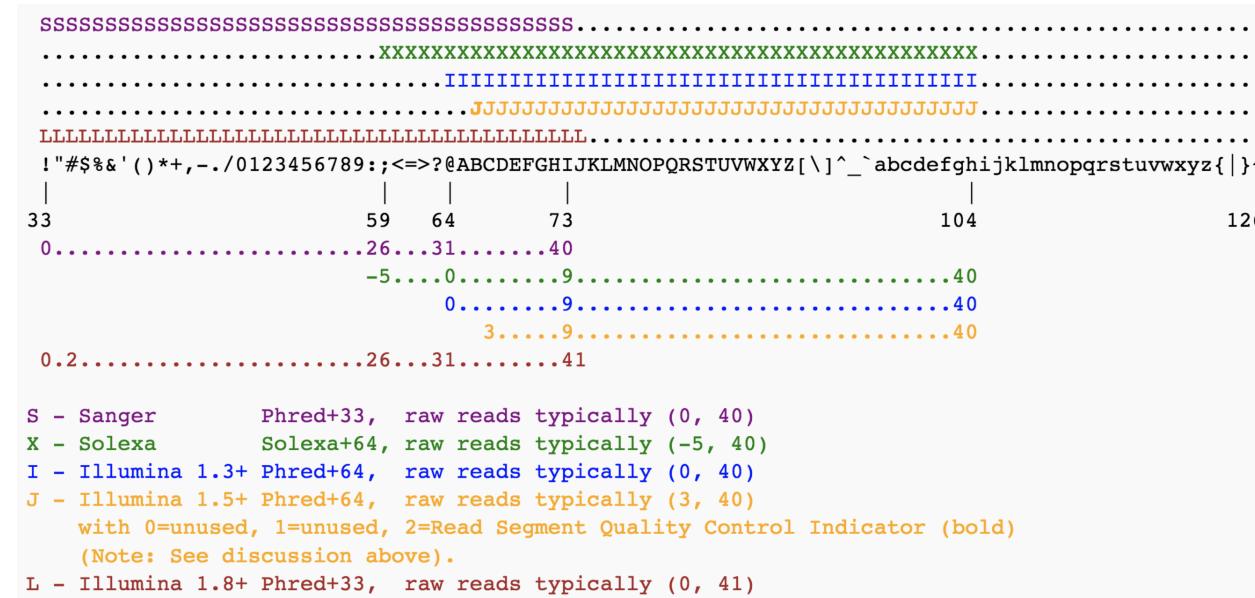
First lines of a FASTQ file

```
@SRR1055095.1 HWI-ST156:397:D09NJACXX:5:1101:1222:1915/1
NGCTGCTGGACTCCGAAGATGGCGGTATATCATCCACTGCTGACTCTN
+
#1=DDFFFHFHHHJJHIIGHIJJGIJAFFDHGGGIJJGJIJJJJIIH#
@SRR1055095.2 HWI-ST156:397:D09NJACXX:5:1101:1245:1920/1
NCTTTTCTTGTTCTCATCATCTTCAGGAGGGAGGGTCATCCTTGTGN
+
#1=BDB:?:FFDFFDF?EF<FFF>B>?C@CF<1??CFB:09;09BFE9DB#
```

Repeat this tens of millions of times, and you'll have *one* sample

# The raw data: sequencing reads

Different quality encodings exist



# The raw data

Demo: quality control report, from FastQC

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good\\_sequence\\_short\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html)

Your next best friend: MultiQC

Common sequence artefacts in NGS data:

- read errors
- base calling errors
- small insertions and deletions
- poor quality reads
- primer/adapter contamination

Solutions: Quality trimming & filtering (wide range of QC tools available)

# Reference files

Reference genome sequences (in `fasta` format), required for genome alignment.

Think of the alignment as the address of each read in a 3-billion houses street, where the elements along the street can also end up repeating themselves.

## What's out there?

Ensembl: <http://www.ensembl.org/info/data/ftp/index.html>

Gencode (human & mouse): <https://www.gencodegenes.org/>

UCSC: <http://hgdownload.cse.ucsc.edu/downloads.html>

iGenome: [http://support.illumina.com/sequencing/sequencing\\_software/igenome.html](http://support.illumina.com/sequencing/sequencing_software/igenome.html)

# Reference files

Some critical points:

Be consistent!

Different chromosome identifiers!

Reference genomes and annotations are continuously refined, extended and improved

Keep track of version and be consistent!

Naming of genes can vary across versions, in some databases!

# Reference files

## An example:

[http://ftp.ensembl.org/pub/release-106/fasta/homo\\_sapiens/dna/](http://ftp.ensembl.org/pub/release-106/fasta/homo_sapiens/dna/)

... and one level up...

[http://ftp.ensembl.org/pub/release-106/fasta/homo\\_sapiens/](http://ftp.ensembl.org/pub/release-106/fasta/homo_sapiens/)

# Reference files

## The GTF format

```
chr1    unknown exon    11874    12227    .        +        .        gene_id "DDX11L1"; gene_name "DDX11L1";
transcript_id "NR_046018"; tss_id "TSS16107";
chr1    unknown CDS     3427347  3427466 .        -        2        gene_id "MEGF6"; gene_name "MEGF6";
p_id "P34437"; transcript_id "NM_001409"; tss_id "TSS31177";
```

- One line per "feature" (exon, transcript, gene, CDS, 3'UTR, 5'UTR, ...)
- One feature = 9 columns of data, plus optional track definition lines
- Essential for releasing annotation information

seqname – name of chromosome/scaffold

source – data/program source

feature – feature type name

start – positions of the feature

end

score – floating point value

strand – forward or reverse

frame – 0|1|2

attribute – semicolon-separated list of tag-value pairs, providing additional information

# Bioconductor



Home      Install      Help      Developers      About

Search:

## About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

## Bioc2020

Get the latest updates on the [Bioc2020 Conference!](#)

- Registration is Now Open! [Register Today!](#)
- Call for Abstracts! If you are interested in presenting a workshop, poster, or talk please [submit your proposal](#). Deadline March 3rd.
- Apply for [Travel Scholarships](#). Deadline March 3rd.

## Bioconductor is hiring!

Bioconductor is hiring for a [full-time position](#) on the Bioconductor Core

### Install »

- Discover [1823 software packages](#) available in *Bioconductor* release 3.10.

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

### Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

### Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment](#) packages
- [Docker](#) and [Amazon](#) machine images
- Latest [release announcement](#)
- Use Bioconductor in the [AnVIL](#). Bioconductor [AnVIL Project Updates](#)
- [Community Slack](#) sign-up
- [Support site](#)

### Develop »

Contribute to *Bioconductor*

- [Developer resources](#)
- [Use Bioc 'devel'](#)
- 'Devel' packages
- [Package guidelines](#)
- [New package submission](#)
- [Git source control](#)
- [Build reports](#)

# Bioconductor - soon your best friend

- an open source project
- a repository of packages, focused on bioinformatics/computational biology
- a open development platform and community

Currently (June 2023)

- 2230 software packages
- 912 annotation packages
- 419 experiment packages
- 27 workflows
- 3 books

**Aim:** interdisciplinary research, collaboration and rapid development of scientific software

## Documentation

- function manual pages, most of them with runnable examples
- package vignettes - mandatory here!
- workflows, documenting full analyses spanning multiple tools
- a very active support site

# The real deal: Bioconductor's community

sign up / log in • about • faq • rss 

 Bioconductor  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION    LATEST    NEWS    JOBS    TUTORIALS    TAGS    USERS

Limit ▾    Sort ▾    Search

0 votes    2 answers    26 views    Use ddCT with gene specific efficiencies?

ddct efficiency gene specific efficiencies

written 10 days ago by bettenbrock • 0 • updated 24 minutes ago by Zhang, Jitao David • 110

0 votes    0 answers    18 views    How to use panelcn.mops to detect CNVs from whole genome data by getting count windows from BED file?

cn.mops cnv panelcn.mops cnv detection

written 3 days ago by metzgerlukas • 0

0 votes    0 answers    23 views    multifactor desing with interaction LRT or Wald test?

deseq2 wald lrt

written 7 hours ago by jnaviapelaez • 0

0 votes    1 answer    25 views    Meaning of log2 fold change in 2 x 4 interaction

deseq2

written 2 days ago by julia.chariker • 10 • updated 20 hours ago by Michael Love • 27k

0 votes    0 answers    30 views    DropletUtils::swappedDrops more cells than using Read10x with filtered\_feature\_bc\_matrix files

dropletutils single cell rna seq 10x genomics

written 1 day ago by gil.stelzer • 0

0 votes    1 answer    25 views    package to analyze generic protein microarray data

microarray protein analysis proteomeprofiler

written 1 day ago by hcnbox • 0 • updated 1 day ago by Gordon Smyth • 40k

1 vote    1 answer    40 views    In DESeq2, how do you interpret results based on the order of variables in the "contrast" argument?

deseq2

written 1 day ago by gpreising • 0 • updated 1 day ago by Kevin Blighe • 460

3 votes    1 answer    51 views    Conversion of LogFC to FC

limma logfc fc

written 1 day ago by nia • 10 • updated 1 day ago by Gordon Smyth • 40k

Recent...

Replies

- C: Multiple testing across ... by fl • 0
- A: Use ddCT with gene speci... by Zhang, Jitao David • 110
- C: How to use panelcn.mops ... by Kevin Blighe • 460
- C: Conversion of LogFC to FC by Aaron Lun • 25k
- C: DESeq2: experiment with ... by Michael Love • 27k

Votes

- Error while uploading Bioc...
- A: Error while uploading Bi...
- C: Error while uploading Bi...
- A: Multiple testing across ...
- C: Conversion of LogFC to FC

Awards • All »

- Autobiographer  to jacknick1996 • 0
- Autobiographer  to smithcarter240191 • 0
- Autobiographer  to rk6415153 • 0
- Autobiographer  to lawyersforlandlords • 0
- Autobiographer  to casinolife24 • 0
- Scholar  to Mike Smith • 4.2k

Locations • All »

- Spain, 3 minutes ago
- Switzerland, 24 minutes ago
- EMBL Heidelberg / de.NBI, 27 minutes ago

# Data processing



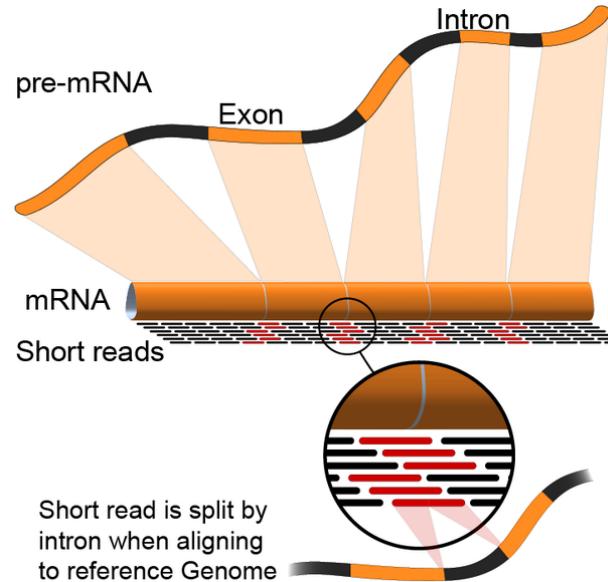
# Data processing

Turning millions of text lines into properly structured abundance tables

Our aim: we often want to compare abundance (expression) of genes or other features between conditions

Splice-aware genome alignment vs "direct" transcript mapping and quantification

# Alignment: not just simple mapping



For RNA-seq data, we need a splice-aware aligner

Common choices:

- STAR
- HISAT2

# Alignment

File-format: [sam](#) (compressed into [bam](#))

```
SRR1055095.6079377 353 chr1    11167    0      50M    =    11751    634    CGCCCCTTGCTTGAGCCGGGCAC
TACAGGACCCGCTTGCTCACGGTGAA    CCCFFFFFFHHHHJJJJJJJJJJJJJJIGJJJJJJJJJJJJJJJDHJJCEHH
AS:i:-10    XN:i:0    XM:i:2    XO:i:0    XG:i:0    NM:i:2    MD:Z:48C0T0    YT:Z:UU    NH:i:20
CC:Z:chrY    CP:i:59361513    HI:i:0
```

Again: repeat this, one read at a time!

Entries:

QNAME - Query NAME of the read or the read pair

FLAG - Bitwise FLAG (pairing, strand, mate strand, etc.)

RNAME - Reference sequence NAME

POS - 1-Based leftmost POSition of clipped alignment

MAPQ - MAPping Quality (Phred-scaled)

CIGAR - Extended CIGAR string (operations: MIDNSHP)

MRNM - Mate Reference NaMe ('=' if same as RNAME)

MPOS - 1-Based leftmost Mate POSition

ISIZE - Inferred Insert SIZE

SEQ - Query SEQuence on the same strand as the reference

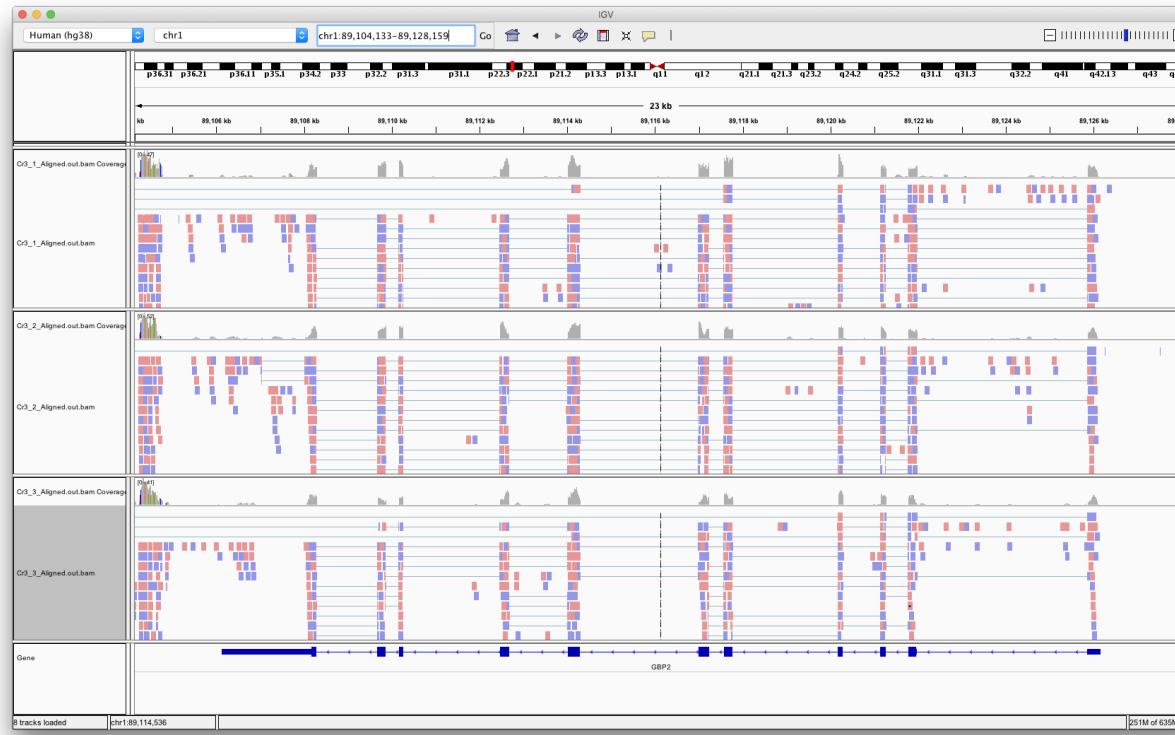
QUAL - Query QUALity (ASCII-33=Phred base quality)

Tags: used to store info about alignment

# Before quantification

... and actually, always: Do visualize your data!

Options: UCSC Genome Browser, IGV, IGB - <http://software.broadinstitute.org/software/igv/>

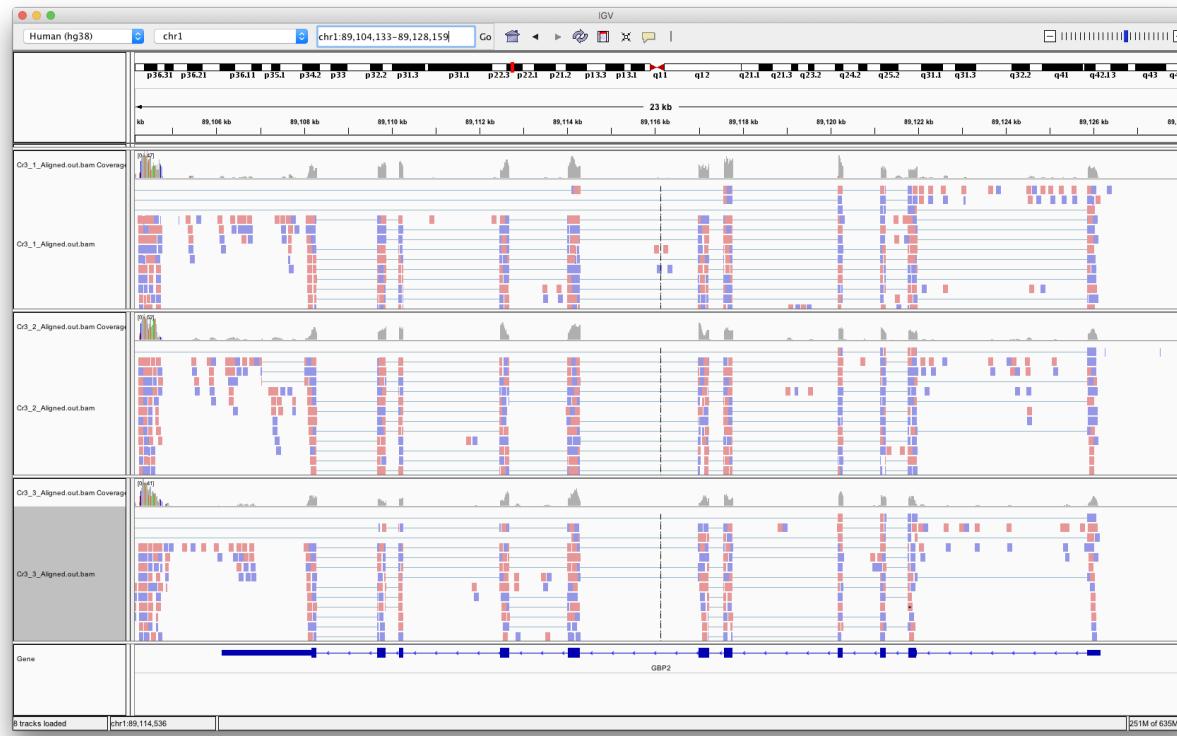


# Quantification

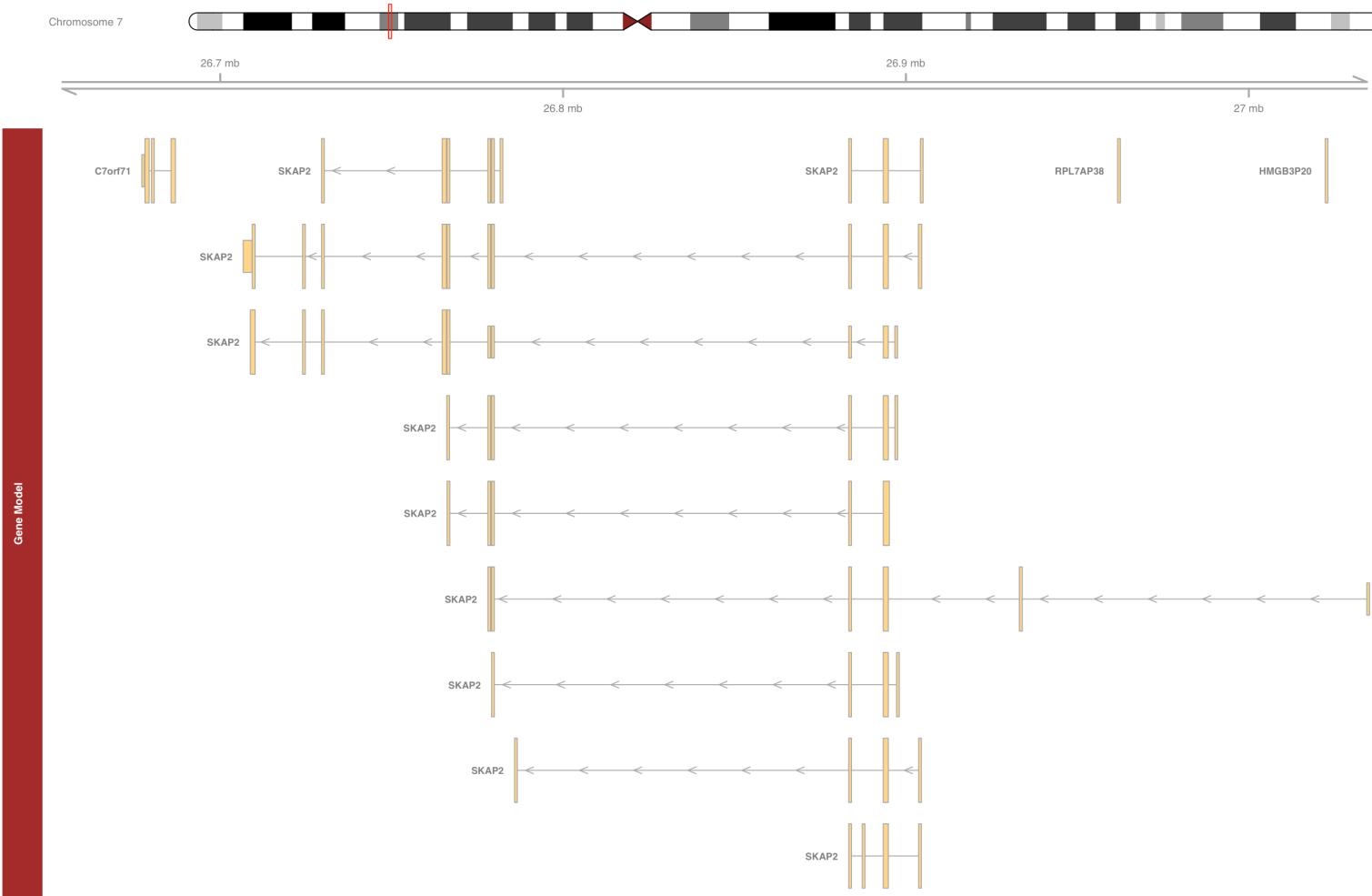
Ingredients: BAM data + GTF annotation file

Output: number of reads overlapping known features (discrete, positive, skewed)

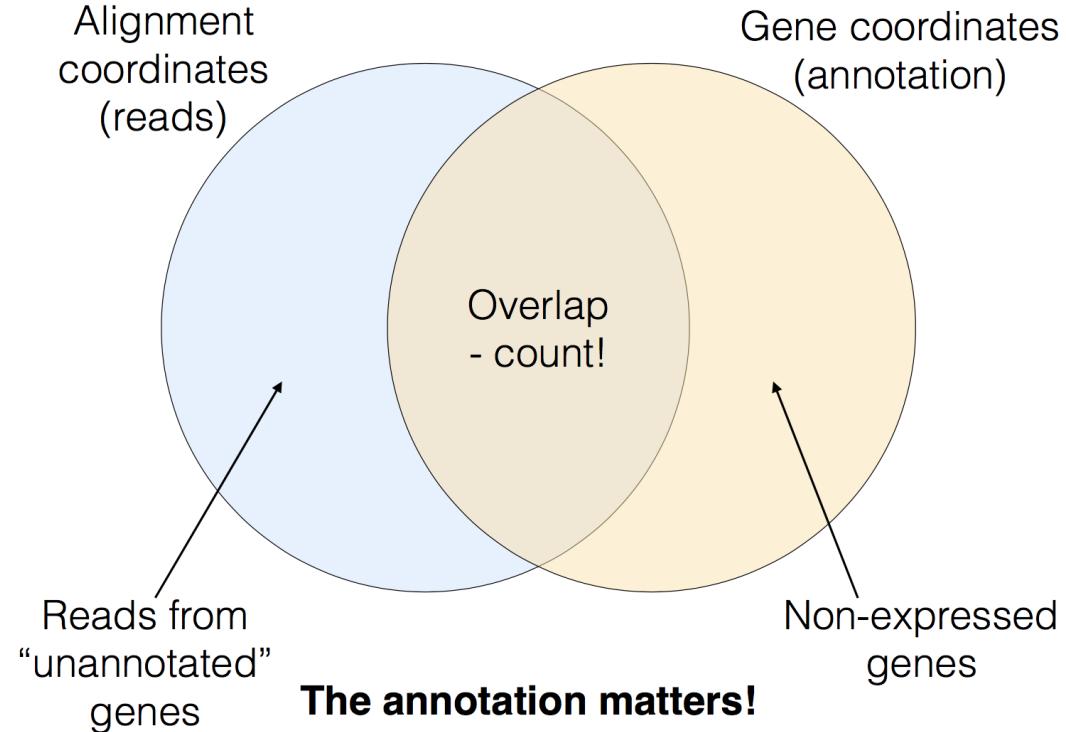
Gene-level counts, often obtained by genome alignment + overlap counting



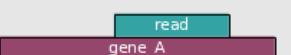
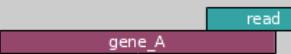
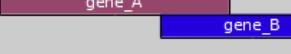
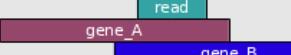
# The role of annotation



# The annotation matters!



# The annotation matters!

	union	intersection _strict	intersection _nonempty
 A single read overlaps a single gene (gene_A).	gene_A	gene_A	gene_A
 A single read overlaps a single gene (gene_A), but only a portion of the read is within the gene's boundaries.	gene_A	no_feature	gene_A
 A single read spans across two genes (gene_A and gene_A).	gene_A	no_feature	gene_A
 Two separate reads overlap two separate genes (gene_A and gene_A).	gene_A	gene_A	gene_A
 A single read overlaps two genes (gene_A and gene_B).	gene_A	gene_A	gene_A
 A single read overlaps two genes (gene_A and gene_B), and both genes are annotated as nonunique.	ambiguous (both genes with --nonunique all)	gene_A	gene_A
 A single read overlaps two genes (gene_A and gene_B), where gene_A is unique and gene_B is nonunique.	ambiguous (both genes with --nonunique all)		
 A single read overlaps two genes (gene_A and gene_B), and both genes are annotated as unique.	alignment_not_unique (both genes with --nonunique all)		

# Alignment-free quantifications

Some recently developed methods:

- salmon (Patro et al, Nat Methods 2017)
- kallisto (Bray et al, Nat Biotechnol 2016)

return...

- Transcript-level counts and TPM (transcripts-per-million) estimates, which can be summed up to get
- Gene-level counts and TPM estimates

Pros & cons

- considerably faster than traditional alignment+counting -> allow bootstrapping
- more highly resolved estimates (transcripts rather than gene) + can be aggregated
- can use a slightly larger fraction of the reads since multi-mapping reads are not excluded
- don't return precise alignments (bam files, for e.g. visualization in genome browser)

# Which way to go?

Based on genome alignment - mainly gene-level quantification: combine exons, "ignoring" splice variants

- Simple, powerful, yet in some cases inaccurate
- Tools:
  - `htseq-count`, `featureCounts` for estimating expression levels (counts)
  - `edgeR`, `DESeq2`, `voom+limma` for statistical modeling

Based on transcriptome mapping - transcript- and gene-level quantification: 'assign' reads (or rather, estimate most likely expression level) based on probabilistic modeling

- Potentially cleaner, but high degree of uncertainty on the transcript level!
- Tools:
  - `bitSeq`, `RSEM`, `salmon`, `kallisto` for (pseudo)alignment/quantification
  - `DESeq2`, `edgeR`, `voom+limma`, `swish`, `DRIMseq`, `DEXSeq`, `sleuth` for modeling (depending on the question of interest)

# What it would look like - STAR + featureCounts:

... due to time constraints

Index the genome

```
$ STAR --runThreadN 24 \
    --runMode genomeGenerate \
    --genomeDir my_genome \
    --genomeFastaFiles my_genome.fa \
    --sjdbGTFfile my_genes.gtf \
    --sjdbOverhang 99
```

# What it would look like - STAR + featureCounts:

... due to time constraints

Map each file

```
$ STAR --runThreadN 24 \
    --runMode alignReads \
    --genomeDir my_genome \
    --readFilesIn my_sample_read1.fastq.gz \
                  my_sample_read2.fastq.gz \
    --readFilesCommand zcat \
    --outFileNamePrefix output/S1/ \
    --outSAMtype BAM SortedByCoordinate \
    --quantMode GeneCounts
```

# What it would look like - STAR + featureCounts:

 SRR1039508	▶	 SRR1039508_Aligned.sortedByCoord.out.bam
 SRR1039509	▶	 SRR1039508_Log.final.out
 SRR1039512	▶	 SRR1039508_Log.out
 SRR1039513	▶	 SRR1039508_Log.progress.out
 SRR1039516	▶	 SRR1039508_ReadsPerGene.out.tab
 SRR1039517	▶	 SRR1039508_SJ.out.tab
 SRR1039520	▶	
 SRR1039521	▶	

# What it would look like - STAR + featureCounts:

... due to time constraints

Quantify

```
featureCounts(files = bamfiles,  
              annot.ext = "my_genes.gtf",  
              isGTFAnnotationFile = TRUE,  
              GTF.featureType = "exon",  
              GTF.attrType = "gene_id",  
              useMetaFeatures = TRUE,  
              isPairedEnd = TRUE,  
              strandSpecific = 0)
```

Directly generates a count matrix in your R session.

# What it would look like - salmon

... due to time constraints

Create an index of the transcriptome

```
$ salmon index -i my_transcripts.idx \
    -t <(cat my_transcripts.fasta my_genome.fasta) \
    -d chromosome_names.txt
```

The genome acts as a 'decoy' sequence, to collect reads truly arising from intronic or intergenic locations.

# What it would look like - salmon

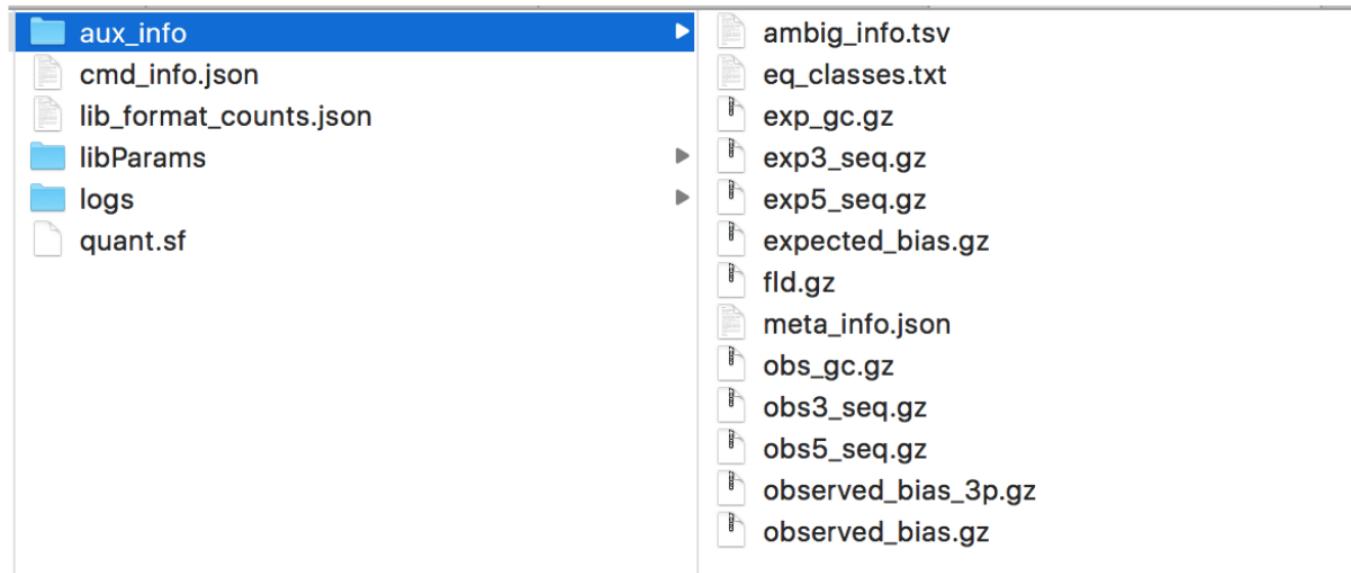
... due to time constraints

Quantify a sample at the transcript level

```
$ salmon quant -i my_transcripts.idx -l A \  
    -1 my_sample_read1.fastq.gz -2 my_sample_read2.fastq.gz \  
    -p 10 -o results/sample1 --validateMappings \  
    --numBootstraps 30 --seqBias --gcBias
```

# What it would look like - salmon

... due to time constraints



# What it would look like - salmon

... due to time constraints

Salmon  
[quant.sf]

Name	Length	EffectiveLength	TPM	NumReads
ENST00000406070	2025	1869.81	0	0
ENST00000446844	2227	2071.81	0.137334	3.71695
ENST00000599620	686	530.936	0	0
ENST00000471557	505	350.256	0.731211	3.3457
ENST00000338761	1456	1300.81	0	0
ENST00000417509	1444	1288.81	7.58582e-08	1.27717e-06
ENST00000484946	610	455.039	2.87905	17.1142
ENST00000490656	660	504.969	1.46703	9.67744
ENST00000439537	1161	1005.81	1.47611	19.3952
ENST00000493251	641	485.994	0.597774	3.79512
ENST00000460127	408	253.708	0	0

# Importing salmon quantifications into R

You can follow (offline) the instructions of the `tximport` package -  
<https://bioconductor.org/packages/tximport/>

`tximeta`: another precious assistant on the way to be consistent and to keep track of provenance identification (we'll see it in action during the exercises)

# What does our data look like now?

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
<b>ENSG00000000003</b>	679	448	873	408	1138	1047	770	572
<b>ENSG00000000005</b>	0	0	0	0	0	0	0	0
<b>ENSG00000000419</b>	467	515	621	365	587	799	417	508
<b>ENSG00000000457</b>	260	211	263	164	245	331	233	229
<b>ENSG00000000460</b>	60	55	40	35	78	63	76	60
<b>ENSG00000000938</b>	0	0	2	0	1	0	0	0
<b>ENSG00000000971</b>	3251	3679	6177	4252	6721	11027	5176	7995
<b>ENSG00000001036</b>	1433	1062	1733	881	1424	1439	1359	1109
<b>ENSG00000001084</b>	519	380	595	493	820	714	696	704
<b>ENSG00000001167</b>	394	236	464	175	658	584	360	269
<b>ENSG00000001460</b>	172	168	264	118	241	210	155	177
<b>ENSG00000001461</b>	2112	1867	5137	2657	2735	2751	2467	2905
<b>ENSG00000001497</b>	524	488	638	357	676	806	493	475
<b>ENSG00000001561</b>	71	51	211	156	23	38	134	172
<b>ENSG00000001617</b>	555	394	905	415	727	697	618	599
<b>ENSG00000001626</b>	10	2	9	2	10	6	5	5
<b>ENSG00000001629</b>	1660	1251	2259	1079	2462	2514	1888	1660
<b>ENSG00000001630</b>	59	54	66	23	84	87	31	59
<b>ENSG00000001631</b>	729	692	943	475	1034	1163	731	744
<b>ENSG00000002016</b>	201	161	256	99	268	257	160	137
<b>ENSG00000002079</b>	3	0	3	1	4	0	0	1
<b>ENSG00000002330</b>	206	174	184	111	194	260	156	177

# Some challenges in RNA-seq data analysis

- 1 - Choosing an appropriate statistical distribution
- 2 - Normalization between samples
- 3 - Few samples available make it difficult to estimate parameters (e.g., variance)
- 4 - Many genes, many tests - high dimensionality

# Some challenges in RNA-seq data analysis - 1

## Choosing an appropriate statistical distribution

Variance depends on the mean count

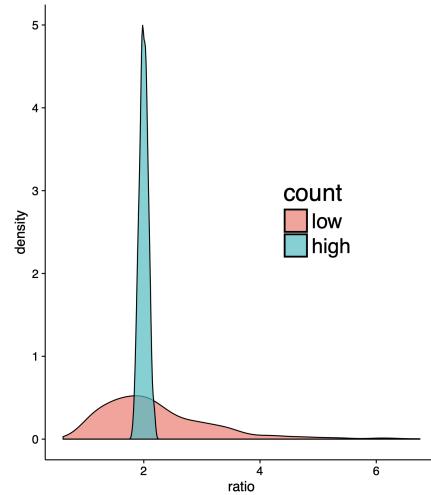
Counts are non-negative and often highly skewed

This means you can't just use t-tests, ANOVA - no prob, `glm`s to the rescue!

Poisson -> negative binomial, better captures variability across biological replicates

# "Why do we not just take the ratios?"

Fold changes, relative abundances



Ex: ratio between two Poisson distributed variables

Low: mean = 20 vs mean = 10

High: mean = 2000 vs mean = 1000

Which one would you trust more? Why?

This goes back to having appropriate statistical frameworks that nicely model your datasets (and how these get generated)

# Some challenges in RNA-seq data analysis - 2

## Normalization between samples

Observed counts depend on:

- abundance level
- gene/transcript length
- sequencing depth
- sequencing biases

"As-is" estimates not directly comparable across samples

Normalization aims to ensure our expression estimates are

- comparable across features (genes, isoforms, etc)
- comparable across libraries (different samples)
- on a human-friendly scale (interpretable magnitude)

Most RNA-seq methods (e.g., edgeR, DESeq2, voom) need raw counts (or equivalent) as input

Don't provide these methods with (e.g.) RPKMs, FPKMs, TPMs, CPMs, log-transformed counts, normalized counts, ...

Read documentation carefully!

# Digression: Normalization expression units

- RPKM/FPKM (Reads/Fragments per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset
- TPM (transcripts per million)
  - normalizes to transcript copies instead of reads - gives an idea of the proportion of transcripts
  - Corrects for cases where the total RNA output differs between samples
  - More appropriate for between sample comparisons (the sum of all TPMs in each sample are the same)

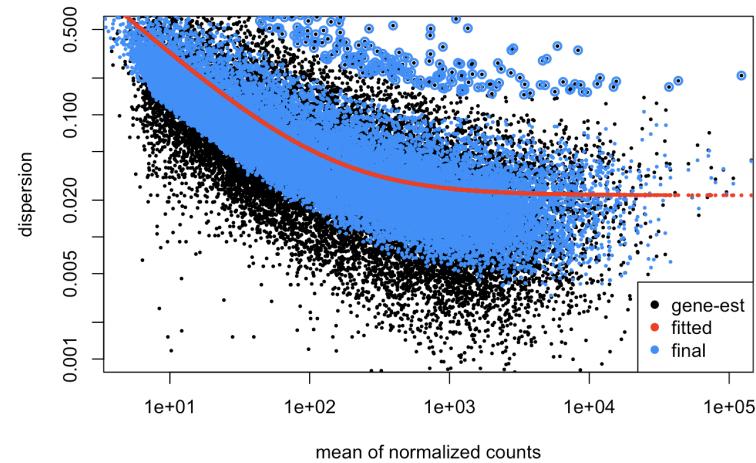
For DE analysis you have to work with discrete counts...

... and for comparisons you can use normalized counts (median ratio/TMM methods are robust across all genes!)

# Some challenges in RNA-seq data analysis - 3

Few samples available make it difficult to estimate parameters (e.g., variance)

You can take advantage of the large number of genes



-> Shrink the gene-wise estimates towards a center value defined by the observed distribution of dispersions across...

- all genes (common dispersion estimate)
- genes with similar expression (trended dispersion estimate)

# Some challenges in RNA-seq data analysis - 4

**Many genes, many tests - high dimensionality**

FDR for multiple test correction

Some more ideas:

- filter out genes that have little chance of showing significance (without looking at the test results)
- independent hypothesis weighting

... all nicely implemented for the DESeq framework

Do not test AND filter on logFC post-hoc!

**Think** of the null you're testing against -  $\beta = 0$  by default, useful to adapt if you want to focus on larger effect sizes

# Exploratory analysis and visualization

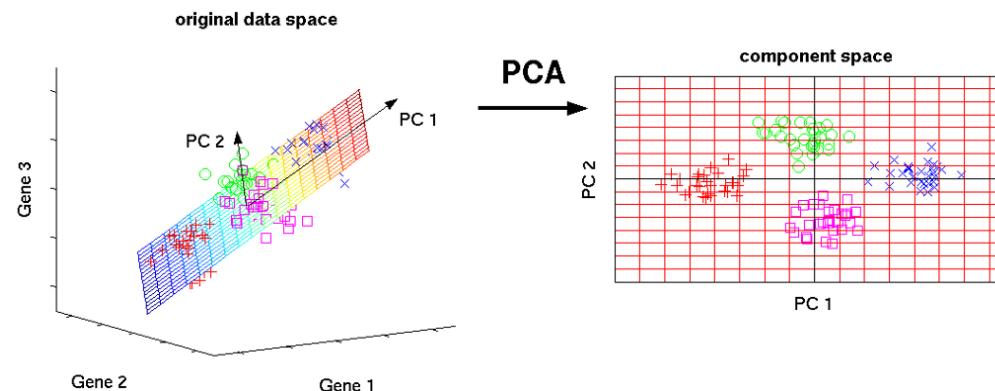
Is the data of good quality?

Quality as "fitness for purpose": DE

Very important: Transforming the data is often required for better further explorations (here: variance stabilization, regularized logarithm...)

One notable example, commonly used: Principal Components Analysis.

The data points (here, the samples) are projected onto the 2D plane such that they spread out in the two directions that explain most of the differences (the variability)



# Exploratory analysis and visualization

"Perspective matters"



# Gene identifiers

Take for example **GBP2**, guanylate binding protein 2 (4 transcripts available)

Ensembl ID: ENSG00000162645 (human), ENSMUSG00000028270 (mouse)

<http://www.ensembl.org/id/ENSG00000162645>

Entrez ID: 2634

HGNC ID: 4183

RefSeq ID: NM\_004120

UCSC ID: uc001dmz.3

Official symbol: GBP2

Synonyms: - (sometimes makes the ambiguity even bigger!!!)

Gene symbols can change over time!

Typically, no 1:1 mapping between different ID types

# Working with Excel

Don't do it - and not just because I do like R

Correspondence | [Open Access](#) | Published: 23 June 2004

## Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

[Barry R Zeeberg](#), [Joseph Riss](#), [David W Kane](#), [Kimberly J Bussey](#), [Edward Uchio](#), [W Marston Linehan](#), [J Carl Barrett](#) & [John N Weinstein](#) 

[BMC Bioinformatics](#) **5**, Article number: 80 (2004) | [Cite this article](#)

**113k** Accesses | **43** Citations | **515** Altmetric | [Metrics](#)

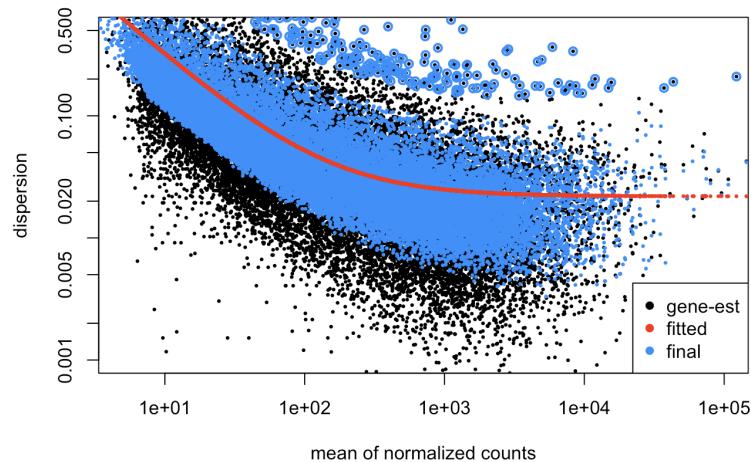
# Working with Excel

There are excellent reasons *not* to do it

# Differential expression analysis, with DESeq2

There is one main function, `DESeq`, that does the following

- estimation of size factors
- estimation of the dispersion values for each gene
- fitting of the generalized linear model and performing statistical testing



# What do our results look like?

	<b>id</b>	<b>baseMean</b>	<b>log2FoldChange</b>	<b>IfcSE</b>	<b>stat</b>	<b>pvalue</b>	<b>padj</b>
<b>1</b>	ENSG00000152583	997.43977	-4.5749187	0.18405609	-24.856111	2.220933e-136	4.003898e-132
<b>2</b>	ENSG00000165995	495.09291	-3.2910618	0.13317370	-24.712551	7.839410e-135	7.066444e-131
<b>3</b>	ENSG00000120129	3409.02938	-2.9478099	0.12143769	-24.274258	3.666925e-130	2.203577e-126
<b>4</b>	ENSG00000101347	12703.38706	-3.7669954	0.15543799	-24.234715	9.583815e-130	4.319425e-126
<b>5</b>	ENSG00000189221	2341.76725	-3.3535799	0.14178235	-23.653014	1.098955e-123	3.962392e-120
<b>6</b>	ENSG00000211445	12285.61515	-3.7304027	0.16583058	-22.495263	4.618318e-112	1.387651e-108
<b>7</b>	ENSG00000157214	3009.26322	-1.9767725	0.08998232	-21.968454	5.769975e-107	1.486016e-103
<b>8</b>	ENSG00000162614	5393.10168	-2.0356649	0.09418231	-21.614091	1.323849e-103	2.983294e-100
<b>9</b>	ENSG00000125148	3656.25278	-2.2109786	0.10564078	-20.929214	2.902397e-97	5.813824e-94
<b>10</b>	ENSG00000154734	30315.13547	-2.3456037	0.11580806	-20.254235	3.259713e-91	5.876611e-88
<b>11</b>	ENSG00000139132	1223.46614	-2.2289030	0.11283944	-19.752872	7.578284e-87	1.242012e-83
<b>12</b>	ENSG00000162493	1099.62587	-1.8912170	0.09577959	-19.745511	8.767336e-87	1.317146e-83
<b>13</b>	ENSG00000134243	5510.95826	-2.1957116	0.11200768	-19.603224	1.451321e-85	2.012647e-82
<b>14</b>	ENSG00000179094	776.59667	-3.1917499	0.16396292	-19.466291	2.120834e-84	2.731028e-81
<b>15</b>	ENSG00000162692	508.17023	3.6926606	0.19018590	19.416059	5.646022e-84	6.785765e-81
<b>16</b>	ENSG00000163884	561.10717	-4.4591282	0.23486369	-18.986027	2.225442e-80	2.507517e-77
<b>17</b>	ENSG00000178695	2649.85015	2.5281746	0.13443012	18.806607	6.667027e-79	7.070186e-76
<b>18</b>	ENSG00000198624	2057.19173	-2.9184357	0.16129634	-18.093626	3.577533e-73	3.583098e-70
<b>19</b>	ENSG00000107562	25136.30757	1.9116698	0.10574810	18.077582	4.786167e-73	4.541317e-70
<b>20</b>	ENSG00000148848	1365.17399	1.8145431	0.10048365	18.058094	6.813522e-73	6.141709e-70

I would like to understand more what is going on here

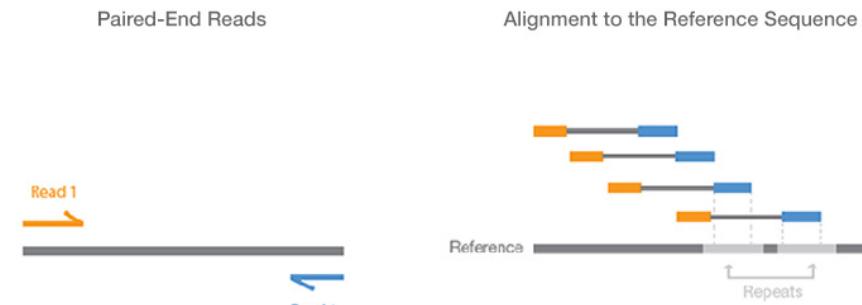
# Practical tips for your (next) RNA-seq dataset

## Single- vs paired-end sequencing

Each fragment can be sequenced from one end only, or from both ends

Single-end cheaper and faster

Paired-end provide improved ability to localize the fragment in the genome and resolve mapping close to repeat regions - less multimapping reads



---

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

# Practical tips for your (next) RNA-seq dataset

## Strand-specificity

In “standard” protocols, we don’t know from which strand a read stems

Various “strand-specific” protocols allow us to keep this information

Strand-specificity leads to lower number of ambiguous reads (overlapping multiple genes)

# Practical tips for your (next) RNA-seq dataset

## Sequencing depth and number of replicates

Recommendation: At least 3 biological replicates - Conesa et al, Genome Biol, 2016

Recommendation: At least 6 biological replicates - Schurch

Diminishing returns in increasing the read depth - go for replicates given a fixed budget!

# I am not kidding you



**zack chiang**  
@z\_chiang · [Follow](#)



doing bulk sequencing analysis in 2022



7:28 PM · May 31, 2022



530

Reply

Copy link

[Read 3 replies](#)

# Practical tips for your (next) RNA-seq dataset

Bulk or single cell?

That depends on your question!

Did someone say heterogeneity?

# What do our results look like?

	<b>id</b>	<b>baseMean</b>	<b>log2FoldChange</b>	<b>IfcSE</b>	<b>stat</b>	<b>pvalue</b>	<b>padj</b>
<b>1</b>	ENSG00000152583	997.43977	-4.5749187	0.18405609	-24.856111	2.220933e-136	4.003898e-132
<b>2</b>	ENSG00000165995	495.09291	-3.2910618	0.13317370	-24.712551	7.839410e-135	7.066444e-131
<b>3</b>	ENSG00000120129	3409.02938	-2.9478099	0.12143769	-24.274258	3.666925e-130	2.203577e-126
<b>4</b>	ENSG00000101347	12703.38706	-3.7669954	0.15543799	-24.234715	9.583815e-130	4.319425e-126
<b>5</b>	ENSG00000189221	2341.76725	-3.3535799	0.14178235	-23.653014	1.098955e-123	3.962392e-120
<b>6</b>	ENSG00000211445	12285.61515	-3.7304027	0.16583058	-22.495263	4.618318e-112	1.387651e-108
<b>7</b>	ENSG00000157214	3009.26322	-1.9767725	0.08998232	-21.968454	5.769975e-107	1.486016e-103
<b>8</b>	ENSG00000162614	5393.10168	-2.0356649	0.09418231	-21.614091	1.323849e-103	2.983294e-100
<b>9</b>	ENSG00000125148	3656.25278	-2.2109786	0.10564078	-20.929214	2.902397e-97	5.813824e-94
<b>10</b>	ENSG00000154734	30315.13547	-2.3456037	0.11580806	-20.254235	3.259713e-91	5.876611e-88
<b>11</b>	ENSG00000139132	1223.46614	-2.2289030	0.11283944	-19.752872	7.578284e-87	1.242012e-83
<b>12</b>	ENSG00000162493	1099.62587	-1.8912170	0.09577959	-19.745511	8.767336e-87	1.317146e-83
<b>13</b>	ENSG00000134243	5510.95826	-2.1957116	0.11200768	-19.603224	1.451321e-85	2.012647e-82
<b>14</b>	ENSG00000179094	776.59667	-3.1917499	0.16396292	-19.466291	2.120834e-84	2.731028e-81
<b>15</b>	ENSG00000162692	508.17023	3.6926606	0.19018590	19.416059	5.646022e-84	6.785765e-81
<b>16</b>	ENSG00000163884	561.10717	-4.4591282	0.23486369	-18.986027	2.225442e-80	2.507517e-77
<b>17</b>	ENSG00000178695	2649.85015	2.5281746	0.13443012	18.806607	6.667027e-79	7.070186e-76
<b>18</b>	ENSG00000198624	2057.19173	-2.9184357	0.16129634	-18.093626	3.577533e-73	3.583098e-70
<b>19</b>	ENSG00000107562	25136.30757	1.9116698	0.10574810	18.077582	4.786167e-73	4.541317e-70
<b>20</b>	ENSG00000148848	1365.17399	1.8145431	0.10048365	18.058094	6.813522e-73	6.141709e-70

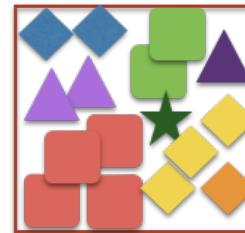
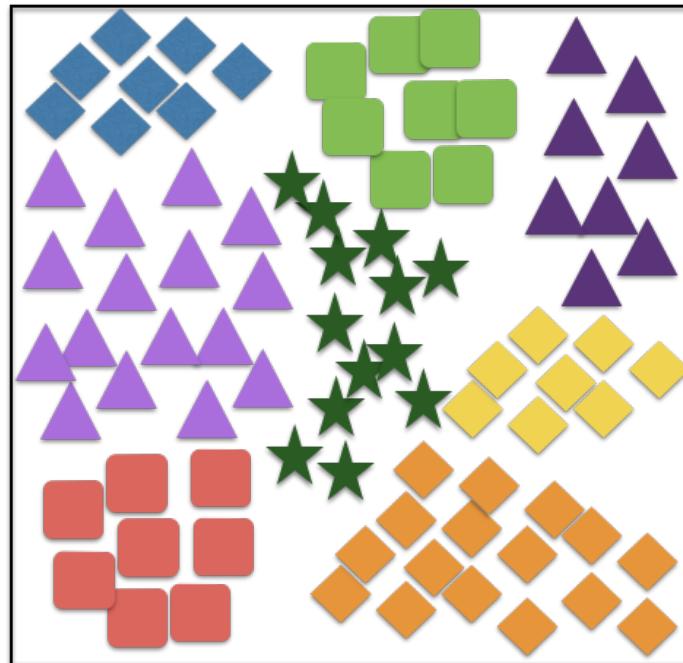
I need to understand more what is going on here



# Functional enrichment analysis

Our problem, at a glance

All known genes in a species  
(categorized into groups)



DEGs

# Functional enrichment analysis

Test whether known biological functions or processes are over-represented (= enriched) in an experimentally-derived gene list, e.g. a list of differentially expressed (DE) genes.

**Example:** Transcriptomic study, in which 12671 genes have been tested for differential expression between two sample conditions and 529 genes were found DE

Among the DE genes, 28 are annotated to a specific functional gene set, which contains in total 170 genes. This setup corresponds to a 2x2 contingency table:

```
deTable <- matrix(  
  c(28, 142, 501, 12000),  
  nrow = 2,  
  dimnames = list(c("DE", "Not_DE"),  
                 c("In_gene_set", "Not_in_gene_set")))
```

```
deTable  
#           In_gene_set Not_in_gene_set  
# DE             28          501  
# Not_DE        142         12000
```

```
fisher.test(deTable, alternative = "greater")  
#  
#      Fisher's Exact Test for Count Data
```

# Some terminology: Gene sets, pathways, networks

**Gene sets** are simple lists of usually functionally related genes without further specification of relationships between genes. **Any *a priori*** classification of genes into biologically relevant groups. Sets do not need to be exhaustive or disjoint

**Pathways** can be interpreted as specific gene sets, typically representing a group of genes that work together in a biological process.

**Gene regulatory networks** describe the interplay and effects of regulatory factors (such as transcription factors and microRNAs) on the expression of their target genes.

Is the expression of genes in a gene set associated with the experimental conditions?

Genes categories	Organism-specific Background	DE results	Over-represented?
Functional category 1	35/13000	25/1000	Likely
Functional category 2	56/13000	4/1000	Unlikely
Functional category 3	90/13000	8/1000	Unlikely
Functional category 4	15/13000	10/1000	Likely
...			
...			

# Functional enrichment analysis

Primarily developed and applied on transcriptomic data, now extended and applied also in other fields of genomic and biomedical research (proteomic and metabolomic data, genomic regions, ...)

Many methods available!

- Over-representation analysis (ORA) - are differentially expressed (DE) genes in the set more common than expected? Based on (variations) of the 2x2 contingency table method
- Functional class scoring (FCS) - summarize gene set (= functional class) scores by summarizing statistic of DE of genes in a set, and compare to null
- Pathway topology (PT) - explicitly taking into account interactions between genes as defined in signaling pathways and gene regulatory networks

Topology-based methods appear to be most realistic, but: features that are not-detectable on the transcriptional level + insufficient network knowledge

**Cautious interpretation of results is required to derive valid conclusions!**

# Collections of gene sets

## Gene Ontology ([GO](#)) Annotation (GOA)

- CC Cellular Components
- BP Biological Processes
- MF Molecular Function

## Pathways

- [MSigDb](#)
- [KEGG](#)
- [reactome](#)
- [PantherDB](#)
- ...

GO and KEGG annotations are most frequently used for the enrichment analysis of functional gene sets - likely due to their long-standing curation and availability for a wide range of species

# Software for functional enrichment

- DAVID
- GSEA and its variations
- inside Bioconductor: [reactome](#), [topGO](#), [pathview](#), [GOSeq](#), ...
- Ingenuity Pathway Analysis

Not straightforward, but amazing way of helping generating hypothesis for the bench scientist

**Friendly tip:** wrappers for topGO and goseq are in [pcaExplorer](#) and in [ideal!](#)

## General recommendations

Never forget to visualize your data!

Pick candidates, sets of candidates, plot instead of guessing!

The choice of a background matters!

# Background matters

... with a slight modification of the example above

```
deTable_detected <- matrix(  
  c(8, 142, 501, 12000),  
  nrow = 2,  
  dimnames = list(c("DE", "Not_DE"),  
                 c("In_gene_set", "Not_in_gene_set")))  
fisher.test(deTable_detected, alternative = "greater")
```

versus...

```
deTable_allgenes <- matrix(  
  c(8, 142, 501, 32000),  
  nrow = 2,  
  dimnames = list(c("DE", "Not_DE"),  
                 c("In_gene_set", "Not_in_gene_set")))  
fisher.test(deTable_allgenes, alternative = "greater")
```

# Data analysis: where's the real bottleneck?

- ✓ Efficient methods to process your data (entire workflows available, see e.g. [the Bioconductor workflow packages](#))
- ✓ Compelling ways of visualizing your data (better QC, better hypotheses, better answers → see <http://bioconductor.org/packages/iSEE/> & <http://bioconductor.org/packages/iSEEu/> )
- ✓ Powerful framework to communicate, interactively and reproducibly

## 🔥 Data interpretation



... just not in plain sight

Show me your data...

	A	B	C	D	E	F	G	H	I	J	K
1	SAMEA103885102	SAMEA103885347	SAMEA103885043	SAMEA103885392	SAMEA103885182	SAMEA103885136	SAMEA103885413	SAMEA103884967	SAMEA103885368	SAMEA103885218	SAMEA103885319
2	55,69716914	96,38981386	67,26753412	142,6835597	186,1485543	122,5926771	170,5541862	203,0472731	70,5820175	30,98336005	71,87276974
3	0	0	0	0	0	0	0	0	0	0	0
4	964,8925057	924,83214	1036,106869	1046,57092	1014,635976	1083,422465	969,604741	982,6266614	954,7538272	1035,036116	1099,503914
5	351,9054782	684,638914	387,1752238	820,970061	375,1488073	948,9162924	503,55028	1231,24557	300,7527914	911,6795373	889,3737712
6	677,4205719	221,0877036	276,2026831	97,82962948	395,4077973	175,5987827	179,7898584	133,993082	675,5371526	153,7212284	319,7823597
7	132,782876	198,7016694	662,811222	1522,435008	386,9046409	560,2933294	1257,995559	5812,826356	909,7667115	2395,917116	2048,353449
8	362,6435931	5375,365498	650,3552061	7471,729136	116,9579367	5036,571099	223,3816218	5507,177186	66,70469095	5451,589642	480,5118025
9	5185,999104	4426,618781	5870,757295	5878,363213	3769,061292	3829,066212	2882,333961	3365,101977	3953,382129	3625,846869	4932,212822
10	4064,476563	10213,97627	2388,342021	10237,20614	3863,969772	13282,99842	2921,831249	11699,4626	2120,792061	6985,444675	2949,988682
11	996,2886541	1107,437138	1481,523638	1905,441941	1296,389901	1195,809922	1705,234892	1943,638311	981,0211341	1280,849815	1550,697703
12	78,82859697	104,2567382	130,0656527	75,32368946	90,65332179	69,80311374	172,8622743	73,59486838	47,6298952	86,1652492	
13	372,4875108	375,1582833	336,9448447	378,4351496	484,7559515	589,1766985	548,4876464	415,3602864	281,9806048	533,7955594	265,9931667
14	1004,363486	804,9305608	2122,334216	566,7309848	894,7410645	609,4980038	1607,317126	355,0574378	1234,02879	589,4720116	1880,361894
15	47,6088729	190,9341388	47,12634333	206,4662057	145,2740402	422,2343534	150,387129	287,7915443	53,8085684	313,9134375	92,61472368
16	10,2940199	10,73067656	18,42333204	34,72412059	11,87265036	24,24830727	21,76451809	77,31764486	6,892555164	4,824180733	11,23090998
17	8,461862162	0,100956323	6,053534158	14,24249154	8,595001149	0	0,22640779	17,29978818	4,094109524	5,538925825	21,58950506
18	1265,474946	1617,700447	1669,609838	2297,800891	1288,313487	1497,771426	1576,57678	2747,84197	890,4337923	1037,283091	1936,845356
19	2742,714674	1367,003771	5053,573294	2427,122201	1532,174172	1042,918499	2445,006435	1099,595415	1781,289062	913,0693244	2293,72266
20	739,0239221	721,4533065	978,5249905	869,4506791	758,6712728	687,2941888	958,8405469	1131,158723	723,1258083	891,5201189	1172,627644
21	153,0850266	223,6814932	90,69677231	99,67707173	165,2491213	181,6727726	165,8323252	124,8435309	135,6806311	172,4092075	154,5518966
22	0	0	0	0	10,63374428	1,275177865	4,407102394	0,415747599	0,615201563	4,740536892	0
23	294,9241388	200,997003	193,0178533	259,2214914	272,6918871	292,386033	180,4896341	179,036124	382,9993779	271,7909482	229,4676362
24	10311,29607	40953,9043	14546,93203	48197,23868	12049,09318	61320,54376	15556,91086	48401,85979	6945,254193	62689,66844	24693,50787
25	14252,63718	8840,178525	18210,18028	12299,01927	11043,55437	8344,372672	11131,06854	8353,885536	9092,021277	6232,986277	11739,0848
26	2129,428152	471,0255994	287,0384441	86,35857737	2472,120332	559,3936962	486,4665709	66,4070027	5038,903217	1219,154589	445,1902436
27	0	3,371313144	5,038960845	1,962785986	19,06817927	25,57466411	19,61201231	32,07074799	2,764897276	7,007072669	3,240100076
28	0,4707608	0	0	0	0	0	0	0	0	0	0
29	0,83189235	0,56534712	0,780066877	1,377008507	1,834964003	3,243938183	1,613445267	35,93939273	21,80390352	0,697877051	0,497286622
30	660,247462	697,0714736	376,575126	468,8681819	1120,360629	875,9863134	416,0590986	737,4089693	1126,114691	873,0030589	707,054169
31	15314,39018	12764,13479	5147,168455	7605,221356	17247,28895	19126,9456	5236,812531	11044,68781	16448,27488	19508,61894	5852,803153
32	2828,408307	2687,797465	2800,493773	6277,1227	2275,27131	2454,51924	1339,12723	2779,114942	2188,442779	2400,119315	1476,802287
33	40,23728704	299,899605	56,13562799	568,1032424	385,3034226	2407,236342	279,7861222	2043,400943	69,42871741	1174,143223	83,05725953
34	13078,03109	14673,85241	12869,05451	14992,04765	11871,1049	16338,89938	10146,36181	10380,14905	10500,81202	14966,33552	10032,07621

Show me your data...

	A	B	C	D	E	F	G	H	I	J	K		
1	SAMEA103885102	SAMEA103885347	SAMEA103885043	SAMEA103885392	SAMEA103885182	SAMEA103885136	SAMEA103885413	SAMEA103884967	SAMEA103885368	SAMEA103885218	SAMEA103885319		
2	55,69716914	96,38981386	67,26753412	142,6835597	186,1485543	122,5926771	170,5541862	203,0472731	70,5820175	30,98336005	71,87276974		
3	0	0	0	0	0	0	0	0	0	0	0		
4	964,8925057	924,83214	1036,106869	1046,57092	1014,635976	1083,422465	969,604741	982,6266614	954,7538272	1035,036116	1099,503914		
5	351,9054782	684,638914	387,1752238	820,970061	375,1488073	948,91	A	B	C	D	E		
6	677,4205719	221,0877036	276,2026831	97,82962948	395,4077973	175,59	id	baseMean	logFoldChange	IfcSE	stat		
7	132,782876	198,7016694	662,811222	1522,435008	386,9046409	560,29	1	ENSG00000125347	30585,07489	5,560123732	0,218266321		
8	362,6435931	5375,365498	650,3552061	7471,729136	116,9579367	5036,5	2	ENSG00000162645	36790,59205	6,665803772	0,286573165		
9	5185,999104	4426,618781	5870,757295	5878,363213	3769,061292	3829,0	3	ENSG00000111181	688,5779196	4,710591861	0,196102275		
10	4064,476563	10213,97627	2388,342021	10237,20614	3863,969772	13282,	4	ENSG00000174944	1324,397476	9,863903776	0,471512737		
11	996,2886541	1107,437138	1481,523638	1905,441941	1296,389901	1195,8	5	ENSG00000100336	2515,478907	8,708505005	0,214209592		
12	78,82859697	104,2567382	130,0656527	75,32368946	90,653	6	ENSG00000145365	3637,222141	5,193447761	0,236854233	7,1512E-79	3,43439E-75 P2RY14	
13	372,4875108	375,1582833	336,9448447	378,4351496	484,7559515	589,17	7	ENSG00000137496	7129,418865	4,05866209	0,177364585	3,85308E-70	1,14346E-66 TIFA
14	1004,363486	804,9305608	2122,334216	566,7309848	894,7410645	609,49	8	ENSG00000120425	7129,418865	17,24505541	1,21896E-66	3,10068E-63 L18BP	
15	47,6088729	190,9341388	47,12634333	206,4662057	145,2740402	422,23	9	ENSG00000154451	1714,98786	10,37647789	0,551586848	16,9990962	8,33958E-65 1,85618E-61 GBP5
16	10,2940199	10,73067656	18,42333204	34,72412059	11,87265036	24,248	10	ENSG00000129244	682,8305656	7,191566714	0,381506097	16,2292733	3,13145E-59 6,19539E-56 ATP1B2
17	8,461862162	0,100956323	6,053534158	14,24249154	8,595001149		11	ENSG00000131389	5147,824664	5,593758623	0,28477475	16,13120061	1,53997E-58 2,74208E-55 HLA-DPA1
18	1265,474946	1617,700447	1669,609838	2297,800891	1288,313487	1497,7	12	ENSG00000204257	1909,416235	4,071574682	0,190921436	16,08816042	3,08879E-58 4,99991E-55 HLA-DMA
19	2742,714674	1367,003771	5053,573294	2427,122201	1532,174172	1042,9	13	ENSG00000100342	4112,437557	7,13529667	0,39590635	15,49683826	3,64388E-54 5,40691E-51 APOL1
20	739,029221	721,4533065	978,5249905	869,4506791	758,6712728	687,29	14	ENSG00000162654	52929,49435	9,752769157	0,569217591	15,37684234	2,34089E-53 3,2063E-50 GBP4
21	153,0850266	223,6814932	90,69677231	99,6770173	165,2491282	181,67	15	ENSG00000133321	1393,428038	7,448881999	0,419830443	15,36053451	3,01082E-53 3,82933E-50 PLAAT4
22	0	0	0	10,63374428	1,275177865	4,4071	16	ENSG00000128284	13363,55118	7,977215573	0,454494945	15,35158013	3,45663E-53 4,10325E-50 APOL3
23	294,9241388	200,997003	193,0178533	259,2214914	272,6918871	292,3	17	ENSG00000134470	4468,963176	4,300572585	0,220226433	14,98717724	8,90586E-51 9,91111E-48 L15RA
24	10311,29607	40953,9043	14546,93203	48197,23868	12049,09318	61320,	18	ENSG00000172399	714,5095698	7,829885987	0,463112753	14,74778213	3,17943E-49 3,33017E-46 MYO2Z
25	14252,637178	8840,178525	18210,18028	12299,01927	11043,55437	8344,3	19	ENSG00000164509	700,3132408	8,395184784	0,536171751	13,79256698	2,82525E-43 2,7948E-40 L31RA
26	2129,428152	471,0255994	287,0384441	86,35857737	2472,120332	559,39	20	ENSG00000204267	614,1301184	3,740655413	13,74439071	5,50344E-43	5,15759E-40 GBP1P1
27	0	3,371313144	5,038960845	1,962785986	19,06817927	25,574	21	ENSG00000204267	10442,43781	3,458265659	0,179235202	13,71530609	8,22187E-43 7,31993E-40 TAP2
28	0,4707608	0	0	0	0	0	22	ENSG00000254838	135,9997527	5,591075237	0,335403469	13,68821633	1,19404E-42 1,01244E-39 GVINP1
29	0,83189235	0,56534712	0,780066877	1,377008507	1,834964003	3,2439	23	ENSG00000131203	92938,1422	9,545152592	0,644384499	13,26095306	3,89909E-40 3,15578E-37 ID01
30	660,247462	697,0714736	376,575126	468,8681819	1120,360629	875,98	24	ENSG00000213886	2086,328471	10,26938285	0,703626324	13,17372947	9,62317E-37 UBD
31	15314,39018	12764,13479	5147,168455	7605,221356	17247,28895	19126	25	ENSG00000100911	11111,0667	3,358971492	0,179410258	13,14847609	1,73639E-39 1,28825E-36 PSME2
32	2828,408307	2687,797465	2800,493773	6277,1227	2275,27131	2454,	26	ENSG00000168994	38700,36687	5,356767076	0,337264864	12,91793941	3,56563E-38 2,53958E-35 TAP1
33	40,23728704	299,899605	56,13562799	568,1032424	385,3034226	2407,2	27	ENSG00000242574	2265,524717	3,819583423	0,219870422	12,82384142	1,20575E-37 8,25756E-35 HLA-DMB
34	13078,03109	14673,85241	12869,05451	14992,04765	11871,1049	16338,	28	ENSG00000234518	312,0912993	8,276859324	0,576720452	12,61765436	1,68774E-36 1,11303E-33
							29	ENSG00000163568	1455,22973	7,804253083	0,543387364	12,52191997	5,6648E-36 3,60241E-33 AIM2
							30	ENSG00000213626	4433,054879	-4,42792907	0,275004543	-12,46499068	1,15885E-35 7,11536E-33 LBH
							31	ENSG00000183734	893,1217877	4,237240545	0,260041233	12,4489509	1,41695E-35 8,4101E-33 ASCL2
							32	ENSG000002070547	2816,1070595	3,748796822	0,200633386	12,3533619	4,67175E-35 2,6834E-32 CD40
							33	ENSG00000140105	166892,85	5,82727503	0,391766026	12,32183169	6,91066E-35 3,84535E-32 WARS1
							34	ENSG00000117228	98501,16773	7,686166048	0,552174564	12,10870939	9,48831E-34 5,11966E-31 GBP1
							35	ENSG000002044731	730,7135805	5,027915974	0,334446879	12,04351489	2,09821E-33 1,09885E-30 C4A
							36	ENSG0000019582	53539,51495	4,152771883	0,264525014	11,91861512	9,46691E-33 4,81622E-30 CD74
							37	ENSG00000204252	147,6369201	5,782657331	0,402965448	11,86865362	1,72217E-32 8,51804E-30 HLA-DOA
							38	ENSG000002070547	221,0902311	5,370582109	0,370389876	-11,79995024	3,90542E-32 1,87946E-29 LINC01235
							39	ENSG00000089041	14232,26023	4,656176735	0,316525516	11,55096999	2,79919E-31 3,42025E-28 P2RX7
							40	ENSG00000179583	1521,7220208	6,112248838	0,449889179	11,36335141	6,36549E-30 2,90625E-27 CIITA
							41	ENSG00000121308	1206,206203	5,996911594	0,439943587	11,35807349	6,76193E-30 3,01008E-27 BCL2L14
							42	ENSG00000136436	4841,032355	2,828125249	0,161138537	11,34505304	7,84786E-30 3,40827E-27 CALCOCO2
							43	ENSG00000170989	704,6955516	-5,365857478	0,386740233	-11,28886293	1,48951E-29 6,31482E-27 51PR1
							44	ENSG00000168899	715,8817981	4,743931938	0,333167369	11,23739082	2,67162E-29 1,1063E-26 VAMP5
							45	ENSG00000153012	528,5976347	-6,370842542	0,483211518	-11,11488931	1,06182E-28 4,297E-26 LGI2
							46	ENSG00000140511	4173,841235	7,091736603	0,55079617	11,05987466	1,96372E-28 7,60129E-26 HAPLN3
							47	ENSG00000185338	4154,862612	7,689753422	0,60471596	11,06161974	1,92588E-28 7,60129E-26 SOCS1
							48	ENSG00000174749	1397,798733	2,925589988	0,174477271	11,03628579	2,553372E-28 9,67481E-26 FAM241A
							49	ENSG00000188820	8001,349732	5,681784849	0,426951103	10,96562302	5,59141E-28 2,07418E-25 CALHM6

Show me your data...

	A	B	C	D	E	F	G	H	I	J	K
1	SAMEA103885102	SAMEA103885347	SAMEA103885043	SAMEA103885392	SAMEA103885182	SAMEA103885136	SAMEA103885413	SAMEA103884967	SAMEA103885368	SAMEA103885218	SAMEA103885319
2	55,69716914	96,38981386	67,26753412	142,6835597	186,1485543	122,5926771	170,5541862	203,0472731	70,5820175	30,98336005	71,87276974
3	0	0	0	0	0	0	0	0	0	0	0
4	964,8925057	924,83214	1036,106869	1046,57092	1014,635976	1083,422465	969,604741	982,6266614	954,7538272	1035,036116	1099,503914
5	351,9054782	684,638914	387,1752238	820,970061	375,1488073	948,91					
6	677,4205719	221,0877036	276,2026831	97,82962948	395,4077973	175,59					
7	132,782876	198,7016694	662,811222	1522,435008	386,9046409	5036,5	506,29	982,6266614	954,7538272	1035,036116	1099,503914
8	362,6435931	5375,365498	650,3552061	7471,729136	116,9579367						
9	5185,999104	4426,618781	5870,757295	5878,363213	3769,061292	3829,0					
10	4064,476563	10213,97627	2388,342021	10237,20614	3863,969772	13282,					
11	996,2886541	1107,437138	1481,523638	1905,441941	1296,389901	1195,8					
12	78,82859697	78,75996066	104,2567382	130,0656527	75,32368946	90,653					
13	372,4875108	375,1582833	336,9448447	378,4351496	484,7559515	589,17					
14	1004,363486	804,9305608	2122,334216	566,7309848	894,7410645	609,49					
15	47,6088729	190,9341388	47,12634333	206,4662057	145,2740402	422,23					
16	10,2940265	10,73067656	18,42333204	34,72421059	11,87265036	24,248					
17	8,461862162	0,100956323	6,053534158	14,24249154	8,595001149						
18	1265,474946	1617,700447	1669,609838	2297,800891	1288,313487	1497,7					
19	2742,714674	1367,003771	5053,573294	2427,122021	1532,174172	1042,9					
20	739,029221	721,4533065	978,5249905	869,4506791	758,6712728	687,29					
21	153,0850266	223,6814932	90,69677231	99,67707173	165,2491282	181,67					
22	0	0	0	0	10,63374428	1,275177865	4,4071				
23	294,9241388	200,997003	193,0178533	259,2214914	272,6918871	292,3					
24	10311,29607	40953,9043	14546,93203	48197,23868	12049,09318	61320,					
25	14252,637178	8840,178525	18210,18028	12299,01927	11043,55437	8344,3					
26	2129,428152	471,0255994	287,0384441	86,35857737	2472,120332	559,39					
27	0	3,371313144	5,038960845	1,962785986	19,06817927	25,574					
28	0,4707608	0	0	0	0						
29	0,83189235	0,56534712	0,780066877	1,377008507	1,834964003	3,2439					
30	660,247462	697,0714736	376,575126	468,8681819	1120,360629	875,98					
31	15314,39018	12764,13479	5147,168455	7605,221356	17247,28895	19126					
32	2828,408307	2687,797465	2800,493773	6277,1227	2275,27131	2454,					
33	40,23728704	299,899605	56,13562779	568,1032424	385,3034226	2407,2					
34	13078,03109	14673,85241	12869,05451	14992,04765	11871,1049	16338,					

	A	B	C	D	E	F	G	H	I	J	K
1	id	baseMean	logFoldChange	IfcSE	stat	pvalue	padj				SYMBOL
2	ENSG00000125347	30585,07489	5.560123732	0.218266321	20,89247536	6,26847E-97	1,11616E-92	IRF1			
3	ENSG00000162645	36790,59205									
4	ENSG00000111181	688,5779196									
5	ENSG00000174944	1324,397476									
6	ENSG00000100336	2551,478907									
7	ENSG00000145365	3637,222141									
8	ENSG00000137497	7129,418865									
9	ENSG00000154541	1714,98786									
10	ENSG00000129244	682,8305656									
11	ENSG00000123189	5147,824664									
12	ENSG000000204257	1909,416235									
13	ENSG00000100342	4112,437557									
14	ENSG00000162654	52929,494435									
15	ENSG00000133321	1393,428038									
16	ENSG00000128284	1363,55118									
17	ENSG00000134470	4468,961376									
18	ENSG00000172399	714,5095698									
19	ENSG00000164509	700,3132408									
20	ENSG00000025492	1295,305924									
21	ENSG000000204267	10442,43781									
22	ENSG000000254838	135,9997527									
23	ENSG00000131203	92938,1422									
24	ENSG000000213886	2086,328471									
25	ENSG00000100911	11111,0667									
26	ENSG00000168394	38700,36687									
27	ENSG000000242574	2265,524717									
28	ENSG000000234518	312,0912993									
29	ENSG00000163568	1455,22973									
30	ENSG000000213626	4433,054879									
31	ENSG00000183734	893,1217877									
32	ENSG00000101017	28016,70595									
33	ENSG00000140105	166892,85									
34	ENSG00000117228	98501,16773									
35	ENSG000000244731	730,7135805									
36	ENSG00000019582	53359,71495									
37	ENSG000000204252	147,6369201									
38	ENSG000000270547	221,0902311									
39	ENSG000000089041	14232,26023									
40	ENSG00000179583	1521,720208									
41	ENSG00000121380	1206,206203									
42	ENSG00000136436	4841,032355									
43	ENSG00000170989	704,6955516									
44	ENSG000000168899	715,8817981									
45	ENSG000000153012	528,5976347									
46	ENSG000000140511	4173,841235									
47	ENSG000000185338	4154,862612									
48	ENSG000000174749	1397,798733									
49	ENSG000000188820	8001,349732									

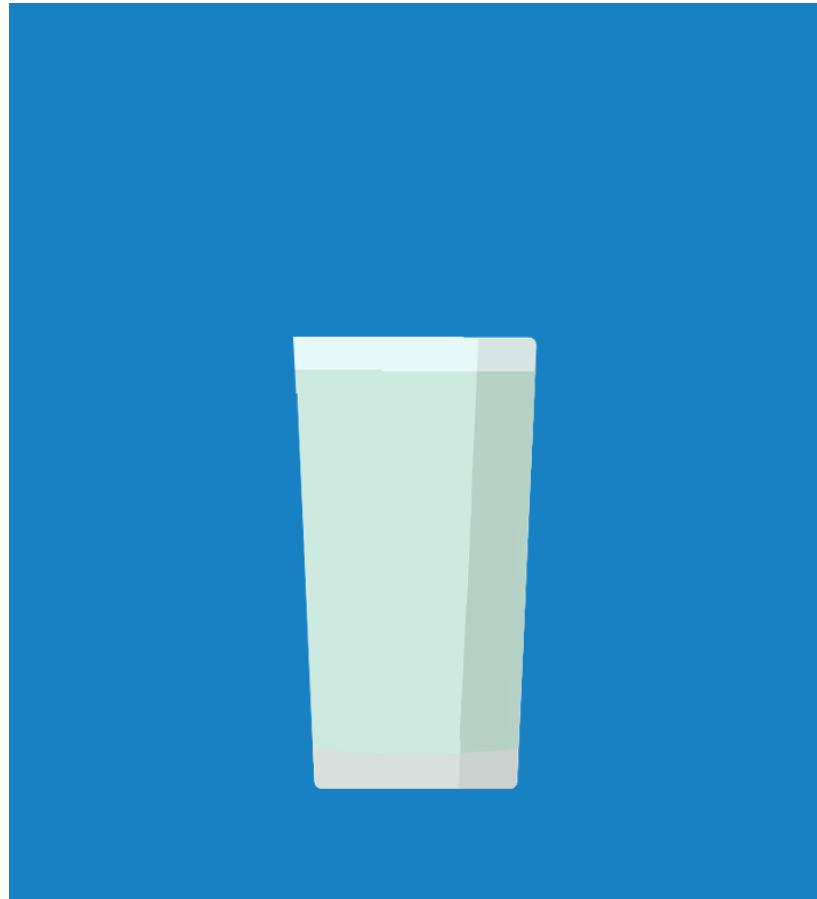
Show me your data...

A	B	C	D	E	F	G	H	I	J	K
1 SAMEA103885102	SAMEA103885347	SAMEA103885043	SAMEA103885392	SAMEA103885182	SAMEA103885136	SAMEA103885413	SAMEA103884967	SAMEA103885368	SAMEA103885218	SAMEA103885319
2 55,69716914	96,38981386	67,26753412	142,6835597	186,1485543	122,5926771	170,5541862	203,0472731	70,5820175	30,98336005	71,87276974
3 0	0	0	0	0	0	0	0	0	0	0
4 964,8925057	924,83214	1036,106869	1046,57092	1014,635976	1083,422465	969,604741	982,6266614	954,7538272	1035,036116	1099,503914
5 351,9054782	684,638914	387,1752238	820,9700671	375,1488073	948,91					
6 677,4205719	221,0877036	276,2026831	97,82962948	395,4077973	175,59					
7 132,782876	198,7016964	662,811222	1522,435008	386,9046409	560,29					
8 362,6435931	5375,365498	650,3552061	7471,7292136	116,9579367	5036,5					
9 5185,999104	4426,618781	5870,757295	5878,363213	3769,061292	3829,0					
10 4064,476563	10213,97627	2388,342021	10237,20614	3863,969772	13282,					
11 996,2886541	1107,4371318	1481,523638	1905,441941	1296,389901	1195,8					
12 78,82859697	78,75996066	104,2567382	130,0656527	75,32368946	90,653					
13 372,4875108	375,1582833	336,9448447	378,4351496	484,7559515	589,17					
14 1004,3634846	804,9305068	2122,332416	5667,7309849	894,7410645	609,49					
15 47,6088729	190,9341388	47,12634333	206,4662057	145,2740402	422,23					
16 10,2940199	10,73067656	18,42333204	34,72412059	11,87265036	24,248					
17 8,461862162	0,100956323	6,053534158	14,24249154	8,595001149						
18 1265,474946	1617,700447	1669,609838	2297,800891	1288,313487	1497,7					
19 2742,714674	1367,003771	5053,573294	2427,122201	1532,174172	1042,9					
20 739,0239221	721,4533065	978,5249905	869,4506791	758,						
21 153,0850266	223,6814932	90,69677231	99,67707173	165,						
22 0	0	0	0	10,63374428	1,27					
23 294,9241388	200,997003	193,0178533	259,2214914	211,						
24 10311,29607	40953,9043	14546,93203	48197,23868	1204						
25 14252,63718	8840,178525	18210,18028	12299,01927	1104						
26 2129,428152	471,0255994	287,0384441	86,35857737	2472						
27 0	3,371313144	5,038960845	1,962785986	19,0						
28 0,4707608	0	0	0	10,63374428	1,27					
29 0,83189235	0,56534712	0,780066877	1,377008507	1,83						
30 660,247462	697,0714736	376,575126	468,8681819	1120						
31 15314,39018	12764,13479	5147,168455	7605,221356	1724						
32 2828,408307	2687,797465	2800,493773	6277,1227	227						
33 40,23728704	299,899605	56,13562779	568,1032424	385,						
34 13078,03109	14673,85241	12869,05451	14992,04765	118						
35 1	A	B	C	D	E	F	G	H	I	K
36 2	1	gs_id	gs_description	gs_pvalue	2	gs_genes	gs_de_count	gs_bg_count	Expected	DE_count
37 3	2	GO:0022250	adaptive immune response	9,2E-23	2	B2M,BCL6,BTN3A1,BTN3A2,BTN3A3,C1Q8,C1R,C1RL,C1S,C2,C3,C4A,C4B	89	323	17,93	89
38 4	3	GO:0060333	interferon-gamma-mediated signaling pathway	1,2E-20	2	B2M,CAMK2D,CITA,GBP1,GBP2,AHL-A,HLA-B,HLA-C,HLA-DPA1,HLA-DPB1	34	82	4,55	34
39 5	4	GO:0060337	type I interferon signaling pathway	4,8E-13	2	GBP2,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,IFI27,IFI35,IFI7,IFI73,IFI76	25	80	4,44	25
40 6	5	GO:0019885	antigen processing and presentation of endogenous peptide	4,2E-12	2	B2M,ERAP2,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,TAP1,TAP2,TAPBP	11	14	0,78	11
41 7	6	GO:0051607	defense response to virus	2E-10	2	AIM2,APOBEC3A,APOBEC3D,APOBEC3G,CD40,CXCL10,CXCL9,DXD60,DT40	41	206	11,44	41
42 8	7	GO:0001916	positive regulation of T cell mediated cytotoxicity	5,6E-10	2	B2M,CD1A,CD1C,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,IL12A,IL12B1,F	12	23	1,28	12
43 9	8	GO:0002475	antigen processing and presentation via MHC class Ib	5,8E-10	2	B2M,CD1A,CD1C,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,TAP2	10	15	0,83	10
44 10	9	GO:0062270	DNA replication initiation	1,9E-08	2	CNE2,CDC45,CD6,CDC6,CDT1,MCM10,MCM2,MCM4,MCM5,MCM6,MCM7	13	35	1,94	13
45 11	10	GO:0034341	response to interferon-gamma	2,2E-08	2	ACOD1,B2M,CALCO2,CAMK2D,CCL13,CCL15,CCL18,CCL5,CCL7,CCL8,C	58	174	9,66	58
46 12	11	GO:0069545	inflammatory response	2,9E-08	2	ACOD1,ADAM8,ADGRS5,ADORA3,AGT,AIM2,AOC3,APOL2,APOL3,APPBL2	87	589	32,7	87
47 13	12	GO:0050852	T cell receptor signaling pathway	9,4E-08	2	BTN2A2,BTN3A1,BTN3A2,BTN3A3,CD28,CEACAM1,CTLA4,CYLD,FYBL,GBP	28	165	9,16	28
48 14	13	GO:0045087	innate immune response	1,1E-07	2	ACOD1,ADAM8,AIM2,APOBEC3A,APOBEC3D,APOBEC3G,APOL1,APOL2,B2	136	710	39,42	136
49 15	14	GO:0069585	complement activation, classical pathway	2,5E-07	2	C1Q8,C1R,C1RL,C2,C3,C4A,C4B,SERPING1	9	19	1,05	9
50 16	15	GO:0007272	double-strand break repair via break-induced replication	4,1E-07	2	CDC45,GINS2,MCM2,MCM4,MCM5,MCM6,MCM7	7	11	0,61	7
51 17	16	GO:0042832	defense response to protozoan	4,4E-07	2	BATF2,CD40,GBP2,GBP4,GBP6,GBP7,IL12A,SLC11A1,TPSN32	9	20	1,11	9
52 18	17	GO:0071222	cellular response to lipopolysaccharide	7,1E-07	2	ACOD1,CARD16,CARD17,CCL24,CD274,CD80,CMPK2,CX3CR1,CXCL10,CXC	27	171	9,49	27
53 19	18	GO:0030449	regulation of complement activation	8,5E-07	2	C1Q8,C1R,C1S,C2,C3,C4A,C4B,CFB,CFH,CR1L,SERPING1	11	33	1,83	11
54 20	19	GO:0070998	chemokine-mediated signaling pathway	0,0000012	2	CCL13,CCL15,CCL18,CCL5,CCL7,CCL8,CCL9,CCL10,CXCL11,CX	15	63	3,5	15
55 21	20	GO:0002479	antigen processing and presentation of exogenous peptide	0,0000014	2	B2M,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,NCF1,PSMB10,PSMB8,PSMB	16	72	4	16
56 22	21	GO:0019886	antigen processing and presentation of endogenous peptide	0,0000016	2	CD74,CTSS,HLA-DMA,HLA-DMB,HLA-DOA,HLA-DOB,HLA-DPA1,HLA-DPB1,I	18	90	5	18
57 23	22	GO:0305933	neutrophil chemotaxis	0,0000019	2	CCL13,CCL15,CCL18,CCL5,CCL7,CCL8,CDD00N,CD74,CXADR,CXCL10,CX	17	82	4,55	17
58 24	23	GO:0071356	cellular response to tumor necrosis factor	0,0000041	2	ACOD1,AIM2,BRCA1,CARD16,CASP4,CCL13,CCL18,CCL19,CCL20,CCL7,CCL	36	246	13,66	36
59 25	24	GO:0042102	positive regulation of T cell proliferation	0,0000044	2	CDC5,CD74,CD28,CD80,HLA-DMB,HLA-DPA1,HLA-E,IL12A,IL1	16	78	4,33	16
60 26	25	GO:0045071	negative regulation of viral genome replication	0,0000067	2	APOBEC3A,APOBEC3D,APOBEC3G,CCL5,CCL8,IFIT5,IFIT7,IFITM1,IFITM2,IFITM3,	13	55	3,05	13
61 27	26	GO:0032729	positive regulation of interferon-gamma production	0,0000067	2	HLA-DPA1,HLA-DPB2,IL12A,IL12B1,IL12RB2,IL27,KLRK1,PE4B,RIPK2,S	13	55	3,05	13
62 28	27	GO:0007186	G protein-coupled receptor signaling pathway	0,0000076	2	ADGRES,ADGRS5,ADORA3,AGT,C3,CCL13,CCL15,CCL18,CCL5,CCL7,CCL8,C	59	594	32,98	59
63 29	28	GO:0071346	cellular response to interferon-gamma	0,0000085	2	ACOD1,B2M,CAMK2D,CCL13,CCL15,CCL18,CCL5,CCL7,CCL8,CITA,EDN1,I	49	158	8,77	49
64 30	29	GO:0050830	defense response to Gram-positive bacterium	0,00001	2	GBP2,GBP4,GBP6,GBP7,GSMDM,H2B2,IFH4B,IFH4B1,KLRK1,P2RX7,RIPK	12	49	2,72	12
65 31	30	GO:0000083	regulation of transcription involved in G1/S transition of mi	0,000012	2	BACH1,CDC45,CDC6,CDT1,DHFR,E2F1,ORC1,RRM2,TYMS	9	28	1,55	9
66 32	31	GO:0050729	positive regulation of inflammatory response	0,000019	2	ADAM8,AGT,C3,CCL18,CD28,GBP5,HLA-E,IDO1,IL15,JAK2,KARS1,LRRK2,N	22	119	6,61	22
67 33	32	GO:0042531	positive regulation of tyrosine phosphorylation of STAT prot	0,00003	2	CCL5,CD40,IL12A,IL15,IL13RA1,JAK2,LY6,OSM,PARP14,STAP2,TNFSF18	11	46	2,55	11
68 34	33	GO:0035455	response to interferon-alpha	0,000062	2	IFIT2,IFIT3,IFITM1,IFITM2,IFITM3,LAMP3,PYHIN1	7	20	1,11	7
69 35	34	GO:0050718	positive regulation of interleukin-1 beta secretion	0,000072	2	AIM2,GSDMD,LIRRA2,NLRP12,NOD2,P2RX7,RIPK2,TLR8	8	27	1,5	8
70 36	35	GO:0032693	negative regulation of interleukin-10 production	0,00009	2	CD274,DL1,IL10,IL11,JAK3,PDCD11LG2,TNFSF21	6	15	0,83	6
71 37	36	GO:0000076	Arrest at G1/S transition checkpoint	0,000099	2	CDC45,CDC6,CDT1,CDP5,ORC1,TICRR	6	15	0,83	6
72 38	37	GO:2000425	regulation of apoptosis by caspase cascade	0,0001	2	C3,C4A,CD3,CD4,CD30,LF	5	10	0,56	5
73 39	38	GO:0071347	cellular response to tumor necrosis factor	0,0005	2	C1D1,CC13,C15,C18,CD28,CD29,CD30,CD31,CD32,CD33,CD34,CD35,CD36,CD37,CD38,CD39,CD3	21	157	8,72	21

	A	B	C	D	E	F	G	H	I	J	K			
1	SAMEA103885102	SAMEA103885347	SAMEA103885043	SAMEA103885392	SAMEA103885182	SAMEA103885136	SAMEA103885413	SAMEA103884967	SAMEA103885368	SAMEA103885218	SAMEA103885319			
2	55,69716914	96,38981386	67,26753412	142,6835597	186,1485543	122,5926771	170,5541862	203,0472731	70,5820175	30,98336005	71,87276974			
3	0	0	0	0	0	0	0	0	0	0	0			
4	964,8925057	924,83214	1036,106869	1046,57092	1014,635976	1083,422465	969,604741	982,6266614	954,7538272	1035,036116	1099,503914			
5	351,9054782	684,638914	387,1752238	820,970061	375,1488073	948,91	A	B	C	D	E			
6	677,4205719	221,0877036	276,2026831	97,82962948	395,4077973	175,59	id	baseMean	logFoldChange	IfcSE	stat			
7	132,782876	198,7016694	662,811222	1522,435008	386,9046409	5036,5	ENSG00000125347	30585,07489	5.560123732	0.218266321	20,89247536	pvalue		
8	362,6435931	5375,365498	650,3552061	7471,729136	116,9579367	3829,0	ENSG00000162645	36790,59205	1.11616E-92	IRF1	padj			
9	5185,999104	4426,618781	5870,757295	5878,363213	3769,061292	3829,0	ENSG00000111181	688,5779196			SYMBOL			
10	4064,476563	10213,97627	2388,342021	10237,20614	3863,969772	13282,	Category	Term	Count	%	PValue			
11	996,2886541	1107,437138	1481,523638	1905,441941	1296,389901	1195,8	ENSG0000010174944	1324,397476			Genes			
12	78,75996066	104,2567382	130,0656527	75,32368946	90,653	6	ENSG00000100336	2551,478907			List			
13	372,4875108	375,1582833	336,9448447	378,4351496	484,7559515	589,17	ENSG00000145365	3637,222141			Total			
14	1004,363486	804,9305608	2122,334216	566,7309848	894,7410645	609,49	ENSG00000137496	7129,418865			Pop			
15	47,6088729	190,9341388	47,12634333	206,4662057	145,2740402	422,23	ENSG00000129244	682,8305656			Hits			
16	10,2940199	10,73067656	18,42333204	34,724212059	11,87265036	24,248	ENSG000001203189	5147,8246644			Pop Total			
17	8,461862162	0,100956323	6,053534158	14,24249154	8,595001149	12	ENSG000000204257	1909,416235			Fold Enrichment			
18	1265,474946	1617,700447	1669,609838	2297,800891	1288,313487	1497,7	ENSG000000100342	4112,437557			Bonferroni			
19	2742,714674	1367,003771	5053,573294	2427,122201	1532,174172	1042,9	ENSG00000162654	52292,49435						
20	739,0239221	721,4533065	978,5249905	869,4506791	758,									
21	153,0850266	223,6814932	90,69677231	99,67707173	165,									
22	0	0	0	0	10,63374428	1,27								
23	294,9241388	200,997003	193,0178533	259,2214914	272,	1	gs_id	gs_description	gs_pvalue	gs_genes	s_de_count			
24	10311,29607	40953,9043	14546,93203	48197,23868	1204	2	GO:0002250	adaptive immune response	9.2E-23	B2M,BCL6,BTN3A1,BTN3A2,BTN3A3,C1QB,C1R,C1L,C1S,C2,C3,C4A,C4B	s_bg_count			
25	14252,637178	8840,178525	18210,18028	12299,01927	1104	3	GO:0060333	interferon-gamma-mediated signaling pathway	1.2E-20	B2M,CAMK2D,CITA,GBP1,GBP2,HLA-A,HLA-C,HLA-DPA1,HLA-IRF1	Expected			
26	2129,428152	471,0255994	287,0384441	86,35857737	2472	4	GO:0060337	type I interferon signaling pathway	4.8E-13	GPBP2,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,IFI27,IFI35,IFI72,IFI73,IFI74,IFI75,IFI76,IFI77,IFI78,IFI79,IFI7	DE_count			
27	0	3,37131344	5,038960845	1,962785986	19,0	5	GO:0019885	antigen processing and presentation of endogenous peptide	4.2E-12	B2M,ERAP2,HLA-A,HLA-B,HLA-C,HLA-E,HLA-F,HLA-G,TAP1,TAP2,TAPP2	11	14	0.78	11
28	0,4707608	0	0	0	0	6	GO:0051607	defense response to virus	2E-10	AIM2,APOBECE3A,APOBECE3D,CD40,CXCL10,CXCL9,DXD60,DTX	41	206	11,44	41
29	0,83189235													
30	660,247462													
31	15314,39018													
32	2828,408307													
33	40,23728704													
34	13078,03109													

Show me your data...

# A cocktail recipe



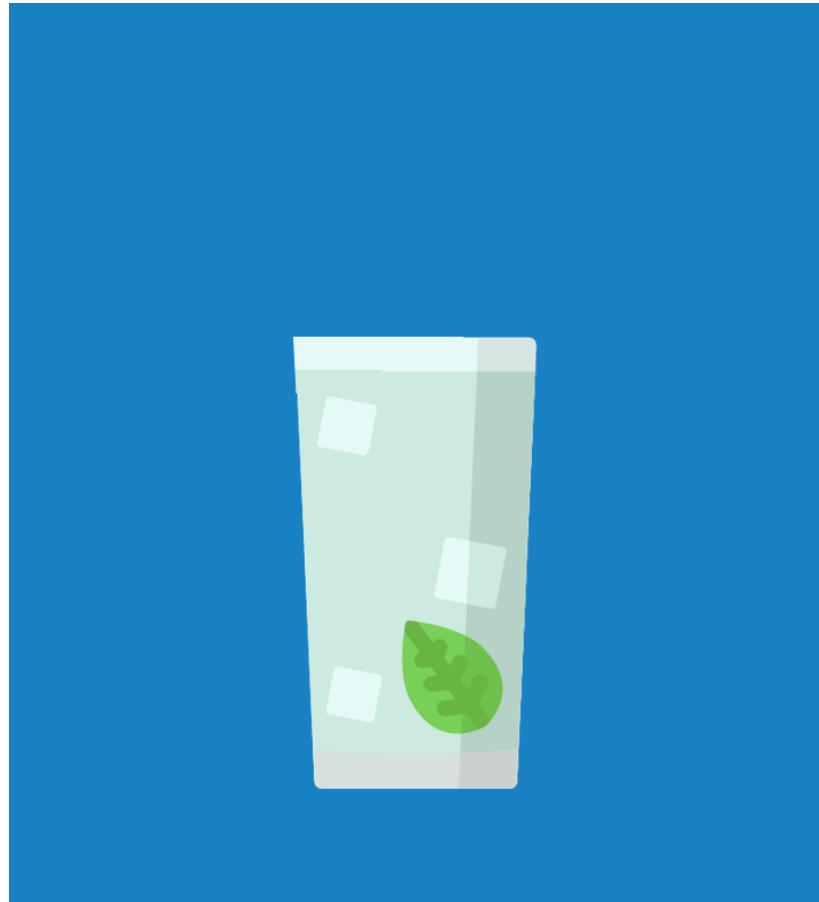
*You might (and should) have a set of standardized objects for your datasets and results (Excel does **not** count)*

# A cocktail recipe



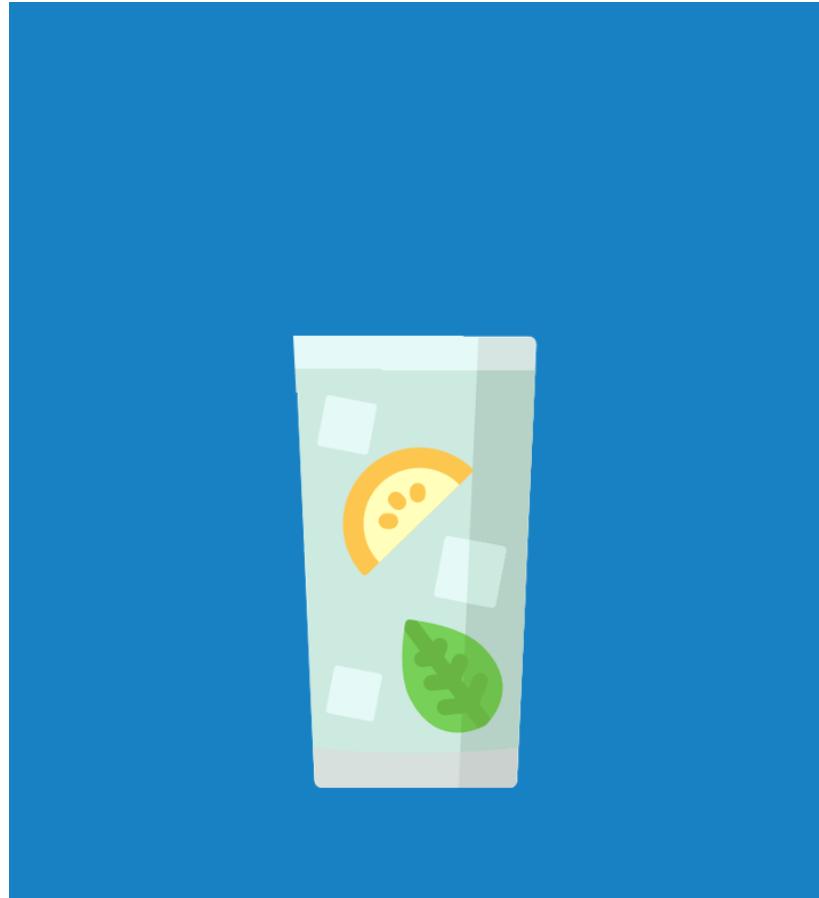
*A container for the expression matrix, `dds`*

# A cocktail recipe



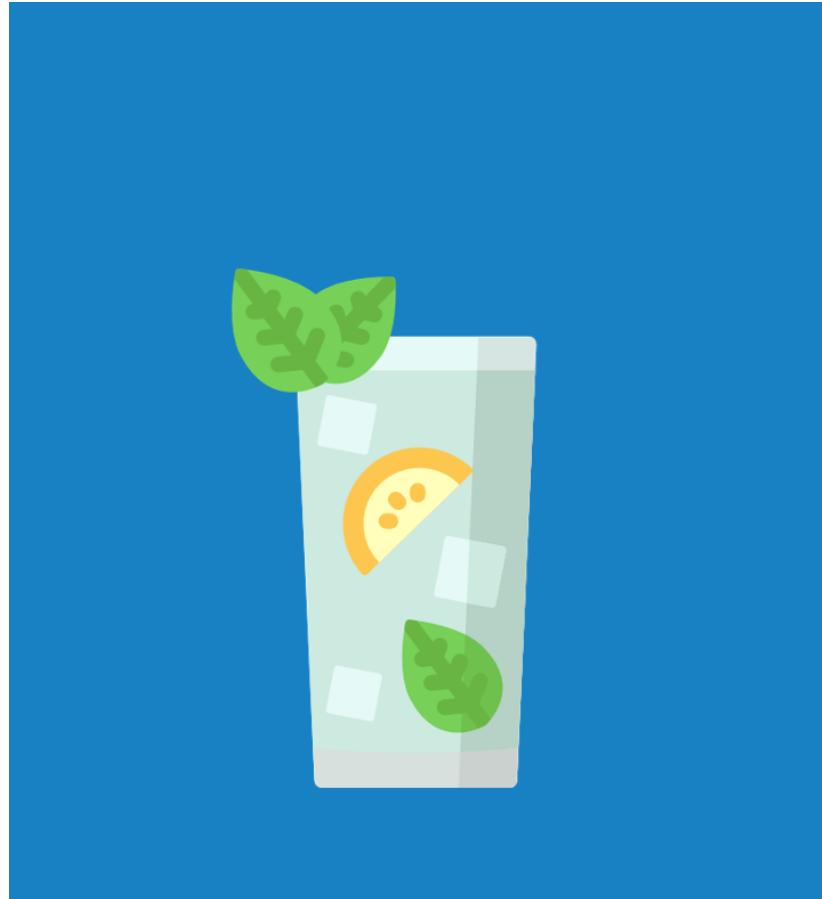
*A container for the results of differential expression, `res_de`*

# A cocktail recipe



*A table for the result of functional enrichment, `res_enrich`*

# A cocktail recipe



*Decorate with an annotation table, [anno\\_df](#)*

# A cocktail recipe



*Shaken, not stirred...*

# A cocktail recipe



GeneTonic is now on Bioconductor!

http://127.0.0.1:5184 | Open in Browser | [C](#)

~/Development/GeneTonic - Shiny

 GeneTonic [Bookmark](#)   

# GeneTonic

## First steps with a full glass

Overview on the provided input

Expression Matrix + DE results +

Functional analysis results + Annotation info +

58294 genes x 24 samples  
dds object 

928 DE genes  
res object 

500 functional categories  
func enrich object 

2 feature identifiers for 58294 features  
annotation object 

 GeneTonic is a project developed by [Federico Marini](#) in the Bioinformatics division of the [IMBEI](#) - Institute for Medical Biostatistics, Epidemiology and Informatics  
License: [MIT](#) - The GeneTonic package is developed and available on [GitHub](#)

Welcome!

Gene-Geneset

Enrichment Map

DEview

GeneSets

Bookmarks

About

## Overview on the provided input

Expression Matrix

+

DE results

+

Functional analysis results

+

Annotation info

+

58294 genes x 24 samples



dds object

928 DE genes



res object

500 functional categories



func enrich object

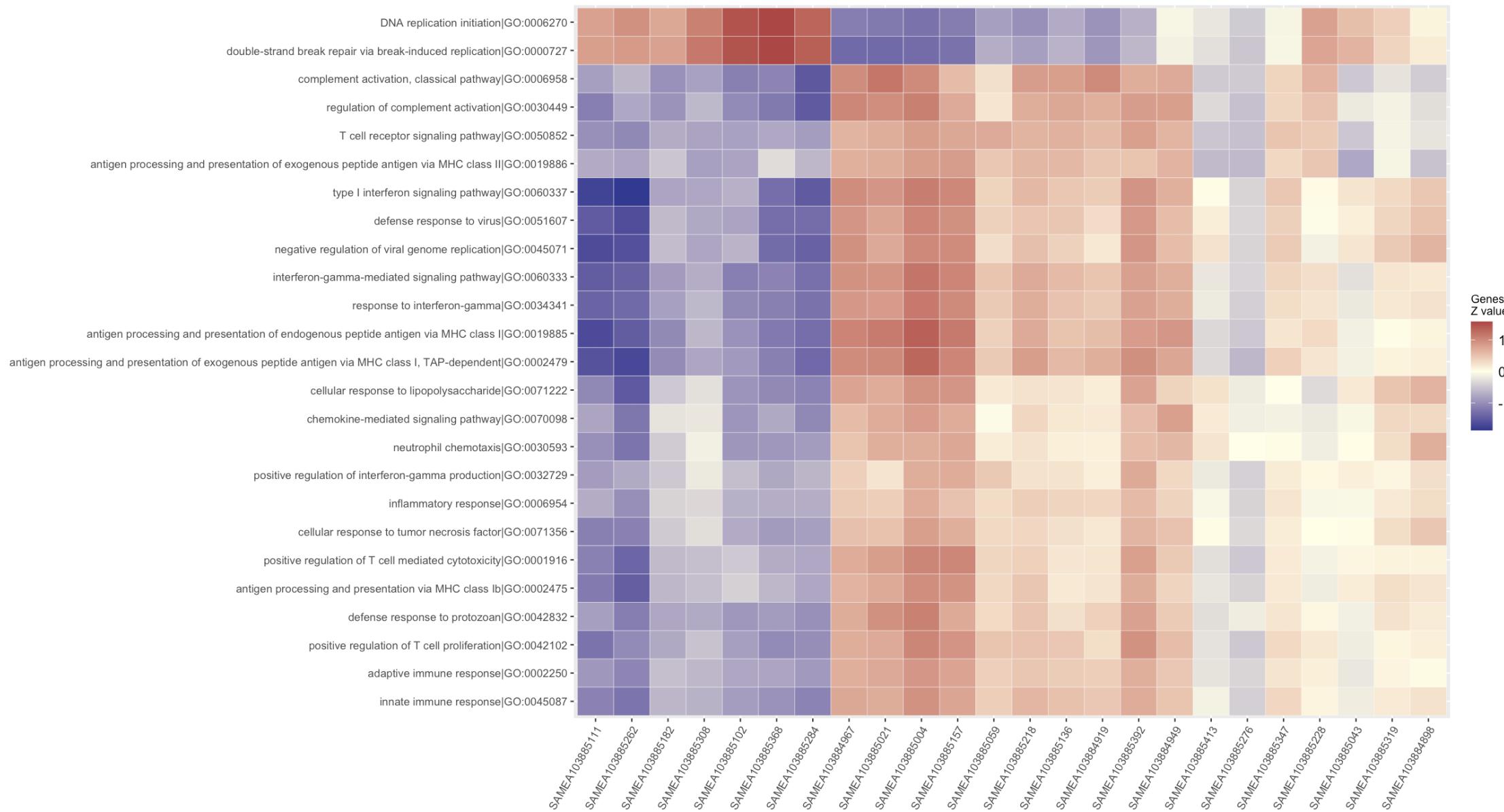
2 feature identifiers for 58294 features

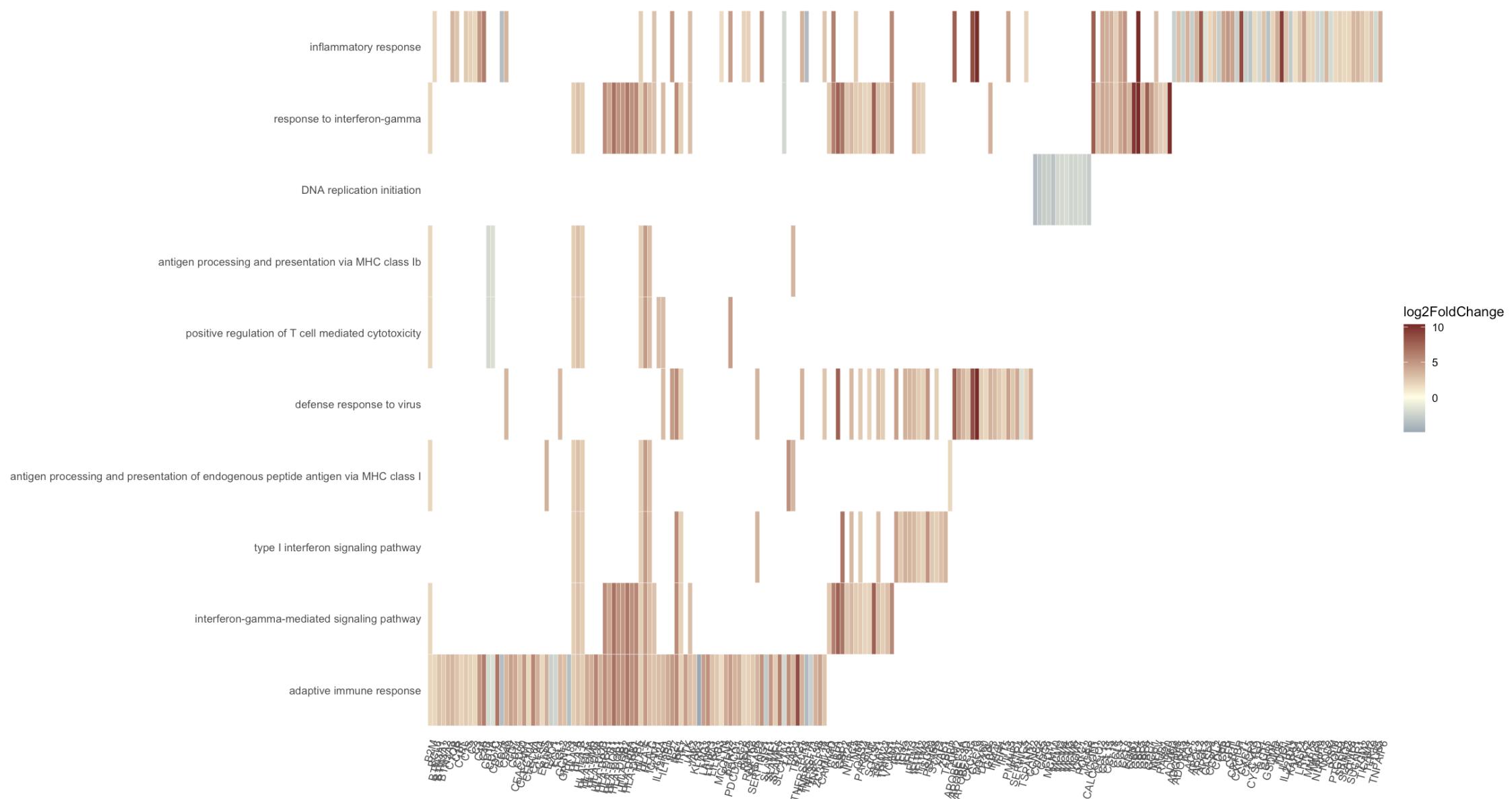


annotation object



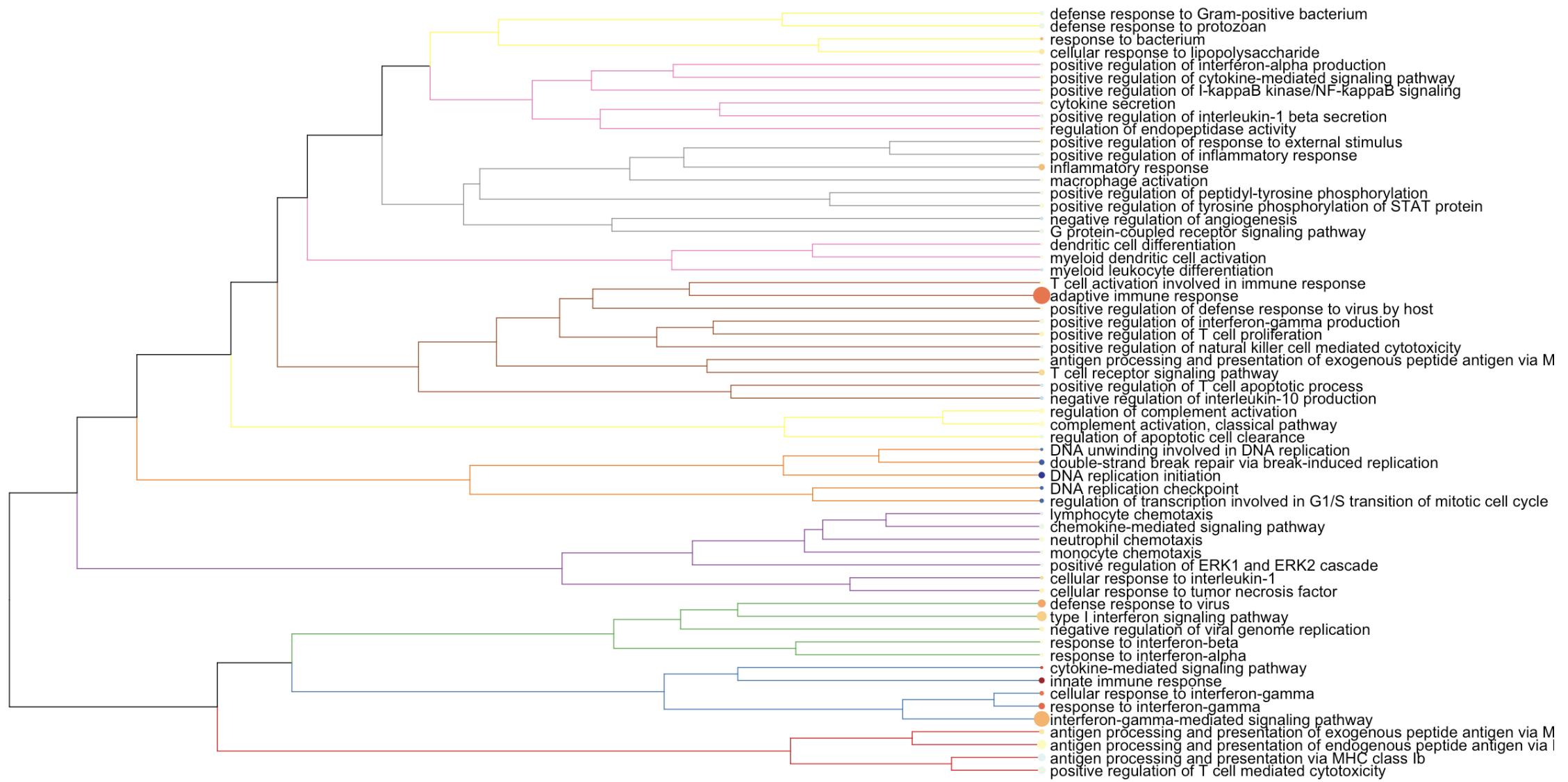




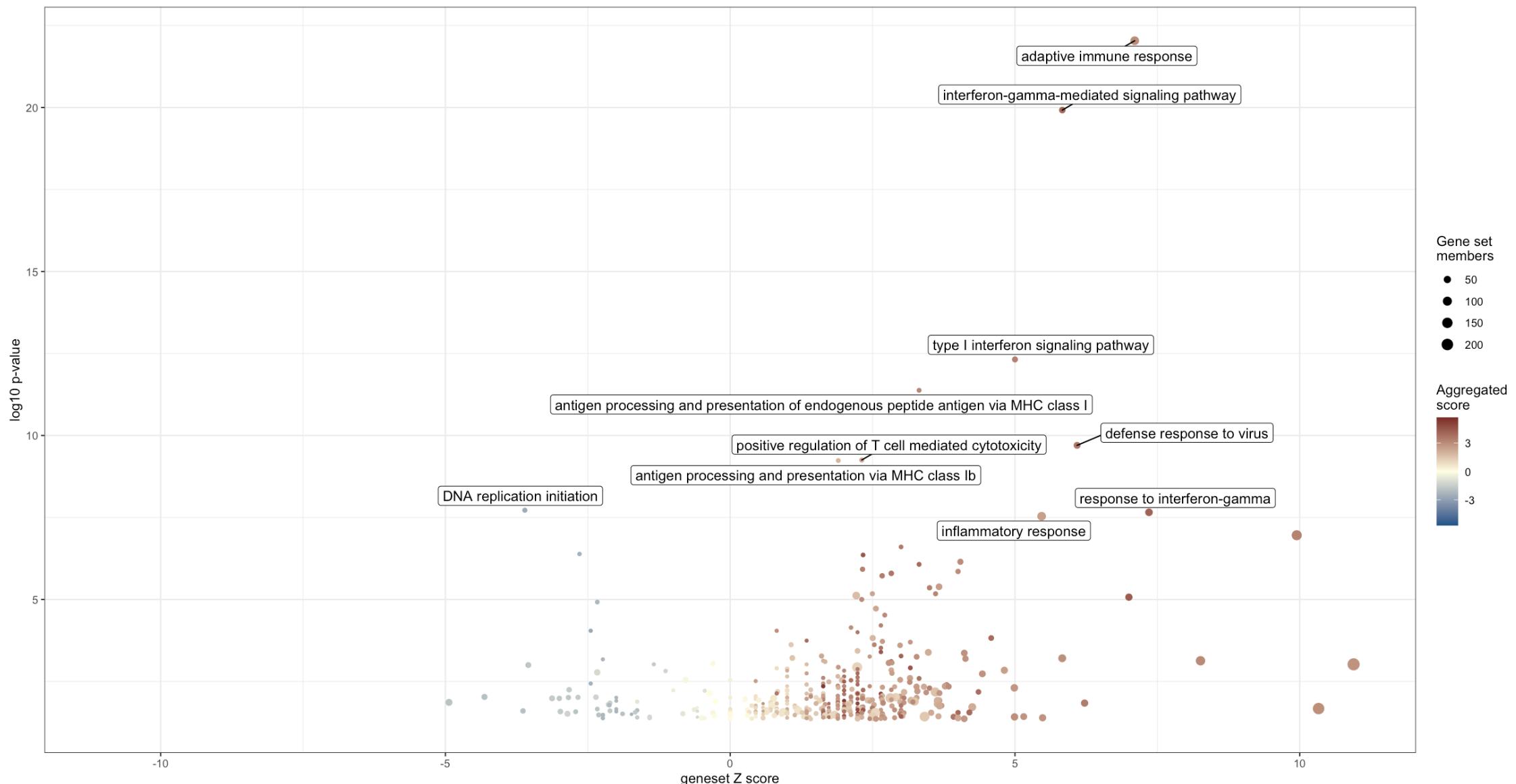


Geneset MDS plot - condition IFNg vs naive

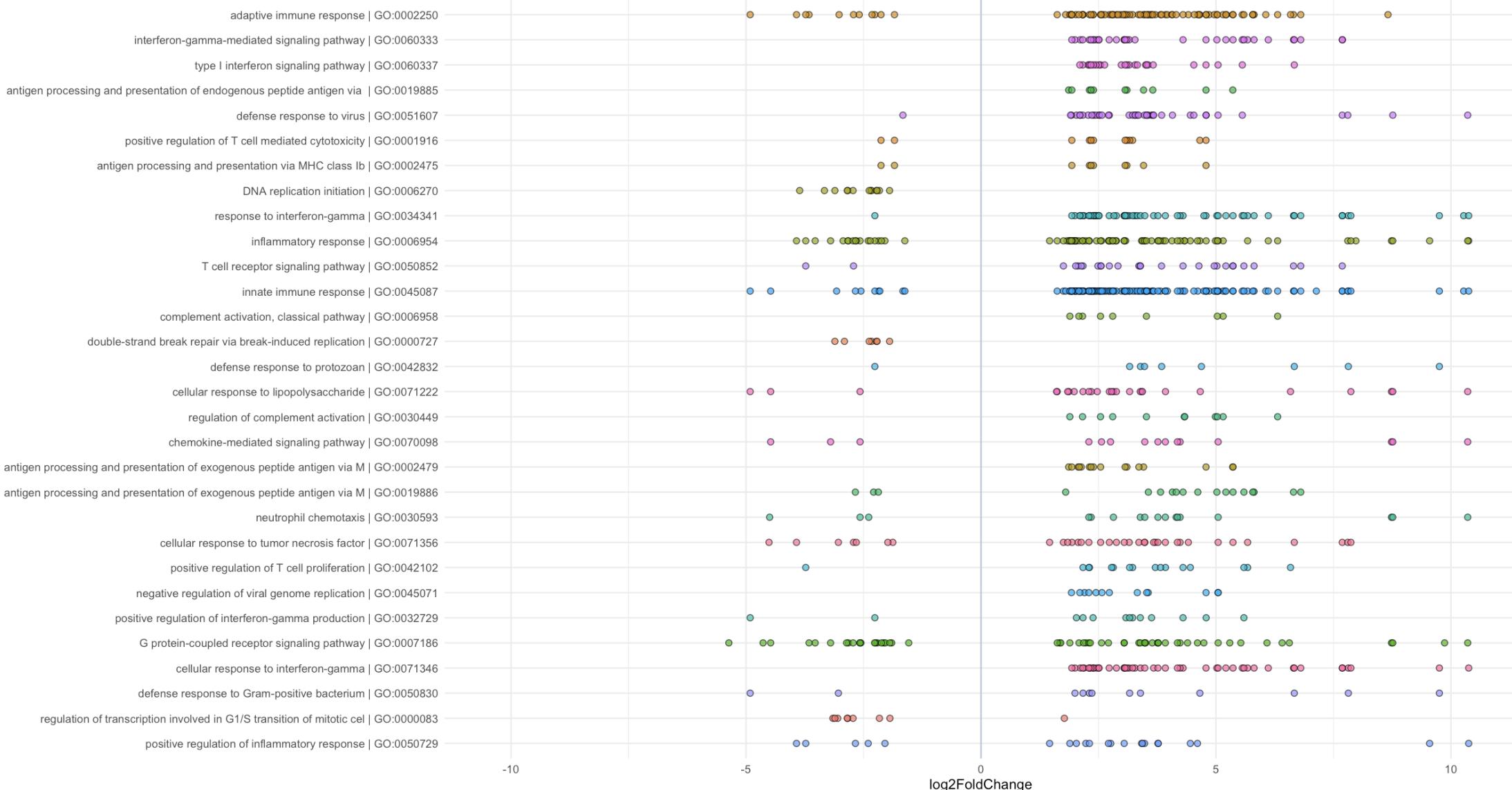




## Geneset volcano



## Enrichment overview - condition IFNg vs naive



Welcome!

Gene-Geneset

Enrichment Map

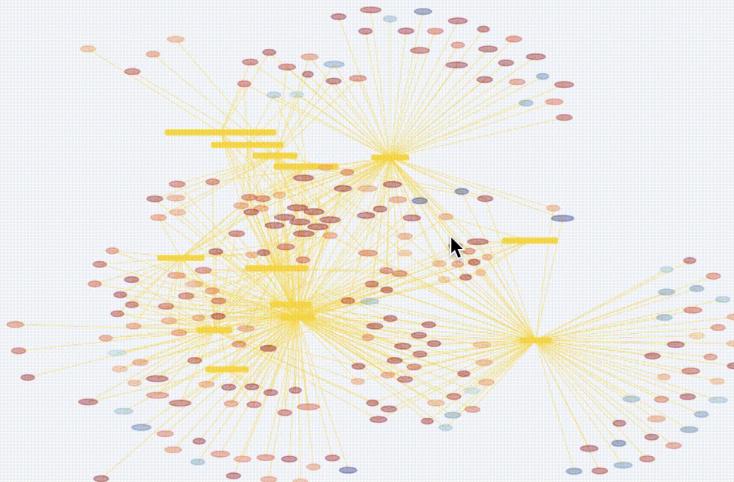
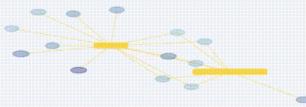
DView

GeneSets

Bookmarks

About

Select by id



Genesetbox

Please select a gene set.



Genebox

Please select a gene/feature.



# The interplay of genes and genesets

# A quick demo

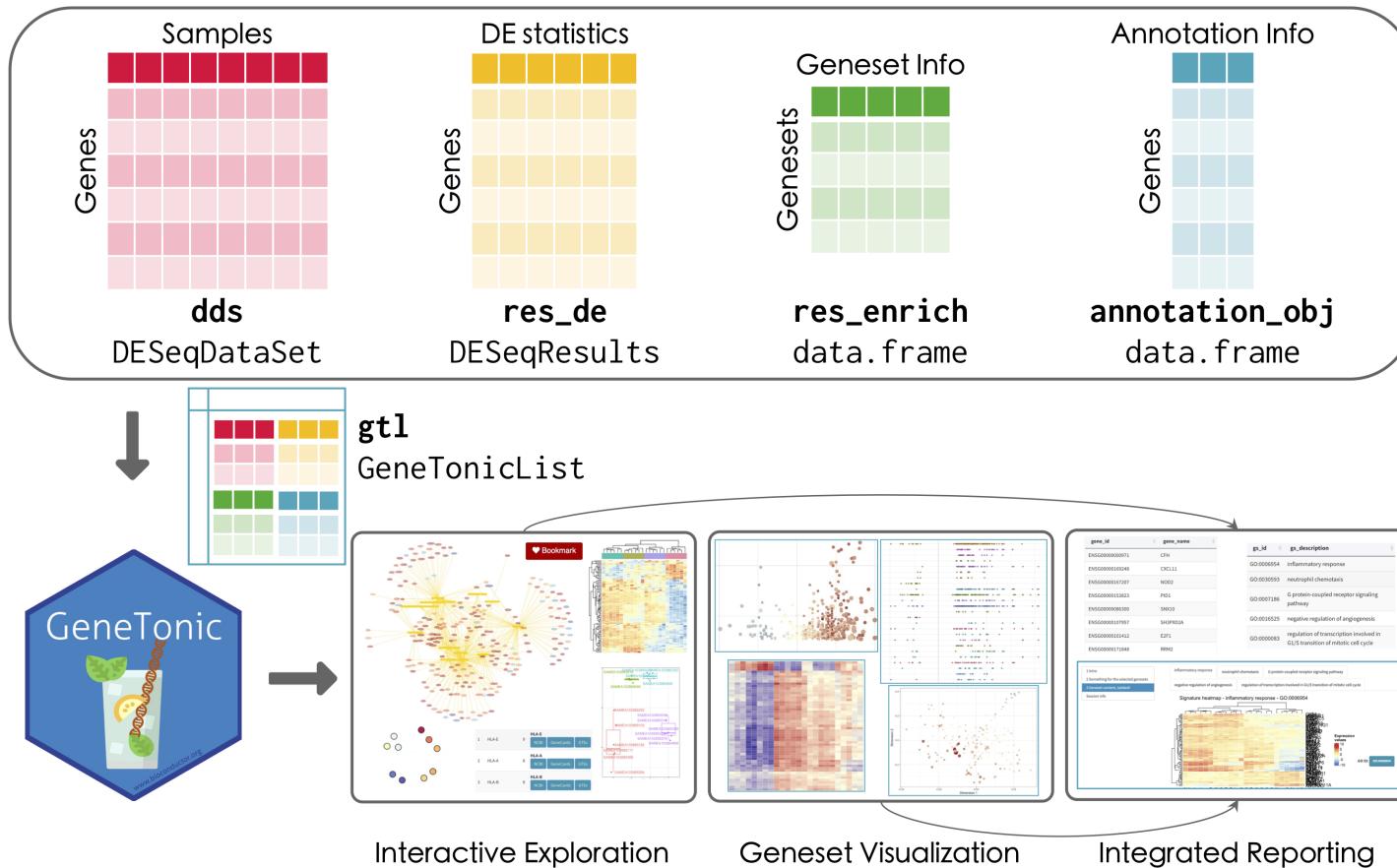
```
BiocManager::install("GeneTonic")
library("GeneTonic")
example(GeneTonic, ask = FALSE)
```



# GeneTonic



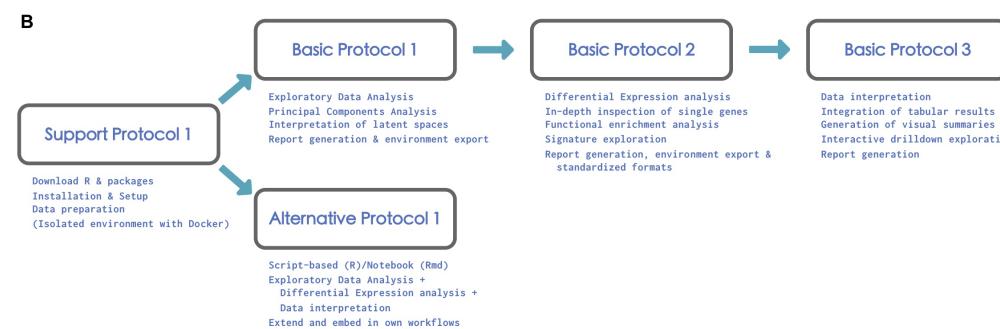
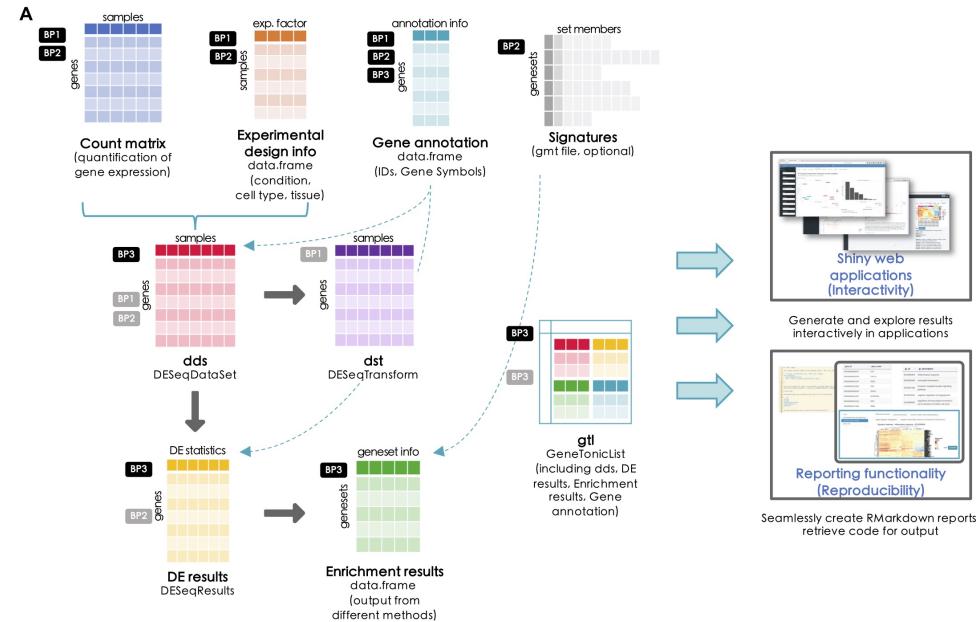
A way to enjoy the data interpretation...



# Suddenly i SEE



# Putting it all together



# Interfaces for RNA-seq

## Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective FREE

Alicia Poplawski, Federico Marini, Moritz Hess, Tanja Zeller, Johanna Mazur,  
Harald Binder

*Briefings in Bioinformatics*, Volume 17, Issue 2, March 2016, Pages 213–223,  
<https://doi.org/10.1093/bib/bbv036>

Published: 23 June 2015 Article history ▾

Interfaces have been developed for making such analysis steps accessible to life scientists without extensive knowledge of command line tools.

Systematic search and evaluation of such interfaces

Definition of criteria for evaluation (ease of configuration, documentation, usability, computational demand, reporting)

There's no **one tool fits all** winner!



Does the analysis/exploration/interpretation of your data spark joy?

# Lessons learnt for me (so far)

**It takes two to tango**

Results & Interpretation

Bioinformatician & Experimental scientist

Reproducibility & Interactivity

*Stop using Excel for performing Bioinformatics data analysis :)*

**Our contributions to the Bioconductor project:**

`pcaExplorer` - Exploratory Data Analysis

`ideal` - Differential Expression analysis

`GeneTonic` - Interpretation of DE results

`iSEE` - Exploration of single cell datasets

...

`countsimQC` - comparing count data sets

`iCOBRA` - calculation and visualization of performance metrics

`ExploreModelMatrix`

`tximeta` - tx importing with metadata, *for free*

`alevinQC`

# The essential - Recap

- Expression quantification
- Data exploration: principal components analysis, gene plots
- Differential analysis: DE modeling, design, effect size, variability, significance
- Functional interpretation: gene sets, pathways, biological themes

# Practical session

Check again the README file at [https://github.com/imbeimainz/MSE\\_GenEpi\\_2023](https://github.com/imbeimainz/MSE_GenEpi_2023)

## References:

- Marini and Binder (2016) - Development of Applications for Interactive and Reproducible Research: a Case Study (Genomics Computational Biology) [10.18547/gcb.2017.vol3.iss1.e39](https://doi.org/10.18547/gcb.2017.vol3.iss1.e39)
- Marini and Binder (2019) - pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components (BMC Bioinformatics) [10.1186/s12859-019-2879-1](https://doi.org/10.1186/s12859-019-2879-1)
- Marini, Linke, Binder (2020) - ideal: an R/Bioconductor package for Interactive Differential Expression Analysis (BMC Bioinformatics) [10.1186/s12859-020-03819-5](https://doi.org/10.1186/s12859-020-03819-5)
- Rue-Albrecht, Marini, Soneson, Lun (2018) - iSEE: Interactive SummarizedExperiment Explorer (F1000 Research) <https://doi.org/10.12688/f1000research.14966.1>
- Marini, Ludt, Linke, Strauch (2021) - GeneTonic: an R/Bioconductor package for streamlining the interpretation of RNA-seq data (BMC Bioinformatics) <https://doi.org/10.1186/s12859-021-04461-5>
- Ludt, Ustjanzew, Binder, Strauch, Marini (2022) - Interactive and Reproducible Workflows for Exploring and Modeling RNA-seq Data with pcaExplorer, Ideal, and GeneTonic (Current Protocols) <https://doi.org/10.1002/cpz1.411>

## References: (cont'd)

- Srivastava, Malik, Sarkar, Zakeri, Almodaresi, Soneson, Love, Kingsford, Patro (2020) - Alignment and mapping methodology influence transcript abundance estimation (Genome Biology)  
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02151-8>
- Conesa, Madrigal, Tarazona, Gomez-Cabrero, Cervera, McPherson, Szcześniak, Gaffney, Elo, Zhang, Mortazavi (2016) - A survey of best practices for RNA-seq data analysis (Genome Biology)  
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8>

## Resources

The whole `rnaseqGene` workflow is detailed here: <https://www.bioconductor.org/packages/rnaseqGene/>

**... thank you for your attention!**

[marinif@uni-mainz.de](mailto:marinif@uni-mainz.de) -  @FedeBioinfo  
[charlotte.soneson@fmi.ch](mailto:charlotte.soneson@fmi.ch) -  @CSoneson

