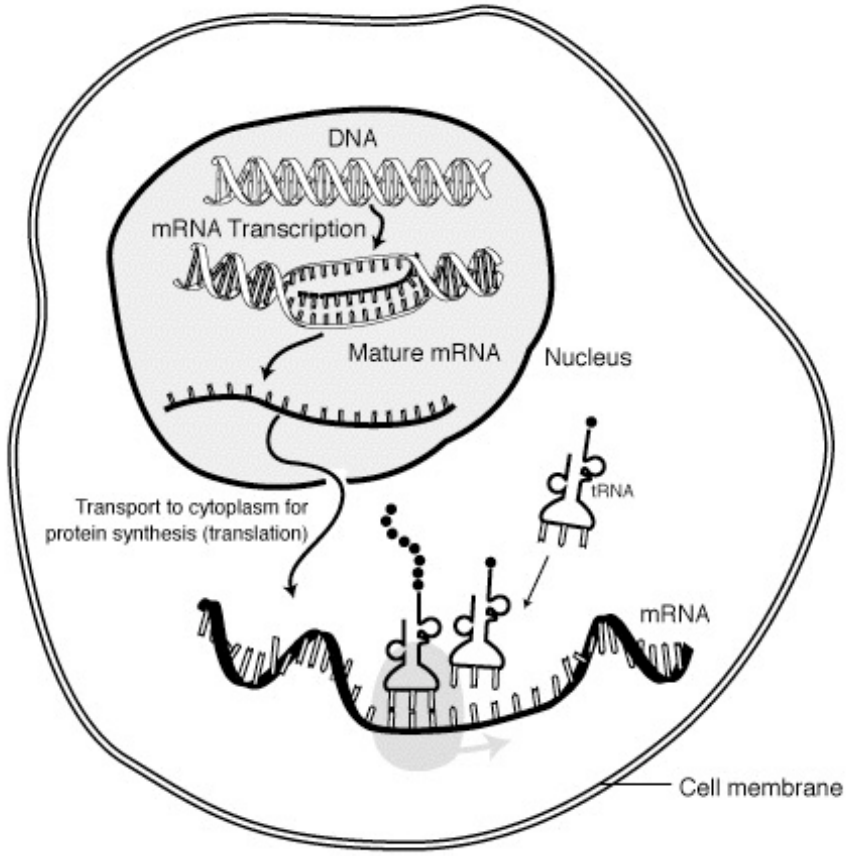
RNA-seq

What is RNA-seq and how does it work?

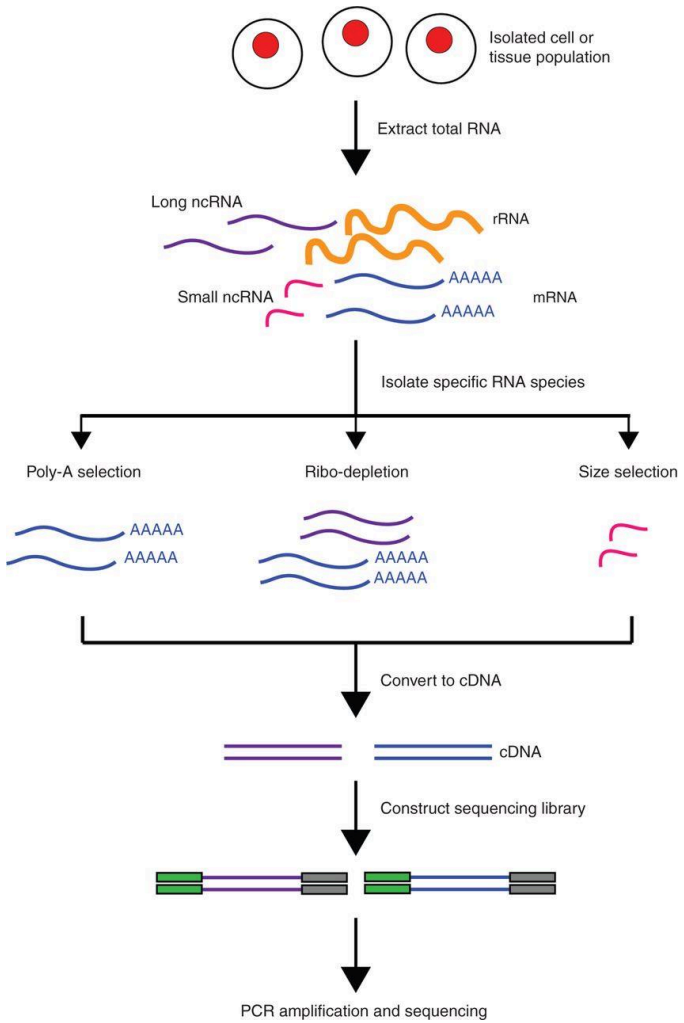
What is RNA-seq?

- RNA-seq = RNA sequencing
- Use next generation sequencing to quantify RNA in a cell
- Mainly when people do RNA-seq they're concerned with mature messenger RNA, because they want to know what genes are expressed



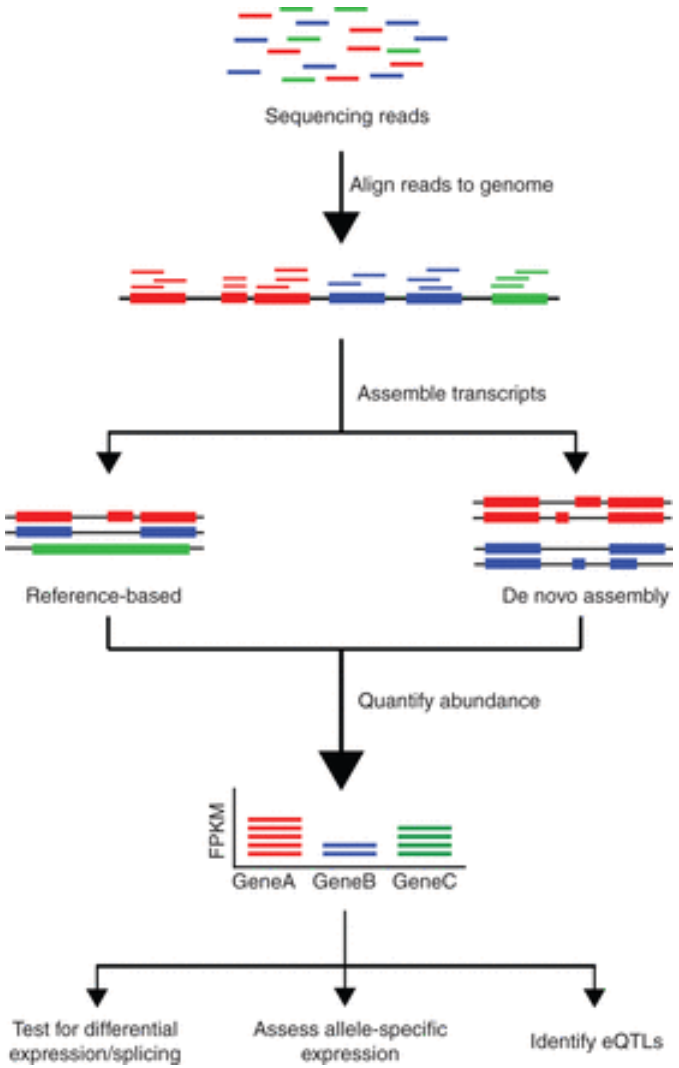
Stark 2019, Image: Wikipedia

RNA-seq workflow



- RNA is extracted from cell
- Only mRNA (sometimes lincRNA) enriched
- Converted to cDNA
- Construct library
- Sent to sequencing

RNA-seq from sequencing reads to expression



unaligned reads

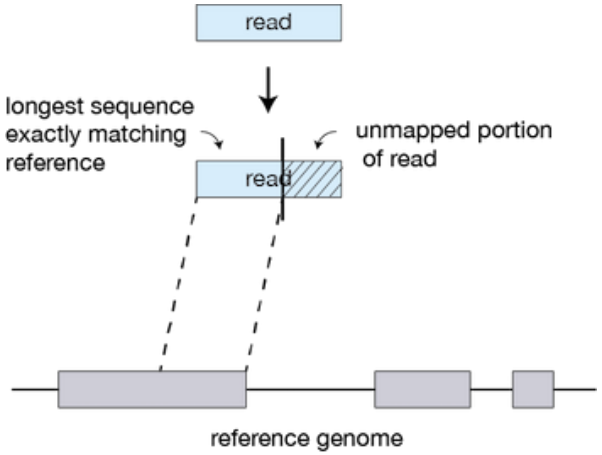
alignment to genome
(while aware of exon structure)

If we care about alternative transcripts,
we can assemble transcriptome

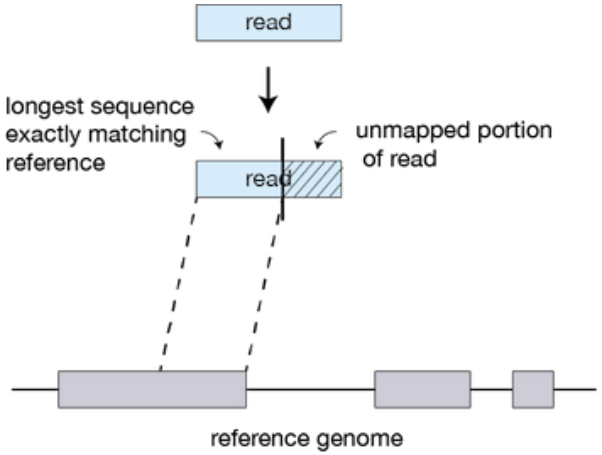
count reads per gene

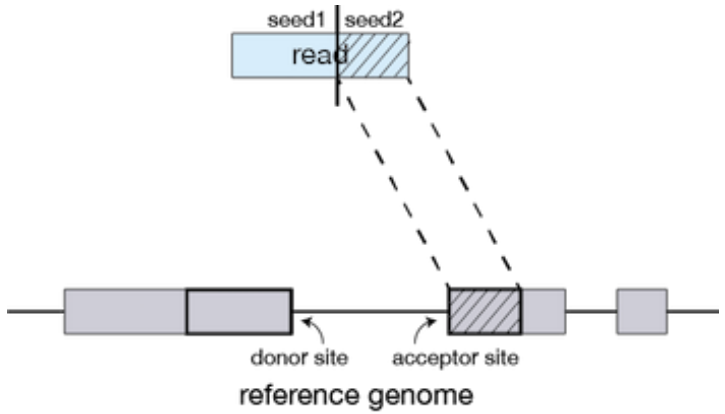
Actual analysis
differential gene expression / allele specific expression / expression QTL

RNA-seq split reads mapping

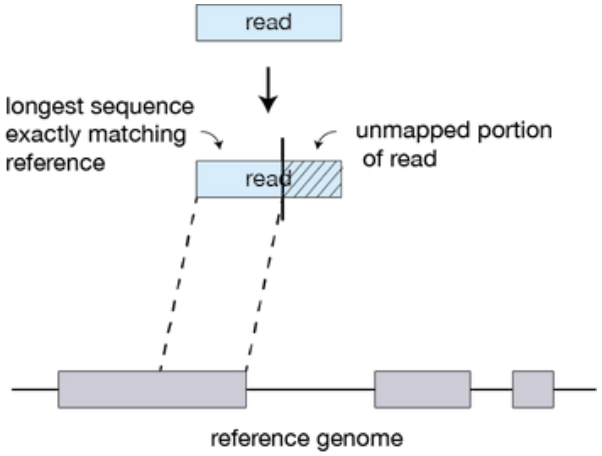


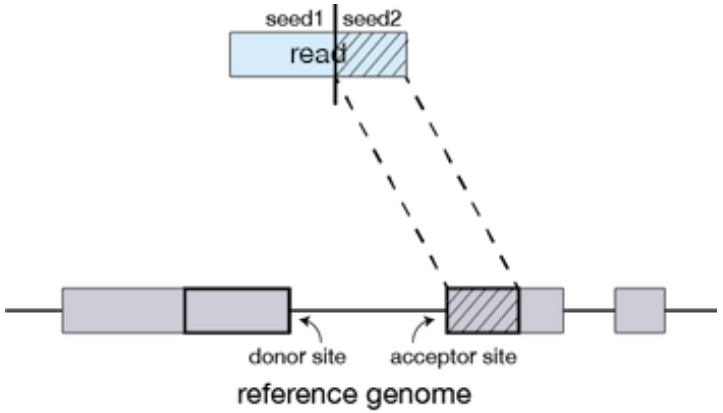
RNA-seq split reads mapping

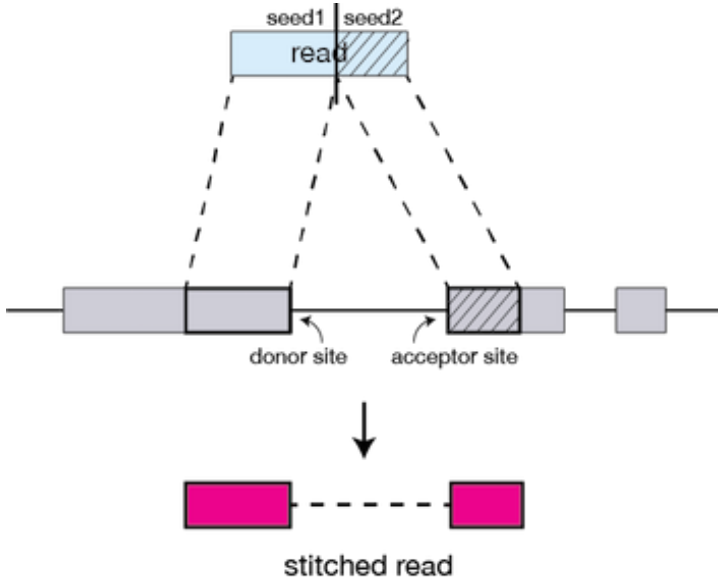




RNA-seq split reads mapping

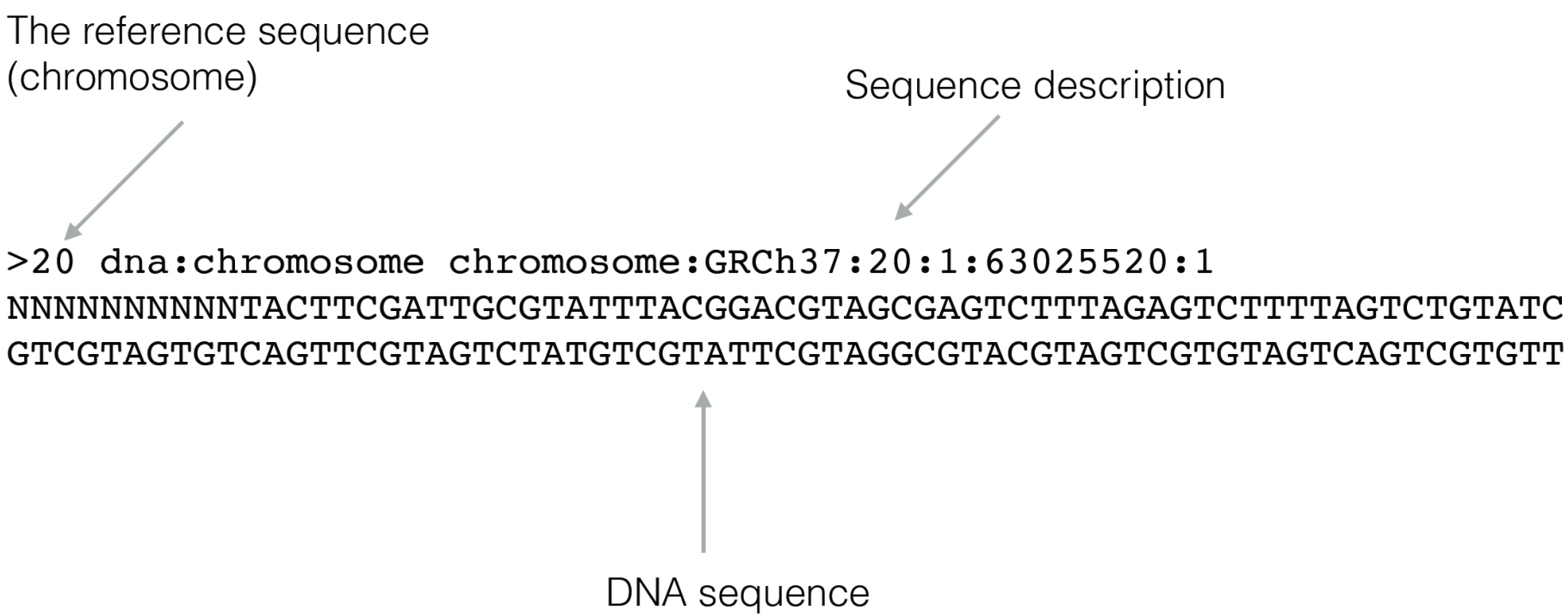






FASTA format

We start with a reference genome to map to



FASTQ format

and sequence reads to map

DNA sequence

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GTCGTAGTGTCAGTTCGTAGTCTATGTCGTATTCGTAGGCGTACGTAGTCGTGTAGTCAGTCGTGTT
+
!#!#!!*!#^^!#(#*&^*$^^#%&$*$&(^&$^^^^$)$*&$**$))))$*&^$*##$))$)))))
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Quality scores

$Q=-10 \log_{10}(P_{err})$

(Some) of the Many Gene IDs

- Ensembl ids
  - ENS[species prefix][feature type prefix][a unique eleven digit number].[version number]
  - Human: ENSG00000168769, Mouse: ENSMUST00000101509.9
- HUGO Gene Naming Consortium sets gene names like BRCA1
- Other ids:
  - Entrez
  - UCSC
  - NCBI
  - RefSeq
  - OMIM
  - Uniprot
- If you don't know what type of id it is, google the id

Gene name errors are widespread in the scientific literature

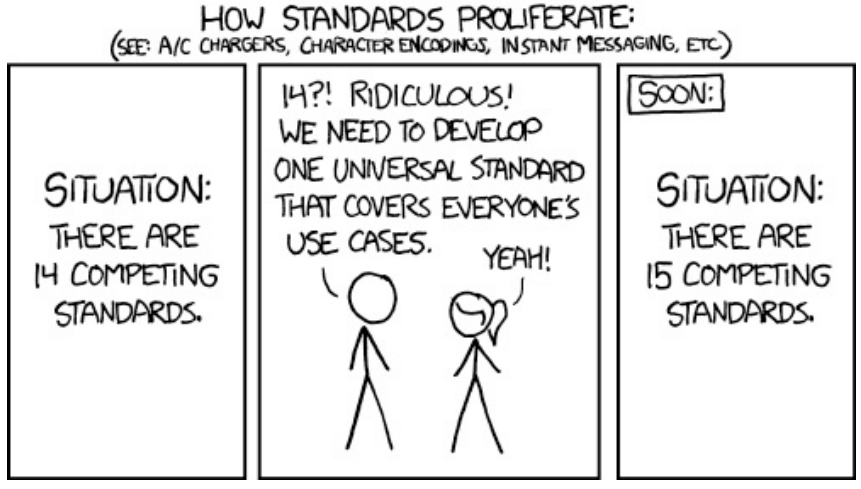
Mark Ziemann, Yotam Eren & Assam El-Osta

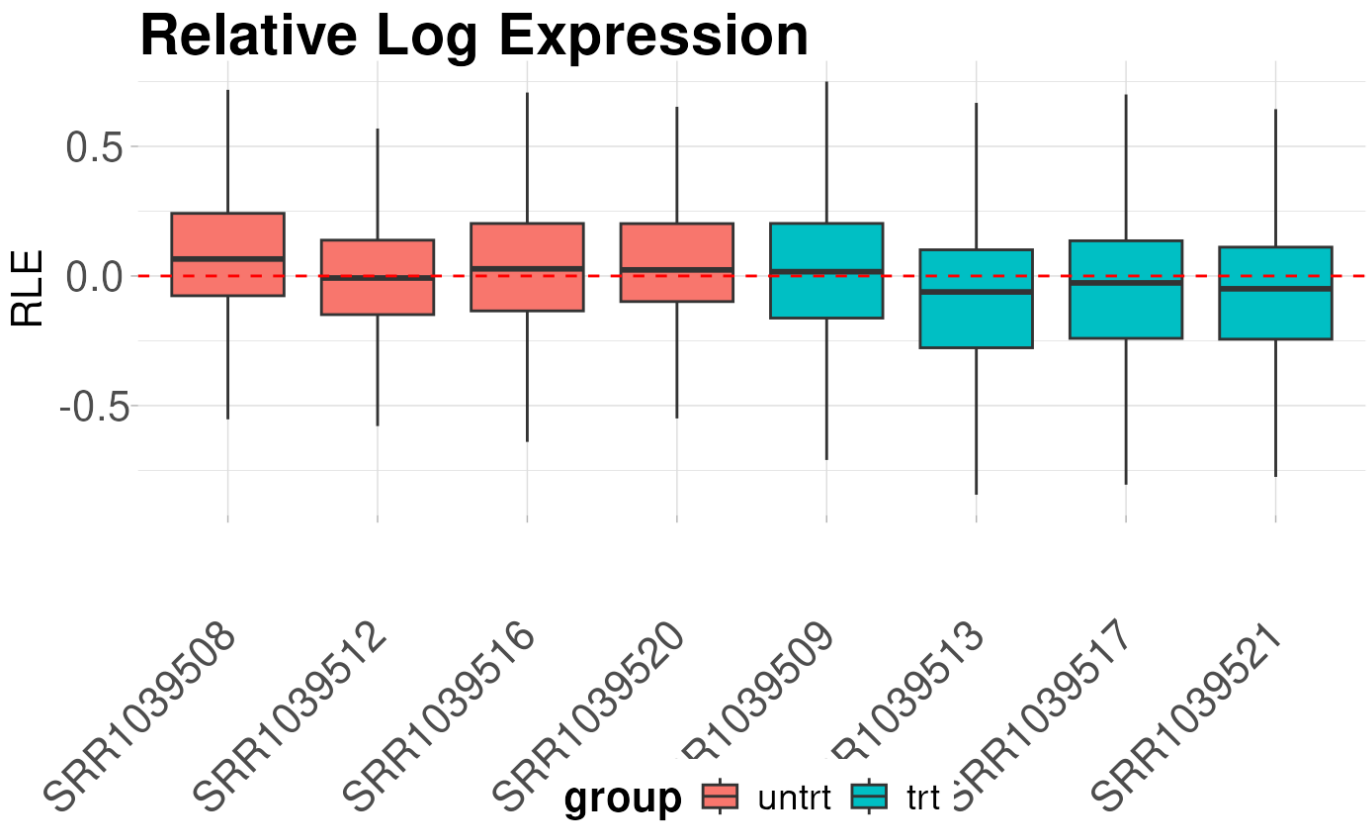
Genome Biology **17**, Article number: 177 (2016) | Cite this article

134k Accesses | **53** Citations | **3209** Altmetric | Metrics

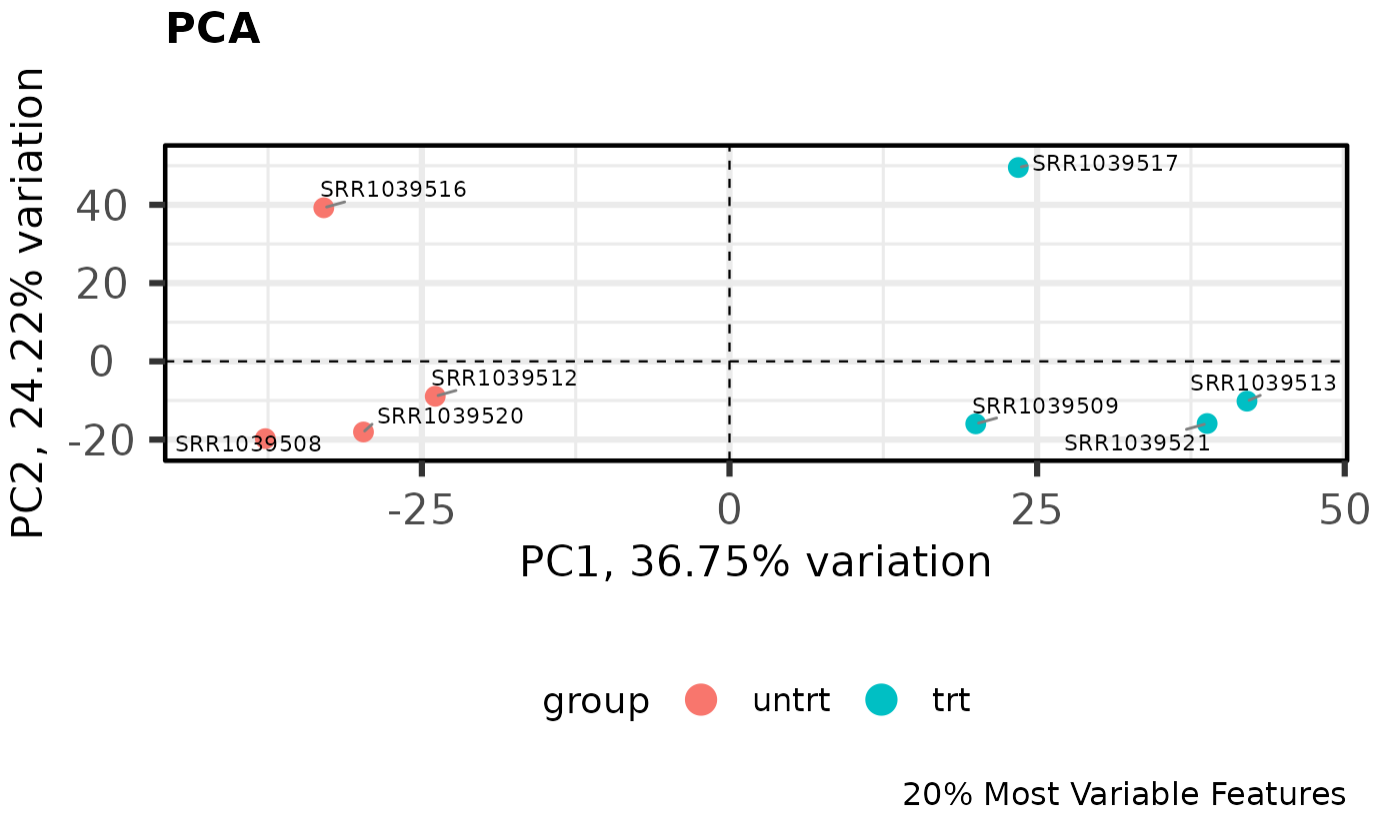
Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

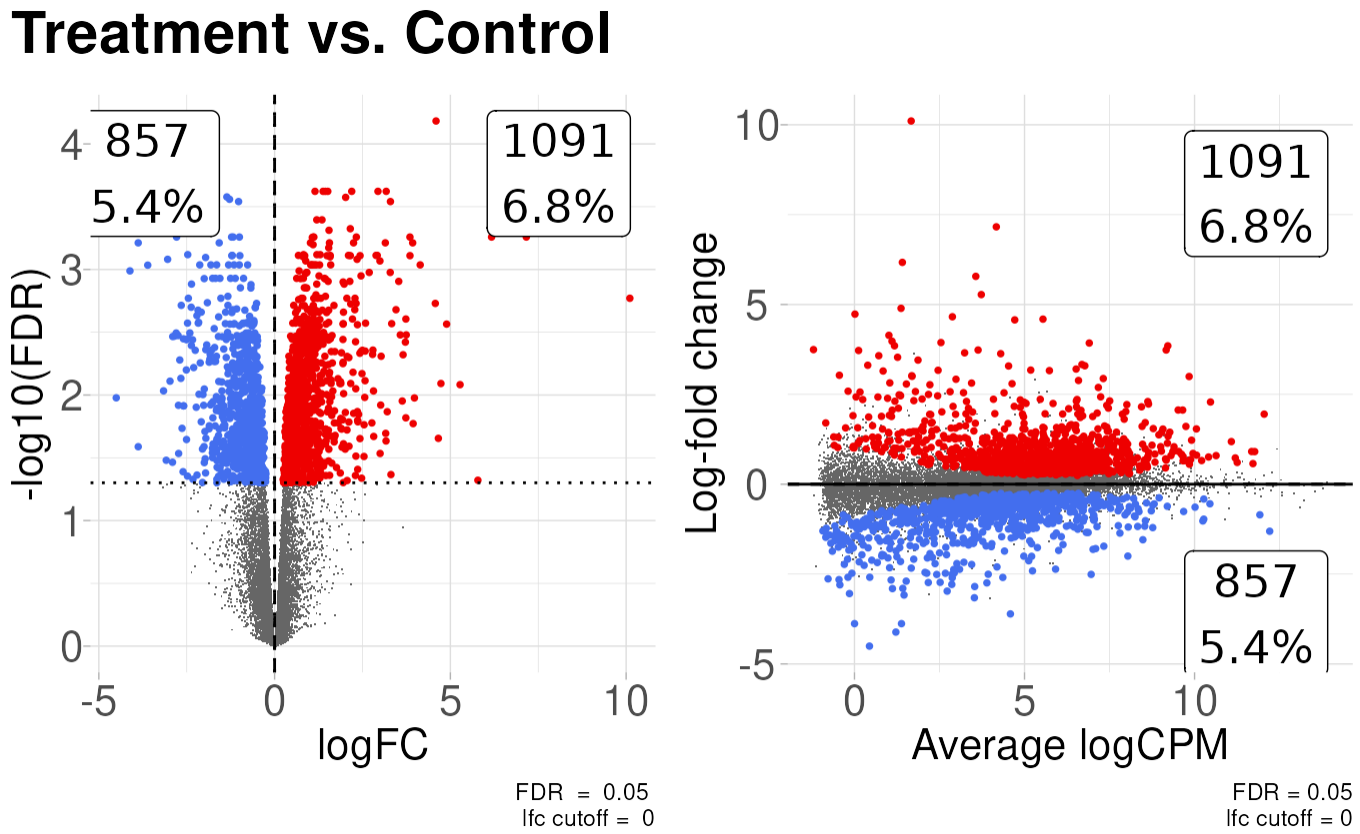



Relative Log Expression
RLE
0.5
0.0
-0.5
SRR1039508
SRR1039512
SRR1039516
SRR1039520
SRR1039509
SRR1039513
SRR1039517
SRR1039521
group
untrt
trt

Plotting DE results: PCA

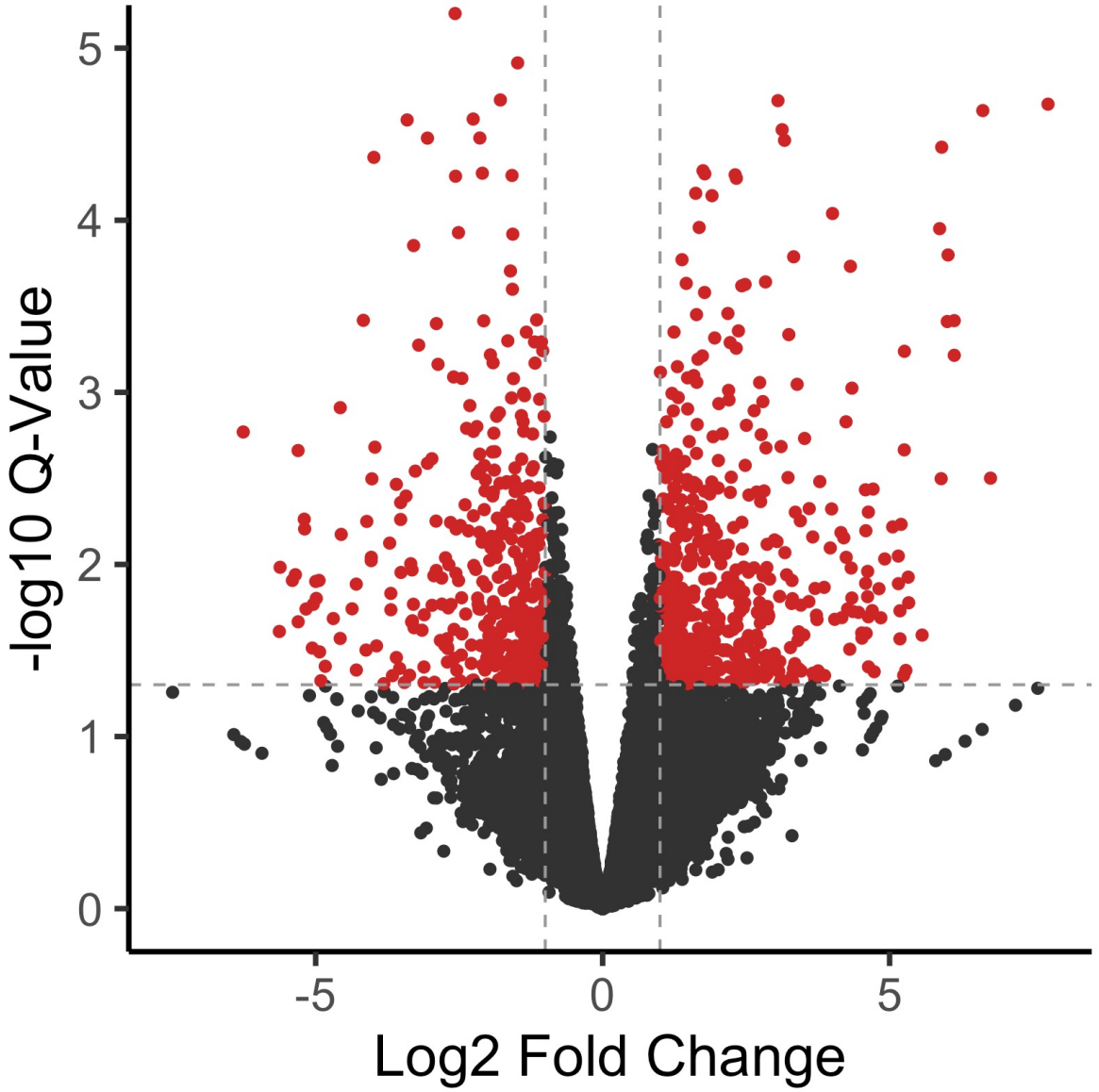


Plotting DE results: Volcano and MA plot



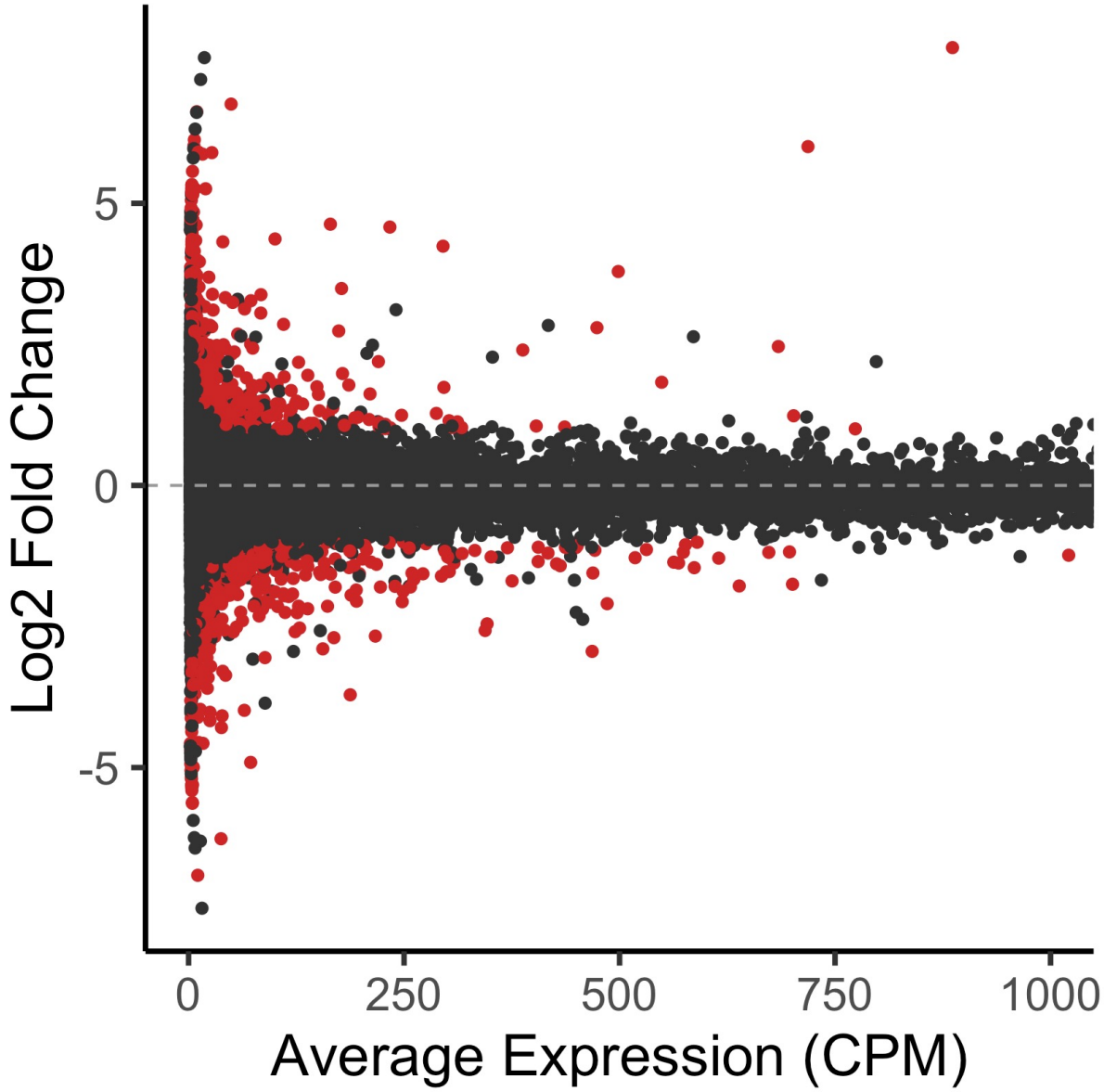
Volcano Plot

- X-axis shows the average log2 fold change for the variable of interest
- Y-axis negative log10 corrected p-value
- Overall shows how the expression in the treatment group changes and how statistically significant the changes are

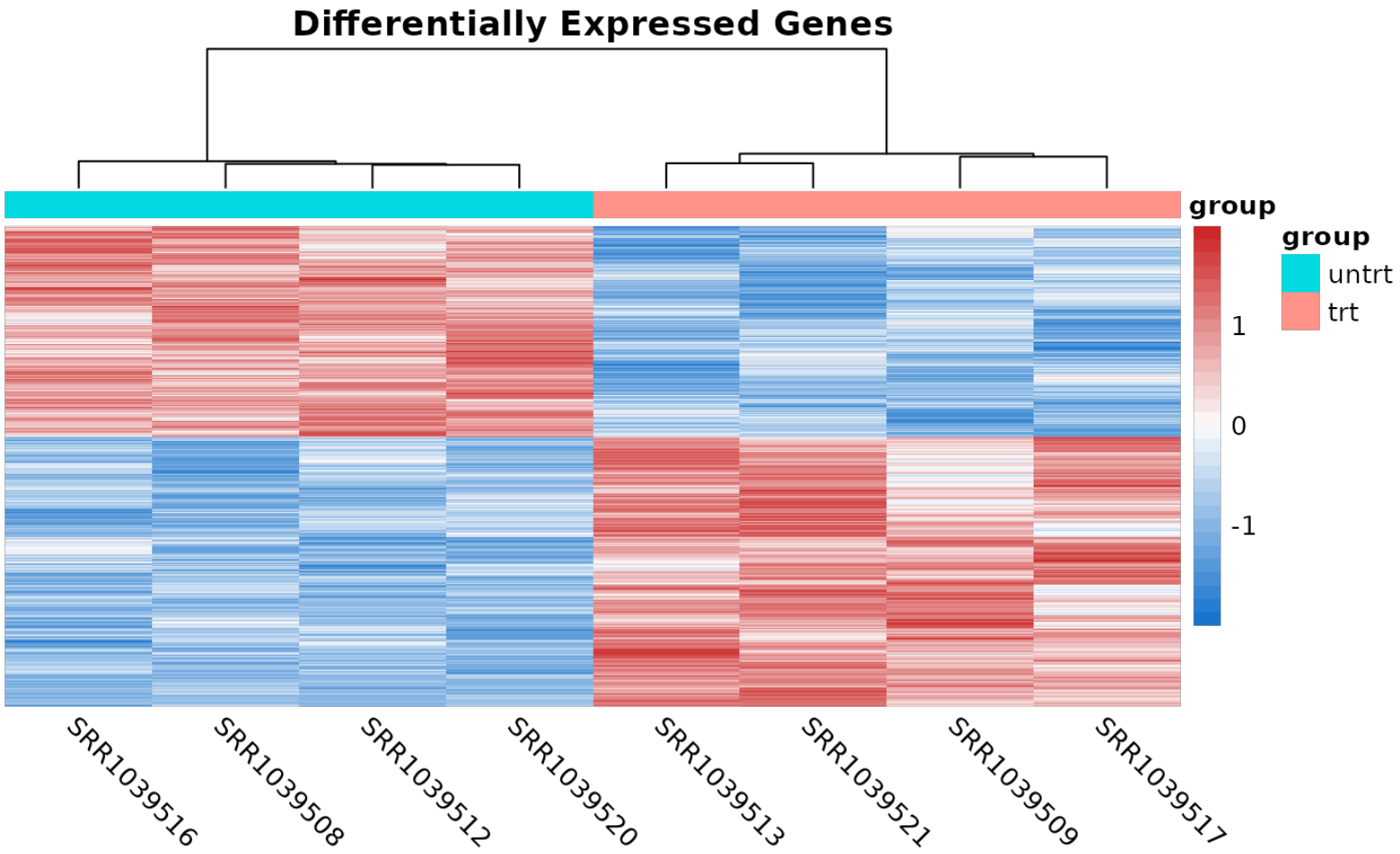


MA Plot

- X-axis shows the average expression for all samples
- Y-axis shows the log2 fold change for the variable of interest
- Overall shows how the expression in the variable group changes relative to all other samples



Plotting DE results: Heatmap



UpSet Plot

- New variation on a venn diagram because venn diagrams are not great at visualizing more than 2 or 2 sets https://en.wikipedia.org/wiki/Venn_diagram
- Points and line on the bottom show the overlap between the conditions on the left
- Bar height and counts show the number of features that overlap