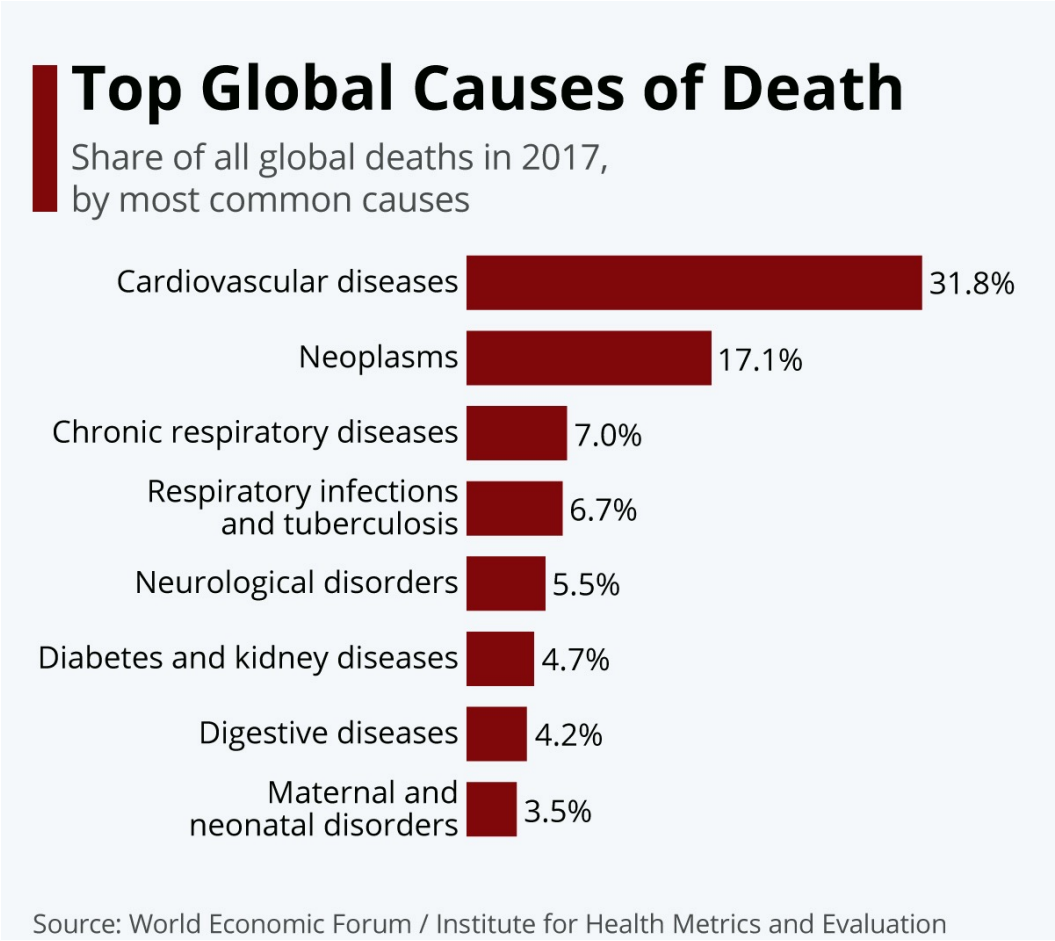
A Short Introduction to Transcriptomics

Kristoffer Vitting-Seerup, PhD

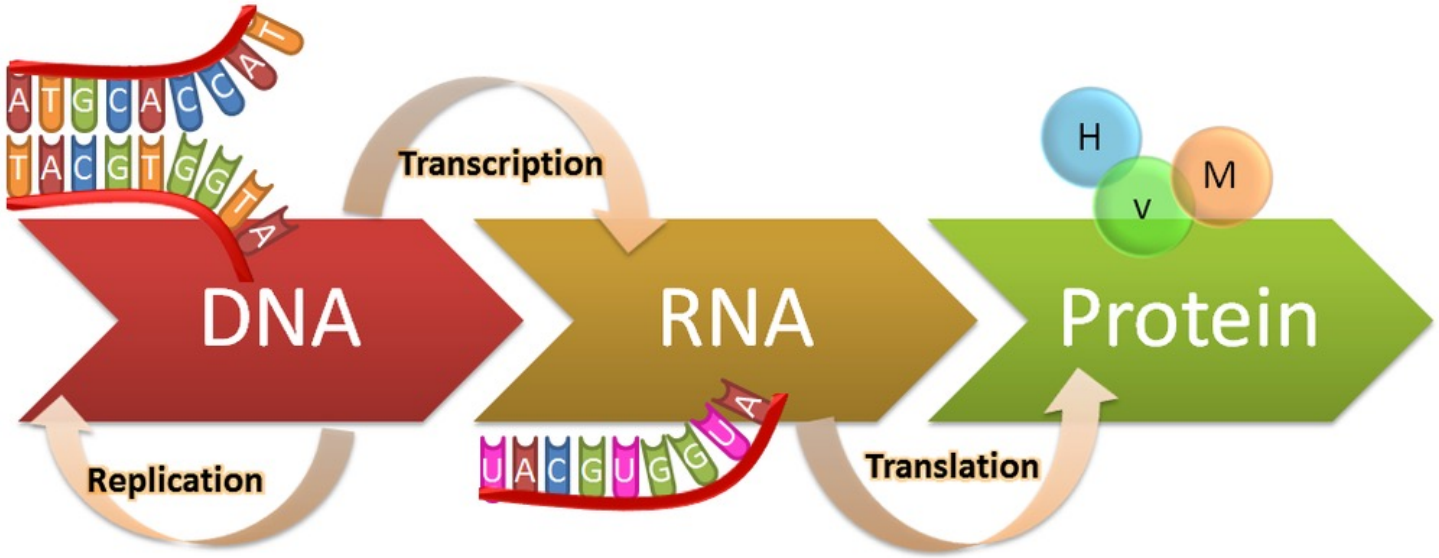
Assistant Professor in Bioinformatics

krivi@dtu.dk



To treat most of these we need to understand the molecular mechanisms and how they are change by disease

Aim: Profile difference between healthy and sick

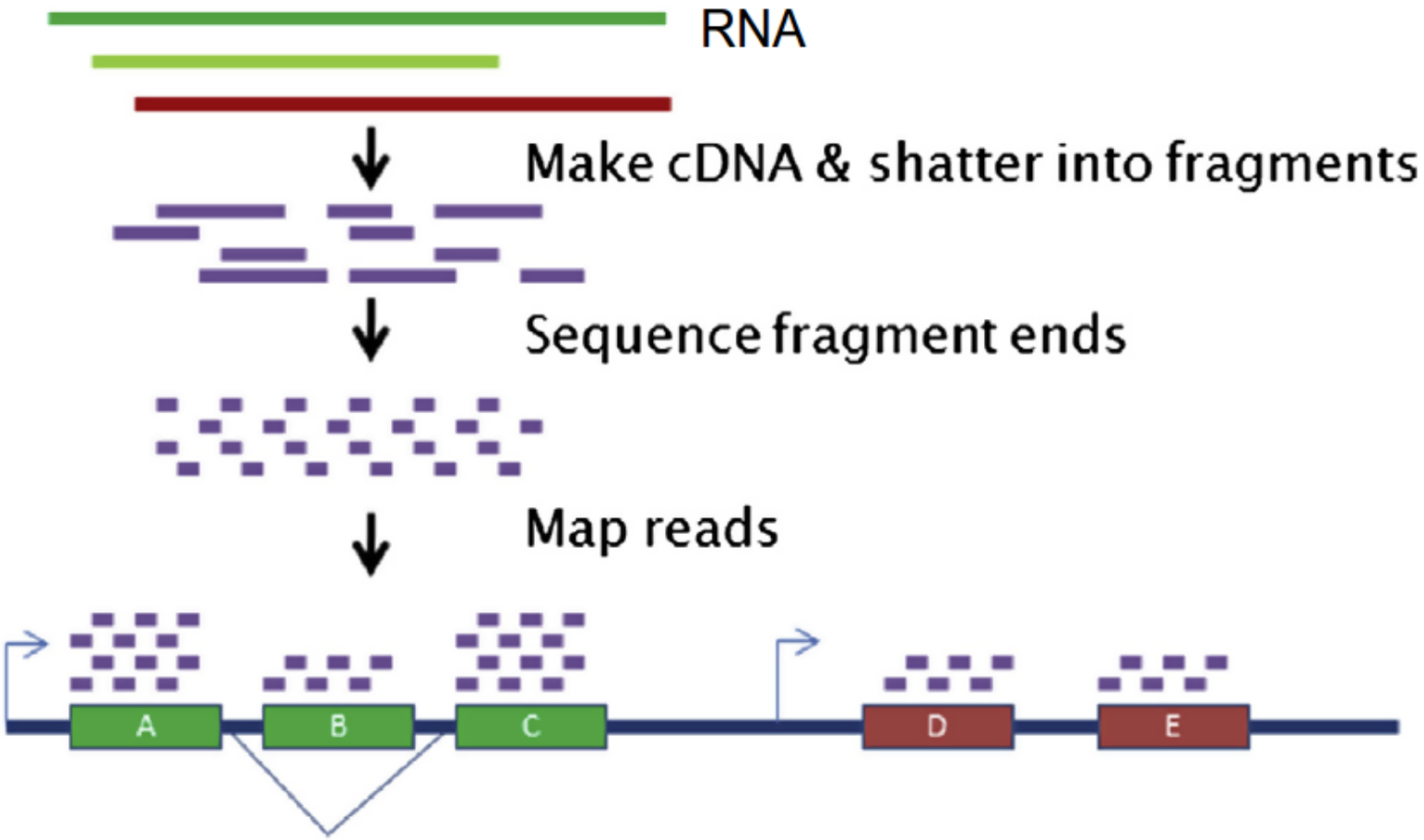


Solution: Measure all <u>DNA</u>, <u>RNA</u> or all <u>protein</u> in healthy and sick

Genomics **Transcriptomics** Proteomics

RNA-sequencing 101

RNA-sequencing

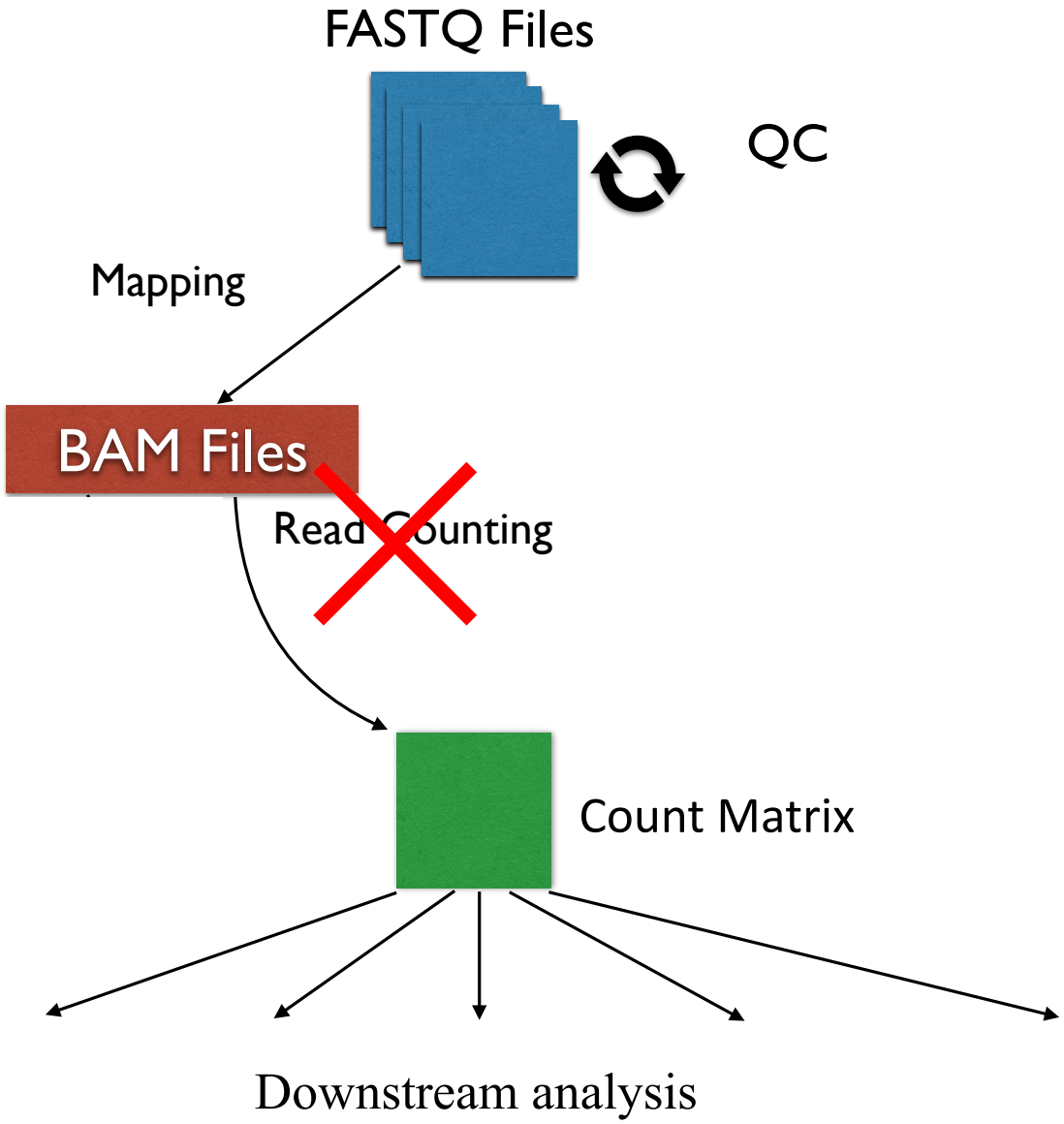


Count read mapped to each Gene

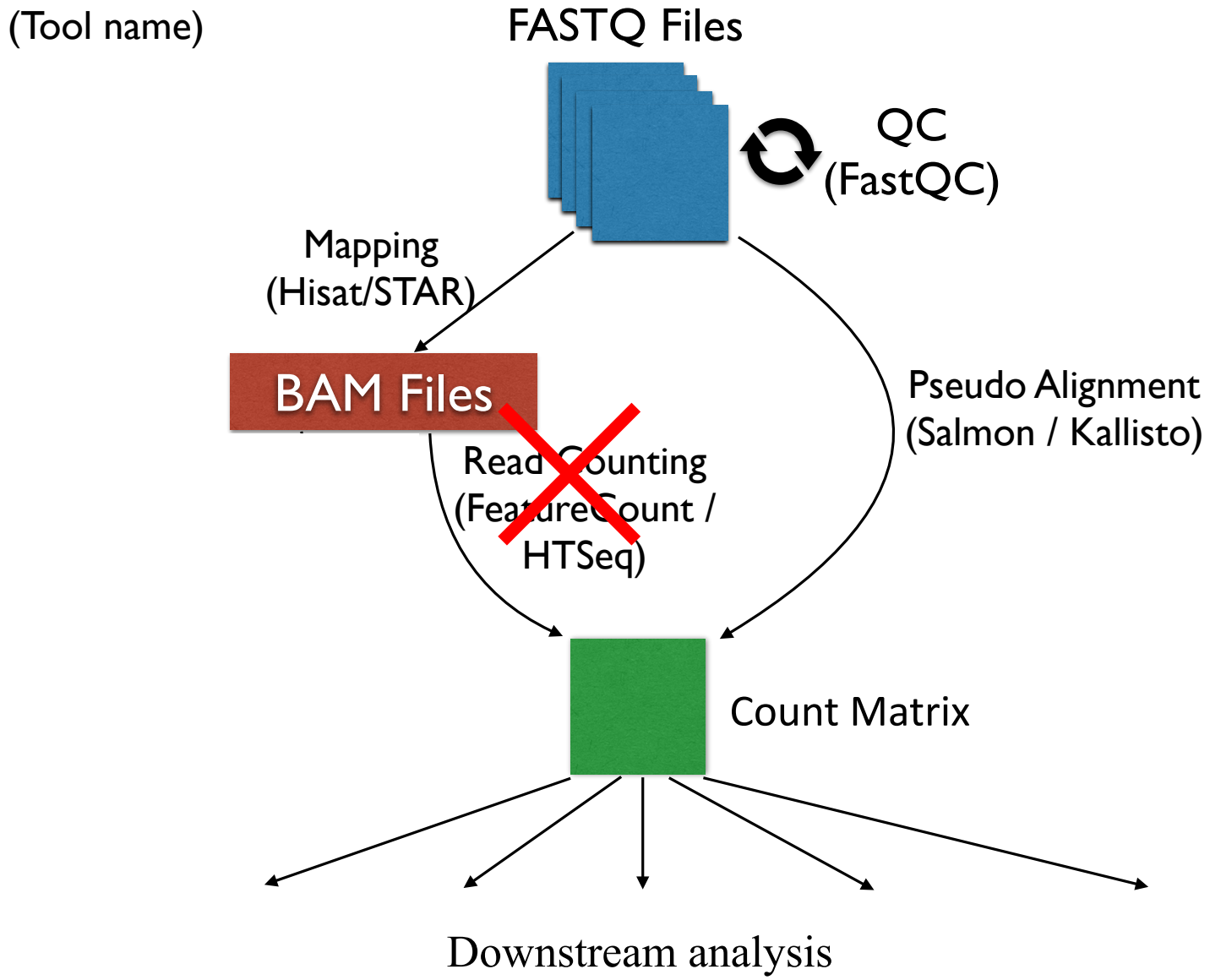
=

Count Matrix

More detailed workflow



More detailed workflow



Gene Count Matrix

The results from a single sample RNA-seq datasets:

Identifier	sample1
DDR1	884
RFC2	422
HSPA6	621
PAX8	658
GUCA1A	426
UBA7	524
THRA	564
PTPN21	909
CCL5	771
CYP2E1	315
EPHB3	362
ESRRA	911
CYP2A6	409
GAS6	368
MMP14	3
TRADD	102
FNTB	368
PLD1	661

Read count of each “feature”

Typically, 20,000 – 50,000 genes

Gene Count Matrix

Samples →

Genes ↓

	sample1	sample2	sample3	sample4	sample5
DDR1	884	81	622	946	541
RFC2	422	858	305	830	28
HSPA6	621	130	401	221	110
PAX8	658	176	573	999	296
GUCA1A	426	171	723	659	64
UBA7	524	579	19	199	510
THRA	564	110	119	384	488
PTPN21	909	668	810	781	852
CCL5	771	838	217	837	816
CYP2E1	315	978	289	718	583
EPHB3	362	488	359	850	558
ESRRA	911	724	41	476	264
CYP2A6	409	683	286	998	690
GAS6	368	222	212	642	747
MMP14	3	833	878	552	511
TRADD	102	479	844	327	43
FNTB	368	284	591	729	747
PLD1	661	489	706	859	587

RNA-sequencing Count Matrix

- 3 minutes with neighbour:
- You are analysing 2 genes (gene A and B) in two conditions (condition 1 and 2) on the basis of a RNA-seq experiment that resulted the following number of reads:

	Condition 1	Condition 2
Gene A	1000	3000
Gene B	2000	4000

- Question: Is the following statement correct:
  “Both gene A and B are more expressed in condition 2”
  Explain why/why not.

RNA-sequencing Count Matrix

Solution

- Question: Is the following statement correct:
  "Both gene A and B are more expressed in condition 2"
  Explain why/why not.

- We cannot state any such thing since we do not know the sequencing depth (library size)

- Solution: Normalise for library size (divide by library size)

- Similarly there are many other things to normalize for:
  Gene length, GC content, number of reads per fragment

RNA-seq Normalization

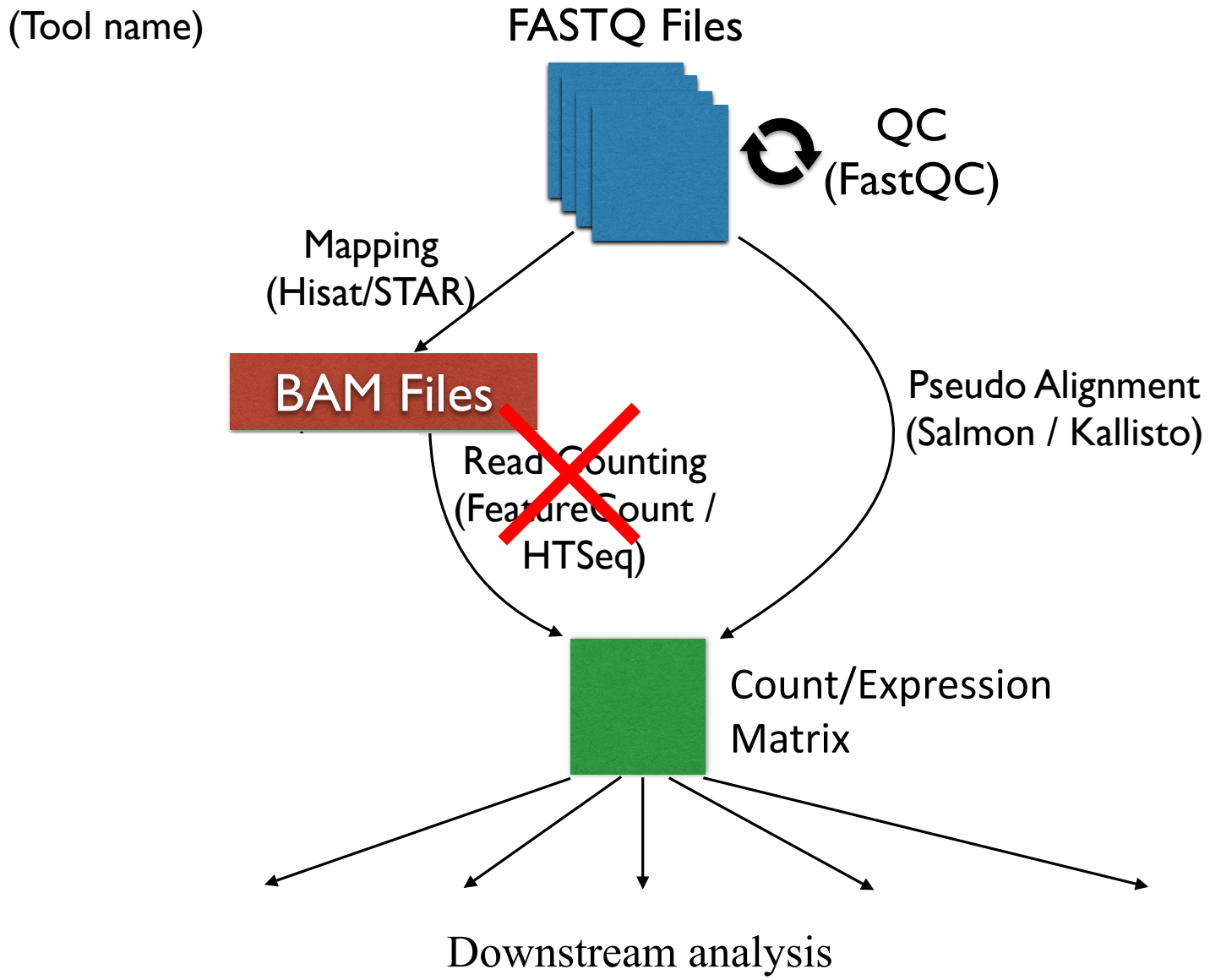
Solution

- Raw read count: Reads per genes

- RPM/CPM: Reads per million reads mapped (normalized for library size)

- TPM: Transcripts per million (normalized for library size and gene-length)

More detailed workflow



Downstream Analysis

Two of the hundreds of possible uses

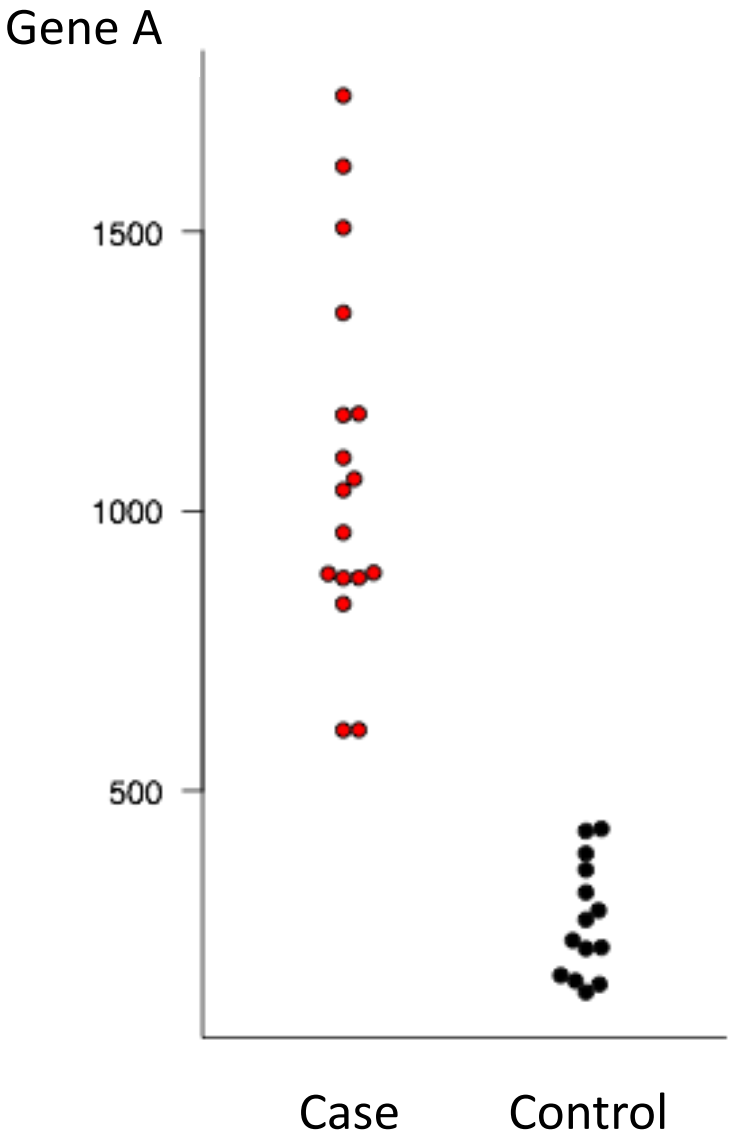
What question do you want to answer?

Typically, we use transcriptomics to compare between two or more groups, generally referred to as a case/control study.

Examples include:

- Disease vs. normal
- Drug treatment vs. control
- Good prognosis vs. bad prognosis
- Timepoint 1 vs timepoint 2

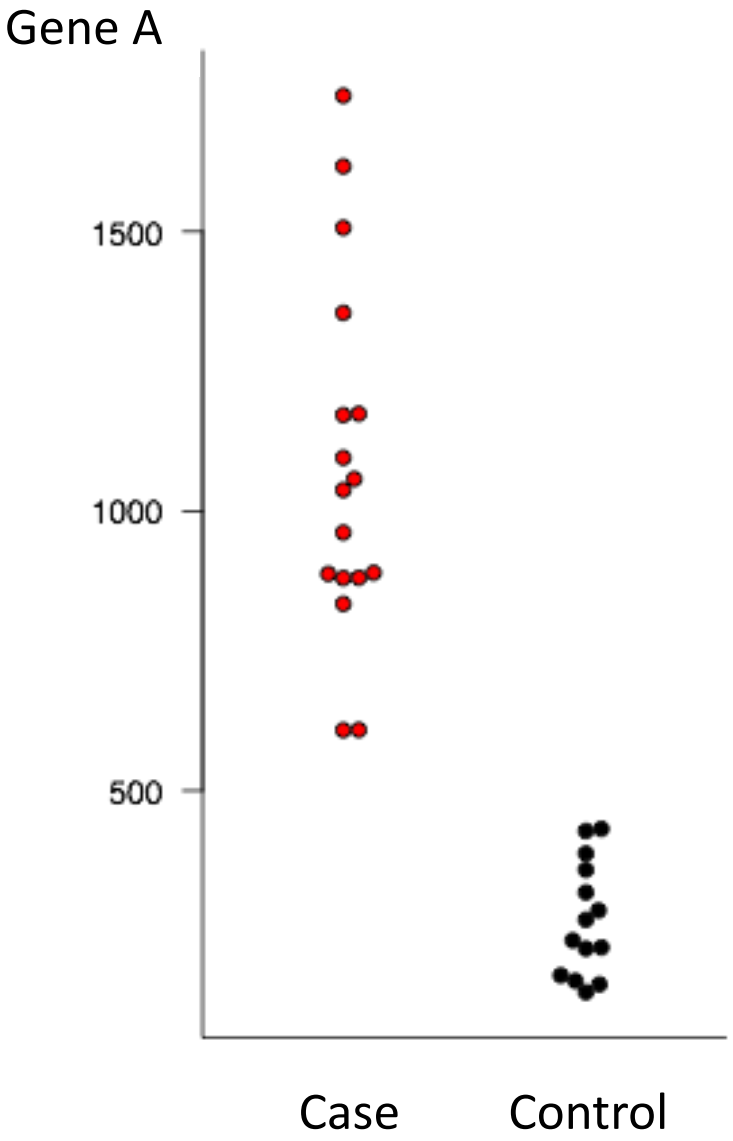
Case/Control Study



2 Min with neighbor:

- What would we like to know/summarize?

Between Group Comparison



Solution

What would we like to know/summarize?

Measures of interest:

- Effect Size:
  Fold Change
- Certainty of observed difference:
  P-value (statistical model)

Differential Expression Analysis

- A nice, normalized expression matrix:

	Healthy 1	Healthy 2	Healthy 3	Disease 1	Disease 2	Disease 3
Gene A	10	12	9	9	8	13
Gene B	0	5	1	25	26	22
Gene C	35	45	55	0	8	10

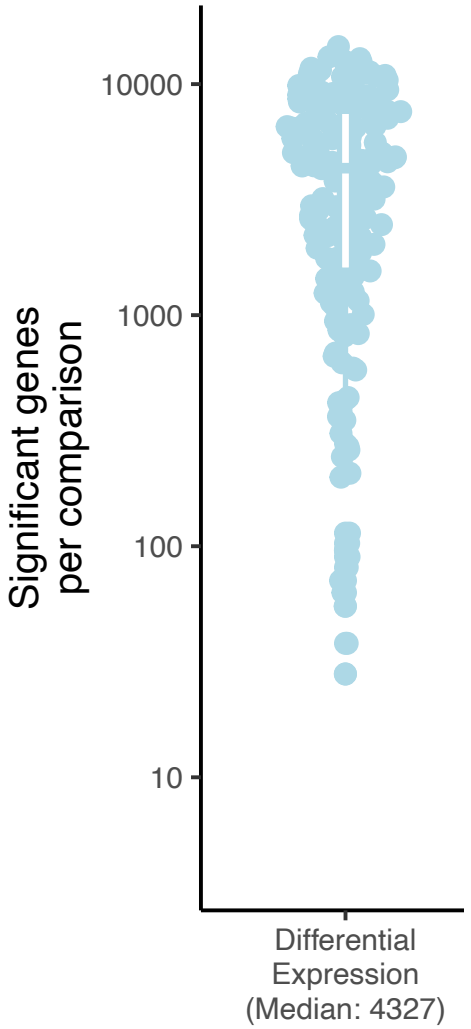
Modern bioinformatic tool jointly estimates both log2FCs and P-values using (generalized) linear modeling

Differential Expression Analysis

- Done by advanced well-tested bioinformatics tools
  - DESeq2
  - EdgeR
  - Voom-limma
- Count matrix as input (do normalization internally)
- They use (generalized) linear models
  - Can take unwanted effects into account

In the real world

- We recently did a systematic analysis of 100 RNA-seq datasets
- On average thousands of genes change significantly between conditions (!)
- How do you make sense of such a list?



Dam *et. al.* 2022

Gene-sets

- Collection of genes that have something in common
  - Participate in the same process (e.g. cell cycle)
  - Have the same molecular function (e.g. DNA binding)
  - Cellular location (e.g. nucleus)
  - Identified by Kristoffer (e.g. what I just found in my data)
- Many databases
  - Gene Oncology (GO-terms) (http://geneontology.org/)
  - MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/)
  - Enrichr (https://maayanlab.cloud/Enrichr/#libraries)

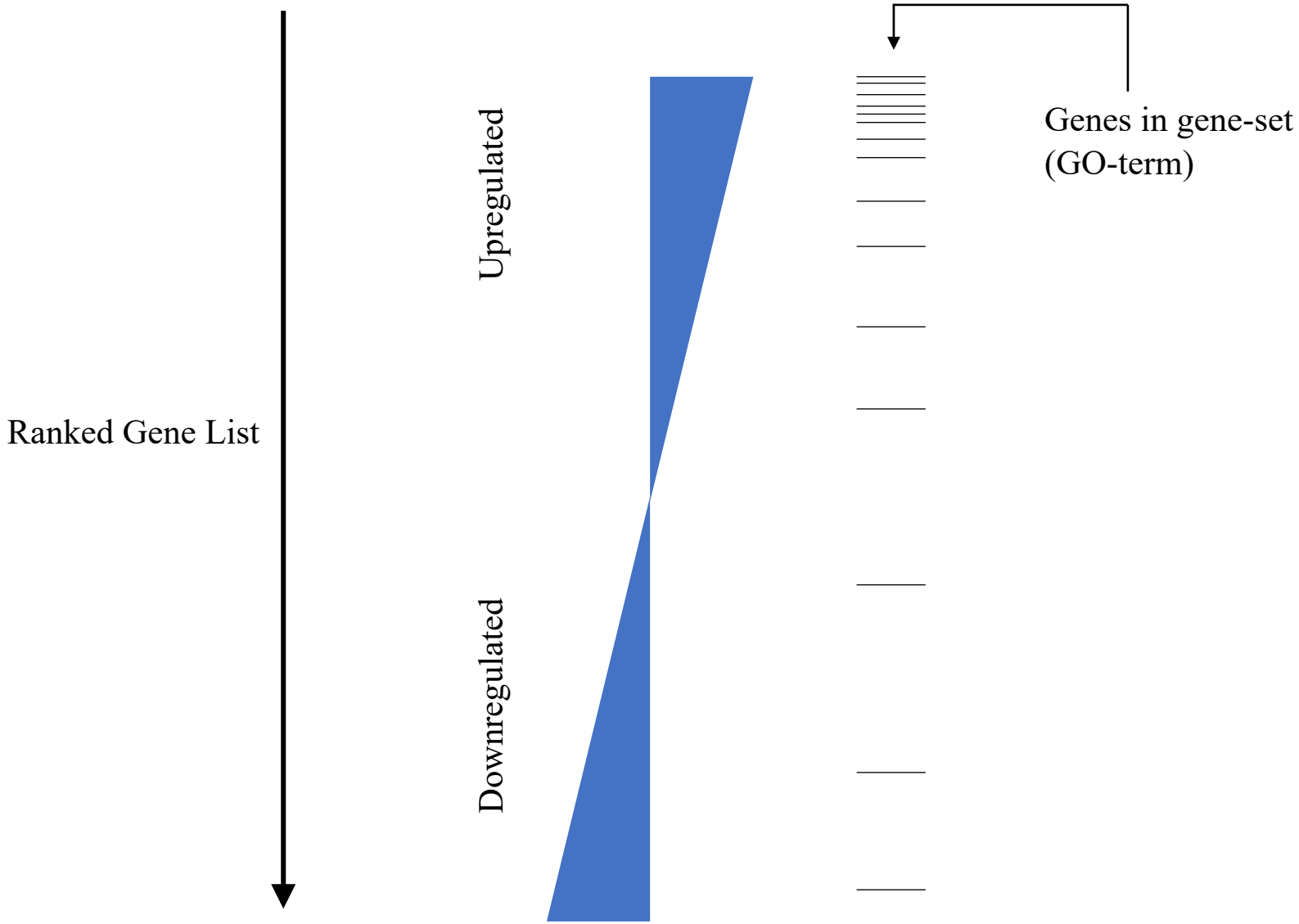
Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) can be done in two ways:

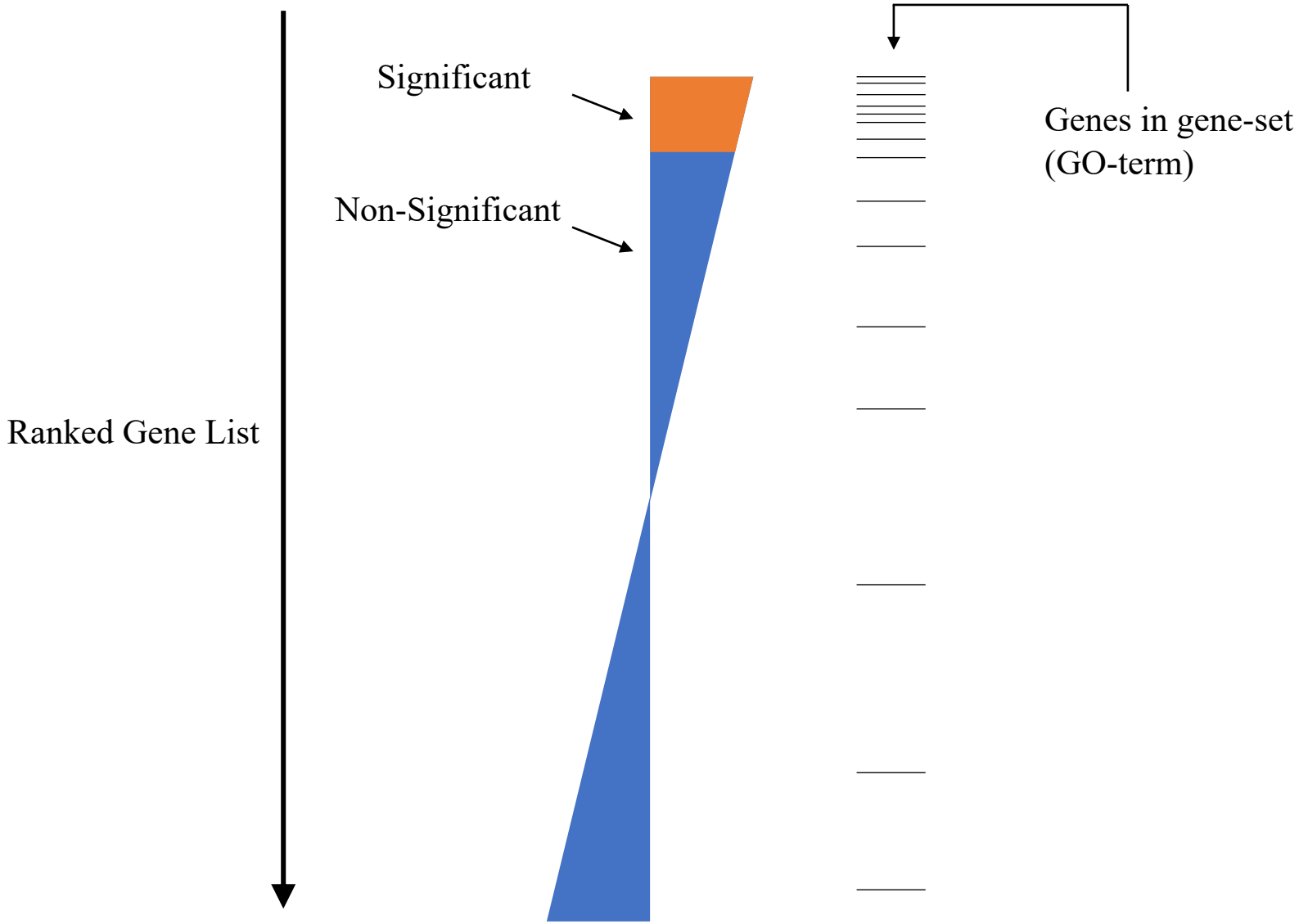
1. Overrepresentation Analysis (OA/ORA)
2. Functional Class Scoring (FSC)

Confusingly these are both referred to as GSEA

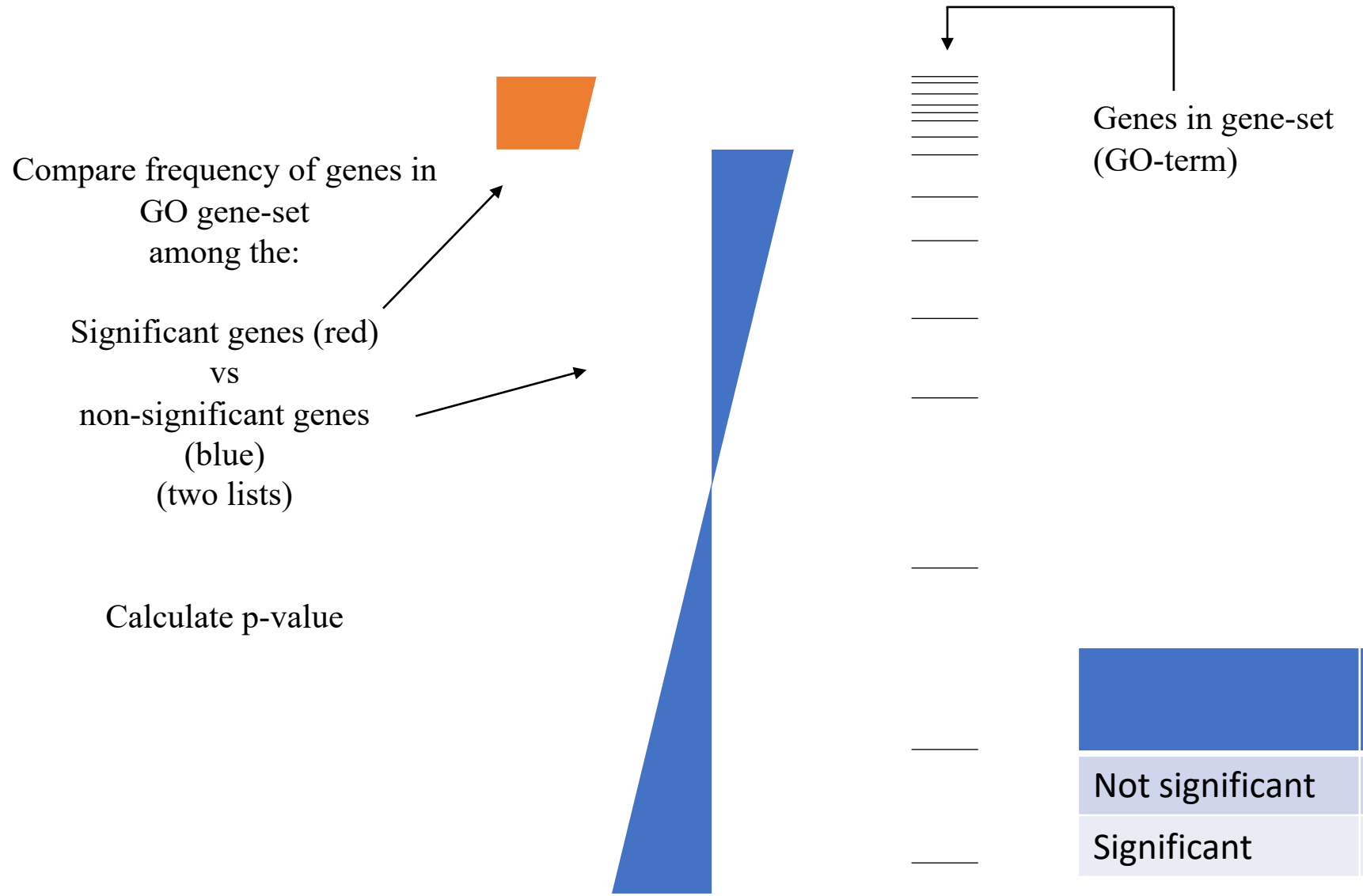
Overrepresentation Analysis



Overrepresentation Analysis



Overrepresentation Analysis



	Not in gene-set	In gene-set
Not significant	1000	100
Significant	10	10

Note on Overrepresentation Analysis

Can be use for anything where you can divide observations into 4 groups based on two binary categories

- Are significant genes enriched for genes in a gene-set?
- Are people with horn-rimmed glasses also typically taller than 2m?
- Are people with biking helmets enriched amongst students at DTU?
- Etc…

GSEA

- Literally hundreds of tools for doing it!

- R packages
  - fgsea
  - clusterProfiler
  - limma
  - gProfiler
  - **pairedGSEA**
- Web tools
  - http://geneontology.org/
  - https://david.ncifcrf.gov/
  - https://biit.cs.ut.ee/gprofiler/gost
- Pay attention to your background!

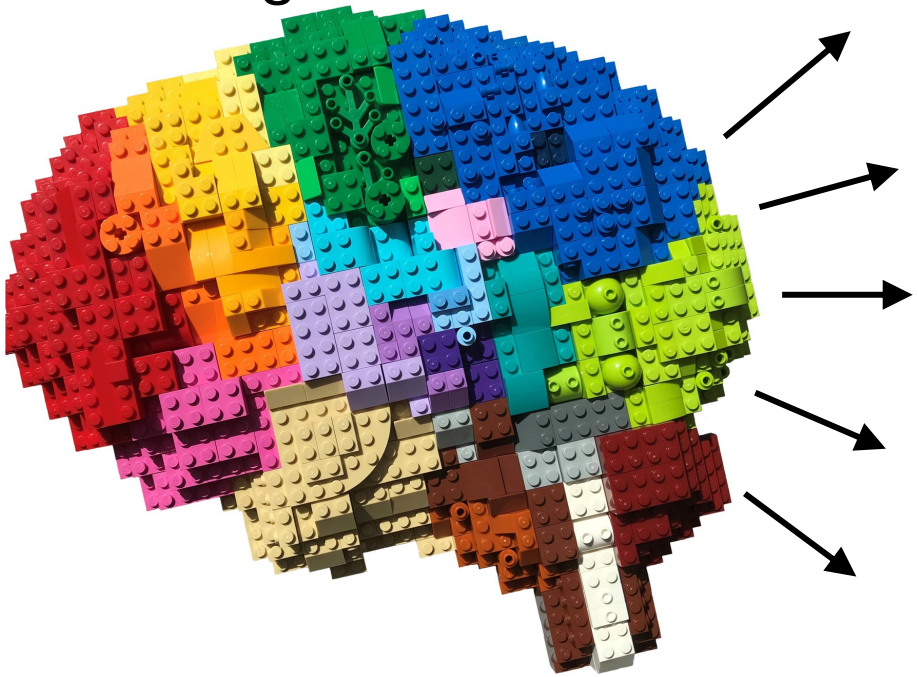
RNA-sequencing 101 – Done!

Extention #1

Bulk vs Single cell vs Spatial

Transcriptomics

Organ/Tissue



Bulk RNA-seq

Single Cell RNA-seq

Modified from @BoXia7, https://twitter.com/adj_23/status/1261928476811997184

Bulk RNA-seq

- 2 Min with neighbors:
  What potential problems are there with just measuring the average signal (compared to single cell)?

Bulk RNA-seq

Single Cell RNA-seq

Bulk RNA-seq

- 2 Min with neighbors:
  What potential problems are there with just measuring the average signal (compared to single cell)?
- Signal you find could be because there is more of a cell type or the expression in a cell type is upregulated
- Bulk is poor for detecting changes occurring in a small fraction of cells

Bulk RNA-seq

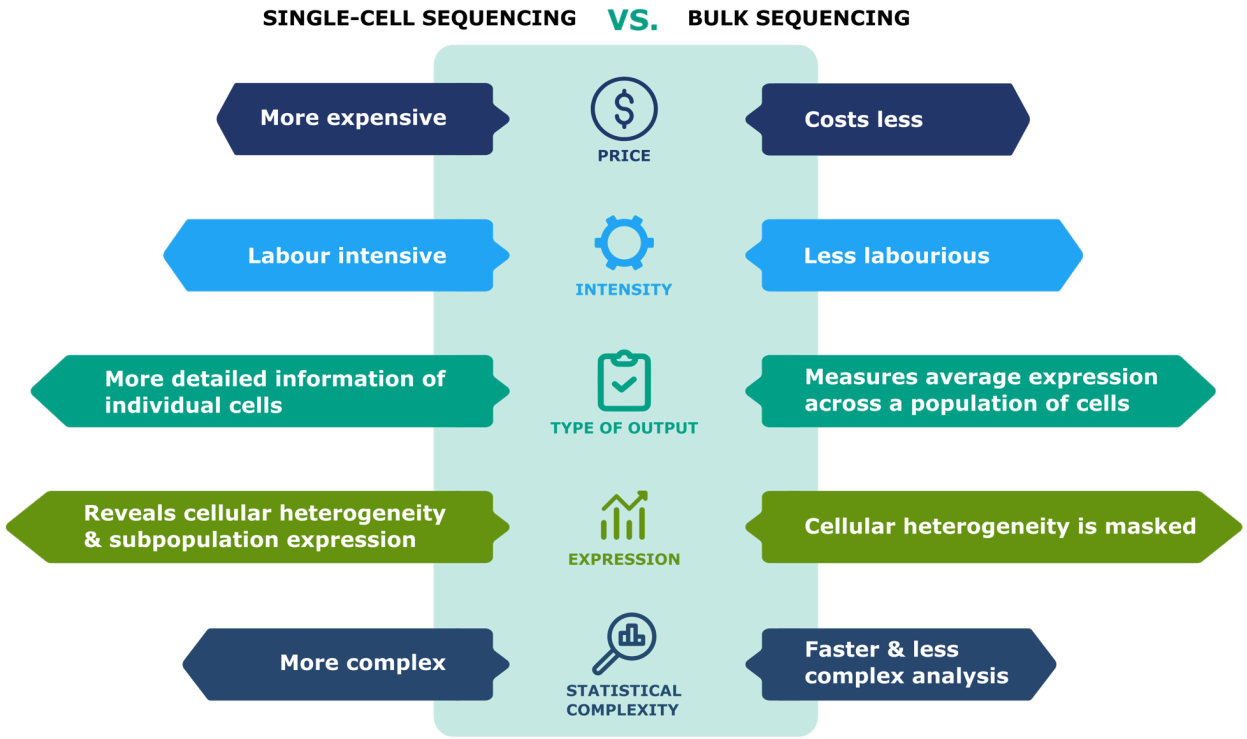
Solution

Single Cell RNA-seq

Bulk RNA-seq

Solution

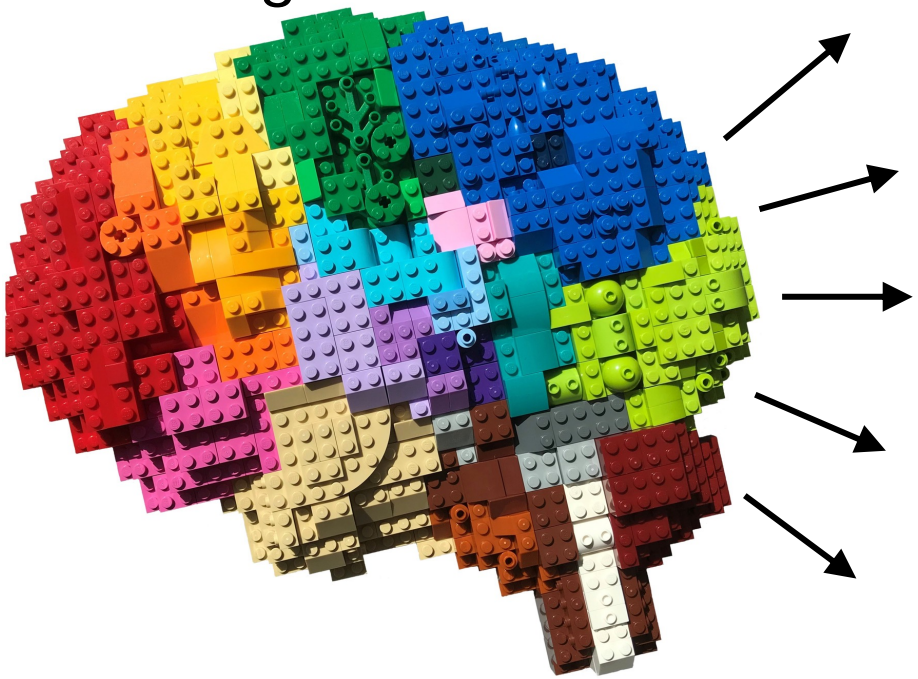
- 2 Min with neighbors:
  What potential problems are there with just measuring the average signal (compared to single cell)?
- Signal you find could be because there is more of a cell type or the expression in a cell type is upregulated
- Bulk is poor for detecting changes occurring in a small fraction of cells



https://lizard.bio/blog/single-cell-vs-bulk/

Transcriptomics

Organ/Tissue

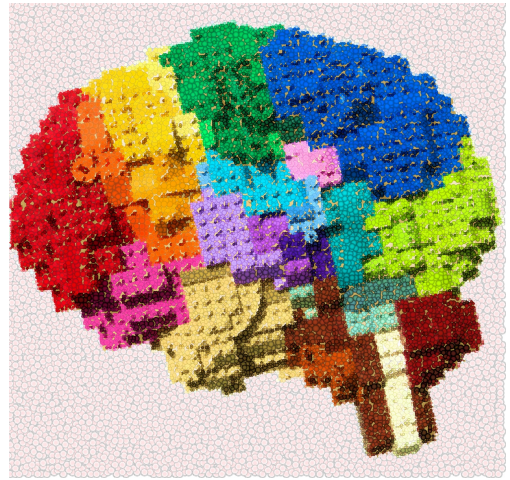


Bulk RNA-seq

Spatial RNA-seq

Single Cell RNA-seq

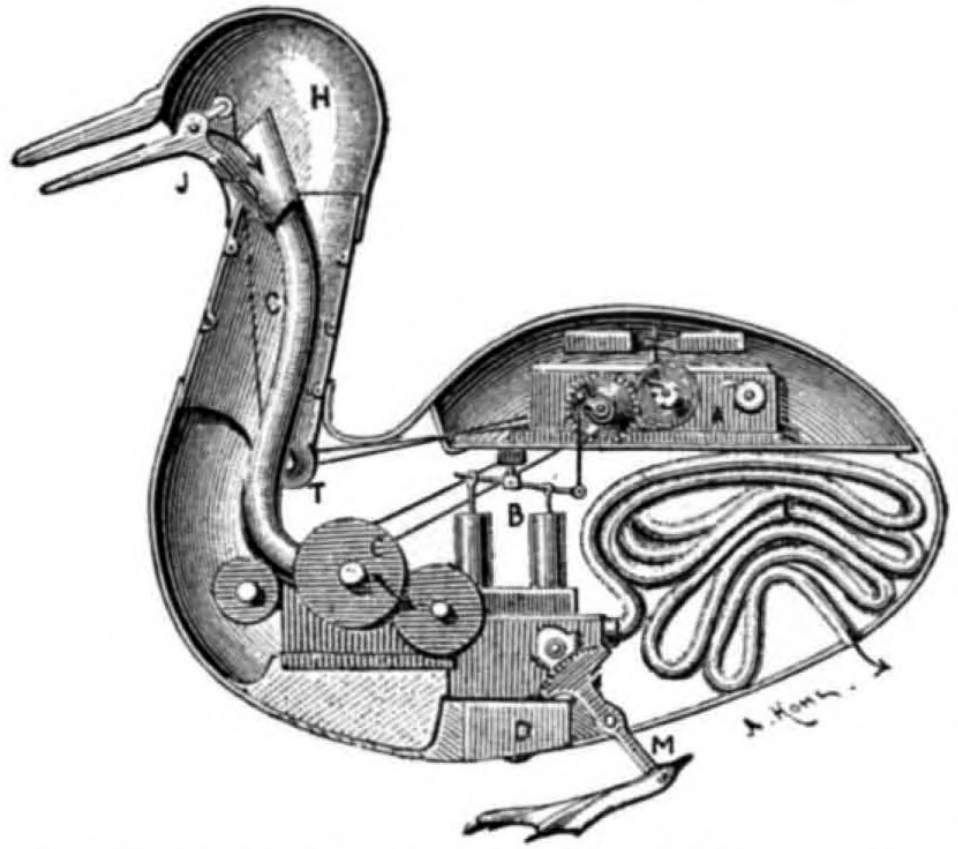
(SC) Spatial RNA-seq



Modified from @BoXia7, https://twitter.com/adj_23/status/1261928476811997184

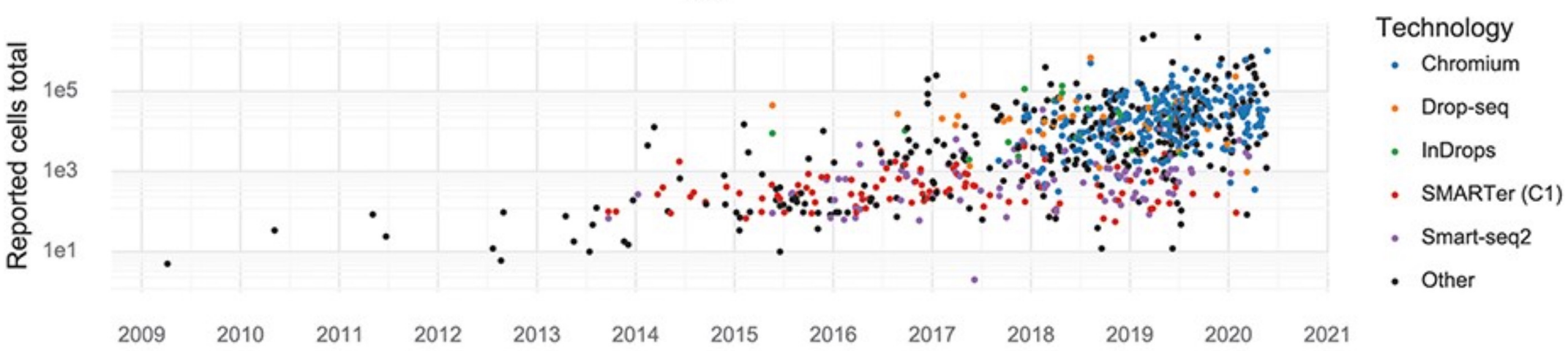
Bulk RNA-seq

- The workhorse for the last almost 20 years
- Cannot be understated how important it has been!
- Have pushed the reductionistic → integrative (holistic?) research paradigm
- Will continue to be relevant due to limitations of single-cell / spatial



(Spatial) Single Cell RNA-sequencing

- What everybody wants to do!
- Enables exciting insights into both normal and disease states
- Many limitations including much harder to analyze(!)
  - Only the 300-5000 highest expressed genes (fewer for spatial)
  - Very labor intensive and expensive!
- Is the subject of a the advanced master 22102 I’m running



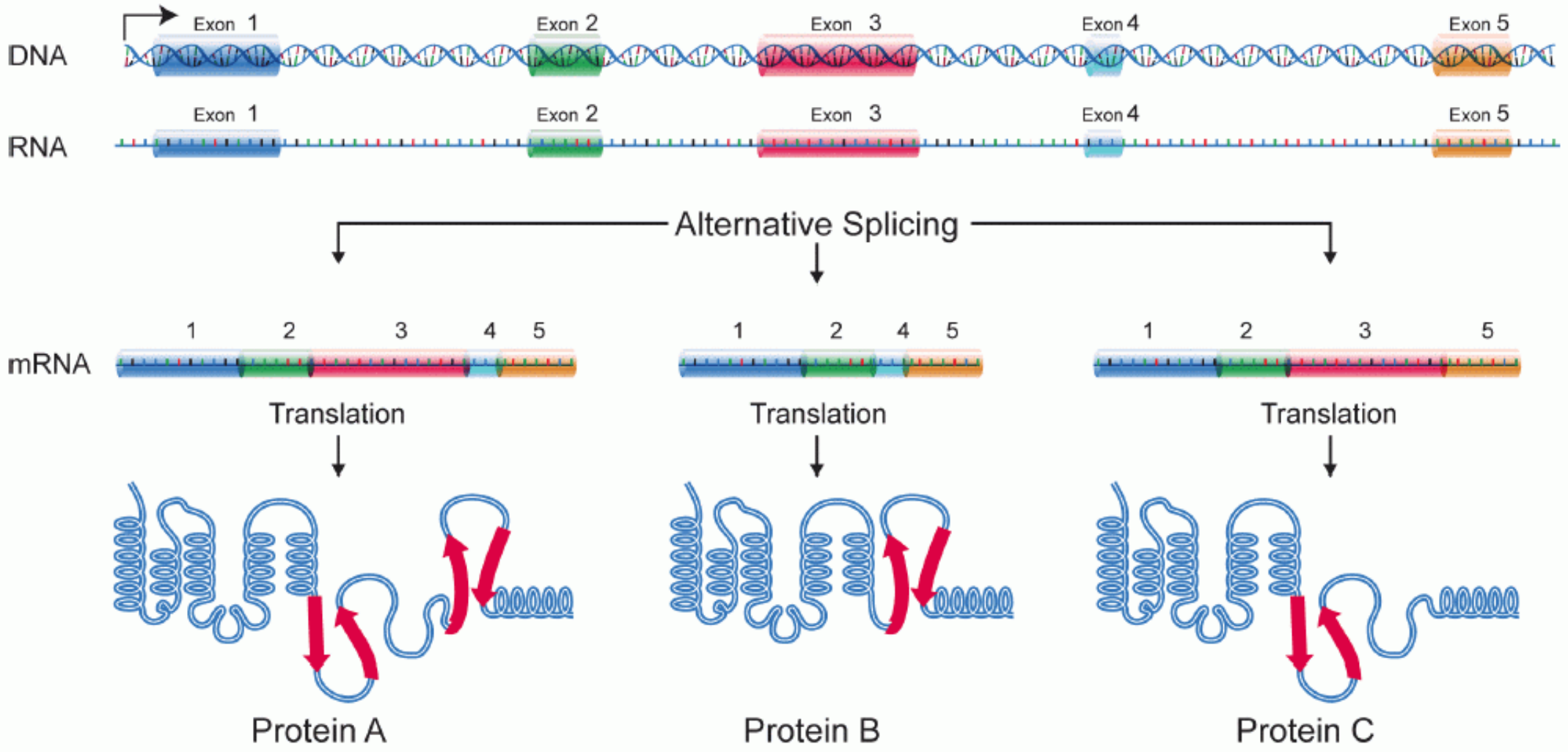
Single Cell Multi Omics

Joint analysis of multiple modalities (DNA, RNA, etc) pushes us towards a holistic research paradigm

Extention #2

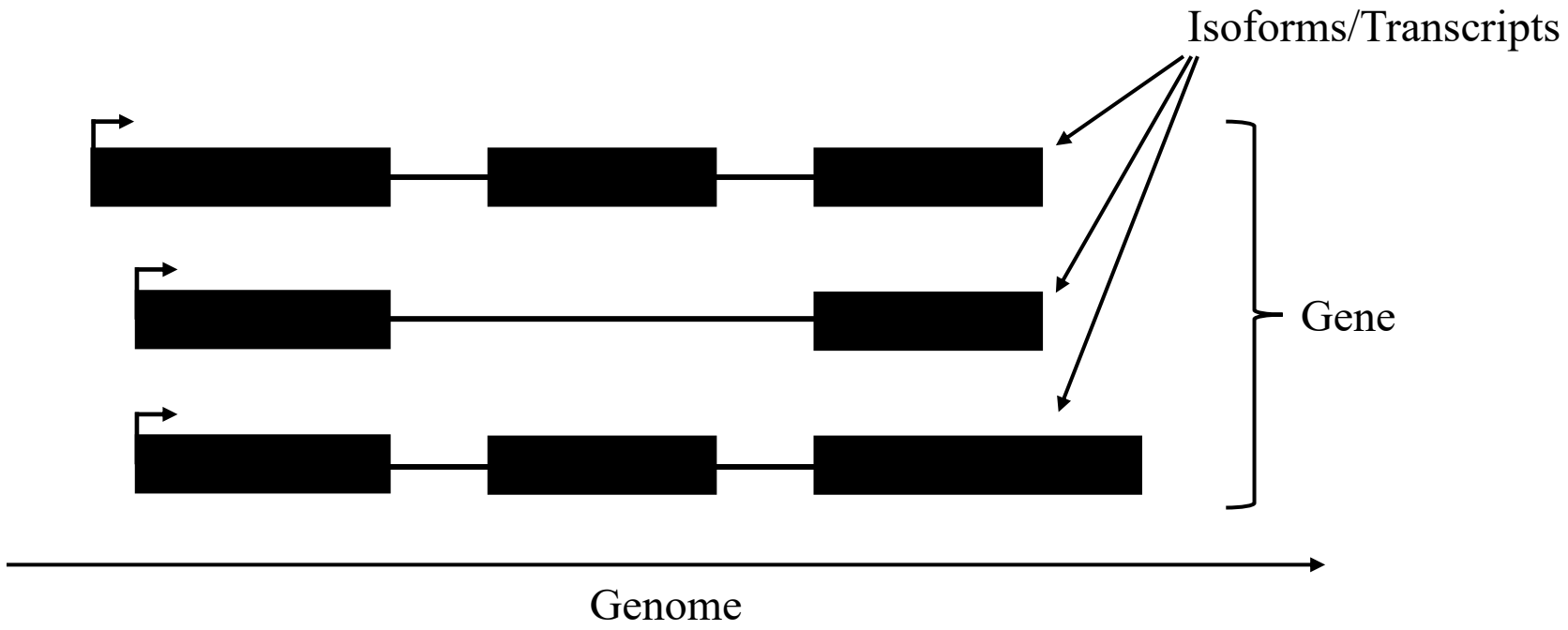
There is no such thing as a “gene”

Alternative Splicing



Genes vs Isoforms

The terms isoforms and transcripts are (unfortunately) used interchangeably



Isoforms/Transcripts can easily be quantified from RNA-seq data
via pseudo-alignment

Genes vs Isoforms

Take 2 min with your neighbor and discuss:
What would you gain by profiling the transcriptome with isoform resolution (instead of gene resolution)?

Genes vs Isoforms

Solution

Take 2 min with your neighbor and discuss:
What would you gain by profiling the transcriptome with isoform resolution (instead of gene resolution)?

A few answers:

- Alternative splicing
- Isoform Switches
  - Isoforms often have different functions
  - Sequence Analysis (e.g. protein domains)
- Improved Gene-level quantification

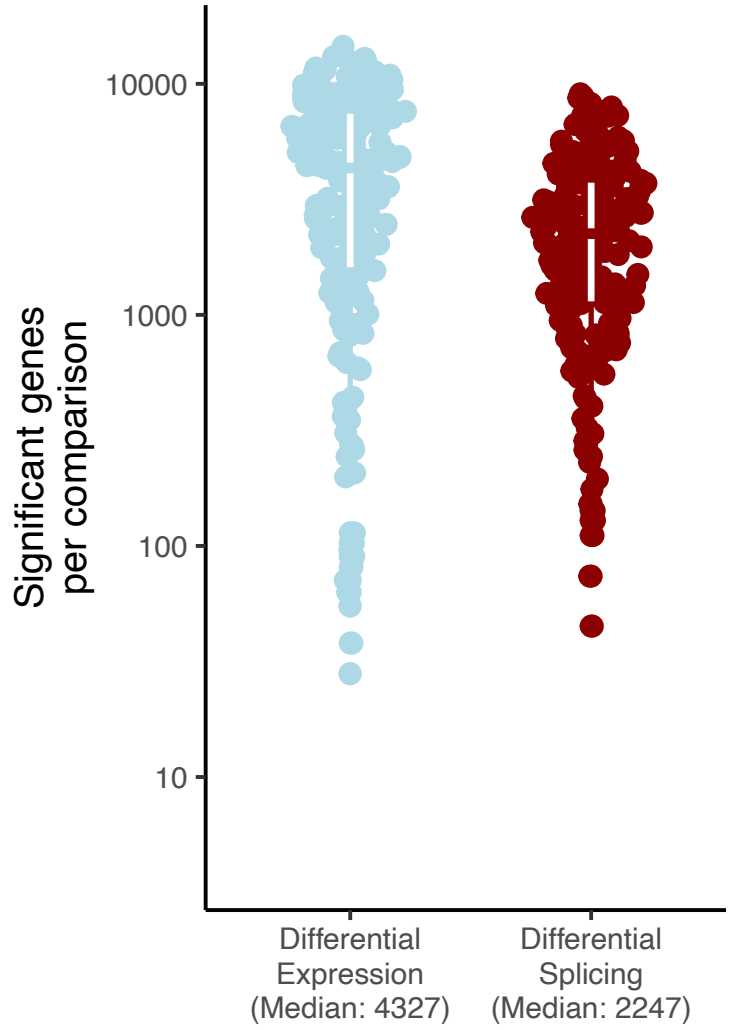
Analysis of isoforms

- Mostly called differential transcript usage (DTU)
- Good tools:
  - DEXSeq (in family with DESeq2)
  - limma
  - satuRn
- Can be done at two levels
  - Gene level: This gene have changes in isoform usage (unknown which)
  - Transcript level: This isoform has changed usage
- Long read RNA-seq is really useful (both PacBio and ONT)

Isoforms have different function

- Opposite effects in apoptosis
- Interact with different proteins
- Located different places in the cell
- P53

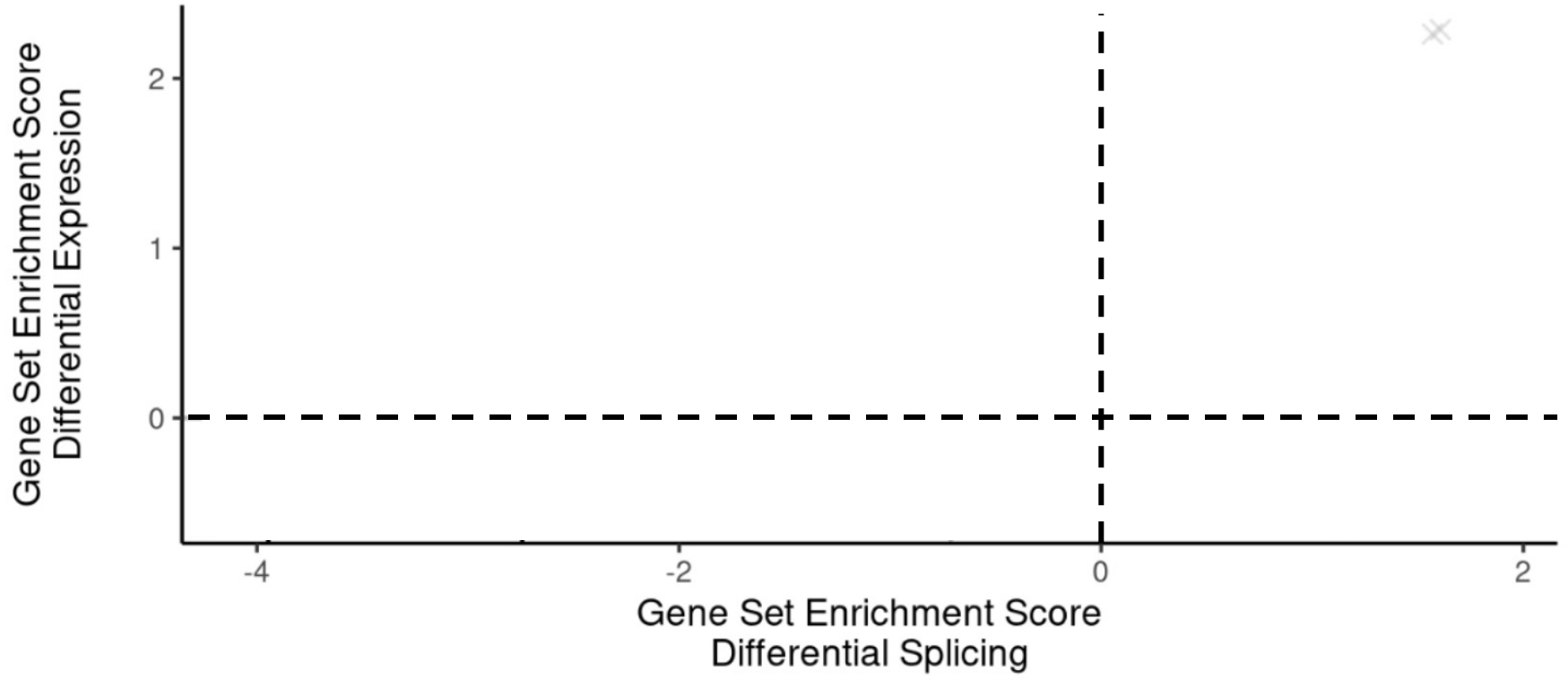
Differential Splicing is Omnipresent



93% of all multi-isoform genes

Dam *et. al.* 2022

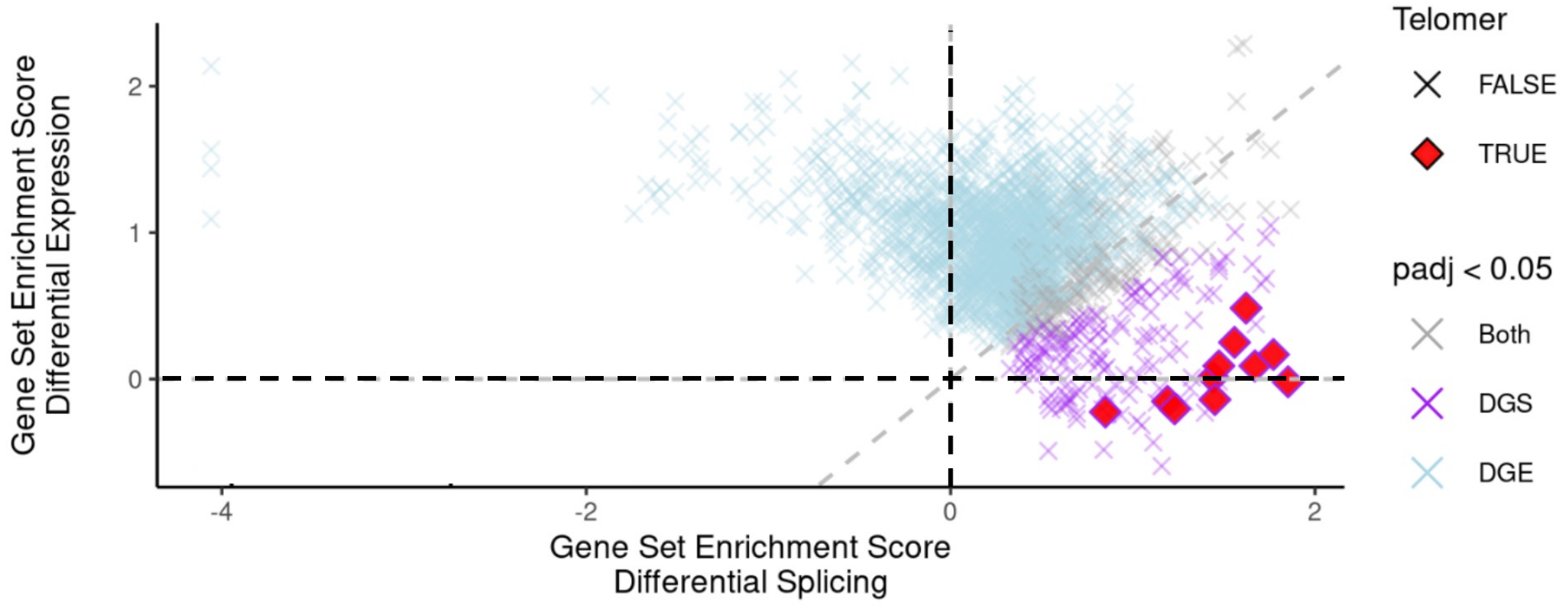
Splicing Mediate Distinct Biological Signals



pairedGSEA R package

Dam *et. al.* 2022

Splicing Mediate Distinct Biological Signals



pairedGSEA R package

Dam *et. al.* 2022

Isoforms are important

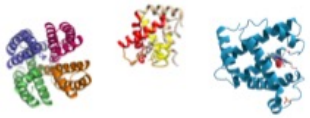
- For analysis of high-throughput data – including single cell analysis
- In clinical settings
  - Diagnosis
  - Treatment
- In genetics
- In most diseases – especially cancer

Exceptionally understudied!

Transcriptomics Enable high-level Integrative Analysis

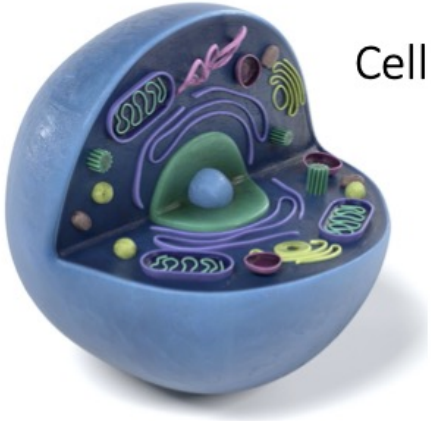
DNA

Transcripts



Proteins

Pathways



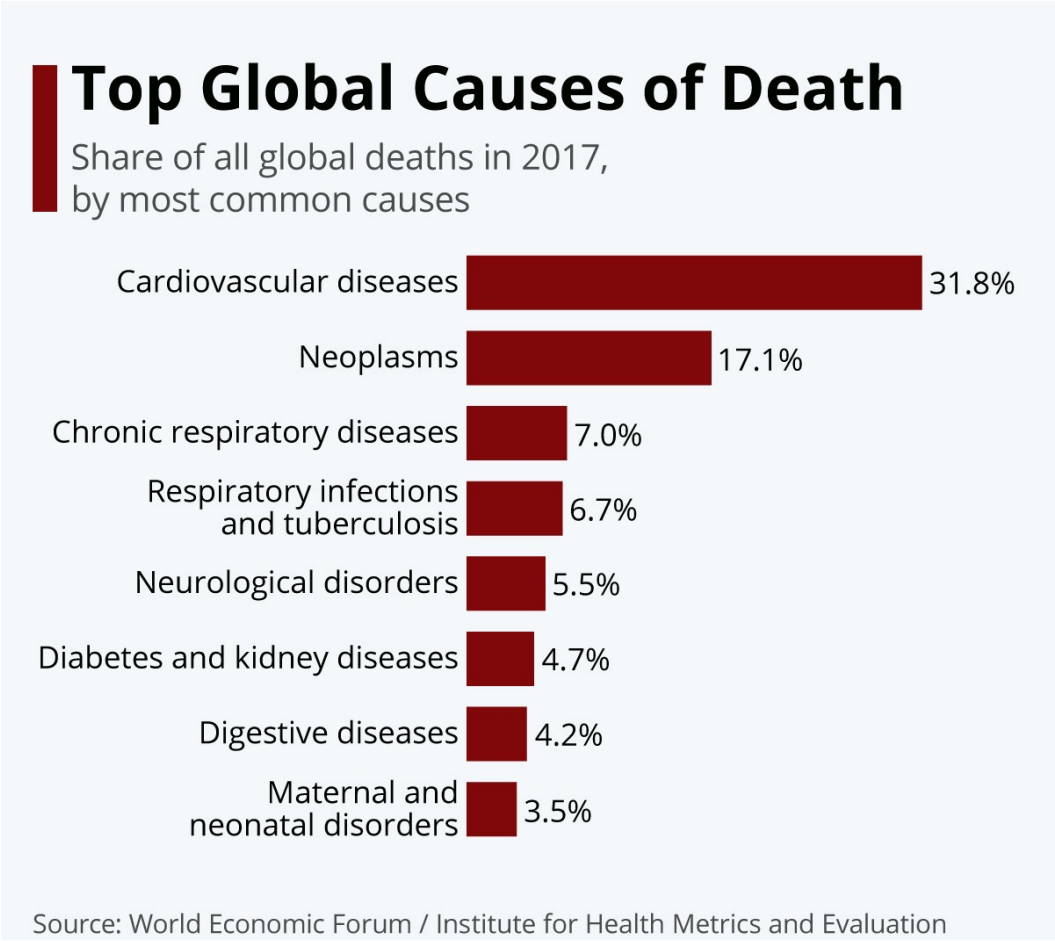
Cell

Cellular interactions

Tissue/Cancer organization

Functional phenotype

Original Slide by Lars Rønn Olsen



To treat most of these we need to understand the molecular mechanisms and how they are change by disease

Come work with me

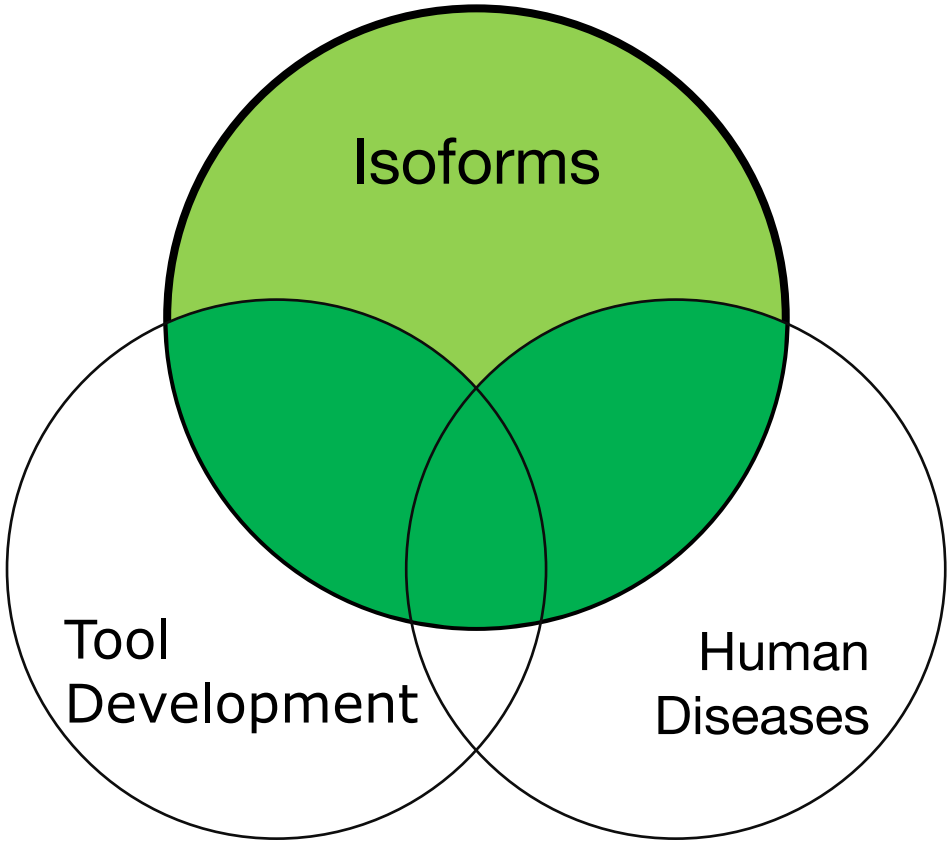
Projects

- BSc
- MSc (30-60 ETCS)
- Special Courses

Technologies

- RNA-seq (bulk/SC)
- Proteomics

I aim to *inspire* and *enable* everybody to study *isoforms*



Main take-aways

- Transcriptomics profile the RNA content of a cell
  - Bulk
  - Single Cell
  - Spatial
- The resulting count/expression matrix enables downstream analysis
  - Differential Expression
  - Gene set enrichment analysis
- Isoforms are important and overlooked
- Enable high-level Integrative Analysis

Assignment Time!