Extracting information from omics data

2pm Tuesday 9 April 2024; Data-driven approaches to understanding dementia

Charlotte Capitanchik, PhD
Postdoctoral Research Associate UK DRI @ KCL w/ Prof. Jernej Ule
Part-time Bioinformatics Lead @ Goodwright Ltd.

(Gabriel) Mateus Bernado Harrington, PhD
Postdoctoral Research Associate UK DRI @ Cardiff w/ Prof. Caleb Webber

King's College London

Cardiff University
Prifysgol Caerdydd

UK Dementia Research Institute

Outline

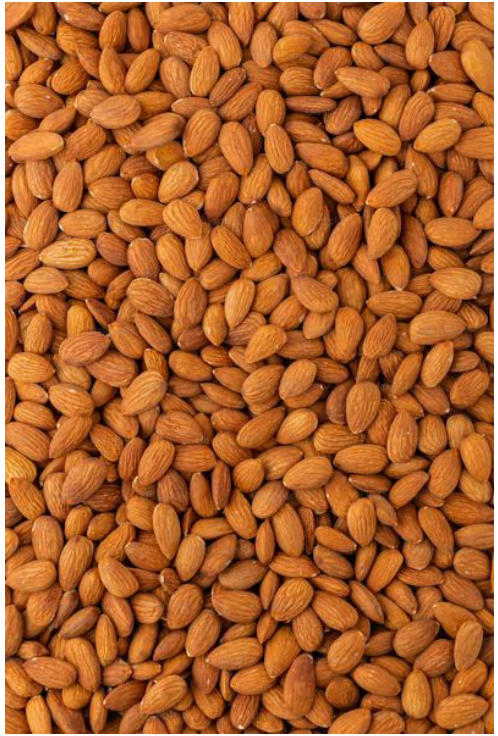
1. Overview of omics methods
2. Experimental design, Reproducibility
3. Pre-processing and pipelines
4. Data Visualisation
5. Multi-omic integration

All the Omics (well not all...)

1. Genomics
    a. Genotyping
    b. Whole genome seq
2. Transcriptomics
    a. Microarrays
    b. Bulk RNA-seq (short read or long read)
        i. Random primed/3' end sequencing/5' end sequencing
    c. Single Cell/Nuclei
    d. Metabolic sequencing (e.g. SLAM-seq)
    e. RNA-protein binding (e.g. iCLIP, TRIBE)
    f. RNA structure (e.g. PARIS2)
    g. RNA modifications (e.g. miCLIP, from direct RNA seq, DART-seq)
    h. Ribosome profiling
3. Epigenomics
    a. DNA-Protein binding/histone modifications (ChIP-Seq, CUT&Tag, CUT&RUN)
    b. ATAC-Seq (Assay for Transposase-Accessible Chromatin using Sequencing)
    c. Bisulfite sequencing for m5C

All the Omics continued (still not all…)

1. Proteomics
    a. Tandem Mass Tag
    b. Targeted/Shotgun
    c. phosphoproteomics
    d. Protein-protein interactions (AP-MS)
    e. RNA-protein interactions (RBPome, OOPs)
    f. DNA-protein interactions…
2. Metabolomics
3. Lipidomics

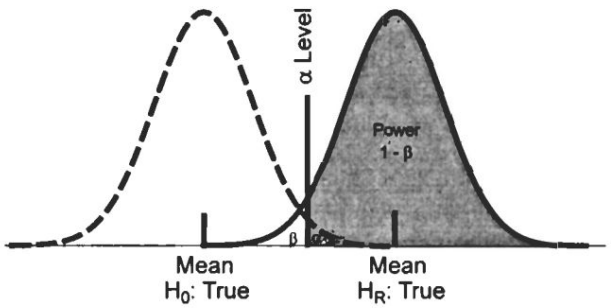


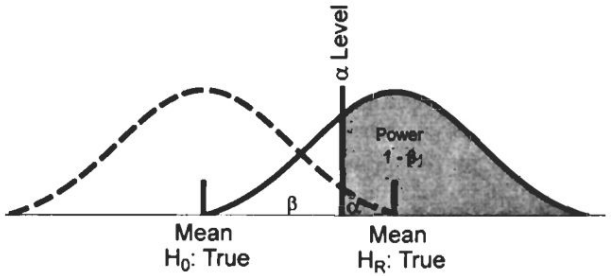
Experimental design - Controls

- This is a big topic worthy of careful consideration (and it’s worth remembering you won’t always have power to design the experiments you want to…), but one big point to highlight:
- **Controls**!
- Everyone always includes a negative control, but it’s shockingly rare to see a positive control, use them if you can!

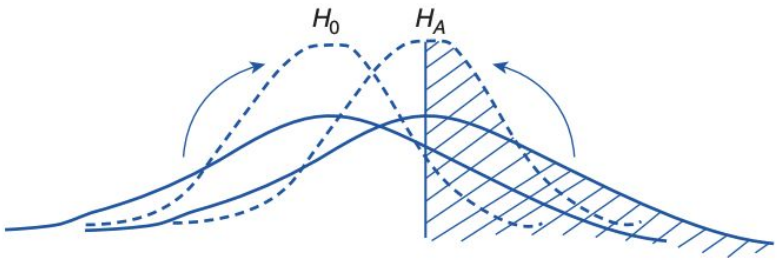
What impacts your ability to detect differences between conditions?

A.k.a. what variables affect statistical power?

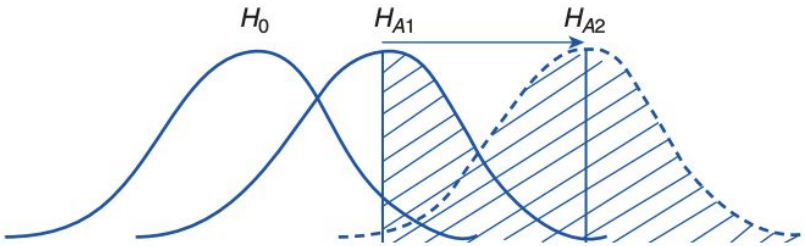




⬆️ Variance
⬇️ Power



⬆️ Sample size
⬆️ Power



⬆️ Effect size
⬆️ Power

Practical implications

- If your effect size is **big**, you will need less replicates (samples) to observe it
- If you have a lot of variability between samples you can increase your power by making more replicates (beware 'N-hacking', but probs not that big a deal in omics? See Reinagel 2023, PLoS Biology.)
- In omics we can also think of read depth as "sample size" - eg. if there is a small difference between DEG (effect size) you will need higher read depth (sample size) to reliably detect it…
- People often talk of "signal-to-noise ratio" = if variance is high, power is reduced; the impact is worse if effect size is small.

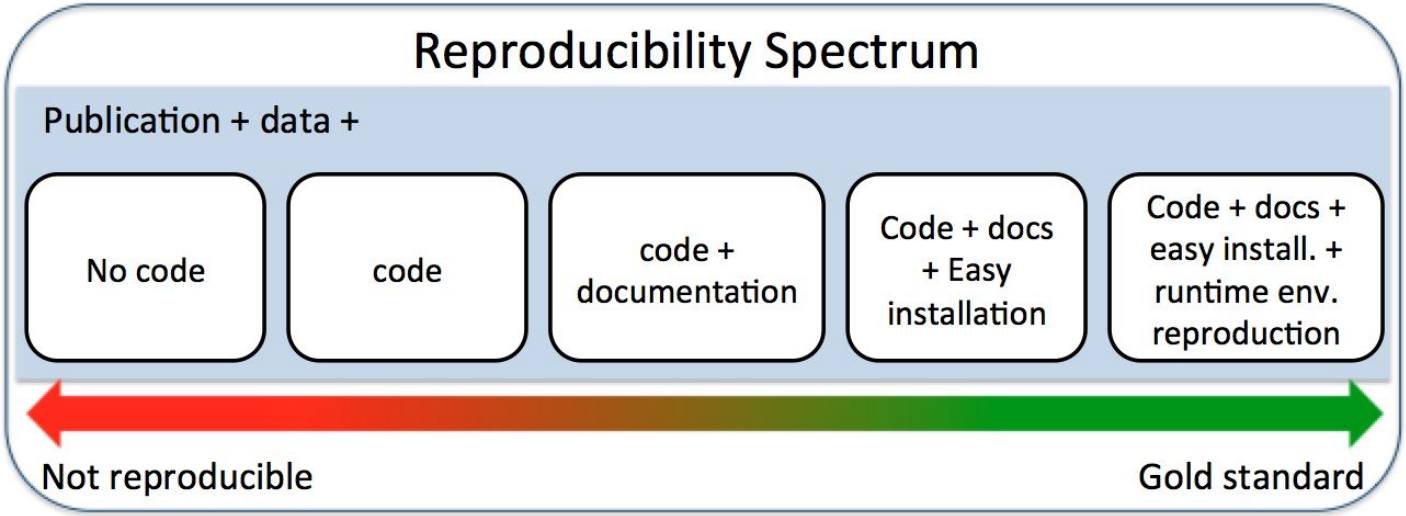
Reinagel, Pamela. "Is N-Hacking Ever OK? The consequences of collecting more data in pursuit of statistical significance." *PLoS Biology* 21.11 (2023): e3002345.

Reproducibility

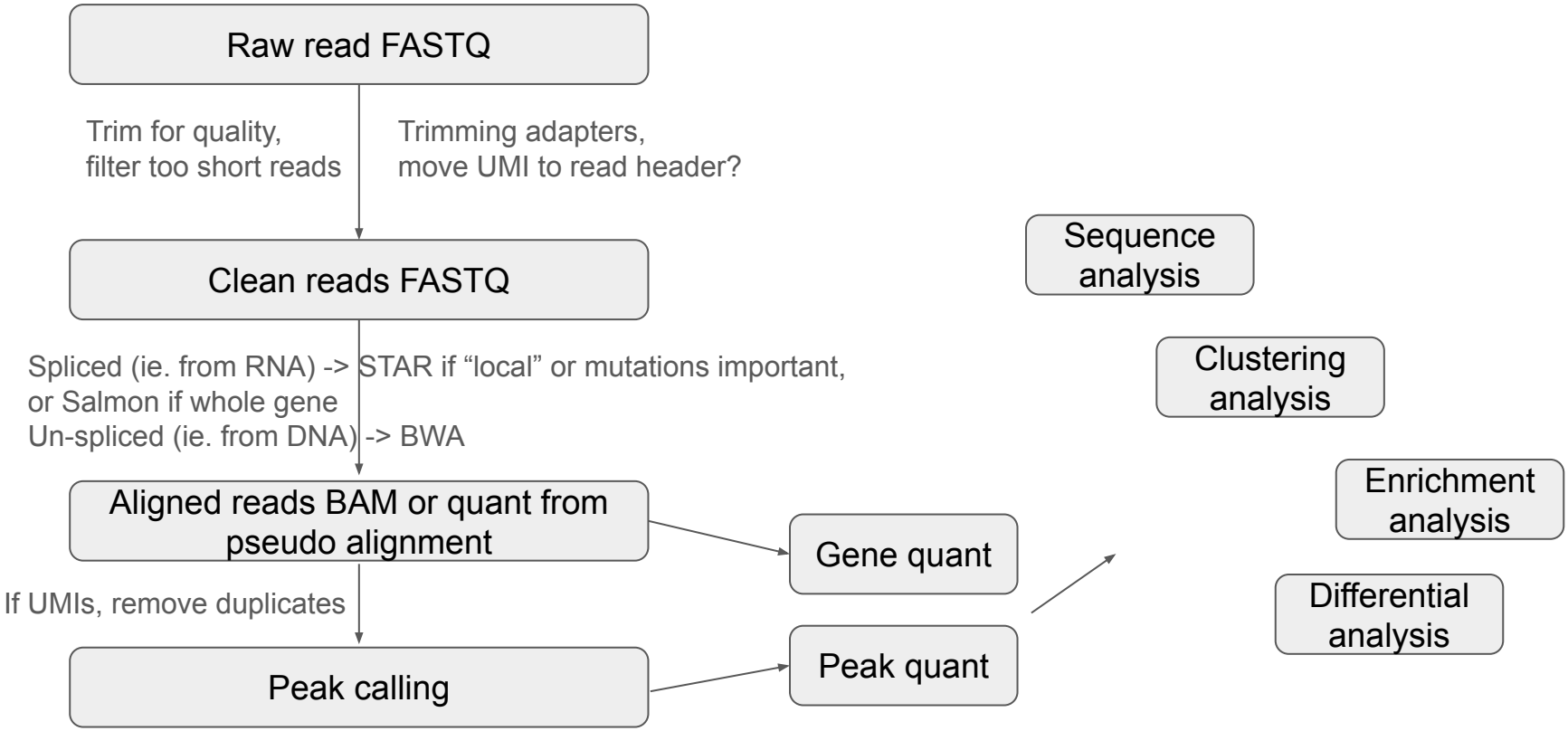
- Reproducibility isn't just about lofty ideals of doing science that isn't a total load of rubbish
- If someone asked you to reproduce a figure you made 6 months ago from scratch, how easily could you do it?
- If you solved a niche problem 2 years ago and need to solve it again, how easily could you find and redeploy your solution?

Reproducibility

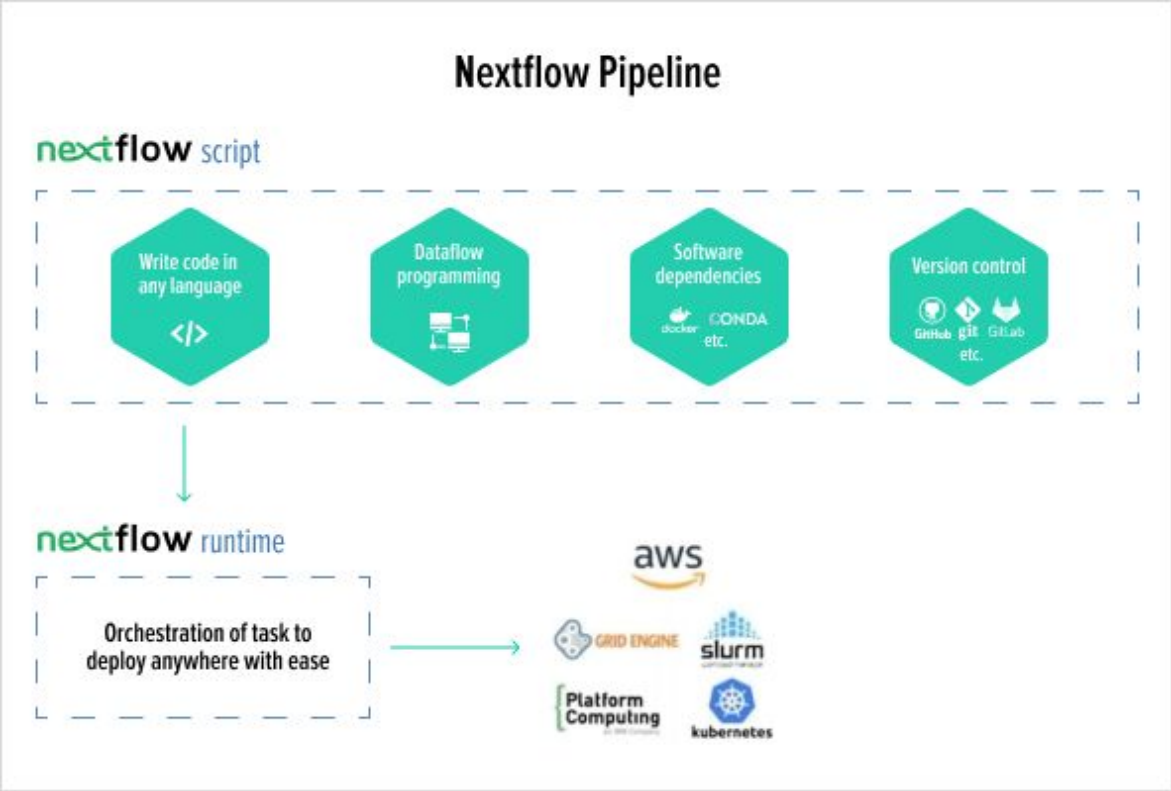
- Do future you a solid, and make sure your work can be reproduced, at least by you, ideally by others too!
- The Turing Way is a great guide to reproducibility if you want to learn more



Analysing xyz-Seq



Nextflow is a workflow language



Nf-core has lots of pipelines and modules ready to go

nf-core

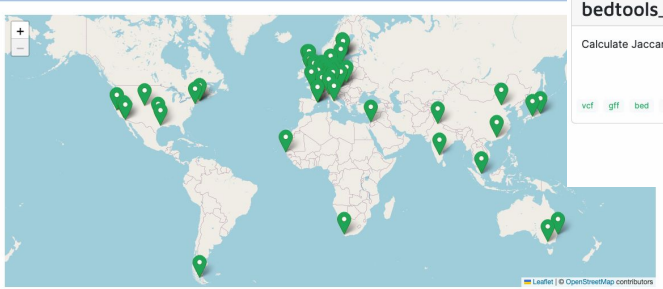
- 92 pipelines
- 1058 modules

Organisations

Some of the organisations running nf-core pipelines are listed below, along with a key person who you can contact for advice.

Note

Is your group missing? Please submit a pull request to add yourself! It's just a few lines in a simple YAML file.



Home | Pipelines | Resources | Docs | Events | About | Search | Join nf-core

nf-core/modules

Browse the 1201 modules that are currently available as part of nf-core.

bedtools | Name

bedtools_bamtobed

Converts a bam file to a bed12 file.

bam bed bedtools bamtobed converter

Included in: cutandrun radseq ssds

bedtools_closest

For each feature in A, finds the closest feature (upstream or downstream) in B.

bedtools closest bed vcf gff

bedtools_complement

Returns all intervals in a genome that are not covered by at least one interval in the input BED/GFF/VCF file.

bed gff vcf complement bedtools intervals

Included in: cutandrun

bedtools_coverage

computes both the depth and breadth of coverage of features in file B on the features in file A

bedtools coverage bam bed gff vcf histogram

Included in: radseq

bedtools_genomecov

Computes histograms (default), per-base reports (-d) and BEDGRAPH (-bg) summaries of feature coverage (e.g., aligned sequences) for a given genome.

bed bam genomecov bedtools histogram

Included in: cutandrun hicar nascent +1 more pipelines

bedtools_getfasta

extract sequences in a FASTA file based on intervals defined in a feature file.

bed fasta getfasta

Included in: readsimulator viralrecon

bedtools_groupby

Groups features in a BED file by given column(s) and computes summary statistics for each group to another column.

bed groupby bedtools

bedtools_intersect

Allows one to screen for overlaps between two sets of genomic features.

bed intersect overlap

Included in: circrna cutandrun nascent +1 more pipelines

bedtools_jaccard

Calculate Jaccard statistic b/w two feature files.

vcf gff bed jaccard intersection union statistics

bedtools_makewindows

Makes adjacent or sliding windows across a genome or BED file.

bed windows fai chunking

Included in: phaseimpute radseq ssds

bedtools_map

Allows one to screen for overlaps between two sets of genomic features.

bed vcf gff map bedtools

bedtools_maskfasta

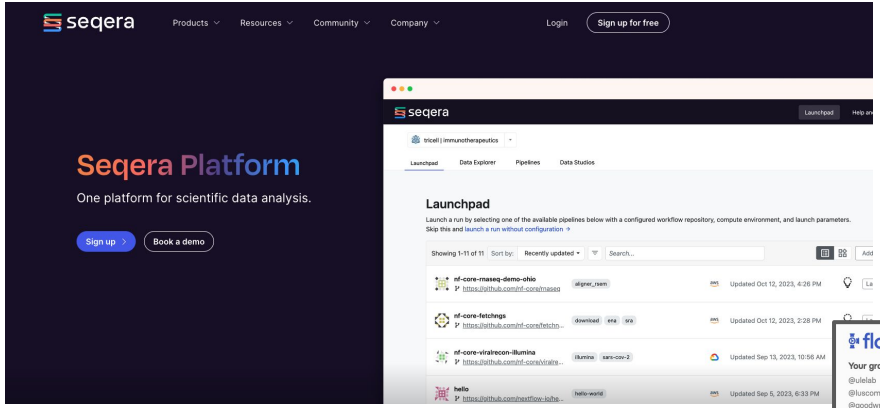
masks sequences in a FASTA file based on intervals defined in a feature file.

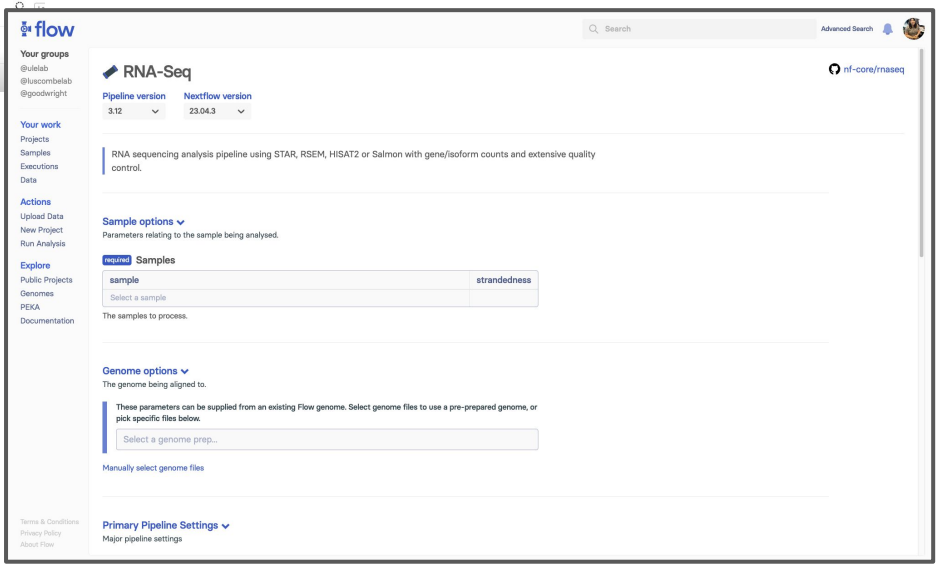
bed fasta maskfasta

Included in: callingcards viralrecon

Previous 1 2 Next

Nextflow pipelines can also be run through a GUI





app.flow.bio

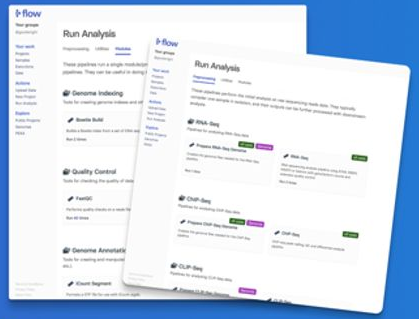
flow A Nextflow-based bioinformatics analysis platform and open database

1. Upload your experimental data
2. Automatically store standardised data
3. Select parameters and run a pipeline in 1-click
4. Get your research insight and visualisations
5. Add your results to a growing public database

Collate and view analysis history

View input/output files of Nextflow analyses in real time, along with curated data outputs

Run Nextflow pipelines down to individual modules (+ add custom pipelines)



View terminal history, stdout, and Nextflow logs

View analysis metadata and parameter history

Try now at flow.bio

No programming required - run Nextflow pipelines in the browser at the click of a button

goodwright software for scientists

King's College London

UK Dementia Research Institute

UCL

The Francis Crick Institute

Batch correction

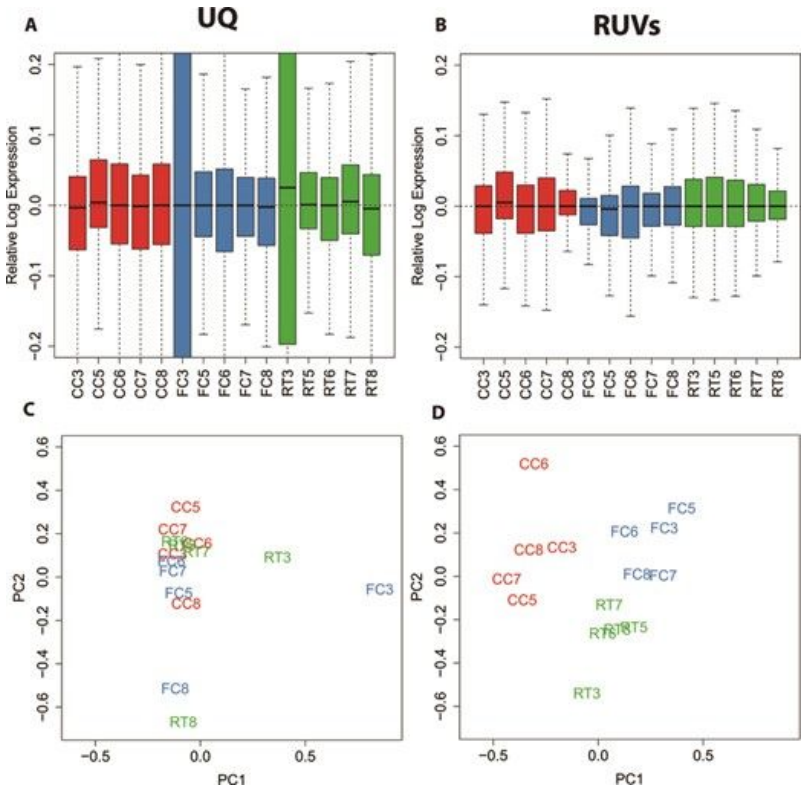
- Omics data from brains is notoriously variable
- Data can be confounded by:
  - known variables, eg. time of day, time to autopsy, RIN, fixation, differences in dissection, laboratory, experimental batch, batch of library preparation
  - unknown variables eg. any of the above that weren’t measured/recorded/communicated to you, GREMLINS
  - Dropout - eg. loss of rare cell types due to sample prep

Batch correction

It is always best to try and reduce variation experimentally

BUT

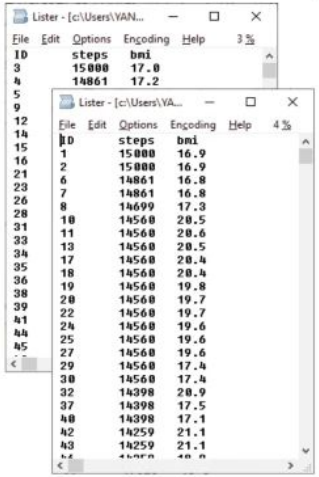
If you can't it is possible to remove known and unknown confounders with statistical modelling 🌈



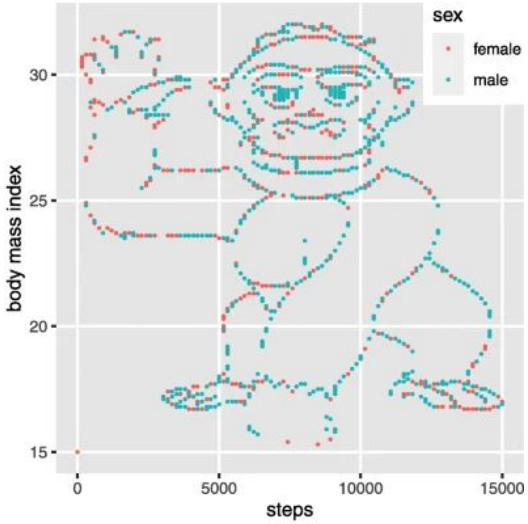
Peixoto, Lucia, et al. "How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets." *Nucleic acids research* 43.16 (2015): 7664-7674.

Always visualise your data

a



b

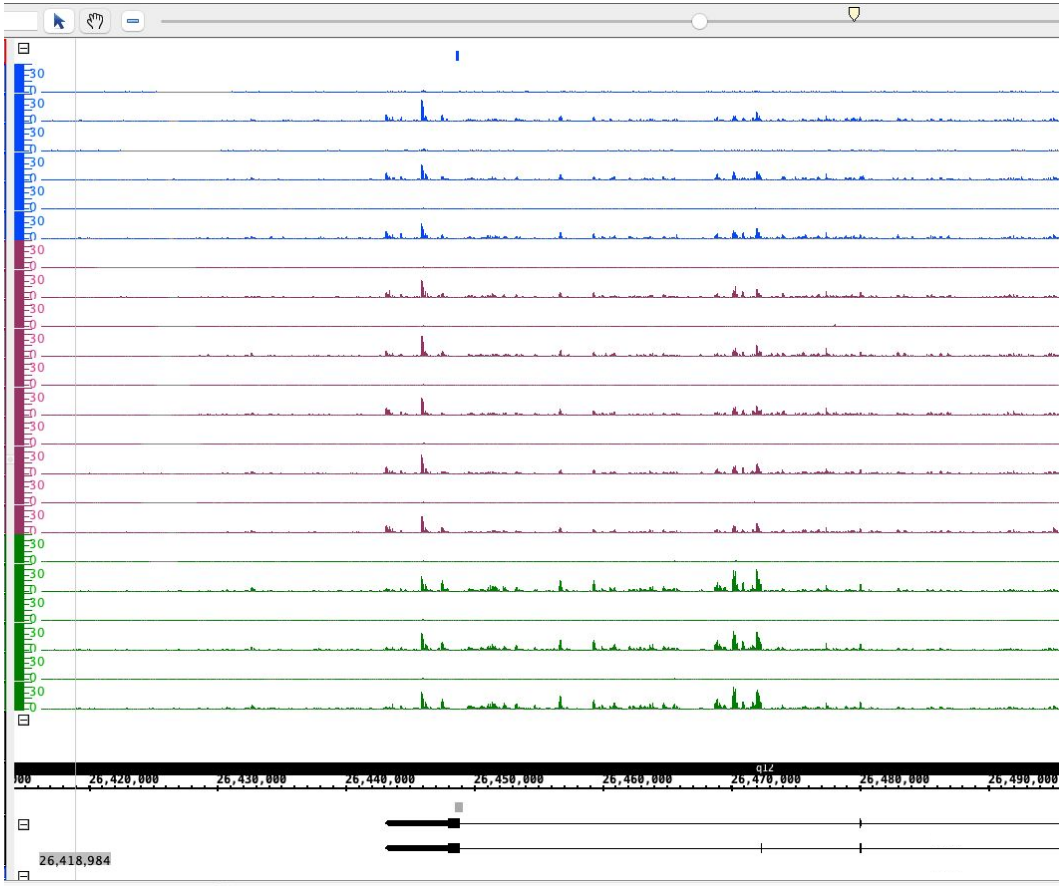


c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

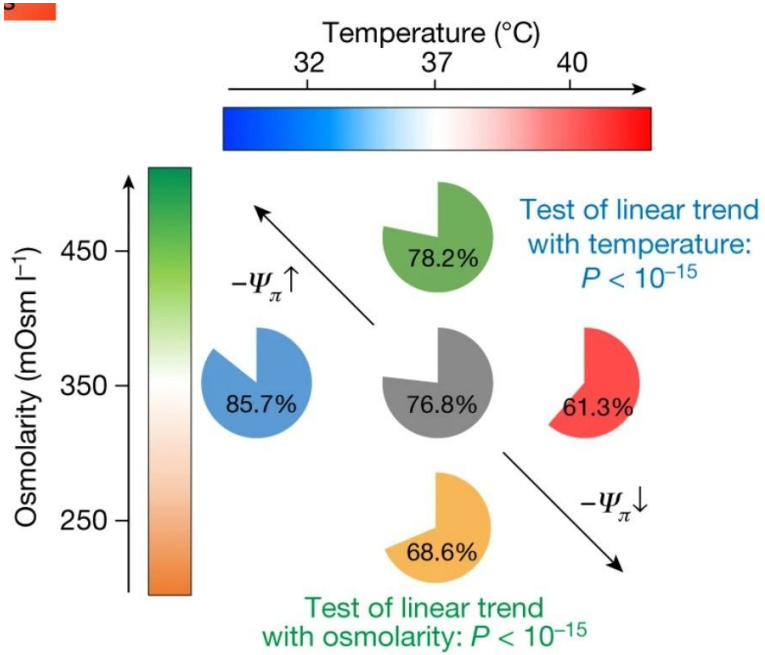
Yanai, Itai, and Martin Lercher. "A hypothesis is a liability." *Genome biology* 21 (2020): 1-5.

Always visualise your data



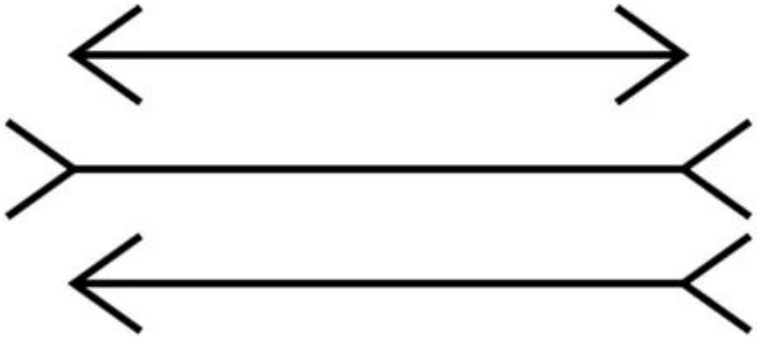
Responsible data visualisation

- With visualisation we aim to be: **clear, accurate & attractive** - in practise difficult to be all three!



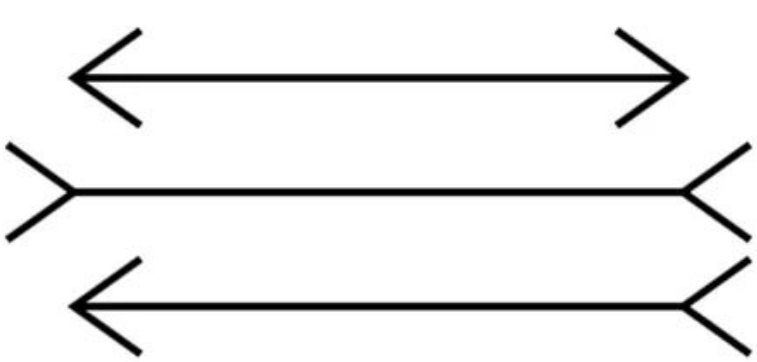
Responsible data visualisation

- With visualisation we aim to be: **clear, accurate & attractive** - in practise difficult to be all three!
- It's important to understand that visualisation is **rarely neutral**
- Humans have **biases in perception**, hence optical illusions!

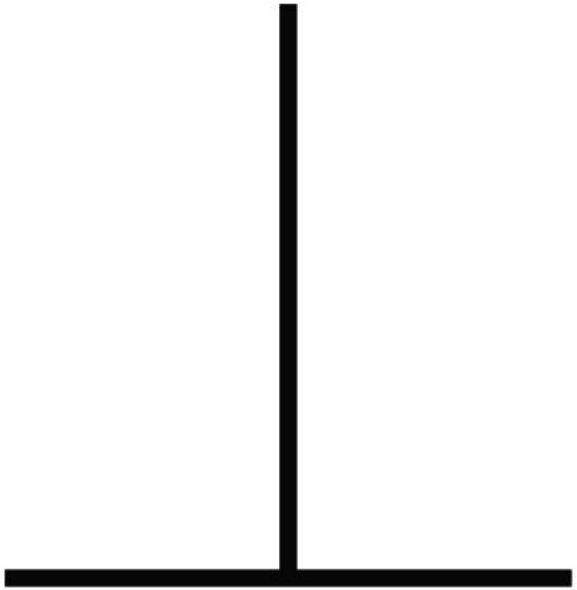


Responsible data visualisation

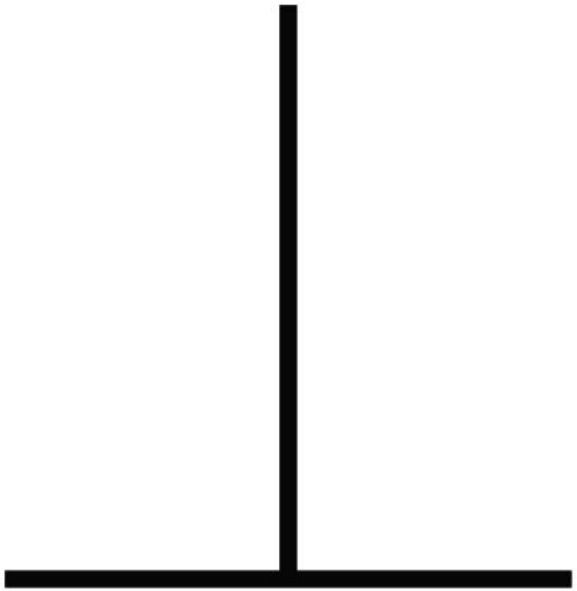
- With visualisation we aim to be: **clear, accurate & attractive** - in practise difficult to be all three!
- It's important to understand that visualisation is **rarely neutral**
- Humans have **biases in perception**, hence optical illusions!

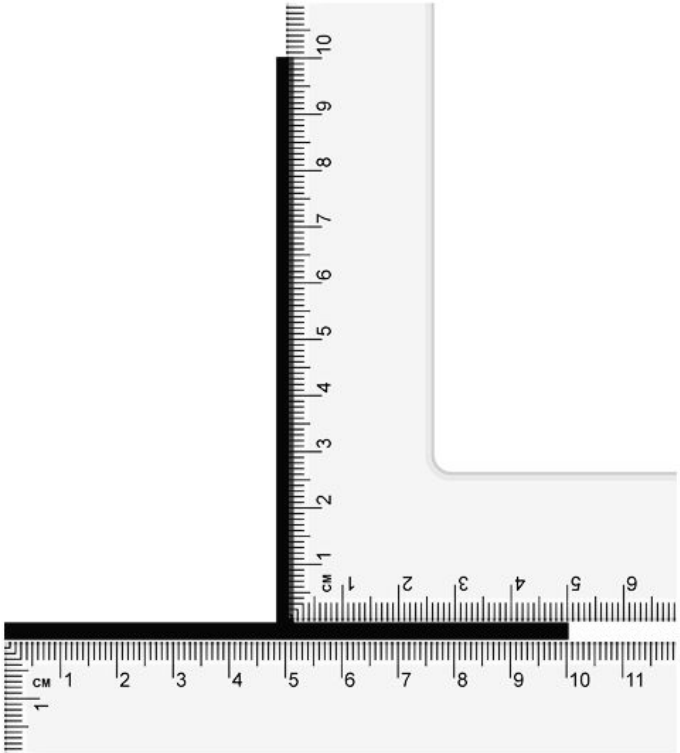


Responsible data visualisation



Responsible data visualisation





Responsible data visualisation

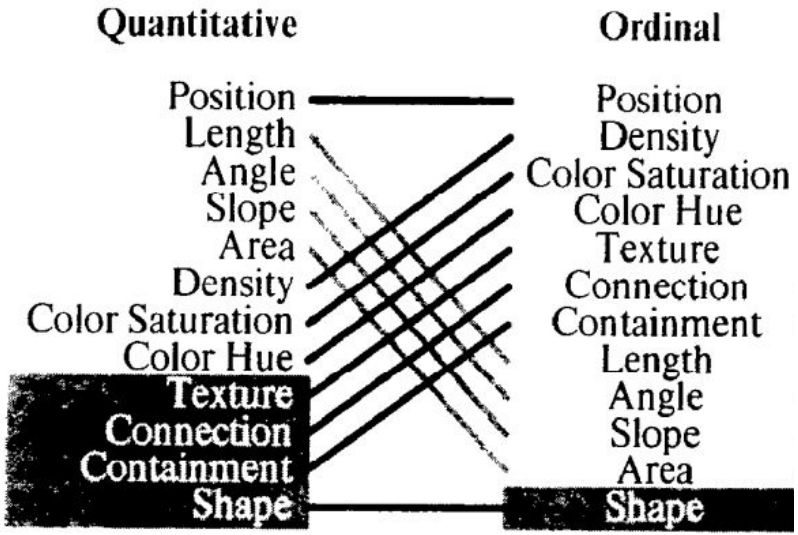
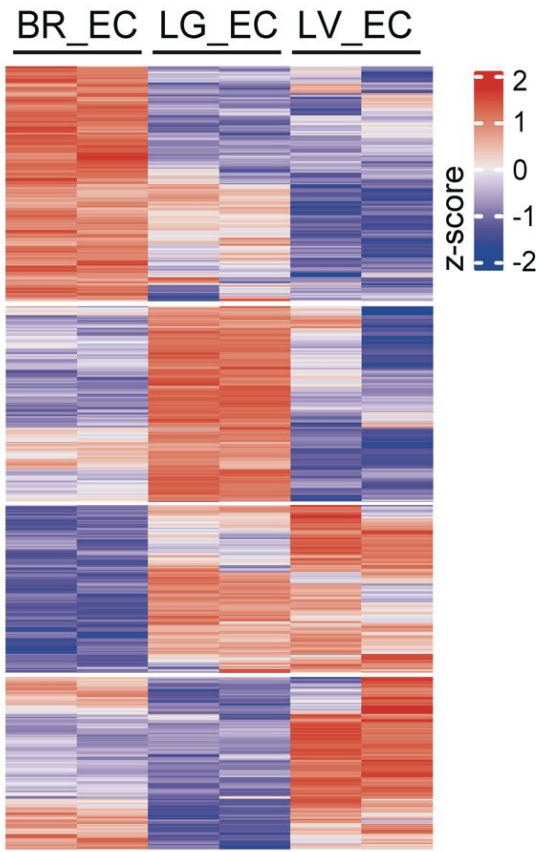
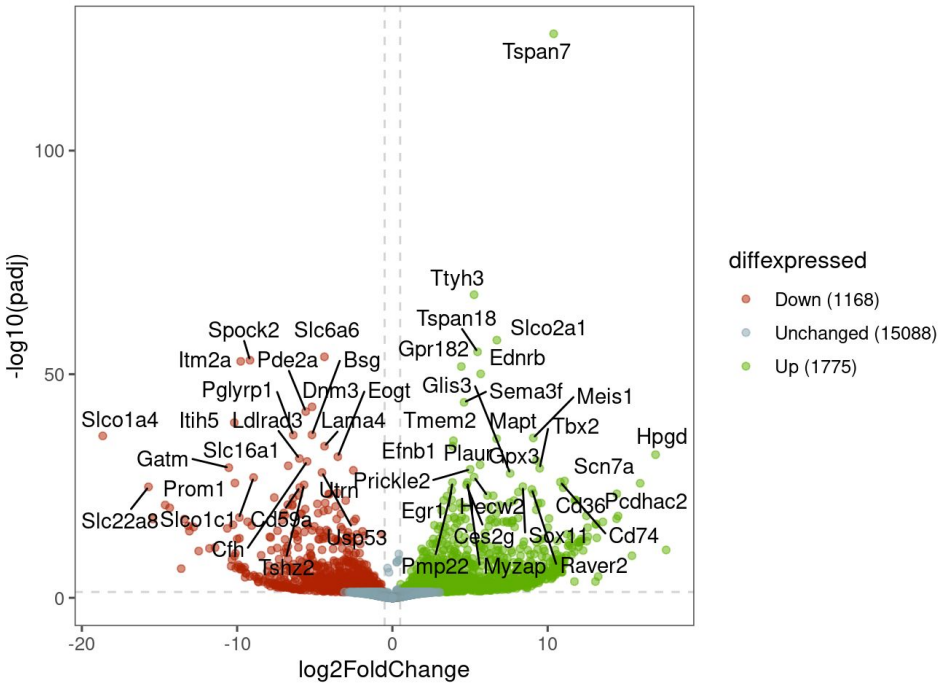


Figure 9. Ranking of perceptual tasks. The columns are for three different types of information. Tasks higher in the chart are perceived more accurately than tasks lower in the chart. The tasks shown in gray are not relevant to that type of information.

Mackinlay, Jock. "Applying a theory of graphical presentation to the graphic design of user interfaces." *Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software*. 1988.

Heatmap vs. volcano plot for gene expression

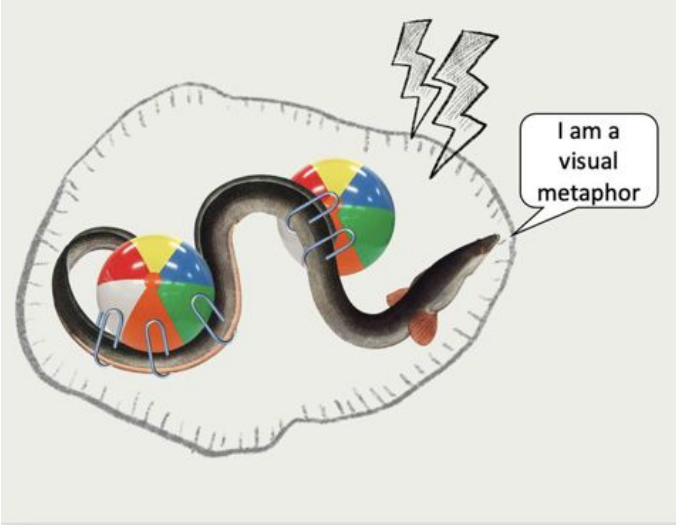




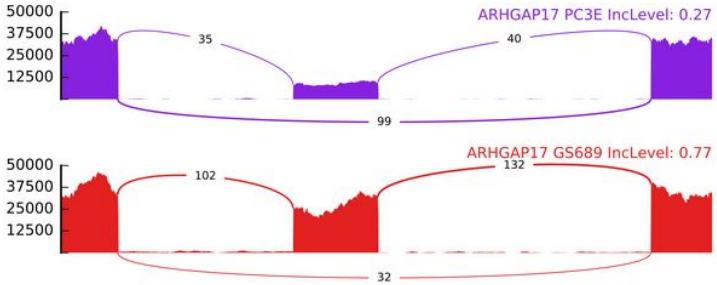
Multi-omic Data Integration

Ah yes, that thing I put in my grant…

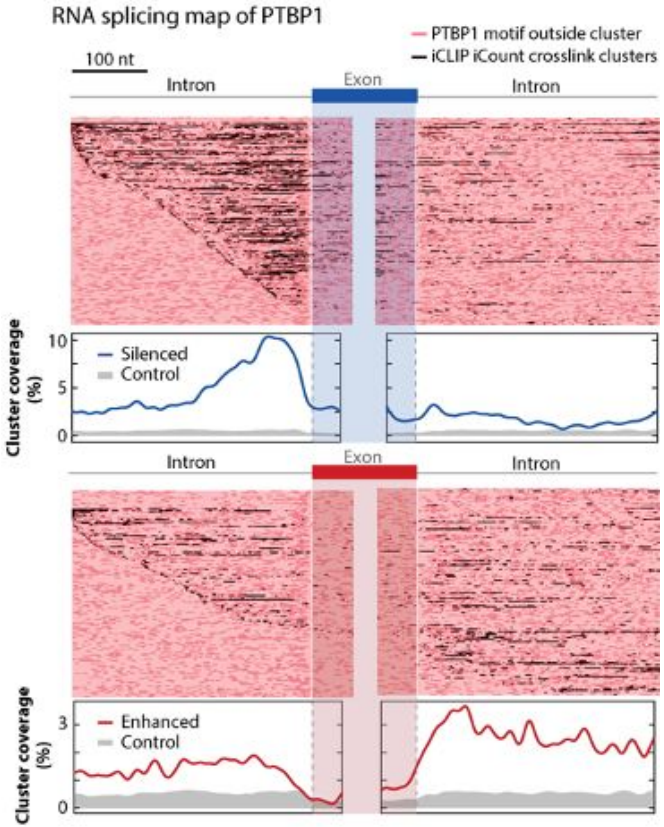
One example of integration of iCLIP and RNA-Seq…



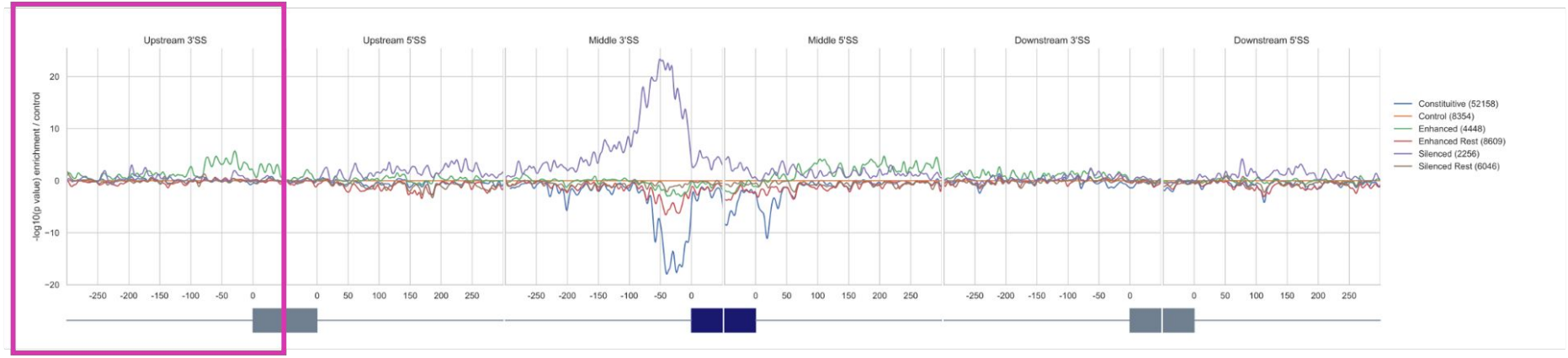
+



RNA splicing map



RNA splicing map



Ptbp1 RNA map

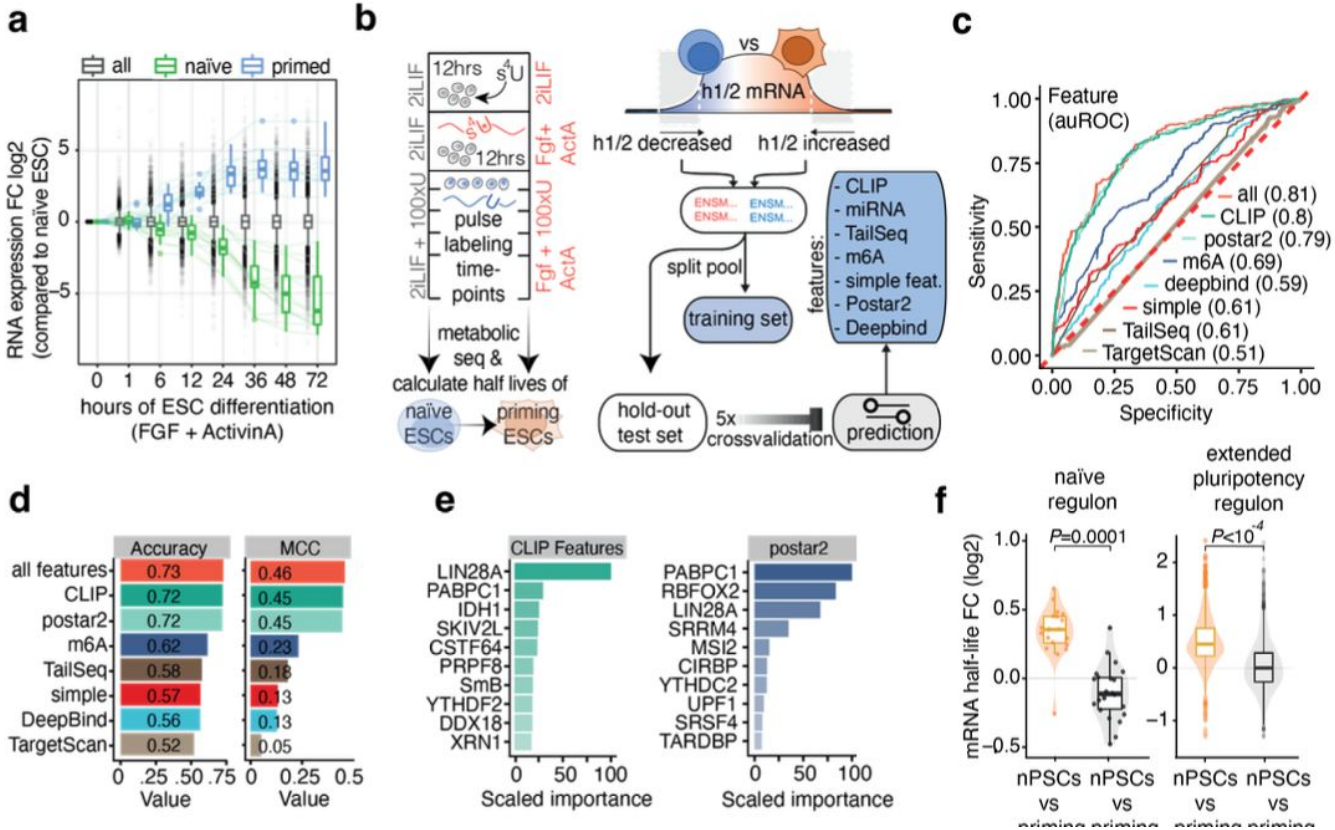
Modelling omics data

1. Chuck in all your data as features and try to predict something e.g. SVM, random forest, neural net…, can identify most predictive features.

2. Using only sequence as input can you predict something measured by omics data eg. gene expression, mRNA half-life, ribosome occupancy…, use backpropogation methods to figure out which sequences were important.

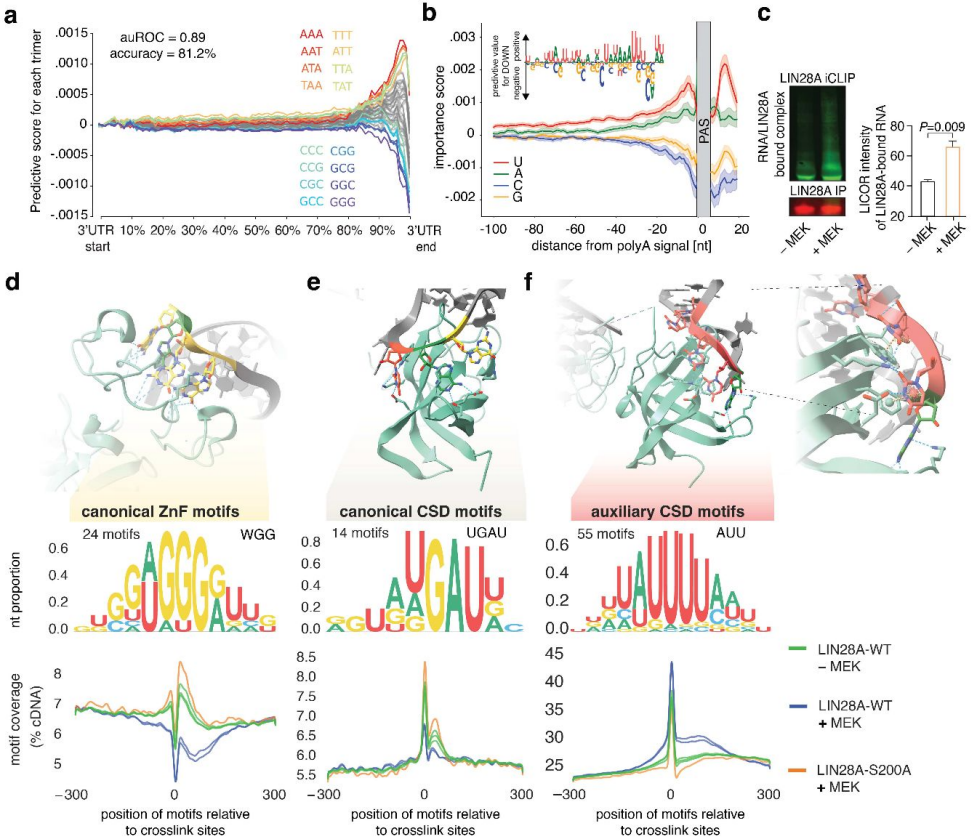
3. “Foundation models” training deep learning models to predict omics data tracks, use backpropogation methods to figure out which sequences were important or attention layer of transformer models.

Modelling omics data (1)



Modic, Miha , et al. "Poised PABP-RNA hubs implement signal-dependent mRNA decay in development." (2023).

Modelling omics data (2)



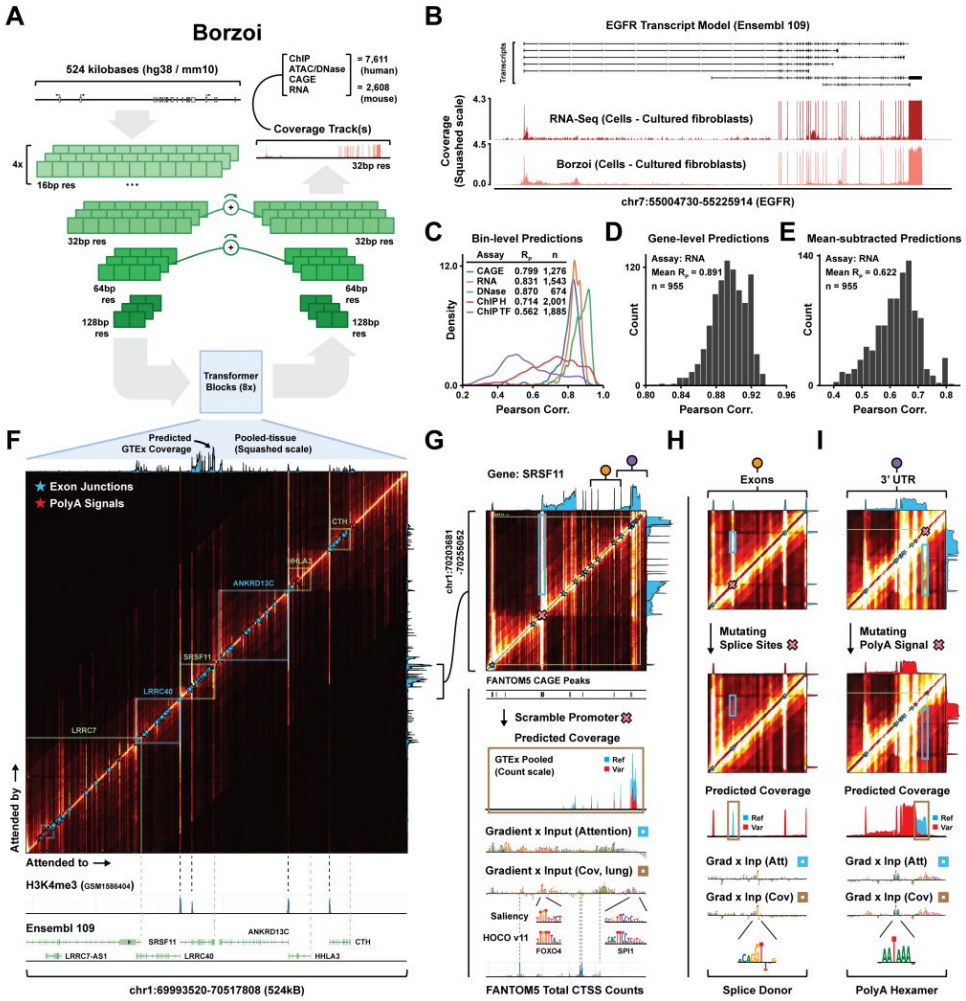
Modic, Miha , et al. "Poised PABP-RNA hubs implement signal-dependent mRNA decay in development." (2023).

Modelling omics data (3)

Borzoi models are convolutional neural networks trained to predict RNA-seq coverage at 32bp resolution given 524kb input sequences

GTEx + ENCODE

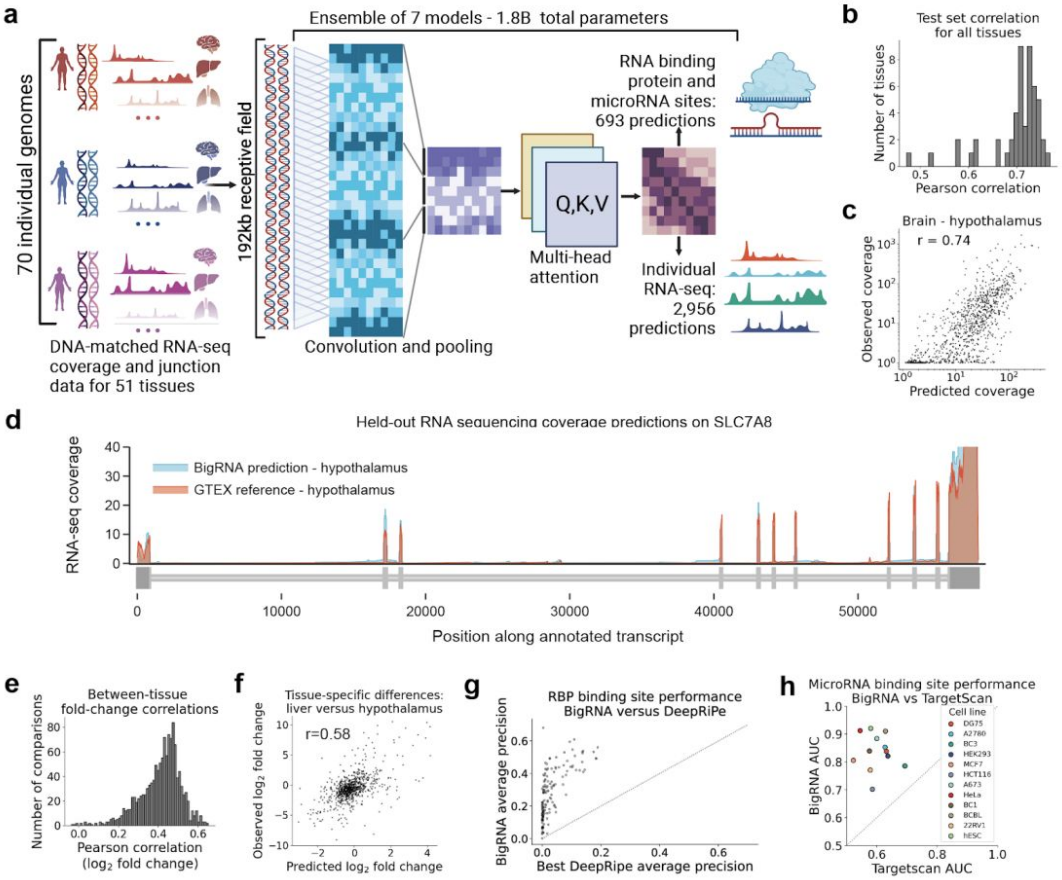
https://github.com/calico/borzoi



Linder, Johannes, et al. "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation." *bioRxiv* (2023): 2023-08.

Modelling omics data (3)

BigRNA
GTEx



Celaj, Albi, et al. "An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics." *bioRxiv* (2023): 2023-09..

Thank you for listening

See you in 30 minutes for the practical session 💪🐦

Ensure you have a GitHub account, you will need it!