knowledge AI in bio-data analytics

Nicos Angelopoulos

The Pirbright Institute
Head of Computational Biology
& co-lead of bioinformatics

https://stoics.org.uk/~nicos
nicos@stoics.org.uk

24.05.29

overview

- Stream 1. prior knowledge for Bayesian machine learning
- Stream 2. applied knowledge representation for biological big data analytics
- Stream 3. Bayesian networks for cancer and biological datasets
- Stream 4. single cell RNA for pig immunity

Stream 1. (York) Bayesian machine learning

How can we incorporate existing (biological) knowledge in the analysis of new experimental data

Bayesian methods allow for the incorporation of prior knowledge and expectations, although often applications use agnostic priors

Bayesian machine learning theory

Bayes' Theorem

$$p(M|D) = \frac{p(D|M)p(M)}{\sum_M p(D|M)p(M)}$$

Metropolis-Hastings

$$\alpha(M_i, M_*) = min\left\{\frac{q(M_*, M_i)P(D|M_*)P(M_*)}{q(M_i, M_*)P(D|M_i)P(M_i)}, 1\right\}$$

Stream 1. (York) Bims

A probabilistic programming framework for Bayesian machine learning of structured statistical models (classification trees and Bayesian networks).

Allows the encoding of prior information in the form of a probabilistic logic program.

- Theory (York, 2000-5, KR paper 2017)
- Applications (Edinburgh, 2006-8, IAH 2009, NKI 2013)

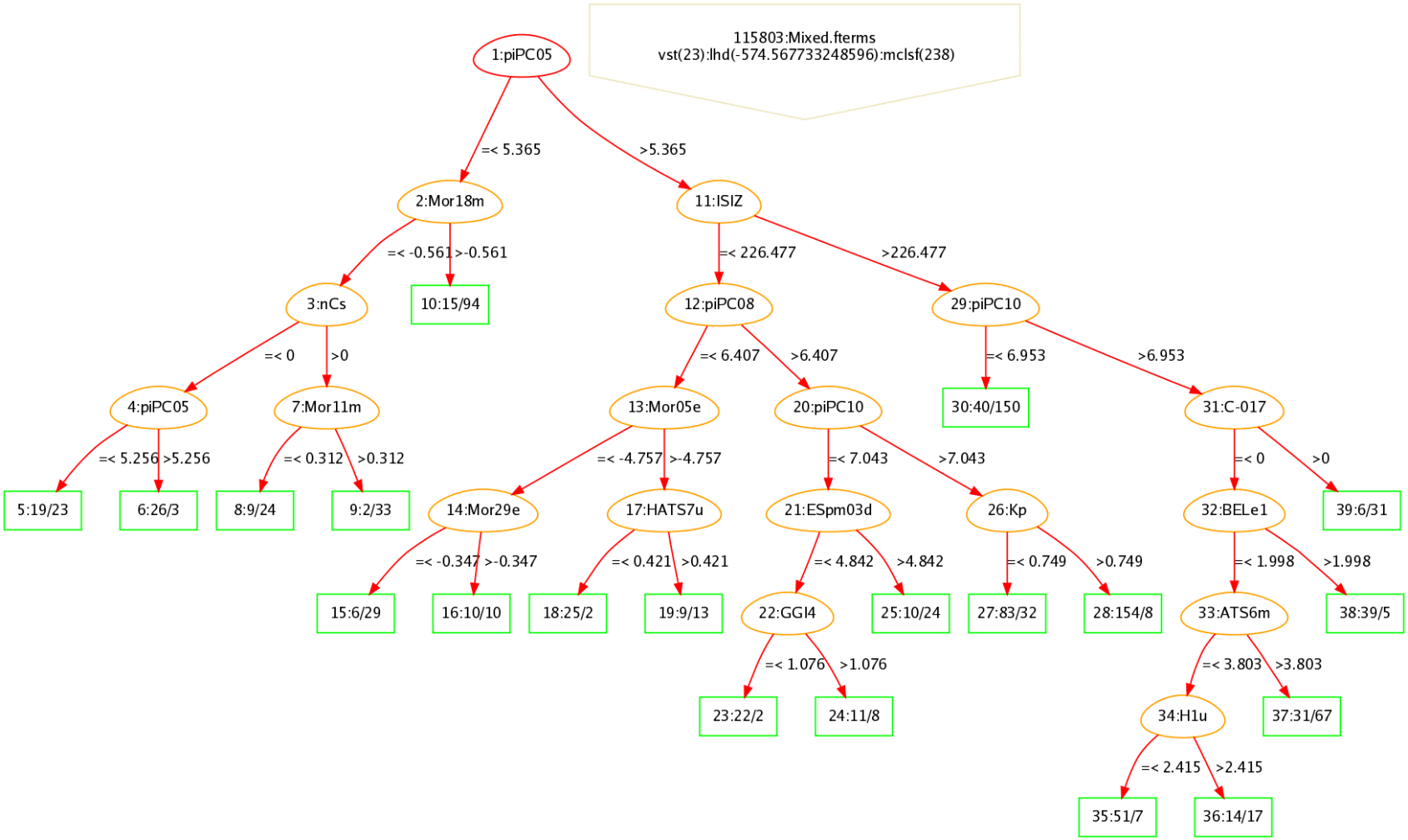
Learning binding molecules

Edinburgh: Pyruvate kinase interactors improve chances of discovering binding molecules based on examples from screened library of chemicals

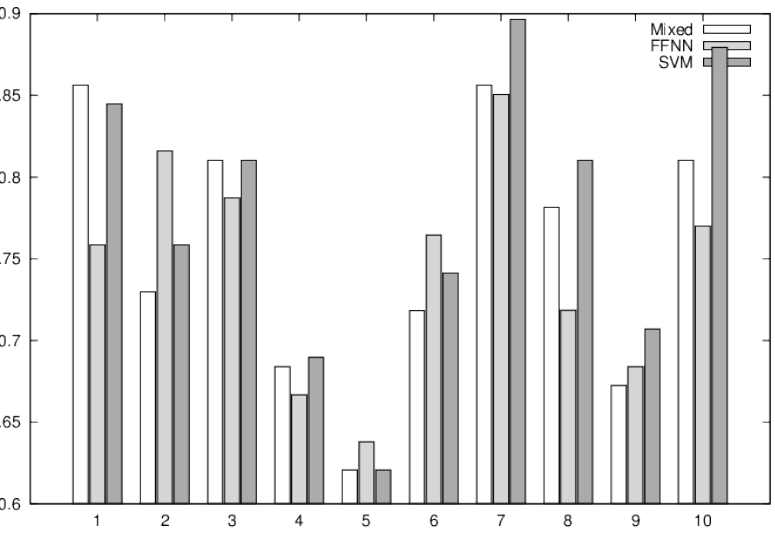
pyruvate kinase affinity data

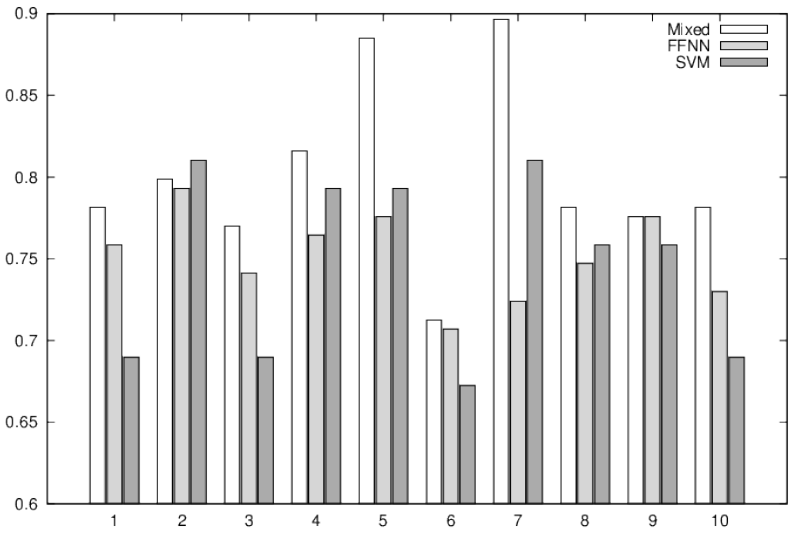
582 Active and 582 Inactive, with 1100 property descriptors for each molecule. Compared to Feed Forward NNs and SVMs.

best likelihood model



ten-fold validation





$$Sensitivity = \frac{T^+}{T^+ + F^-}$$

$$Specificity = \frac{T^-}{T^- + F^+}$$

Stream 2. (Imperial) Knowledge-based data analytics

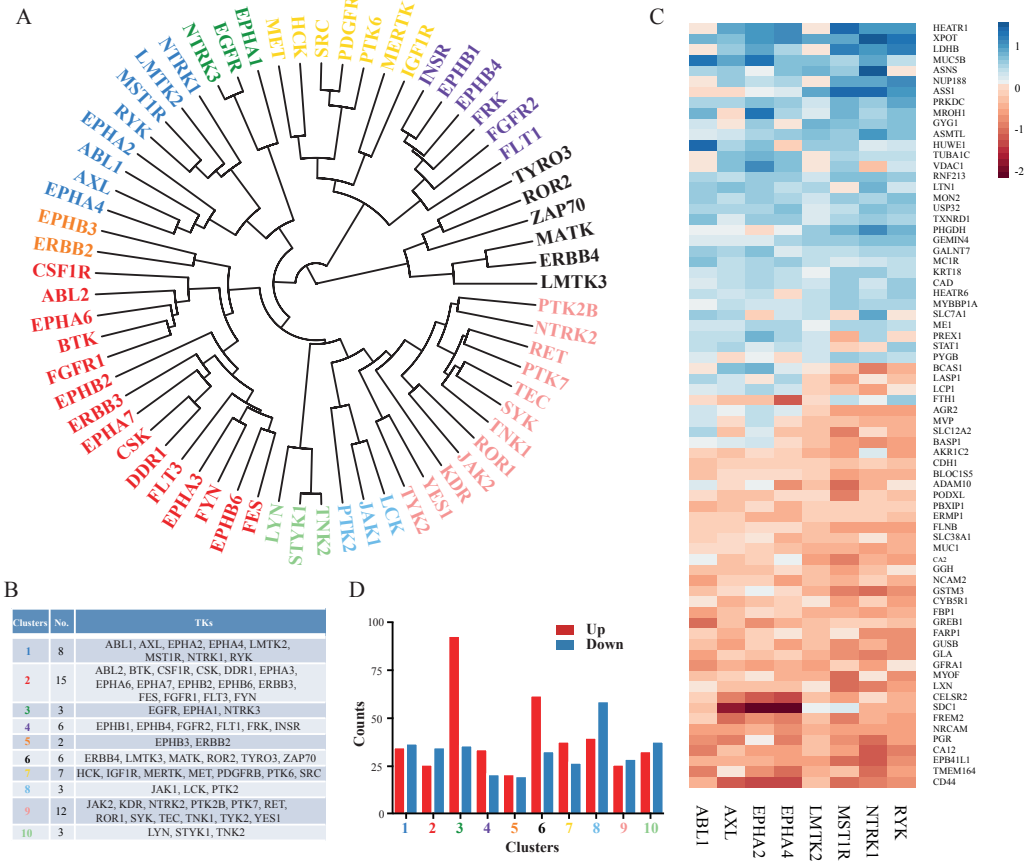
tkSilac: tyrosine kinase screen

- MCF7 cell line
- 33 SILAC runs
- 65/66 expressed tyrosine kinases
- 4739 proteins quantified in some experiment
- 1000 proteins quantified in 60 or more TK KO

Molecular and Cellular Proteomics (MCP) 2015

Tyrosine kinase screen (Imperial)

Figure 4



Clusters	No.	TKs
1	8	ABL1, AXL, EPHA2, EPHA4, LMTK2, MST1R, NTRK1, RYK
2	15	ABL2, BTK, CSF1R, CSK, DDR1, EPHA3, EPHA6, EPHA7, EPHB2, EPHB6, ERBB3, FES, FGFR1, FLT3, FYN
3	3	EGFR, EPHA1, NTRK3
4	6	EPHB1, EPHB4, FGFR2, FLT1, FRK, INSR
5	2	EPHB3, ERBB2
6	6	ERBB4, LMTK3, MATK, ROR2, TYRO3, ZAP70
7	7	HCK, IGF1R, MERTK, MET, PDGFRB, PTK6, SRC
8	3	JAK1, LCK, PTK2
9	12	JAK2, KDR, NTRK2, PTK2B, PTK7, RET, ROR1, SYK, TEC, TNK1, TYK2, YES1
10	3	LYN, STYK1, TNK2

Figure 2

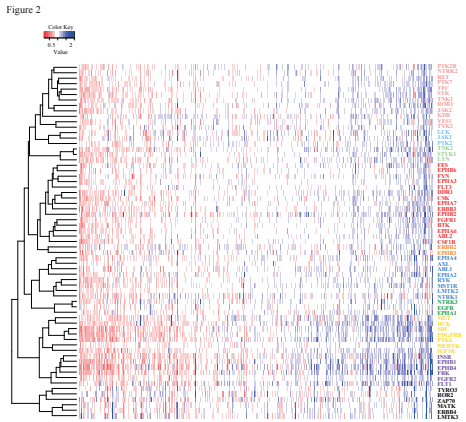


Fig. 2. Heatmap of quantified proteins after TK silencing. The overall pattern of regulation is shown in the heatmap of quantified values. After normalized to siControl, values of fold changes are all above 0, with value 1 showing that the expression levels of the specific protein are not altered after silencing TKs. For each knockdown (rows) the quantified value for an identified protein is plotted in red for down regulated proteins (below 1), white for non-differential and non-identified and blue for up-regulated proteins (above 1). The row labels indicate the knock out experiment and the colors correspond to the clusters described below.

Figure 6

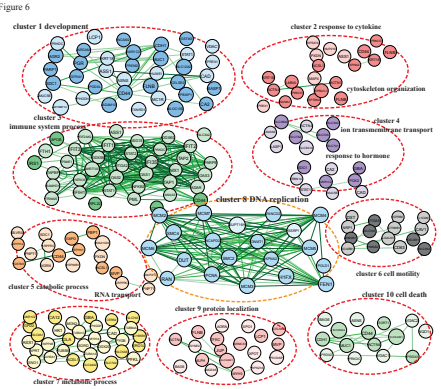
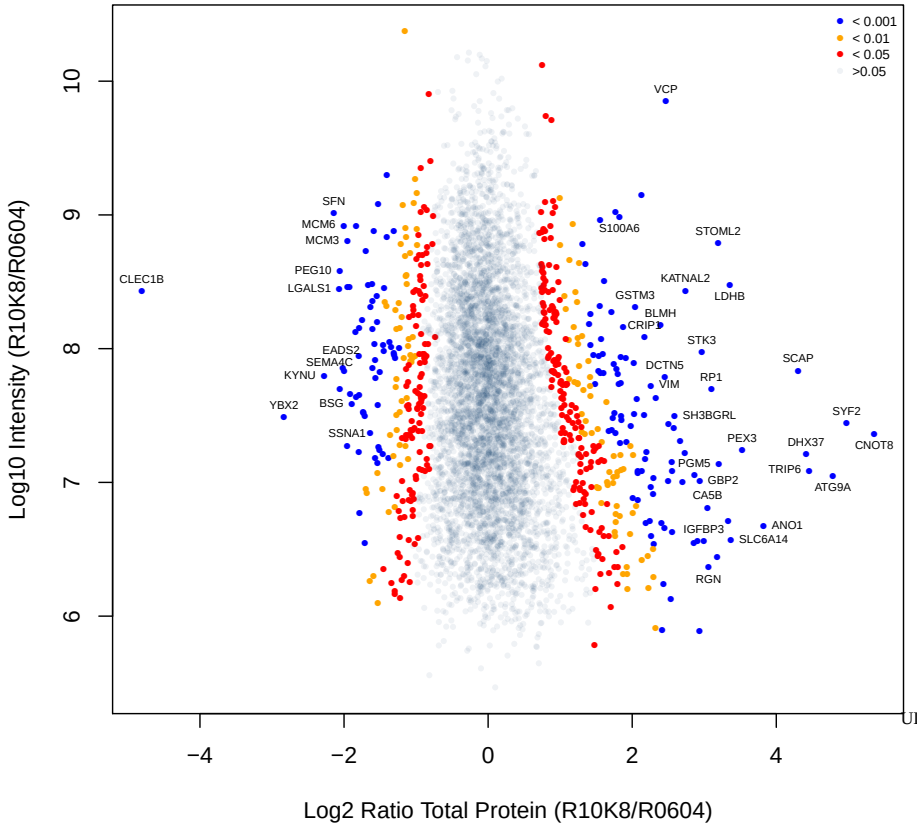
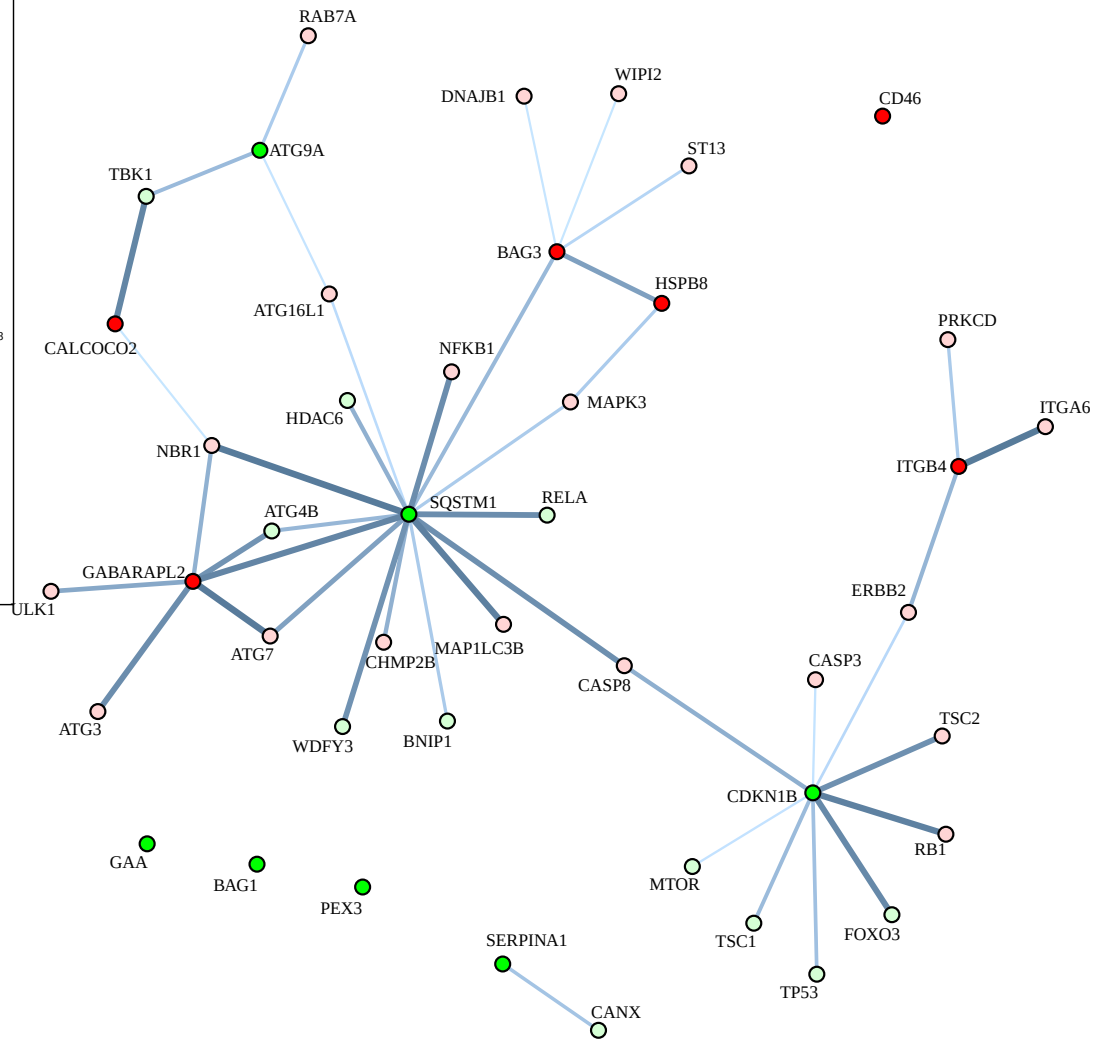


Fig. 6. Representatives of defined functional networks in each classified TK cluster. The functional networks were generated using GO analysis combined with the STRING platform. Proteins in lighter color are up-regulated, whereas brighter color indicates down-regulation. Arrows show the interactions between connected proteins. Representative defined functional networks associated with their clusters are shown here. The color coding and the number for each cluster are indicated as above.

Fig. 4. Hierarchical clustering of the 65 TKs expressed in MCF7 cells. A, Hierarchical clustering of the 65 TKs was performed using R's hclust function. The complete linkage method which aims to identify similar clusters based on overall cluster measure was used. 10 distinctive clusters were obtained and the complete dendrogram is shown with the labels colored for these clusters. B, Full list of the TKs included in each cluster . The color coding of the clusters is used throughout to identify the analysis relevant to the corresponding clusters. C, Heatmap of the proteomic quantifications (log2 values of normalized fold changes against control) for the downstream effects (significantly up- or down-regulated proteins, Significant B test $p < 0.05$) after silencing TKs in cluster 1. D, Number of proteins significantly up or down-regulated in each identified cluster. x-axis shows 10 different clusters and y-axis indicates the counts.

herceptin resistance (BT474HR) — ATG9A / autophagy





proteomics data analytics (Imperial)

tyrosine kinase screen
Molecular and Cellular Proteomics (MCP) 2015

KSR1:
Breast Cancer Res. and Treat., 2015

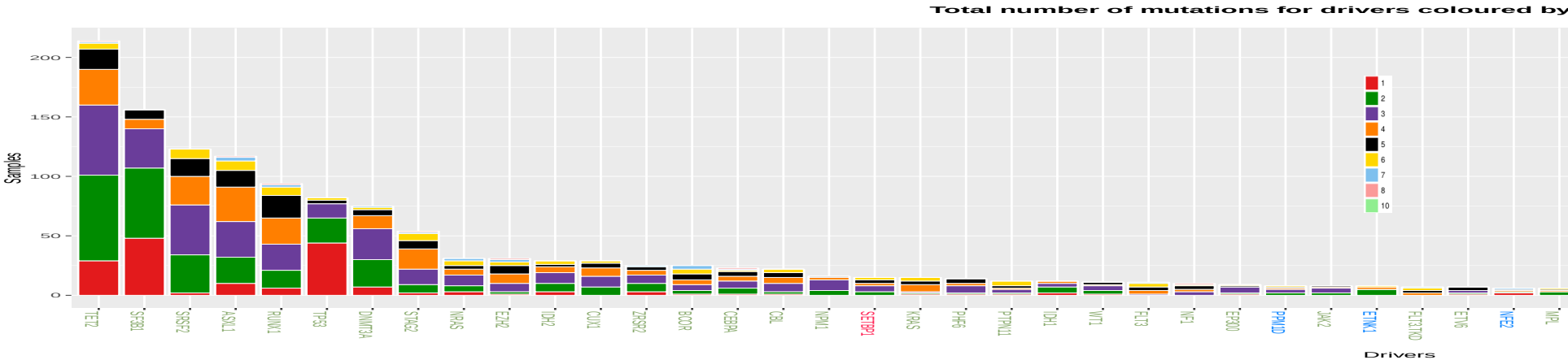
ATG9A:
Oncotarget 2016

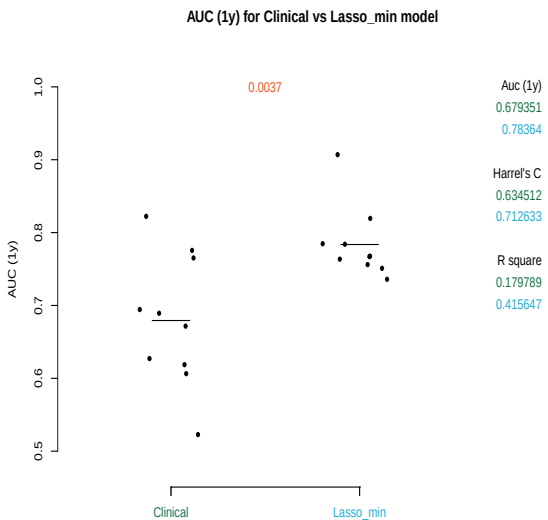
Prolog libraries[1]:
Real (> 580), proSQLite (> 840), bio_analytics

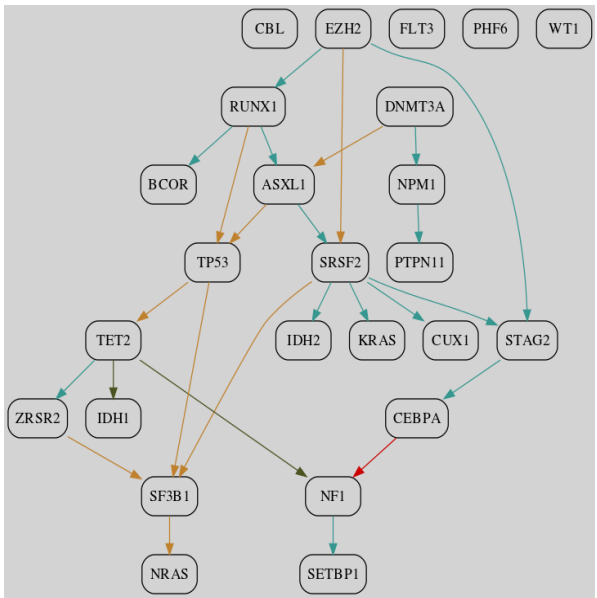
bio_db (currently: 91 tables, 55 M records
on human, mouse, chicken)

[1] work started at NKI

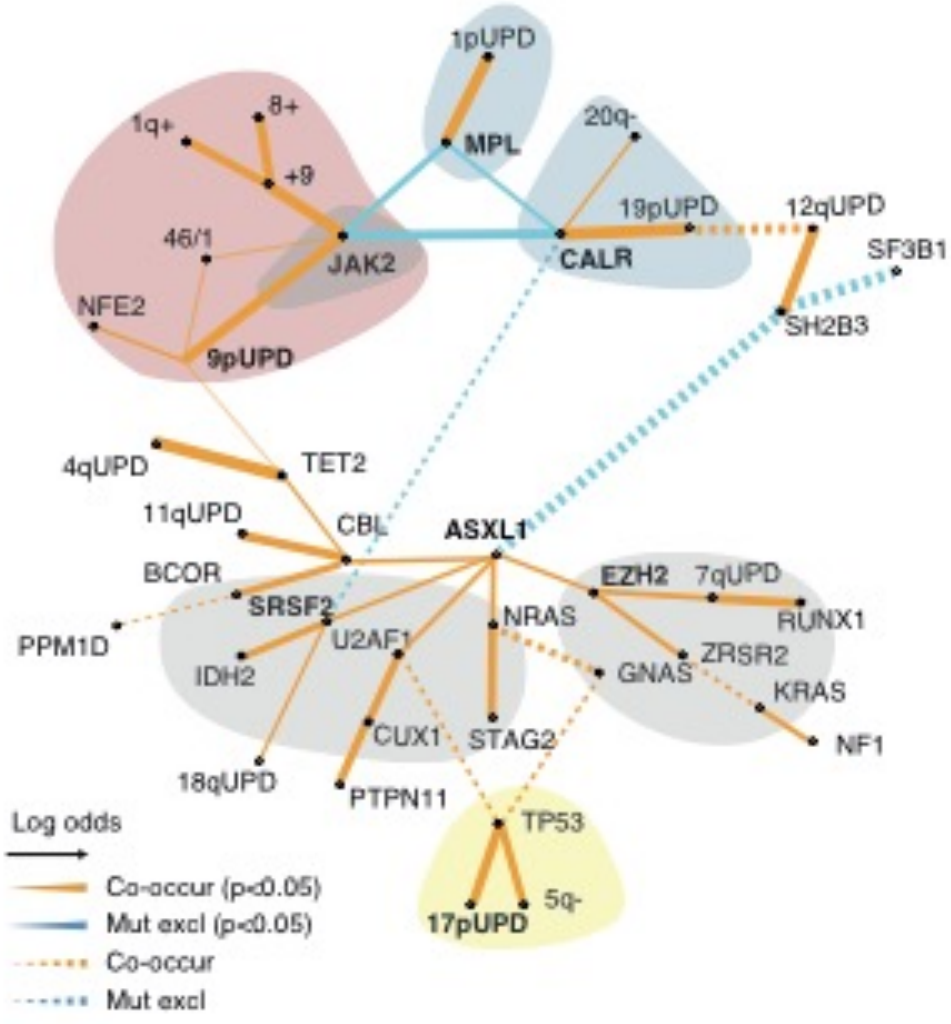
Stream 3. (Sanger) Bayesian networks in cancer genomics





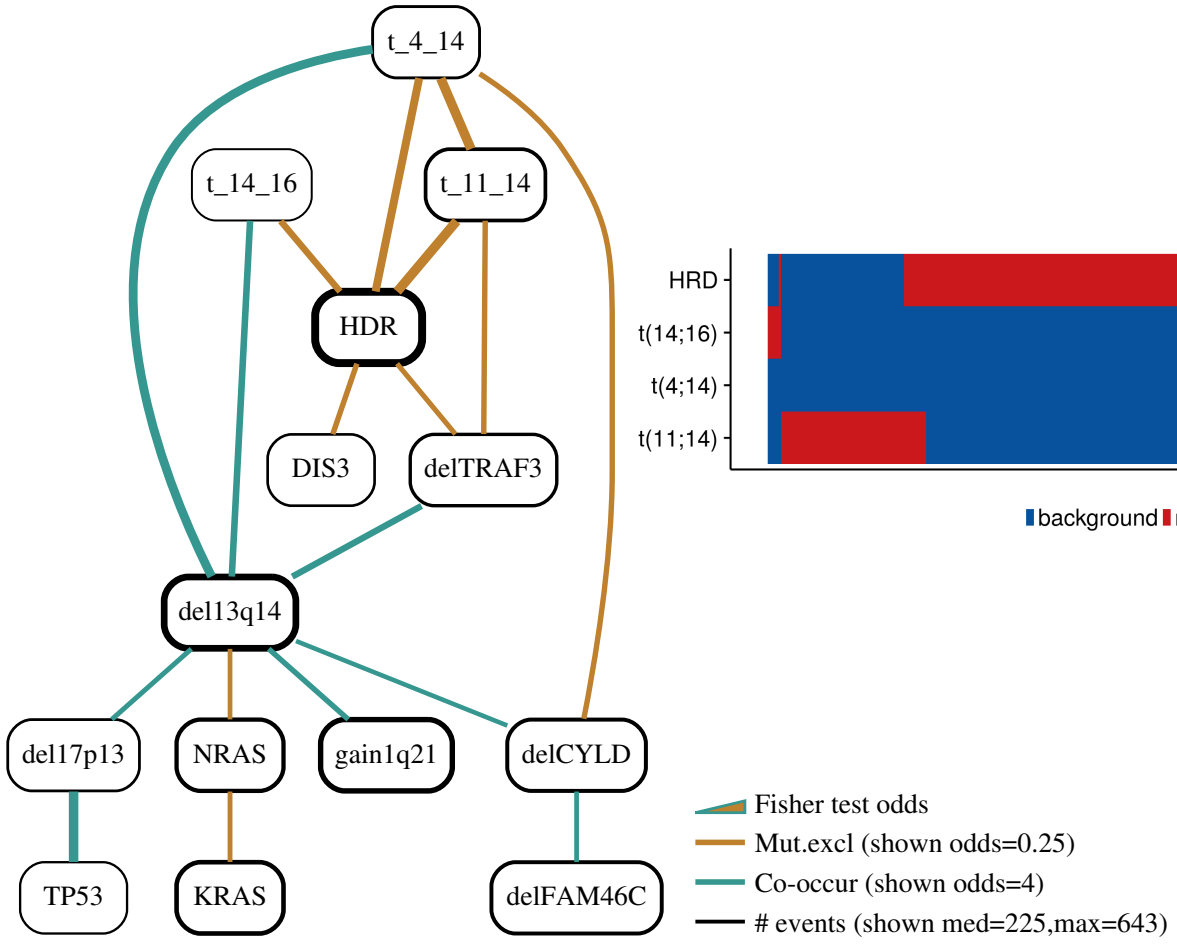


MPN: myeloproliferative neoplasms



New England Journal of Medicine, October 2018

myeloma structural variations



Nature Communications, August 2019

BNs in cancer genomics

- MPN published in New England J. of Medicine, Oct, 2018
- multiple myeloma: in Nature Communications (3rd author), Aug, 2019
- colorectal: January 2020
  (with Dutch collaborators - J. of Clin. Oncology)
- 1st author methods paper:
  Communications Biology (April 2022)

Renal carcinoma, Bayesian estimate

Cell

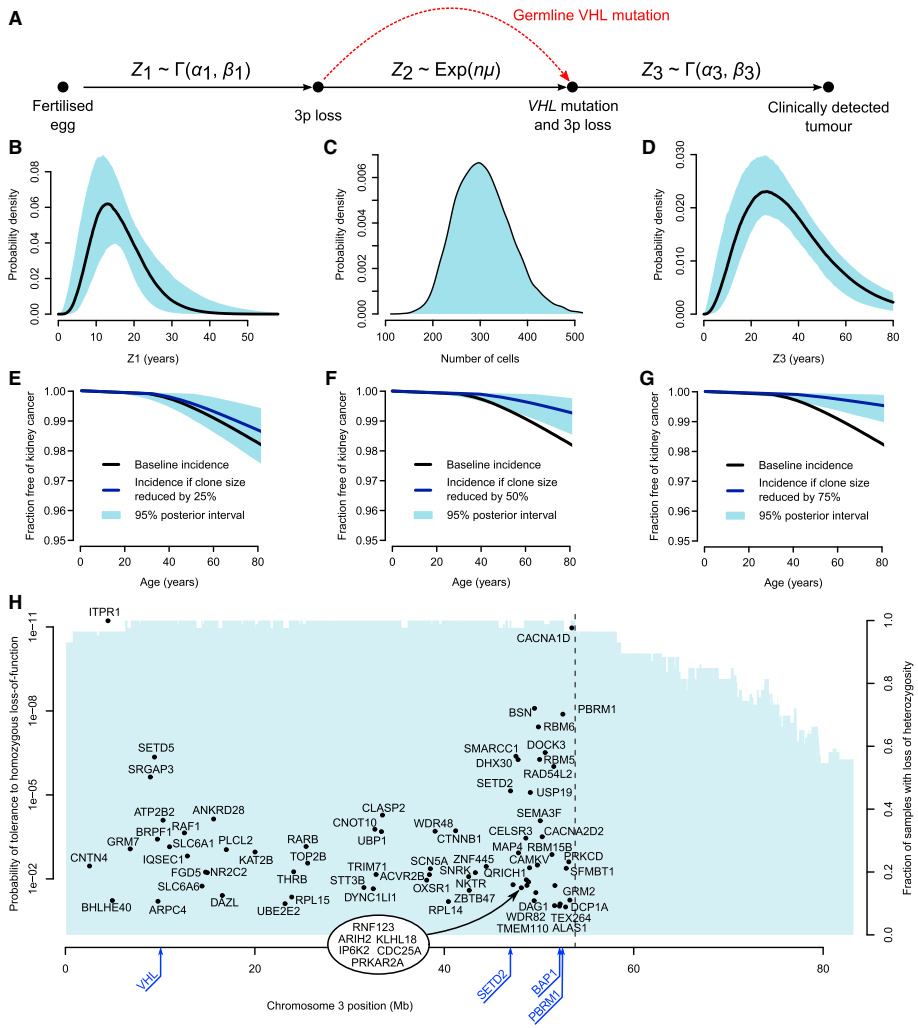


Figure 7. Mathematical Modeling of Clear Cell Renal Cell Carcinoma Evolution

(A) Schematic depicting how the age of incidence of renal cell carcinoma may be modeled as the sum of waiting times; Z_1 representing the time to 3p loss, Z_2 representing the time to *VHL* inactivation, and Z_3 representing the time from bi-allelic loss of *VHL* to clinically detected tumor. Z_1 and Z_3 are modeled by gamma distributions and Z_2 by an exponential distribution of the product of n, the number of cells with 3p loss and μ, the calculated *VHL* mutational rate.

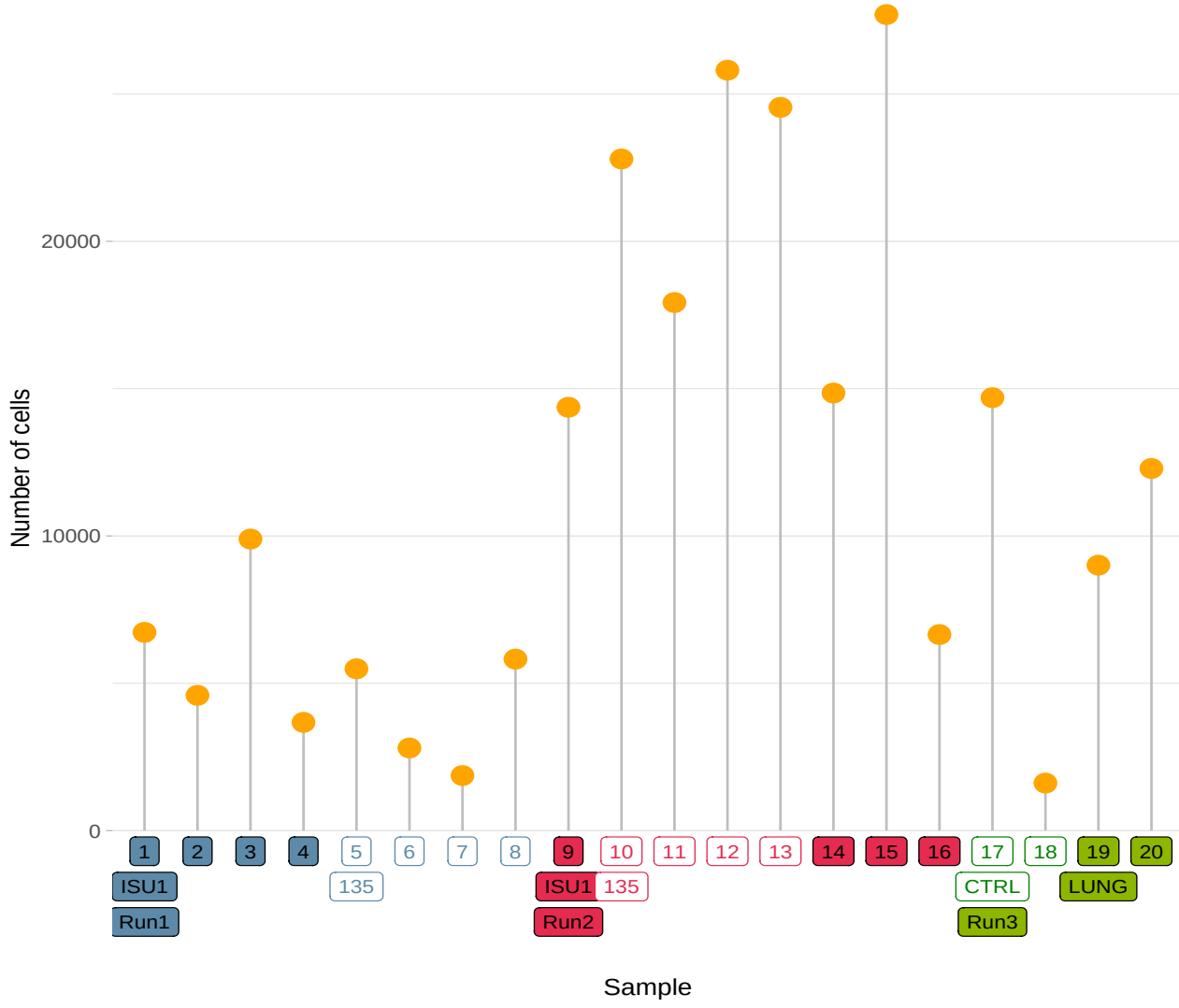
(B–D) The posterior distribution of the waiting times for Z_1 (B), the number of cells with 3p loss (C), and the waiting time for Z_3 (D) with 95% posterior intervals.

Stream 4. single cell RNAseq (scRNA)

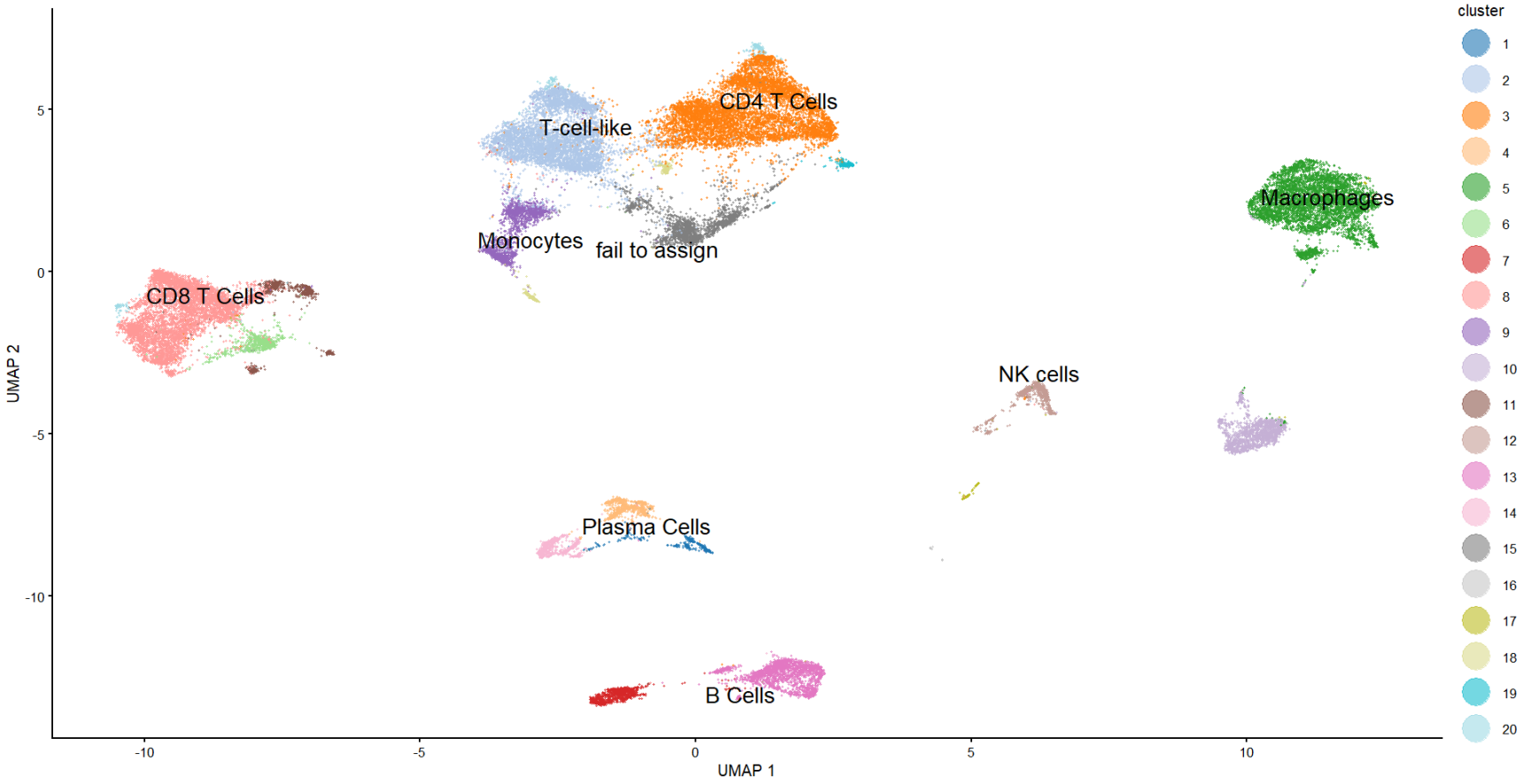
Ability to interrogate expression at a single cell level, but

. . . at major cost to depth

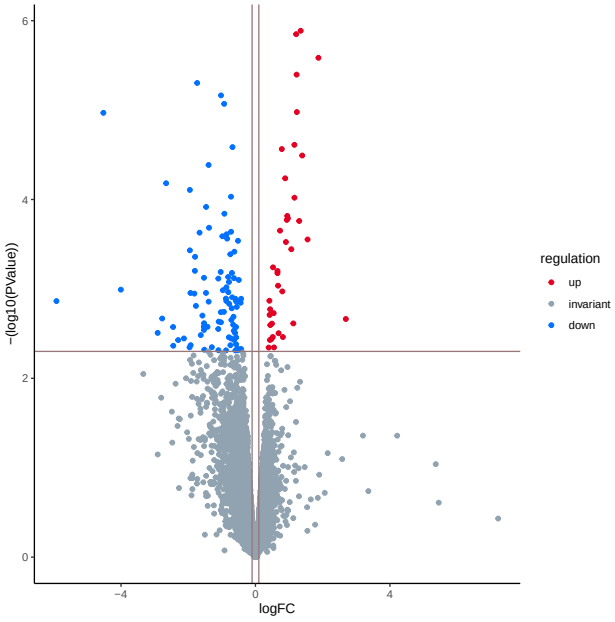
scRNA-seq pig mucosal immunity - all runs

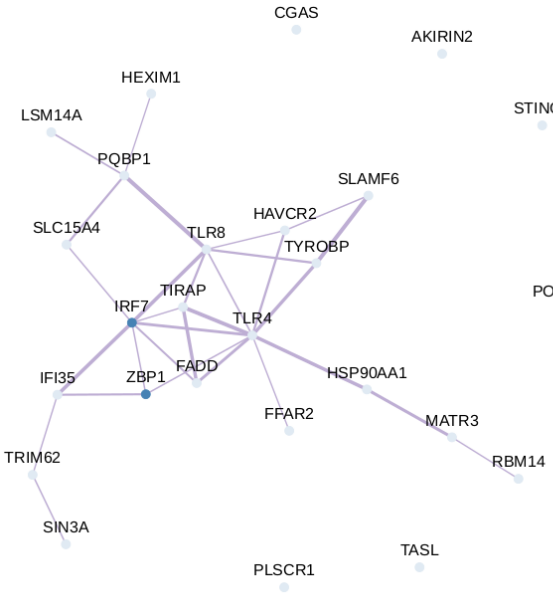


pig immunity - cell type identification

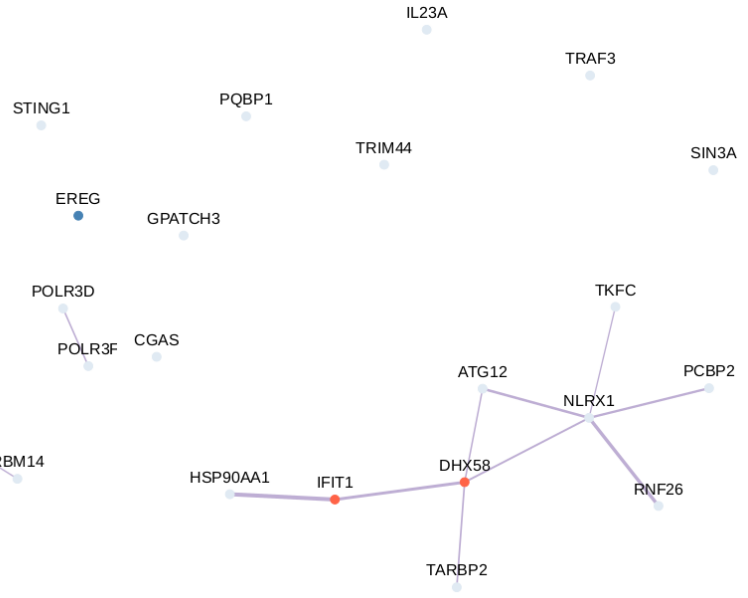


Differential expression



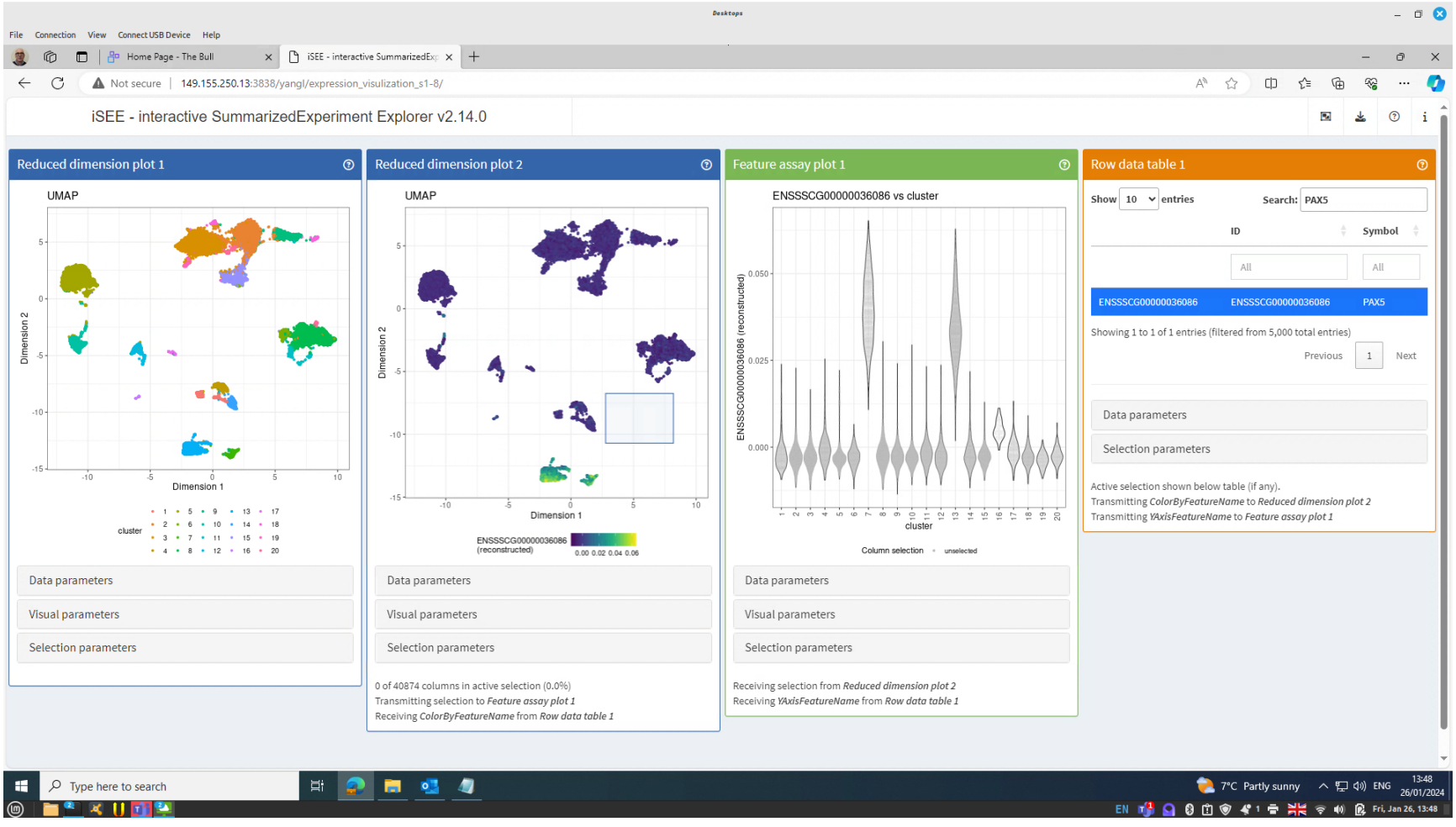


Positive Reg of Innate Immune Rensp.



Reg of Defense Response to Virus

mucosal immunity - scRNA clusters and cells



bottom line - scRNA

generate rich datasets

some loss of depth

some good tools exist

but it is not as straight forward as just applying an R package

truly multi-disciplinary: interactions are necessary

truly multi-disciplinary: appreciation lacking

scientific vision of computational biology

- work in close collaboration with experimental and clinical groups
- get involved early in the formulation of scientific question
- iterative, refining process, (forming a common language)
- properly resource analysis and computational tasks
- data management life cycle
- policies on data and analytics
- harmonious development and use of resources
- simple and robust solutions

Thank you