

knowledge AI in computational biomedicine

Nicos Angelopoulos

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

<https://stoics.org.uk/~nicos>
nicos@email.stoics.org.uk

25.08.28

Personal background

BSc Computer Science & Statistics

MSc Advanced Computer Science (AI)

PhD Computer Science

2000-2018 Researcher in projects and labs

2019-2025 Independent academic positions:

- ▶ Essex (CS)
- ▶ Cardiff (Medicine)
- ▶ Pirbright (livestock immunity)

Data analytics, algorithms, machine learning and software engineering for research in biomedicine with emphasis on cancer.

Open source code for analytics- including machine learning.

Linux/Unix user for over 35 years.

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, Sanger 2022)

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

(Imperial) Knowledge-based data analytics

Figure 2

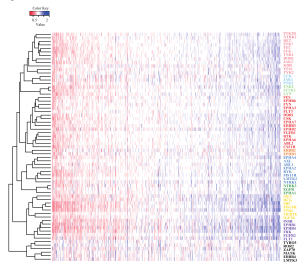


Figure 3

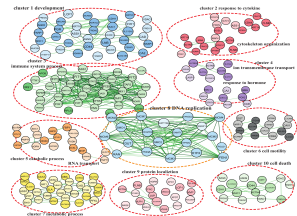
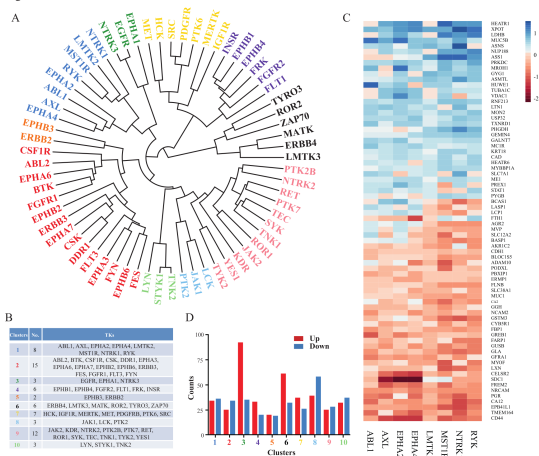
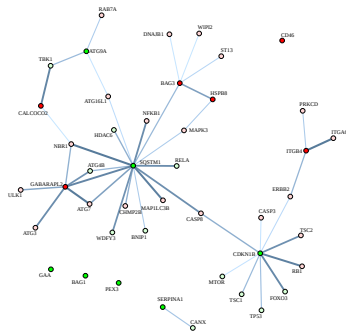
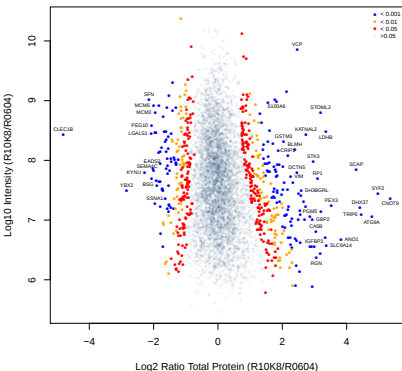


Figure 4



herceptin resistance (BT474HR) — ATG9A / autophagy



tyrosine kinase screen

Molecular and Cellular Proteomics (MCP) 2015

KSR1: Breast Cancer Res. and Treat., 2015

ATG9A: Oncotarget 2016

proteomics data analytics (Imperial)

tyrosine kinase screen

Molecular and Cellular Proteomics (MCP) 2015

KSR1:

Breast Cancer Res. and Treat., 2015

ATG9A:

Oncotarget 2016

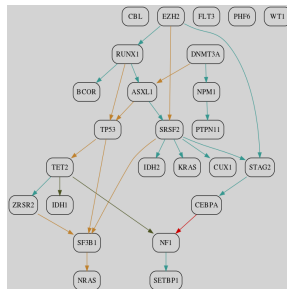
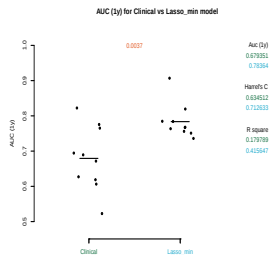
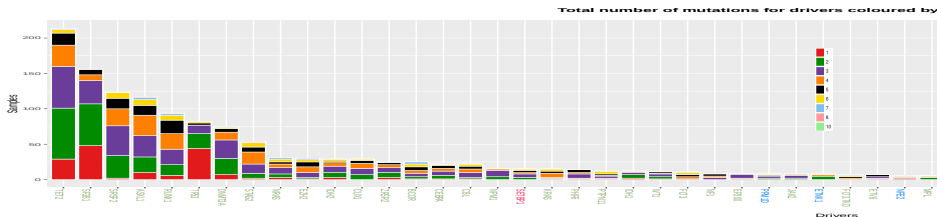
Prolog libraries¹:

Real (600), proSQLite (> 1200), bio_analytics

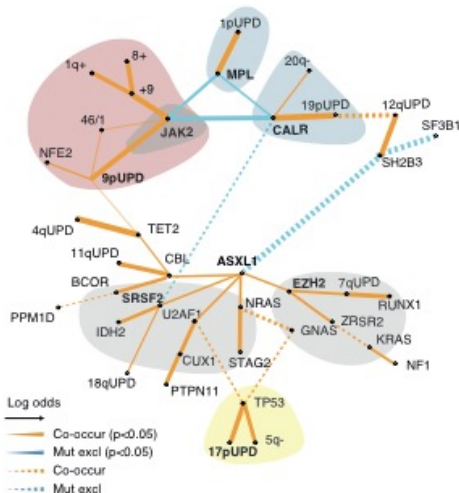
bio_db (currently: 91 tables, 55 M records
on human, mouse, chicken)

¹work started at NKI

(Sanger) Bayesian networks in cancer genomics

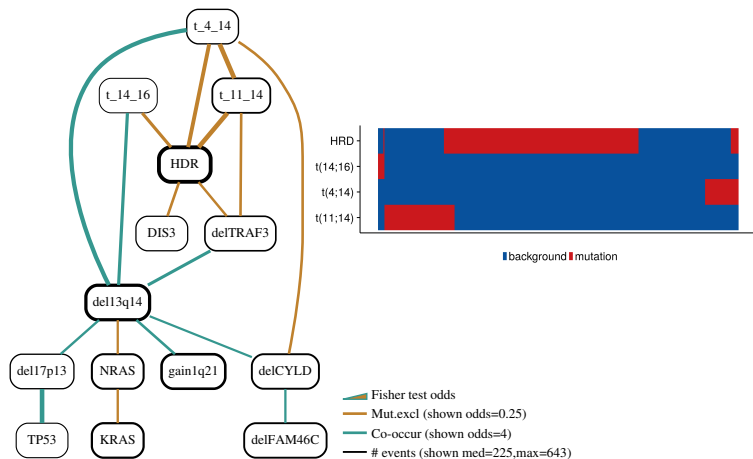


MPN: myeloproliferative neoplasms



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 (cited > 680)
- ▶ multiple myeloma: in Nature Communications (3rd author), Aug, 2019 (cited > 300)
- ▶ colorectal: January 2020 (with Dutch collaborators - J. of Clin. Oncology)
- ▶ 1st author methods paper:
Communications Biology (April 2022)

Renal carcinoma, Bayesian estimate



(cited > 550)

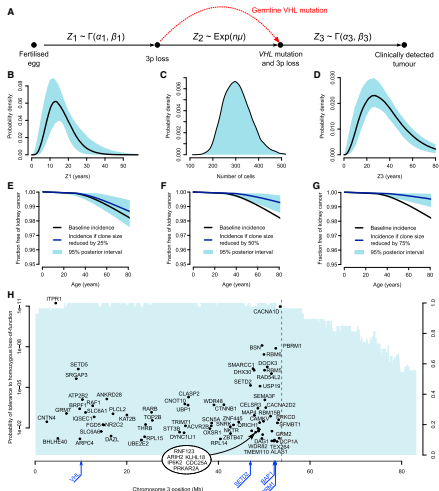
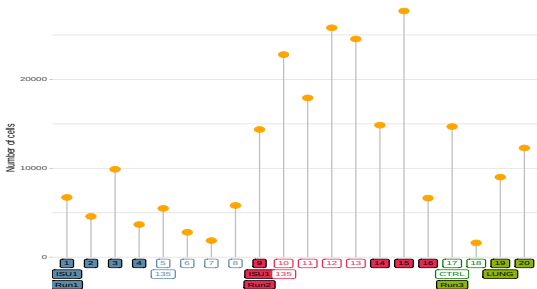


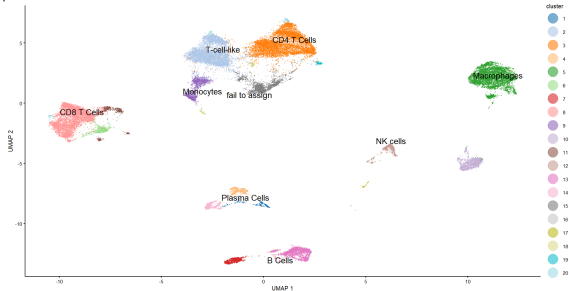
Figure 7. Mathematical Modeling of Clear Cell Renal Carcinoma Evolution

(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_2 are modeled by gamma distributions and Z_3 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_1), and the waiting time for Z_3 (D) with 95% posterior intervals.

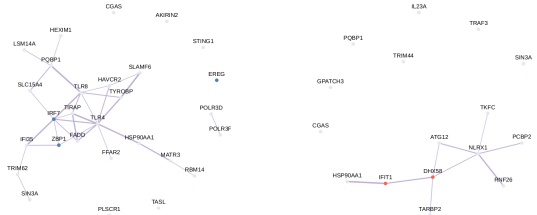
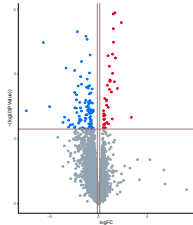
scRNA-seq pig mucosal immunity - all runs



Sam

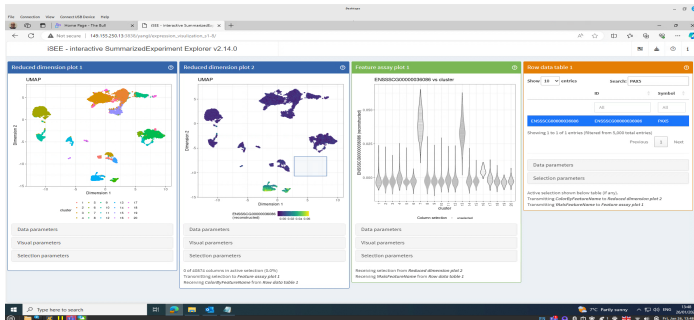


Differential expression



+ve Reg of Innate Imm.

Reg of Defense Response to Virus



experience with data science



Data analytics, algorithms, machine learning and engineering for research software run on a variety of computing systems.

Open source code for analytics- including machine learning.

Academic lead of HPC system at Pirbright

- ▶ instigated project based code development with use of github
- ▶ trained bioinformatician and biologists in the use of HPC
- ▶ data management for sequencing facility

Linux/Unix user for over 35 years.

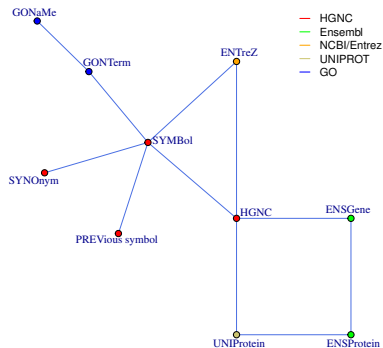
experience with big databases

Programmatic access to best biological databases

- ▶ NCBI/Entrez
- ▶ Ensembl
- ▶ Uniprot
- ▶ HGNC
- ▶ Gene Ontology

across a number of species

- ▶ human
- ▶ mouse
- ▶ chicken
- ▶ porcine
- ▶ bovine

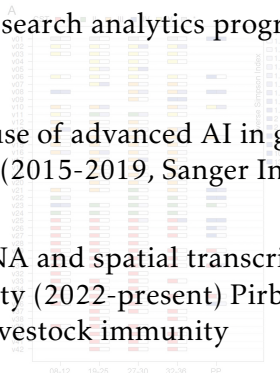


Research analytics programme building

Building data-rich research analytics programmes based on experience

(a) the use of advanced AI in genomic cancer cohorts (2015-2019, Sanger Institute)

(b) scRNA and spatial transcriptomics for immunity (2022-present) Pirbright Institute, B-cell livestock immunity



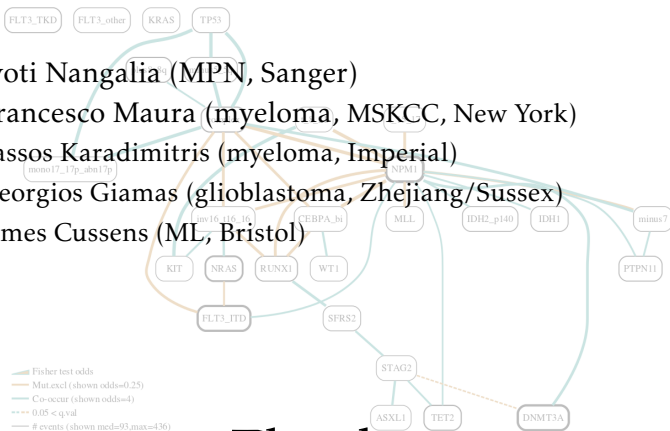
computational biology roadmap

- ▶ contribute access to computational resources and explainable AI and ML expertise to all groups
- ▶ get involved early in
the formulation of scientific question
- ▶ iterative,
refining process,
forming a common language
- ▶ cultivate data analytics within
biology and medicine
- ▶ empower causality in biomedicine via explainable AI



External collaborators

- ▶ Jyoti Nangalia (MPN, Sanger)
- ▶ Francesco Maura (myeloma, MSKCC, New York)
- ▶ Tassos Karadimitris (myeloma, Imperial)
- ▶ Georgios Giamas (glioblastoma, Zhejiang/Sussex)
- ▶ James Cussens (ML, Bristol)



Thank you