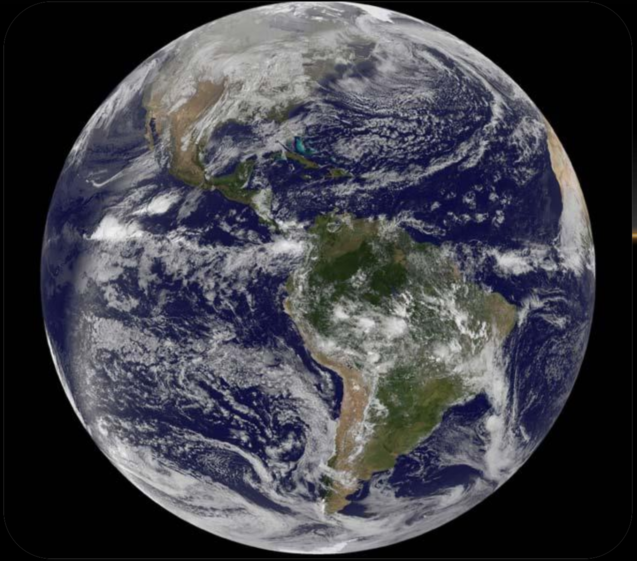
Systems Immunology & Immune Oncology
A Data-Centric View



https://pics-about-space.com/pics-new-world-nasa?p=2

Magnus Fontes

Disclosures & Affiliations

- General Manager of Institut Roche, France https://institut.roche.com/
- Adjunct professor of mathematics, Lund University, Sweden https://portal.research.lu.se/en/persons/magnus-fontes
- Co-founder of the bioinformatics software company Qlucore www.qlucore.com

A brief personal history of computing

Slide rule

Hp-35 (1972)

©FLT-PICA, Stockholm, Sverige/Sweden. 1998-05-20 FLT-PICA code 20360 ***Arkivbild/File Picture 1980***
ABC-80 dator fran Luxor
FLT PICA

ABC80 (1978)
16 kB RAM

Sparc Station
Beginning of the 90’s
Up to 500 MB in RAM

?

Nearly a century-and-a-half ago, Louis Pasteur envisioned
a world where "Institut Pasteur" scientists across the globe
would share their research and knowledge in service of humanity...

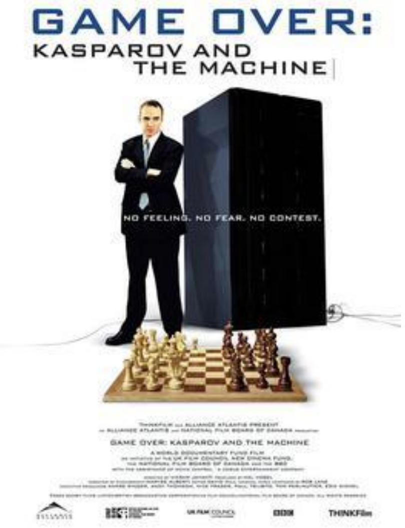
Pasteur Global Health Genomics Center
https://vimeo.com/171747507

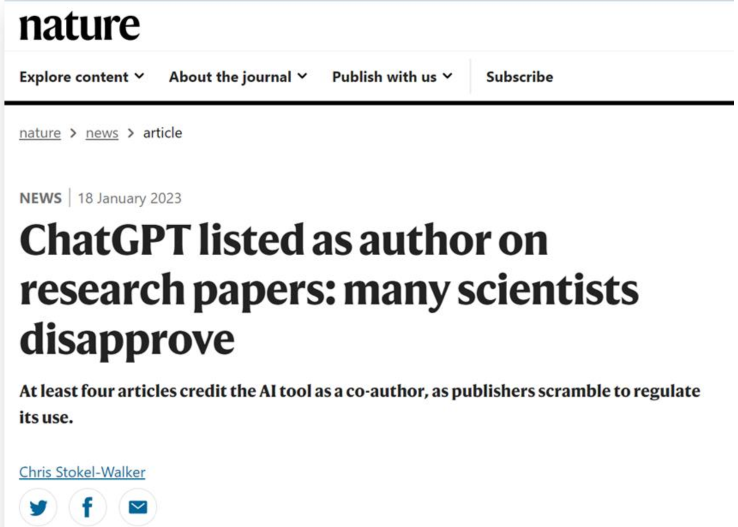
...standing united "on the edge of (global) mysteries"
and aspiring to "lift the veil" on the origin of diseases.

Digitalized Biomedicine and Healthcare

The Human Machine Partnership has placed us in a position to deliver even more on the vision of Louis Pasteur “standing united and aspiring to lift the veil”
through Data driven collaborative modeling

Multidisciplinary & Collaborative research through Human-Machine Partnerships


GAME OVER:
KASPAROV AND
THE MACHINE
NO FEELING. NO FEAR. NO CONTEST.
GAME OVER: KASPAROV AND THE MACHINE
THINKFilm



NEWS | 18 January 2023

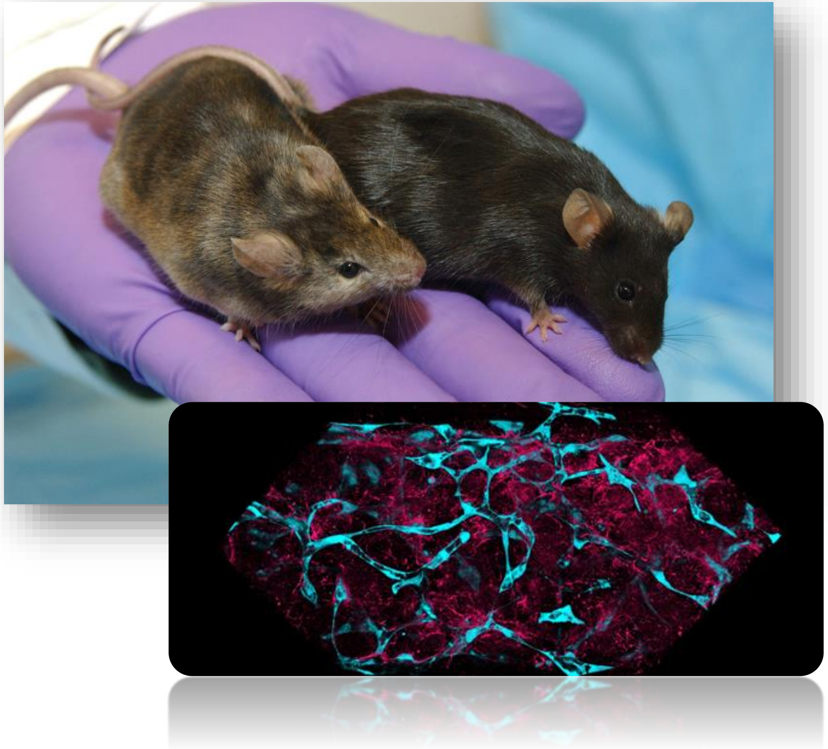
ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

Chris Stokel-Walker

- Individual Human Knowledge → Connected Global Knowledge
- Legal & Ethical & Psychological considerations
- How do we use our new tools? Repurposing Generative AI?

Do you have a model for that?



$$\frac{dx}{dt} = \alpha x - \beta x y,$$

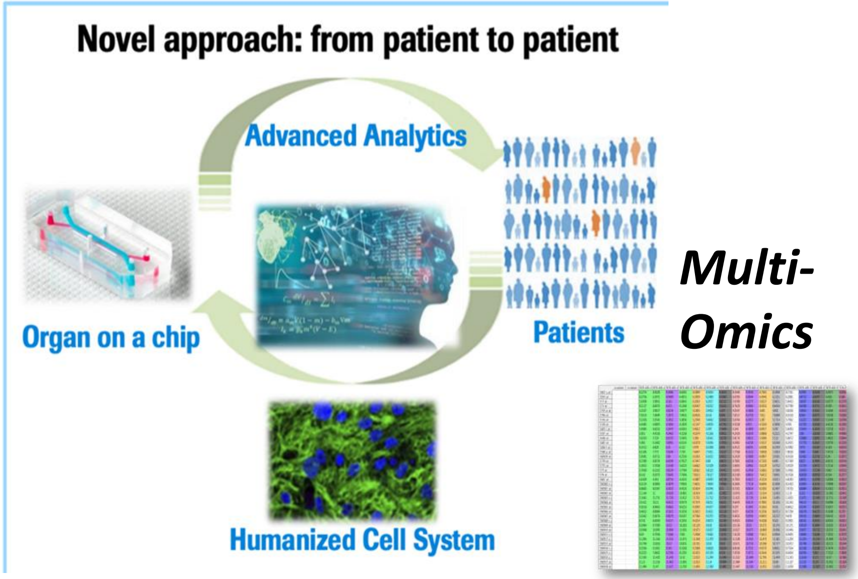
$$\frac{dy}{dt} = \delta x y - \gamma y,$$

Transformation of Biomedical R&D

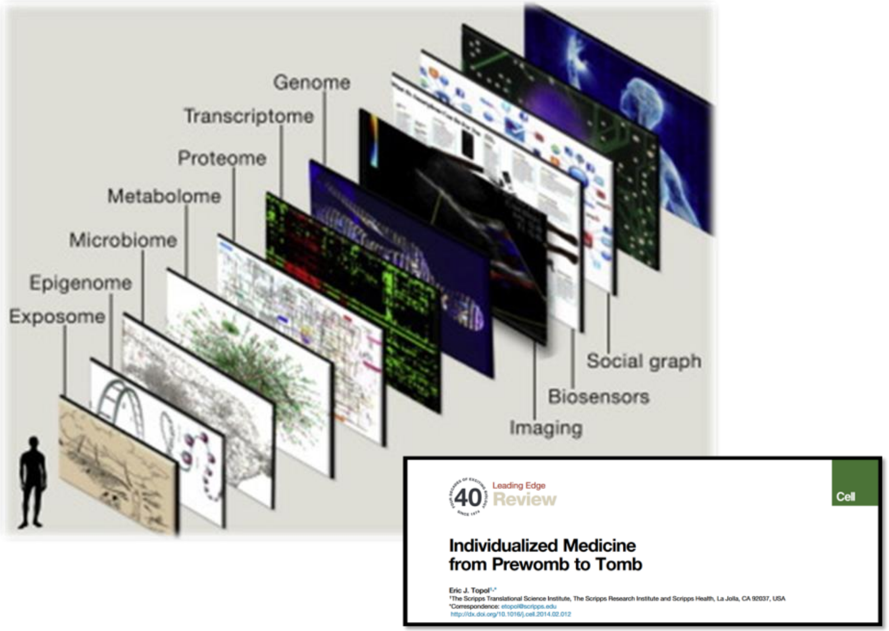
- Reverse Translational Research
- Exvivo biological systems – Biological Avatars
- Insilico Systems – Digital Avatars
- Advanced Analytics through AI and Human-Machine partnerships

Transformation of Healthcare:

- Longitudinal and deep precision sampling
- Early detection and diagnosis
- Treating early disease
- Personalization & Combination therapies
- Clinical decision support via Personalized Biological + Digital Avatars



Multi-Omics



Individualized Medicine from Prewomb to Tomb

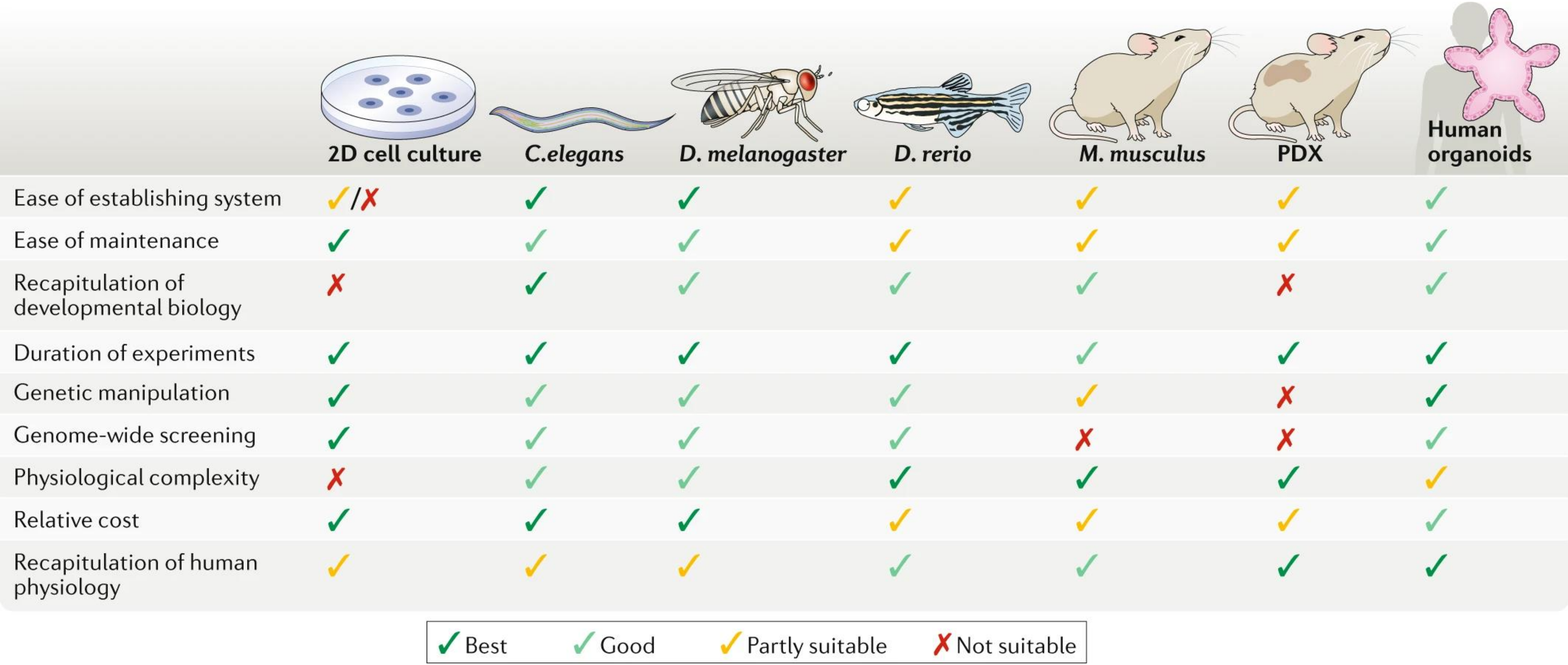
Eric J. Topol[1,*]

[1]The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA

*Correspondence: etopol@scripps.edu

http://dx.doi.org/10.1016/j.cell.2014.02.012

Biological Avatars – Exvivo Systems → Modeling holistic dynamics

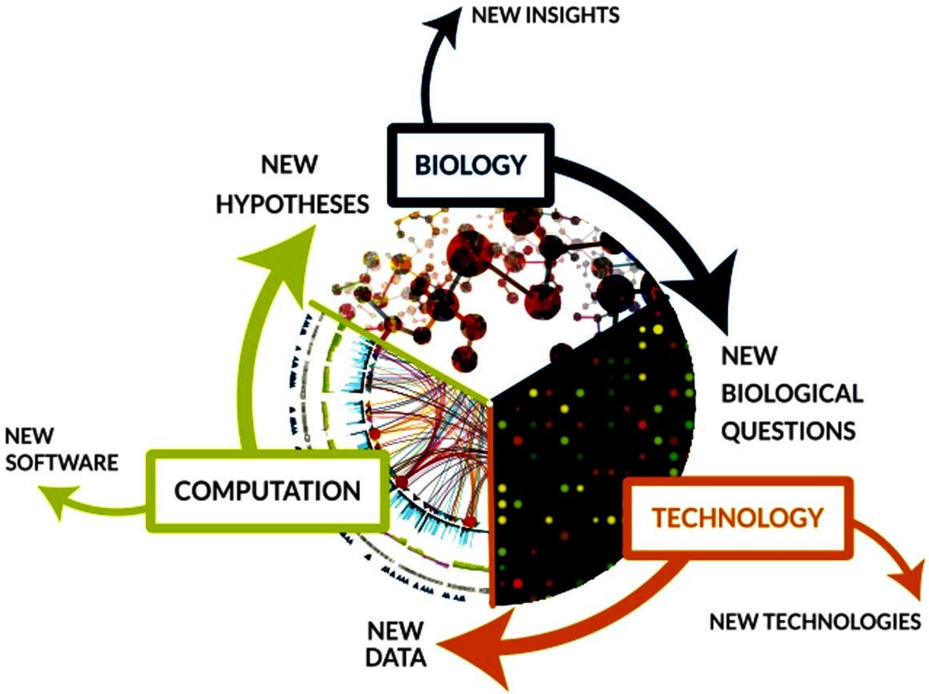


	2D cell culture	*C.elegans*	*D. melanogaster*	*D. rerio*	*M. musculus*	PDX	Human organoids
Ease of establishing system	(✓)/✗	✔	✔	(✓)	(✓)	(✓)	✓
Ease of maintenance	✔	✓	✓	(✓)	(✓)	(✓)	✓
Recapitulation of developmental biology	✗	✔	✓	✓	✓	✗	✓
Duration of experiments	✔	✔	✔	✔	✓	✔	✔
Genetic manipulation	✔	✓	✓	✓	(✓)	✗	✔
Genome-wide screening	✔	✓	✓	✓	✗	✗	✓
Physiological complexity	✗	✓	✓	✔	✔	✔	(✓)
Relative cost	✔	✔	✔	(✓)	(✓)	(✓)	✓
Recapitulation of human physiology	(✓)	(✓)	(✓)	✓	✓	✔	✔

✔ Best ✓ Good (✓) Partly suitable ✗ Not suitable

Kim, J., Koo, BK. & Knoblich, J.A. Human organoids: model systems for human biology and medicine. Nat Rev Mol Cell Biol 21, 571–584 (2020). https://doi.org/10.1038/s41580-020-0259-3

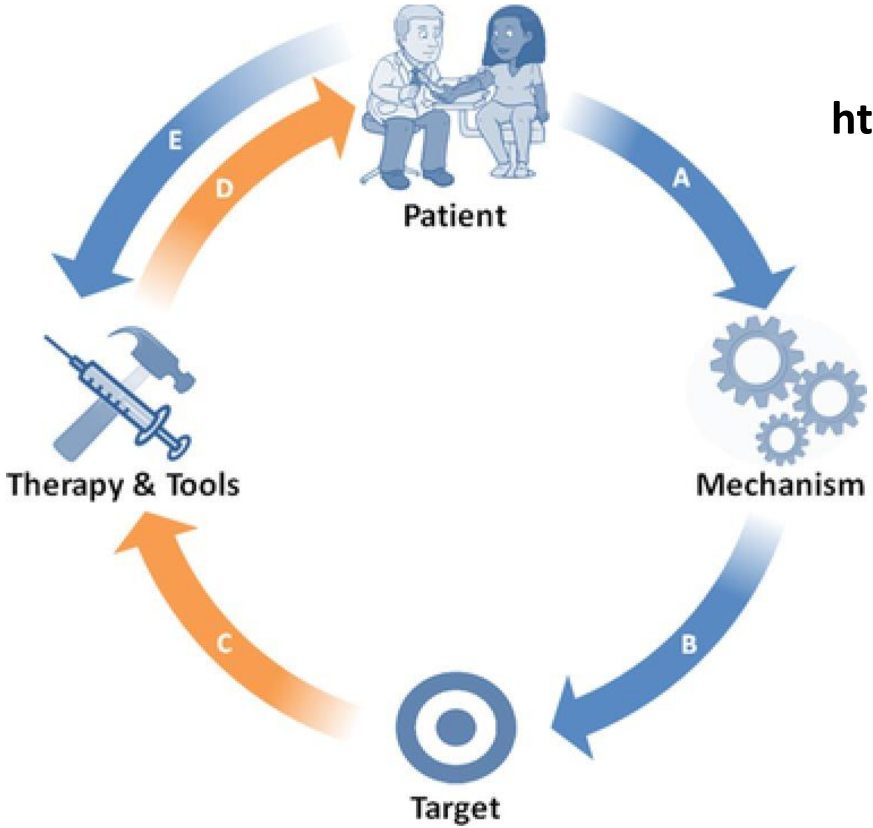
Digital Avatars -- Computational Systems → Predicitive modeling



https://www.nibib.nih.gov/science-education/science-topics/computational-modeling

Reverse Translational Resarch – « *The ground truth* »

The ultimate model for human health and disease is human health and disease



Clin Transl Sci. 2018 Mar; 11(2): 98–99.
Published online 2018 Feb 9. doi: 10.1111/cts.12538

https://www.gene.com/stories/reverse-translation

"In a clinical trial of one of our cancer immunotherapies, our scientists observed differences in people who responded to the medicine. By digging deeper into the biology of the non-responders, the team discovered that some of them had an up-regulated gene signature associated with a protein called TGF-beta. With this knowledge, they then explored the phenomenon in pre-clinical models. When they combined the cancer immunotherapy with an investigational antibody that blocks TGF-beta, it resulted in improved anti-tumor activity in pre-clinical models that mimic the biology of some non-responders."

Computational Mathematical Modeling through Compartmentalized Integrative Grey Box Modeling

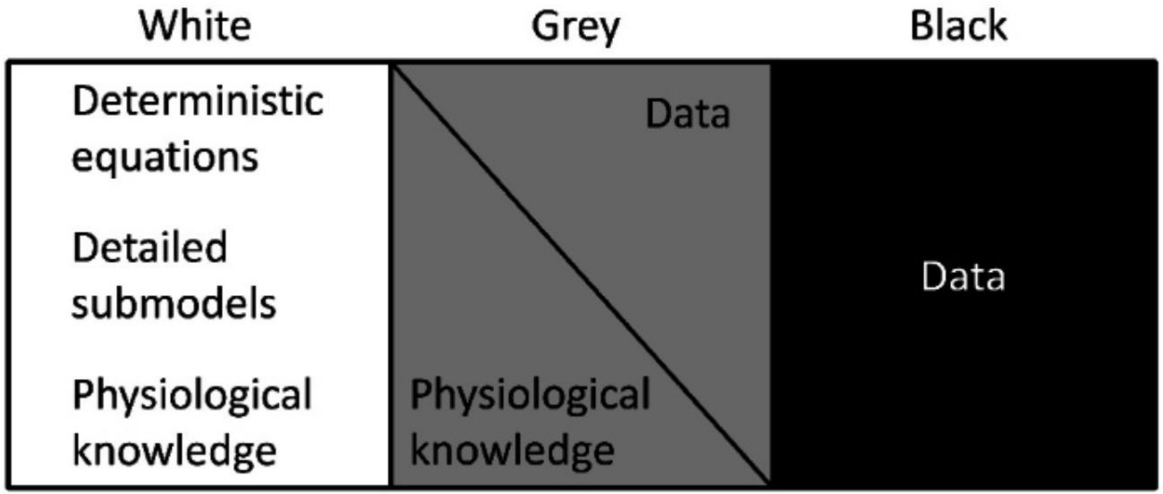
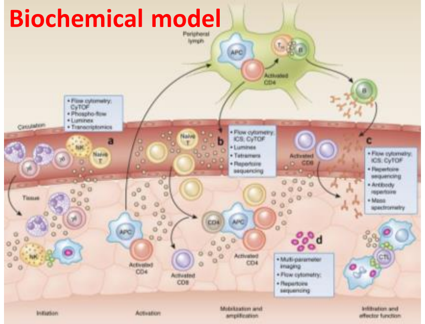


Illustration of the concept of grey-box modeling. White-box models are based mainly on knowledge about the system. Black-box models are built on statistical information from the data. Grey-box modeling combines the two approaches.

J Diabetes Sci Technol. 2013 Mar; 7(2): 431–440.
Published online 2013 Mar 1. doi: 10.1177/193229681300700220

The Grey Box Modeling Loop Connecting Data & Models


Biochemical model

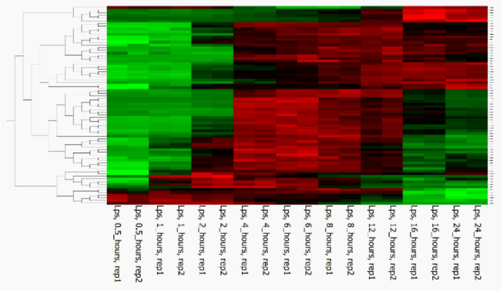
Mechanistic mathematical model

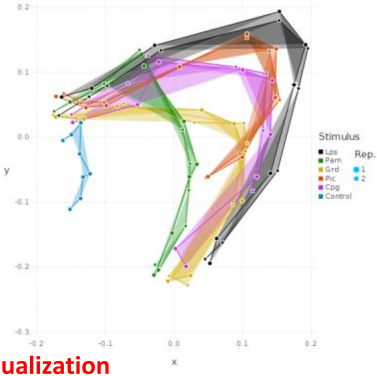
$$\frac{dx}{dt} = \alpha x - \beta x y,$$

$$\frac{dy}{dt} = \delta x y - \gamma y,$$

"A Systems Biology Workbench"

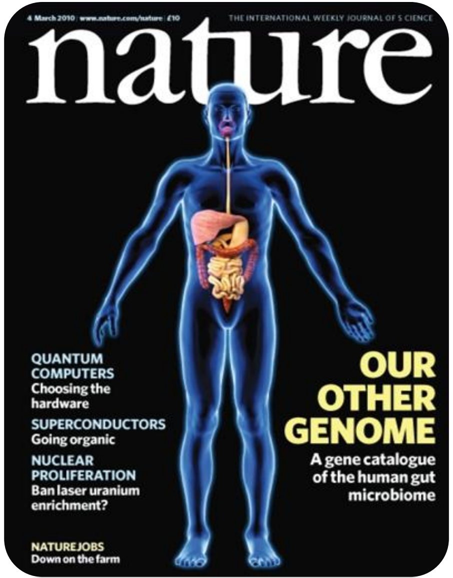
A holistic engine for scientific discovery and innovation





Statistical learning and visualization

The human model system


4 March 2010 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
QUANTUM COMPUTERS
Choosing the hardware
SUPERCONDUCTORS
Going organic
NUCLEAR PROLIFERATION
Ban laser uranium enrichment?
NATUREJOBS
Down on the farm
OUR OTHER GENOME
A gene catalogue of the human gut microbiome

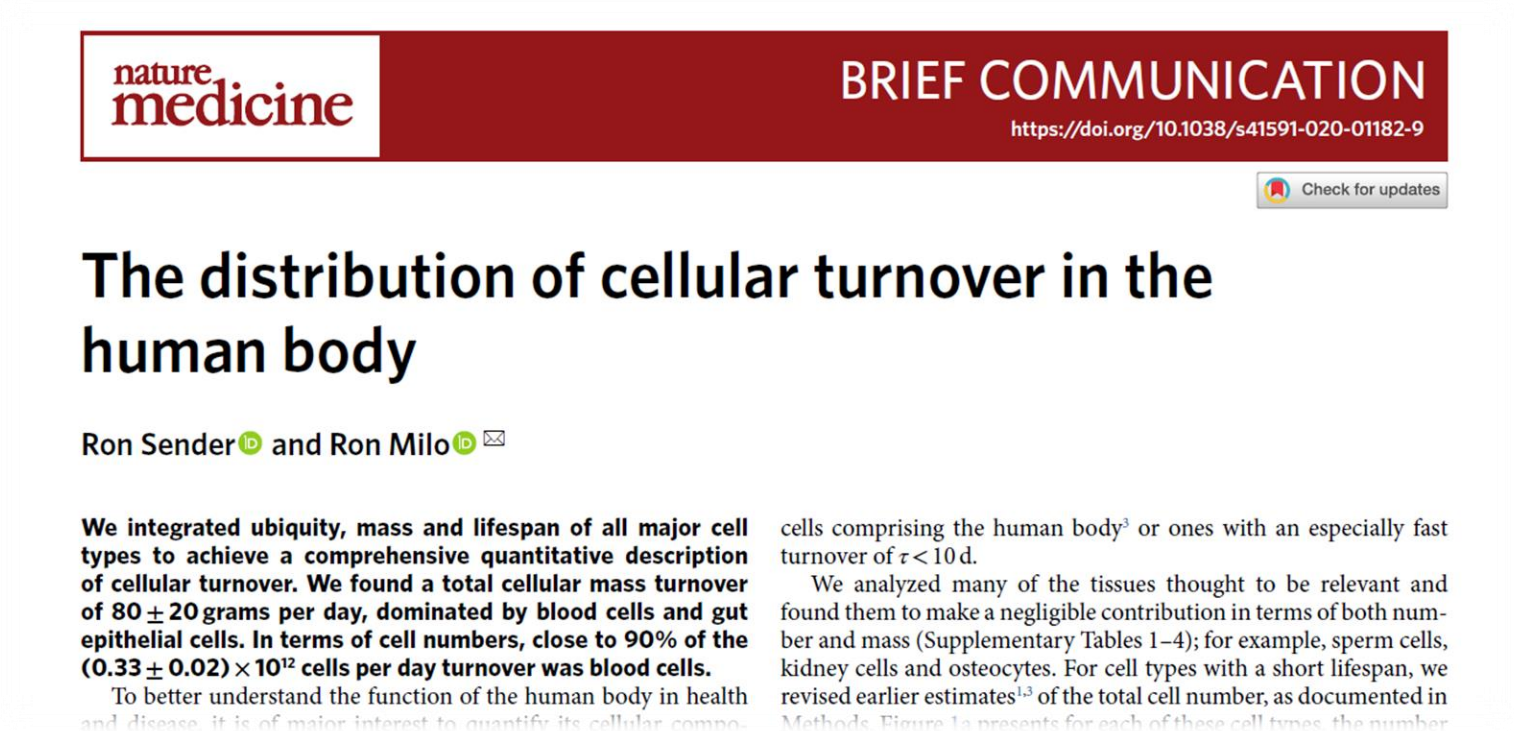
Nature March 2010

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
First results from the Human Microbiome Project highlight the healthy variation in our microbial selves
PAGES 194, 207 & 215
FELLOW TRAVELLERS
MATERIALS
FEELING THE PRESSURE
CLIMATE CHANGE
GET USED TO UNCERTAINTY
Climate modelling faces its limits
SOLAR SYSTEM
TITAN'S ELUSIVE METHANE
Tropical lakes on Saturn's enigmatic moon?
NATURE.COM/NATURE
14 June 2012

Nature June 2012

- 3 × 10^13 human cells
- As many bacterial cells
- Around 10 times as many viruses
- Around 10^10 proteins per human cell

A very dynamical system...



nature medicine

BRIEF COMMUNICATION

https://doi.org/10.1038/s41591-020-01182-9

Check for updates

The distribution of cellular turnover in the human body

Ron Sender and Ron Milo

We integrated ubiquity, mass and lifespan of all major cell types to achieve a comprehensive quantitative description of cellular turnover. We found a total cellular mass turnover of 80 ± 20 grams per day, dominated by blood cells and gut epithelial cells. In terms of cell numbers, close to 90% of the $(0.33 \pm 0.02) \times 10^{12}$ cells per day turnover was blood cells.

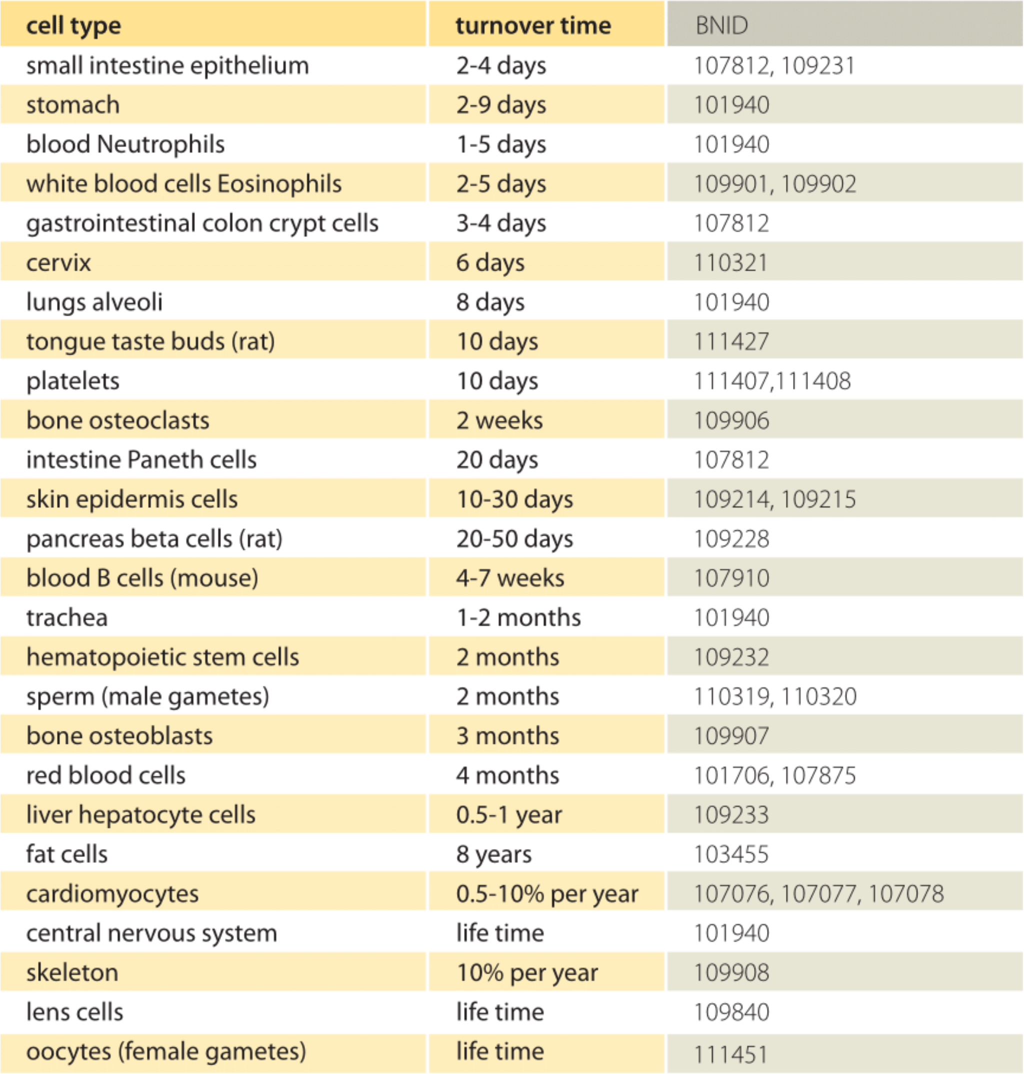
To better understand the function of the human body in health and disease, it is of major interest to quantify its cellular compo- … cells comprising the human body[3] or ones with an especially fast turnover of $\tau < 10$ d.

We analyzed many of the tissues thought to be relevant and found them to make a negligible contribution in terms of both number and mass (Supplementary Tables 1–4); for example, sperm cells, kidney cells and osteocytes. For cell types with a short lifespan, we revised earlier estimates[1,3] of the total cell number, as documented in Methods. Figure 1a presents for each of these cell types, the number

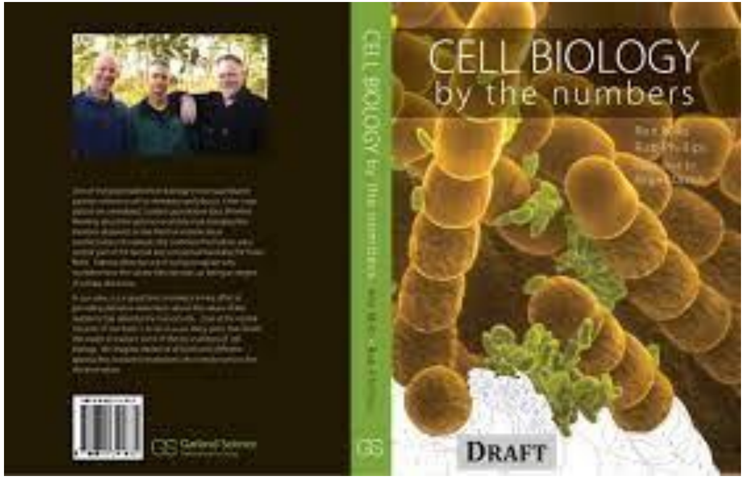
Approximately 4 million human cells die per second in an average human

Bionumbers, e.g. Lifespan of a cell:

https://bionumbers.hms.harvard.edu/search.aspx

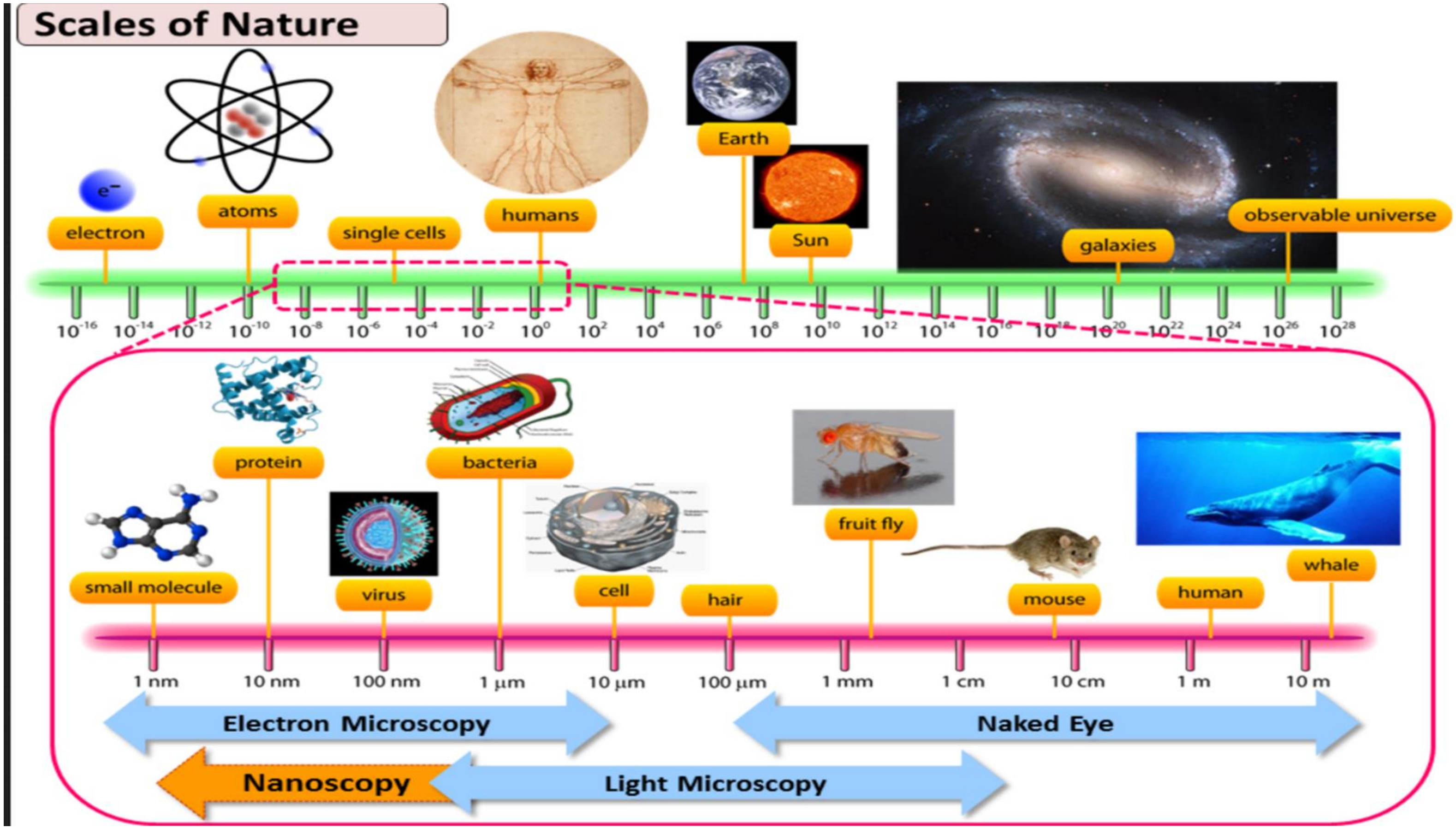


cell type	turnover time	BNID
small intestine epithelium	2-4 days	107812, 109231
stomach	2-9 days	101940
blood Neutrophils	1-5 days	101940
white blood cells Eosinophils	2-5 days	109901, 109902
gastrointestinal colon crypt cells	3-4 days	107812
cervix	6 days	110321
lungs alveoli	8 days	101940
tongue taste buds (rat)	10 days	111427
platelets	10 days	111407,111408
bone osteoclasts	2 weeks	109906
intestine Paneth cells	20 days	107812
skin epidermis cells	10-30 days	109214, 109215
pancreas beta cells (rat)	20-50 days	109228
blood B cells (mouse)	4-7 weeks	107910
trachea	1-2 months	101940
hematopoietic stem cells	2 months	109232
sperm (male gametes)	2 months	110319, 110320
bone osteoblasts	3 months	109907
red blood cells	4 months	101706, 107875
liver hepatocyte cells	0.5-1 year	109233
fat cells	8 years	103455
cardiomyocytes	0.5-10% per year	107076, 107077, 107078
central nervous system	life time	101940
skeleton	10% per year	109908
lens cells	life time	109840
oocytes (female gametes)	life time	111451



http://book.bionumbers.org/

Some (log) length scales of Life

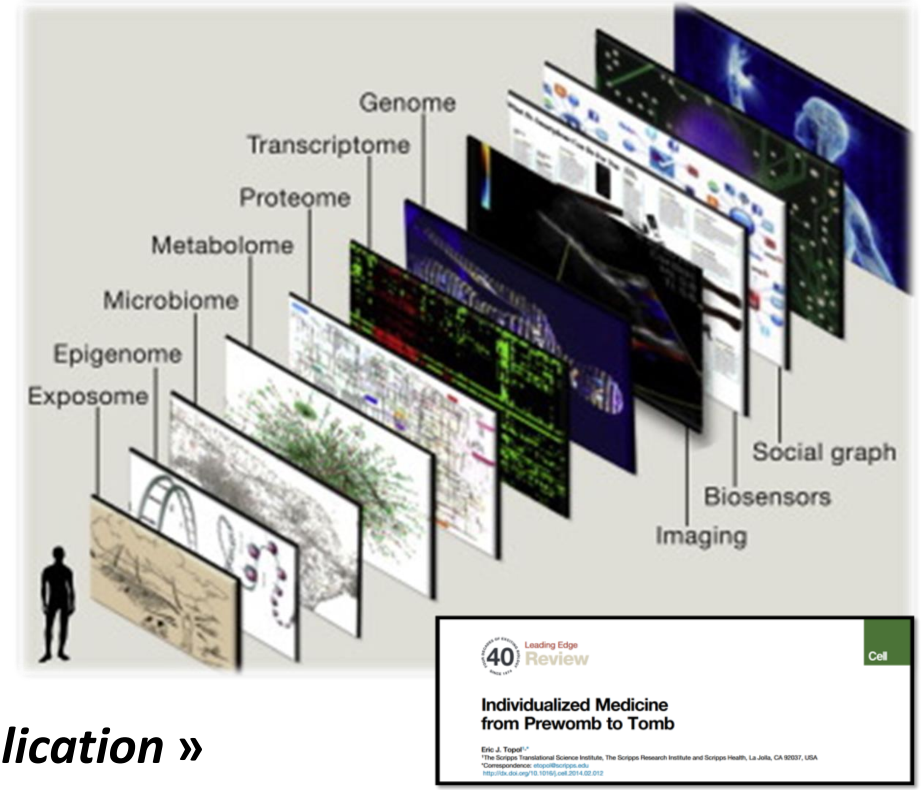


http://www.bates.edu/gould-research-lab

A view of our hierarchically organized system
Organizational levels of life:

- …
- Molecules ←→
- Cells ←→
- Tissues←→
- Organs ←→
- Organ complexes ←→
- Organisms
- Populations
- ….

The focus depends on « *the target application* »



Leading Edge Review

Cell

Individualized Medicine from Prewomb to Tomb

Eric J. Topol[1],*

[1] The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA

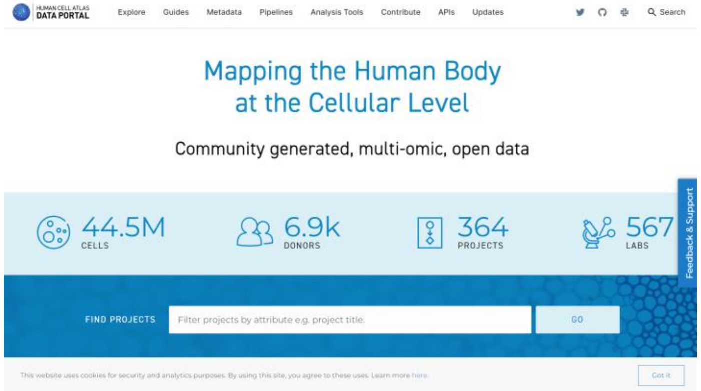
*Correspondence: etopol@scripps.edu

http://dx.doi.org/10.1016/j.cell.2014.02.012

Focusing on Cells – The basic units of Life

HUMAN
CELL
ATLAS

https://www.humancellatlas.org/



Mapping the Human Body
at the Cellular Level

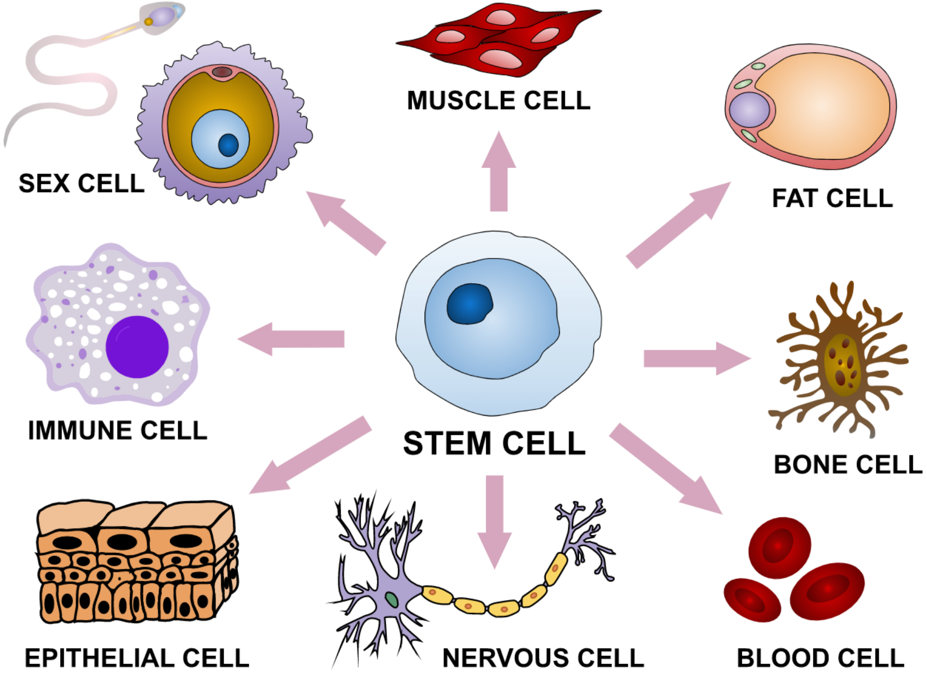
Community generated, multi-omic, open data

This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more here

https://data.humancellatlas.org/

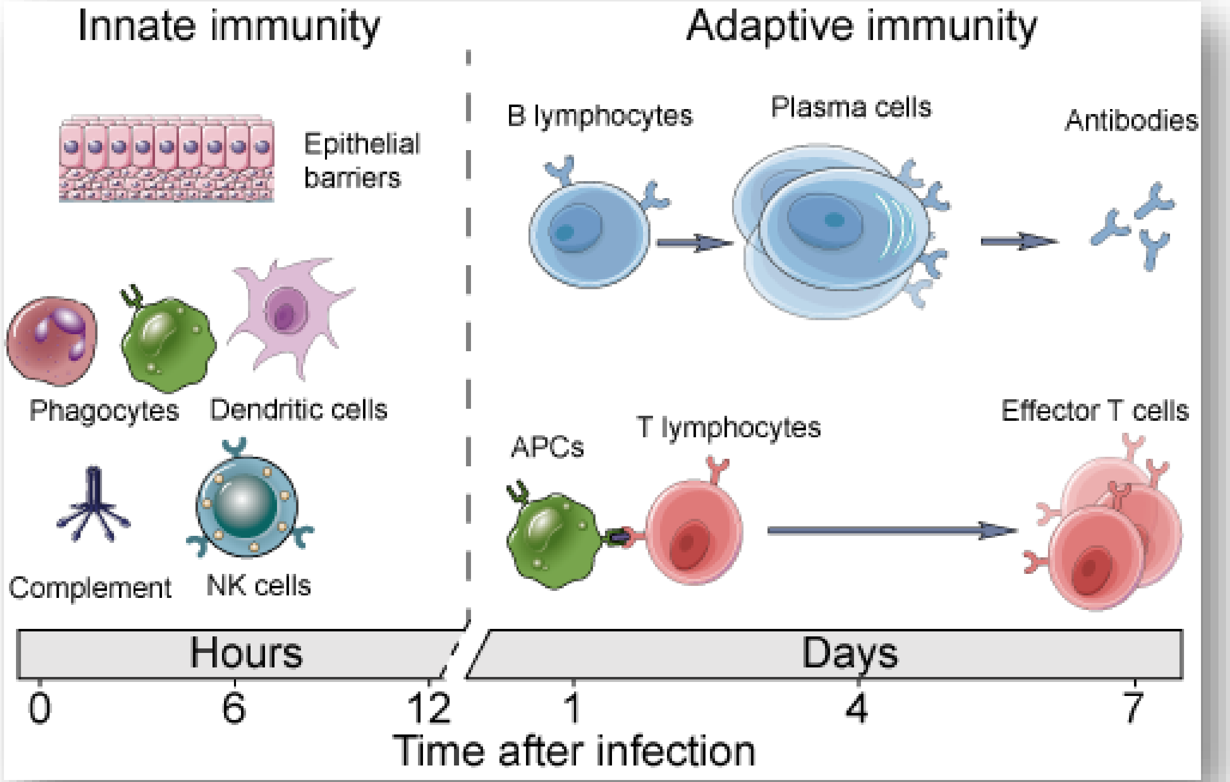
On June 15th 2023 Aviv Regev received the L’Oreal-Unesco for women in science award *“for her pioneering work applying mathematics and computer science to revolutionize cell biology.“*

Cells are specialized and adapted to their neighborhood



From Wikicommons
Date 10 June 2019
Author Haileyfournier

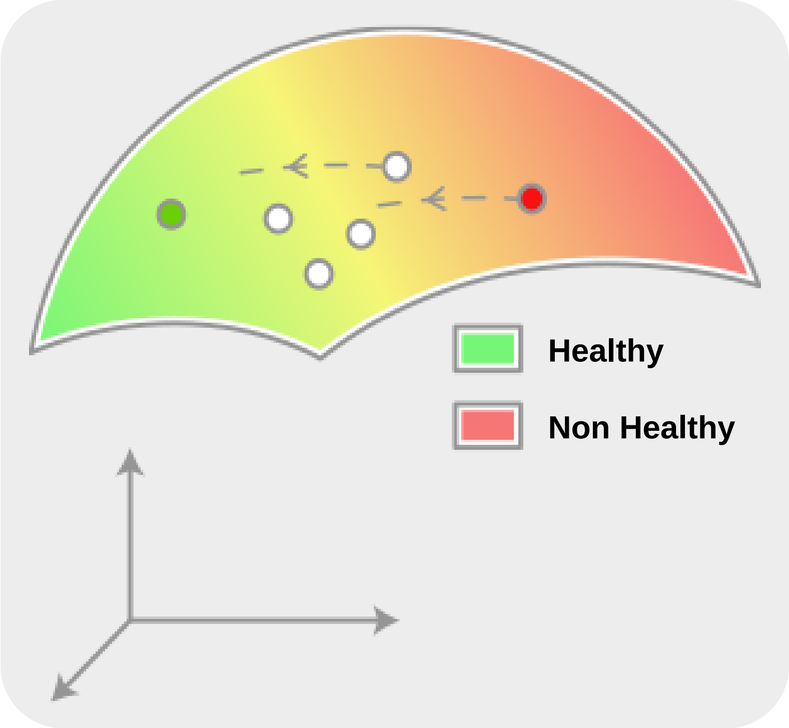
The Human Immune System



https://www.creative-diagnostics.com/innate-and-adaptive-immunity.htm

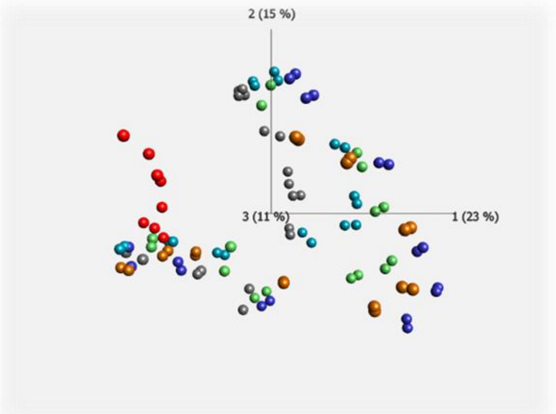
The Immunity State Space

The Immunity State Space is constructed from « precision measurements » of molecular, cellular and higher order biological activities connected with the Immune System.
The goal is to first identify domains of health and non-health and then find interventions to push patients back to healthy states.

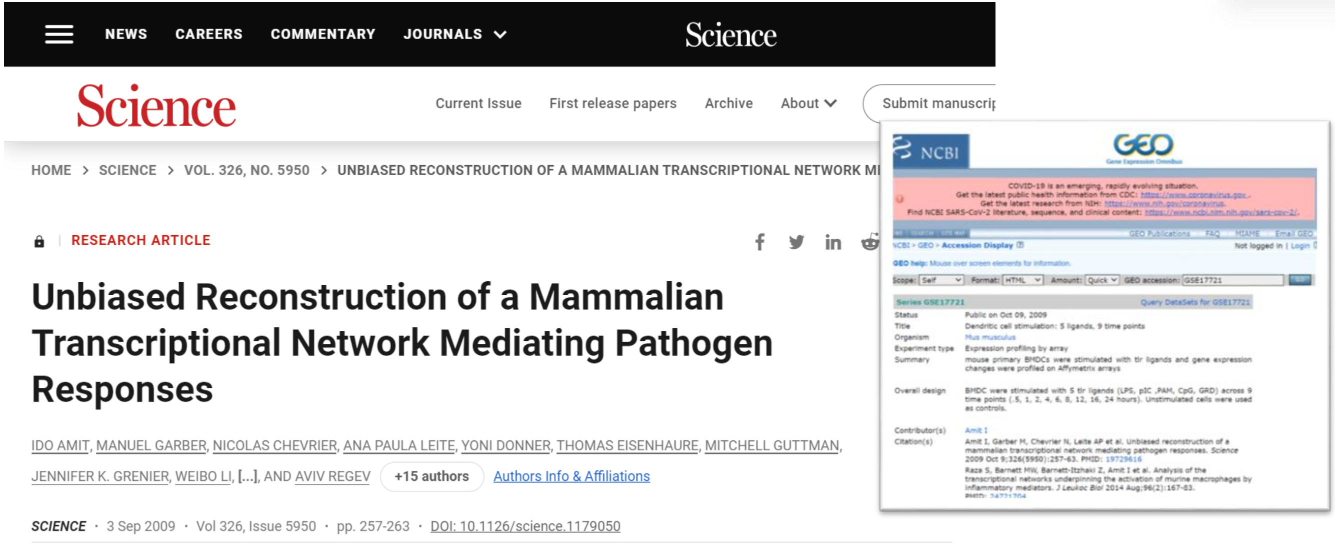


Computational modeling experience tells us that « *Life operates in low dimension* »

A typical example: Invitro stimulated mouse dendritic cells 0-0.5-2-4-6-8-12-16- 24h bulk mRNA expression for all 10716 genes (two samples per time point) result in « clear » 1 dimensional trajectories embedded in low (<7) dimensional space:



🟥 Control
🟩 Cpg
🟦 Grd
🟦 Lps
⬜ Pam
🟧 Pic



Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses

IDO AMIT, MANUEL GARBER, NICOLAS CHEVRIER, ANA PAULA LEITE, YONI DONNER, THOMAS EISENHAURE, MITCHELL GUTTMAN, JENNIFER K. GRENIER, WEIBO LI, [...], AND AVIV REGEV +15 authors Authors Info & Affiliations

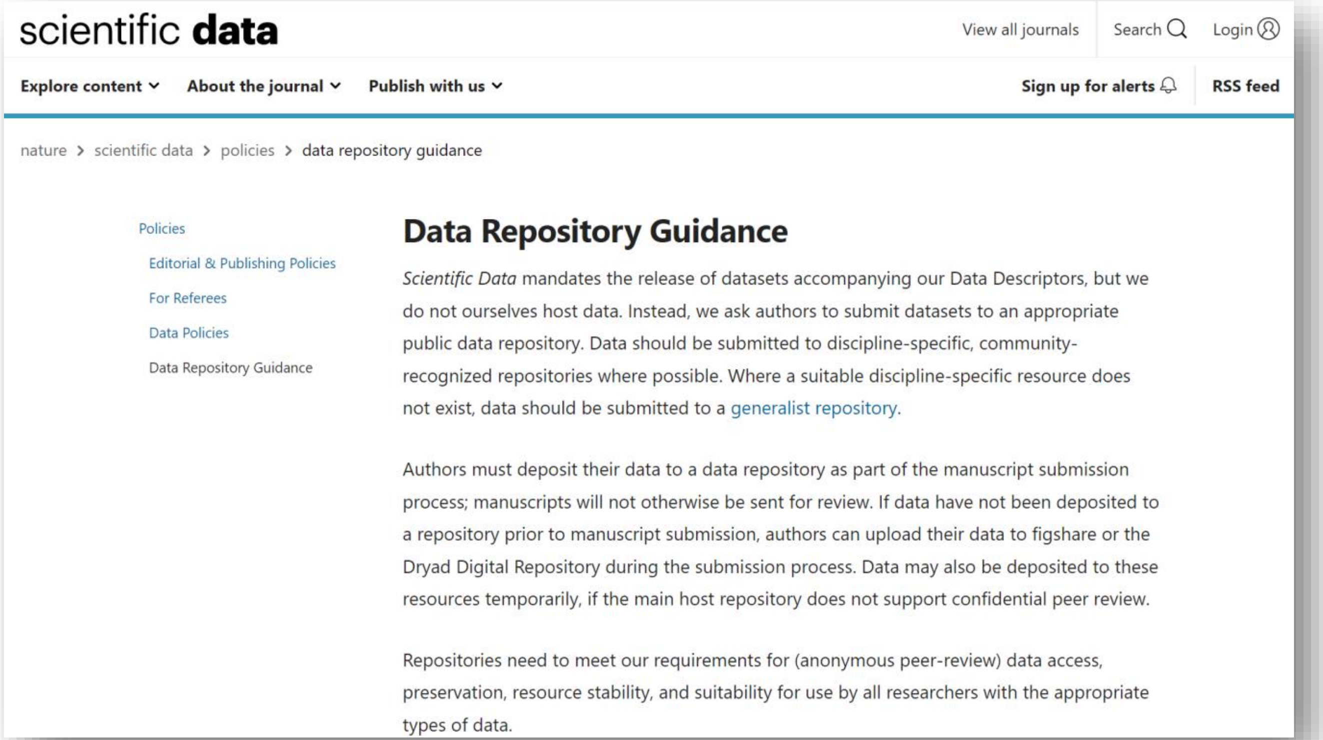
SCIENCE · 3 Sep 2009 · Vol 326, Issue 5950 · pp. 257-263 · DOI: 10.1126/science.1179050

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE17721

Series GSE17721	Query DataSets for GSE17721
Status	Public on Oct 09, 2009
Title	Dendritic cell stimulation: 5 ligands, 9 time points
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	mouse primary BMDCs were stimulated with tlr ligands and gene expression changes were profiled on Affymetrix arrays
Overall design	BMDC were stimulated with 5 tlr ligands (LPS, pIC ,PAM, CpG, GRD) across 9 time points (.5, 1, 2, 4, 6, 8, 12, 16, 24 hours). Unstimulated cells were used as controls.
Contributor(s)	Amit I
Citation(s)	Amit I, Garber M, Chevrier N, Leite AP et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 2009 Oct 9;326(5950):257-63. PMID: 19729616 Raza S, Barnett MW, Barnett-Itzhaki Z, Amit I et al. Analysis of the transcriptional networks underpinning the activation of murine macrophages by inflammatory mediators. *J Leukoc Biol* 2014 Aug;96(2):167-83. PMID: 24721704

Where to find and deposit biomedical data?



Data Repository Guidance

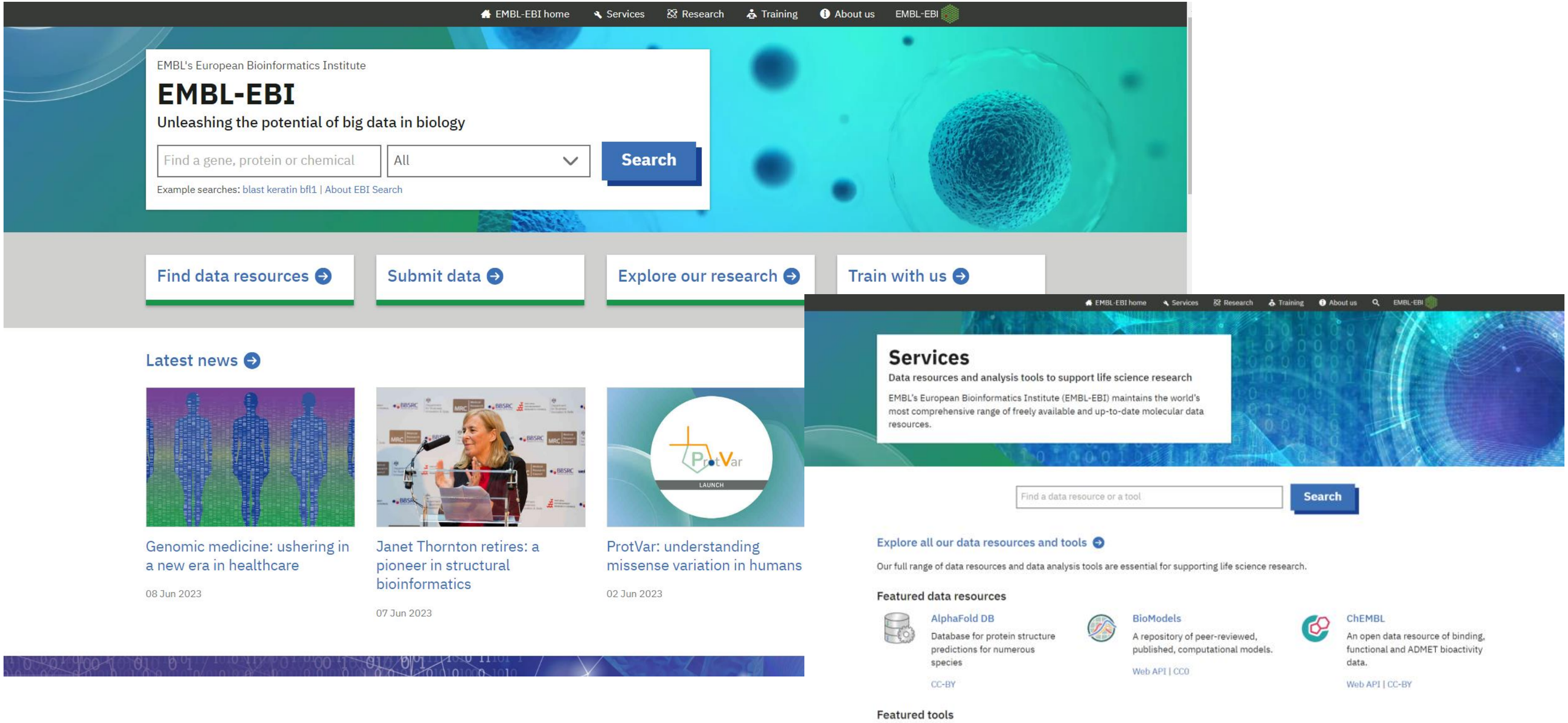
Scientific Data mandates the release of datasets accompanying our Data Descriptors, but we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible. Where a suitable discipline-specific resource does not exist, data should be submitted to a generalist repository.

Authors must deposit their data to a data repository as part of the manuscript submission process; manuscripts will not otherwise be sent for review. If data have not been deposited to a repository prior to manuscript submission, authors can upload their data to figshare or the Dryad Digital Repository during the submission process. Data may also be deposited to these resources temporarily, if the main host repository does not support confidential peer review.

Repositories need to meet our requirements for (anonymous peer-review) data access, preservation, resource stability, and suitability for use by all researchers with the appropriate types of data.

https://www.nature.com/sdata/policies/repositories

https://www.ebi.ac.uk/



EMBL-EBI home · Services · Research · Training · About us · EMBL-EBI

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical | All | **Search**

Example searches: blast keratin bfl1 | About EBI Search

Find data resources · Submit data · Explore our research · Train with us

Latest news

Genomic medicine: ushering in a new era in healthcare
08 Jun 2023

Janet Thornton retires: a pioneer in structural bioinformatics
07 Jun 2023

ProtVar: understanding missense variation in humans
02 Jun 2023

EMBL-EBI home · Services · Research · Training · About us · EMBL-EBI

Services

Data resources and analysis tools to support life science research

EMBL's European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Find a data resource or a tool | **Search**

Explore all our data resources and tools

Our full range of data resources and data analysis tools are essential for supporting life science research.

Featured data resources

AlphaFold DB
Database for protein structure predictions for numerous species
CC-BY

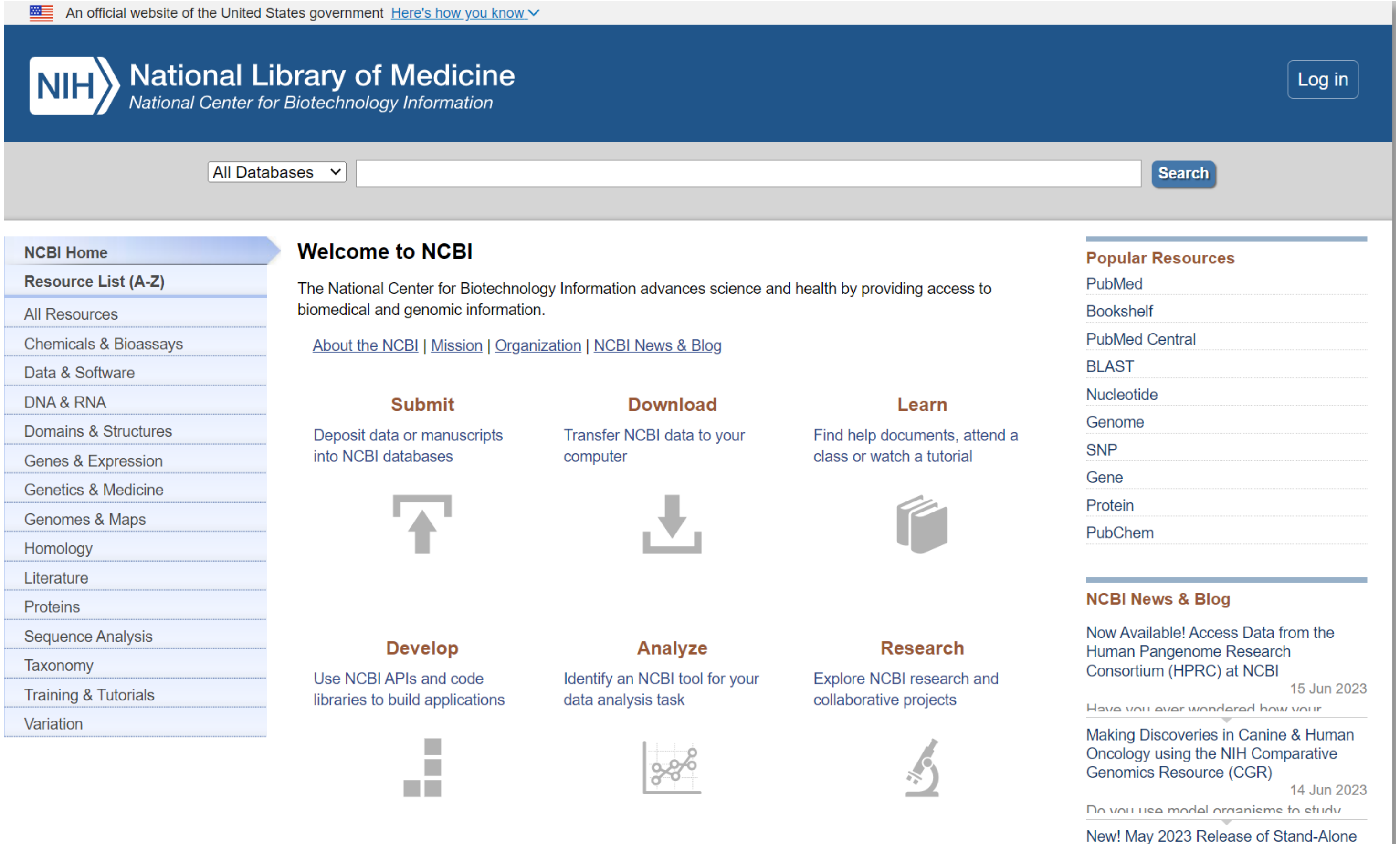
BioModels
A repository of peer-reviewed, published, computational models.
Web API | CC0

ChEMBL
An open data resource of binding, functional and ADMET bioactivity data.
Web API | CC-BY

Featured tools

https://www.ebi.ac.uk/services

https://www.ncbi.nlm.nih.gov/



An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases | Search

- **NCBI Home**
- **Resource List (A-Z)**
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

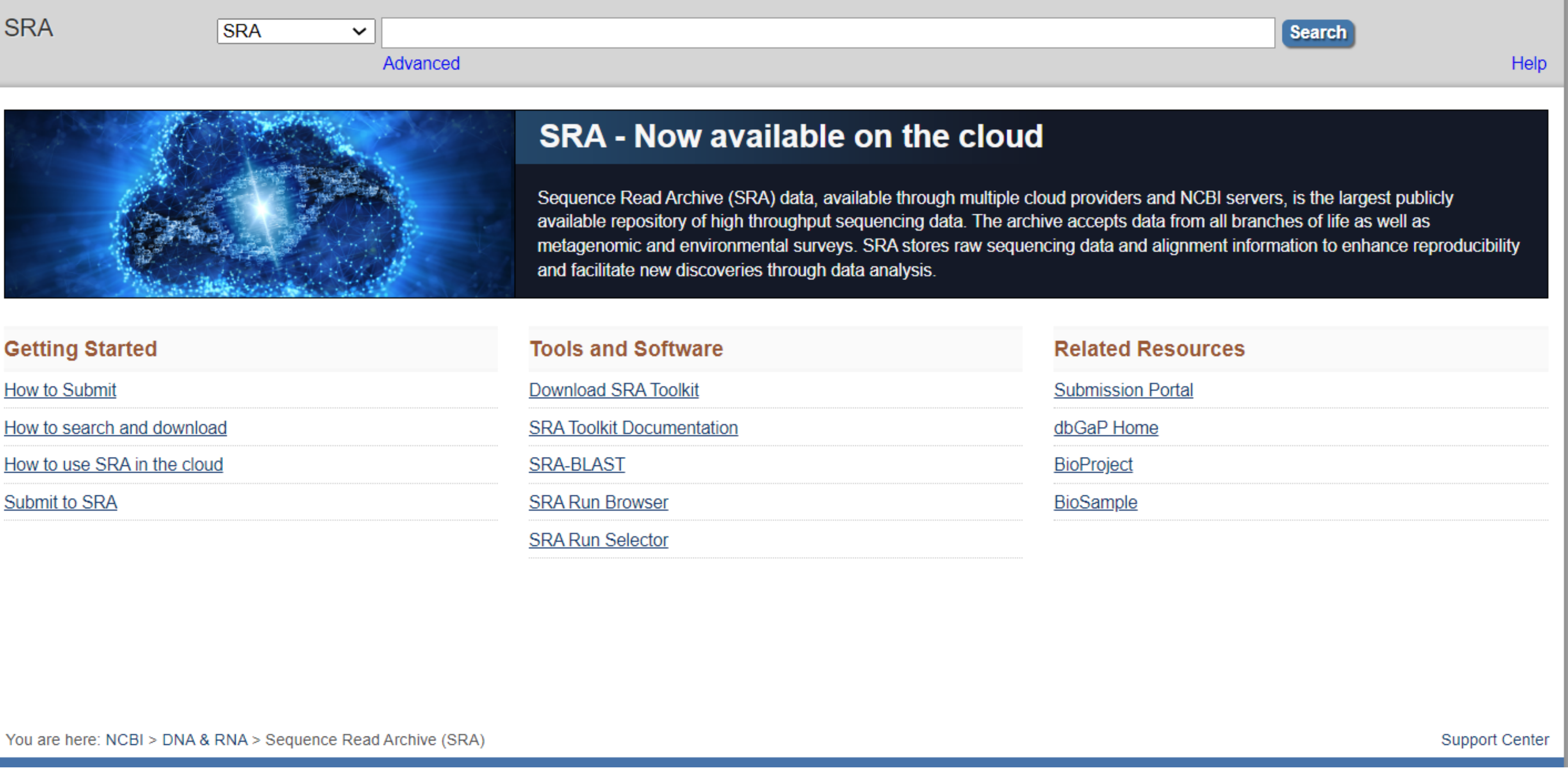
Popular Resources
- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Now Available! Access Data from the Human Pangenome Research Consortium (HPRC) at NCBI
15 Jun 2023
Have you ever wondered how your

Making Discoveries in Canine & Human Oncology using the NIH Comparative Genomics Resource (CGR)
14 Jun 2023
Do you use model organisms to study

New! May 2023 Release of Stand-Alone



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

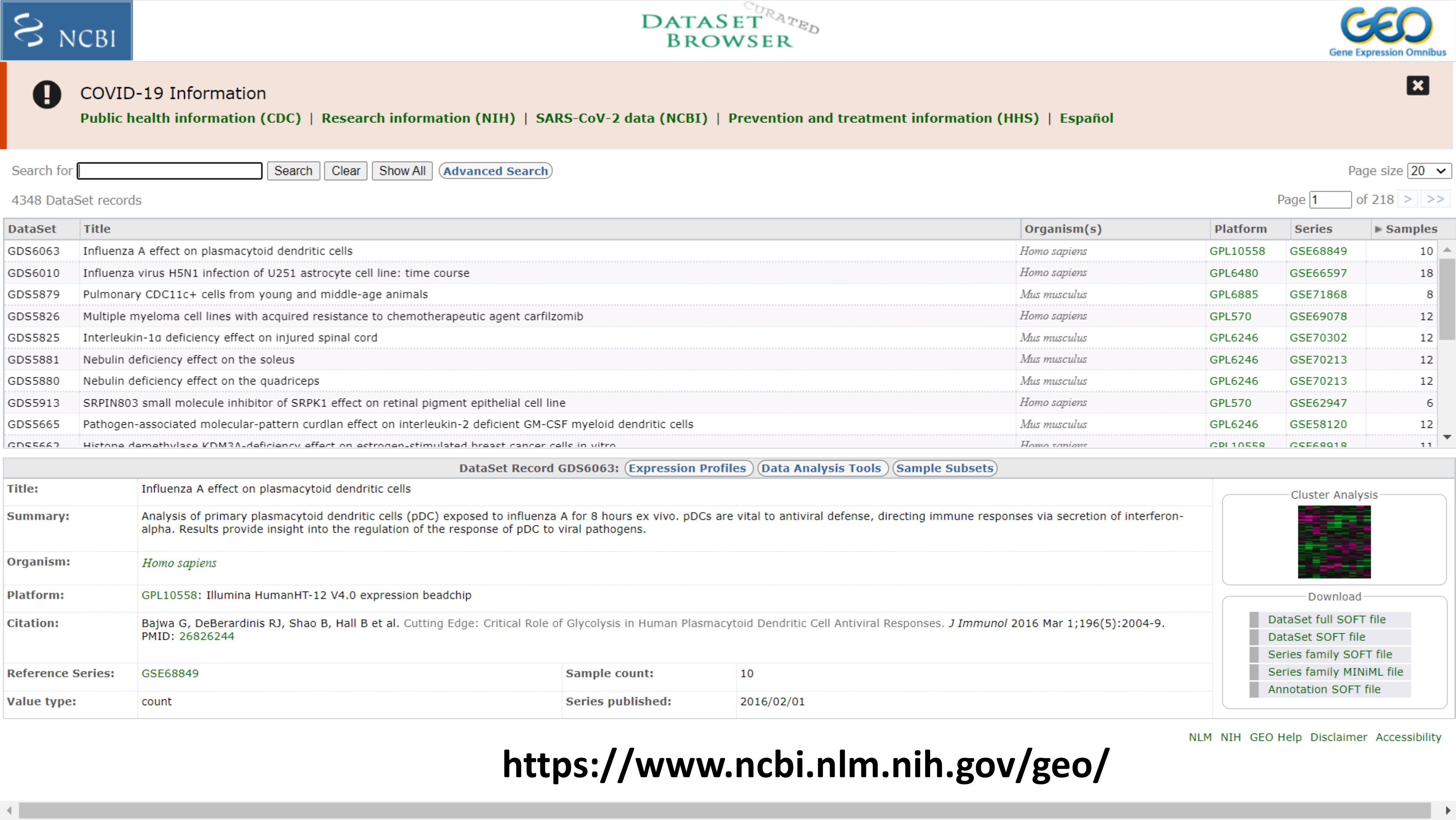
Related Resources

- Submission Portal
- dbGaP Home
- BioProject
- BioSample

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)

Support Center

https://www.ncbi.nlm.nih.gov/sra



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

4348 DataSet records

Page 1 of 218

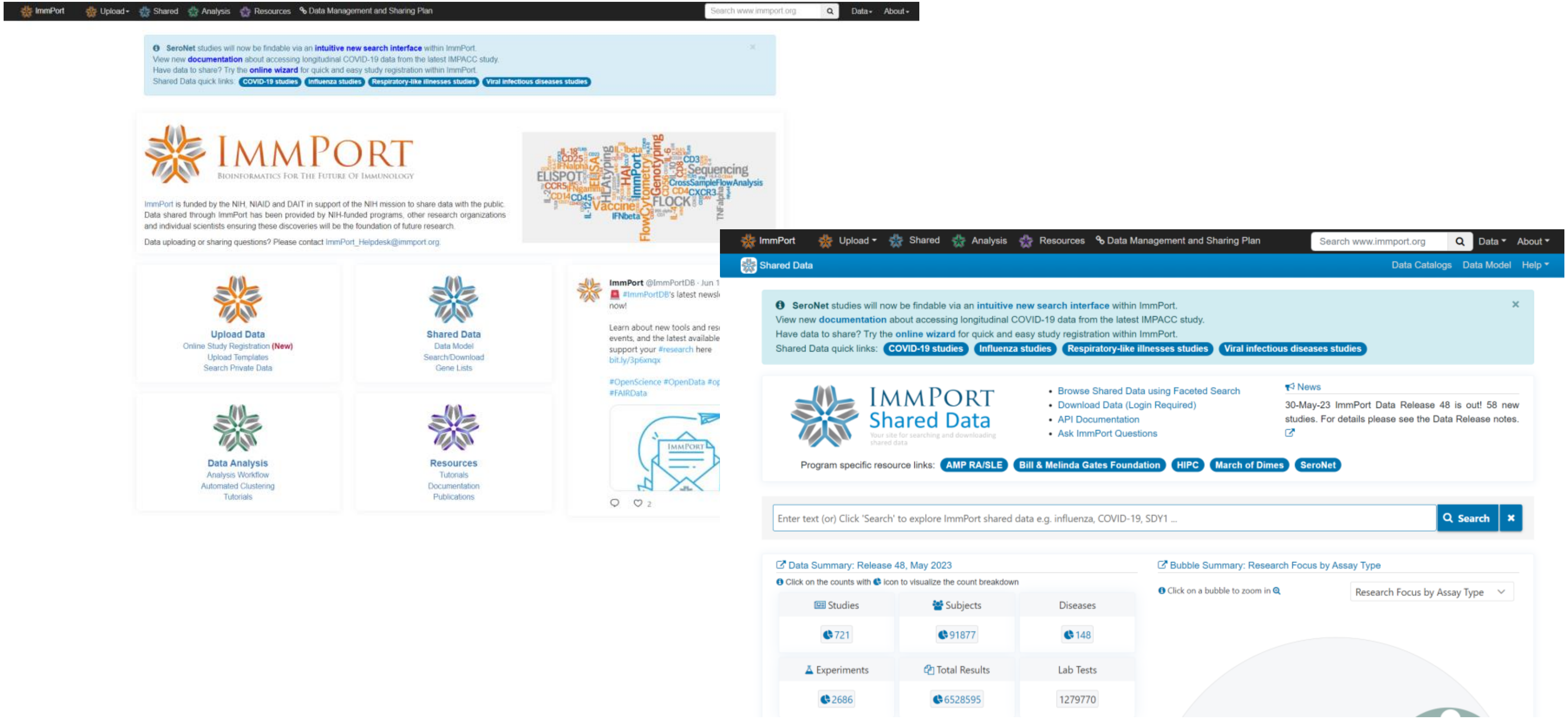
DataSet	Title	Organism(s)	Platform	Series	▸ Samples
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6010	Influenza virus H5N1 infection of U251 astrocyte cell line: time course	*Homo sapiens*	GPL6480	GSE66597	18
GDS5879	Pulmonary CDC11c+ cells from young and middle-age animals	*Mus musculus*	GPL6885	GSE71868	8
GDS5826	Multiple myeloma cell lines with acquired resistance to chemotherapeutic agent carfilzomib	*Homo sapiens*	GPL570	GSE69078	12
GDS5825	Interleukin-1α deficiency effect on injured spinal cord	*Mus musculus*	GPL6246	GSE70302	12
GDS5881	Nebulin deficiency effect on the soleus	*Mus musculus*	GPL6246	GSE70213	12
GDS5880	Nebulin deficiency effect on the quadriceps	*Mus musculus*	GPL6246	GSE70213	12
GDS5913	SRPIN803 small molecule inhibitor of SRPK1 effect on retinal pigment epithelial cell line	*Homo sapiens*	GPL570	GSE62947	6
GDS5665	Pathogen-associated molecular-pattern curdlan effect on interleukin-2 deficient GM-CSF myeloid dendritic cells	*Mus musculus*	GPL6246	GSE58120	12
GDS5662	Histone demethylase KDM3A-deficiency effect on estrogen-stimulated breast cancer cells in vitro	*Homo sapiens*	GPL10558	GSE68918	11

DataSet Record GDS6063: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Influenza A effect on plasmacytoid dendritic cells		
Summary:	Analysis of primary plasmacytoid dendritic cells (pDC) exposed to influenza A for 8 hours ex vivo. pDCs are vital to antiviral defense, directing immune responses via secretion of interferon-alpha. Results provide insight into the regulation of the response of pDC to viral pathogens.		
Organism:	*Homo sapiens*		
Platform:	GPL10558: Illumina HumanHT-12 V4.0 expression beadchip		
Citation:	Bajwa G, DeBerardinis RJ, Shao B, Hall B et al. Cutting Edge: Critical Role of Glycolysis in Human Plasmacytoid Dendritic Cell Antiviral Responses. *J Immunol* 2016 Mar 1;196(5):2004-9. PMID: 26826244		
Reference Series:	GSE68849	**Sample count:**	10
Value type:	count	**Series published:**	2016/02/01

https://www.ncbi.nlm.nih.gov/geo/

https://www.immport.org/home



SeroNet studies will now be findable via an **intuitive new search interface** within ImmPort.
View new **documentation** about accessing longitudinal COVID-19 data from the latest IMPACC study.
Have data to share? Try the **online wizard** for quick and easy study registration within ImmPort.
Shared Data quick links: **COVID-19 studies** **Influenza studies** **Respiratory-like illnesses studies** **Viral infectious diseases studies**

ImmPort is funded by the NIH, NIAID and DAIT in support of the NIH mission to share data with the public. Data shared through ImmPort has been provided by NIH-funded programs, other research organizations and individual scientists ensuring these discoveries will be the foundation of future research.

Data uploading or sharing questions? Please contact ImmPort_Helpdesk@immport.org.

Upload Data
Online Study Registration (New)
Upload Templates
Search Private Data

Shared Data
Data Model
Search/Download
Gene Lists

Data Analysis
Analysis Workflow
Automated Clustering
Tutorials

Resources
Tutorials
Documentation
Publications

ImmPort @ImmPortDB · Jun 1
🚨 #ImmPortDB's latest newsl
now!

Learn about new tools and res
events, and the latest available
support your #research here
bit.ly/3p6xnqx

#OpenScience #OpenData #op
#FAIRData

SeroNet studies will now be findable via an **intuitive new search interface** within ImmPort.
View new **documentation** about accessing longitudinal COVID-19 data from the latest IMPACC study.
Have data to share? Try the **online wizard** for quick and easy study registration within ImmPort.
Shared Data quick links: **COVID-19 studies** **Influenza studies** **Respiratory-like illnesses studies** **Viral infectious diseases studies**

- Browse Shared Data using Faceted Search
- Download Data (Login Required)
- API Documentation
- Ask ImmPort Questions

News

30-May-23 ImmPort Data Release 48 is out! 58 new studies. For details please see the Data Release notes.

Program specific resource links: **AMP RA/SLE** **Bill & Melinda Gates Foundation** **HIPC** **March of Dimes** **SeroNet**

Enter text (or) Click 'Search' to explore ImmPort shared data e.g. influenza, COVID-19, SDY1 ...

Data Summary: Release 48, May 2023

Click on the counts with icon to visualize the count breakdown

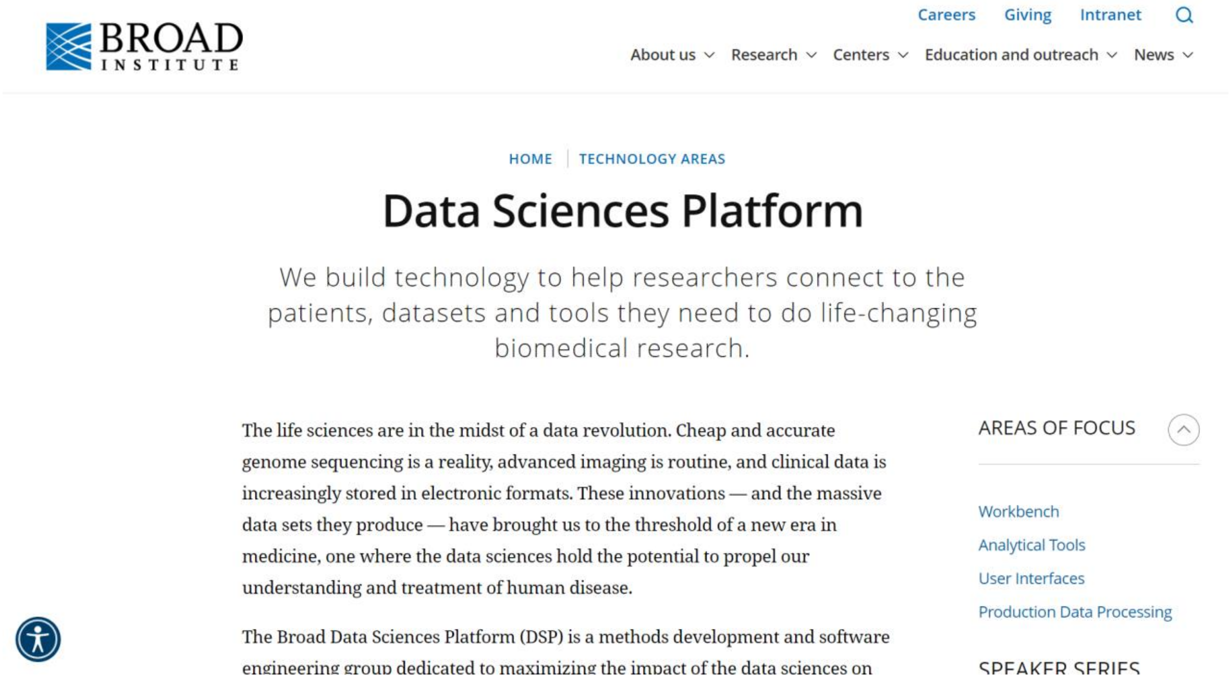
Studies	Subjects	Diseases
721	91877	148
Experiments	**Total Results**	**Lab Tests**
2686	6528595	1279770

Bubble Summary: Research Focus by Assay Type

Click on a bubble to zoom in

https://www.immport.org/shared/home

https://www.broadinstitute.org/



Data Sciences Platform

We build technology to help researchers connect to the patients, datasets and tools they need to do life-changing biomedical research.

The life sciences are in the midst of a data revolution. Cheap and accurate genome sequencing is a reality, advanced imaging is routine, and clinical data is increasingly stored in electronic formats. These innovations — and the massive data sets they produce — have brought us to the threshold of a new era in medicine, one where the data sciences hold the potential to propel our understanding and treatment of human disease.

The Broad Data Sciences Platform (DSP) is a methods development and software engineering group dedicated to maximizing the impact of the data sciences on

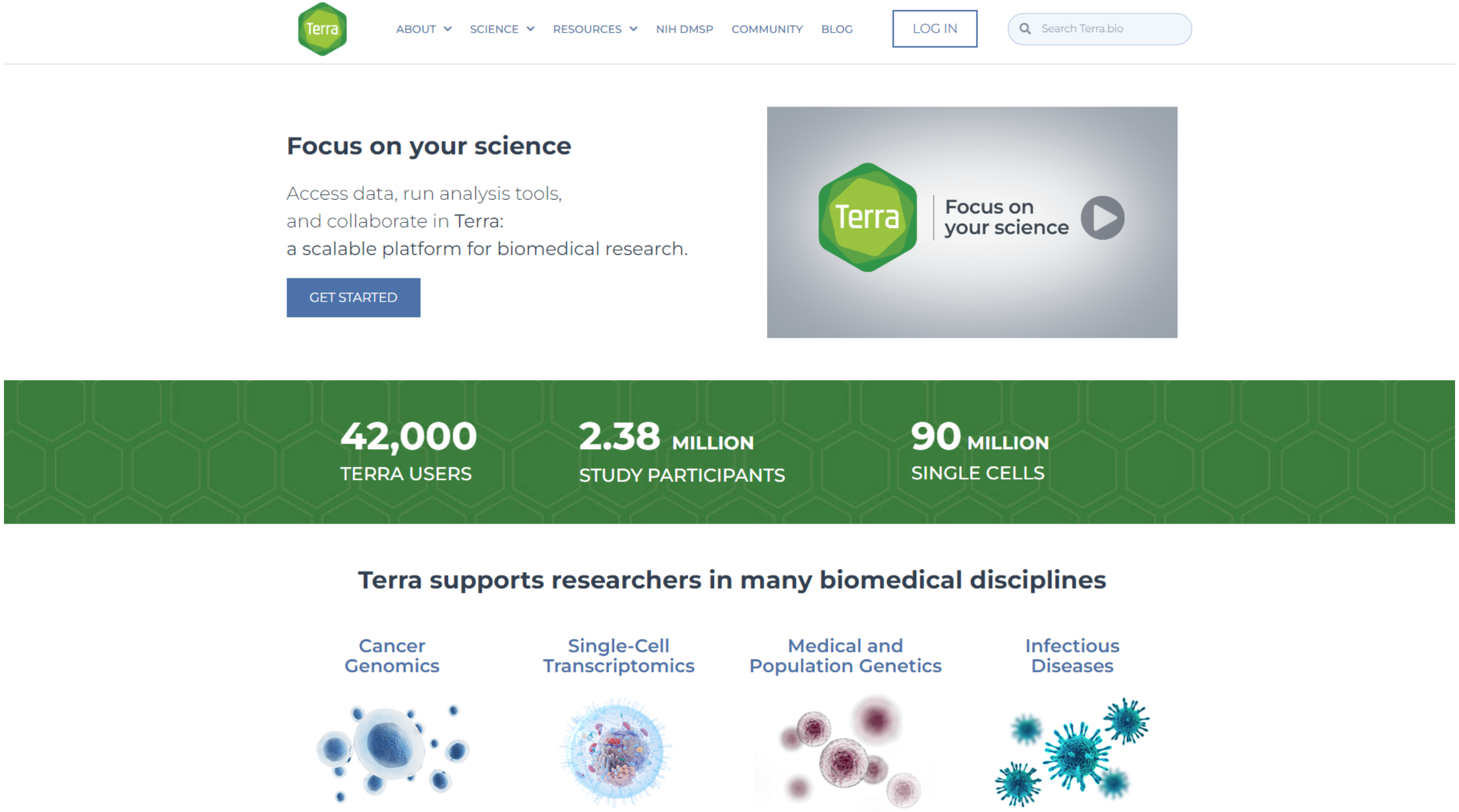
AREAS OF FOCUS

- Workbench
- Analytical Tools
- User Interfaces
- Production Data Processing

SPEAKER SERIES

https://www.broadinstitute.org/data-sciences-platform

https://terra.bio/ - A joint effort between the Broad Institute & Microsoft & Verily



Focus on your science

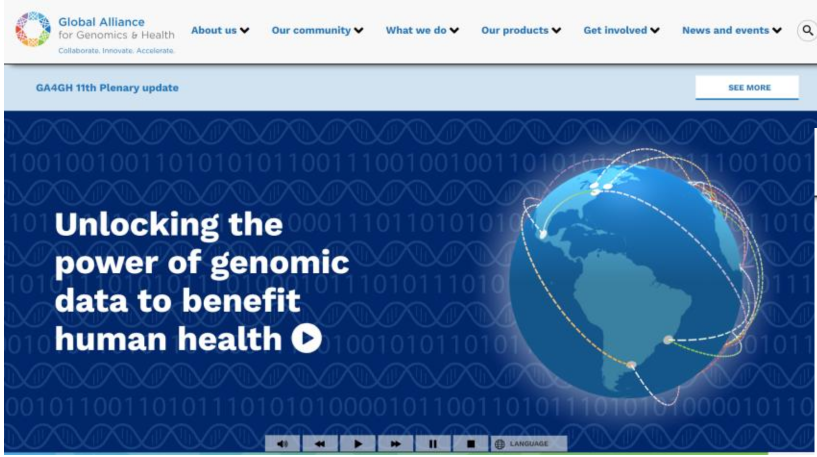
Access data, run analysis tools,
and collaborate in Terra:
a scalable platform for biomedical research.

42,000 TERRA USERS	2.38 MILLION STUDY PARTICIPANTS	90 MILLION SINGLE CELLS

Terra supports researchers in many biomedical disciplines

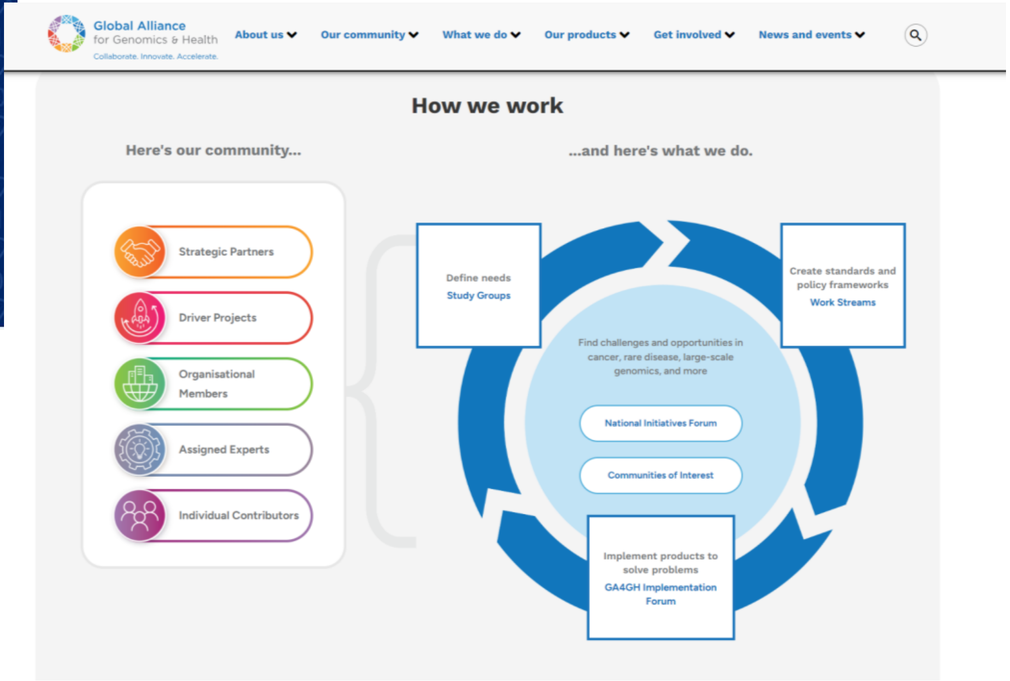
- Cancer Genomics
- Single-Cell Transcriptomics
- Medical and Population Genetics
- Infectious Diseases

https://www.ga4gh.org/



GA4GH 11th Plenary update

Unlocking the power of genomic data to benefit human health



How we work

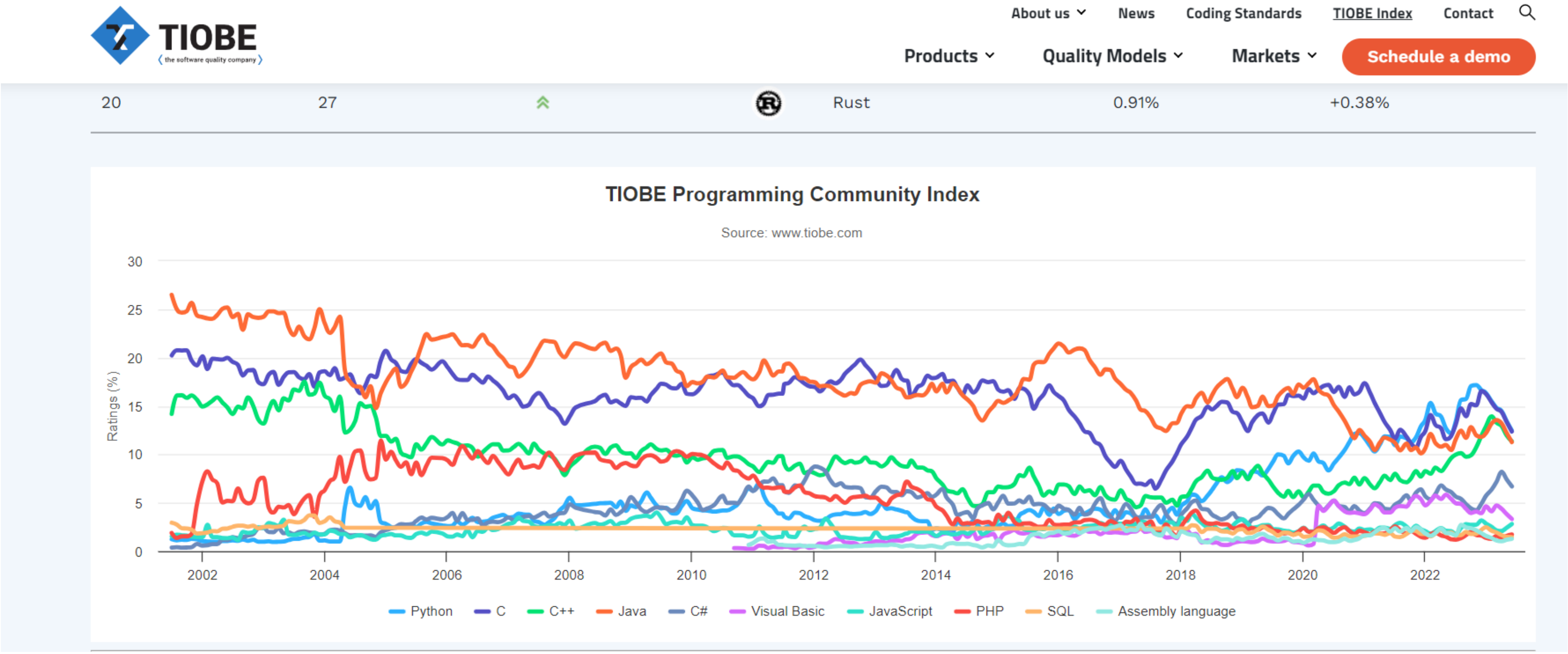
Here's our community...

...and here's what we do.

Find challenges and opportunities in cancer, rare disease, large-scale genomics, and more

Formed in 2013, the Global Alliance for Genomics and Health (GA4GH) unites an international community dedicated to advancing human health through genomic data. We build technical standards and policy frameworks and tools that will expand responsible, voluntary, and secure use of genomic and other related health data.

Interacting with Data - Software tools

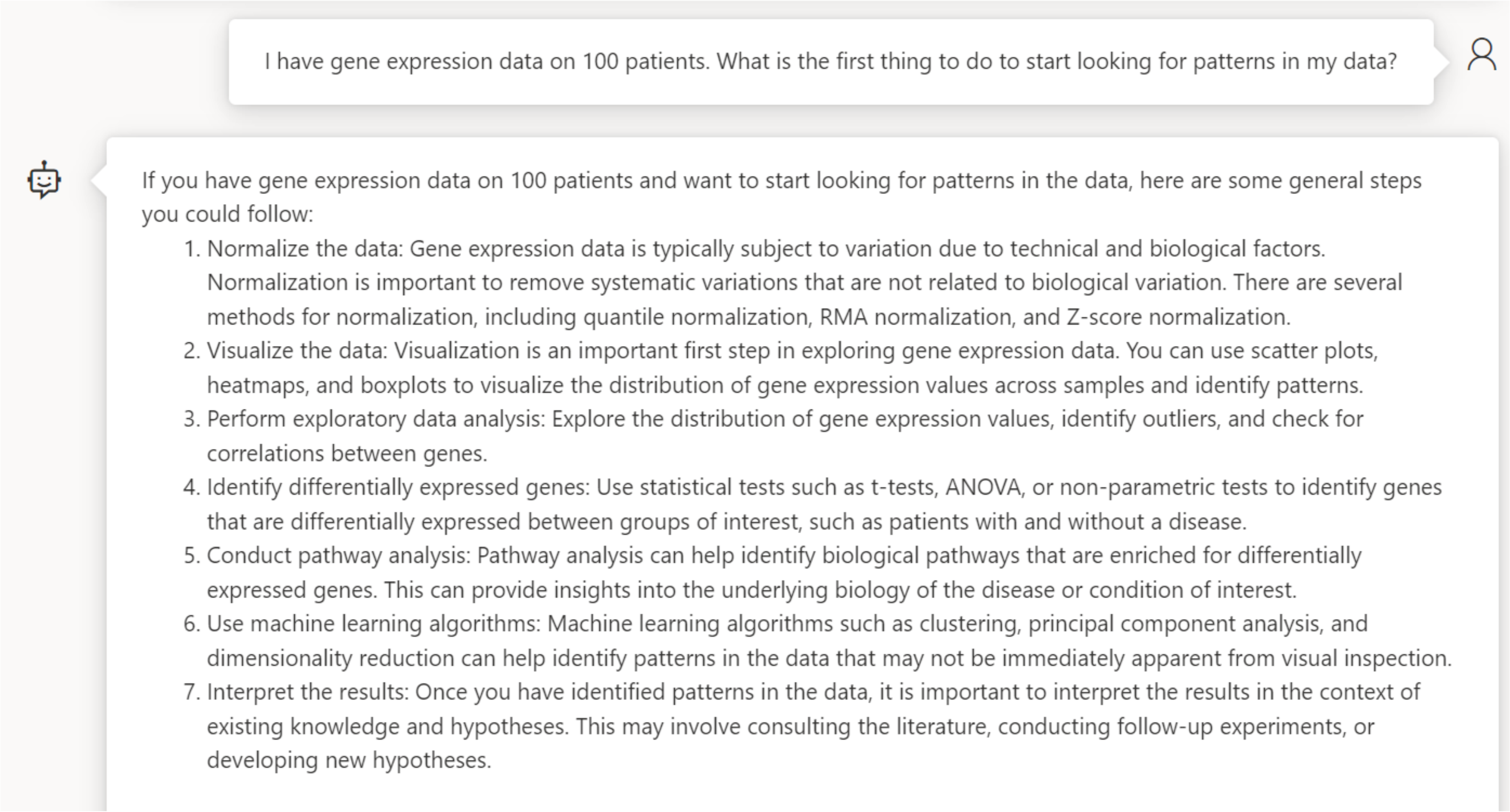


20	27	⇑	Ⓡ	Rust	0.91%	+0.38%

TIOBE Programming Community Index

Source: www.tiobe.com

https://www.tiobe.com/tiobe-index/

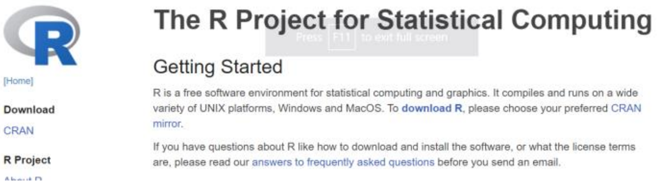


I have gene expression data on 100 patients. What is the first thing to do to start looking for patterns in my data?

If you have gene expression data on 100 patients and want to start looking for patterns in the data, here are some general steps you could follow:

1. Normalize the data: Gene expression data is typically subject to variation due to technical and biological factors. Normalization is important to remove systematic variations that are not related to biological variation. There are several methods for normalization, including quantile normalization, RMA normalization, and Z-score normalization.
2. Visualize the data: Visualization is an important first step in exploring gene expression data. You can use scatter plots, heatmaps, and boxplots to visualize the distribution of gene expression values across samples and identify patterns.
3. Perform exploratory data analysis: Explore the distribution of gene expression values, identify outliers, and check for correlations between genes.
4. Identify differentially expressed genes: Use statistical tests such as t-tests, ANOVA, or non-parametric tests to identify genes that are differentially expressed between groups of interest, such as patients with and without a disease.
5. Conduct pathway analysis: Pathway analysis can help identify biological pathways that are enriched for differentially expressed genes. This can provide insights into the underlying biology of the disease or condition of interest.
6. Use machine learning algorithms: Machine learning algorithms such as clustering, principal component analysis, and dimensionality reduction can help identify patterns in the data that may not be immediately apparent from visual inspection.
7. Interpret the results: Once you have identified patterns in the data, it is important to interpret the results in the context of existing knowledge and hypotheses. This may involve consulting the literature, conducting follow-up experiments, or developing new hypotheses.

What tools do I use?
→ *Python* and *R are* still dominating in bioinformatics



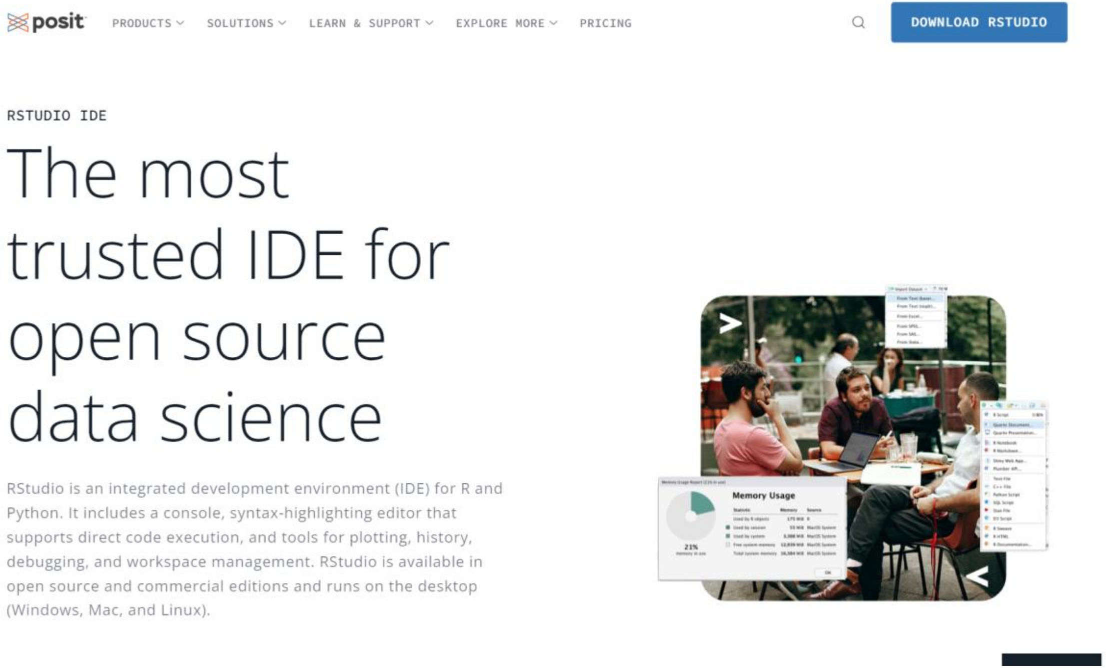
The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

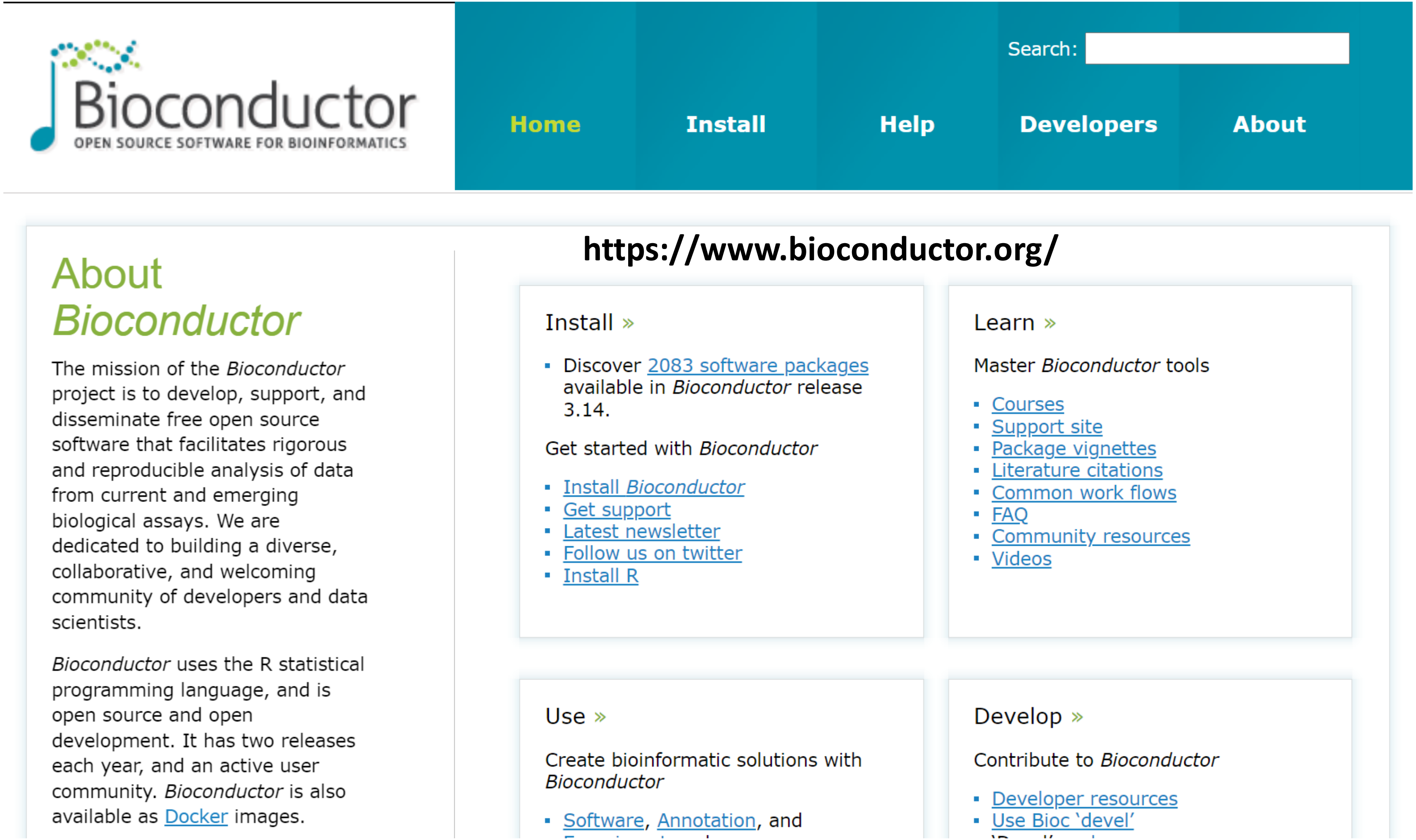
https://www.r-project.org/



RSTUDIO IDE

The most trusted IDE for open source data science

RStudio is an integrated development environment (IDE) for R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging, and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux).



https://www.bioconductor.org/

About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as Docker images.

Install »

- Discover 2083 software packages available in *Bioconductor* release 3.14.

Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

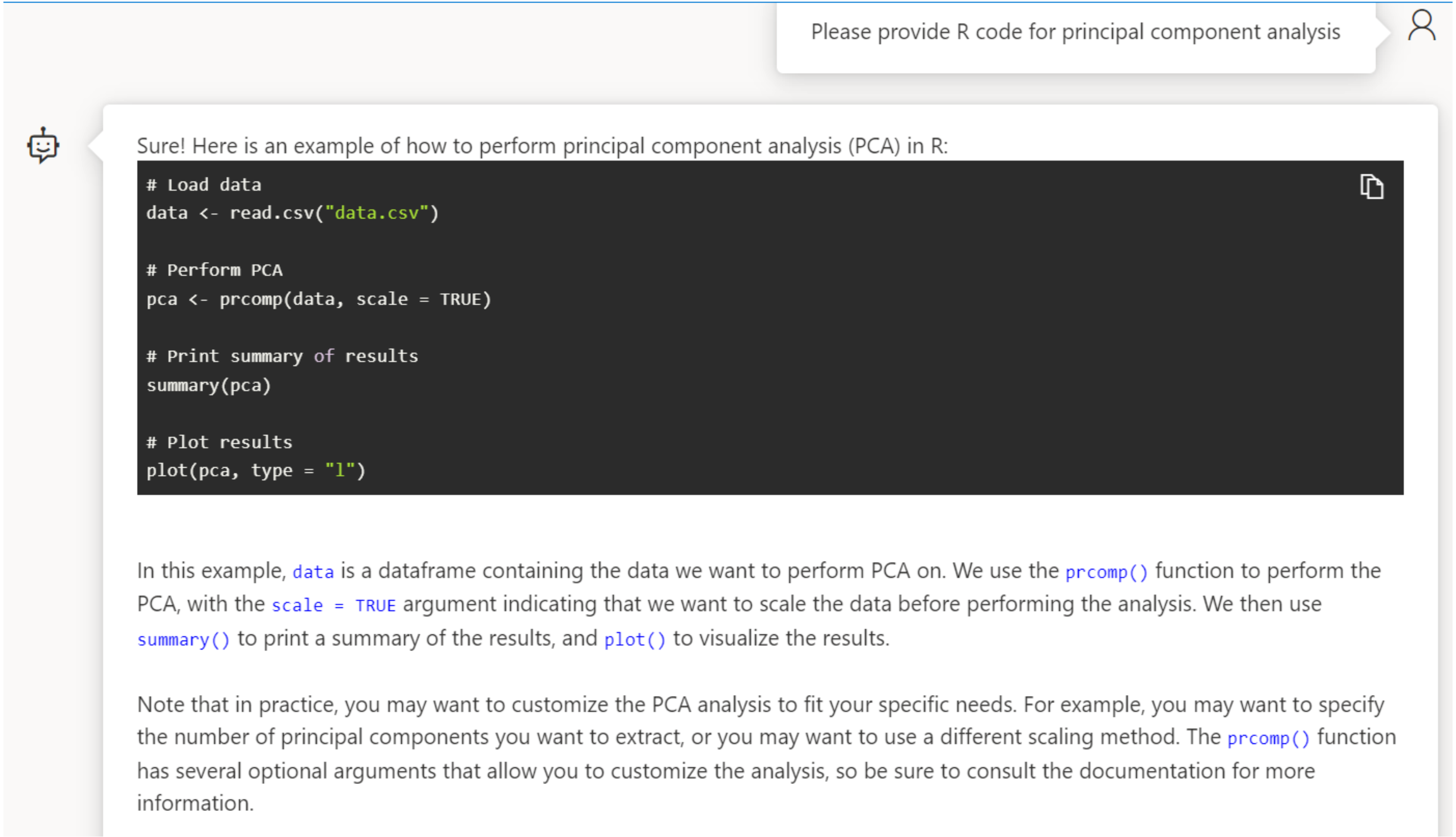
Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and

Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'



Please provide R code for principal component analysis

Sure! Here is an example of how to perform principal component analysis (PCA) in R:

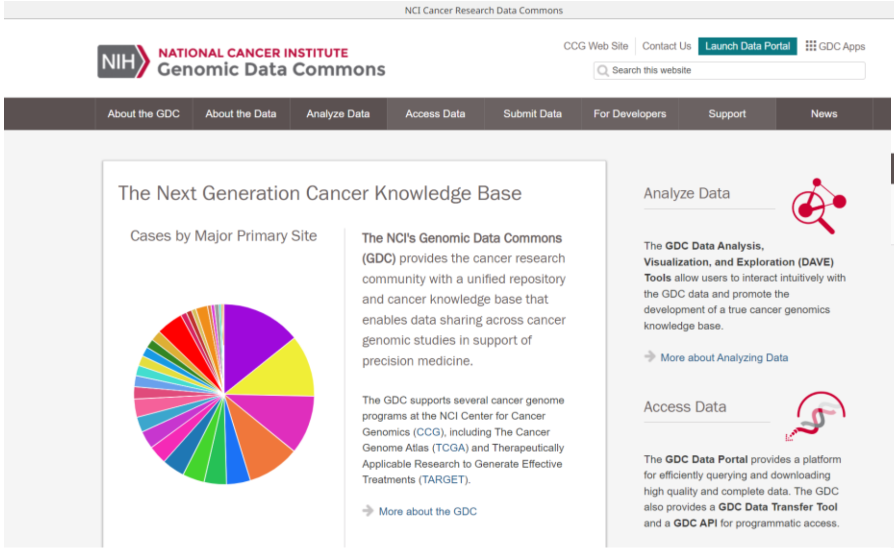
```
# Load data
data <- read.csv("data.csv")

# Perform PCA
pca <- prcomp(data, scale = TRUE)

# Print summary of results
summary(pca)

# Plot results
plot(pca, type = "l")
```

In this example, `data` is a dataframe containing the data we want to perform PCA on. We use the `prcomp()` function to perform the PCA, with the `scale = TRUE` argument indicating that we want to scale the data before performing the analysis. We then use `summary()` to print a summary of the results, and `plot()` to visualize the results.

Note that in practice, you may want to customize the PCA analysis to fit your specific needs. For example, you may want to specify the number of principal components you want to extract, or you may want to use a different scaling method. The `prcomp()` function has several optional arguments that allow you to customize the analysis, so be sure to consult the documentation for more information.

Interactive data analysis using e.g. RShiny web-applications

https://gdc.cancer.gov/



The Next Generation Cancer Knowledge Base

Cases by Major Primary Site

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified repository and cancer knowledge base that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

More about the GDC

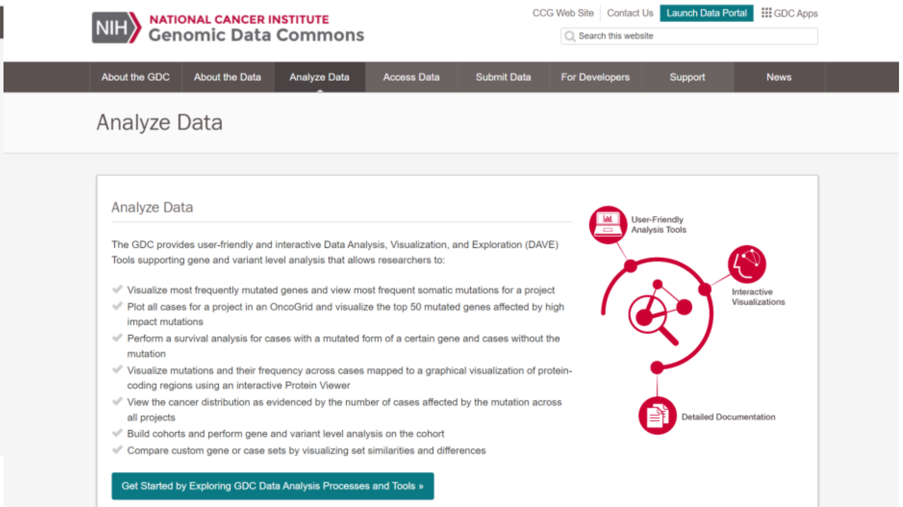
Analyze Data

The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

More about Analyzing Data

Access Data

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. The GDC also provides a **GDC Data Transfer Tool** and a **GDC API** for programmatic access.



Analyze Data

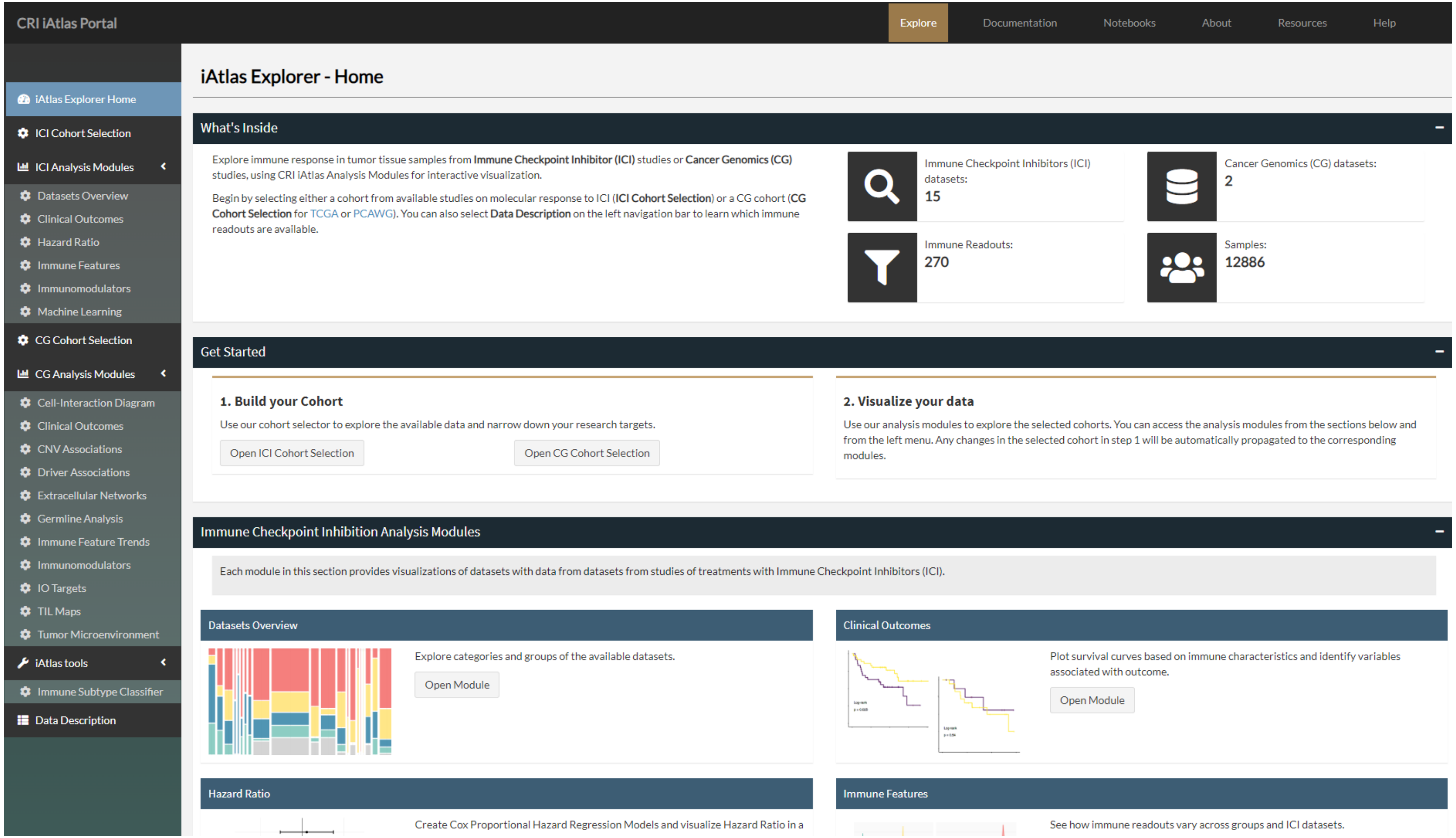
Analyze Data

The GDC provides user-friendly and interactive Data Analysis, Visualization, and Exploration (DAVE) Tools supporting gene and variant level analysis that allows researchers to:

- ✓ Visualize most frequently mutated genes and view most frequent somatic mutations for a project
- ✓ Plot all cases for a project in an OncoGrid and visualize the top 50 mutated genes affected by high impact mutations
- ✓ Perform a survival analysis for cases with a mutated form of a certain gene and cases without the mutation
- ✓ Visualize mutations and their frequency across cases mapped to a graphical visualization of protein-coding regions using an interactive Protein Viewer
- ✓ View the cancer distribution as evidenced by the number of cases affected by the mutation across all projects
- ✓ Build cohorts and perform gene and variant level analysis on the cohort
- ✓ Compare custom gene or case sets by visualizing set similarities and differences

https://gdc.cancer.gov/analyze-data

https://cri-iatlas.org/



iAtlas Explorer - Home

What's Inside

Explore immune response in tumor tissue samples from **Immune Checkpoint Inhibitor (ICI)** studies or **Cancer Genomics (CG)** studies, using CRI iAtlas Analysis Modules for interactive visualization.

Begin by selecting either a cohort from available studies on molecular response to ICI (**ICI Cohort Selection**) or a CG cohort (**CG Cohort Selection** for TCGA or PCAWG). You can also select **Data Description** on the left navigation bar to learn which immune readouts are available.

Immune Checkpoint Inhibitors (ICI) datasets: 15

Cancer Genomics (CG) datasets: 2

Immune Readouts: 270

Samples: 12886

Get Started

1. Build your Cohort

Use our cohort selector to explore the available data and narrow down your research targets.

2. Visualize your data

Use our analysis modules to explore the selected cohorts. You can access the analysis modules from the sections below and from the left menu. Any changes in the selected cohort in step 1 will be automatically propagated to the corresponding modules.

Immune Checkpoint Inhibition Analysis Modules

Each module in this section provides visualizations of datasets with data from datasets from studies of treatments with Immune Checkpoint Inhibitors (ICI).

Datasets Overview

Explore categories and groups of the available datasets.

Clinical Outcomes

Plot survival curves based on immune characteristics and identify variables associated with outcome.

Hazard Ratio

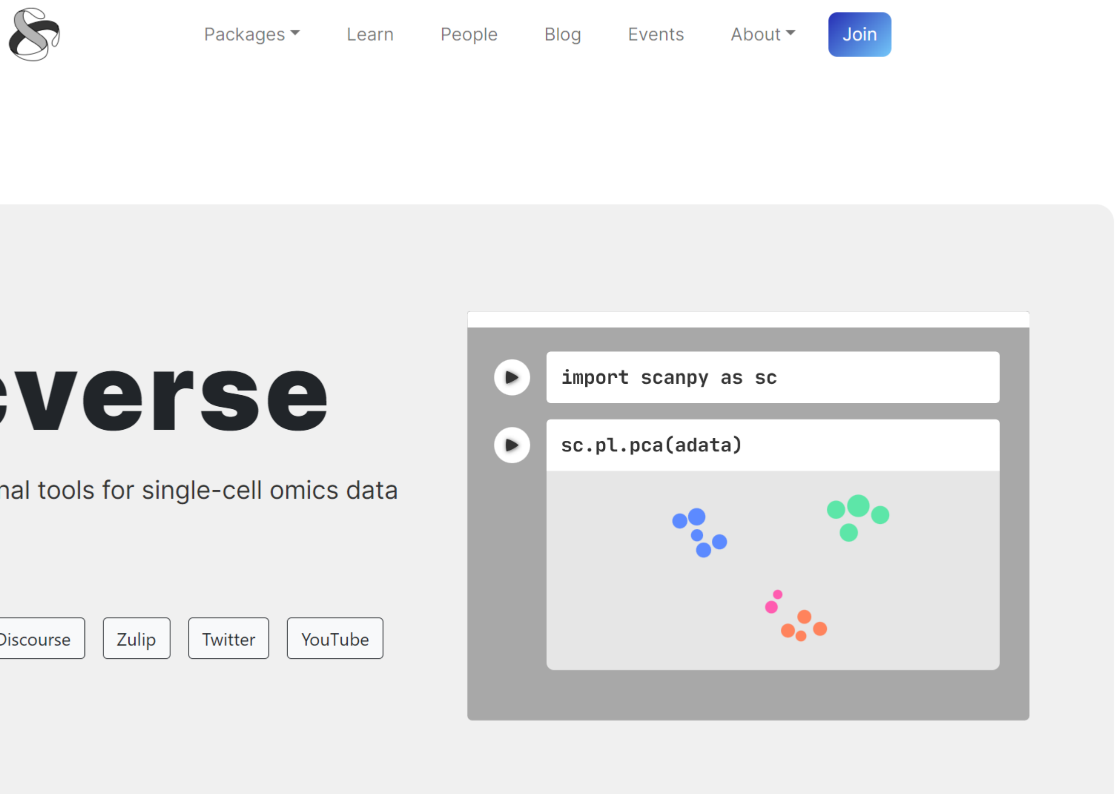
Create Cox Proportional Hazard Regression Models and visualize Hazard Ratio in a

Immune Features

See how immune readouts vary across groups and ICI datasets.

https://isb-cgc.shinyapps.io/iatlas/

scverse
A library of single cell analysis algorithms and computational models



scverse

Foundational tools for single-cell omics data analysis

```
import scanpy as sc
```

```
sc.pl.pca(adata)
```

https://scverse.org/

MISSION

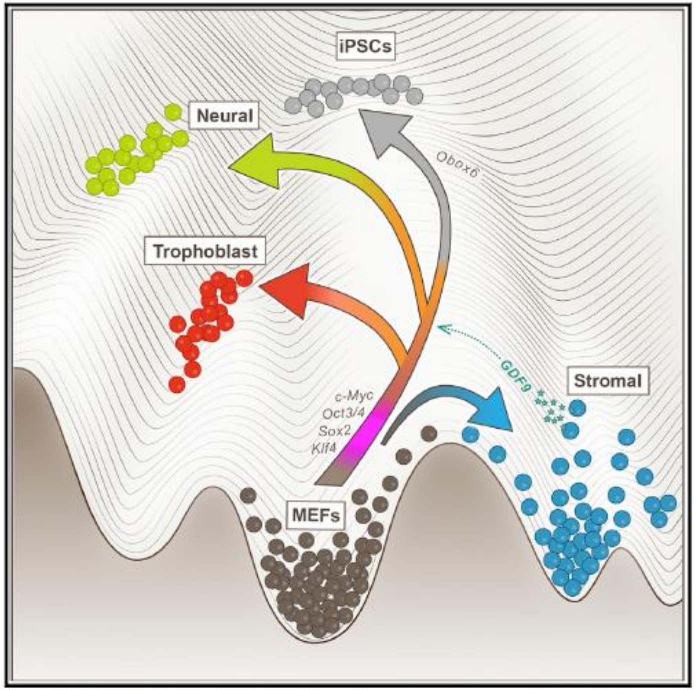
scverse is a consortium of foundational tools (mostly in Python) for omics data in life sciences. It has been founded to ensure the long-term maintenance of these core tools.

Cell

Resource

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Graphical Abstract



Authors

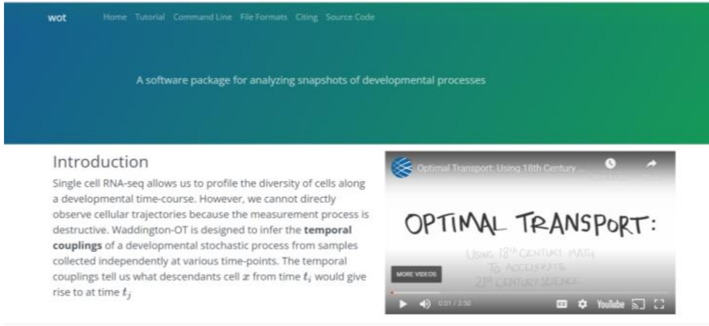
Geoffrey Schiebinger, Jian Shu,
Marcin Tabaka, ..., Rudolf Jaenisch,
Aviv Regev, Eric S. Lander

Correspondence

jianshu@broadinstitute.org (J.S.),
aregev@broadinstitute.org (A.R.),
lander@broadinstitute.org (E.S.L.)

In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.

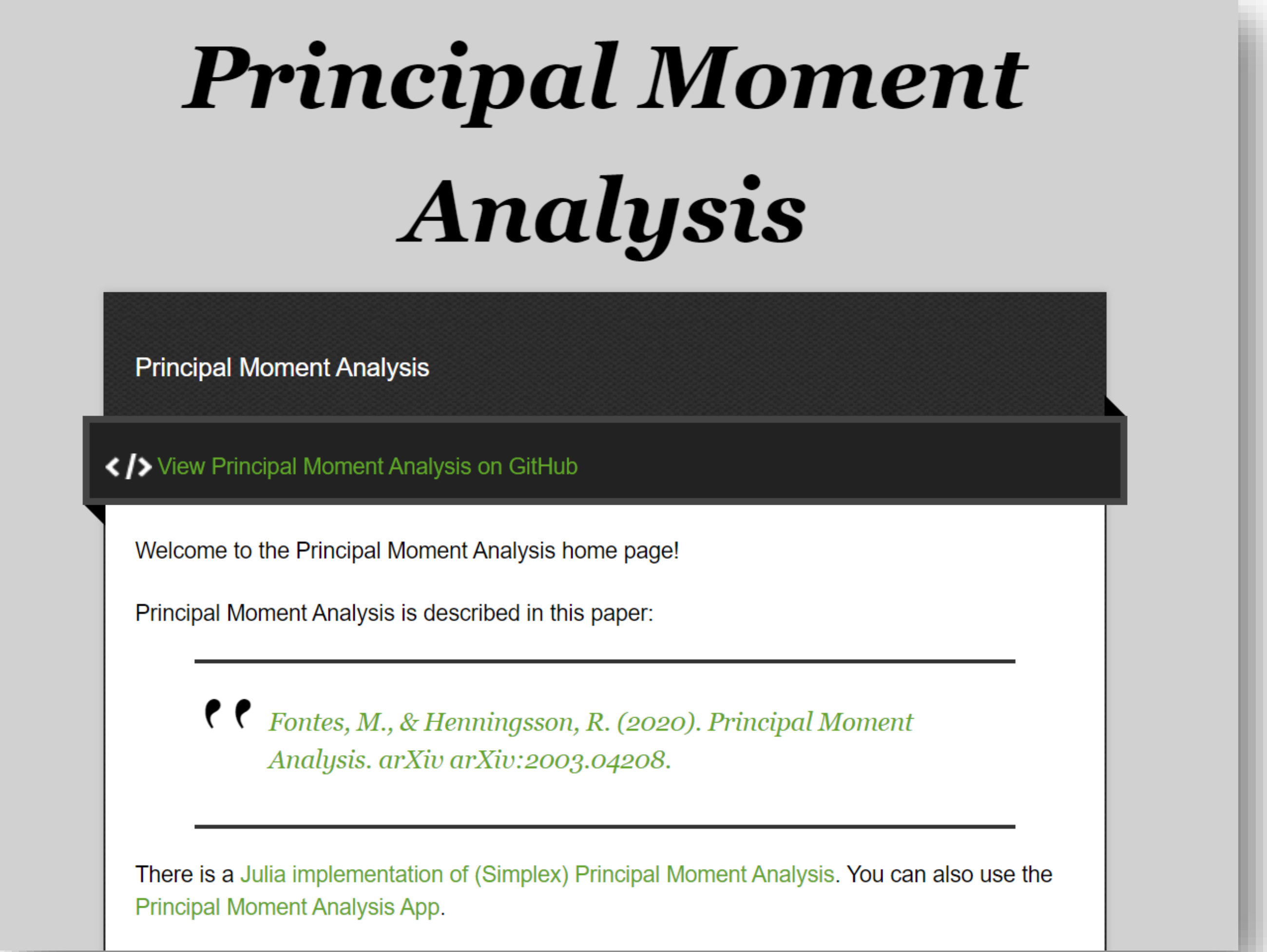


A software package for analyzing snapshots of developmental processes

Introduction

Single cell RNA-seq allows us to profile the diversity of cells along a developmental time-course. However, we cannot directly observe cellular trajectories because the measurement process is destructive. Waddington-OT is designed to infer the **temporal couplings** of a developmental stochastic process from samples collected independently at various time-points. The temporal couplings tell us what descendants cell x from time t_i would give rise to at time t_j

https://broadinstitute.github.io/wot/



Principal Moment Analysis

Principal Moment Analysis

View Principal Moment Analysis on GitHub

Welcome to the Principal Moment Analysis home page!

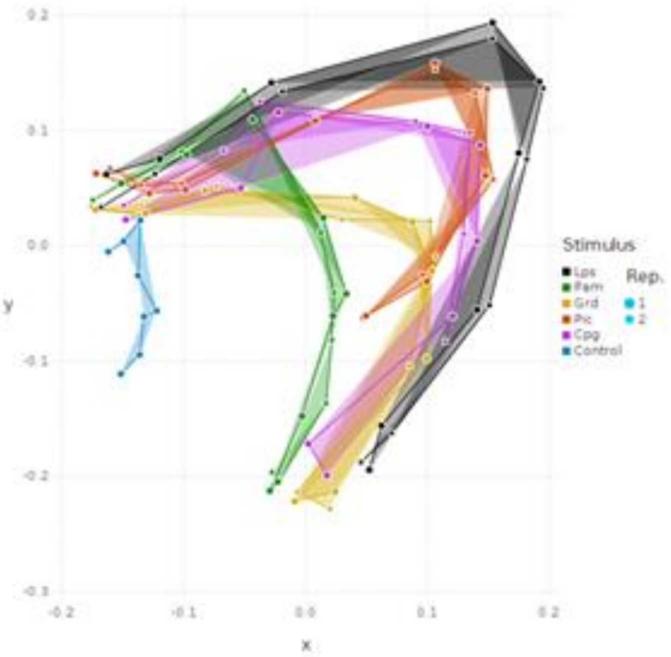
Principal Moment Analysis is described in this paper:

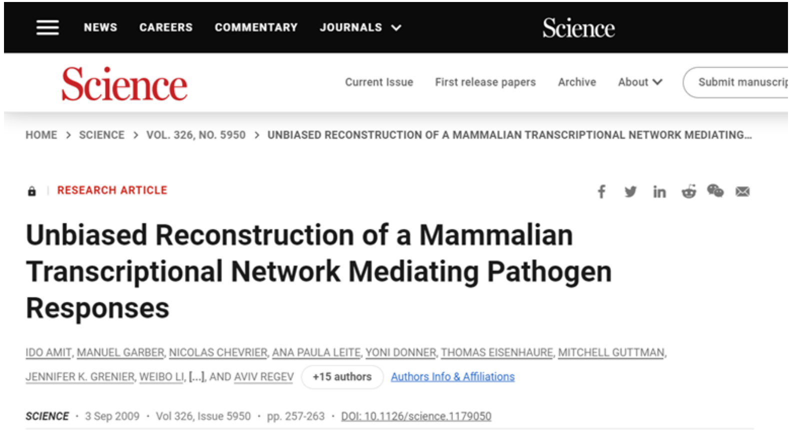
> *Fontes, M., & Henningsson, R. (2020). Principal Moment Analysis. arXiv arXiv:2003.04208.*

There is a Julia implementation of (Simplex) Principal Moment Analysis. You can also use the Principal Moment Analysis App.

PMA for dimension reduction and visualizations of State Spaces

https://principalmomentanalysis.github.io/




NEWS
CAREERS
COMMENTARY
JOURNALS
Science
Science
Current Issue
First release papers
Archive
About
Submit manuscrip
HOME > SCIENCE > VOL. 326, NO. 5950 > UNBIASED RECONSTRUCTION OF A MAMMALIAN TRANSCRIPTIONAL NETWORK MEDIATING...
+15 authors
Authors Info & Affiliations

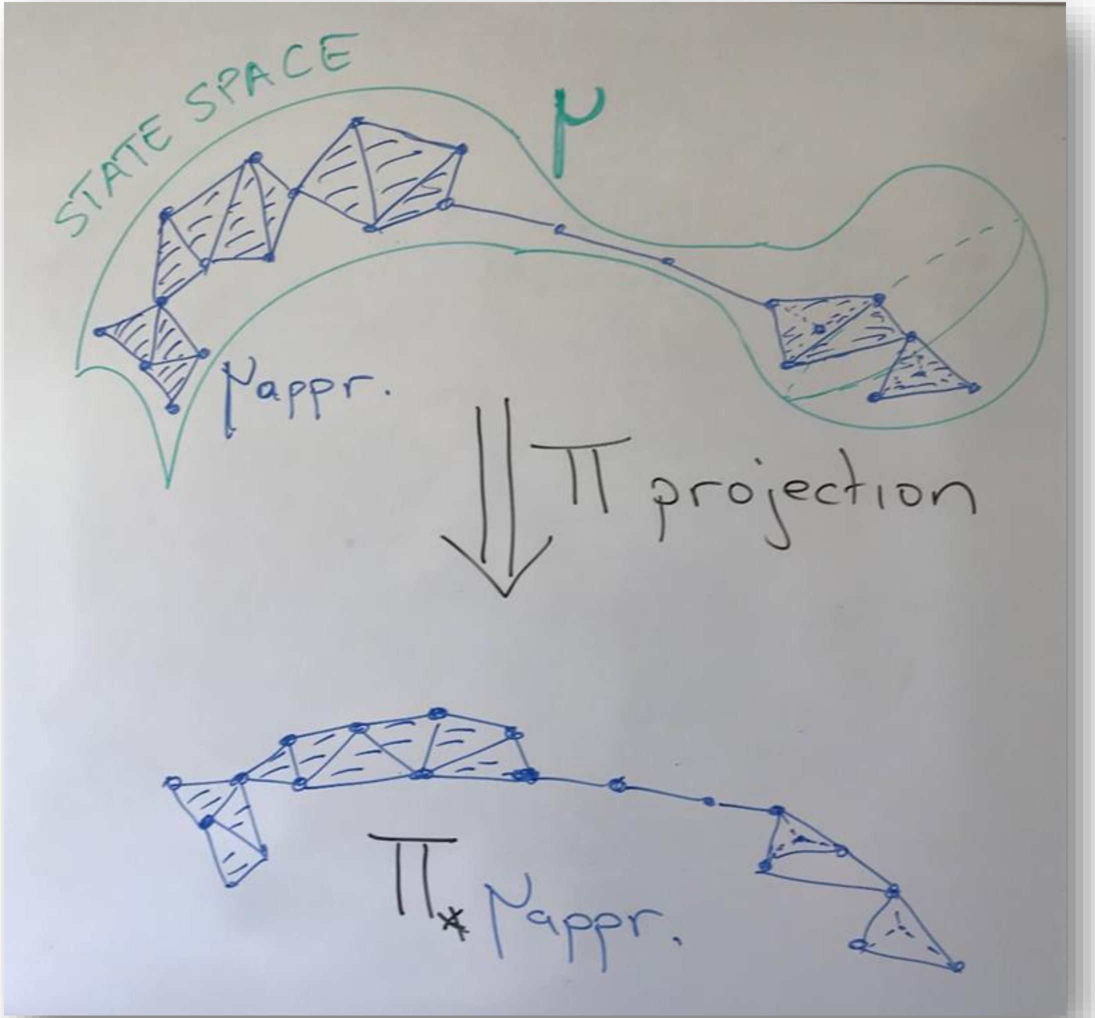
RESEARCH ARTICLE

Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses

IDO AMIT, MANUEL GARBER, NICOLAS CHEVRIER, ANA PAULA LEITE, YONI DONNER, THOMAS EISENHAURE, MITCHELL GUTTMAN, JENNIFER K. GRENIER, WEIBO LI, [...], AND AVIV REGEV

SCIENCE · 3 Sep 2009 · Vol 326, Issue 5950 · pp. 257-263 · DOI: 10.1126/science.1179050

PMA is a framework that incorporates the sample distribution and lower dimensional (noise reduced) approximations of it.



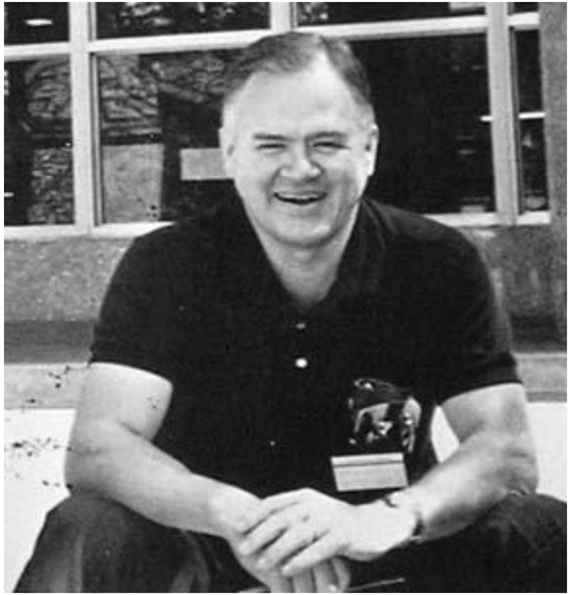
Take home messages concerning PMA

- Fast (*sample based PMA* is as fast as corresponding PCA)
- Robust (equivalent to local infinite pseudo sample bootstrap)
- Statistically and conceptually sound ("optimal" approximation of underlying probability measure respecting intrinsic local dimensionality and sampling density, quality etc)
- Possible to supervise using annotation information and expert knowledge
- Immediate: *PMA projection score* generalizes.
- Immediate: Kernel PMA

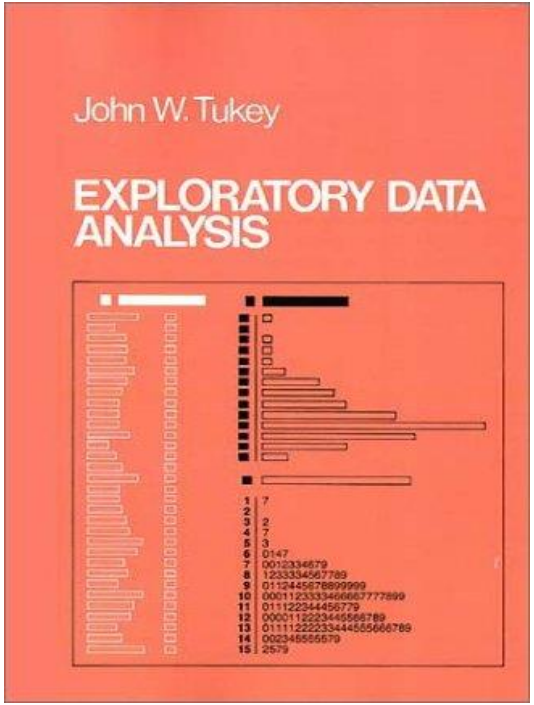
STATISTICAL LEARNING & VISUALIZATION → ACQUIRING INSIGHTS AROUND PATTERNS IN DATA RELATED TO BIOLOGICAL VARIATION

Design and Photo: Emilia Fontes 2016

Exploratory Data Analysis vs Confirmatory Data Analysis



John Wilder Tukey (1915-2000)
Inventor of the FFT, the Box plot and the word "bit".



John W. Tukey

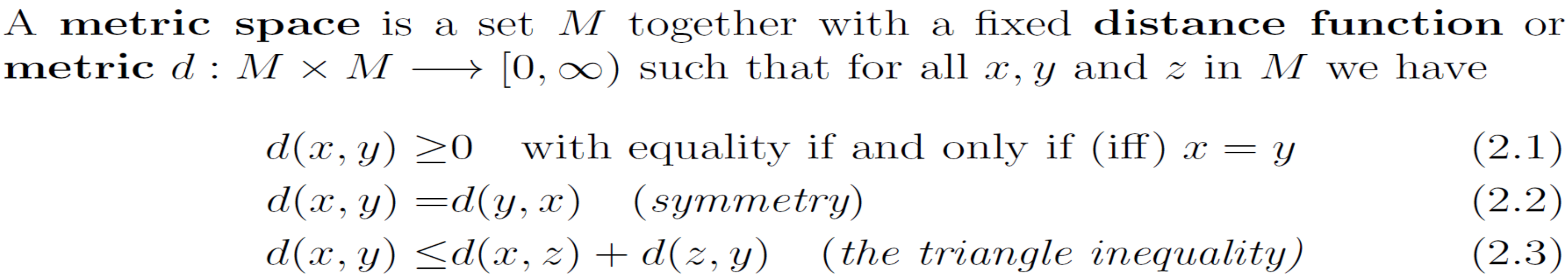
EXPLORATORY DATA ANALYSIS

1977

First question: What should it mean to be similar?

-> Choice of similarity measure or distance function

Example genuine metrics:



A **metric space** is a set M together with a fixed **distance function** or **metric** $d : M \times M \longrightarrow [0, \infty)$ such that for all x, y and z in M we have

$$d(x,y) \geq 0 \quad \text{with equality if and only if (iff) } x = y \tag{2.1}$$

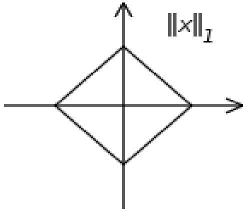
$$d(x,y) = d(y,x) \quad \textit{(symmetry)} \tag{2.2}$$

$$d(x,y) \leq d(x,z) + d(z,y) \quad \textit{(the triangle inequality)} \tag{2.3}$$

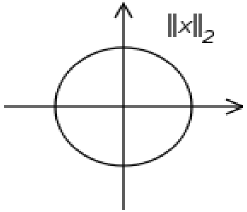
Examples:

- **Different types of edit distances** for e.g. sequencing data
- L2 or ***Euclidean distance*** for quantitative data
- L1 or ***Manhattan*** *(or Taxicab)* ***distance***

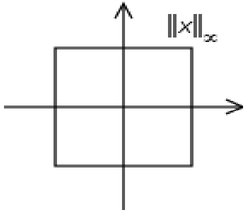
Effects of different similarity measures or distance functions



The “unit sphere” in the L^1 norm. A constraint formulated with this norm favors *sparse vectors*. The idea behind the *Lasso*, see Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).



The normal “unit sphere” in the standard L^2 norm

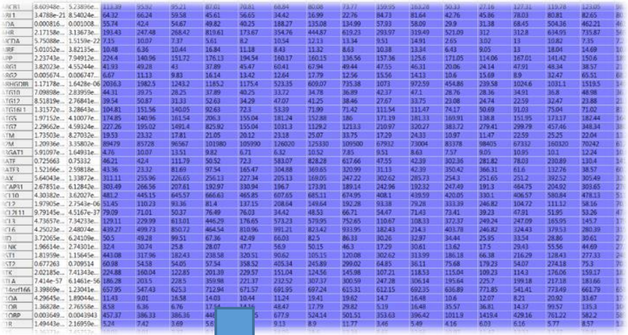


The “unit sphere” in the L^∞ norm

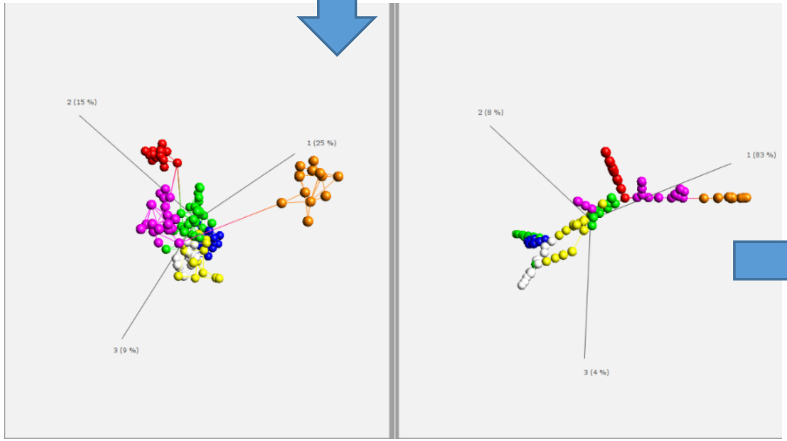
Feature extraction and selection to “understand” phenotypic variation

Sample space population

Measurements



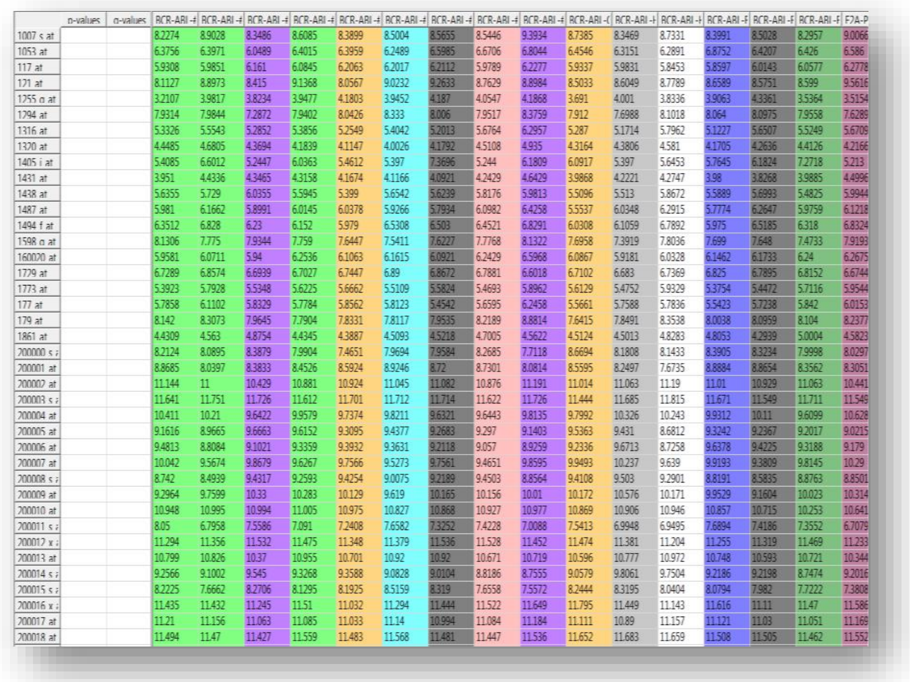
Data



Data Analysis

→ clustering & classification → Biological Insights

How to explore MULTI-OMICS data?



	p-values	q-values	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -#	BCR-ABL -(	BCR-ABL -I	BCR-ABL -I	BCR-ABL -F	BCR-ABL -F	BCR-ABL -F	E2A-P
1007 s at			8.2274	8.9028	8.3486	8.6085	8.3899	8.5004	8.5655	8.5446	9.3934	8.7385	8.3469	8.7331	8.3991	8.5028	8.2957	9.0066
1053 at			6.3756	6.3971	6.0489	6.4015	6.3959	6.2489	6.5985	6.6706	6.8044	6.4546	6.3151	6.2891	6.8752	6.4207	6.426	6.586
117 at			5.9308	5.9851	6.161	6.0845	6.2063	6.2017	6.2112	5.9789	6.2277	5.9337	5.9831	5.8453	5.8597	6.0143	6.0577	6.2778
121 at			8.1127	8.8973	8.415	9.1368	8.0567	9.0232	9.2633	8.7629	8.8984	8.5033	8.6049	8.7789	8.6589	8.5751	8.599	9.5616
1255 g at			3.2107	3.9817	3.8234	3.9477	4.1803	3.9452	4.187	4.0547	4.1868	3.691	4.001	3.8336	3.9063	4.3361	3.5364	3.5154
1294 at			7.9314	7.9844	7.2872	7.9402	8.0426	8.333	8.006	7.9517	8.3759	7.912	7.6988	8.1018	8.064	8.0975	7.9558	7.6289
1316 at			5.3326	5.5543	5.2852	5.3856	5.2549	5.4042	5.2013	5.6764	6.2957	5.287	5.1714	5.7962	5.1227	5.6507	5.5249	5.6709
1320 at			4.4485	4.6805	4.3694	4.1839	4.1147	4.0026	4.1792	4.5108	4.935	4.3164	4.3806	4.581	4.1705	4.2636	4.4126	4.2166
1405 i at			5.4085	6.6012	5.2447	6.0363	5.4612	5.397	7.3696	5.244	6.1809	6.0917	5.397	5.6453	5.7645	6.1824	7.2718	5.213
1431 at			3.951	4.4336	4.3465	4.3158	4.1674	4.1166	4.0921	4.2429	4.6429	3.9868	4.2221	4.2747	3.98	3.8268	3.9885	4.4996
1438 at			5.6355	5.729	6.0355	5.5945	5.399	5.6542	5.6239	5.8176	5.9813	5.5096	5.513	5.8672	5.5889	5.6993	5.4825	5.9944
1487 at			5.981	6.1662	5.8991	6.0145	6.0378	5.9266	5.7934	6.0982	6.4258	5.5537	6.0348	6.2915	5.7774	6.2647	5.9759	6.1218
1494 f at			6.3512	6.828	6.23	6.152	5.979	6.5308	6.503	6.4521	6.8291	6.0308	6.1059	6.7892	5.975	6.5185	6.318	6.8324
1598 g at			8.1306	7.775	7.9344	7.759	7.6447	7.5411	7.6227	7.7768	8.1322	7.6958	7.3919	7.8036	7.699	7.648	7.4733	7.9193
160020 at			5.9581	6.0711	5.94	6.2536	6.1063	6.1615	6.0921	6.2429	6.5968	6.0867	5.9181	6.0328	6.1462	6.1733	6.24	6.2675
1729 at			6.7289	6.8574	6.6939	6.7027	6.7447	6.89	6.8672	6.7881	6.6018	6.7102	6.683	6.7369	6.825	6.7895	6.8152	6.6744
1773 at			5.3923	5.7928	5.5348	5.6225	5.6662	5.5109	5.5824	5.4693	5.8962	5.6129	5.4752	5.9329	5.3754	5.4472	5.7116	5.9544
177 at			5.7858	6.1102	5.8329	5.7784	5.8562	5.8123	5.4542	5.6595	6.2458	5.5661	5.7588	5.7836	5.5423	5.7238	5.842	6.0153
179 at			8.142	8.3073	7.9645	7.7904	7.8331	7.8117	7.9535	8.2189	8.8814	7.6415	7.8491	8.3538	8.0038	8.0959	8.104	8.2377
1861 at			4.4309	4.563	4.8754	4.4345	4.3887	4.5093	4.5218	4.7005	4.5622	4.5124	4.5013	4.8283	4.8053	4.2939	5.0004	4.5823
200000 s ;			8.2124	8.0895	8.3879	7.9904	7.4651	7.9694	7.9584	8.2685	7.7118	8.6694	8.1808	8.1433	8.3905	8.3234	7.9998	8.0297
200001 at			8.8685	8.0397	8.3833	8.4526	8.5924	8.9246	8.72	8.7301	8.0814	8.5595	8.2497	7.6735	8.8884	8.8654	8.3562	8.3051
200002 at			11.144	11	10.429	10.881	10.924	11.045	11.082	10.876	11.191	11.014	11.063	11.19	11.01	10.929	11.063	10.441
200003 s ;			11.641	11.751	11.726	11.612	11.701	11.712	11.714	11.622	11.726	11.444	11.685	11.815	11.671	11.549	11.711	11.549
200004 at			10.411	10.21	9.6422	9.9579	9.7374	9.8211	9.6321	9.6443	9.8135	9.7992	10.326	10.243	9.9312	10.11	9.6099	10.628
200005 at			9.1616	8.9665	9.6663	9.6152	9.3095	9.4377	9.2683	9.297	9.1403	9.5363	9.431	8.6812	9.3242	9.2367	9.2017	9.0215
200006 at			9.4813	8.8084	9.1021	9.3359	9.3932	9.3631	9.2118	9.057	8.9259	9.2336	9.6713	8.7258	9.6378	9.4225	9.3188	9.179
200007 at			10.042	9.5674	9.8679	9.6267	9.7566	9.5273	9.7561	9.4651	9.8595	9.9493	10.237	9.639	9.9193	9.3809	9.8145	10.29
200008 s ;			8.742	8.4939	9.4317	9.2593	9.4254	9.0075	9.2189	9.4503	8.8564	9.4108	9.503	9.2901	8.8191	8.5835	8.8763	8.8501
200009 at			9.2964	9.7599	10.33	10.283	10.129	9.619	10.165	10.156	10.01	10.172	10.576	10.171	9.9529	9.1604	10.023	10.314
200010 at			10.948	10.995	10.994	11.005	10.975	10.827	10.868	10.927	10.977	10.869	10.906	10.946	10.857	10.715	10.253	10.641
200011 s ;			8.05	6.7958	7.5586	7.091	7.2408	7.6582	7.3252	7.4228	7.0088	7.5413	6.9948	6.9495	7.6894	7.4186	7.3552	6.7079
200012 x ;			11.294	11.356	11.532	11.475	11.348	11.379	11.536	11.528	11.452	11.474	11.381	11.204	11.255	11.319	11.469	11.233
200013 at			10.799	10.826	10.37	10.955	10.701	10.92	10.92	10.671	10.719	10.596	10.777	10.972	10.748	10.593	10.721	10.344
200014 s ;			9.2566	9.1002	9.545	9.3268	9.3588	9.0828	9.0104	8.8186	8.7555	9.0579	9.8061	9.7504	9.2186	9.2198	8.7474	9.2016
200015 s ;			8.2225	7.6662	8.2706	8.1295	8.1925	8.5159	8.319	7.6558	7.5572	8.2444	8.3195	8.0404	8.0794	7.982	7.7222	7.3808
200016 x ;			11.435	11.432	11.245	11.51	11.032	11.294	11.444	11.522	11.649	11.795	11.449	11.143	11.616	11.11	11.47	11.586
200017 at			11.21	11.156	11.063	11.085	11.033	11.14	10.994	11.084	11.184	11.111	10.89	11.157	11.121	11.03	11.051	11.169
200018 at			11.494	11.47	11.427	11.559	11.483	11.568	11.481	11.447	11.536	11.652	11.683	11.659	11.508	11.505	11.462	11.552

- **N=10-10^5 Samples**
- **P=10-10^10 Variables**

How to find relevant structure?

Problems:

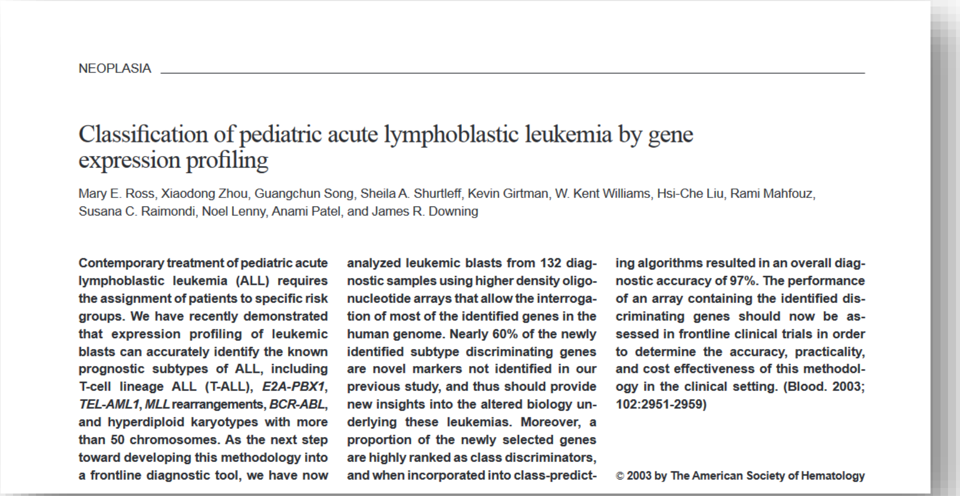
- **Noise**
- **Artifacts: technical, batch effects,..**
- **High Dimensions**
- **Often few samples and many variables**

Visualize and Analyze the data

- **Choice of “*similarity measure*”**
- **Choice of dimension/model-reduction method = choice of “*objective function*”**

Classification of pediatric acute lymphoblastic leukemia by gene expression profiling
Mary E. Ross et al.
Blood 2003 102:2951-2959; doi:10.1182/blood-2003-01-0338
Data available at https://www.stjuderesearch.org/site/data/ALL3/

Example Data: Affymetrix chip 22282 ProbeIDs and 132 Samples



NEOPLASIA

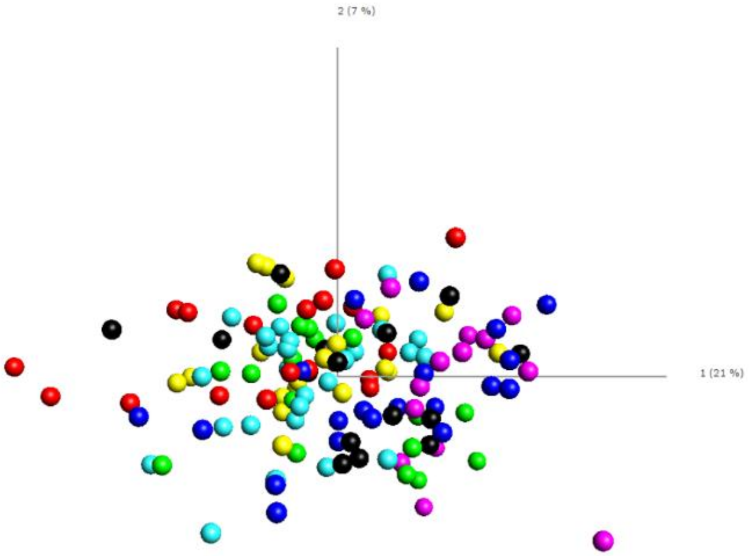
Classification of pediatric acute lymphoblastic leukemia by gene expression profiling

Mary E. Ross, Xiaodong Zhou, Guangchun Song, Sheila A. Shurtleff, Kevin Girtman, W. Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C. Raimondi, Noel Lenny, Anami Patel, and James R. Downing

Contemporary treatment of pediatric acute lymphoblastic leukemia (ALL) requires the assignment of patients to specific risk groups. We have recently demonstrated that expression profiling of leukemic blasts can accurately identify the known prognostic subtypes of ALL, including T-cell lineage ALL (T-ALL), *E2A-PBX1*, *TEL-AML1*, *MLL* rearrangements, *BCR-ABL*, and hyperdiploid karyotypes with more than 50 chromosomes. As the next step toward developing this methodology into a frontline diagnostic tool, we have now analyzed leukemic blasts from 132 diagnostic samples using higher density oligonucleotide arrays that allow the interrogation of most of the identified genes in the human genome. Nearly 60% of the newly identified subtype discriminating genes are novel markers not identified in our previous study, and thus should provide new insights into the altered biology underlying these leukemias. Moreover, a proportion of the newly selected genes are highly ranked as class discriminators, and when incorporated into class-predicting algorithms resulted in an overall diagnostic accuracy of 97%. The performance of an array containing the identified discriminating genes should now be assessed in frontline clinical trials in order to determine the accuracy, practicality, and cost effectiveness of this methodology in the clinical setting. (Blood. 2003; 102:2951-2959)

© 2003 by The American Society of Hematology

Dimension reduction techniques: *Classical PCA*

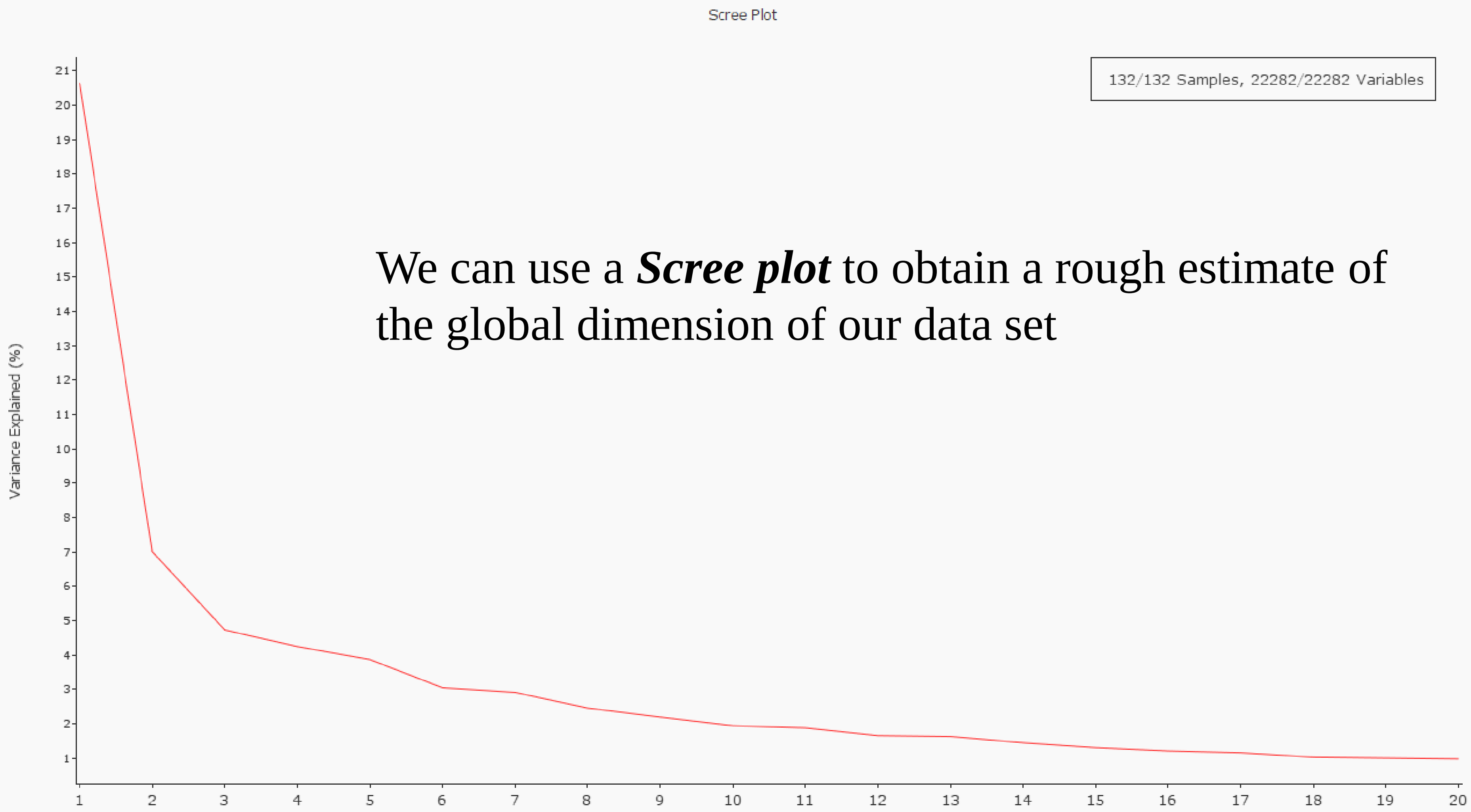


ALL (Leukemia) samples
mRNA bulk

22282 variables 132 samples

PCA based on the correlation matrix;
Mean centering and normalizing all variables to unit variance

Distances: Euclidean **Objective function**: Total Variance



We can use a ***Scree plot*** to obtain a rough estimate of the global dimension of our data set

Fontes and Soneson *BMC Bioinformatics* 2011, **12**:307
http://www.biomedcentral.com/1471-2105/12/307

BMC Bioinformatics

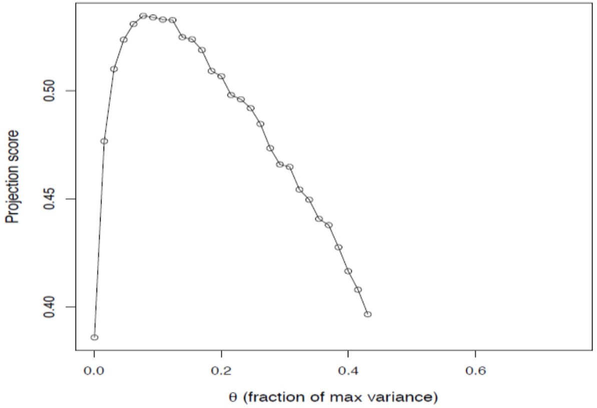
RESEARCH ARTICLE Open Access

The projection score - an evaluation criterion for variable subset selection in PCA visualization

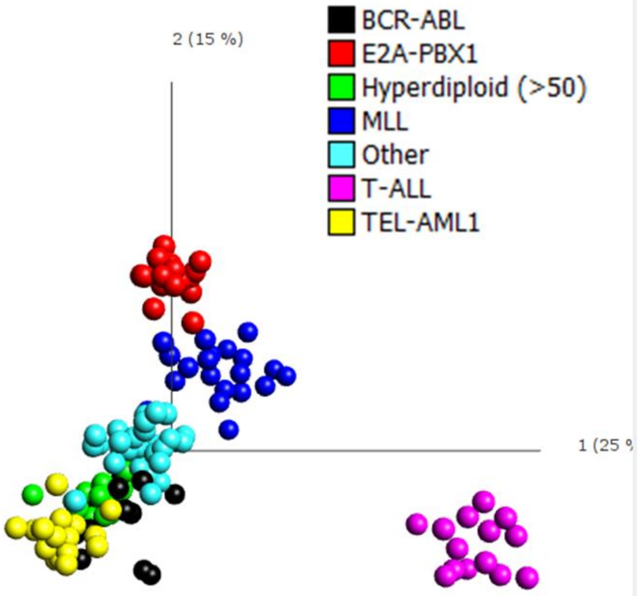
Magnus Fontes* and Charlotte Soneson

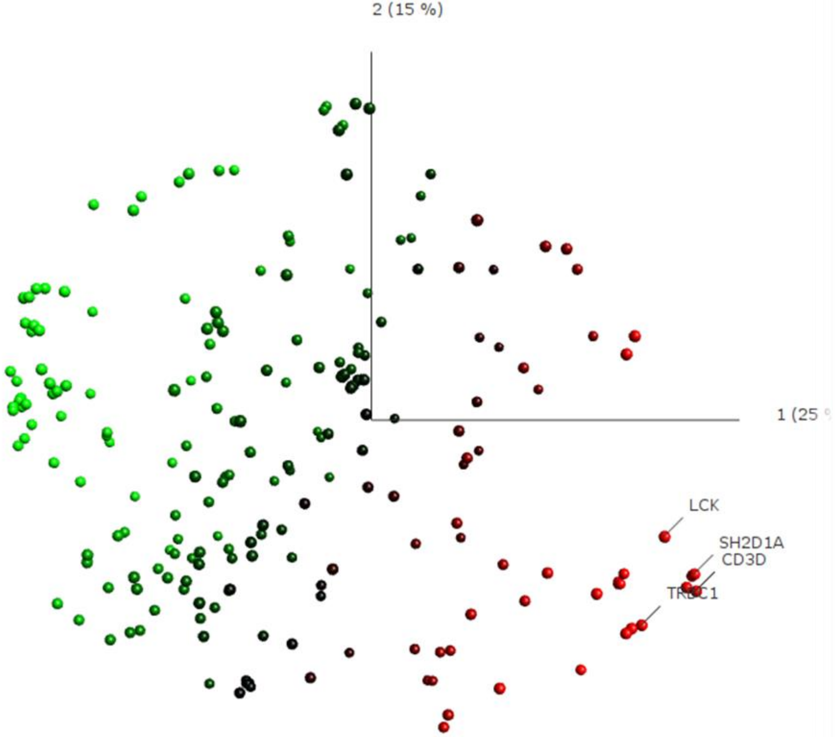
$$\alpha_2(\boldsymbol{\Lambda}_{\mathrm{X}}, S) = \frac{\sum_{k\in S}\lambda_k^2}{\sum_{k=1}^{r}\lambda_k^2}.$$

$$\begin{aligned}\tau(\phi_m(\mathbf{X}), S, \mathcal{P}_{\phi_m(\mathrm{X})}) = {} & (\alpha_2(\boldsymbol{\Lambda}_{\phi_m(\mathrm{X})}, S))^{1/2} \\ & - \mathbb{E}_{\mathcal{P}_{\phi_m(\mathrm{X})}}\left[(\alpha_2(\boldsymbol{\Lambda}_{\phi_m(\mathrm{X})}, S))^{1/2}\right].\end{aligned}$$

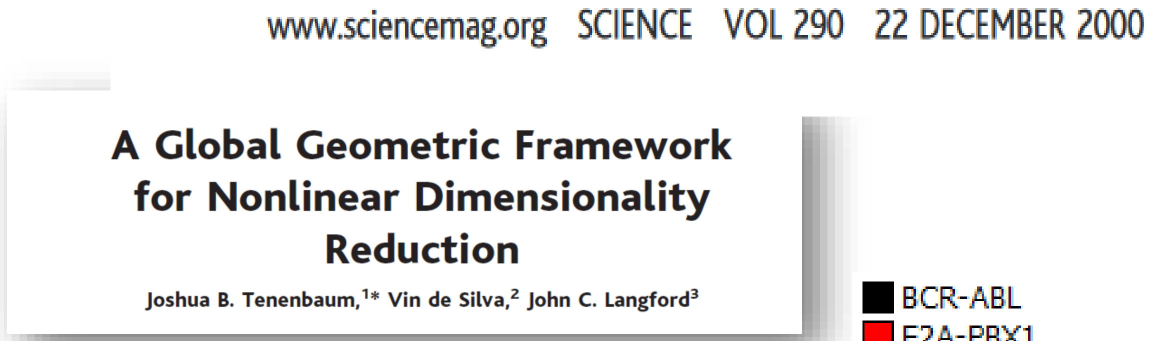


Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000, **24**:227-235.





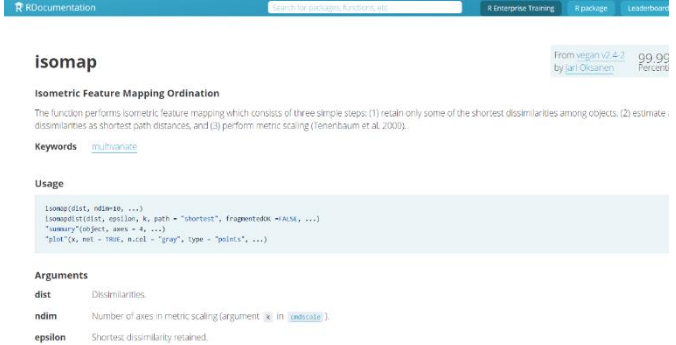
Duality of PCA → *Bi-plot* after optimization over *projection score*



www.sciencemag.org SCIENCE VOL 290 22 DECEMBER 2000

A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum,[1]* Vin de Silva,[2] John C. Langford[3]



isomap

From vegan v2.4-2 by Jari Oksanen

99.99 Percenti

Isometric Feature Mapping Ordination

The function performs isometric feature mapping which consists of three simple steps: (1) retain only some of the shortest dissimilarities among objects, (2) estimate dissimilarities as shortest path distances, and (3) perform metric scaling (Tenenbaum et al. 2000).

Keywords multivariate

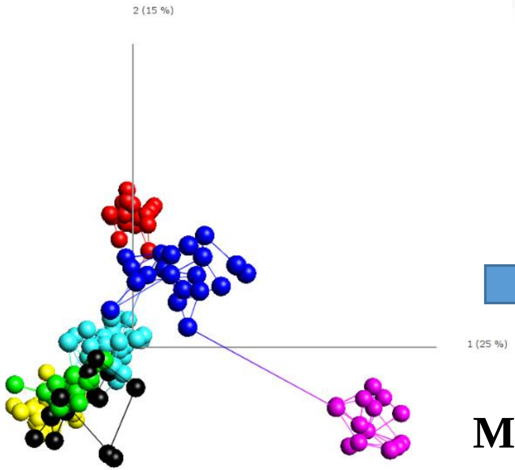
Usage

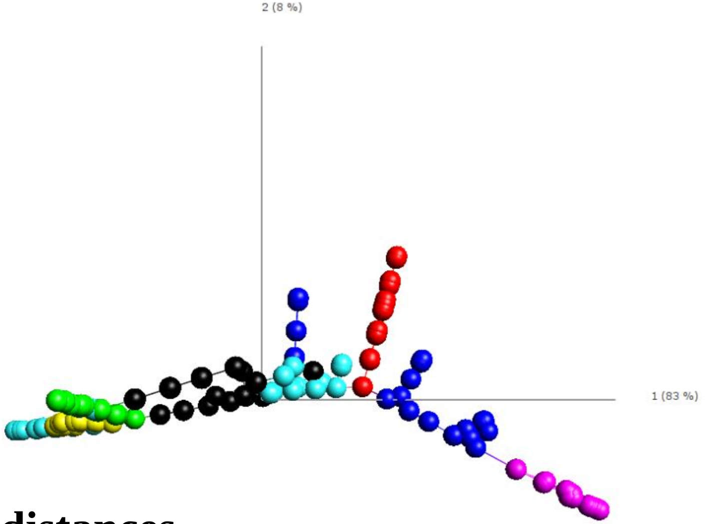
```
isomap(dist, ndim=10, ...)
isomapdist(dist, epsilon, k, path = "shortest", fragmentedOK =FALSE, ...)
"summary"(object, axes = 4, ...)
"plot"(x, net = TRUE, n.col = "gray", type = "points", ...)
```

Arguments

dist	Dissimilarities.
ndim	Number of axes in metric scaling (argument `k` in `cmdscale`).
epsilon	Shortest dissimilarity retained.

⬛ BCR-ABL
🟥 E2A-PBX1
🟩 Hyperdiploid (>50)
🟦 MLL
▢ Other
🟪 T-ALL
🟨 TEL-AML1





MDS based on graph distances

ISOMAP

Stochastic Neighbor Embedding (SNE)

Original distances

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

Original SNE

Sam Roweis and Geoffrey Hinton

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

Distances in reduced space

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}.$$

tSNE

Van der Maaten

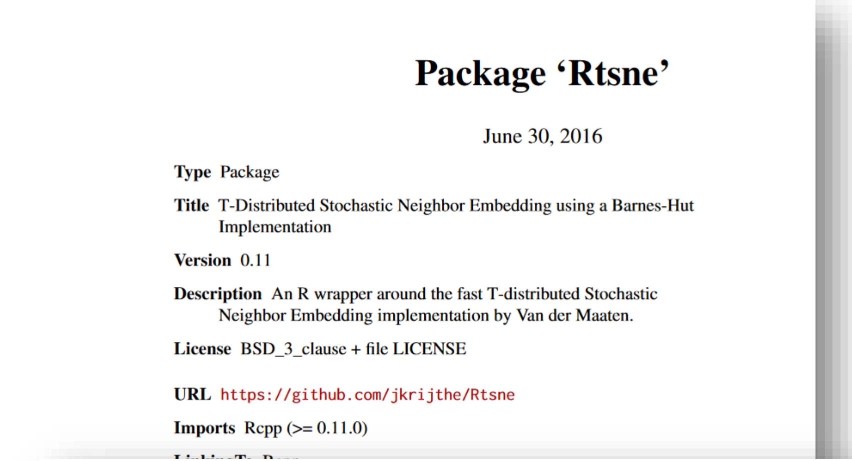
The cost function C is given by

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

Kullback Leibler

Relative Entropy

Stochastic Neighbor Embedding (SNE) and tSNE



Package 'Rtsne'

June 30, 2016

Type Package

Title T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation

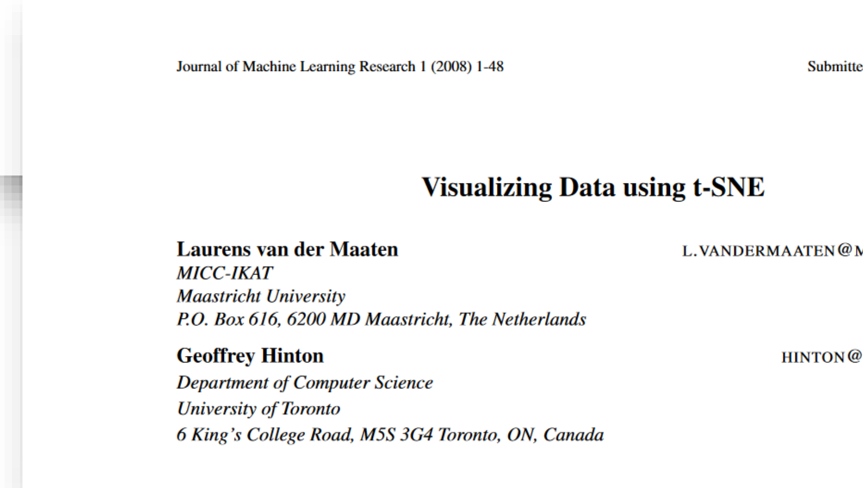
Version 0.11

Description An R wrapper around the fast T-distributed Stochastic Neighbor Embedding implementation by Van der Maaten.

License BSD_3_clause + file LICENSE

URL https://github.com/jkrijthe/Rtsne

Imports Rcpp (>= 0.11.0)

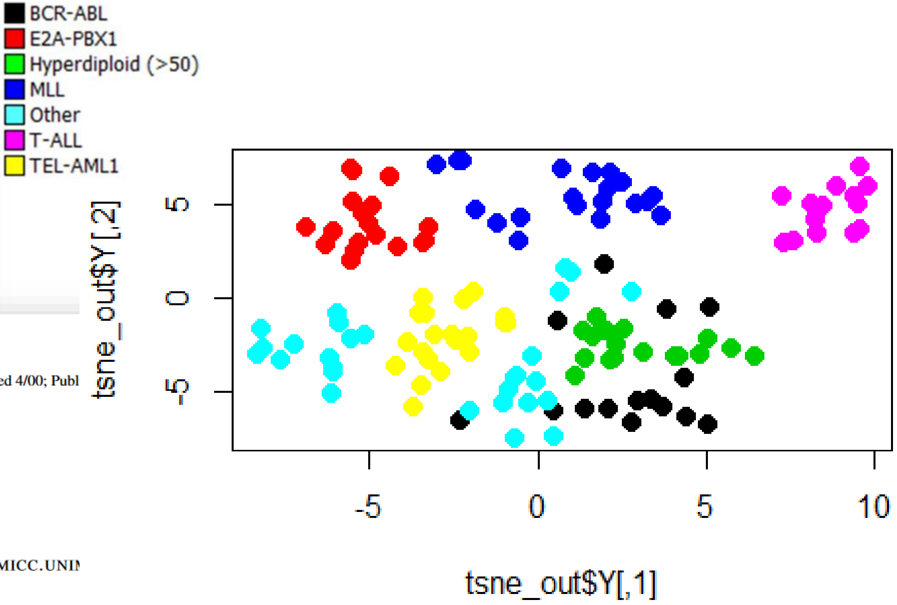


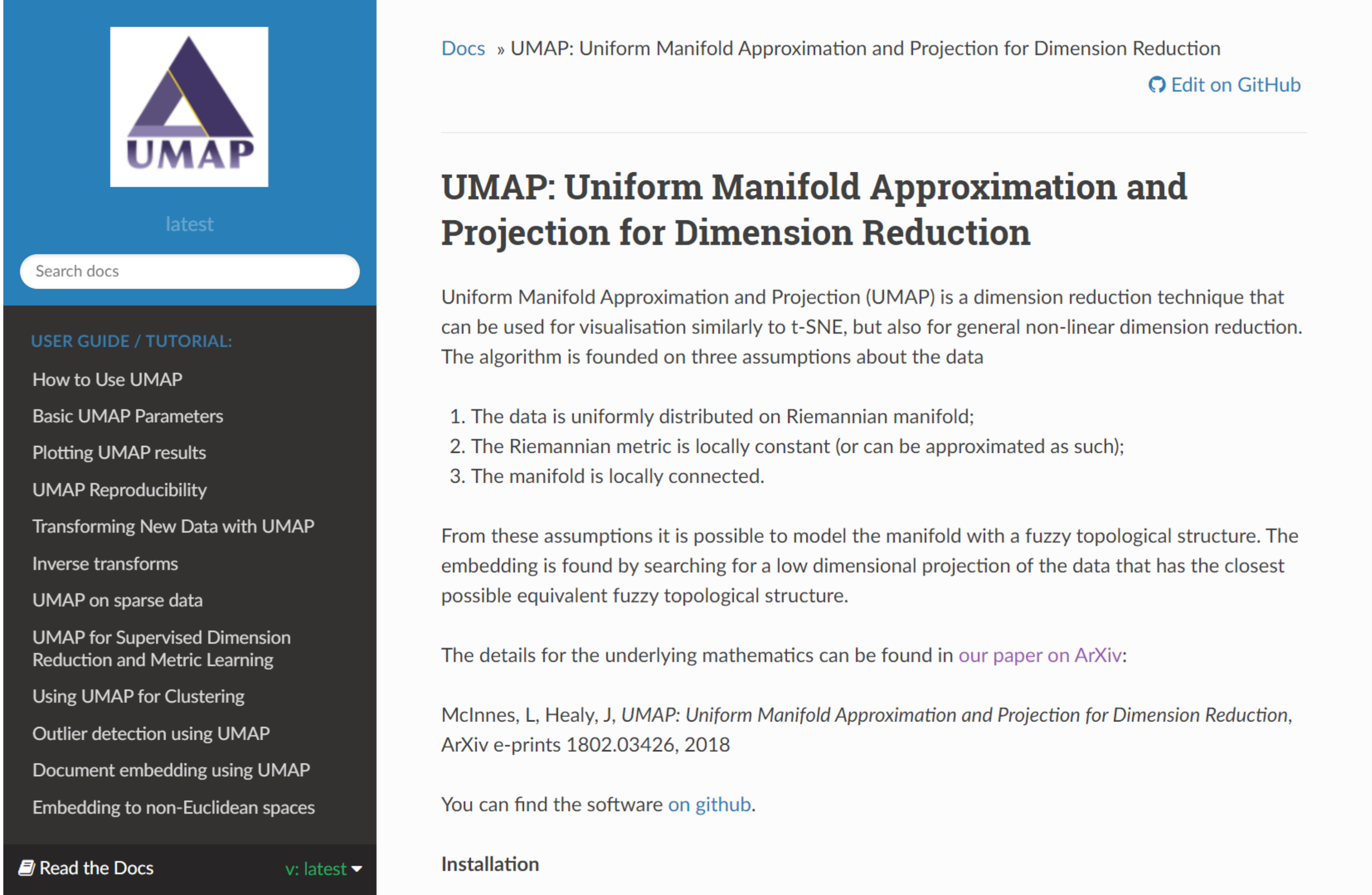
Journal of Machine Learning Research 1 (2008) 1-48 Submitted 4/00; Publ

Visualizing Data using t-SNE

Laurens van der Maaten L.VANDERMAATEN@MICC.UNI
MICC-IKAT
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada




UMAP
latest
Search docs
USER GUIDE / TUTORIAL:
How to Use UMAP
Basic UMAP Parameters
Plotting UMAP results
UMAP Reproducibility
Transforming New Data with UMAP
Inverse transforms
UMAP on sparse data
UMAP for Supervised Dimension Reduction and Metric Learning
Using UMAP for Clustering
Outlier detection using UMAP
Document embedding using UMAP
Embedding to non-Euclidean spaces
Read the Docs
v: latest
Docs » UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
Edit on GitHub

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. The algorithm is founded on three assumptions about the data

1. The data is uniformly distributed on Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected.

From these assumptions it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

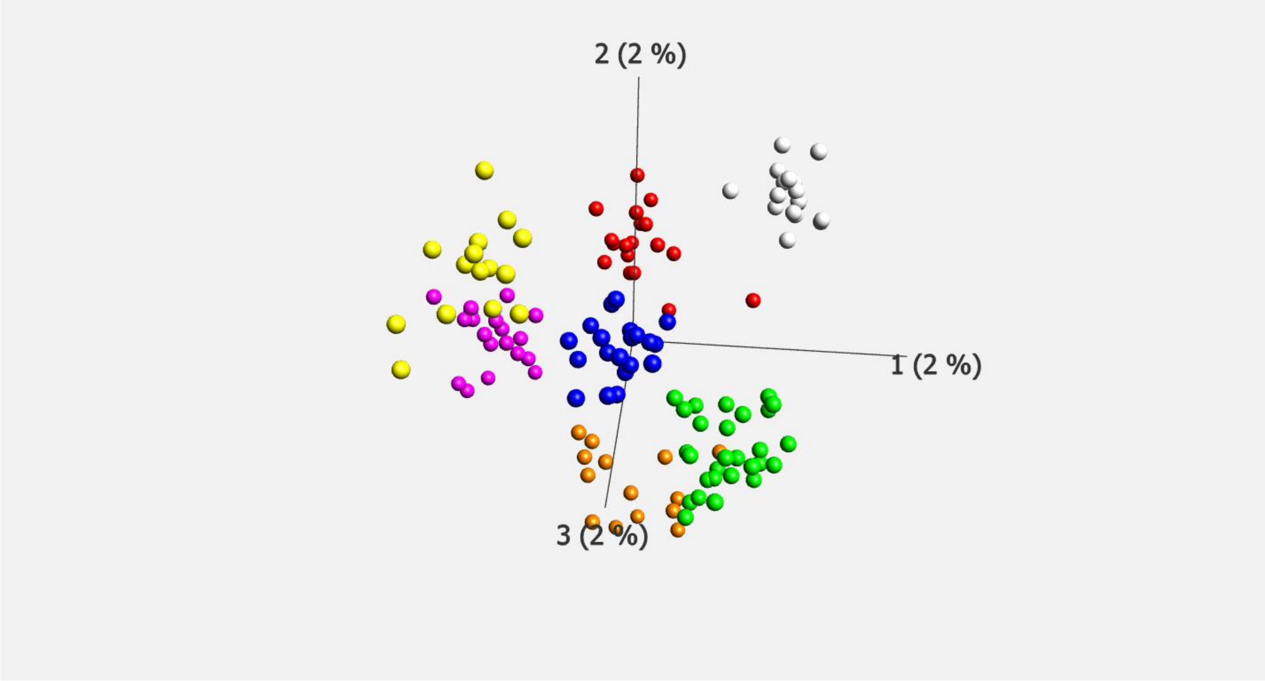
The details for the underlying mathematics can be found in our paper on ArXiv:

McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

You can find the software on github.

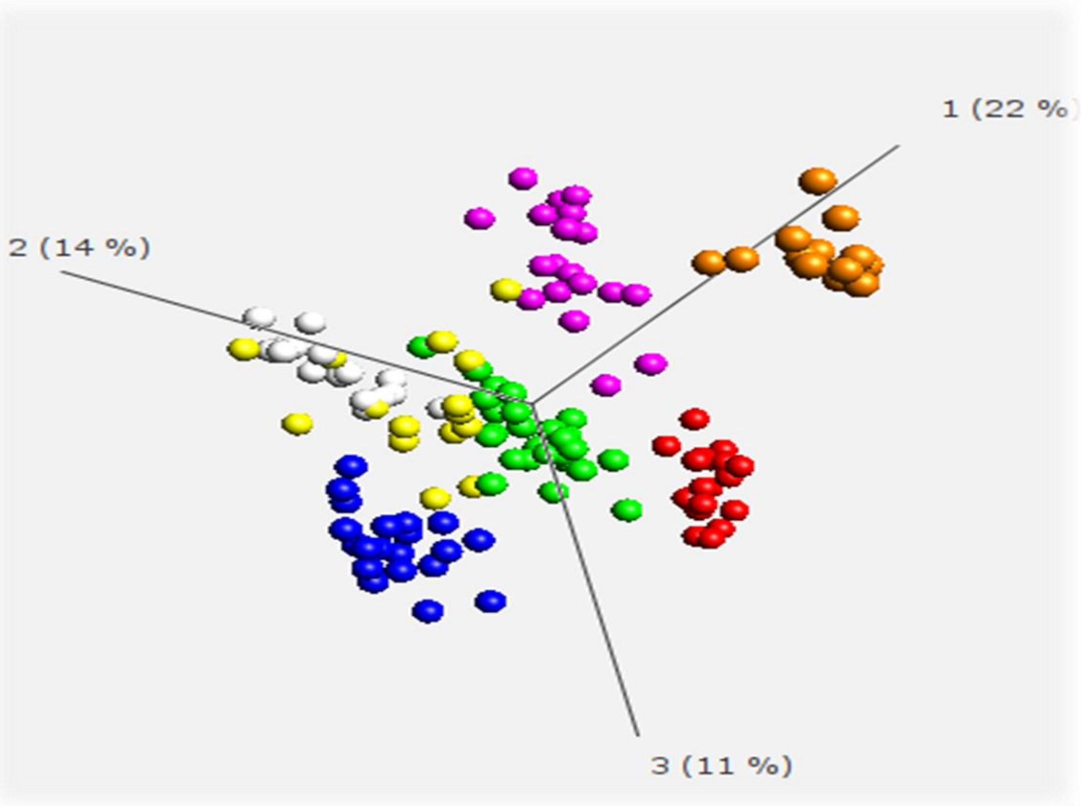
Installation

Dangers with exploratory analyses: ANOVA on Random data



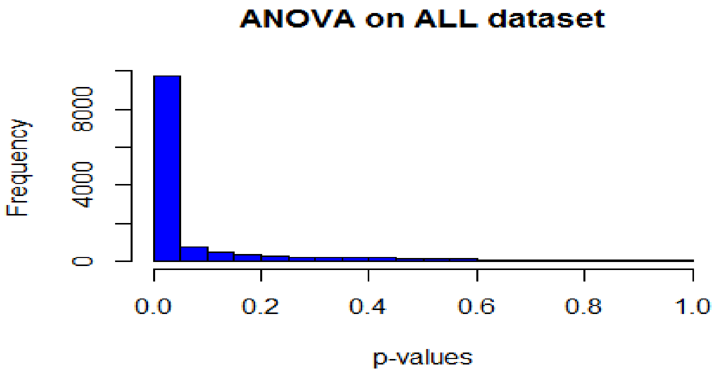
Filtering using ANOVA on a 22282*132 dataset with significance threshold p=0.05 on random data resulting in 1108 “*significant discoveries*” + PCA visualization.
*Note that 0.05*22282 =1114.1*

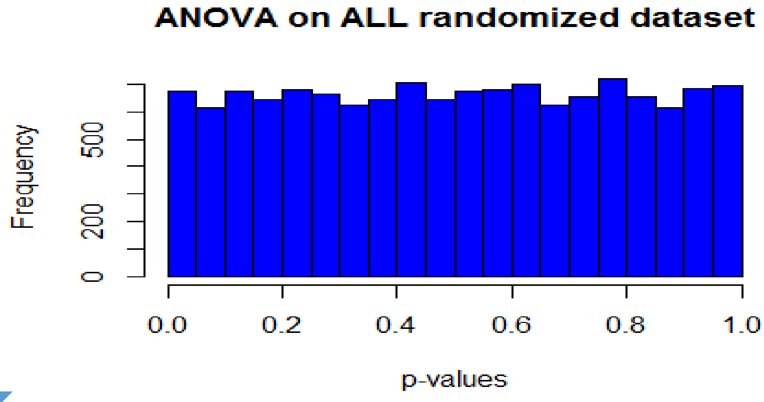
ANOVA on ALL dataset

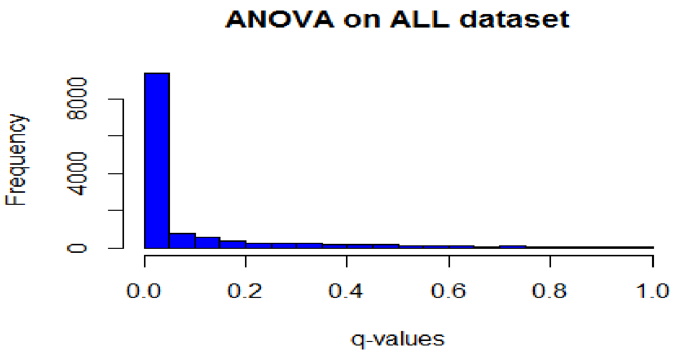


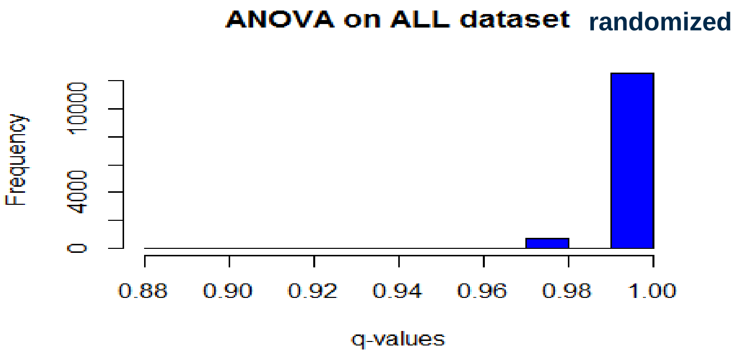
Filtering down to 1000 Variables (p=6e-13)

The p-value distributions

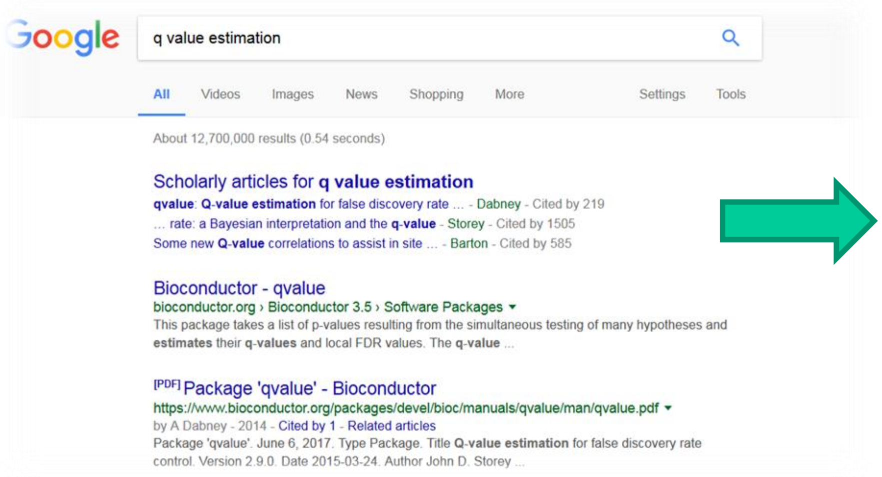








Google search for *q value estimation*



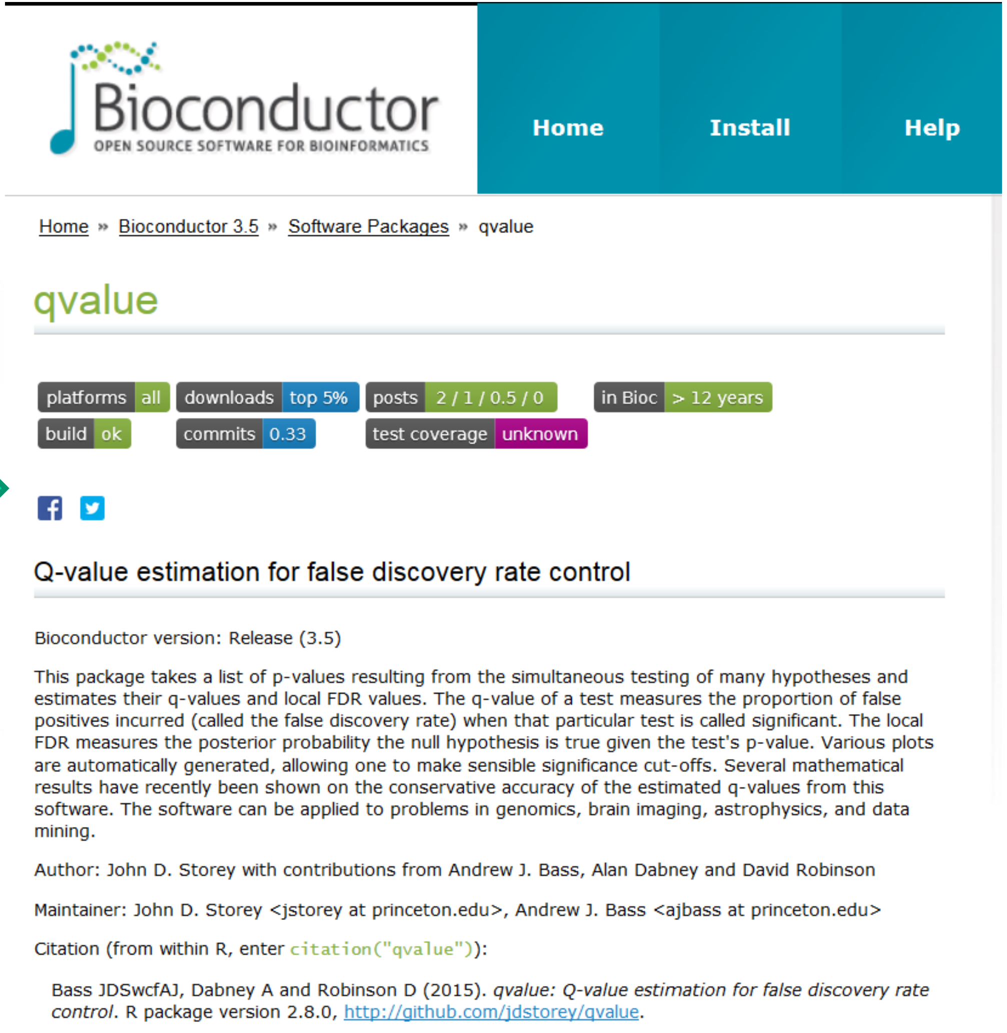
About 12,700,000 results (0.54 seconds)

Scholarly articles for **q value estimation**
qvalue: **Q-value estimation** for false discovery rate ... - Dabney - Cited by 219
... rate: a Bayesian interpretation and the **q-value** - Storey - Cited by 1505
Some new **Q-value** correlations to assist in site ... - Barton - Cited by 585

Bioconductor - qvalue
bioconductor.org › Bioconductor 3.5 › Software Packages ▾
This package takes a list of p-values resulting from the simultaneous testing of many hypotheses and **estimates** their **q-values** and local FDR values. The **q-value** ...

[PDF] Package 'qvalue' - Bioconductor
https://www.bioconductor.org/packages/devel/bioc/manuals/qvalue/man/qvalue.pdf ▾
by A Dabney - 2014 - Cited by 1 - Related articles
Package 'qvalue'. June 6, 2017. Type Package. Title **Q-value estimation** for false discovery rate control. Version 2.9.0. Date 2015-03-24. Author John D. Storey ...

lead to R implementation in *Bioconductor* →



Home » Bioconductor 3.5 » Software Packages » qvalue

qvalue

Q-value estimation for false discovery rate control

Bioconductor version: Release (3.5)

This package takes a list of p-values resulting from the simultaneous testing of many hypotheses and estimates their q-values and local FDR values. The q-value of a test measures the proportion of false positives incurred (called the false discovery rate) when that particular test is called significant. The local FDR measures the posterior probability the null hypothesis is true given the test's p-value. Various plots are automatically generated, allowing one to make sensible significance cut-offs. Several mathematical results have recently been shown on the conservative accuracy of the estimated q-values from this software. The software can be applied to problems in genomics, brain imaging, astrophysics, and data mining.

Author: John D. Storey with contributions from Andrew J. Bass, Alan Dabney and David Robinson

Maintainer: John D. Storey <jstorey at princeton.edu>, Andrew J. Bass <ajbass at princeton.edu>

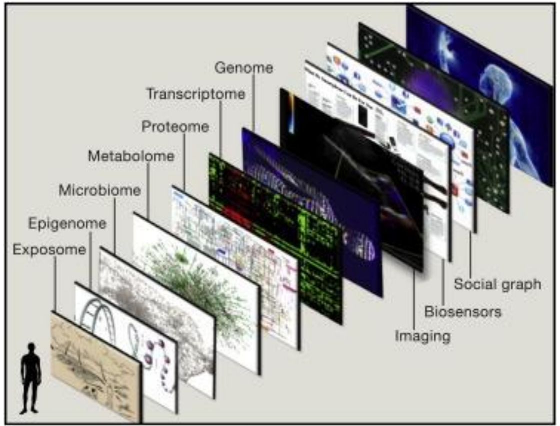
Citation (from within R, enter `citation("qvalue")`):

Bass JDSwcfAJ, Dabney A and Robinson D (2015). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.8.0, http://github.com/jdstorey/qvalue.

Recap: General directions in Translational research

Emerging Technology & Data & Insights for

- Longitudinal Sampling:
  Precision Diagnostics & Medicine: Mapping of « Patient journeys » & « Trajectories in Biomedical State Space » integrating different data modalities (Omics, Imaging, RWD, ….)
- Deep phenotyping sampling:
  Integrative and Holistic analyses using emerging bio-technologies



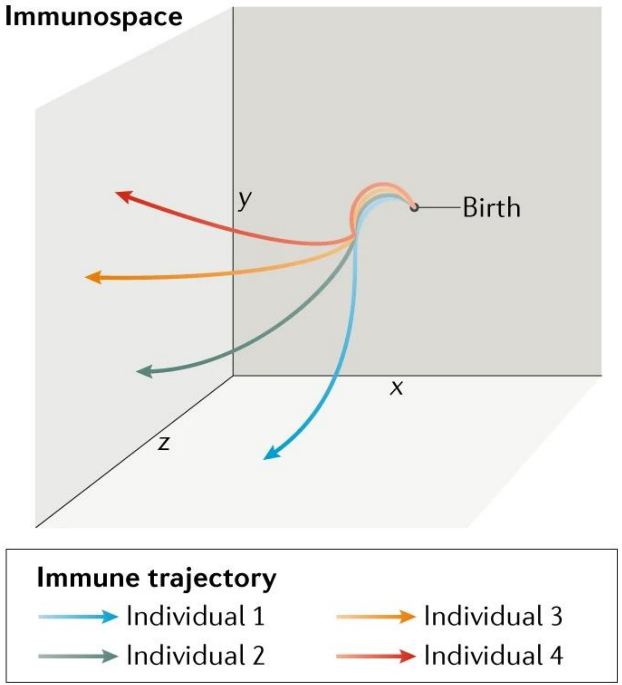
Individualized Medicine
from Prewomb to Tomb

Eric J. Topol[1],*
[1]The Scripps Translational Science Institute, The Scripps Research Institute and Scripps Health, La Jolla, CA 92037, USA
*Correspondence: etopol@scripps.edu
http://dx.doi.org/10.1016/j.cell.2014.02.012

Multidisciplinary collaborative data & advanced analytics driven biomedical research in order to:

- Better understanding the Dynamics & Complexity:
  Static snapshots will be complemented with Dynamic systems control approaches leading to dynamic & composite biomarkers for response, resistance, safety, quality of life …
- Better Health State Monitoring & Early Detection and dynamical control:
  Controlling and maintaining health in order to prevent disease will lead to treatment paradigm shifts.

General Systems Immunology for Biomedicine: Predicting and controlling dynamics under perturbations

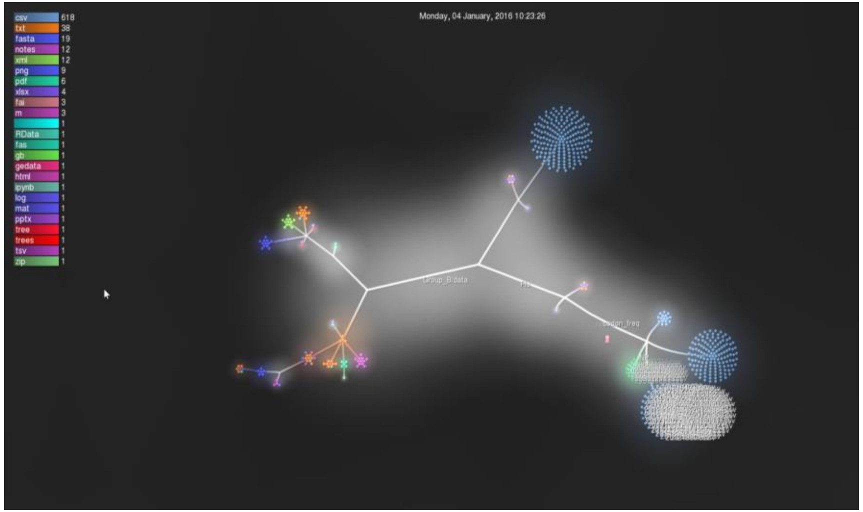


- “Human immune systems are relatively stable within individuals over the course of weeks to months, but incredibly variable between individuals”
- “…induced responses to pathogens differ markedly among different age groups and… these differences are unique to different kinds of stimuli”
- Functional gene expression responses of human blood cells to common pathogens differ broadly across age groups.
- Divergent immune cell composition with advanced age is associated with chromatin changes that are induced by environmental influences over the course of life.

Petter Brodin
Nature Reviews | Immunology volume 19 | February 2019 | 87

Infrastructure connecting the dots through Human-Machine partnerships Biomedicine& Data & Advanced Analytics → New Insights

- Connecting data bases
- Connecting IT platforms & tools
- Connecting collaborative research infrastructure (wiki, fora, …)
- Connecting scientists from bench to computational mathematics
- Incentivize sharing through provenance & tracking



Drive cultural change ⟷ **Collaboration & Sharing**