Computational Analysis of Transcriptomics Data

National Institute of Immunology, New Delhi

Day 2

21-05-2022

Tutor

Rintu Kutum

Faculty Fellow, Department of Computer Science

Data Scientist, Mphasis Lab for Machine Learning & Computational Thinking

Ashoka University

Young Professional

ISO TC 215 - Health Informatics - Genomics

Bureau of Indian Standards, MHD

Member & Instructor

Reproducibility for Everyone (R4E)

eLife Initiative on Reproducible Research (2018-2020)

Reproducibility for Everyone

Contribution in the field of Large-scale Biomedical Research (2012-2021)

International DREAM Challenges

(2013-2021)

DREAM-Toxicogenetics, 40th rank

Got the exposure to large scale biomedical and genomics datasets. Deployed random forests algorithm to predict toxicity of drugs in humans

Eduati F et al., **Nature Biotechnology**, 2015

DREAM-Preterm Birth, *3rd Rank*

Developed ensembled machine learning algorithms to predict preterm birth from multi-omics data

Tarca AL et al., **Cell Reports Medicine**, 2021

DREAM-Single Cell Signaling in Breast Cancer, *9th Rank*

Developed semi-supervised biased-GLM approach to predict missing single cell markers

Gabor A et al., **Mol Sys Bio** (2021)

PrecisionFDA Challenges

(2017-2021)

Contributed to the machine learning algorithms to tackle mislabelling of samples

Proteogenomics (Yoo S. et al., **Patterns**, 2021)

ASHOKA UNIVERSITY

Contributions in the field of Genomics (2014-2020)

International DREAM Challenges

(2013-2021)

DREAM-Toxicogenetics, 40th rank

Got the exposure to large scale biomedical and genomics datasets. Deployed random forests algorithm to predict toxicity of drugs in humans

Eduati F et al., **Nature Biotechnology**, 2015

DREAM-Preterm Birth, *3rd Rank*

Developed ensembled machine learning algorithms to predict preterm birth from multi-omics data

Tarca AL et al., **Cell Reports Medicine**, 2021

DREAM-Single Cell Signaling in Breast Cancer, *9th Rank*

Developed semi-supervised biased-GLM approach to predict missing single cell markers

Gabor A et al., **Mol Sys Bio** (2021)

PrecisionFDA Challenges

(2017-2021)

Contributed to the machine learning algorithms to tackle mislabelling of samples

Proteogenomics (Yoo S. et al., **Patterns**, 2021)

Population Genetics

Unsupervised random forests (RF) algorithm to decipher genetic structure based on CNV

Narang A. et al, GBE (2014)

Asthma

Unsupervised and supervised RF algorithm to find discover subgroups and associated biomarkers from metabolites of children with Asthma

Sinha A. et al, **JTM** (2017)

Tuberculosis

Data science approach to discover drug tragets from 1800 Mtb clinical strains

Kaur D. et al, **Sci. Report** (2017)

Performed batch correction, preprocessing, and differential proteomics analysis

Menon D. et al, **ACS infectious diseases** (2019)

Genomewide investigation of codon-bias in the toxin-antitoxin genes of 1800 Mtb clinical strains (THSTI) Talwar S. et al, **mSystems** (2020)

Celiac Disease

Discovery of transcriptomics biomarkers for Celiac Patients (AIIMS)

Acharya P., ***Kutum R.***, et al., **Clinical and Translational Gastroenterology** (2018)

Telomere Biology

Contributed to the data analysis of ChIP-seq Data

Mukherjee AK. et al, **JBC** (2019)

Cardiovascular Disease

Contributed to the data analysis and machine learning approaches to discover features (genes) from the methylome

Ghose S. et al, Gene (2019)

Epilepsy

Performed transcriptomics data analysis of patients with epilepsy to find biomarkers

Rawat C. et al, **Sci. Report**, (2020)

COVID-19 [Public Health]

Performed the statistical modeling and data visualization of PanIndia sero-survey

Naushin S. et al, **eLife**, 2021

ASHOKA UNIVERSITY

Contributions in the field of Public Health (2021)

International DREAM Challenges

(2013-2021)

DREAM-Toxicogenetics, 40th rank

Got the exposure to large scale biomedical and genomics datasets. Deployed random forests algorithm to predict toxicity of drugs in humans

Eduati F et al., **Nature Biotechnology**, 2015

DREAM-Preterm Birth, *3rd Rank*

Developed ensembled machine learning algorithms to predict preterm birth from multi-omics data

Tarca AL et al., **Cell Reports Medicine**, 2021

DREAM-Single Cell Signaling in Breast Cancer, *9th Rank*

Developed semi-supervised biased-GLM approach to predict missing single cell markers

Gabor A et al., **Mol Sys Bio** (2021)

PrecisionFDA Challenges

(2017-2021)

Contributed to the machine learning algorithms to tackle mislabelling of samples

Proteogenomics (Yoo S. et al., **Patterns**, 2021)

Population Genetics

Unsupervised random forests (RF) algorithm to decipher genetic structure based on CNV

Narang A. et al, GBE (2014)

Asthma

Unsupervised and supervised RF algorithm to find discover subgroups and associated biomarkers from metabolites of children with Asthma

Sinha A. et al, **JTM** (2017)

Tuberculosis

Data science approach to discover drug tragets from 1800 Mtb clinical strains

Kaur D. et al, **Sci. Report** (2017)

Performed batch correction, preprocessing, and differential proteomics analysis

Menon D. et al, **ACS infectious diseases** (2019)

Genomewide investigation of codon-bias in the toxin-antitoxin genes of 1800 Mtb clinical strains (THSTI) Talwar S. et al, **mSystems** (2020)

Celiac Disease

Discovery of transcriptomics biomarkers for Celiac Patients (AIIMS)

Acharya P., ***Kutum R.***, et al., **Clinical and Translational Gastroenterology** (2018)

Telomere Biology

Contributed to the data analysis of ChIP-seq Data

Mukherjee AK. et al, **JBC** (2019)

Cardiovascular Disease

Contributed to the data analysis and machine learning approaches to discover features (genes) from the methylome

Ghose S. et al, Gene (2019)

Epilepsy

Performed transcriptomics data analysis of patients with epilepsy to find biomarkers

Rawat C. et al, **Sci. Report**, (2020)

COVID-19 [Public Health]

Performed the statistical modeling and data visualization of PanIndia sero-survey

Naushin S. et al, **eLife**, 2021

ASHOKA UNIVERSITY

Contribution in the field of Ayurgenomics (2014-2020)

International DREAM Challenges

(2013-2021)

DREAM-Toxicogenetics, 40th rank

Got the exposure to large scale biomedical and genomics datasets. Deployed random forests algorithm to predict toxicity of drugs in humans

Eduati F et al., **Nature Biotechnology**, 2015

DREAM-Preterm Birth, *3rd Rank*

Developed ensembled machine learning algorithms to predict preterm birth from multi-omics data

Tarca AL et al., **Cell Reports Medicine**, 2021

DREAM-Single Cell Signaling in Breast Cancer, *9th Rank*

Developed semi-supervised biased-GLM approach to predict missing single cell markers

Gabor A et al., **Mol Sys Bio** (2021)

PrecisionFDA Challenges

(2017-2021)

Contributed to the machine learning algorithms to tackle mislabelling of samples

Proteogenomics (Yoo S. et al., **Patterns**, 2021)

Population Genetics

Unsupervised random forests (RF) algorithm to decipher genetic structure based on CNV

Narang A. et al, GBE (2014)

Asthma

Unsupervised and supervised RF algorithm to find discover subgroups and associated biomarkers from metabolites of children with Asthma

Sinha A. et al, **JTM** (2017)

Tuberculosis

Data science approach to discover drug tragets from 1800 Mtb clinical strains

Kaur D. et al, **Sci. Report** (2017)

Performed batch correction, preprocessing, and differential proteomics analysis

Menon D. et al, **ACS infectious diseases** (2019)

Genomewide investigation of codon-bias in the toxin-antitoxin genes of 1800 Mtb clinical strains (THSTI) Talwar S. et al, **mSystems** (2020)

Celiac Disease

Discovery of transcriptomics biomarkers for Celiac Patients (AIIMS)

Acharya P., ***Kutum R.***, et al., **Clinical and Translational Gastroenterology** (2018)

Telomere Biology

Contributed to the data analysis of ChIP-seq Data

Mukherjee AK. et al, **JBC** (2019)

Cardiovascular Disease

Contributed to the data analysis and machine learning approaches to discover features (genes) from the methylome

Ghose S. et al, Gene (2019)

Epilepsy

Performed transcriptomics data analysis of patients with epilepsy to find biomarkers

Rawat C. et al, **Sci. Report**, (2020)

Ayurgenomics

1. Performed unsupervised RF algorithm to decipher underlying combinatorial nature of *Prakriti* questionnaire
2. Performed supervised RF algorithm to build predictive models of pure *Prakriti* types

Tiwari P., ***Kutum R.***, Sethi T., et al, **PLoS One**, (2017)

Performed data analysis and visualization of the genetic variations from the exome sequencing data of pure *Prakriti* types

Abbas T., ***Kutum R***, et al, (bioRxiv) under review, (2020)

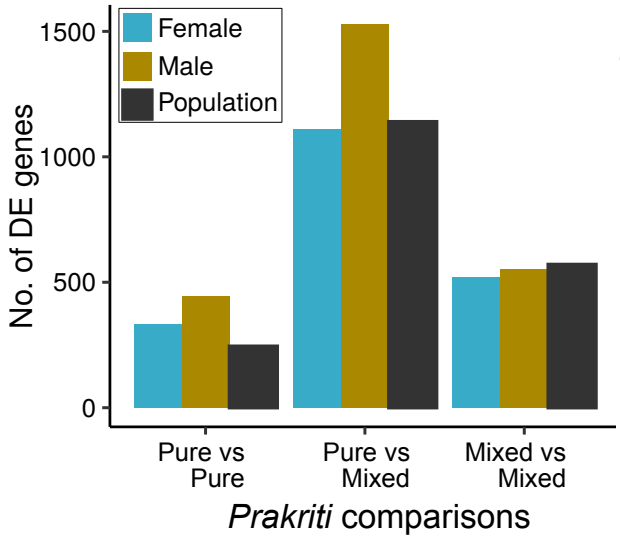
COVID-19 [Public Health]

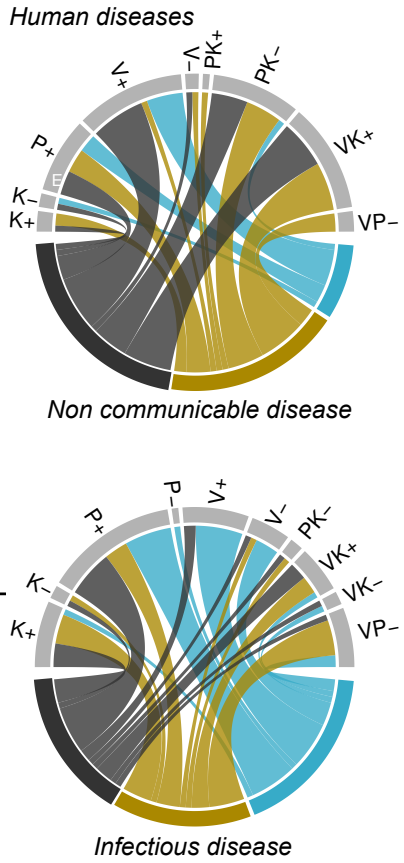
Performed the statistical modeling and data visualization of PanIndia sero-survey

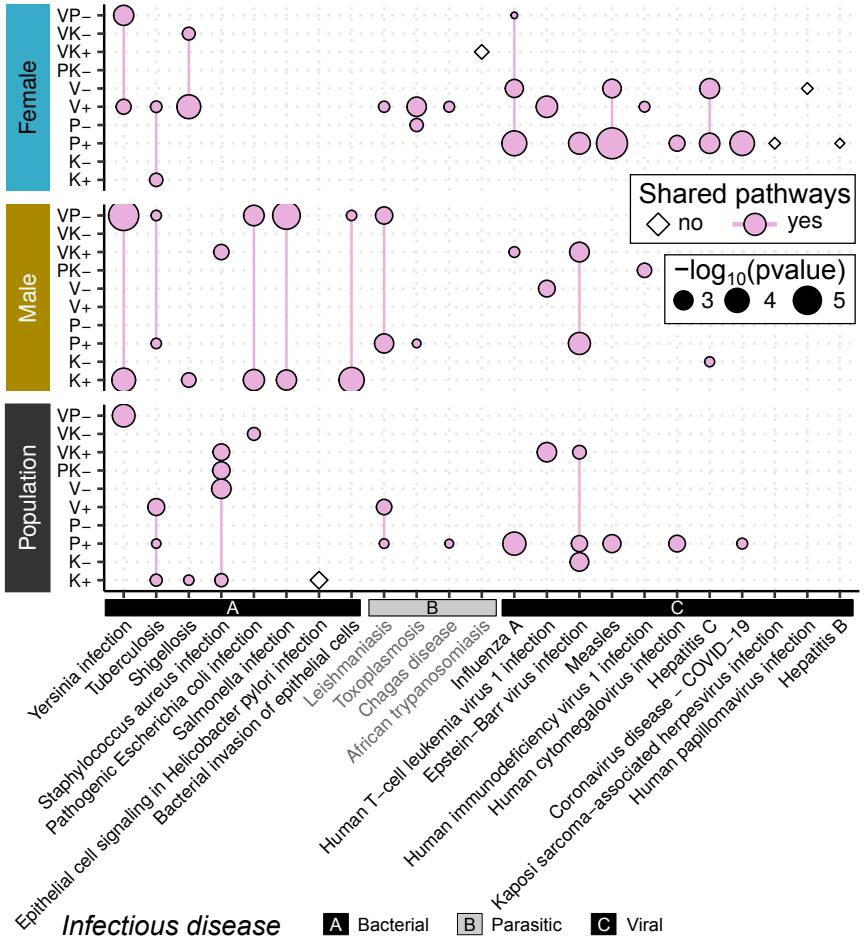
Naushin S. et al, **eLife**, 2021

ASHOKA UNIVERSITY

Interindividual transcriptomic differences across *Prakriti* types observed at the population & gender level

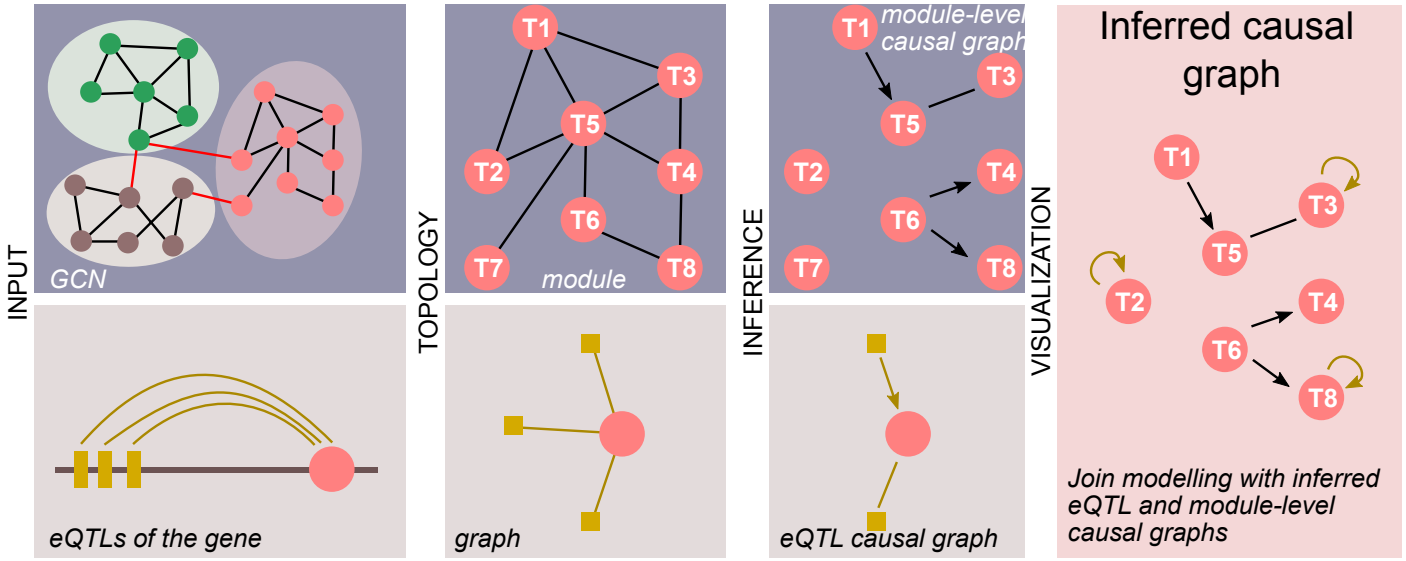






Developed "Knowledge Induced Causal Graph (*kicg*)" framework

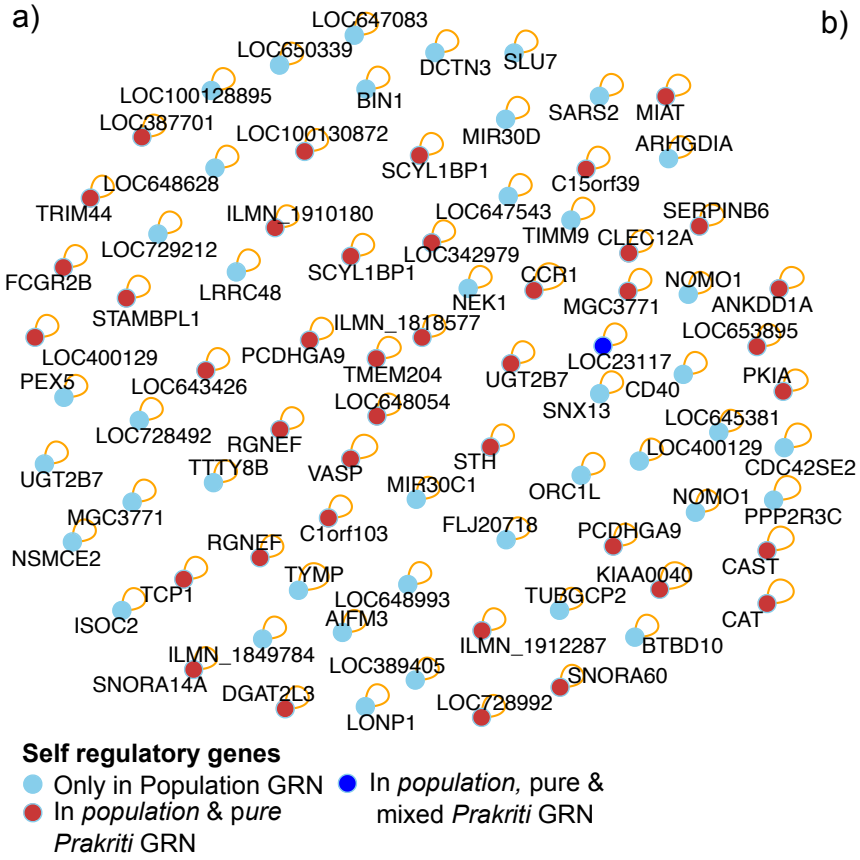
kicg: A framework to jointly model *module-level* and *eQTL-based* inference of causal graph (gene regulatory network) using ***MRPC*** algorithm and ***Bayesian network structure learning*** (bnlearn package) algorithm.



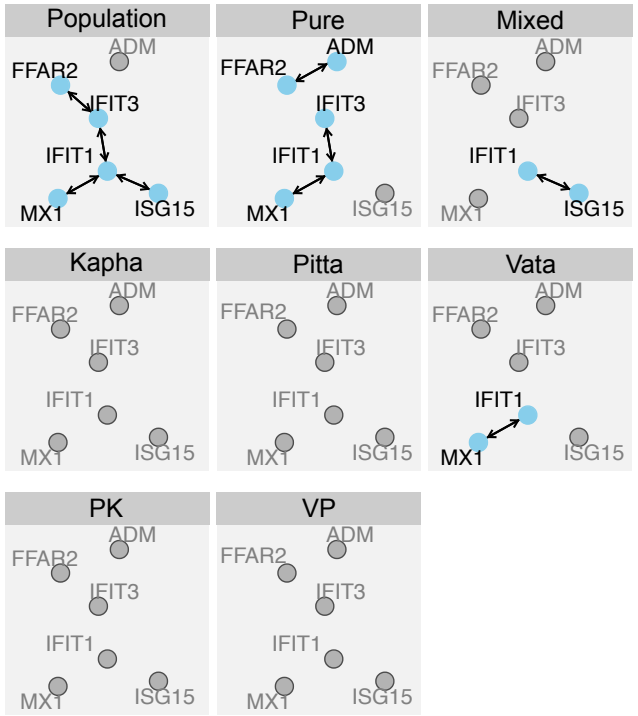
(Soon it will be available as ***'kicg'*** *R package*)

Causal gene networks inferred using 'kicg' framework for *Prakriti* types & the population

a)



b) Comparison of the MX1-gene set observed in the population GRN in *Prakriti* group GRNs



Differential connectivity in the MX1-geneset (subnetwork) was observed across population and *Prakriti* GRNs

Genomic Surveillance of COVID-19 Variants With Language Models and Machine Learning

Sargun Nagpal[1†], Ridam Pal[1†], Ashima[1‡], Ananya Tyagi[1‡], Sadhana Tripathi[1‡], Aditya Nagori[1], Saad Ahmad[1], Hara Prasad Mishra[1], Rishabh Malhotra[1], Rintu Kutum[1,2*] and Tavpritesh Sethi[1,3*]

[1]Indraprastha Institute of Information Technology Delhi, New Delhi, India
[2]Ashoka University, Sonipat, India
[3]All India Institute of Medical Sciences, New Delhi, India

Highlights

- We developed a genomic surveillance model for SARS-CoV-2 genome sequences, ***Strainflow***, where sequences were treated as documents with words (codons) to learn the codon context of 0.9 million spike genes using the *skip-gram algorithm*

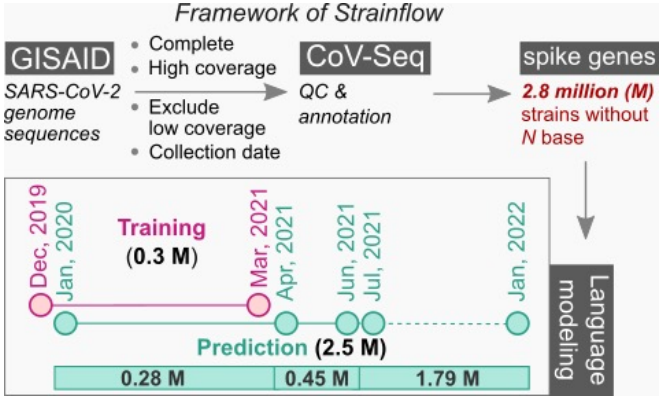
- *Machine Learning* modeling of the entropy of the latent dimensions helped us to develop an ***epidemiological early warning system*** for the ***COVID-19 caseloads***

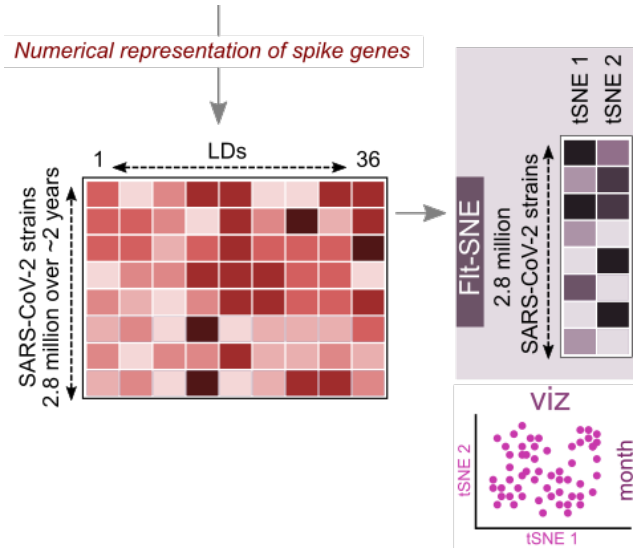
https://www.frontiersin.org/articles/10.3389/fgene.2022.858252/full

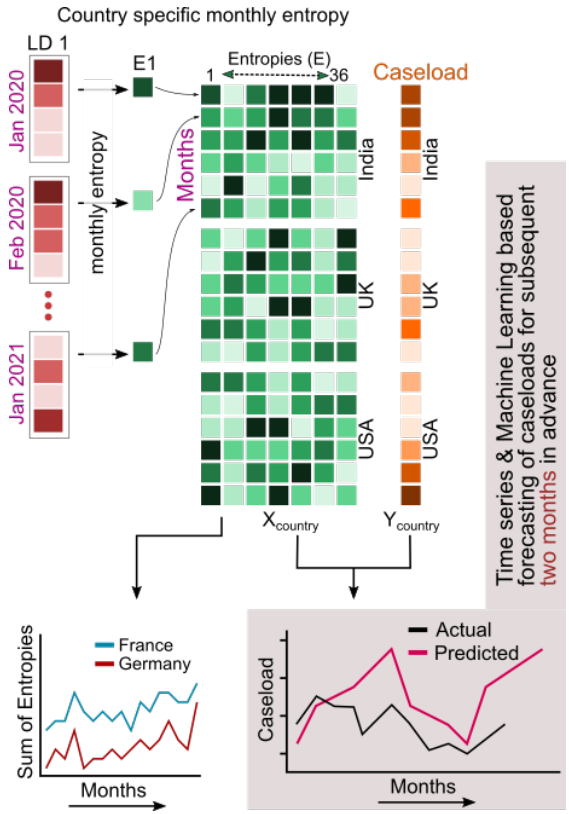
Genomic Surveillance of COVID-19 Variants With Language Models and Machine Learning

Sargun Nagpal[1†], Ridam Pal[1†], Ashima[1‡], Ananya Tyagi[1‡], Sadhana Tripathi[1‡], Aditya Nagori[1], Saad Ahmad[1], Hara Prasad Mishra[1], Rishabh Malhotra[1], Rintu Kutum[1,2*] and Tavpritesh Sethi[1,3*]

[1]Indraprastha Institute of Information Technology Delhi, New Delhi, India
[2]Ashoka University, Sonipat, India
[3]All India Institute of Medical Sciences, New Delhi, India







https://www.frontiersin.org/articles/10.3389/fgene.2022.858252/full

CSIR-IGIB internship trainer

Taught scientific computing for Genomics (R and python)

Trained 22 scholars from various institutes namely IIT Madras, KIIT University, VIT University, BITS (Pilani, Goa & Hyderabad), NIPER Kolkata, Gautam Buddha University, JAYPEE Information Technology (Solan) and AMITY University

2018 – *Hands-on workshop on R Programming Language*, ANDC, Delhi University, New Delhi
2018 – 5 days FDP on '*R programming for Biologists*', Manav Rachna International Institute of Research and Studies, Haryana
2019 – *Hands-on workshop on R Programming Language*, North East Hill University, Meghalaya

eLife FEATURE ARTICLE

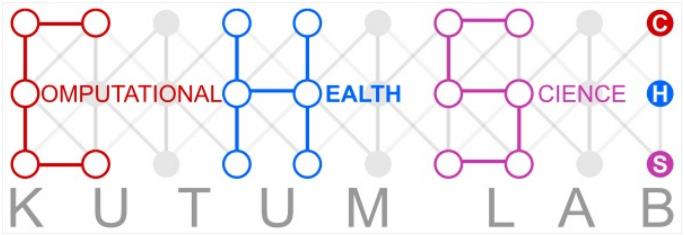
SCIENCE FORUM

A community-led initiative for training in reproducible research

Auer S et al., eLife 2021; https://elifesciences.org/articles/64719

Reproducibility for Everyone

https://www.repro4everyone.org/



Research Plans
Computational Health Science
Team Science

Themes

- Compute for Scholars (C4S) [Open Training, Nurture, Open Community, Support]
- Compute for Public Health (C4PH) [Open Training, contribute and collaborate with the experts the field]
- Computational Spatial Biology [Open Training, Benchmark, Develop]
  - Benchmarking ML algorithms in single cell biology (ML4SCB)
  - How do regulatory interactions within a cell facilitate cell-to-cell communication in the ecosystem of tissue (spatial biology)? (ML4SB)
- Facilitate Patient-centric Eco-Informatics to assist clinicians, researchers, & key partners

[Assist & Develop Technology for Patients with the stakeholders]

Community Science

Benchmark and Develop novel interpretable ML/AI algorithms for Single Cell Genomics

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

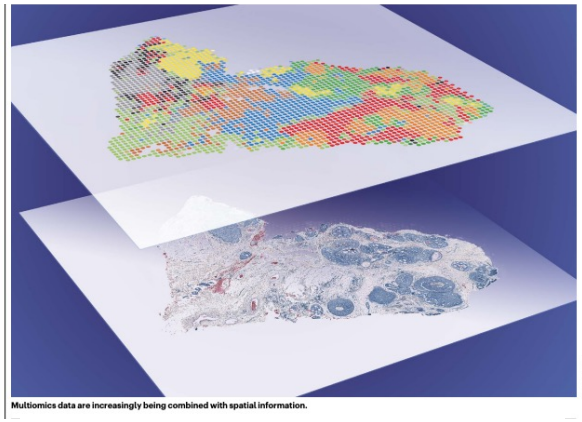
nature › technology features › article

TECHNOLOGY FEATURE | 19 July 2021 | Correction 21 July 2021 | Correction 24 August 2021

Single-cell analysis enters the multiomics age

A rapidly growing collection of software tools is helping researchers to analyse multiple huge '-omics' data sets.

Jeffrey M. Perkel



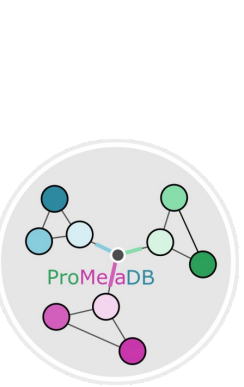
Multiomics data are increasingly being combined with spatial information.

Dozens of tools have been developed to achieve this, and many are indexed on the community-driven 'awesome-multi-omics' and 'awesome-single-cell' lists on GitHub.

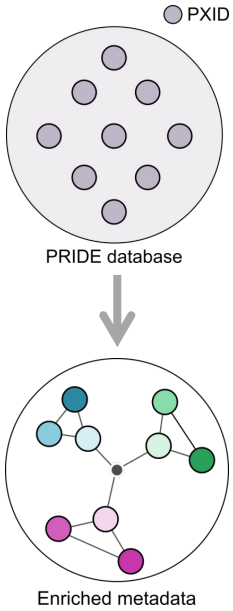
seandavi/awesome-single-cell: 2018-06-20-1

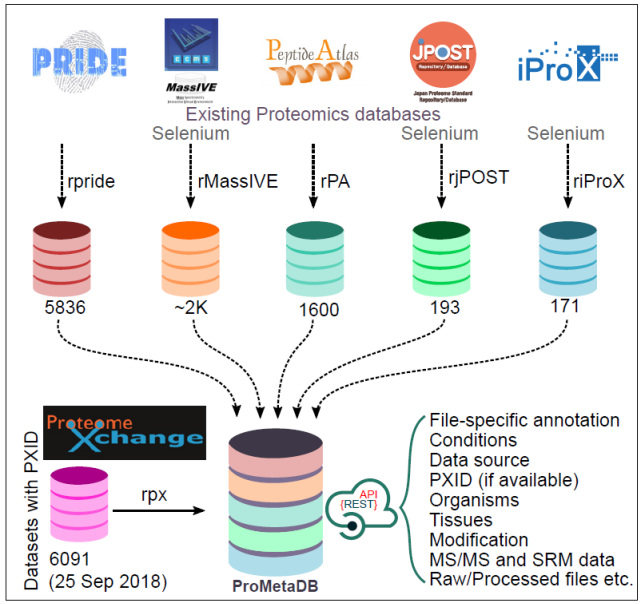
Sean Davis; Rintu Kutum; Luke Zappia; Jon Sorenson; Vladimir Kiselev; Poirion olivier; Olga Botvinnik; Keegan Korthauer; Anthony Gitter; Kieran R Campbell; Peter Hickey; Mustafa Anil Tuncel; MikeDMorgan; markrobinsonuzh; Catalina Vallejos; Zhe Wang; Nathan Salomonis; dyl4nm4rsh4ll; Daniel Wells; Yun YAN; Viktor Petukhov; Thomas Wolfgruber; Robert Aboukhalil; Patrick Roelli; Michael Kelly; Joshua Welch; John Reid; David DeTomaso; AlessandraDM; Alex Wolf

List of software packages for single-cell data analysis, including RNA-seq, ATAC-seq, etc.



An unified resource for proteomics metadata



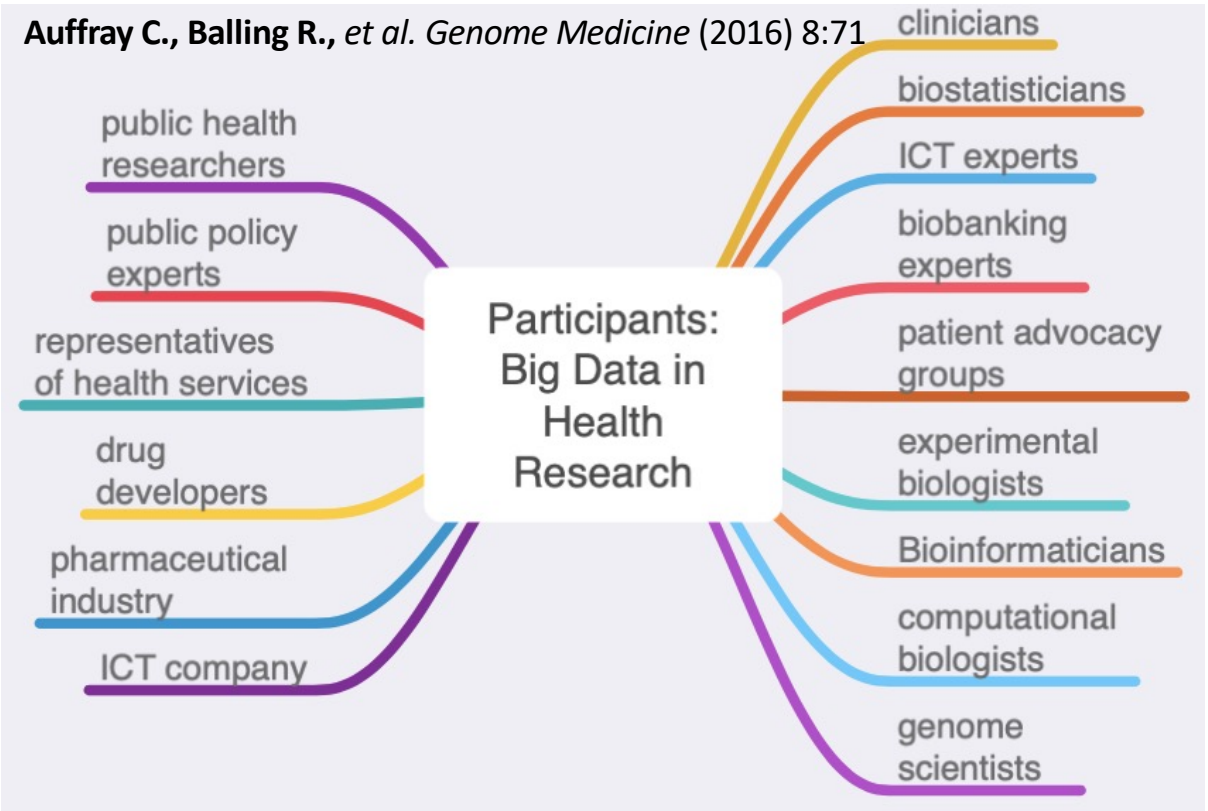


https://github.com/rintukutum/prometadb
https://prometadb.igib.res.in/prometadb/index.html

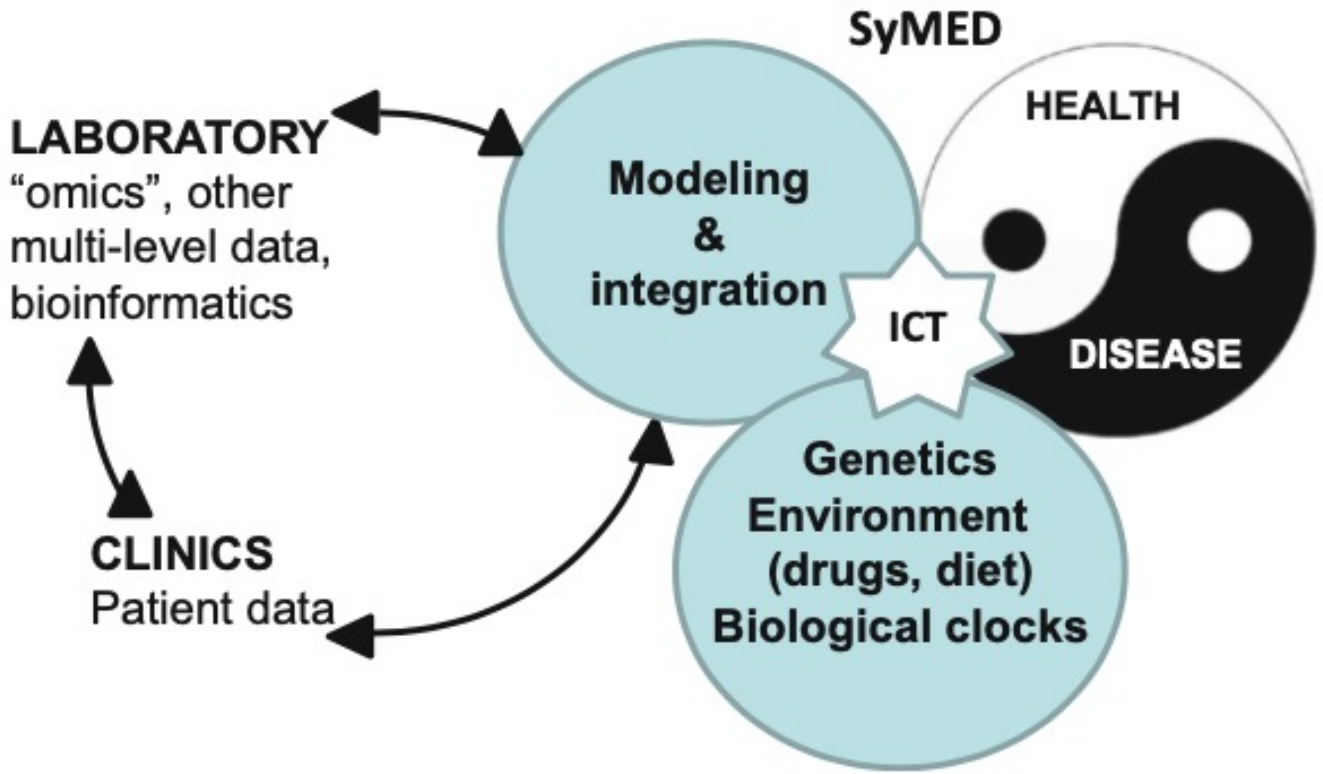
Contribution of Genome Informatics in Biomedical Reseearch

"Big data in health" encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points.

Auffray C., Balling R., *et al. Genome Medicine* (2016) 8:71



Systems Medicine Education and Training



Review > Methods Mol Biol. 2016;1386:73-86. doi: 10.1007/978-1-4939-3283-2_5.

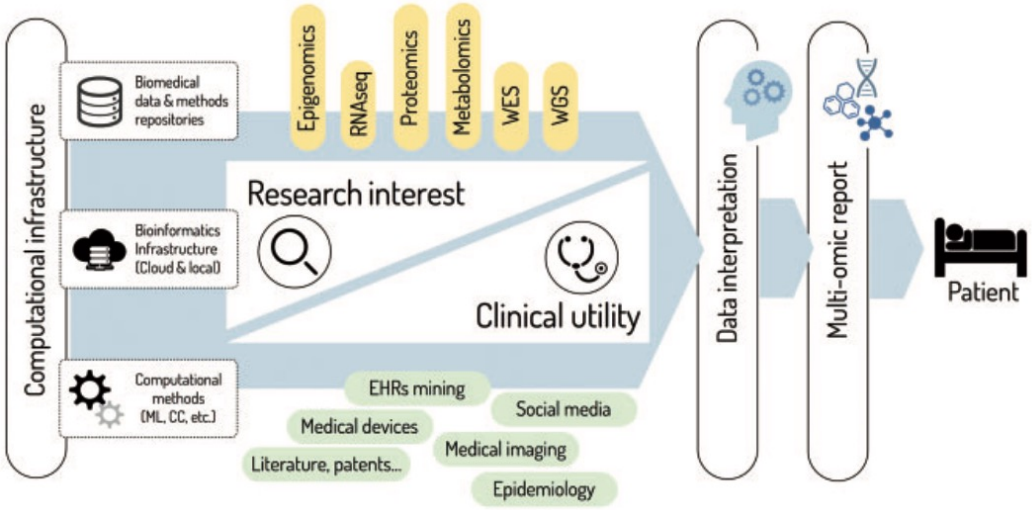
Training in Systems Approaches for the Next Generation of Life Scientists and Medical Doctors

Damjana Rozman [1], Jure Acimovic [2], Bernd Schmeck [3 4]

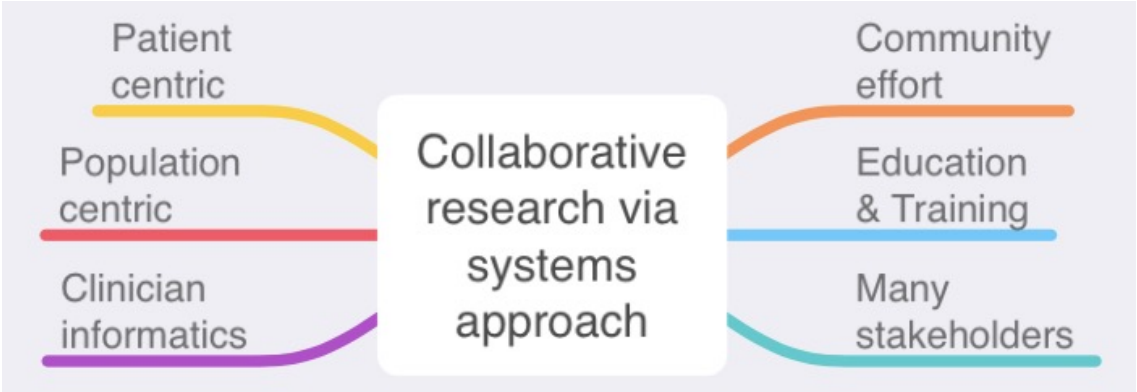
Affiliations + expand

PMID: 26677180 DOI: 10.1007/978-1-4939-3283-2_5

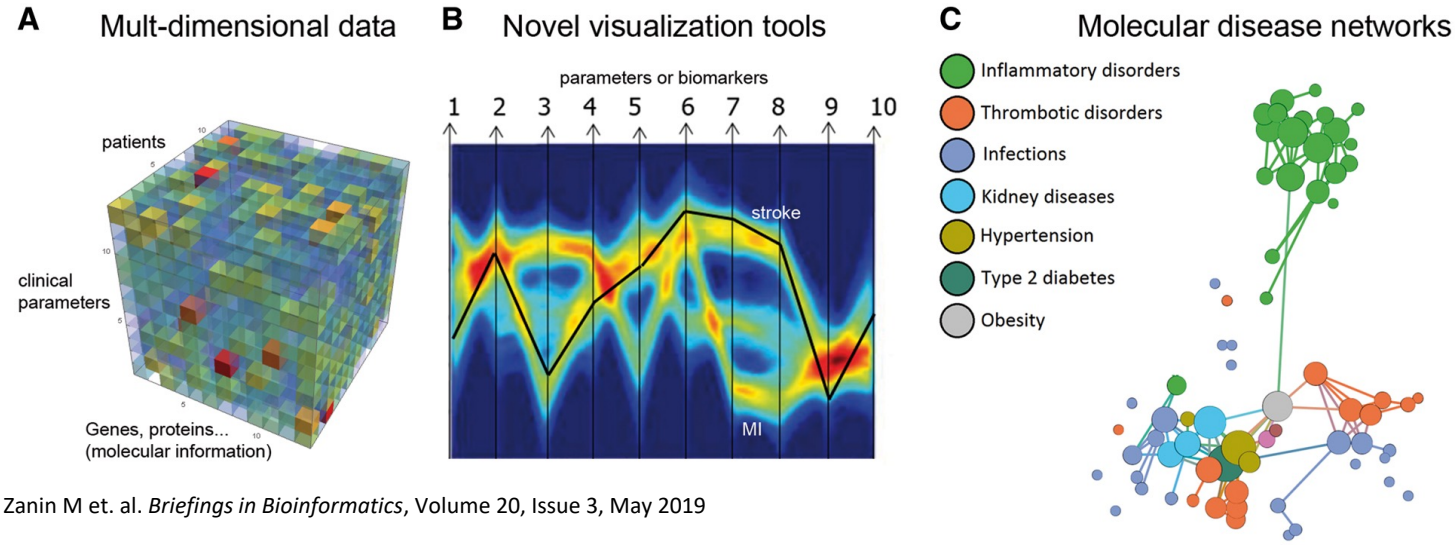
Precision medicine needs pioneering clinical bioinformaticians



Gómez-López G. et. al. *Briefings in Bioinformatics*, May 2019



Community effort endorsing *multiscale* modelling, *multiscale* data science and *multiscale* computing for systems medicine



Zanin M et. al. *Briefings in Bioinformatics*, Volume 20, Issue 3, May 2019

- Most diseases are characterized by **a complex interplay of multiple factors**, which is the focus of the new **systems medicine** approach.
- Most disease definitions are 18th/19th century and organ- or symptom-based and may mechanistically combine entirely different phenotypes.
- A necessary step is the **identification of barriers and challenges** of multiscale systems medicine **through a collective effort.**

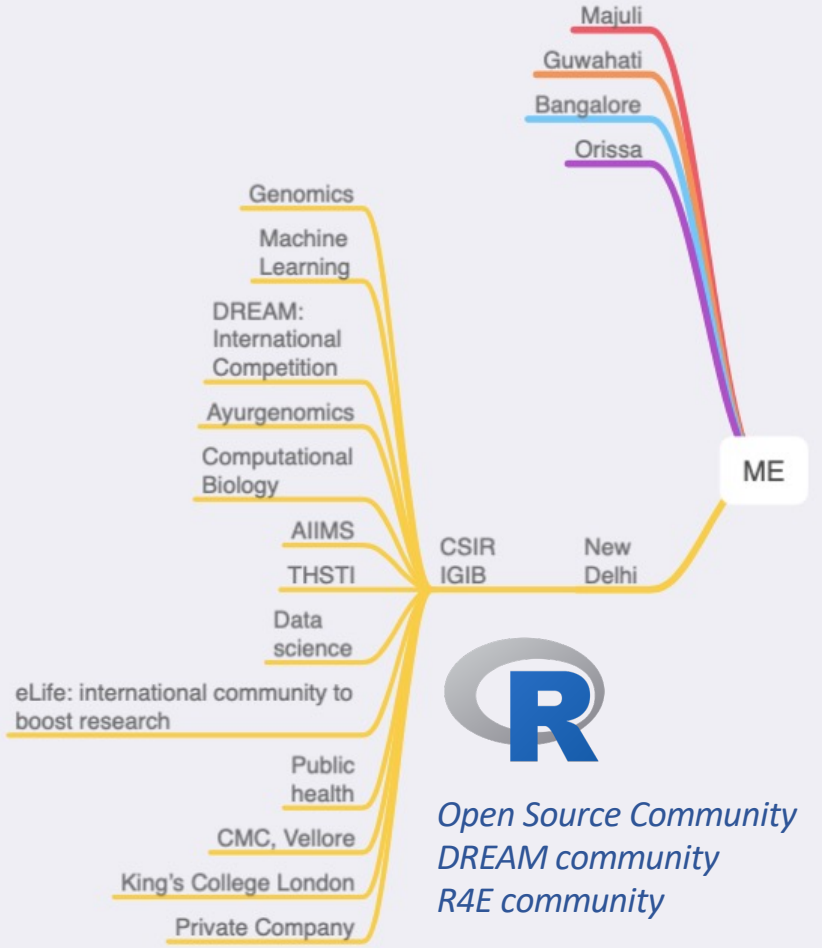
Inter-disciplinary mentors (2012-2022)

Dr. Tavpritesh Sethi

MBBS -> CS + Ayurveda + Physiology

Rishi Das Roy, PhD

CS -> Biological Science



M.Sc. Dissertation mentor

Dr. Anurag Agrawal

Ph.D. supervisor

Dr. Debasis Dash

Dr. Bhavana Prasher

Research mentors

Dr. Mitali Mukerji

Dr. Rajesh S. Gokhale

Prof. Samir K. Brahmachari

Dr. Rohit Vashisht (UCSF)

Dr. Sheetal Gandotra (IGIB)

Dr. Shantanu Sengupta (IGIB)

Dr. Shantanu Chowdhury (IGIB)

Dr. Govind K. Makharia (AIIMS)

Dr. Neena Malhotra (AIIMS)

Dr. Amit Kumar Pandey (THSTI)

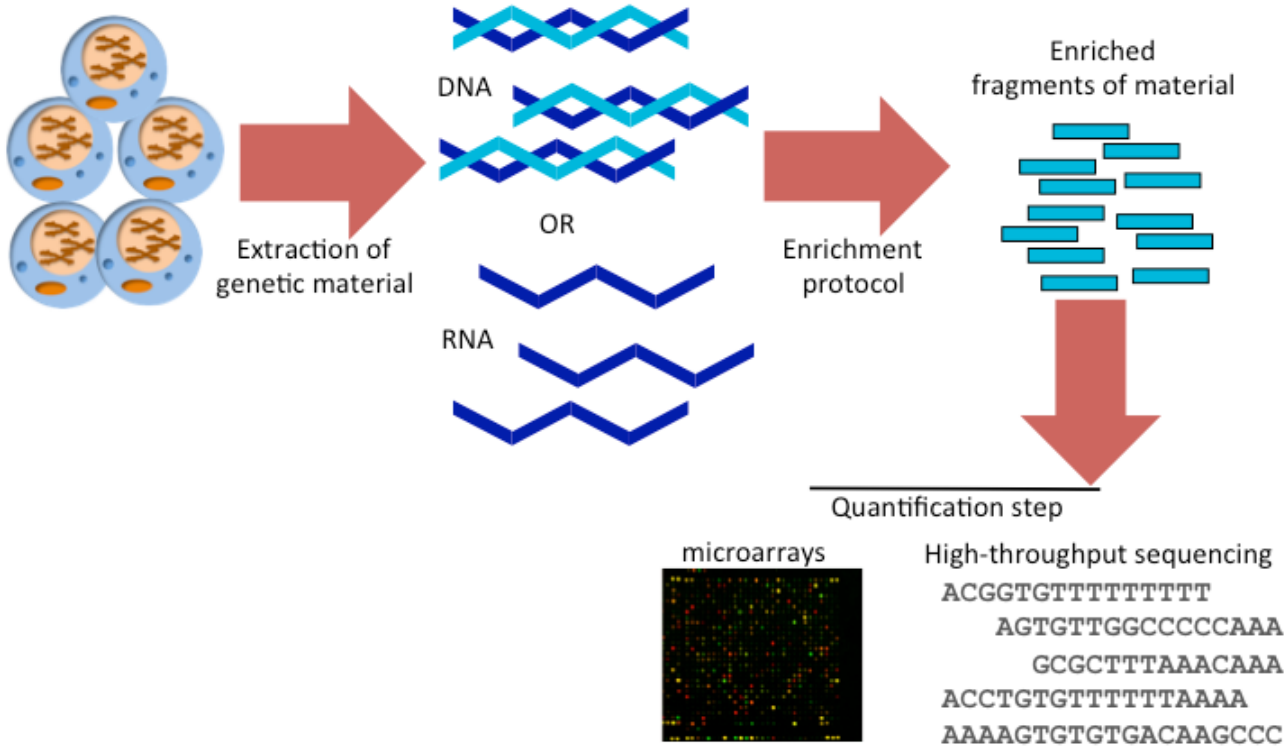
eLife Community & many …

Current mentors

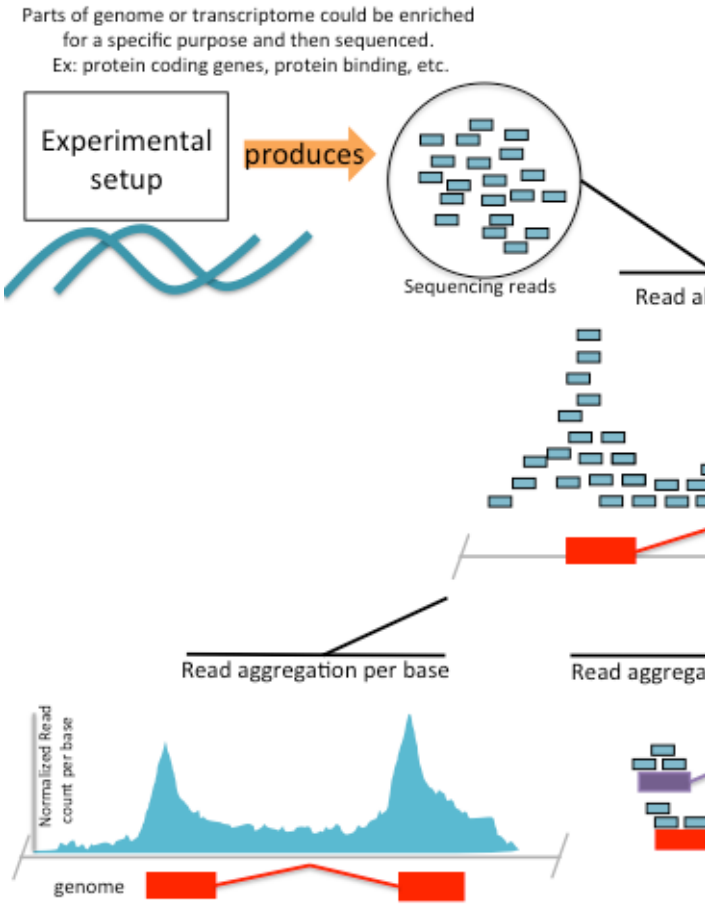
A. Prof LS Shashidhara, **B.** Prof Subhashis Banerjee, **C.** Prof Alok Bhattacharya, **D.** Prof Anurag Agrawal

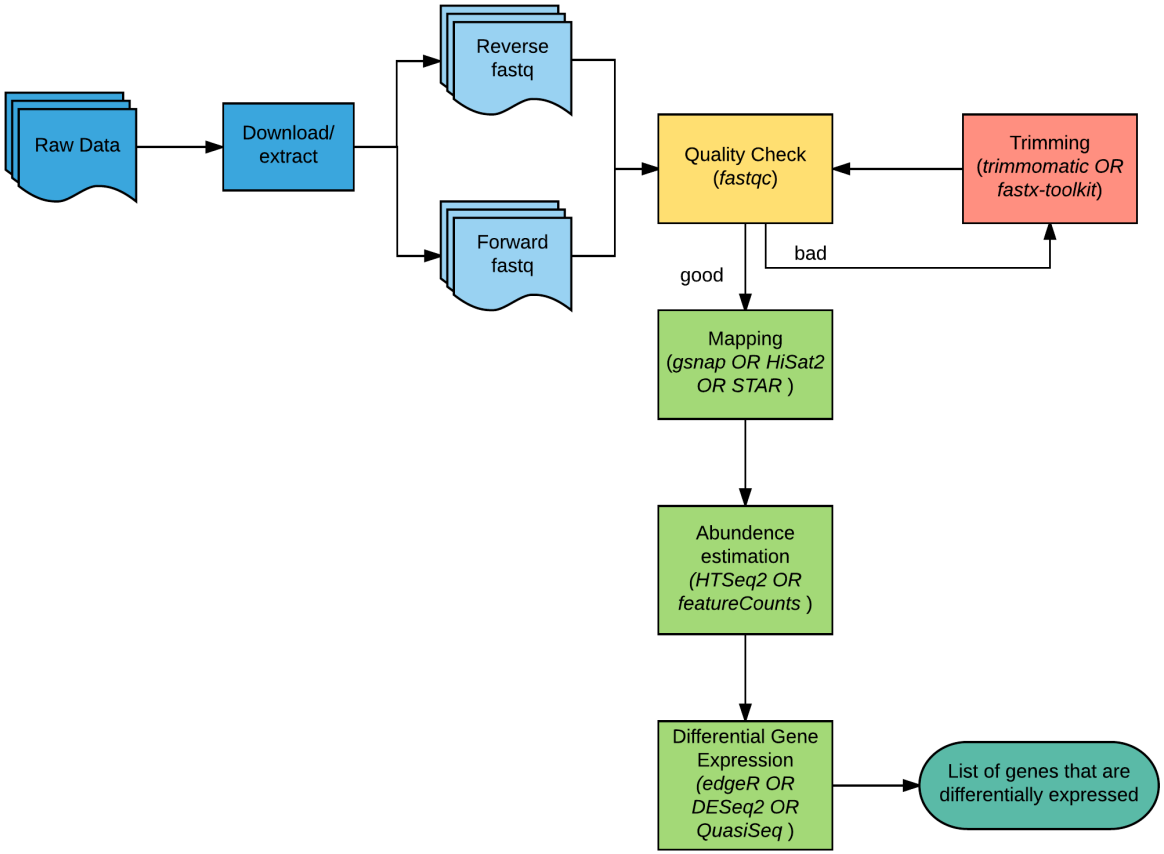
ASHOKA UNIVERSITY

Common steps of high-throughput assays in genome biology



High-throughput sequencing summary and workflow





THANK YOU