Omics Data Integration in Biomedical Research and Precision Medicine

Jornadas SEFORI, CITIC-UGR

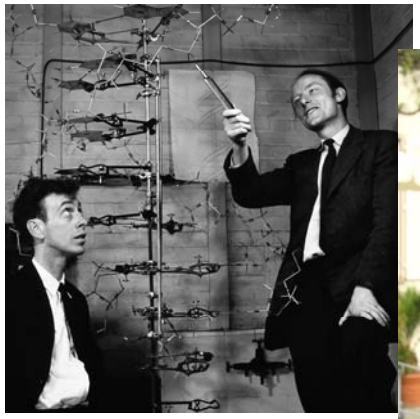
Oct 2018

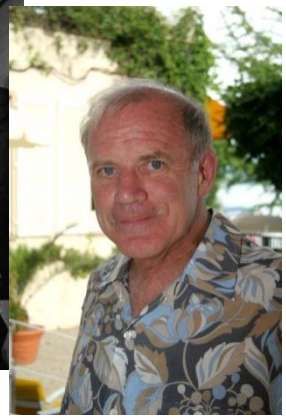
Pedro Carmona-Saez

Genyo Bioinformatics Unit

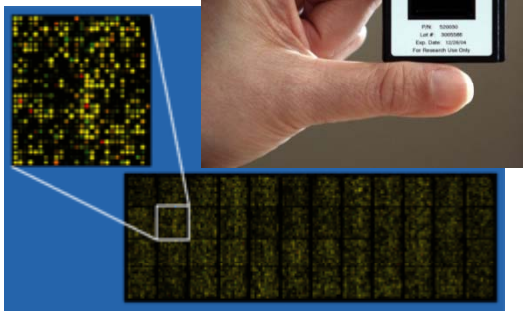
Biomedical research guided by the evolution of technology

1960-80





Biomedical research guided by the evolution of technology



1960-80 → **Late 90s**

REPORT

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

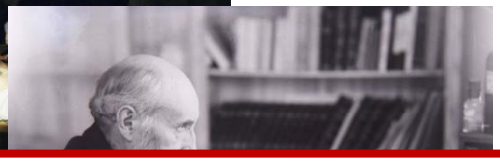
Mark Schena[1], Dari Shalon[1], Ronald W. Davis[2], Patrick O. Brown[3]

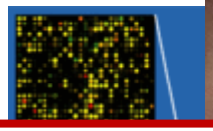
+ Author Affiliations

[1] These authors contributed equally to this work.
[2] Present address: Synteni, Palo Alto, CA 94303, USA.
[3] To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

Science 20 Oct 1995:
Vol. 270, Issue 5235, pp. 467-470
DOI: 10.1126/science.270.5235.467

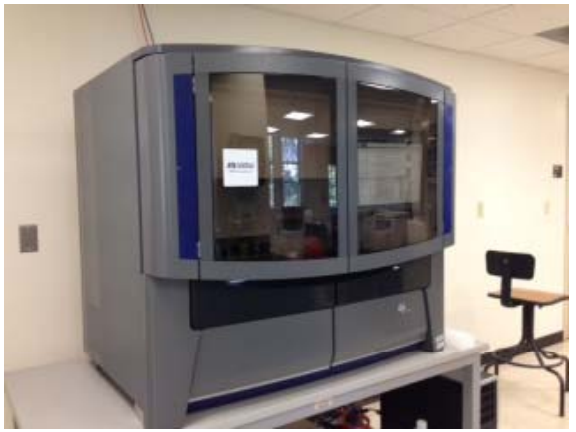
Biomedical research guided by the evolution of technology



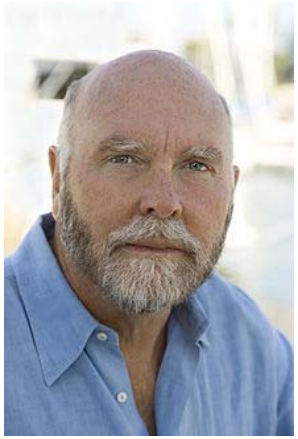


Archaeomics
Epigenomics
Morphomics
Regulomics
Toxicogenomics
Phosphoproteomics
Proteomics
Kinomics
Alternatomics
Glycomics
Behavioromics
Metabolomics
Lipoproteomics
Secretomics
Genomics
Transcriptomics
Lipidomics
Fluxomics
Interactomics
Orfeomics

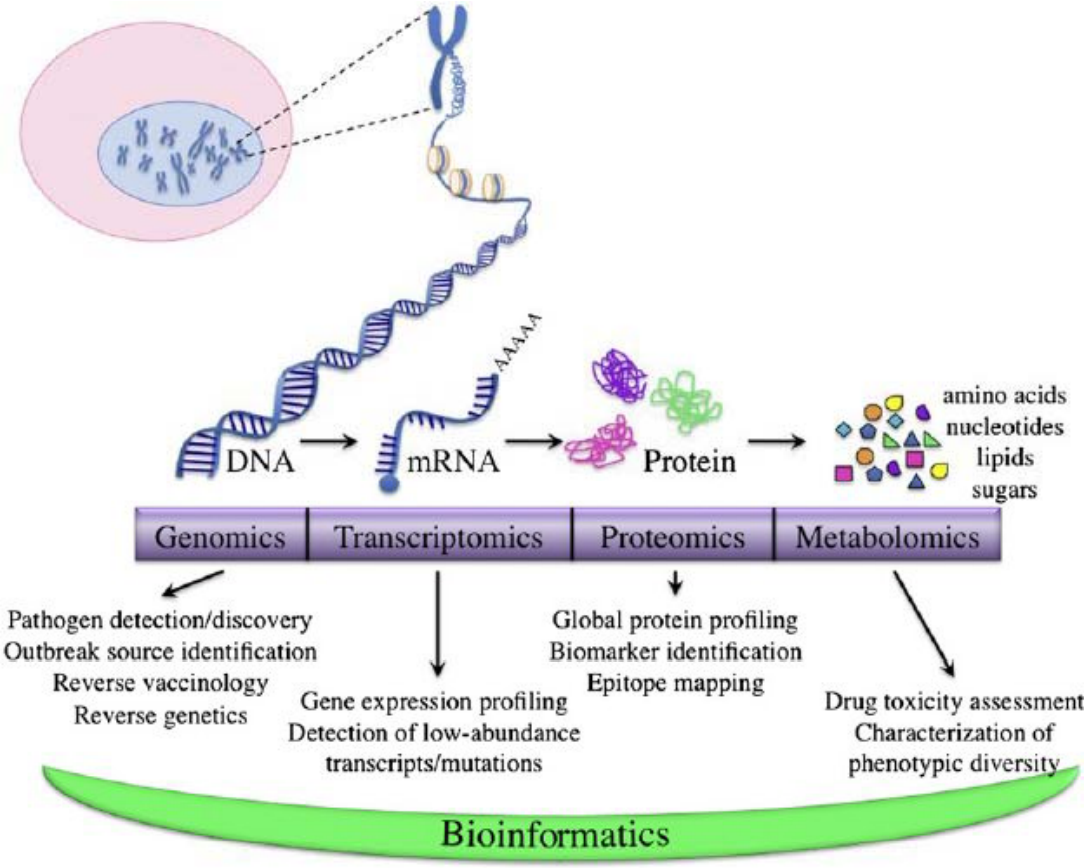
Biomedical research guided by the evolution of technology

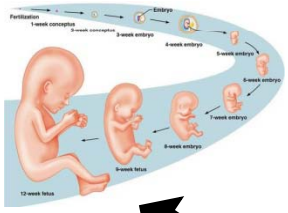


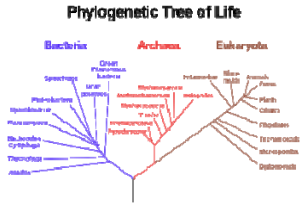
2001 — **2005**



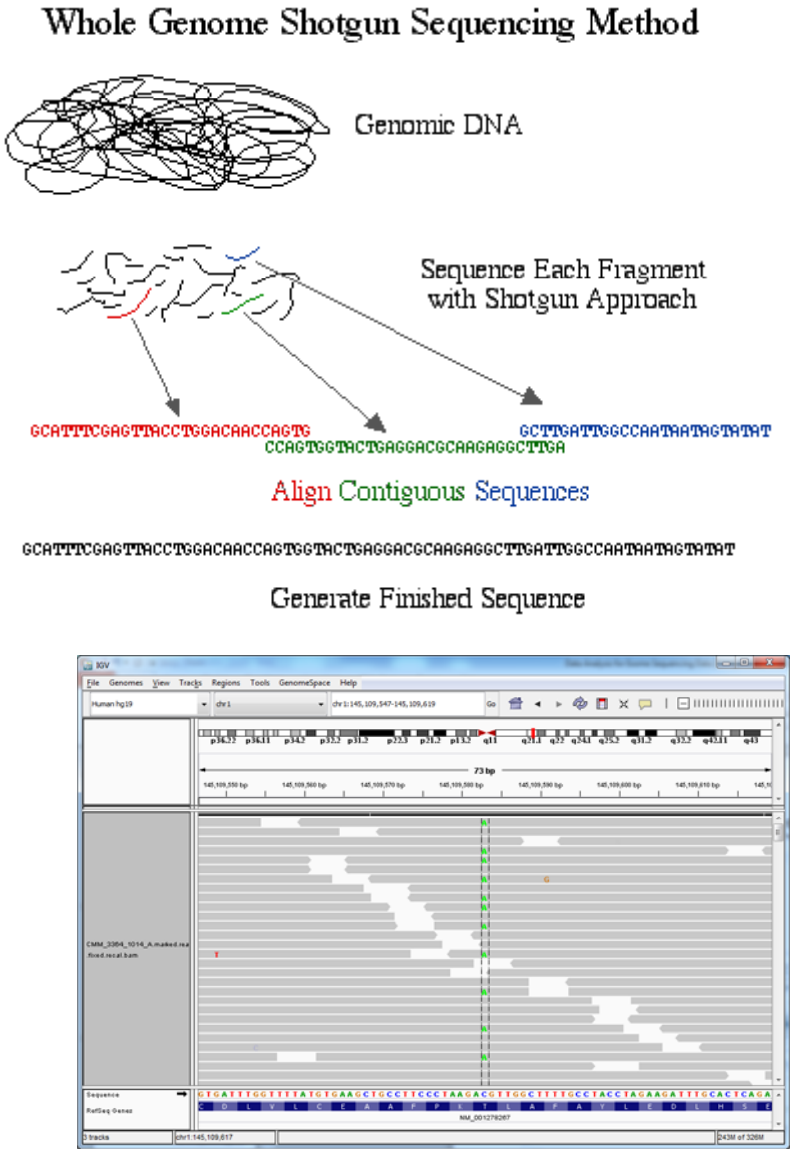
The code of life

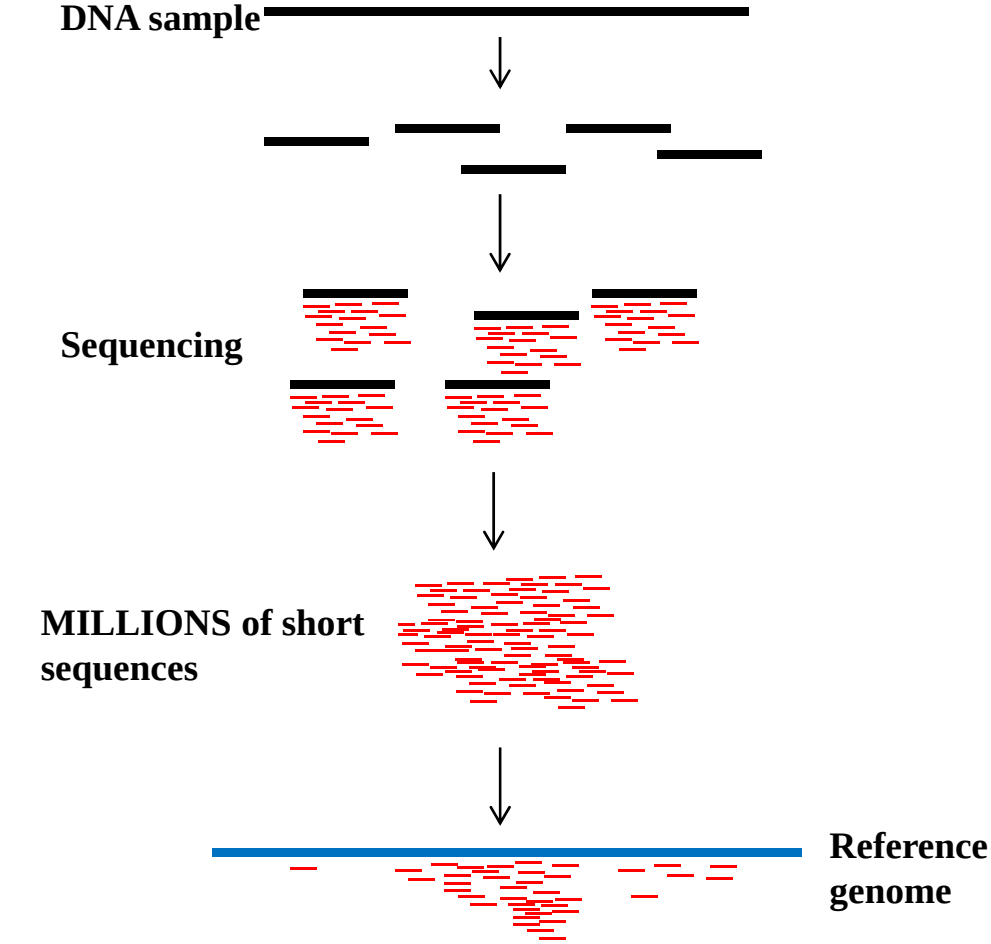






Next Generation Sequencing





Next Generation Sequencing

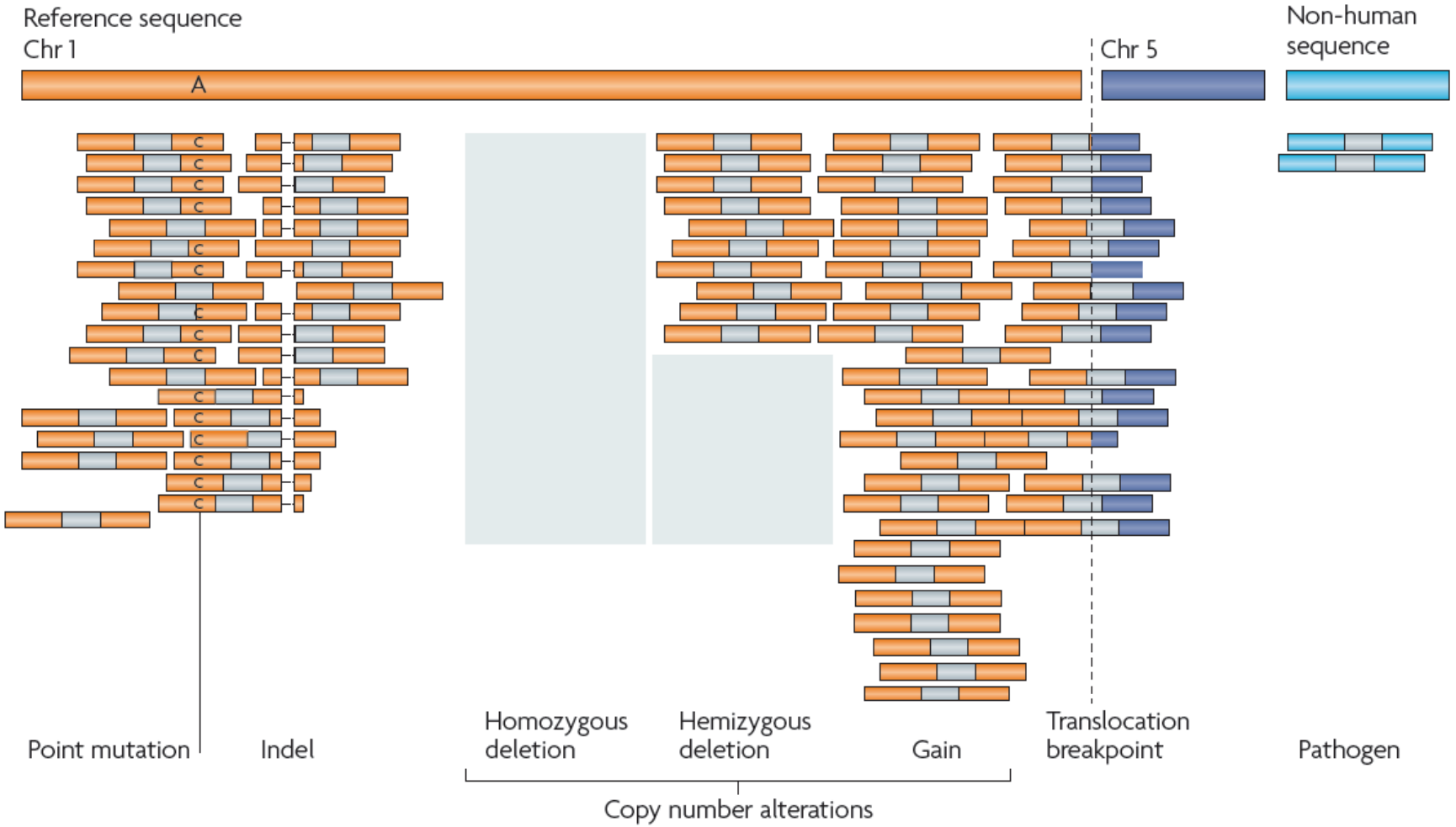


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

Meyerson et al. NRG 2010

Pricing trend

(2001) The Human genome project:

- 13 years

- 23 labs

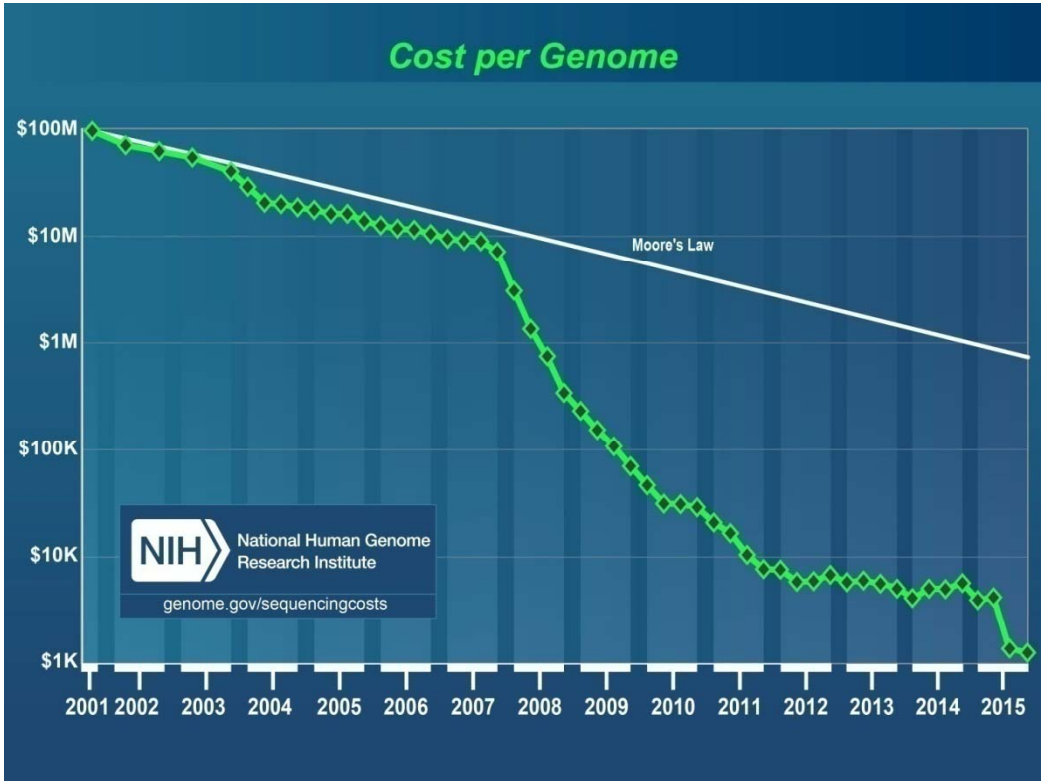
- $500 Million.

• A Human genome today:

- 1 day

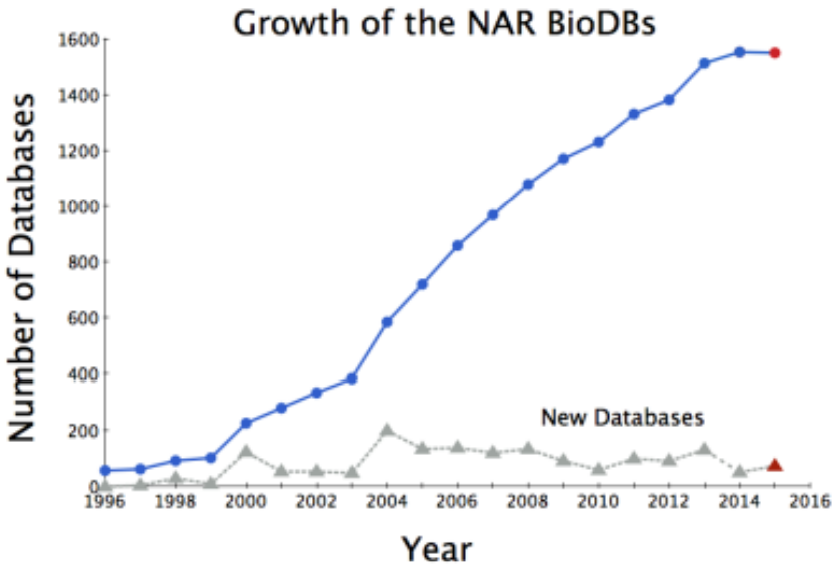
-1 machine

- $1,000



Increasing Information and Resources

Exponential growth of Biological Databases and information



Growth of Omics Data (EBI Repositories)

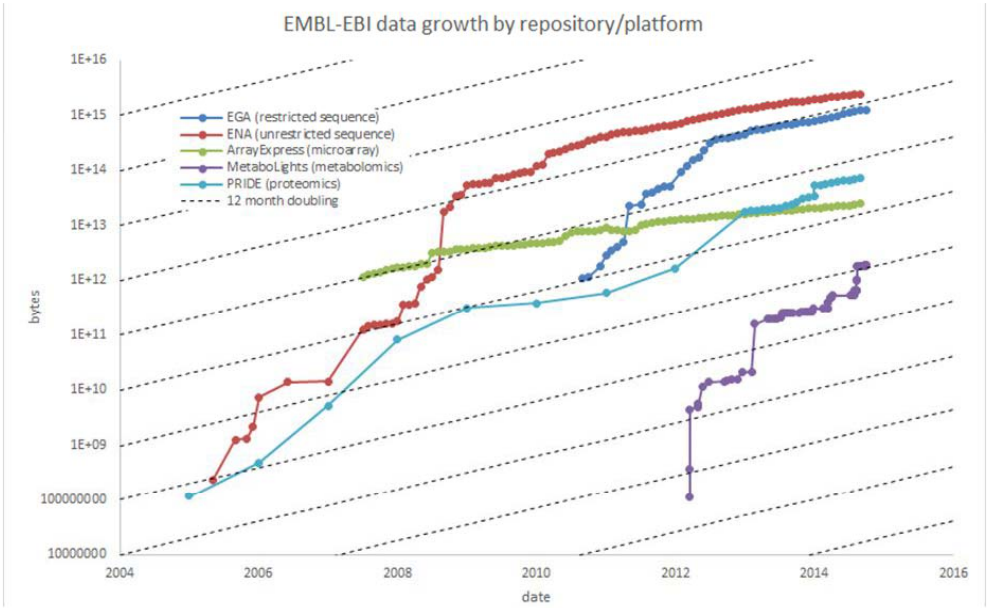


Illustration: Christoph Steinbeck, EBI

Big Data in Biology

PERSPECTIVE

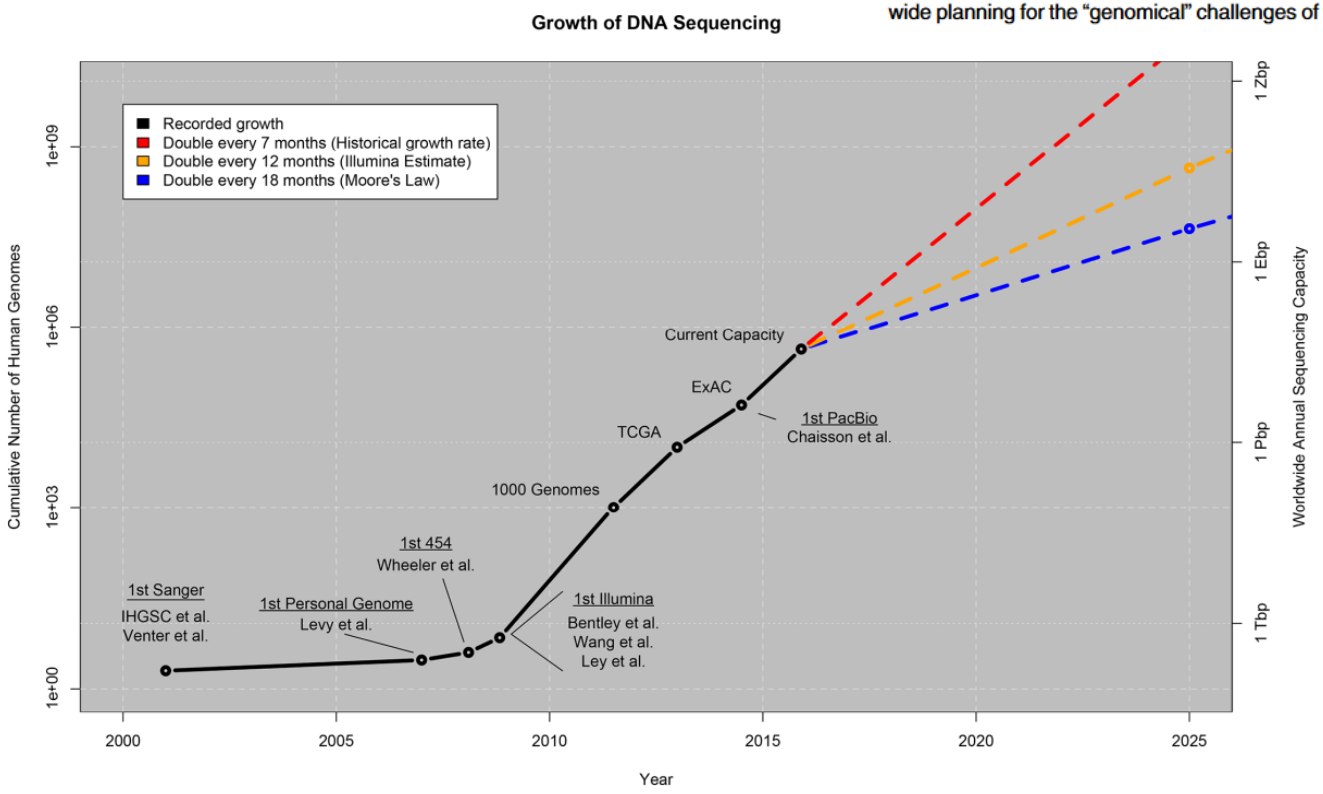
Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

PLOS Biology 2015

Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.



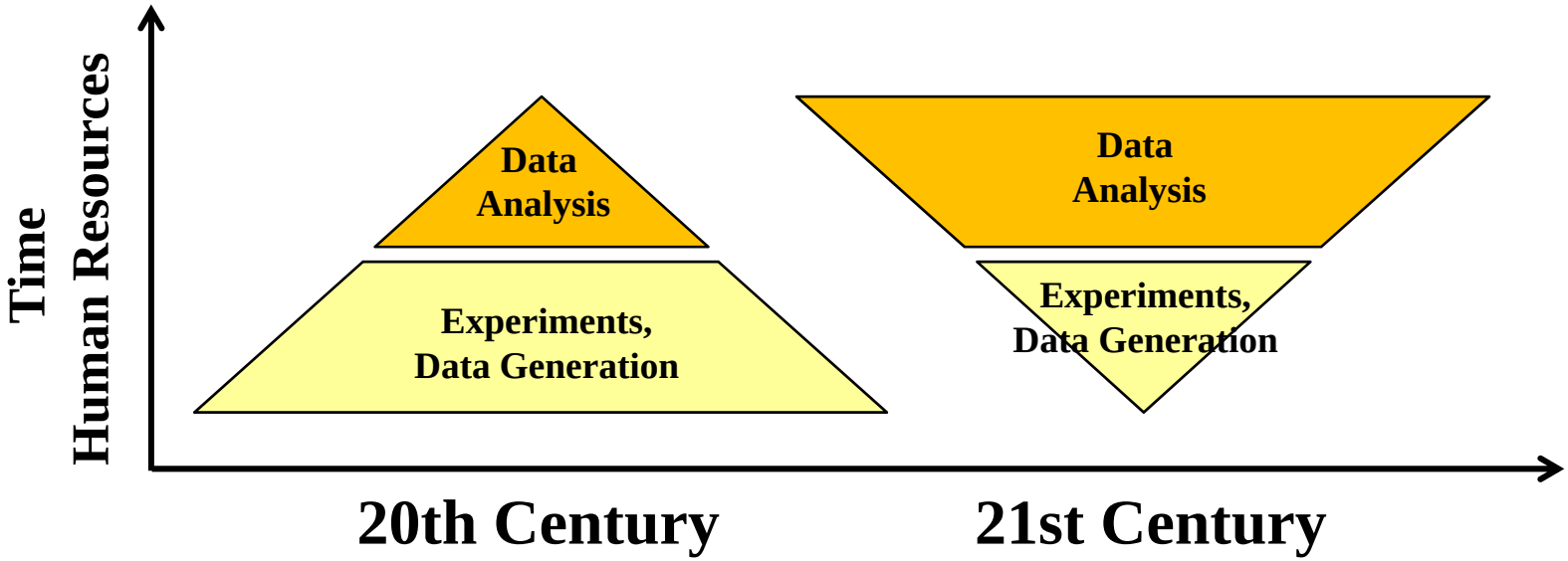
Biological research has changed

Mardis *Genome Medicine* 2010, 2:84
http://genomemedicine.com/content/2/11/84

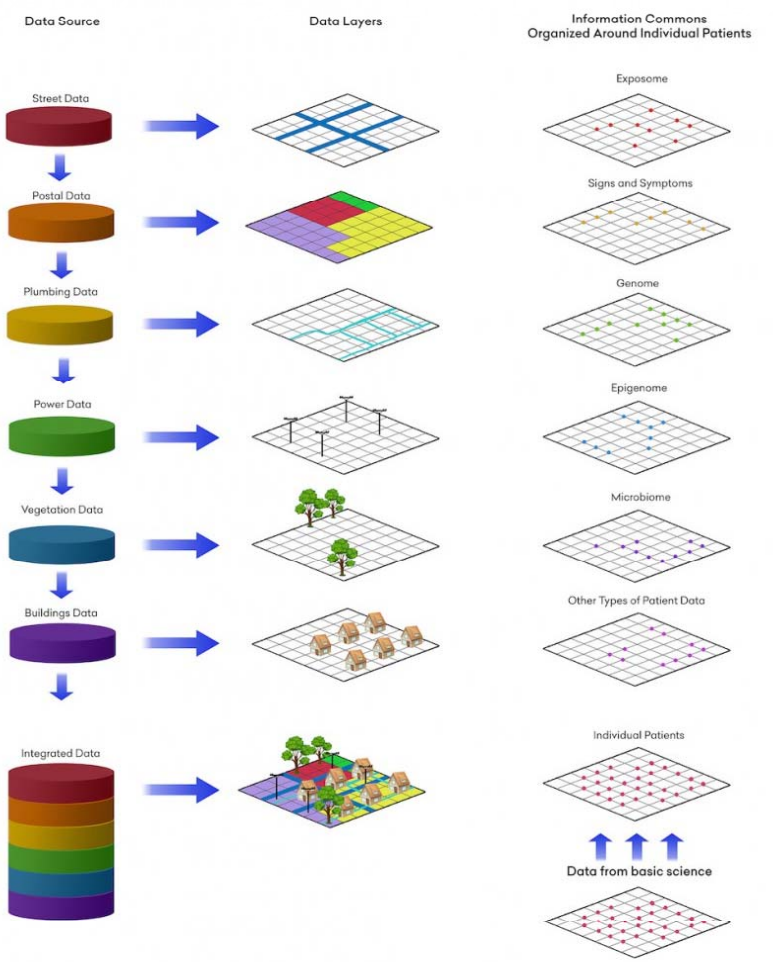
MUSINGS

The $1,000 genome, the $100,000 analysis?

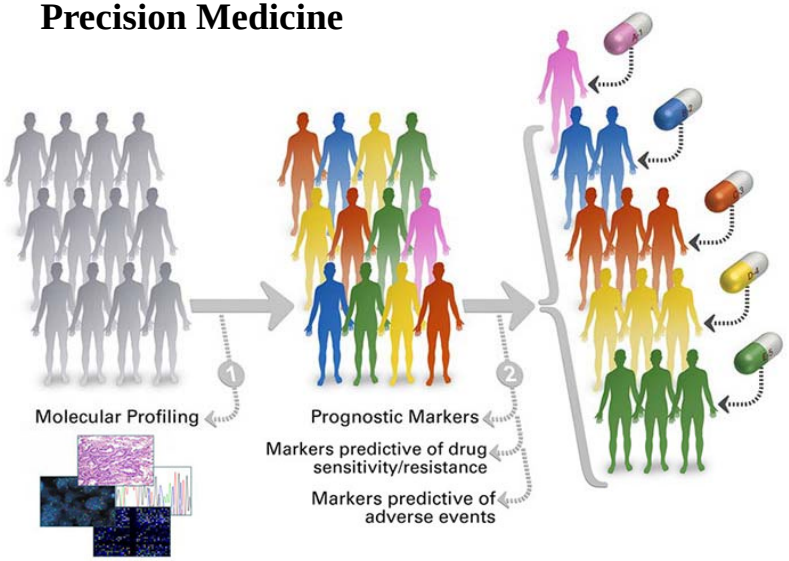
Elaine R Mardis*



Precision Medicine



Precision Medicine



Traditional Approach

What are the Challenges?

Storage

- Fast and efficient storage systems to query large collections of ‘omics data.

Distribution and data access

- Cloud-computing systems such as EasyGenomics (BGI) or “Embassy” as part of ELIXIR project

Analysis

- Algorithms/methods to **extract knowledge and information from the data**
- Integration of biological domain expertise, large-scale machine learning systems and efficient computing infrastructure

A COMPLEX TASK

Biological systems are complex. Genes may carry out different functions in different cell types / tissues, even antagonist functions.

- **Signal-to-noise ratio and confounding.** High throughput biological data have low signal-to-noise ratio and many variables that make it difficult to distinguish signals from random patterns.

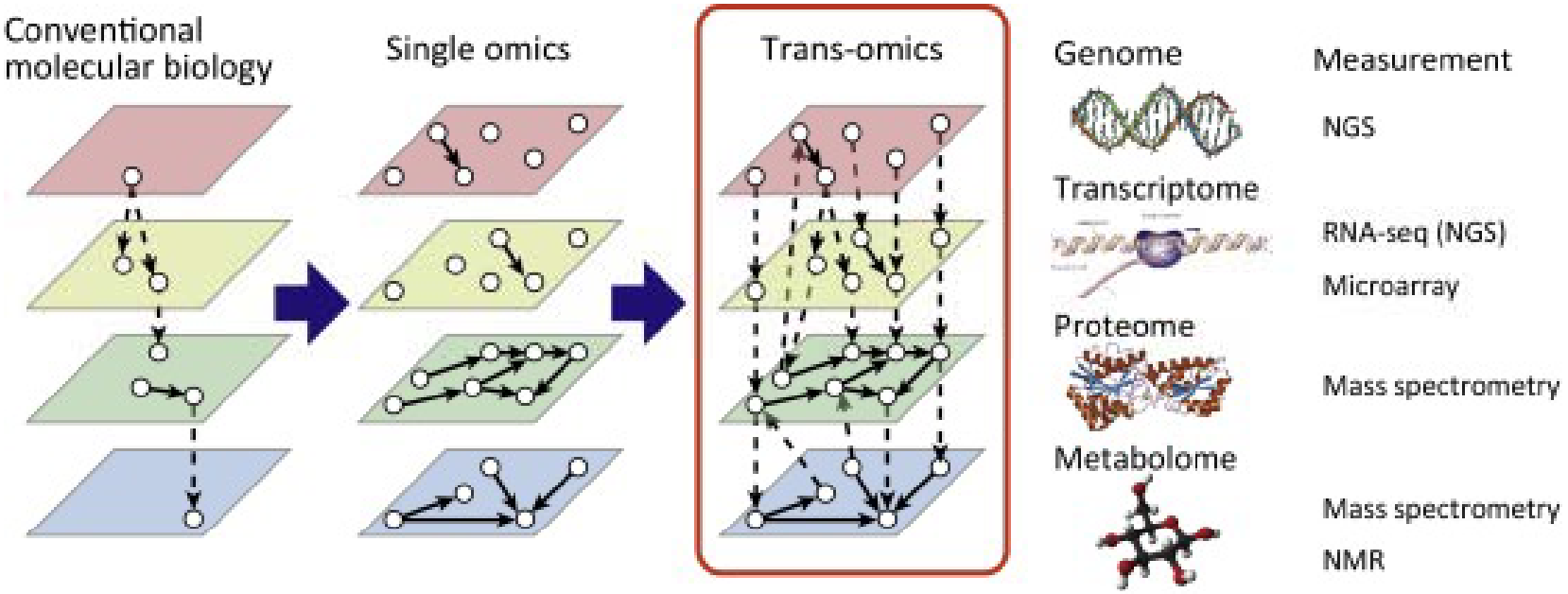
- **High-dimensionality**, large number of variables and small number of samples

- **Biological systems are dynamic**

Integrating Omics Data

- **Integrating different studies/datasets with the same type of data**. Integrate data with the thousands of other published datasets and look for similar patterns.

- **Integrarting multiple data types.** Integrate different omics data (genomics, transcriptomics, metabolomics, …)

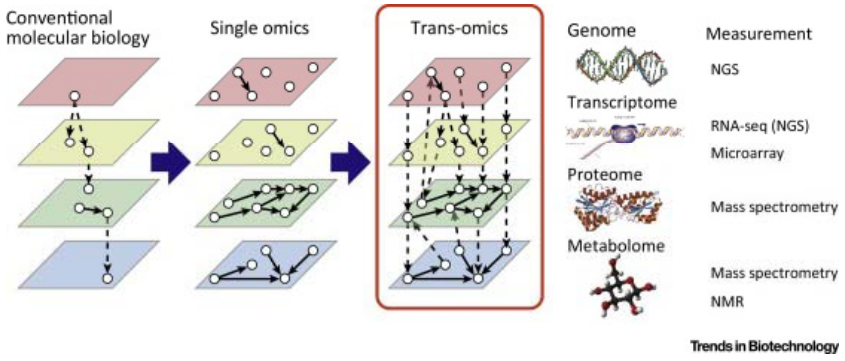


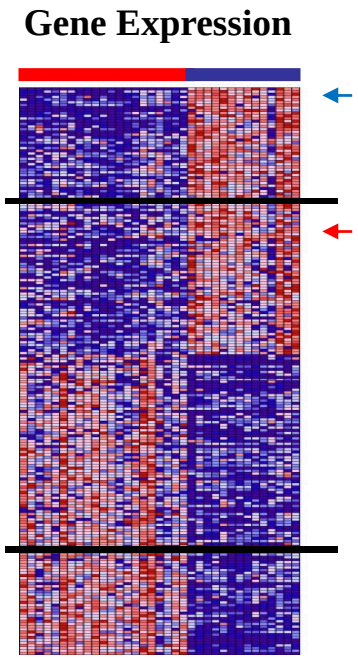
Multi-omics integration for Biomarker Discovery

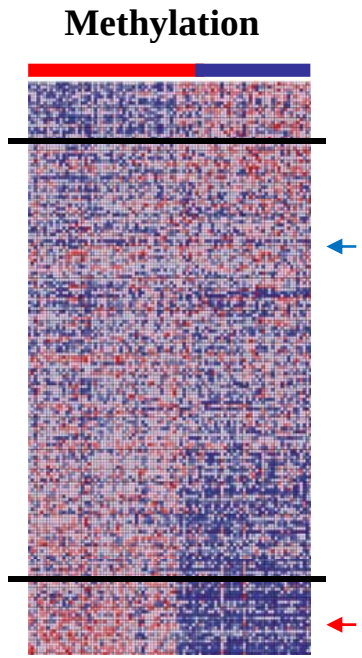
An increasing number of projects include measurements of the same samples from multiple omics techniques.

Integrating different OMICS layers will potentially provide:

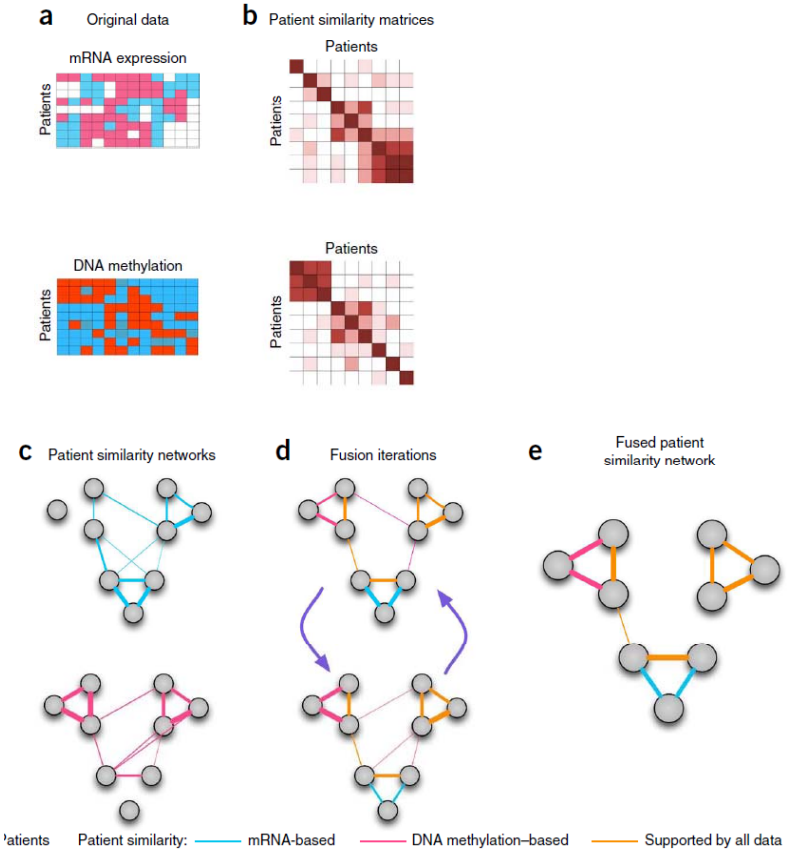
- More comprehensive view of biological processes
- More robust patterns
- Better predictions and classifications



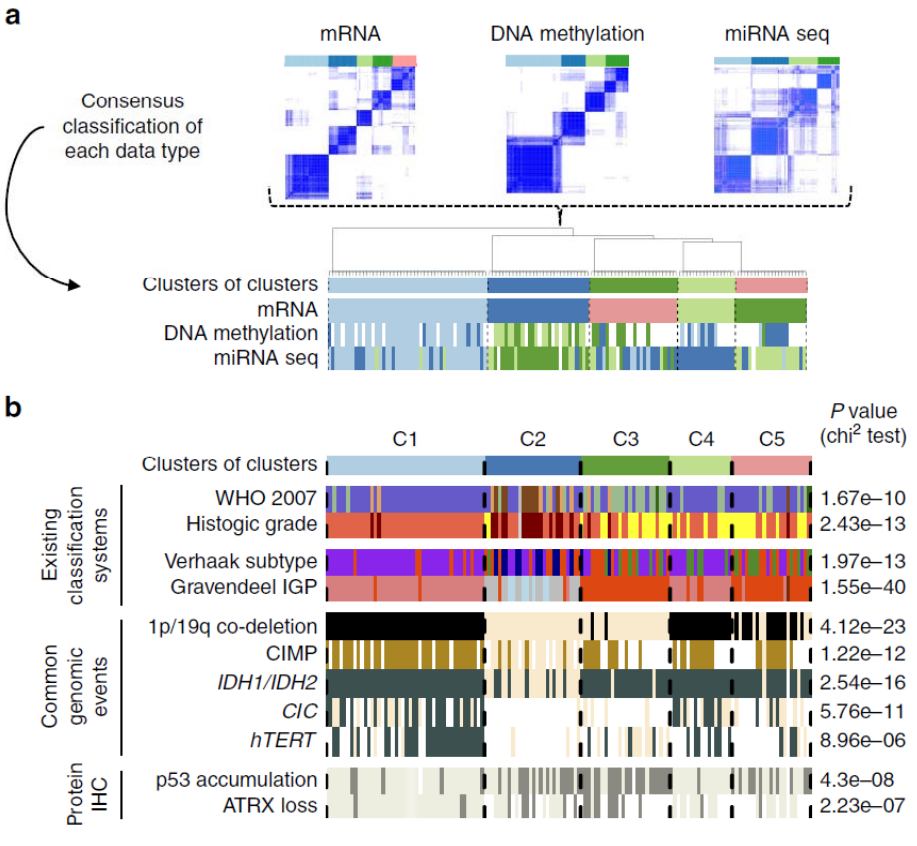




Multi-omics integration for Classification and Clustering



Wang et al. Nature Methods 2014



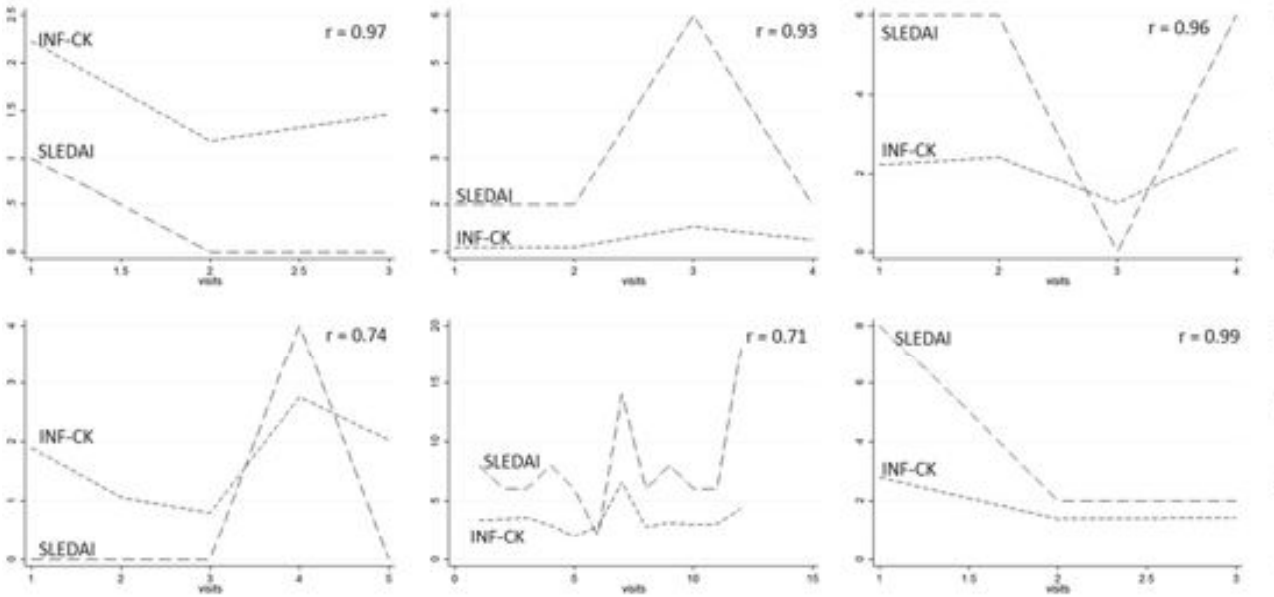
Kamoun et al. Nature Comm 2016

Cluster of clusters approach merges consensus clustering probability matrix to infer groups of sample with similar patterns across data types

Systemic lupus erythematosus

SLE is an autoimmune diseases are characterized by immune responses to self antigens that result in tissue damage

The disease course is unpredictable, with periods of remission and flares that lead to cumulative damage over time



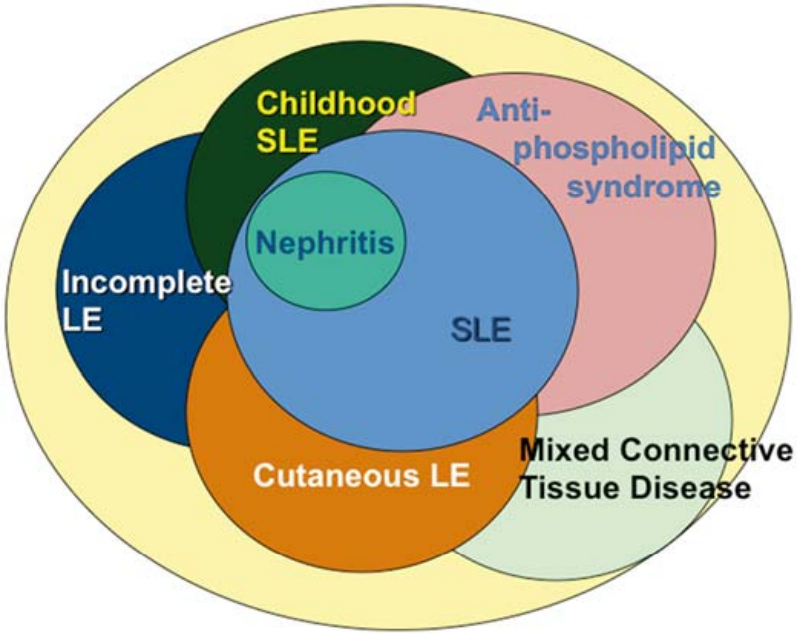
Connelly KL et al. Sci. Reports 2018

Systemic lupus erythematosus

Diagnosis is primarily clinical and remains challenging because of the heterogeneity of SLE

Only one drug (belimumab) has been approved for use in SLE in the past 60 years

Unlike in cancer, genetic traits do not seem to be generally useful as diagnostic biomarkers at present and there is an urgent need for better biomarkers and better treatments



Systemic lupus erythematosus

Published in final edited form as:
Cell. 2016 April 21; 165(3): 551–565. doi:10.1016/j.cell.2016.03.008.

Personalized Immunomonitoring Uncovers Molecular Networks That Stratify Lupus Patients

Romain Banchereau[1,7], **Seunghee Hong**[1,7], **Brandi Cantarel**[1], **Nicole Baldwin**[1], **Jeanine Baisch**[1], **Michelle Edens**[1], **Alma-Martina Cepika**[1], **Peter Acs**[1], **Jacob Turner**[1], **Esperanza Anguiano**[1], **Parvathi Vinod**[1], **Shaheen Kahn**[2], **Gerlinde Obermoser**[1], **Derek Blankenship**[1], **Edward Wakeland**[2], **Lorien Nassi**[2,3], **Alisa Gotte**[2,3,4], **Marilynn Punaro**[2,3], **Yong-Jun Liu**[1,5], **Jacques Banchereau**[6], **Jose Rossello-Urgell**[1], **Tracey Wright**[2,3], and **Virginia Pascual**[1,3,*]

[1]Baylor Institute for Immunology Research, Dallas, TX 75204, USA

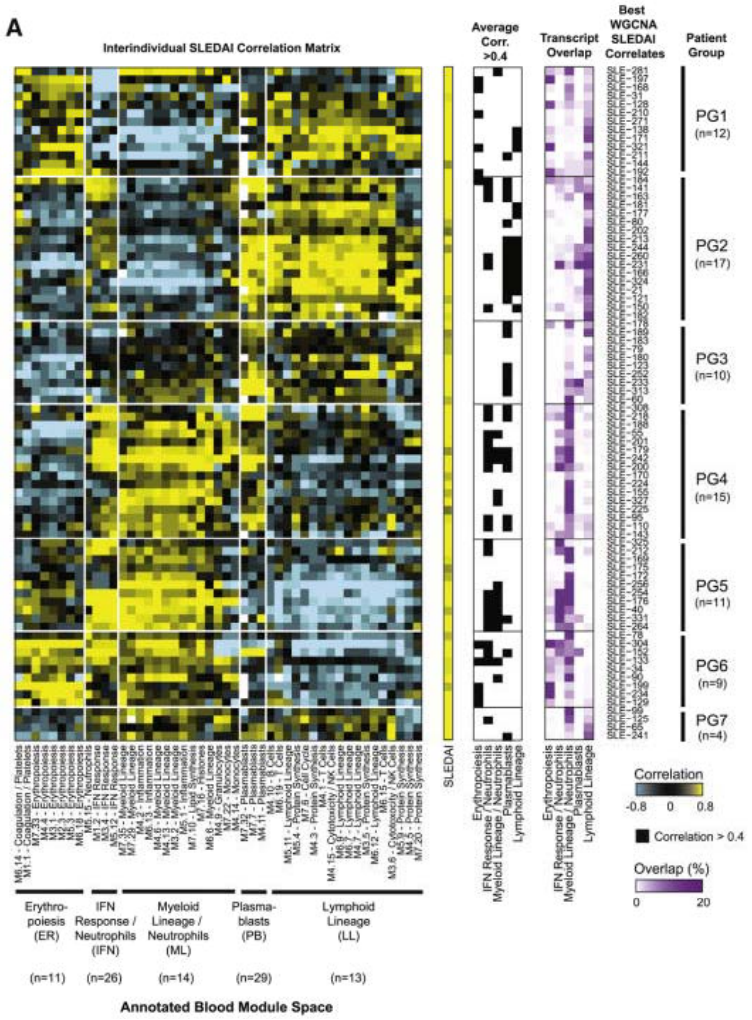
[2]UT Southwestern Medical Center, Dallas, TX 75235, USA

[3]Texas Scottish Rite Hospital for Children, Dallas, TX 75219, USA

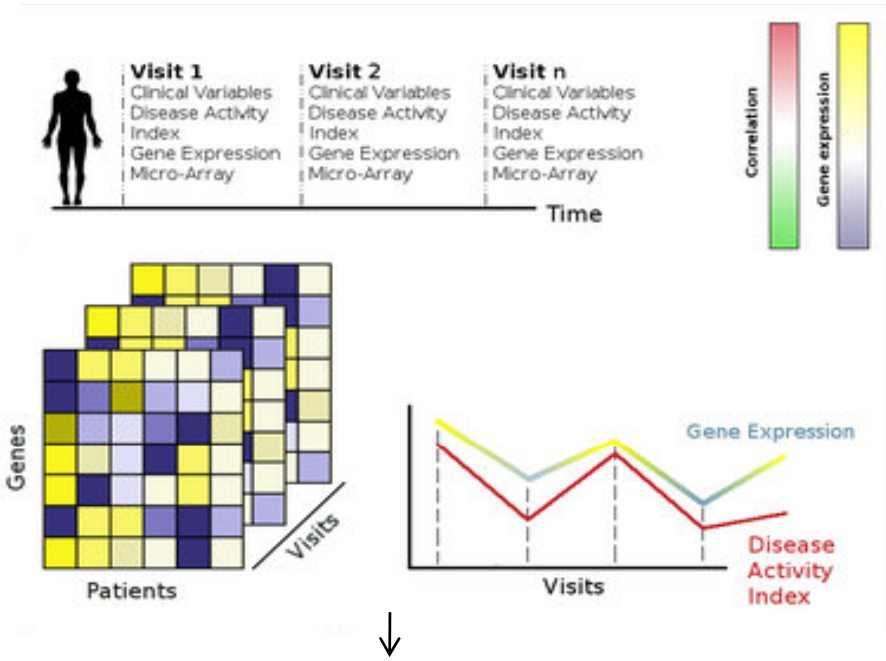
[4]Vanderbilt University School of Medicine, Nashville, TN 37232, USA

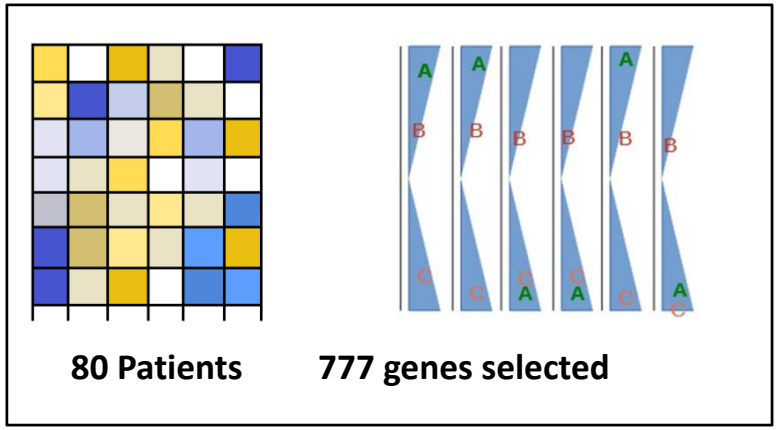
[5]MedImmune, Gaithersburg, MD 20878, USA

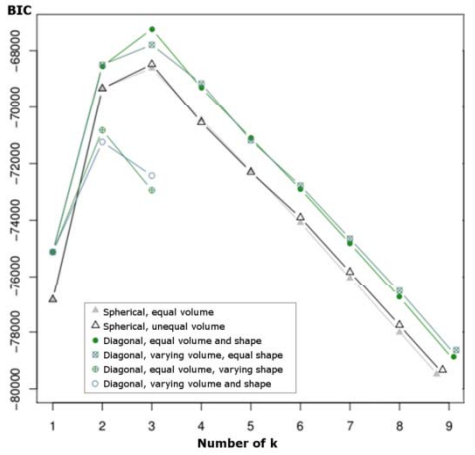
[6]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA

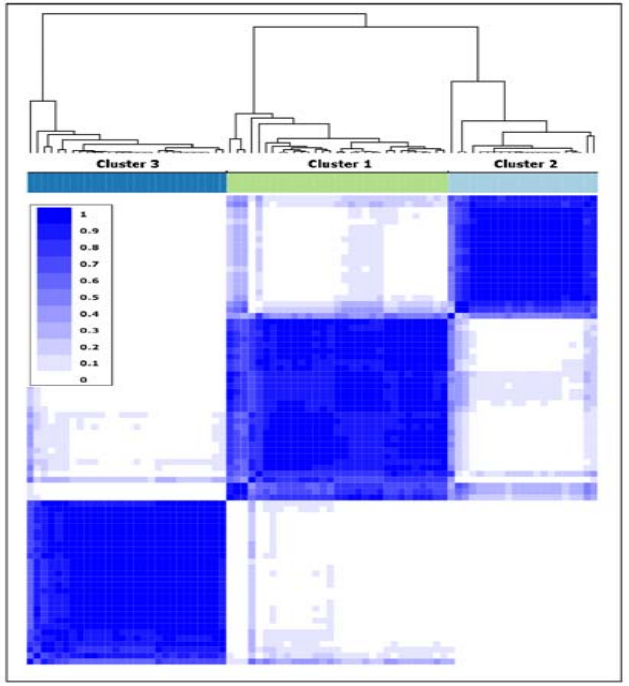


Systemic lupus erythematosus









Systemic lupus erythematosus

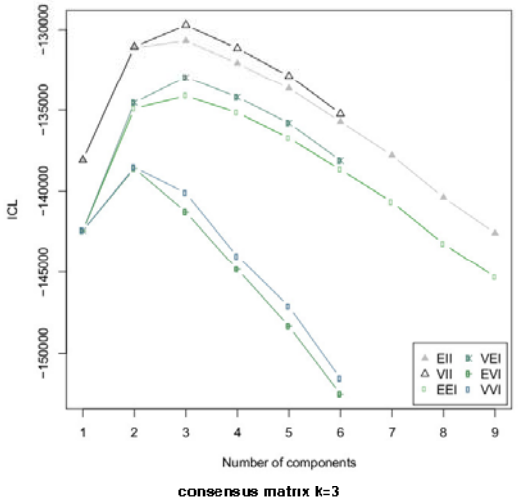
Two independent Datasets

Dataset 1. Pediatric Patients

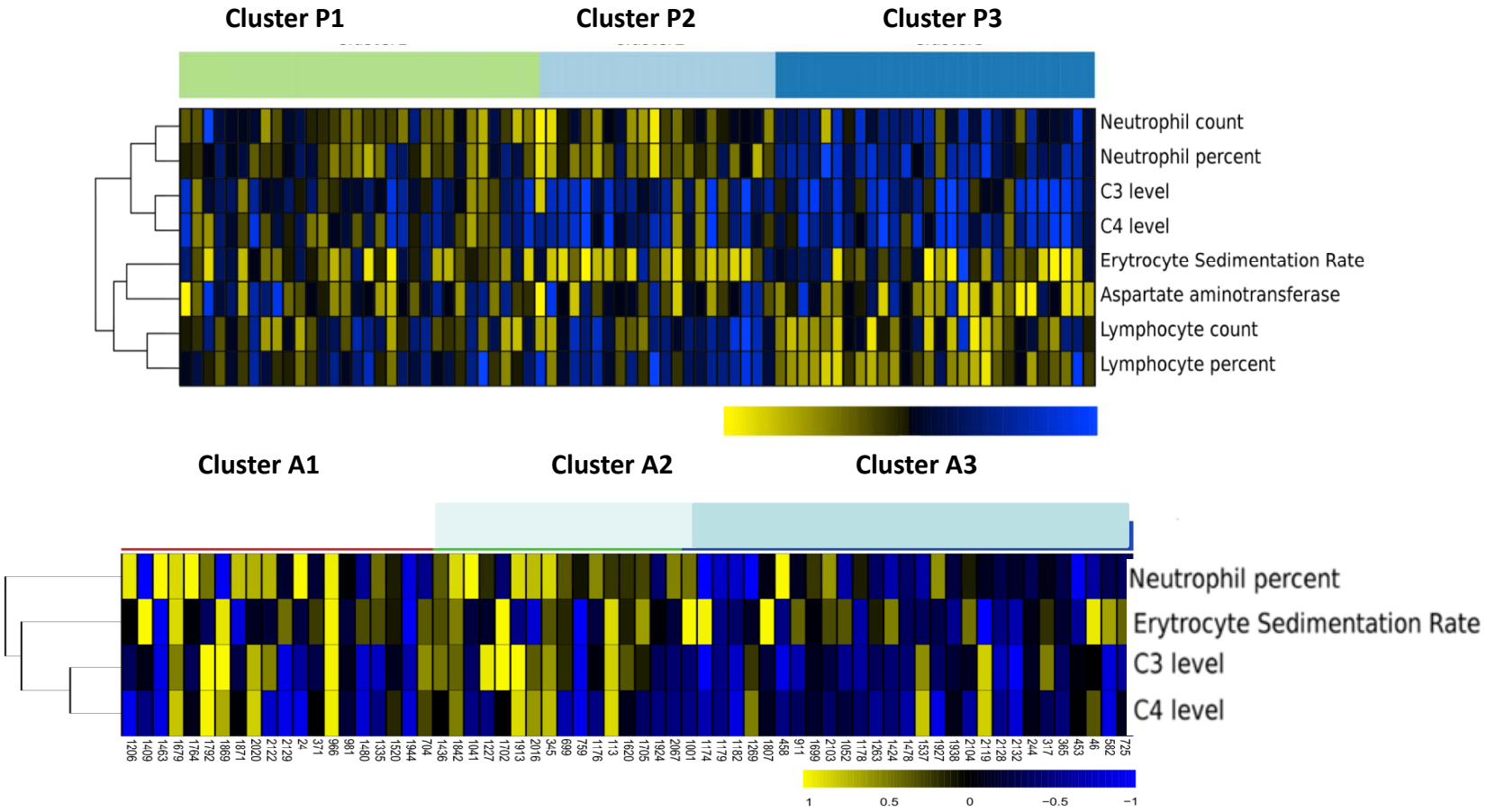
- **Banchereau et al. Cell. 2016**
- Longitudinal study **with 158** patients **and 46 Healthy** Samples
- **997 samples in total**
- Illumina GE Arrays

Dataset 2. Adult Patients

- Hopkins lupus cohort
- Longitudinal study with 301 patients **and 20 Healthy Sam**ples
- **747 samples in total**
- Affymetrix GE Arrays



Systemic lupus erythematosus



The percentage of neutrophils decreased with disease activity in cluster A3 and increased in clusters A2 and A1.

Toro-Domínguez D, et al. *Longitudinal Stratification of Gene Expression Reveals Three SLE Groups of Disease Activity Progression*. **Arthritis Rheumatol. 2018**

TF regulons

Most SLE loci occur in likely gene regulatory Regions (Maurano et al. Science 2012)

nature genetics

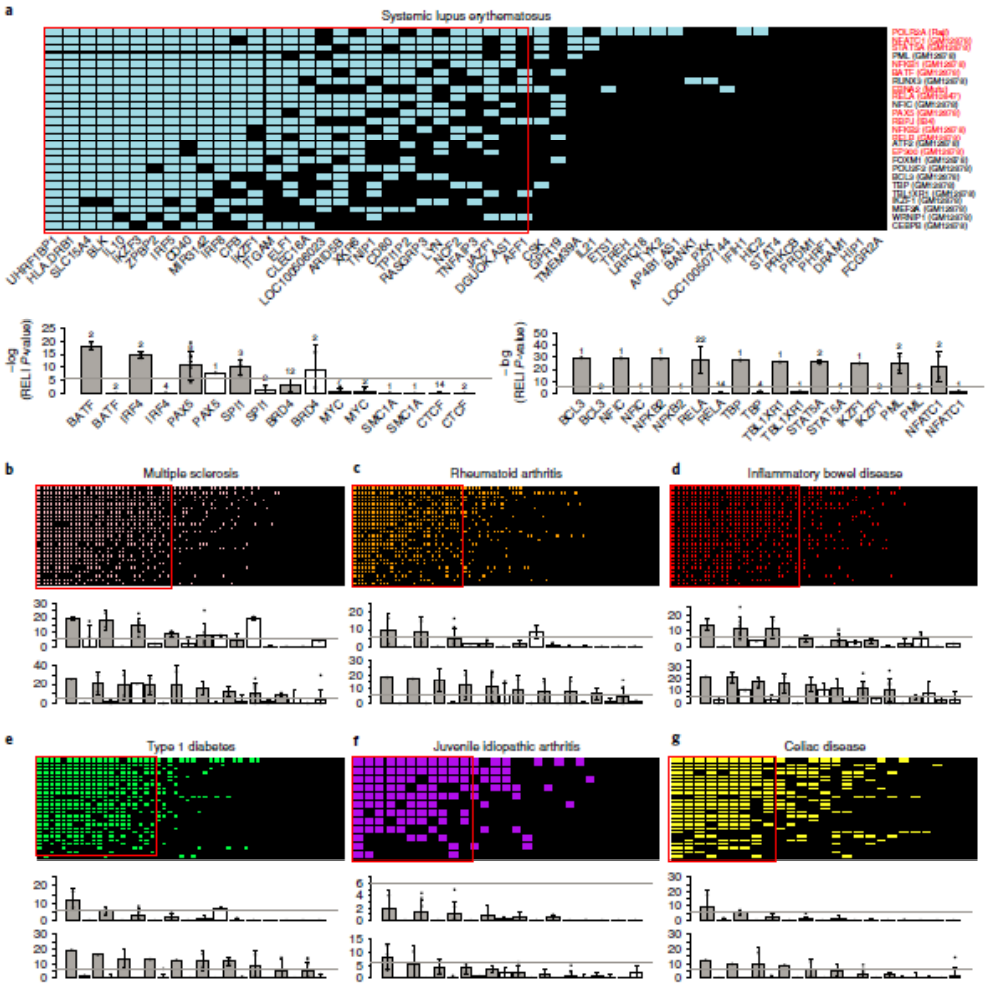
ARTICLES

https://doi.org/10.1038/s41588-018-0102-3

Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity

John B. Harley[1,2,3,4,5,9*], Xiaoting Chen[1,9], Mario Pujato[1,9], Daniel Miller[1], Avery Maddox[1], Carmy Forney[1], Albert F. Magnusen[1], Arthur Lynch[1], Kashish Chetal[6], Masashi Yukawa[7], Artem Barski[4,7,8], Nathan Salomonis[4,6], Kenneth M. Kaufman[1,2,4,5], Leah C. Kottyan[1,4*] and Matthew T. Weirauch[1,3,4,6*]

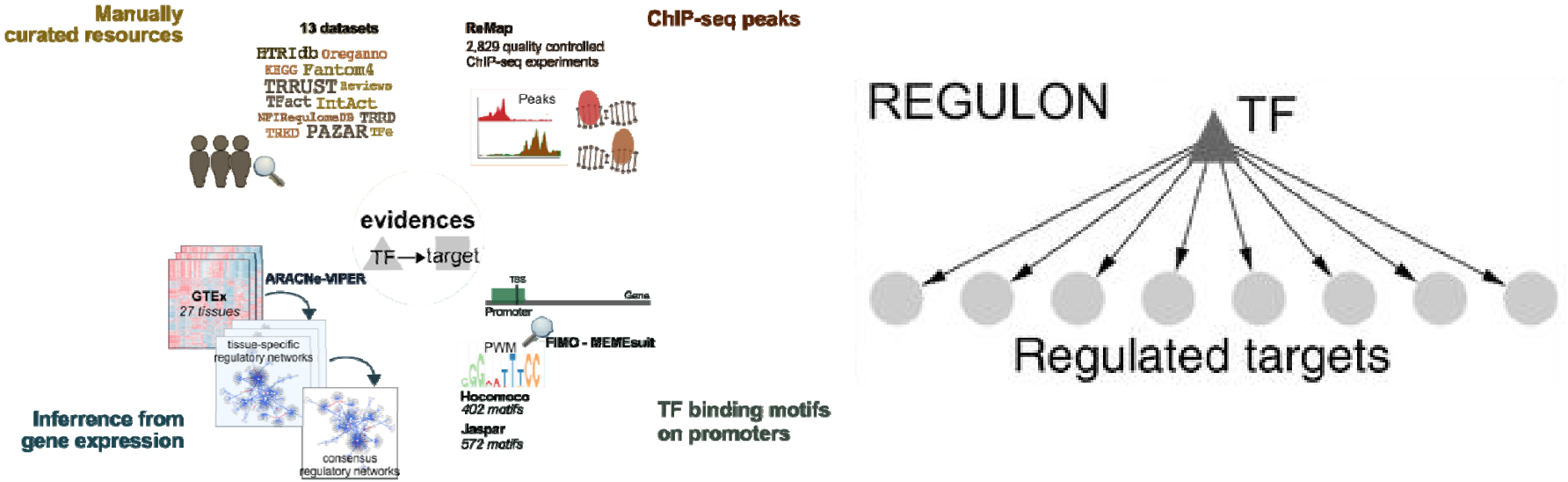
Harley et al. Nat Genet. 2018 May;50(5):699-707



TF regulons

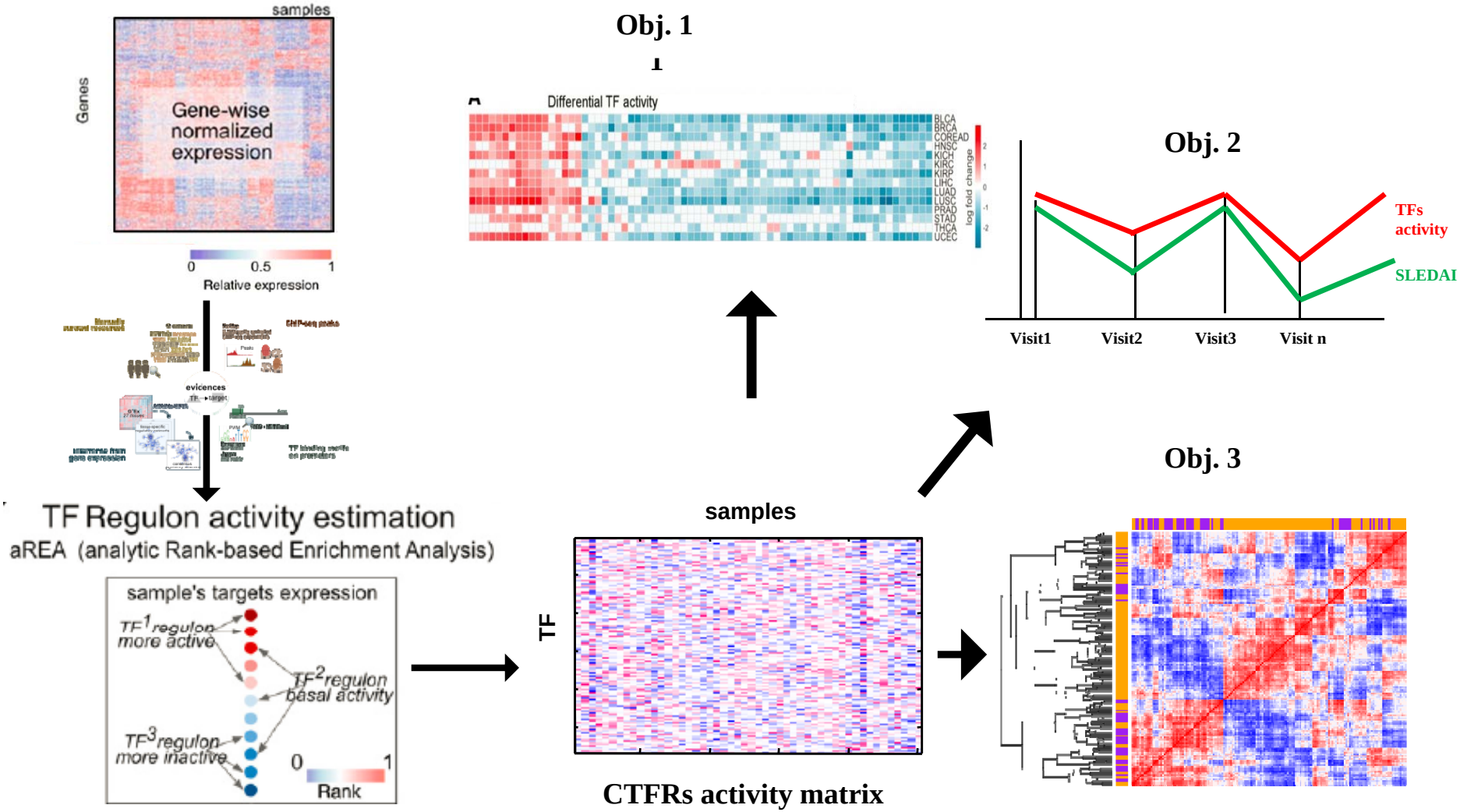
Activities of TF activities can be estimated computationally from the gene expression levels of their direct targets (the so-called TF regulon)

The assumption behind is that the level of protein activity of a TF is reflected on the transcript levels of its targeted genes



Garcia-Alonso et al

TF regulons



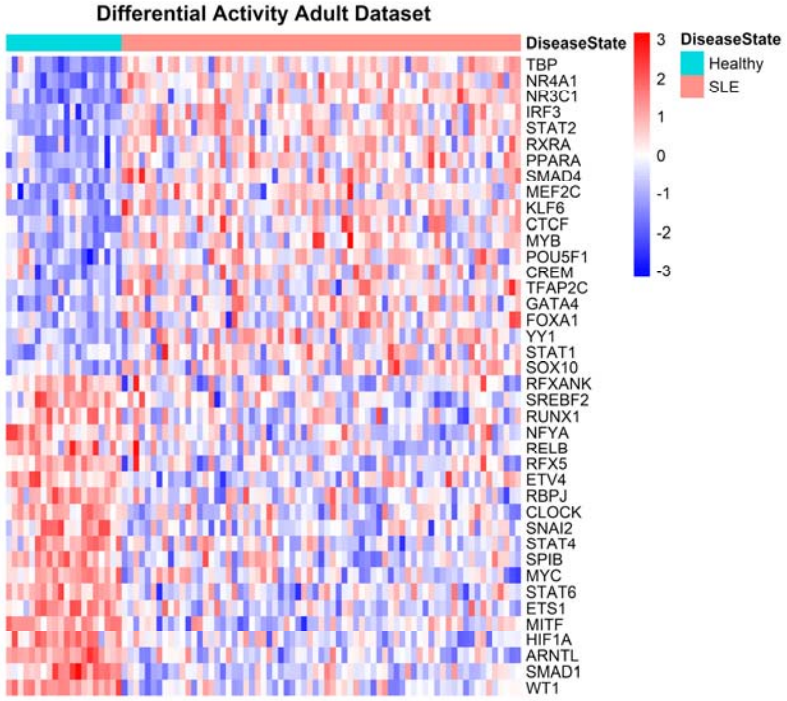
Adapted from Garcia-Alonso et al

Systemic lupus erythematosus

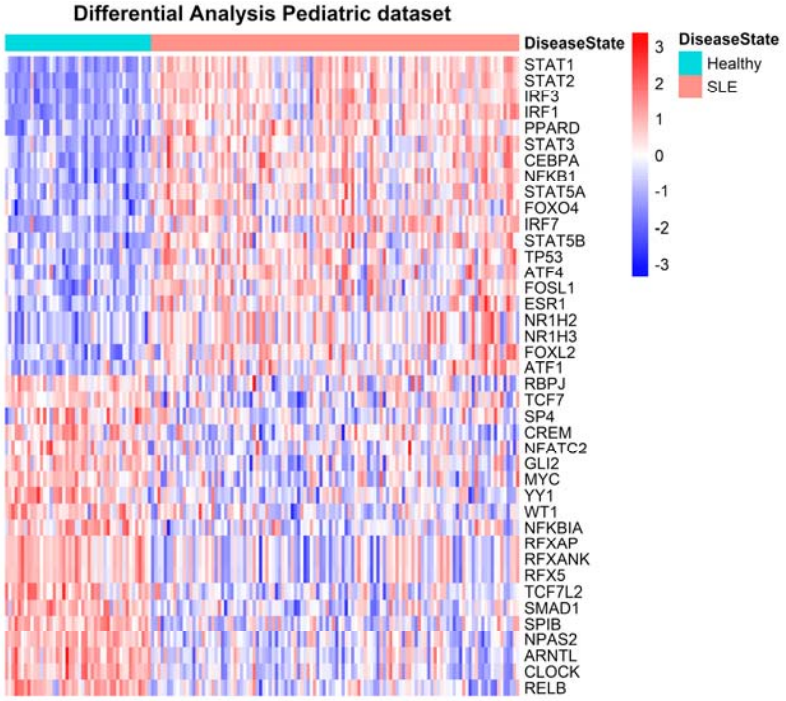
High Activity
Low Activity

Differential Activity Analysis among lupus and healthy samples where lupus samples had a high value of SLEDAI

adjPval < 0.05

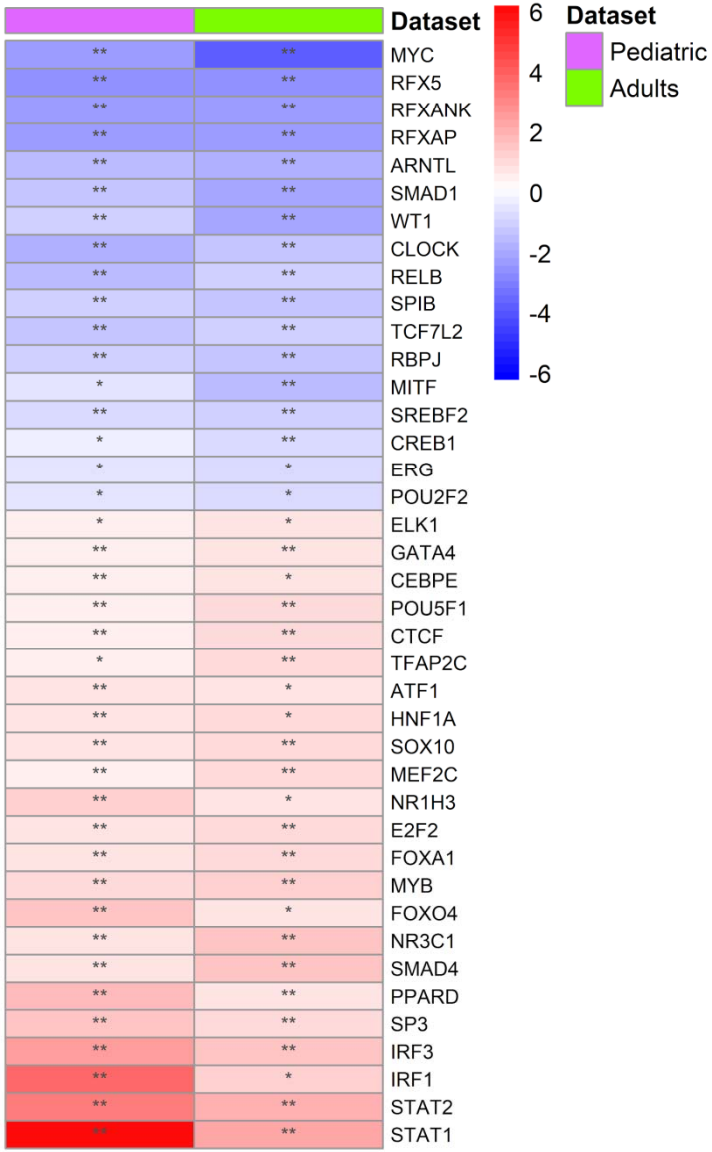


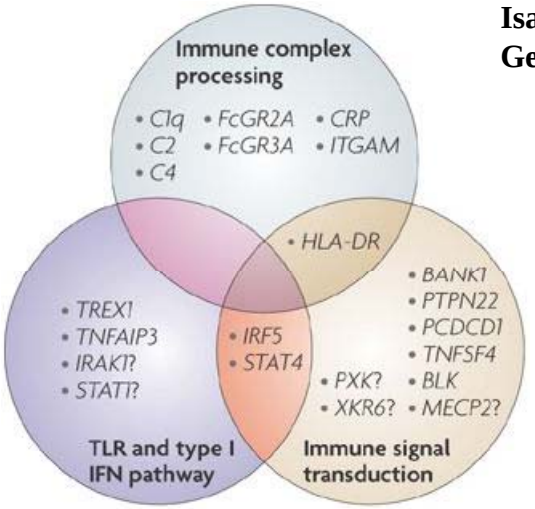
Up Activity	33
Down Activity	39



Up Activity	81
Down Activity	30

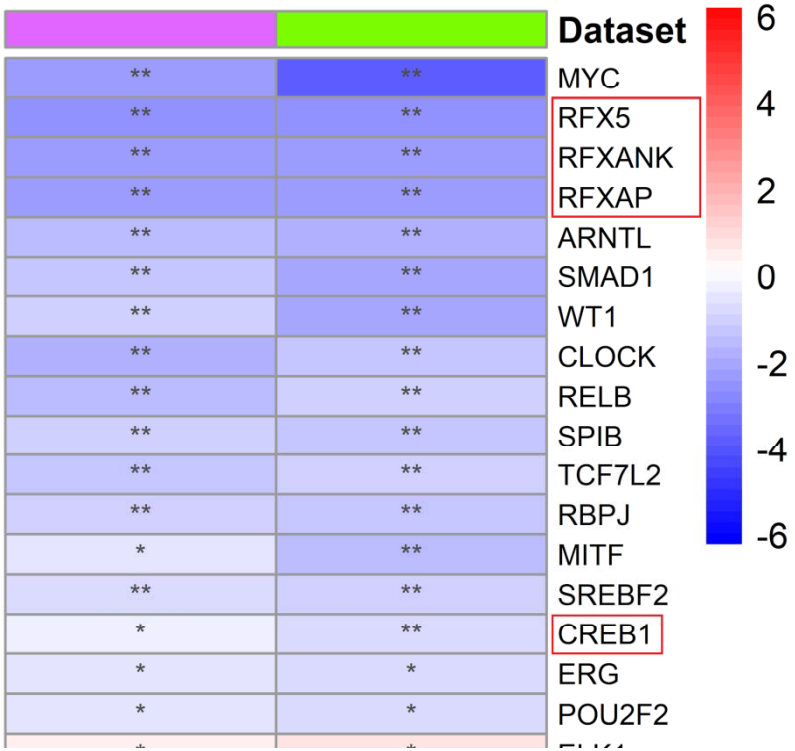
Systemic lupus erythematosus

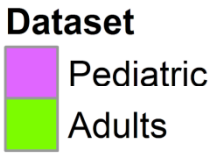




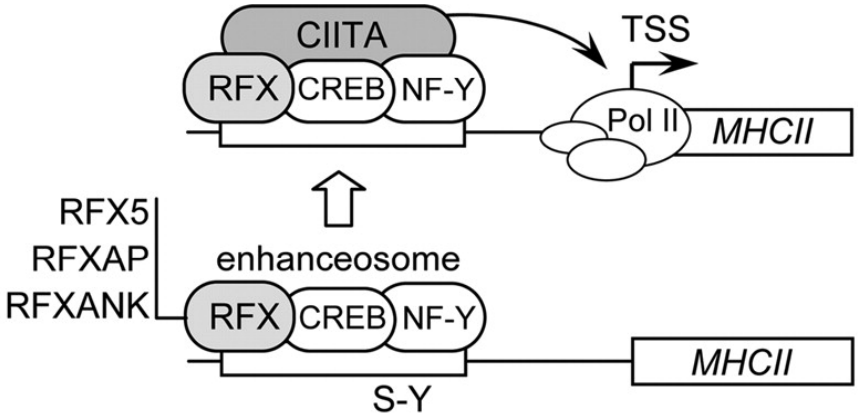
Isaac. Nature Reviews Genetics 2009

Systemic lupus erythematosus

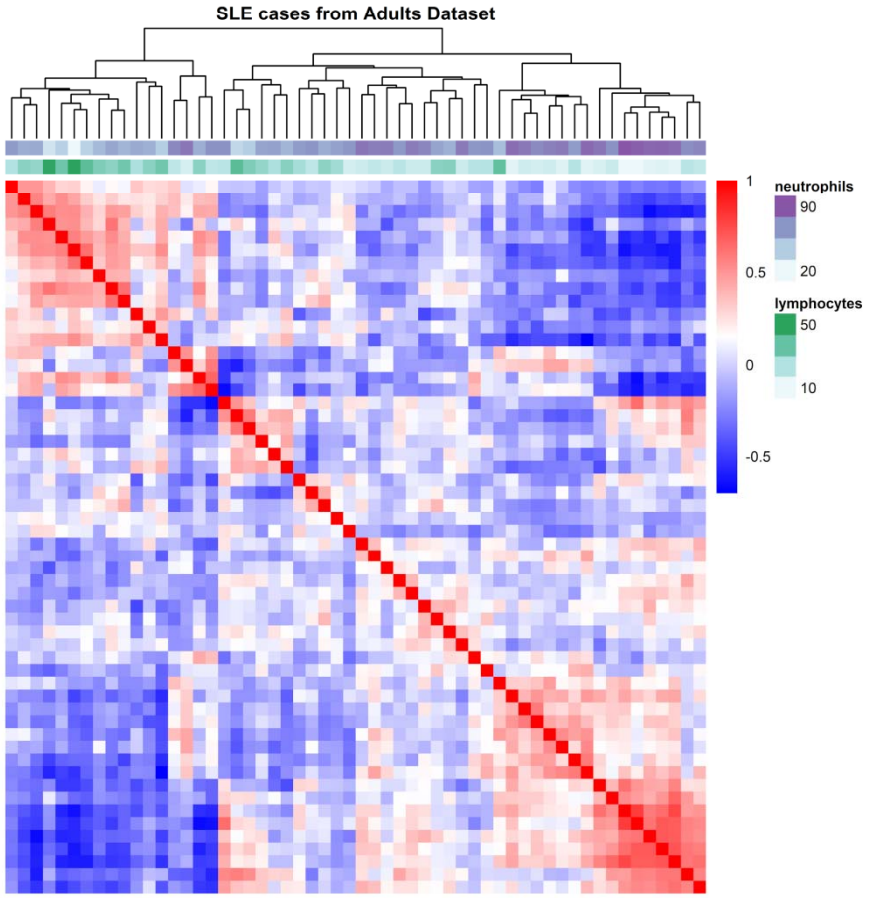


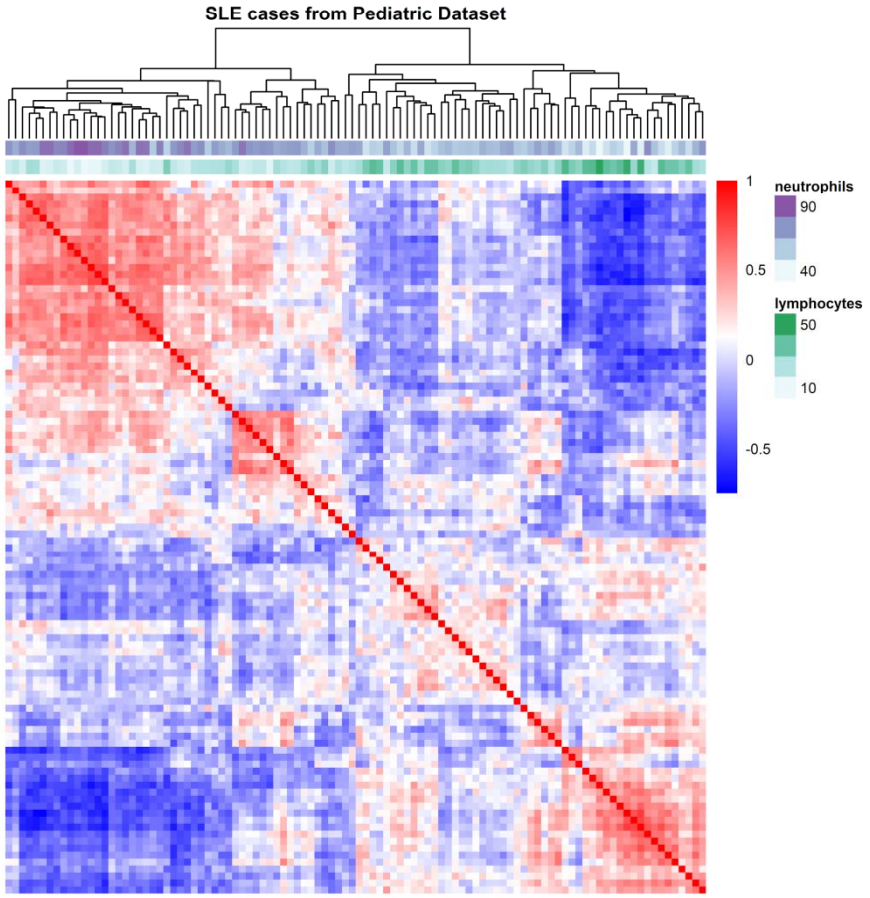


Seguín-Estévez et al., 2009

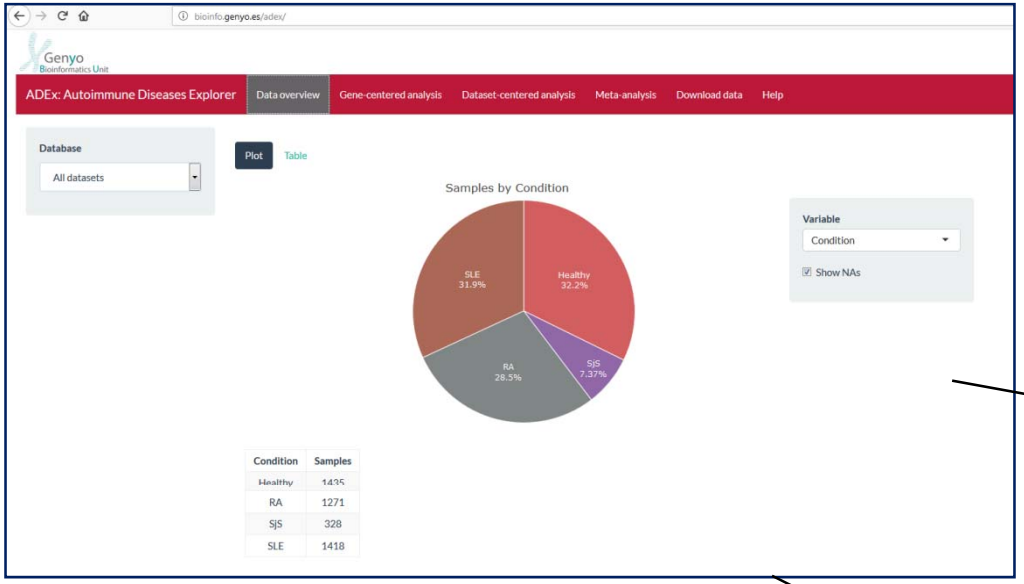


Systemic lupus erythematosus

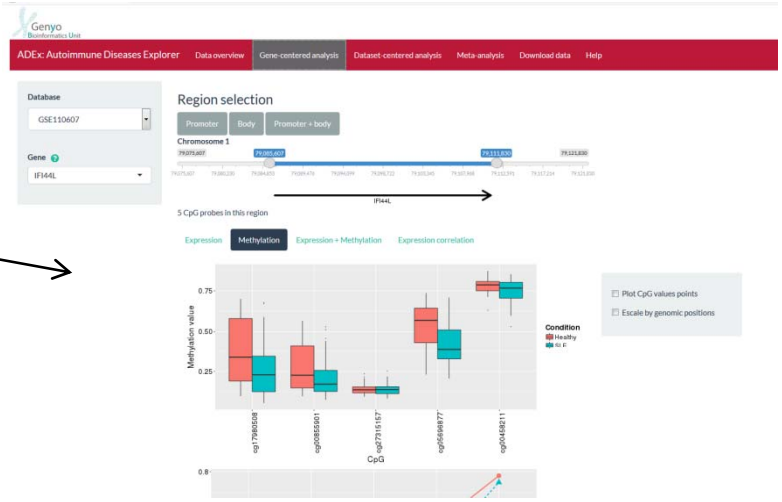


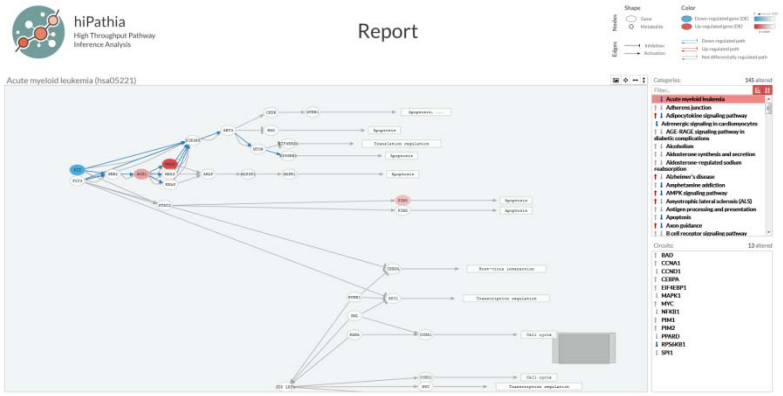


ADEx: Autoimmune Diseases Explorer

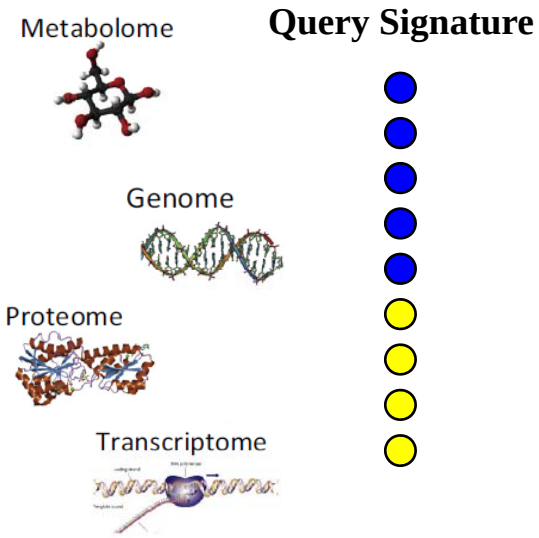


http://bioinfo.genyo.es/adex/





Integrating datasets: Connecting Phenotypes

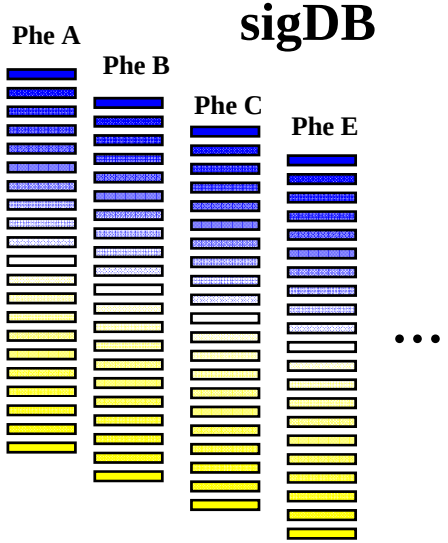


Query Signature

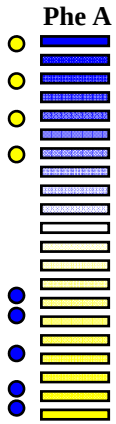
Finding conditions with similar trends markers

Common mechanisms,
Common Pathways
etc

sigDB

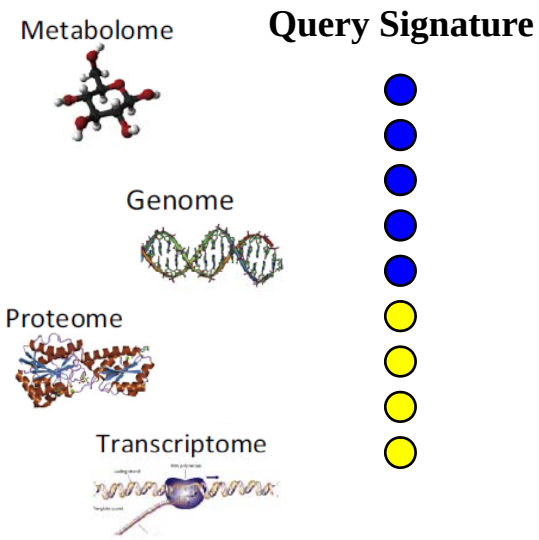


Finding conditions with inverse trends markers



Phenotype reversion,
Candidate drugs
etc

Integrating datasets: Connecting Phenotypes



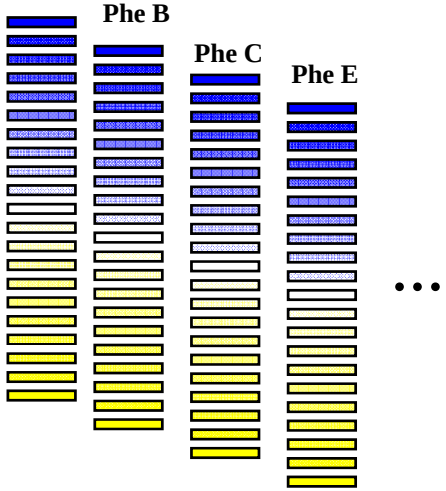
Query Signature

Phe C

Lamb et al. Science 2006

sigDB

Phe A **Phe B** **Phe C** **Phe E**



Over-expressed genes

Under-expressed genes

…

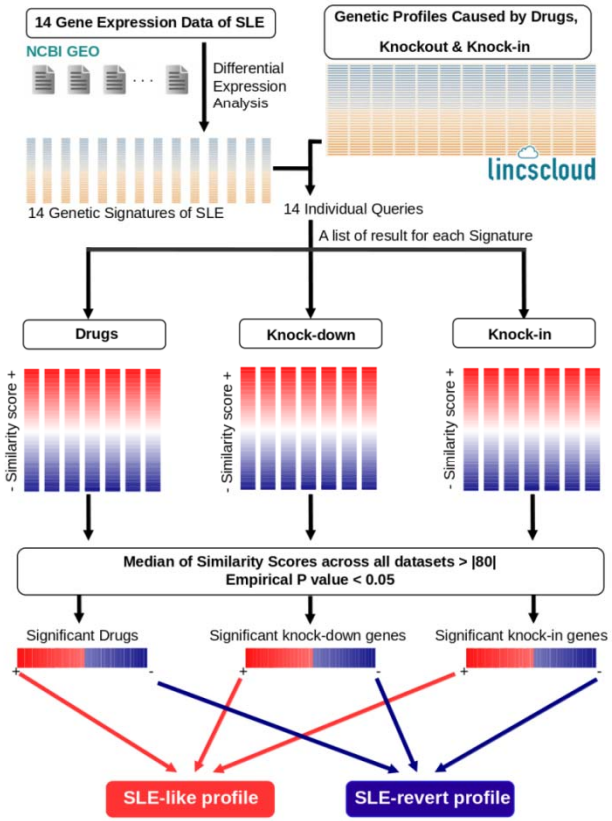
Phe A

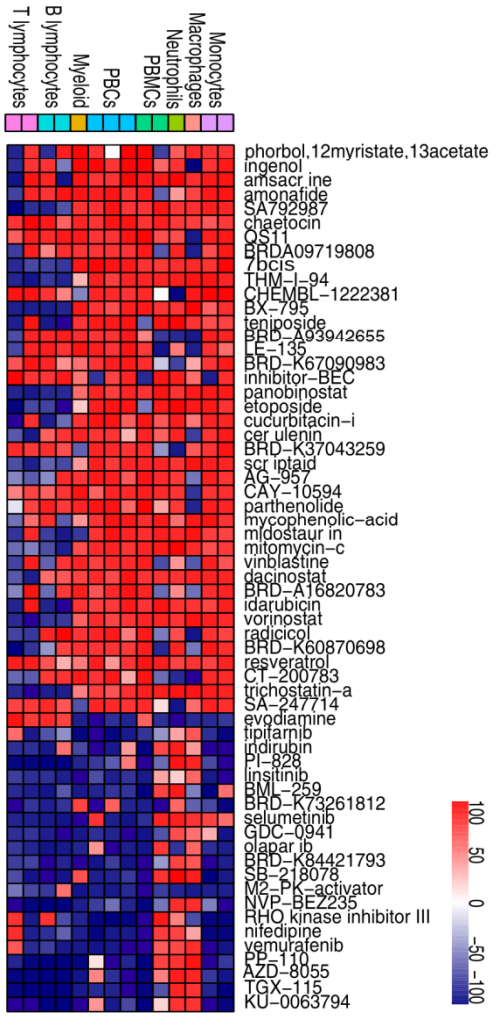
lincscloud

GENETIC REAGENTS	22,119	CHEMICAL REAGENTS	20,413
knock down	18,492	tool compounds	14,339
over expression	3,492	drugs and bioactives	5,585
variant	135	other	489

Drug repurposing in SLE

We queried Lincscloud to connect compounds and SLE signatures. We obtained a list of drugs, knock-in and knockout genes with significant similarity scores with respect to the SLE signatures





Toro-Domínguez et al. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. **Arthritis Res Ther. 2017**

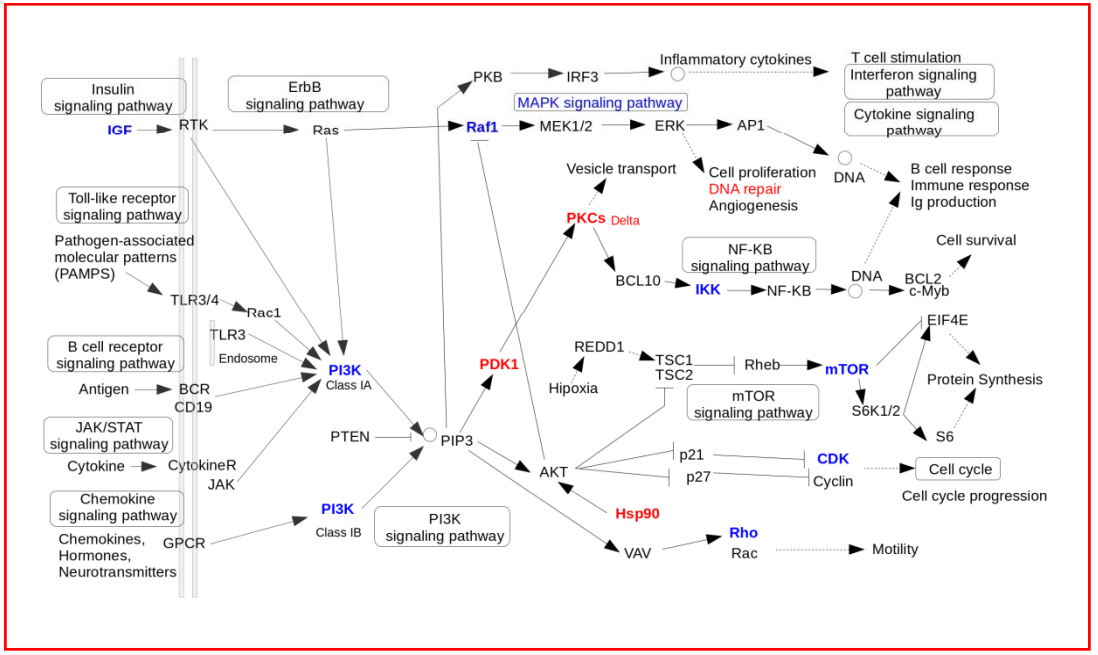
Drug repurposing in SLE

-	PI3K	Inhibitor	PI828 - GDC0941 - NVP-BEZ235 PP110 - TGX115	4.915E-06
-	mTOR	Inhibitor	NVP-BEZ235 - AZD8055 - TGX115 Ku0063794	1.792E-05
-	CDK	Inhibitor	BML259 - Indirubin	1.463E-02
-	IKBalfa	Inhibitor	Evodiamine	
-	Farnesyltransferase	Inhibitor	Tipifarnib	
-	IGF1R	Inhibitor	Linsitinib	
-	MAP2K1	Inhibitor	Selumetinib	
-	CHK1	Inhibitor	SB218078	
-	Piruvate kinase	Inhibitor	M2PK Activator	
-	Rho kinase	Inhibitor	Rho kinase inhibitor III	
-	Voltage dependent calcium channel	Inhibitor	Nifedipine	
-	Braf	Inhibitor	Vemurafenib	

Table 3: Significant knock-down and knock-in genes obtained.

Score	Type of experiment	Genes
+	Knock-in	*Ifinb1 - Ifng - Cd40 - Bcl10 - Klf6 - Lyn - Tyrap*
+	Knock-down	*Clcn3 - Ppp1r14b - Lmnb2 - Tbx2 - Pmm2 - Myc - Atp6v1f - Max - Pepd - Puf60 - Phb2 - Akr1a1 - Btg1 - Abhd2 - Tfdp1 - Pax8 - Fosl2 - Nt5e - Rrm1 - Nr2f6 - Ramp1 - Ryk - Cish - Ppp2r1a - Cd14 - Ufd1l - Htra1 - Slc35a1 - Twf2 - Nnt - Homer2 - Hs2st1 - Znf768 - Ggt1 - Dffb -Hspa2 - Prkdc - Arpc5 - Nfkbia - Slc39a8 - Thap11 - Gstp1 - Etv1 - Gcat - Kiaa0907 - Dlx3 - Elk1 - Pias4 - Meox2 - Gper - Nras - Tceb3c - Kif2c - Polr2f - Ctbp2 - Chaf1b - Cep55 - Hook2 - Znf8 - Ndufb7 - Nisch - Hoxc10 - Aqp12a - Yes1 - Psmd5 - Jag1 - Mdh2 - Polr2i - Ddf1 - Hras - Hdac10 - Slc25a14 - Med7 - Hmgcr - Pdxp - Fdx1 -Nipbl - Prkag3 - Ppia - Eif2ak3 - B4galt1 - Uck2 - Jun - Med4 - Ybx1 - Bub1b - Crcp - Med1 - Hdac11 - Sbno1*
-	Knock-down	*Mitf - Etfa - Pip4k2b - Vrk2 - Spen - Nsdhl - Znf586 - Gmpda1 - Six4 - Parn - Dusp14 - Iqgap1 - Lrrk2 - Gpr123 - Sf1 - Fez2 - Ipmk - Sat1 - Elf4 - Rptor - Eif4e - Arl3 - Kars - Csnk1a1 - Sptlc2 - Men1 - Snx17 - Vegfc - Ppp3ca - Bnip3 - Erbb3 - Ero1l - Copb2 - Serpinc1 - Ak4 - Hla_a - Pik3ca - Pik3c2a - Igf2r - Lypla1 - Stx4 - Atm - Espl1 - Igf1r - St3gal5 - Mtor - Grn - Hsp90aa1 - Prpf4b - Tm9sf3*

PI3K molecular signaling pathway



Toro-Domínguez et al. Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. **Arthritis Res Ther. 2017**

PLOS | MEDICINE

Browse Publish About Search

advanced search

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

68,836 Save | 3,556 Citation | 2,605,883 View | 10,479 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

Check for updates

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Related PLOS Articles

Why Current Publication Practices May Distort Science

Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland

Why Most Published Research Findings Are False: Problems in the Analysis

Most Published Research

John P. A. Ioannidis

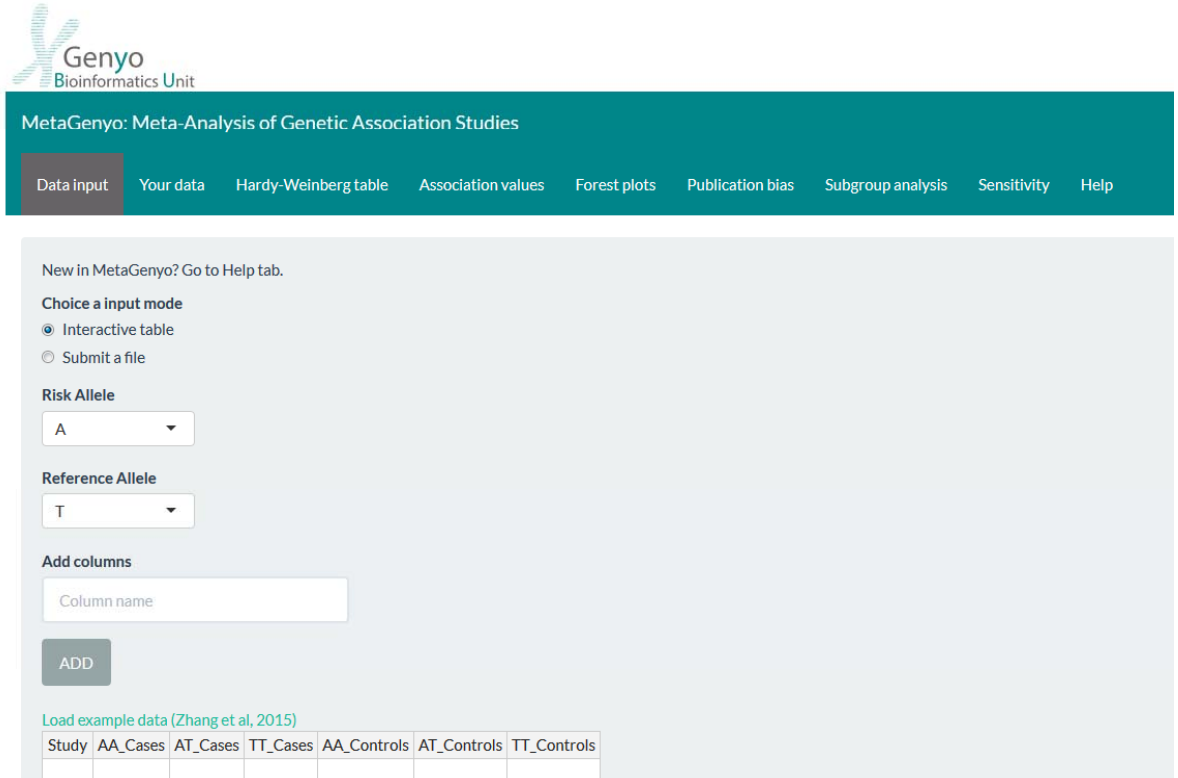
Professor of Medicine, Stanford University (previously at U Ioannina, Greece)
Verified email at stanford.edu - Homepage

Evidence-based medicine research methods meta-analysis clinical epidemiology genetic epidemiology

Cited by		VIEW ALL
	All	Since 2013
Citations	181686	132586
h-index	173	131
i10-index	794	674

MetaGenyo: Meta-Analysis of Genetic Association Studies

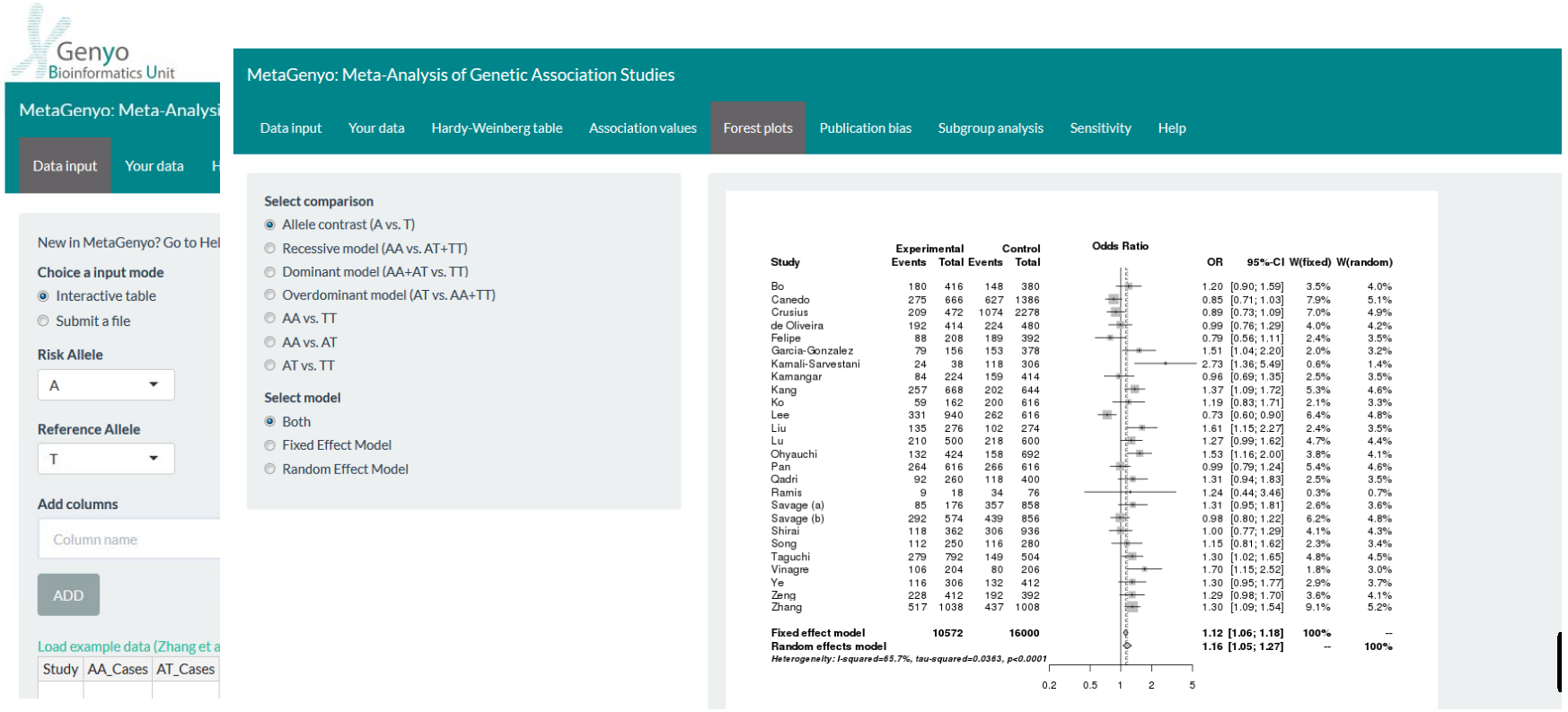
http://bioinfo.genyo.es/metagenyo/



Martorell-Marugan J, Toro-Dominguez D, Alarcon-Riquelme ME, Carmona-Saez P. *MetaGenyo: A web tool for meta-analysis of genetic association studies*. **BMC Bioinformatics. 2017**

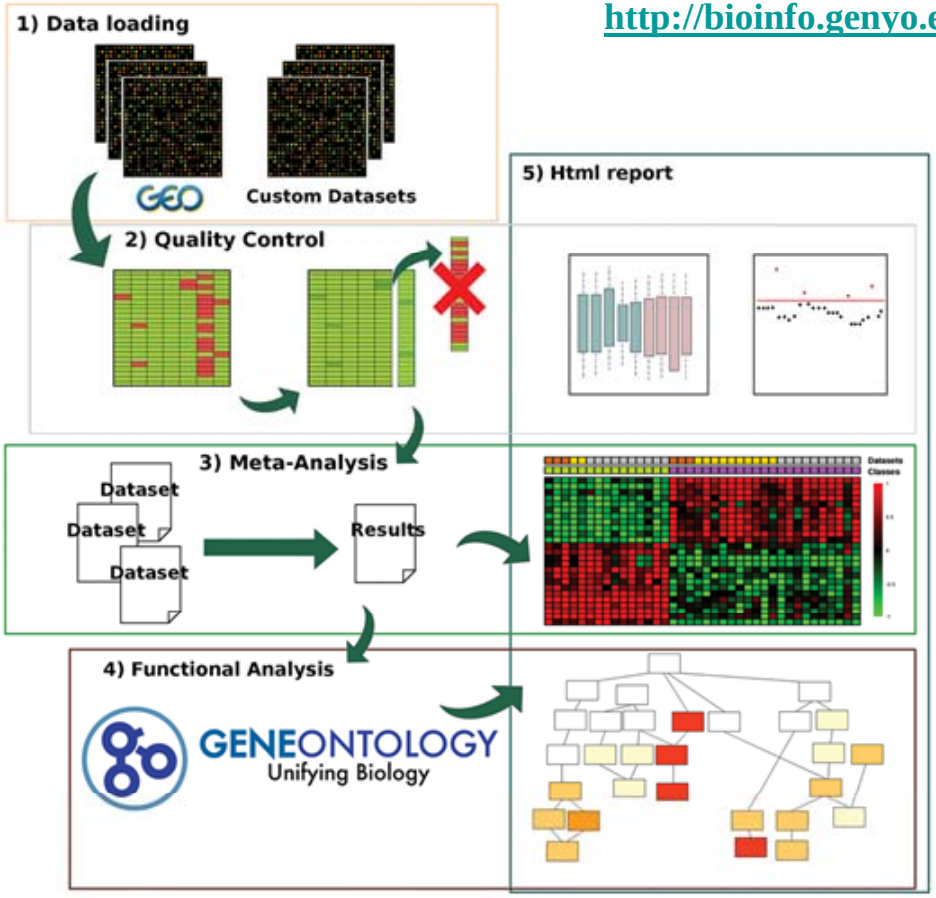
MetaGenyo: Meta-Analysis of Genetic Association Studies

http://bioinfo.genyo.es/metagenyo/



Martorell-Marugan J, Toro-Dominguez D, Alarcon-Riquelme ME, Carmona-Saez P. *MetaGenyo: A web tool for meta-analysis of genetic association studies*. **BMC Bioinformatics. 2017**

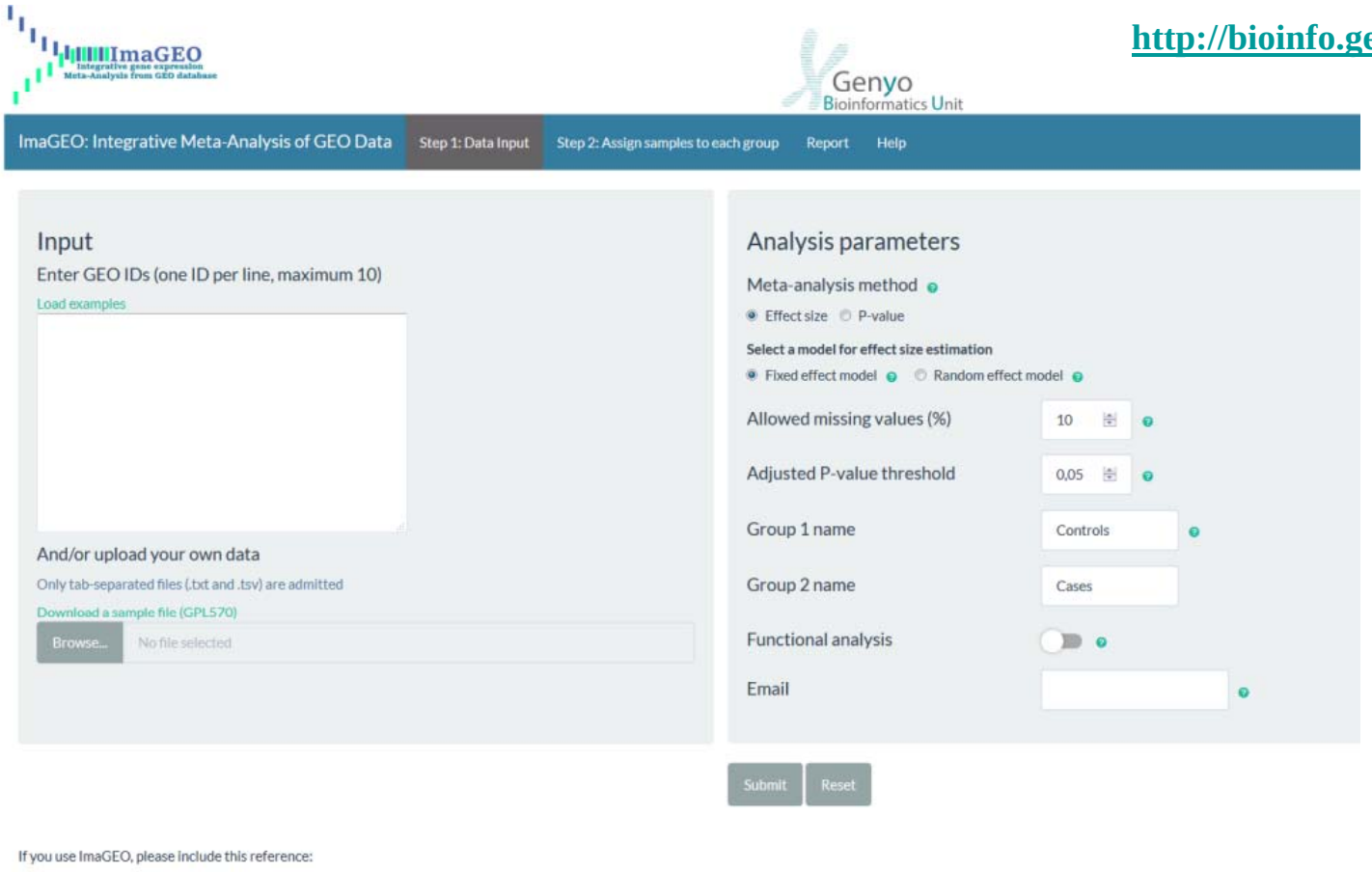
IMAGEO : Integrative Meta-Analysis from GEO Data



http://bioinfo.genyo.es/imageo/

D. Toro-Domínguez, J. Martorell-Marugán, R. López-Dominguez, A. García-Moreno, V. González-Rumayor, M. E Alarcón-Riquelme, P. Carmona-Sáez. *ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database*. **Bioinformatics 2018**

IMAGEO: Integrative Meta-Analysis from GEO Data

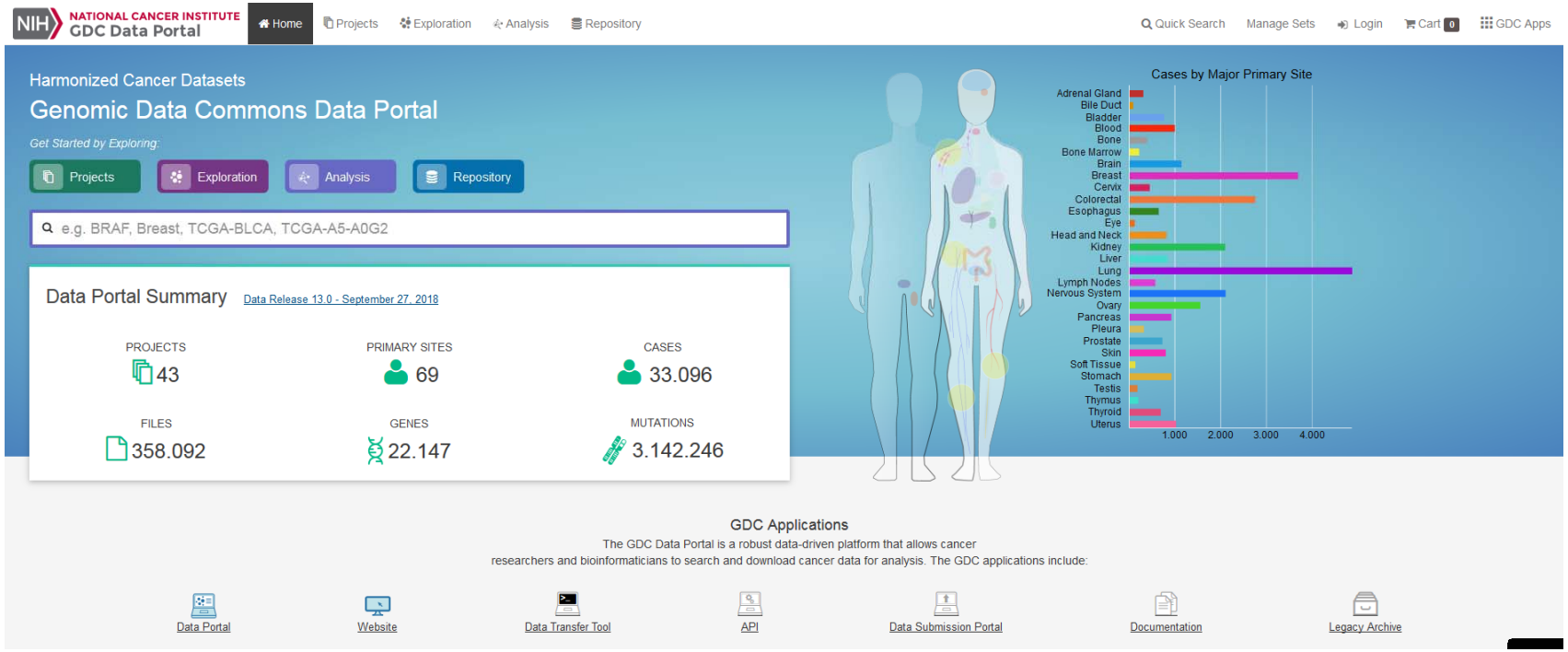


http://bioinfo.genyo.es/imageo/

D. Toro-Domínguez, J. Martorell-Marugán, R. López-Dominguez, A. García-Moreno, V. González-Rumayor, M. E Alarcón-Riquelme, P. Carmona-Sáez. *ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database*. **Bioinformatics 2018**

WE NEED LONG TERM INITIATIVES

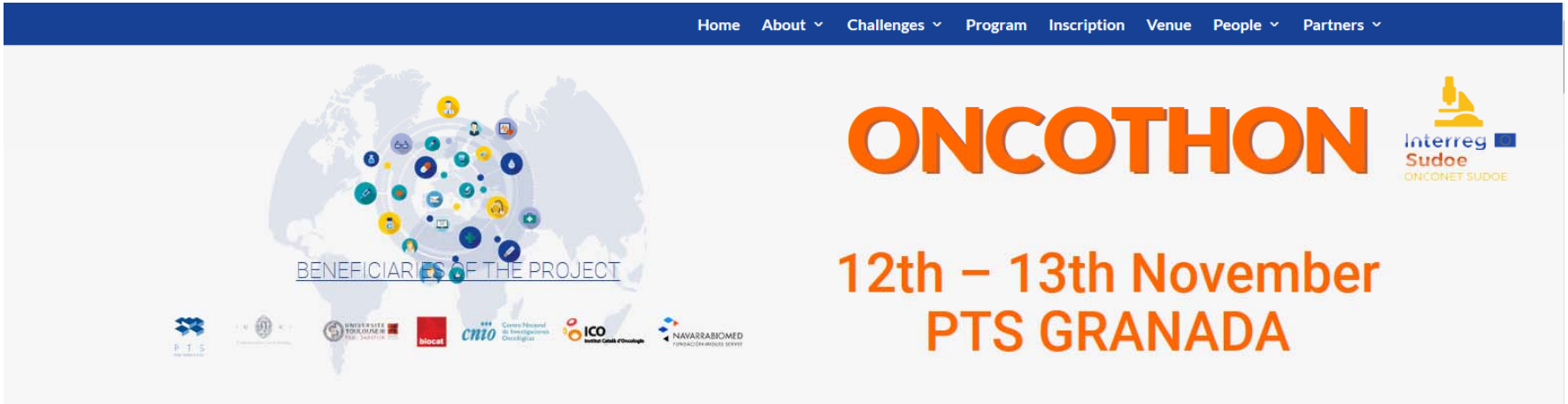
SUSTAINABLE RESEARCH



Bioinformatics Week in Granada

http://jbi2018.ugr.es/

http://oncothon.ptsgranada.com/



Follow us in twitter:

@pcarmonasaez

@BioinfoGenyo

Dr. Pedro Carmona

Jordi Martorell

Daniel Toro

Raúl López

Adrian García

Dr. Marta Alarcon. GENyO

Dr. Julio Saez-Rodriguez. U. of Heidelberg

Dr. Michelle Petri. Johns Hopkins Lupus Center

Dr. Joaquín Dopazo. Bioinformatics Area. FPS

The problem: Characterization of Gen2.2

Denditric Cells are antigen presenting cells with a key role in the immune response

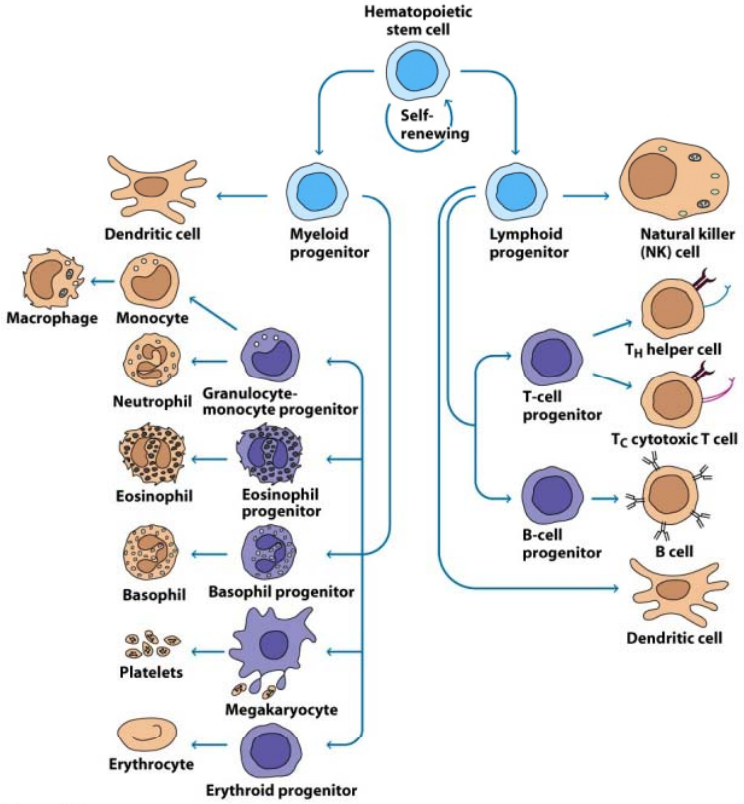


Figure 2-2
Kuby IMMUNOLOGY, Sixth Edition

Various DC precursors, such as DC1 (myeloid origin) or plasmacytoid dendritic cells (**pDCs**).

pDC are difficult to isolate (less than 0.5% of the circulating cells). Therefore, good cells models are required

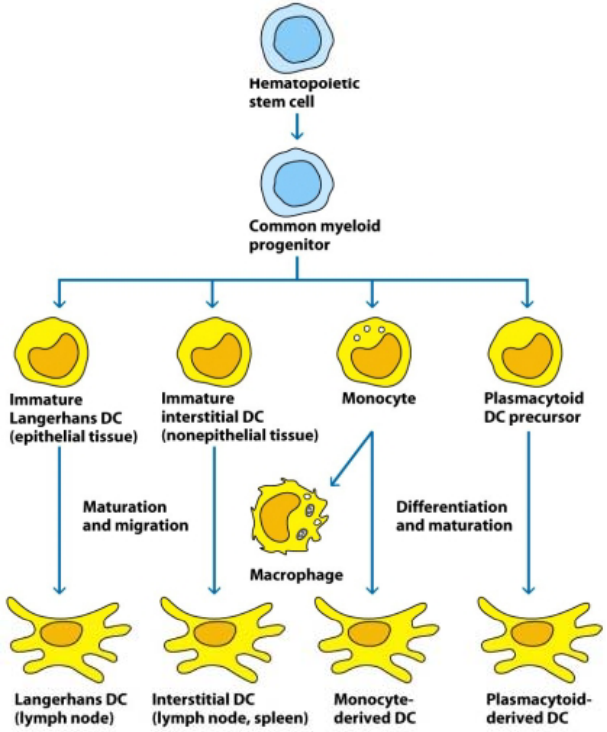
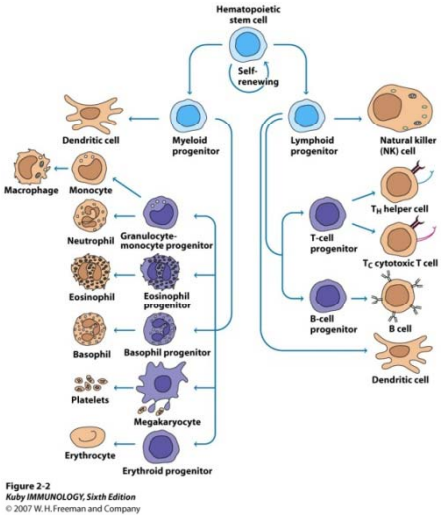


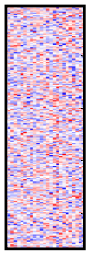
Figure 2-10
Kuby IMMUNOLOGY, Sixth Edition

Connecting Phenotypes: Similar profiles

Compare gene expression profiles of Gen2.2 with previously published immune system gene expression datasets to get insights into similarities among cell lines



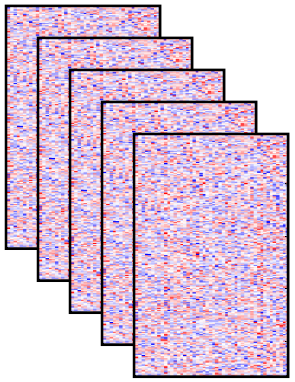
Gene 2.2 expression profile
Three replicates Illumina HT-V4

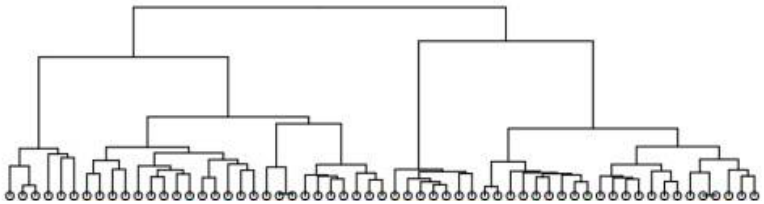


Gene expression compendium of Immune cell popluations

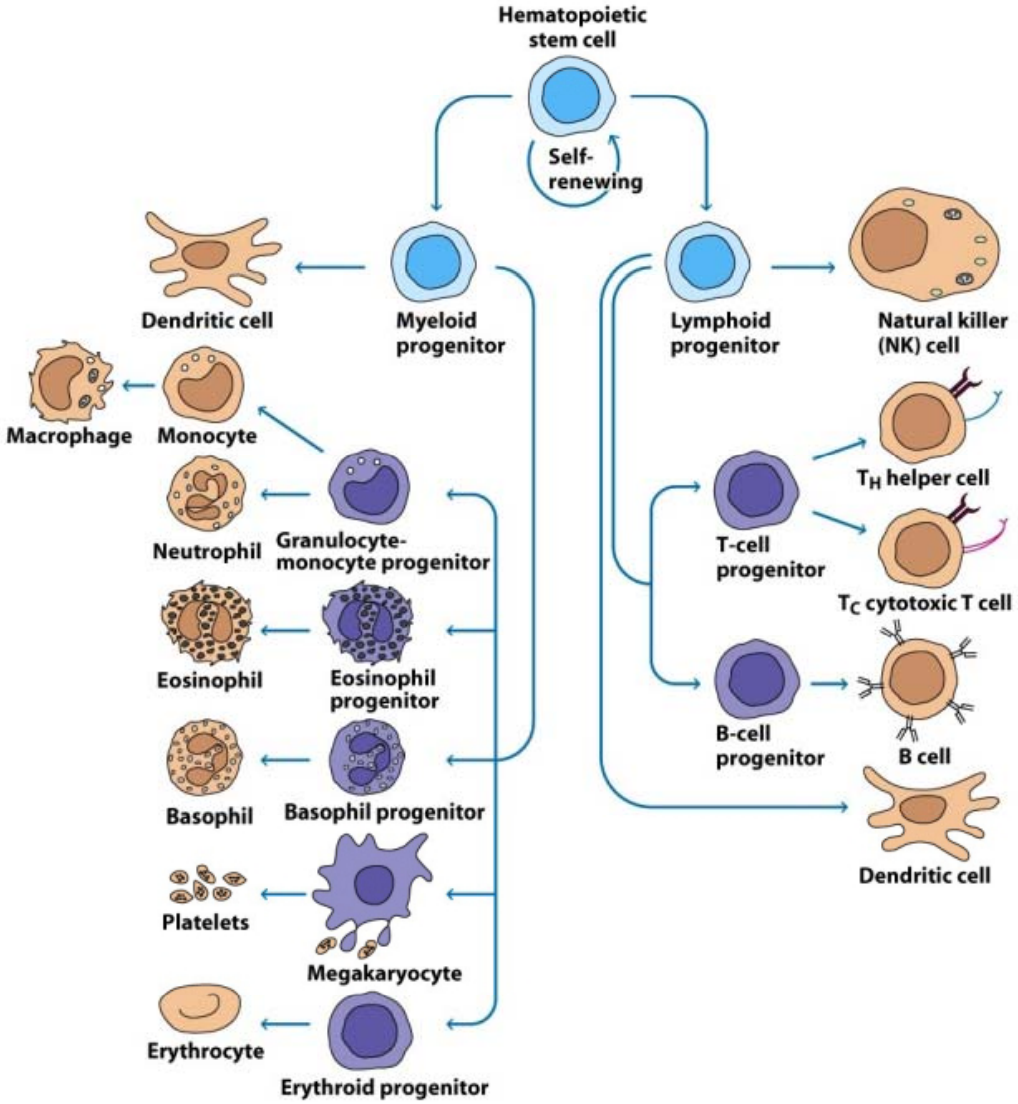
Rank-based Normalization
Join Datasets by Common Genes
Combined Analysis

GSE12507
GSE15215
GSE28490
GSE28491
GSE35457



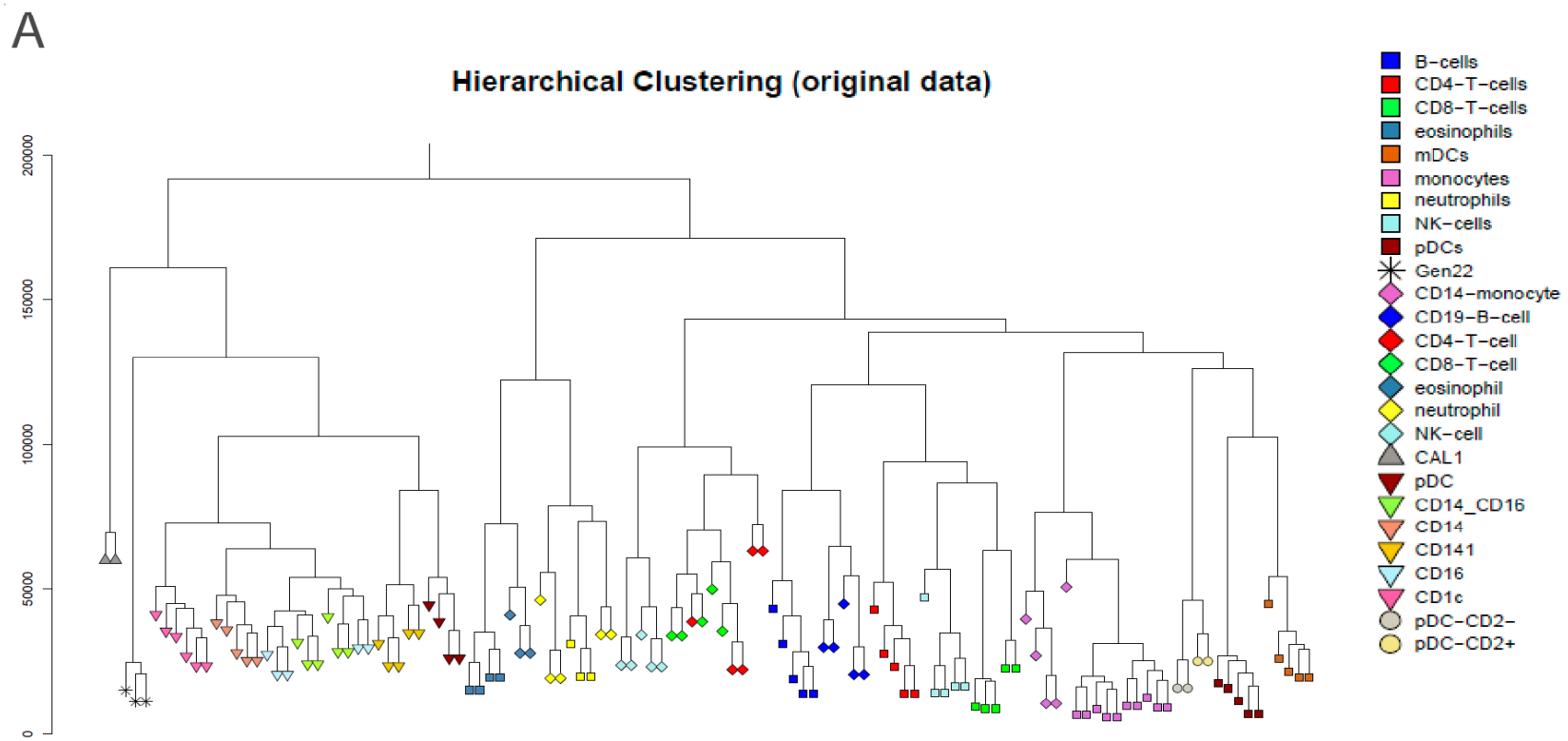


Connecting Phenotypes: Similar profiles



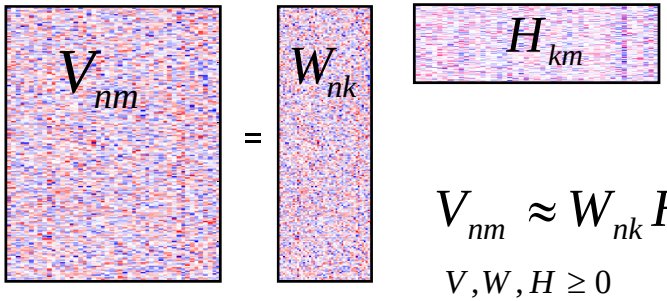
GEO Datasets	Cell Types
GSE28490	Monocytes
GSE28491	B cells
	CD4+ T cells
	NK cells
	CD8+ T cells
	Eosinophils
	mDC
	Neutrophils
	pDC
GSE12507	pDC cell line (CAL1)
	immature T cell line (MOLT4)
GSE15215	pDC CD2-
	pDC CD2+
GSE35457	pdc
	cd14_cd16-mono
	cd14-mono
	cd141-dc
	cd16-mono
	cd1c-dc

Connecting Phenotypes: Similar profiles



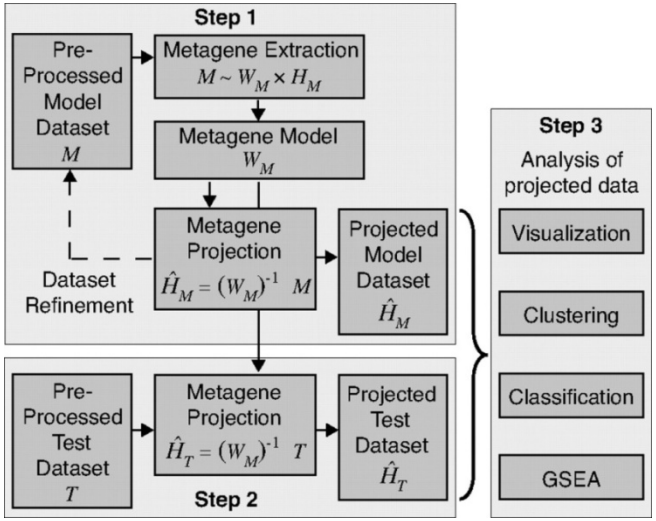
Carmona-Sáez P, et al. Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. **Bioinformatics. 2017**

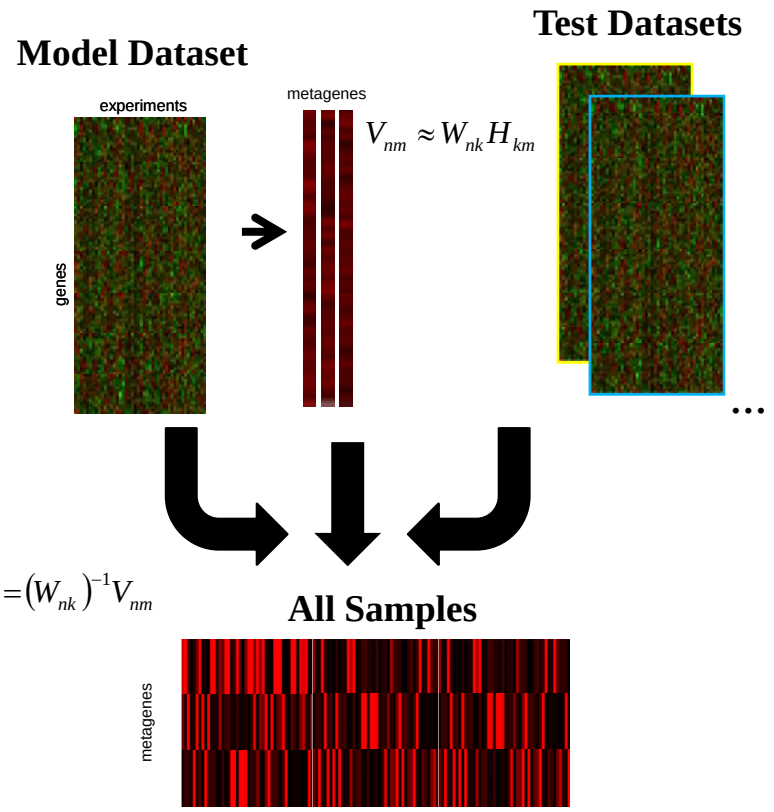
Connecting Phenotypes: Similar profiles



$$V_{nm} \approx W_{nk} H_{km}$$

$$V, W, H \geq 0$$





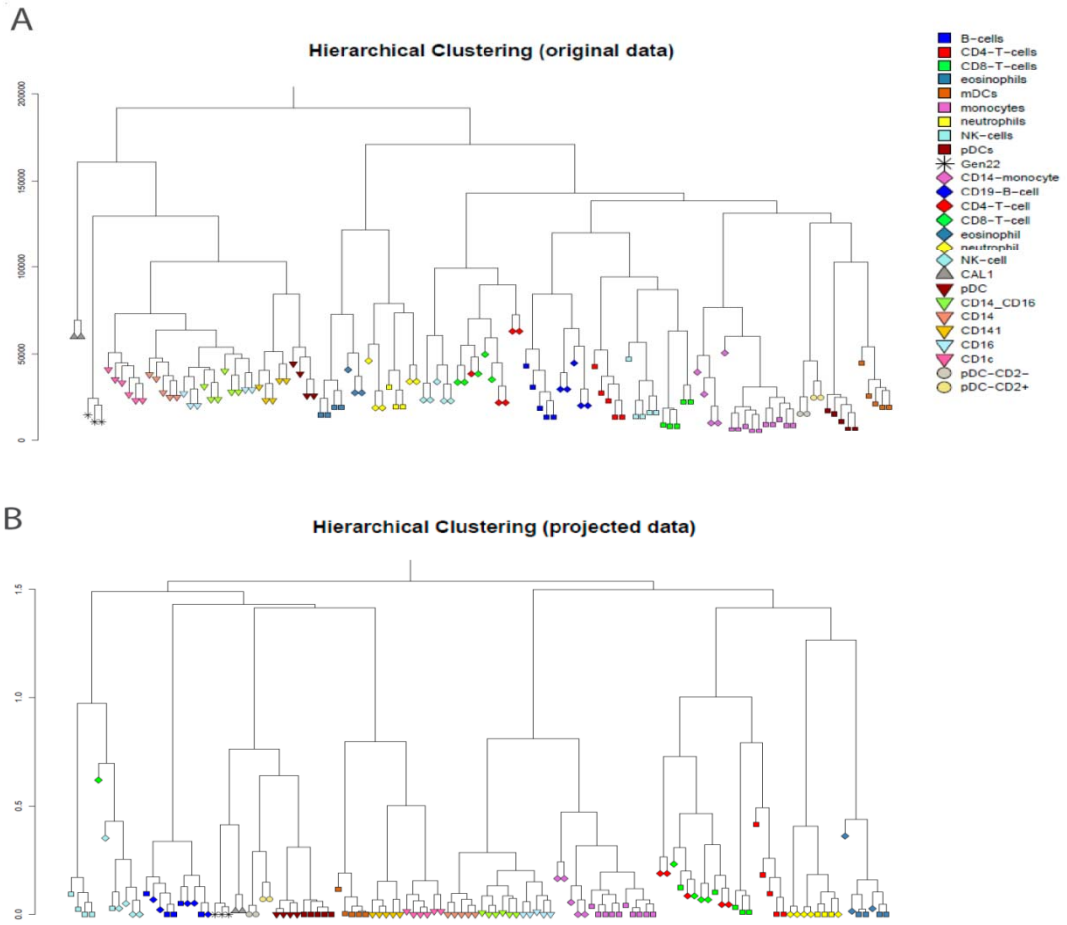
Metagene projection for cross-platform, cross-species characterization of global transcriptional states

Pablo Tamayo*, Daniel Scanfeld*, Benjamin L. Ebert*, Michael A. Gillette*†, Charles W. M. Roberts‡, and Jill P. Mesirov*§

*Eli and Edythe L. Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02141; †Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; and ‡Department of Pediatric Oncology, Dana–Farber Cancer Institute, Boston, MA 02115

Communicated by Edward M. Scolnick, The Broad Institute, Cambridge, MA, February 6, 2007 (received for review December 7, 2006)

Connecting Phenotypes: Similar profiles



Carmona-Sáez P, et al. Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. **Bioinformatics. 2017**