BIOS 516 Introduction to Large-Scale Biomedical Data Analysis

Lecture 2

Steve Qin

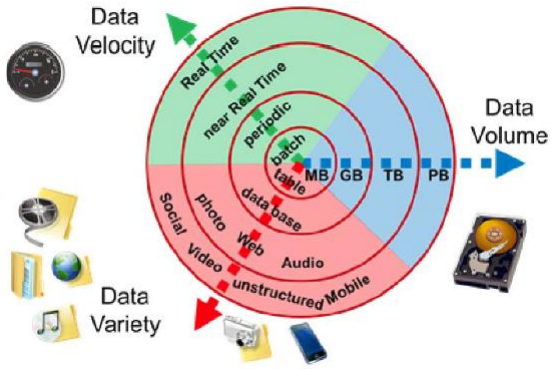
September 7, 2021

Background

- What is Big Data?
- What is the big deal?
- Where can I find biomedical Big Data?
- How can we take advantage of it?

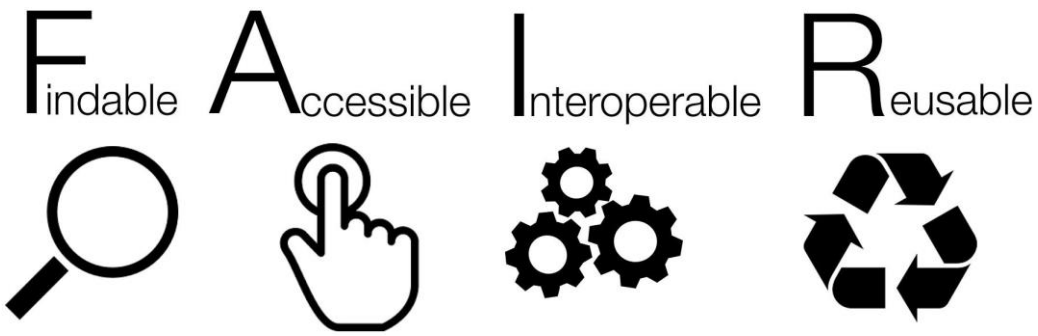
BigData

- Volume
- Variety
- Velocity



https://en.wikipedia.org/wiki/Big_data

FAIR principals

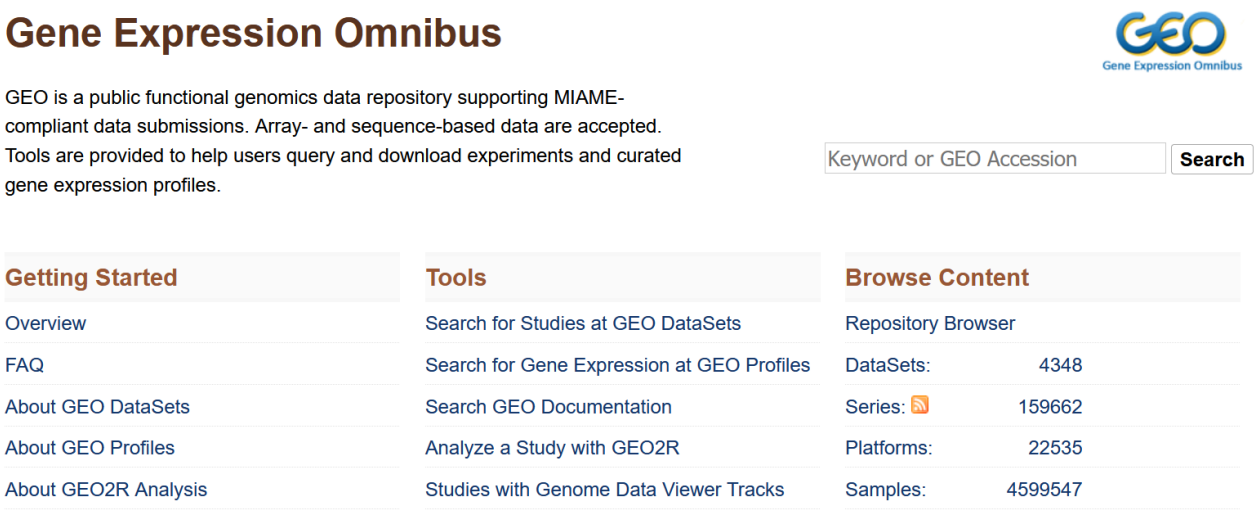


Where is biomedcial BigData?

- GEO, ArrayExpress
- mSigDB
- ENCODE
- TCGA,
- GTEx
- PheGenI, GWAS catalog
- 1000 Genomes,
- UKBB
- …

GEO, SRA, ArrayExpress and GSA

- Repositories
- For sharing high-throughput experimental data, often required by publishers.
  - Originally designed to share microarray data
- Data uploaded by members of the whole research community
- Capture and display rich metadata, and enables query of the metadata.
  - e.g., mouse brain, K562 cell lines.
- No quality control, honor system
- Totally free. No registration is required.
- Operate like a collection of supplemental data of papers.



MIAME and MINSEQE guidelines

MIAME (Minimum Information About a Microarray Experiment)

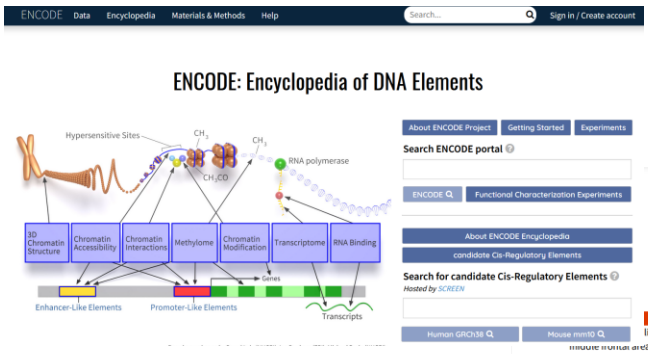
MINSEQE (Minimum Information About a Next-generation Sequencing Experiment)

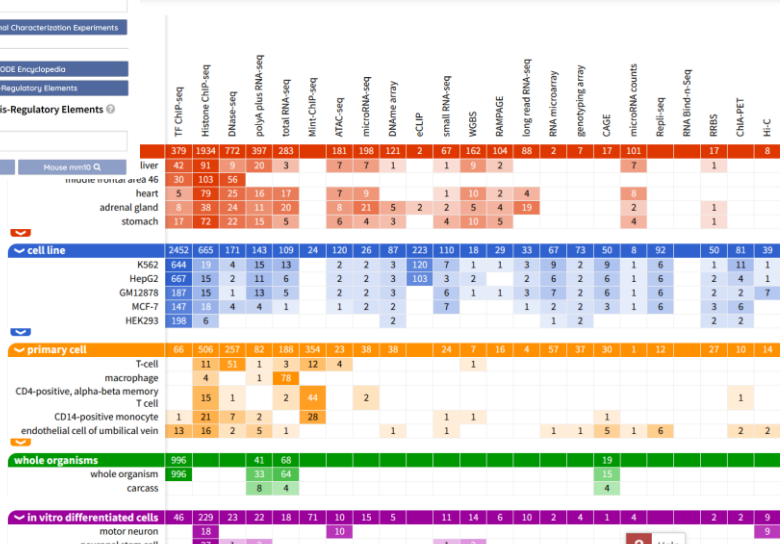
- Raw data for each assay (e.g., CEL or FASTQ files)
- Final processed (normalized) data for the set of assays in the study (e.g., the gene expression data count matrix used to draw the conclusions in the study)
- Essential sample annotation (e.g., tissue, sex and age) and the experimental factors and their values (e.g., compound and dose in a dose response study)
- Experimental design including sample data relationships (e.g., which raw data file relates to which sample, which assays are technical, which are biological replicates)
- Sufficient annotation of the array or sequence features examines (e.g., gene identifiers, genomic coordinates)
- Essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

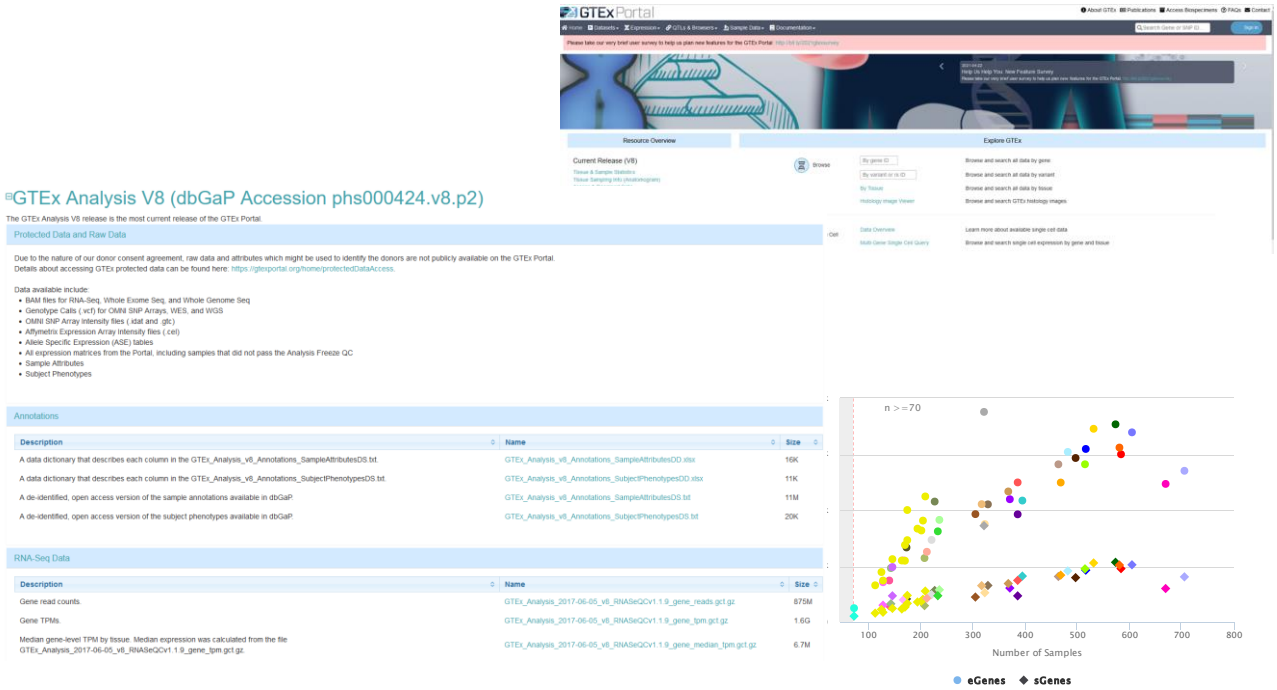
An example: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5366693

Web portal of large consortia

- ENCODE, TCGA, GTEx, 1000 Genomes …
- Only data produced by members of the consortia, follow a set of protocols.
- High quality data, often with extensive quality control.
- Publicly available, popular in the research community.
- Often used as benchmark data.

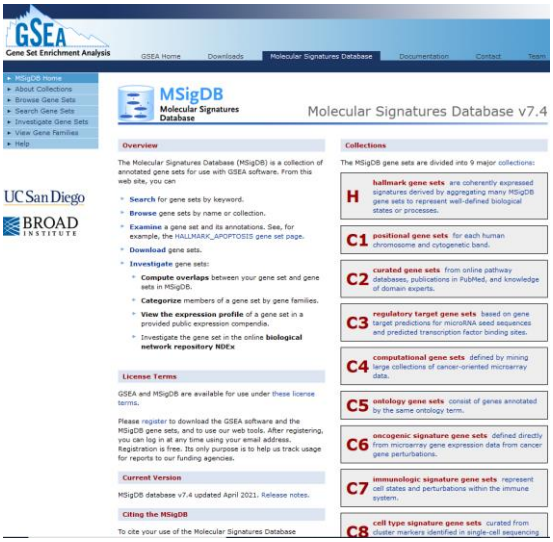


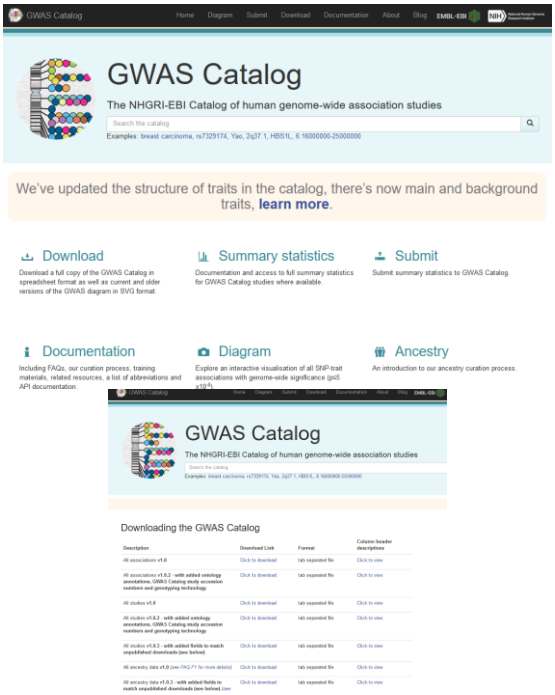




Specialty databases

- mSigDB, GWAS catalog
- Data often not produced by the database owner.
- Designed for secondary or tertiary analyses.
- Data parsed, collected, and often processed and QCed.
- The database just serve as an access point for the collection.
- Often easy to query and free to download.





Various ways to use these data

- Construct informative prior for Bayesian inference
- Build null distributions
- Features to be used in ML algorithms
- Develop supervised ML models
  - As positive training data
  - As negative training data
- Mine novel biological knowledge

Use Big Data to construct informative prior and null distribution

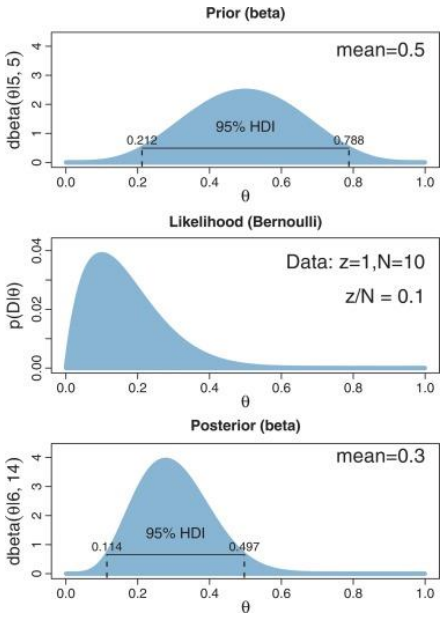
Bayesian inference

Bayes' Theorem

- Bayes' theorem is most commonly used to estimate the state of a hidden, causal variable $\boldsymbol{H}$ based on the measured state of an observable variable $\boldsymbol{D}$:

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D)}$$

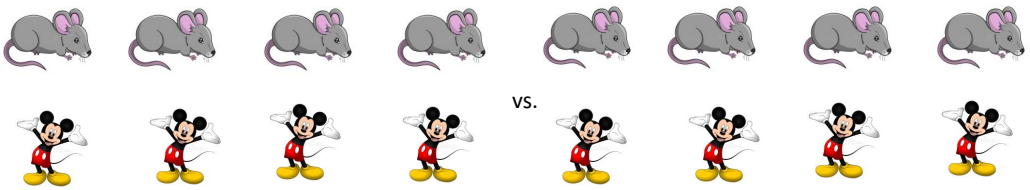
Likelihood, Prior, Posterior, Evidence



Detection of DE genes

- A classical problem in gene expression microarray study: detect differentially expressed (DE) genes.
- DE genes: genes from various samples are expressed differentially in different cell types, tissues, developmental stages or diseases.
- Typically the number of replicates is rather low.

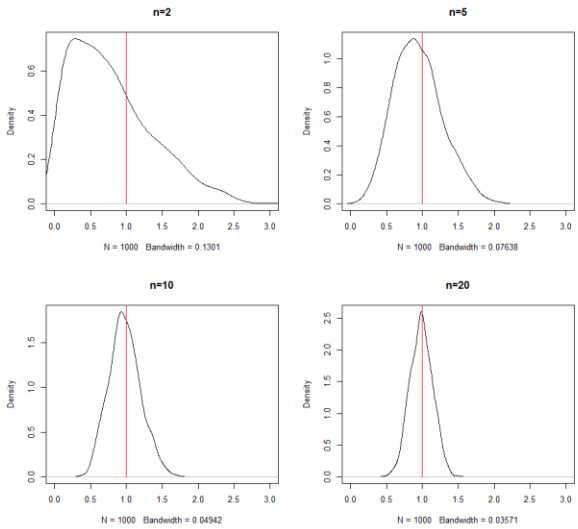
We wish:



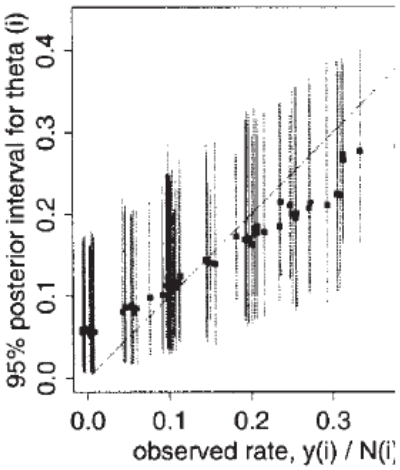
But in reality, we often have:

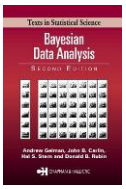
vs.

The problem

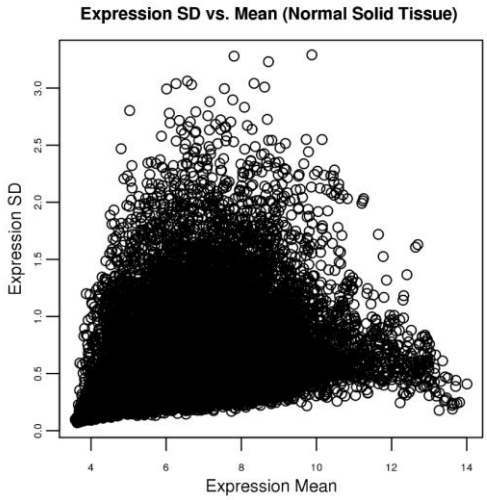


How does hierarchical model work

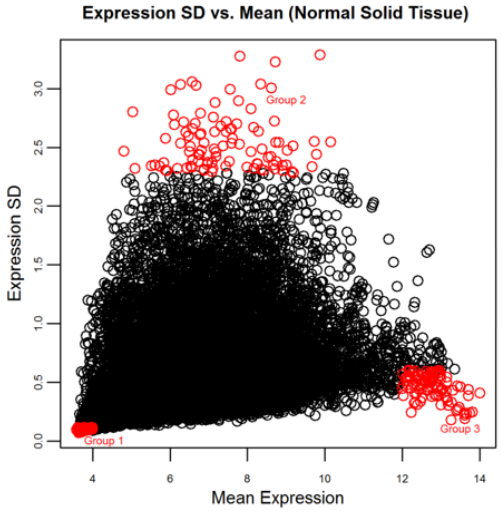




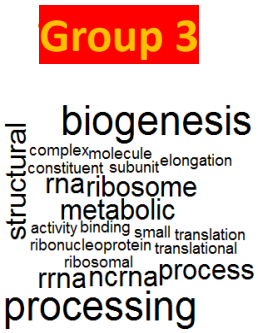
Std dev vs mean

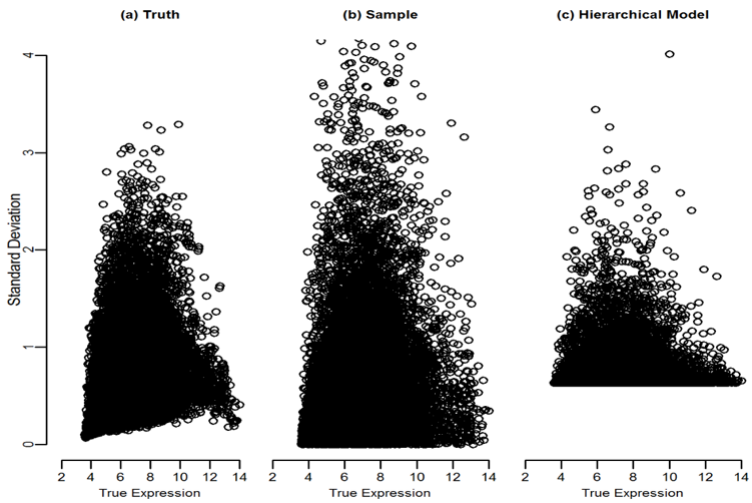


Std dev vs mean



Diverse functions





Drawbacks of hierarchical models

- Restrict to current dataset.
- May overcorrect, especially at the low end.
- Inflated variance means much less discovery power—conservative.

Public databases

- 4,600,533 samples
- 159,703 series

- 74,706 experiments
- 2,557,032 assays

A microarray compendium

CORRESPONDENCE

A global map of human gene expression

To the Editor:
Although there is only one human genome sequence, different genes are expressed

- 5,372 samples
- 206 different studies
- From 163 different labs
- Normalized

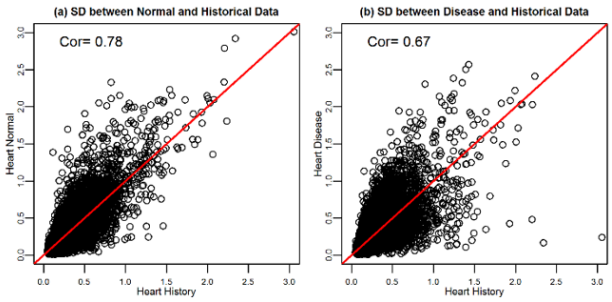
Lukk et al. 2010.

The global gene expression map

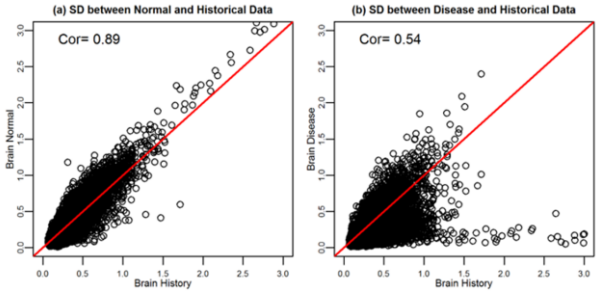
4 meta groups		15 groups	
Group	# of samples	Group	# of samples
cell line	1259	blood neoplasm cell line	166
		non neoplastic cell line	262
		solid tissue neoplasm cell line	831
disease	765	blood non neoplastic disease	388
		solid tissue non neoplastic disease	377
neoplasm	2315	breast cancer	672
		germ cell neoplasm	71
		leukemia	567
		nervous system neoplasm	112
		non breast carcinoma	288
		non leukemic blood neoplasm	334
		other neoplasm	167
		sarcoma	104
normal	1033	normal blood	467
		normal solid tissue	566

Standard deviations from different studies

heart



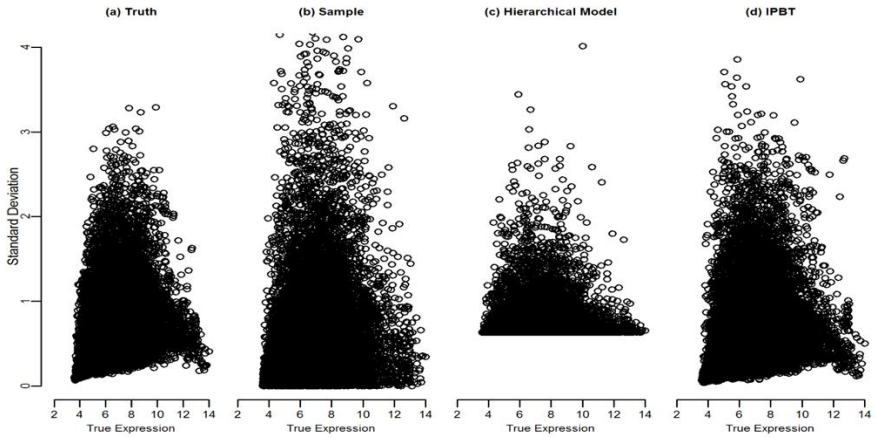
brain



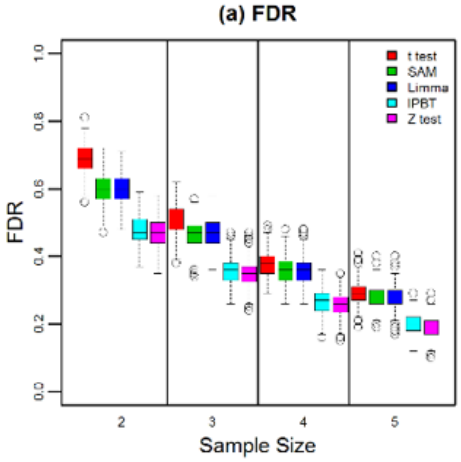
Informative Prior Bayesian Test (IPBT)

- Use historical data to build gene-specific, informative priors.
- Conduct Bayesian inference on σ_i, the standard deviation of gene *i*.
- Either calculate a Bayes factor or test statistics of an adjusted *t*-test and rank genes based on that.

Compare variance estimates

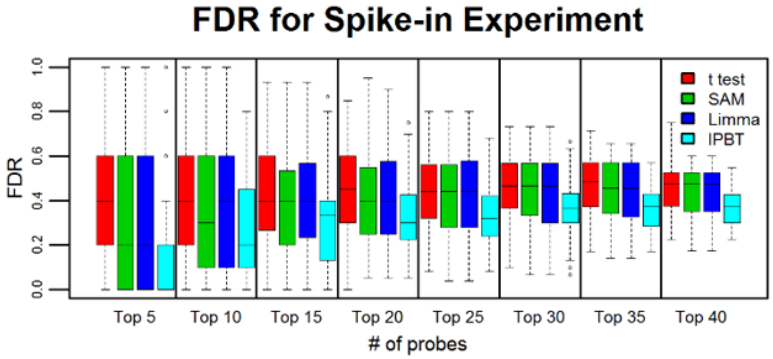


FDR boxplot



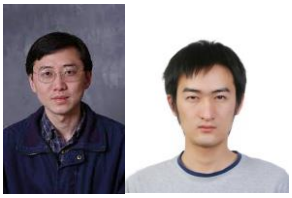
Spike-in experiment

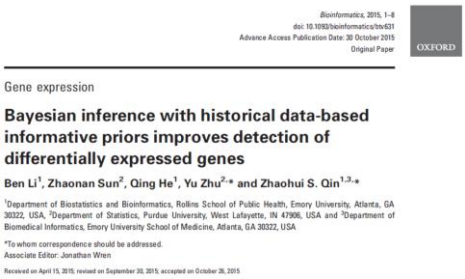
- FDR when first k declare significant



Summary

- Gene-specific properties such as variance can be captured by exploiting existing data that are public-available.
- Utilizing historical data in detecting differentially expressed genes is a better alternative than classical hierarchical model.
- Using informative prior can overcome difficulties faced in low-sample size inference problems.
- It is possible to reduce the number of replicates.





Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv631
Advance Access Publication Date: 30 October 2015
Original Paper

OXFORD

Gene expression

Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes

Ben Li[1], Zhaonan Sun[2], Qing He[1], Yu Zhu[2,*] and Zhaohui S. Qin[1,3,*]

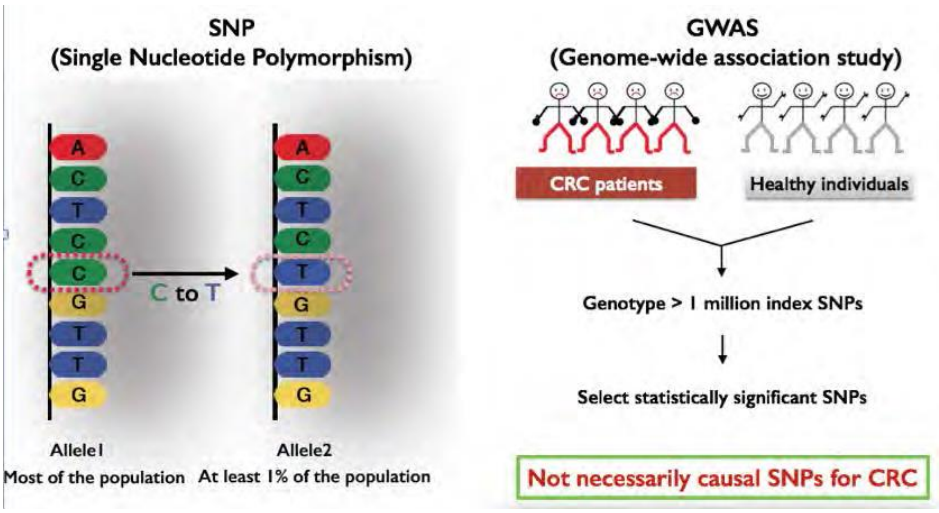
[1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, [2]Department of Statistics, Purdue University, West Lafayette, IN 47906, USA and [3]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

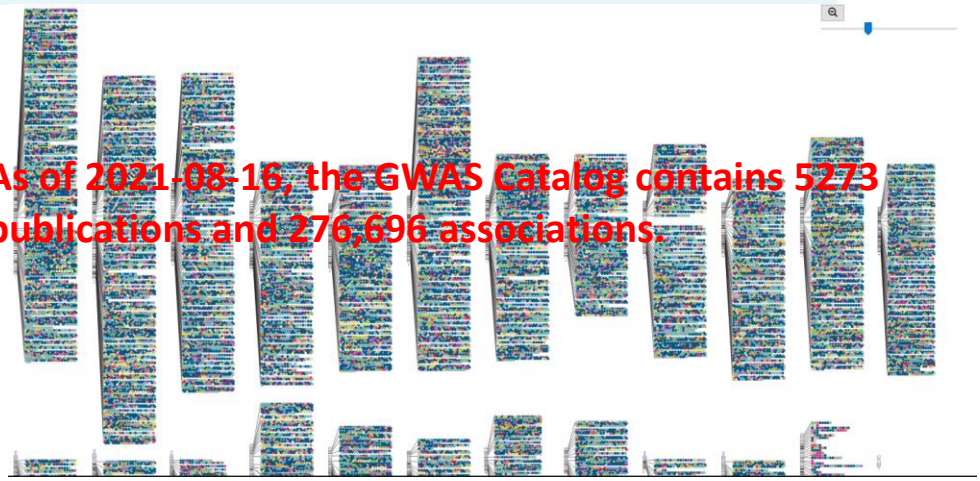
Received on April 15, 2015; revised on September 30, 2015; accepted on October 26, 2015

Use Big Data to annotate different parts of the genome

Genome-wide Association Studies



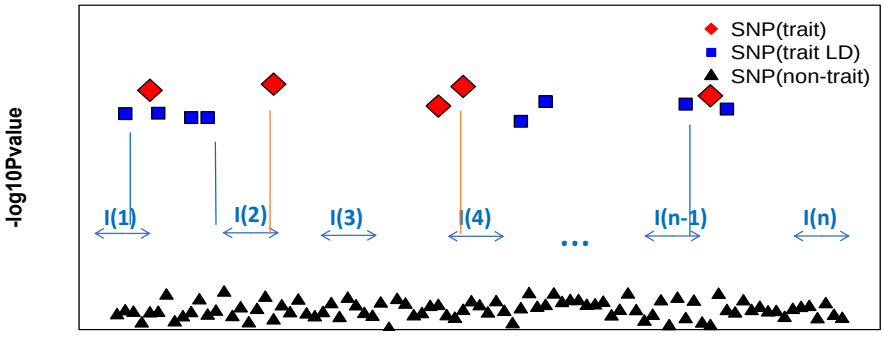
From Peggy J. Farnham.



As of 2021-08-16, the GWAS Catalog contains 5273 publications and 276,696 associations.

traseR: trait-associated SNP enrichment analysis

- The goal is to link GWAS SNPs to genomic loci to uncover potential connections



Example query result (H3K4me1 peaks in T cell)

Table 1. Top-ranked traits for peripheral T cell H3K4me1 peaks

Trait	*P* value	OR	#taSNP hits	#taSNP
All	2.7e-48	1.5	1846	30 553
Behcet syndrome	4.4e-23	6.3	59	274
Diabetes mellitus, type 1	1.7e-11	5.0	33	185
Lupus erythematosus	6.2e-09	3.9	32	223
Arthritis, rheumatoid	1.4e-07	5.1	20	112
Multiple sclerosis	1.6e-05	2.9	26	236
Autoimmune diseases	5.2e-05	15.9	6	15

	inside	outside
#SNP(trait LD)	87	326
#SNP(non-trait)	165,441	3,812,459

traseR: trait-associated SNPs

- Easy-to-use bioinformatics tools that is capable of uncovering potential connections between genomic loci and complex diseases through known GWAS variants.
- Provide annotation to interesting genomic loci found by experiments.

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btv741
Advance Access Publication Date: 18 December 2015
Applications Note

OXFORD

Genome analysis

traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals

Li Chen[1] and Zhaohui S. Qin[2,3]*

[1]Department of Mathematics and Computer Science, [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322 USA and [3]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.
Associate editor: John Hancock

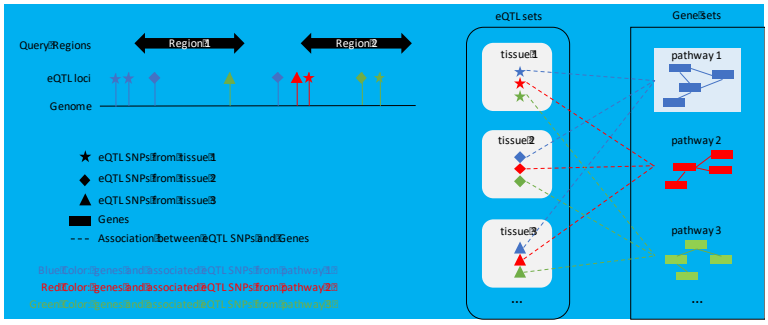
Received on 16 October 2015; revised on 27 November 2015; accepted on 12 December 2015

Loci2path

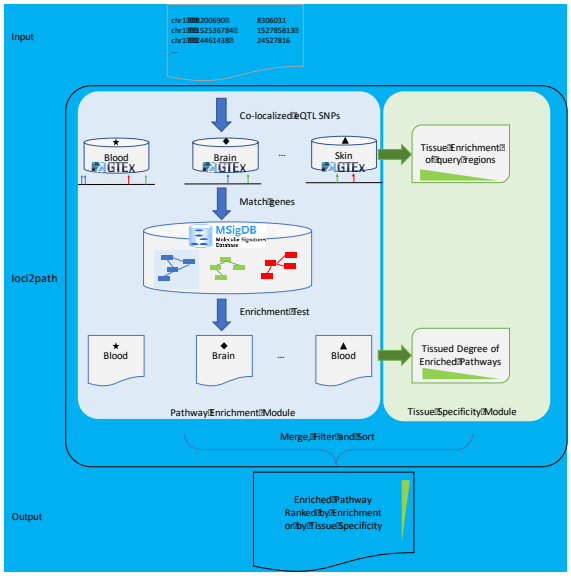
- Annotating a given genomic locus or a set of genomic loci for the non-coding part of the genome.
- Takes advantage of the newly emerged, genome-wide and tissue-specific expression quantitative trait loci (eQTL) information to help annotate a set of genomic intervals in terms of transcription regulation.
- key advantages
  - no longer rely on proximity to link a locus to a gene which has shown to be unreliable;
  - provide the regulatory annotation under the context of specific tissue types which is important.

GTEx

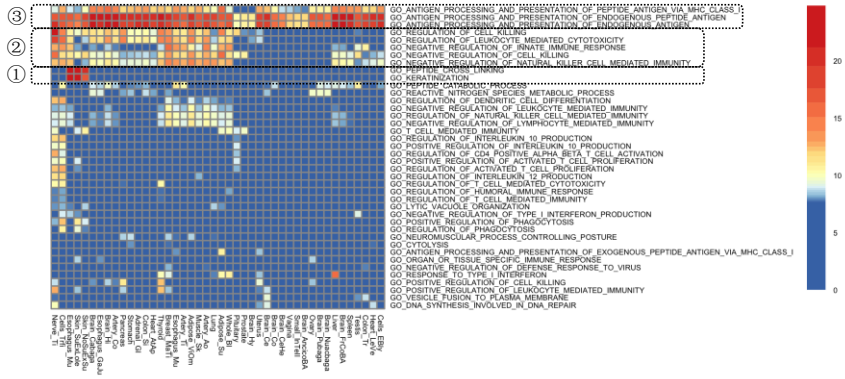
How does it work?



Workflow



Result - Psoriasis



Summary

- Annotate non-coding variants using eQTL resources
- Tissue specificity provide additional information
- Tissue degree suggest different categories of pathways involved in the pathogenesis of psoriasis
- Enrichment in pathways across immune diseases reveals common gene sets of shared disease risks

Bioinformatics, 36(3), 2020, 690–697
doi: 10.1093/bioinformatics/btz669
Advance Access Publication Date: 27 August 2019
Original Paper

OXFORD

Genome analysis

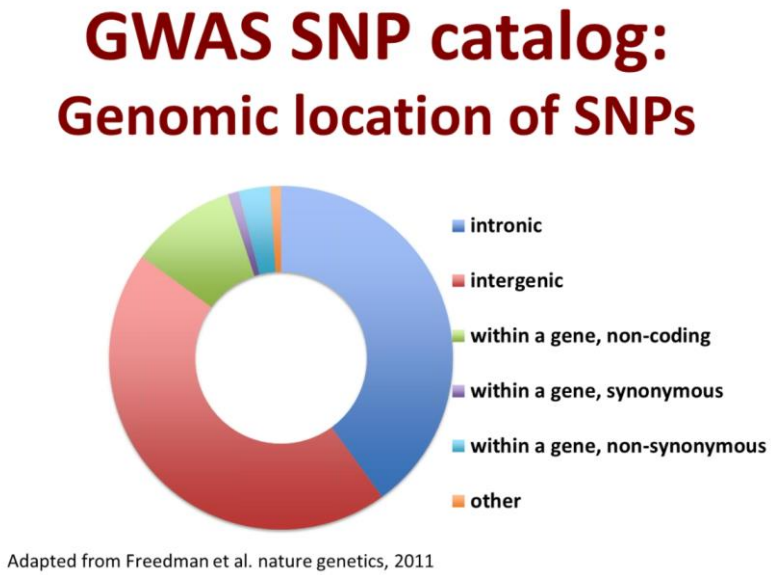
Regulatory annotation of genomic intervals based on tissue-specific expression QTLs

Tianlei Xu [1], Peng Jin [2] and Zhaohui S. Qin [3,*]

Use Big Data as features and training data in ML

Goal: annotate non-coding variants

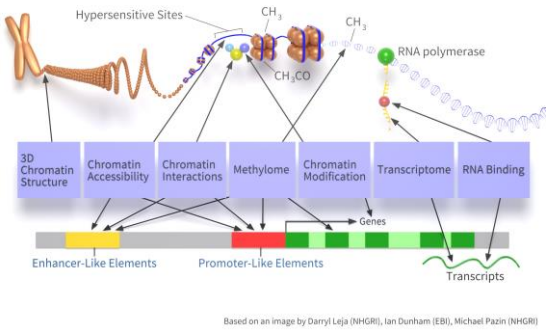
- Many computational tools developed for coding variants
  - SIFT
  - PolyPhen
- Method is needed to annotate the majority (90%) of GWAS-identified variants of complex diseases which are non-coding



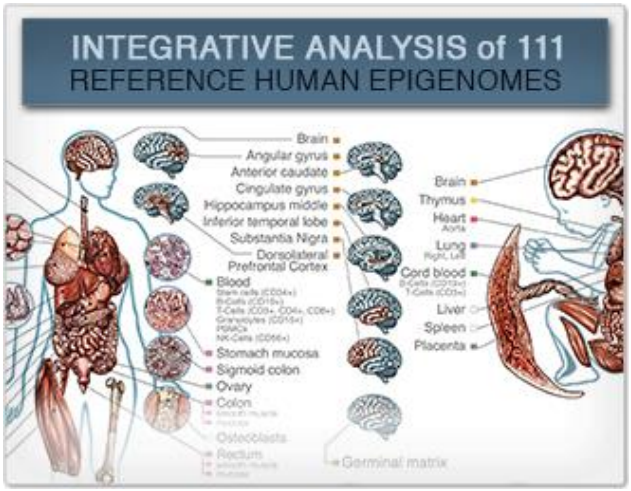
Adapted from Freedman et al. nature genetics, 2011

ational Human Genome Research Institute

ENCODE



ROADMAP epigenomics PROJECT



RegulomeDB 2.0

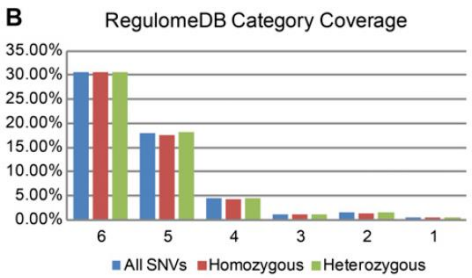
Table 2. RegulomeDB variant classification scheme

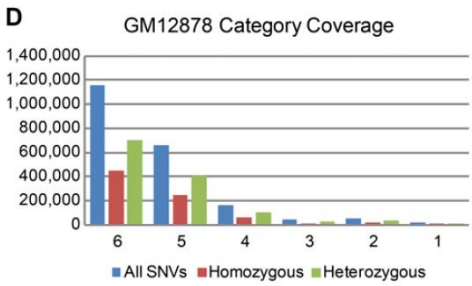
Category	Description
	Category scheme
	Likely to affect binding and linked to expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

0.55%

1.48%

1.16%





Boyle et al. 2012

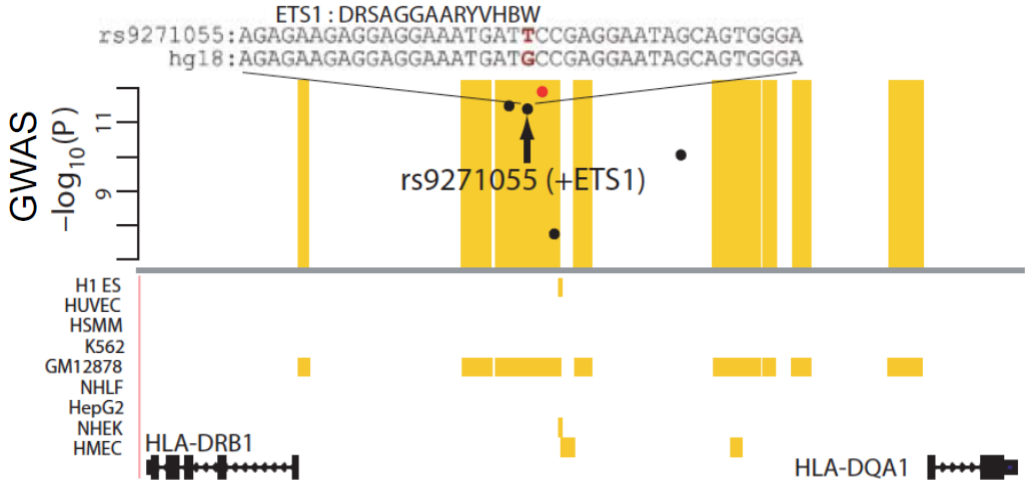
HaploReg v4.1

Nucleic Acids Research 2012

HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants

Lucas D. Ward[1,2,*] and Manolis Kellis[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology and [2]The Broad Institute of MIT and Harvard. Cambridge, MA 02139, USA



Existing methods for annotating noncoding variants

CADD

TECHNICAL REPORTS

nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,3], Daniela M Witten[2,3], Preti Jain[1,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

GWAVA

BRIEF COMMUNICATIONS

Functional annotation of noncoding sequence variants

Graham R S Ritchie[1,2], Ian Dunham[1], Eleftheria Zeggini[1] & Paul Flicek[1,2]

Eigen/EigenPC

TECHNICAL REPORTS

nature genetics

A spectral approach integrating functional genomic annotations for coding and noncoding variants

Iuliana Ionita-Laza[1,8], Kenneth McCallum[2,8], Bin Xu[2] & Joseph D Buxbaum[3–7]

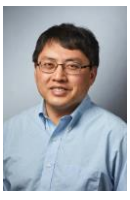
GenoCanyon

SCIENTIFIC REPORTS

OPEN

A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data

Qiongshi Lu[1], Yiming Hu[1], Jiehuan Sun[1], Yunmei Cheng[1], Kai-Hei Cheung[1,2,3] & Hongyu Zhao[1,2,3]



PAFA

Genome Medicine

METHOD Open Access

Prioritization and functional assessment of noncoding variants associated with complex diseases

Lin Zhou[1,2] and Fangqing Zhao[1,2,3*]

PINES

Genome Biology

SOFTWARE Open Access

PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants

Corneliu A. Bodea[1,2,8], Adele A. Mitchell[3], Alex Bloemendal[4], Aaron G. Day-Williams[1,5], Heiko Runz[2,5] and Shamil R. Sunyaev[2,3,6]*

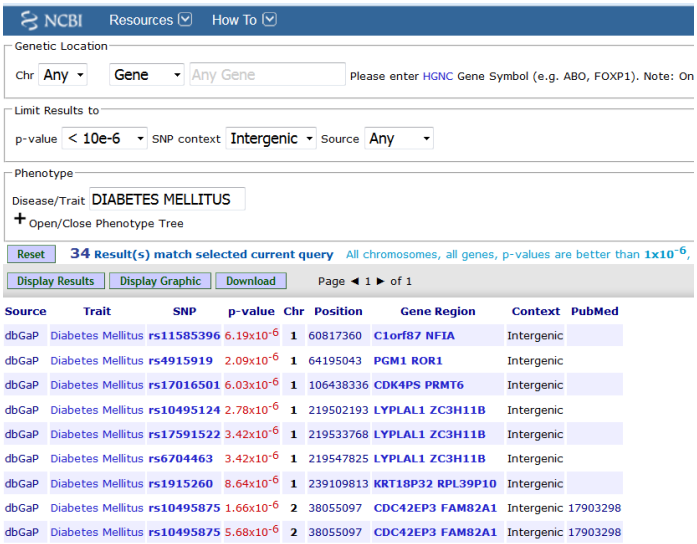
Information sources for identifying non-coding variants?

- Phylogenetic conservation
  - PhastCon scores
  - GERP scores
- Genomic profile
  - Whether it overlap with any known transcription factor binding motif?
- Epigenomic profile
  - TF binding
  - Histone modification
  - DNA methylation
  - …

DIVAN: DIsease-specific Variant ANnotation

- A unique model for each disease/phenotype
  - 45 diseases from 12 categories.
- Using trait-associated SNPs identified by GWAS as training data
- Using genomic and epigenomic data as features
- Use machine learning techniques to distinguish risk variants from benign variants.

GWAS SNP collection (1)



NCBI Resources How To

Genetic Location

Chr Any Gene Any Gene Please enter HGNC Gene Symbol (e.g. ABO, FOXP1). Note: On

Limit Results to

p-value < 10e-6 SNP context Intergenic Source Any

Phenotype

Disease/Trait DIABETES MELLITUS

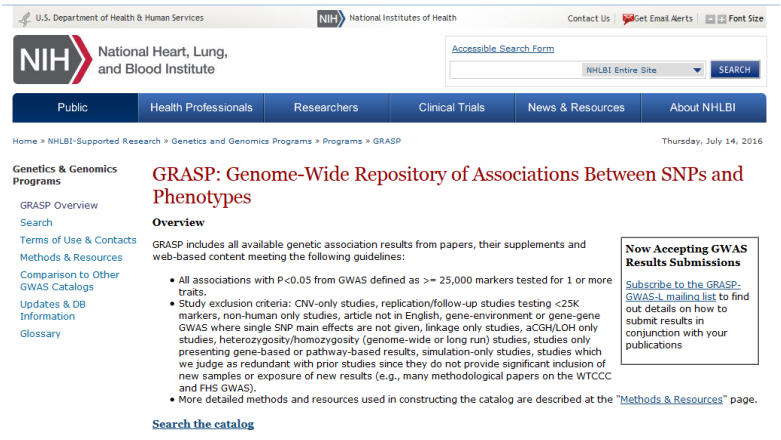
+ Open/Close Phenotype Tree

Reset **34 Result(s) match selected current query** All chromosomes, all genes, p-values are better than $1x10^{-6}$,

Display Results Display Graphic Download Page ◄ 1 ► of 1

Source	Trait	SNP	p-value	Chr	Position	Gene Region	Context	PubMed
dbGaP	Diabetes Mellitus	rs11585396	$6.19x10^{-6}$	1	60817360	C1orf87 NFIA	Intergenic	
dbGaP	Diabetes Mellitus	rs4915919	$2.09x10^{-6}$	1	64195043	PGM1 ROR1	Intergenic	
dbGaP	Diabetes Mellitus	rs17016501	$6.03x10^{-6}$	1	106438336	CDK4PS PRMT6	Intergenic	
dbGaP	Diabetes Mellitus	rs10495124	$2.78x10^{-6}$	1	219502193	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs17591522	$3.42x10^{-6}$	1	219533768	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs6704463	$3.42x10^{-6}$	1	219547825	LYPLAL1 ZC3H11B	Intergenic	
dbGaP	Diabetes Mellitus	rs1915260	$8.64x10^{-6}$	1	239109813	KRT18P32 RPL39P10	Intergenic	
dbGaP	Diabetes Mellitus	rs10495875	$1.66x10^{-6}$	2	38055097	CDC42EP3 FAM82A1	Intergenic	17903298
dbGaP	Diabetes Mellitus	rs10495875	$5.68x10^{-6}$	2	38055097	CDC42EP3 FAM82A1	Intergenic	17903298

GWAS SNP collection (2)



U.S. Department of Health & Human Services NIH National Institutes of Health Contact Us Get Email Alerts Font Size

NIH National Heart, Lung, and Blood Institute

Accessible Search Form NHLBI Entire Site SEARCH

Public | Health Professionals | Researchers | Clinical Trials | News & Resources | About NHLBI

Home » NHLBI-Supported Research » Genetics and Genomics Programs » Programs » GRASP Thursday, July 14, 2016

Genetics & Genomics Programs

- GRASP Overview
- Search
- Terms of Use & Contacts
- Methods & Resources
- Comparison to Other GWAS Catalogs
- Updates & DB Information
- Glossary

GRASP: Genome-Wide Repository of Associations Between SNPs and Phenotypes

Overview

GRASP includes all available genetic association results from papers, their supplements and web-based content meeting the following guidelines:

- All associations with P<0.05 from GWAS defined as >= 25,000 markers tested for 1 or more traits.
- Study exclusion criteria: CNV-only studies, replication/follow-up studies testing <25K markers, non-human only studies, article not in English, gene-environment or gene-gene GWAS where single SNP main effects are not given, linkage only studies, aCGH/LOH only studies, heterozygosity/homozygosity (genome-wide or long run) studies, studies only presenting gene-based or pathway-based results, simulation-only studies, studies which we judge as redundant with prior studies since they do not provide significant inclusion of new samples or exposure of new results (e.g., many methodological papers on the WTCCC and FHS GWAS).
- More detailed methods and resources used in constructing the catalog are described at the "Methods & Resources" page.

Now Accepting GWAS Results Submissions

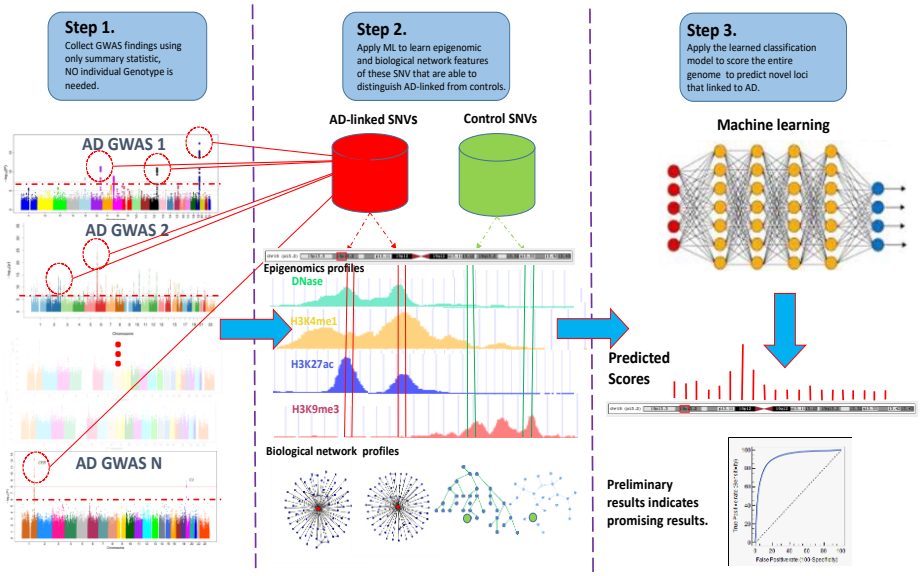
Subscribe to the GRASP-GWAS-L mailing list to find out details on how to submit results in conjunction with your publications

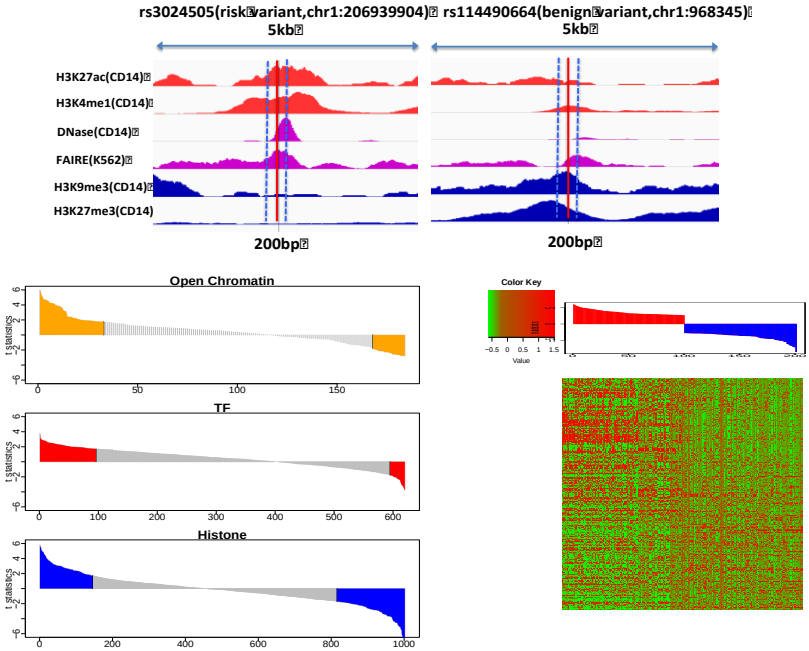
Search the catalog

Epigenomic features utilized

Data Source	cell line	TF/Histone	feature
REMC DNase	73	-	73
REMC Histone	109	31	735
ENCODE DNase	80	-	80
ENCODE FAIRE	31	-	31
ENCODE TF(HAIB)	19	76	293
ENCODE TF(SYDH)	31	100	279
ENCODE Histone	18	42	267
ENCODE RNA Polymerase	31	2	49
Total	**261**	**217**	**1806**

DIVAN

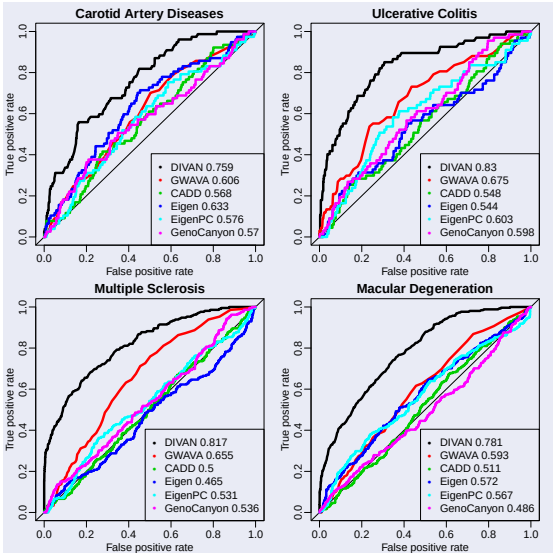




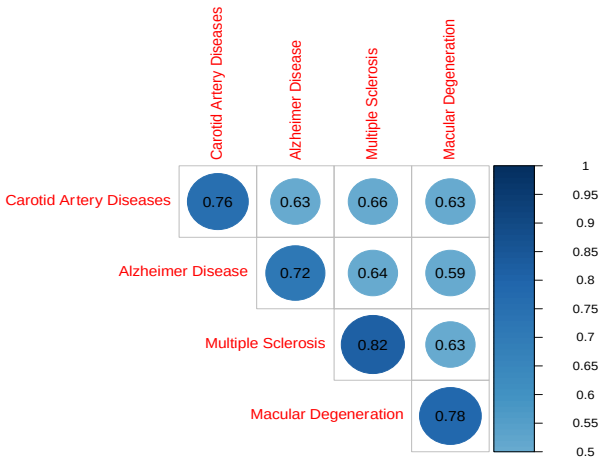
Choose GWAS variants

- We choose 45 diseases/phenotypes spanning 12 disease/phenotype classes, with at least 50 disease-SNP associations from Association Results Browser
- Benign SNPs are sampled from the 1000 Genomes (Phase I) with same distance (SNP to nearest TSS) distribution as risk SNPs

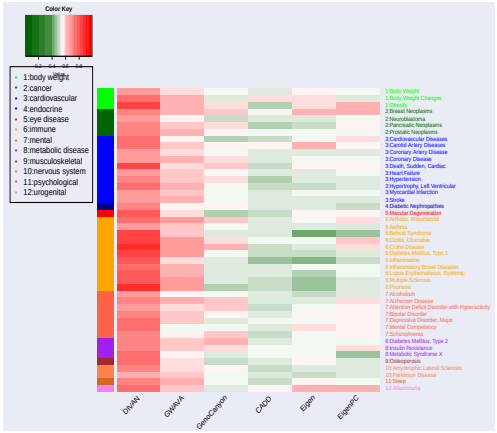
Performance comparison on four diseases

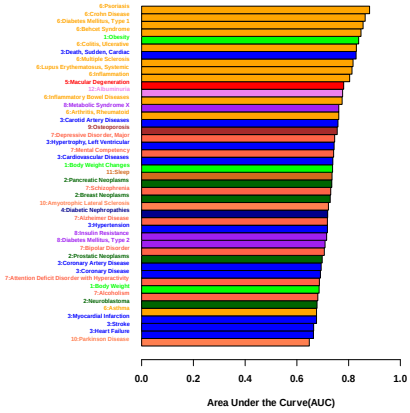


The importance of disease-specificity



AUC values of the 45 diseases tested



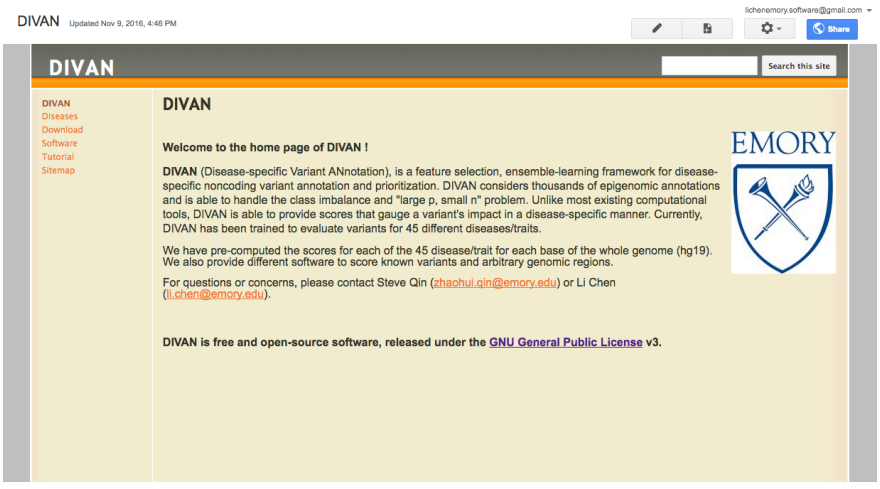


AUC from 0.66 to 0.88 with median 0.74

Immune diseases are best predicted

DIVAN website

https://sites.google.com/site/emorydivan/



Summary

- Disease-specific risk variant identification is feasible
- Training data obtained from GWAS results and 1000 Genomes databases.
- Features are collected from genomics profiling data stored in ENCODE, REMC.

Chen *et al. Genome Biology* (2016) 17:252
DOI 10.1186/s13059-016-1112-z

Genome Biology

METHOD Open Access

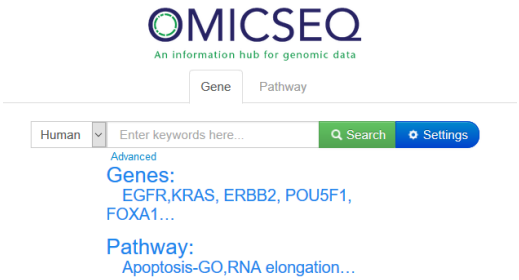
DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles

Li Chen[1], Peng Jin[2] and Zhaohui S. Qin[3,4*]

Extract insights from Big Data

Omicseq:

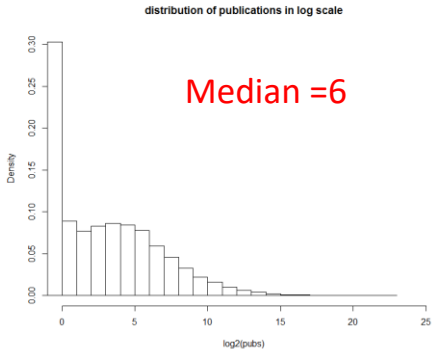
An omics data search engine



Literature is the major source of biomedical knowledge

- Accurate (with quality control)
- Specific and definitive
- With established infrastructure and technology to conduct effective literature mining

How much do we know?



- For an obscure gene, little information is known in the literature.

69

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed sumo1p1

Create RSS Create alert Advanced

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed Commons
Reader comments
Trending articles

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary **Sort by**: Most Recent **Send to**

See **SUMO1P1** SUMO1 pseudogene 1 in the Gene database

Search results

Items: 3

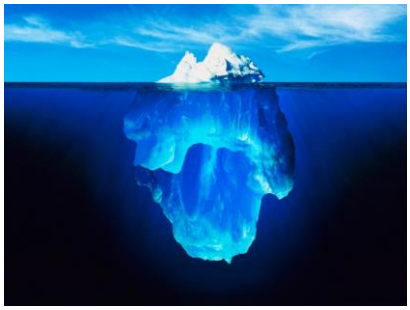
1. Phenotype and Tissue Expression as a Function of Genetic Risk in Polycystic Ovary Syndrome.
Pau CT, Mosbruger T, Saxena R, Welt CK.
PLoS One. 2017 Jan 9;12(1):e0168870. doi: 10.1371/journal.pone.0168870. eCollection 2017.
PMID: 28068351 **Free PMC Article**
Similar articles

2. Cross-ethnic meta-analysis of genetic variants for polycystic ovary syndrome.
Louwers YV, Stolk L, Uitterlinden AG, Laven JS.
J Clin Endocrinol Metab. 2013 Dec;98(12):E2006-12. doi: 10.1210/jc.2013-2495. Epub 2013 Oct 8.
PMID: 24106282
Similar articles

3. Genome-wide study identifies PTPRO and WDR72 and FOXQ1-**SUMO1P1** interaction associated with neurocognitive function

Literature is limited

- Only interesting (for authors and the journal, not necessarily for all audience) and significant findings were reported
- Mundane events, like most TF binding, gene expression changes do not make it to the papers
- Polished yet subjective and selective



71

Genotype-Tissue expression (GTEx)

COSMIC
Catalogue of somatic mutations in cancer

ROADMAP epigenomics PROJECT

modENCODE

National Human Genome Research Institute
ENCODE

BLUEPRINT epigenome

GSA Genome Sequence Archive

FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

GEO Gene Expression Omnibus

dbGaP
GENOTYPES and PHENOTYPES

The eMERGE Network
electronic Medical Records & Genomics
A consortium of biorepositories linked to electronic medical records data

1000 Genomes
A Deep Catalog of Human Genetic Variation

SEUVADIS CONSORTIUM

CCLE Cancer Cell Line Encyclopedia

The Cancer Genome Atlas
Understanding genomics to improve cancer care

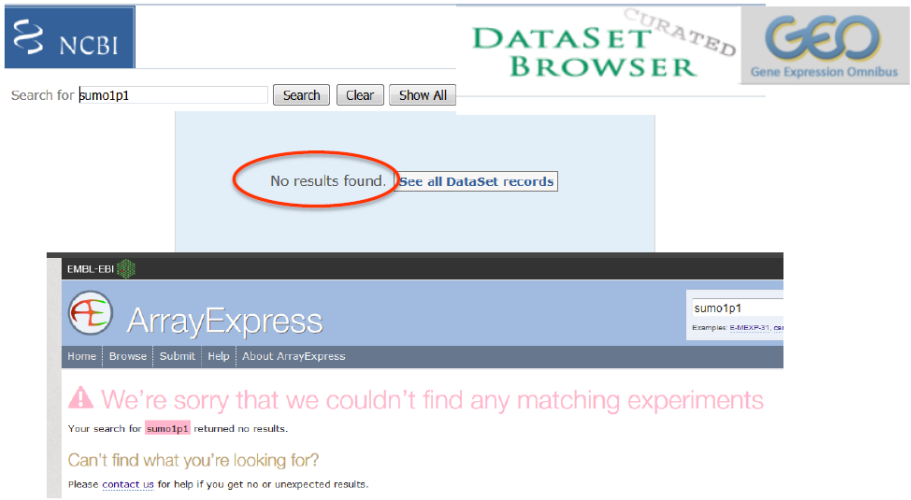
ICGC

International HapMap Project

NIH LINCS PROGRAM

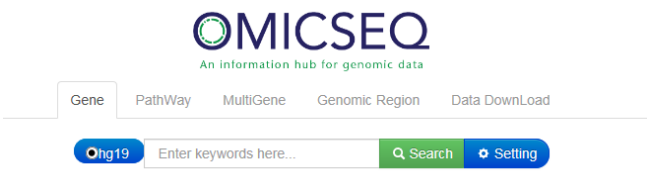
72

Perform a search in public data repositories



Our goals

- To develop a website that links to ALL the biomedical data that ever produced
- The database only stores data that are processed and ready-to-use
- To build a search engine from which one can get information on any gene.
- Do not rely on the metadata.



- Collect different variety of –omics data.
- Develop standardized protocols to process different types of data.
- Store these processed data in databases.
- Collect metadata.
- Develop a query engine for dataset searching.
- Develop a ranking algorithm “TrackRank”.
- Facilitating easy downloading of the processed, ready-to-analyze data.

Data types to be included

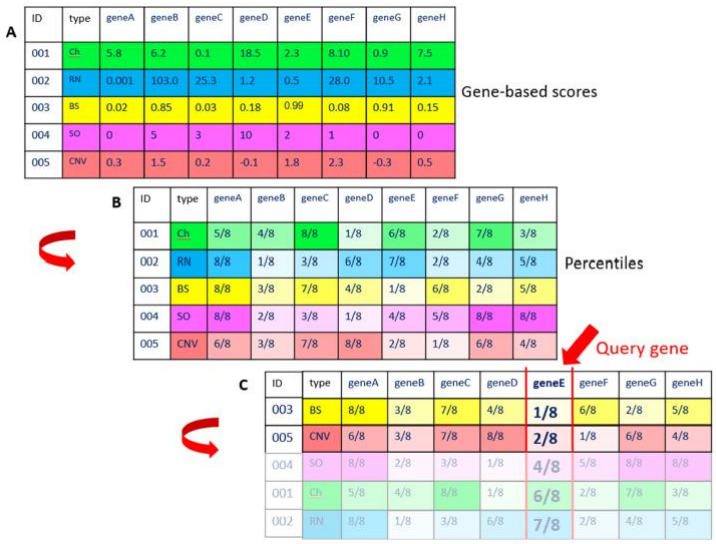
- From experimental assays
  - FPKM values from RNA-seq,
  - Read counts at promoters from ChIP-seq,
  - P-values of detecting DE genes using microarray,
  - Pausing index from GRO-seq,
  - Average methylation level at the promoters from BS-seq

Compare across data types (I)

- Our hypothesis is that a gene "plays an important role in a dataset" if its score ranks at the top among all genes in the genome.
- We developed trackRank algorithm to rank datasets using this idea.

Raw data → Summary stat/score → Rank

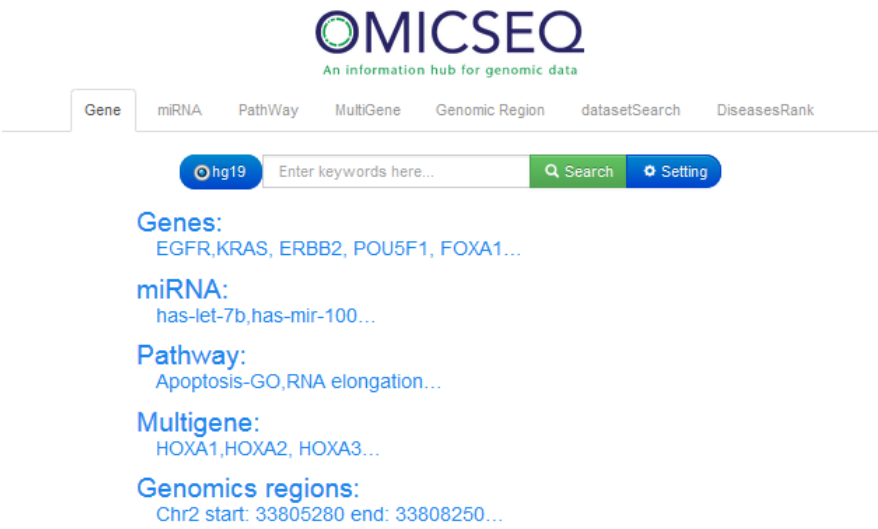
77



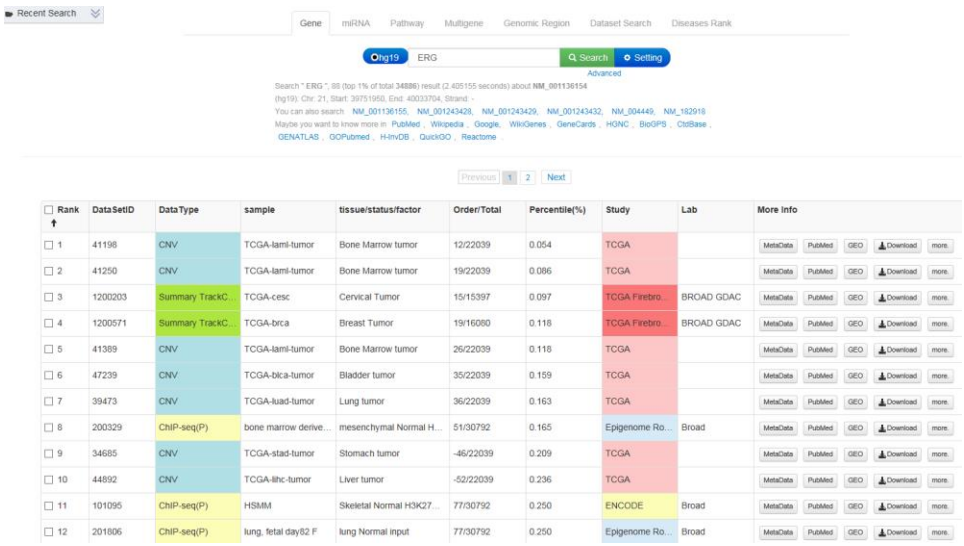
ID	type	geneA	geneB	geneC	geneD	geneE	geneF	geneG	geneH
001	Ch	5.8	6.2	0.1	18.5	2.3	8.10	0.9	7.5
002	RN	0.001	103.0	25.3	1.2	0.5	28.0	10.5	2.1
003	BS	0.02	0.85	0.03	0.18	0.99	0.08	0.91	0.15
004	SO	0	5	3	10	2	1	0	0
005	CNV	0.3	1.5	0.2	-0.1	1.8	2.3	-0.3	0.5

ID	type	geneA	geneB	geneC	geneD	geneE	geneF	geneG	geneH
001	Ch	5/8	4/8	8/8	1/8	6/8	2/8	7/8	3/8
002	RN	8/8	1/8	3/8	6/8	7/8	2/8	4/8	5/8
003	BS	8/8	3/8	7/8	4/8	1/8	6/8	2/8	5/8
004	SO	8/8	2/8	3/8	1/8	4/8	5/8	8/8	8/8
005	CNV	6/8	3/8	7/8	8/8	2/8	1/8	6/8	4/8

ID	type	geneA	geneB	geneC	geneD	**geneE**	geneF	geneG	geneH
003	BS	8/8	3/8	7/8	4/8	**1/8**	6/8	2/8	5/8
005	CNV	6/8	3/8	7/8	8/8	**2/8**	1/8	6/8	4/8
004	SO	8/8	2/8	3/8	1/8	**4/8**	5/8	8/8	8/8
001	Ch	5/8	4/8	8/8	1/8	**6/8**	2/8	7/8	3/8
002	RN	8/8	1/8	3/8	6/8	**7/8**	2/8	4/8	5/8



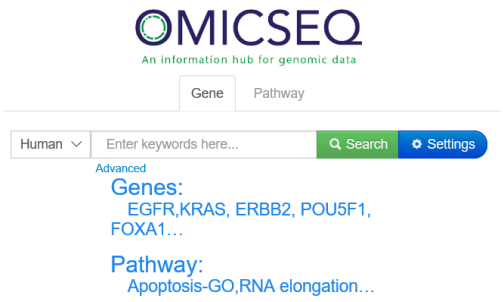
Omicseq result page



80

Summary

- Developed Omicseq: a omics data search engine, and a biological knowledge discovery tool.
- Does not rely on metadata
- Powered by trackRank algorithm
- Powerful resource for data mining
- Try it http://www.omicseq.org



Nucleic Acids Research, 2017 1
doi: 10.1093/nar/gkx258

Omicseq: a web-based search engine for exploring omics datasets

Xiaobo Sun[1], William S. Pittard[2], Tianlei Xu[1], Li Chen[1], Michael E. Zwick[3], Xiaoqian Jiang[4], Fusheng Wang[5,6] and Zhaohui S. Qin[2,7,*]

[1]Department of Mathematics and Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA, [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA, [3]Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA, [4]Health Science Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [5]Department of Biomedical Informatics, Stony Brook University, HSC L3-043, Stony Brook, NY 11794, USA, [6]Department of Computer Science, Stony Brook University, Computer Science Building, Stony Brook, NY 11794, USA and [7]Department of Biomedical Informatics, Emory University School of Medicine, 36 Eagle Row, Atlanta, GA 30322, USA

Received February 20, 2017; Revised March 27, 2017; Editorial Decision March 31, 2017; Accepted April 04, 2017

Ways to handle Big Data

Distributed systems to handle Big Data

OXFORD (GIGA)n SCIENCE

GigaScience, 7, 2018, 1–16

doi: 10.1093/gigascience/giy052
Advance Access Publication Date: 10 May 2018
Technical Note

TECHNICAL NOTE

Optimized distributed systems achieve significant performance improvement on sorted merging of massive VCF files

Xiaobo Sun[1], Jingjing Gao[2], Peng Jin[3], Celeste Eng[4], Esteban G. Burchard[4], Terri H. Beaty[5], Ingo Ruczinski[6], Rasika A. Mathias[7], Kathleen Barnes[8], Fusheng Wang[9,*], Zhaohui S. Qin[2,10,*] and CAAPA consortium[11]

Summary

- Genomics Big data widely available.
- There are many different ways to utilize these Big Data
- If carefully designed, These Big Data give us opportunity to gain insights and make new discoveries.
- Statistics and ML thinking is required to use these resources effectively.
- Need latest informatics methods to enables the use of Big Data.

References

- Li B, Sun Z, He Q, Sun Z, Zhu Y, Qin ZS (2015) Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics*. **32.** 682-689.
- Chen L, Qin ZS. (2015) traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics.* **32.** 1214-1216.
- Chen L, Jin P, Qin ZS**.** (2016) DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol.* **17.** 252.
- Sun X, Pittard WS, Xu T, Chen L, Zwick ME, Jiang X, Wang F, Qin ZS. (2017) Omicseq: A web-based search engine for exploring omics datasets. *Nucleic Acids Research.* **45.** W445-W452.
- Sun X, Gao J, Jin P, Eng C, Burchard EG, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Wang F, Qin ZS on behalf of CAAPA consortium (2018) Optimized Distributed Systems Achieve Significant Performance Improvement on Sorted Merging of Massive VCF Files. *Gigascience*. **7.**
- Xu T, Jin P, Qin ZS. (2020) Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics.* **36.** 690-697.

Thank you

Questions:

zhaohui.qin@emory.edu