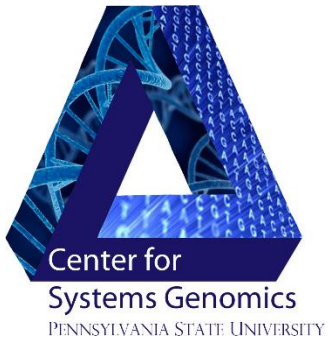
Bioinformatics pipelines, workflows, and resources

Marylyn D. Ritchie, PhD
Professor, Biochemistry & Molecular Biology
Center for Systems Genomics
The Pennsylvania State University

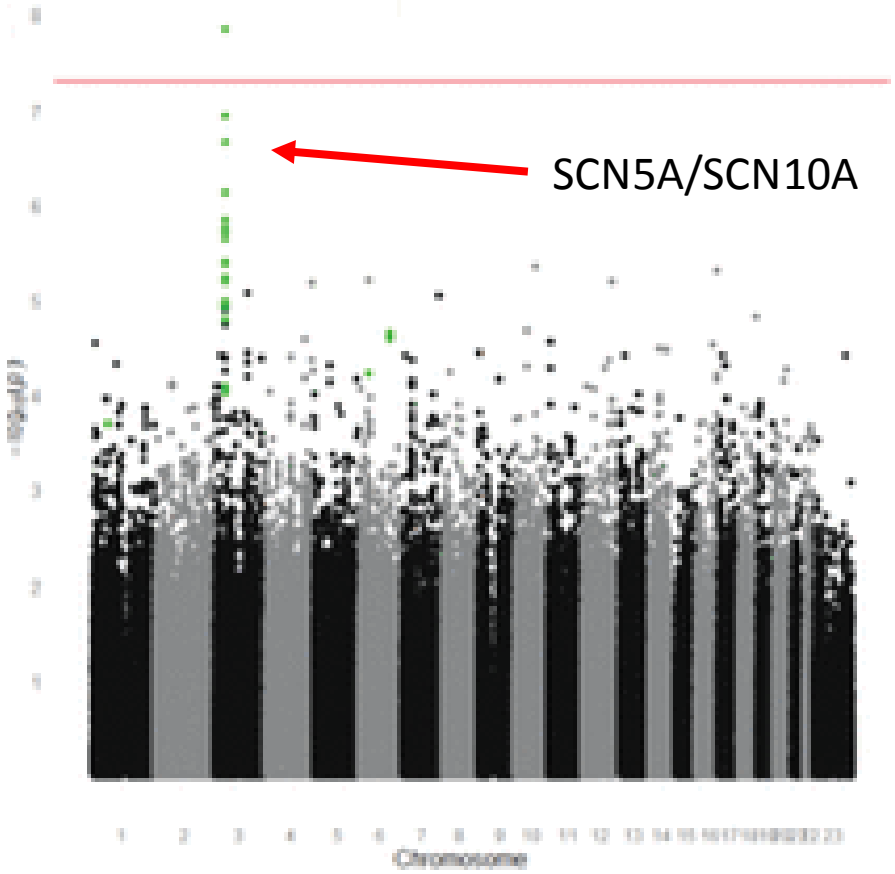


PENNSTATE

1855

GWAS: Genome-Wide Association Study

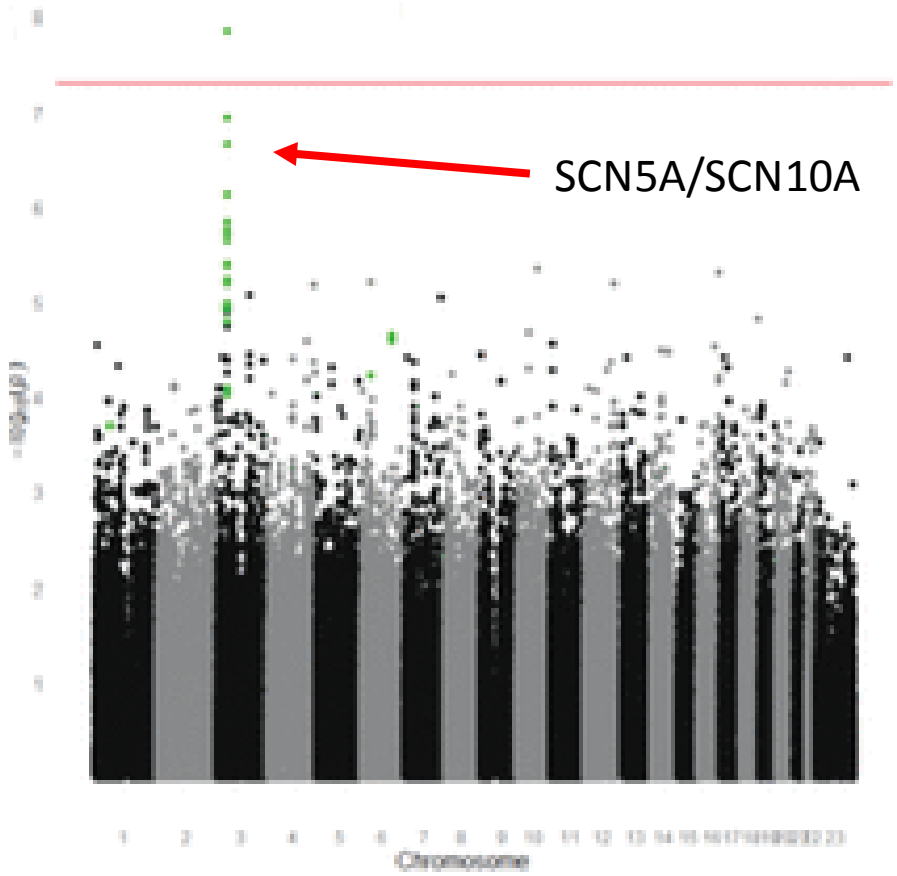
Genome-wide association analysis of QRS duration using sex-adjusted linear regression



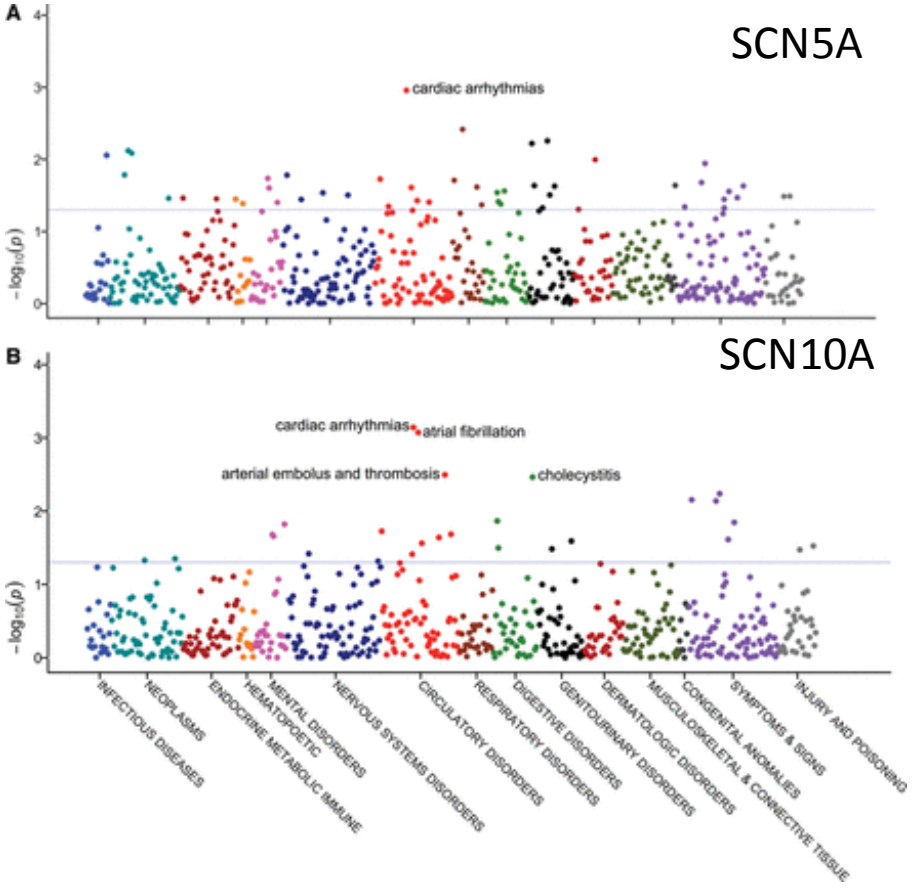
American Heart Association.
Ritchie M D et al. Circulation. 2013;127:1377-1385

PheWAS: Phenome-Wide Association Study

Genome-wide association analysis of QRS duration using sex-adjusted linear regression



Phenome-wide association analysis of SCN5A/SCN10A using logistic regression

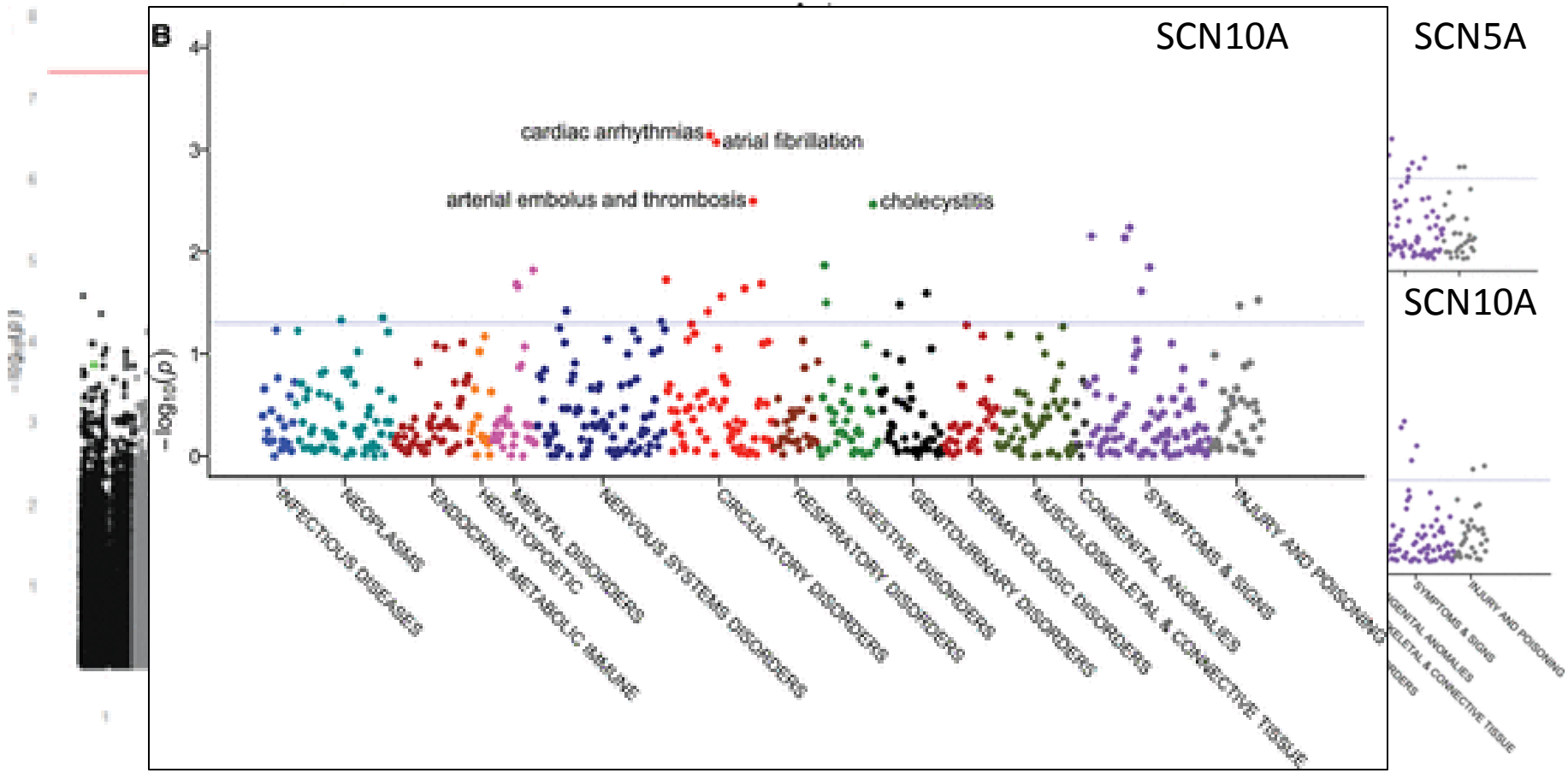


American Heart Association. Ritchie M D et al. Circulation. 2013;127:1377-1385

PheWAS: Phenome-Wide Association Study

Genome-wide association analysis of QRS duration using sex-adjusted linear regression

Phenome-wide association analysis of SCN5A/SCN10A using logistic regression



Ritchie M D et al. Circulation. 2013;127:1377-1385

PheWAS

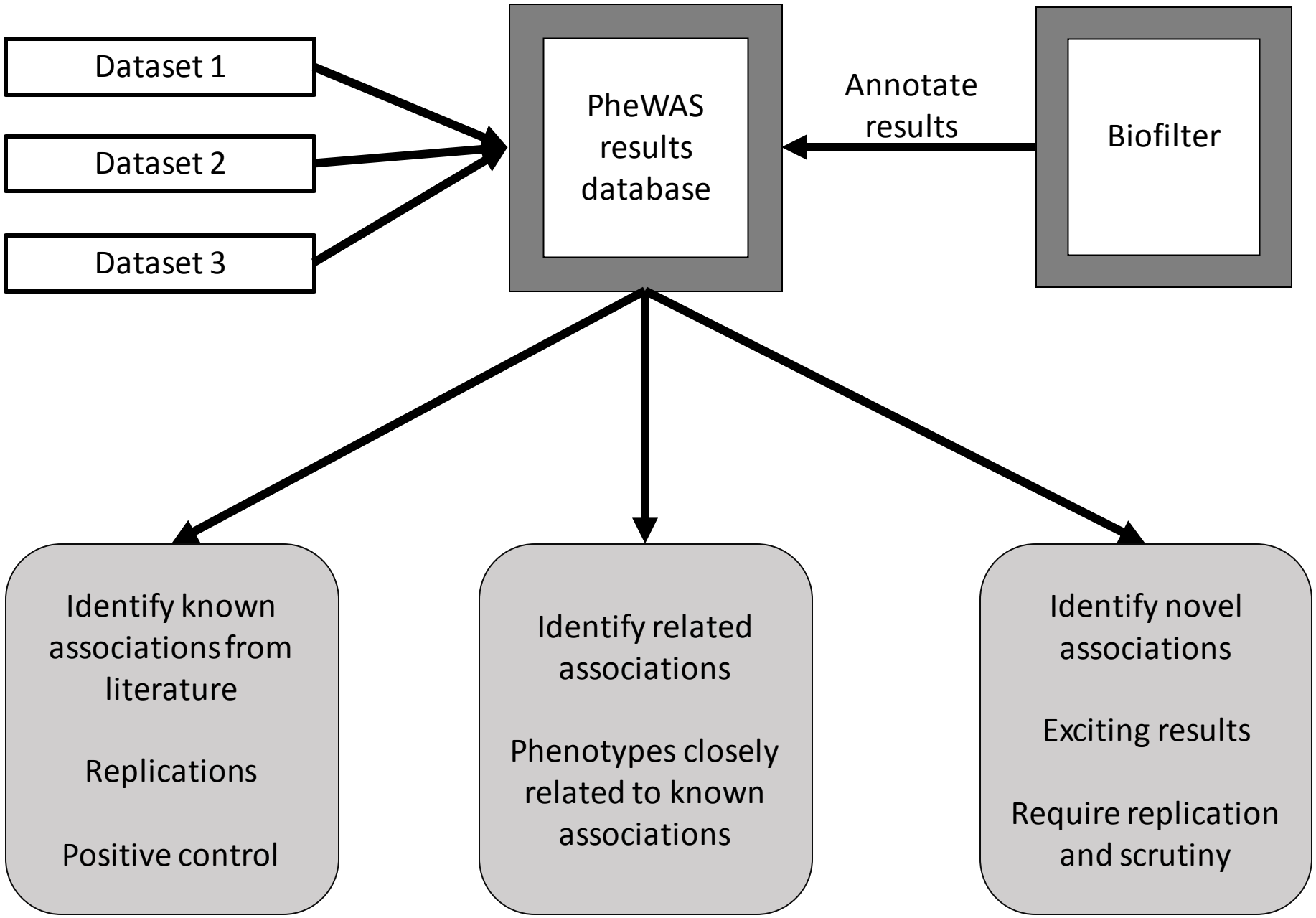
- Comprehensively calculating the association between:
  - Small to large number of SNPs
  - Diverse range of phenotypes from phenotypically rich datasets
- PheWAS has the potential to:
  - Provide novel mechanistic insights
  - Foster hypothesis generation
  - Identify pleiotropy
- PheWAS can be performed as part of:
  - Epidemiology cohort studies
  - Animal breeding studies
  - Prospective clinical trials
  - DNA Biobanks linked to electronic health records

Genetic Epidemiology 35 : 410–422 (2011)

The Use of Phenome-Wide Association Studies (PheWAS) for Exploration of Novel Genotype-Phenotype Relationships and Pleiotropy Discovery

S.A. Pendergrass,[1] K. Brown-Gentry,[1] S.M. Dudek,[1] E.S. Torstenson,[1] J.L. Ambite,[2] C.L. Avery,[3] S. Buyske,[4,5] C. Cai,[2] M.D. Fesinmeyer,[6] C. Haiman,[7] G. Heiss,[3] L.A. Hindorff,[8] C.-N. Hsu,[2] R.D. Jackson,[9] C. Kooperberg,[6] L. Le Marchand,[10] Y. Lin,[6] T.C. Matise,[5] L. Moreland,[11] K. Monroe,[7] A.P. Reiner,[6,12] R. Wallace,[13] L.R. Wilkens,[10] D.C. Crawford,[1,14] and M.D. Ritchie[1,14]*

- Analysis pipeline for PheWAS
- Establish phenotype classes for analysis
- Data visualization for synthesizing results
- Challenges in interpreting results
  - Multiple testing
  - Correlations between phenotypes (as well as LD)
  - Replication suggested to avoid false positive results


Dataset 1
Dataset 2
Dataset 3
PheWAS results database
Annotate results
Biofilter
Identify known associations from literature
Replications
Positive control
Identify related associations
Phenotypes closely related to known associations
Identify novel associations
Exciting results
Require replication and scrutiny

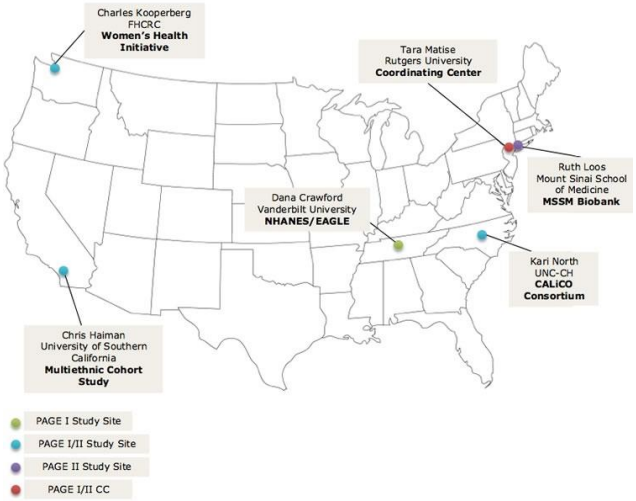
PheWAS in Epidemiology Cohorts

OPEN ACCESS Freely available online

PLOS | GENETICS

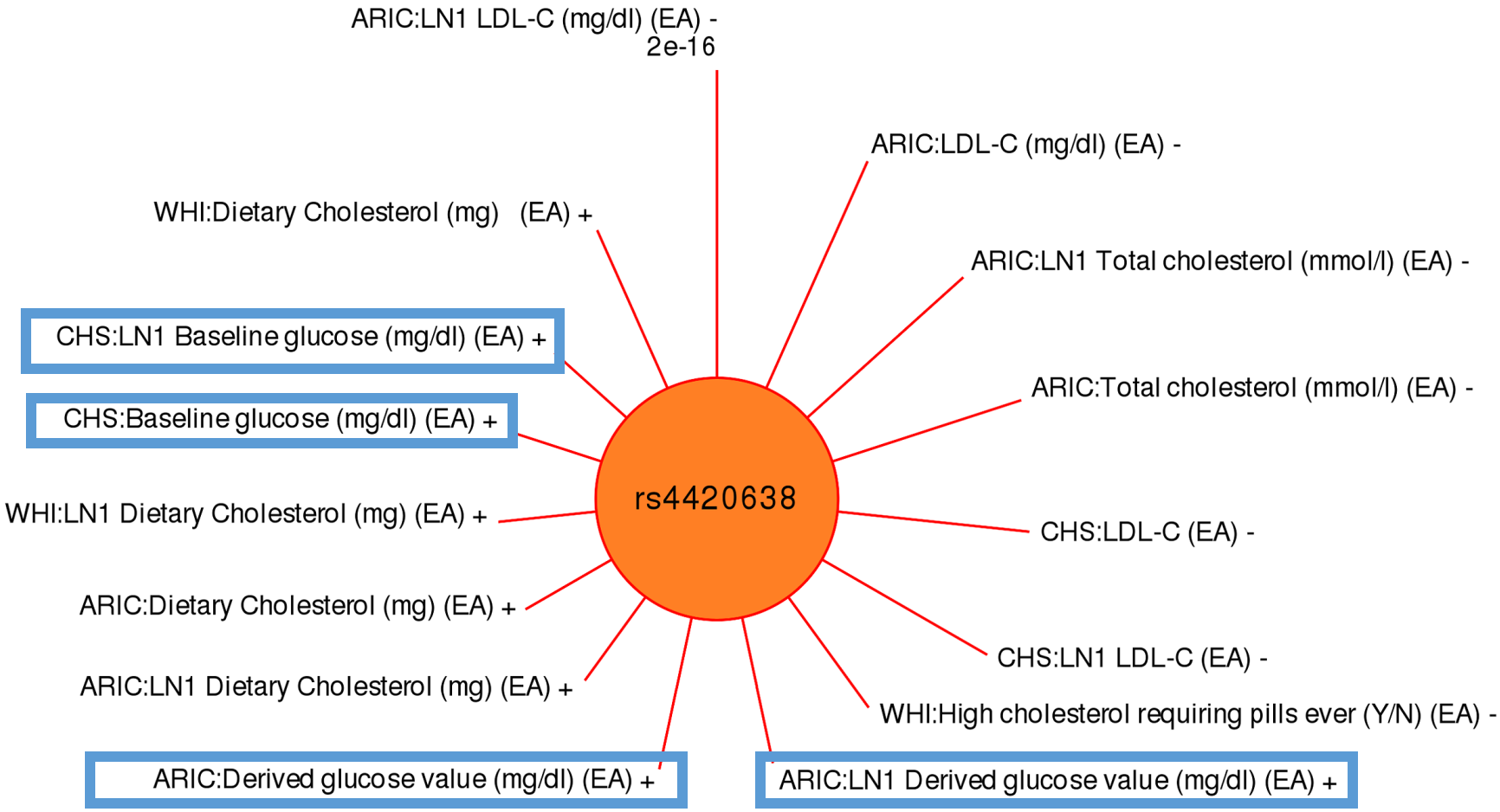
Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network

Sarah A. Pendergrass[1], Kristin Brown-Gentry[2], Scott Dudek[2], Alex Frase[1], Eric S. Torstenson[2], Robert Goodloe[2], Jose Luis Ambite[3], Christy L. Avery[4], Steve Buyske[5,6], Petra Bůžková[7], Ewa Deelman[3], Megan D. Fesinmeyer[8], Christopher A. Haiman[9], Gerardo Heiss[4], Lucia A. Hindorff[10], Chu-Nan Hsu[3], Rebecca D. Jackson[11], Charles Kooperberg[8], Loic Le Marchand[12], Yi Lin[8], Tara C. Matise[5], Kristine R. Monroe[9], Larry Moreland[13], Sungshim L. Park[12], Alex Reiner[8,14], Robert Wallace[15], Lynn R. Wilkens[12], Dana C. Crawford[2,16], Marylyn D. Ritchie[1]*



				Maximum Sample Size[3]				Minimum Sample Size[3]					
Study[1]	PMID[2]	Age Range	Sex	EA	AA	H	API	EA	AA	H	API	# SNPs[4]	# Phenotypes[5]
ARIC	[36]	45–64	M/F	11,068	4,007	NA	NA	17	7	NA	NA	69	612
CHS	[37]	65–100 at baseline	M/F	4,487	820	NA	NA	151	116	NA	NA	46	341
EAGLE	[38]	12–95	M/F	2,628	2,107	2,071	NA	7	16	15	NA	236	327
MEC	[39]	45–75	M/F	3,893	4,749	6,863	6,810	33	27	40	13	74	63
WHI	[40]	50–79 at baseline	F	13,334	4,274	2,023	927	14	5	7	5	94	3,363

APOE/APOC1/C1P1/C2/C4, rs4420638, Coded Allele A



SNP rs4420638 has previously been associated with LDL cholesterol levels, triglycerides, Alzheimer's disease, coronary artery disease

Pendergrass et al. PLoS Genet. 2013;9(1):e1003087

Plot generated by PheWAS-View

Diabetes. 2014 Sep;63(9):3149-58. doi: 10.2337/db13-1800. Epub 2014 Apr 10.

Pleiotropic Effects of Lipid Genes on Plasma Glucose, HbA1c, and HOMA-IR Levels.

Li N[1], van der Sijde MR[2]; LifeLines Cohort Study Group, Bakker SJ[3], Dullaart RP[4], van der Harst P[5], Gansevoort RT[3], Elbers CC[6], Wijmenga C[2], Snieder H[7], Hofker MH[8], Fu J[9].

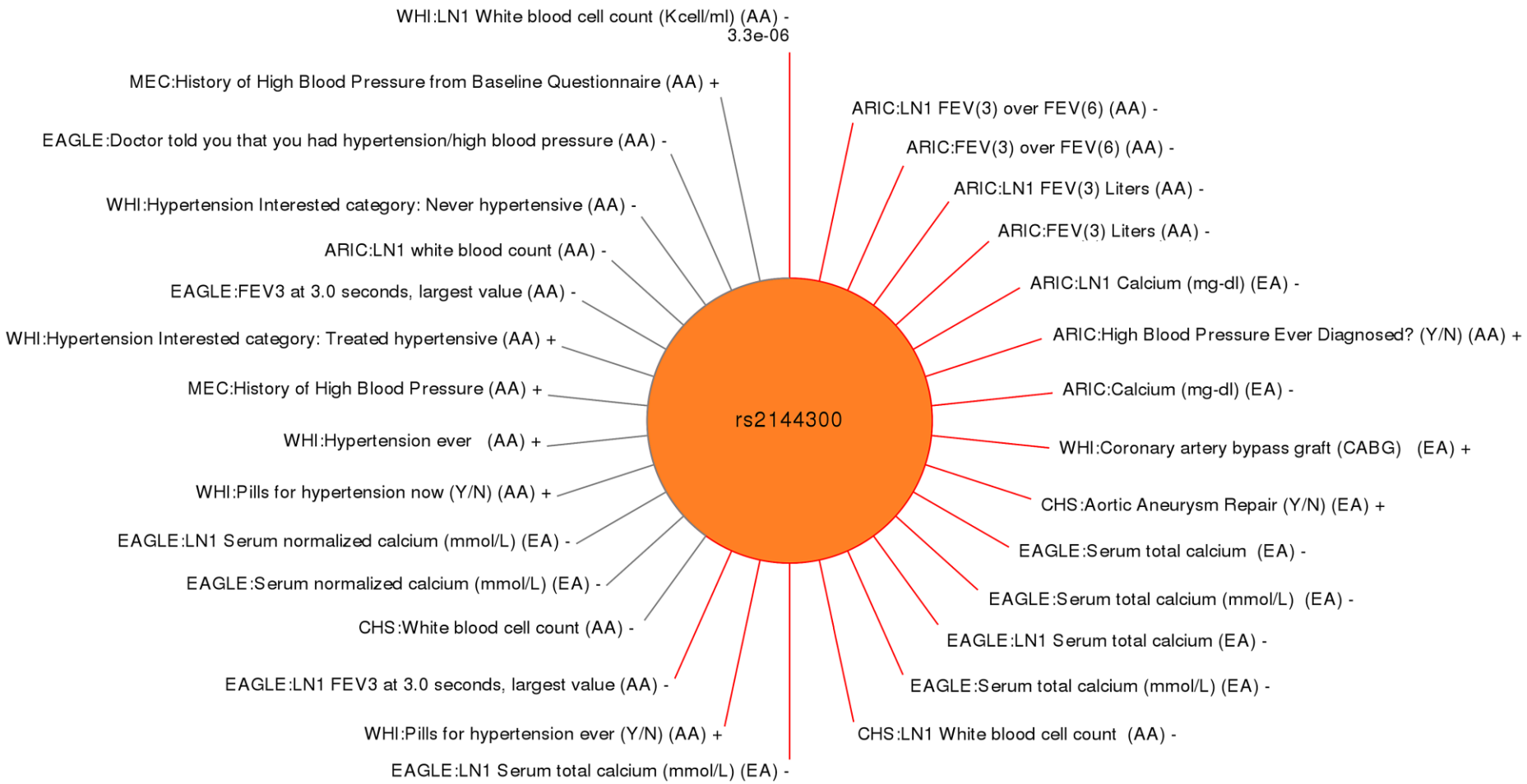
⊕ Author information

Abstract

Dyslipidemia is strongly associated with raised plasma glucose levels and insulin resistance (IR), and genome-wide association studies have identified 95 loci that explain a substantial proportion of the variance in blood lipids. However, the loci's effects on glucose-related traits are largely unknown. We have studied these lipid loci and tested their association collectively and individually with fasting plasma glucose (FPG), glycated hemoglobin (HbA1c), and IR in two independent cohorts: 10,995 subjects from LifeLines Cohort Study and 2,438 subjects from Prevention of Renal and Vascular Endstage Disease (PREVEND) study. In contrast to the positive relationship between dyslipidemia and glucose traits, the genetic predisposition to dyslipidemia showed a pleiotropic lowering effect on glucose traits. Specifically, the genetic risk score related to higher triglyceride level was correlated with lower levels of FPG ($P = 9.6 \times 10(-10)$ and $P = 0.03$ in LifeLines and PREVEND, respectively), HbA1c ($P = 4.2 \times 10(-7)$ in LifeLines), and HOMA of estimated IR ($P = 6.2 \times 10(-4)$ in PREVEND), after adjusting for blood lipid levels. At the single nucleotide polymorphism level, 15 lipid loci showed a pleiotropic association with glucose traits ($P < 0.01$), of which eight (CETP, MLXIPL, PLTP, GCKR, APOB, APOE-C1-C2, CYP7A1, and TIMD4) had opposite allelic directions of effect on dyslipidemia and glucose levels. Our findings suggest a complex genetic regulation and metabolic interplay between lipids and glucose.

PMID: 24722249 [PubMed - in process]

GALNT2, rs2144300, Coded Allele C



SNP published for HDL cholesterol and triglycerides in 2008

SNP published for circulating myeloperoxidase levels (serum) in 2013

Pendergrass et al. PLoS Genet. 2013;9(1):e1003087

Plot generated by PheWAS-View

PheWAS

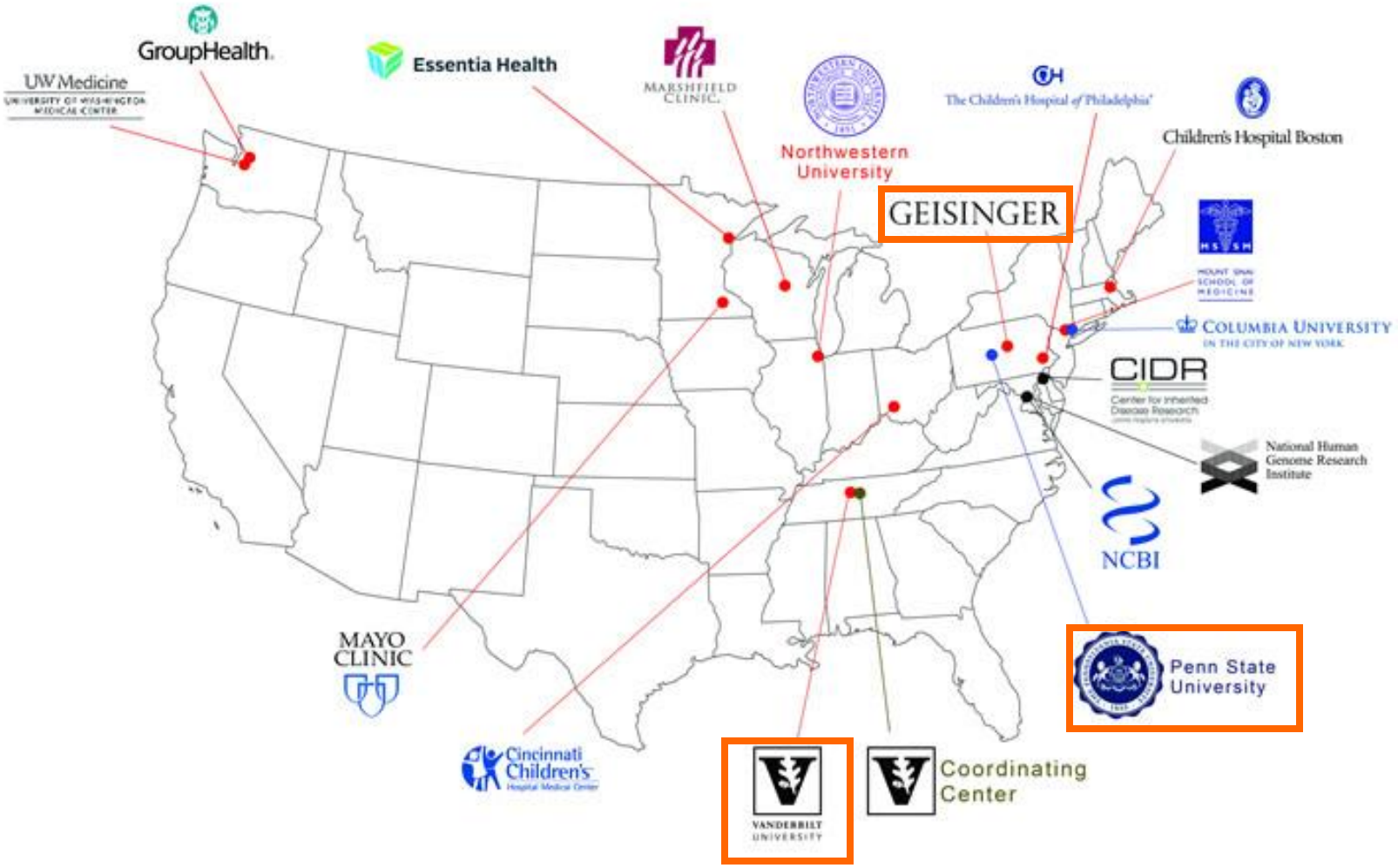
- Comprehensively calculating the association between:
  - Small to large number of SNPs
  - Diverse range of phenotypes from phenotypically rich datasets
- PheWAS has the potential to:
  - Provide novel mechanistic insights
  - Foster hypothesis generation
  - Identify pleiotropy
- PheWAS can be performed as part of:
  - Epidemiology cohort studies
  - Animal breeding studies
  - Prospective clinical trials
  - DNA Biobanks linked to electronic health records

Electronic Health Record Based PheWAS

- Electronic Health Record Data (EHR)
  - Billing codes
    - International Statistical Classification of Diseases and Related Health Problems (ICD-9 codes)
    - Can create "case/control" status based on the presence/absence of diagnosis ICD-9 codes
      - Two instances of an ICD-9 code = case, otherwise control
  - Clinically measured lab variables

Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Ritchie et al. *AJHG* 2010; *PMID: 20362271*

PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations Denny, Ritchie et al. *Bioinformatics* 2011; *PMID: 20335276*

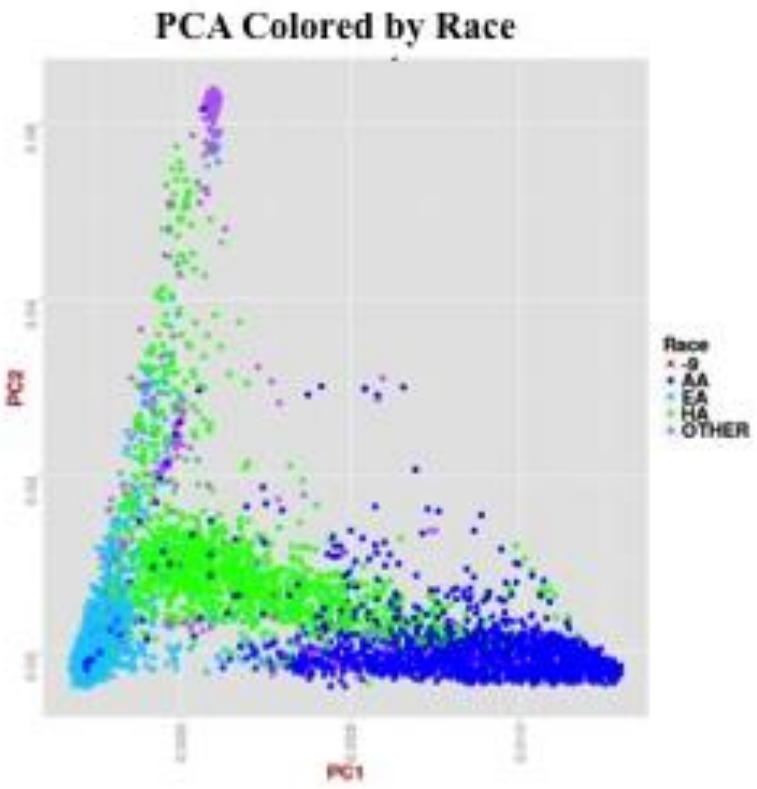


Part of the eMERGE network (Electronic Medical Records in Genomics)

- Vanderbilt University *BioVU*
- Geisinger Health System *MyCode*

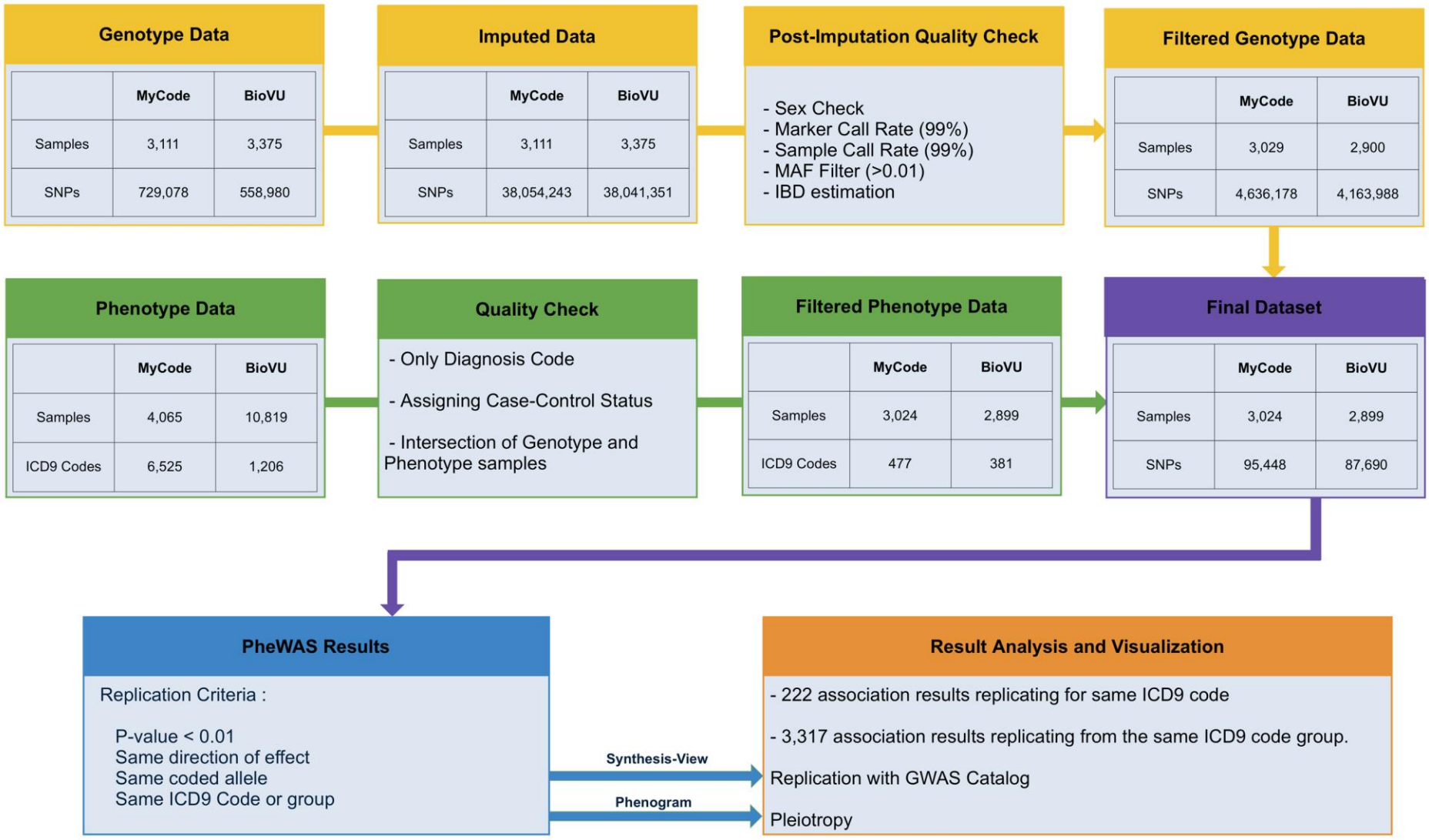
Electronic Health Record Based PheWAS

IMPUTE2 Imputed Data	
Dataset	# Genotyped Samples
Merged eMERGE-I 1M	2,634
Merged eMERGE-I 660	16,029
★ Geisinger	3,111
Group Health/ U Wash	731
Marshfield/Essentia/PSU	616
Mayo Clinic	3,121
Mount Sinai	6,290
Northwestern	2,951
★ Vanderbilt	7,616
CCHMC/BCH	5,346
CHOP	6,850
Total – All IMPUTE2 Imputed Samples	55,292

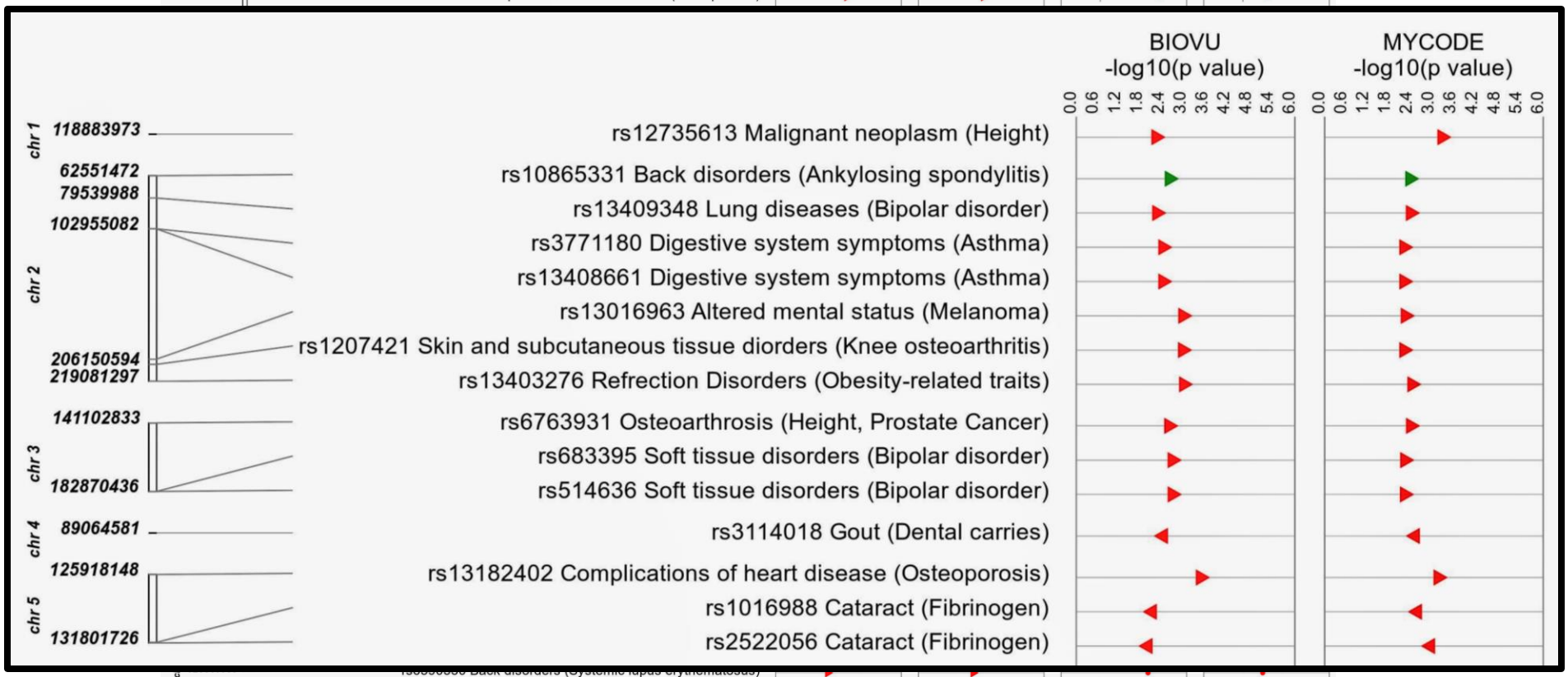


Verma S et al., Frontiers in Genetics, in revision

Electronic Health Record Based PheWAS - ImmunoPheWAS



Verma A, Okula A, Pendergrass SA, in preparation

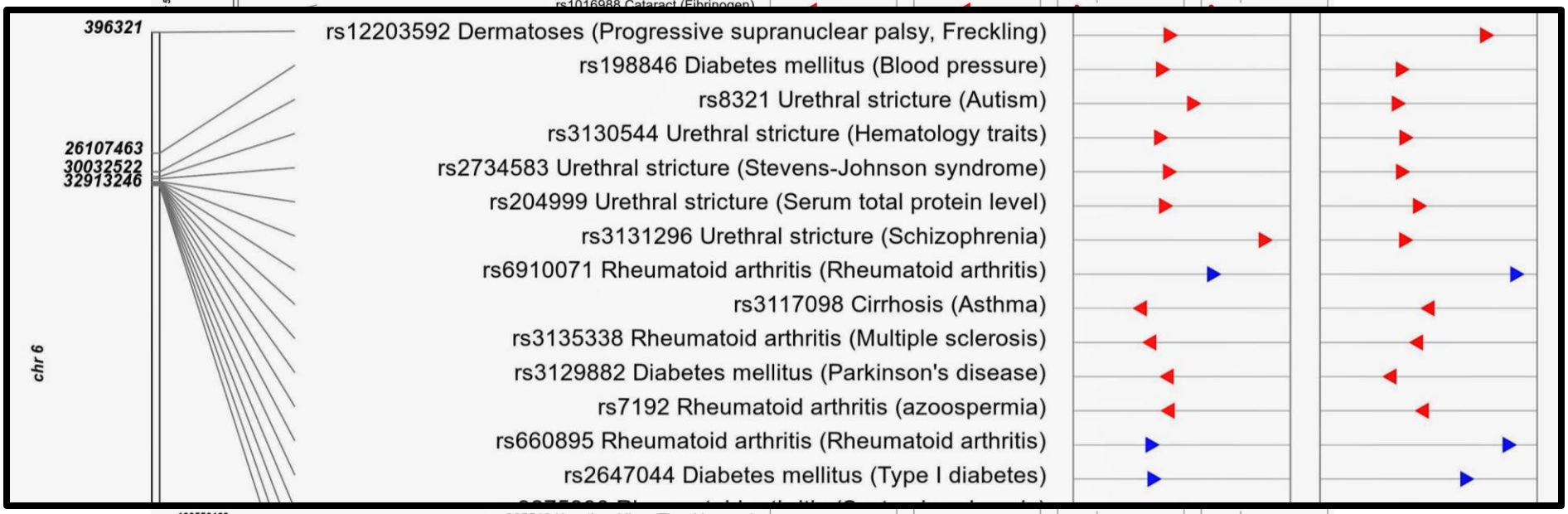


known

related

novel

Plot generated by Synthesis-View



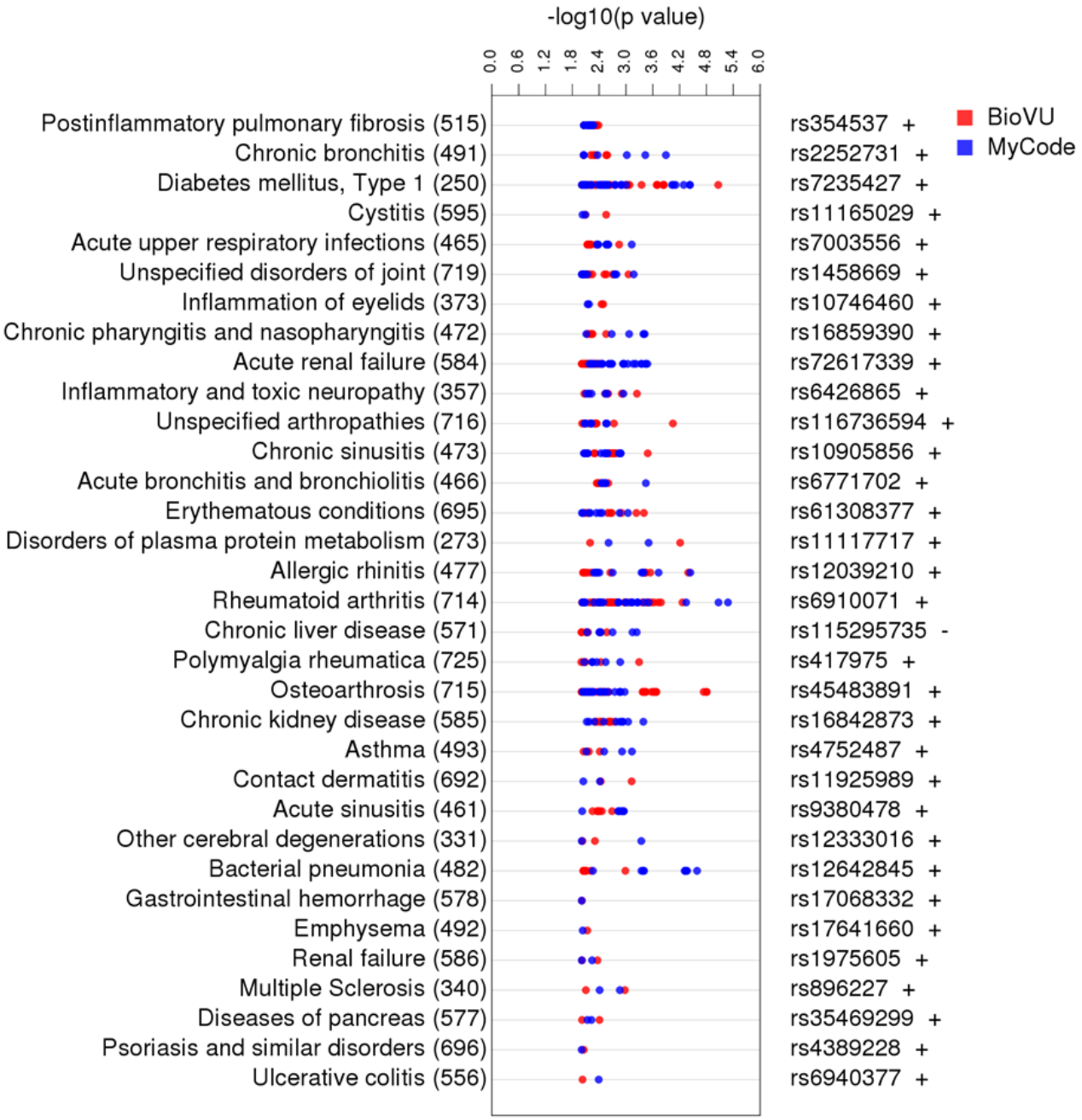
known

related

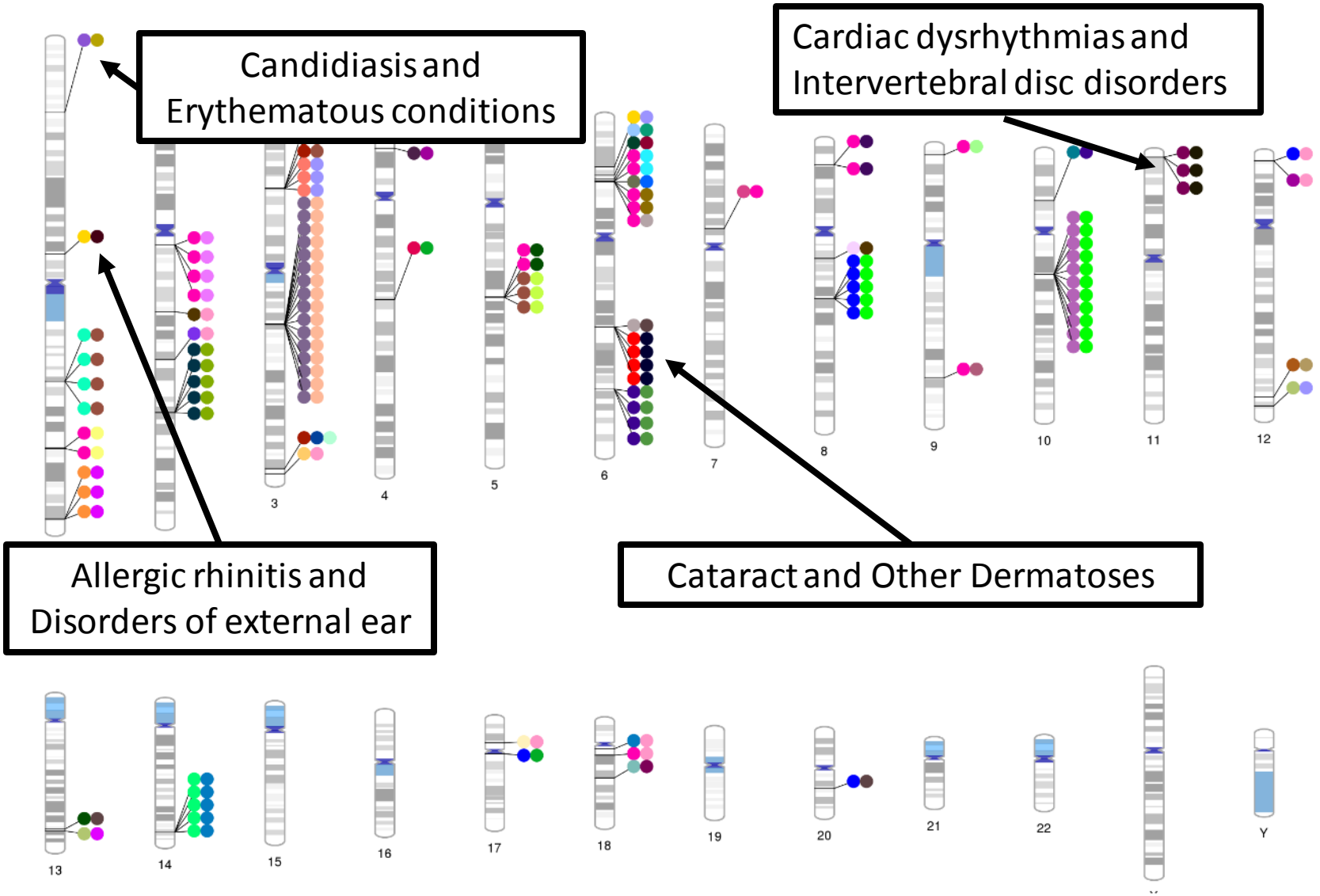
novel

Plot generated by Synthesis-View

Immune/Autoimmune Diagnosis Code Replications



Evidence of pleiotropy - ImmunoPheWAS



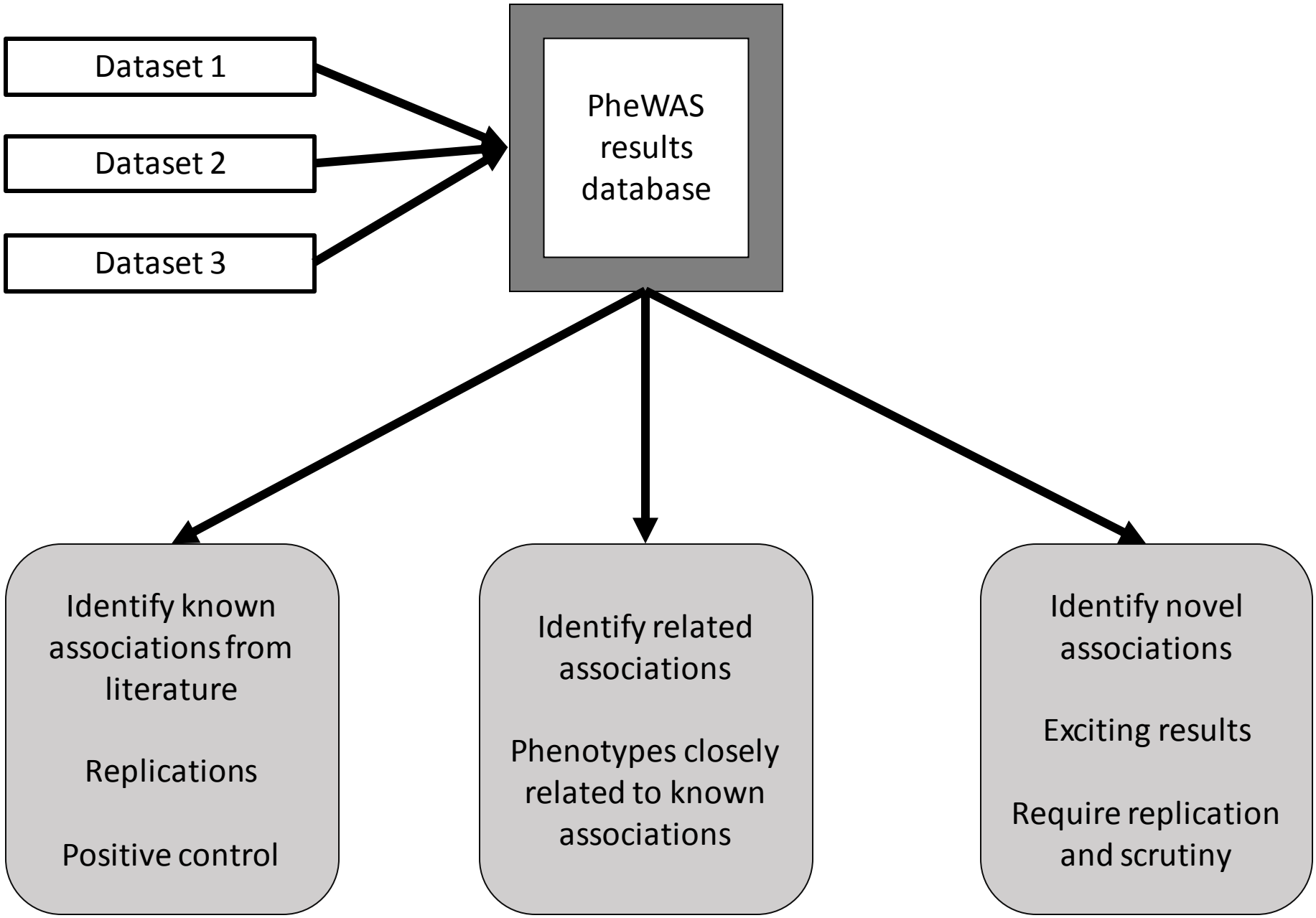
Verma A, Okula A, Pendergrass SA, in preparation

Plot generated by Phenogram

EHR-only PheWAS

- PheWAS does not REQUIRE genotype data
- EHR phenotypes/variables can be the independent variables → associate with one or more clinical outcomes of interest

Post-surgical Sepsis	Lab	Lab	Lab	Lab	Vital	Viral	Vital	ICD9	ICD9	CPT
Control										
Control										
Case										
Case										
Case										
Control										
Control										


Dataset 1
Dataset 2
Dataset 3
PheWAS results database
Identify known associations from literature
Replications
Positive control
Identify related associations
Phenotypes closely related to known associations
Identify novel associations
Exciting results
Require replication and scrutiny

EHR-only PheWAS

- PheWAS does not REQUIRE genotype data
- EHR phenotypes/variables can be the independent variables → associate with one or more clinical outcomes of interest
- Use statistical or machine learning methods to look for patterns of clinical variables that are predictive of adverse clinical outcomes
- Use one EHR as a training set and others as validation

Building pipelines and workflows

- Commercial and free software exists for setting up pipelines and workflows
- Important considerations:
  - Will I be running the same thing every time?
    - You do not want to do the same thing repeatedly without some automation in the pipeline.
  - How much human interaction is needed during the pipeline?
    - Will you need to implement stopping points to look at the data?
  - Who all will be using this pipeline?
    - Is it for one person, a lab, a unit, an institution?

Galaxy

- http://main.g2.bx.psu.edu/

Galaxy

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use the free public server

Get Galaxy

Install locally or in the cloud

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State,

http://www.taverna.org.uk/

Taverna

myGrid

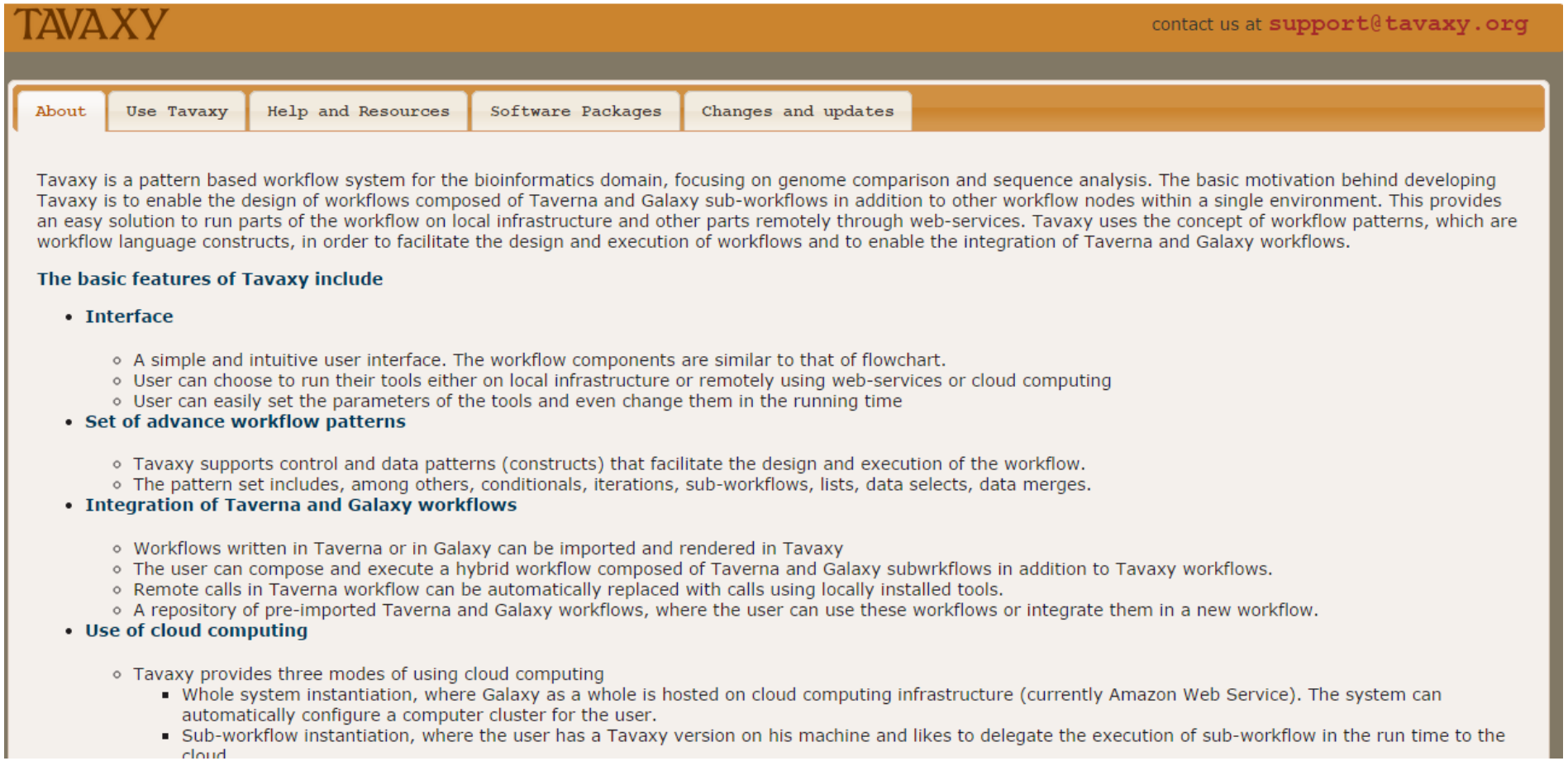
Google™ Custom Search

Introduction | Documentation | Download | Developers | Cite | Collaborations | News | About

Genome and gene expression

Taverna is being used by various projects and researchers for the analysis of genomes and gene expression. These include:

- Next Generation Sequencing using Taverna 2 Server on Amazon cloud
- TavernaPBS – next generation sequencing analysis using a computational cluster that uses a PBS queuing system and Taverna 2 Workbench
- Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR) – workflows to associate mouse genome and phenome
- Developmental Gene Expression Map (DGEMap)- analysis of human gene expression during development
- Graves disease – identification of genes responsible
- MicroArray analysis using R – statistical analysis of gene expression
- SIGENAE – development of workflows to analyse breeding animal data
- Trypanosomiasis – identification of genes responsible for sleeping sickness
- Williams-Beuren syndrome – automation and confirmation of gene characterization
- Integration of plant genome resources (PLANET)
- Annotation of genomes
- Shared Genomics



http://www.tavaxy.org/

Getting Genetics Done

Getting Things Done in Genetics & Bioinformatics Research

About GGD

Many resources (e.g. Nature Reviews Genetics) offer a 10,000-foot view of the current trends in the field, reviews of various technologies, and guidelines on how to effectively design, analyze, and interpret experiments in human genetics and bioinformatics research. By comparison very few resources focus on the mundane, yet critical know-how for those on the ground actually *doing* the science (i.e. grad students, postdocs, analysts, and junior faculty). Getting Genetics Done aims to fill that gap by featuring software, code snippets, literature of interest, workflow philosophy, and anything else that can boost productivity and simplify getting things done in human genetics research.

I highly encourage you to leave comments. You can comment on any post without having to register, so please share your thoughts with us and other readers. By its very nature, there may be much easier or better ways to do things than what is covered here, so feel free to share your own ideas.

All the written content is licensed under cc-by-sa and all the code here is open source (GPLv2). So feel free to wholesale copy, modify, share, or redistribute anything you find here, just include a link back to the site.

GettingGeneticsDone.blogspot.com can also be reached by going to GettingGeneticsDone.com. You can keep up with GGD by signing up for an e-mail whenever new content is posted, subscribing to the RSS feed, or following @genetics_blog on Twitter. If you would like to contribute, or if you have any topics, suggestions, praise, complaints, or anything else you'd like to share, please email me.

Thank you, and welcome!

Stephen Turner

SEARCH GGD

Search

THE GGD TEAM

Stephen Turner is an assistant professor of public health sciences and director of the Bioinformatics Core at the University of Virginia.

Will Bush is an assistant professor in the Vanderbilt University Center for Human Genetics Research

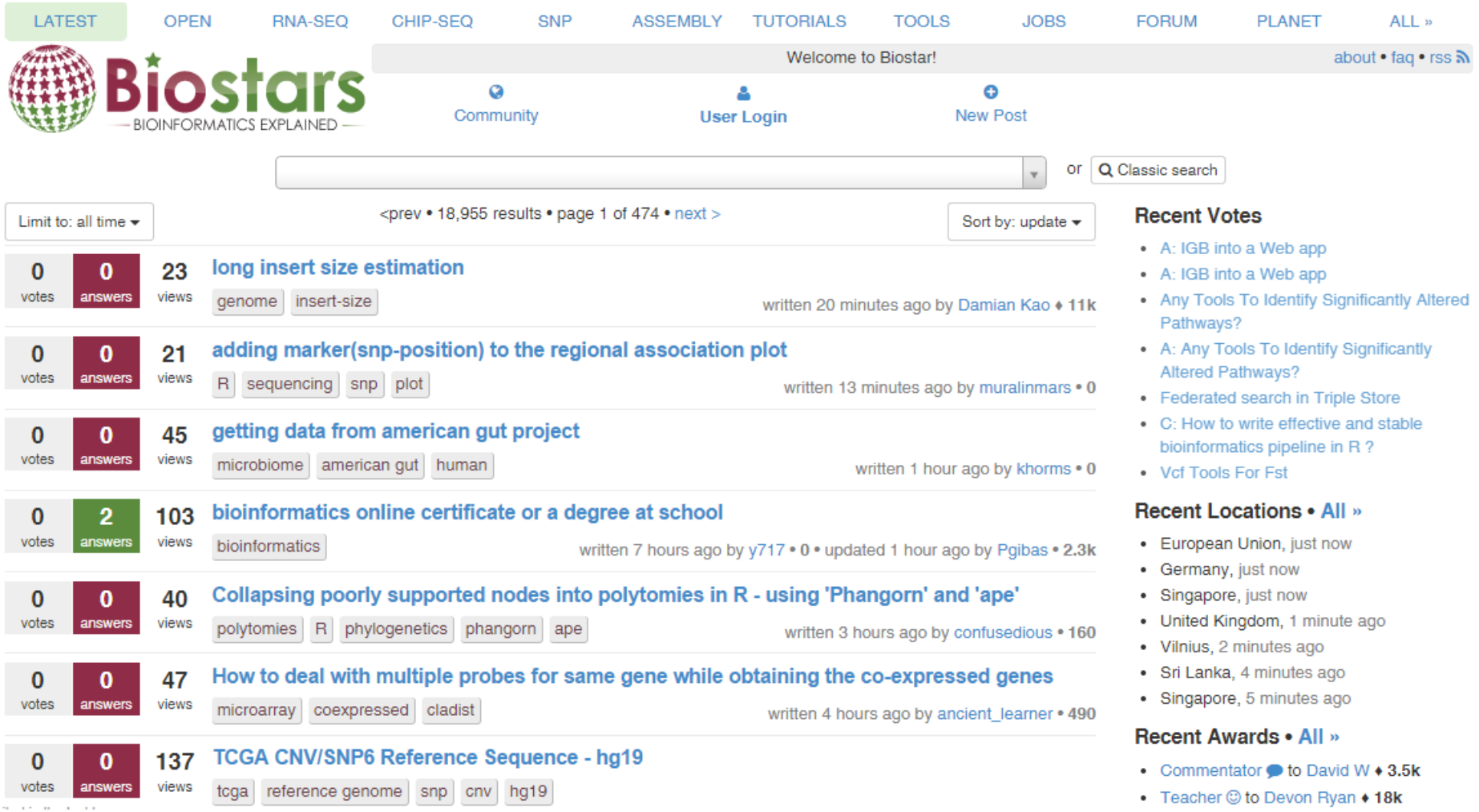
About GGD
@genetics_blog on Twitter
Subscribe via e-mail
Subscribe via RSS

ARCHIVE

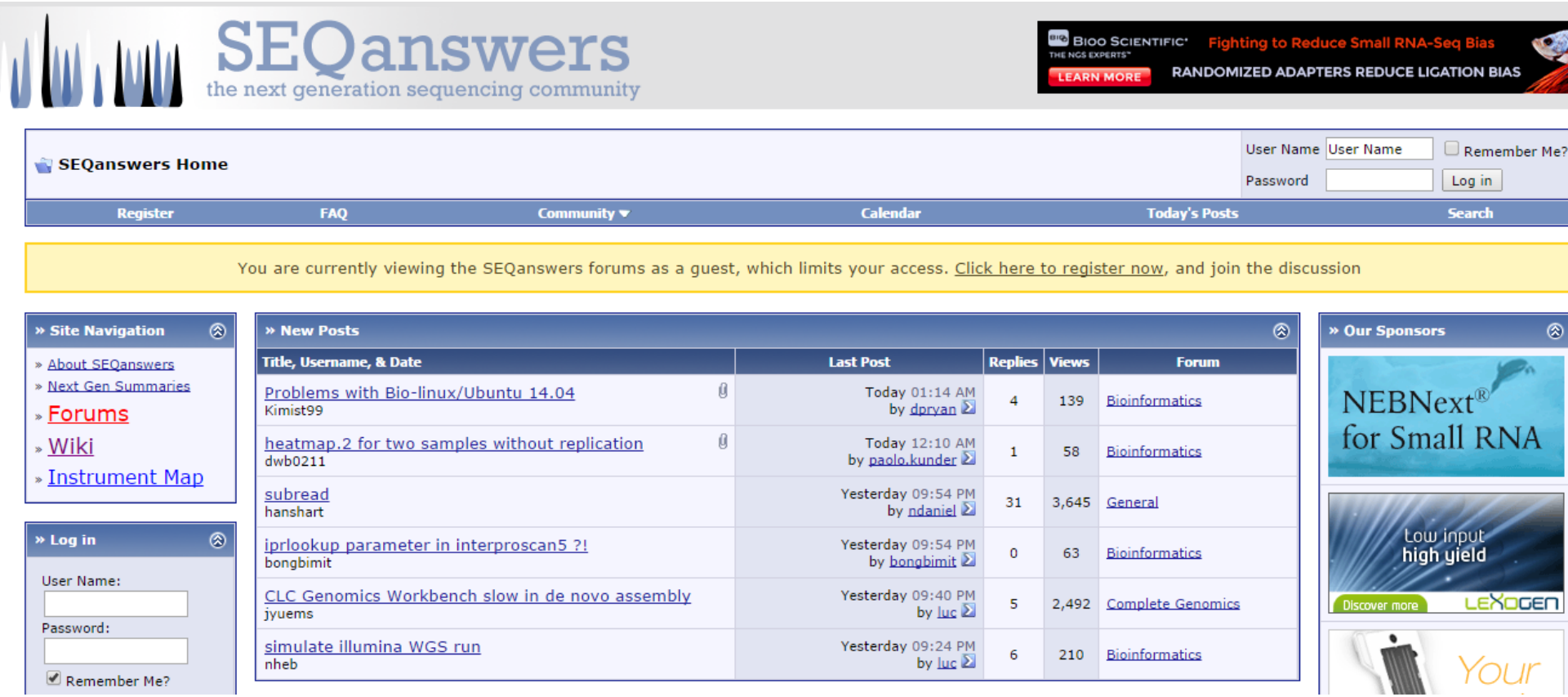
▼ 2014 (20)

▶ September (2)

http://gettinggeneticsdone.blogspot.com



https://www.biostars.org/



http://seqanswers.com/

Graduate Certificate in Applied Bioinformatics

REQUEST INFO APPLY NOW

World Campus was very inclusive —there were people constantly checking in with me. It was a very pleasant surprise!

Angelik Leslie

Bachelor of Arts in Law and Society

Lead the Application of Genomic Information

With ongoing public attention on the Human Genome Project, understanding, analyzing, and interpreting genomic data continue to be a focus across many disciplines. By merging biology, computer science, and information technology, bioinformatics allows us to combine mathematics and computers to gain a better understanding of biological processes.

Program Summary

Credits Required	11
Tuition per Credit	$930

< Back to Graduate Degree Programs

Admissions Requirements
Degree Requirements
Course Descriptions
Tuition and Fees
Career Opportunities
FAQ

Course Schedule
Our Community
The Experience

> Request Information
> Contact Us
> Apply Now

Bioinformatics

Advanced Academic Programs > Academics > Graduate Degree Programs > Bioinformatics

Click here to RSVP for the Nov. 5 Baltimore Open House

Click here to RSVP for the Nov. 11 Online Open House

Johns Hopkins University offers an innovative graduate degree program that prepares students for success in the field of bioinformatics. Drawing upon the strengths of the Johns Hopkins' Biotechnology and Engineering, the program creates a rigorous bioinformatics curriculum that brings together the computer science, biosciences and bioinformatics disciplines.

Quick Stats

Course Locations	Baltimore, MD; Rockville, MD; Online
Available 100% Online	Yes.
Entry Terms	Fall, Spring or Summer semester
Degree Requirements	11 courses

+ Request Info

+ Share This Page

Follow @JHUBiotech on Twitter.

Northeastern University
College of Science

Search

Bioinformatics – Master of Science

About the Program | CO-OP Internship | How to Apply | ALIGN Master's Degree Program | FAQs | Contact Information

Georgia Tech Bioinformatics
Interdisciplinary Graduate Program

Welcome | Prospective Students | Program Requirements | Research | People | News | Events | Employers | Atlanta | Login

UC San Diego

Bioinformatics Algorithms (Part 1)

This course is the first in a two-part series covering some of the algorithms underlying bioinformatics. It will cover some of the algorithms underlying the following fundamental topics in bioinformatics: assembling genomes, comparing DNA and protein sequences, finding regulatory motifs, analyzing genome rearrangements, identifying proteins, and many other topics.

About the Course

The sequencing of the human genome fueled a computational revolution in biology. As a result, modern biology produces as many new algorithms as any other fundamental realm of science. Accordingly, the newly formed links between computer science and biology affect the way we teach applied algorithms to computer scientists.

Genome sequencing is just one of hundreds of biological problems that have become inextricable from the computational methods required to solve them. In this course, we

Sessions

Oct 20th 2014 - Jan 12th 2015 ▾

Join for Free

Earn a Verified Certificate

Stanford University

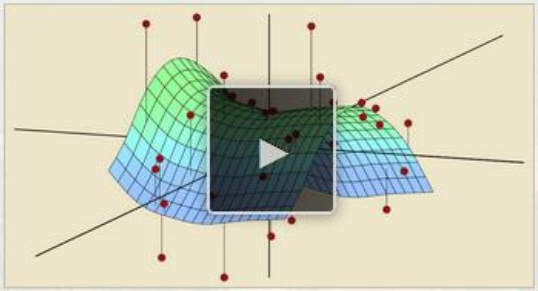
ABOUT

Register

Log in

Statistical Learning

REGISTER FOR STATLEARNING



How To Apply

Selection Process and Stipends

Travel

General Information

Campus Information

PROGRAMMING FOR BIOLOGY

October 13 - 28, 2014

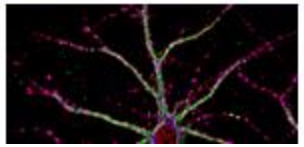
Application Deadline: July 15, 2014

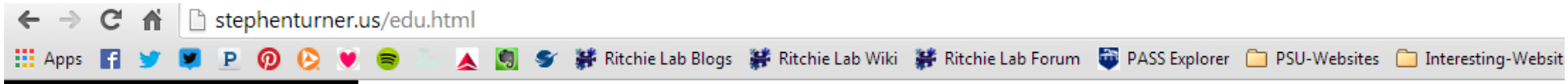
Instructor:
Simon Prochnik, *DOE Joint Genome Institute*

See the Roll of Honor - who's taken the course in the past

Web-based tools are no longer enough for today's biologist who needs to access and analyze large datasets from myriad sources in desparate formats. The need to design and program custom analysis pipelines is becoming ever more important as new technologies increase the already exponential rate at which biological data is generated. Designed for lab biologists with little or no programming experience, students will leave the two-week **Programming for Biology** course with the bioinformatics and scripting skills necessary to exploit this abundance of biological data. The prerequisite for the course is basic knowledge of UNIX; some scripting experience is helpful. Lectures and problem sets from previous years that cover this background material are available online and students can study this material before starting the course.

The course teaches Perl, a scripting language that is easy to learn and efficient to use. Perl also has a vast array of ready-





Stephen D. Turner, Ph.D.

- CV
- E-mail
- Twitter
- Blog
- Github
- Slides
- Edu

Bioinformatics Workshops & Training

Jump to: In-person courses, online material, or other resources.

In-person Courses

Last updated: 2014-03-18

Note that many of the courses shown below also have course materials from past years publicly available online.

Cold Spring Harbor Courses: meetings.cshl.edu/courses.html

Cold Spring Harbor has been offering advanced workshops and short courses in the life sciences for years. Relevant workshops includ Genetics of Complex Human Diseases, Statistical Methods for Functional Genomics, Programming for Biology, Computational and Co Unlike most of the others below, you won't find material from past years' CSHL courses available online.

Canadian Bioinformatics Workshops: bioinformatics.ca/workshops

Bioinformatics.ca through its Canadian Bioinformatics Workshops (CBW) series began offering one and two week short courses in bioi proteomics in 1999. The more recent workshops focus on training researchers using advanced high-throughput technologies on the lat computational biology to deal with the new data. Course material from past workshops is freely available online, including both audio/vi include microarray analysis, RNA-seq analysis, genome rearrangements, copy number alteration, pathway analysis, genome visualizat functional annotation, data analysis using R, statistics for metabolomics, and much more.

UC Davis Bioinformatics Training Program: training.bioinformatics.ucdavis.edu

The UC Davis Bioinformatics Training program offers several intensive short bootcamp workshops on RNA-seq, data analysis and visu a focus on Amazon's computing resources. They also offer a week-long Bioinformatics Short Course, covering in-depth the practical th next-generation sequencing techniques. Every course's documentation is freely available online, even if you didn't take the course.

MSU NGS Summer Course: http://bioinformatics.msu.edu/ngs-summer-course-2014

Other web resources…

- http://biocyc.org/
- http://metacyc.org/

Acknowledgements

Yuki Bradford, analyst
Scott Dudek, software developer
Alex Frase, software developer
Molly Hall, PhD student
Dokyoon Kim, Post-doc
Ruowang Li, PhD student
Anastasia Lucas, undergrad researcher
Anna Okula, PhD student
Donna Panasiti, administrative asst.
Sarah Pendergrass, Research Scientist
Suzy Unger, program coordinator
Anurag Verma, bioinformatics programmer
Shefali Verma, bioinformatics analyst
John Wallace, software developer