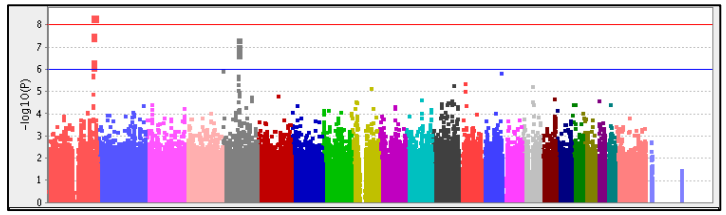
PheWAS Implementation and methods

Robert Carroll, PhD
Department of Biomedical Informatics
Robert.Carroll@vanderbilt.edu

Introduction to PheWAS

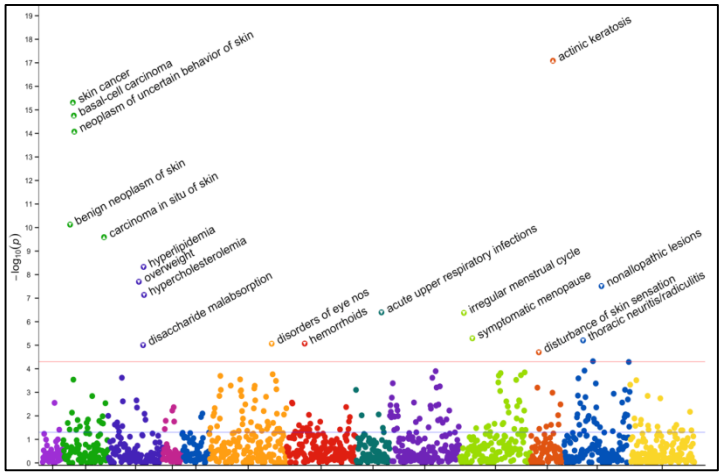
The genome-wide association study

Target phenotype → association P value



The phenome-wide association study

Target genotype → association P value



PheWAS requirement: A large cohort of patients with genotype data and many diagnoses

Methods of grouping doing a PheWAS

- EMR data
  - Billing codes:
    - **International Classification of Disease** (ICD) – WHO standard for diagnoses, signs, and symptoms
    - Current Procedural Terminology (CPT) – procedure codes
  - Text
    - NLP
    - N-grams
- Observational cohort data
  - Must have a lot of phenotypes; most have a focus
  - Framingham, NHANES, 23&Me are good targets

"PheWAS" – Phenome-wide association study

Genotypes

PheWAS analysis

(~1,000 phenotypes, selected SNPs)

Electronic Medical Record → Phenotype mapping → ~700-1600 Clinical phenotypes (& controls)

VanderbiltBioVU

The *eMERGE* Network
electronic Medical Records & Genomics
A consortium of biorepositories linked to electronic medical records data for conducting genomic studies

DNA Biobanks linked to medical records

Denny et al. *Bioinformatics.* 2010

ICD codes

- International Classification of Disease (ICD)
- At Vanderbilt we currently use ICD-10 CM
  - Transitioned completely in Oct 2015
  - Much of the world uses the similar ICD-10
- The large majority of our historical data is in ICD-9 CM
  - Currently many systems still double code data in ICD-9 and ICD-10
- Diagnostic codes:
  - ICD-9-CM: ~13,500
  - ICD-10: ~68,000

ICD9 codes

- 3-digit codes (000-999): diagnoses, signs, symptoms
- 2-digit codes (00-99): procedures
- V-codes and E-codes

Grouping	Examples	Count
Chapter	390-459.99 DISEASES OF THE CIRCULATORY SYSTEM	20
Section	401-405.99 HYPERTENSIVE DISEASE 390-392.99 ACUTE RHEUMATIC FEVER	120
Category (3-digit)	401 Essential Hypertension 402 Hypertensive heart disease	900+
Fully-specified (3-5 digits)	401.9 Benign essential hypertension 402.11 Benign hypertensive heart disease with heart failure	~13,500

ICD10 codes

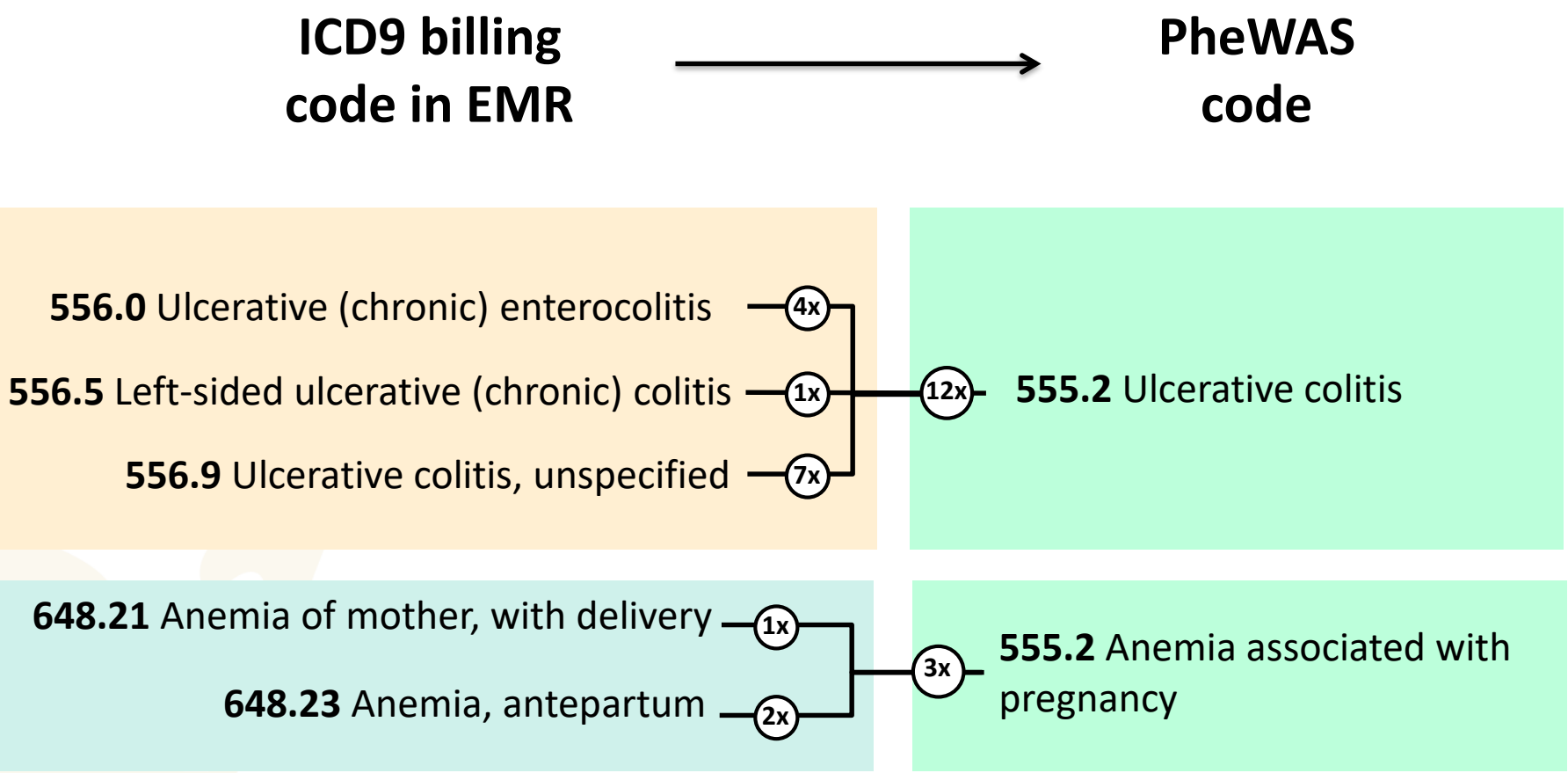
- Start with a character
  - No overlap with ICD9 E and V codes if properly specified
- 21 Chapters
  - They join and cross letter codes
    - CHAPTER II - Neoplasms (C00-D49)
    - CHAPTER III - Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)
- There exist billing maps between ICD9 and ICD10
- Phecode map development is in progress

The problem with billing codes

- False positives
  - Diagnoses evolve over time -- physicians may initially bill for suspected diagnoses that later are determined to be incorrect
  - Wrong code entered (easier to find or remember)
  - Physicians may bill for a different condition if it pays for a given treatment
    - psoriatic arthritis and rheumatoid arthritis
- False negatives:
  - Outpatient billing limited to 4 diagnoses/visit
  - Outpatient billing done by physicians (e.g., takes too long to find the unknown ICD9)
  - Inpatient billing done by professional coders:
    - omit codes that don't pay well
    - can only code problems actually explicitly mentioned in documentation

EMR Phenotyping with ICD-9 billing codes

Phenome-wide association study (PheWAS)

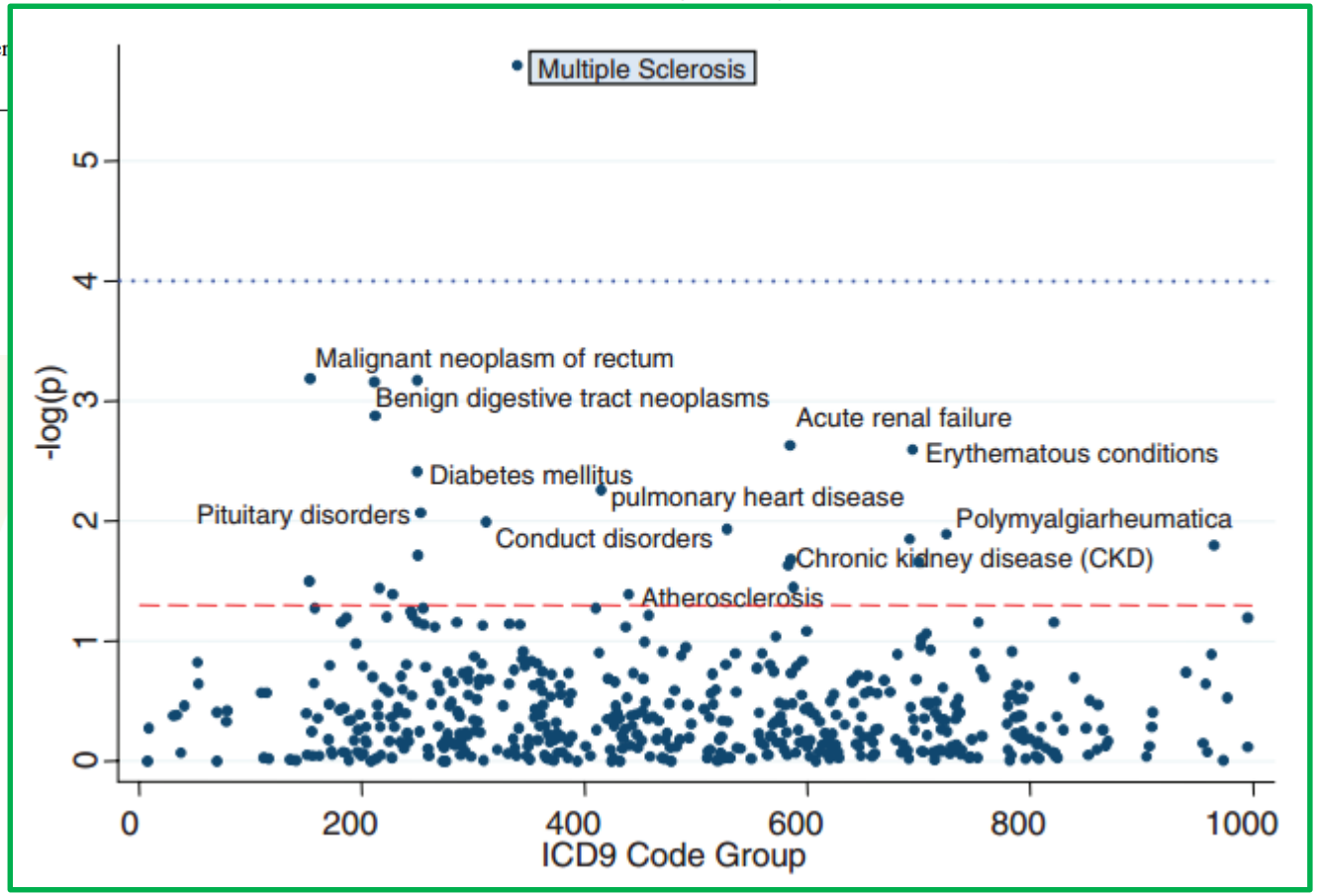


Original PheWAS

Table 2. Diseases previously associated with the five SNP studied and current PheWAS ORs

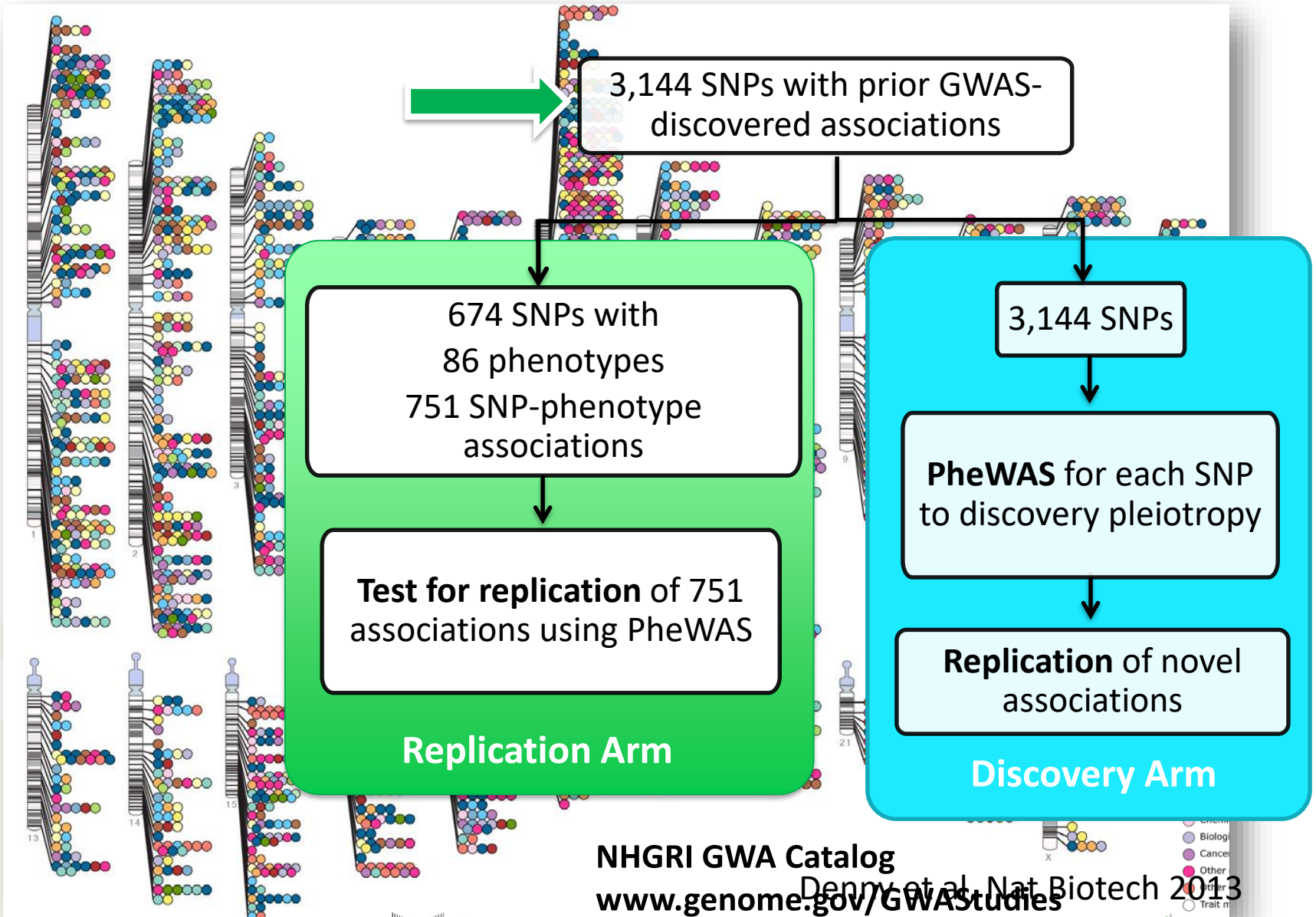
SNP	Gene/region	Disease	Cases	Previous OR	PheWAS *P*-value	PheWAS OR
rs3135388	DRB1*1501	MS	89	1.99[a]	2.77×10^{-6}	2.24 (1.56–3.16)
		SLE	141	2.06[b]	0.51	1.13 (0.79–1.58)
rs17234657	Chr. 5	CD	200	1.54[c]	0.00080	1.57 (1.19–2.04)
rs2200733	Chr. 4q25	AF and flutter	606	1.75[d]	0.14	1.15 (0.95–1.39)
rs1333049	Chr. 9p21	CAD				
		Carotid ather				
rs6457620	Chr. 6	RA[g]				

N = 6,005



Denny et al., Bioinformatics 2010

PheWAS of “all” NHGRI GWAS Catalog SNPs



Denny et al. Nat Biotech 2013

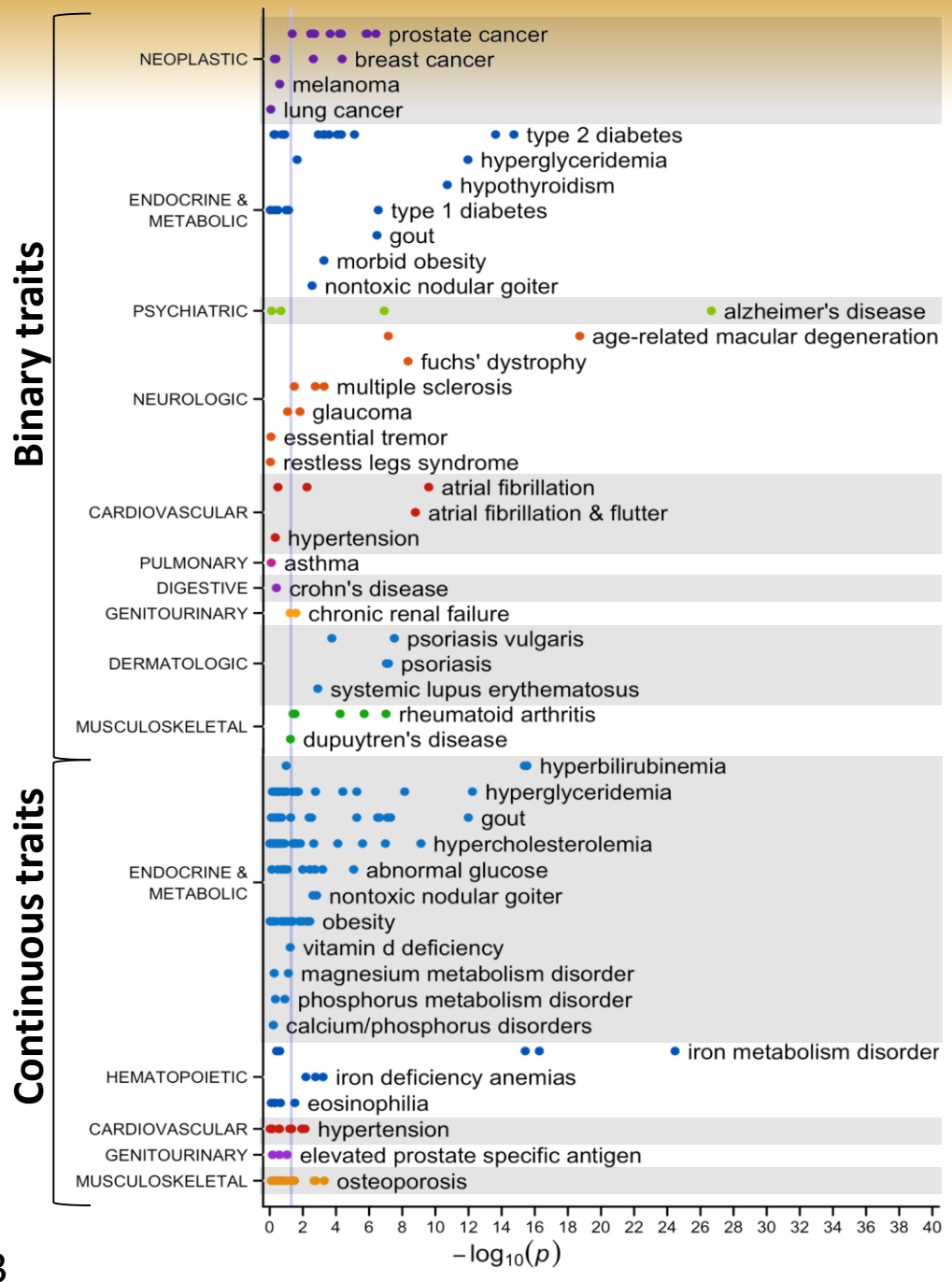
PheWAS Population

- 13,835 European-Ancestry individuals from 5 eMERGE sites with available GWAS data
- 2,080,550 unique dates of interaction with the EMR
- Mean follow-up of 15.7 ± 10.3 years

Replications of NHGRI GWAS associations via PheWAS

P-value for replication:

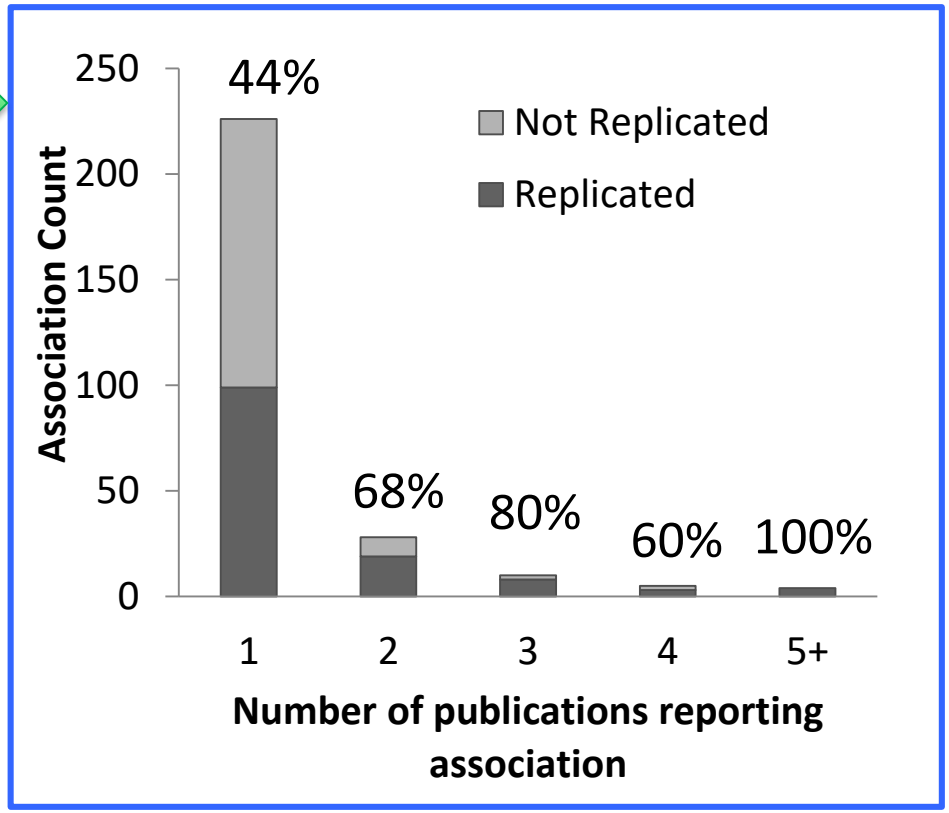
- All - 210/751: 2×10^{-98}
- Powered - 51/77: 3×10^{-47}



Nat Biotech 2013

Factors associated with replication

- Number of prior publications
- Exactness of phenotype match
- SNP location/functional status NOT associated

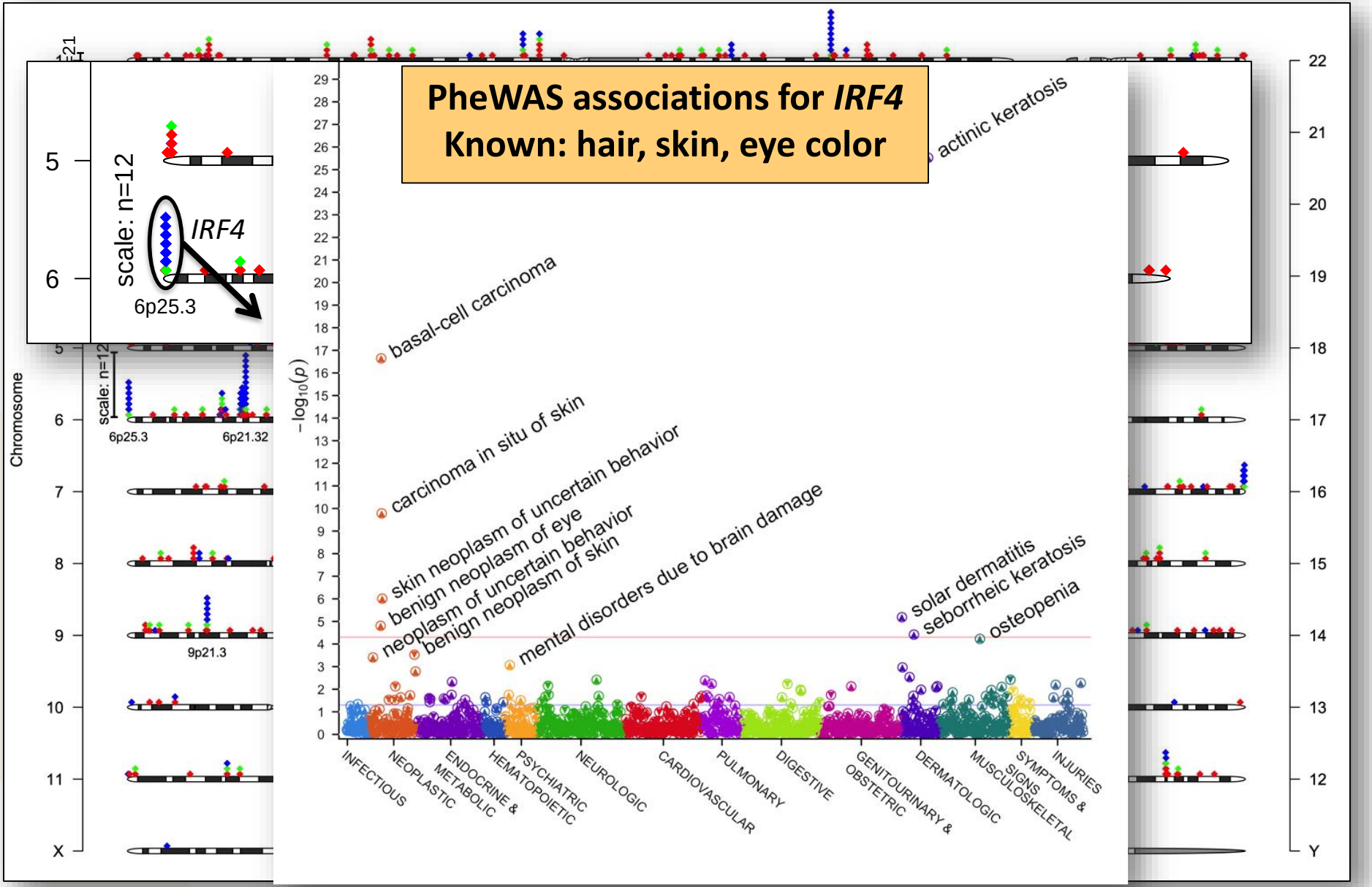


Denny et al, Nat Biotech 2013

PheWAS of all GWAS "hits"

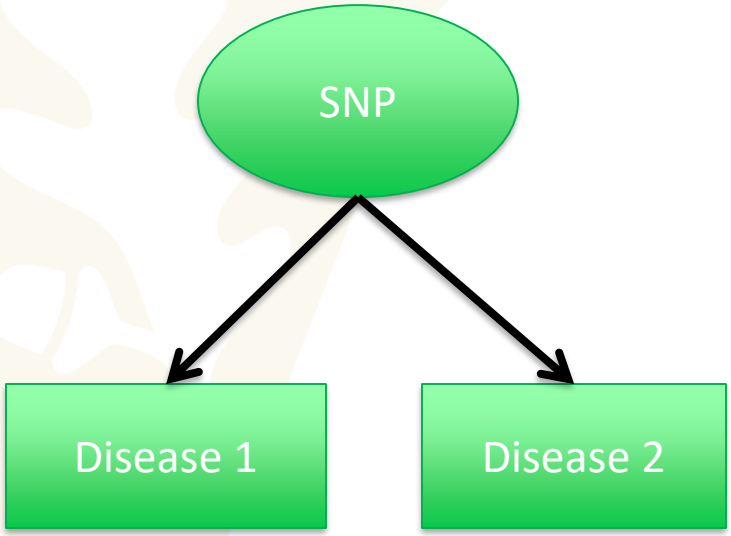
Each dot=one phenotype

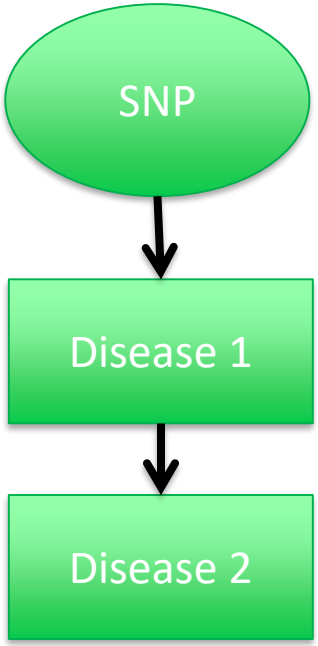
- GWA catalog association only
- GWA catalog association replicated by PheWAS
- New association found by PheWAS



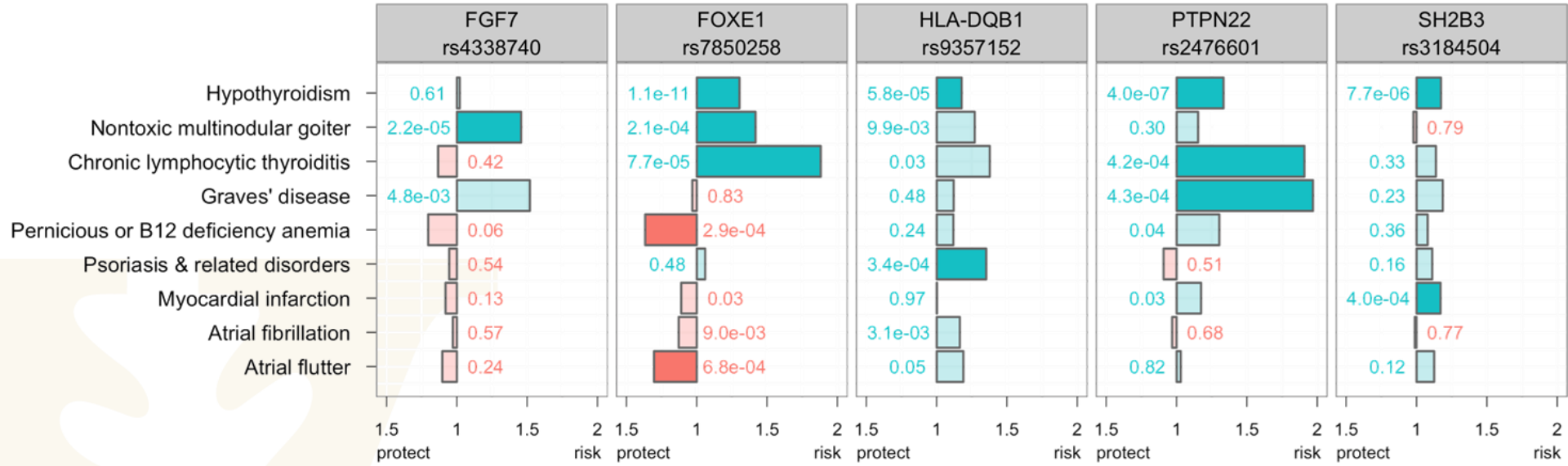
Genetics and Risk

- Pleiotropy: Single genetic changes may have multiple phenotypic effects.
- A single disease may have many associated genetic variants.
- A disease may have many comorbid diseases



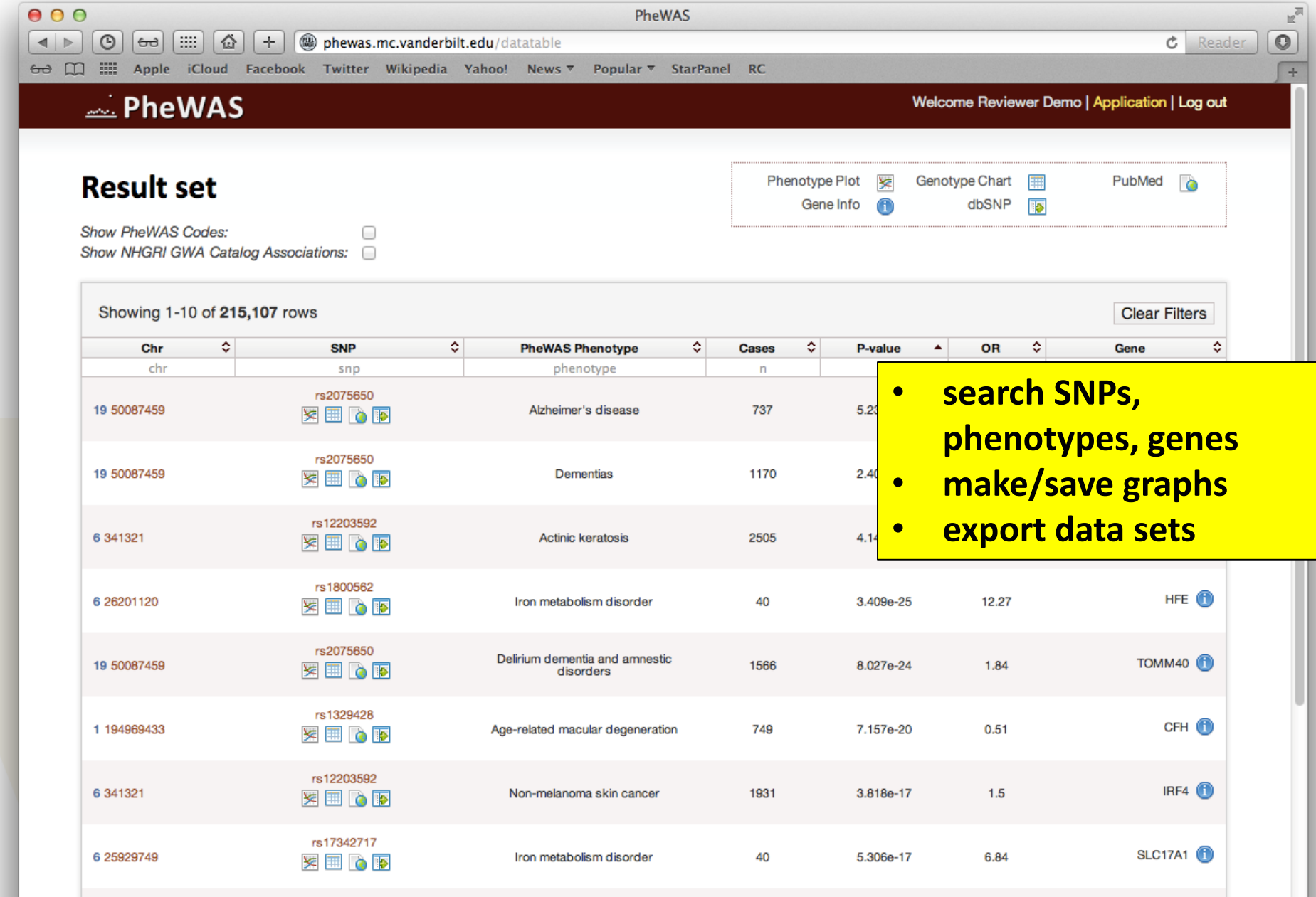


Pleiotropy in Thyroid Diseases



phewascatalog.org

PheWAS results for >3000 SNPs identified in GWAS studies



- **search SNPs, phenotypes, genes**
- **make/save graphs**
- **export data sets**

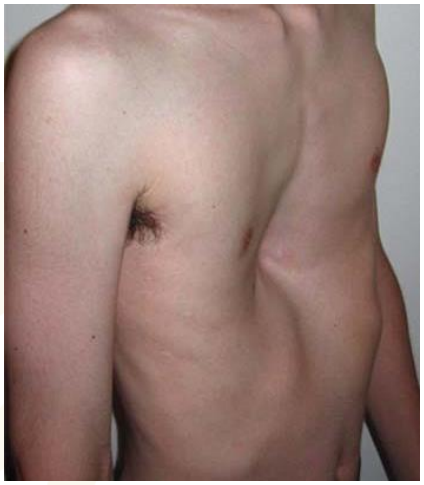
Other uses and types of phenome scanning

- Other structured data
  - Labs
  - Vitals
  - Report measurements (ECGs, echos, etc.)
- Natural language data
- Environmental (“E-WAS”)

Using PheWAS to explore a physical exam finding

Pectus excavatum

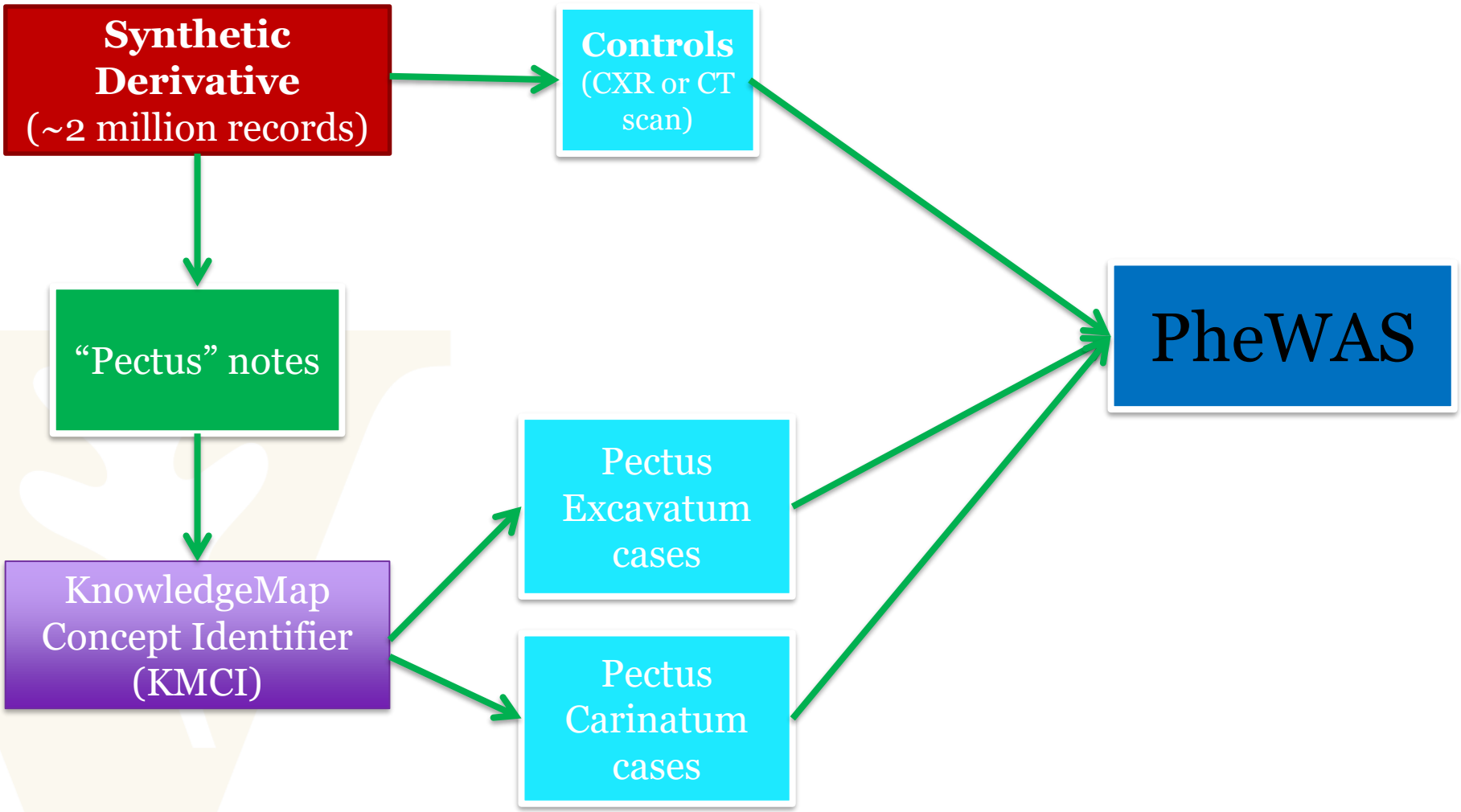
Cardiac issues if extreme



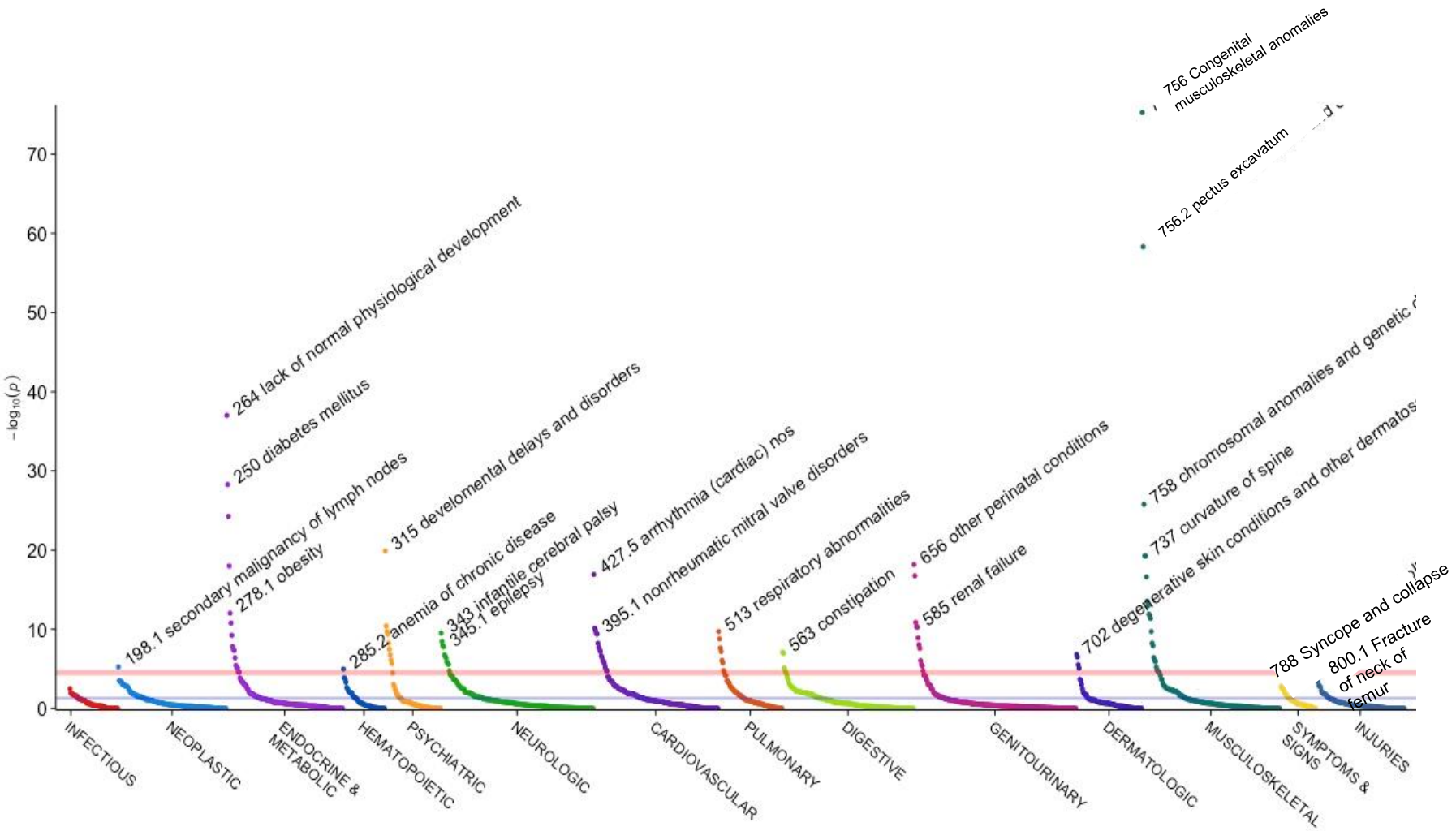
Pectus carinatum

Thought incidental

Study Design

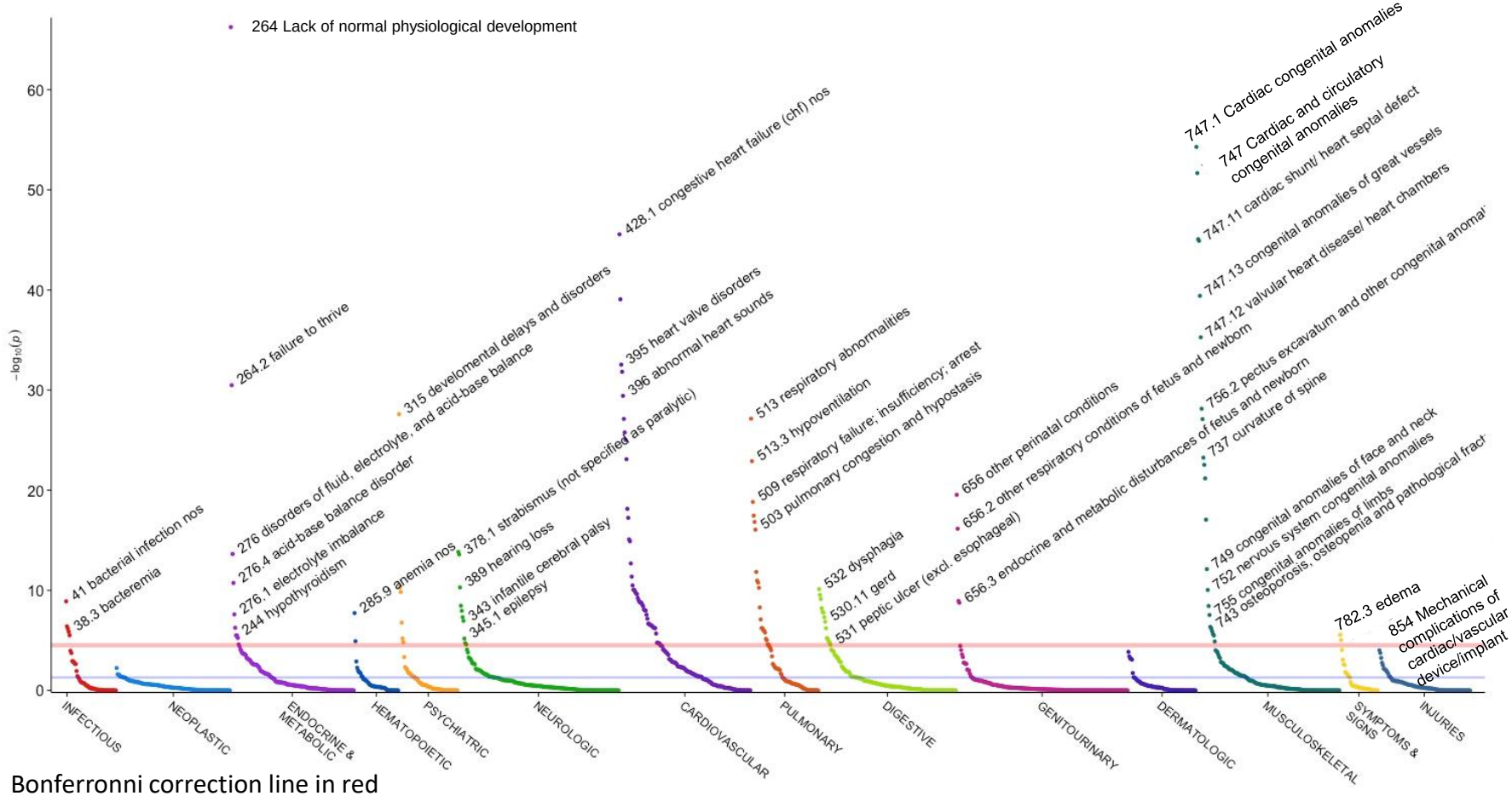


Pectus Excavatum

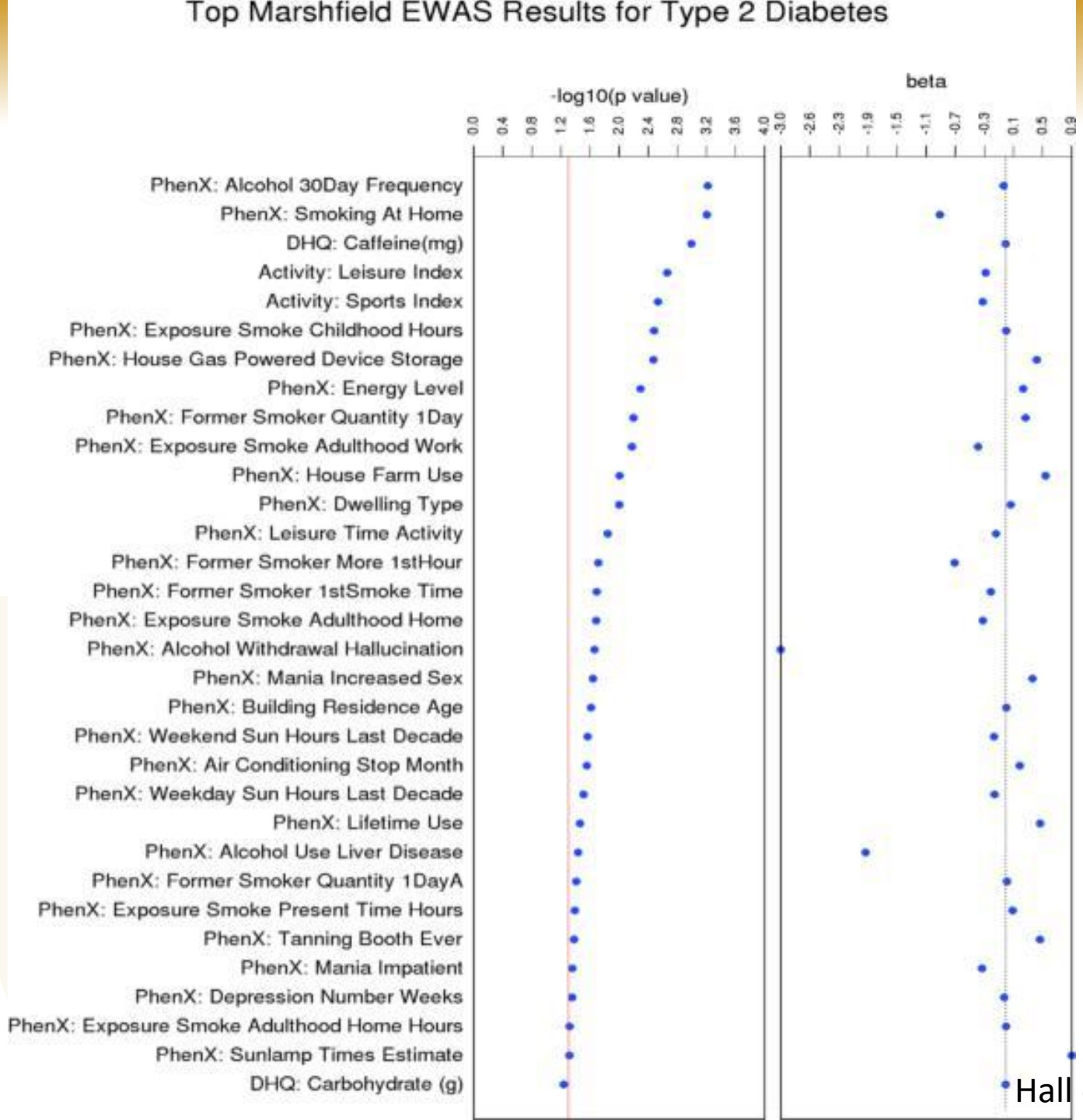


Bonferronni correction line in red

Pectus Carinatum



Bonferronni correction line in red



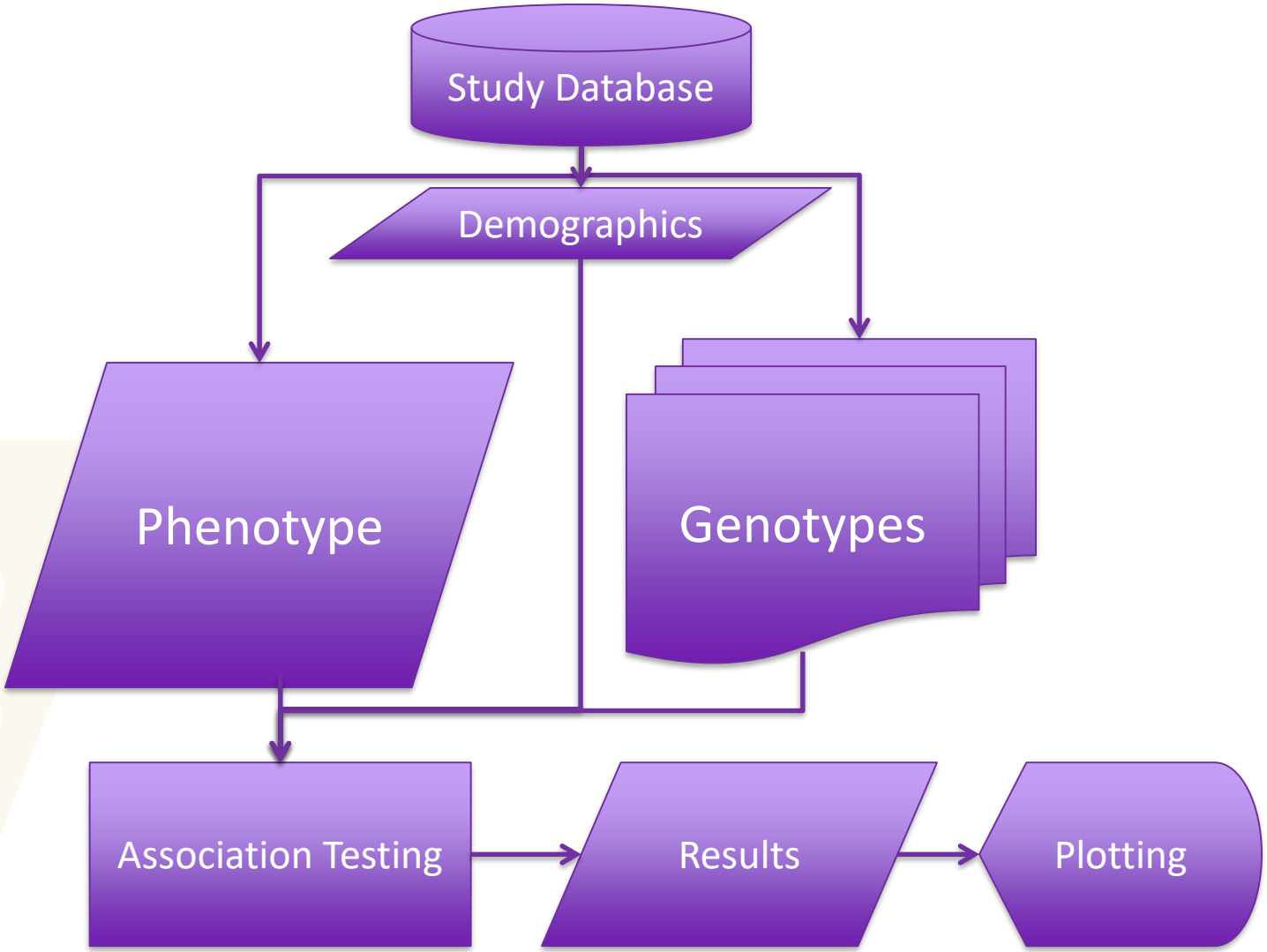
Hall et al, Pac Bio Symp 2014

Summary

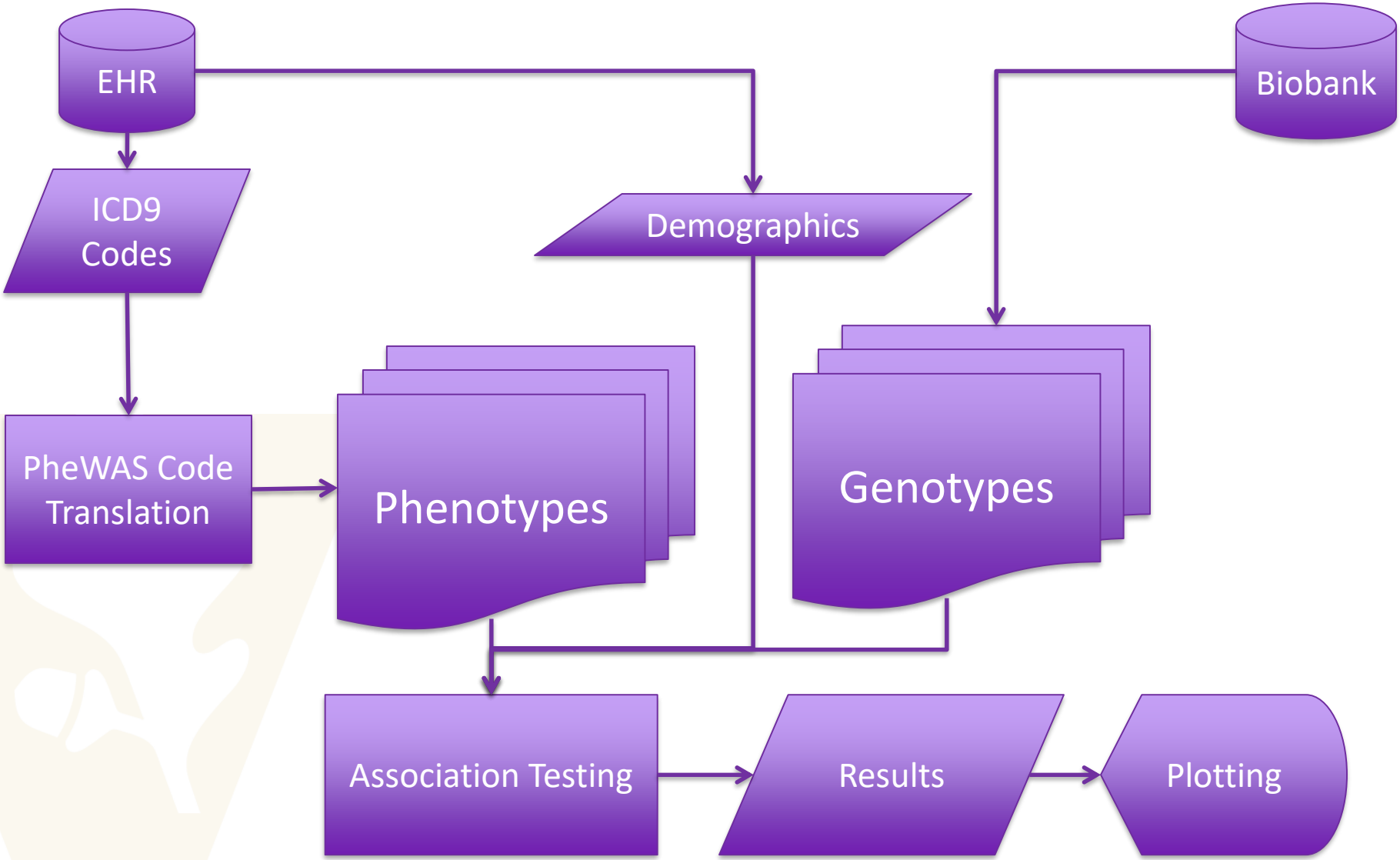
- EHR-linked DNA biobanks can be used for genomic and pharmacogenomic discovery. They can be cost efficient and fast. Big populations are (will be) needed for genomic discovery, deciphering rare variants, and drug-drug interactions.
- Tools to provide access to data, algorithms, and results (Research repositories, PheKB.org, phewascatalog.org).
- Phenotype algorithms are typically portable across EHR systems, healthcare settings, NLP systems, etc.
- Think about confounding.

R PHEWAS PACKAGE

GWAS: Genome Wide Association Study



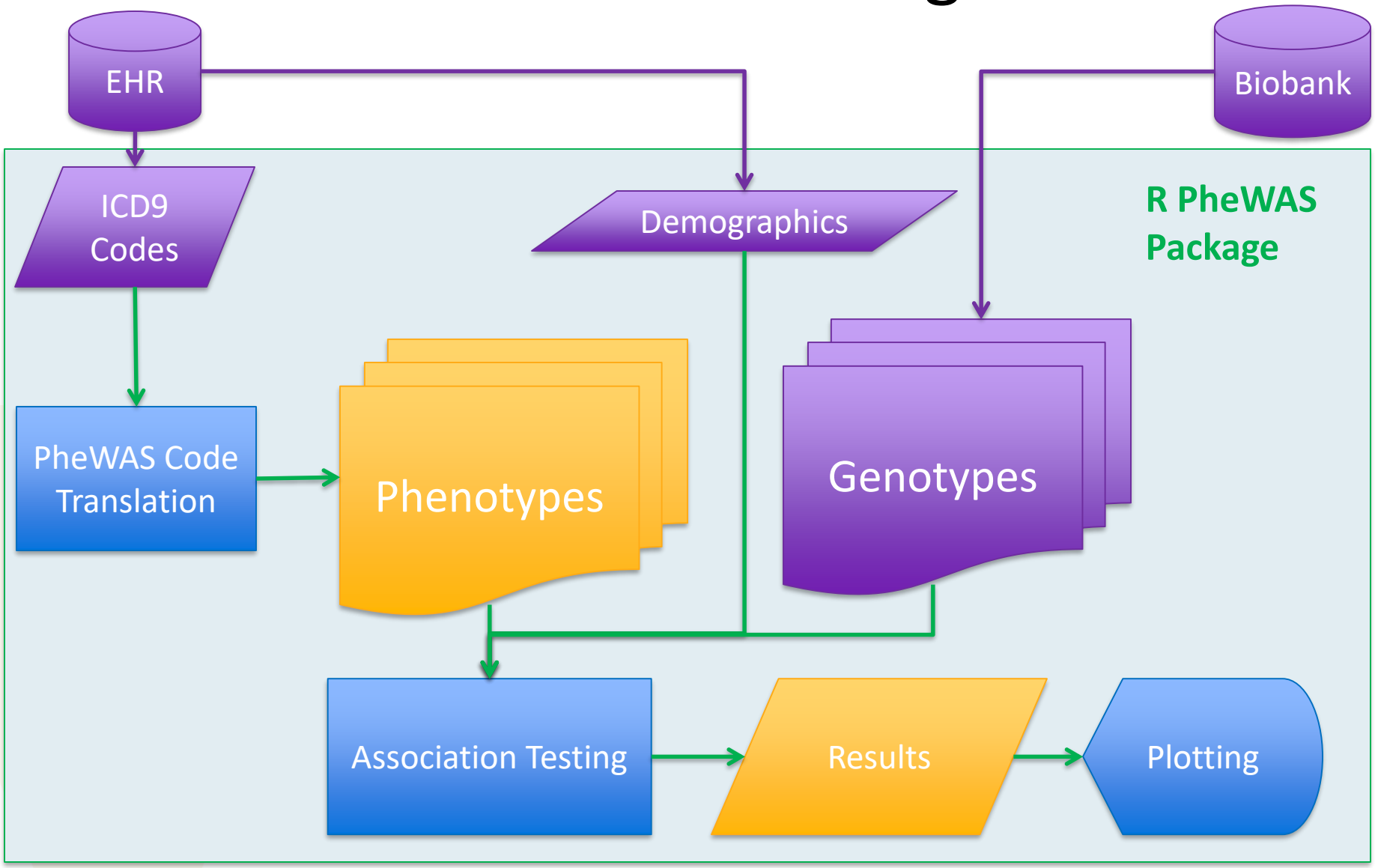
PheWAS: Phenome-Wide Association Study



PheWAS phenotype examples

PheWAS Code	Phenotype	Case ICD-9	Controls exclude:
250	Diabetes	250.*	T1D, T2D, secondary DM
250.1	Type 1 Diabetes	250.01 250.11 250.21 …	“
250.2	Type 2 Diabetes	250.00 250.10 250.20 …	“
714	RA and inflammatory arthritis	714.*	RA, Psoriatic arthritis, Lupus
714.1	RA	714.0, 714.3	“
714.2	Felty’s syndrome	714.3	“

R PheWAS Package



Data Creation

- generateExample() will generate a practice data set.
- It doesn't reflect the nuances and inter-relation of PheWAS codes, but it can be nice to try things out.

Data Creation

- Need a data frame of attributes for regression.
- Phenotype variables can be boolean or numeric.
- Genotype data is best formatted as allele counts.
  - plink --recodeA
  - genotypes=read.table(“genotypes.raw”,head=T)
- Covariates are in tabular form as well.

Create PheWAS Table

- createPhewasTable combines a few steps.
  - Translate: Converts ICD-9s to PheWAS codes. Optional.
  - Aggregation: Combines overlapping codes
  - Exclusions: Excludes individuals from analysis if they share a similar diagnosis, e.g., Type 1 and 2 Diabetes. Optional.
- ICD-9 code data has a triplet format:
  - ID: Which individual?
  - ICD-9: Which code?
  - Count: How many of this code for this ID?
- It also can help for other phenotypes.
  - Find the max for a set of different lab values.
  - Log transform the sum of code counts
- Or, one can skip this step.
  - It's the slowest function for a typical PheWAS.
  - SQL aggregation and perl reshaping are faster.

ID	ICD9	count
1	250.01	1
1	411.2	3
2	714.1	32
	…	

PheWAS method

- phewas(phenotypes, genotypes)
  - Takes tables of different categories
    - Outcomes (eg, phenotypes)
    - Predictors (eg, genotypes)
    - Covariates (eg, age and gender)
    - Adjustments for comparison of results (eg, none vs. BMI)
  - Or use one data frame with vectors of names
- Performs logistic or linear regression
- Can also perform chi-square and t-tests

Data Shapes

One data frame and names

- data

335	411	rs1234	rs4321
T	F	0	1
NA	T	2	1
F	F	2	0

- phenotypes

```
c("335","411")
```

- genotypes

```
c("rs1234","rs4321")
```

Several data frames

- phenotypes

id	335	411
1	T	F
2	NA	T
3	F	F

- genotypes

Id	rs1234	rs4321
1	0	1
2	2	1
3	2	0

PheWAS method options

- Easy to use multithreading with parallel
  - phewas(…, cores=4)
  - Single threaded uses lapply
- Can also return complete models (can cause issues)
- additive.genotypes
  - Calculate allele frequencies
  - Calculate HWE p-value
- Calculate significance thresholds
  - Supply an alpha
  - Returns easy to use T/F variables for each
  - Alpha, Bonferroni, and FDR
  - Returns the details in the object's attributes

PheWAS execution

- The alternate method is to use the data parameter with name vectors in the phenotype, genotype, and covariates parameters.
- > mydata=merge(phenotypes,genotypes)
- > results=p hewas(phenotypes=names(phenotypes)[-1],genotypes=c("rs1234","rs5678"), data=mydata)

Phenotype-only PheWAS

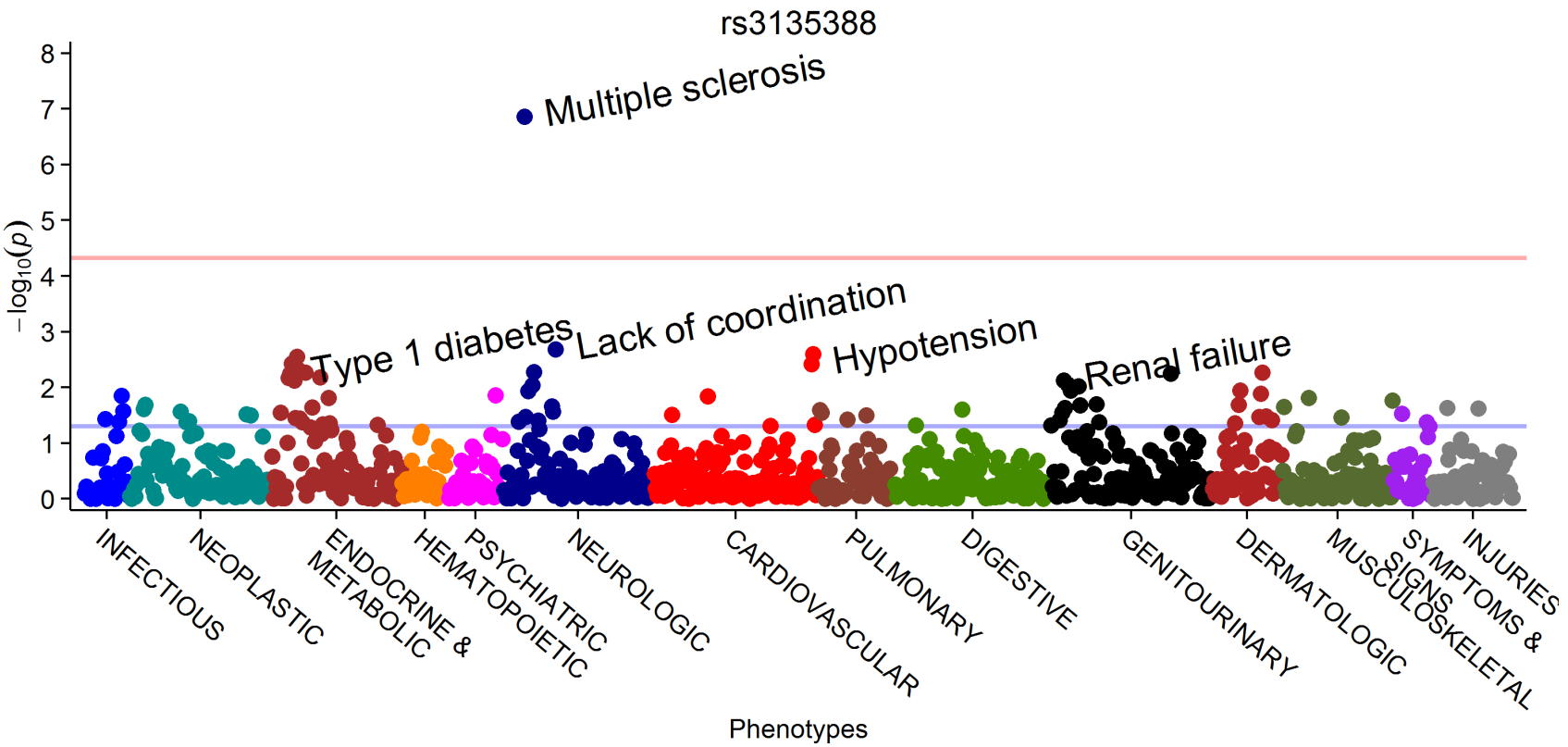
- The phewas function can be used for more than just generic PheWAS.
- In the following example, outcomes and predictors are used for a phenotype only analysis.
- Note that these parameters are simply alternate names for phenotypes and genotypes, respectively.
- max.a1c.results=phewas(outcomes=phenotypes, predictors=csv.phenotypes[,c("id","max.a1c")])

PheWAS Meta-analysis

- The phewasMeta method can assist in meta-analysis of multiple PheWAS, e.g., if one has multiple genotype platforms of data to analyze. It wraps the metagen method of the meta package.
- results.omni1=phewas(phenotypes=phenotypes.omni1,genotypes=genotypes.omni1)
- results.omni1$study="Omni 1"
- results.omni.express=phewas(phenotypes=phenotypes.omni.express, genotypes=genotypes.omni.express)
- results.omni.express$study="Omni Express"
- results.merged=rbind(results.omni1,results.omni.express)
- results.meta=phewasMeta(results.merged)

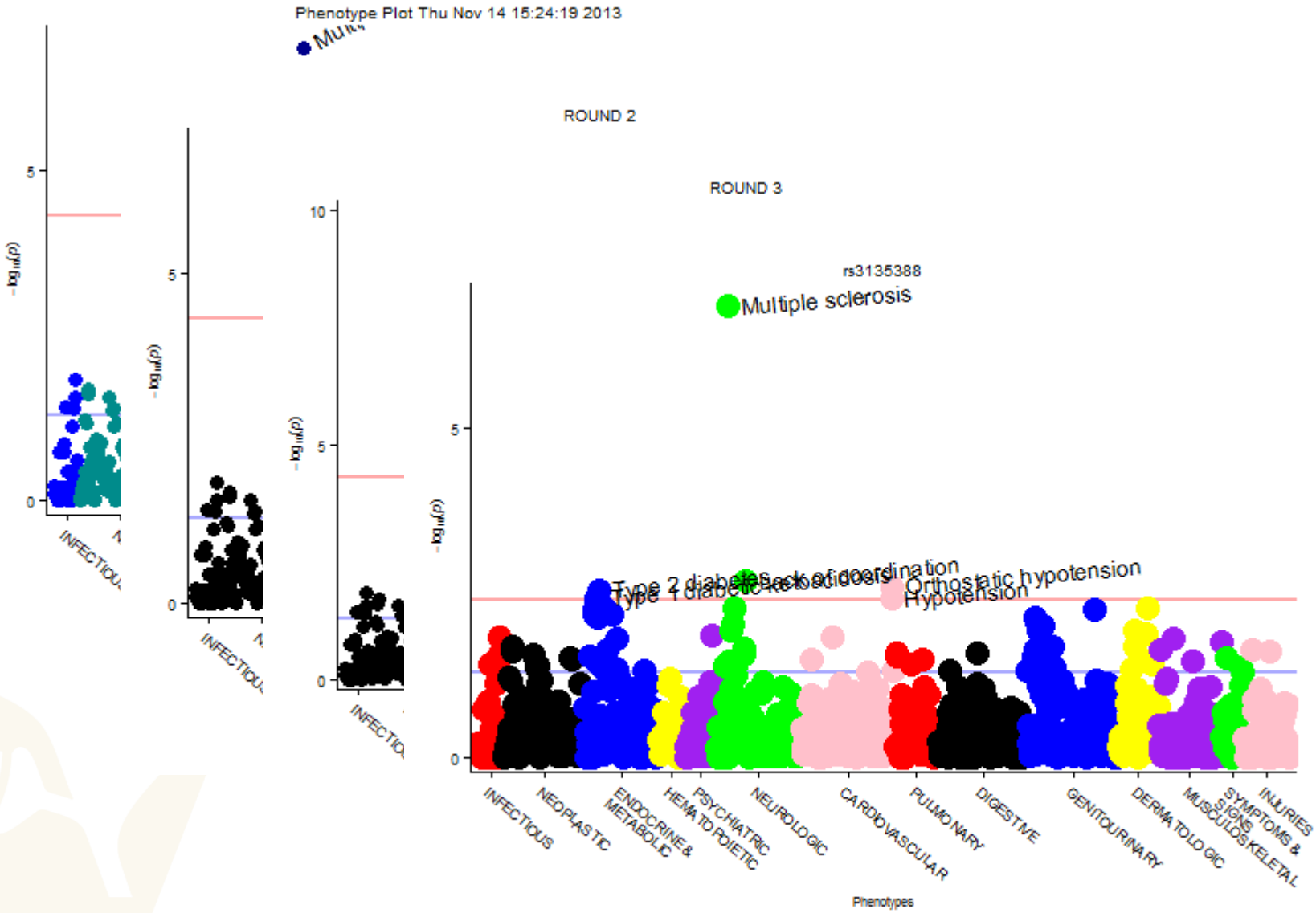
phewasManhattan

- Simple method for plotting right from phewas()



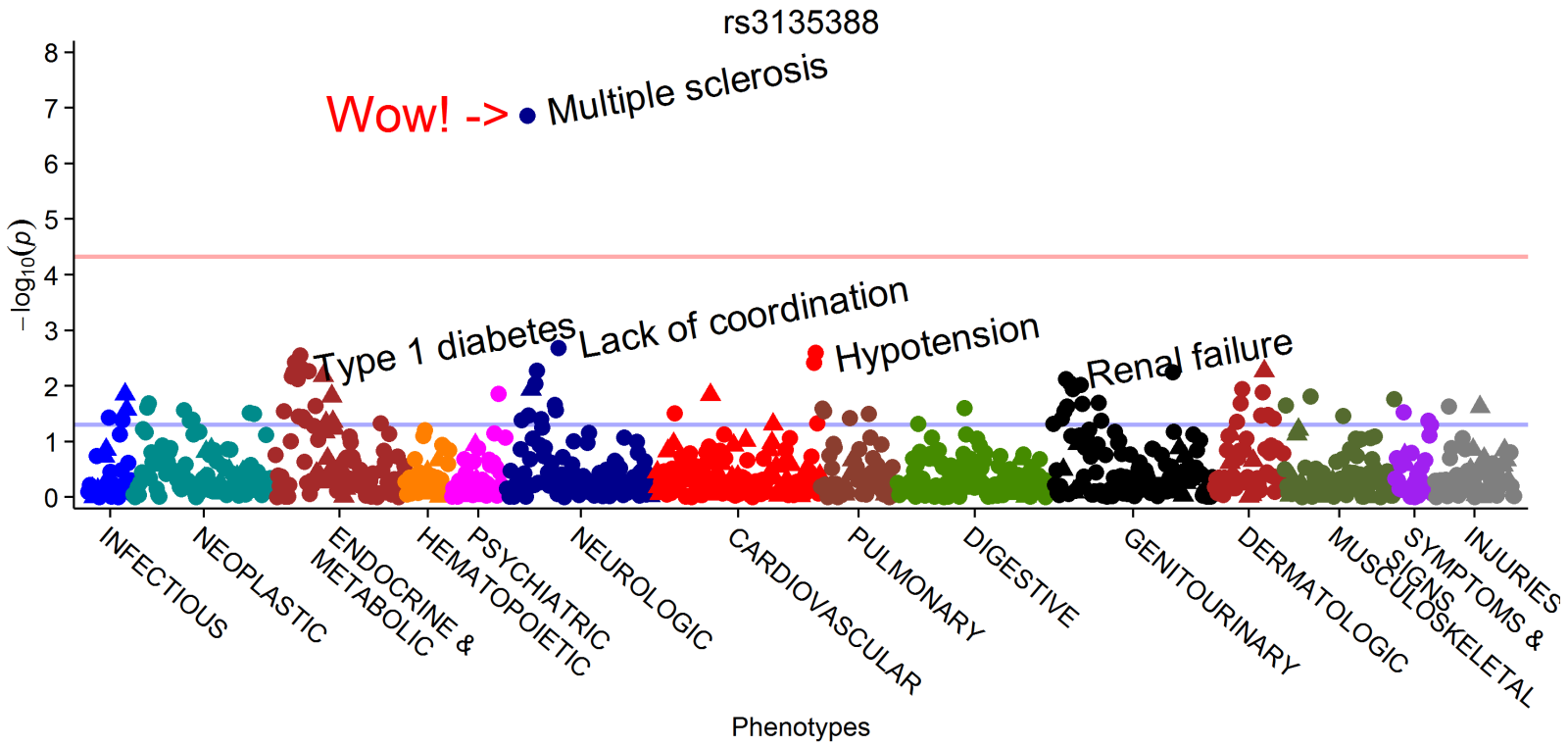
Other Phenotype Plots

- phenotypeManhattan
  - Intended for any p-value plotting.
  - Allows for adding one's own descriptions.
  - Works for GWAS results, too.
- phenotypePlot
  - Most options are documented in this function.
  - Set colors, descriptions, annotations, groupings, and more.
  - These options can be used from the higher level functions as well.


Phenotype Plot Thu Nov 14 15:24:19 2013
ROUND 2
ROUND 3
rs3135388
Multiple sclerosis
Type 1 diabetic ketoacidosis
Lack of coordination
Orthostatic hypotension
Hypotension
INFECTIOUS
NEOPLASTIC
ENDOCRINE & METABOLIC
HEMATOPOIETIC
PSYCHIATRIC
NEUROLOGIC
CARDIOVASCULAR
PULMONARY
DIGESTIVE
GENITOURINARY
DERMATOLOGIC
MUSCULOSKELETAL
SYMPTOMS & SIGNS
INJURIES
Phenotypes

ggplot2

- These plotting functions return ggplot2 objects.
- If one doesn't see the right option, modify the plot!
- plot+aes(shape=HWE_p<.1)+annotate("text",label="Wow! ->", x=175,y=6.9,colour="red",hjust=0,size=8)



Summary

- R PheWAS is a native R package
  - Portable across platforms
  - Extensible
  - Integrated documentation
- Designed to be robust to input data types
- A “classic” PheWAS analysis and plot is only a few function calls away

Github

- https://github.com/PheWAS/PheWAS
- Github allows users to collaborate on open source projects.
- This includes bug reporting and more.
- It also is the best way to access the most recent version of the package.

Installing R PheWAS

- install.packages("devtools")
- install.packages(c("dplyr","tidyr","ggplot2","MASS","meta","ggrepel","DT"))
- devtools::install_github("PheWAS/PheWAS")
- library(PheWAS)

Accessing Documentation

- ?PheWAS
- The ? operator will allow you to look up help for most functions and data objects in PheWAS
- ?? can help find items if you can’t remember the name.
- vignette("PheWAS-package")

Contact

- Feel free to contact me:
  Robert.Carroll@Vanderbilt.edu