Leveraging Big Data to Fuel Precision Health Research

Jennifer “Piper” Below
Vanderbilt Genetics Institute
AIDPH, 2018

Outline

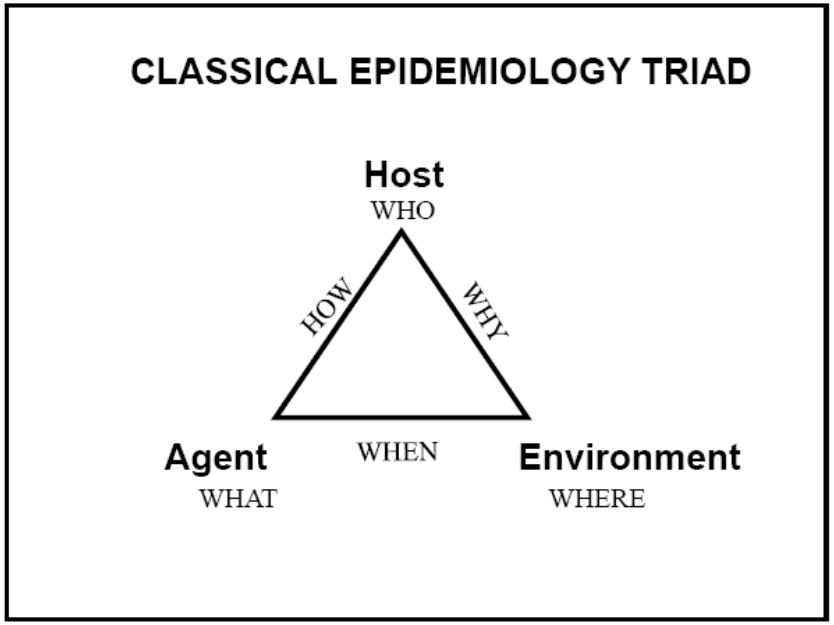
- Genetic epidemiology
  - Objectives
  - Questions
  - Applications
  - Challenges
- Family-based genomic studies
- Genome-wide association studies
- Functionally oriented analyses
- Leveraging biobanks and EHR repositories for phenome-wide characterizations

Defining Genetic Epidemiology

"Classical Epidemiology"

- Understand patterns of disease occurrence in human populations and the factors that influence these patterns.

<u>Applies well to infectious disease</u>
- External agent
- Susceptible host
- Environment that brings the host and agent together (e.g. route of transmission of agent from source to the host)



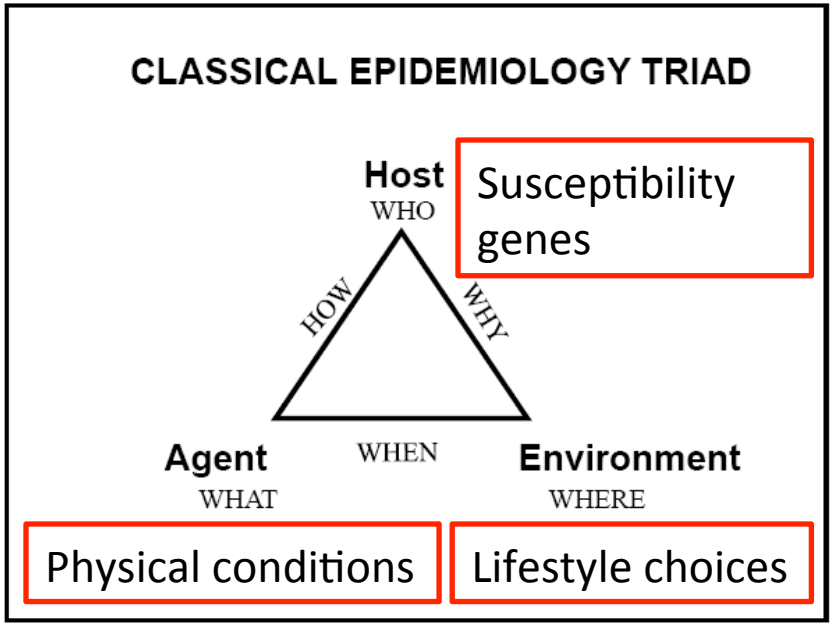
Strom S. 2008

Defining Genetic Epidemiology

"Classical Epidemiology"

- Understand patterns of disease occurrence in human populations and the factors that influence these patterns.

Genetic epidemiology
Agent: physical conditions necessary for disease
Host: intrinsic factors that influence susceptibility
Environment: extrinsic factors



Strom S. 2008

Defining Genetic Epidemiology

- Genetic variation is the exposure of interest

- It is intrinsic
- Sometimes causal
  - Necessary or sufficient?

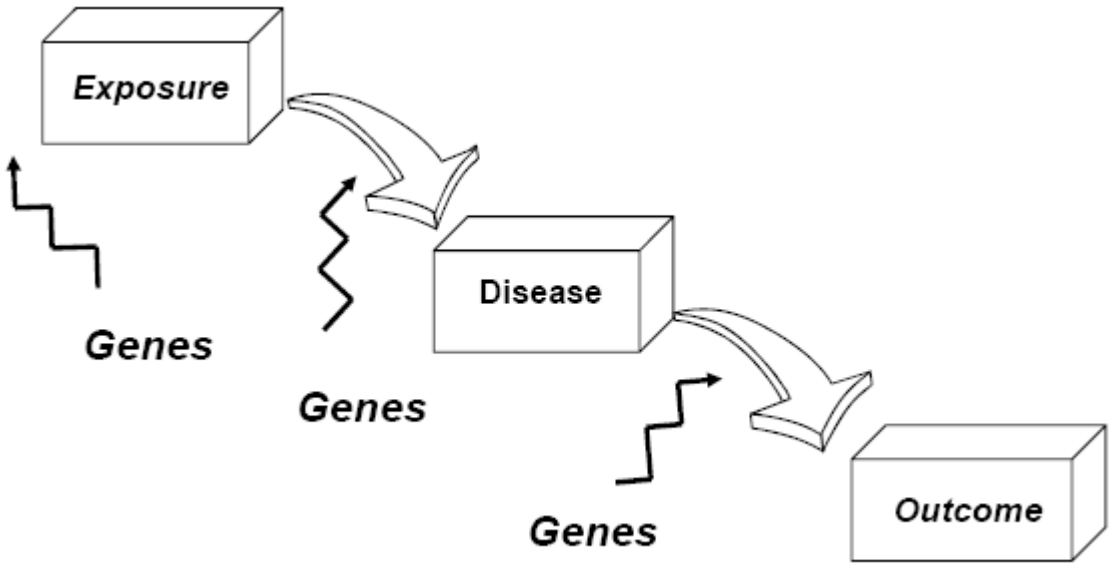
Defining Genetic Epidemiology

"Genetic Epidemiology"

- Assesses the contribution of genetic *AND* environmental factors to better understand the etiology, distribution and control of disease in families and populations

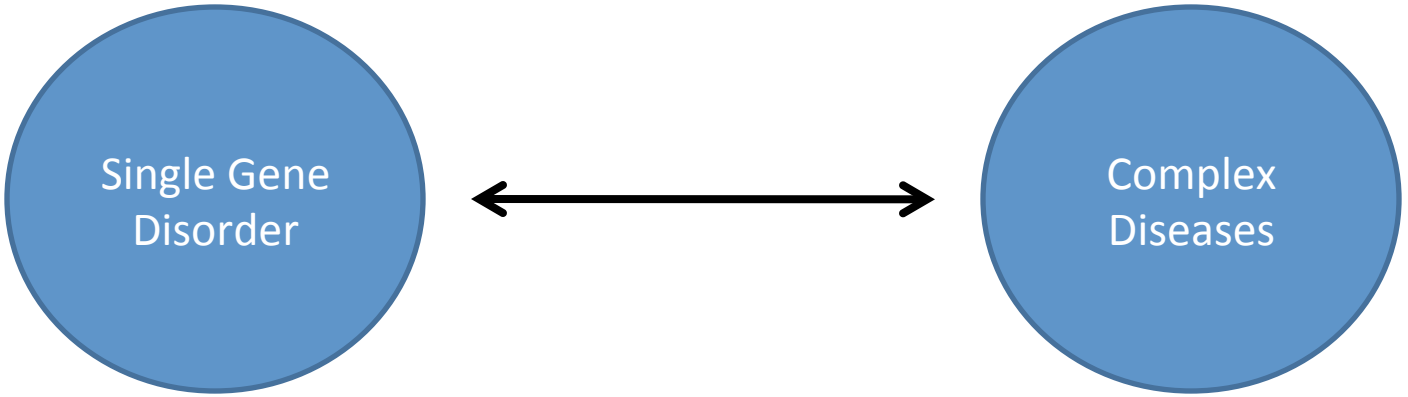
Genetic Epidemiology Objectives

- Genetic epidemiology enhances our understanding of the pathogenesis of disease



Strom S. 2008

Genetic Epidemiology Spectrum



- Rare
- Often caused by a single mutation in a single gene with big effects

- Common in the general population
- Modest effects of multiple genes
- Influential interactions among genes and environment

Genetic Epidemiology Objectives

- Understand disease etiology
- Use genetic markers to directly measure risk (vs surrogate information i.e., family history)
- Reduce heterogeneity of disease classification in descriptive studies

Steps for genetic epidemiologic research

1. Establish that there is a genetic component to the disorder: heritability
2. Establish the relative size of that genetic effect in relation to other sources of variation in disease risk (i.e., environmental effects)
3. Identify the gene(s) responsible for the genetic component

Genetic Epidemiology Questions

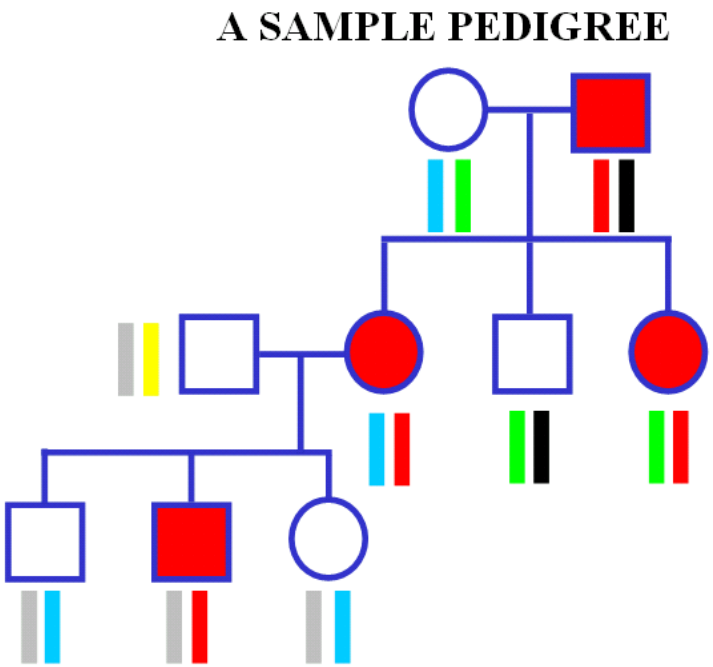
- Establish that there is a genetic component to the disease
  - Is there familial clustering?
  - Could be shared genes or environments
- Size of the genetic effect?
- Evidence for a particular genetic model?
  - Dominant, recessive, polygenic
- Where is the disease gene?
- What are the important variants in the gene?
- How does the gene contribute to disease in the general population?
  - Variant frequency, magnitude of risk, environmental interactions

Human Disease Genetics

- Linkage studies
  - Search for co-transmission of genomic regions and disease in family members (under a given model)
- Association studies
  - Search for genetic differences between population members who have the disease (cases) and those who do not (controls)

Familial clustering

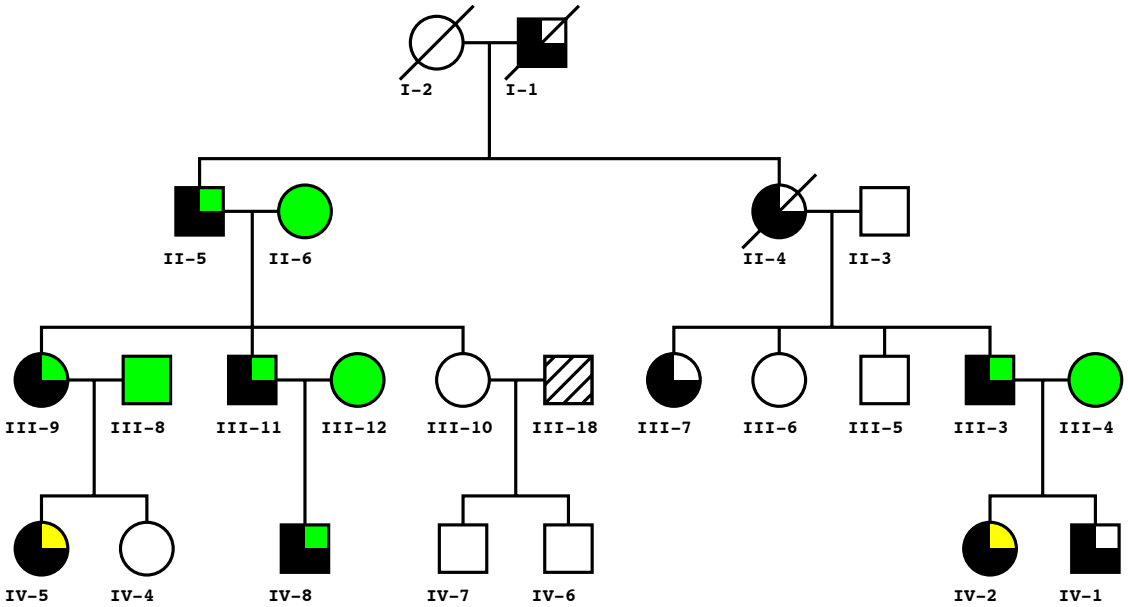
- Conventional wisdom: “disease runs in the family”
- Approach to determining whether there is a genetic component:
  1) Does the phenotype aggregate in families?

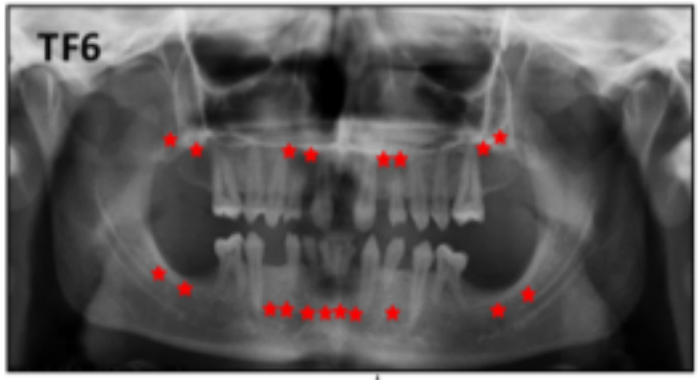


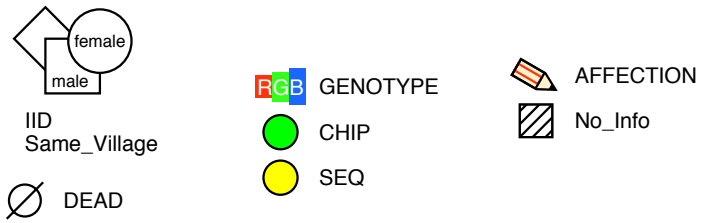
Familial clustering

- Does the disease or trait cluster in families?
  - Could be shared genes or environments
- Approach:
- 1) Does the phenotype aggregate in families?
- 2) Compare disease frequency in relatives of affected individuals with the general population or with disease frequency in relatives of unaffected individuals

Familial Tooth Agenesis







Pathogenic mutations causing tooth agenesis identified by familial genomic analysis (Dinckan *et al*)

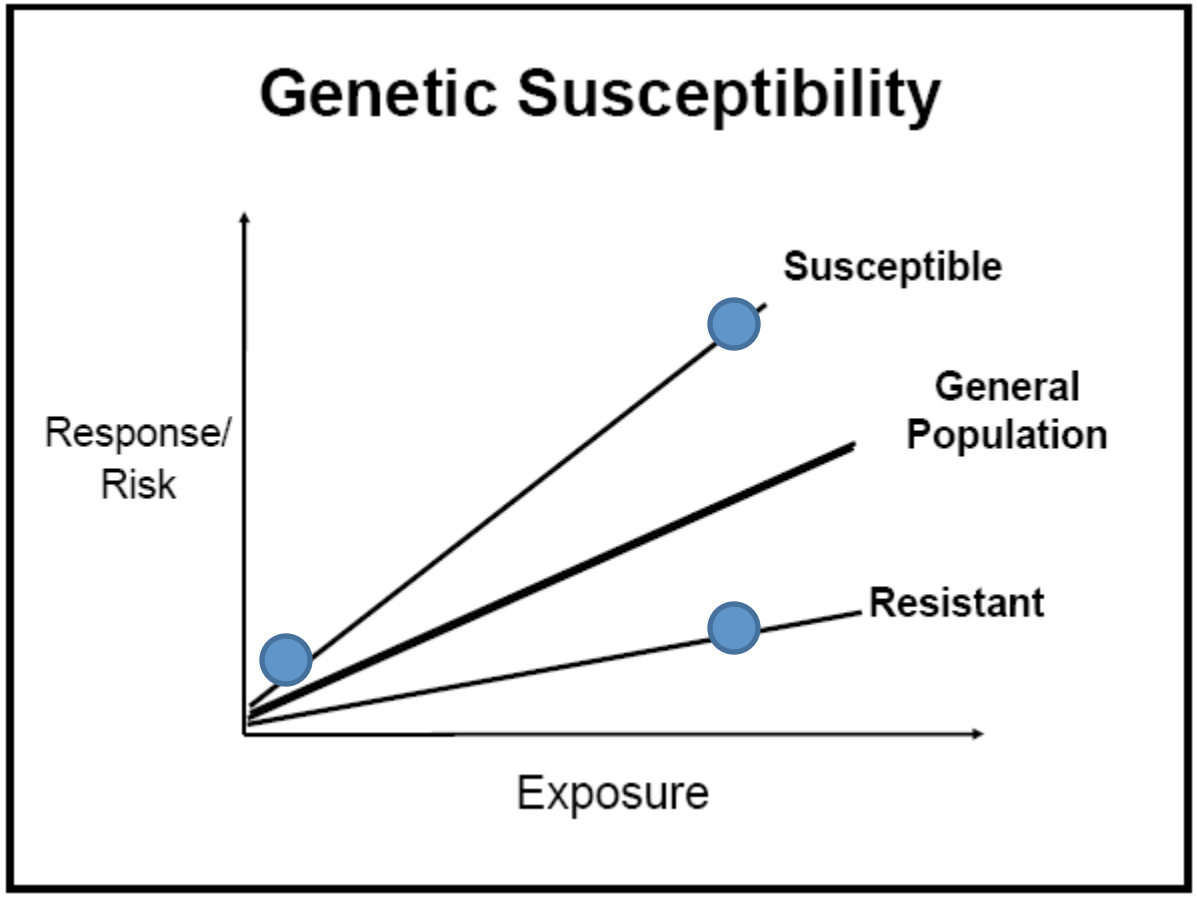
Family	Gene	Zygosity	Nucleotide Change	Protein Change	ExAC
1	*WNT10A*	Hom	NM_025216.2:c.697G>T	p.E233*	0
2	*WNT10A*	Het	NM_025216.2:c.682T>A	p.F228I	0.01
3	*WNT10A*	Het	NM_025216.2:c.682T>A	p.F228I	0.01
	LAMA3	Het	NM_000227.4:c.1097G>A	p.R366H	2.6×10^{-4}
4	*WNT10A*	Hom	NM_025216.2:c.433G>A	p.V145M	2.5×10^{-5}
5	*WNT10A*	Hom	NM_025216.2:c.433G>A	p.V145M	2.5×10^{-5}
6	*WNT10A*	Hom	NM_025216.2:c.433G>A	p.V145M	2.5×10^{-5}
7	*LRP6*	Het	NM_002336.2:c.3607+3_6del	?	0
8	*KREMEN1*	Hom	NM_032045.4:c.146C>G	p.T49R	0
9	*KREMEN1*	Hom	NM_032045.4:c.773_778del	p. F258_P259del	0
10	*DKK1*	Het	NM_012242.2:c.548-4G>T	?	5.8×10^{-4}
	LAMA3	Het	NM_000227.4:c.2798G>T	p.G933V	0
	COL17A1	Het	NM_000494.3:c.3277+3G>C	?	5.5×10^{-5}
11	*ANTXR1*	Hom	NM_032208.2:c.1312C>T	p.R438C	3.6×10^{-4}
12	*TSPEAR*	Hom	NM_144991.2:c.(1726G>T; 1728delC)	p.V576Lfs*37	2.5×10^{-5}
13	*TSPEAR*	Hom	NM_144991.2:c.1877T>C	p.F626S	1.2×10^{-4}
14	*LAMB3*	Hom	NM_000228.2:c.547C>T	p.R183C	5.1×10^{-4}
15	*BCOR*	Het	NM_001123383.1:c.1651G>A	p.D551N	1.4×10^{-4}
	WNT10A	Hom	NM_025216.2:c.682T>A	p.F228I	0.01

Causes of familial aggregation

- Genetic
  - Mendelian (single gene mutation)
  - Non-Mendelian (polygenic or multifactorial)
- Non-Genetic
  - Common exposure to an etiologic agent
  - Coincidence/chance
- Gene*Environment Interaction
  - Disease occurs at a higher frequency among exposed susceptible individuals than among non-exposed susceptible individuals or exposed non-susceptible individuals

Strom S. 2008

G*E



Strom S. 2008

Familial clustering

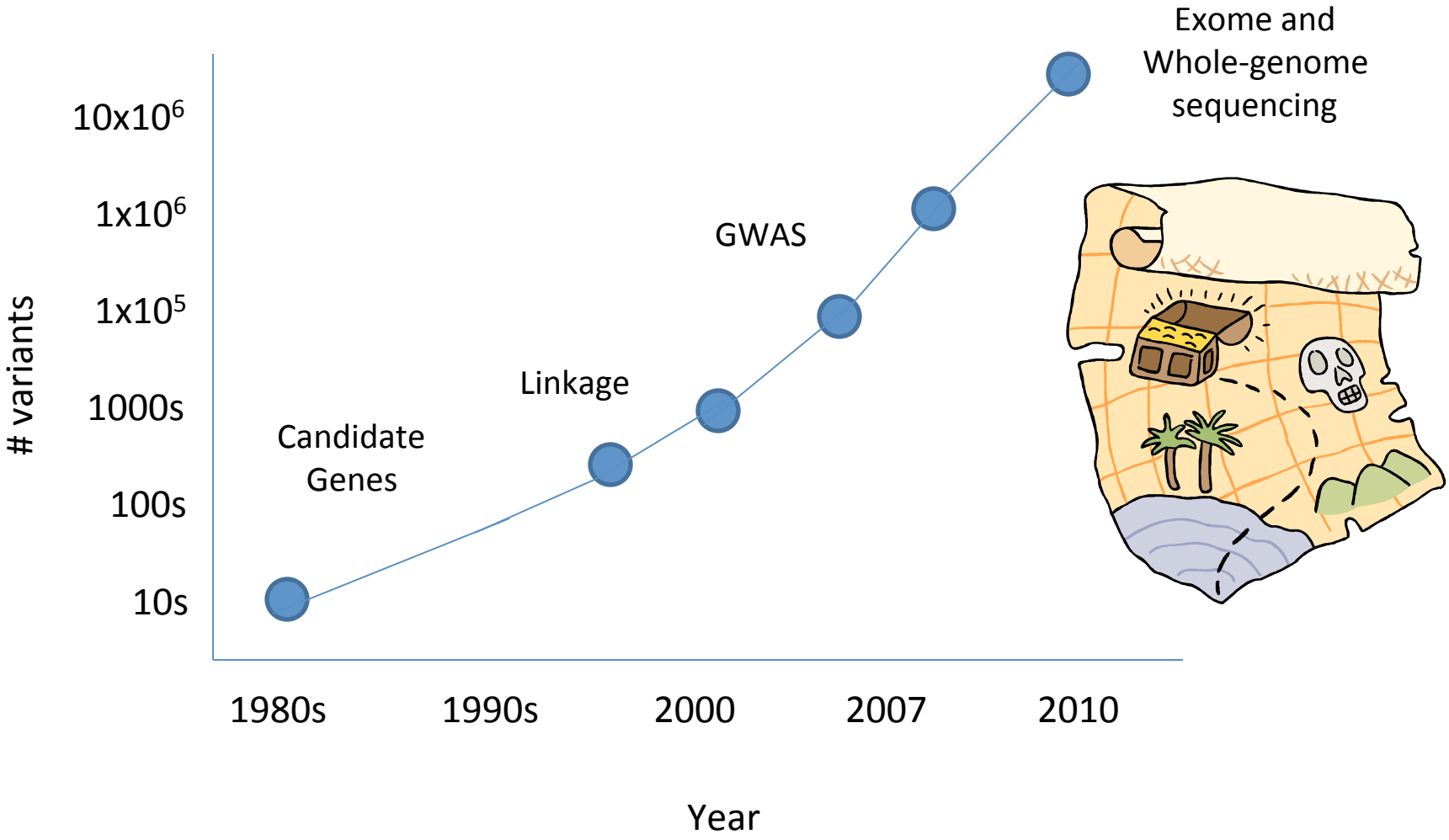
- Does the disease or trait cluster in families?
  - Could be shared genes or environments
- Approach:
- 1) Does the phenotype aggregate in families?
- 2) Compare disease frequency in relatives of affected individuals with the general population or with disease frequency in relatives of unaffected individuals

- Other evidence aggregating in families
  - Earlier age of onset among familial versus non-familial cases
  - Stronger phenotypic correlations between parents and biologic versus adopted children

Genetic Epidemiology Questions

- Establish that there is a genetic component to the disease (familial aggregation)
- Is there evidence for a genetic effect?
- Is there evidence for a particular genetic model?
  - Dominant, recessive, polygenic
- Where is the disease gene?
- What are the important variants in the gene?
- How does the gene contribute to disease in the general population?
  - Variant frequency, magnitude of risk, environmental interactions

Data growth in human genetics



Genetic association study design

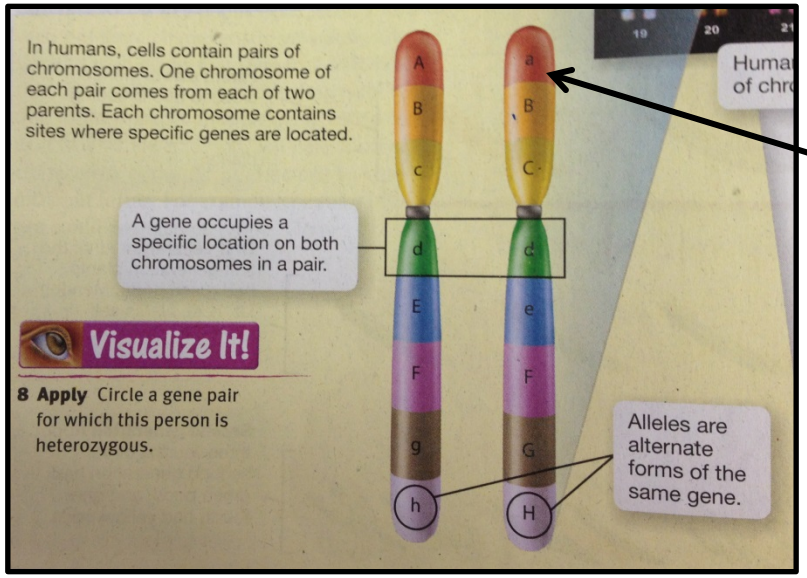
Affected

Unaffected

Identify SNP genotypes

statistics

Determine whether a particular SNP allele occurs more often among cases compared to controls



What are the genotypes here?

Genetic association study design

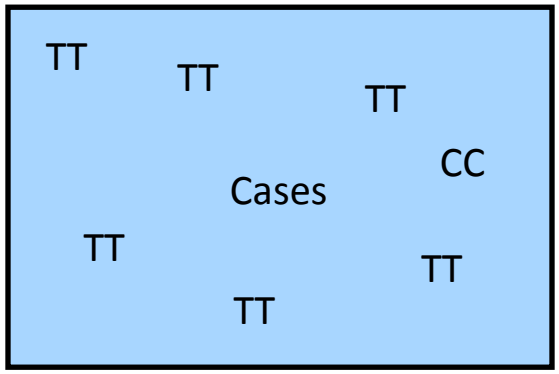
Genetic Association Studies

Advantages:

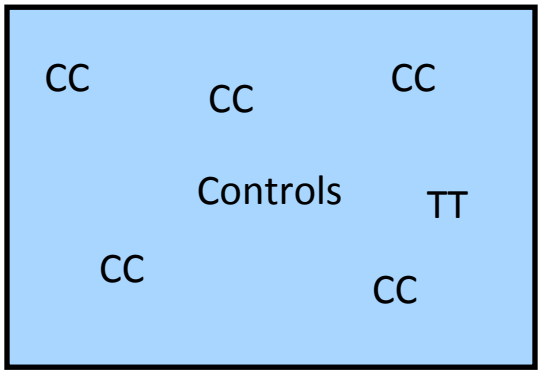
- Ease of data collection relative to family-based studies
- Individuals in control group could be used as controls for other studies

Disadvantages:

- Power depends on effect size of the variant on risk, allele frequency, and sample size
- Hard to interpret biological significance when variants are non-coding
- Subject to bias due to population stratification, if not modeled appropriately



Freq T = 90%



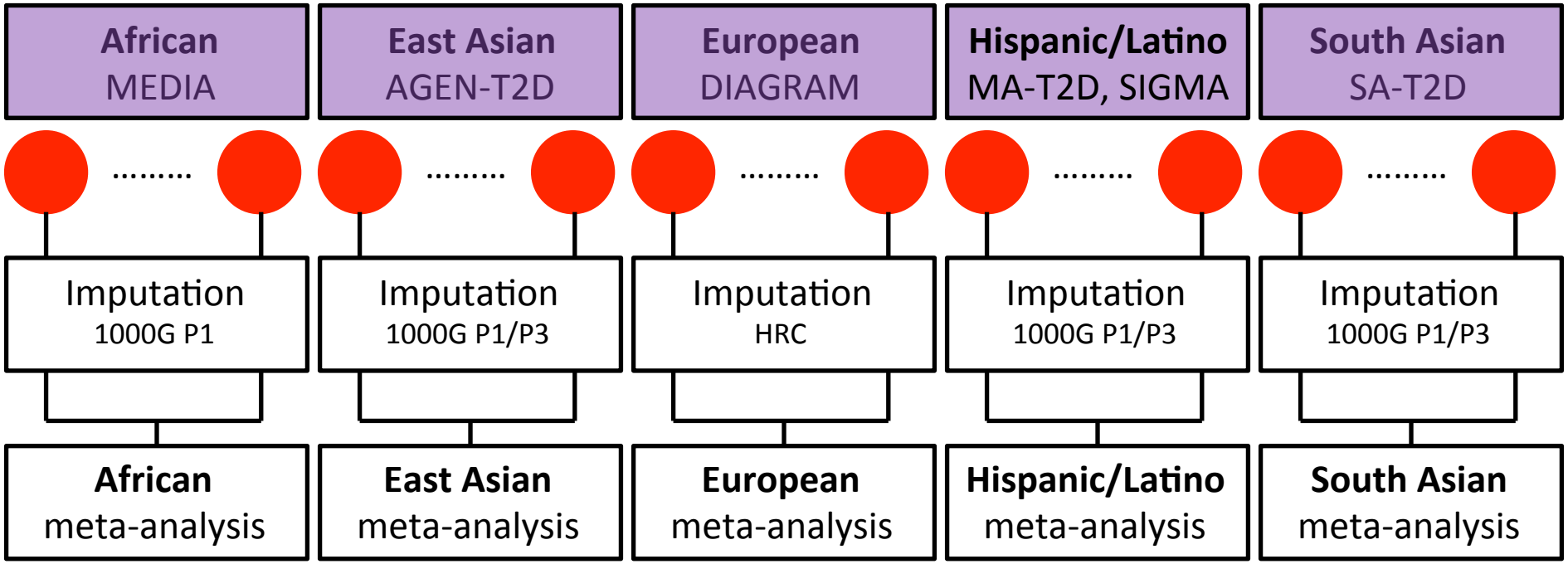
Freq T = 15%

Leveraging diversity in GWAS, an example

- Genome-wide association studies (GWAS) of type 2 diabetes (T2D) have been extremely successful in multiple ancestry groups.
- However, at the majority of these loci, the variants and transcripts through which these effects on T2D are mediated are unknown.
- Higher-density reference panels from diverse populations will improve the utility of imputation in fine-mapping studies.

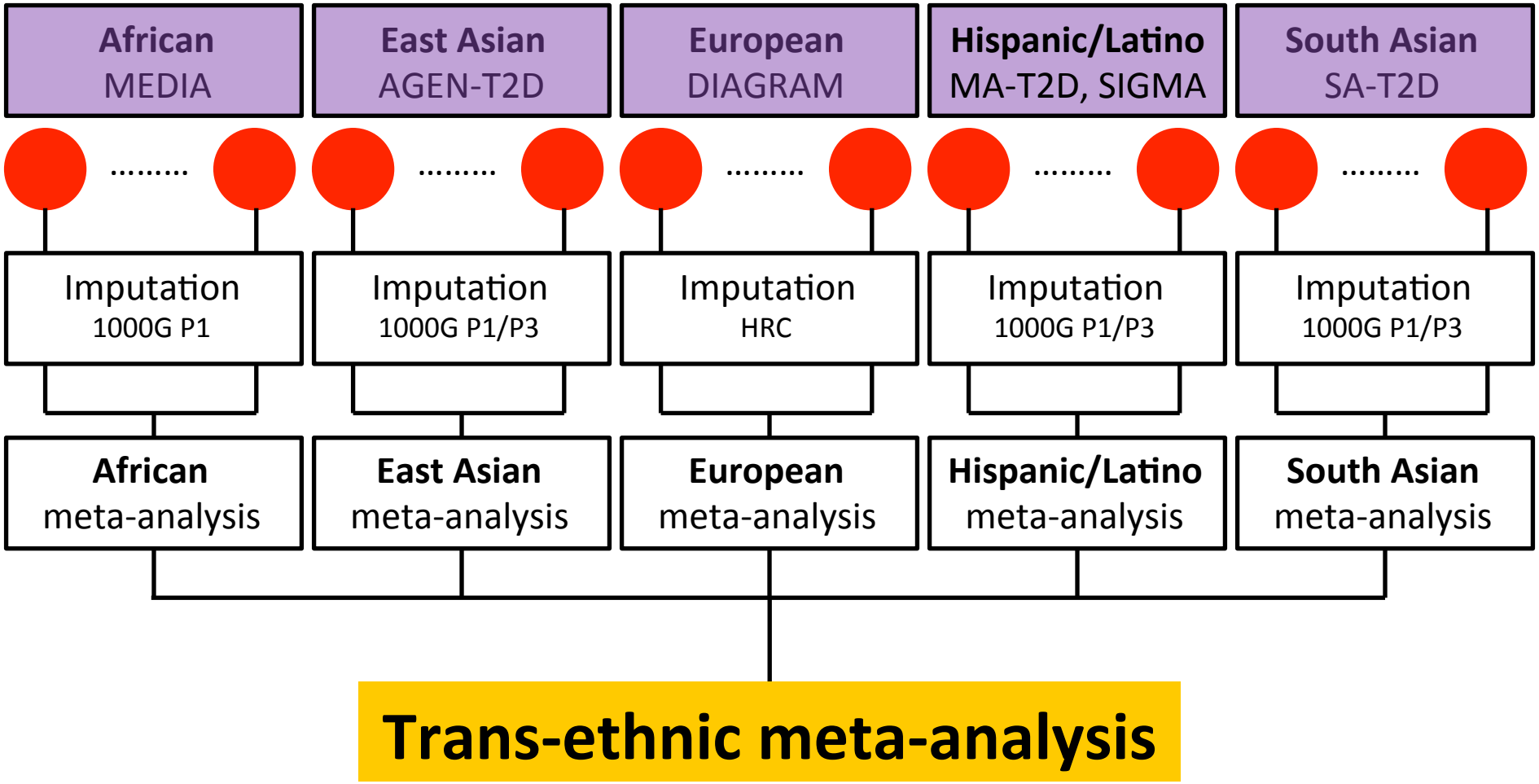
DIAMANTE

(DIAbetes Meta-ANalysis of Trans-Ethnic association studies)



DIAMANTE

(DIAbetes Meta-ANalysis of Trans-Ethnic association studies)



Aims

- Discover novel variants and loci specific to understudied populations
- Increase power to detect novel loci for T2D susceptibility across populations.
- By taking advantage of differences in the patterns of linkage disequilibrium between ethnicities, improve fine-mapping resolution of causal variants.

To better identify causal molecular mechanisms and genes within T2D GWAS loci.

Hispanic/Latino effort

- Hispanic/Latino populations are disproportionately affected by T2D as well as other cardiometabolic diseases (serum lipid levels)
- Presents unique opportunity to work toward correcting a major health disparity while improving our understanding of genetics of these traits

Largest meta-analyses in H/L to date

- **T2D: ~13k cases and ~21.5k controls**
- **Lipids: ~25-28k samples**

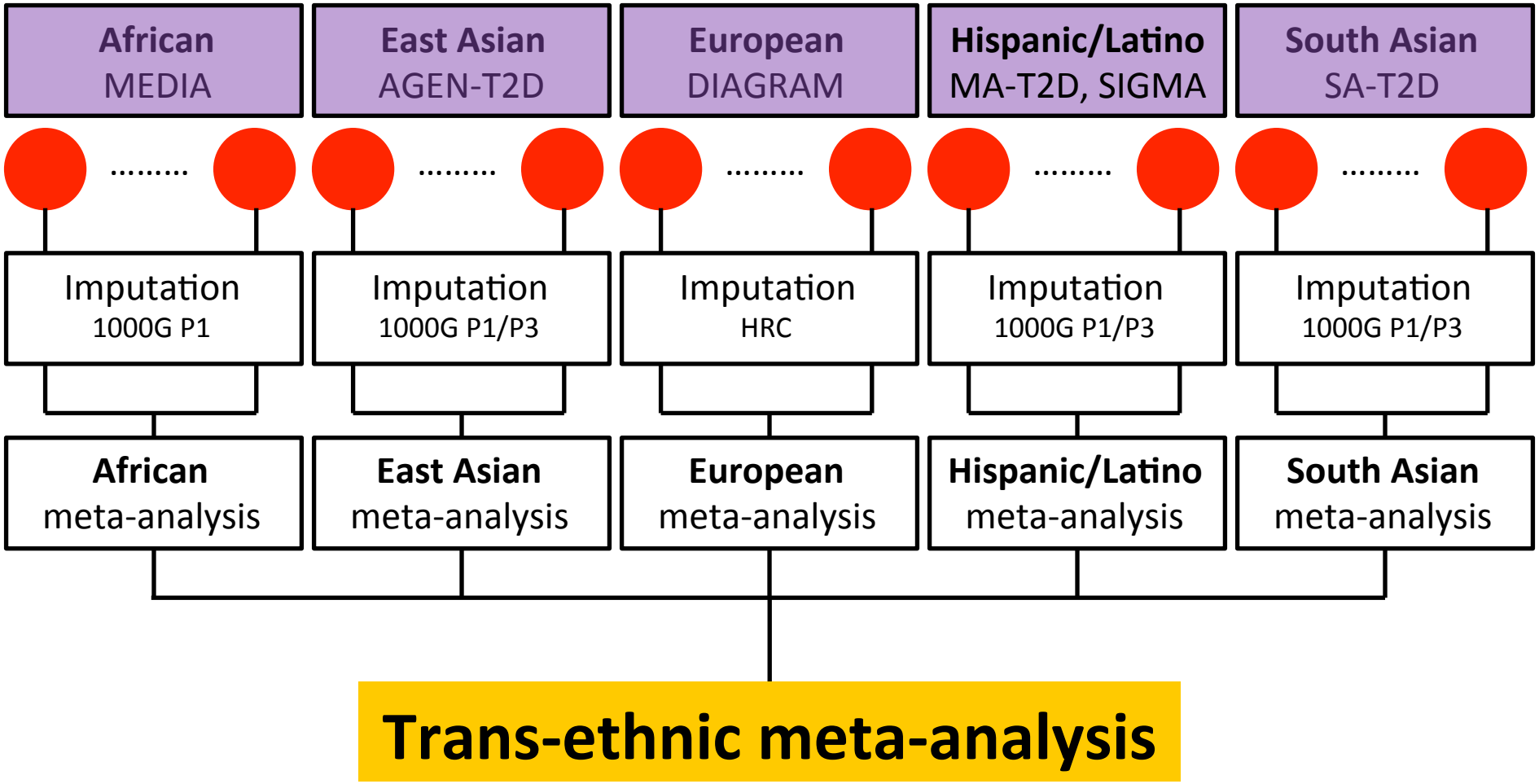
	N by trait					
study	**T2D cases**	**T2D controls**	**HDL cholesterol**	**LDL cholesterol**	**total cholesterol**	**triglycerides**
BioMe	1,358	2,538	-	-	-	-
Genetics of Latinos Diabetic Retinopathy (GOLDR)	500	86	597	596	597	597
Hispanic Community Health Study / Study of Latinos (HCHS/SOL)	2,460	5,185	12,730	12,467	12,731	12,730
Mexican-American Hypertension-Insulin Resistance (HTN-IR)	112	620	732	729	733	730
Insulin Resistance Atherosclerosis Study Family Study (IRASFS)	-	-	1,019	1,007	1,020	1,019
Insulin Resistance Atherosclerosis Study (IRASc)	-	-	176	171	176	176
Los Angeles Latino Eye Study (LALES)	1,217	1,963	-	-	-	-
Mexican American Study of CAD (MACAD)	41	765	730	704	726	735
Multi-Ethnic Study of Atherosclerosis (MESA)	268	1,224	1,422	1,387	1,419	1,422
Mexico City sample 1 (MC1)	960	339	1,278	1,233	1,268	1,277
Mexico City sample 2 (MC2)	895	888	1,783	1,783	1,783	1,783
Non-Insulin-Dependent Diabetes Mellitus (NIDDM)	52	190	241	241	244	244
San Antonio Family Heart Study (SA)	312	270	-	-	-	-
Slim Initiative in Genomic Medicine for the Americas (SIGMA)	4,245	4,068	2,890	1,154	2,064	3,738
Starr County Health Studies (SC)	449	170	544	531	544	544
Women's Health Initiative (WHI)	282	3,205	3,361	3,361	3,361	3,361
total	13,151	21,511	27,503	25,364	26,666	28,356

Summary of H/L results

- Identified a total of 29 potentially novel gene/locus associations
- 72 known genes for T2D and lipids traits were replicated
- Gene ontology enrichment analysis of MetaXcan lipid trait results identified pathways most significantly enriched:
  - negative regulation of lipoprotein lipase activity
  - negative regulation of very-low-density lipoprotein particle clearance
  - chylomicron remnant clearance
- Many of the known hits never previously observed in a HL population

DIAMANTE

(DIAbetes Meta-ANalysis of Trans-Ethnic association studies)



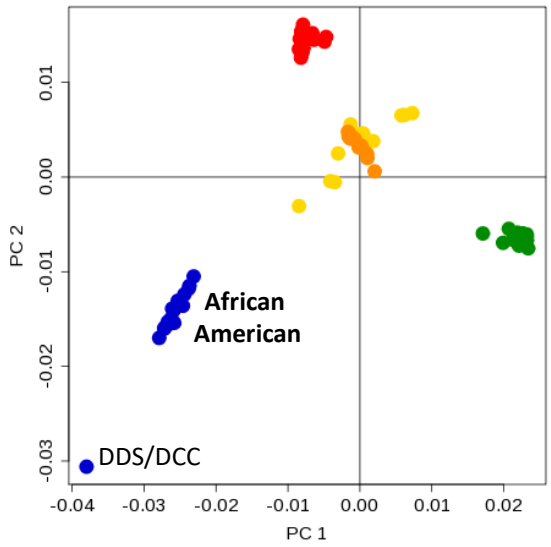
Subjects

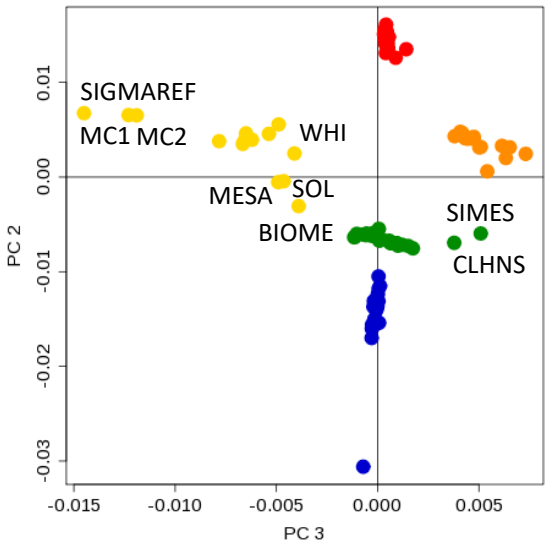
Ancestry group	GWAS	Cases	Controls	Effective sample size
African	22	15,487	23,709	32,593 (7.1%)
East Asian	37	46,696	143,172	105,333 (23.0%)
European	32	80,154	853,816	251,740 (55.0%)
Hispanic/Latino	14	12,385	21,423	27,417 (6.0%)
South Asian	15	16,540	32,952	40,737 (8.9%)
Total	**120**	**171,262**	**1,075,072**	**457,820**

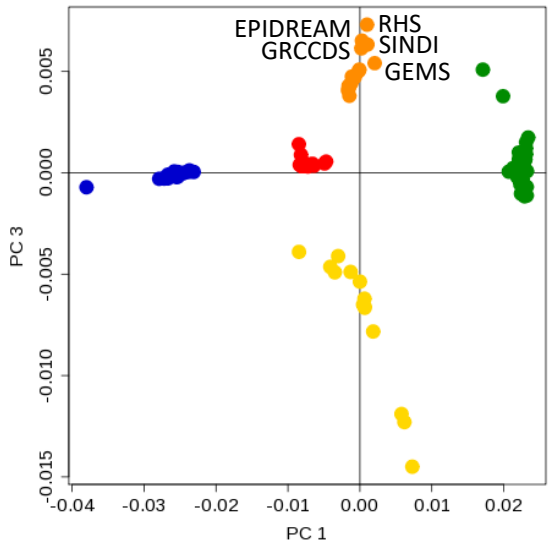
- **Association summary statistics obtained from each study (98 GWAS).**
- **Focused on 19,829,461 autosomal bi-allelic SNVs that are shared across reference panels with MAF ≥0.5% in at least one ancestry group, and AF diff <20% in this analysis.**

Collecting and harmonizing genomic data from 120 different study groups…

PC Analysis







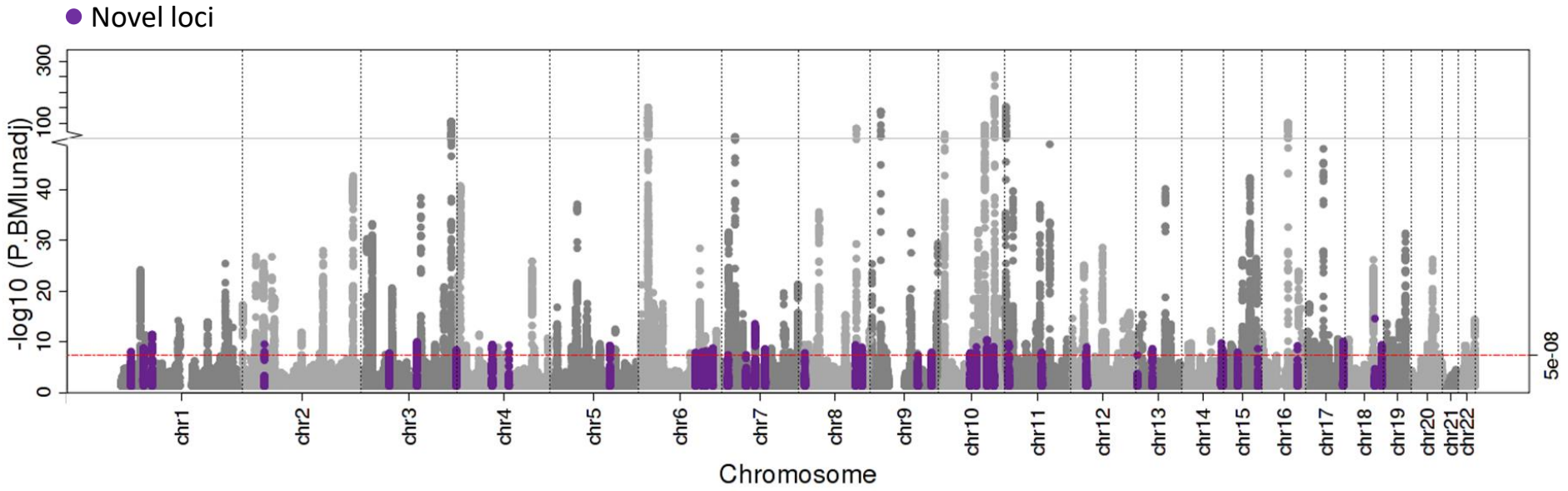
- Matrix of genome-wide mean allele frequency differences between each pair of studies.
- Multi-dimensional scaling to obtain 3 PCs.

Meta-regression

208 loci attain genome-wide significant association

40 novel loci

168 known loci



- Loci with lead SNVs mapping within 1Mb are combined as a single locus.
- Consider SNVs reported in at least 80% of total effective sample size.
- Three axes of genetic variation included in meta-regression as covariates to account for ancestry.
- Double genomic control correction: λ_{META}=1.238.

Dissection of distinct association signals at T2D susceptibility loci

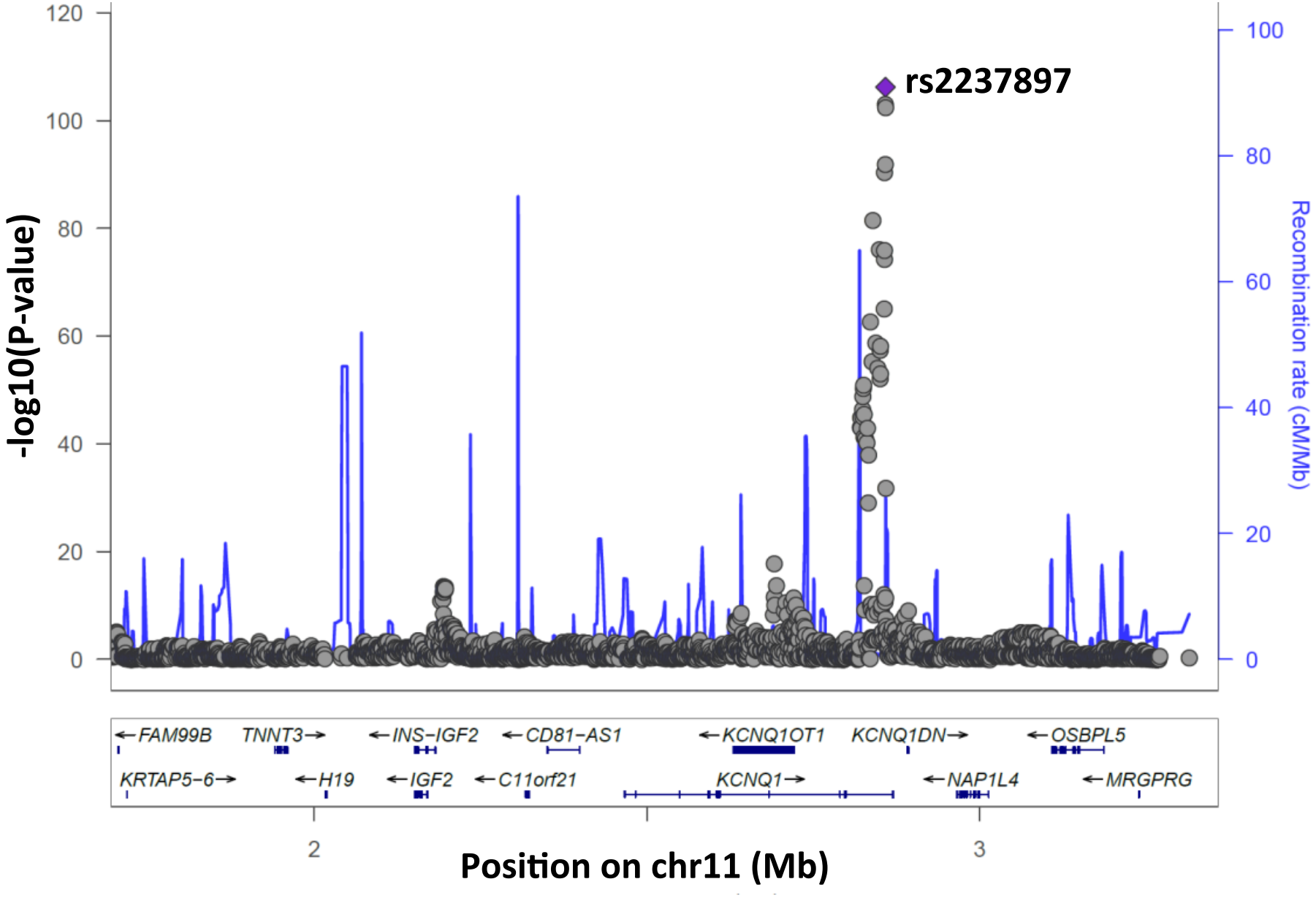
- Each GWAS matched to ethnic-specific group in 1000G Phase 3 reference panel.
- Iterative approximate conditional analysis, implemented in GCTA: forward selection of index variants for each loci:
  - Approximate conditional analysis in each study
  - Trans-ethnic meta regression of conditional association summary statistics in MR-MEGA
  - Continued iteratively until residual association of $p \geq 10^{-5}$
- Total of 342 distinct association signals at locus-wide significance $p < 10^{-5}$ at 208 distinct loci.

Dissection of association signals

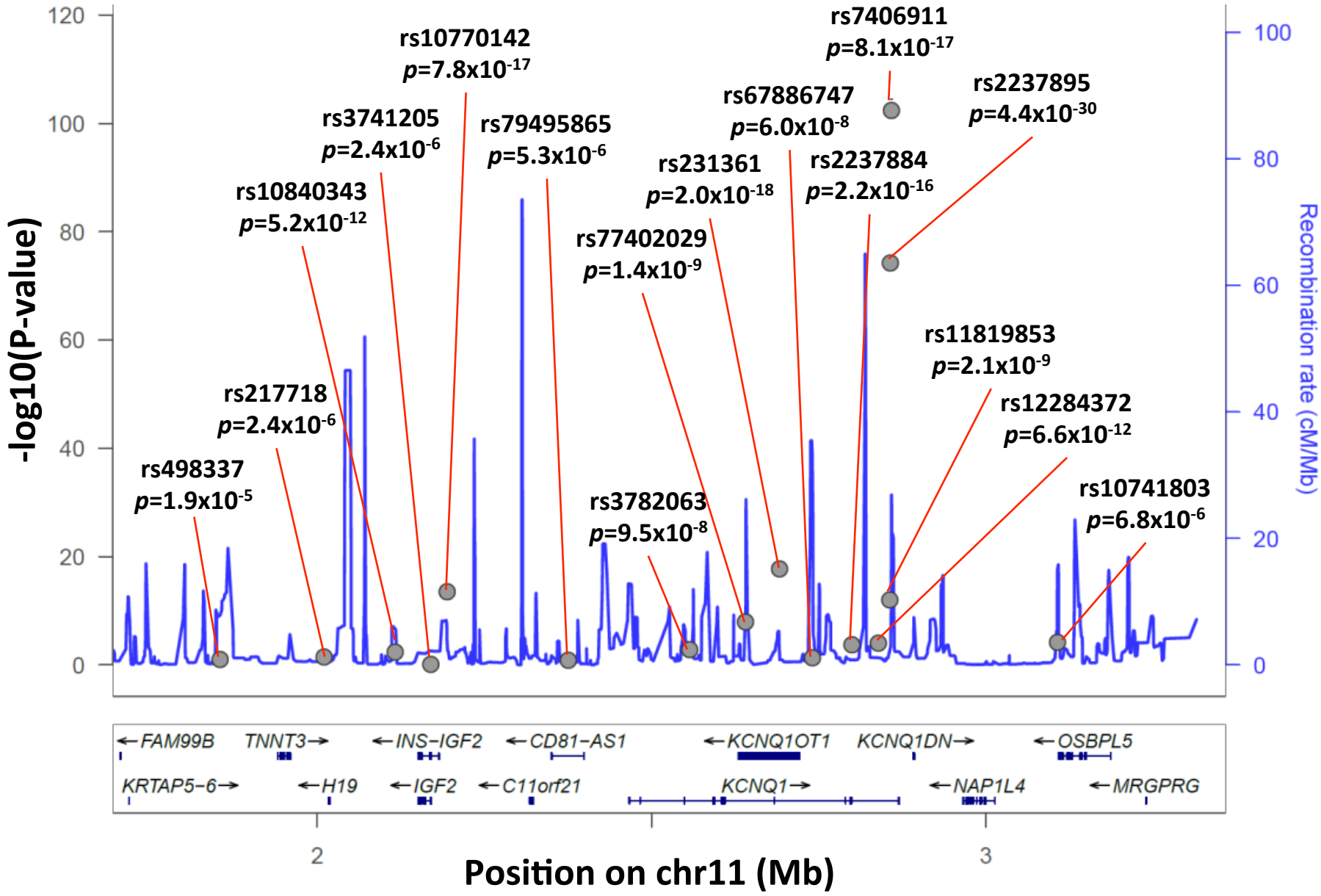
Number of distinct signals	Loci
1	148
2	31
3	17
4	7
5 or more	5
Total	**208**

- 29% of loci have evidence of distinct signals of association at locus-wide significance.

INS-IGF2 and *KCNQ1*



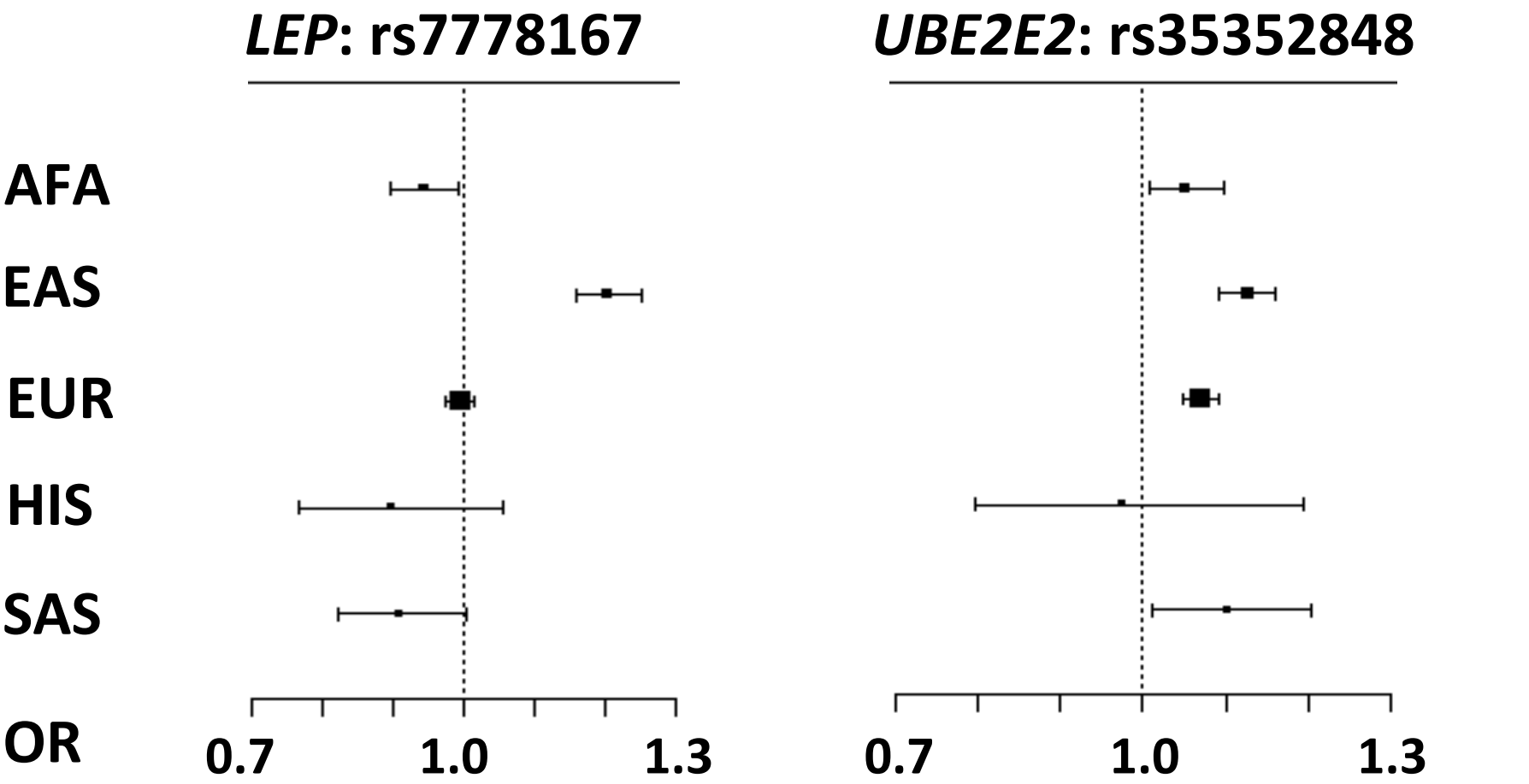
INS-IGF2 and *KCNQ1*



Heterogeneity in allelic effects on T2D correlated with ancestry

- Meta-regression partitions heterogeneity into
  - (i) correlated with ancestry and
  - (ii) residual
- Allelic effects on T2D risk of index variants were predominantly consistent across populations
- Significant evidence of heterogeneity correlated with ancestry for 55 association signals (16.1%, $p<0.00014$)

Heterogeneous allelic effects



p-Ancestry-het=8.2x10^{-16}
p-Residual-het=0.080

p-Ancestry-het=4.2x10^{-11}
p-Residual-het=0.018

Fine-mapping distinct T2D association signals

- Defined credible sets of variants that accounted for 99% of the posterior probability of driving each distinct association signals.
- Substantial improvement in fine-mapping resolution over previous efforts.
- 99% credible sets at 139 (40.6%) signals include ten or fewer variants.
- 99% credible sets for 36 (10.5%) distinct signals contained only one variant
- Lead SNVs at 70 signals have >80% posterior probability of driving association.

Enrichment of T2D association signals

1. GENCODE coding regions

   Credible set variants were highly significantly enriched in coding exons: p=1.4x10^{-5}, OR=5.23.

Enrichment of T2D association signals

1. GENCODE coding regions

   Credible set variants were highly significantly enriched in coding exons: p=1.4x10^{-5}, OR=5.23.

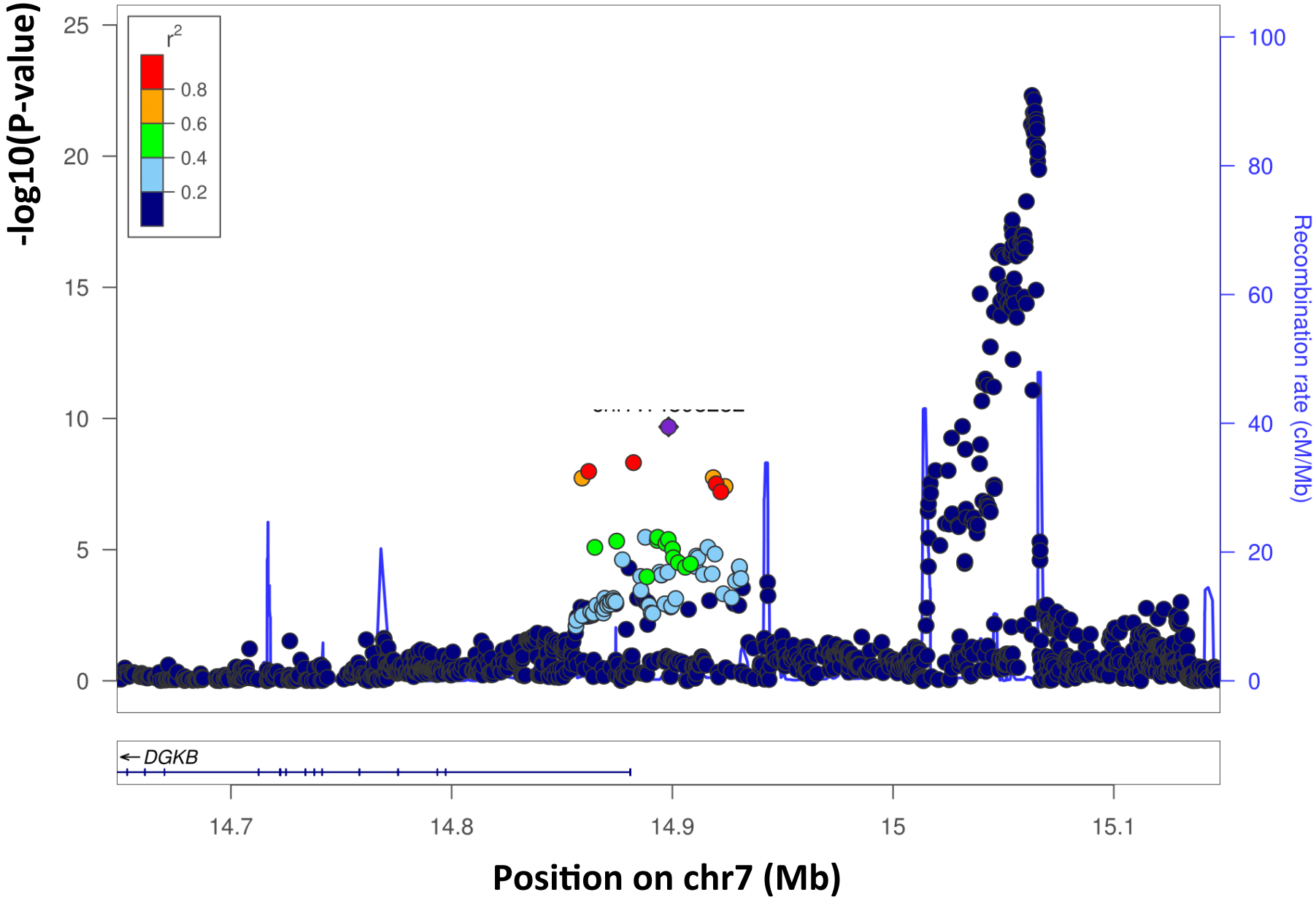
2. Genomic annotations of active enhancer and promoter elements and ChIP-seq binding sites.

   Significant joint enrichment for transcription factor binding site (TFBS) for

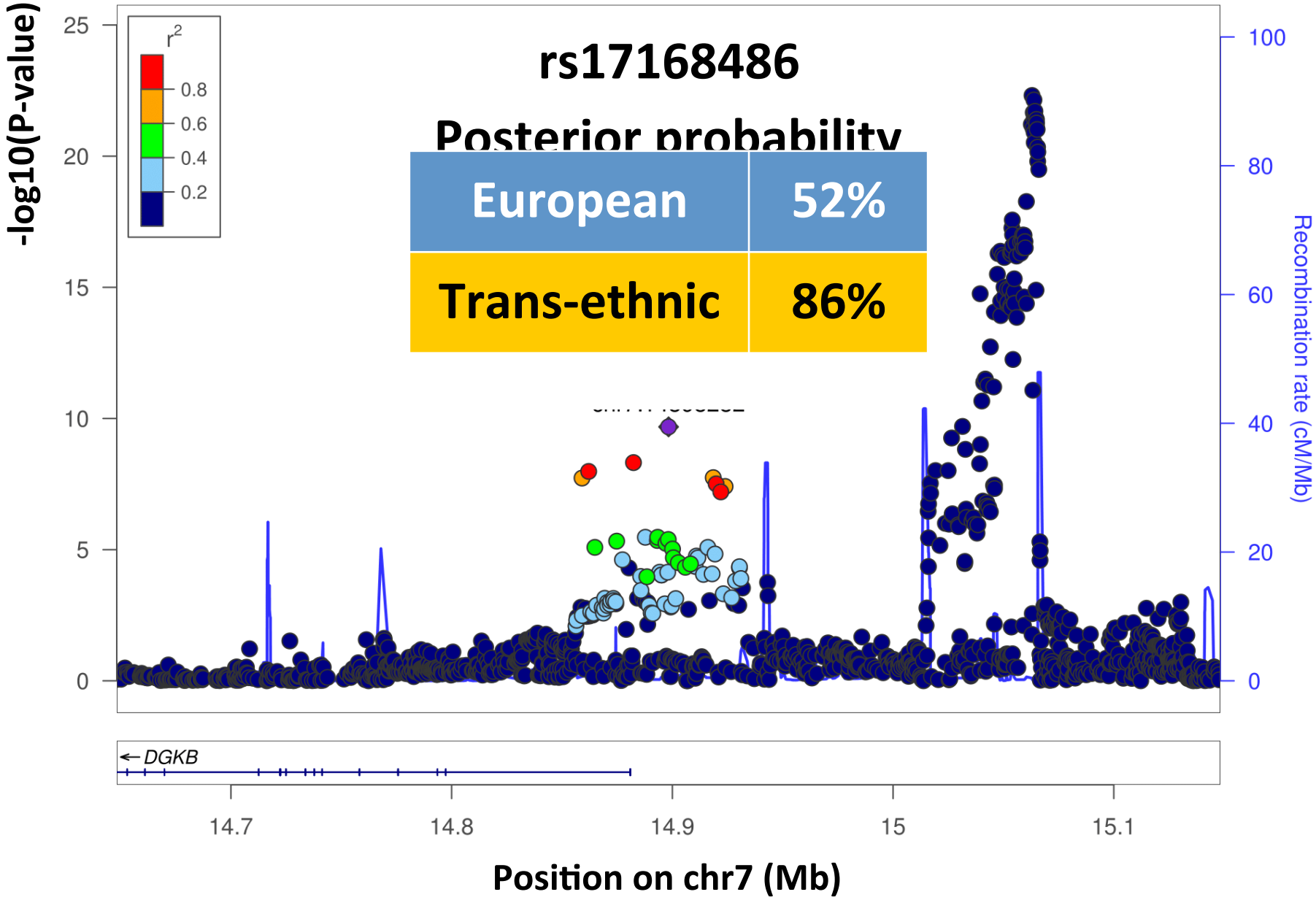
   PDX1: p=2.6x10^{-6}; OR 9.52.

   FOXA2: p=1.8x10^{-5}; OR 5.82.

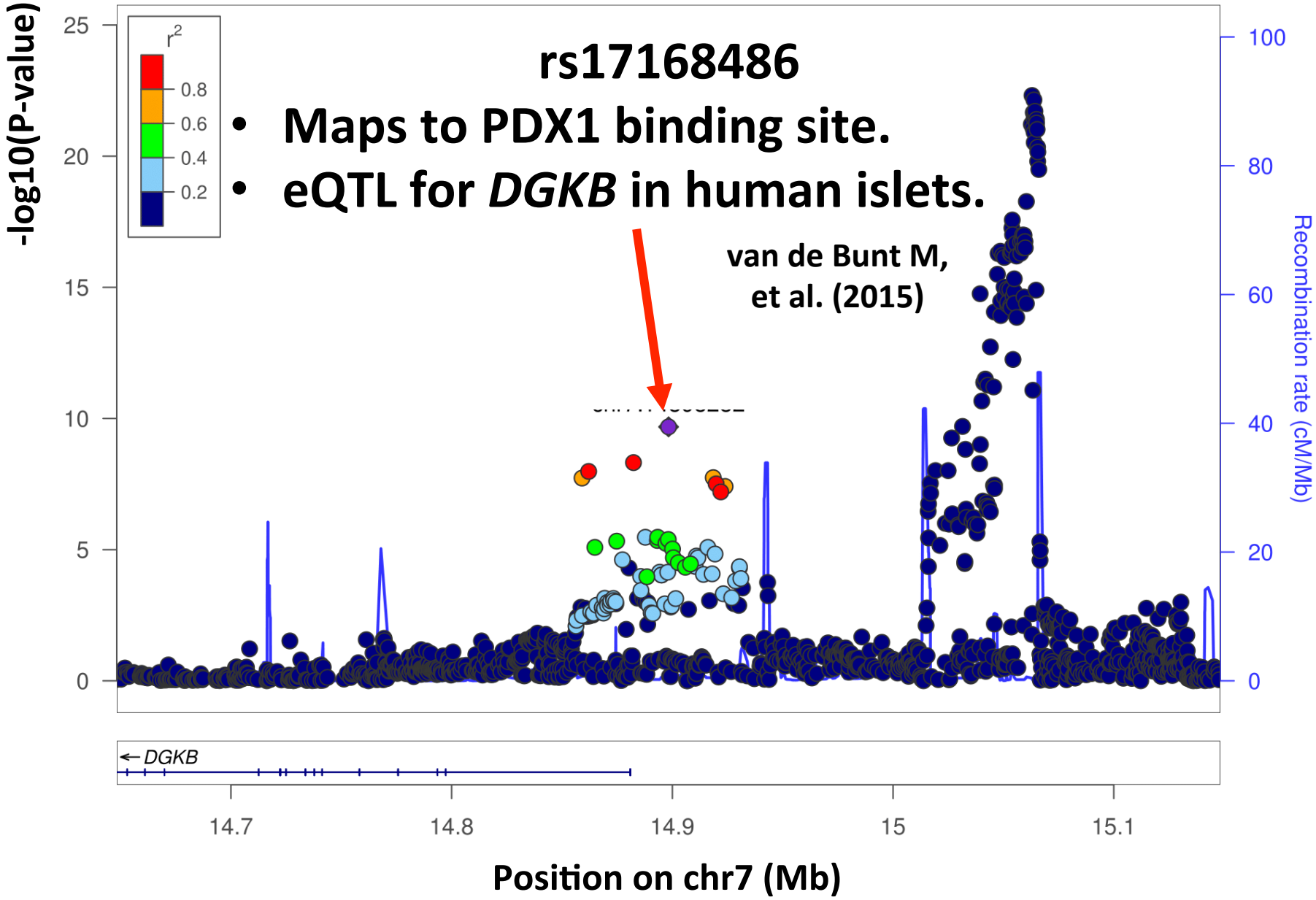
DGKB


-log10(P-value)
25
20
15
10
5
0
r2
0.8
0.6
0.4
0.2
100
80
60
40
20
0
Recombination rate (cM/Mb)
DGKB
14.7
14.8
14.9
15
15.1
Position on chr7 (Mb)

DGKB


rs17168486
Posterior probability
European 52%
Trans-ethnic 86%
-log10(P-value)
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
25
20
15
10
5
0
100
80
60
40
20
0
DGKB
14.7
14.8
14.9
15
15.1
Position on chr7 (Mb)

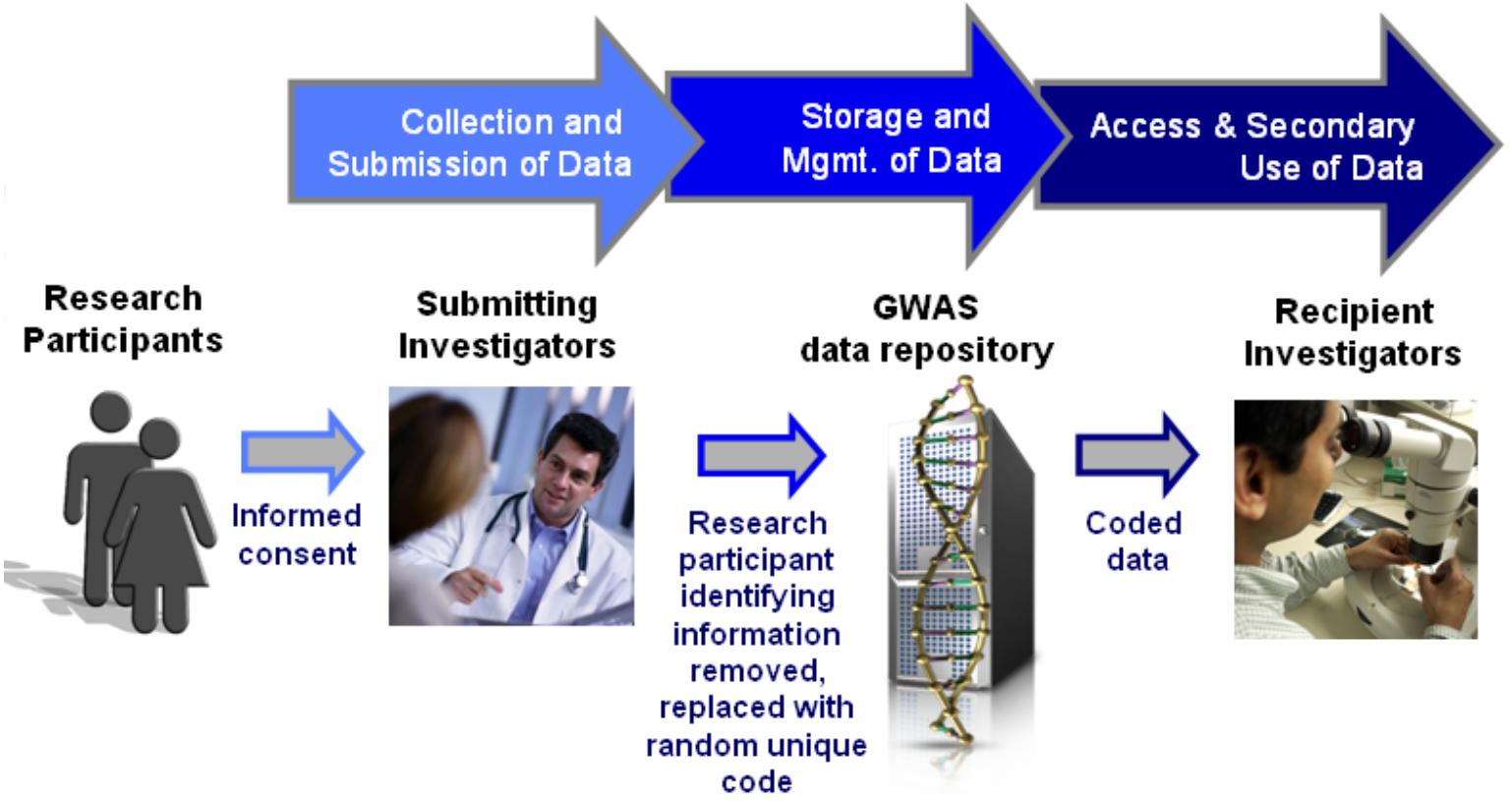
DGKB


rs17168486
• Maps to PDX1 binding site.
• eQTL for *DGKB* in human islets.
van de Bunt M, et al. (2015)
-log10(P-value)
Recombination rate (cM/Mb)
r²
0.8
0.6
0.4
0.2
25
20
15
10
5
0
100
80
60
40
20
0
←DGKB
14.7
14.8
14.9
15
15.1
Position on chr7 (Mb)

Summary

- Trans-ethnic meta-regression identified 40 novel loci for T2D susceptibility.
- Distinct association signals at several T2D loci, including 11 at *KCNQ1* and 5 at *INS-IGF2*.
- First strong evidence of heterogeneity in allelic effects between ancestry groups.
- Substantially improved fine-mapping resolution.
- Enrichment of T2D association signals in coding exons and TFBS for PDX1 and FOXA2.
- All this was only possible because we brought together data for more than **1,250,000** people in a trans-ethnic analysis… what is the path for other phenotypes?

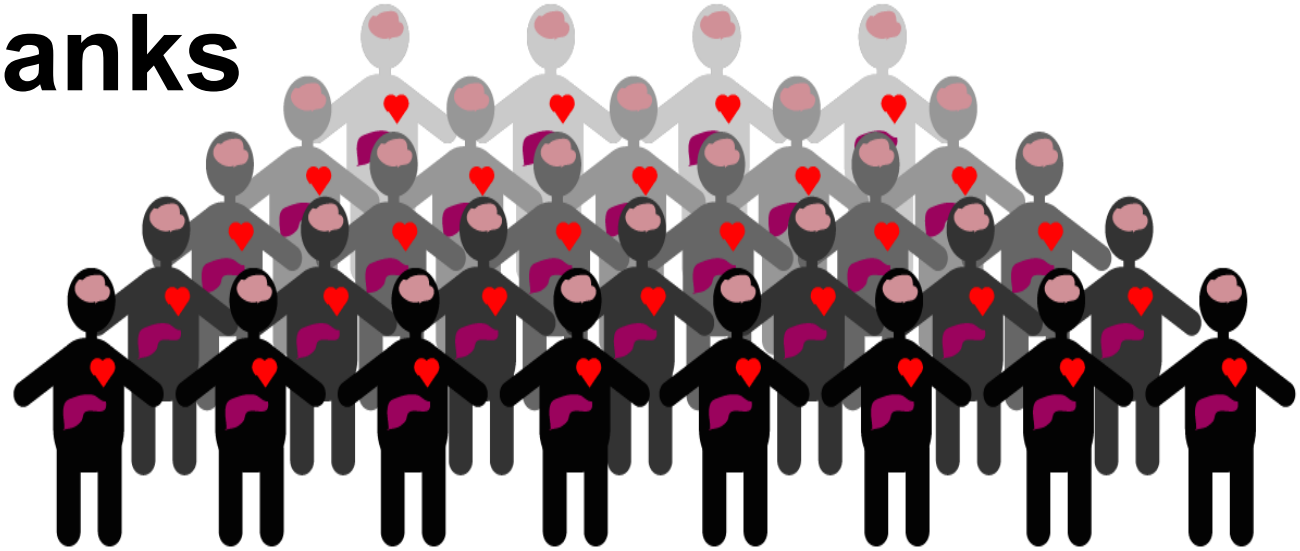
GWAS Repository Process



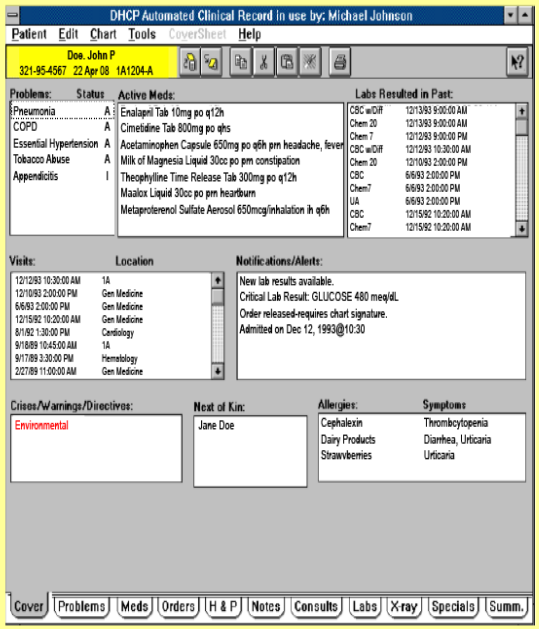
What can we do uniquely well in biobank data?

How can biobank strengths improve understanding of mechanisms of disease?

Biobanks



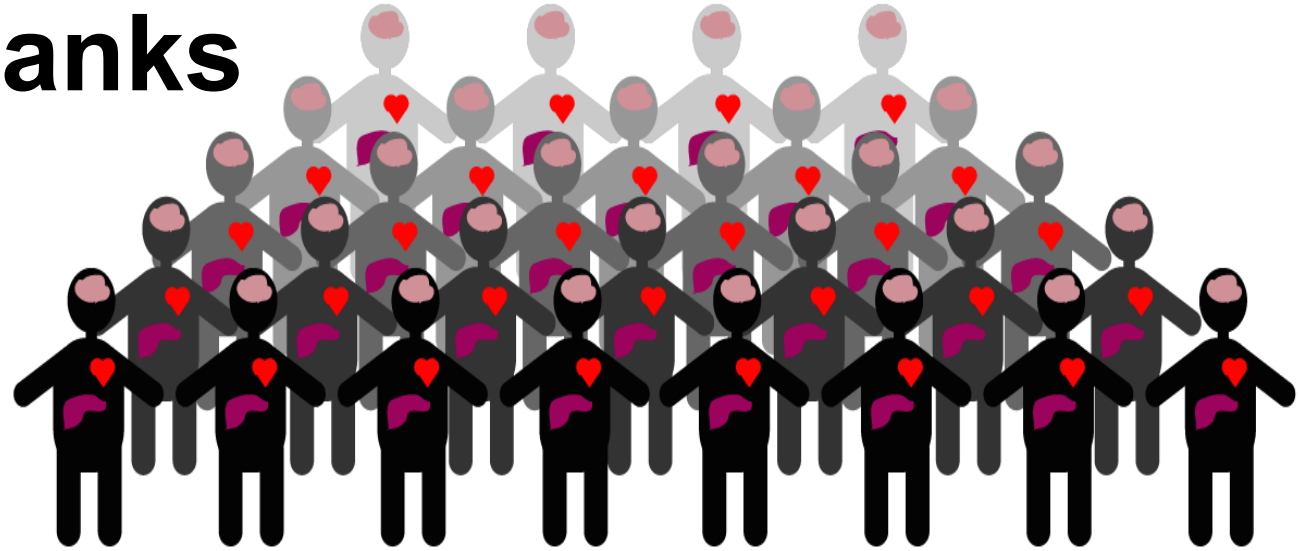
EHR



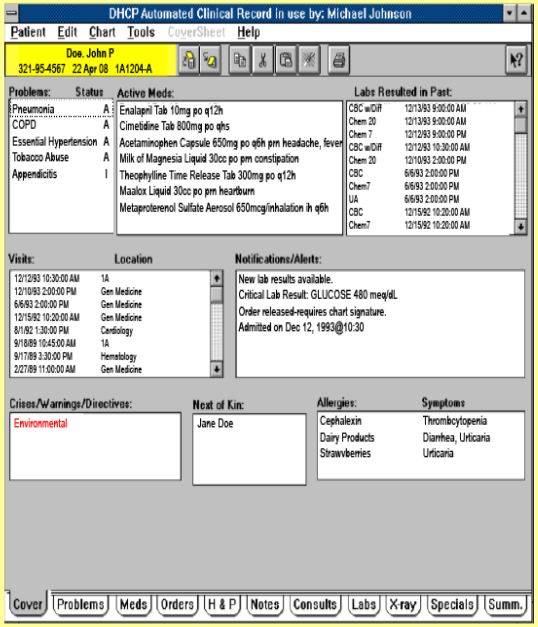
-OMICS

- Genome variation
- Transcriptomes
- Metabolomes
- Methylomes

Biobanks



EHR

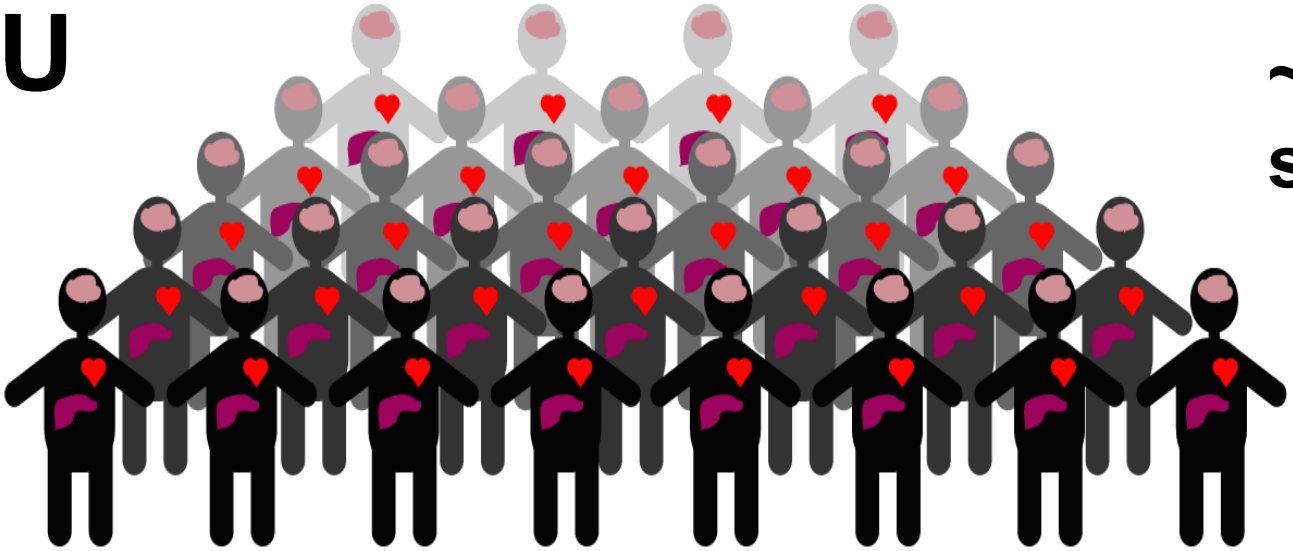


-OMICS

- Genome variation
- Transcriptomes
- Metabolomes
- Methylomes

Measure
Impute

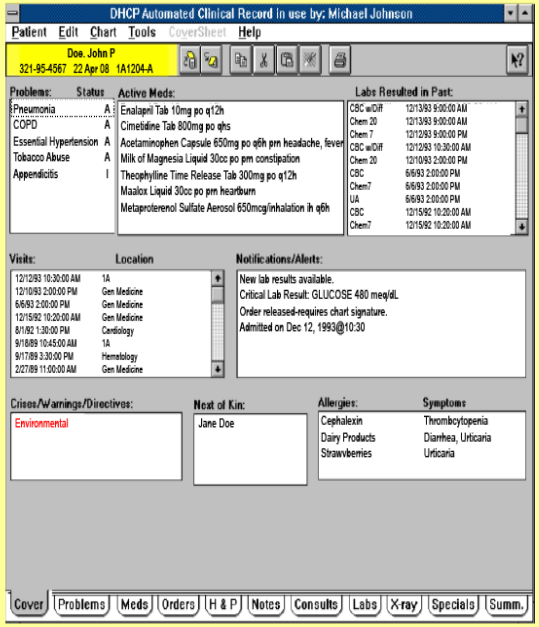
BioVU



~250,000 samples

~10-15 yrs of EHR

EHR



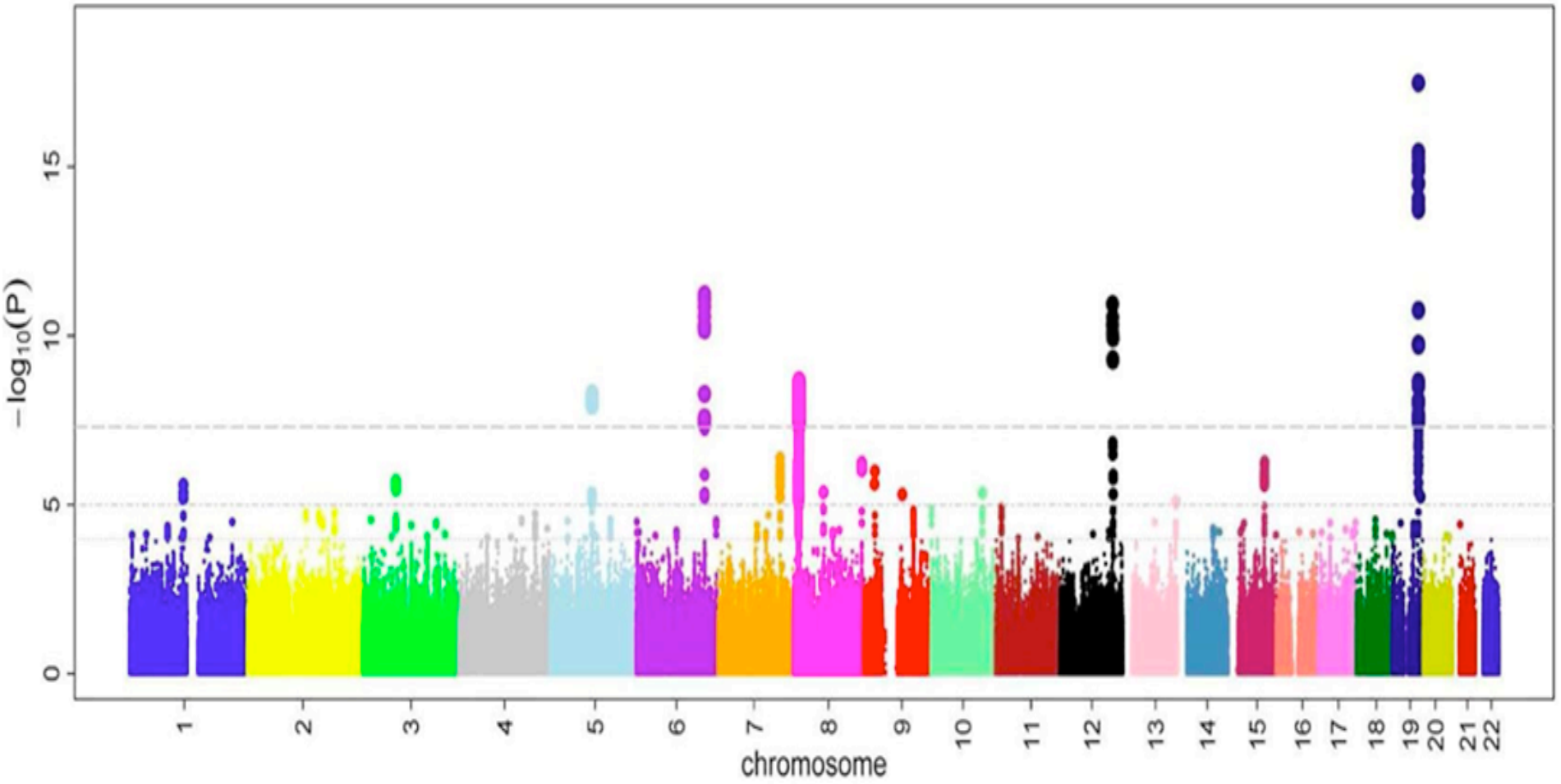
-OMICS

- Genome variation
- Transcriptomes
- Metabolomes

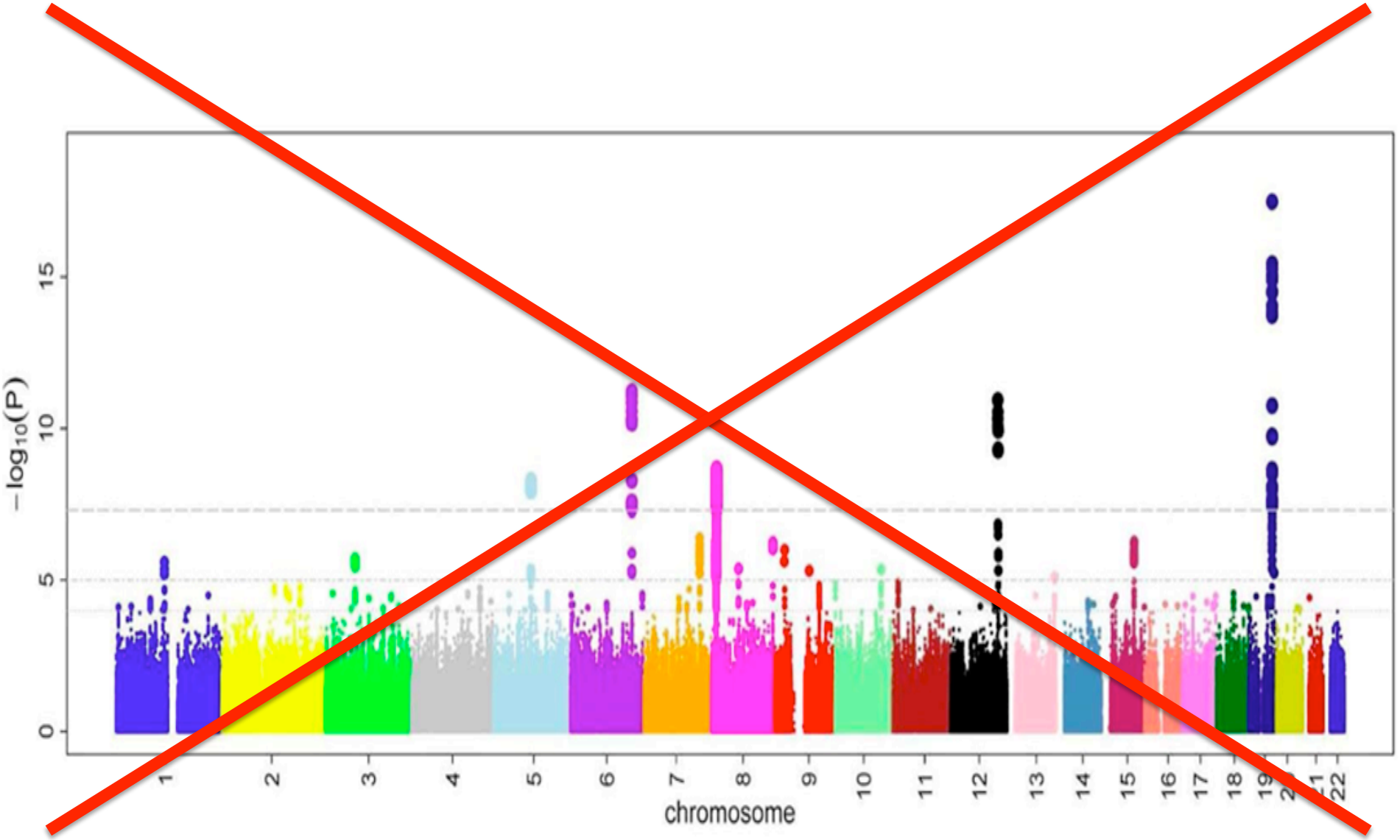
120K

Imputed

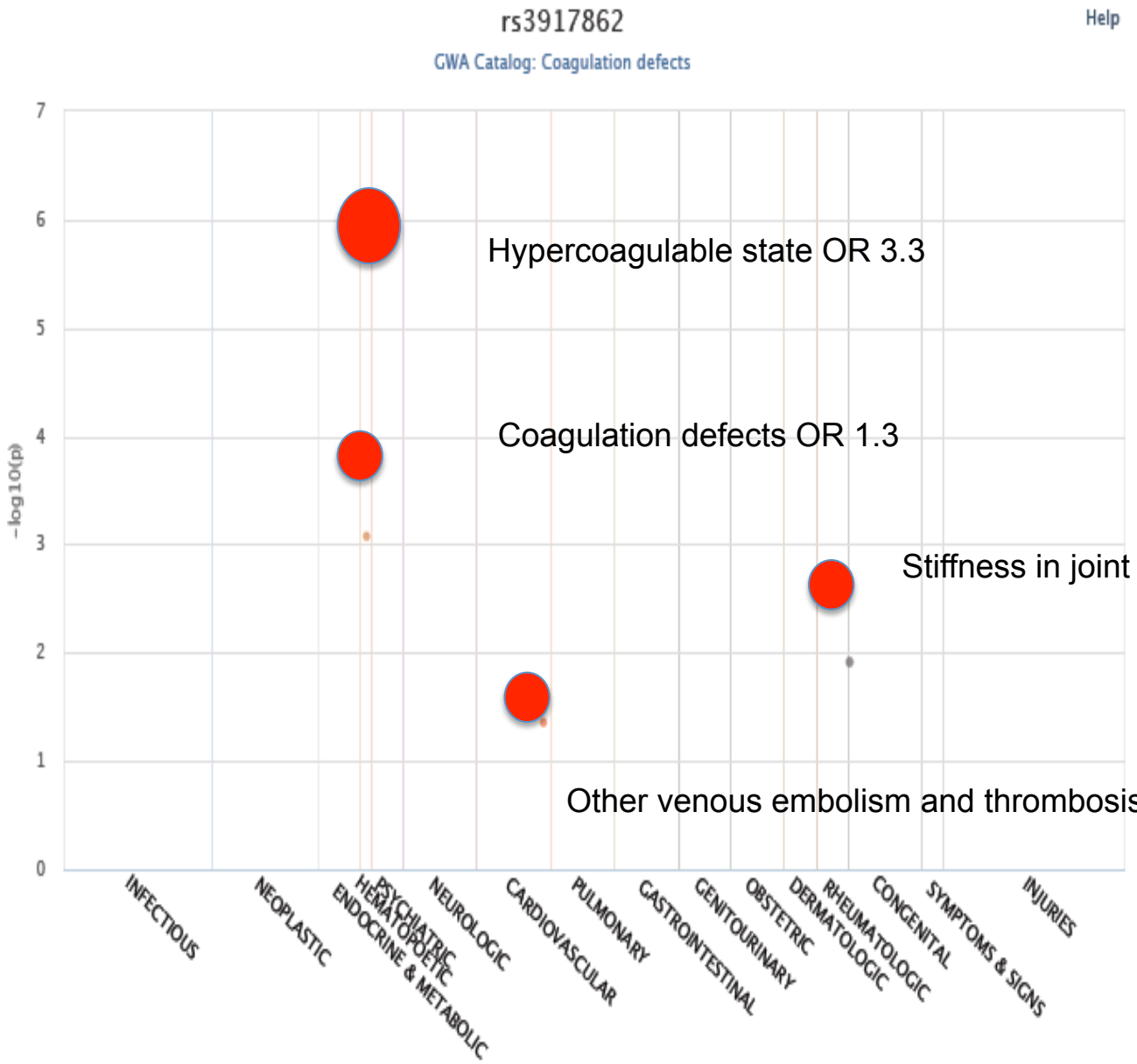
What can we do uniquely well in BioVU?



What can we do uniquely well in BioVU?



What can we do uniquely well in BioVU?



Phenome-wide Association Study (PheWAS) of Neanderthal Variants

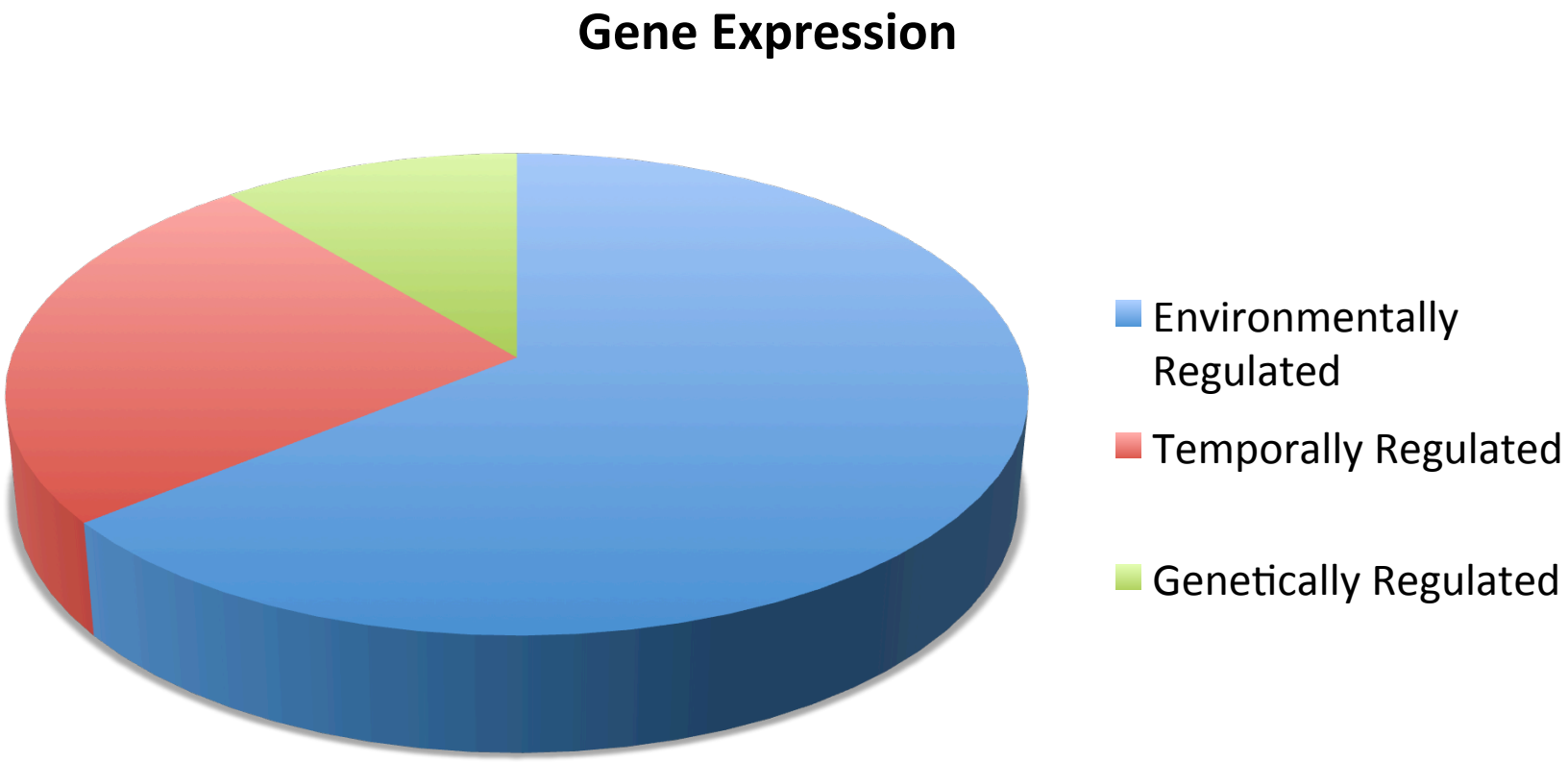
GWAS: What variants are associated with this phenotype?

PheWAS: What phenome is associated with this variant?

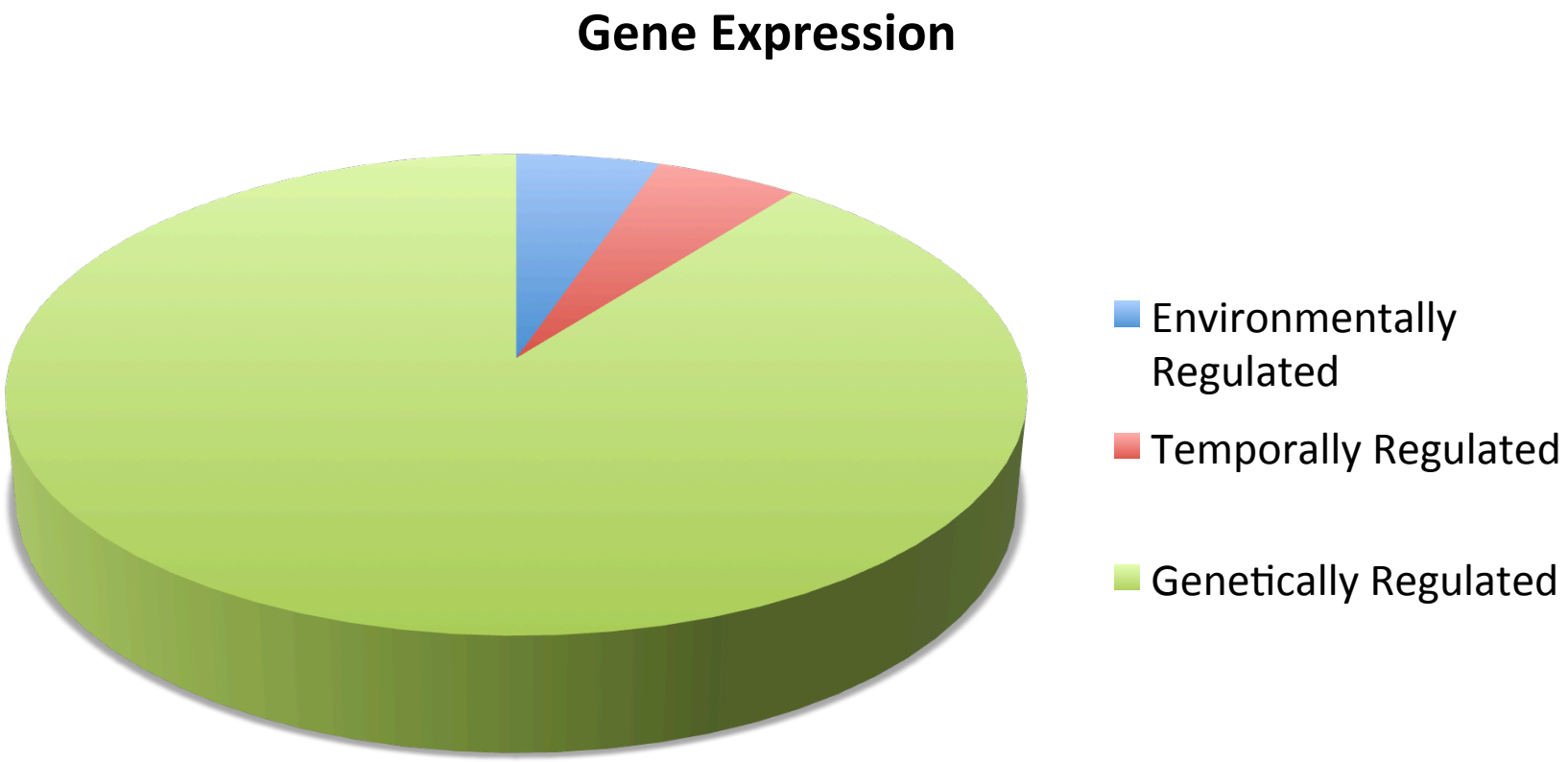
Problems with GWAS

- Common variants (as opposed to non-synonymous coding variation) tend to be uncoupled from gene activity- hard to interpret
- Burden of test correction is high
- Effect sizes per SNP are usually very small
- Association signals tend to spans multiple genes either or no genes

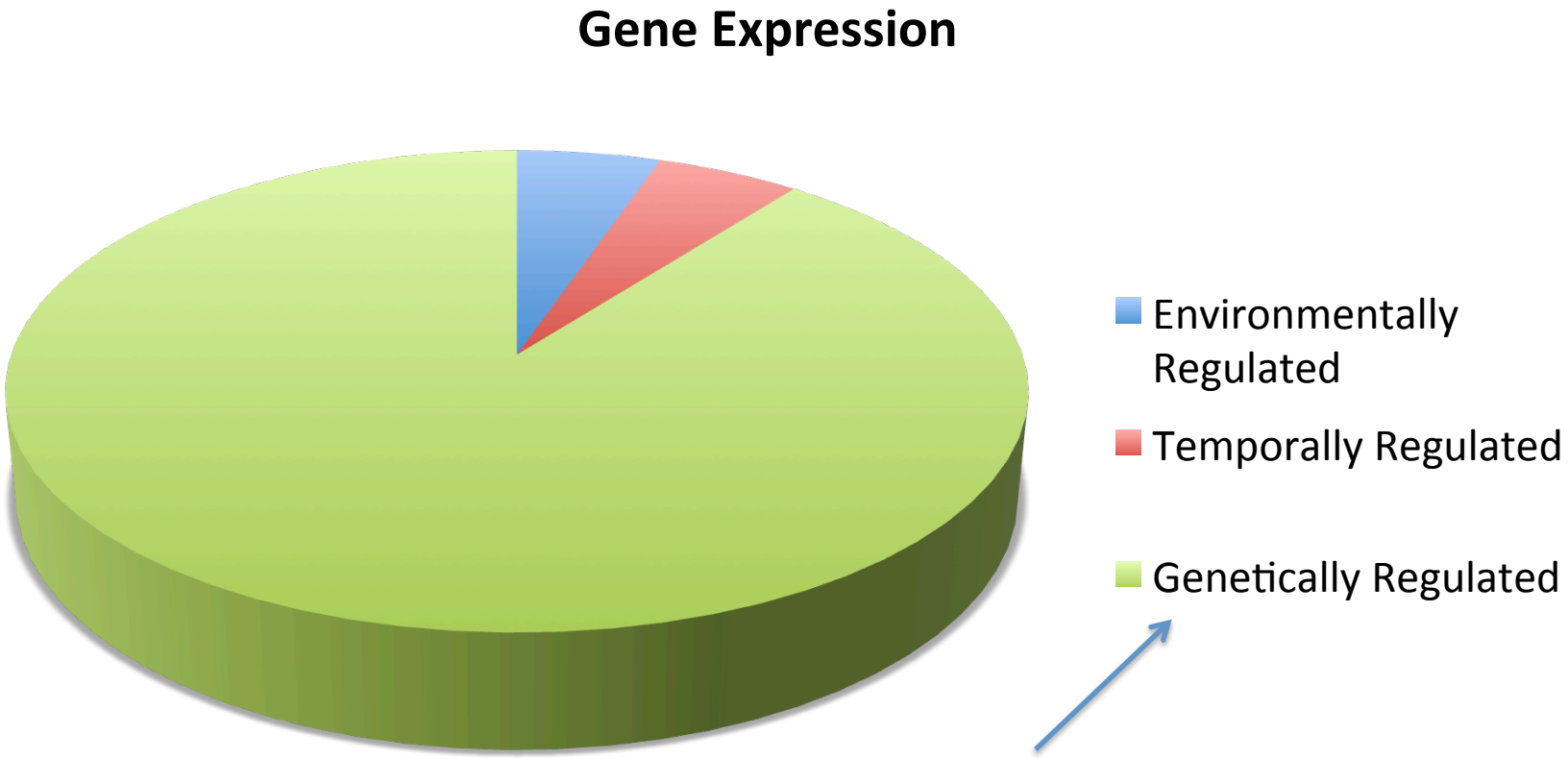
Thinking outside of the SNP



Thinking outside of the SNP

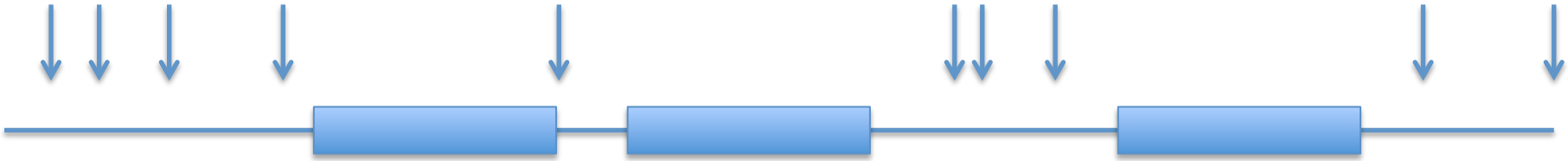


Thinking outside of the SNP

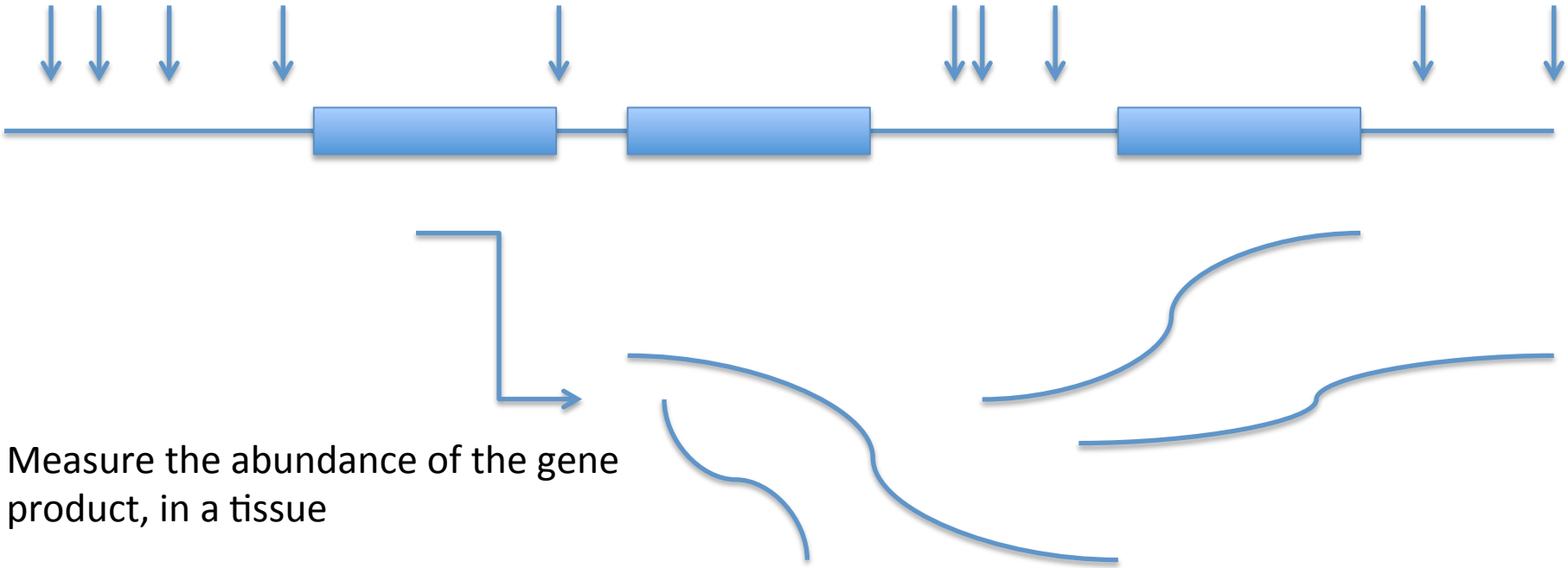


Maybe we can estimate this from genotypes

Intro to PrediXcan



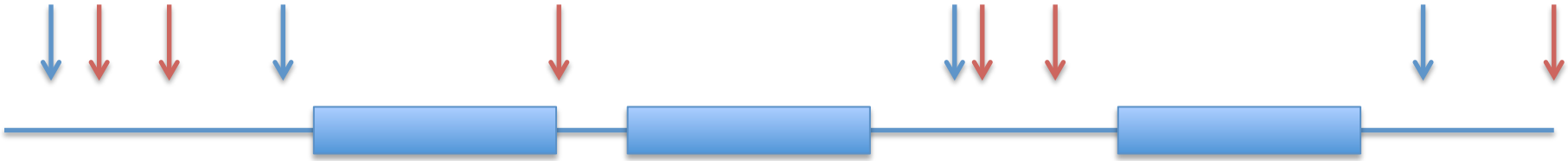
Intro to PrediXcan



Measure the abundance of the gene product, in a tissue

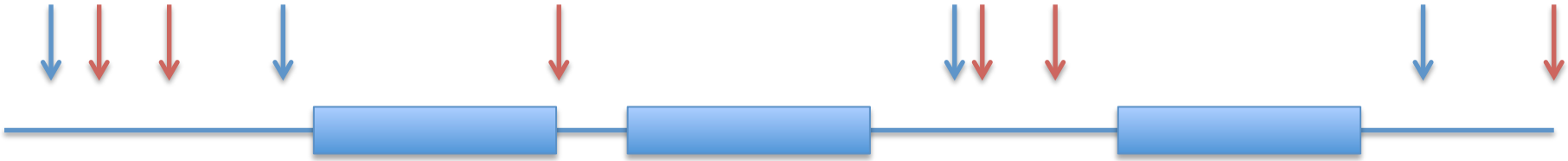
Intro to PrediXcan

Are associated with expression in some tissue



Intro to PrediXcan

Are associated with expression in some tissue



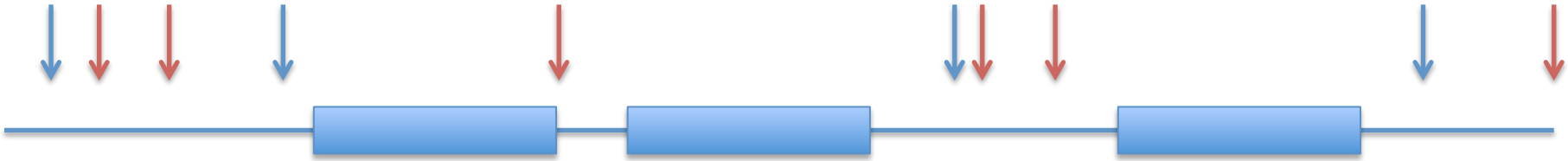
Additive model of gene expression trait trained in reference transcriptome data sets

$$T = \underbrace{\sum_k w_k X_k}_{\text{GReX}} + \varepsilon$$

Weights stored in PredictDB

Intro to PrediXcan

Are associated with expression in some tissue



Additive model of gene expression trait trained in reference transcriptome data sets

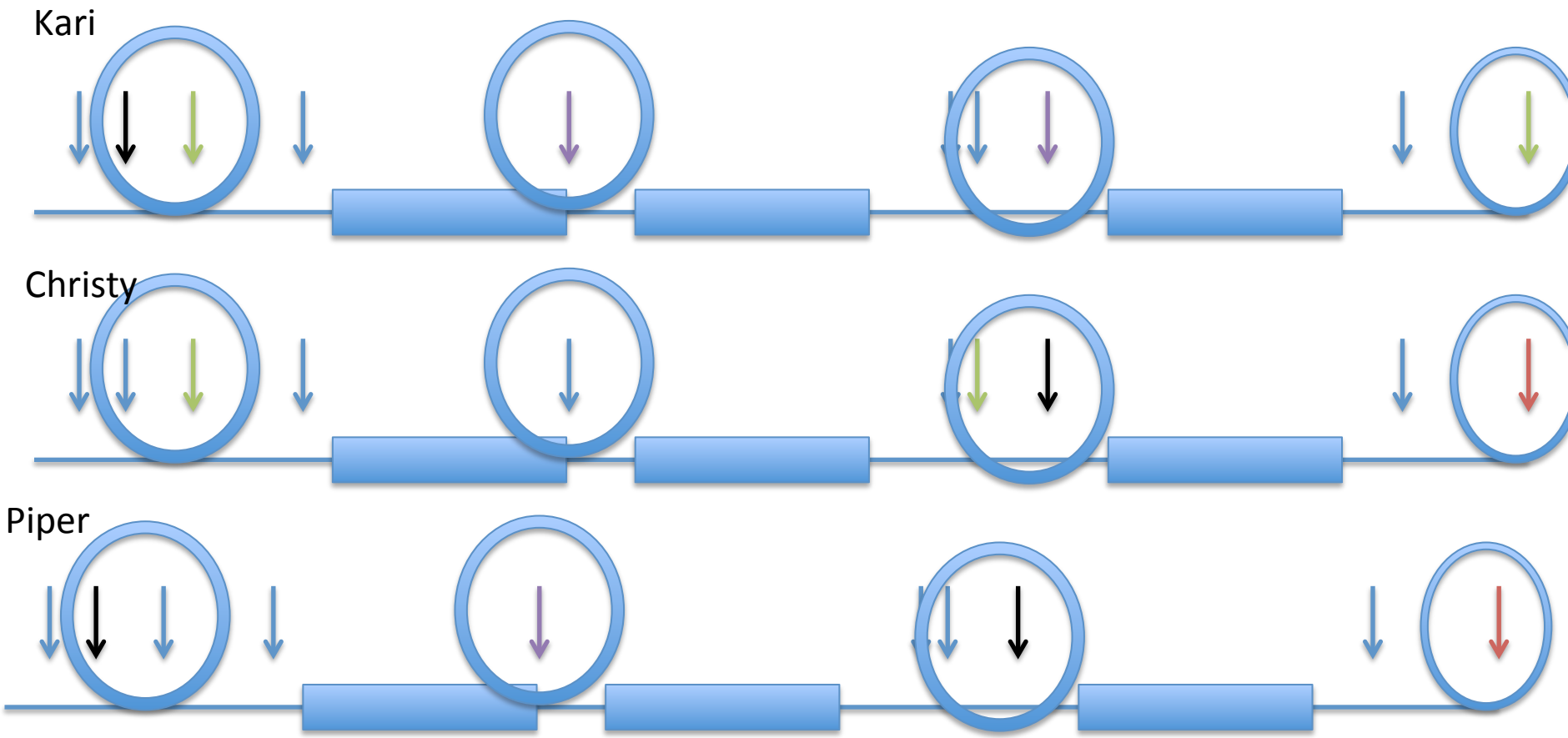
$$T = \sum_k w_k X_k + \varepsilon$$

GReX

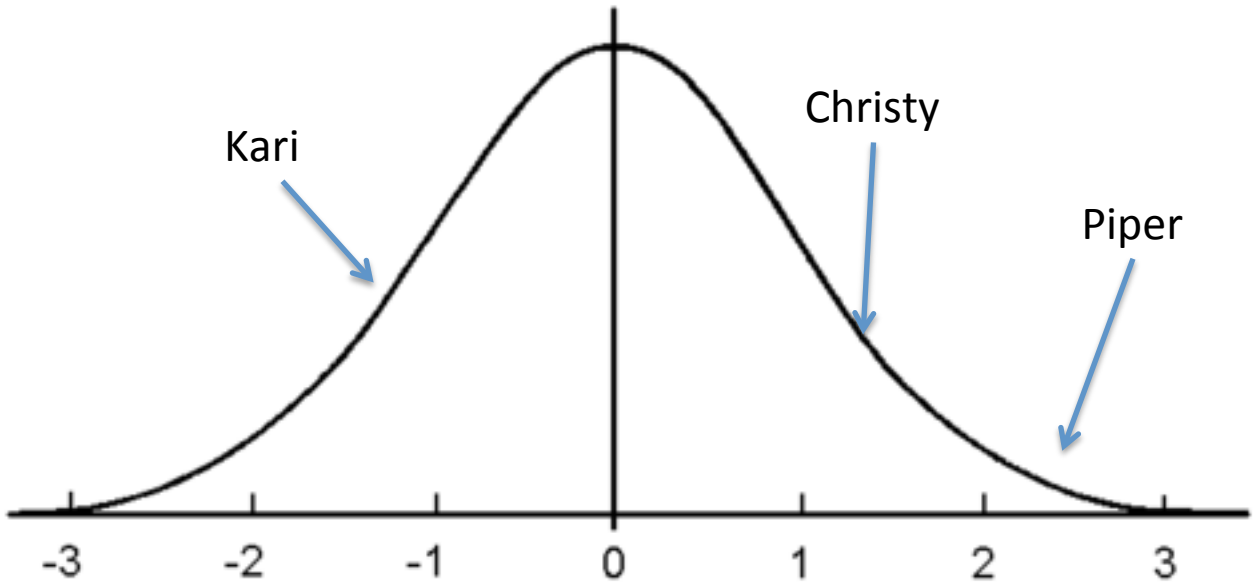
Weights stored in PredictDB

Done, in ~41 tissues in the GTEx project

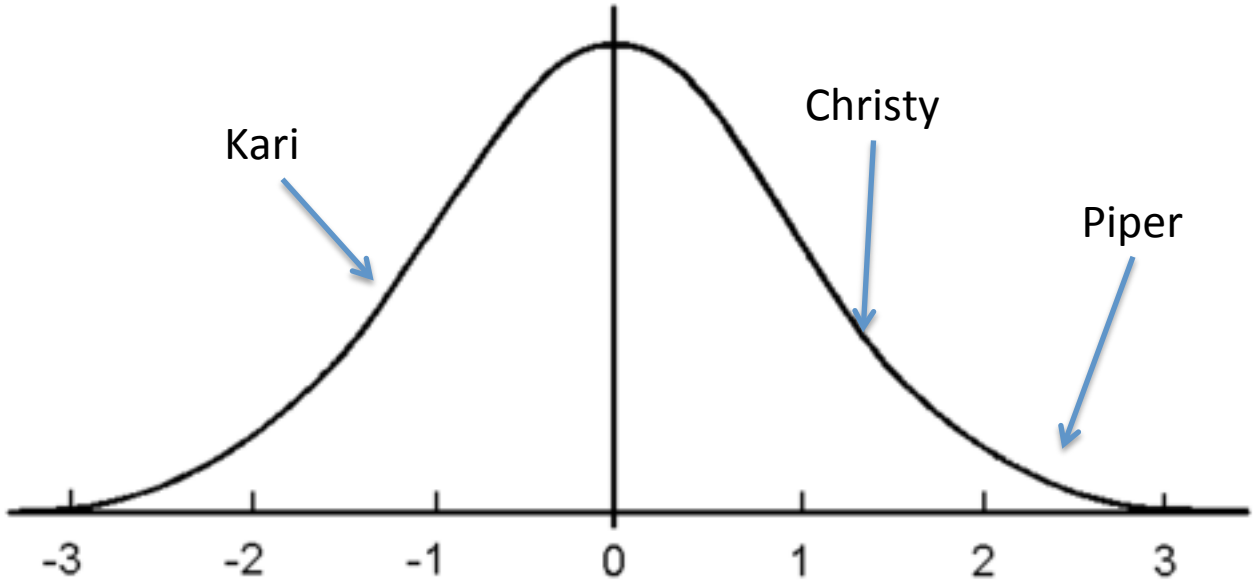
Intro to PrediXcan

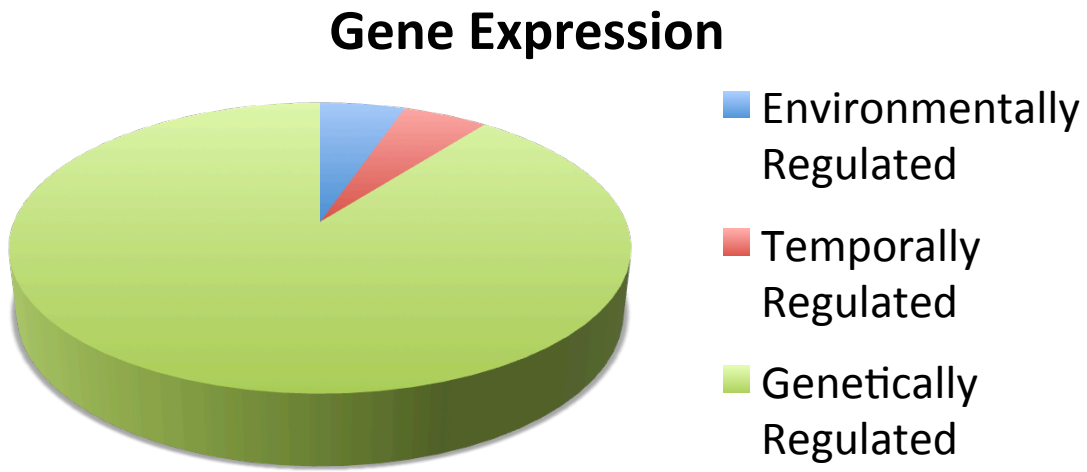


Intro to PrediXcan

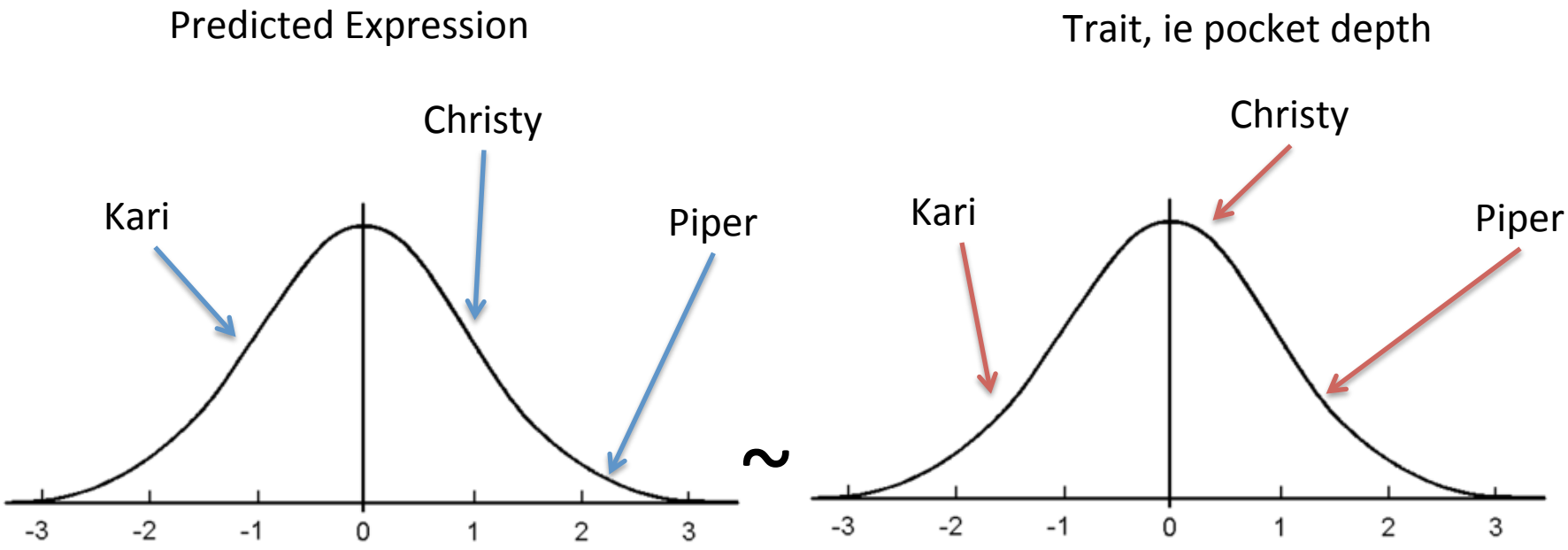


Intro to PrediXcan



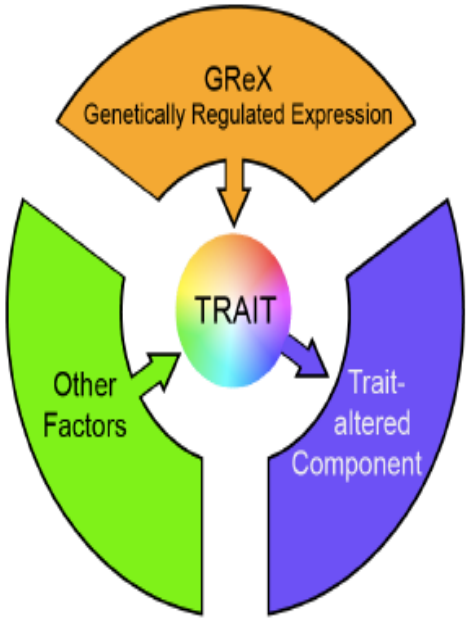


Intro to PrediXcan



Expression driven gene-based analyses

- Directly interpretable, functionally oriented results
- Reduce multiple testing burden from millions of SNPs to thousands of genes
- Aggregate statistical evidence over a large number of variants which each might contribute only a small effect



PrediXcan

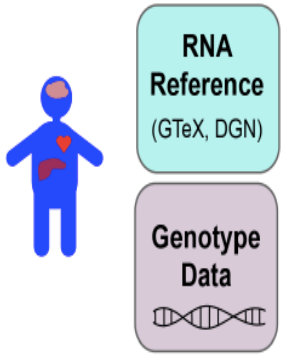
(Gamazon et al., 2015, Nat Genet)

https://github.com/hakyimlab/PrediXcan

Reference panel: GTEx

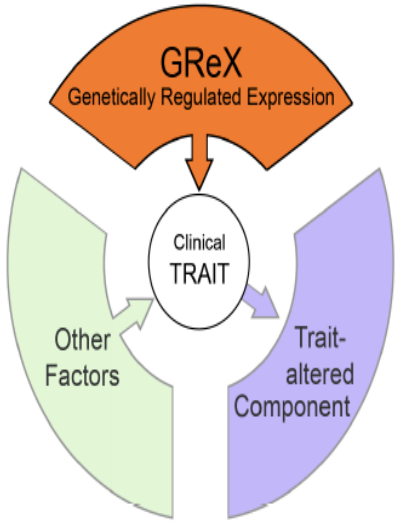
GTEx: WGS and RNA-Seq in 44 tissues from ~450 (>950) subjects

A. Reference Panel of Measured Transcriptome & Genome Variation



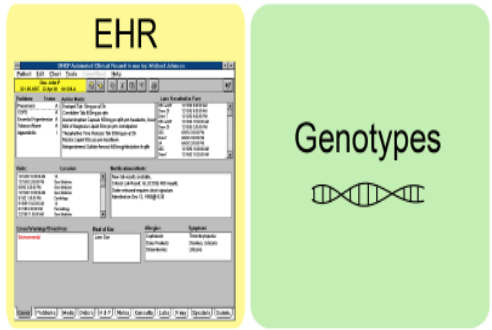
Transcriptome Prediction

B. Genetic Prediction of Gene Expression



High quality prediction performance in 18,483 genes and lncRNAs

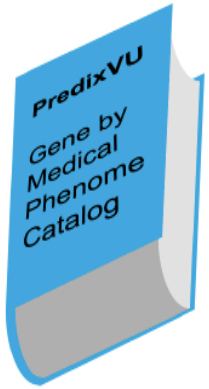
PrediXcan Application to BioVU



C. Biobank with Genotype Data
(Whole Genome Sequence or Dense Genotyping Array)
Linked to Electronic Health Recods (EHR)

BioVU: >250K subjects, linked to EHR going back on average 10-15 years; ~50K with genome interrogation, >120K by 2018

Creation of PredixVU



D. Results Portal
for Query by Gene or Phenome

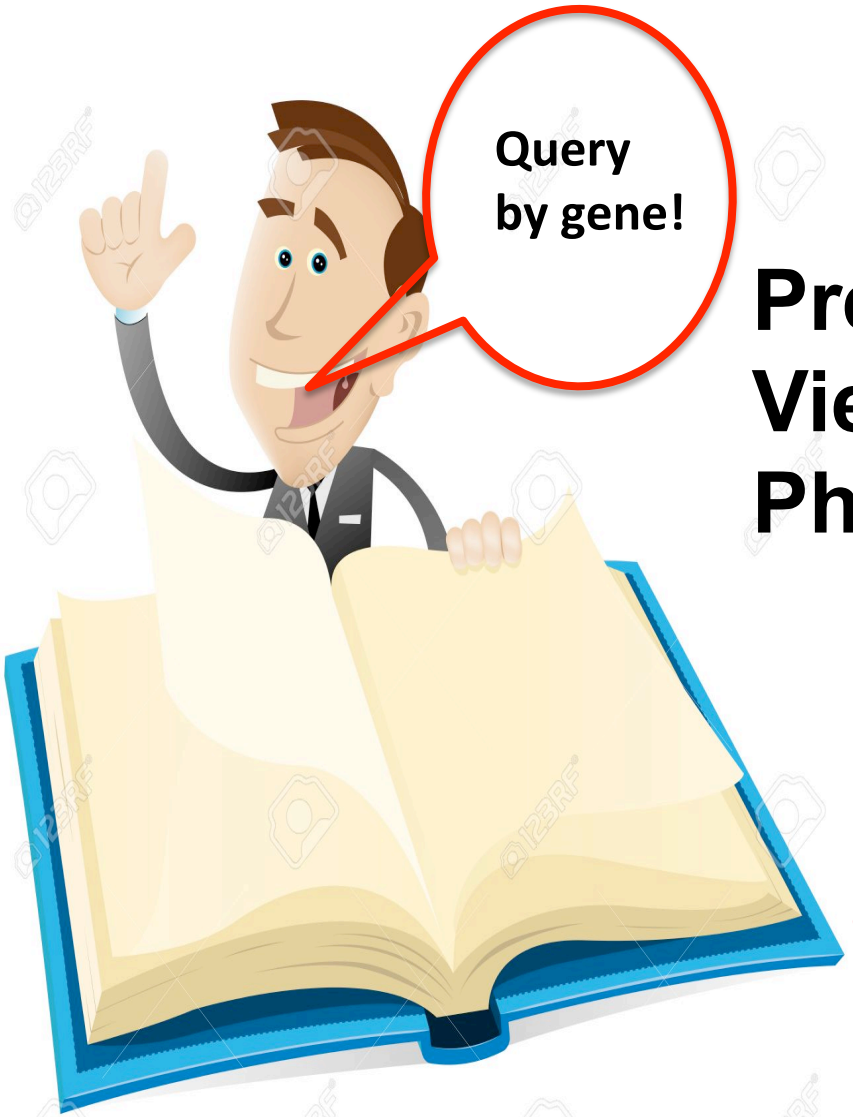
PredixVU catalog is a discovery engine for gene-phenotype relationships and iteration to primary mechanism of disease

Gene-based PheWAS

What does this gene do?

What does the natural variation in the expression of this gene associate with across the medical phenome?

Genome X Transcriptome X EHR



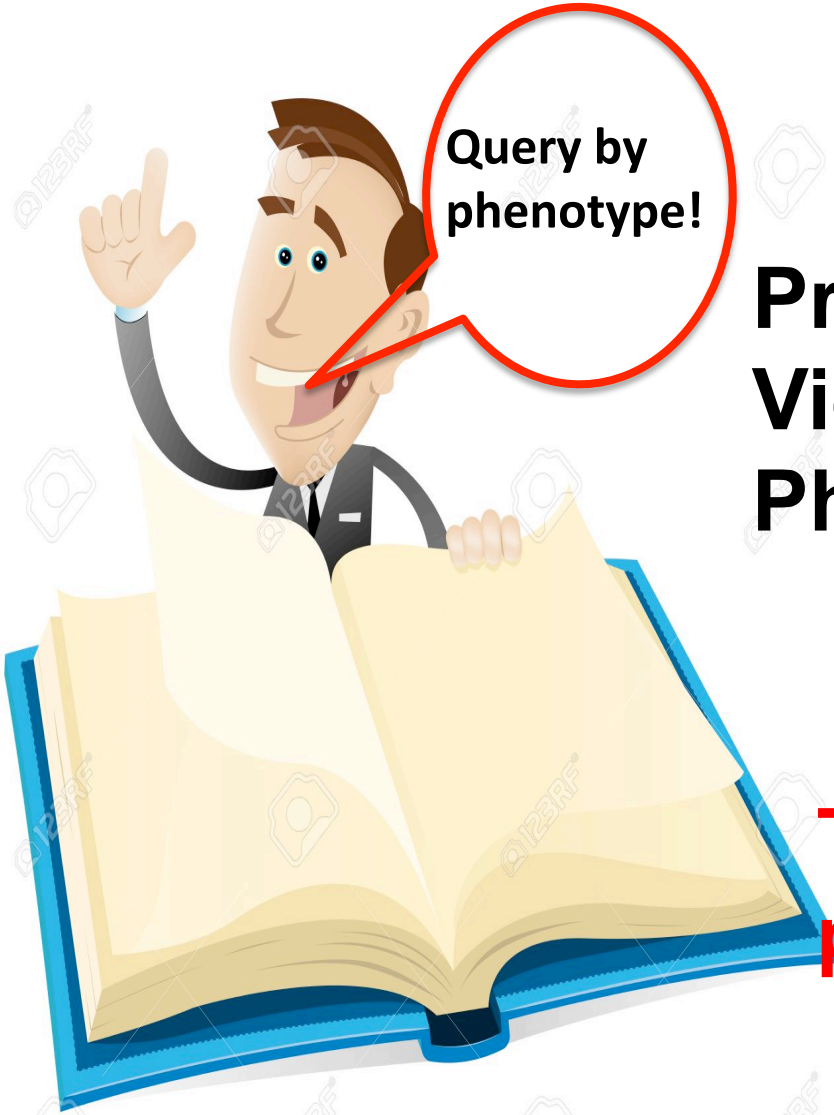
PredixVU: A Catalog for Viewing Gene-Based Phenome-Wide Association

… or gene set, pathway, network …

Reduced GReX *CFTR*

Tissue	P-value	Trait	Cases	Controls
Brain-Hypothalamus	2.32E-39	Cystic fibrosis	71	9033
Heart-LeftVentricle	5.37E-39	Pseudomonal pneumonia	105	6217
Heart-LeftVentricle	2.04E-32	Bronchiectasis	124	6820
Heart-LeftVentricle	8.39E-30	Other disorders of metabolism	88	8608
Heart-LeftVentricle	4.16E-27	MRSA pneumonia	82	6217
Brain-Hypothalamus	1.51E-26	Pseudomonal pneumonia	105	6217
Heart-LeftVentricle	2.13E-21	Bronchopneumonia/lung abscess	71	6217
Heart-LeftVentricle	2.46E-21	Diseases of pancreas	337	8624
Heart-LeftVentricle	4.94E-19	Bacterial pneumonia	385	6217
Brain-Hypothalamus	2.15E-17	Diseases of pancreas	337	8624
Heart-LeftVentricle	1.56E-13	Chronic sinusitis	589	6193
Heart-LeftVentricle	3.67E-13	Nutritional marasmus	72	6138
Heart-LeftVentricle	2.06E-12	Failure to thrive (childhood)	73	6138
Heart-LeftVentricle	4.53E-11	Secondary diabetes mellitus	80	4936
Heart-LeftVentricle	5.58E-11	Intestinal malabsorption non-celiac	72	5956
Heart-LeftVentricle	6.63E-11	Nasal polyps	49	6193

Genome X Transcriptome X EHR



PredixVU: A Catalog for Viewing Gene-Based Phenome-Wide Association

Test pleiotropy, investigate phenome relationships, …

Query by Phenome: Gingival and periodontal diseases

tissue	beta	p-value	gene
Cells-Transformedfibroblasts	-0.690925	4.78E-07	CTNS
Brain-CerebellarHemisphere	-0.84831	6.35E-07	CTNS
Muscle-Skeletal	-1.69745	7.39E-07	CTNS
Brain-Cortex	-1.00725	7.53E-07	CTNS
Heart-AtrialAppendage	20.4918	7.64E-07	PELI3
Brain-Caudate-basalganglia	-1.12867	8.11E-07	CTNS
Brain-Cerebellum	-0.971678	8.56E-07	CTNS
Brain-Anteriorcingulatecortex-BA24	-3.15379	8.57E-07	CTNS
Brain-Hypothalamus	-1.19931	9.34E-07	CTNS
Brain-Cerebellum	-0.0742653	1.09E-06	MC1R
Adipose-Subcutaneous	-33.3475	1.11E-06	RPL41
Brain-Nucleusaccumbens-basalganglia	-0.909724	1.12E-06	CTNS
Brain-FrontalCortex-BA9	-0.799975	1.21E-06	CTNS
Heart-AtrialAppendage	-1.04589	1.44E-06	CTNS
Lung	47.1514	1.48E-06	OR6C2
SmallIntestine-TerminalIleum	-0.845769	1.58E-06	CTNS
Esophagus-GastroesophagealJunction	-0.780144	1.63E-06	CTNS
Liver	-1.036	1.78E-06	CTNS
Heart-AtrialAppendage	10.8251	1.82E-06	INHBC
Thyroid	-0.671328	1.87E-06	CTNS

CTNS **-** lysosomal cystine transporter

- **Ascertained in PrediXcan on BioVU for gingival and periodontal diseases**
- **In BioVU, also associated with**
  - **Gingivitis**
  - **Dermatophytosis / Dermatomycosis**
  - **Influenza**

Query by Phenome: Dental caries

tissue	beta	p-value	gene
SmallIntestine-TerminalIleum	0.078875	1.01E-08	INVS
Brain-Cortex	-1.59343	1.20E-07	COX6A2
Breast-MammaryTissue	-25.1841	7.55E-07	LRP2
Brain-CerebellarHemisphere	2.2705	7.85E-07	USP33
Esophagus-Mucosa	0.528235	1.38E-06	NDUFA6
Muscle-Skeletal	0.665914	1.57E-06	NDUFA6
Adipose-Subcutaneous	0.497227	2.32E-06	NDUFA6
Heart-AtrialAppendage	2.00369	2.89E-06	DCC
Cells-EBV-transformedlymphocytes	-19.514	2.89E-06	XRCC6
Brain-Cerebellum	2.01339	3.40E-06	POLH
Brain-CerebellarHemisphere	0.594764	4.24E-06	NDUFA6
Artery-Tibial	0.430468	5.57E-06	NDUFA6
Skin-NotSunExposed-Suprapubic	0.486097	5.89E-06	NDUFA6
Artery-Coronary	-1.6148	6.39E-06	RANGAP1
SmallIntestine-TerminalIleum	0.956041	7.42E-06	NDUFA6

Where do we go from here?

- **Comparing PrediXcan results for BioVU and UKBioBank**
  - **Seeking candidates for functional studies!**

INVS - inversin, cilia function

- **Ascertained in PrediXcan on BioVU for dental caries**
- **In BioVU, also associated with**
  - **Diseases of hard tissues of teeth**
  - **Gingivitis**
  - **Noninfectious gastroenteritis**
  - **Other local infections of skin and subcutaneous tissue**
  - **Gingival and periodontal diseases**
  - **Functional digestive disorders**
- **In UKBioBank associated with**
  - **Hip circumference**
  - **Weight**
  - **BMI**
  - **Leg fat mass**
  - **Metabolic rate**

COX6A2 **-** Cytochrome c oxidase (COX), the terminal enzyme of the mitochondrial respiratory chain

- **Ascertained in PrediXcan on BioVU for dental caries**
- **In BioVU, also associated with**
  - **Diseases of hard tissues of teeth**
  - **Nausea and vomiting**
  - **Congenital anomalies of intestine**
  - **Diseases of the jaws**
  - **Other disorders of intestine**
- **In UKBioBank associated with**
  - **Waist circumference**
  - **Arm fat percentage**
  - **BMI**
  - **Leg fat mass**

NDUFA6 **-** enzyme of the mitochondrial membrane respiratory chain

- **Ascertained in PrediXcan on BioVU for dental caries**
- **In BioVU, also associated with**
  - **Diseases of hard tissues of teeth**
  - **Loss of teeth or edentulism**
  - **Other diseases of the teeth and supporting structures**
  - **Diabetes mellitus**
  - **Varicose veins of lower extremity**
- **In UKBioBank associated with**
  - **Body size at age 10**
  - **Arm fat**
  - **Impedance measures**
  - **Treatment with lipitor**

Summary

- Existing big data resources provide exciting opportunities to generate hypotheses for genetic epidemiology
- Ancestral heterogeneity is powerful when used correctly
- Sample sizes need to be large to find genetic effects in complex traits- this will mean a lot of data sharing
- Functionally orienting analyses can improve interpretation and power (fewer tests!)
- We really need to work on getting good oral health measures into EHRs and DNA databanks
- Integrating additional *omics (microbiome, measured transcriptome) and environment (diet, SES, education) will help us tackle GxE effects down the road

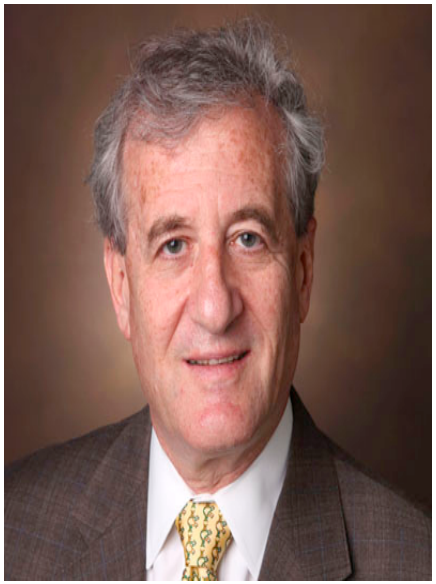
Thank you!

- My lab:
  - Lauren Petty
  - Hung-Hsin Chen

Acknowledgements

- **Participating cohorts, PIs and analysts:**
- **MEDIA (African): ARIC**: Nisa Maruthur, Mandy Li; **BioMe**: Ruth Loos, Claudia Schurmann, Michael Preuss; **CARDIA**: Myriam Fornage, Edmond Kabagambe; **CFS**: Sanjay Patel, Brian Cade; **CHS**: Bruce Psaty, David Siscovick, Rich Jensen; **eMERGE**: Geoff Hayes, Yoonjung Joo; **FamHS**: Ingrid Borecki , Ping An; **GeneSTAR**: Diane Becker, Lisa Yanek; **GENOA**: Lawrence Bielak, Patricia Peyser, Sharon Kardia; **HANDLS**: Michael Nalls, Salman Tajuddin; **Health ABC**: Gregory Tranah, Steve Cummings, Daniel Evans, Aude Nicolas; **HUFS/AADM**: Charles Rotimi, Daniel Shriner, Adebowale Adeyemo, Guanjie Chen; **JHS**: James Wilson, Leslie Lange, Laura Raffield; **MESA**: Jerome Rotter, Xiuqing Guo, Ida Chen, Jie Yao; **SIGNET-REGARDS**: Michele Sale, Wei-Min Chen, Mary Cushman; **WFSM**: Donald Bowden, Maggie Ng, Poorva Mudgal, Jacob Keaton, Barry Freedman; **WHI**: Simin Liu, Brian Chen, Katie Chan, Ian Pan; **Zulu**: Mark McCarthy, Anubha Mahajan, Meng Sun.
- **AGEN-T2D (East Asian): BES**: Jost Jonas, Yaxing Wang; **CAGE-Amagasaki CAGE-GWAS**: Fumihiko Takeuchi, Norihiro Kato; **CAGE-KING**: Masahiro Nakatochi, Mitsuhiro Yokota, Norihiro Kato; **CHNS**: Cassie N. Spracklen Karen L. Mohlke; **CLHNS**: Ying Wu, Karen L. Mohlke; **EHIME NAGAHAMA**: Takahisa Kawaguchi, Yasuharu Tabara; **HKDR**: Claudia Tam, Ronald Ma, Juliana Chan; **KARE**: Young Jin Kim; Sanghoon Moon, Bong-Jo Kim; **MESA**: Jie Yao, Xiuqing Guo, Jerry Rotter; **SBCS SWHS**: Jirong Long, Xiao-ou Shu; **SCES SiMES**: Ching-Yu Cheng, Tien-Yin Wong; **SCHS**: Mark Pereira, Myron Gross, Woon Puay Koh, Jian Min Yuan; **DC/SP2**: Xueling Sim, E-Shyong Tai; **SNUH**: Soo-Heon Kwak, Kyong Soo Park; **SSH**: Yoon Shin Choo; **TaiChi-G**: Xiuqing Guo, Ida Chen, Jerry Rotter, Wayne Sheu; **TWT2D**: Chien-Hsiun Chen, Li-Chang Chang; **BBJ**: Momoko Hirokoshi.
- **DIAGRAM (European): BIOME**: Ruth Loos, Michael Preuss; **deCODE**: Valgerður Steinþórsdóttir, Unnur Þorsteinsdóttir; **DGDG**: Cecile Lecoeur, Philippe Froguel; **DGI**: Anubha Mahajan, Mark I McCarthy, Leif Groop; **EGCUT**: Andrew Morris, Andres Metspalu; **FHS**: Achilleas N Pitsillides, Josee Dupuis; **FUSION**: Daniel Taliun, Michael Boehnke; **GCKD**: Matthias Wuttke, Anna Köttgen; **GENOA**: Lawrence Bielak, Patricia Peyser; **GERA**: James Cook, Andrew Morris; **GoDARTS**: Anubha Mahajan, Colin Palmer; **GOMAP**: William Rayner, Eleftheria Zeggini; **HPFS**: Andrew Morris; **INTERACT**: Jian'an Luan, Claudia Langenberg; **KORA**: Clemens Baumbach, Harald Grallert; **MESA**: Jie Yao, Xiuqing Guo, Jerry Rotter; **METSIM**: Daniel Taliun, Michael Boehnke, Markku Laakso; **NHS**: Andrew Morris; **NUGENE**: Andrew Morris; **PIVUS**: Anubha Mahajan, Andrew Morris, Lars Lind; **RS1/RS2/RS3**: Jana Nano, Abbas Dehghan; **UK Biobank**: Anubha Mahajan, Mark I McCarthy; **ULSAM**: Anubha Mahajan, Andrew Morris, Erik Ingelsson; **WTCCC**: Anubha Mahajan, Mark I McCarthy.
- **MA-T2D, SIGMA (Hispanic/Latino): BioMe**: Claudia Schurmann, Ruth J.F. Loos; **GOLDR**: Yang Hai, Jerome I. Rotter; **HCHS/SOL**: Misa Graff, Kari E. North; **HTN-IR**: ; Yang Hai, Jerome I. Rotter **LALES**: Darryl Nousome, Roberta McKean-Cowdin; **MACAD**: Yang Hai, Jerome I. Rotter; **MESA/MESA Family**: Jie Yao, Jerome I. Rotter; **Mexico City:** Lauren E. Petty, Jennifer E. Below, Esteban J. Parra; **NIDDM**: Yang Hai, Jerome I. Rotter; **SIGMA T2D**: Josep M. Mercader, Jose Florez; **Starr County**: Lauren E. Petty, Jennifer E. Below, Craig L. Hanis; **WHI**: Misa Graff, Kari E. North.
- **SA-T2D(South Asian): LOLIPOP**: Weihua Zhang, John C Chambers, Jaspal S Kooner; **PROMIS**: Jung-Jin Lee, Danish Saleheen; **RHS**: Fumihiko Takeuchi, Norihiro Kato; **SINDI**: Xueling Sim, Tai E Shyong; **EPIDREAM/INTERHEART:** Amel Lamri, Sonia Anand**; GRC-CDS**: Meraj Ahmad, Divya Sri Priyanka T, Giriraj R Chandak; **SDS**: Richa Saxena, Bishwa R Sapkota, Dharambir K Sanghera; **GEMS**: Lin Tong, Habibul Ahsan; **INDICO**: Anubha Mahajan, Dwaipayan Bhardwaj; **UKBB**: Anubha Mahajan, Mark McCarthy.
- **DIAMANTE Trans-ethnic:**
- **Analysts:** Jennifer E Below, Kyle J Gaulton, Hidetoshi Kitajima, Anubha Mahajan, Andrew P Morris, Maggie Ng, Lauren E. Petty, Xueling Sim, Daniel Taliun, Weihua Zhang.
- **PIs:** Mike Boehnke, John Chambers, Mark McCarthy, Andrew P Morris, Jerry Rotter, E-Shyong Tai.
- **NIH funding:** U01-DK105535, R01-DK66358 and R01-DK78616.

Eric Gamazon



Dan Roden

Lisa Bastarache

Josh Denny

Anuar Konkashbaev

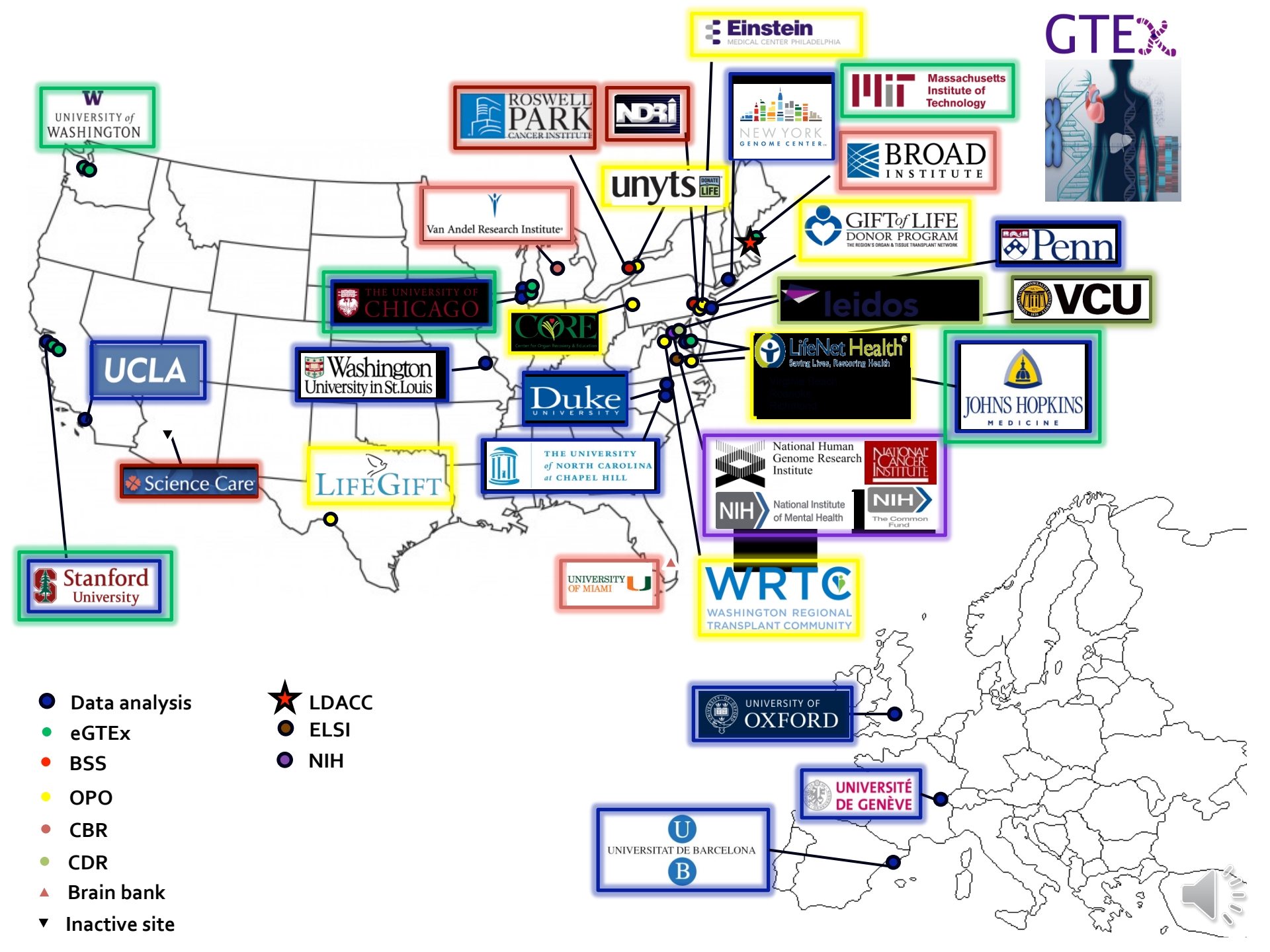
Haky Im

Gokhan Unlu

Jibril Hirbo

Jess Brown

Ela Knapik


Einstein
MEDICAL CENTER PHILADELPHIA
GTEx
UNIVERSITY of WASHINGTON
ROSWELL PARK CANCER INSTITUTE
NDRI
NEW YORK GENOME CENTER
Massachusetts Institute of Technology
BROAD INSTITUTE
unyts
Van Andel Research Institute
GIFT of LIFE DONOR PROGRAM
Penn
THE UNIVERSITY OF CHICAGO
leidos
VCU
CORE
UCLA
Washington University in St. Louis
LifeNet Health
Saving Lives, Restoring Health
JOHNS HOPKINS MEDICINE
Duke UNIVERSITY
National Human Genome Research Institute
NATIONAL CANCER INSTITUTE
NIH
National Institute of Mental Health
NIH
The Common Fund
Science Care
LIFEGIFT
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Stanford University
UNIVERSITY OF MIAMI
WRTC
WASHINGTON REGIONAL TRANSPLANT COMMUNITY
UNIVERSITY OF OXFORD
UNIVERSITÉ DE GENÈVE
UNIVERSITAT DE BARCELONA
Data analysis
eGTEx
BSS
OPO
CBR
CDR
Brain bank
Inactive site
LDACC
ELSI
NIH