

Real-world applications of clinical AI + genomics

Ruthie Johnson, PhD

Berkowitz Postdoctoral Fellow
ruth_johnson@hms.harvard.edu

2/8/24



HARVARD
MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics

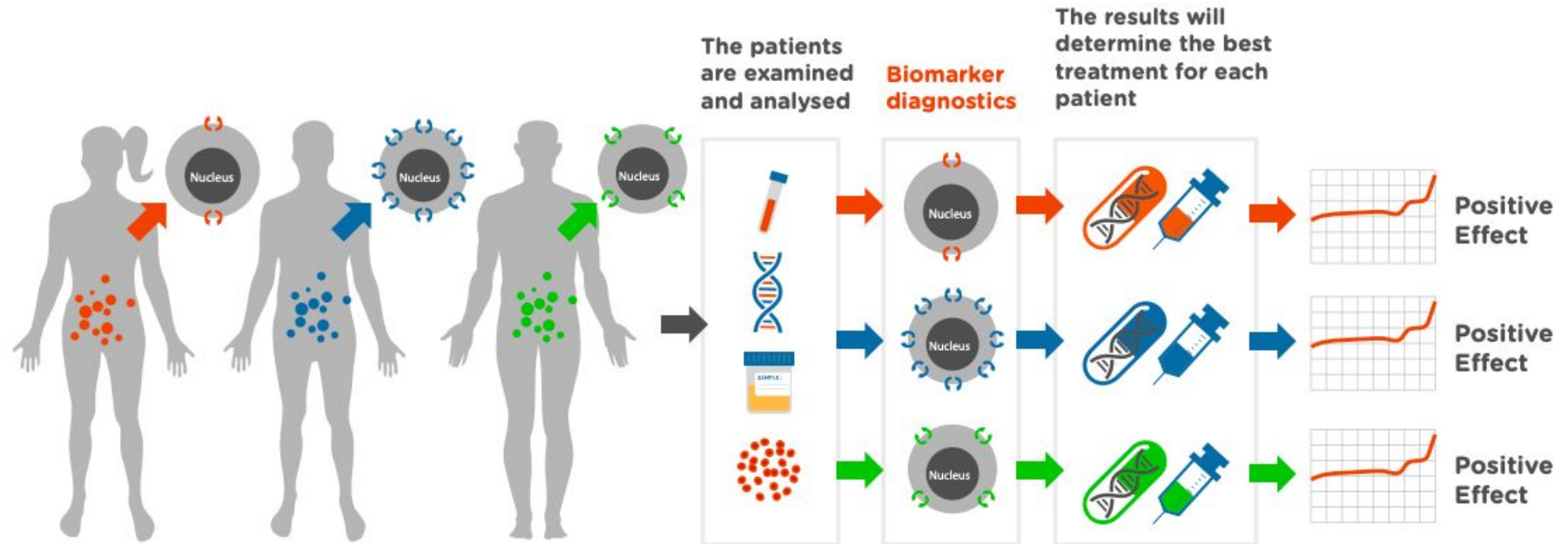
Overview

- **A brief primer on genetics**
- **Break (?)**
- **Assessing the interplay between genetic ancestry and disease risk**
 - *Leveraging genomic diversity for discovery in an EHR-linked biobank-- the UCLA ATLAS Community Health Initiative (Johnson et al. Genome Medicine 2022)*
- **Predicting rare disease through EHR signatures**
 - *Electronic health record signatures identify undiagnosed patients with Common Variable Immunodeficiency Disease (Johnson et al. medRxiv 2022)*

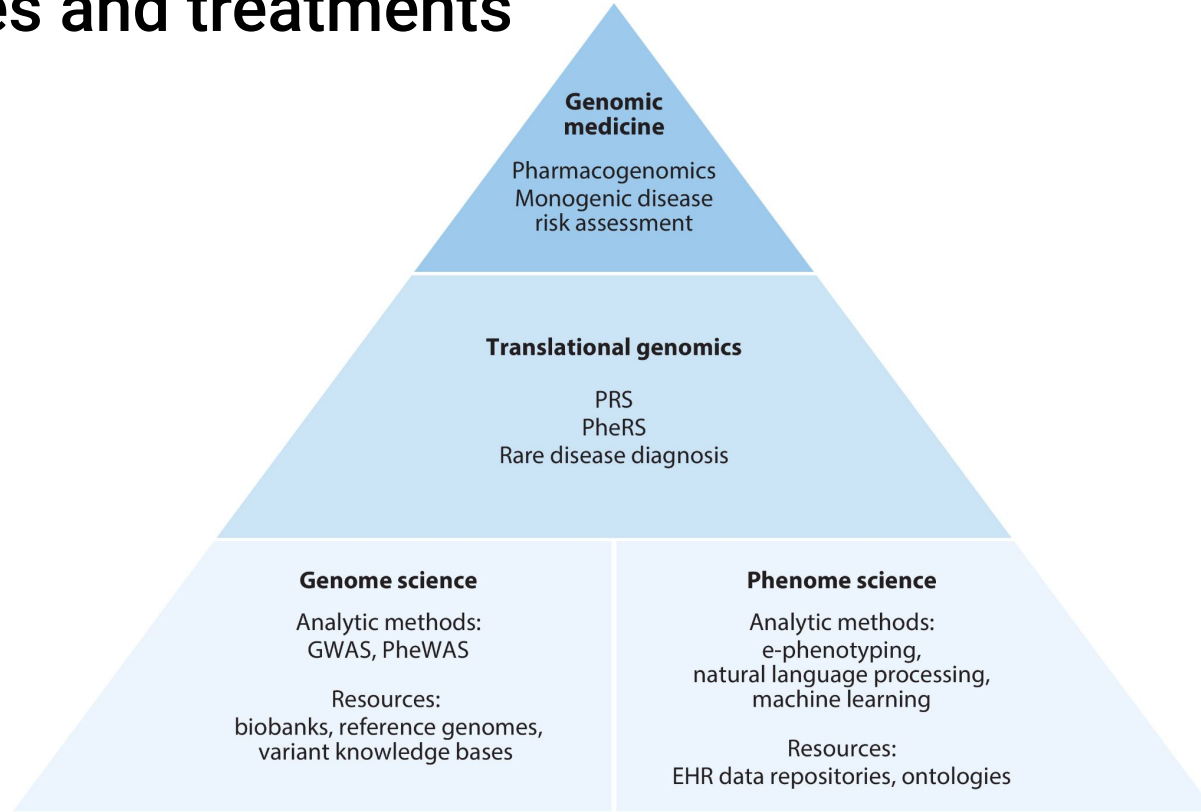
What does precision medicine with genomics entail?

INNOVATIVE MEDICINE: PERSONALISED MEDICINE

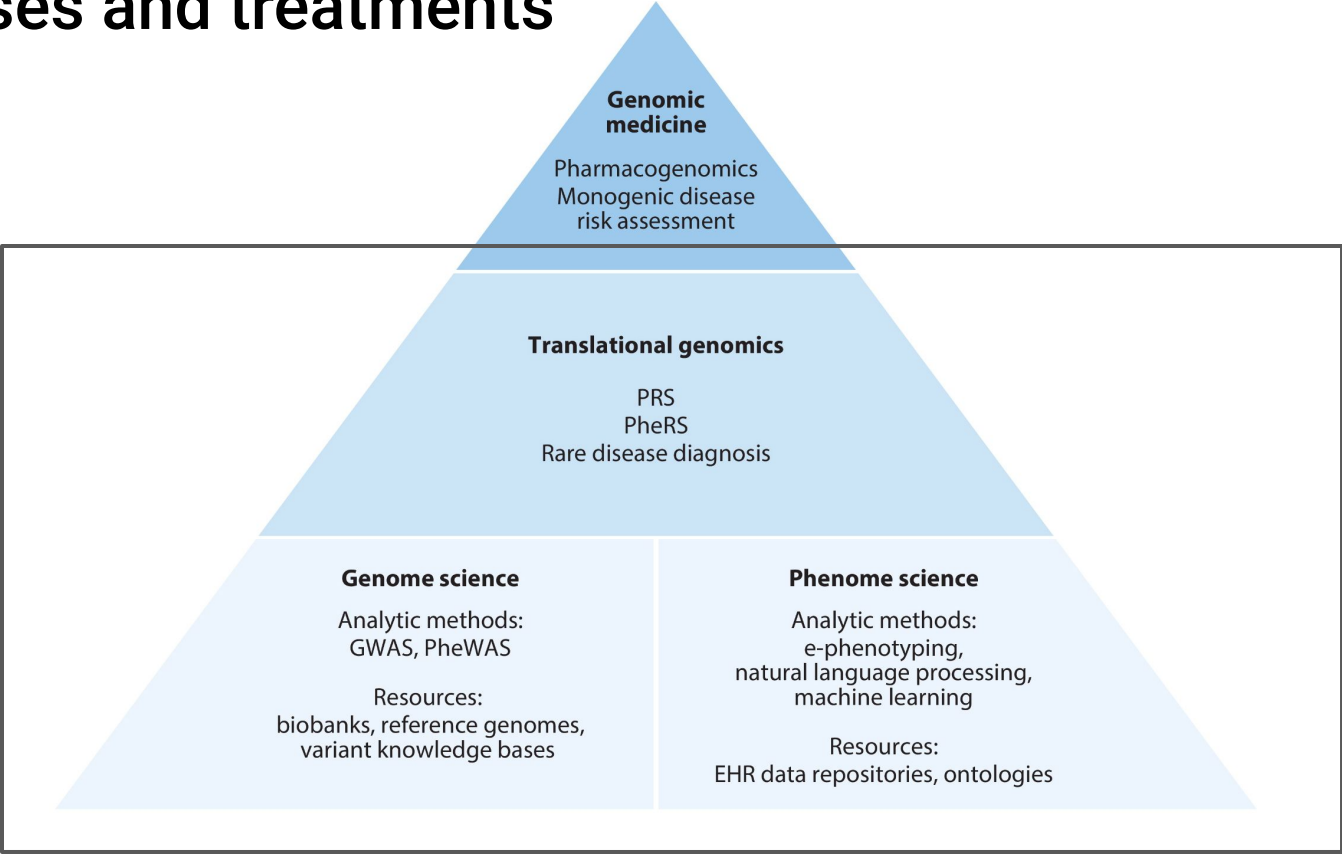
Cancer patients with e.g. colon cancer receive a personalised therapy based on their biomarkers



Genomic medicine is a key component for individualized diagnoses and treatments



Genomic medicine is a key component for individualized diagnoses and treatments

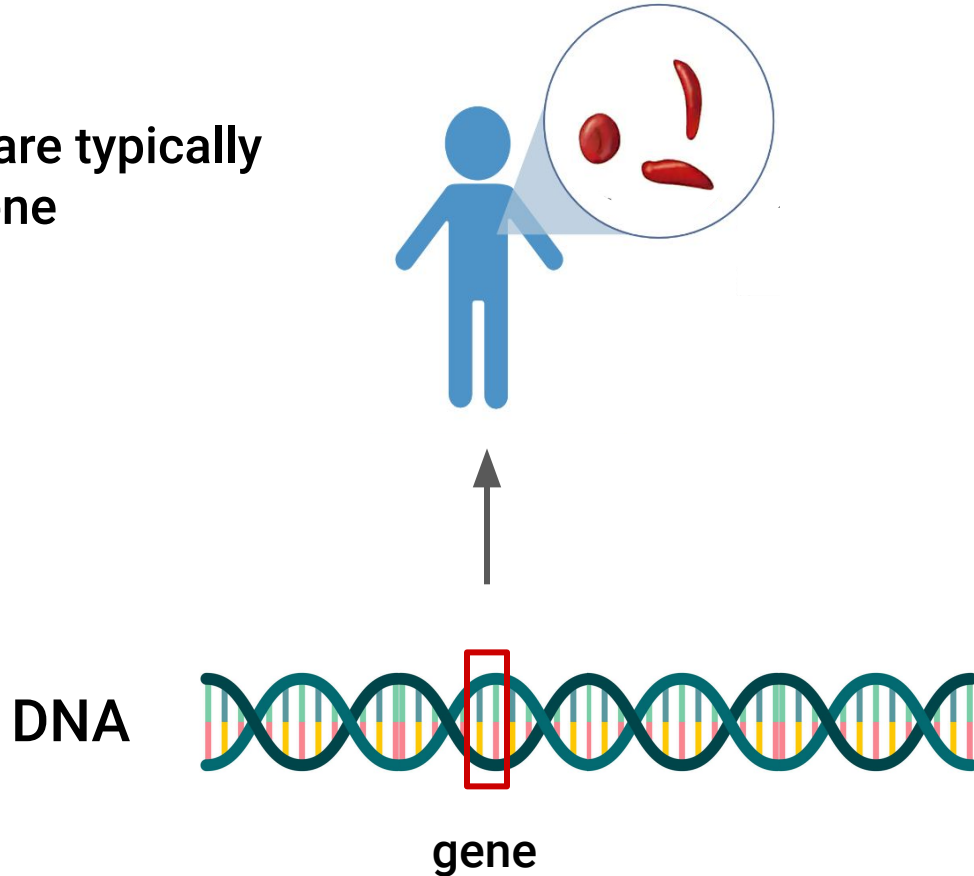


A key goal of genomic medicine is identifying genes that cause a disease

Monogenic diseases are typically caused by a single gene

- Cystic fibrosis
- Sickle cell anemia
- Huntington disease
- Duchenne muscular dystrophy

many, many more!

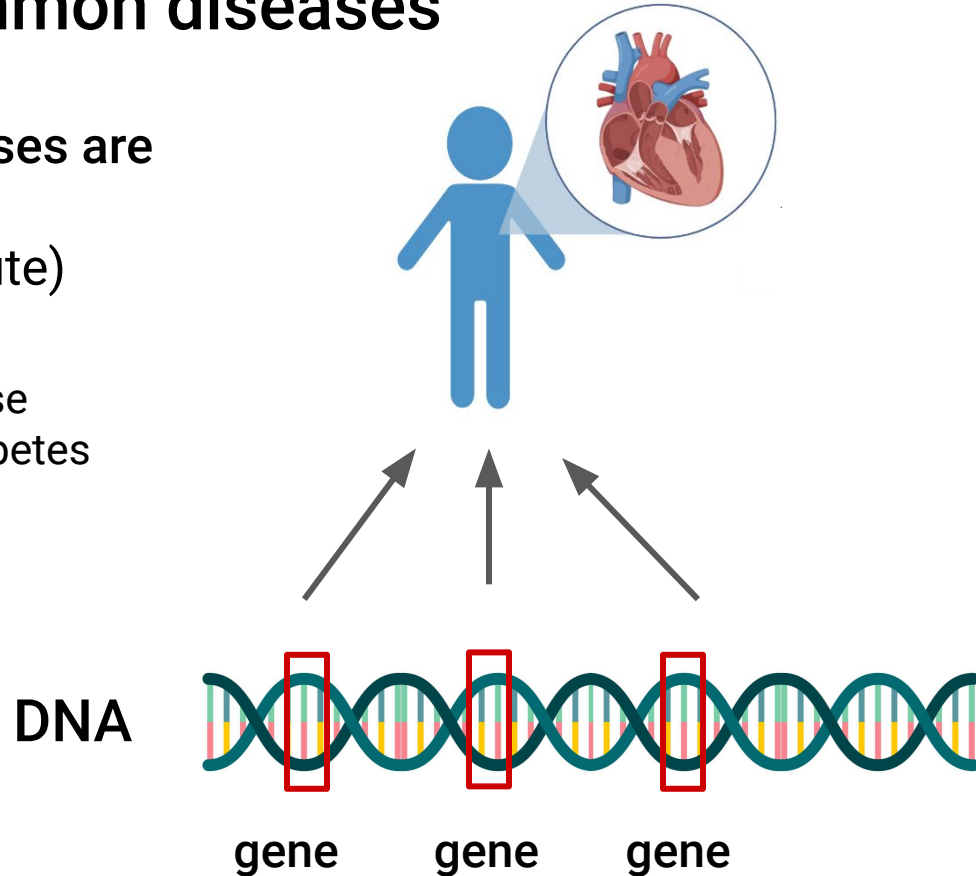


Numerous genes across the genome contribute to disease risk for most common diseases

Complex traits/diseases are *polygenic*
(many genes contribute)

- Coronary heart disease
- Type I and Type II diabetes
- Breast cancer
- Height and BMI

many, many more!



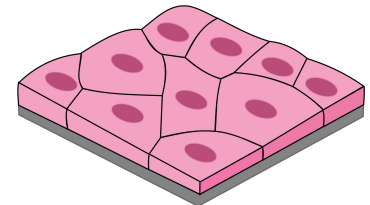
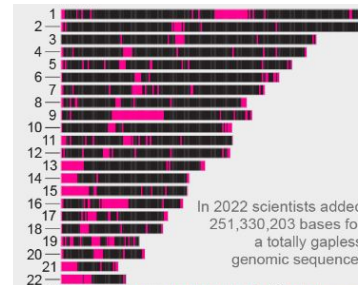
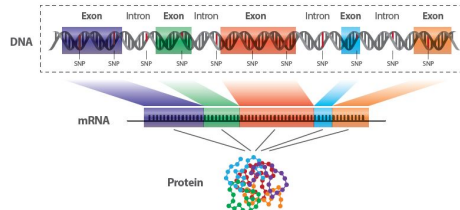
Commonly measured biospecimens and biomarkers

Genotyping
~650K common SNPs and small indels
~\$100
studying complex traits/diseases and common genetic variation

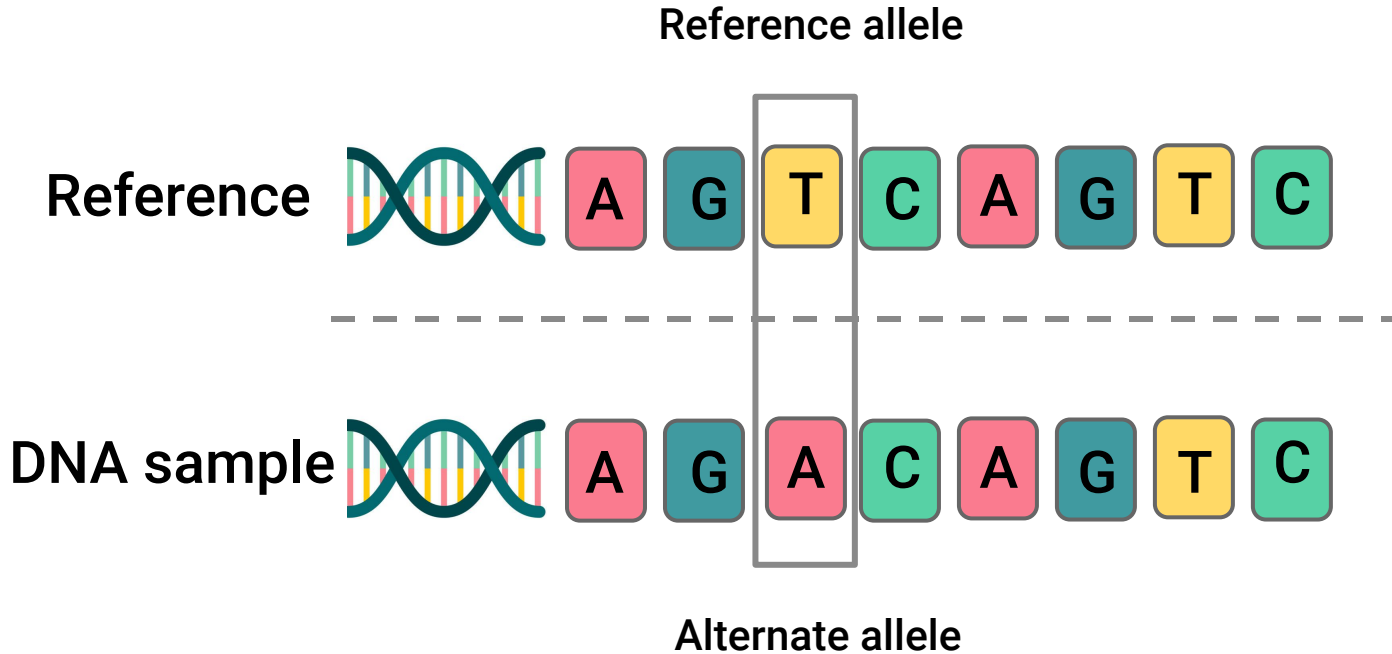


Commonly measured biospecimens and biomarkers

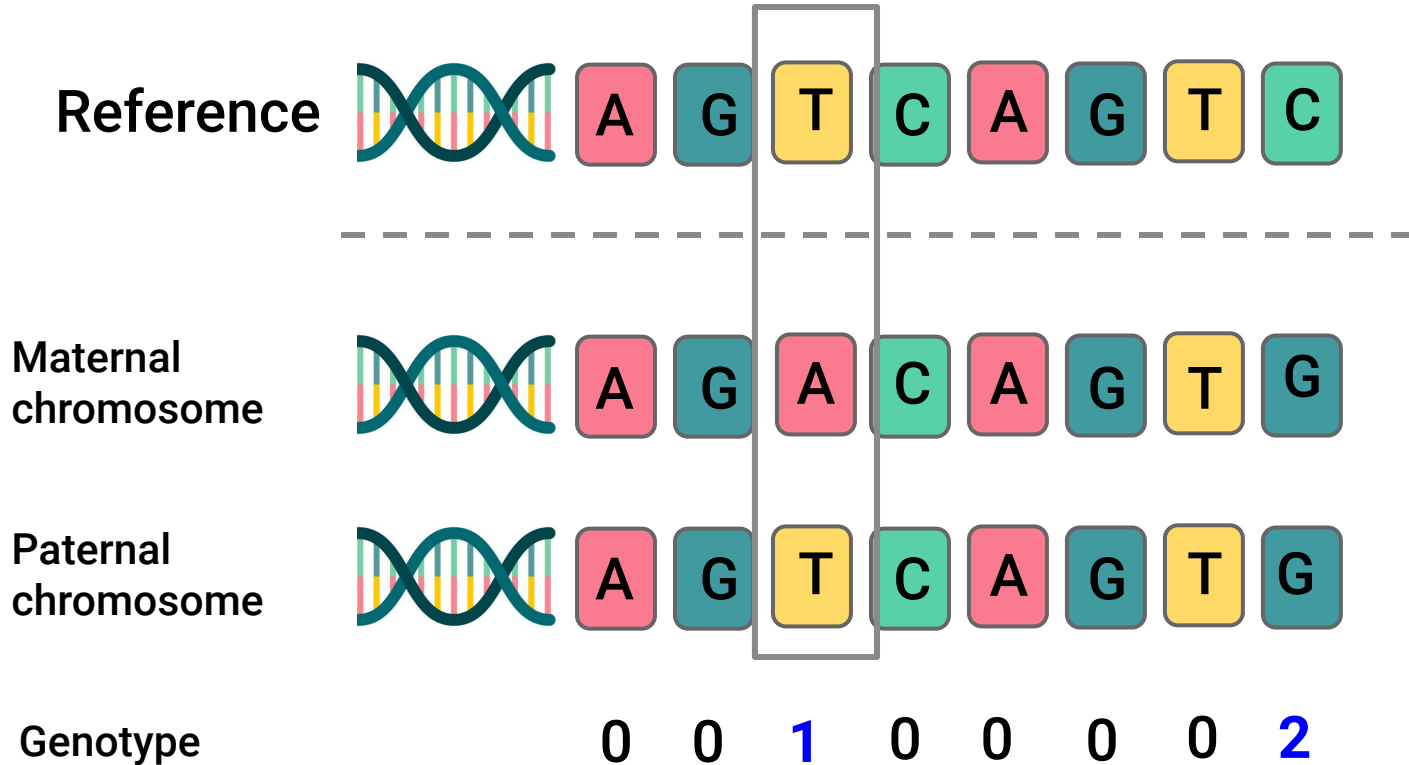
Genotyping	Exome sequencing	Whole genome sequencing	RNA-sequencing
~650K common SNPs and small indels	exons (protein coding regions)	all variants (including very rare) in exons and introns and large structural variants	Captures cellular content of RNAs
~\$100	~\$8K - \$11K	~\$10K - \$20K	~10K - \$50K+
studying complex traits/diseases and common genetic variation	Studying rare genetic diseases	Ultra rare genetic diseases, including de novo mutations	Understanding transcriptome, i.e. connecting genes to functional proteins



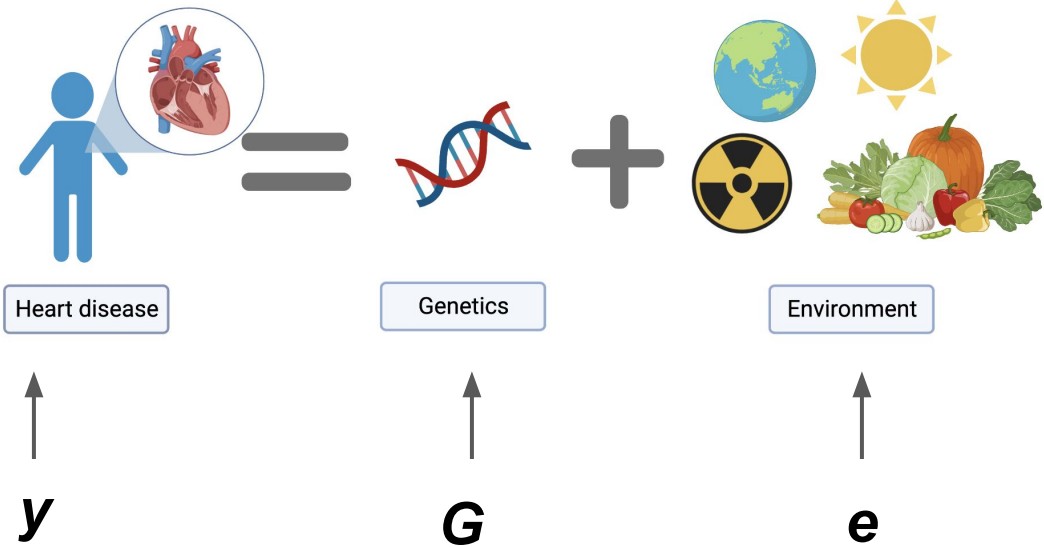
Single nucleotide polymorphisms (SNP) are “common” point mutations across the genome (minor allele frequency > 1%)



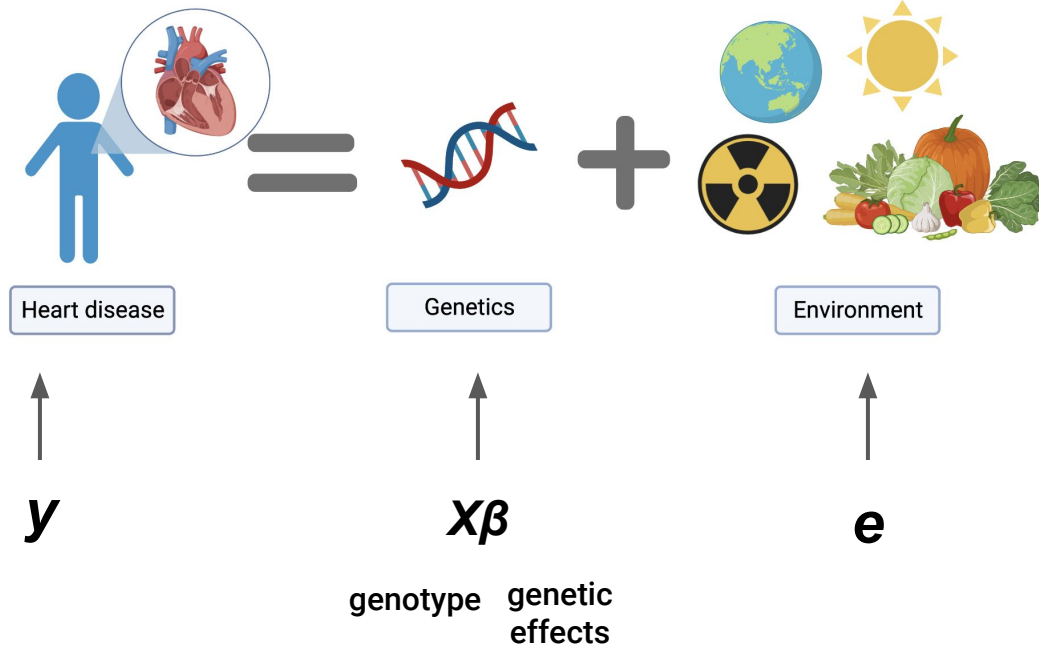
Single nucleotide polymorphisms (SNP) are “common” point mutations across the genome (minor allele frequency > 1%)



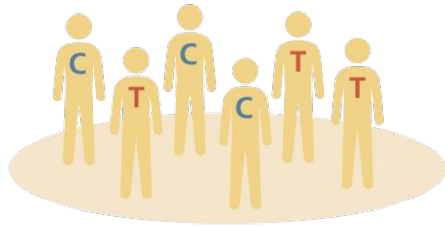
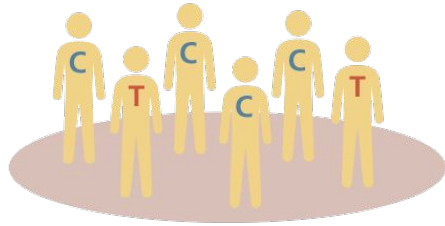
Disease phenotypes are a combination of genetic and environmental components



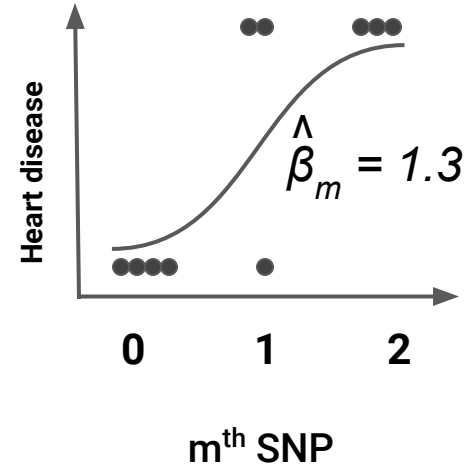
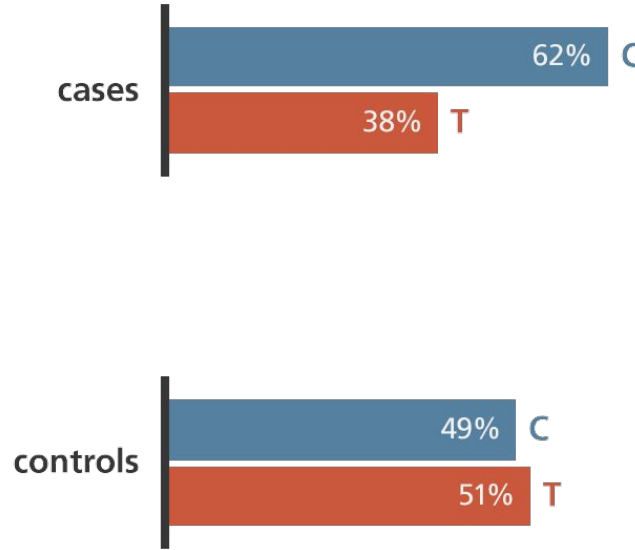
Some SNPs have no physiological effect, while others are linked to changes in phenotype



Genome-wide association studies (GWAS) aims to estimate the effects of the SNPs affecting a given phenotype



Hypothesis test at the m^{th} SNP



Genome-wide association studies (GWAS) aims to estimate the effects of the SNPs affecting a given phenotype

$$\beta \sim N(0, I\sigma_g^2)$$

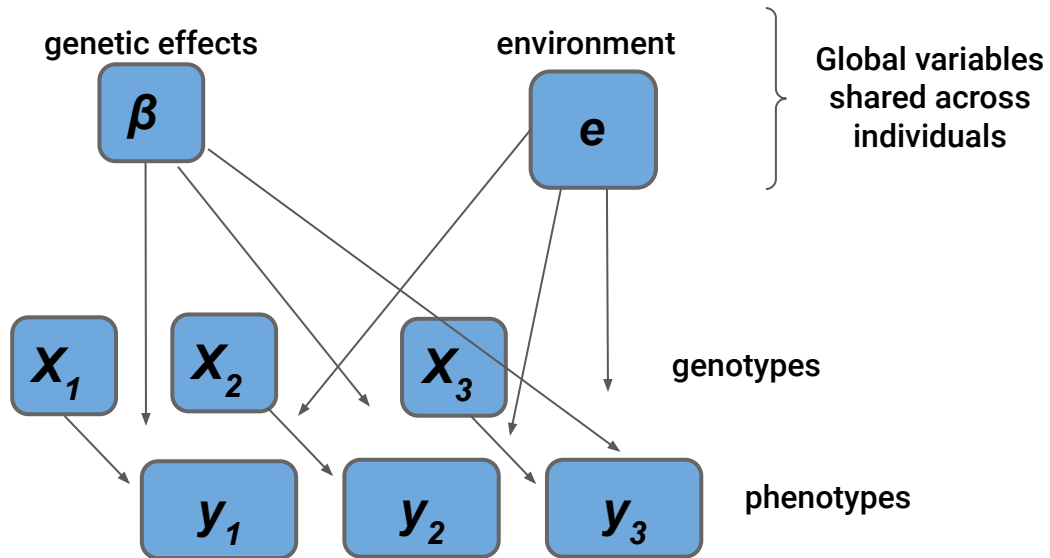
, for heart disease

$$e \sim N(0, \sigma_e^2)$$

$$y_i = x_i \beta + e$$

, for i^{th} individual

(assumes phenotypes $[y]$ are i.i.d.)



Genome-wide association studies (GWAS) aims to estimate the effects of the SNPs affecting a given phenotype

$$\beta \sim N(0, I\sigma_g^2)$$

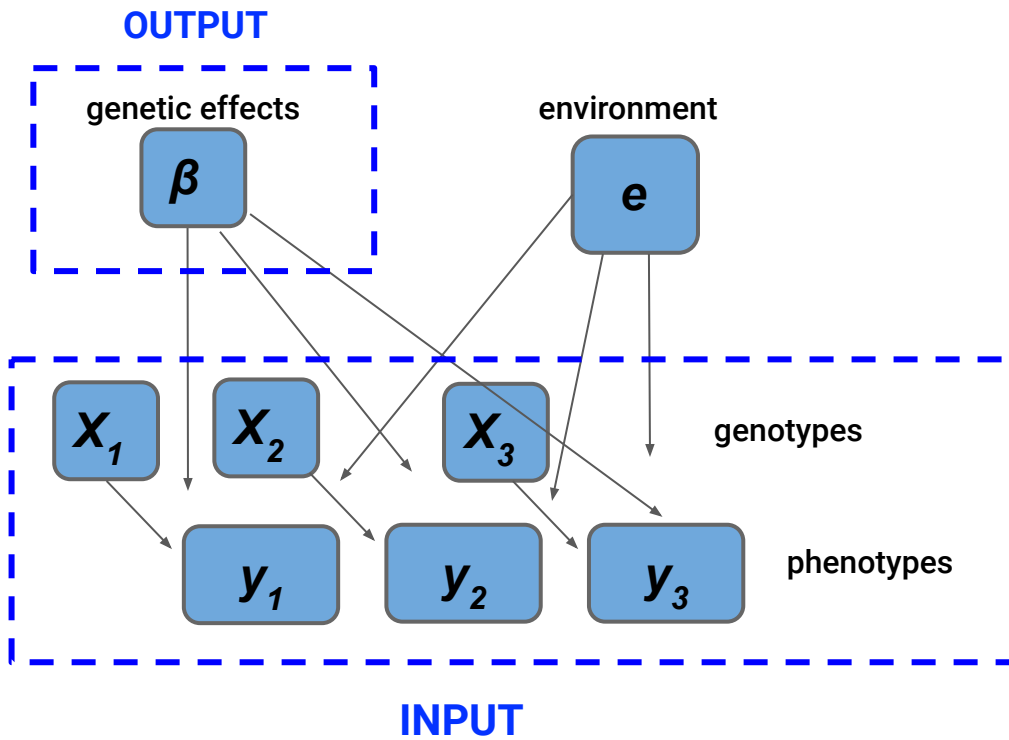
, for heart disease

$$e \sim N(0, \sigma_e^2)$$

$$y_i = x_i \beta + e$$

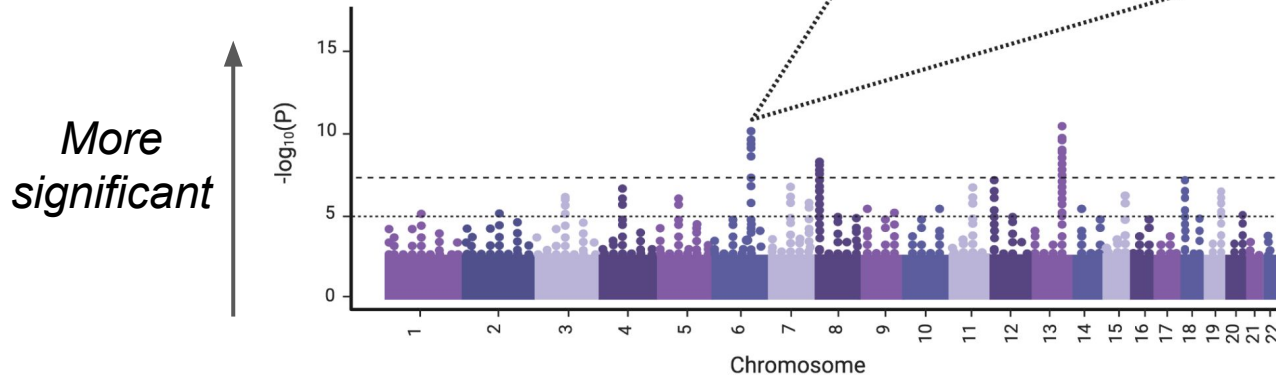
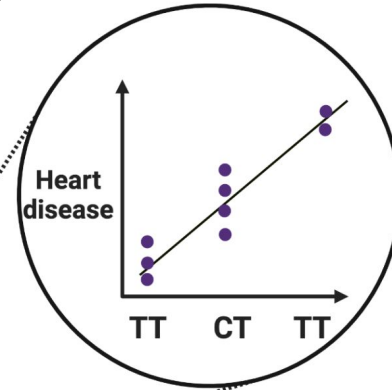
, for i^{th} individual

(assumes phenotypes $[y]$ are i.i.d.)



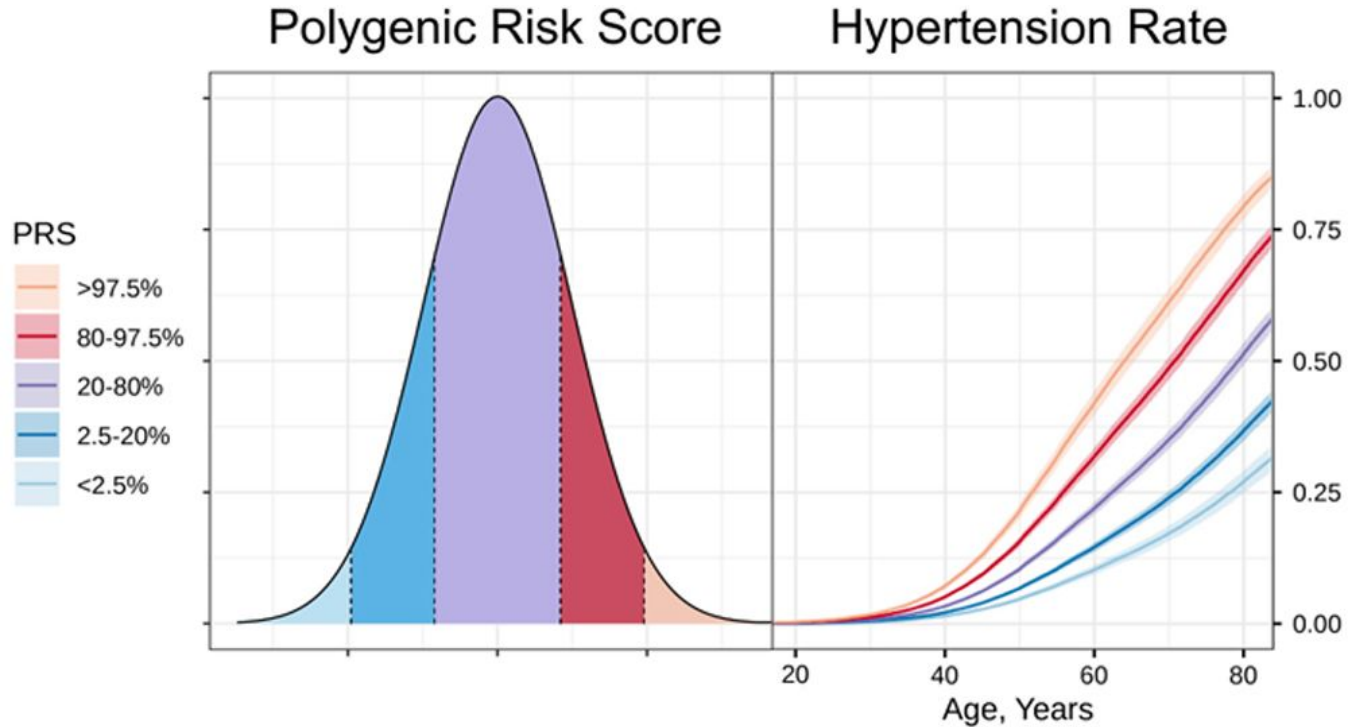
Disease risk is spread throughout the genome

- Coronary heart disease: 250+ regions
- Type I and Type II diabetes: 60+ and 500+ regions
- Breast cancer: 200+ regions
- Height and BMI: 700+ and 250+ regions



→ Identified variants are **NOT** always causal! Functional validation is needed to confirm causality.

Polygenic risk scores provide individual-level predictions to identify patients with heightened disease risk



Wow, if genetics can help us predict disease... why isn't *[insert favorite direct-to-consumer genetics company]* in all of the clinics?

- It is unclear how much additional risk information PRS provides over current risk assessment methods
- The majority of diseases have a much smaller genetic component compared to the effect of environmental factors

Wow, if genetics can help us predict disease... why isn't *[insert favorite direct-to-consumer genetics company]* in all of the clinics?

- It is unclear how much additional risk information PRS provides over current risk assessment methods
- The majority of diseases have a much smaller genetic component compared to the effect of environmental factors
- PRS has relatively poor sensitivity and specificity making it challenging to administer as a clinical prediction tool

“Typical sensitivity for a polygenic score is **10-15%** (meaning that only 10-15% of people who will develop the disease will have a high polygenic score)⁷—for example, a polygenic score developed to detect women at **>17%** lifetime risk of breast cancer has a **sensitivity of 39%** (it will identify 39% of the women who will go on to develop breast cancer, but miss 61% of them) and a **specificity of 78%** (22% of women who will not go onto develop breast cancer will be classified as having a “high risk score”)” - Sud et al. BMJ 2023

Wow, if genetics can help us predict disease... why isn't *[insert favorite direct-to-consumer genetics company]* in all of the clinics?

- It is unclear how much additional risk information PRS provides over current risk assessment methods
- The majority of diseases have a much smaller genetic component compared to the effect of environmental factors
- PRS has relatively poor sensitivity and specificity making it challenging to administer as a clinical prediction tool

“Typical sensitivity for a polygenic score is **10-15%** (meaning that only 10-15% of people who will develop the disease will have a high polygenic score)⁷—for example, a polygenic score developed to detect women at **>17%** lifetime risk of breast cancer has a **sensitivity of 39%** (it will identify 39% of the women who will go on to develop breast cancer, but miss 61% of them) and a **specificity of 78%** (22% of women who will not go on to develop breast cancer will be classified as having a “high risk score”)” - Sud et al. BMJ 2023

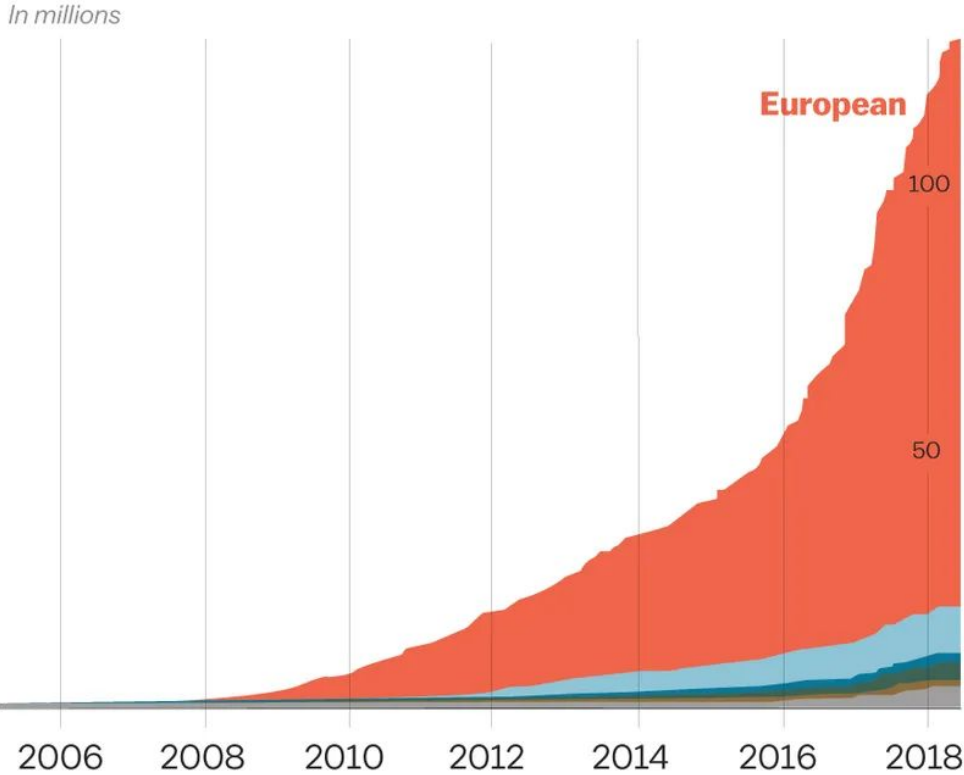
- **Is it equitable? PRS does not have uniform performance across all patient populations.**

Take 5 min break

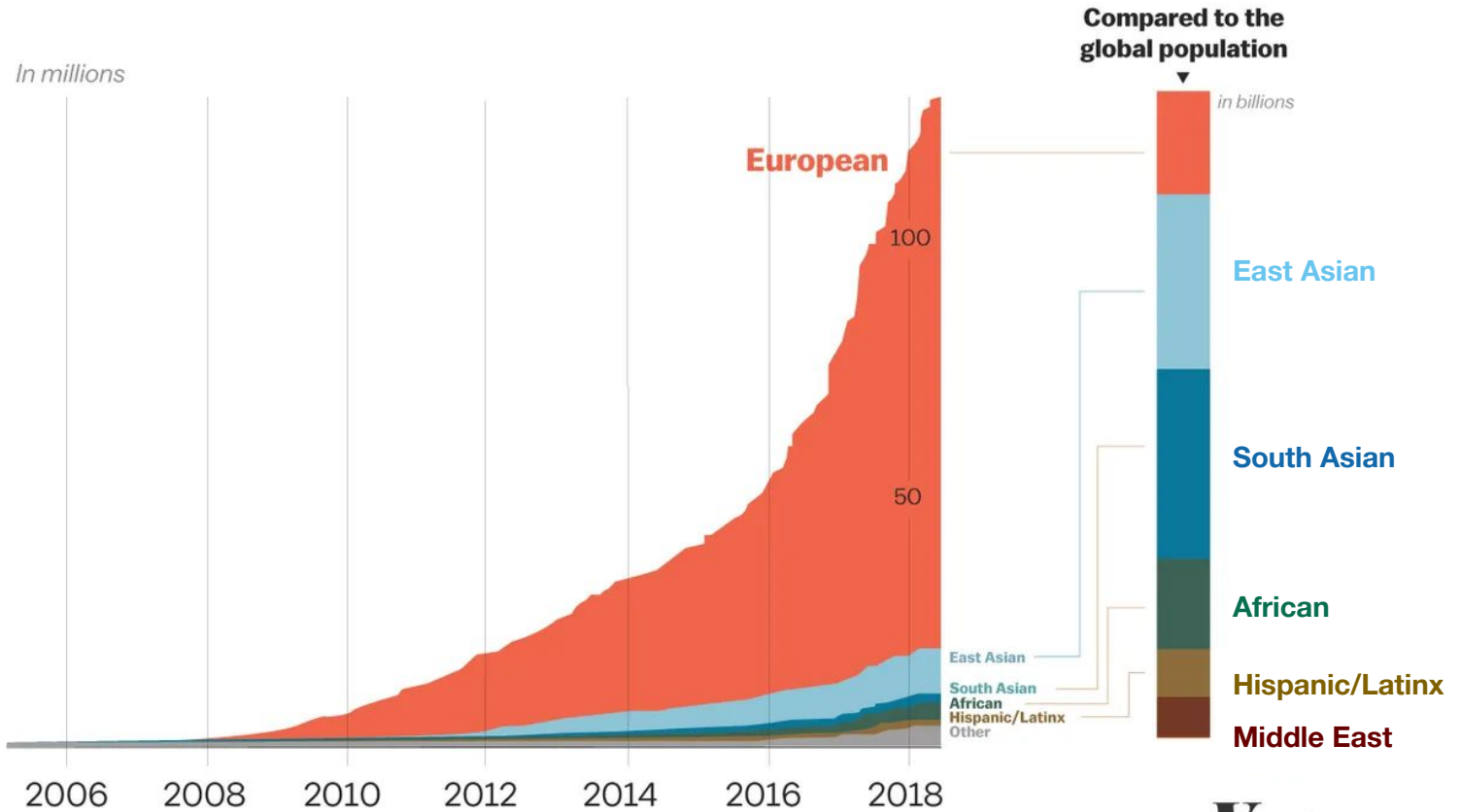
Part 2

Assessing the interplay between genetic ancestry and disease risk

Majority of genetic studies focus on European ancestry individuals



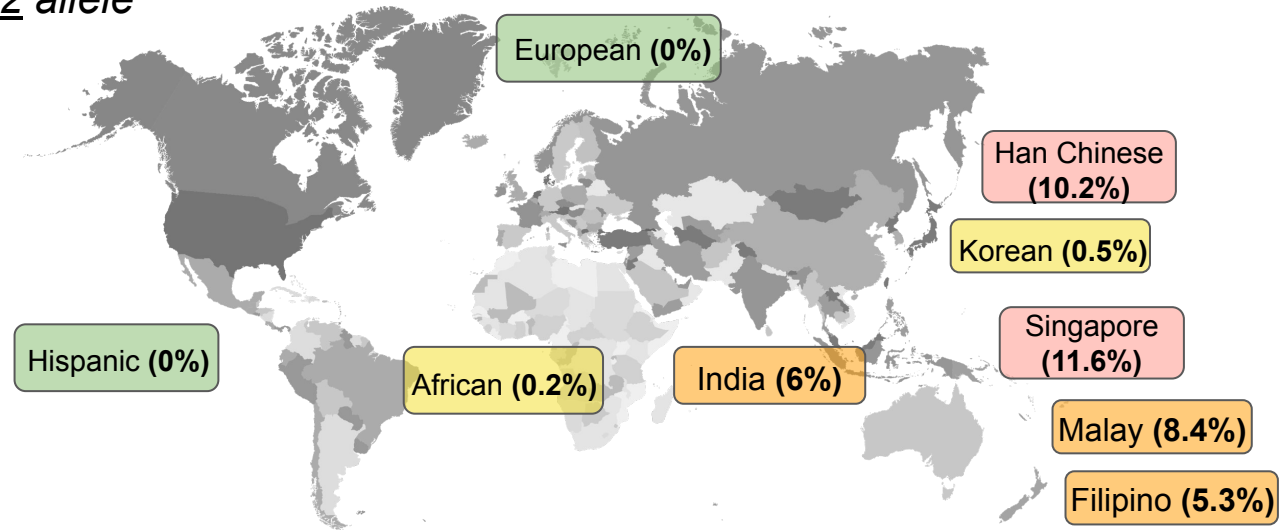
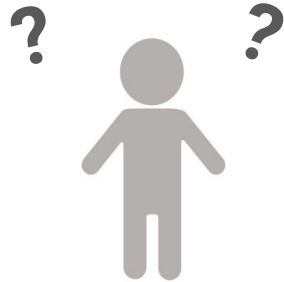
Majority of genetic studies focus on European ancestry individuals



Martin et al. Nat Genet 2019

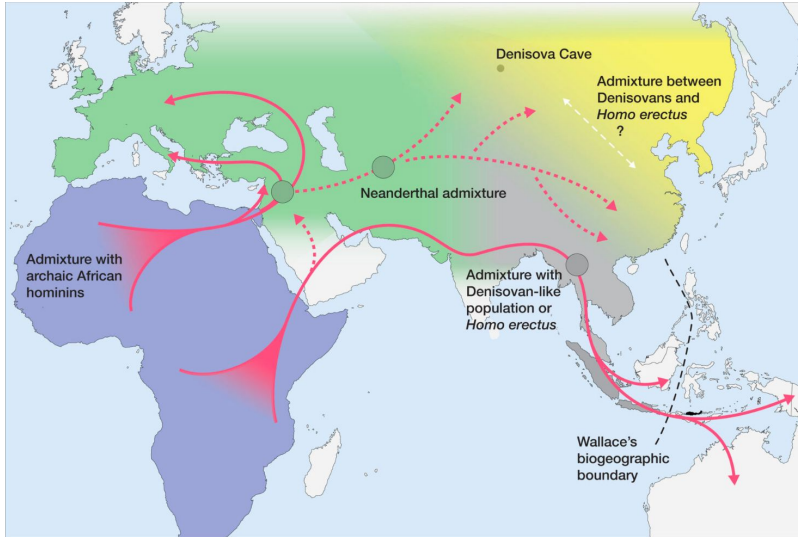
Explicitly considering genetic ancestry is key to precision medicine efforts

- Genetic ancestry provides specific information about key patterns of genetic variation, making it an important factor in numerous healthcare decisions
 - e.g. Carbamazepine is highly associated with adverse side effects in individuals with the HLA allele B*1502 allele

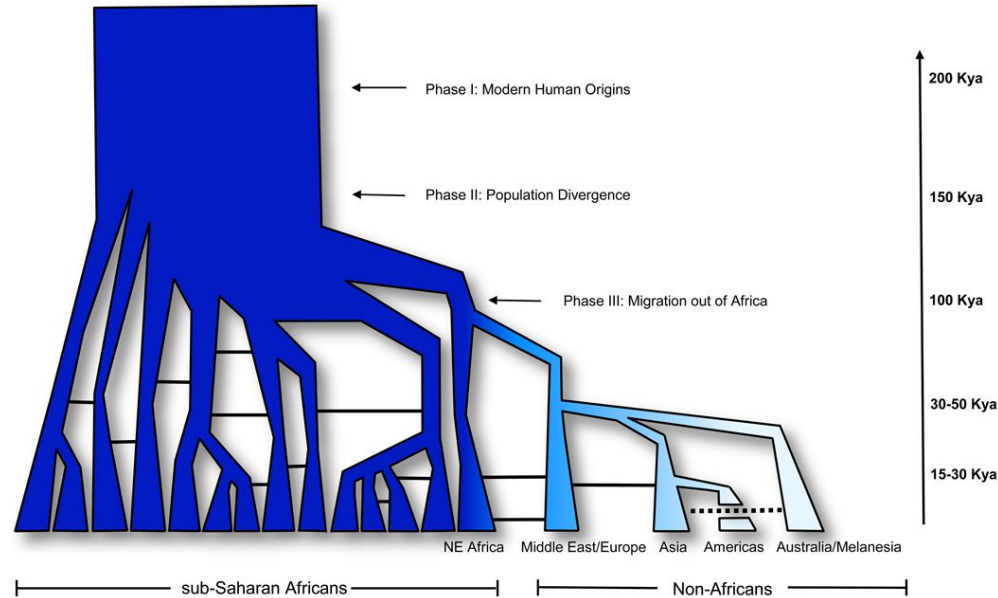


HLA allele B*1502 allele frequency

Evolutionary forces created a variety of genetic landscapes across continents

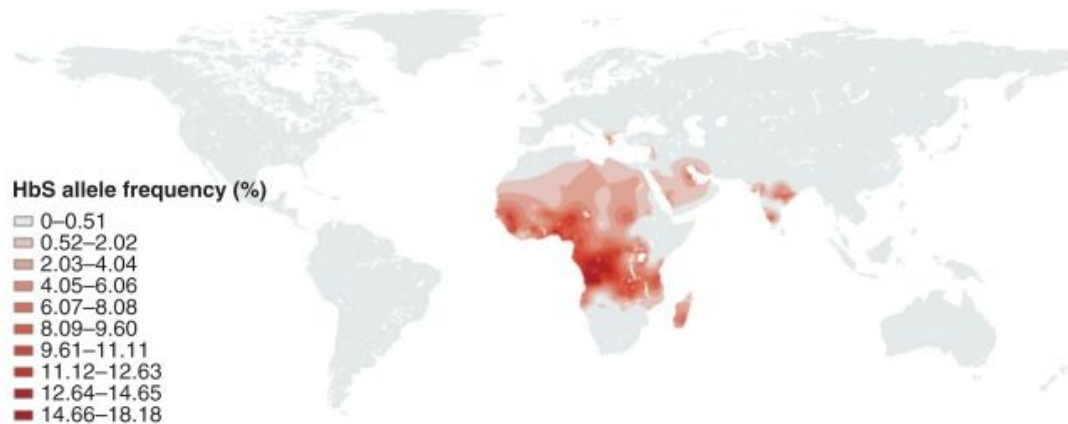


Historical patterns of migration influenced the global distribution of genetic variation through gene flow and genetic drift

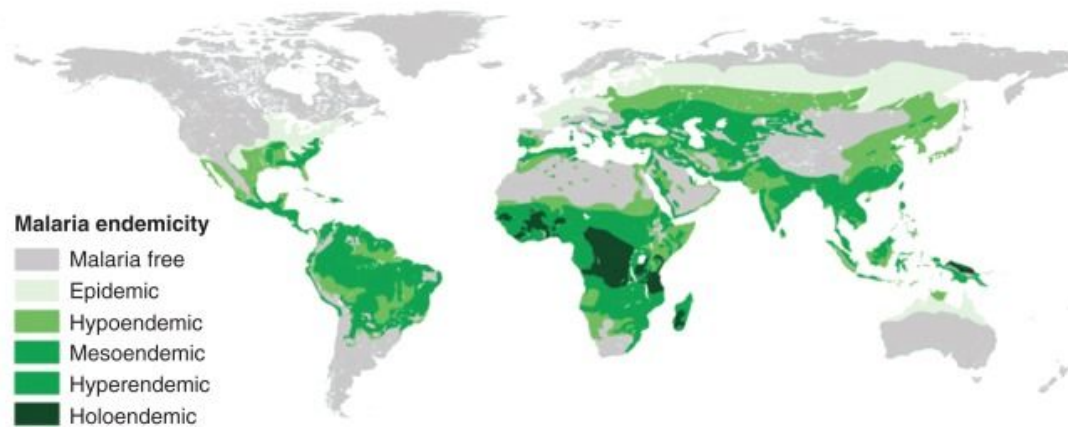


The out-of-Africa migration led to a bottleneck effect that reduced genetic variation across non-African ancestry populations

Differential genetic architecture across ancestries affects disease risk across populations



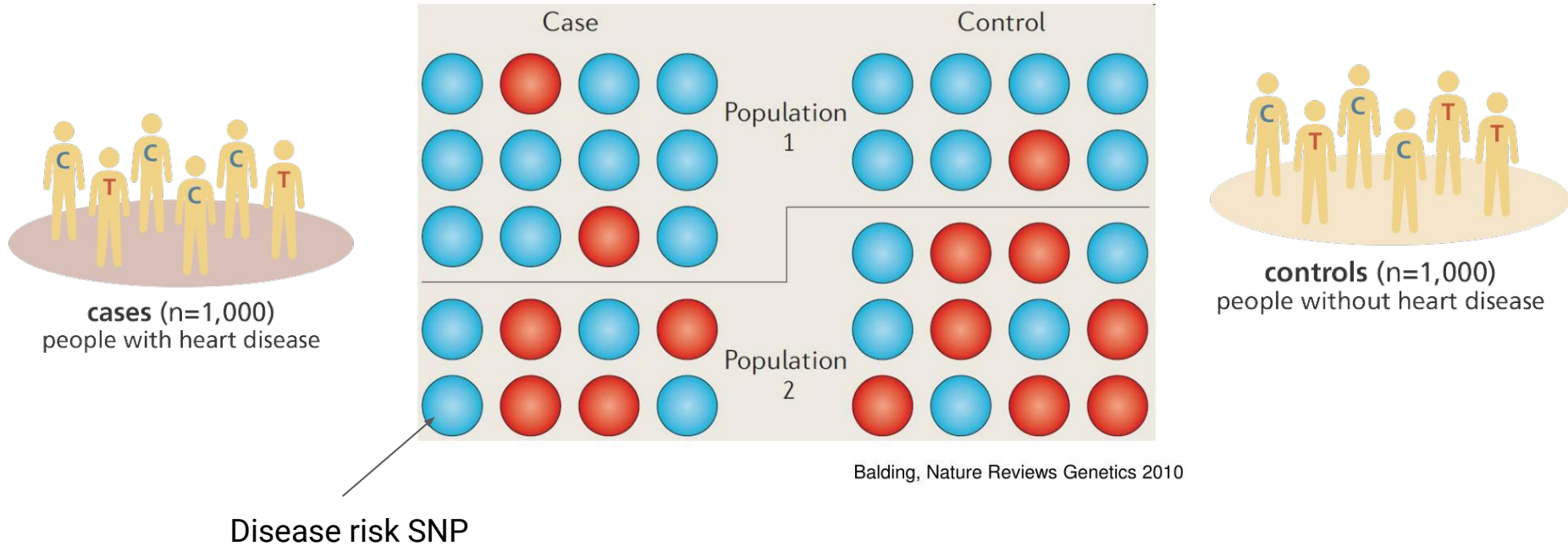
HbS allele frequency



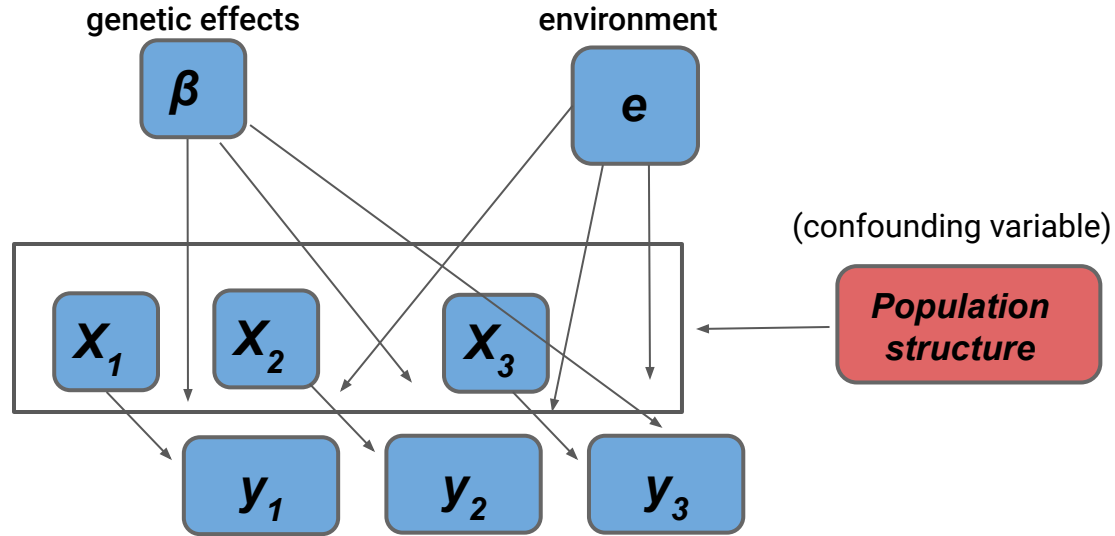
Malaria endemicity



Population structure can lead to spurious associations



Population structure confounds the association between genotypes and phenotypes



Principal component analysis captures population structure

Individual A	2	0	1	0	0	1
Individual B	0	2	0	0	0	1
Individual C	0	0	1	0	0	0
Individual D	0	0	1	0	2	2
⋮						
	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6

PCA
→

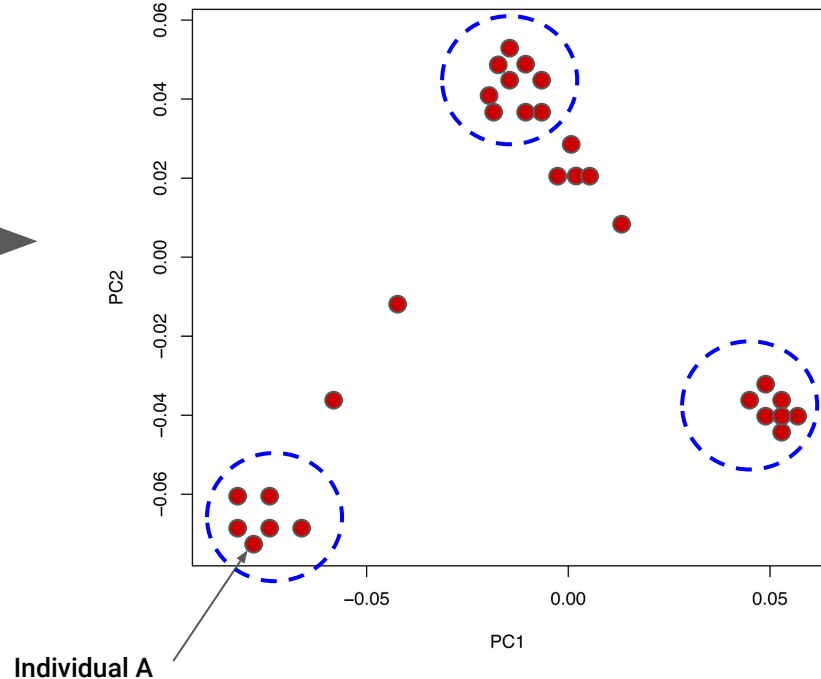
Individual A	0.1	0.4	1	0.9
Individual B	0	0.2	0.5	0.7
Individual C	0.6	0.3	0.2	0
Individual D	0	0.7	0.1	0.5
⋮				
	PC 1	PC 2	PC 3	PC 4

PCA is a dimensionality reduction technique that aims to maximize the variance of the data represented in the top principal components (vectors) → reconstruct the information represented in the data with the fewest dimensions as possible

Principal component analysis captures population structure

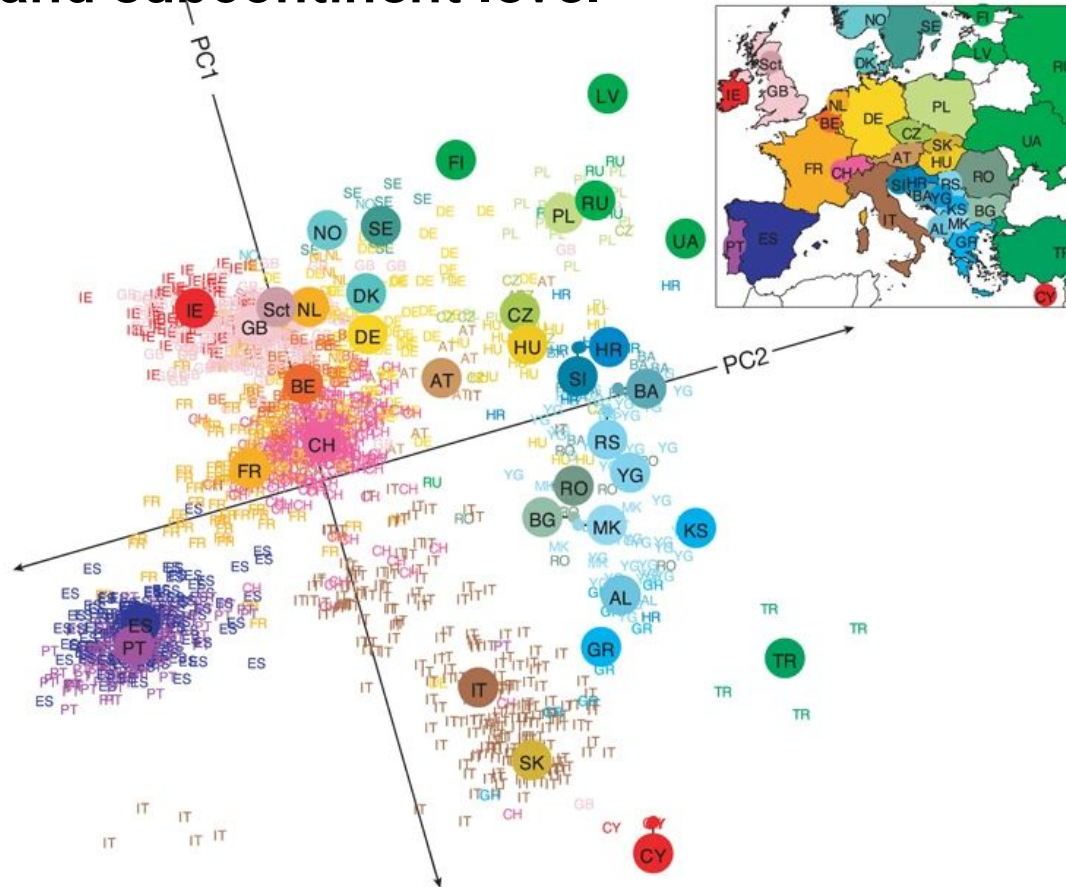
Individual A	2	0	1	0	0	1
Individual B	0	2	0	0	0	1
Individual C	0	0	1	0	0	0
Individual D	0	0	1	0	2	2
⋮						
	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6

PCA
→

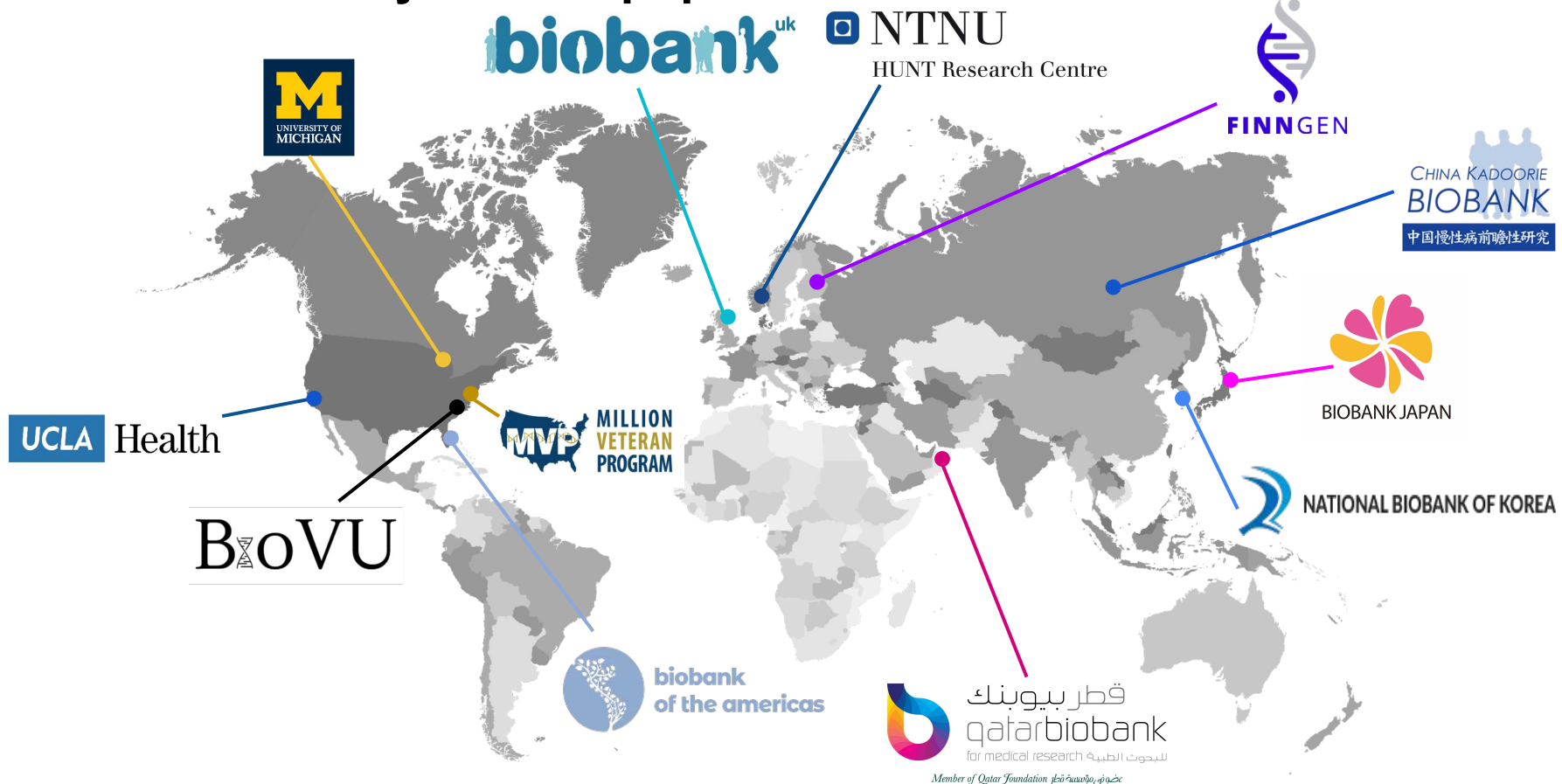


PCA is a dimensionality reduction technique that aims to maximize the variance of the data represented in the top principal components (vectors) → reconstruct the information represented in the data with the fewest dimensions as possible

Principal component analysis captures population structure at the continental and subcontinent level

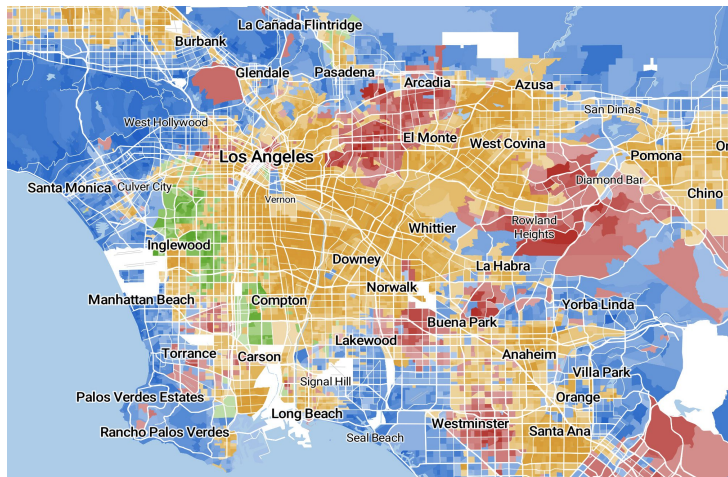


EHR-linked biobanks provide the opportunity to study disease risk across ancestrally diverse populations

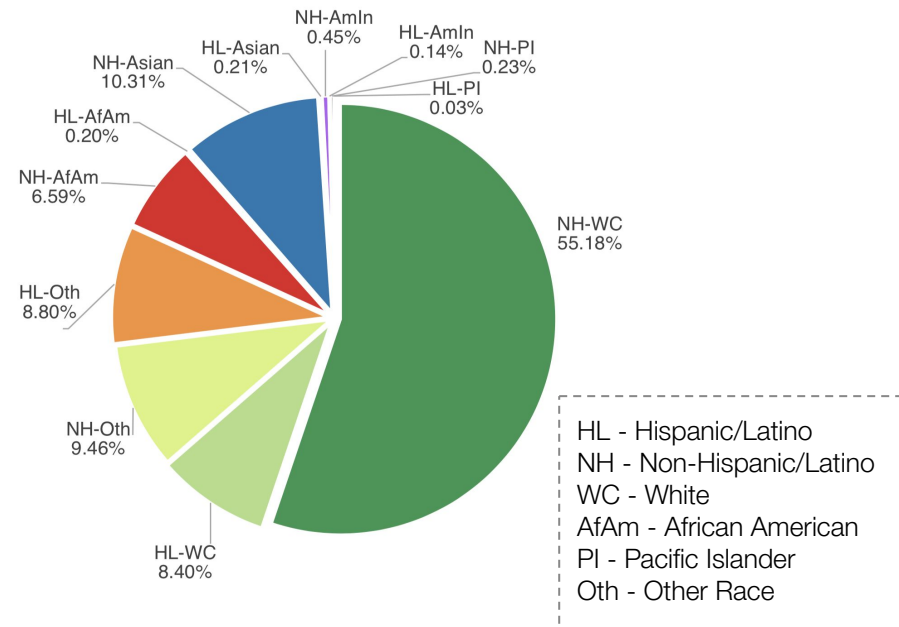


UCLA ATLAS captures the vibrant diversity of Los Angeles

Los Angeles Census



ATLAS self-identified race/ethnicity

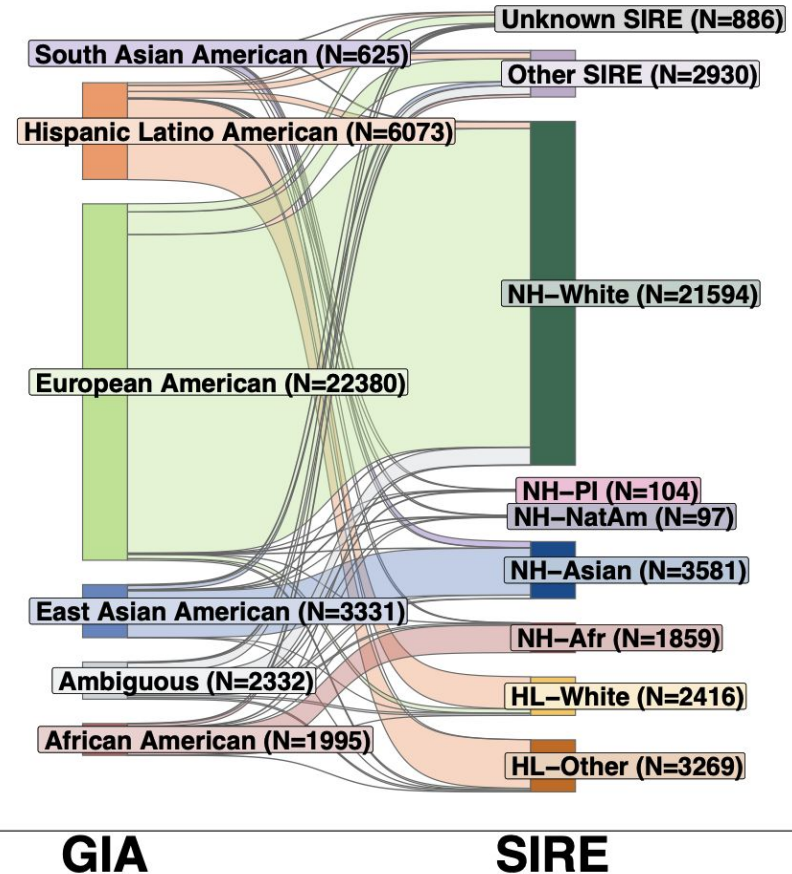


Within ATLAS, about 40% of individuals self-identify as a race other than White, with appreciable sample sizes in the Hispanic Latino and Asian American populations

Self-identified race/ethnicity (SIRE) and genetically inferred ancestry (GIA) are not analogous

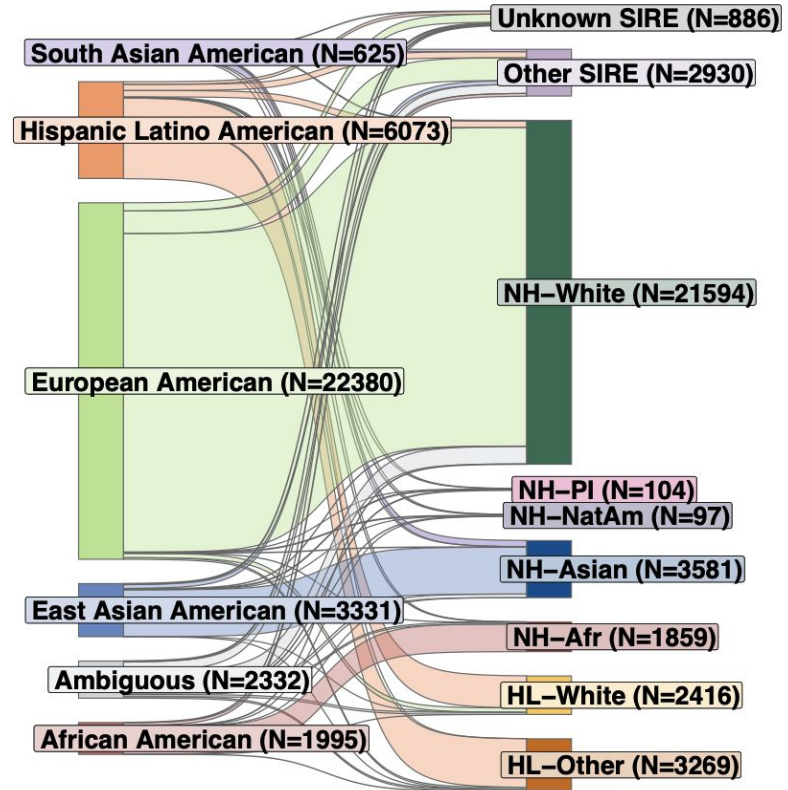
Genetically inferred ancestry (GIA): genetic characterization of individuals within a group who likely share recent biological ancestors as inferred by a method of choice and a given reference panel

Self-identified Race and Ethnicity (SIRE) have no direct biological implications



No clear 1:1 correspondence between SIRE and GIA

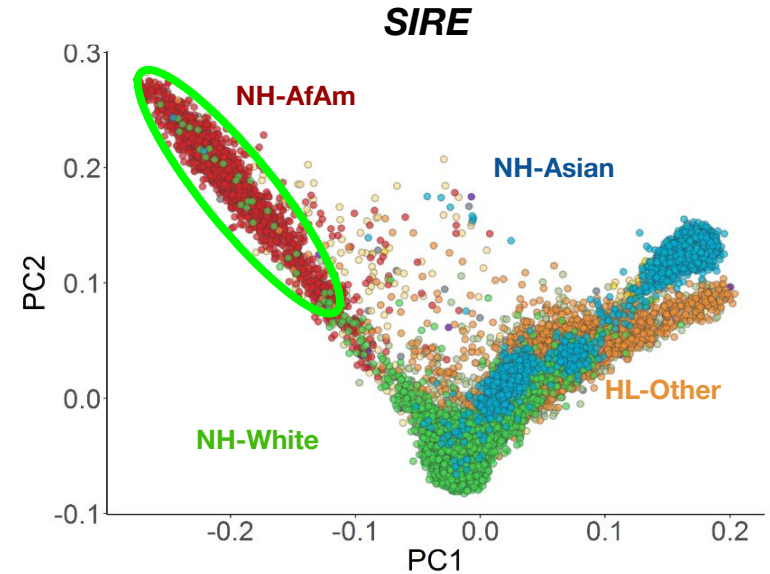
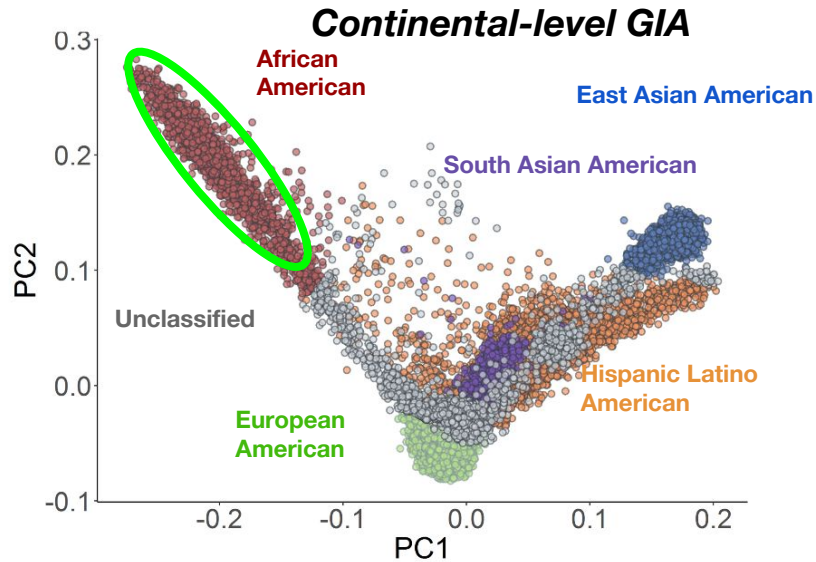
- Hispanic Latino American GIA group splits into multiple SIREs
- Significant proportion of individuals in the European American GIA group self-identify as one of the multiple other SIREs



GIA

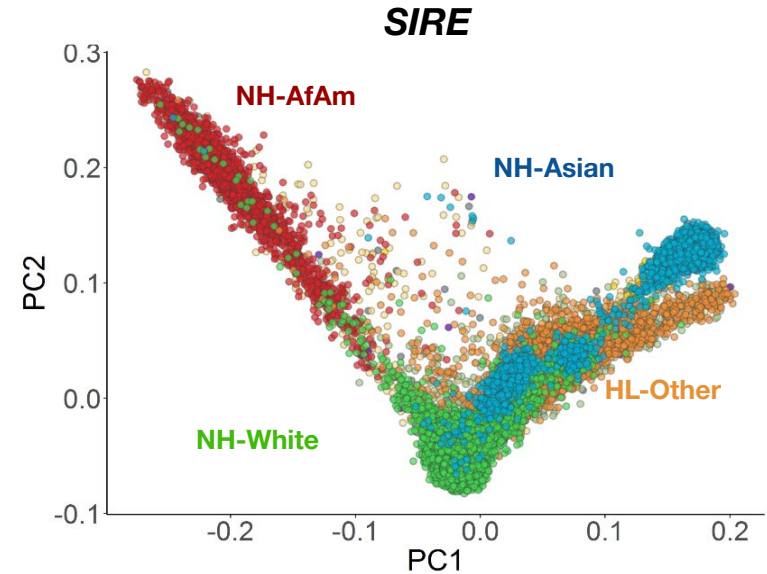
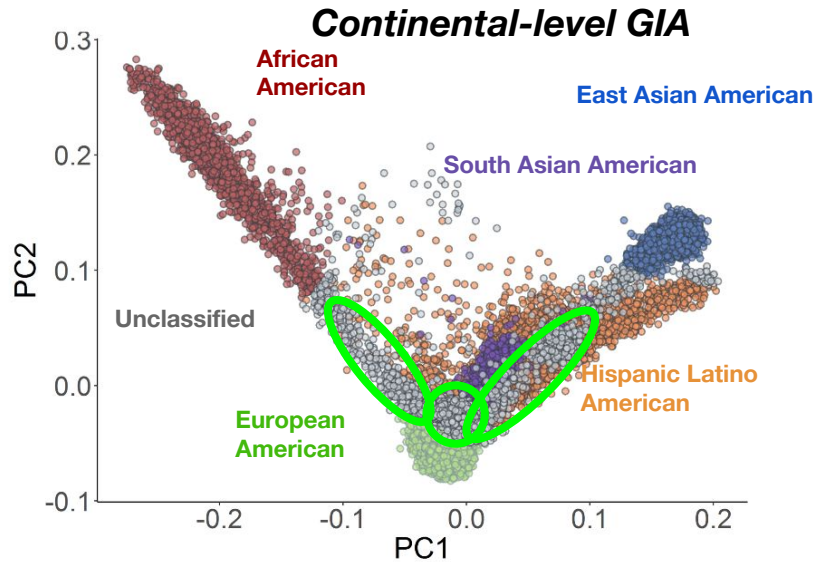
SIRE

PCA reveals notable differences between GIA and SIRE



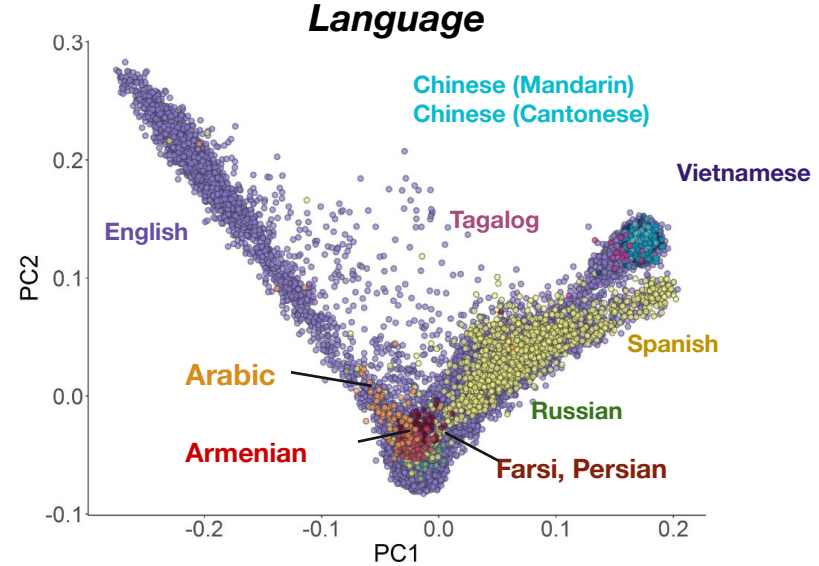
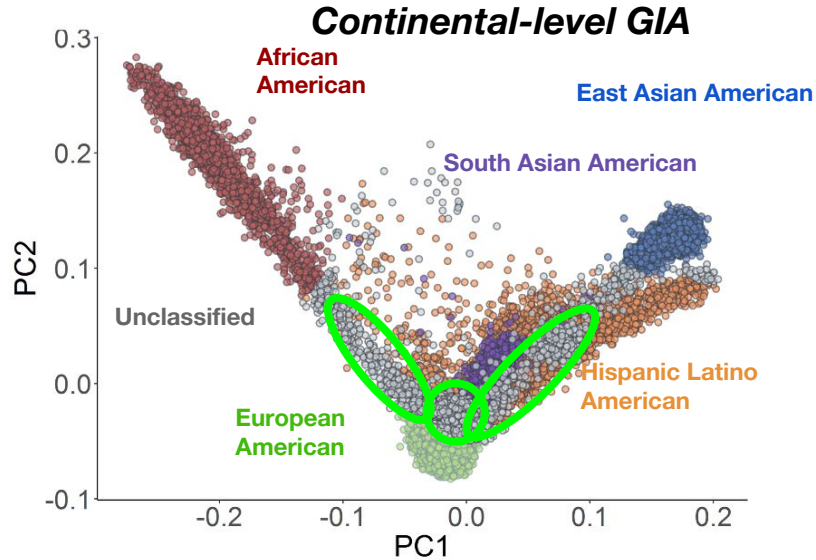
- Cline between African and European ancestry, and those who self-identify as African American along almost all of PC2
- GIA form a much tighter cluster, leaving many of the individuals who self-identified as African American outside this boundary.

PCA reveals notable differences between GIA and SIRE



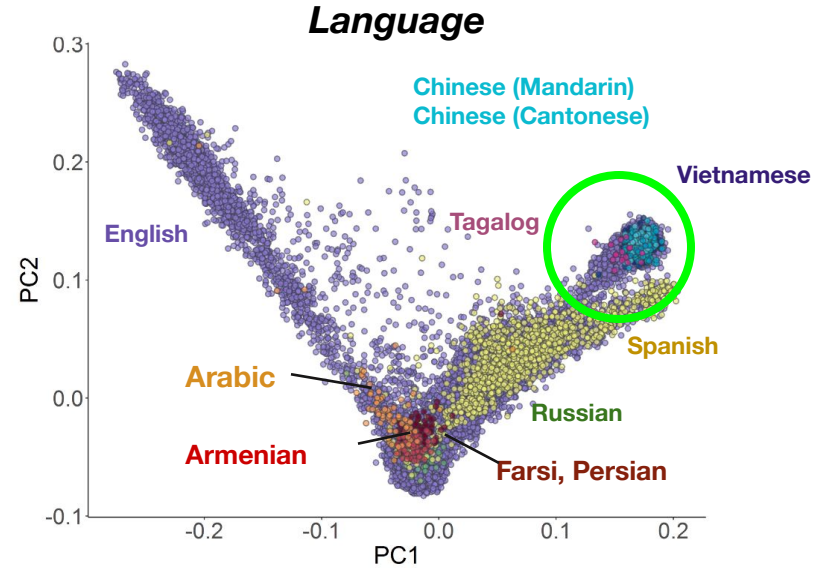
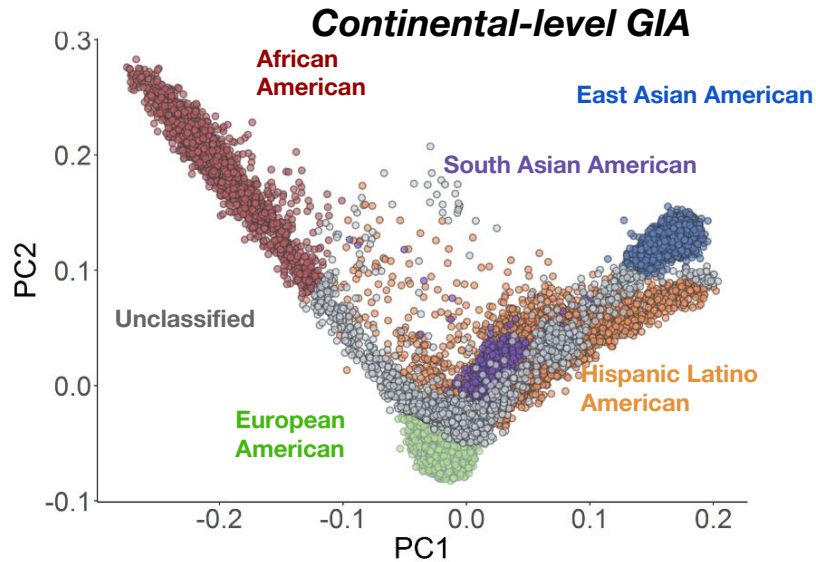
- There are also a large number of individuals that could not be assigned a GIA cluster and race/ethnicity information does not reveal any patterns either

Projecting individuals' preferred language onto PCs reveals individuals likely with Middle Eastern ancestry



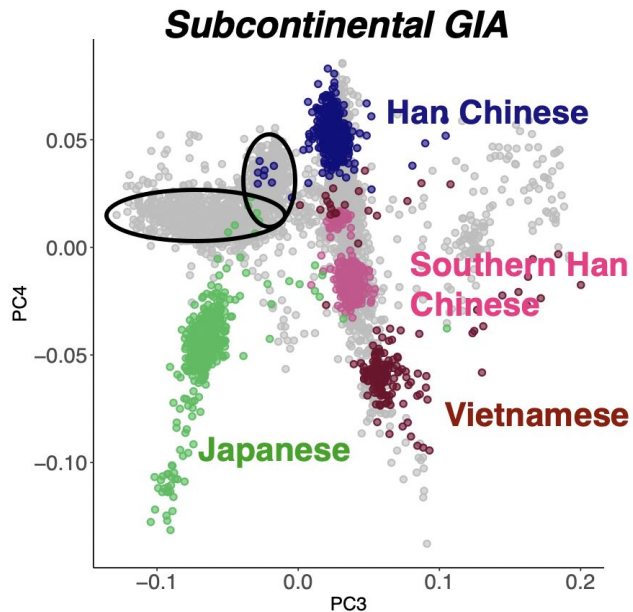
- Additional EHR information such as **“Language”** can help elucidate uncharacterized GIA groups

Projecting individuals' preferred language onto PCs reveals substructure within continental GIA clusters



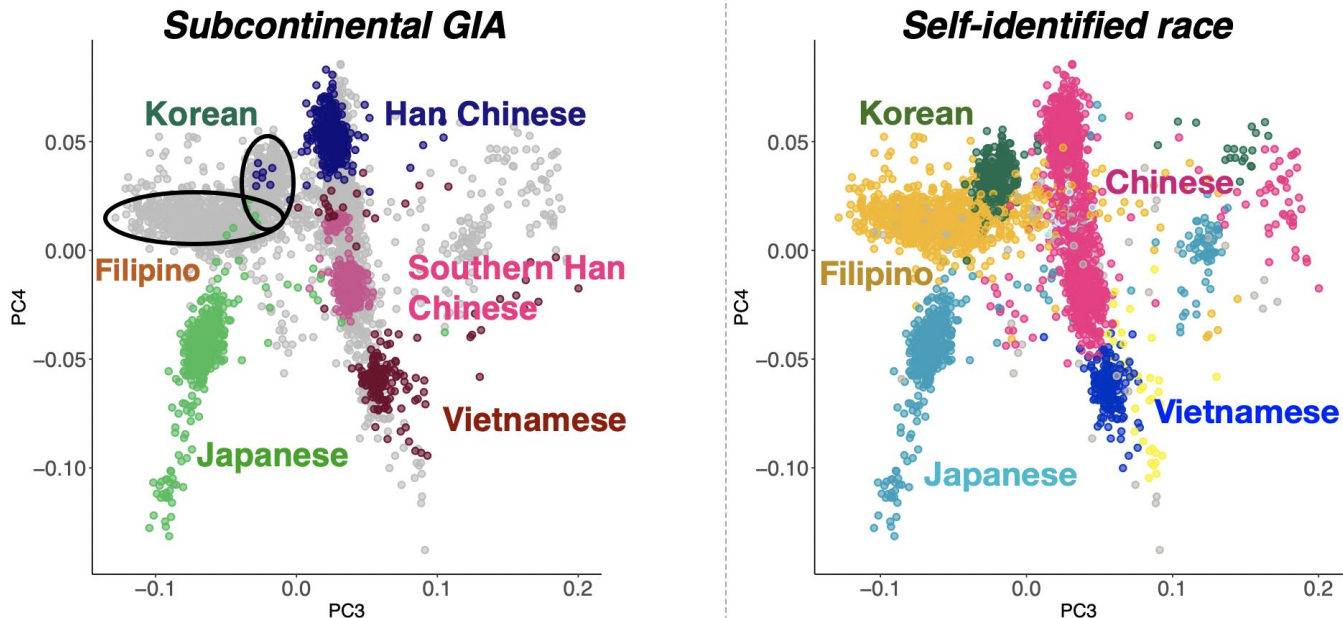
- Within the East Asian American GIA group, there are a variety of different languages represented, such as **Mandarin**, **Cantonese**, **Vietnamese**, and **Tagalog**

PCA identifies fine-scale population structure within the East Asian American GIA group



Using information from 1000 Genomes, we can see distinct clusters of individuals of Japanese, Vietnamese, and Chinese descent, but there are two distinct clusters that could not be characterized.

PCA identifies fine-scale population structure within the East Asian American GIA group

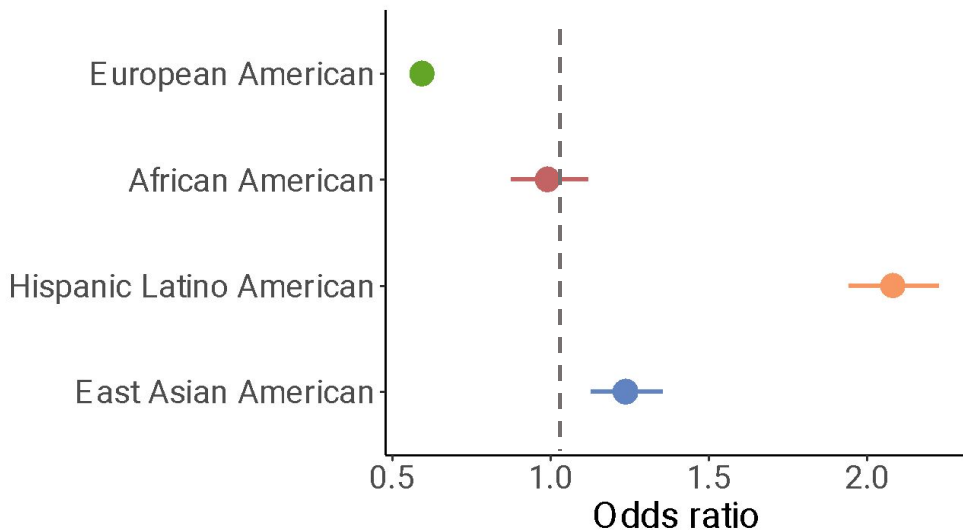


Self-identified race information projected onto these clusters reveals that these are likely individuals of Korean and Filipino descent

Associations between GIA and phenotypes remain even after accounting for SIRE

$$\text{logit}(\text{phecode}) = \beta_0 + \beta_1 \text{genetic_ancestry_group} + \beta_2 \text{sex} + \beta_3 \text{age} + \beta_4 \text{SIRE} \text{ [over all ATLAS individuals]}$$

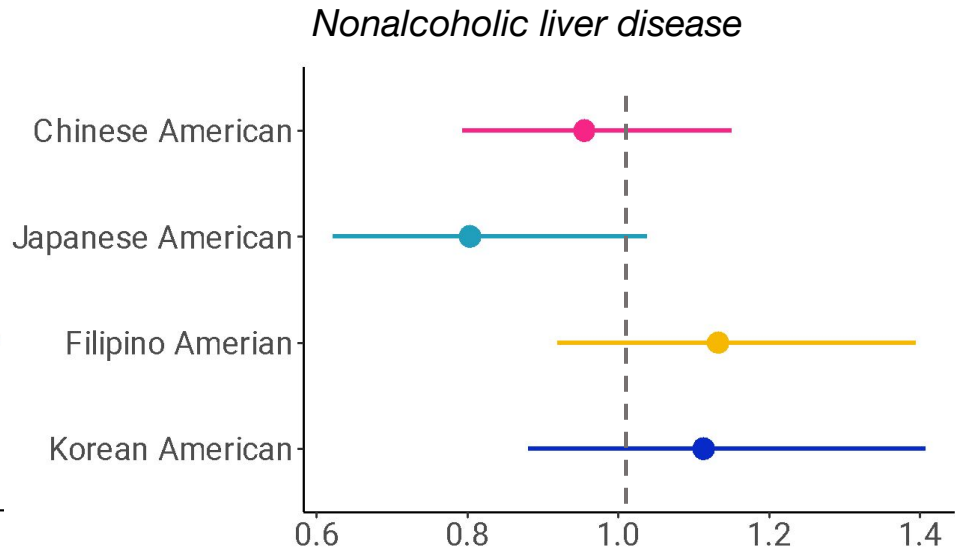
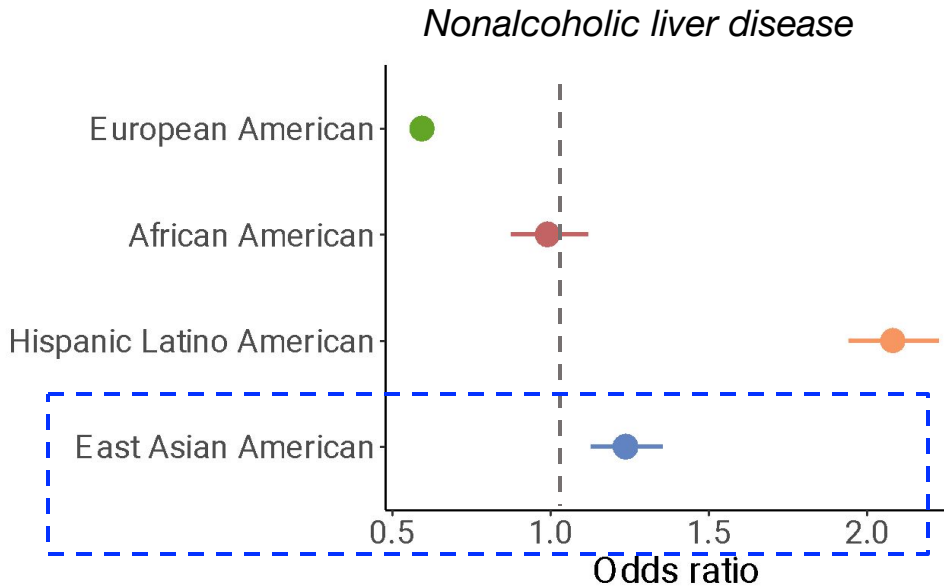
Nonalcoholic liver disease



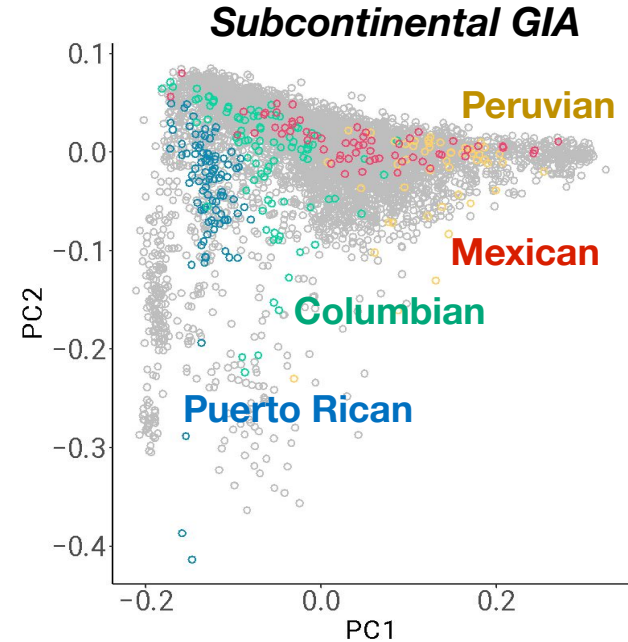
Associating each GIA group with disease status across 1,800 EHR-derived phenotypes (phecodes) yields a total of **259 significant associations** even after accounting for SIRE ($p\text{-value} < 1.12 \times 10^{-5}$)

Extensive genetic diversity within populations is intertwined with disease risk

- Enrichment in the East Asian American group is driven by the Filipino and Korean American groups
- Potential protective effect in the Chinese and Japanese American groups

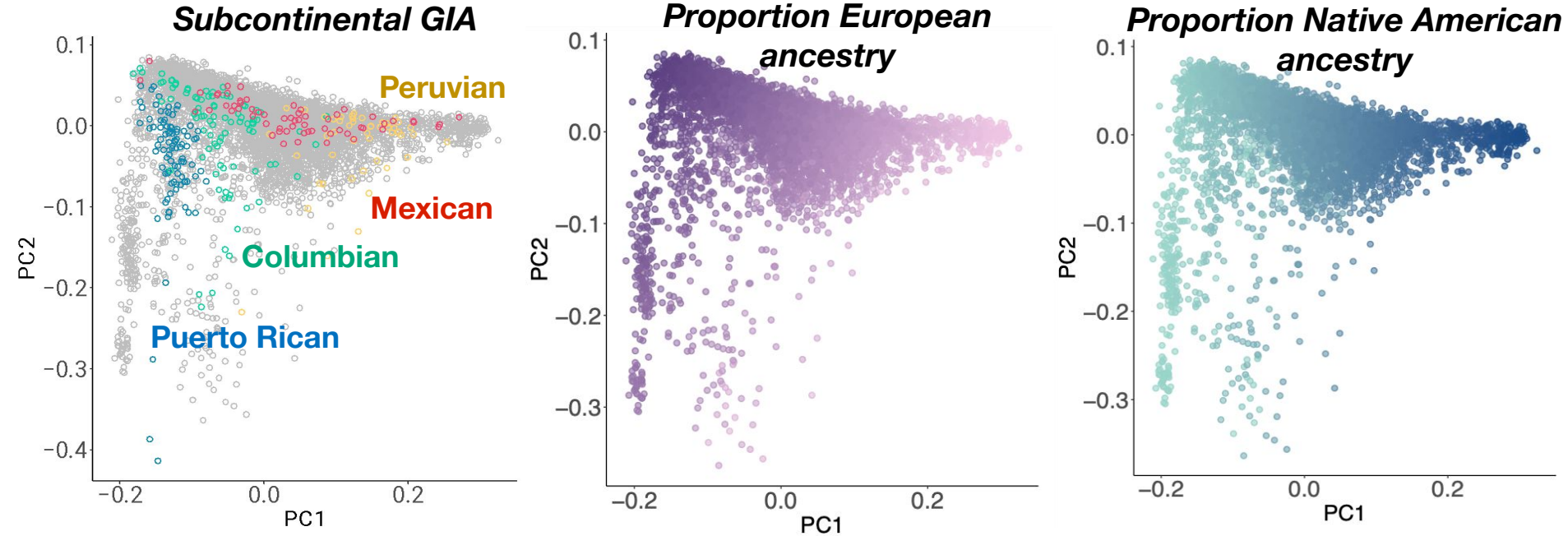


Characterizing genetic ancestry as a continuum is particularly relevant for admixed populations



Neither reference panel nor demographic information can elucidate any clusters of population structure in the Hispanic Latino American GIA group

Population structure beyond discrete clusters in the Hispanic Latino American GIA group

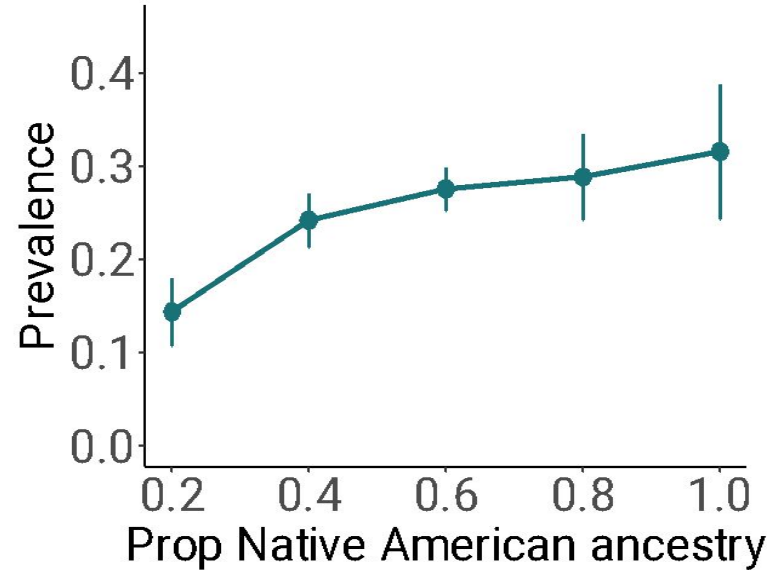
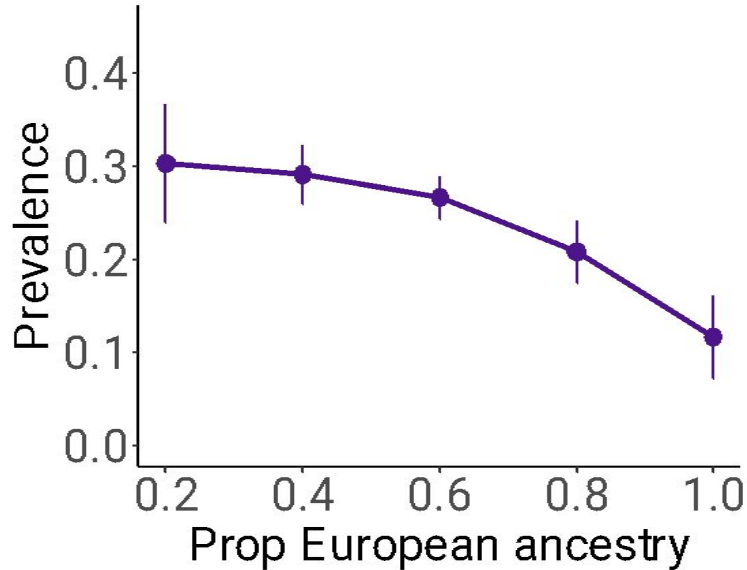


Population substructure is better characterized by the clines of European and Native American ancestry along PC1

Disease prevalence varies with genetic admixture proportions

424 significant ancestry proportion - phenotype associations out of 1,800 phecodes x 4 ancestry tests: European, African, East Asian, Native American (p -value $< 2.08 \times 10^{-5}$)

Nonalcoholic liver disease



Considering the actual proportion of ancestry when assessing disease risk can be more informative.

Conclusions

- There are **marked differences between race/ethnicity and genetically inferred ancestry**, emphasizing that the populations defined by these two criteria are not analogous
- There is **substantial disease risk heterogeneity across subgroups** of the same continental genetic ancestry group, both across subcontinental ancestry and genetic admixture
- Association analyses show possible **differential genetic architecture across populations**

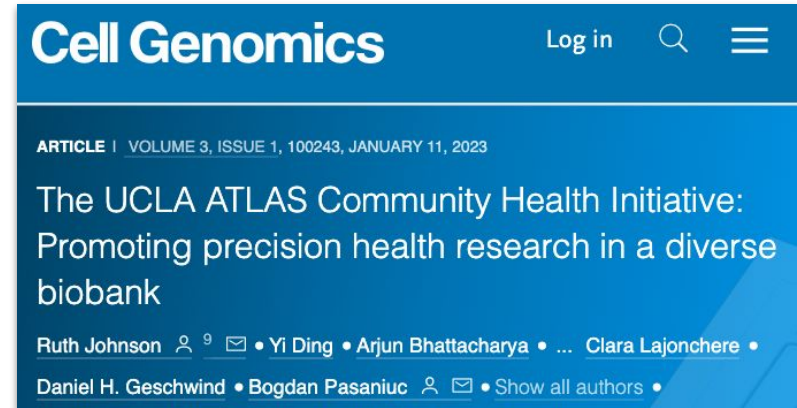


Genome Medicine

Home About Articles Submission Guidelines

Research | [Open Access](#) | [Published: 09 September 2022](#)

Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative







Cell Genomics

Log in 🔍 ☰

ARTICLE | [VOLUME 3, ISSUE 1, 100243, JANUARY 11, 2023](#)

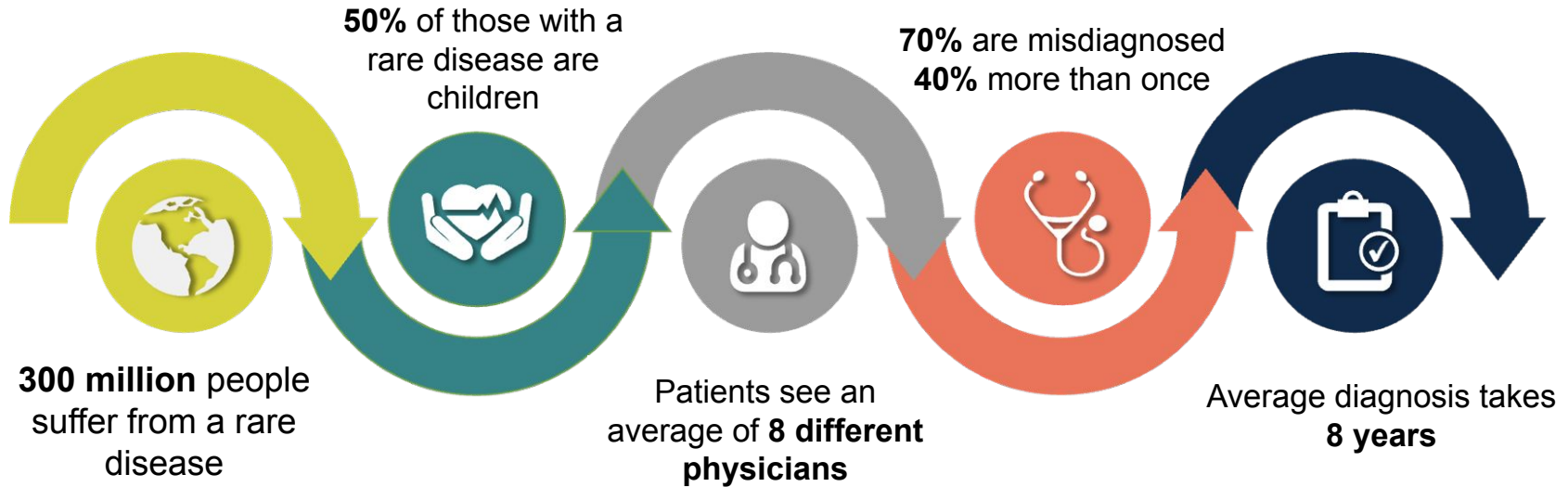
The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank

Ruth Johnson  ⁹  • Yi Ding • Arjun Bhattacharya • ... Clara Lajonchere • Daniel H. Geschwind • Bogdan Pasaniuc   • [Show all authors](#)

Part 3

Predicting rare disease through EHR signatures

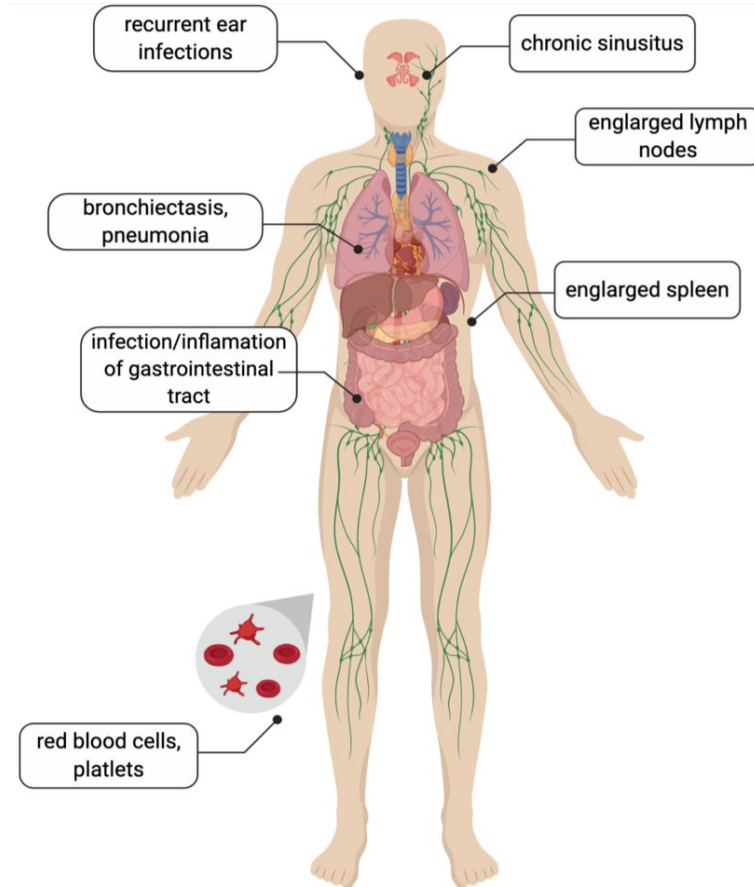
Current diagnostic odyssey for rare diseases is often prolonged by years due to misdiagnosis



Diagnostic odyssey causes the biggest delay in initiating treatment for rare disease patients

CVID is a rare, heterogenous immunodeficiency disorder

- Common Variable Immunodeficiency Disorders (CVID) is broadly characterized by recurrent viral and bacterial infections, but clinical manifestations are very heterogeneous
- **Occurs 1 in 25,000 to 1 in 50,000 people**
- Genetic basis of CVID is highly variable and largely unknown
- Majority of cases have an unknown cause and there are currently no specific mutations associated with a diagnosis



Heterogeneity of clinical manifestations leads to a diagnostic delays of **5-15 years**

- Clinical phenotypes of CVID intersect with virtually all medical specialties, making it difficult to pin down the immunogenic basis of the diagnosis
- Guidelines for recognizing CVID are very broad and limited as no single lab test can definitively determine a diagnosis
- Patients get 'lost' in specialty clinics where only a subset of their symptoms are treated

10 Warning Signs of Primary Immunodeficiency

Primary Immunodeficiency (PI) causes children and adults to have infections that come back frequently or are unusually hard to cure. 1:500 persons are affected by one of the known Primary Immunodeficiencies. **If you or someone you know is affected by two or more of the following Warning Signs, speak to a physician about the possible presence of an underlying Primary Immunodeficiency.**

- 1 Four or more new ear infections within 1 year.
- 2 Two or more serious sinus infections within 1 year.
- 3 Two or more months on antibiotics with little effect.
- 4 Two or more pneumonias within 1 year.
- 5 Failure of an infant to gain weight or grow normally.
- 6 Recurrent, deep skin or organ abscesses.
- 7 Persistent thrush in mouth or fungal infection on skin.
- 8 Need for intravenous antibiotics to clear infections.
- 9 Two or more deep-seated infections including septicemia.
- 10 A family history of PI.



Jeffrey Modell Foundation | Curing PI Worldwide



Presented as a public service by:

Funding was made possible in part by grant 5H75DP22546-05 from the United States Centers for Disease Control and Prevention (CDC)



National Institute of Allergy and Infectious Diseases (NIAID)



National Institute of Child Health and Human Development (NICHD)



These warning signs were developed by the Jeffrey Modell Foundation Medical Advisory Board. Consultation with Primary Immunodeficiency experts is strongly suggested. © 2009 Jeffrey Modell Foundation
For information or referrals, contact the Jeffrey Modell Foundation: 866-INFO-4-PI | info4pi.org

Aggregating phenotypes prioritizes patients with CVID

Acute upper respiratory infections
of multiple or unspecified sites –
All: 13%



Chronic sinusitis – All:
All: 4.45%



Asthma –
All: 10%



Bronchiectasis -
All: 0.6%



→ some phenotypes are relatively common in the general patient population, but are even more highly enriched in the CVID population.

Aggregating phenotypes prioritizes patients with CVID

Acute upper respiratory infections
of multiple or unspecified sites –
All: 13% (CVID: 24%)



Chronic sinusitis – All:
All: 4.45% (CVID: 48%)



Asthma –
All: 10% (CVID: 42%)



Bronchiectasis -
All: 0.6% (CVID: 23%)



→ some phenotypes are relatively common in the general patient population, but are even more highly enriched in the CVID population.

Aggregating phenotypes prioritizes patients with CVID

Acute upper respiratory infections
of multiple or unspecified sites –
All: 13% (CVID: 24%)



Chronic sinusitis – All:
All: 4.45% (CVID: 48%)



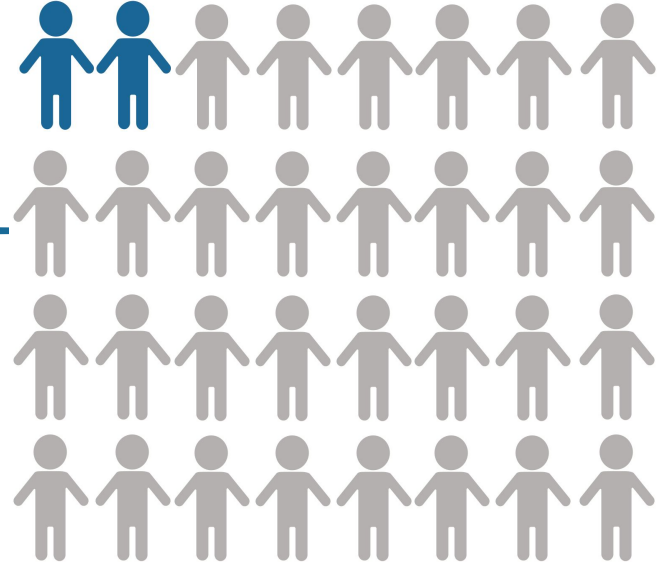
Asthma –
All: 10% (CVID: 42%)



Bronchiectasis -
All: 0.6% (CVID: 23%)



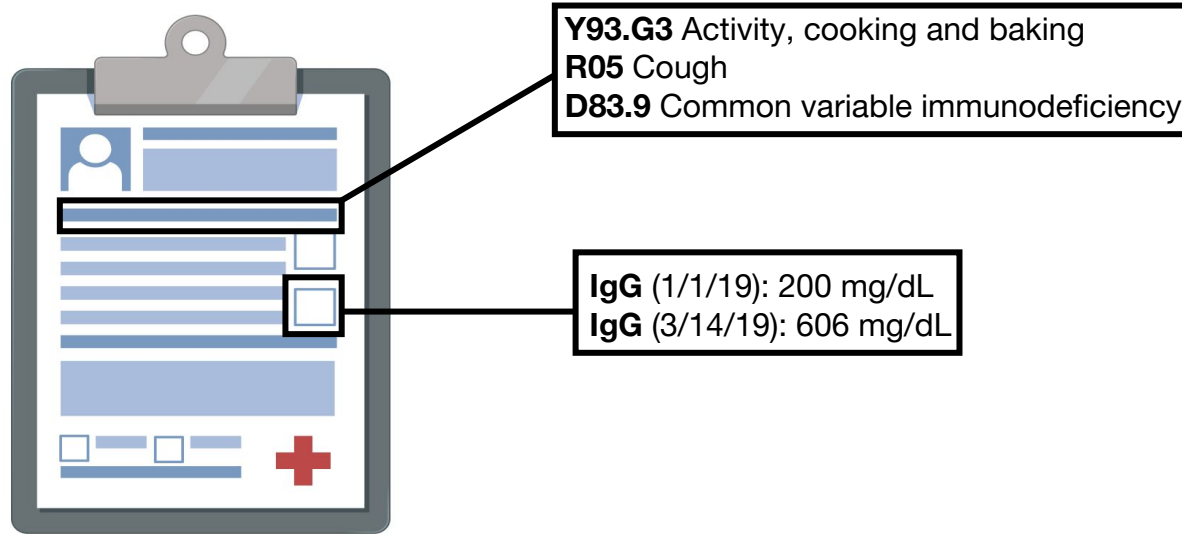
Combination of all four –
All: 0.02% (CVID: 4%)



→ some phenotypes are relatively common in the general patient population, but are even more highly enriched in the CVID population.

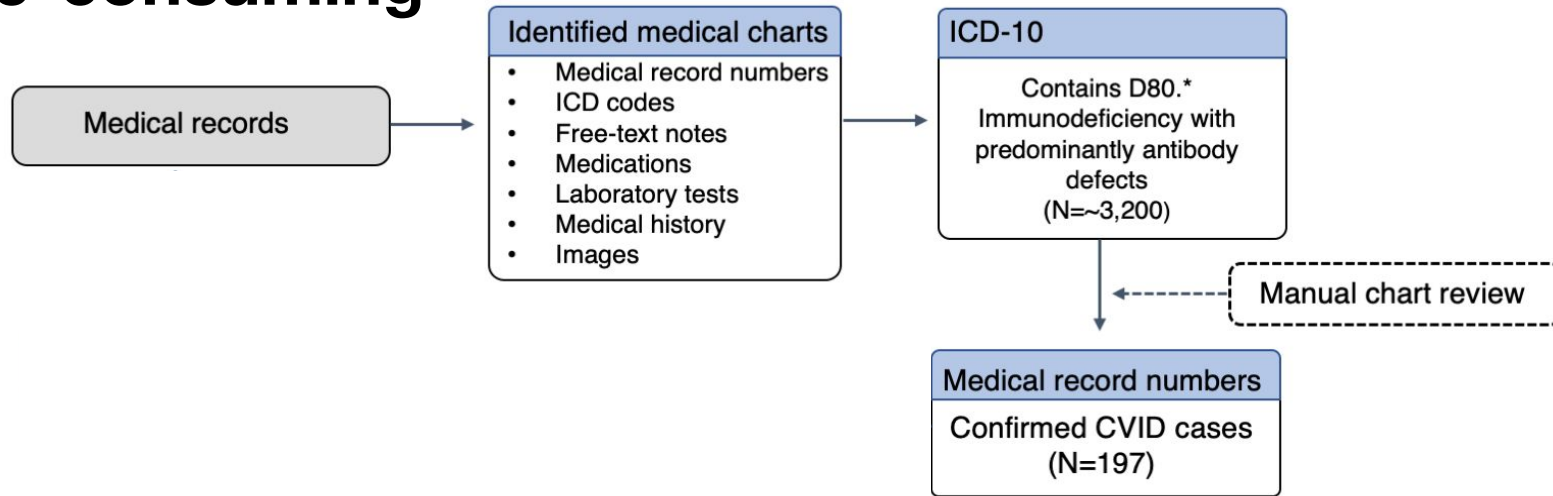
EHR-signatures describe key characteristics of a disease and how it is represented in the EHR

A major bottleneck is identifying a set of EHR-derived features that characterize COVID-- no one test or feature within the medical data that definitively describes individuals with COVID



Need to look at the combination of various phenotypes in the medical record, not just the absence or presence of a single set

Obtaining high-quality labeled cases is challenging and time-consuming



- Initial set of patients with any type of immunodeficiency are selected and then are manually reviewed to determine the diagnosis
- Extreme case data imbalance: **197** cases, **1 million** controls
- A key concern is overfitting, where the model can simply 'memorize' the cases because there are so few of them and so many features in the EHR

Feature selection to identify a set of features to accurately predict CVID

OMIM clinical description

```
# 607594
IMMUNODEFICIENCY, COMMON VARIABLE, 1; CVID1

INHERITANCE
- Autosomal recessive

RESPIRATORY
Airways
- Bronchiectasis
- Bronchitis, recurrent

HEAD & NECK
Head
- Sinusitis, recurrent
Ears
- Otitis media, recurrent
Eyes
- Conjunctivitis

Lung
- Pneumonia, recurrent
```

- Utilize existing clinical databases that act as a proxy for learned information regarding CVID phenotype patterns

Feature selection to identify a set of features to accurately predict CVID

OMIM clinical description

607594
IMMUNODEFICIENCY, COMMON VARIABLE, 1; CVID1

INHERITANCE - Autosomal recessive	RESPIRATORY <i>Airways</i> - Bronchiectasis - Bronchitis, recurrent
HEAD & NECK <i>Head</i> - Sinusitis, recurrent	<i>Lungs</i> - Pneumonia, recurrent
<i>Ears</i> - Otitis media, recurrent	
<i>Eyes</i> - Conjunctivitis	

HP:0002090

HP:0000246

- Utilize existing clinical databases that act as a proxy for learned information regarding CVID phenotype patterns
- Clinical descriptions are annotated with HPO terms which is mapped to diagnosis codes listed in the EHR

Feature selection to identify a set of features to accurately predict CVID

OMIM clinical description

607594
IMMUNODEFICIENCY, COMMON VARIABLE, 1; CVID1

INHERITANCE - Autosomal recessive	RESPIRATORY <i>Airways</i> - Bronchiectasis - Bronchitis, recurrent
HEAD & NECK <i>Head</i> - Sinusitis, recurrent	<i>Lungs</i> - Pneumonia, recurrent
<i>Ears</i> - Otitis media, recurrent	
<i>Eyes</i> - Conjunctivitis	

HP:0000246

HP:0002090



Diagnosis codes

- Hypothyroidism (244.0)
- Adrenal hypofunction (255.2)
- Other arthropathies (716.0)
- Psoriasis (696.4)
- Acquired hemolytic anemias (283.0)

34 EHR features

(Bastarache et al. Science 2018)

- Utilize existing clinical databases that act as a proxy for learned information regarding CVID phenotype patterns
- Clinical descriptions are annotated with HPO terms which is mapped to diagnosis codes listed in the EHR
- The OMIM database provides 34 EHR-derived features without ever looking at the training data

UCLA-specific data capture unique phenotyping patterns

OMIM clinical description

607594
IMMUNODEFICIENCY, COMMON VARIABLE, 1; CVID1

INHERITANCE
- Autosomal recessive

HEAD & NECK
Head
- Sinusitis, recurrent
Ears
- Otitis media, recurrent
Eyes
- Conjunctivitis

RESPIRATORY
Airways
- Bronchiectasis
- Bronchitis, recurrent
Lungs
- Pneumonia, recurrent

HP:0000246

HP:0002090

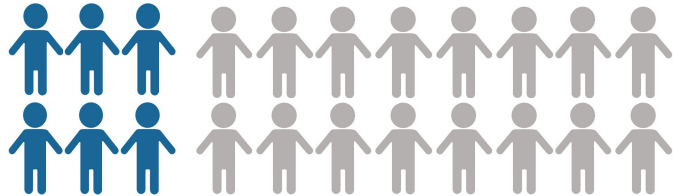


Diagnosis codes

- Hypothyroidism (244.0)
- Adrenal hypofunction (255.2)
- Other arthropathies (716.0)
- Psoriasis (696.4)
- Acquired hemolytic anemias (283.0)

34 EHR features

(Bastarache et al. Science 2018)



1. Chronic sinusitis (All: 4.45%, CVID: 48%)
2. Asthma (All: 10%, CVID: 42%)
- ...
10. Bronchiectasis (All: 0.6%, CVID: 23%)

+10 EHR features

PheNet scores reflect how closely patients' EHR matches patterns of CVID

Score weights are inferred by performing a marginal regression for each feature

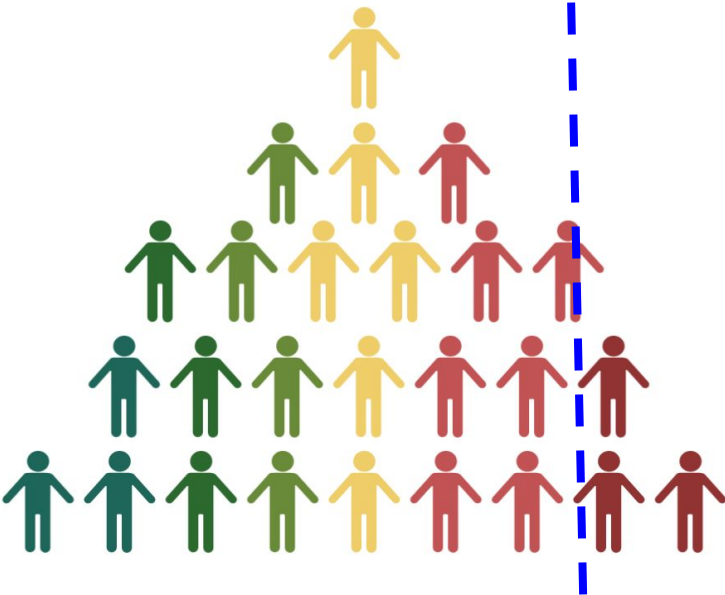
$$\mathbf{x}_1 \beta_{\text{Sinusitis}} + \mathbf{x}_2 \beta_{\text{Pneumonia}} + \mathbf{x}_3 \beta_{\text{Asthma}} + \dots + \mathbf{x}_N \beta_{\text{IgG}} = 0.99$$



PheNet model maintains interpretability of the results

Clinical predictions require a lot of trust and transparency for both clinicians and patients

$$x_1 \beta_{\text{Sinusitis}} + x_2 \beta_{\text{Pneumonia}} + x_3 \beta_{\text{Asthma}} + \dots + x_N \beta_{\text{IgG}} = 0.99$$



Patient follow-up

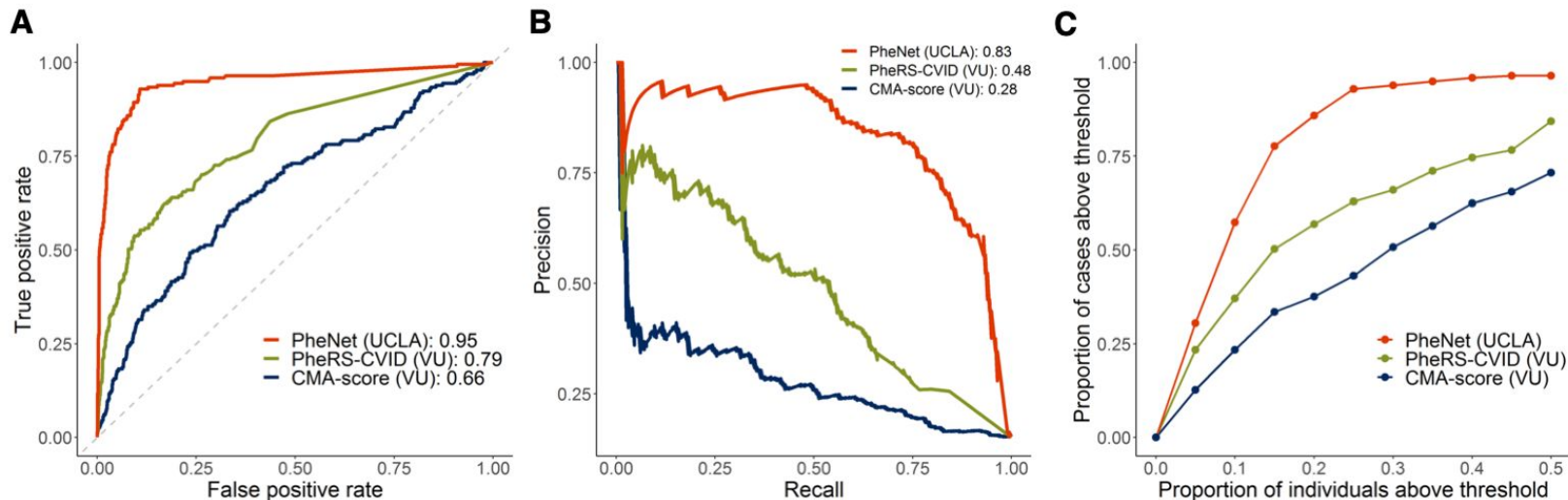
Dear Provider,

XXXX XXX XX XX X XXXX X
XX XX XX XX X XX:
- Sinusitis
- Low IgG

XX XX XX XX X XXX XXXX
XX

xx

PheNet outperforms previous state-of-the-art methods

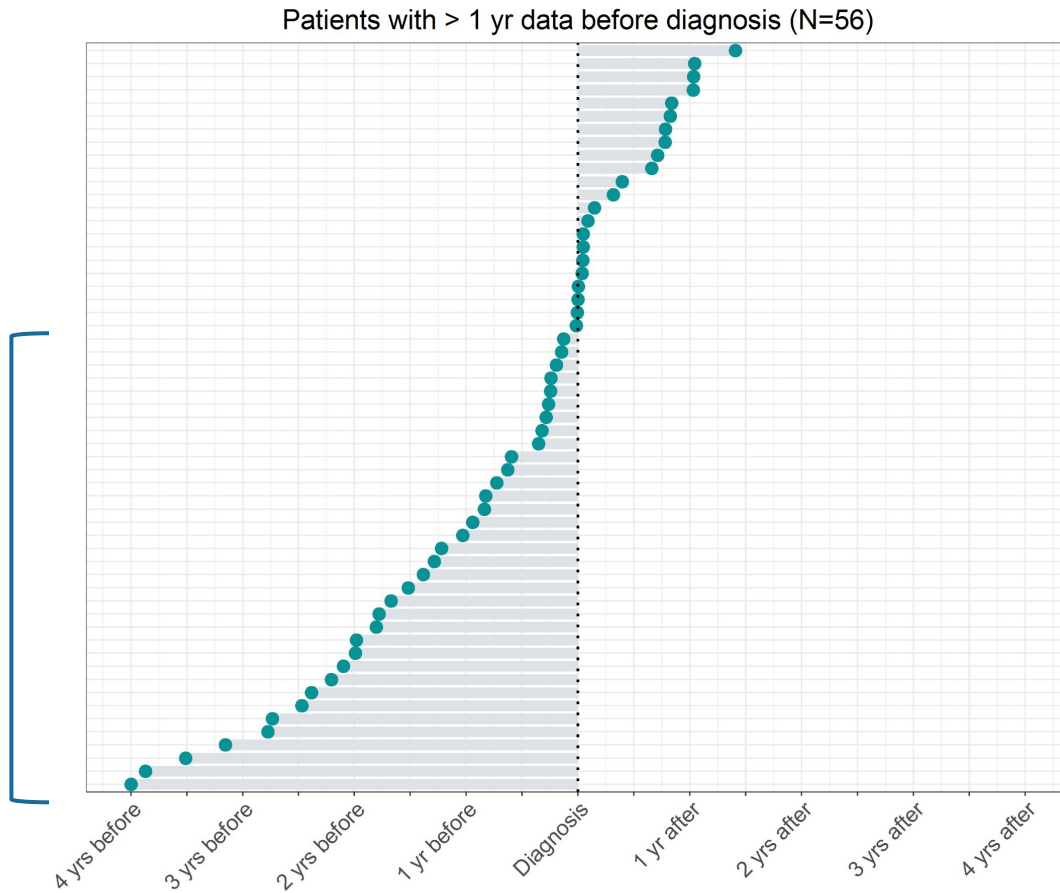


- PheNet performs **17%-31% better** when comparing AUC-ROC and **42%-66% better** when comparing AUC-PR
- Top 10% of individuals with the highest PheNet score captures **60% of CVID cases** whereas previous methods only captures **24%-45% of cases**.

Retrospective study shows PheNet can identify CVID patients before their formal clinical diagnosis

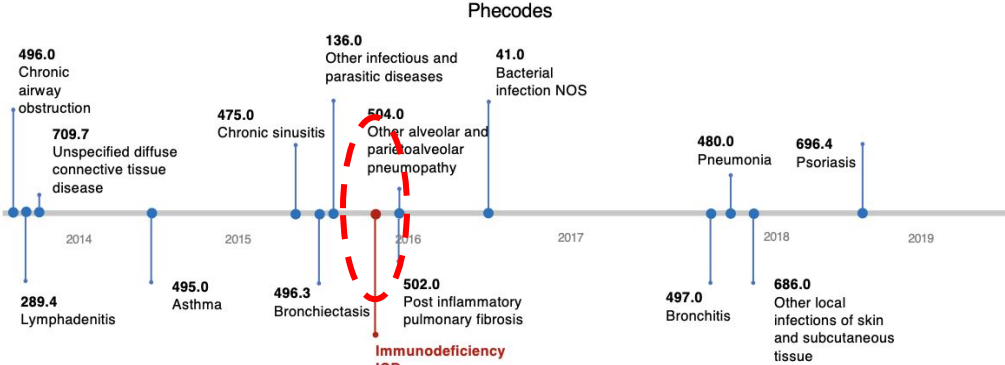
- PheNet would have identified **64% of individuals** with CVID before their original diagnosis
- Average gap between the date of diagnosis and the date identified by PheNet **244 days** (SD: 374).

Identified by PheNet prior to original clinical diagnosis

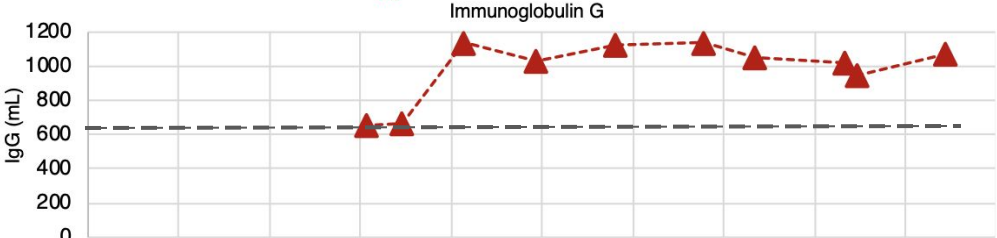


Example patient shows patterns of CVID months before diagnosis

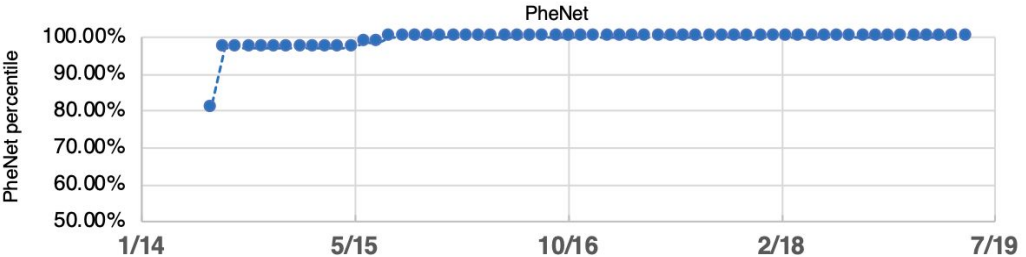
Diagnosis codes



Antibody laboratory tests

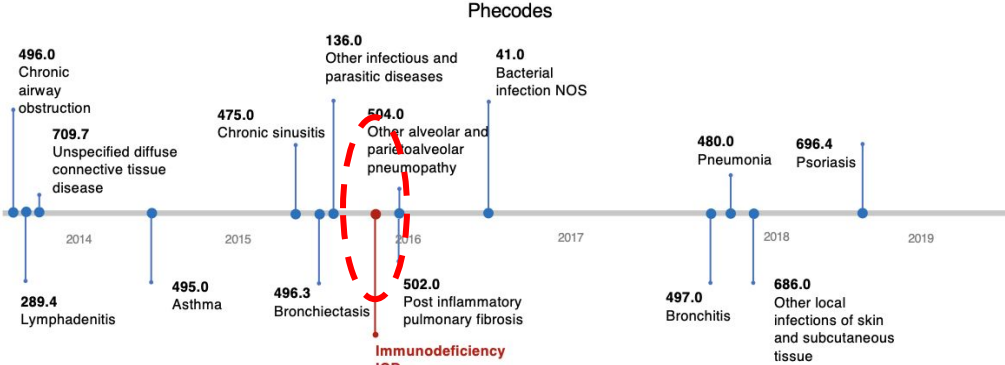


Phenotype risk score percentile

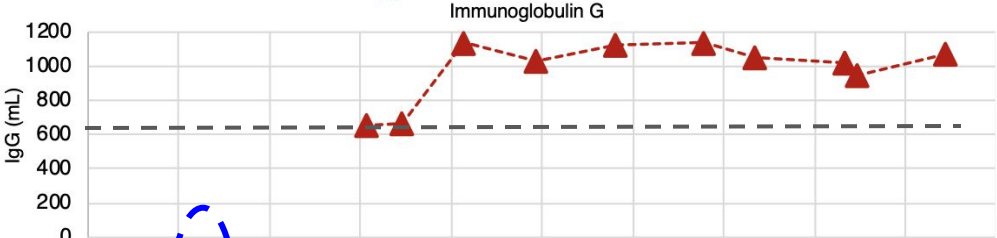


Example patient shows patterns of CVID months before diagnosis

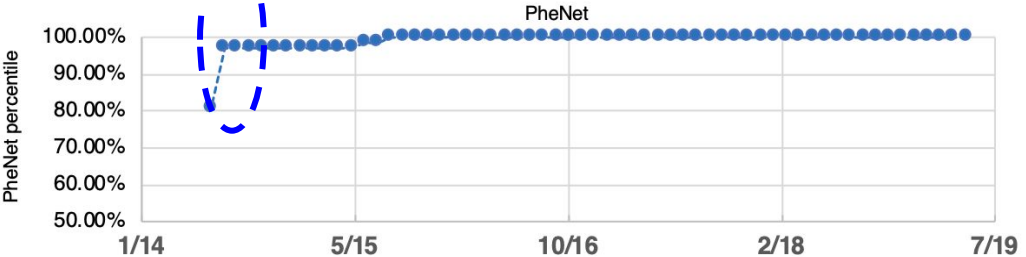
Diagnosis codes



Antibody laboratory tests

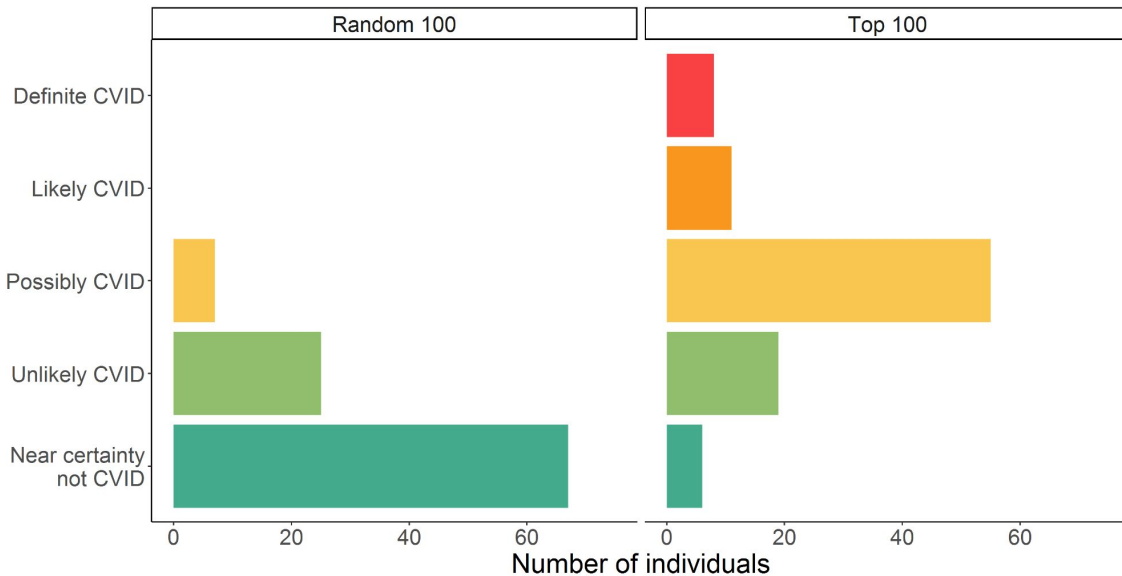


Phenotype risk score percentile



Top ranked PheNet patients have probable CVID according to an immune specialist blinded chart review

- Top 100 PheNet patients and random 100 patients were given to an immunologist for a blind chart review
- From the top 100 ranked individuals, **73% highly probable (scores 1-5)** as having CVID and **8% positively diagnosed (score 5)** with CVID



PheNet identifies prospective CVID patients across 5 UC institutions through a \$5 million grant



National Institute of Allergy and Infectious Diseases

Collaborative multi-site project to speed the identification and management of rare genetic immune diseases

Butte, Manish J. Pasaniuc, Bogdan

University of California Los Angeles, Los Angeles, CA, United States

UC DAVIS
HEALTH

UC Davis ★ Sacramento

UCSF Health

UC Santa Cruz

UC Berkeley

UC Merced

UCLA Health

UC Santa Barbara

UCLA

UC Riverside

UC Irvine

UCI Health

UC San Diego

UC San Diego Health

Coordinating a multi-site collaboration

Run in Python in Azure VM

UCLA Data Discovery
Repository (de-identified)

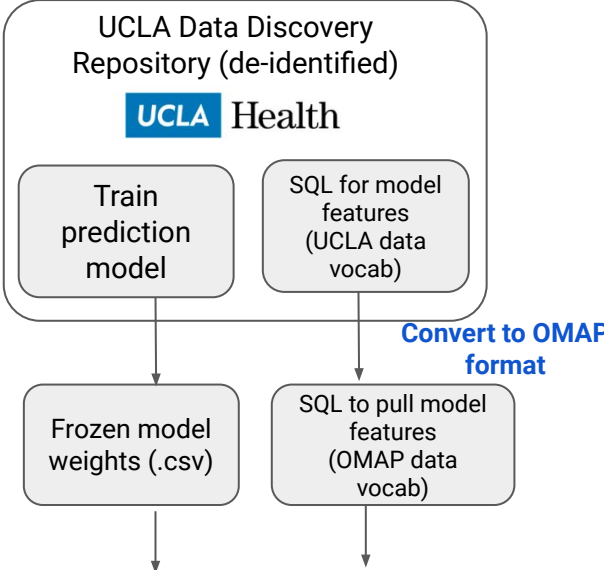
UCLA Health

Train
prediction
model

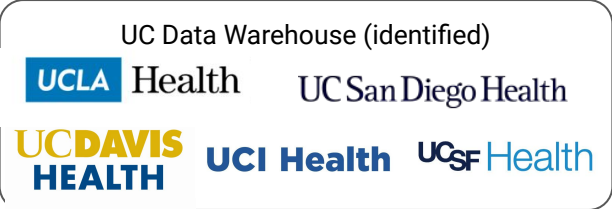
SQL for model
features
(UCLA data
vocab)

Coordinating a multi-site collaboration

Run in Python in Azure VM



Convert to OMAP format



Run in Python in Databricks environment

Coordinating a multi-site collaboration

Run in Python in Azure VM

UCLA Data Discovery Repository (de-identified)

UCLA Health

Train prediction model

SQL for model features (UCLA data vocab)

Convert to OMAP format

Frozen model weights (.csv)

SQL to pull model features (OMAP data vocab)

UC Data Warehouse (identified)

UCLA Health UC San Diego Health

UCDAVIS HEALTH **UCI** Health **UCSF** Health

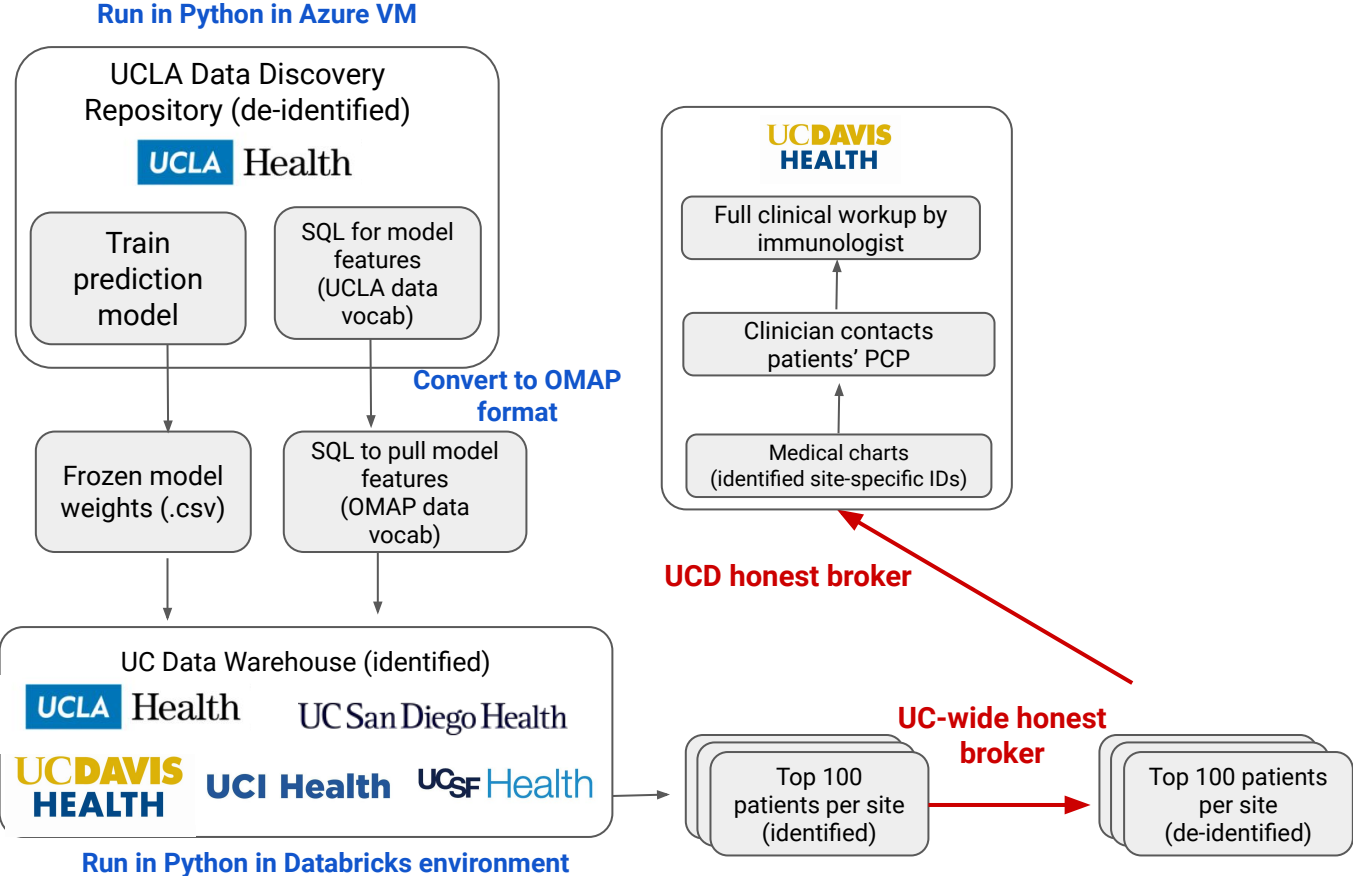
Run in Python in Databricks environment

Top 100 patients per site (identified)

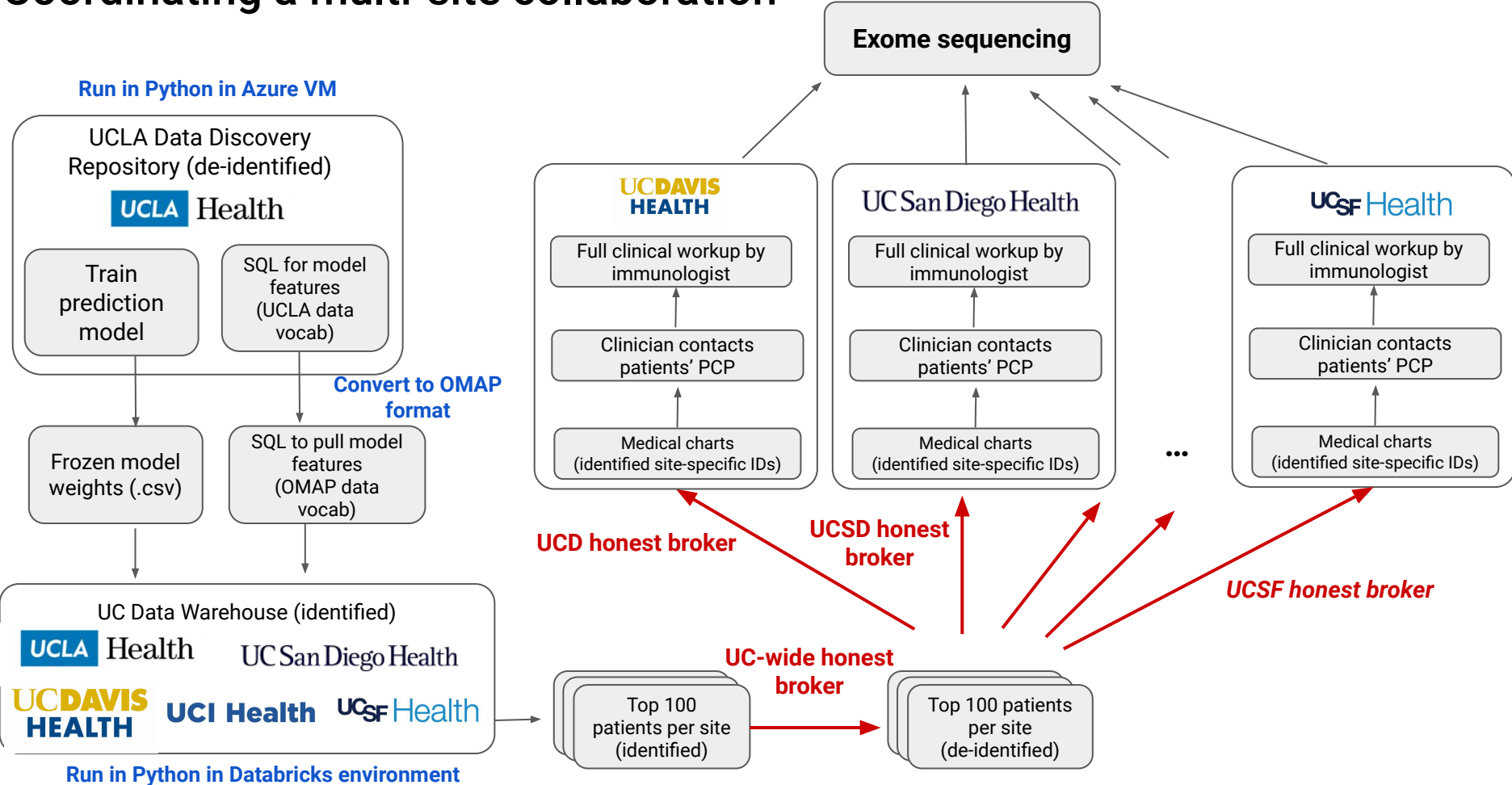
UC-wide honest broker

Top 100 patients per site (de-identified)

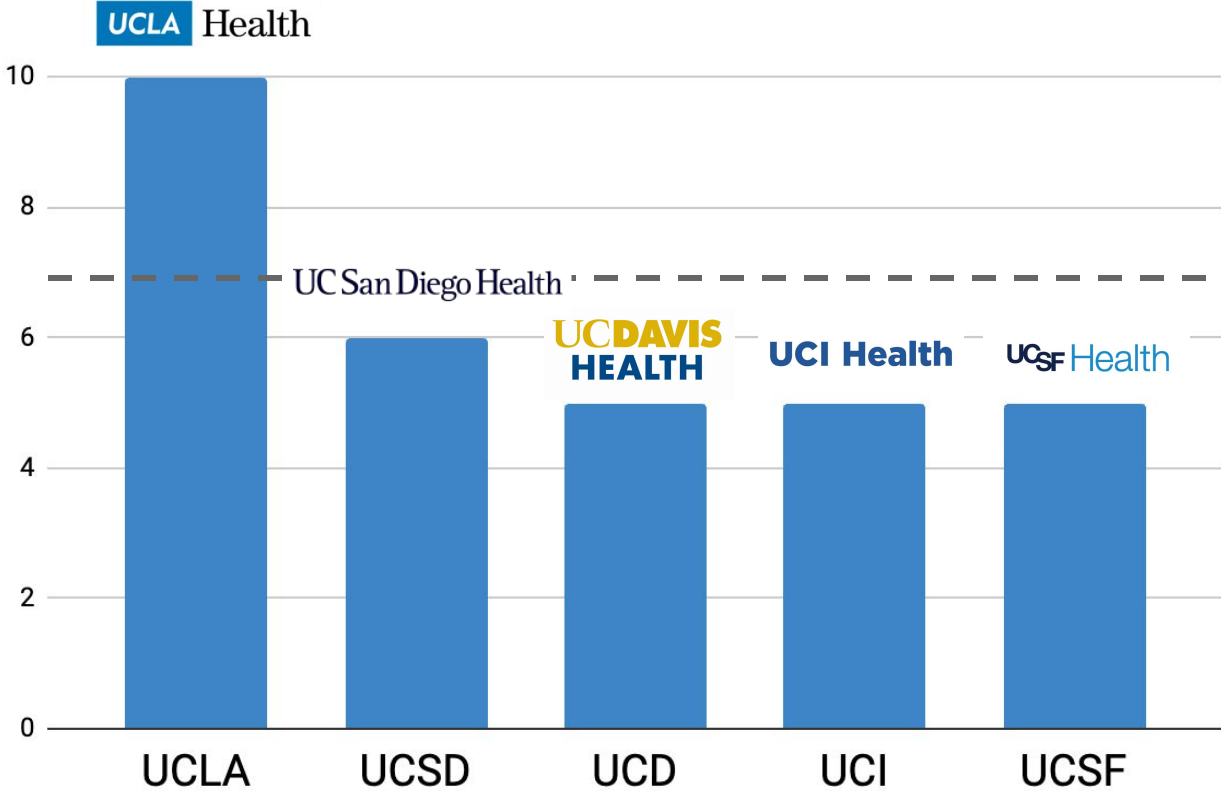
Coordinating a multi-site collaboration



Coordinating a multi-site collaboration



PheNet identifies prospective CVID patients across 5 UC institutions through a \$5 million grant





Patients that have visited for a full immunological evaluation

Year 1 goal

Conclusions

- EHR-signatures **leverage common patterns of phenotypes** to prioritize patients with rare disorders
- **64% of CVID patients** could have been identified by PheNet more than **8 months earlier** than they had been clinically diagnosed
- PheNet is validated across **5 additional UC health systems** to identify new CVID patients

Electronic health record signatures identify undiagnosed patients with Common Variable Immunodeficiency Disease

 Ruth Johnson, Alexis V. Stephens, Sergey Knyazev, Lisa A. Kohn, Malika K. Freund, Leroy Bondhus, Brian L. Hill,  Tommer Schwarz, Noah Zaitlen,  Valerie A. Arboleda, Manish J. Butte, Bogdan Pasaniuc

Collaborators

- Manish J. Butte
- Alexis Stephens
- Yi Ding
- Vidhya Venkateswaran
- Arjun Bhattacharya
- Alec Chiu
- Tommer Schwarz
- Alex Flynn-Carroll
- Malika Freund
- Lingyu Zhan
- Jonatan Hervoso
- Kangcheng Hou
- Kathryn S. Burch
- Christa Caggiano
- Brian Hill
- Humza Khan
- Yang Wu
- Ziqi Xu
- Sergey Knyazev
- Yung-Han (Tina) Chang
- Claudia Giambartolomei
- Sandra Lapinska
- Rachel Mester
- Ella Petter
- Helen Shang
- Valerie A. Arboleda
- Nadav Rakocz
- Brunilda Balliu
- Jae Hoon Sul
- Ariel Wu
- Noah Zaitlen
- Eran Halperin
- Clara Lajonchere
- Daniel H. Geschwind



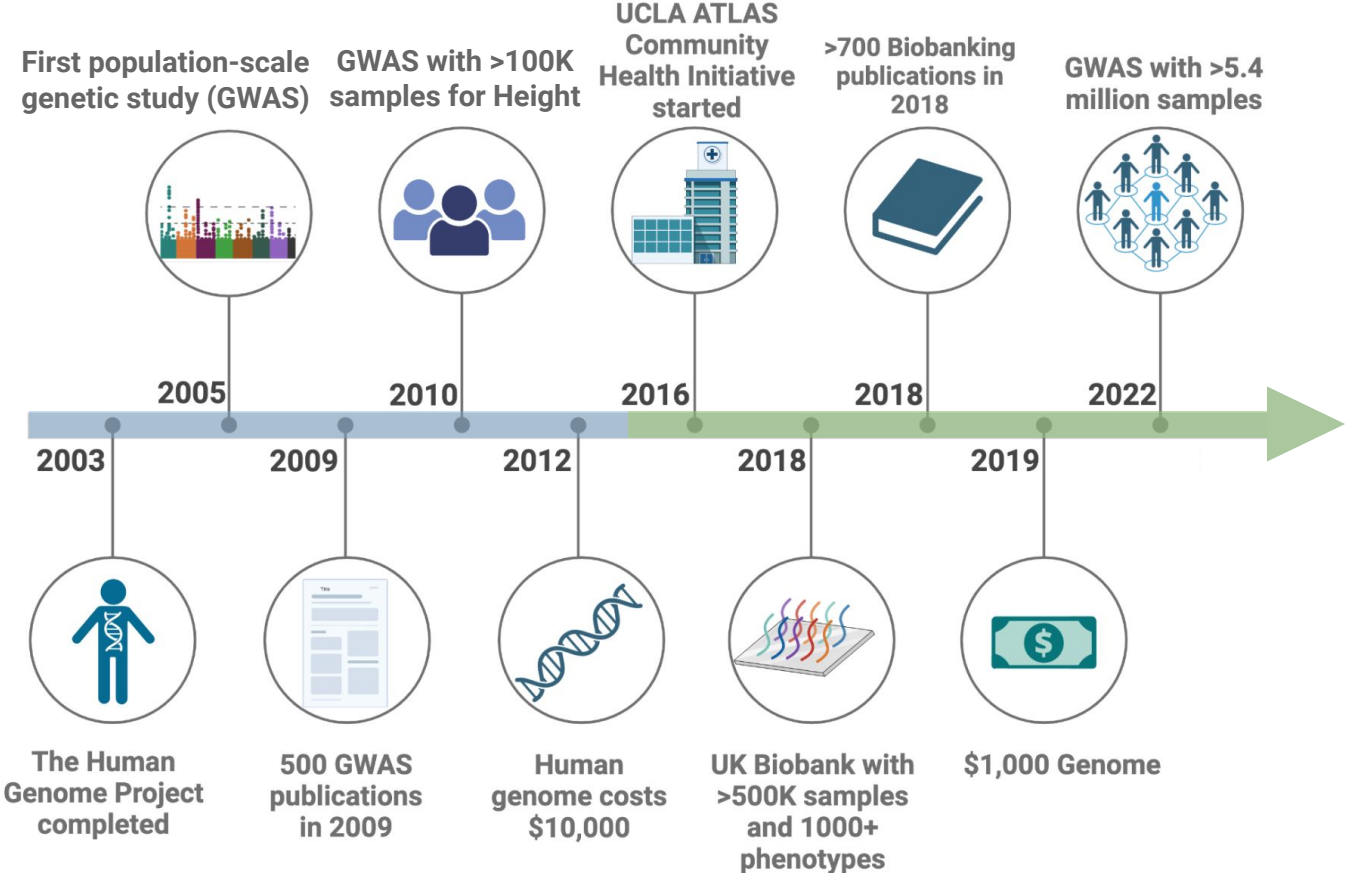
UCLA patients

- Maryam Ariannejad
- Joe DeYoung
- Vivek Katakwar
- Nicholas Mancuso
- Huwenbo Shi
- Megan Roytman
- Gleb Kichaev
- Rob Brown
- Utku Acar MD
- Maria Garcia-Lloret
- UCLA Precision Health Data Discovery Repository Working Group
- UCLA Precision Health ATLAS Working Group
- Office of Health Informatics and Analytics (OHIA)

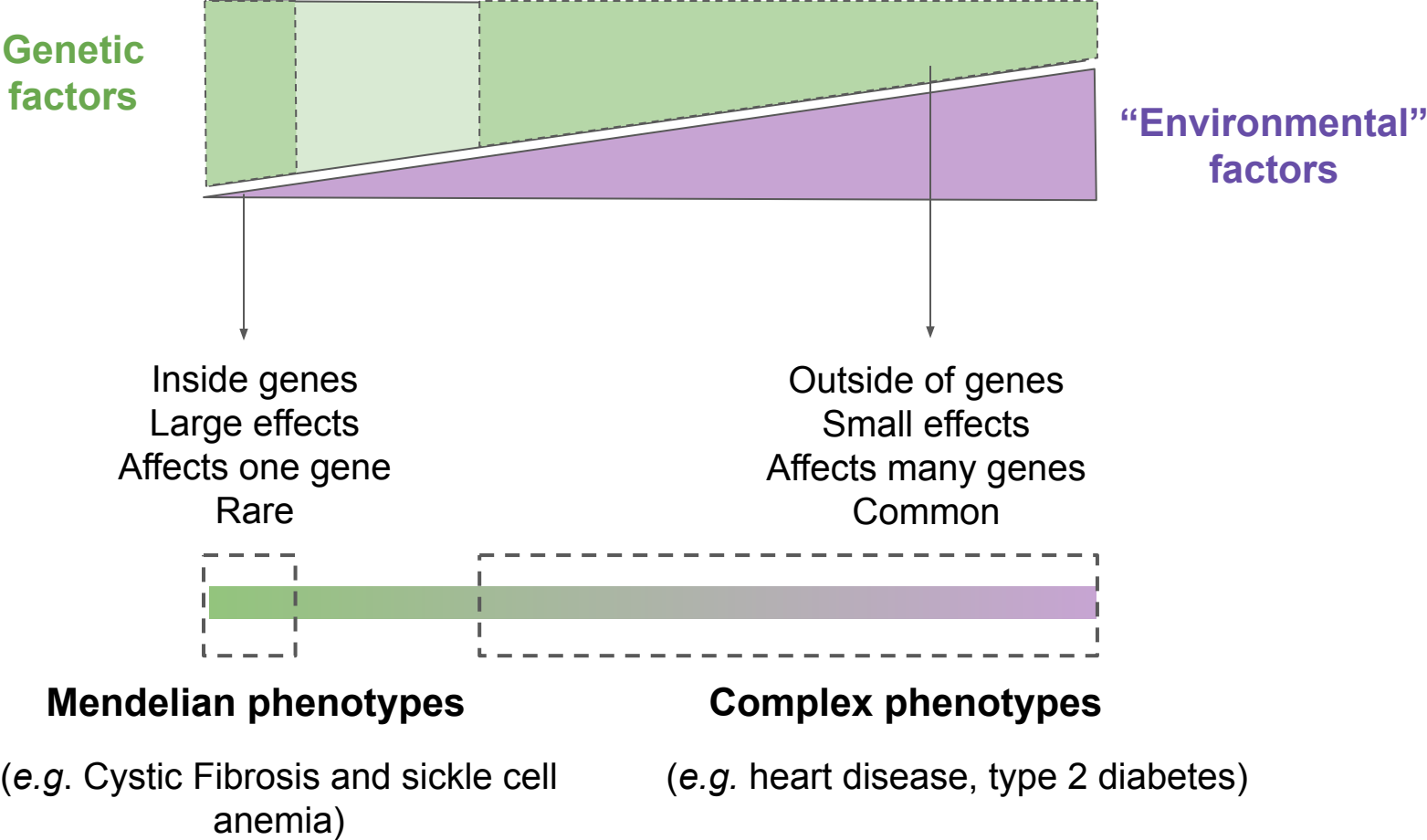
Questions, Comments, Concerns?

ruth_johnson@hms.harvard.edu

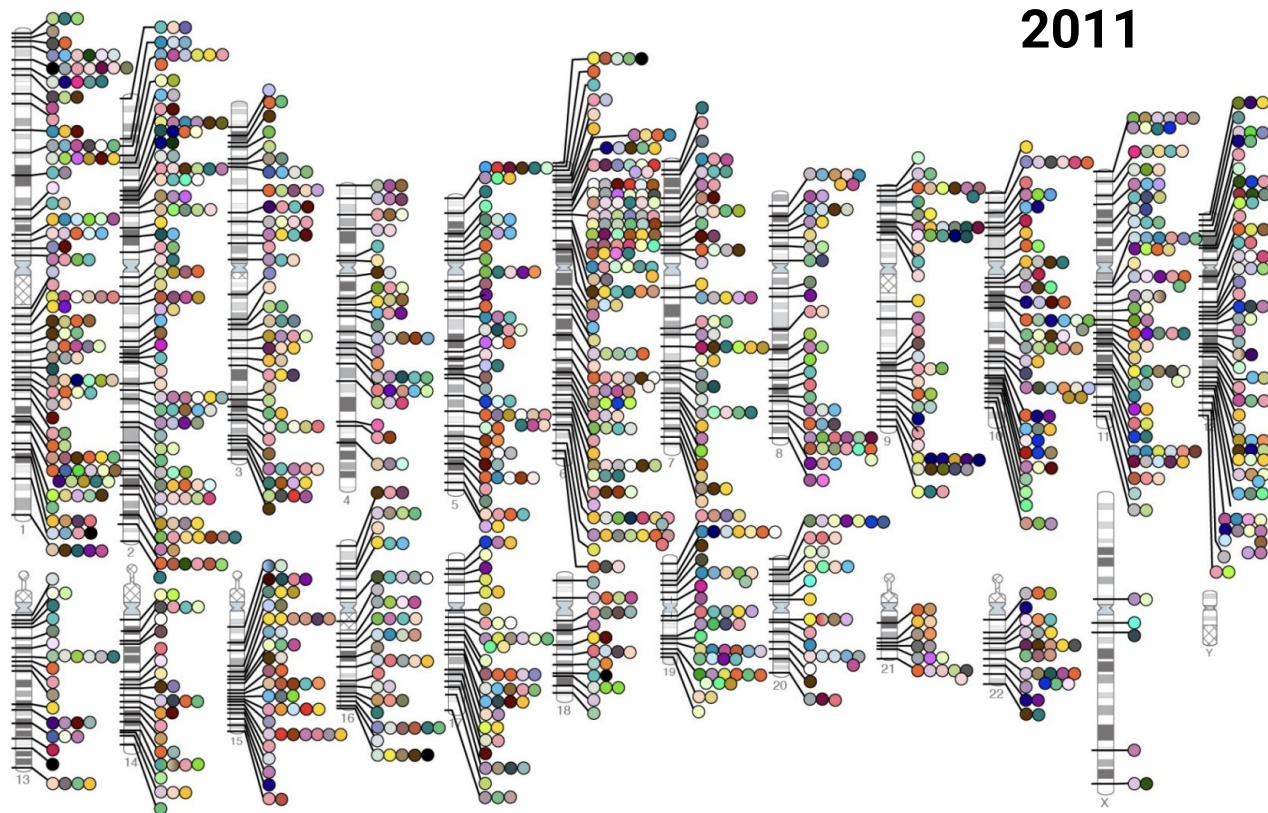
A brief recap of genomics since the Human Genome Project...



Genetics contribute to the whole spectrum of disease risk



Hundreds of thousands of genetic risk regions have been identified through GWAS

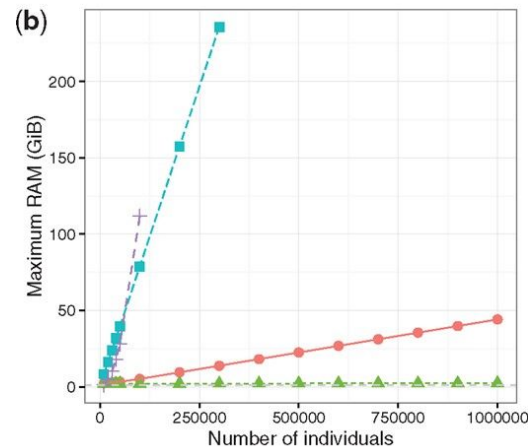
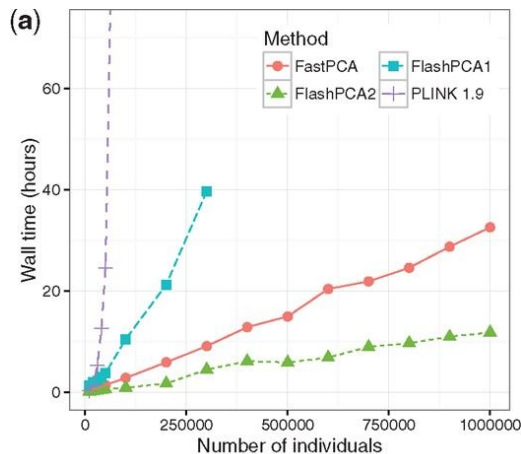


Hundreds of thousands of genetic risk regions have been identified through GWAS



PCA is extremely computationally intensive

100,000 individuals x
650,000 SNPs



**FlashPCA2: principal component analysis of
Biobank-scale genotype datasets** 

Gad Abraham ✉, Yixuan Qiu, Michael Inouye