Penn Genetics Summer Short Course in Computational Genomics

Educational Objectives

Data science

- Read, interpret, and critique the scientific literature
- Ingest, manipulate, summarize, and evaluate large datasets
- Develop reproducible workflows
- Utilize a high-performance compute cluster for data analysis

Genetics

- Understand key concepts in complex trait genetics
- Perform GWAS meta-analysis and downstream computational analyses
- Possess a framework for interpreting GWAS findings to biological insights

Deliverables

- GWAS meta-analysis
- Downstream analysis of summary statistics
  - Significant loci
  - Prioritized variants and genes
  - Biological meaning

Structure

- Weekly meetings (Wednesday 1-3 [except Juneteenth – Tuesday June 18])
  - **Didactic**
    - Presentation
      - Theory / background
      - Practical advice
    - Coding demonstration
  - Alternating with **Discussion Section**
    - Project updates
    - Troubleshooting
- Literature review
- Projects
- Mentors

Genetic alchemy: From base data to noble

Scott M. Damrauer, MD
Vice-Chair of Clinical Research, Department of Surgery
William Maul Measey Associate Professor of Surgery II
Associate Professor of Genetics

Penn Medicine

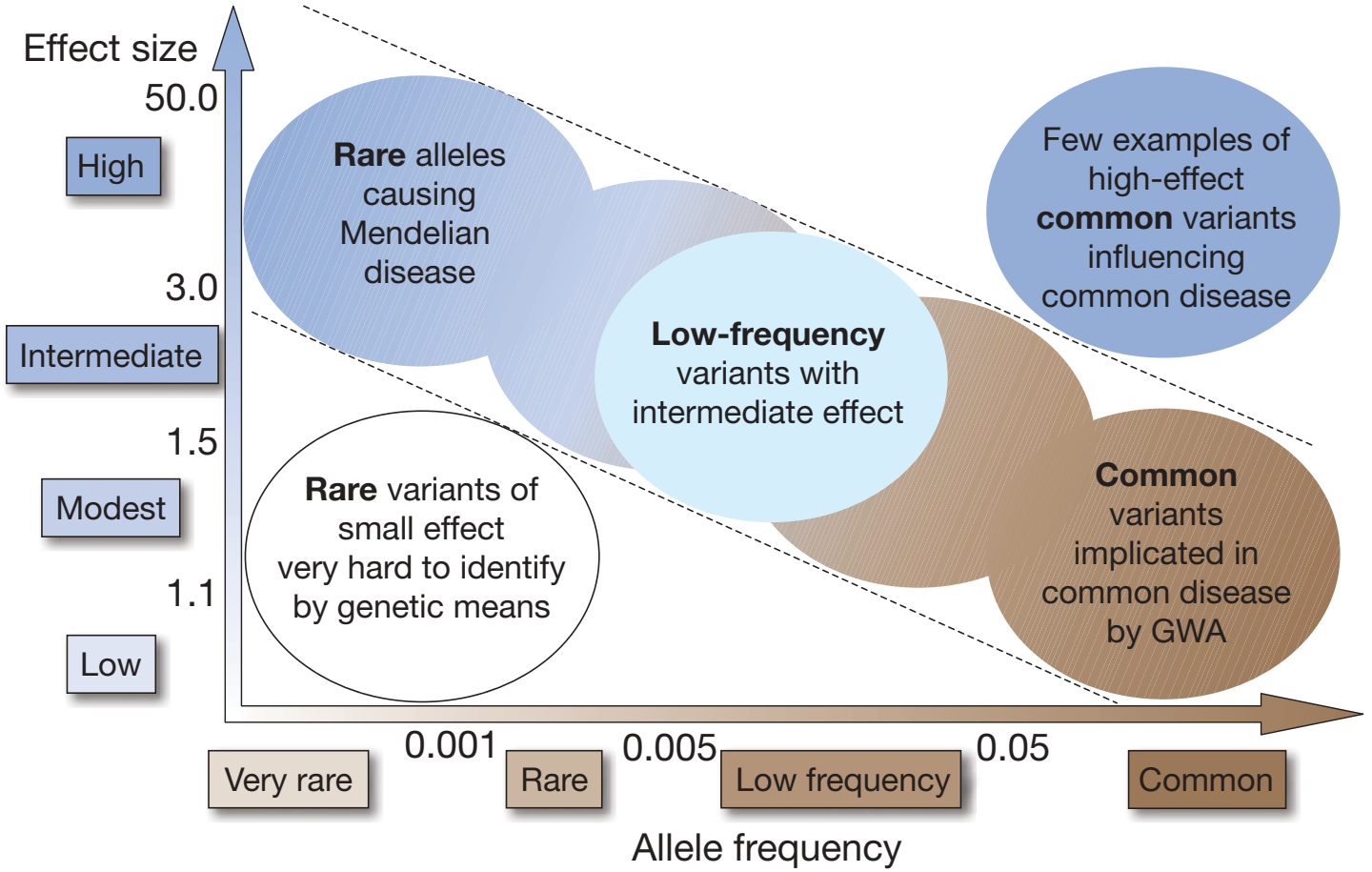
VA HEALTH CARE Defining EXCELLENCE in the 21st Century


GUINNESS WORLD RECORDS 2013
DISCOVER A WORLD OF NEW RECORDS
3D


BRIGHAM
45
YOUNG
7'6"
7 ft.
6'6"
6 ft.
5'6"
5 ft.
4'6"
4 ft
3'6"
3 ft.
2'6"
2 ft.
1'6"
1 ft.

David Teniers the Younger (1610–1690), The Alchemist, ca. 1643–45. Oil on panel, 20 1/8 x 28 in (51 x 71 cm). Herzog Anton Ulrich Museum, Braunschweig (139)

Common variant genetics for risk prediction



Manolio et al. *Nature* 2009

All humans are at least 99.5% genetically similar

- Most common of form genetic variation occurs when a single nucleotide differs between individuals
  - This is a **Single Nucleotide Polymorphism (SNP)**
  - On average ~ 1/300 base pairs
- Each individual has 2 copies of each gene / “alleles”

International HapMap Consortium, *Nature* 2003

Testing genetic association

Does the "A" allele Increase the risk for Disease relative to "G"?

Testing genetic association in populations

Quantifying associations with linear outcomes

T-test or ANOVA to compare means

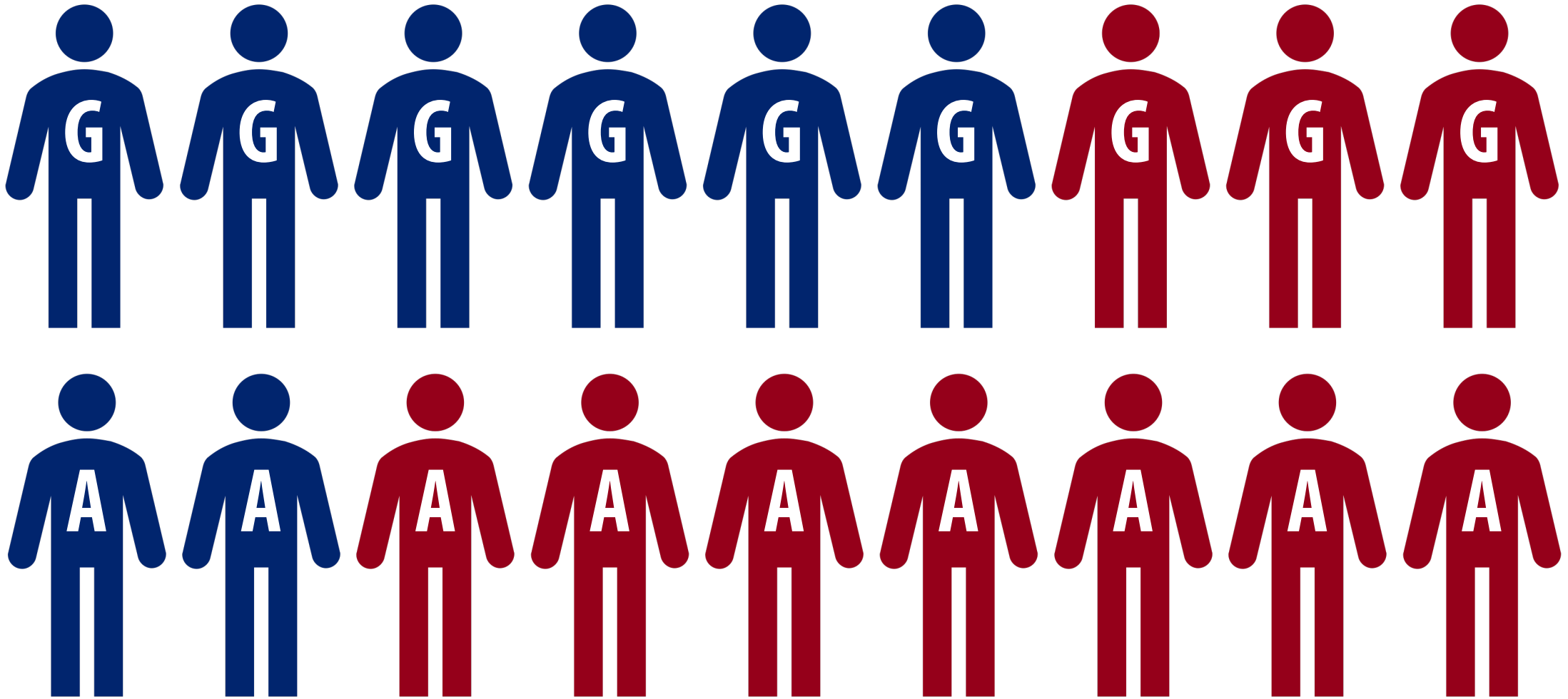
- Hypothesis testing

Linear regression to quantify association

- Hypothesis testing
- Modeling

- Multiple exposures (Multivariable)

Testing genetic association in populations



Quantifying associations with categorical outcomes

Chi Squared

- Unequal proportions from a r x c table
- Hypothesis testing
- Odds ratio

Can also use

- Fisher's Exact Test

Logistic regression

- Hypothesis testing
- Modeling

- Multivariable

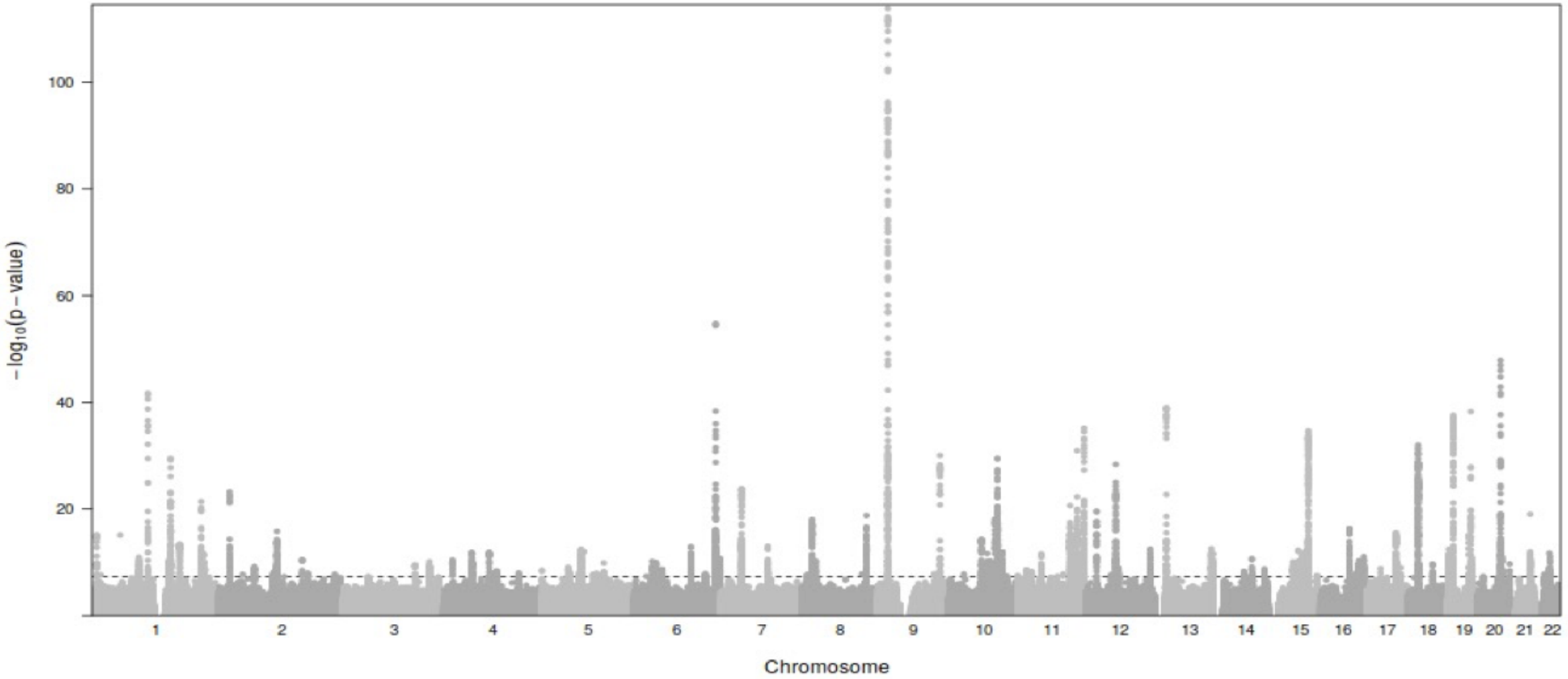
Genome Wide Association Study (GWAS)

$$logit(P|X,G) = \beta_0 + \boldsymbol{\beta_G \cdot G} + \beta_{age} \cdot age + \beta_{sex} \cdot sex + \sum_{k=1}^{10} \beta_{PC} \cdot PC_k$$

- Repeat the model for every genetic variant in the panel
  - ~300,000 to 400,000 for chip based GWAS
  - ~ 50 million times for imputed GWAS

- P-values adjusted for multiple testing
  - Genome-wide significance $P < 5x10^{-8}$

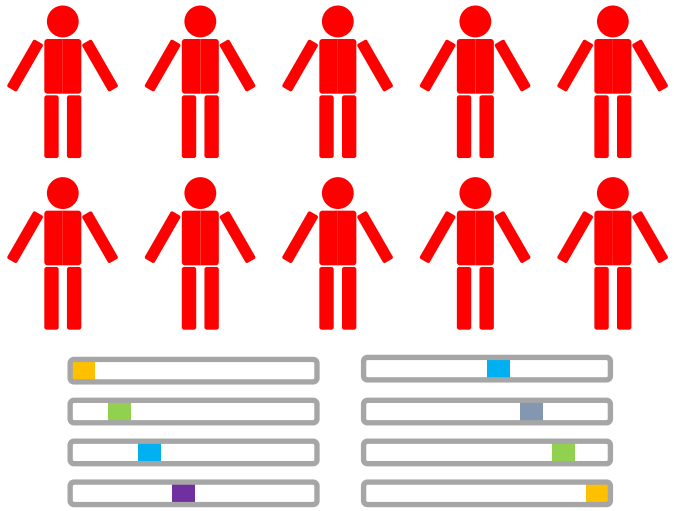
AAAgen meta-analysis

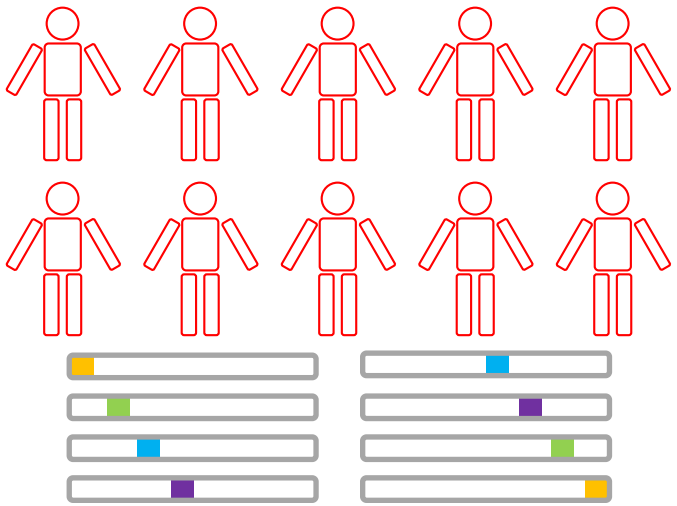


Roychowdhury et al. *Nature Genetics* 2023

ASCERTAINED case-control genetic association study

- Case control cohort based on disease status
- Perform genotyping on all participants
- Impute from directly genotyped data to more dense genetic data based on reference panels
- Perform association testing sequentially for each variant



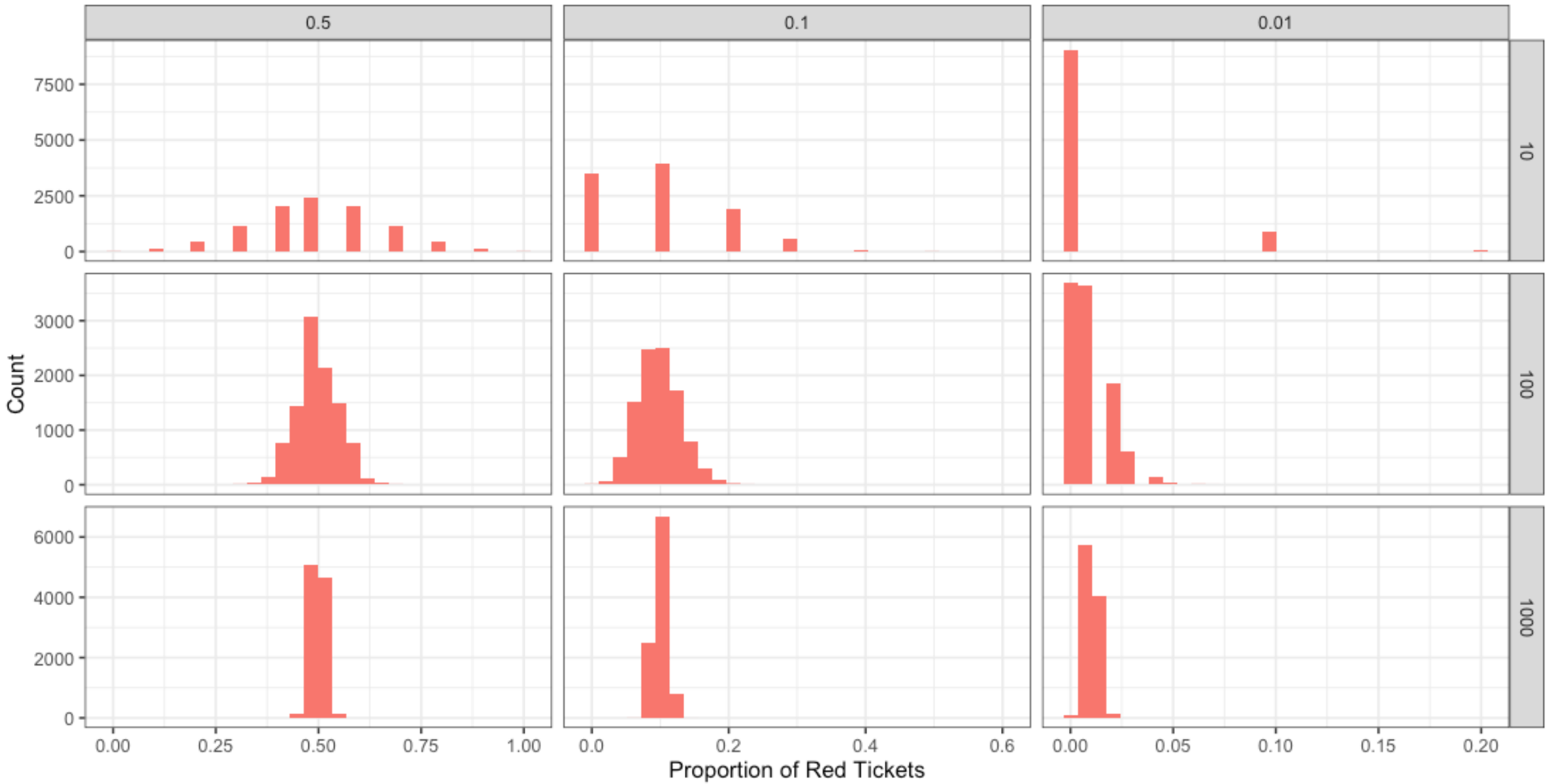


Sample size

- Traditionally thought of as
  - Acceptable risk of a false negative (power, beta, probability of a type 2 error)
  - Acceptable false positive rate (alpha, probability of a type 1 error)
  - Magnitude of effect
  - Variance (how precisely you can measure the effect)

- In the setting of an observational study, also depends on:
  - Rate of exposure in the population

Drawing lottery tickets


0.5
0.1
0.01
10
100
1000
Count
Proportion of Red Tickets
0
2500
5000
7500
1000
2000
3000
2000
4000
6000
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.00
0.05
0.10
0.15
0.20

What percentage of smokers have lung cancer?

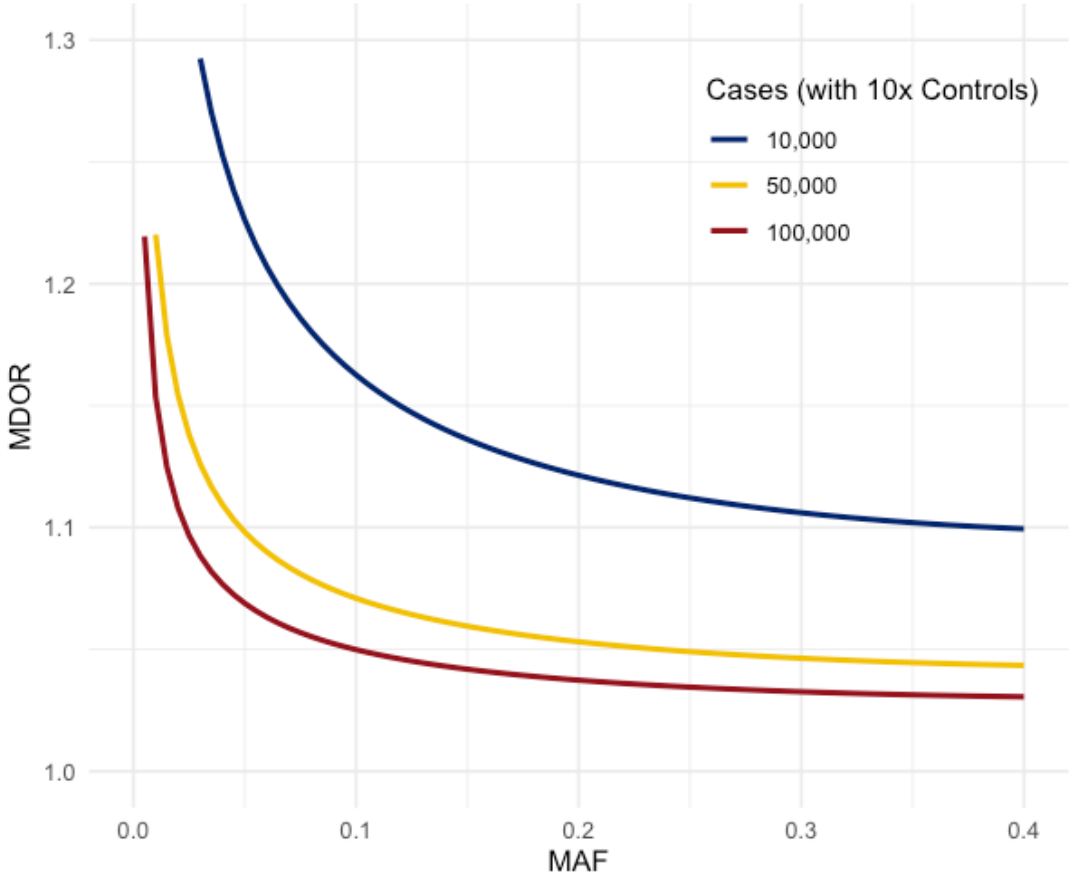
Does smoking increase the risk of lung cancer?

GWAS power

- Bigger sample size is always better
  - This refers to cases more than controls
  - After 10 controls per case, the value of more controls becomes less

- For the “power” calculation, we solve for the minimum detectable odds ratio (MDOR)
  - genpwr package in R

MDOR for binary trait

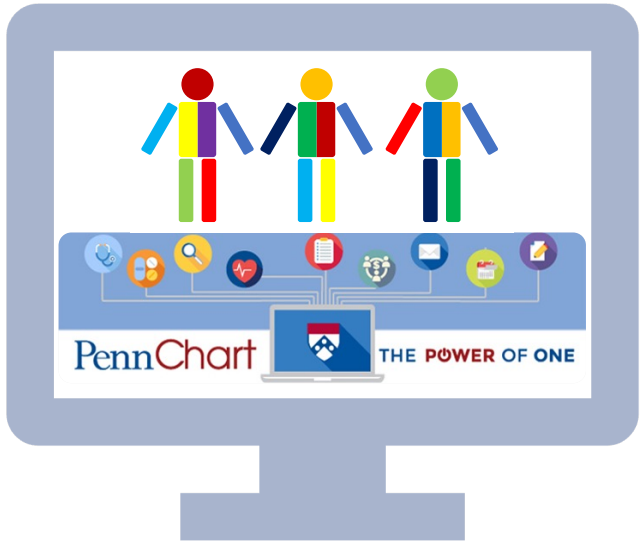


- 10 x controls per case
- 80% Power
- Alpha = GWS
- Additive genetic model

EHR based genetics

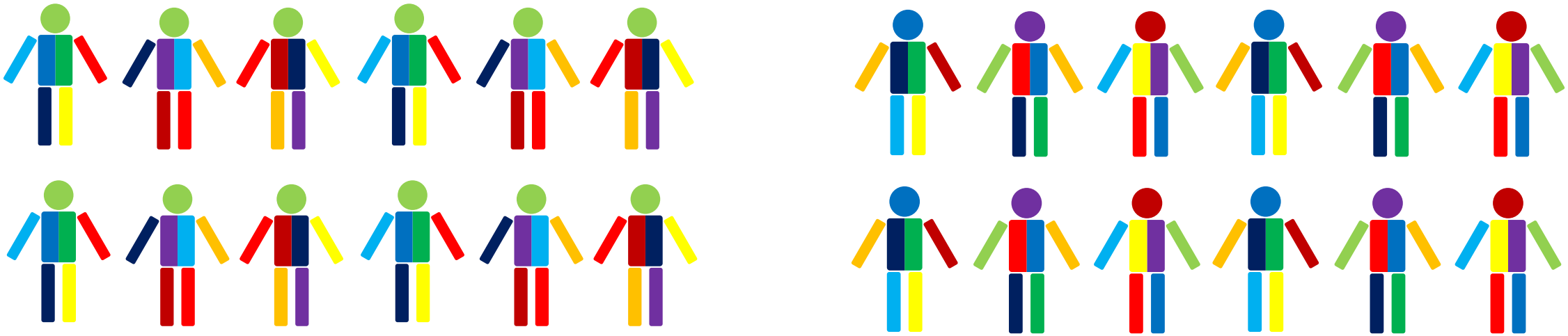
Genetic data is acquired ONCE per person

EHR captures the full range of biomedical "traits"



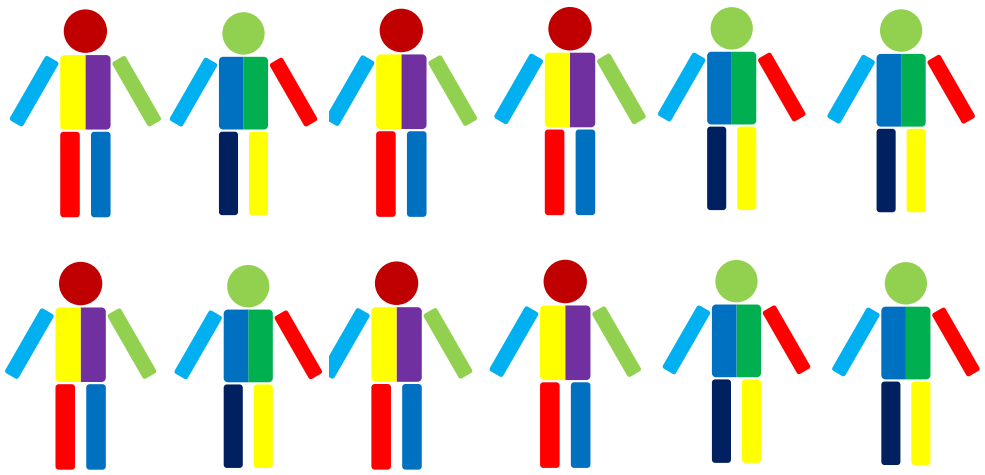
EHR based genetics

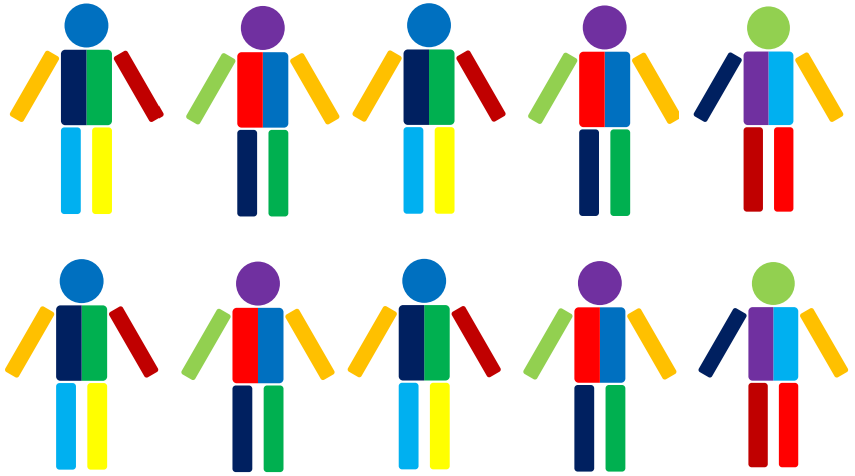
- "**Phenotyping**" performed on all aspects of data
- Structured: ICD-9/10 codes, CPT codes, Vital signs, labs
- Semi-structured: Text reports
- Unstructured: Office notes, Images



EHR based genetics

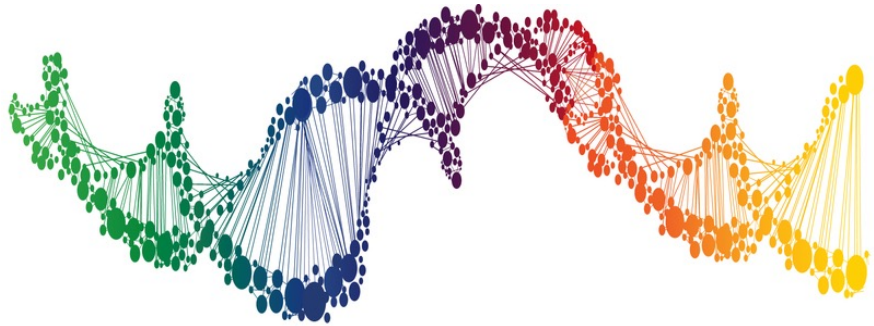
- "Phenotyping" performed on all aspects of data
- Structured: ICD-9/10 codes, CPT codes, Vital signs, labs
- Semi-structured: Text reports
- Unstructured: Office notes, Images





Phenome-wide association study

GWAS looks for associations with a specific phenotype across the entire genome



PheWAS looks for associations with a specific genotype across the entire phenome

The Biobank Continuum

VA Million Veteran Program

90% Male | 10% Female

80K + (or 8%) Hispanic

180K + (or 18%) Black

MVP has the largest cohort of people of African population of any research program in the world

25% minority

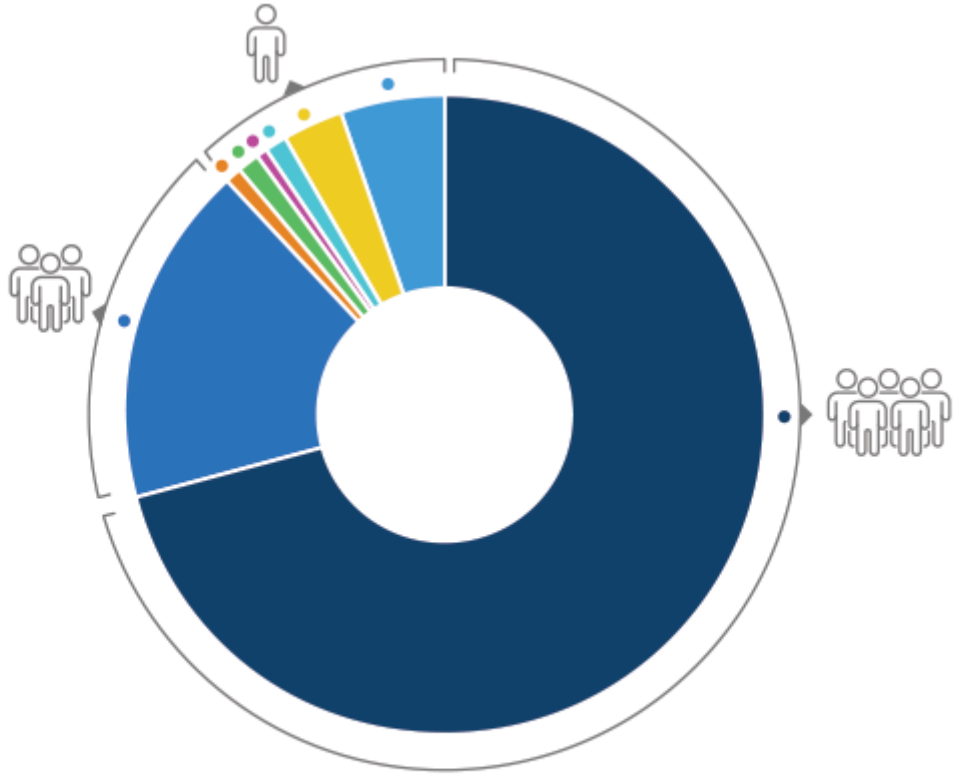
Racial and ethnic groups represented

Average age is 66

with largest cohort (33%) aged 70–79 (83% over 50)

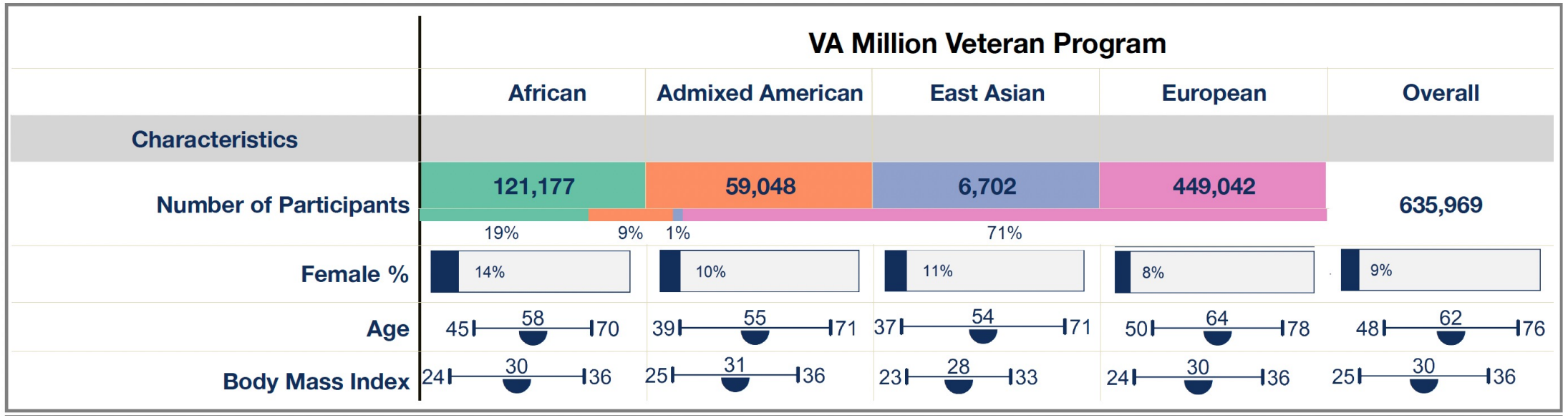
1,023,500+ enrolled

(as of March 13, 2024)



- White **70.99%**
- Black or African American **17.32%**
- American Indian or Alaskan Native **0.75%**
- Asian **1.07%**
- Native Hawaiian or Other Pacific Islander **0.47%**
- Other **1.35%**
- Multi-Racial **2.88%**
- Missing **5.18%**

VA MVP – Study Participants



Analysis pipeline

Genotypes

42 Million SNPs (MAC > 20)

Phenotypes

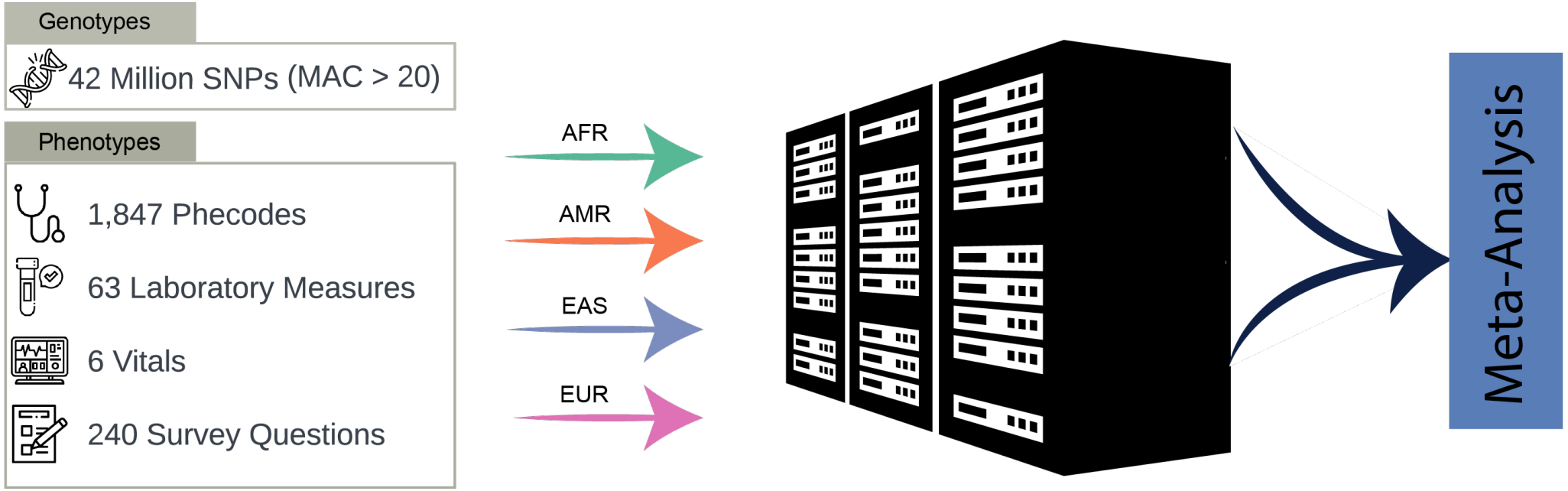
1,847 Phecodes

63 Laboratory Measures

6 Vitals

240 Survey Questions

Analysis pipeline



Estimated wall time of **8 years** on existing compute infrastructure

Analysis pipeline

Genotypes

42 Million SNPs (MAC > 20)

Phenotypes

- 1,847 Phecodes
- 63 Laboratory Measures
- 6 Vitals
- 240 Survey Questions

AFR

AMR

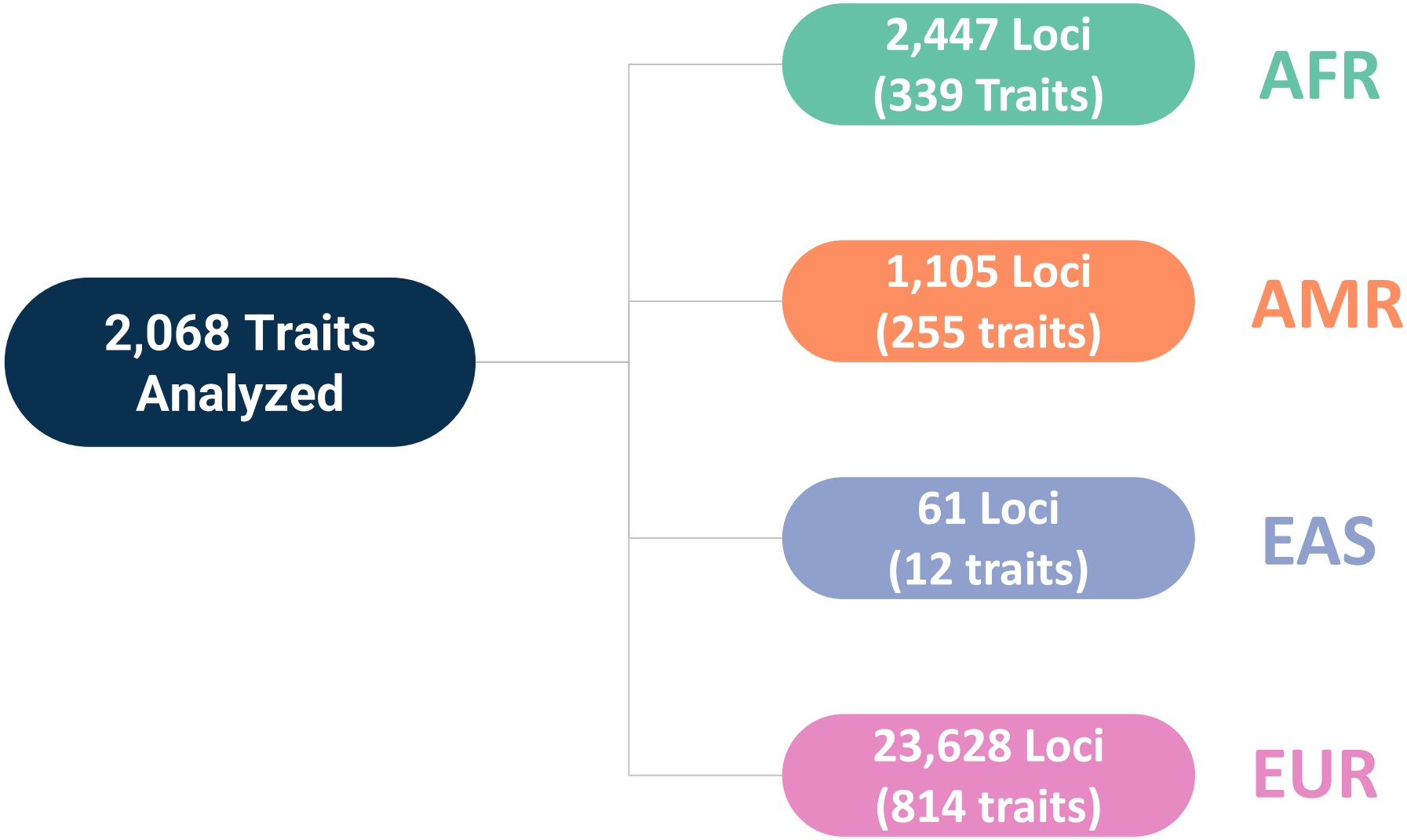
EAS

EUR

Meta-Analysis

Actual wall time of **14 days** on Summit

Genome-wide loci within each population group



Study wide p-value threshold 4.6×10^{-11}

Diversity and Scale: Genetic Architecture of 2,068 Traits in the VA Million Veteran Program

Anurag Verma†[1,2,3], Jennifer E Huffman†[4,5,6], Alex Rodriguez†[7], Mitchell Conery†[8], Molei Liu†[9], Yuk-Lam Ho[4], Youngdae Kim[10], David A Heise[11], Lindsay Guare[2], Vidul Ayakulangara Panickan[12], Helene Garcon[4], Franciel Linares[13], Lauren Costa[14], Ian Goethert[15], Ryan Tipton[16], Jacqueline Honerlaw[4], Laura Davies[17], Stacey Whitbourne[6,14,18], Jeremy Cohen[11], Daniel C Posner[4], Rahul Sangar[14], Michael Murray[14], Xuan Wang[12], Daniel R Dochtermann[19], Poornima Devineni[19], Yunling Shi[19], Tarak Nath Nandi[7], Themistocles L Assimes[20], Charles A Brunette[21,6], Robert J Carroll[22], Royce Clifford[23,24], Scott Duvall[25,26], Joel Gelernter[27,28], Adriana Hung[29], Sudha K Iyengar[30], Jacob Joseph[31,32], Rachel Kember[33,34], Henry Kranzler[33,34], Daniel Levey[35,27], Shiuh-Wen Luoh[36,37], Victoria C Merritt[23], Cassie Overstreet[27], Joseph D Deak[38,39], Struan F A Grant[40,41,42,43], Renato Polimanti[38], Panos Roussos[44], Yan V Sun[45], Sanan Venkatesh[44], Georgios Voloudakis[44], Amy Justice[35,46,47], Edmon Begoli[48], Rachel Ramoni[49], Georgia Tourassi[50], Saiju Pyarajan[19], Philip S Tsao[20,51], Christopher J O'Donnell[52], Sumitra Muralidhar[49], Jennifer Moser[49], Juan P Casas[4], Alexander G Bick[53], Wei Zhou[54,55,56], Tianxi Cai‡[12], Benjamin F Voight‡[1,8,43,57], Kelly Cho‡[6,14,18], Michael J Gaziano‡[6,14,18], Ravi K Madduri‡[7], Scott M Damrauer‡[1,43,58,59], Katherine P Liao‡[60,61]

Other sources of GWAS summary statistics

- Pan UK Biobank project -- https://pan.ukbb.broadinstitute.org
- FinnGen -- https://r10.finngen.fi
- All of US
- Biobank Japan

- Published, phenotype specific studies
  - Open GWAS project -- https://gwas.mrcieu.ac.uk
  - GWAS catalog -- https://www.ebi.ac.uk/gwas/

Penn Genetics Summer Short Course in Computational Genomics