SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

Session 5:
Assessing Genetic Variation, Imputation, Principal Component Analysis

Podcast on the Human Genome Project
Interview with Dr. Eric Green, NHGRI Director

- https://geneticsunzipped.com/blog/2020/10/22/s322-the-past-present-and-future-of-the-human-genome-project

Assessing Genetic Variation: Genotyping vs. Sequencing

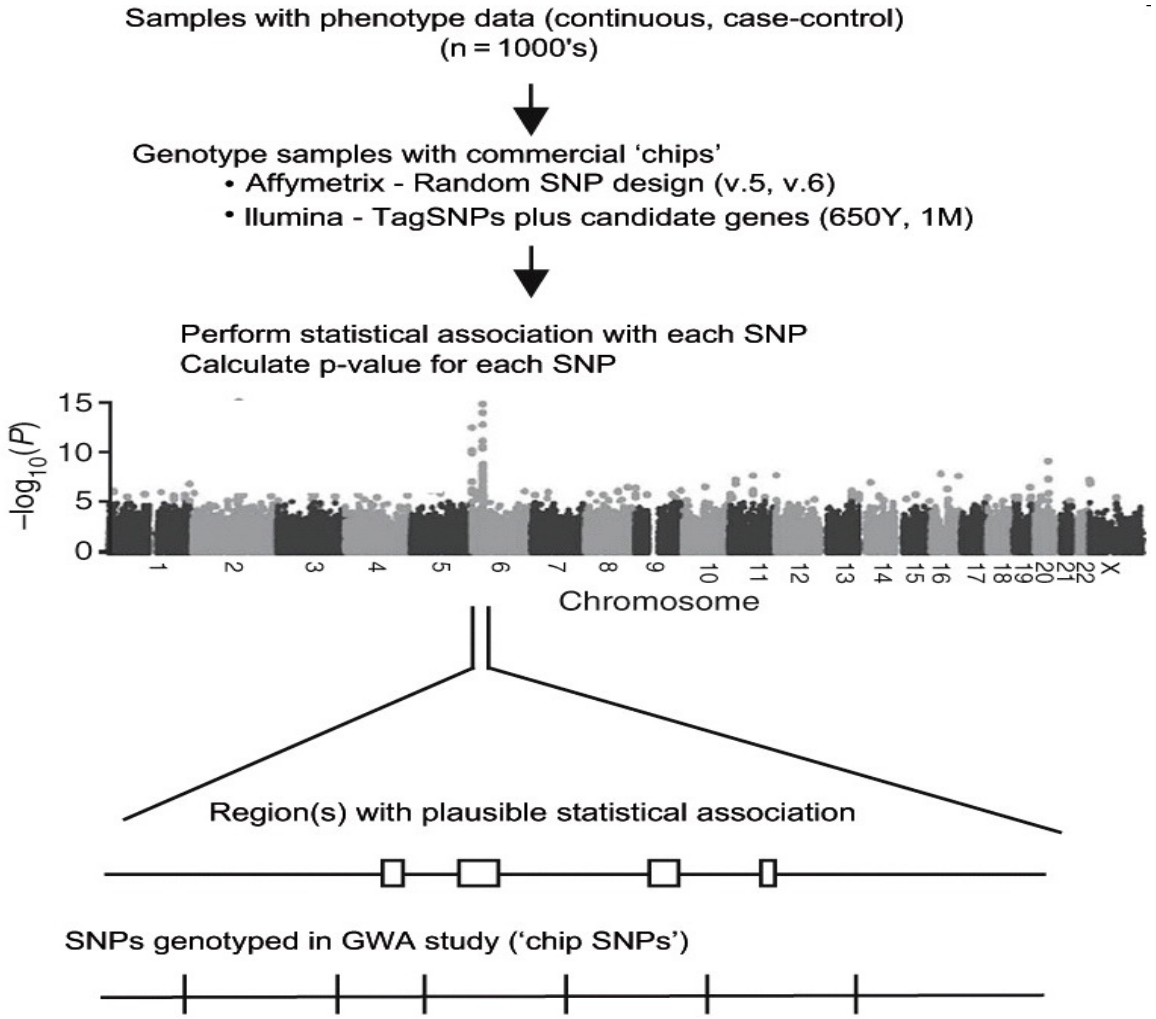
Genotyping: Target a particular genetic variant and ”measure” it

Sequencing: Target a region (could be the whole genome) and “measure” the entire region (all base-pairs)

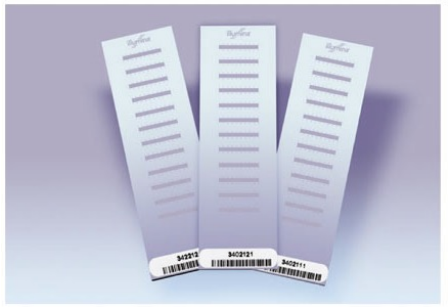
From a bioinformatic/analysis point of view, genotyping data is much easier to handle.

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Genome-wide association studies (GWAS)



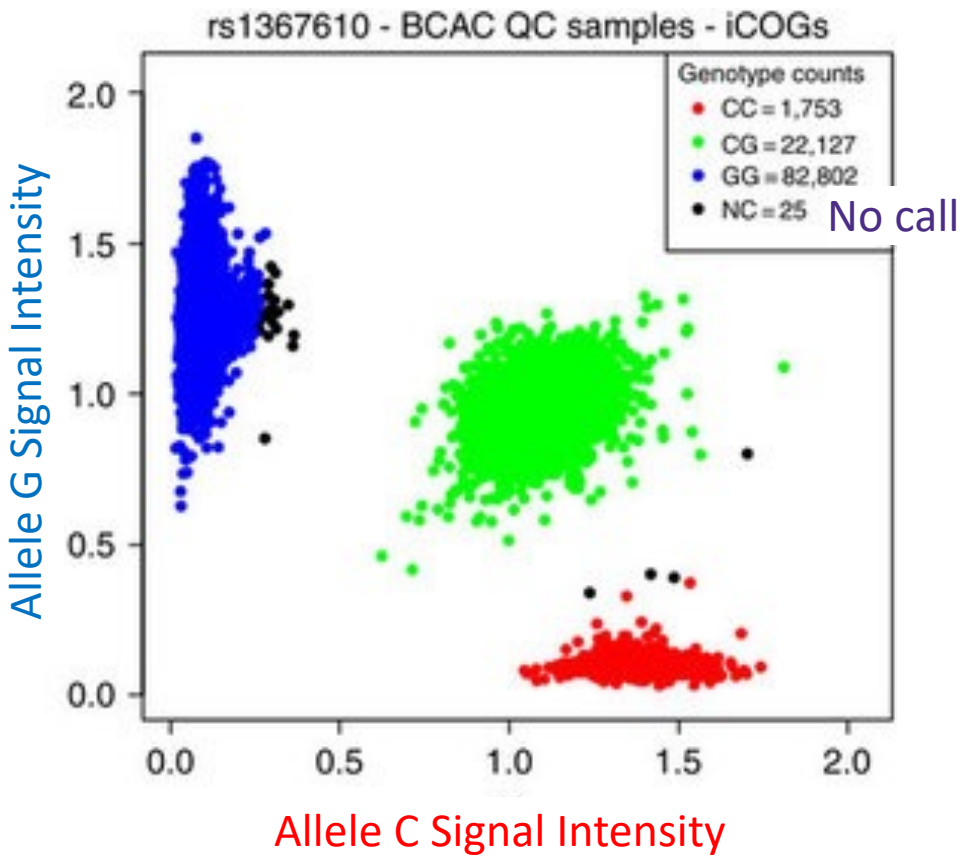
Screen across the genome for SNPs that are associated with trait (agnostic approach)

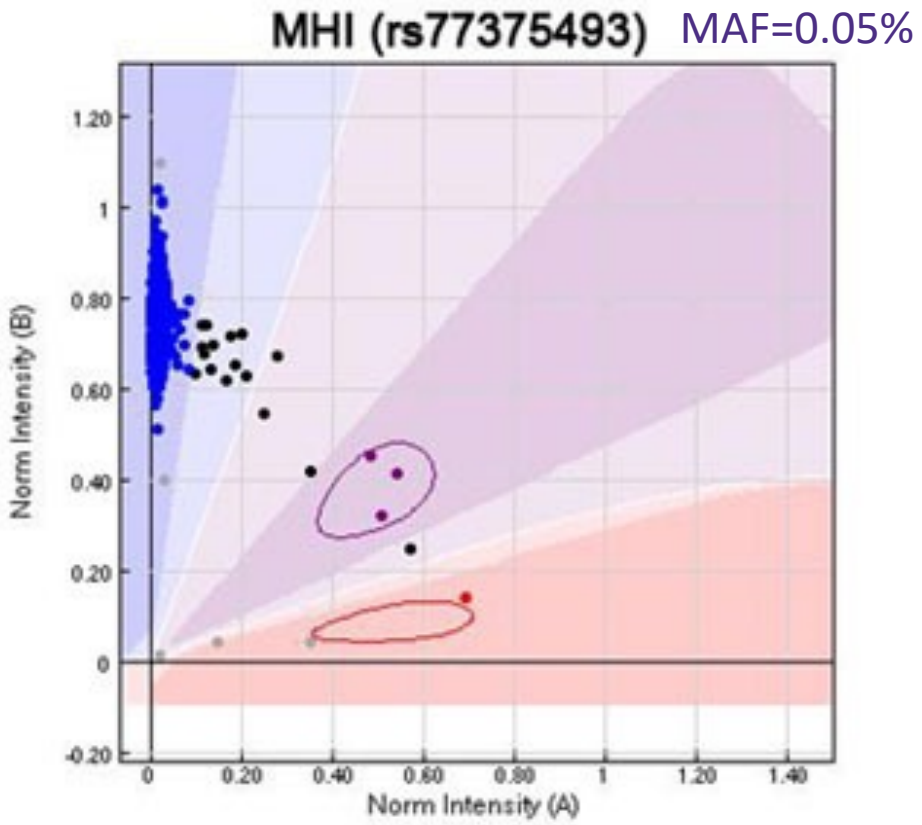


Rieder et al, Drug Metab Rev 2008

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Genotyping Output





Li, Nat Comm 2014

Auer, Nat Genet 2014

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Pricing (CIDR, March 2023)

Illumina Genotyping – GWAS, PGx, PRS					
Global		**Screening**		**Consortium-Developed**	
Global Diversity Array	$110-$130	Global or Asian	$75-$100	Oncoarray	$85-$110
Pharmacogenetics		**Other Consortium Developed Arrays**			
Global Diversity Array + PGx	$175	Exome Beadchip, DrugDev, H3AfricaArray, ImmunoArray, PsychArray, QC		Inquire for pricing	
***PLUS OPTIONS: Custom content can be added to most GWAS and Consortium arrays. Please Inquire for pricing.**					

Affymetrix Genotyping - GWAS and Custom	
UK Biobank 821K Axiom Array	~$150 - $210
Custom Array (up to 750K SNPs)	~$180 - $240
Custom Array (up to 50K SNPs)	~$120 - $170

https://cidr.jhmi.edu/xtras/shared/documents/pricing.pdf

Imputation (I)

- Cost efficient: Can assess more SNPs than genotyped
  - Input: 550,000 SNPs in 10,000 individuals
  - Reference panel: 2,504 individuals from the 1,000 Genomes project (>80M markers excluding singletons)
  - Output: Imputed data for >80M markers for your 10,000 individuals
- Maximizes sample size
  - Fills in missing values for already genotyped SNPs
- Allows us to combine data from existing platforms that genotype different SNPs

	HumanHap	Affy 6.0	OmniExpress
HumanHap	459,999	126,959	260,661
Affy 6.0		668,283	168,223
OmniExpress			565,810

Lindström, PLoS One 2017

* 75,285 markers are on all 3 platforms

Most imputation methods work under the framework that individual haplotypes are all unique but expected to share contiguous, mosaic stretches with other haplotypes in the sample.

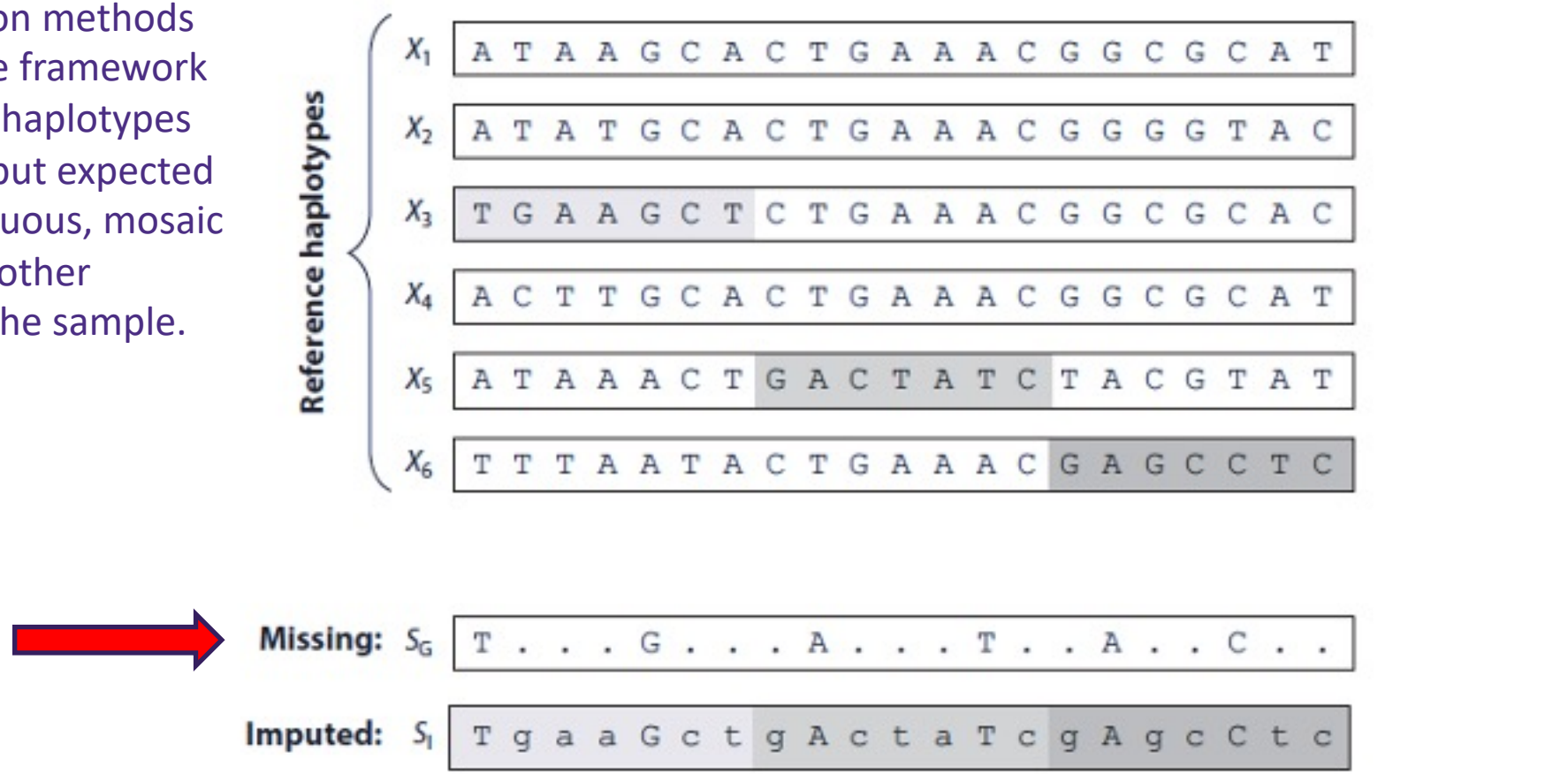


Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype (S_G) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in S_G are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype X_3. In the second segment, the alleles A and T imply that haplotype X_5 might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes X_3, X_5, and X_6. Consequently, the missing sites can be imputed to obtain the final imputed haplotype, S_I.

Das, Ann Rev of Genomics and Hum Genet 2018

Imputation (II)

> Many imputation algorithms employ a Hidden Markov Model (HMM) method

> Software: MACH/minimac, IMPUTE, Beagle

> Outputs:

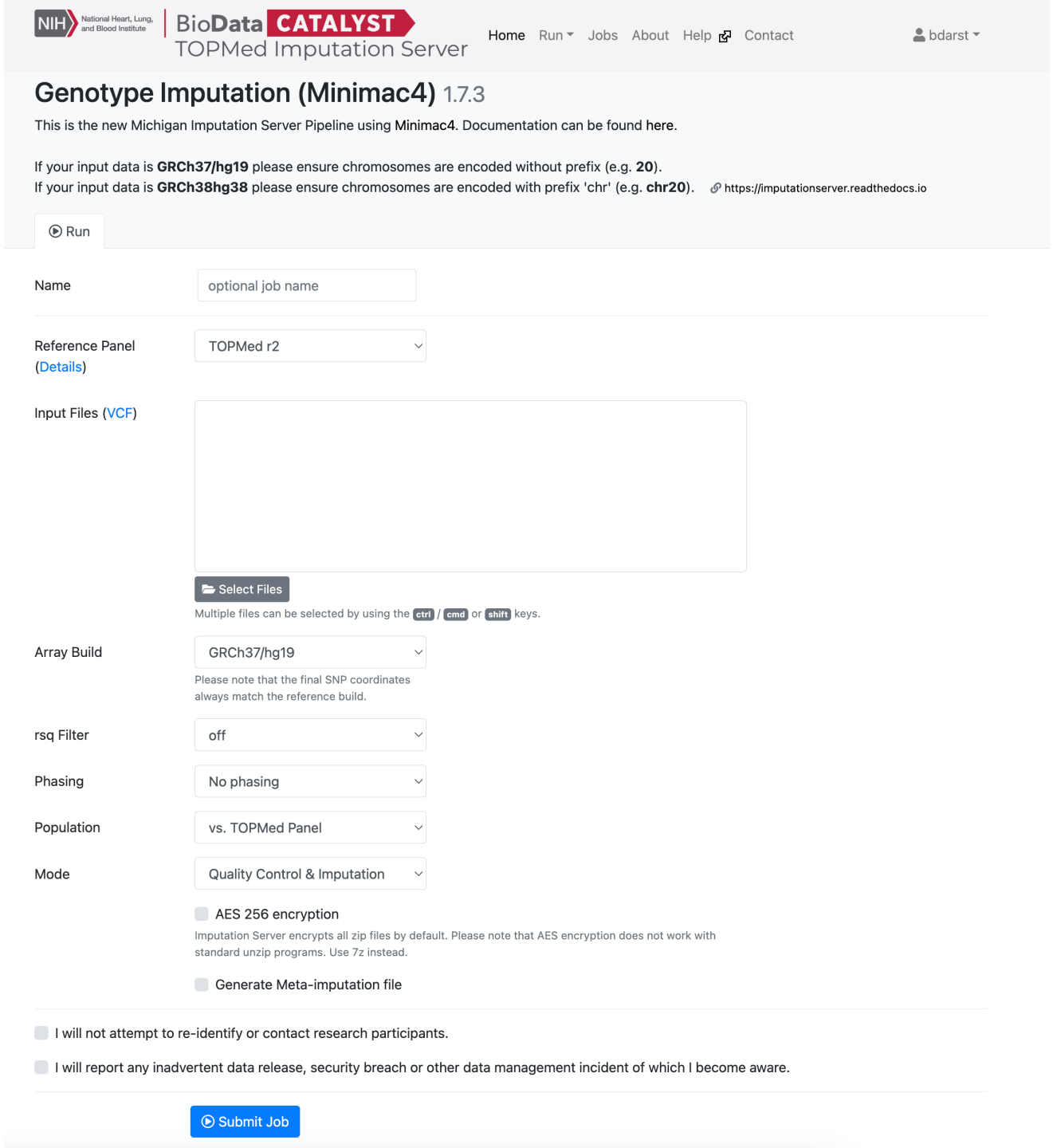
- Posterior probabilities for each potential genotype with three data points per SNP/individual
- "Dosage" of each imputed genotype ranging between 0-2, representing copies of the reference allele (continuous number)

For more information about imputation, read Marchini and Howie, Nat Rev Genet 2010

Imputation (III)

> The imputation quality score r^2 measures how well a SNP was imputed.
  - Ranges between 0 and 1.
  - Typically, a cut-off of 0.30 or so will flag most of the poorly imputed SNPs, but only a small number (<1%) of well imputed SNPs.

> Factors that affect imputation quality:
  - Number of genotyped SNPs in your data
  - **Size of reference panel**
  - Similarity in genetic ancestry between reference and study samples
  - Allele frequency



TOPMed Imputation Server

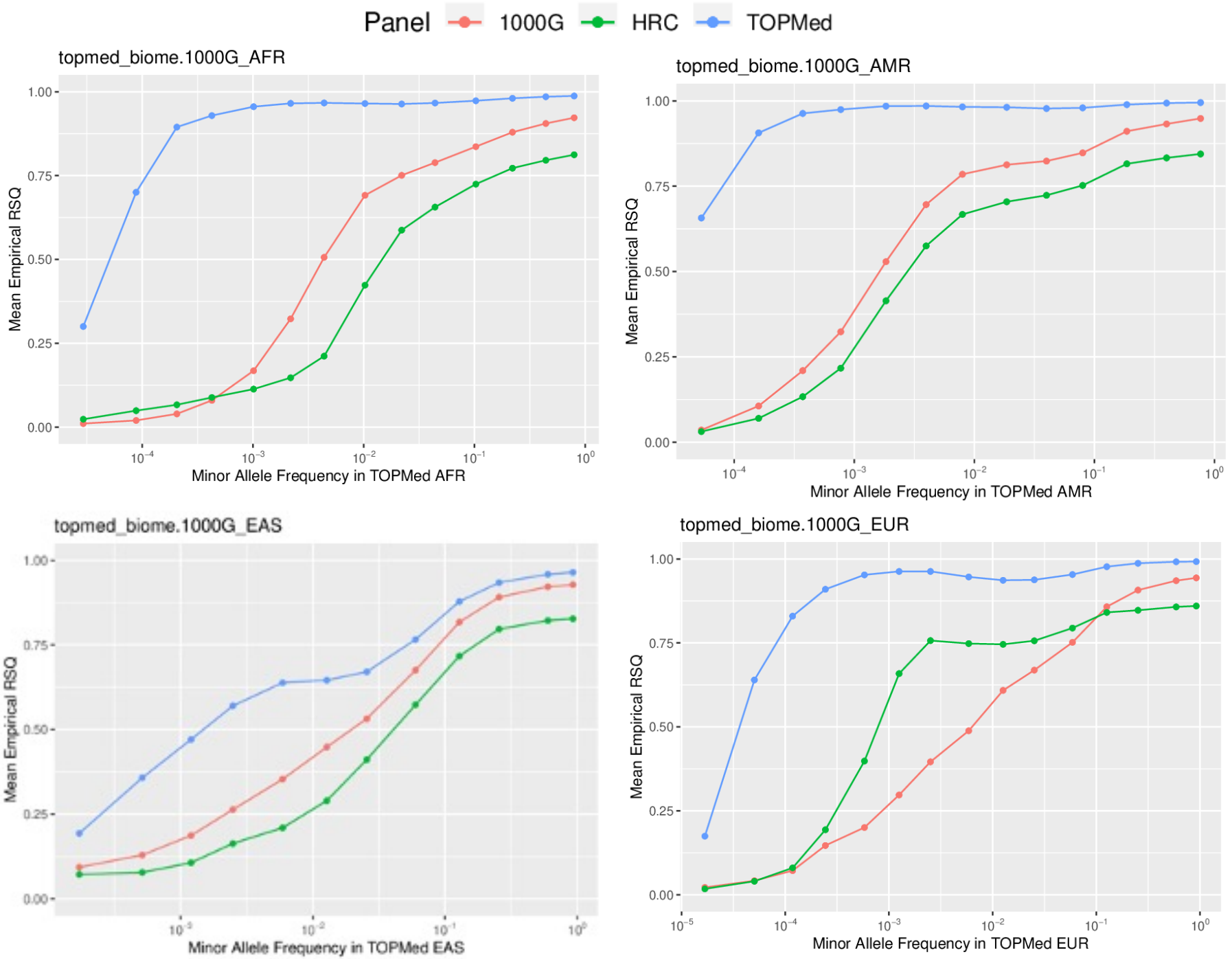
Version r2 includes **97,256** reference samples and **308,107,085** SNVs and indels distributed across the 22 autosomes and the X chromosome.

Population	N
African	24,267
Admixed American	17,085
European	47,159
East Asian	1,184
South Asian	644
Not assigned	6,917
Total	**97,256**

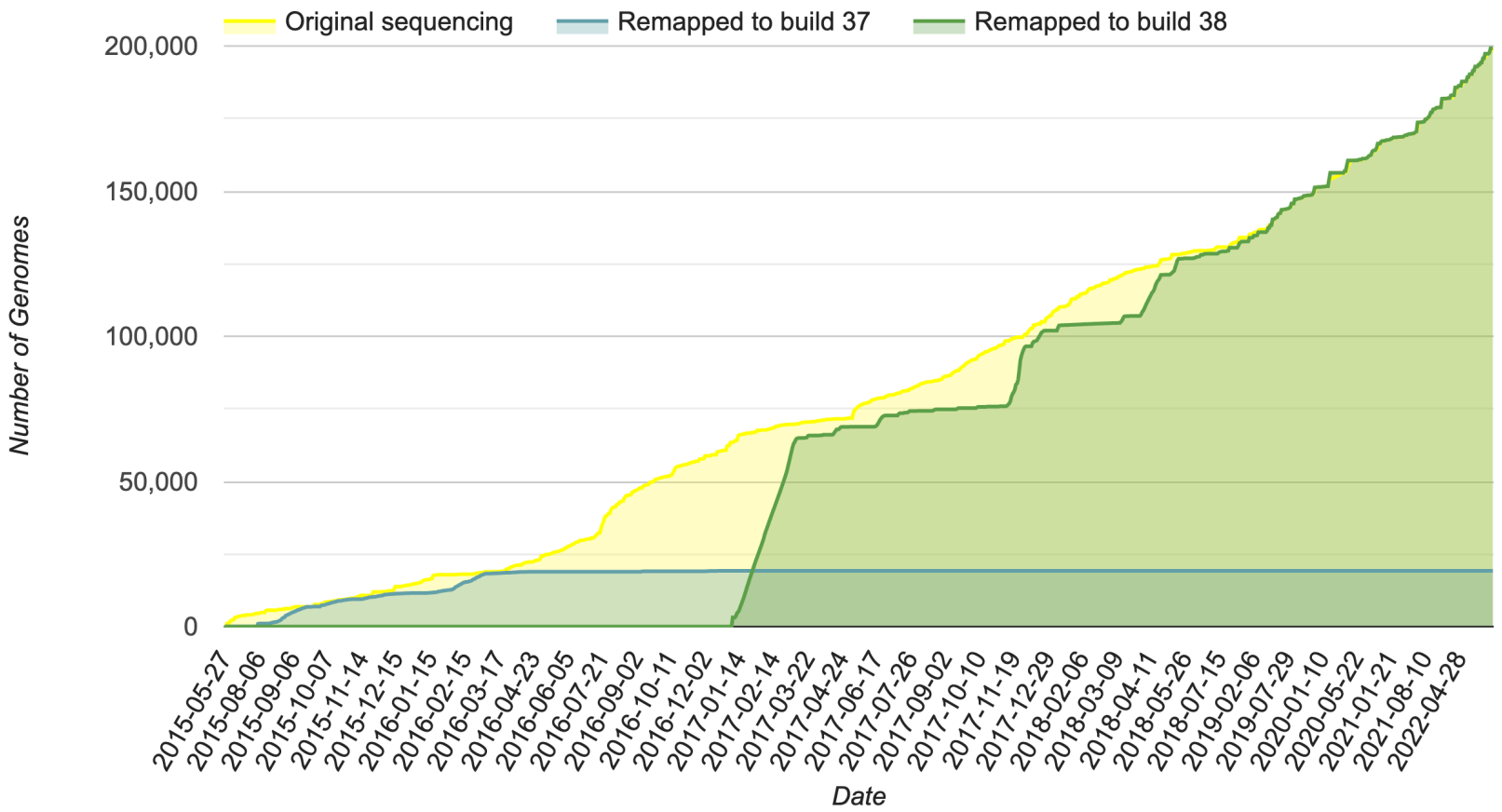
Taliun et al., Nature 2021

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Reference Panels	N	Ancestry
HapMap	60	European
1000 Genomes Phase 1	1,092	Mixed
1000 Genomes Phase 3	**2,504**	**Mixed**
CAAPA	883	African
TopMed	**97,256**	**Mixed**
GAsP	1,654	Asian
ChinaMap	10,155	Asian
HRC	**32,470**	**European**
AFAM	2,269	African



Imputation Reference Panels: TOPMed whole-genome sequencing



EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Importance of diversity in imputation reference panels

Imputation of a rare African ancestry-specific *HOXB13* variant

Received: 14 October 2019 | Accepted: 24 January 2020
DOI: 10.1002/pros.23960

ORIGINAL ARTICLE

The Prostate WILEY

Mutation *HOXB13* c.853delT in Martinican prostate cancer patients

Régine Marlin PhD[1] | Morgane Créoff MD[2] | Sylvie Merle MD[3] | Magalie Jean-Marie-Flore[1] | Mickaelle Rose[2] | Sarah Malsa[2] | Xavier Promeyrat MD[2] | François Martin MD[2] | Georges Comlan MD[2] | Nicolas Rabia MD[2] | Taoufiq Taouil MD[2] | Irfane Issoufaly MD[2] | Patrick Escarmant MD[2] | Vincent Vinh-Hung MD[2] | Odile Béra MD[1,2]

- Exome sequencing of *HOXB13* in 46 early-onset PCa cases
- rs77179853 (X285K) carried by 3 cases
  →stop loss deletion, RAF=0.2% in 1KGP, AFR only

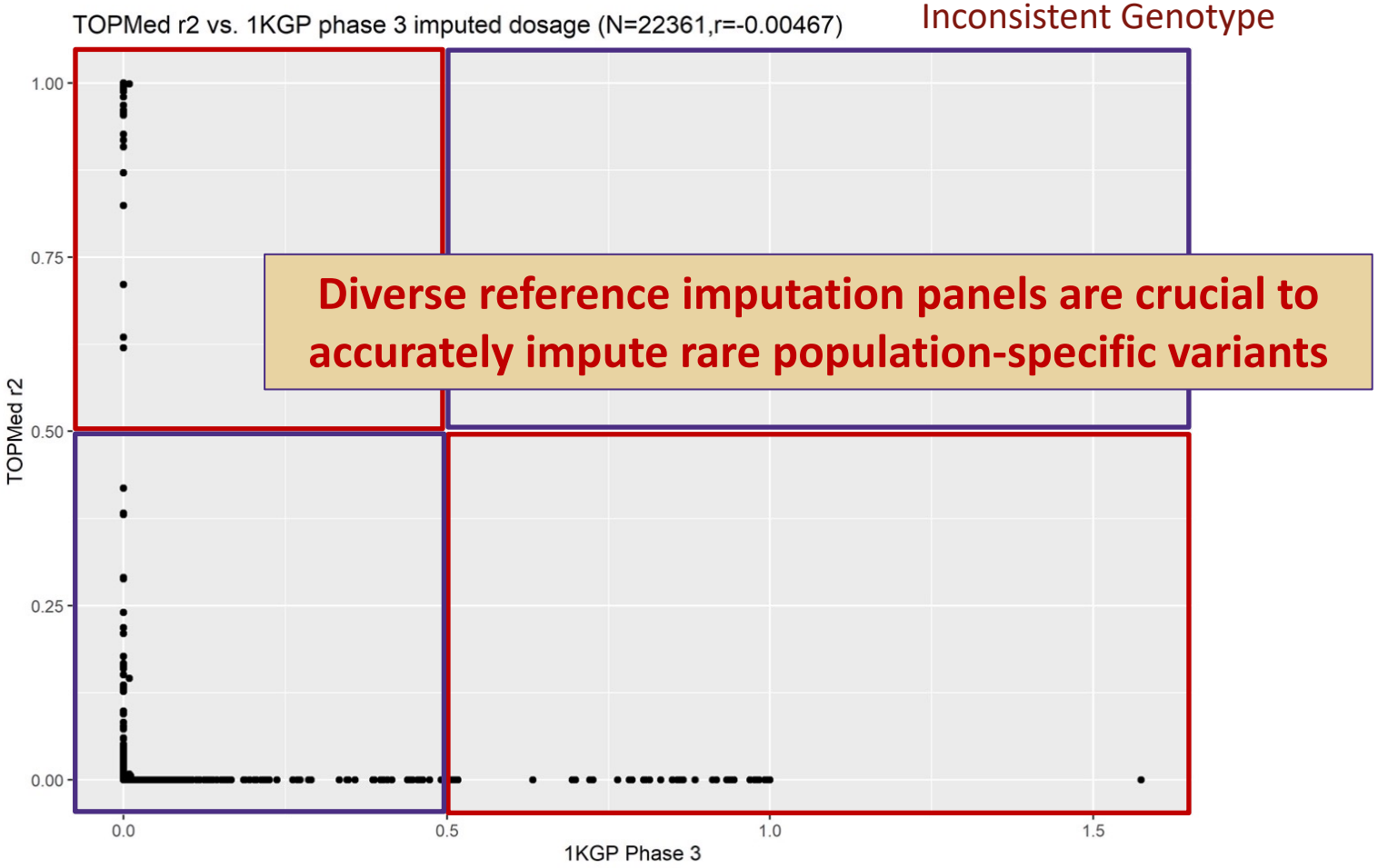
Imputation of rs77179853 into large-scale African ancestry GWAS data

			1000 Genomes Project Phase 3 (**3 carriers/2,504 participants**)			TOPMed Freeze 8 (**126 carriers/97,256 participants**)		
Genotyping Array	# Controls	# Cases	Info	Control Freq	Case Freq	Info	Control Freq	Case Freq
AAPC1M	4,642	4,822	0.819	0.24%	0.15%	0.921	0.13%	0.17%
ONCO-AAPC	3,953	4,231	0.748	0.20%	0.21%	0.918	0.11%	0.34%
H3 (California/Uganda Study)	1,048	1,590	0.684	0.15%	0.13%	0.949	0.15%	0.23%
HumanOmni (NCI Ghana Prostate Study)	634	640	0.753	0.16%	0.10%	0.967	0.49%	1.15%
MADCaP	396	405	0.819	0.25%	0.19%	0.941	0.12%	0.88%
	10,673	11,688						

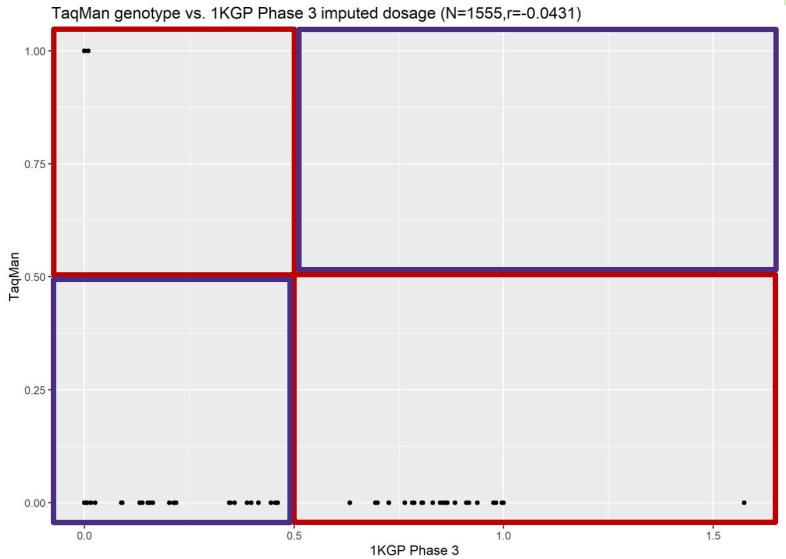
Darst et al., *Eur Urol* (2022)

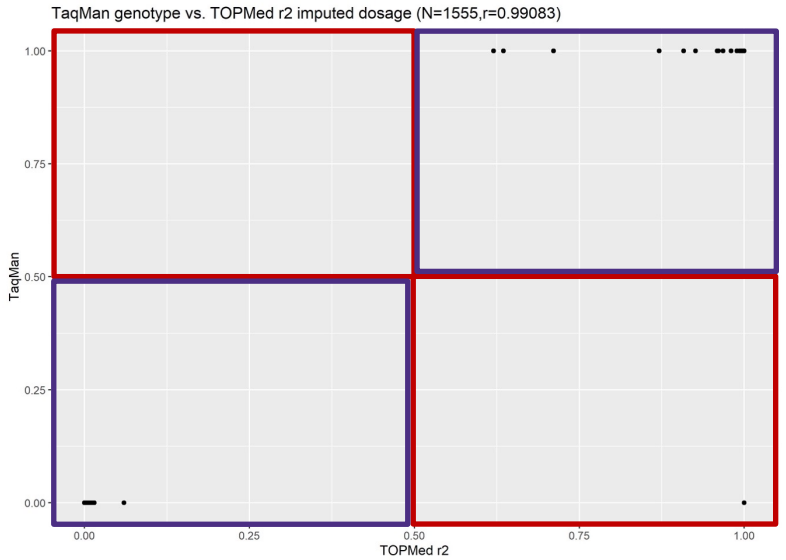
Importance of diversity in imputation reference panels

Imputation of a rare African ancestry-specific *HOXB13* variant



Diverse reference imputation panels are crucial to accurately impute rare population-specific variants





Subsequently found: Prostate cancer OR=2.42 (95% CI=1.52-3.87), P=$2x10^{-4}$
Metastatic prostate cancer OR=5.08 (95% CI= 1.88-13.7), P=0.001

Darst et al., *Eur Urol* (2022)

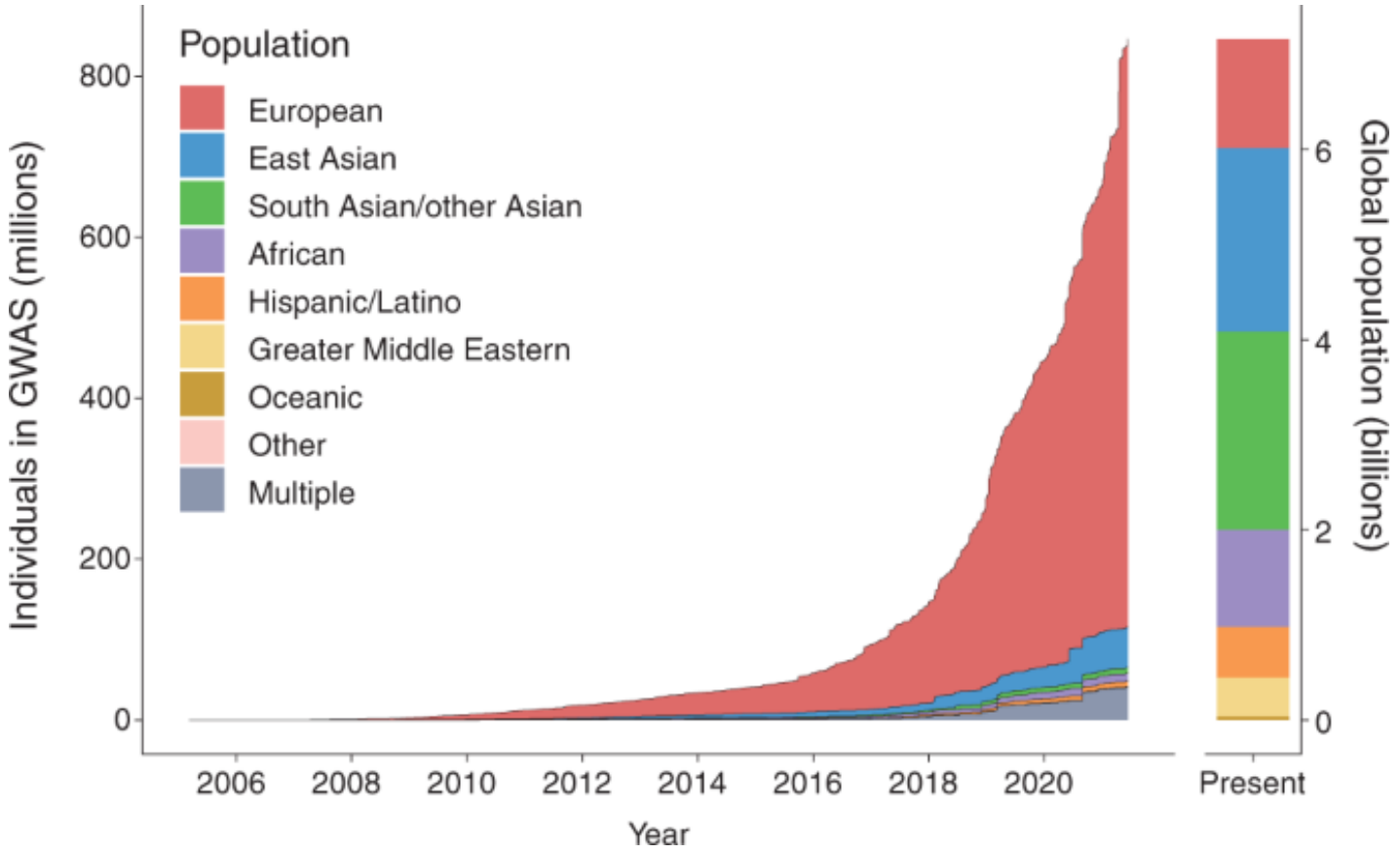
Breakout Room Discussion:

- Explore the breakdown of genetic ancestry in GWAS as reported on the website https://gwasdiversitymonitor.com.
  - What populations seem over- and under-represented in genetic studies?
  - What consequences can this have?

- What are your ideas for how we can we increase the diversity of study participants in genetic epidemiology?

Genomics is failing on diversity

An analysis by **Alice B. Popejoy** and **Stephanie M. Fullerton** indicates that some populations are still being left behind on the road to precision medicine.



Popejoy and Fullerton, Nature 2016

Martin, Nature Genetics 2019

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

The Multi-Ethnic Genotyping Array (MEGA) – 1.8M markers

Abbreviated reference	Approximate SNP allocation	Content description	Parameters informing content
		Backbone content	
Infinium HumanCore BeadChip	250,000	Included for backwards compatibility	Highly informative GWAS tag SNPs for EUR or ASN ancestries
African Diaspora Consortium Power Chip	700,000	Augmented GWAS coverage for African ancestries	692 individuals sequenced by CAAPA, highly informative for variants with MAF>2%
Improved cross-population tagging content	300,000	Augmented GWAS coverage for diverse ancestries	New tagging strategy developed by PAGE using 1KGP Phase 3 sequencing, highly informative for variants with MAF<2%
Multiethnic exonic content	400,000	Exome markers for diverse populations	Derived from WGS/WES data from > 36,000 individuals in diverse ethnic groups, emphasizes loss of function and splice variants
NHGRI GWAS catalog	11,631	Markers (tag SNPs) from published GWAS	Includes tag SNPs not reaching genome-wide significance ($p<5x10^{-8}$), and SNPs in high LD
SNPs in publications	5,874	SNPs listed in UCSC browser track	Mentioned by rsid number in $\geq$ 4 publications
Clinical and pharmacogenetic	17,000	All clinically relevant SNPs	Domain expert opinion and those annotated as deleterious
		PAGE Hand Curated Custom Content	
Validated regulatory SNPs	2,500	Regulatory variants with *in vitro* differential function in the literature	Differential EMSA, most with differential luciferase or equivalent
Enhanced GWAS	20,000	Improved tag SNP coverage for candidate genes/regions	Minimum r^2 of 0.8 rather than mean r^2 of 0.6 used for backbone
Enhanced Exome	16,000	Improved exonic coverage for candidate genes/regions	All available exonic variants
Fine-mapping	16,000	Fine-mapping coverage for GWAS catalog reports	All SNPs tagged at $r^2 > 0.6$ in reference population from primary GWAS report
OMIM/Clinvar [a]	Overlaps backbone content	Clinically relevant SNPs related to traits of interest	E.g. hyperlipidemia (*LPL*, *LDLR*, etc.), BMI (*MC4R*, etc.)

Bien, PLOS ONE 2016

Sequencing

- Capture ALL base-pairs in our region of interest
  - Whole genome sequencing, whole exome sequencing, targeted sequencing (e.g., follow up a GWAS signal)

- More expensive and requires more bioinformatics support than genotyping

- Exome and targeted sequencing have important limitations – they require an initial capture step to target the region(s) of interest.
  - Exome sequencing is often easier than targeted sequencing as it is not as ad hoc (i.e., GWAS region), and the exome has less repetitive regions than the genome as a whole

The Human Genome Project (1990-2003) set out to sequence every base pair in the human DNA

$2.7 billion

Led by Craig Venter (Celera Genomics)

Led by Francis Collins (NIH)

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

SCIENCE
VOLUME 376|ISSUE 6588|1 APR 2022

COVER

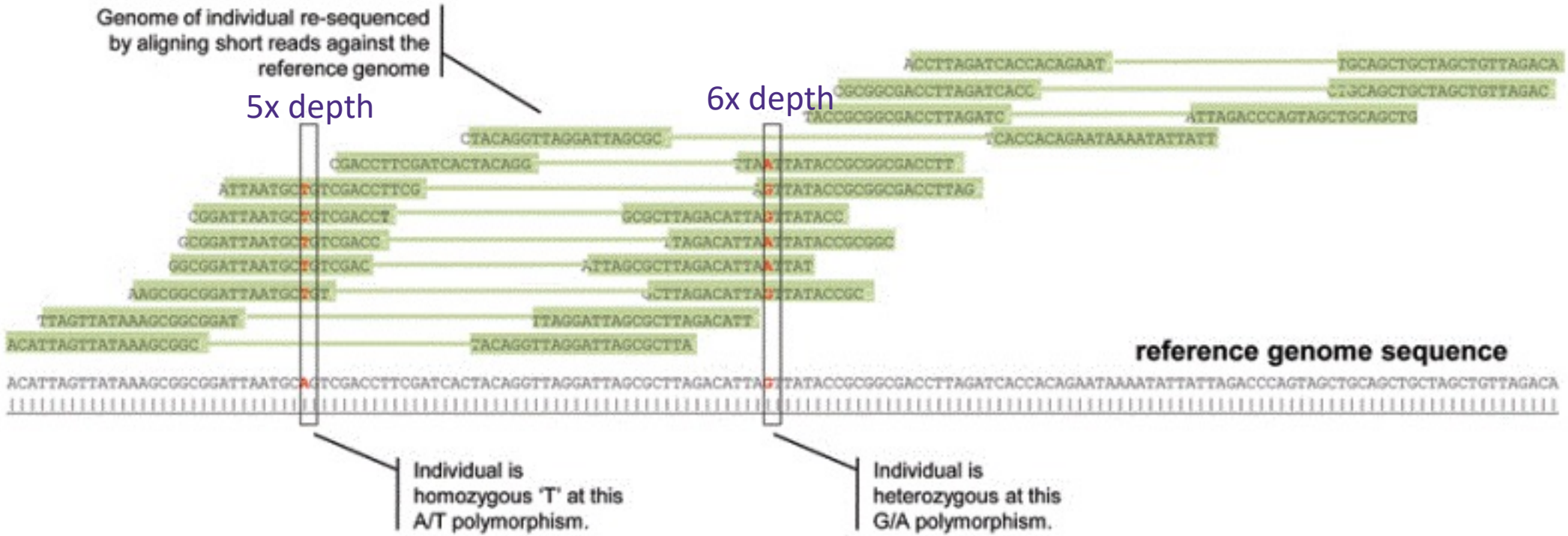
The Telomere-to-Telomere (T2T) Consortium has completed a challenging 8% of the human genome left unresolved by the initial Human Genome Project. In this data visualization, each chromosome begins at bottom right and wraps around, with chromosomes X and 1 through 22 arranged from the outside in (chromosome Y is not shown). The newly completed regions are highlighted in red.

~200M bp of novel sequence (total: 3,117,292,070 bp)
115 new protein coding genes (total 19,969 genes)

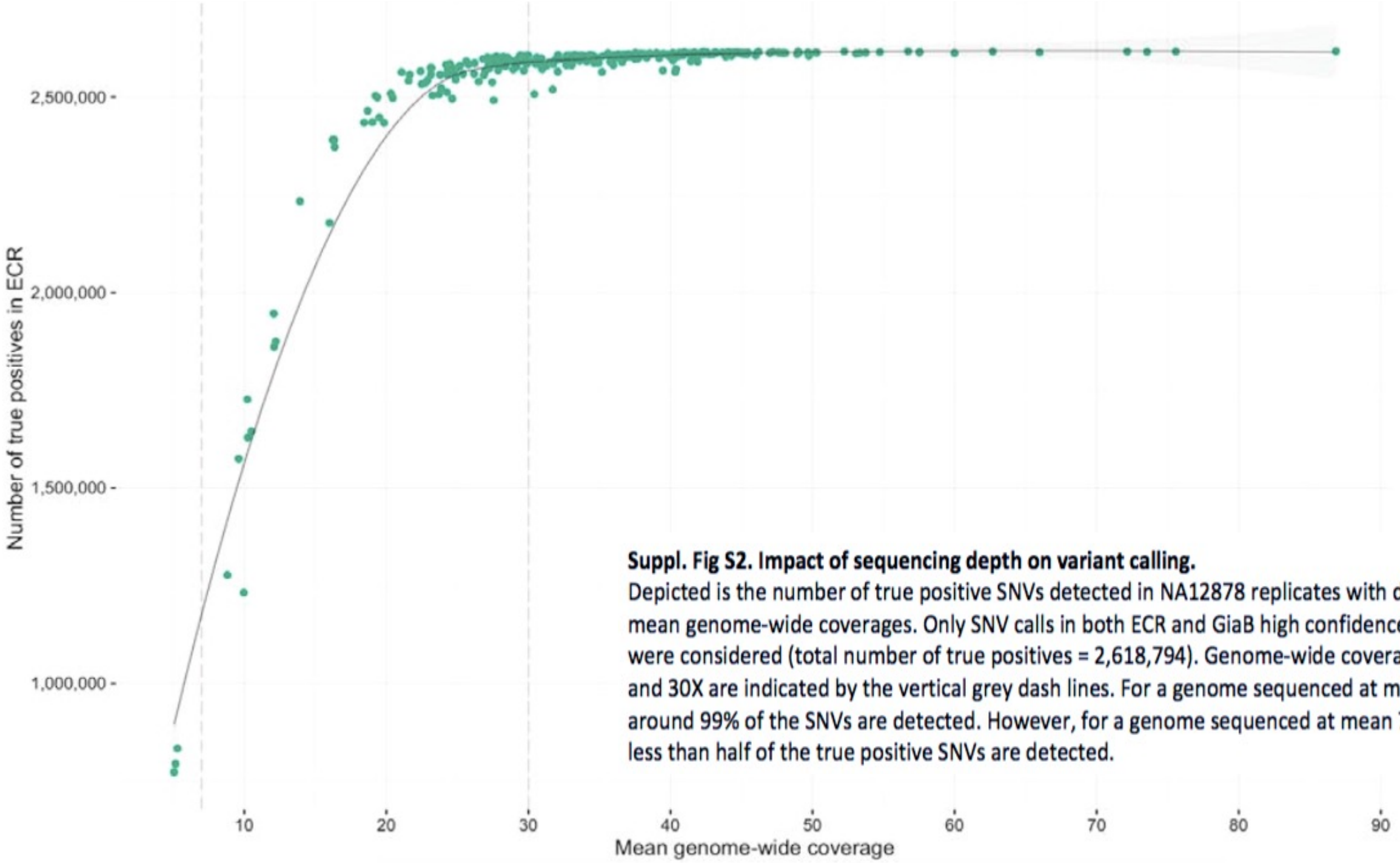
"Although CHM13 represents a complete human haplotype, it does not capture the full diversity of human genetic variation. To address this bias, the Human Pangenome Reference Consortium has joined with the T2T Consortium to build a collection of high-quality reference haplotypes from a diverse set of samples."

Sequencing alignment and depth

> Depth: The number of times a base-pair is sequenced



Paired end sequencing: Distance between each end is known, making alignment algorithms easier.

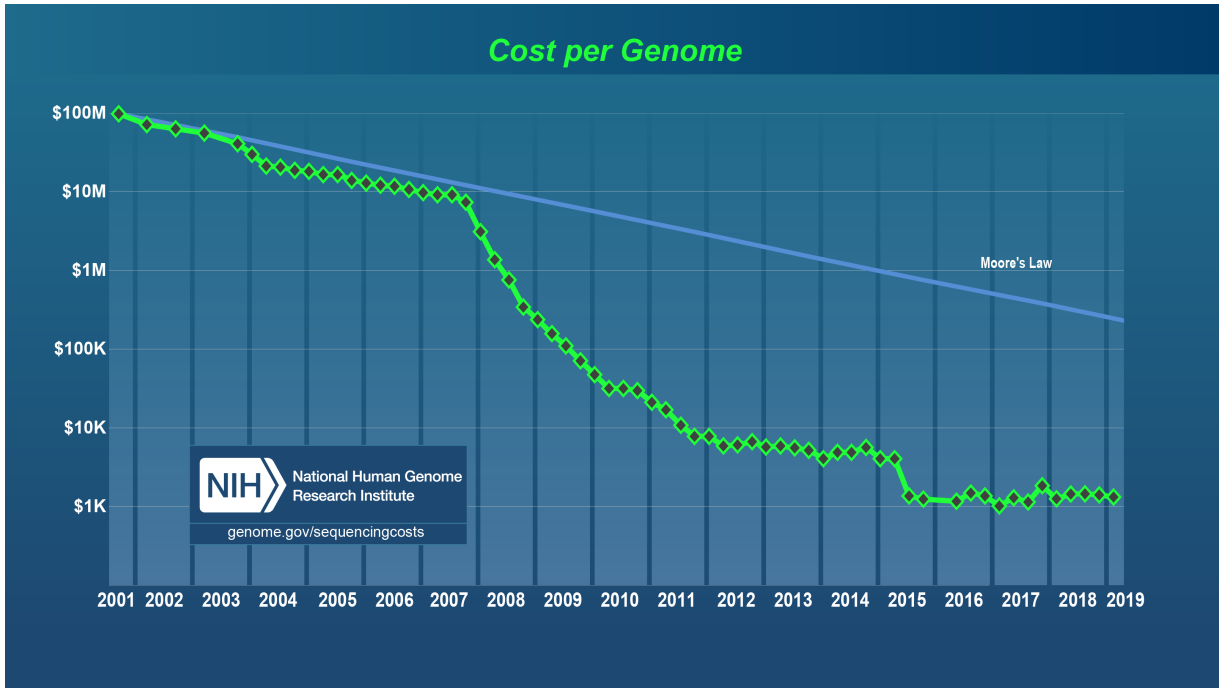


Suppl. Fig S2. Impact of sequencing depth on variant calling.
Depicted is the number of true positive SNVs detected in NA12878 replicates with different mean genome-wide coverages. Only SNV calls in both ECR and GiaB high confidence regions were considered (total number of true positives = 2,618,794). Genome-wide coverages of 7X and 30X are indicated by the vertical grey dash lines. For a genome sequenced at mean 30X, around 99% of the SNVs are detected. However, for a genome sequenced at mean 7X coverage, less than half of the true positive SNVs are detected.

Practical roadblocks to genome sequencing

Sequencing cost per genome is currently ~$1,000

Sequencing one genome generates ~200 GB data



Pricing Sequencing (CIDR, March 2023)

Illumina Sequencing		
Whole Genome, low pass 4X*		Inquire for pricing
Whole Genome (30X)	>96 samples	$1,000 (saliva DNA source $1,250)
Whole Exome	>90% @ 20X	~$300-$450 sample number dependent
Whole Exome, FFPE DNA source, mean 100X		$625-$850 Sample number dependent
Whole Exome Plus Custom content		Inquire for pricing
Custom Targeted (500 kb – 34 Mb options)		~$150 - $1000
Custom Targeted (amplicon; 10 – 250kb)		~$80-~$200
***Please Inquire for other options. If FFPE DNA Source, costs increase ~ 25%.**		

https://cidr.jhmi.edu/xtras/shared/documents/pricing.pdf

SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

Estimating ancestry using genetic data

Definitions

What is Race?

A sociopolitically constructed system for classifying and ranking human beings according to subjective beliefs about shared ancestry based on perceived innate biological similarities; the system varies globally.

What is Ethnicity?

A sociopolitically constructed system for classifying human beings according to claims of shared heritage often based on perceived cultural similarities (e.g., language, religion, beliefs); the system varies globally.

National Academies of Sciences, Engineering, and Medicine. 2023. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field.

What is ancestry?

Genealogical ancestry

Genealogical ancestry describes information about your ancestors from whom you are biologically descended. If one of your ancestors belonged to a particular group X, you might say that you have some "X" ancestry. For example, if one of your four grandparents was Swedish you might describe yourself as "one fourth Swede".

Genetic ancestry

The paths through an individual's family tree by which they have inherited DNA from specific ancestors. Genetic ancestry can be thought of in terms of lines extending upwards in a family tree from an individual through their genetic ancestors. Shared genetic ancestry arises from having genetic ancestors in common (that is, overlapping lines of ancestry). In practice, shared genetic ancestry is typically inferred by some measure(s) of genetic similarity.

The difference between ***genealogical*** and ***genetic*** ancestry can be illustrated by full siblings. Full siblings have identical genealogical ancestry but differ in their genetic ancestry, due to differences in transmission of chromosomal segments from their parents.

Mathieson I and Scally A. PLoS Genet, 2020

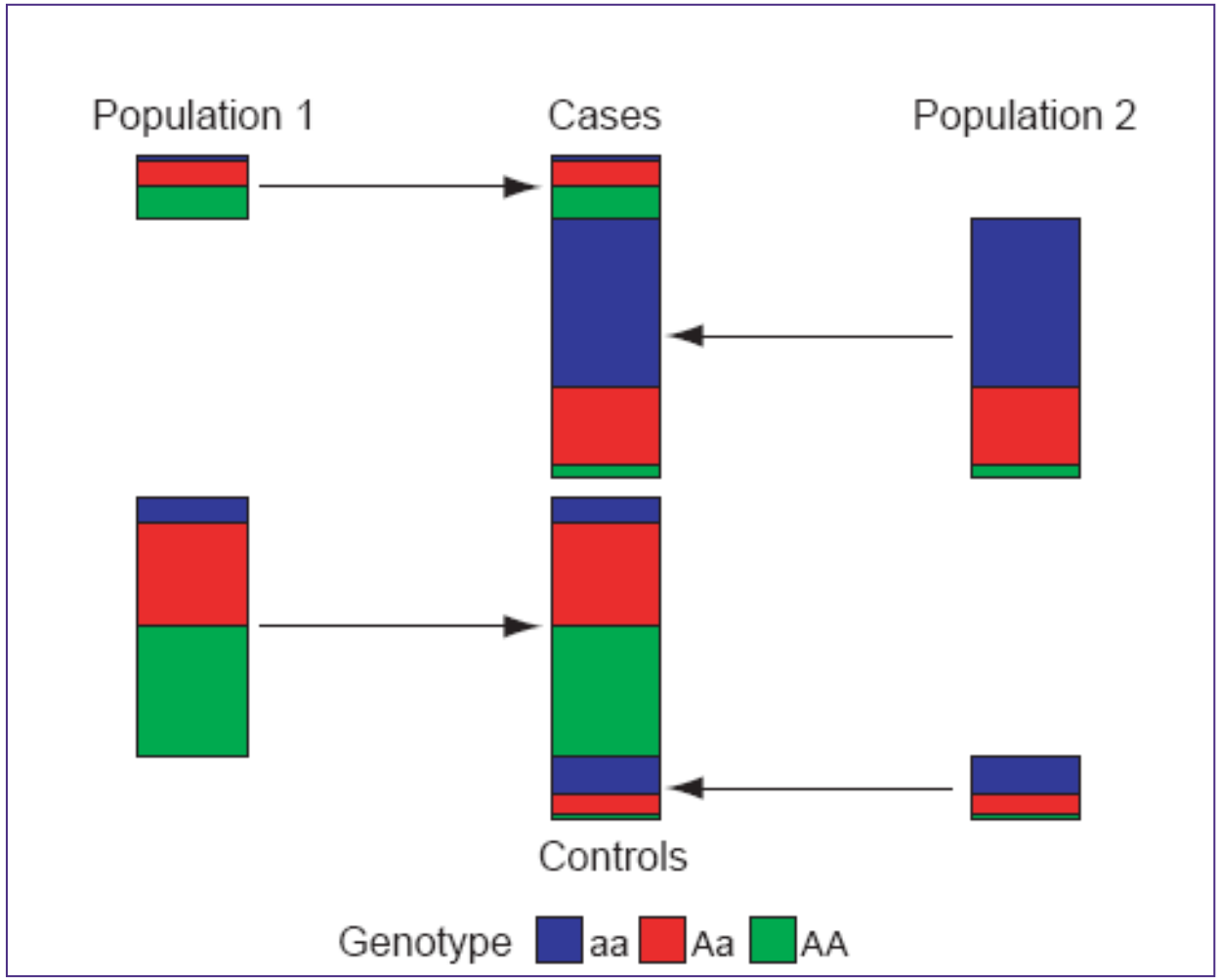
National Academies of Sciences, Engineering, and Medicine. 2023. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field.

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Assume we conduct a case-control GWAS…

- Our cases were collected in Africa
- Our controls were collected in Asia
- If we identify multiple alleles that are significantly more common in cases compared to controls, do we believe that these results are due to association with disease or due to population differences?

Population Stratification - Confounding by ancestry



Group differences in ancestry AND outcome

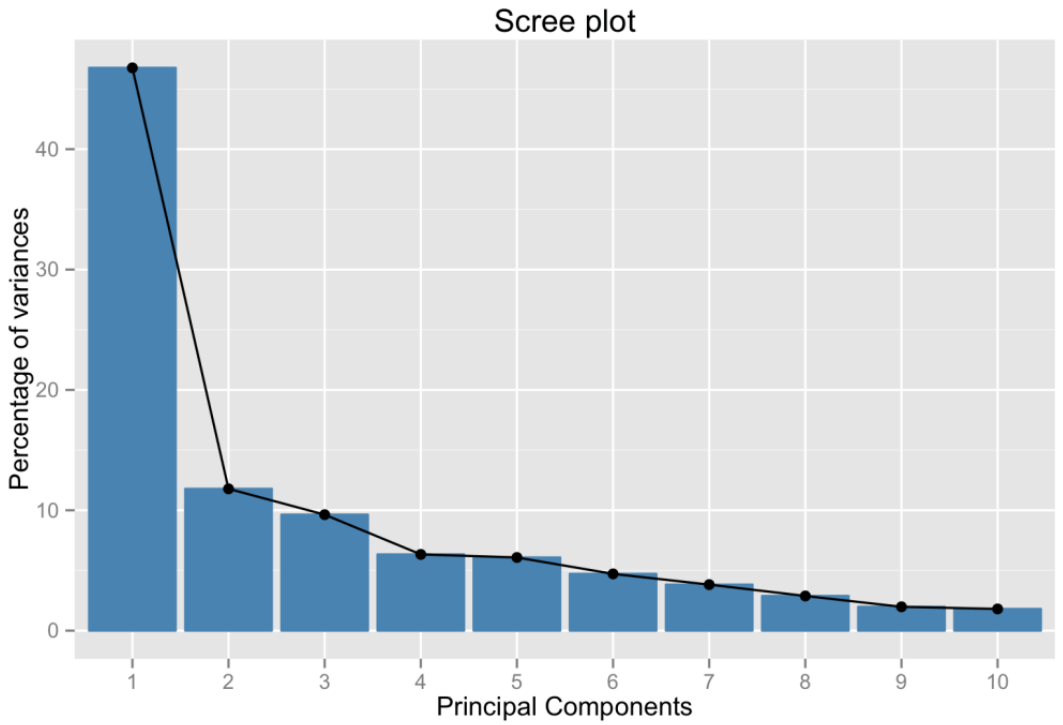
We can use genetic data to derive ancestry-informed covariates and adjust for these in our association studies.

Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

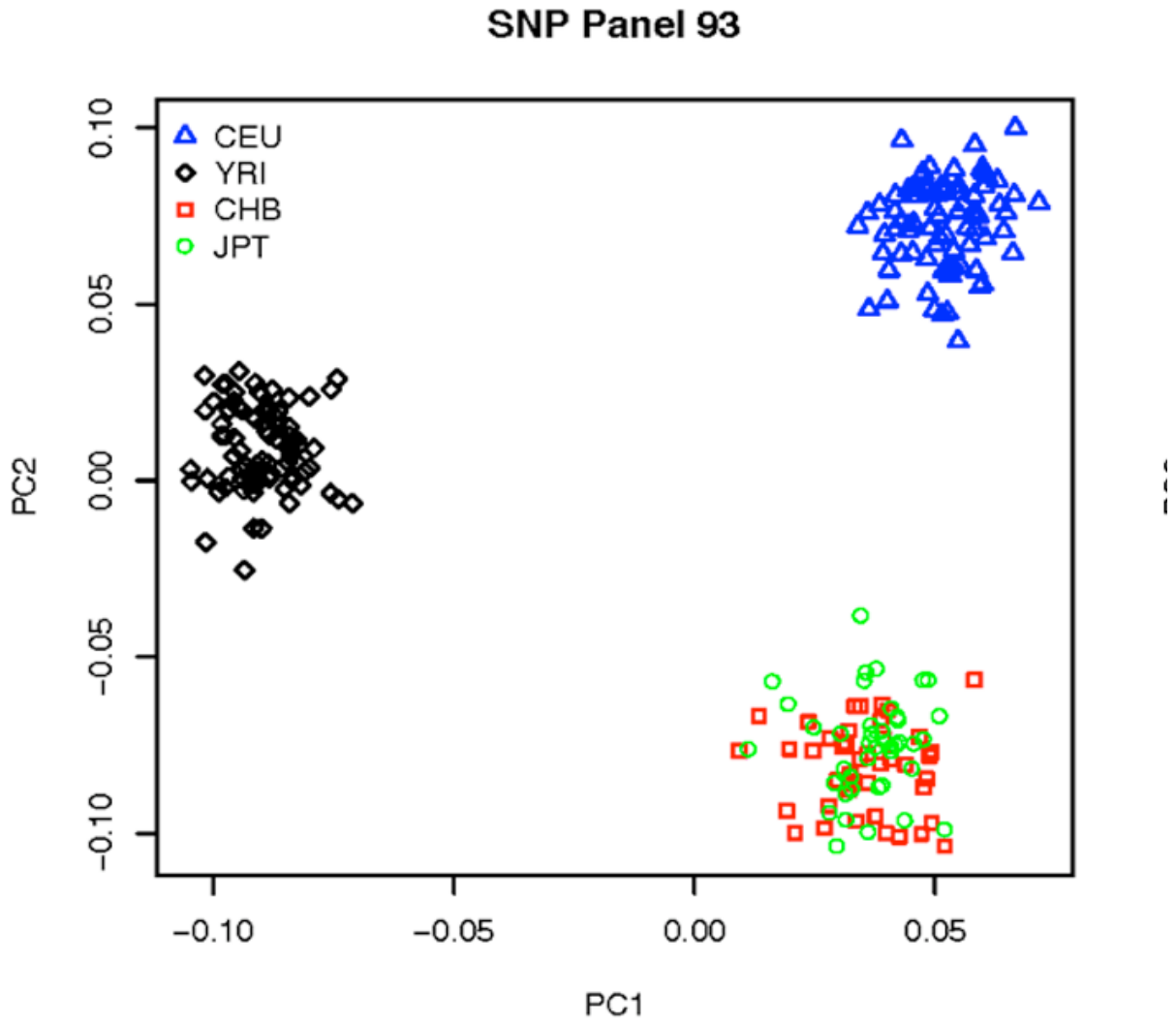
Principal Component Analysis (PCA)

- Reduces the dimension of the data from MANY SNPs to a small set of principal components (PCs) that can explain most of the variation in the data
- The first PC (PC1) is constructed to explain as much of the variation as possible, the second PC (PC2) is constructed to explain as much of the remaining variation as possible, …
- The more correlation in the data (i.e., between SNPs), the fewer PCs are needed to explain high proportions of the variation.
- Each PC is a linear combination of all SNPs, and PCs are indepenent of each other



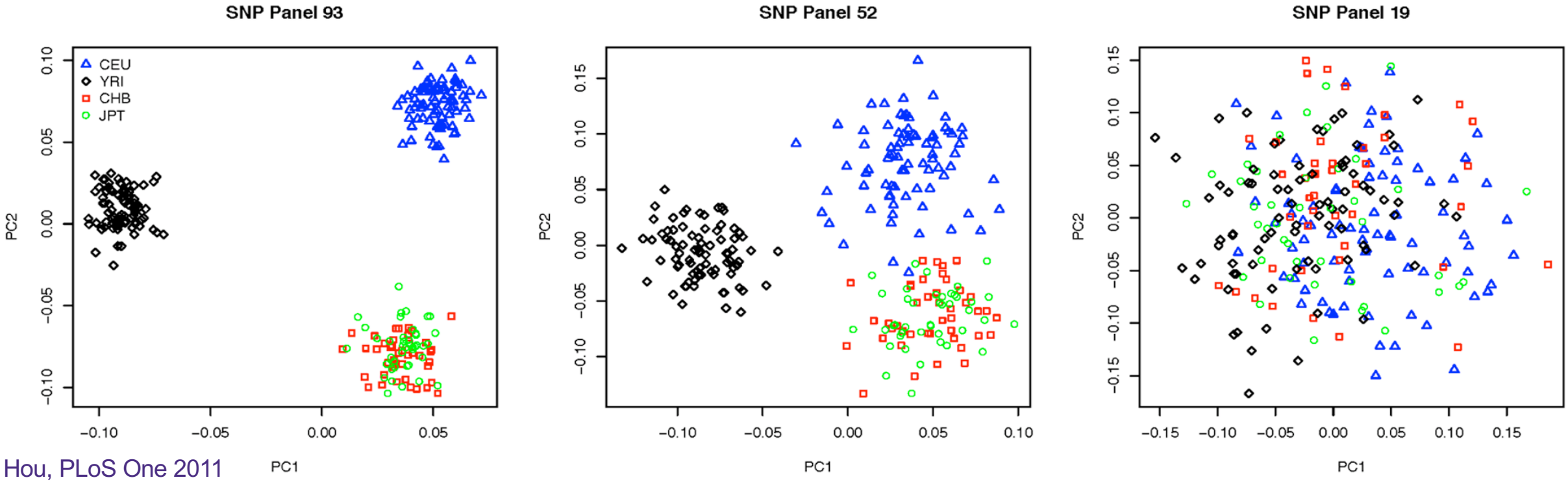
EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

The first two PCs can help distinguish populations between continents



Hou, PLoS One 2011

The first two PCs can help distinguish populations between continents



Hou, PLoS One 2011

Ancestry informative markers (AIMs): Few variants selected to capture major variation between select populations

Nowadays, ideally, we use ~10K – 100K variants (depending on the study goal) selected from GWAS data

--Variant selection criteria: common (e.g., MAF>5%), independent (e.g., r^2<10%), and either genotyped directly or imputed with r^2>0.9

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 depending on how many variant alleles an individual carries (e.g., AA- 0, AG – 1, GG - 2)
- Multiply that genotype value by the loading value for each SNP
- Sum over all SNPs to get a final PC value for that individual

Individual	SNP1 Loading = 4	SNP2 Loading = 0.3	SNP3 Loading = -2	SNP4 Loading = 1	PC1 total
A	2	1	1	0	
B	1	0	2	1	

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 depending on how many variant alleles an individual carries (e.g., AA- 0, AG – 1, GG - 2)
- Multiply that genotype value by the loading value for each SNP
- Sum over all SNPs to get a final PC value for that individual

Individual	SNP1 Loading = 4	SNP2 Loading = 0.3	SNP3 Loading = -2	SNP4 Loading = 1	PC1 total
A	2*4 = 8	1*0.3 = 0.3	1*-2 = -2	0*1 = 0	
B	1*4 = 4	0	2*-2 = -4	1*1=1	

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 depending on how many variant alleles an individual carries (e.g., AA- 0, AG – 1, GG - 2)
- Multiply that genotype value by the loading value for each SNP
- Sum over all SNPs to get a final PC value for that individual

Individual	SNP1 Loading = 4	SNP2 Loading = 0.3	SNP3 Loading = -2	SNP4 Loading = 1	PC1 total
A	2*4 = 8	1*0.3 = 0.3	1*-2 = -2	0*1 = 0	6.3
B	1*4 = 4	0	2*-2 = -4	1*1=1	1

Include PCs in your genetic association study

$$\mathbf{Y = \beta_G * genotype + \beta_1 PC1 + \beta_2 PC2 + \beta_3 PC3 + \beta_4 PC4 + \beta_5 PC5}$$

> Accounts for underlying patterns in the population that are not truly associated with a particular phenotype but may appear to be so due to differences in allele frequency and trait distribution associated with ancestry.

> PCA has become a standard tool to investigate genetic ancestry patterns in genome-wide data. We can use PCs for

- Population genetics
- Identify "population outliers"
- Identify any other structure that is not obvious

Breakout Activity

What populations are separated by PC1? And by PC2?

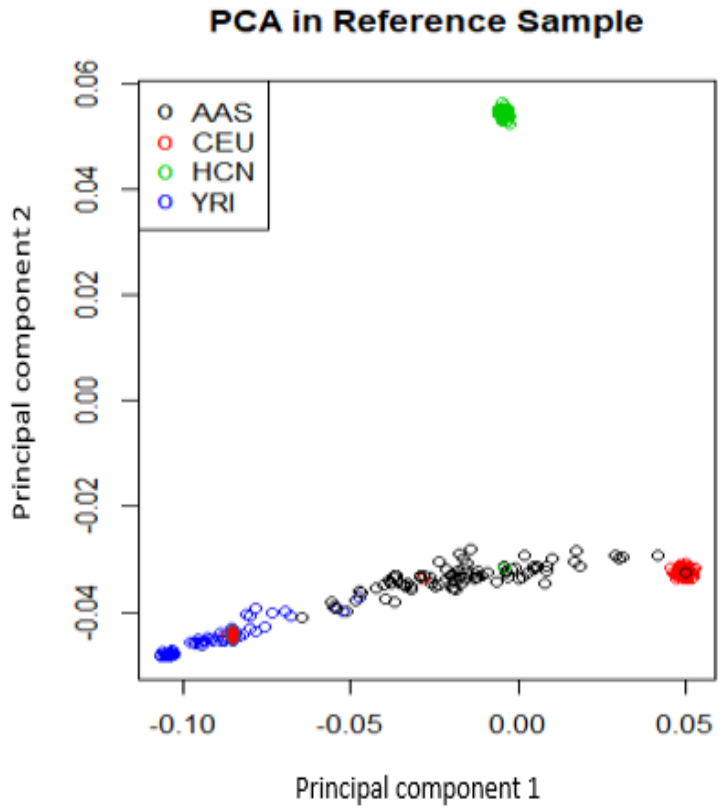
Why do we see clustering of three populations (blue, red, green), while the black circles are spread across the PC1 axis?

Notice the red dot in the lower left corner among the blue dots. What might be happening here?

Where on this plot might you see people who describe their ancestry as Chinese American (ancestors from both European and Chinese populations)?

Can we use PCs to sufficiently account for the observed population structure across these four populations in regression models?

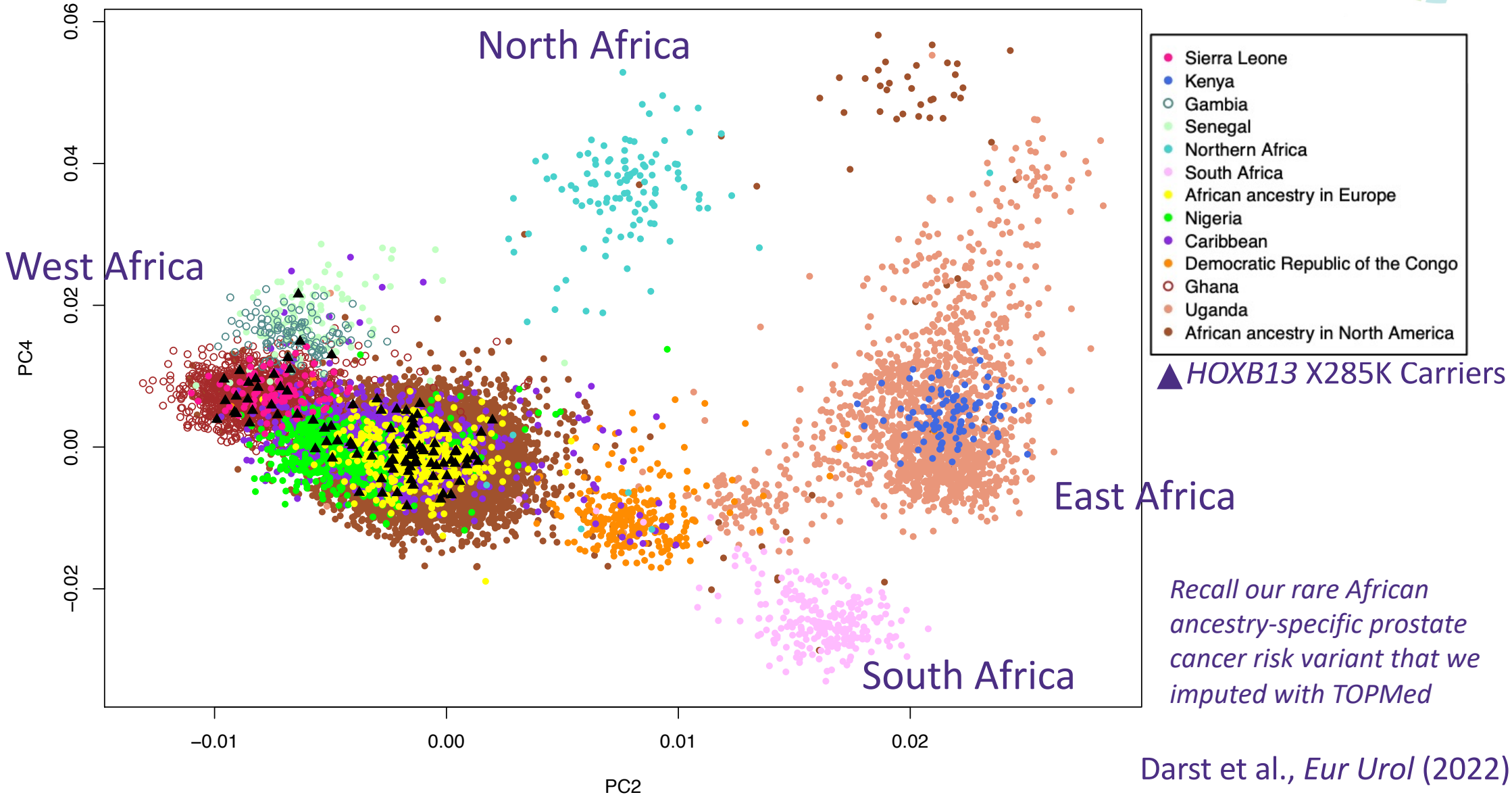
What are pros/cons of using self-described race vs genetic ancestry in epidemiology studies? Think of what each can tell you based on the questions you are trying to ask.



*PCA plot of African Americans from the Southeast (**AAS**), Europeans from Utah (**CEU**), Yorubans from Nigeria (**YRI**), and Han Chinese from Beijing (**HCN**). Each dot represents one person, and each person is color-coded based on their self-described race.*

What about population structure within continents?

Population structure in Africa

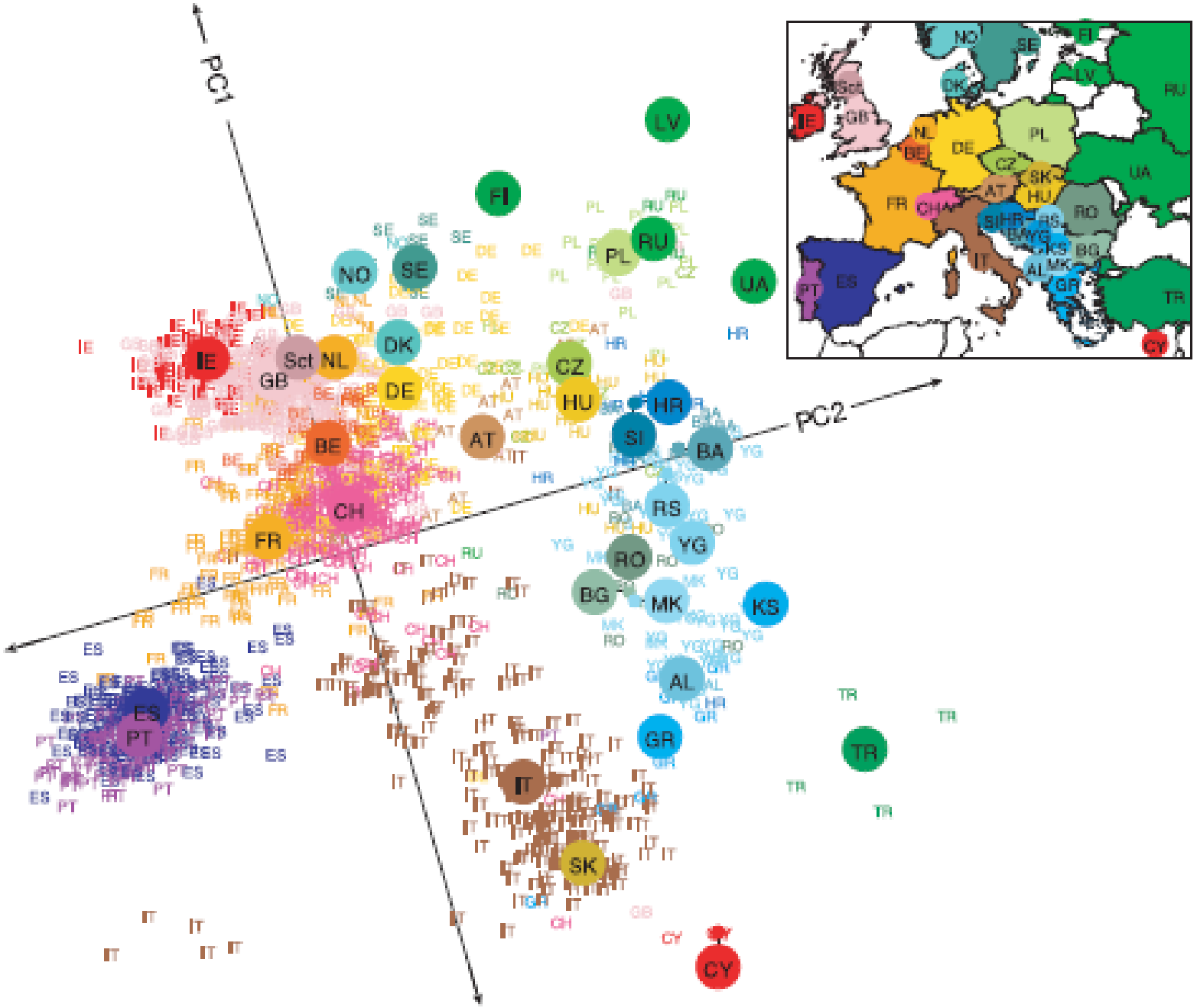


Recall our rare African ancestry-specific prostate cancer risk variant that we imputed with TOPMed

Darst et al., *Eur Urol* (2022)

Population Structure in Europe

1,387 samples
~200,000 SNPs



Novembre et al, Nature 2008

23andMe Ancestry Composition

Reference data

400,000 reference individuals with ancestry from >150 countries/2,000 subregions

People who report four grandparents all born in the same country