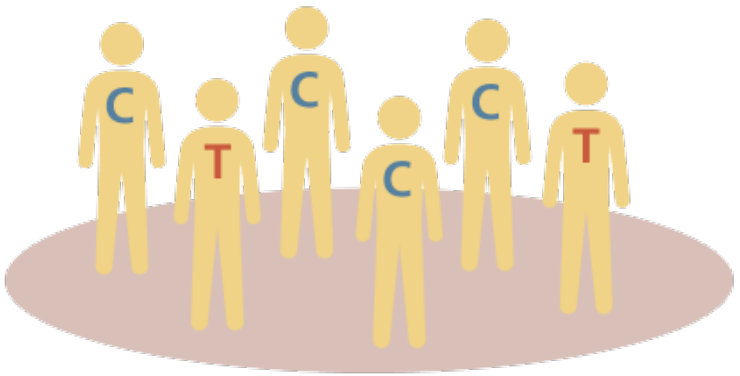
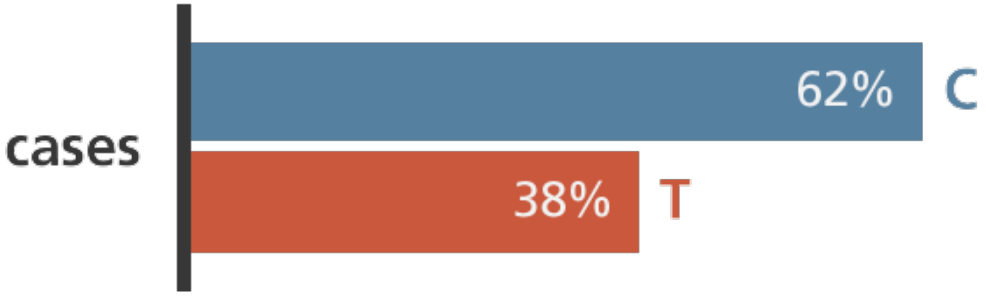
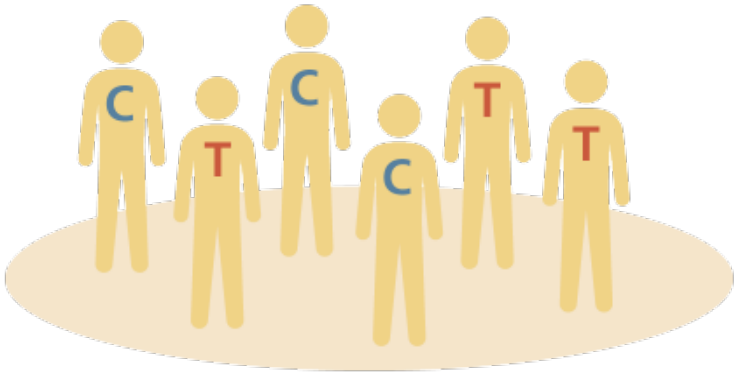
Session 6:
Study designs for genetic association studies

SNP



cases (n=1,000)
people with heart disease



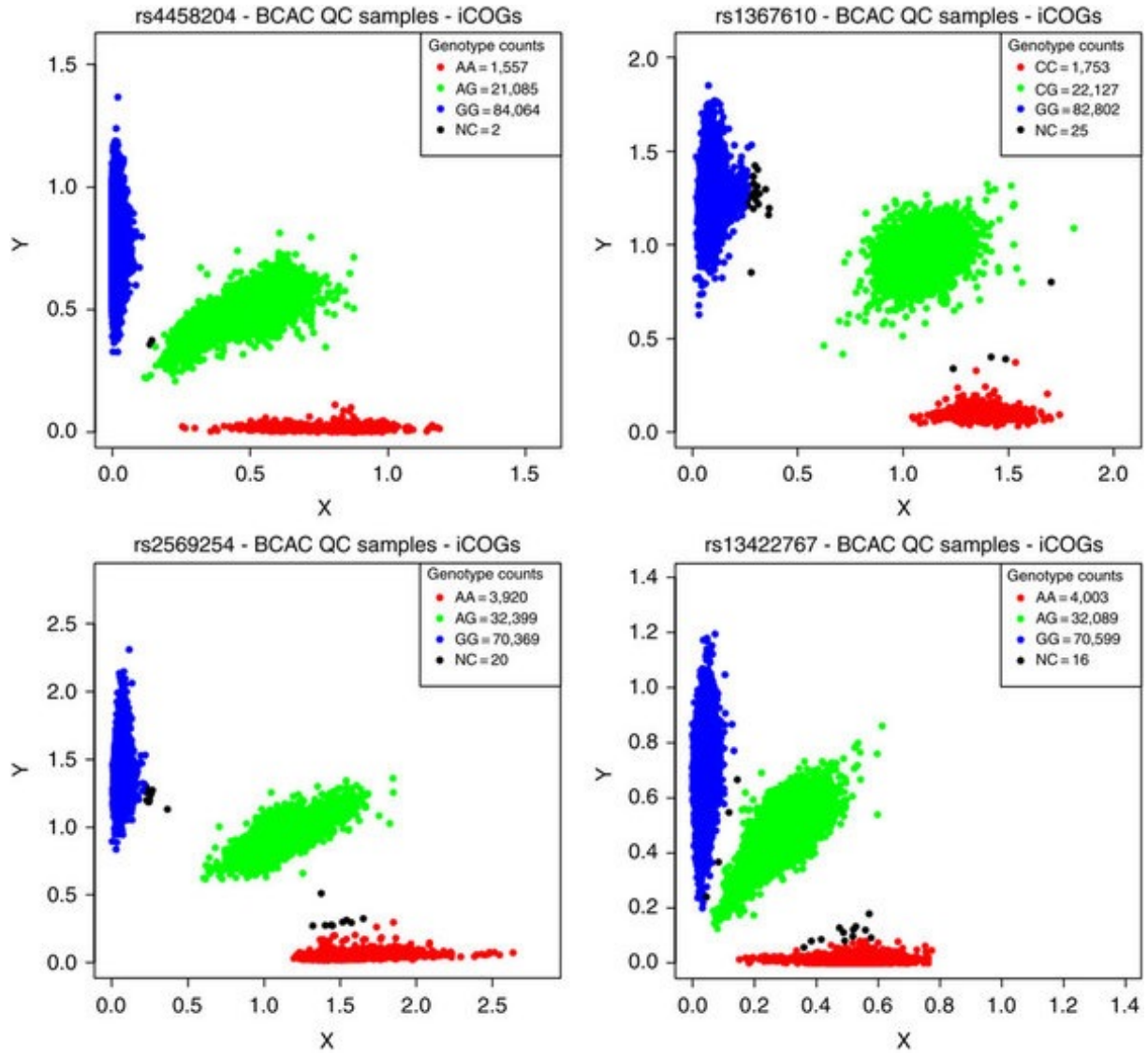


controls (n=1,000)
people without heart disease

Genotyping vs. sequencing

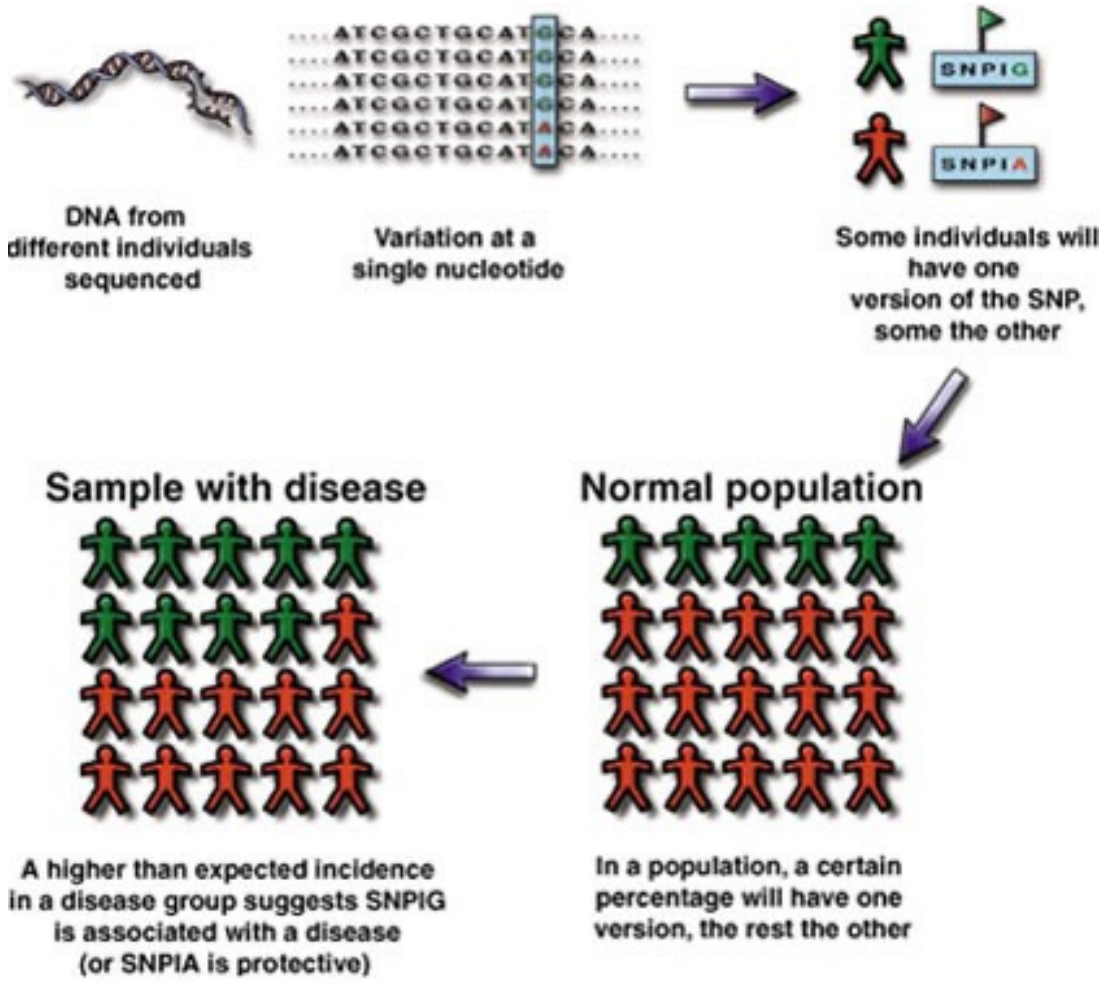
- Genotyping: Target a particular genetic variant and ”measure” it
- Sequencing: Target a region (could be the whole genome) and “measure” the entire region (all base-pairs)
- From a bioinformatic/analysis point of view, genotyping data is much easier to handle.

Genotyping Output



Li, Nat Comm 2014

Genetic association studies using SNPs



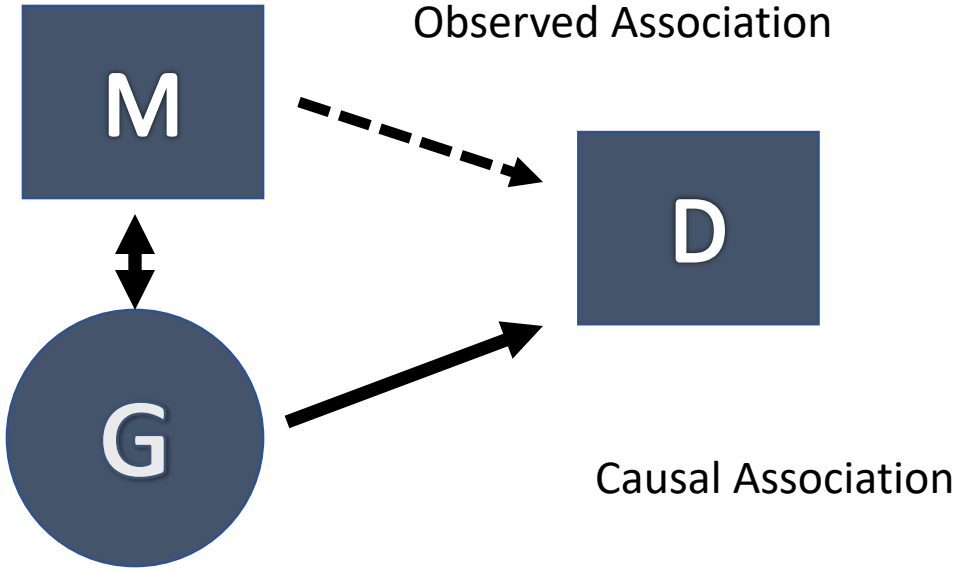
Why we like SNPs:

- Abundant in the genome
- Easy to measure

In the early era of genetic epidemiology (-2008)

- The human genome had not been mapped so we did not know where to find genetic variation
- Genotyping was expensive: only a handful of SNPs were genotyped
  - SNPs were often coding variants/had a known function
- Candidate gene studies – pick your favorite gene!
  - Choose which SNPs to genotype in the entire population - if you choose your SNPs carefully, they could explain most of the variation in the gene (LD!)

The use of “tags” (proxy markers)



G and M are in LD with each other

An association between M and D suggests there may be a causal marker near M (also called indirect association)

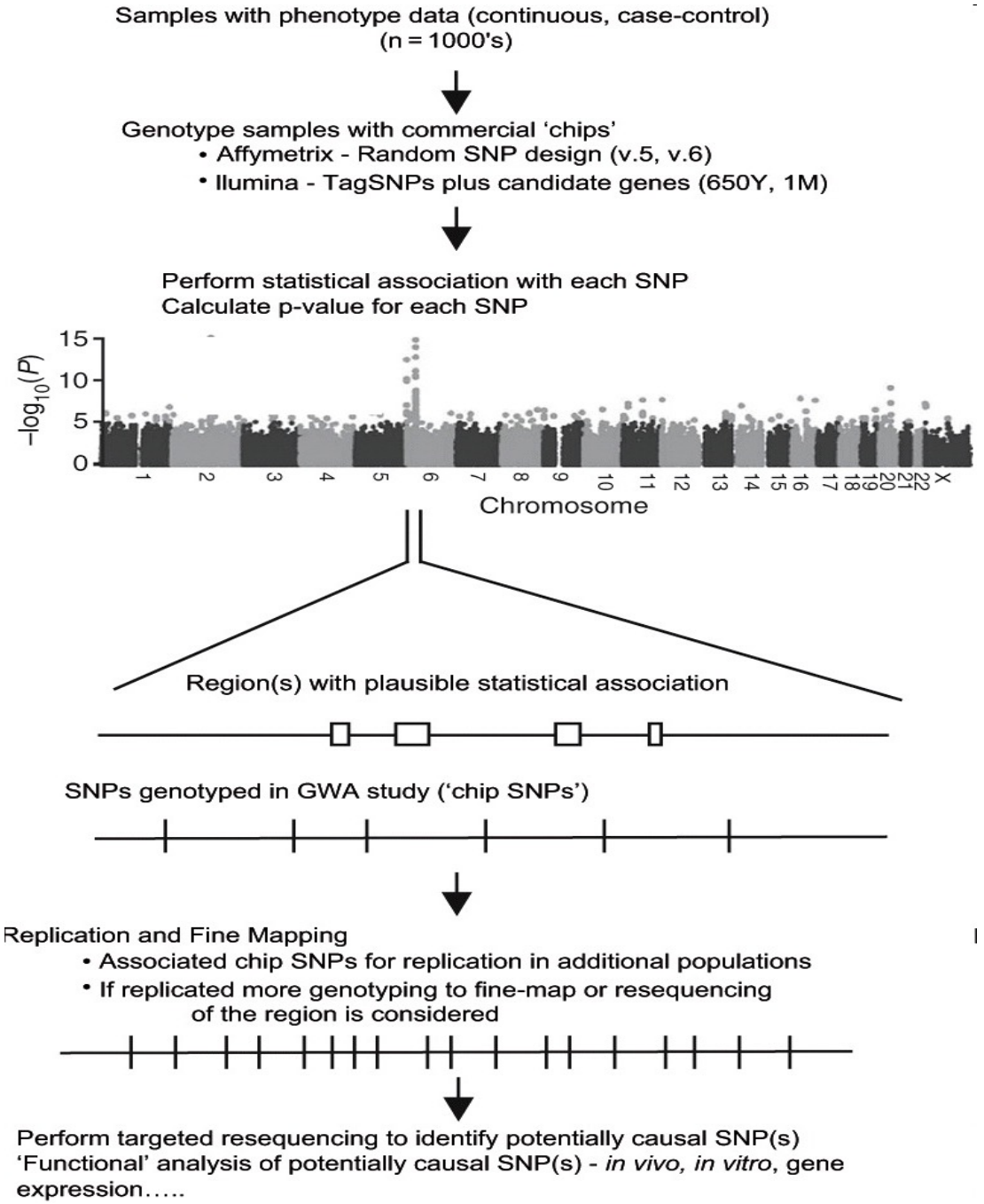
If the r^2 between M and G is 0.5 you need to double your sample size to obtain the same power as if you had measured G directly

When there is strong LD in a region, we will have very limited loss of power in our association studies even though we are only genotyping a few SNPs.
Caveat: Rare variation (<5%) will not be captured

Genome-wide association studies (GWAS)

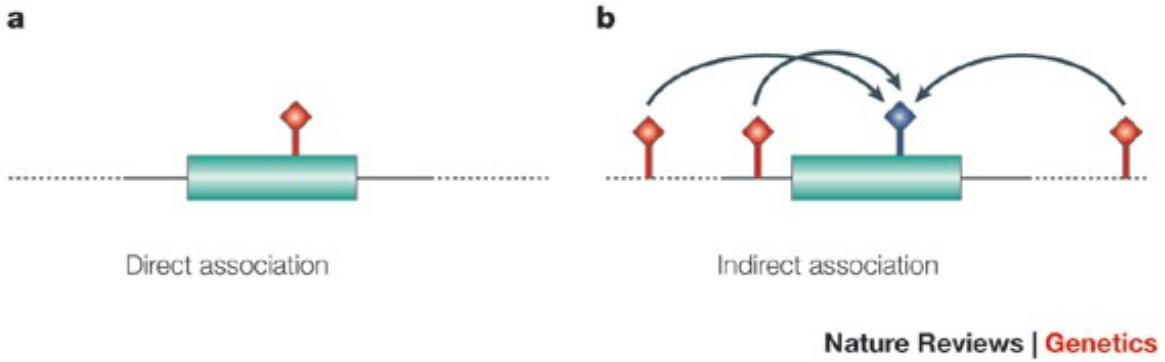
Screen the genome for SNPs that are associated with your trait (agnostic approach)

Rieder et al. 2008

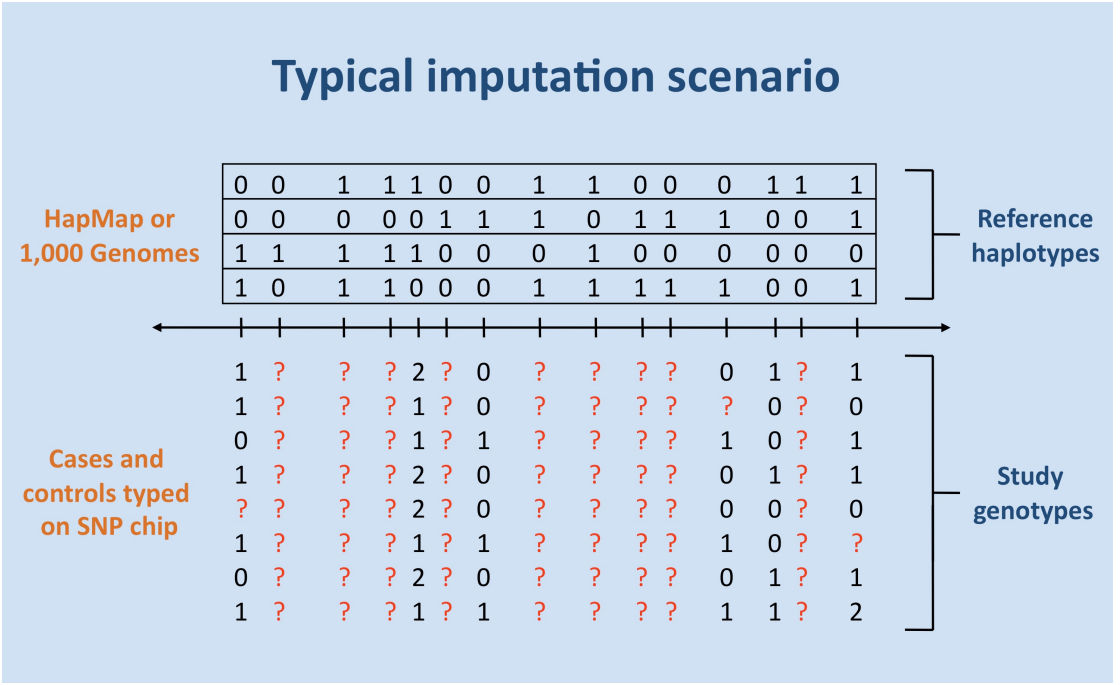


Genetic association studies rely heavily on LD

1) Indirect association



2) Imputation



Hirschhorn & Daly. *Nature Reviews Genetics 2005,* http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Imputation (I)

- Cost efficient
  - Can assess more SNPs than we genotyped

- Maximizes our sample size
  - Fills in missing values for already genotyped SNPs

- Allows us to combine existing data from different arrays that genotype different SNPs

Imputation (II)

- We can infer genotypes for SNPs we did not genotype (or failed in the lab)
  - **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

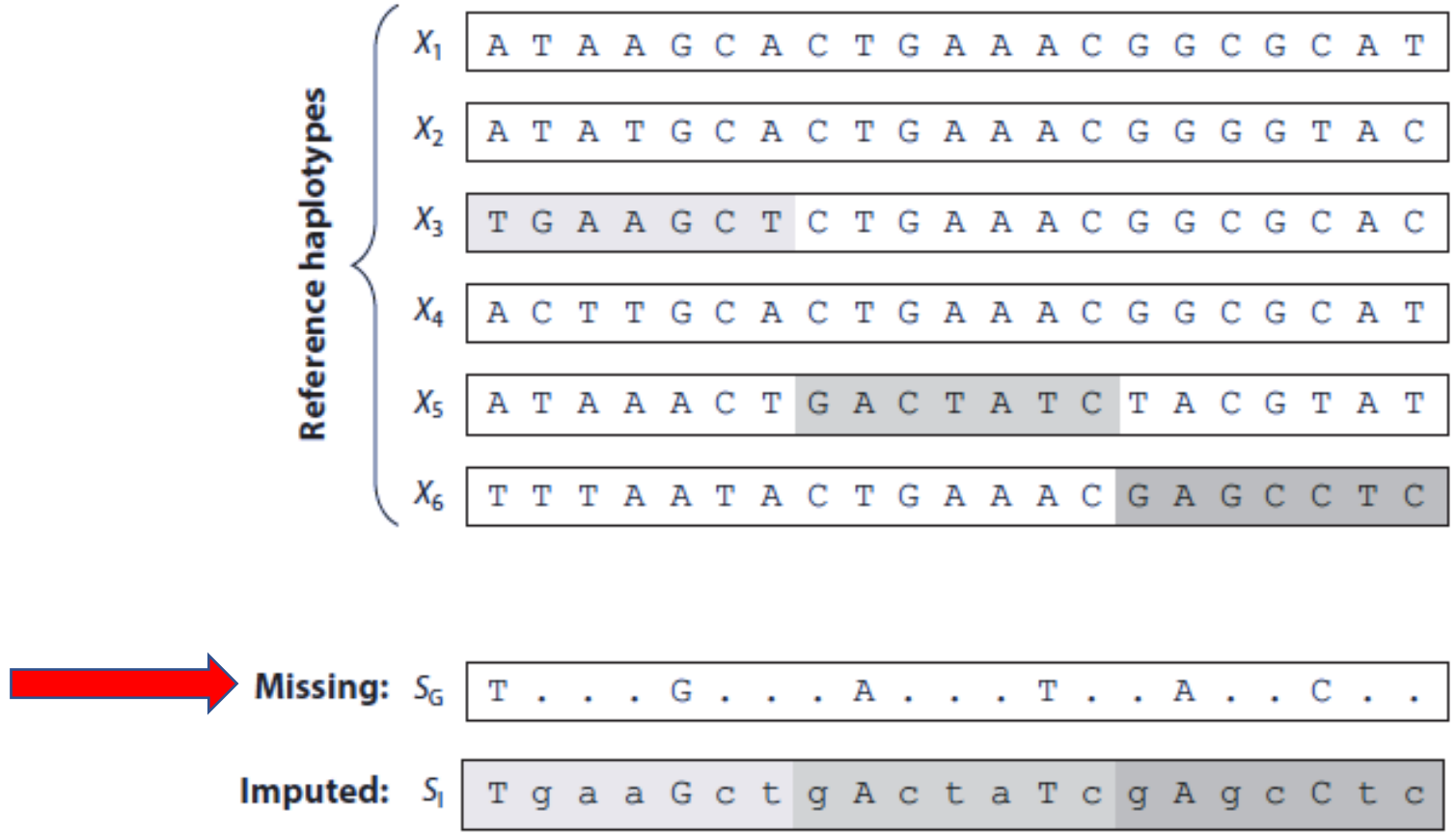


Figure 2

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 (III)

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

- Software: MACH, minimac, IMPUTE2, Beagle, PLINK

- Outputs:
  - Posterior probabilities for each potential genotype with three data points per SNP/individual [IMPUTE and BEAGLE]
  - “Dosage” of each imputed genotype ranging between 0-2, representing copies of the reference allele (continuous number) [MACH and BEAGLE].

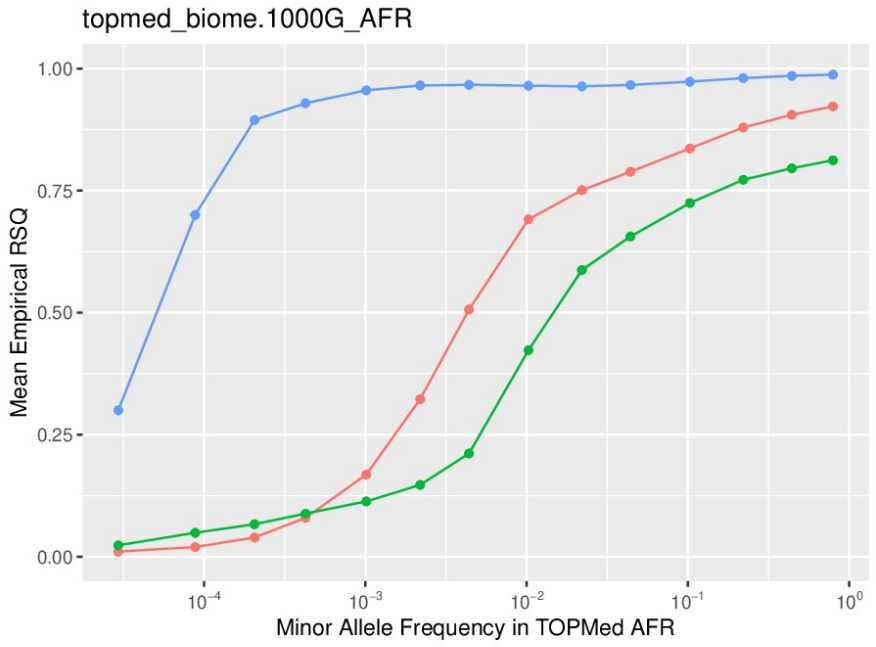
For more information about imputation, read Marchini and Howie, Nat Rev Genet 2010

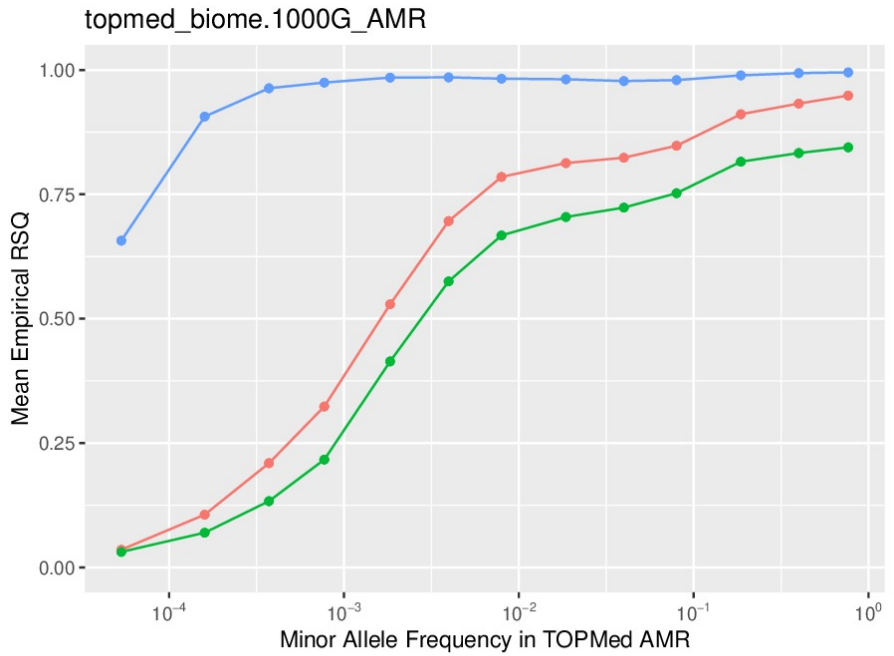
Imputation (IV)

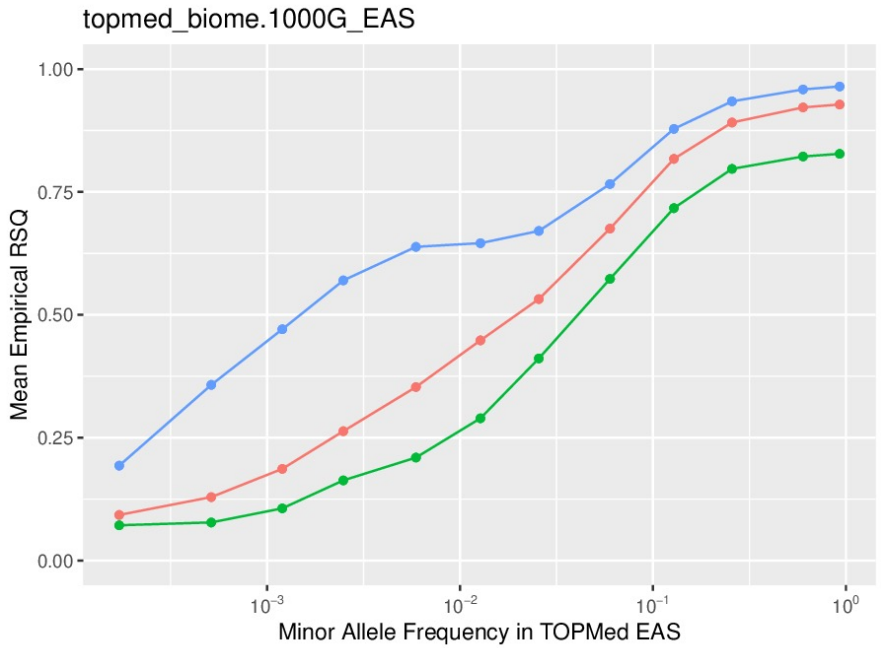
- The imputation quality score r^2 measures how well a SNP was imputed.
  - Ranges between 0 and 1.
  - Typically, a cut-off of 0.30 or so will flag most of the poorly imputed SNPs, but only a small number (<1%) of well imputed SNPs.

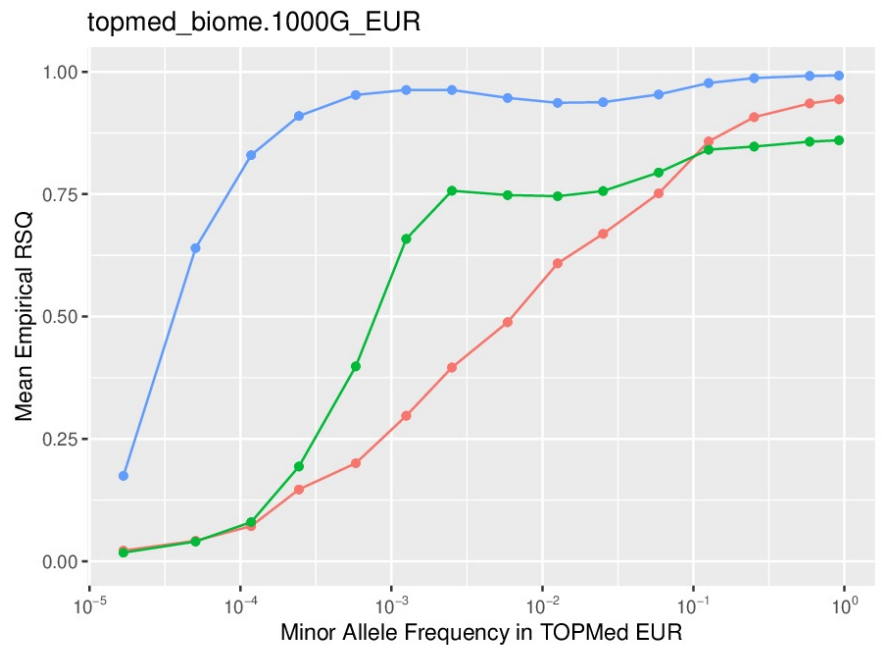
- Factors that affect imputation quality:
  - Number of genotyped SNPs in your data
  - Size of reference panel
  - Similarity in genetic ancestry between reference and study samples
  - Allele frequency

Reference Panels	N	Ancestry
HapMap	60	EUR
1000 Genomes Phase 3	2,504	Mixed
CAAPA	883	African American
HRC	32,470	EUR
TopMed	97,256	Mixed

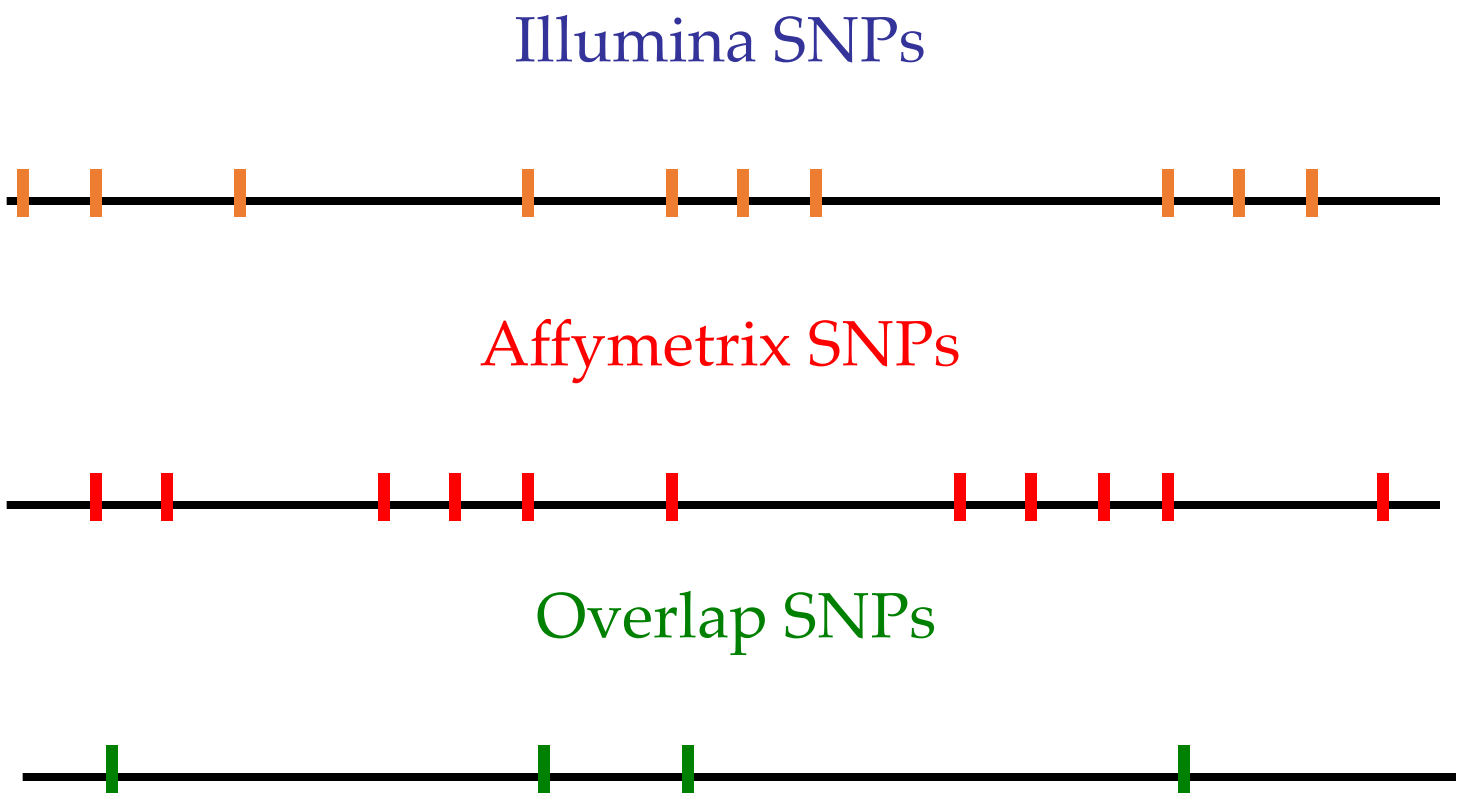








Imputation for studying SNPs across platforms



Imputation for studying SNPs across platforms

1000G SNPs

Illumina SNPs

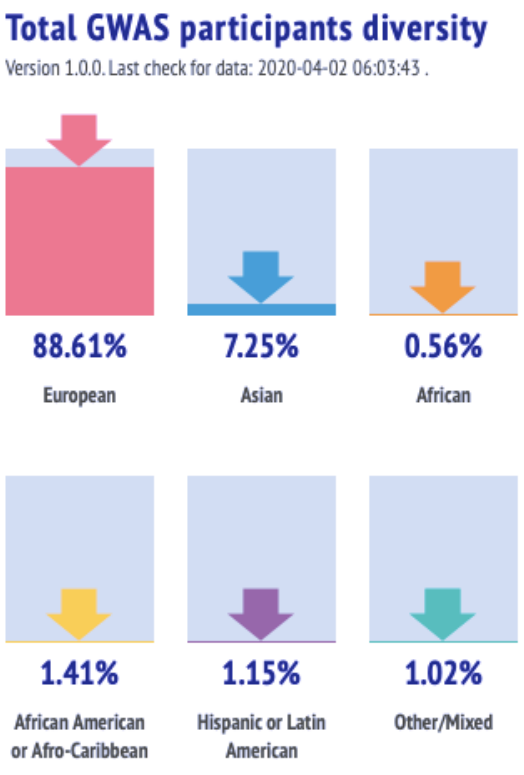
Affymetrix SNPs

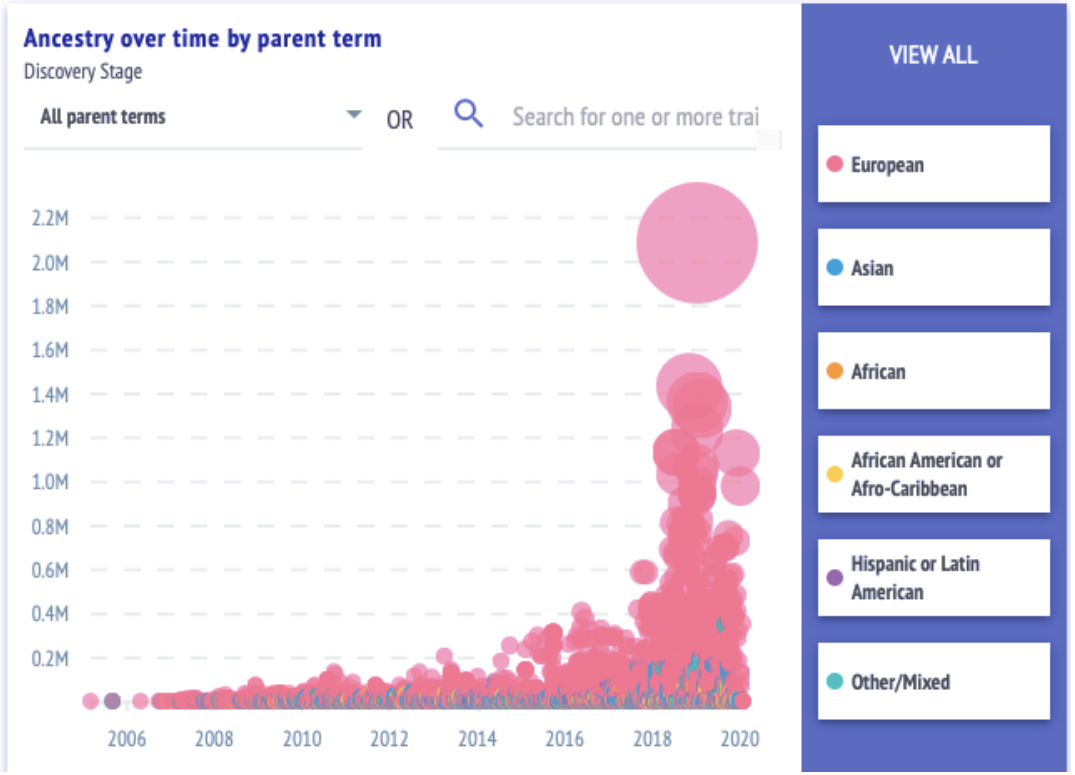
Overlap SNPs

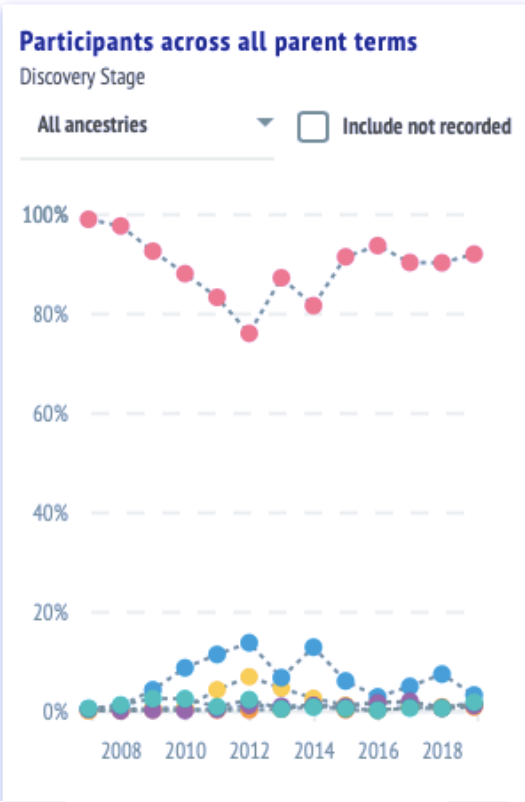
Limitations with traditional genome-wide genotyping arrays

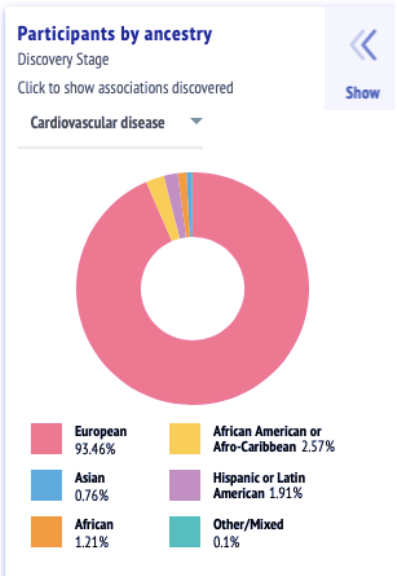
- Genotyping arrays are often designed to capture genetic variation in populations of European ancestry

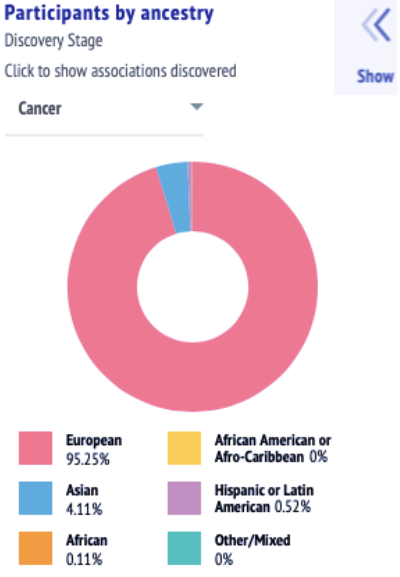
- Only capture common SNPs

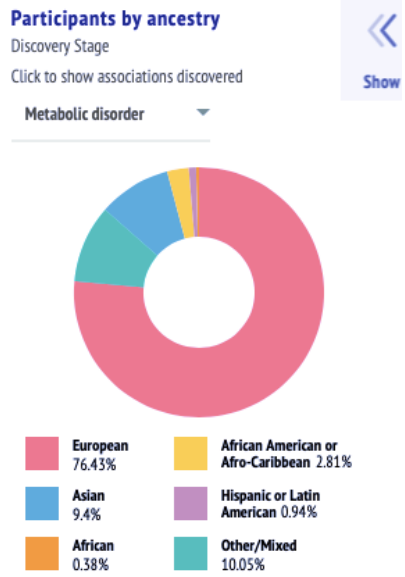












Popejoy and Fullerton, Nature 2016

https://gwasdiversitymonitor.com

Breakout Room Discussion:

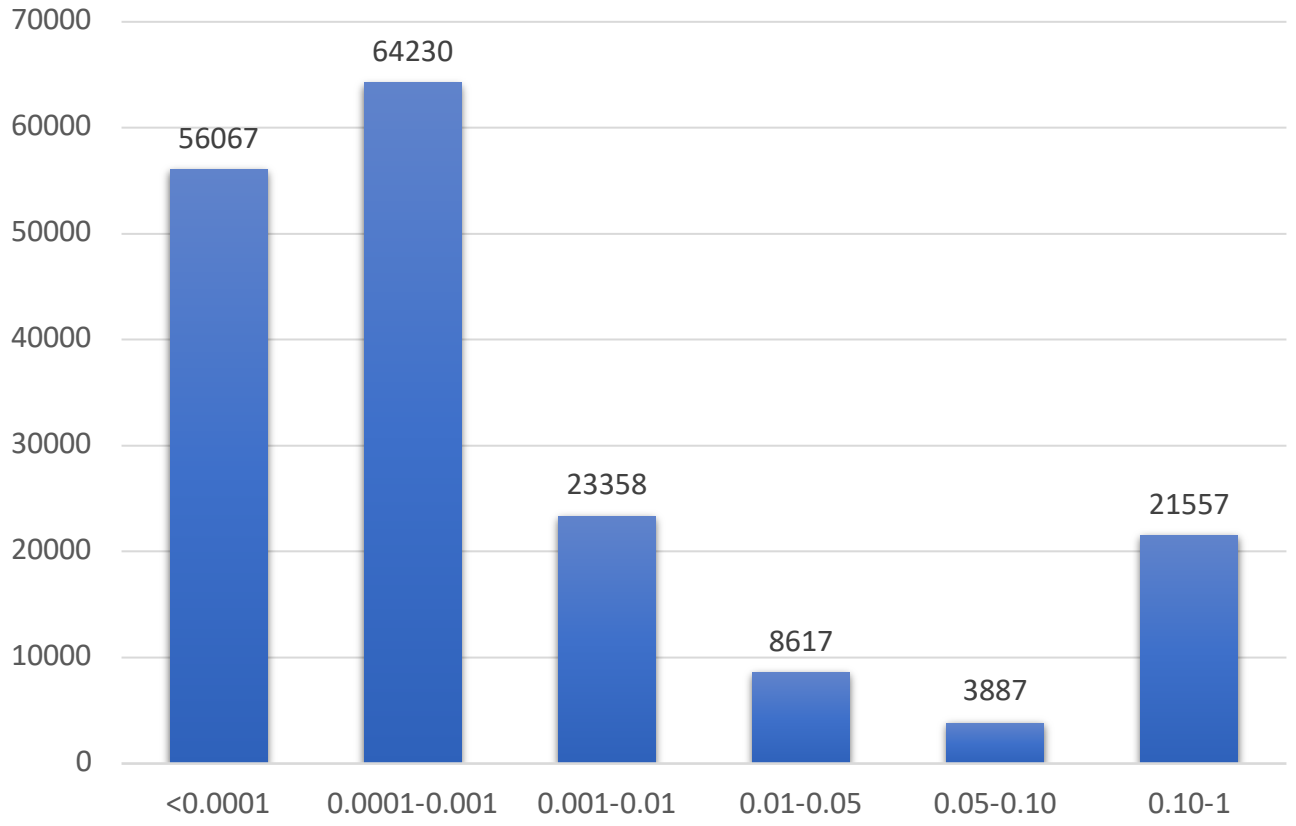
- Explore the breakdown of genetic ancestry in GWAS as reported on the website https://gwasdiversitymonitor.com. What do you notice about recent trends? What populations seem over- and under-represented in genetic studies?

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

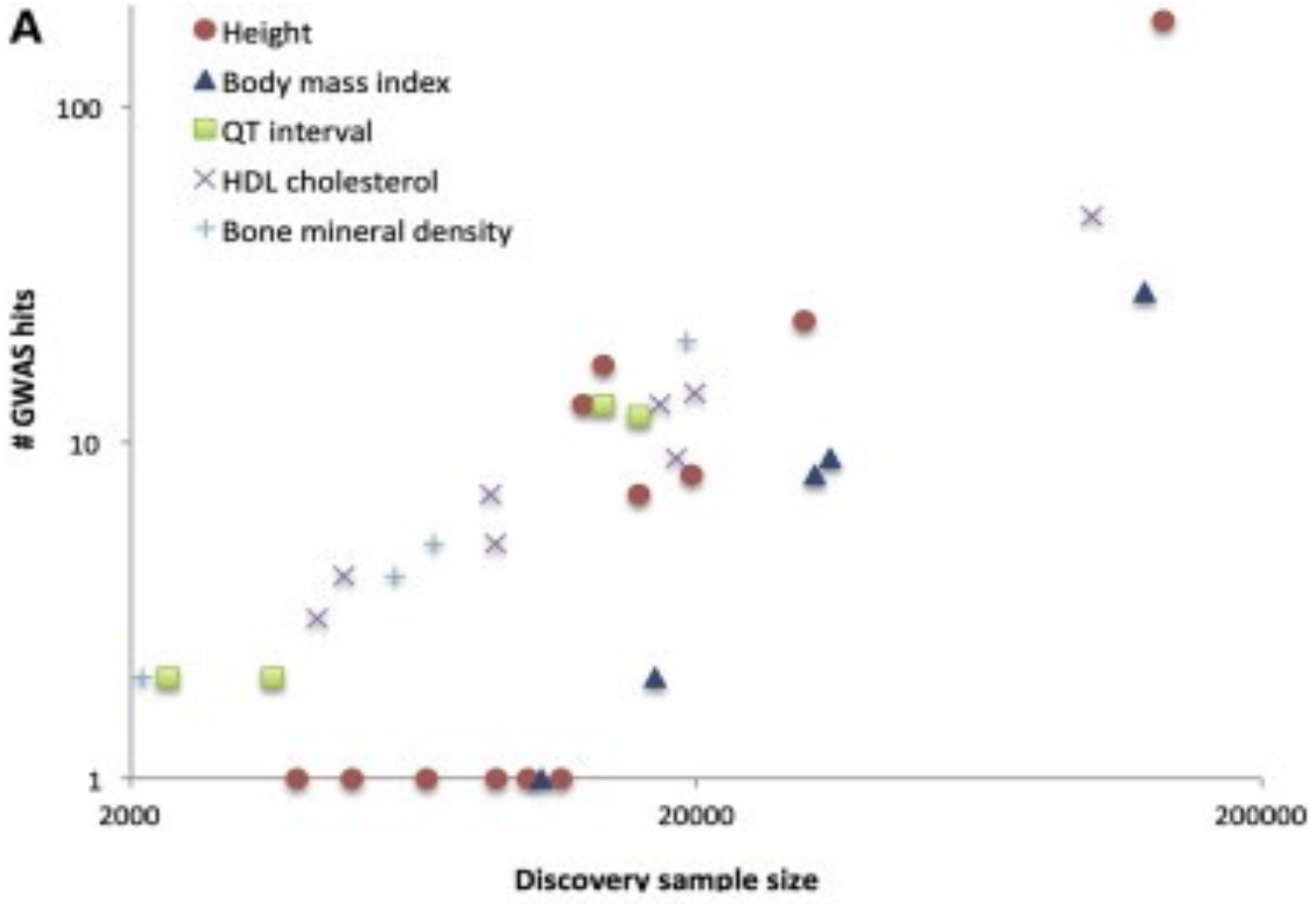
The exome array (~240,000 genetic variants)

- Design based on exome and whole-genome sequencing data from > 10,000 individuals (75% European ancestry)

MAF distribution of exome array data in the Women's Genomic Health Study (n=22,618 European Ancestry women. 58,000 variants were monomorphic)



Customized large-scale genotyping arrays

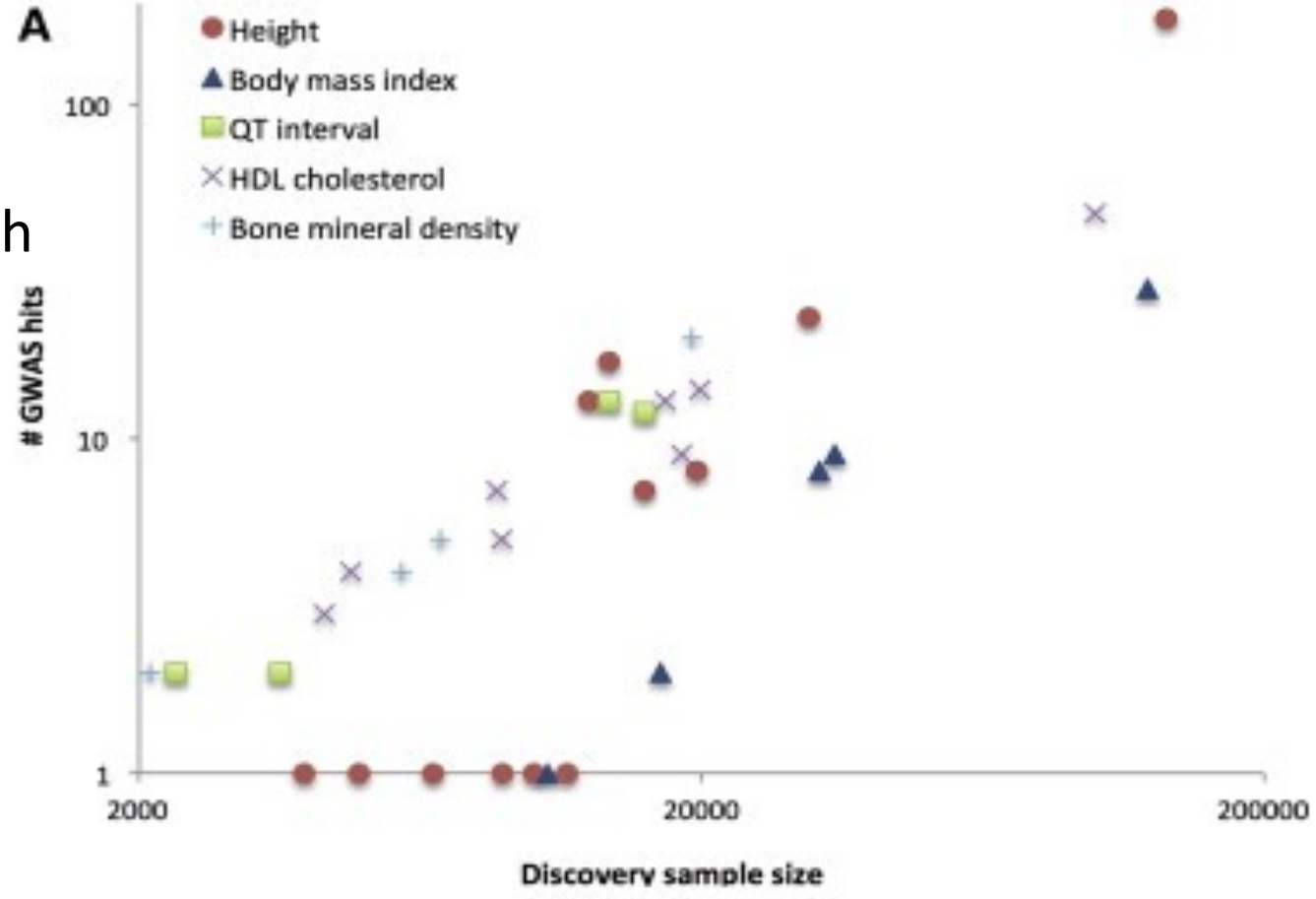


Visscher, Am J Hum Genet 2012

Customized large-scale genotyping arrays

- Idea: Can we design a custom array with 100,000s of SNPs and reduce the price if we commit to genotyping MANY subjects?

- Cost of these arrays are approximately 20% of GWAS arrays, thus enabling far more subjects to be genotyped. Genotyping using a uniform array has also enabled direct comparison across phenotypes.



Visscher, Am J Hum Genet 2012

Customized large-scale genotyping arrays

- **MetaboChip**
  - Custom array designed to test ~200,000 SNPs of interest for metabolic and cardiovascular disease traits.
  - Genotyped in > 100,000 subjects
- **ImmunoChip**
  - Custom array designed to test 195,806 SNPs for immune-mediated diseases.
  - Genotyped in > 150,000 subjects
- **Cardiochip**
  - Custom array that contains 50,000 SNPs across 2,000 genes associated with cardiovascular disease.
  - Genotyped in > 210,000 subjects
- **OncoArray**
  - Custom array designed to test ~500,000 SNPs related to multiple cancers: breast, colorectal, lung, ovary and prostate.
  - Genotyped in > 400,000 subjects

Combination arrays

- Emerged over the last few years
  - Includes both GWAS and exome array SNPs
  - Often allows for custom content
  - Target biobanks (e.g., UK Biobank)

Pricing (CIDR, March 2021)

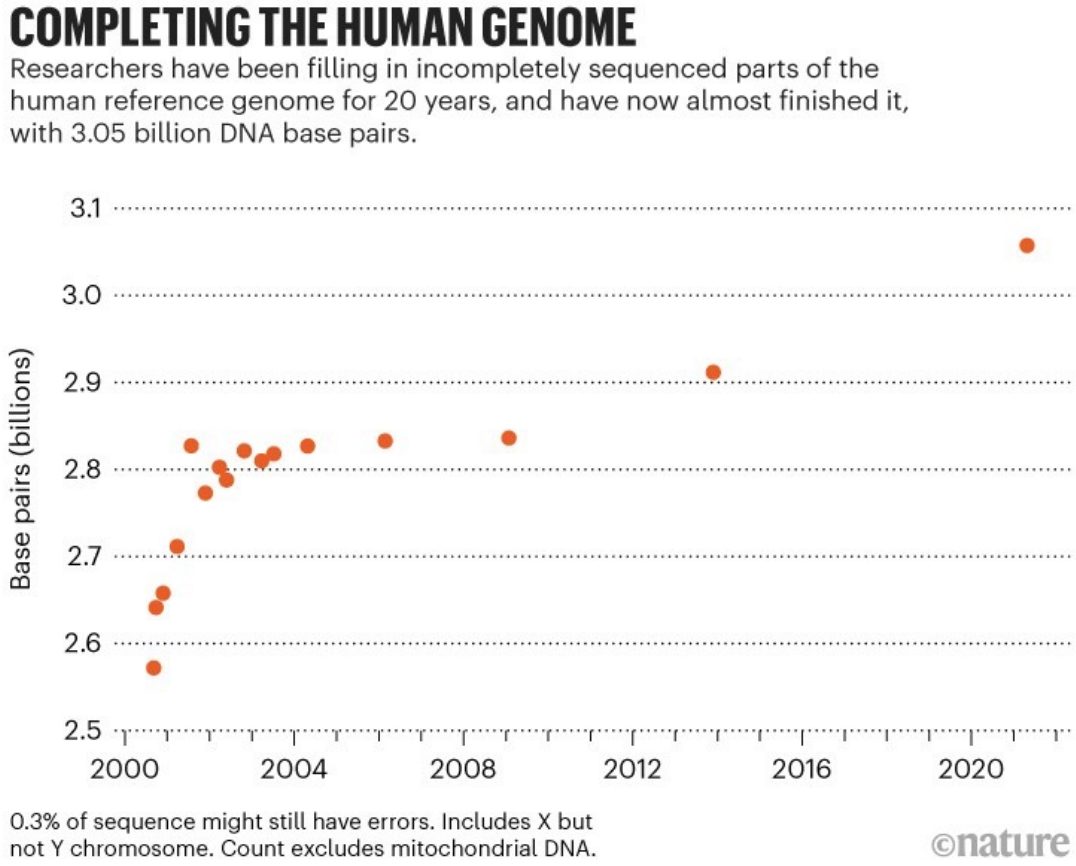
Affymetrix Genotyping - GWAS and Custom	
UK Biobank 821K Axiom Array	~$150 - $210 Inquire for pricing, sample number dependent
Custom Array (up to 750K SNPs)	~$180 - $240 Inquire for pricing
Custom Array (up to 50K SNPs)	~$120 - $170 Inquire for pricing

Illumina Genotyping – GWAS					
Global		**Screening**		**Consortium-Developed**	
Global Diversity Array or MEGA	$110-$130	Global or Asian	$75-$100	Oncoarray	$85-$110
Other Consortium Developed Arrays					
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.**					

https://www.cidr.jhmi.edu/services/pricing.pdf

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

$2.7 billion

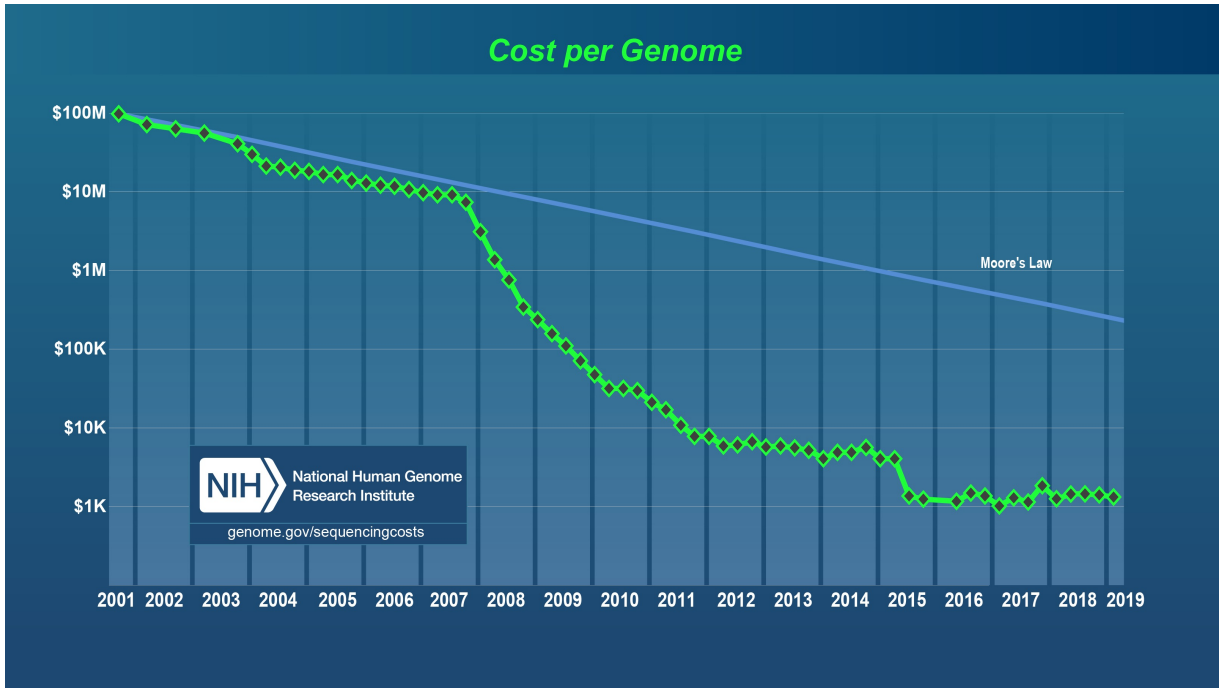


https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.full

Practical roadblocks to genome sequencing

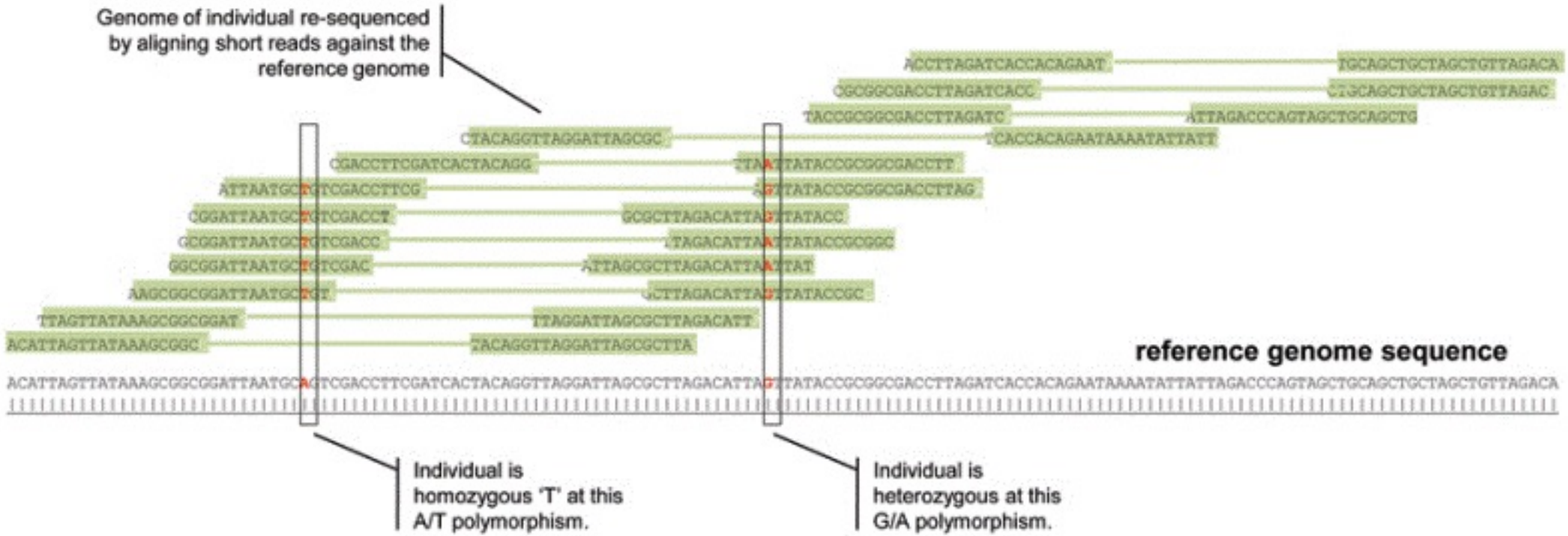
Sequencing cost per genome is currently ~$1,000

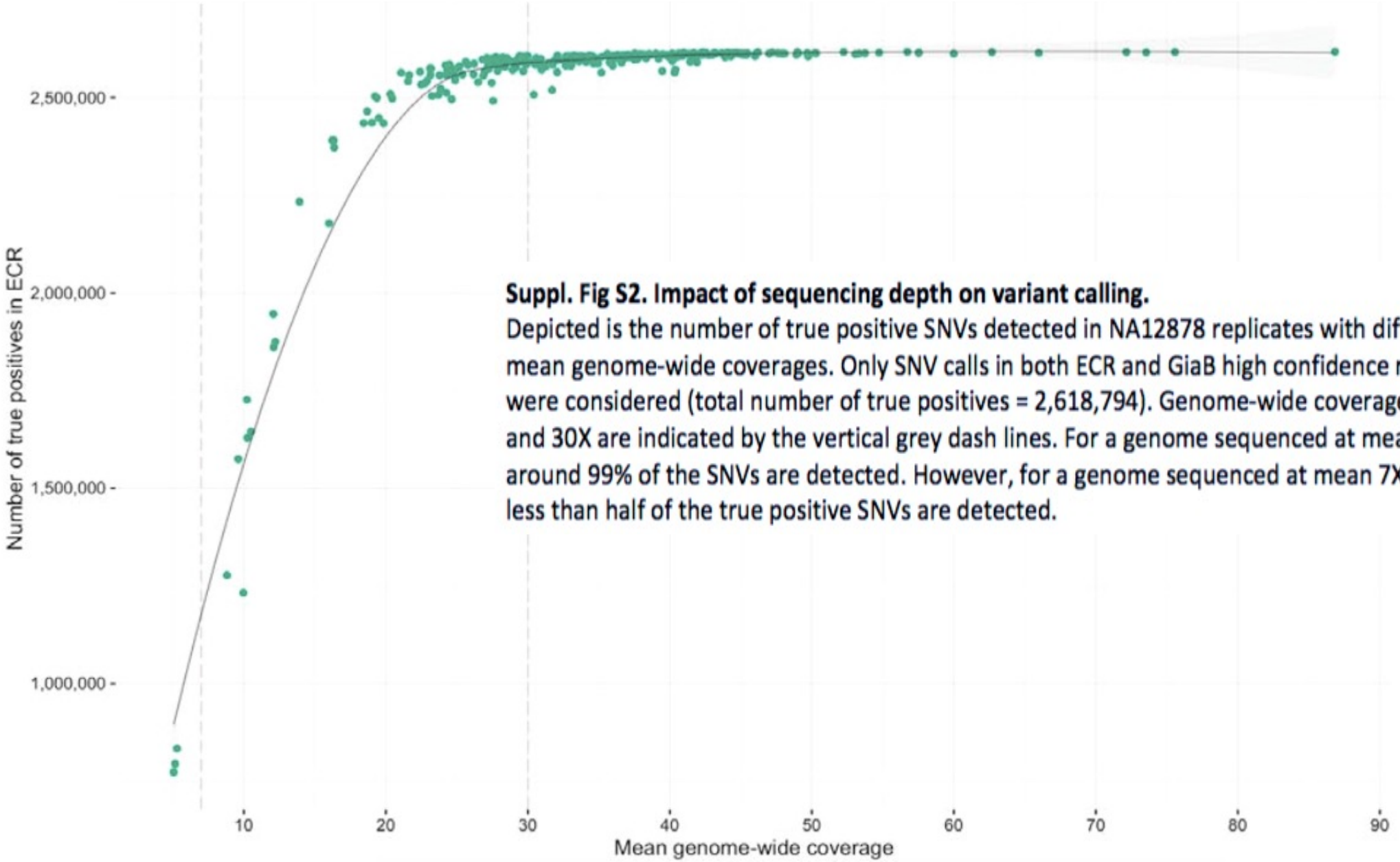
Sequencing one genome generates ~200 GB data



Sequencing alignment and depth

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





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.

Pricing Sequencing (CIDR, March 2021)

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 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://www.cidr.jhmi.edu/services/pricing.pdf