SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

Session 6:
Study designs for genetic association studies

or…How to assess genetic variation

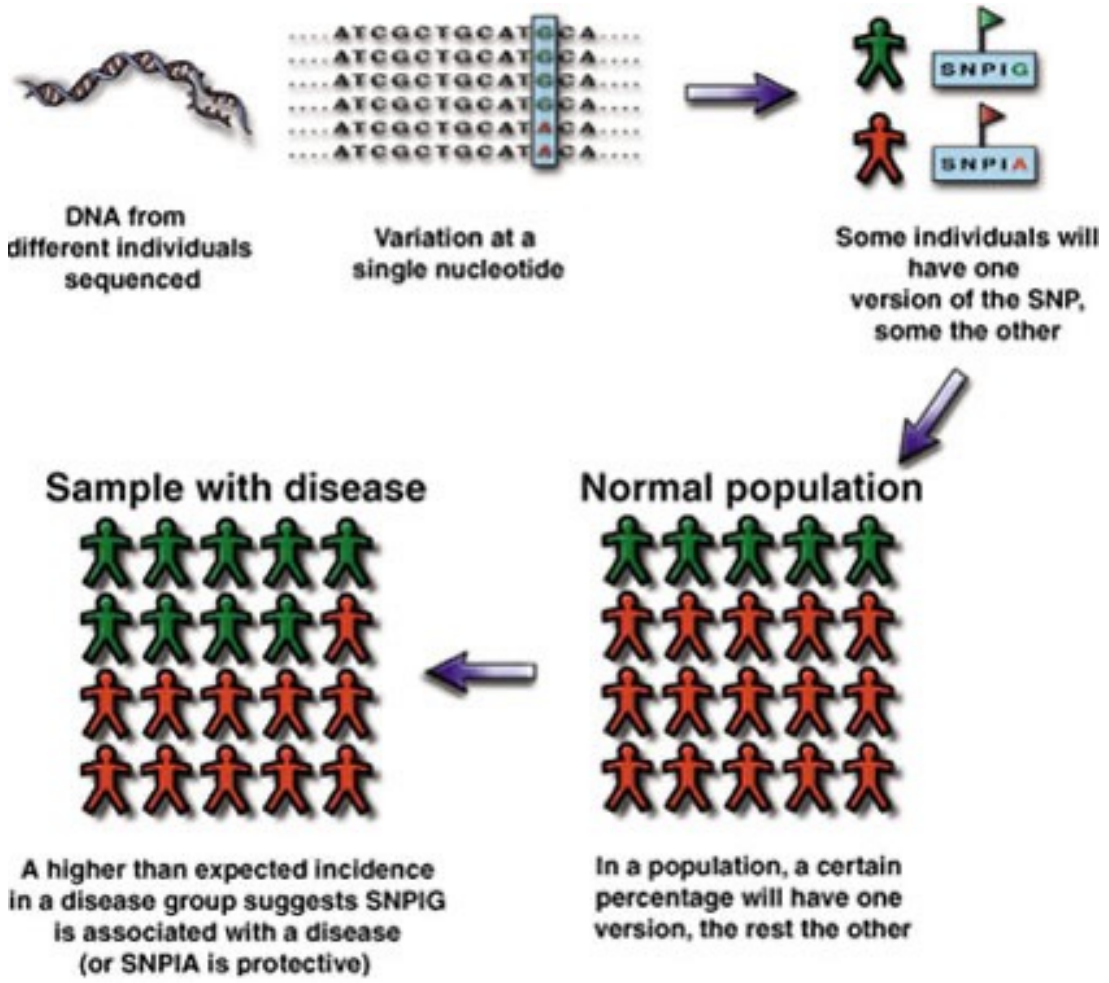
Genotyping vs. Sequencing

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)

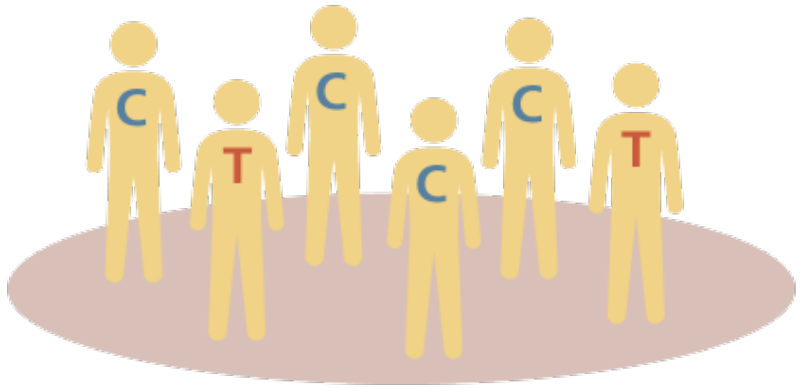
From a bioinformatic/analysis point of view, genotyping data is much easier to handle.

Genetic association studies using SNPs

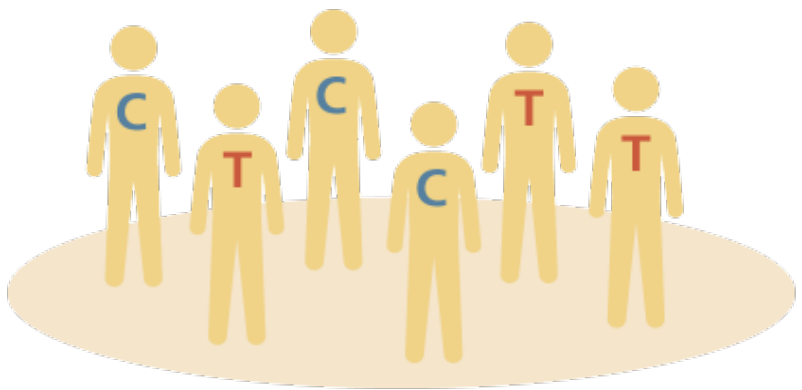


Why we like SNPs:

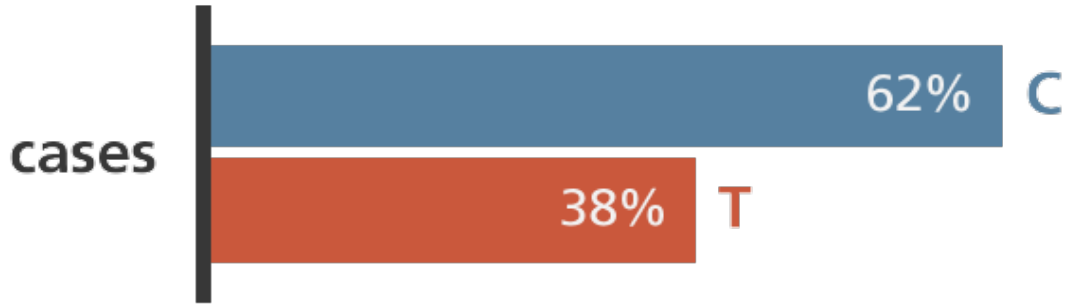
- Abundant in the genome
- Easy to measure



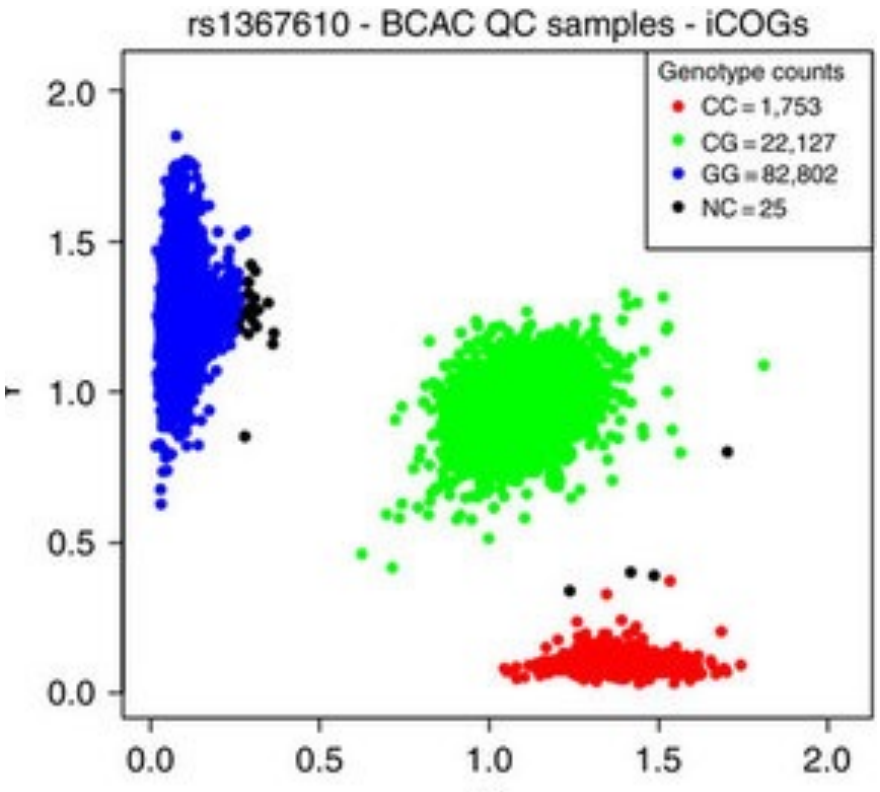
cases (n=1,000)
people with heart disease

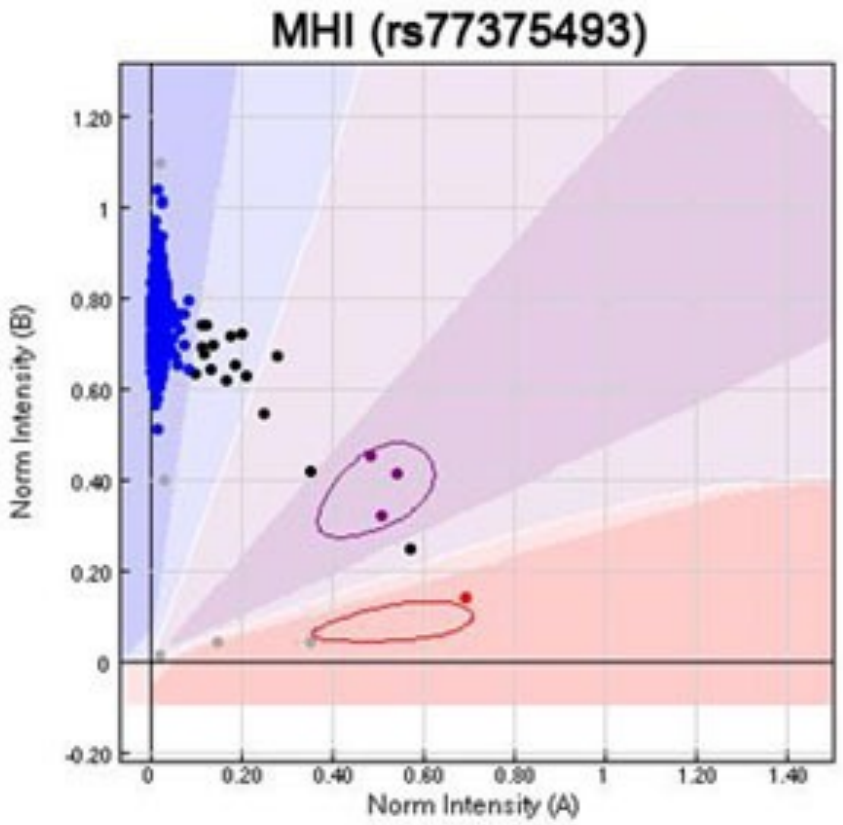


controls (n=1,000)
people without heart disease



Genotyping Output





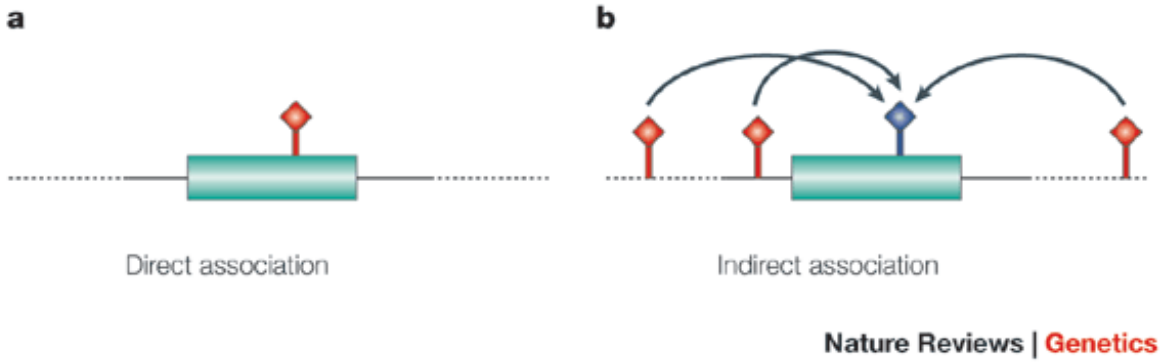
Li, Nat Comm 2014

Auer, Nat Genet 2014

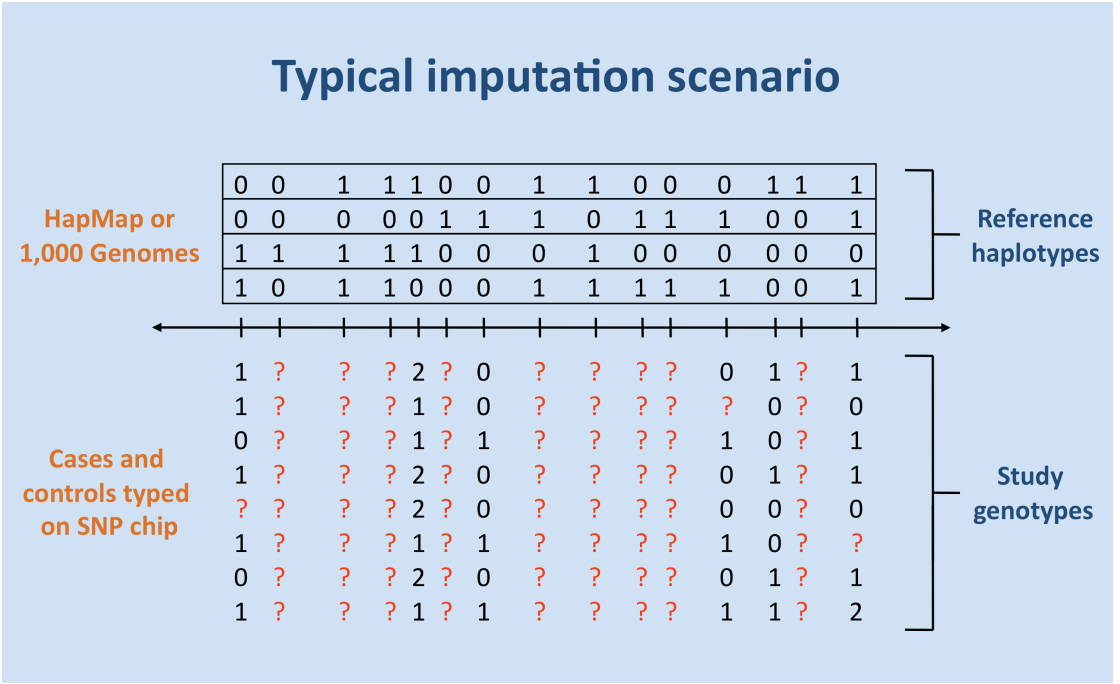
EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Genetic association studies rely heavily on LD

1) Indirect association



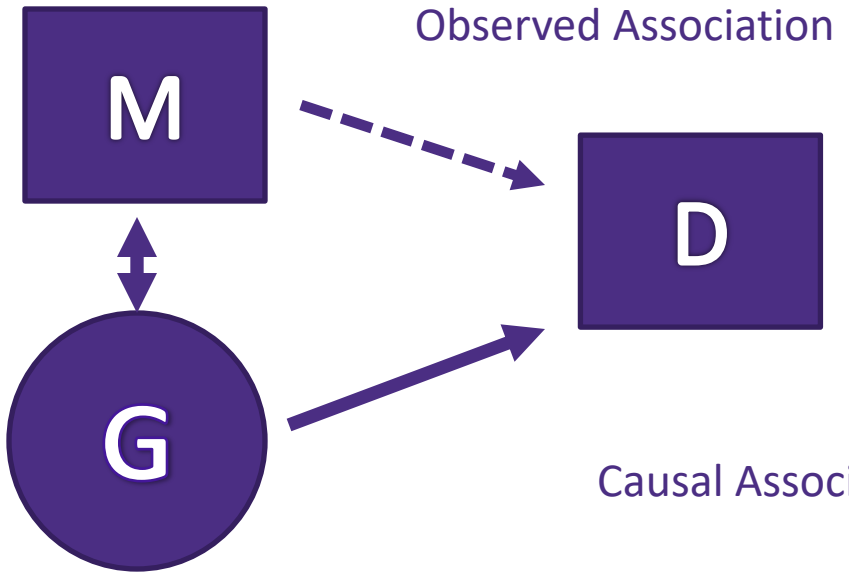
2) Imputation



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

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

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)

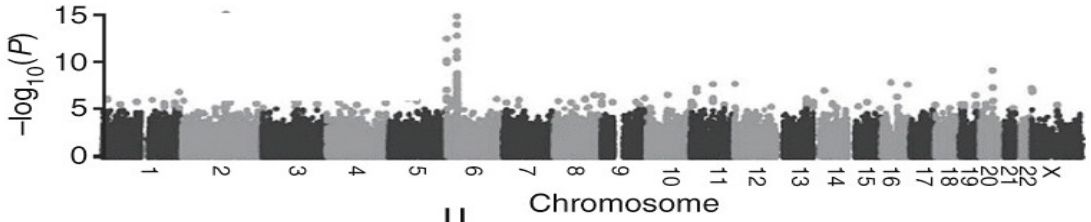
Screen the genome for SNPs that are associated with your trait (agnostic approach)

Rieder et al. 2008

Samples with phenotype data (continuous, case-control)
(n = 1000's)

Genotype samples with commercial 'chips'
- Affymetrix - Random SNP design (v.5, v.6)
- Illumina - TagSNPs plus candidate genes (650Y, 1M)

Perform statistical association with each SNP
Calculate p-value for each SNP



Region(s) with plausible statistical association

SNPs genotyped in GWA study ('chip SNPs')

Replication and Fine Mapping
- Associated chip SNPs for replication in additional populations
- If replicated more genotyping to fine-map or resequencing of the region is considered

Perform targeted resequencing to identify potentially causal SNP(s)
'Functional' analysis of potentially causal SNP(s) - *in vivo, in vitro*, gene expression…..

Imputation (I)

- Cost efficient and maximizes our sample size.
  - Can assess more SNPs than we genotyped
  - Fills in missing values for already genotyped SNPs

> 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

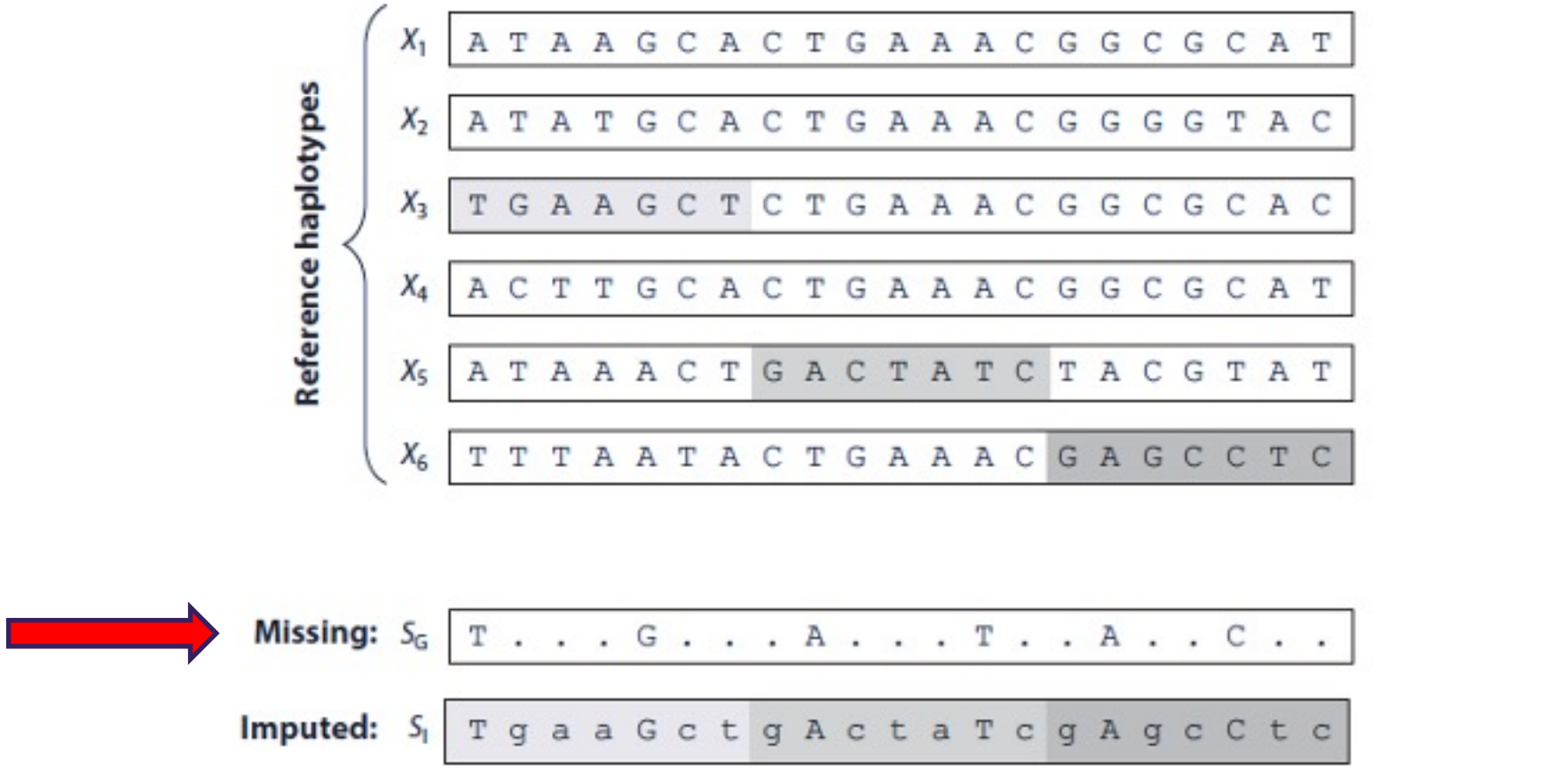


Figure 2

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 (II)

- 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].

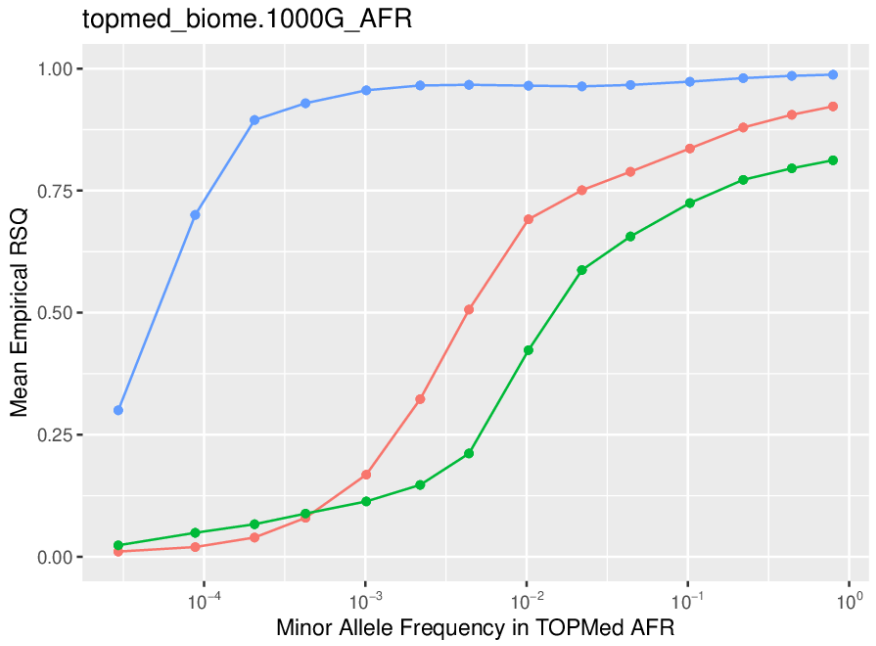
For more information about imputation, read Marchini and Howie, Nat Rev Genet 2010

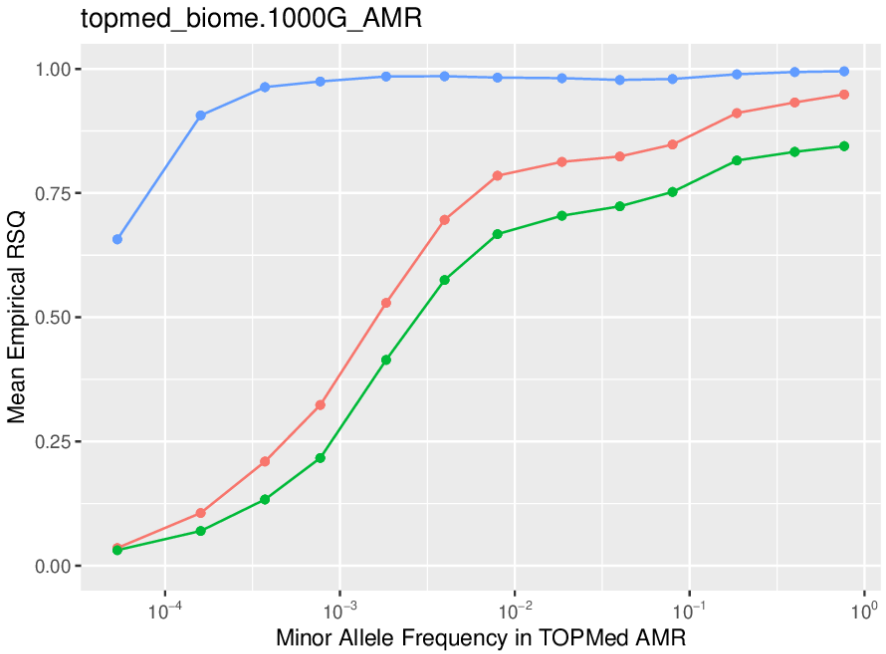
Imputation (III)

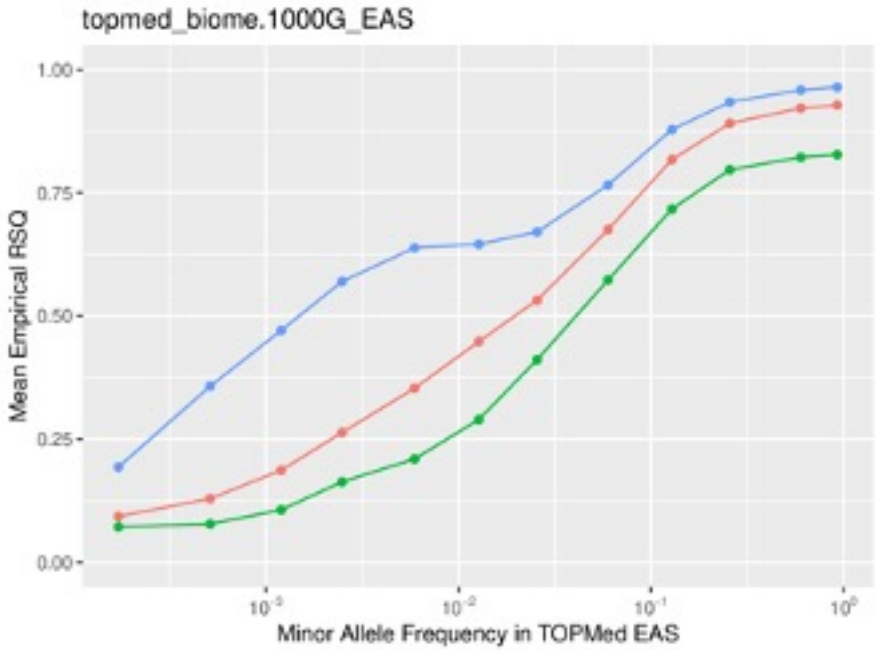
- The imputation quality score r^2 measures how well a SNP was imputed.
  - Ranges between 0 and 1.
  - Typically, a cut-off of 0.30 will flag most of the poorly imputed SNPs, but only a small number (<1%) of well imputed SNPs.

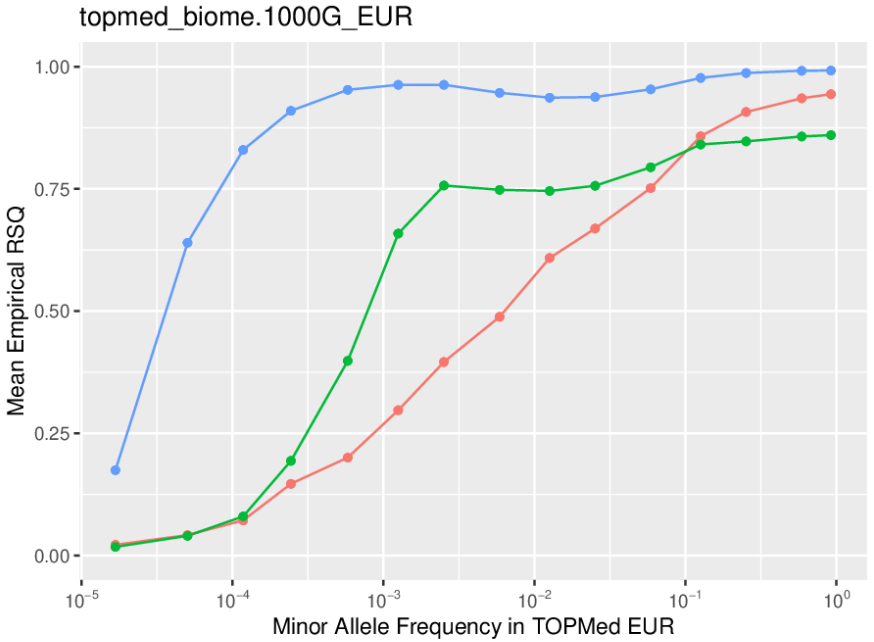
- Factors that affect imputation quality:
  - Number of genotyped SNPs in your data
  - Size of reference panel
  - Similarity in genetic ancestry between reference and study samples
  - Allele frequency

Reference Panels	N	Ancestry
HapMap	60	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



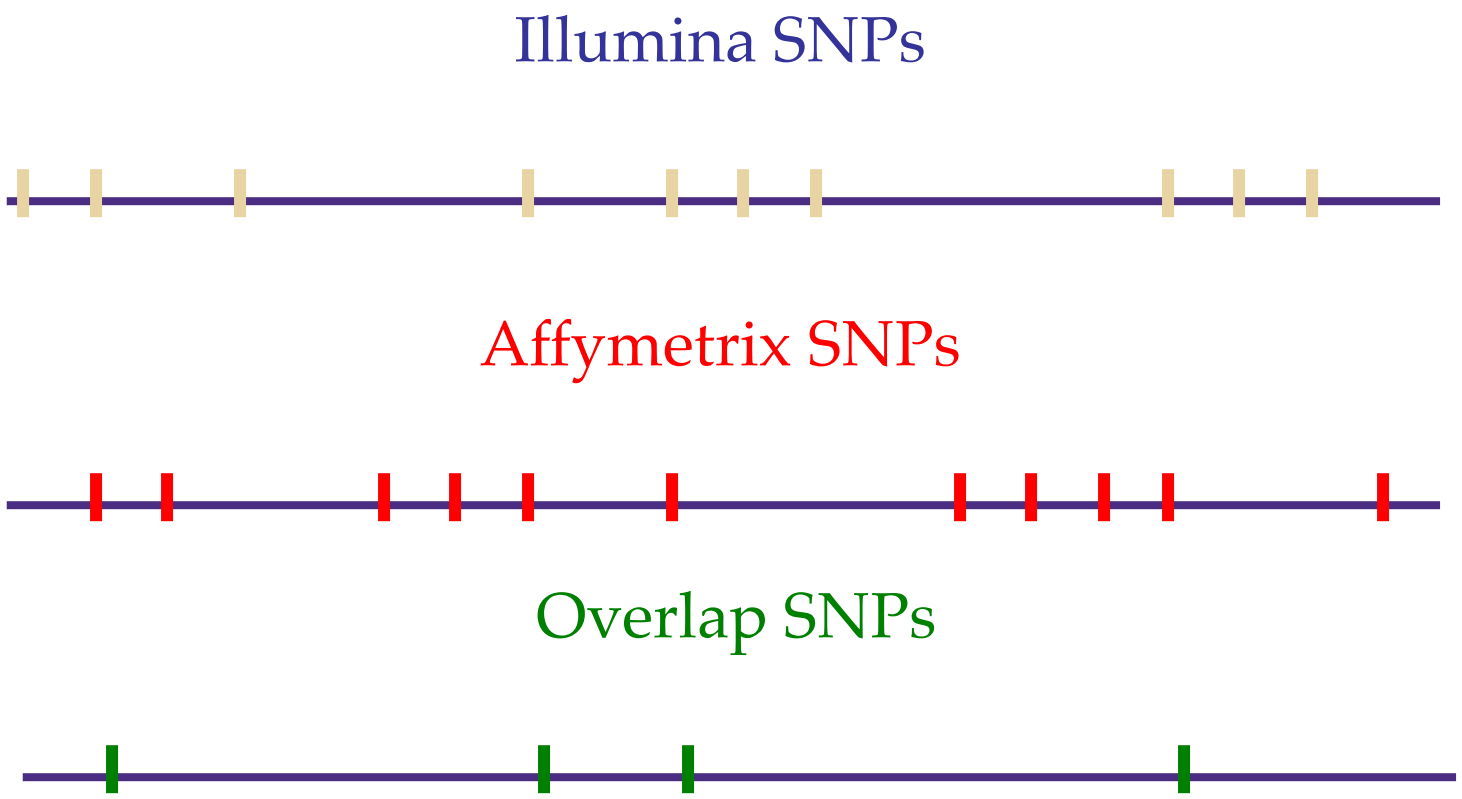






Panel 1000G HRC TOPMed

Imputation for studying SNPs across platforms



Imputation for studying SNPs across platforms

TopMed SNPs

Illumina SNPs

Affymetrix SNPs

Overlap SNPs

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

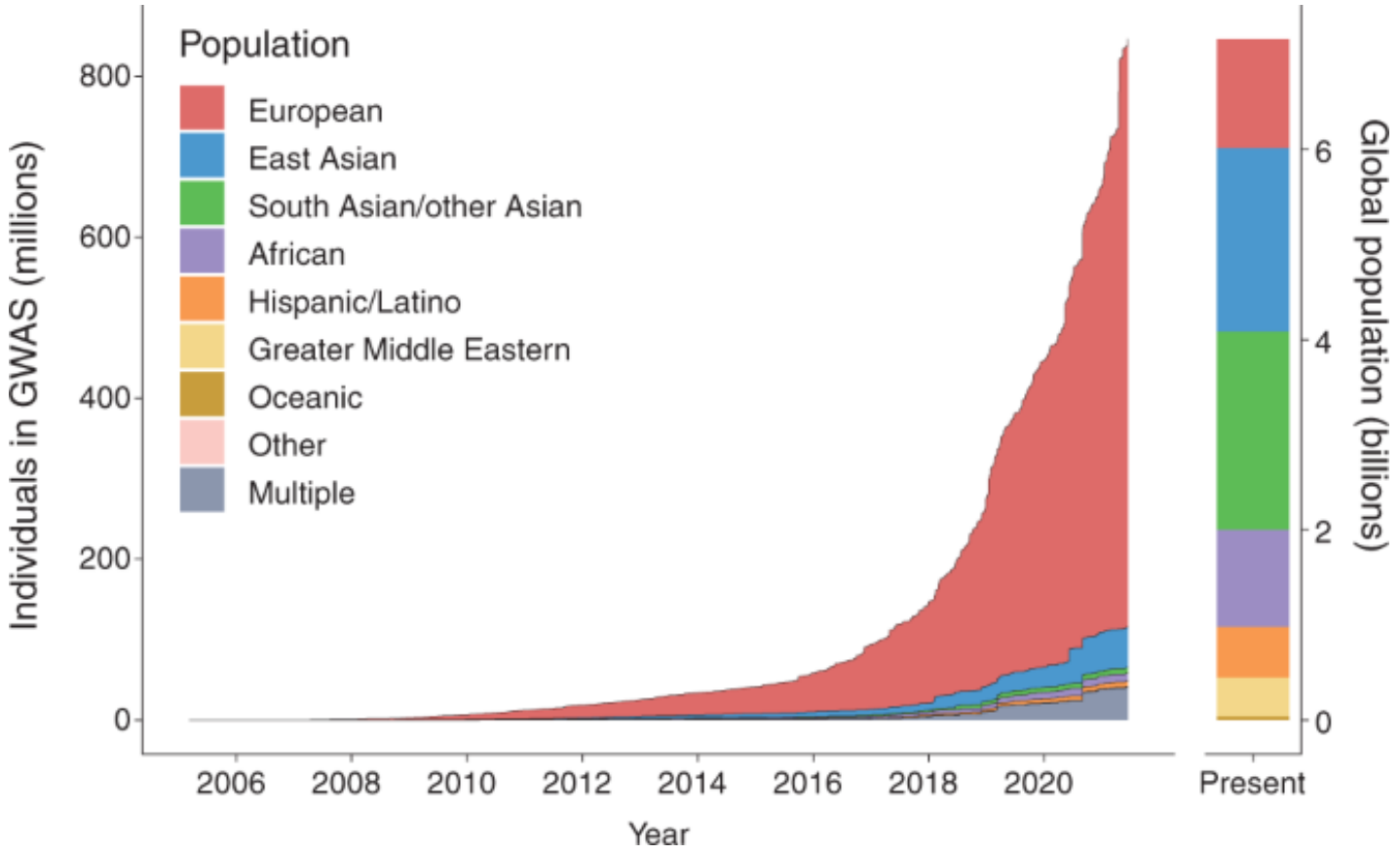
Limitations with traditional genome-wide genotyping arrays

> Genome-wide genotyping arrays have traditionally been designed to capture genetic variation in populations of European ancestry

> Only capture common SNPs

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

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?

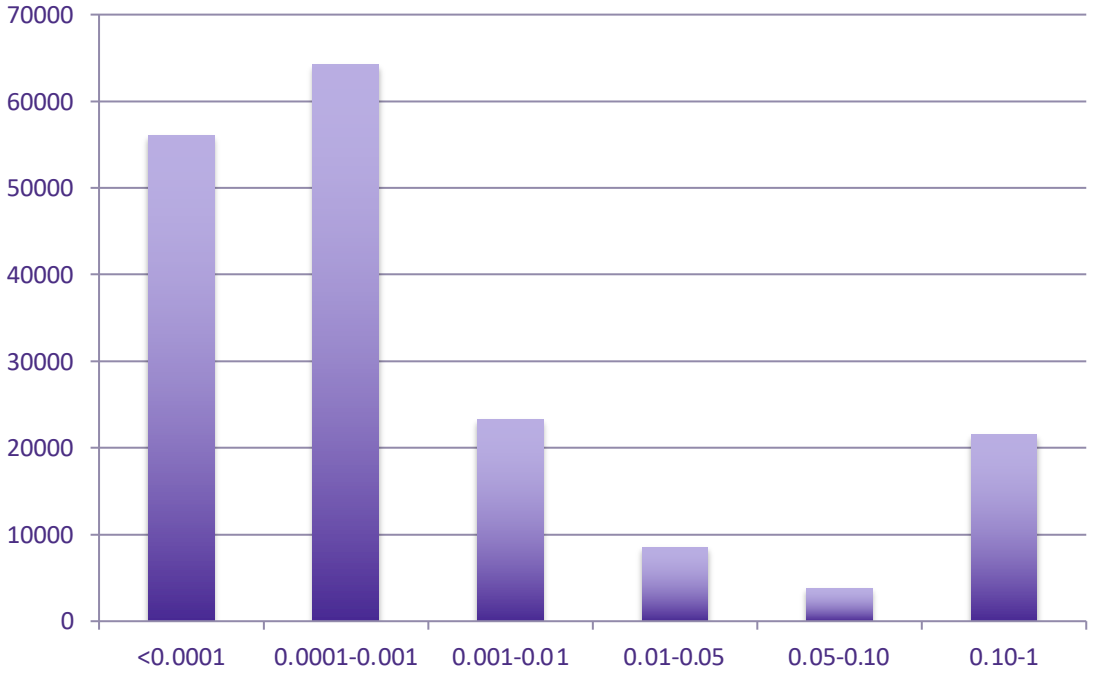
The Multi-Ethnic Genotyping Array (MEGA) – 1.8M markers

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

The exome array (~240,000 genetic variants)

- Design based on exome and whole-genome sequencing data from across the world (at the time mostly unpublished data)
  - 9000 samples of European ancestry, 2000 samples of African ancestry, 500 samples each of Hispanic and Asian ancestry



Minor allele frequency distribution of exome array data in the Women's Genomic Health Study (n=22,618, European ancestry)

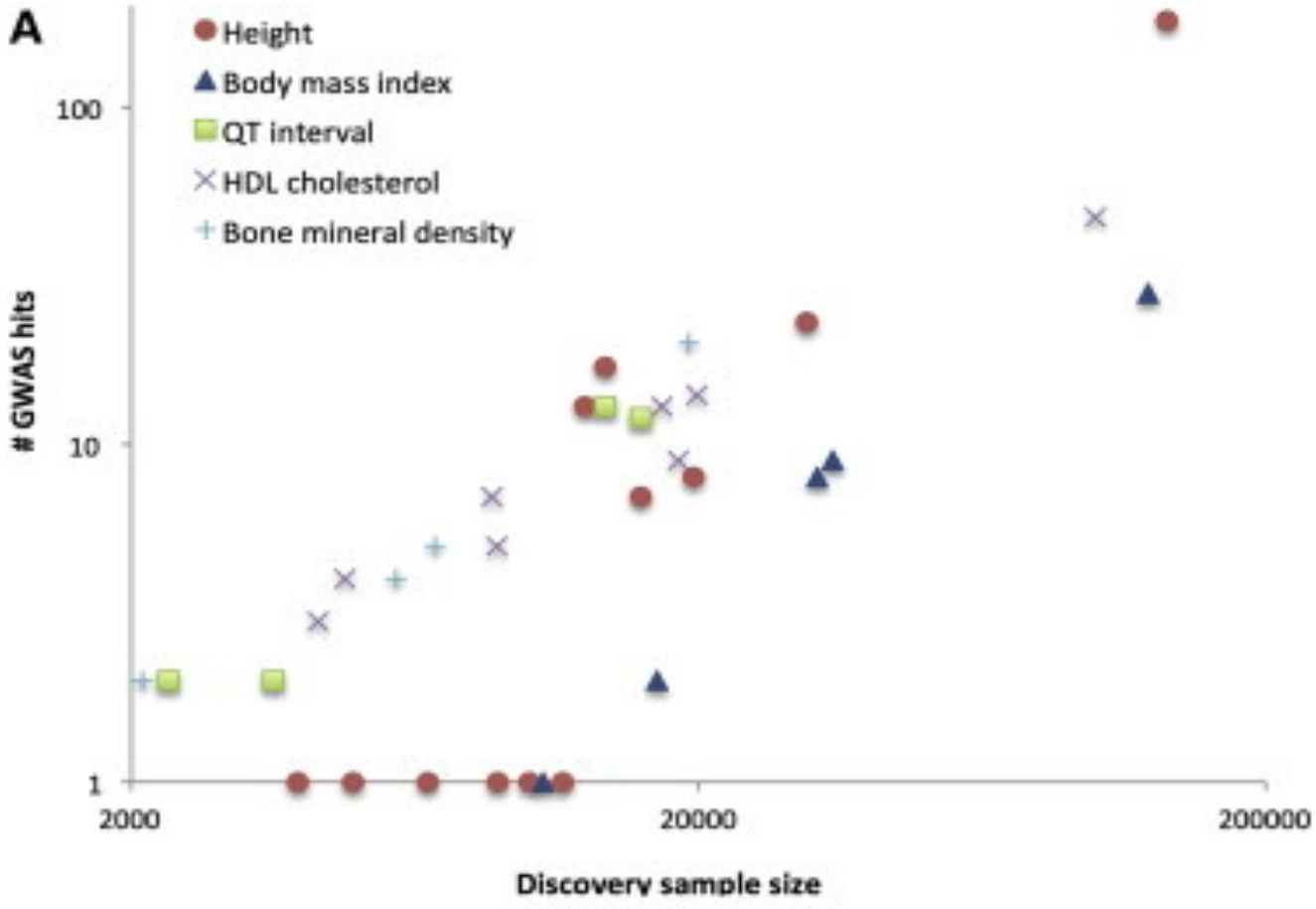
~58,000 variants were monomorphic

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Customized large-scale genotyping arrays

- Can we design a custom array with 100,000s of SNPs and reduce the price per sample 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**
  - Includes both GWAS and exome array SNPs and also allows for custom content. Target biobanks (e.g., UK Biobank)

Pricing (CIDR, March 2022)

Illumina Genotyping – GWAS					
Global		**Screening**		**Consortium-Developed**	
Global Diversity Array	$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.**					

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

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

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)
- Identify variants that might be unique to your subjects (i.e., breast cancer cases)
- More expensive and requires more IT 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

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

SCIENCE
VOLUME 376|ISSUE 6588|1 APR 2022

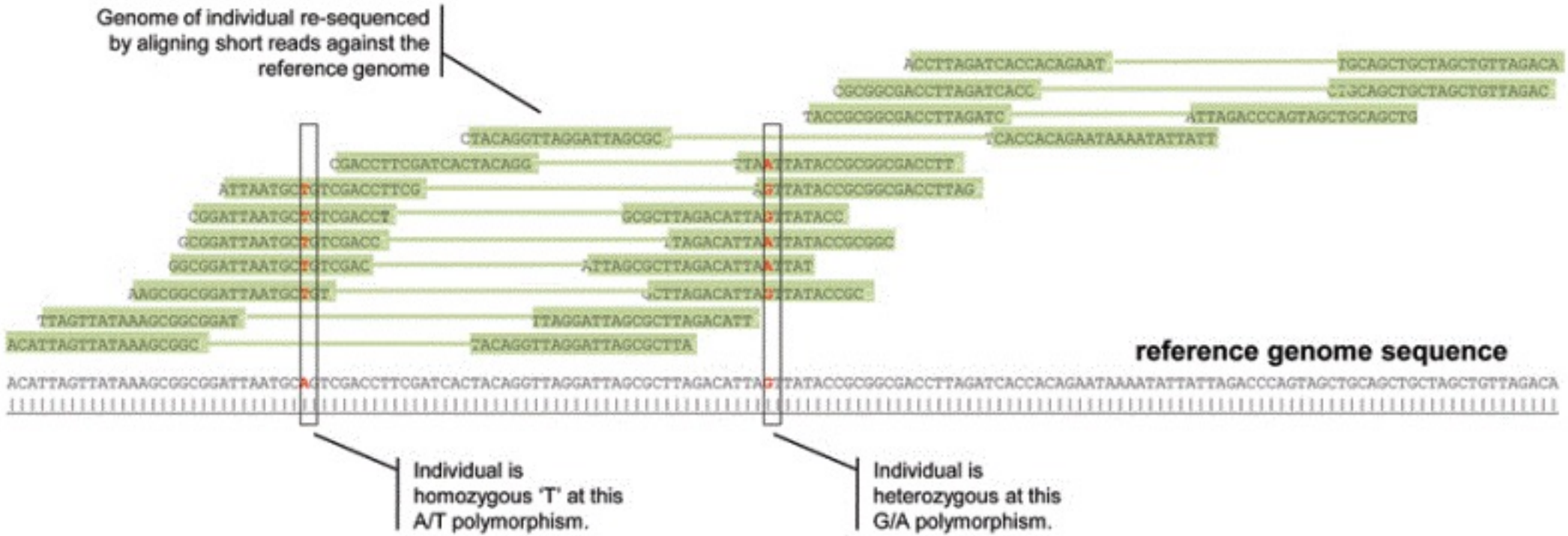
COVER

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.

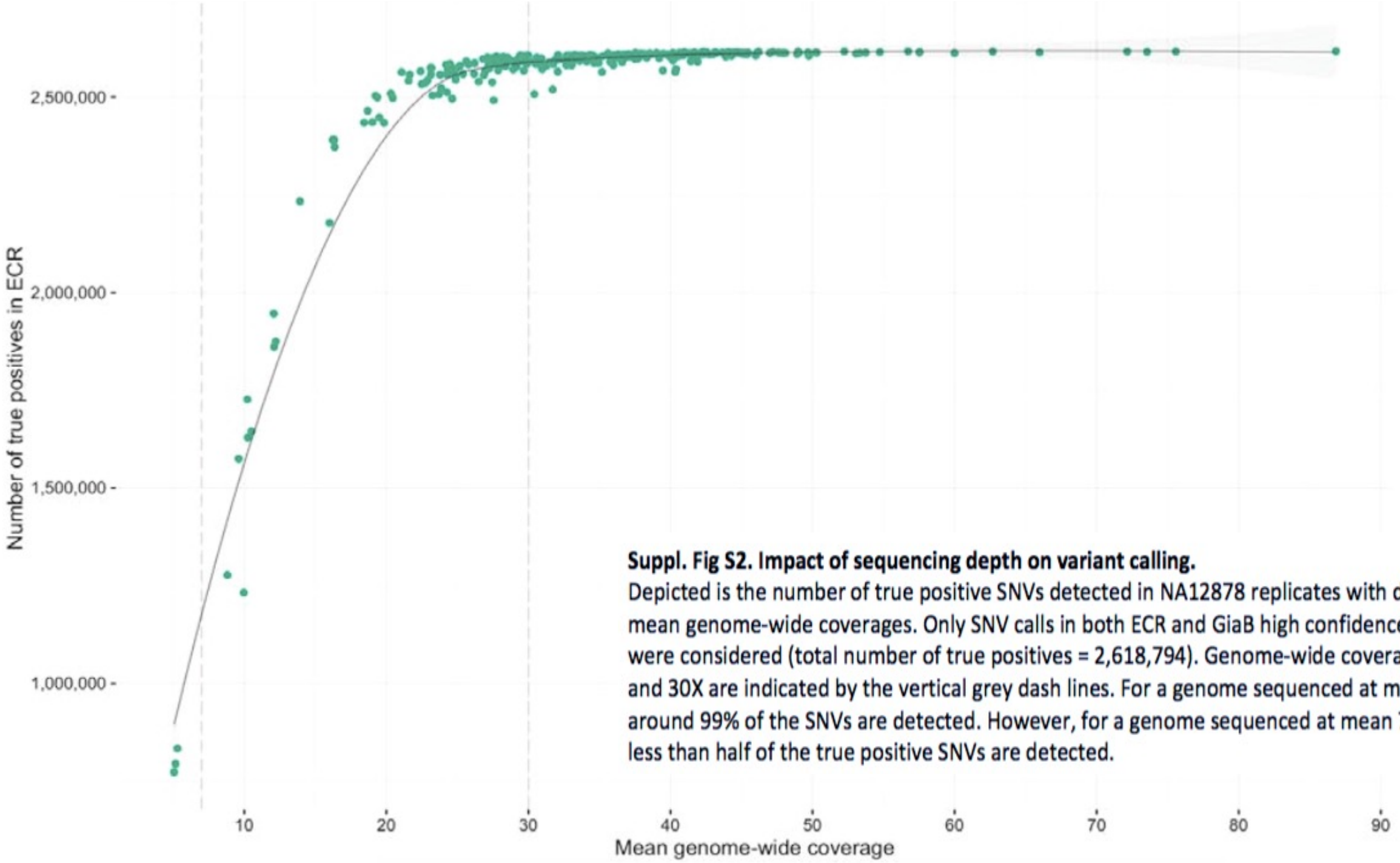
EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Sequencing alignment and depth

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



EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

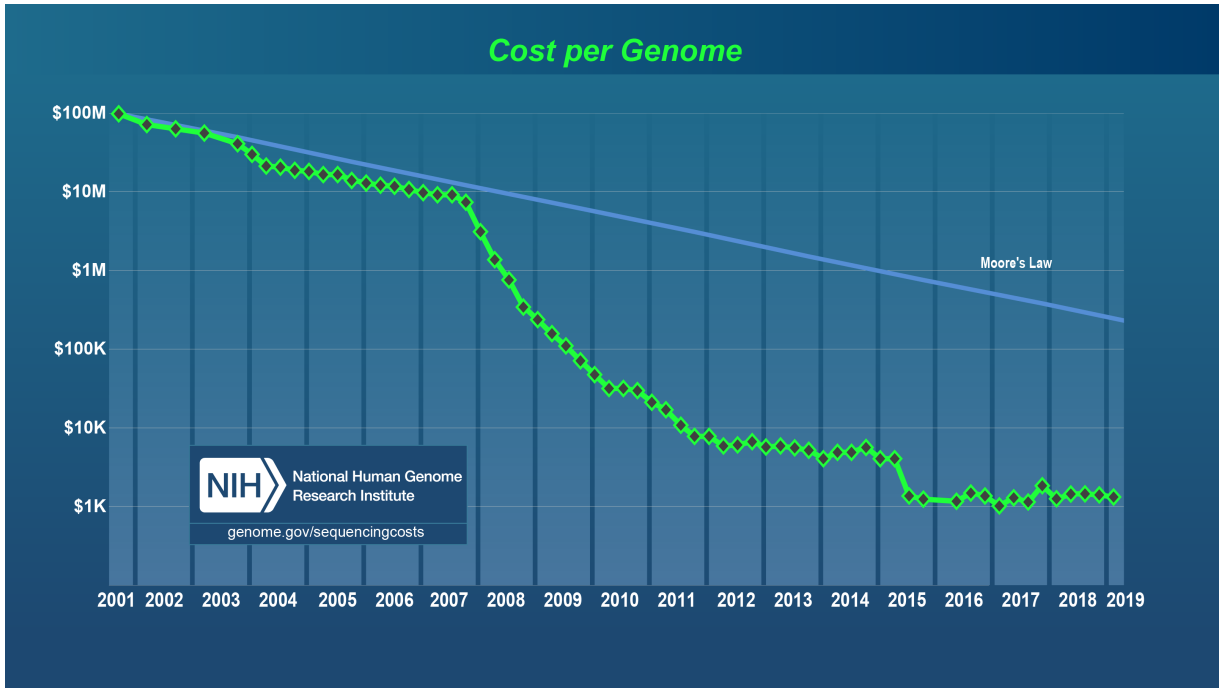


Suppl. Fig S2. Impact of sequencing depth on variant calling.
Depicted is the number of true positive SNVs detected in NA12878 replicates with different mean genome-wide coverages. Only SNV calls in both ECR and GiaB high confidence regions were considered (total number of true positives = 2,618,794). 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 2022)

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