# Genome-wide Association Studies

BIOS 770

02/12/2025

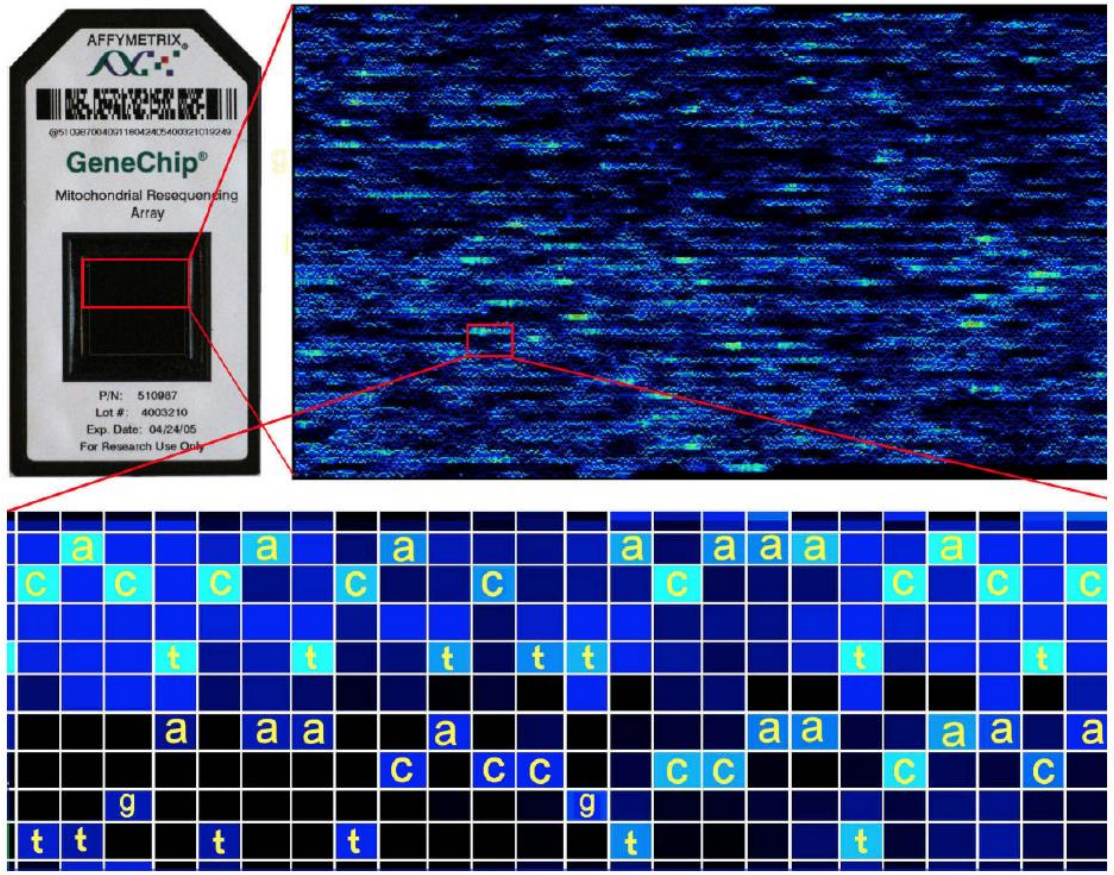
Jingjing Yang (jingjing.yang@emory.edu)

# Outline

- Genotype Calling and Imputation
- Single Variant Test (common variants with MAF > 1%)
  - Dichotomous Trait
    - Logistic regression model based test
  - Quantitative Trait
    - Linear regression model based test
- Visualization GWAS Results by Manhattan plot and LocusZoom plot

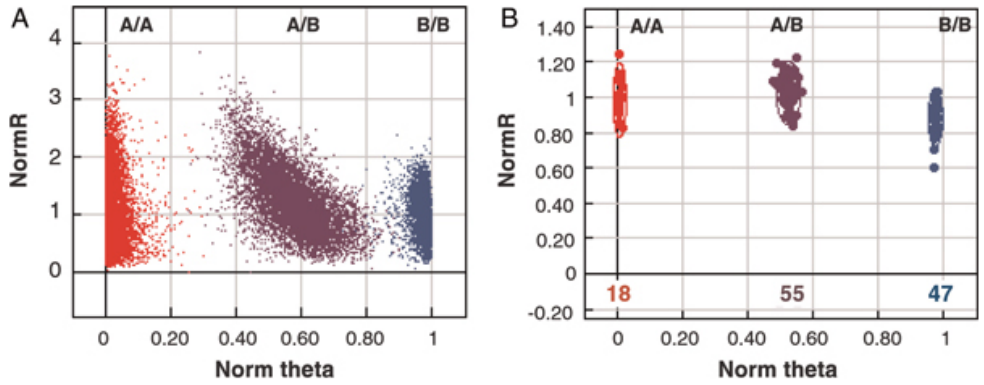
# DNA Microarrays

- Old generation of technology (Affymetrix or Illumina)
- Still widely used for cheap cost, ~$200 / sample
- Can be customized with densely spaced SNPs for target genes
- Genotype 0.5 Million ~ 1 Million SNPs

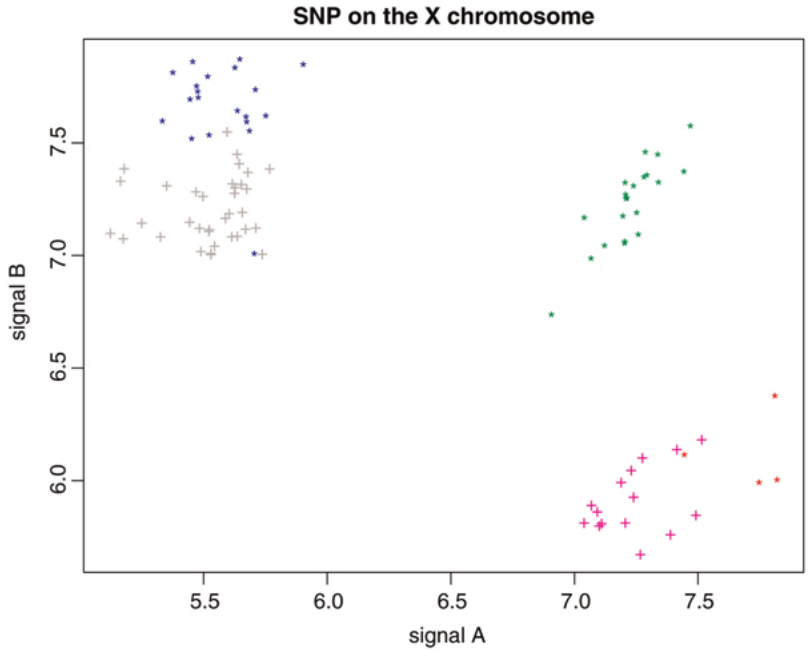


# Variant Calling for Microarray Genotyping

- Each SNP has three possible genotypes: AA, BB, AB with two alleles A and B
- Summarize the probe intensities for each allele and SNP, and then make a call based on the summarized intensities




**Normalised and summarised allele intensities from the Illumina BeadChip array**. The intensities are shown in transformed polar coordinates: the theta-coordinate represents the angle from the x-axis (the angle from the x-axis to the vector [A, B] of the two allele intensities), and the R-coordinate represents the copy number (the length of the vector). (A) Intensities for a single nucleotide polymorphism (SNP) from 120 arrays, clearly separating the intensities into three groups (A/A, A/B, B/B). (B) Data from 317,000 SNPs (from the same 120 arrays). This plot clearly indicates that signal strength varies considerably with the SNP, a factor that must be taken into account when genotyping individual SNPs and deriving copy numbers. The figure is reproduced with the permission of Gunderson *et al*. [15]

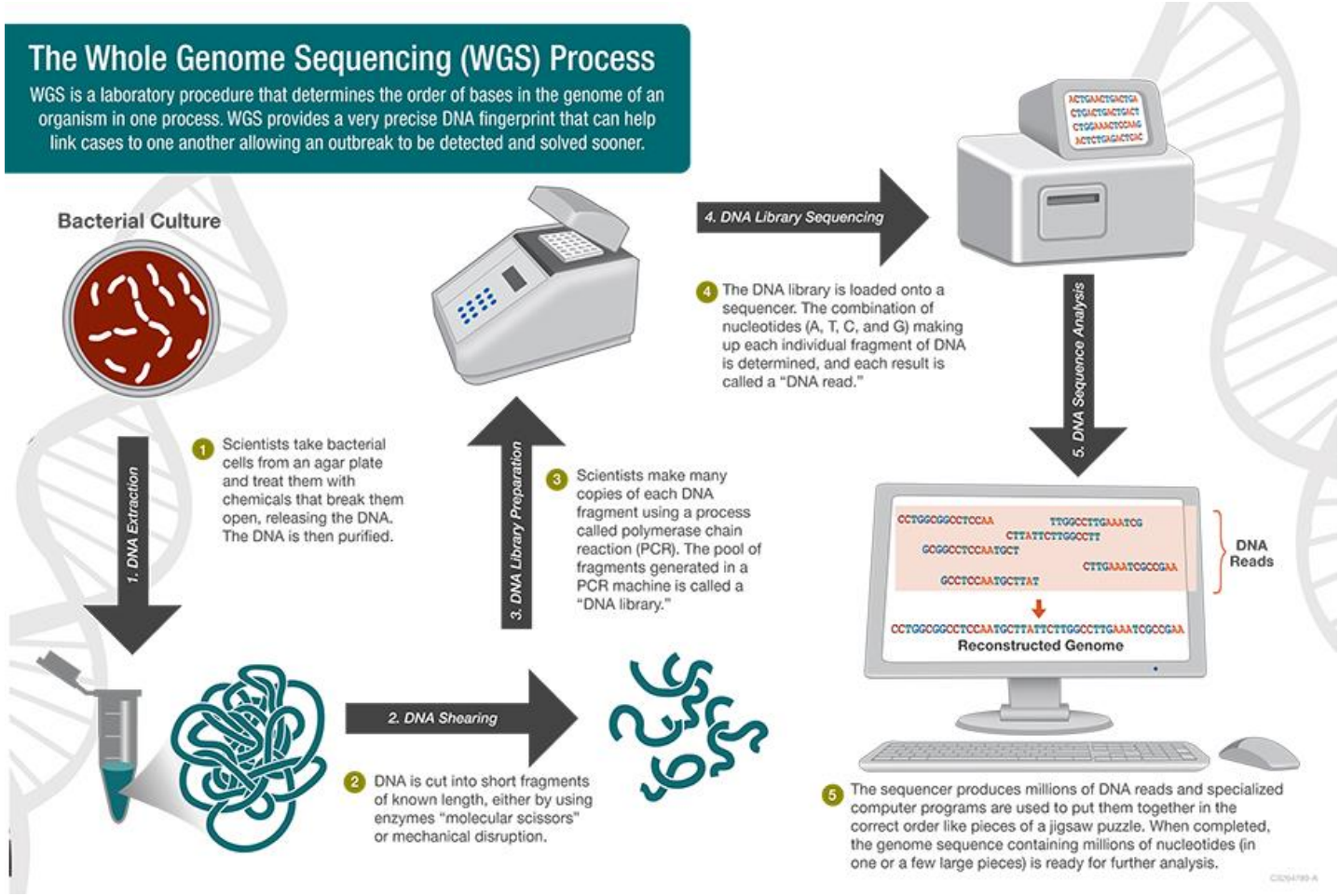



**Normalised and summarised allele intensities from the Affymetrix GeneChip array**. Each SNP is represented by a pair of intensity values (A, B) for the A and B alleles, respectively (here, on a log-scale). An X chromosome SNP is shown, clearly indicating separation into distinct genotype clusters. The plot also shows that different copy numbers can be distinguished. Males are haploid for the particular SNP (ie either AY or BY) and show up as homozygous but with reduced allele intensity. Grey: BY; blue: BB; green: AB; red: AA; and pink: AY.

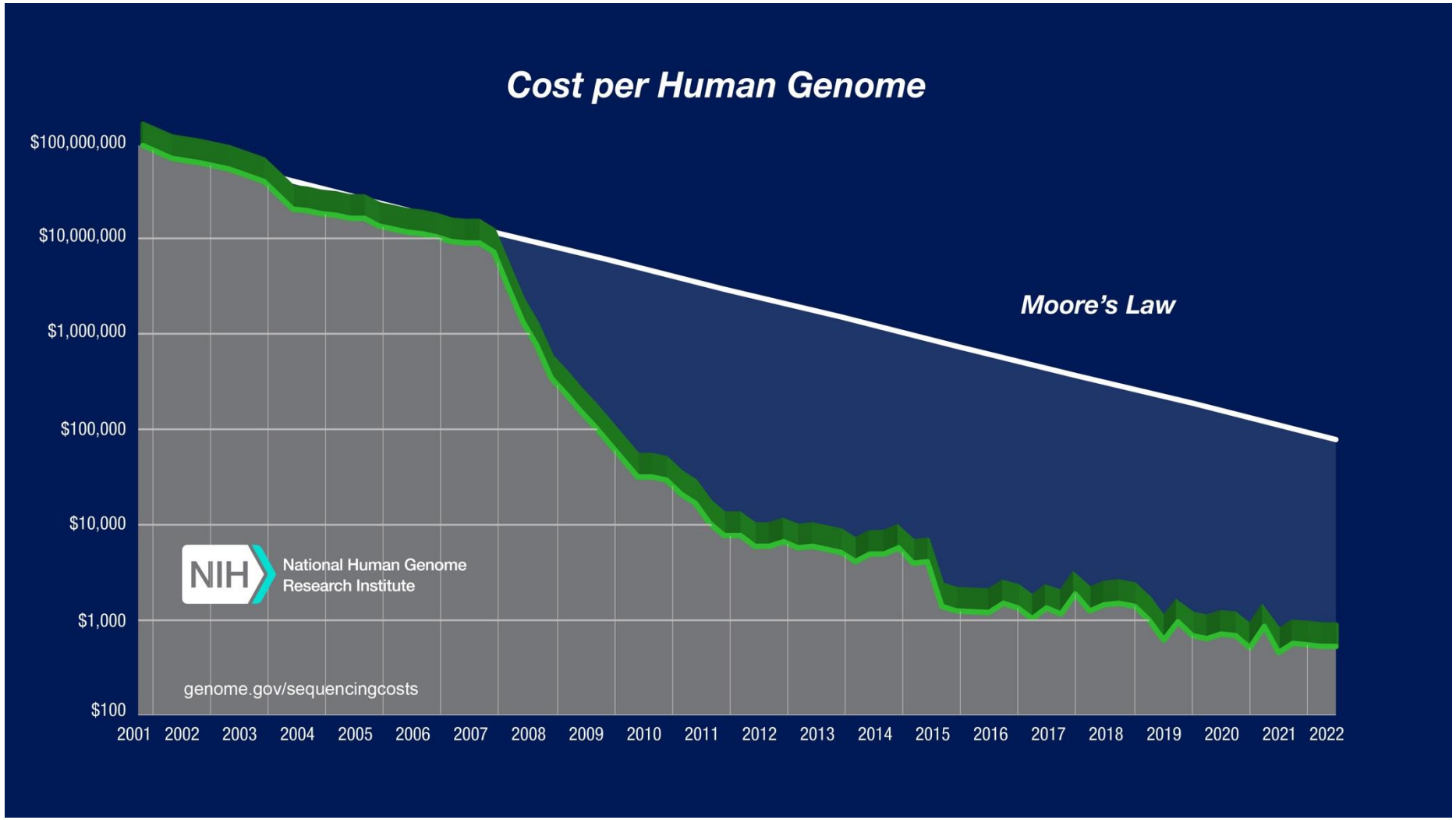
Lamy P. et. al. Human Genomics, 2011.

# Whole Genome Sequencing (WGS)

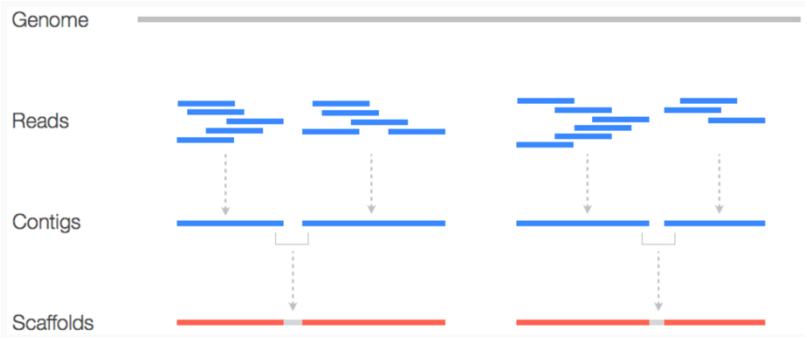
- Next-generation sequencing technology
- Introduction video of Illumina Sequencing technology: https://www.youtube.com/watch?v=fCd6B5HRaZ8
- Profile >10M SNPs



# Whole Genome Sequencing Now Costs ~$600/sample



# WGS Analysis Workflow




Raw read quality control — FastQC

Data preprocessing — Cutadapt: Read trimming

Bowtie2 — Alignment

Variant calling — GATK :
- Multiple-sequence realignment
- Quality score recalibration
- SNP genotyping
- Indel discovery and genotyping

Genome assembly — SPAdes Genome Assembler: For Illumina sequencing. Align overlapping reads to form longer contigs (larger contiguous sequences) and order the contigs into scaffolds (a framework of the sequenced genome).

Genome annotation — VEP: Annotate biological functions of called variants.

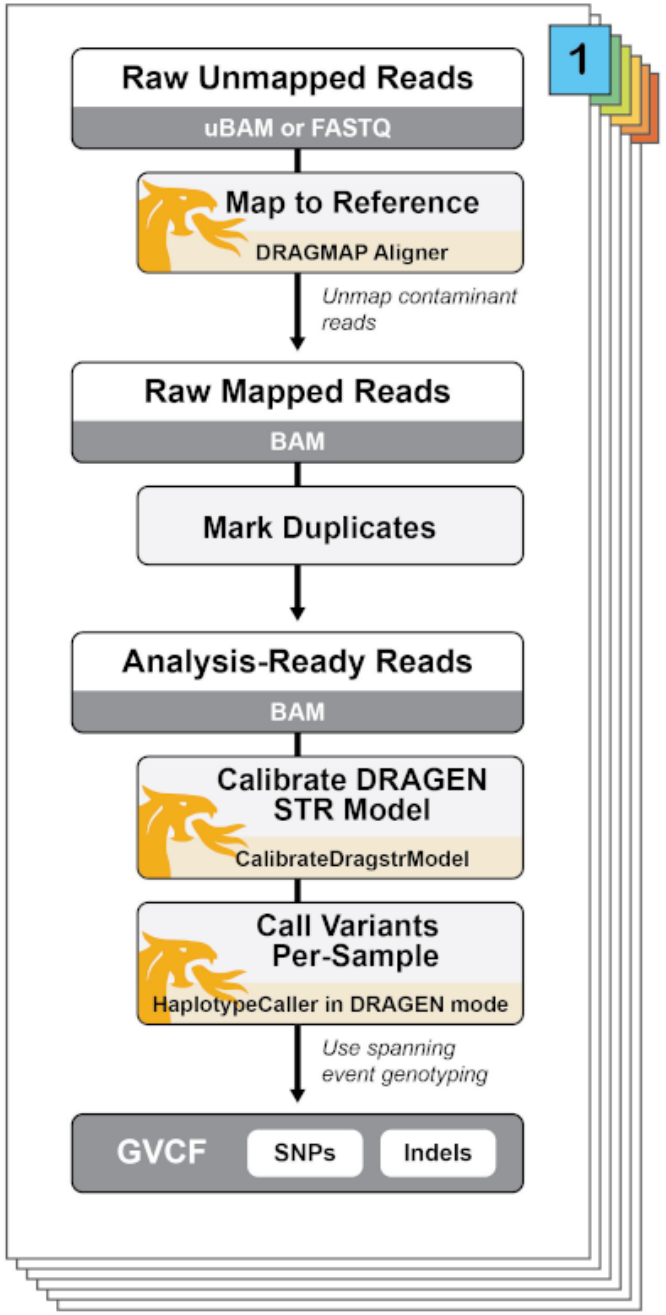
Other advanced analyses

Figure 1. Bioinformatics workflow of whole genome sequencing.

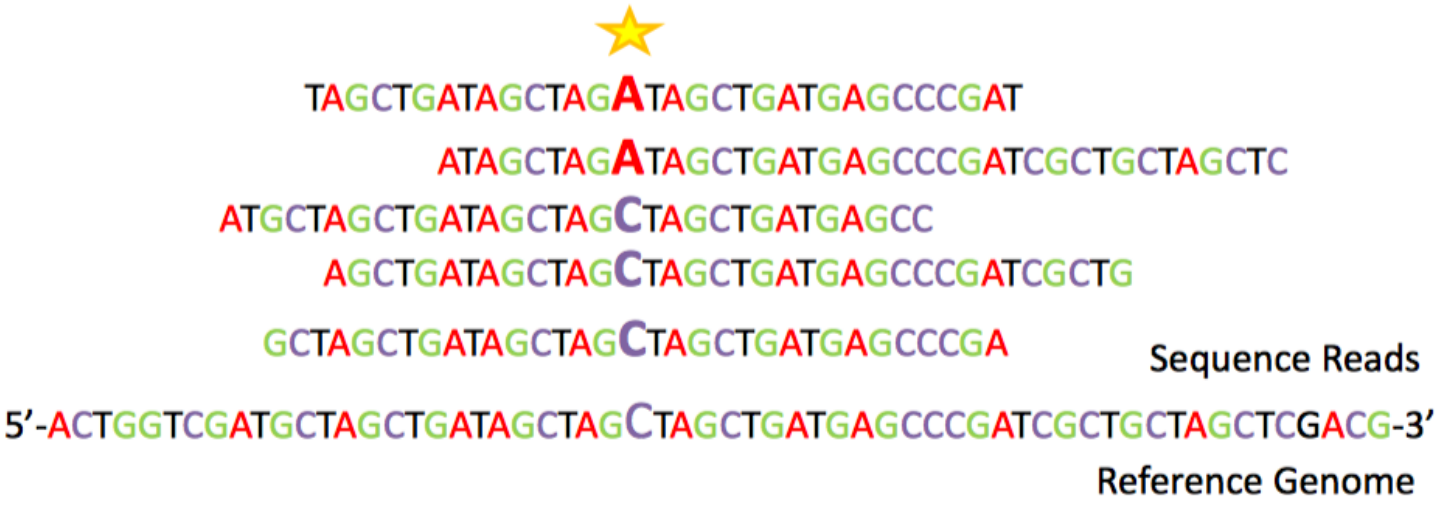
https://www.cd-genomics.com/bioinformatics-workflow-for-whole-genome-sequencing.html

# DRAGEN-GATK Whole Genome Germline Pipeline for Variant Discovery

- https://app.terra.bio/#workspaces/warp-pipelines/DRAGEN-GATK-Whole-Genome-Germline-Pipeline
- https://broadinstitute.github.io/warp/docs/Pipelines/Whole_Genome_Germline_Single_Sample_Pipeline/README



# Variant Calling for Sequence Genotyping



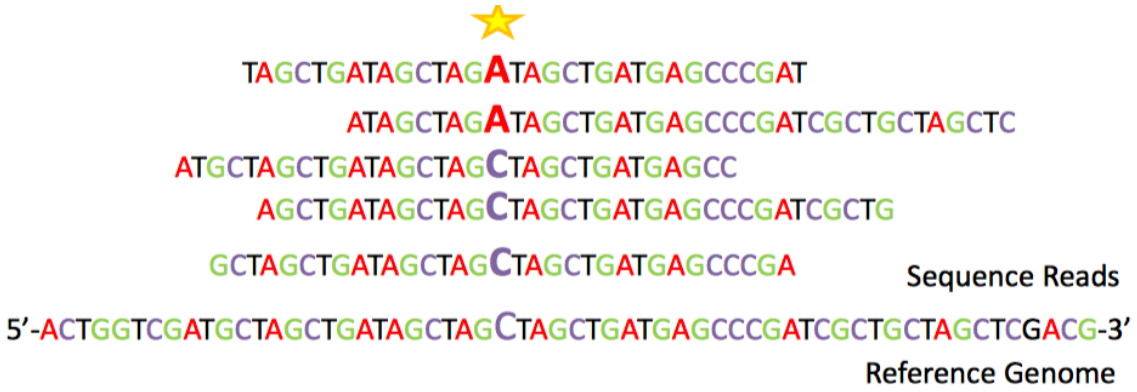

$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

## Ingredients That Go Into Prior

- Most sites don't vary
  - P(non-reference base) ~ 0.001
- When a site does vary, it is usually heterozygous
  - P(non-reference heterozygote) ~ 0.001 * 2/3
  - P(non-reference homozygote) ~ 0.001 * 1/3
- Mutation model
  - Transitions account for most variants (C↔T or A↔G)
  - Transversions account for minority of variants

# Variant Calling for Sequence Genotyping




- Assume the probability to sequence "C" given "A", or sequence "A" given "C" is 0.001:

P(reads | AA) = dbinom(3, size = 5, prob = 0.001) = 9.98e-9
P(reads | AC) = dbinom(3, size = 5, prob = 0.5) = 0.312
P(reads | CC) = dbinom(2, size = 5, prob = 0.001) = 9.97e-6

- Assume population based prior allele frequency P(A) = 0.2: Prior(AA) = 0.04; Prior(AC) = 0.32; Prior (CC) = 0.64
  - Posterior(AA) < 0.001
  - Posterior(AC) = 0.999
  - Posterior(CC) < 0.001

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced

# Genotype data format

- Single nucleotide polymorphism (SNP) : One reference allele and one alternative allele, e.g., AG, AA, GG
- Variant Call Format (VCF): Text file recording one SNP per row

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                              FORMAT       NA00001         NA00002          NA00003
20     14370    rs6054257  G     A      29    PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
20     1110696  rs6040355  A     G,T    67    PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS    NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4        0/2:17:2         1/1:40:3
```

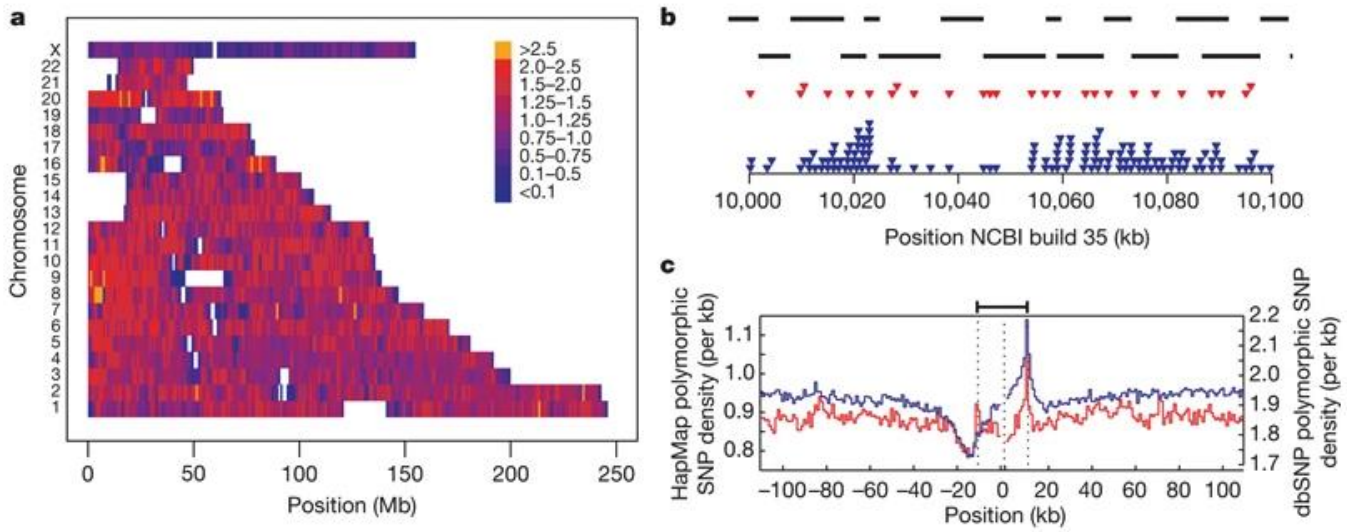
# Human genetic variants and sample sizes over past 20 years

| Year | No. of Samples | No. of Markers | Publication |
|---|---|---|---|
| Ongoing | 120,000 | 600 million | NHLBI Precision Medicine Cohorts / TopMed |
| 2016 | 32,488 | 40 million | Haplotype Reference Consortium (Nature Genetics) |
| 2015 | 2,500 | 80 million | The 1000 Genomes Project (Nature) |
| 2012 | 1,092 | 40 million | The 1000 Genomes Project (Nature) |
| 2010 | 179 | 16 million | The 1000 Genomes Project (Nature) |
| 2010 | 100,184 | 2.5 million | Lipid GWAS (Nature) |
| 2008 | 8,816 | 2.5 million | Lipid GWAS (Nature Genetics) |
| 2007 | 270 | 3.1 million | HapMap (Nature) |
| 2005 | 270 | 1 million | HapMap (Nature) |
| 2003 | 80 | 10,000 | Chr. 19 Variation Map (Nature Genetics) |
| 2002 | 218 | 1,500 | Chr. 22 Variation Map (Nature) |
| 2001 | 800 | 127 | Three Region Variation Map (Am J Hum Genet) |
| 2000 | 820 | 26 | T-cell receptor variation (Hum Mol Genet) |

# HapMap Project (2002 - 2010)

- Goal: Develop a haplotype map of the human genome.
- The HapMap is valuable by reducing the number of SNPs required to examine the entire genome for association with a phenotype from the 10 million SNPs that exist to roughly 500,000 tag SNPs.
- **Phase II HapMap** characterizes **>3.1M SNPs (~1.3M from Phase I)** genotyped in **270** individuals from **4 populations**, including 25-35% common SNPs (~10M), one per kilobase.
  - 30 trios of Yoruba (**YRI**) in Ibadan, Nigeria : African
  - 30 trios of Centre d'Etude du Polymorphisme Humain (CEPH) collection (**CEU**) living in Utah : European
  - 45 unrelated Han Chinese individuals in Beijing (**CHB**) + 45 unrelated Japanese in Tokyo (**JPT**) : Asian

The International HapMap Consortium. Nature 449, 851-861 (2007).




**a**, SNP density across the genome. Colours indicate the number of polymorphic SNPs per kb in the consensus data set. Gaps in the assembly are shown as white. **b**, Example of the fine-scale structure of SNP density for a 100-kb region on chromosome 17 showing Perlegen amplicons (black bars), polymorphic Phase I SNPs in the consensus data set (red triangles) and polymorphic Phase II SNPs in the consensus data set (blue triangles). Note the relatively even spacing of Phase I SNPs. **c**, The distribution of polymorphic SNPs in the consensus Phase II HapMap data (blue line and left-hand axis) around coding regions. Also shown is the density of SNPs in dbSNP release 125 around genes (red line and right-hand axis). Values were calculated separately 5′ from the coding start site (the left dotted line) and 3′ from the coding end site (right dotted line) and were joined at the median midpoint position of the coding unit (central dotted line).

# Figure 1: SNP density in the Phase II HapMap.

The International HapMap Consortium. Nature 449, 851-861 (2007).

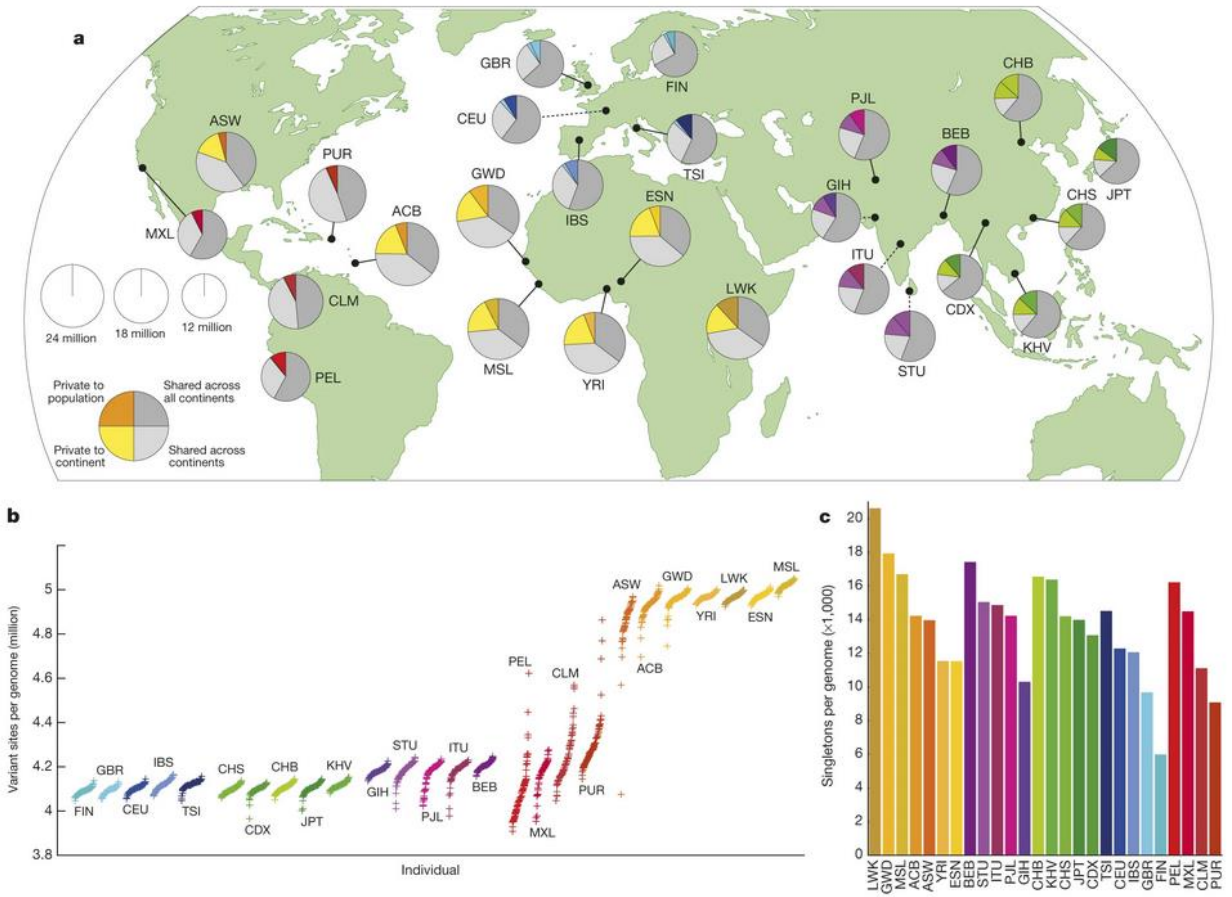
# 1000 Genome Project (2008 - 2015)

Goal: Find most genetic variants with MAF > 1% in populations across the world.

- First project to sequence (~8X coverage) the genomes of a large number of people (n = 2504)
- Largest public catalogue of human variation and genotype data: http://www.internationalgenome.org/
- 26 different populations under 5 super populations:
  - **AFR**, African
  - **AMR**, Ad Mixed American
  - **EAS**, East Asian
  - **EUR**, European
  - **SAS**, South Asian

The 1000 Genomes Project Consortium. Nature 526, 68-74 (2015).

From: A global reference for human genetic variation




**a,** Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. **b,** The number of variant sites per genome. **c,** The average number of singletons per genome.

# Table 1. Median autosomal variant sites per genome in 1000 genome project

| | AFR | | AMR | | EAS | | EUR | | SAS | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Samples** | **661** | | **347** | | **504** | | **503** | | **489** | |
| **Mean coverage** | **8.2** | | **7.6** | | **7.7** | | **7.4** | | **8.0** | |
| | **Var. sites** | **Singletons** | **Var. sites** | **Singletons** | **Var. sites** | **Singletons** | **Var. sites** | **Singletons** | **Var. sites** | **Singletons** |
| SNPs | 4.31M | 14.5k | 3.64M | 12.0k | 3.55M | 14.8k | 3.53M | 11.4k | 3.60M | 14.4k |
| Indels | 625k | - | 557k | - | 546k | - | 546k | - | 556k | - |
| Large deletions | 1.1k | 5 | 949 | 5 | 940 | 7 | 939 | 5 | 947 | 5 |
| CNVs | 170 | 1 | 153 | 1 | 158 | 1 | 157 | 1 | 165 | 1 |
| MEI (Alu) | 1.03k | 0 | 845 | 0 | 899 | 1 | 919 | 0 | 889 | 0 |
| MEI (L1) | 138 | 0 | 118 | 0 | 130 | 0 | 123 | 0 | 123 | 0 |
| MEI (SVA) | 52 | 0 | 44 | 0 | 56 | 0 | 53 | 0 | 44 | 0 |
| MEI (MT) | 5 | 0 | 5 | 0 | 4 | 0 | 4 | 0 | 4 | 0 |
| Inversions | 12 | 0 | 9 | 0 | 10 | 0 | 9 | 0 | 11 | 0 |
| Nonsynon | 12.2k | 139 | 10.4k | 121 | 10.2k | 144 | 10.2k | 116 | 10.3k | 144 |
| Synon | 13.8k | 78 | 11.4k | 67 | 11.2k | 79 | 11.2k | 59 | 11.4k | 78 |
| Intron | 2.06M | 7.33k | 1.72M | 6.12k | 1.68M | 7.39k | 1.68M | 5.68k | 1.72M | 7.20k |
| UTR | 37.2k | 168 | 30.8k | 136 | 30.0k | 169 | 30.0k | 129 | 30.7k | 168 |
| Promoter | 102k | 430 | 84.3k | 332 | 81.6k | 425 | 82.2k | 336 | 84.0k | 430 |
| Insulator | 70.9k | 248 | 59.0k | 199 | 57.7k | 252 | 57.7k | 189 | 59.1k | 243 |
| Enhancer | 354k | 1.32k | 295k | 1.05k | 289k | 1.34k | 288k | 1.02k | 295k | 1.31k |
| TFBSs | 927 | 4 | 759 | 3 | 748 | 4 | 749 | 3 | 765 | 3 |
| Filtered LoF | 182 | 4 | 152 | 3 | 153 | 4 | 149 | 3 | 151 | 3 |
| HGMD-DM | 20 | 0 | 18 | 0 | 16 | 1 | 18 | 2 | 16 | 0 |
| GWAS | 2.00k | 0 | 2.07k | 0 | 1.99k | 0 | 2.08k | 0 | 2.06k | 0 |
| ClinVar | 28 | 0 | 30 | 1 | 24 | 0 | 29 | 1 | 27 | 1 |

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

The 1000 Genomes Project Consortium. Nature 526, 68-74 (2015).

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program, Taliun D. et. al. Nature 2021.

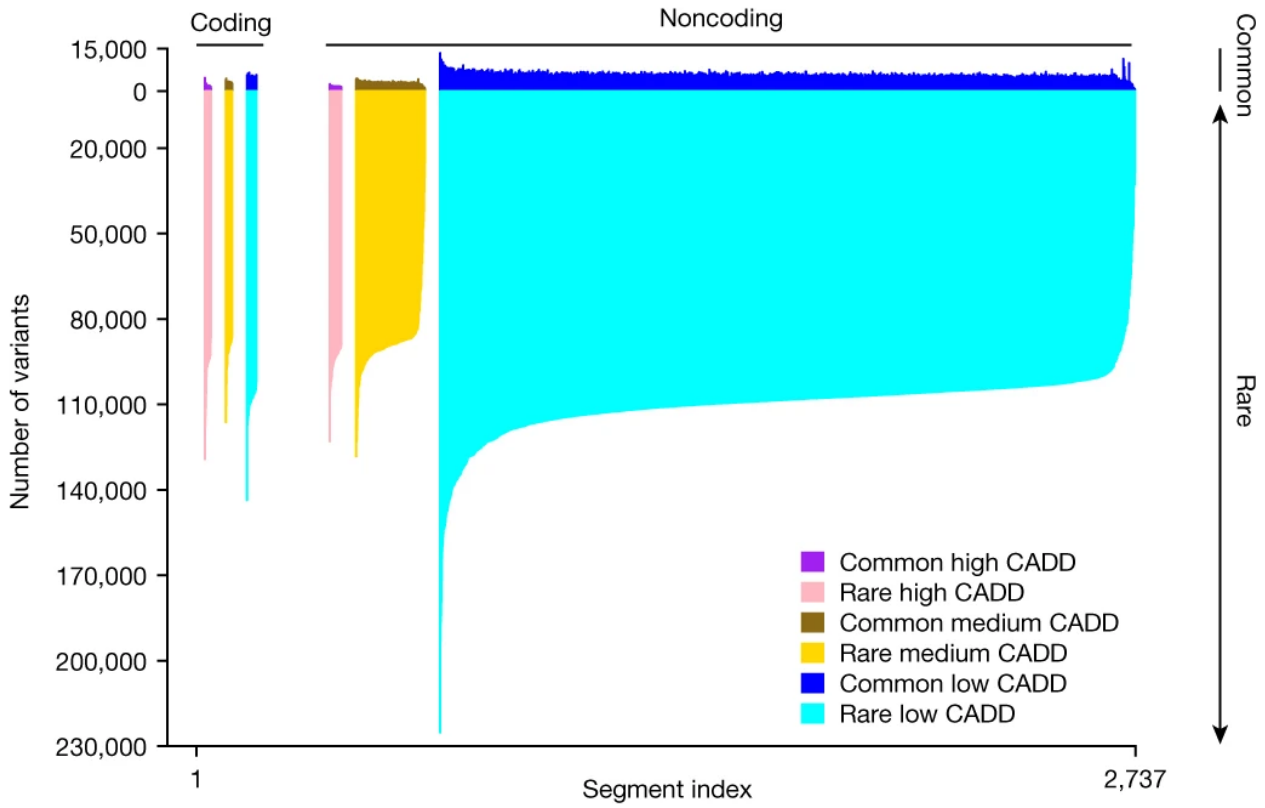
| | All unrelated individuals ($n$ = 40,722) | | Per individual | | | |
|---|---|---|---|---|---|---|
| | Total | Singletons (%) | Average | 5th percentile | Median | 95th percentile |
| **Total variants** | **384,127,954** | **203,994,740 (53)** | **3,748,599** | **3,516,166** | **3,563,978** | **4,359,661** |
| **SNVs** | 357,043,141 | 189,429,596 (53) | 3,553,423 | 3,335,442 | 3,380,462 | 4,125,740 |
| **Indels** | 27,084,813 | 14,565,144 (54) | 195,176 | 180,616 | 183,503 | 233,928 |
| **Novel variants** | **298,373,330** | **191,557,469 (64)** | **29,202** | **20,312** | **24,106** | **44,336** |
| **SNVs** | 275,141,134 | 177,410,620 (64) | 25,027 | 17,520 | 20,975 | 36,861 |
| **Indels** | 23,232,196 | 14,146,849 (61) | 4,175 | 2,747 | 3,145 | 7,359 |
| **Coding variation** | **4,651,453** | **2,523,257 (54)** | **23,909** | **22,158** | **22,557** | **27,716** |
| **Synonymous** | 1,435,058 | 715,254 (50) | 11,651 | 10,841 | 11,056 | 13,678 |
| **Nonsynonymous** | 2,965,093 | 1,648,672 (56) | 11,384 | 10,632 | 10,856 | 13,221 |
| **Stop/essential splice** | 97,217 | 60,347 (62) | 474 | 425 | 454 | 566 |
| **Frameshift** | 104,704 | 71,577 (68) | 132 | 112 | 127 | 165 |
| **In-frame** | 51,997 | 29,110 (56) | 102 | 85 | 99 | 128 |

Novel variants are taken as variants that were not present in dbSNP build 149, the most recent dbSNP version without TOPMed submissions.

[Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program](), Taliun D. et. al. Nature 2021.

## Fig. 1: Distribution of genetic variants across the genome.

From: Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program




Common (allele frequency ≥ 0.5%) and rare (allele frequency < 0.5%) variant counts are shown above and below the *x* axis, respectively, within 1-Mb concatenated segments (see Methods). Segments are stratified by CADD functionality score, and sorted based on their number of rare variants according to the functionality category. There were 22 high CADD, 22 medium CADD and 34 low CADD coding segments, and 40 high CADD, 238 medium CADD and 2,381 low CADD noncoding segments. Noncoding regions of the genome with low CADD scores (<10, reflecting lower predicted function) have the largest levels of common and rare variation (noncoding plot region, dark and light blue, respectively), followed by low CADD coding regions (coding plot region, dark and light blue, respectively). Overall, the vast majority of human genomic variation comprises rare variation.

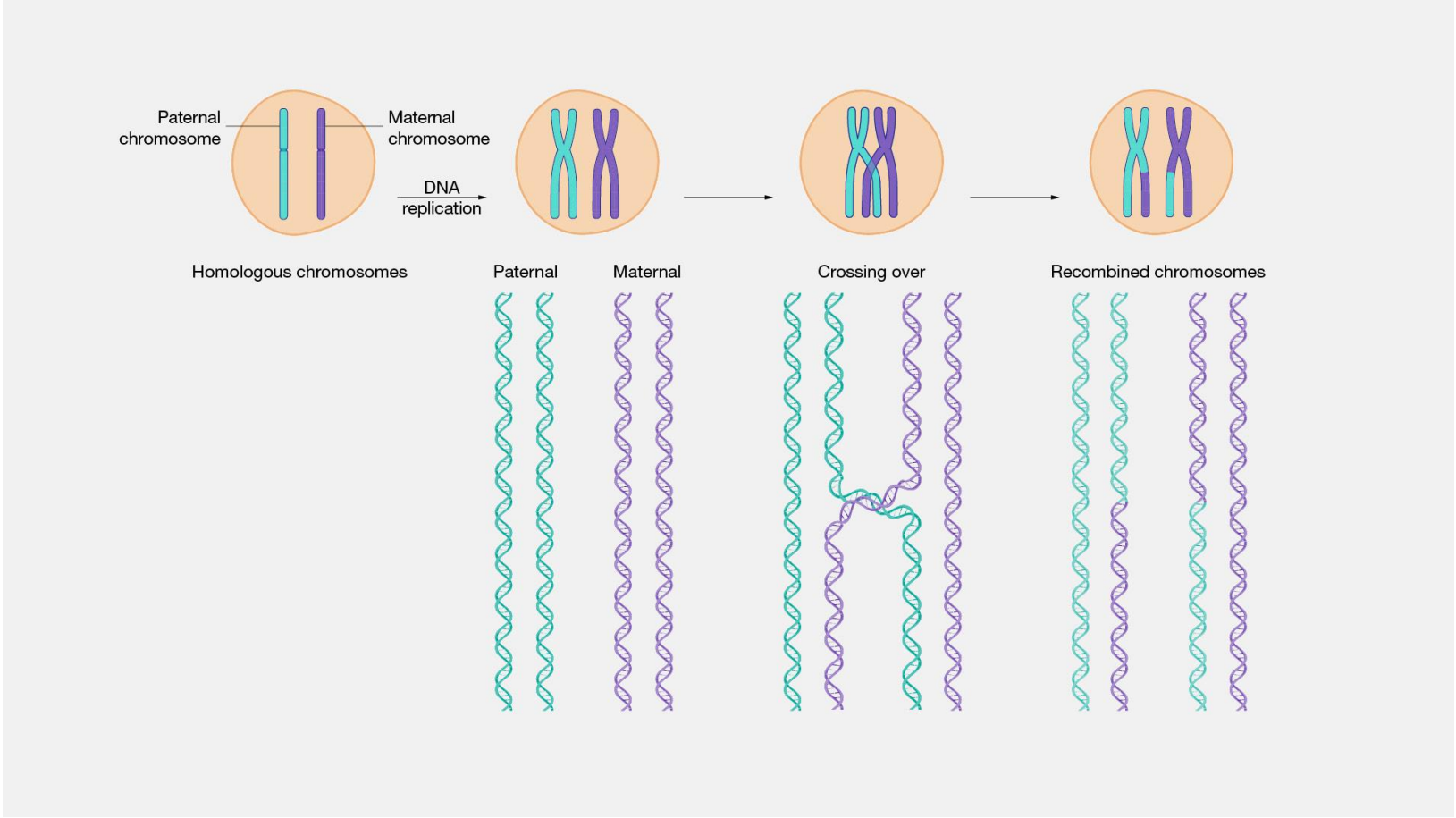
# Impute Microarray Genotype Data

- Using WGS reference panel, e.g., 1000 Genome, TOPMed
- Fill in SNP genotypes for those not genotyped by Microarray
- Genotype imputation has become a standard tool in GWAS
  - Can only impute variants observed in a reference panel. Reference panels with millions of deeply sequenced individuals are available.
  - Result in ~10M imputed common variants
  - Improve GWAS power
  - Facilitating fine-mapping and meta-analysis
  - Facilitating GWAS results interpretation

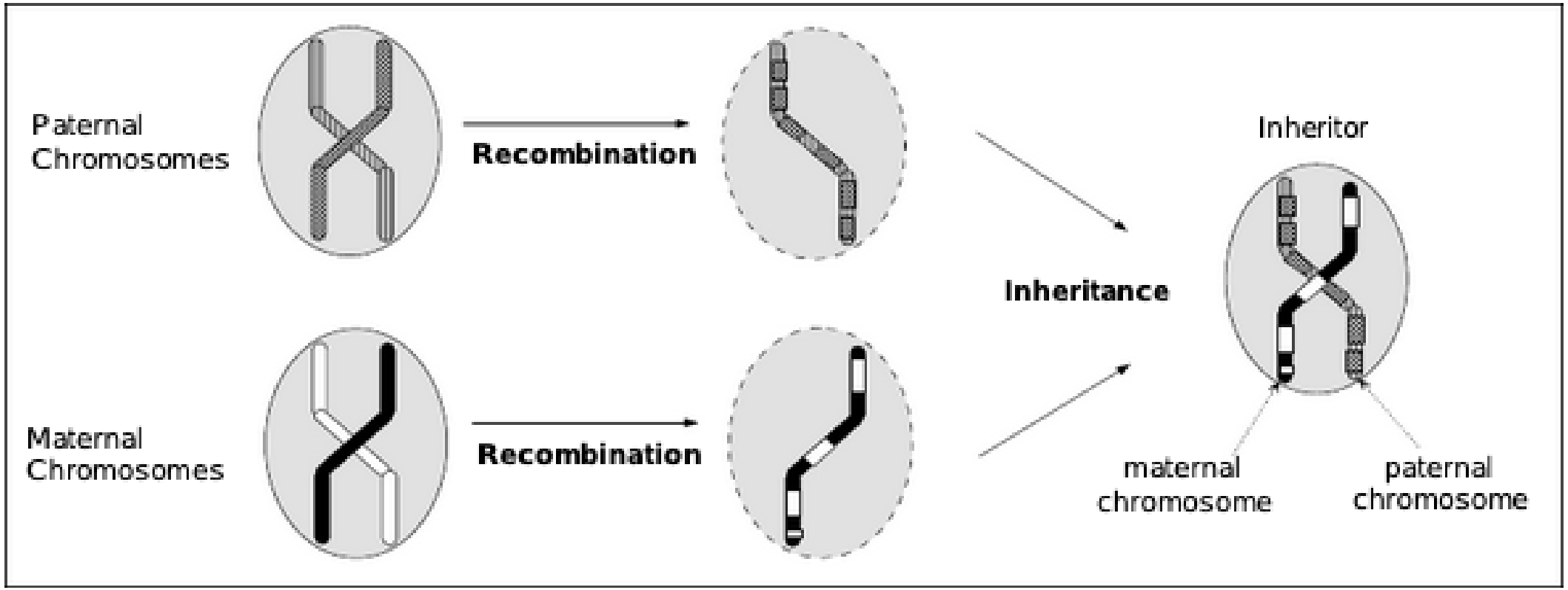
# Genotype Imputation Intuition

- Any two individuals, even if unrelated, can share short stretches of chromosome derived from a distant common ancestor.
- Observed genotypes from Microarray can be used to identify DNA segments shared between the study sample and a reference panel of sequenced genomes.
- A study haplotype can be represented as a mosaic of short segments of related haplotypes found in the reference panel.
- Points where the reference haplotype template changes represent historical recombination events.
- Points where the observed target allele differs from the template allele represent historical mutation events, gene conversion events, genotype error, or erroneously assigned matches.

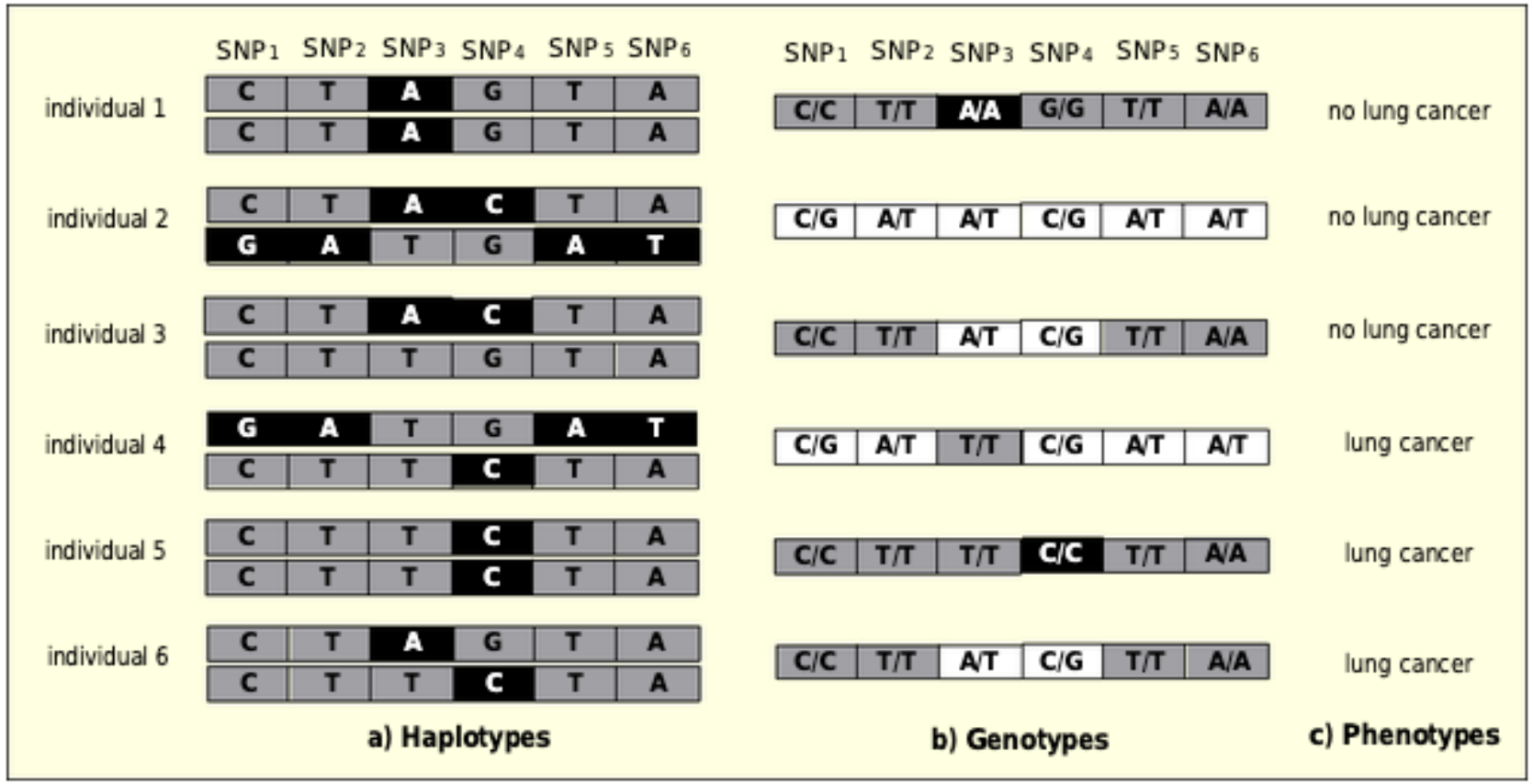
**Homologous recombination** is a type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. During the formation of egg and sperm cells (meiosis), paired chromosomes from the male and female parents align so that similar DNA sequences can cross over, or be exchanged, from one chromosome to the other. This exchanging of DNA is an important source of the genomic variation seen among offspring.



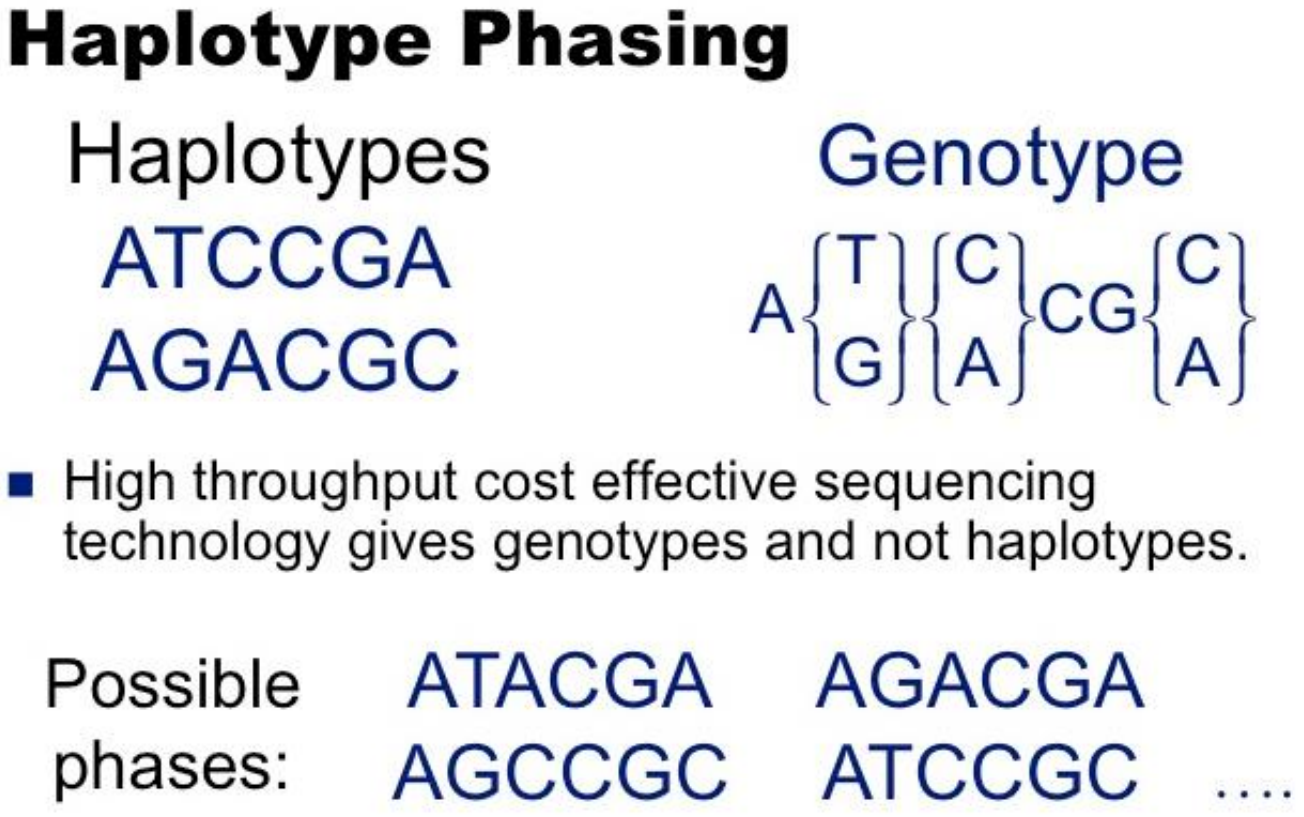
# Recombination and Inheritance



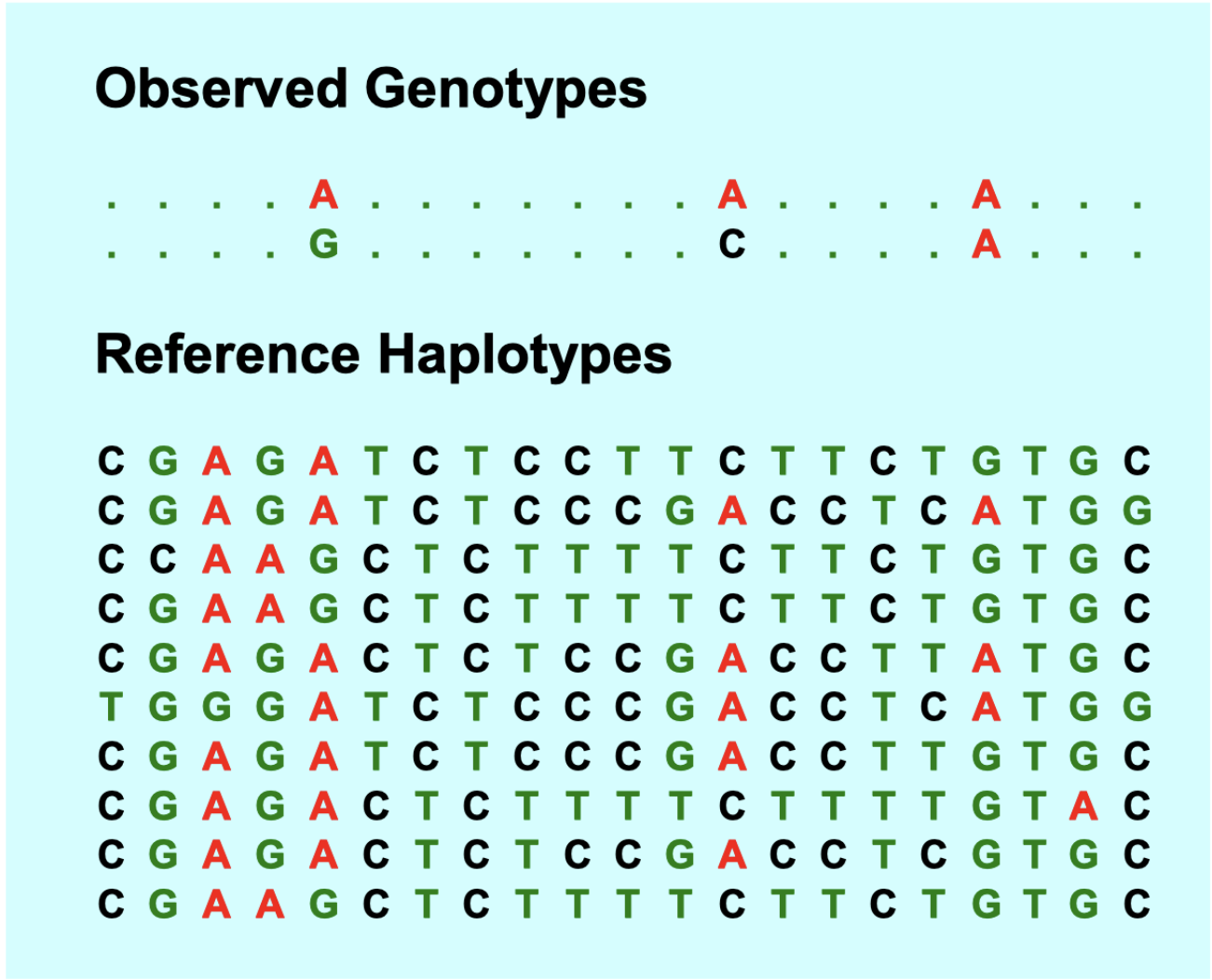
# Haplotypes, Genotypes, and Phenotypes



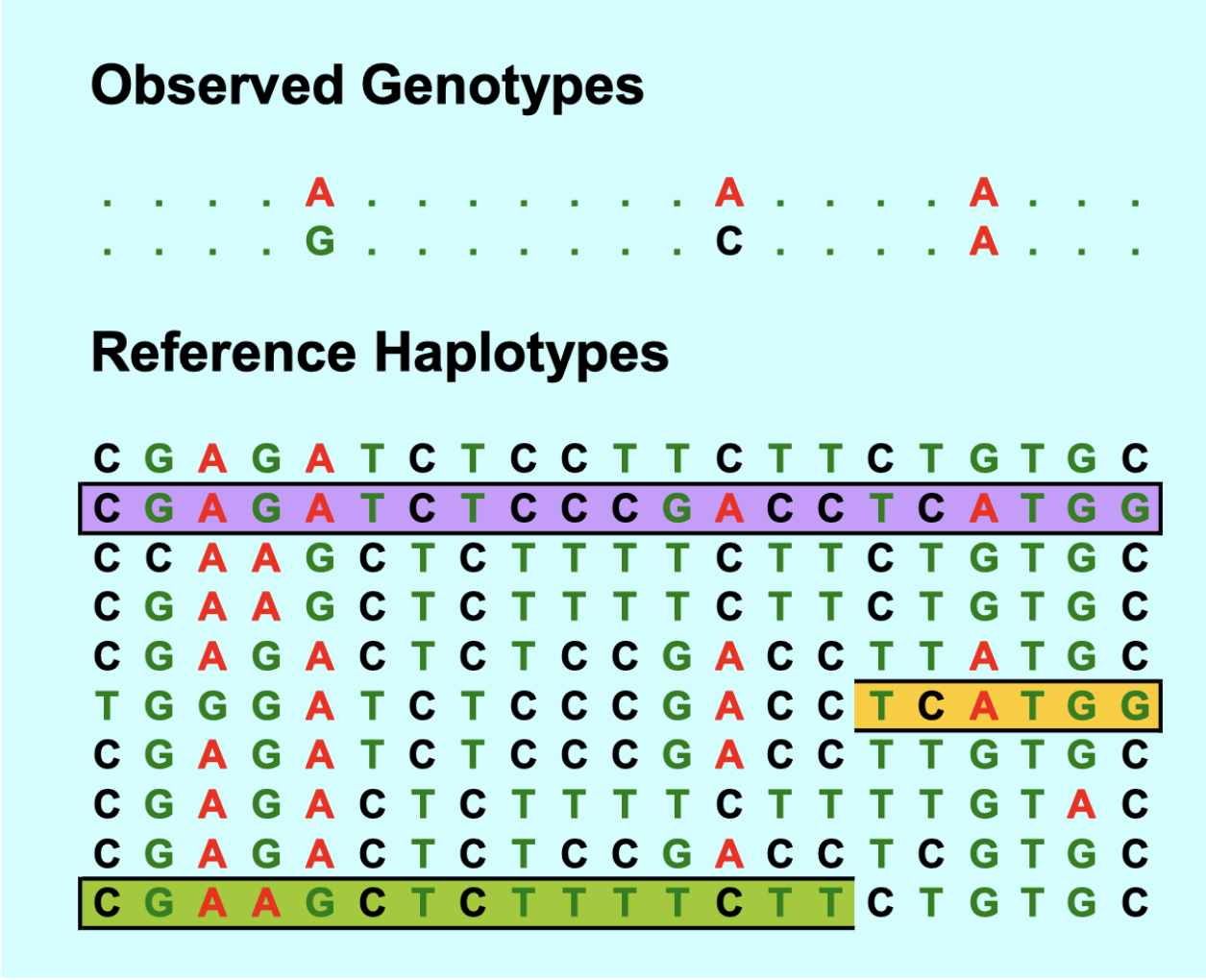
The problem of Haplotype Inference referred to as Haplotype Phasing. Genotyping technologies obtain “genotype” information on SNPs which mixes the genetic information from both chromosomes. However, many genetic analyses require “haplotype” information (like genotype imputation) which is the genetic information on each chromosome (see Figure).



# Observed Genotypes



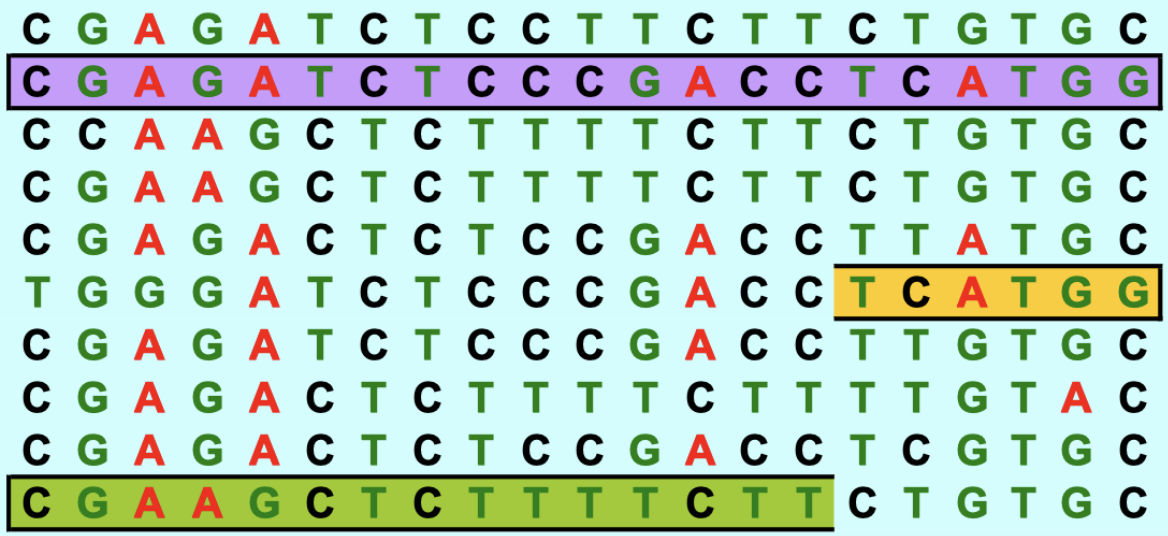
# Identify Match Among Reference



# Phase Chromosome, Impute Missing Genotypes

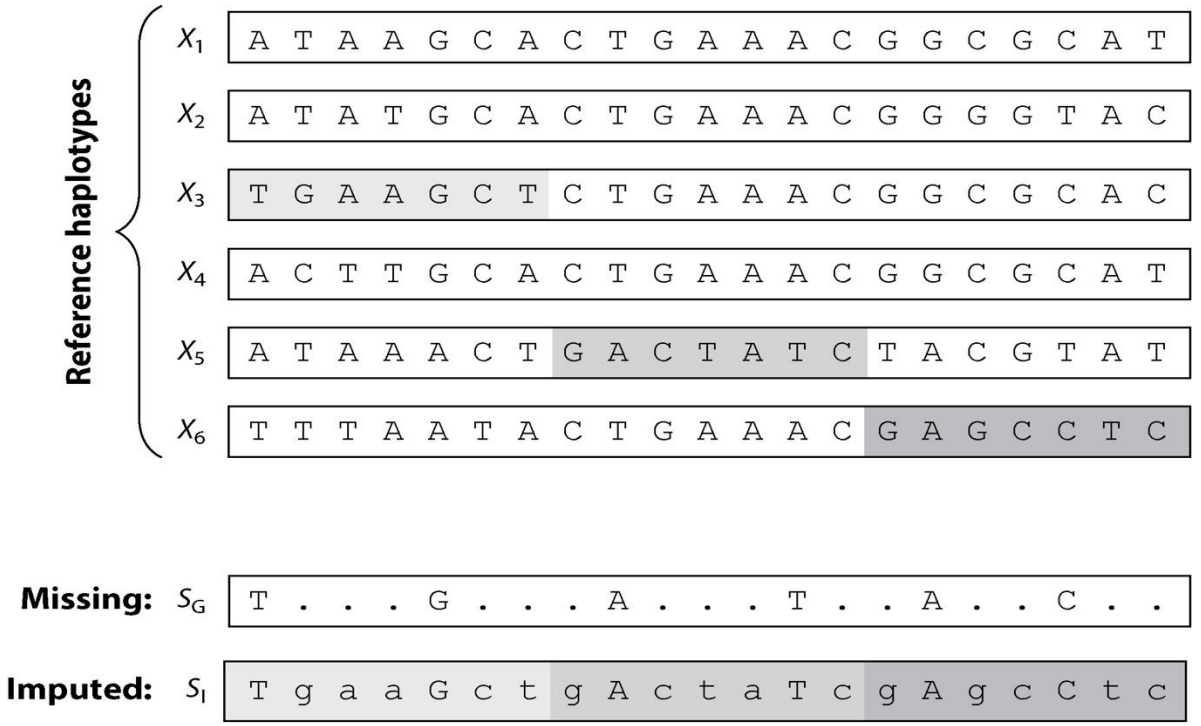
**Observed Genotypes**


**Reference Haplotypes**



# Li and Stephens Hidden Markov Mode

- Hidden Markov Model (HMM)
  - Observed genotypes of unknown phase in a study sample represent the observed data of the HMM.
  - Underlying and unobserved set of phased genotypes represent the hidden states of the HMM.
  - Probability of a template switch between markers is determined by the HMM transition probabilities, related to population recombination rate.
  - Probability that an observed allele differs from the template is determined by the HMM emission probabilities.

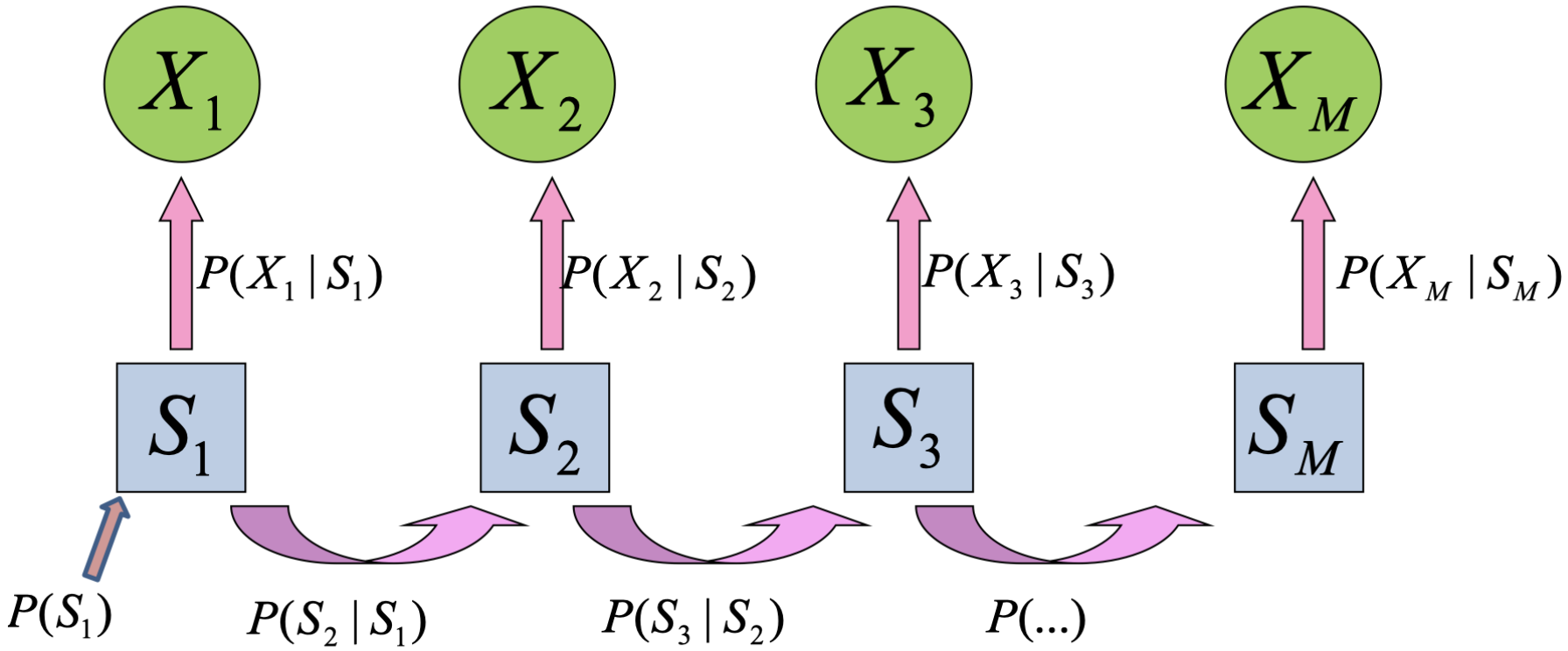



Das S, et al. 2018.
*Annu. Rev. Genom. Hum. Genet.* 19:73–96

**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, $S_I$.

# Markov Model




The final ingredient connects template states along the chromosome …

# Possible States

- A state *S* selects pair of template haplotypes
  - Consider $S_i$ as vector with two elements ($S_{i,1}$, $S_{i,2}$)
- With *H* possible haplotypes, $H^2$ possible states
  - $H(H+1)/2$ of these are distinct
- A recombination rate parameter describes probability of switches between states
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a, S_{i+1,2} = b))$ $(1-\theta)^2$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b))$ $(1-\theta)\theta/H$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b^*))$ $(\theta/H)^2$

# Emission Probabilities

- Each value of *S* implies expected pair of alleles
- Emission probabilities will be higher when observed genotype matches expected alleles
- Emission probabilities will be lower when alleles mismatch
- Let *T(S)* be a function that provides expected allele pairs for each state *S*

# Emission Probabilities

$$P(G_j|S_j) = \begin{cases} (1-\varepsilon_j)^2+\varepsilon_j^2, & T(S_j)=G_j \text{ and } G_j \text{ is heterozygote,} \\ 2(1-\varepsilon_j)\varepsilon_j, & T(S_j)\neq G_j \text{ and } G_j \text{ is heterozygote,} \\ (1-\varepsilon_j)^2, & T(S_j)=G_j \text{ and } G_j \text{ is homozygote,} \\ (1-\varepsilon_j)\varepsilon, & T(S_j) \text{ is heterozygote and } G_j \text{ homozygote,} \\ \varepsilon_j^2, & T(S_j) \text{ and } G_j \text{ are opposite homozygotes.} \end{cases}$$

# HMM

- The probability of each possible unobserved path through the HMM hidden states (reference haplotypes) can be calculated.
  - Penalized when path switches reference haplotypes via HMM transition probability
  - Penalized when reference allele on the path differs from the observed allele via HMM emission probabilities
- Probability that the unobserved path goes through a particular HMM state (state probability per reference haplotype) can be calculated by HMM forward-backward algorithm
- The probability (Z) that the target haplotype (study sample) carries a particular allele is the sum of the state probabilities corresponding to reference haplotypes that carry the allele
  - Z is also the expected number of a particular allele

# Phasing

- Pre-phasing genotype data of the study sample greatly reduce computation burden of genotype imputation
  - First pre-phasing (haplotype estimation) of the genotypes of study samples
  - Imputation into the estimated study haplotypes
- Reduce the complexity of the imputation step from quadratic to linear in the number of reference haplotypes
  - Allowing matches to be found by comparing against phased sample haplotypes rather than against all pairs of sample haplotypes
  - Reduce cost for exploring multiple reference panels
  - Benefits from advanced phasing methods

# Techniques for Computation Efficiency

- Storing reference data in memory
  - Burrows-Wheeler Transform
  - M3VCF format :exploits local redundancy among haplotypes by only storing unique allele sequences along with a map
    - Reduce >90% computation time compared with using VCF format with >100K reference samples
    - Allowing reference haplotypes to be locally clustered
  - Binary reference format (bref) : Because of the bulk of alleles with low nonmajor allele frequency in reference panel, only store a list of reference haplotypes that carry the minor allele (one list per allele)
    - Searching the lists of haplotypes to find the allele on a given haplotype
    - If haplotype is not found in any list, the haplotype carries the major allele
    - Reduces >30% computation time with >100K reference samples

Minimac3: Das S. et. al. Nat. Genet. 2016

# Techniques for Computation Efficiency

- Clustering identical reference haplotype segments
  - Conduct local clustering ahead for the reference panel
  - Same allele sequence can be carried by many reference haplotypes in short regions
  - Reduce state space for non-boundary regions
- Imputation via linear interpolation
  - HMM state probabilities are calculated for genotyped markers of the study sample
  - HMM state probabilities at imputed markers are estimated by linear interpolation on genetic distance
- One can cluster reference haplotypes that have identical allele sequences between two genotyped markers before linear interpolation

Minimac3: Das S. et. al. Nat. Genet. 2016

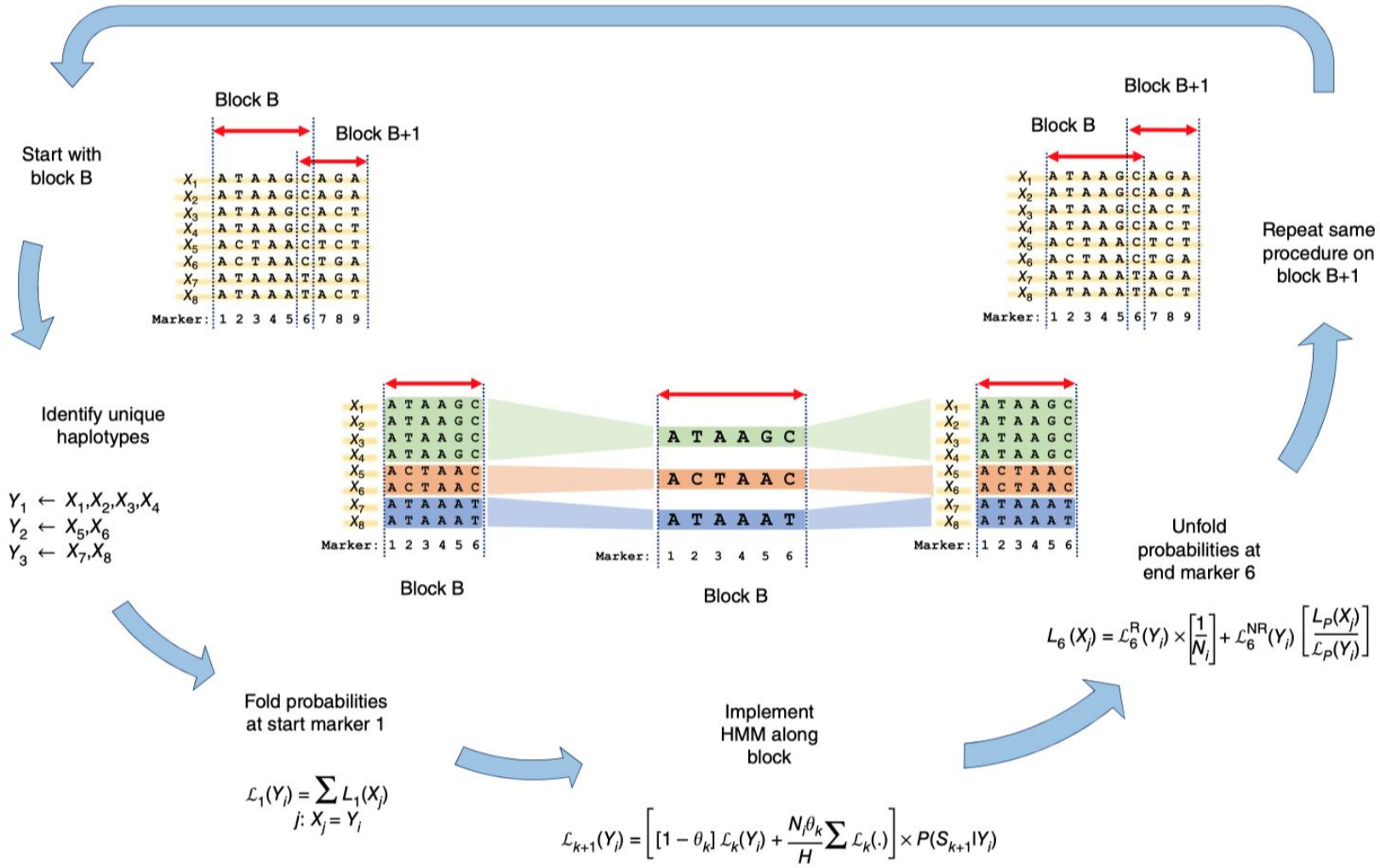



**Figure 1** Overview of state space reduction. We consider a chromosome region with $M = 9$ markers and $H = 8$ haplotypes: $X_1$, $X_2$, …, $X_8$. We break the region into consecutive genomic segments (blocks) and start by analyzing block B from marker 1 to marker 6. In block B, we identify $U = 3$ unique haplotypes: $Y_1$, $Y_2$, and $Y_3$ (colored in green, red, and blue, respectively). Given we know the left probabilities of the original state space at marker 1 (that is, $L_1(X_1)$, …, $L_1(X_8)$), we fold them to get the left probabilities of the reduced state space at marker 1: $\mathcal{L}_1(Y_1)$, $\mathcal{L}_1(Y_2)$, and $\mathcal{L}_1(Y_3)$. We implement HMM on the reduced state space ($Y_1$, $Y_2$, and $Y_3$) from marker 1 to marker 6 to get $\mathcal{L}_6(Y_1)$, $\mathcal{L}_6(Y_2)$, and $\mathcal{L}_6(Y_3)$. We next unfold the left probabilities of the reduced state space at marker 6 to obtain the left probabilities of the original state space: $L_6(X_1)$, …, $L_6(X_8)$. We repeat this procedure on the next block, starting with $L_6(X_1)$, …, $L_6(X_8)$, to finally obtain $L_9(X_1)$, …, $L_9(X_8)$.

[Minimac3: Das S. et. al. Nat. Genet. 2016]

**Table 1** Genotype imputation tools that employ a hidden Markov model (HMM)

| Tool | Year | Description of state space | Computational complexity | HMM parameter functions |
|---|---|---|---|---|
| FastPHASE | 2006 | All genotype configurations from a fixed number of localized haplotype clusters | Maximization-step linear in number of haplotypes, quadratic in number of clusters | Depends on recombination and mutation rates; parameters are fit using an expectation–maximization algorithm |
| IMPUTE | 2007 | All genotype configurations from all reference haplotypes | Quadratic in number of haplotypes | Depends on a fine-scale recombination map that is fixed and provided internally by the program |
| Beagle | 2007 | All genotype configurations from a variable number of localized haplotype clusters | Quadratic in number of haplotypes | Empirical model with no explicit parameter functions |
| IMPUTE2 | 2009 | All reference haplotypes | Phasing quadratic in number of haplotypes, imputation linear in number of haplotypes | Same as IMPUTE |
| MaCH | 2010 | All genotype configurations from all reference haplotypes | Quadratic in number of haplotypes | Depends on recombination rate, mutation rate, and genotyping error; parameters are fit using a Markov chain Monte Carlo or expectation–maximization algorithm |
| Minimac and Minimac2 | 2012 | All reference haplotypes | Linear in number of haplotypes | Same as MaCH |
| Minimac3 | 2016 | All unique allele sequences observed in reference data in a small genomic segment | Linear in number of haplotypes | Same as MaCH, but parameter estimates are precalculated and fixed |
| Beagle 4.1 | 2016 | All reference haplotypes at genotyped markers | Linear in number of haplotypes | Depends on recombination rates and error rates, which are precalculated and fixed |
| Minimac4 | 2017 | Collapsed allele sequences from reference data that match at genotyped positions in small genomic segments | Linear in number of haplotypes | Same as Minimac3 |
| IMPUTE4[a] | 2017 | All possible reference haplotypes | Linear in number of haplotypes | Same as IMPUTE2 |
| Beagle 5.0 | 2018 | A user-specified number of reference haplotypes | Linear in number of haplotypes | Same as Beagle 4.1 |

This table describes the typical state space and parameter functions used to model the Li and Stephens framework. Minimac and IMPUTE2 were the first tools to use the prephasing approach. Minimac3 and Beagle 4.1 exploit local haplotype redundancy to reduce the size of the state space and hence the computational burden.

[a]IMPUTE4 uses the same HMM as IMPUTE2; however, to reduce memory usage and increase speed, it uses compact binary data structures and takes advantage of high correlations between inferred copying states in the HMM to reduce computation.

Das S. et. al. Annual Reviews. 2018.

# Measuring Imputation Accuracy

- Imputation methods estimate a probability distribution for the allele carried by each haplotype per imputed marker
- Posterior genotype probabilities can be derived under HWE
- Expected allele dose (dosage) of the imputed genotype is given by the sum of the posterior allele probabilities for each haplotype —– used for follow-up GWAS
- Imputation $r^2$ : Squared correlation between the true and estimated dose of an allele across all imputed samples
  - Can be estimated from posterior allele probabilities without knowing the true allele
  - Threshold 0.3 is commonly used

## ESTIMATING $r^2$

One attractive feature of $r^2$, the squared correlation between true and imputed allele dose, is that it can be estimated from posterior allele probabilities without knowing the true allele on each chromosome. Here, we derive an estimate of $r^2$ in terms of the posterior allele probabilities.

Let $X$ be 1 if a chromosome carries the allele of interest and be 0 otherwise, and let $Z$ be the estimated posterior allele probability that $X = 1$. Then $r^2$ is defined to be the squared correlation of $X$ and $Z$. We say that the posterior allele probabilities are correctly calibrated if $E[X \mid Z] = Z$. If the posterior allele probabilities are correctly calibrated, we can use the law of total expectation and the fact that $X^2 = X$ to obtain

$$\begin{aligned} E[X^2] &= E[X] = E[E[X \mid Z]] = E[Z] \\ \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= E[Z] - E[Z]^2 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(X, Z) &= E[XZ] - E[X]E[Z] \\ &= E[E[XZ \mid Z]] - E[E[X \mid Z]]\, E[Z] \\ &= E[Z^2] - E[Z]E[Z] \\ &= \mathrm{Var}(Z). \end{aligned}$$

Consequently,

$$\begin{aligned} r^2 &= \frac{(\mathrm{Cov}(X, Z))^2}{\mathrm{Var}(X)\mathrm{Var}(Z)} \\ &= \frac{\mathrm{Var}(Z)}{\mathrm{Var}(X)} \\ &= \frac{E[Z^2] - E[Z]^2}{E[Z] - E[Z]^2} \end{aligned}$$

If there are $n$ imputed chromosomes and $z_i$ is the estimated reference allele probability in the $i$th haplotype, one can estimate $E[Z^k]$ as $E[Z^k] \approx (1/n)\sum z_i^k$ and $r^2$ as

$$r^2 \approx \frac{n\sum z_i^2 - (\sum z_i)^2}{n\sum z_i - (\sum z_i)^2}.$$

Das S. et. al. Annual Reviews. 2018.

**Table 2** The most commonly used public reference panels to date

| Reference panel | Number of reference samples | Number of sites (autosomes + X chromosome) | Average sequencing coverage | Ancestry distribution | Publicly available | Indels available | Reference |
|---|---|---|---|---|---|---|---|
| International HapMap Project phase 3 | 1,011 | 1.4 million | NA[a] | Multiethnic | Yes | No | **47** |
| 1000G phase 1 | 1,092 | 28.9 million | 2–6× | Multiethnic | Yes | Yes | **1** |
| 1000G phase 3 | 2,504 | 81.7 million | 7× genomes, 65× exomes | Multiethnic | Yes | Yes | **3** |
| UK10K Project | 3,781 | 42.0 million | 7× genomes, 80 × exomes | European | Yes | Yes | **89** |
| HRC | 32,470 | 40.4 million | 4–8 ×[b] | Predominantly European[c] | Partially[d] | No | **69** |
| TOPMed | 60,039 | 239.7 million | 30× | Multiethnic | Partially[e] | Yes | **71** |

Abbreviations: 1000G, 1000 Genomes Project; HRC, Haplotype Reference Consortium; indel, insertion or deletion; NA, not applicable; TOPMed, Trans-Omics for Precision Medicine.

[a]The International HapMap Project phase 3 data were genotyped on the Illumina Human1M and Affymetrix 6.0 SNP arrays.
[b]The HRC panel was obtained by combining sequencing data across many low-coverage (4–8×) and a few high-coverage sequencing studies.
[c]The only non-European samples in the HRC panel are through the 1000G reference panel (which was a contributing study).
[d]Most of the HRC samples (~27,000) are available for download through controlled access from the European Genome-Phenome Archive.
[e]Some of the TOPMed samples (~18,000) are available for download through controlled access from the Database of Genotypes and Phenotypes (dbGaP).

Das S. et. al. Annual Reviews. 2018.

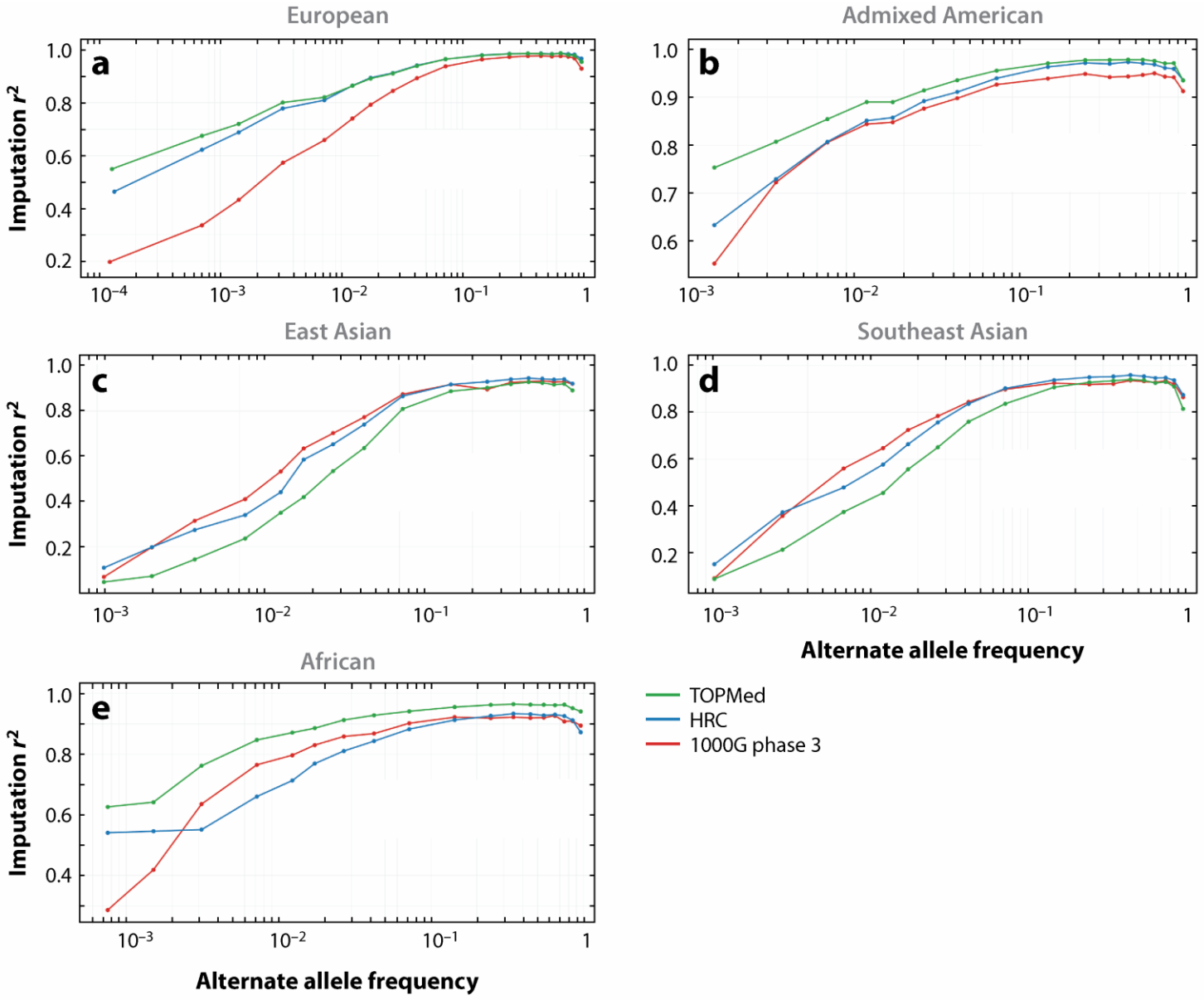


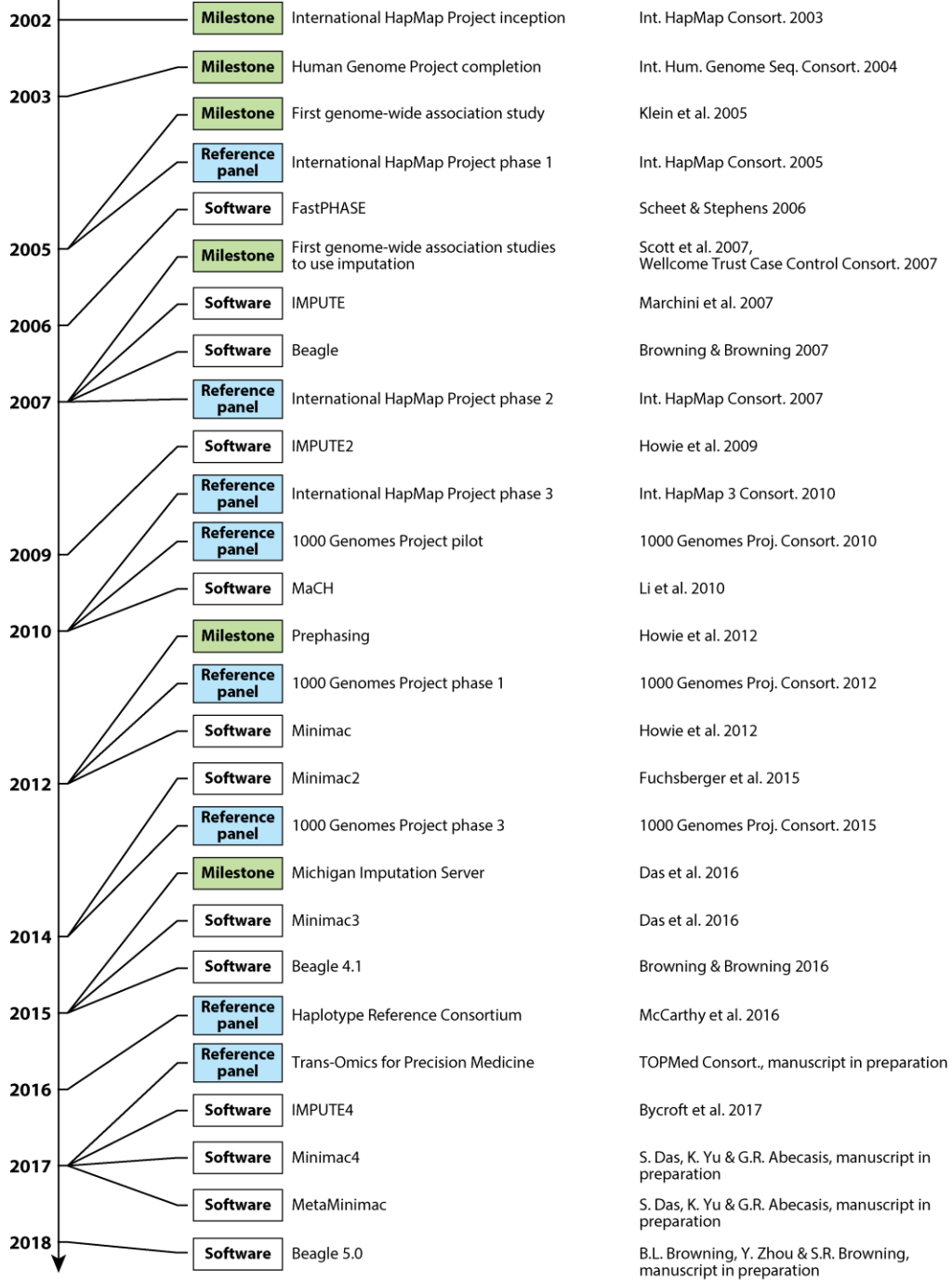

**Figure 5** Imputation accuracy for five ancestries: (*a*) European, (*b*) admixed American, (*c*) East Asian, (*d*) Southeast Asian, and (*e*) African. We extracted 10 samples from each of these ancestries from the 1000 Genomes Project (1000G) phase 3 data, masked all variants except those on the Illumina 1M chip, and imputed them using the Trans-Omics for Precision Medicine (TOPMed) (with 18,000 samples), Haplotype Reference Consortium (HRC), and 1000G phase 3 (after removing overlaps) reference panels. The aggregate $r^2$ (measuring the imputation accuracy) is plotted as a function of the alternate allele frequency.

Das S. et. al. Annual Reviews. 2018.

Das S, et al. 2018.
*Annu. Rev. Genom. Hum. Genet.* 19:73–96

# Summary of Genotype Imputation

- Using WGS reference panel, e.g., 1000 Genome, TOPMed
- Fill in SNP genotypes for those not genotyped by Microarray
- Check imputation $r^2$ : accept imputed genotypes with $r^2$ > e.g., 0.3
- Result in ~10M common variants
- Imputed genotype data
  - Dosage format — expected number of minor alleles with domain [0, 2]
  - Genotype format — number of minor alleles with values 0, 1, or 2
  - Genotype with the highest estimated probability will be reported




Das S, et al. 2018.
*Annu. Rev. Genom. Hum. Genet.* 19:73–96

**Figure 1** A brief time line summarizing the major developments in genotype imputation. Each major development has been categorized as a milestone (*green*), a reference panel (*blue*), or software (*white*).

[Das S. et. al. Annual Reviews. 2018.]

# Factors Affecting Genotype Imputation Accuracy

- Size of reference panel
- Density of genotyping array
- Minor allele frequency of variant being imputed (in the reference panel)
- Haplotype accuracy in reference and study samples
- Sequencing coverage of reference panel (ancestry matches)

# What is Association Studies?

- Test associations between markers/SNPs/genes and the trait of interest
- Test whether the trait and genotype are independent

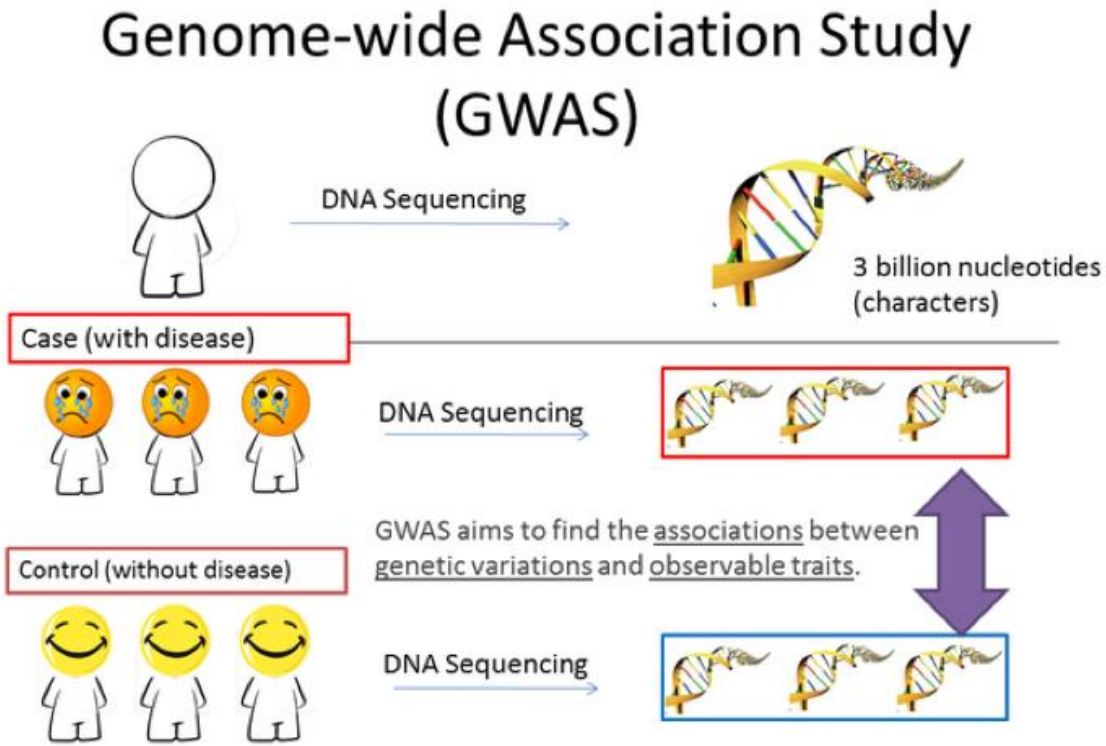
- **Population Data: Generalized linear regression model based tests**

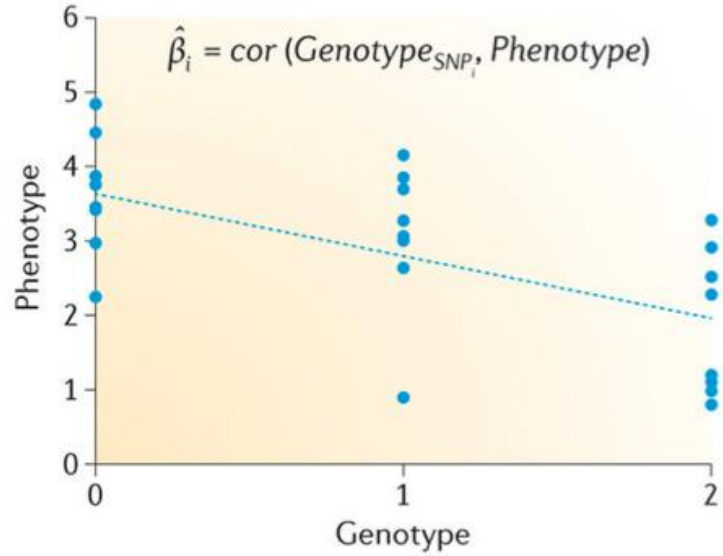
- Family Data

# Population-based Association Studies

- Phenotype(s) of interest
  - Dichotomous trait, e.g., case/control
  - Quantitative trait, e.g., Height, BMI, Lipids
  - Mendelian vs. Complex phenotypes
- Number of markers tested
  - May range from 1 to ~10 million
  - Candidate gene study (often appear as replication study)
  - Genome-wide association study (GWAS)

# GWAS




From `Quora.com` and Pasaniuc B & Price AL, Nat. Rev. 2017

# Phenotype and covariate data

- Phenotype data
  - Dichotomous traits : 0 / 1
  - Quantitative traits : observed continuous quantitative values
- Covariate data
  - Gender
  - Age
  - BMI
  - Batches, etc.

# Single Variant GWAS

- Test one SNP per time
- Test genome-wide variants independently
- Suitable for common SNPs with minor allele frequency (MAF) > 1%, or 0.1%

# Logistic Regression Model for Studying Dichotomous Phenotype

- $Y$ = dichotomous phenotype
- $X$ = a coding for the genotype

| Genotype | Codominant | Dominant | Recessive | Additive |
|---|---|---|---|---|
| AA | $X = (0,1)^{\mathrm{T}}$ | $X = 1$ | $X = 1$ | $X = 2$ |
| Aa | $X = (1,0)^{\mathrm{T}}$ | $X = 1$ | $X = 0$ | $X = 1$ |
| aa | $X = (0,0)^{\mathrm{T}}$ | $X = 0$ | $X = 0$ | $X = 0$ |

Assume a logistic regression model:

$$\log\left[\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right] = \beta_0 + \alpha C + \beta_1 X$$

where $\beta_0$ is the intercept, $\alpha$ is the coefficient for covariates $C$, and $\beta_1$ is the genetic effect-size (i.e., log(Odds-Ratio) ).

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

# Test Statistic

- Wald Test: $Z = \frac{\widehat{\beta_1}}{Standard\_Error(\widehat{\beta_1})} \sim N(0,1)$ under $H_0$
- Chi-square Test: $X^2 = \frac{\widehat{\beta_1}^2}{Var(\widehat{\beta_1})} \sim Chi\_Square$ with df=1 under $H_0$
- How to obtain p-value?

# Advantages of Logistic Regression Model

- Account for confounding covariates (C), e.g., age, gender, BMI, smoking
- Flexible for various genetic models
- Flexible for testing multiple markers in the same model (modeling LD)
- Equivalent to the corresponding Chi-square test using contingency tables, if not modeling covariates
- Allow gene-environment interactions
- Without the assumption of HWE

# Study Quantitative Trait

- Linear regression model
  - $Y = \beta_0 + \alpha C + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$
  - $Y$ represents the quantitative trait values
  - $X$ represents the genotype data (0, 1, 2) for additive genetic model
  - $C$ represents the confounding covariates or other environmental variables
  - $\epsilon$ represents the error term, other unknown factors
- $H_0: \beta_1 = 0\ ; H_a: \beta_1 \neq 0$
- P-values can be obtained by Wald Test, T-test, Score test, Log likelihood test, etc.

# Genome-wide Association Study (GWAS)

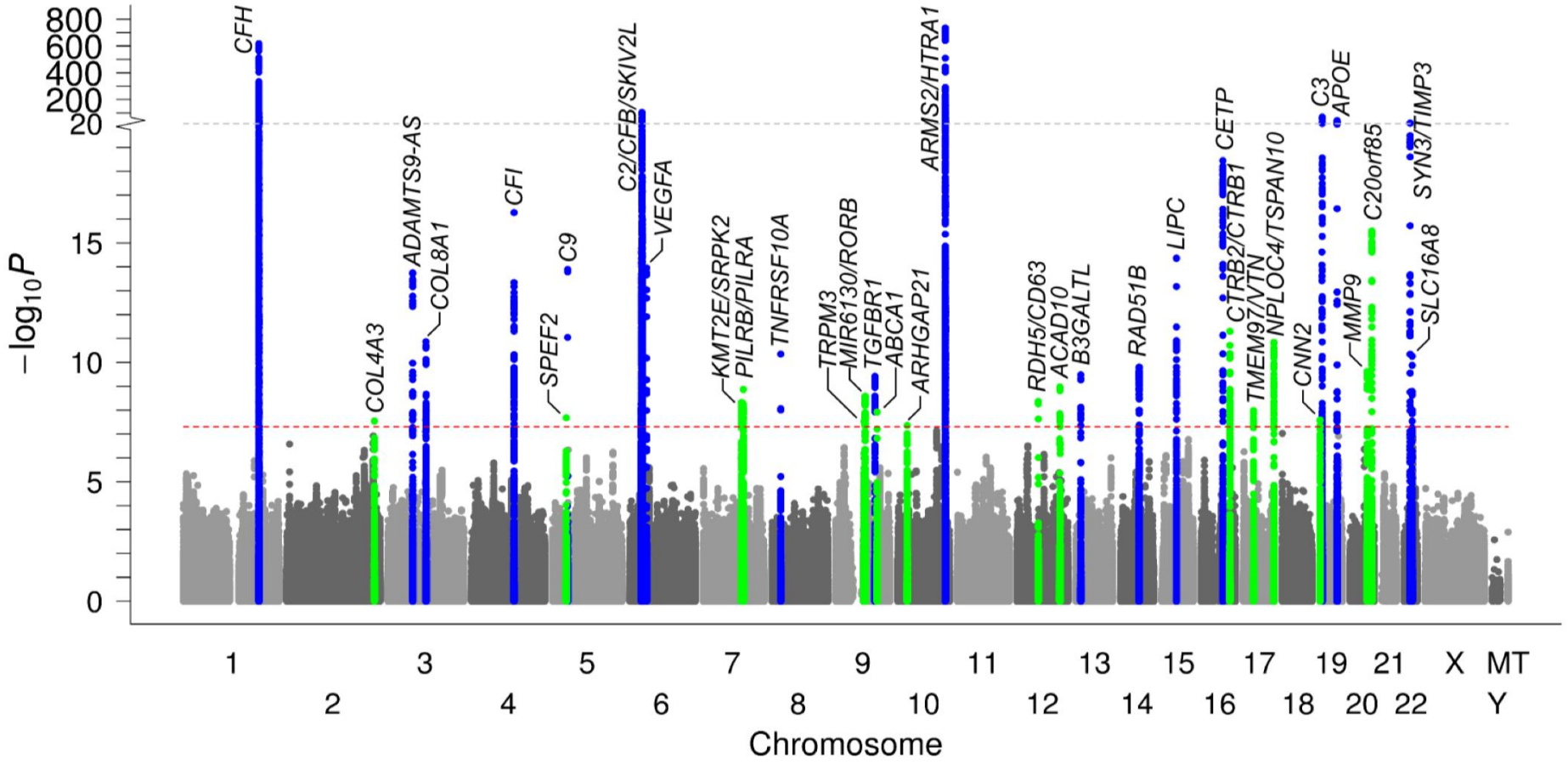
**GWAS**: independent single-variant tests across all genome-wide variants

- Quality control (QC) of the study dataset
- Choose a model/test for the phenotype of interest (e.g., linear regression model for quantitative traits, logistic regression model for dichotomous traits, other association tests from previous lecture)
- Significance level $\alpha = 5 \times 10^{-8}$
- Report nearby genes of significant SNPs

# Visualize GWAS Results by Manhattan Plot

– Scatter plot of $-\log 10$(p-values) across all genome-wide variants
– Visualize signal peaks

# GWAS Results

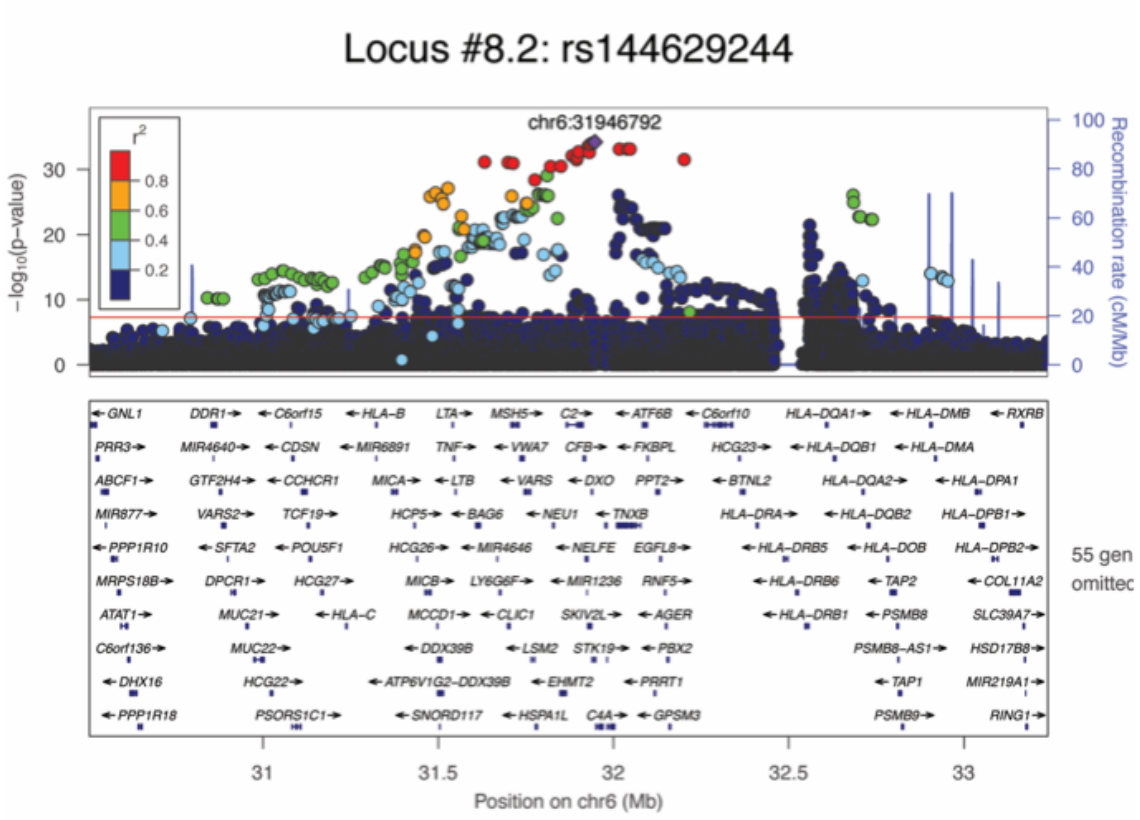



Fritsche L.G. et al. Nat Genet, 2016.

**18 known AMD loci** and **16 novel AMD loci**

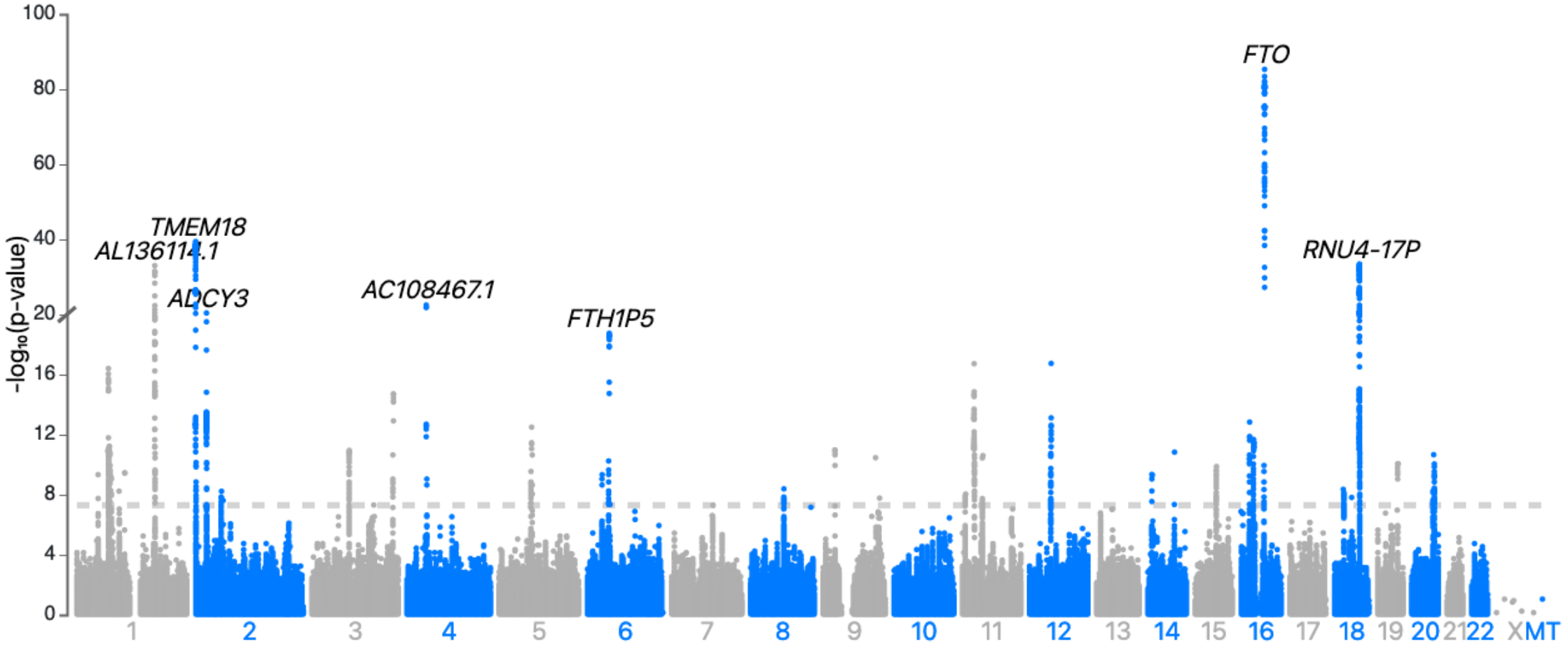
# Visualize GWAS Loci by Locus Zoom Plot

– Zoom into the peak region with gene annotations
– Visualize $r^2$ between the specified significant (purple diamond) signal and its neighbor SNPs
– Visualize recombination rate

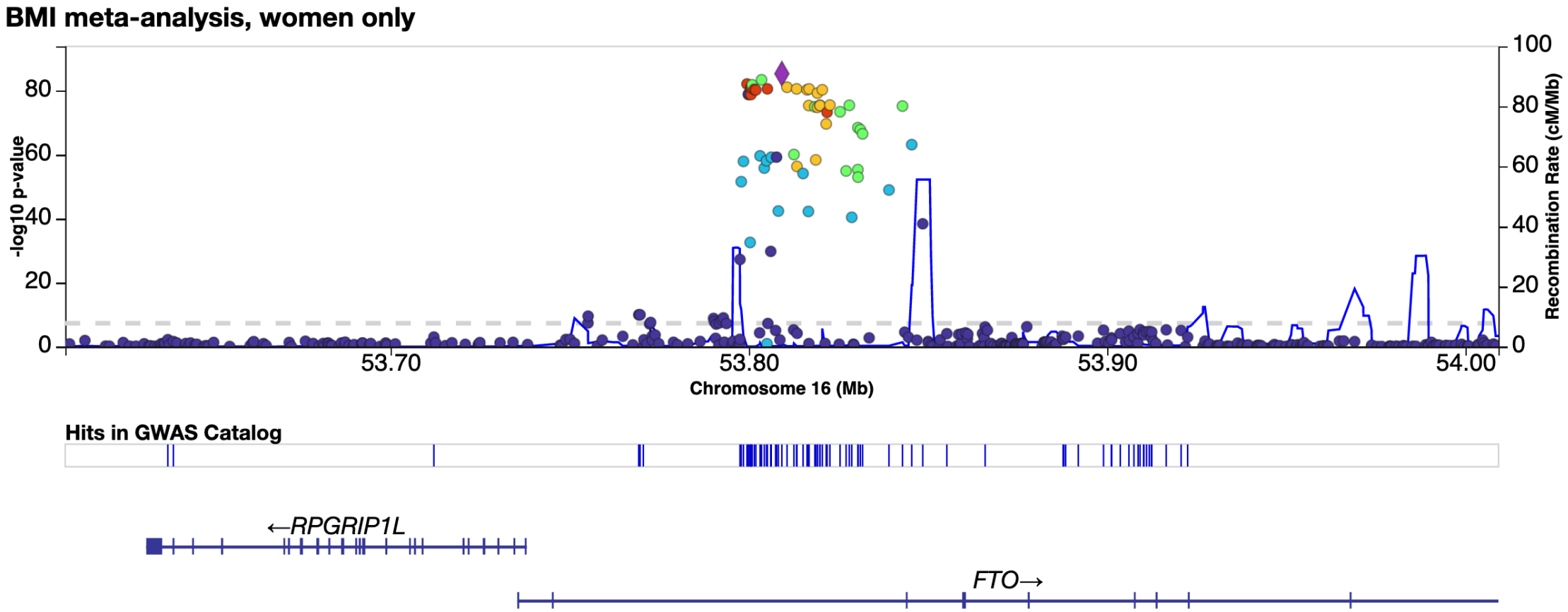



Fritsche L.G. et al. Nat Genet, 2016.

# LocusZoom Visualization of GWAS of BMI , Women only



# LocusZoom Visualization of GWAS of BMI



# Example GWAS Discoveries

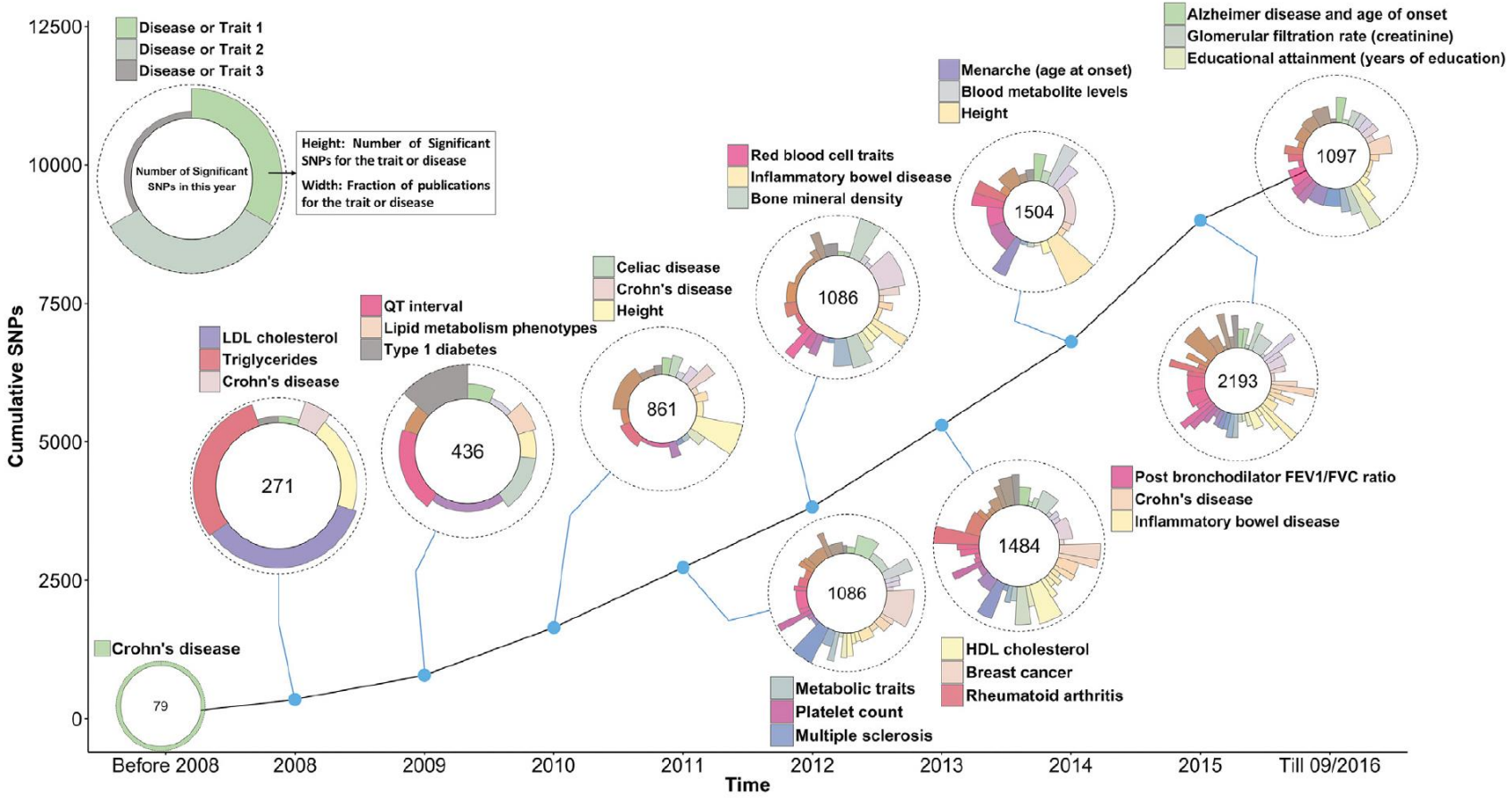
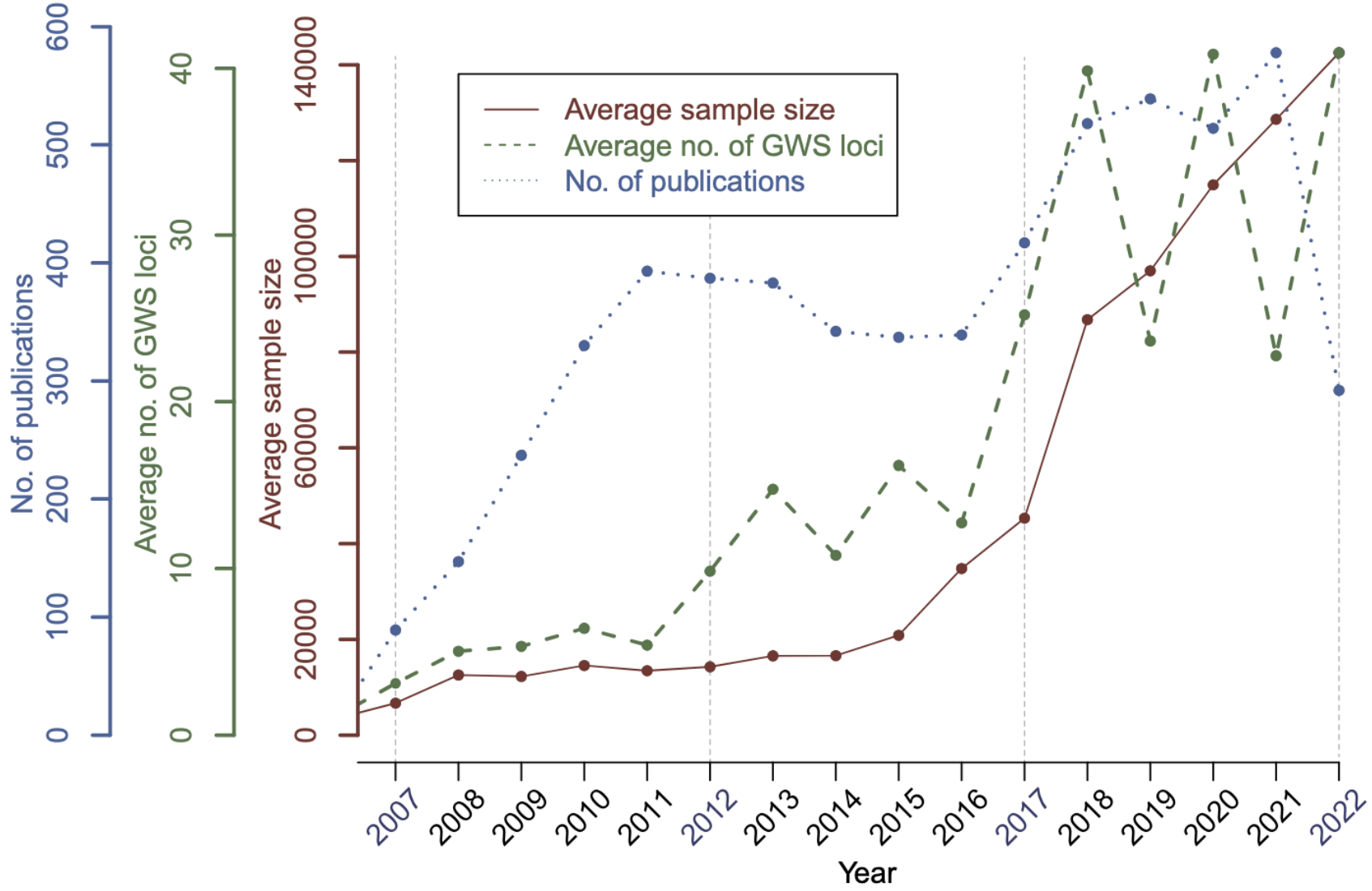



**Figure 2. GWAS SNP-Trait Discovery Timeline**

Data used for generating the graph were taken from the GWAS Catalogue.[10] SNPs and traits were selected according to the following filters. SNPs were selected with a p value $< 5 \times 10^{-8}$. For each trait with two or more selected SNPs, SNPs were removed if they had an LD $r^2 > 0.5$ (calculated from 1000 Genomes phase 3 data) with another selected SNPs and their p value was larger. For each year of discovery, only the top three traits and diseases with the largest number of SNPs are labeled in the circle.

Visscher P.M. et al. AJHG 2017.




Abdellaoui A, et. al. 15 years of GWAS discovery: Realizing the promise. AJHG. 2023.

**Figure 1. Average sample size and average number of genome-wide significant (GWS) loci per publication for each year during the 15 years history of GWAS discoveries**

The data were extracted from 5,771 GWAS publications that used a genome-wide genotyping array and shared their summary statistics on GWAS Catalog before November 8, 2022.

# GWAS Catalogue Diagram

## Published Genome-Wide Associations as of July 2019

p≤5X10-8 for 17 trait categories

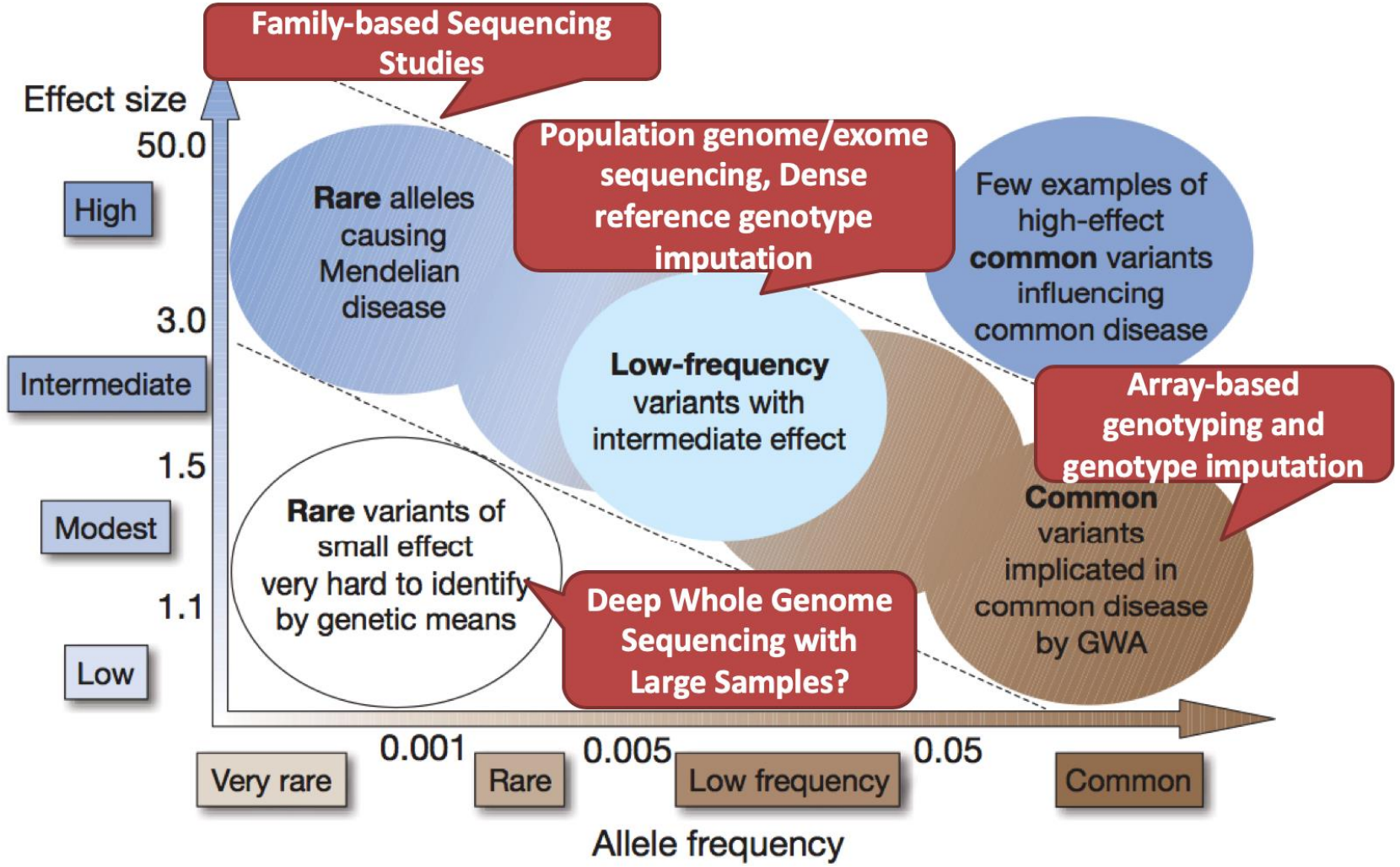

National Human Genome Research Institute

EMBL-EBI

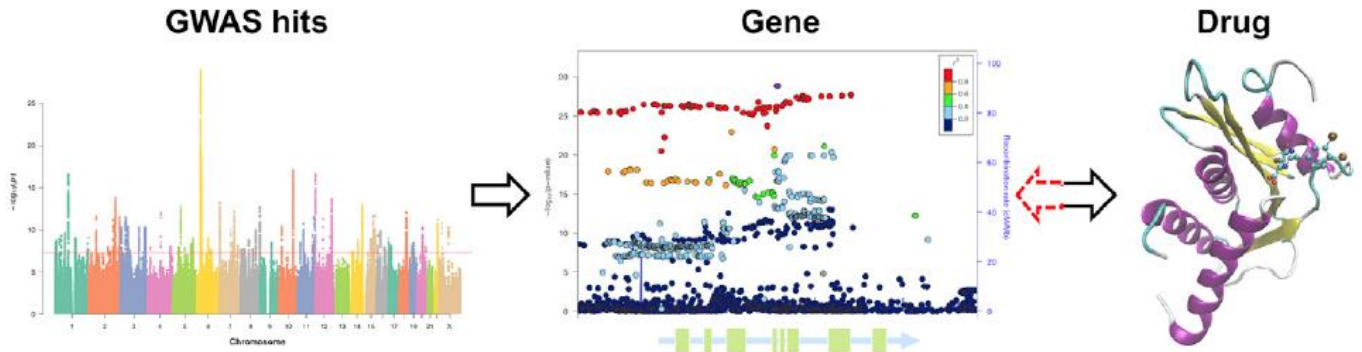
NHGRI-EBI GWAS Catalog
www.ebi.ac.uk/gwas

# Genetic architecture of complex traits

MAF Spectrum and Genetic Effect Sizes



# Example links between GWAS discoveries and drug developments




| Trait | Gene with GWAS hits | Known or candidate drug |
|---|---|---|
| **Type 2 Diabetes** | *SLC30A8/KCNJ11* | ZnT-8 antagonists/Glyburide |
| **Rheumatoid Arthritis** | *PADI4/IL6R* | BB-Cl-amidine/Tocilizumab |
| **Ankylosing Spondylitis(AS)** | *TNFR1/PTGER4/TYK2* | TNF-inhibitors/NSAIDs/fostamatinib |
| **Psoriasis(Ps)** | *IL23A* | Risankizumab |
| **Osteoporosis** | *RANKL/ESR1* | Denosumab/Raloxifene and HRT |
| **Schizophrenia** | *DRD2* | Anti-psychotics |
| **LDL cholesterol** | *HMGCR* | Pravastatin |
| **AS, Ps, Psoriatic Arthritis** | *IL12B* | Ustekinumab |

Visscher P.M. et al. AJHG 2017.

# GWAS Tools

- Michigan Imputation Server
  - https://imputationserver.sph.umich.edu

- GWAS Tool
  - PLINK: https://www.cog-genomics.org/plink/2.0/
  - EPACTS: https://genome.sph.umich.edu/wiki/EPACTS

- GWAS Results Visualization and Manhattan/LocusZoom Plot Tool
  - https://my.locuszoom.org/

# Outline of Next Lecture

- Quality Control
  - Genotype Quality Control
  - Sample Relatedness: Kingship Coefficient
- Population Stratification
  - Genomic Control Factor
  - Genotype Principal Components Analysis
  - Meta-analysis
- Linear Mixed Model (LMM)