Lecture 6: Introduction to Genome-wide Association Studies

10/05/2021

Jingjing Yang, PhD
Assistant Professor of Human Genetics and Biostatistics
Office: Whitehead Biomedical Research Building, Suite 305K
Email: jingjing.yang@emory.edu

Outline

1. Genetic variants and GWAS
2. Linkage disequilibrium
3. Statistical methods for single variant GWAS
4. Population stratification
5. Meta-analysis methods

Genetic terms

- Genetic markers, variants, e.g., SNPs, Indel (Insertion, Deletion), Copy number variation (CNV), Structure variation (SV, ¿1KB)
- Minor allele frequency (MAF)
- Common Variants: Genetic variants (e.g., SNPs) with MAF$> 5\%$.
- Genes
- Phenotypes or traits
- Genotype, quantified as values in $[0, 2]$ or $0, 1, 2$
- Linkage disequilibrium (LD)

See videos about introductions to the basic principles of genetics, e.g., genes, SNPs, phenotypes, as provided by 23&me: https://www.23andme.com/gen101/

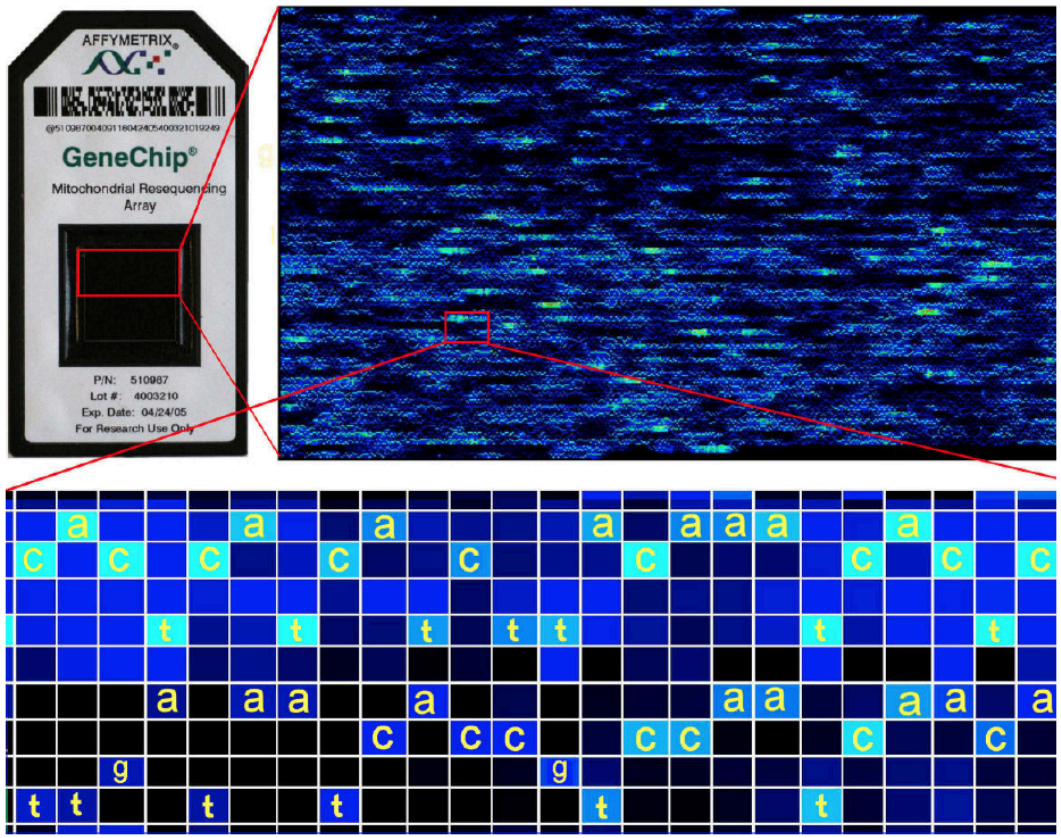
Studies of Human Genetics

- 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)

DNA Microarrays (CHIP)

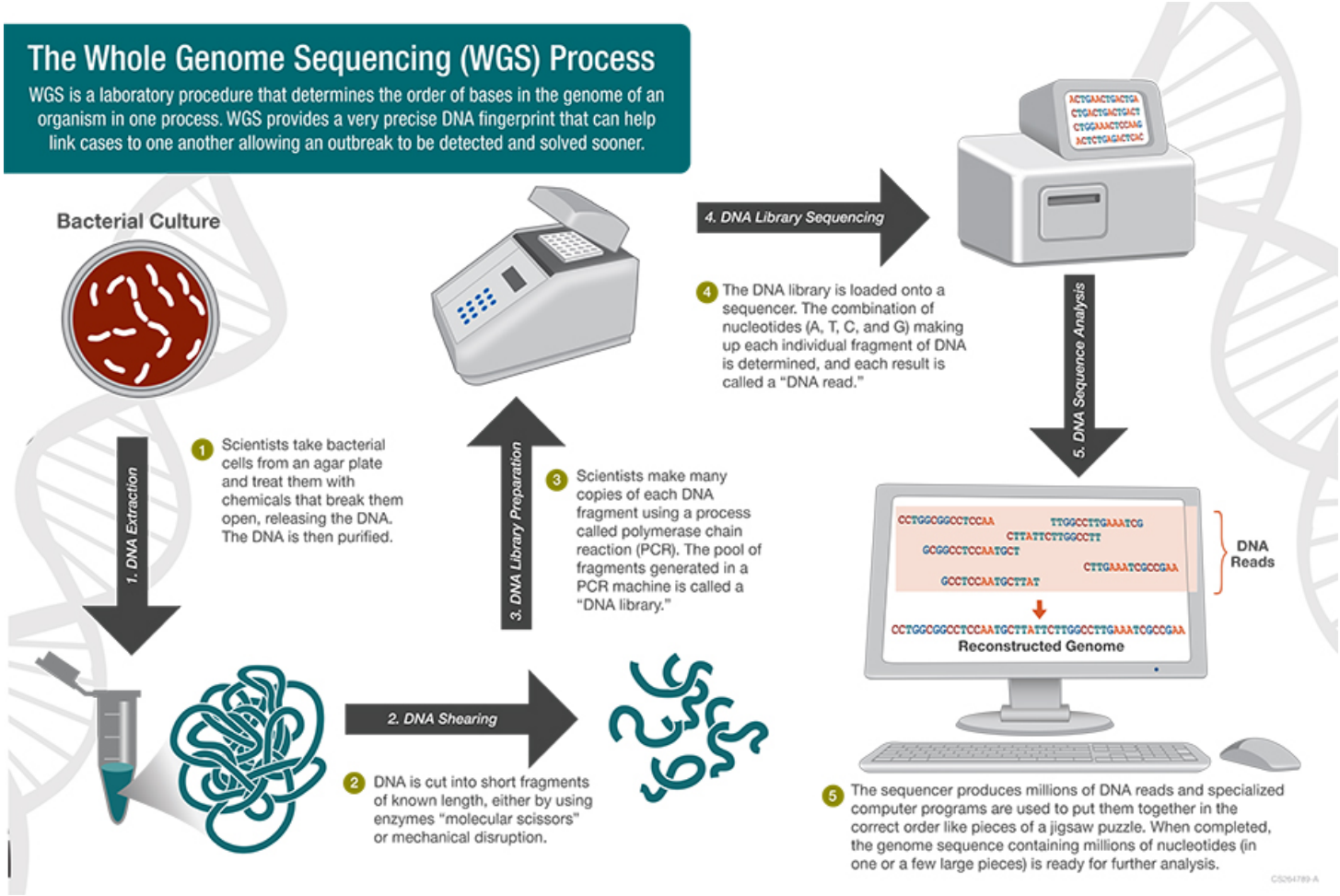
- Microarrays (Illumina and Affymetrix) are used to genotype $0.5M - 1M$ SNPs across the whole genome
- LD-information of the HapMap project has been incorporated so that the chips provide adequate coverage of the entire human genome for most ethnicities.
- Customize chips with densely spaced SNPs within known genes regions



Whole Genome Sequencing

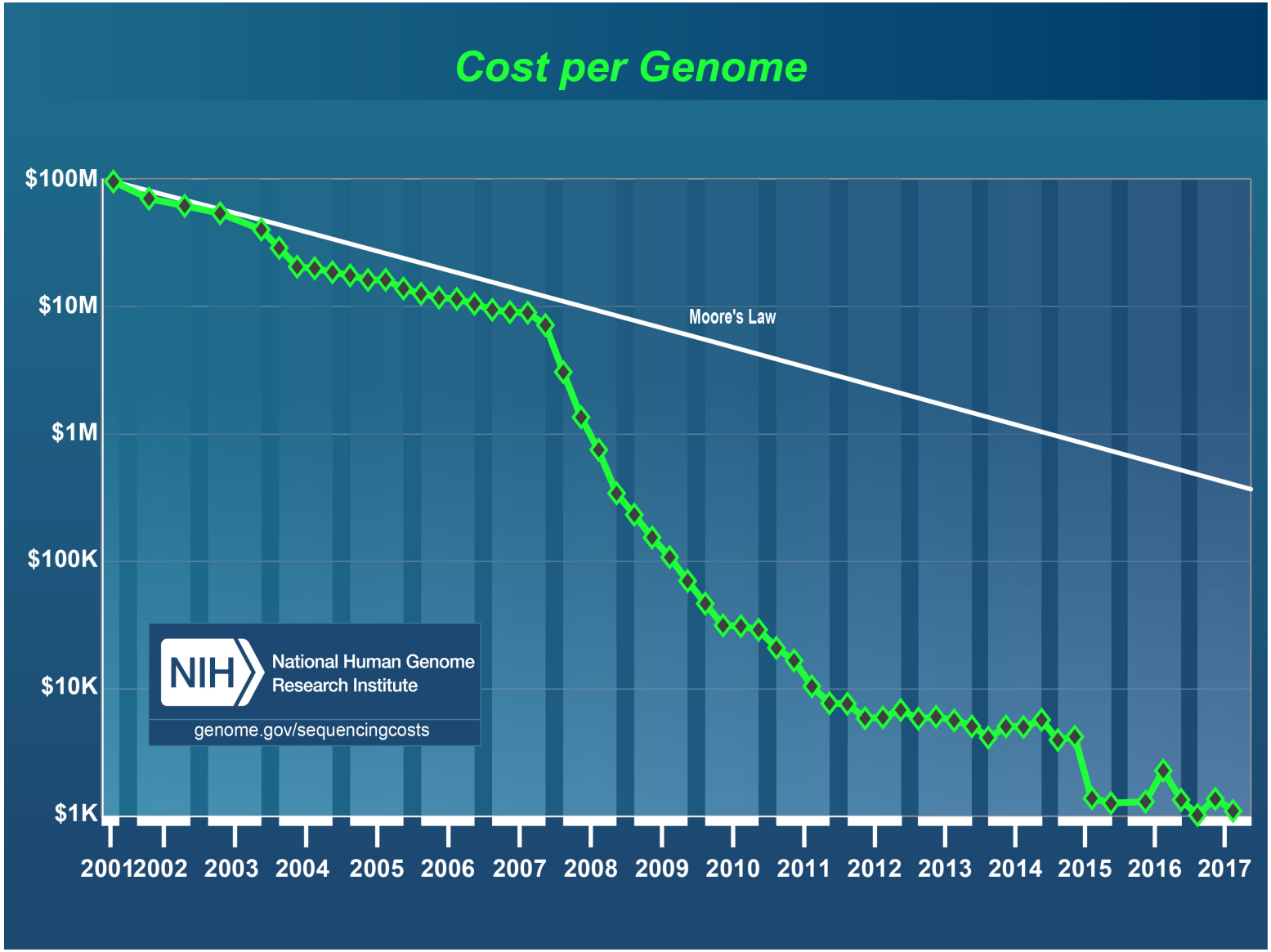
Introduction video of Illumina Sequencing technology

https://www.youtube.com/watch?v=fCd6B5HRaZ8



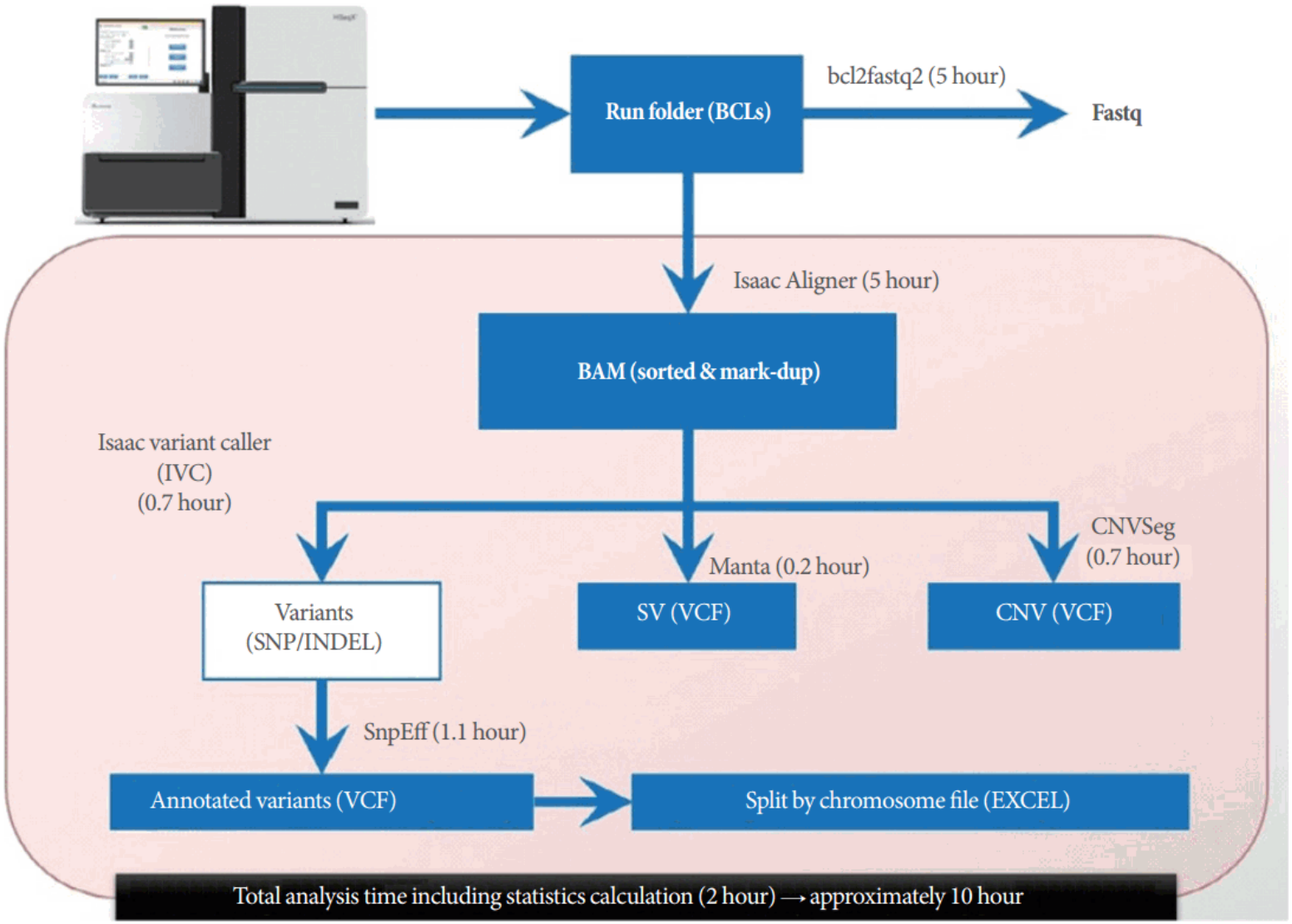
https://www.cdc.gov/pulsenet/pathogens/wgs.html

21st Century Sequencing Costs



http://genome.gov/sequencingcostsdata

WGS Analysis Workflow



DOI:10.5213/inj.1632742.371

1000 Genome Project 2008 - 2015

GOAL: Find most genetic variants with MAF $\geq 1\%$ in populations across the world.

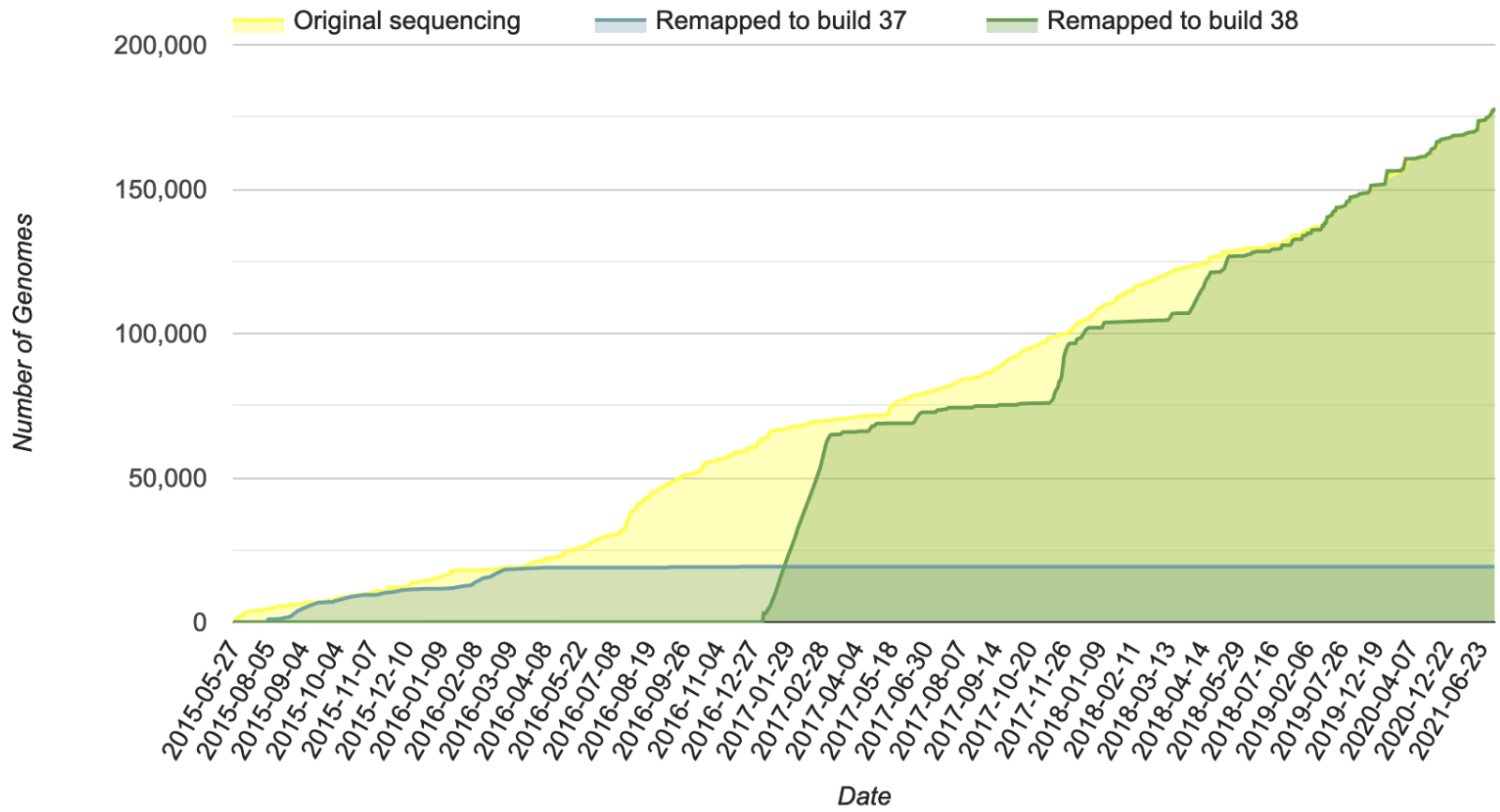
- First project to sequence the genomes of a large number of people (2,504 samples)
- Largest public catalogue of human variation and genotype data, http://www.internationalgenome.org/
- 26 Different populations under 5 super populations
  - AFR: African
  - AMR: Admixed American
  - EAS: East Asian
  - EUR: European
  - SAS: South Asian

TOPMed WGS Project 2015 - Present

- NIH National Heart, Lung, and Blood Institute (NHLBI) sponsored the Trans-Omics for Precision Medicine (TOPMed) program https://topmed.nhlbi.nih.gov/
- Deep (30x coverage) whole genome sequencing for all of the collected samples from ongoing disease-specific research projects
- WGS data are generated by seven sequencing centers
- University of Washington group is designated as the Data Coordinating Center (DCC) and will coordinate phenotype information
- University of Michigan group is designated as the Informatics Research Center (IRC) with responsibility for creating a unified variant call set
- The sequence and genotype data will be deposited to dbGaP https://www.ncbi.nlm.nih.gov/gap

Summary of total sequencing progress over time

This chart shows a summary of the total genomes received by IRC over time



Samples that have completed QC: 178,156 (as of 6/23/2021).

$> 10^{16}$ sequenced bases, $> 100\times$ more data than the 1000 Genome Project.

http://nhlbi.sph.umich.edu

10^{16} sequenced bases

US corn production in 2014: 1.3 x 10^{15} kernels

Image: Patrick Porter @ Smug Mug

Table 1 Number of variants in 40,722 unrelated individuals in TOPMed

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

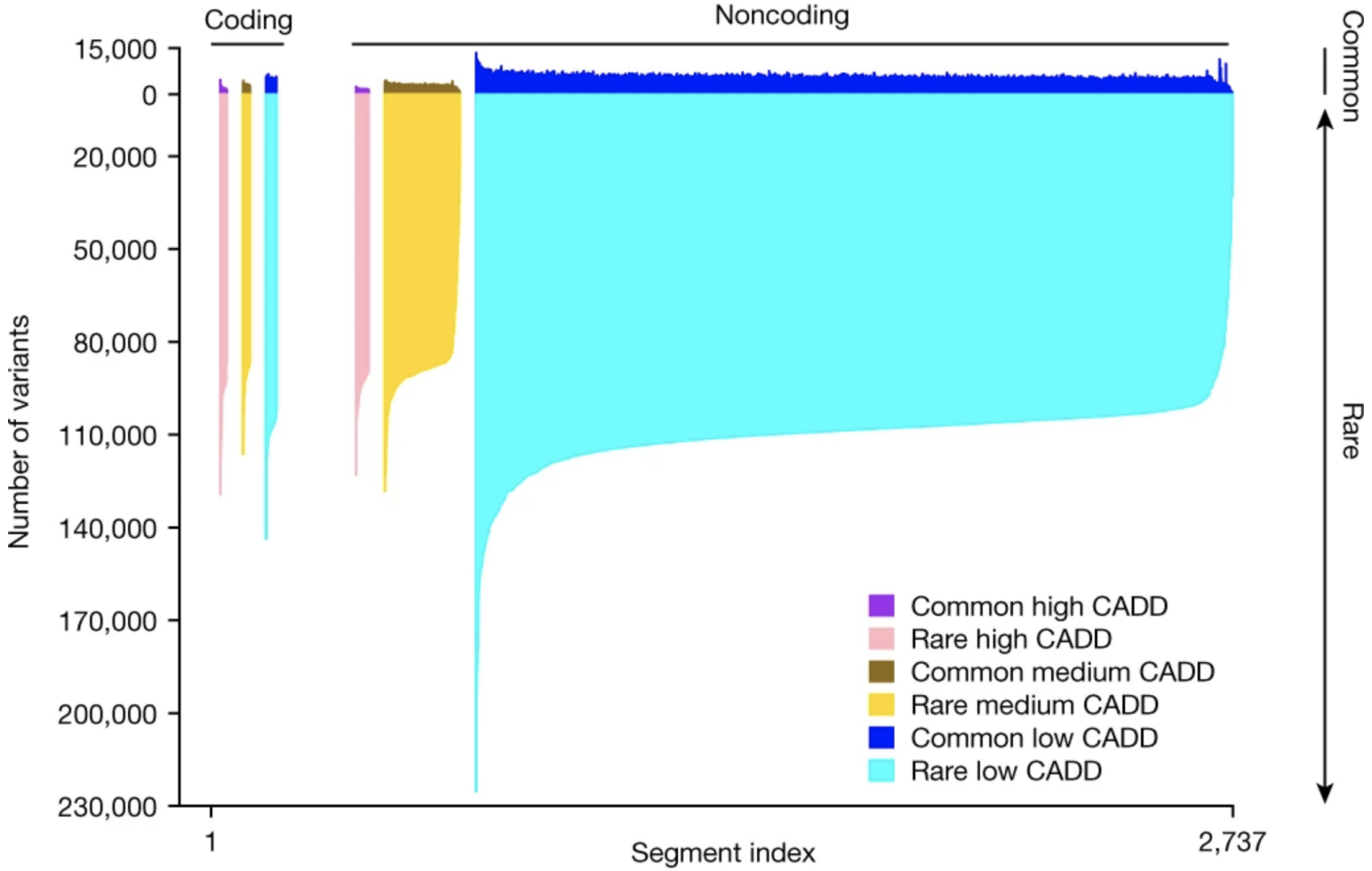
	All unrelated individuals (n = 40,722)		Per individual			
	Total	Singletons (%)	Average	5th percentile	Median	95th percentile
Total variants	**384,127,954**	**203,994,740 (53)**	**3,748,599**	**3,516,166**	**3,563,978**	**4,359,661**
SNVs	357,043,141	189,429,596 (53)	3,553,423	3,335,442	3,380,462	4,125,740
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.

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



Taliun D. et. al., Nature, 2021.

Genome-Wide Association Study (GWAS)

- **Key Goals of GWAS**
  - Test associations between each genetic variant or gene across the whole genome and the phenotype of interest
  - Understand the biological function of these associated loci (Challenging)
  - Germ line risk prediction for diseases
- **Rationale**
  - Most traits and diseases have complex genetic etiology: Many genetic variants make small contributions (Polygenic)
  - Significant genetic variants could be just correlated (in LD) with the true causal ones
  - Large sample size and whole genome sequencing data might be needed to ensure enough power for identifying risk variants or genes

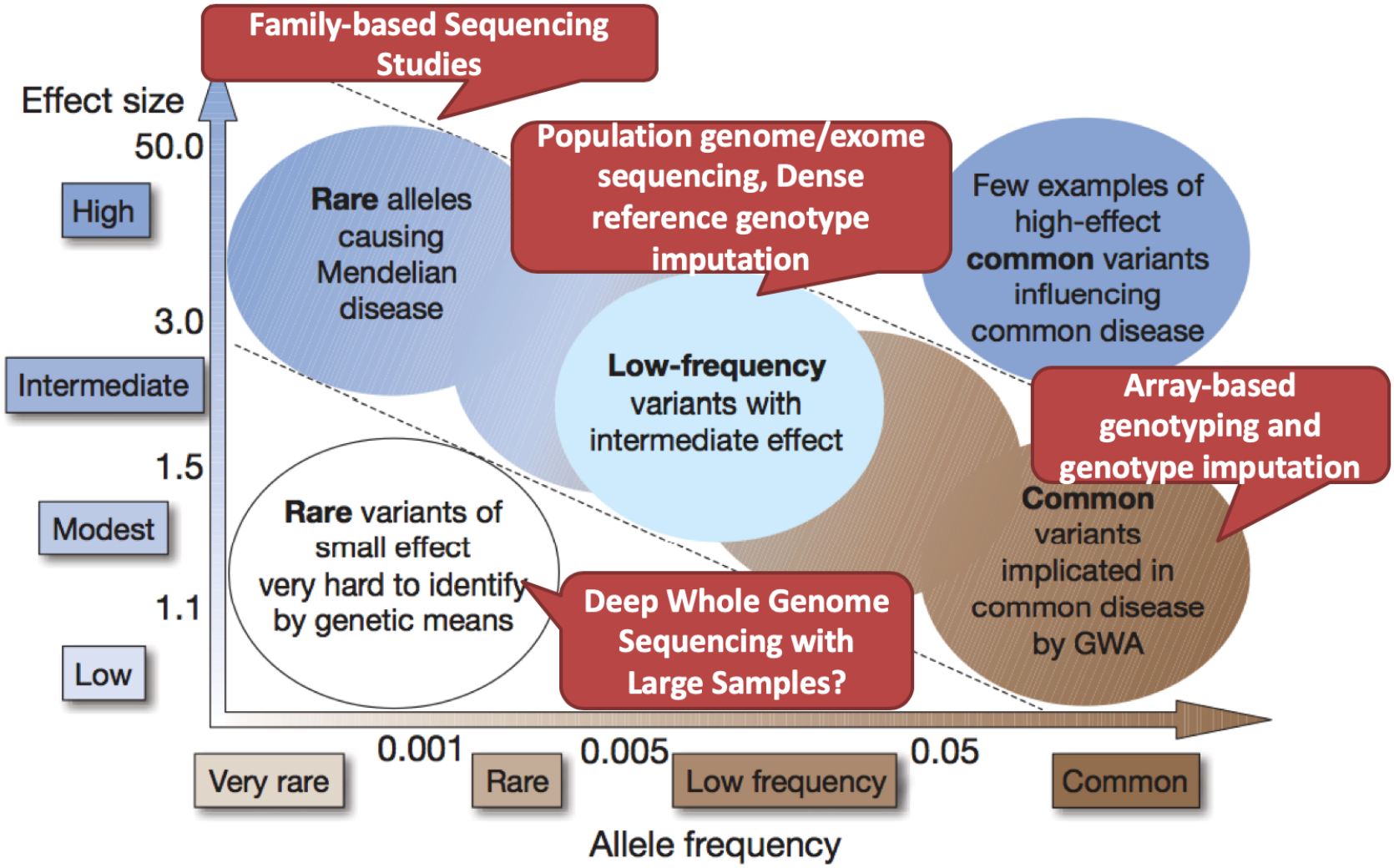
Types of Association Studies

- Quantitative and Dichotomous (i.e., Case-control studies) traits
- Family-based association study
- Population-based association study (our main focus in this lecture)

GWAS Procedure

- Quality control (QC) of the study dataset: missing rate, HWE p-value, ancestry
- Choose a model/test for the phenotype of interest (e.g., linear regression model for quantitative traits, logistic regression model for dichotomous traits)
- Significance level $\alpha = 5 \times 10^{-8}$
- Annotate biological functions and nearby genes of the significant SNPs
- Investigate the biological functions of significant SNPs or genes

Genetic architecture of complex traits



Sources of Association

- Causal association — best
  - Genetic marker alleles influence susceptibility
- Linkage disequilibrium — useful
  - Genetic marker alleles associated with other nearby alleles that influence susceptibility
- Population stratification — misleading
  - Genetic marker is unrelated to disease alleles

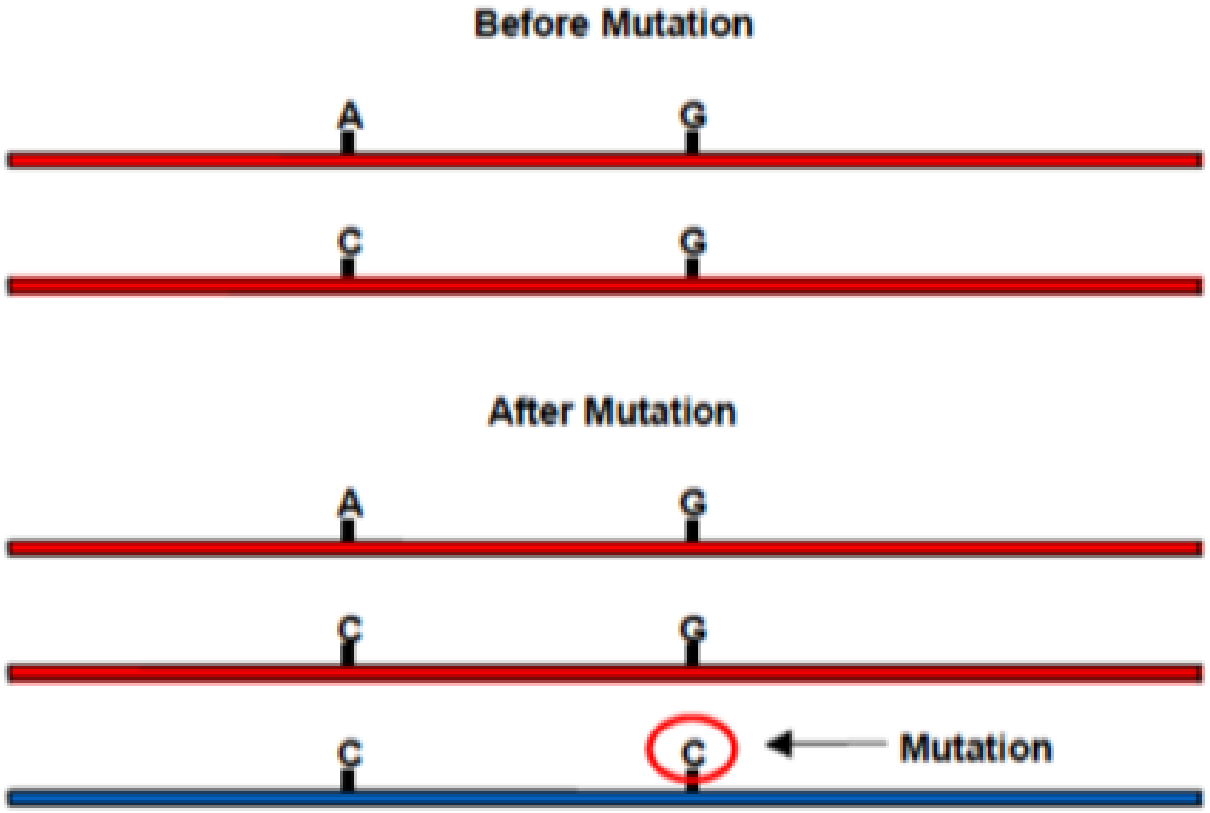
Linkage Disequilibrium (LD)

- **Linkage Disequilibrium (LD)** is the non-random association of alleles at different loci in a given population..
- Nearby markers are likely to be correlated, why?
- Origin of LD?

Linkage Disequilibrium (LD)

- Consider the history of two neighboring single nucleotide polymorphism (SNP)
- SNPs exist today arose through ancient mutation events…

- One SNP arose first and then the other …



Linkage Disequilibrium (LD)

- Recombination generates new arrangements for the ancestral alleles

Linkage Disequilibrium (LD)

- Chromosomes are mosaics

- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection

- Combinations of alleles at very close markers reflect ancestral haplotypes

Linkage Disequilibrium (LD)

With observed frequency p_A and p_B for two alleles A and B at two markers and frequency p_{AB} for alleles A and B appear together:

$$D_{AB} = p_{AB} - p_A p_B$$

- Define a random variable X_A to be the number of allele A present at the first marker, $0, 1, 2$
- Define a random variable X_B to be the number of allele B present at the second marker, $0, 1, 2$
- Correlation between these two random variables is given by

$$r_{AB} = \frac{Cov(X_A, X_B)}{\sqrt{Var(X_A)Var(X_B)}} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_b)}}$$

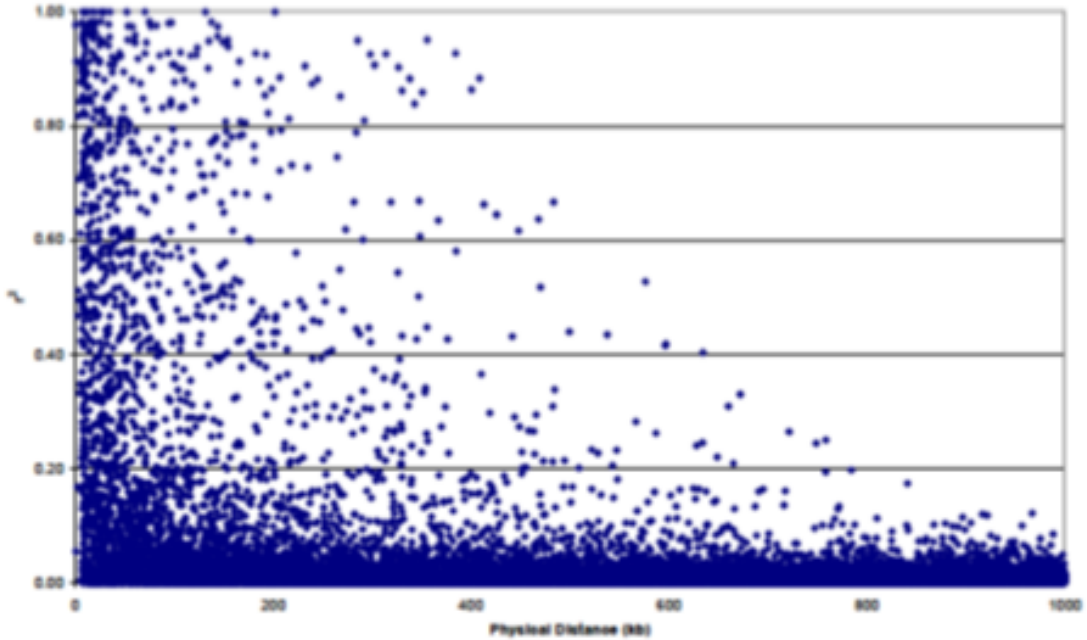
- r^2 between these two random variables is given by

$$r^2_{AB} = \frac{D^2_{AB}}{p_A(1-p_A)p_B(1-p_b)}$$

Linkage Disequilibrium (LD)

Genotype data for multiple samples from a population

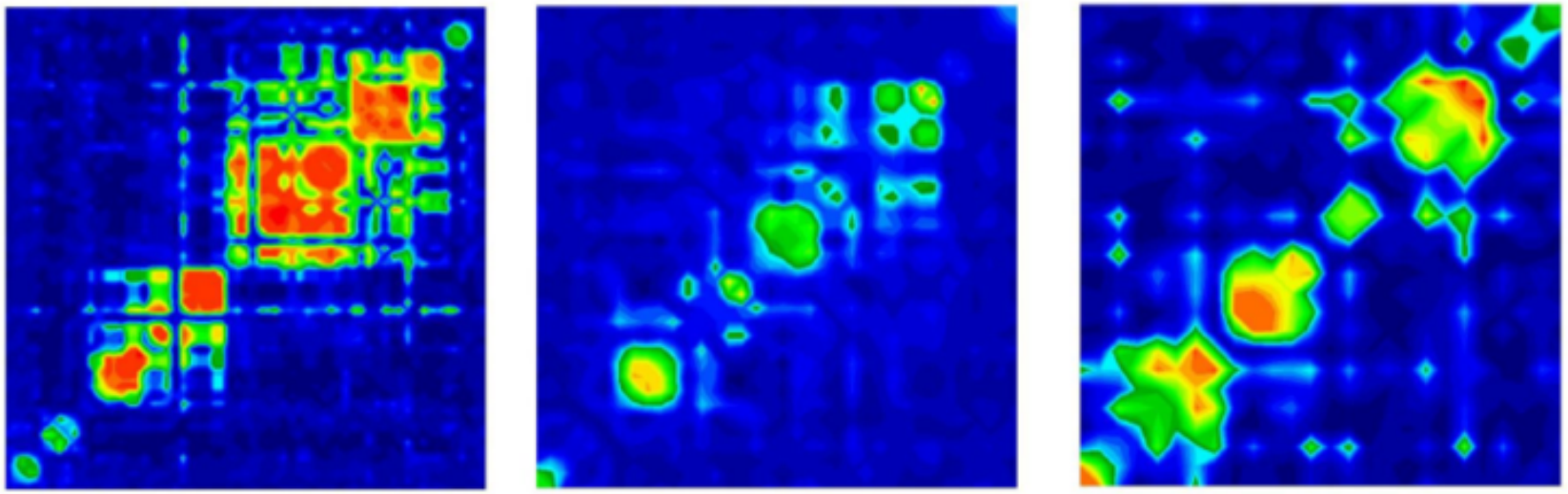
- SNP1: x1 = (0, 1, 2, 1, 0, 0, …)
- SNP2: x2 = (1, 1, 2, 0, 0, 0, …)
- r^2 = (correlation(x1, x2))^2
- Raw r^2 from CHR22



Dawson et al, *Nature*, 2002

Linkage Disequilibrium (LD)

Linkage Disequilibrium in Three Regions



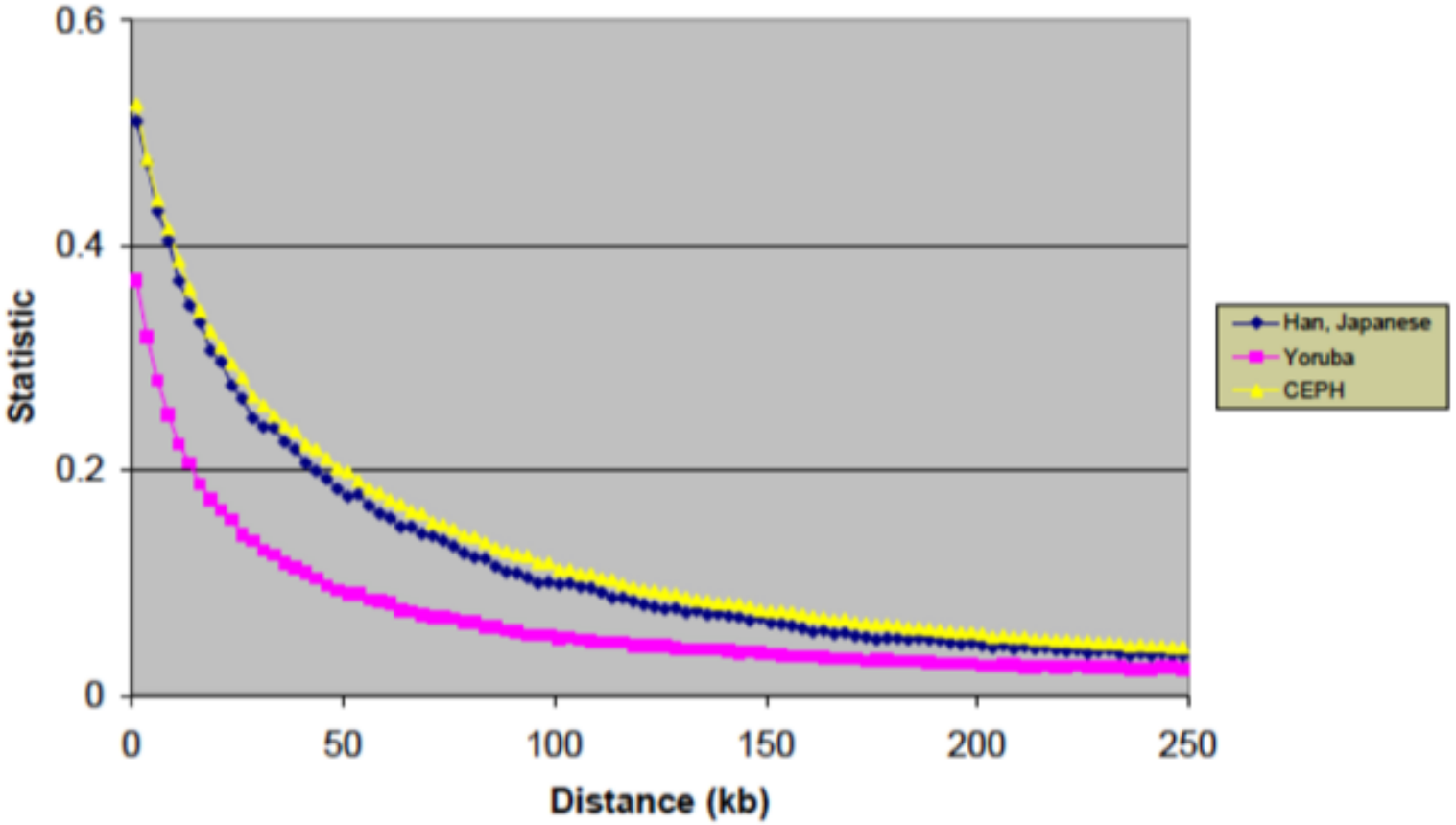
2q13
(63 markers)

13q13
(38 markers)

14q11
(26 markers)

Abecasis et al, *Am J Hum Genet*, 2001

Comparing Populations ...

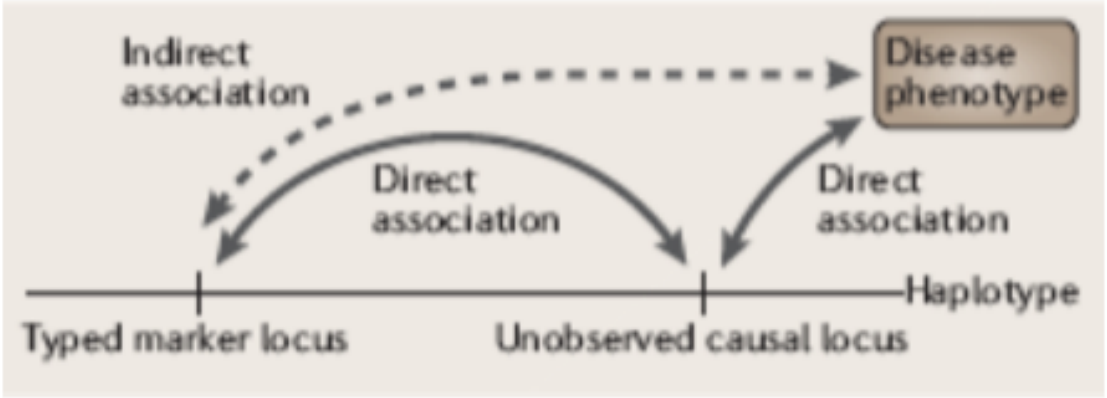


LD extends further in CEPH and the Han/Japanese than in the Yoruba

International HapMap Consortium, *Nature*, 2005

Why LD is Important for Association Studies?

- SNPs in strong LD with disease variant are good proxies for disease variant



Balding, 2006

- If testing (unobservable) disease variant for association would yield chi-squared statistic X^2, testing variant in LD yields r^2X^2
- Model LD among multiple markers in joint tests to improve power

Statistical Methods for Single Variant Association Test

1. Contingency table based tests (only for dichotomous traits)
   (a) Genotypic Association test (2-*df* test)
   (b) Genotypic Association test with dominant/recessive disease models
   (c) Allelic Association test
2. Regression based tests
   (a) Logistic regression based tests for dichotomous traits
   (b) Linear regression based tests for quantitative traits

Genotypic Association Test

- Compare genotype frequencies in cases and controls in a 2×3 table
- Not assuming any specific disease model

	AA	Aa	aa	Total
Case	n_{10}	n_{11}	n_{12}	$n_{1.}$
Control	n_{00}	n_{01}	n_{02}	$n_{0.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n

The genotype/codominant test: D – disease status; G – genotype

$$H_0 : \Pr(D = 1|Geno = \text{AA}) = \Pr(D = 1|Geno = \text{Aa}) = \Pr(D = 1|Geno = \text{aa})$$

H_1 : At least one inequality holds

The standard $2\ df$ Pearson χ^2 test of independence for a 2×3 table is:

$$X_G^2 = \sum_{i=0,1} \sum_{j=0,1,2} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 2$$

– $O_{ij} = n_{ij}$: observed count in the cell

– $E_{ij} = n_{i.} n_{.j} / n$: expected count under independence: $np_{D=i} p_{G=j} = n(n_{i.}/n)(n_{.j}/n)$

Dominant or Recessive Disease Models

- Compare frequencies of AA or Aa with aa in cases and controls in a 2×2 table
- Assume dominant or recessive Mendelian disease model
- More powerful than genotype test if the disease model is true

With dominant disease model:

	AA or Aa	aa	Total
Case	$n_{10} + n_{11}$	n_{12}	$n_{1.}$
Control	$n_{00} + n_{01}$	n_{02}	$n_{0.}$
Total	$n_{.0} + n_{.1}$	$n_{.2}$	n

$$H_0 : \Pr(D = 1|\text{AA}) = \Pr(D = 1|\text{Aa}) = \Pr(D = 1|\text{aa})$$
$$H_1 : \Pr(D = 1|\text{AA or Aa}) \neq \Pr(D = 1|\text{aa})$$

The standard $1\ df$ Pearson χ^2 test of independence for a 2×2 table is:

$$X_D^2 = \sum_{i=0,1} \sum_{j=0,1} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 1$$

How to obtain E_{ij}?

Allelic Association Test

- Compare frequencies of alleles A and a in cases and controls in a 2×2 table
- **Assume additive disease model**: the risk associated with the heterozygote genotype is intermediate between the two homozygotes. (mostly used model)
- Assume HWE: allele frequencies in a population will remain constant from generation to generation, with random mating and in the absence of other evolutionary influences (selection, mutation, genetic drift)
- The allele test is the most powerful test for additive model.

	A	a	Total
Case	$n_{1A} = 2n_{10} + n_{11}$	$n_{1a} = n_{11} + 2n_{12}$	$2n_{1.}$
Control	$n_{0A} = 2n_{00} + n_{01}$	$n_{0a} = n_{01} + 2n_{02}$	$2n_{0.}$
Total	$n_{.A} = 2n_{.0} + n_{.1}$	$n_{.a} = n_{.1} + 2n_{.2}$	$2n$

The allele test:

$$H_0 : \Pr(\mathrm{A}|D = 1) = \Pr(\mathrm{A}|D = 0)$$

The standard $1\ df$ Pearson χ^2 test of independence for a 2×2 table is:

$$X_L^2 = \sum_{i=0,1} \sum_{j=0,1} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 1$$

Measure of Association Strength: Odds Ratio

	Exposed (E)	Not Exposed ($\bar{E}$)
Case (D)	a	b
Control ($\bar{D}$)	c	d

Odds ratio:

$$\begin{aligned} OR &= \frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(\bar{D}|\bar{E})} \\ &= \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})} \\ &= ad/bc \end{aligned}$$

- Exposed = carry certain genotype
- Counts pertain to individuals, not alleles.

Genotype Model ($\bar{E}$=aa)

	AA	Aa	aa
Case	n_{10}	n_{11}	n_{12}
Control	n_{00}	n_{01}	n_{02}

$OR_{het} = (n_{11}n_{02})/(n_{01}n_{12})$

$OR_{hom} = (n_{10}n_{02})/(n_{00}n_{12})$

Dominant Model ($\bar{E}$=aa)

	AA or Aa	aa
Case	$n_{10} + n_{11}$	n_{12}
Control	$n_{00} + n_{01}$	n_{02}

$OR_D = [(n_{10} + n_{11})n_{02}]/[(n_{00} + n_{01})n_{12}]$

Allele Model ($\bar{E}$=a)

	A	a
Case	$2n_{10} + n_{11}$	$n_{11} + 2n_{12}$
Control	$2n_{00} + n_{01}$	$n_{01} + 2n_{02}$

$OR_L = [(2n_{10} + n_{11})(n_{01} + 2n_{02})]/[(2n_{00} + n_{01})(n_{11} + 2n_{12})]$

Trend Model

estimate OR by maximum likelihood

OR_T: logistic regression

Calculating Confidence Interval for Odds Ratio

In large samples and when OR is estimated from the contingent table, $\log(\widehat{OR})$ is approximately normally distributed, with estimated variance

$$\widehat{\text{Var}}[\log(OR)] \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d},$$

where a, b, c, d are the cells contributing to the estimation of OR.

A $(1-\alpha)100$th confidence interval for the population OR :

$$\exp^{\log(\widehat{OR}) \pm z_{(1-\alpha/2)} \sqrt{\widehat{\text{Var}}[\log(OR)]}}$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)100$th percentile of the standard normal.

- 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 β_0 is the intercept, α is the coefficient for covariates C, and β_1 is the genetic effect-size (i.e., log(Odds-Ratio)).

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Logistic Regression vs. Contingency Table

- Likelihood ratio test of logistic regression $\approx$ chi-square tests for appropriate contingency tables.
- The estimated coefficients $=$ log of the corresponding odds ratios.
- For the additive model, the trend test $\approx$ likelihood ratio test from logistic regression with additive coding for X.
- Because the logistic regression operate on variables defined for individuals, not chromosomes, there is no underlying assumption about HWE.

Extension to other phenotypes:

- The phenotype Y can be a count or a continuous outcome.
- The generalized linear model is given by

$$g[\mathrm{E}(Y|X)] = \beta_0 + \alpha C + \beta_1 X$$

  where $g(.)$ is a link function.

-

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Additional Factors Important for GWAS: batch effects, population stratification

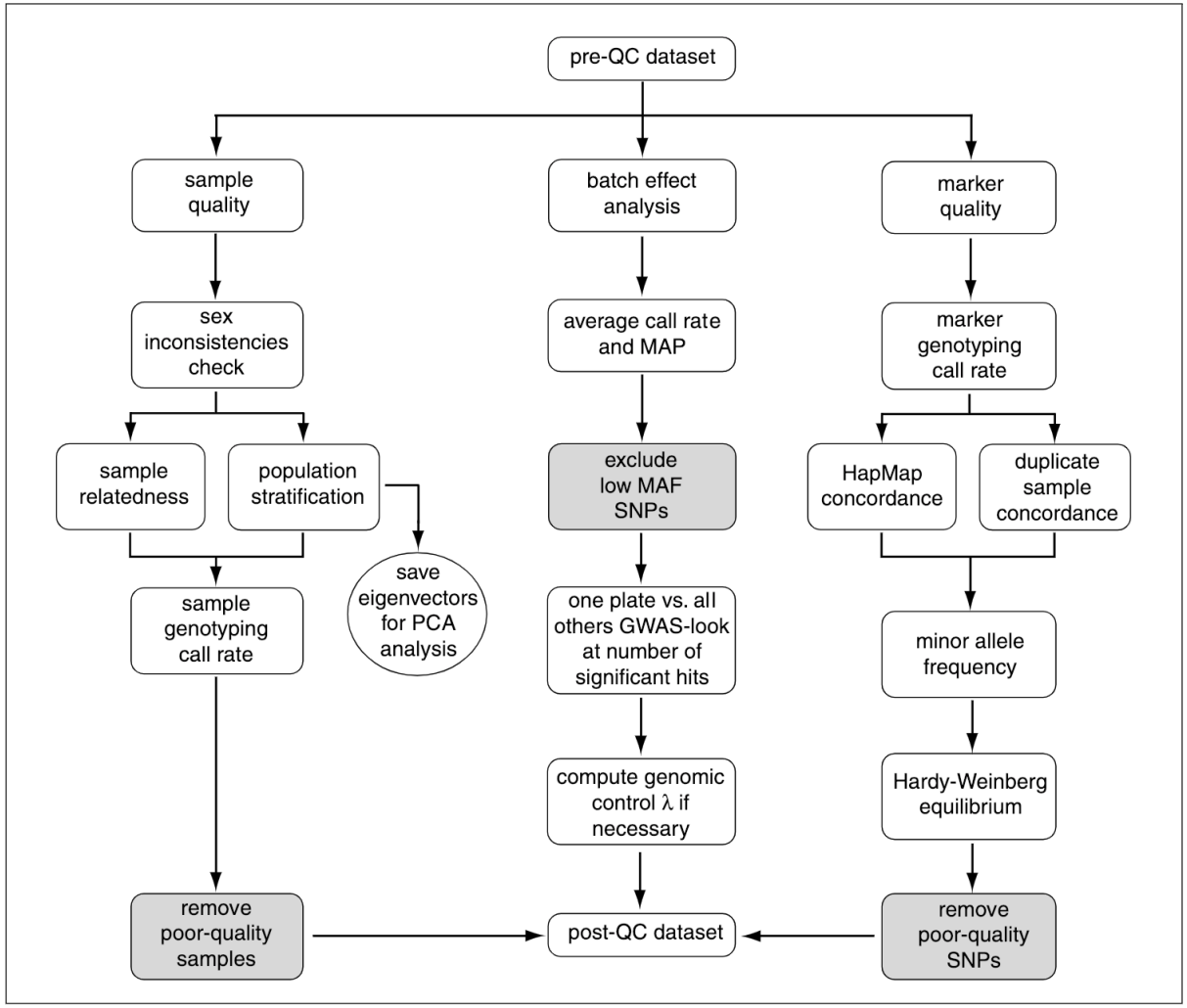


Figure 1.19.1 A flowchart overview of the entire GWAS QC process. Each topic is discussed in detail in the corresponding section in the text. Squares represent steps, ovals represent input or output data, and trapezoids represent filtering of data.

DOI:10.1002/0471142905.hg0119s68

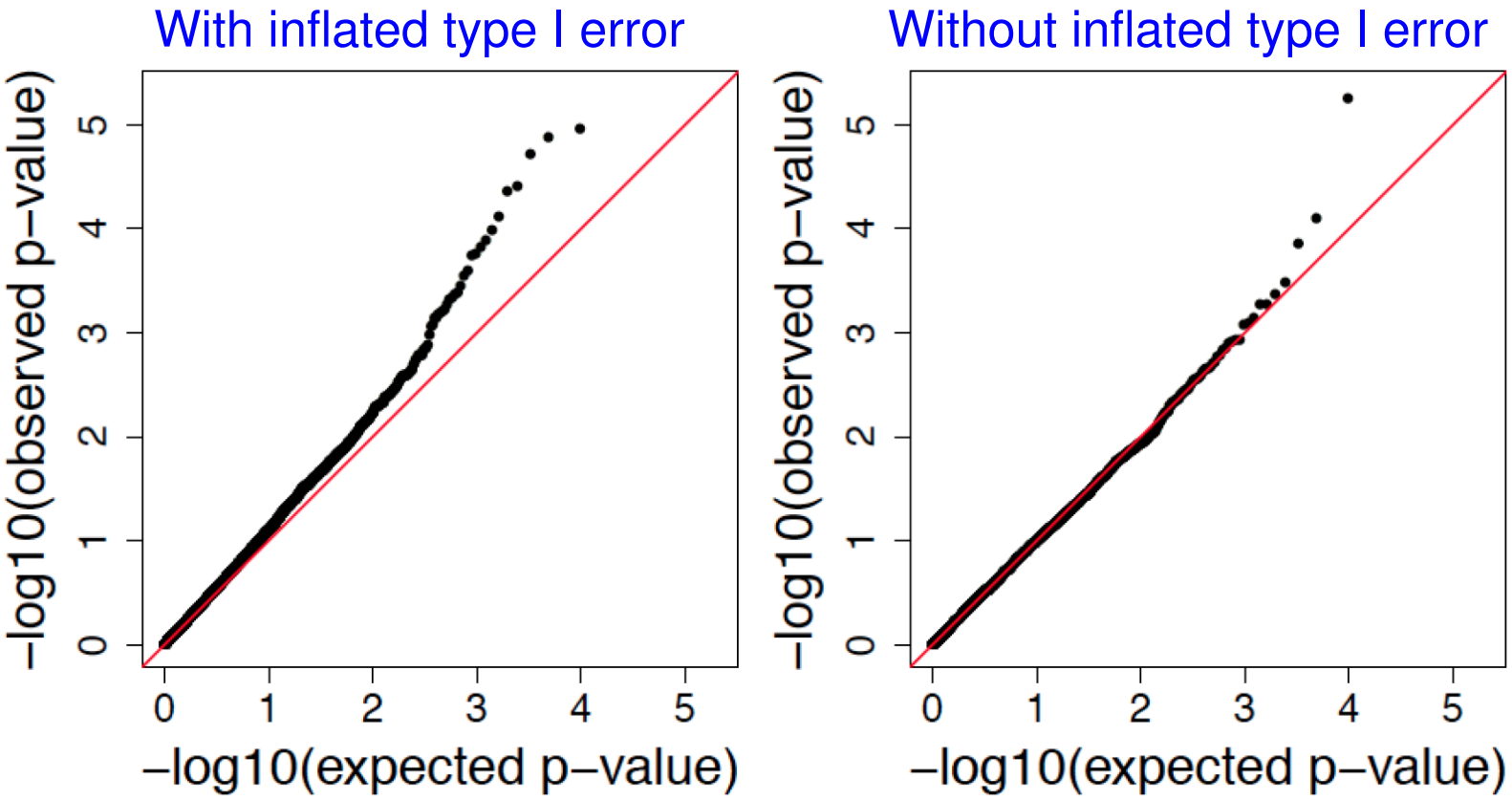
Quantile-Quantile (QQ) Plot

– Obtained $-\log 10$(p-values) from GWAS

– Sort all $-\log 10$(p-values) from most significant to least

– Pair these with the expected values of order statistics of a Uniform$(0, 1)$ distribution

– Under NULL hypothesis (no association), p-values follow a Uniform$(0, 1)$ distribution

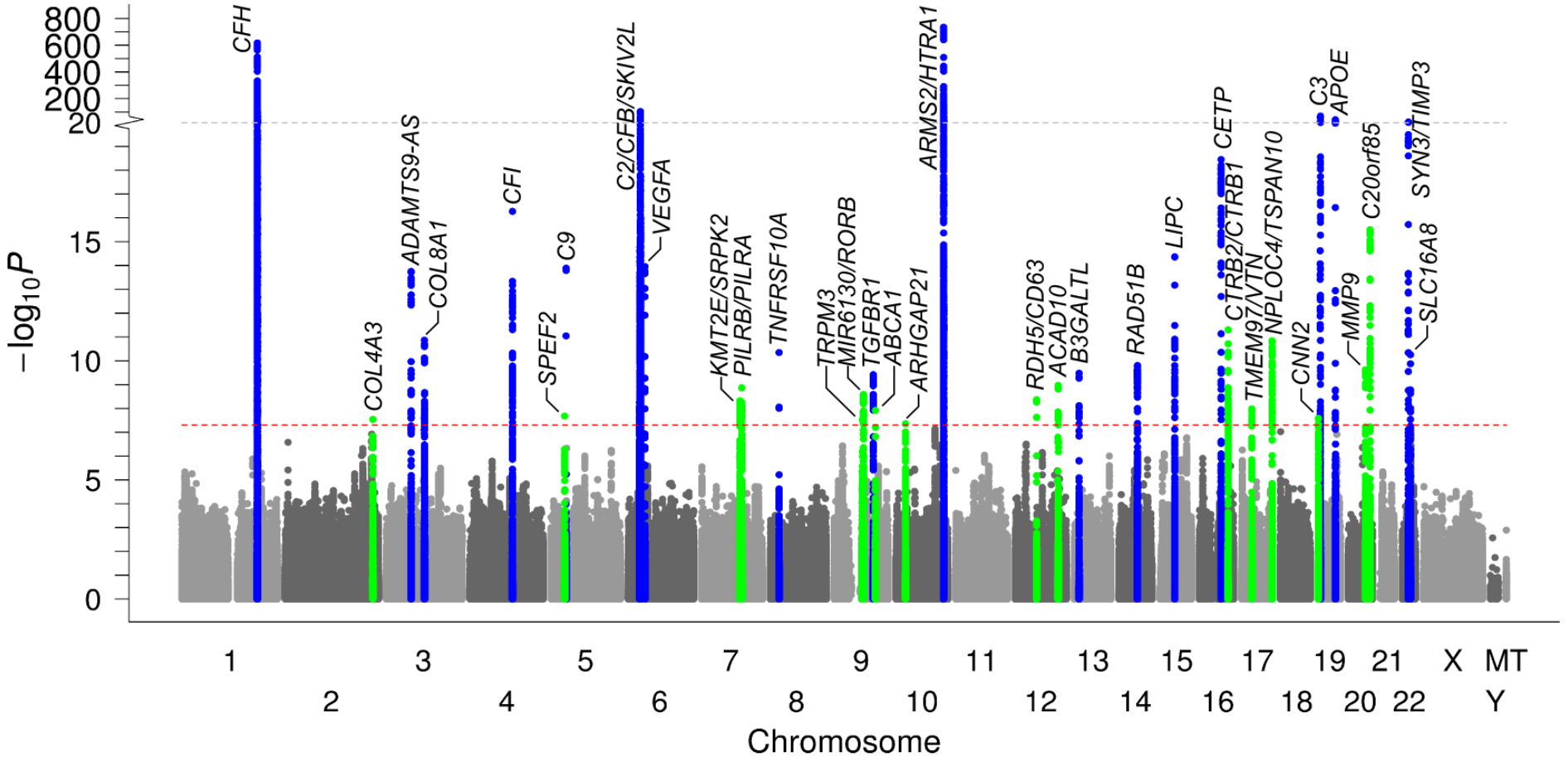


Visualize GWAS Results: Manhattan Plot

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

– Visualize signal peaks

GWAS Results



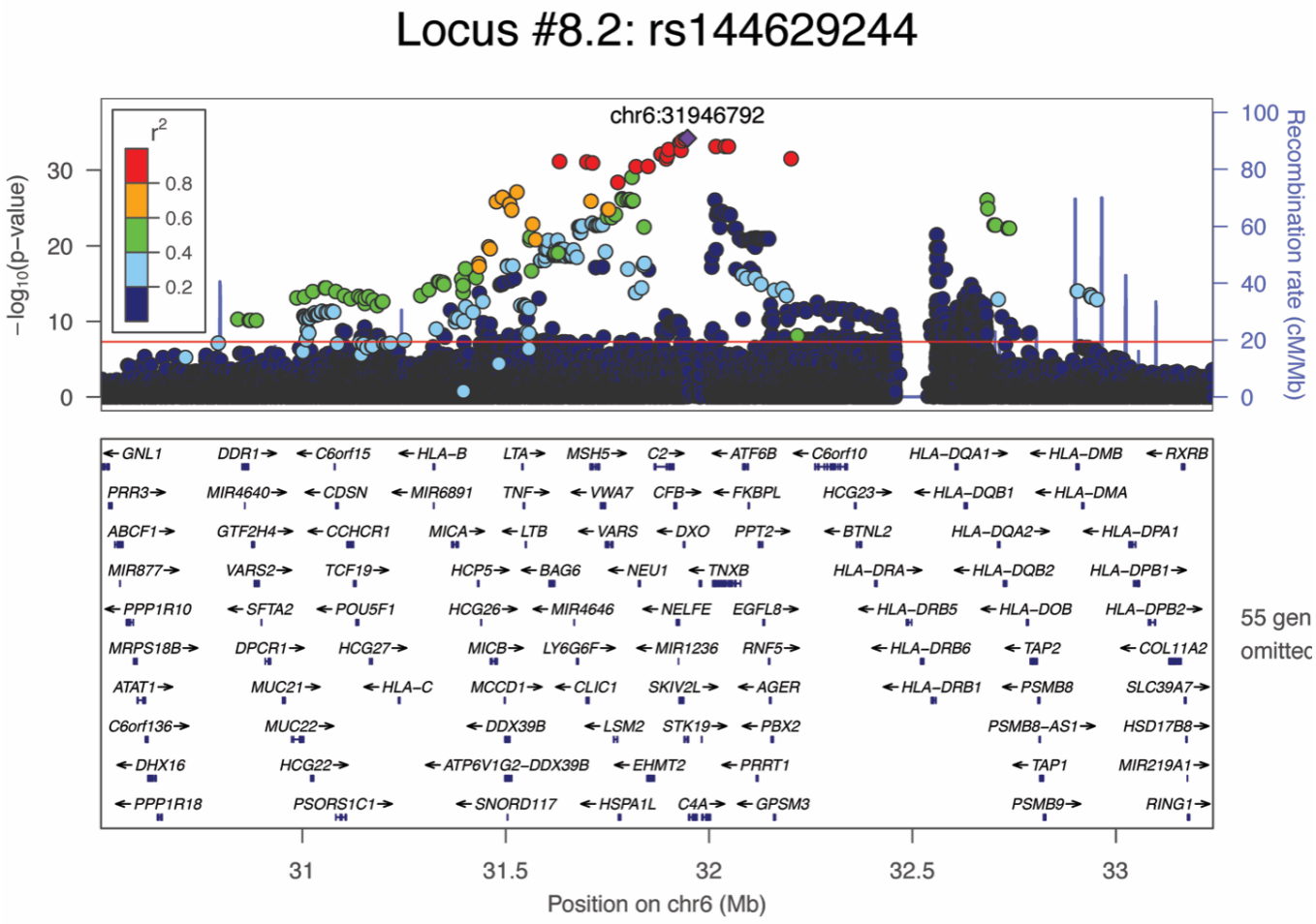
18 known AMD loci and **16 novel AMD loci**

Visualize GWAS Results: Locus Zoom Plot

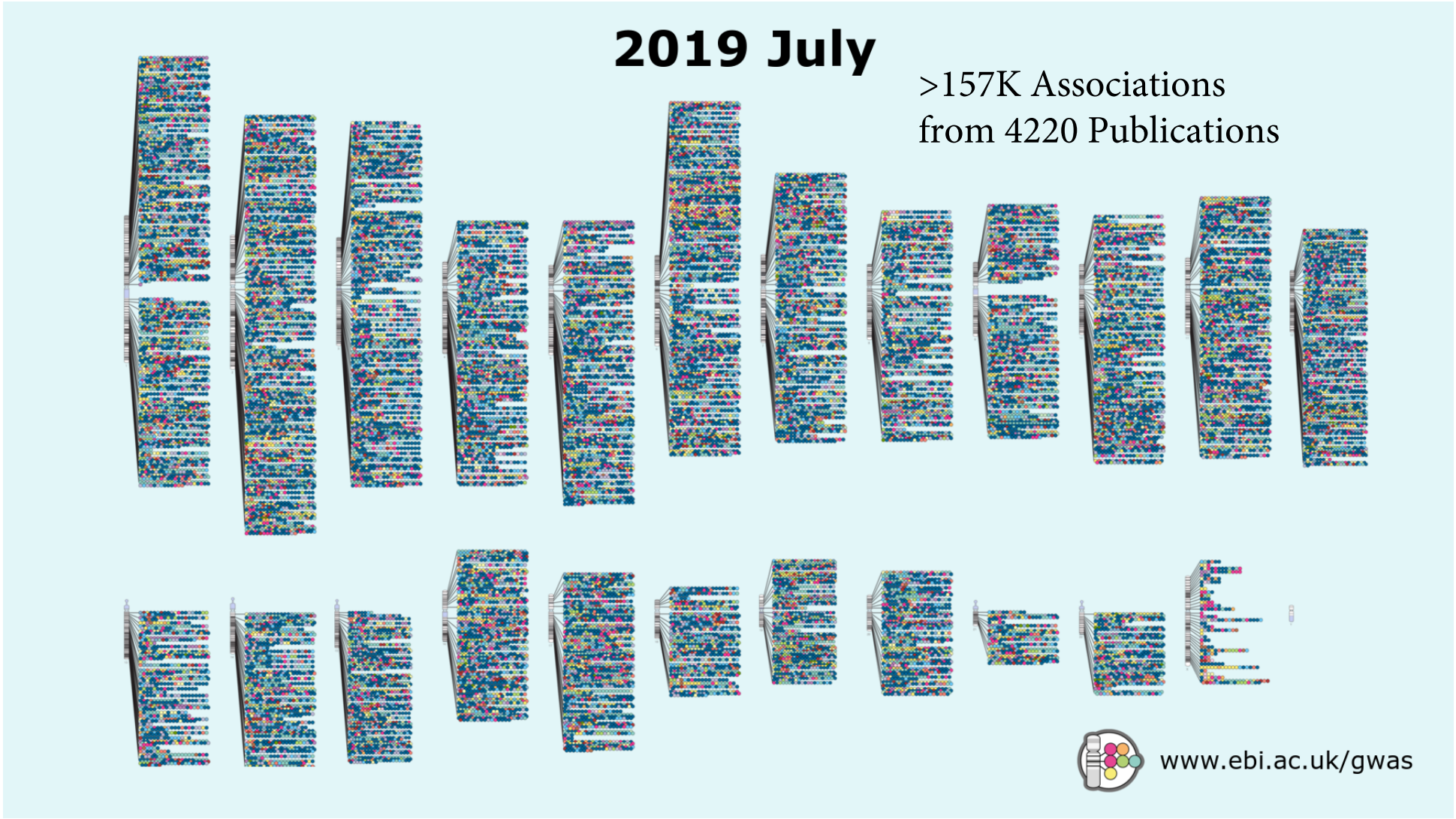
– Zoom into the peak region with gene annotations

– Visualize r^2 between the specified significant (purple diamond) signal and its neighbor SNPs

– Visualize recombination rate



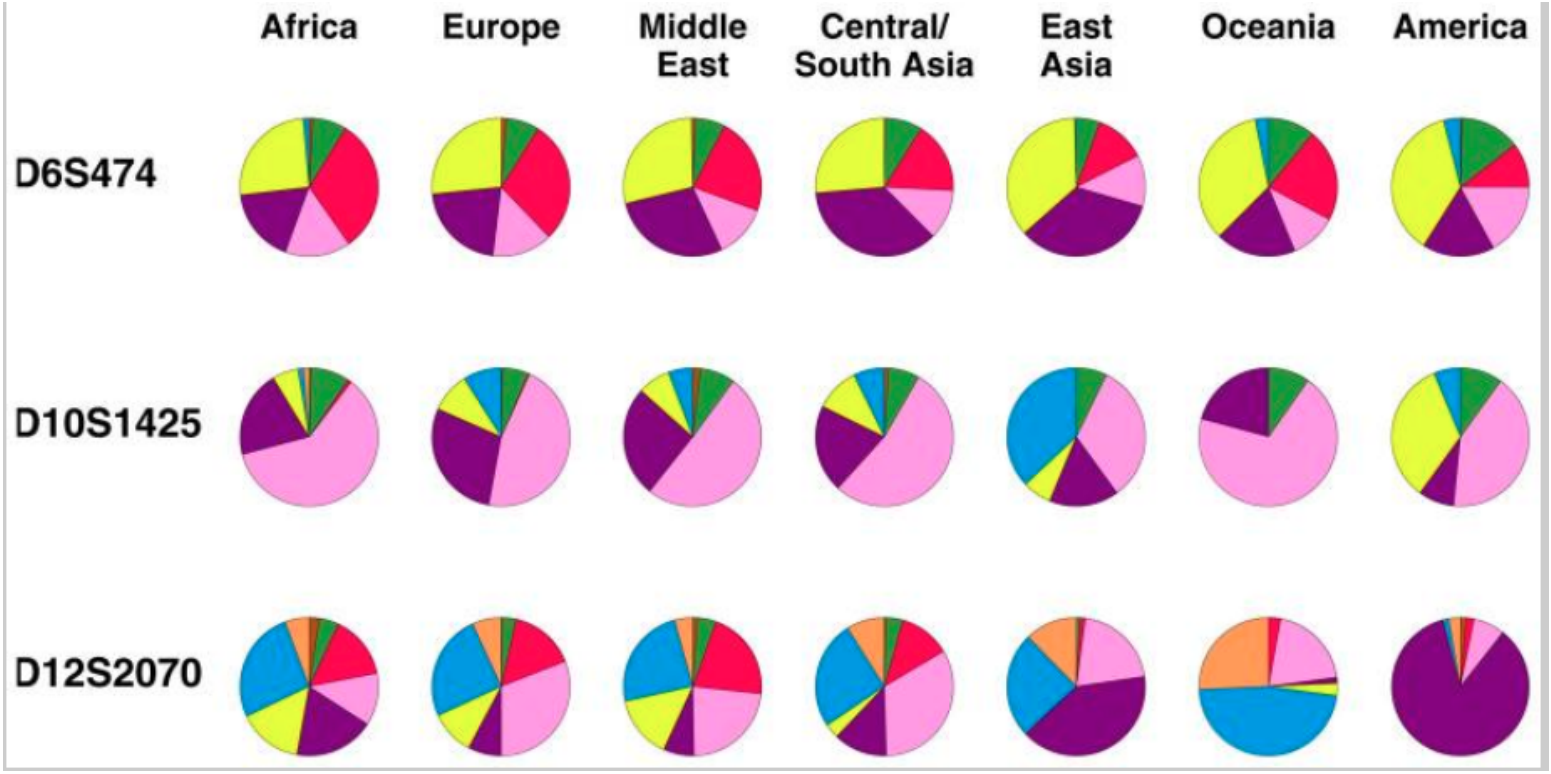
NHGRI-EBI GWAS Catalog



https://www.ebi.ac.uk/gwas/

Population Stratification

Population stratification (or population structure) is the presence of a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry.



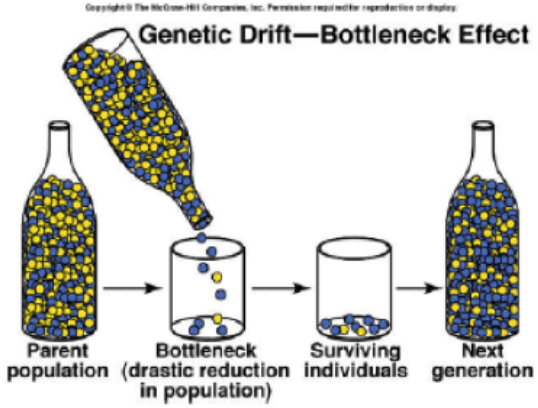
Allele frequencies at three microsatellite loci (Rosenberg N.A., Hum Biol. 2011). Each of the three loci has exactly eight alleles. In most of the pie charts, one or more alleles is rare or absent.

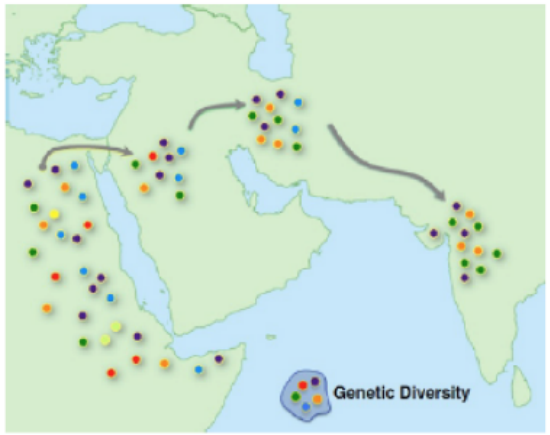
Population Stratification Causes

Basic cause of population stratification is non-random mating, often due to human migration and physical separation.

Human migration:

National Geographic





Henn *et al.* (2012) *PNAS*

Population Stratification Consequences

1. Population stratification is a major con-founder in genetic association studies, which can lead to false significant association that are not due to a disease locus;
2. Often lead to inflated false positive findings for studies including a mixture of different subpopulations;
3. Often seen when case-control ratio (or disease prevalence) is different across subpopulations, or when phenotypes differ among subpopulations.

How to Address Population Stratification

Straightforward approach:

- Carefully select samples such that cases and controls are ethnically matched
- Stratify analyses by ethnicity and then combine results by meta-analysis

Potential problems:

- Self-report is not always reliable
- Considerable variability exists even within race

Widely used approach:

- Account for inflated false-positive rate (genomic control factor)
- Adjust for ancestry quantified by genetic markers (Principal Components Analysis)

Alternative approach:

- Family-based association analysis

Genomic Control Factor

Genomic Control Factor is used to control for systematic inflation of type I error.

The idea is that the statistic T is inflated by an inflation factor λ (i.e., genomic control factor) so that

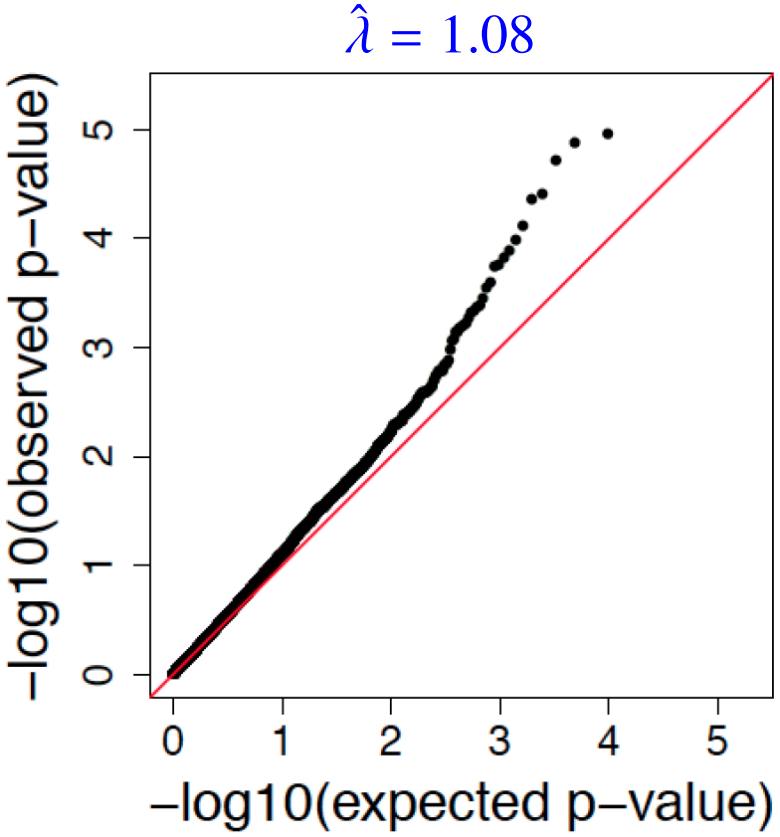
$$T \sim \lambda \chi_1^2$$

where λ can be estimated by

$$\hat{\lambda} = \text{median}(T_1, T_2, \ldots, T_M)/0.456$$

- M is the number of independent tests, though in practice all tests are included.
- The denominator is the median of χ_1^2 distribution.
- $\hat{\lambda}$ should be 1 under H_0.

Genomic Control Factor



Divide all chi-square test statistics T by the estimated inflation (GC) factor to get corrected test statistics

$$T/\hat{\lambda} \sim \chi^2_1$$

under H_0 of no association.

Genomic Control Factor

Limitation:

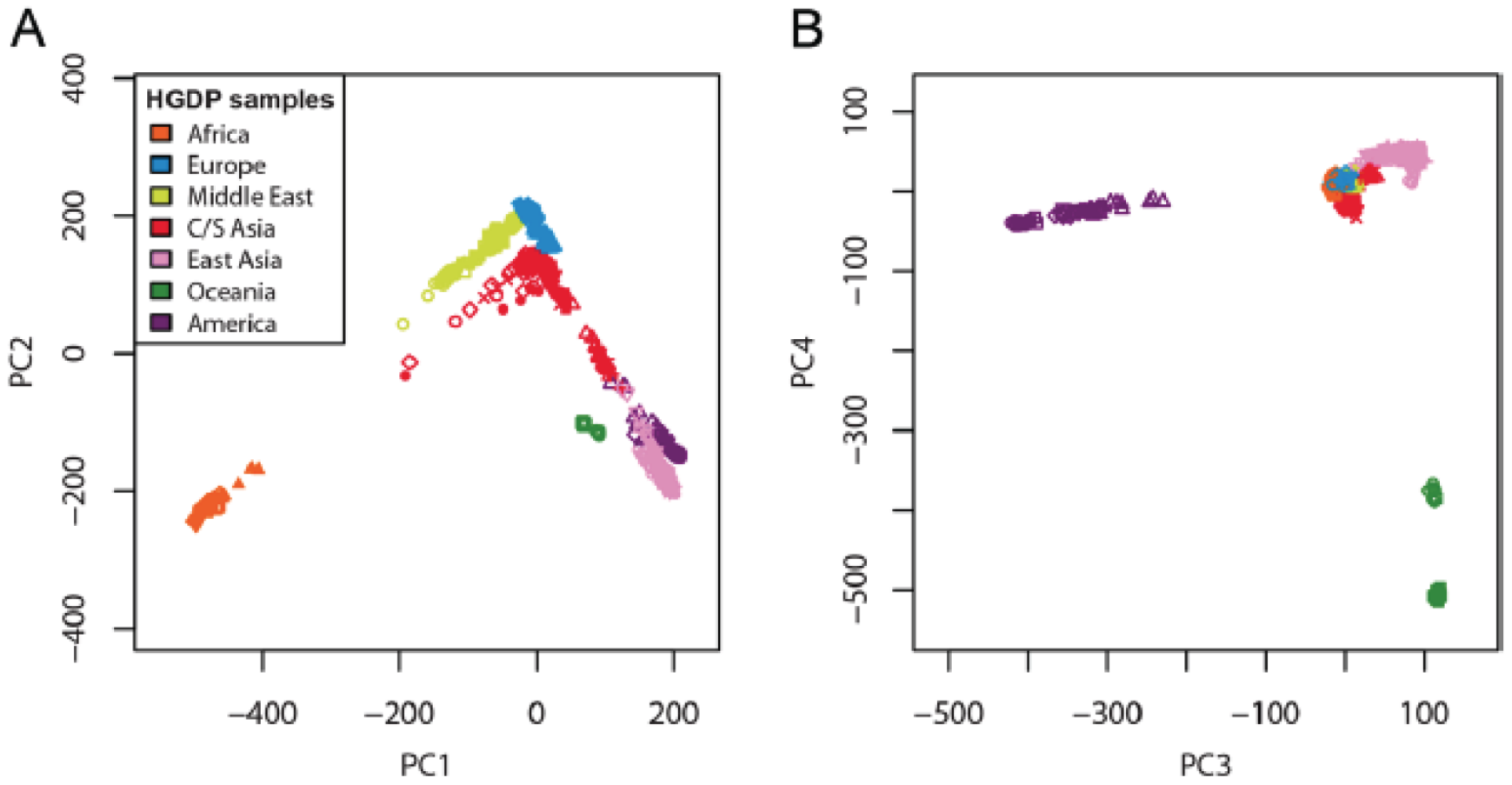
- Genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor.
- However, some markers differ in their allele frequencies across ancestral populations more than others.
- Thus, the uniform adjustment applied by genomic control may be insufficient at markers having unusually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss in power

Principal Component Analysis (PCA)

The principal component analysis (PCA) has become one of the standard ways to adjust for population stratification in population-based GWAS (Price et. al., Nature Genetics, 2006).

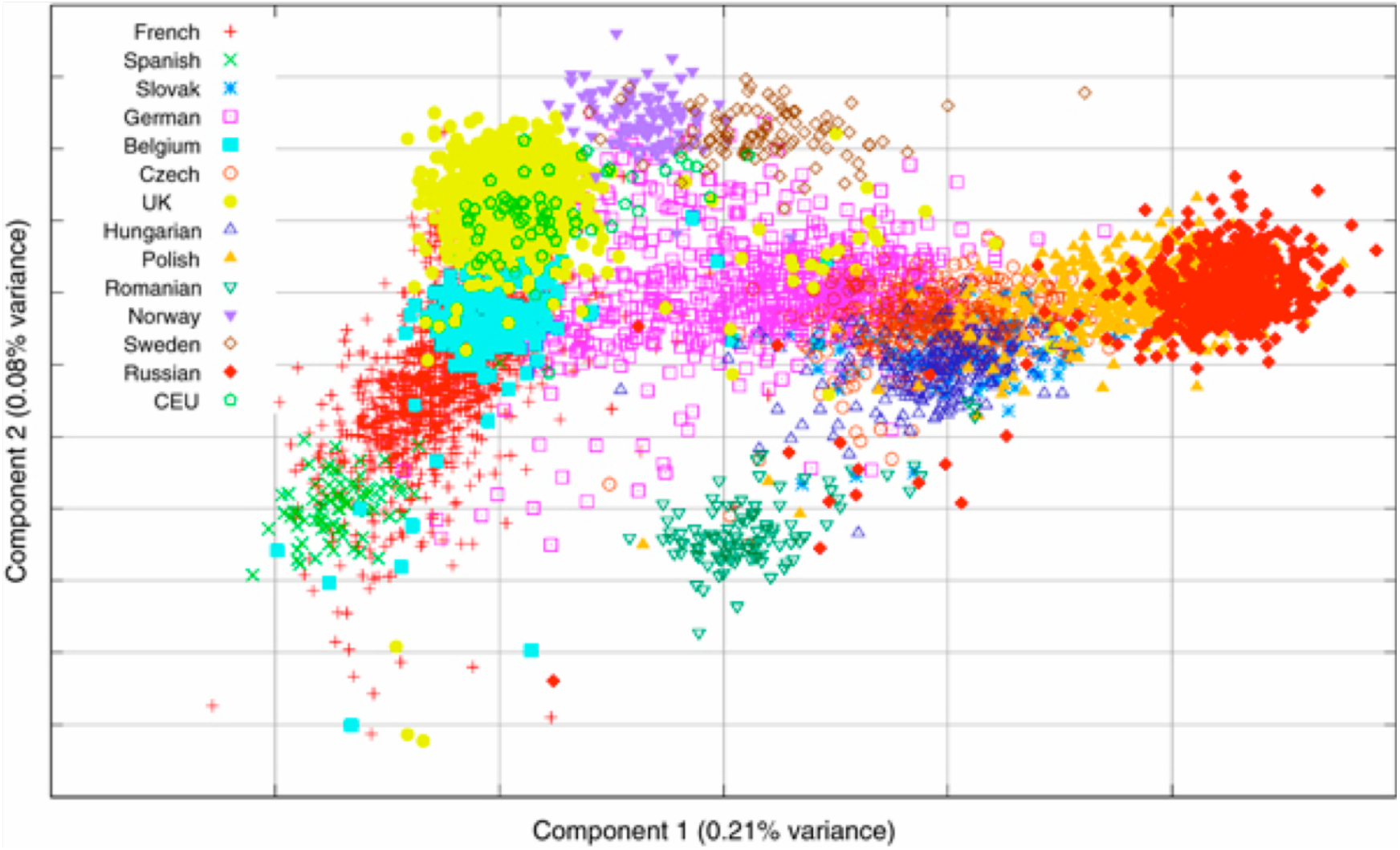
- Apply PCA to genotype data to obtain top principal components (PCs) that explain most genotype data variation
- GWAS tool PLINK (Purcel at. al., 2007) can be used to generate top PCs, https://www.cog-genomics.org/plink/
- Top PCs will reflect sample ancestry
- Include top PCs as covariates in GWAS

Top PCs often reflect geographic distribution (e.g, PC1 - PC4 as follows)



Li et al. Science. 2008; Jakobsson et al. Nature. 2008.

PC1 vs. PC2 among European samples



Heath et al. 2008.

PCA Algorithm

Notation:

- M : Number of SNPs
- N : Number of subjects
- $Z = (z_{ij})$: an $M \times N$ matrix of standardized genotyped coded for the additive model for the ith SNP in the $j$the subject, i.e.,

$$z_{ij} = (X_{ij} - \bar{X}_{i.}) / \sqrt{2\hat{p}_i(1-\hat{p}_i)}$$

where $\hat{p}_i$ denotes the MAF of the ith SNP.

Algorithm:

- Compute the $N \times N$ variance-covariance matrix as $\Sigma = Z^{\mathrm{T}}Z/(N-1)$.
- Compute the eigenvalue decomposition of Σ: e.g., using R function `eigen`
- Select the top K eigenvalues that are significantly large ($K = 5$ or 10) by a scree plot.
- Include the K eigenvectors (PCs) as additional covariates in the generalized linear regression models that are used for GWAS.

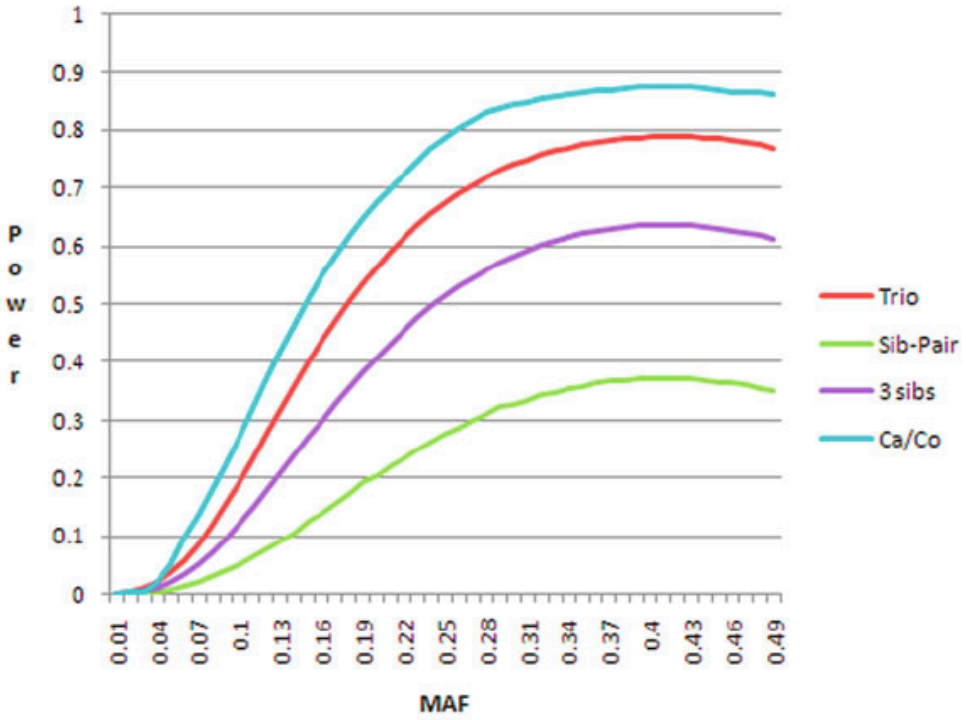
Family-Based Association Methods

- **Intuition:** Under the null hypothesis of no association, an affected child is equally likely to inherit either allele at the tested marker locus; allele not inherited by the affected child serves as a matched control.

  - **Transmission disequilibrium test (TDT):** Father-Mother-AffectedChild trios. Spielman *et al.* 1993. *American Journal of Human Genetics*
  - **Discordant alleles tests:** Affected-Unaffected siblings
  - **Family-based association test (FBAT):** General tests that can be used for both dichotomous and quantitative traits
  - **Quantitative TDT (QTDT):** Variance-Components based test of transmission distortion for quantitative traits

Power for Family-Based & Population-Based Designs

Common disease (prevalence of 10%)



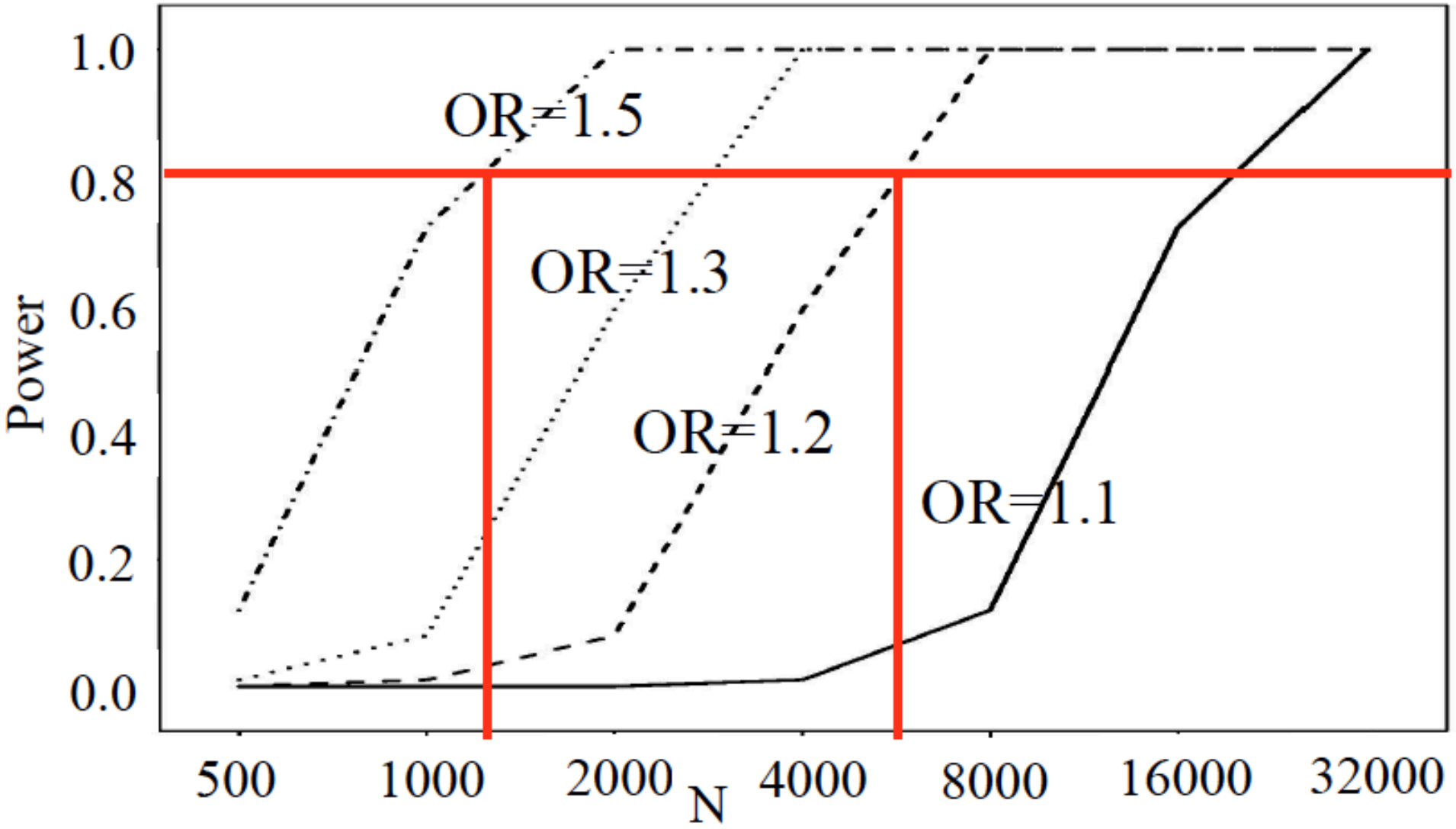
- Sib-Pair: discordant sib pairs
- 3 sibs: discordant sib trios (one discordant sib pair and one additional sibling)
- Power is estimated for 1500 families or 1500 cases and 1500 controls under an additive mode of inheritance and an odds-ratio of 1.4.

Meta-Analysis

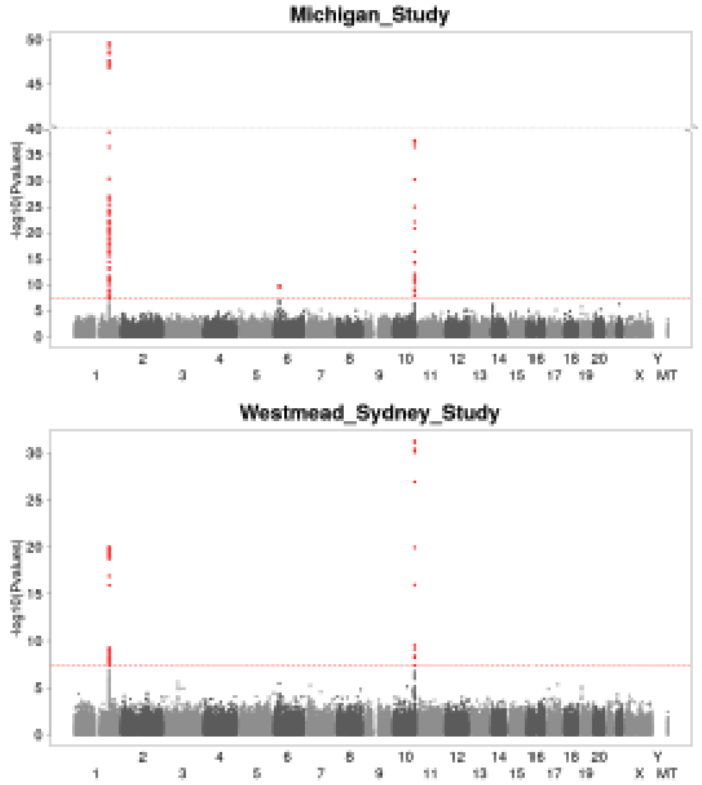
- Combine summary statistics (e.g., p-values, odds ratios, effect-sizes) across multiple studies for the same phenotype
- Improve power for increasing total sample size
- Address between study variances (due to population stratification, study design)
- Avoid the hassle of sharing individual-level genotype/phenotype/covariate data (e.g., privacy protocols)
- Yang et. al. (2017) showed that meta-analysis with summary results can be statistically equivalent to joint analysis using individual-level data

Why Meta-Analysis?

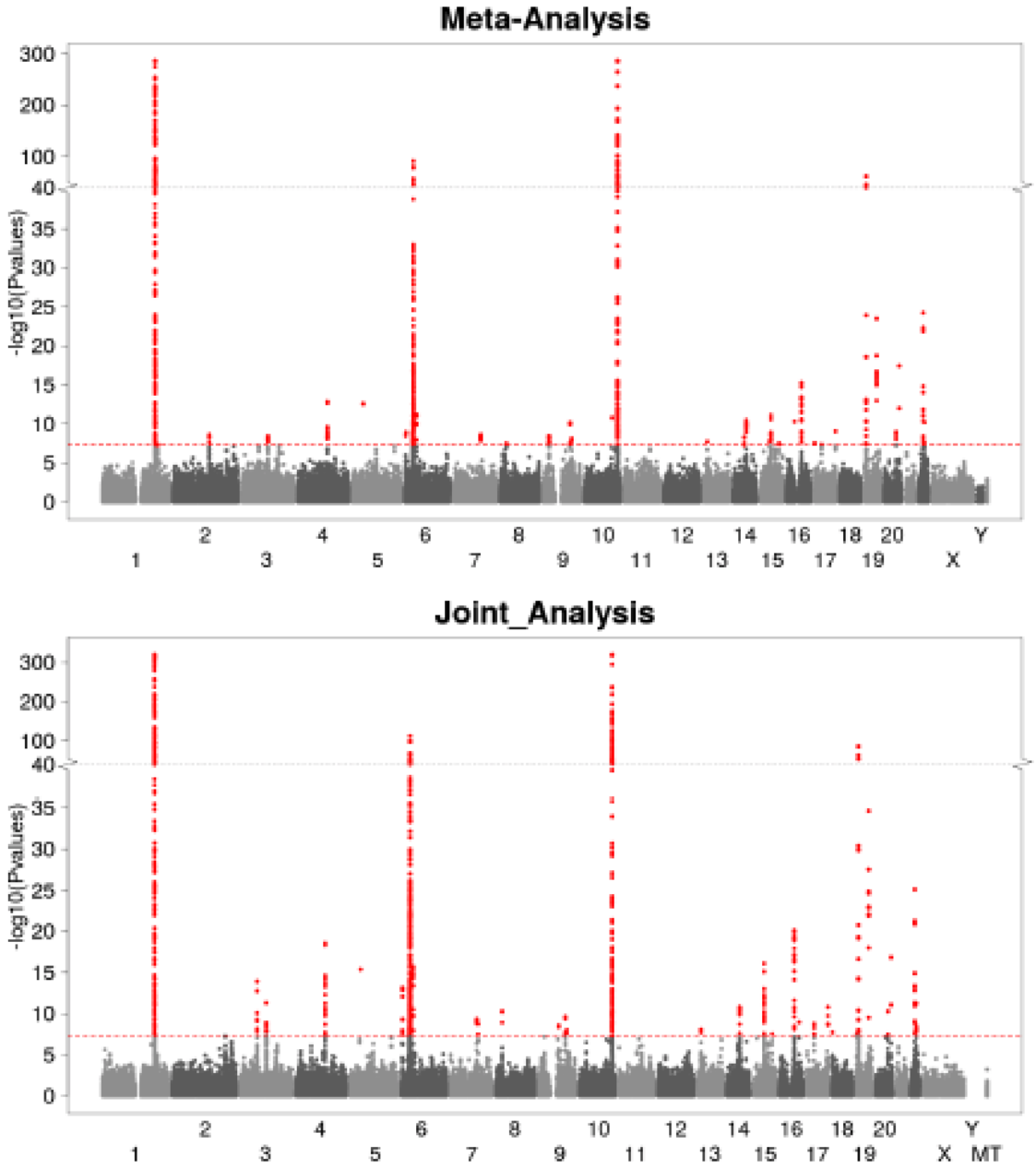
Additive model, N cases, N controls, MAF = .3, $\alpha = 5 \times 10^{-8}$



Why Meta-Analysis?



Example two individual studies of AMD.



Meta-Analysis Methods

- Fisher's Method: combining p-values
- Stouffer's Z-score method
- Fixed Effect Model: combining standardized effect-sizes
- Software: METAL (Willer et. al., 2010, Bioinformatics.)
  https://genome.sph.umich.edu/wiki/METAL_Documentation

Fisher's method

Given summary statistics from individual studies of the same genetic variant

- p_k: p-value from the kth study, $k = 1, \ldots, K$

The test statistic

$$-2 \sum_k \log(p_k) \sim \chi^2_{(2K)}$$

Derivation:

- Under the null, each p_k follows $U[0, 1]$
- The $-\log$ of a uniformly distributed value follows an exponential distribution
- Scaling a value that follows an exponential distribution by a factor of two yields a quantity that follows a χ^2 distribution with 2 df
- The sum of K independent χ^2 values follows a χ^2 distribution with $2K$ df

Stouffer's Z-score method

Given summary statistics from individual studies of the same genetic variant

- n_k: sample size of the kth study
- p_k: p-value from the kth study
- β_k: effect-size for the kth study

Then, we obtain

- $Z_k = \text{sign}(\beta_k)\Phi^{-1}(1 - p_k/2)$, where Φ is standard normal CDF.
- $w_k = \sqrt{n_k}$: weight

Stouffer's Z statistic is given by

$$\frac{\sum_k w_k Z_k}{\sqrt{\sum_k w_k^2}} \sim N(0, 1)$$

Inverse-variance estimator

Given summary statistics from individual studies of the same genetic variant

- $\hat{\beta}_k$: genetic effect-size from the kth study
- v_k: variance of $\hat{\beta}_k$ from the kth study

Then, consider

- $\beta_{meta} = \frac{\sum_k w_k \beta_k}{\sum_k w_k}, \; w_k = 1/v_k$
- $V_{beta} = \frac{1}{\sum_k w_k}$
- Inverse-variance weighting

The Wald test statistic is given by

$$\frac{\beta_{meta}}{\sqrt{V_{meta}}} \sim N(0,1)$$

Post-GWAS Analysis

- Replication study with independent datasets
- Fine-mapping GWAS loci while accounting for functional annotation (Yang et. al. 2017 AJHG; Schaid et. al., 2018, Nature Reviews Genetics)
- Biological interpretation with gene ontology/pathway analysis
- Biological replication (e.g., CRISPER-CAS9)

GWAS Tools

- PLINK (Purcel et. al., 2007): https://www.cog-genomics.org/plink/, data preparation, QC, GWAS, generate top PCs
- EPACTS: https://github.com/statgen/EPACTS, GWAS with genotyped and imputed doseage data, Manhattan plot, QQ plot
- Locuszoom (Boughton et. al., 2021): https://my.locuszoom.org/, Manhattan plot, Locus zoom plot, visualize other public GWAS results
- METAL (Willer et. al., 2010, Bioinformatics.): https://genome.sph.umich.edu/wiki/METAL_Documentation, meta-analysis with GWAS summary statistics (Z-scores, p-values, effect sizes, standard deviation of effect sizes)
- DAVID: https://david.ncifcrf.gov/, Gene ontology analysis

References

- Price A.L. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics volume 38, pages 904-909 (2006).
- Purcell, S. et. al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet.81, 559-575. 2007.
- Cristen J. Willer, Yun Li, Gonçalo R. Abecasis; METAL: fast and efficient meta-analysis of genome-wide association scans, Bioinformatics, Volume 26, Issue 17, 1 September 2010, Pages 2190-2191.
- Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G. and International Age-Related Macular Degeneration Genomics Consortium, 2017. A scalable Bayesian method for integrating functional information in genome-wide association studies. The American Journal of Human Genetics, 101(3), pp.404-416.
- Yang, J, Chen, S, Abecasis, G, IAMDGC. Improved score statistics for meta-analysis in single-variant and gene-level association studies. Genetic Epidemiology. 2018; 42: 333– 343. https://doi.org/10.1002/gepi.22123.
- Schaid, D.J., Chen, W. and Larson, N.B., 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 19(8), pp.491-504.