Lecture 8: Introduction to Rare Variant Analysis and Collapsing Tests

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2017

Lecture Overview

1. Limitations of GWAS
2. Sequencing and Rare Variants
3. Rationale for Rare Variant Analysis
4. Challenges
5. Collapsing/Burden Tests
6. Supervised Collapsing/Burden Tests

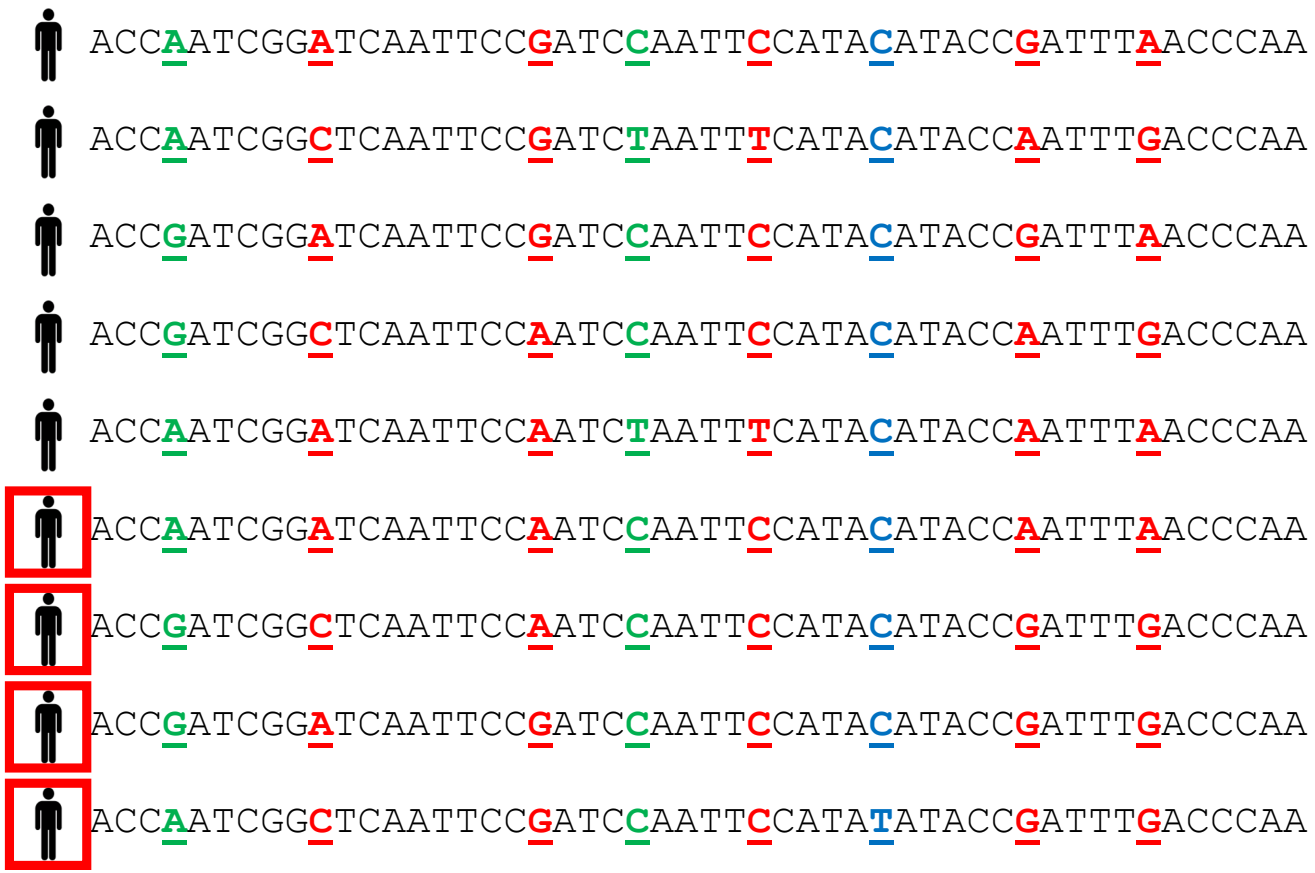
GWAS: Missing Heritability

- GWAS focus on common variants (MAF $\geq 5\%$) whose effects are small with RR$\approx$1.2-1.5.
- Missing heritability: Significant GWAS SNPs explain a small proportion of disease heritability.
- Possible reasons:
  - GxG and GxE interactions?
  - Many common causal variants: Each with a small effect?
  - Epigenetics?
  - Rare variants?

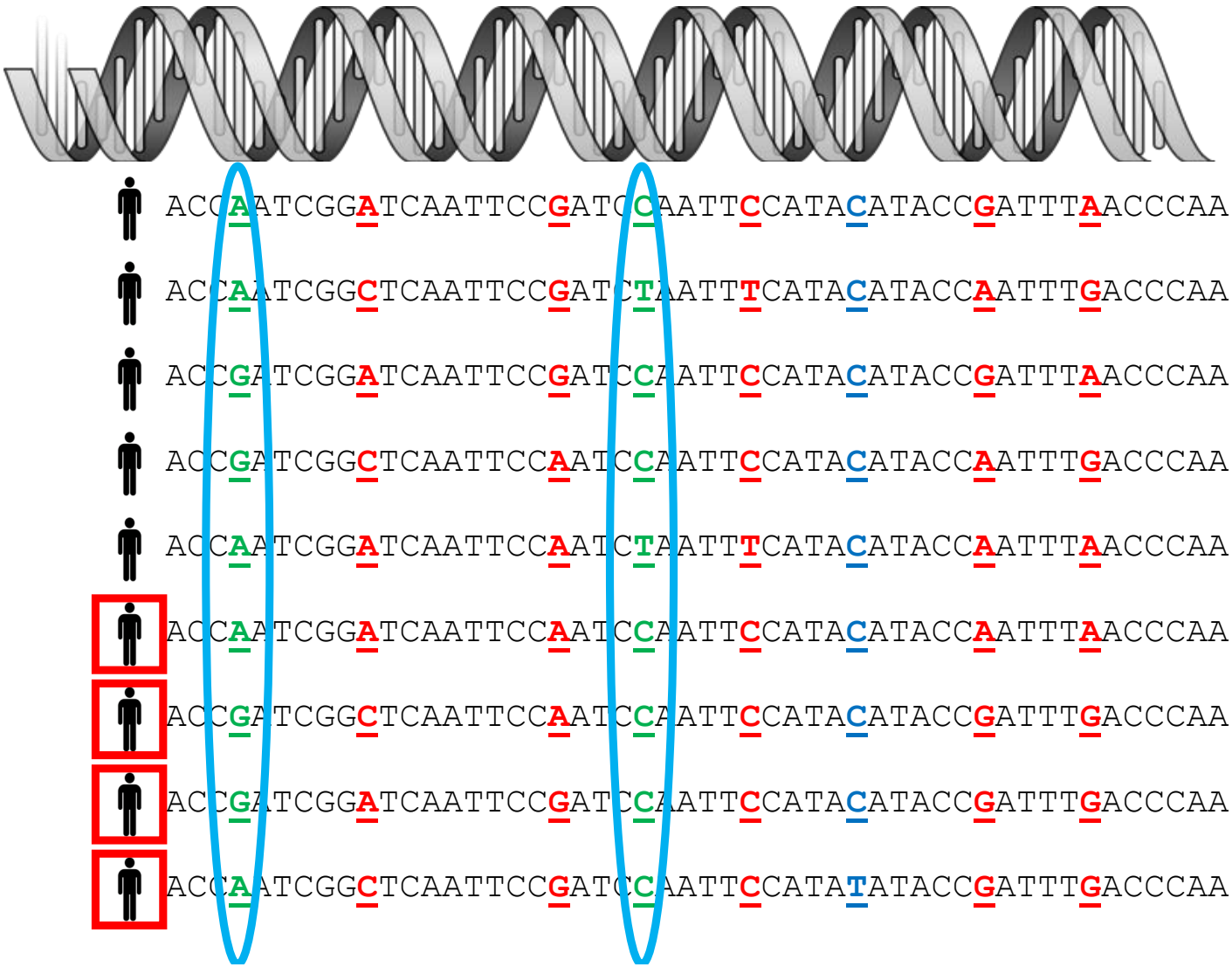
Next Generation Sequencing (NGS)

- Genotype all basepairs (bps) in a gene, the whole exome, or the whole genome (3 billion bps).
- Allow to identify all SNPs or other types of variants. No need to rely on LD to tag untyped causal SNPs.

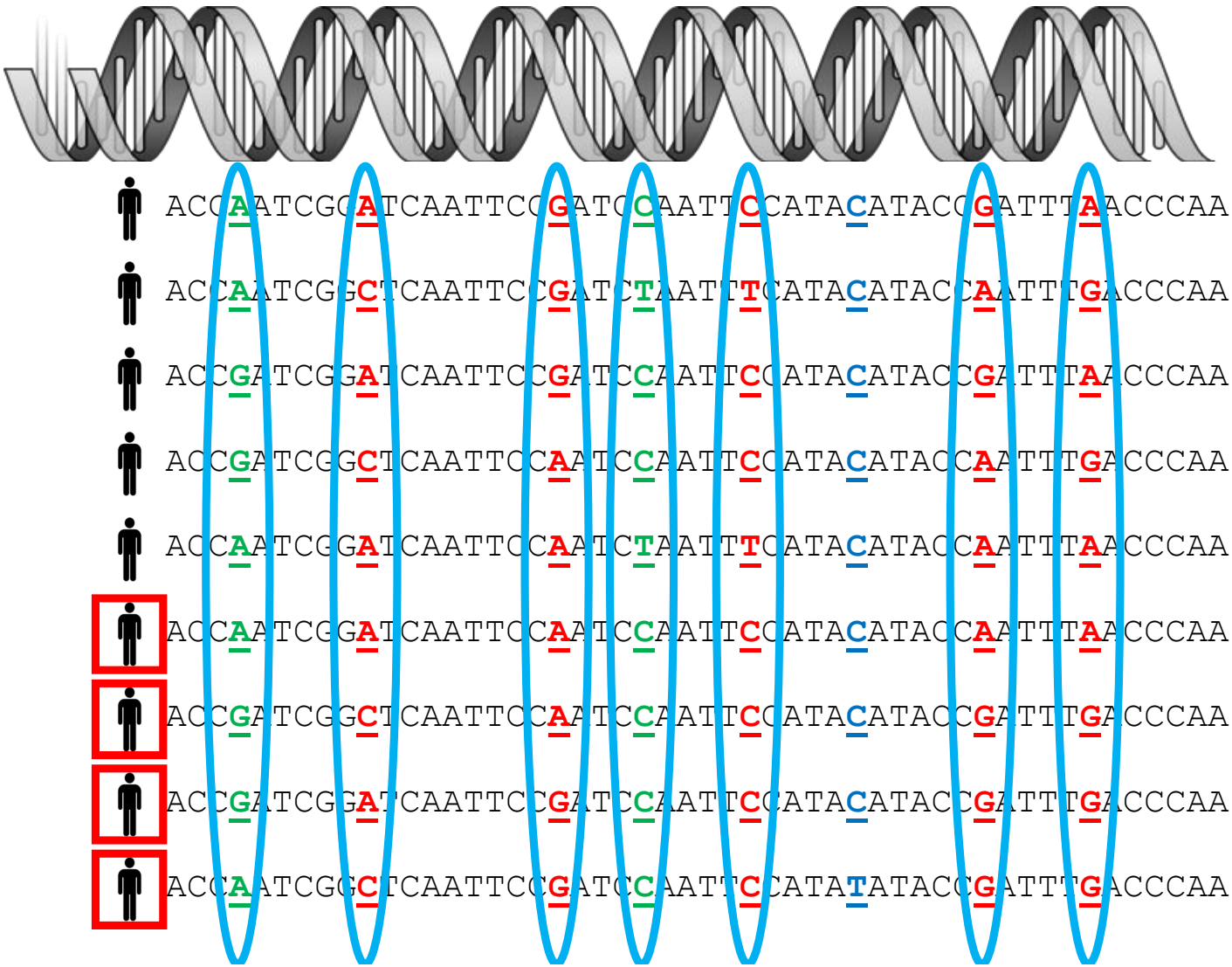
Genetic Association Studies



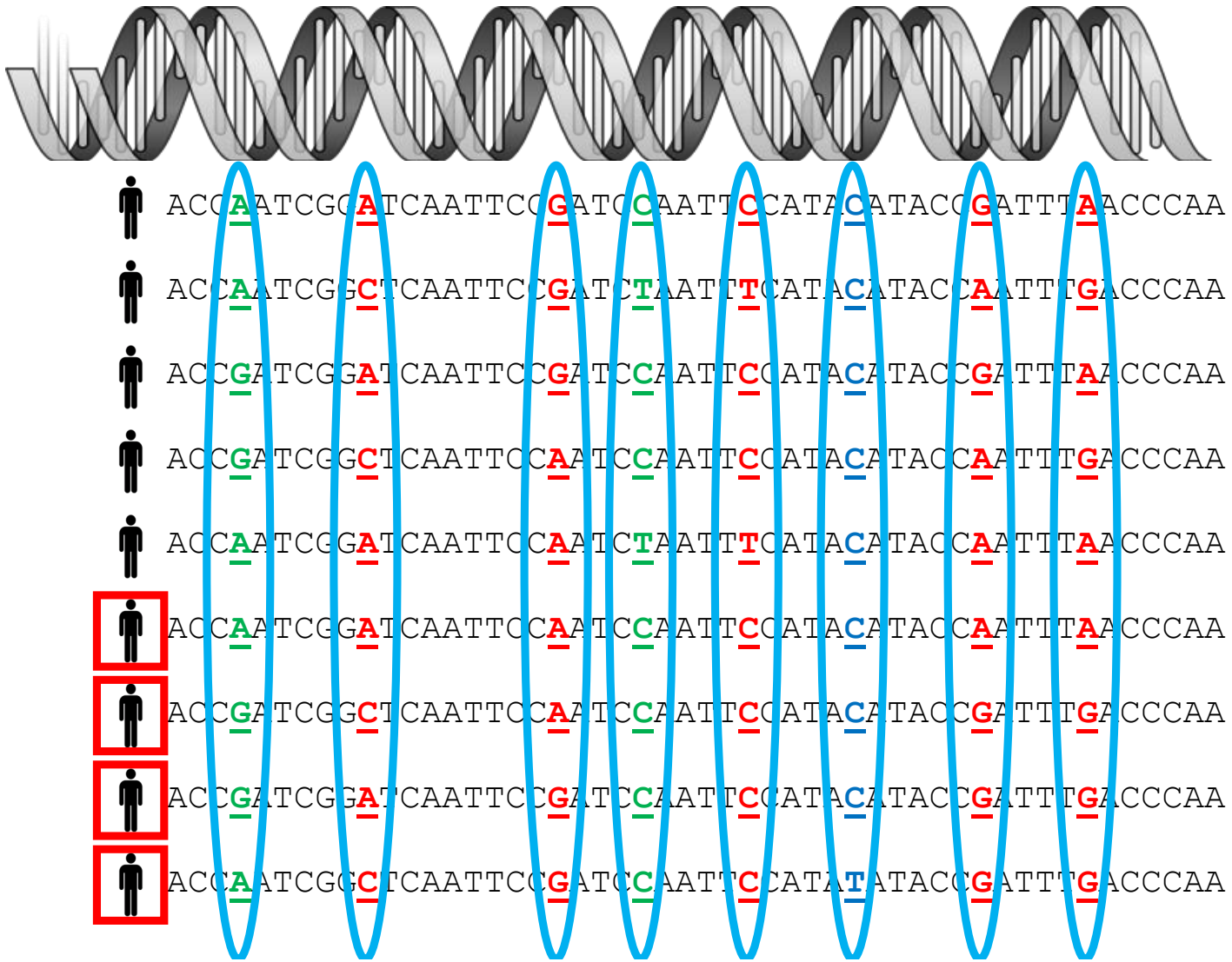
GWAS: A few years ago



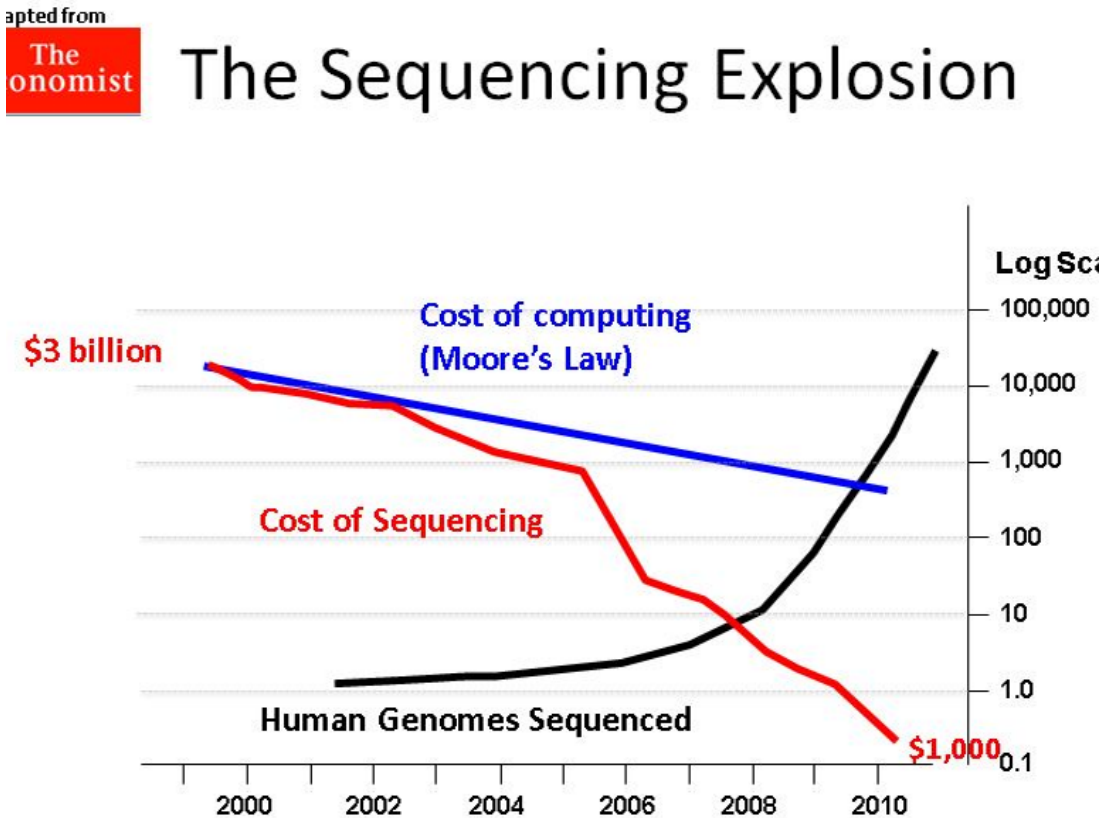
GWAS: current (+ imputation)



Sequencing


ACCAATCGGATCAATTCCGATCCAATTCCATACATACCGATTTAACCCAA
ACCAATCGGCTCAATTCCGATCTAATTTCATACATACCAATTTGACCCAA
ACCGATCGGATCAATTCCGATCCAATTCCATACATACCGATTTAACCCAA
ACCGATCGGCTCAATTCCAATCCAATTCCATACATACCAATTTGACCCAA
ACCAATCGGATCAATTCCAATCTAATTTCATACATACCAATTTAACCCAA
ACCAATCGGATCAATTCCAATCCAATTCCATACATACCAATTTAACCCAA
ACCGATCGGCTCAATTCCAATCCAATTCCATACATACCGATTTGACCCAA
ACCGATCGGATCAATTCCGATCCAATTCCATACATACCGATTTGACCCAA
ACCAATCGGCTCAATTCCGATCCAATTCCATATATACCGATTTGACCCAA

Sequencing Cost Has Dropped Dramatically



Massively parallel sequencing

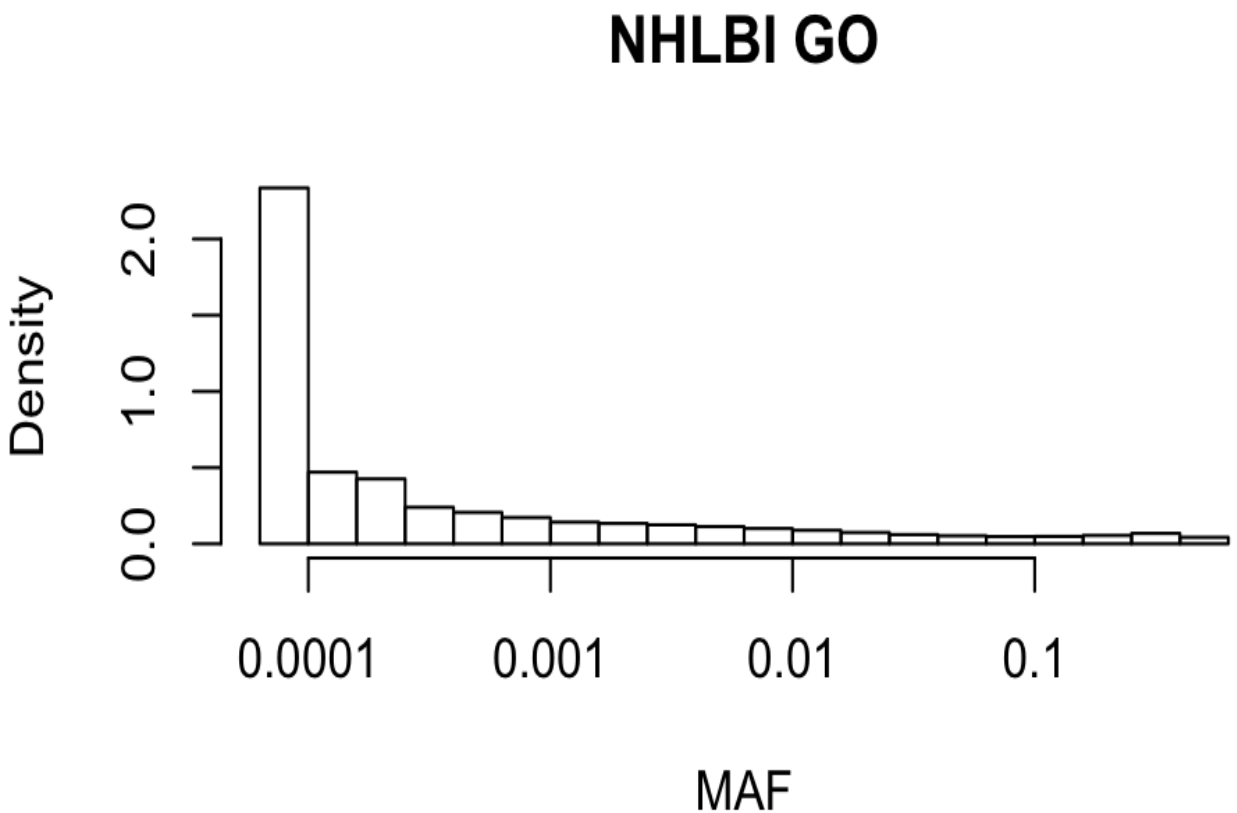
- Illumina can achieve $\$1,000$ per whole genome.

Common vs Rare variants

- Common Variants (Common SNPs):
  - MAF $> 0.01 \sim 0.05$.
  - Often high correlation with adjacent SNPs (Strong Linkage Disequilibrium(LD)).
- Rare Variants (Rare SNPs):
  - MAF $\leq 0.01 \sim 0.05$.
  - Relatively new mutations.
  - Often weak correlation with other SNPs.

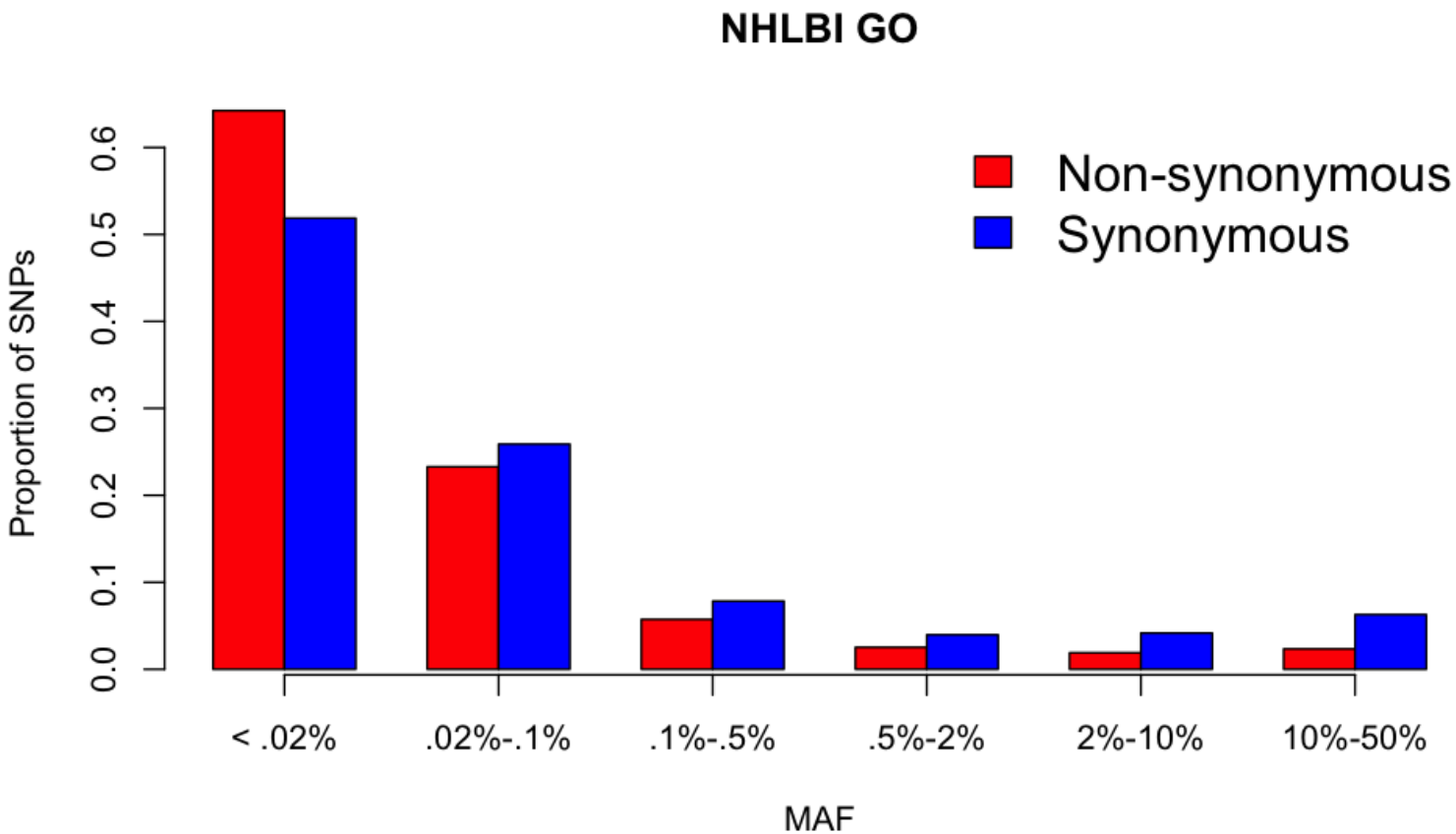
Why rare variants?

- Most of human variants are rare



Why rare variants?

- Functional variants tend to be rare.



Challenges in Association Studies for Rare Variants

- Compared to common variant studies, individual SNP analysis in rare variant studies is seriously underpowered.

How many subjects are needed to observed a rare variant?

- Sample size required to observe a variant with MAF$=p$ with at least θ chance

$$n > \frac{ln(1-\theta)}{2ln(1-p)}$$

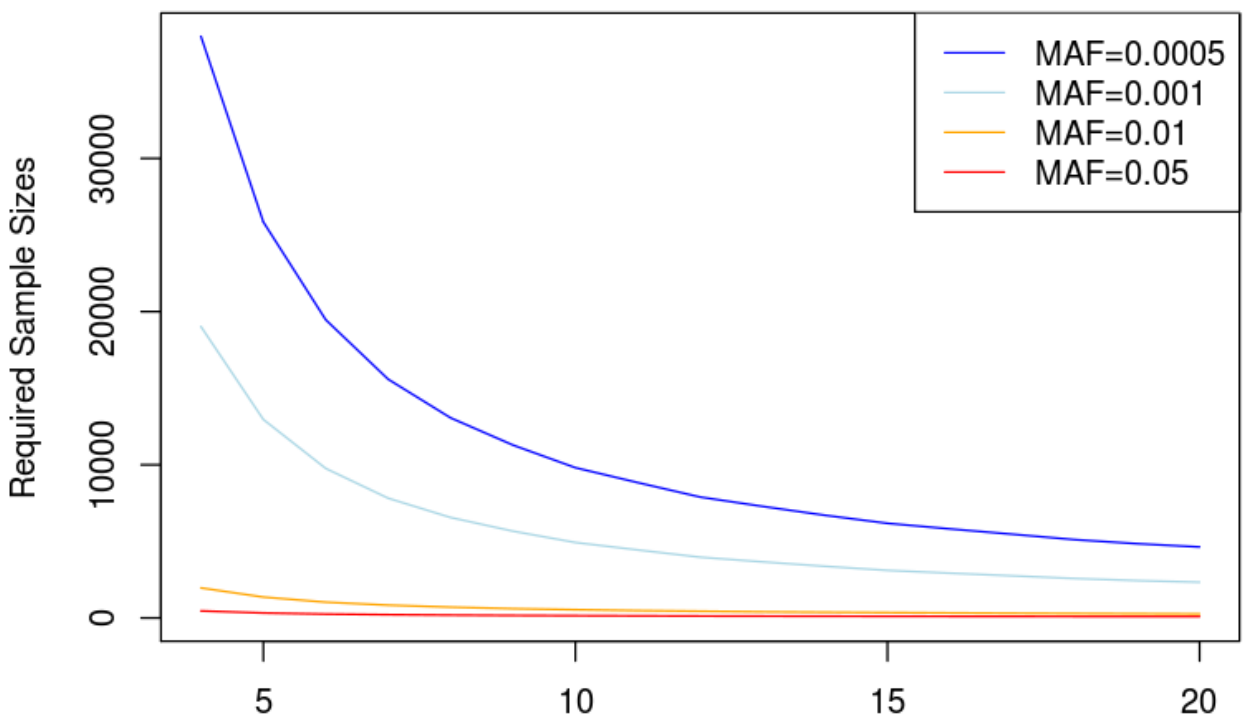
- For $\theta = 99.9\%$, the required minimum sample size is

MAF	0.1	0.01	0.001	0.0001
Minimum n	33	344	3453	34537

- Large samples are required to observe rare variants.

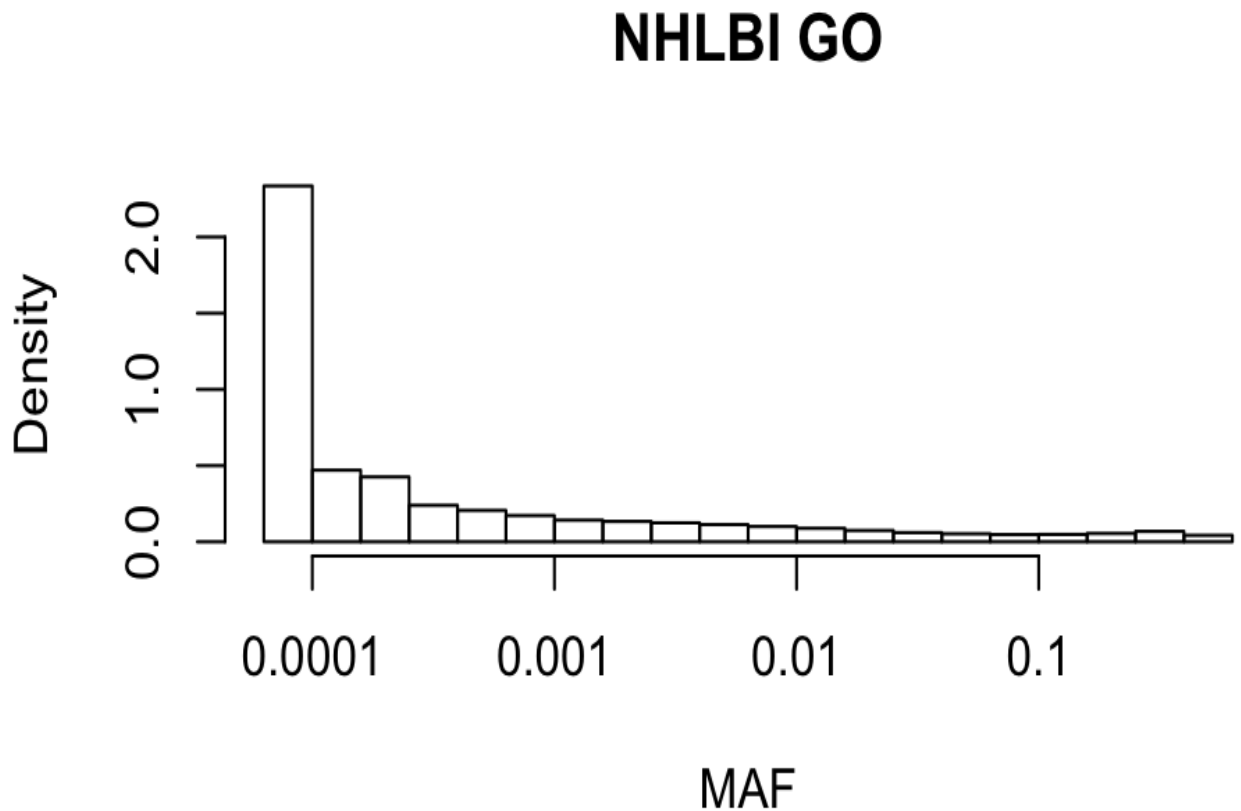
How many subjects are needed to achieve 80% of power ($\alpha = 10^{-6}$) by single variant test?

- Single variant test is not powerful to identify rare variant effects.



Multiple Testing:

- A lot more rare variants than common variants → larger multiple testing burden



Challenges in Association Studies for Rare Variants

- Individual rare variant tests are underpowered
- Need cost-effective study designs to genotype a large number of individuals
- Need powerful statistical methods and strategies to test for associations
  - Region based analysis: genes, moving windows, networks/pathways
  - Integrate with bioinformatics: Incorporate functional information

Region Based Analysis of Rare Variants

- Single variant test is not powerful to identify rare variant associations
- Gene (or Region) based tests
- Strategy:
  - Identify all observed variants within a sequenced (sub)-region.
  - Regions: gene, regulatory region, ...
  - Test the joint effect of rare/common variants while adjusting for covariates.

Regression Models

- p variants in a certain region.
- SNPs in a region $\mathbf{G}_i = (g_{i1}, g_{i2}, \ldots, g_{ip})'$, $(g_{ij} = 0, 1, 2)$
- Covariates $\mathbf{X}_i$: age, gender, PC scores (for population stratification).
- Continuous/binary traits:

$$\mu_i / logit(\mu_i) = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \mathbf{G}_i'\boldsymbol{\beta}$$

- Test of no genetic region effect:

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) = 0$$

Major Classes of Tests

- Burden/Collapsing tests
- Supervised/Adaptive Burden/Collapsing tests
- Variance component (similarity) based tests
- Omnibus tests: hedge against difference scenarios

Note: "Burden" tests sometimes refers to collapsing tests or to any region based test — inconsistent notation.

Burden Tests

- Aggregate rare variant information in a region into a summary dose variable
  - Binary Collapsing: CAST
  - CMC
  - Count Collapsing: MZ (GRANVIL)
  - Weighted Sum Test
- Most powerful if all rare variants are causal variants with the same effect sizes (and association directions).

Burden Tests- Principle

- If p is large, multivariate test $\boldsymbol{\beta} = 0$ is not powerful.
- Collapsing: Suppose $\beta_1 = \cdots = \beta_{\boldsymbol{p}} = \beta$

$$\mu_i / logit(\mu_i) = \alpha_0 + \mathbf{X}_i^T \alpha + C_i \beta$$

- $C_i = g_{i1} + \cdots + g_{ip}$: genetic burden/score
- Test $H_0 : \beta = 0$ (df=1)

Burden Tests

- Collapse rare variants

Y	G_1	G_2	G_3	G_4
1	1	0	0	0
1	0	1	0	0
1	0	0	1	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Burden Tests

- Collapse rare variants

Y	G_1	G_2	G_3	G_4		C
1	1	0	0	0		1
1	0	1	0	0		1
1	0	0	1	1	→	2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
0	0	0	0	0		0
0	0	0	0	0		0
0	0	0	0	0		0

Burden Tests

- Many different types of tests exist based on different C_i
- Existence of any rare variants can cause loss of function of a region (ex. CAST)

$$C_i = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{p} g_{ij} > 0 \\ 0 & \text{if} \quad \sum_{j=1}^{p} g_{ij} = 0 \end{cases}$$

- Dominant genetic model (ex. MZ-test)

$$C_i = \sum_{j=1}^{p} I(g_{ij} > 0)$$

Weighted Burden

- Assume that rarer variants have larger effects
- Suppose $\beta_j = w(MAF_j)\beta$.
  - Ex: $w(MAF_j) = 1/\sqrt{MAF_j(1 - MAF_j)}$ (Madsen and Browning).
- $C_i = w_1 g_{i1} + \cdots + w_p g_{ip}$
  - Weighted count of rare variants, where $w_j = w(MAF_j)$.

Burden test - CMC test

Li and Leal (2008) *AJHG*

- There exists many variations of burden tests.
- CMC test
  - Group variants based on their MAFs
  - Collapse each group using CAST approach
  - Conduct Hoteling's T-test

Burden tests - Original Weighted Sum

Madsen and Browning (2009) *Plos Genetics*

- Assume binary trait without covariates
- Control only MAFs and rank sum test
  - Weight: $w_j = 1/\sqrt{q_j^u(1-q_j^u)}$, where q_j^u is the estimated MAF from control samples.
  - Test statistic:

$$T_{wst} = \sum_{i \in case} rank(C_i), \quad C_i = \sum w_j g_{ij}$$

  - P-values from normal approximation:

$$Z = (T_{wst} - \widehat{\mu})/\widehat{\sigma}$$

Power of Burden Tests

- Power of burden tests depends on
  - Number of associated variants
  - Number of non-associated variants
  - Direction of the effects.
- Powerful if most variants are causal and have effects in the same direction.

Burden vs. Single Variant Test

	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

[Li and Leal (2008) AJHG]

- Power from simulated data
- Combining variants can greatly increase the power.

Burden Test: Mixed effect directions

- Lose power if variants have positive and negative effects.

Y	G_1	G_2	G_3	G_4
1	1	0	0	0
1	0	1	0	0
1	0	0	0	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0	0	0	0	0
0	0	0	1	0
0	0	0	0	1

Burden Test: Mixed effect directions

- Lose power if variants have positive and negative effects.

Y	G_1	G_2	G_3	G_4		C
1	1	0	0	0		1
1	0	1	0	0		1
1	0	0	0	0	→	0
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
0	0	0	0	0		0
0	0	0	1	0		1
0	0	0	0	1		1

Burden Test: Mixed effect directions

- Several methods have been developed to estimate association directions and incorporate them in the burden test framework.
  - Adaptive Sum Test
  - Estimated regression coefficient (EREC) test

Adaptive sum test

Han F and Pan W. (2010) *Hum Hered*

- Model:

$$C_i = \sum_{j=1}^{p} w_j g_{ij}$$

$$logit(Pr(Y = 1)) = \alpha_0 + C_i \beta$$

- Fit individual SNP models

$$logit(Pr(Y = 1)) = \alpha_0 + g_j \beta_j$$

  - Assign $w_j = -1$ if $\widehat{\beta}_j < 0$ and the p-value is small
  - $w_j = 1$ otherwise.

Adaptive sum test

- Compute p-values using permutation.
- Step-up procedure assign $w_j = 0$ if g_j is unlikely associated with the trait (Hoffmann *et al.*Plos One, 2010)

Estimated regression coefficient (EREC) test

Lin DY. and Tang Z. (2011) *AJHG*

- Estimate regression coefficient β and use it as a weight.

$$C_i = \sum_{j=1}^{p} w_j g_{ij}, \quad w_j = \widehat{\beta}_j$$

  - Motivation: True β_j is the optimal weight
  - Estimate $\widehat{\beta}_j$ by fitting individual SNP regression models
  - Use $w_j = \widehat{\beta}_j + \delta$ when the sample size is small (n$<$ 2000)

Estimated regression coefficient (EREC) test

- Calculate

$$C_i = \sum_{j=1}^{p} w_j g_{ij}, \quad w_j = \widehat{\beta}_j$$

- Test statistic:

$$T_{EREC} = \sum_{i=1}^{n} C_i (y_i - \widehat{\mu}_{0,i}).$$

  - Use score test statistics
  - P-values from the parametric bootstrap.

Estimated regression coefficient (EREC) test

- Cons:
  - Individual SNP regression models are difficult to fit for very rare variants.
  - The constant δ is arbitrary.

Adaptive burden test

- Adaptive burden tests have robust power.
- Compute p-values through permutation or bootstrap
  - Computationally intensive