An Introduction to Genome-Wide Association Studies (GWAS)

Shweta Ramdas

23 June 2021

shwetar@pennmedicine.upenn.edu

How do we find the genes responsible for a disease?

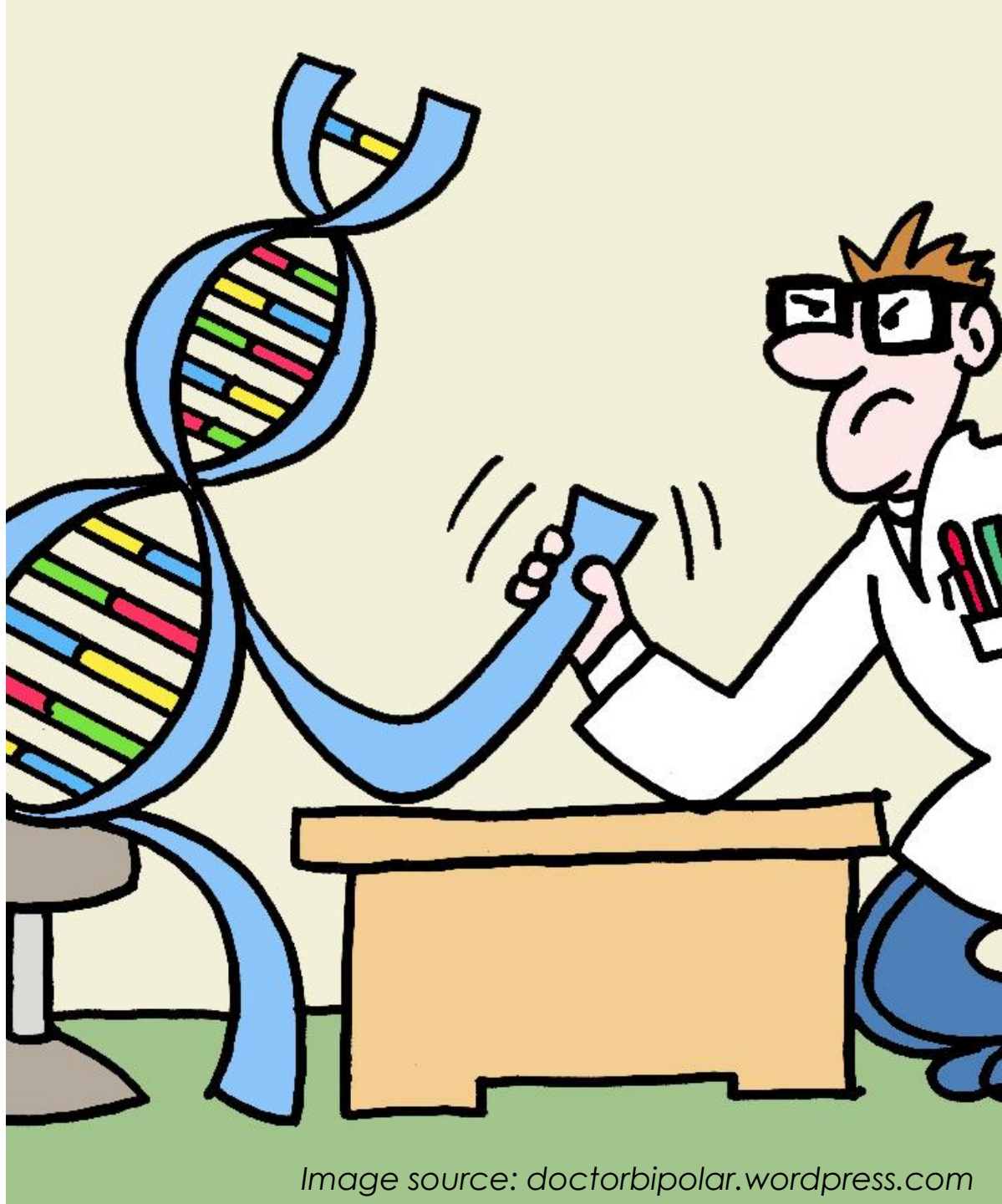


Image source: doctorbipolar.wordpress.com

How do we find the genes responsible for a disease *using genetic data from human populations*?

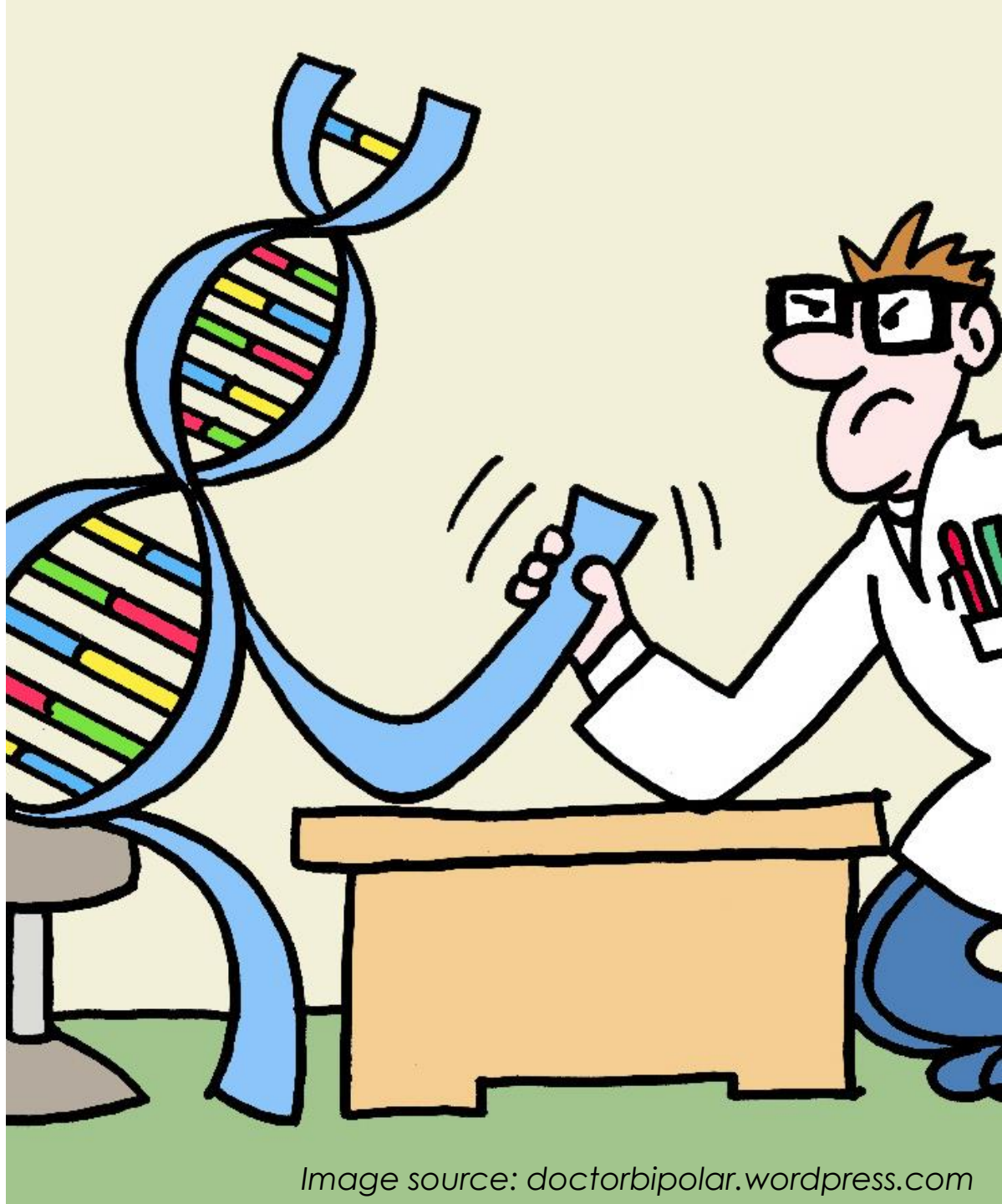


Image source: doctorbipolar.wordpress.com

Knowing your genetic risk for a disease

		My Risk	Population Risk	
Colorectal Cancer	★★★★	8.9%	5.6%	1.60x
Rheumatoid Arthritis	★★★★	4.6%	2.4%	1.94x
Type 1 Diabetes	★★★★	2.1%	1.0%	2.08x
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.43%	0.36%	1.21x
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.28%	0.23%	1.22x
Bipolar Disorder	★★★★	0.15%	0.10%	1.44x

Genome-wide association study (GWAS)

- Studies aiming to find **genetic differences between individuals** that influence **susceptibility to diseases** (or other traits).

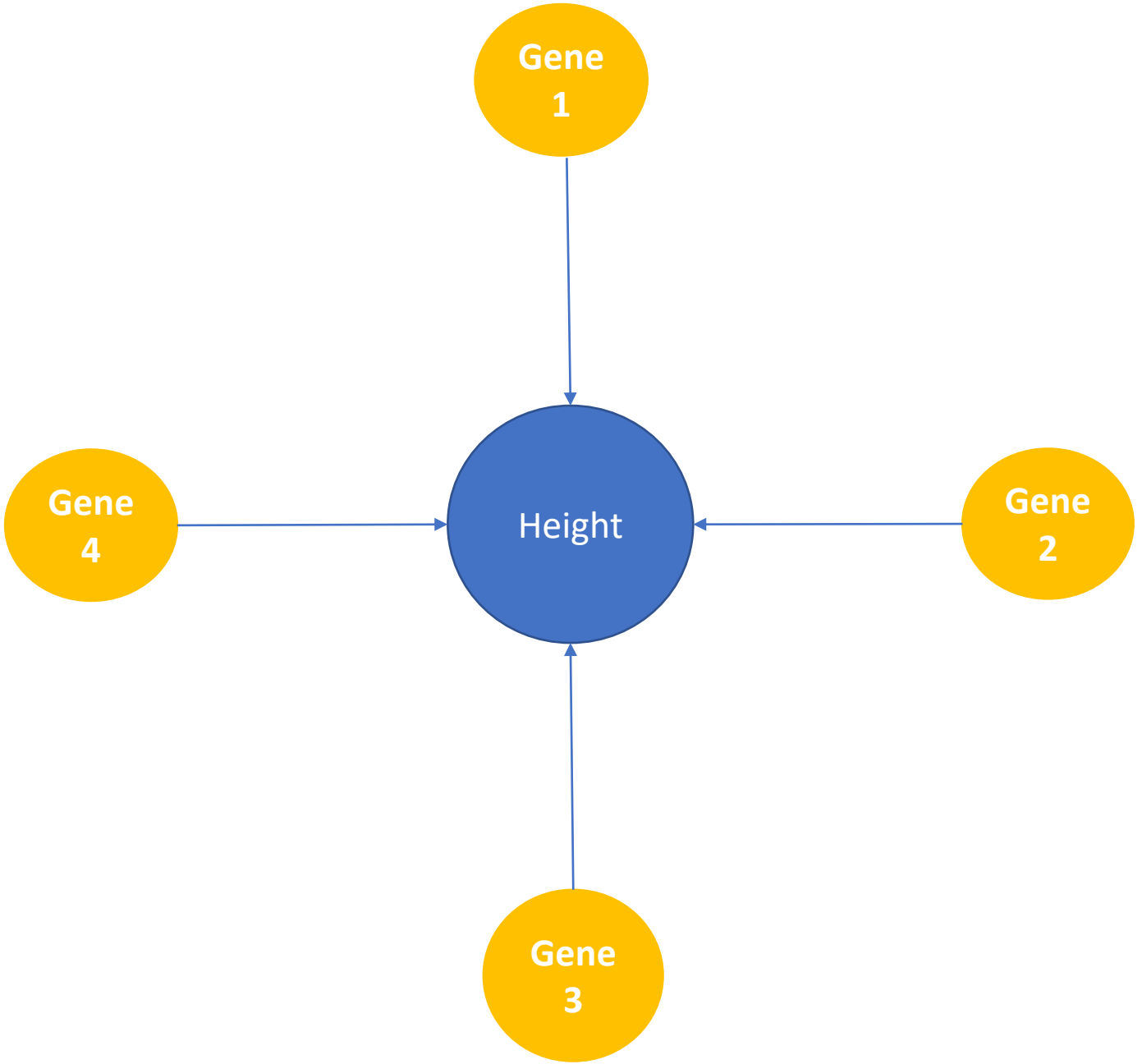
What diseases do we study?

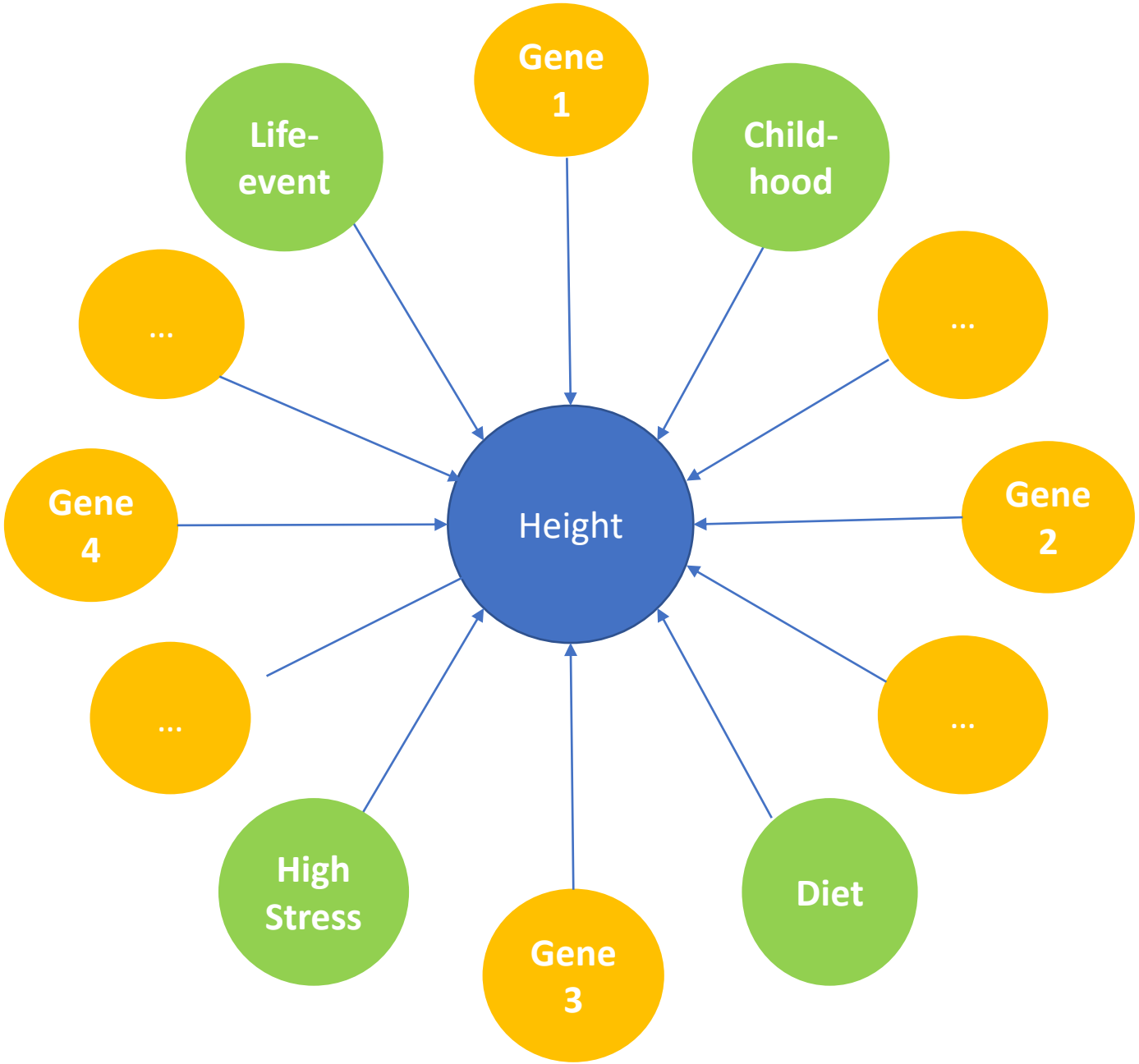
What diseases do we study?

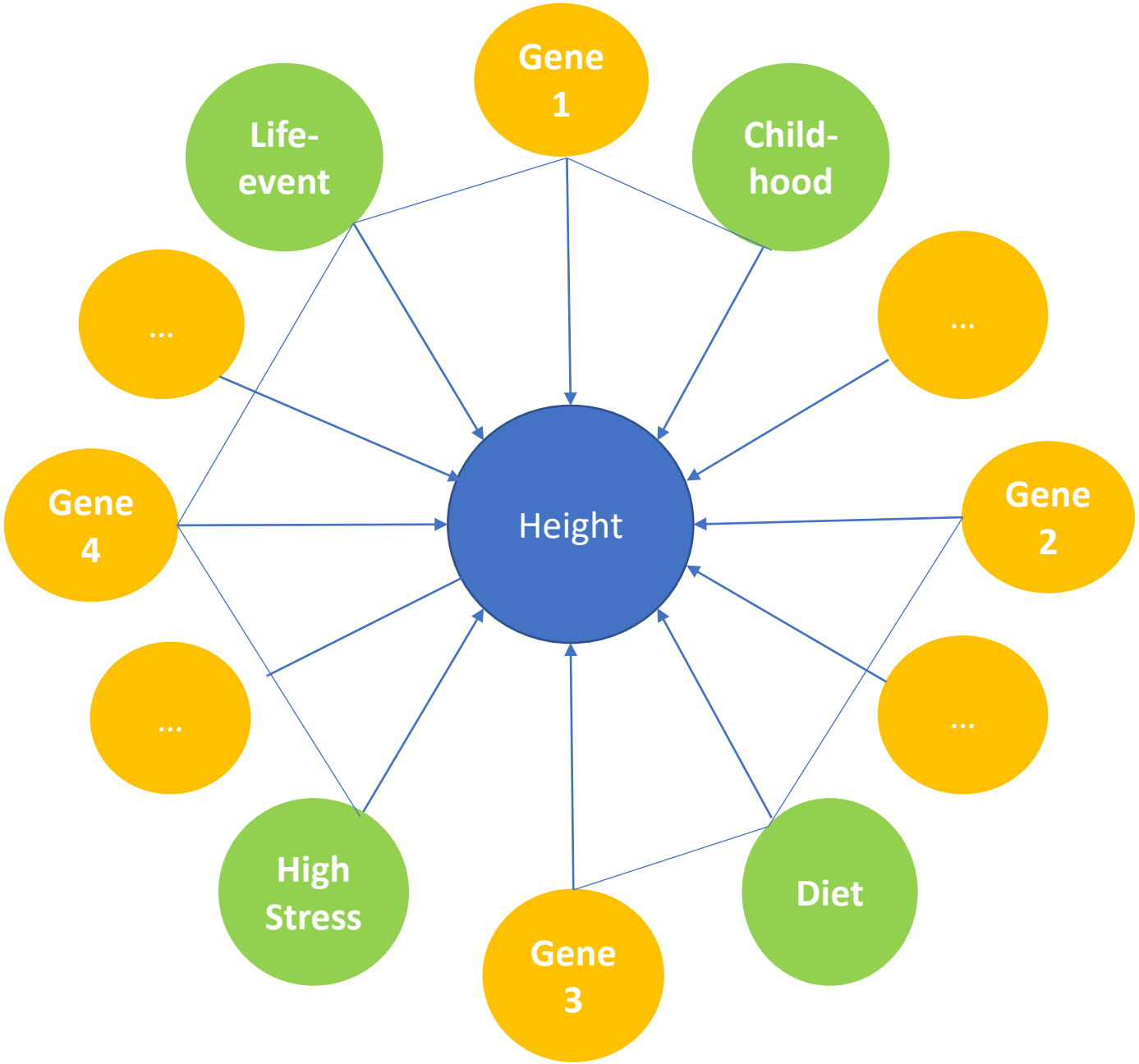
- **Mendelian**: One gene-one disease

What diseases do we study?

- **Mendelian**: One gene-one disease
- **Complex trait**: many genes + environment = disease
  - Different people with the same trait may have different causal factors


Gene 1
Gene 4
Height
Gene 2
Gene 3


Gene 1
Life-event
Child-hood
...
...
Gene 4
Height
Gene 2
...
...
High Stress
Gene 3
Diet

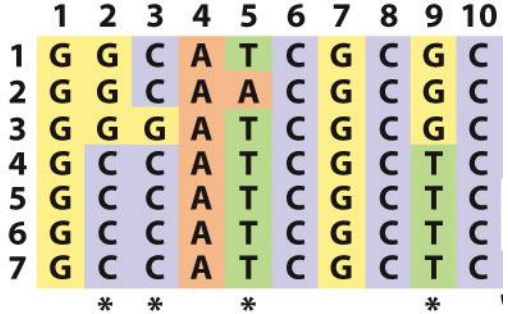

Gene 1
Life-event
Child-hood
...
...
Gene 4
Height
Gene 2
...
...
High Stress
Gene 3
Diet

How do we find disease genes in the genome?

- Core principle:
  - Collect a large group of people, half with a disease, half without
  - Sample their DNA
  - Look for **genetic differences** between the two groups

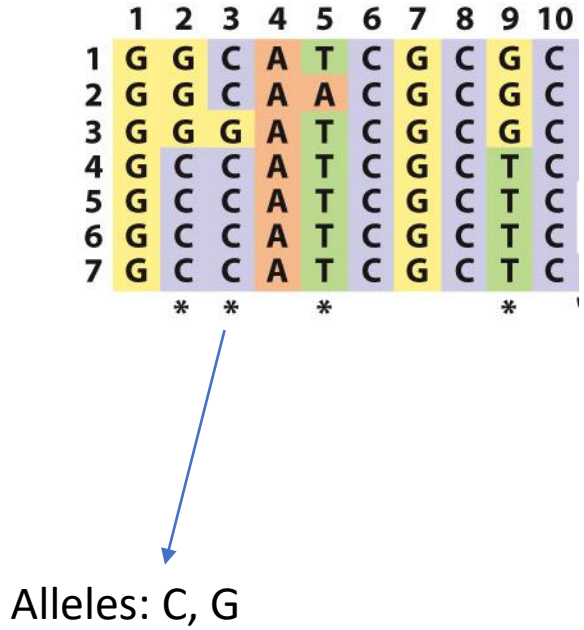
Defining genetic variation in the human genome

- Human genome is _____ bases long
- Genetic variant: any position in the genome that varies among individuals
  - **Single Nucleotide Polymorphism (SNP)**
- Allele: The nucleotides at a given genetic variant
- Genotype: The nucleotides an individual carries at a given position



Defining genetic variation in the human genome

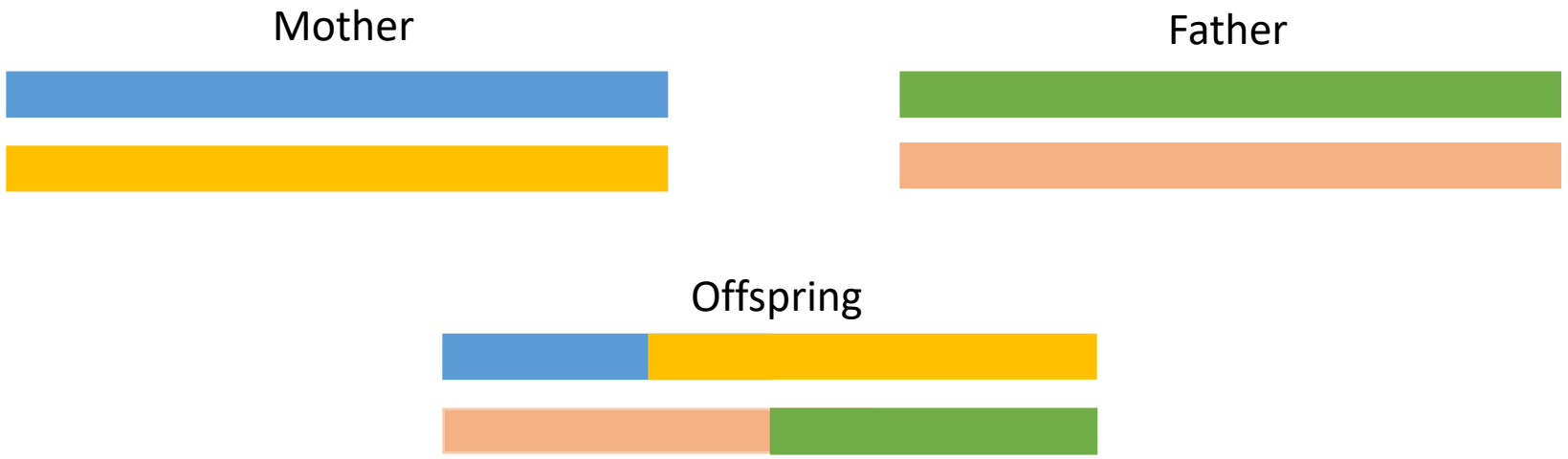
- Human genome is _____ bases long
- Genetic variant: any position in the genome that varies among individuals
  - **Single Nucleotide Polymorphism (SNP)**
- Allele: The nucleotides at a given genetic variant
- Genotype: The nucleotides an individual carries at a given position



Do we test all positions in the genome for association with the disease?

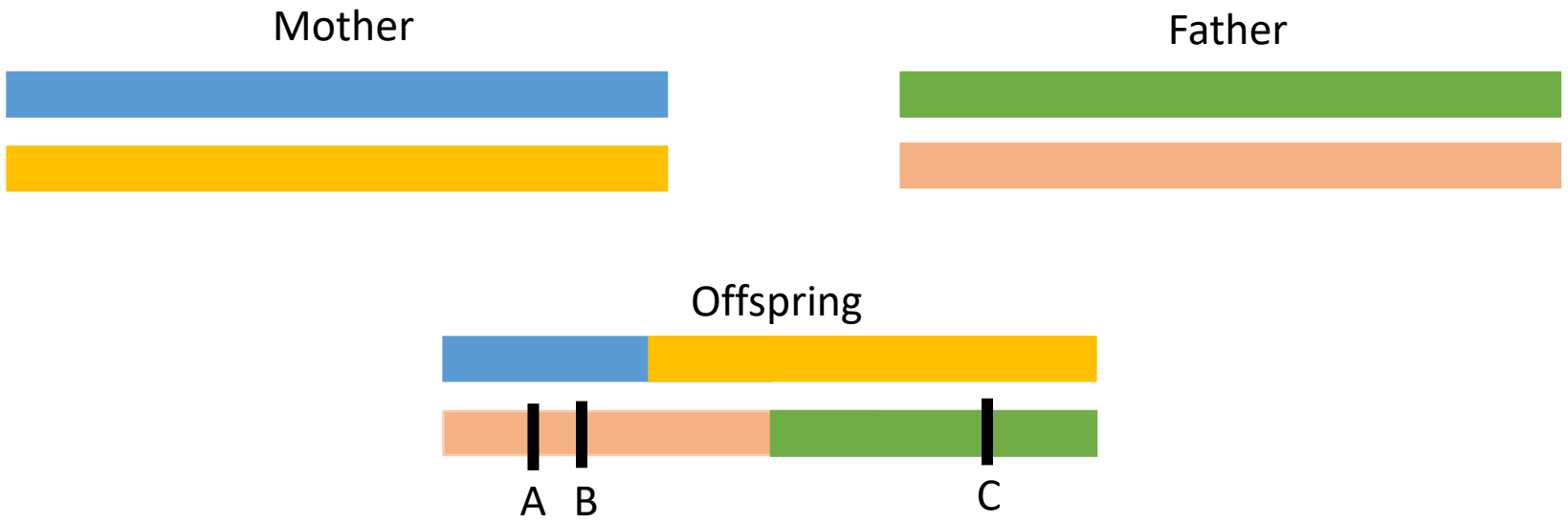
- Sequencing every position in the genome is expensive
- Can we select a subset of positions to test?

Genetic recombination and linkage



- Offspring gets a mosaic of the two maternal chromosomes, and a mosaic of the two paternal chromosomes
- Physically close locations in a genome have similar genotypes across individuals
  - They ‘travel together’ in individuals

Genetic recombination and linkage



- Offspring gets a mosaic of the two maternal chromosomes, and a mosaic of the two paternal chromosomes
- Physically close locations in a genome have similar genotypes across individuals
  - They ‘travel together’ in individuals

Linkage allows us to use ‘tag SNPs’

- Can use a single marker as a ‘signpost’ for an entire region
- These signposts are called ‘tag SNPs’
- We use 1 million tag SNPs in GWAS
  - 0.03% of the total genome

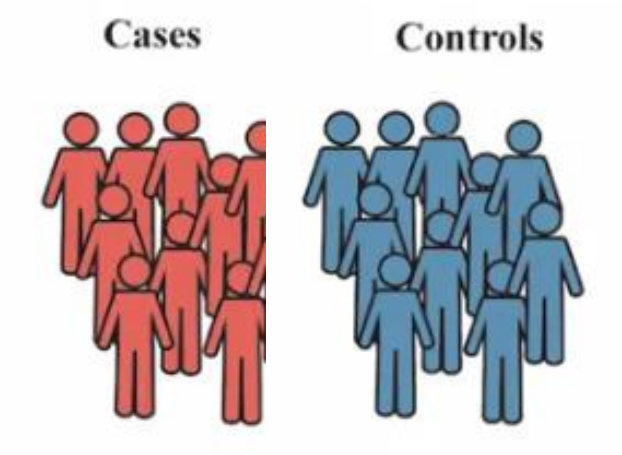
Tag SNP

Linkage allows us to use ‘tag SNPs’

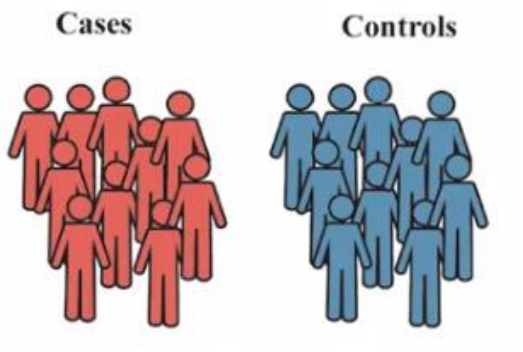
- Can use a single marker as a ‘signpost’ for an entire region
- These signposts are called ‘tag SNPs’
- We use 1 million tag SNPs in GWAS
  - 0.03% of the total genome

Tag SNP

What data from a GWAS looks like

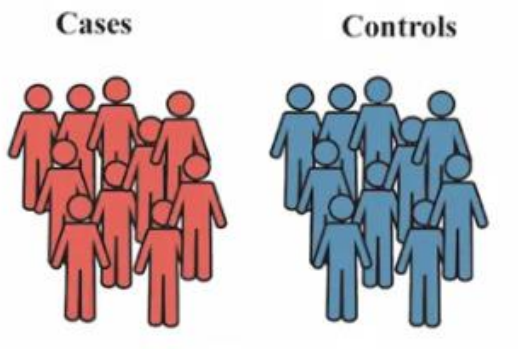


What data from a GWAS looks like



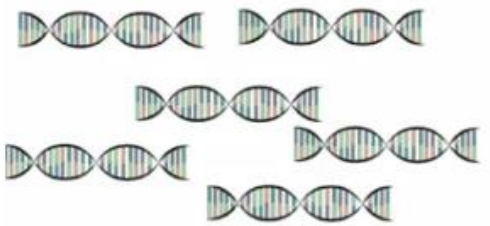
Collect DNA from each sample

What data from a GWAS looks like

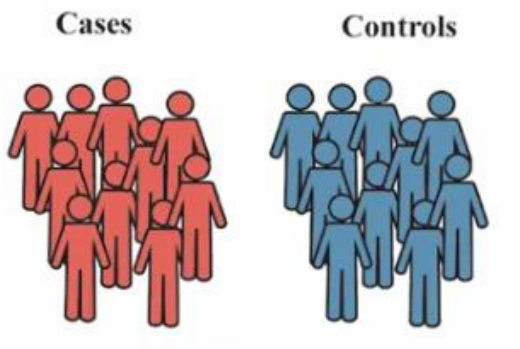


Collect DNA from each sample

Fragment DNA into pieces

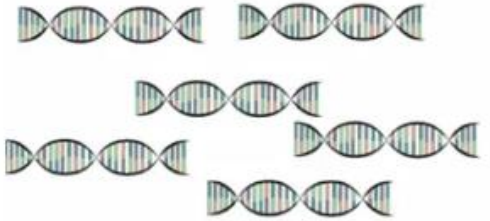


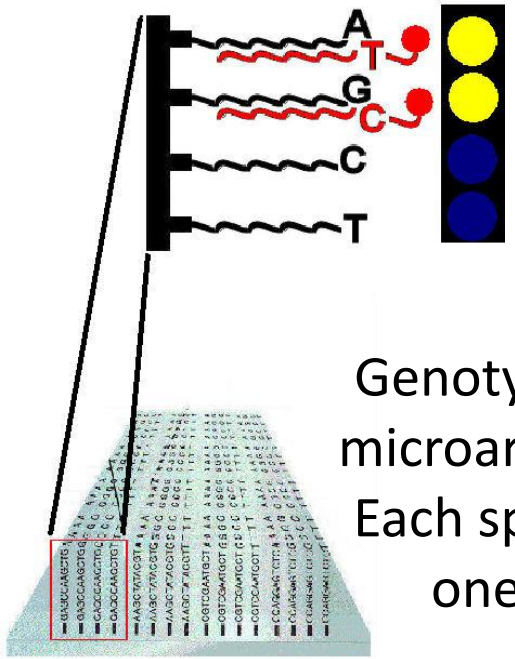
What data from a GWAS looks like



Collect DNA from each sample

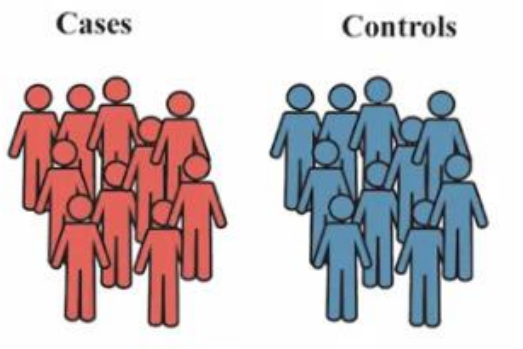
Fragment DNA into pieces





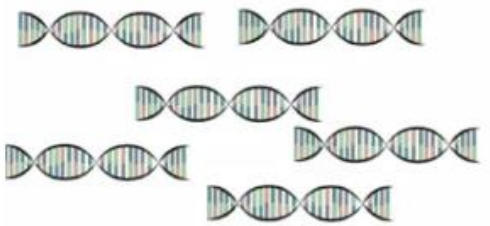
Genotype with microarray chip: Each spot tests one SNP

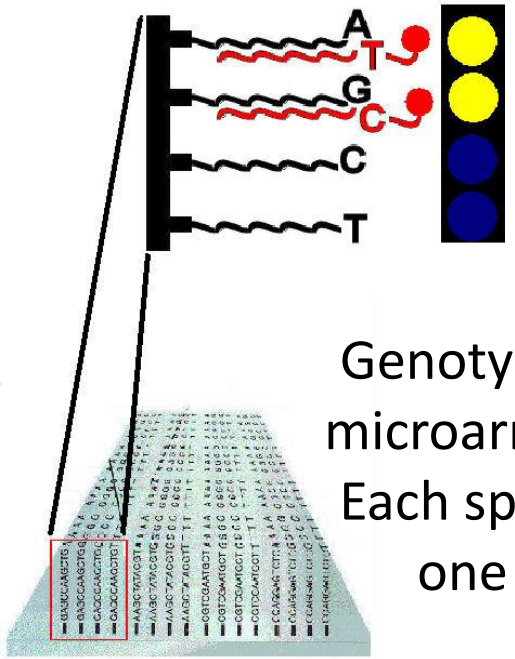
What data from a GWAS looks like



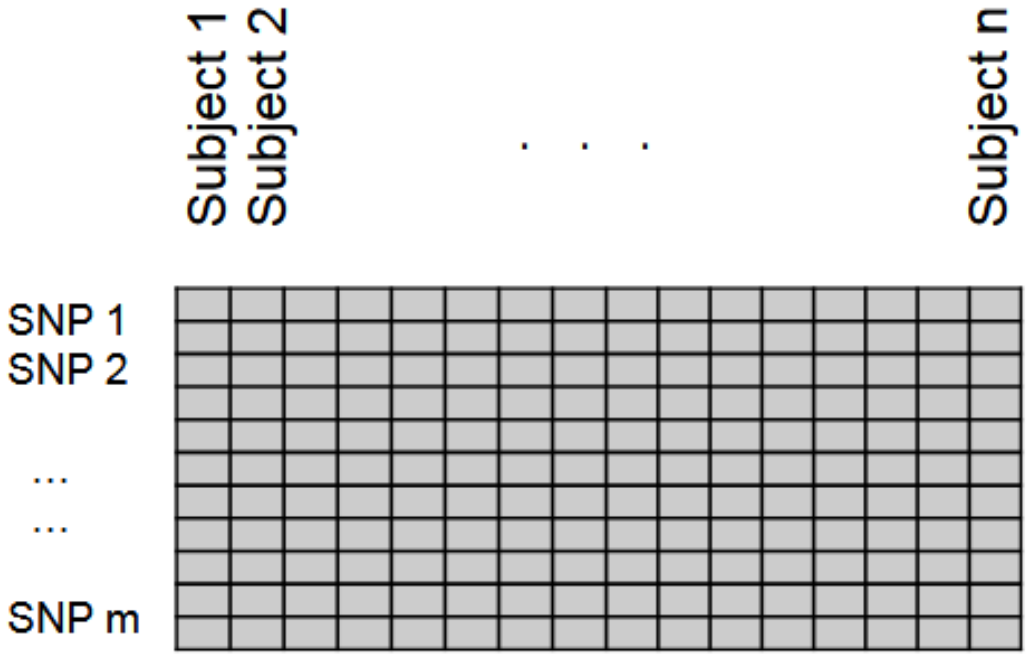
Collect DNA from each sample

Fragment DNA into pieces

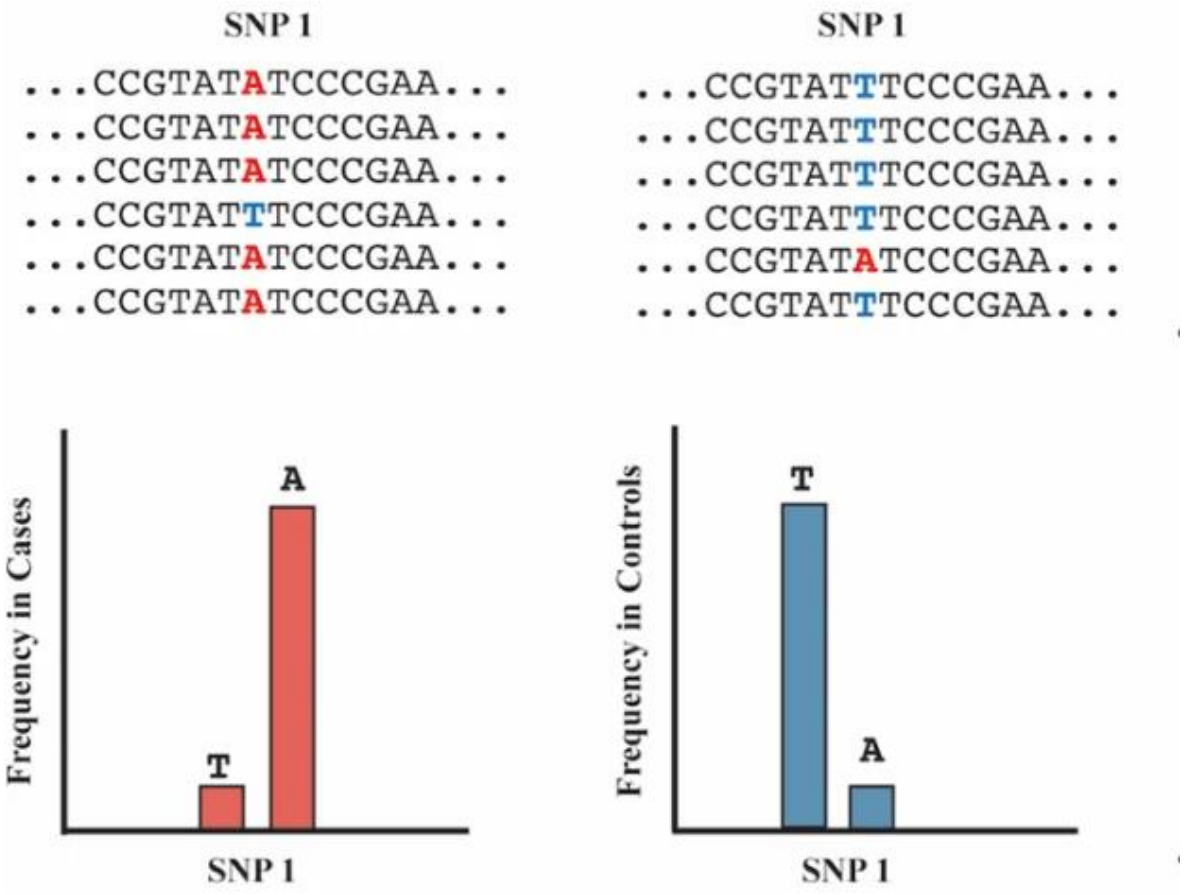




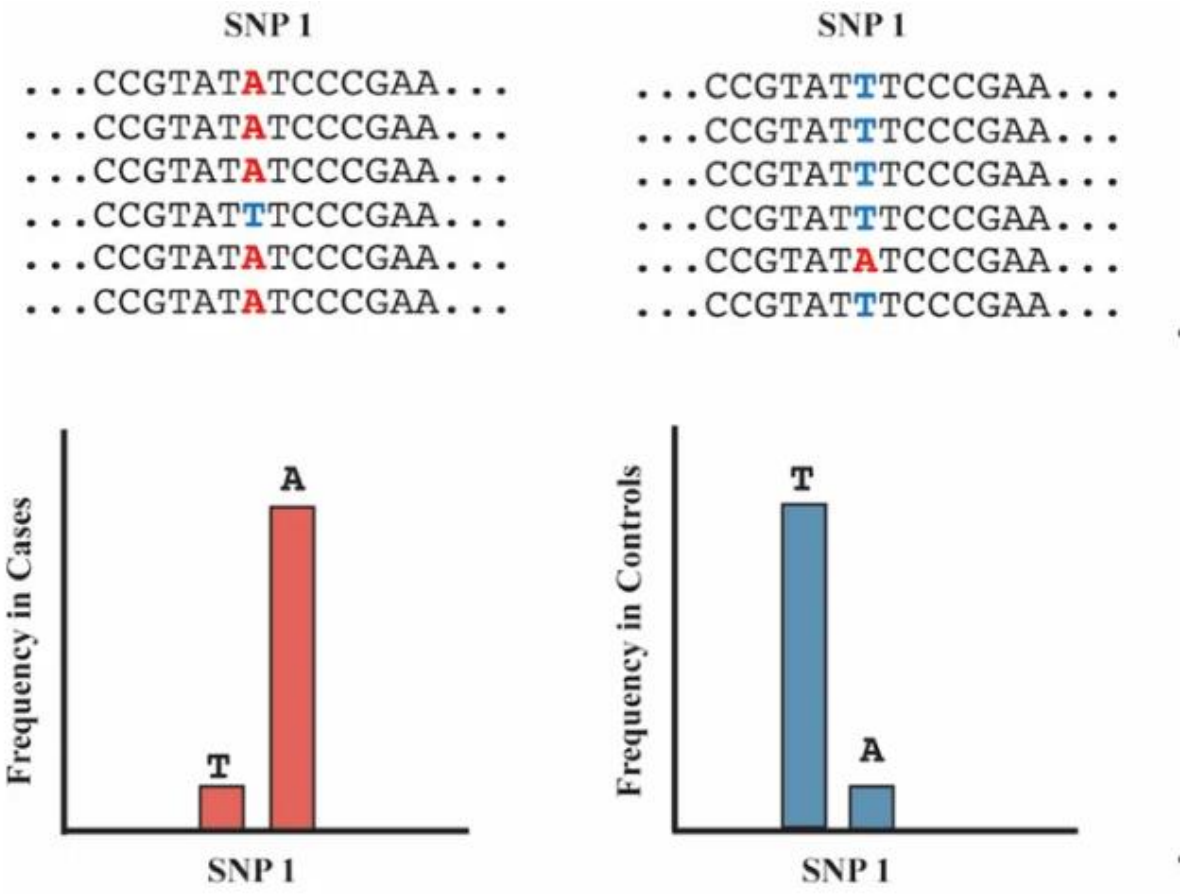
Genotype with microarray chip: Each spot tests one SNP



2. Testing each SNP for association



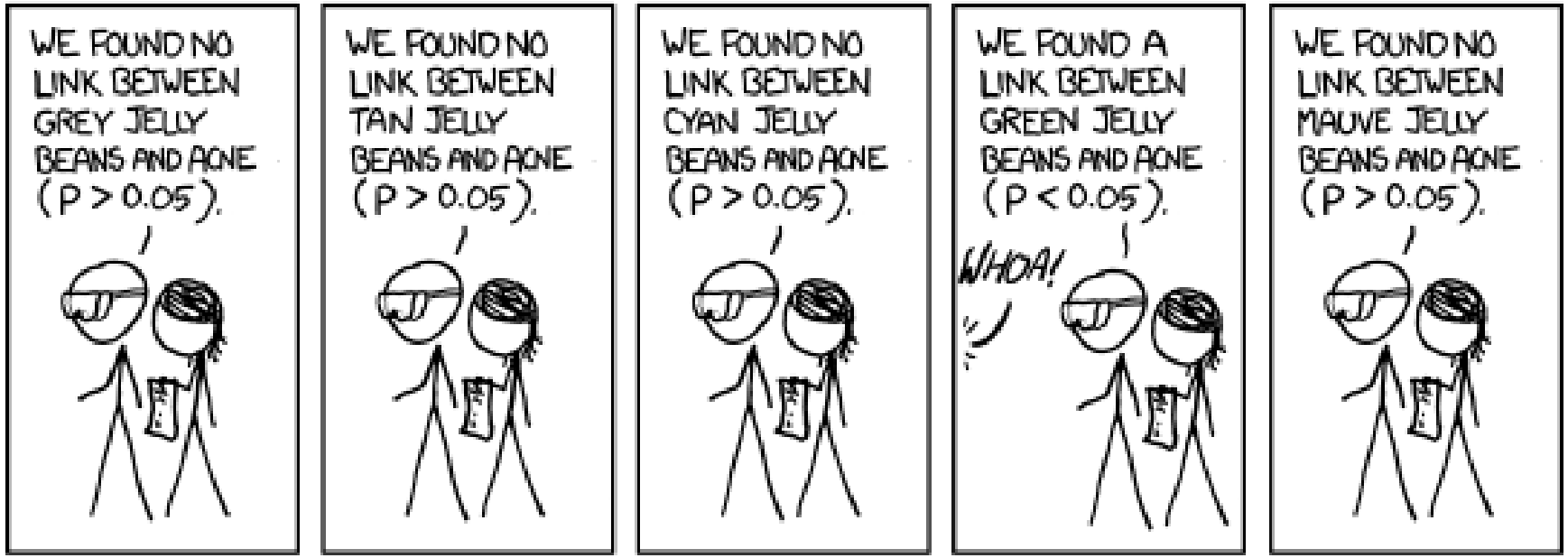
2. Testing each SNP for association



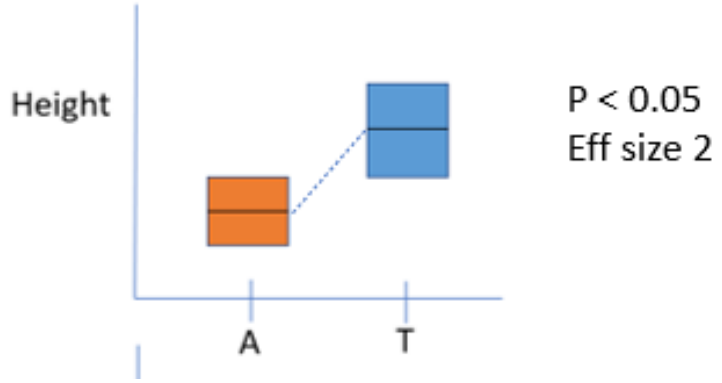
- Test significance of association with allele at SNP1 with disease status
  - Chi-square test/linear model
- Check P-value to test significance of association
- $P < 0.05$: significant
- $P > 0.05$: model is not significant

P-value threshold for GWAS significance

- The more tests you run, the more likely one of these tests reaches P < 0.05 **just by chance**
- **P-value threshold: 5e-08 (0.05/1,000,000)**
  - Correcting for **'multiple testing'**



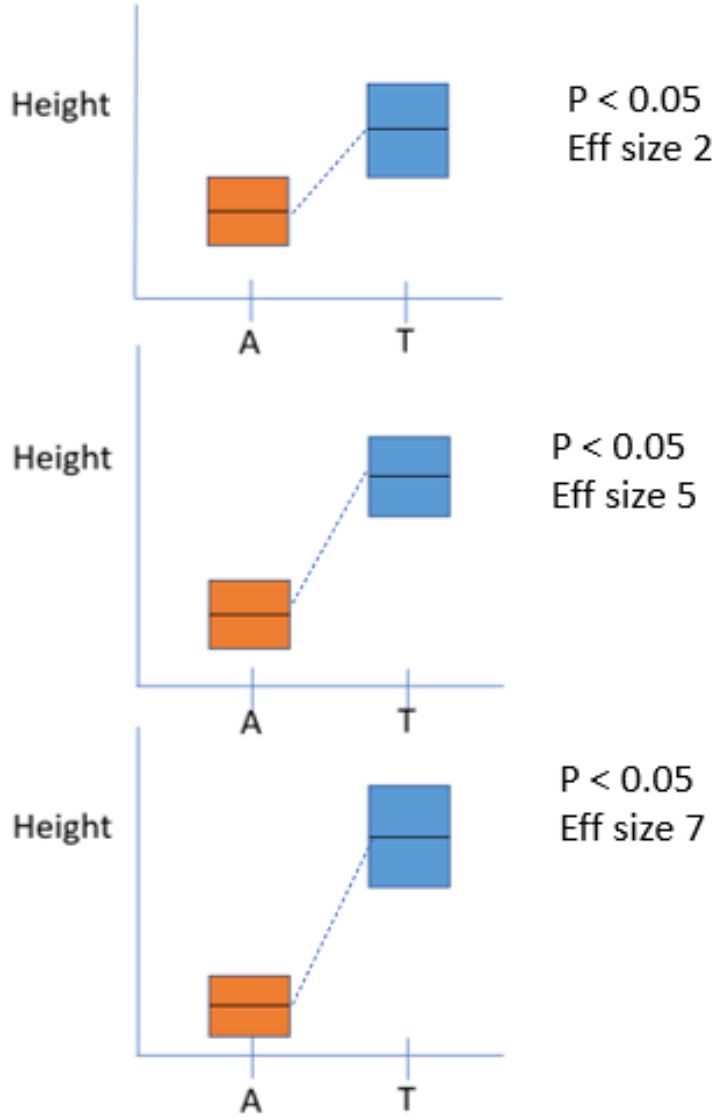
2b. Metrics of interest



- P-value:
  - How significant is the association between the SNP and the trait?
- Effect size:
  - What is the **impact** of the SNP on the trait?
  - Does having a ‘T’ allele increase your risk of disease by 2-fold? 20 fold? 200-fold?

2b. Metrics of interest

- P-value:
  - How significant is the association between the SNP and the trait?
- Effect size:
  - What is the impact of the SNP on the trait?
  - Does having a ‘T’ allele increase your risk of disease by 2-fold? 20 fold? 200-fold?



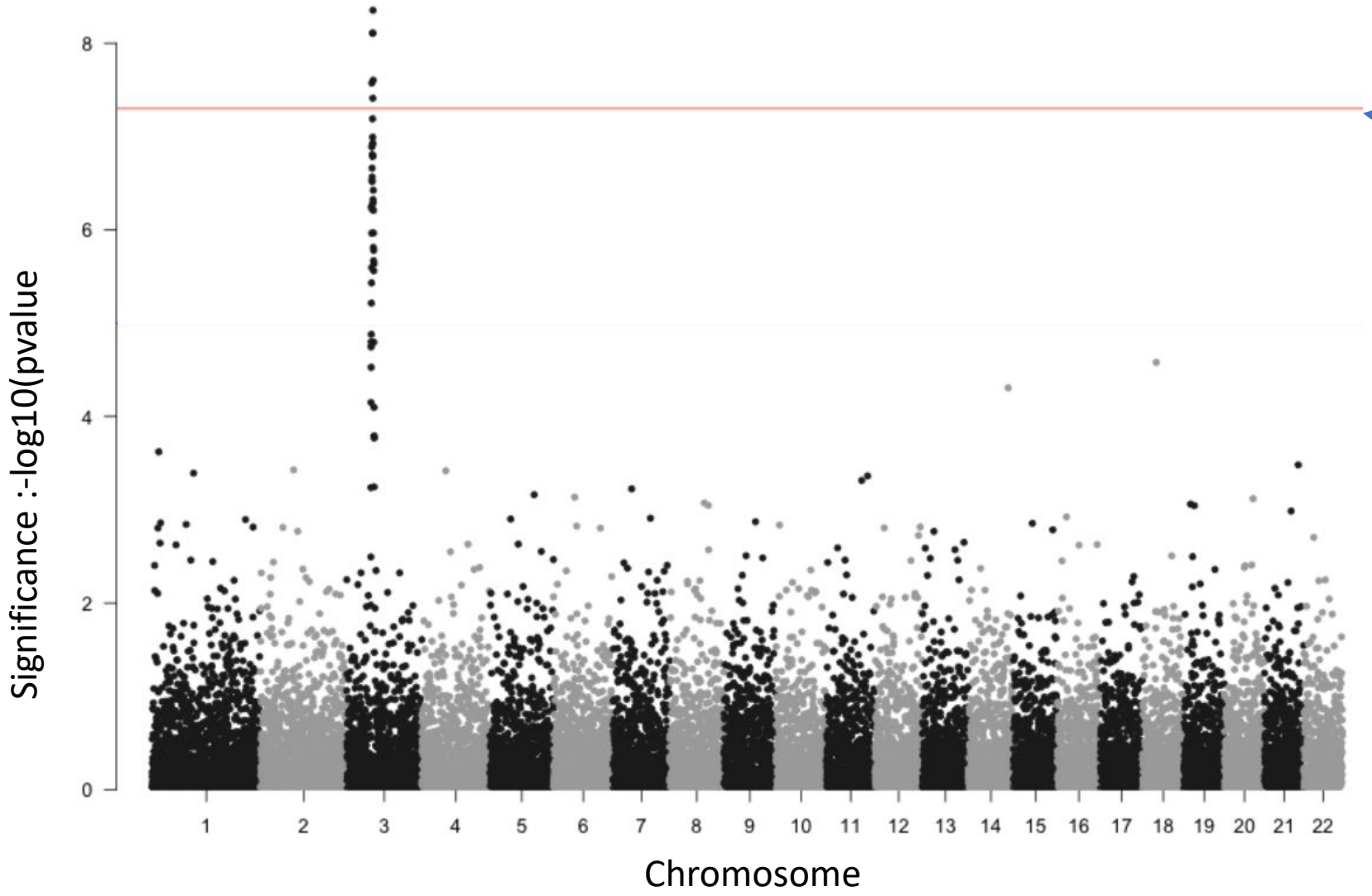
		My Risk	Population Risk	
Colorectal Cancer	★★★★	8.9%	5.6%	1.60x
Rheumatoid Arthritis	★★★★	4.6%	2.4%	1.94x
Type 1 Diabetes	★★★★	2.1%	1.0%	2.08x
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.43%	0.36%	1.21x
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.28%	0.23%	1.22x
Bipolar Disorder	★★★★	0.15%	0.10%	1.44x

- Colorectal Cancer:
  - 1 SNP with P-value < 5e-08
  - I have the ‘risk allele’, increasing my risk 1.6 fold

Summary of steps in a GWAS

- Collect samples (cases and controls)
- Genotype samples across 1 million SNPs
- Perform statistical test for each SNP
- Filter those SNPs with **P-value < 5e-08**
  - (0.05 * 1,000,000)

Visualizing GWAS results in a Manhattan plot



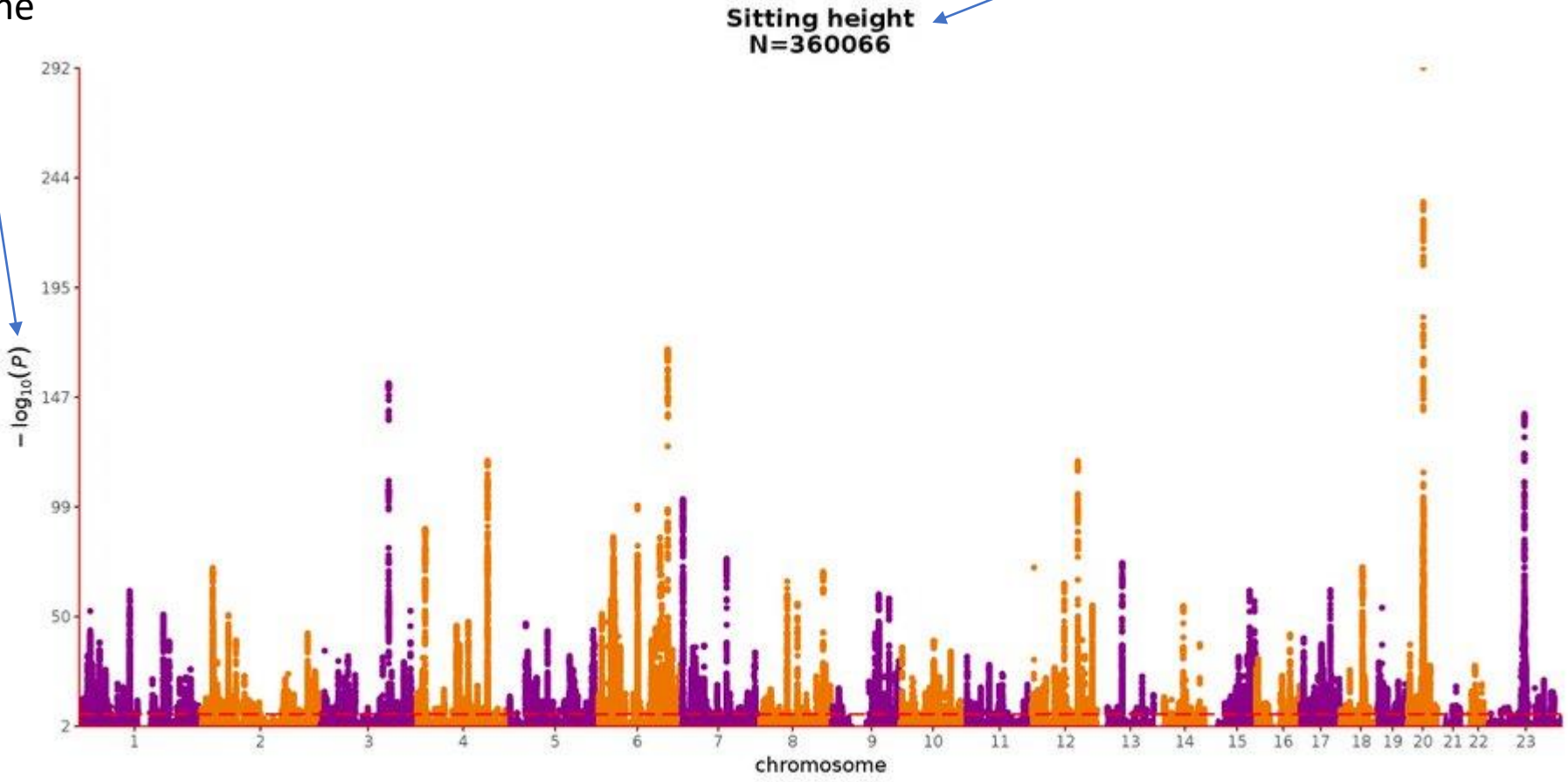
Threshold for significance ($P < 5e\text{-}08$)

- Each point on the plot = 1 SNP

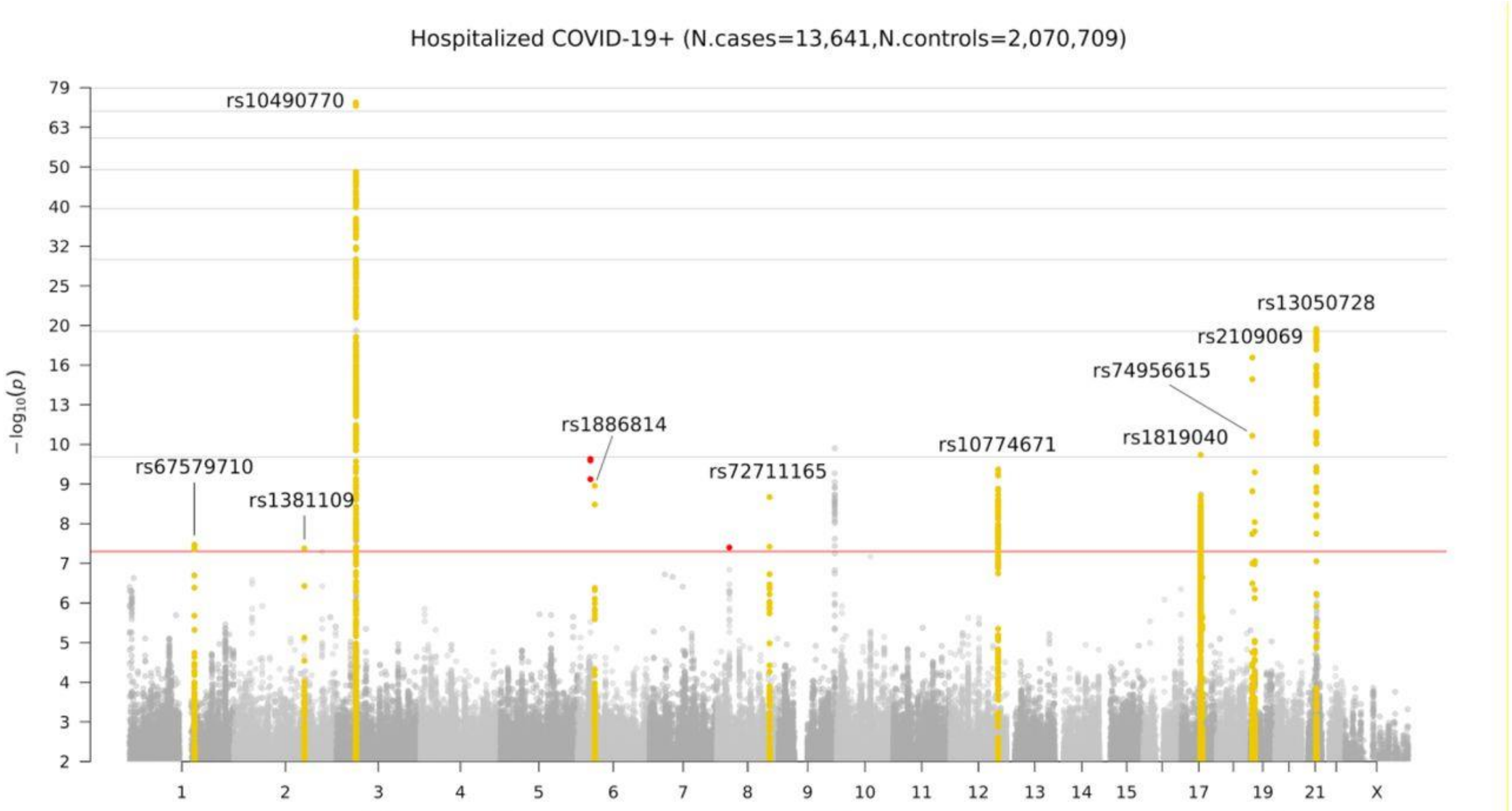
GWAS of height shows signals across the genome

Y-axis represents significance, the higher the more significant

Sample size of gwas



Threshold for significance (P < 5e-08)


Hospitalized COVID-19+ (N.cases=13,641,N.controls=2,070,709)
rs10490770
rs67579710
rs1381109
rs1886814
rs72711165
rs10774671
rs74956615
rs1819040
rs2109069
rs13050728
$-\log_{10}(p)$
79
63
50
40
32
25
20
16
13
10
9
8
7
6
5
4
3
2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
X

What do GWAS tell us about biology?

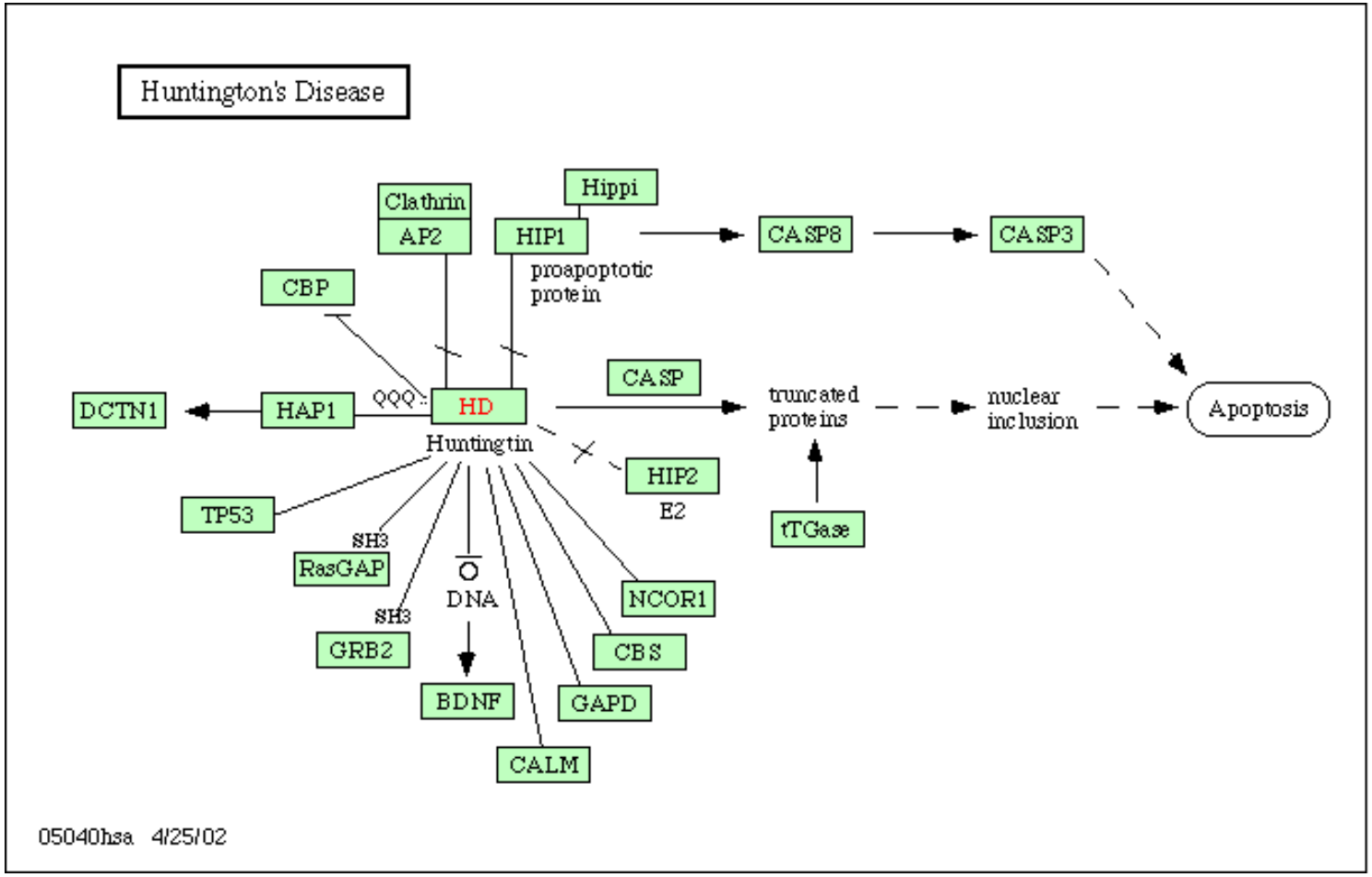
Biology of human traits from GWAS

Biology of human traits from GWAS

- Most traits are highly **polygenic**
  - More than 100 genetic variants associated with schizophrenia, more than 200 with height, more than 500 with cholesterol
  - Cumulative genetic risk of an individual is **additive effect of all associated variants**

How are there so many risk factors?

- Genes act in pathways → Changes in multiple pathways lead to risk of disease



How do we compute a person's genetic risk for disease?

GWAS result:

- One significant SNP (P < 5e-08)
- SNP rs10454798 with alleles A,C
  - Effect size for allele A = 0.01
  - Effect size for allele C = 0

How do we compute a person's genetic risk for disease?

GWAS result:

- One significant SNP (P < 5e-08)
- SNP rs10454798 with alleles A,C
  - Effect size for allele A = 0.01
  - Effect size for allele C = 0

- My genotype at SNP rs10454798
  - AA (2 copies of the risk allele)
- My 'added' risk: 2* 0.01 = 0.02

- A person with a GG genotype has no added risk
- A person with an AG genotype has added risk of 0.01

Question: Calculate the genetic risk for an individual for ‘handedness’

- GWAS results

SNP	Effect allele	
rs3798220	G	-0.47
rs8099	T	0.512

SNP	My Genotype	Risk from this SNP
rs3798220	AG	
rs8099	AA	

Total risk =

Biology of human traits from GWAS

- Most traits are highly **polygenic**
  - More than 100 genetic variants associated with schizophrenia, more than 200 with height, more than 500 with cholesterol
- Individual genetic variants have **small effect sizes**

Biology of human traits from GWAS

- Most traits are highly **polygenic**
  - More than 100 genetic variants associated with schizophrenia, more than 200 with height, more than 500 with cholesterol
- Individual genetic variants have **small effect sizes**
- Most causal variation is **non-coding**

Most causal variation is non-coding

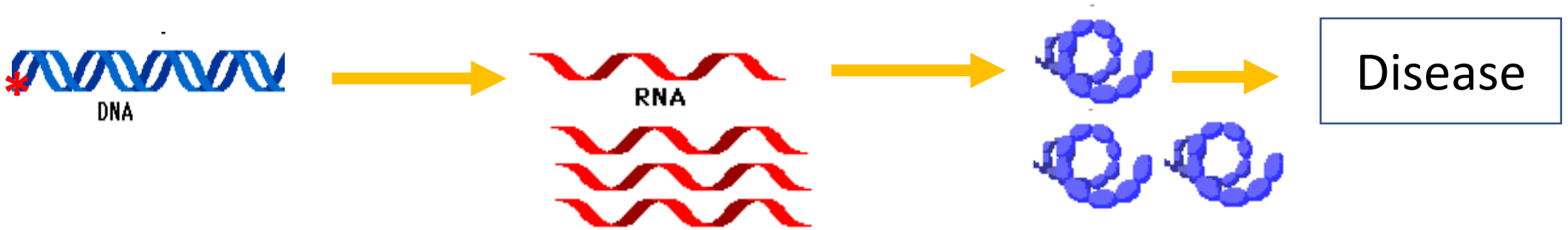
Most causal variation is non-coding

Previous model:

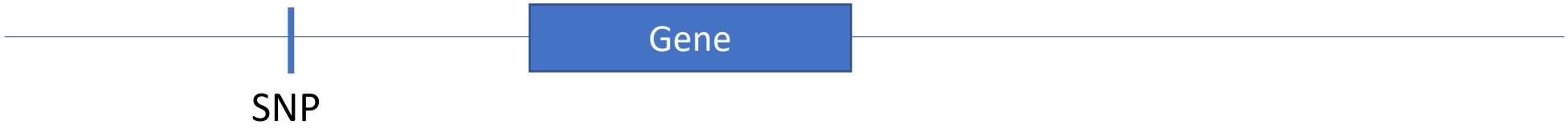
Most causal variation is non-coding

Previous model:

Current model:

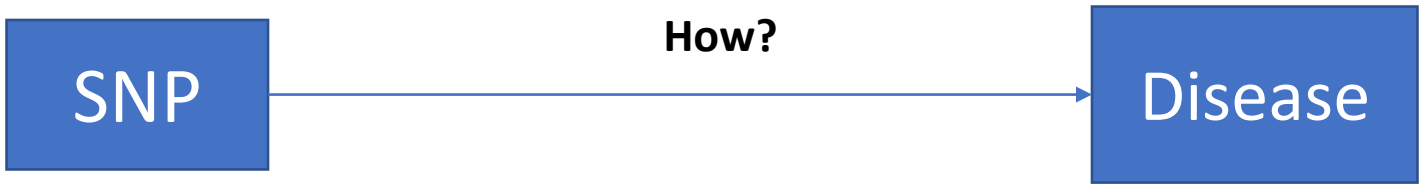


Regulatory variation is usually outside a gene



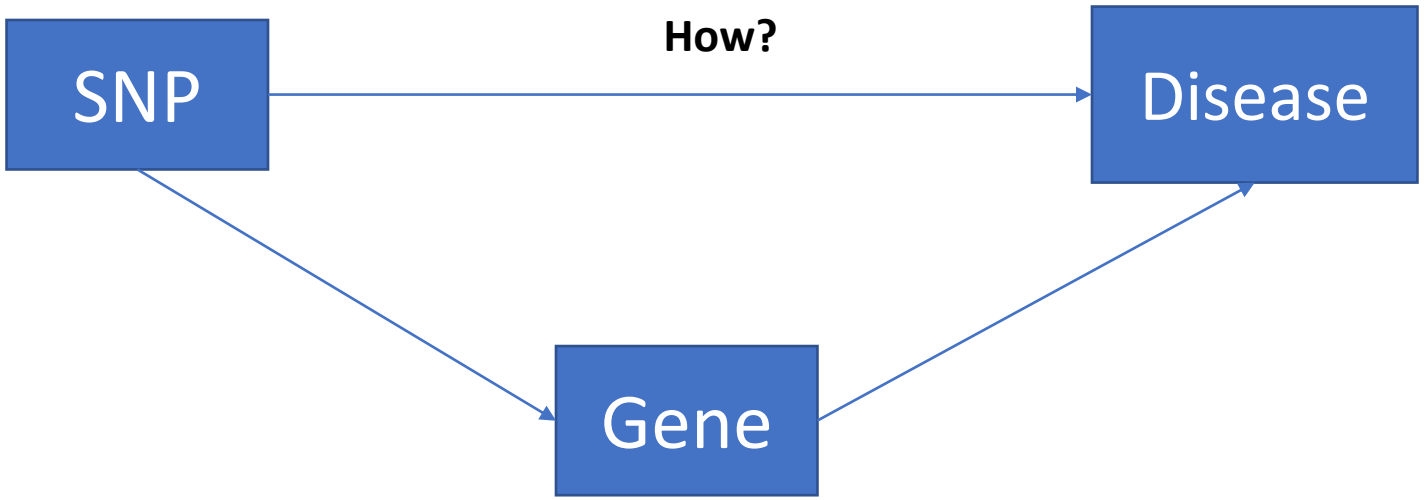
What GWAS can/cannot find

- What is the causal gene?
  - If most causal SNPs are regulatory, how do we find out which genes are involved in the disease?



What GWAS can/cannot find

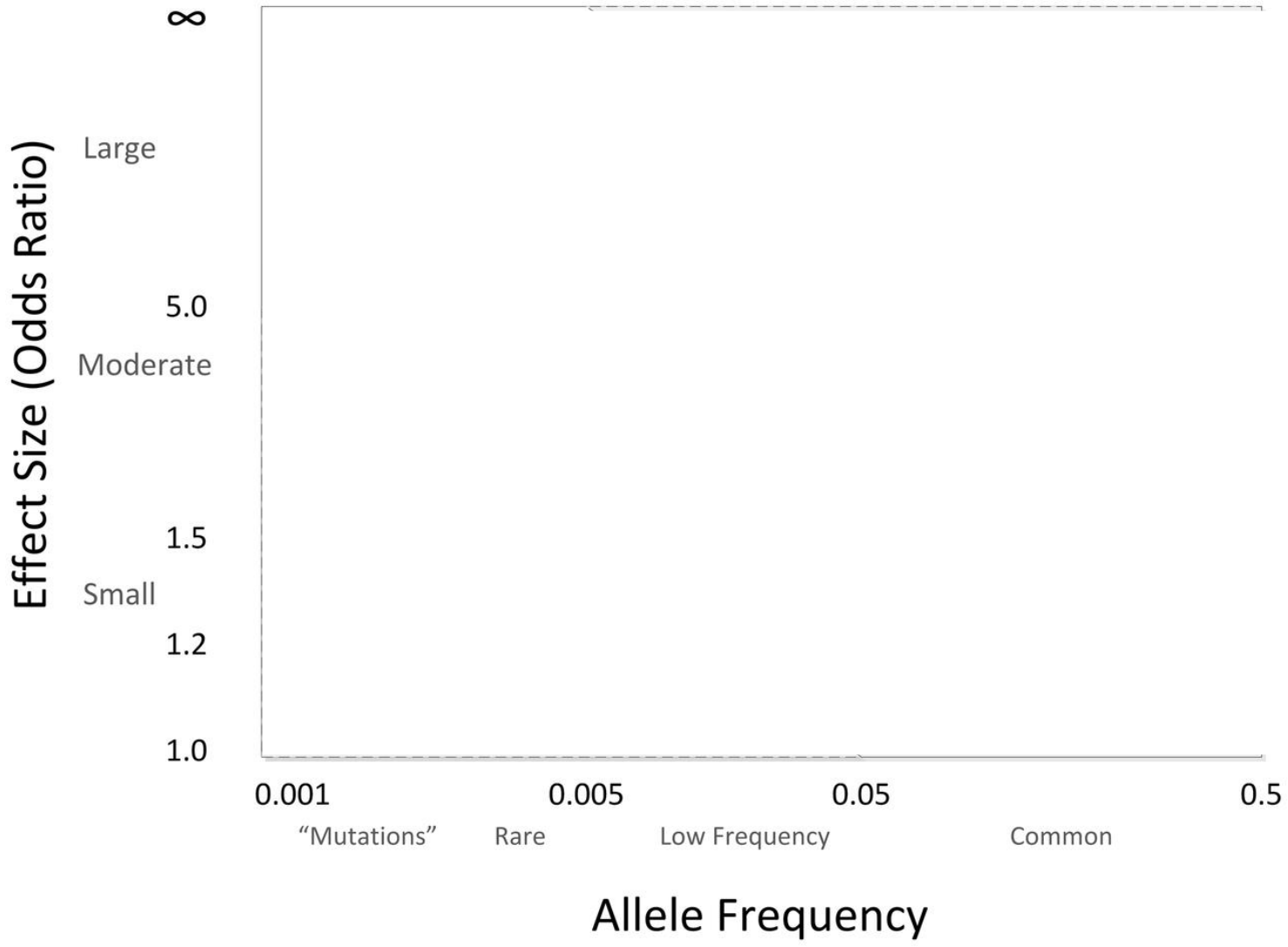
- What is the causal gene?
  - If most causal SNPs are regulatory, how do we find out which genes are involved in the disease?



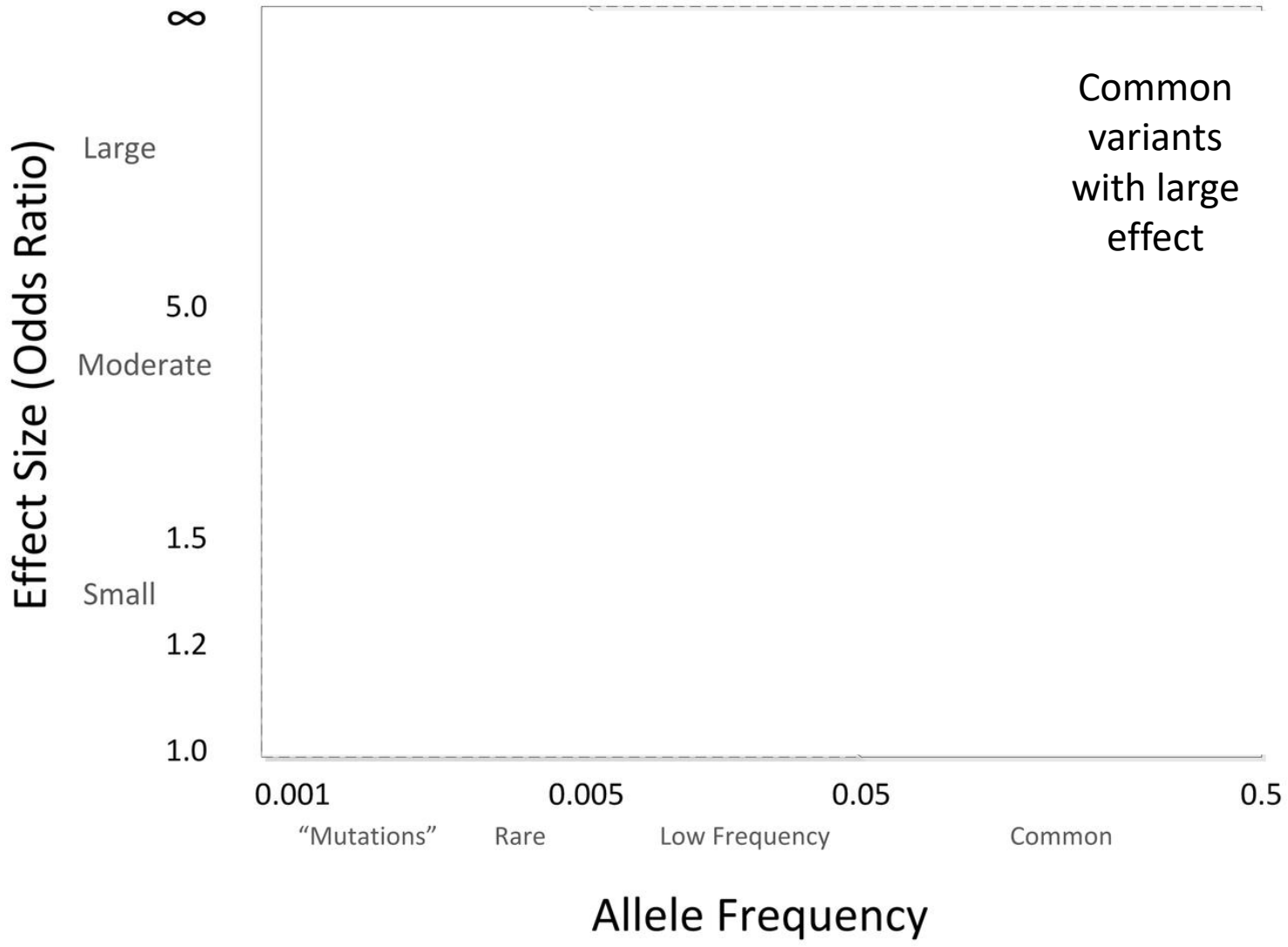
What GWAS can/cannot find

- What is the causal variant?
  - It is difficult to distinguish between variants in LD
  - The true causal variant is difficult to identify
  - The top associated variant is a proxy for the causal variant

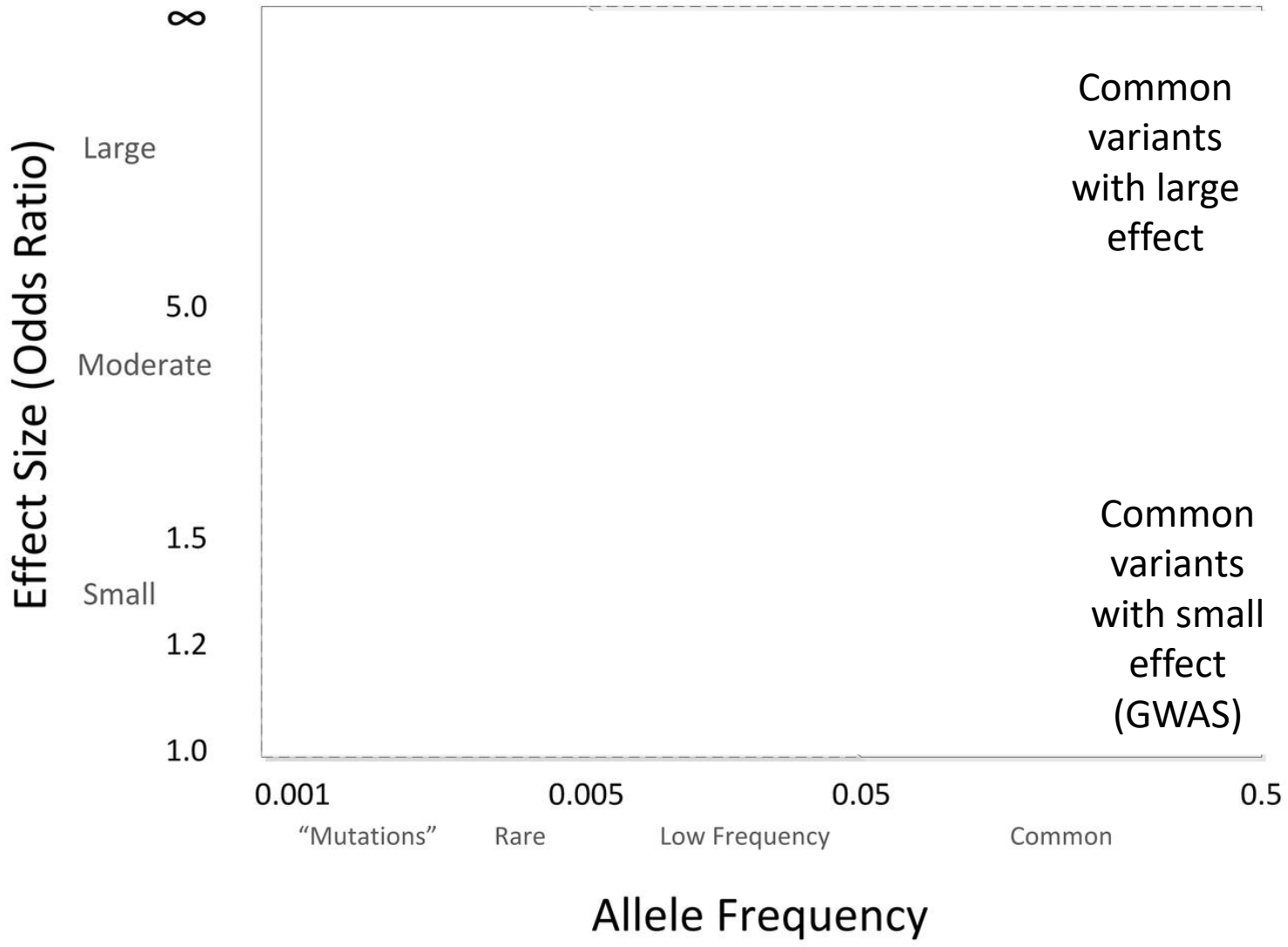
What GWAS can/cannot find



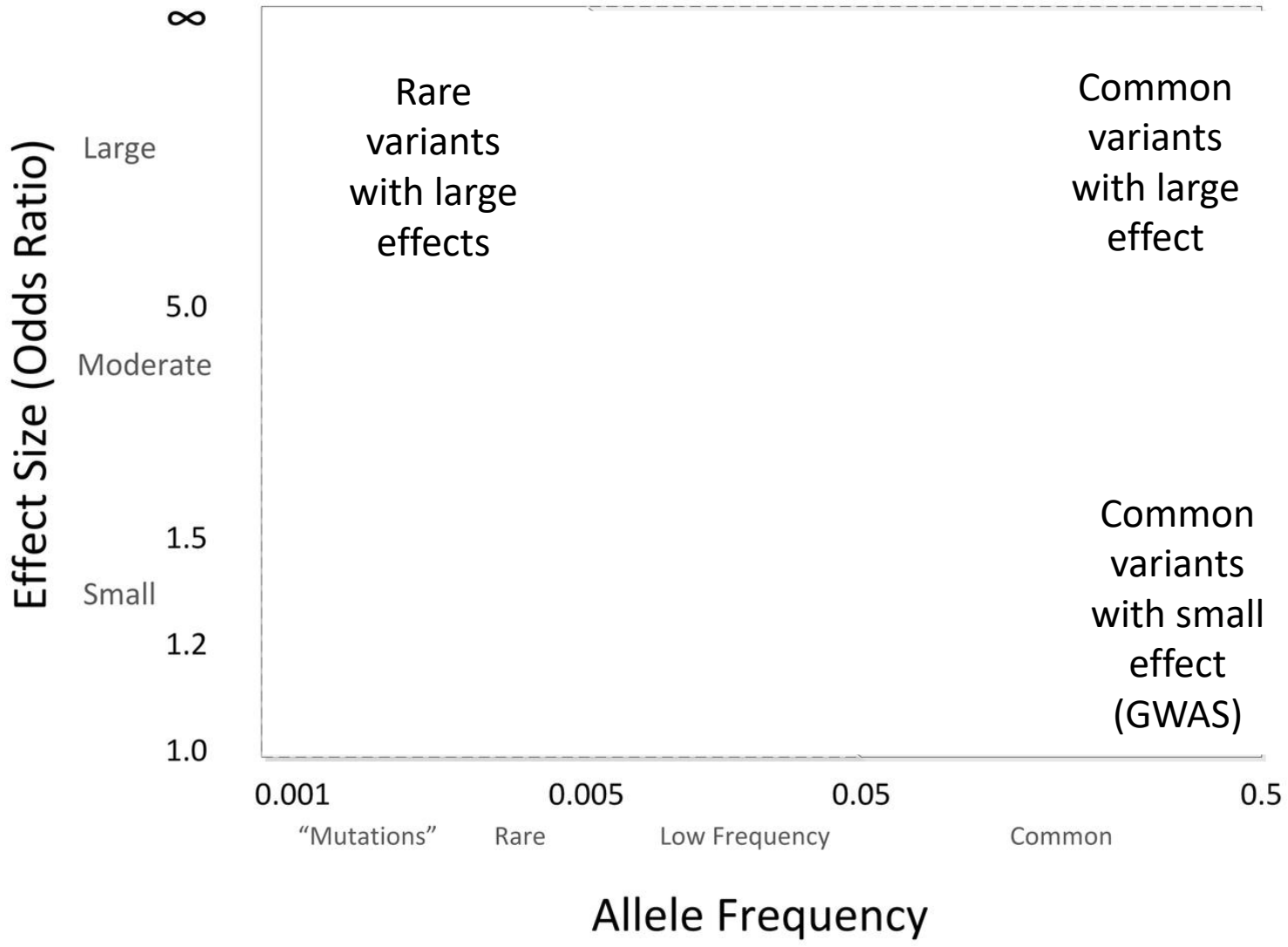
What GWAS can/cannot find



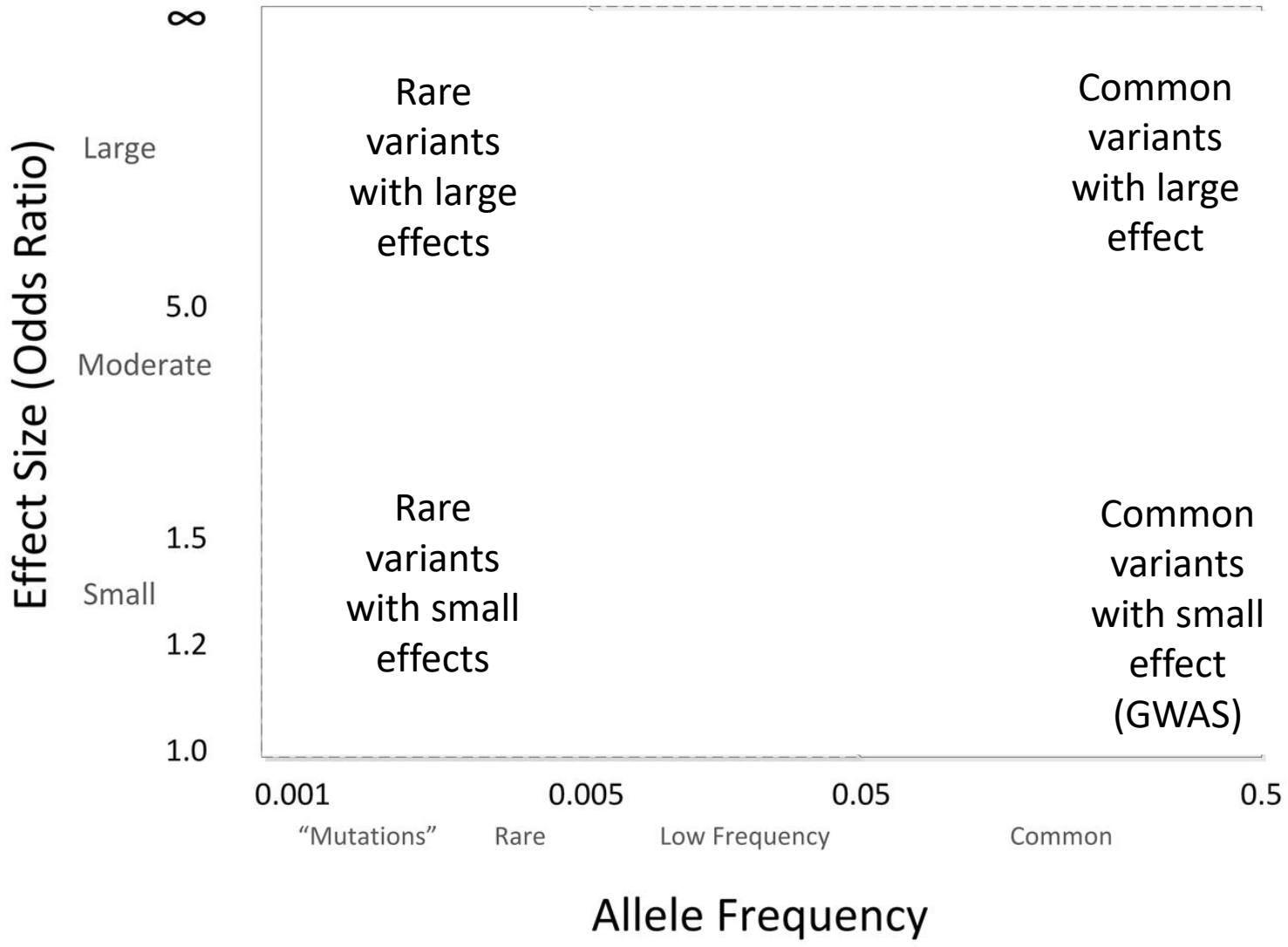
What GWAS can/cannot find



What GWAS can/cannot find



What GWAS can/cannot find



More GWAS: GWAS and Population Stratification

- Most GWAS are conducted on European populations
  - Why?
- How transferrable are the results from a European GWAS into GWAS of another ancestry?

Review

- GWAS allow us to study the genetic basis of _____ traits, and help us find _____ variants of _____ effects.
- What does a GWAS result look like?
  - For each variant: P-value, Effect size
  - Effect size of 0 = no effect on phenotype
- GWAS show us that
  - Such traits are highly polygenic.
  - Most causal variants are non-coding

Some questions to think about

- Hypothesis-free experimental approach
  - ‘Hypothesis-generating’ research

Questions?

- shwetar@pennmedicine.upenn.edu