Odds Ratio and G.W.A.S.

“Genome-Wide Association Studies”

What is G.W.A.S?

- Genome-wide **association** studies
  - Finding associations between particular alleles or genetic information and the state of individuals, ie. disease/no disease or tall/short
  - One caveat that is underappreciated:

____

Because SNPs are sprinkled throughout the genome, they are used as ‘markers’ for linked loci that are probably the actual contributing cause of state whereas the SNPs are probably not actually responsible for the state

____

What is G.W.A.S?

- GWAS explained in cartoon format:

  https://www.broadinstitute.org/visuals/explainer-genome-wide-association-studies

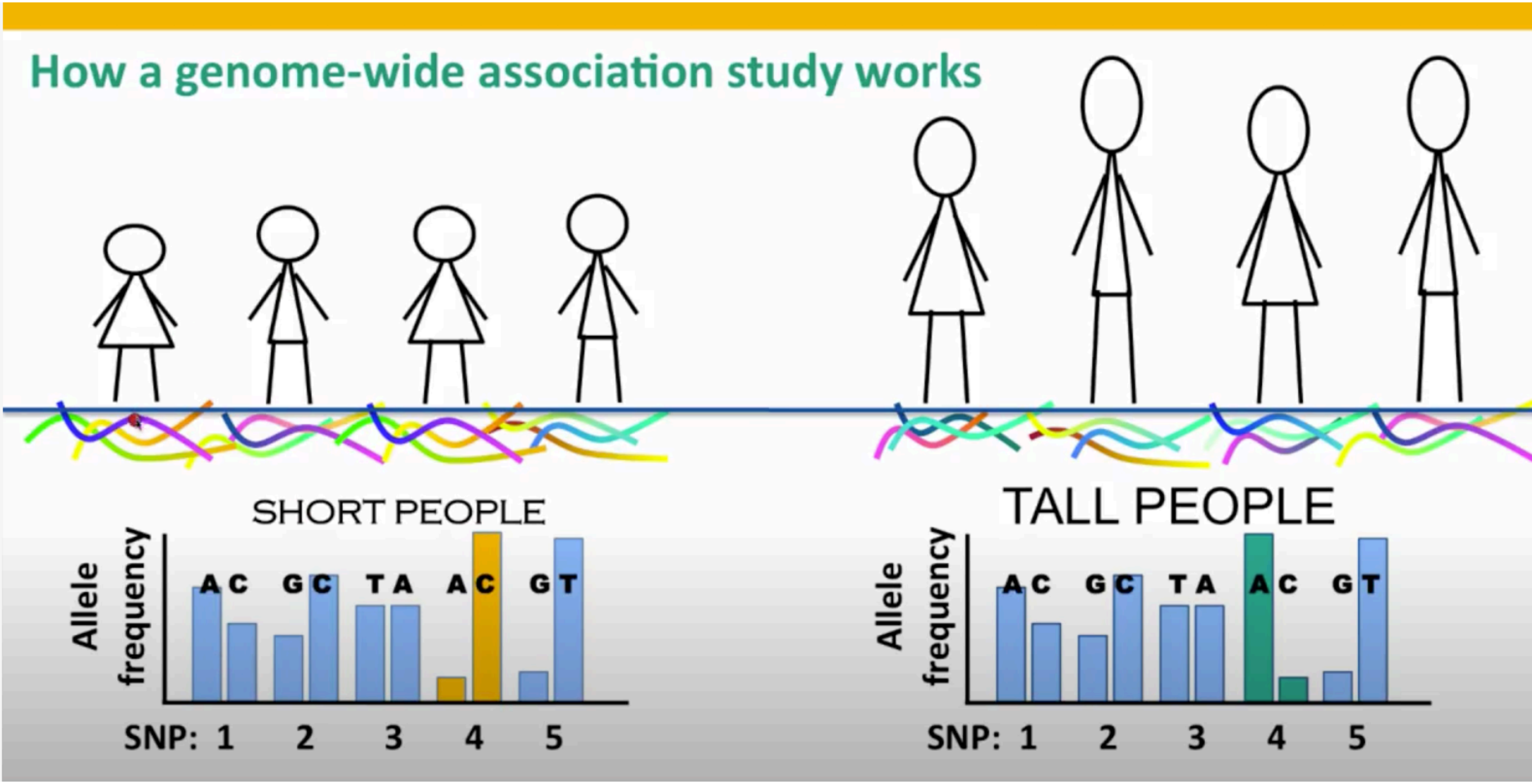
- GWAS explained in a short video in the fantastic series called “Useful Genetics” by Professor Rosie Redfield:

  https://www.youtube.com/watch?v=-WrmAvL7I1Y&list=PLgh8WcYegg44s-NIxYPVHQsCa0-bCR7-G&index=5

  - She uses the cartoon of extremely tall group of people versus extremely short people to illustrate how these SNPs are identified

What is G.W.A.S?

- GWAS explained in a short video in the fantastic series called “Useful Genetics” by Professor Rosie Redfield: https://www.youtube.com/watch?v=-WrmAvL7I1Y&list=PLgh8WcYegg44s-NIxYPVHQsCa0-bCR7-G&index=5



How a genome-wide association study works

What is G.W.A.S?

- There are sophisticated tools for finding these associations (that include accounting for testing multiple hypotheses simultaneously), but odds ratio is what they are based on, fundamentally.
- <u>Assumptions:</u>
  - Each SNP is an independent test
  - Associations are tested by comparing the frequency of the SNP state in two samples: individuals with the disease (or who are short) and in individuals without the disease (or who are tall)
  - You can also compare the frequency of the three genotypes (AA, Aa, or aa) between these two groups

What is G.W.A.S?

- Reminder: Odds Ratio basics
  - Measures the strength of association
  - **Odds = P/(1-P)**
    - Probability of winning is 50% means odds = 50/(1-50) = 1
    - Probability of winning is 75%, odds = 75/25 = 3
  - Surprise! Conditional probability again…. No one expects conditional probability to show up but it does, all the time. Repeatedly.

$$\textbf{Odds Ratio} = \frac{\textbf{odds(Disease|Allelic State)}}{\textbf{odds(Disease|Alternate Allelic State)}}$$

What is G.W.A.S?

- Reminder: Odds Ratio basics

$$\text{Odd Ratio} = \frac{\text{odds(Disease|Allelic State)}}{\text{odds(Disease|Alternate Allelic State)}}$$

Example: OR for getting a particular disease when you have a genotype of TT (compared to AT at the same locus?)

P(Disease|Genotype is ‘TT’) =0.75

P(Disease|Genotype is ‘AT’) =0.25

$$\text{OR} = \frac{(0.75/0.25)}{(0.25/0.75)} = \frac{3}{1/3} = 9$$

What is G.W.A.S?

- Reminder: Odds Ratio basics

$$\textbf{Odd Ratio} = \frac{\textbf{odds(Disease|Allelic State)}}{\textbf{odds(Disease|Alternate Allelic State)}}$$

Example: OR for getting a particular disease when you have a genotype of TT (with unknown alternate allele) compared to a known population-wide risk of 0.4? ← this is the way to do this to be informative for a particular individual result.

P(Disease|Genotype is ‘TT’) =0.75

P(Disease|Genotype is unknown) =0.25

$$\textbf{OR} = \frac{(0.75/0.25)}{(0.4/0.6)} = \frac{3}{2/3} = 9/2 = 4.5$$

Example: There is a known **rs1234567** that is associated with a particular cancer (important note: this associations is, as always, present in specific populations, as we will discuss). C is presumed to be the risk allele in this case (it is present in 56.5% of the cases and only 48.9% of the controls)

	"CC"	"CT"	"TT"
Cases:	250	375	150
Controls:	460	940	500

Cases:

C = 2*250+375=875

T = 375+2*150=675

Controls:

C = 2*460+940= 1860

T = 940+2*500 = 1940

Example: There is a known **rs1234567** that is associated with a particular cancer (important note: this associations is, as always, present in specific populations, as we will discuss). C is presumed to be the risk allele in this case (it is present in 56.5% of the cases and only 48.9% of the controls)

	C	T
Cases:	**875**	**675**
Controls:	**1860**	**1940**

OR_c = odds(disease|C)/odds(disease|T)

$$OR_c = \frac{\left(875/1550\right)\left(675/1550\right)}{\left(1860/3810\right)\left(1940/3810\right)} = \frac{875/675}{1860/1940} = \frac{875*1940}{675*1860} = 1.35$$

	C	T
Cases:	**875**	**675**
Controls:	**1860**	**1940**

OR_c = odds(disease|C)/odds(disease|T)

$$\boxed{OR_c = \frac{\left(\frac{875}{1550}\right)\left(\frac{675}{1550}\right)}{\left(\frac{1860}{3810}\right)\left(\frac{1940}{3810}\right)} = \frac{\frac{875}{675}}{\frac{1860}{1940}} = \frac{875*1940}{675*1860} = 1.35}$$

Using the formula (given in the textbook), we can calculate 95% and 99% confidence intervals:

$\ln(1.35) - 1.96(0.06064187) < \ln(OR) < \ln(1.35) + 1.96(0.06064187)$

$0.18 < \ln(OR) < 0.42$

convert back to OR using $e^{0.18}$ and $e^{0.42}$:

$1.20 < OR < 1.52$

This interval does not contain an odds ratio of ‘1’, so the hypothesis of no effect can be rejected with 95% confidence

The problem with this OR calculation is that is done with respect to the low risk allele (in this case, the T):

$$\boldsymbol{OR_c = odds(disease|C)/odds(disease|T)}$$

This example keeps our genetic data in the same format as what you encounter in our textbook (case: control)

What you actually need, to make this a useful predictor for a particular individual, is to account for prevalence of the disease (like we saw with prostrate and breast cancer examples when we looked at Bayes' theorem) and the genotype (or Allele) frequencies in a particular population: CEU, JPN, etc

This is a bit beyond the scope of this course, but it is important to note that for a OR to be meaningful for genetic data, you need to have a sense of prevalence of the disease in the particular population that you are sampling:

P(D) = prevalence = P(D|CC)P(CC) + P(D|Cc)P(cc) + P(D|cc)P(cc) **= 0.10**

If, there is a particular disease (myocardial infarction) associated with homozygous rs1234567 and we know that it is present in the sampled Caucasian population at 10%, and that the risk genotype is present in the same population at 24.8%, we can then get an odds ratio for the CC genotype:

OR_{CC} = odds(disease|CC)/odds(disease in avg population)
= (0.248/1-0.248)/(0.1/0.9) = **2.97**

Instead of using OR, you can convert these into Likelihood ratios. We will see likelihood much, much later in this course, but you can use them like so:

LR_C = P(C|Disease)/P(C|Control) = (875/1550)/(1860/3800)

= 1.15

LR_T = P(T|Disease)/P(T|Control) = (675/1550)/(1940/3800)

=0.85

Since Likelihood is conditioned on the disease or the control, it already takes into consideration the prevalence of the disease and is a bit easier to apply directly to individuals.

Finally, you can also multiply OR that are suspected to contribute to a disease together (including environmental contributors such as diet, exercise, smoking etc).

$$\underbrace{OR_{C*}OR_{C}*OR_{A}*..}_{\textbf{GENETIC}}\underbrace{..*OR_{smoking}*OR_{pollution}*}_{\textbf{ENVIRONMENTAL}}OR_{diet}$$

Contributions: