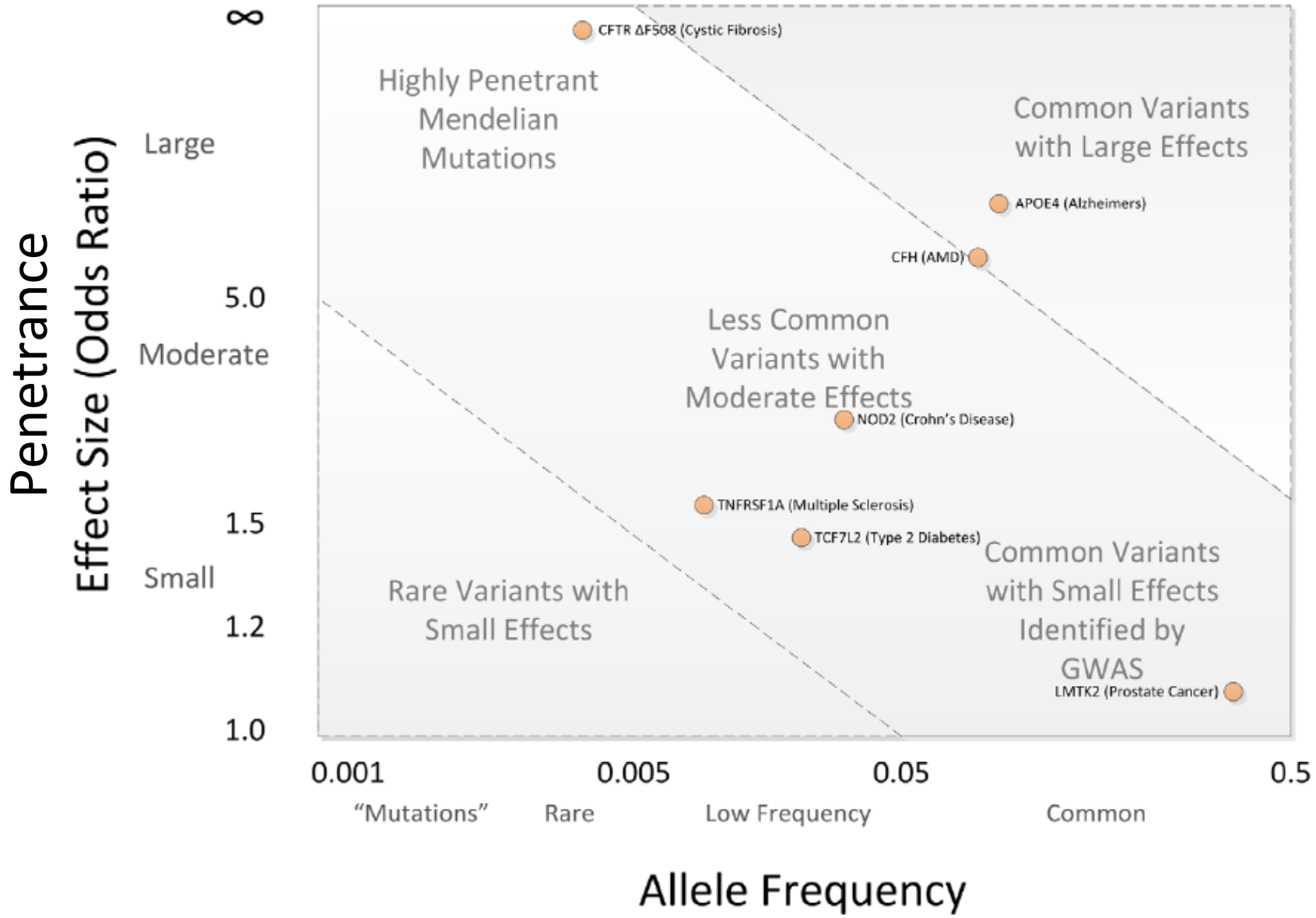
# Genome Wide Association Studies

# What are we trying to do?

- Correlate changes in genome with phenotype of interest.

# GWAS!


Penetrance
Effect Size (Odds Ratio)
∞
Large
5.0
Moderate
1.5
Small
1.2
1.0
CFTR ΔF508 (Cystic Fibrosis)
Highly Penetrant Mendelian Mutations
Common Variants with Large Effects
APOE4 (Alzheimers)
CFH (AMD)
Less Common Variants with Moderate Effects
NOD2 (Crohn's Disease)
TNFRSF1A (Multiple Sclerosis)
TCF7L2 (Type 2 Diabetes)
Common Variants with Small Effects Identified by GWAS
LMTK2 (Prostate Cancer)
Rare Variants with Small Effects
0.001
0.005
0.05
0.5
"Mutations"
Rare
Low Frequency
Common
Allele Frequency

# How tell if a variant is correlated to a phenotype?

- Correlation
- Probability
- Odds ratio/Relative Risk

# Different kinds of Phenotypes

- Categorical
  - Have diabetes or don't
  - Have phenylketonuria or don't
- Quantitative
  - Gradient of disease or biomarker of interest
    - Age of onset of Alzheimer's
    - Blood Pressure
    - HDL concentrations

# Covariance and Correlation

Want:

- Degree of association between 2 variables, X &Y
- Given observations $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$
  - Covariance (sum of O-E divided by number -1)

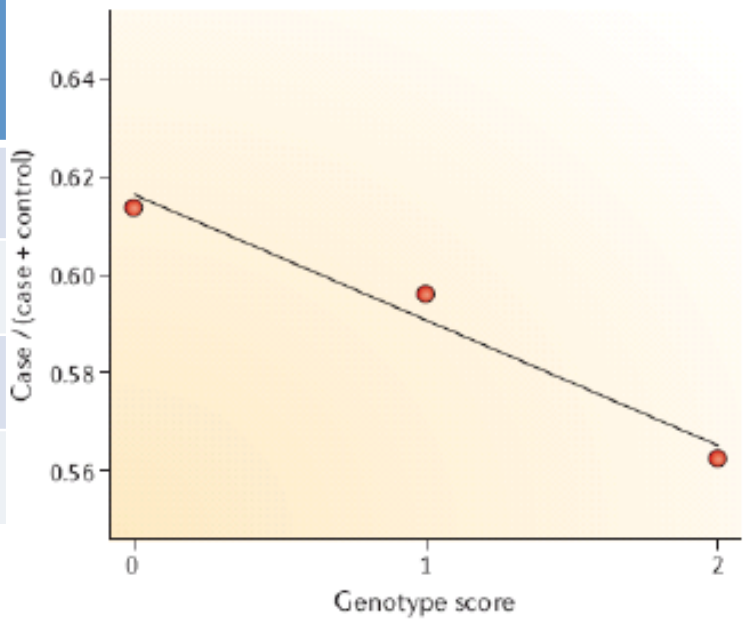
$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

  - Correlation coefficient

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

# Case/Control Study (Categorical)

| Genotype | Case | Control | Total |
|---|---|---|---|
| AA | $N_{case,AA}$ | $N_{control,AA}$ | $N_{AA}$ |
| Aa | $N_{case,Aa}$ | $N_{control,Aa}$ | $N_{Aa}$ |
| aa | $N_{case,aa}$ | $N_{control,aa}$ | $N_{aa}$ |
| Total | $N_{case}$ | $N_{control}$ | N |

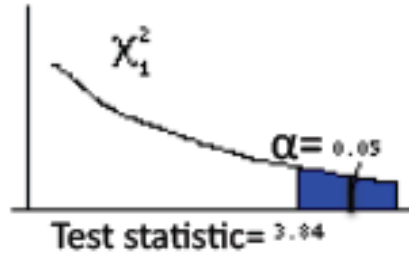



$X^2$ test with 1 or 2 df (1 if single allele, 2 if genotype)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

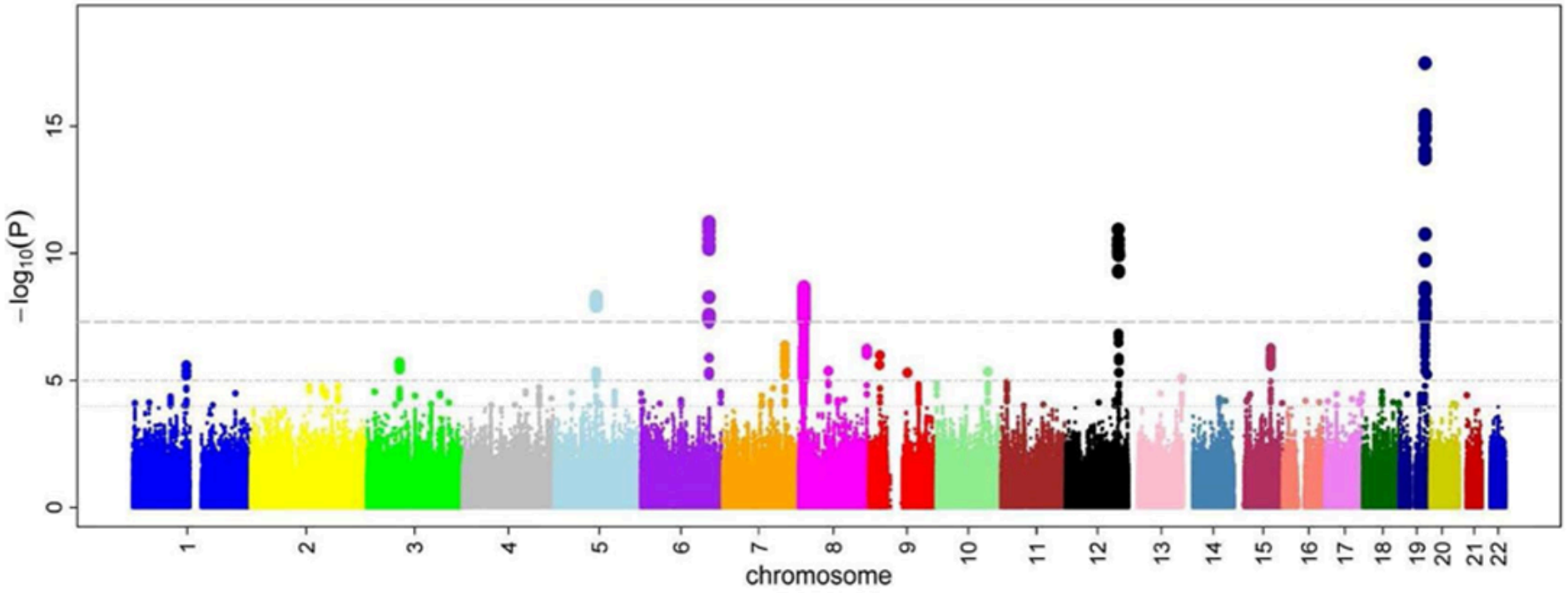
$O_i$ = observed frequency for $i^{th}$ outcome
(the value can be read off of the contingency table)
$E_i$ = expected frequency for $i^{th}$ outcome
(the value can be obtained as described in the previous slides)
$n$ = total number of outcomes

The probability distribution of this statistic is given by the chi-square distribution.




Using chi-square test, we can test how well observed values fit expected values computed under the independence hypothesis

# Manhattan Plot




GWAS for microcirculation

M. Kamran Ikram et al - Ikram MK et al (2010) Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo. PLoS Genet. 2010 Oct 28;6(10):e1001184. doi:10.1371/journal.pgen.1001184.g001

# Let’s do a GWAS

| Allele | Short Hair | Long Hair | Totals | Expected |
|---|---|---|---|---|
| T | 3 | 12 | 15 | 7.5 |
| G | 9 | 0 | 9 | 4.5 |
| Totals | 12 | 12 | 24 | |

| Allele | Short Hair | Long Hair | Expected |
|---|---|---|---|
| T | 3 | 12 | 7.5 |
| G | 9 | 0 | 4.5 |

$$\chi^2 = \sum_{i=o}^{\infty} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \left(\frac{(3-7.5)^2}{7.5}\right) + \left(\frac{(9-4.5)^2}{4.5}\right) + \left(\frac{(12-7.5)^2}{7.5}\right) + \left(\frac{(0-4.5)^2}{4.5}\right)$$

$$\chi^2 = 14.4$$

# Significance

| | P | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **DF** | **0.995** | **0.975** | **0.20** | **0.10** | **0.05** | **0.025** | **0.02** | **0.01** | **0.005** | **0.002** | **0.001** |
| **1** | 0.0000393 | 0.000982 | 1.642 | 2.706 | 3.841 | 5.024 | 5.412 | 6.635 | 7.879 | 9.550 | 10.828 |
| **2** | 0.0100 | 0.0506 | 3.219 | 4.605 | 5.991 | 7.378 | 7.824 | 9.210 | 10.597 | 12.429 | 13.816 |
| **3** | 0.0717 | 0.216 | 4.642 | 6.251 | 7.815 | 9.348 | 9.837 | 11.345 | 12.838 | 14.796 | 16.266 |
| **4** | 0.207 | 0.484 | 5.989 | 7.779 | 9.488 | 11.143 | 11.668 | 13.277 | 14.860 | 16.924 | 18.467 |
| **5** | 0.412 | 0.831 | 7.289 | 9.236 | 11.070 | 12.833 | 13.388 | 15.086 | 16.750 | 18.907 | 20.515 |

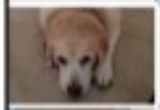
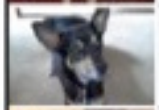
By alleles – 1 df

$X^2 = 14.4$

Therefore $p<0.001$

# What we actually care about

- How much does this allele/genotype increase your risk of a disease/phenotype
- Odds ratio (OR) or Risk ratio (RR)
  - Risk ratio depends on penetrance

| Dog | Coat Length | chr32 7420804 | chr32 7472206 | chr32 7473337 | chr32 7479580 | chr32 7482967 | chr32 7490570 | chr32 7492364 |
|---|---|---|---|---|---|---|---|---|
|  | Short Coat | TC | AA | GT | TT | AG | TT | CG |
|  | Short Coat | TC | AA | GG | TT | GG | CC | GG |
|  | Short Coat | CC | GA | GT | CT | AG | CT | |
|  | Short Coat | TC | GA | GG | CT | AA | TT | CG |
|  | Short Coat | CC | GA | GT | CT | AG | CT | GG |
|  | Short Coat | TC | AA | GG | TT | GG | CC | CG |
|  | Long Coat | CC | AA | TT | TT | GG | TT | GG |
|  | Long Coat | TC | AA | TT | TT | GG | TT | GG |
|  | Long Coat | TC | AA | TT | TT | GG | TT | GG |
|  | Long Coat | CC | AA | TT | TT | GG | CT | CG |
|  | Long Coat | TT | AA | TT | TT | GG | TT | GG |
|  | Long Coat | CC | AA | TT | TT | GG | TT | GG |

**Chr32 7490570**

| Alelle | Short | Long | Totals | Expected |
|---|---|---|---|---|
| T | 6 | 11 | 17 | 8.5 |
| C | 6 | 1 | 7 | 3.5 |
| Totals | 12 | 12 | 24 | |

$$\chi^2 = \sum_{i=o}^{\infty} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \left(\frac{(6-8.5)^2}{8.5}\right) + \left(\frac{(11-8.5)^2}{8.5}\right) + \left(\frac{(6-3.5)^2}{3.5}\right) + \left(\frac{(1-3.5)^2}{3.5}\right)$$

$$\chi^2 = 5.0$$

Still significant 0.05>p>0.025

| Alelle | Short | Long | Totals |
|---|---|---|---|
| T | 6 | 11 | 17 |
| C | 6 | 1 | 7 |
| Totals | 12 | 12 | 24 |

$$P(Short \mid T) = \frac{P(Short \,\&\, T)}{P(T)} = \frac{6}{17} = 0.35$$

$$P(Short \mid C) = \frac{P(Short \,\&\, C)}{P(C)} = \frac{6}{7} = 0.85$$

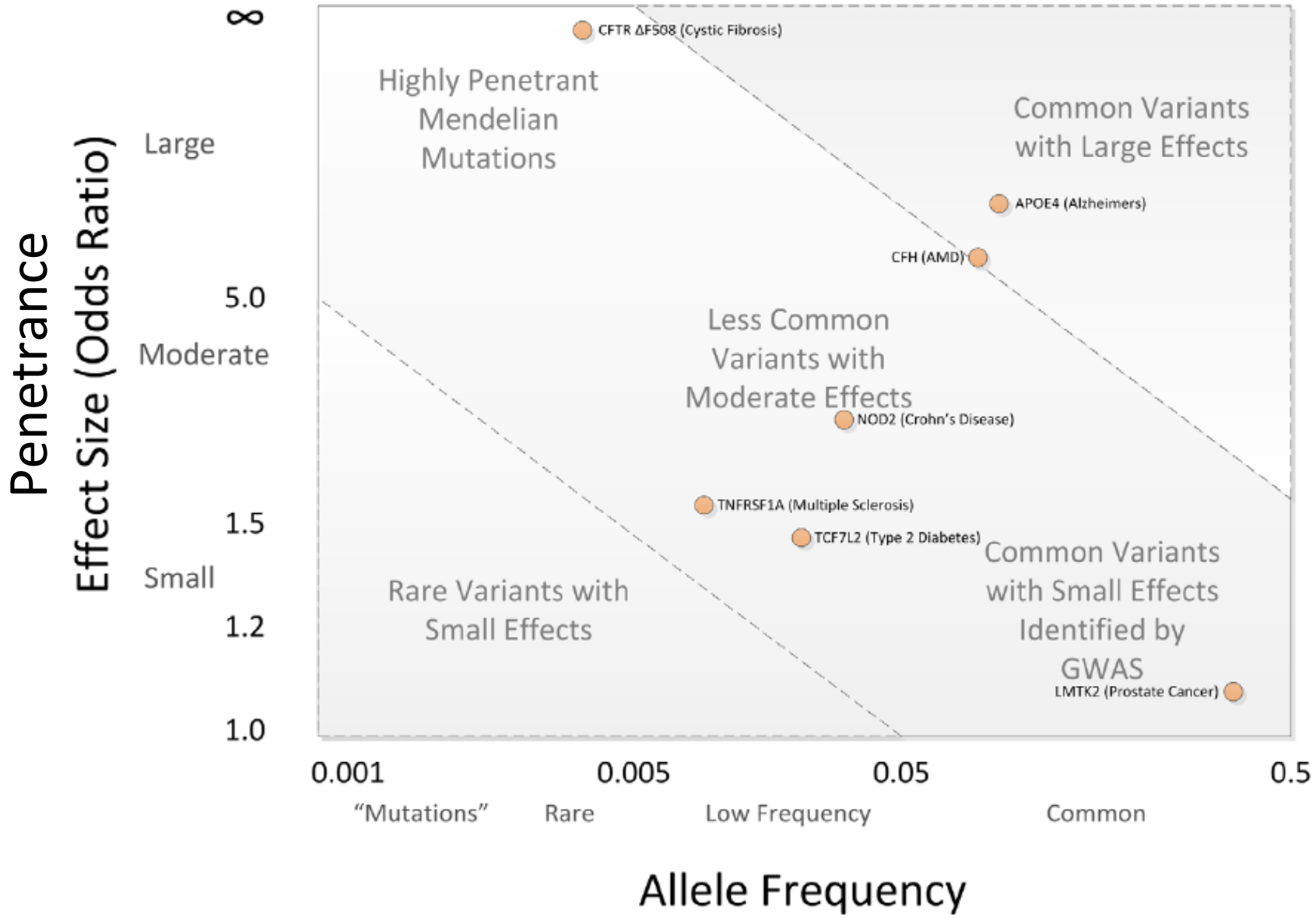
$$Odds(T) = \frac{P(Short \mid T)}{1 - P(Short \mid T)} = 0.538$$

$$Odds(C) = \frac{P(Short \mid C)}{1 - P(Short \mid C)} = 5.66$$

$$OR = \frac{Odds(C)}{Odds(T)} = 10.5$$

# Quantitative Phenotypes

- Can't use $\chi^2$
- Need to use a regression on the data
  - Linear is a good first step
- Lots of ongoing research on how to improve this


Penetrance
Effect Size (Odds Ratio)
∞
Large
5.0
Moderate
1.5
Small
1.2
1.0
CFTR ΔF508 (Cystic Fibrosis)
Highly Penetrant Mendelian Mutations
Common Variants with Large Effects
APOE4 (Alzheimers)
CFH (AMD)
Less Common Variants with Moderate Effects
NOD2 (Crohn's Disease)
TNFRSF1A (Multiple Sclerosis)
TCF7L2 (Type 2 Diabetes)
Common Variants with Small Effects Identified by GWAS
LMTK2 (Prostate Cancer)
Rare Variants with Small Effects
0.001
0.005
0.05
0.5
"Mutations"
Rare
Low Frequency
Common
Allele Frequency

# Why do GWAS usually find common variants with moderate to large effects?

- Need large numbers of individuals to find rare variants (statistical power)
- Need moderate effect to reach statistical significance
- If very high effect, can still find, like Chr32 7473337

# Some problems with GWAS

- Clinical phenotypes aren't always good enough
  - May be multiple pathophysiologies that have different causes
  - Clinical categories may be too broad to distinguish
- SNPs sometimes not clear how they relate
- Often find nothing
  - Moving toward pathway analysis instead of gene by gene

# Questions?