Quantitative Understanding in Biology

Basics of genome-wide association study (GWAS) analysis

Jason Mezey
jgm45@cornell.edu
March 7, 2013 (Th) 5:30-7PM

Goals for today

- Motivation: example of a successful GWAS and why we should care
- Structure and statistics of a GWAS analysis
- GWAS analysis issues
- History and future of GWAS

A successful GWAS...

Macular degeneration

From Wikipedia, the free encyclopedia

Age-related macular degeneration (AMD) is a medical condition which usually affects older adults and results in a loss of vision in the center of the visual field (the macula) because of damage to the retina. It occurs in "dry" and "wet" forms. It is a major cause of blindness and visual impairment in older adults (>50 years). Macular degeneration can make it difficult or impossible to read or recognize faces, although enough peripheral vision remains to allow other activities of daily life.

Starting from the inside of the eye and going towards the back, the three main layers at the back of the eye are the retina, which contains the nerves; the choroid, which contains the blood supply; and the sclera, which is the white of the eye.

The macula is the central area of the retina, which provides the most detailed central vision.

In the dry (nonexudative) form, cellular debris called drusen accumulates between the retina and the choroid, and the retina can become detached. In the wet (exudative) form, which is more severe, blood vessels grow up from the choroid behind the retina, and the retina can also become detached. It can be treated with laser coagulation, and with medication that stops and sometimes reverses the growth of blood vessels.[1][2]

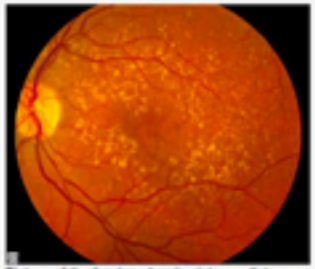
Although some macular dystrophies affecting younger individuals are sometimes referred to as macular degeneration, the term generally refers to age-related macular degeneration (AMD or ARMD).

Age-related macular degeneration begins with characteristic yellow deposits (drusen) in the macula, between the retinal pigment epithelium and the underlying choroid. Most people with these early changes (referred to as age-related maculopathy) have good vision. People with drusen can go on to develop advanced AMD. The risk is higher when the drusen are large and numerous and associated with disturbance in the pigmented cell layer under the macula. Large and soft drusen are related to elevated cholesterol deposits and may respond to cholesterol-lowering agents.

Contents [hide]

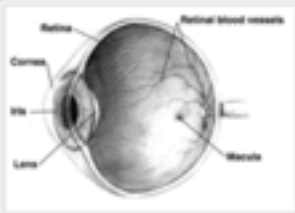
Macular degeneration

Classification and external resources



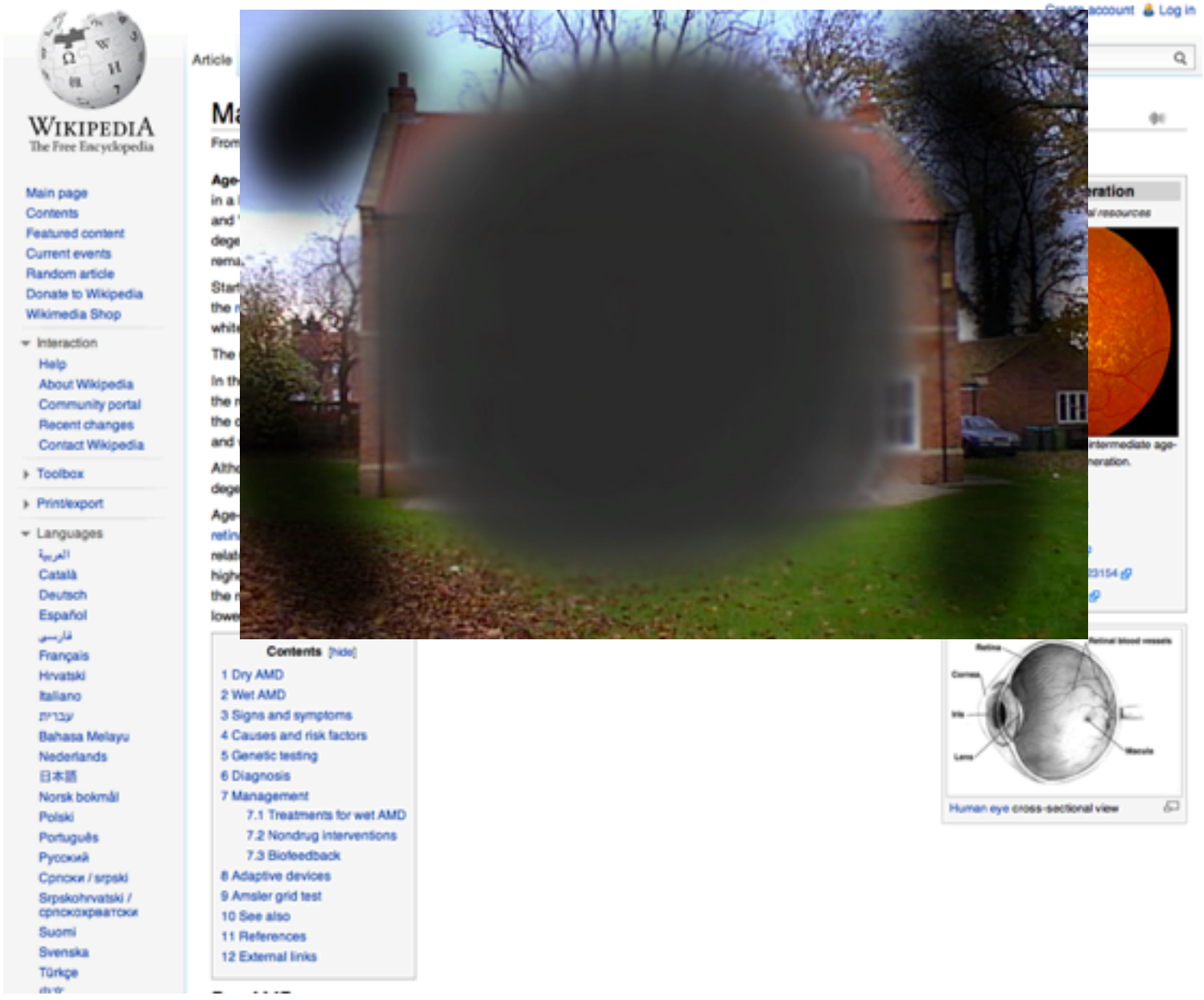
Picture of the fundus showing intermediate age-related macular degeneration.

ICD-10	H35.3
ICD-9	362.50
DiseasesDB	11948
MedlinePlus	001000
eMedicine	article/1223154
MeSH	D008268

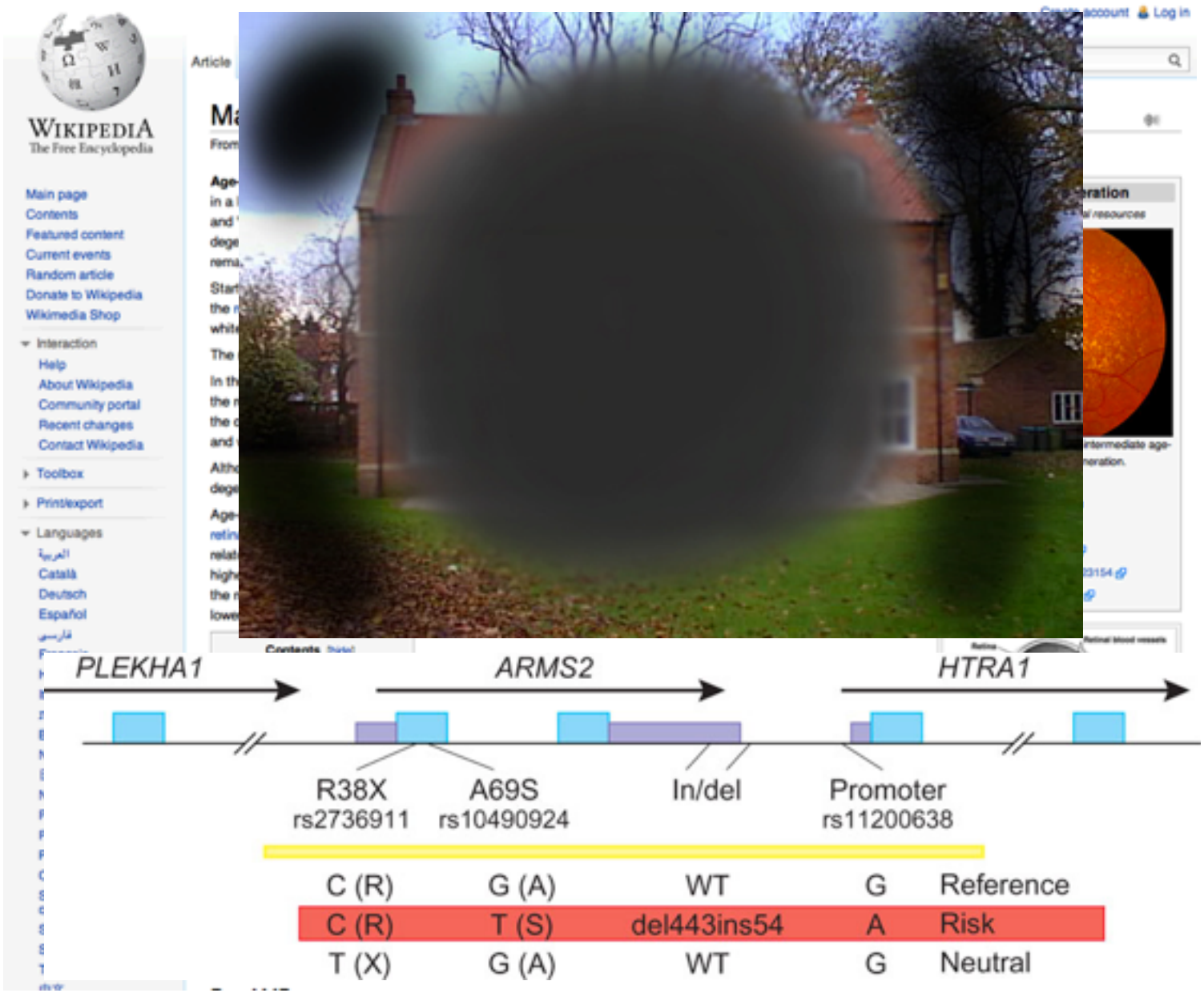


Human eye cross-sectional view

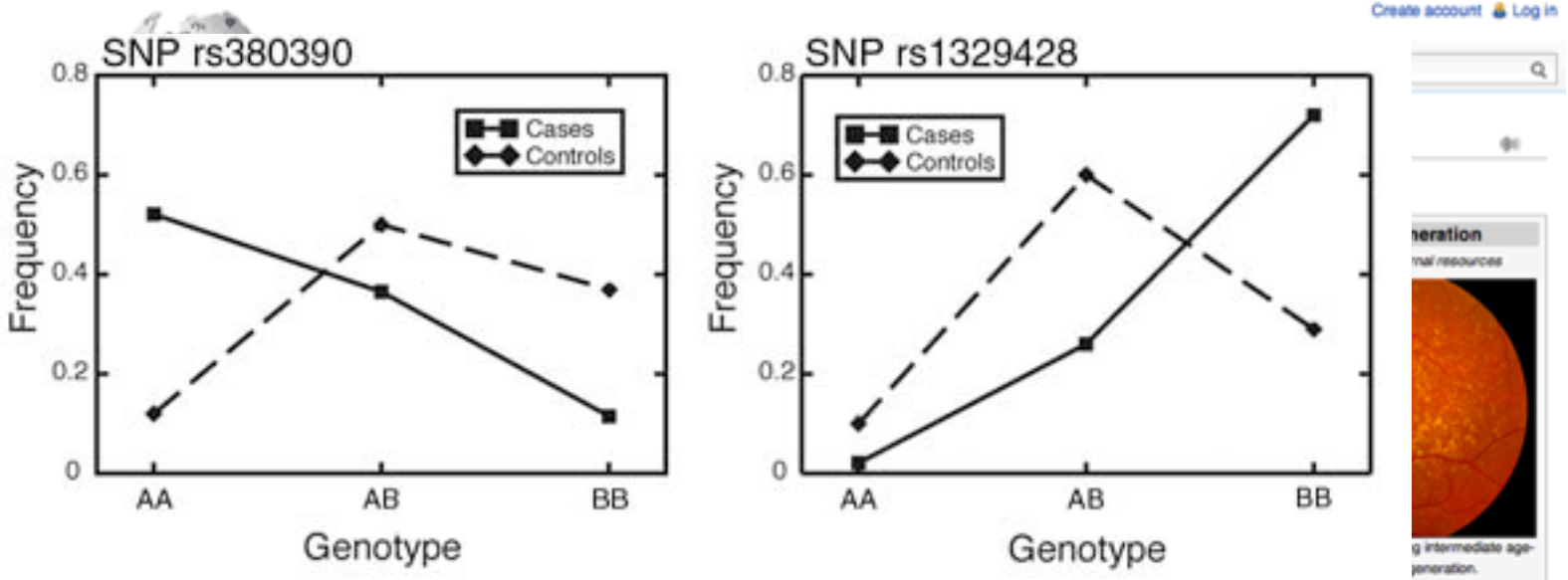
A successful GWAS...

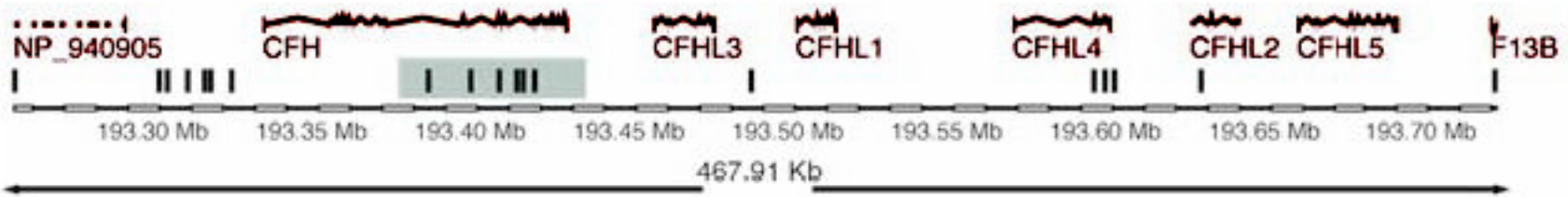


A successful GWAS...

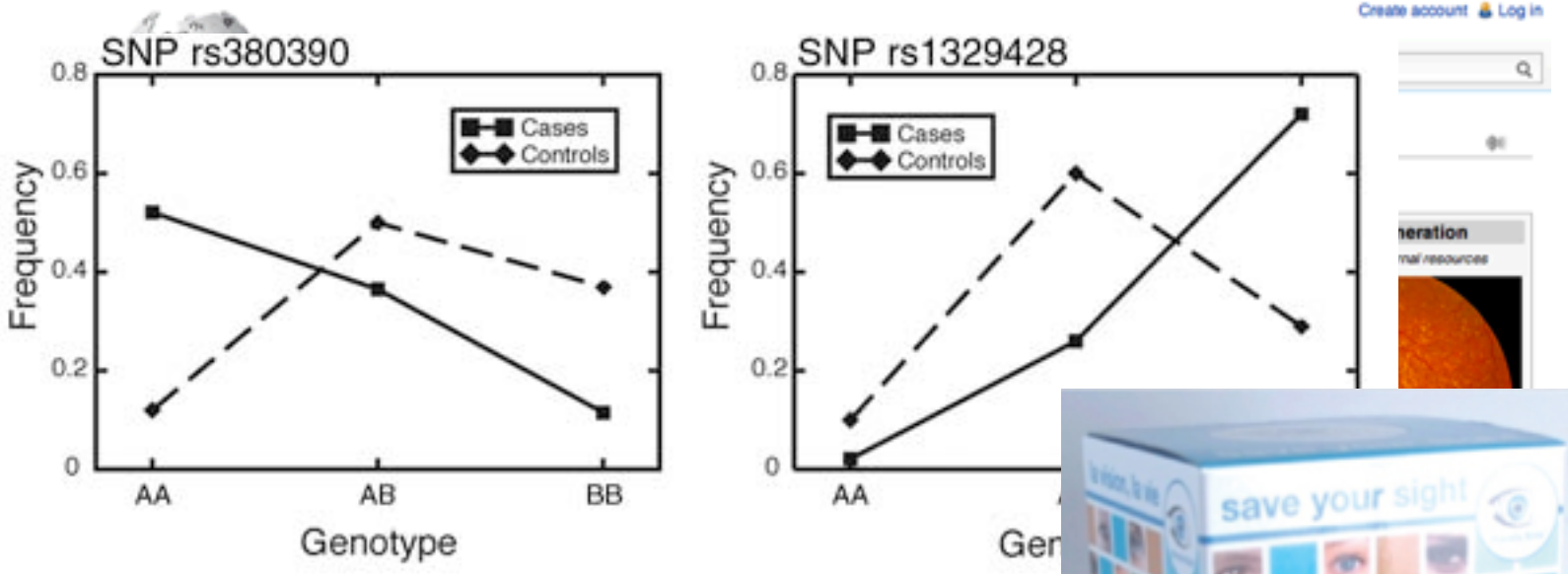


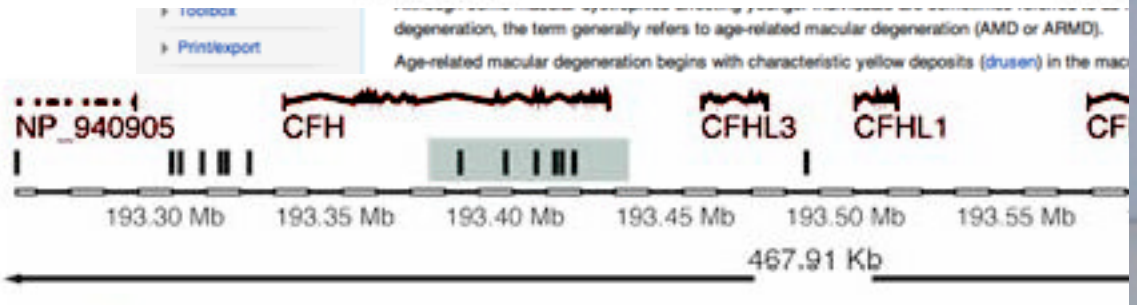
A successful GWAS...



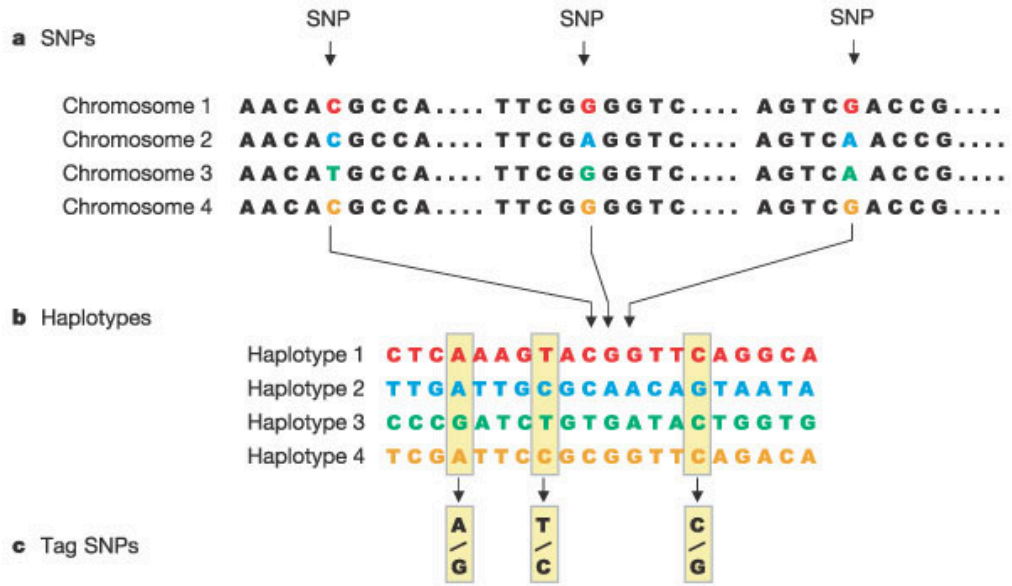


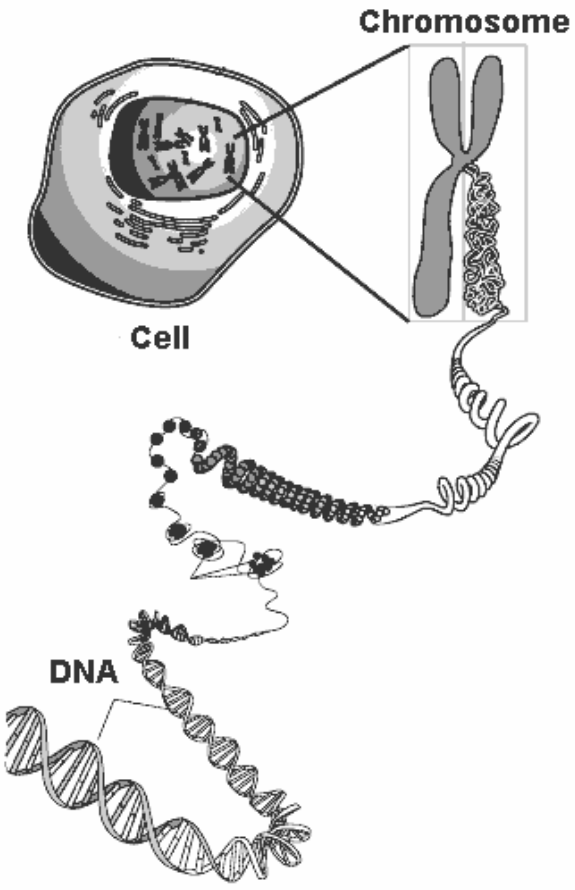
A successful GWAS...





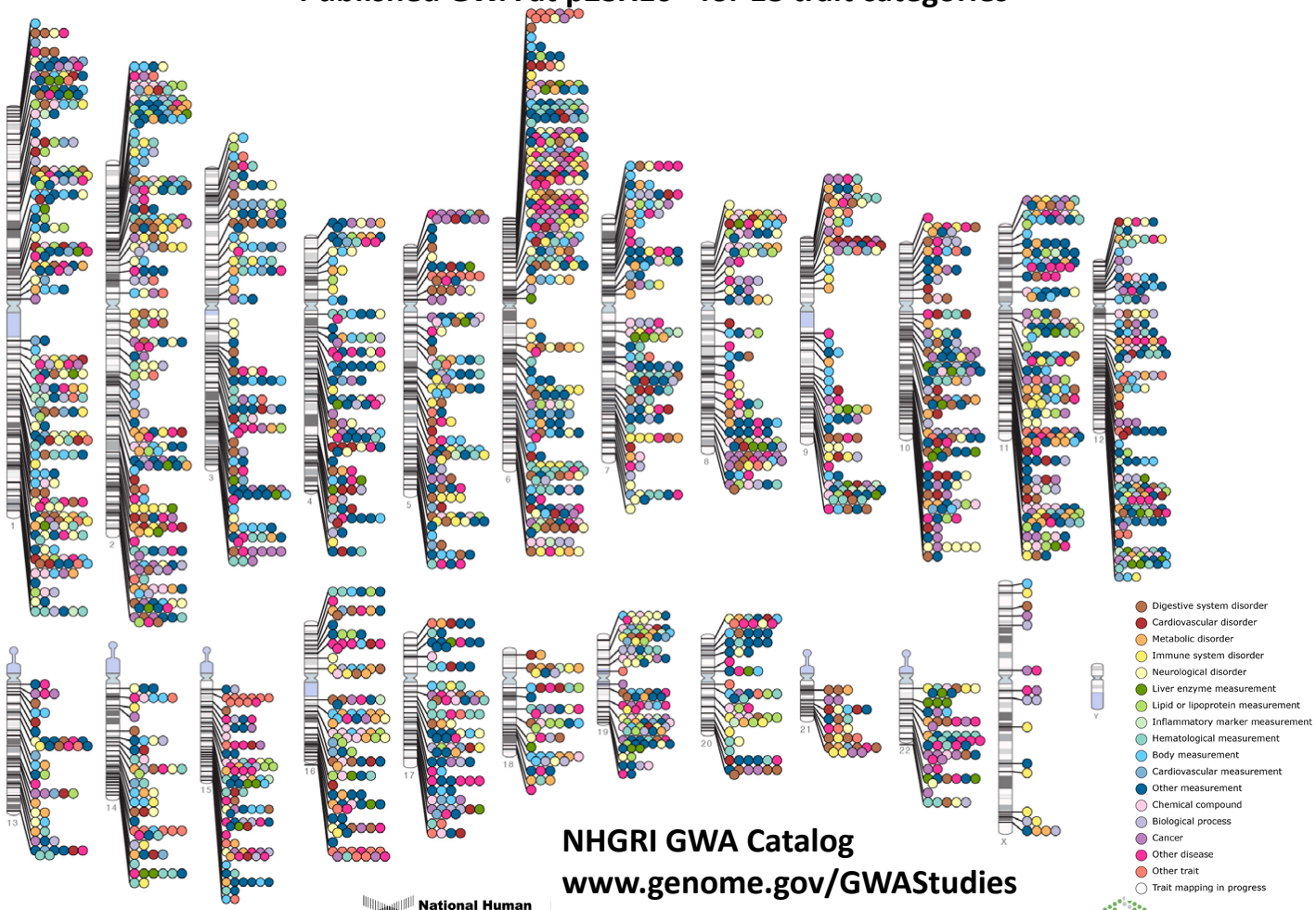
What do we know about disease / quantitative loci?





What do we know about disease / quantitative loci?

Published Genome-Wide Associations through 07/2012
Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories



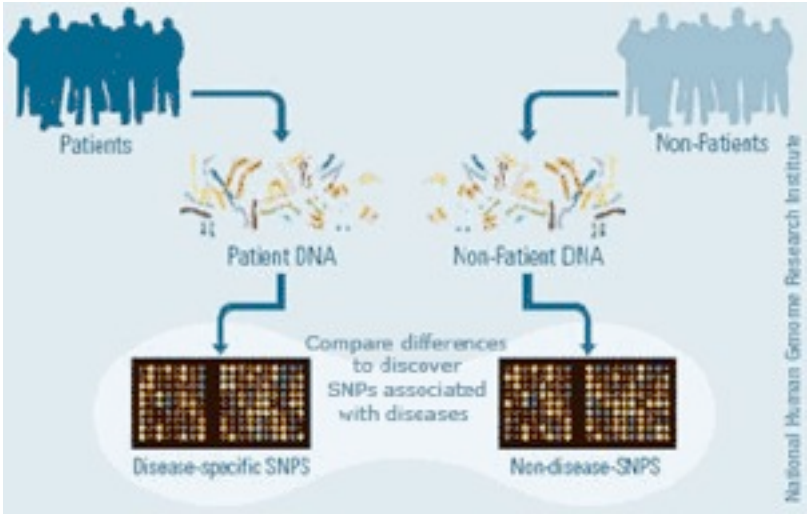
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

GWAS structure

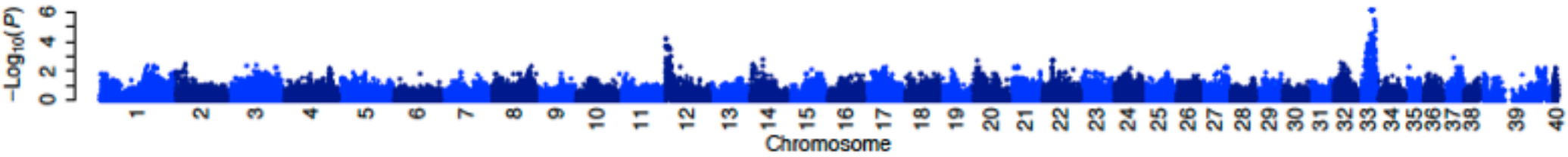
- Genome-wide association studies (GWAS) are used to map the genomic location of disease loci:



- Associations can be identified by assessing the correlation of each genetic marker independently for an association with phenotype:

$$H_0 : Cov(Y, X) = 0$$

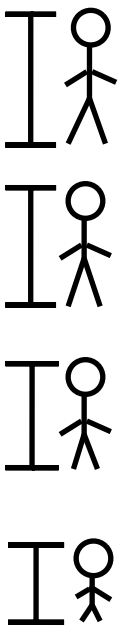
$$H_A : Cov(Y, X) \neq 0$$



Manhattan plot of individual marker analysis p-values.

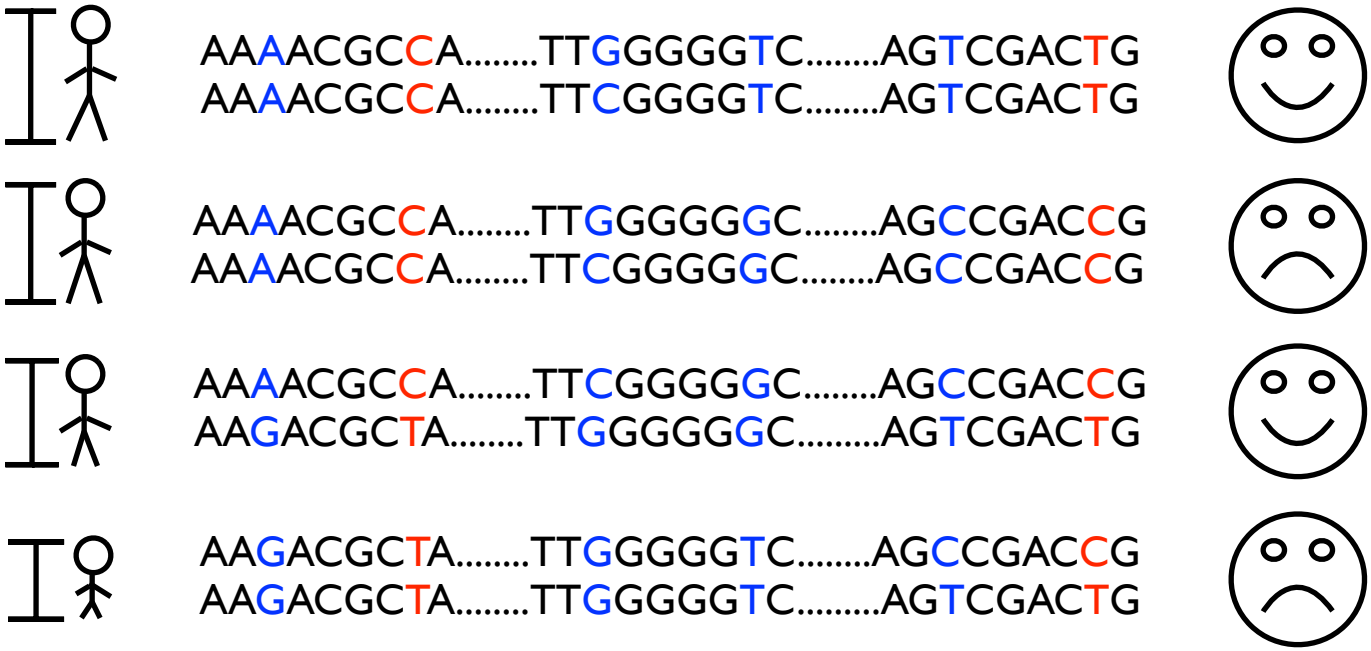
Association analysis intuition

Association analysis intuition

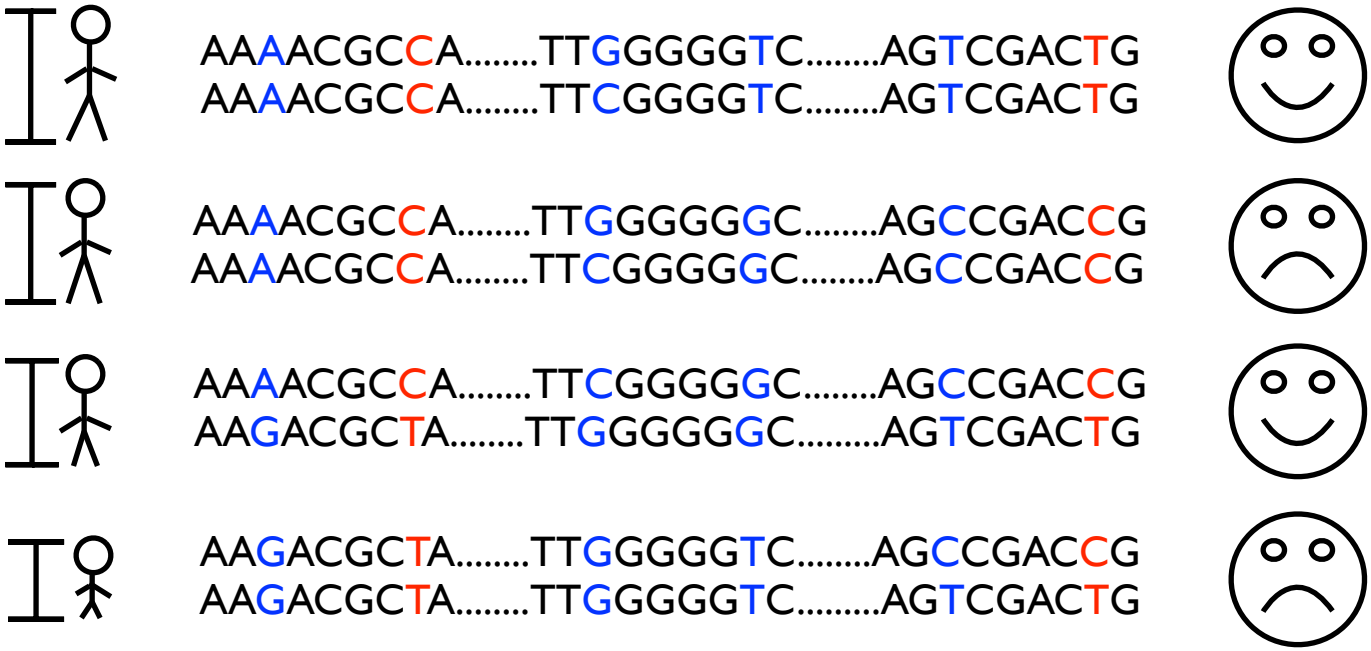


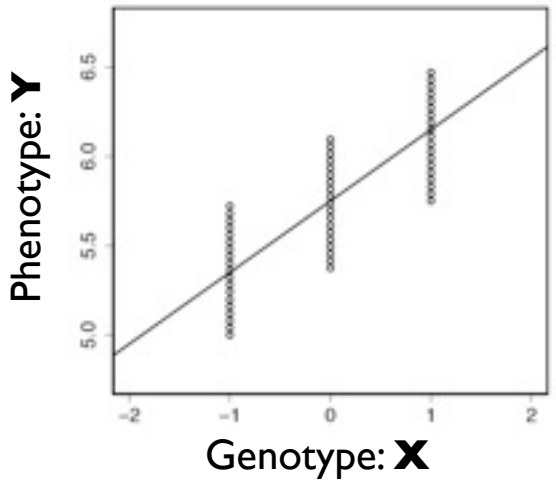
Association analysis intuition

Association analysis intuition

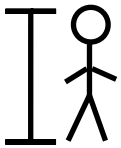


Association analysis intuition

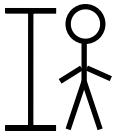




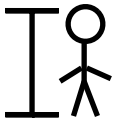
Association analysis intuition



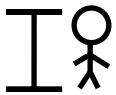
AAAACGCCA.....TTGGGGGTC.....AGTCGACTG
AAAACGCCA.....TTCGGGGTC.....AGTCGACTG



AAAACGCCA.....TTGGGGGGC.....AGCCGACCG
AAAACGCCA.....TTCGGGGGC.....AGCCGACCG

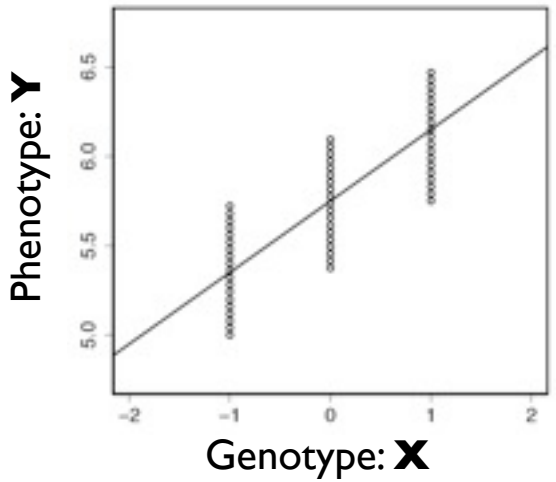


AAAACGCCA.....TTCGGGGGC.....AGCCGACCG
AAGACGCTA.....TTGGGGGGC.....AGTCGACTG

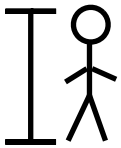


AAGACGCTA.....TTGGGGGTC.....AGCCGACCG
AAGACGCTA.....TTGGGGGTC.....AGTCGACTG

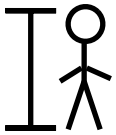
$$\mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \epsilon$$



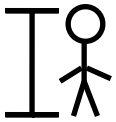
Association analysis intuition



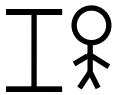
AAAACGCCA.....TTGGGGGTC.....AGTCGACTG
AAAACGCCA.....TTCGGGGTC.....AGTCGACTG



AAAACGCCA.....TTGGGGGGC.....AGCCGACCG
AAAACGCCA.....TTCGGGGGC.....AGCCGACCG

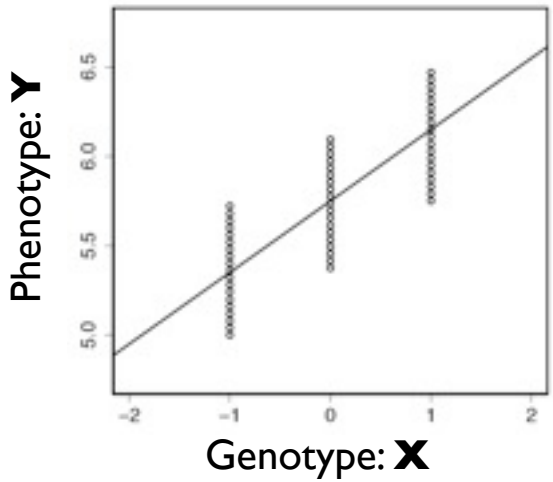


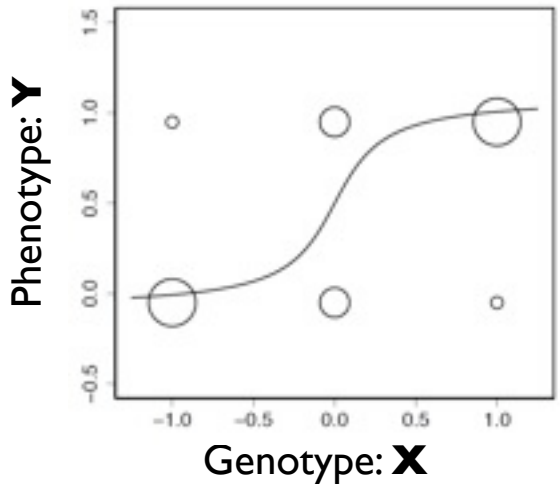
AAAACGCCA.....TTCGGGGGC.....AGCCGACCG
AAGACGCTA.....TTGGGGGGC.....AGTCGACTG



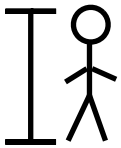
AAGACGCTA.....TTGGGGGTC.....AGCCGACCG
AAGACGCTA.....TTGGGGGTC.....AGTCGACTG

$$\mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \epsilon$$

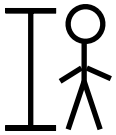




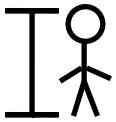
Association analysis intuition



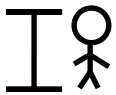
AAAACGCCA.....TTGGGGGTC.....AGTCGACTG
AAAACGCCA.....TTCGGGGTC.....AGTCGACTG



AAAACGCCA.....TTGGGGGGC.....AGCCGACCG
AAAACGCCA.....TTCGGGGGC.....AGCCGACCG

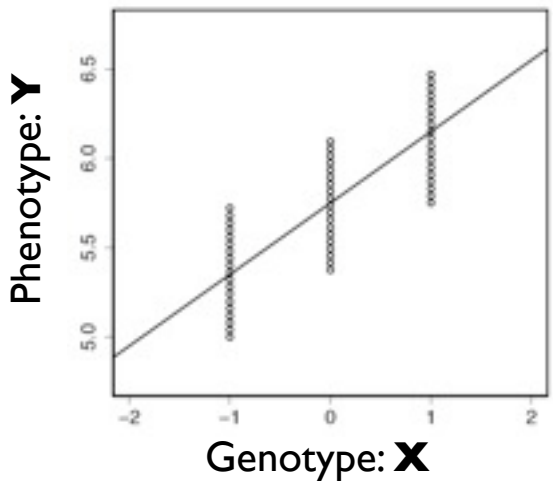


AAAACGCCA.....TTCGGGGGC.....AGCCGACCG
AAGACGCTA.....TTGGGGGGC.....AGTCGACTG

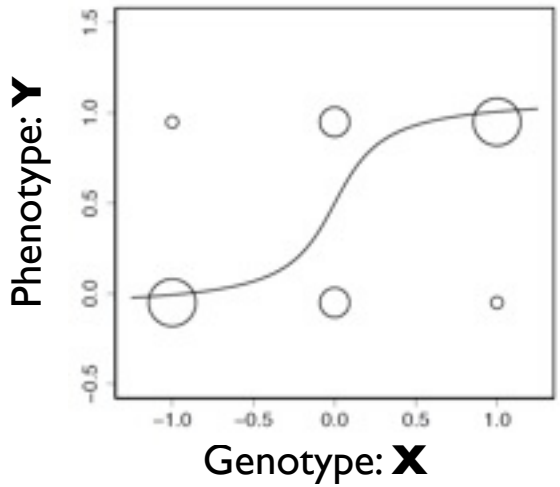


AAGACGCTA.....TTGGGGGTC.....AGCCGACCG
AAGACGCTA.....TTGGGGGTC.....AGTCGACTG

$$\mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \epsilon$$

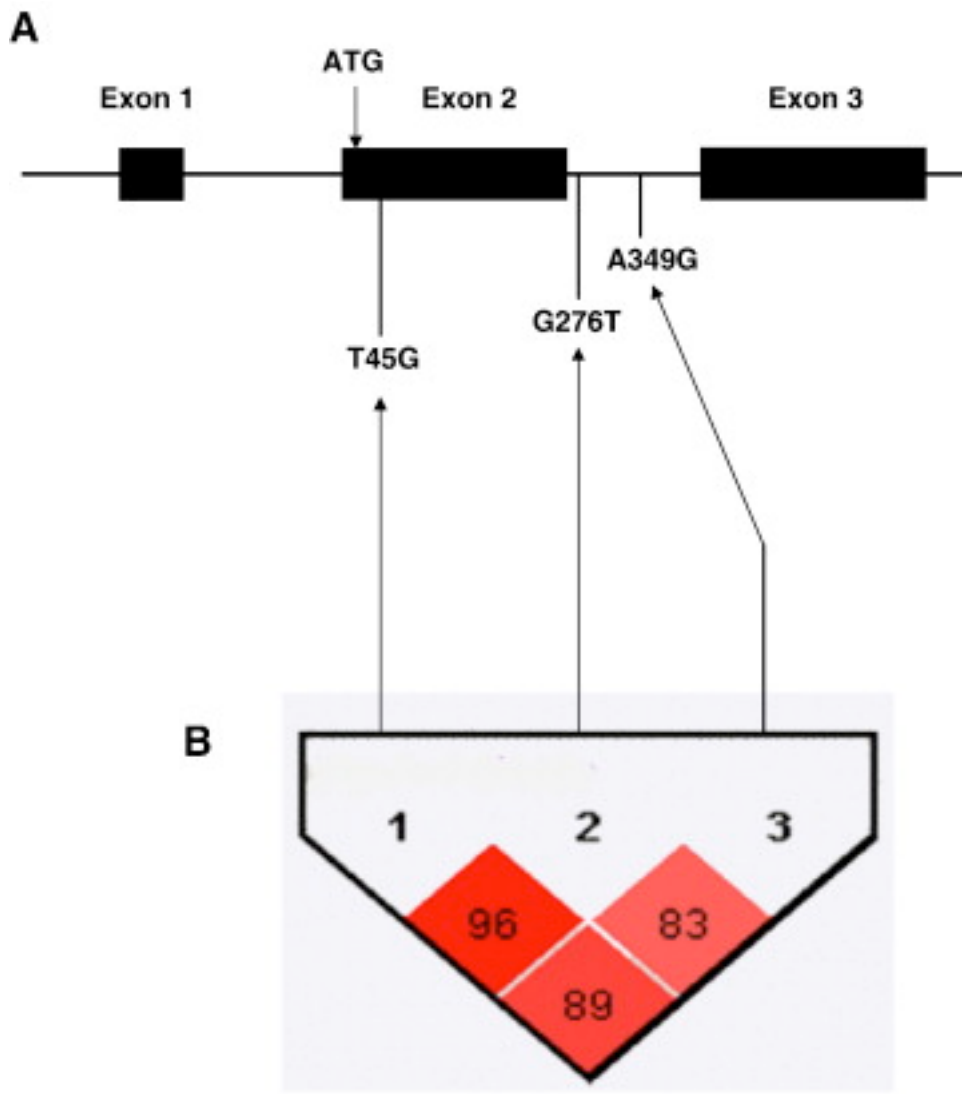


$$\mathbf{Y} = \gamma^{-1}(\beta_0 + \mathbf{X}_1\beta_1)$$



Note that association is really an association!!

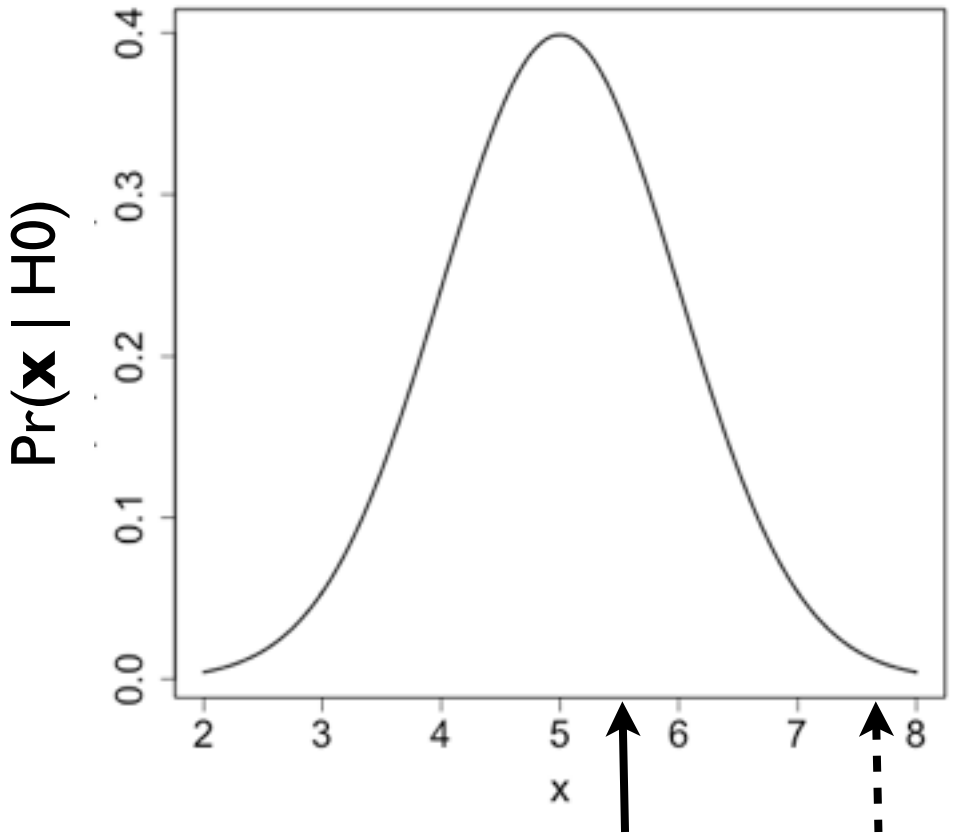
- If we test a (non-causal) genotype that is correlated with the causal genotype AND if correlated genotypes are in the same position in the genome THEN we can identify the genomic position of the casual genotype (!!) = association
- This is the case in genetic systems (why!?)
- Do we know which genotype is causal in this scenario?



Core of association: p-value

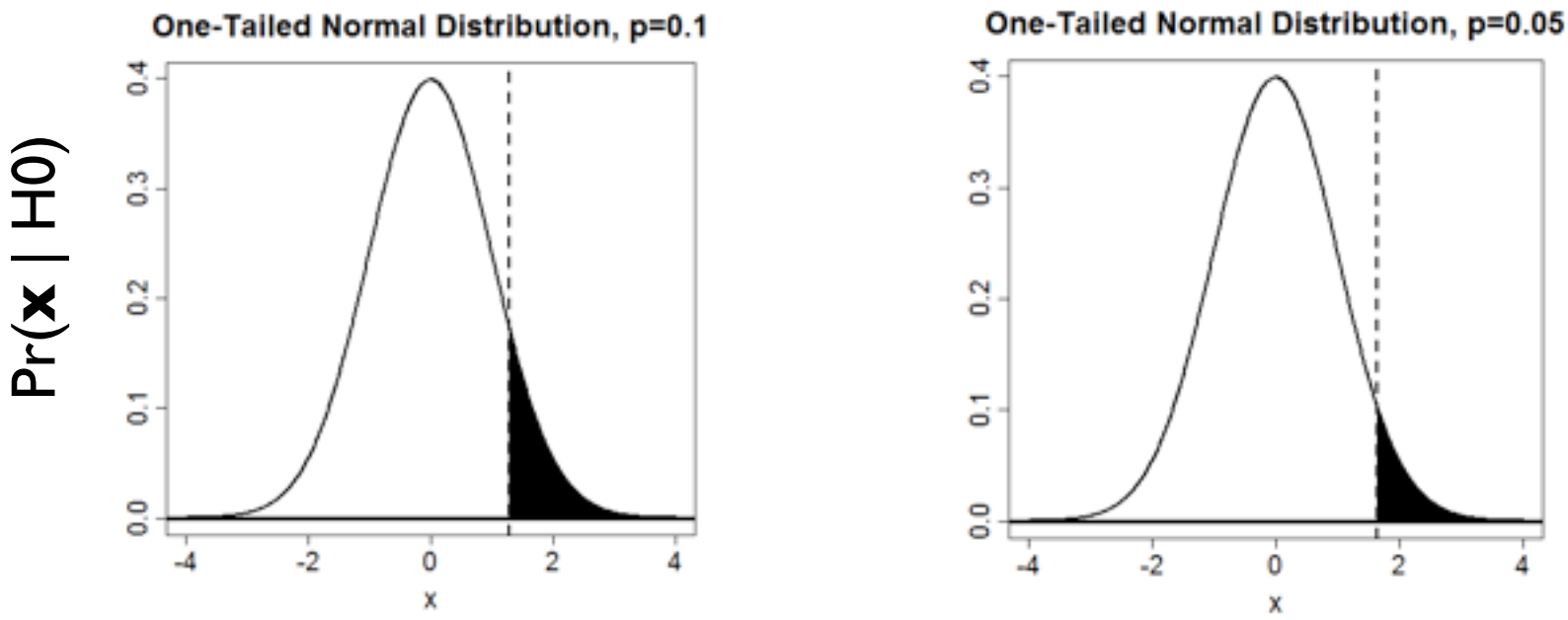
- Assume that you want to know if an underlying probably model is not a correct description of your system (a hypothesis! that we will call H0), e.g. in an association our H0 could be the regression slope of phenotype on genotype is zero

- In general you measure a value "x" generated by your system - we can assess our hypothesis H0 by considering the probability of observing "x" conditional on H0 bring correct (=true) - *note that this distribution need not be normal!!*



Intuition (!!): p-values

- **p-value** - the probability of obtaining a value of a statistic $T(\mathbf{x})$, *or more extreme*, conditional on H0 being true
- In our case, our statistic is "x" and if we assume a "one-tailed test" (we will get to this in a moment) our p-value could be:



- To really understand this, we need probability and statistics...

p-value 1

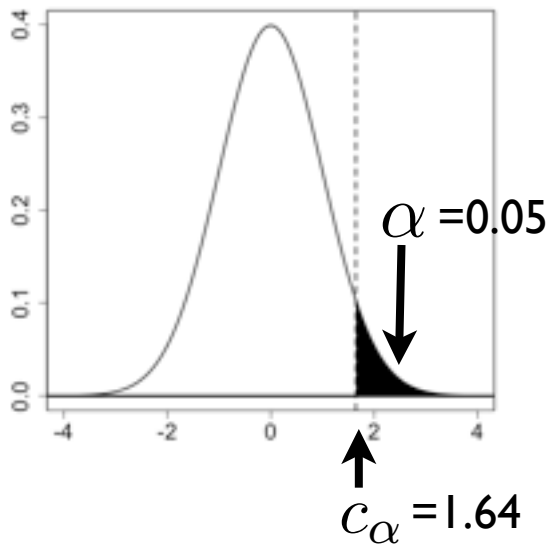
- We quantify our intuition as to whether we would have observed the value of our statistics given the null is true with a *p-value*
- **p-value** - the probability of obtaining a value of a statistic *T*(**x**), *or more extreme*, conditional on H0 being true
- Formally, we can express this as follows:

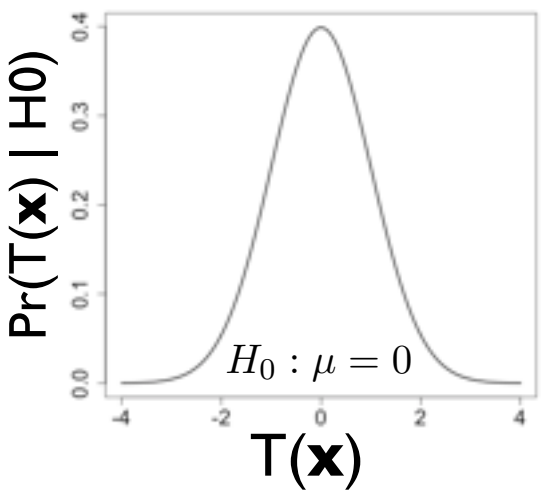
$$pval = Pr(|T(\mathbf{x})| \geqslant t | H_0 : \theta = c)$$

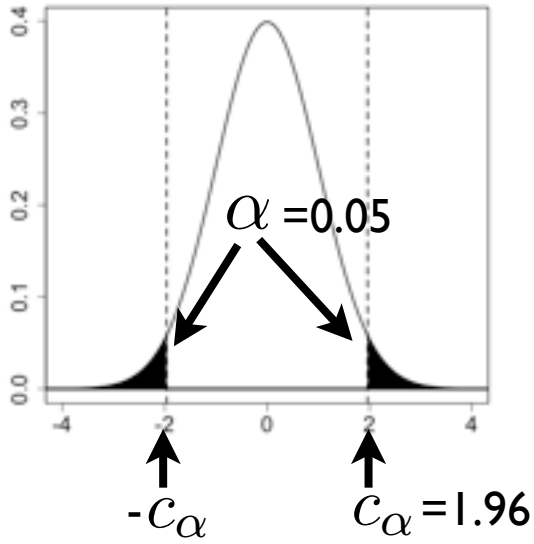
- Note that a p-value is a function on a statistic (!!) that takes the value of a statistic as input and produces a p-value as output in the range [0, 1]:

$$pval(T(x)) : T(x) \rightarrow [0, 1]$$

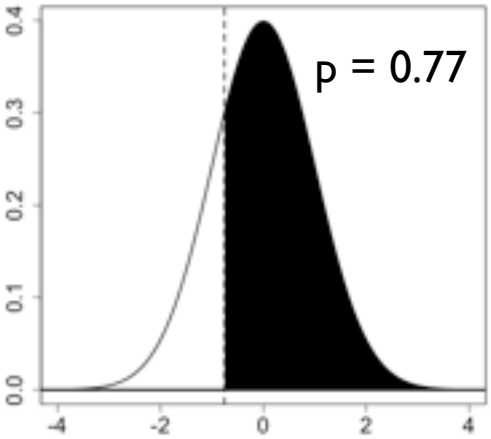
Assume H0 is correct (!): $\mu = 0$





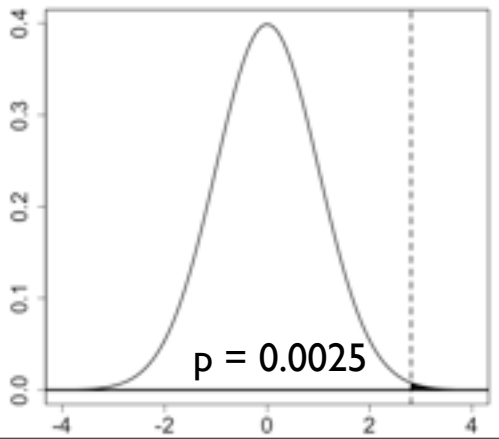


one-sided test



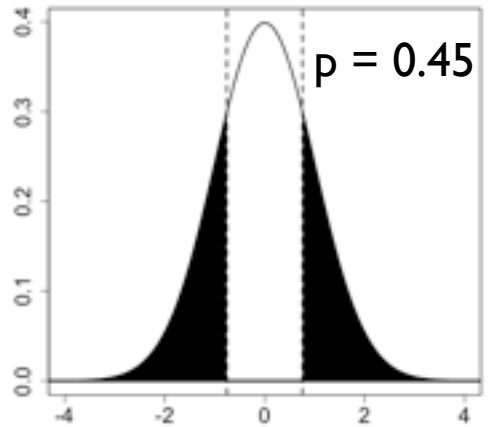
Sample I:

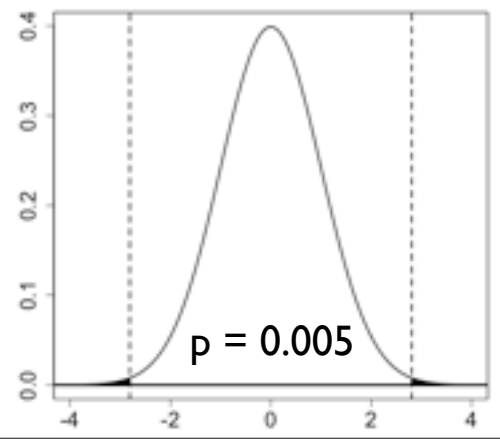
T(**x**)= -0.755



Sample II:

T(**x**)= 2.8



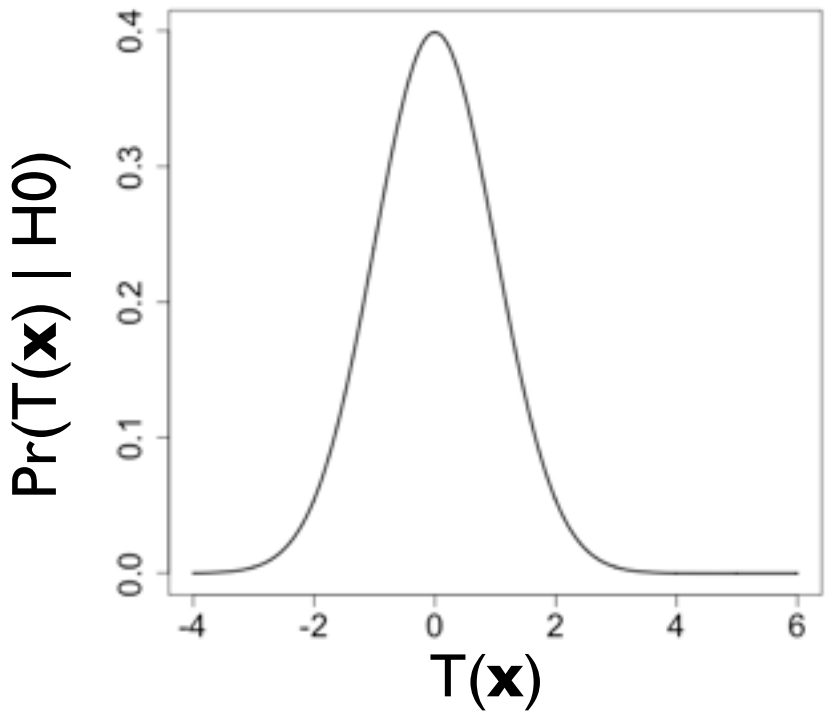


two-sided test

Results of hypothesis decisions I: when H0 is correct (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

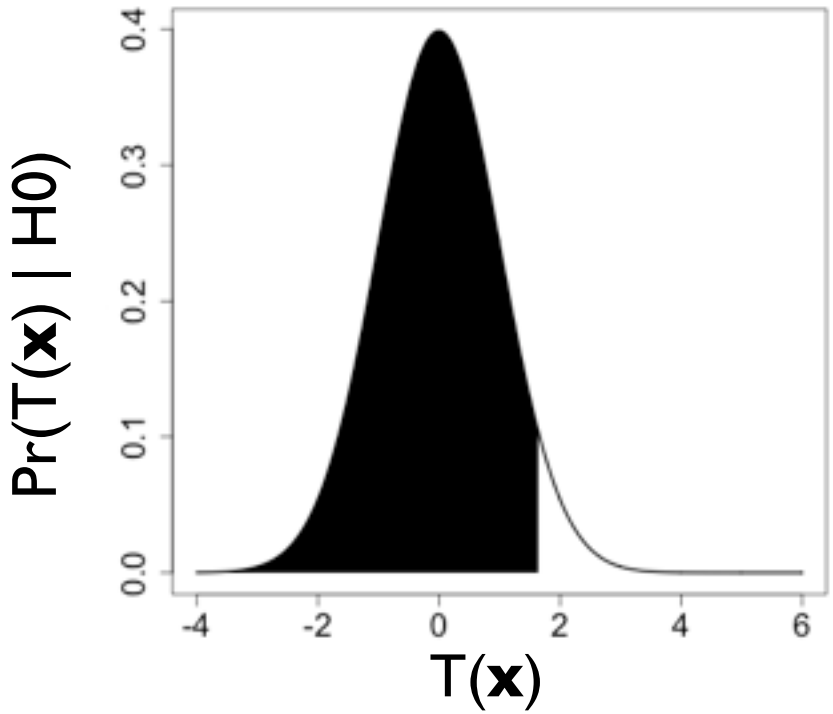
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



Results of hypothesis decisions I: when H0 is correct (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

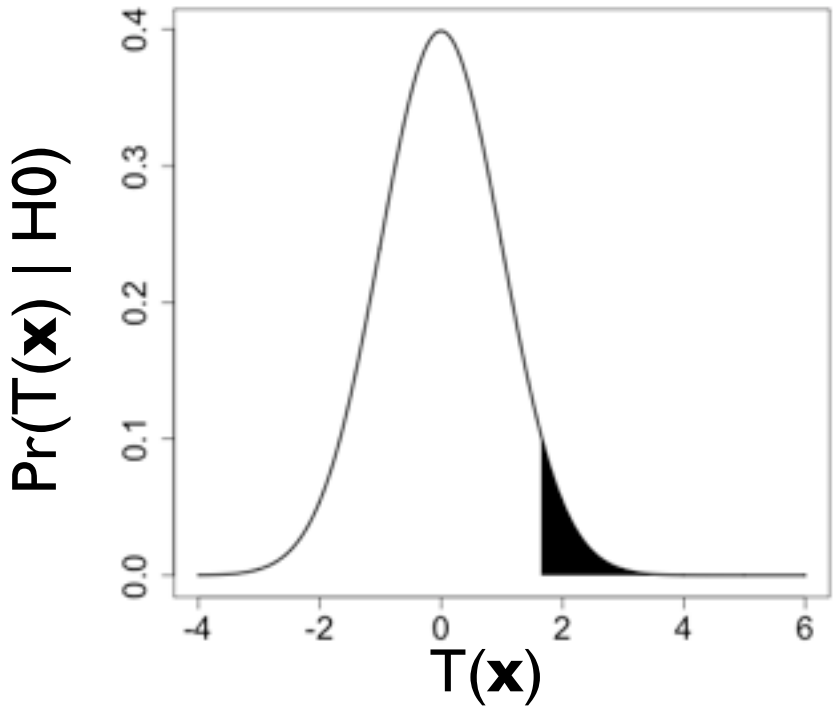
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



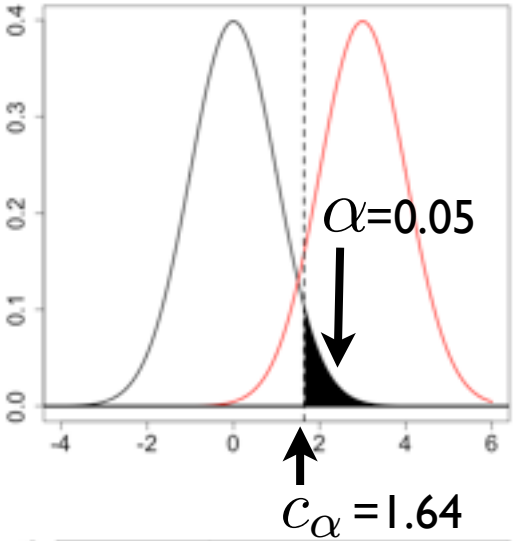
Results of hypothesis decisions I: when H0 is correct (!!)

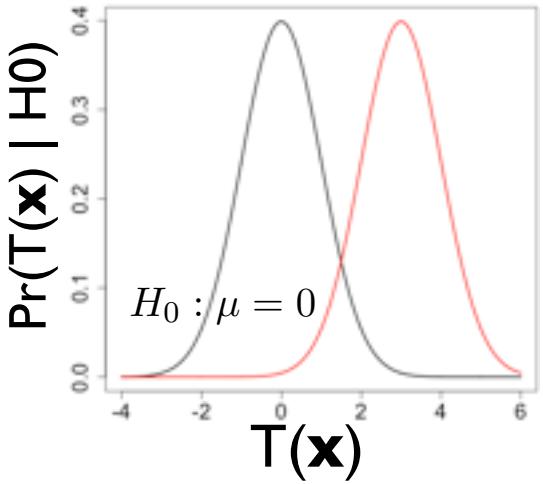
- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

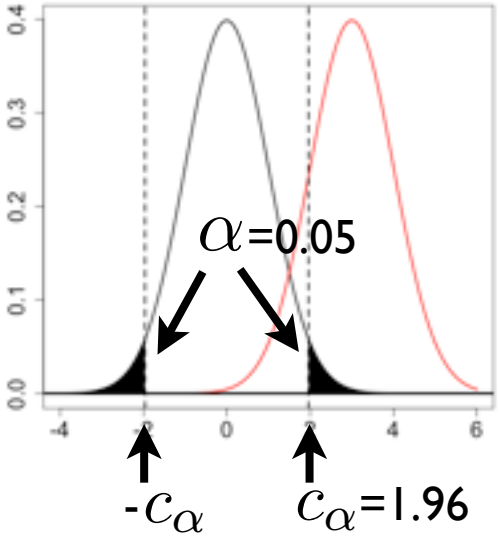
	H_0 is true
cannot reject H_0	1-α, (correct)
reject H_0	α, type I error



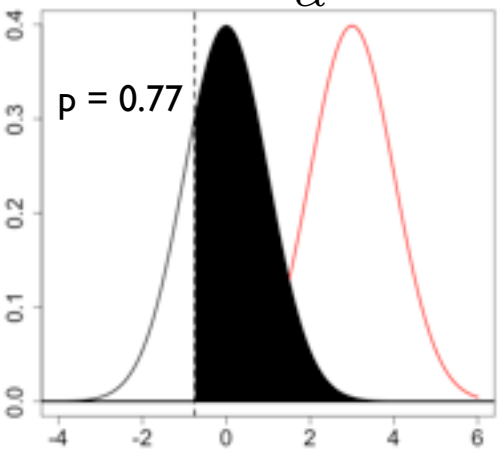
Assume H0 is wrong (!): $\mu = 3$

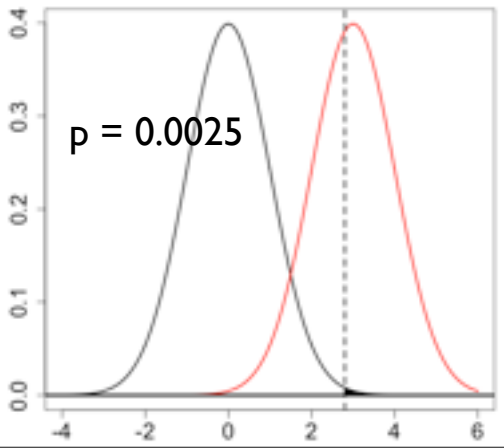






one-sided test



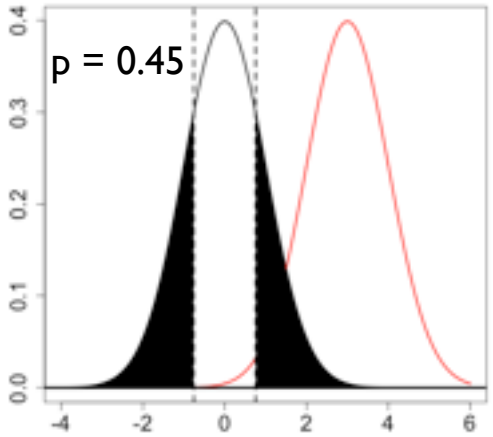


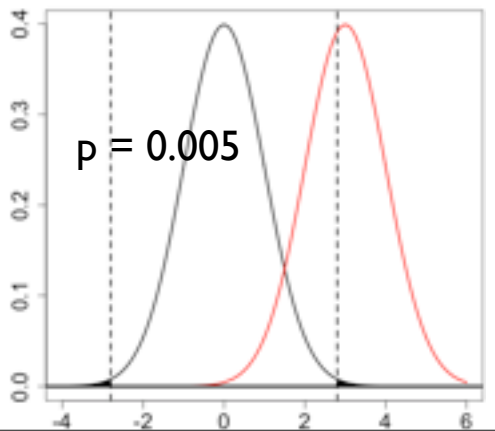
Sample I:

T(**x**)= -0.755

Sample II:

T(**x**)= 2.8



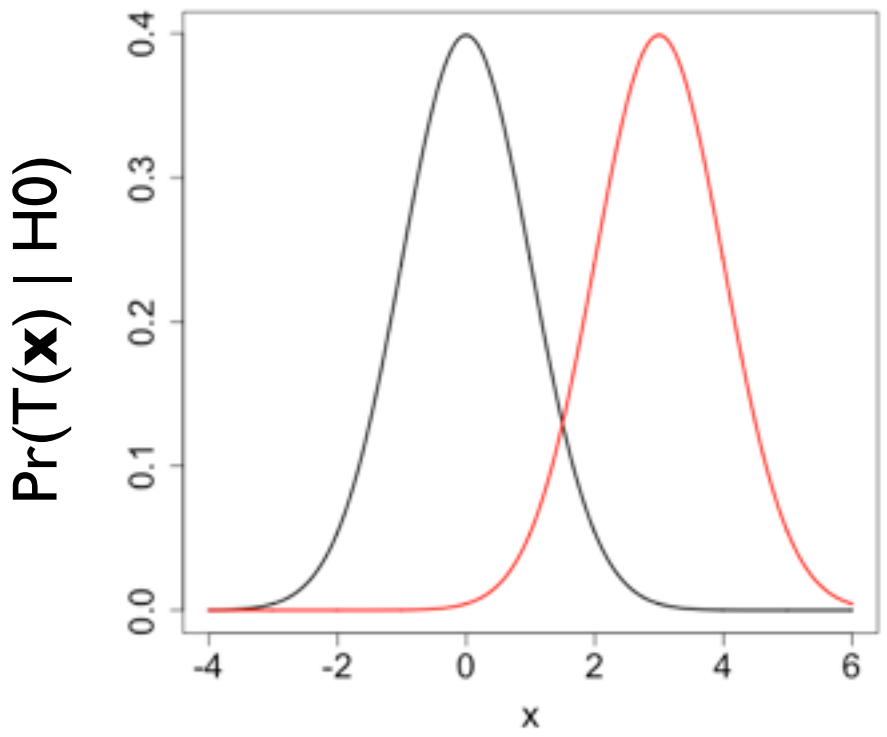


two-sided test

Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

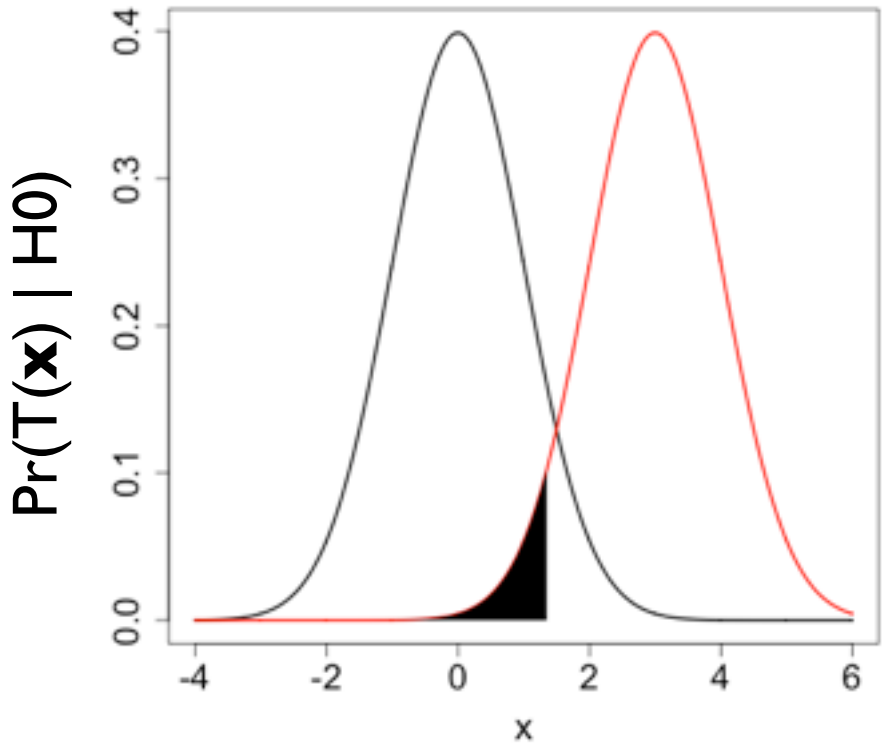
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

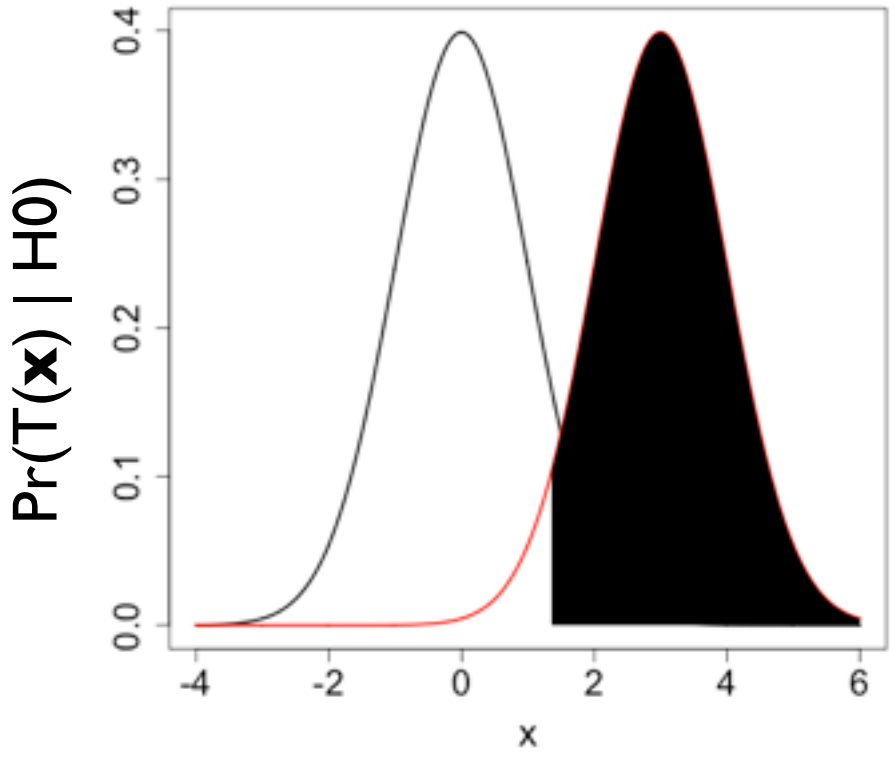
	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Results of hypothesis decisions II: when H0 is wrong (!!)

- There are only two possible decisions we can make as a result of our hypothesis test: *reject* or cannot *reject*

	H_0 is true	H_0 is false
cannot reject H_0	1-α, (correct)	β, type II error
reject H_0	α, type I error	$1-\beta$, power (correct)



Important concepts I

- REMEMBER (!!): there are two possible outcomes of a hypothesis test: we reject or we cannot reject
- We never know for sure whether we are right (!!)
- If we cannot reject, this does not mean H0 is true (why?)
- Note that we can control the level of type I error because we decide on the value of α

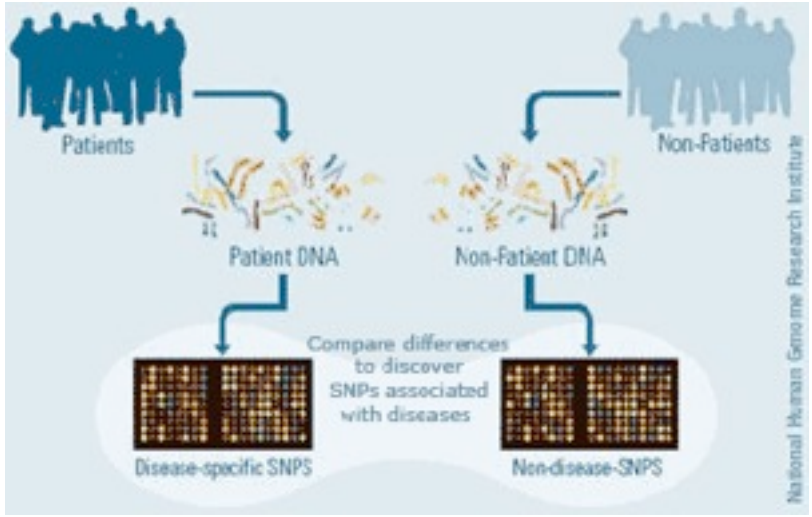
Final general concept 11

- Note that since we have induced a probability model on our r.v. -> sample -> statistic, and a p-value is a function on a statistic, we also have a probability distribution on our p-values
- This is the possible p-values we could obtain over an infinite number of different samples (sets of experimental trials)!
- This distribution is always (!!) the uniform distribution on [0,1] (regardless of the statistic or hypothesis test):

$$Pr(pval) \sim U[0,1]$$

GWAS structure

- Genome-wide association studies (GWAS) are used to map the genomic location of disease loci:

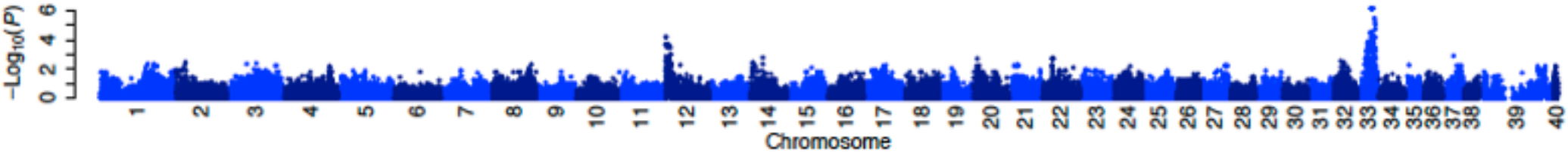


- Associations are identified by calculating p-values when considering an H0 of no association between phenotype and genotype for each:

$$\mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \epsilon$$

$$H_0 : \beta_a = 0 \cap \beta_d = 0$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$



Quantile-Quantile (QQ) plots

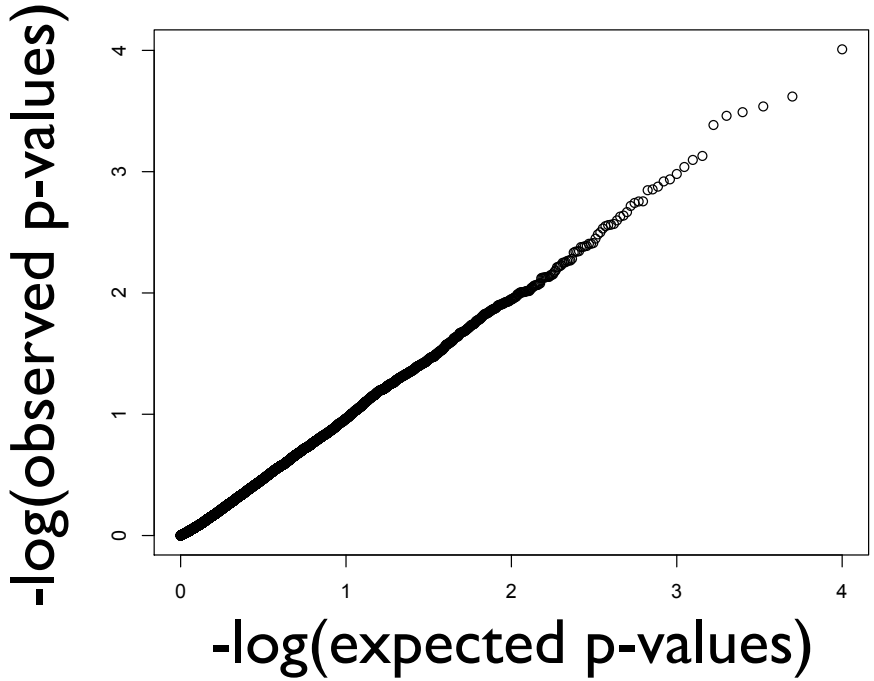
- An essential tool for detecting the problems in a GWAS is a Quantile-Quantile (QQ) plot
- **quantile** - regular, equally spaced intervals of a random variable that divide the random variable into units of equal distribution
- A Quantile-Quantile (QQ) plot (in general) plots the observed quantiles of one distribution versus another OR plots the observed quantiles of a distribution versus the quantiles of the ideal distribution
- In GWAS we use a QQ plot to plot our the quantile distribution of observed p-values (on the y-axis) versus the quantile distribution of expected p-values (what distribution is this!?)

Quantile-Quantile (QQ) plots

- How to construct a QQ plot for a GWAS:
  - If you performed *N* tests, take the -log (base 10) of each of the p-values and put them in rank order from smallest to largest
  - Create a vector of *N* values evenly spaces from 1 to 1 / *N* (how do we do this?), take the -log of each of these values and rank them from smallest to largest
  - Take the pair of the smallest of values of each of these lists and plot a point on an x-y plot with the observed -log p-value on the y-axis and the spaced -log value on the x-axis
  - Repeat for the next smallest pair, for the next, etc. until you have plotted all *N* pairs in order

Quantile-Quantile (QQ) plots

- In an ideal GWAS case where there ARE NO causal polymorphisms, your QQ plot will be a line:

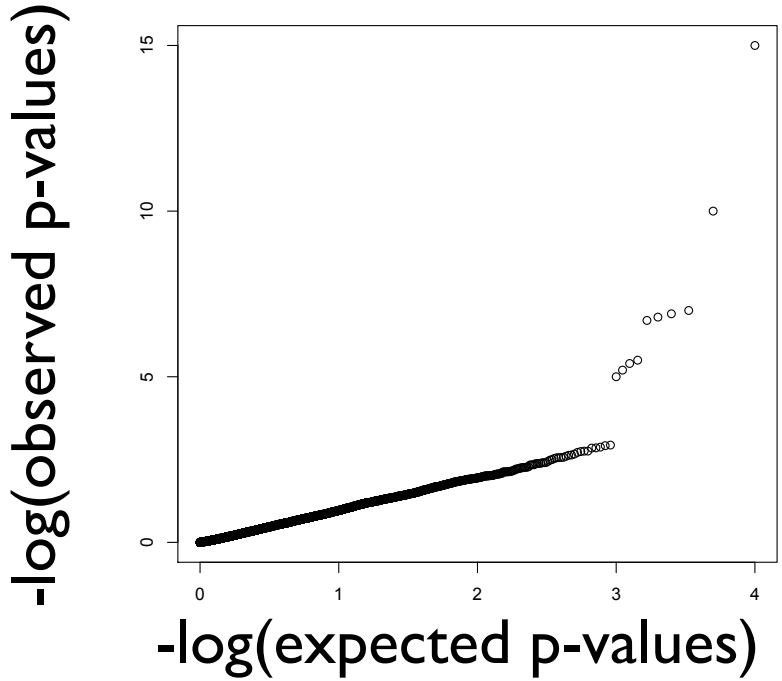


- The reason is that we will observe a uniform distribution of p-values from such a case and in our QQ we are plotting this observed distribution of p-value versus the expected distribution of p-values: a uniform distribution (where both have been -log transformed)

- Note that if you GWAS analysis is correct but you do not have enough power to detect positions of causal polymorphisms, this will also be your result (!!), i.e. it is a way to assess whether you can detect any hits in your GWAS (!!)

Quantile-Quantile (QQ) plots

- In an ideal GWAS case where there ARE causal polymorphisms, your QQ plot will be a line with a tail (!!):

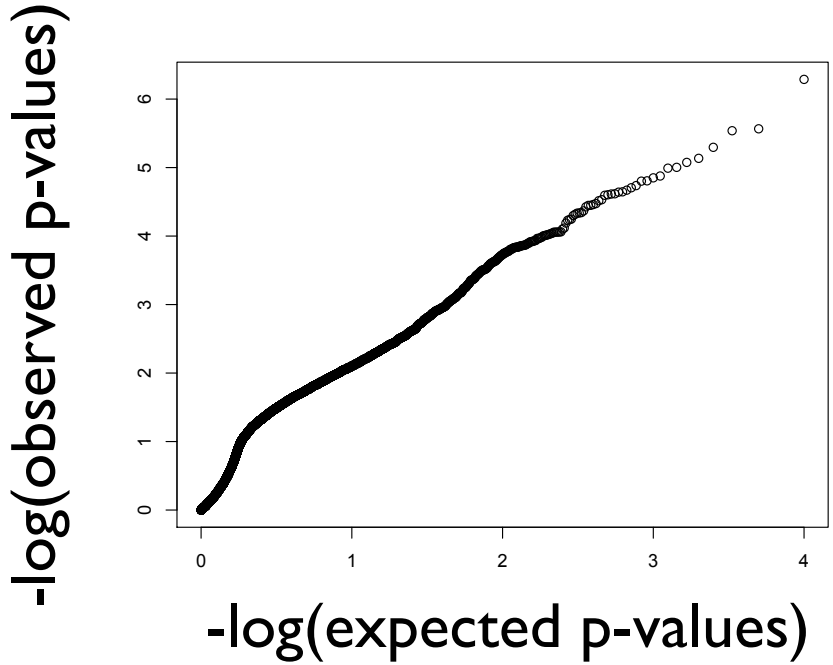


- This happens because most of the p-values observed follow a uniform distribution (i.e. they are not in LD with a causal polymorphism so the null hypothesis is correct!) but the few that are in LD with a causal polymorphism will produce significant p-values (extremely low = extremely high -log(p-values)) and these are in the "tail"

- This is ideally how you want your QQ-plot to look - if it does, you are in good shape!

Quantile-Quantile (QQ) plots

- In practice, you can find your QQ plot looks different than either the "null GWAS" case or the "ideal GWAS" case, for example:



- This indicates that something is wrong (!!!!) and if this is the case, you should not interpret any of your significant p-values as indicating locations of causal polymorphisms (!!!!)
- Note that this means that you need to find an analysis strategy such that the result of your GWAS produces a QQ plot that does NOT look like this (note that this takes experience and many tools to do consistently!)
- Also note that unaccounted for covariates can cause this issue and the most frequent culprit is unaccounted for population structure (see lab!)

Deciding on which p-values indicate associations: Type 1 error

- Say your QQ looks good - how do we decide on a cutoff for determining which p-values are significant?
- Recall that Type 1 error is the probability of incorrectly rejecting the null hypothesis when it is correct
- We can control Type 1 error by setting it to a specified level but recall there is a trade-off: if we set it to low, we will not make a Type 1 error but we will also never reject the null hypothesis, even when it is wrong
- In general we like to set a conservative Type 1 error for a case where we perform many tests (why is this!?)
- To do this, we have to deal with the *multiple testing problem*

Multiple Testing

- Say that we perform *N* hypothesis tests
- Recall that if we set a Type I error to a level (say 0.05) this is the probability of incorrectly rejecting the null hypothesis
- If we performed *N* tests that were independent, we would therefore expect to incorrectly reject the null *N**0.05 and if N is large, we would therefore make LOTS of errors (!!)
- This is the multiple testing problem = the more tests we perform the greater the probability of making a Type I error

Correcting for multiple tests I

- Since we can control the Type I error, we can correct for the probability of making a Type I error due to multiple tests
- There are two general approaches for doing this: those that involve a *Bonferroni correction* and those that involve a correction based on the estimate the *False Discovery Rate* (FDR)
- Both are different techniques for controlling Type I error but in practice, both set the Type I error to a specified level (!!)

Correcting for multiple tests II

- A Bonferroni correction sets the Type I error for the entire set of N tests using the following approach: for a desired type I error α set the Bonferroni Type I error α_B to the following:

$$\alpha_B = \frac{\alpha}{N}$$

- We therefore use the Bonferroni Type I error to assess EACH of our *N* tests
- For example, if we have *N*=100 and we want an overall Type I error of 0.05, we require a test to have a p-value less than 0.0005 to be considered significant

Correcting for multiple tests III

- A False Discovery Rated (FDR) based approach (there are many variants!) uses the expected number of false positives to set (=control) the type 1 error
- For *N* tests and a specified Type 1 error, the FDR is defined in terms or the number of cases where the null hypothesis is rejected R:

$$FDR = \frac{N * \alpha}{R}$$

- Intuitively, the FDR is the proportion of cases where we reject the null hypothesis that are false positives
- We can estimate the FDR, e.g. say for *N*=100,000 tests and a Type I error of 0.05, we reject the null hypothesis 10,000 times, the FDR = 0.5
- FDR methods for controlling for multiple tests (e.g. Benjamini-Hochberg) set the Type 1 error to control the FDR to a specific level, say FDR=0.01 (what is the intuition at this FDR level?)

History and future of GWAS

- Modern age of GWAS started in ~2007 with the Wellcome Trust Case Control Consortium
- Since then, several hundred association hits have been been replicated (thousands have not been replicated)
- Of these several hundred, in a few cases we have a great candidate gene (and in even fewer a causal polymorphism!)
- So, one issue is how many of these hits are correct and actually useful

Another issue...

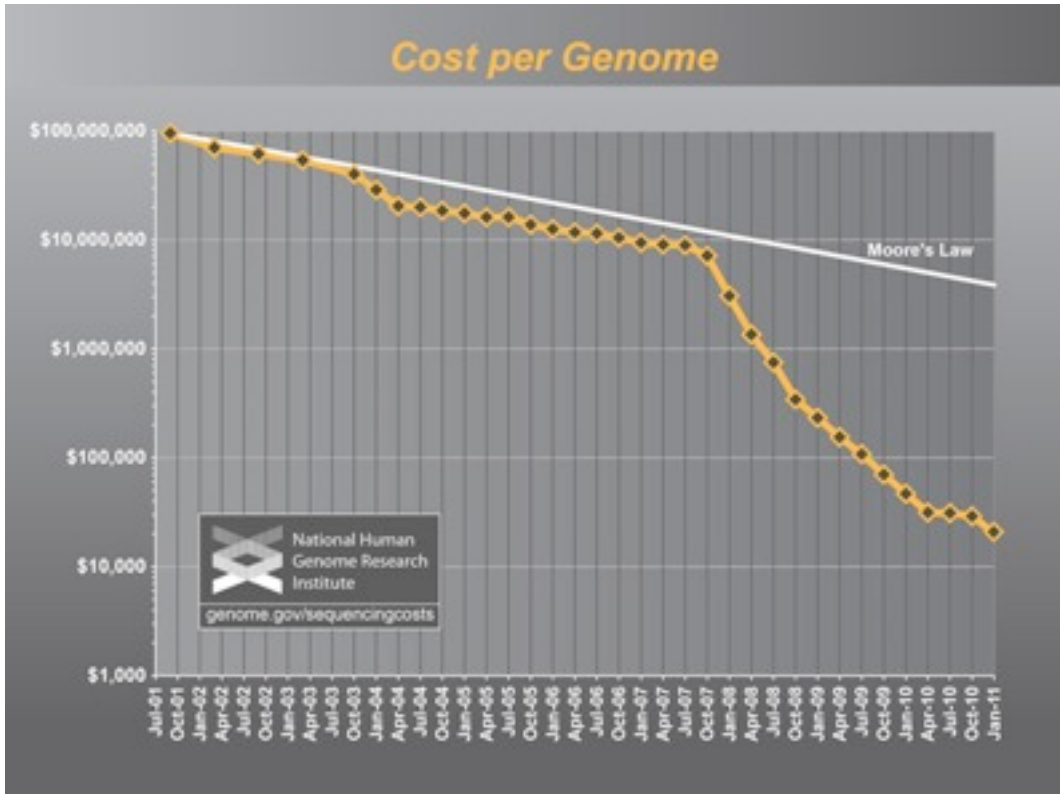
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

The next GWAS wave: consortiums and next-gen

wellcome trust

CHARGE - S

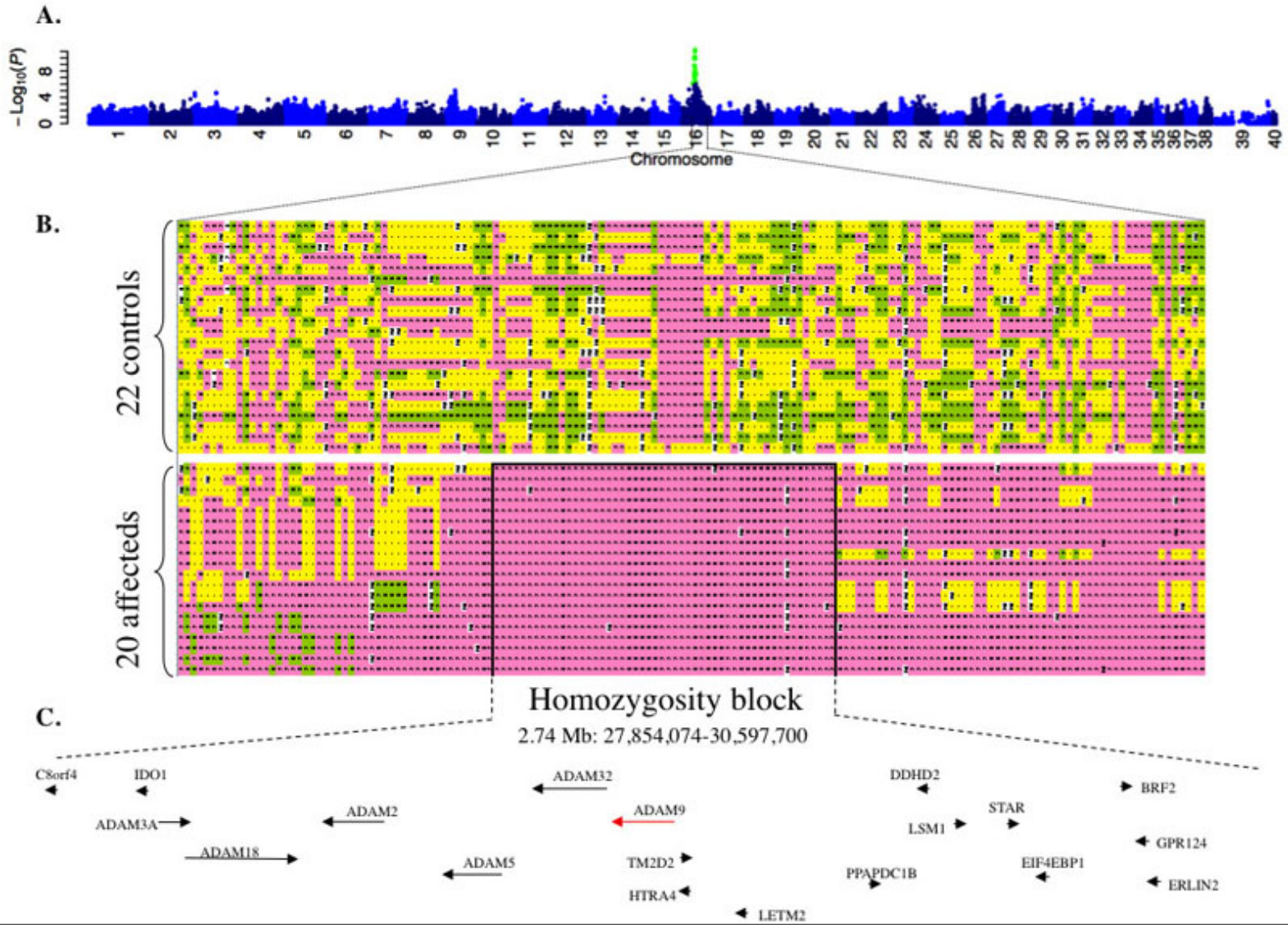


Importance of GWAS

- Regardless of what we find in the next wave, people will always do GWAS analyses, i.e. it is a basic experiment and analysis that provides information on genetics!!
- They will always be useful for traits with simple genetics, i.e. Mendelian
- They are also the core of eQTL (xQTL) analyses

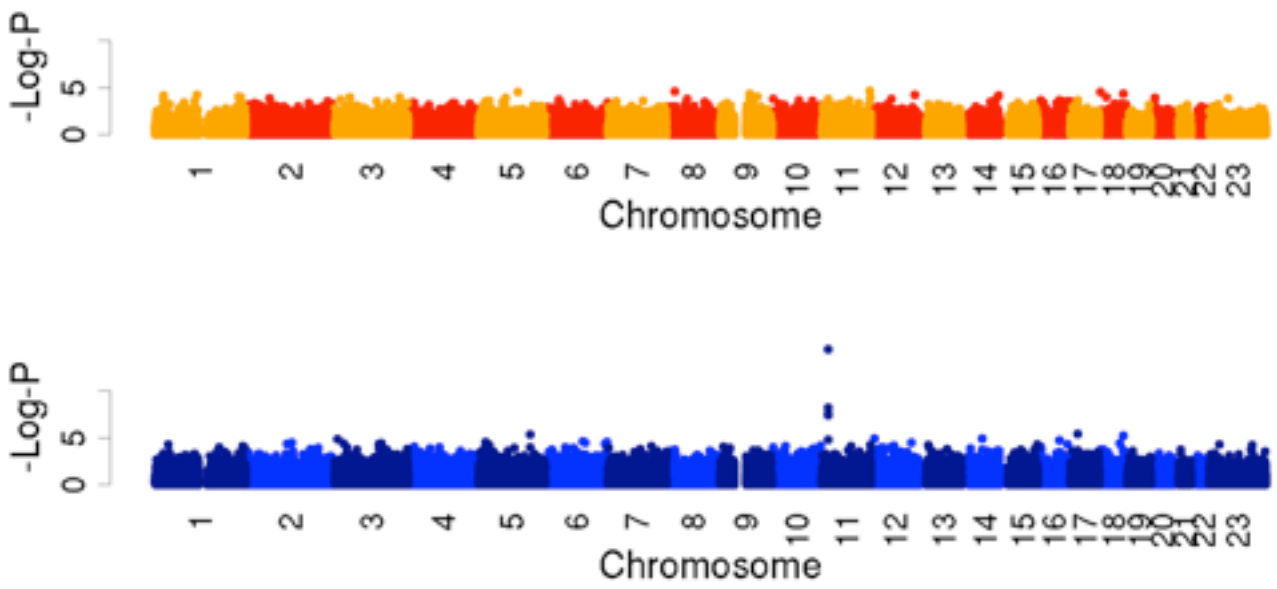
Mendelian disease: Canine cone-rod dystrophy (crd3) in Glen of Imaal Terriers

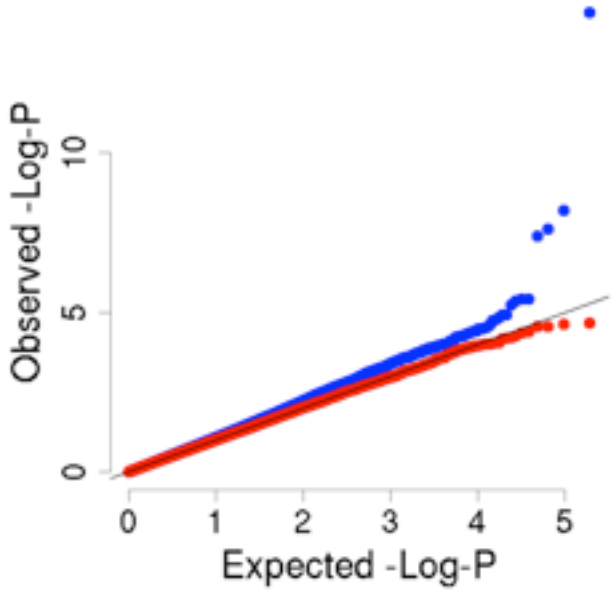
Molecular Vision 2010; 16:1549-1569 <http://www.molvis.org/molvis/v16/a167>



expression Quantitative Trait Locus (eQTL) analysis

- Lung Cancer. 2009 Feb;63(2):180-6. Genetic variants in GTF2H1 and risk of lung cancer: a case-control analysis in a Chinese population.





That’s it!