

Introduction to association studies

...including (but not limited to) genome-wide association studies (GWAS)

Heather J. Cordell

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

`heather.cordell@ncl.ac.uk`



Overview and Aims

- To provide a reminder of the basic concepts, definitions, terminology etc. in genetics
- Historical perspective on genetic studies (of **monogenic** diseases)
 - Familial aggregation/segregation analysis
 - Linkage analysis of family (pedigree) data
- Association analysis (for **complex** genetic diseases)
 - Families or unrelated individuals
 - Candidate genes/candidate variants or genome-wide
 - GWAS using SNP chips
 - Next-generation sequencing studies

Basic Genetics

- Series of molecules (nucleotides, bases) arranged in a double helix structure
 - A Adenine
 - C Cytosine
 - G Guanine
 - T Thymine
- For our purposes, we can consider DNA as a long string of bases
ACCTGTGTGCCCAATGGCGTCCCATACTATCGG

Basic Genetics

- Series of molecules (nucleotides, bases) arranged in a double helix structure

- A Adenine
- C Cytosine
- G Guanine
- T Thymine

- For our purposes, we can consider DNA as a long string of bases

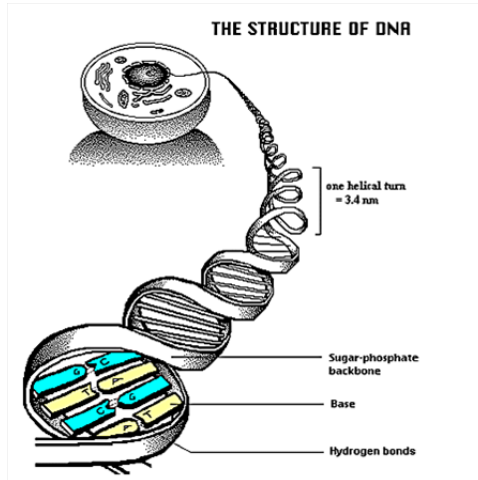
ACCTGTGTGCCCAATGGCGTCCCATACTATCGG

- Actually, 2 such strings, known as the 'forward' and 'reverse' strands
 - With a redundancy in pairing, such that A always pairs with T, and G with C

ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
TGGACACACGGGTACCGCAGGGTATGATAGCC

- (This redundancy means we don't need to show both strings)

Structure of DNA



Genetic variation

- The sequences of unrelated humans are 99.9% identical
- Differences are mostly **single nucleotide polymorphisms** (SNPs) (= single base changes)

Three DNA sequences

```
ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
ACCTGTGCGCCCA GTGGCGTCCCATACTATCGG
ACCTGTGCGCCCAATGGCGTCCCATAGTATCGG
```

- Different sequences are said to possess different **alleles** at these positions

Genetic variation

- As well as SNPs

ACCTGTGTGCCCCAATGGCGTCCCATACTATCGG
ACCTGTGCGCCCA GTGGCGTCCCATAGTATCGG

other types of variation include deletions or inversions

ACCTGTGTGCCCCAATGGCGTCCCATACTATCGG
ACCTGTGTGCCCCA-----ATACTATCGG
ACCTGTGTGCCCCACCTGCGGTAAATACTATCGG

Genetic variation

- As well as SNPs

```
ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
ACCTGTGCGCCAGTGGCGTCCCATAGTATCGG
```

other types of variation include deletions or inversions

```
ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
ACCTGTGTGCCCA-----ATACTATCGG
ACCTGTGTGCCCACCTGCGGTAAATACTATCGG
```

- Or differences in the number of repeats e.g. copy number variants (CNVs) or short tandem repeats (STRs) (sometimes called microsatellites)

```
ACCTG AGTT AGTT AGTT AGTT AGTT ATACTATCGG
ACCTG AGTT AGTT AGTT ---- ---- ATACTATCGG
```


Alleles and genotypes

- Each person has two alleles at each genetic position (=location, locus)
 - One inherited from their father, one from their mother
- Their genotype at a locus is the **combination** of alleles they possess

Two individuals

Person 1	ACCTGTGTGCCCAATGGCGTCCCATACTATCGG
	ACCTGTGCGCCCAATGGCGTCCCATACTATCGG
Person 2	ACCTGTGCGCCCA GTGGCGTCCCATACTATCGG
	ACCTGTGCGCCCA GTGGCGTCCCATAGTATCGG

- The term **haplotype** denotes alleles on the same sequence (inherited from the same parent)

Mendelian inheritance

- Within a family, inheritance depends on the physical proximity of the loci on the DNA strands
 - Alleles at loci that are physically closer tend to get transmitted together (i.e. in coupling)
 - Alleles at loci that are sufficiently far apart (or on different chromosomes) are transmitted independently

Parental transmission

Parent	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
	ACCTGTG C GCCCA T TGGCGTCCCATA A TATCGG
Child 1	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
Child 2	ACCTGTG T GCCCA A TGGCGTCCCATA A TATCGG
Child 3	ACCTGTG T GCCCA T TGGCGTCCCATA A TATCGG

Mendelian inheritance

- Within a family, inheritance depends on the physical proximity of the loci on the DNA strands
 - Alleles at loci that are physically closer tend to get transmitted together (i.e. in coupling)
 - Alleles at loci that are sufficiently far apart (or on different chromosomes) are transmitted independently

Parental transmission

Parent	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
	ACCTGTG C GCCCA T TGGCGTCCCATA A TATCGG
Child 1	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
Child 2	ACCTGTG T GCCCA A TGGCGTCCCATA A TATCGG
Child 3	ACCTGTG T GCCCA T TGGCGTCCCATA A TATCGG

Mendelian inheritance

- Within a family, inheritance depends on the physical proximity of the loci on the DNA strands
 - Alleles at loci that are physically closer tend to get transmitted together (i.e. in coupling)
 - Alleles at loci that are sufficiently far apart (or on different chromosomes) are transmitted independently

Parental transmission

Parent	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
	ACCTGTG C GCCCA T TGGCGTCCCATA A TATCGG
Child 1	ACCTGTG T GCCCA A TGGCGTCCCATA C TATCGG
Child 2	ACCTGTG T GCCCA A TGGCGTCCCATA A TATCGG
Child 3	ACCTGTG T GCCCA T TGGCGTCCCATA A TATCGG

Alleles and loci

- A **locus** or **marker** = location on a chromosome (on the genome)
- Said to be **polymorphic** if different forms of genetic material (i.e. different **alleles**) can exist at that location
 - E.g. at a SNP it might be possible to have an A or a G
 - At a repeat marker it might be possible to have the sequence AGTT, or AGTT AGTT, or AGTT AGTT AGTT
- Often we label the alleles alphabetically (e.g. A,B,C,D; a,b,c,d) or numerically (e.g. 1,2,3,4)
 - An individual with the same alleles at a locus (e.g. AA) is said to be **homozygous**
 - An individual with the different alleles at a locus (e.g. AB) is said to be **heterozygous**

Measuring genotypes

- Most genetic epidemiological studies do not measure full genome sequences
 - Too expensive/complicated for routine use in genetic epidemiological studies
 - Starting to be used in small-scale studies
 - Or in large-scale projects such as:
 - The international **1000 Genomes Project**
 - The UK Department of Health **100,000 Genomes Project**
 - The NHLBI Trans-Omics for Precision Medicine (TOPMed) program

Measuring genotypes

- Most genetic epidemiological studies do not measure full genome sequences
 - Too expensive/complicated for routine use in genetic epidemiological studies
 - Starting to be used in small-scale studies
 - Or in large-scale projects such as:
 - The international **1000 Genomes Project**
 - The UK Department of Health **100,000 Genomes Project**
 - The NHLBI Trans-Omics for Precision Medicine (TOPMed) program
- Most genetic epidemiological studies involve genotyping a subset of known genetic variants
 - Specific **candidate** genes (or loci)
 - Or else a set of **known polymorphic** markers (SNPs, microsatellites) spaced at intervals across the genome
 - Chosen based on surveys of human genetic variation such as HapMap and 1000 Genomes

- The **phenotype** is the characteristic or trait (e.g. eye colour, height, occurrence of a diabetes) that results from having a specific genotype
- Simple **Mendelian** or **monogenic** disorders show a close correspondence between genotype (at a single genetic locus) and phenotype
 - In **dominant** disorders, only one 'disease' allele is required for an individual to get the disease
 - In **recessive** disorders, two 'disease' alleles are required for an individual to get the disease

Penetrances

- The **penetrance** is the probability of being diseased, given genotype

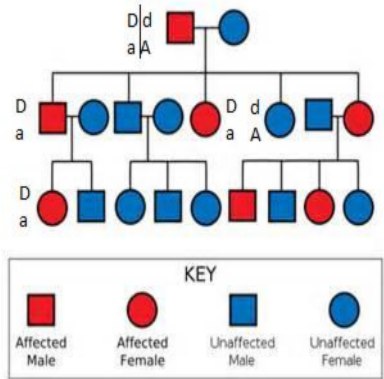
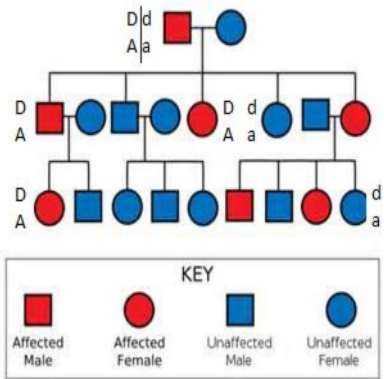
	Dominant	Recessive	Incomplete Penetrance	Genotype Relative Risk (GRR) or odds ratio (OR)
dd	0	0	0.1	1.0
dD	1	0	0.5	5.0
DD	1	1	0.8	8.0

= factor by which your
baseline penetrance
should be multiplied
(for 0,1,2 copies of D)

Parametric linkage Analysis

- Traditionally, genetic determinants of (Mendelian) disease have been identified using **parametric linkage analysis**
 - Relies on ascertaining (a small set of) large families (pedigrees) each containing a number of affected individuals
 - Idea is to examine co-segregation (co-transmission) of disease phenotype and alleles at one or more genetic marker loci
 - Under the assumption that disease **phenotype** reflects an underlying disease **genotype**

Linkage analysis



Parametric linkage analysis

- Calculate likelihood (probability) of observed genotype and phenotype data in (a small set of) large families
 - Under the assumption that the disease is caused by a disease locus situated at recombination fraction θ from a genotyped marker locus
 - Under some assumed mode of inheritance (e.g. dominant, recessive)
- Estimate θ by maximum likelihood techniques
- Test for linkage using likelihood ratio test (LOD score) of the null hypothesis that $\theta = 0.5$
 - “Convincing” evidence for linkage is usually taken as a LOD of 3
 - Corresponds to a likelihood ratio of 1000, i.e. data is 1000 times more likely under the alternative hypothesis than under the null hypothesis

Genetics of common diseases

- **Parametric linkage analysis** has been a highly successful strategy for identifying (localising) genes involved in rare monogenic (single-gene) disorders
 - e.g. Huntingdon's disease, Cystic Fibrosis
- Less successful for common complex disorders
 - Hard to find large pedigrees showing clear disease segregation
 - Complex modes of inheritance: many interacting genetic and environmental factors
 - ⇒ No one-to-one correspondence between genotype and phenotype

Genetics of common diseases

- **Parametric linkage analysis** has been a highly successful strategy for identifying (localising) genes involved in rare monogenic (single-gene) disorders
 - e.g. Huntingdon's disease, Cystic Fibrosis
- Less successful for common complex disorders
 - Hard to find large pedigrees showing clear disease segregation
 - Complex modes of inheritance: many interacting genetic and environmental factors
 - ⇒ No one-to-one correspondence between genotype and phenotype
- **Non-parametric linkage analysis** (e.g. affected sib pair studies) uses a simpler approach
 - Tries to determine whether members of a family with “similar” trait values (e.g. both affected with disease) tend to inherit genetic material in common from their common ancestors
 - More often than would be expected by chance
- However it also has only proved useful in a few instances...

Success of non-parametric linkage

- Type 1 diabetes: confirmed the roles of HLA and insulin genes (Davies et al. 1994)
- Crohn's disease: *NOD2* / *CARD15* gene implicated (Hugot et al. 2001)
- Age-related macular degeneration:
 - Complement factor H gene identified through a combination of approaches, including follow-up of significant regions from non-parametric linkage scan (Haines et al. 2005)
- All of these findings have subsequently been identified through association studies
- Risch and Merikengas (1996) showed that for common genetic variants of small effect, association analysis has greater power

Association vs linkage

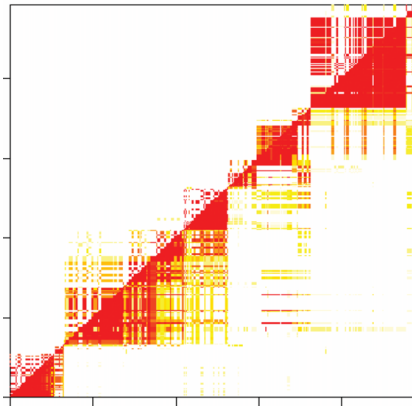
- Linkage studies measure correlation between alleles at the marker (test) locus and disease status **within** families
 - Caused by a lack of recombination between alleles at the marker locus and the underlying disease locus
 - In linkage studies, we test for this **lack of recombination** directly
- Association studies measure correlation between alleles at the marker (test) locus and disease status **across** families
 - Caused by historical lack of recombination between marker and disease alleles over many generations
 - Correlations across families operate even across **unrelated individuals** (=families of size 1)
 - Motivates the use of population-based association studies of unrelated individuals

Correlation across families?

- Why should the same marker allele be correlated with disease status across many different families?
 - This implies that it tends to occur together with (i.e. on the same haplotype as) the disease allele
- One explanation is that this marker allele is itself, in fact, the causal (disease) allele
- Another explanation arises due to a phenomenon known as **linkage disequilibrium** (LD)
 - Alleles at loci that are close together tend to show correlation with one another
 - Suppose SNP 1 with alleles *A* and *C*, SNP 2 with alleles *G* and *T*
 - If SNPs 1 and 2 are in LD, then people who tend to have *A* allele(s) at SNP 1, also tend to have *G* allele(s) at SNP 2 (for example)
 - Occurs due to historical lack of recombination over many generations

Visualising LD

- Plot showing LD measures r^2 (upper) and D' (lower):



- LD across the genome shows a block-like pattern, boundaries correspond to 'recombination hotspots'

(Skip to slide 29)

- Linkage disequilibrium (LD) corresponds to correlation between alleles at two or more loci, at the population level
 - Suppose 2 diallelic loci: locus A with alleles A and a , locus B with alleles B and b
 - If loci A and B are in LD, then people who tend to have A allele(s) at locus A, also tend to have B allele(s) at locus B (for example)
- Most easily described by looking at the haplotype frequencies of the 4 possible haplotypes:

$$A - B, A - b, a - B, a - b$$

Linkage equilibrium (LE)

- Two loci are in linkage *equilibrium* when the probability of observing a certain allele at one locus does not depend on which allele is observed at the other locus
- For our diallelic loci, this means

$$p_{AB} = p_A p_B$$

- p_{AB} is the population frequency (probability) of haplotype $A - B$
- p_A is the population allele frequency of allele A (at locus A)
- p_B is the population allele frequency of allele B (at locus B)

Counting haplotypes

- Suppose we could observe a set of haplotypes (e.g. 12 haplotypes from 6 people)

ID	SNP1	SNP2
1a	A	B
1b	A	B
2a	A	B
2b	a	B
3a	a	B
3b	A	B
4a	A	b
4b	A	B
5a	A	B
5b	A	B
6a	a	B
6b	A	b

$$p_A = 9/12, \quad p_a = 3/12$$
$$p_B = 10/12, \quad p_b = 2/12$$

$$p_{AB} = 7/12, \quad p_{Ab} = 2/12,$$
$$p_{aB} = 3/12, \quad p_{ab} = 0$$

Under LE, we expect $p_{AB} = p_A p_B$

Under LD, we expect $p_{AB} \neq p_A p_B$

Deviation from linkage equilibrium

- The *deviation* from linkage equilibrium is the difference

$$D_{AB} = p_{AB} - p_A p_B$$

- In our example, $D_{AB} = 7/12 - (10/12) \times (9/12) = -0.042$
- The sign of D is not usually considered important
 - Indicates whether haplotype $A - B$ is more or less frequent than expected
 - i.e. does allele A tend to go with allele B , or with allele b
- For 2 diallelic loci it turns out that

$$|D_{AB}| = |D_{Ab}| = |D_{aB}| = |D_{ab}| = |D|, \text{ say}$$

- The range of possible values for $|D|$ depends on the allele frequencies
- Therefore a normalised value D' is often used

$$D' = |D|/D_{\max}$$

where D_{\max} refers to the maximum value $|D|$ could take given the observed allele frequencies

- If $D \geq 0$, D_{\max} is given by the smaller of $p_A p_b$ and $p_a p_B$
- If $D < 0$, D_{\max} is given by the smaller of $p_A p_B$ and $p_a p_b$

- D' varies between 0 and 1
 - 0 meaning no LD
 - 1 meaning 'complete' LD
- For 2 diallelic loci, if $D' = 1$, it means that at least one of the four haplotypes

$A - B$

$A - b$

$a - B$

$a - b$

does not occur

- Turns out to be a good indicator of historical recombination

- An arguably more useful measure of the correlation between alleles at two loci is the (squared) correlation coefficient
 - Denote the allele at locus A as a random variable X (= 1 or 0 according to whether allele is A or a)
 - Denote the allele at locus B as a random variable Y (= 1 or 0 according to whether allele is B or b)
 - Then the correlation coefficient

$$r = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

and r^2 is the square of this

Formula for r^2

- It turns out that

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b}$$

- Estimate in our example:

$$r^2 = \frac{(7/12 - 90/144)^2}{(9/12) \times (3/12) \times (10/12) \times (2/12)} = 0.067$$

(correlation coefficient $r = \sqrt{0.067} = 0.258$)

(Resume here)

- r^2 ranges between 0 (no correlation) and 1 (perfect correlation)
- If $r^2 = 1$, only two of the four possible haplotypes occur
- r^2 has a useful interpretation in terms of power
 - Suppose a sample size of n (e.g. people/chromosomes) is required to detect an association at the causal locus A
 - Then a sample size of n/r^2 is required to detect an association if locus B is typed instead of A.
 - E.g. if 500 cases and controls were required to detect association at A, then $500/0.067 = 7500$ would be required to detect this association via genotyping locus B

Decay of LD

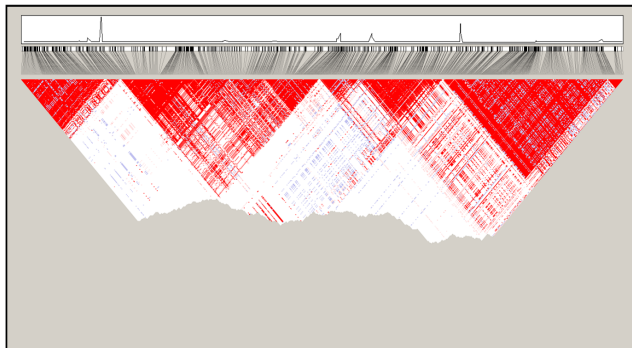
- Since LD is caused by a lack of recombination, and the probability of recombination θ increases with distance, the strength of LD between loci is expected to decline with distance.
- It can be shown that, if in one generation the LD between two loci is represented by D , then in the next generation this measure takes value $(1 - \theta)D$
- So after n generations

$$D_n = (1 - \theta)^n D_0$$

- LD will thus decay over time (and will tend to 0)
- However, the closer the loci, the smaller the rate of decay

Visualising linkage disequilibrium (LD)

- We often visualise LD through a plot, showing the 'LD blocks':



Tagging SNPs

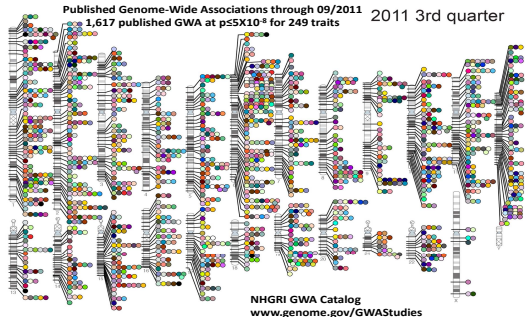
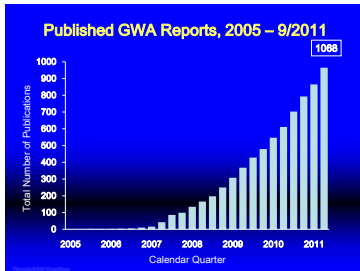
- If two SNPs are in strong LD, then the genotype at SNP2 will be well-predicted by the genotype at SNP1
- If genotype at SNP1 is correlated with a disease phenotype, then genotype at SNP2 will also be correlated with disease phenotype
 - \Rightarrow don't need to genotype both SNPs in order to detect the association with disease
- Using a sample of haplotypes from a reference sample (e.g. HapMap, 1000 Genomes), we can pick a reduced set of SNPs to genotype that provide us with (almost) as much information as the full set

Tagging SNPs

- Often used in candidate gene studies to reduce costs
 - Given a candidate gene or region of interest, need to pick tagging SNPs
 - E.g. using Haploview software
 - <http://www.broadinstitute.org/haploview/haploview>
- Some (early) genome-wide platforms chose to focus on tagging SNPs
 - Improves efficiency (given a fixed number of SNPs)
 - At the expense of reduced redundancy
- Nowadays we have very dense genome-wide platforms containing millions of SNPs
 - **But**, there has been a resurgence of interest in developing **cheaper** SNP-chips with **fewer** (carefully chosen) SNPs
 - e.g. UK Biobank Axiom Array used in the UK Biobank project
 - Through **imputation**, we can infer the genotypes at SNPs that were not actually genotyped

Success of GWAS

- Over the last 10-15 years, there have been a slew of high-profile GWAS, in a variety of different diseases
- Highly successful: many hundreds of associations (between genotype and phenotype) detected



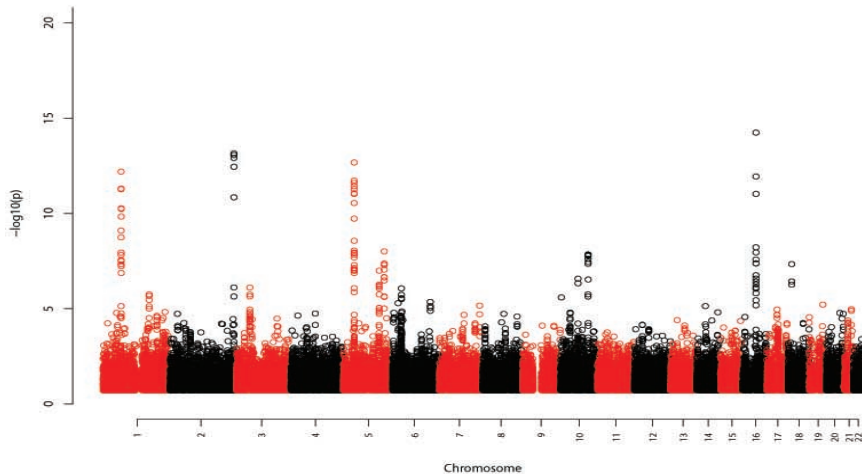
Testing for association

- Most methods produce a **test statistic** and a **p value** indicating how significant the association/correlation between a given SNP and phenotype is
 - i.e. how likely it was to have occurred by chance
- In GWAS, we require stringent significance levels (e.g. $p = 5 \times 10^{-8}$) to overcome the multiple testing problem incurred when we test many SNPs throughout the genome
 - Or, in a Bayesian framework, stringent Bayes Factors to account for the low prior probability that any particular SNP is associated
 - If testing 1 million SNPs using $p = 0.05$, we would obtain 50,000 'significant' results just by chance!
 - We therefore need to use large sample sizes (1000s of individuals) to have sufficient power

Testing for association

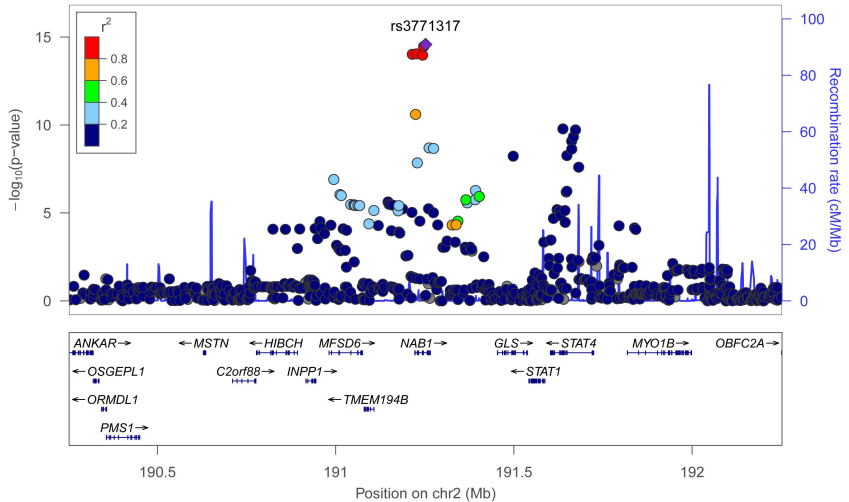
- Most methods produce a **test statistic** and a **p value** indicating how significant the association/correlation between a given SNP and phenotype is
 - i.e. how likely it was to have occurred by chance
- In GWAS, we require stringent significance levels (e.g. $p = 5 \times 10^{-8}$) to overcome the multiple testing problem incurred when we test many SNPs throughout the genome
 - Or, in a Bayesian framework, stringent Bayes Factors to account for the low prior probability that any particular SNP is associated
 - If testing 1 million SNPs using $p = 0.05$, we would obtain 50,000 'significant' results just by chance!
 - We therefore need to use large sample sizes (1000s of individuals) to have sufficient power
- At any location showing 'significant' association, we expect several SNPs in the same region to show association with phenotype
 - Due to correlation (LD) between neighbouring SNPs

Manhattan Plots



Close-up of hit region

Plotted SNPs



Disappointment with GWAS

- GWAS point us to genomic regions (loci) highly likely to harbour disease genes
 - We still don't know the functional (causal) variant, in most cases
 - Indeed, the causal variant may well not even have been genotyped (but SNPs that are correlated with it have been genotyped)
- SNPs identified through GWAS generally have small ORs (< 1.5), suggesting their effects are not very 'important'
 - As we increase sample size, we detect more and more 'significant' SNPs with smaller and smaller effect sizes (ORs)
 - But the SNPs identified do not have strong predictive value (e.g. for predicting disease status)

Disappointment with GWAS



- Problem of 'missing heritability'
 - SNPs identified through GWAS do not fully account for observed correlations in phenotype between close relatives
 - Suggesting there are additional genetic factors to be found...

Is this disappointment warranted?

- GWAS are best considered as a **hypothesis generating** exercise
 - Identifying 'candidate' genomic regions for further investigation
 - Possibly via different types of experiment
 - And potentially pointing us to new biology
 - Ankylosing spondylitis (IL-23 pathway)
 - Schizophrenia (calcium signalling)
 - Inflammatory bowel disease
(IL-23 pathway, autophagy pathway, innate immunity)

Is this disappointment warranted?

- GWAS are best considered as a **hypothesis generating** exercise
 - Identifying 'candidate' genomic regions for further investigation
 - Possibly via different types of experiment
 - And potentially pointing us to new biology
 - Ankylosing spondylitis (IL-23 pathway)
 - Schizophrenia (calcium signalling)
 - Inflammatory bowel disease
(IL-23 pathway, autophagy pathway, innate immunity)
- Identifying 'all' genetic factors involved in a disease is not a realistic goal
 - And, indeed, an unnecessary goal, provided those genetic factors you have identified improve understanding of disease mechanisms

Is this disappointment warranted?

- GWAS are best considered as a **hypothesis generating** exercise
 - Identifying 'candidate' genomic regions for further investigation
 - Possibly via different types of experiment
 - And potentially pointing us to new biology
 - Ankylosing spondylitis (IL-23 pathway)
 - Schizophrenia (calcium signalling)
 - Inflammatory bowel disease
(IL-23 pathway, autophagy pathway, innate immunity)
- Identifying 'all' genetic factors involved in a disease is not a realistic goal
 - And, indeed, an unnecessary goal, provided those genetic factors you have identified improve understanding of disease mechanisms
- See discussion in
 - Visscher et al. (2012) AJHG 90:7-24 "Five Years of GWAS Discovery"
 - Visscher et al. (2017) AJHG 101:5-22 "10 Years of GWAS Discovery: Biology, Function, and Translation"
 - Abdellaoui et al. (2023) AJHG 110:179-194 "15 Years of GWAS Discovery: Realizing the promise"