Genome-Wide Association Studies (GWAS)

Carolin Kosiol
Institute of Population Genetics
Vetmeduni Vienna
<carolin.kosiol@vetmeduni.ac.at>

Spezielle Statistik in der Biomedizin
WS 2014/15

GWAS I

- We now know how to assess the null hypothesis as to whether a polymorphism has a causal effect on our phenotype
- Occasionally we will assess this hypothesis for a single genotype
- In quantitative genomics, we generally do not know the location of causal polymorphisms in the genome
- We therefore perform a hypothesis test of *many genotypes throughout the genome*
- This is a genome-wide association study (GWAS)

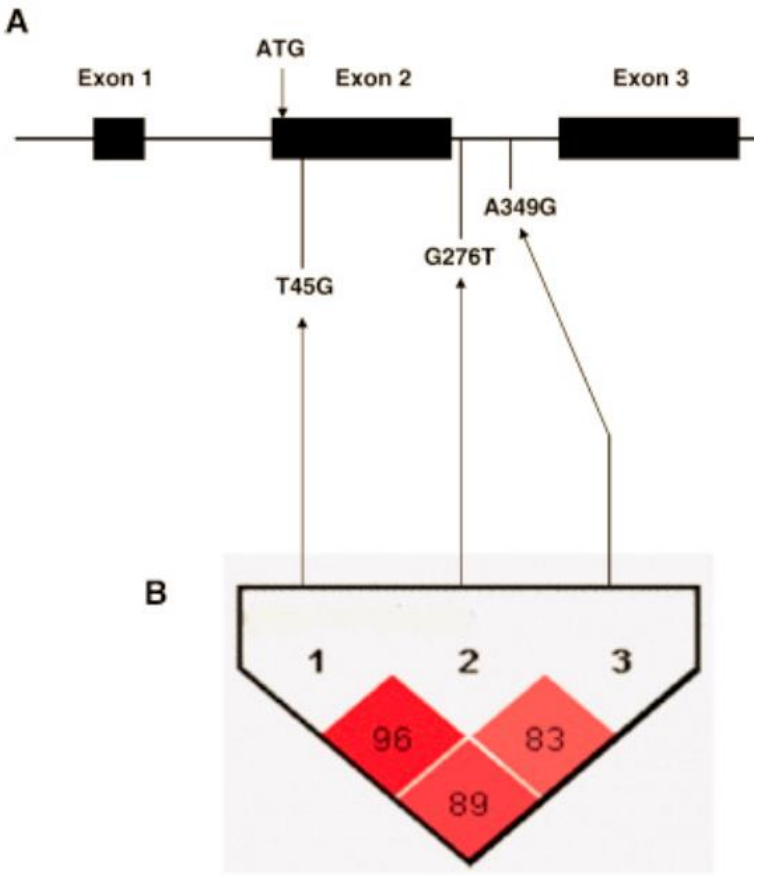
GWAS II

- Analysis in a GWAS raises (at least) two issues we have not yet encountered:
  - An analysis will consist of many hypothesis tests (not just one)
  - We often do not test the causal polymorphism (usually)

- Note that this latter issue is a bit strange (!?) - how do we assess causal polymorphisms if we have not measured the causal polymorphism?

Solution: correlation among genotypes

- 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.
- This is the case in genetic systems.
- Do we know which genotype is causal in this scenario?



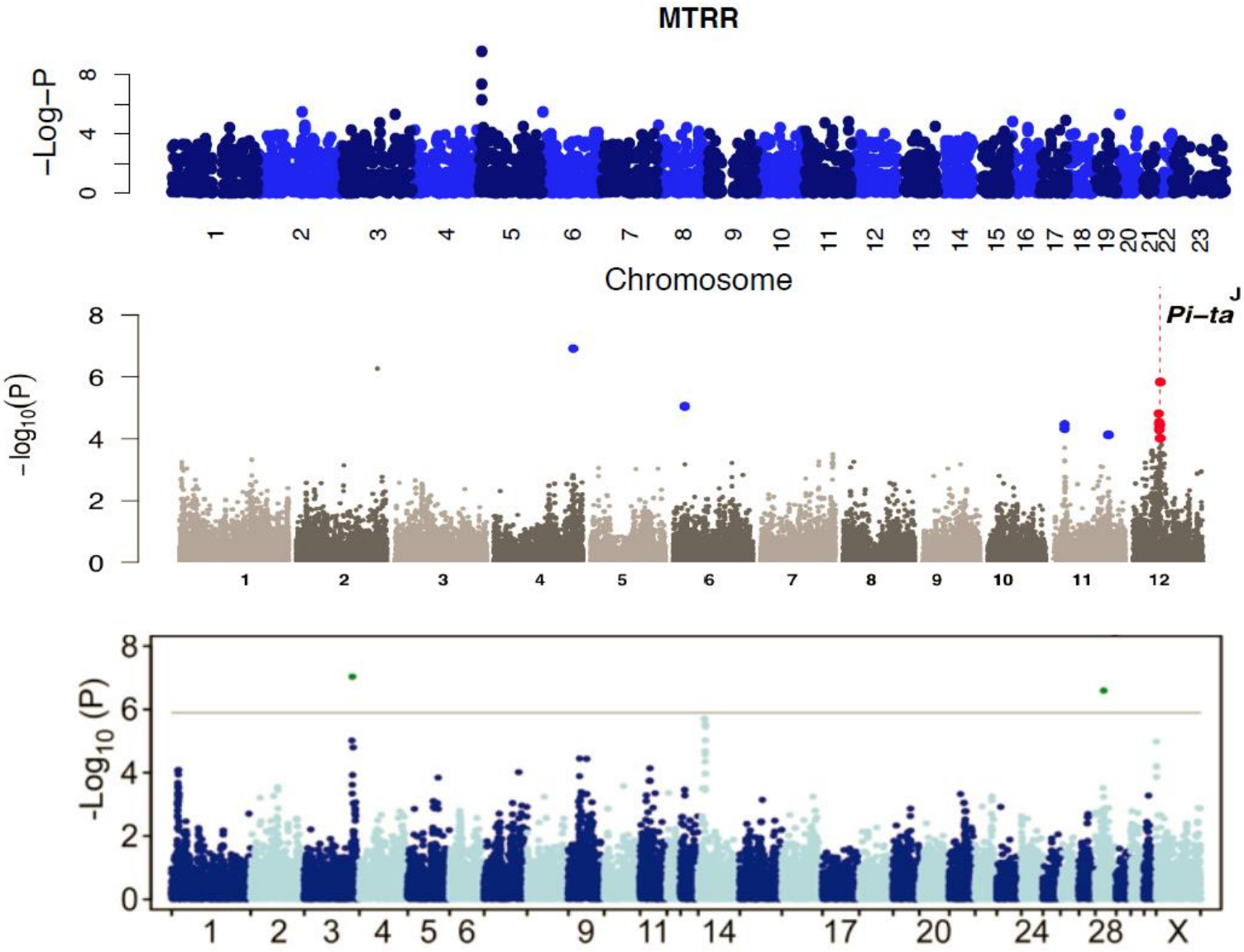
GWAS III

- For a typical GWAS, we have a phenotype of interest and we do not know any causal polymorphisms (loci) that affect this phenotype (but we would like to find them!)
- In an “ideal” GWAS experiment, we measure the phenotype and *N* genotypes THROUGHOUT the genome for *n* independent individuals.
- To analyze a GWAS, we perform *N* independent hypothesis tests.
- When we reject the null hypothesis, we assume that we have located a position in the genome that contains a causal polymorphism (not the causal polymorphism!), hence a GWAS is a *mapping* experiment

Manhattan Plot

- We will consider a number of visualization tools for analyzing GWAS data
- For the moment, we will introduce the *Manhattan plot*
- This is a plot of genotypes on the x-axis and on the y-axis the -log p-values (base 10) (why?) resulting from each hypothesis test of each genotype
- Each “point” on the plot is therefore a single p-value corresponding to a single measured genotype
- We are looking for sets of points with high -log p-value = the position of a causal polymorphism

Manhattan plot: examples



GWAS Definitions

- **Association analysis** - any analysis involving a statistical assessment of a relation between genotype and phenotype, e.g. a hypothesis test involving a multiple regression model
- **Mapping analysis** - an association analysis
- **Linkage disequilibrium (LD) mapping** - an association analysis (we will define LD next lecture)
- **Segregating** - any locus where there is more than one allele in the population
- **Genetic marker** - any segregating polymorphism we have measured in a GWAS, i.e. SNPs genotyped in a GWAS
- **Tag SNP** - a SNP correlated with a causal polymorphism
- **Locus** or **Genetic Locus** - a position in the genome (which may refer to a single polymorphism or an entire genomic segment, e.g. that contains the coding region of a gene

Issues for successful mapping of causal polymorphisms in GWAS

- For GWAS, we are generally concerned with correctly identifying the position of as many causal polymorphisms as possible (True Positives) while minimizing the number of cases where we identify a position where we think there is a causal polymorphism but there is not (False Positive)
- We are less concerned with cases where there is a causal polymorphism but we do not detect it.
- Issues that affect the number of True Positives and False Positives that we identify in a GWAS can be statistical and experimental (or a combination)

Statistical Issues 1: Type 1 error

- Recall that Type 1 error is the probability of incorrectly rejecting the null hypothesis when it is correct
- A Type 1 error in a GWAS produces a false positive
- 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 (e.g. if Type 1 error is to low, we will not detect ANY causal polymorphisms)
- In general we like to set a conservative Type 1 error for a GWAS
- To do this, we have to deal with the *multiple testing problem*

Statistical Issues II: Multiple Testing

- Recall that when we perform a GWAS, we perform *N* hypothesis tests (where *N* is the number of measured genotype markers)
- Also recall that if we set a Type 1 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 he greater the probability of making a Type 1 error

Correcting for multiple tests I

- Since we can control the Type I error, we can correct for the probability of making a Type 1 error due to multiple tests
- There are two general approaches for doing this in a GWAS: 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 1 error but in practice, both set the Type I error to a specified level.

Bonferroni corrections

- A Bonferroni correction sets the Type I error for the entire GWAS using the following approach: for a desired type 1 error set the Bonferroni Type 1 error to the following:

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

- We therefore use the Bonferroni Type I error to asses each of our N tests in a GWAS
- For example, if we have N=100 in our GWAS and we want an overall GWAS Type I error of 0.05, thus we require a test to have a p-value of 0.0005 to be considered significant

False Discovery Rate (FDR)

- For N tests and specified Type I error, the FDR is defined as the number of cases where the null hypothesis is rejected R:
- Intuitively, the FDR is the proportion of cases where we reject the null hypothesis that are false positives

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

- We can estimate the FDR for a GWAS, 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.

Correcting for multiple tests II

- Since the lower the Type 1 error the lower the power of our test, if we set the Type 1 error too low due to a very large N, we might not get any hits even when there are clear causal polymorphisms.
- In general, a Bonferroni correction sets a lower overall GWAS Type I error than FDR approaches (what are the trade-offs and why would we choose one over the other?)
- Both Bonferroni and FDR approaches make the implicit assumption that all tests are independent (which we know not to be the case in GWAS!)

Experimental issues that produce false positives

- Type 1 errors can produce a false positives (= places we identify in the genome as containing a causal polymorphism / locus that do not)
- However, there are experimental reasons why we cancorrectly reject the null hypothesis (= we do not make a Type 1 error) but we still get a false positive:
  - Cases of disequilibrium when there is no linkage
  - Genotyping errors
  - Unaccounted for covariates (upcoming lectures)
  - There are others...

Combined statistical / experimental issues that affect power I

- Recall that *power* is defined as the probability of correctly rejecting the null hypothesis when it is false
- Also recall that we cannot control power directly because it depends on the true parameter value(s) that we do not know!
- Also recall that we can indirectly control power by setting our Type 1 error, where there is a trade-off between Type 1 error and power (what is this trade-off!?)
- There are also a number of issues that affect power that are a function of the GWAS experiment

Combined statistical / experimental issues that affect power II

- Power tends to increase with the increasing size of the true effect of the genotype on phenotype
- Power tends to increase with increasing sample size *n*
- Power tends to increase as the Minor Allele Frequency (MAF) increases
- Power tends to increase as the LD between a causal polymorphism and the genotype marker being tested increases (i.e. as the correlation between the causal and marker genotype increase)
- Power also depends on other factors including the type of statistical test applied, etc.
- Can any of these be controlled?

An issue specific to GWAS: resolution

- **Resolution** - the region of the genome indicated by significant tests for a set of correlated markers in a GWAS
- Recall that we generally consider a set of contiguous significant markers (a “skyscraper” on a Manhattan plot) to indicate the location of a single causal polymorphism (although it need not indicate just one!)
- Note that the marker with the most significant p-value within a set is not necessarily closest to the causal polymorphism
- In practice, we often consider a set of markers with highly significant p-values to span the region where a causal polymorphism is located
- In general, resolution in a GWAS is limited by the level of LD, which means there is a trade-off between resolution and the ability to map causal polymorphisms and that there is a theoretical limit to the resolution of a GWAS experiment

That’s it for now

- After the two week break: We will continue with generalized linear models as well as more methods in statistical genetics such as case & control studies.