Variant-level association analysis

Human Genomic Epidemiology – Asia

Pak Sham

14th June 2022

INTRODUCTION

Variant-based association tests

- After rigorous data quality control, the next step in GWAS is **variant-based association testing**.
- The purpose is to identify SNPs which are associated with the phenotype.
- These associations will include both **direct** and **indirect** association.

Learning objectives

- Understand the principles and practice of variants-based association tests in genome-wide association studies.
  - Variant-based association tests
  - Controlling for confounding
  - Controlling for related subjects
  - Multiple testing adjustment
  - Visual displays
  - Power calculation

ASSOCIATION TESTS

Pearson chi-square tests

Genotype	Cases	Controls
AA	m_{AA}	n_{AA}
Aa	m_{Aa}	n_{Aa}
aa	m_{aa}	n_{aa}

Allele	Cases	Controls
A	$2m_{AA}+m_{Aa}$	$2n_{AA}+n_{Aa}$
a	$2M_{aa}+m_{Aa}$	$2n_{aa}+n_{Aa}$

- For disease phenotypes, the simplest association tests are based on **chi-squared statistics** for equal genotype or allele frequencies in cases and controls

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- O_i and E_i are observed and expected counts in cell i. Expected cell counts are the products of the corresponding row and column totals divided by the grand total.
- Testing for equal genotype frequencies gives a chi-squared test with 2 degrees of freedom (df)
- Testing for equal allele frequencies gives a chi-squared test with 1 df.

Limitations of chi-squared tests

- The genotype-based chi-squared test loses statistical power (more false negatives) when the effects of alleles are approximately additive
- The allele-based chi-squared test is liberal (more false positives) when the 2 alleles in a genotype are not independent, i.e. when there is Hardy-Weinberg Disequilibrium (HWD)
- The allele-based chi-squared test can be generalized to allow for HWD. This test is equivalent to a **Cochran-Armitage trend test**.
- However, this test still has the limitation that it does not allow for potential confounders.

Analysis of variance

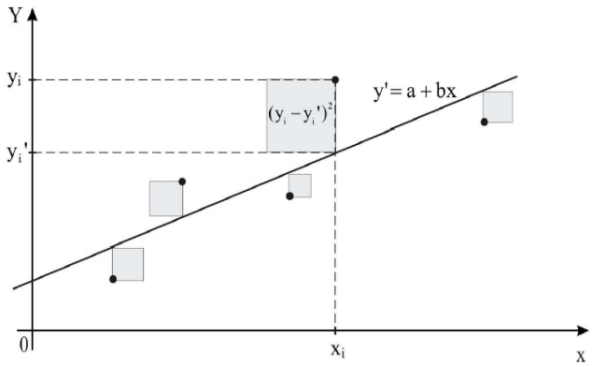
- For quantitative phenotypes, a simple association test would be **analysis of variance** (ANOVA), which is a test of the null hypothesis that 2 or more groups have the same mean value of a trait.
- For 3 genotypes, ANOVA gives an F test with numerator df of 2, and denominator df on n-3 (where n is the overall sample size).
- Like Pearson chi-squared tests, ANOVA also suffers from not being able to allow for potential confounders.

Regression models

- A **linear regression** model predicts the value of a dependent variable (Y) by a linear combination of multiple independent variables (X_1, X_2, X_3, …., X_p)

$$Y' = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_pX_p$$

- The **regression coefficients** (b's) are estimated by minimizing the sum of squared residuals $(Y\text{-}Y')^2$ over the observations (Method of **Least Squares**)



- Inclusion of potential confounding variables into the model allows adjustment for their effects while estimating the effects of genotype

Logistic regression

- When the dependent variable is binary (0, 1) such as the presence or absence of a disease, **logistic regression** is more appropriate than linear regression.
- In logistic regression the log-odds of disease is modelled as a linear combination of predictor variables

$$\ln(p/(1-p)) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_pX_p$$

- The regression coefficients are estimated by the maximizing the log-likelihood function of the data, which is the sum of the log-likelihood of each observation, which in turn is equal to ln(*p*) *for an affected subject and* ln(*1-p*) for an unaffected subject (Method of **Maximum Likelihood**)
- Each **regression coefficient** can be interpreted as the **logarithm of the odds ratio** for an independent variable
- Note: a regression coefficient of 0 translates to an odds ratio of 1

Coding for genetic effects

Genotypes are coded to have numerical values depending on the presumed mode of inheritance of the variant.

Genotype	Dominant	Recessive	Additive
AA	1	1	2
Aa	1	0	1
aa	0	0	0

(Reference allele a)

- The additive (0,1,2) coding, also known as "**allele dosage**", is the default coding for GWAS, since the additive model is usually appropriate for variants of small effect size.
- After a variant is detected to be associated with the phenotype, dominant and recessive modes of inheritance can be explored.

Coding for imputed genotypes

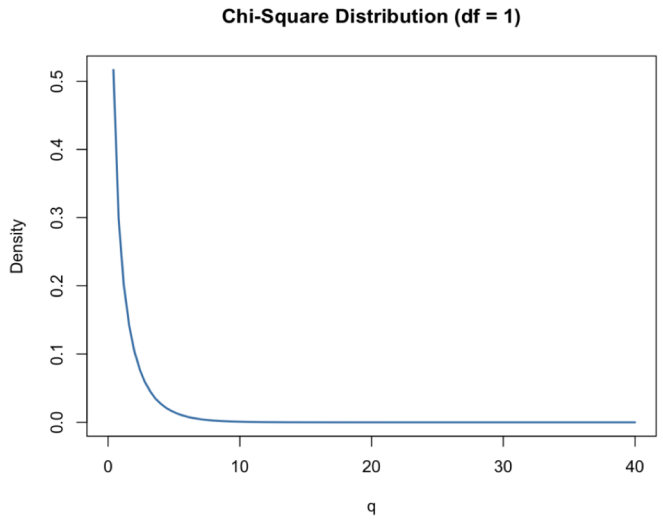
- The additive coding simplifies association analysis of imputed genotypes
- Imputation calculates the probabilities of the different possible genotypes for the untyped SNP in each subject, based on the genotypes on the available typed SNPs.
- Suppose that the 3 imputed genotypic probabilities for a subject are p_{AA}, p_{Aa} and p_{aa}. The imputed allele dosage for allele A is $p_{AA} + (1/2)p_{Aa}$
- This can be entered as the numerical value of the imputed genotype of the subject in a regression analysis.

Coding for X-chromosome variants

- As males only has one X chromosome, their genotype for SNP on the X chromosome is entirely derived from the maternal gamete, and is hemizygous (A,a).
- If genotype a is coded 0, then should genotype A be coded as 1, so that it is equivalent to genotype Aa in females, or 2, so that it is equivalent to the AA genotype in females, or somewhere in between?
- There is no clear answer to this as the appropriate coding may differ for different variants, although a coding of 2 is perhaps more common.
- One way of avoiding the issue is to perform GWAS stratified by sex, and then to perform a meta-analysis to combine the sex-specific tests into a single test (e.g. by Fisher's method for combining p-values).

Association test

- Having obtained an estimate of the regression coefficient of the genotype (*b*) and its standard error (s), by ordinary least squares or maximum likelihood, a test statistic for association can be calculated as $X^2=(b/s)^2$
- Under the null hypothesis E(b)=0, and b/s has a standard normal distribution in large samples. The test statistic $(b/s)^2$ therefore has a chi-square distribution with 1 df (which has mean 1 and variance 2).



Odds ratios and risk ratios

- One of the attractions of logistic regression analysis for case-control studies is that the regression coefficients can be interpreted as log-odds-ratios.
- Risk ratios cannot be directly estimated from case-control studies. However, odds ratios can be converted to risk ratios if the population prevalence of the disease is known
- When the disease prevalence is very low, the risk ratio is closely approximated by the odds ratio (recall odds = risk/(1-risk); when risk is close to 0, 1-risk is close to 1, so that odds ~ risk).

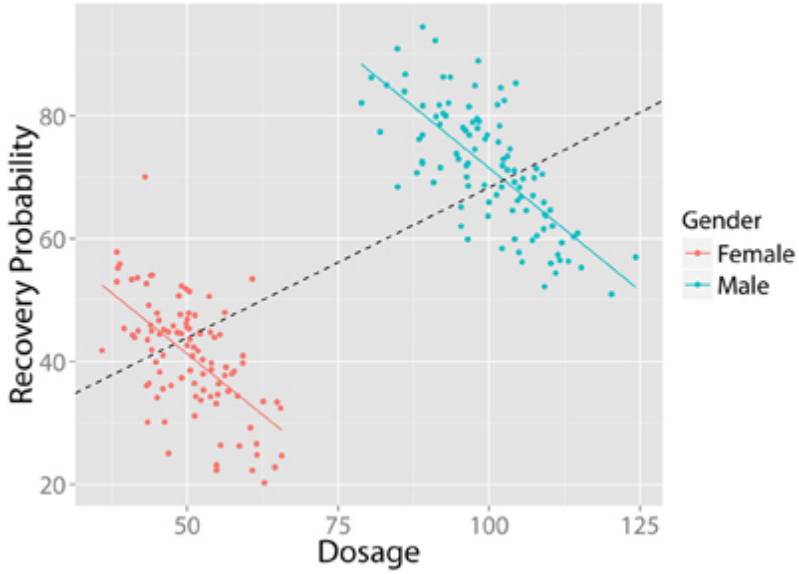
Linear regression for binary traits

- Although logistic regression is considered more appropriate for binary traits than linear regression, the 2 types of regression produce almost equivalent results when the effect size of predictor variables are very small (which is true for GWAS of complex traits).
- Linear regression is much faster than logistic regression, and more readily generalized to more sophisticated methods (e.g. linear mixed models, LASSO). Thus it is not unusual for linear regression and its generalizations to be used even for binary traits, when computational burden is an issue.

CONTROLLING FOR CONFOUNDERS

Confounding

- Mismatching between cases and controls in age, sex or other variables can lead to biased effect size estimates.
- **Simpson's paradox** (the direction of effect being reversed when a "confounding" variable is not taken into account).



Within gender: positive correlation

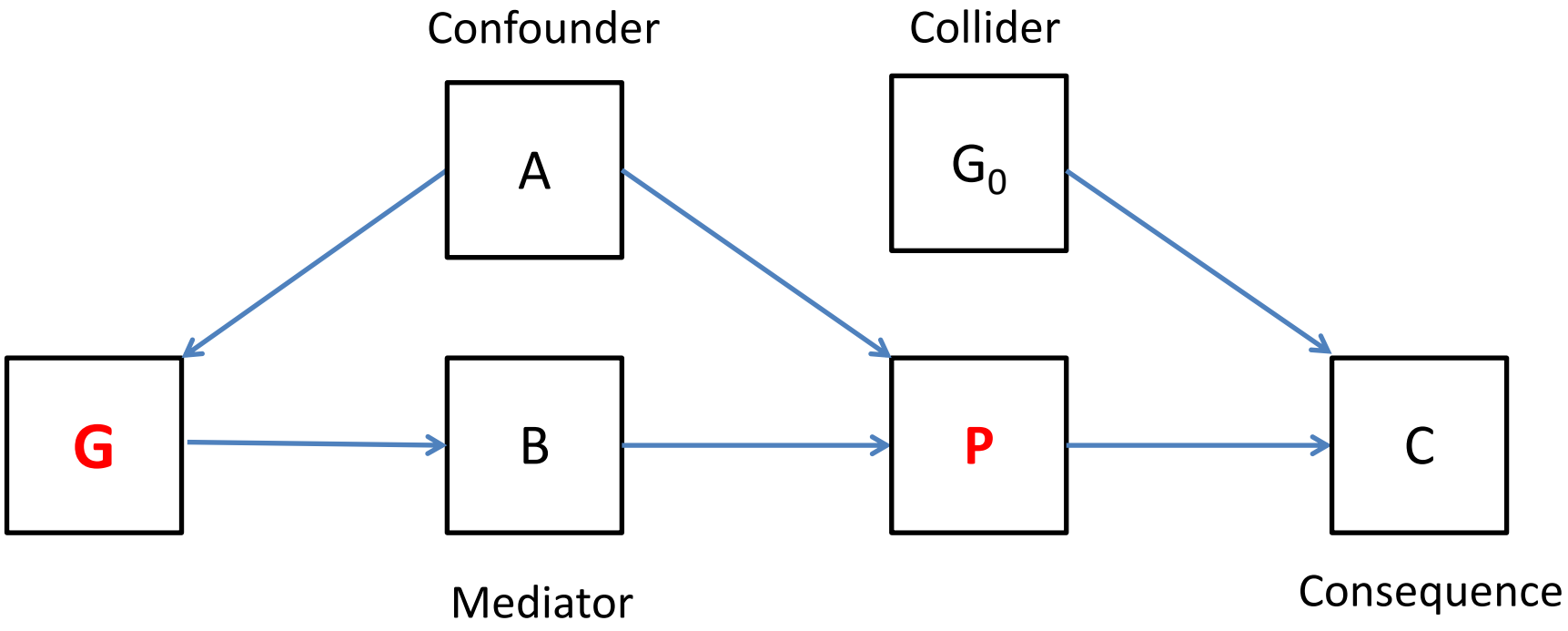
Overall: negative correlation

- A confounding variable is correlated with both the dependent variable (disease) and independent variable (genotype)

Covariates in regression

- A covariate is a predictor variable which is not the variable of interest in the analysis; in GWAS the variable of interest is the SNP genotype.
- Inclusion of a confounder as covariates in a regression analysis will remove its effect from the regression coefficient estimate of the SNP genotype, so that it more accurately represents the causal effect of the genotype on the phenotype.
- The inclusion of too many covariates in relation to the sample size can cause "overfitting" (fitting to noise rather than signal). However, the number of covariates that can be included in GWAS should be large, given that typical GWAS sample sizes are in the 1000's.
- Nevertheless, most GWAS's include very few covariates, usually only **age** **sex** and **principal components**. Why?

What to include as covariate



- Not covarying for A will bias the estimate of the causal effect of G on P.
- Covarying for B or C will under-estimate the causal effect of G on P.
- Covarying for C can also lead to spurious association (of G_0) due to collider bias
- Few variables will qualify as A, since genotype is determined at conception.

Population stratification

- The **population ancestry** of a subject is one variable that can influence their genotype.
- If population ancestry is recorded in a GWAS, then the recorded ancestry can be included as covariates in regression.
- However, population ancestry is often not reliably recorded – **hidden population stratification**.
- As an uncontrolled confounder, hidden population can cause a systematic inflation of association test statistics, resulting in spurious (i.e. false) associations.
- Underlying population stratification can be revealed by **principal components analysis** (PCA); a number of **principal components** can be then included as covariates in regression analysis to control for hidden population stratification.

Genomic control method

- **Genomic control** method attempts to “shrink” inflated association studies due to hidden population stratification or other problems
- The genomic control method defines **Lambda** = Median of observed chi-squared test statistics / Median of chi-squared distribution (i.e. 0.456).
- Corrected test statistics = Observed test statistics / Lambda
- The choice of the median (rather than the mean) for calculating Lambda is that the median is less influenced by a few highly significant SNPs, which may reflect true associations rather than a systematic inflation of test statistics.
- Genomic control correction may not be necessary if adjustment for principal components has brought lambda close to 1.

CONTROLLING FOR RELATED SUBJECTS

Related individuals

- Classical regression models assume **independent observations**
- This means the “errors” (residual variation after allowing for the influences of all predictors) are uncorrelated among the subjects.
- Violation of this assumption can invalidate association tests, producing too many false positives.
- When some subjects in a GWAS are related to each other, this violates the independent observations assumption.
- Part of the sample quality control is to identify groups of closely related individuals in the data. It is usual to retain only one subject from each group of closely related individuals for association analysis

Linear mixed models

- The exclusion of genetically related subjects causes lose of information, especially when many individuals in the sample are related to each other.
- Genetic relationships between individuals can be modeled as a **random effect** in a **linear mixed model** (LMM).
- LMM includes fixed effects, which are measured predictors (e.g. SNP genotype, principal components), as in standard regression models.
- However, LMM also includes random effects, which are unmeasured (i.e. latent) variables.
- The **correlations of random effects** among individuals are assumed to be known. Since latent variables have no specified units of measurement, the effect size of a random effect is usually measured by variance explained, rather than a regression coefficient.

Examples of random effects

- **Household** – each household is associated with a random effect which is assumed to have a normal distribution across households. The household effect is the same for residents of the same household, but uncorrelated for residents in different households. Whether individuals share the same household or not can be specified by a square correlation matrix, where the ij'th element is 1 if individuals i and j share the same household, and 0 otherwise.
- In the classical twin design, **additive genetic effects** and **shared environmental effects** are both random effects. The correlation of additive genetic effects is 1 for MZ twins and 0.5 for DZ twins. The correlation of shared environmental effects is 1 for both MZ and DZ twins.

LMM for GWAS

- LMM can accommodate unmeasured variables which may lead to phenotypic correlations among observations can be included as random effects.
- For GWAS, the most important random effect is the **residual polygenic effect** not explained by the specific SNP being tested. However, other random effects, such as household effect, can also be included.
- For individuals of known genetic relationship (e.g. first cousins), their correlation in polygenic effect can be specified as twice their kinship coefficient.
- However, people are often unaware of more distant relationships.

LMM for GWAS

- Fortunately, kinship coefficients between individuals can be estimated from the genotype data on a large number of SNPs.
- In practice, twice the kinship matrix is approximated by a **genetic relationship matrix** (GRM), where each element is the sample covariance between the standardized allele dosages across all SNPs in 2 individuals

$$g_{ij} = \frac{\sum_{k=1}^{m} G_{ik} G_{jk}}{m}$$

- Since random effects should be uncorrelated with the predictor of interest in the LMM, when a specific SNP is being tested as a fixed effect, it should be excluded from the GRM calculation. To avoid having to calculate a large number of GRMs, it is common practice to calculate GRMs excluding SNPs from an entire chromosome.

LMM for GWAS

- The **phenotypic variance explained** by a random effect is typically estimated by **restricted maximum likelihood** (REML)
  - Residual polygenic effect: σ_g^2
  - Residual error: σ^2
- The **residual phenotypic covariances** explained by random effects are:
  - Residual polygenic effect: $\sigma_g^2\mathbf{G}$ (**G**=GRM)
  - Residual error: $\sigma^2\mathbf{I}$ (**I**=Identity matrix)
  - Total: $\mathbf{\Sigma} = \sigma_g^2\mathbf{G} + \sigma^2\mathbf{I}$
- The effect of a SNP genotype can then be estimated and tested using **generalized least squares** (GLS):

$$\hat{\beta} = (\mathbf{x}^T\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}^T\Sigma^{-1}\mathbf{y} \qquad Var(\hat{\beta}) = (\mathbf{x}^T\Sigma^{-1}\mathbf{x})^{-1}$$

- Variations of this methodology have been proposed and implemented (e.g. BOLT-LMM)

MULTIPLE TESTING ADJUSTMENT

Impact of multiple testing

- The p-value of a statistical test is the probability that the observed value of the test statistic is exceeded, if the null hypothesis is true.
- This means that, when the null hypothesis is true, the probability of obtaining a p-value of less than a particular value is equal to that value. e.g. Prob($p<0.05$)=0.05, Prob($p<0.01$)=0.01.
- Thus, when a large number of SNPs are tested, many tests for which H_0 is true will have significant p-values, if the p-value threshold is set at conventional levels. e.g. in 100,000 tests, 5,000 are expected to have $p < 0.05$, by chance.
- Genuine positive findings may be “swamped” by chance positive results.

Bonferroni correction

- Raises the critical significance level according to the number of tests. For n independent tests, set significance level to $\alpha = 0.05 / n$

n	α
1	0.05
10	0.005
etc ……	

- For n independent tests, all under H_0, the probability that at least 1 test is significant is $1 - (1 - \alpha)^n$. Controlling this "family-wise" error rate to 0.05 means setting the critical significance level α such that $1 - (1 - \alpha)^n = 0.05$, which gives $\alpha = 1 - (1 - 0.05)^{1/n} \approx 0.05 / n$

Genome-wide significance

- In GWAS, appropriate multiple testing correction may need to take account of all common polymorphisms in the genome, regardless of the number of SNPs actually tested in the study.
- This led to the adoption of fixed genome-wide significance e.g. 5×10^{-8}
- The value of 5×10^{-8} takes into account the total number of SNPs in the genome and their LD relationship. Based on European-ancestry genetic data, statistical modelling showed that a p-value smaller than this is expected to arise by chance once every 20 genome scans.
- The 5×10^{-8} genome-wide significance level may not be sufficiently stringent
  - For populations with greater genetic diversity and weaker LD
  - As more imputed rare SNPs are tested by GWAS

False discovery rate - intuition

- The Bonferroni adjustment, or the fixed genome-wide significance threshold, may over correct when the null hypothesis is false for multiple tests. In this scenario it may be useful to consider the false discovery rate.
- Suppose that a study has performed 100 tests, and 20 of these are significant at $p<0.05$. How many of these 20 significant results would you guess constitute true discoveries (as against false positives)?
- By chance, one would expect 5 out of 100 tests to be significant at $p<0.05$. Therefore one might guess that 15/20 of the significant results to be true discoveries (or in other words, 5/20 to be false discoveries).

False discovery rate

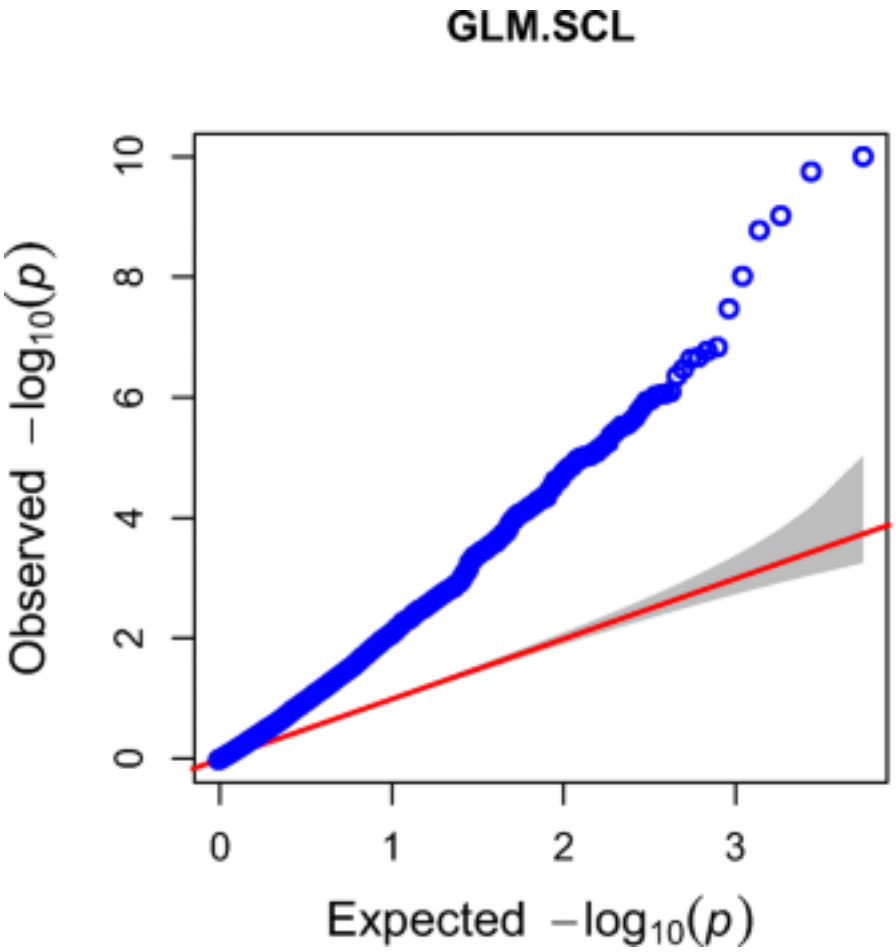
- More generally, if the null hypothesis (H_0) is true for a proportion π_0 of m tests, then the expected number of tests for which H_0 is true is $m\pi_0$. Among these, a proportion α are expected to have p-value less than α. Therefore the expected number of false positives is $\alpha m\pi_0$
- If R tests are actually significant, then the expected proportion of false positive results (i.e. the FDR) is $\alpha m\pi_0/R \sim \alpha m/R$, if π_0 is close to 1.
- In order to control the FDR at a particular desired level, say f, we set $\alpha m/R \leq f$, i.e. the critical $\alpha = (R/m)\, f$
- Thus, the critical significance threshold becomes more stringent, as we progress from the least significant to the most significant test.
- The q-value of a test is the smallest FDR for which the test can be considered significant

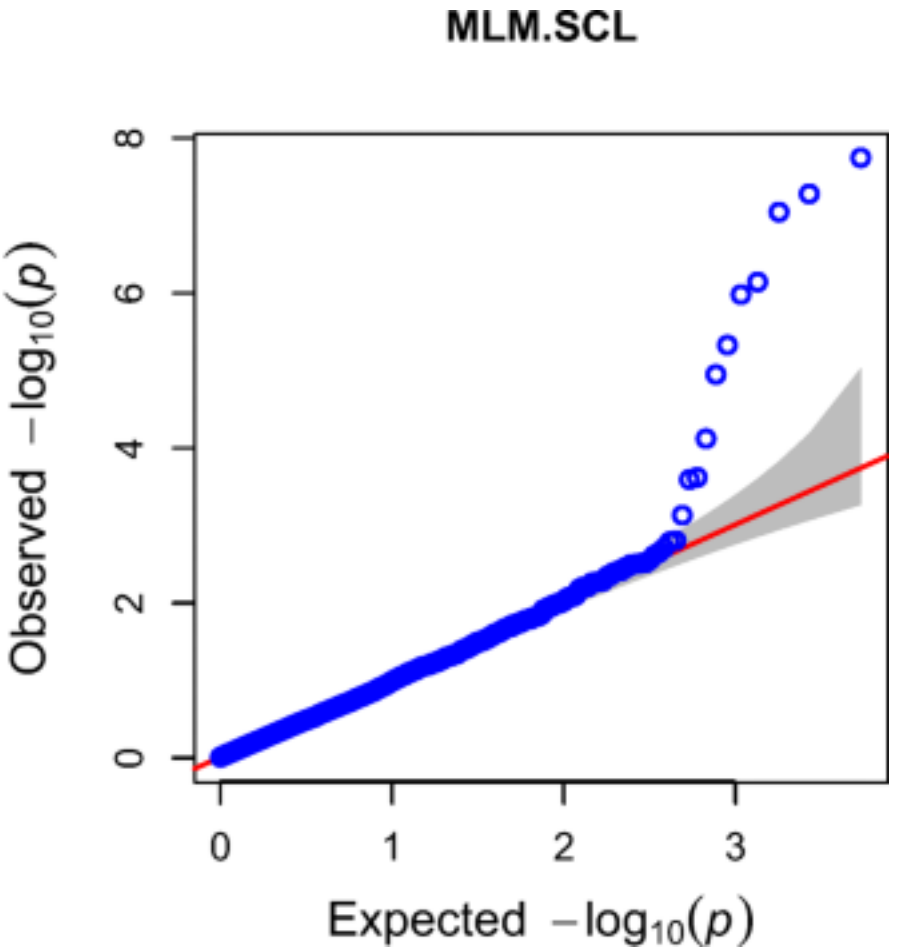
VISUAL DISPLAYS

Quantile-Quantile (QQ) Plots

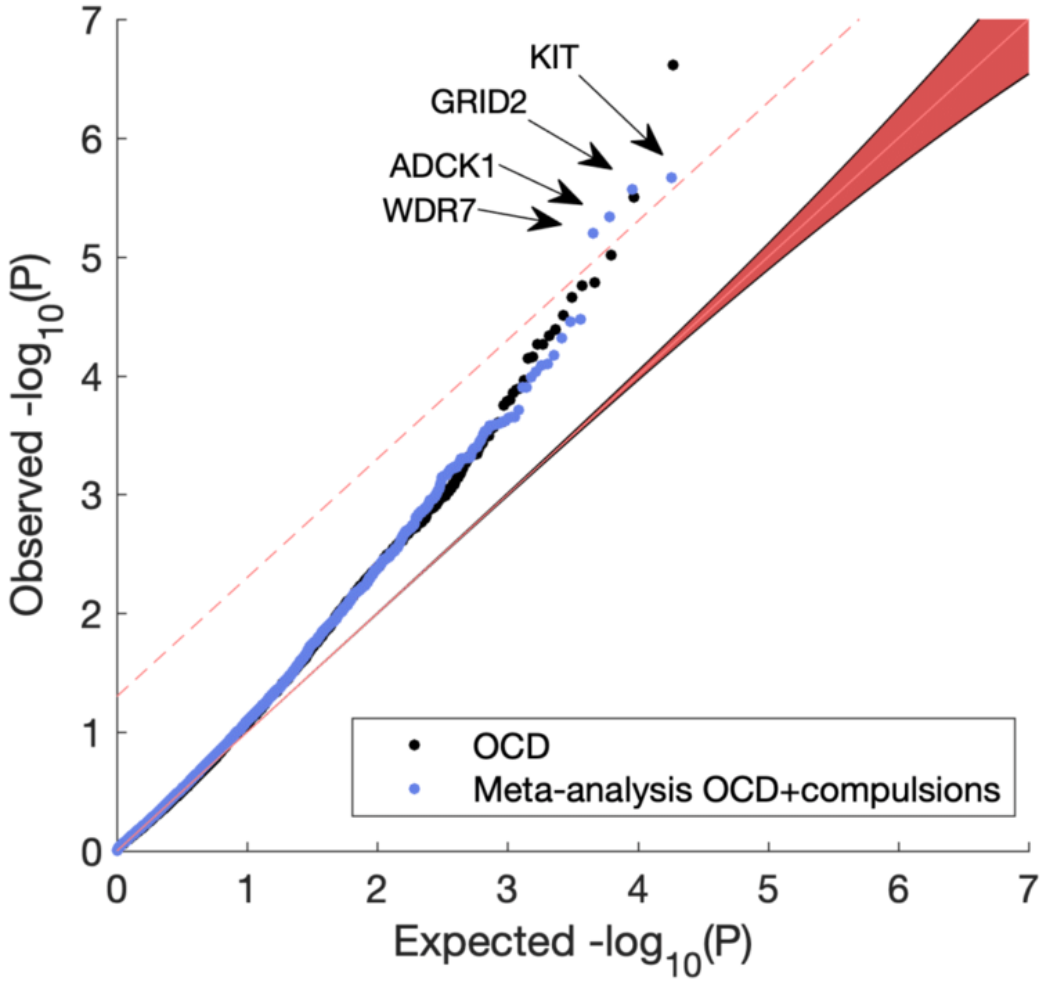
- QQ plots are effective for visualizing the overall pattern of p-values from a GWAS.
- Plots observed $-\log_{10}(p)$, ranked in magnitude, against their expected values according to the null hypothesis (i.e. p-values having uniform distribution between 0 and 1)
- If the null hypothesis is true for all SNPs and the tests are behaving appropriately then the plot should follow a straight line at 45° from the origin
- If the plot exceeds the null line by more than $\log_{10}(20)=1.3$ at some SNP, then this SNP and SNPs with smaller p-values are significant at FDR<1/20.
- Deviation from the null line suggests
  - Inflated false positive rate, if the deviation starts from the origin
  - True associations, if the plot at first conforms to the null line and only deviates when the expected $-\log_{10}(p)$ reaches at least 1 (although with increasing sample size, deviation from the null line may begin earlier)

Good and bad QQ plots





QQ plot and FDR

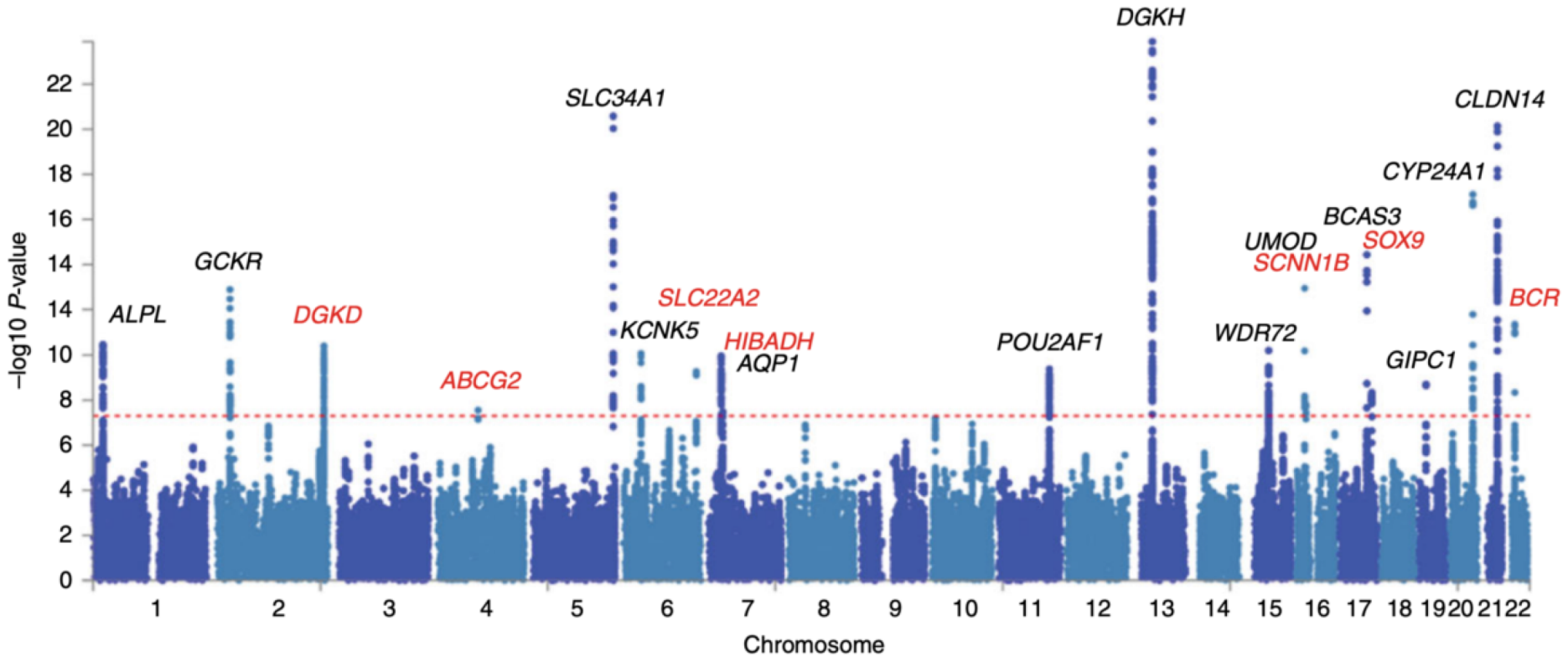


The SNPs beyond the point where the plot crossed the line y=x+1.3 are significant at FDR of 0.05

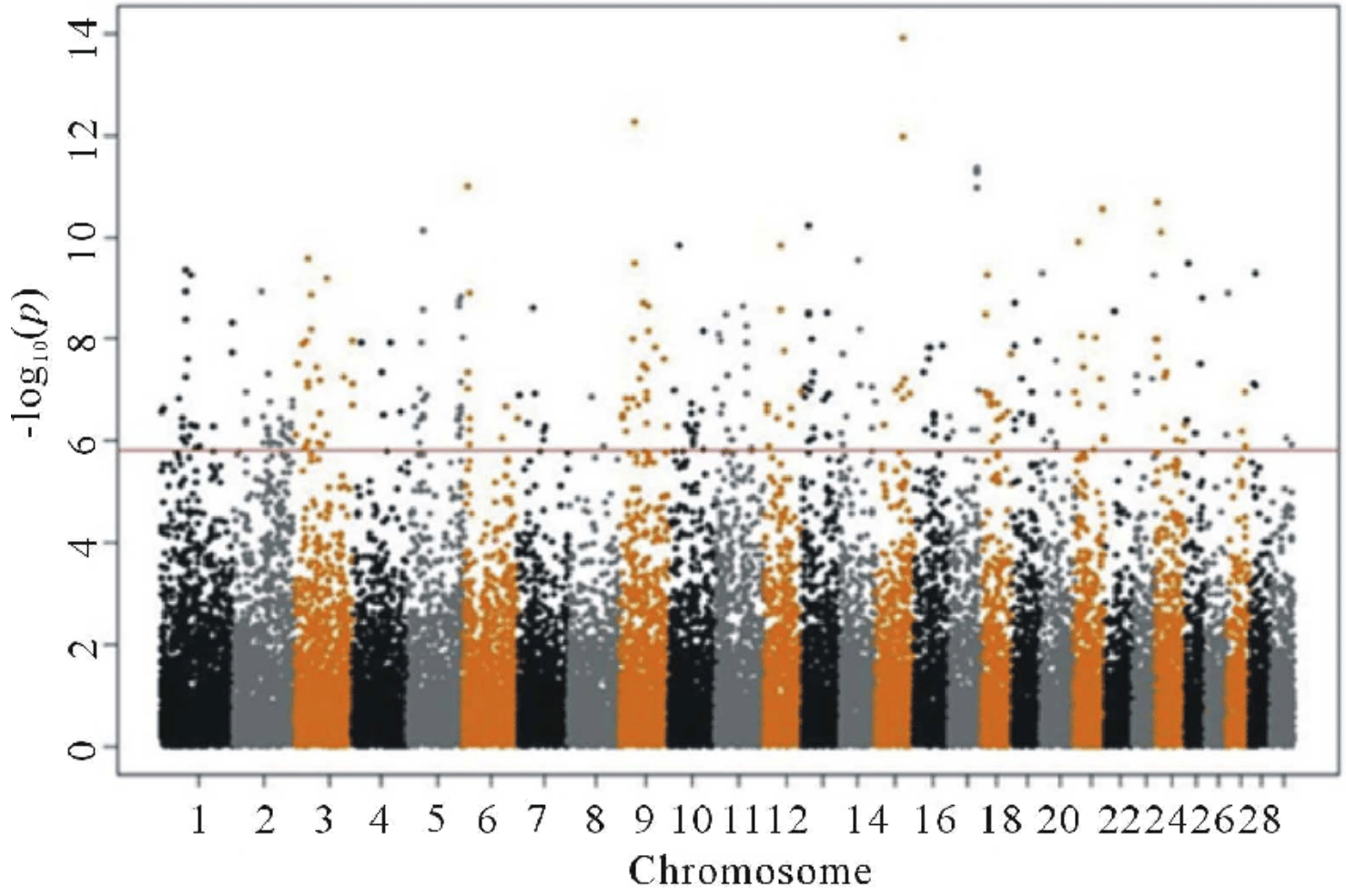
Manhattan and locus zoom plots

- Plots $-\log_{10}(p)$ values of SNPs against their genomic locations
- **Manhattan plot**
  - Whole-genome scale, by chromosome
  - True association signals usually form tight clusters of significant SNPs LD, appearing as thin columns of points. Some breaking through the genome-wide significance threshold
  - Random scattering of significant SNPs which do not cluster in thin columns is suggestive of genotyping errors or other technical problems.
- **Locus zoom plot**
  - Zoom into a region containing a significant association signal
  - Help to correlate the positions of significant SNPs with the positions of protein-coding genes and other DNA elements.

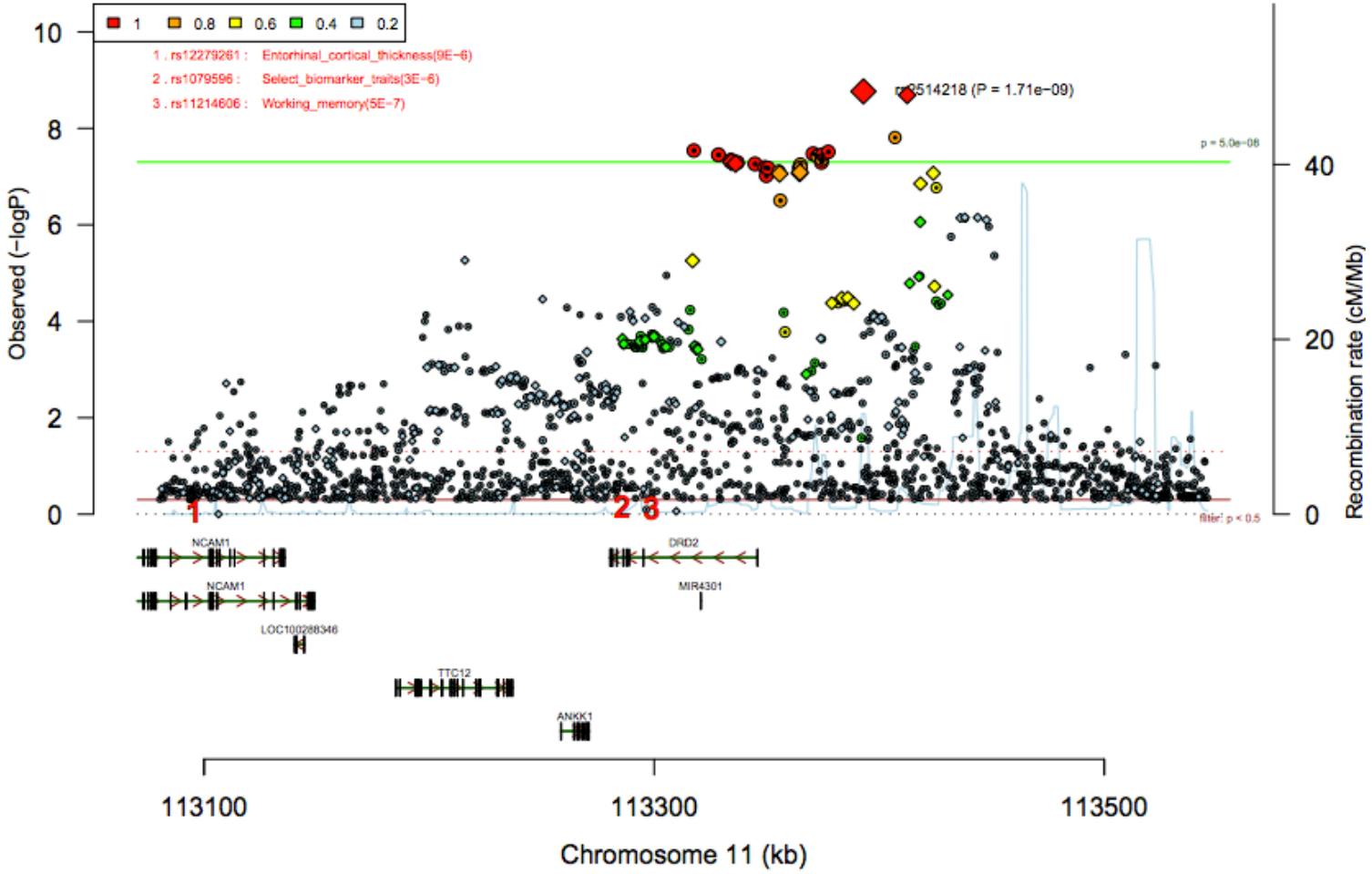
Good Manhattan plot



Poor Manhattan plot



Locus zoom plot



POWER CALCULATION

Statistical power

- Classical hypothesis testing requires only the null hypothesis to be clearly defined.
- A clearly defined alternative hypothesis was later introduced, to calculate the probability of a type 2 error (not rejecting the null hypothesis when the alternative hypothesis is true).
- Statistical power is the probability of rejecting the null hypothesis under an assumed alternative hypothesis (1 – type 2 error probability)

Power calculation

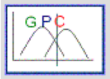
- Power calculation involves obtaining the distribution of the test statistic under the alternative hypothesis (including the assumed effect size), for a specified study design, statistical test and sample size.
- For a chi-squared test, the distribution of the test statistic under an alternative hypothesis can usually be approximated by a non-central chi-squared distribution, which is characterized by a non-centrality parameter (NCP).
- Analytic power usually proceeds by first calculating NCP for a given sample size and effect size. Knowing the NCP will then allow power to be calculated for any specified significance threshold (such as $5x10^{-8}$). It is useful to note that NCP is often directly proportion to the sample size and to the square of the effect size.
- In some complicated scenarios power calculation requires computer simulation.

Analytical power calculation

Genetic Power Calculator (GPC)

http://zzz.bwh.harvard.edu/gpc/

- Calculates statistical power for association analysis of discrete traits (case-control and case-parents) and continuous traits (singletons and sibships)
- Interactive input of sample size and assumed parameter values under alternative hypothesis (e.g. effect size, allele frequencies, linkage disequilibrium)



Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships, and other common tests. Suggestions, comments, etc to *Shaun Purcell*.

If you use this site, please reference the following Bioinformatics article:

```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

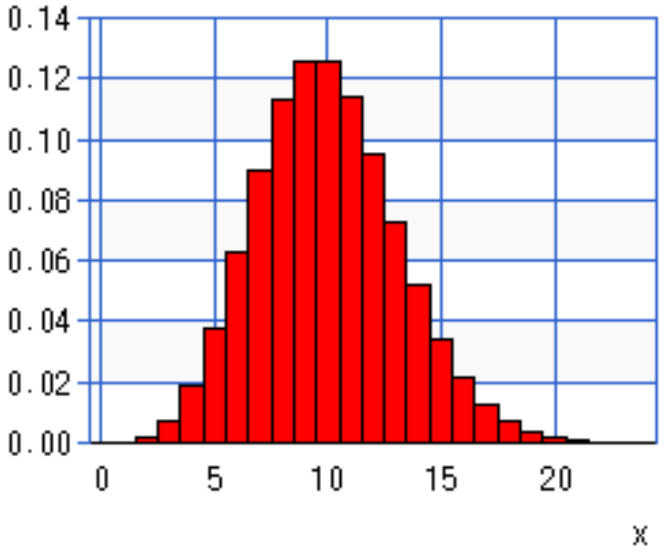
Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Instructions for power calculations

VC model calculations are based upon formula derived in Sham et al (2000) [AJHG, 66, 1616-1630]. Users of this site who are unsure of the nature of the VC tests and power calculations are **strongly** advised to consult this article.

Power under polygenicity

- Many SNPs contribute to complex traits
- A GWAS has multiple chances of detecting true associations
- Suppose a trait has 1,000 independent causal SNPs, and a study has only 1% power to detect each of these SNPs.
- The number of significant causal SNPs follows a binomial distribution with n=1,000 and p=0.01



- Study likely to detect 3 to 23 causal SNPs.
- These SNPs are no different from the other causal SNPs.
- Power of independent replication of each SNP is only 1%, with same sample size and p-value threshold

How to increase power

- Increase sample size
- Improve accuracy of trait measurement
- Repeated measures (average out fluctuations)
- Reduce residual variation (e.g. age, sex)
- Joint analysis of multiple correlated phenotypes
- Select subjects at either extremes of trait values
- Increase SNP density (greater LD, improved imputation)
- Consider each p-value in relation to overall distribution of p-values - False Discovery Rate (FDR)
- Stratify SNPs into functional classes and perform separate FDR on each class

THANK YOU