

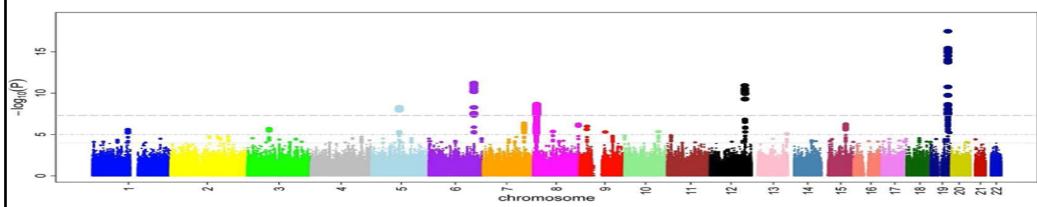
Complex Trait Genomic Analysis



Miguel Pérez-Enciso

(miguel.perez@uab.es)

Genome Wide Association Studies



Outline

1. Why disequilibrium?
2. What could go wrong? (almost everything)
3. Testing: False Discovery Rate (FDR)
4. Population Structure
5. Mixed Models

The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Geneticists have made substantial progress in identifying the genetic basis of many human diseases, at least those with conspicuous determinants. These successes include Huntington's disease, Alzheimer's disease, and some forms of breast cancer. However, the detection of genetic factors for complex diseases—such as schizophrenia, bipolar disorder, and diabetes—has been far more complicated. There have been numerous reports of genes or loci that might underlie these disorders, but few of these findings have been replicated. The modest nature of the gene effects for these disorders likely explains the contradictory and inconclusive claims about their identification. Despite the small effects of such genes, the magnitude of their attributable risk (the proportion of people affected due to them) may be large because they are quite frequent in the population, making them of public health significance.

age analysis we have chosen for this argument is a popular current paradigm in which pairs of siblings, both with the disease, are examined for sharing of alleles at multiple sites in the genome defined by genetic markers. The more often the affected siblings share the same allele at a particular site, the more likely the site is close to the disease gene. Using the formulas in (1), we calculate the expected proportion Y of alleles shared by a pair of affected siblings for the best possible case—that is, a closely linked marker locus (recombination fraction $\theta = 0$) that is fully informative (heterozygosity = 1) (2)—as

$$Y = \frac{1+w}{2+w} \text{ where } w = \frac{pq(\gamma-1)^2}{(p\gamma+q)^2}$$

If there is no linkage of a marker at a particular site to the disease, the siblings

linkage analysis for loci conferring GRR of about 2 or less will never allow identification because the number of families required (more than ~2500) is not practically achievable.

Although tests of linkage for genes of modest effect are of low power, as shown by the above example, direct tests of association with a disease locus itself can still be quite strong. To illustrate this point, we use the transmission/disequilibrium test of Spielman *et al.* (3). In this test, transmission of a particular allele at a locus from heterozygous parents to their affected offspring is examined. Under Mendelian inheritance, all alleles should have a 50% chance of being transmitted to the next generation. In contrast, if one of the alleles is associated with disease risk, it will be transmitted more often than 50% of the time.

For this approach, we do not need families with multiple affected siblings, but can focus just on single affected individuals and their parents. For the same model given above, we can calculate the proportion of heterozygous parents as $pq(\gamma+1)/(p\gamma+q)$ (4). Similarly, the probability for a heterozygote parent to transmit the high risk A allele is just $\gamma/(1+\gamma)$. Association tests can also be performed for pairs of affected siblings. When the locus is associated with disease, the transmission excess

Science, 1996. Cited by 5700

		Locus 2	
		B	b
Locus 1	A	p_{AB}	p_{Ab}
	a	p_{aB}	p_{ab}

haplotype frequencies

Remember:
 $E(p_{AB}) = p_A \cdot p_B$

disequilibrium

$$p_{AB} = p_A \cdot p_B - D$$

$$D = p_{AB} - p_A \cdot p_B$$

allelic frequencies

Disequilibrium measures (i): D'

$$D' = D / D_{MAX} = (p_{AB} - p_A p_B) / D_{MAX}$$

$$D_{MAX} = \min(p_A p_b, p_B p_a) \text{ if } D > 0$$

$$D_{MAX} = \max(p_A p_B, p_a p_b) \text{ if } D < 0$$

- Lewontin (1964).
- $D'=1$ indicates lack of recombination, intermediate values are less easy to interpret.
- Very sensitive to rare alleles.

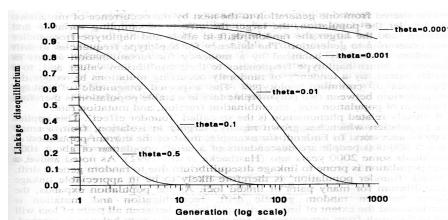
Disequilibrium measures (ii): r^2

$$r^2 = D^2 / p_A p_B p_a p_b$$

- The squared correlation between alleles at both haplotypes.
- It is 1 iff both alleles at the two loci are identical in all haplotypes.
- nr^2 is the power in a Chi-2 test to detect disequilibrium.
- More related to distance than D' .

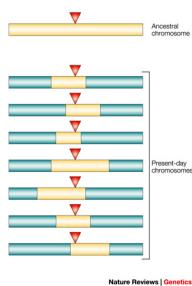
Why disequilibrium?

- Linkage analysis maps the loci with very poor resolution because the number of recombinations analyzed is very low.
- Disequilibrium is rapidly eroded by recombination



Ergo

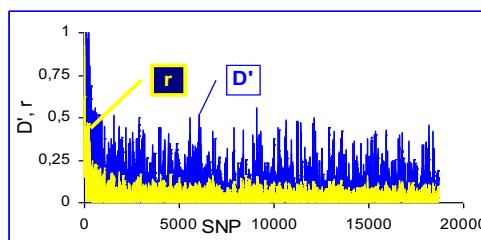
- Two loci that are in strong disequilibrium should be close physically.
- PRINCIPLE OF LD MAPPING: The locus in strongest LD with a given trait should be the causative locus or the closest to the causative locus.



Ardlie et al. 2002

What can go wrong?

- Disequilibrium pattern is highly variable.
- Depends on metrics and on allele frequencies
- LD is affected by all evolutionary forces, including, and most frequently, admixture.



VERY important remarks

- It is the existence of recombination that makes it linkage disequilibrium mapping possible. LD mapping cannot be applied to mtDNA or Y chromosome.
- THERE EXISTS A LARGE STOCHASTIC SAMPLING IN THE AMOUNT OF LINKAGE DISEQUILIBRIUM.

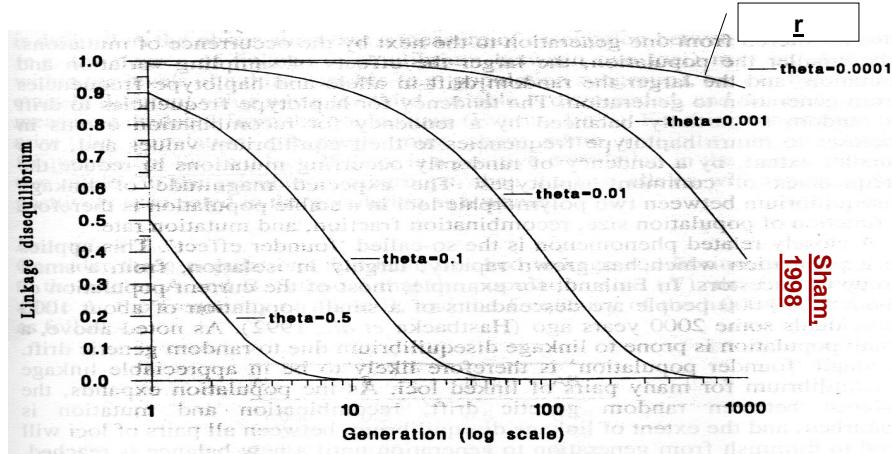
Forces that affect disequilibrium

- Recombination
- Drift
- Mutation
- Admixture
- Selection

Any evolutive force influences LD!

Forces that affect disequilibrium: Recombination

$$D_{t+1} = (1 - r) D_t$$



Measuring association

- Qualitative traits (disease)
 - Chi-² tests
 - Logistic regression
- Quantitative traits
 - Linear models

Measuring association (i) Chi-2 test

	AA	Aa	aa	Sum
Case	n+AA	n+Aa	n+aa	n+
Control	n-AA	n-Aa	n-aa	n-
Sum	nAA	nAa	naa	N

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n_i n_j / N)^2}{n_i n_j / N} \quad 1 \text{ d.f.}$$

Measuring association (ii): logistic regression

- Logistic regression is a way of analyzing count data (e.g., binary).
- Is a class of the so called generalized linear methods.

Typical regression
 $y = Xb + e$

Logistic regression
 $P = \exp(Xb) / [1-\exp(Xb)]$

Measuring association (ii): logistic regression

y	SNP	P(y=Y)	
1	AA	$1 / [1 + \exp(\beta_{AA})]$	$L = \prod P(y_i=Y)$
1	AA	$1 / [1 + \exp(\beta_{AA})]$	
0	aa	$\exp(\beta_{aa}) / [1 + \exp(\beta_{aa})]$	$\exp(Xb)$ is the odds ratio of the effect, or increase in risk, e.g., $\exp(\beta_{AA} - \beta_{aa})$ is the increased risk of AA vs. aa genotypes
0	Aa	$\exp(\beta_{Aa}) / [1 + \exp(\beta_{Aa})]$	
1	aa	$1 / [1 + \exp(\beta_{aa})]$	
0	Aa	$\exp(\beta_{Aa}) / [1 + \exp(\beta_{Aa})]$	

Typical regression

$$y = Xb + e$$

Logistic regression

$$P = 1 / [1 + \exp(Xb)]$$

Measuring association : Linear models

$$y = Xb + e$$

- Each SNP genotype can be coded as class, dominance and additive effects are estimated simultaneously.
- A regression can be carried out on the number of, say, A alleles (0, 1, 2) assuming additivity.

Advantages

- ✓ Easy to implement.
- ✓ No strong assumption about the mode of inheritance.
- ✓ It can be stratified (say by breed).
- ✓ Logistic regression allows for a very flexible modeling strategy.

Disadvantages

- † Sensitive to a number of uncontrolled factors (admixture, random drift). Leads to spurious association.
- † Loss of power is very rapid when the no. of alleles increase.
- † Chi-2 uses very little information (only one marker at a time, no family structure accounted for).

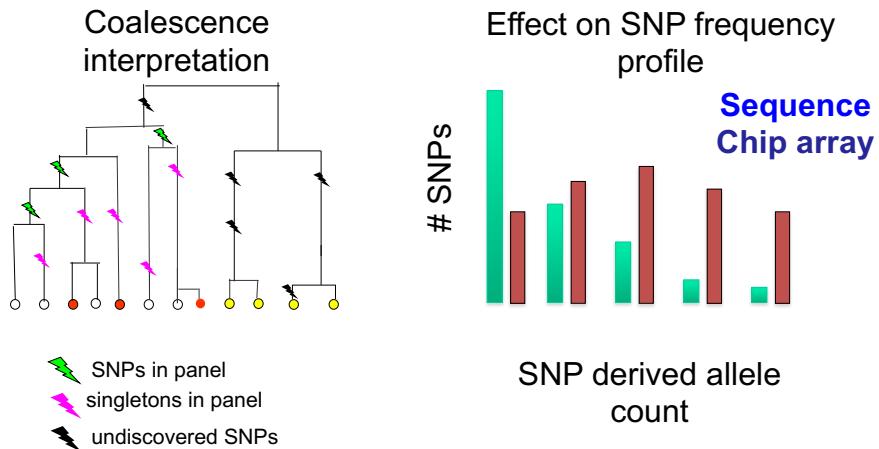
Whole genome association studies (WGAS)

- RATIONALE: Using a sufficiently dense genotyping should allow us to identify markers that are in strong disequilibrium with the causative mutation(s) AND are not false positives.
- Currently, arrays allow genotyping hundred thousand / millions of SNPs and have made feasible these studies.
- A catalogue of GWAS:
<http://www.genome.gov/gwastudies/>
- <http://www.ncbi.nlm.nih.gov/gap>

Common software for GWAS

- Plink: managing data, simple statistics
<https://www.cog-genomics.org/plink2>
- GCTA: compute marker relationship matrix
<http://www.complextraitgenomics.com/software/qcta/download.html>
- SNPassoc: R for association
<http://cran.r-project.org/web/packages/SNPassoc/index.html>
- GenABEL: complex, mixed model modeling
<http://www.genabel.org/packages/GenABEL>

SNPs in arrays are subject to ascertainment bias



How about sequence?

- It avoids SNP ascertainment bias
- The QTNs are in the data
- Usually, used with imputed data
- Main issue are rare variants. Quality control remain a main issue here.

Rare-Variant Association Analysis: Study Designs and Statistical Tests

Seunggeung Lee,¹ Gonçalo R. Abecasis,¹ Michael Boehnke,¹ and Xihong Lin^{2,*}

Author information ► Copyright and License information ►

This article has been cited by other articles in PMC.

Abstract

Go to:

Despite the extensive discovery of trait- and disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants can explain additional disease risk or trait variability. An increasing number of studies are underway to identify trait- and disease-associated rare variants. In this review, we provide an overview of statistical issues in rare-variant association studies with a focus on study designs and statistical tests. We present the design and analysis pipeline of rare-variant studies and review cost-effective sequencing designs and genotyping platforms. We compare various gene- or region-based association tests, including burden tests, variance-component tests, and combined omnibus tests, in terms of their assumptions and performance. Also discussed are the related topics of meta-analysis, population-stratification adjustment, genotype imputation, follow-up studies, and heritability due to rare variants. We provide guidelines for analysis and discuss some of the challenges inherent in these studies and future research directions.

Summary of Statistical Methods for Rare-Variant Association Testing

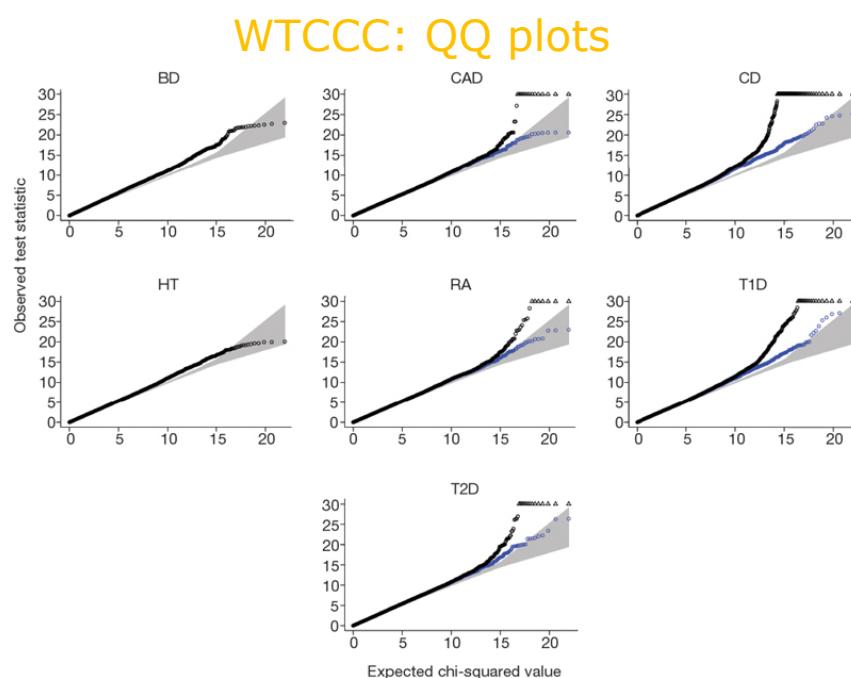
	Description	Methods	Advantage	Disadvantage	Software Packages ^a
Burden tests	collapse rare variants into genetic scores	ARIEL test, ⁵⁰ CAST, ⁵¹ CMC method, ⁵² MZ test, ⁵³ WSS ⁵⁴	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum ⁵⁵ , Step-up, ⁵⁶ EREC test, ⁵⁷ VT, ⁵⁸ KBAC method, ⁵⁹ RBT ⁶⁰	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT ⁶¹ , SSU test, ⁶² C-alpha test ⁶³	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O, ⁶⁴ Fisher method, ⁶⁵ MiST ⁶⁶	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test ⁶⁷	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

The Wellcome Trust Case Control Consortium (WTCCC)

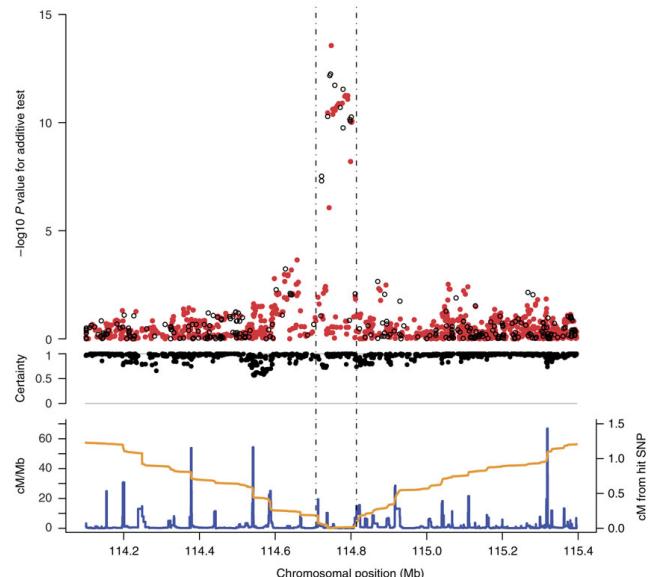
- 3000 control in one group of blood donors and a control group.
- 2000 cases in 7 common diseases: Bipolar disorder, Coronary artery disease, Chron's disease, Hypertension, Rheumatoid arthritis, Type I diabetes, Type II diabetes.
- UK wide caucasian origin.
- 50 research groups.
- 500k SNP chip genotyped.
- Data publicly available under some restrictions.

<http://www.wtccc.org.uk/>

Superseded by
larger initiatives
such as
UK biobank



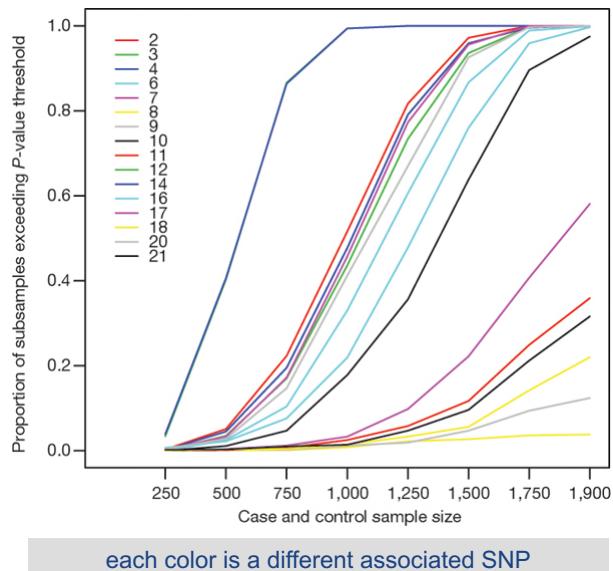
WTCCC: How the results look like



WTCCC: General results

- The increased associated risk is rather small (odds-ratio), in the order 1.2 - 1.5
- 25 regions with $P < 10^{-7}$
 - 12 previously known
 - 10 have been confirmed
 - 1 gave equivocal results
 - 5 more have been replicated

WTCCC: Influence of sample size



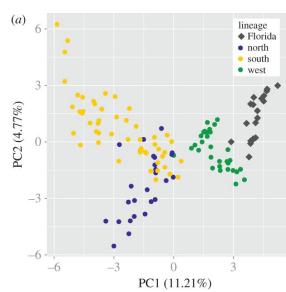
For the 16 SNPs in Table 3 with P values below 5×10^{-7} , we randomly generated 1,000 subsets of our full data set corresponding to case-control studies with different numbers of cases, and the same number of controls (x axis). The y axis gives the proportion of subsamples in which that SNP achieved a P value below 5×10^{-7} . SNPs are numbered according to the row in which they occur in Table 3

Warning!!!!

One of the main risks of GWAS is uncorrected structure in the data, since admixture inflates false positive rate!

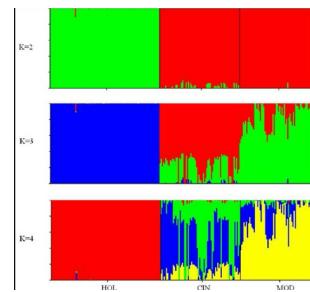
Detecting structure

Principal component analyses



- Compute % of variance explained by each eigenvalue
- PCA is highly sensitive to uneven sampling / population

STRUCTURE software



- Estimate optimum number of clusters

Correcting structure

- Principal components as cofactors
- Mixed models with genomic relationship

Principal Components

Since PCA reflects structure, the idea is to include the first eigenvectors as fixed effects in the model.

$$y = \mathbf{X}\beta + \Sigma \mathbf{q}' \mathbf{m} + \lambda \mathbf{g} + e$$

$$\mathbf{G} = \mathbf{M}\mathbf{M}'; \quad \mathbf{G} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$$

Price, A. L. et al (2006). Nature Genetics, 38, 904–909.

Mixed Models

It is a generalization of the previous approach, where all PC are considered. G is the genomic relationship matrix.

$$y = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \lambda \mathbf{g} + e$$

$$\mathbf{u} \sim N(0, \mathbf{G})$$

$$\mathbf{G} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$$