Genome-wide association studies: methods and applications

Ksenia Krasheninnikova

St.Petersburgh University of the Russian Academy of Sciences

November 24, 2012

Agenda

- What is GWAS?
- Methods
- Statistical Approaches to Data Processing
- GWAS Projects at Theodosius Dobzhansky Center for Genome BioInformatics (St.Petersburg)

What is GWAS?

- GWAS - Genome-Wide Association Studies
- Goal: to reveal the risk alleles for genetically complex disorders
- Idea: to compare two groups: people with the disease (cases) and similar people without (controls)
- The first successful GWA study was published in 2005

GWAS - essentials

- Large study samples (f.e. WTCC (2007): 14,000 cases - 3,000 controls)
- Polymorphic alleles which cover the genome adequately
- Statistically powerful analytic methods for data retrieval
  - prior data: HapMap Project
    http://hapmap.ncbi.nlm.nih.gov

Example Project

- Paper: Genome-wide association study of 14,000 cases of seven common deseases and 3,000 shared controls, June 2007
- Who: Wellcome Trust Case Control Consortium
- View: To explore the utility, design and analyses of GWA studies
- Experiment: study over 7 deseases:
  - bipolar disorder (BD)
  - coronary artery desease (CAD)
  - Crohn's desease (CD)
  - hypertension (HT)
  - rheumatoid arthritis (RA)
  - type 1 diabetes (T1D)
  - type 2 diabetes (T2D)

Obtaining Study Samples

- Cases and Controls
- "Sufficiently large" - 2,000 individuals for each desease and 3,000 combined controls (genotyoed with Affymetrics Chip)
- Two control groups
  - 1,500 from the British Cohort (58C)
  - 1,500 - blood donors in the WTCCC project

Two controls

- To assess bias in ascertaining control samples
- Check for effects of differential genotyping errors

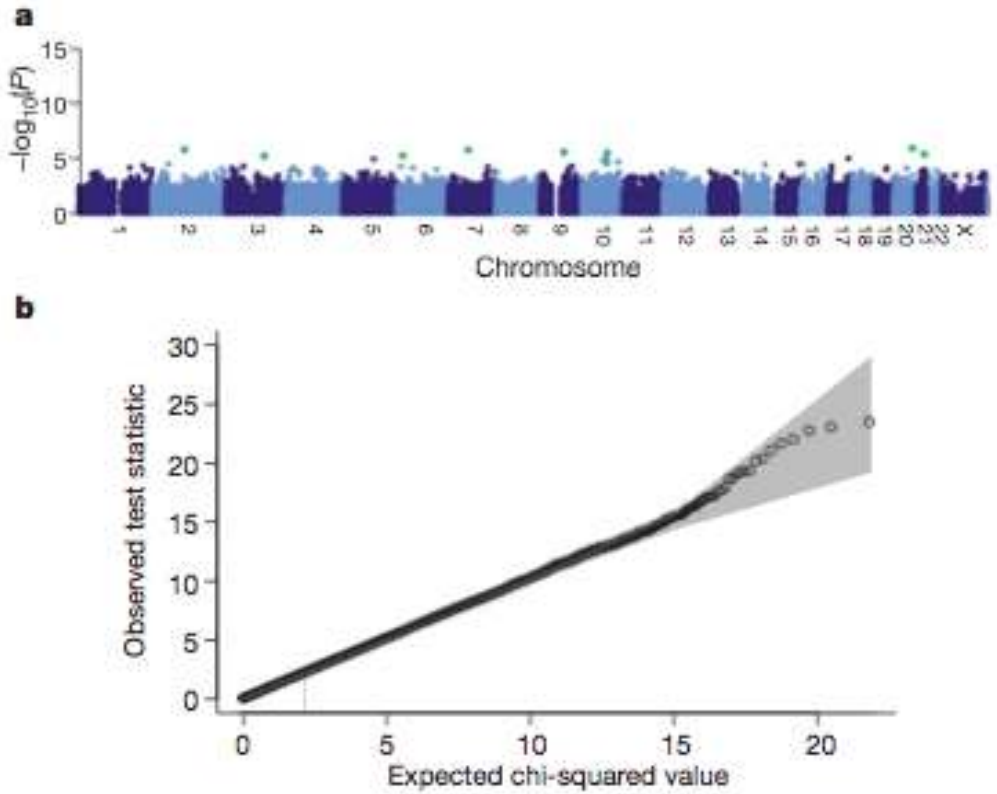


Figure : Genome-wide scan for allele frequency differences between controls.

Linkage Disequillibrium (LD)

- LD occurs when genotypes at the two loci are not independent.
- However, it is very likely that the marker investigated (in this case the SNP) is in LD with the genes causing the disease - *statistical* point of view.
- GWAS is searching for another kind of LD: which arises from *physical* linkage, where the marker is found at a chromosomal locus that is near the genetic difference actually causing the disease.

Hidden Population Structure

- Hidden population structure
  - British population: several waves of immigration from southern and northern Europe
- Exclude 153 individuals with non-european ancestry
- Study population heterogenity with statistical tests of independence
- Is it a bottleneck?

Bottlenecks - Hidden Population Structure

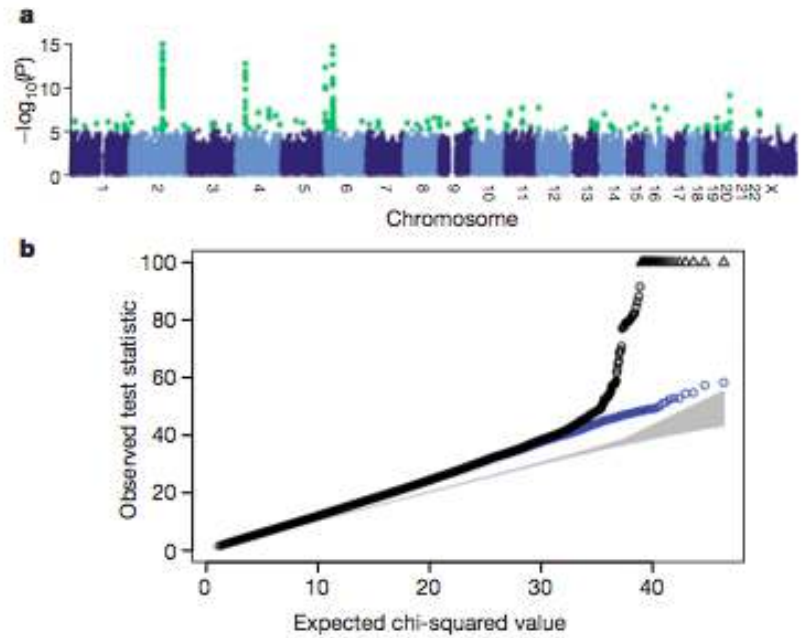
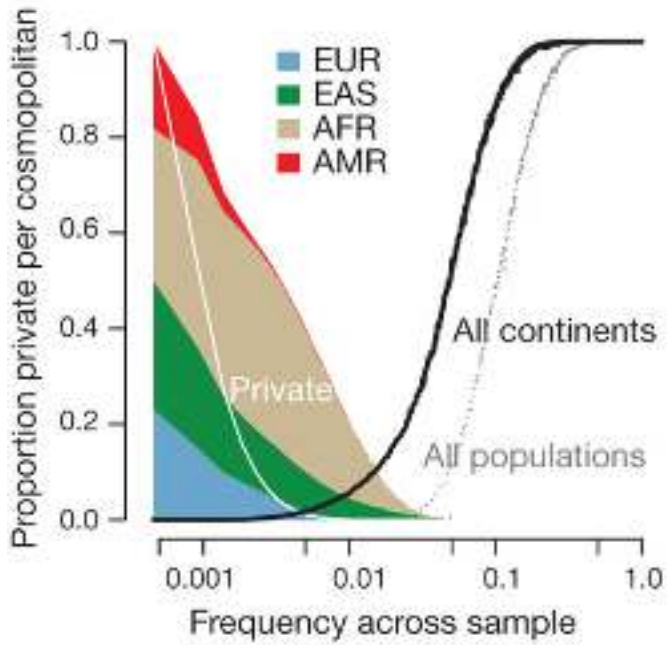


Figure : Genome-wide picture of geographic variation.

- Wide-spread small differences in allele frequencies are evident as an increased slope of the line (Fig. 2b);
- A few loci show much larger differences (Fig. 2a)

Hidden Population Structure - Regions of Variance

- LCT (lactase)
- Majorhistocompatibility complex
- within-UK differentiation at 4p14
  - British population: several waves of immigration from southern and northern Europe
- exclude 153 individuals with non-european ancestry



Data Analysis Approaches

- Tests for independence
  - trend test for independence of alleles frequencies
  - trend test aka genotypic test - independence of genotypes frequencies
- Bayesian hierarchical modelling
  - incorporation of other results
- Linear regression, logistic regression
  - search for epistasis within a single GWAS study
- Pathways
  - accumulation of the effects of genetic variants

Statistical Approaches Overview - Trend Test

	B = 1	B = 2	B = 3	Sum
A = 1	N_{11}	N_{12}	N_{13}	R_1
A = 2	N_{21}	N_{22}	N_{23}	R_2
Sum	C_1	C_2	C_3	N

Figure : Contingency table

- trend statistic: $T = \sum_{i}^{k} t_i(N_{1i}R_2 - N_{2i}R_1)$
- Pearson chi-squared test vs.chi-squared test: orders the effect of B
- Null hypothesis: $P(A = 1|B = 1) == P(A = 1|B = k)$

Trend Test - Example

	Genotype aa	Genotype Aa	Genotype AA	Sum
Controls	20	20	20	60
Cases	10	20	30	60
Sum	30	40	50	120

Figure : Example of genotypic trend test

- t = (0, 1, 1): to test whether allele a is recessive to allele A

Data Analysis of WTCC Paper - Trend Test

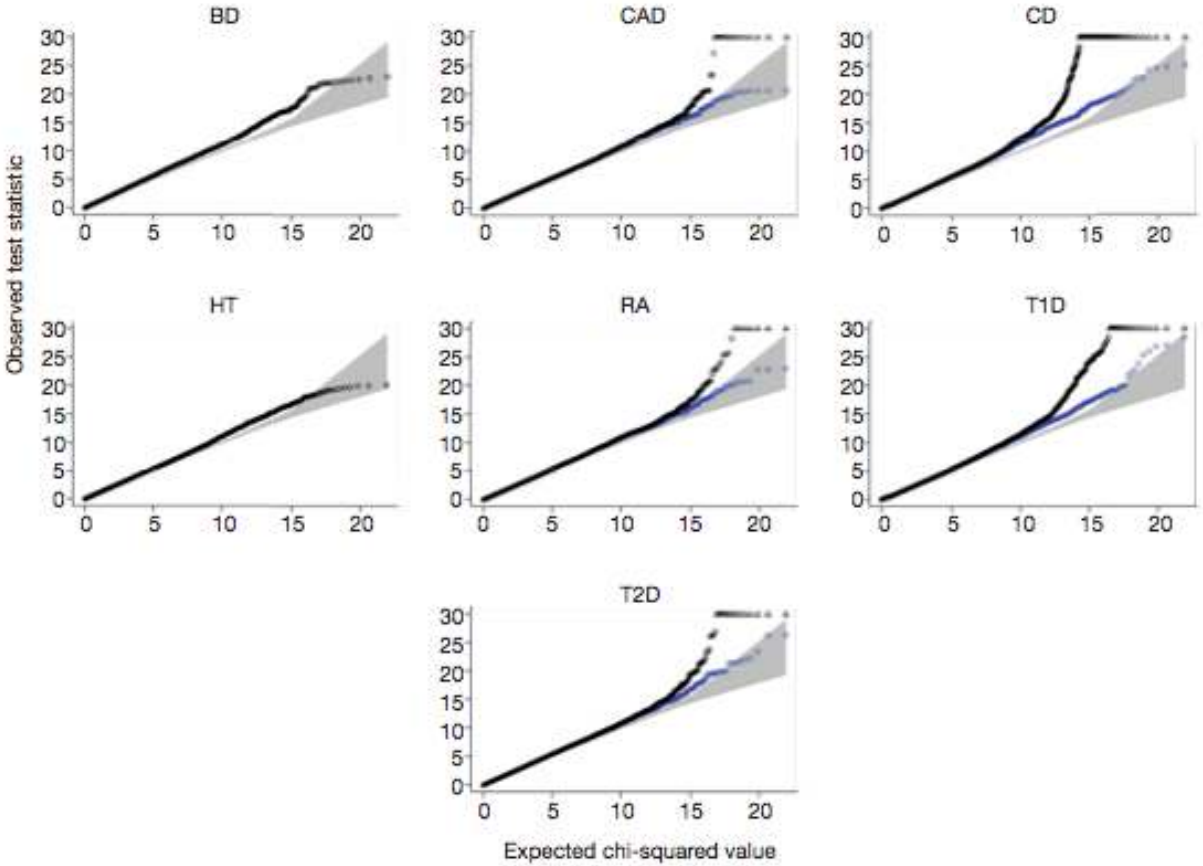


Figure : Quantile-quantile plots for seven genome-wide scans.

Statistical Approaches Overview - Bayesian Hierarchical Models

- Goal: posterior probability p(SNP,Disorder|Data) of association for each SNP
- To estimate prior distribuition: HapMap linkage disequillibrium data

Statistical Approaches Overview - Modelling Epistasis

- Epistastatic interaction - a process, where the alleles of one gene influence the effects of alleles of another on a trait value or risk of disease
- Build a general linear model: $\mu_i = \delta + \beta_1 x_i + \beta_2 y_i + \gamma x_i y_i$
  - i denotes individual, x_i and y_i - predictors, β_1 and β_2 are their main effect sizes, δ is the intercept, and γ is the effect size for the interaction.
- Turn to the logistic regression (estimating odds): $\frac{P(z_i=1|g_i)}{P(z_i=0|g_i)} = e^{\mu_i}$
- Single SNP: $\mu_i = 1_{\{g_i=1/1\}}\alpha_1 + 1_{\{g_i=1/2\}}\alpha_2 + 1_{\{g_i=2/2\}}\alpha_3$

Statistical Approaches Overview - Modelling Epistasis (2)

- Full epistatic model for two SNPs:
  $\mu_i = 1_{\{g_i=1/1,h_i=1/1\}}\gamma_1 + 1_{\{g_i=1/1,h_i=1/2\}}\gamma_2 + 1_{\{g_i=1/1,h_i=2/2\}}\gamma_3 +$
  $1_{\{g_i=1/2,h_i=1/1\}}\gamma_4 + 1_{\{g_i=1/2,h_i=1/2\}}\gamma_5 + 1_{\{g_i=1/2,h_i=2/2\}}\gamma_6 +$
  $1_{\{g_i=2/2,h_i=1/1\}}\gamma_7 + 1_{\{g_i=2/2,h_i=1/2\}}\gamma_8 + 1_{\{g_i=2/2,h_i=2/2\}}\gamma_9$
- Additive model - effects of loci are independent:
  $\mu_i = 1_{\{g_i=1/1\}}\alpha_1 + 1_{\{g_i=1/2\}}\alpha_2 + 1_{\{g_i=2/2\}}\alpha_3 + 1_{\{g_i=1/1\}}\beta_1 + 1_{\{g_i=1/2\}}\beta_2$
- F-test: ANOVA - if means are equal
- Note: The assessment of all pairwise or higher-order interactions is computationally prohibitive

Statistical Approaches Overview - GWAS Pathways Analysis

The genes that coordinate to achieve a specific task are grouped together in the same pathway.

1. Select one or more pathways for the GWASPA (prior hypothesis or results of GWAS)
2. Select a database (Pathguide, Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO)) to delinate the genes in the pathway
3. Assign the GWAS SNPs to known genes within the selected pathway, as given in the selected database
4. Pathway scoring system - bias in distribution of SNPs
5. Statistical approach to aggregate the results

Results: pathways in Parkinson's disease, bipolar disorder

Ins and Outs according to WTCC

- Hypothesis-free - looks at very many SNPs simultaneously rather than focusing on a set of suggested loci
- If there are genetic loci influencing the trait where the rare allele has a frequency under 5%, or even under 1%, the GWAS technique is unlikely to be able to detect these loci.

GWAS Projects at Theodosius Dobzhansky Center for Genome BioInformatics

- "CCR5-32 genotyping of the SP Injection Drug Users cohort"
- "GWAS from the Botswana Harvard Partnership cohort" (Andrey Shevchenko)
  - Idea: to investigate the influence of host genotypic variants in selected genes (AIDS Restriction Genes (ARGs)) on the susceptibility to HIV
- GWAS data representation (Gaik Tamazian)

The End