Iceland Liechtenstein Norway grants

Introduction to Genome Wide Association Studies

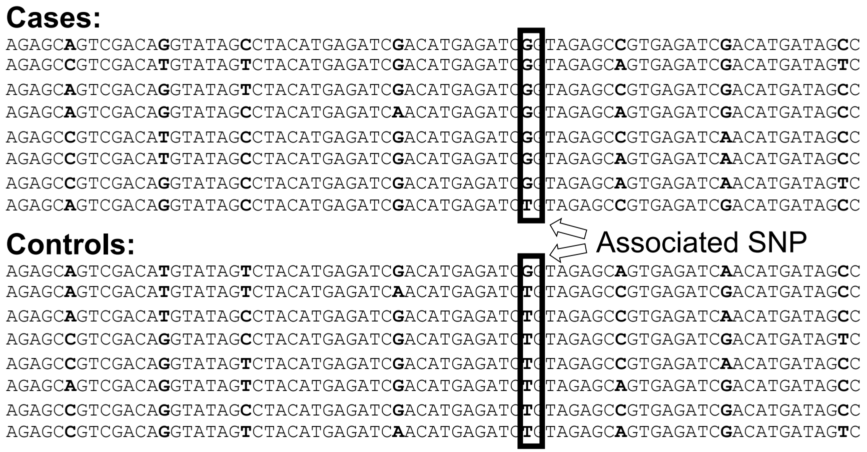
Genome wide association studies

- Goal: find connections between:
  - A phenotype: height, type-I diabetes, etc., known to be heritable
  - Whole-genome genotype
- Specific goals are distinct:
  1. **Identify statistical connections between points (or areas) in the genome and the phenotype**
     - **Drive hypotheses for biological studies of specific genes/regions in specific context**
  2. Generate insights on genetic architecture of phenotype
     - Many small genetic effects dispersed across the genome?
     - Few large effects concentrated in one area (MHC?)
  3. Build statistical models to predict phenotype from genotype
     - “Show me your genome and I will tell you what diseases you will get”

Methodology

- Collect *n* subjects with known phenotype (usually *n* in range 10^3-10^4)
- Measure each one in *m* genomic locations (“representing **common** variation in the whole genome”)
  - Usually SNPs: Single Nucleotide Polymorphisms
  - Typically *m* in range 10^5-10^6
  - Recently moving to whole genome sequencing ($m = 3*10^9$ but realistically same information)
- Now we can think of our data as X_{n*m} matrix with subjects as rows, SNPs as columns,
  - X_{ij} is in {0,1,2} (genotype at single locus)
  - Also given extra vector Y_n of phenotypes
- Our first task: association testing
  - Find SNPs (columns in X) that are statistically associated with Y
  - Can be thought of as *m* separate statistical tests run on this matrix

Can you find the associated SNP?



Cases:

AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

AGAGC**C**GTCGACA**T**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**G**ACATGATAG**T**C

AGAGC**A**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**A**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**G**ACATGATAG**C**C

AGAGC**C**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C

AGAGC**C**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**C**C

AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**T**C

AGAGC**A**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

Controls:

AGAGC**A**GTCGACA**T**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**G**GTAGAGC**A**GTGAGATC**A**ACATGATAG**C**C

AGAGC**A**GTCGACA**T**GTATAG**T**CTACATGAGATC**A**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

AGAGC**A**GTCGACA**T**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C

AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**T**C

AGAGC**C**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**A**ACATGATAG**C**C

AGAGC**A**GTCGACA**G**GTATAG**T**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**A**GTGAGATC**G**ACATGATAG**C**C

AGAGC**C**GTCGACA**G**GTATAG**C**CTACATGAGATC**G**ACATGAGATC**T**GTAGAGC**C**GTGAGATC**G**ACATGATAG**C**C

AGAGC**C**GTCGACA**G**GTATAG**T**CTACATGAGATC**A**ACATGAGATC**T**GTAGAGC**A**GTGAGATC**G**ACATGATAG**T**C

Disease association analysis of a single SNP

	Genotype 0	Genotype 1	Genotype 2	Total
Y=0 (healthy)	N_{00}	N_{01}	N_{02}	N_0
Y=1 (sick)	N_{10}	N_{11}	N_{12}	N_1
Total	M_0	M_1	M_2	n

ow our problem is one of testing:

$_0$: No connection between disease and SNP ⇔ the rows and columns of the table are independent

bvious approach: χ^2 test for 3x2 table (2-df)

ther alternatives: logistic regression, trend test,… (dealing with genotype as numeric)

his approach generates *m* (≈10^6) total hypotheses tests and p values

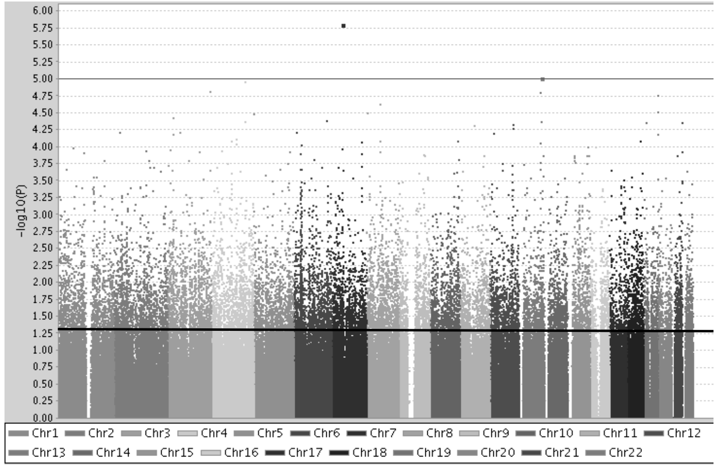
“Manhattan plot” of GWAS results

at happens if we use a p-value
shold of α=0.05 (black line) to
lare results as significant?

would get about 10^6x0.05 =
false discoveries

ution: be very selective in what
ults we declare as significant.
his plot the threshold is the
nge line at $\alpha=10^{-5}$

Declaring only one association
hr7



The multiplicity problem in GWAS

What is a statistically sound choice of a threshold for declaring an association?

•Family wise error rate (FWER): the probability of making even one false discovery out of our *m* tests

•Controlling FWER: the well known Bonferroni correction, perform each test at level *α = 0.05/m*

- For *m* = 10^6 this gives *α* = 5 x 10^{-8}

•Leading journals (Nature Genetics) require a p value smaller than 5 x 10^{-8} to publish GWAS results

- Implicitly require Bonferroni for 10^6 – super conservative!
- Lesson learned in blood, from findings that did not replicate and were eventually deemed false!

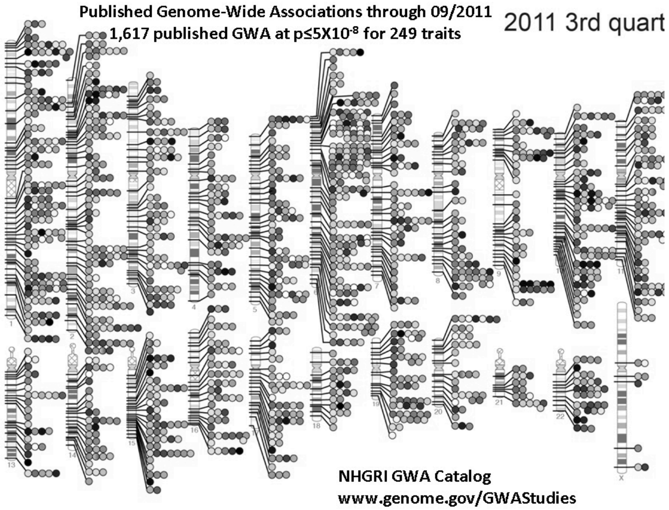
GWAS promise and history

- We know of many highly heritable traits and diseases including
  - Height
  - Heart Disease
  - Many cancers
- The GWAS promise: we will identify the genetic basis for this heritability
- First GWAS in 2005, since then:
  Thousands of studies, hundreds of thousands of individuals, hundreds of billions of SNPs genotyped, many billions of $$$ invested
- Was the promise fulfilled?

Yes: we found a lot of associations, learned some biology

Lessons learned:

- A few of strongest associations are in coding regions
- Most associations are in regulatory elements
- Some are in gene deserts

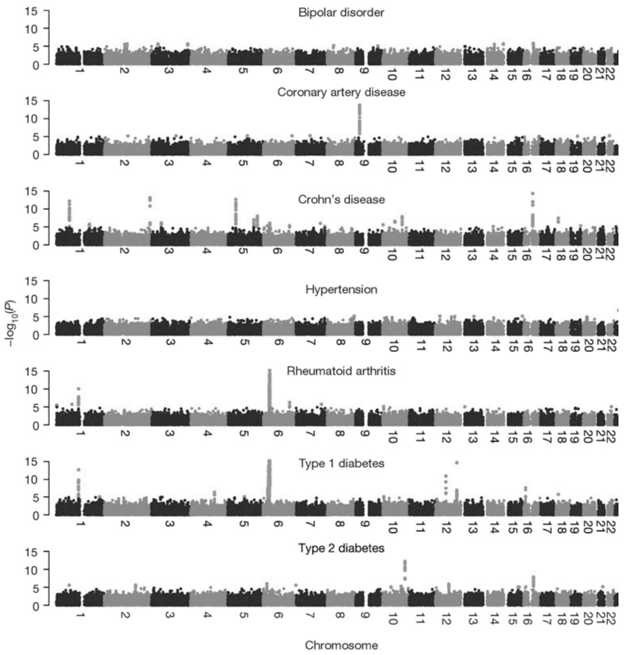


Published Genome-Wide Associations through 09/2011
1,617 published GWA at p≤5X10^{-8} for 249 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies

Results of famous WTCCC study of seven diseases on 14,000 cases and 3,000 shared controls (Nature, 2007)

Total found: 13 significant findings at level 5*10^{-8}

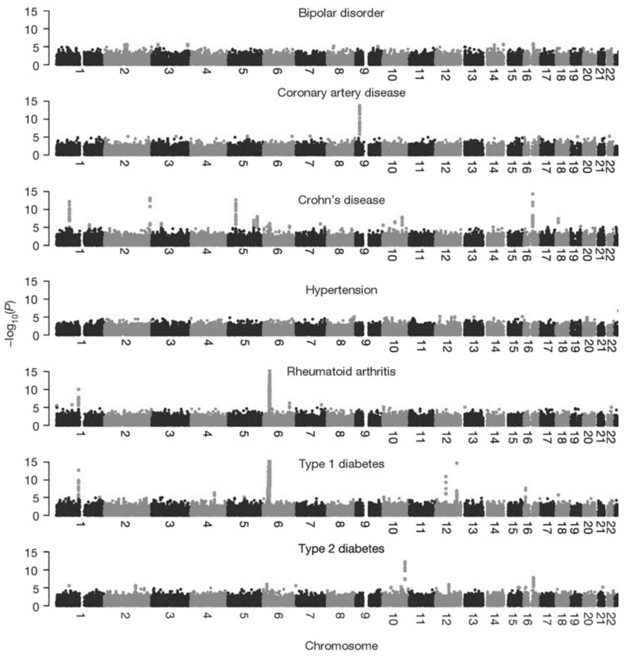


Our GWAS findings do not explain heritability

- Height:
  - From twins and family study, about 80% of height variability is heritable
  - Huge height GWAS (n>40K) found SNPs explaining ~10% of height variability
- Diseases: Schizophrenia, heart disease, cancers,…
  - Heritability: 30%-80%
  - For none of these, GWAS gives more than 5%-10%
- Basically, for all complex traits investigated a major gap remains!

Results of famous WTCCC study of seven diseases on 14,000 cases and 3000 shared controls (Nature, 2007)

Total found: 13 significant findings at level 5*10^{-8}
Heritability explained: small for all except T1D



Where is the missing heritability? Theories:

1. Rare variants not covered by GWAS : Every family has its own mutation
   - We know some examples in cancer (BRCA)
2. Complex associations/epistasis: combinations of SNPs
   - Problem: 10^6 SNPs is 10^{12} pairs
3. Lack of power: the effects are weak, we need much more data
   - Or statistical approaches that aggregate more smartly
4. Epigenetic effects: heritability is not in the genome at all

To some extent, all these theories have been tested, some have provided interesting answers (still hotly debated)

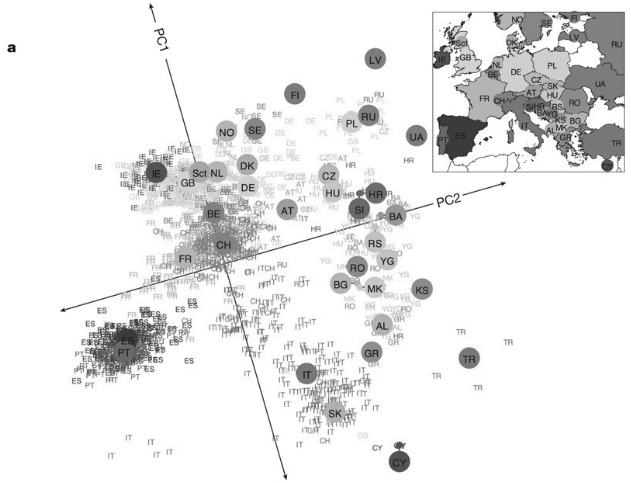
The importance of genetic structure

- Genetic structure: not everyone in the population is from same genetic background
  - Some people are more genetically similar than others
  - Israel: Ashkenazi Jews, Mizachi Jews, Arabs,…
  - US: Caucasian, Black, Hispanic
- Particularly interesting: admixed populations
  - African/Hispanic Americans: mixture of African, European and Native American ancestry
  - Proportions may vary significantly between “African American” individuals
- Many SNPs in the genome have different distribution between Africans and Europeans
  - Most not due to selection/adaptation but due to random drift

Genetic structure and GWAS

- Many traits have strong population association
  - In the US, diabetes much more common among blacks
  - In Israel, Crohn’s disease is much more common among Ashkenazi Jews
- Now, say that we sampled diabetes cases in some hospitals in US + controls in the same hospitals, performed GWAS
  - % of blacks in cases will be higher than in controls (because of high prevalence)
  - What will our GWAS show?
- Every SNP which differs in distribution between Europeans and Africans will be statistically associated with the disease
  - Only because of structure/stratification in our sample!

Even homogeneous population has some structure: Genes mirror geography within Europe



J Novembre *et al.* *Nature* **000**, 1-4 (2008) doi:10.1038/nature07331

Genetic risk prediction from GWAS

- The vision, the doctor will have a “desktop predictor”
  - Input: patient’s genome
  - Output: risk for one (or many) diseases

- Building prediction models is a very different use of GWAS information
  - Non-genetic risk factors that are correlated with the genome (like diet) are also legitimate for prediction
  - Don’t need to name the SNPs that are responsible for risk (⇒ can use structure)
  - Don’t necessarily need a biologist in the loop

- We have accumulating evidence that we may be able to do much better prediction than our identified significant associations only can offer
  - Advanced methods can take advantage of weaker associations, signal from rare variants, environmental effects, etc.