Genome-wide association study (GWAS)

- **What is GWAS?** An approach that involves rapidly scanning markers across a person’s complete set of DNA or genome, to find genetic variations associated with a particular disease.

- **Why is it possible?** Human Genome Project (2003) and International HapMap Project (2007)

- **Why GWAS?**
  - Compared to candidate gene study, GWAS permit a comprehensive scan of the genome in an unbiased fashion and thus have the potential to identify totally novel susceptibility factors.
  - Compared to family based study, GWAS allow the identification of disease genes with only modest increases in risk, a severe limitation in linkage studies.

A genome-wide association study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants
Laura J. Scott, et al.
Science 316, 1341 (2007);
DOI: 10.1126/science.1142382

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

A genome-wide association study identifies novel risk loci for type 2 diabetes

Population resources – trios or case-control samples

Whole-genome genotyping

Genome-wide association

Fine mapping

Gene mining

Gene sequencing & polymorphism identification

Identification of causative SNPs

Pathway analysis & target identification
Basic idea

- Genotyping cases and controls at $10^6$ SNP markers throughout the genome.
- Test associations between the genotypes at each locus and disease status.

0, 1, 2 represent different SNP genotypes, for example, 0=$DD$, 1=$Dd$, 2=$dd$, where $d$ is the disease allele.
Affymetrix GeneChip SNP Genotyping

250 ng Genomic DNA

Xba  Xba  Xba

Restriction enzyme digestion

Adapter ligation

Single Primer PCR Amplification

Fragmentation (DNaseI) and Labeling with a biotinylated nucleotide analogue using terminal deoxynucleotidyl transferase

Hybridization, wash and scan
SNP probe design

5' Genomic Sequence SNP TAG 3'

SNP probe = 25 bases

Perfect Match Mismatch

Allele 'A'

Perfect Match Mismatch

Allele 'B'

Quartet
SNP data scoring approach

Basically a clustering problem

Xiangning Chen et al. Genome Res. 1999; 9: 492-498

Summary of SNP data preprocessing

- **Chip Image → scan to get intensity data**
- **Normalization:**
  - Goal: to make the distribution of the probe intensities for each array in a set of arrays the same.
  - Quantile Normalization (one of many):
    - Motivation: if all \( n \) data vectors have the same distribution, then plotting the quantiles in \( n \) dimensions gives a straight line along the line given by unit vector.
    - By projecting the points of the \( n \) dimensional quantile plot onto the diagonal to make a set of data have the same distribution.

- **Genotype calling algorithm:**
  - **MPAM**: Modified Partitioning Around (Medoids Liu et al., 2003)
    - Only SNPs with 2 or 3 clearly separated clusters are selected by MPAM and the SNPs exhibiting a high degree of misclassification were discarded from the 10 K array.
  - **DM**: dynamic model-based algorithm (Di et al., 2005)
    - Exhibiting a higher degree of misclassification for known heterozygous bases than for known homozygous bases.
  - **BRLMM**: Robust Linear Model with Mahalanobis distance classifier (Rabbee and Speed, 2006).
    - Using a large training sample, but only use the perfect match probe.
  - **SNiPer**: (Huentelman, M.J., 2005)
Coverage of Common Variants by Whole-genome Products

Coverage Mostly Provided by Pairwise Correlations

Affymetrix Mapping 500K GeneChip

Illumina HumanHap300 BeadChip

Adapted from the Candidate Gene Resource Steering Committee Meeting July 25, 2006
Example of coverage of the genome

Adapted from the Candidate Gene Resource Steering Committee Meeting July 25, 2006
New Affy release

- Affymetrix® Genome-Wide Human SNP Array 6.0 (features 1.8 million genetic markers)
- More than 906,600 SNPs:
  - Unbiased selection of 482,000 SNPs;
  - Selection of additional 424,000 SNPs
  - Tag SNPs
  - SNPs from chromosomes X and Y
  - Mitochondrial SNPs
  - New SNPs added to the dbSNP database
  - SNPs in recombination hotspots
- More than 946,000 copy number probes

Can your lab afford it?

For 1,000 cases and 1,000 controls

<table>
<thead>
<tr>
<th>Date</th>
<th>SNPs</th>
<th>Samples</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/15/05</td>
<td>500K</td>
<td>2,000</td>
<td>$3.2M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>($1,600/sample)</td>
</tr>
<tr>
<td>7/25/06</td>
<td>500K</td>
<td>2,000</td>
<td>$1.06M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>($530/sample)</td>
</tr>
<tr>
<td>1/07</td>
<td>500K</td>
<td>2,000</td>
<td>$0.82M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>($410/sample)</td>
</tr>
<tr>
<td>10/08</td>
<td>500K</td>
<td>2,000</td>
<td>$0.5M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>($250/sample)</td>
</tr>
</tbody>
</table>
Challenges with GWAS

- **Assumption:** The linkage disequilibrium across the gene is high enough that even if we do not select the “causal” SNP, we will detect an association in one or several of the nearby SNPs.

- **Coverage issue:** LD is surprisingly low across much of the genome and it varies tremendously.
  - It will be hard to assemble a good set of SNPs that “covers” the genome.
  - This process must be repeated for “all” distinct populations and distinct populations cannot be mixed for fear of admixture.

- **Next Generation sequencing technology:** Sequence the whole genome with chip technology. Increase the coverage.
Statistical Challenges

- Data processing
  - Normalization
  - Genotype calling
  - Data managing (tens to hundreds of GB)

- Data analysis
  - Single SNP analysis or multi-locus analysis?
  - Population substructure (Heterogeneity)?
  - Gene-gene interaction?
  - Gene-environment interaction?
  - Multiple testing (large $p (>500K)$, small $n (<2000))$?
  - A useful software: plink

- Biological interpretation
  - Involve multidisciplinary efforts (genetics, statistics, computer science,...)
A gene-centric approach

- Genes are the functional units. The sequence information and function of genes are highly consistent across diverse populations (Neale and Sham, 2004)

- Treat each gene as a testing unit:
  - Reduce genotyping burden
  - Release multiple testing burden: dealing first with the multiple variants within a gene and then with the multiple genes in the genome (Neale and Sham, 2004)
  - Improve reproducibility.
The idea

- Focus on the distribution of joint genotypes of multiple SNPs in a gene.

- The joint distribution captures the LD information of multiple SNPs and hence is robust when SNP interaction exists.

- Define a non-linear transformation of joint genotype frequency through entropy measure and test the difference of entropy between cases and controls.

Cui et al. (2008) *Genetics.*
“Heterogeneity” of complex diseases

“polygenic with genetic Heterogeneity”

Gene +

Gene -

Epistasis

Gene +

Gene -

complex phenotype

Salt intake

Psychosocial Stress

Diet

others

“Environmental factors”

Adaped from Monika Stoll
Gene-centric gene-gene interactions

- Gene-gene (G×G) interactions (or epistasis) play pivotal roles in determining trait variation and individual disease risk.

- The identification of G×G interaction has been traditionally sought by focusing on pairwise SNP×SNP interaction.

\[ y = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \]

- Genes are the functional units in living organism. It makes more sense to study gene-gene interaction (or epistasis) by considering each gene as a unit.

- We propose a gene-centric approach for G×G interaction using a kernel machine method.
The illustration

Single SNP interaction

Gene centric gene-gene interaction

Statistical models for G×E interaction

- **Traditional linear model for GxE interaction**
  \[ Y = \alpha + \alpha_1 X + \beta_l T_l + \sigma^2 \varepsilon, \quad 1 \leq l \leq d \]  
  \[ Y = \alpha + \alpha_1 X + \beta_l T_l + \delta_l XT_l + \sigma^2 \varepsilon, \quad 1 \leq l \leq d \]  
  \[ = \alpha + \alpha_1 X + (\beta_l + \delta_l X) T_l + \sigma^2 \varepsilon \] (1)

- **Allow non-linear G×E interaction**
  \[ Y = \alpha + \alpha_1 X + \beta_l (X) T_l + \sigma(X) \varepsilon, \quad 1 \leq l \leq d \] (2)

- **Assuming linear for \( \beta(X) = \beta + \delta X \), reduce to model (2), assuming \( \beta(X) = \beta \), reduce to model (1).**

- **A more general model:**
  \[ Y = \alpha(X) + \beta_l (X) T_l + \sigma(X) \varepsilon, \quad 1 \leq l \leq d \]

Ma et al., (2011) *Bioinformatics*
Gene set analysis

- Sets of genes that belong to a common pathway or Gene ontology term may function together to explain disease signal;

- Pathway or GO enrichment analysis
  - Fisher’s exact test (assuming a hypogeometric distribution under the null);
  - P-value combination method;
  - Kolmogorov running mean test;
Summary

- Single SNP analysis
- Multiple loci analysis
- Haplotype analysis
- Gene-based analysis
- Pathway-based analysis
- Gene-gene interaction
- Gene-environment interaction
- Rare variants analysis
- Multiple testing issues