THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Design and Analysis of Genetic Association Studies

Hemant K Tiwari, Ph.D.
Professor & Head
Section on Statistical Genetics

Department of Biostatistics
School of Public Health

Association Analysis

- Linkage Analysis used to be the first step in gene mapping process
- Closely located SNPs to disease locus may co-segregate due to linkage disequilibrium i.e. allelic association due to linkage.
- The allelic association forms the theoretical basis for association mapping

Linkage vs. Association

- **Linkage analysis is based on pedigree data (within family)**
- **Association analysis is based on population data (across families)**

- **Linkage analyses rely on recombination events**
- **Association analyses rely on linkage disequilibrium**

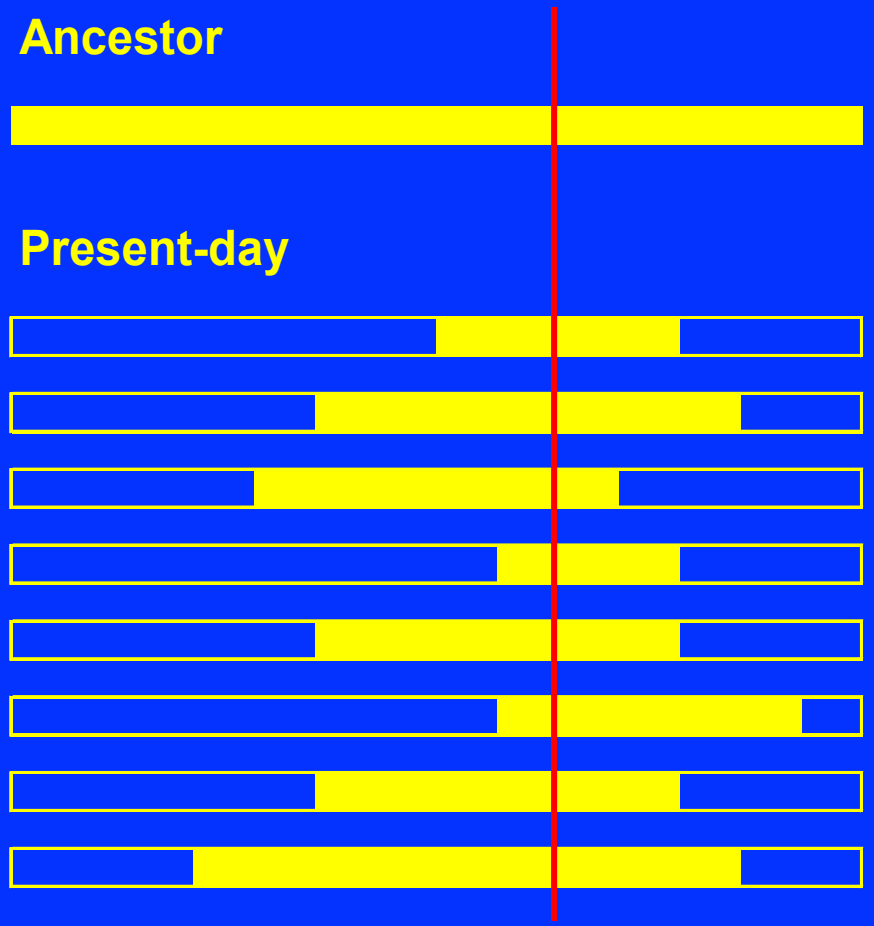
- **The statistic in linkage analysis is the count of the number of recombinants and non-recombinants**
- **The statistical method for association analysis is "statistical correlation" between Allele at a locus with the trait**

Linkage Disequilibrium

- Over time, meiotic events and ensuing recombination between loci should return alleles to equilibrium.
- But, marker alleles initially close (genetically linked) to the disease allele will generally remain nearby for longer periods of time due to reduced recombination.
- This is disequilibrium due to linkage, or "linkage disequilibrium" (LD).

Linkage Disequilibrium (LD)

- Chromosomes are mosaics
- Tightly linked markers
  - Alleles associated
  - Reflect ancestral haplotypes
- Shaped by
  - Recombination history
  - Mutation, Drift



Tishkoff and Verrelli (2003)

Measures of LD

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

Classical Definition of Disequilibrium Coefficient D_{AB}

$$\delta_{AB} = p_{AB} - p_A p_B = p_{AB} - p_{ab}$$

- Based on definition of δ_{AB}

$$p_{AB} = p_A p_B + \delta_{AB}$$

$$p_{Ab} = p_A p_b - \delta_{AB}$$

$$p_{aB} = p_a p_B - \delta_{AB}$$

$$p_{ab} = p_a p_b + \delta_{AB}$$

Commonly used other measures of disequilibrium: D' and r^2

$$D' = \begin{cases} \dfrac{\delta}{\min(p_1 q_2, p_2 q_1)} & \text{if } \delta > 0 \\ \dfrac{\delta}{\min(p_1 q_1, p_2 q_2)} & \text{if } \delta < 0 \end{cases}$$

$$r^2 = \frac{\delta^2}{p_1 p_2 q_1 q_2}$$

Patterns of LD in Human Genome

- The human genome has been portrayed as a series of high Linkage Disequilibrium (LD) regions separated by short segments of very low LD.
- In the high LD regions alleles tend to be correlated with one another.
- The high LD alleles tend to be transmitted from one generation to the next with a low probability of recombination.
- Such alleles can sometimes be used to infer the state of nearby loci
- The high LD regions are often referred to as blocks
- Blocks exhibit low haplotype diversity and most of the common haplotypes can be defined by relatively small number of SNPs (3-5)

Haplotype Blocks

- Within the haplotype block (in the high LD regions) haplotype diversity is low.
- Thus, only a few SNPs should be necessary to identify the haplotype structure within these regions or blocks
- These SNPs are called tag SNPs

Haplotype Blocks

- A *haplotype block* is a discrete (does not overlap another block) chromosome region of high LD and low haplotype diversity.

- They are blocks of the common haplotypes that represent a particular region of the chromosomes in a population

Haplotype Blocks

- Blocks extend many (>100) kbs
- All alleles within blocks are in strong associations.
- There are no associations between blocks.
- In each block, only a few (4-5) haplotypes account for the majority (90%) of variation.
- In each block, only a few SNPs are required to map the majority of haplotype variation.
- Blocks boundaries correspond to recombination hot-spots

HapMap Project

- Formally initiated in October 2002
- The HapMap Project is a huge international effort among scientist in Japan, UK, Canada, China, USA, and Nigeria
- Their goal was to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain
- Funded in part by grants from the NIH

HapMap II samples

- Study involves a total of 270 DNA samples representing peoples from around the world:
  - Northern and Western European
  - Yoruba (African)
  - Japanese
  - Han Chinese
- Promises to provide an important basis to carry out candidate-gene, linkage-based and genome-wide association studies

HapMap 3 samples

label	Population Sample	# Samples
ASW	African ancestry in Southwest US	90
CEU	Utah residents with northern & western ancestry from CEPH collection	180
CHB	Han Chinese in Beijing, China	90
CHD	Chinese in Metropolitan Denver, Colorado	100
GIH	Gujarati Indians in Houston, Texas	100
JPT	Japanese in Tokyo, Japan	91
LWK	Luhya in Webuye, Kenya	100
MEX	Mexican ancestry in Los Angeles, CA	90
MKK	Maasai in Kinyawa, Kenya	180
TSI	Toscani in Italy	100
YRI	Yoruba in Ibadan, Nigeria	180

1000 Genomes Project

- The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied using sequencing. (http://www.1000genomes.org/about)
- The plan for the full project is to sequence about 2,500 samples at 4X coverage.
- 1092 human genomes from 14 populations are available (Nature 491, 56–65 (01 November 2012) doi:10.1038/nature11632)

Association Study Design

Population-based association tests

- Cases-Control Design
- Ascertain two groups of individuals from the population: unrelated affected cases and unrelated unaffected controls.
- Can use standard statistical tests to compare the relative frequencies of alleles (genotypes) at a single marker locus in cases and controls (Chi-square test, logistic regression)
- Potentially subject to confounding by population admixture or stratification

Association Study Design

Family-based association tests

- Ascertain small nuclear families and extended pedigrees containing affected and unaffected individuals
- Use transmission of marker alleles from parents to offspring.
- Standard statistical tests to compare transmissions of marker alleles to affected and unaffected offspring (TDT, sibTDT, Pedigree TDT, TRANSMIT, etc.)
- Not confounded by admixture or stratification if conditioned on parents
- Valid test of linkage and association

Genome-wide Association Studies (GWAS)

- To scan 1 to 2.5 M SNPs of many people to find genetic variations associated with a disease

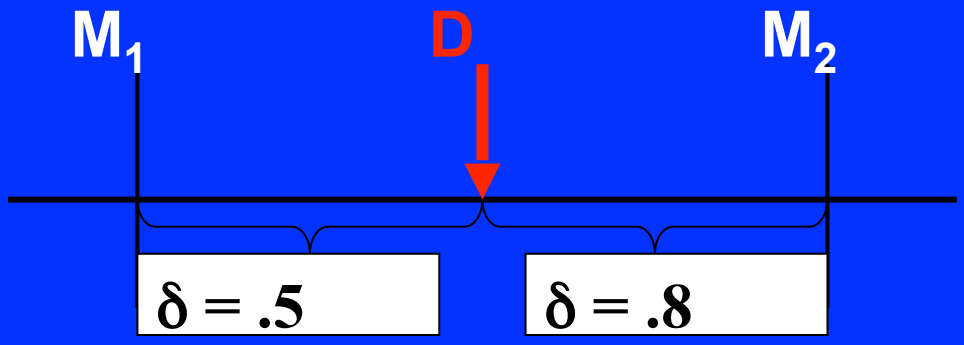
- GWAS are particularly useful in finding genetic variant that contribute to common, complex diseases, such as asthma, cardiovascular diseases, cancer, diabetes, obesity, and mental disorders.

Source: http://www.genome.gov/20019523#1

http://www.genome.gov/26525384

Why GWAS will enable us to find disease genes?

- It utilizes linkage disequilibrium between SNPs and putative gene loci.



- The coverage of the genome by SNPs has to be excellent

- Availability of genome-wide SNPs chip

First Successful GWAS on Age-Related Macular degeneration
Science: March 10, 2005

Complement Factor H Polymorphism in Age-Related Macular Degeneration

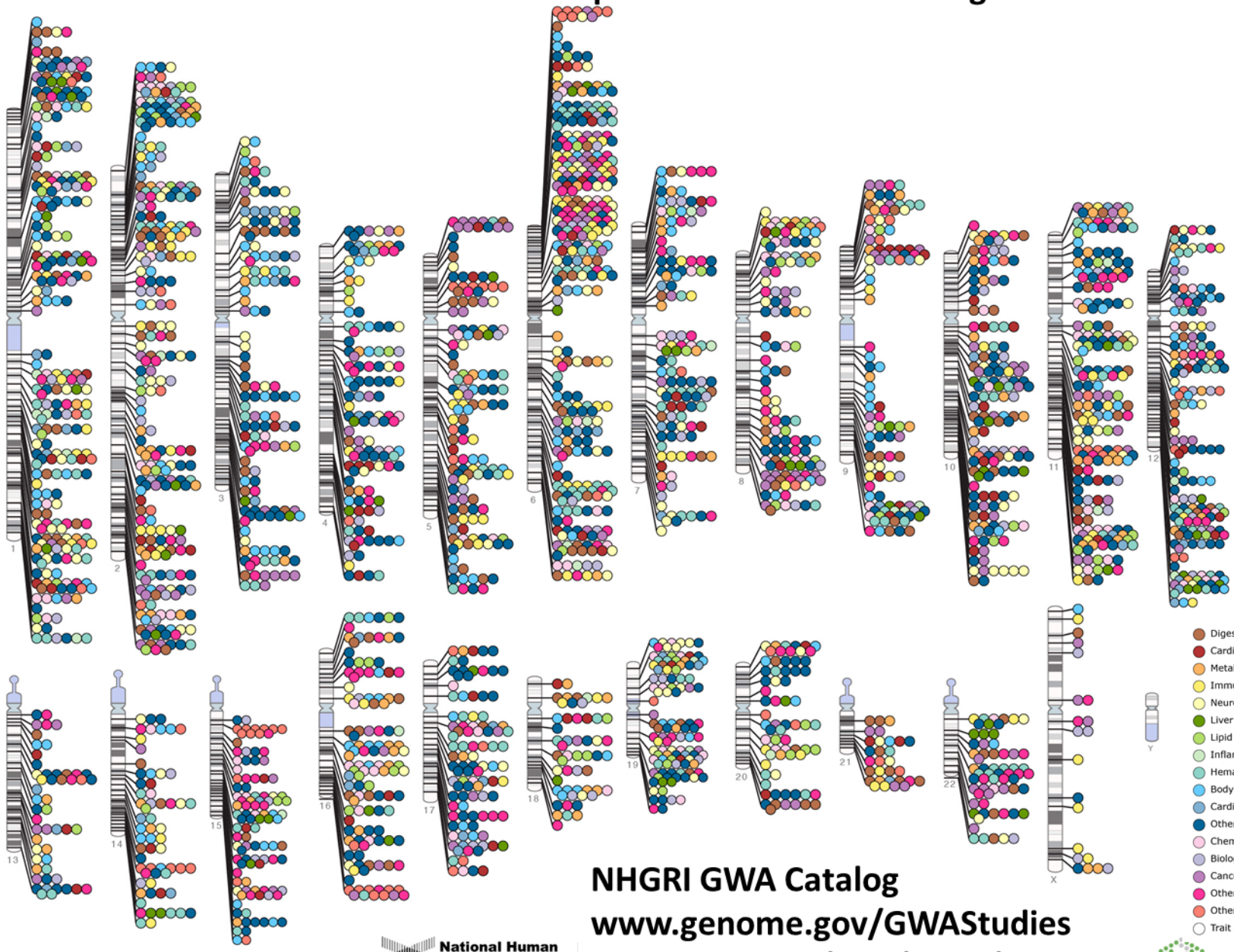
Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7†]

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

Using 96 cases and 50 controls Klein et al. (2005) found CFH gene on chromosome 1 ($p=4x10^{-8}$, OR=4.60) using 100K affy chip

Published Genome-Wide Associations through 07/2012

Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories



NHGRI GWA Catalog

www.genome.gov/GWAStudies

National Human Genome Research Institute

www.ebi.ac.uk/fgpt/gwas/

EMBL-EBI

What steps needed for GWAS

- Use appropriate design
  - Pedigrees, case-control, unrelated individuals
- Determine the sample size
  - Power
- Choose SNP genotyping platform
  - Affy, Illumina, Perlegen
- Perform QC (HWE, Mendelian errors, outliers, etc.)
- Imputation
- Choose appropriate Association test

Quality Control (QC)

- The first step of GWAS analysis is the quality control of the genotypic and phenotypic data. There are number of procedures needed to ensure the quality of genotype data both at the genotyping laboratory and after calling genotypes using statistical approaches.
- The QC and association analysis of GWAS data can be performed using the robust, freely available, and open source software PLINK developed by Purcell *et al*. (2007)

Quality Control (QC)

- **Sex Inconsistency:** It is possible that self-reported sex of the individual is incorrect. Sex inconsistency can be checked by comparing the reported sex of each individual with predicted sex by using X-chromosome markers' heterozygosity to determine sex of the individual empirically.
- **Relatedness and Mendelian Errors:** Another kind of error that can occur in genotyping is due to sample mix-up, cryptic relatedness, duplications, and pedigree errors such as self-reported relationships that are not accurate. The relationship errors can be corrected by consulting with the self-reported relationships and/or using inferred genetic relationships.

Quality Control (QC)

- **Batch Effects:** For GWAS, samples are processed together for genotyping in a batch. The size and composition of the sample batch depends on the type of the commercial array, for example, an Affymetrix array can genotype up to 96 samples, and an Illumina array can genotype up to 24 samples. To minimize batch effects, samples should be randomly assigned plates with different phenotypes, sex, race, and ethnicity.
- The most commonly used method is to compare the average minor allele frequencies and average genotyping call rates across all SNPs for each plate. Most genotyping laboratories perform batch effect detection and usually re-genotype the data if there is a batch effect or a plate discarded when there is a large amount of missing data.

Quality Control (QC)

- **Marker and sample genotyping efficiency or call rate:** Marker genotyping efficiency is defined as the proportion of samples with a genotype call for each marker. If large numbers of samples are not called for a particular marker, that is an indication of a poor assay, and the marker should be removed from further analysis. A threshold for removing markers varies from study to study depending on the sample size of the study. However, usual recommended call rates are approximately 98% to 99%.

Quality Control (QC)

- **Population stratification:** There are a number of methods proposed to correct for population substructure. Three commonly used methods to correct for the underlying variation in allele frequencies that induces confounding due to population stratification:
  - genomic control
  - structured association testing
  - principal components (Most Commonly Used Method)

Population Stratification

- Population stratification: Sample consists of divergent populations
- Case-control studies can be affected by population stratification

False positive due to admixture

Population 1

	Allele A	Allele B	Total
Affected	64	16	80
Unaffected	16	4	20
Total	80	20	

OR=1.0 (CI 0.29-3.4), p-value=1

Population 2

	Allele A	Allele B	Total
Affected	4	16	20
Unaffected	16	64	80
Total	20	80	

OR=1.0 (CI 0.29-3.4), p-value=1

Combine both population with equal proportion

	Allele A	Allele B	Total
Affected	68	32	100
Unaffected	32	68	100
Total	100	100	

OR=4.5 (CI 2.5-8.2), (p-value $=6.6 \times 10^{-7}$)

True association can be masked due to admixture

Population 1

	Allele A	Allele B	Total
Affected	20	80	100
Unaffected	80	20	100
Total	100	100	

OR=0.06, p-value = 4.4×10^{-14}

Population 2

	Allele A	Allele B	Total
Affected	80	20	100
Unaffected	20	80	100
Total	100	100	

OR=16.0, p-value = 4.4×10^{-14}

Combine both population with equal proportion

	Allele A	Allele B	Total
Affected	100	100	200
Unaffected	100	100	200
Total	200	200	

OR=1, p-value = 1

How to correct for stratification

- Stratification can be adjusted in your analysis by using.
  - Family-based design
    - TDT in family-based association
  - Population-based design
    - Admixture mapping: Structured Association Testing, Genomic Control, Regional Admixture mapping, Principal Components Method

Quality Control (QC)

- Principal components analysis (PCA) uses thousands of markers to detect population stratification and Principal Components (PCs) then can be used to correct for stratification by modeling PCs as covariates in the model
- PCs can be calculated using a program Eigenstrat (Patterson et al., 2006; Price et al., 2006). There are two issues with using PCA, (1) how many SNPs to use, and (2) how many PCs should be included as covariates in the association analysis.

Quality Control (QC)

- **Hardy-Weinberg equilibrium (HWE) filter:** The HWE test compares the observed genotypic proportion at the marker versus the expected proportion. Deviation from HWE at a marker locus can be due to population stratification, inbreeding, selection, non-random mating, genotyping error, actual association to the disease or trait under study, or a deletion or duplication polymorphism. However, HWE is typically used to detect genotyping errors. SNPs that do not meet HWE at a certain threshold of significance are usually excluded from further association analysis.

Quality Control (QC)

- **Marker allele frequency filter:** It is also important to discard SNPs based on minor allele frequency (MAF). Most GWAS studies are powered to detect a disease association with common SNPs (MAF $\geq$ 0.05). The rare SNPs may lead to spurious results due to the small number of homozygotes for the minor allele, genotyping errors, or population stratification

Genotype Imputation

- It is common to impute missing SNP data, e.g. from 1 M SNPs to 2.5 M SNPs using either HapMap or 1000 Genomes data
- There are number of programs available to perform imputation
  - IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
  - MACH (http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html)
  - BEAGLE (http://faculty.washington.edu/browning/beagle/beagle.html)

Why so much interest in imputing missing genotypes?

- Inexpensive "*in silico*" genotyping strategies
- Estimate genotypes for individuals related to those in GWAS sample
- Estimate additional genotypes for individuals in the GWAS sample
  - Facilitate comparisons across studies
  - Improve coverage of the genome (more genotypes better the coverage)

Family Data Imputation

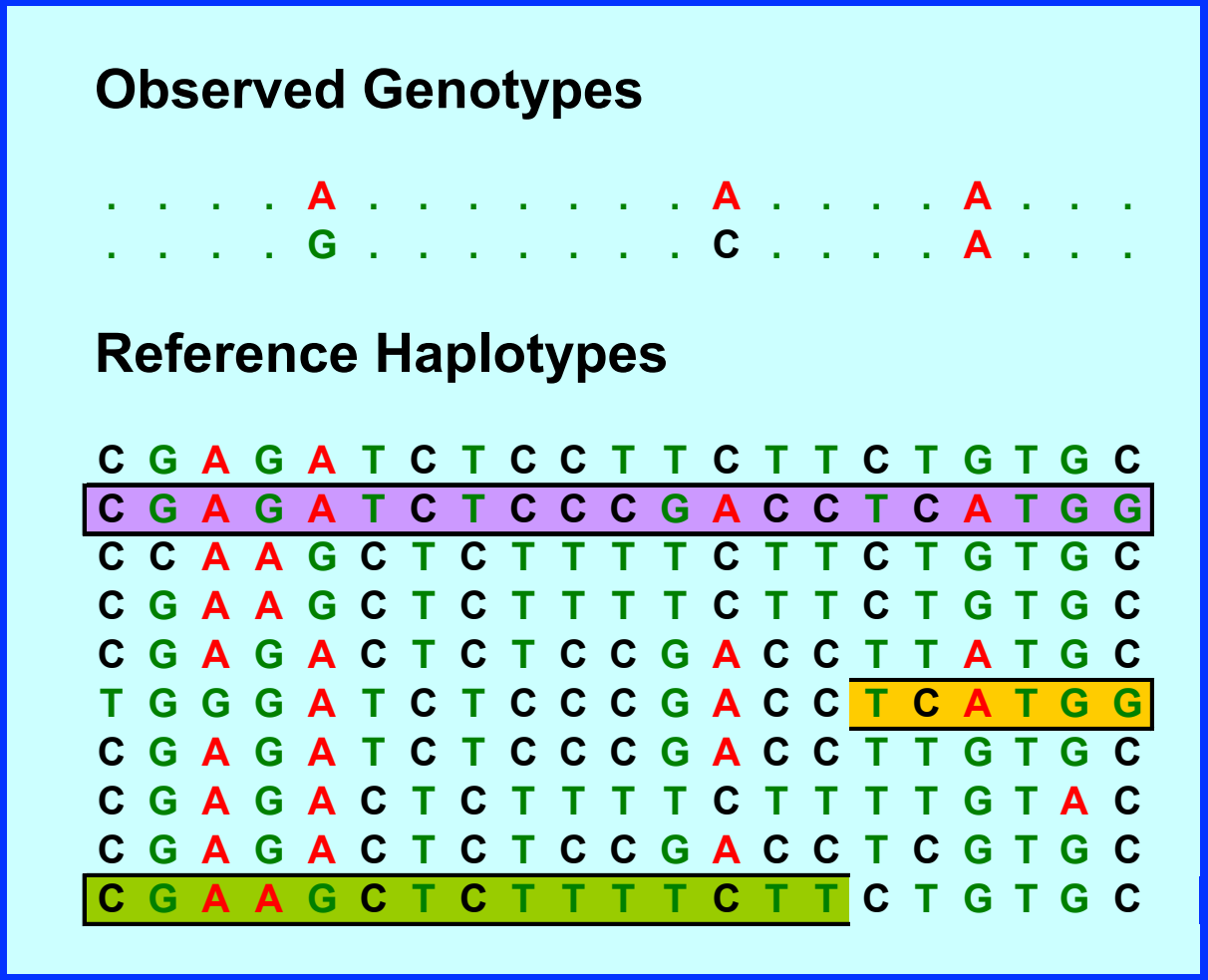
- Much easier
- Can get very accurate genotypes
- Based on the

  P (missing genotype | IBD sharing within haplotypes)

Population Data

- In pedigrees, we expect relatively long stretches of shared chromosome
- In population sample, these stretches will typically be much shorter
- But, this should not stop us for imputing!
- We can borrow the information from known haplotype data sets (HapMap, 1000 Genomes)

Identify match among reference



Courtesy Gonçalo Abecasis

Phase chromosomes, impute missing genotypes

Observed Genotypes

```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Courtesy of Gonçalo Abecasis

Issues with Imputation

- Requires large scale computing resources
- Need to assess quality of imputation
  - Compare imputed genotypes to actual genotypes
- Error rates are higher than for genotyped SNPs
- Works less well for rarer alleles
- Best to take account of uncertainty imputed SNPs in analysis

Analysis Procedures

- One Stage procedure
  - All markers are typed on all samples
  - Replication is left for others
- Two Stage procedure
  - All markers are typed on all samples at stage 1
  - Replication study is performed at stage 2 as a replication study on a different sample & only significant SNPs from stage 1 are used
- Replication
  - Replication is must from a protection for false positives
  - Most of the journals require replication

Study Designs & Methods for GWAS

	Details	Advantages	Disadvantages	Statistical analysis method
Cross-sectional	Genotype and phenotype (ie, note disease status or quantitative trait value) a random sample from population	Inexpensive. Provides estimate of disease prevalence	Few affected individuals if disease rare	Logistic regression, χ^2 tests of association or linear regression
Cohort	Genotype subsection of population and follow disease incidence for specified time period	Provides estimate of disease incidence	Expensive to follow-up. Issues with drop-out	Survival analysis methods
Case-control	Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample	No need for follow-up. Provides estimates of exposure effects	Requires careful selection of controls. Potential for confounding (eg, population stratification)	Logistic regression, χ^2 tests of association
Extreme values	Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample	Genotype only most informative individuals hence save on genotyping costs	No estimate of true genetic effect sizes	Linear regression, non-parametric, or permutation approaches
Case-parent triads	Genotype affected individuals plus their parents (affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)	Robust to population stratification. Can estimate maternal and imprinting effects	Less powerful than case-control design	Transmission/disequilibrium test, conditional logistic regression or log-linear models
Case-parent-grandparent septets	Genotype affected individuals plus their parents and grandparents	Robust to population stratification. Can estimate maternal and imprinting effects	Grandparents rarely available	Log-linear models
General pedigrees	Genotype random sample or disease-outcome based sample of families from general population. Phenotype for disease trait or quantitative trait	Higher power with large families. Sample may already exist from linkage studies	Expensive to genotype. Many missing individuals	Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test
Case-only	Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease-outcome based sample	Most powerful design for detection of interaction effects	Can only estimate interaction effects. Very sensitive to population stratification	Logistic regression, χ^2 tests of association
DNA-pooling	Applies to variety of above designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis	Potentially inexpensive compared with individual genotyping (but technology still under development)	Hard to estimate different experimental sources of variance	Estimation of components of variance

HJ Cordell, DG Clayton. Genetic association studies. Lancet 2005; 366: 1121-31

Statistical Methods & Software for Genetic Association Studies

	Approach	Reference	Software	URL
Logistic regression	Model log odds of disease as linear function of underlying genotype variables	20, 74, 20	Standard statistical package (eg, Stata, SAS, S-Plus, R)	http://www.stata.com/ http://www.sas.com/ http://www.insightful.com/products/splus/ http://www.r-project.org/
χ^2 test of association	Test for independence of disease status and genetic risk factor	20	Standard statistical package	See above
Linear regression	Model quantitative trait as linear function of underlying genotype variables	75	Standard statistical package	See above
Survival analysis	Model survivor function or hazard as function of underlying genotype variables	20, 52	Standard statistical package	See above
Transmission/ disequilibrium test	Test departure of transmission of alleles from heterozygous parents to affected offspring from null hypothesis of half	71, 76–78	Various (eg, Genehunter, RC-TDT, Genassoc, Transmit, Unphased	http://fhcrc.org/labs/kruglyak/Downloads/index.html http://www.uni-bonn.de/~umt70e/soft.htm http://www-gene.cimr.cam.ac.uk/clayton/software/ http://www.mrc-bsu.cam.ac.uk/personal/frank/
Conditional logistic regression	Calculate conditional probability of affected offspring genotypes, given parental genotypes	54, 60, 79, 80	Genassoc Unphased	http://www-gene.cimr.cam.ac.uk/clayton/software/ http://www.mrc-bsu.cam.ac.uk/personal/frank/
Log linear models	Model counts of genotype combinations for mother, father, and affected offspring	57, 58, 59	Standard statistical package	See above
Pedigree disequilibrium test	Test departure of transmission of alleles to affected pedigree members from null expectation	81, 82	Pedigree disequilibrium test Unphased	http://www.chg.duke.edu/software/pdt.html http://www.mrc-bsu.cam.ac.uk/personal/frank/
Family-base association test	Tests for association or linkage between disease phenotypes and haplotypes by utilising family-based controls	83–86	Family-based association test	http://www.biostat.harvard.edu/~fbat/fbat.htm
Quantitative transmission/ disequilibrium test	Linkage disequilibrium analysis of quantitative and qualitative traits based on variance components	87, 88	Quantitative transmission/ disequilibrium test	http://www.sph.umich.edu/csg/abecasis/QTDT/
DNA pooling	Test for differences in allele frequencies in different pooled samples while estimating components of variance due to experimental error	61, 89–91	Standard statistical package	See above

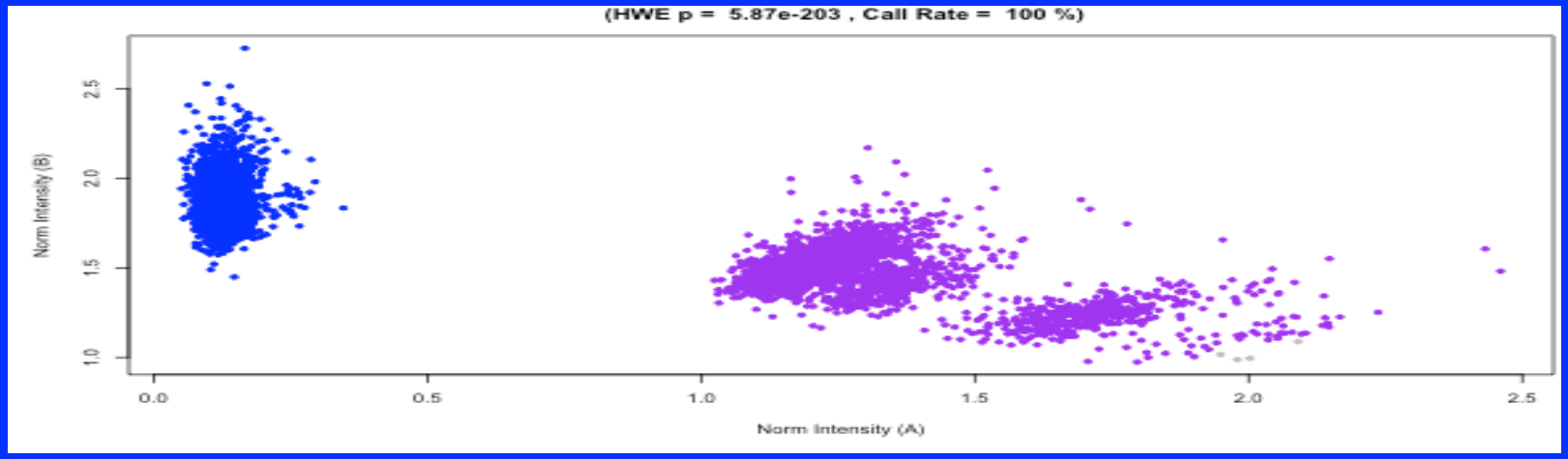
The references are those from the following paper:

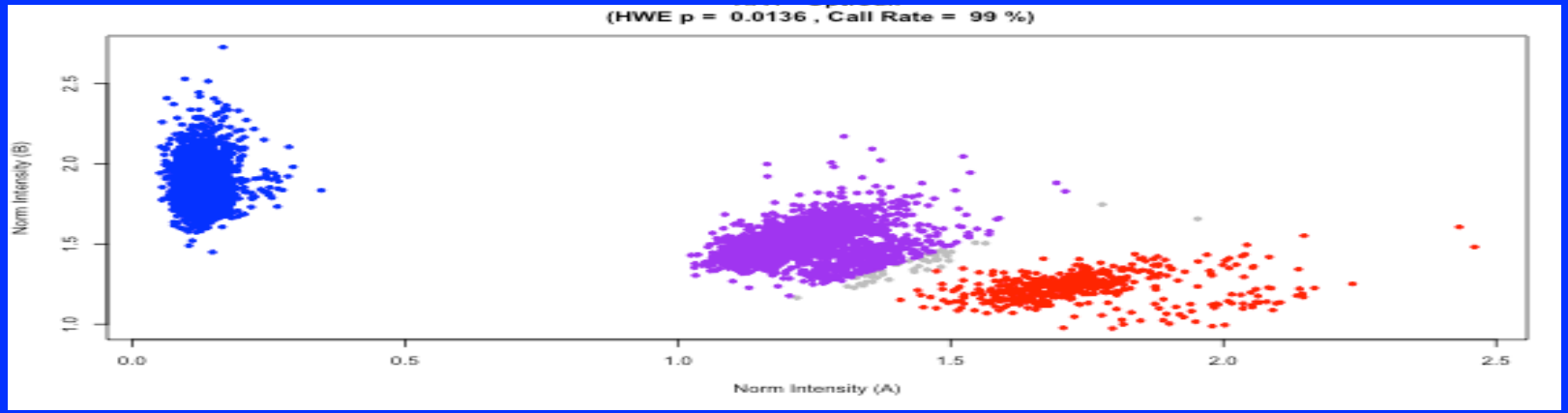
HJ Cordell, DG Clayton. Genetic association studies. Lancet 2005; 366: 1121-31

Commonly Used Software

- FBAT
  - Family based association analysis
- PLINK
  - Whole genome association analysis toolset
- SAGE (ASSOC)
  - Statistical Analysis for Genetic Epidemiology
- LMEKIN in R
  - Mixed-model procedure to analyze familial data
- STRUCTURE
  - Population structure inference
- EIGENSTRAT
  - Detects and corrects for population stratification in genome-wide association studies

After Association Analysis QC (Cluster Plots)





Life After Linkage & GWAS

- Copy number variations (CNVs)
  - Duplications, deletions
- Next Generation Sequencing
- Whole-genome methylation
  - Modification of a molecule by the addition of a methyl group
- Metabolomics
- Microbiome
- RNA-Seq
- CHiP-Seq