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

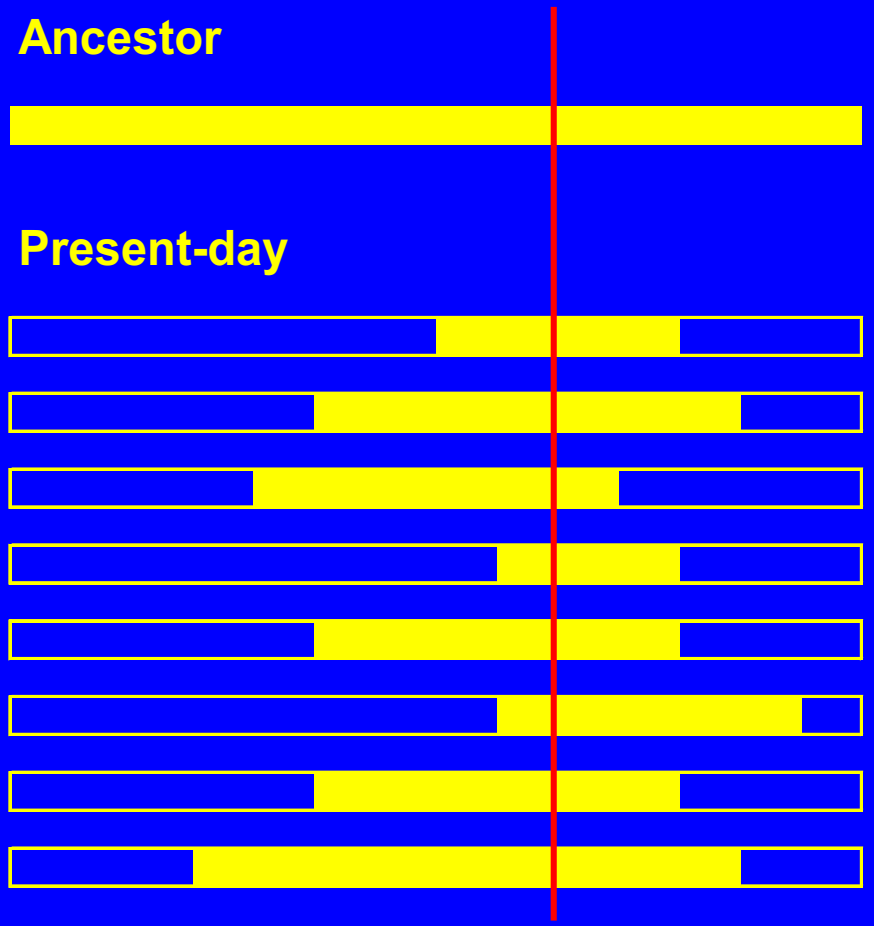
- **Linkage analysis is based on pedigree data (within family)**
- **Association analysis is based on population data (across families)**

- **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 (LD)

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



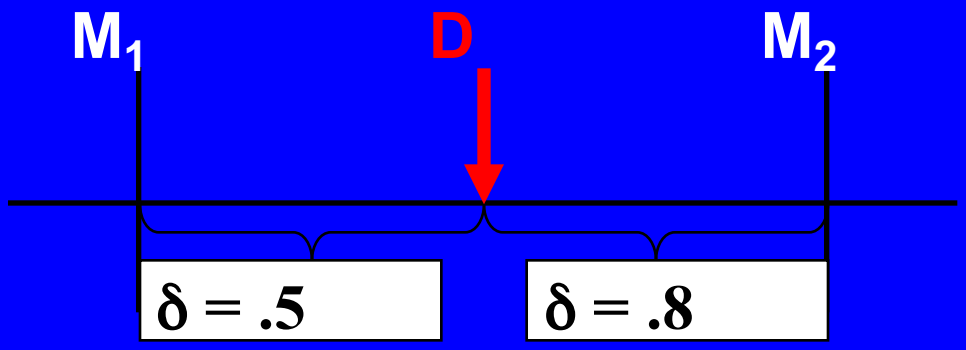
Tishkoff and Verrelli (2003)

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).

Why GWAS 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

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

- 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

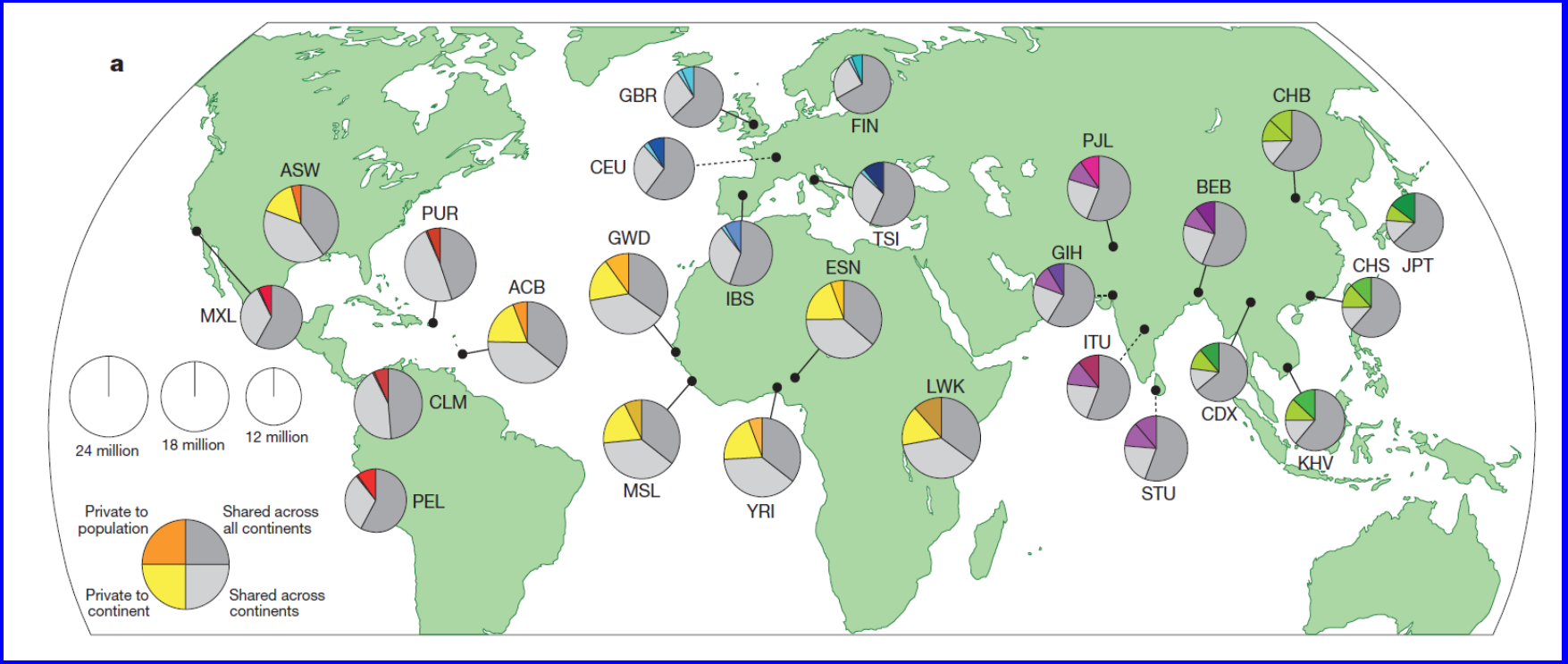
HapMap 3 samples

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 was 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 was to sequence about 2,500 samples at 4X coverage.
- 2504 human genomes from 26 populations are available (The 1000 Genomes Project Consortium. *A global reference for human genetic variation*. Nature, Vol 526, 68–74 (01 October 2015); An inregrated map of structural variation in 2504 genomes, Nature, Vol 526, 75–81 (01 October 2015))

Population	Samples
Chinese Dai in Xishuangbanna, China(**CDX**)	93
Han Chinese in Bejing, China (**CHB**)	103
Japanese in Tokyo, Japan (**JPT**)	104
Kinh in Ho Chi Minh City, Vietnam (**KHV**)	99
Southern Han Chinese, China (**CHS**)	105
Total East Asian Ancestry (EAS)	**504**
Bengali in Bangladesh (**BEB**)	86
Gujarati Indian in Houston,TX (**GIH**)	103
Indian Telugu in the UK (**ITU**)	102
Punjabi in Lahore,Pakistan (**PJL**)	96
Sri Lankan Tamil in the UK (**STU**)	102
Total South Asian Ancestry (SAS)	**489**
African Ancestry in Southwest US (**ASW**)	61
African Caribbean in Barbados (**ACB**)	96
Esan in Nigeria (**ESN**)	99
Gambian in Western Division, The Gambia (**GWD**)	113
Luhya in Webuye, Kenya (**LWK**)	99
Mende in Sierra Leone (**MSL**)	85
Yoruba in Ibadan, Nigeria (**YRI**)	108
Total African Ancestry (AFR)	**661**
British in England and Scotland (**GBR**)	91
Finnish in Finland (**FIN**)	99
Iberian populations in Spain (**IBS**)	107
Toscani in Italia (**TSI**)	107
Utah residents with Northern and Western European ancestry (**CEU**)	99
Total European Ancestry (EUR)	**503**
Colombian in Medellin, Colombia (**CLM**)	94
Mexican Ancestry in Los Angeles, California (**MXL**)	64
Peruvian in Lima, Peru (**PEL**)	85
Puerto Rican in Puerto Rico (**PUR**)	104
Total Americas Ancestry (AMR)	**347**
Total	**2504**



The 1000 Genomes Project Consortium. *A global reference for human genetic variation*. *Nature*, Vol 526, 68–74 (01 October 2015)

What steps needed for GWAS

- Use appropriate design
  - Pedigrees, case-control, unrelated individuals (population sample)
- Determine the sample size
  - Power
- Choose SNP genotyping platform
  - Affy or Illumina
  - Perform QC (HWE, Mendelian errors, outliers, etc.)
- Imputation of genotype data using 1000 Genomes
- Choose appropriate Association test
- QC after association testing

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

Power Analysis

- To calculate power, assume that there are 20 causal SNPs (out of 1M tested) associated with the disease in case-control design with same effect size assessed by Genetic Relative Risk (GRR) or OR. Further, the power is calculated based on number of SNPs to be detected out of 20 at Genome-wide level ($5x10^{-8}$).

Sample Size (case-control pairs)	Minimum MAF	Minimum Genetic Relative Risk (GRR)	# of variants to be detected	Power
2000-2000	0.04	1.6	At least 1 variant	98.51%
		2.2	All 20 variants	93.98%
3000-3000	0.04	1.5	At least 1 variant	99.65%
		1.9	All 20 variants	87.42%

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/gwastudies/
http://www.ebi.ac.uk/gwas/

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

Complement Factor H Polymorphism in Age-Related Macular Degeneration

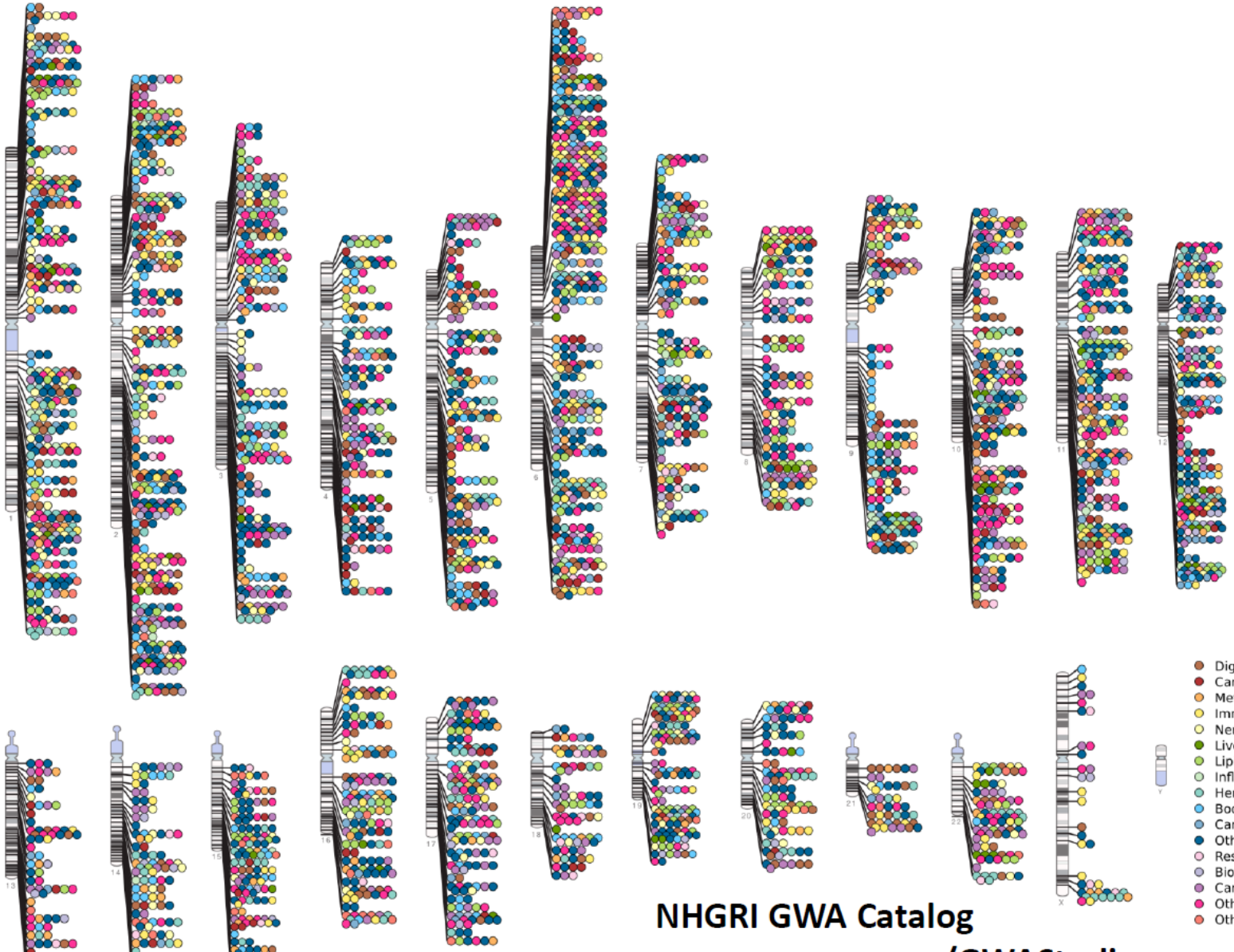
Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7†]

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 12/2013

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog

www.genome.gov/GWAStudies

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

National Human Genome Research Institute

EMBL-EBI

Quality Control (QC)

- The first step of GWAS analysis is the quality control of the *phenotypic* and *genotypic* 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)

Phenotype QC

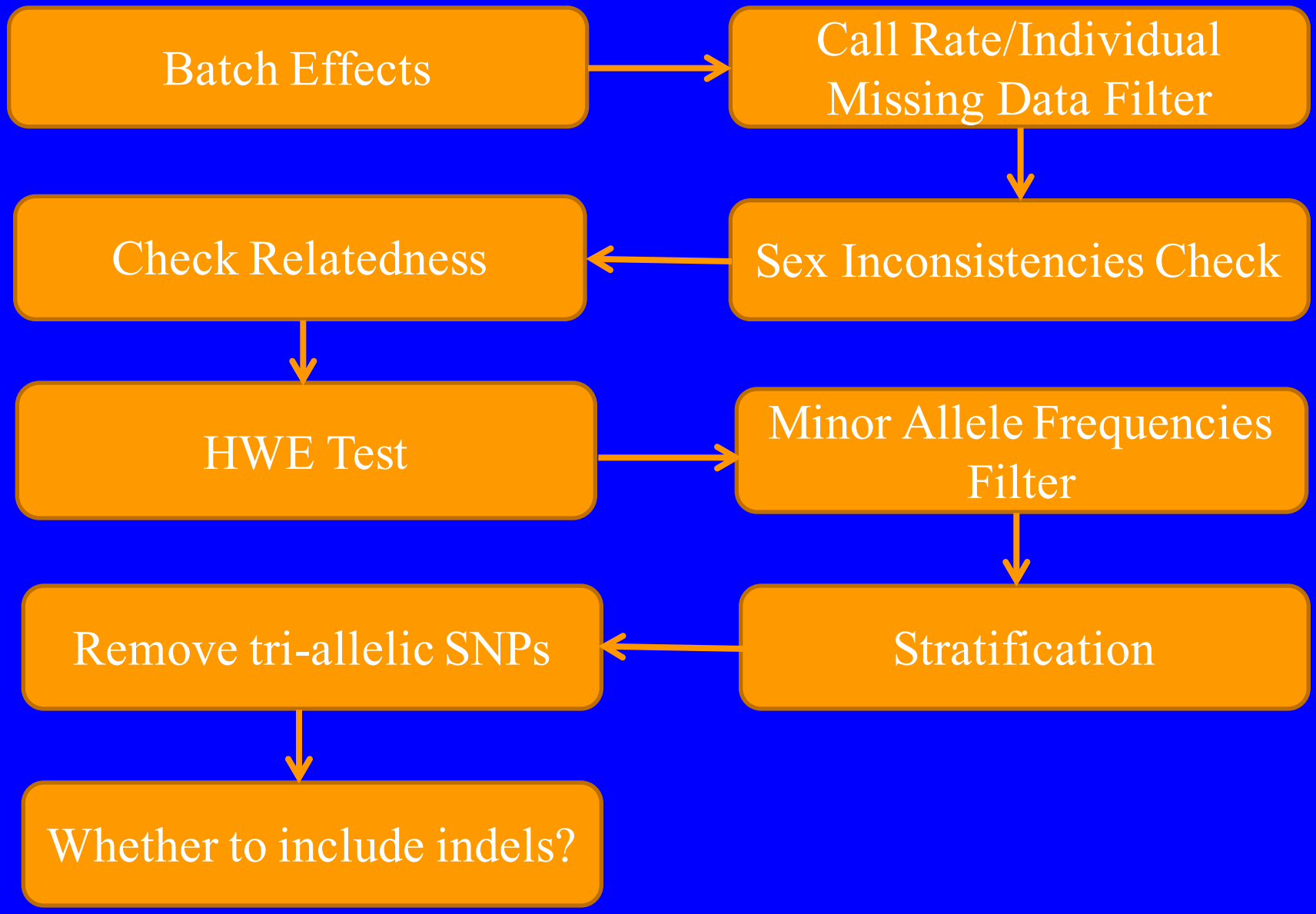
- The phenotype is extremely critical to good genetic studies
  - Precise
  - The closer to a gene product the easier it is to detect an association
  - Highly variable

Wojczynski MK, Tiwari HK. Definition of phenotype. Adv Genet. 2008; 60:75-105. Review.

Phenotype data management and QC

- Create a standard report with descriptive analysis, including and especially histogram of distribution, moments (especially mean, variance, skewness, kurtosis) of continuous covariates. For dichotomous covariates, describe frequencies and check if the frequencies are different in two categories.
  - Consider whether box plots will be helpful.
- Check distribution of quantitative traits to be analyzed. If they are not normally distributed then transform them (log, inverse, box-cox, etc.) to make normal.
- Regarding outlier detection, check for univariate outliers using 4SD criteria. If needed, check for multivariate outliers using Mahalanobis distance.
- If needed, impute missing phenotype data using multiple imputation.

QC for Genetic Data Pipeline



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. To minimize batch effects, samples should be randomly assigned on plates with different phenotypes, sex, race, and ethnicity.
- The most commonly used method to detect batch effect 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): Call Rate & Missingness in Individuals

- Marker genotyping efficiency or call rate 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%**.
- Individuals missing more than **10%** of SNPs are usually deleted

Quality Control (QC): Sex Inconsistency check

- 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. **Use genetic sex, whenever sex is modeled in the analysis.**

Quality Control (QC): Relatedness Check

- Another kind of error that can occur in genotyping is due to sample mix-up or contamination, sample swaps, cryptic relatedness, and sample duplications. The relationship errors can be corrected by using identical by descent/ identical in state distribution of all individuals.

Quality Control (QC): HWE Test

- 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. We will use HWE p-value<0.001, the same as the HapMap project. but those SNPs which deviate from HWE at significance level of 0.001 will be labeled as "suspect" and will be tested using cluster plots or intensity plots to re-examine the MAF, and will be excluded from further investigation if they will be excluded if they still do not meet HWE.

Quality Control (QC): MAF 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. We may remove SNPs with MAF<0.01. This criterion ensures that at least 30 cases and 30 controls are present, respectively with MAF $\geq$0.01.

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).
  - Use scree plot to determine the number of PCs to be included or
  - investigate correlation between phenotypes and PCs and include PCs showing significant correlation with the phenotype.

SNPs with more that 2 alleles

- All tri-allelic SNPs will be excluded

Decide whether to include indels

- In first pass of GWAS, we usually do not include indels

Genotype Imputation

- It is common to impute missing SNP data, e.g. from 1 M or 2.5 M SNPs 7-9 M SNPs using either 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)

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

Example Imputation

- We followed the two-stage procedure proposed by Abecasis Group documented in the 1000 Genomes Imputation Cookbook (http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook) .
- First stage involves the use of MACH software/library for pre-phasing.
- Second stage uses Minimac for imputation.
- The version of the **1000G reference panel** is recommended and distributed through the imputation programs MACH/MINIMAC website. http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2013-09.html.
- The original Phase 1 release 1000G reference panel contains 38 M SNVs which does not include singletons and monomorphic sites, therefore there are a total of 27.5 Million SNPs available to be imputed

Quality Control after Imputation

	Reference Panel SNVs	Post Imputation R^2 and EAF Cutoff Distribution					
Chrom	Total SNVs	EAF=0.01 R^2 = 0.1	EAF=0.01 R^2 =0.3	EAF=0.03 R^2 = 0.1	EAF=0.03 R^2 = 0.1	EAF=0.05 R^2 =0.1	EAF=0.05 R^2 = 0.3
1	2155157	691797	681787	553271	553271	487871	484856
2	2346858	753568	745060	602021	602021	531053	529018
3	1966659	642297	635900	516191	516191	457481	455671
4	1968167	660091	651765	532885	532885	468870	466516
5	1808082	585163	579575	467848	467848	414029	412929
6	1755856	606952	601193	492397	492397	433583	431362
7	1599382	530217	519825	429625	429625	377046	373271
8	1557424	499586	494626	400371	400371	353567	352475
9	1187730	386265	380586	311331	311331	274530	272702
10	1262742	429071	424248	348335	348335	307532	305924
11	1356880	450613	444321	364802	364802	324866	322690
12	1314327	437958	431759	349042	349042	308497	306757
13	987739	337520	334248	270993	270993	237994	237165
14	904349	296219	292201	236018	236018	207566	206256
15	812545	257689	253534	205225	205225	181596	180651
16	865998	274706	266040	219507	219507	193408	190250
17	753172	243107	233897	194885	194885	173603	170908
18	783008	261110	256801	209193	209193	184992	183586
19	603516	207897	192183	169589	169589	150619	143194
20	617694	202569	199032	162495	162495	143377	142400
21	377553	126070	123708	103609	103609	92684	92051
22	365644	124827	120350	100631	100631	88935	86957
Total	27,350,482	9,005,292	8,862,639	7,240,264	7,240,264	6,393,699	6,347,589

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

- Genome-wide Efficient Mixed Model Association (GEMMA)
  - Family based association analysis
  - (http://www.xzlab.org/software.html)
- FBAT or PBAT
  - Family based association analysis
- GenABEL or ProbABEL (http://www.genabel.org/)
  - Family based association analysis
- PLINK
  - Whole genome association analysis toolset
- SAGE (ASSOC)
  - Statistical Analysis for Genetic
- LMEKIN in R
  - Mixed-model procedure to analyze familial data

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