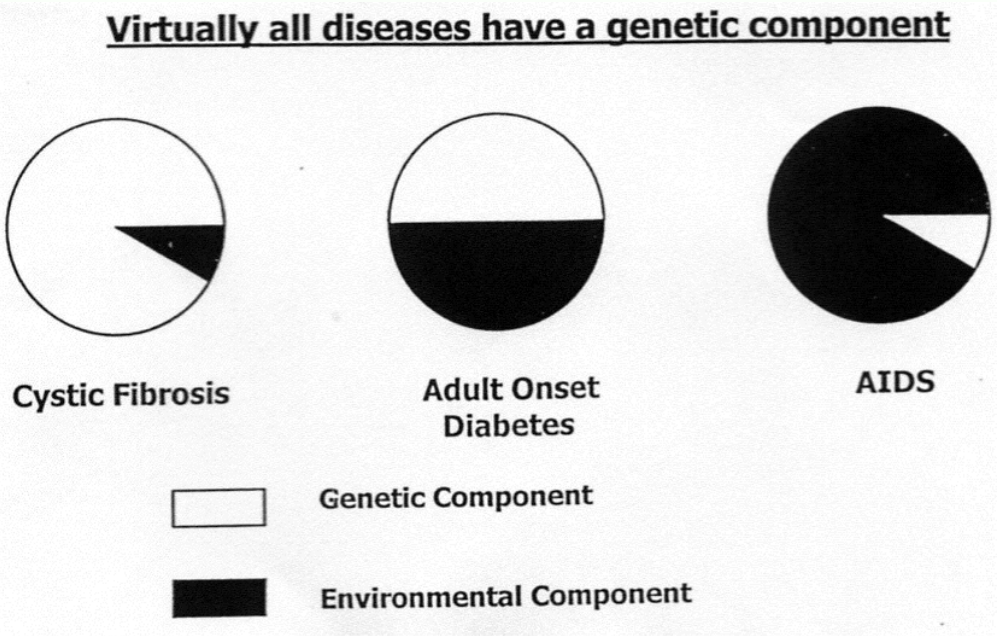
Biomarker Discovery, Validation and Implementation (BRIM 620/820)

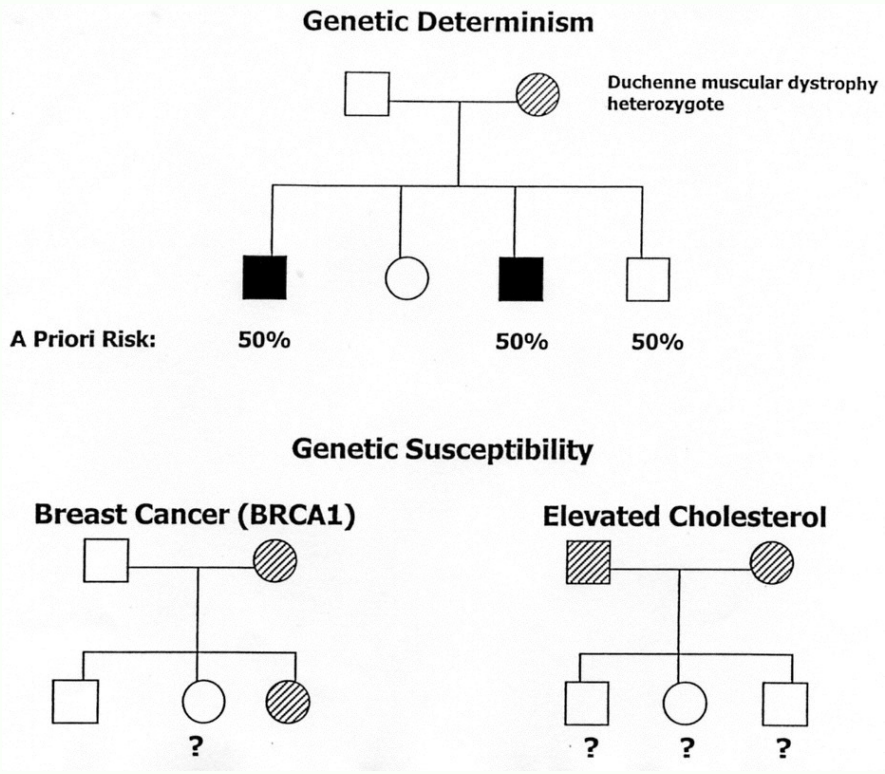
Genome-Wide Association Studies (GWAS) and Identification of Disease-Susceptibility Genes

George T. Cicila, Ph.D.
Physiology and Pharmacology

February 1, 2016

george.cicila@utoledo.edu
382 Block Health Sciences





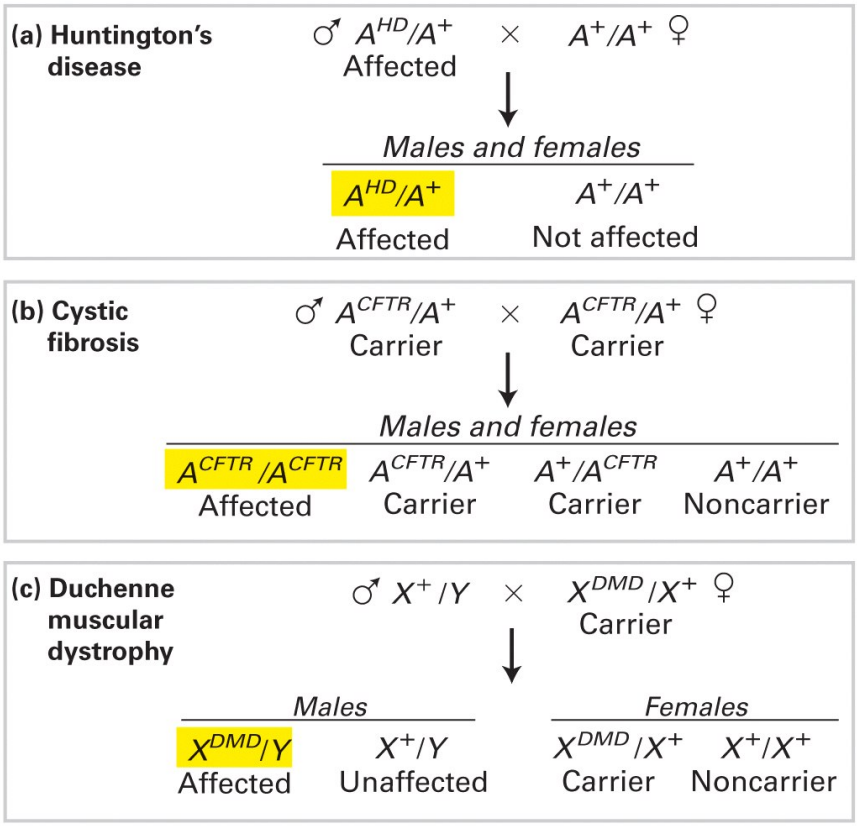
Genetic Architecture of Disease

- Most Diseases Have a Genetic Component
- **Mendelian Disorders**: Identification of causal variant straightforward
- Map inheritance of disease phenotype in affected families to genomic regions of shared inheritance in affected individuals, narrow region by identifying recombination, identify variants in resident genes
- Linkage

Autosomal Dominant

Autosomal Recessive

X-Linked Recessive



Genetic Architecture of Disease

- Most Diseases Have a Genetic Component
- Most Diseases do **NOT** show Mendelian patterns of inheritance
- Multifactorial—complex, multi-genic
- Support for involvement of genetic factors
- Cases cluster in families
- Families also share environmental factors
- Family studies, Twin studies
  - phenotype concordance provide estimate of heritability

We will focus on two kinds of traits

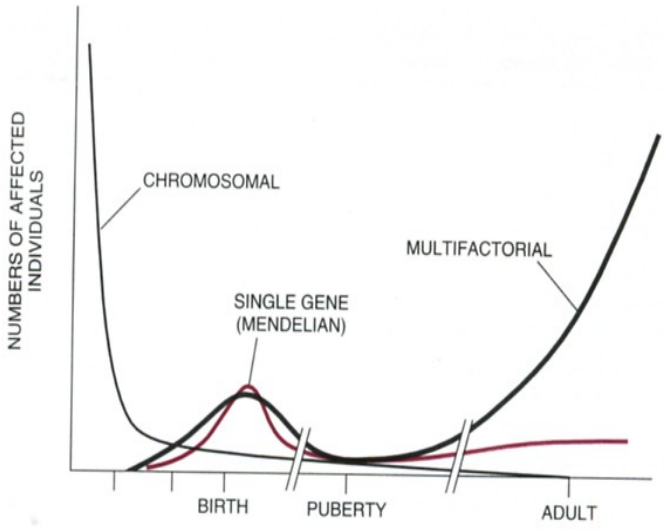
Mendelian

- Variation in phenotype largely due to inheritance of single genes
- Alleles for these diseases result in discrete phenotypes
- Dominant or recessive patterns of inheritance for phenotype

Multifactorial or Complex

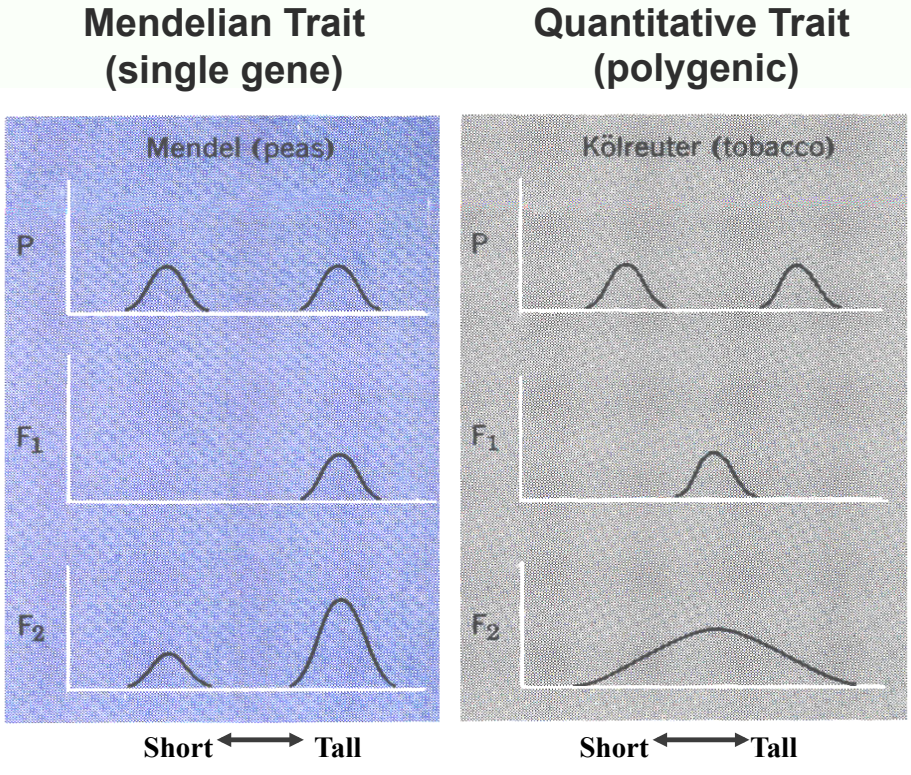
- Variation in phenotype due to many genes (polygenic)
- Phenotypes show continuous variation, sum of the effects of all contributing loci
- Phenotype often normally distributed

Temporal Changes in Chromosomal, Mendelian, and Multifactorial Diseases/Disorders



Multifactorial Inheritance

- Phenotypic traits (commonly quantitative) resulting from the **interaction** of multiple environmental factors with multiple genes
- Complex, multifactorial traits do **NOT** demonstrate simple, Mendelian patterns of inheritance
- Risk should be increased for sibs of patients showing severe expression of the trait



Genetic Markers

Genetic markers are used to test *cosegregation* between alleles at marker and trait loci

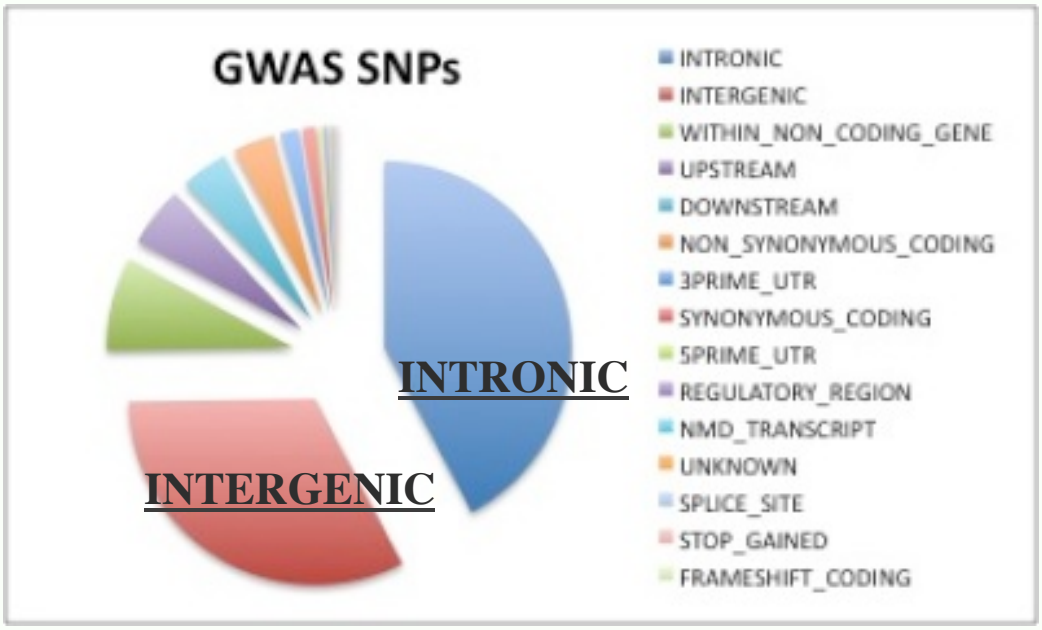
- Restriction-Fragment Length Polymorphisms (**RFLP**)

- Simple Sequence Repeats (SSRs) or **Microsatellite markers** (short tandem DNA repeats)

- Single nucleotide polymorphisms (**SNPs**).

Often used in groups of linked markers to define **haplotypes**

Markers closest to the disease gene show strongest correlation with disease patterns in affected families.

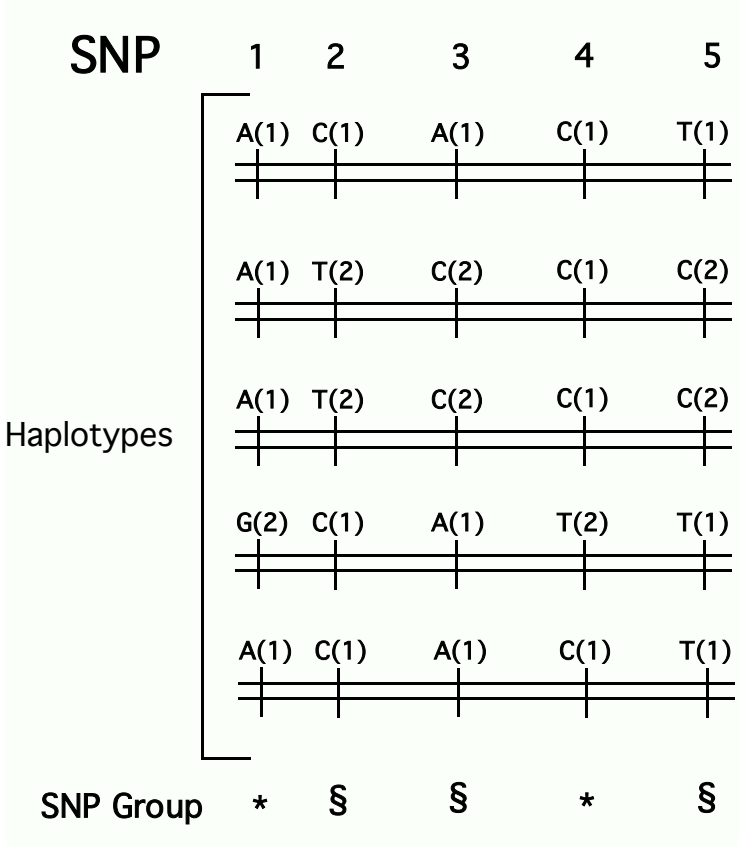


WikiGenes: Principles for the post-GWAS functional characterisation of risk loci.

file:///Volumes/HP%20v125w/BRIM%20GWAS%20lecture/new%20GWAS%20articles/WikiGenes%20-%20Principles%20for%20the%20post-GWAS%20functional%20characterisation%20of%20risk%20loci.webarchive

Identifying Disease Susceptibility Genes

- Begins with linkage or association of loci with disease in segregating populations (laboratory studies) and/or large families or populations (humans) using polymorphic markers.
- For some diseases, presence of the aberrant phenotype is very closely associated with the alleles of a particular locus or set of loci (**haplotype**),
- these are said to be in **linkage disequilibrium.**



Adapted from Cho JH (2010) Genome-wide association studies: Present status and future directions.

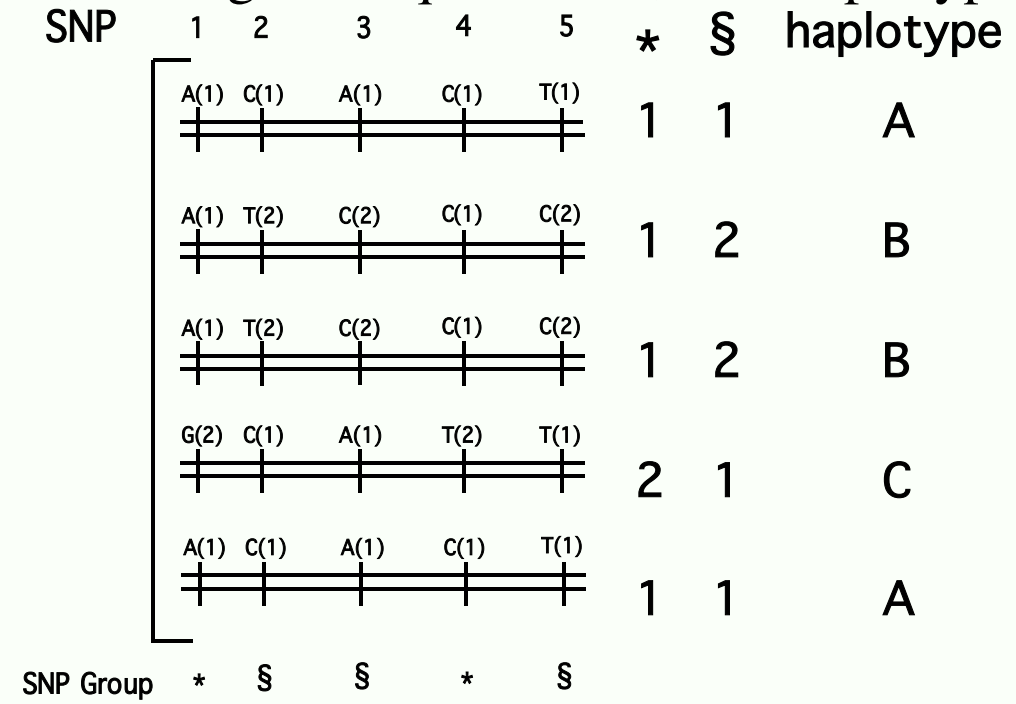
1 = common SNP, 2 = less common SNP

SNPs 1 and 4 are in complete linkage disequilibrium (i.e., correlated) = SNP group *

Likewise, SNPs 2, 3, and 5 are in complete linkage disequilibrium = SNP group §

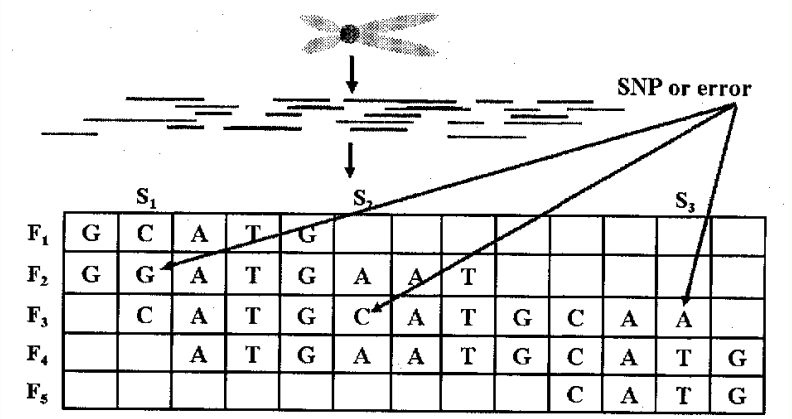
All variation can be sampled by genotyping with marker from each SNP group

SNPs, Linkage Disequilibrium, and Haplotypes



Here there are 5 Loci (SNPs), 2 SNP Groups, and 3 Haplotypes

Haplotype assembly via shotgun sequencing



S_1 S_2 S_3

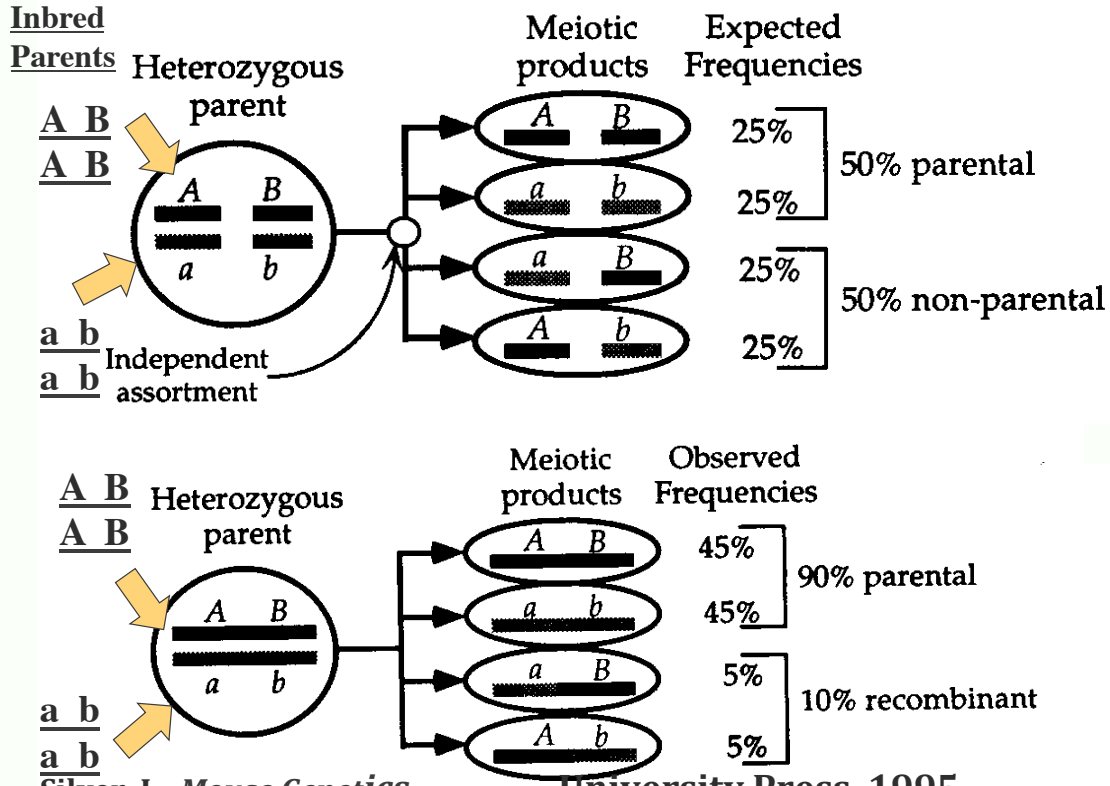
G C A T G C A T G C A A G HAPLOTYPE 1 C C A

G G A T G A A T G C A T G HAPLOTYPE 2 G A A

Huang Y-T, Chang C-J, & Chao K-M (2011) "The extent of linkage disequilibrium and computational challenges of single nucleotide polymorphisms in genome-wide association studies" *Curr. Drug Metabol.* 12:498-506.

Linkage Analysis

- Extremely successful in identifying genes responsible for traits showing Mendelian Inheritance
- Some notable successes in identifying sequence variants that affect susceptibility to common disease
  - *INS* in type I diabetes mellitus
  - *BRCA1* and *BRCA2* in breast cancer
  - *APOE* in Alzheimer's disease
- Problem: identify chromosomal regions, not genes



Silver, L., *Mouse Genetics*, University Press, 1995
[available online at http://www.informatics.jax.org/silverbook/]

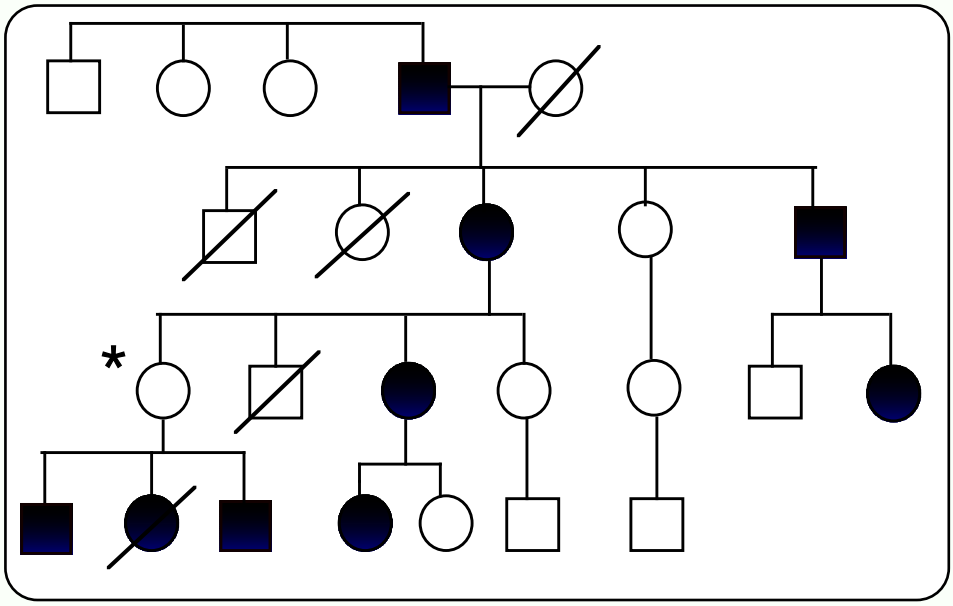
Inheritance of Quantitative Traits in Humans

- Linkage methods harder to employ
- Difficulty finding enough (and large enough) families
- Genetic Heterogeneity:
  - Different genes (loci)--can affect same trait,
  - Different alleles (of same gene)--can have different effects on trait of interest
- Incomplete Penetrance
- Variable Expressivity

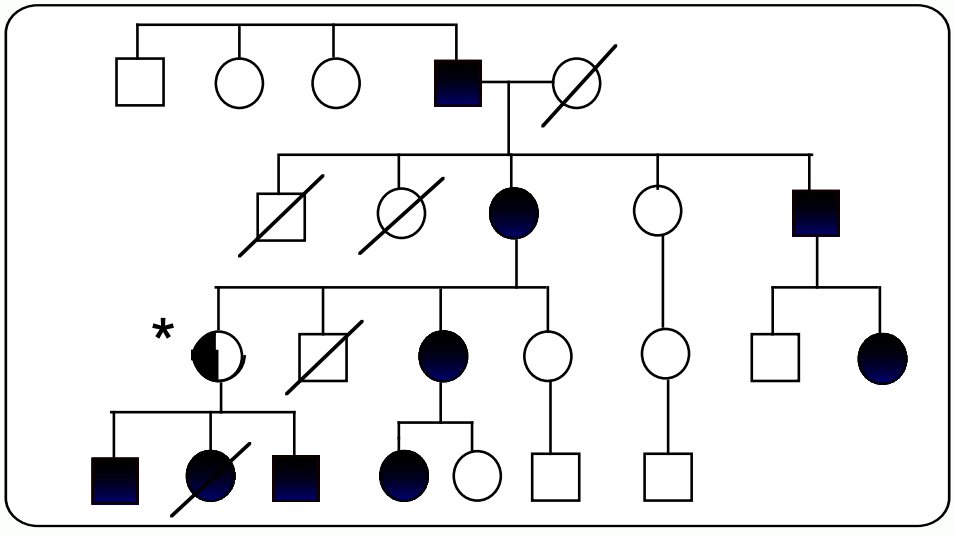
Penetrance can be either. . .

- **Complete: Individuals carrying a defective gene express the mutant phenotype (All Express)**
- **Incomplete: Individuals carrying defective gene (mutant genotype) may or may not express mutant phenotype (All or None)**

Incomplete Penetrance

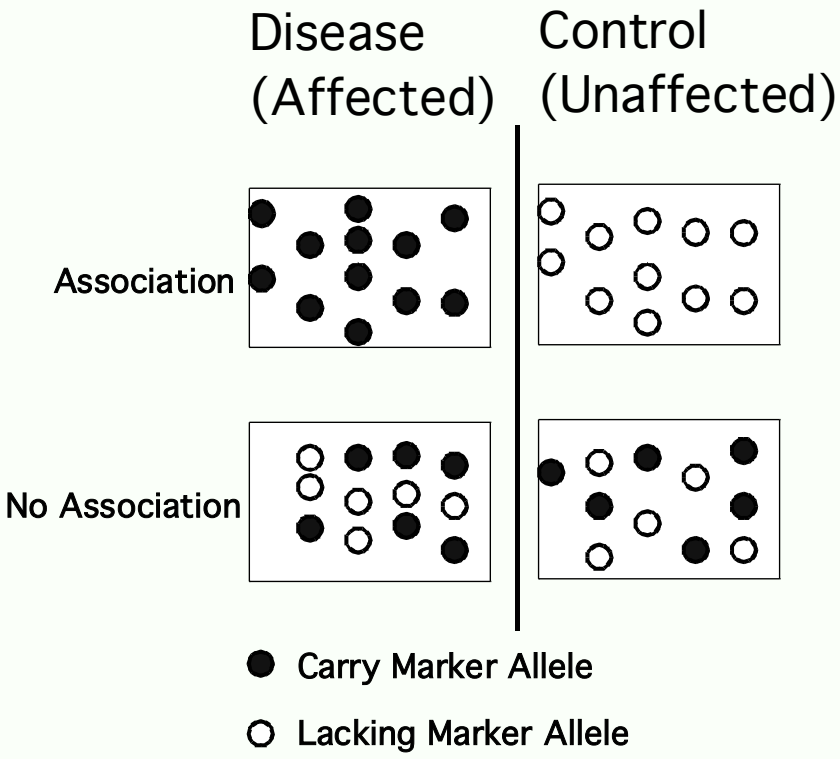


Incomplete Penetrance



Note: while III-1 (marked with *) is unaffected, she must carry the dominant allele

Rationale Underlying Association Studies



Association Between a Marker Locus and a Trait Can Be Due to:

1. Marker allele causing the observed major gene effect.

2. An allele, in linkage disequilibrium with marker locus, causing the major gene effect on a trait-dependent on population history

3. Association by chance in a heterogenous population

Genome-Wide Association Studies (GWAS)

- Used to associate loci with disease trait in populations (instead of families)
- Take advantage of linkage disequilibrium
- Identify candidate genes/loci, but to date, rarely identify causative variant
- Loci identified usually have modest effects

Key Requirements for GWAS

- 1) Sufficiently large sample sizes
- 2) Sufficiently high number of polymorphic markers
- 3) Sufficiently powerful analytic methods

Marker Selection

- Selections that take linkage disequilibrium into account will achieve greater coverage (at least of index population)
- May face difficulties with other populations, especially with populations having African ancestry
- Other markers (*e.g.*, HapMap SNPs) not included in commercial arrays, can have have missing genotype data at untyped variants imputed using data from populations that have been extensively genotyped/ re-sequenced

Types of Genome-Wide Association Studies (GWAS)

- **Case-Control**
- **Cohort**
- **Trio**

Genome-Wide Association Studies (GWAS)

- **Case-Control:** Compare allele frequencies between patients with the disease and a disease-free group
- PRO: Least expensive, easiest to recruit. Optimal for studying rare disease
- CON: Most assumptions (*i.e.* most patients recruited from clinics, may introduce bias)
- Cases and controls need to be from same population

Genome-Wide Association Studies (GWAS)

- **Trio:** includes the affected case participant and both parents. Only offspring phenotyped, only "affected offspring-trio"s studied
- Estimate frequency allele transmitted from heterozygous parent to affected offspring
- Transmission Disequillibrium Test (TDT)
- PRO: Not susceptible to
  - population stratification or
  - genetic differences between cases and controls
- CON:
  - Sensitive to genotyping error, distort transmission,
  - Difficult to recruit, especially for disorders with older ages of onset

Genome-Wide Association Studies (GWAS)

- **Cohort:** collect extensive baseline information on a large group of individuals, then follow through time to identify the affected
- PRO: cases are free of survival bias & more representative of spectrum of disease effects than in case-control studies
- CON: large sample size and long follow-up (expensive). Poorly-suited for rare disease

Characteristics of the Three Classes of Association Studies

Table 1. Study Designs Used in Genome-wide Association Studies

	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) and free of survival bias Direct measure of risk Fewer biases than case-control studies Continuum of health-related measures available in population samples not selected for presence of disease	Controls for population structure; immune to population stratification Allows checks for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

Pearson, TA and Manolio, TA (2008) How to interpret a genome-wide association study. *JAMA* 299(11):1335-1344.

Issues with GWAS

- 1) sample size
- 2) latent population substructure
- 3) family based vs. case-control
- 4) potential to use historical control genotypes to substitute/supplement for newly typed controls

Issues with Case-Control

- **Selection of Subjects for Cases**
  - Enrichment for specific-disease-predisposing alleles
    - Efforts to minimize phenotypic heterogeneity
    - Focus on extreme and/or familial cases
  - to improve study power, especially when have cost-constraints
  - Genetic architecture

Issues with Case-Control

- **Selection of Subjects for Controls**
  - Loss of power if unable to exclude latent diagnoses of phenotype (misclassification)
  - This is bigger problem with common traits such as obesity or hypertension
- Can select "hyper-normal" group for control by applying more stringent selection for cases
  - *i.e.*, early onset or extreme phenotype (while excluding monogenic forms)
  - However, can result in inadvertent selection effects

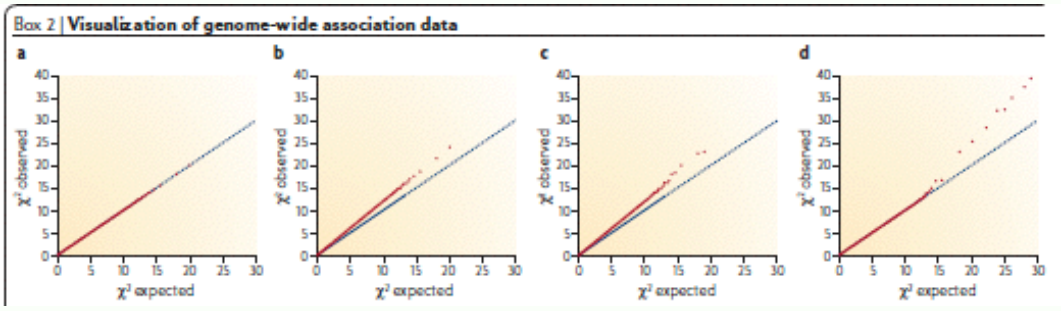
Issues with Case-Control

- **Latent Population Substructure**
  - Can inflate type I error rate
  - Generate claims of spurious "associations" around variants informative of that substructure
- **Population Stratification**: presence of individuals with different ancestral/demographic histories—markers informative of these might be confounded with disease status, leading to spurious "associations"
- **Cryptic Relatedness:** Despite allowances for known family relationships, individuals in sample have residual, non-trivial relatedness. Violates assumptions of independence

Issues with Case-Control

- **Latent Population Substructure**
- Inclusion of parental controls (*i.e.*, family-based association study) best control for this
  - Relatively inefficient vs. case-control
    - Genotype 3 individuals (case, parents) to study 4 alleles (2 transmitted, 2 non-transmitted) vs.
    - Genotype 2 individuals (case, control) to study 4 alleles (2 from case, 2 from control)
- Use Principal components analyses to phenotype unrelated markers throughout genome—define cryptic population substructure

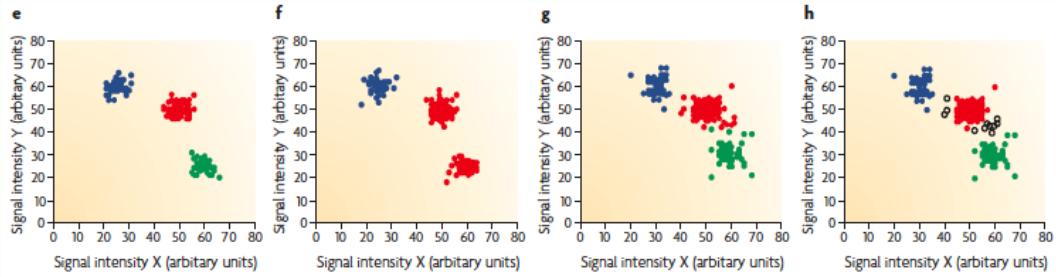
Quartile-Quartile (Q-Q) Plots



a) Observed data closely conforms to expectation—*i.e.*, no evidence of association.
b) Inflation of observed findings across the distribution—*i.e.*, indicative of population stratification or cryptic relatedness.
c) Evidence of population substructure, but with suggestion of excess of strong association.
d) Little evidence of population substructure, but with strong excess of association.
Blue line = null hypothesis, red circles = idealized GWA test results

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356-369.

Signal Intensity (cluster) Plots



These display idealized plots based on ~200 genotypes
e) Three clusters are well-defined and individual genotypes accurately called (*i.e.*, three colors.
f) Clusters are well-defined but allele-calling error leads to two clusters assigned the same genotype.
g) Overlap between clusters result in failure to call certain genotypes (*i.e.*, open circles in h). Here all failed genotypes are homozygotes or heterozygotes for the green allele

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356-369.

Significance in GWAS

- The magnitude of the number of comparisons in a GWAS will result in both
- False Positive Results (Type 1 errors) or
- If multiple comparisons is overly conservative (or power inadequate) –
- False Negative Results (Type 2 errors)

Type I Errors

- Probability of a type 1 error is controlled by setting the significance level, α
- Probability of at least one Type 1 error in a study is a function of both α and the number of observations (n) = $1 - (1 - \alpha)^n$
- thus, for candidate gene studies and small GWAS, unlikely **NOT** to commit a type 1 error

Type I errors

- Is correction for number of SNPs correct? Are all the SNPs independent?
- With linkage disequilibrium, this could lead to over-correction
- If know set of informative SNPs, could correct as follows
- $\alpha_{GWAS} = \alpha / n_{informative}$
- Relationship between SNPs (or statistical testing of SNPs) relate to GWAS studied
  - Variations/alternatives to permutation testing
  - Principal components analysis
  - Analysis of underlying linkage disequillibrium structure in genome

see Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, and O'Brien SJ. (2010) "Accounting for multiple comparisons in a genome-wide association study (GWAS)" *BMC Genomics* 11:724.

Type I Errors

- Typically GWAS in populations of European descent, use Bonferroni correction for an estimated 1 million independent variants in human genome
  - For $\alpha = 0.05$, $P < 5 \times 10^{-8}$
  - For $\alpha = 0.01$, $P < 1 \times 10^{-8}$
- However, such avoidance of type 1 errors could inflate Type 2 errors

- What to do about Type I errors due to multiple comparisons?
- P-value adjustments for multiple comparisons
- Using q values (false discover rate),
- Two-stage analyses
- Genotype imputation

Validation and Replication

- Best practice for determining whether a primary association is reproducible
- Independent replication samples to study
  - Same allele or haplotype (or well-established proxy)
  - Same phenotype
  - Same genetic model
- Otherwise may be testing multiple hypotheses

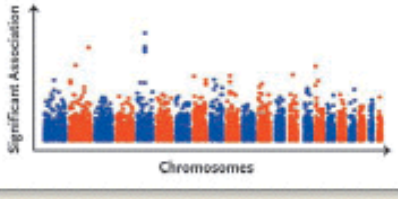
Replication Study Design

- 1) Multi-Stage Design
  - First stage GWAS to identify
  - Second stage to test subset of markers
  - Replication sample size—need to consider "winner's curse", over-estimate of the true effect size of 1° SNPs
- 2) Joint Analysis of two independent populations (best power)
  - Distribute test statistics across data from both stages
  - Best if samples are from similar populations and if there are little genetic heterogeneity differences across the groups

Meta-Analysis

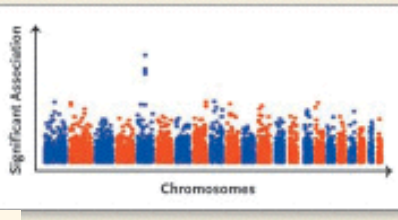
- Single SNP GWAS can be considered a means of identifying preliminary genetic information
- Meta-analysis seeks to pool information from multiple GWAS (with comparable test statistics) to increase the odds of finding true positives
- HapMap data set can be used to combine data from different platforms (*e.g.*, Affymetrix and Illumina)
- Can infer missing genotypes on other platforms by "imputation"

Raw GWAS data (in Manhattan Plot)



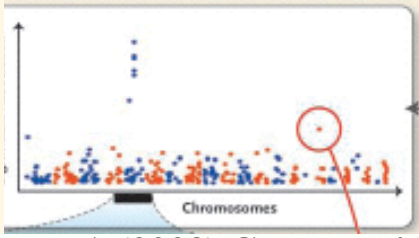
QC and data cleaning

Genome-wide Association with "clean data"



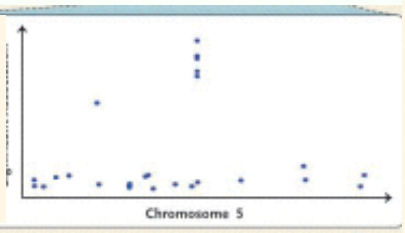
Selection of variants for replication

Test of replication: selected SNPs are genotyped in independent cohort or case-control set

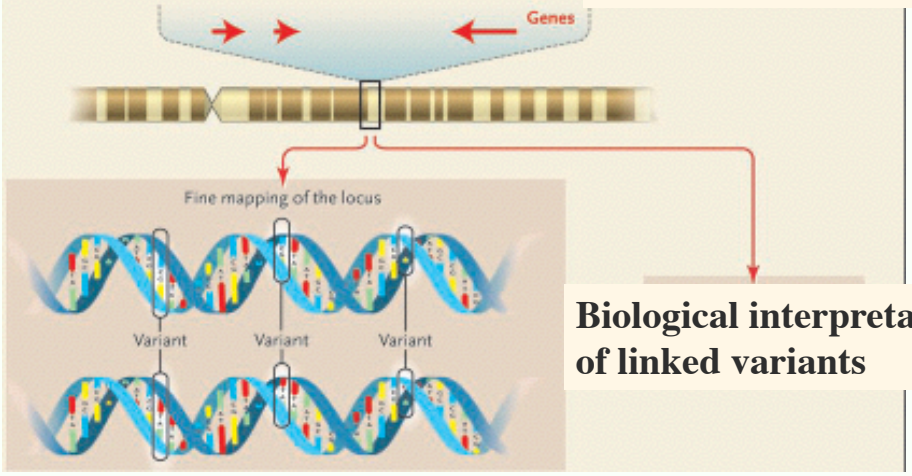


Hardy J and Singleton A (2009) Genomewide association studies and human disease. *N. Engl. J. Med.* 360:1759-1768.

Selection of variants and data mining at an unequivocally associated locus



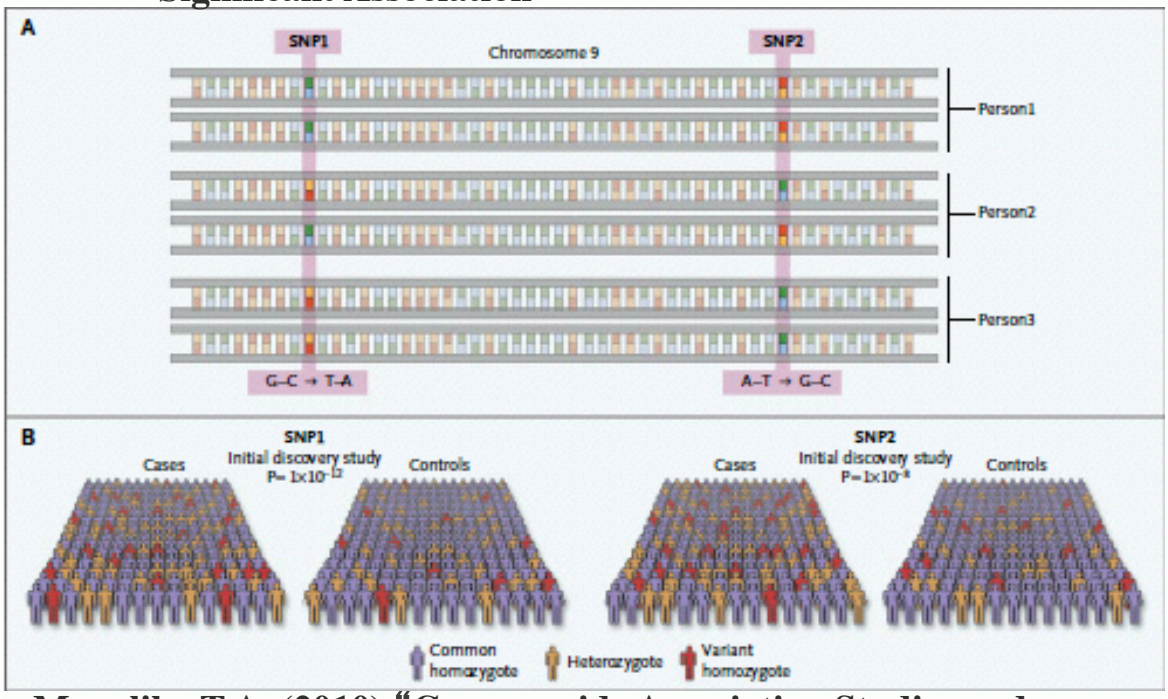
SNPs marking suggestive loci are selected for further tests of replication by genotyping the relevant SNPs in additional cohorts, trios, or case-control sets



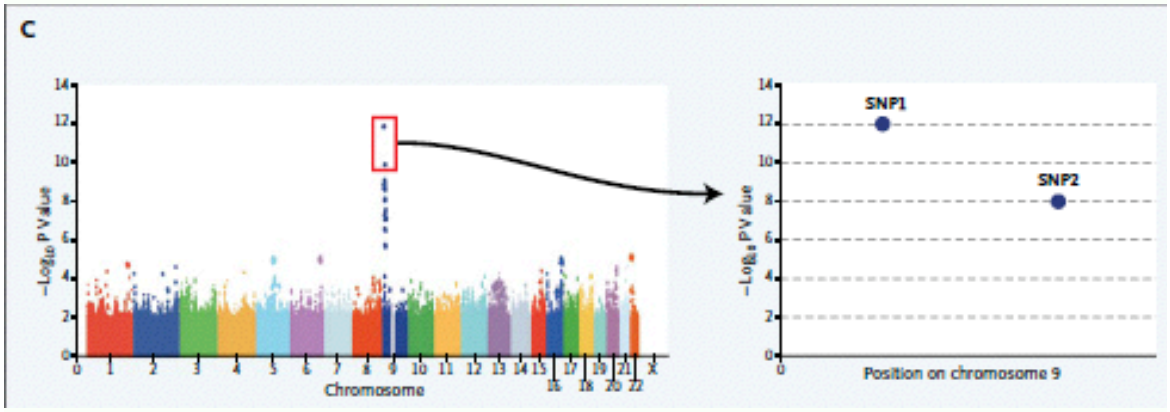
Biological interpretation of linked variants

Hardy J and Singleton A (2009) Genomewide association studies and human disease. *N. Engl. J. Med.* 360:1759-1768.

Three individuals, 2 SNPs in small portion of HSA9
Significant Association

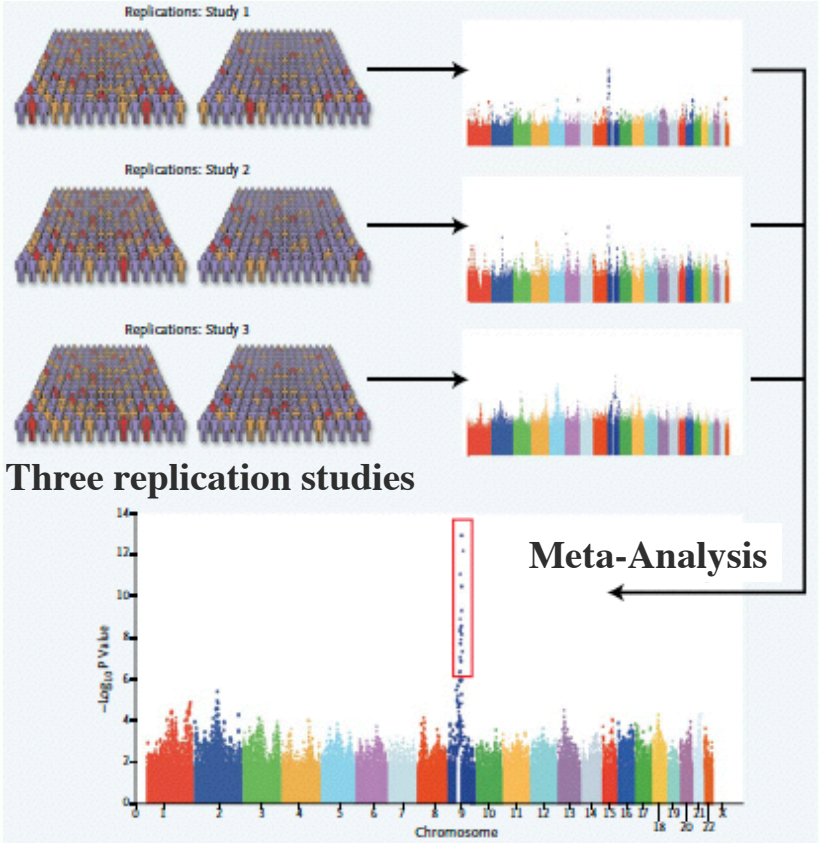


Manolilo, T.A. (2010) "Genomewide Association Studies and Assessment of Risk of Disease." *N. Engl. J. Med.* 363(2):166-176



All SNPs "surviving" a quality control screen are shown (right) the two HSA9 SNPs, having P-values of 10^{-12} and 10^{-9}, respectively

Manolilo, T.A. (2010) "Genomewide Association Studies and Assessment of Risk of Disease." *N. Engl. J. Med.* 363(2):166-176



GWAS and Meta-analysis

Manolilo, T.A. (2010) "Genomewide Association Studies and Assessment of Risk of Disease." *N. Engl. J. Med.* 363(2): 166-176

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

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

The Wellcome Trust Case Control Consortium*

- ~2000 patients for each disease and ~3000 controls (shared with each disease population)
- 24 independent association signals ($P<5x10^{-7}$)
  - 1 for bi-polar, 1 for CAD, 9 for Crohn's, 3 for RA, 7 for Type 1 Diabetes, 3 for Type 2 Diabetes, 0 for hypertension
- 58 loci, with single-point ($10^{-5} < P < 5x10^{-7}$)

Problem: Which SNP is Causative?

- Genome-wide association studies frequently identify associations with many highly correlated single-nucleotide polymorphisms (SNPs) in a chromosomal region,
- due in part to linkage disequilibrium, among SNPs.
- Makes it difficult to determine which SNP (within a group) is the likely causative or functional variant
- Association signals may encompass one gene, multiple genes or be confined to "gene desert"
- If involves protein-encoding gene, might not be a missense mutation, but be a non-coding variant that alters gene expression
- There are vast array of small and large non-coding RNAs
- Regulatory elements may be located 100,000 – 1,000,000 bp from the gene regulated

IL23R and Inflammatory Bowel Disease

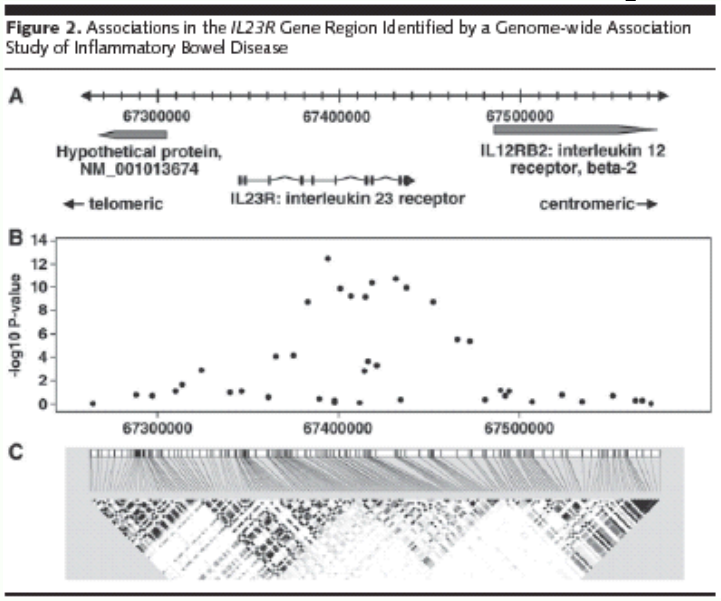


Figure 2. Associations in the *IL23R* Gene Region Identified by a Genome-wide Association Study of Inflammatory Bowel Disease

-log10 *P* values for association with inflammatory bowel disease plotted for each SNP genotyped in the region.Those reaching a pre-specified value of -log10≥7 are presumed to be associated with disease.

Duerr RH, Taylor KD, Brant SR, et al. A genomewide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science*. 2006;314(5804):1461-1463.

IL23R and Inflammatory Bowel Disease

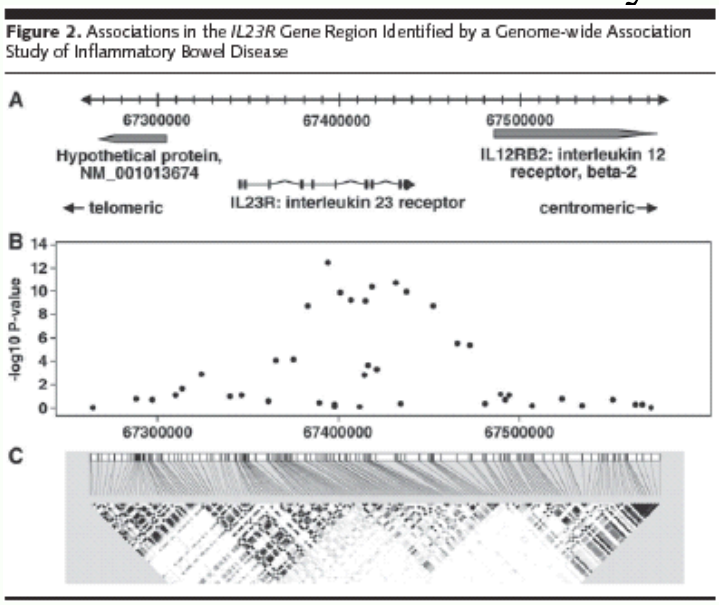
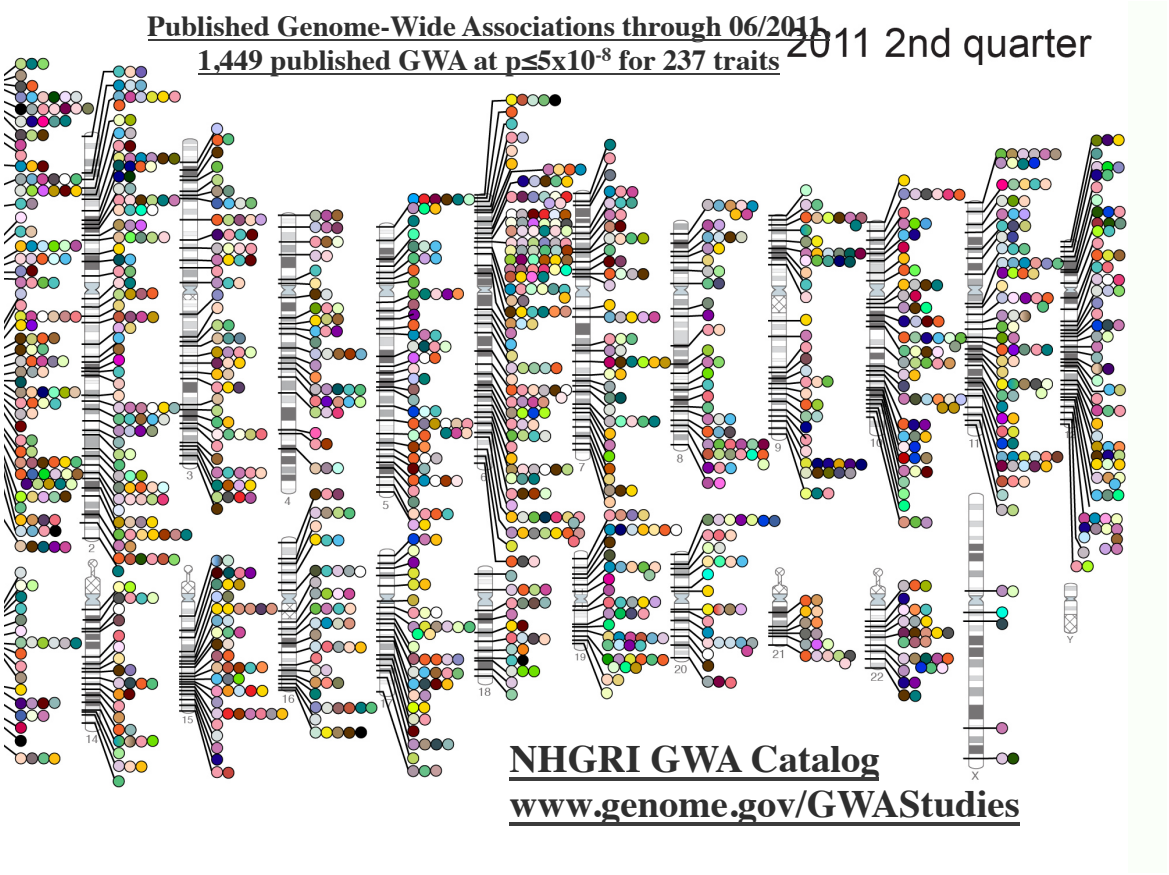


Figure 2. Associations in the *IL23R* Gene Region Identified by a Genome-wide Association Study of Inflammatory Bowel Disease

Pair-wise linkage disequilibrium estimates between SNPs (measured as r^2) are plotted, with higher r^2 values indicated by darker shading. This region contains 4 triangles of linkage disequilibrium. Two *IL23R* linkage disequilibrium regions contain SNPs associated with inflammatory bowel disease.The *IL12RB2* region does not.

Duerr RH, Taylor KD, Brant SR, *et al*. A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science*. 2006;314(5804):1461-1463.



Update: NHGRI-EBI GWAS Catalog

- Welter D, *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42:D1001-D1006
- http://www.ebi.ac.uk/gwas/
- Most up-to-date diagram < http://www.ebi.ac.uk/gwas/diagram >
- Downloadable spreadsheet

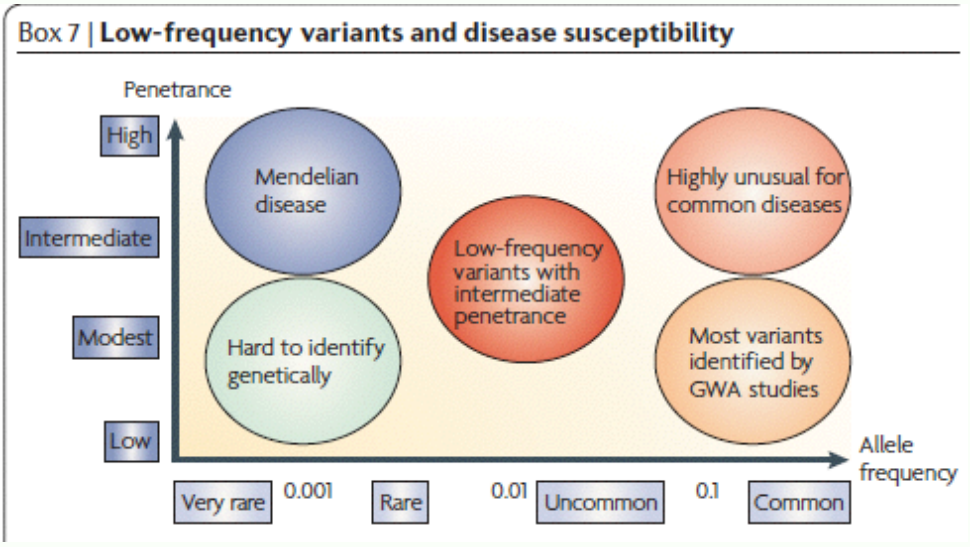
Are the variants responsible for multifactorial diseases rare or common?

- When GWAS began, the common disease – common variant (CDCV) hypothesis dominated
- CDCV now refuted, in light of the "missing heritability problem"
- GWAS currently explain a small amount of the inferred genetic variance for almost all phenotypes examined
  - age-related macular degeneration and
  - type 1 diabetes are exceptions,
  - complement factor H and the major histocompatibility complex variants, respectively, account for ≈ 50% of the attributable risk for both
- Most of the detectable odds ratios are between 1.1 and 1.3 (*i.e.*, common SNPs are in linkage disequilibrium that increase carrier's disease risk between 10-30% over the risk in non-carriers)

Are the variants responsible for multifactorial diseases rare or common?

- As of June 2011 (shown previously), 1,449 GWA with 237 traits/diseases on all chromosomes (excepting the Y)
- While some may be in linkage disequilibrium with rare variants, it is more likely that most are common variants
- Insufficient data to determine now, when more genomes sequenced, will be more clear

Penetrance vs. Disease Susceptibility



Utility of Common (vs. Rare) Allelic Variants

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5):356-369.

Going beyond single-SNP GWAS

- Meta-Analysis
- Epistasis within SNP studies (Variable Expressivity and Reduced Penetrance)
- Pathway Analysis + GWAS
- Copy Number Variant (CNV) polymorphisms
- Next-Generation Sequencing (DNA-Seq, RNA-Seq)
- Gene x Environment Interactions?

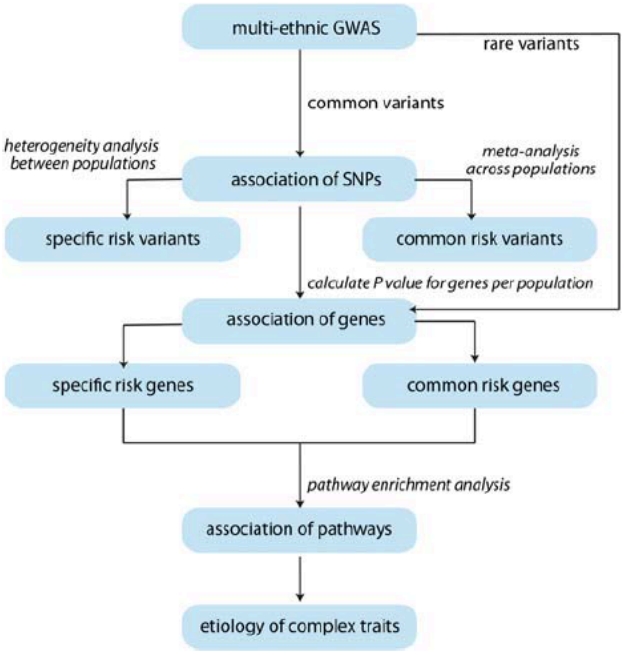


Figure 4. The three-stage framework of a multi-ethnic GWA study.

Fu J *et al*. (2011) "Multi-ethnic studies in complex traits" *Hum. Molec. Genet.* **20**:R206-R213

Overview of Genotyping and Sequencing Technologies

	# of Markers	Advantages	Disadvantages
HapMap based genotyping platform	Several hundred thousand	Definitive identification of disease susceptibility loci for complex diseases; high quality genotypes @ modest cost	Possibility that uncommon, major effect variants not tested
1,000 genomes-based genotyping platform	Several million	More comprehensive assaying of less common variants	Multiple testing burden is increased; genotyping accuracy and statistical power to detect association may be reduced
Genome-wide sequencing of DNA (DNA-Seq)	Exome sequencing (≈35 million bp); Whole genome sequencing (≈3 billion bp)	Individualized, more comprehensive assaying of less common variants	Presently, costs and analytic hurdles are prohibiting widespread use; currently applied to carefully-selected cases
Genome-wide sequencing of RNA (RNA-Seq)	variable	New insight into transcriptome including non-coding RNAs; allele-specific differences in gene expression can be defined	Sequencing space dominated by common transcripts

Adapted from Cho JH (2010) Genome-wide association studies: Present status and future directions, *Gastoenterology* 138:1558-1672.

Benefits of GWAS

- No requirement for initial hypothesis
- Uses digital and additive data that can be mined and augmented without degradation
- Encourages formation of collaborative consortia, can continue with subsequent analyses
- Provides data on the ancestry of each subject, assists in matching case subjects with control subjects
- Provides data on both sequence and copy-number variations

Adapted from Hardy J and Singleton A (2009) Genomewide association studies and human disease. *N. Engl. J. Med.* 360:1759-1768.

Limitations of GWAS

- False positive and false-negative results
- Insensitivity to rare / structural variants
- Requirement of large sample sizes
  - Increasing sample sizes can remedy the first three
- Genotyping errors
  - Confirm different tests (real-time PCR, mass-spec)
- Lack of information on gene function
  - Identifies loci, not genes
- Possible biases due to inappropriate selection of cases and controls
  - Disease heterogeneity ('lumpers' and 'splitters')

Adapted from Wang T-H and Wang H-S (2009) "A genome-wide association study primer for clinicians *Taiwan J. Obstet. Gyencol.* 48 (2):89-95.

Misconceptions of GWAS

- Thought to provide data on all genetic variability associated with disease, when in reality only common alleles with large effects are identified
- Thought to screen out alleles having a small effect size, when in reality such findings may still be very useful in determining pathogenic biochemical / pathophysiological pathways, even though low-risk alleles may be of little predictive value

Adapted from Hardy J and Singleton A (2009) Genomewide association studies and human disease. *N. Engl. J. Med.* 360:1759-1768.

What to look for in a GWAS

- Were phenotyping parameters well-described and defined? Population studied?
- Were cases and controls comparable?
- Was genotyping conducted so that most variation detected? Sufficient QC?
- Was the study large enough to detect associations of modest effect?
- Were expected associations detected (replicating previous results)?
- Was the criterion for significance sufficiently rigorous to prevent detection of spurious associations?
- Were the results replicated in an independent population? Was this population similar in geographic origin? Were the phenotyping parameters similar?

What to look for in a GWAS

- **Was there evidence that the identified gene polymorphism(s) were related to differences in function?**

Box 2. Ten Basic Questions to Ask About a Genome-wide Association Study Report[a]

1. Are the cases defined clearly and reliably so that they can be compared with patients typically seen in clinical practice?
2. Are case and control participants demonstrated to be comparable to each other on important characteristics that might also be related to genetic variation and to the disease?
3. Was the study of sufficient size to detect modest odds ratios or relative risks (1.3-1.5)?
4. Was the genotyping platform of sufficient density to capture a large proportion of the variation in the population studied?
5. Were appropriate quality control measures applied to genotyping assays, including visual inspection of cluster plots and replication on an independent genotyping platform?
6. Did the study reliably detect associations with previously reported and replicated variants (known positives)?
7. Were stringent corrections applied for the many thousands of statistical tests performed in defining the *P* value for significant associations?
8. Were the results replicated in independent population samples?
9. Were the replication samples comparable in geographic origin and phenotype definition, and if not, did the differences extend the applicability of the findings?
10. Was evidence provided for a functional role for the gene polymorphism identified?

[a]For a more detailed description of interpretation of genome-wide association studies, see NCI/NHGRI Working Group on Replication in Association Studies.[28]

Pearson, TA and Manolio, TA (2008) How to interpret a genome-wide association study. *JAMA* 299 (11):1335-1344.

Box 1. Terms Frequently Used in Genome-wide Association Studies

Alleles
Alternate forms of a gene or chromosomal locus that differ in DNA sequence

Candidate gene
A gene believed to influence expression of complex phenotypes due to known biological and/or physiological properties of its products, or to its location near a region of association or linkage

Copy number variants
Stretches of genomic sequence of roughly 1 kb to 3 Mb in size that are deleted or are duplicated in varying numbers

False discovery rate[59,60]
Proportion of significant associations that are actually false positives

False-positive report probability[61]
Probability that the null hypothesis is true, given a statistically significant finding

Functional studies
Investigations of the role or mechanism of a genetic variant in causation of a disease or trait

Gene-environment interactions
Modification of gene-disease associations in the presence of environmental factors

Genome-wide association study
Any study of genetic variation across the entire human genome designed to identify genetic association with observable traits or the presence or absence of a disease, usually referring to studies with genetic marker density of 100 000 or more to represent a large proportion of variation in the human genome

Genotyping call rate
Proportion of samples or SNPs for which a specific allele SNP can be reliably identified by a genotyping method

Haplotype
A group of specific alleles at neighboring genes or markers that tend to be inherited together

HapMap[12,13]
Genome-wide database of patterns of common human genetic sequence variation among multiple ancestral population samples

Hardy Weinberg equilibrium
Population distribution of 2 alleles (with frequencies p and q) such that the distribution is stable from generation to generation and genotypes occur at frequencies of p^2, 2pq, and q^2 for the major allele homozygote, heterozygote, and minor allele homozygote, respectively

Linkage disequilibrium
Association between 2 alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance

Mendelian disease
Condition caused almost entirely by a single major gene, such as cystic fibrosis or Huntington's disease, in which disease is manifested in only 1 (recessive) or 2 (dominant) of the 3 possible genotype groups

Minor allele
The allele of a biallelic polymorphism that is less frequent in the study population

Minor allele frequency
Proportion of the less common of 2 alleles in a population (with 2 alleles carried by each person at each autosomal locus) ranging from less than 1% to less than 50%

Modest effect
Association between a gene variant and disease or trait that is statistically significant but carries a small odds ratio (usually <1.5)

Non-Mendelian disease (also "common" or "complex" disease)
Condition influenced by multiple genes and environmental factors and not showing Mendelian inheritance patterns

Nonsynonymous SNP
A polymorphism that results in a change in the amino acid sequence of a protein (and therefore may affect the function of the protein)

Platform
Arrays or chips on which high-throughput genotyping is performed

Polymorphic
A gene or site with multiple allelic forms. The term *polymorphism* usually implies a minor allele frequency of at least 1%

Population attributable risk
Proportion of a disease or trait in the population that is due to a specific cause, such as a genetic variant

Population stratification (also "population structure")
A form of confounding in genetic association studies caused by genetic differences between cases and controls unrelated to disease but due to sampling them from populations of different ancestries

Power
A statistical term for the probability of identifying a difference between 2 groups in a study when a difference truly exists

Single-nucleotide polymorphism
Most common form of genetic variation in the genome, in which a single-base substitution has created 2 forms of a DNA sequence that differ by a single nucleotide

Tag SNP
A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs so that it can serve as a proxy for these SNPs on large-scale genotyping platforms

Trio
Genetic study design including an affected offspring and both parents

Abbreviation: SNP, single-nucleotide polymorphism.

Pearson, TA and Manolio, TA (2008) How to interpret a genome-wide association study. *J. Am. Med. Assoc.* 299 (11):1335-1344.

Glossary

Annotation catalog: A map denoting the function of specific genomic regions, such as sites to which noncoding RNA or transcription factors bind.
Common disease–common variant hypothesis: The hypothesis that genetic influences on susceptibility to common diseases are attributable to a limited number of variants present in more than 1% to 5% of the population.
Complex condition: A condition caused by the interaction of multiple genes and environmental factors. Examples of complex conditions, which are also called multifactorial diseases, are cancer and heart disease.
Copy-number variation: Variation from one person to the next in the number of copies of a particular gene or DNA sequence. The full extent to which copy-number variation contributes to human disease is not yet known.
Fine mapping: An experimental approach to narrowing a genomewide association signal by typing all known SNPs in the haplotype block containing the tag SNP. If successful, this approach results in the identification of a subsegment of the block that has a stronger association than the surrounding areas.
Gene deserts: Large intergenic regions.
Haplotype: A set of DNA variations, or polymorphisms, that tend to be inherited together. A haplotype can refer to a combination of alleles or to a set of single-nucleotide polymorphisms found on the same chromosome.
Heritability: The proportion of interindividual differences (variance) in a trait that is the result of genetic factors; often estimated on the basis of parent–offspring correlations for continuous traits or the ratio of the incidence in first-degree relatives of affected persons to the incidence in first-degree relatives of unaffected persons.
Intergenic regions: Segments of DNA that do not contain or overlap genes.
Introns: The portions of a gene that are removed (spliced out) before translation to a protein. Introns may contain regulatory information that is critical to appropriate gene expression.
Inversion: A chromosomal segment that has been broken off and reinserted in the same place, but with the genetic sequence in reverse order.
Linkage disequilibrium: An association between two alleles located near each other on a chromosome, such that they are inherited together more frequently than would be expected by chance.
Low-depth coverage: A preliminary strategy in DNA sequencing whereby each base pair is sequenced a minimum of 2 to 4 times rather than the 20 to 30 times that is characteristic of complete (high-depth) sequencing.
Minor-allele frequency: The proportion of the less common of two alleles in a population (with two alleles carried by each person at each autosomal locus), ranging from <1% to <50%.
Noncoding RNAs: Segments of RNA that are not translated into amino acid sequences but may be involved in the regulation of gene expression.
Nonsynonymous single-nucleotide polymorphism: A polymorphism that results in a change in the amino acid sequence of a protein (and may therefore affect the function of the protein).
Rare variant: A genetic variant with a minor-allele frequency of less than 1%. Rare variants are typically single-nucleotide substitutions but can also be structural variants.
RNA interference: The inhibition of gene expression by noncoding RNA molecules.
Single-nucleotide polymorphism (SNP): A single-nucleotide variation in a genetic sequence; a common form of variation in the human genome.
Structural variant: A genetic variant involving the insertion, deletion, duplication, translocation, or inversion of segments of DNA up to millions of bases in length.
Tag SNP: A readily measured SNP that is in strong linkage disequilibrium with multiple other SNPs, so that it can serve as a proxy for these SNPs on large-scale genotyping platforms.
1000 Genomes Project: An international collaboration formed to produce an extensive public catalog of human genetic variation, including SNPs and structural variants and the haplotypes on which they occur.
Transcription factor: A protein that binds to gene regulatory regions in DNA and helps to control gene expression.
Translocation: A chromosomal segment that has been broken off and reinserted in a different place in the genome.

- Manolio TA (2010) Genomewide association studies and the assessment of disease risk. *N. Engl. J. Med.* **363**(2): 166-176.