

Genetic Epidemiology

Association Studies and Power Considerations

Karen L. Edwards, Ph.D.

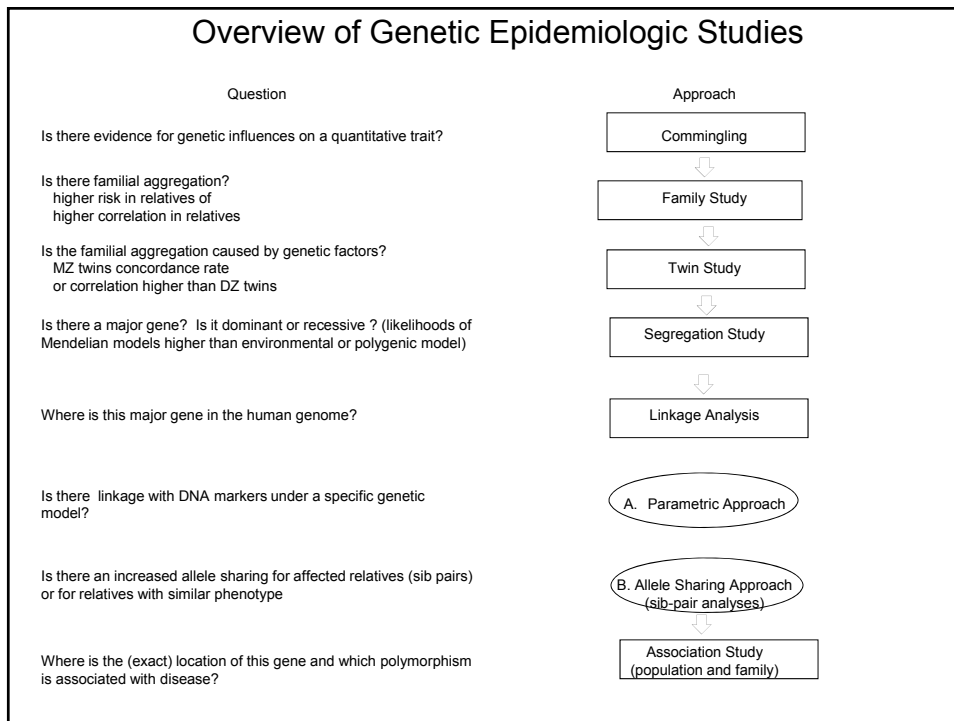
Professor

Department of Epidemiology and
Genetic Epidemiology Research Institute

School of Medicine

University of California, Irvine
Irvine, CA

Overview of Genetic Epidemiologic Studies



Linkage, Review

Cosegregation of two loci in related individuals

- 2 loci are linked if they are transmitted together from parent to offspring more often than expected under law of independent assortment
- During meiosis, recombination occurs with a probability of less than 50% ($\Theta < 0.5$)
- Linkage extends over larger regions of the genome than LD

Good for localization – Not as good at fine mapping

- Marker and disease loci do not need to be in the same gene – we estimate how close they are with theta (Θ)
- One of the most important tools in genetic epi

Linkage Disequilibrium

Linkage Disequilibrium (allelic association)

- 2 loci (alleles) are in LD if across the population they are together on the same haplotype more often than expected by chance
- Depends on Θ (recombination fraction and number of generations)
 - Diminished by a factor of $1-\Theta$ per generation

Foundation on which genetic association studies are based

Complimentary to linkage studies

Epidemiologic Study Design: Review

- Traditional epi studies evaluate the relationship between an exposure and an outcome or disease in a population
- Use a range of statistical methods and approaches to evaluate evidence for the association
 - Odds ratio
 - Relative risk

Epidemiologic Study Design: Review

- Assess a relationship
Exposure → Disease
- Case-control studies
 - Cases are individuals with new disease (incident)
 - Controls are individuals drawn from the same population without disease (population at risk)
 - Selection of cases and controls is very important
 - Need to account for factors that might obscure this relationship
 - Adjust or match for these factors
 - Measure of association in case-control studies is the odds ratio

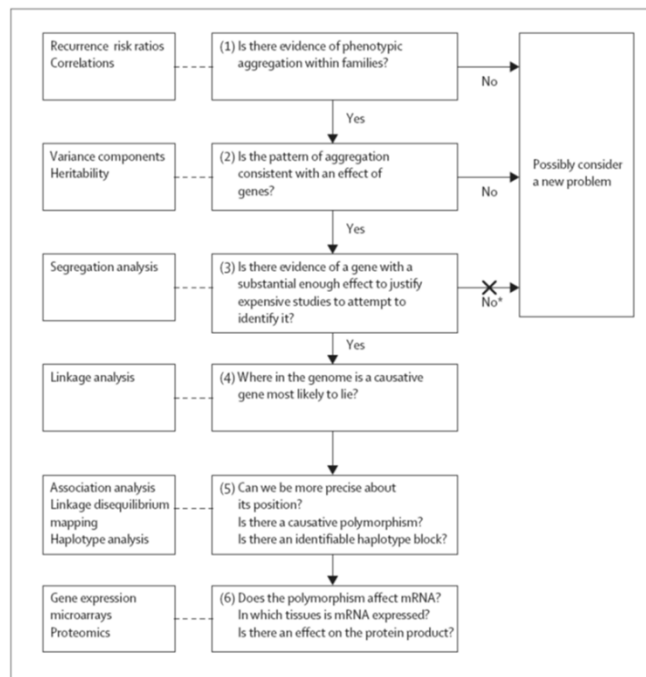
Calculating the Odds Ratio

Contingency (or 2 x 2) Table

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$OR = (a/c) / (b/d)$$

$$= (a*d) / (b*c)$$

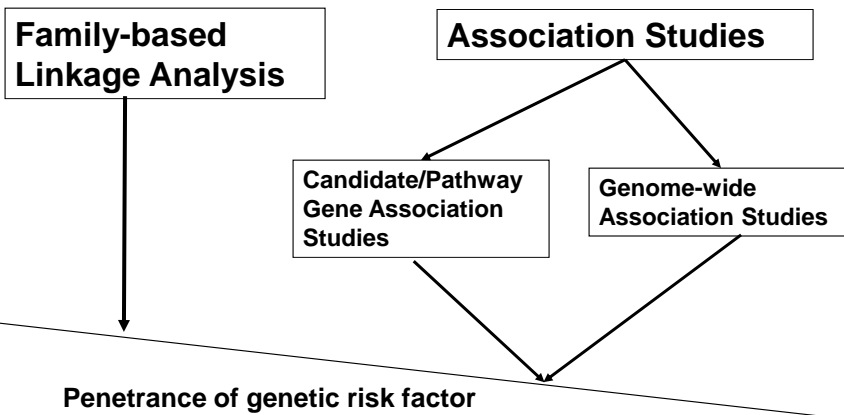


Epidemiology to Genetic Epidemiology

- **Exposure**
- Disease / Outcome
- Some unique challenges in genetic epi studies



Study Design to Investigate Heritability of Common Diseases



Genetic Association Studies: Context

- The search for disease susceptibility genes is conducted using two main methods:
 - The linkage approach in which evidence is sought for co-segregation between a LOCUS and a putative disease locus, using family data
 - linkage analysis is a powerful tool for detecting the presence of a disease locus in a chromosomal region
 - Not efficient at discriminating between small differences in recombination frequency
 - requires data on a large number of informative gametes
 - Genetic Association studies

Genetic Association Studies

- Candidate gene and genome-wide association studies
- Often case-control study design
- Basic idea: Test whether genetic polymorphisms (alleles) are associated with disease status

Association approach

- Evidence is sought for an association between a particular ALLELE and disease in a population
- There should be some evidence that the trait is under genetic control before conducting an association study
- Often used as a followup to linkage to narrow a region of interest (fine mapping), or to evaluate a specific candidate gene(s)

Why Do Association Studies in Unrelated Individuals?

- May be more powerful for detecting loci with smaller effects
- Fine mapping
- Does not require family data
 - Faster
 - Cheaper

Genetic Association Studies

- Despite the popularity, there are many challenges in conducting genetic association studies
 - Interpretation is not always clear
 - Replication has proven difficult
 - Power
 - Gene x environment interactions
 - Gene x gene interactions
 - Confounding
 - Multiple testing

Possible explanations for observing an association

- The marker is part of the pathologic process and is the cause of the disease
 - In this case, the same positive association would be expected to occur in “all” populations
- Linkage disequilibrium (LD) between the marker and the susceptibility gene
 - Usually what we are detecting
- Generally interpreted to mean linkage

Possible explanations for observing an association, cont

- Confounding
 - Genetic ancestry is the most important confounder to consider
 - Population stratification
 - other genetic and environmental factors such as religion, geographic location
- Chance
 - Multiple testing problems with large numbers of markers

Population Structure and Population Stratification

- **Population structure:** heterogeneity in genetic ancestry
- **Population stratification:** systematic difference in population structure between cases and controls
- One form of population stratification is confounding by genetic ancestry

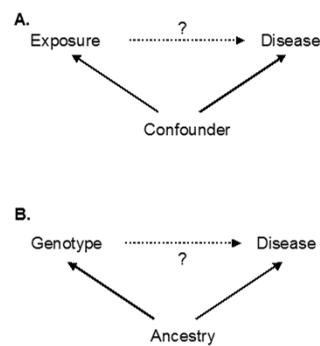


Figure 5-1. Relationship between (A) confounding and (B) population stratification. Directed acyclical graphs (DAGs) demonstrating how a confounding factor is associated with both exposure and disease outcome (A), and, similarly, genetic ancestry can be associated with both genotype and disease outcome (B).

Allele frequencies vary across populations



Humans on the move. Worldwide genetic variation at a neutral marker. Allele frequencies of one randomly chosen microsatellite marker reveal common alleles shared in all populations and the gradual and arbitrary differences in allele frequencies across geographic regions. Populations shown in this example are Yoruba and Bantu (Africa); French, Russians, Palestinians, and Pakistani Brahui (Eurasia); Han Chinese, Japanese, and Yakut (East Asia); New Guineans (Oceania); and Maya and Karitianans (America). From King and Motulsky (2002), *Science*, 298: 2342-2344.

Population Stratification

- Example from Knowler et al.,

Pima ancestry	%Gm Haplotype	Gm Haplotype	%NIDDM
Total (crude)	6.0	Present	8%
		Absent	29%

Population Stratification

- Example from Knowler et al.,

Pima ancestry	%Gm Haplotype	Gm Haplotype	%NIDDM
Total (crude)	6.0	Present	8%
		Absent	29%
None	65.6	Present	17.8%
		Absent	19.9%
50%	42.2	Present	28.3%
		Absent	28.8%
100%	1.6	Present	35.9%
		Absent	39.3%

Methods for dealing with population stratification

- Restrict to homogeneous population
- Family based study designs/analysis
- Adjust associations for substructure and admixture
 - Using self-reported information on race/ethnicity
 - Using unlinked genetic markers
 - Genomic control
 - Structured association
 - Principal Component Analysis

Population Stratification: Bottom Line

- Population stratification is often cited as a major limitation of genetic association studies
 - Does not strike fatal blow for association studies
 - The impact of this form of confounding may have been exaggerated
 - Methods exist for controlling for stratification
 - Should not be ignored
 - Most associations will be small- may be impacted by relatively small amount of confounding
 - Case-control studies should address this issue in their methods and/or discussion.

How do we know if we have confounding in our sample?

- One approach is to evaluate if the marker alleles are in Hardy-Weinberg equilibrium (HWE)
- HW genotype frequencies
 - $p^2 + 2pq + q^2$
 - Depend on allele frequencies
- Evaluate HWE in the control group
 - expect the marker to deviate from HWE among the case group if there is an association between marker and disease, particularly for a rare dominant disease susceptibility allele

Example:

- The locus for red cell phosphatase has three alleles, A, B and C. Based on a random sample of 178 individuals, the frequencies of the genotypes were as follows:

Observed

AA	AB	AC	BB	BC	CC	Total
17	86	5	61	9	0	178

Are these data consistent with HWE?

$f(A)$

$f(B)$

$f(C)$

Possible explanations for a significant deviation from HWE

- Misclassification of alleles/genotype
- Non-random mating in the population
 - one form of non-random mating is population stratification
 - assortative mating
 - consanguineous mating
- Differential survival, natural selection
- Migration, mutation, genetic drift
- Sampling

Basic study design using cohort or case-control approach

Genotype	Cohort		Case-Control		OR
	Disease risk	Relative Risk (RR)	Frequency in cases	Frequency in controls	
NN	I_0	1	A_1	B_1	1
NS	I_1	I_1/I_0	A_2	B_2	A_2B_1/A_1B_2
SS	I_2	I_2/I_0	A_3	B_3	A_3B_1/A_1B_3

N = normal allele, S = susceptibility allele

Alleles vs. Genotypes?

- Can consider the genotype or a particular allele as the exposure of interest
 - Assumes independence (HWE) if using alleles
 - Departures from HWE can affect the Type 1 error rate (false positive), resulting in either an inflated or deflated Type 1 error (Schaid and Jacobsen, AJE 1999;149:706-11).
 - Can correct for deviations from HWE to reduce chance of a false positive association

Interpretation of the OR in Gen Epi Studies

- Odds ratio is used to describe the relationship and strength of the association in epidemiologic studies

- Interpretation of the OR in gen epi studies is similar:

Odds of disease in those with a particular genotype or genetic variant vs. the odds of disease in those with the reference genotype

- Range is the same: 0 to infinity
- However, risks are generally small in genetic epi studies: OR 1.2 – 2.0 are common
 - That is, for an OR=1.2 a particular genotype is associated with a 20% increase odds of disease compared to those with the reference genotype

Summary: Points to Consider

- Maintenance of LD depends on population history and is affected by the recombination fraction (Θ), such that the magnitude of allelic association (disequilibrium) decays at a rate of $1-\Theta$ / generation in a large, stable randomly mating population
 - It is generally accepted that, for most human populations and most regions of the genome, substantial linkage disequilibrium is only likely to occur between loci with a recombination fraction of less than 1%. Thus, LD mapping is most useful for fine mapping over small distances or for recent mutations.

Summary: Points to Consider

- Different alleles maybe associated with disease in different populations
 - random markers can be used, but more meaningful results are often obtained with candidate genes and/or functional mutations
- Adjustment for multiple comparisons is not straightforward
 - Bonferroni correction is considered conservative because markers are not independent, and are often highly correlated
 - False Discovery Rate
 - Staged study designs

Family Based Tests of Association

Family Based Tests of Association

- Family based tests of association are robust to the effects of population stratification
- Associations identified using case-control approaches should be followed-up by a family based test
- One of the first family based tests to be widely used was the Transmission Disequilibrium Test (TDT)
- Many extensions of the TDT have been developed
 - Qualitative traits
 - Quantitative traits

Transmission disequilibrium test (TDT)

- Developed by Spielman et al (1993)
- Not affected by population stratification
- Not affected by departures from HWE
- Uses family data to avoid finding associations due strictly to population stratification
- Provides a test of Linkage AND association for a sample of trios

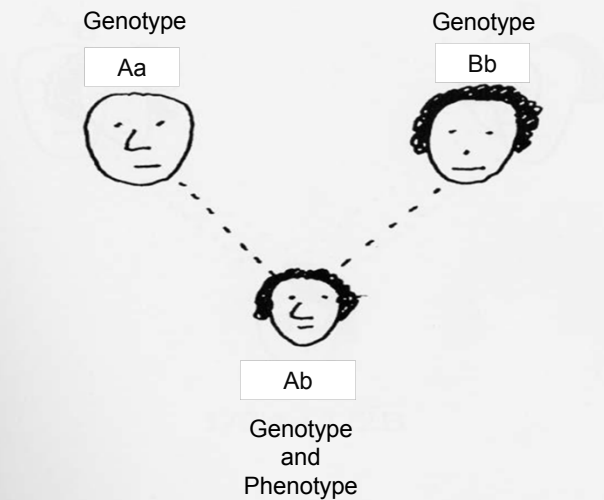
Transmission disequilibrium test (TDT)

- The basic idea behind the classic TDT (and any of its derivatives) is to:
 - look for preferential transmission of a parental marker allele to an affected offspring
 - use non-transmitted alleles from heterozygous parents as "controls"
 - Requires data on trios
 - Trios consist of two parents and an affected offspring
 - Phenotype or disease status of parents is not relevant

Formalities of the TDT

- The data consists of:
 - Genotype information for parents and offspring
 - Phenotype/Disease information for the affected child for the classic TDT
- The hypotheses for data consisting of trios with exactly one affected child are as follows:
 - Ho: no linkage or no association
 - Ha: linkage AND association
- For data containing trios with more than 1 affected child, the hypotheses are:
 - Ho: no linkage
 - Ha: linkage
 - However, the test will only be powerful in the presence of association

TDT: Data Required and General Concept



Extensions of the TDT

- **Extended to many scenarios, including:**
 - multiallelic markers
 - simultaneous use of several markers
 - quantitative traits
 - X chromosome markers
 - pedigrees
 - C-TDT
 - parent of origin effects
 - GxE

Summary: Issues to consider

- Having parental genotype information generally provides more power than using sibship information
 - Only families with heterozygous parents are informative
 - Single SNPs may not be as informative, but will depend on allele frequencies
- Larger sibships provide more information than smaller sibships
- Since association is expected over short distances ($<2\text{cM}$), then it makes sense to either:
 - use a dense set of markers in a specific region of interest OR
 - test markers that have alleles corresponding to functional mutations
 - must also consider the issue of multiple testing

Power and Sample Size Considerations: The Basics

Power and Sample Size

- Critical part of study design
- Can either estimate power or sample size
- Computed by specifying model parameters
 - Can be estimated for Mendelian disorders
 - Generally unknown for complex diseases
- Deal with uncertainty by considering a range of the parameter values
 - Can report “worst-case scenario”
 - Show power over the range of values indicating median power and/or sample size
- Number of software programs

Power and Type 1 Error

- For any question you have 2 hypotheses:
 - H_0 : There is no association between disease x and marker y
 - H_a : There is an association between disease x and marker y
- Power is related to Type 1 Error
- Both give probabilities of positive results, but under 2 different settings (H_0 and H_a)

Power and Type 1 Error

- ▶ Power is the probability that your study will show the association given the alternative hypothesis is true
 - ▶ That is, when H_a is true: There is an association between disease x and marker y
- ▶ Type 1 error is the probability that your study will show the association when the null hypothesis is true
 - ▶ That is when H_0 is true: There is no association between disease x and genotype y)

Degree of LD (r^2) and power

- r^2 impacts power, such that

$$N_2 = N_1 / r^2$$

Where N_1 is the sample size required, and N_2 is the new sample size required

- For example
 - ▶ When $r^2=1.0$,
 - ▶ $N_1 = N_2 = 1,000$
 - ▶ In contrast, when $r^2=0.2$,
 - ▶ $N_2 = 1,000/0.2 = 5,000$

Assumptions for Power Calculations

- Power depends on
 - Linkage disequilibrium (in association studies)
 - Relatedness of individuals (for some designs)
 - Pedigree or family structure
 - Effect size
 - Measurement error (genotype and phenotype)
 - Penetrance
 - Frequency of the high risk allele
 - Genetic model (dominant, recessive, codominant)
 - Prevalence of disease
 - Type of test (allelic, genotypic or trend test)
 - Number of independent tests performed
 - Alpha or type 1 error level

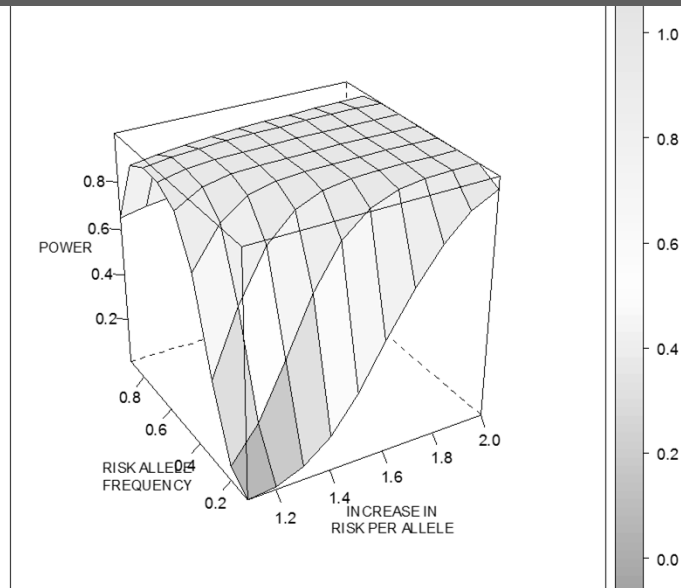
Multiple testing

- Issue of type 1 error (false positive)
- Methods to deal with multiple testing
 - Bonferroni correction (overly conservative with large numbers of markers)
 - False Discovery rates (FDR)
 - Staged study designs

Software Tools

- Genetic Power Calculator (many others, including Quanto)
 - Case-control, TDT and VC linkage
 - Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149-150
- FBAT and PBAT
 - Family based association testing
 - Laird N, Horvath S, Xu X. Implementing a unified approach to family based tests of association. *Genetic Epidemiol* 2000:S36-42.
- Pawe 3D
 - Visualize power for genetic association studies
 - Gordon D, Haynes C, Blumenfeld J, Finch SJ (2005) PAWE-3D: visualizing Power for Association With Error in case/control genetic studies of complex traits. *Bioinformatics* 21:3935-3937.
- CaTS
 - Genetic association studies, GWAS and candidate gene
 - Skol AD, Scott LJ, Abecasis GR, Boehnke M. *Nat Genetic* 2006;38:209-13

Heat Map showing impact of allele frequency and effect size on Power of a genetic association study

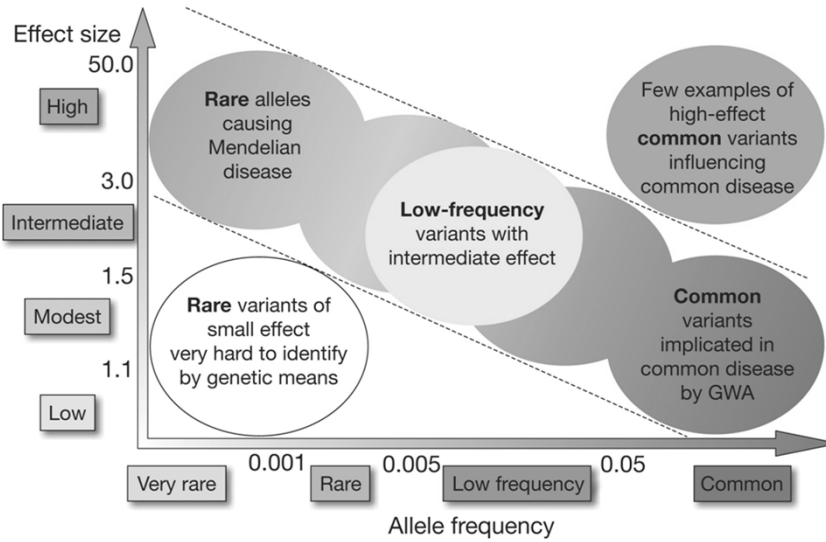


GENOME WIDE ASSOCIATION STUDIES (GWAS)

Outline

- What is a Genome Wide Association Study (GWAS)
- Points to consider in Conducting and Interpreting GWAS
- Post-GWAS Research
- Impact of GWAS findings

Genetic Variation and Disease Susceptibility



Manolio et al. Nature 2009; 461: 747-753

WHAT IS A GENOME WIDE ASSOCIATION STUDY (GWAS)

GWAS DEFINITION

- A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.
- Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.
- Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

<http://www.genome.gov/20019523>

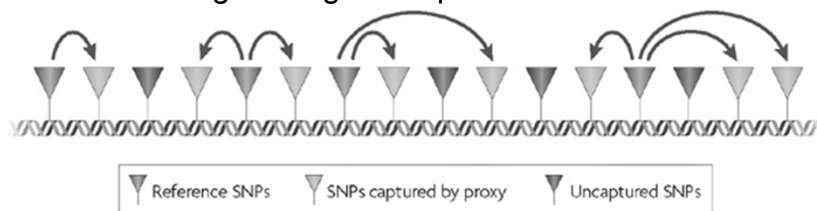
Tools/Discoveries that Made GWAS Possible

- First draft of human genome completed June 2000
- Identification and characterization of common genetic variation
- Advances in genotyping technology, with reduction in costs



Some Key Concepts for GWAS

- Focus on common genetic variants (typically minor allele frequency >5%)
- Single Nucleotide Variants (SNPs) are directly genotyped across the genome
- SNPs that are genotyped will capture unmeasured variants through linkage disequilibrium.



Nature Reviews | Genetics

Kruglyak Nature Reviews Genetics 2008; 9: 314-318.

POINTS TO CONSIDER IN CONDUCTING AND INTERPRETING GWAS

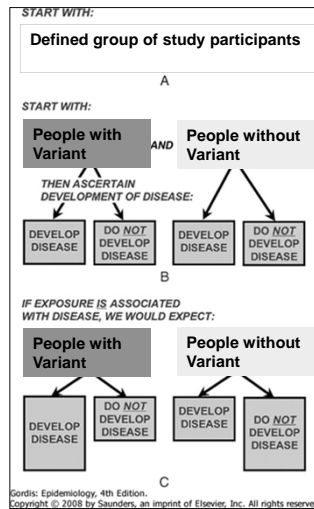
Study Designs Used in GWAS

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

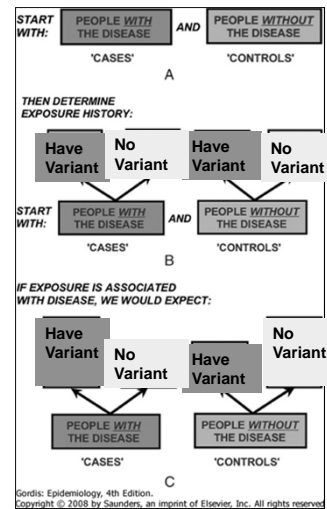
	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly 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 & Manolio JAMA 2008; 299:1335-44

Cohort

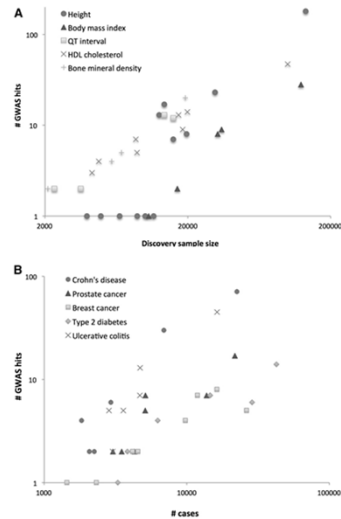


Case-Control



Sample Size

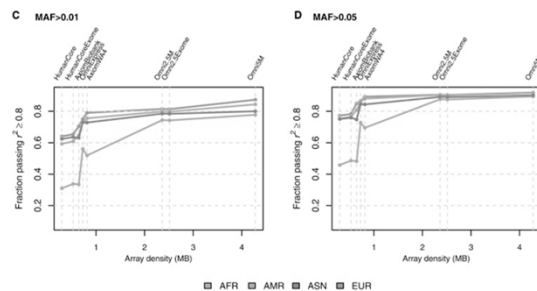
- Variants identified by GWAS have modest effect sizes
- Very large sample sizes are needed to detect variants
- Sample size often achieved through meta-analysis in consortia



Visscher et al. AJHG 2012; 90: 7–24.

Genomic Coverage of GWAS Chips

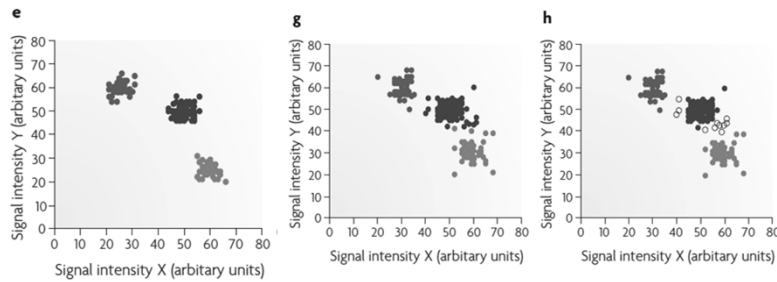
- estimated by the percent of common SNPs having an r^2 of 0.8 or greater with at least 1 SNP on the platform.
- Platforms comprising 500,000 to 1,000,000 SNPs capture ~67-89% of common SNPs in populations of European and Asian ancestry and 46-66% in populations of African ancestry.



Nelson et al. G3 (Bethesda) 2013; 3: 1795–1807.

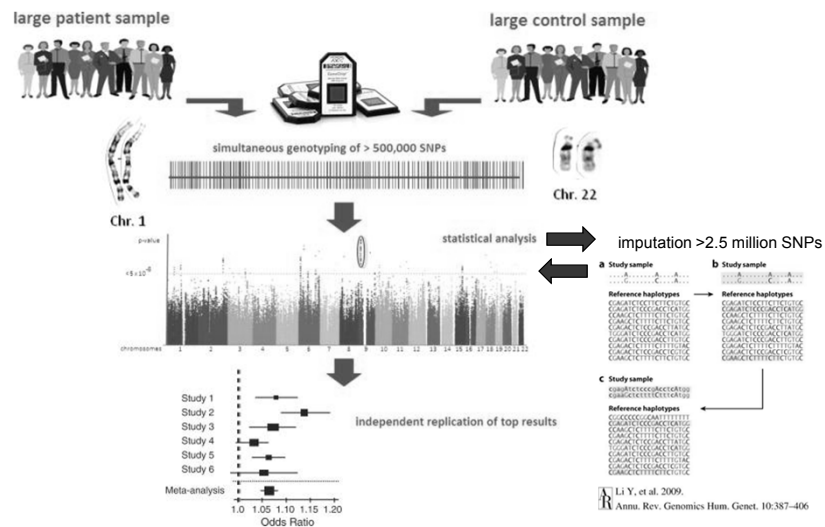
Genotyping and Quality Control in GWAS

- Genotype “calling” is based on intensities for the two alleles at each genetic marker
- Genotyping errors, must be diligently sought and corrected.
- Established quality control features should be applied both on a per-sample and a per-SNP basis.



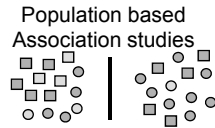
McCarthy et al. Nat Rev Genet 2008; 9:356-369

Schematic of Typical GWAS



Schunkert H et al. Eur Heart J 2010; 31: 918-925

Common Model: Logistic Regression



$$\text{logit}(p) = \alpha + \beta_1 \text{dose} + \beta_{pc} + \dots$$

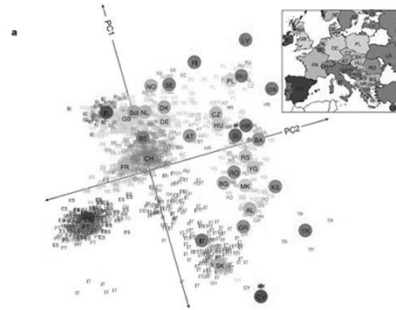
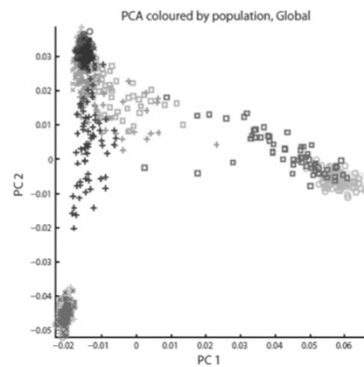
$$OR = e^{\beta_1}$$

- dose=output estimate of # of alternate alleles from imputation
- pc=principle components from principle component analysis (PCA)

Principal components analysis

- A dimensionality reduction technique used to infer continuous axes of variation.
- For GWAS based on SNP x Individual matrix
- The *first principal component (pc1)* is the linear combination of x-variables that has maximum variance
- *pc2* is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with constraint that correlation between pc1 and pc2 is 0
- Continue, with constraint that all pcs are orthogonal
- Standard calculation in programs such as Eigenstrat, Plink, R, etc.

Captures inter- and intra-continental variability



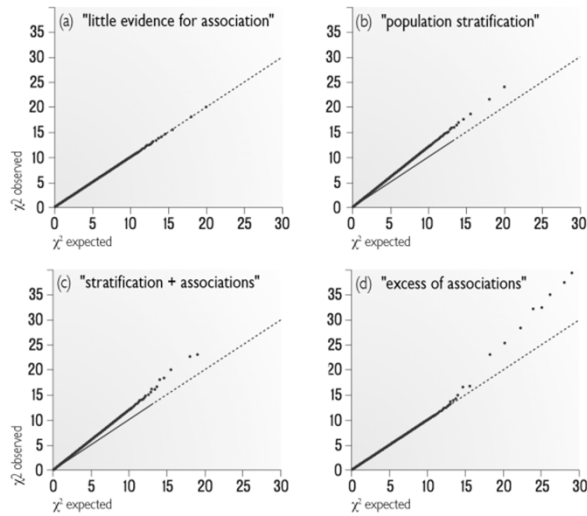
Population Stratification

- Population substructure in GWAS data, because allele frequencies differ in different populations
- Population stratification= confounding by population substructure
 - Example: Lactase gene associated with height in European populations
- Methods can be used to control for population stratification
- Most common method: adjust for top pcs from principle components analysis

$$\text{logit}(p) = \alpha + \beta_1 \text{dose} + \beta_{pc} + \dots$$

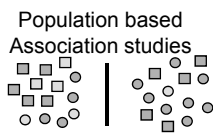
$$OR = e^{\beta_1}$$

Q-Q plots



(modified by Josh Bis from McCarthy et al., Nature Reviews Genetics, May 2008)

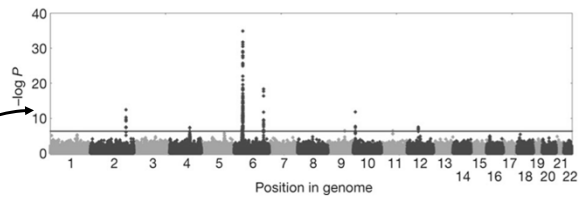
Manhattan Plot



Compare genotypes in cases and controls

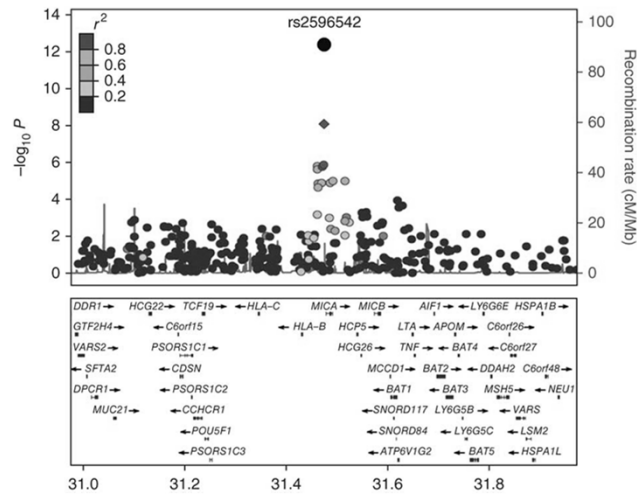
Odds ratio for an allele:
1.35, $p = 6.3 \times 10^{-10}$

$-\log_{10}(p) = 9.2$



Chromosome	Position (Mb)	Gene	$-\log_{10}(P)$
1	100	CCNE1	11.5
1	100	CCNE2	11.5
1	100	CCNE3	11.5
1	100	CCNE4	11.5
1	100	CCNE5	11.5
1	100	CCNE6	11.5
1	100	CCNE7	11.5
1	100	CCNE8	11.5
1	100	CCNE9	11.5
1	100	CCNE10	11.5
1	100	CCNE11	11.5
1	100	CCNE12	11.5
1	100	CCNE13	11.5
1	100	CCNE14	11.5
1	100	CCNE15	11.5
1	100	CCNE16	11.5
1	100	CCNE17	11.5
1	100	CCNE18	11.5
1	100	CCNE19	11.5
1	100	CCNE20	11.5
1	100	CCNE21	11.5
1	100	CCNE22	11.5
1	100	CCNE23	11.5
1	100	CCNE24	11.5
1	100	CCNE25	11.5
1	100	CCNE26	11.5
1	100	CCNE27	11.5
1	100	CCNE28	11.5
1	100	CCNE29	11.5
1	100	CCNE30	11.5
1	100	CCNE31	11.5
1	100	CCNE32	11.5
1	100	CCNE33	11.5
1	100	CCNE34	11.5
1	100	CCNE35	11.5
1	100	CCNE36	11.5
1	100	CCNE37	11.5
1	100	CCNE38	11.5
1	100	CCNE39	11.5
1	100	CCNE40	11.5
1	100	CCNE41	11.5
1	100	CCNE42	11.5
1	100	CCNE43	11.5
1	100	CCNE44	11.5
1	100	CCNE45	11.5
1	100	CCNE46	11.5
1	100	CCNE47	11.5
1	100	CCNE48	11.5
1	100	CCNE49	11.5
1	100	CCNE50	11.5
1	100	CCNE51	11.5
1	100	CCNE52	11.5
1	100	CCNE53	11.5
1	100	CCNE54	11.5
1	100	CCNE55	11.5
1	100	CCNE56	11.5
1	100	CCNE57	11.5
1	100	CCNE58	11.5
1	100	CCNE59	11.5
1	100	CCNE60	11.5
1	100	CCNE61	11.5
1	100	CCNE62	11.5
1	100	CCNE63	11.5
1	100	CCNE64	11.5
1	100	CCNE65	11.5
1	100	CCNE66	11.5
1	100	CCNE67	11.5
1	100	CCNE68	11.5
1	100	CCNE69	11.5
1	100	CCNE70	11.5
1	100	CCNE71	11.5
1	100	CCNE72	11.5
1	100	CCNE73	11.5
1	100	CCNE74	11.5
1	100	CCNE75	11.5
1	100	CCNE76	11.5
1	100	CCNE77	11.5
1	100	CCNE78	11.5
1	100	CCNE79	11.5
1	100	CCNE80	11.5
1	100	CCNE81	11.5
1	100	CCNE82	11.5
1	100	CCNE83	11.5
1	100	CCNE84	11.5
1	100	CCNE85	11.5
1	100	CCNE86	11.5
1	100	CCNE87	11.5
1	100	CCNE88	11.5
1	100	CCNE89	11.5
1	100	CCNE90	11.5
1	100	CCNE91	11.5
1	100	CCNE92	11.5
1	100	CCNE93	11.5
1	100	CCNE94	11.5
1	100	CCNE95	11.5
1	100	CCNE96	11.5
1	100	CCNE97	11.5
1	100	CCNE98	11.5
1	100	CCNE99	11.5
1	100	CCNE100	11.5

Regional Plots

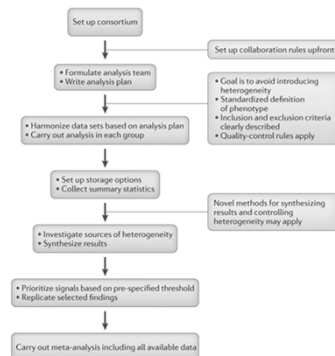
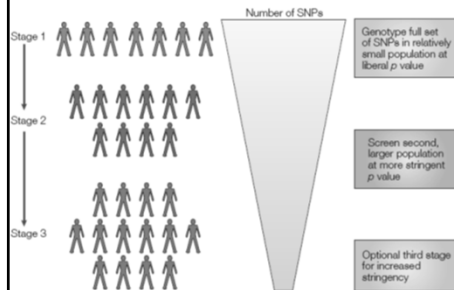


Replicating genotype-phenotype associations

What constitutes replication of a genotype-phenotype association, and how best can it be achieved?

NCI-NHGRI Working Group on Replication in Association Studies

The study of human genetics has recently



Nature Reviews | Genetics

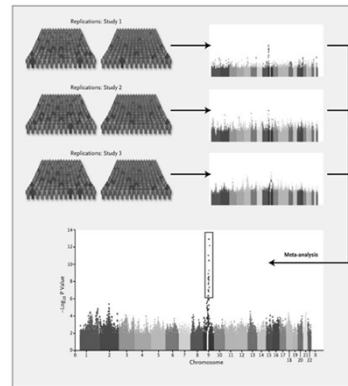
NCI-NHGRI Working Group on Replication in Association Studies. *Nature* 2007; 447:655-660.

Hirschhorn & Daly, *Nat Rev Genet* 2005; 6:95-108.

Evangelou & Ioannidis, *Nat Rev Genet* 2013; 14: 379-389.

Meta-Analysis

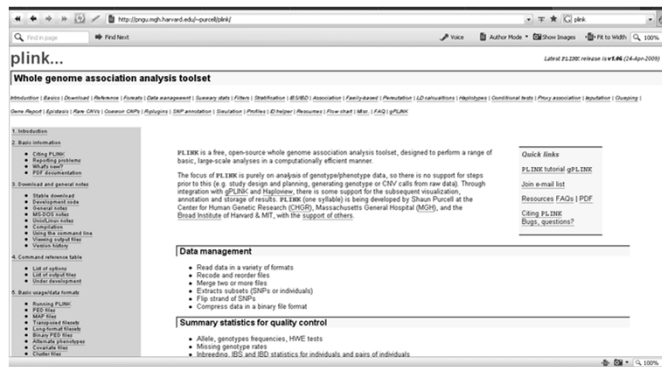
- Large sample sizes required because of small effect sizes, p-value threshold, misclassification inherent in using tagSNPs, etc.
- Meta-analysis are often used to combine information across studies.
- Meta-analysis combines information across studies, creating a weighted average of study specific estimates.



Plink

<http://pngu.mgh.harvard.edu/~purcell/plink/>

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.



Limitations of GWAS

- Countries of recruitment dominated by Europe and North America.
 - Starting to be addressed through studies of other ancestries.
- Possible biases due to case and control selection and genotyping errors
 - Addressed through standards for study design, QC and analysis
- The potential for false-positive results
 - Addressed through replication, meta-analysis and the use of strict genome-wide significance thresholds.
- Lack of information on gene function
 - Addressed through post-GWAS functional follow-up studies
- Insensitivity to rare variants and structural variants
 - Addressed through alternative study designs

“Missing” Heritability

FEATURE PERSONAL GENOMES NATURE 464



the case of the missing heritability

Scientists opened up the human genome, they expected to find the genetic component on traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on places where the missing loot could be stashed away.

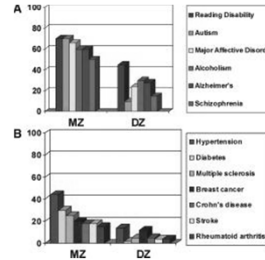
- Additional studies needed
- Impact of other types of genetic variation
 - Less common/rare variants
 - Copy number and structural variants
 - Epigenomic variability
- Interactions (effect modification)
 - Gene-environment
 - Gene-gene (pairwise and networks)
- Limitations of study design and disease definitions
- Current heritability estimates may be overestimated

Maher Nature 2008; 456: 18-21.
Manolio et al. Nature 2009; 461: 747-753.

Assess the Heritability of a Trait

Twin Studies

- Compare trait in monozygotic and dizygotic twins
- Greater concordance in monozygotic twins reflects genetic similarity

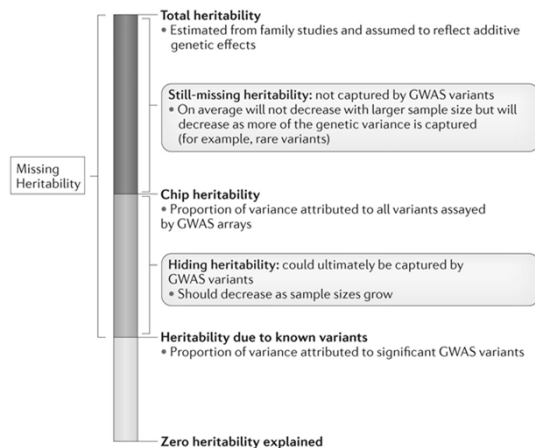


Wong A H et al. Hum. Mol. Genet. 2005;14:R11-R18

Cancer Site	Heritable Factors	Environmental Factors	
		Shared	Non-shared
Prostate	0.42 (0.29-0.50)	0 (0-0.09)	0.58 (0.50-0.67)
Colorectal	0.35 (0.10-0.48)	0.05 (0-0.23)	0.60 (0.52-0.70)
Bladder	0.31 (0.00-0.45)	0 (0-0.28)	0.69 (0.53-0.86)
Breast	0.27 (0.04-0.41)	0.06 (0-0.22)	0.67 (0.56-0.76)
Lung	0.26 (0.00-0.49)	0.12 (0-0.34)	0.62 (0.51-0.73)

Source: Scandinavian Twin Registry, Lichtenstein et al. New Engl J Med 2000

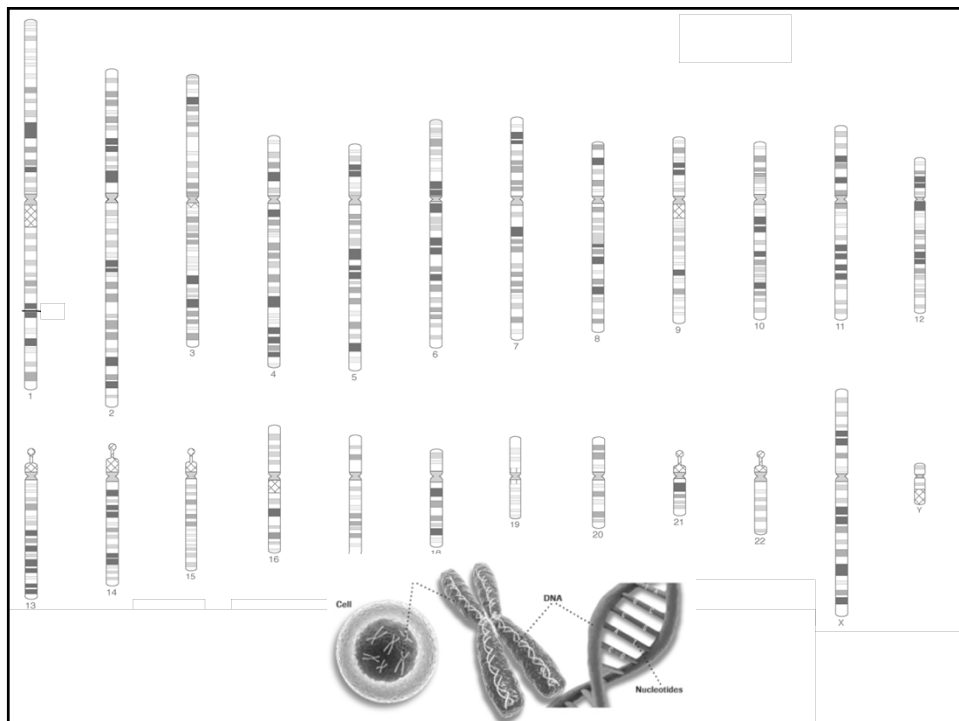
Contribution of Genetic Variants to Disease Heritability

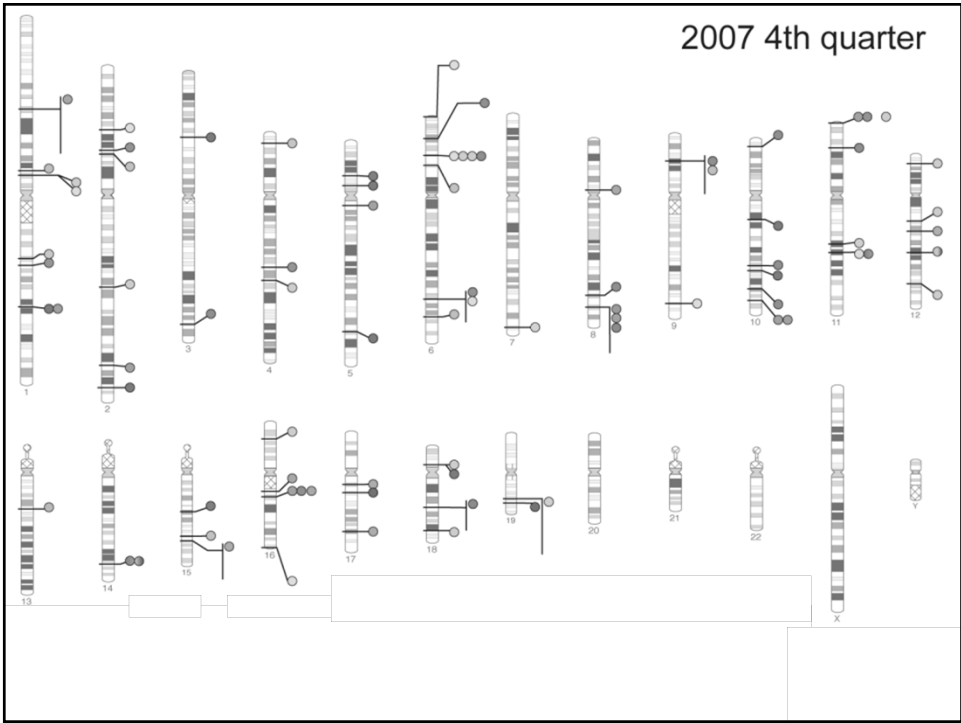
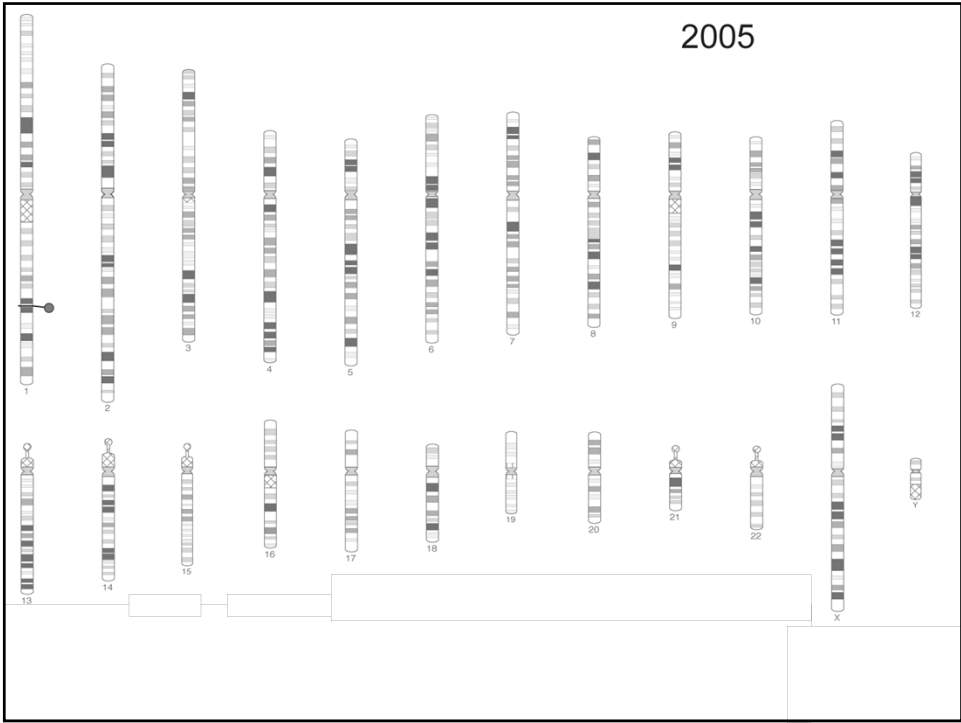


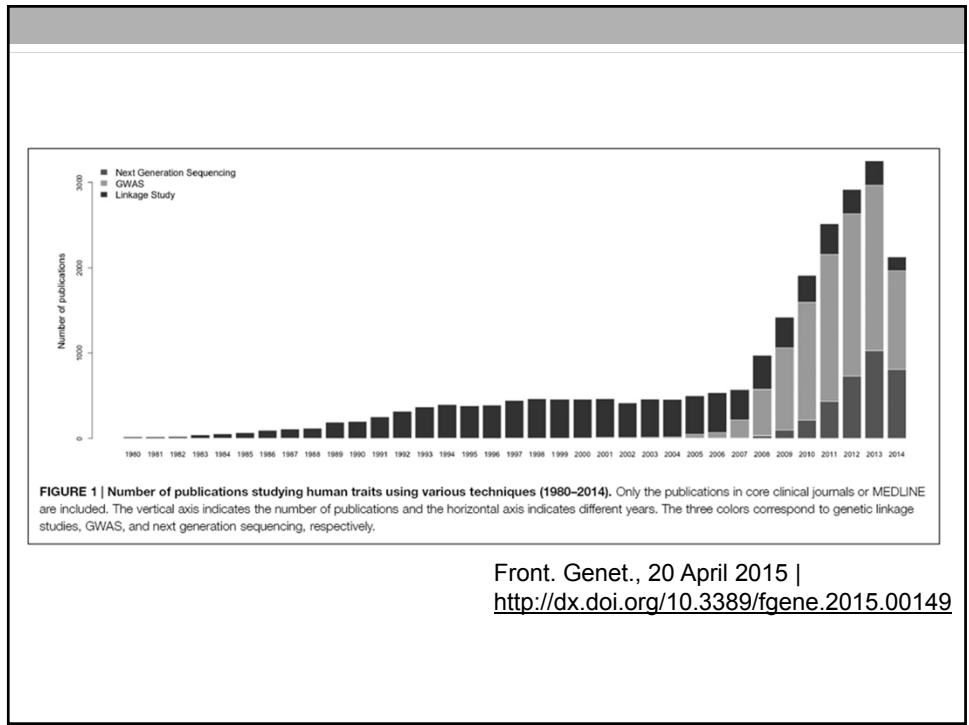
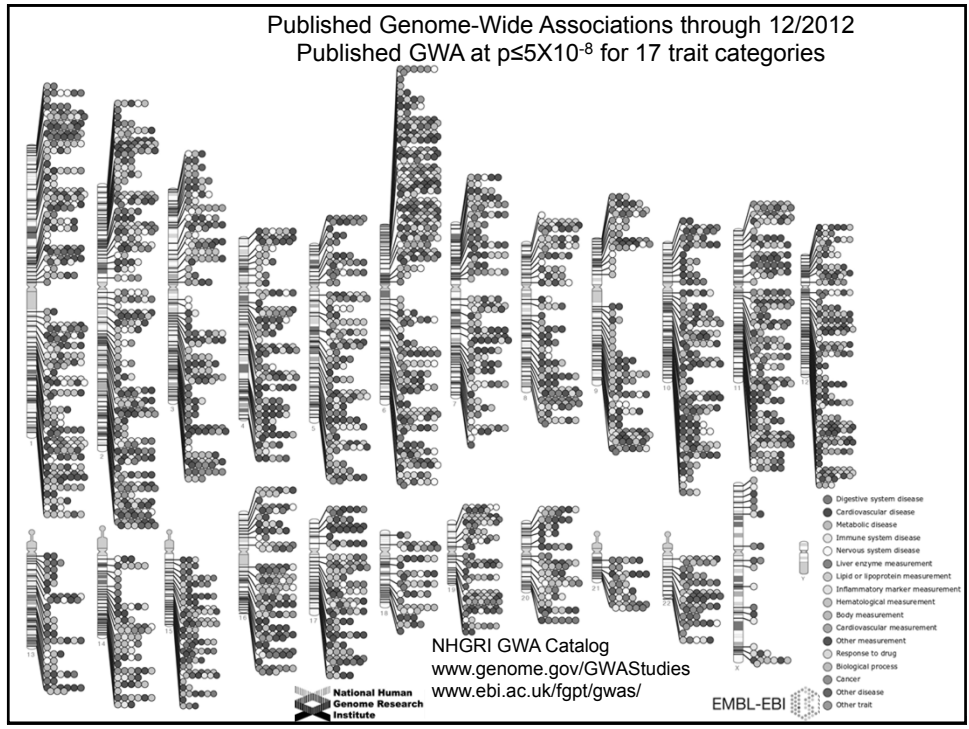
Nature Reviews | Genetics

<http://www.nature.com/nrg/journal/v15/n11/full/nrg3786.html>

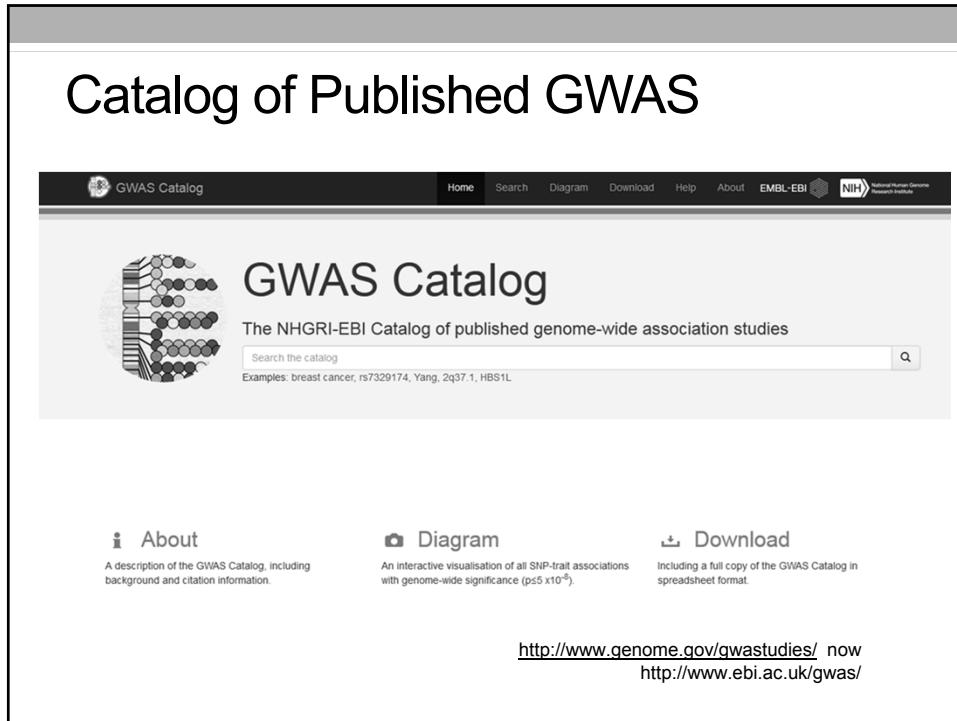
GWAS FINDINGS







Catalog of Published GWAS



The screenshot shows the top portion of the GWAS Catalog website. At the top, there is a navigation bar with the following items: "GWAS Catalog" (with a logo), "Home", "Search", "Diagram", "Download", "Help", "About", "EMBL-EBI", and "NIH National Human Genome Research Institute". Below the navigation bar is a large header area. On the left is a circular logo composed of a grid of dots of varying shades of gray. To the right of the logo, the text reads "GWAS Catalog" in a large font, followed by "The NHGRI-EBI Catalog of published genome-wide association studies" in a smaller font. Below this text is a search input field with the placeholder text "Search the catalog" and a search button icon. Underneath the search field, there are example search terms: "Examples: breast cancer, rs7329174, Yang, 2q37.1, HBS1L". At the bottom of the header area, there are three columns of links, each with an icon and a brief description:

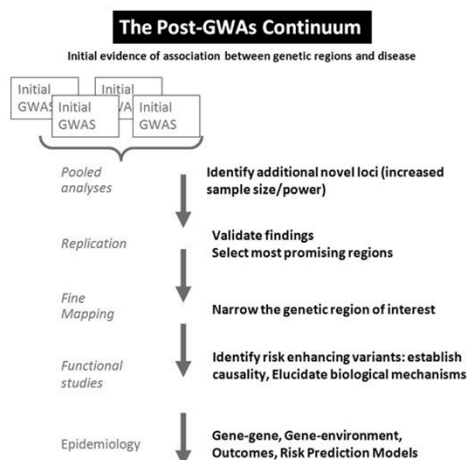
- About** (info icon): A description of the GWAS Catalog, including background and citation information.
- Diagram** (camera icon): An interactive visualisation of all SNP-trait associations with genome-wide significance ($p \leq 5 \times 10^{-8}$).
- Download** (download icon): Including a full copy of the GWAS Catalog in spreadsheet format.

At the bottom right of the page, there are two URLs: <http://www.genome.gov/gwastudies/> and <http://www.ebi.ac.uk/gwas/>.

POST-GWAS RESEARCH

The Post-GWAS Continuum

- Follow-up studies and analysis to capitalize on and expand GWAS findings
- Each step builds on the knowledge gained from the preceding studies



Nature iCOGS

- Large scale results for breast, ovarian and prostate cancer
- Collaborative Oncological Gene-environment Study (COGS)
- Published Online March 27, 2013
- Simultaneous publication of 13 papers, commentaries, editorials and hypertexted essays. Includes:
 - Commentary: Public health implications from COGS and potential for risk stratification and screening
 - Primer: Risk prediction and population screening for breast, ovarian and prostate cancers

nature | iCOGS

Home | Primers | About | Sponsor

Foreword
iCOGS collection provides a collaborative model
Nature Genetics is pleased to present the iCOGS Focus comprising a collection of 13 papers from COGS, representing a significant advance in our understanding of genetic susceptibility to three hormone-related cancers: breast, ovarian and prostate cancer. We hope that you will find this Focus issue, as well as the accompanying F Focus online, a useful guide to this milestone in genetic epidemiology.

Research Highlights

Breast and ovarian cancer in BRCA1 mutation carriers | Breast cancer in BRCA2 mutation carriers | Breast cancer associations in east Asian women | Fine mapping of the 11q13 breast cancer susceptibility locus | Fine mapping of prostate cancer associations at TERT locus

Commentary

Public health implications from COGS and potential for risk stratification and screening
Commentary

Turning of COGS moves forward findings for hormonally mediated cancers
Commentary

Research articles

Large-scale genotyping identifies 41 new loci associated with breast cancer risk
Douglas Easton, Pip Hall et al.

Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array
Reinhold Erbe et al.

GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer
Paul Pharoah, Joellen Schildkraut, Thomas Stiles et al.

Primers

These Primers, or hypertext essays, provide a guided tour through the entire collection of 13 coordinated COGS publications. This new publishing format interfaces editorial analysis with threads, which include a series of direct quotations from relevant sections of the original research publications or editorially written highlights.

Common variation and heritability estimates for breast, ovarian and prostate cancers

Shared susceptibility loci for breast, prostate and ovarian cancers

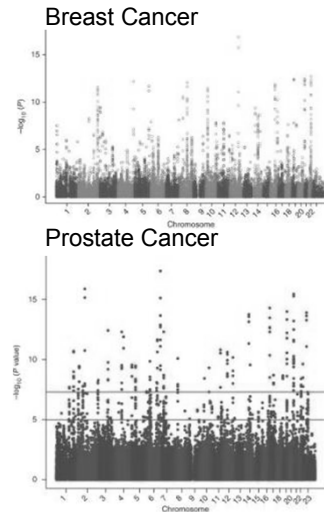
Key COGS Findings

Breast Cancer

- GWAS meta-analysis of 10,052 cases and 12,575 controls
- Replication in 45,290 cases and 41,880 controls
- Identified 41 new loci
- Top 5% and 1% of risk distribution have 2.3 fold and 3 fold higher risk than average population.

Prostate Cancer

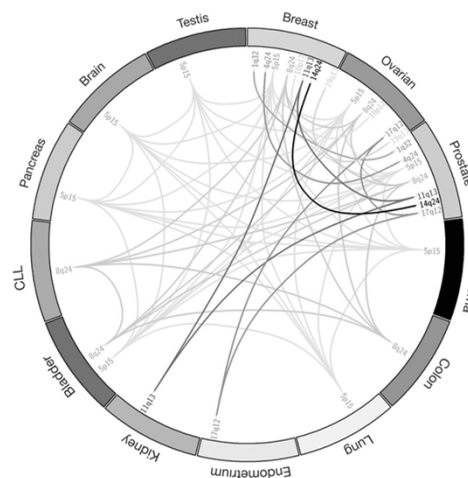
- GWAS meta-analysis of 11,085 cases and 11,463 controls
- Replication in 25,074 cases and 24,272 controls
- Identified 23 new loci
- Top 1% of risk distribution has 4.7 higher risk than average population.



Michailidou et al. Nature Genetics 2013; 45, 353–361.
Eeles et al. Nature Genetics 2013; 45, 385–391.

Pleiotropy in COGS and other Cancer GWAS

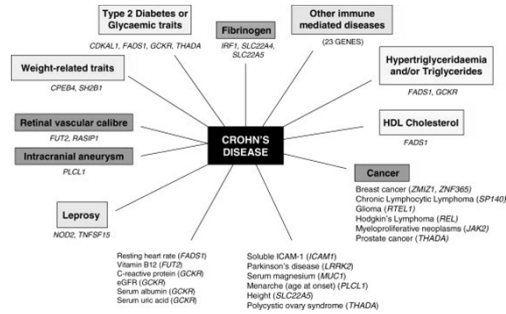
- Pleiotropy= a single locus influencing two or more traits.
- Several findings from GWAS are shared among different cancer types.



Sakoda et al. Nature Genetics 2013; 45, 345–348.

Pleiotropy in GWAS

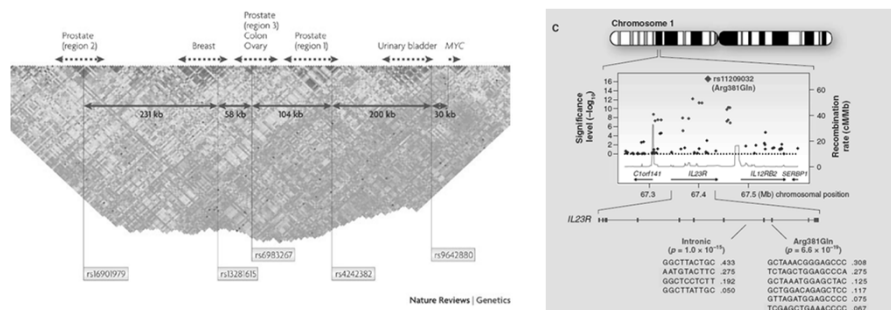
- Numerous examples of GWAS findings impacting more than one trait.
- Pleiotropy scans can identify novel loci.



Am J Hum Genet. Nov 11, 2011; 89(5): 607-618.

Fine-Mapping

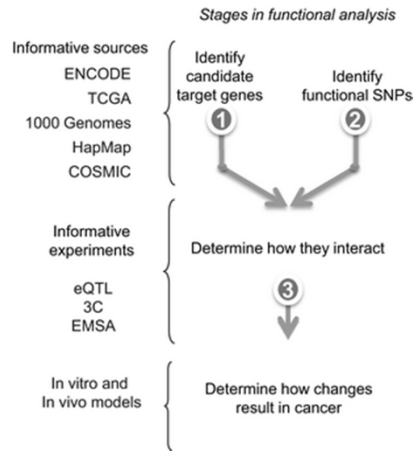
- Genotype additional SNPs to narrow down the region of interest
- Targeted resequencing, to gain additional information on sequence variation in the area of interest



Ioannidis et al. Nat Rev Genet 2009; 10: 318-329.
Altshuler Science. 2008; 322: 881-8.

Post-GWAS Biological Studies

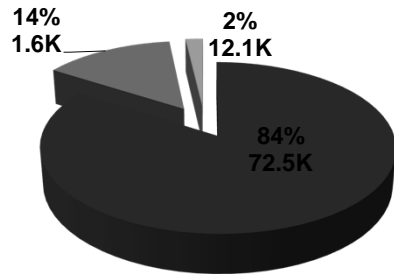
- Identification of risk-modifying variants
- Determination of biological mechanism of risk-enhancement
- Examination of functional consequences of variant



Monteiro & Freedman. J Int Med 2013; 274: 414-424.

Mendelian Traits

Human Genetic Mutation Database (HGMD)

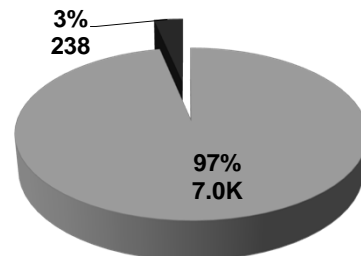


- Missense/Nonsense
- Splicing
- Regulatory

<http://www.hgmd.cf.ac.uk/>

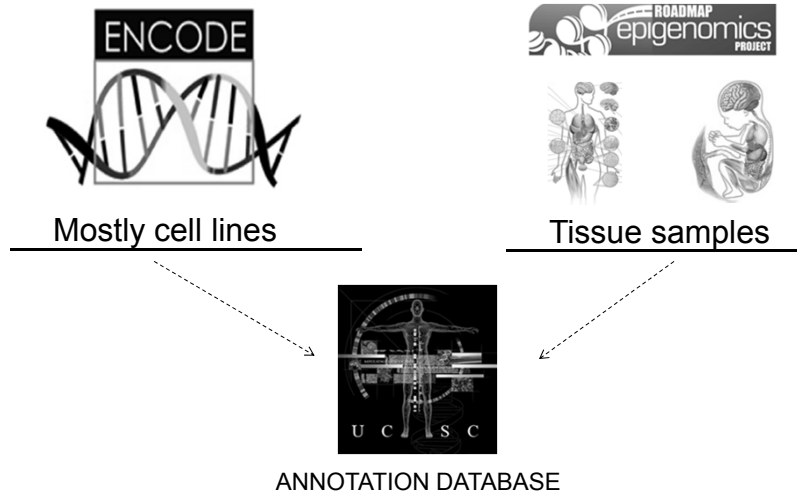
Complex Traits

Genome-Wide Association Study (GWAS) Catalog

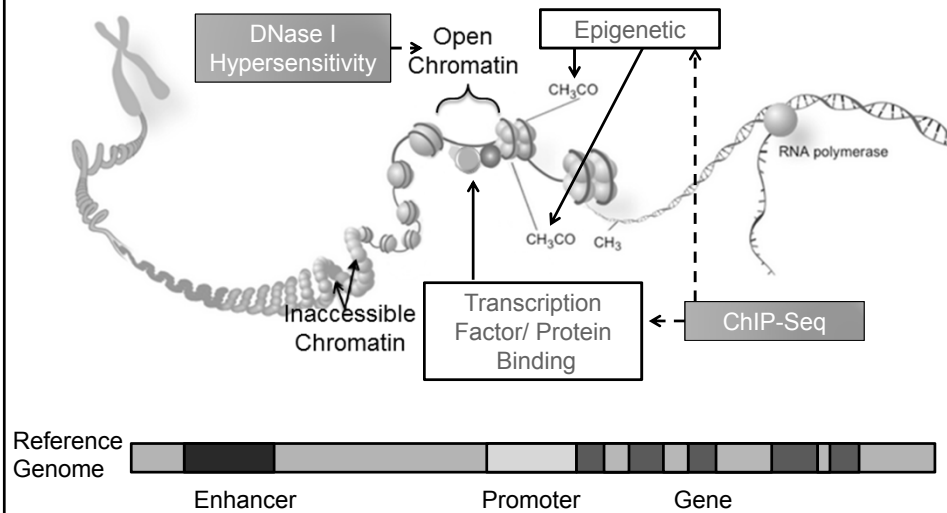


<http://www.genome.gov/gwastudies/>

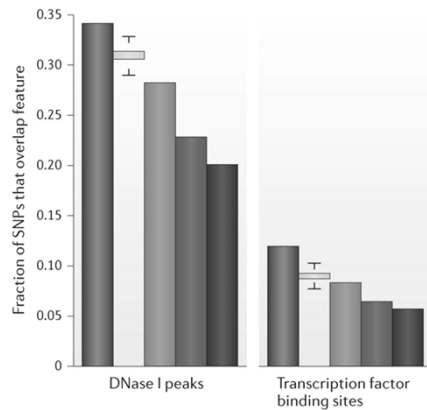
Genome-Wide Functional Annotation



Functional Attributes of Regulatory Regions



Variants in Regulatory Regions



Nature Reviews | Genetics

- Use of ENCODE data
- SNPs identified in GWAS studies (red bar) often lie in enhancers or other regulatory elements.
- Smaller fraction of control SNP sets overlap with these features (blue bars)
 - SNPs on Illumina 2.5M chip
 - SNPs in 1000 Genomes
 - SNPs from 24 personal Genomes

Manolio Nat Rev Genet. 2013; 14: 549-58.

Using Functional Models in Follow-up

ARTICLE

doi:10.1038/nature16549

Schizophrenia risk from complex variation of complement component 4

Aswin Sekar^{1,2,3}, Allison R. Bialas^{4,5}, Heather de Rivera^{1,2}, Avery Davis^{1,2}, Timothy R. Hammond⁴, Nolan Kamitaki^{1,2}, Katherine Tooley^{1,2}, Jessy Presumey⁵, Matthew Baum^{1,2,3,4}, Vanessa Van Doren¹, Giulio Genovese^{1,2}, Samuel A. Rose², Robert E. Handsaker^{1,2}, Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Mark J. Daly^{2,6}, Michael C. Carroll⁵, Beth Stevens^{2,4} & Steven A. McCarroll^{1,2}

Schizophrenia is a heritable brain illness with unknown pathogenic mechanisms. Schizophrenia's strongest genetic association at a population level involves variation in the major histocompatibility complex (MHC) locus, but the genes and molecular mechanisms accounting for this have been challenging to identify. Here we show that this association arises in part from many structurally diverse alleles of the complement component 4 (*C4*) genes. We found that these alleles generated widely varying levels of *C4A* and *C4B* expression in the brain, with each common *C4* allele associating with schizophrenia in proportion to its tendency to generate greater expression of *C4A*. Human *C4* protein localized to neuronal synapses, dendrites, axons, and cell bodies. In mice, *C4* mediated synapse elimination during postnatal development. These results implicate excessive complement activity in the development of schizophrenia and may help explain the reduced numbers of synapses in the brains of individuals with schizophrenia.

Take Home Points

- Cancer and other complex diseases are influenced by a combination of genetic and environmental factors.
- Genome-wide association studies (GWAS) can be used to identify common genetic variants associated with complex diseases.
- GWAS have evolved standards for study design, analysis, replication and interpretation.
- GWAS, to date, have identified over 2,000 variants associated with over 300 traits, including hundreds of variants associated with common cancers.
- Post-GWAS research includes discovery and replication, biological and functional follow-up and epidemiologic studies.
- GWAS have revealed new biology of complex diseases, and some GWAS findings are readily translatable to clinical care.

Acknowledgements

Colleagues

- NCI Epidemiology and Genomics Program
 - Muin Khoury, Acting Associate Director
- NHGRI Division of Genomic Medicine
 - Teri Manolio, Director

Slides

- Liz Gillanders, NCI
- Teri Manolio, NHGRI
- Lucia Hindorff, NHGRI
- Chris Amos, Dartmouth
- Josh Bis, University of Washington

GENE-ENVIRONMENT INTERACTIONS

Gene-Environment Interactions

- Complex diseases result from an interplay of genetic and environmental factors
- Why study Gene-environment Interactions (GxE)?
 - Studies of GxE may help identify and characterize genetic and environmental effects.
 - Studies of GxE may improve our understanding of biological mechanisms.
 - Studies of GxE may identify sub-groups for targeted interventions or screening.
- Term “GxE” is often used for both biological and statistical interactions. As with other studies of interactions the two concepts are often conflated.

What is Meant by Interaction?

- **Biological Interaction**

- The interdependent operation of two or more biological causes to produce, prevent or control an effect
- Interdependency among the biologic mechanisms of actions for two or more exposures through common pathways, protein complexes or biological products.

- **Statistical Interaction**

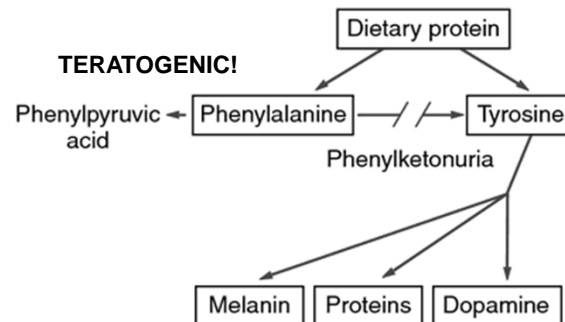
- The observed joint effects of two factors differs from that expected on the basis of their independent effects
- Deviation from additive or multiplicative joint effects

- **Effect Modification** (or Effect Measure Modification)

- Differences in the effect measure for one factor at different levels of another factor
- Example: OR differs for males vs. females; AR differs for pre-menopausal and post-menopausal women, etc.

Biological Interaction Example: PKU

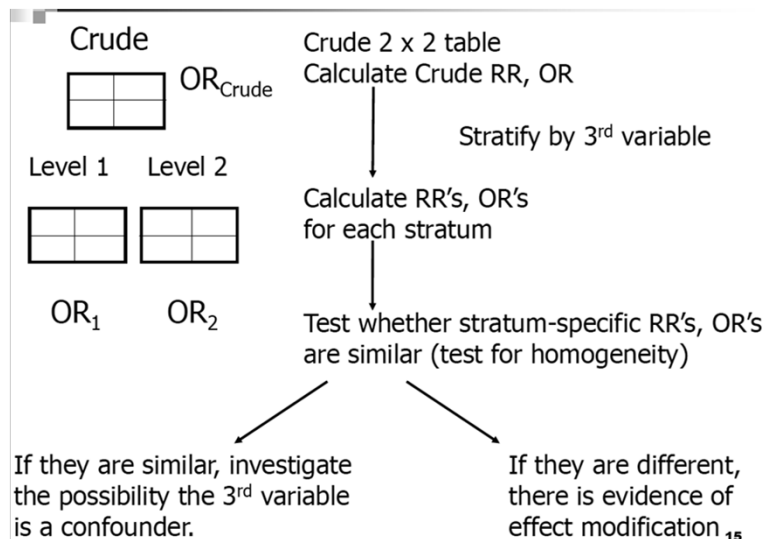
- Phenylketonuria
- Interaction between diet and genetic factor
- Can modify diet to address outcomes.



Statistical GxE Interaction

- This lecture will focus on methods for statistical interaction/effect modification.
- Keep in mind these interactions often do not have straightforward biologic interpretation, although some argue for links.
 - Non additive effects may imply non-independence of biologic mechanism of actions
 - Weinberg (1986), VanderWeele (2008-)
 - Multiplicative model may correspond to independent effects on multiple steps of a multi-step carcinogenic model
 - Siemiatycki and Thomas (1981)

Effect Measure Modification



Venous Thrombosis

- Generally manifests as thrombosis of deep leg veins or pulmonary embolism
- Incidence in women age 20-49 yrs is ~ 2 /10,000 persons/yr
- Case fatality rate is ~ 1% to 2%
- Association between oral contraceptive pill (OCP) and VT: Incidence of VT is ~12 to 34 / 10,000 in OCP users

Factor V Leiden Mutations

- R506Q mutation – amino acid substitution
- Geographic variation in mutation prevalence
 - Frequency of the mutation in populations of European descent is ~2% to 10%
 - Rare in African and Asians
- Relative risk of VT among carriers
 - 3- to 7-fold higher than non-carriers
- Is there a gene-environment interaction?

OCP, Factor V Leiden Mutations and Venous Thrombosis

Strata	Cases	Controls
G+E+	25	2
G+E-	10	4
G-E+	84	63
G-E-	36	100

Total 155 169

OR (95% CI)

OR for G in E+
 $(25 \times 63) / (2 \times 84)$
 9.4 (2.1-41.1)

OR for G in E-
 $(10 \times 100) / (4 \times 36)$
 6.9 (1.8-31.8)

Lancet 1994;344:1453

Alternative way of looking at ORs

Strata	Cases	Controls
G+E+	25	2
G+E-	10	4
G-E+	84	63
G-E-	36	100

Total 155 169

OR (95% CI)

34.7 (7.8, 310.0)

6.9 (1.8, 31.8)

3.7 (1.2, 6.3)

Reference

Lancet 1994;344:1453

Interactions are Scale Dependent

	G=0	G=1
E=0	1.0	RR _G
E=1	RR _E	RR _{GE}

Multiplicative model

No Interaction: $RR_{GE} = RR_G \times RR_E$

Relative-risk associated with E is the same by levels of G and reverse

Interaction Relative Risk = $RR_{GE} / (RR_G \times RR_E)$

Additive model

No Interaction: $RR_{GE} = RR_G + RR_E - 1$

Risk-difference associated with E is the same by levels of G and reverse

Relative Excess Risk due to Interaction (RERI) = $RR_{GE} - RR_G - RR_E + 1$

Expectations Using Different Scales

Measurement Scale and Interaction Effect	Cohort Study	Case-control study*
Multiplicative Scale		
No Interaction	$RR_{GE} = RR_G \times RR_E$	$OR_{GE} = OR_G \times OR_E$
Synergistic Interaction	$RR_{GE} > RR_G \times RR_E$	$OR_{GE} > OR_G \times OR_E$
Antagonistic Interaction	$RR_{GE} < RR_G \times RR_E$	$OR_{GE} < OR_G \times OR_E$
Additive Scale		
No Interaction	$RR_{GE} = RR_G + RR_E - 1$	$OR_{GE} = OR_G + OR_E - 1$
Synergistic Interaction	$RR_{GE} > RR_G + RR_E - 1$	$OR_{GE} > OR_G + OR_E - 1$
Antagonistic Interaction	$RR_{GE} < RR_G + RR_E - 1$	$OR_{GE} < OR_G + OR_E - 1$

* Formulas for the ORs are approximations based on the approximation of the OR to the RR

Adapted from "Genetic Epidemiology: Methods and Applications". Austin 2013.

OCP, Factor V Leiden Example

	G=0	G=1
E=0	1.0	RR _G =6.9
E=1	RR _E =3.7	RR _{GE} =34.7

Multiplicative model

Interaction Relative Risk: $RR_{GE}/RR_G \times RR_E$
 $34.7 / 6.9 \times 3.7 = 1.4$

Additive model

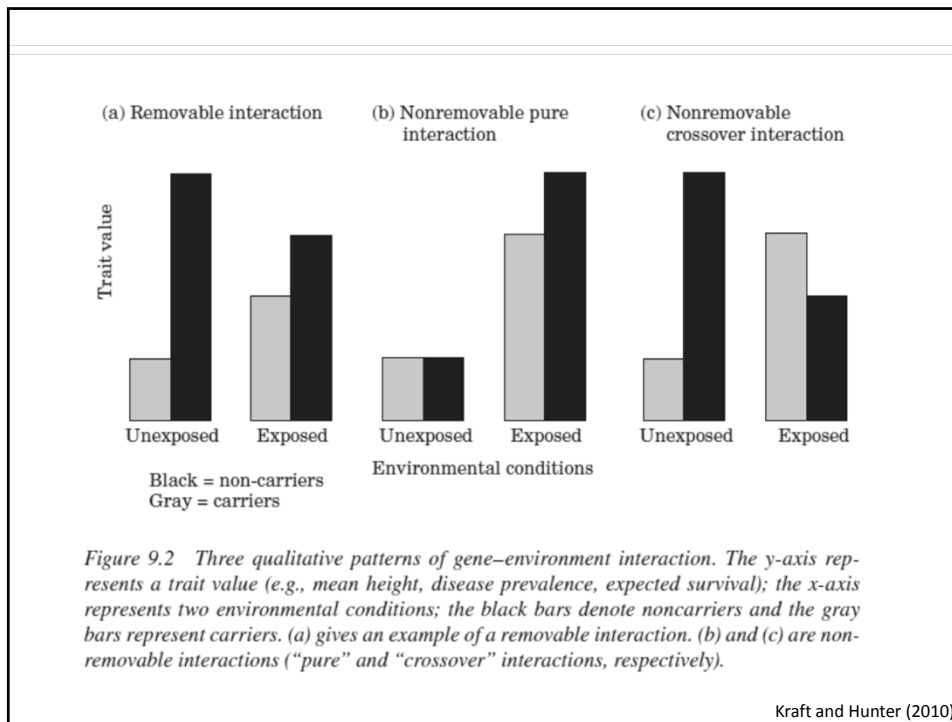
Relative Excess Risk due to Interaction (RERI): $RR_{GE} - RR_G - RR_E + 1$
 $34.7 - (6.9 + 3.7 - 1) = 25.1$

NAT2, smoking and bladder Cancer

(Garcia-Closas et al., Lancet, 2005)

	NAT2 rapid/intermediate	NAT2 slow
Never-smoker	1.0	0.9 (0.6-1.3)
Ever-smoker	2.9 (2.0-4.2)	4.6 (3.2-6.6)

No effect of NAT2 in the absence of smoking



Multiplicative vs. Additive Interactions

- Multiplicative
 - widely used in practice
 - partly due to popularity of logistic regression models
 - do not necessarily have mechanistic interpretation
 - large sample size is needed to ensure sufficient power
 - has been the focus of recent methodologic developments
 - case-only, empirical-Bayes, two-stage etc.
- Additive model
 - much less widely used (although
 - has direct relevance for evaluation of targeted intervention and links with mechanistic interaction under the sufficient component framework
 - Power is often higher than tests for multiplicative interaction

Aside: Interaction in a Regression Setting

$$G \begin{cases} 1 \text{ if carrier} \\ 0 \text{ if non-carrier} \end{cases} \quad E \begin{cases} 1 \text{ if exposed} \\ 0 \text{ if unexposed} \end{cases}$$

Risk of disease

$$p_{GE} = b_0 + b_g G + b_e E + b_{ge} GE$$

Log odds of disease

$$\log \frac{p_{GE}}{1-p_{GE}} = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE$$

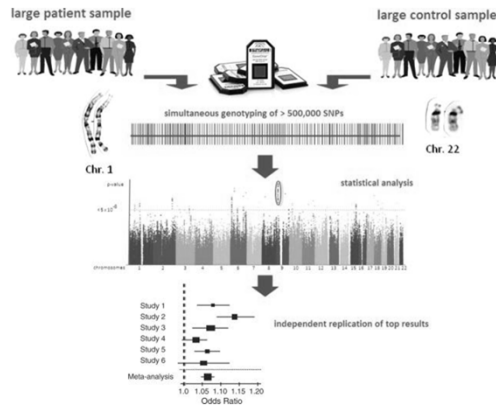
Test for "additive interaction:" H_0 is $b_{ge}=0$

Test for "(multiplicative) interaction:" H_0 is $\beta_{ge}=0$ (Interaction OR $e^{\beta_{ge}}=1$)

IN CLASS EXERCISE

Gene-Environment-Wide Interaction Study

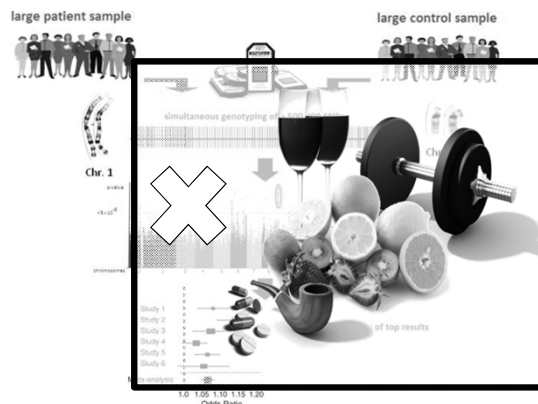
- “GEWIS”
- Motivated by discovery
- Builds on the genome-wide association study model
- Gene (G) x environmental factor (E) on a SNP-by-SNP basis across the genome



Schunkert et al.. Eur Heart J. 2010; 31: 918-925.

Gene-Environment-Wide Interaction Study

- “GEWIS”
- Motivated by discovery
- Builds on the genome-wide association study model
- Gene (G) x environmental factor (E) on a SNP-by-SNP basis across the genome



Schunkert H et al. Eur Heart J 2010;eurheartj.ehq038

Schunkert et al.. Eur Heart J. 2010; 31: 918-925.

Some of the Challenges in “GEWIS”

- Power for discovery:
 - False-negative findings
 - Individual studies with low sample sizes
 - Multiple comparisons (multiple G, E and models)
- Characterizing and modeling non-genetic risk factors:
 - Time dependency
 - Measurement error
 - Multi-faceted
- Interpretation of significant findings:
 - Biological plausibility in an agnostic approach
 - Heterogeneity and replication
 - Translation to clinical or public health relevance

Thomas D. Nat Rev Genet. 2010; 11: 259-72;
Dempfle A. et al., Eur J Hum Genet 2008; 16: 1164-1172.

Goals

- Identify methods with high power
- Reduce number of false positives

Approaches for GEWIS

- Multifactor dimension reduction, and other machine learning techniques
- Pathway/hierarchical models
- Family based tests
- Additive models
- **Logistic regression-based tests for multiplicative interactions**

Methods for GxE

- See full table in Hutter et al. Genet Epidemiol. 2013 Nov;37(7):643-57. doi: 10.1002/gepi.21756.

Table 1. Overview of analytical methods for characterization and discovery of G × E interactions

Method	Highlights	Reference
Sufficient component models	<ul style="list-style-type: none"> • Framework where the presence of interaction in the additive scale can be used as evidence of overlap of biologic actions through a common underlying pathway. • Useful in characterization motivated by understanding biological mechanisms. 	[VanderWeele, 2009; VanderWeele and Robins, 2007].
Test for qualitative interaction	<ul style="list-style-type: none"> • Tests for qualitative interactions. • Useful in characterization motivated by understanding nature of interaction. 	[Gail and Simon, 1985]
Goodness of fit tests	<ul style="list-style-type: none"> • Simultaneously test multiple terms including G × G and G × E. • Useful when building parsimonious models for risk assessment in public health contexts. 	[Hosmer et al., 1997]
Unconditional logistic regression	<ul style="list-style-type: none"> • Standard method for analysis. • Robust to assumptions about G-E correlation. 	[Breslow and Day, 1980]
Case only	<ul style="list-style-type: none"> • Efficient method for analysis of multiplicative interaction odds ratio. • Exploits, and is highly sensitive to, assumption of G-E independence. • Useful for improved power for discovery of G × E interaction. 	[Piegorisch et al., 1994]
Maximum likelihood estimation method	<ul style="list-style-type: none"> • Exploits G × E independence assumption in the analysis of case-control data. • Allows efficient estimation of all parameters from logistic regression model. Useful for both discovery and characterization. For discovery, the method could be used for joint test for genetic effects and G × E interaction. 	[Chatterjee and Carroll, 2005]

Logistic Regression Based Methods for Multiplicative GxE

Method	Key Details
Case-control	Robust model; Does not assume G-E independence; low power for discovery.
Case-only	Gains in power and efficiency under G-E independence.
Data-adaptive estimators (e.g. Empirical Bayes and Bayesian Model Averaging)	Increased power versus case-control and improved control of type 1 error versus case-only.
Two-step procedures	Screening step and testing step. Maintains type 1 error and provides power gain under many settings.
Joint-test of genetic main effect and GxE (2 degree of freedom tests)	Tests null hypothesis that genetic marker is not associated with disease in any stratum defined by exposure.

Modified from Mukherjee et al. Am J. of Epidemiology. 2012; 175(3): 177-190.

Case-only Design

- Case-only approach tests the association between the genotype and exposure in the cases only.
- Has higher statistical power than standard case-control method with same number of cases.
- Relies on assumption that genetic and environmental factors are independent in the source population.
- Increased false-positive rate if assumption is violated.

2x2x2 Representation of Unmatched Case-Control Study Examined by Standard Test for GxE Interaction

Environment	Gene			
	G=1		G=0	
	D=1	D=0	D=1	D=0
E=1	<i>a</i>	<i>b</i>	<i>e</i>	<i>f</i>
E=0	<i>c</i>	<i>d</i>	<i>g</i>	<i>h</i>
OR (D-E)	<i>ad / cb</i>		<i>eh / gf</i>	
OR (G×E)	<i>adfg / bceh</i>			

$$\text{OR(GxE)} = \text{OR(G-E|D=1)}/\text{OR(G-E|D=0)}.$$

Assuming $\text{OR(G-E|D=0)}=1$ greatly reduces the variability in OR(GxE) .
The case-only estimate of OR(GxE) is ag/ce .

Piegorsch (1994)

Extensions of Case-only method.

- The gain in power comes from the assumption of G-E independence, not the fact that only cases are used.
- Can build assumption into the analysis of case-control data.
 - allow for estimation of main effects
 - Allow for estimates/tests of interaction effects other than multiplicative odds model.
 - See Han et al. AJE 2012.
- “Hedge” methods weighted towards case-only method if data supports independence assumption, towards case-control method if assumption appears to be violated.
 - Empirical Bayes, model averaging methods
 - Mukherjee et al 2012; Li and Conti 2009
- Use of case-only design and/or G-E independence assumption in new methods for large-scale GxE analysis

Two-Step Methods

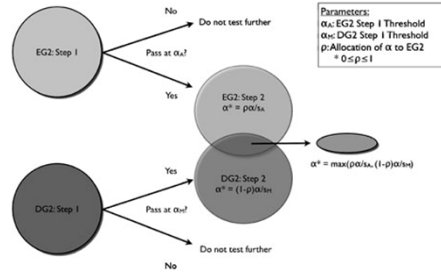
Step 1: Screening Step

Prioritize SNPs for testing:

- Correlation between G and E in full sample of cases and controls
- Marginal association between G and outcome (D)
- Hybrid approaches

Step 2: Testing Step

Test for interaction in prioritized SNPs with appropriate significance levels.



Hybrid approach proposed by Murcraey et al 2011.

Murcraey CE, et al. Am J Epidemiol 2009; 169: 219-226.
 Murcraey CE, et al. Genet Epidemiol 2011; 35: 201-210.
 Kooperberg C and Leblanc M. Genet Epidemiol. 2008; 32: 255-63.

Modules Framework for GxE Methods

Module A: Screening

- No Screening
- Marginal (G-D association)
- Correlation (G-E)
- Hybrid approaches

Module B: Multiple Comparisons

- Bonferroni testing
- Permutations
- Weighted hypothesis testing

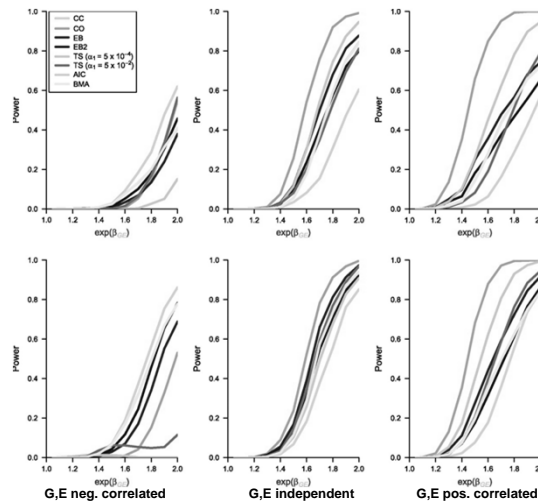
Module C: Testing

- Case-control
- Case-only
- Empirical Bayes
- Bayesian Model Averaging

Modified from Hsu et al. Genetic Epidemiology 2012; in press.

Power Considerations

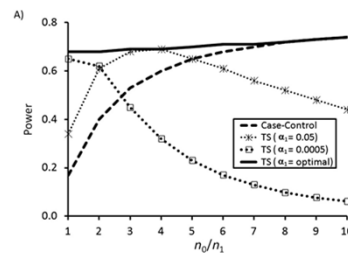
- Rule of thumb is that tests of interactions need sample sizes 4 times larger than tests of main effects.
- All methods require large sample sizes (on the order of 10,000 cases) for reasonable effect sizes.
- The most powerful method depends on assumptions on underlying interaction.
- Hybrid and cocktail methods tend to be relatively powerful over a wider variety of types of interactions.



Mukherjee B et al. Am. J. Epidemiol. 2012;175:177-190

Practical Considerations

- Choosing optimal alpha/weights
- Case:control ratio
- Linkage disequilibrium between top SNPs
- Computational needs



Empirical power for two-step methods for different alpha thresholds as a function of the ratio of cases to controls (n_0/n_1). $N_1=2,000$; $R_{ge}=1.8$; $\Pr(E)=0.5$.

Thomas D, et al. Am. J. Epidemiol. 2012;175: 203-7.

Software for analysis

Software	Good for	URL
PLINK	GWAS, data handling, GE test, joint test	http://pngu.mgh.harvard.edu/~purcell/plink/
ProbABEL	GWAS, computes robust variance-covariance matrix	http://www.genabel.org/packages/ProbABEL
GxEscan	R script incorporating multiple GWAS GxE tests	http://biostats.usc.edu/software
Multassoc	Test a group of SNPs taking interaction with other G, E into account	http://dceg.cancer.gov/tools/analysis/multassoc
R	Flexible, write your own scripts	http://www.r-project.org/
METAL	Meta-analysis	http://www.sph.umich.edu/csg/abecasis/metal/

EPIDEMIOLOGY OF GXE

Different Motivations for Studying GxE

DISCOVERY

- Identify novel loci
- Focus on variants that would not be found in marginal search alone
- Priority given to power
- Hypothesis generating

CHARACTERIZATION

- Describe interaction
- Focus on putative and established variants
- Priority given to descriptive model
- Provides etiologic insight

Genetic Epidemiology with a Capital “E”

Thomas DC (2000)

- Focus on population-based research
- Joint effects of genes and the environment
- Incorporation of underlying biology

Khoury MJ (2011)

- Large scale harmonized cohorts and consortia
- Multilevel factors (includes GxE and more) across the lifestyle
- Incorporation of underlying biology
- Integrating, evaluating and translating knowledge

Sources of Bias in Epidemiology

- **Selection Bias**
 - Arises from issues in case/control ascertainment
- **Information Bias**
 - Arises from measurement error or misclassification in assessing factors of interest.
- **Confounding**
 - Arises when there is an extraneous disease risk factor that is also associated with exposure and not in the causal pathway.

Box 1 | Major sources of bias that affect case-control and prospective cohort studies

Biases that relate to subject selection

Prevalence-incidence or survival bias. Selection of existing cases that are currently available for study will miss fatal and short episodes, and might miss mild or silent cases¹⁵.

Non-response (or respondent) bias. Differential rates of refusal or non-response to inquiries between cases and disease-free comparison subjects¹⁵.

Diagnosis bias. Also known as diagnostic suspicion bias. Knowledge of a subject's exposure to a putative cause of disease can influence both the intensity and outcome of the diagnostic process¹⁵.

Referral or admission-rate bias. Factors related to the probability of referral. Cases who are more likely to receive advanced care or to be hospitalized — such as those with greater access to health care or with co-existing illnesses — can distort associations with other risk factors in clinic-based studies, unless the same referral or admission biases are operative in disease-free comparison subjects¹⁵.

Surveillance bias. If a condition is mild or likely to escape routine medical attention, cases are more likely to be detected in people who are under frequent medical surveillance¹⁵.

Biases that relate to measuring exposures and outcomes

Recall bias. Questions about specific exposures might be asked more frequently of cases, or cases might search their memories more intensively for potential causative exposures.

Family information bias. The flow of family information about exposures or illnesses can be stimulated by, or directed to, a new case in its midst¹⁵.

Exposure suspicion bias. Knowledge of a patient's disease status can influence the intensity and outcome of the search for exposure to a putative cause¹⁵.

Manolio et al. Nat Rev Genet. 2006. 7: 812-820.

Sources of Bias in G and GxE

Method	Key Considerations
Selection Bias	<ul style="list-style-type: none"> • Issues of poor control selection and incomplete case ascertainment. • Need to consider non-respondants, people who refuse or are unable to provide DNA/data
Information Bias	<ul style="list-style-type: none"> • Errors in questionnaire, specimen handling • Highlights importance of lab QC • Can impact type I and type II error for GxE
Confounding	<ul style="list-style-type: none"> • Population stratification for G • "Traditional" factors for E • Under certain conditions "confounders" can bias the interaction term (see, for example, Tchetgen Tchtgen and VanderWeele 2012).

- **Concerns of all three of these factors increase when examining GxE in existing genetic studies that used "convenient controls".**
- **Presence of these biases may contribute to disparate findings in literature and issues in replication.**

Modified from Garcia-Closas et al. in *Human Genome Epidemiology*. 2004.

Challenges to Investigations of GXE

"Data harmonization, population heterogeneity, and imprecise measurements of exposures across studies"
Khoury et Al, 2012

"Establishing the existence of and interpreting GXE interactions is difficult for many reasons, including, but not limited to, the selection of theoretical and statistical models and the ability to measure accurately both the G and E components." Boffetta et Al, 2012



The Fiddler Crab Analogy*



- Issues of imbalance in how we look at G and E
- Complications with how we look at E:
 - Distribution of E
 - Measurement error
 - Multi-faceted

* Credited to Chris Wild (CEBP, 2005. 14: 1847-1850) via Duncan Thomas



Measurement Error

- Environmental factors are often complex, multifaceted and difficult to measure.
- Measurement error can lead to both type I and type II error for GxE.
- Statistical methods to correct misclassification exist, but are infrequently used for GxE.
- Measurement error has strong impact on power to detect GxE.
- Additional issues arise when considering GxE across multiple studies


Using Traditional “Environmental Data” in Consortium Settings

- Large sample-sizes needed to detect GxE: often need to combine data across studies.
- **Data harmonization** is the process of combining information on key data elements from individual studies in a manner that renders them inferentially equivalent.

Harmonization Resources for Phenotype Data

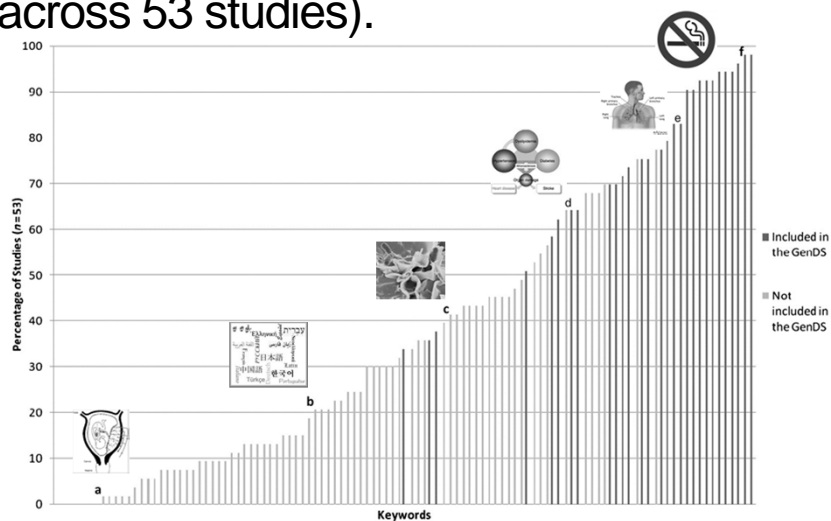
- Identify and document a set of core variables
- Assess the potential to share each variable between studies
- Define appropriate data processing algorithms
- Process and synthesize real data.



- Develop a recommended minimal set of high priority measures
- Toolkit provides *standard* measures related to complex diseases, phenotypic traits and environmental exposures

Fortier I et al. Int. J. Epidemiol. 2011;40:1314-1328
Hamilton CM et al. Am J Epidemiol. 2011; 174:253-60.

Some variables are more “harmonizable” than others (DataSHaPER approach across 53 studies).

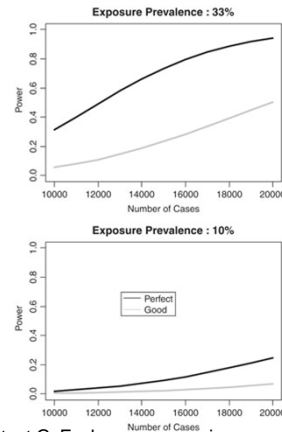


Fortier I et al. Int. J. Epidemiol. 2011;40:1314-1328

Trade-offs in Data Harmonization



- Cost of collecting rich phenotype information can put restrictions of sample size for detailed measures
- Genetic and environmental heterogeneity are likely present in large samples from multiple studies
- Combining across studies may require identifying the “least common denominator”
- Harmonization can induce misclassification and heterogeneity.



Power to detect GxE when exposure is measured perfectly or via a good proxy (77% specificity and 99% sensitivity). Interaction OR=1.35, type 1 error= 5×10^{-8} .

Bennett SN, et al. Genet Epidemiol. 2011; 35: 159-173.

Mega- vs. Meta-Analysis

Mega-Analysis

- Analyze all samples in a single model (AKA “pooled” analysis)
- Facilitates more complex models (rather than meta-analysis of a defined set of parameter estimates).
- Can introduce confounding/bias, particularly if case/control numbers differ.
- Requires all data to be accessible by a single analyst

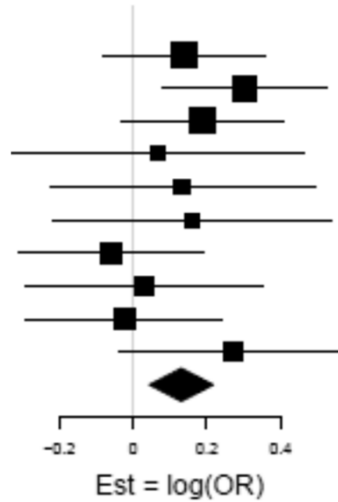
Meta-Analysis

- Analyze each study separately and then combine study specific estimates.
- Focus often only on meta-analysis of interaction term; may not fully capture joint effects.
- Question of comparability of what is being meta-analyzed.
- Explicitly shows between study heterogeneity.

Heterogeneity

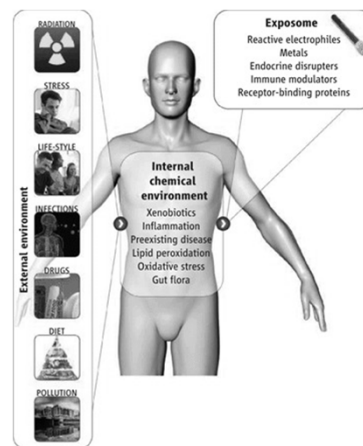
- “If explanations can be found for heterogeneity, there is an opportunity for insights about the complexity of the disease, but spurious inconsistency due to methodological or data-quality differences will just add confusion”

- Thomas 2010



Beyond Data Harmonization

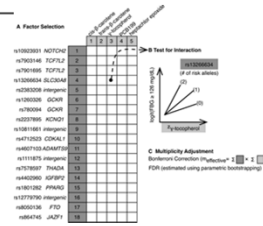
- We may be missing key environmental factors
- Measuring the environment often does not have the same “economy of scale”
- The multifactorial and dynamic nature of exposure/risk can complicate the study of environmental factors



Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus

Chirag J. Patel · Rong Chen · Keiichi Kodama ·
John P. A. Ioannidis · Atul J. Butte

Hum Genet. 2013. Epub ahead of print.

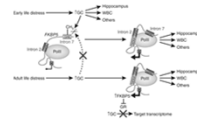


How do environments talk to genes?

Moshe Szyf

A report elucidates the widely recognized, but poorly understood, concept of gene-environment interaction, finding a molecular mechanism in the case of post-traumatic stress disorder: demethylation of a glucocorticoid response element in the stress response regulator *FKBP5* that depends on both the risk allele and childhood trauma.

Nat Neurosci. 2013 Jan;16(1):2-4.



VIEWS

IN FOCUS

Integrative Cancer Epidemiology—The Next Generation

Margaret R. Spitz¹, Neil E. Caporaso², and Thomas A. Sellers³

Summary: We outline an integrative approach to extend the boundaries of molecular cancer epidemiology by integrating modern and rapidly evolving “omics” technologies into state-of-the-art molecular epidemiology. In this way, one can comprehensively explore the mechanistic underpinnings of epidemiologic observations in cancer risk and outcome. We highlight the exciting opportunities to collaborate across large observational studies and to forge new interdisciplinary collaborative ventures. *Cancer Discov.* 2(12): 1087–90. ©2012 AACR.

Cancer Discov. 2012 Dec;2(12):1087-90.

Summary

- Gene-environment wide interaction studies are used for discovery and characterization.
- Remember distinction between biological and statistical interaction.
- Important to consider scale (additive vs. multiplicative)
- Large sample sizes are needed for GxE studies, particularly for GEWIS.
- Data harmonization allows core variables to be combined across studies.
- We need to give the “E” similar, if not more, attention than we give the “G” for GxE analysis.

IN CLASS EXERCISE

In Class Exercise:

- You continue to work with collaborators on the FAKE study. They decide to follow-up on their candidate gene study with a genome-wide association study (GWAS). They were only able to afford genome-wide genotyping on a subset of the subjects, so they decide to reach out to their collaborators in the Meta-Analysis of Diet and Environment for Understanding Phenotypes (MADE-UP) consortia. The next page has “table 1” for the 8 studies in this consortia. Brainstorm with your group about the following:
 - What are potential issues/challenges that you might encounter in analyzing this data?
 - What are solutions might you use for some of these challenges?
 - What additional information would be most helpful for you to have?

	Study 1: Cohort		Study 2: Case-control		Study 3: Case-control		Study 4: Cohort		Study 5: Case-control		Study 6: Case-control		Study 7: Cohort		Study 8: Cohort	
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control
N	931	1,435	1,410	1,666	2,031	2,044	69	238	465	465	5,450	5,475	1,064	1,202	1,381	1,303
% Female	32.0%	43.2%	32.1%	44.1%	19%	16.5%	34.8%	26.2%	26.7%	26.8%	56%	56%	60%	55%	0%	0%
Mean Age (yrs)	65.5	65.8	65.1	67.5	59.8	61.3	58.1	57.8	62.4	62.8	64.0	64.2	61.3	62.8	65.4	65.4
% Strawberry eaters	47.7%	45.8%	45%	40%	65.2%	56.2%	60.9%	65.2%	55.4%	55.6%	59.3%	52.1%	58.2%	59.0%	65.3%	66.4%
% Rhubarb eaters	21.6%	15.1%	25.6%	24.5%	36.7%	34.5%	12.1%	7.1%	14.1%	10.2%	NA	NA	28.4%	33.4%	14.9%	10.7%
Instrument for dietary assessment	FFQ	FFQ	FFQ	FFQ	5 Q survey	5 Q survey	FFQ	FFQ	24 hour recall	24 hour recall	5 Q survey	5 Q survey	FFQ	FFQ	24 hour recall	24 hour recall
Country	USA	USA	USA	USA	China	China	Japan	Japan	Germany	Germany	USA	USA	Canada	Canada	USA	USA
Genotyping Platform	Illumina 550K		Affymetrix 6.0		Illumina 550K		Illumina 1M		Illumina Omni Express		Illumina Omni Express		Affymetrix Axiom-CEU		Affymetrix Axiom-CEU	

What is old is new

Rare Variants

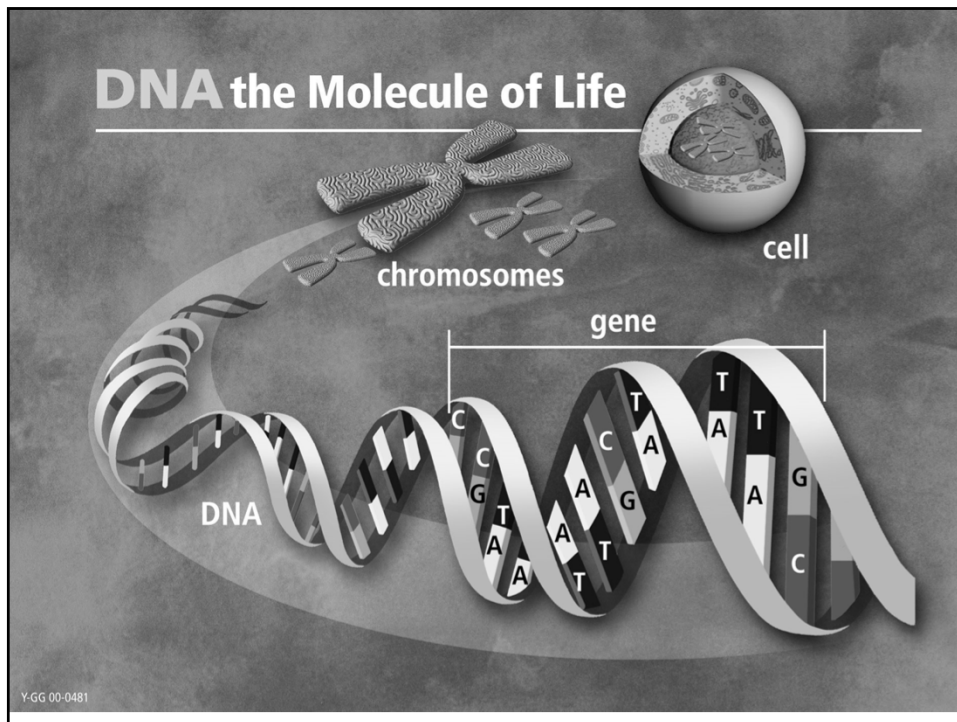


Karen L. Edwards, Ph.D.
Professor
Department of Epidemiology and
Genetic Epidemiology Research Institute
University of California, Irvine
Irvine, CA
kedward1@uci.edu

This session



- Why study rare variants and where do they come from?
- Association analysis of rare variants
- Using new technology in family studies

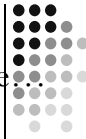


Why Study Rare Variants?

- Problem of “missing heritability”
 - GWAS studies have thus far focused on common SNPs
 - Have identified over 500 strong independent SNP associations
 - However, most common variants (SNPs) identified only explain a small proportion of the total genetic variance of complex diseases
 - 10-20% depending on the disease
 - These associations tend to be with non functional variants, and not causal polymorphisms
 - There are additional susceptibility loci to be found



Nice thought piece on the rare versus common variant debate



 GENOME-WIDE ASSOCIATION STUDIES

Rare and common variants: twenty arguments

Greg Gibson

Abstract | Genome-wide association studies have greatly improved our understanding of the genetic basis of disease risk. The fact that they tend not to identify more than a fraction of the specific causal loci has led to divergence of opinion over whether most of the variance is hidden as numerous rare variants of large effect or as common variants of very small effect. Here I review 20 arguments for and against each of these models of the genetic basis of complex traits and conclude that both classes of effect can be readily reconciled.

NATURE REVIEWS | GENETICS

VOLUME 13 | FEBRUARY 2012 | 135

A Paradigm Shift in Genetic Epi?



- Common Variant-Common Disease (CDCV) hypothesis
 - Common diseases are due to common genetic variation
 - Basis for most GWAS studies
- Common Disease—Rare Variant (CDRV) hypothesis
 - Multiple rare DNA sequence variations, each with relatively high penetrance and “large” effects, are the major contributors to genetic susceptibility to complex disease

What is a rare variant?

Table 1 | Potential frequencies of causal variants in complex traits

Variant class	Minor allele frequency		Implications for analysis
Very common	Between 5 and 50%	GWAS	Amenable to association analysis using current genome-wide association methods
Less common	Between 1 and 5%		Amenable to association analysis using variants catalogued in the 1000 Genomes Project
Rare (but not private)	Less than 1% but still polymorphic in one or more major human populations		Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families
Private	Restricted to probands and immediate relatives		Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants

From: Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews. Genetics 2010, 11:415–25.

LESS COMMON
 $1\% < \text{MAF} < 5\%$

RARE
 $\text{MAF} < 1\%$

PRIVATE
Unique to
Proband

From Dr. S. Santorico – UCD Dept of Statistics

Rare Variants

- Genetic architecture of most complex traits has not been fully described
 - Rare variants ($\text{MAF} < 1\%$) are “common” and make up most of the polymorphic sites in the human genome.
- Rare variants may have larger effect sizes, explain some of the missing heritability, and should identify new susceptibility loci for both common and Mendelian disorders

Significance of Rare Variants



Discovering the genetic basis of common diseases, such as diabetes, heart disease, and schizophrenia, is a key goal in biomedicine. Genomic studies have revealed thousands of common genetic variants underlying disease, but these variants explain only a portion of the heritability. Rare variants are also likely to play an important role, but few examples are known thus far, and initial discovery efforts with small sample sizes have had only limited success.

Zuk et al., www.pnas.org/cgi/doi/10.1073/pnas.1322563111

Challenges of studying Rare Variants



- They are rare!
 - Impacts power and sample size
- Definition of rare varies
 - In general, a minor allele frequency (MAF) of less than 1% is considered rare
 - MAF between 0.1% and 3% are defined as rare
 - MAF <0.1% as novel
 - In contrast to GWAS where the MAF for most variants is about 5% or greater
 - Private mutations may be found in a single individual or family
 - Rare variants are not generally in LD with common variants and may have different population histories

Considerations in Rare Variants Analysis



What to Sequence and Who

Sequencing Depth

Analyzing data

Rare variant association study (RVAS) vs. Common variant association study (CVAS – aka GWAS)

Filtering

Using Annotation

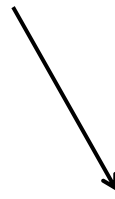
Rare vs. Common Variants Analysis



Genome Wide Association Study (GWAS)

RVAS

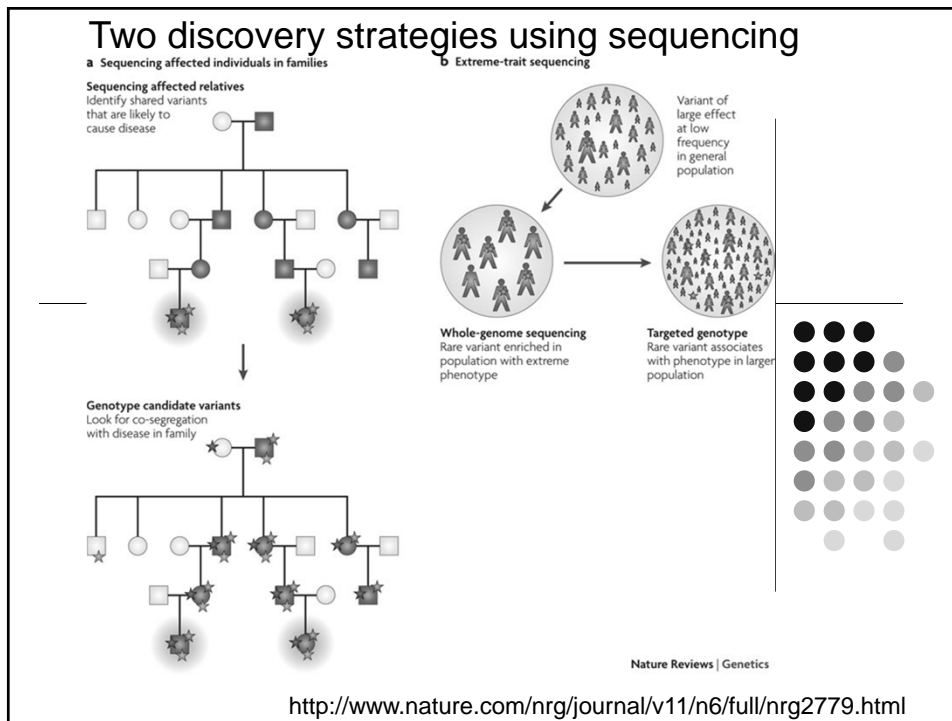
CVAS



Sequencing: Who, what and why?



- Sequencing is still “expensive”
 - Unrelated cases and controls
 - Families
 - Extremes – affected and unaffected
- Sequencing for discovery
 - Whole Exome Sequencing (WES) (coding regions)
 - Whole Genome Sequencing (WGS)
 - Targeted regions
- Followup with targeted genotyping of identified rare variants in larger samples

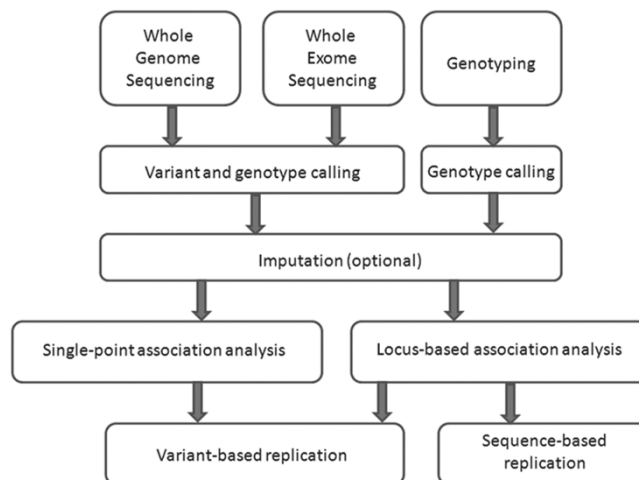
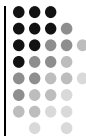


Sequencing Depth



- Most current sequencing platforms generate millions of short sequence reads
 - High-depth reads (e.g. 30x) to exhaustively identify variation
 - Decreased sequencing depth studies are increasing – requires more samples – detection and calling accuracy can be compromised.
- Reads are then aligned to a reference genome
- Variant calling is performed to identify sites at which one or more samples differ from the reference sequence
- Focus is on SNPs, copy number variation is less straightforward at this point

Overview of steps taken in the search for low-frequency and rare variants affecting complex traits



Human Molecular Genetics, 2013 R1–R6
doi:10.1093/hmg/ddt376

Rare Variant Reference Panels



- 1000 Genomes Project
 - Catalog of common and uncommon variation identified through WGS and exome sequencing across several global populations
- NHLBI Exome Sequencing Project (NHLBI-ESP) (<http://esp.gs.washington.edu>)
 - WES of 6500 samples in phenotyped sets from the USA.
- UK10K Project (www.uk10k.org)
 - High-depth WES of 6000 and low-depth WGS of 4000 well-phenotyped individuals from the UK

Rare Variant Association Analysis



- Statistical considerations for analyzing rare variants are important
- Testing for associations are challenging due to rareness and the large number of rare variants
- Approaches
 - Single variant analysis
 - Single-point analysis of rare variants is under-powered
 - Do not have enough copies of the rare variant allele in most association studies
 - Alternative is a multivariate approach that combines information across multiple rare variant sites within a defined region
 - Defined regions of the genome may include
 - gene (locus) - for exome or candidate gene studies
 - or other functional unit
 - defined genomic region- such as a sliding window for whole genome studies
- Numerous locus-specific statistical approaches have been developed
- Correcting for multiple comparison is still needed

Statistical approaches for analysis of rare variants



- Many Approaches have been developed:
 - Collapsing and Aggregation Methods (Burden tests)
 - Non-Burden tests
- Collapsing methods/Burden tests
 - Aggregate information across multiple variants into a single quantity to evaluate cumulative effects (burden) of multiple variants in a defined genomic region of interest
 - Test for trait association with an accumulation of rare minor alleles
 - Vary in the way they collapse variants
 - Assume all collapsed variants are associated with the disease and variants can be either deleterious or protective
- Non-Burden tests
 - Multivariate tests that combine single-variant test statistics
 - Make no assumption about direction and magnitude of effect of each rare variant – more flexible and more powerful in some scenarios
 - Sequence Kernel Association Test (SKAT)

Rare Variant Methods, cont



- Vary in way variants are collapsed
 - Model the phenotype using a regression approach
 - as a function of the proportion or count of rare variants in the defined region at which an individual has the minor allele (Burden test)
 - Or as a function of the presence or absence of a minor allele at any rare variant site within the locus or region of interest - (Collapsing method)
 - Limitation is that we ignore directionality (eg both deleterious and protective variants are treated in the same way)
 - Assume equal contribution from each variant
 - Most powerful when most variants are causal and in the same direction (eg deleterious)
- Weighted aggregation tests – weight each variant based on other evidence, these weights contribute to the “burden”
- SKAT tests are more powerful when most variants are not causal or when the effects of causal variants are in different directions – a regression framework
- A unified approach between the collapsing methods and SKAT has been developed
 - SKAT-O ; maintains power under both scenarios



A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic

Bo Eskerod Madsen¹, Sharon R. Browning^{2*}

Abstract

Resequencing is an emerging tool for identification of rare disease-associated mutations. Rare mutations are difficult to tag with SNP genotyping, as genotyping studies are designed to detect common variants. However, studies have shown that genetic heterogeneity is a probable scenario for common diseases, in which multiple rare mutations together explain a large proportion of the genetic basis for the disease. Thus, we propose a weighted-sum method to jointly analyse a group of mutations in order to test for groupwise association with disease status. For example, such a group of mutations may result from resequencing a gene. We compare the proposed weighted-sum method to alternative methods and show that it is powerful for identifying disease-associated genes, both on simulated and Encode data. Using the weighted-sum method, a resequencing study can identify a disease-associated gene with an overall population attributable risk (PAR) of 2%, even when each individual mutation has much lower PAR, using 1,000 to 7,000 affected and unaffected individuals, depending on the underlying genetic model. This study thus demonstrates that resequencing studies can identify important genetic associations, provided that specialised analysis methods, such as the weighted-sum method, are used.

Citation: Madsen BE, Browning SR (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet 5(2): e1000384. doi:10.1371/journal.pgen.1000384



Improving power

- Filtering based on likelihood of function
- Alternatively, could incorporate weights according to probability of being functional
 - Good weight choices can improve power
- Based on MAF – under assumption that rarer variants are more likely to be deleterious according to natural selection theory
 - Implemented in a number of different tests and based on internal information from your sample
- Functional annotation predictions
 - Weights are based on external information
 - GERP or PhastCons- Measures of Conservation
 - PolyPhen-2 – computational predictions that a variant is likely to be damaging
 - CADD – Combined Annotation Dependent Depletion – a measure of deleteriousness

Searching for missing heritability: Designing rare variant association studies

Or Zuka,b,1, Stephen F. Schaffnera, Kaitlin Samochaa,c,d, Ron Doa,e, Eliana Hechter, Sekar Kathiresana,e,f,g, Mark J. Dalya,c, Benjamin M. Nealea,c, Shamil R. Sunyaeva,h, and Eric S. Landera,i,j,2

Genetic studies have revealed thousands of loci predisposing to hundreds of human diseases and traits, revealing important biological pathways and defining novel therapeutic hypotheses. However, the genes discovered to date typically explain less than half of the apparent heritability. Because efforts have largely focused on common genetic variants, one hypothesis is that much of the missing heritability is due to rare genetic variants. Studies of common variants are typically referred to as genomewide association studies, whereas studies of rare variants are often simply called sequencing studies. Because they are actually closely related, we use the terms common variant association study (CVAS) and rare variant association study (RVAS). In this paper, we outline the similarities and differences between RVAS and CVAS and describe a conceptual framework for the design of RVAS. We apply the framework to address key questions about the sample sizes needed to detect association, the relative merits of testing disruptive alleles vs. missense alleles, frequency thresholds for filtering alleles, the value of predictors of the functional impact of missense alleles, the potential utility of isolated populations, the value of gene-set analysis, and the utility of de novo mutations. The optimal design depends critically on the selection coefficient against deleterious alleles and thus varies across genes. The analysis shows that common variant and rare variant studies require similarly large sample collections. In particular, a well-powered RVAS should involve discovery sets with at least 25,000 cases, together with a substantial replication set.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1322563111/-/DCSupplemental.

Important Considerations

- Population stratification is an important consideration
 - Rare variants show increased population specificity
 - Rare variants can show stronger patterns of population stratification than common variants
 - Most of the rare variant tests allow adjustment for covariates including PCA's
 - Some studies have shown that genomic control and PCA have not been effective at controlling population stratification
- Underscores the need for attention to study design
 - Case and Control Selection
 - Replication

Unrelated Individuals and Family Studies



- Case-control association studies will require large sample sizes
 - Burden and non-burden tests increase the overall MAF, but power is still a concern
- Family studies are making a come back
 - Variants that are rare in the population will be “enriched” in families where the variant is causal
 - Incorporation of new technology is a focus
 - Analytic approach varies
 - Discovery
 - Followup on previous linkage regions
 - Combine linkage and association testing

Main Points to Remember



- Emerging Area
 - Methods are evolving
 - Families and Unrelated individuals
 - No consensus yet on approach
 - As data / evidence accumulates we will likely see more “standardized” approaches as with GWAS
 - Functional information and annotation will also continue to improve
- Basic factors still need to be considered
 - Appropriate selection of your sample
 - Adjustment for covariates, including population stratification
 - Adjustment for multiple comparisons
- Recurring themes
 - What is old is new
 - Emerging methods that build on fundamentals

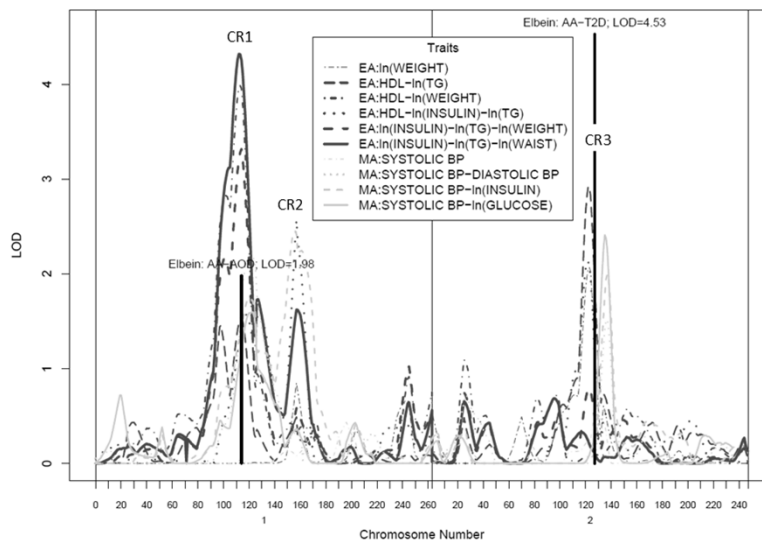
Example: Identifying susceptibility genes for metabolic syndrome in a multi-ethnic family study



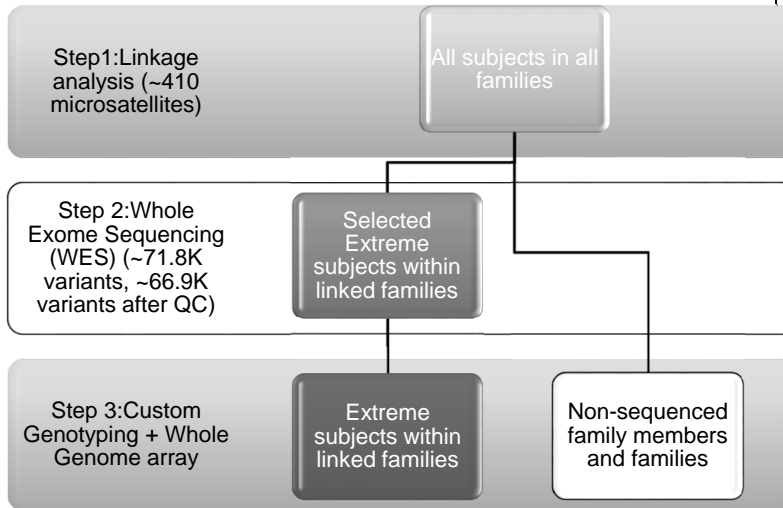
Multivariate Linkage Analyses



Chromosomes 1-2 : Results from Linkage Analysis in the GENNID EA, MA and AA Families



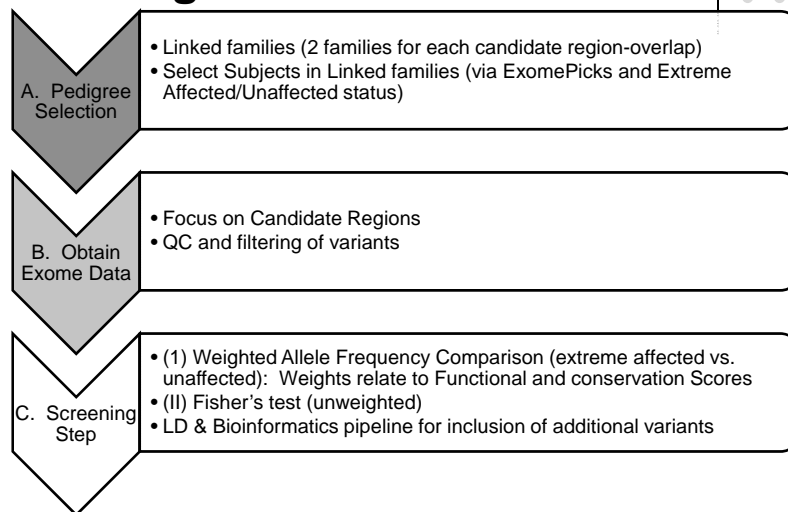
Project Overview



7/5/2016

31

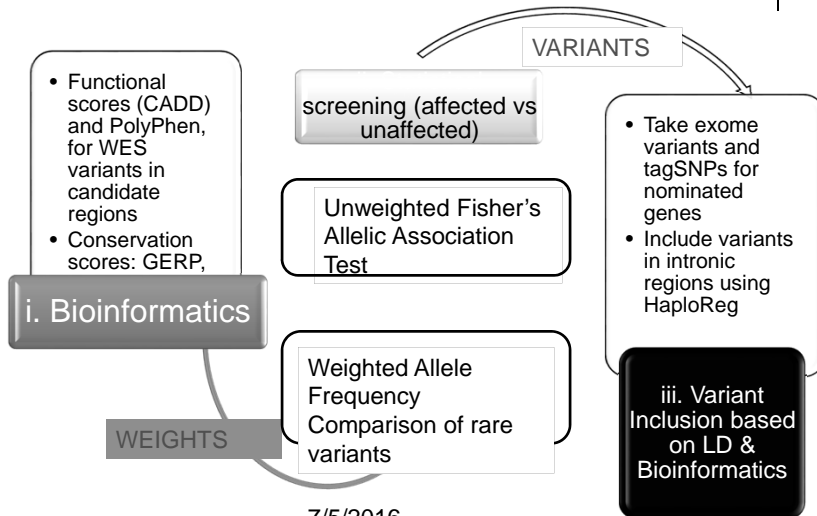
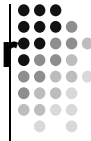
Aim 2: Whole Exome Sequencing and Filtering



7/5/2016

32

Framework for selecting variants for custom genotyping



7/5/2016

33