SISG 2014

SISG Module 16: Genetic Epidemiology

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Genetic Epidemiology

SISG Module 16

July 16-18, 2014

Instructors: Carolyn Hutter & Karen Edwards

Big Picture Learning Goals

- Familiarity with major study designs used in genetic epidemiology
- Familiarity with major issues associated with each approach
- Aware of software and web resources used in genetic epidemiology

Course Objectives

- The objective of this course is to provide an introduction to methods and applications of genetic epidemiology.
- Students will be exposed to basic concepts and principles of genetic epidemiology, including:
  - study designs for family based and population based studies
  - analytical methods used in studies of linkage and association
  - modern approaches to gene-environment interactions and rare variant analysis
  - key web resources for analysis and interpretation
  - relevant literature in the field

Class Structure

Day	Time	Lead	Topics
Wednesday	1:30-3:00	Carolyn	Class Intro
		Carolyn	Epi 101
	3:00-3:30 Break		
	3:30-5:00	Karen	Overview of Genetic Epi & Family Studies
Thursday	8:30-10:00	Karen	Family Studies II
	10:00-10:30 Break		
	10:30-11:00	Karen	Q&A
	11:00-12:00	Carolyn	Linkage Disequilibrium
	12:00-1:30 Lunch		
	1:30-3:00	Karen	Association Studies
	3:00-3:30 Break		
	3:30-5:00	Carolyn	Genome Wide Association Studies

Class Structure Continued

Friday	Time	Lead	Topics	
	8:30-10:00	Carolyn	GxE	
	10:00-10:30 Break			
	10:30-11:00	Carolyn	Harmonization	
	11:00-12:00	Karen	Sequencing Studies	
	12:00-1:30 Lunch			
	1:30-3:00	Stephanie	Functional	
	3:00-3:30 Break			
	3:30-4:30	All	Journal Club	
	4:30-5:00	Carolyn & Karen	Wrap-up	

- This module will include a combination of lectures, in class tutorials and assignments, small group interactive activities and readings.

Class introductions

- Break into groups of 2-3
- Introduce yourselves to one-another, and you will introduce your group members to the class.
- Items to include:
  - Name
  - "Day-job"
  - Main objective for taking this course
  - Thing most excited to learn in over the next 3 days

Introduction to Epidemiology

Definitions, Objectives and Historic Examples

"And it was so typically brilliant of you to have invited an epidemiologist."

THINK-PAIR-SHARE ACTIVITY

1. Define "Epidemiology"
2. Give an example of what an epidemiologist does

Definitions of Epidemiology

Objectives of Epidemiology

1. Identify disease etiology
2. Determine the burden of disease
3. Study the natural history of disease
4. Evaluate preventive and therapeutic measures
5. Provide foundation for public policy

http://www.webmd.boots.com/a-to-z-guides/tc/smoking-what-will-happen

Ignaz Semmelweis - 1846

- Childbed Fever
  - Major cause of death post childbirth
  - Theories included putrid air, solar influences, etc.
- Obstetrical Clinics of the Allgmeine Krankenhaus
  - 1st Clinic
    - Physicians and Medical Students
    - Performed autopsies at start of the day
    - Mortality: 16%
  - 2nd Clinic
    - Midwives
    - No autopsies
    - Mortality: 7%

© 2005 Elsevier

- Suggested transmission of disease from cadavers to women
- Noted a colleague died from similar infection after being punctured during an autopsy
- Implemented policy that physicians and students wash hands and scrub nails after autopsy, before contact with patients:

Edward Jenner - 1796

- Smallpox
  - 400,000 people died per year in Europe
  - Devastating when introduced to Americas (biological warfare)
  - Case fatality rate of 40-60%
- Variolation and vaccination
  - Originally infected healthy individuals with material from smallpox patients.
  - Jenner noted that dairy maids, who were exposed to cowpox, did not develop smallpox.

Smallpox Eradication

- 1967
  - ~15 million cases per year
  - ~2 million deaths
  - WHO starts efforts to eradicate smallpox
- 1980
  - WHO certifies that smallpox has been eradicated
  - Last natural case in 1977
  - Last US vaccinations in 1972
- 2001
  - Increased concern about smallpox and bioterrorism

John Snow- 1854

- Cholera
  - Severe bacterial infection
  - Miasmatic theory of disease
- London in the 1800s
  - Multiple cholera pandemics 1831-1854
  - 1949 John Snow published that cholera was caused by water

Gordis: Epidemiology, 4th Edition.
Copyright © 2008 by Saunders, an imprint of Elsevier, Inc. All rights reserved

"Father of Modern Epidemiology"

- Combined field work with statistical methods to examine water sources used by people with cholera.
- **Broad Street Pump**
  - Collected records on 83 deaths
  - 73 lived closest to pump
  - Expected 14 cases in that area
  - Interviewed and determined remaining 10 also drank from pump
  - Followed up other cases
  - Noted unexpected non-cases
    - Workhouse with only 5/535 inmates
    - Brewery where workers drank beer

John Snow's Map

http://scienceblogs.com/significantfigures/index.php/2013/03/11/200-years-of-dr-john-snow-a-significant-figure-in-the-world-of-water/

http://www.youtube.com/watch?v=Pq32LB8j2K8

"Father of Modern Epidemiology"

- Combined field work with statistical methods to examine water sources used by people with cholera.
- "Grand Experiment"
  - Multiple companies supplied same neighborhoods
  - Noted that the Lambeth Company moved intake to a less polluted source
  - Went house to house to house to collect info:

	1849 cholera epidemic	1853-54 cholera epidemic
Lambeth Waterworks Company	London intake	Seething Wells intake
Southwark and Vauxhall Water Company	London intake	London intake

Table 1-5. Deaths from Cholera per 10,000 Houses, by Source of Water Supply, London, 1854

Water Supply	Number of Houses	Deaths from Cholera	Deaths per 10,000 Houses
Southwark and Vauxhall Co.	40,046	1,263	315
Lambeth Co.	26,107	98	38
Other districts in London	256,423	1,422	56

Data adapted from Snow J: On the mode of communication of cholera. In Snow on Cholera: A Reprint of Two Papers by John Snow, M.D. New York, The Commonwealth Fund, 1936.

Richard Doll and Bradford Hill- 1952

- Smoking and lung cancer
  - 1920s health care workers noted that many lung cancer patients also smoked
  - Incidence of lung cancer in men over 45 rose 6 fold from 1930 to 1945
  - Cars or other industrial changes.
- Experimental design
  - Case-control study
    - Looked at hospital patients with and without cancer.
  - Cohort study
    - Prospectively followed >40,000 physicians

BRITISH MEDICAL JOURNAL

LONDON SATURDAY DECEMBER 13 1952

A STUDY OF THE AETIOLOGY OF CARCINOMA OF THE LUNG

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, C.B.E., Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

Relationship between death rate, expressed as a percentage of the rate for all men, and the amount smoked, for five disease groups.

BMJ 2004;328:1519

US Lung Cancer Trends

*Age-adjusted to 2000 US standard population.

Source: Death rates: US Mortality Data, 1960-2009, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention. Cigarette consumption: US Department of Agriculture, 1900-2007.

Descriptive and Analytic Epidemiology

Descriptive vs. Analytical Epidemiology

Descriptive Epidemiology

- Includes activities related to characterizing the distribution of diseases within a population

Analytical Epidemiology

- Concerns activities related to identifying possible causes for the occurrence of diseases

Objectives of Descriptive Epidemiology

- To evaluate trends in health and disease and allow comparisons among countries and subgroups within countries
- To provide a basis for planning, provision and evaluation of services
- To identify problems to be studied by analytic methods and to test hypotheses related to those problems

Think of this as the standard dimensions used to track the occurrence of a disease.

Some Definitions

- **Endemic**
  - Habitual presence of a disease in a given area
  - Usual prevalence of disease within such an area
- **Epidemic**
  - Occurrence in a community or region of a group of illnesses of similar nature in excess of normal expectancy.
  - Derived from common or propagated source.
  - Often called "outbreak"
- **Pandemic**
  - Epidemic over a wide geographic area.

Key Measures in Epidemiology

Morbidity Measures

- Incidence (new cases)
  - Cumulative incidence (proportion)
  - Incidence rate (new cases per unit time)
- Prevalence (Existing cases)
  - Existing cases
  - Point prevalence (proportion)
  - Period prevalence

Mortality Measures

- Mortality rate (number of deaths per unit time)
- Cumulative mortality (proportion)
- Case-fatality rate (proportion of subjects with disease that die from that disease)
- Proportionate mortality (proportion of people who die, who died from a specific disease- be careful when using this measure)

The denominator is a key part of measures in epidemiology.

Goal in Analytical Epidemiology

- Test a hypothesis about relationship between exposure(s) and disease(s)

- Consider Internal Validity
  - Ideal: Free from bias in design, implement, analyze and interpretation
  - Reality: We need to address biases

- Consider External Validity
  - Ideal: Generalizable
  - Reality: Applies to study population, infer more broadly

Study Designs for Analytical Epidemiology

Experimental vs. Observational Studies in Analytical Epidemiology

EXPERIMENTAL

- Randomized trials
- Community interventions

OBSERVATIONAL

- Cross-sectional studies
- Case-control studies
- Cohort studies

Randomized Trials

- RCTs
  - Randomized Clinical Trials
  - Randomized Control Trials
- Often used for studies of treatment/drugs in relation to prognosis
- Can also be used in other settings, including studies of risk and prevention
- Key elements:
  - Study population is defined
  - Study population is randomly assigned to two (or more) study "arms"
  - Outcomes are compared for the different arms

When can an RT be used?

- The exposure is potentially modifiable by an investigator
  - Need to be able to randomly assign
- **<u>Equipoise</u>** = there is genuine uncertainty about which intervention strategy is superior
  - Need to be ethical
- The primary outcomes are relatively common and occur relatively quickly
  - Need to be able to follow-up, and have enough outcomes to see the difference.

Cohort and Case-Control Studies

- Observational study designs used in epidemiology
- Consider strengths, limitations and sources of bias
- Cohort Study
  - Steps: Define cohort, take baseline measurements and exposure status, ascertain outcome information, compare incidence in exposed and unexposed
  - Can be prospective or retrospective
- Case-Control Study
  - Steps: Identify cases, identify/select controls, collect exposure history prior to disase onset, compare odds of exposure in cases to odds in controls
  - Key step is identifying an appropriate control group

© 2005 Elsevier

Cohort

Case-Control

Downloaded from: StudentConsult (on 29 September 2013 04:59 PM)
© 2005 Elsevier

Cohort Study

Strengths

- Temporal sequence
- Allows measurement of incidence rate and risk
- Can examine multiple outcomes
- Allows examination of rare exposures
- Minimizes information bias for exposure
- Minimizes survivor bias

Limitations

- Ineffective for rare diseases
- Often requires large sample
- Often requires long time to complete (lag-time)
- Expensive
- Attrition/sensitive to loss to follow-up
- May have differential ascertainment of outcome
- Confounding can occur

Case-Control Study

Strengths

- Can examine multiple exposures
- Allows examination of rare diseases
- Minimizes information bias for exposure
- Applicable when long lag time
- Compared to cohort studies, they are often smaller and require less time and money

Limitations

- Control selection can be difficult
- Recall limitations and recall bias
- Sample size issues for rare exposures
- Cannot directly estimate incidence rates, relative risks or attributable risk.

Measures of Association

Note: Some slides in this lecture come from:
http://www.teachepi.org/documents/courses/fundamentals/Pai_Lecture4_Measures%20of%20Effect%20and%20Impact.pdf
Others from University of Washington EPI 420 materials

Main Measures of Association

- Relative Risk
  measure of the relative *probability* of developing disease based on exposure status
- Attributable Risk
  measure of the amount of excess disease incidence attributed to the exposure of interest
- Odds Ratio
  measure of the relative *odds* of exposure based on disease status (can approximate the RR)

The 2x2 Table For Count Data

	Disease - yes	Disease - no	Column total (Margins)
Exposure - yes	a	b	a+b
Exposure - no	c	d	c+d
Row total (Margins)	a+c	b+d	a+b+c+d

Relative Risk (RR) For Count Data

- Used in Randomized trials Cohort studies
- Based on cumulative incidence measure
- AKA: Risk Ratio
- If no association RR=1

	Disease		
Exposure	+	-	Total
+	a	b	a+b
-	c	d	c+d

$$\textbf{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{\text{(Incidence of Disease in Exposed)}}{\text{(Incidence of Disease in Unexposed)}}$$

Attributable Risk (AR) for Count Data

- Used in Randomized trials and Cohort studies
- AKA: Risk Difference*
- Difference in risk between exposed and unexposed
- If no association AR=0

	Disease		
Exposure	+	-	Total
+	a	b	a+b
-	c	d	c+d

AR = a/(a+b) - c/(c+d) = $\text{Incidence}_{(\text{exposed})} - \text{Incidence}_{(\text{unexposed})}$

* Note: Some argue that you should use the term risk difference when testing for association, and only use "Attributable Risk" for when you have established causality.

Odds Ratio (OR) for Count Data

- Used primarily in Case Control Studies (also in Cohort)
- AKA: Relative Odds
- Good *estimate* of RR
- If no association OR=1

	Disease		
Exposure	+	-	Total
+	a	b	a+b
-	c	d	c+d

$$\textbf{OR} = \frac{a/c}{b/d} = \frac{a*d}{b*c} = \frac{\text{(Odds of Exposure among Cases)}}{\text{(Odds of Exposure among Controls)}}$$

What are Odds?

- Odds: Ratio of ways an event can occur to ways the event can not occur.
  - Odds of 1:1 indicate both options are easily likely.
  - When rolling dice, probability of getting a 2 is 1/6
  - Odds of getting a 2 is 1:5.
- Used in epidemiology, because of situations where we calculate odds ratios
  - Case-control studies
  - Logistic regression
- If P=probability of event, then odds= $\frac{P}{1-P}$
- If P is very small, 1-P≈1 and odds=

Odds Ratio in Cohort Study and Case-Control Study Reduce to Same Calculation

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Not Exposed	c	d

$$OR = \frac{\text{odds that an exposed person develops disease}}{\text{odds that a non-exposed person develops disease}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

A

	CASES (with disease)	CONTROLS (without disease)
History of exposure	a	b
No history of exposure	c	d

$$OR = \frac{\text{odds that a case was exposed}}{\text{odds that a control was exposed}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

B

Gordis: Epidemiology, 4th Edition.
Copyright © 2008 by Saunders, an imprint of Elsevier, Inc. All rights reserved

Confidence Intervals and p-values

- Presentation so far has focused on point estimates
- Gives information on magnitude of association
- Statistical software will also provide estimate of confidence intervals and p-values
- Important to consider precision and statistical significance, along with estimate of magnitude of association.

Odds Ratio Estimates Relative Risk When Disease is Rare

- The OR will be a good estimate of the RR if the outcome is rare.
- If the outcome is common, and association is positive, then the OR will overestimate the RR
- This overestimation can be quite large for common outcomes.

BMJ 1997;No 7121 Volume 315

Bias, Confounding, Effect Modification, and Causal Inference

Association and Causality

- An exposure and outcome are associated if there is a differential distribution:
  - Incidence of outcome differs for exposed and unexposed group; or
  - Prevalence of exposure differs between cases and controls
- An exposure is causal for the outcome if the presence (or absence) of the exposure directly or indirectly influences whether the outcome occurs.

Sources of Bias in Epidemiology

Bias = Systematic error in the design, conduct or analysis of a study that results in a <u>mistaken estimate of an exposure's effect on the risk of disease</u>

Phillips, C.V. 2003. Epidemiology. 14(4):459-466.

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.

*Some argue confounding is not technically a bias

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 cases[19].

Non-response (or respondent) bias. Differential rates of refusal or non-response to inquiries between cases and disease-free comparison subjects[19].

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 process[19].

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 subjects[19].

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 surveillance[20].

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 midst[19].

Exposure suspicion bias. Knowledge of a patient's disease status can influence the intensity and outcome of the search for exposure to a putative cause[19].

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

Confounding

- Confounding is a key topic in epidemiology
- A confounder is often defined as a factor that is:
  1. A risk factor for disease
  2. Associated with exposure
  3. Not a direct result of exposure
- Confounding can lead to "spurious" associations

Example of Confounding

- Birth order and Down syndrome
- Birth order is associated with Down syndrome, later order children with higher risk
  - Maternal age is associated with birth order
  - Maternal age is associated with Down Syndrome
- Stratifying on maternal age, there is no longer evidence of an association between birth order and Down syndrome

Approaches to Handling Confounding

In Design of Study	In Analysis of Data
• Randomization	• Standardization
• Restriction	• Adjustment

• Matching
  • Group Matching
  • Individual Matching
	• Stratification

Stratified Analysis

Crude

OR_{Crude}

Stratum 1 Stratum 2

OR_1 OR_2

Crude 2 x 2 table
Calculate Crude OR (or RR)

Stratify by Confounder

Calculate OR's
for each stratum

If stratum-specific OR's are similar,
calculate adjusted RR (e.g. MH)

If Crude OR =/= Adjusted OR,
confounding is likely

If Crude OR = Adjusted
OR, confounding is
unlikely

What is Meant by Interaction?

- **Biological Interaction**
  - The interdependent operation of two or more biological causes to produce, prevent or control an effect
  - Two causes interact on a biological level to cause a disease or outcome
- **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.

Guidelines for Judging Whether an Association is Causal

- Temporal relationship (exposure should proceed outcome)
- Strength of association (size of odds ratio or relative risk)
- Dose-response relationship
- Cessation of exposure leads to reduction in outcome
- Replication of finding (multiple independent studies)
- Biological plausibility
- Consistency with other knowledge
- Consideration of alternative explanations (ability to rule them out)
- Specificity of the association

Interpreting and Applying Results from Epidemiologic Studies

Gordis: Epidemiology, 4th Edition.
Copyright © 2008 by Saunders, an imprint of Elsevier, Inc. All rights reserved

COMMENT

Twenty tips for interpreting scientific claims

This list will help non-scientists to interrogate advisers and to grasp the limitations of evidence, say **William J. Sutherland, David Spiegelhalter** and **Mark A. Burgman.**

Calls for the closer integration of science in political decision-making have been commonplace for decades. However, there are serious problems in the application of science to policy — from energy to health and environment to education.

One suggestion to improve matters is to encourage more scientists to get involved in politics. Although laudable, it is unrealistic to expect substantially increased political involvement from scientists. Another proposal is to expand the role of chief scientific advisers', increasing their number, availability and participation in political processes. Neither approach deals with the core problem of scientific illiteracy among many who vote in parliaments.

Perhaps we could teach science to politicians? It is an attractive idea, but which busy politician has sufficient time? In practice, policy makers almost never read scientific papers or books. The research relevant to the topic of the day — for example, mitochondrial replacement, bovine tuberculosis or nuclear-waste disposal — is interpreted for them by advisers or external advocates. And there is rarely, if ever, a beautifully designed double-blind, randomized, replicated, controlled experiment with a large sample size and unambiguous conclusion that tackles the exact policy issue.

In this context, we suggest that the immediate priority is to improve policy-makers' understanding of the imperfect nature of evidence, to improve their ability to intelligently interrogate experts and advisers, and to understand the quality, limitations and biases of evidence. We term these interpretive scientific skills. These skills are more accessible than those required to understand the fundamental science itself, and can form part of the broad skill set of most politicians.

To this end, we suggest 20 concepts that should be part of the education of civil servants, politicians, policy advisers and journalists — and anyone else who may have to interact with science or scientists. Politicians with a healthy scepticism of scientific advocates might simply prefer to arm themselves with this critical set of knowledge.

We are not so naive as to believe that improved policy decisions will automatically follow. We are fully aware that scientific judgement itself is value-laden, and that bias and context are integral to how evidence is interpreted and applied. What we offer is a simple list of ideas that could help decision-makers to parse how evidence can contribute to a decision, and potentially to avoid undue influence by those with vested interests. The harder part — the social acceptability of different policies — remains in the hands of politicians and the broader political process.

Of course, others will have slightly different lists. Our point is that a wider

© 2012 Macmillan Publishers Limited. All rights reserved

Future Directions in Epidemiology

American Journal of Epidemiology
Vol. 177, No. 4
DOI: 10.1093/aje/kws414
Advance Access publication:
January 7, 2013
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Point-Counterpoint

Point: Is There a Future for Innovative Epidemiology?

Lewis H. Kuller*

* Correspondence to Dr. Le
Bellefield Avenue, Room 55

Initially submitted May 1

Cancer Epidemiology, Biomarkers & Prevention

Hypothesis/Commentary

Cancer Epidemiology in the 21st Century

Transforming Epidemiology for 21st Century Medicine and Public Health

Muin J. Khoury[1,5], Tram Kim
Stephen J. Chanock[2], Robe
Robert A. Hiatt[1,5], Robert N.
Jeffrey A. Meyerhardt[9], Oluf
Daniela Seminara[1], David F.
Ann Zauber[21], and Sheri D.

American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2012.
Vol. 175, No. 7
DOI: 10.1093/aje/kws138
Advance Access publication:
March 12, 2012

Commentary

Cardiovascular Epidemiology in a Changing World—Challenges to Investigators and the National Heart, Lung, and Blood Institute

Paul D. Sorlie, Diane E. Bild, and Michael S. Lauer*

Summary

- Epidemiology is the study of the distribution and determinants of health-related states in populations
- Historic examples demonstrate objectives of epidemiology
- Study design is a key component of epidemiology
- Relative risks, risk differences and odds ratios are used to measure association
- It is important to consider and address bias in epi studies
- Selection bias and information bias are two main classes of bias
- Understanding confounding and effect modification are important in studies of association
- Future directions are transforming the field of epidemiology

Definitions of Epidemiology

- Greek Etymology
  - Epi - upon, among, on, over
  - Demos- people, populance
  - Logos- study, word, discourse, count

- the study of the distribution and determinants of health-related states in specified populations, and the application of this study to control health problems - Last
- the study of how disease is distributed in populations and the factors that influence or determine this distribution – Gordis
- a branch of medical science that deals with the incidence, distribution, and control of disease in a population – Merriam-Webster
- Epidemiology is the study (or the science of the study) of the patterns, causes, and effects of health and disease conditions in defined populations. - Wikipedia

Genetic Epidemiology

Introduction

Karen L. Edwards, Ph.D.

Professor

Department of Epidemiology and

Genetic Epidemiology Research Institute

School of Medicine

University of California, Irvine
Irvine, CA

Lecture Outline

- Introduction to Genetic Epidemiology
  - Define genetic epidemiology
  - Terms and concepts important in genetic epidemiology

  http://www.genome.gov/Glossary/
  http://www.cdc.gov/excite/library/glossary.htm
  - Overview of study designs
  - Collecting family history information and pedigree drawing

Genetic Epidemiology

- Goals
  - To discover and characterize genetic susceptibility to health and disease in human populations
  - To identify interactions between genetic and environmental factors
- Use family based studies and studies of unrelated individuals
- Apply principals of epidemiology, biostatistics and genetics/genome science
- Rapidly evolving field

Percent of Obese (BMI ≥ 30) in U.S. Adults

Percent of Obese (BMI ≥ 30) in U.S. Adults

Percent of Obese (BMI ≥ 30) in U.S. Adults

Genetic Epidemiology

Exposure → Disease / Outcome

Genotype Phenotype

Phenotype and Genotype

- Phenotype and Genotype are the key components in genetic epidemiologic studies

- Phenotype (trait) – observed characteristics that are usually the focus of a genetic epidemiologic study
  - Not always a direct reflection of genotypes
  - Examples: Blood pressure, body weight, cholesterol level, eye color, heart disease, diabetes, Parkinson's disease, cancer, longevity
  - Quantitative vs. qualitative (discrete) trait

Chromosomes, DNA, Genes

- Chromosomes are made up of DNA and are long strands of "genes"
  - Humans have about 20,000 genes in their genome
  - Genes have both coding (exon), noncoding (intron) and regions upstream that affect expression (promoter region)
    - Promoter region – a sequence of DNA found near beginning of a gene and needed to turn a gene on or off
    - Exons – contain stretches of DNA that code for proteins
    - Introns come in between the exons – intervening sequences, do not code for proteins
  - Genes control growth, development, health and disease
  - Genes are turned on and off in different patterns and at different stages of development = gene regulation

Genotype

- The unique genetic information of an individual
- Each gene has its own specific location on the chromosome
- Genes come in pairs, one version of each gene is inherited from your mother and one from your father (allele)
- Variations in the underlying DNA can result in differences between individuals, and may underlie a specific phenotype
  - Different ways of measuring the genotype and alleles
  - Single nucleotide polymorphism – most SNPs are not themselves functional, but mark the functional variations that affect disease risk
  - Sequencing is now common
- Factors that affect the expression of the gene (such as environment) are also important to consider (Gene x environment (GxE))

Genotype

- Each person inherits 1 chromosome from their biologic father and one from their biologic mother
- 23 pairs of chromosomes (a total of 46 chromosomes)
  - 22 of the pairs look the same in males and females
  - The sex chromosomes differ in males (X,Y) and females (X, X)
- Genotype – the unique genetic information from an individual
  - The genetic contribution to the phenotype
  - Genotype can refer to a collection of genes or the two alleles of a particular gene
  - Humans have about 20,000 genes

A pair of chromosomes

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.

Identifying genetic effects: Overview

Approaches to understanding genetic influences: Overview of Genetic Epidemiologic Studies

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation?	Family Study
Is the familial aggregation caused by genetic factors?	Twin Study
Is there a major gene? Is it dominant or recessive ?	Segregation Study
Where is this major gene in the human genome?	Linkage Analysis
Is there linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there an increased allele sharing for affected relatives (sib pairs) or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the exact location of this gene and which polymorphism is associated with disease?	Association Study (population and family)

- Most human traits have a skewed distribution – which could be consistent with a genetic effect
- Body weight as our example

Basic idea behind Commingling Analysis:

- If a single gene has an effect on the variation in a quantitative trait (body weight), each genotype has a particular distribution associated with it
- The overall population distribution results from the commingling of these genotype-specific sub-distributions
- Assume a gene with alleles B and b: have 3 possible genotypes in the population:
  - BB - homozygous
  - Bb - heterozygous
  - bb - homozygous

A pair of chromosomes from an individual person

Body weight gene (alleles b and B - heterozygous)

- Is the frequency distribution of body weight consistent with the influence of a gene with either dominant or recessive effects?

Commingling Analysis Summary:

- Why is it used: To provide preliminary evidence for a single gene that influences a quantitative trait (e.g. body weight, blood pressure, cholesterol level, blood glucose level).
- A statistical modeling approach that does not measure the genotype, but assumes genetic principals in the model
- Unrelated individuals – faster and easier

Overview of Genetic Epidemiologic Study Design

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation? higher risk in relatives of cases	Family Study
Is the familial aggregation caused by genetic factors? MZ twins concordance rate or correlation higher than DZ twins	Twin Study
Is there a major gene? Is it dominant or recessive ?	Segregation Study
Where is this major gene in the human genome?	Linkage Analysis
Is there a linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there increased allele sharing for affected relatives or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the disease causing gene and which polymorphism is associated with disease?	Association Study (population and family-based)

General comments about twin studies

- One of the first approaches used to evaluate evidence for genetic influences on traits
- Evaluate both genetic and environmental influences on traits
- Measure of interest is the heritability of the trait
  - Proportion of total variance in the quantitative trait due to additive genetic effects
  - Population specific
- Evaluate evidence for genetic influences on different types of traits
  - qualitative traits – diabetes
  - quantitative traits – blood glucose

A twin approach to unraveling epigenetics

Jordana T. Bell and Tim D. Spector

Trends Genet. 2011 March ; 27(3): 116–125.

Abstract

The regulation of gene expression plays a pivotal role in complex phenotypes, and epigenetic mechanisms such as DNA methylation are essential to this process. The availability of next generation sequencing technologies allows us to study epigenetic variation at an unprecedented level of resolution. Even so, our understanding of the underlying sources of epigenetic variability remains limited. Twin studies have played an essential role in estimating phenotypic heritability, and these now offer an opportunity to study epigenetic variation as a dynamic quantitative trait. High monozygotic twin discordance rates for common diseases suggest that unexplained environmental or epigenetic factors could be involved. Recent genome-wide epigenetic studies in disease-discordant monozygotic twins emphasize the power of this design to successfully identify epigenetic changes associated with complex traits. We describe how large-scale epigenetic studies of twins can improve our understanding of how genetic, environmental and stochastic factors impact upon epigenetics, and how such studies can provide a comprehensive understanding of how epigenetic variation affects complex traits.

Overview of Genetic Epidemiologic Study Design

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation? higher risk in relatives of cases	Family Study
Is the familial aggregation caused by genetic factors? MZ twins concordance rate or correlation higher than DZ twins	Twin Study
Is there a major gene? Is it dominant or recessive ?	Segregation Analysis
Where is this major gene in the human genome?	Linkage Analysis
Is there a linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there increased allele sharing for affected relatives or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the disease causing gene and which polymorphism is associated with disease?	Association Study (population and family-based)

Overview of Genetic Epidemiologic Study Design

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation? higher risk in relatives of cases	Family Study
Is the familial aggregation caused by genetic factors? MZ twins concordance rate or correlation higher than DZ twins	Twin Study
Is there a major gene? Is it dominant or recessive ?	Segregation Study
Where is this major gene in the human genome?	Linkage Analysis
Is there a linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there increased allele sharing for affected relatives or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the disease causing gene and which polymorphism is associated with disease?	Association Study (population and family-based)

Association Studies

- Evaluate the association between a particular genetic variant and the trait (disease) in a population
- Focuses on unrelated individuals – usually case-control study
- Risk is typically estimated by the odds ratio (OR)
  - Compares the frequency of the genetic variant in those with disease to those without the disease
  - Measure of the strength of an association
    - OR=1 is no effect, OR>1 is increased risk, OR< is decreased risk
- A follow-up or replication study is an important, but challenging aspect of association studies

Summary

- Genetic Epidemiology - the genetic and environmental aspects of disease in human populations
- Use a variety of study designs to identify and evaluate evidence of genetic effects and impact on disease risk in populations
- Integrates epidemiology, genetics, genomics and biostatistics

Genetic Epidemiology

Collecting Family Data

Collecting Family Data

- Collecting family data is time consuming and expensive
- Need for complete and extended pedigrees
  - Local relatives vs. all relatives
  - Collection of phenotype data
- Need for accurate description of biologic relationships
- Confidentiality and IRB issues in collecting family data

Collecting family data

- IRB Issues
  - Confidentiality of information
    - Publication of pedigree information, genetic status
  - Sensitive information
    - Non-paternity, adoptions, abortions, medical conditions
- General approaches to data collection
  - Proband contact
  - Individual family members as contacts

Collecting family data

- Phenotype Information
  - Survey
    - Proband only
      - Pro: quick and inexpensive
      - Con: lack of knowledge about some relatives
    - Relatives
    - Critical information
      - Mother and Father ID for ALL related individuals
  - Measurements
    - Collecting blood / tissue samples
    - Physical measurements (height, weight, etc)
    - Tools for standardized measures (www.phenxtoolkit.org)
      - PhenX Tool Kit – a catalog of high-priority measures for consideration and inclusion in genetic epi studies

Family Studies: Family Health History, Segregation and Linkage Analysis

Karen L. Edwards, Ph.D.

Professor

Dept of Epidemiology and

Genetic Epidemiology Research Institute

School of Medicine
University of California Irvine
Seattle, WA

Overview of Genetic Epidemiologic Study Design

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation? higher risk in relatives of higher correlation in relatives	Family Study
Is the familial aggregation caused by genetic factors? MZ twins concordance rate or correlation higher than DZ twins	Twin Study
Is there a major gene? Is it dominant or recessive ? (likelihoods of Mendelian models higher than environmental or polygenic model)	Segregation Analysis
Where is this major gene in the human genome?	Linkage Analysis
Is there a linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there an increased allele sharing for affected relatives (sib pairs) or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the disease causing gene and which polymorphism is associated with disease?	Association Study (population and family-based)

Family Health History: Application to public health

Advantages:

- Reflects multiple genetic, environmental, behavioral factors and interactions
  - No genetic test can do this
- Family history is a predictor of most diseases (diabetes, cancers, CVD)
- Effective (public health) interventions exist for many of these diseases
  - Quitting smoking, maintaining ideal body weight, diet, exercise
- Overcomes one of the most important barriers - getting people interested in learning and talking about their health

Goal: Use family history information to motivate behavior change and promote a healthy lifestyle for primary prevention of disease

- More personalized health messages that " fit within pre-existing beliefs about current health status, possible causes and risk factors, course of the disease, magnitude of and potential consequences of the risk, and ways to reduce the risk" See Claassen *et al. BMC Public Health 2010,* ***10:248***

Genetic Epidemiology

Segregation Analysis

Complex Segregation Analysis (CSA)

- A modeling approach used to determine whether there is evidence for a single gene that underlies a trait or disease
  - Also provides information on mode of inheritance
    - Dominant, Recessive or Codominant
- General method for evaluating the transmission of a trait within pedigrees
  - Mendelian transmission

CSA, cont

- Information from CSA is useful in model based (parametric) linkage methods
  - LOD method linkage analysis depends on the specification of a reasonable model, including an approximation of the mode of inheritance
    - Assumes the existence of a Mendelian trait

The goal

- To test for compatibility with Mendelian expectations by estimating parameters for a range of genetic models
- CSA can provide the statistical evidence for Mendelian control of a trait or disease
  - As with all methods so far, this evidence can be used to support a genetic cause of the disease, but is not definitive
  - Simultaneously considers major locus, polygenic and environmental effects

The Approach

- A variety of models are fit to the family data and compared using a likelihood ratio test (for nested models)
  - The null hypothesis is that the data DO fit with some model of inheritance (genetic or not)- a "goodness of fit" approach

The Models

- The models are formed by estimating and restricting a specified set of parameters
  - The most general model, where all parameters are estimated
  - Single locus models with no polygenic inheritance and differing modes of inheritance
  - Polygenic model, with no single locus effect
  - Mixed model, both single gene and polygenic components
  - Nongenetic model or "environmental model"

Parameters: single locus component

- Means (*u*) for each subdistribution
- Variance of each subdistribution
- Allele frequencies
- Transmission probabilities - should conform to Mendelian expectations
  - t_1 = P(AA parents transmits A allele to offspring) = 1.0
  - t_2= P(Aa parents transmits A allele to offspring) = 0.5
  - t_3 = P(aa parents transmits A allele to offspring) = 0.0

Parameters : Polygenic component

- Heritability (h^2)
  - proportion of variance due to additive genetic effects
  - Not a single major gene
  - Can reflect "residual genetic effects" not accounted for by a single major locus
  - Sometimes referred to as multifactorial component

Model Testing

- Hypothesis testing for nested models using the LRT (likelihood ratio test)
- LRT = -2 [In L(reduced model) - In L(full model)]
  - LRT is distributed as a chi square with the degrees of freedom (df) equal to the difference in the number of estimated parameters
  - The likelihood of each model is proportional to the probability of the data, given the model and family structure

Model testing, cont

- To compare non-nested models
  - use the AIC to compare (not test) models to support a particular model over another
    - AIC= -2(ln likelihood) + 2(number of estimated parameters)
    - Calculate the AIC for each competing model and select the one with the smallest AIC as being the most parsimonious

Interpretation: Inferring A Major Gene

- To infer a major gene
  - reject nongenetic models
  - accept a major gene model (single or mixed model)
  - should always test transmission probabilities in CSA of quantitative traits to safeguard against false inference of a major gene

Ascertainment Correction

- Ideal probands would be newly diagnosed, population based (incident) cases
- Should correct for ascertainment unless pedigrees (probands) are selected from a random, population based sample
- Correction for ascertainment is not straightforward and is not usually done
  - Estimators for population parameters (allele frequency and heritabilty) will be most affected

Review Table

Other Issues to Consider

- Nonpaternity seems to have little effect on the ability to select models
- Can adjust for covariate effects
- Can also consider adjusting for other known genetic factors affecting your trait of interest

Important Limitations in CSA

- Implicit assumption of etiologic homogeneity
- Power is difficult to estimate as there is no single nongenetic alternative model, but instead a range of competing models
- Sample size
  - Larger extended kindreds with several generations are generally better than small nuclear families
  - generally requires a large amount of data, with more complex models requiring more data

Summary of CSA

- Does not require genotype data
- Can be time consuming to complete analyses
- Information from CSA is useful for a variety of reasons
  - Preliminary data, estimates for linkage analyses, choice of phenotype
- Assumes the existence of a Mendelian trait

In Class Exercise

Pedigree drawing demo using Progeny software tool

Draw your pedigree

Standardized Human Pedigree Nomenclature: Update and Assessment of the Recommendations of the National Society of Genetic Counselors.
Authors: Bennett, French, Resta, Lochner Doyle

- Standard format and nomenclature for drawing pedigrees
- Pedigrees convey lots of information
  - Picture is worth a 1000 words
  - Sensitive information and how to display?

J Genet Counsel (2008) 17:424–433

Bennett article - some key points

- A medical pedigree is a graphic presentation of a family's health history and genetic relationships
- A pivotal tool in the practice of medical genetics / genetic epi research
- Interpreting a pedigree should be a standard competency of all health professionals
- Pedigrees should not contain information about which a subject had no prior knowledge.
  - a person who had presymptomatic or susceptibility genetic testing through research should not find out about increased or decreased disease risk status from a publication

In Class Exercise: Pedigree Drawing

Let me start with my great-great grandparents: Jim and Ann Flight.

They had two children: Kathy, and Gerry.

Kathy died in a car accident along with her father Jim.

Gerry married Kate Doe.

Kate and Gerry had one child, Kathy

Kathy Flight married David Dewey and they had my dad, Bob. My dad took his mother's maiden name because David had an affair with someone named Maggie Braun.

After Jim's death, Ann married Paul Wright. Ann and Paul had one child: Tom Wright.

Tom Wright married Kaisa Stone.

Tom and Kaisa had one daughter: Heather. Heather Wright was wed to Peter Meter and had one child, Jean. Jean married Bob Flight and they had me Jane Flight.

Genetic Epidemiology

Linkage Analysis

Linkage Analysis, overview

- Linkage
  - Location of genetic loci sufficiently close together on a chromosome that they do not segregate independently
  - linkage is a property of loci (not alleles), and evaluation involves all alleles at the marker locus
  - the specific alleles segregating in one family may differ from alleles at the same locus segregating in a <u>different</u> family

Linkage vs. Association

- Linkage
  - Cosegregation of a disease or trait with a specific chromosomal region in multiple families
  - Genetic linkage is the tendency of two loci to be inherited together (e.g. loci are on the same chromosome)
  - Property of two loci (genes or locations)
- Association
  - Presence of a disease or trait with a specific allele in a gene or marker (in unrelated subjects) – probably due to linkage disequilibrium

Linkage Analysis –background

- The aim of linkage analysis is to infer the relative position of two or more loci
  - Examining patterns of allele sharing or cosegregation of marker and disease in relatives
  - The location of one locus is known (the marker), the other is unknown (the disease causing gene)
  - Alleles of loci on the same chromosome can violate Mendels's law of independent assortment (linkage)
- Evidence of linkage between a known marker and a putative gene for a disorder is the ultimate statistical evidence for a genetic component in disease etiology

General Approaches to Linkage Analysis

- Genome Wide Scan
  - Isolate a gene solely on the basis of it's chromosomal location, without regard to it's biochemical function.
  - This is often referred to as the "positional genetic" approach (i.e. genome screens are often referred to positional cloning)
- Candidate gene approach
  - Select candidate genes based on their function or other known properties

Required data for family studies

- At least pairs of related individuals
- Accurate pedigree structure / biological relationships
  - Nuclear family vs. extended kindred
- Phenotype data – quantitative or categorical
- Genotype data
  - Location of markers (marker map)

Genetic Markers

- A genotype (measurable "trait") that is genetically determined, can be accurately classified, has a simple, unequivocal pattern of inheritance (*and polymorphic).*

- Types of genetic markers
  - Polymorphic markers – lots of alleles / variation
    - Variable number of tandem repeats (VNTR)
    - Microsatellites, (e.g. CA repeats), very polymorphic
  - Single nucleotide polymorphisms (SNP's) - 2 allele markers, very common
  - Sequence data – exome or whole genome

Statistical Analysis: LOD based Linkage Analysis

- Involves comparison of likelihoods of observing the segregation pattern of 2 loci under specific models, including
  - Under the null hypothesis of no linkage
    - Independent assortment – loci recombine as if on different chromosomes
  - Alternative hypotheses of linkage
    - differ in the extent of crossing over (i.e. different values of recombination events)

LOD Score

- LOD score = log (base 10) of the odds of linkage vs. no linkage (not an odds ratio!)
  - LOD score $\geq$ 3, supports linkage, corresponds to a genome-wide type 1 error rate of 0.05 (depends on number of markers tested)
  - LOD score $\leq$ -2, used to exclude a chromosomal region
    - Exclusion mapping
- add LOD scores from all families to obtain LOD score for your sample
  - Assumes families are independent

Linkage Mapping of CVD Risk Traits in the Isolated Norfolk Island Population

Abstract: To understand the underlying genetic architecture of cardiovascular disease (CVD) risk traits, we undertook a genome-wide linkage scan to identify CVD quantitative trait loci (QTLs) in 377 individuals from the Norfolk Island population. The central aim of this research focused on the utilization of a genetically and geographically isolated population of individuals from Norfolk Island for the purposes of variance component linkage analysis to identify QTLs involved in CVD risk traits.

-The ancestral origins of the Norfolk Island are well documented and originated from divergent founding paternal and maternal lineages, European and Tahitian, respectively.

-1,574 residents

-Exhaustive genealogical documents indicate that the population grew from a limited number of initial founders (nine males, twelve females) and in relative isolation in the early generations of population expansion

- Evidence of the Island's strict immigration laws are obvious by the limited numbers of surnames, resulting in the worlds only telephone directory which includes nicknames to differentiate between individuals with the same name

Hum Genet. 2008 December ; 124(5): 543–552. doi:10.1007/s00439-008-0580-y.

Linkage Mapping of CVD Risk Traits in the Isolated Norfolk Island Population

The Norfolk Island genealogy dates back approximately ten generations to the initial founders and contains 6379 individual entries linked together within 2185 nuclear families. The complexity of the island's heritage is evident considering 5750 individuals reside within a single multifamily pedigree exhibiting 1661 marriages and 1233 founders.

Methods: Substantial evidence supports the involvement of traits such as systolic and diastolic blood pressures (SBP and DBP), high-density lipoprotein-cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), body mass index (BMI) and triglycerides (TG) as important risk factors for CVD pathogenesis. In addition to the environmental influences of poor diet, reduced physical activity, increasing age, cigarette smoking and alcohol consumption, many studies have illustrated a strong involvement of genetic components in the CVD phenotype through family and twin studies. We undertook a genome scan using 400 markers spaced approximately 10cM in 600 individuals from Norfolk Island. Genotype data was analyzed using the variance components methods of SOLAR.

Results: Our results gave a peak LOD score of 2.01 localizing to chromosome 1p36 for systolic blood pressure and replicated previously implicated loci for other CVD relevant QTLs.

Hum Genet. 2008 December ; 124(5): 543–552. doi:10.1007/s00439-008-0580-y.

Sib-Pair Linkage Analysis

- Sib pairs are generally easier to collect, tend to be more closely matched for age and environment than other relative pairs
- Qualitative trait: under linkage, Affected relative pairs should share alleles IBD (inherited from a common ancestor within the pedigree), more often than expected under Mendelian expectations
- Quantitative trait: relative pairs should show a correlation between the magnitude of their phenotypic difference and the number of alleles shared IBD

Quantitative sib-pair linkage

- A regression approach
  - Regress the squared within-pair difference of a quantitative trait on the number of marker alleles shared IBD
- Null hypothesis - the slope of the squared within pair difference is zero
- The alternative hypothesis is that under linkage, the slope is negative.

Identity by descent vs. Identity by state

- IBS- two alleles at a given locus are identical in state if they represent the same allelic variant at that locus
- IBD- two alleles at a given locus are IBD if they were transmitted from a common ancestor –ie they represent copies of the same ancestral DNA

Quantitative Sib-pair linkage results

Linkage Disequilibrium

Outline

- Linkage disequilibrium (LD)
  - Definition of linkage disequilibrium
  - Importance of disequilibrium
  - Measures of disequilibrium
- SNP selection
  - Public resources
  - Tag SNP selection programs
- Imputation

Definitions

SNP1: rs3822050 and SNP2: rs10517002

- **Allele**
  - Different versions of DNA sequence at a given location

SNP1: **C** and **T**
SNP2: **C** and **A**

- **Genotype**
  - The two alleles in an individual at a given locus

SNP1: **C/C, C/T or T/T**
SNP2: **C/C, C/A or A/A**

- **Haplotype**
  - A series of alleles along a single chromosome

- **Diplotype**
  - a set of haplotype pairs in an individual

What is Linkage Disequilibrium?

Linkage Disequilibrium: Two loci that are in linkage disequilibrium are inherited together more often than would be expected by chance.

Zondervan & Cardon, 2004

Systematic studies of common genetic variants are facilitated by the fact that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as **linkage disequilibrium**

The international HapMap consortium, 2005

Linkage Disequilibrium refers to the nonindependence of alleles at different sites.

Pritchard and Przeworski 2001

Linkage Equilibrium

SNP1: **C/T**
SNP2: **C/A**

haplotype frequencies in population match what is expected based on allele frequencies
Example: frequency of C-A haplotype equals frequency of C allele at SNP 1 * frequency of A allele at SNP 2

Linkage Disequilibrium

SNP1: **C/T**
SNP2: **C/A**

haplotype frequencies in population differ from what is expected based on allele frequencies

It is a Matter of Scale

"*Nothing in biology makes sense except in the light of evolution*"

-Theodosius Dobzhansky, 1973

Novembre J, Ramachandran S. 2011.
Annu. Rev. Genomics Hum. Genet. 12:245–74

Current Haplotypes Arose from Ancient Mutation Events

1. Ancestral state has no variation at either SNP position.

2. Mutation leads to first SNP

3. Asecond mutation leads to second SNP

4. Recombination or recurrent mutation needed for all four haplotypes

Haplotypes

The International HapMap Consortium. Nature | Vol 437 | 27Octob

Focus on Pairwise LD

	A	a	
B	p_{AB}	p_{aB}	$\mathbf{p_B}$
b	p_{Ab}	p_{ab}	$\mathbf{p_b}$
	$\mathbf{p_A}$	$\mathbf{p_a}$	

If loci are independent, then we expect

$$p_{AB} = p_A * p_B$$

$$p_{Ab} = p_A * p_b$$

$$p_{AB} = p_A * p_B$$

$$p_{AB} = p_A * p_B$$

Measuring LD for pairs of sites- D

	A	a	
B	p_{AB}	p_{aB}	$\mathbf{p_B}$
b	p_{Ab}	p_{ab}	$\mathbf{p_b}$
	$\mathbf{p_A}$	$\mathbf{p_a}$	

One important measure of LD is

$$\mathbf{D_{AB} = p_{AB} - p_A p_B}$$

Notice that D=0 if and only the two sites are independent

A disadvantage of D is that the range of possible values depends greatly on the marginal allele frequencies.

Measuring LD for pairs of sites- D'

Lewontin (1964) proposed an adjusted statistic that has range [-1, 1]:

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

D' = D/max(D), where max(D) is dependent on the marginal allele frequencies

If $D_{AB}>0$: $D'_{AB} = D_{AB}/(\min(P_aP_B, P_AP_b))$
If $D_{AB}<0$: $D'_{AB} = D_{AB}/(\min(P_AP_B, P_aP_b))$

Properties of D'

- D' favored in medical genetics
  - D'=0 implies independence
  - |D'|<1 implies that there has been recombination between the two sites in the history of the sample (or recurrent mutation)
  - |D'=1| implies "complete LD"
    - No historic recombination
    - Neither site has experienced recurrent mutation or gene conversion
    - Genotypes not perfectly correlated (unequal allele frequency)
    - D' inflated in smaller samples

Measuring LD for pairs of sites- r^2

Along with D', the other most widely used statistic is r^2:

$$r^2 = D_{AB}^2 / (p_A * p_B * p_a * p_b)$$

r^2 has range [0,1]. Its value is 1 if just 2 of the 4 haplotypes are present.

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

r^2 is intimately connected to the power of association mapping [Pritchard & Przeworski 2001]

Properties of r^2

- r^2 favored in population genetics
  - r^2 =0 implies independence
  - r^2 =1 implies "perfect LD"
    - Marker loci have identical allele frequencies
    - Genotype is perfectly correlated
  - Related to power if ($N_2=N_1/r^2$)
    - where N_1 is sample size needed for directly genotyped SNP, N_2 is sample size needed to test tagged SNP and r2 is the LD between the directly genotyped SNP and the tagged SNP).
    - Assume need 1,000 for directly genotyped SNP, examples of sample size needed for tagged SNPs, depending on r^2
      - r^2=1.0, N_1= N_2=1,000
      - r^2=0.2, N_2 = 1,000/0.2 = 5,000

Complete LD

- No detected recombination between SNPs
- Unequal allele frequency
- Allelic association is as strong as possible
  - 3 haplotypes observed
  - Genotype is not perfectly correlated
  - D'=1, r^2<1

Slides from C. Carlson

Perfect LD

- No detected recombination between SNPs
- Equal allele frequency
- Allelic association is as strong as possible
  - 2 haplotypes observed
  - Genotype is <u>perfectly</u> correlated
  - D'=1, r^2=1

Slides from C. Carlson

What factors affect LD?

- Mutation
- Historical recombination
- Natural selection
- Founder effects
- Migration
- Random drift
- Population admixture

LD over time

- Recombination assorts SNPs on haplotypes.
- Under assumption of random mating and a large population, LD will break down over time.

Applications of LD

- LD is the *sine qua non* of genetic association studies:
  - We are interested in testing for an association between disease status and causal mutations
  - If all polymorphisms were independent at the population level, association studies would have to examine every one of them.
  - Instead we can test a subset and get information on all of them.

- LD is also used in studies of human history, natural selection and the biology of recombination

LD Across a Gene

Genotype at one site can predict genotype at another site

SNP Selection

- We use information about allele frequencies and LD across the genome to make informed choices as to which variants to genotype
  - Identify SNPs in region of interest
  - Interested in minimal set of SNPs needed to capture variation in region.

Identify variation for your region

- Option 1: sequence individuals in your sample for the entire gene/region of interest
- Option 2: sequence a subset of individuals to identify variation in your region
- Option 3: Use public databases to identify known variation in your region

SNP Database Resources

- NCBI SNP Database, dbSNP
  - http://www.ncbi.nlm.nih.gov/SNP/
- International HapMap Project
  - http://www.hapmap.org/
- NHLBI Program for Genomic Applications (http://www.nhlbi.nih.gov/resources/pga/)
  - SeattleSNPs (http://pga.mbt.washington.edu/)
  - InnateImmunity (http://innateimmunity.net/)
- 1,000 genomes project
  - http://www.1000genomes.org
- Exome variant server (EVS)
  - **ADD LINK**

Tag SNPs

- tagSNPs
  - SNPs are selected based on their pair wise ability to predict genotype of untyped SNPs
  - Based on an r2 concept of LD structure
  - Example program: LDSelect
- haplotype-tagging SNPs (htSNPs)
  - SNPs are selected to optimize resolution of existing haplotypes
  - Based on a D' concept of LD structure
  - Example program: Haploview, HaploBlockfinder
- Multi-marker tagSNPs
  - Use tagSNP concept, but extend past pair wise LD
  - Example program: tagger

Tag SNPs – using r^2 information

Think-Pair-Share Exercise:
Which SNPs are in high LD?
How many SNPs would you need to genotype to effectively capture the variation across the region?

After Carlson *et al.* (2004) *AJHG* **74**:106

Tag SNPs – using r^2 information

Tags:

Test for association:

After Carlson *et al.* (2004) *AJHG* **74**:106

Tag SNPs are Population Specific

European-Americans
CRP

African-Americans
CRP

Frazer, KA et al.
Human genetic variation and its contribution to complex traits
Nature Reviews Genetics 10, 241-251 (April 2009)

Thousand Genomes and GVS Tutorial

Limitations of tag SNPs

- Ultimately, we are interested in identifying common polymorphisms that are causally associated with disease risk, we cannot determine if signal is from the tagSNP or from a correlated SNP.
- What happens if your tagSNP fails in the genotyping/QC stage?

Imputation

- We also use LD information to impute genotype information.
- Common example is in genome-wide association studies.
  - Example: 300,000 genotyped SNPs can be used to infer 2.4 million SNPs in HapMap data
- Exploits LD and haplotype information.

Imputation with family data

Figure 1

Genotype imputation within a sample of related individuals. (*a*) The observed data, which consist of genotypes at a series of genetic markers. In this case, a subset of markers has been typed in all individuals (*red*), whereas the remaining markers have been typed in only a few individuals (*black*, in individuals in the top two generations of the pedigree). (*b*) The process of inferring information on identity-by-descent by examining markers for which genotypes are available in all individuals. Each IBD segment that appears in more than one individual is assigned a unique color. For example, a segment marked in blue is shared between the first individual in the grandparental generation at the top of the pedigree, the first individual in the parental generation, and individuals 3 and 4 in the offspring generation at the bottom of the pedigree. (*c*) Observed genotypes and IBD information have been combined to fill in a series of genotypes that were originally missing in the offspring generation.

Imputation with Population Data

Observed Data (GWAS Target Panel)

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Data (Sequencing data)

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


Identify Matching Sequences

Observed Data (GWAS Target Panel)

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Data (Sequencing data)

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

Infer Missing Data

Observed GWAS and Inferred Data

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

Reference Data (Sequencing data)

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

Imputation

- Imputation can be computationally intensive
- When possible, use a reference data from the same source population as your imputation sample (i.e. match on ancestral groups)
- Imputation output:
  - A “best guess” genotype (i.e. TT)
  - Probability of each genotype (pr(TT), pr(TA), pr(AA)
  - A quality score (typically a r2 measure)
- Often better to use probabilities, or inferred dosage based on probabilities, in analysis, rather than “best guess”
  - If coded 0=AA, 1=TA, 2=TT; dosage=pr(TA)+2*pr(TT)

Imputation Programs

- IMPUTE2
  - http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- Beagle
  - http://faculty.washington.edu/browning/beagle/beagle.html
- MaCH/minimac
  - http://genome.sph.umich.edu/wiki/MaCH:_1000_Genomes_Imputation_Cookbook
  - http://genome.sph.umich.edu/wiki/Minimac

Summary

- Linkage disequilibrium (LD) refers to the nonindependence of alleles at different sites in the genome
- LD is shaped by population genetic forces
- We exploit LD information in genetic epidemiology
  - Selecting tagSNPs for association studies
  - Imputation in GWAS studies
- LD complicates interpretation of association studies

Tag SNPs – using r^2 information

Tags:

SNP 1
SNP 3
SNP 6

3 in total

Test for association:

SNP 1 captures 1 & 2
SNP 3 captures 3 & 5
SNP 6 captures 4 & 6

After Carlson *et al.* (2004) *AJHG* **74**:106

Picking tag SNPs using multimarker r2

Tags:

SNP 1
SNP 3

2 in total

Test for association:

SNP 1 captures 1+2
SNP 3 captures 3+5
SNP 1 and 3 in combo also captures 4 and 6

http://www.broad.mit.edu/mpg/tagger

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

Over of Genetic Epidemiologic Studies

Question	Approach
Is there evidence for genetic influences on a quantitative trait?	Commingling
Is there familial aggregation? higher risk in relatives of higher correlation in relatives	Family Study
Is the familial aggregation caused by genetic factors? MZ twins concordance rate or correlation higher than DZ twins	Twin Study
Is there a major gene? Is it dominant or recessive ? (likelihoods of Mendelian models higher than environmental or polygenic model)	Segregation Study
Where is this major gene in the human genome?	Linkage Analysis
Is there linkage with DNA markers under a specific genetic model?	A. Parametric Approach
Is there an increased allele sharing for affected relatives (sib pairs) or for relatives with similar phenotype	B. Allele Sharing Approach (sib-pair analyses)
Where is the (exact) location of this gene and which polymorphism is associated with disease?	Association Study (population and family)

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 <u>population</u> 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 cosegregation 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 polymorphisms (alleles) in a gene 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 <u>the</u> 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 Haplotvpe	Gm Haplotype	%NIDDM
Total (crude)	6.0	Present	8%
		Absent	29%

Population Stratification

- Example from Knowler et al.,

Pima ancestry	%Gm Haplotvpe	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

	Cohort		Case-Control		
Genotype	Disease risk risk	Relative Risk (RR)	Frequency in cases	Frequency in controls	OR
NN	I_o	1	A_1	B_1	1
NS	I_1	I_1/I_o	A_2	B_2	A_2B_1/A_1B_2
SS	I_2	I_2/I_o	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-Θ / 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 (<2cM), 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 uncertainity by considering a range of the parameter values
  - Can report "worst-case scenario"
  - Show power <u>over the range of values</u> indicating median power and/or sample size
- Number of software programs

Power and Type 1 Error

- For any question you have 2 hypotheses:
  - Ho: There is no association between disease x and marker y
  - Ha: 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 (Ho and Ha)

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 Ha 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 Ho 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

Nature Reviews | Genetics

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

Schematic of Typical GWAS

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

POINTS TO CONSIDER IN CONDUCTING AND INTERPRETING GWAS

Points to Consider for GWAS

- Study Designs Used in GWAS
- Genomic Coverage of GWAS Chips
- Genotyping and Quality Control in GWAS
- Replication
- Analysis and Presentation of GWAS Results

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

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

- Genomic coverage of GWAS genotyping platforms is often estimated by the percent of common SNPs having an r^2 of 0.8 or greater with at least 1 SNP on the platform.
- Genotyping platforms comprising 500,000 to 1,000,000 SNPs have been estimated to capture 67% to 89% of common SNP variation in populations of European and Asian ancestry and 46% to 66% of variation in individuals of recent 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, especially if occurring differentially between cases and controls, are an important cause of spurious associations and 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

PERSPECTIVE

DRINKING FROM THE FIRE HOSE — STATISTICAL ISSUES IN GENOMEWIDE ASSOCIATION STUDIES

STATISTICS AND MEDICINE

Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

David J. Hunter, M.B., B.S., and Peter Kraft, Ph.D.

Related article, page 443

"There have been few, if any, similar bursts of discovery in the history of medical research…"

The past 3 months have seen the publication of a series of

ating the need for guessing which genes are likely to harbor variants

The main problem with this strategy is that, because of the

ost stud-

ained in

samples

ower to

and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of as-

lated to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.[2] Such findings promise to open up new avenues of research, through both the discovery of new genes rele-

generate P values as small as 10^{-7}. In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as 10^{-7} in an initial study. On the other hand, a "statistically significant" finding

Hunter DJ and Kraft P, *N Engl J Med* 2007; 357:436-39

Slide courtesy of T. Manolio

Schematic of Typical GWAS

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

Manhattan Plot

Population based
Association studies

Compare genotypes in cases and controls

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

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

Nature 466, 113–117 (01 July 2010)

Regional Plots

Nature Genetics 43, 539–546 (2011)

Q-Q plots

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

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

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.

Obtaining the Summary Estimate

Weighted average of the individual study estimates

Using the Odds Ratio (OR) as our example:

Assume we have *k* studies, and refer to the study specific OR for the *i*th study as OR_i

- Assign each study a study specific weight w_i
- Calculate the summary OR (OR_{meta}) as:

$$\ln(OR_{meta}) = \frac{\sum_{i=1}^{k} w_i \ln(OR_i)}{\sum_{i=1}^{k} w_i}$$

Obtaining the Summary Estimate

- Similarly, there are formulas for calculating the standard error, 95% confidence intervals and p-values.
- Need to choose how to calculate weights.

  Multiple options: Inverse Variance is most commonly used

  – **Fixed Effects and Random Effects:** Estimate different parameters

Fixed Effects and Random Effects

Fixed Effects

Estimate: the average effect size

Source of variance: within study variance

Random Effects

Estimate: the mean of a distribution of effect sizes

Sources of variance: within study variance and between study variance

Borenstein, M. et al. Research Synthesis Methods 2010; 87-111

Inverse Variance Weights

Fixed Effects

$$w_i = \frac{1}{\sigma_i^2}$$

Random Effects

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}$$

Within study variance (standard error squared) Different value for each study

Between study variance The same value for each study in the meta-analysis

$$\ln(OR_{meta}) = \frac{\sum_{i=1}^{k} w_i \ln(OR_i)}{\sum_{i=1}^{k} w_i}$$

Inverse Variance Weights

Fixed Effects

$$w_i = \frac{1}{\sigma_i^2}$$

Random Effects

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}$$

If there is no between study variance, the random effects reduces to fixed effects

$$\ln(OR_{meta}) = \frac{\sum_{i=1}^{k} w_i \ln(OR_i)}{\sum_{i=1}^{k} w_i}$$

Meta-Analysis Results: The Forest Plot

Seven Prostate Cancer Loci. Nat Genet. 2011 July 10; 43(8): 785–791.

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

- 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

- 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.

GWAS FINDINGS

2005
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

2006
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

2007 4th quarter

2008 4th quarter

2009 4th quarter

2010 4th quarter

Published Genome-Wide Associations through 02/2012,
1,184 publications on >200 Traits

2011 2nd quarter

NHGRI GWA Catalog
www.genome.gov/GWAStudies

Published Genome-Wide Associations through 12/2012
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Type 2 Diabetes Timeline

44 SNPs account for ~ 10% heritability

Slide courtesy of Josh Bis. Data from Wheeler E , Barroso I. Briefings in Functional Genomics 2011;10:52-60

GWAS Findings in the Literature

- More than 1,700 publications and 11,000 SNPs
- Associations with >300 complex diseases and traits.
- Countries of recruitment dominated by Europe and North America.

Pace of genome-wide association study publications since 2005.

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

Traits in the GWAS Catalog

Slide courtesy of L. Hindorff and H. Junkins

NHGRI Catalog of Published GWAS

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

NHGRI Catalog of Published GWAS

genome.gov
National Human Genome Research Institute
National Institutes of Health

SEARCH

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

Home > Research Funding > Research Funding Divisions > Division of Genomic Medicine > **GWAS Catalog**

Share Print

Division of Genomic Medicine

A Catalog of Published Genome-Wide Association Studies

Division Staff | Funding Opportunities | Genomic Medicine Activities | GWAS Catalog | Meetings & Workshops | Potential Sample Collections for Sequencing | Programs | Publications | Trans-NIH Sequencing Inventory

Current uses of and future directions for the Genome-Wide Association Studies Catalog

On Thursday, July 18th, 2013, the Division of Genomic Medicine held a webinar to highlight current uses and explore priorities and future directions for the GWAS catalog. See archived video and presentations.

Additional information has been added to the HTML catalog columns below. For a description of column headings for the HTML catalog, go to: Catalog Heading Descriptions

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits

Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

View the Interactive Diagram | View the Full Catalog | Download the Catalog | Search the Catalog

The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed

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

Search By:

Journal: Select Journal

First Author: (last name)

Disease/Trait: (string search)

Tip: Expand your search by using the OR operator (returns results with either term), or narrow your search using the AND operator (returns results with both terms).

or

β2-Glycoprotein I (β2-GPI) plasma levels
5-HTT brain serotonin transporter levels
AB1-42
Abdominal aortic aneurysm
Acenocoumarol maintenance dosage
Activated partial thromboplastin time
Acute graft versus host disease
Acute lung injury
Acute lymphoblastic leukemia (childhood)
Addiction
Adiponectin levels
Adiposity

Tip: Hold Ctrl-key to select multiple entries.

Chromosomal Region: (e.g., "13q21.31")

Gene: (e.g., "LRP5")

SNP: (e.g., "rs20755555")

The SNP data in the catalog has been mapped to dbSNP Build 137 and Genome Assembly, GRCh37/hg19.

OR greater than:

p-Value threshold: Enter the exponent. For example, enter "5" for $p<10^{-5}$

Search | Reset

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

http://epi.grants.cancer.gov/pgwas/index.html

Discovery and Replication

- Finding of new associations through pooled analyses
- Independent replication of associations
- Fine Mapping of association signals

Biological Studies

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

Epidemiology

- Evaluation of gene-gene and gene-environment interactions
- Assessment of penetrance and population attributable risk
- Development of complex risk models and evaluation of their clinical validity

Clinical and Prevention Applications

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

www.nature.com/icogs/

Key COGS Findings

Breast Cancer

- GWAS meta-analyis 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

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.

Variants in Regulatory Regions

- 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.

Post-GWAS Gene-Environment Studies

DISCOVERY	CHARACTERIZATION
• Identify novel loci	• Describe interaction
• Focus on variants that would not be found in marginal search alone	• Focus on putative and established variants
• Priority given to power	• Priority given to descriptive model
• Hypothesis generating	• Provides etiologic and public health insight

Hutter et al. Genet Epi. 2013; 37: 643–657.

BRINGING GENOME-WIDE ASSOCIATION FINDINGS INTO CLINICAL USE

Biological Insights From GWAS

- Reveled new biology
- Identified a small number of potential therapeutic targets
- Facilitated basic research in human genetics and genomics

Table 1. Examples of Previously Unsuspected Associations between Certain Conditions and Genes and the Related Metabolic Function or Pathway, According to Genomewide Association Studies.

Condition	Gene	Function or Pathway	Source of Data
Age-related macular degeneration	*CFH*	Complement-mediated inflammation	Klein et al.[25]
Coronary disease	*CDKN2A, CDKN2B*	Cell-cycle regulator	Helgadottir et al.[36]
Childhood asthma	*ORMDL3*	Unknown	Moffatt et al.[37]
Type 2 diabetes	*CDKAL1*	Cell-cycle regulator	Scott et al.[3]
Crohn's disease	*ATG16L1*	Autophagy	Rioux et al.[38]

Table 2. Examples of Loci Shared by Conditions or Traits Previously Thought to Be Unrelated, According to Genomewide Association Studies.

Gene	Conditions Sharing Associations	Source of Data
CDKN2A, CDKN2B	Coronary disease	Helgadottir et al.[36]
	Type 2 diabetes	Scott et al.[3]
	Invasive melanoma	Kamb et al.[43]
ORMDL3	Childhood asthma	Moffatt et al.[37]
	Crohn's disease	Barrett et al.[27]
CDKAL1	Type 2 diabetes	Scott et al.[3]
	Prostate cancer	Steinthorsdottir et al.[44]
LRRK2	Parkinson's disease	Paisán-Ruíz et al.[45]
	Crohn's disease	Barrett et al.[27]
KITLG	Testicular carcinoma	Rapley et al.[46]
	Blond or brown hair	Sulem et al.[47]
C10orf67	Sarcoidosis	Franke et al.[48]
	Celiac disease	Franke et al.[48]
JAZF1	Height	Johansson et al.[49]
	Type 2 diabetes	Zeggini et al.[50]
	Prostate cancer	Thomas et al.[17]

Manolio TA. N Engl J Med 2010;363:166-176.
Visscher PM et al. AJHG 2012. 90: 7-24.

Characteristics of GWAS Readily Translatable to Clinical Care

Application	Key characteristics	Example
Risk prediction	Heritability is high	T1DM loci
	Large proportion of heritability is explained	
	Genotyping can be targeted to high-risk group (such as that defined by positive family history)	
	Genotyping scores substantially increase predictive value	
	Early detection is important	
	Preventive strategies are available	
Disease subtyping or classification	Clinical syndrome has multiple subtypes with varying prognosis or response to treatment aetiologies	*CRP* (or *HNF1A* typing) for MODY
	Subtypes are not readily discernible by clinical examination	
	Compared to other indicators, assay of biomarker (genetic or non-genetic) is easier, cheaper, more available, more reliable and more sensitive	
	Subtyping affects treatment selection	
	Subtyping identifies increased risk in family members	
Drug development	Loss of function variants have a useful biological effect	*ITPA* variants and ribavirin toxicity
Drug toxicity	Variants are common	*SLCO1B1* variants and statins
	Effect sizes are large	
	Alternative drugs or increased monitoring are available	

CRP, C-reactive protein; GWAS, genome-wide association study; *HNF1A*, HNF1 homeobox A; *ITPA*, inosine triphosphatase; MODY, maturity-onset diabetes of the young; *SLCO1B1*, solute carrier organic anion transporter family member 1B1; T1DM, type 1 diabetes mellitus.

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

Risk Prediction and Disease Classification

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

Prostate Cancer Example

Using GWAS findings to Inform Screening Strategies

- Men 45-79
- Age screening
  - 63% of men eligible for screening
  - 96% of cases in screened population
- Polygenic screening
  - 53% of men eligible for screening
  - 93% of cases in screened population

Pashayan & Pharoah. Per Med 2012; 9: 451-455.

Drug Development and Toxicity

- Pharmacogenetic variants may be more amenable to GWAS discovery.
  - Influence of common variants
  - Larger effect sizes
- Examples:
  - Use of antiviral agent ribavirin in treatment of hepatitis C.
    - GWAS of change in hemoglobin levels during ribavirin treatment of hepatitis C lead to identified association with *ITPA* variants.
    - May lead to biomarkers and novel therapeutics to avoid

Marollo NatRev Genet. 2013; 14: 549-58.

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 reveled 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.

http://medical-dictionary.thefreedictionary.com/phenylketonuria; Scriver CR (2007) Hum Mutat

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*63)/(2*84)
9.4 (2.1-41.1)

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

Lancet 1994;344:1453

Alternative way of looking at ORs

Strata	Cases	Controls	OR (95% CI)
G+E+	25	2	34.7 (7.8, 310.0)
G+E-	10	4	6.9 (1.8, 31.8)
G-E+	84	63	3.7 (1.2, 6.3)
G-E-	36	100	Reference
Total	155	169	

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	$RR_{GE}=RR_G x RR_E$	$OR_{GE}=OR_G x OR_E$
Synergistic Interaction	$RR_{GE}>RR_G x RR_E$	$OR_{GE}>OR_G x OR_E$
Antagonistic Interaction	$RR_{GE}<RR_G x RR_E$	$OR_{GE}<OR_G x OR_E$
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 x 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

Figure 9.2 Three qualitative patterns of gene–environment interaction. The y-axis represents a trait value (e.g., mean height, disease prevalence, expected survival); the x-axis represents two environmental conditions; the black bars denote noncarriers and the gray bars represent carriers. (a) gives an example of a removable interaction. (b) and (c) are nonremovable interactions ("pure" and "crossover" interactions, respectively).

Kraft and Hunter (2010)

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
  - 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

Relative Risk and Attributable Risk

- For a **fixed attributable risk**, the relative risk will be largest in the group with the lower background rate.
- In this case there is no statistical interaction on an additive scale, but statistical interaction on a multiplicative scale.

AR=2
RR=2

AR=2
RR=1.67

Relative Risk and Attributable Risk

- For a **fixed relative risk**, the attributable risk will be largest in the group with the higher background rate.
- In this case there is no statistical interaction on an multiplicative scale, but statistical interaction on an additive scale.

AR=2
RR=2

AR=3
RR=2

Aside: Interaction in a Regression Setting

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

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 $e^{\wedge}\beta_{ge}=1$)

IN CLASS EXERCISE

Power for GxE

- Rule of thumb is that tests of interactions need sample sizes 4 times larger than tests of main effects.
- Most modern studies of GxE also have to address multiple testing/multiple comparison issues.
- Sample sizes needed to reliably detect gene-environment interaction are typically quite large
- Replication is important, but challenging for GxE
- Statistical methods have been developed to improve precision and power for GxE testing.

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

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

OR(GxE) = OR(G-E|D=1)/OR(G-E|D=0).

Assuming 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.
  - Emperical 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

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	• 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	• Tests for qualitative interactions. • Useful in characterization motivated by understanding nature of interaction.	[Gail and Simon, 1985]
Goodness of fit tests	• 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	• Standard method for analysis. • Robust to assumptions about G-E correlation.	[Breslow and Day, 1980]
Case only	• 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.	[Piegorsch et al., 1994]
Maximum likelihood estimation method	• 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.

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 Murcray et al 2011.

Murcray CE, et al. Am J Epidemiol 2009; 169: 219-226.
Murcray 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	Module B: Multiple Comparisons	Module C: Testing
• No Screening • Marginal (G-D association) • Correlation (G-E) • Hybrid approaches	• Bonferroni testing • Permutations • Weighted hypothesis testing	• Case-control • Case-only • Empirical Bayes • Bayesian Model Averaging

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

Comparing Methods: Type 1 Error

Simulation results comparing methods:
CC=case-control; CO=case only; EB=Empirical bayes
Bonf=no screening; Corr=G-E screen; Marg=G-D screen.
Top=threshold; Wt=weighted hypothesis testing

QQ plots for genome-wide tests for G-E for BMI (E) and Diabetes (D) in the Nurses Health Study.

Hsu et al. Genetic Epidemiology 2012; in press.

Cornelis M C et al. Am. J. Epidemiol. 2012;175:191-202

Power Considerations

- 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

Emperical power for two-step methods for diffferent 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 power calculations

Software	Good for	URL
Quanto	Joint test, GE test, family-based designs, case-control, continuous outcome	http://hydra.usc.edu/gxe/
Power	Additive interaction	http://dceg.cancer.gov/tools/design/power
Gemis	Misclassification in E	http://www.hsph.harvard.edu/peter-kraft/software/

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	CHARACTERIZATION
• Identify novel loci	• Describe interaction
• Focus on variants that would not be found in marginal search alone	• Focus on putative and established variants
• Priority given to power	• Priority given to descriptive model
• Hypothesis generating	• 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

Slide by Muin Khoury

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 cases[19].

Non-response (or respondent) bias. Differential rates of refusal or non-response to inquiries between cases and disease-free comparison subjects[19].

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 process[19].

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 subjects[19].

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 surveillance[20].

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 midst[19].

Exposure suspicion bias. Knowledge of a patient's disease status can influence the intensity and outcome of the search for exposure to a putative cause[19].

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

Sources of Bias in G and GxE

Method	Key Considerations
Selection Bias	• 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	• Errors in questionnaire, specimen handling • Highlights importance of lab QC • Can impact type I and type II error for GxE
Confounding	• 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 <u>difficult</u> 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

"Your weight problem is partly genetic and partly Boston Cream pie."

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

Two "reaction norms" (i.e. gene-environment interaction patterns) after Lewontin (1971) Am J Hum Genet

Genetic effect, environmental effect, and gene-environment interaction all depend on what part of the E distribution you've sampled: this has implications for discovery and replication

Kraft and Hunter. Human Genome Epidemiology, 2nd Edition. 2009.
Modified from Slide by Peter Kraft.

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

Minimum number of cases (case:control ratio = 1) required to detect a 2-fold interaction with 80% power as a function of the true prevalence of the environmental factor and the sensitivity/specificity

Garcia-Closas M et al. CEBP. 1999;8:1043-1050

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 of “G”

- Imputation can allow for information on a shared set of variants, even if studies used different genotyping platforms.
- Strand alignment errors and allele flipping
- The “minor” allele in one sample may not be the same in other samples, particularly:
  - when the MAF is near 0.5; or
  - when the samples differ in genetic ancestry

Li, Willer, Sanna, Abecasis. Annu Rev Genomics Hum Genet. 2009;10:387-406

Environmental Data Harmonization

- **Goal: To create a standardized, comprehensive database to be used for analysis involving environmental and epidemiologic variables**
  - Common data elements
  - ‘Inferentially consistent’
  - Standardized variable definitions and permissible values
  - Consistent coding
- **Process: Iterative process**
  - Data planning
  - Data acquisition
  - Data extraction, transform and load (ETL)
  - Data QC and delivery

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=5x10-8.

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 multifactoral and dynamic nature of exposure/risk can complicate the study of environmental factors

Rappaport SM and Smith MT. Science 2010; 330: 460-461.

Advances in Measuring E

- PQA4: As modern measurement technologies improve, are there better ways to objectively ascertain exposure to cancer risk?

http://provocativequestions.nci.nih.gov/

Hum Genet
DOI 10.1007/s00439-012-1258-z

ORIGINAL INVESTIGATION

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. Spitz[1], Neil E. Caporaso[2], and Thomas A. Sellers[3]

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.

Gene-enviromentalomics #badomics

FEATURE NEWS

'Omes that matter

See Nature 494, 416–419 (2013)

CROSSWORD

ACROSS
2. The 'ome joining the rest together.
3. All the data generated by clinical trials.
5. All the protein kinases encoded in a genome.
7. A map of the brain's neurons and neural connections.
9. All of the metabolites in an organism.
13. The interaction network for approved drugs and targets.
16. The genome's influence on a patient's response to drugs.
18. The part of the genome that codes for proteins.
19. All the regulatory components in a cell.
21. The structures and functions of all membranes in an organism.
22. The phenotypic variability of an organism.
23. Genetic material recovered from uncultured microorganisms in environmental samples.
24. A database of basic nuclear proteins.
25. The totality of human speech components.

DOWN
1. All life is there.
2. The entirety of molecular interactions in a cell.
3. All the transcribed RNA of a cell or tissue.
4. Isaac Asimov's phrase for chance findings in the genome.
6. All the nascent polypeptide chains in a cell.
8. The part of the genome involved in regulating nutrient metabolism.
10. All the nucleic acids, lipids, proteins, sugars and other molecules in a system.
11. All the proteins expressed at a certain time by a certain genome.
12. All the cellular processes involved in toxicity.
14. The metals and metal-containing molecules in a cell or a tissue.
15. All the processes relating to force and mechanical systems in a cell.
17. A measure of everything an individual is exposed to in a lifetime.
20. Where the trouble began (*Nature passim*).

*Answers available at go.nature.com/mdar*a9

28 FEBRUARY 2013 | VOL 494 | NATURE | 1

THE GOOD, THE BAD AND THE UGLY

Does your 'ome meet the criteria?

GOODOME
Encapsulates a new focus (Interactome: all interactions between biomolecules)

Refers to a comprehensive collection (Transcriptome: everything transcribed from DNA to RNA)

Easy to say (Phenome: comprehensive physical characteristics of an organism)

Easy to understand (Lipidome: all an organism's fatty molecules)

BADOME
Renames existing field (Nutriome: study of nutrients)

Limited in scope (Museome: sequenced DNA from objects in museum archives)

Unpronounceable (tRNome: collection of transfer RNAs)

Obscure (Predatasome: genes used by predatory proteobacteria while invading other bacteria)

Baker M. Nature. 2013. 494, 416–419.

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 Session 2:

- 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	

GENE-ENVIRONMENT INTERACTIONS

What is old is new

Rare Variants and Burden tests

Karen L. Edwards, Ph.D.

Professor

Department of Epidemiology and

Genetic Epidemiology Research Institute

School of Medicine

University of California, Irvine
Irvine, CA

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

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, are the major contributors to genetic susceptibility to complex disease

Rare Variants

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

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

Filtering

Using Annotation

Sequencing: Who, what and why?

- Sequencing is still “expensive”
  - Families
  - Extreme affected
- 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

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

http://www.nature.com/nrg/journal/v11/n6/full/nrg2779.html

Sequencing Depth

- Most current sequencing platforms generate millions of short sequence reads
  - Most are High-depth 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

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

http://www.nature.com/nrg/journal/v11/n11/full/nrg2867.html

Rare Variant Association Analysis

- Statistical considerations for analyzing rare variants are important
- 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 gene (locus) or other functional genomic region
    - Numerous locus-specific statistical approaches have been developed

Statistical approaches for analysis of rare variants

- Fall into 2 broad categories:
  - Collapsing Methods
  - Methods based on similarities among individual sequences
- Collapsing methods
  - Aggregate information across multiple variants within a region of interest into a single quantity
  - 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 can be deleterious or protective.
- Methods based on similarities among individual sequences
  - Multivariate tests that combine single-variant test statistics
  - Make no assumption about direction of effect of each rare variant – more flexible

Brief Aggregate Testing Overview

- "Burden Test"
  - mean burden between cases and controls
  - Deleterious SNPs & most causal variants
- SKAT
  - variance of burden between cases and controls
  - Protective / deleterious SNPs & many non-causal
- MiST
  - Utilizes mean and variance
  - Allows for unknown underlying genetic model

- Aggregate rare variants
  - **Gene**
  - Pathway
- How can we aggregate outside genes?

Sun, Zheng, Hsu. A unified mixed-effects model for rare-variant association in sequencing studies (submitted).

Are Diseases Influenced by Rare Variants Familial

- Family studies are making a come back
  - Variants that are rare in the population will be "enriched" in families where the variant is causal
  - Analytic approach varies
    - Followwp on previous linkage regions
    - Combine linkage and association testing

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

Multivariate Linkage Analyses

Project Overview

Aim 2: Whole Exome Sequencing and Filtering

Framework for selecting variants for custom genotyping

USING FUNCTIONAL ANNOTATION TO INFORM ASSOCIATION ANALYSES

Fred Hutchinson Cancer Research Center
Stephanie Rosse, Ph.D.
srosse@fhcrc.org

Outline

- Define ‘functional’ variation
- Describe available genome-wide functional annotations
- Walk through example of a Colorectal Cancer risk locus
- Live example of the UCSC Genome Browser
- Example annotation of rs6983267

Mendelian Traits

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

Mendelian Traits

Complex Traits

Human Genetic Mutation Database (HGMD)

Missense/Nonsense
Splicing
Regulatory

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

Genome-Wide Association Study (GWAS) Catalog

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

Sequence Variation

Single nucleotide polymorphism (SNPs)

- Average 1 in every 300 nucleotides (~10M SNPs)
- 1.2M coding (most observed only in a single individual)
- The rest is non-coding RNA, regulatory DNA, Repeat elements, or sequence without known function
- Strong negative selection against deleterious variation

HOW DO WE DEFINE FUNCTION?

Coding versus Non-coding

Coding Function

Coding Function

- Missense and Nonsense are likely to disrupt the structure of the encoded protein
- Functional Annotations: Polyphen2, SIFT, GERP to predict deleteriousness of an amino acid change
- **Deleterious mutations are 'functional'**

Deleterious Missense/Nonsense

Clinical significance: Likely Pathogenic

Condition(s):

Name: Maturity-onset diabetes of the young, type 2 (MODY2)

Non-coding Function

Deleterious missense and nonsense

GLUCOKINASE

Transcription Factor

Non-coding Function

Deleterious missense and nonsense

GLUCOKINASE

Transcription Factor

- Regulatory polymorphism can disrupt gene expression
- <u>Function Annotations:</u> motif analysis uses conservation to predict functional variation in binding site. Not highly sensitive or specific.
- Prediction??
  - Quantitative vs. Qualitative

Qualitative Effect

Quantitative Effect

GLUCOKINASE

'…GCK promoter mutation…modulates GCK gene expression through loss of allele-specific Sp1 binding…' Gasperikova et al, *Diabetes*, 2009

Qualitative Effect

Quantitative Effect

Definition of 'function' in non-coding:

1. Allelic differences in gene expression using a model system (e.g. human cell line)
   - ❑ Literature review – publication bias
   - ❑ <1000 examples
2. In a predicted regulatory element using epigenetics
   - ❑ Equivalent to saying 'in a gene'

GLUCOKINASE

'…GCK promoter mutation…modulates GCK gene expression through loss of allele-specific Sp1 binding…' Gasperikova et al, *Diabetes*, 2009

Mendelian Traits

Complex Traits

Human Genetic Mutation Database (HGMD)

Missense/Nonsense
Splicing
Regulatory

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

Genome-Wide Association Study (GWAS) Catalog

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

Outline

- Define ‘functional’ variation
- Describe available genome-wide functional annotations
- Walk through example of a Colorectal Cancer risk locus
- Live example of the UCSC Genome Browser
- Example annotation of rs6983267

Functional Attributes of Regulatory Regions

Image modified from Dr. Stamatoyannopoulos

Functional Attributes of Regulatory Regions

Image modified from Dr. Stamatoyannopoulos

Functional Attributes of Regulatory Regions

Image modified from Dr. Stamatoyannopoulos

Functional Attributes of Regulatory Regions

Functional Attributes of Regulatory Regions

Genome-Wide Functional Annotation

Mostly cell lines

Tissue samples

ANNOTATION DATABASE

ChIP-Seq Method

DNase-seq

Uniformly Processed Roadmap & ENCODE

TISSUES/CELL TYPES

Epigenetic Profile	ChIP-seq Histone	DNase-seq
Embryonic Stem Cell (ESC), other stem cells (eg induced & hematopoietic)	30	8
Blood and immune	23	7
Lung fibroblasts (fetal)	1	1
Muscle satellite	1	1
Epithelial	8	4
Fat Adipose nuclei	1	0
Thymus, Brain, Neurosphere	14	3
Muscle, Heart, Sm. Muscle	14	4
Digestive	12	4
Other (kidney, lung, placenta amnion, liver)	11	6

Regulatory Segmentation

ChromHMM (Ernst et al. 2011)

1. Labels each assay (ChIPseq, DNase) as high to low in **200bp bins** (genome-wide)

2. Runs a 25-state Hidden Markov Model

3. Assigns interpretable labels to each hidden state (eg, CTCF, weak enhancer, enhancer, transcribed region, enhancer)

Segway (Hoffman et al. 2011)

1. Uses a Dynamic Bayesian Network approach using **base-pair** resolution

2. Trained over the ENCODE pilot regions (1% of the genome), and fitted over the whole genome

Emerging annotations

CADD: Genome-wide Measurement of Deleteriousness

Combined Annotation Dependent Depletion

(**CADD**)

High Resolution Regulatory Elements

DHS + ChIP-seq

Comparison of Biochemical Assays

Annotation	Whole Genome?	Disease Specific?	Nucleotide Resolution
GERP	Yes	No	Yes
PhastCons	Yes	No	Few bps
ChromHmm	Yes	Yes	Segment (~1.2kb)
Segway	Yes	Yes	Segment (~200)
Histone	Yes	Yes	Peaks (~1-2kb)
TFBS	Yes	Yes	Peaks(<100)
Open Chromatin	Yes	Yes	Peaks (~200)
PolyPhen2	Non-synonymous	No	Yes
SIFT	Non-synonymous	No	Yes
CADD	Yes	Not yet	Yes

Outline

- Define ‘functional’ variation
- Describe available genome-wide functional annotations
- Walk through example of a Colorectal Cancer risk locus
- Live example of the UCSC Genome Browser
- Example annotation of rs6983267

FUNCTIONAL ANNOTATION COLORECTAL CANCER SUSCEPTIBILITY LOCI

GWAS Catalog:

SNP	Date	Author	N	Region	Gene	P value	OR [95% CI]
rs4775984	02/13	Peters	12,696	15q13.3	GREM1	5e-7	1.12 [1.08-1.19]

“Fine-mapping of common genetic variants associated with colorectal tumor risk identified potential functional variants” -Du et al.

Turning on Tracks

Non-coding Annotation Tools and Framework

UCSC Genome Browser

OMIM: Jaeger, E. et al. **Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1**. Nature Genet. 44: 699-703, 2012

Annotation Tools

UCSC Genome Browser

Position

Variants in LD (r2>0.5)

Gene Annot.

Histone Modification in Colon Mucosa

Annotation Tools

Functional Class	Subclass	Evidence	Tools	Annotation
Regulatory	Promoter Enhancer Silencer Repressor Etc.	Position	UCSC genes, OMIM	Near Candidate Promoter
		Histone Mod.	ChIPseq	
		Open Chromatin	DHS, FAIRE	
		Protein Bound	ChIPseq TF, DHS footprint	

UCSC Genome Browser

Position

Variants in LD (r2>0.5)

Gene Annot.

Histone Mod.

DHS

TFBS

Testable Hypothesis

Testable Hypothesis

Testable Hypothesis

*Kodach LL, et al.(2008). "The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers". *Gastroenterology* **134** (5): 1332–41.

Annotation Tools

Annotation Tools

Annotation Tools

UCSC Genome Browser

Annotation Tools

Annotation Tools

UCSC Genome Browser

Annotation Tools

Functional Class	Subclass	Evidence	Tools	Annotation
Regulatory	Promoter Enhancer Silencer Repressor Etc.	Position	UCSC genes, OMIM	Near Candidate
		Histone Mod.	ChIPseq	Promoter
		Open Chromatin	DHS, FAIRE	Accessible DNA
		Protein Bound	ChIPseq TF, DHS footprint	Suz12 TFBS
		Altered binding motif	ConSite, JASPAR, Haploreg	Altered Binding Motifs
		Conservation	PhastCons	Weak
Post-transcription		micro RNA	TargetScanS 5.2	NA
Splice Site		Donor/ acceptor site motif	BDGP, ESEfinder, RescueESE, ESRSearch, PESX, Ensembl	NA

Outline

- Define ‘functional’ variation
- Describe available genome-wide functional annotations
- Walk through example of a Colorectal Cancer risk locus
- Live example of the UCSC Genome Browser
- Example annotation of rs6983267

This demo will show limitations

Outline

- Define ‘functional’ variation
- Describe available genome-wide functional annotations
- Walk through example of a Colorectal Cancer risk locus
- Live example of the UCSC Genome Browser
- Example annotation of rs6983267

The distal enhancer on 8q24

Supplement

List of Annotation Tools

Annotation Database	Website
UCSC Genome Browser	http://genome.ucsc.edu
Haploreg	http://www.broadinstitute.org/mammals/haploreg
Annovar	http://www.openbioinformatics.org/annovar/
Annotations	**Website**
Transcriptional	
ENCODE ChIPseq Histone Mod.	http://genome.ucsc.edu/cgi-bin/hgHubConnect
Roadmap epigenomics ChIPseq	http://vizhub.wustl.edu/VizHub/RoadmapReleaseAll.txt
Transcription Factors	
ENCODE TFBS ChIPseq	http://genome.ucsc.edu/cgi-bin/hgHubConnect
JASPAR	http://jaspar.genereg.net/
CONSITE	http://consite.genereg.net/
Haploreg	http://www.broadinstitute.org/mammals/haploreg
Transfac	http://www.gene-regulation.com/pub/databases.html
Splice Site	
BDGP	http://www.fruitfly.org/
Functional Segmentation	
ChromHMM	http://compbio.mit.edu/ChromHMM/
Segway	http://noble.gs.washington.edu/proj/segway/
Post-Transcriptional	
miRNA TargetScanS 5.2	http://genome.ucsc.edu/

MODULE WRAP UP

SISG Module 16

Carolyn Hutter

Karen Edwards

Epidemiology 101

- Epidemiology is the study of the distribution and determinants of health-related states in populations
- Study design is a key component of epidemiology
- Relative risks, risk differences and odds ratios are used to measure association
- It is important to consider and address bias in epi studies
- Understanding confounding and effect modification are important in studies of association

	Disease		
Exposure	+	-	Total
+	a	b	a+b
-	c	d	c+d

Overview of Genetic Epidemiology

- Different study designs in genetic epidemiology answer different questions about genetic basis of disease.
- Collecting family data and family history data is important but challenging.
- Pedigrees provide visual representations of family data.

Segregation and Linkage Analysis

- Complex segregation analysis is a modeling approach for evaluating the transmission of a trait within a pedigree.
- Linkage analysis focuses on the co-segregation between a marker locus and a putative disease locus using family data.

Linkage Disequilibrium

- LD refers to nonindependence of alleles
- Measures of LD include D, D' and r^2
- LD is exploited for association studies, imputation and GWAS
- LD in turn causes challenges in interpreting association findings and identifying the causal allele(s)

Association Studies

- Focus on association between a particular allele and disease in a population.
- Need to consider potential impact of population stratification.
- Family based association studies.
- Need to consider power and sample size in design and analysis of association studies.

Population based Association studies

Genome Wide Association Studies

- GWAS perform association analysis on markers across the genome.
- Standards for QC, analysis and presentation.
- GWAS have been highly successful at identifying common variants with modest effect sizes associated with disease.
- Post-GWAS studies focus on functional characterization and epidemiologic interpretation.
- Increasing examples of applications of GWAS findings.

Gene-Environment Interactions

- Focus on interplay between genetic and environmental factors
- Need to consider scale (additive and multiplicative)
- Novel methods to improve power
- Attention needs to be given to measuring and analyzing both the G and the E

Sequencing Studies

- Next-generation sequencing allows us to examine rare and low-frequency variants.
- Gene-level tests are sometimes needed to evaluate this data.
- Family based study designs can be useful for identifying rare variants associated with disease phenotypes.

Functional Studies

- We are often interested in potential function of variants associated with disease.
- Existing genome-wide functional annotations from ENCODE, epigenomics Roadmap and other sources can be used to facilitate interpreation.

SISG Module 16
Optional Exercise: Measures of Association

Table 1: Association between binge drinking and atherosclerosis

	Disease Status	
Exposure Status	Case	Control
Exposed (+)	106	585
Unexposed (-)	186	2165

1. What is the incidence of outcome in the exposed group? In the unexposed group?
2. What is the RR for the association between exposure and outcome? The AR? The OR?

Table 2: Association between, stratified by gender

Men	**Disease Status**	
Exposure Status	Case	Control
Exposed (+)	89	374
Unexposed (-)	118	801

Women	**Disease Status**	
Exposure Status	Case	Control
Exposed (+)	17	211
Unexposed (-)	68	1364

3. What is the RR in men? What is the RR in women?
4. Is there evidence for confounding or effect modification?

SISG Module 16
Optional Exercise: Measures of Association
ANSWER KEY

Table 1: Association between binge drinking and atherosclerosis

	Disease Status	
Exposure Status	Case	Control
Exposed (+)	106	585
Unexposed (-)	186	2165

1. What is the incidence of outcome in the exposed group? In the unexposed group?

ANSWER: Incidence in exposed = 106/(106+585)=15.3%
Incidence in unexposed=186/(186+2165)=7.9%

2. What is the RR for the association between exposure and outcome? The AR? The OR?

ANSWER: RR = 1.9 (1.6 – 2.4); AR=7.4%; OR = 2.1 (1.6 – 2.7)
RR = 1.9 (1.6 – 2.4)

Table 2: Association between, stratified by gender

Men	**Disease Status**	
Exposure Status	Case	Control
Exposed (+)	89	374
Unexposed (-)	118	801

Women	**Disease Status**	
Exposure Status	Case	Control
Exposed (+)	17	211
Unexposed (-)	68	1364

3. What is the RR in men? What is the RR in women?

ANSER: RR in men= 1.50 (1.16-1.93); RR in women= 1.57 (0.94-2.62)

4. Is there evidence for confounding or effect modification?

ANSER: The RR is similar between men and women, so there is not strong evidence for effect modification. The RR in the stratified analysis is lower (closer to the null of 0) than in the combined analysis, indicating evidence of confounding.

How to build a pedigree using the Online Pedigree Drawing tool by Progeny

I. Opening the Online Pedigree Drawing tool by Progeny

1) Open a browser and type into the address field:
www.progenygenetics.com/online-pedigree
(Note: You should make sure that pop-ups are enabled by your browser)

2) Scroll to the bottom and click on the button "I Accept - Begin Pedigree".

3) On the toll bar at the top of the page, click on the button labeled "New".

4) When prompted: "Confirm Create New Family" click "Yes".

II. Drawing a pedigree

1) **Add your first family member (proband).** You should now be looking at a blank page. The white space is called the canvas. Start by adding your proband to the canvas. This can be done in several ways:

i) Right click on the canvas, select "Add Individual" and then choose either "Add Male" or "Add Female".

ii) On the tool bar at the top of the page, click on the button labeled "Palette". A small window should appear titled "Add". This is your palette; it allows you to add individuals to the canvas. Click on either the "Male" or "Female" button and then click on the canvas.

2) **Add descriptions** about your proband. Double click on the icon that you just added and window should appear titled "Add/Edit Subtext…". This box allows you to add descriptions of individuals that will appear below their icon in the pedigree. Click on the button labeled ""Add Row" and in the field titled "Subtext Line 1" write whatever information you wish to record (can either use this to indicate disease status, Yr of birth or death, etc). Click the button labeled "Save" and then exit the "Add/Edit Subtext…" window. Remember to consider confidentiality – you should not enter full names on the pedigree.

3) Now that your first family member is represented on the canvas, you can **add additional relatives** to your proband. This can be done in several ways:

i) Click on the icon for the family member you wish to add relatives to. The icon will now become highlighted and will be shown with four plus signs around it. Left clicking on the blue plus sign that points up will add a father to the selected individual, clicking on the yellow plus sign that points up will add a mother. Left clicking on the plus sign to the right that is blue will add a brother, clicking on the tallow will add a sister, and clicking on the green will add a spouse. Left clicking on the blue plus sign that points down will add a son, clicking on the yellow will add a daughter.

ii) Relationships can also be added to the canvas by right clicking on an individual, selecting “Add Relation” and choosing the relation that you would like to add.

iii) If two individuals are already represented on the canvas, they can be attached. This is done by selecting one of the individuals by clicking on their icon. When selected, an individual has up to five black boxes around it called handles. By clicking on a handle you can drag a line from one individual to another. Each handle connects individuals by a different relationship:

- Left Side or Right Side Handle – assign as a spouse only.
- Top Middle Handle – assign as a child only.
- Top Right Corner or Top Left Corner Handle – assign as a sibling only or twin. (Holding the control key on your keyboard while adding a sibling will connect the two individuals as twins)
- Bottom Handle – assign as a parent only.

Feel free to add as many relationships as you wish, and other demographic or other descriptions for each subject as desired.

4) To indicate that an individual is deceased, right click on that individual, choose “Change Status” and select “Deceased”.

III. Formatting

1) Individuals can be moved around the canvas by clicking on their icon and dragging them. When an individual is moved, its relationships are preserved.

2) To move multiple individuals at once, lasso the icons by clicking on the white canvas and drag a box around the desired icons. This will highlight the individuals, allowing you to click on one of the highlighted individuals and move all of them as a group.

3) You can also select or search for specified relatives of an Individual by right clicking on an Individual icon and choosing the “Select” option. This way you can search and highlight all relatives, ancestors, descendants, parents, siblings, children, spouses and aunts and uncles.

4) When excessive amounts of icon moving have been done in Progeny, the pedigree can begin to look sloppy. If this happens, it is often useful to click on the “Smart Draw” button on the toolbar at the top of the page. This button helps to clean up the pedigree by rearranging all of the icons in neat fashion while maintaining relationships. It often helps to click on “Smart Draw” more than once.

IV. Deleting Individuals and Relationships

1) To delete an individual form the pedigree, simply select that person and click the delete button on your key board. A dialog box will open to ask if you are sure that you want to delete that individual. Multiple individuals can be deleted at once by selecting all the individuals you want to delete and then clicking the delete button on your key board.

2) To delete a relationship or detach a person, highlight the line connecting the individuals by clicking on the line with your left mouse button. After selecting the line, you should see two black dots at the end of each line indicating that the line has been selected. Press the delete key on your keyboard. The line will disappear, breaking the relationship between the two individuals.

V. Pedigree Properties

By right clicking on an individual's icon and choosing properties, you can access the properties window, which allows you to make changes to the setting of that individual and of the entire pedigree. The properties window has five tabs: Progeny, Family, Individual, Relationships, and Dimensions.

1) **The Progeny tab** controls the options that will be set for all pedigrees (global settings). Each option you select will be the default for all the pedigrees you create.

- Show Grid – displays a grid on the canvas
- Snap to Grid – position icons on grid lines only.
- Show Page Breaks – (Print Preview) shows you where the page breaks are for printing.
- Auto Draw – automatically smart draws (executes the SmartDraw algorithm) the pedigree when you add a person.
- Live Dragging of icons – lines move with the icon when you reposition the icons.
- Show Symbols Legend – have symbols legend appear on pedigree or not.
- Show Subtext Legend – have subtext legend appear on pedigree or not.
- One Click Add - enables the yellow arrows used for adding individuals to the pedigree.
- Show Header – have header appear on pedigree or not.
- Show Footer – have footer appear on pedigree or not.
- Show Haplotypes – displays haplotypes with the pedigree

2) **The Individuals tab** controls the properties of an individual.

- Change Gender – Female, Male, and Unknown.
- Attributes – Show someone as Deceased, Proband, No Issue, Infertile, SAB.
- Adopted – Show individual as being Adopted In or Adopted Out
- Marked By – Place an asterisk, minus, or plus next to the individual. (* - +)
- Icon Text - Type in any text you want to display inside the icon.
- Proband Arrow – Change Position and Arrow Type of the proband

- Symbols – Update the icon symbol for the individual

3) **The Relationships tab** controls the properties of the relationship lines connecting each of the individuals to their relatives. To change the nature of any relationship, click on the line first and then access the Properties Window or if the Properties window is already displayed, just click and highlight the required relationship line. This will bring up the Line Tab shown below with the current settings of the selected line shown.

4) **The Dimensions tab** is used to change the dimensions of the pedigree icons and subtext, access the Properties Window and click on the Dimensions Tab. A window appears allowing you to select any of the objects by clicking with the left mouse button. To move an object to a desired position, simply drag it to the spot and release. To alter the size of the object, place the pointer over the edges until the pointer changes to a sliding arrow. Drag the edge to increase or decrease its size.

VI. Copying and Pasting Pedigrees

Pedigrees can be copied and pasted within the Progeny Online Tool or into other Windows applications such as Microsoft Word or PowerPoint for publications. There are two ways of copying pedigree contents:

1) One way is to use the Copy Pedigree option shown by right-clicking on the canvas. This copies the whole pedigree contents to the clipboard and creates a Windows Metafile that can be pasted into any other Windows application.

2) The second method is to copy Individuals within the pedigree from one Progeny Anywhere application to another (or from one Online Pedigree to another). Using this method, you need to lasso the individuals in the pedigree that you intend to copy. The selected individuals will be highlighted in yellow.

To copy the currently highlighted individuals to the Windows clipboard, do one of the following; right-click anywhere on the canvas and choose Copy Individuals option, or press the Ctrl-C keys on your keyboard. When the selected Individuals have been successfully copied to the clipboard, the icons will be highlighted in green.

VII. Saving and Opening Pedigrees

1) To save a pedigree, simply right-click on the canvas and choose Save. This will save the current changes made to the pedigree. To save the pedigree to a different file, choose Save As and then name the new pedigree file whatever you want. The file can be reopened using any Progeny pedigree software.

2) To open a pedigree that was previously saved, choose Open and then select the file of the previously saved pedigree.

1000 Genomes Browser: The 1000 Genomes Browser provides a visual access to the data and information gathered in the 1000 Genomes Project.

1. Go to the main page for the 1000 Genomes (http://www.1000genomes.org/). Click on the "About" tab.
    a. **QUESTION:** What is the stated Goal of this Project?
    b. **QUESTION:** How many samples were sequenced in total? How many of African Ancestry?
2. Go to the 1000 Genomes Browser Webpage (http://browser.1000genomes.org/index.html)
3. Click on "Browse Human"
4. Search for "UCHL1"
5. Click on "Region in Detail" under 3. UCHL1 - ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)
6. Click on "Linkage Data"
7. Click on "Select Populations"
8. Choose the "1000Genomes:phase_1_CEU" and "1000Genomes:phase_1_YRI"
    a. QUESTION: Which population seems to have more LD among variants in this region?
9. For more details, go to the 1000 Genomes Tutorials:
    a. http://www.1000genomes.org/using-1000-genomes-data
    b. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/1000genomes_browser_main_project_20110521/The_1000_Genomes_Browser_Tutorial.ensembl_65.doc

Genome variation Server: The Genome Variation Server (GVS) enables rapid access to human genotype data found in dbSNP, and provides tools for analysis of genotype data. It is particularly useful for examining LD between variants in particular genes or regions, and for selecting tagSNPs.

1. **Go to the website http://gvs.gs.washington.edu/GVS138/**
2. **Search for "UCHL1" under gene name**
3. **Select "HapMap-CEU"**
4. **Click the tab that says "Display Linkage Disequilibrium", and open the graphical display**
5. **Click the tab that says "Display tagSNPs", and open the table view.**
    a. **QUESTION: How many variants are there in total for this gene in the HapMap CEU population?**
    b. **QUESTION: How many tagSNPs would you need to genotype to capture the variation in this gene?**
6. **Click the "Display SNP Summary" tab and see what information is provided.**
7. **Repeat 3-6, but selecting the HapMap-YRI rather than the HapMap-CEU population.**
8. **Repeat the above for a gene of interest.**
9. **Go back to the main website. Click on the side tab that says "BatchGVS138" Batch GVS allows you to perform large automated searches.**
10. **Click on How to use Batch GVS to see details and examples for using this feature in Batch Mode "http://gvs.gs.washington.edu/GVSBatch138/HelpHowToUse.jsp"**

SISG Module 16
Optional Exercise: Linkage Disequilibrium

Table 1: Haplotype and allele frequencies

		Locus 2		
		B	b	
Locus 1	A	0.04	0.50	0.54
	a	0.27	0.19	0.46
		0.31	0.69	1.0

1. The table above provides hypothetical data, showing allele and haplotype frequencies from a population sample. Note, the designation of capital vs. lower case letters are arbitrary. At locus 2 the "b" allele is more frequent than the "B" allele in this population.
    a. Using the notation presented in the slides, state the values for p_A, p_a, p_B, p_b, p_{AB}, p_{Ab}, p_{aB}, and p_{ab} for this population sample.
    b. Describe in words what p_A and p_b represent.
    c. What is the most frequent haplotype in this sample? The least frequent?
    d. Based on this data , calculate D, r^2 and D' between Locus 1 and Locus 2.

2. Assume Locus 1 is the (untyped) disease causing SNP, and Locus 2 is the SNP you are genotyping in your case-control study. You have performed power calculations assuming that you had genotyped the disease causing allele. These power calculations indicate a sample size of 500 cases and 500 controls. Estimate the sample size you will actually need for your study.

ANSWER KEY

Table 1: Haplotype and allele frequencies

		Locus 2		
		B	b	
Locus 1	A	0.04	0.50	0.54
	a	0.27	0.19	0.46
		0.31	0.69	1.0

1. The table above provides hypothetical data, showing allele and haplotype frequencies from a population sample. Note, the designation of capital vs. lower case letters are arbitrary. At locus 2 the "b" allele is more frequent than the "B" allele in this population.
   a. Using the notation presented in the slides, state the values for p_A, p_a, p_B, p_b, p_{AB}, p_{Ab}, p_{aB}, and p_{ab} for this population sample.
   ANSWER: p_A=0.54, p_a=0.46, p_B=0.31, p_b=0.69, p_{AB}=0.04, p_{Ab}=0.50, p_{aB}=0.27, and p_{ab} =0.19
   b. Describe in words what p_A and p_b represent.
   ANSWER: p_A is the frequency of the A allele at locus 1.
   p_b is the frequency of the b allele at locus 2
   c. What is the most frequent haplotype in this sample? The least frequent?
   ANSWER: The most frequent haplotype is the Ab haplotype (50% of the population. The least frequent is the AB haplotype (4% of the population).
   d. Based on this data, calculate D_{AB}, r^2 and D' between Locus 1 and Locus 2.
   ANSWER:
   D_{AB} =$\mathbf{p_{AB} - p_A p_B}$=0.04-(0.54*0.46)= -0.12
   r^2 =$\mathbf{D^2 / (p_A\ p_B\ p_a\ p_b)}$**=(**-0.12^2)/(0.54*0.46*0.31*0.69)=0.27
   D'**=**If D_{AB}<0: $D'_{AB} = D_{AB}/(\min(P_AP_B,P_aP_b))$=-0.12/min(0.54*0.31, 0.46*0.69)
   =-0.12/min(0.16, 0.31)=-0.12/0.16=-0.75

2. Assume Locus 1 is the (untyped) disease causing SNP, and Locus 2 is the SNP you are genotyping in your case-control study. You have performed power calculations assuming that you had genotyped the disease causing allele. These power calculations indicate a sample size of 500 cases and 500 controls. Estimate the sample size you will actually need for your study.

ANSWER: N_2=N_1/r^2

N_2=500/0.27=1851 cases and 1851 controls.

In-Class Exercise: Calculating genetic, environmental, and gene-environment interaction effects

The Fargo-Akron-Kalamazoo Epidemiology (FAKE) study is a prospective cohort study of 80,000 adults from the U.S. Midwest, recruited at primary care visits to several large HMOs. All participants contributed a blood sample and completed a diet questionnaire at enrollment. FAKE investigators hypothesize that subjects carrying a nonsense mutation in *rhubarbulin 1* (*RBLN1-X*) will be more sensitive to the effects of heavy rhubarb consumption (equivalent to more than three slices of rhubarb pie per day), which can lead to chronic diarrhea (CD).[1] The following table shows the distribution of *RBLN1-X*, heavy rhubarb consumption, and incidence of CD over the first 5 years of follow up (assume no censoring).

RBLN1-X	Rhubarb pie	Cases	Controls	TOTAL
Non-Carrier	Unexposed	6,990	50,660	57,650
Carrier	Unexposed	2,468	11,939	14,407
Non-Carrier	Exposed	1,293	5,098	6,391
Carrier	Exposed	461	1,091	1,552
	TOTAL	11,212	68,788	80,000

1. If the investigators had not collected data on *RBLN1-X*, would they have detected an association between rhubarb consumption and CD? I.e. what is the marginal measure of association for CD comparing exposed to unexposed?

2. If the investigators had not collected data on rhubarb consumption, would they have detected an association between *RBLN1-X* and CD? I.e. what is the marginal measure of association for CD comparing carriers to non-carriers?

3. What are the genotype-stratum-specific effects of rhubarb consumption?

4. Would you say there is gene-environment interaction?

*Modified from exercise developed by Peter Kraft.

[1] I doubt that excessive rhubarb consumption can cause CD. But according to wikipedia, rhubarb can be used as a laxative.

In Class Exercise: GxE Harmonization

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 7 studies in this consortia. Brainstorm with your group about the following:

1. What are potential issues/challenges that you might encounter in analyzing this data?
2. What are solutions might you use for some of these challenges?
3. What additional information would be most helpful for you to have?

Table 1: Overview of Subject Characteristics

	Study 1: Cohort		**Study 2: Case-control**		**Study 3: Case-control**		**Study 4: Cohort**		**Study 5: Case-control**		**Study 6: Cohort**		**Study 7: Cohort**	
	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	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%	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	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%	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%	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	FFQ	FFQ	24 hour recall	24 hour recall
Country	USA	USA	USA	USA	China	China	Japan	Japan	Germany	Germany	Canada	Canada	USA	USA
Genotyping Platform	Illumina 550K		Affymetrix 6.0		Illumina 550K		Illumina 1M		Illumina Omni Express		Affymetrix Axiom-CEU		Affymetrix Axiom-CEU	

nature genetics

Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions

Chen Wu[1,2,13], Peter Kraft[2,13], Kan Zhai[1,13], Jiang Chang[1,13], Zhaoming Wang[3,4], Yun Li[5], Zhibin Hu[6], Zhonghu He[7], Weihua Jia[8], Christian C Abnet[3], Liming Liang[2], Nan Hu[3], Xiaoping Miao[9], Yifeng Zhou[10], Zhihua Liu[1], Qimin Zhan[1], Yu Liu[1], Yan Qiao[1], Yuling Zhou[1], Guangfu Jin[6], Chuanhai Guo[7], Changdong Lu[11], Haijun Yang[11], Jianhua Fu[8], Dianke Yu[1], Neal D Freedman[3], Ti Ding[12], Wen Tan[1], Alisa M Goldstein[3], Tangchun Wu[9], Hongbing Shen[6], Yang Ke[7], Yixin Zeng[8], Stephen J Chanock[3,4], Philip R Taylor[3] & Dongxin Lin[1]

We conducted a genome-wide association study (GWAS) and a genome-wide gene-environment interaction analysis of esophageal squamous-cell carcinoma (ESCC) in 2,031 affected individuals (cases) and 2,044 controls with independent validation in 8,092 cases and 8,620 controls. We identified nine new ESCC susceptibility loci, of which seven, at chromosomes 4q23, 16q12.1, 17q21, 22q12, 3q27, 17p13 and 18p11, had a significant marginal effect ($P = 1.78 \times 10^{-39}$ to $P = 2.49 \times 10^{-11}$) and two of which, at 2q22 and 13q33, had a significant association only in the gene–alcohol drinking interaction (gene-environment interaction P ($P_{G \times E}$) $= 4.39 \times 10^{-11}$ and $P_{G \times E} = 4.80 \times 10^{-8}$, respectively). Variants at the 4q23 locus, which includes the *ADH* cluster, each had a significant interaction with alcohol drinking in their association with ESCC risk ($P_{G \times E} = 2.54 \times 10^{-7}$ to $P_{G \times E} = 3.23 \times 10^{-2}$). We confirmed the known association of the *ALDH2* locus on 12q24 to ESCC, and a joint analysis showed that drinkers with both of the *ADH1B* and *ALDH2* risk alleles had a fourfold increased risk for ESCC compared to drinkers without these risk alleles. Our results underscore the direct genetic contribution to ESCC risk, as well as the genetic contribution to ESCC through interaction with alcohol consumption.

© 2012 Nature America, Inc. All rights reserved.

ESCC ranks as the tenth most prevalent cancer in the world, with marked regional variation and a particularly high incidence in certain regions of China. Previous molecular epidemiological studies using a candidate gene approach have implicated a set of genetic variations that confer susceptibility to ESCC, primarily variations that are related to alcohol metabolism[1–6]. The GWAS has emerged as a powerful and successful tool to identify common disease alleles by using high-throughput genotyping technology to interrogate a large number of tagging SNPs that serve as surrogates for untested common SNPs across the genome. In studies published thus far, GWAS of cancers of the upper aerodigestive tract, including ESCC in individuals of European[7,8] and Japanese ancestry[9], have shown that variants in *ADH* genes and/or *ALDH2* are associated with risk of ESCC; in addition, these studies have shown an interaction for these loci with alcohol. Two GWAS showed that variants in *PLCE1* and, perhaps, *C20orf54* are associated with risk of ESCC in Chinese populations[10,11].

We recently reported a multistage GWAS of ESCC that was based on genotyping 666,141 SNPs in 2,031 cases and 2,044 controls with a second replication stage in 6,276 cases and 6,165 controls and identified three new loci that are associated with susceptibility to ESCC[12]. In this previous study, we attempted to replicate 29 SNPs with $P \leq 10^{-7}$. Because of our use of this stringent P value threshold, it is possible that some true ESCC-associated loci with moderate effect sizes were overlooked[13]. However, such loci may be detected by dense genotyping or analyzing larger sample sizes[14]. Furthermore, in our published GWAS report, we observed that three variants at 12q24 conferred ESCC risk through a gene-lifestyle interaction, with a pronounced elevation of risk among alcohol users[12]. Alcohol intake is an important risk factor that contributes to the development of ESCC in Asian and other populations[15]. These findings underscore the fact that ESCC is a complex disease and that its etiology is related to environmental exposures, multiple genetic loci and gene-environment interactions.

[1]State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), US National Institutes of Health, Bethesda, Maryland, USA. [4]Core Genotyping Facility, NCI-Frederick, SAIC-Frederick, Frederick, Maryland, USA. [5]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [6]Department of Epidemiology and Biostatistics, Cancer Center, Nanjing Medical University, Nanjing, Jiangsu, China. [7]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University School of Oncology, Beijing Cancer Hospital and Institute, Beijing, China. [8]State Key Laboratory of Oncology in Southern China, Sun Yat-Sen University Cancer Center, Guangzhou, Guangdong, China. [9]Key Laboratory for Environment and Health (Ministry of Education), School of Public Health, Huazhong University of Sciences and Technology, Wuhan, Hubei, China. [10]Laboratory of Cancer Molecular Genetics, Medical College of Soochow University, Suzhou, Jiangsu, China. [11]Anyang Cancer Hospital, Anyang, Henan, China. [12]Shanxi Cancer Hospital, Taiyuan, Shanxi, China. [13]These authors contributed equally to this work. Correspondence should be addressed to D.L. (lindx72@cicams.ac.cn).

Received 5 March; accepted 16 August; published online 9 September 2012; doi:10.1038/ng.2411

Table 1 Characteristics of cases with ESCC and controls who participated in this study

	GWAS[a]		Replication 1[a]		Replication 2[b]		Combined sample		High-risk cohort[c]	
	Cases (*N* = 2,031)	Controls (*N* = 2,044)	Cases (*N* = 3,571)	Controls (*N* = 3,602)	Cases (*N* = 4,521)	Controls (*N* = 5,018)	Cases (*N* = 10,123)	Controls (*N* = 10,664)	Cases (*N* = 1,410)	Controls (*N* = 1,656)
Age, mean (s.d.)	59.8 (9.8)	61.3 (8.5)	60.5 (8.9)	55.7 (12.7)	60.1 (9.0)	51.5 (13.5)	60.2 (9.1)	54.8 (12.9)	58.1 (8.1)	57.8 (9.2)
Sex										
Male, *N* (%)	1,627 (80.1)	1,706 (83.5)	2,653 (74.3)	2,374 (65.9)	3,380 (74.8)	3,955 (78.8)	7,660 (75.7)	8,035 (75.3)	919 (65.2)	1,222 (73.8)
Female, *N* (%)	404 (19.9)	338 (16.5)	918 (25.7)	1,228 (34.1)	1,141 (25.2)	1,063 (21.2)	2,463 (24.3)	2,629 (24.7)	491 (34.8)	434 (26.2)
Smoking status										
Nonsmoker, *N* (%)	706 (34.8)	895 (43.8)	1,604 (44.9)	2,082 (57.8)	2,115 (46.8)	2,542 (50.7)	4,425 (43.7)	5,519 (51.8)	551 (39.1)	577 (34.8)
Smoker, *N* (%)	1,325 (65.2)	1,149 (56.2)	1,967 (55.1)	1,520 (42.2)	2,406 (53.2)	2,476 (49.3)	5,698 (56.3)	5,145 (48.2)	859 (60.9)	1,079 (65.2)
Drinking status										
Nondrinker, *N* (%)	886 (43.6)	1,139 (55.7)	1,982 (55.5)	2,307 (64.0)	2,115 (46.8)	2,886 (57.5)	4,983 (49.2)	6,332 (59.4)	1,112 (78.9)	1,373 (82.9)
Drinker, *N* (%)	1,145 (56.4)	905 (44.3)	1,589 (44.5)	1,295 (36.0)	2,406 (53.2)	2,132 (42.5)	5,140 (50.8)	4,332 (40.6)	298 (21.1)	283 (17.1)

[a]Cases and controls were recruited from Beijing region. [b]Cases and controls were recruited from Jiangsu, Henan and Guangdong provinces. [c]This case-control set, derived from Shanxi province, where the ESCC incidence and mortality rates are among the highest in China[11], had considerably lower percentage of drinkers in both the case and control categories compared with the percentages in other groups.

Because some ESCC susceptibility loci act in an environment-responsive manner, true associations might not be detected by GWAS without accounting for environmental risk factors[16]. Thus, to discover these susceptibility loci, incorporation of environmental risk factors in the context of GWAS may yield additional regions that are worthy of follow-up studies.

Here we report a new, multistage GWAS of ESCC in a total of 10,123 cases with ESCC and 10,664 controls (**Table 1**). We also report, to our knowledge, the first genome-wide gene-environment interaction analysis of ESCC that incorporates alcohol drinking. We replicated results from these GWAS in an additional case-control panel from a high-risk population.

RESULTS

New loci associated with susceptibility to ESCC

To identify new susceptibility loci for ESCC, we analyzed 169 promising SNPs (with $10^{-7} < P < 10^{-4}$ in our previous GWAS; **Supplementary Table 1**) in replication 1 comprising 3,571 cases and 3,602 controls. We further verified the 18 SNPs with $P < 0.01$ in replication 2 comprising 4,521 cases and 5,018 controls. We found that 15 SNPs were significantly associated with ESCC risk in the replication 2 samples in the same direction as in the genome-wide scan and replication 1 ($P = 2.20 \times 10^{-3}$ to $P = 1.67 \times 10^{-24}$). A joint analysis of the genome-wide scan data together with the samples from replications 1 and 2 showed that these 15 associations reached genome-wide significance (all $P \leq 2.49 \times 10^{-11}$; **Tables 2** and **3**).

Eight of the significant makers were located at chromosome 4q23, which harbors a cluster of seven alcohol dehydrogenase superfamily genes (*ADH* genes). The top marker in this region was rs1042026 (odds ratio (OR) = 1.35, 95% CI 1.29–1.41, $P_{combined} = 1.78 \times 10^{-39}$), and the other seven SNPs in this region are in moderate linkage disequilibrium (LD) with rs1042026 (r^2 = 0.30–0.49), all of which provided significant marginal associations in the combined dataset ($P_{combined} = 1.26 \times 10^{-29}$ to $P_{combined} = 2.75 \times 10^{-20}$) (**Fig. 1a**). After conditioning on rs1042026, the association *P* values for the other seven SNPs increased by over 13 orders of magnitude, suggesting

Table 2 Nine SNPs with significant marginal effects only on ESCC risk in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; risk allele	Genome-wide discovery 2,031 cases, 2,044 controls			Replication 1 3,571 cases, 3,602 controls			Replication 2 4,521 cases, 5,018 controls			Combined sample 10,123 cases, 10,664 controls		
	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs4785204; chr. 16; 48,661,235; *HEATR3*; T	0.25	1.30 (1.17–1.43)	3.05×10^{-7}	0.27	1.23 (1.14–1.33)	9.45×10^{-8}	0.25	1.22 (1.14–1.30)	3.07×10^{-8}	0.26	1.24 (1.18–1.29)	2.24×10^{-20}
rs7206735; chr. 16; 48,706,009; *HEATR3*; C	0.28	1.29 (1.17–1.42)	3.49×10^{-7}	0.29	1.21 (1.12–1.30)	6.97×10^{-7}	0.28	1.18 (1.10–1.26)	1.31×10^{-6}	0.28	1.20 (1.15–1.26)	1.97×10^{-16}
rs6503659; chr. 17; 37,150,790; *HAP1*; A	0.12	1.39 (1.22–1.58)	5.11×10^{-7}	0.15	1.20 (1.09–1.31)	1.00×10^{-4}	0.12	1.27 (1.16–1.39)	2.36×10^{-7}	0.13	1.27 (1.20–1.34)	2.73×10^{-16}
rs2239815; chr. 22; 27,522,670; *XBP1*; T	0.38	1.28 (1.16–1.40)	1.72×10^{-7}	0.39	1.12 (1.04–1.20)	2.10×10^{-3}	0.35	1.17 (1.10–1.25)	9.17×10^{-7}	0.37	1.18 (1.13–1.23)	3.88×10^{-15}
rs2239612; chr. 3; 188,275,936; *ST6GAL1*; T	0.17	1.35 (1.20–1.51)	3.27×10^{-7}	0.20	1.20 (1.11–1.30)	1.15×10^{-5}	0.18	1.17 (1.08–1.26)	1.00×10^{-4}	0.19	1.21 (1.15–1.27)	5.74×10^{-14}
rs17761864; chr. 17; 2,118,387; *SMG6*; A	0.14	1.38 (1.22–1.56)	2.16×10^{-7}	0.15	1.16 (1.06–1.27)	1.60×10^{-3}	0.13	1.16 (1.06–1.27)	9.00×10^{-4}	0.14	1.21 (1.14–1.28)	2.21×10^{-11}
rs2847281; chr. 18; 12,811,593; *PTPN2*; C	0.15	1.33 (1.19–1.50)	1.37×10^{-6}	0.17	1.16 (1.06–1.27)	9.00×10^{-4}	0.14	1.14 (1.05–1.24)	2.20×10^{-3}	0.16	1.20 (1.14–1.26)	2.49×10^{-11}
rs4822983[a]; chr. 22; 27,445,066; *CHEK2*; T	0.19	1.46 (1.31–1.62)	1.02×10^{-8}	0.21	1.22 (1.12–1.32)	1.82×10^{-6}	0.19	1.24 (1.15–1.34)	2.06×10^{-8}	0.20	1.27 (1.21–1.34)	1.94×10^{-22}
rs1033667[a]; chr. 22; 27,460,300; *CHEK2*; T	0.26	1.33 (1.20–1.46)	1.91×10^{-8}	0.27	1.17 (1.09–1.26)	3.72×10^{-5}	0.24	1.26 (1.18–1.36)	3.69×10^{-11}	0.25	1.25 (1.19–1.30)	4.85×10^{-22}

The *P* values shown are two sided and were calculated by the additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates. [a]Discovered by imputation analysis. Chr., chromosome; MAF, minor allele frequency in controls; OR, odds ratio for the minor allele.

© 2012 Nature America, Inc. All rights reserved.

Table 3 Eight SNPs with a significant marginal effect on ESCC risk and the interaction of genes and alcohol drinking in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; substitution	Subgroup	Genome-wide discovery			Replication 1			Replication 2			Combined sample		
		MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs1042026; chr. 4; 100,447,489; *ADH1B*; G>A	Case_control	0.26	1.29 (1.17–1.42)	1.51×10^{-7}	0.23	1.29 (1.19–1.39)	1.33×10^{-10}	0.22	1.44 (1.34–1.54)	1.67×10^{-24}	0.23	1.35 (1.29–1.41)	1.78×10^{-39}
	Nondrinker	0.24	1.11 (0.96–1.28)	1.63×10^{-1}	0.23	1.15 (1.04–1.28)	6.80×10^{-3}	0.21	1.31 (1.18–1.45)	4.91×10^{-7}	0.22	1.20 (1.12–1.27)	4.21×10^{-8}
	Drinker	0.27	1.44 (1.26–1.63)	3.69×10^{-8}	0.24	1.45 (1.29–1.63)	3.05×10^{-10}	0.22	1.58 (1.43–1.75)	1.94×10^{-19}	0.24	1.51 (1.42–1.61)	2.34×10^{-36}
	G × E		1.31 (1.08–1.59)	5.20×10^{-3}		1.28 (1.09–1.49)	2.00×10^{-3}		1.20 (1.04–1.38)	1.12×10^{-2}		1.27 (1.16–1.38)	2.54×10^{-7}
rs3805322; chr. 4; 100,276,021; *ADH4*;A>G	Case_control	0.48	0.79 (0.73–0.86)	1.89×10^{-7}	0.49	0.80 (0.75–0.86)	2.83×10^{-10}	0.48	0.84 (0.79–0.89)	1.25×10^{-8}	0.48	0.81 (0.78–0.85)	2.92×10^{-24}
	Nondrinker	0.49	0.94 (0.82–1.06)	2.93×10^{-1}	0.49	0.85 (0.77–0.93)	4.00×10^{-4}	0.48	0.95 (0.86–1.04)	2.35×10^{-1}	0.49	0.90 (0.85–0.95)	2.00×10^{-4}
	Drinker	0.47	0.68 (0.60–0.77)	1.23×10^{-9}	0.47	0.76 (0.68–0.84)	4.12×10^{-7}	0.47	0.76 (0.69–0.83)	1.47×10^{-9}	0.47	0.74 (0.70–0.78)	4.28×10^{-24}
	G × E		0.72 (0.60–0.85)	2.00×10^{-4}		0.88 (0.76–1.01)	7.48×10^{-2}		0.81 (0.71–0.91)	7.00×10^{-4}		0.81 (0.75–0.88)	5.58×10^{-7}
rs17033; chr. 4; 100,447,968; *ADH1B*; A>G	Case_control	0.11	1.40 (1.23–1.59)	2.80×10^{-7}	0.12	1.36 (1.23–1.50)	1.83×10^{-9}	0.10	1.44 (1.31–1.58)	1.98×10^{-14}	0.11	1.41 (1.33–1.50)	1.26×10^{-29}
	Nondrinker	0.11	1.31 (1.08–1.58)	6.50×10^{-3}	0.11	1.27 (1.11–1.45)	6.00×10^{-4}	0.10	1.35 (1.18–1.55)	1.29×10^{-5}	0.11	1.31 (1.21–1.43)	1.88×10^{-10}
	Drinker	0.12	1.47 (1.24–1.76)	1.69×10^{-5}	0.12	1.48 (1.27–1.72)	4.43×10^{-7}	0.10	1.50 (1.31–1.72)	4.02×10^{-9}	0.11	1.51 (1.38–1.64)	2.10×10^{-20}
	G × E		1.13 (0.87–1.46)	3.70×10^{-1}		1.18 (0.97–1.45)	1.03×10^{-1}		1.11 (0.92–1.33)	2.87×10^{-1}		1.14 (1.01–1.28)	3.23×10^{-2}
rs17028973; chr. 4; 100,541,809; *ADH7*; C>T	Case_control	0.34	1.26 (1.15–1.38)	4.61×10^{-7}	0.33	1.22 (1.14–1.31)	2.48×10^{-8}	0.31	1.29 (1.21–1.38)	7.67×10^{-15}	0.32	1.26 (1.21–1.32)	2.53×10^{-28}
	Nondrinker	0.32	1.16 (1.01–1.33)	3.05×10^{-2}	0.33	1.08 (0.98–1.19)	1.02×10^{-1}	0.31	1.17 (1.07–1.29)	1.10×10^{-3}	0.32	1.14 (1.07–1.21)	1.36×10^{-5}
	Drinker	0.36	1.35 (1.19–1.53)	2.29×10^{-6}	0.32	1.43 (1.28–1.59)	1.60×10^{-10}	0.31	1.42 (1.29–1.55)	2.58×10^{-13}	0.33	1.41 (1.33–1.50)	5.76×10^{-29}
	G × E		1.18 (0.99–1.42)	7.15×10^{-2}		1.34 (1.16–1.55)	7.20×10^{-5}		1.18 (1.04–1.34)	1.15×10^{-2}		1.24 (1.14–1.35)	3.04×10^{-7}
rs1614972; chr. 4; 100,477,178; *ADH1C*; T>C	Case_control	0.26	1.27 (1.15–1.40)	1.34×10^{-6}	0.25	1.26 (1.17–1.36)	1.62×10^{-9}	0.24	1.32 (1.23–1.41)	7.87×10^{-15}	0.25	1.28 (1.23–1.34)	8.02×10^{-28}
	Nondrinker	0.25	1.16 (1.00–1.33)	4.71×10^{-2}	0.25	1.16 (1.04–1.28)	5.20×10^{-3}	0.24	1.17 (1.06–1.29)	2.80×10^{-3}	0.25	1.16 (1.09–1.23)	6.18×10^{-6}
	Drinker	0.27	1.36 (1.19–1.56)	5.64×10^{-6}	0.25	1.40 (1.25–1.58)	1.85×10^{-8}	0.24	1.46 (1.32–1.62)	1.38×10^{-13}	0.25	1.42 (1.33–1.52)	2.39×10^{-26}
	G × E		1.20 (0.99–1.45)	7.06×10^{-2}		1.23 (1.06–1.44)	7.90×10^{-3}		1.25 (1.09–1.44)	1.60×10^{-3}		1.24 (1.13–1.36)	2.54×10^{-6}
rs1229977; chr. 4; 100,421,437; *ADH1A*; *C*>T	Case_control	0.10	1.37 (1.20–1.57)	3.22×10^{-6}	0.11	1.31 (1.19–1.45)	1.52×10^{-7}	0.11	1.33 (1.21–1.46)	4.41×10^{-9}	0.11	1.33 (1.25–1.41)	2.75×10^{-20}
	Nondrinker	0.10	1.19 (0.97–1.46)	9.89×10^{-2}	0.11	1.23 (1.07–1.41)	2.90×10^{-3}	0.10	1.25 (1.09–1.44)	1.80×10^{-3}	0.10	1.22 (1.12–1.33)	4.89×10^{-6}
	Drinker	0.11	1.49 (1.25–1.79)	1.28×10^{-5}	0.11	1.41 (1.21–1.65)	1.28×10^{-5}	0.11	1.43 (1.25–1.63)	2.00×10^{-7}	0.11	1.45 (1.33–1.58)	8.27×10^{-17}
	G × E		1.29 (0.99–1.70)	6.20×10^{-2}		1.17 (0.95–1.43)	1.46×10^{-1}		1.13 (0.93–1.36)	2.14×10^{-1}		1.19 (1.06–1.35)	4.30×10^{-3}
rs1789903; chr. 4; 100,481,064; *ADH1C*; C>G	Case_control	0.09	1.41 (1.22–1.62)	3.35×10^{-6}	0.09	1.31 (1.18–1.47)	1.04×10^{-6}	0.08	1.40 (1.26–1.55)	1.53×10^{-10}	0.09	1.37 (1.28–1.46)	9.46×10^{-21}
	Nondrinker	0.09	1.27 (1.03–1.58)	2.74×10^{-2}	0.10	1.13 (0.98–1.31)	9.21×10^{-2}	0.08	1.34 (1.15–1.56)	2.00×10^{-4}	0.09	1.24 (1.13–1.36)	6.67×10^{-6}
	Drinker	0.09	1.54 (1.26–1.88)	2.02×10^{-5}	0.09	1.55 (1.30–1.84)	6.37×10^{-7}	0.09	1.45 (1.25–1.68)	6.83×10^{-7}	0.09	1.50 (1.36–1.65)	1.16×10^{-16}
	G × E		1.25 (0.93–1.67)	1.34×10^{-1}		1.40 (1.12–1.75)	3.20×10^{-3}		1.08 (0.88–1.33)	4.64×10^{-1}		1.23 (1.08–1.41)	1.90×10^{-3}
rs1893883; chr. 4; 100,343,739; *ADH6*; C>G	Case_control	0.14	1.31 (1.16–1.48)	1.23×10^{-5}	0.14	1.30 (1.18–1.43)	3.61×10^{-8}	0.13	1.39 (1.28–1.52)	3.98×10^{-14}	0.14	1.34 (1.27–1.42)	2.69×10^{-25}
	Nondrinker	0.13	1.21 (1.01–1.45)	4.34×10^{-2}	0.14	1.21 (1.06–1.37)	3.40×10^{-3}	0.13	1.37 (1.21–1.56)	1.18×10^{-6}	0.13	1.26 (1.16–1.36)	8.48×10^{-9}
	Drinker	0.15	1.39 (1.18–1.63)	9.57×10^{-5}	0.15	1.41 (1.22–1.63)	2.04×10^{-6}	0.14	1.45 (1.29–1.64)	1.76×10^{-9}	0.15	1.42 (1.31–1.54)	2.84×10^{-18}
	G × E		1.17 (0.92–1.49)	2.02×10^{-1}		1.19 (0.99–1.44)	7.11×10^{-2}		1.03 (0.87–1.23)	7.09×10^{-1}		1.14 (1.02–1.27)	2.20×10^{-2}

The *P* values shown are two sided and were calculated by an additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates for the subgroups of case_control (individuals included in the case-control study), nondrinker and drinker. The *P* values for the gene × environment interaction were calculated by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression with age, sex, smoking as covariates. MAF, minor allele frequency in the controls; OR, odds ratio for the minor allele; G × E, gene × environment interaction.

© 2012 Nature America, Inc. All rights reserved.

Figure 1 Regional plots of the association results for genotyped and imputed SNPs and the recombination rates within nine significant susceptibility loci. (**a**–**i**) The significant loci are located in chromosomes 4q23 (**a**),16q12.1 (**b**),17q21 (**c**), 22q12 (**d**), 3q27 (**e**), 17p13 (**f**), 18p11 (**g**), 13q33 (**h**) and 2q22 (**i**). For each plot, the $-\log_{10} P$ values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The estimated recombination rates (cM/Mb) from the HapMap Project (NCBI Build 36) are shown as light blue lines, and the genomic locations of genes within the regions of interest in the NCBI Build 36 human assembly were annotated from the UCSC Genome Browser and are shown as arrows. SNPs shown in red, orange, green, light blue and blue have $r^2 \geq 0.8$, $r^2 \geq 0.6$, $r^2 \geq 0.4$, $r^2 \geq 0.2$ and $r^2 < 0.2$ with the tag SNP, respectively. Purple diamonds represent associations of tag SNPs identified in the GWAS stage.

© 2012 Nature America, Inc. All rights reserved.

that the association signals of these other seven SNPs probably point toward the same locus, which is marked by the top SNP (rs1042026) (**Supplementary Table 2**). An imputation analysis in the initial GWAS identified associations for 111 SNPs within a 2-Mb region centered on rs1042026 ($P \leq 10^{-4}$), but none of these SNPs was more significant than the index marker, rs1042026, and a conditional analysis did not provide evidence for a second, independent susceptibility allele in this region (**Supplementary Table 3**).

The markers rs4785204 and rs7206735 at 16q12.1 were also strong signals that were associated with ESSC risk (OR = 1.24, 95% CI 1.18–1.29, $P_{combined} = 2.24 \times 10^{-20}$ for rs4785204 and OR = 1.20, 95% CI 1.15–1.26, $P_{combined} = 1.97 \times 10^{-16}$ for rs7206735; **Fig. 1b**). These two SNPs are located in close proximity to each other and are in moderate LD ($r^2 = 0.41$ in controls); after conditioning on rs4785204, rs7206735 was no longer genome-wide significant (**Supplementary Table 2**). An imputation analysis identified 40 untyped SNPs clustering in two blocks with high LD tagged by these markers that reached a significance of $P \leq 10^{-4}$; again, after conditioning on the index SNP (rs4785204), there was little evidence of a second susceptibility allele in this region (**Supplementary Table 3**).

We found new susceptibility loci for rs6503659 at 17q21 (OR = 1.27, 95% CI 1.20–1.34, $P_{combined} = 2.73 \times 10^{-16}$) and rs2239815 at 22q12 (OR = 1.18, 95% CI 1.13–1.23, $P_{combined} = 3.88 \times 10^{-15}$). Although we detected residual associations at many imputed SNPs in the region tagged by rs6503659, none of them was more significant than the index marker, and conditional analyses did not suggest the presence of a second susceptibility allele in this region (**Fig. 1c** and **Supplementary Table 3**). However, of the 36 imputed SNPs with $P \leq 10^{-4}$ in the region tagged by rs2239815, 8 comprised a separate significant block that was only in weak LD with rs2239815 ($r^2 = 0.21$–0.39) (**Fig. 1d**). We selected the top two imputed SNPs from this block, rs4822983 and rs1033667, both of which had *P* values that were smaller than that of the genotyped index SNP, rs2239815, in the initial GWAS for further replication in all samples. We found that each of these two SNPs was significantly associated with ESCC risk (OR = 1.27, 95% CI 1.21–1.34, $P_{combined} = 1.94 \times 10^{-22}$ for rs4822983 and OR = 1.25, 95% CI 1.19–1.30, $P_{combined} = 4.85 \times 10^{-22}$ for rs1033667; **Table 2**). After conditioning on rs4822983 in the combined sample, evidence for the associations between rs2239815 and rs1033667 and ESCC dropped by over ten orders of magnitude (**Supplementary Table 2**); similarly, after conditioning on rs4822983 in the initial GWAS, there was little evidence of a second susceptibility marker among the imputed SNPs in this region (**Supplementary Table 3**).

The SNP rs2239612 at 3q27 was also associated with ESCC risk (OR = 1.21, 95% CI 1.15–1.27, $P_{combined} = 5.74 \times 10^{-14}$), all seven imputed SNPs in this region were in strong LD with rs2239612 ($r^2 = 0.94$–0.99), and we identified no other significant LD block in this region (**Fig. 1e** and **Supplementary Table 3**). An additional two new identified markers were rs17761864 at 17p13 (OR = 1.21, 95% CI 1.14–1.28, $P_{combined} = 2.21 \times 10^{-11}$) and rs2847281 at 18p11 (OR = 1.20, 95% CI 1.14–1.26, $P_{combined} = 2.49 \times 10^{-11}$). An imputation analysis identified

Table 4 Two SNPs significantly associated with ESCC risk revealed by a SNP × alcohol drinking interaction analysis in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; substitution	Subgroup	Genome-wide discovery			Replication 1			Replication 2			Combined sample		
		MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs9288520; chr. 2; 217,189,516; *IGFB2*; G>A	Nondrinker	0.37	0.69 (0.60–0.79)	6.54×10^{-8}	0.33	0.87 (0.79–0.95)	3.40×10^{-3}	0.34	0.82 (0.74–0.90)	2.60×10^{-5}	0.34	0.81 (0.77–0.86)	4.72×10^{-12}
	Drinker	0.31	1.18 (1.03–1.34)	1.77×10^{-2}	0.30	1.08 (0.97–1.21)	1.58×10^{-1}	0.32	1.06 (0.96–1.17)	2.34×10^{-1}	0.31	1.09 (1.02–1.16)	1.00×10^{-2}
	G × E	0.34	1.71 (1.41–2.06)	2.69×10^{-5}	0.32	1.24 (1.07–1.43)	3.70×10^{-3}	0.33	1.29 (1.13–1.47)	2.00×10^{-4}	0.33	1.33 (1.22–1.45)	4.39×10^{-11}
rs17450420; chr. 13; 103,837,147; *SLC10A2*; A>G	Nondrinker	0.05	0.62 (0.44–0.87)	6.00×10^{-3}	0.05	0.76 (0.61–0.93)	8.20×10^{-3}	0.04	0.89 (0.72–1.09)	2.59×10^{-1}	0.05	0.78 (0.68–0.89)	2.00×10^{-4}
	Drinker	0.03	1.74 (1.27–2.37)	6.00×10^{-4}	0.05	1.27 (0.99–1.62)	5.78×10^{-2}	0.04	1.22 (0.99–1.51)	6.73×10^{-2}	0.04	1.34 (1.16–1.54)	4.65×10^{-5}
	G × E	0.04	2.76 (1.75–4.37)	1.37×10^{-5}	0.05	1.68 (1.22–2.31)	1.50×10^{-3}	0.04	1.42 (1.07–1.90)	1.68×10^{-2}	0.05	1.70 (1.41–2.06)	4.80×10^{-8}

The *P* values shown are two sided and were calculated by an additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates for the subgroups of nondrinker and drinker. *P* values for the gene × environment interaction were calculated by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression with age, sex, smoking as covariates. MAF, minor allele frequency in the controls; OR, odds ratio for the minor allele; chr., chromosome; G × E, gene × environment interaction.

15 associated SNPs with weaker signals than those of rs17761864 and rs2847281 (**Fig. 1f**,**g** and **Supplementary Table 3**).

Risk lock found by genome-wide gene-environment analysis

We performed a genome-wide gene-environment interaction analysis using previous genome-wide–association scan data by testing whether the per-allele odds ratio for each SNP differed between ever drinkers and never drinkers. A quantile-quantile plot of the observed versus expected Wald χ^2 1-degree-of-freedom test for interaction showed no evidence for inflation ($\lambda = 1.004$; **Supplementary Fig. 1a**). There were 25 promising SNPs associated with ESCC risk at significance levels ranging from $P_{G \times E} = 1.42 \times 10^{-23}$ to $P_{G \times E} = 9.88 \times 10^{-5}$ (**Supplementary Fig. 1b**). Among them, 15 SNPs were located at 12q24, including rs11066015 ($P_{G \times E} = 1.42 \times 10^{-23}$), rs11066280 ($P_{G \times E} = 1.25 \times 10^{-17}$) and rs2074356 ($P_{G \times E} = 3.38 \times 10^{-16}$), which our previous report showed all interact with alcohol drinking to promote ESCC risk[12]. rs11066015 is in strong LD with rs671 ($r^2 = 0.79$), a functional SNP in *ALDH2* (encoding aldehyde dehydrogenase-2) that is known to be associated with both a flushing response to alcohol intake and ESCC risk in a drinking-behavior–specific manner[4,7–9]. We then performed a fast-track replication in the replication 1 samples of the ten remaining tag SNPs located in regions other than 12q24. Of these ten SNPs, eight did not replicate (all $P_{G \times E} > 0.05$; **Supplementary Table 4**), and we did not evaluate them further. Additional replication (replication 2) of rs9288520 at 2q22 and rs17450420 at 13q33 verified their associations with ESCC risk (combined-sample $P_{G \times E} = 4.39 \times 10^{-11}$ and $P_{G \times E} = 4.80 \times 10^{-8}$, respectively). The minor allele of rs9288520 was associated with reduced risk of ESCC in all nondrinkers (OR = 0.81, 95% CI 0.77–0.86, $P = 4.72 \times 10^{-12}$) but was associated with increased risk in all drinkers (OR = 1.09, 95% CI 1.02–1.16, $P = 0.01$). Similarly, the minor allele of rs17450420 was associated with reduced ESCC risk in nondrinkers (OR = 0.78, 95% CI 0.68–0.89, $P = 0.0002$) but with increased risk in drinkers (OR = 1.34, 95% CI 1.16–1.54, $P = 4.65 \times 10^{-5}$; **Table 4**). Neither of these two SNPs was significantly related to drinking status in cases, controls or the combined sample (data not shown). To increase the spectrum of variants tested, we performed an imputation analysis in the GWAS set and identified two and nine imputed SNPs, respectively, in the two regions that showed significant interactions with drinking ($P_{G \times E} < 10^{-4}$); however, none of these SNPs was more significantly related to cancer than the index markers in each region, rs17450420 and rs9288520 (**Fig. 1h**,**i**).

Alcohol use and interaction in ESCC susceptibility

Stratified analyses showed that the nine newly associated SNPs at 22q12, 17q21, 17p13, 16q12.1, 3q27 and 18p11 were all significantly associated with ESCC risk in the same direction in both drinkers and nondrinkers; the associations did not differ significantly between subgroups categorized by alcohol-drinking status (**Supplementary Table 5**). The associations for the eight SNPs (rs1042026, rs3805322, rs17028973, rs1614972, rs17033, rs1229977, rs1789903 and rs1893883) at 4q23 differed by alcohol use, with higher risk in drinkers than in nondrinkers (interaction $P = 2.54 \times 10^{-7}$ to $P = 3.23 \times 10^{-2}$; **Table 3**), which is consistent with previously published epidemiologic data[4,7,9,15,17]. An analysis of the joint effects of drinking, rs1042026 in *ADH1B* and rs11066015 in *ALDH2* on risk of developing ESCC identified that the odds in drinkers carrying risk alleles at both *ADH1B* (GA or AA genotype) and *ALDH2* (GA or AA genotype) was approximately fourfold higher than that in drinkers carrying the nonrisk *ADH1B* G and *ALDH2* G alleles and was more than threefold higher than that in nondrinkers carrying the risk alleles. The effect sizes of

Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

© 2012 Nature America, Inc. All rights reserved.

the *ADH1B* and *ALDH2* variants for ESCC risk in nondrinkers were not large (**Fig. 2** and **Supplementary Table 6**).

Replication of susceptibility loci in a high-risk population

We next examined whether these significant loci were also associated with susceptibility to ESCC in 1,410 cases and 1,656 controls obtained from a high-risk population in Shanxi province, China, as described in a previous GWAS[11]. We found that among the 18 SNPs listed in **Tables 2–4**, 4 showed significant association (**Supplementary Table 7**) in this independent dataset. rs2239815 (OR = 1.24, 95% CI 1.12–1.38, $P = 3.23 \times 10^{-5}$), rs4822983 (OR = 1.28, 95% CI 1.07–1.54, $P = 0.0082$) and rs1033667 (OR = 1.35, 95% CI 1.20–1.51, $P = 2.95 \times 10^{-7}$) at 22q12 and rs2239612 (OR = 1.15, 95% CI 1.01–1.30, $P = 0.0343$) at 3q27 were all associated with increased ESCC risk in this group, as was observed in the other groups. A gene-drinking interaction analysis showed that rs1614972 in *ADH1C* at 4q23 had evidence for replication (OR = 1.37, 95% CI 1.03–1.82, $P_{G \times E} = 0.0281$). In this case-control group, there was some evidence for an interaction between rs2847281 at 18p11 (*PTPN2*) and alcohol drinking (OR = 1.55, 95% CI 1.12–2.15, $P_{G \times E} = 0.0083$).

DISCUSSION

In a multistage GWAS, we identified nine new susceptibility loci associated with ESCC risk across three independent study groups comprising a total of 10,123 cases and 10,664 controls. Among these loci, three had a significant interaction with alcohol drinking, an important lifestyle risk factor in the development of ESCC. We also confirmed some of our findings in an independent study from a high-risk population. To the best of our knowledge, this is one of the largest studies to explore gene-environment interactions for risk of developing ESCC by incorporating alcohol-drinking status into the primary GWAS stage 1 analysis.

Among the six regions with a notable marginal effect for risk of ESCC, two at 16q12.1 tagged by rs4785204 and rs7206735 contain the *TMEM188*, *HEATR3* and *PAPD5* genes, which are interesting and plausible candidate genes worthy of follow-up studies. The variant rs6503659 is located 13,595 bp downstream of *JUP* and 6,366 bp upstream of *HAP1* at 17q21. *JUP* encodes γ-catenin, a cytoplasmic protein that has a similar structure and function to β-catenin and serves as a cell-to-cell attachment molecule through its interaction with E-cadherin[18,19]. The role of γ-catenin in cancer is complex and dependent on the cellular context. Functional loss of γ-catenin results in tumor invasion or metastasis, and γ-catenin is an important part of Wnt signaling[20–22]. Therefore, subtle changes in γ-catenin expression caused by genetic variation could potentially influence the differentiation or invasion of certain transformed cells, resulting in cancer formation. *HAP1* produces huntingtin-associated protein-1, a binding partner of the Huntington's disease protein huntingtin. *HAP1* is involved in vesicular transport, gene transcription regulation, membrane receptor trafficking and other functions such as calcium release and protein aggregation[23,24]. However, the function of *HAP1* in cancer is not clear.

In a further imputation analysis, we identified two LD blocks at 22q12 that contain *XBP1* and *CHEK2*, which encode X-box binding protein 1 and a cell-cycle checkpoint kinase, respectively. *XBP1* is an important part of the unfolded protein response that is involved in the regulation of endoplasmic reticulum stress–mediated apoptosis, and aberrant expression of *XBP1* has been implicated in cancer development and progression, as well as in resistance to drugs[25–28]. *CHEK2* is responsible for preventing DNA-damaged cells from entering into mitosis, a crucial step to avert cancer development. It has therefore been considered as a candidate cancer susceptibility gene[29]. Markers near *CHEK2* have been found previously to be promising signals in the NCI GWAS[11] that we used for replication in this study. In view of the probable important roles of *XBP1* and *CHEK2* in cancer, it is plausible that genetic variations influencing the functions of these genes may confer susceptibility to ESCC.

rs2239612 is located at 3q27 in *ST6GAL1*, which encodes ST6 β-galactosamide α-2,6-sialyltransferase. Previous studies have shown that *ST6GAL1* is upregulated in many types of human cancers, and elevated expression of *ST6GAL1* is also correlated with tumor invasiveness and metastasis[30–33]. rs17761864 is located at 17p13 in *SMG6* (also known as *EST1A*), whose product is an essential factor in nonsense-mediated mRNA decay and telomere maintenance[34,35]; however, whether this gene has a role in cancer is currently unknown. The variant rs2847281 is located at 18p11 in *PTPN2*, encoding non-receptor type 2 protein tyrosine phosphatase, which not only influences the development of the immune system but is also linked to a number of autoimmune diseases and cancer[36,37].

In this study, we performed gene-environment interaction analyses by testing for differences in the per-allele odds ratios between ever drinkers and never drinkers. These analyses identified three genomic regions that had significant interactions with alcohol consumption to promote risk of developing ESCC. Notably, on chromosome 4q23 there is a region that harbors a cluster of seven genes encoding alcohol dehydrogenase (ADH) family proteins ((listed 5′ to 3′) *ADH7*, *ADH1C*, *ADH1B*, *ADH1A*, *ADH6*, *ADH4* and *ADH5*). ADHs oxidize alcohol to acetaldehyde, a carcinogen that is probably important in the etiology of alcohol-related cancers[38]. Drinkers with the fast ADH metabolizer genotype produce more acetaldehyde and are expected to have an elevated risk of these cancers. However, in this study, we were unable to determine the exact contribution of individual variants because of the LD pattern over the region covering the *ADH* genes. Deep sequencing of this region is warranted to map candidate genes and variants for follow-up functional analyses. Using a genome-wide gene-environment interaction analysis, we found that the most significant interaction region was for variants at 12q24 harboring *ALDH2*, which encodes aldehyde dehydrogenase-2 that, in turn, detoxifies acetaldehyde to acetate. The directions of our associations reported here are consistent with those reported in our previous GWAS[12] and other published studies[4,8,9]. Furthermore, in the present study, we evaluated the joint effects of *ADH1B* and *ALDH2* variants and drinking on ESCC risk and found that individuals who carried both of the risk alleles of *ADH1B* and *ALDH2* and were classified as alcohol drinkers had the highest risk. These findings clearly indicate a gene-environment interaction between alcohol use and genetic variation in the alcohol-metabolizing pathway for developing ESCC. Because ADHs oxidize alcohol to carcinogenic acetaldehyde, which is then detoxified by aldehyde dehydrogenases, it is anticipated that individuals with the combination of the fast alcohol metabolizer genotype and the slow acetaldehyde metabolizer genotype would be most susceptible to ESCC. These results strongly highlight the potential importance of reducing alcohol use in individuals carrying high-risk alleles to reduce ESCC risk.

We also identified associations with ESCC risk for rs9288520, located upstream of *IGFBP2* at 2q22, and rs17450420, located in a gene desert upstream of *SLC10A2* at 13q33. These two variants did not show marginal effects but were significantly associated with risk when alcohol drinking was incorporated into the genome-wide gene-environment interaction analysis. Compared to common alleles, the minor alleles of these two SNPs were associated with decreased risk of ESCC in nondrinkers and increased risk in drinkers. *IGFBP2* produces

© 2012 Nature America, Inc. All rights reserved.

insulin-like growth factor binding protein 2, which is involved in cell proliferation, migration and apoptosis, and elevated serum IGFBP2 concentrations have been detected in patients with various types of cancer[39]. Interestingly, it has been shown that *IGFBP2* RNA is overexpressed in the placenta and fetal lungs of rats fed with alcohol, and this overexpression is associated with ethanol-induced growth retardation[40]. *SLC10A2* encodes a sodium/bile acid cotransporter and has been suggested to be associated with alcohol dependence[41].

Because of the stringent *P* values we required for statistical significance to prevent false-positive findings in the GWAS, additional associations with promising *P* values were not confirmed in the present study, underscoring the need to continue the search for new loci[13]. Therefore, it is important to undertake complementary strategies to discover additional variants, particularly when some genetic effects are dependent on environmental exposure and may show a substantial effect only when a specific environmental exposure is present. Indeed, by replication of more potential associated SNPs in expanded samples and by performing a genome-wide gene-environment interaction analysis, we extended our GWAS results with the discovery of nine new ESCC susceptibility loci.

We also replicated the results in an additional case-control group from Shanxi province, a region with extremely high rates of ESCC in China[11]. We verified that four of the nine loci identified in the GWAS and replication samples also had significant marginal genetic effects on ESCC risk in this high-risk population; however, we found only modest evidence for a gene–alcohol drinking interaction in this population. This apparent inconsistency probably reflects differences in environmental exposures between general and high-risk populations. It is well known that in general populations, alcohol drinking and tobacco smoking are the major risk factors for ESCC[15,17]. However, in some high-risk regions of the Shanxi and Henan provinces in China, alcohol drinking has little or no association with ESCC risk[42,43], whereas other factors such as nutritional deficiencies, family history and certain chemical carcinogens in the diet are strongly associated with this type of cancer[44–46]. In this context, it is therefore not surprising to observe different genetic risks between general and high-risk populations. These differences also emphasize the importance of further analyses of interactions between genetic variants and the specific environmental factors in high-risk populations.

In conclusion, we have identified nine new susceptibility loci for ESCC in Chinese populations, extending our previous findings and advancing the understanding of the genetic etiology of ESCC. The newly identified susceptibility loci warrant follow-up fine-mapping and functional studies. Furthermore, the risk variants in the alcohol metabolism pathway that we have confirmed in this large study might be useful for identifying high-risk individuals for the prevention of ESCC in the Chinese population, particularly where alcohol consumption is a possible health risk.

URLs. R, http://www.r-project.org/; MACH, http://www.sph.umich.edu/csg/abecasis/MACH/index.html; LocusZoom, http://csg.sph.umich.edu/locuszoom/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was funded by the National High-Tech Research and Development Program of China (2009AA022706 to D.L.), the National Basic Research Program of China (2011CB504303 to D.L. and W.T.), the National Natural Science Foundation of China (30721001 to D.L., Q.Z. and Z.L.) and the Intramural Research Program of the US National Institutes of Health, NCI and the Division of Cancer Epidemiology and Genetics.

AUTHOR CONTRIBUTIONS

D.L. was the overall principle investigator of the study who conceived the study and obtained financial support, was responsible for study design, oversaw the entire study, interpreted the results and wrote parts of and synthesized the paper. C.W. performed overall project management, oversaw laboratory analyses, performed statistical analyses and drafted the initial manuscript. P.K. oversaw statistical analyses, interpreted the results and reviewed the manuscript. Y. Li and L.L. performed the imputation analysis and reviewed the manuscript. K.Z., J.C., Y.Q., Yuling Zhou and Y. Liu performed laboratory analyses. Z. Hu, G.J. and H.S. were responsible for subject recruitment and sample preparation of Nanjing samples. Z. He, C.G., C.L., H.Y. and Y.K. were responsible for subject recruitment and sample preparation of Henan samples. W.J., J.F. and Y. Zeng were responsible for subject recruitment and sample preparation of Guangzhou samples. X.M. and T.W. provided some of the control samples. Yifeng Zhou was responsible for subject recruitment of the additional validation cohorts. D.Y. and W.T. performed subject recruitment and sample preparation of Beijing samples. Q.Z. and Z.L. provided some of the financial support and reviewed the manuscript. Z.W., C.C.A., N.H., N.D.F., T.D., A.M.G., S.J.C. and P.R.T. performed subject recruitment, sample preparation, laboratory analysis and statistical analysis of Shanxi samples and reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2411.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sun, T. *et al.* Polymorphisms of death pathway genes *FAS* and *FASL* in esophageal squamous-cell carcinoma. *J. Natl. Cancer Inst.* **96**, 1030–1036 (2004).
2. Zhang, X. *et al.* Identification of functional genetic variants in *cyclooxygenase-2* and their association with risk of esophageal cancer. *Gastroenterology* **129**, 565–576 (2005).
3. Sun, T. *et al.* A six-nucleotide insertion-deletion polymorphism in the *CASP8* promoter is associated with susceptibility to multiple cancers. *Nat. Genet.* **39**, 605–613 (2007).
4. Lewis, S.J. & Smith, G.D. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1967–1971 (2005).
5. Hiyama, T., Yoshihara, M., Tanaka, S. & Chayama, K. Genetic polymorphisms and esophageal cancer risk. *Int. J. Cancer* **121**, 1643–1658 (2007).
6. Akbari, M.R. *et al.* Candidate gene association study of esophageal squamous cell carcinoma in a high-risk region in Iran. *Cancer Res.* **69**, 7994–8000 (2009).
7. Hashibe, M. *et al.* Multiple *ADH* genes are associated with upper aerodigestive tract cancers. *Nat. Genet.* **40**, 707–709 (2008).
8. McKay, J.D. *et al.* A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet.* **7**, e1001333 (2011).
9. Cui, R. *et al.* Functional variants in *ADH1B* and *ALDH2* coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology* **137**, 1768–1775 (2009).
10. Wang, L.D. *et al.* Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at *PLCE1* and *C20orf54*. *Nat. Genet.* **42**, 759–763 (2010).
11. Abnet, C.C. *et al.* A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42**, 764–767 (2010).
12. Wu, C. *et al.* Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat. Genet.* **43**, 679–684 (2011).
13. Panagiotou, O.A. & Ioannidis, J.P. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **41**, 273–286 (2012).
14. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
15. Islami, F. *et al.* Alcohol drinking and esophageal squamous cell carcinoma with focus on light-drinkers and never-smokers: a systematic review and meta-analysis. *Int. J. Cancer* **129**, 2473–2484 (2011).
16. Hunter, D.J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
17. Gao, Y.T. *et al.* Risk factors for esophageal cancer in Shanghai, China. I. Role of cigarette smoking and alcohol drinking. *Int. J. Cancer* **58**, 192–196 (1994).

© 2012 Nature America, Inc. All rights reserved.

18. Aberle, H., Schwartz, H. & Kemler, R. Cadherin-catenin complex: protein interactions and their implications for cadherin function. *J. Cell Biochem.* **61**, 514–523 (1996).
19. Bullions, L.C. & Levine, A.J. The role of β-catenin in cell adhesion, signal transduction, and cancer. *Curr. Opin. Oncol.* **10**, 81–87 (1998).
20. Morin, P.J. *et al.* Activation of β-catenin–Tcf signaling in colon cancer by mutations in β-catenin or APC. *Science* **275**, 1787–1790 (1997).
21. Kolligs, F.T. *et al.* γ-catenin is regulated by the APC tumor suppressor and its oncogenic activity is distinct from that of β-catenin. *Genes Dev.* **14**, 1319–1331 (2000).
22. Simcha, I. *et al.* Suppression of tumorigenicity by plakoglobin: an augmenting effect of N-cadherin. *J. Cell Biol.* **133**, 199–209 (1996).
23. Li, X.J. *et al.* A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378**, 398–402 (1995).
24. Wu, L.L. & Zhou, X.F. Huntingtin associated protein 1 and its functions. *Cell Adh. Migr.* **3**, 71–76 (2009).
25. Kim, R., Emi, M., Tanabe, K. & Murakami, S. Role of the unfolded protein response in cell death. *Apoptosis* **11**, 5–13 (2006).
26. Koong, A.C., Chauhan, V. & Romero-Ramirez, L. Targeting XBP-1 as a novel anti-cancer strategy. *Cancer Biol. Ther.* **5**, 756–759 (2006).
27. Romero-Ramirez, L. *et al.* XBP1 is essential for survival under hypoxic conditions and is required for tumor growth. *Cancer Res.* **64**, 5943–5947 (2004).
28. Shuda, M. *et al.* Activation of the *ATF6*, *XBP1* and *grp78* genes in human hepatocellular carcinoma: a possible involvement of the ER stress pathway in hepatocarcinogenesis. *J. Hepatol.* **38**, 605–614 (2003).
29. Antoni, L., Sodha, N., Collins, I. & Garrett, M.D. CHK2 kinase: cancer susceptibility and cancer therapy—two sides of the same coin? *Nat. Rev. Cancer* **7**, 925–936 (2007).
30. Dall'Olio, F., Chiricolo, M. & Lau, J.T. Differential expression of the hepatic transcript of β-galactoside α-2,6-sialyltransferase in human colon cancer cell lines. *Int. J. Cancer* **81**, 243–247 (1999).
31. Wang, P.H. *et al.* Enhanced expression of α 2,6-sialyltransferase ST6Gal I in cervical squamous cell carcinoma. *Gynecol. Oncol.* **89**, 395–401 (2003).
32. Recchi, M.A. *et al.* Multiplex reverse transcription polymerase chain reaction assessment of sialyltransferase expression in human breast cancer. *Cancer Res.* **58**, 4066–4070 (1998).
33. Pousset, D., Piller, V., Bureaud, N., Monsigny, M. & Piller, F. Increased α 2,6 sialylation of N-glycans in a transgenic mouse model of hepatocellular carcinoma. *Cancer Res.* **57**, 4249–4256 (1997).
34. Eberle, A.B., Lykke-Andersen, S., Mühlemann, O. & Jensen, T.H. SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nat. Struct. Mol. Biol.* **16**, 49–55 (2009).
35. DeZwaan, D.C. & Freeman, B.C. The conserved Est1 protein stimulates telomerase DNA extension activity. *Proc. Natl. Acad. Sci. USA* **106**, 17337–17342 (2009).
36. Doody, K.M., Bourdeau, A. & Tremblay, M.L. T-cell protein tyrosine phosphatase is a key regulator in immune cell signaling: lessons from the knockout mouse model and implications in human disease. *Immunol. Rev.* **228**, 325–341 (2009).
37. Dubé, N. & Tremblay, M.L. Involvement of the small protein tyrosine phosphatases TC-PTP and PTP1B in signal transduction and diseases: from diabetes, obesity to cell cycle, and cancer. *Biochim. Biophys. Acta* **1754**, 108–117 (2005).
38. World Cancer Research Fund/American Institute for Cancer Research. *Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective* (AICR, Washington, DC, 2007).
39. Hoeflich, A. *et al.* Insulin-like growth factor-binding protein 2 in tumorigenesis: protector or promoter? *Cancer Res.* **61**, 8601–8610 (2001).
40. Fatayerji, N., Engelmann, G.L., Myers, T. & Handa, R.J. *In utero* exposure to ethanol alters mRNA for insulin-like growth factors and insulin-like growth factor-binding proteins in placenta and lung of fetal rats. *Alcohol. Clin. Exp. Res.* **20**, 94–100 (1996).
41. Edenberg, H.J. *et al.* Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol. Clin. Exp. Res.* **34**, 840–852 (2010).
42. Gao, Y. *et al.* Risk factors for esophageal and gastric cancers in Shanxi Province, China: a case-control study. *Cancer Epidemiol.* **35**, e91–e99 (2011).
43. He, Z. *et al.* Prevalence and risk factors for esophageal squamous cell cancer and precursor lesions in Anyang, China: a population-based endoscopic survey. *Br. J. Cancer* **103**, 1085–1088 (2010).
44. Mark, S.D. *et al.* Prospective study of serum selenium levels and incident esophageal and gastric cancers. *J. Natl. Cancer Inst.* **92**, 1753–1763 (2000).
45. Gao, Y. *et al.* Family history of cancer and risk for esophageal and gastric cancer in Shanxi, China. *BMC Cancer* **9**, 269 (2009).
46. Lu, S.H. *et al.* Relevance of N-nitrosamines to esophageal cancer in China. *J. Cell Physiol. Suppl.* **4**, 51–58 (1986).

© 2012 Nature America, Inc. All rights reserved.

ONLINE METHODS

Study subjects. This study was an extension of our previous GWAS in which the genome-wide scan sample comprised 2,031 cases with ESCC and 2,044 controls and the replication samples comprised 6,276 cases and 6,165 controls. The sources and characteristics of these study subjects were described previously[12]. To further increase our statistical power for validation, we added an additional 1,816 cases and 2,455 controls in the present study. These cases and controls were recently recruited from the Han Chinese population through collaboration with multiple hospitals in Beijing and Jiangsu province, China. A diagnosis of ESCC was confirmed by either histopathologic or cytologic analyses, as described previously[12]. Demographic characteristics of the subjects, including age, sex, smoking status and drinking status, were obtained from each patient's medical records. Control subjects were selected on the basis of a physical examination and were frequency matched for age and sex to the cases with ESCC, as previously described[12]. All the cases and controls for each of the replication cohorts were sampled from the same locality and the same population to assure minimal population stratification. In replication 1, a total of 3,571 cases and 3,602 controls were collected from the Beijing region, and in replication 2, 4,521 cases and 5,018 controls were recruited from the Jiangsu, Henan and Guangdong provinces. This study also included an additional validation cohort consisting of 1,410 cases with ESCC and 1,656 controls from a study conducted in a population at high risk for ESCC in Shanxi, China, as described previously[11]. For this study, alcohol drinking status was assessed by a detailed questionnaire[42]. For the present analysis, individuals were classified as drinkers if they reported drinking any form of alcohol at least twice a week; otherwise, they were defined as nondrinkers. Individuals who reported smoking more than 100 cigarettes in their life or smoking tobacco in a pipe more than 100 times were defined as smokers; all others were defined as nonsmokers. The distributions of the selected characteristics among the cases and controls for each of the study sets examined in the genome-wide scan and in each replication are shown in **Table 1**. At recruitment, informed consent was obtained from each subject, and the study was approved by the institutional review boards of the Chinese Academy of Medical Sciences Cancer Institute, Peking University, SunYat-Sen University Cancer Center, Nanjing Medical University, the Medical College of Soochow University, Shanxi Cancer Hospital and the US NCI.

SNP selection and genotyping for replication. In replication 1, we selected SNPs with marginal significance ($10^{-7} < P \leq 10^{-4}$) for the genetic association analysis and SNPs with $P \leq 10^{-4}$ for the genome-wide gene × drinking interaction analysis. All selection was based on our previous GWAS scan results[12]. We adopted a two-step approach to select these SNPs. First, we excluded those SNPs with MAF < 0.01 in both cases and controls and those with genotype frequencies not conforming to Hardy-Weinberg equilibrium (HWE) in the controls ($P < 0.01$). Second, we computed the correlation coefficient (r) of each pair of adjacent SNPs on the same chromosome to assess LD status. SNPs with $r^2 > 0.8$ were considered to be in one LD block, and we thus selected the most significant SNP (with the lowest P value) in the block for replication. Using these criteria, we selected 175 SNPs for the genetic association analysis and 12 SNPs for the genome-wide gene × drinking interaction analysis in replication 1. Genotyping in replication cohort 1 was accomplished with an Illumina GoldenGate Assay of 187 attempted SNPs (Illumina). We filtered out SNPs with call rate <95% or with genotype frequencies in controls departing from HWE ($P < 0.01$). Finally, 169 and 10 genotyped SNPs passed quality control and were included in the final genetic association analysis and the final gene × drinking interaction analysis, respectively. We next selected SNPs with association at a significance of $P < 0.01$ for the replication 2 analysis. With this criterion, 18 SNPs for the genetic association analysis and 2 SNPs for the gene × drinking interaction analysis were selected and genotyped using a TaqMan genotyping platform (ABI 7900HT Real Time PCR system, Applied Biosystems) in replication 2.

For genotyping quality control, we implemented several measures in the replication assays, including (i) case and control samples were mixed in the plates, (ii) persons who performed the genotyping assays were not aware of the case or control status of the samples, (iii) both positive and negative (no DNA) control samples were included on every 384-well assay plate and (iv) replication of nearly 10% of the total DNA samples (400 in the GWAS scan and 700 in replication 1) was performed using the TaqMan genotyping platform (with duplication concordances of 99.92% and 99.99%, respectively).

Statistical analyses. For the GWAS, associations between genotypes and risk of developing ESCC were analyzed by an additive model in a logistic regression (genotypic trend effect with a 1-degree-of-freedom test) framework with age, sex, smoking, drinking and the first three principal components from EIGENSTRAT as covariates[12]. SNPs imputed using the GWAS scan data were included in this logistic regression model using SNP 'dosages' (the expected allele counts). Conditional association analyses were conducted by including in the unconditional logistic regression model the most significant SNP on 4q23, 16q12.1 or 22q12 and examining the association between each of the remaining SNPs and risk of ESCC. For the analysis of the gene × drinking interaction, we tested the interaction between each SNP and drinking status by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression based on the equation $Y = \beta_0 + \beta_1 \times \text{SNP} + \beta_2 \times \text{drinking status} + \beta_3 \times (\text{SNP} \times \text{drinking status})$. Here, Y is the logit of the probability of being a case, β_0 is a constant, β_1 and β_2 are the main effects of SNP and dinking status, respectively, and β_3 is the interaction term to be tested. We further performed stratified analyses of significant SNPs identified by a two-phase replication strategy in different cohorts: we used case or control status as the outcome and tested the associations in the nondrinker and drinker groups. Sex, age, smoking and first three principal components served as covariates in both the genome-wide gene × drinking interaction analysis and the stratified analysis. The odds ratios calculated are presented for the minor allele of each SNP. For fine mapping of the significant regions, we used MACH software (see URLs) to impute untyped markers using LD and haplotype information from the HapMap II CHB + JPT populations as the reference set. To identify susceptibility genes underlying the various associations, we analyzed the LD patterns around the risk-associated SNPs and determined LD blocks where the risk-associated SNPs were located. We then investigated the gene or genes covered by the LD blocks. The LD structures and haplotype block plots were generated using Haploview v4.1 software (see URLs). Significant regions were plotted using the online tool LocusZoom (see URLs).

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2411

Joint Linkage and Association Analysis with Exome Sequence Data Implicates *SLC25A40* in Hypertriglyceridemia

Elisabeth A. Rosenthal,[1,*] Jane Ranchalis,[1] David R. Crosslin,[2] Amber Burt,[1] John D. Brunzell,[3] Arno G. Motulsky,[1,2] Deborah A. Nickerson,[2] NHLBI GO Exome Sequencing Project, Ellen M. Wijsman,[1,2,4] and Gail P. Jarvik[1,2]

Hypertriglyceridemia (HTG) is a heritable risk factor for cardiovascular disease. Investigating the genetics of HTG may identify new drug targets. There are ~35 known single-nucleotide variants (SNVs) that explain only ~10% of variation in triglyceride (TG) level. Because of the genetic heterogeneity of HTG, a family study design is optimal for identification of rare genetic variants with large effect size because the same mutation can be observed in many relatives and cosegregation with TG can be tested. We considered HTG in a five-generation family of European American descent (n = 121), ascertained for familial combined hyperlipidemia. By using Bayesian Markov chain Monte Carlo joint oligogenic linkage and association analysis, we detected linkage to chromosomes 7 and 17. Whole-exome sequence data revealed shared, highly conserved, private missense SNVs in both *SLC25A40* on chr7 and *PLD2* on chr17. Jointly, these SNVs explained 49% of the genetic variance in TG; however, only the *SLC25A40* SNV was significantly associated with TG ($p = 0.0001$). This SNV, c.374A>G, causes a highly disruptive p.Tyr125Cys substitution just outside the second helical transmembrane region of the SLC25A40 inner mitochondrial membrane transport protein. Whole-gene testing in subjects from the Exome Sequencing Project confirmed the association between TG and *SLC25A40* rare, highly conserved, coding variants ($p = 0.03$). These results suggest a previously undescribed pathway for HTG and illustrate the power of large pedigrees in the search for rare, causal variants.

Introduction

Cardiovascular disease (CVD) is the leading cause of death in the United States and poses a significant morbidity and cost for treatment after cardiac events. CVD is associated with the correlated traits of high LDL, low HDL, high total cholesterol, high triglyceride (TG) (defined as $200 \leq TG < 500$ mg/dl in adults[1]), hypertension, diabetes, and metabolic syndrome. Furthermore, CVD is associated with environmental variables that can be confounded with lipid levels, such as obesity, poor diet, lack of exercise, and smoking.

Hypertriglyceridemia (HTG), defined as TG > 500 mg/dl in adults,[1] is a risk factor for CVD, independent of high LDL and low HDL.[2–7] Although HDL and TG levels are highly correlated, an independent role of HDL level in CVD etiology has been challenged by recent Mendelian randomization studies and the failure of cholesteryl ester transfer protein inhibitors to reduce vascular events.[8,9] Conversely, Mendelian randomization suggests a causal role of TG in CVD.[10] Elevated TG has been implicated in both microvascular and macrovascular endothelial damage with associated atherosclerosis.[6] Within the United States, ~16% of adults of European origin have high TG levels, indicating a need for further intervention.[7] However, studies of TG level and lipid metabolism have been difficult.[6,7] One reason for this difficulty is the existence of high within-individual variation of TG measurement that expands with increasing TG. High TG is also associated with high LDL and low HDL, making it difficult to tease apart the effect of specific lipids on CVD risk within studies.

There are currently few pharmacological treatments for elevated TG. The most common treatment, fibrates, effectively reduces elevated TG and reduces the risk for cardiovascular events.[11,12] Unfortunately, some 5% of individuals stop using fibrates because of side effects.[13] Other potential drugs, targeting different parts of the metabolic pathway, have been found to have intolerable complications such as fatty liver or to actually raise the risk of cardiovascular events.[13]

In order to find additional effective treatments, studies of TG need to be undertaken. Focusing on the genetic control of elevated TG may remove some of the confounding with LDL and HDL and lead to new drug targets. TG is known to be heritable and there are several known genetic mutations that influence TG levels, most notably those in the structural loci for ApoA5 and ApoC3.[14–21] In mice, expression of both *Apoa5* and *Apoc3* are associated with TG levels.[22–25] Whereas circulating levels of ApoA5 are negatively associated with TG levels, ApoC3 levels are positively correlated with TG. However, there is conflicting evidence in humans for an association between CVD and single-nucleotide variants (SNVs) within *APOC3* (MIM 606368) and *APOA5* (MIM 107720).[26–30] These and other known genetic variants explain only ~10% of the genetic

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [3]Division of Metabolism, Endocrinology, and Nutrition, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [4]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
*Correspondence: erosen@uw.edu
http://dx.doi.org/10.1016/j.ajhg.2013.10.019. ©2013 by The American Society of Human Genetics. All rights reserved.

variation in TG,[20,21] which may explain the conflicting evidence indicating a relationship between regulatory SNVs and CVD.

The genetic heterogeneity in the etiology of high TG makes large family studies the optimal design for identification of novel TG loci with large effect sizes.[31] This design allows for the study of numerous people with an identical mutation and the ability to study cosegregation as well as association. For these reasons, we set out to discover genes underlying elevated TG in a single large family, described as familial combined hyperlipidemia (FCHL [MIM 144250]) in the 1980s and resampled in the 2000s, yielding five generations, four of which have phenotype and marker data. FCHL is genetically heterogeneous and characterized by variable atherogenic lipoprotein levels in multiple family members,[32–35] making it an ideal diagnosis to identify novel lipid level genes. This family includes four related individuals, including the proband, with TG measurement in the top 1/2 percentile of adults in the United States,[36] and previously reported genetic variation associated with TG does not explain the elevated levels. Segregation analyses supported at least one major gene for high TG in this family.

By using linkage analysis and whole-exome sequence data, we detected two candidate genes of interest. Although whole-exome sequencing has proven powerful in discovering underlying loci for simple Mendelian traits and de novo mutations, it has not yet been as successful in uncovering major contributing loci to complex traits. With family data, we have the opportunity to observe the same very rare variant in multiple family members and to evaluate cosegregation of a single variant with a phenotype. In this paper, we show that reduction of the search space by using linkage analysis in a large family, and then focusing on rare exome variants, can be used to successfully discover multiple major contributing loci for a heterogeneous, complex quantitative trait. This allows single gene burden testing in a second cohort, eliminating the need for multiple gene tests, which increases the power to detect an association.

Subjects and Methods

Subjects and Phenotype

The family presented here was ascertained in the 1980s as part of a cohort of four families of European American descent that had the diagnosis of FCHL. Observed TG in these families was previously shown to be associated with increased risk of CVD mortality.[37] The proband had fasting triglycerides >2,900 mg/dl[38] and normal LDL-C levels. The family was subsequently ascertained and diagnosed with FCHL, based on hypertriglyceridemia and hypercholesterolemia in first-degree relatives. The proband was originally hypothesized to represent a homozygote for FCHL, thought to be a Mendelian trait at that time. Re-collection of samples as well as sampling of additional family members, particularly in the youngest generation, took place in the 2000s. The final pedigree contains 5 generations of 121 individuals and includes five large sibships (n = 5, 5, 6, 7, and 9). Fasting TG was measured on 85 individuals (74 in the 1980s and 41 in the 2000s) via standard enzymatic methods, and reported in mg/dl.[39] The same methods were employed in the 1980s and 2000s. TG ranged between 22 and 2,926 mg/dl, with a median of 111 mg/dl for individuals not taking statin drugs. Twenty-six family members had TG > 200 mg/dl and 8 had TG > 500 mg/dl. Four related subjects in the second and third generations had TG > 1,000 mg/dl, making this family a good candidate for linkage analysis. In addition, none of the subjects were taking lipid-lowering medications in the 1980s and only nine were taking such medication in the 2000s. All subjects, or their representative, gave written, informed consent and this study was approved by the University of Washington Human Subject Review Board (FWA #00006878).

TG Adjustment

Fasting TG level is a highly right-skewed phenotype that is associated with age, sex, and body mass index in typical European American populations. Because of the ascertainment, the TG in this data set does not follow typical distributions and adjusting the TG via a linear model with the family data would not result in a reliable phenotype. To increase the power to detect linkage, we corrected for this ascertainment by adjusting TG (adjTG) by using the age- and sex-based means and standard deviations from The Lipid Research Clinics Population Studies Data Book, Volume 1,[40] added the constant 8 to avoid negative values, and then log transformed the data (log-adjTG) to reduce skew.

Because TG was measured in the 1980s and 2000s for 30 individuals, we generated three overlapping phenotype data sets. Data set "1980only" contains phenotypes from the 1980s only (n = 74, all statin-free). Data set "1980plus" contains phenotypes from the 1980s and phenotypes for only the newly sampled individuals from the 2000s who are not taking lipid-lowering medications (n = 85 statin-free). Data set "2000plus" contains all phenotypes from the 2000s for individuals not taking lipid-lowering medications and phenotypes from the 1980s for individuals who only have phenotype at this time point or who were taking statin drugs in the 2000s (n = 85 statin-free). Phenotypes in data sets 1980plus and 2000plus differ for 21 individuals.

Genotyping

The family was extensively genotyped via highly polymorphic markers, common SNVs, and next-generation exome sequencing. Quality control on all genotype data as well as alignment to the Rutgers Build 35 map was performed as described previously.[41,42] In brief, STR markers from Marshfield Panel 9 (1980s), Prevention Genetics Set 13 (2000s), and selected regions from Decode (1980s and 2000s) were genotyped in the family (n = 82). In addition, ~50K SNVs from the Illumina HumanCVD Bead chip were genotyped in the family (2000s, n = 64).[43] Finally, the exome was sequenced on the Solexa platform for 16 individuals selected for their phenotypes and informativeness for imputing the genotypes of others.[41,44] Novel SNVs of interest, detected with exome sequencing, were genotyped in the entire pedigree via custom TaqMan Genotyping Assays with the Applied Biosystems 7900HT real-time PCR system[45] or Veracode 384-plex Bead Plates.[46]

Joint Segregation and Linkage Analysis

Given that log-adjTG is a complex trait with an unknown mode of inheritance in these data, we used an iterative Bayesian oligogenic

Table 1. Percent Variance of log-adjTG Explained in the Family, by Previously Reported SNVs

Chr	Gene (SNV, Proxy)	MIM #	SNV	Proxy	% Vtot (SD)	% Vg (SD)	p Value
1	*DOCK7*	604774	rs2131925	rs1748195	3 (3)	14 (16)	0.84
1	*GALNT2*	602274	rs1321257	rs4846914	4 (3)	23 (20)	0.15
2	*APOB*	107730	rs1042034	NA	4 (3)	16 (17)	0.73
2	*GCKR*	600842	rs1260326	NA	4 (3)	18 (19)	0.09
3	*MSL2,*[a] *PCCB*	614802, 232050	rs645040	rs3821445	3 (2)	17 (19)	0.34
4	*AFF1*	159557	rs442177	rs3775214	2 (2)	13 (17)	0.68
7	*TBL2*	605842	rs17145738	NA	5 (4)	20 (18)	0.04
8	*LPL,*[a] *LPL*	609708	rs12678919	rs12679834	2 (2)	14 (19)	0.64
8	*TRIB1*[a]	609461	rs2954029	rs17321515	3 (2)	13 (16)	0.60
10	*CYP26A1*[a]	602239	rs2068888	rs4418728	4 (3)	24 (21)	0.25
11	*FADS1, FADS2*	606148, 606149	rs174546	rs1535	3 (2)	14 (18)	0.82
11	*APOA5*	606368	rs3135506	NA	2 (2)	14 (18)	0.67
15	*LIPC*	151670	rs6074	rs871804	4 (3)	18 (17)	0.17
15	*LIPC*	151670	rs261342	NA	2 (2)	15 (19)	0.94
16	*CTF1*[a]	600435	rs11649653	NA	3 (3)	14 (16)	0.47
16	*CETP*	118470	rs7205804	rs1532625	3 (3)	21 (21)	0.99
19	*APOE,*[a] *APOC1*[a]	107741, 107710	rs439401	NA	5 (3)	20 (18)	0.12
19	*SUGP1, GATAD2A*	607992, 614997	rs10401969	rs3794991	1 (2)	10 (17)	0.70
20	*PLTP*	172425	rs4810479	NA	7 (5)	30 (18)	0.56

Estimated percent total variance (Vtot), percent genetic variance (Vg), and p value for pedigree adjusted association of log-adjTG and known or suspected pathogenic SNVs in the family. Proxy SNVs from the CVD chip were used in place of the reported SNVs that did not exist on the CVD chip. Proxy SNVs are correlated with the original SNV ($r^2 > 0.8$). *APOA5* (rs3135506) was genotyped separately in the family. The second column from left gives the gene name for the SNV and its proxy, if necessary. Gene names are from NCBI and aliases are as follows: *DOCK7* = *ANGPTL3*, *AFF1* = *KHL8*, *TBL2* = *MLXIPL*, *SUGP1* = *CILP2*. SD indicates standard deviation.
[a]SNV is near but not within a gene.

joint segregation and linkage analysis approach from the package Loki2.4.7.[47] This approach has many advantages for these data. First, Loki is capable of handling multipoint analysis on a large pedigree. Second, the user does not supply a trait model; rather, a prior distribution on the genetic effects is supplied, allowing the trait model, including the number of QTLs, to vary within a plausible model space. Third, missing trait-locus genotypes can be sampled at each iteration, given the marker segregation in the pedigree. Fourth, genotype covariates can be included in the model, allowing for estimation of the genetic variation explained by a SNV at each iteration, including the information from the estimated missing genotypes. Over an entire run, the average total variation (Vtot) and genetic variation (Vg) explained can then be used to estimate the percent Vtot and Vg explained by that SNV. In addition to modeling the possibility of multiple QTLs, variance resulting from any polygenic effects is captured by Vg. Because calculating p values is time consuming with this method, we calculated the Bayes factor (BF) for 2 cM intervals across the genome, which compares the posterior odds to the prior odds that a QTL is located in the region. We consider a maxBF > 25 as suggestive of linkage.[41,48] All identified linkage regions were jointly analyzed, allowing the regions to compete for the modeled QTLs.

Because the trait under study is skewed, it is important to specify a prior distribution that is expansive enough to include a likely model but not so broad that the space cannot be adequately sampled. Furthermore, the detected regions of linkage should be consistent given differing prior distributions. By using methods described previously,[49–51] we used several values for the underlying additive genetic effects, ranging between 1 and 8 times the standard deviation (SD) of log-adjTG in order to verify consistency of detected linkage regions. Here we report the results for the prior distribution with mean additive genetic effect = 2.7 times the SD of log-adjTG. All runs contained 1,000 burn-in iterations followed by 200,000 iterations (1 chromosome) or 500,000 iterations (2 chromosomes jointly), of which every 10th iteration was saved. The prior distribution on the number of QTLs followed a curtailed Poisson(1) distribution with a maximum of 15. The total map length was assumed to be 3,000 cM. The prior distribution on the allele frequency for all QTLs is set to Uniform(0,1) and cannot be changed by the user.

Candidate Locus Analysis

Candidate loci were assessed for the estimated percent Vtot and Vg that they explained in the family. These candidate variants were chosen from two sources (Tables 1 and 2 and Table S1 available online). First, candidate SNVs known or thought to be associated with TG,[21,52] including *APOA5* variant rs3135506, were assessed within the family primarily via the CVD chip data. If the reported SNV was not on the CVD chip, we attempted to find a correlated

Table 2. Distribution of Novel SNVs under the Linkage Signals on Chr7 and Chr17

Chr.	7	17
# novel sites	53	20
Intergenic	2	1
Intronic	4	1
3′ UTR	1	1
5′ UTR	1	0
Synonymous	23	2
Splice	0	1
Missense	22	14
GERP > 3	12	6
Shared	1	4
Liver expressed	1	2

The final row indicates the number of novel SNVs that are missense or splice sites, have Genomic Evolutionary Rate Profiling score (GERP) > 3, and are shared by at least two relatives with high log-adjTG. Novel is defined as not existing in dbSNP134.

proxy SNV ($r^2 > 0.8$) within 500 bases of the reported SNV.[53,54] *APOA5* SNV rs3135506 was genotyped separately. A Bonferroni adjustment was used for the testing of these known candidate loci.

Candidate rare SNVs were chosen from the exome data in the linkage regions. Exome sites were filtered based on novelty, coding effect, conservation across mammals, and sharing of the rare allele among the four individuals with TG > 1,000. Because the genetic basis of TG in this family had yet to be discovered, we hypothesized that the high levels of TG were due to novel variation. We defined a variant as novel if it did not have an rsID in dbSNP134 at the time of exome sequencing (April 2012)[55] and occurred at most once in the NHLBI Exome Sequencing Project (ESP) data on 6,500 exomes. Only sites that had an effect on the coding sequence (nonsense, missense, or splice) were kept. Conservation across mammals was calculated by the Genomic Evolutionary Rate Profiling (GERP) score;[56] positive values indicate conservation and negative values indicate lack of conservation. Only sites with a GERP > 3 were considered because we expect only mutations at locations with high conservation across mammals would be pathogenic for high TG. Finally, the rare allele had to exist in at least two of the four individuals with TG > 1,000. All pedigree members were genotyped for the novel sites selected with these criteria to confirm their estimated effect. Variance components analysis in the package SOLAR[57] was used to calculate p values for association, in a mixed model analysis that adjusted for correlation among related individuals through the kinship matrix, which captures any underlying shared polygenic effect.

Validation of Novel Gene Association

We sought confirmation of association with TG for any gene containing a novel SNV that remained associated with the trait after genotyping of the full pedigree. By using TG data from unrelated individuals in the PennCATH, TRIUMPH, Cleveland Clinic, ARIC, CARDIA, CHS, FHS, JHS, MESA, and WHI cohorts in the dbGaP posted ESP data, we combined all rare (MAF < 0.5%) missense, nonsense, and splice SNVs with GERP > 4.8 into a single genotype factor (1 = presence of a minor allele and 0 = absence of all minor alleles) in individuals with measured TG. We used a GERP cutoff of 4.8 because this is the 75th percentile for GERP of all identifiable LDL-raising pathogenic SNVs in *LDLR* (MIM 606945)[55,56,58,59] (data not shown). We assumed that individuals missing genotypes at these rare SNVs were noncarriers. We performed this two-sided whole-gene test via a linear model for log(TG) on the {0, 1} genotype factor, using the package R with and without adjusting for race (first three principal components or ethnic group), age, and sex. Statin and fibrate usage is not contained in this data. However, because these medications reduce TG levels, their absence from the model serves to make the test more conservative.

Results

Phenotype

The adjTG is a stable phenotype over time in this data set. The correlation of adjTG values ~20 years apart, in the 21 individuals with data at both time points and not on statins, is highly significant ($r^2 = 0.78$, $p = 2.7 \times 10^{-5}$) with positive slope coefficient ($\beta = 0.5$). adjTG level appears to have larger variance at higher levels, as expected. The adjusted values were independent of age and sex, as expected.

Known TG SNVs

There were 18 SNVs, or their proxies, on the CVD chip of the total 35 (Table 1) that were reported to have an effect on TG.[21,52] These 18 SNVs and the *APOA5* SNV, rs3135506, explained at most 7% Vtot in log-adjTG in this family (Table 1). In addition, there was no association with the cumulative genetic risk score[21] and log-adjTG ($p = 0.20$). rs4810479, in the 5′ region of *PLTP* (MIM 172425), explained the most Vtot in log-adjTG, 7% (SD = 5%), but was not significantly associated with log-adjTG ($p = 0.56$), adjusting for pedigree structure. The *APOA5* SNV, rs3135506, explained 2.1% Vtot (SD = 2%). SNV rs17145738 in *TBL2* ([MIM 605842] alias *MLXIPL*) and near one of our linkage signals (see below), explained 5% Vtot in log-adjTG (SD = 4%) and had the lowest p value for association, $p = 0.04$, adjusting for pedigree structure. This result is not significant given the Bonferroni adjustment for multiple testing (Bonferonni cutoff of 0.003). Furthermore, the minor allele is associated with lower log-adjTG, indicating that this common SNV is unlikely to explain the high TG observed in this family. Because these known SNVs did not explain much of the Vtot in log-adjTG and were not significantly associated with log-adjTG, we performed linkage analysis to identify regions of interest.

Joint Segregation and Linkage Analysis Results

Segregation analysis, both alone and in conjunction with linkage analysis, shows evidence of at least one QTL underlying log-adjTG. Given the prior mean additive genetic effect = 2.7 times the SD of log-adjTG, the posterior probability of at least one QTL is 0.97, 0.88, and 0.998 for data

Figure 1. Evidence of Linkage, as Measured by the Bayes Factor, on Chromosomes 7 and 17 for the Three Data Sets 1980only, 1980plus, and 2000plus
Location is given in cM and spans the entire length of the chromosomes. The horizontal dashed line at 25 indicates the cutoff for evidence for linkage.

sets 1980only, 1980plus, and 2000plus. The posterior probability of exactly one QTL is 0.83, 0.54, and 0.23 for data sets 1980only, 1980plus, and 2000plus. In addition, the three overlapping data sets show support for log-adjTG linkage to chr7 and chr17 (Figure 1). Data sets 1980only and 1980plus support linkage to both chromosomes. Data set 2000plus shows support for only chr17. Using differing prior distributions on the genetic effects did not change the locations of the linkage signals (data not shown). Because data set 1980plus shows support for both regions and contains more phenotype data than does 1980only, we continued the analysis with this data set. The posterior predicted model for 1980plus is a recessive mode of inheritance.

Novel SNVs

Within and near these linkage regions, we found five novel SNVs that met our criteria for candidacy, labeled N7, N17a, N17b, N17c, and N17d, that were private to this family (Tables 2 and 3). Variant N7 is shared by all four IDs with TG > 1,000. Although chr17 contained four SNVs of interest, they were shared by three different pairs of IDs with TG > 1,000: N17a at one pair, N17b at another, and N17c and N17d at a different pair. These five SNVs explain 19%, 7.8%, 4.8%, 5.3%, and 5.3% of Vtot, respectively, in data set 1980plus. It is expected that the estimated %Vtot explained will decrease upon full genotyping in the family, because the most informative people have measured or imputable genotypes and the branch of the family without variability of the SNV of interest will lower the Vtot explained. Therefore, we chose to pursue full genotyping of N7 and N17a only because they explain >7% of Vtot with just the exome genotypes, which is more than the known SNVs in *PLTP*, *APOA5*, and *TBL2* noted above. These SNVs were in *SLC25A40* (MIM 610821) and *PLD2* (MIM 602384), respectively. Additionally, no known pathogenic or novel likely pathogenic variation was found in the known TG genes *LPL* (MIM 09708), *APOC2* (MIM 608083), *APOA5*, *GPIHBP1* (MIM 612757), and *LMF1* (MIM 611761) or in the lipid-related genes *CETP* (MIM 118470) and *LDLR* in the 16 exomes.

Jointly analyzing the chromosomes and their associated novel SNVs, genotyped in the full family, clarifies the relationship among these two regions (Figures 2 and S1). As expected, given the complex model, adjustment for each SNV impacts the linkage evidence for the other region.

Table 3. Percent Variance of log-adjTG Explained in the Family, by Candidate Novel SNVs

Variant	N7	N17a	N17b	N17c	N17d
Chr	7	17	17	17	17
Pos.	87477251	4711152	6021379	2290581	4496467
Gene	*SLC25A40*	*PLD2*	*WSCD1*	*MNT*	*SMTNL2*
RefSeq	NM_018843.3	NM_002663.4	NM_015253.1	NM_020310.2	NM_001114974.1; NM_198501.2
NT change	c.374A>G	c.85A>T	c.1246A>G	c.1363G>C	c.730+1G>T; c.298+1G>T
Protein change	p.Tyr125Cys	p.Thr29Ser	p.Arg416Gly	p.Val455Leu	NA
Coding effect	missense	missense	missense	missense	splice
GERP	5.1	5.1	4.3	3.9	5.0
%Vtot as exome variant	19	7.8	4.8	5.3	5.3
%Vg as exome variant	49	23	22	27	27
maxBF	1.8	48	103	72	66

Estimated percent total variance (Vtot) and genetic variance (Vg) in log-adjTG explained by each of the candidate novel SNVs within the linkage regions, using only individuals that have exome data. MaxBF are given for when the novel variant is included as a marker and covariate in the joint linkage and association analysis. Missing genotype data are imputed for other family members. Physical positions are from build 37. In dbSNP137, rsIDs have been assigned to SNVs N17b, N17c, and N17d as follows: N17b = rs200724890, N17c = rs201365025, N17d = rs202160684. There are two RefSeq numbers for *SMTNL2* because it has two isoforms.

Figure 2. Evidence of Linkage in Data Set 1980plus, as Measured by the Bayes Factor, on Chromosomes 7 and 17 when Analyzed Jointly, and with either SNV N7 or N17 as a Genotype Covariate
Location is given in cM and spans the entire length of the chromosomes. The horizontal dashed line at 25 indicates the original cutoff for evidence for linkage.

When both chromosomes are analyzed jointly, the linkage evidence on chr7 was reduced (maxBF = 8), but the linkage to chr17 remained (maxBF = 207). Linkage results are similar when including the fully genotyped SNVs N7 and N17a as markers only, in the analysis, indicating that when neither of the SNVs is included as a genotype covariate, the posterior distribution favors chr17 over chr7. However, only SNV N7 explained all evidence for linkage (chr17 maxBF reduced from 207 to 1, and chr7 maxBF reduced from 8 to 0.5) when included as a genotype covariate (Figure 2). When included as a genotype covariate, N17a explained some of the signal on chr17 (maxBF decreases from 207 to 69) and chr7 (maxBF reduces from 8 to 3.5).

The relative strength of each candidate SNV is reflected in the percent Vg explained by each, as well as the significance of the pedigree-adjusted association with log-adjTG. N7 (Ngt = 40) explained 19% of log-adjTG Vtot and 49% of Vg. N17a (Ngt = 78) explained 7% of log-adjTG Vtot and 25% Vg. When both variants are included jointly, they explained 49% of log-adjTG Vg; with N7 and N17a explaining 40% ($p = 0.0001$) and 8.6% ($p = 0.31$) of log-adjTG Vg, respectively. We also explored the significance of the known TG SNVs while adjusting for N7 genotype. In this case, each SNV explained <5% of Vtot and were not statistically significant (unadjusted $p > 0.08$). The *TBL2* SNV remained insignificant (unadjusted $p = 0.08$). Furthermore, this SNV is carried by individuals with log-adjTG below the mean, and these are not the same individuals as those that carry the N7 variant. Although N7 is associated with log-adjTG, it is not associated with LDL-C in this family ($p = 0.83$, Figure S2).

Table 4. Demographic Count Data and Age Distribution for the ESP Cohort, Broken down by Quantile of log(TG)

Log(TG) quantile	M	F	EA	AA	Mean (Min, Max) Age
[1.77,4.36]	341	597	434	504	50.15 (12, 85)
(4.36,4.74]	298	666	461	503	57.60 (18, 84)
(4.74,5.13]	312	596	522	386	58.27 (20, 93)
(5.13,6.84]	343	589	664	268	59.69 (21, 90)

This data set contains 3,770 verified unrelated individuals. Abbreviations are as follows: M, male; F, female; EA, European American; AA, African American.

Confirmation of *SLC25A40* Association

SLC25A40 was significantly associated with log(TG) in a separate cohort by using ESP data (Table 4). There were five rare missense SNVs (Table 5) in *SLC25A40* with GERP > 4.8 identified in ESP subjects with TG data. Notably, the *SLC25A40* SNV in the linkage family had a GERP of 5.1 and each of these five SNVs had GERPs > 5.1; two were predicted to be "probably damaging" by PolyPhen.[60] Six subjects of European American (EA) descent and two of African American (AA) descent carried one of these variants. The two of AA descent carried the same SNV; this SNV did not vary in EA in the entire ESP data. The SNVs found in the EA-descent subjects were also specific to that ancestry group. No subject carried more than one of these *SLC25A40* SNVs. After pooling carrier status for each of these SNVs into a single factor, the whole-gene test gives a significant positive association between these high GERP variants at *SLC25A40* and log(TG) ($\beta = 0.45$, two-sided $p = 0.02$), adjusted for age and sex. This whole-gene association with log(TG) remains significant and positive when adjusted for race, age, and sex ($\beta = 0.42$, two-sided $p = 0.03$). Results are similar when applied to individuals older than 25 and when removing individuals with a rare SNV in the gene, whose GERP score ranges between 2 and 4.8. Although the sample size is too small for stratified testing, the coefficient is positive in both ancestry groups ($\beta = 0.18$ and 1.1 for EA and AA descent, respectively). Although we focused on *SLC25A40* as the best candidate from this region, we sought to rule out associations of other regional genes and log(TG) in the ESP cohort by similar analyses (Table S2). Carrier status at *PLD2* ($p = 0.48$), *WSCD1* ($p = 0.19$), *MNT* ($p = 0.81$ [MIM 603039]), and *SMTNL2* ($p = 0.87$) did not predict log(TG).

Discussion

We identified two genes, *SLC25A40* and *PLD2*, each containing a novel variant that cosegregates with severely high TG levels in a single large family ascertained for segregation of FCHL. Joint analysis of the linkage regions and these variants indicated that the variant in *SLC25A40* is the most probable variant for pathogenicity. However, we

Table 5. Rare SNVs with GERP > 4.8 in *SLC25A40* Found in Individuals in the ESP Data

Position	rsID	NT Change	MAF % (EA/AA/All)	AA Change	GERP	PolyPhen	n
87470986	NA	c.785A>G	0.0118/0.0/0.0078	p.Gln262Arg	5.92	B	1 EA
87483601	NA	c.182G>T	0.0/0.0454/0.0154	p.Gly61Val	5.91	PD	2 AA
87473169	rs145515966	c.641G>C	0.0117/0.0/0.0077	p.Trp214Ser	5.67	PD	1 EA
87473143	rs140104130	c.667T>A	0.0349/0.0/0.0231	p.Trp223Arg	5.67	B	3 EA
87477276	NA	c.349G>A	0.0116/0.0/0.0077	p.Ala117Thr	5.13	B	1 EA

MAF % based on 8,600 European (EA) and 4,406 African American (AA), self-reported ancestry subjects; TG were measured in 2,168 EA and 1,793 AA subjects, with ancestry genetically verified. Abbreviations are as follows: B, benign; PD, probably damaging.

cannot rule out a possible interaction or additive effect between these and other identified variants. No known common pathogenic variants were found to be associated with TG in this family. We were able to replicate the effects of rare, conserved missense mutations only in *SLC25A40* in the separate ESP cohort.

SLC25A40 is solute carrier family 25, member 40, and its product localizes to the mitochondria, where it is involved in membrane transport.[61] Although the gene is ubiquitously expressed, its product is found mainly in the adrenal gland.[62] The novel variant in this gene (RefSeq accession number NM_018843.3), c.374A>G, is located in the seventh exon (of 12), is highly conserved (GERP = 5.1), and causes a p.Tyr125Cys substitution in the 338 amino acid protein. This change is just outside the second helical transmembrane region of the protein. Although there is no PolyPhen prediction, the tyrosine at amino acid 125 is perfectly conserved across vertebrates.[58,63] Furthermore, the Grantham score is 194, indicating a high chemical dissimilarity from the wild-type.[64] This particular mutation causes a new cysteine that neighbors a cysteine at position 124 and increases the number of cysteines to 9. Cysteines play a critical role in protein folding because they are the only amino acids that can create disulfide bonds that stabilize the folded form of the protein.[65] This change of a single amino acid to a cysteine allows for an increased number of folding possibilities, and possibly a lower entropy stabilization, which can disrupt the protein's function.

It is possible that this specific variant is not pathogenic, but is in linkage disequilibrium with a causal variant not detected by our exome sequencing. Further gene sequencing, as well as functional studies, are needed to verify the association. However, we found evidence for a pathogenic effect of high TG for rare, conserved (GERP > 5) missense variants in *SLC25A40* in a separate, unrelated cohort, supporting the idea that *SLC25A40* affects TG levels.

Although our analyses suggest that a *PLD2* missense variant may also impact TG, whole-gene testing in a separate cohort does not support pathogenicity for the gene. Further study may be warranted, because this gene may have biological relevance to TG. *PLD2* is phospholipase D2, whose product catalyzes the hydrolysis of phosphatidylcholine to phosphatidic acid and choline, and is ubiquitously expressed. The novel variant in this gene (RefSeq NM_002663.4), c.85A>T, causes a p.Thr29Ser (out of 934 amino acids) and is highly conserved across mammals.

Three other candidate genes identified on chr17, *WSCD1*, *MNT*, and *SMTNL2*, also were not significantly associated with TG in this separate cohort. Of these three genes, only *SMNTL2* is expressed in liver. *SMNTL2* is smoothelin-like 2, of which very little is currently known. Because neither *WSCD1* nor *MNT* are expressed in liver, they are less likely to be involved in TG regulation. Little is known about *WSCD1*, the WSC domain containing 1 gene. *MNT* encodes the MAX dimerization protein and is thought to repress transcription by binding to DNA binding proteins.[66,67] Reliance on the exome-sequence data rather than full-genome sequencing may have missed other potentially causal variants.

It is worth noting that the size of the linkage signals do not necessarily correlate with the candidate gene with the most evidence for causality. When chr7 and chr17 are analyzed jointly, the posterior distribution prefers a recessive QTL with moderate allele frequency on chr17. This is most probably due to the ascertainment scheme, lack of prior information on the allele frequency, and the observed phenotype segregation. The ascertainment scheme, which is useful for observing a rare allele multiple times, artificially inflates the observed allele frequency. In addition, the uniform prior distribution on the allele frequencies does not allow for a preference for rare alleles. Finally, the skew of the phenotype data in which offspring may have much higher phenotype than their parents results in evidence for a recessive trait. Therefore, segregation analysis showed overwhelming support for a recessive gene with a common allele frequency, which fit the marker and trait segregation at chr17. Further analysis including the exome-sequence data as covariate effects, which included a fixed rare allele frequency in the model, favored a dominant gene as evidenced by the high association of the SNV on chr7 with TG in the family.

The sampling and analysis design used here is a powerful approach for successful identification of novel genes and biological pathways underlying heterogeneous complex traits. Ascertainment of a large family segregating extreme values of a quantitative phenotype reduces the number of

putative underlying highly penetrant loci, increasing the chances of finding a variant with biological relevance in multiple family members. The large size of the pedigree, although a fraction of the size of a typical GWAS sample, provides multiple dimensions of information that, when used jointly, substantially increases the power to detect linkage.[31] Furthermore, correction for the ascertainment of these extreme phenotypes, as in this study, further increases power to detect linkage. The use of linkage analysis further drastically reduces the search space, limiting the multiple testing problem seen with GWASs or exome-wide burden tests. Finally, joint linkage and association in such pedigrees, as when both chromosomes 7 and 17 compete for the modeled QTL and the candidate novel variants are included as covariates, can help determine which of the identified candidate genes has the most promise. One limitation of this approach is that the same rare variant may not be found in a second family or enough unrelated individuals to provide confirmation; however, a single or limited number of gene-based burden tests can be made in an unrelated sample, again avoiding the penalty of exome-wide multiple comparisons.

Although common alleles have been found to explain a small portion of moderate to severe HTG in unrelated samples, we show that rare variation plays a major role in explaining severe HTG in a family previously diagnosed with FCHL and that this knowledge can be used to identify novel genes and biological pathways of interest. Additionally, our linkage results and the replication of a HTG effect in the ESP data, which is not ascertained on FCHL, lend support to an oligogenic inheritance of the lipid traits contributing to the diagnosis of FCHL in this family. Further evidence for this includes the fact that LDL-C level is not associated with the *SLC25A40* variant in these data, the proband had normal LDL-C and apoB levels, we do not detect linkage between LDL-C and this region of chr7 (data not shown) by similar methodology, and we have previously reported linkage between apoB level and chr4 in this family.[42] FCHL families may not have a monogenic disorder as earlier described, but rather a confluence of separate traits in the same family. Furthermore, the complex nature of the trait within this family is borne out by the fact that three carriers of the SLC25A40 variant do not have HTG, although two of them have TG near the 95th percentile (Figure S2), and that one individual with TG $>$ 500 does not carry the SLC25A40 variant, indicating that other genetic variants may be influential. However, the polygenic effects are accounted for in the analysis and have a small impact on the trait, relative to the major gene component. Indeed, *SLC25A40* may be shown to cause familial hypertriglyceridemia (MIM 145750), given that it does not appear to raise LDL-C. We note that our proband had highly elevated TG for a subject with FCHL, which is why the proband was thought to potentially be a FCHL homozygote; thus, the relationship of this locus to a more typical FCHL family with more modest HTG is not yet clear. Neither *SLC25A40* nor *PLD2* (or their respective biological pathways) have been previously implicated in triglyceride levels, to our knowledge. Functional studies, in vitro and in vivo, will need to be carried out to verify that they impact TG levels and to discover the mechanism by which they have an effect. Given the relevance of high TG to CVD, TG-lowering treatments targeted to these pathways may be identified.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, two figures, and two tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

Thanks go to Peter Byers for his helpful comments. Funding for this analysis was provided by National Institutes of Health grants P01 HL030086, T32 GM007454, and R01 HL094976 and the State of Washington Life Sciences Discovery Fund award to the Northwest Institute of Genetic Medicine (grant 265508). The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). Genotyping services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute.

Received: July 25, 2013
Revised: September 12, 2013
Accepted: October 21, 2013
Published: November 21, 2013

Web Resources

The URLs for data presented herein are as follows:

NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
R statistical software, http://www.r-project.org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

References

1. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) (2002). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation *106*, 3143–3421.

2. Hokanson, J.E., and Austin, M.A. (1996). Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies. J. Cardiovasc. Risk *3*, 213–219.
3. Patel, A., Barzi, F., Jamrozik, K., Lam, T.H., Ueshima, H., Whitlock, G., and Woodward, M.; Asia Pacific Cohort Studies Collaboration (2004). Serum triglycerides as a risk factor for cardiovascular diseases in the Asia-Pacific region. Circulation *110*, 2678–2686.
4. Tirosh, A., Rudich, A., Shochat, T., Tekes-Manova, D., Israeli, E., Henkin, Y., Kochba, I., and Shai, I. (2007). Changes in triglyceride levels and risk for coronary heart disease in young men. Ann. Intern. Med. *147*, 377–385.
5. Miller, M., Cannon, C.P., Murphy, S.A., Qin, J., Ray, K.K., and Braunwald, E.; PROVE IT-TIMI 22 Investigators (2008). Impact of triglyceride levels beyond low-density lipoprotein cholesterol after acute coronary syndrome in the PROVE IT-TIMI 22 trial. J. Am. Coll. Cardiol. *51*, 724–730.
6. Kohli, P., and Cannon, C.P. (2012). Triglycerides: how much credit do they deserve? Med. Clin. North Am. *96*, 39–55.
7. Miller, M., Stone, N.J., Ballantyne, C., Bittner, V., Criqui, M.H., Ginsberg, H.N., Goldberg, A.C., Howard, W.J., Jacobson, M.S., Kris-Etherton, P.M., et al.; American Heart Association Clinical Lipidology, Thrombosis, and Prevention Committee of the Council on Nutrition, Physical Activity, and Metabolism; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Nursing; Council on the Kidney in Cardiovascular Disease (2011). Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation *123*, 2292–2333.
8. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet *380*, 572–580.
9. Schwartz, G.G., Olsson, A.G., Abt, M., Ballantyne, C.M., Barter, P.J., Brumm, J., Chaitman, B.R., Holme, I.M., Kallend, D., Leiter, L.A., et al.; dal-OUTCOMES Investigators (2012). Effects of dalcetrapib in patients with a recent acute coronary syndrome. N. Engl. J. Med. *367*, 2089–2099.
10. Sarwar, N., Sandhu, M.S., Ricketts, S.L., Butterworth, A.S., Di Angelantonio, E., Boekholdt, S.M., Ouwehand, W., Watkins, H., Samani, N.J., Saleheen, D., et al.; Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration (2010). Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. Lancet *375*, 1634–1639.
11. Lee, M., Saver, J.L., Towfighi, A., Chow, J., and Ovbiagele, B. (2011). Efficacy of fibrates for cardiovascular risk reduction in persons with atherogenic dyslipidemia: a meta-analysis. Atherosclerosis *217*, 492–498.
12. Bruckert, E., Labreuche, J., Deplanque, D., Touboul, P.J., and Amarenco, P. (2011). Fibrates effect on cardiovascular risk is greater in patients with high triglyceride levels or atherogenic dyslipidemia profile: a systematic review and meta-analysis. J. Cardiovasc. Pharmacol. *57*, 267–272.
13. Wierzbicki, A.S., Hardman, T.C., and Viljoen, A. (2012). New lipid-lowering drugs: an update. Int. J. Clin. Pract. *66*, 270–280.
14. Pennacchio, L.A., Olivier, M., Hubacek, J.A., Krauss, R.M., Rubin, E.M., and Cohen, J.C. (2002). Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. Hum. Mol. Genet. *11*, 3031–3038.
15. Talmud, P.J., Hawe, E., Martin, S., Olivier, M., Miller, G.J., Rubin, E.M., Pennacchio, L.A., and Humphries, S.E. (2002). Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. Hum. Mol. Genet. *11*, 3039–3046.
16. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. *39*, 513–516.
17. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. *40*, 189–197.
18. Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gómez Pérez, F.J., Frazer, K.A., Elliott, P., Scott, J., et al. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat. Genet. *40*, 149–151.
19. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
20. Nilsson, S.K., Heeren, J., Olivecrona, G., and Merkel, M. (2011). Apolipoprotein A-V; a potent triglyceride reducer. Atherosclerosis *219*, 15–21.
21. Johansen, C.T., Kathiresan, S., and Hegele, R.A. (2011). Genetic determinants of plasma triglycerides. J. Lipid Res. *52*, 189–206.
22. Baroukh, N., Bauge, E., Akiyama, J., Chang, J., Afzal, V., Fruchart, J.C., Rubin, E.M., Fruchart-Najib, J., and Pennacchio, L.A. (2004). Analysis of apolipoprotein A5, c3, and plasma triglyceride concentrations in genetically engineered mice. Arterioscler. Thromb. Vasc. Biol. *24*, 1297–1302.
23. Ito, Y., Azrolan, N., O'Connell, A., Walsh, A., and Breslow, J.L. (1990). Hypertriglyceridemia as a result of human apo CIII gene expression in transgenic mice. Science *249*, 790–793.
24. Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science *294*, 169–173.
25. Maeda, N., Li, H., Lee, D., Oliver, P., Quarfordt, S.H., and Osada, J. (1994). Targeted disruption of the apolipoprotein C-III gene in mice results in hypotriglyceridemia and protection from postprandial hypertriglyceridemia. J. Biol. Chem. *269*, 23610–23616.
26. Russo, G.T., Meigs, J.B., Cupples, L.A., Demissie, S., Otvos, J.D., Wilson, P.W., Lahoz, C., Cucinotta, D., Couture, P., Mallory, T., et al. (2001). Association of the Sst-I polymorphism at the APOC3 gene locus with variations in lipid levels, lipoprotein subclass profiles and coronary heart disease risk: the Framingham offspring study. Atherosclerosis *158*, 173–181.
27. Olivieri, O., Bassi, A., Stranieri, C., Trabetti, E., Martinelli, N., Pizzolo, F., Girelli, D., Friso, S., Pignatti, P.F., and Corrocher, R. (2003). Apolipoprotein C-III, metabolic syndrome, and risk of coronary artery disease. J. Lipid Res. *44*, 2374–2381.

28. Vaessen, S.F., Schaap, F.G., Kuivenhoven, J.A., Groen, A.K., Hutten, B.A., Boekholdt, S.M., Hattori, H., Sandhu, M.S., Bingham, S.A., Luben, R., et al. (2006). Apolipoprotein A-V, triglycerides and risk of coronary artery disease: the prospective Epic-Norfolk Population Study. J. Lipid Res. *47*, 2064–2070.
29. Bi, N., Yan, S.K., Li, G.P., Yin, Z.N., and Chen, B.S. (2004). A single nucleotide polymorphism -1131T>C in the apolipoprotein A5 gene is associated with an increased risk of coronary artery disease and alters triglyceride metabolism in Chinese. Mol. Genet. Metab. *83*, 280–286.
30. Dallongeville, J., Cottel, D., Montaye, M., Codron, V., Amouyel, P., and Helbecque, N. (2006). Impact of APOA5/A4/C3 genetic polymorphisms on lipid variables and cardiovascular disease risk in French men. Int. J. Cardiol. *106*, 152–156.
31. Wijsman, E.M. (2012). The role of large pedigrees in an era of high-throughput sequencing. Hum. Genet. *131*, 1555–1563.
32. Jarvik, G.P., Brunzell, J.D., Austin, M.A., Krauss, R.M., Motulsky, A.G., and Wijsman, E. (1994). Genetic predictors of FCHL in four large pedigrees. Influence of ApoB level major locus predicted genotype and LDL subclass phenotype. Arterioscler. Thromb. *14*, 1687–1694.
33. Goldstein, J.L., Schrott, H.G., Hazzard, W.R., Bierman, E.L., and Motulsky, A.G. (1973). Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia. J. Clin. Invest. *52*, 1544–1568.
34. Goldstein, J.L., Dana, S.E., Brunschede, G.Y., and Brown, M.S. (1975). Genetic heterogeneity in familial hypercholesterolemia: evidence for two different mutations affecting functions of low-density lipoprotein receptor. Proc. Natl. Acad. Sci. USA *72*, 1092–1096.
35. Breslow, J.L. (2000). Genetics of lipoprotein abnormalities associated with coronary artery disease susceptibility. Annu. Rev. Genet. *34*, 233–254.
36. Ford, E.S., Li, C., Zhao, G., Pearson, W.S., and Mokdad, A.H. (2009). Hypertriglyceridemia and its pharmacologic treatment among US adults. Arch. Intern. Med. *169*, 572–578.
37. Austin, M.A., McKnight, B., Edwards, K.L., Bradley, C.M., McNeely, M.J., Psaty, B.M., Brunzell, J.D., and Motulsky, A.G. (2000). Cardiovascular disease mortality in familial forms of hypertriglyceridemia: A 20-year prospective study. Circulation *101*, 2777–2782.
38. Chait, A., and Brunzell, J.D. (1992). Chylomicronemia syndrome. Adv. Intern. Med. *37*, 249–273.
39. Warnick, G.R. (1986). Enzymatic methods for quantification of lipoprotein lipids. Methods Enzymol. *129*, 101–123.
40. US Dept of Health and Human Services (1980). The Lipid Research Clinics' Population Studies Data Book, *Volume 1* (Washington, DC: National Institutes of Health).
41. Rosenthal, E.A., Ronald, J., Rothstein, J., Rajagopalan, R., Ranchalis, J., Wolfbauer, G., Albers, J.J., Brunzell, J.D., Motulsky, A.G., Rieder, M.J., et al. (2011). Linkage and association of phospholipid transfer protein activity to *LASS4*. J. Lipid Res. *52*, 1837–1846.
42. Wijsman, E.M., Rothstein, J.H., Igo, R.P., Jr., Brunzell, J.D., Motulsky, A.G., and Jarvik, G.P. (2010). Linkage and association analyses identify a candidate region for apoB level on chromosome 4q32.3 in FCHL families. Hum. Genet. *127*, 705–719.
43. Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., Price, T.S., Glessner, J.T., Galver, L., Barrett, J.C., Grant, S.F.A., Farlow, D.N., et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS ONE *3*, e3583.
44. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
45. De La Vega, F.M., Dailey, D., Ziegle, J., Williams, J., Madden, D., and Gilbert, D.A. (2002). New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. Biotechniques (*Suppl*), 48–50, 52, 54.
46. Lin, C.H., Yeakley, J.M., McDaniel, T.K., and Shen, R. (2009). Medium- to high-throughput SNP genotyping using VeraCode microbeads. Methods Mol. Biol. *496*, 129–142.
47. Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. *61*, 748–760.
48. Gagnon, F., Jarvik, G.P., Badzioch, M.D., Motulsky, A.G., Brunzell, J.D., and Wijsman, E.M. (2005). Genome scan for quantitative trait loci influencing HDL levels: evidence for multilocus inheritance in familial combined hyperlipidemia. Hum. Genet. *117*, 494–505.
49. Yu, D. (2003). Testing the robustness of Markov chain Monte Carlo segregation and linkage analysis when normality assumptions are violated. PhD thesis, University of Washington, Seattle, WA.
50. Wijsman, E.M., Daw, E.W., Yu, C.E., Payami, H., Steinbart, E.J., Nochlin, D., Conlon, E.M., Bird, T.D., and Schellenberg, G.D. (2004). Evidence for a novel late-onset Alzheimer disease locus on chromosome 19p13.2. Am. J. Hum. Genet. *75*, 398–409.
51. Gagnon, F., Jarvik, G.P., Motulsky, A.G., Deeb, S.S., Brunzell, J.D., and Wijsman, E.M. (2003). Evidence of linkage of HDL level variation to *APOC3* in two samples with different ascertainment. Hum. Genet. *113*, 522–533.
52. Goldberg, I.J., Eckel, R.H., and McPherson, R. (2011). Triglycerides and heart disease: still a hypothesis? Arterioscler. Thromb. Vasc. Biol. *31*, 1716–1725.
53. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and de Bakker, P.I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics *24*, 2938–2939.
54. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
55. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
56. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
57. Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. *62*, 1198–1211.
58. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

59. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.
60. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
61. Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. Cell *134*, 112–123.
62. Pontius, J.U., Wagner, L., and Schuler, G.D. (2003). UniGene: a unified view of the transcriptome. In The NCBI Handbook, J. McEntyre and J. Ostell, eds. (Bethesda, MD: National Center for Biotechnology Information).
63. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.
64. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862–864.
65. Kim, P.S., and Baldwin, R.L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. *51*, 459–489.
66. Blackwood, E.M., and Eisenman, R.N. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science *251*, 1211–1217.
67. Kretzner, L., Blackwood, E.M., and Eisenman, R.N. (1992). Myc and Max proteins possess distinct transcriptional activities. Nature *359*, 426–429.

SISG Module 16, 2014
Genetic Epidemiology
Suggested Readings:

General

- Series on Genetic Epidemiology in The Lancet http://www.thelancet.com/series/genetic-epidemiology
- Austin MA. Genetic Epidemiology: Methods and Applications . CABI; 1 edition (August 2013)ISBN-10: 1780641818

Pedigree Drawing

- Bennett RL, et al Recommendations for standardized pedigree nomenclature.AJHG 1995; 56; 745-52
- Bennett RL et al,. Standardized human pedigree nomenclature: Update and assessment of the recommendations of the National Society of genetic counselors. J Genetic Counsel (2008); 17:424-433

Family Based Studies

- Jarvik GP. Complex segregation analysis: uses and limitations. Am J. Hum. Genet. 1998; 63: 942-046.
- Bellis C, Cox HC, Dyer TD, Charlseworth JC, Begley KN, Quinlan S, Lea RA, Heath SC, Blangero J, Griffiths LR. Linkage mapping of CVD risk traits in the isolated Norfolk Island population. Hum. Genet. 2008; 124(5): 543-552.

Linkage Disequilibrium

- Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406
- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.
- Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9, 477-485 (June 2008)

GWAS

- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN,

Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53.

- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008 May;9(5):356-69.

GxE Interactions

- Khoury MJ and Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am J Epidemiol. 2009; 169: 227-30.
- Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet. 2010 Apr;11(4):259-72.
- Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM; NCI Gene-Environment Think Tank. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. Genet Epidemiol. 2013 Nov;37(7):643-57.

Rare Variant/Next Generation Sequencing Association Studies

- Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014 Jan 28;111(4):E455-64.
- Cirulli ET1, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415-25.

Functional Studies

- Ecker JR1, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. Genomics: ENCODE explained. Nature. 2012 Sep 6;489(7414):52-5.
- Maurano MT1, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190-5