Next Generation Statistical Methods for Genome-Wide Association Studies

Course Overview and Intro

06 September 2023

Agenda

- Course format
- Intro to genetic epidemiology
- Course scope
- A few basic definitions

Agenda

- Course format
- Intro to genetic epidemiology
- Course scope
- A few basic definitions

Home Sessions Instructors News

DCEG Statistical Genetics Workshop

CONTENTS

Next Generation Statistical Methods for Genome Wide Association Studies: A Hands-On Course

Background

Genome-wide association studies (GWAS) have revolutionized our understanding of the genetic basis of complex traits and diseases. In the early years of GWAS, data analysis primarily relied on relatively simplistic methods, such as running millions of univariate linear or logistic regressions, one for each genetic variant. However, as the sample sizes for some GWAS have become extremely large and various types of other genomic data have become widely available, analysis of such data has

Latest Posts

Statistical Genetics Workshop Announcement

DCEG Statistical Genetics Workshop schedule for fall 2023

Location: Rm 1106-A/B at the CRL Building, 9615 Medical Center Drive, Rockville, MD 20892/online

Time: 9:30-12:30 EST

Link: https://nih.zoomgov.com/j/1600232059?pwd=aW1NTmRCWXAwajZ0bFN0ZEtTQWhiUT09

Read more

Published: Jul 5, 2023

https://dceg-workshops.github.io/statistical_genetics_workshop/

Course format

The course will consist of nine sessions held from September to December of 2023. Sessions will be held on Wednesdays from 9:30 to 12:30 will include a lecture (1.25 hours, including Q&A) and a 1.5-hour practical tutorial. (See schedule below for specific dates.) Participants are expected to complete background reading before each session (estimated out-of-class time: < 2 hours) and hands-on exercises after each session (estimated out-of-class time: < 2 hours). The course will be hybrid with both in-person and online participants, and all lectures will be recorded and archived for future use. Practical tutorials will be in-person at the Shady Grove NCI campus.

Intended audience

Researchers and analysts with strong quantitative background who are involved or anticipate being involved in analysis of large-scale genome-wide genotyping data. Participants should have basic knowledge of epidemiologic study designs, genetic concepts and terminologies, and statistical methodologies (e.g., hypothesis testing, parameter estimation, regression models, Bayes probability), as well as familiarity with R and command-line interfaces.

By the end of the course, participants will have gained a deep understanding of advanced statistical methods and computational tools for analyzing GWAS data, and will be able to apply these methods to their own research. They will also be familiar with best practices for data management and sharing in GWAS, and will be able to produce reproducible and FAIR-compliant pipelines.

https://dceg-workshops.github.io/statistical_genetics_workshop/

Published: Jul 5, 2023 by Wendy Wong

DCEG Statistical Genetics Workshop schedule for fall 2023

Location: Rm 1106-A/B at the CRL Building, 9615 Medical Center Drive, Rockville, MD 20892/online

Time: 9:30-12:30 EST

Link: https://nih.zoomgov.com/j/1600232059?pwd=aW1NTmRCWXAwajZ0bFN0ZEtTQWhiUT09

- Session 1: Introduction 09/06/23
- Session 2: Basic GWAS analyses 09/20/23
- Session 3: Fine-mapping and colocalization 09/27/23
- Session 4: Heritability, functional enrichment, polygenic scores 10/25/23
- Session 5: Rare variants 11/08/23
- Session 6: Integrative analyses and Mendelian Randomization 11/15/23
- Session 7: GWAS, fine-mapping and PRS in diverse-genetic-ancestry and admixed samples 11/29/23
- Session 8: Genetic mosaicism and clonal hematopoiesis 12/06/23
- Session 9: Functional genomics 12/13/23

https://dceg-workshops.github.io/statistical_genetics_workshop/2023/07/05/save-the-dates.html

Instructors

Kevin Brown, Ph.D.

DCEG, NCI

Dr. Brown received a Ph.D. in Genetics from the George Washington University in Washington, D.C., in 2003. He conducted his postdoctoral training in the Laboratory of Dr. Jeffrey Trent at the Translational Genomic Research Institute (TGen) in Phoenix, Arizona. He subsequently went on to direct his own research program at TGen as an investigator from 2005 to 2010, and served as an adjunct professor in basic medical sciences at the Mayo Clinic Cancer Center, the University of Arizona College of Medicine, and Arizona State University from 2008 to 2010. His work at TGen involved the application of whole-genome familial linkage, candidate gene, and genome-wide association study (GWAS) approaches to identify genetic variants associated with melanoma susceptibility. In 2010, Dr. Brown joined the Laboratory of Translational Genomics (LTG) in the Division of Cancer Epidemiology and Genetics (DCEG) as a tenure-track investigator. He was awarded NIH scientific tenure and promoted to senior investigator in 2018. His research focuses on the genetic underpinnings of melanoma susceptibility.

Kevin Brown's profile

https://dceg-workshops.github.io/statistical_genetics_workshop/instructors

Home Sessions Instructors News

Introduction
Basic GWAS analyses

Se ntroduction

| SESSIONS

Introduction
Basic GWAS analyses

Overview | Lecture | Practical | + Supplemental

Topics Covered

- Why genetic epidemiology?
- Why GWAS and RVAS?
- What we have learned.
- What remains to be done.
- Overview of the course

Practical

- To introduce the Google Colaboratory platform.
- To introduce participants the importance of Quality Control in the analysis pipeline.
- To provide participants with the basic practical skills needed to perform QC of GWAS data using plink.

Recommended Readings/Videos

- Readings on genome-wide association study design and analysis
  - Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., & Lappalainen, T. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9
- Overview of findings from genome-wide association studies
  - Penney, K. L., Michailidou, K., Carere, D. A., Zhang, C., Pierce, B., Lindström, S., & Kraft, P. (2017). Genetic Epidemiology of Cancer. https://doi.org/10.1093/oso/9780190238667.003.0005
  - Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics, 101(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

https://dceg-workshops.github.io/statistical_genetics_workshop/sessions/

Session 1:Introduction

CONTENTS

Overview Lecture Practical Supplemental

This workshop offers an introduction to utilizing Jupyter notebooks through Google Colaboratory. Additionally, it serves as a quick guide to GWAS QC with PLINK. We'll guide you step by step in using PLINK and R to carry out various QC procedures for Genome-Wide Assoication Studies(GWAS).

Objectives

- To introduce the Google Colaboratory platform.
- To introduce participants the importance of Quality Control in the analysis pipeline.
- To provide participants with the basic practical skills needed to perform QC of GWAS data using plink.

Prerequisites

Before starting this workshop, we recommend that you have:

- A Google Account
- A basic understanding of genetics, genome-wide association studies (GWAS), and the role of quality control (QC) in GWAS.
- Familiarity with command-line tools, as we will use the command-line tool plink for the QC steps.

https://dceg-workshops.github.io/statistical_genetics_workshop/sessions/session_1/practical

Product Solutions Open Source Pricing Search or jump to... Sign in Sign up

DCEG-workshops / statgen_workshop_tutorial Public

Notifications Fork 0 Star 0

Code Issues Pull requests Discussions Actions Projects Security Insights

Welcome to statgen_workshop_tutorial Discussions!

Announcements · shukwong

is:open Sort by: Latest activity Label Filter: Open

Categories

- View all discussions
- Announcements
- General
- Ideas
- Polls
- Q&A
- Show and tell

Discussions

1 **Welcome to statgen_workshop_tutorial Discussions!**
shukwong announced 4 hours ago in Announcements 0

Keep the discussion going! Post questions, comments, resources to the workshop discussion page.

https://github.com/DCEG-workshops/statgen_workshop_tutorial/discussions

Questions?

Agenda

- Course format
- Intro to genetic epidemiology
- Course scope
- A few basic definitions

Epidemiology is the study of the distribution and determinants of health-related states in specified populations, and the application of this study to control health problems.

Genetic variation is one of the determinants of health-related states in populations, so…

Genetic epidemiology is the study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with social and environmental factors

How can genetic epidemiology improve human health?

- Genetic epidemiology has contributed to a deeper understanding of disease biology, leading to new therapeutics.
- Genetic epidemiology also has direct clinical and public health applications.

- Genetic epidemiology has contributed to a deeper understanding of disease biology, leading to new therapeutics.
- Genetic epidemiology also has direct clinical and public health applications.

HEALTH

Rare Gene Mutations Inspire New Heart Drugs

By GINA KOLATA MAY 24, 2017

HEALTH

Rare Mutation Ignites Race for Cholesterol Drug

Genetic Connections

By GINA KOLATA JULY 9, 2013

BUSINESS DAY

Aiming to Push Genomics Forward in New Study

By ANDREW POLLACK JAN. 13, 2014

Researchers identify a healthy individual with extremely low levels of LDL cholesterol.

Three siblings of this *proband* also had very low LDL levels.

DNA sequencing determined that these siblings all carried two copies of a rare loss-of-function variant in the *ANGPTL3* gene.

Drugs that mimic the effect of this mutation are more likely to be safe and effective.

https://www.nytimes.com/2017/05/24/health/heart-drugs-gene-mutations.html
http://www.nytimes.com/2013/07/10/health/rare-mutation-prompts-race-for-cholesterol-drug.html
https://www.nytimes.com/2014/01/13/business/aiming-to-push-genomics-forward-in-new-study.html

HEALTH

Rare Gene Mutations Inspire New Heart Drugs

By GINA KOLATA MAY 24, 2017

HEALTH

Rare Mutation Ignites Race for Cholesterol Drug

Genetic Connections
By GINA KOLATA JULY 9, 2013

BUSINESS DAY

Aiming to Push Genomics Forward in New Study

By ANDREW POLLACK JAN. 13, 2014

Researchers determine that genetic variation in *PCSK9* influences LDL levels.

They screen study subjects for new variants that have large effects and identify a healthy individual *homozygous* for an extremely rare loss-of-function variant with extremely low LDL levels.

Drugs that mimic the effect of this mutation are more likely to be safe and effective.

https://www.nytimes.com/2017/05/24/health/heart-drugs-gene-mutations.html
http://www.nytimes.com/2013/07/10/health/rare-mutation-prompts-race-for-cholesterol-drug.html
https://www.nytimes.com/2014/01/13/business/aiming-to-push-genomics-forward-in-new-study.html

RESEARCH ARTICLE | BIOLOGICAL SCIENCES

Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia

REPORT

Correction of Sickle Cell Disease in Adult Mice by Interference with Fetal Hemoglobin Silencing

Fetal hemoglobin (HbF) contributes to variability in the expression of hemoglobinopathies. A genome-wide association study identified a single nucleotide polymorphism near the *BCL11A* associated with HbF levels.

Silencing BCL11A prevents sickle cell disease in mice.

Trials of several approaches to silencing *BCL11A* are now underway.

Uda M(2008) *PNAS*
Xu J (2011) *Science*
Frangoul (2021) *NEJM*
https://www.nytimes.com/2020/01/11/health/sickle-cell-disease-cure.html

- Genetic epidemiology has contributed to a deeper understanding of disease biology, leading to new therapeutics.
- Genetic epidemiology also has direct clinical and public health applications.

Genetics in Public Health

- Predictive: Screening those with rare risk variants
  - *BRCA1*/*BRCA2*, Lynch Syndrome/HNPCC
- Diagnostic: screening newborns
  - Amer. College of Med. Genetics: screen for 29 conditions

Table 2
Newborn screening panel: core panel and secondary targets

MS/MS				
Acylcarnitines		Amino acids		
9 OA	5 FAO	6 AA	3 Hb Pathies	6 Others
CORE PANEL				
IVA	MCAD	PKU	Hb SS*	CH
GA I	VLCAD	MSUD	Hb S/βTh*	BIOT
HMG	LCHAD	HCY*	Hb S/C*	CAH*
MCD	TFP	CIT		GALT
MUT*	CUD	ASA		HEAR
3MCC*		TYR I*		CF
Cbl A,B*				
PROP				
BKT				

List includes congenital hypothyroidism, phenylketonuria, hemoglobinopathies, maple syrup urine disease, cystic fibrosis.

Brice and Zimmern (2010) Khoury et al (ed); Watson (2006) *Genet Med*

Genetics in Public Health

Risk-stratified screening for breast cancer, combining data on questionnaire risk factors (QRF), mammographic density (MD), common risk variants (polygenic risk scores [PRS]), family history, and rare risk variants.

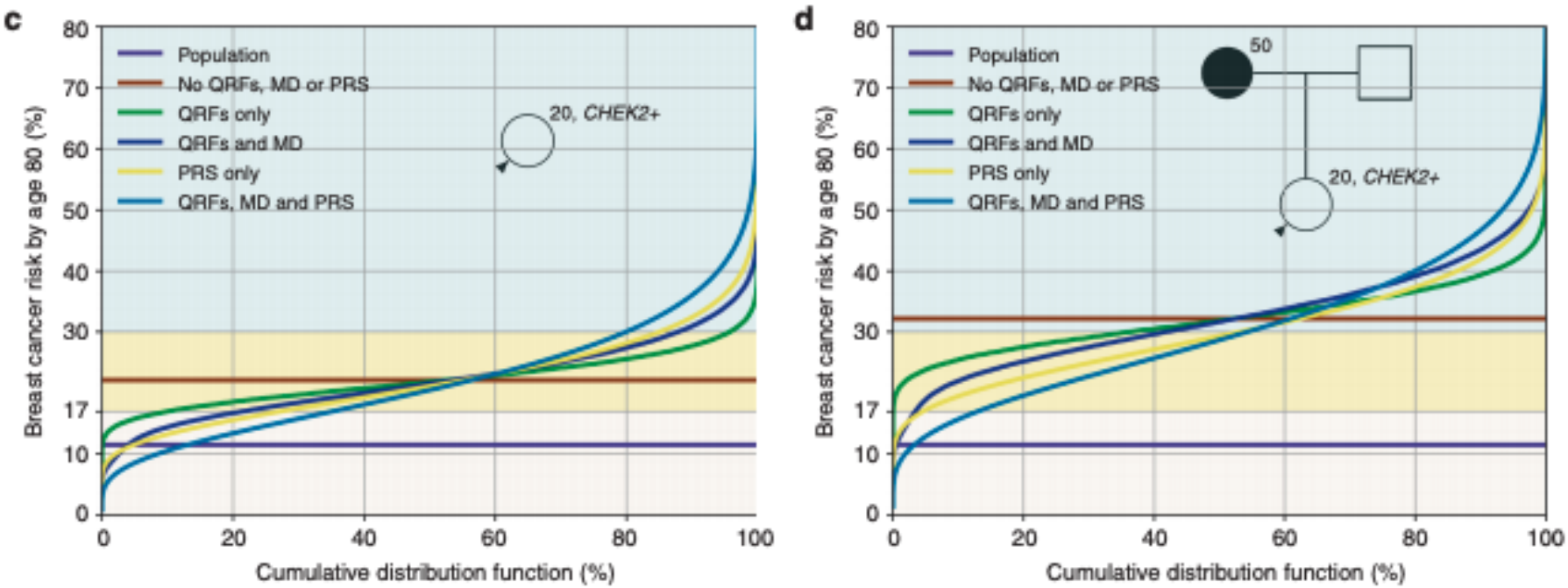


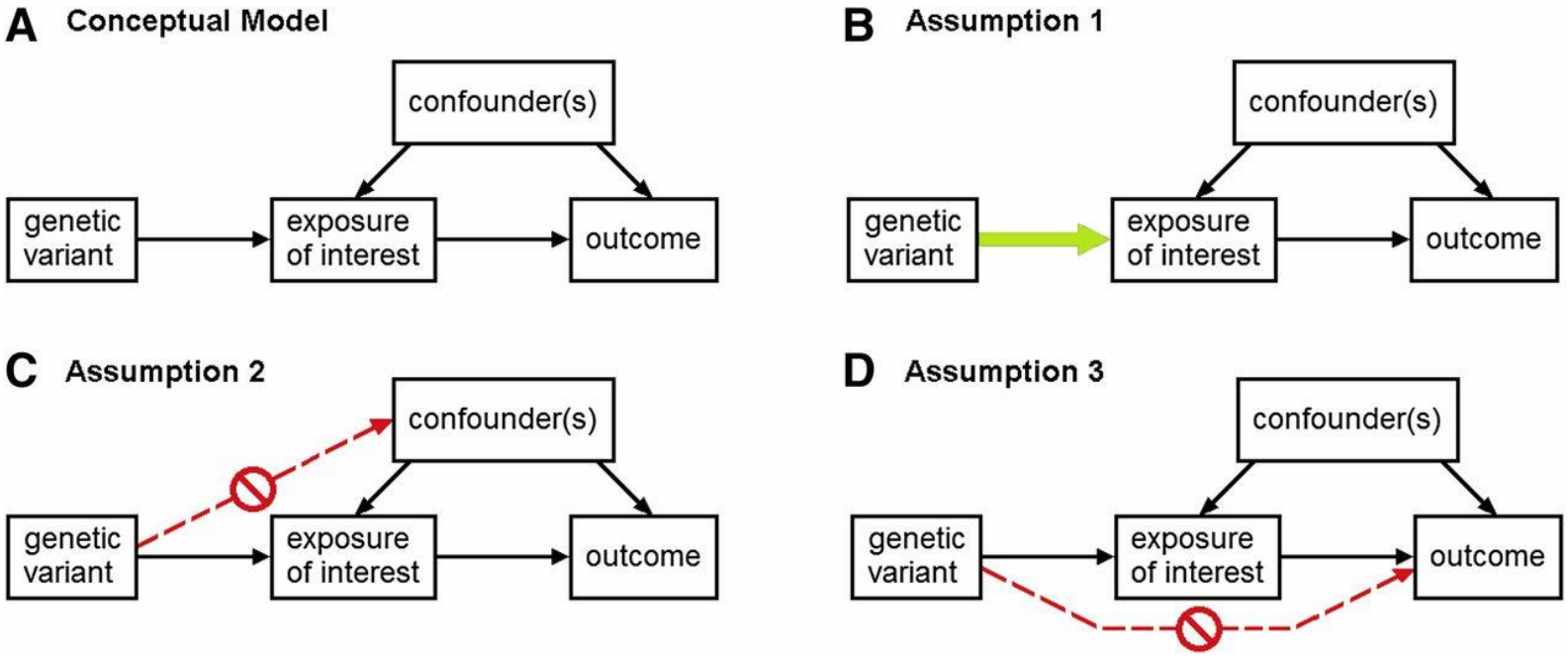
Fig. 3 Predicted Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) breast cancer risk for a female intermediate-risk rare pathogenic variant carrier, on the basis of the different predictors of risk (questionnaire-based risk factors [QRFs], mammographic density [MD], and polygenic risk scores [PRS]). (**a**, **c**) Lifetime risk (age 20 to 80 years) for a *CHEK2* 1100delC carrier with unknown family history; (**b**, **d**) lifetime risk for a *CHEK2* 1100delC carrier with her mother affected at age 50. (**e**, **f**) Risk for a *PALB2* and an *ATM* rare pathogenic variant carrier respectively, both with unknown family history. The backgrounds of the graphs are shaded to indicate the familial breast cancer risk categories based on the National Institute for Health and Care Excellence (NICE) guidelines:[3] (1) near-population risk shaded in pink (<17%), (2) moderate risk shaded in yellow (≥17% and <30%), and (3) high risk shaded in blue (≥30%). Predictions based on UK breast cancer incidence.

Lee A (2019) *Genet Med*

- Genetic epidemiology has contributed to a deeper understanding of disease biology, leading to new therapeutics.
- Genetic epidemiology also has direct clinical and public health applications.
- But wait, there’s more!

Mendelian Randomization

Use genetic variation as an instrumental variable to assess the causal relationship between an exposure and an outcome.



Sekula (2016) JASN
https://jasn.asnjournals.org/content/27/11/3253

Mendelian Randomization

ORIGINAL CONTRIBUTION

Genetic Loci Associated With C-Reactive Protein Levels and Risk of Coronary Heart Disease

Paul Elliott, FRCP
John C. Chambers, PhD
Weihua Zhang, PhD
Robert Clarke, MD
Jemma C. Hopewell, PhD
John F. Peden, PhD
Jeanette Erdmann, PhD
Peter Braund, MSc
James C. Engert, PhD
Derrick Bennett, PhD
Lachlan Coin, PhD
Deborah Ashby, PhD
Ioanna Tzoulaki, PhD
Ian J. Brown, PhD
Shahrul Mt-Isa, BSc
Mark I. McCarthy, FRCP
Leena Peltonen, MD, PhD
Nelson B. Freimer, MD
Martin Farrall, FRCPath
Aimo Ruokonen, MD, PhD
Anders Hamsten, MD
Noha Lim, PhD
Philippe Froguel, MD
Dawn M. Waterworth, PhD
Peter Vollenweider, MD
Gerard Waeber, MD
Marjo-Riitta Jarvelin, MD
Vincent Mooser, MD
James Scott, FRS
Alistair S. Hall, FRCP
Heribert Schunkert, MD
Sonia S. Anand, MD
Rory Collins, FRCP
Nilesh J. Samani, FRCP
Hugh Watkins, FRCP
Jaspal S. Kooner, FRCP

See also pp 49 and 92.

Context Plasma levels of C-reactive protein (CRP) are independently associated with risk of coronary heart disease, but whether CRP is causally associated with coronary heart disease or merely a marker of underlying atherosclerosis is uncertain.

Objective To investigate association of genetic loci with CRP levels and risk of coronary heart disease.

Design, Setting, and Participants We first carried out a genome-wide association (n=17 967) and replication study (n=13 615) to identify genetic loci associated with plasma CRP concentrations. Data collection took place between 1989 and 2008 and genotyping between 2003 and
study of the most closely associated
locus and published data on other
100 823 controls, to investigate th
disease. We compared our finding
vational studies of CRP levels and
associated with CRP levels, we sel
against coronary heart disease amo

Main Outcome Measure Risk

Results Polymorphisms in 5 gene
(% difference per minor allele): SN
interval [CI], −17.6% to −12.0%; P
CI, −14.4% to −8.5%; P=1.3×10
CI, −23.4% to −17.9%; P=1.3×
−16.6% to −10.9%; P=1.9×10
CI, −25.3% to −18.1%; P=8.1×10
cus with coronary heart disease ga
1.01) per 20% lower CRP level. Ou
CRP locus showed no association wi
to 1.02; per 20% lower CRP level,
predicted from meta-analysis of the
heart disease (z score, −3.45; P<.0
1.02 to 1.09; per minor allele), rs4537545 in *ICAM1* (OR, 0.94; 95% CI, 0.91 to 0.97), and rs4420638 in the *APOE-CI-CII* cluster (OR, 1.16; 95% CI, 1.12 to 1.21) were all associated with risk of coronary heart disease.

Conclusion The lack of concordance between the effect on coronary heart disease risk of *CRP* genotypes and CRP levels argues against a causal association of CRP with coronary heart disease.

JAMA. 2009;302(1):37-48 www.jama.com

Coronary heart disease (CHD) is the leading cause of death worldwide.[1] Inflammation plays a key role in the pathogenesis of CHD at every stage from initiation to progression and rupture of the atherosclerotic plaque.[2] C-reactive protein (CRP), an acute-phase protein synthesized primarily by the liver, is currently the most widely used biomarker of inflammation.[3] Observational studies have consistently demonstrated that higher plasma levels of CRP are associated with higher risk of CHD,[4,5] and measurement of CRP has been advocated as a means of improving risk prediction.[6] There is

Author Affiliations are listed at the end of this article.
Corresponding Author: Paul Elliott, FRCP, MRC-HPA Centre for Environment and Health, Department of Epidemiology and Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom (p.elliott@imperial.ac.uk).

Articles

Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study

Benjamin F Voight*, Gina M Peloso*, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, Heribert Schunkert, Nilesh J Samani, Robert Clarke, Jemma C Hopewell, John F Thompson, Mingyao Li, Gudmar Thorleifsson, Christopher Newton-Cheh, Kiran Musunuru, James P Pirruccello, Danish Saleheen, Li Chen, Alexandre F R Stewart, Arne Schillert, Unnur Thorsteinsdottir, Gudmundur Thorgeirsson, Sonia Anand, James C Engert, Thomas Morgan, John Spertus, Monika Stoll, Klaus Berger, Nicola Martinelli, Domenico Girelli, Pascal P McKeown, Christopher C Patterson, Stephen E Epstein, Joseph Devaney, Mary-Susan Burnett, Vincent Mooser, Samuli Ripatti, Ida Surakka, Markku S Nieminen, Juha Sinisalo, Marjo-Liisa Lokki, Markus Perola, Aki Havulinna, Ulf de Faire, Bruna Gigante, Erik Ingelsson, Tanja Zeller, Philipp Wild, Paul I W de Bakker, Olaf H Klungel, Anke-Hilze Maitland-van der Zee, Bas I M Peters, Anthonius de Boer, Diederick E Grobbee, Pieter W Kamphuisen, Vera H M Deneer, Clara C Elbers, [illegible] Wijmenga, W M Monique Verschuren, Jolanda M A Boer, Yvonne T van der Schouw, Asif Rasheed, [illegible], Ron Do, Jose M Ordovas, Gonçalo R Abecasis, Michael Boehnke, Karen L Mohlke, Mark J Daly, [illegible] González, Shaun Purcell, Stacey Gabriel, Jaume Marrugat, John Peden, Jeanette Erdmann, [illegible], Marcus Fischer, Christian Hengstenberg, Andreas Ziegler, Ian Buysschaert, Diether Lambrechts, [illegible] Schatzkin, Diana Rubin, Jürgen Schrezenmeir, Stefan Schreiber, Arne Schäfer, John Danesh, [illegible] Hugh Watkins, Alistair S Hall, Kim Overvad, Eric Rimm, Eric Boerwinkle, Anne Tybjærg-Hansen, [illegible] Pier M Mannucci, Diego Ardissino, David Siscovick, Roberto Elosua, Kari Stefansson, [illegible] Leena Peltonen, Stephen M Schwartz, David Altshuler, Sekar Kathiresan

Dr Sekar Kathiresan, Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA
skathiresan@partners.org

[illegible]rol is associated with reduced risk of myocardial infarction, but whether this [illegible]ng the fact that genotypes are randomly assigned at meiosis, are independent [illegible]modified by disease processes, mendelian randomisation can be used to test [illegible] plasma biomarker with disease is causal.

[illegible] randomisation analyses. First, we used as an instrument a single nucleotide [illegible]lial lipase gene (*LIPG* Asn396Ser) and tested this SNP in 20 studies [illegible]07 controls). Second, we used as an instrument a genetic score consisting of [illegible]ociate with HDL cholesterol and tested this score in up to 12 482 cases of [illegible]ls. As a positive control, we also tested a genetic score of 13 common SNPs [illegible]rol.

[illegible]llele (2·6% frequency) had higher HDL cholesterol (0·14 mmol/L higher, $p=8\times10^{-13}$) but similar levels of other lipid and non-lipid risk factors for myocardial infarction compared with non-carriers. This difference in HDL cholesterol is expected to decrease risk of myocardial infarction by 13% (odds ratio [OR] 0·87, 95% CI 0·84–0·91). However, we noted that the 396Ser allele was not associated with risk of myocardial infarction (OR 0·99, 95% CI 0·88–1·11, p=0·85). From observational epidemiology, an increase of 1 SD in HDL cholesterol was associated with reduced risk of myocardial infarction (OR 0·62, 95% CI 0·58–0·66). However, a 1 SD increase in HDL cholesterol due to genetic score was not associated with risk of myocardial infarction (OR 0·93, 95% CI 0·68–1·26, p=0·63). For LDL cholesterol, the estimate from observational epidemiology (a 1 SD increase in LDL cholesterol associated with OR 1·54, 95% CI 1·45–1·63) was concordant with that from genetic score (OR 2·13, 95% CI 1·69–2·69, $p=2\times10^{-10}$).

Interpretation Some genetic mechanisms that raise plasma HDL cholesterol do not seem to lower risk of myocardial infarction. These data challenge the concept that raising of plasma HDL cholesterol will uniformly translate into reductions in risk of myocardial infarction.

Funding US National Institutes of Health, The Wellcome Trust, European Union, British Heart Foundation, and the German Federal Ministry of Education and Research.

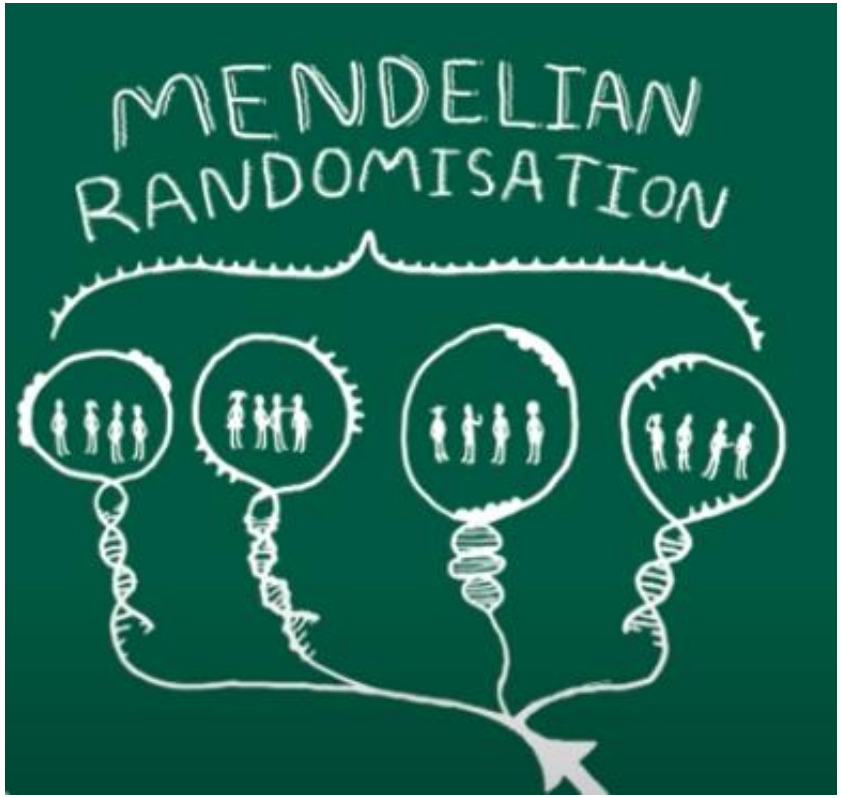
Introduction

Cholesterol fractions such as LDL and HDL cholesterol are among the most commonly measured biomarkers in clinical medicine.[1] Observational studies have shown that LDL and HDL cholesterol have opposing associations with risk of myocardial infarction, with LDL cholesterol being positively associated and HDL cholesterol being inversely associated.[2,3] However, observational studies cannot distinguish between a causal role in the pathological process and a marker of the underlying

Mendelian Randomization studies suggest CRP and HLD levels are not causally associated with coronary heart disease, consistent with clinical trials.

Voight BF (2012) *Lancet*; Elliott (2009) *JAMA*

Mendelian Randomization



A cute 00:02:15 introduction… that glosses over some important pesky details (to be discussed later).

https://www.youtube.com/watch?v=LoTgfGotaQ4

Genetic Epidemiology is Transdisciplinary

The history of genetic epidemiology is a tapestry of observational science, statistical developments, animal and plant breeding experiments, molecular experiments, medicine, epidemiology, technology, and…

Duggal P (2019) Am J Epidemiol

Mendel
(discrete traits)
Galton, Pearson
(continuous traits)
Bateson, Punnett
(Mendel "rediscovered")
Fisher, Wright, Haldane
("the modern synthesis")

Mendelian and Statistical Genetics

Garrod
("inborn errors of metabolism," pharmacogenetics)
Neel
(hemoglobinopathies, radiation)

Clinical Genetics

Morgan, Just, Franklin, Watson, Crick, Hershey, Chase, Daly, McClintock
(structure & role of DNA)
Collins, Lander
(the Human Genome Project, the HapMap, high-throughput genotyping and sequencing technologies)

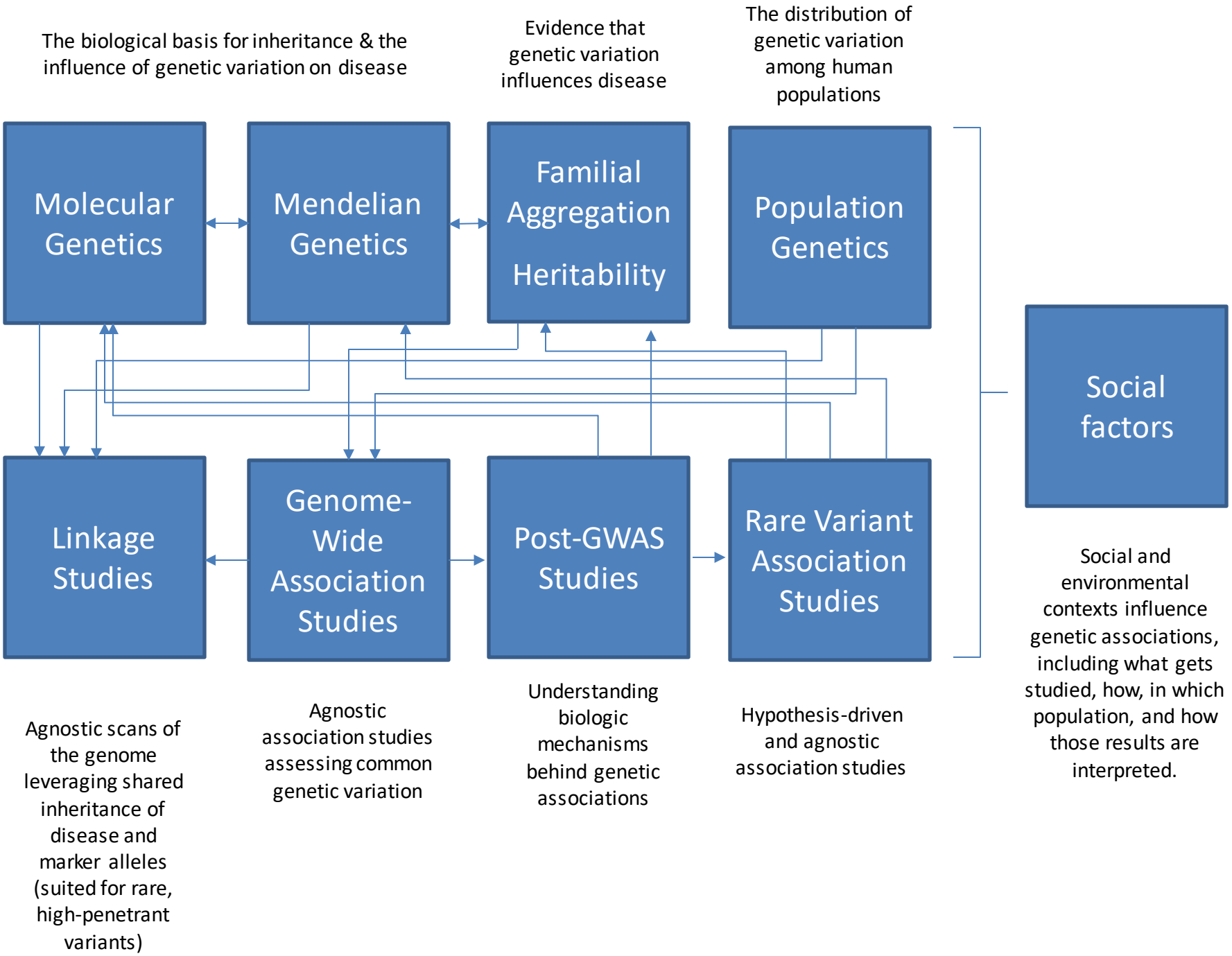
Molecular Genetics

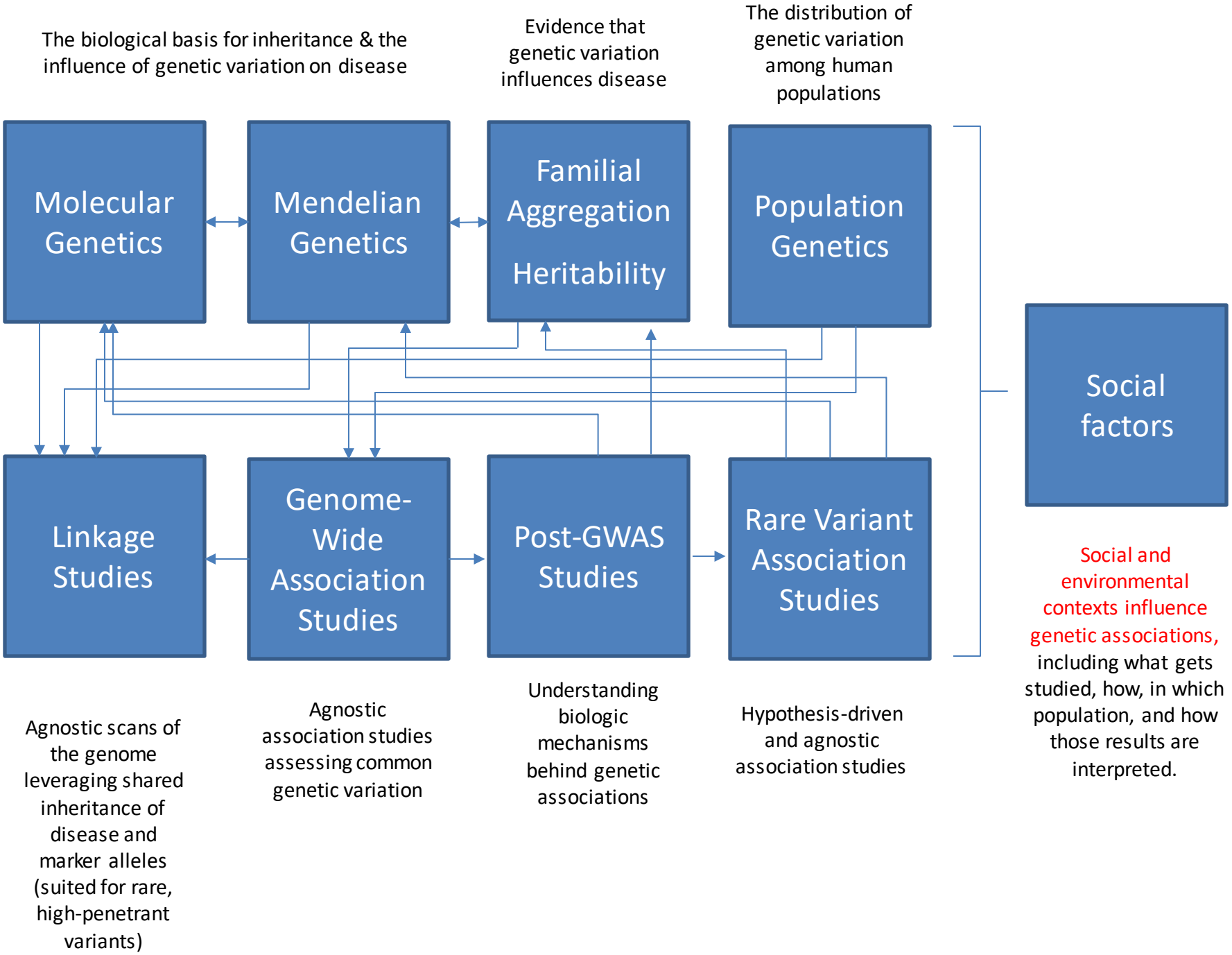
Genetic Epidemiology

"Understanding the function of much that is being revealed will not yield to classical Mendelian genetics but will require the epidemiological approach."

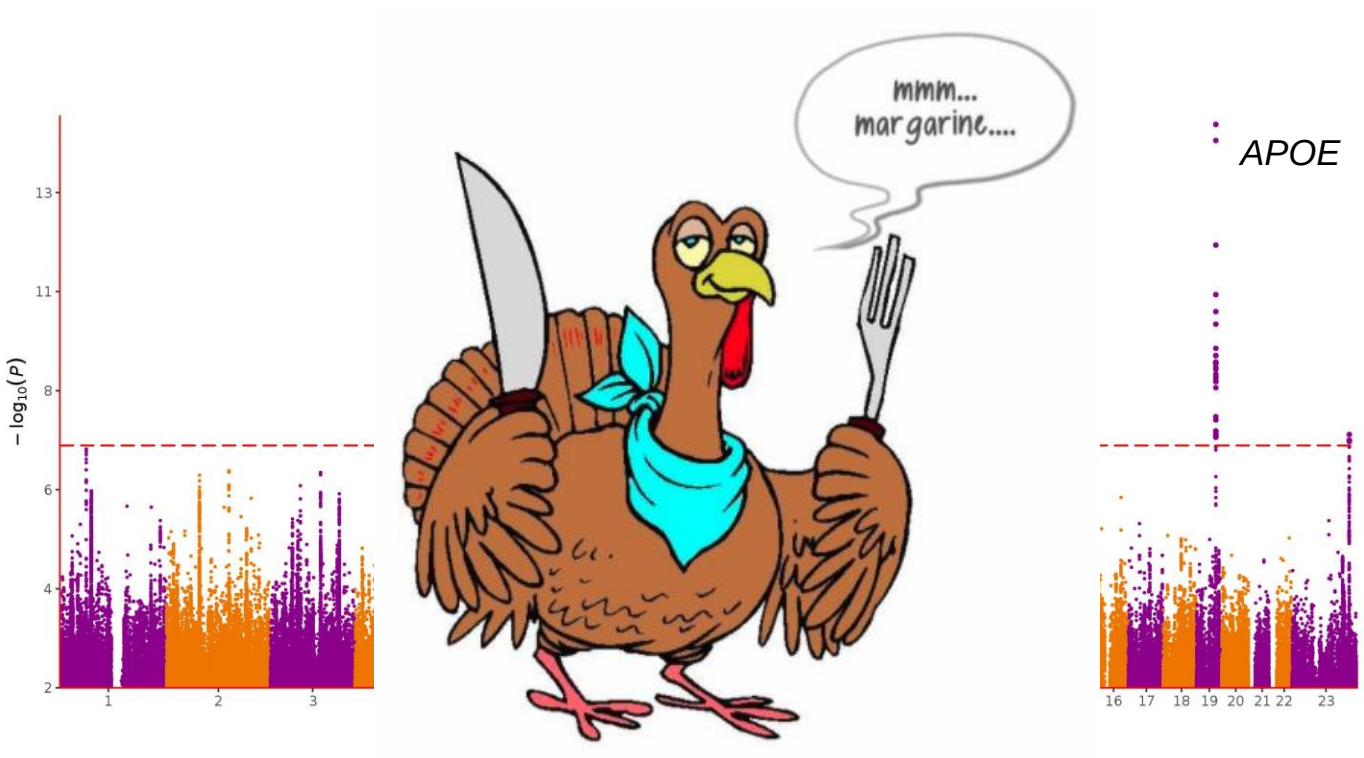
Genetic epidemiology is inseparable from "the concept of multifactorial causation."

Neel (1984) *Genet Epidemiol*


The biological basis for inheritance & the influence of genetic variation on disease
Evidence that genetic variation influences disease
The distribution of genetic variation among human populations
Molecular Genetics
Mendelian Genetics
Familial Aggregation
Heritability
Population Genetics
Social factors
Linkage Studies
Genome-Wide Association Studies
Post-GWAS Studies
Rare Variant Association Studies
Social and environmental contexts influence genetic associations, including what gets studied, how, in which population, and how those results are interpreted.
Agnostic scans of the genome leveraging shared inheritance of disease and marker alleles (suited for rare, high-penetrant variants)
Agnostic association studies assessing common genetic variation
Understanding biologic mechanisms behind genetic associations
Hypothesis-driven and agnostic association studies


The biological basis for inheritance & the influence of genetic variation on disease
Evidence that genetic variation influences disease
The distribution of genetic variation among human populations
Molecular Genetics
Mendelian Genetics
Familial Aggregation
Heritability
Population Genetics
Social factors
Linkage Studies
Genome-Wide Association Studies
Post-GWAS Studies
Rare Variant Association Studies
Social and environmental contexts influence genetic associations, including what gets studied, how, in which population, and how those results are interpreted.
Agnostic scans of the genome leveraging shared inheritance of disease and marker alleles (suited for rare, high-penetrant variants)
Agnostic association studies assessing common genetic variation
Understanding biologic mechanisms behind genetic associations
Hypothesis-driven and agnostic association studies

Example: margarine vs butter use in the UK Biobank



HT @Eric_Fauman
https://twitter.com/Eric_Fauman/status/1101822314163957760?s=20

Example: margarine vs butter use in the UK Biobank

At APOE

- ↑ LDL levels, statin use
- ↑ margarine use
- ↓ butter use

At LPA

- ↑ MI risk
- ↑ margarine use
- ↓ butter use

HT @Eric_Fauman
https://twitter.com/Eric_Fauman/status/1101822314163957760?s=20

Example: margarine vs butter use in the UK Biobank

At APOE

- ↑ LDL levels, statin use
- ↑ margarine use
- ↓ butter use

At LPA

- ↑ MI risk
- ↑ margarine use
- ↓ butter use

"Individuals who learn they have high cholesterol or have a family history of heart disease are more likely to switch to margarine instead of butter."

HT @Eric_Fauman
https://twitter.com/Eric_Fauman/status/1101822314163957760?s=20

Example: margarine vs butter use in the UK Biobank

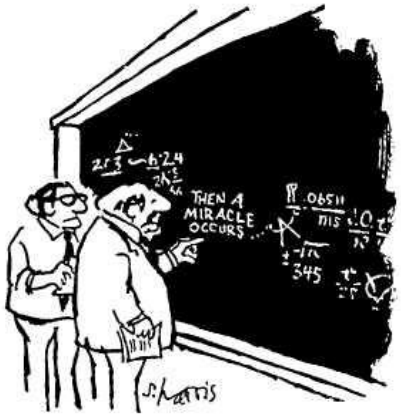
Take home messages

- Social context matters

Example: margarine vs butter use in the UK Biobank

Take home messages

- Social context matters
- Hypothesized biological mechanisms should be concrete & testable

G → 

"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

→ Y

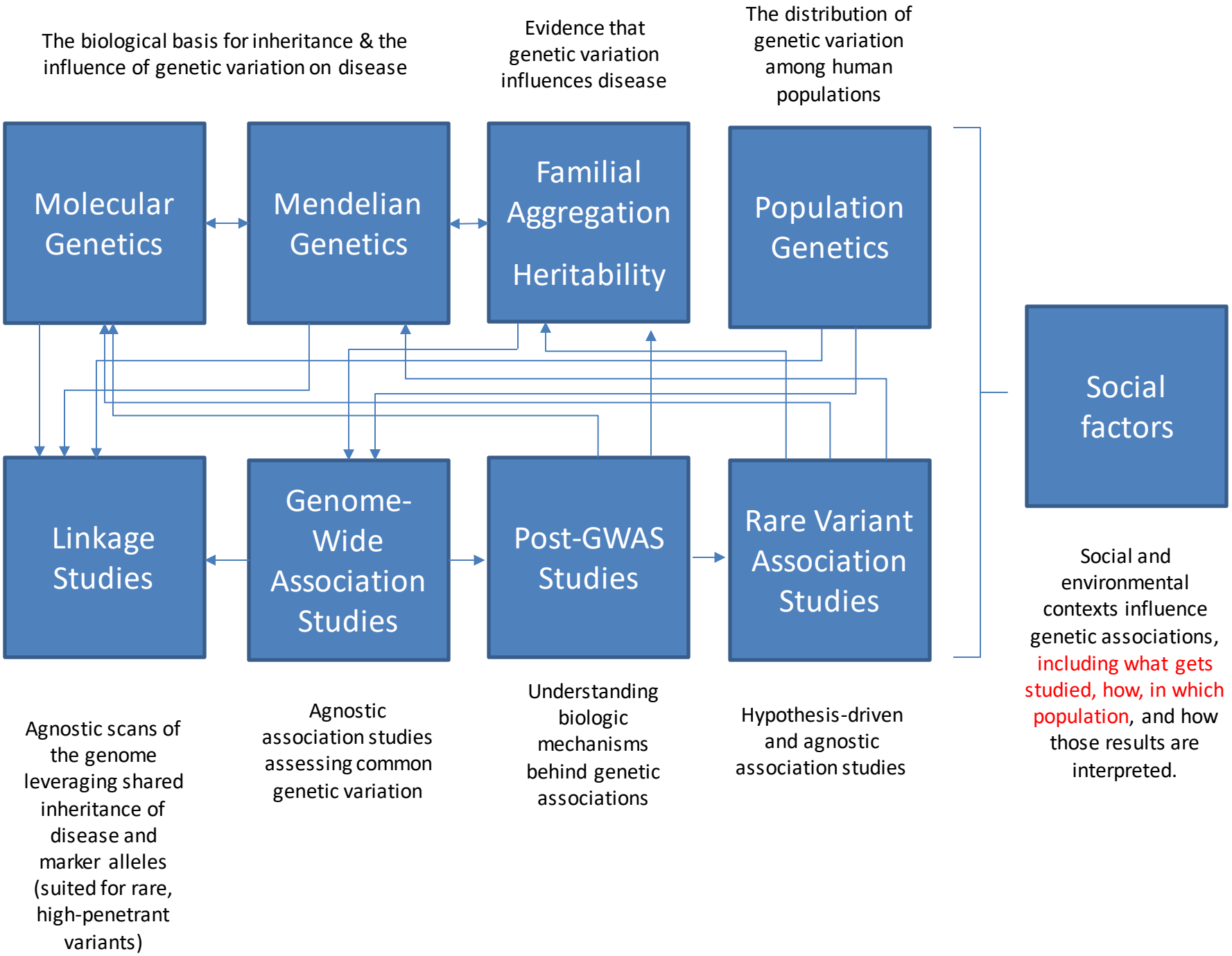
Example: margarine vs butter use in the UK Biobank

Take home messages

- Social context matters
- Hypothesized biological mechanisms should be concrete & testable

"Incorporation of disease biology involves a real understanding of the causal pathways leading from genes and environment to disease causation, and expression of these concepts in our mathematical models for penetrance."

Thomas (2000) *Genet Epidemiol*


The biological basis for inheritance & the influence of genetic variation on disease
Evidence that genetic variation influences disease
The distribution of genetic variation among human populations
Molecular Genetics
Mendelian Genetics
Familial Aggregation
Heritability
Population Genetics
Social factors
Linkage Studies
Genome-Wide Association Studies
Post-GWAS Studies
Rare Variant Association Studies
Social and environmental contexts influence genetic associations, including what gets studied, how, in which population, and how those results are interpreted.
Agnostic scans of the genome leveraging shared inheritance of disease and marker alleles (suited for rare, high-penetrant variants)
Agnostic association studies assessing common genetic variation
Understanding biologic mechanisms behind genetic associations
Hypothesis-driven and agnostic association studies

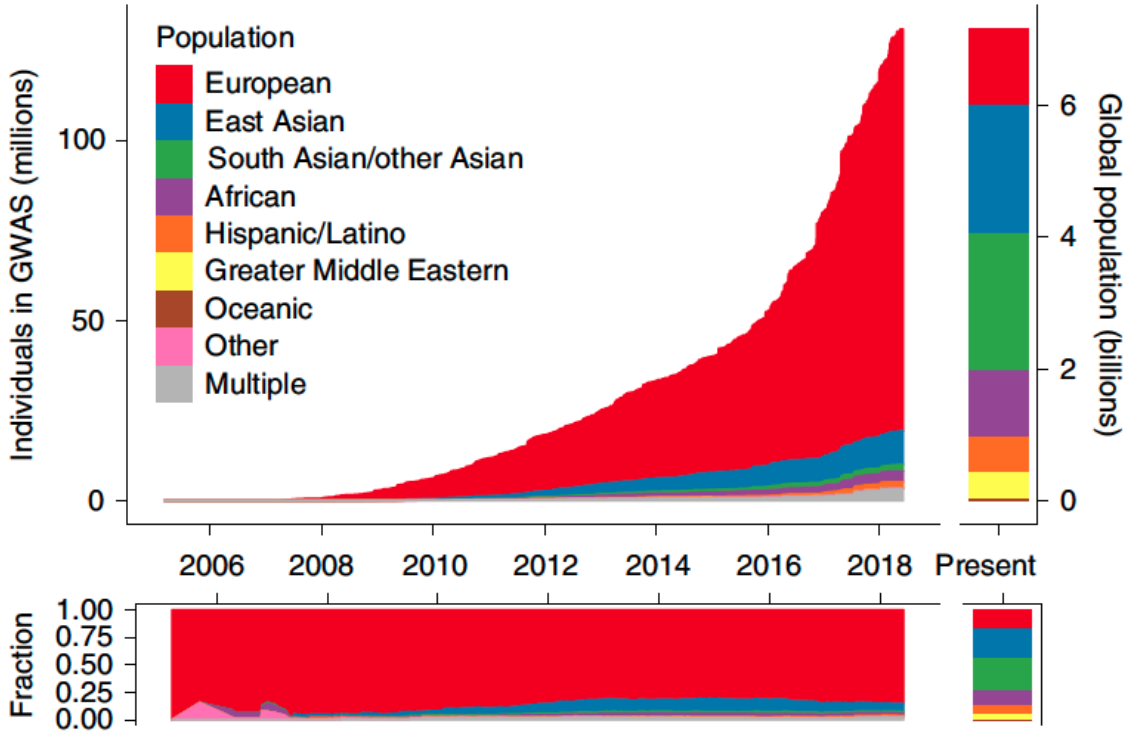
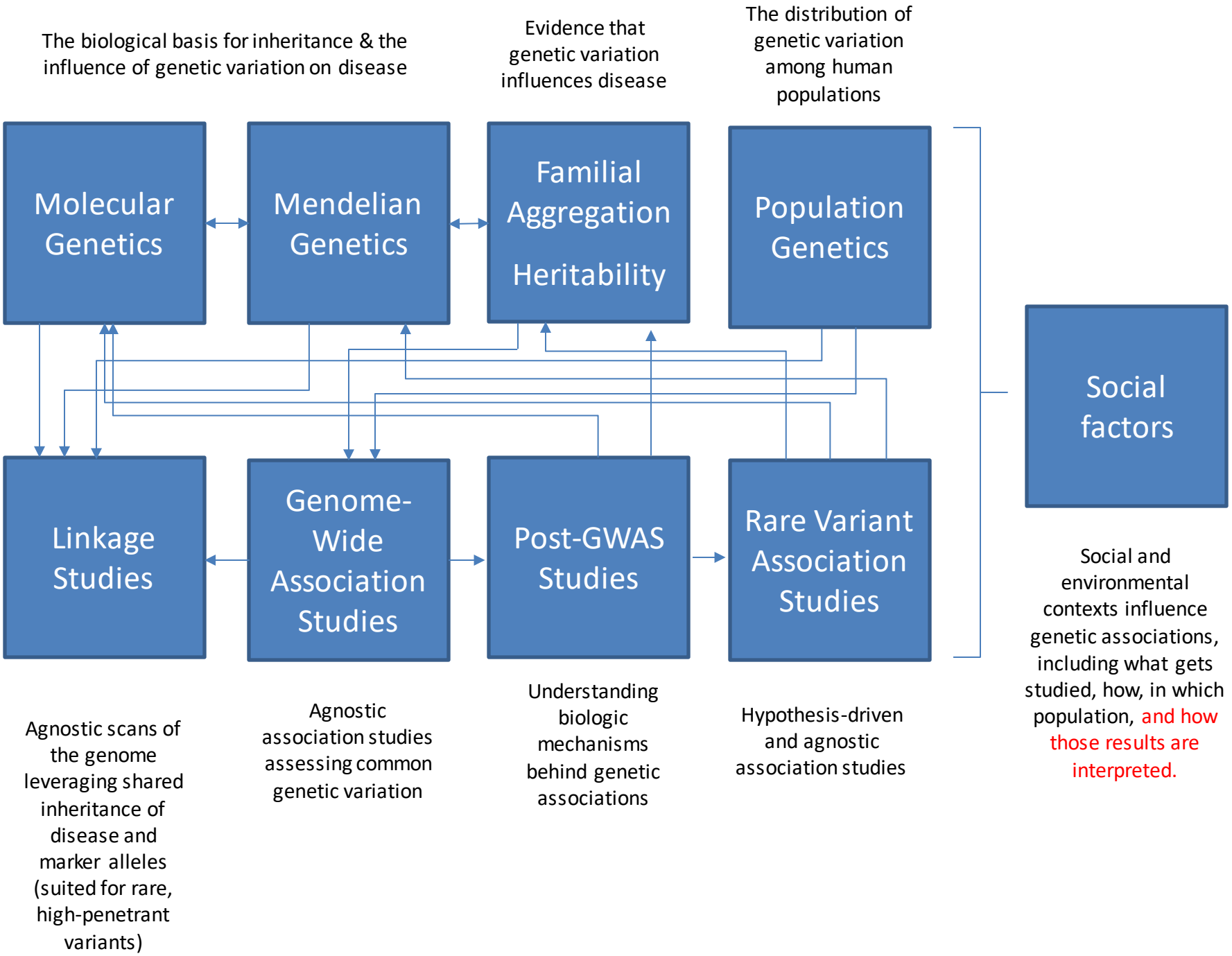


Fig. 1 | Ancestry of GWAS participants over time, as compared with the global population. Cumulative data, as reported by the GWAS catalog[76]. Individuals whose ancestry is 'not reported' are not shown.

This lack of representation limits equity and scientific opportunity.

Martin (2019) Nat Genet


The biological basis for inheritance & the influence of genetic variation on disease
Evidence that genetic variation influences disease
The distribution of genetic variation among human populations
Molecular Genetics
Mendelian Genetics
Familial Aggregation
Heritability
Population Genetics
Social factors
Linkage Studies
Genome-Wide Association Studies
Post-GWAS Studies
Rare Variant Association Studies
Social and environmental contexts influence genetic associations, including what gets studied, how, in which population, and how those results are interpreted.
Agnostic scans of the genome leveraging shared inheritance of disease and marker alleles (suited for rare, high-penetrant variants)
Agnostic association studies assessing common genetic variation
Understanding biologic mechanisms behind genetic associations
Hypothesis-driven and agnostic association studies

The history of genetic epidemiology is a tapestry of observational science, statistical developments, animal and plant breeding experiments, molecular experiments, medicine, epidemiology, technology, and…
prevailing ideas explaining and justifying social hierarchy.

Broader Scientific & Social Context

Blumenbach
On the Natural Variety of Mankind

Georges Cuvier
Louis Agassiz
many others

Galton, Pearson, Fisher
many others

Mendelian and Statistical Genetics

Clinical Genetics

Molecular Genetics

Genetic Epidemiology

"Understanding the function of much that is being revealed will not yield to classical Mendelian genetics but will require the epidemiological approach." Genetic epidemiology is inseparable from "the concept of multifactorial causation."

Essentialist theories of "race"

Eugenics

18th Century — 19th Century — 20th Century — 21st Century

Imperialism

Chattel Slavery

Immigration Restrictions

Forced Sterilization

Jim Crow & Segregation

Revivial of debunked eugenic theories: "Race Realism," "Human Biodiversity"

Why Talk about Eugenics?

- Shaky, pseudoscientific belief that empirical biological evidence exists to justify some individuals as inferior → Shouldn't forget this history → the "backdoor to eugenics" is slippery, and the history of eugenics is not "ancient history" at all
- Eugenic thinking is still very much alive → the historic time of civil unrest makes it all the more important that we keep our history in mind as we do our work
- Rather than assuming our data speaks for itself, we have a duty to consider the language, context, and framing of our work, so that it helps, not harms

Although founded on shaky arguments and weak empirical support, eugenics was firmly within the Western scientific establishment.

"The call is coming from inside the house!"

Slide courtesy of Robbee Wedow

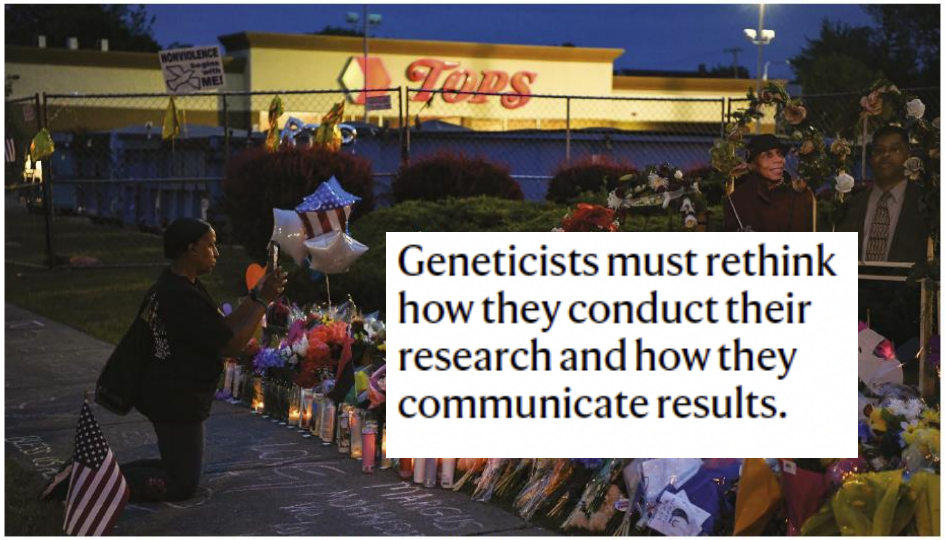
Setting the agenda in research

Comment

Racism in science

Counter the weaponization of genetics research by extremists

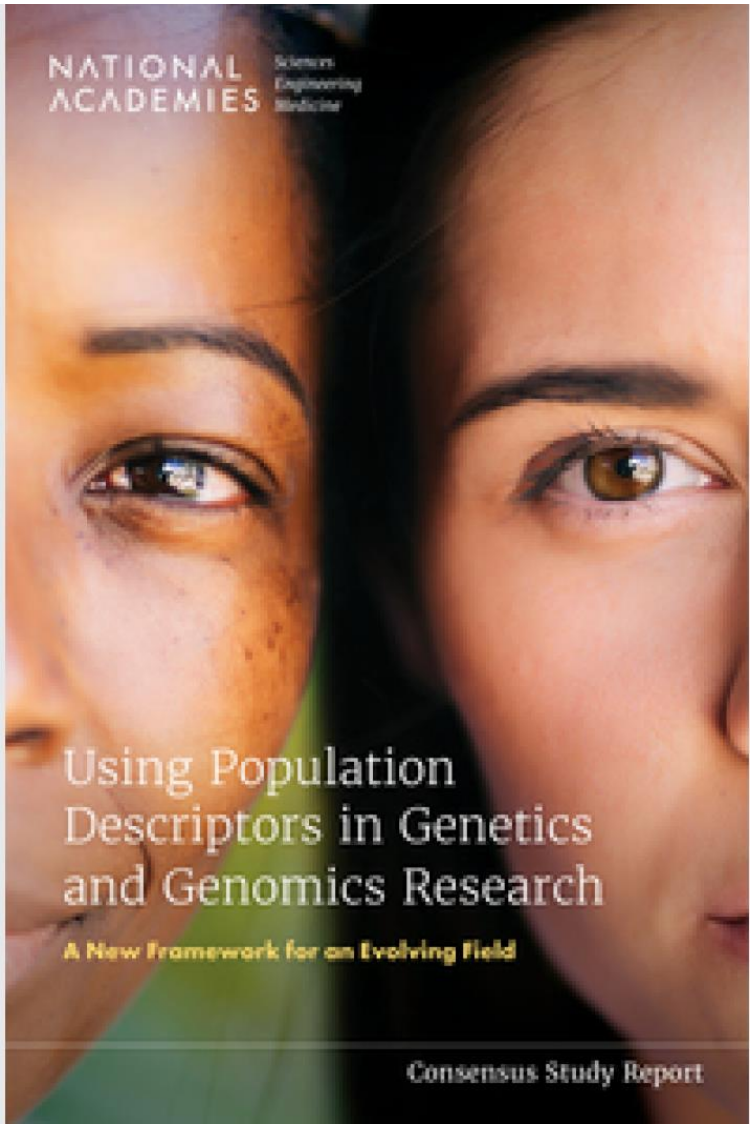
Jedidiah Carlson, Brenna M. Henn, Dana R. Al-Hindi & Sohini Ramachandran



Geneticists must rethink how they conduct their research and how they communicate results.

KENNY HOLSTON/NEW YORK TIMES REDUX/EYEVINE

A memorial to the ten Black people who were killed by a shooter outside a shop in Buffalo, New York, in May 2022.

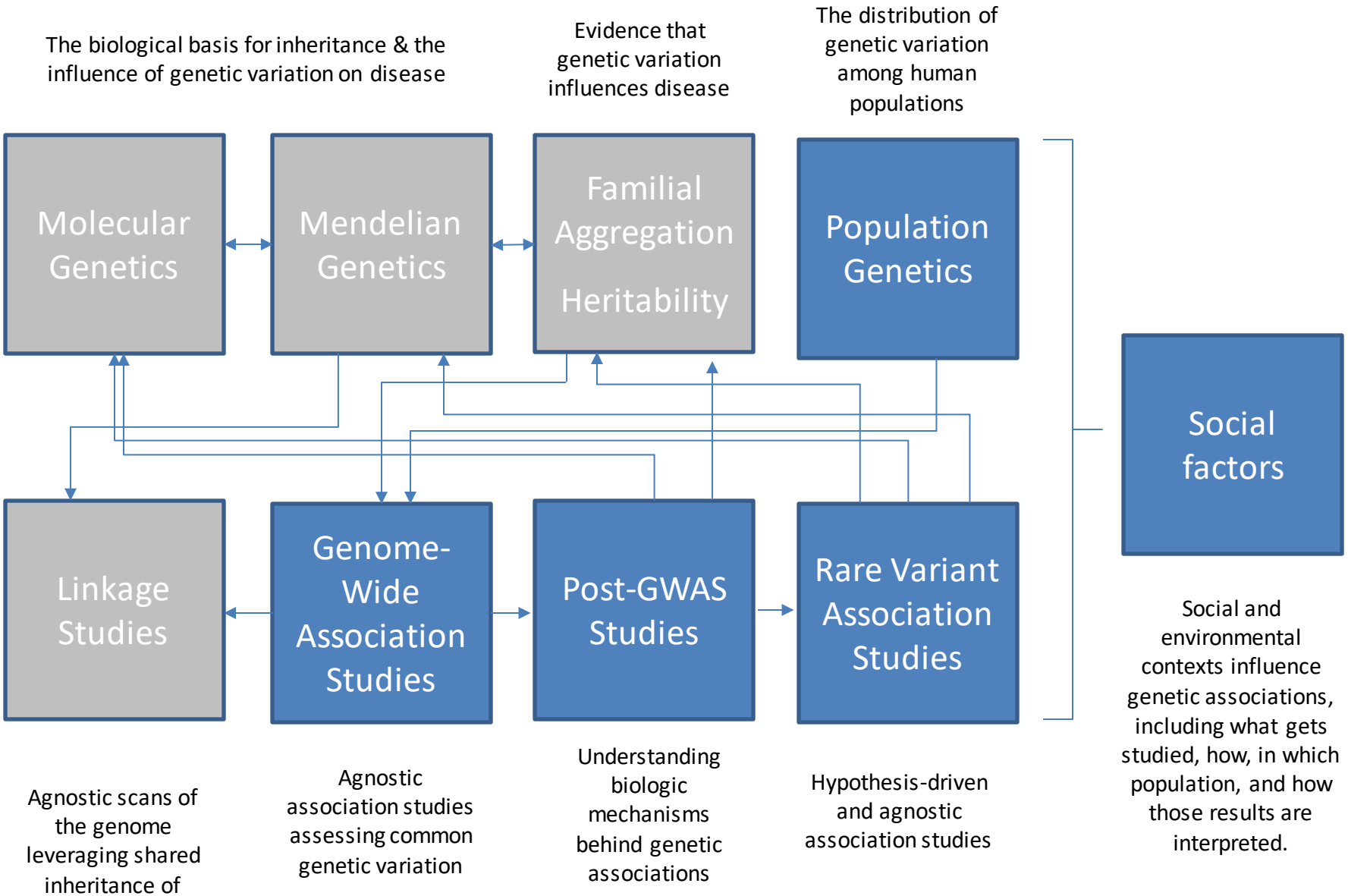


Carlson (2022) *Nature*

Preguntas?

Agenda

- Course format
- Intro to genetic epidemiology
- Course scope
- A few basic definitions

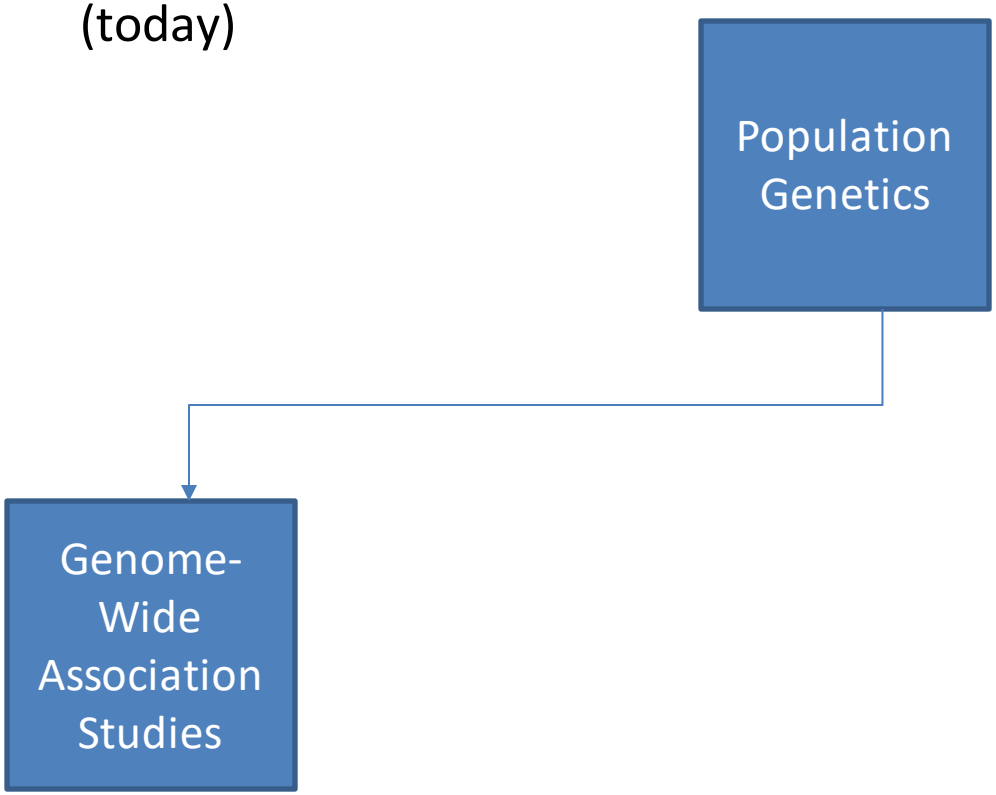


We will not explicitly cover the greyed out topics.

Sue Ingles chapter on Basic Concepts of Molecular Genetics in Duncan Thomas' Statistical Methods for Genetic Epidemiology remains one of my favorite short introductions. Other recommendations? Paste them in the chat or the discussion page.

Session 1
(today)

Basic concepts



Data management. Quality control, quality assessment.

Peter Kraft
Wendy Wong

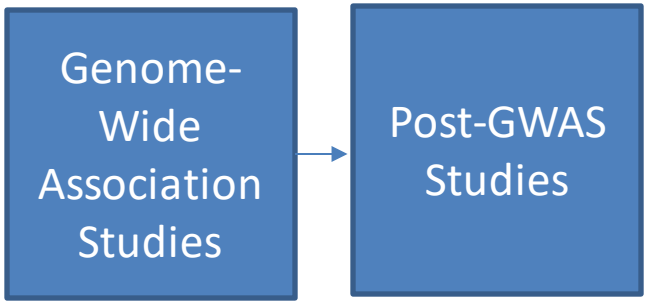
Session 2
(September 20)

Genome-Wide Association Studies

Basic GWAS testing.
Meta-analyses.
Visualizations.
Annotating GWAS results.

Aubrey Hubbard

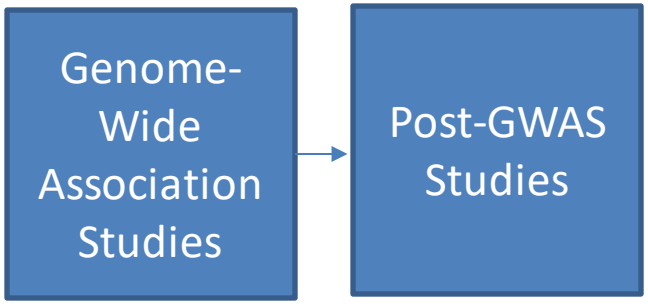
Session 3
(September 27)



Methods to prioritize likely causal *variants*. Statistical fine mapping, colocalization.

Diptavo Dutta

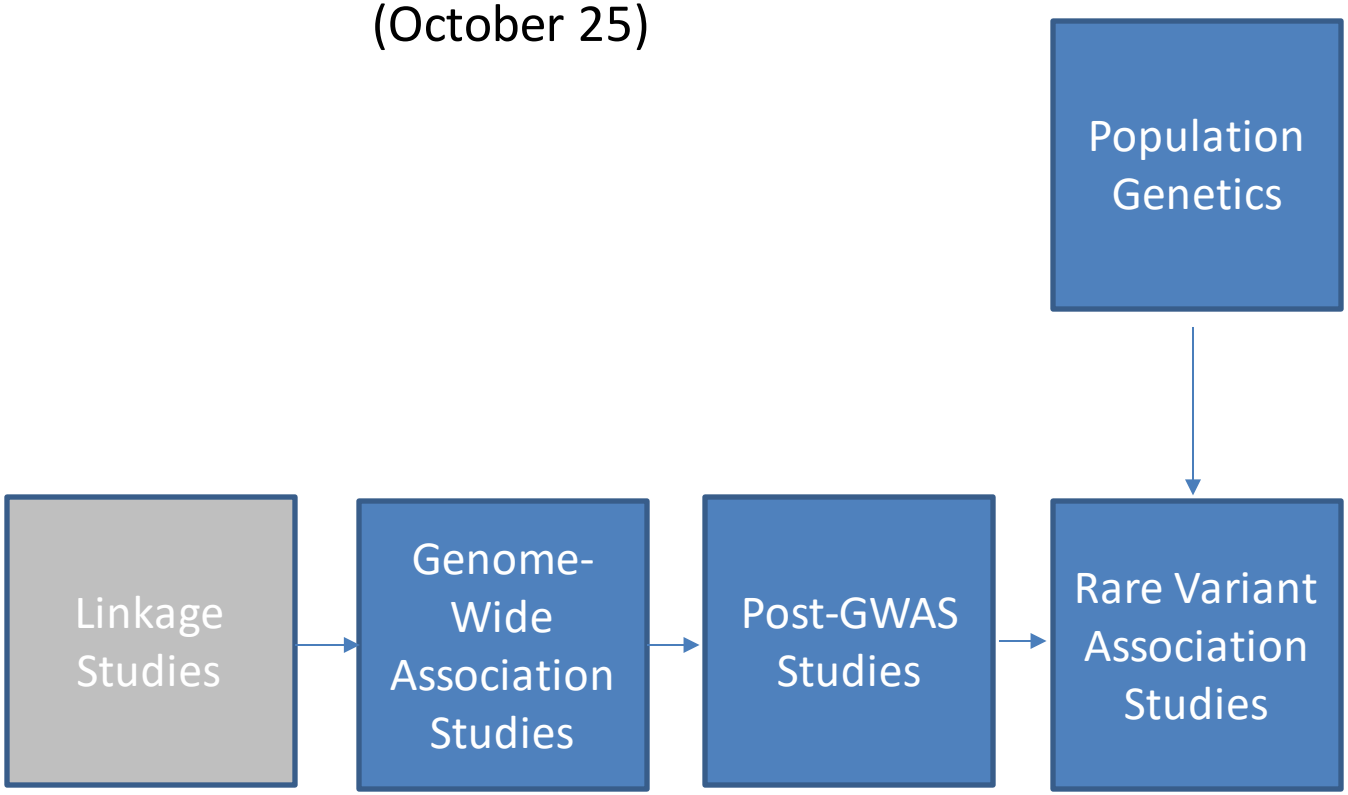
Session 4
(October 25)



Genetic architecture: distribution of effect sizes, number of causal alleles, functional enrichment. Polygenic risk models.

Haoyu Zhang

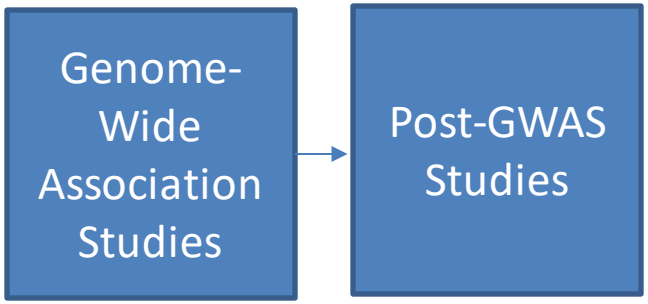
Session 5
(October 25)



Hypothesis-driven and agnostic association studies

Xihao Li
Aubrey Hubbard

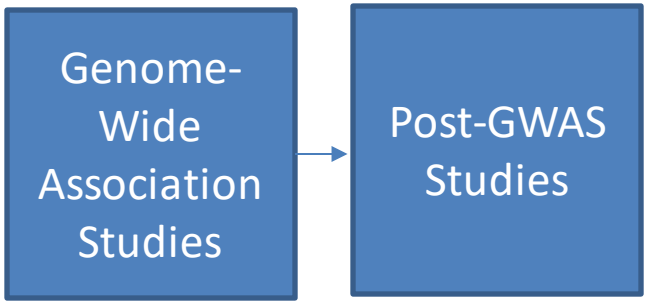
Session 6
(November 15)



Integrative methods (TWAS, PWAS etc.) to prioritize causal *genes.*
Mendelian Randomization.

Guanghao Li
Diptavo Dutta
Sheila Rajagopal
Aubrey Hubbard

Session 7
(November 29)



GWAS, fine-mapping and polygenic risk scores in diverse-genetic-ancestry and recently admixed samples

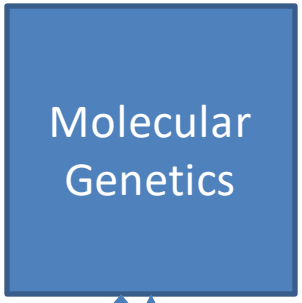
David Conti

Session 8
(December 6)

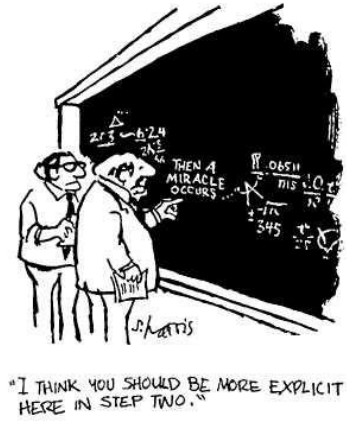
Genetic mosaicism and clonal hematopoiesis

Mitch Machiela
Weiyin Zhou
Sheila Rajagopal
Aubrey Hubbard

Session 9
(December 13)



G



Y

Functional Genomics

Mila Prokunina-Olsson
Kevin Brown

Fragen?

Agenda

- Course format
- Intro to genetic epidemiology
- Course scope
- A few basic definitions

A few basic definitions

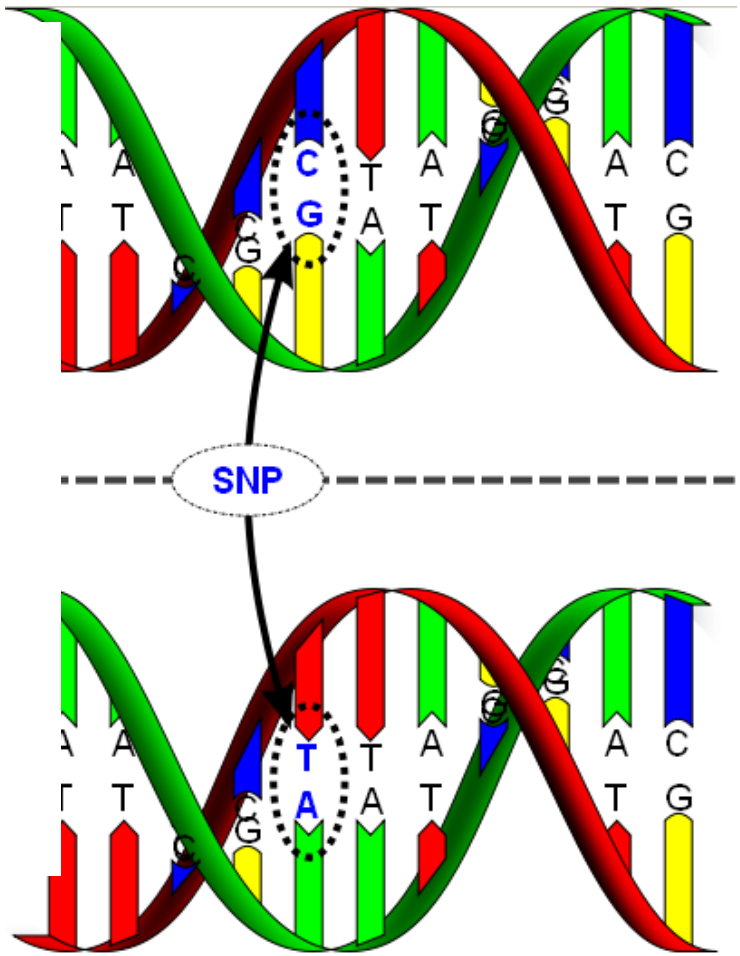
- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

A few basic definitions

- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

Single Nucleotide Polymorphisms

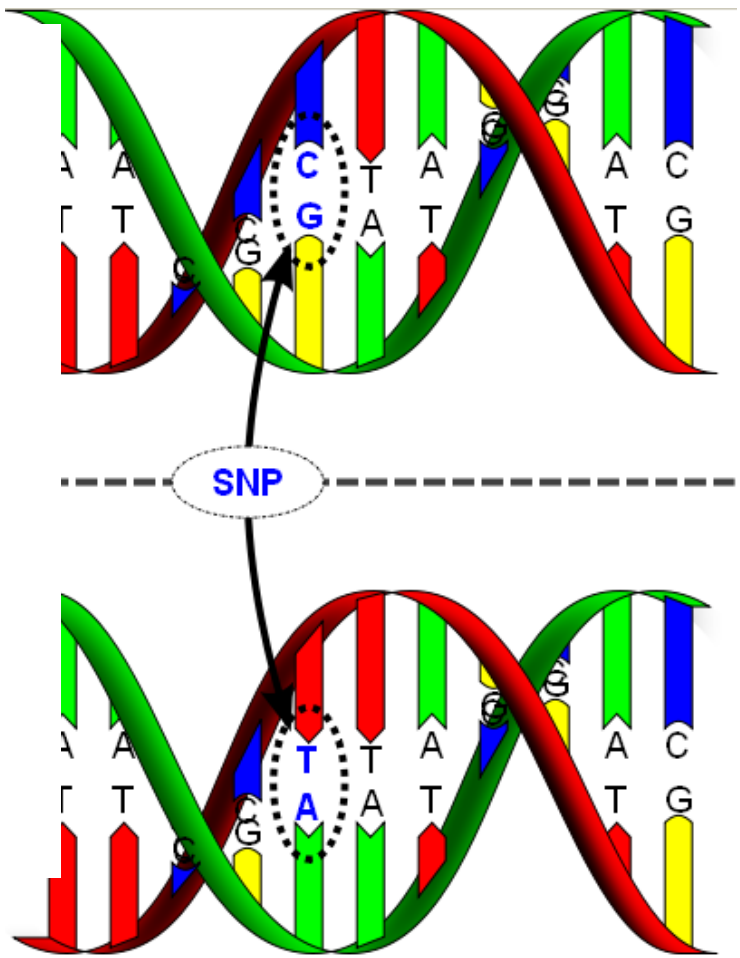
One Allele



Another Allele

Single Nucleotide Polymorphisms

One Allele



Another Allele

SNPs almost always have only two alleles.

A SNP like this one is sometimes written as "a C/T SNP" …

and sometimes written as "a G/A SNP."

To avoid ambiguity you have to specify which strand you are using to define the alleles.

Often the (+) strand of the human reference genome is used. The (+) strand runs 5' to 3,' starting at the telomere of the p arm of each chromosome. The current version of the reference genome is GRCh38/hg38.

SNPs are often referred to by their dbSNP Reference SNP (refSNP) number, or by chromosome and position. E.g. rs671 and chr12:111803962:G:A refer to the same variant.

https://en.wikipedia.org/wiki/Reference_genome
https://www.ncbi.nlm.nih.gov/grc/human

SC Nelson (2020) "Is 'forward' the same as 'plus'?… and other adventures in SNP allele nomenclature" *Trends Genet*

<u>Study A</u>
The G allele of rs8675309 is associated with an increase in circulating rhubarbiol relative to the T allele.

<u>Study B</u>
The C allele of rs8675309 is associated with an increase in circulating rhubarbiol relative to the A allele.

rs8675309 is “strand unambiguous.”

<u>Study A</u>
The G allele of rs90210 is associated with an increase in circulating Luke Perry Factor relative to the C allele.

<u>Study B</u>
The C allele of rs90210 is associated with a decrease in circulating Luke Perry Factor relative to the G allele.

rs90210 is “strand ambiguous.”
(C/G and A/T SNPs are strand ambiguous.)

Be as clear about strand as you can be!

A few basic definitions

- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

Definitions and an example: the ABO blood group

- An allele is the nucleotide sequence at a polymorphic locus (i.e. a small region of the genome, e.g. a single base or a gene) on a chromosome.
- While there are (typically) only at most 2 different alleles carried by an individual at a locus, there can be more than 2 alleles in the population.
- The combination of alleles at a locus carried by an individual is that individual's genotype.
- A familiar example is the ABO blood group, defined by presence of antigens (i.e. structures that induce an immune response) on red blood cells (RBCs).
- The genetics of ABO blood type maps to a single locus. Alleles A and B are codominant with each other and dominant over allele O (no antigen), such that the possible genotypes, i.e. the pairwise combinations of ABO alleles, give rise to phenotypes A, B, AB, or O.

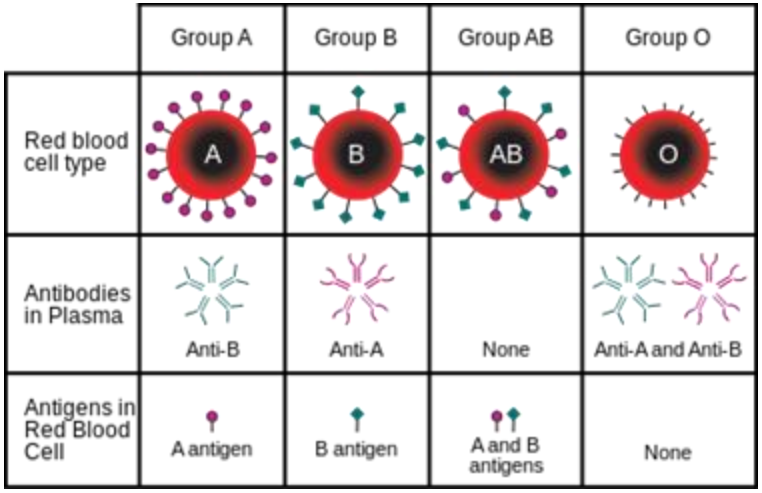


Figure from Wikipedia

Genotype frequencies in the population

A diploid individual (human) has two alleles (maternal and paternal) at an autosomal locus. Assuming there are K distinct alleles in the population,

Let n_{ij}* be the number of individuals with genotype ij,

$$i=1..K, \quad j=i..K.$$

Then the genotype frequency $P_{ij} = n_{ij}/n$,

n is the total number of individuals in the population

$$n=\Sigma\, n_{ij}$$

*NB. The order of i and j is usually arbitrary in this notation such that n_{ij} and n_{ji} refer to the same individuals

Multi allele example: ABO genotypes

$n_{11} = n_{OO}$	1,168
$n_{12} = n_{OA}$	1,080
$n_{13} = n_{OB}$	377
$n_{22} = n_{AA}$	262
$n_{23} = n_{AB}$	186
$n_{33} = n_{BB}$	44
n_{total}	3,117

$P_{11} = P_{OO}$	0.37
$P_{12} = P_{OA}$	0.35
$P_{13} = P_{OB}$	0.12
$P_{22} = P_{AA}$	0.08
$P_{23} = P_{AB}$	0.06
$P_{33} = P_{BB}$	0.01
	1.00

Wolpin (2010) *Cancer Research*

Two allele example: *MTHFR** C677T

MTHFR encodes methylene tetrahydrofolate reductase, an enzyme relevant to chemotherapy with 5-fluorouracil. There are two alleles, C or T, nucleotide position 677 in the gene. These alleles encode different amino acids in the protein at residue 222, either alanine (C) or valine (T), resulting in effects on protein stability, either stable (C) or thermolabile (T).

count of genotypes

$n_{11} = n_{CC}$	334
$n_{12} = n_{CT}$	350
$n_{22} = n_{TT}$	82
n_{total}	766

frequency of genotypes

$P_{11} = P_{CC}$	0.44
$P_{12} = P_{CT}$	0.46
$P_{22} = P_{TT}$	0.10
	1.00

Han (2007) *Carcinogenesis*

Allele frequencies in the population

"Allele frequencies" are the proportions of chromosomes in the population with each of the unique alleles at the variable locus.

Let m_i be the number of copies of allele i in the population.

Then the allele frequency p_i is m_i/m, where $m=2n$ is the total number of chromosomes carrying the variable locus in the population of n individuals.

$$p_i = (2n_{ii} + \sum_{j \neq i} n_{ij})/(2n) = (2\,P_{ii} + \sum_{j \neq i} P_{ij})/2.$$

Example: ABO genotypes to allele frequencies

$n_{11} = n_{OO}$	1,168
$n_{12} = n_{OA}$	1,080
$n_{13} = n_{OB}$	377
$n_{22} = n_{AA}$	262
$n_{23} = n_{AB}$	186
$n_{33} = n_{BB}$	44
n_{total}	3,117

$p_1 = p_O$	0.61
$p_2 = p_A$	0.29
$p_3 = p_B$	0.10
	1.00

Wolpin (2010) *Cancer Research*

Example: *MTHFR* C677T genotypes to allele frequencies

$n_{11} = n_{CC}$	334
$n_{12} = n_{CT}$	350
$n_{22} = n_{TT}$	82
n_{total}	766

$p_1 = p_C$	.66
$p_2 = p_T$	.34
	1.00

Han (2007) *Carcinogenesis*

Notation for alleles for the most common type of variation: diallelic SNPs (K=2)

Single nucleotide polymorphisms (SNPs) – the most common type of variation, typically have 2 unique alleles, i.e. diallelic

For diallelic SNPs, we often use notation that substitutes p and q for p_1 and p_2. Since $p_1+p_2=1$,

$$p_1 = 1- p_2 = p = 1- q$$

However, we still have to specify to which allele p refers! Typically, p refers to the minor allele in the population, so $p < 0.5$.

People also often use "MAF" = minor allele frequency = p

The “N” for genotypes v. alleles

Genotypes

The number (i.e. N) of genotypes at a polymorphic locus in a sample is equal to the **number of individuals**.

Alleles

The number (i.e. N) of alleles is the number of chromosomal regions at a polymorphic locus in the population or, for autosomes, **2x the number of individuals**.

A few basic definitions

- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

Hardy-Weinberg Equilibrium (HWE)

In 1908, Godfrey Hardy and Wilhelm Weinberg independently derived a formula relating the allele frequencies in parents to genotype frequencies in offspring when they achieve equilibrium status.

Formally, HWE occurs under the following assumptions.

- **Random mating**
- **No inbreeding**
- **Infinite population size**
- **Discrete generations**
- **Equal allele frequencies in males and females**
- **No mutation, migration, or selection.**

Even when these assumptions do not hold exactly, HWE often provides a good (and useful) approximation for population genotype frequencies.

Laird and Lange

The Hardy-Weinberg Formula

Using notation from the "Alleles and Genotypes" vignette, the HW formula states that, in general, $P_{ij} = 2\, p_i\, p_j$ if $i \neq j$, and p_i^2 if $i=j$. For the diallelic case:

$$P_{11} = p^2$$
$$P_{12} = 2\, p\, q$$
$$P_{22} = q^2$$

That is, genotype probabilities are given by a multinomial distribution with parameters $(2, p_1,\ldots,p_K)$, or, in the special case of diallelic markers, a binomial* distribution with parameters $(2,p)$.

[*Recall the binomial distribution: $(p + q)^2 = p^2 + 2\, p\, q + q^2$]

We can use HWE to…

- Test for genotyping errors
- Simplify calculations, e.g. in simulations
- Test for underlying deviations from HWE assumptions
  - Random mating
  - No inbreeding
  - Infinite population size
  - Discrete generations
  - Equal allele frequencies in males and females
  - No mutation, migration or selection
- Test for marker-disease association in cases only for some inheritance modes (e.g. dominant, recessive) but low power

A few basic definitions

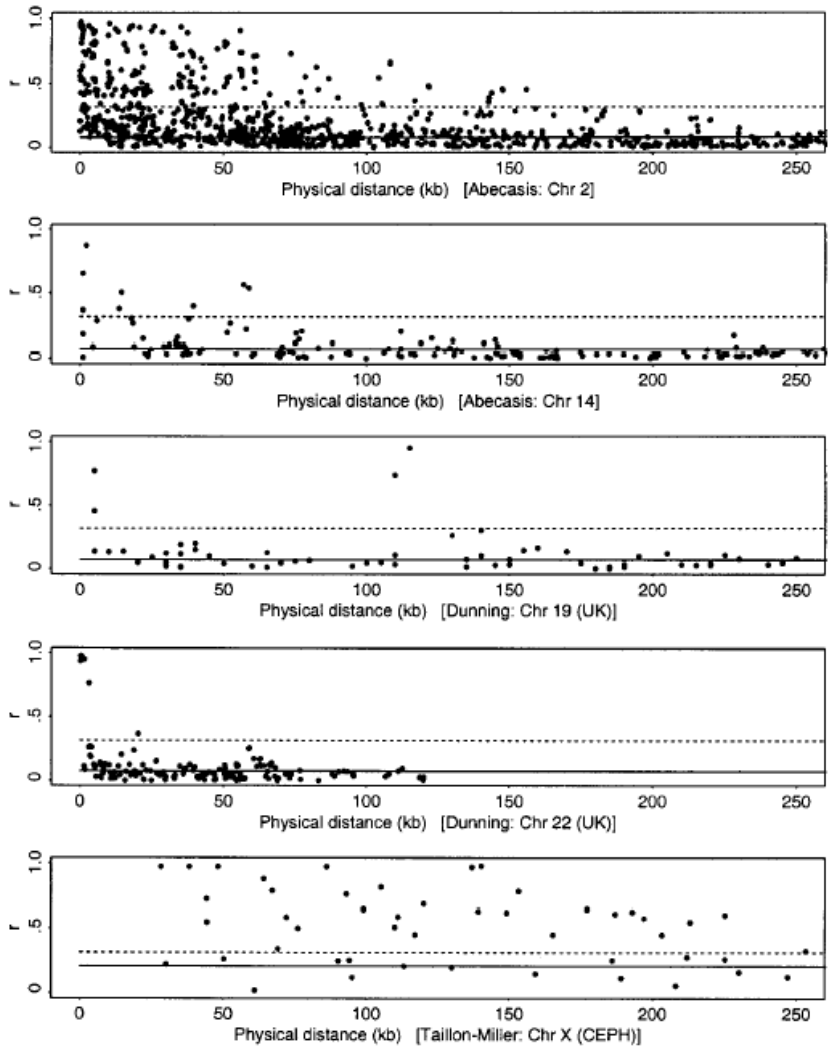
- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

Linkage disequlibrium (LD)

- What are we talking about?
  - The correlation between genetic markers in a population, how it is measured or estimated, and how it is quantified.
  - Typically measured using r^2 (see appendix)
  - High r^2 between two SNPs requires both have similar MAFs
- What are the underlying biological mechanisms?
  - Mutation and recombination in meiosis.
  - Selection
  - Demographics
- Why study?
  - To assess the number of independent tests in the genome.
  - To identify a minimum number of variants that retain coverage of genetic variation in the genome, i.e. through correlation structure.
  - To impute markers that are not directly genotyped
  - To help understand the origins and biological properties of genetic variation in the genome.

Decay of LD: r as a function of physical distance

5 chromosomal regions (2001)



Am. J. Hum. Genet. 69:1–14, 2001

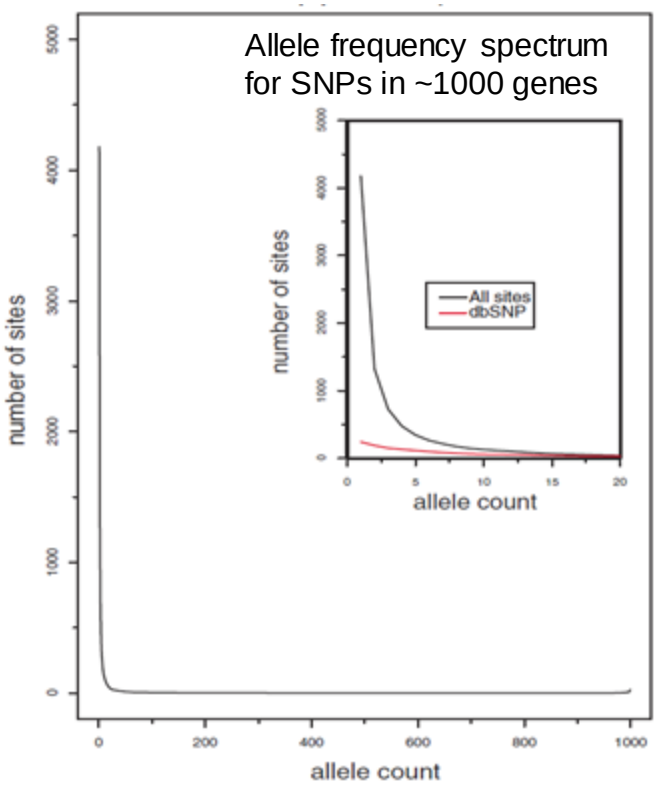
Figure 3 Plots of $\hat{r}$, as a function of physical distance (in kb), for SNP data from five regions (Dunning et al. 2000; Taillon-Miller et al. 2000; Abecasis et al. 2001). On each plot, points above the unbroken line are in significant LD at the .05 level, and points above the dotted line correspond to what Kruglyak (1999) has called "useful LD"; these lines are set at $r = .316$, the equivalent of $r^2 = .1$.

A few basic definitions

- Single nucleotide polymorphisms (SNPs)
- Genotype and allele frequencies (heterozygosity)
- Hardy-Weinberg equilibrium
- Linkage disequilibrium (LD)
- Variation in allele frequencies and linkage disequilibrium

Distribution of allele frequencies

In a sample of 697 subjects



Most SNPs are rare (MAF<1%)

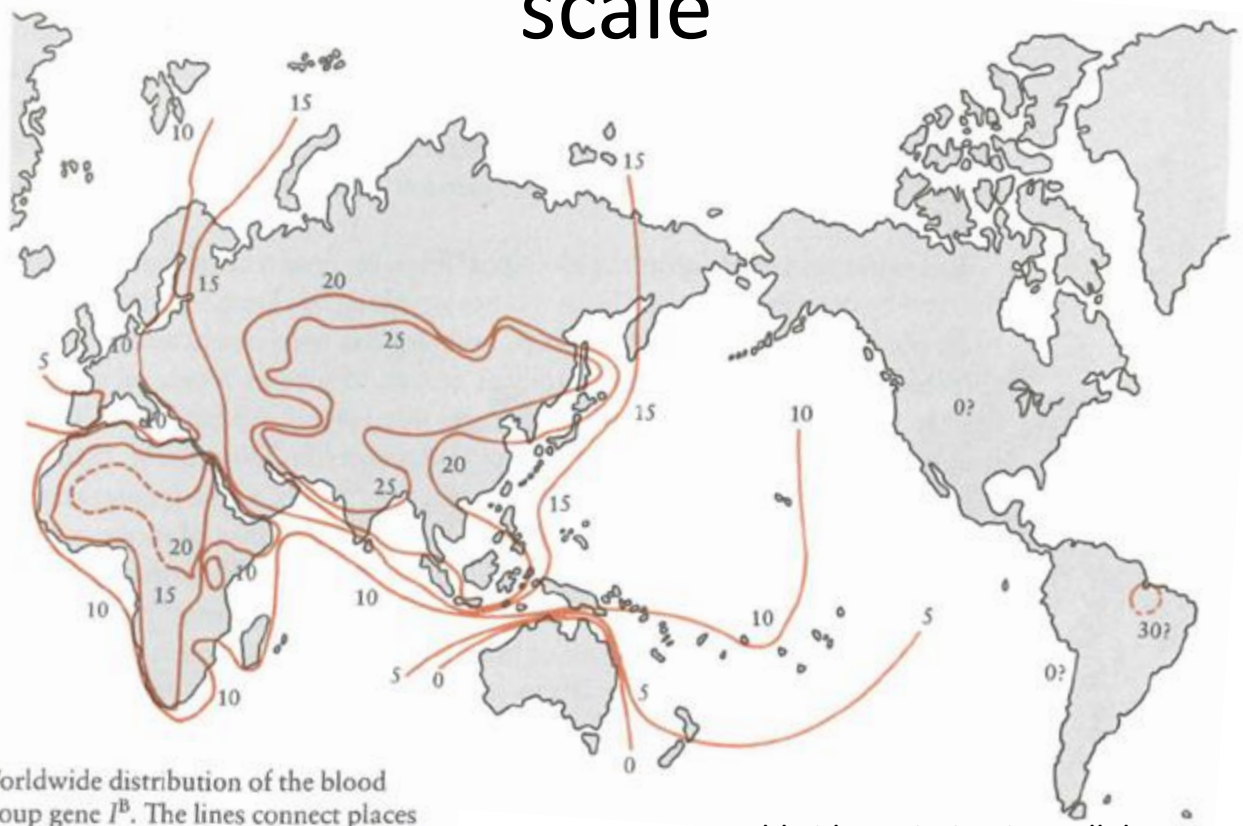
But, most of the differences between unrelated individuals are due to common variants (MAF>5%).

dbSNP is a database to record all SNPs seen in genetic studies (at NCBI/NIH)

Marth (2011) *Genome Biology*

Allele frequencies at individual loci differ across **geographically-defined** populations, more or less in a smooth gradient

Allele frequency variation at the global scale



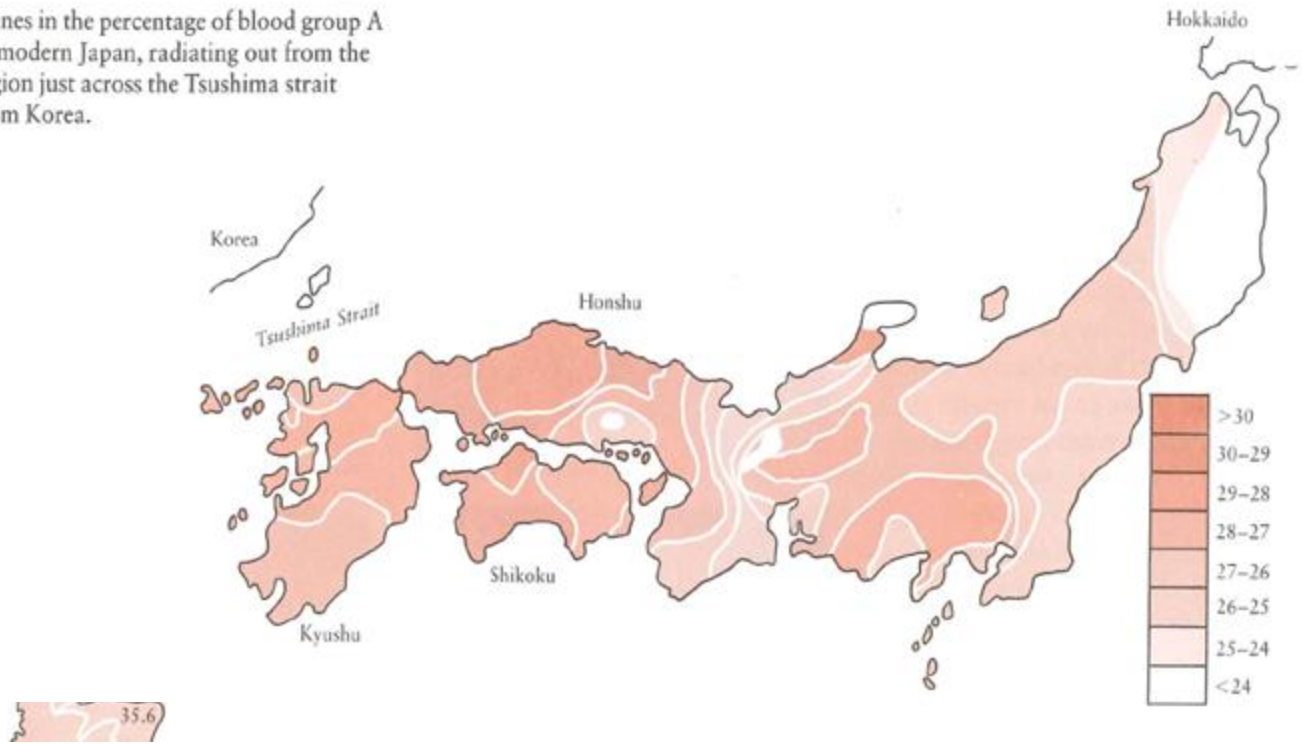
Worldwide distribution of the blood group gene I^B. The lines connect places with equal frequencies of the gene. Note that central Asia is a region of high frequency of I^B, with the frequency decreasing in all directions.

Worldwide variation in B allele frequency

Lewontin (1995) *Human Diversity*

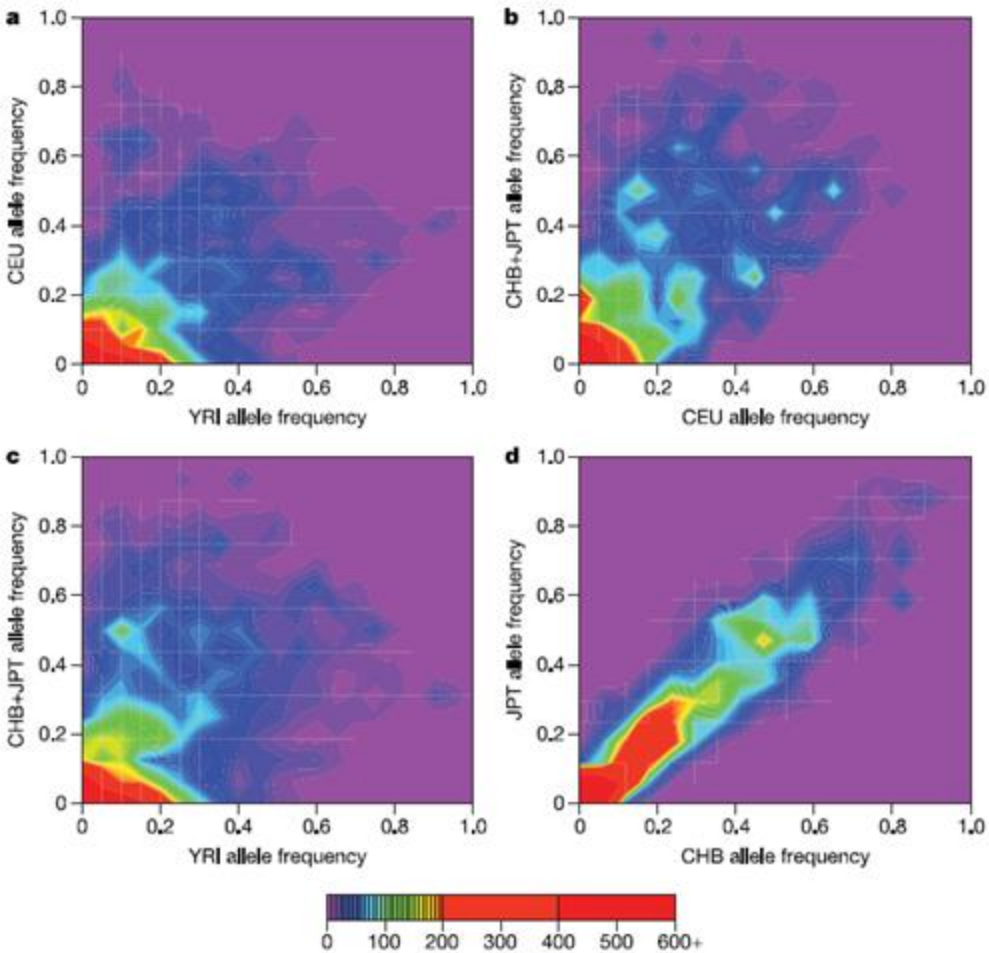
Allele frequency variation at the (more) local scale

Clines in the percentage of blood group A in modern Japan, radiating out from the region just across the Tsushima strait from Korea.



Lewontin (1995) *Human Diversity*

Allele frequency across 4 reference populations by sequencing targeted regions of the genome*



Scatterplots of allele frequencies (of the minor allele across all 3 panels) for SNPs in ENCODE regions in pairs of HapMap reference panels

CEU=CEPH Europeans in Utah
YRI=Yoruba in Ibadan, Nigeria
JPT=Japanese in Tokyo
CHB=Han Chinese in Beijing

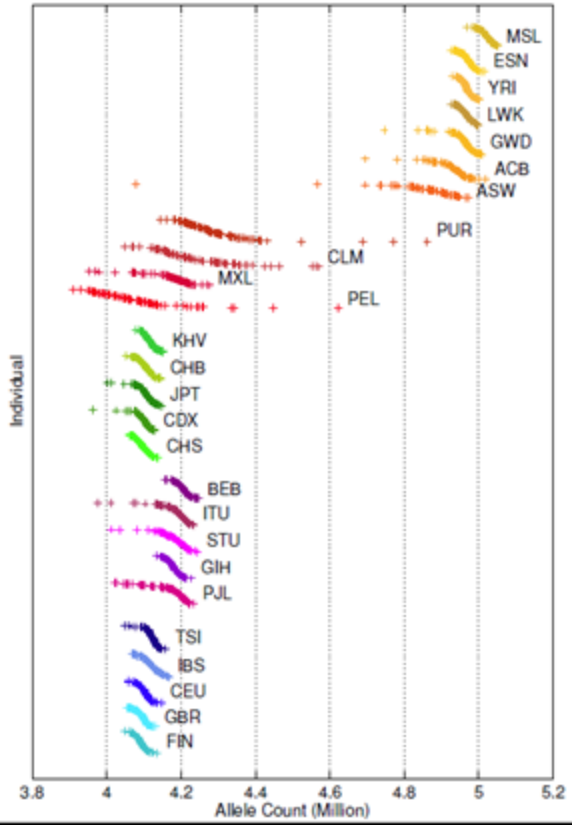
International HapMap Consortium (2005) *Nature*

*regions from the ENCODE project; reference populations from the HapMap project

Critical concept for genetic epidemiology of humans: There is much greater variation, e.g. more SNPs, among African populations than others, consistent with out of Africa hypothesis

Variants per genome

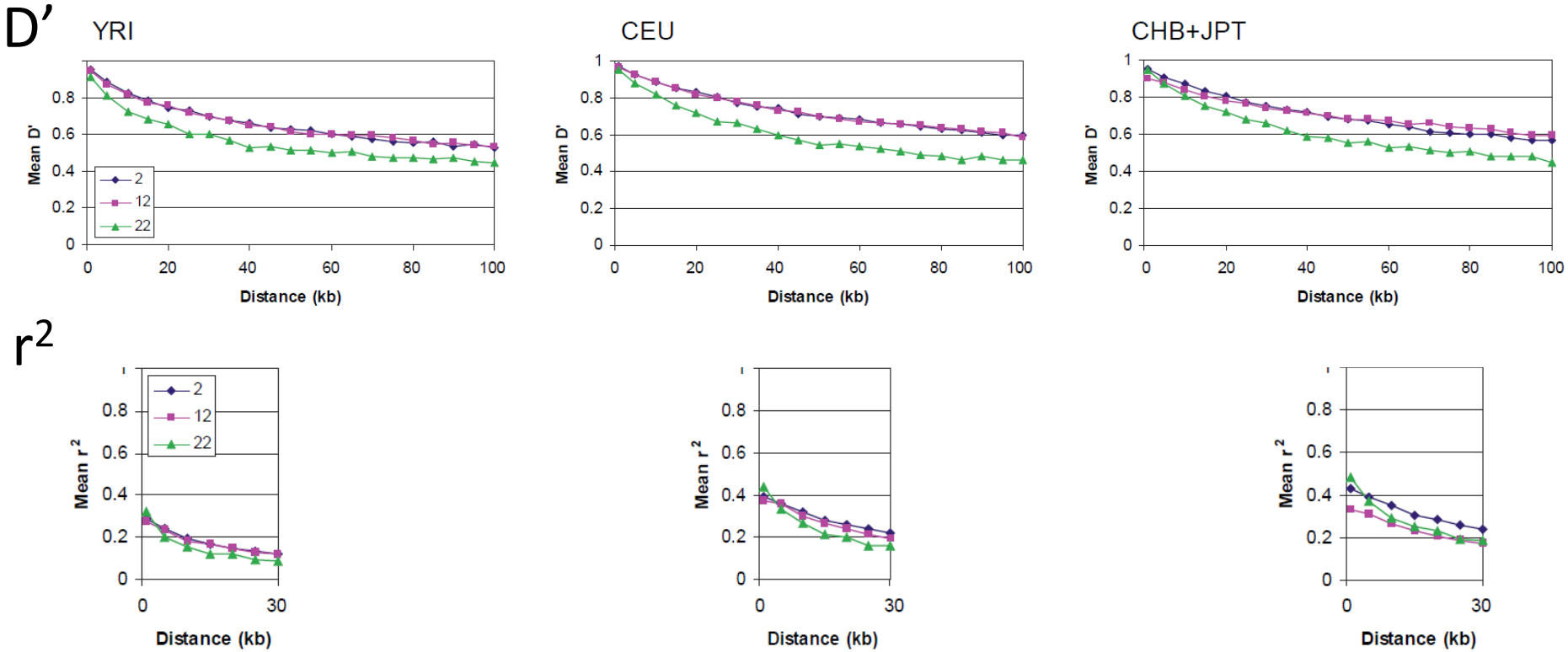
Type	Variant sites / genome
SNPs	$3.8 * 10^6$
Indels	$5.7 * 10^5$
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11



ASHG 2014 Abecasis

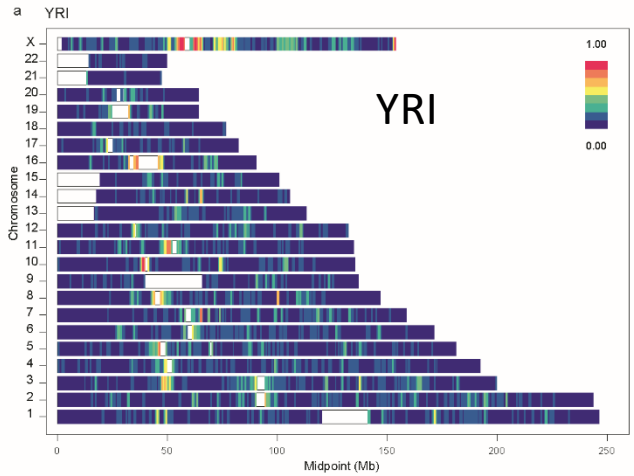
Linkage disequilibrium also varies across populations

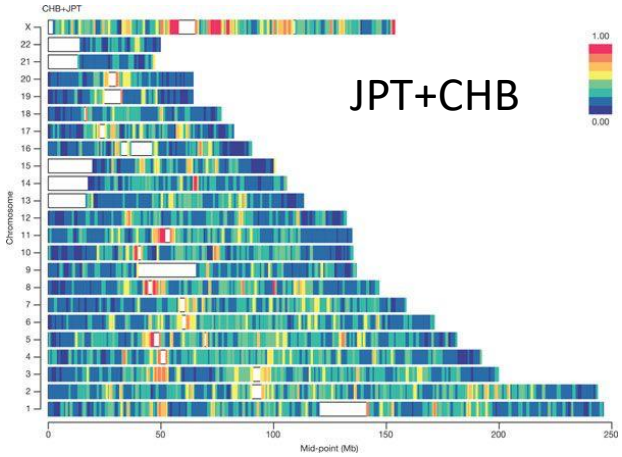
Differences in decay of LD with distance in HapMap samples

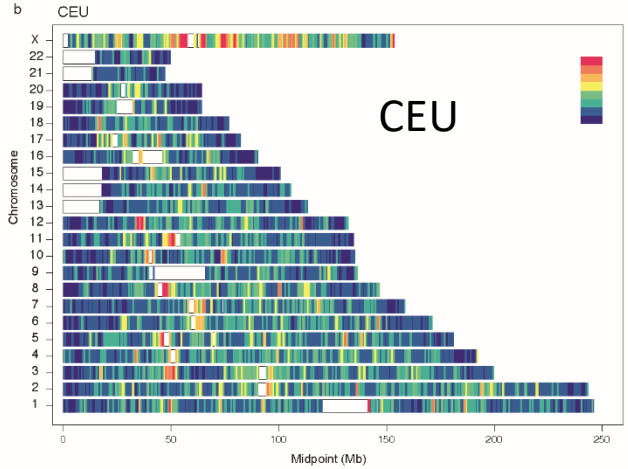


- Mean pairwise D' (top) or r^2 (bottom) by SNP-SNP distance for chromosomes 2 (long), 12 (medium), or 22 (short)
- LD persists longer in CEU (European) and CHB+JPT (Asian) than in YRI (African), especially visible for r^2 measure
- See *Nature* 437:1299 (2005), figure S6

Much less LD in African samples







Authors fit model for the local amount of LD decay (as r^2 in colors) in 30kb in YRI, CEU, JPT+CHB

Much lower values of r^2, i.e. more LD decay per 30kb in YRI than in the others

See *Nature* 437:1299 (2005),

What are some potential reasons/mechanisms for the wide variation in allele frequency and linkage disequilibrium?

A. Stochastic drift with time
B. Population migration, isolation, bottlenecks, or founder effects
C. Selection phenomena – positive or negative
D. Non-random mating
E. Mutation

The time scales for these effects are long, and they can be technically difficult to follow longitudinally. Nevertheless, we can infer something about the past from distributions of current allele frequency.

These differences in allele frequencies and linkage disequilibrium patterns can be used to map samples in terms of genetic similarity. These “map coordinates” can then be used to perform quality control or as covariates to account for potential confounding.

Model-based approaches: STRUCTURE, GRAF-pop

Model-free approach: PCA

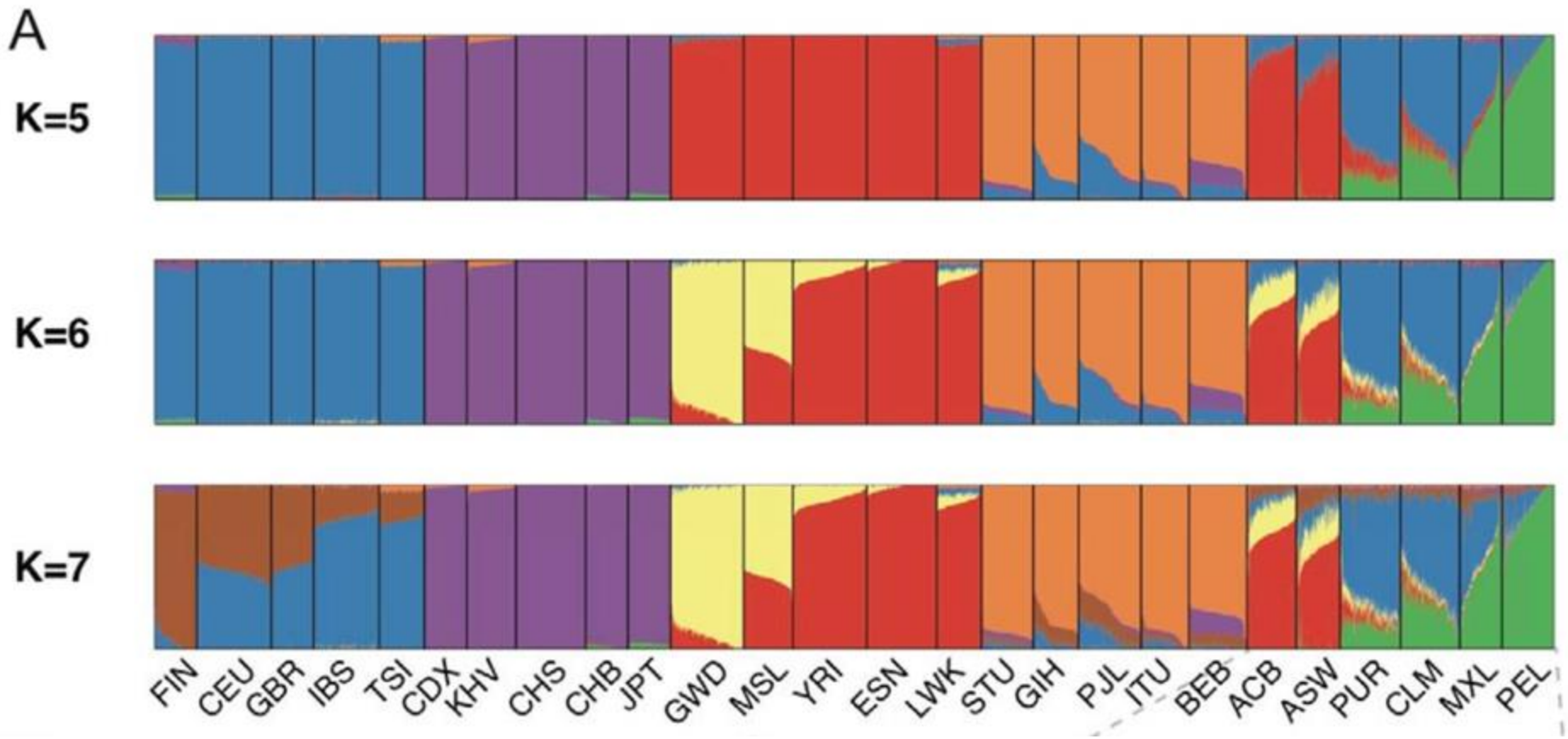
STRUCTURE (GRAF-Pop uses a similar conceptual model)

STRUCTURE calculates the proportion of each individual's genome that "comes from" each of K subpopulations (assuming each is in HWE).

- Underlying model: each allele in an individual can be traced back to one of K ancestral populations
- Under this model, we can estimate the proportion of the individual's alleles that come from each ancestral population.
- Can be run in a supervised fashion, where reference populations are the K subpopulations
- Can be run in an unsupervised fashion, where the subpopulations are inferred from the data, and these are calculated for different values of K (various heuristics for selecting the "optimal" K)

Rosenberg, 2002, Science

STRUCTURE on 1KG



https://www.biorxiv.org/content/10.1101/070797v1.full

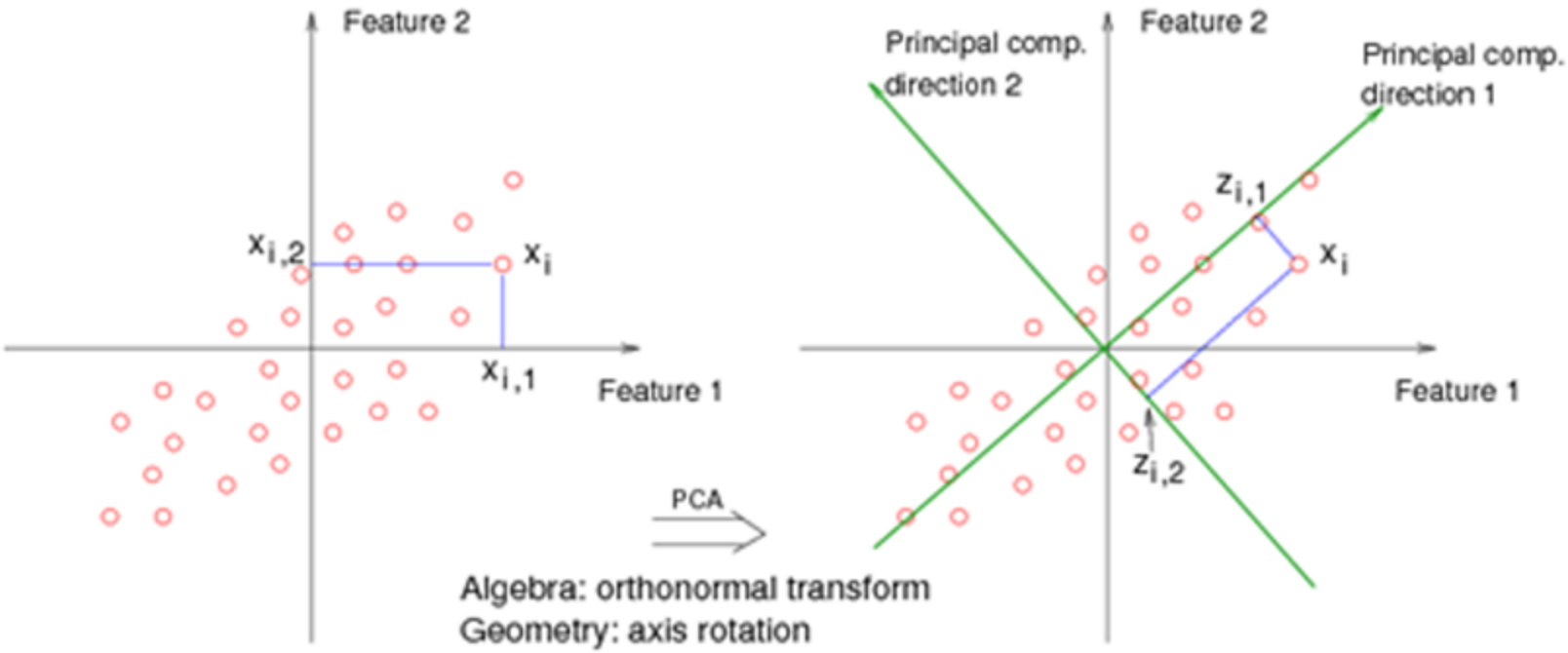
STRUCTURE

Don’t overinterpret these proportions! They depend on the reference samples (or the diversity in the sample at hand). And the “populations” do not necessarily imply shared ancestry.

Further reading: *A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots,* Lawson et al, Nature Communications, 2018

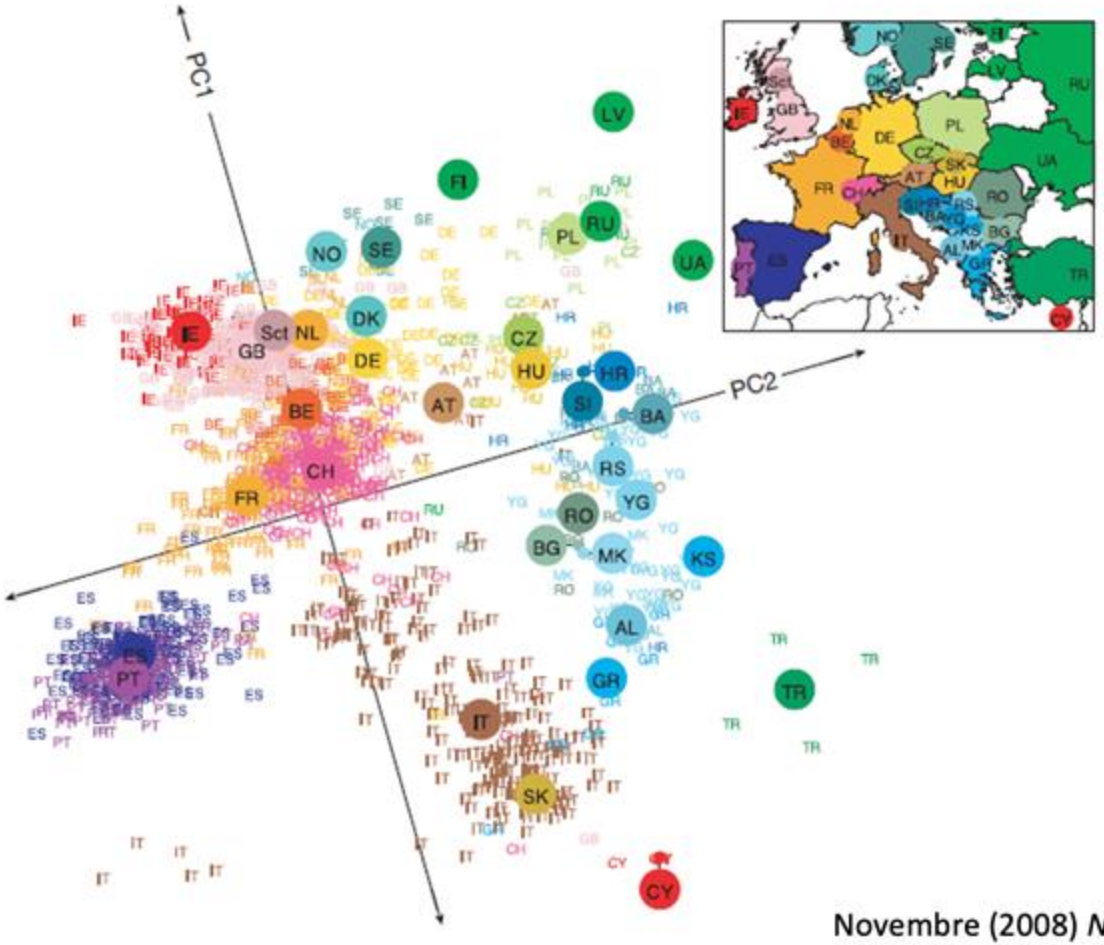
Principal Components (PCs)

Principal components analysis is a general statistical technique that can be used to summarize variation in many variables using a few key summary variables.



Principal Components (PCs) applied to genetic data

When applied to genetic data, PCA can capture latent structure that accounts for patterns of genetic similarity in a set of samples



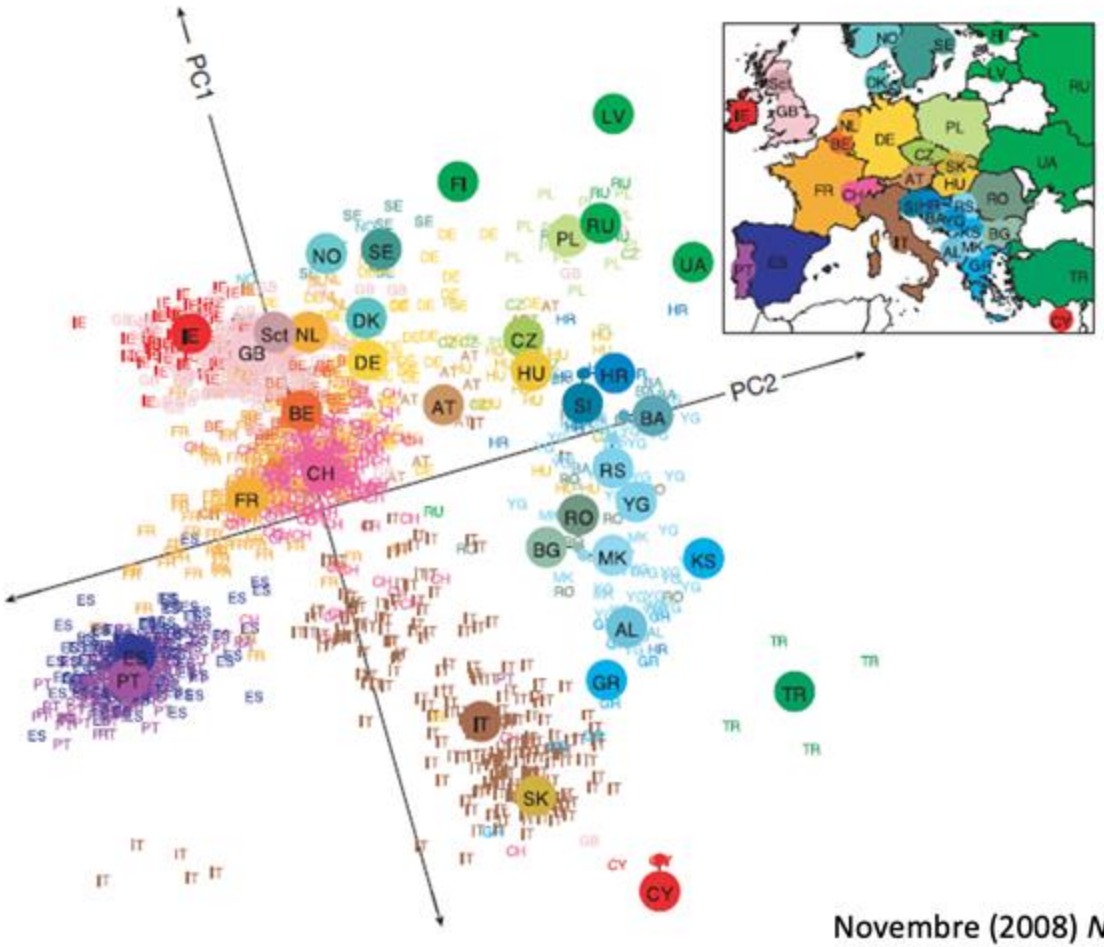
Novembre (2008) *Nature*

Principal Components (PCs) limitations

Nota bene: the picture from the Novembre paper is (by design) “too pretty”

Methods:

“We used a ‘strict consensus’ approach: if all observed grandparents originated from a single country, we used that country as the origin. If an individual’s observed grandparents originated from different countries, we excluded the individual.”



Novembre (2008) *Nature*

Principal Components (PCs) limitations

Remember PCs are just a summary of the data!

- Who is in the data?
- Who is not in the data?

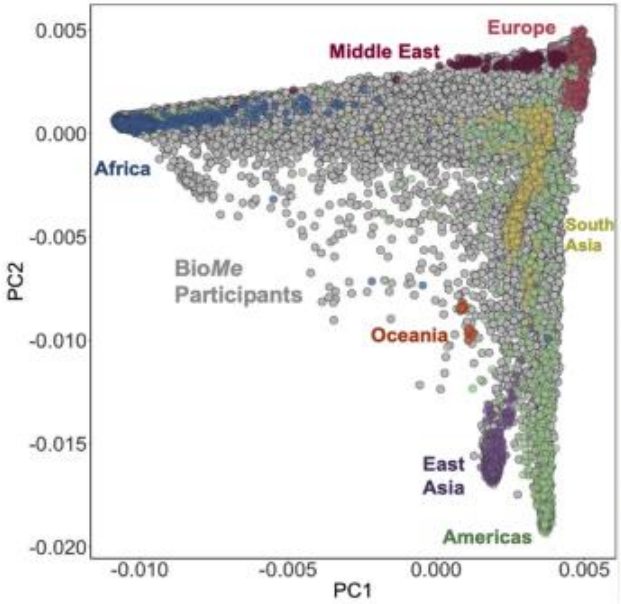


Figure 2. The continuous, category free, nature of genetic variation. *Reproduced from Belbin et al, Towards a fine-scale population health monitoring system. This image shows individuals projected onto the first two principal components of genetic similarity. Colored dots are N=4149 reference panel individuals from 87 populations representing ancestry from 7 continental or subcontinental regions. Gray dots are N=31705 participants from BioMe, a diverse biobank based in New York City. Clearly delineated continental ancestry categories, the islands of color, are shown to be a by-product of sampling strategy. They are not reflective of the diversity in this real-world dataset, made evident by the continuous sea of gray.*

Lewis (2022) *Science*

Enough talk… on to hands-on practice!

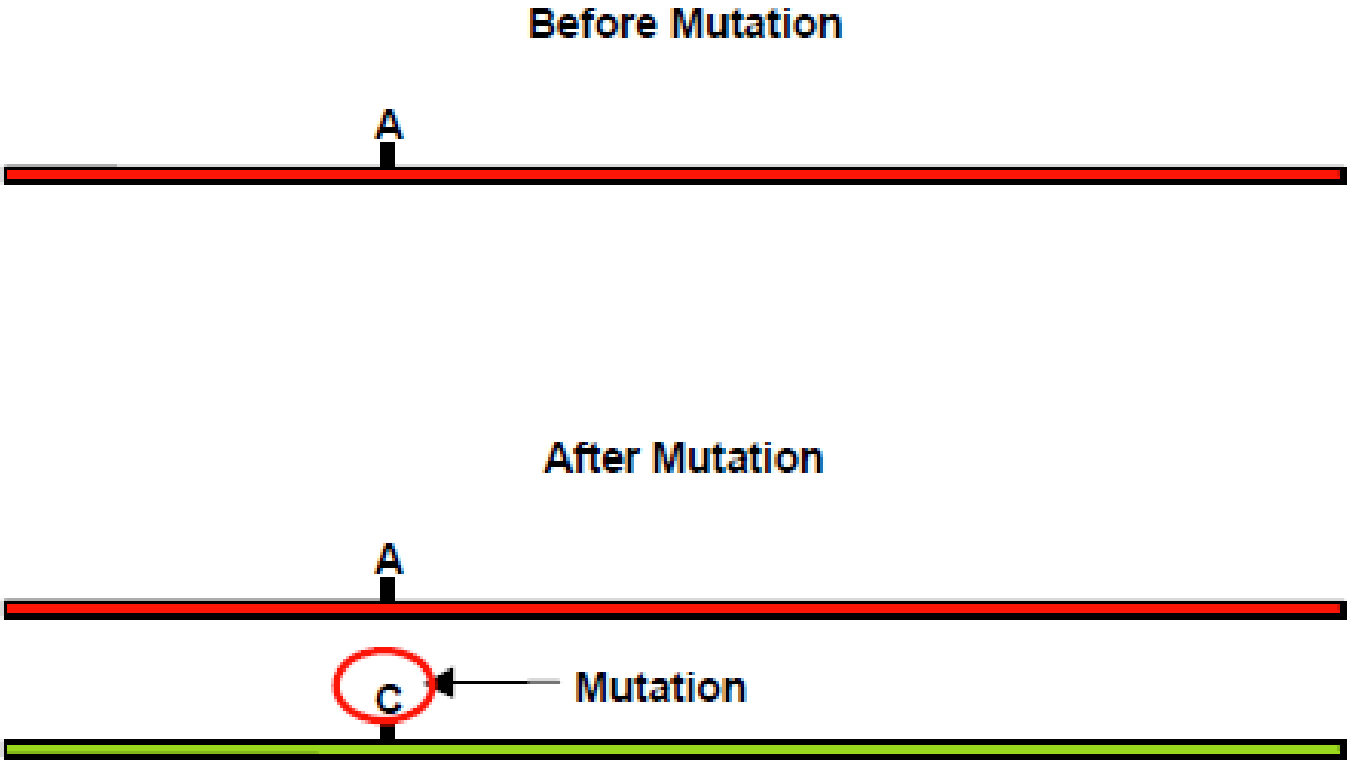
After a break…

Appendix: details on measures of LD

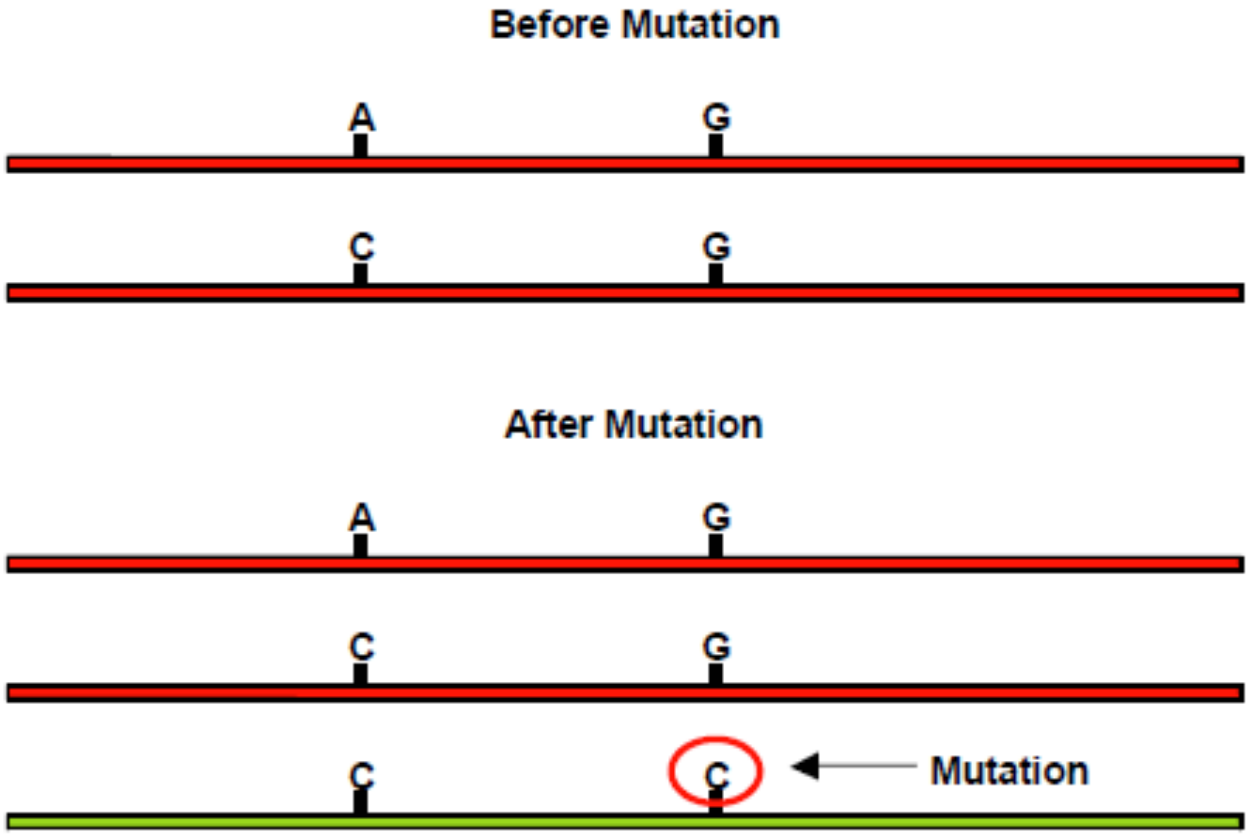
Part 1: Haplotypes and their relationship to LD

Origins of LD

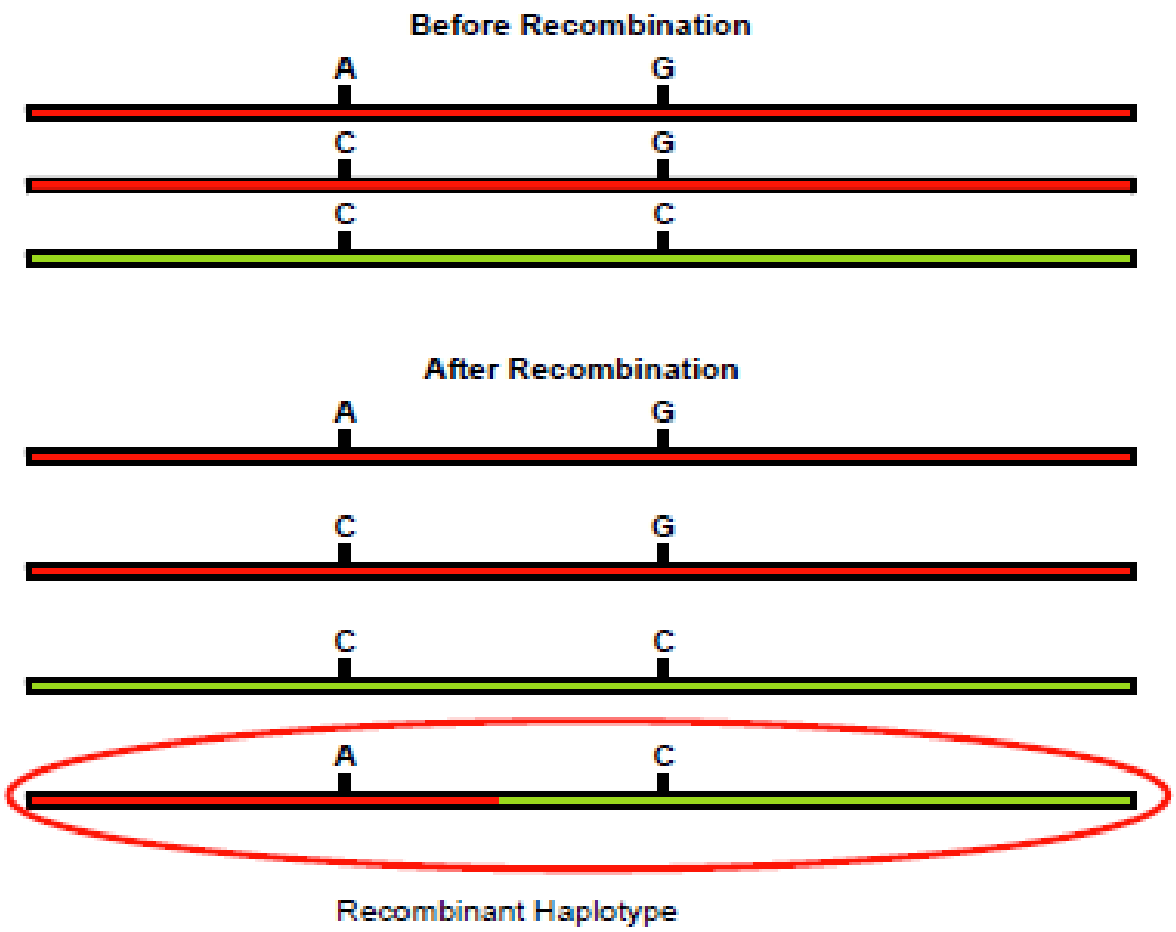
Alleles that exist today reflect ancient genetic events … first one mutation arises



… then the another allele…

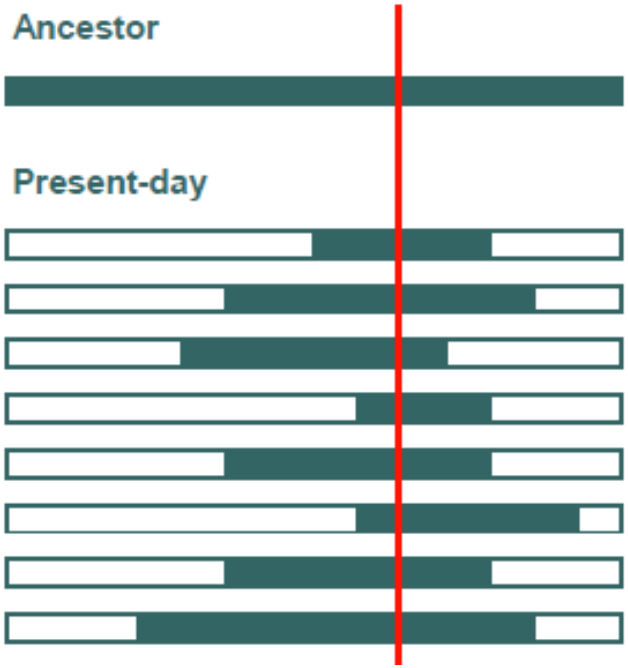


… while recombination (from meiosis) generates new arrangements for ancestral alleles



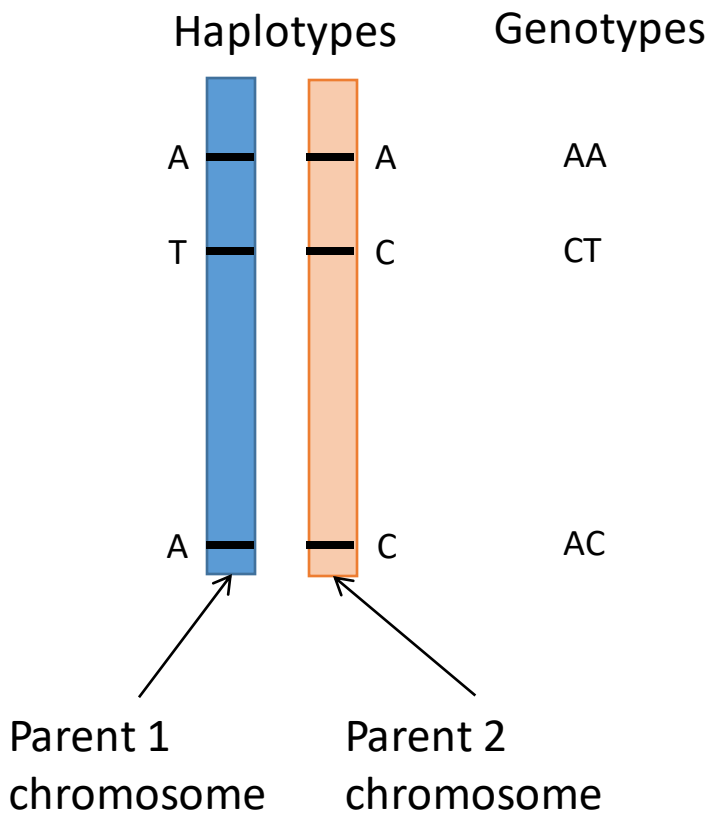
Chromosomes are mosaics of ancestral haplotypes

- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection
- Combinations of alleles at very close markers tend to reflect individual ancestral <u>haplotypes</u>
- Longer range combinations of alleles reflect recombination of ancestral <u>haplotypes</u>



Haplotypes formally defined

Haplotypes are the ordered arrangements of alleles at specific loci <u>along the same chromosome</u>. They are **<u>phased</u>** genotypes, that is, the alleles as they are passed down from parents.



Haplotypes are typically not directly observed by genotyping

- Genotype data typically does not include information about phase of alleles
- Can determine by long-range single molecule sequencing – but experimentally challenging
- More typically, have to infer haplotypes from genotype data

  Example --
  - Genotypes Aa/Gg consistent with haplotype pairs
  - Can estimate haplotypes in a probabilistic framework

$$\begin{matrix} A & a \\ G & g \end{matrix} \quad \text{or} \quad \begin{matrix} A & a \\ g & G \end{matrix}$$

e.g. $\begin{matrix} A & a \\ G & g \end{matrix}$ with probability=0.9 $\begin{matrix} A & a \\ g & G \end{matrix}$ with probability=0.1

- Most popular algorithm: Expectation Maximization[1]
  - Start with a guess for haplotype frequencies
  - Given this guess, calculate expected haplotype counts, assuming haplotypes are in HWE
  - Use these counts to update haplotype frequency estimates
  - Repeat until convergence

[1] Thomas pp. 243-245

Comment about the definition LD

- Two loci are described as being LD if:
  A. Alleles at the two loci are not independently distributed
  B. The two loci are linked (recombination rate $\theta<½$)
- More precisely
  - Condition (A) is called "gametic disequilibrium"
  - Condition (B) is called "linkage"
  - Often people say LD when they mean gametic disequilibrium
- Thus, the standard measures of LD are really measures of gametic disequilibrium
- However, we usually do not compute LD for markers at different chromosomes, and the terms are equivalent for markers on the same chromosome

Part 2: Estimating and quantifying LD

First step in quantifying LD: From genotypes to haplotypes

Two nearby loci in sample of 8 individuals

a	g	a	g	A	g	A	G
A	G	A	G	A	G	A	G
a	g	a	g	A	G	A	G
A	g	A	G	a	g	a	g

Genotypes (data)

	AA	Aa	aa
GG	1	0	0
Gg	1	5	0
gg	0	1	0

Low counts implies low numbers of aG haps

EM algorithm →

Haplotypes (inferred)

	A	a	
G	8	0	50%
g	2	6	50%
	62.5%	37.5%	

No inferred aG haps

Only ambiguous haplotypes. Why?

δ (aka D): a measure of LD related to marginal allele frequencies

- Consider δ defined as follows:

	A	**a**	
G	$p_{AG}=p_A p_G + \delta$	$p_{aG}=p_a p_G - \delta$	p_G
g	$p_{Ag}=p_A p_g - \delta$	$p_{ag}=p_a p_g + \delta$	p_g
	p_A	p_a	1

- p_A & p_G are frequencies of alleles A & G
  - $p_a = 1-p_A$; $p_g = 1-p_G$ allele frequencies of alleles a & g
- δ is a measure of departure from independence
  - No association between A and G $\Rightarrow \delta = 0$
  - The four cells in this table must be between 0 and 1, so:
  - $\delta_{min} = -\min(p_A\, p_G,\ p_a\, p_g)$ and $\delta_{max} = \min(p_A\, p_g,\ p_a\, p_G)$
- δ also sometimes termed "D"

Why is δ a measure of dependence of alleles at two loci?

- Basic idea: two random variables are independent if knowing one gives you no information about the other, i.e. conditioning on one does not change the distribution of the other
- The mathematical definition of independence is

$$\Pr(A,G)=\Pr(A)\Pr(G)$$

- Why is this a good definition? Well, if independent

$$\Pr(A\,|\,G)=\frac{\Pr(A,G)}{\Pr(G)}=\frac{\Pr(A)\Pr(G)}{\Pr(G)}=\Pr(A)$$

- δ is just Pr(A,G)-Pr(A)Pr(G), which equals 0 if no LD.

D’ and r^2: more interpretable LD measures related to δ

δ depends on allele frequencies and may be positive or negative making it difficult to interpret.

Since the marginal frequencies $p_A = p_{AG} + p_{Ag}$ and $p_G = p_{AG} + p_{aG}$:

$$\delta = p_{AG} - p_A p_G = p_{AG} p_{ag} - p_{Ag} p_{aG}$$

Consider:

Measure	Formula
D’ = $\delta/\max(\delta)$	$\frac{p_{AG}p_{ag} - p_{Ag}p_{aG}}{\min(p_A p_G, p_a p_g)}$ if $\delta<0$ $\frac{p_{AG}p_{ag} - p_{Ag}p_{aG}}{\min(p_A p_g, p_a p_G)}$ if $\delta>0$
r^2 (aka Δ^2)	$\frac{(p_{AG}p_{ag} - p_{Ag}p_{aG})^2}{p_A q_A p_G q_G} = \left(\frac{p_{AG} - p_A p_G}{\sqrt{p_A q_A p_G q_G}}\right)^2$

Comments on D' and r^2

- abs(D') = 1 is taken as evidence for no recombination, in the absence of other influences on LD.
- The LD measure "r" is **identical** to the Pearson correlation between alleles at locus X (e.g. A or a) and Y (e.g. G or g), i.e.

$$r = \sqrt{\frac{1}{N}\sum \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_y}},$$

$where$:

$$X_i = allele\ at\ locus\ X = \begin{cases} 0\ if\ X = A \\ 1\ if\ X = a \end{cases}$$

$$Y_i = allele\ at\ locus\ Y = \begin{cases} 0\ if\ Y = G \\ 1\ if\ Y = g \end{cases}$$

- Computation of LD requires estimating frequency of two-locus haplotypes p_{AG}, p_{Ag}, p_{aG}, p_{ag} as on previous slide.

More comments: |D'| and r^2 reflect different properties of LD

- |D'| , absolute value of D prime …
  - ranges from 0 [no LD] to 1 [complete LD]
  - is less sensitive to marginal allele frequencies (D is sensitive but D' is much less sensitive)
  - is directly related to recombination fraction, e.g. D'=1 implies no evidence for historical recombination
- r^2, squared correlation …
  - also ranges from 0 [no LD] to 1
  - r^2=1 if complete LD and allele frequencies the same at two loci [perfect LD]
  - is correlation between alleles on the same chromosome
  - is very sensitive to marginal allele frequencies
    - Far away low frequency variants can have higher r^2 than close common variants (Delvin 1995 Genomics, table 4)
  - is directly related to power for association

D’ and r^2 from haplotype counts instead of frequencies

	A	**a**	
G	n_{11}	n_{10}	$n_{1\bullet}$
g	n_{01}	n_{00}	$n_{0\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 0}$	n

$$\delta = (1/n)^2 (n_{11} n_{00} - n_{10} n_{01})$$

Measure	Formula
D’	$\frac{n_{11}n_{00} - n_{10}n_{01}}{\min(n_{\bullet 1}n_{1\bullet}, n_{\bullet 0}n_{0\bullet})}$ if $\delta<0$ $\frac{n_{11}n_{00} - n_{10}n_{01}}{\min(n_{\bullet 1}n_{0\bullet}, n_{\bullet 0}n_{1\bullet})}$ if $\delta>0$
$\Delta^2 = r^2$	$\frac{(n_{11}n_{00} - n_{10}n_{01})^2}{n_{\bullet 1}n_{\bullet 0}n_{1\bullet}n_{o\bullet}}$

Again r is also Pearson correlation between alleles G and A.

Worked example

Same haplotype table as on previous slide

	A	**a**	
G	8	0	50%
g	2	6	50%
	62.5%	37.5%	

$$\delta = (1/16)^2\ (8\times6 - 2\times0) = 48/256=0.1875 > 0.$$

$$D' = (8\times6 - 0) / \min(8\times6,10\times8) =1$$

$$r^2= (8\times6 - 0)^2 / (10\times6\times8\times8) = 0.6$$

** Tip: If any of the four cells is 0, then |D'|=1. Why? **