Friday Harbor 2017

From Genetics to GWAS
(Genome-wide Association Study)

Sept 7 2017

David Fardo

University of Kentucky

Purpose: prepare for tomorrow's tutorial

- Genetic Variants
- Quality Control
- Imputation
- Association
- Visualization
- Prioritization

OUTLINE

- **Goal**: be able to answer the following questions

- What are some of the historical landmarks of GWAS?
- What is unique about GWAS data and data quality considerations?
- How do you test for genetic association?

TOWARDS GWAS

- Evidence for genetic role ?
  - Population differences
  - Familial aggregation
  - Linkage ?

LINKAGE

Linkage Analysis:
a two cent version

- One cent
  - Use properties of recombination to localize
  - Track transmissions through families

- Second cent
  - Use principle of similarity

  "*Sib-pairs that are phenotypically similar should also be genotypically similar*" -Penrose, 1935
  - Identity by state / descent (IBS/IBD)

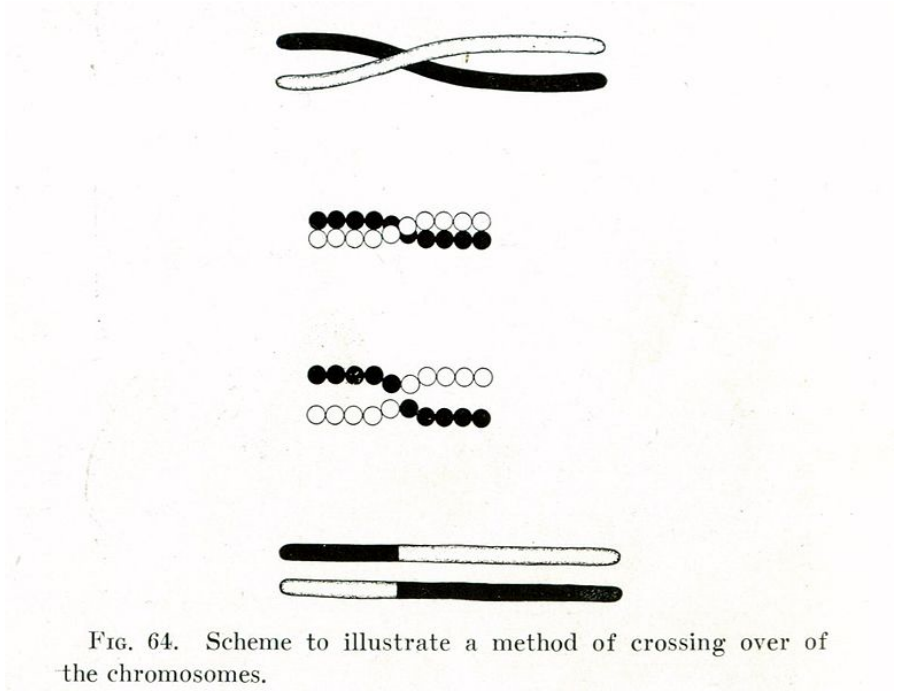


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Thomas Hunt Morgan-

1933 Nobel *"for his discoveries concerning the role played by the chromosome in heredity".*

Recombination

Two Loci: A and B

Two Alleles at each Locus: $\{A_1, A_2\}$, $\{B_1, B_2\}$

Four Possible Haplotypes:

$$\underline{A_1B_1} \quad \underline{A_1B_2} \quad \underline{A_2B_1} \quad \underline{A_2B_2}$$

Ten Possible Diploid Genotypes (sometimes called diplotypes):

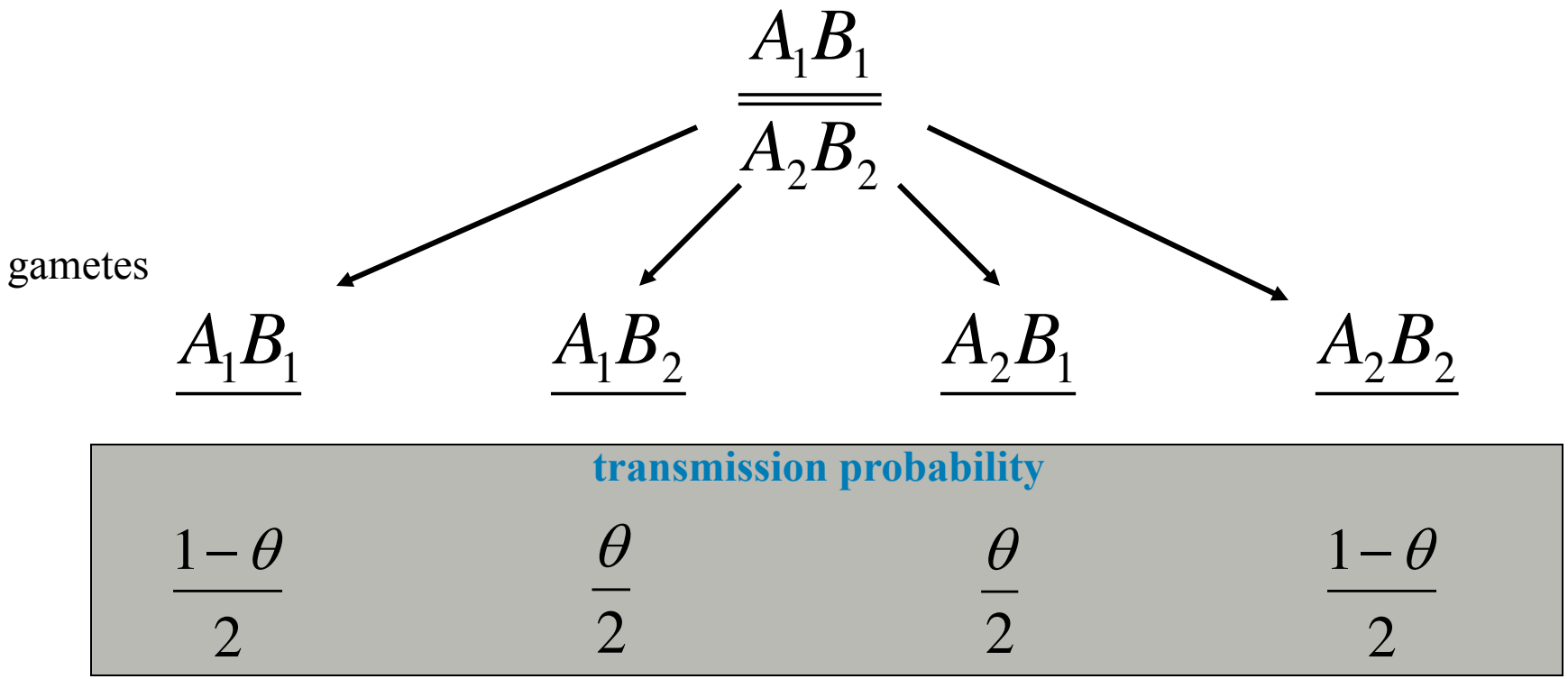
$$\frac{A_1B_1}{A_1B_1} \; \frac{A_1B_1}{A_1B_2} \; \frac{A_1B_1}{A_2B_1} \; \frac{A_1B_1}{A_2B_2} \; \frac{A_1B_2}{A_1B_2} \; \frac{A_1B_2}{A_2B_1} \; \frac{A_1B_2}{A_2B_2} \; \frac{A_2B_1}{A_2B_1} \; \frac{A_2B_1}{A_2B_2} \; \frac{A_2B_2}{A_2B_2}$$

Diploid Genotypes

1	2	3	4	5	6	7	8	9	10
$\frac{A_1B_1}{A_1B_1}$	$\frac{A_1B_1}{A_1B_2}$	$\frac{A_1B_1}{A_2B_1}$	$\frac{A_1B_1}{A_2B_2}$	$\frac{A_1B_2}{A_1B_2}$	$\frac{A_1B_2}{A_2B_1}$	$\frac{A_1B_2}{A_2B_2}$	$\frac{A_2B_1}{A_2B_1}$	$\frac{A_2B_1}{A_2B_2}$	$\frac{A_2B_2}{A_2B_2}$

Recombination only detectable in the **double heterozygotes**

Double Heterozygotes



θ = recombination rate (ranges from 0 to 0.5)

$\theta = 0$: no recombination

$\theta = 0.5$: unlinked

seeblue.

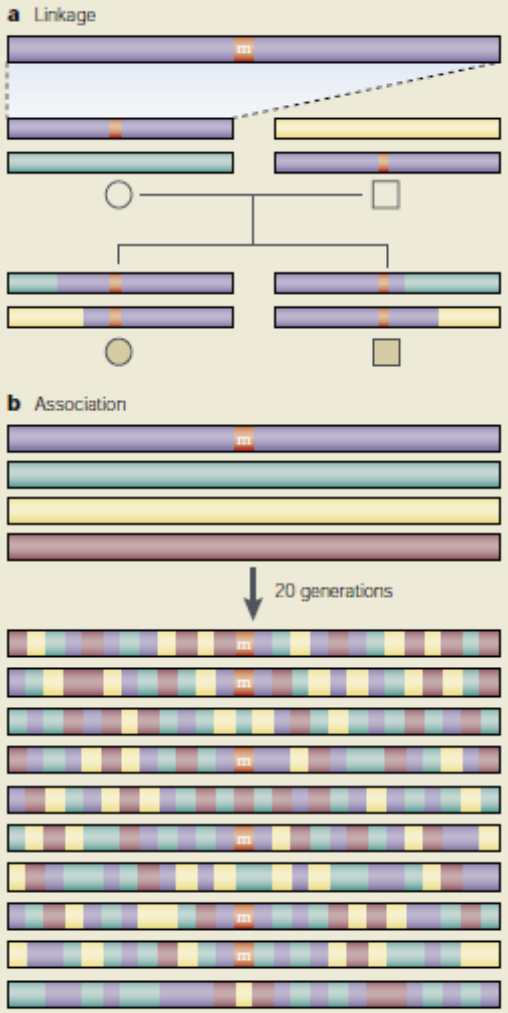
The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

SCIENCE • VOL. 273 • 13 SEPTEMBER 1996

Has the genetic study of complex disorders reached its limits? The persistent lack of replicability of these reports of linkage between various loci and complex diseases might imply that it has. We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

Box 2 | **Linkage versus association**



At a fundamental level, genetic association and linkage analysis rely on similar principles and assumptions[87]. Both rely on the co-inheritance of adjacent DNA variants, with linkage capitalizing on this by identifying haplotypes that are inherited intact over several generations (such as in families or pedigrees of known ancestry), and association relying on the retention of adjacent DNA variants over many generations (in historic ancestries). Thus, association studies can be regarded as very large linkage studies of unobserved, hypothetical pedigrees. In growing populations, such as humans, recombination is the primary force that eliminates linkage and association over generations[88]. When a functional mutation occurs ('m' in the figure) — perhaps one that contributes to disease — it does so on a haplotype of other pre-existing DNA variants. Because linkage focuses only on recent, usually observable ancestry, in whom there have been relatively few opportunities for recombination to occur, disease gene regions that are identified by linkage will often be large, and can encompass hundreds or even thousands of possible genes across many megabases of DNA (figure panel **a**). By contrast, association studies draw from historic recombination so disease-associated regions are (theoretically) extremely small in outbred random mating populations[89], encompassing only one gene or gene fragment (figure panel **b**). Through subsequent generations, as the disease mutation is transmitted, recombination will cause it to be separated from the specific alleles of its original haplotype. Particular DNA variants can remain together on ancestral haplotypes for many generations. This type of non-random association of alleles is known as linkage disequilibrium. It is linkage disequilibrium that provides the genetic basis for most association strategies.

genetic association and linkage analysis rely on similar principles and assumptions[87]. Both rely on the co-inheritance of adjacent DNA variants, with linkage capitalizing on this by identifying haplotypes that are inherited intact over several generations (such as in families or pedigrees of known ancestry), and association relying on the retention of adjacent DNA variants over many generations (in historic ancestries). Thus, association studies can be regarded as very large linkage studies of unobserved, hypothetical pedigrees.

Cardon & Bell, Nat Rev Gen (2001)

LINKAGE DISEQUILIBRIUM (LD)

A definition

Linkage Disequilibrium – allelic association between two genetic loci

What you need to know about LD

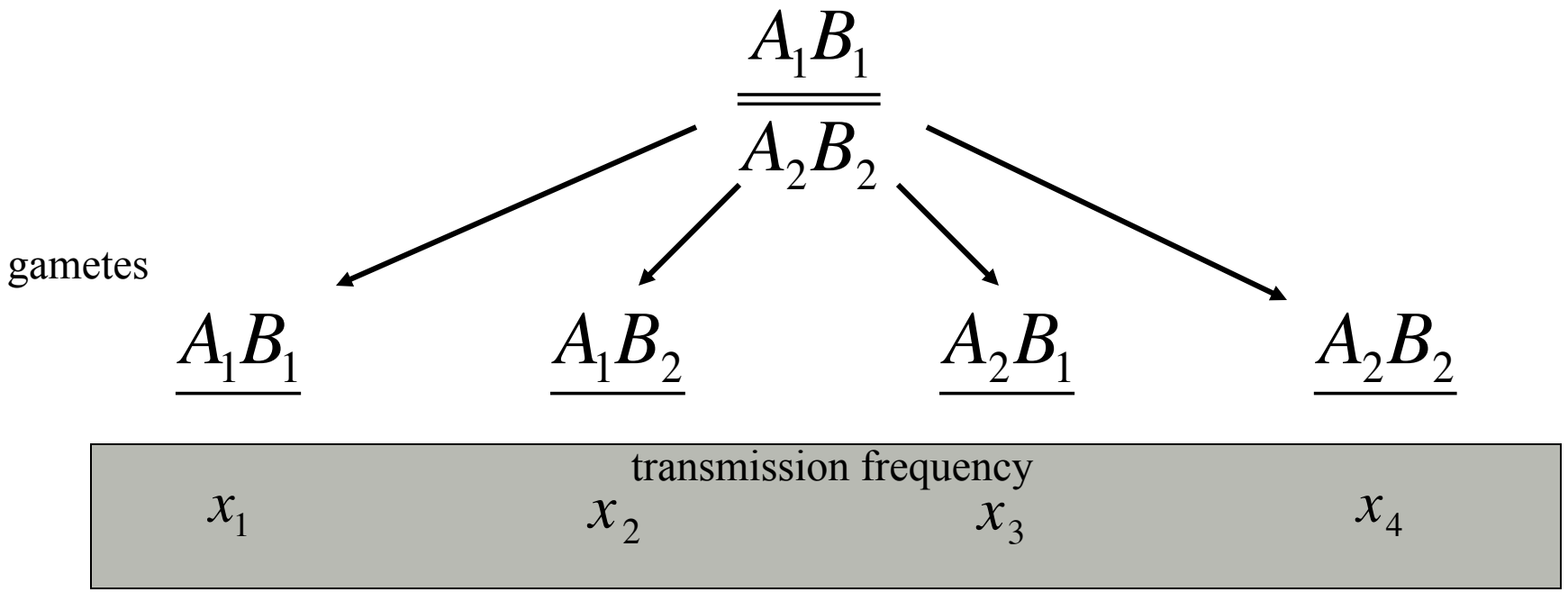
- It can be defined several ways mathematically, each definition with its own pros/cons

  (I will show a couple *briefly*)

- It degrades over generations

- Its properties are used for GWAS

Linkage Disequilibrium



Linkage Disequilibrium

Gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency	x_1	x_2	x_3	x_4

Allele	A_1	A_2	B_1	B_2
Frequency	$p_{A1}=x_1+x_2$	$p_{A2}=x_3+x_4$	$p_{B1}=x_1+x_3$	$p_{B2}=x_2+x_4$

$$D = Observed - Expected$$

$$D = x_1 - p_{A1}p_{B1}$$

$$D = x_1 - (x_1 + x_2)(x_1 + x_3)$$

$$D = x_1x_4 - x_2x_3$$

Linkage Disequilibrium

After one generation of random mating:

$$x_1' = x_1 - \theta D$$
$$x_2' = x_2 + \theta D$$
$$x_3' = x_3 + \theta D$$
$$x_4' = x_4 - \theta D$$

$$D_{t=1} = x_1' x_4' - x_2' x_3'$$
$$D_{t=1} = (1-\theta) D$$

After t generations:

$$D_t = (1-\theta)^t D_0$$

What does this mean?

$$D_t = (1-\theta)^t D_0$$

D_0	theta	t	D_{10}
1	0.5	10	0.001
1	0.1	10	0.35

Normalized LD Parameters

$$D' = \frac{D}{D_{\max}}$$

D_{max} $= min(p_{A1}p_{B2}, p_{A2}p_{B1})$ *if D is positive*
$= min(p_{A1}p_{B1}, p_{A2}p_{B2})$ *if D is negative*

Now, LD ranges from -1 to +1

r^2

Most commonly used LD measure
-- squared correlation coefficient --

$$r^2 = \frac{D^2}{p_{A1}p_{A2}p_{B1}p_{B2}}$$

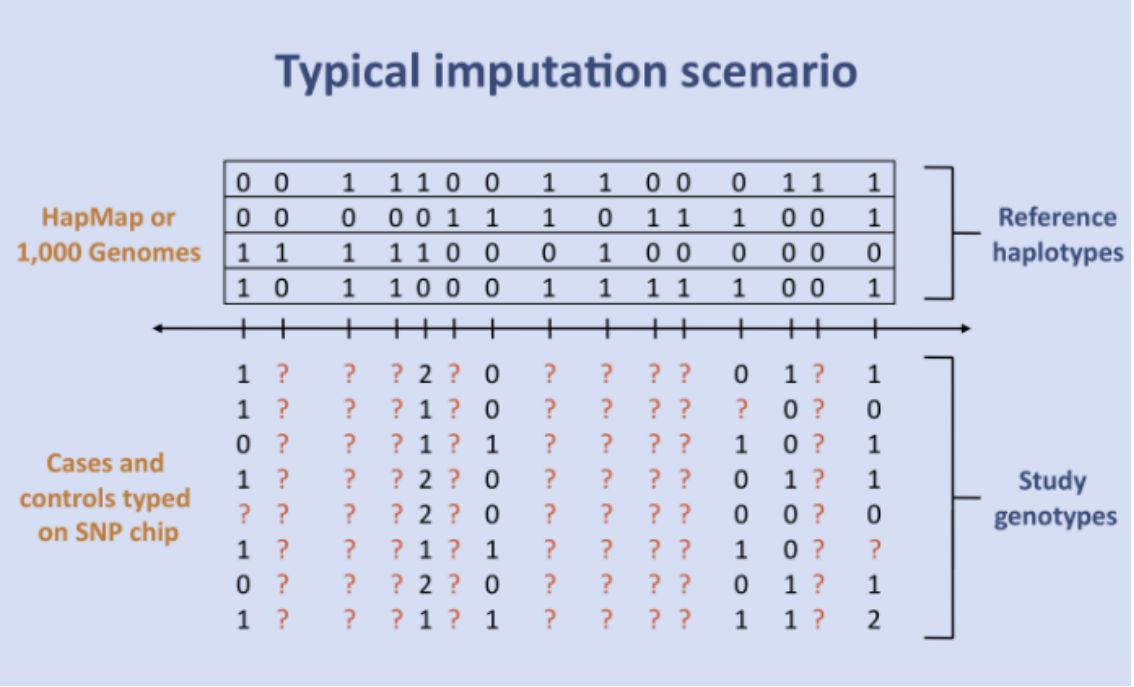
LD take home points

- It can be defined several ways mathematically, each definition with its own pros/cons
- It degrades over generations
- Its properties are used for GWAS

IMPUTATION

Historical context of some large-scale initiatives → towards imputation

- Human Genome Project
  - 2003 (kind of)
  - 2 males, 2 females
- HapMap
  - 2005 / 2007 / 2009
  - initially 269; expanded to ~1400
- 1000 Genomes Project
  - 2010 / 2012 / 2015
  - guess?
- Haplotype Reference Consortium
  - 2016
  - 1st release is ~65k haplotypes
- All of Us (PMI Initiative) ?



http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

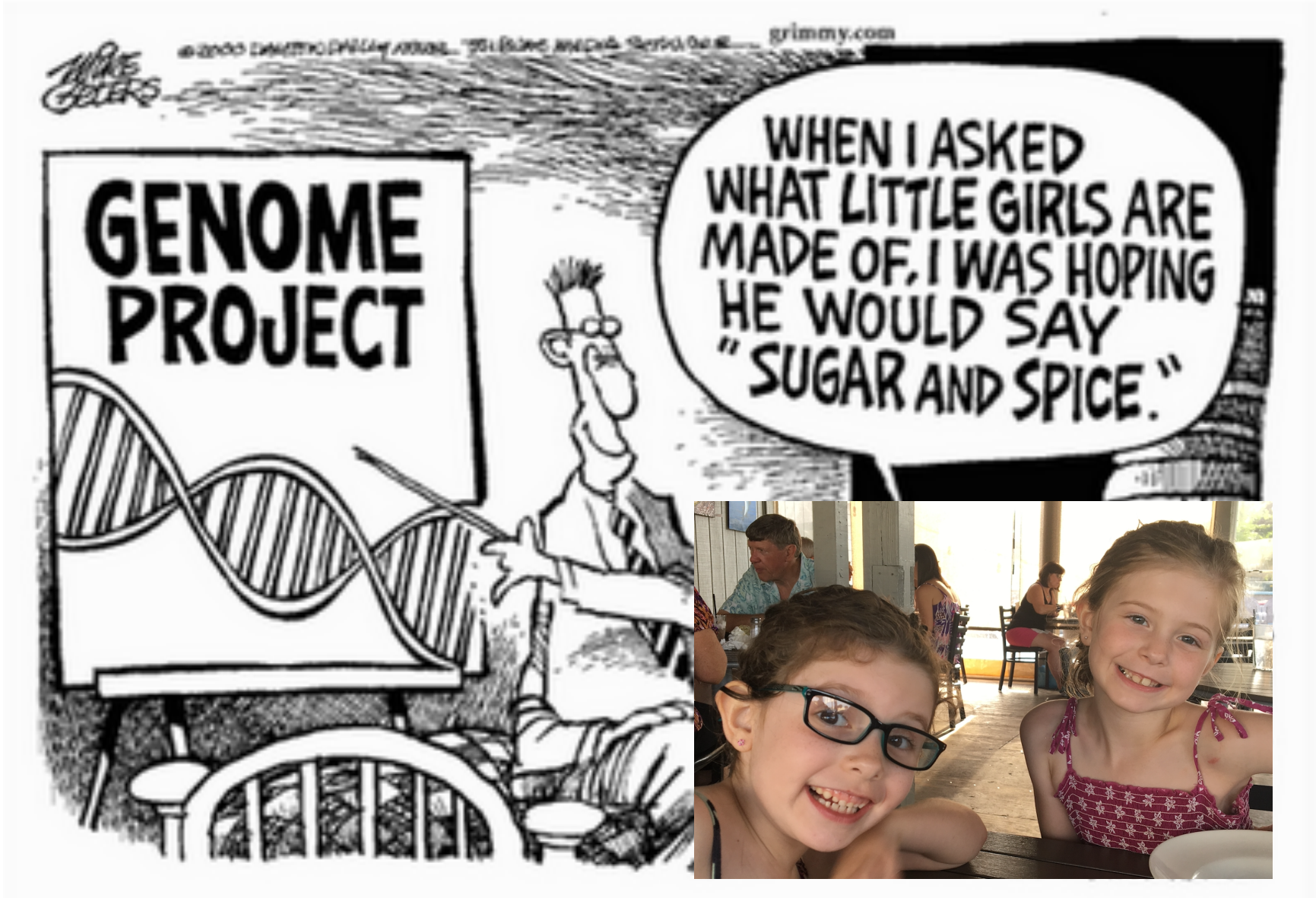
Human Genome Project

Goals:

- identify all the **approximate 30,000 genes** in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Milestones:

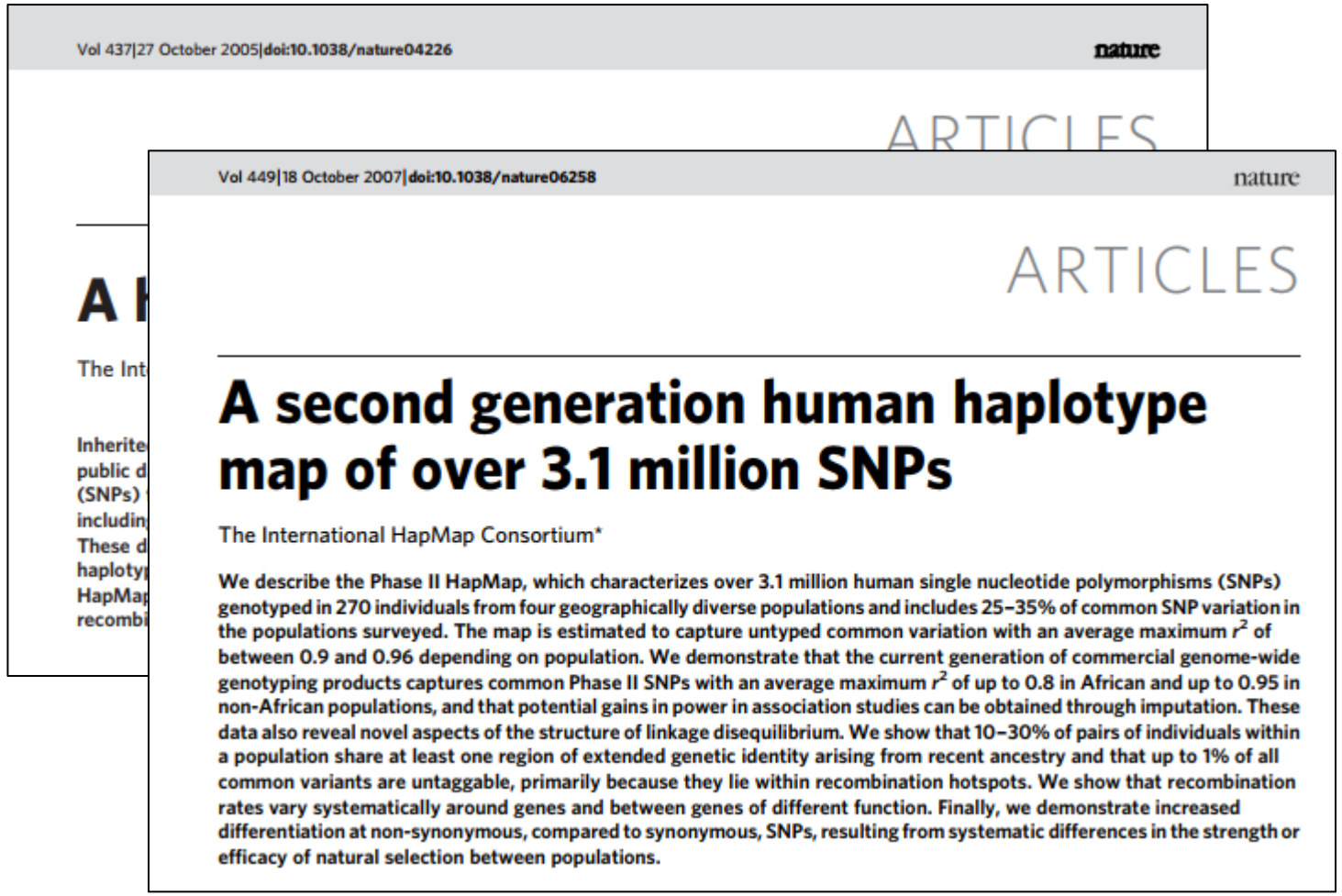
- 1990: Project initiated as joint effort of U.S. Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and Project is **declared finished** two years ahead of schedule


grimmy.com
GENOME PROJECT
WHEN I ASKED WHAT LITTLE GIRLS ARE MADE OF, I WAS HOPING HE WOULD SAY "SUGAR AND SPICE."

HapMap

An NIH program to chart genetic variation within the human genome

- Begun in 2002, the project is a 3-year effort to construct a map of the patterns of SNPs (single nucleotide polymorphisms) that occur across populations in Africa, Asia, and the United States.
- Consortium of researchers from six countries
- Researchers hope that dramatically decreasing the number of individual SNPs to be scanned will provide a **shortcut for identifying the DNA regions** associated with common complex diseases
- Map may also be useful in understanding how genetic variation contributes to responses in environmental factors



Vol 437|27 October 2005|doi:10.1038/nature04226 nature

ARTICLES

Vol 449|18 October 2007|doi:10.1038/nature06258 nature

ARTICLES

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum r^2 of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10–30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.

What is the 1000 Genomes Project ?

- International multi-center collaboration building on HapMap data to establish the most comprehensive catalogue of human genetic variation available
- Phase I: **1,092 complete genomes** from **14 populations** published in *Nature*, October 2012
- Freely accessible public databases
- Final phase of project brings total genotyped to **2504 individuals** from **26 populations** worldwide

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequ
for investigating the relationship between genotype and phenotype. Here we present r
project, designed to develop and compare different strategies for genome-wide sequ
platforms. We undertook three projects: low-coverage whole-genome sequencing
populations; high-coverage sequencing of two mother–father–child trios; and exon
individuals from seven populations. We describe the location, allele frequency and
approximately 15 million single nucleotide polymorphisms, 1 million short insertion
structural variants, most of which were previously undescribed. We show that, becaus
majority of common variation, over 95% of the currently accessible variants found in an
data set. On average, each person is found to carry approximately 250 to 300 loss-of-f
genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrat
to inform association and functional studies. From the two trios, we directly estimate th
substitution mutations to be approximately 10^{-8} per base pair per generation. We ex
signatures of natural selection, and identify a marked reduction of genetic variation in
due to selection at linked sites. These methods and public data will support the next pha

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide
applying whole-genome sequencing to a div
completion of the project, having reconstructed
tion of low-coverage whole-genome sequenc
characterized a broad spectrum of genetic var
polymorphisms (SNPs), 3.6 million short inse
onto high-quality haplotypes. This resource in
ancestries. We describe the distribution of gene
common disease studies.

ARTICLE

OPEN

doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies
impact and disease association.

From Where?

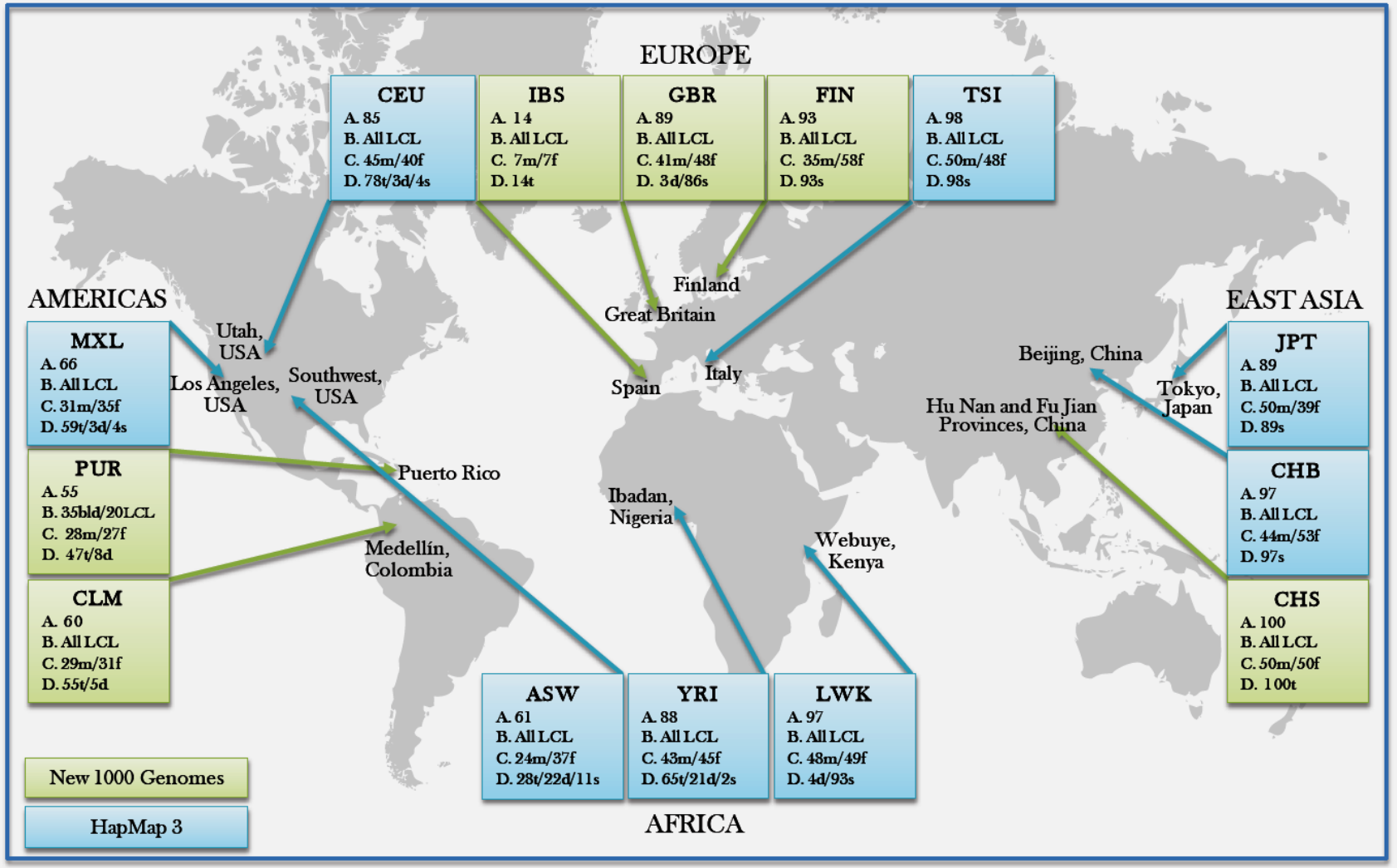


Figure S2. 1000 Genomes Project Phase I populations

Haplotype Reference Consortium

LETTERS

nature genetics

A reference panel of 64,976 haplotypes for genotype imputation

Shane McCarthy[1,98], Sayantan Das[2,3,98], Warren Kretzschmar[4,98], Olivier Delaneau[5], Andrew R Wood[6], Alexander Teumer[7,8], Hyun Min Kang[2,3], Christian Fuchsberger[2,3], Petr Danecek[9], Kevin Sharp[10], Yang Luo[1], Carlo Sidore[11], Alan Kwong[2,3], Nicholas Timpson[12], Seppo Koskinen[13], Scott Vrieze[14,15], Laura J Scott[2,3], He Zhang[16], Anubha Mahajan[4], Jan Veldink[17], Ulrike Peters[18,19], Carlos Pato[20], Cornelia M van Duijn[21], Christopher E Gillies[22], Ilaria Gandin[23], Massimo Mezzavilla[24,25], Arthur Gilly[1], Massimiliano Cocca[23], Michela Traglia[26], Andrea Angius[11], Jeffrey C Barrett[1], Dorrett Boomsma[27], Kari Branham[28], Gerome Breen[29,30], Chad M Brummett[31], Fabio Busonero[11], Harry Campbell[32], Andrew Chan[33,34], Sai Chen[2,3,35,36], Emily Chew[37], Francis S Collins[38], Laura J Corbin[12], George Davey Smith[12], George Dedoussis[39], Marcus Dorr[40,41], Aliki-Eleni Farmaki[39], Luigi Ferrucci[42], Lukas Forer[43], Ross M Fraser[31], Stacey Gabriel[44], Shawn Levy[45], Leif Groop[46–48], Tabitha Harrison[18], Andrew Hattersley[49], Oddgeir L Holmen[50], Kristian Hveem[50], Matthias Kretzler[35,36,51], James C Lee[52,53], Matt McGue[54], Thomas Meitinger[55–57], David Melzer[58], Josine L Min[12], Karen L Mohlke[59], John B Vincent[60–62], Matthias Nauck[8,41], Deborah Nickerson[63], Aarno Palotie[44,64–68], Michele Pato[20], Nicola Pirastu[23], Melvin McInnis[69], J Brent Richards[70–72], Cinzia Sala[26], Veikko Salomaa[13], David Schlessinger[73], Sebastian Schoenherr[43], P Eline Slagboom[74], Kerrin Small[72], Timothy Spector[72], Dwight Stambolian[75], Marcus Tuke[6], Jaakko Tuomilehto[76–79], Leonard H Van den Berg[17], Wouter Van Rheenen[17], Uwe Volker[41,80], Cisca Wijmenga[81], Daniela Toniolo[26], Eleftheria Zeggini[1], Paolo Gasparini[23,25], Matthew G Sampson[22], James F Wilson[32,82], Timothy Frayling[6], Paul I W de Bakker[83,84], Morris A Swertz[81,85], Steven McCarroll[86,87], Charles Kooperberg[18], Annelot Dekker[17], David Altshuler[44,66,88–91], Cristen Willer[16,35,36], William Iacono[54], Samuli Ripatti[92], Nicole Soranzo[1,93,94], Klaudia Walter[1], Anand Swaroop[95], Francesco Cucca[11], Carl A Anderson[1], Richard M Myers[45], Michael Boehnke[2,3], Mark I McCarthy[4,96,97], Richard Durbin[1,99], Gonçalo Abecasis[2,3,99] & Jonathan Marchini[4,10,99] for the Haplotype Reference Consortium

We describe a reference panel of 64,976 human haplotypes at 39,235,157 SNPs constructed using whole-genome sequence data from 20 studies of predominantly European ancestry. Using this resource leads to accurate genotype imputation at minor allele frequencies as low as 0.1% and a large increase in the number of SNPs tested in association studies, and it can help to discover and refine causal loci. We describe remote server resources that allow researchers to carry out imputation and phasing consistently and efficiently.

Over the last decade, large-scale international collaborative efforts have created successively larger and more ancestrally diverse genetic variation resources. For example, in 2007, the International HapMap Project produced a haplotype reference panel of 420 haplotypes at 3.1 million SNPs in three continental populations[1]. More recently, the 1000 Genomes Project has produced a series of data sets built using low-coverage whole-genome sequencing, culminating in 2015 in a reference panel (1000GP3) of 5,008 haplotypes at over 88 million variants from 26 worldwide populations[2]. In addition, several other projects have collected low-coverage whole-genome sequencing data in large numbers of samples that could potentially also be used to build haplotype reference panels[3–5]. A major use of these resources has been to facilitate imputation of unobserved genotypes into genome-wide association study (GWAS) samples that have been assayed using relatively sparse genome-wide microarray chips. As reference panels have increased in number of haplotypes, SNPs and populations, genotype imputation accuracy has increased, allowing researchers to impute and test SNPs for association at ever lower minor allele frequencies (MAFs). A succession of methods developments has provided researchers with the tools to cope with these increasingly larger panels[6–11].

We formed the Haplotype Reference Consortium (HRC; see URLs) to bring together as many whole-genome sequencing data sets as possible to build a much larger combined haplotype reference panel.

A full list of affiliations appears at the end of the paper.

Received 21 December 2015; accepted 18 July 2016; published online 22 August 2016; doi:10.1038/ng.3643

The Haplotype Reference Consortium

OVERVIEW | PARTICIPATING COHORTS | USING THE RESOURCE | CONTACT | SITE LIST

Participating cohorts

A growing list of cohorts/groups that are contributing to the consortium is as follows

HRC COHORTS

	Cohort	# samples in Release 1	Total # samples	Depth	Website	Principal Investigators
1	UK10K	3715	3781	6.5x	http://www.uk10k.org/	Richard Durbin, Nicole Soranzo, George Davey-Smith, Tim Spector, Nick Timpson
2	Sardinia	3445	3514	4x	https://sardinia.irp.nia.nih.gov/	Francesco Cucca, Serena Sanna, Goncalo Abecasis
3	IBD	4478	4478	4x + 2x	http://www.ibdresearch.co.uk/	UK IBD Genetics Consortium
4	GoT2D	2710	2974	4x/Exome	http://www.type2diabetesgenetics.org/infor	Mike Boehnke, David Altshuler, Mark McCarthy
5	BRIDGES	2487	4000	6-8x (12x)		Mike Boehnke, Richard Myers
6	1000 Genomes	2495	2535	4x/Exome	http://www.1000genomes.org/	Richard Durbin, Goncalo Abecasis
7	GoNL	748	748	12x	http://www.nlgenome.nl/	Paul de Bakker
8	AMD	3305	3305	4x		Goncalo Abecasis, Anand Swaroop, Dwight Stambolian
9	HUNT	1023	1254	4x		Cristen Willer, Kristian Hveem
10	SiSu + Kuusamo	1918	1918	4x		Richard Durbin, Aarno Palotie, Samuli Ripatti
11	INGI-FVG	250	250	4-10x	http://www.netgene.it/ita/ingi.asp	Paolo Gasparini, Nicole Soranzo, Nicola Piratsu
12	INGI-Val Borbera	225	225	6x	http://www.netgene.it/ita/ingi.asp	Daniela Toniolo, Nicole Soranzo
13	MCTFR	1325	1339	10x	https://mctfr.psych.umn.edu/	Goncalo Abecasis, Scott Vrieze
14	HELIC	247	2000	4x (1x)	http://www.helic.org/	Eleftheria Zeggini
15	ORCADES	398	399	4x	http://www.orcades.ed.ac.uk/orcades/	Jim Wilson, Richard Durbin
16	inCHIANTI	676	680	7x	http://www.inchiantistudy.net/bindex.html	Tim Frayling, Andrew Wood, Michael Weedon
17	GECCO	1131	3000	4-6x	https://www.fhcrc.org/en/labs/phs/projects/ prevention/projects/gecco.html	Ulrike Peters
18	GPC	697	768	30x		Carlos Pato, Michele Pato, Steven McCarroll
19	Project MinE - NL	935	1250	45x	http://projectmine.com	Jan Veldink, Leonard van den Berg
20	NEPTUNE	403	403	4x	http://www.neptune-study.org/	Matthias Kretzler, Matthew Sampson
	Totals	**32611**	**38821**			

Take home points

- Many **subjects**…
- From many **populations**…
- Assayed for many **variants**

- Quality of reference haplotypes continues to improve

- Data are publicly available

QUALITY CONTROL

Quality Control

- As in ANY analysis, we want quality data
  - Garbage in → garbage out
- So what here is unique?
  - Mendelian inheritance
  - Lab-based protocols
    - Sample duplication for concordance
    - Call rates
    - Chromosomal anomalies
    - …
  - Population genetics, e.g., Hardy-Weinberg Equilibrium testing

- Much research in this area, updated protocols

Hardy-Weinberg Equilibrium

- Allele and genotype frequencies remain constant over time, when…
  - Large population
  - Random mating
  - Sex-independent genotype frequencies
  - No natural selection
  - No migration
  - No mutation
  - No inbreeding

- Implications
  - Can derive **expected** genotype frequencies from allele frequencies
  - If these deviate from realized genotype frequencies, then one of the assumptions may not hold

    OR
  - Genotyping error

    OR
  - Association ?

PROTOCOL

1564 | VOL.5 NO.9 | 2010 | NATURE PROTOCOLS

Data quality control in genetic case-control association studies

Carl A Anderson[1,2], Fredrik H Pettersson[1], Geraldine M Clarke[1], Lon R Cardon[3], Andrew P Morris[1] & Krina T Zondervan[1]

[1]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]GlaxoSmithKline, King of Prussia, Pennsylvania, USA. Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk) or K.T.Z. (krinaz@well.ox.ac.uk).

Published online 26 August 2010; doi:10.1038/nprot.2010.116

This protocol details the steps for data quality assessment and control that are typically carried out during case-control association studies. The steps described involve the identification and removal of DNA samples and markers that introduce bias. These critical steps are paramount to the success of a case-control study and are necessary before statistically testing for association. We describe how to use PLINK, a tool for handling SNP data, to perform assessments of failure rate per individual and per SNP and to assess the degree of relatedness between individuals. We also detail other quality-control procedures, including the use of SMARTPCA software for the identification of ancestral outliers. These platforms were selected because they are user-friendly, widely used and computationally efficient. Steps needed to detect and establish a disease association using case-control data are not discussed here. Issues concerning study design and marker selection in case-control studies have been discussed in our earlier protocols. This protocol, which is routinely used in our labs, should take approximately 8 h to complete.

Curr Protoc Hum Genet. 2011 Jan;Chapter 1:Unit1.19. doi: 10.1002/0471142905.hg0119s68.

Quality control procedures for genome-wide association studies.

Turner S[1], Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD.

Author information

1 Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, Tennessee, USA.

http://bioconductor.org/packages/release/bioc/html/GWASTools.html

Home » Bioconductor 3.5 » Software Packages » GWASTools

GWASTools

platforms all | downloads top 20% | posts 2 / 2 / 2 / 0 | in Bioc 6 years
build ok | commits 0.83 | test coverage 71%

Tools for Genome Wide Association Studies

Bioconductor version: Release (3.5)

Classes for storing very large GWAS data sets and annotation, and functions for GWAS data cleaning and analysis.

Author: Stephanie M. Gogarten, Cathy Laurie, Tushar Bhangale, Matthew P. Conomos, Cecelia Laurie, Caitlin McHugh, Ian Painter, Xiuwen Zheng, Jess Shen, Rohit Swarnkar, Adrienne Stilp, Sarah Nelson

Maintainer: Stephanie M. Gogarten <sdmorris at u.washington.edu>, Adrienne Stilp <amstilp at u.washington.edu>

Citation (from within R, enter `citation("GWASTools")`):

Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, Nelson SC, Rice K, Shen J, Swarnkar R, Weir BS and Laurie CC (2012). "GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies." *Bioinformatics*, **28**(24), pp. 3329-3331. doi: 10.1093/bioinformatics/bts610.

GWAS Data Cleaning

GENEVA Coordinating Center
Department of Biostatistics
University of Washington

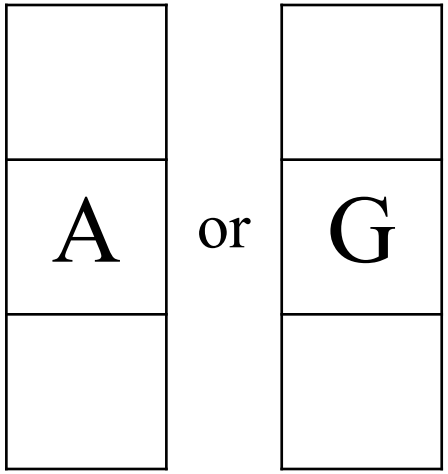
April 24, 2017

Contents

Modes of Inheritance

Coding Genotypes

- Assume a **biallelic** marker (SNP)

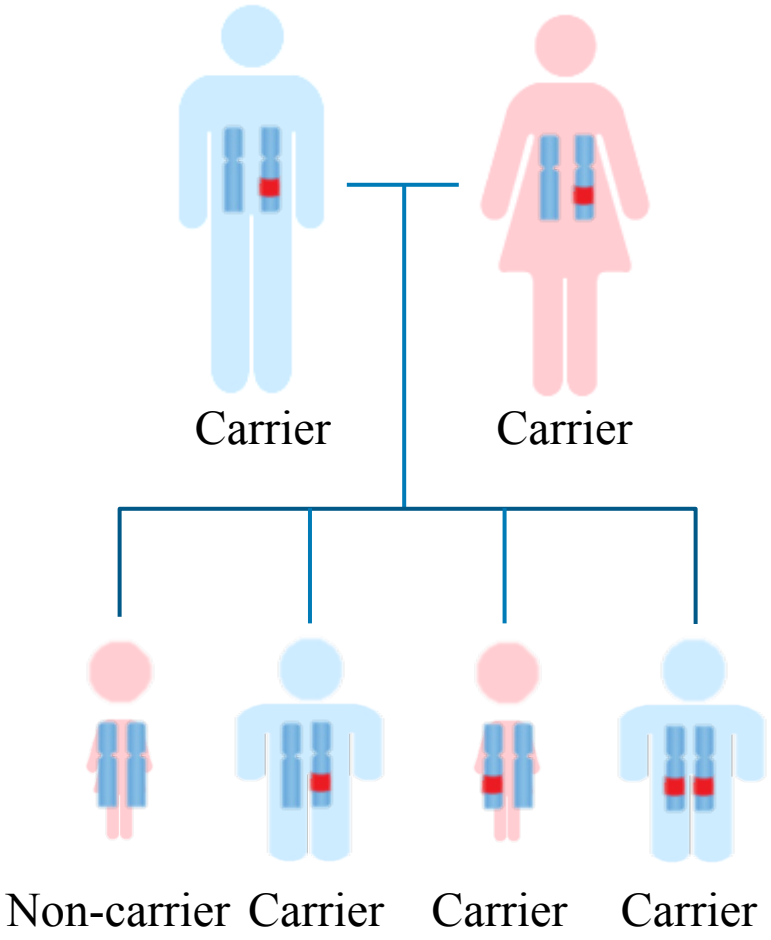


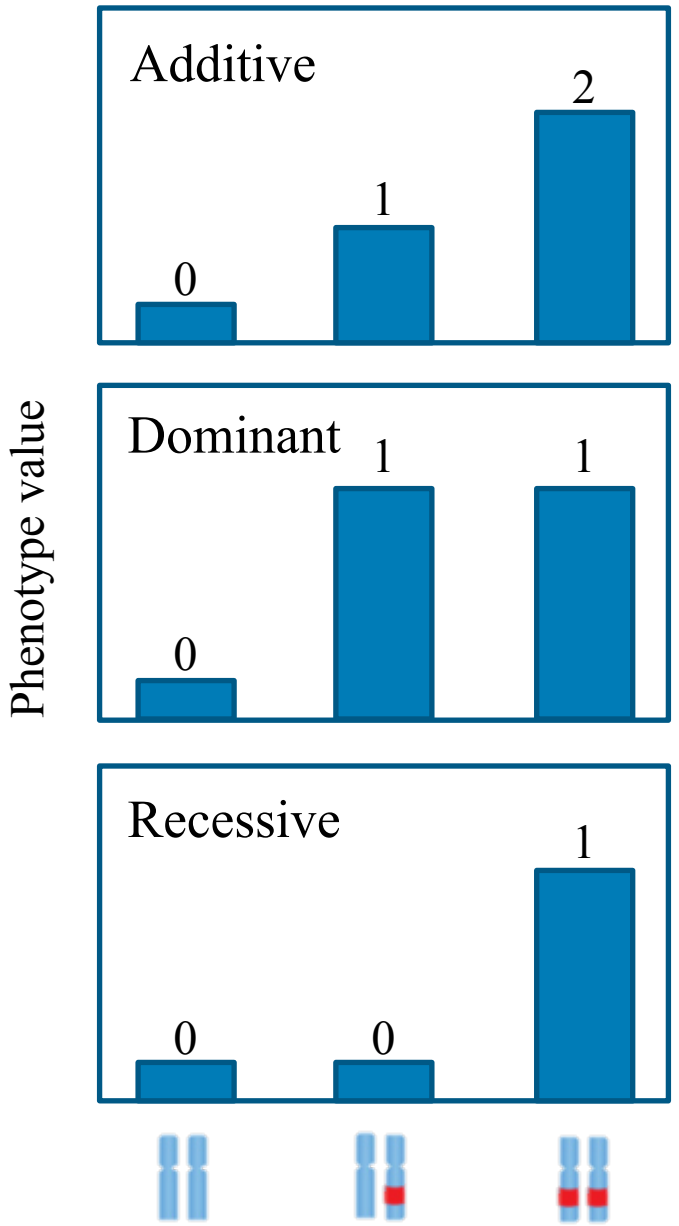
Each chromosome will have one of the two possible alleles

FID	IID	A1	A2
0	0001	A	A
0	0002	A	G
0	0003	G	G
0	0004	A	A
⋮	⋮	⋮	⋮

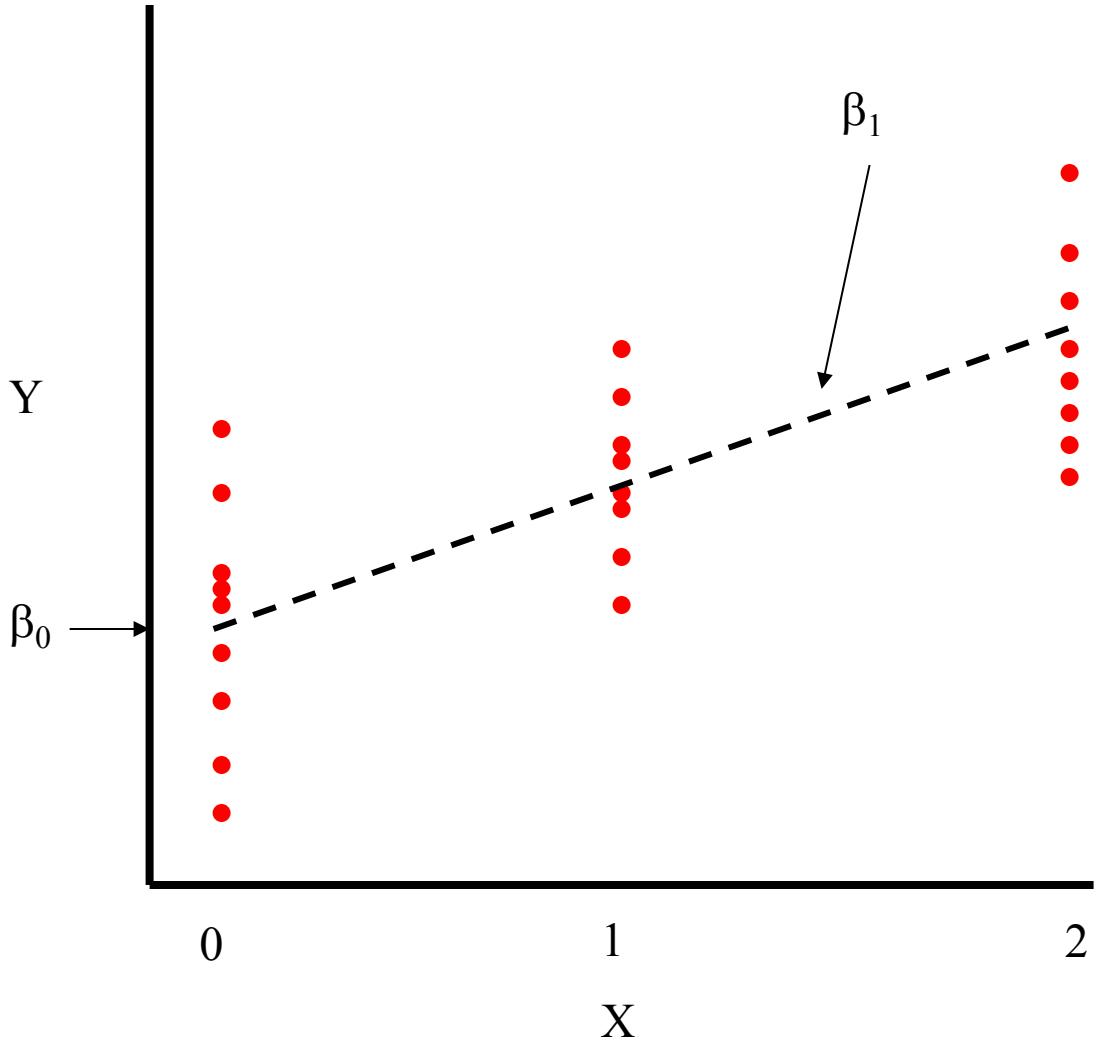
Mode of inheritance (MOI)

A pattern of how a disease is transmitted in families

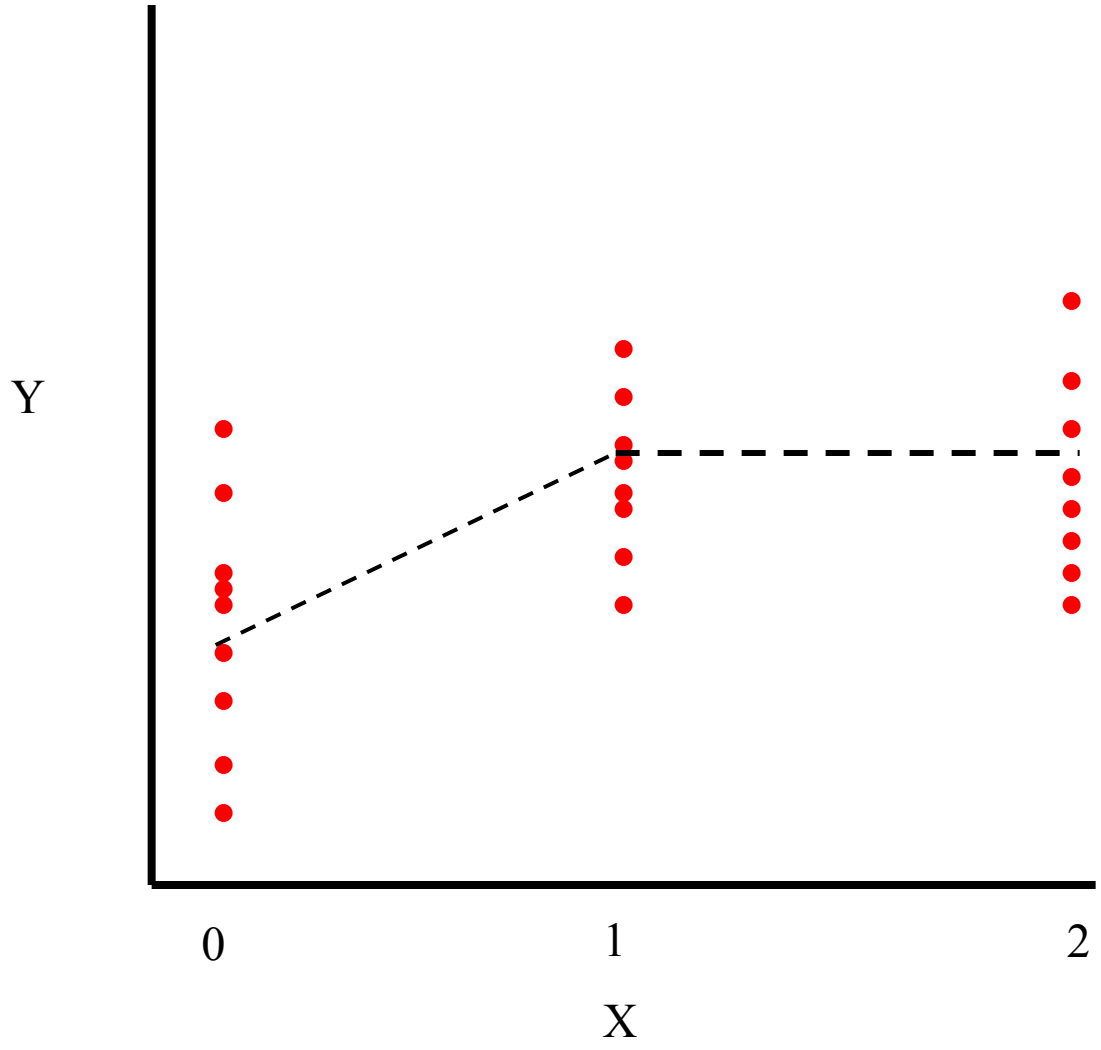




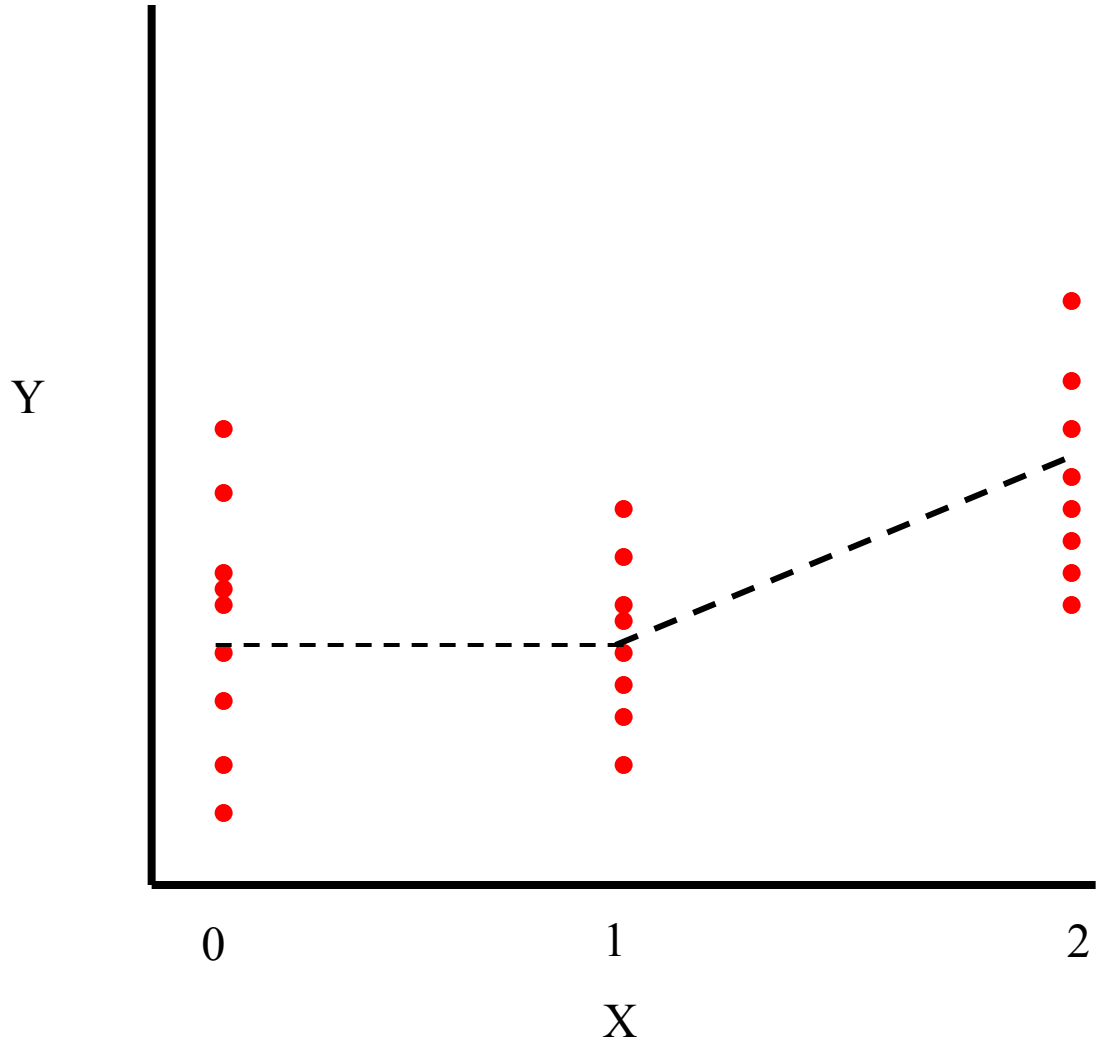
Additive Mode



Dominant Mode



Recessive Mode



FID	IID	A1	A2
0	0001	A	A
0	0002	A	G
0	0003	G	G
0	0004	A	A
⋮	⋮	⋮	⋮

A: risk allele

Additive

FID	IID	G1
0	0001	2
0	0002	1
0	0003	0
0	0004	2
⋮	⋮	⋮

Dominant

FID	IID	G1
0	0001	1
0	0002	1
0	0003	0
0	0004	1
⋮	⋮	⋮

Recessive

FID	IID	G1
0	0001	1
0	0002	0
0	0003	0
0	0004	1
⋮	⋮	⋮

Tests for Association

University of Kentucky

Tests for Association

- Discrete Traits
  - Cochrane-Armitage Trend Test
  - Alleles Test
  - General RxC Contingency Table (Chi-square)
- Other Types
  - Continuous
  - Time-to-event
  - Multivariate

Cochrane-Armitage

	Copies of Allele			
	0	**1**	**2**	
Case	A_0	A_1	A_2	M_1
Control	U_0	U_1	U_2	M_0
	N_0	N_1	N_2	N

$$\chi_1^2 = \frac{N[N(A_1 + 2A_2) - M_1(N_1 + 2N_2)]^2}{M_1(N - M_1)[N(N_1 + 4N_2) - (N_1 + 2N_2)^2]}$$

Alleles Test

	+	-	
Case	A_1+2A_2	A_1+2A_0	$2M_1$
Control	U_1+2U_2	U_1+2U_0	$2M_0$
	N_1+2N_2	N_1+2N_0	$2N$

$$\chi_1^2 = \frac{2N[2N(A_1+2A_2)-2M_1(N_1+2N_2)]^2}{2M_1 2(N-M_1)[2N(N_1+2N_2)-(N_1+2N_2)^2]}$$

Note: Variance (denominator) assumes HWE!!!

General Chi-Square

Copies of Allele				
	0	**1**	**2**	
Case	A_0	A_1	A_2	$\mathbf{M_1}$
Control	U_0	U_1	U_2	$\mathbf{M_0}$
	$\mathbf{N_0}$	$\mathbf{N_1}$	$\mathbf{N_2}$	**N**

$$\chi_2^2 = \frac{[A_0 - E(A_0)]^2}{E(A_0)} + \frac{[A_1 - E(A_1)]^2}{E(A_1)} + \frac{[A_2 - E(A_2)]^2}{E(A_2)} + \frac{[U_0 - E(U_0)]^2}{E(U_0)} + \frac{[U_1 - E(U_1)]^2}{E(U_1)} + \frac{[U_2 - E(U_2)]^2}{E(U_2)}$$

Logistic Regression

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 X$$

$$\log it(p_x) = \beta_0 + \beta_1 X$$

Model Interpretation

Additive model (*X=0, 1 or 2*)

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 X \qquad \ln(\beta_1) = \text{one - unit increase}$$

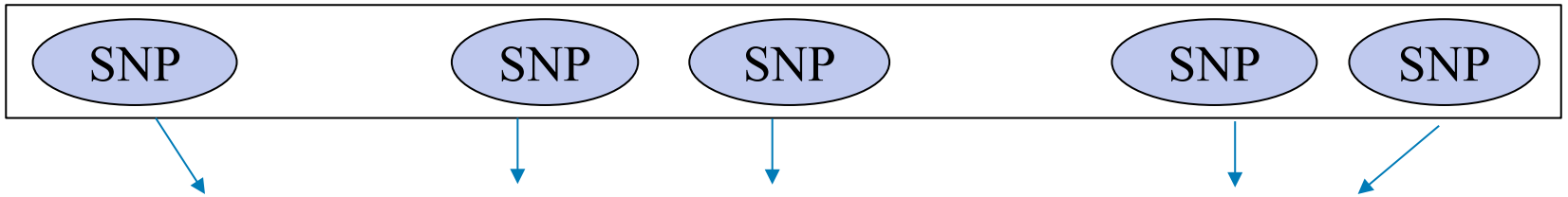
Note: This is analogous to an odds ratio (OR) from a 2x3 table

Genotype Model (indicator variables $G_i = 0$ or 1)

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 G_1 + \beta_2 G_2$$

Subjects with $G_0 = 1$ are the reference group

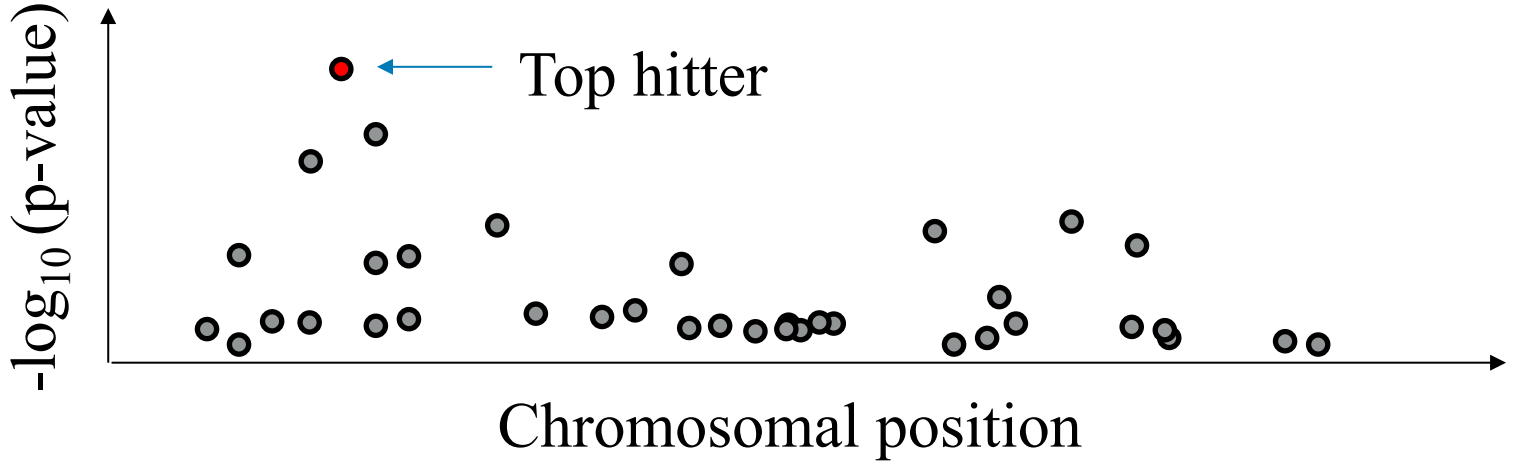
OR for subjects with G_2 compared to $G_0 = e^{\beta_2}$



$$E(y_i) = \beta_0 + \underline{\beta_1 X_{ij}} \quad \text{or} \quad \log\frac{p_i}{1-p_i} = \beta_0 + \underline{\beta_1 X_{ij}}$$

P-value

P-value



Association take home points

- Many ways to seek out and test for a genetic association
- Mode of inheritance, while somewhat a misnomer in complex disease genetics, reflects our assumptions on how genotype influences phenotype
- We will focus largely on the flexible frameworks of linear and logistic regression

Population Stratification

University of Kentucky

Genetic Associations

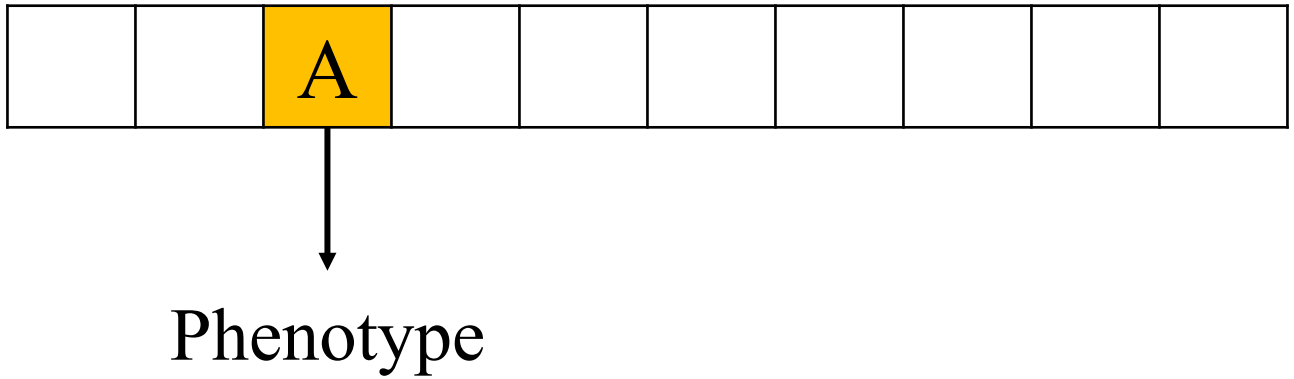
- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times, you'll see ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Yellow fingers associated with lung cancer…
  - e.g. Population stratification

Genetic Associations

- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times, you'll see ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Yellow fingers associated with lung cancer…
  - e.g. Population stratification

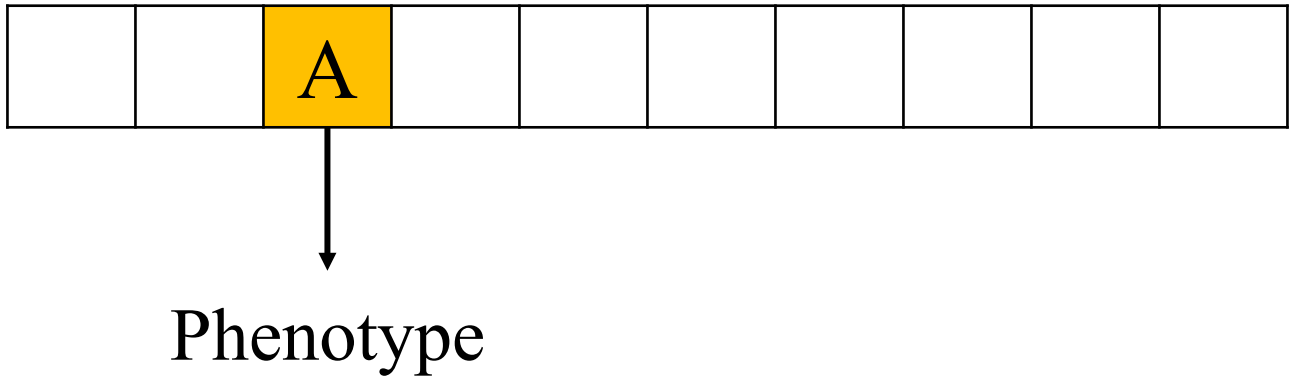
Truth

A genetic association test finds

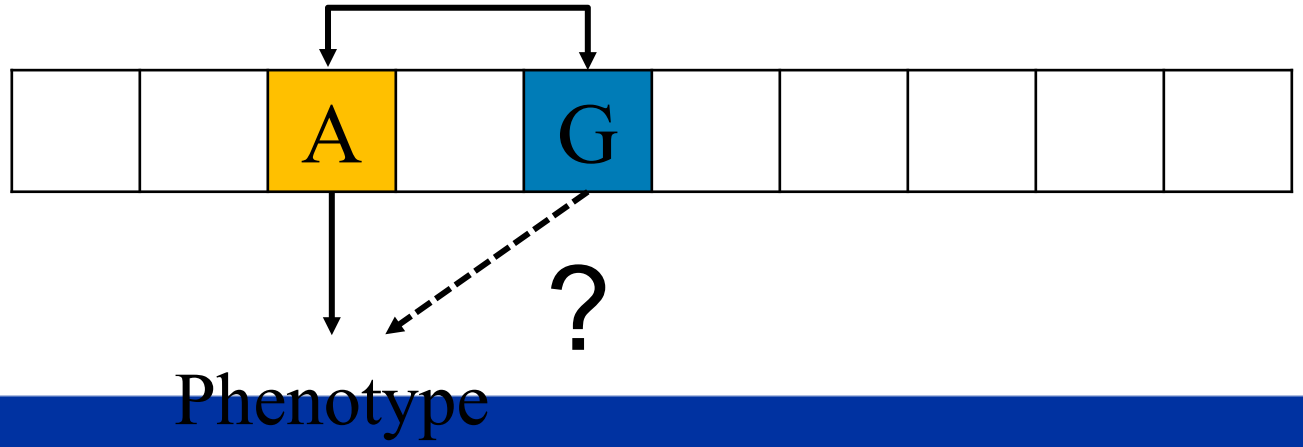


Truth

A genetic association test finds



LD: highly correlated



Genetic Associations

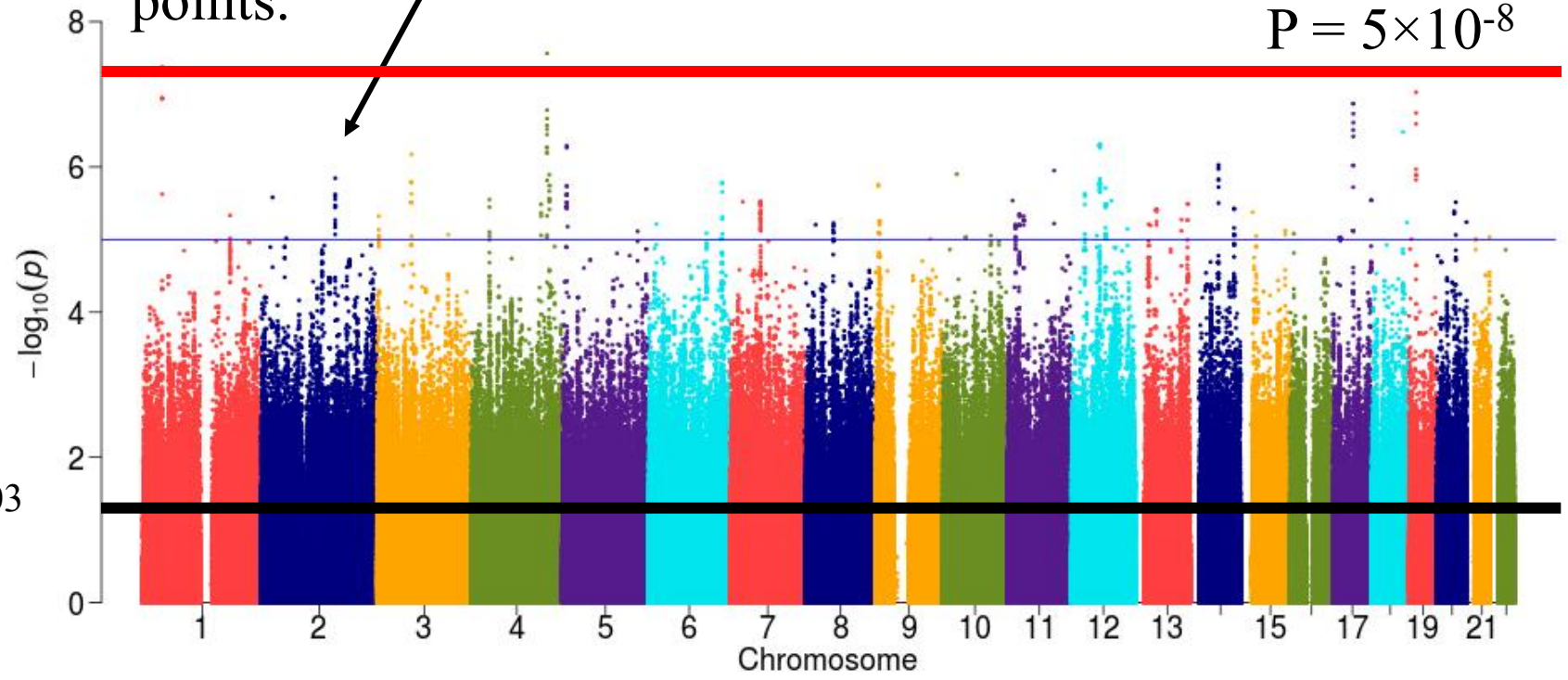
- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times, you'll see ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Yellow fingers associated with lung cancer…
  - e.g. Population stratification

Chance

- Each point represents the association for each locus.
- There are more than 6 million points.

Genome wide significance level

$P = 5 \times 10^{-8}$



$p = 0.05 = 10^{-1.30103}$

Manhattan plot: It is a scatter plot used to display the p-values in genome-wide association studies (GWAS)

https://www.ebi.ac.uk/gwas/diagram

Genetic Associations

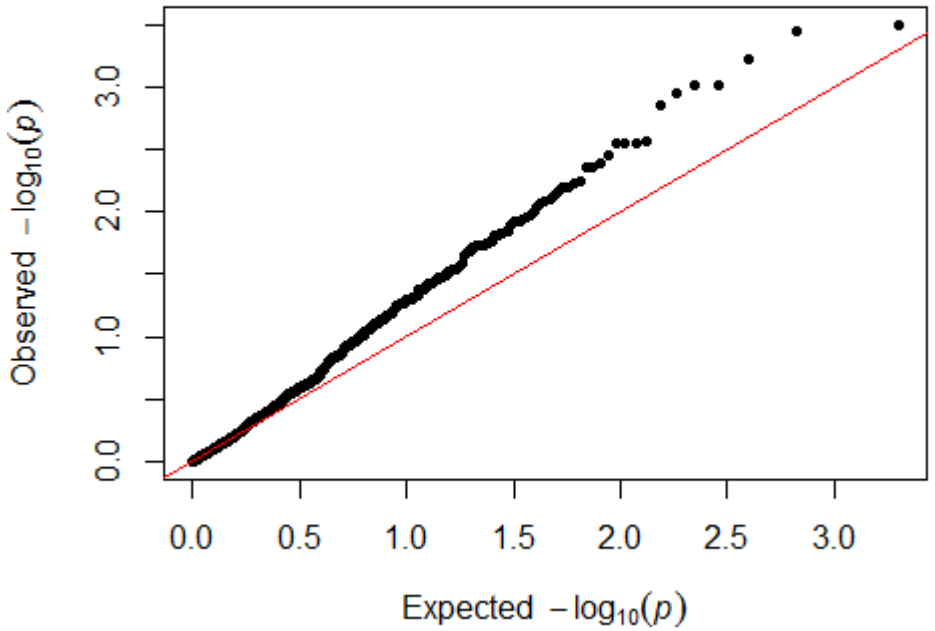
- Truth
  - Causal locus (direct)
  - In LD with causal locus (indirect)
- Chance
  - If you test 100 times, you'll see ~ 5 tests < 0.05
  - The association is due to chance - no causal underpinning
- Bias
  - Association is not causal
  - Yellow fingers associated with lung cancer…
  - e.g. Population stratification

Quantile-quantile (QQ) Plots

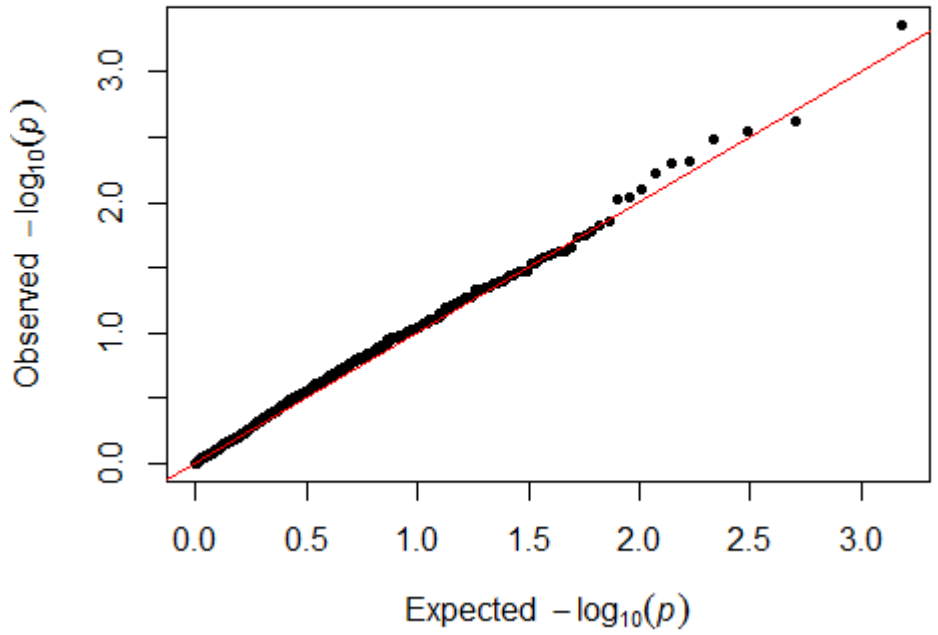
- Good way of seeing what's going on overall
  - Any "real" hits?
  - Any systematic problems?

- In GWAS, MOST SNPs will **not** be associated with whatever phenotype is examined, i.e., they are from the null distribution

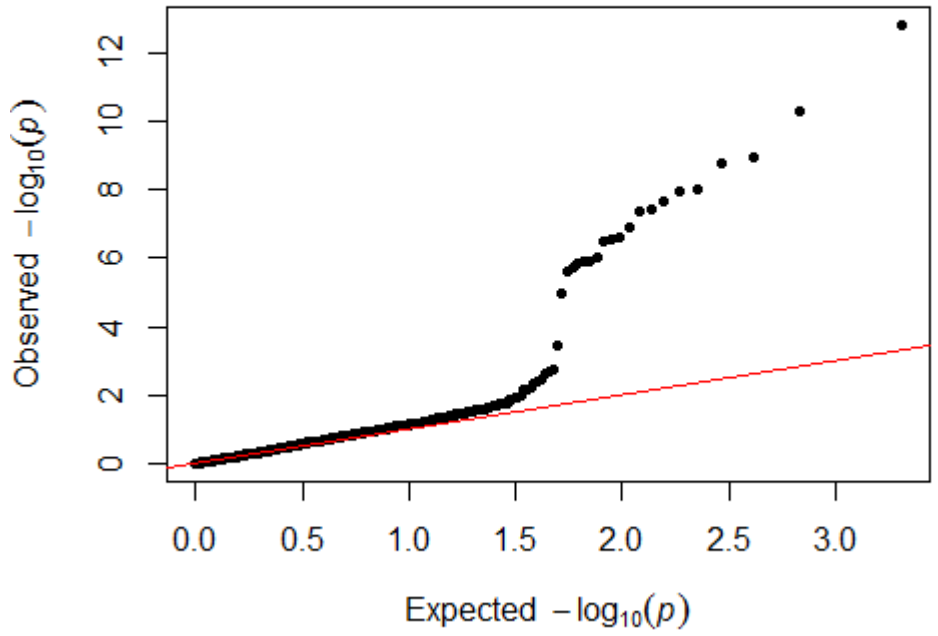
Quantile-quantile (QQ) Plots



Quantile-quantile (QQ) Plots

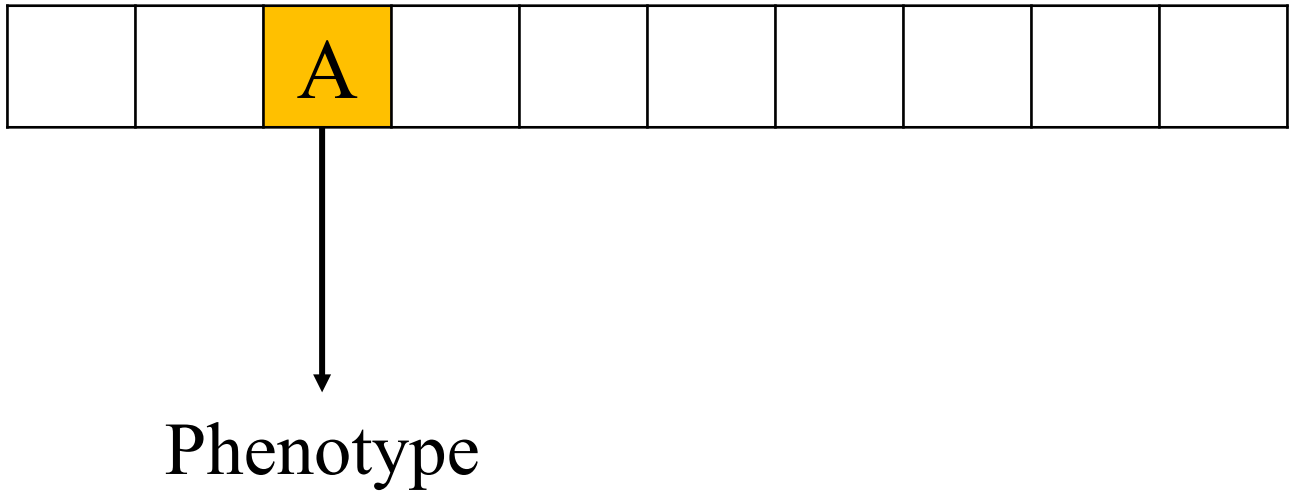


Quantile-quantile (QQ) Plots



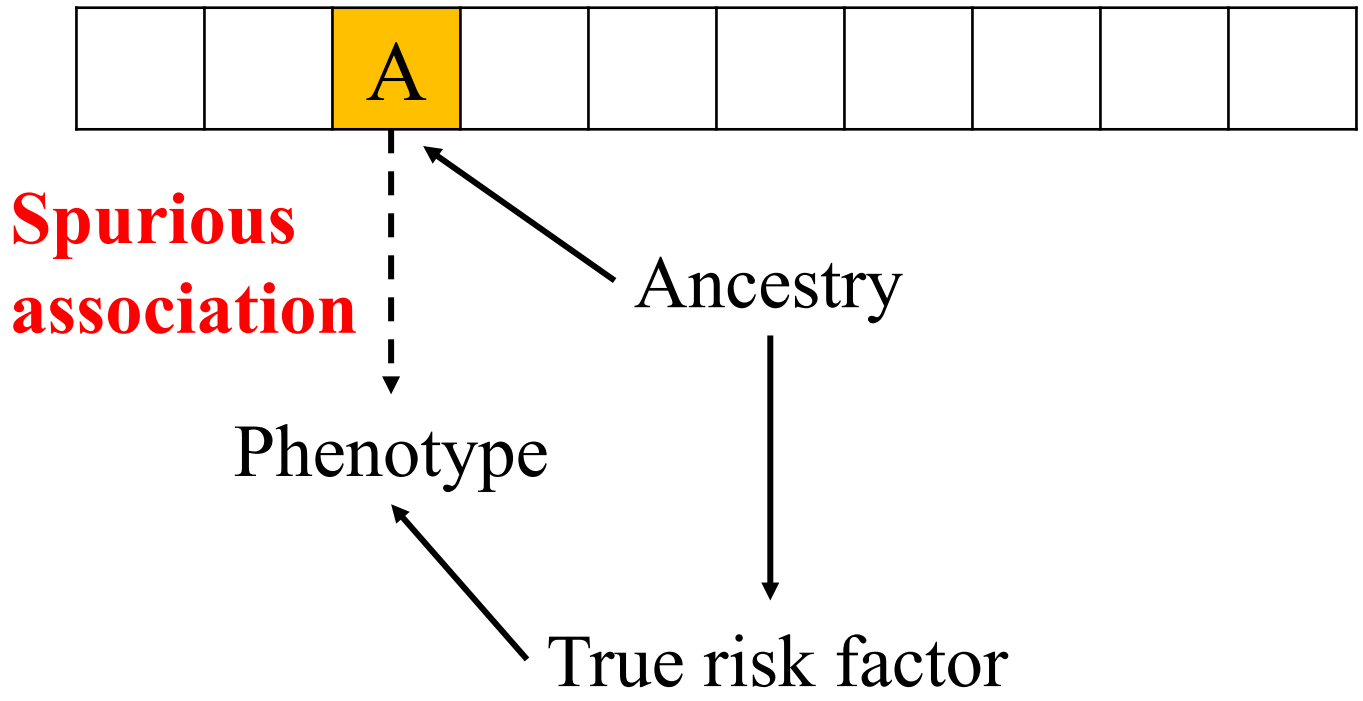
Bias

A genetic association test finds



Bias

A genetic association test finds



Stratification

- Essentially a confounder!
- Yellow fingers associated with lung cancer…
- How does it happen?

Famous Example
Knowler et al (1988)

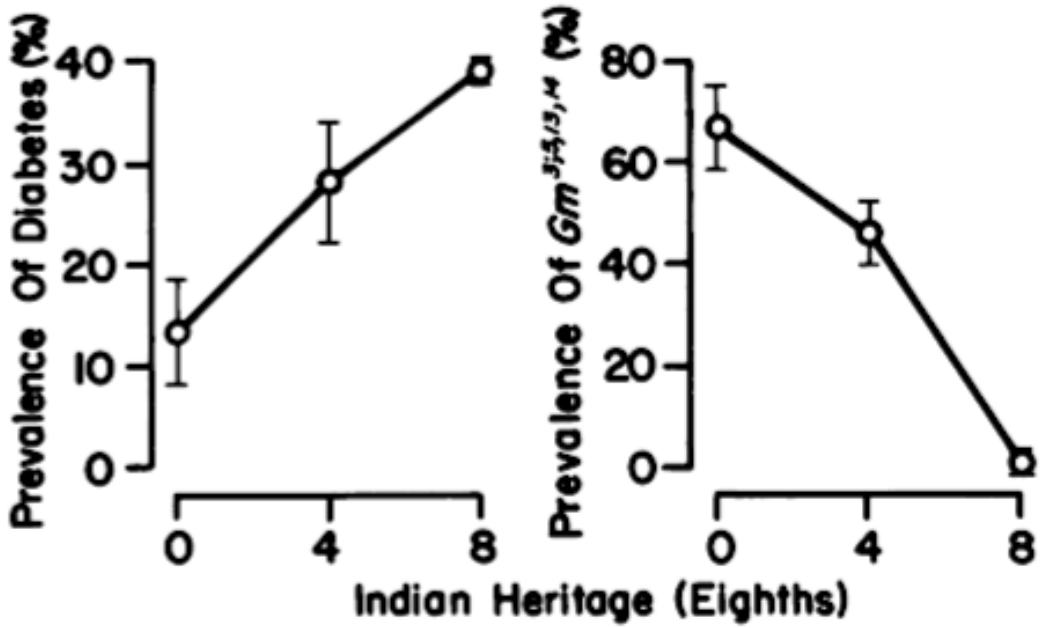
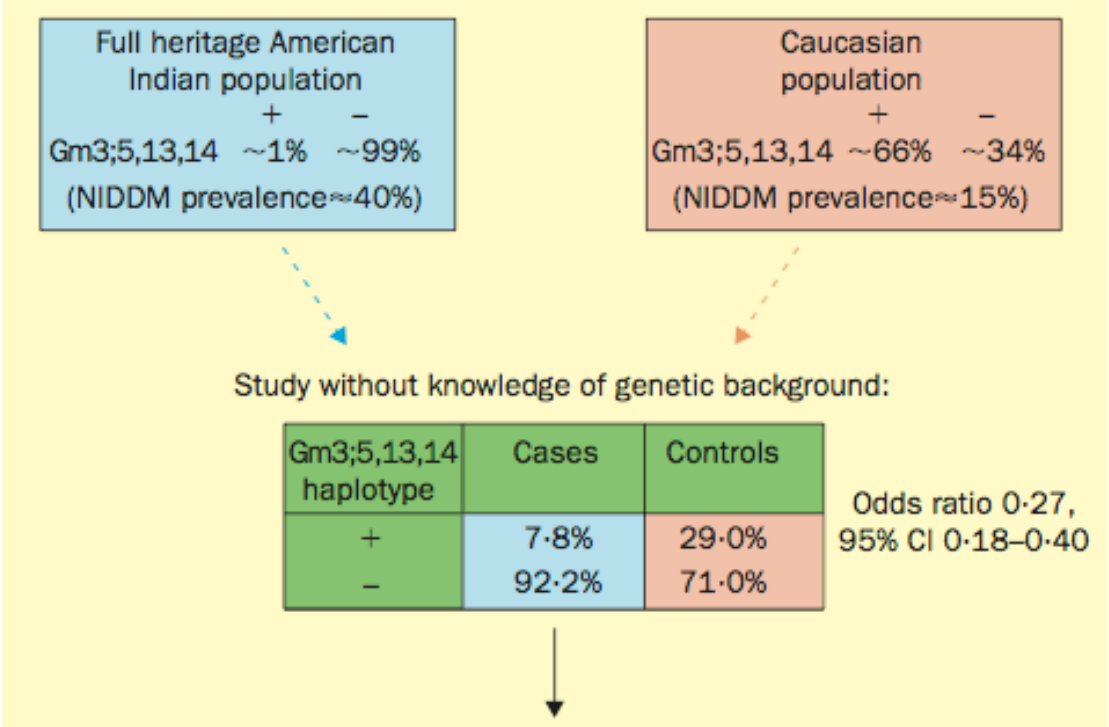


Figure 3 Age-adjusted prevalence (±1 standard error) of diabetes (left) and of $Gm^{3;5,13,14}$ (right), according to Indian heritage, among residents of the Gila River Indian Community.

Cardon et al (2003)



Stratification Happens

- Historical strategies to deal with it
  - Self-Reported Ancestry
    - Match (design) or Adjust (analysis)
  - Use other genetic markers (ancestry informative)
    - Genomic Control
    - STRUCTURE
    - PCA/Eigenstrat
    - Use a family-based design

    - More later

Thank you !

Questions ?

Acknowledgements

Harvard T. F. Chan School of Public Health
- Christoph Lange
- Pete Kraft

University of Colorado
- Matt McQueen

University of Kentucky
- Yuri Katsumata
- Jennifer Daddysman

University of Florida
- Tyler Smith

University of Washington
- Paul Crane
- Joey Mukherjee

Vanderbilt University
- Tim Hohman

Brigham Young University
- Keoni Kauwe

National Institute of Aging

R01 AG042437
K25 AG043546
R01 AG057187
P30 AG028383
R01 AG054060