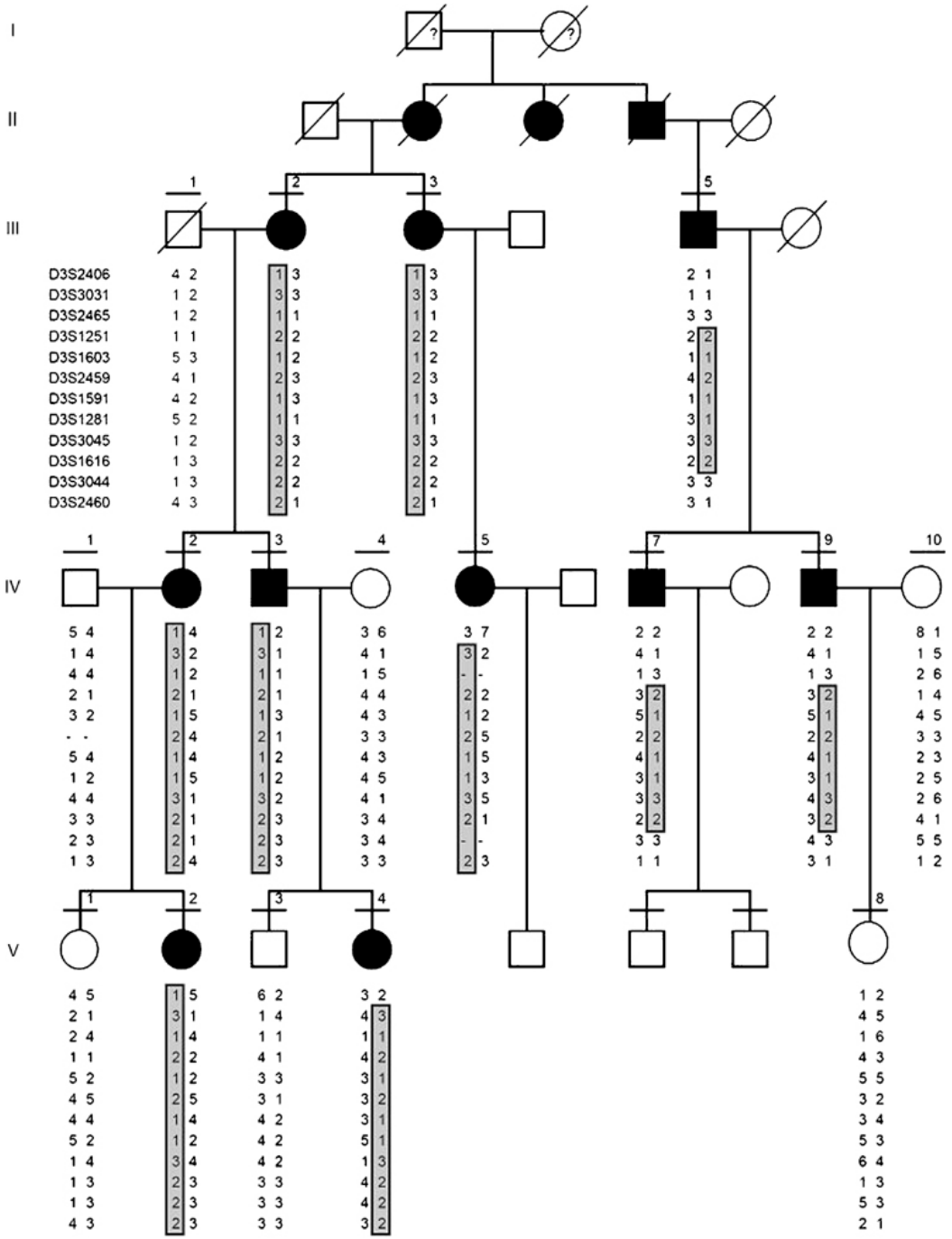
Introduction to Genome-Wide Association Studies (GWAS)

JSM 2018

Eric Sobel, Janet Sinsheimer, Hua Zhou, and others



How to Map Genes

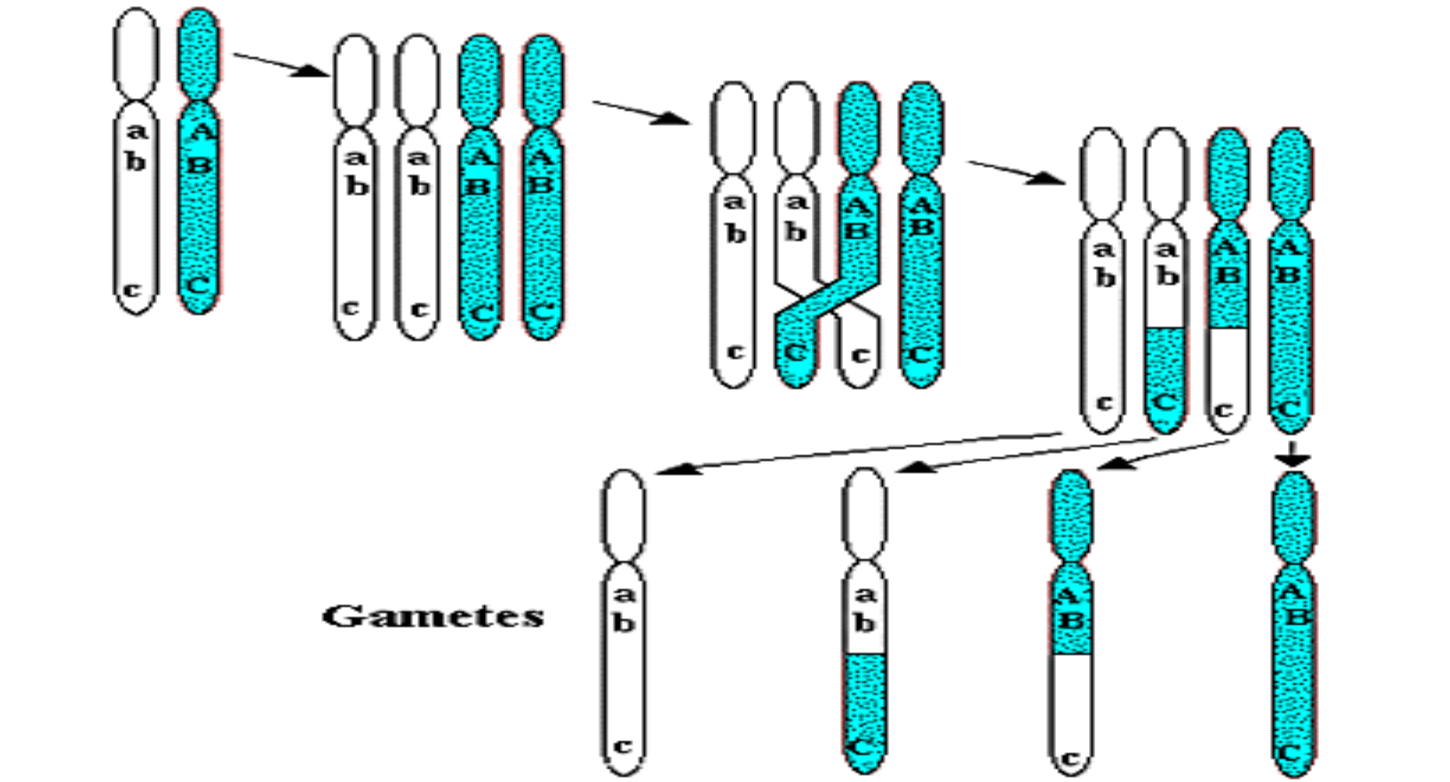
If a genetic variant is segregating, then so are markers near-by, which we can genotype.

European Journal of Human Genetics (2008) 16:265–269

Fifth finger camptodactyly maps to chromosome 3q11.2–q13.12 in a large German kindred.

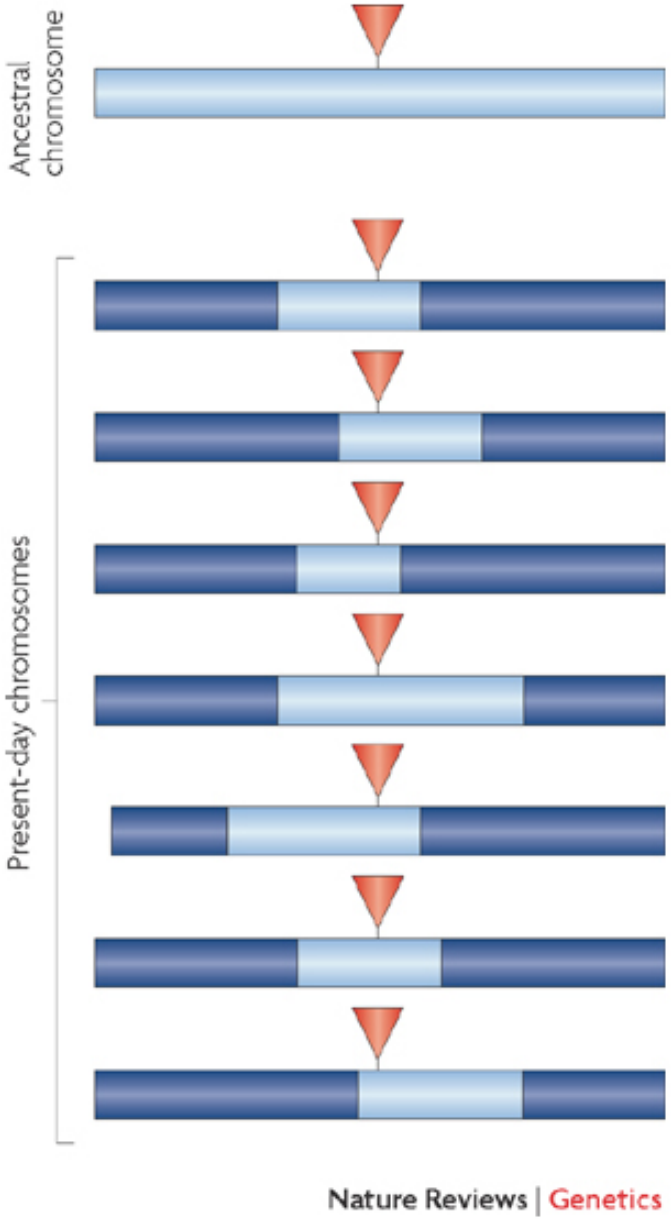
Sajid Malik, Jörg Schott, Julia Schiller, Anna Junge, Erika Baum, Manuela Koch

Pairs of Homologous Chromosomes



Crossing-over and recombination during meiosis

Another view of LD



A new mutation is in complete LD with its background chromosome

Many generations later the region of complete LD is much smaller, but partial LD can still be observed around the mutation site.

Linkage Disequilibrium

Linkage disequilibrium is a measure that expresses the extent to which alleles at two loci are nonrandomly associated within a population.

Linkage disequilibrium implies a shared ancestry between alleles at two loci that has not been eroded by recombination.

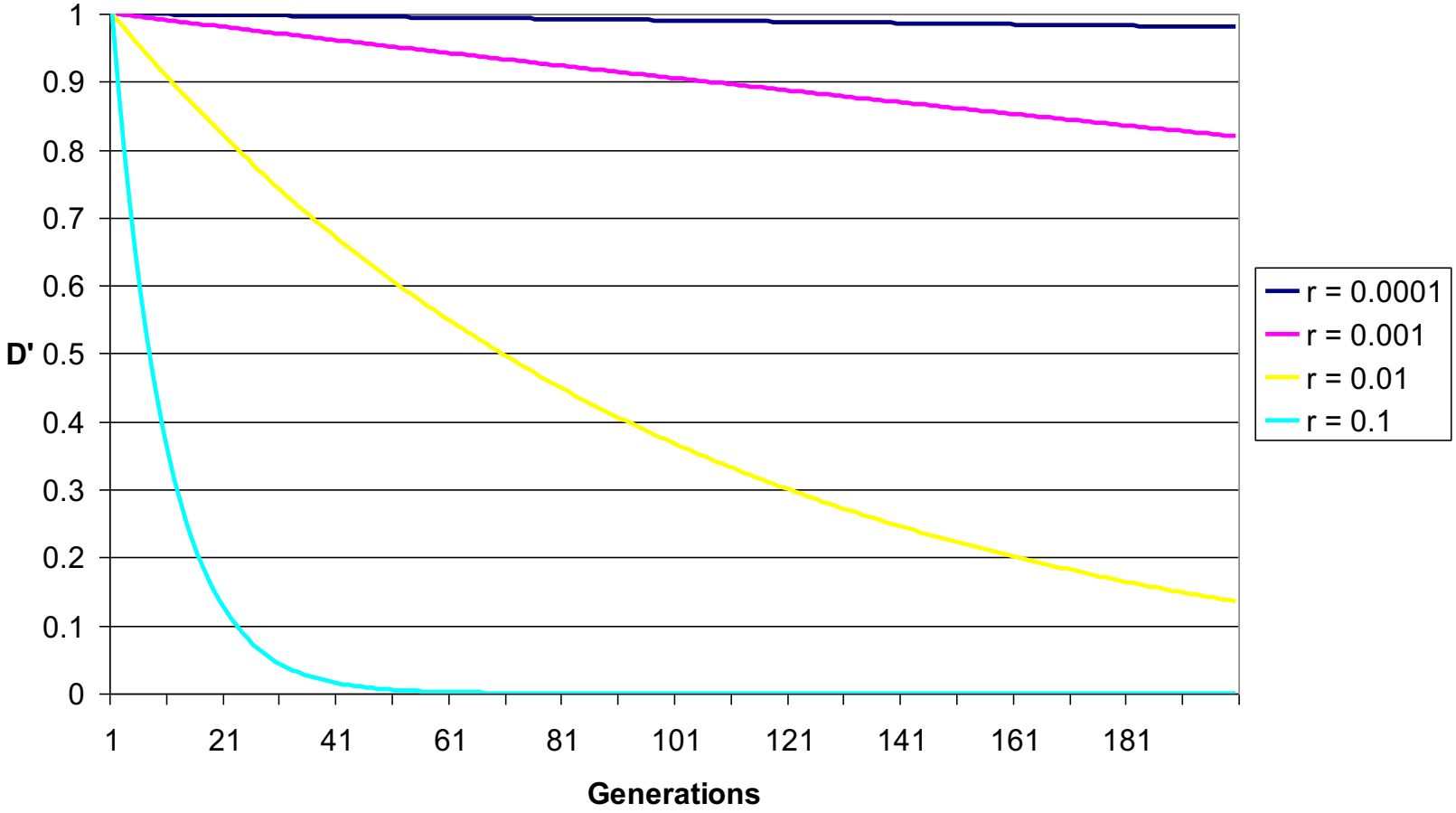
Factors Influencing LD:

LD throughout the genome reflects the population history, breeding patterns and geographic structure.

LD in specific genomic regions reflects the history of natural selection, gene conversion and mutation.

How these local and global factors affect LD between a specific pair of loci will depend on the recombination rate.

Linkage Disequilibrium and Recombination Rate: Simulation Results



Recombination Hotspots and Haplotype Blocks

The figure shows results from one of the first studies to report on haplotype block patterns in the human genome.

Pattern of pairwise LD across a 217 kb region of the MHC class II region.

L is the likelihood ratio of the test of linkage disequilibrium.

Pattern shows stretches on strong LD bounded by regions of high recombination (i.e., recombination hotspots).

Haplotype blocks typically extend from a few kb to hundreds of kb.

Some portions of the genome show no haplotype block pattern.

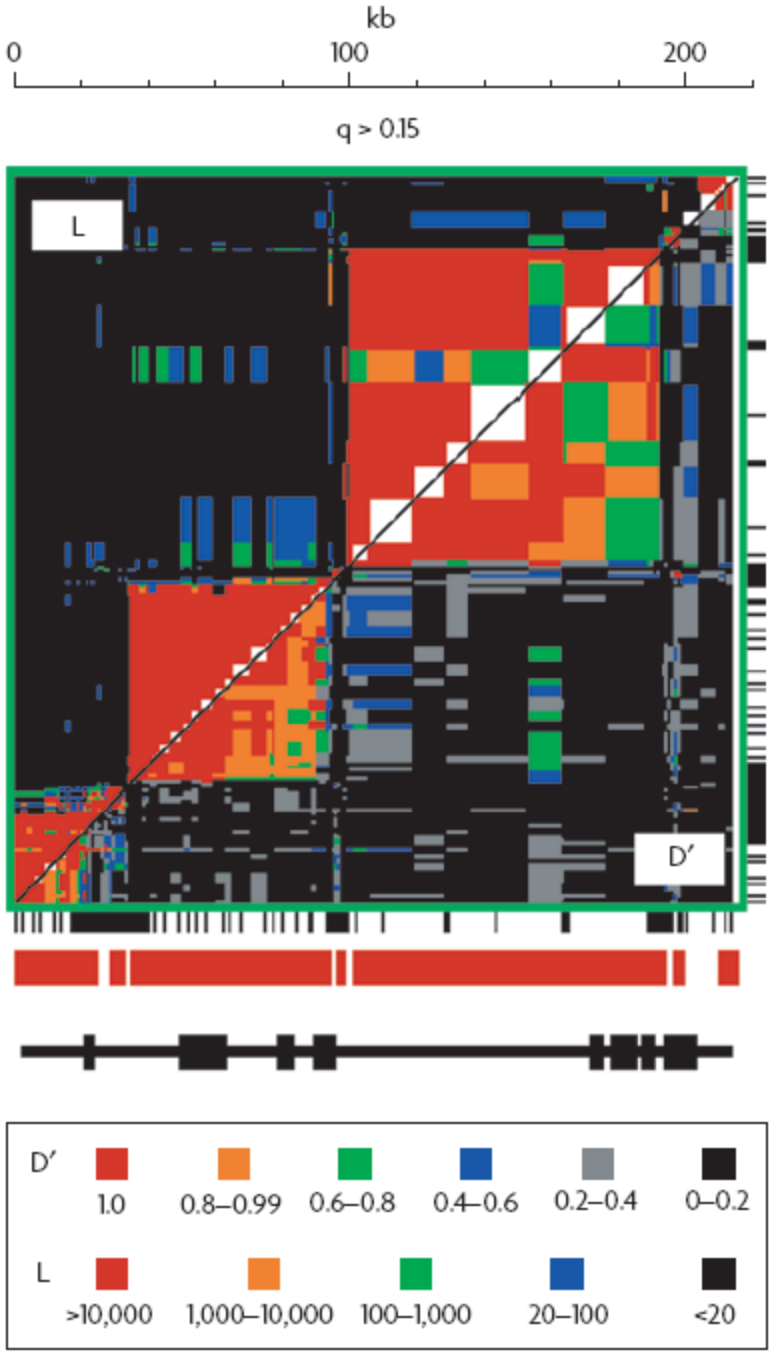
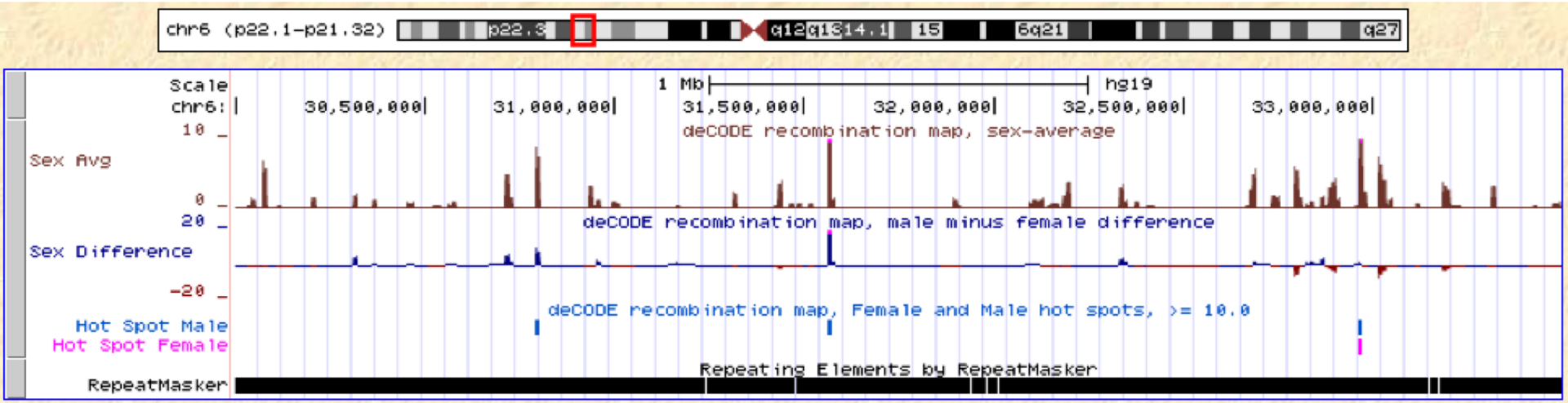


Figure from: Jeffreys AJ et al. (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genet. 29: 217-222

The Recombination Rate is not Uniform

The HLA region shown by the UCSC Genome Browser



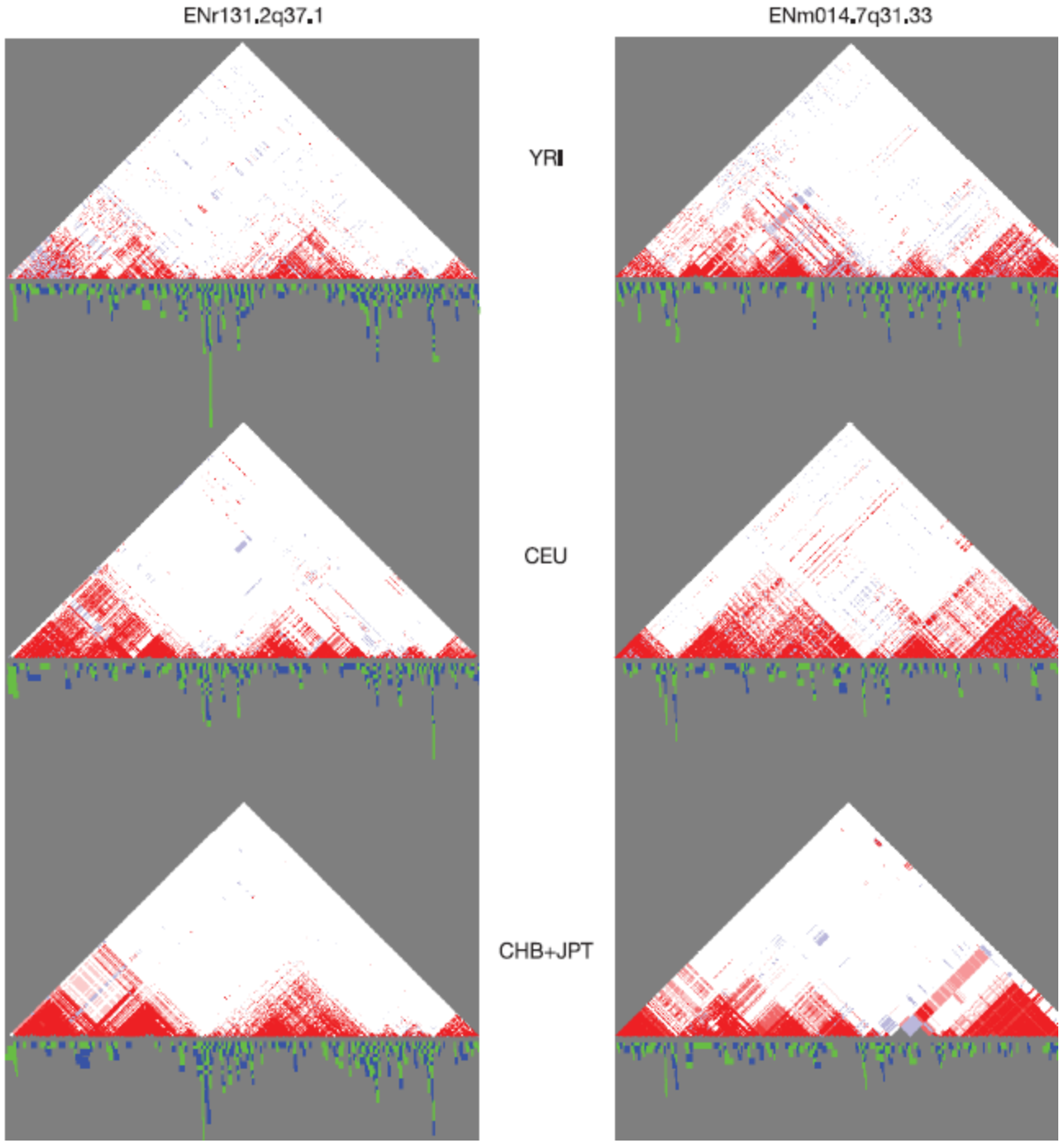
Linkage Disequilibrium: Effective Population Size and Time

Assuming an allele has a history of neutrality (not affected by selection) and the other alleles in the immediate vicinity are also neutral, the frequency of the allele is correlated with its age.

The demographic history of a population, i.e., how the effective population size has changed over time, affects the rate of LD decay.

Small, slow-growing populations will show higher genome-wide levels of LD than large, rapidly-growing populations.

African populations tend to show lower levels of LD than other populations because of the longer period of significant effective population size.



AFRICAN

EUROPEAN

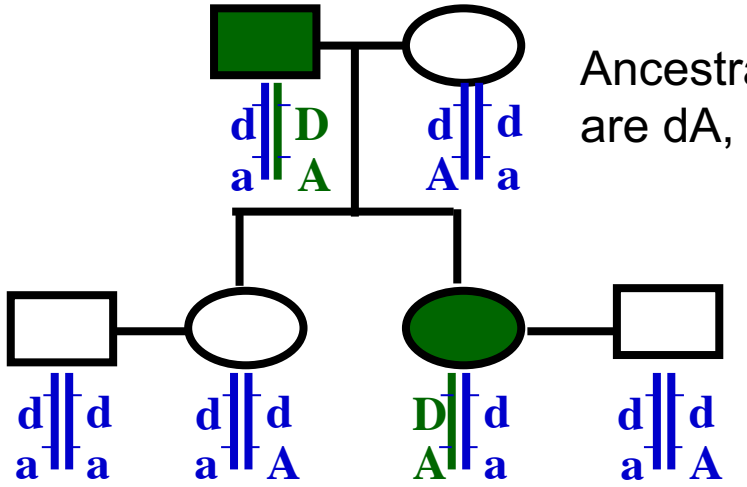
EAST ASIAN

International HapMap Consortium

NATURE|Vol 437|27 October 2005

Linkage Disequilibrium (LD)

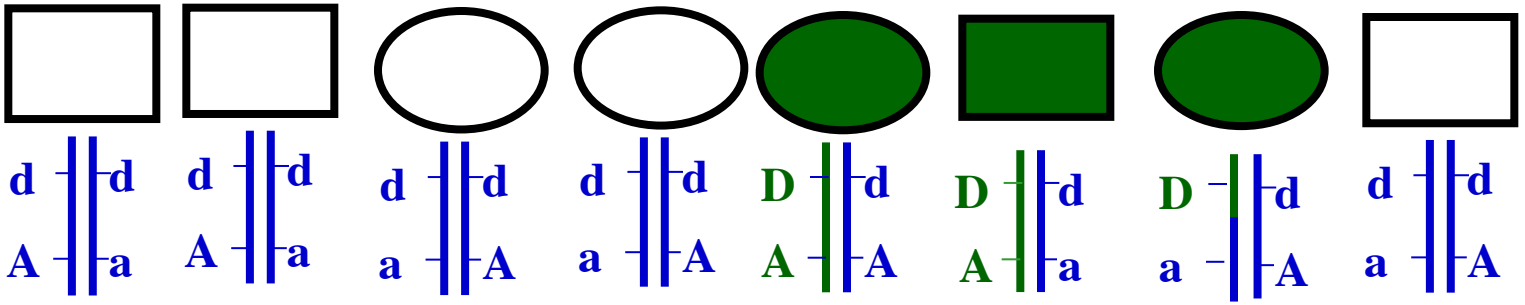
One of the population founders carries an allelic variant that increases risk of a disease. The disease gene is very close to a marker.



Ancestral haplotypes are dA, da, and DA **but not Da**

Note that **D** is associated with **A.** **P(A|D) is close to one.**

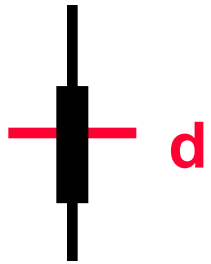
For many generations there is an occasional recombination event between the two loci.



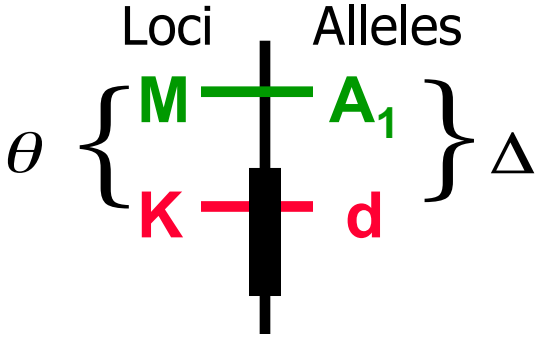
The P(A|D) has decreased, as has the degree of association between **D** and **A**, but still P(A|D) > P(A) and P(haplotype: A D) > P(A) x P(D)

Three causes for allelic association

- best: allele increases disease susceptibility
  - **candidate gene studies**



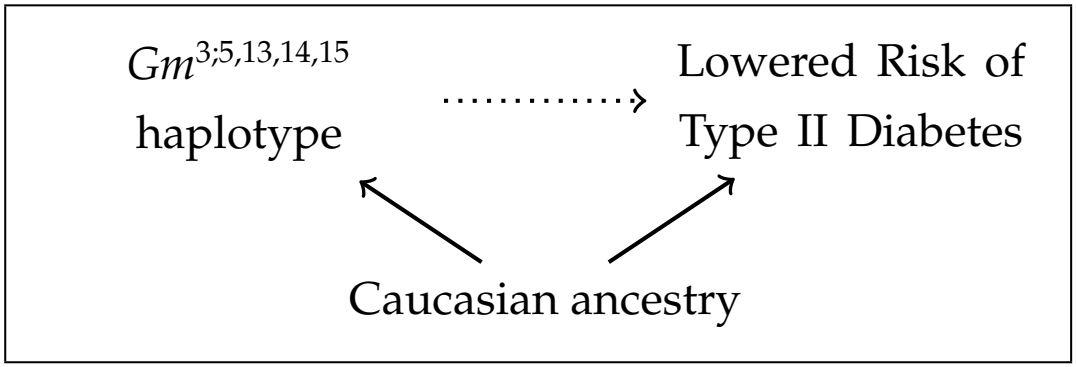
- good: some subjects share common ancestor
  - **linkage disequilibrium studies**



- bad: association due to population stratification
  - **family-based studies offer protection**

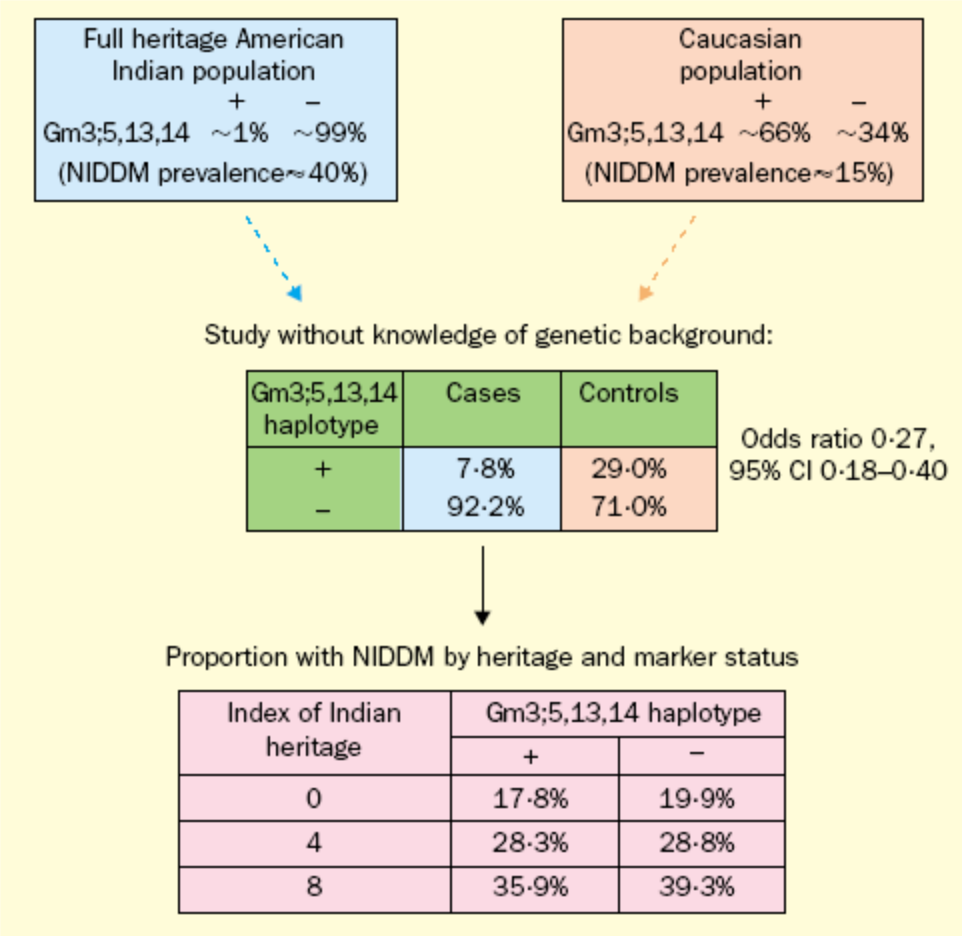
An example of population stratification

- The classic Pima Indian Diabetes study [Knowler, 1988]



- Lessons from this study:
  - Ancestry needs to be accounted for or else we risk false positive associations.
  - Ancestry is fractional, not discrete. Classical stratified designs inadequate.
- Another lesson: Self-reported ancestry often inaccurate. "Cryptic ancestry."

Classic Example of Spurious Association Due to Population Admixture

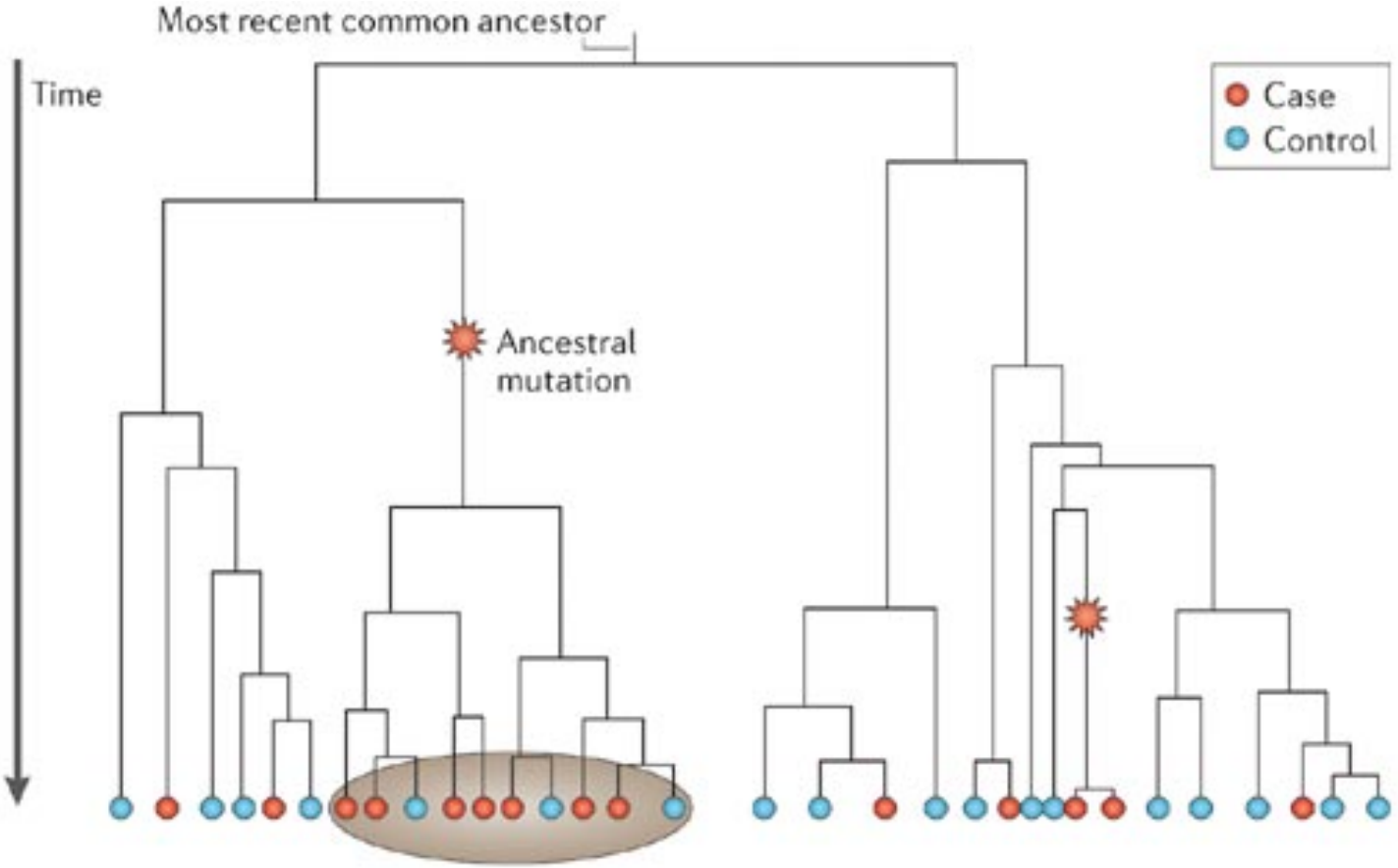


Population stratification in a study of Gm3;5,13,14 haplotype in a genetically admixed sample of native Americans of the Pima and Papago tribes

Disease prevalence and allele frequencies both differ between European and native American populations; the haplotype is not found in full-blood native Americans. The noted association of the Gm3;5,13,14 haplotype with reduced risk of non-insulin-dependent diabetes mellitus (NIDDM) is attributable to ancestral population of origin rather than to linkage disequilibrium between the disease and marker loci.

From LR Cardon and LJ Palmer 2003 Lancet 361:589-604

Population-based Association Studies

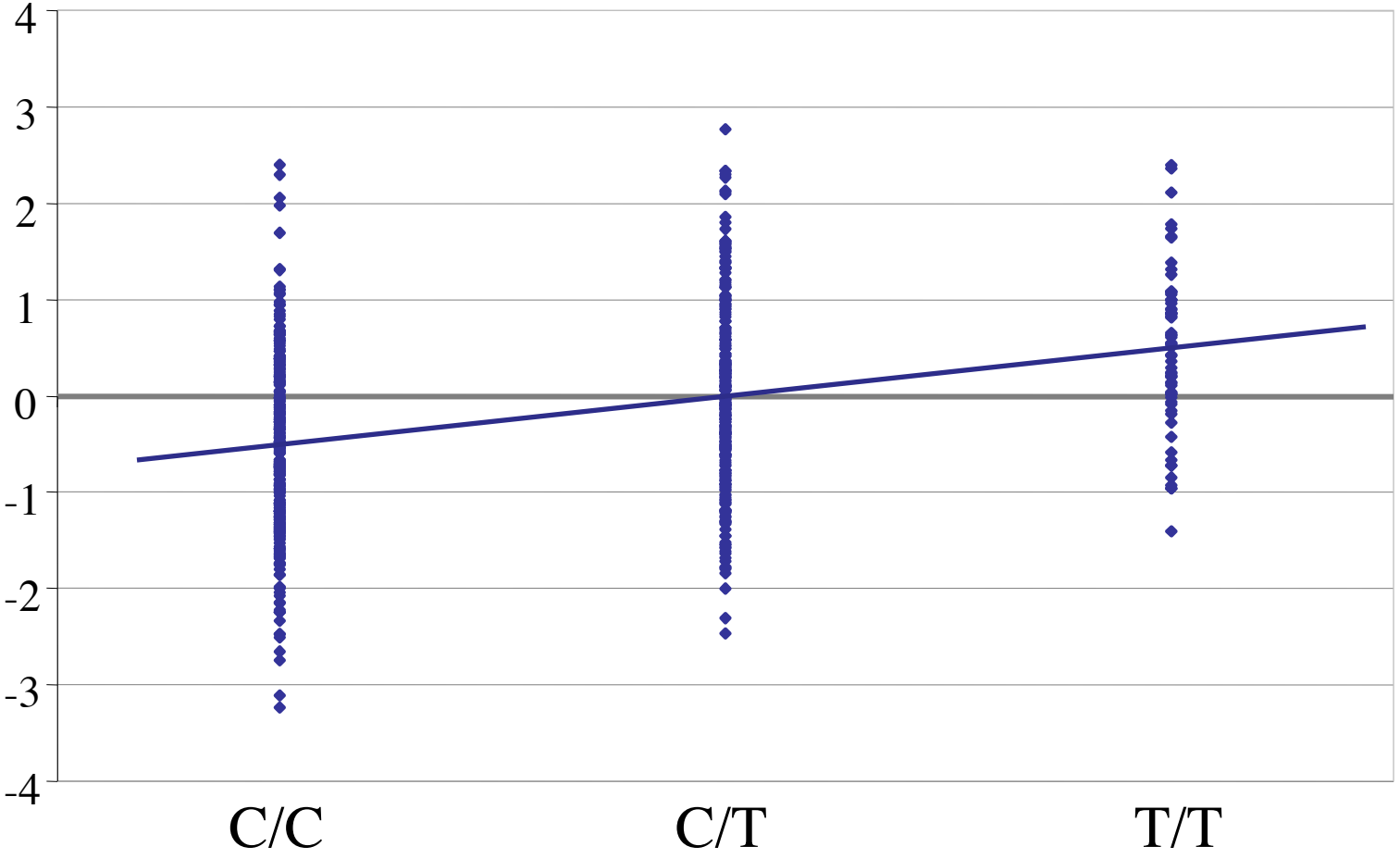


from **A tutorial on statistical methods for population association studies** by DJ Balding
Nature Reviews Genetics **7**, 781-791 (October 2006)

Two Types of Association Tests

- **Population-based** association tests
  - Cases and Controls are unrelated
    - Cross-classify by genotype
    - Use χ^2 test or logistic regression
  - Quantitative trait on population sample
    - Turn genotype into 0, 1, or 2 (count of reference allele)
    - Use linear regression
- **Family-based** association tests
  - Cases and Controls are related: parents, sibs etc
    - often based on allele transmission rates
    - example: Transmission Disequilibrium Test (TDT)
  - Quantitative trait on pedigree samples
    - Variance Components Analysis

Association between a Locus and a Quantitative Trait in a population-based sample



GWAS Model

In GWAS we are trying to find the best fit for the model:

$$Y = \mu + \beta X + \alpha Z + \varepsilon$$

where our data consists of the trait value ***Y***, the measured covariates ***X*** (such as BMI, smoking status, age, ethnicity, etc.), and the SNP genotypes ***Z***. We find estimates for the overall trait mean μ, the covariate effect size β, and the SNP effect size α, that together minimize the normal error term ε.

Historical Overview: CDCV

- In studies of *common, complex* traits (e.g., obesity, CVD, asthma, and neuropsychiatric disease) linkage methods are not as successful as with rare traits.
- One explanation is the Common Disease – Common Variant hypothesis (CDCV).

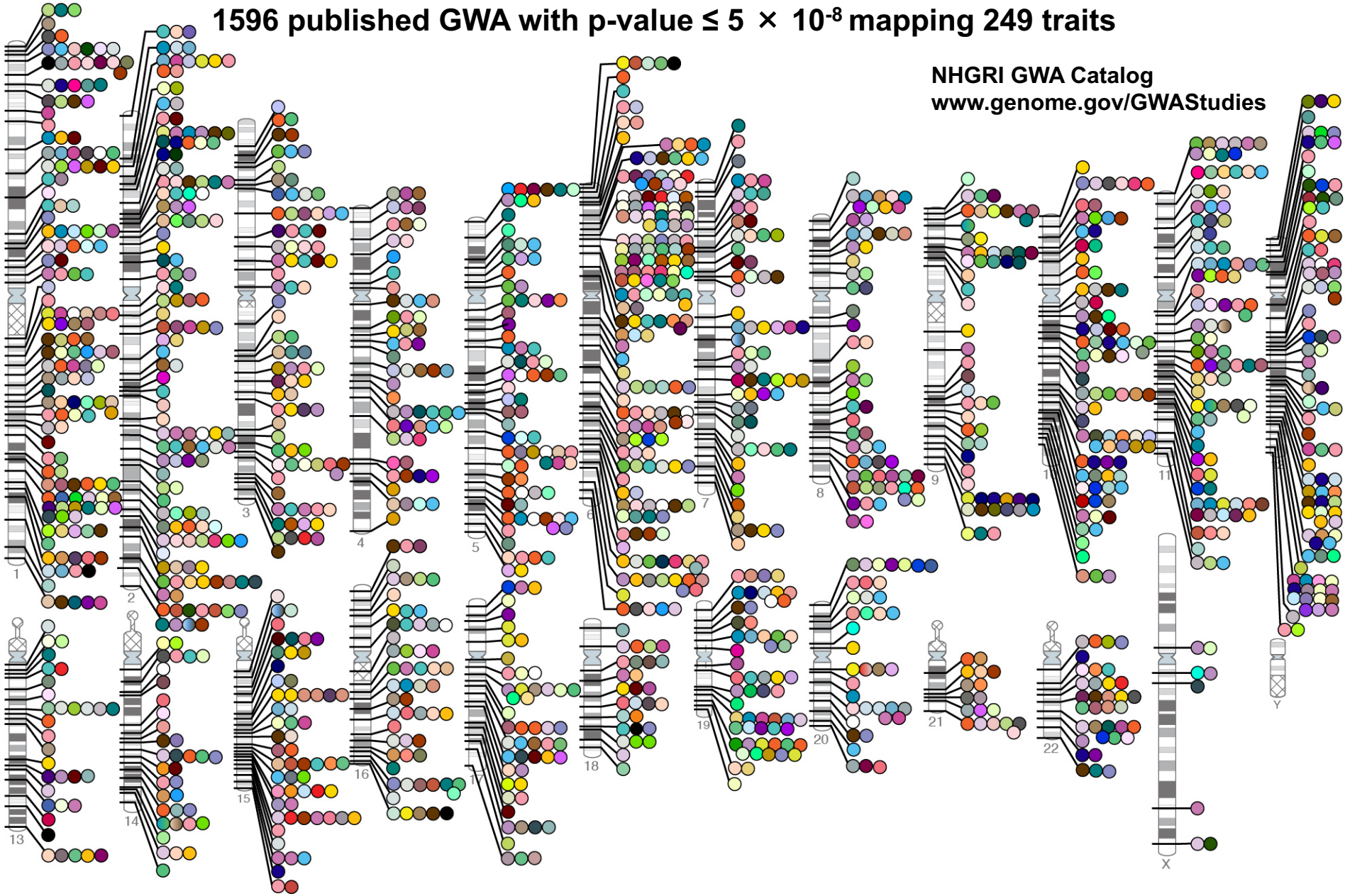
Historical Overview: CDCV

- CDCV posits that for common diseases there are at least several *common* gene-variants, and maybe many, all with relatively minor affect on the trait (*oligogenic model*), as well as possibly an overall background genetic affect (*polygenic model*).
- Due to their minor effect the variants have escaped selection pressure, allowing the trait to remain common.

Historical Overview: CDCV

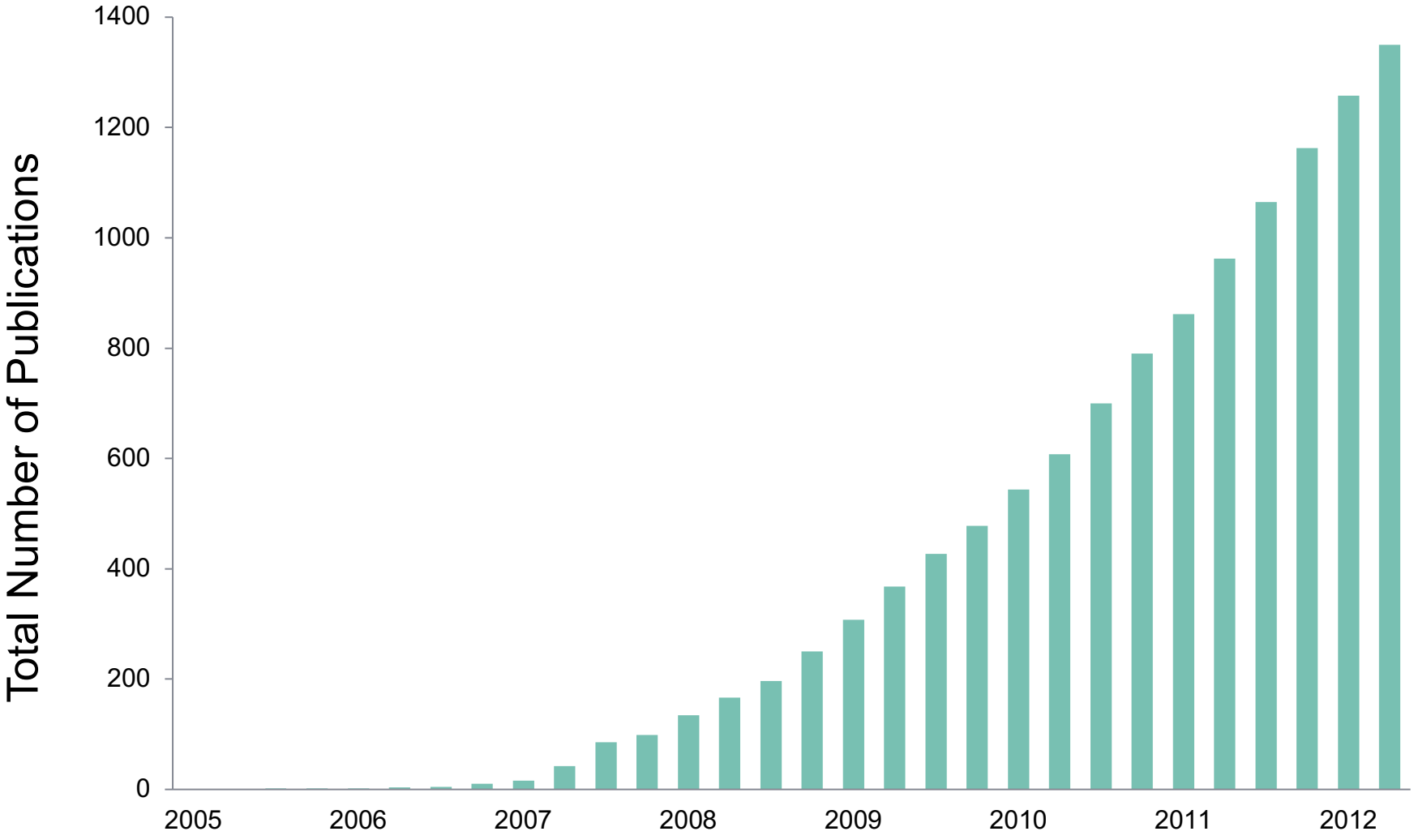
- CDCV hypothesis was popular in late 1990s and early 2000s.
- CDCV drove the push for genome-wide association technologies and studies.
- Many novel, replicated, common, susceptibility variants have been found based on these new technologies.

Published Genome-Wide Associations through 09/2011, 1596 published GWA with p-value ≤ 5×10^{-8} mapping 249 traits




Abdominal aortic aneurysm
Acute lymphoblastic leukemia
Adhesion molecules
Adiponectin levels
Age-related macular degeneration
AIDS progression
Alcohol dependence
Alopecia areata
Alzheimer disease
Amyloid A levels
Amyotrophic lateral sclerosis
Angiotensin-converting enzyme activity
Ankylosing spondylitis
Arterial stiffness
Asparagus anosmia
Asthma
Atherosclerosis in HIV
Atrial fibrillation
Attention deficit hyperactivity disorder
Autism
Basal cell cancer
Behcet's disease
Bipolar disorder
Biliary atresia
Bilirubin
Bitter taste response
Birth weight
Bladder cancer
Bleomycin sensitivity
Blond or brown hair
Blood pressure
Blue or green eyes
BMI, waist circumference
Bone density
Breast cancer
Butyrylcholinesterase levels
C-reactive protein
Calcium levels
Cardiac structure/function
Cardiovascular risk factors
Carnitine levels
Carotenoid/tocopherol levels
Carotid atherosclerosis
Celiac disease
Celiac disease and rheumatoid arthritis
Cerebral atrophy measures
Chronic lymphocytic leukemia
Chronic myeloid leukemia
Cleft lip/palate
Coffee consumption
Cognitive function
Conduct disorder
Colorectal cancer
Corneal thickness
Coronary disease
Cortical thickness
Creutzfeldt-Jakob disease
Crohn's disease
Crohn's disease and celiac disease
Cutaneous nevi
Cystic fibrosis severity
Dermatitis
DHEA-s levels
Diabetic retinopathy
Dilated cardiomyopathy
Drug-induced liver injury
Drug-induced liver injury (amoxicillin-clavulanate)
Endometrial cancer
Endometriosis
Eosinophil count
Eosinophilic esophagitis
Epirubicin-induced leukopenia
Erectile dysfunction and prostate cancer treatment
Erythrocyte parameters
Esophageal cancer
Essential tremor
Exfoliation glaucoma
Eye color traits
F cell distribution
Fibrinogen levels
Folate pathway vitamins
Follicular lymphoma
Fuch's corneal dystrophy
Freckles and burning
Gallstones
Gastric cancer
Glioma
Glycemic traits
Graves disease
Hair color
Hair morphology
Handedness in dyslexia
HDL cholesterol
Heart failure
Heart rate
Height
Hemostasis parameters
Hepatic steatosis
Hepatitis
Hepatitis B vaccine response
Hepatocellular carcinoma
Hirschsprung's disease
HIV-1 control
Hodgkin's lymphoma
Homocysteine levels
HPV seropositivity
Hypospadias
Idiopathic pulmonary fibrosis
IFN-related cytopeni
IgA levels
IgE levels
Inflammatory bowel disease
Insulin-like growth factors
Intracranial aneurysm
Iris color
Iron status markers
Ischemic stroke
Juvenile idiopathic arthritis
Keloid
Kidney stones
LDL cholesterol
Leprosy
Leptin receptor levels
Liver enzymes
Longevity
LP (a) levels
LpPLA(2) activity and mass
Lung cancer
Magnesium levels
Major mood disorders
Malaria
Male pattern baldness
Mammographic density
Matrix metalloproteinase levels
MCP-1
Melanoma
Menarche & menopause
Meningioma
Meningococcal disease
Metabolic syndrome
Migraine
Moyamoya disease
Multiple sclerosis
Myeloproliferative neoplasms
Myopia (pathological)
N-glycan levels
Narcolepsy
Nasopharyngeal cancer
Natriuretic peptide levels
Neuroblastoma
Nicotine dependence
Obesity
Open angle glaucoma
Open personality
Optic disc parameters
Osteoarthritis
Osteoporosis
Otosclerosis
Other metabolic traits
Ovarian cancer
Pancreatic cancer
Pain
Paget's disease
Panic disorder
Parkinson's disease
Periodontitis
Peripheral arterial disease
Personality dimensions
Phosphatidylcholine levels
Phosphorus levels
Photic sneeze
Phytosterol levels
Platelet count
Polycystic ovary syndrome
Primary biliary cirrhosis
Primary sclerosing cholangitis
PR interval
Progranulin levels
Progressive supranuclear palsy
Prostate cancer
Protein levels
PSA levels
Psoriasis
Psoriatic arthritis
Pulmonary funct. COPD
QRS interval
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Refractive error
Renal cell carcinoma
Renal function
Response to antidepressants
Response to antipsychotic therapy
Response to carbamazepine
Response to clopidogrel therapy
Response to hepatitis C treat
Response to interferon beta therapy
Response to metaformin
Response to statin therapy
Restless legs syndrome
Retinal vascular caliber
Retinol levels
Rheumatoid arthritis
Ribavirin-induced anemia
Schizophrenia
Serum metabolites
Skin pigmentation
Smoking behavior
Speech perception
Sphingolipid levels
Statin-induced myopathy
Stevens-Johnson syndrome
Stroke
Sudden cardiac arrest
Suicide attempts
Systemic lupus erythematosus
Systemic sclerosis
T-tau levels
Tau AB1-42 levels
Telomere length
Testicular germ cell tumor
Thyroid cancer
Thyroid volume
Tooth development
Total cholesterol
Triglycerides
Tuberculosis
Type 1 diabetes
Type 2 diabetes
Ulcerative colitis
Urate
Urinary albumin excretion
Urinary metabolites
Uterine fibroids
Venous thromboembolism
Ventricular conduction
VEGF levels
Vertical cup-disc ratio
Vitamin B12 levels
Vitamin D insuffiency
Vitamin E levels
Vitiligo
Warfarin dose
Weight
White cell count
White matter hyperintensity
YKL-40 levels

Published GWAS with p-value ≤ 10^{-5}

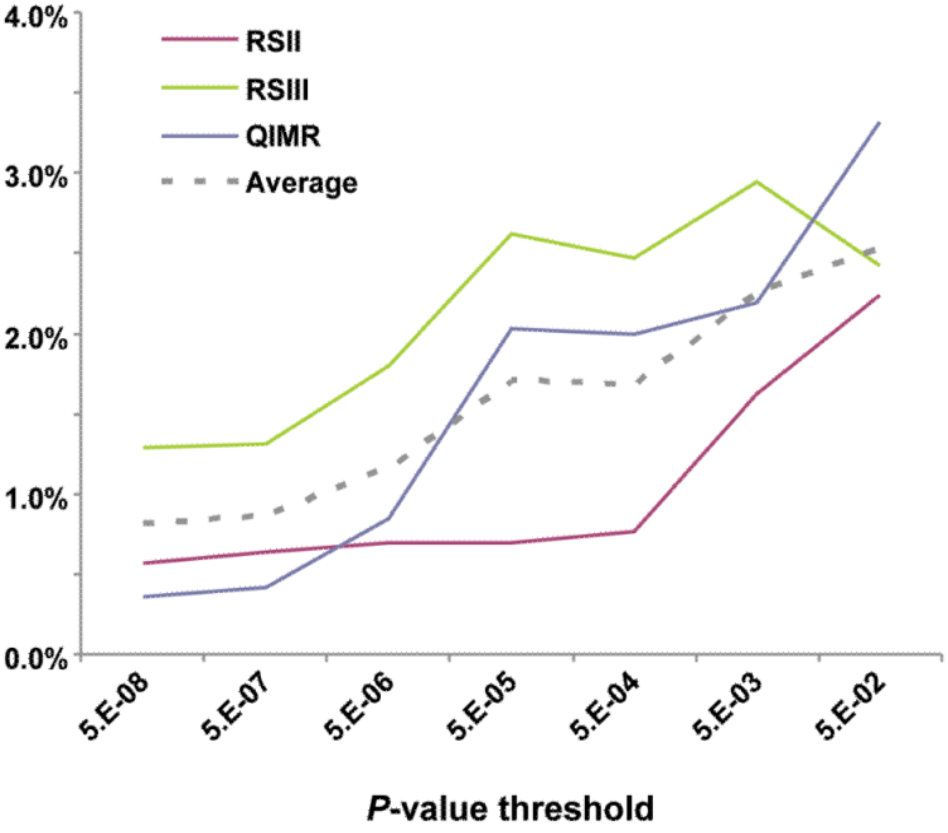


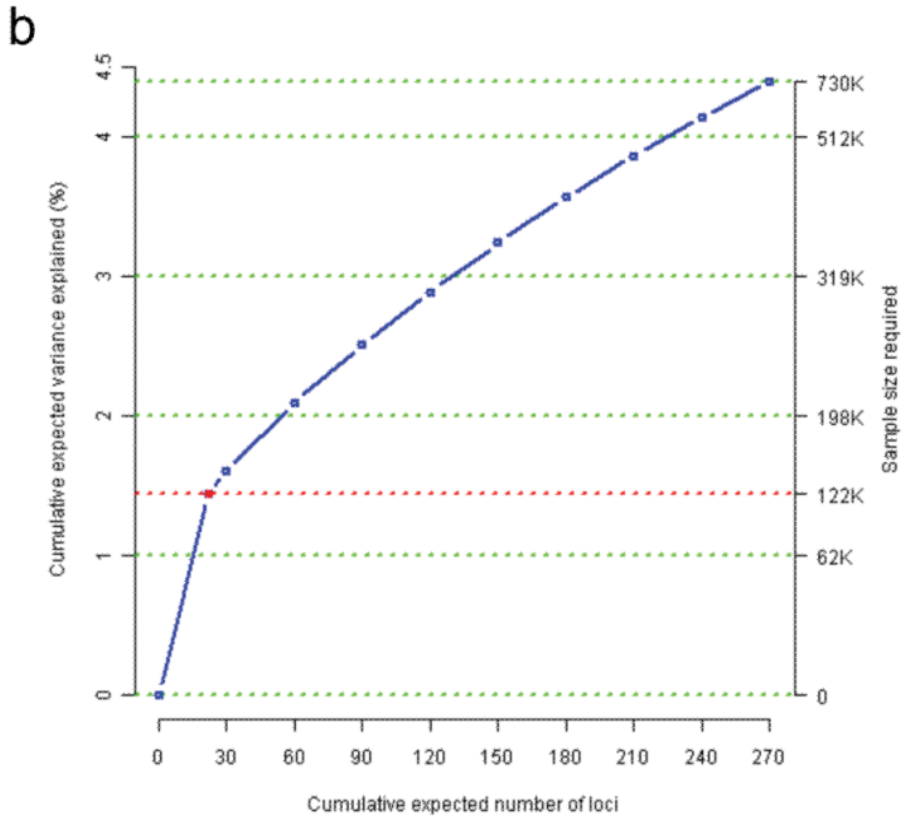
GWAS Results

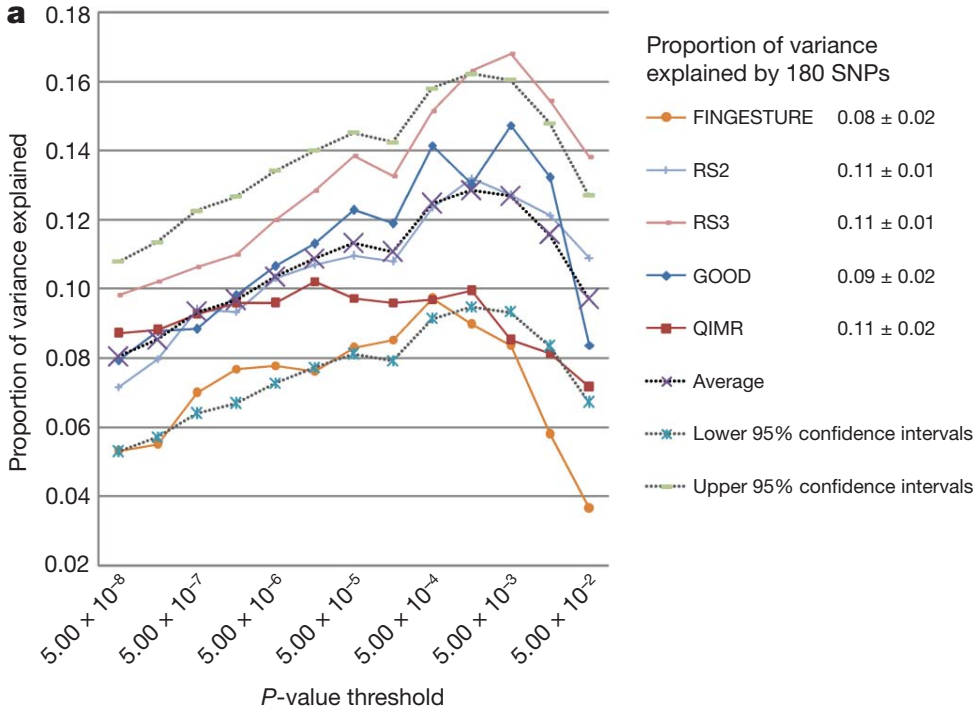
- Many common variants relevant to common traits have been found via GWAS.
- These are often not replicated in other studies on the same or different populations. (Why? The winner's curse.)
- The small effect sizes (usually < 5%) attributable to these regions clearly indicate there are many other causes for these common traits.
- Where is the genetic "dark matter" hiding?

GWAS almost always finds SNPs with small effect sizes

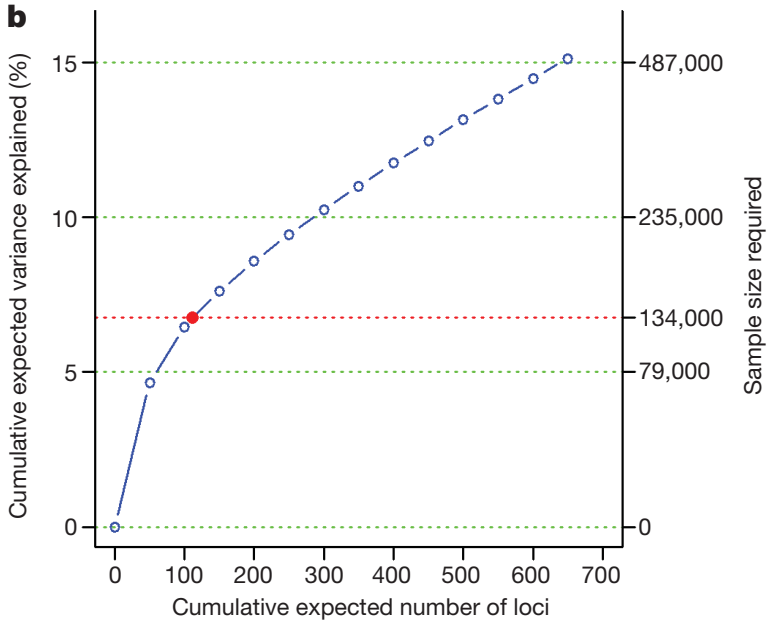
In a study of nearly 250,000 individuals, over 30 SNPs were found that influence normal *BMI*. However, all have a very small effect. [Speliotes et al. Nature Genetics, Nov 2010.]







GWAS almost always finds SNPs with small effect sizes



In a study of nearly 200,000 individuals, nearly 200 SNPs were found that influence normal *height*. Again, all have a very small effect. [Lango-Allen et al. Nature Oct 2010]

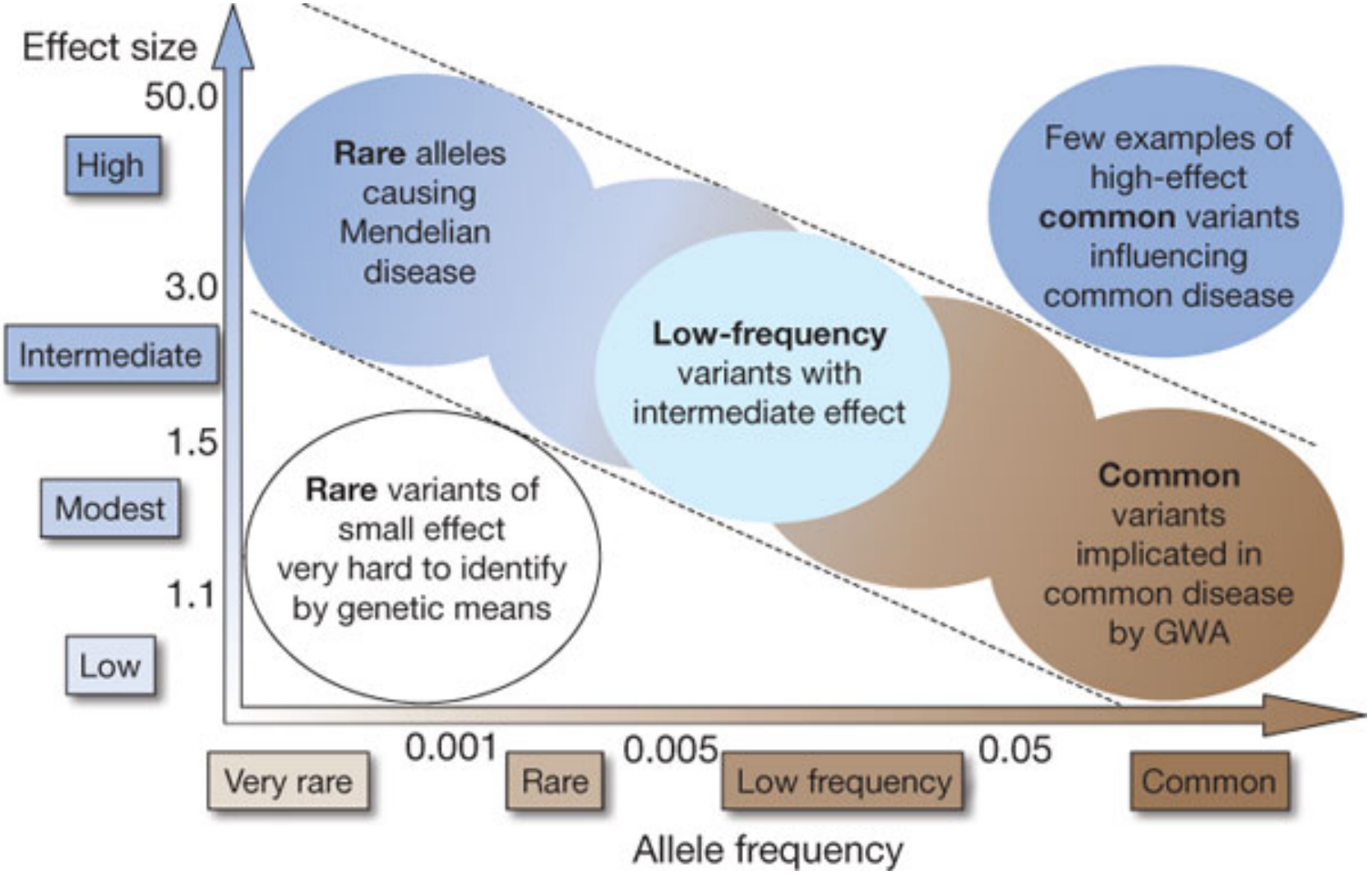
Even Bigger Data Arriving Soon

- The VA's Million Veteran Program (www.research.va.gov/mvp/) already contains GWAS data (657,459 SNPs) on 359,964 veterans. Simply storing the compressed genotype data requires 55 GB.
- A recent initiative sequenced 10,545 human genomes at 30 – 40x coverage at a cost under $2000 per genome. They identified > 150 million variants, the vast majority of which are rare or de novo.
- The iPOP (integrative personal omics profile) study followed a single individual for 401 days and collected transcriptome, proteome, metabolome, microbiome, epigenome, exposome, and clinical tests at 20 time points, along with > 100x coverage whole genome sequence. This type of "omics" profiling yields a dynamic picture of the extensive changes between healthy and diseased.

Dark Matter of Genetic Epidemiology?

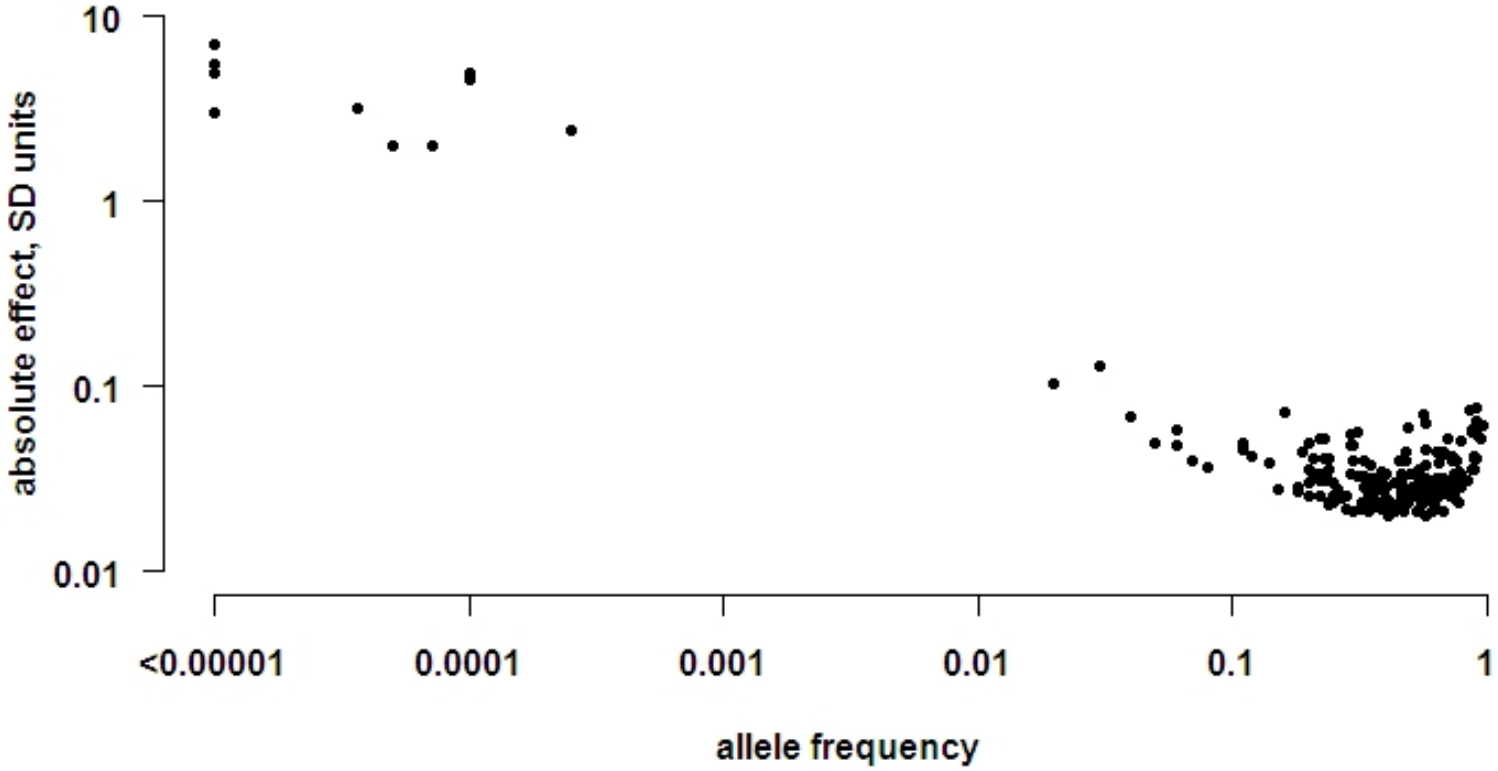
- copy number variants (CNVs)
- epigenetic effects (e.g., methylation patterns)
- rare variants (CDRV hypothesis)
- polygenes of small effect
- variation across populations
- interactions among genes and between genes and environment
- miscalculation of size of total genetic effect

Association and Linkage are Complementary Methods



TA Manolio, *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Example: Variants Found for Body Size



Mutations with intermediate effect (0.1 to 1 SD units) and low frequency (0.01 to 0.001) are not detected efficiently by either linkage or genome-wide association studies. Shown are the allele effect size (y-axis) and frequency (x-axis) for GWAS results and for the sample of mutations described [in the paper].

Kemper, Visscher, and Goddard (2012) *Genome Biology* **13**:244 [doi:10.1186/gb-2012-13-4-244]

Example GWAS

- We will use a simulated data set of 2200 unrelated people and 10,000 SNPs (single nucleotide polymorphisms, i.e., easily assayed, relatively common, variants) spread across the genome.
- The trait was simulated as influenced by two of the SNPs and their interaction.