Evaluating potential bias in and interpreting results from epidemiologic designs for genome-wide association studies

Ellen M. Wijsman, Ph.D.
Dept. Biostatistics and Div. Medical Genetics
University of Washington

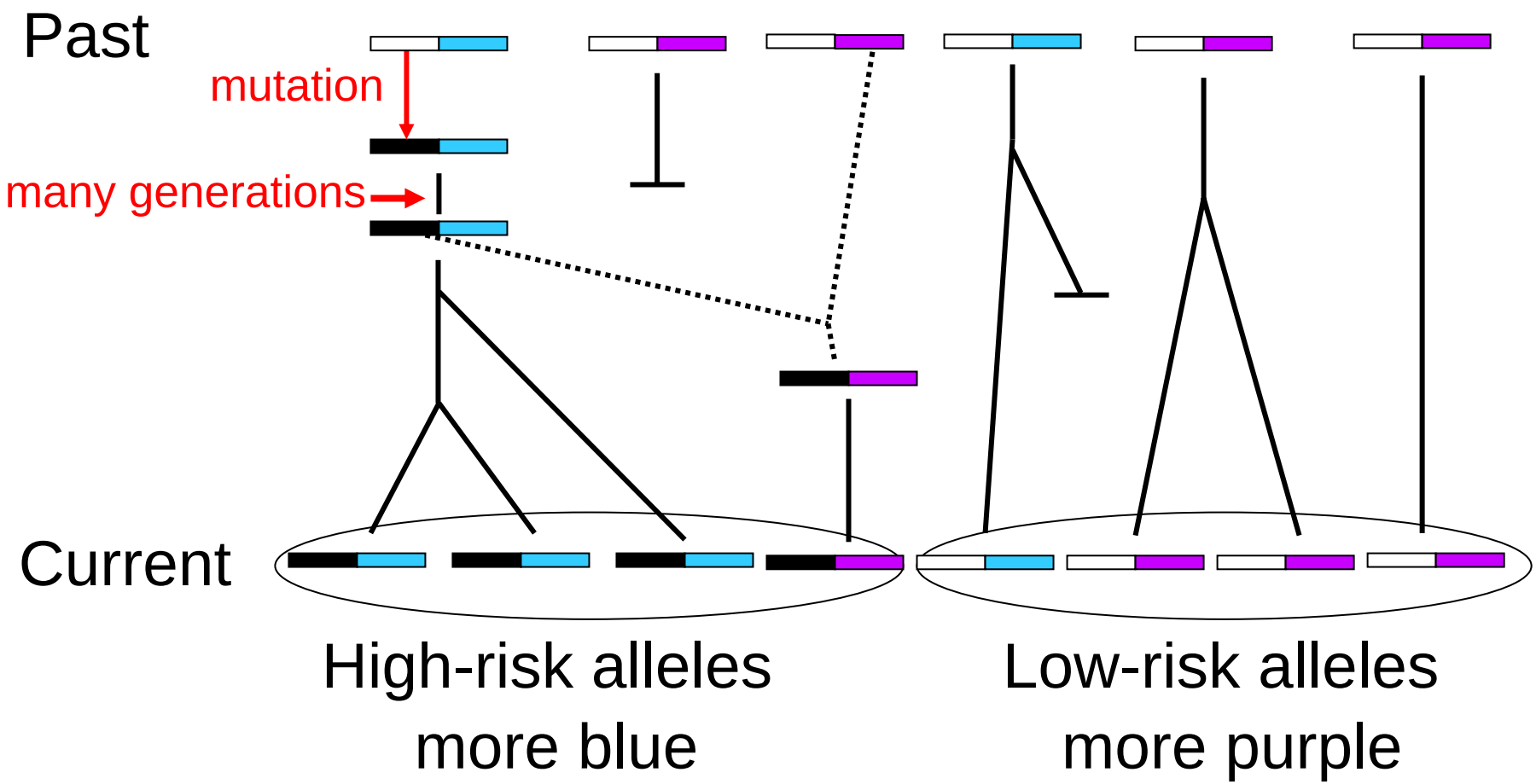
Genetic association studies

- Associations depend on gene histories: markers *and* traits
- Gene histories introduce data structure
- Good design requires an understanding of the potential causes of data structure
- Design, analysis and interpretation must accommodate data structure

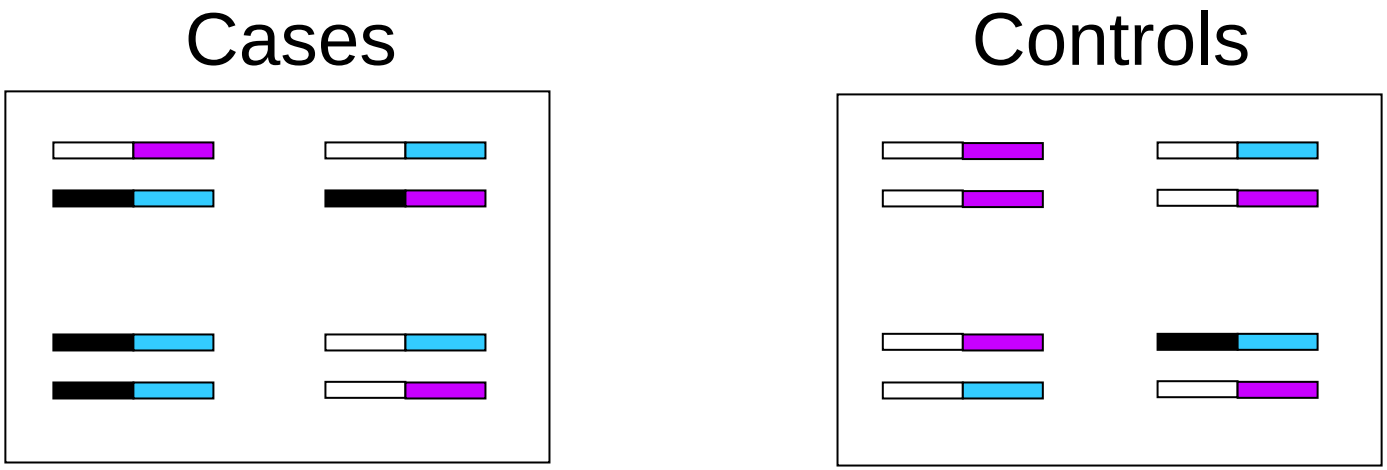
Causes of genetic data structure

- Chromosome history
  - Linkage disequilibrium
- Non-random mating and population history
  - Population structure
- Finite population size
  - Cryptic relatedness
- Sampling through cases
  - Cryptic relatedness

Chromosome history produces linkage disequilibrium (LD)



We sample people, not alleles



Ability to detect association depends on:

- Trait mode of inheritance
  - Genotype penetrances
  - Locus/allelic heterogeneity
- Distance between marker and trait locus
- Age of mutation

(Chapman & Wijsman 1998 AJHG 63:1872-1885)

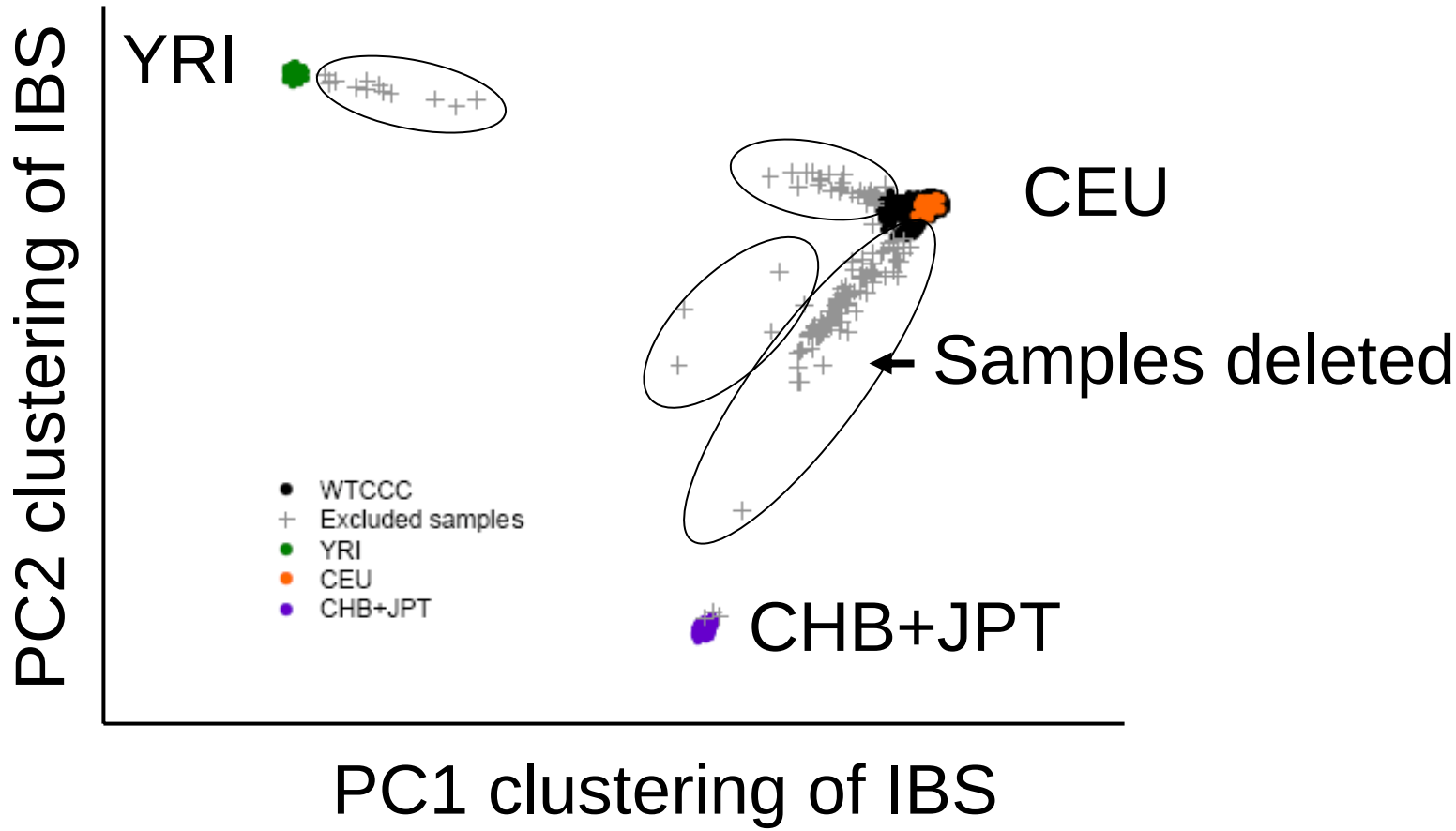
Choice of population

- Large and outbred (e.g., US, Britain)
  - ➢High heterogeneity (genetic and environmental)
  - ➢Weaker association
  - ➢Large available sample sizes
    - – Many choices for subgrouping
- More isolated populations (e.g, Finland)
  - ➢Less heterogeneity
    - – Fewer disease alleles
    - – Less environmental variation
  - ➢Stronger association
  - ➢Smaller available sample sizes

Examples

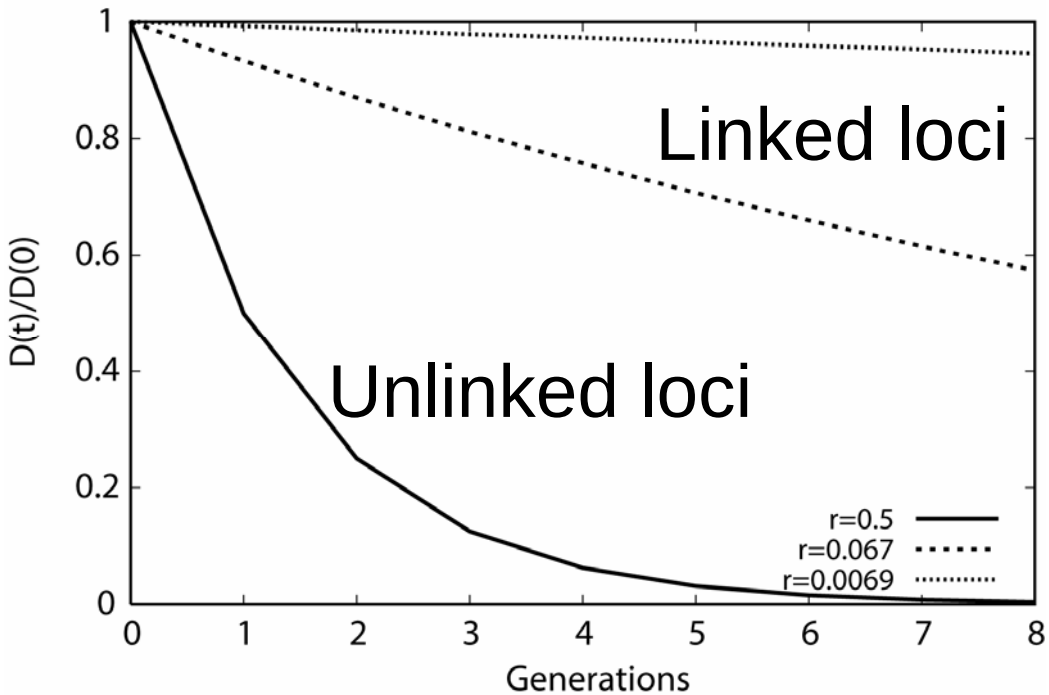
- Great Britain, Welcome Trust (WTCCC)
  - Caucasian: Total population ~60 million
  - 2000 of each of 7 case populations
  - 3000 common controls
  - SNP genome scan
- Guam CC (Univ. of WA, UCSD, Guam)
  - Chamarro: Total population ~45,000
  - 140 cases with neurodegenerative disease
  - 88 elderly unaffected controls
  - STRP genome scan

Ancestry is not always accurate



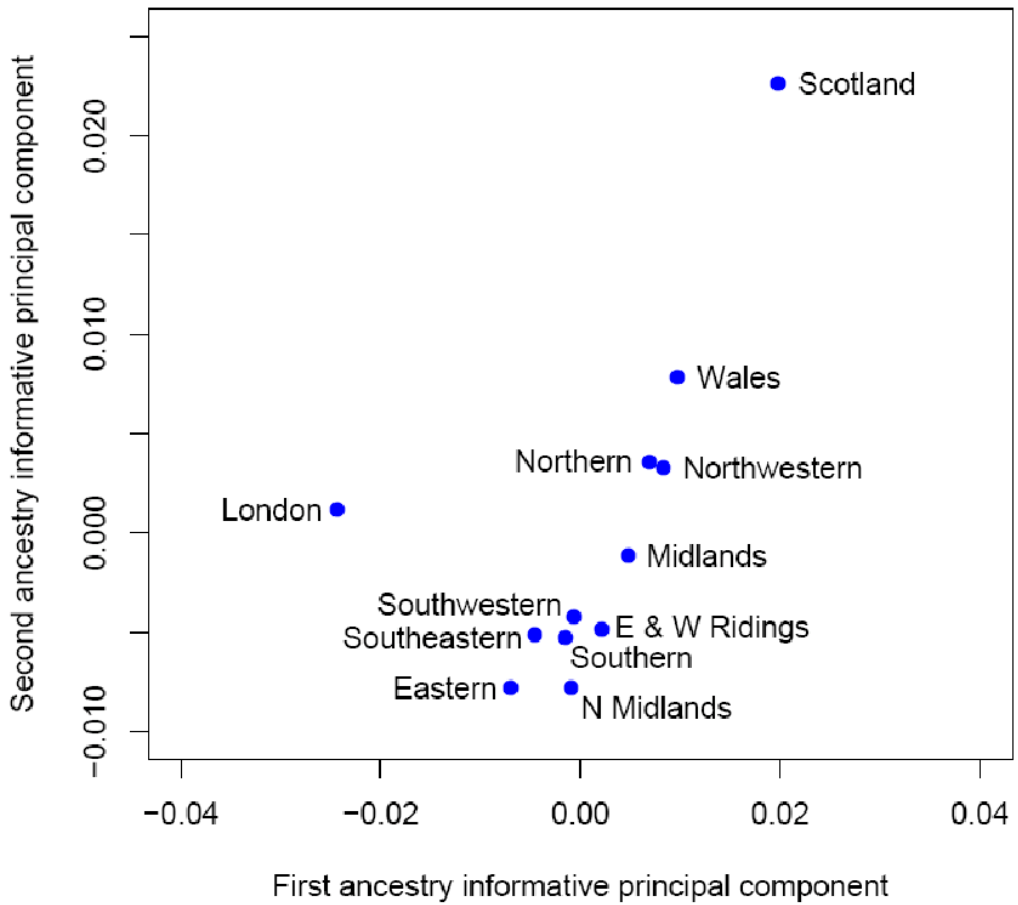
(WTCCC 2007, Nature 447:661-678)

LD decays under random mating



- True random mating rarely occurs
  - Geographical location associated with genotype
  - 1800's: Spouses' birthplaces avg. 6-10 km apart in Europe, US
- Elimination of LD takes longer
- Some geographic substructure is typical

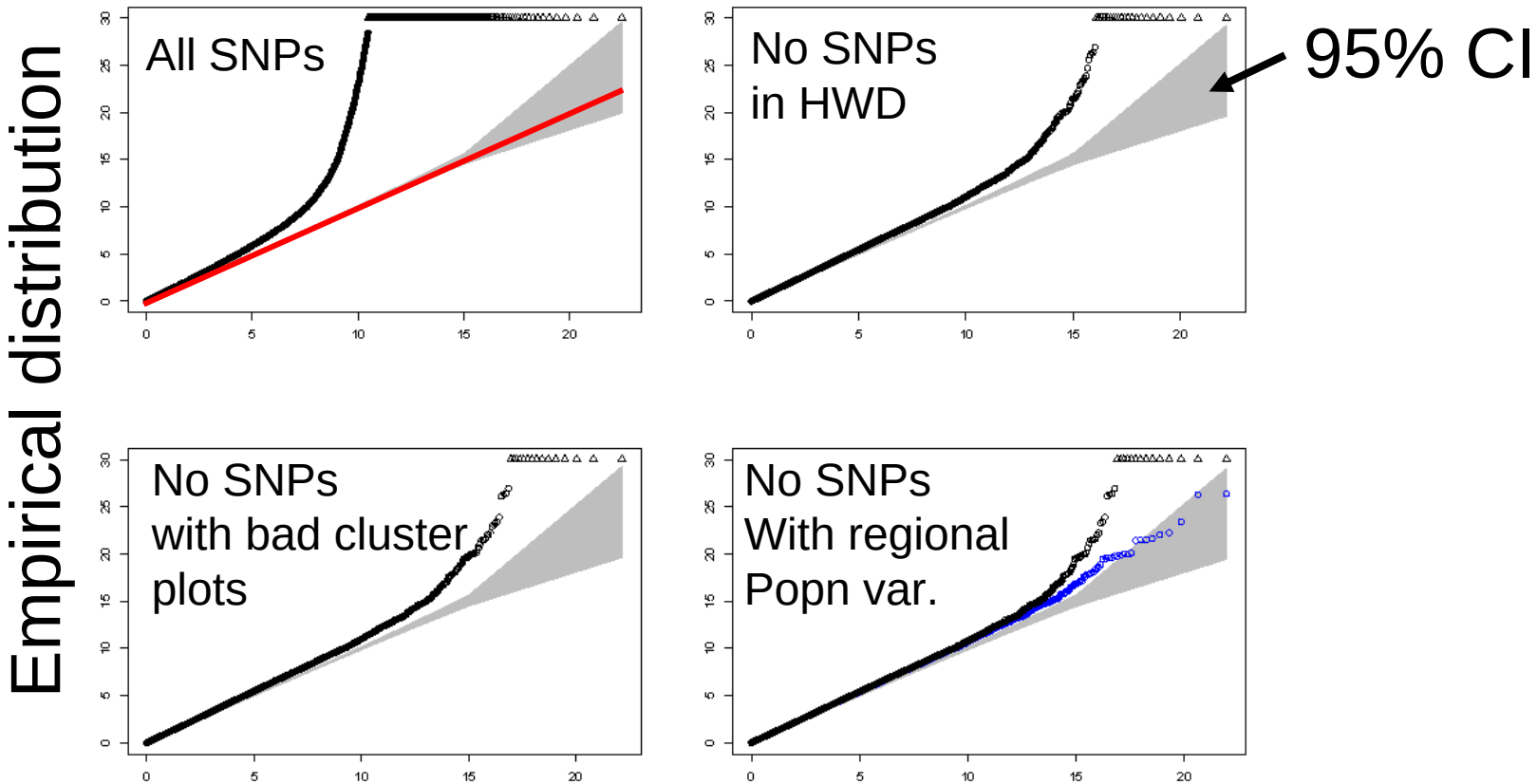
Population structure is unavoidable



(WTCCC 2007, Nature 447:661-678)

Extensive analysis required to minimize spurious association

GWAS trend test results, Type 2 diabetes

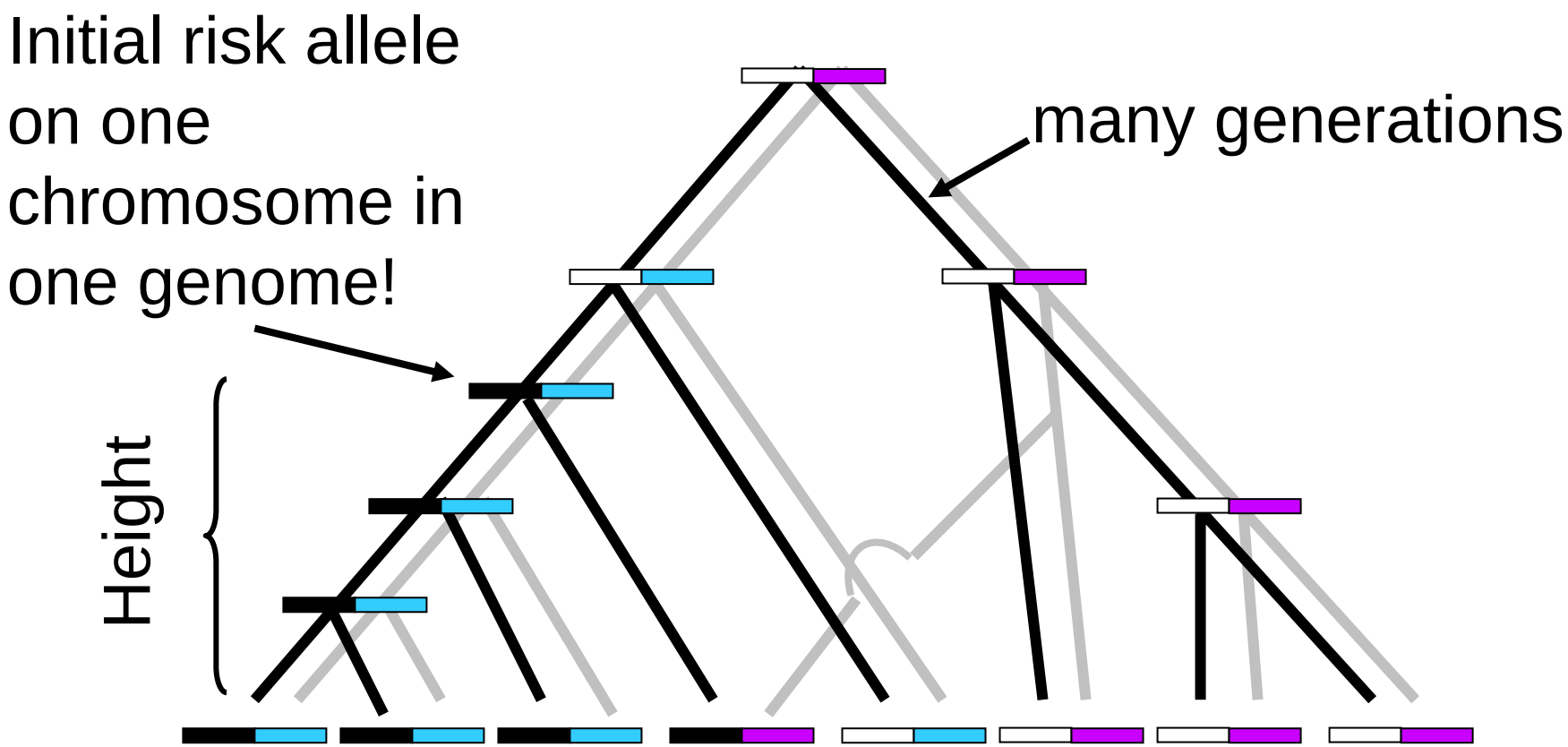


(WTCCC 2007, Nature 447:661-678)

Human history

- Population structure
  - Frequent waves of migration/conquest
  - Low spousal birth distances: nonrandom mating
- World population increase is recent
  - 1 AD: ~300 million
  - 1650: ~500 million
  - 1850: ~1.2 billion
  - 2000: ~6 billion
- Many or most human risk alleles are recent
  - >5% of humans *ever* born are alive today
  - Surviving risk alleles had even faster growth rate
    (Thompson & Neel 1997 AJHG 60:197-204)
  - Many risk alleles have a “short” genealogical tree

Genealogy of chromosomes



- Short “tree” among cases: cases tend to be related
- Shorter trees among rapidly expanding populations

(Voight and Pritchard 2005 PLOS Genet 1:e32)

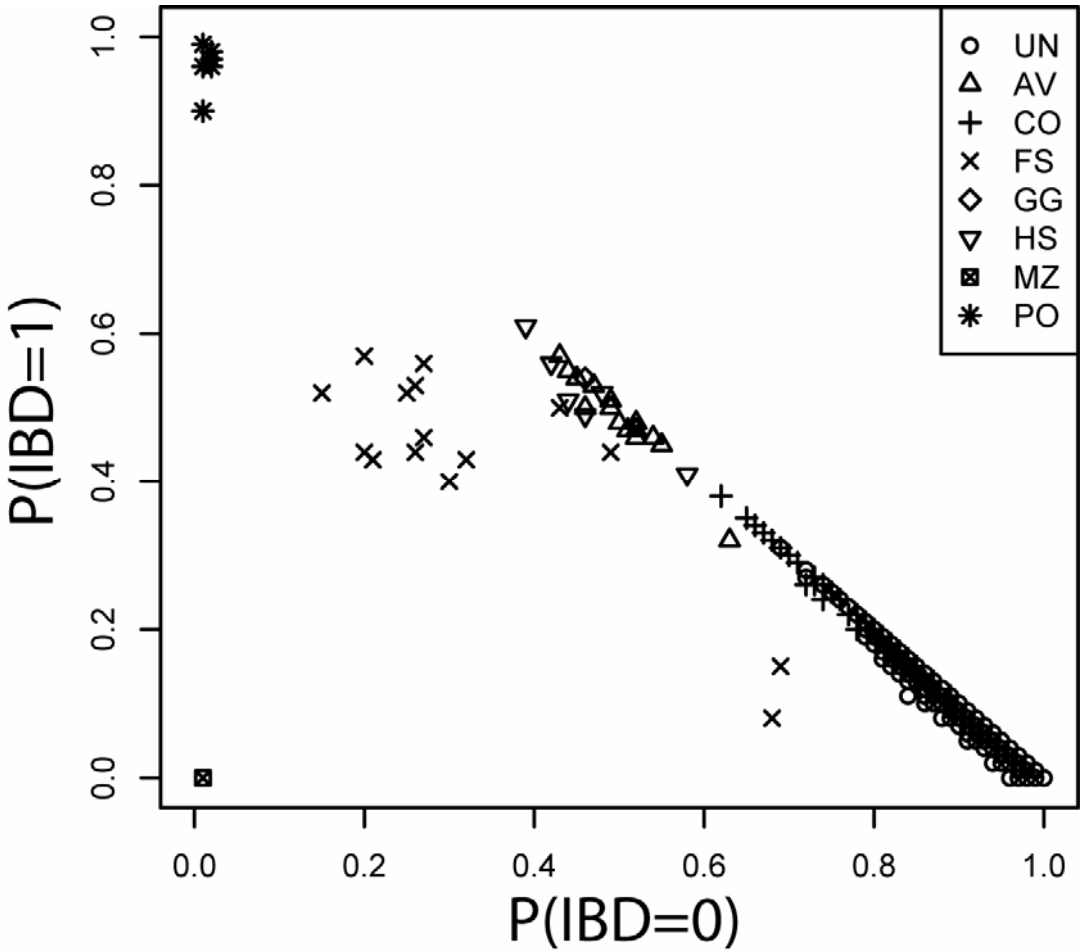
Cryptic relatedness

- Cases and controls drawn from one population
- Sampling through (rarer) cases selects a short branch of the gene (coalescent) tree
- The short tree leads to cases being more related than controls
- In finite populations, controls may also be related (also short tree!)
- Consequence: correlated data, giving inflated variance over that assumed
  - leads to incorrect p-values in statistical tests

(Voight & Pritchard 2005 PLOS Genet 1:e32)

Finite samples include relatives

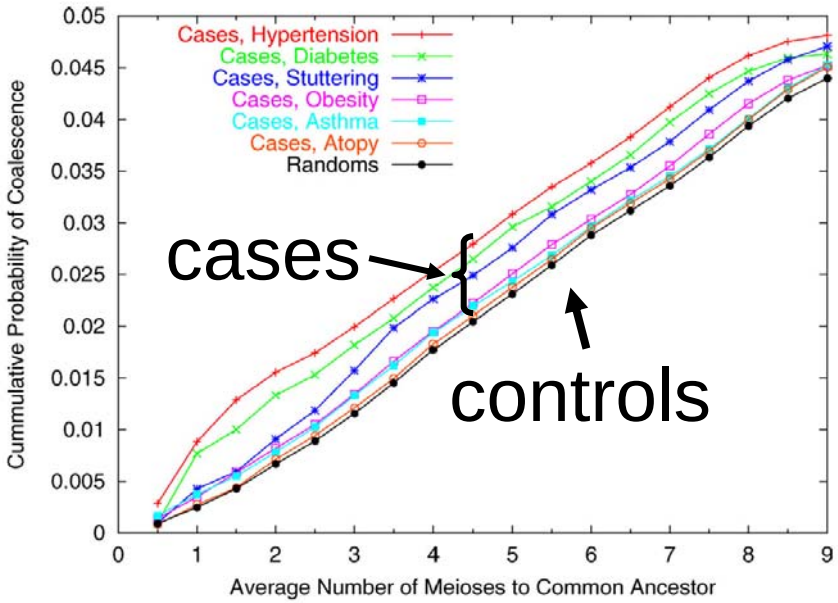
Guam CC



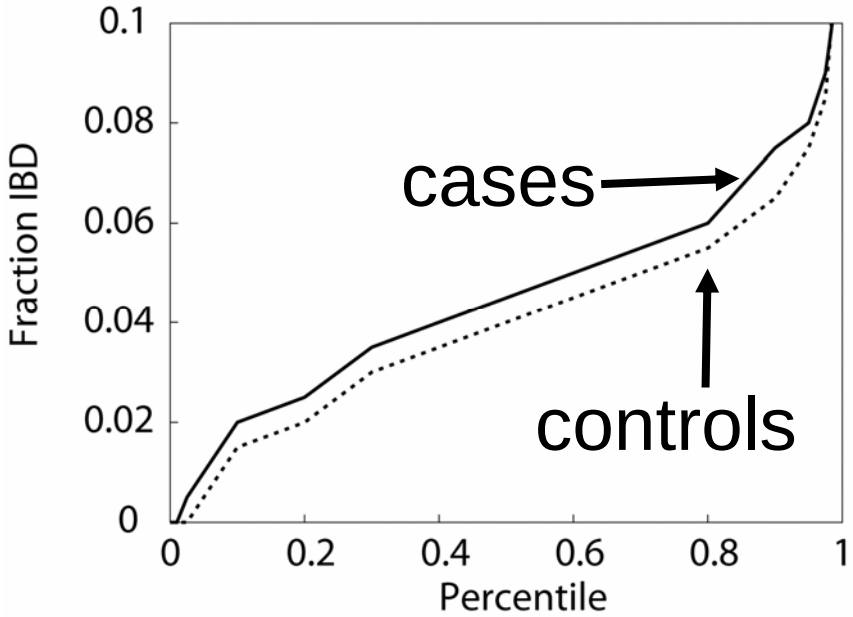
No dichotomy in relationship inferences

Cases show excess relatedness

Hutterites
known relationships



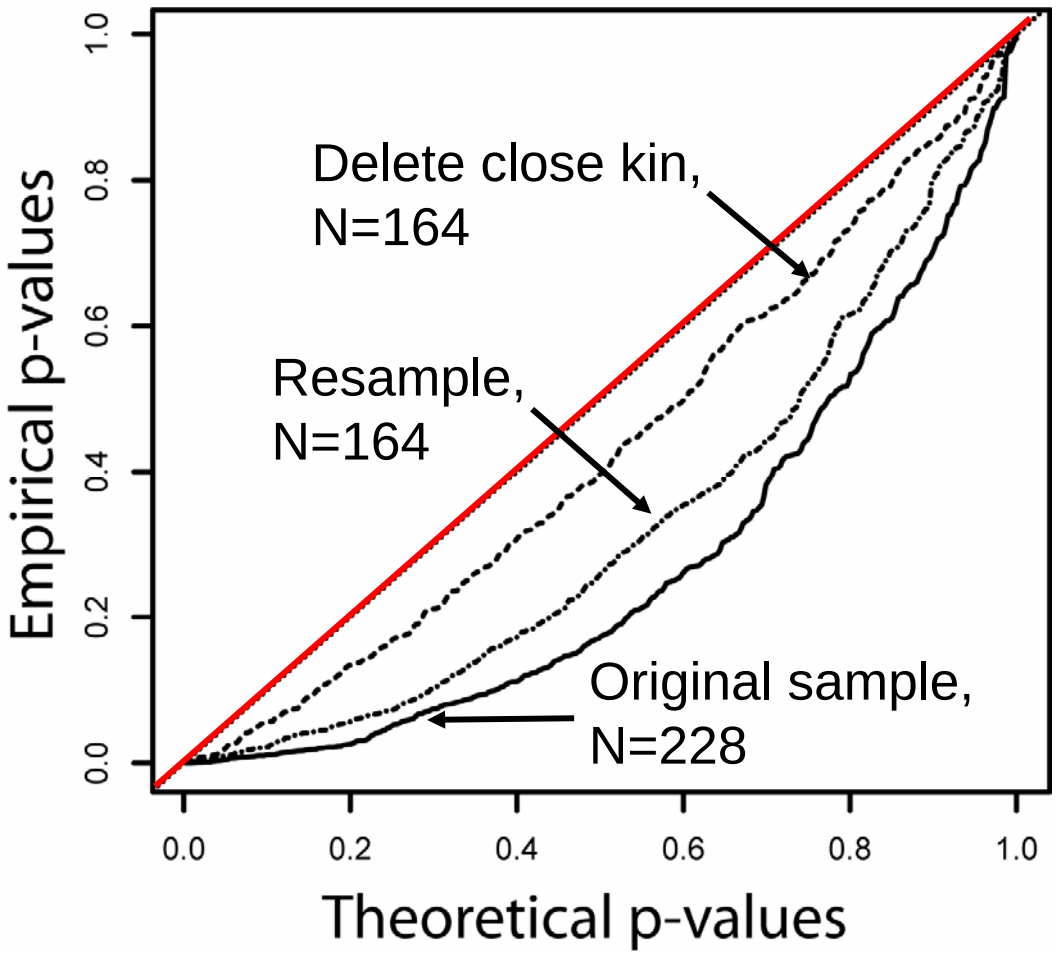
Guam CC
estimated relatedness



(Voight & Pritchard, 2005
PLOS Genet 1:e32)

Relatedness affects tests

Guam CC: Fisher's exact test



Comments and Summary

- Stringent test significance levels required
  - Accuracy of tail of distribution of test statistic is important
  - If inaccurate, how to interpret results?
- Violation of assumptions leads to erroneous distributions of test statistics
  - Leads to incorrect inference/interpretation
- Data structure is unavoidable: violates assumptions
  - Population substructure
  - Cryptic relatedness
- Careful evaluation of effects of possible violation of assumptions/distributions is important
  - Internal consistency of data/results can be evaluated
- Analyses that incorporate the data structure are critical
  - No amount of careful design will completely eliminate the structure