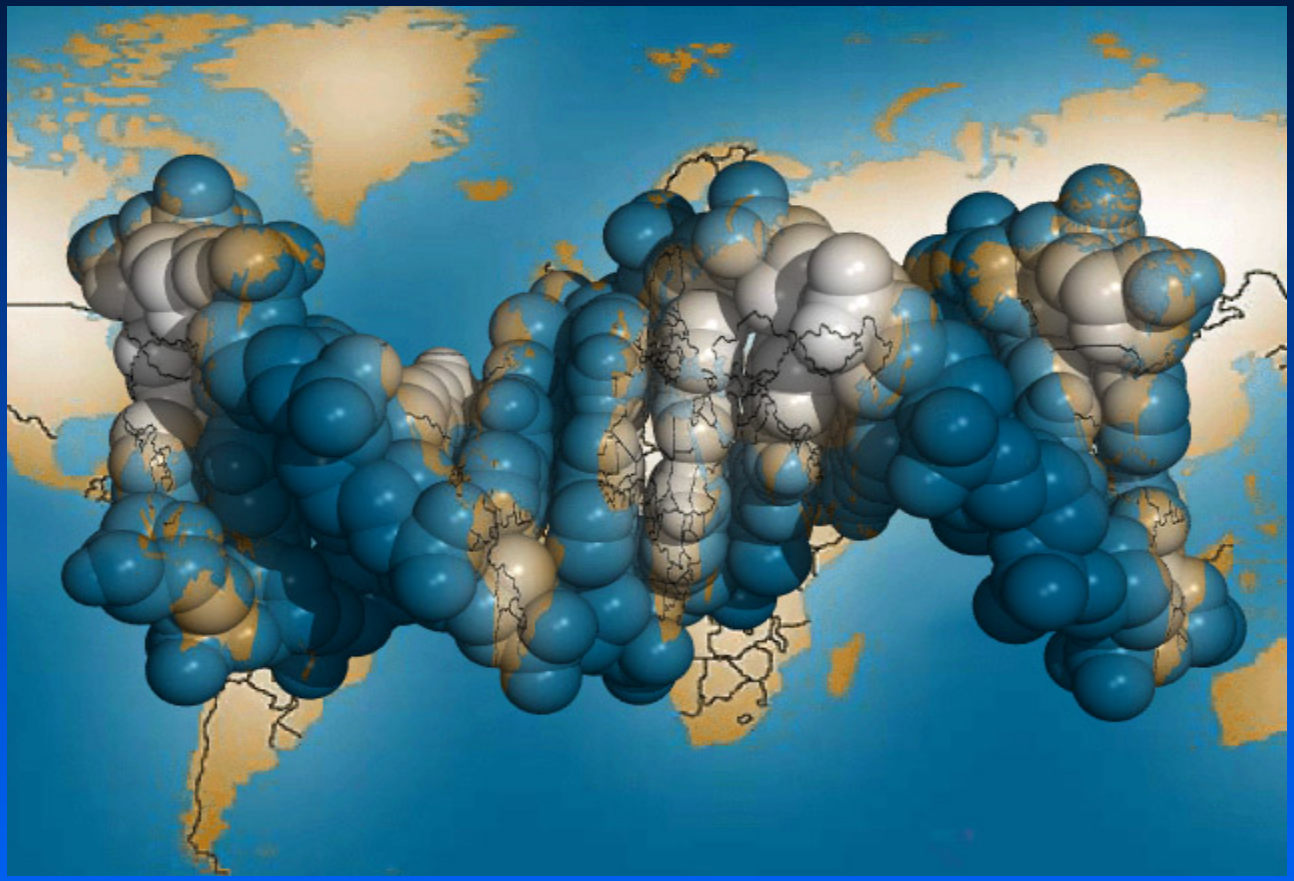
Ancestral Diversity



Lynn B. Jorde
Department of Human Genetics
University of Utah School of Medicine
29 June 2012

Overview

- Background on human genetic variation
- How sequencing has changed things

How is genetic variation distributed among continental populations?

	60 STRs	30 RSPs	100 *Alus*	75 L1s	250K SNP	
Between individuals, within continents	90%	87%	86%	88%	88%	
Between continents (F_{ST})	10%	13%	14%	12%	12%	

F_{ST}: proportion of variation attributed to population subdivision

Jorde *et al*., 2000, *Am. J. Hum. Genet.*
J. Xing *et al*., 2009, *Genome Res*.

How is genetic variation distributed among continental populations?

	60 STRs	30 RSPs	100 *Alus*	75 L1s	250K SNP	Skin pigment-ation
Between individuals, within continents	90%	87%	86%	88%	88%	10%
Between continents (F_{ST})	10%	13%	14%	12%	12%	90%

Jorde *et al*., 2000, *Am. J. Hum. Genet.*

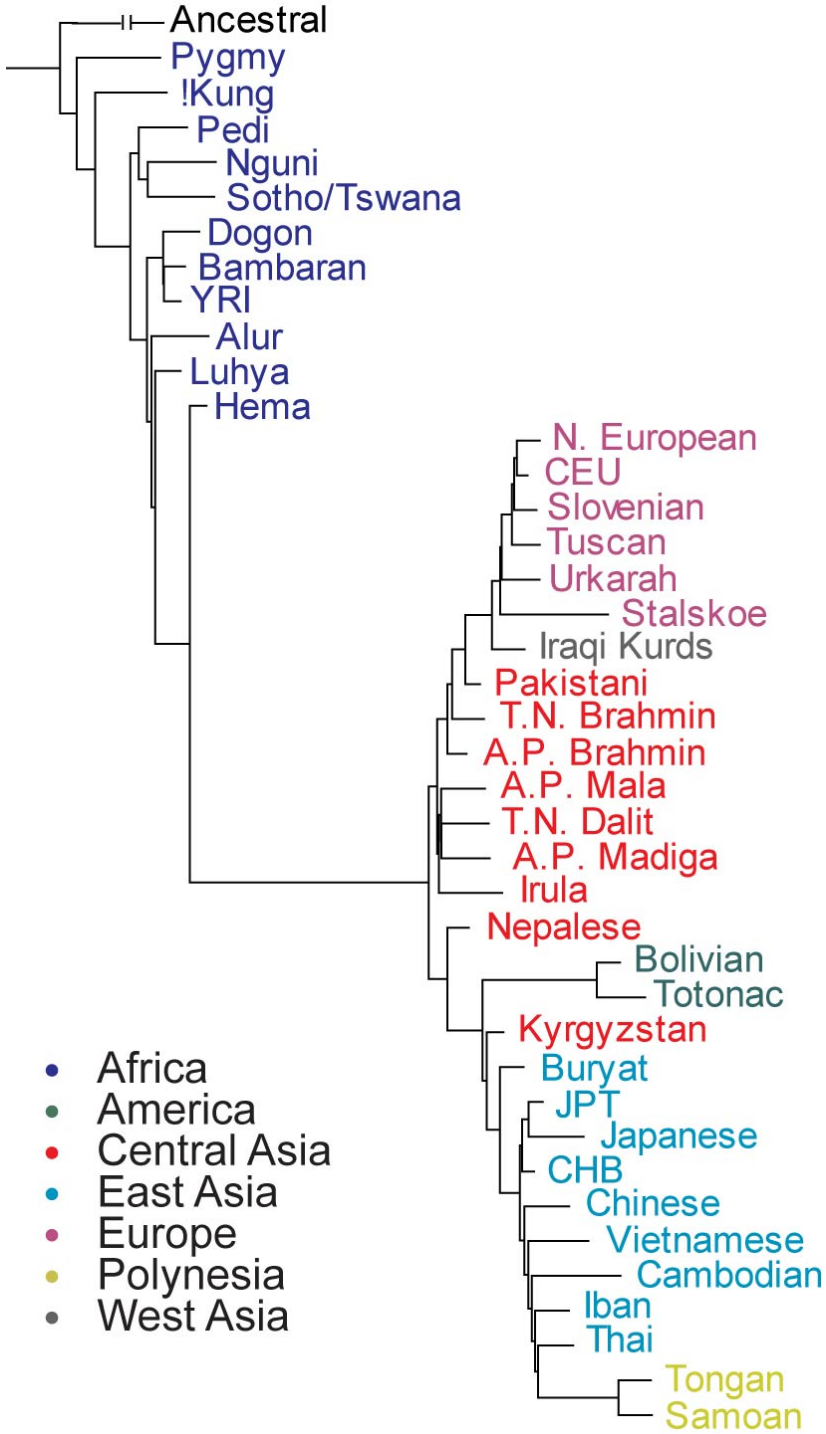
J. Xing *et al*., 2009, *Genome Res.*

% SNPs shared among four major regions (Africa, Europe, E. Asia, India): 250K chip results for ~1,000 samples

Minor allele present in:	
All 4 groups	78.6%
At least 3 groups	88.0%
At least 2 groups	92.1%
Africa only	7.4%
Any non-African group	0.5%

No SNPs were fixed present in one population, fixed absent in another

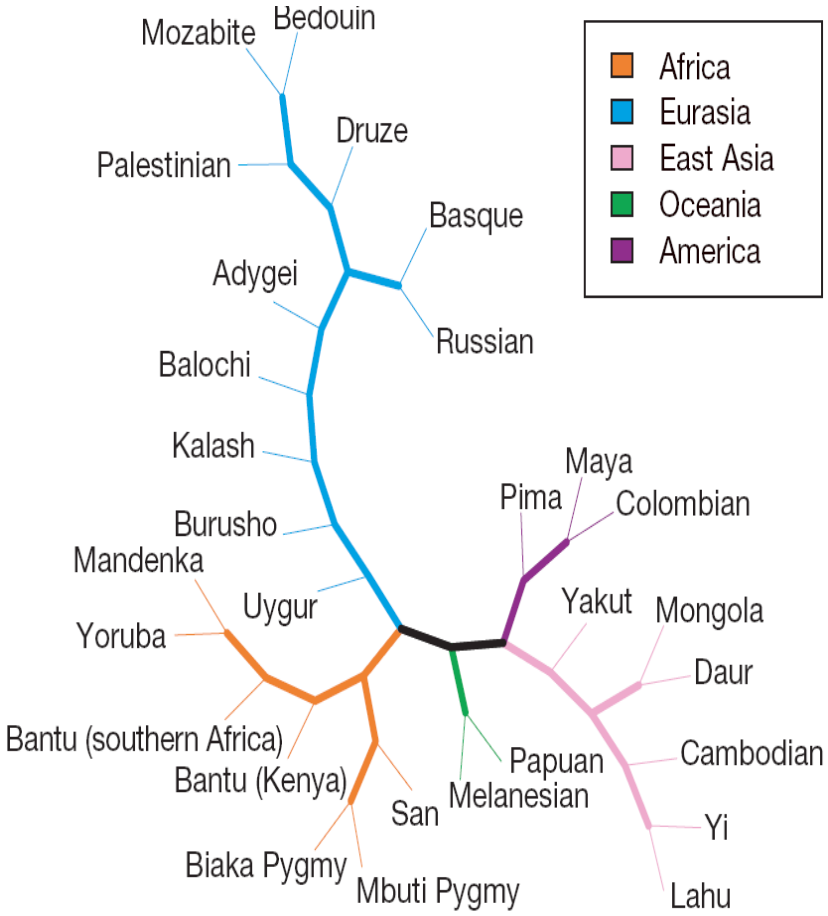
J. Xing *et al*., 2010, *Genomics*



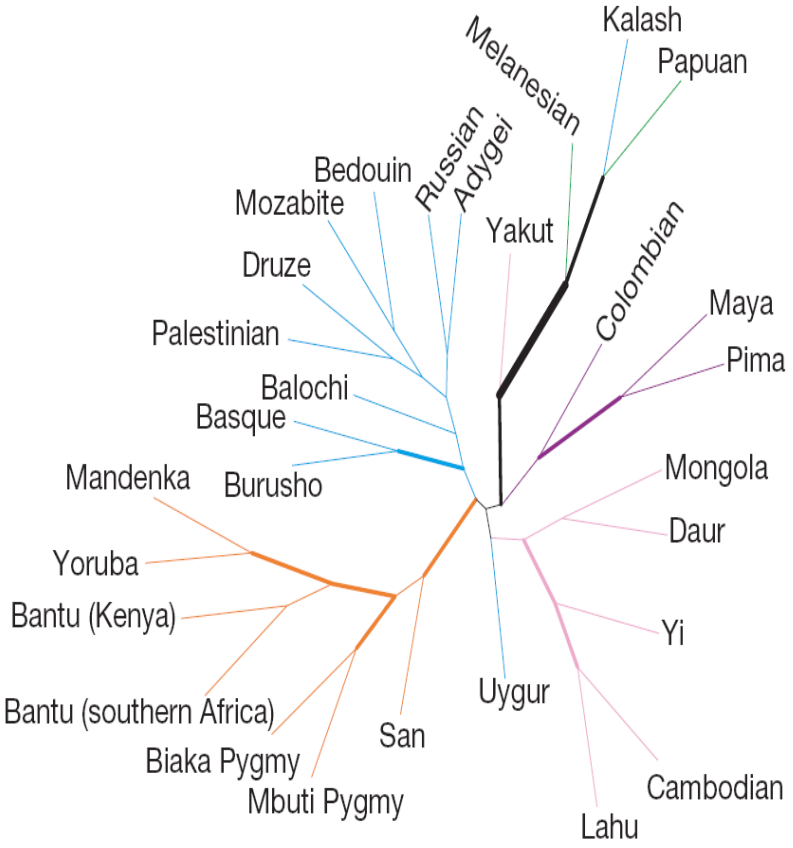
40 populations, ~250K SNPs

Xing et al., 2010, *Genomics*

Population relationships in HGDP sample



525,910 SNPs

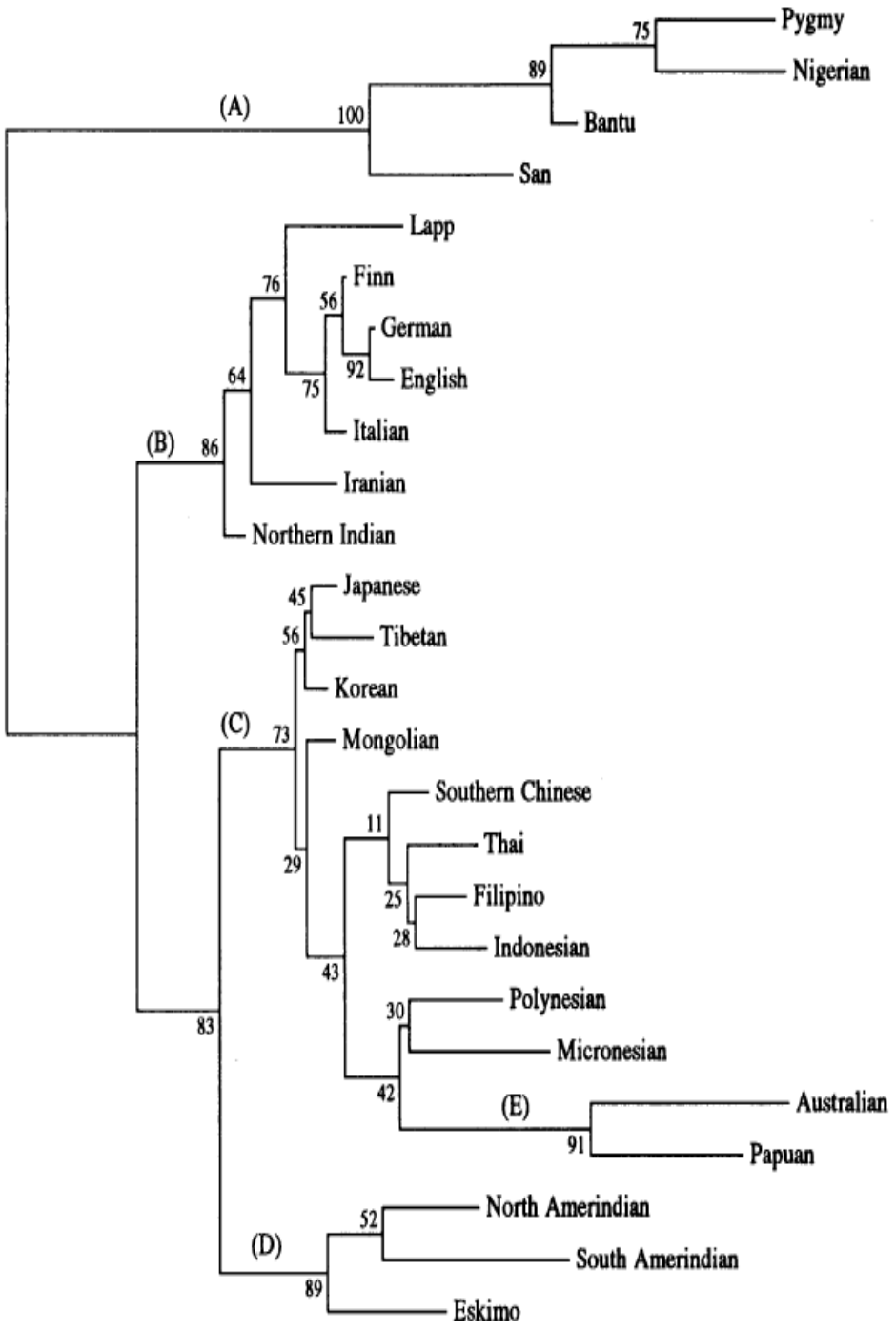


396 copy number variants (CNVs)

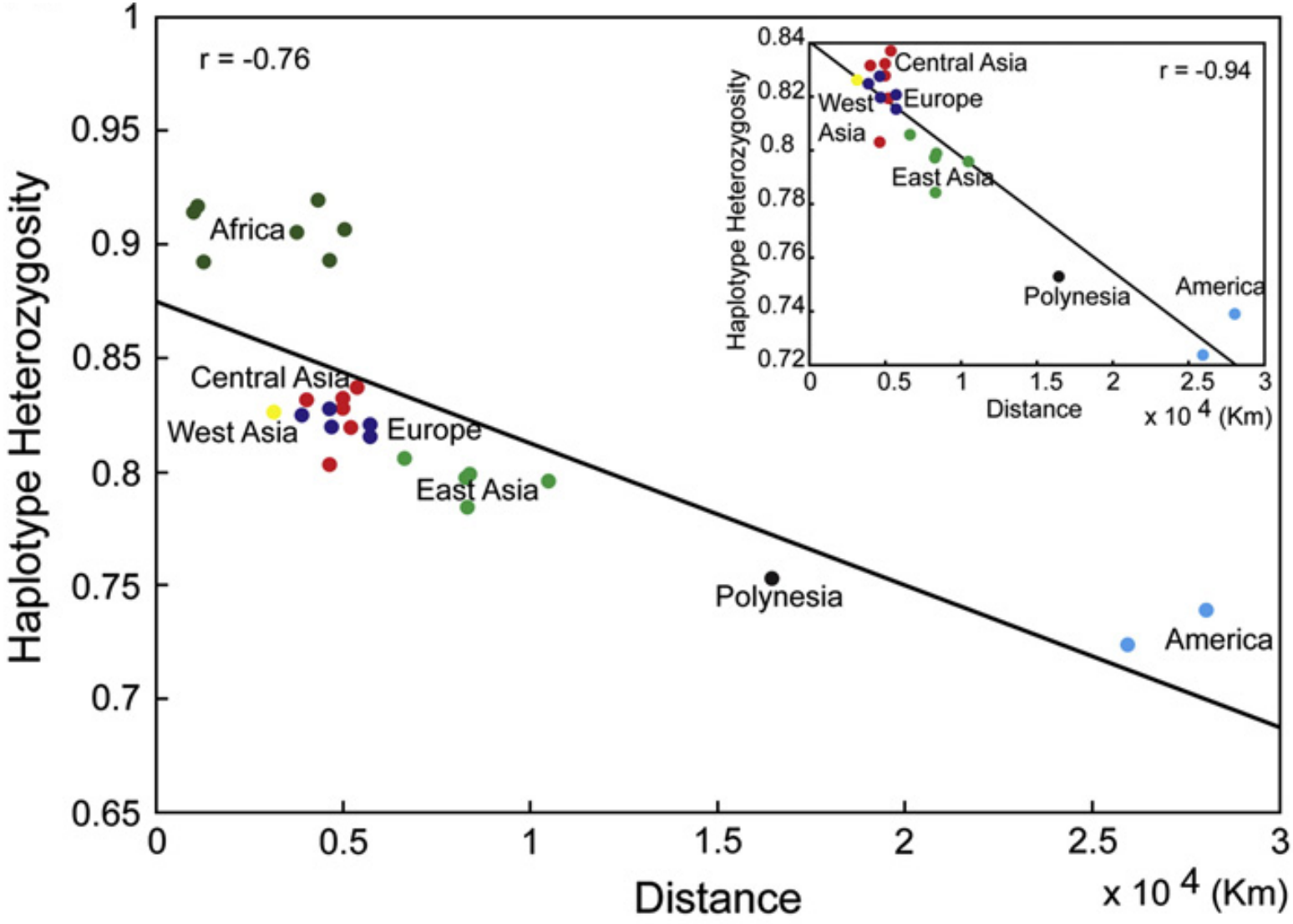
Jakobsson et al., 2008, *Nature* 451: 998-1003

Human population relationships, based on 29 blood group and protein polymorphisms

(Nei and Roychoudhury, 1993, *Mol. Biol. Evol.*)

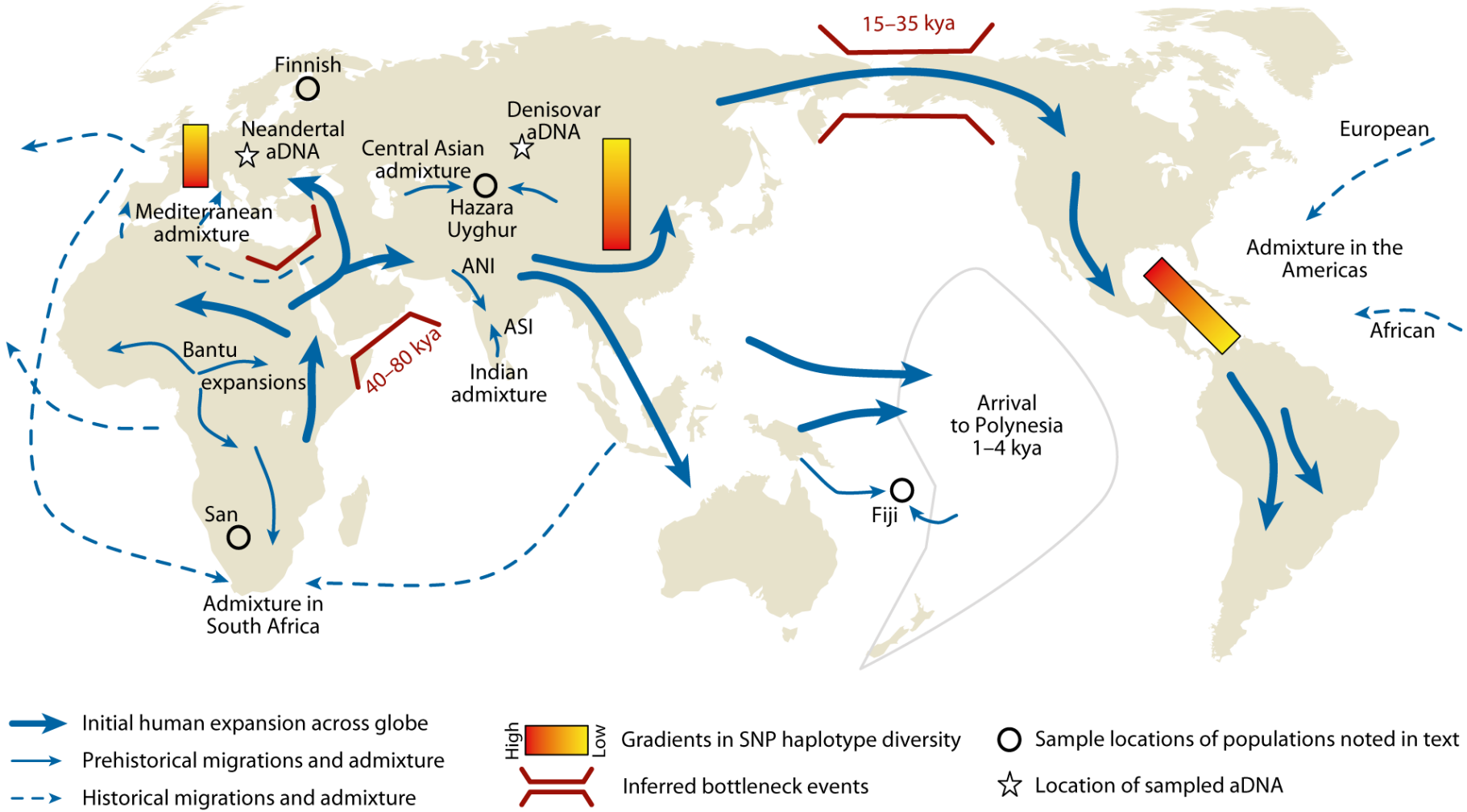


Haplotype diversity declines with distance from Africa



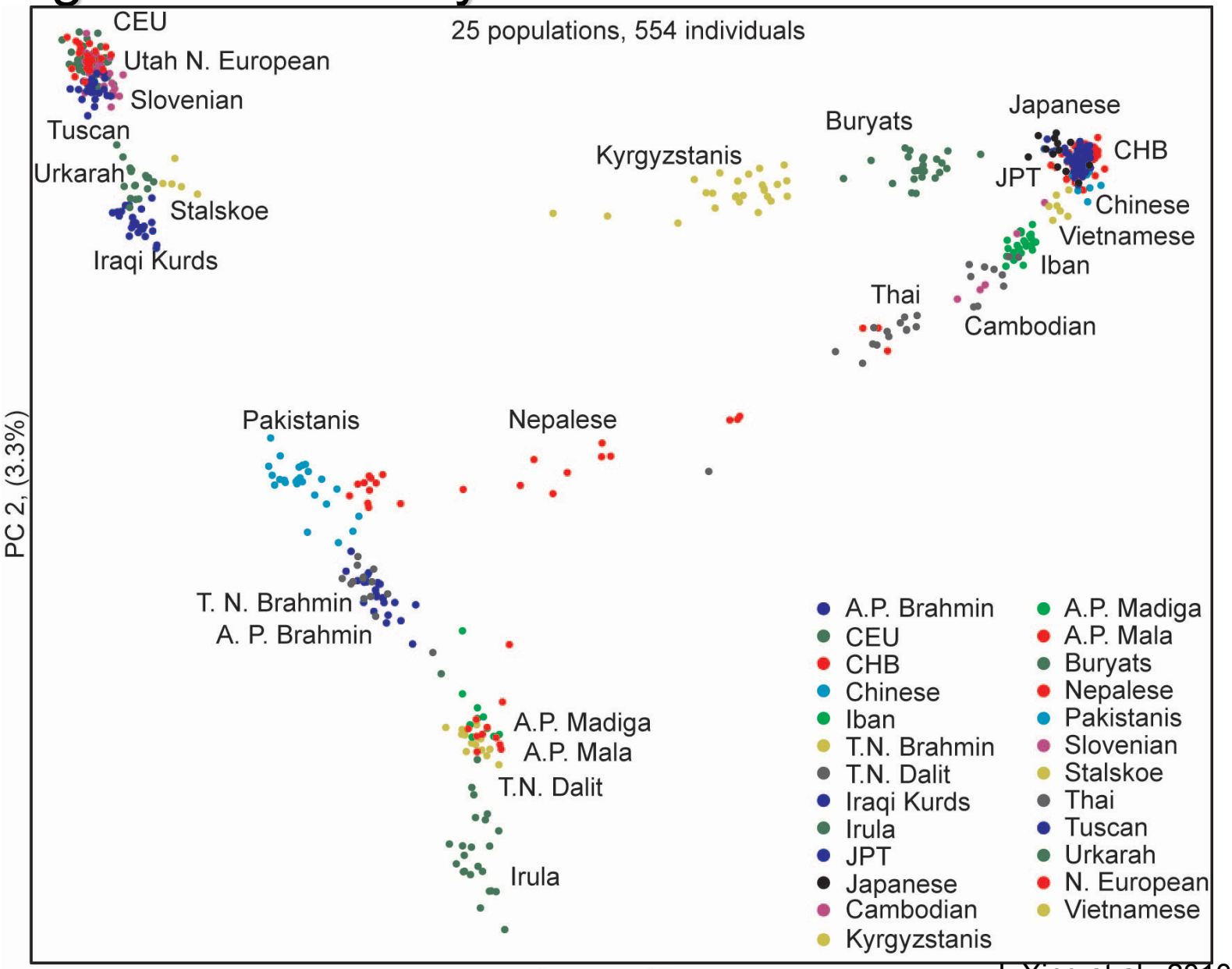
J. Xing et al., 2010, *Genomics*

Recent African origin of anatomically modern humans



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

Principal components analysis displays **individual** genetic similarity in 2D: each dot = 1 individual

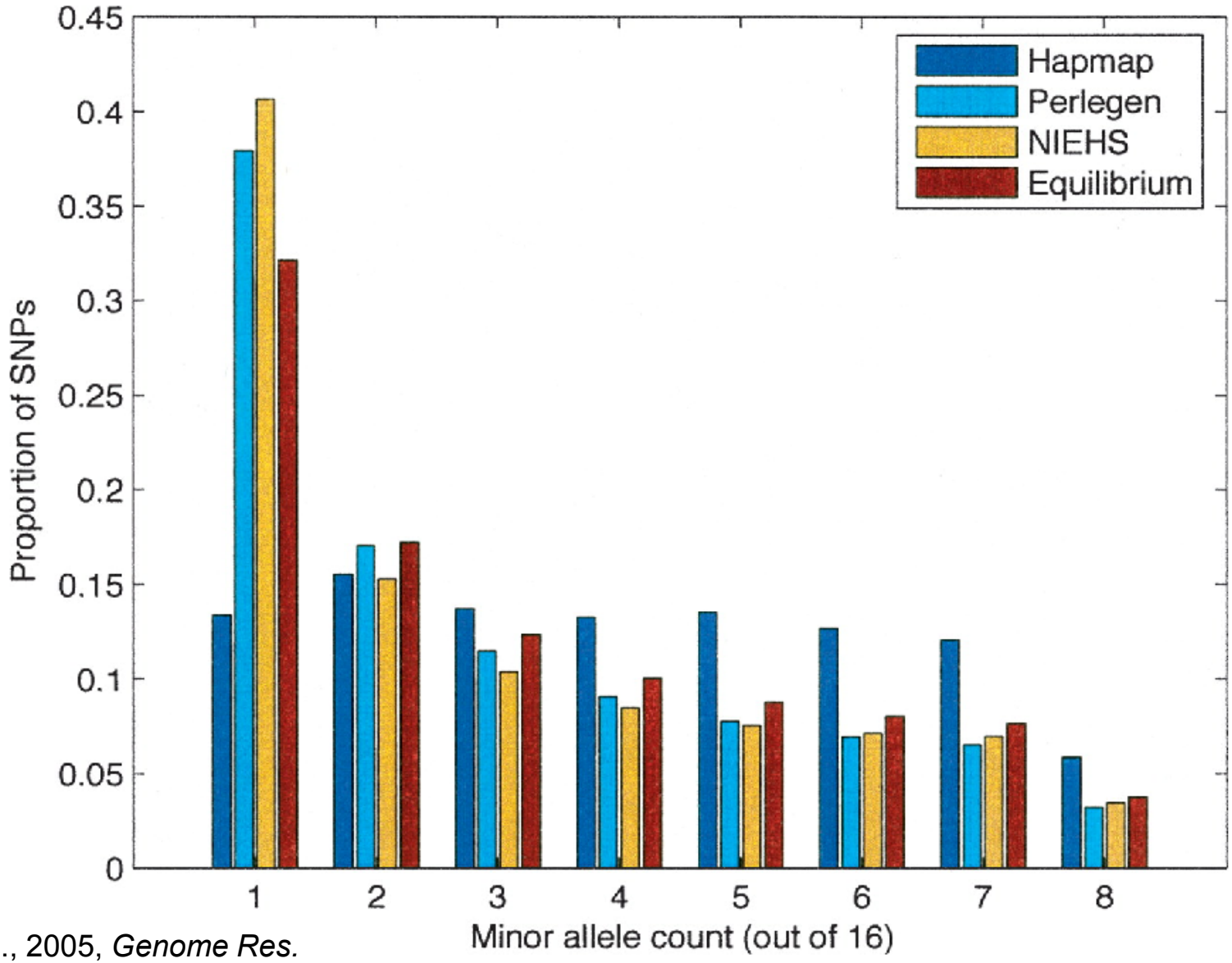


J. Xing et al., 2010, *Genomics*

Microarray-based SNPs portray population relationships accurately but are biased

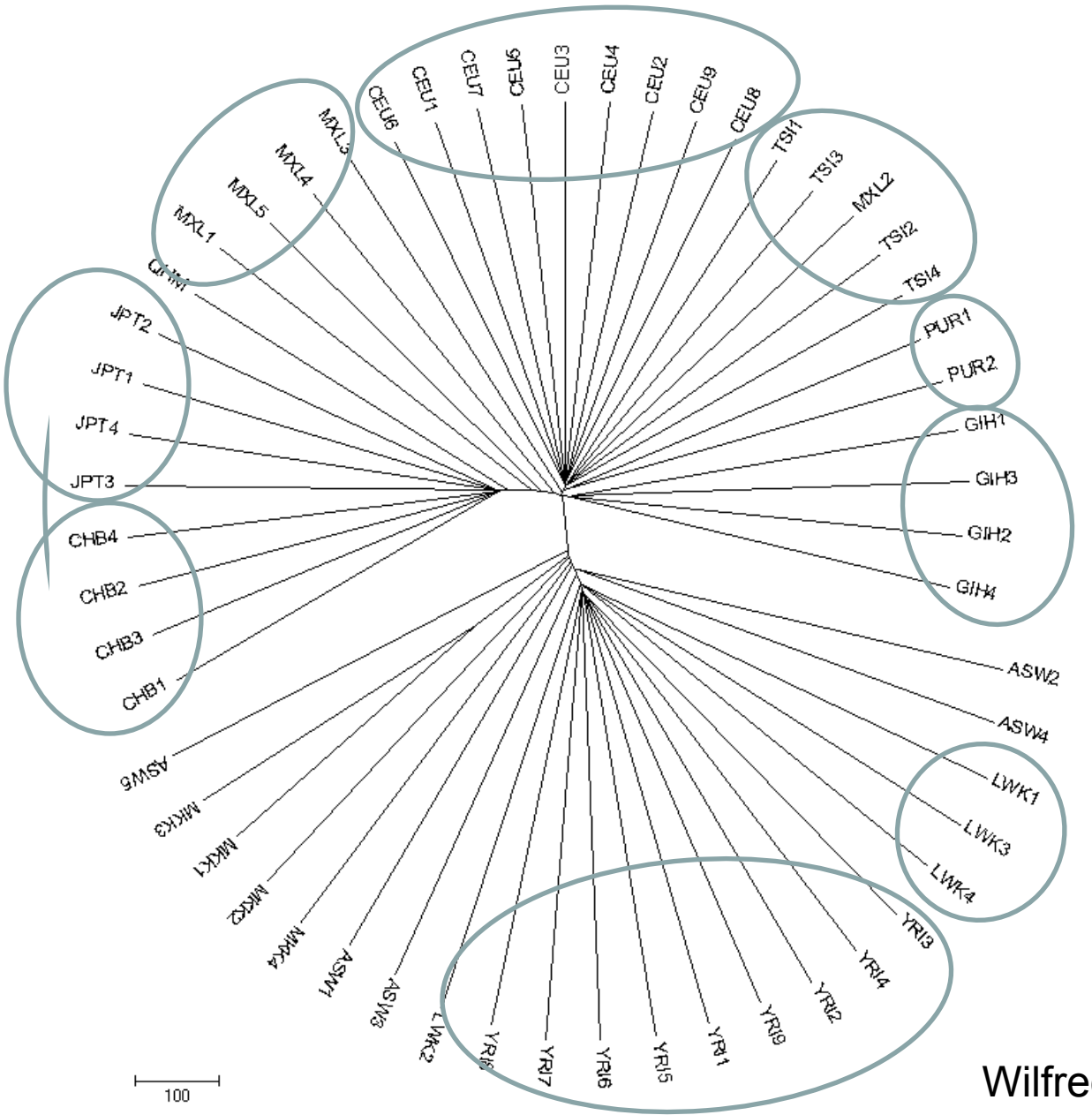
- Microarray SNPs are selected for higher frequency and diversity in Europeans
- Complete DNA sequences are unbiased and include information about rare variants

The effect of ascertainment bias on allele frequencies



Clark et al., 2005, *Genome Res.* 15: 1496-1502

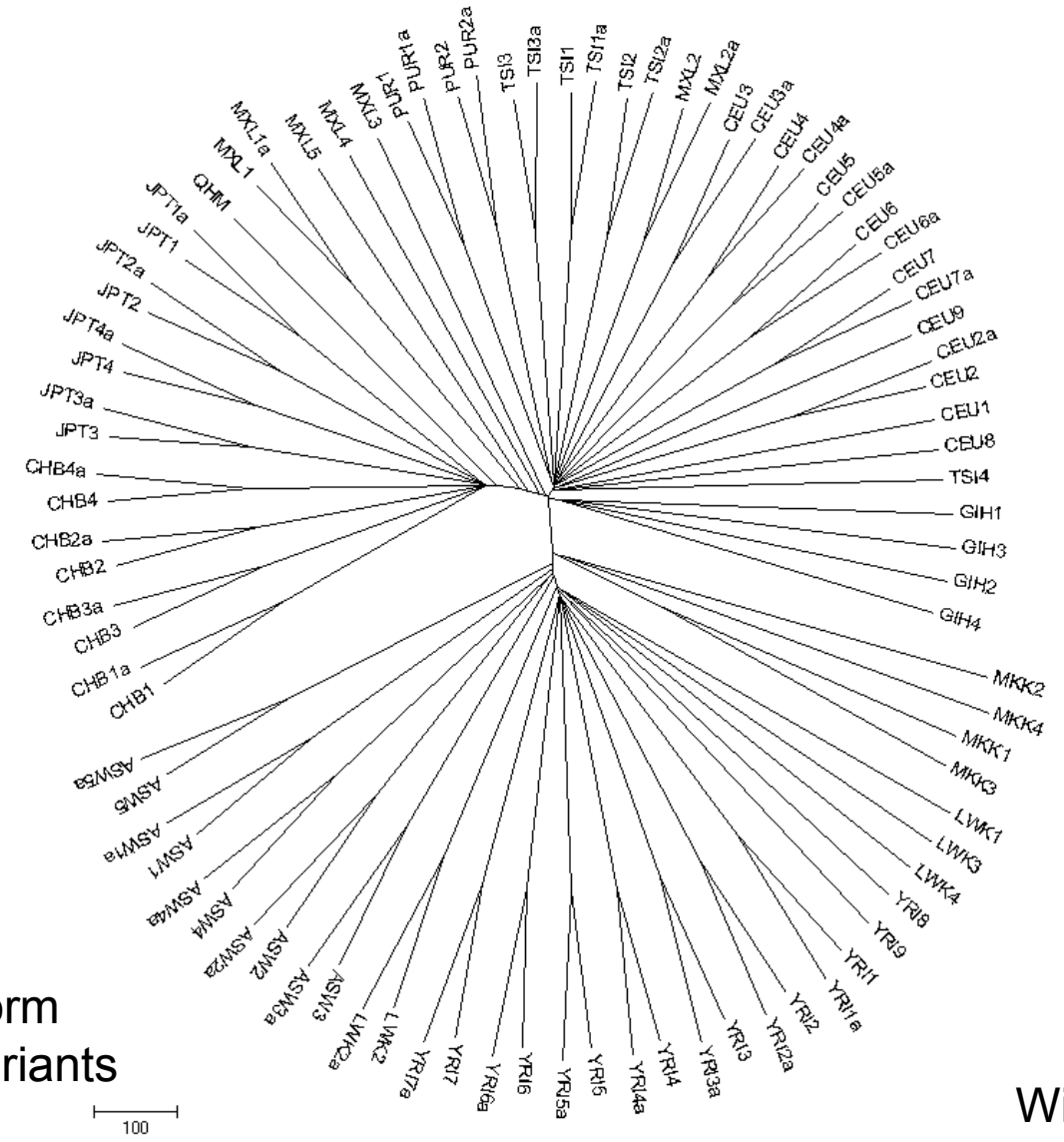
Individual network based on CGI WGS data: 54 individuals



CHB: Chinese
JPB: Japanese
MXL: Mexican
CEU: Utah CEPH
TSI: Tuscan
PUR: Puerto Rican
GIH: Gujarati
ASW: African-American
LWK: Luhya
YRI: Yoruban
MKK: Maasai

Wilfred Wu, MD

Complete Genomics vs. 34 1000 Genomes sequences (Phase 1)

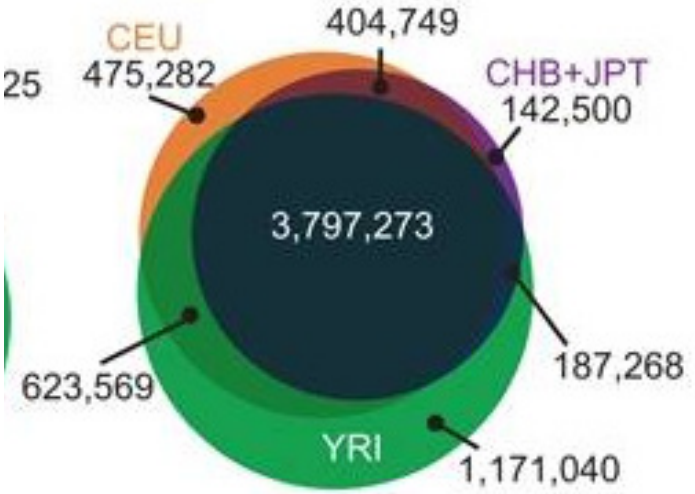


Average between-platform difference = 348,000 variants

Wilfred Wu, MD

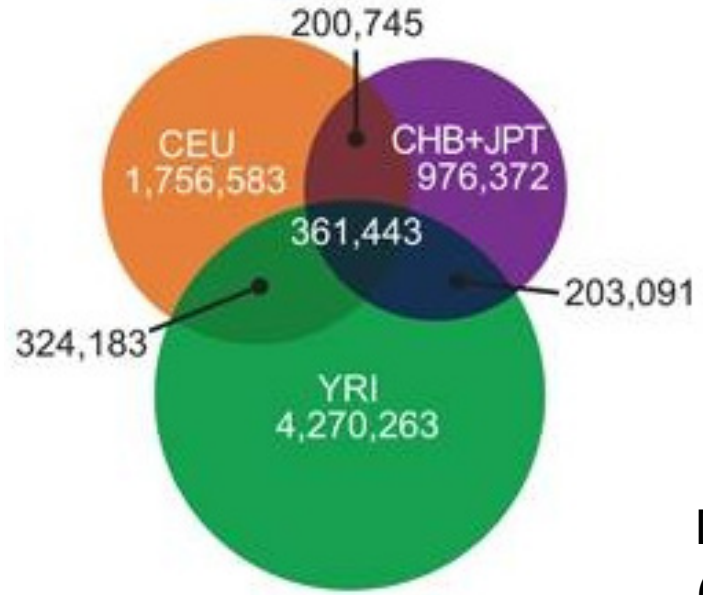
Rare SNPs are much more likely to be population-specific

Common SNPs previously identified in dbSNP (build 129)



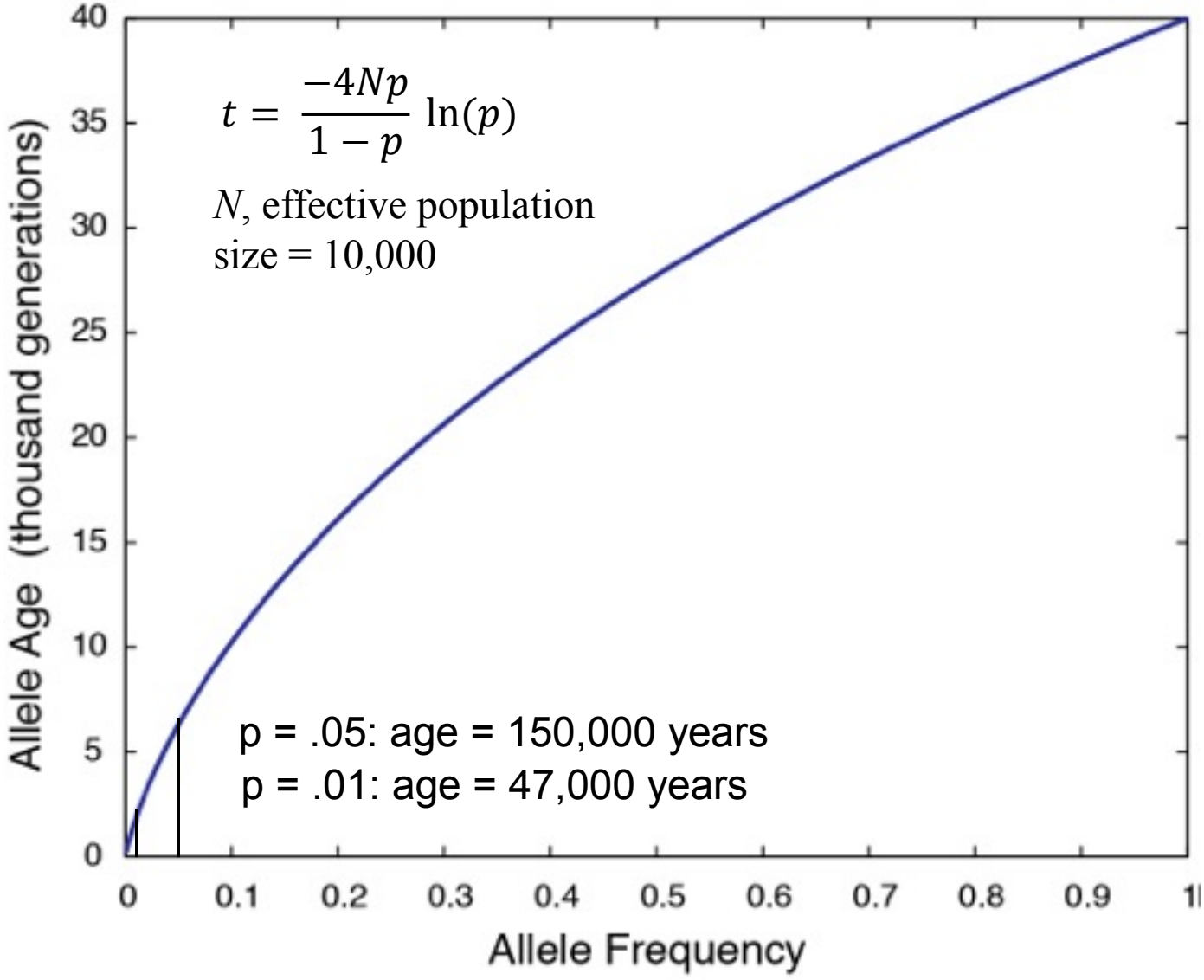
Average allele frequency difference between populations: 15%

New rarer SNPs identified by sequencing

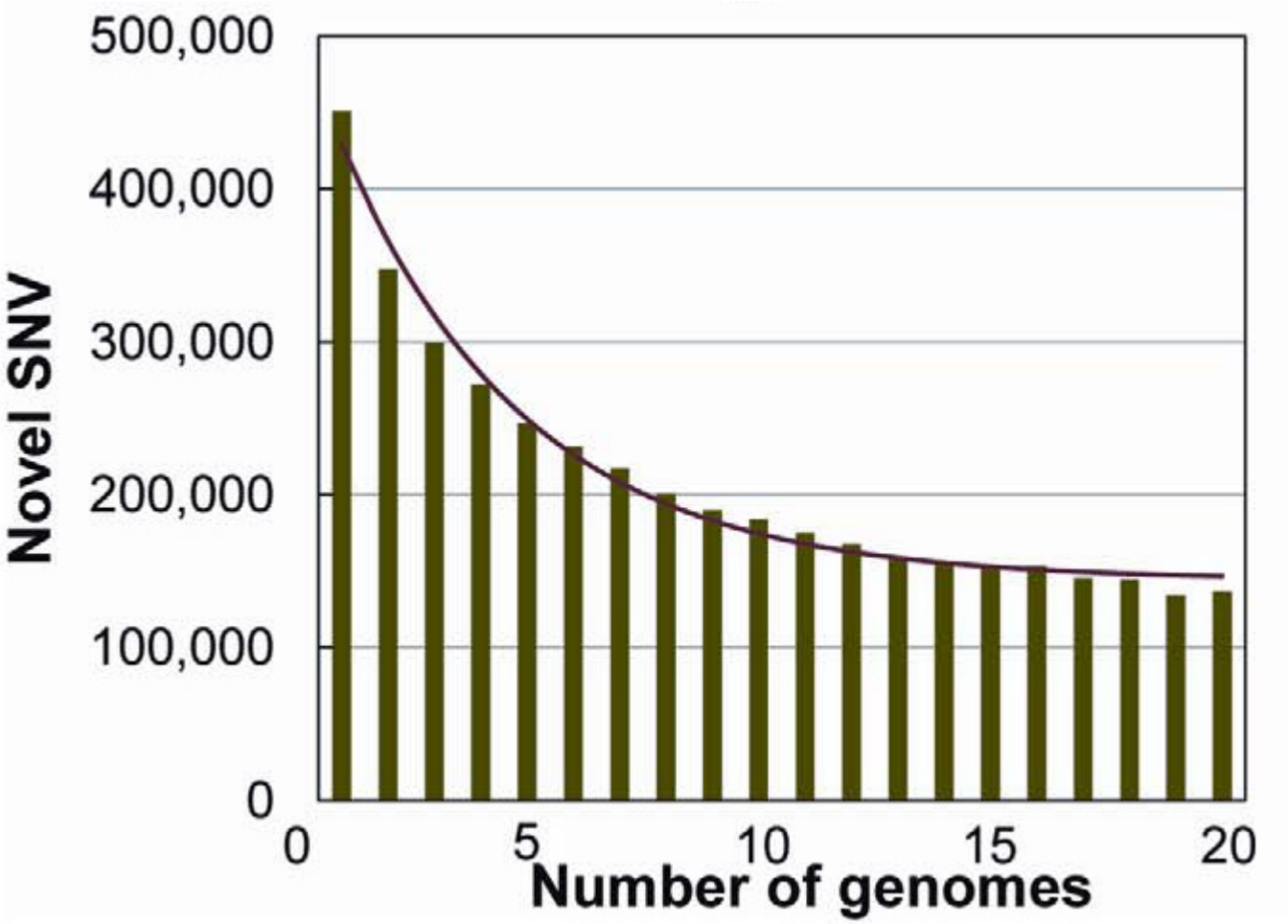


Durbin *et al.,* 2010, *Nature* *(*1000 Genomes Project)

Allele age, t, as a function of frequency



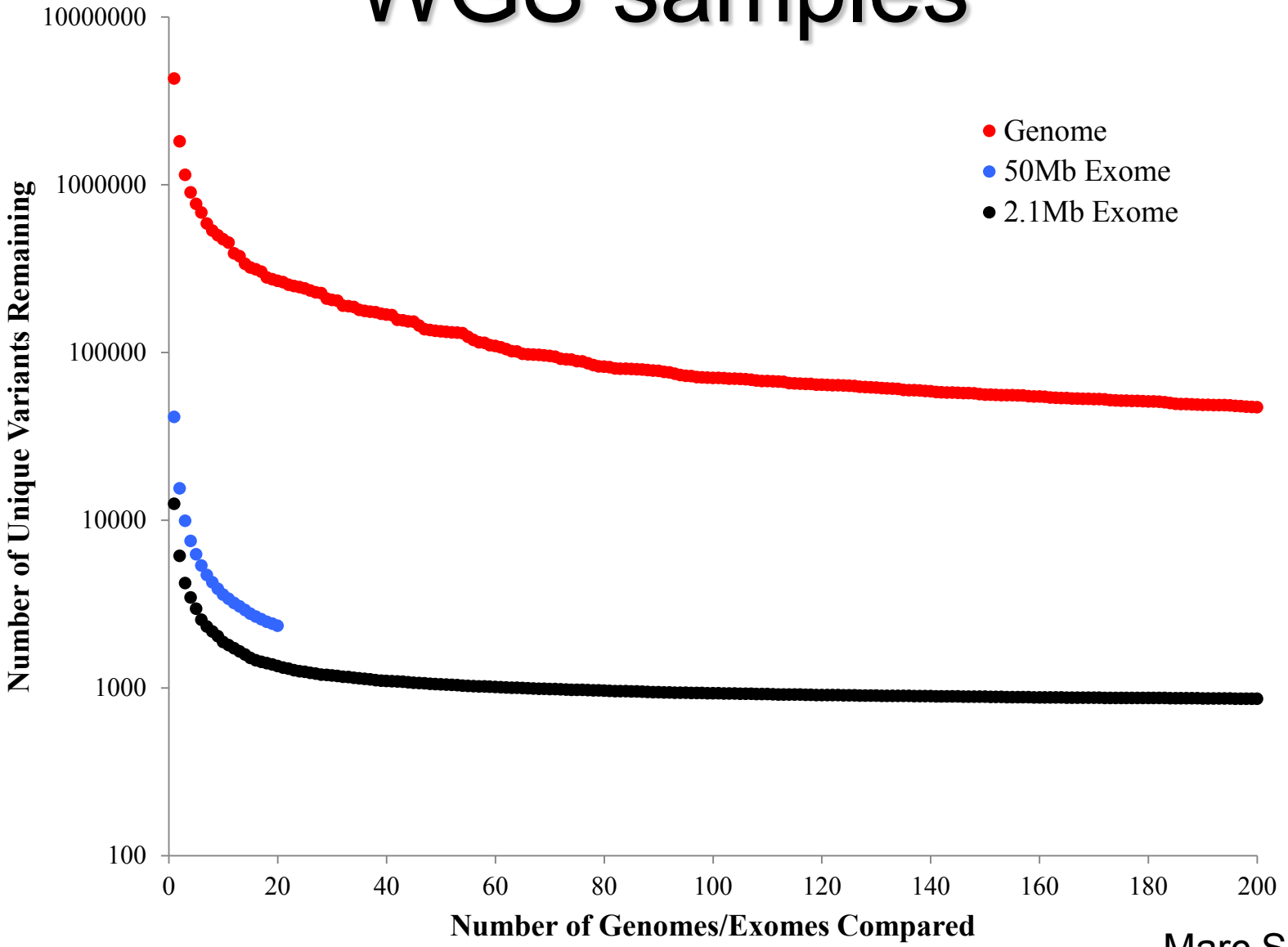
We expect many novel variants with each whole-genome sequence



20 whole-genome sequences

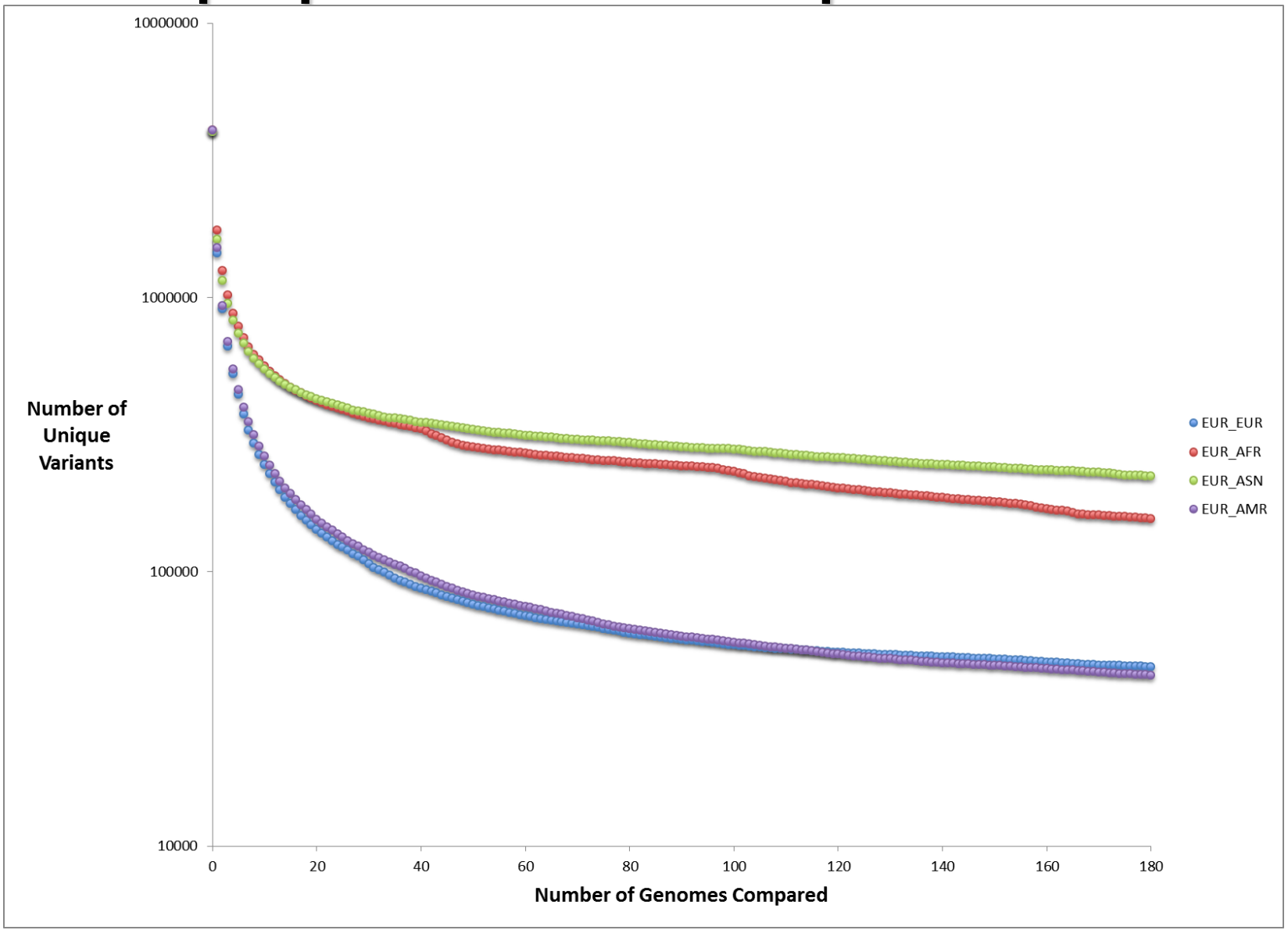
Pelak et al., 2010, *PLoS Genet.*

Number of novel variants in 200 WGS samples

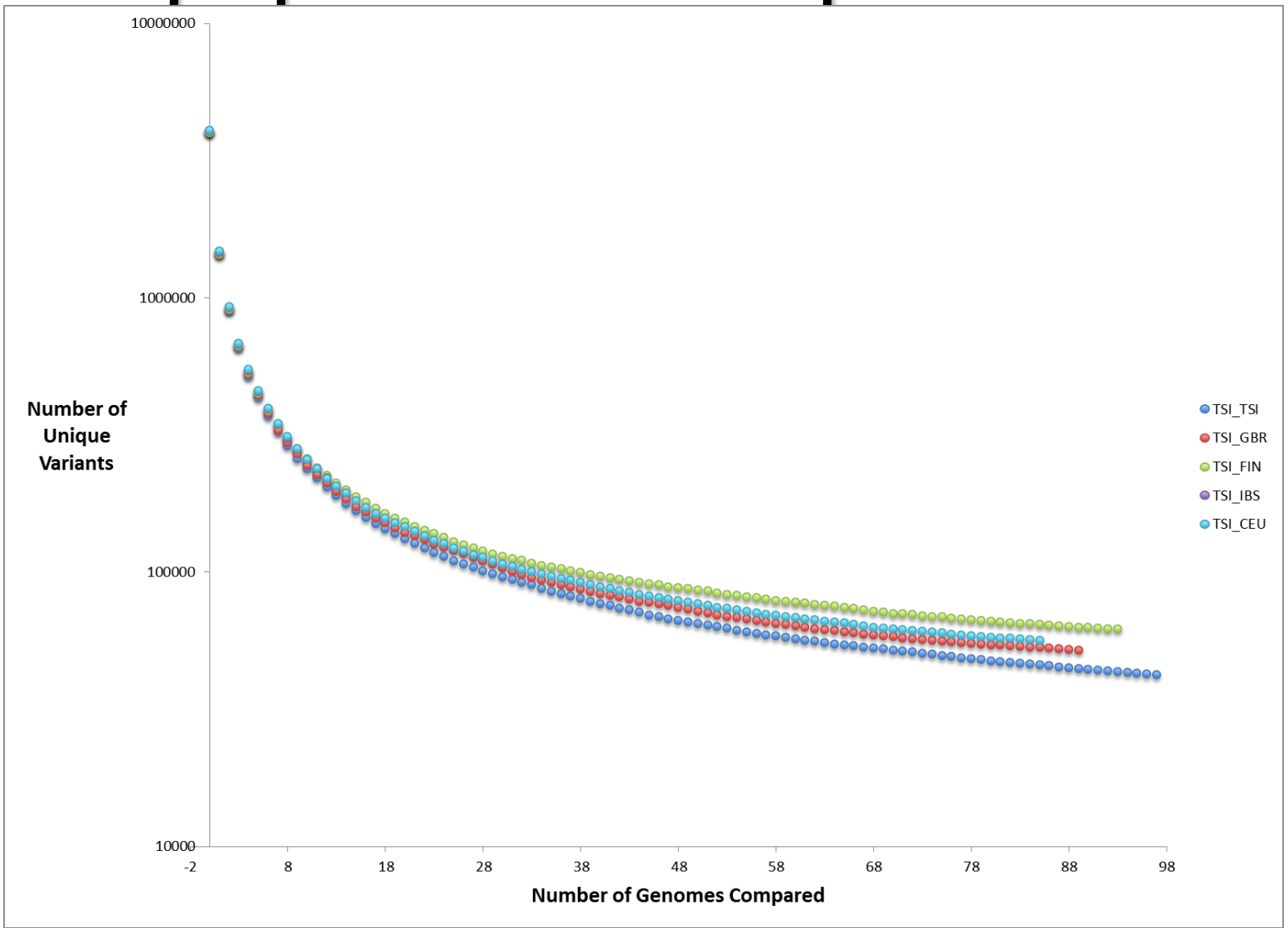


Marc Singleton

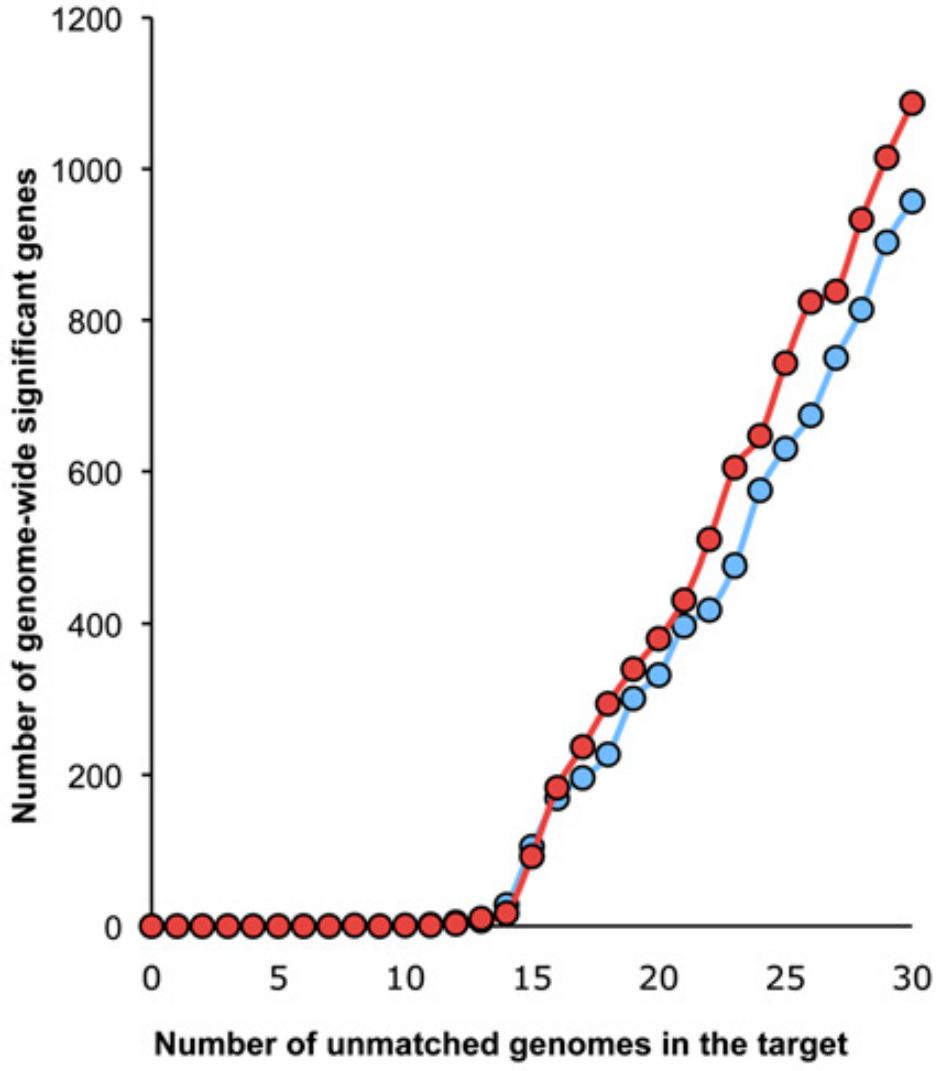
Novel variants in between-population comparisons



Novel variants in between-population comparisons



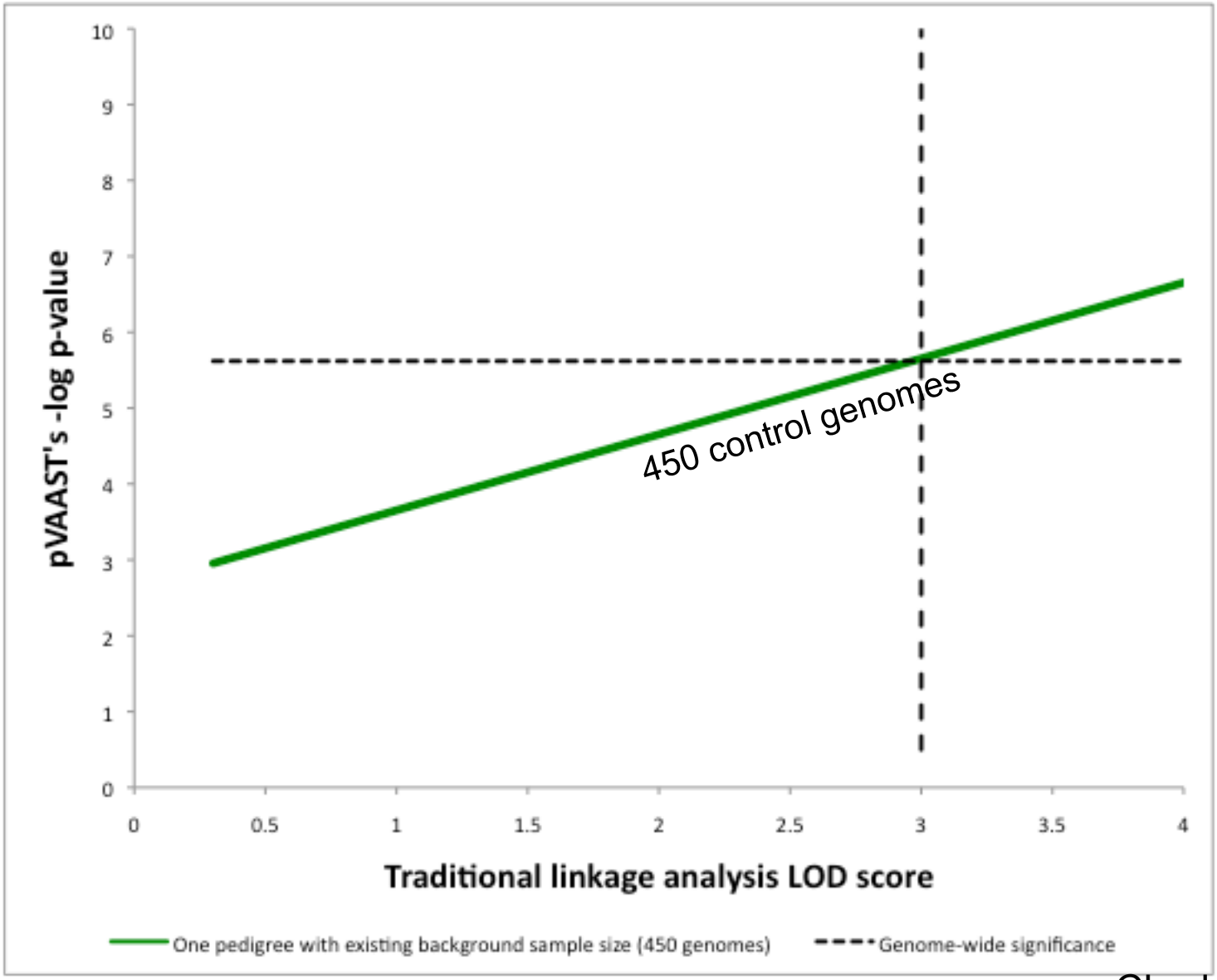
False-positive results increase dramatically with inaccurate case-control matching: WGS data



Comparison of 30 European disease cases with mixtures of European and African controls: VAAST analysis of WGS data

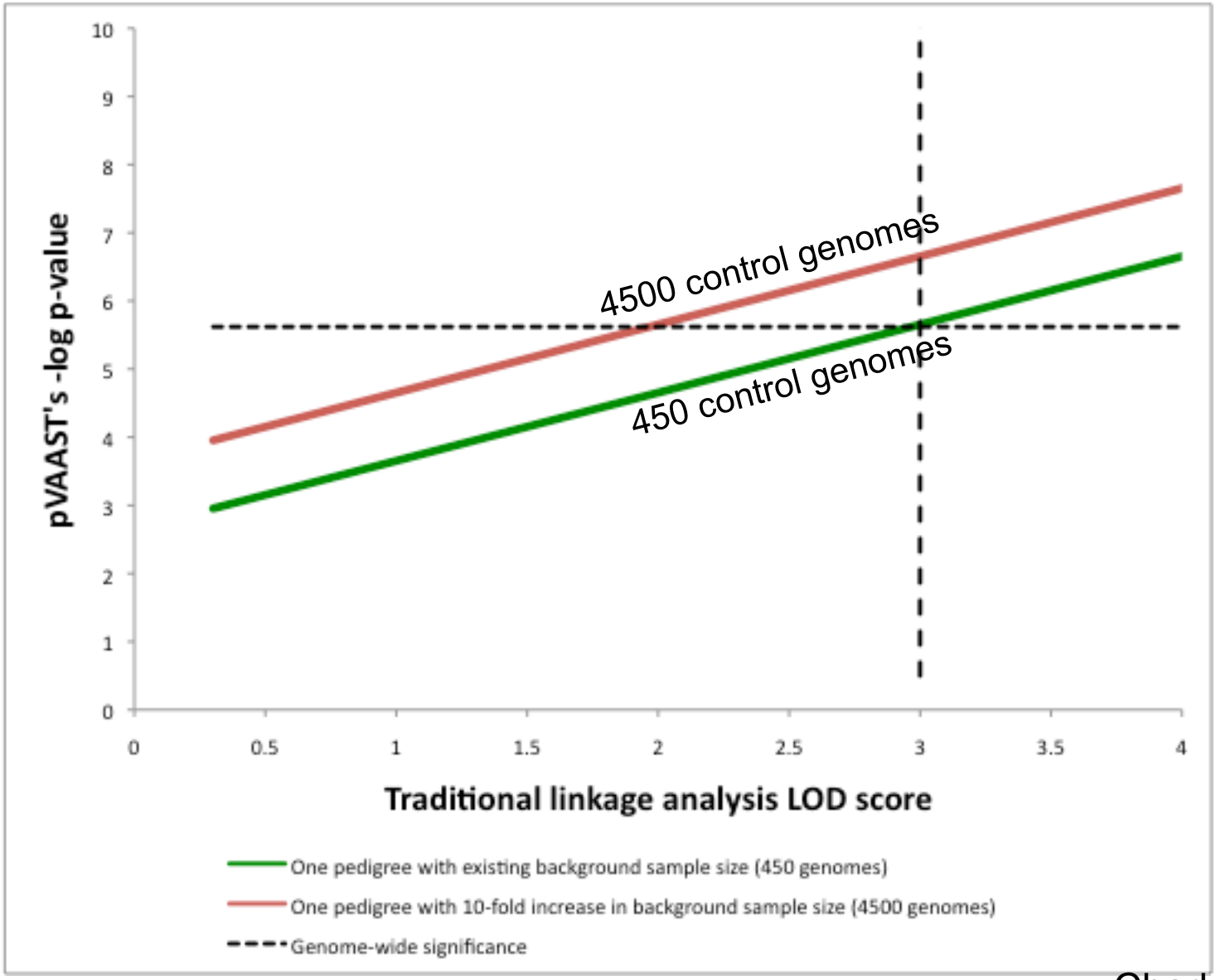
Yandell, et al., 2011, *Genome Res.*

LOD score needed for genome-wide significance for detection of *GATA4* as a disease-causing mutation



Chad Huff, PhD

LOD score needed for genome-wide significance for detection of *GATA4* as a disease-causing mutation



Chad Huff, PhD

Conclusions

- Because of population specificity of rare variants, more sequencing in more populations is needed

- We need to do experiments to determine how closely one needs to match control genomes to minimize false positive findings

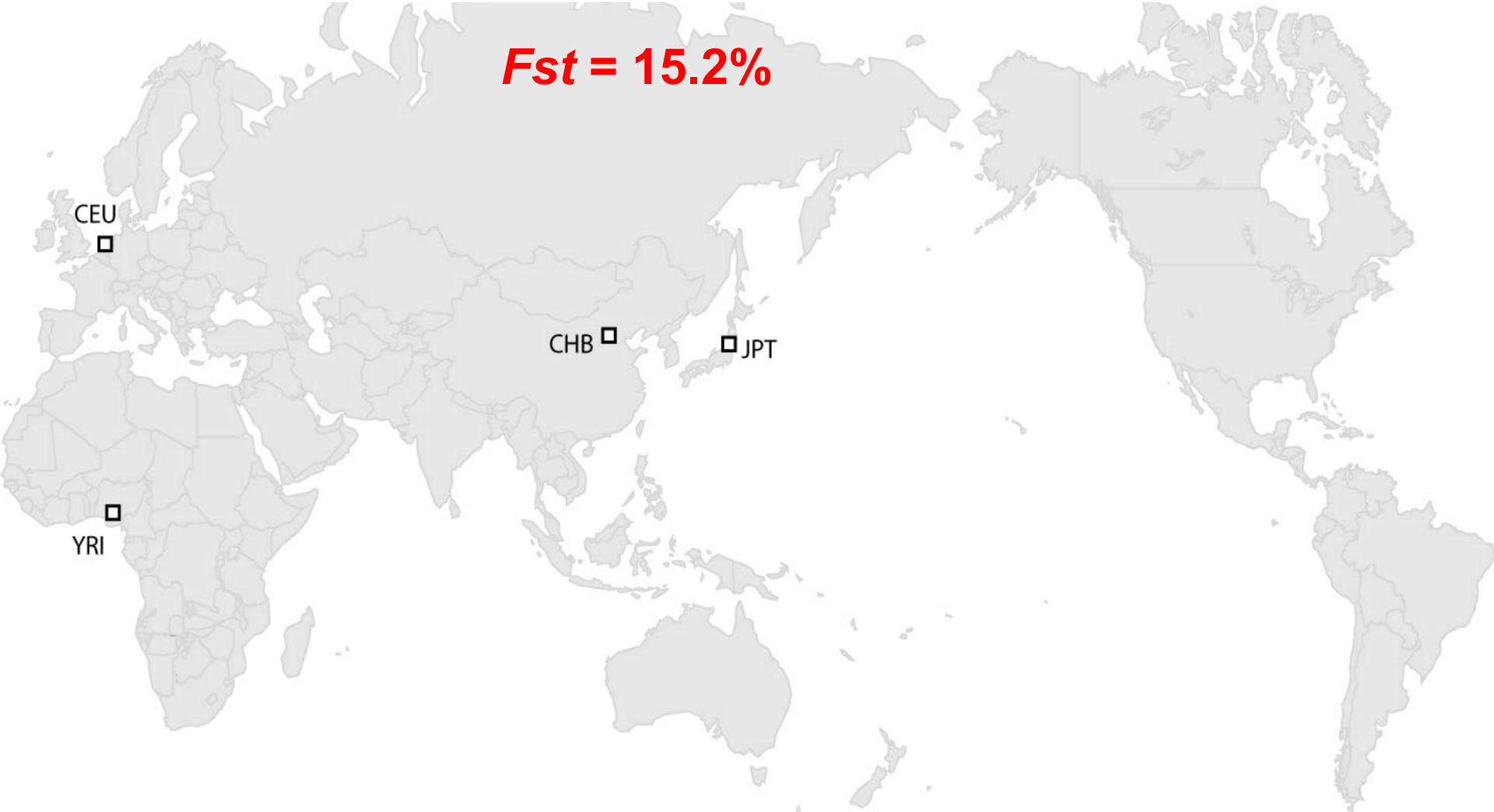
Acknowledgments

University of Utah: Jinchuan Xing, Dave Witherspoon, Chad Huff, Tatum Simonson, Steve Guthery, Scott Watkins, Yuhua Zhang, Bob Weiss, Alan Rogers

LSU: Mark Batzer

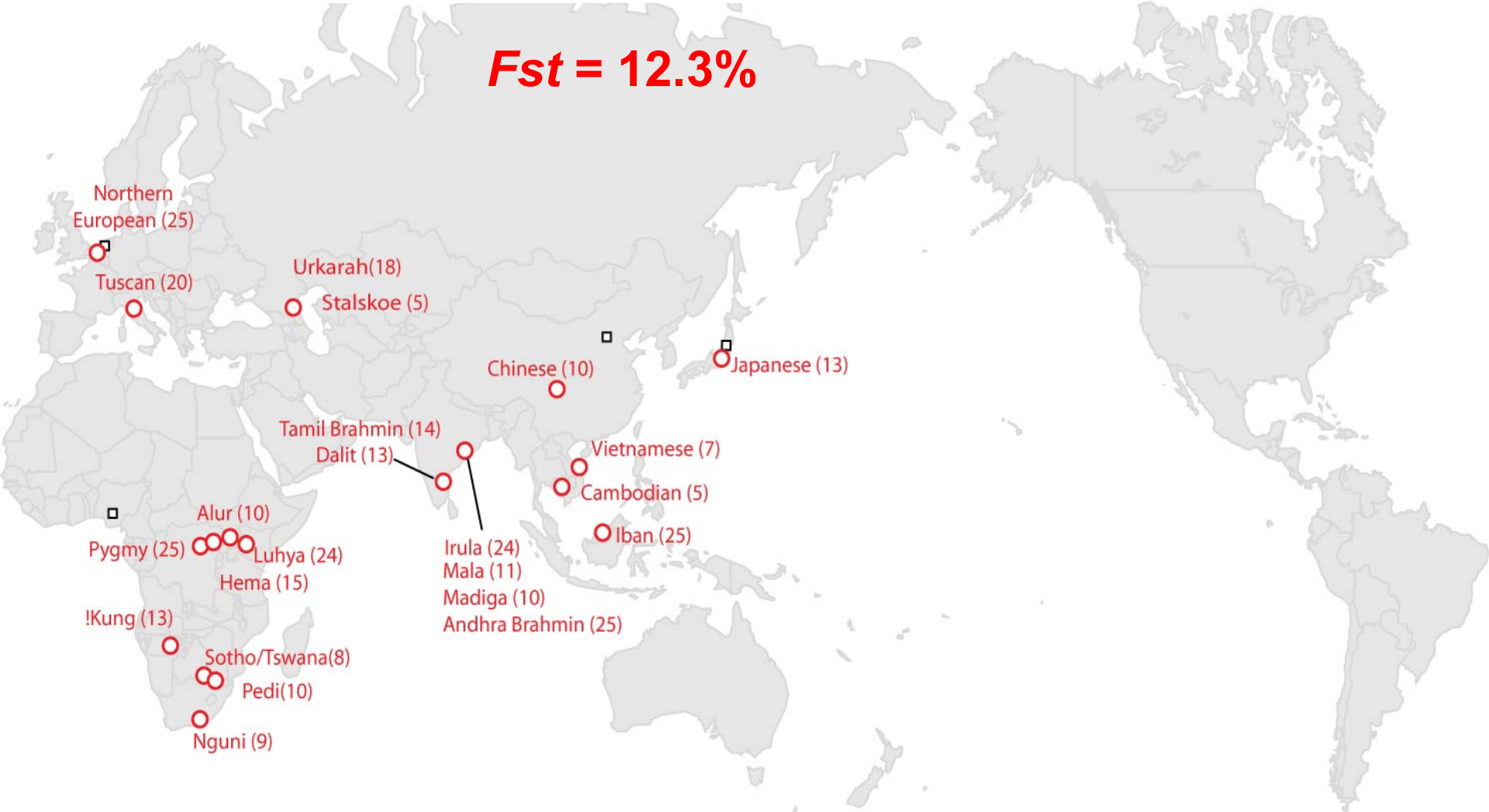
Limited sampling produces high F_{ST} values

HapMap II, 210 individuals, 4 populations



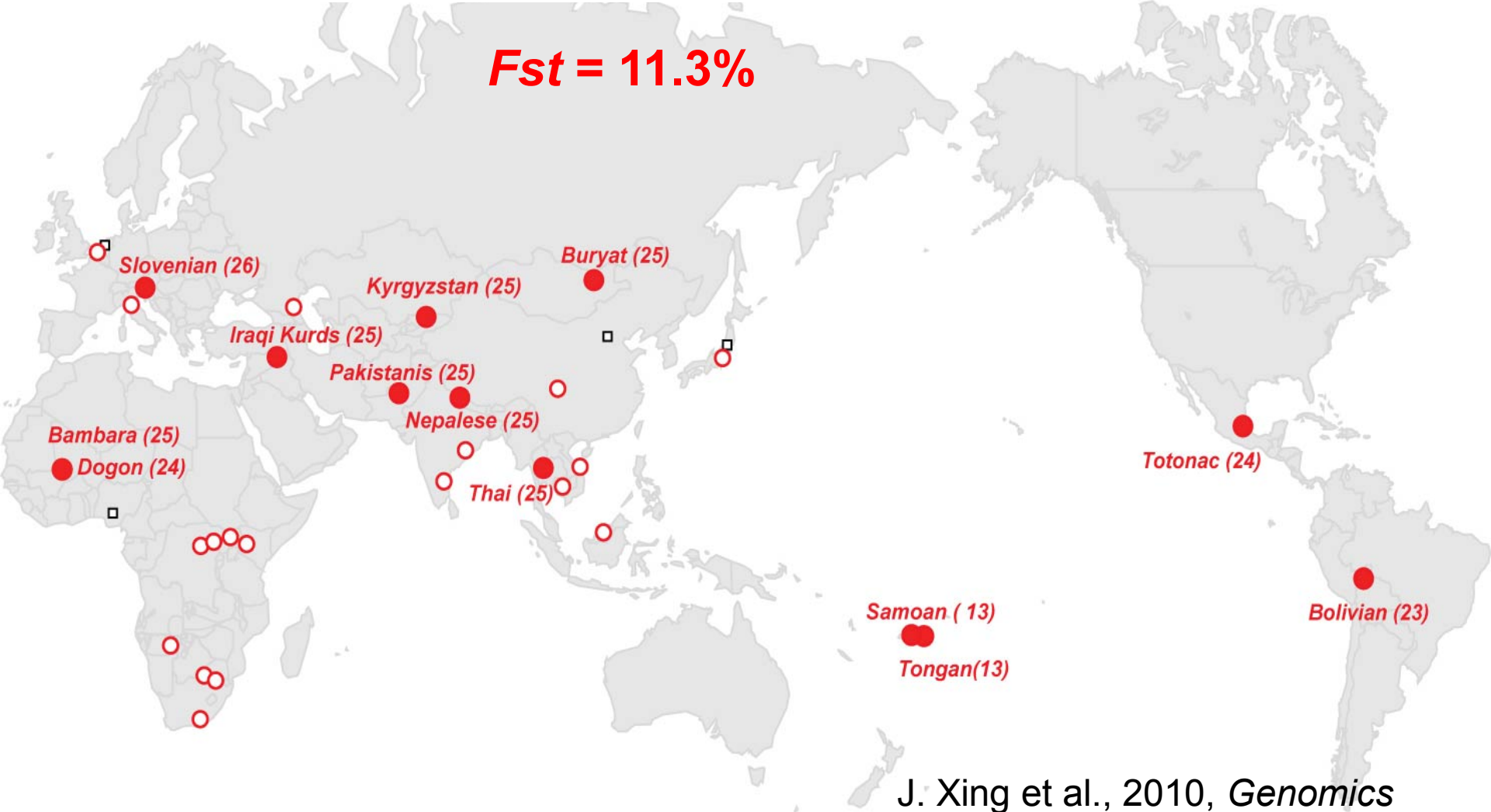
Reduced genetic differentiation (F_{ST}) with more even sampling.

554 individuals, 27 populations

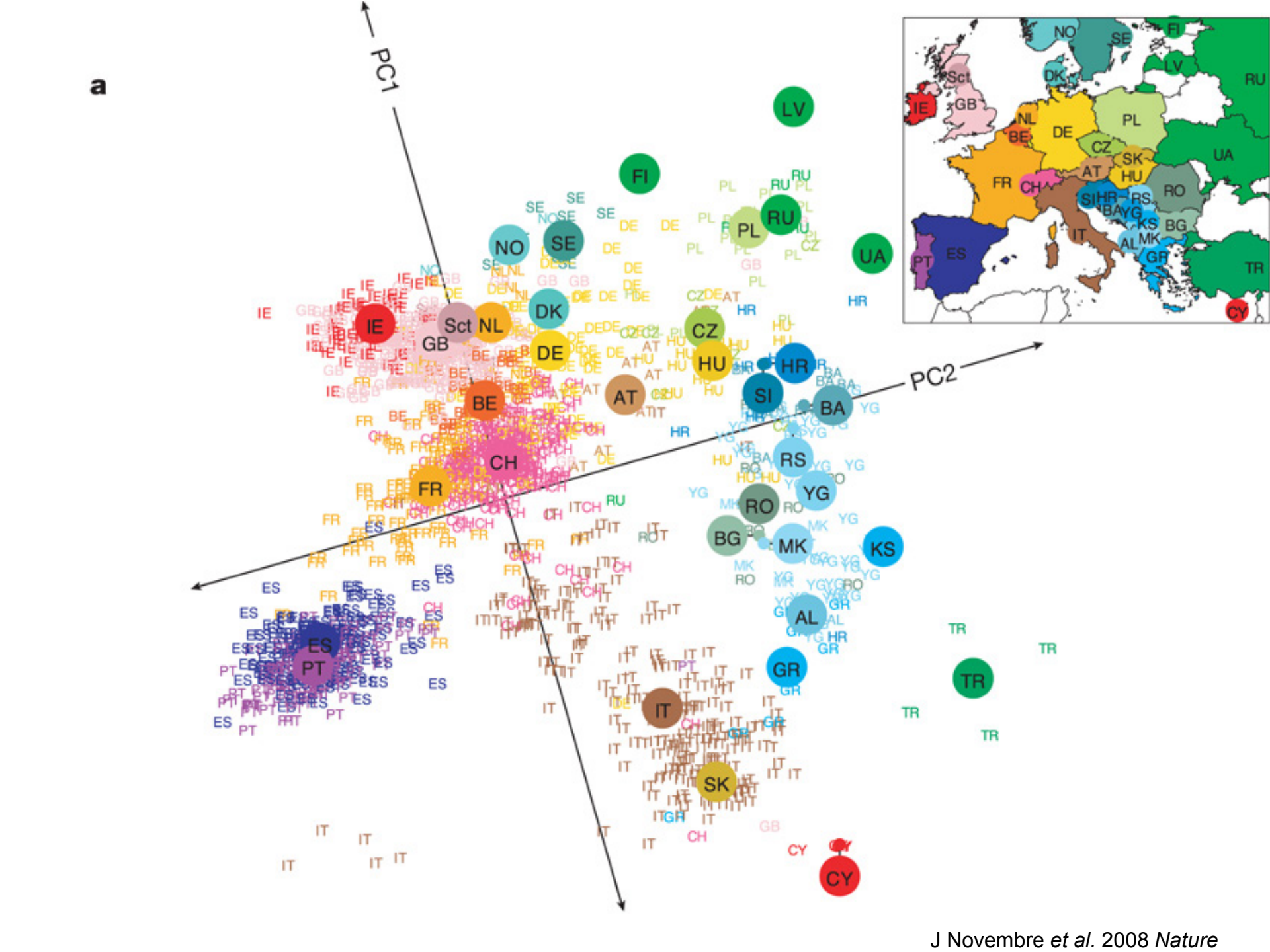


Reduced genetic differentiation (F_{ST}) with more even sampling.

850 individuals, 40 populations

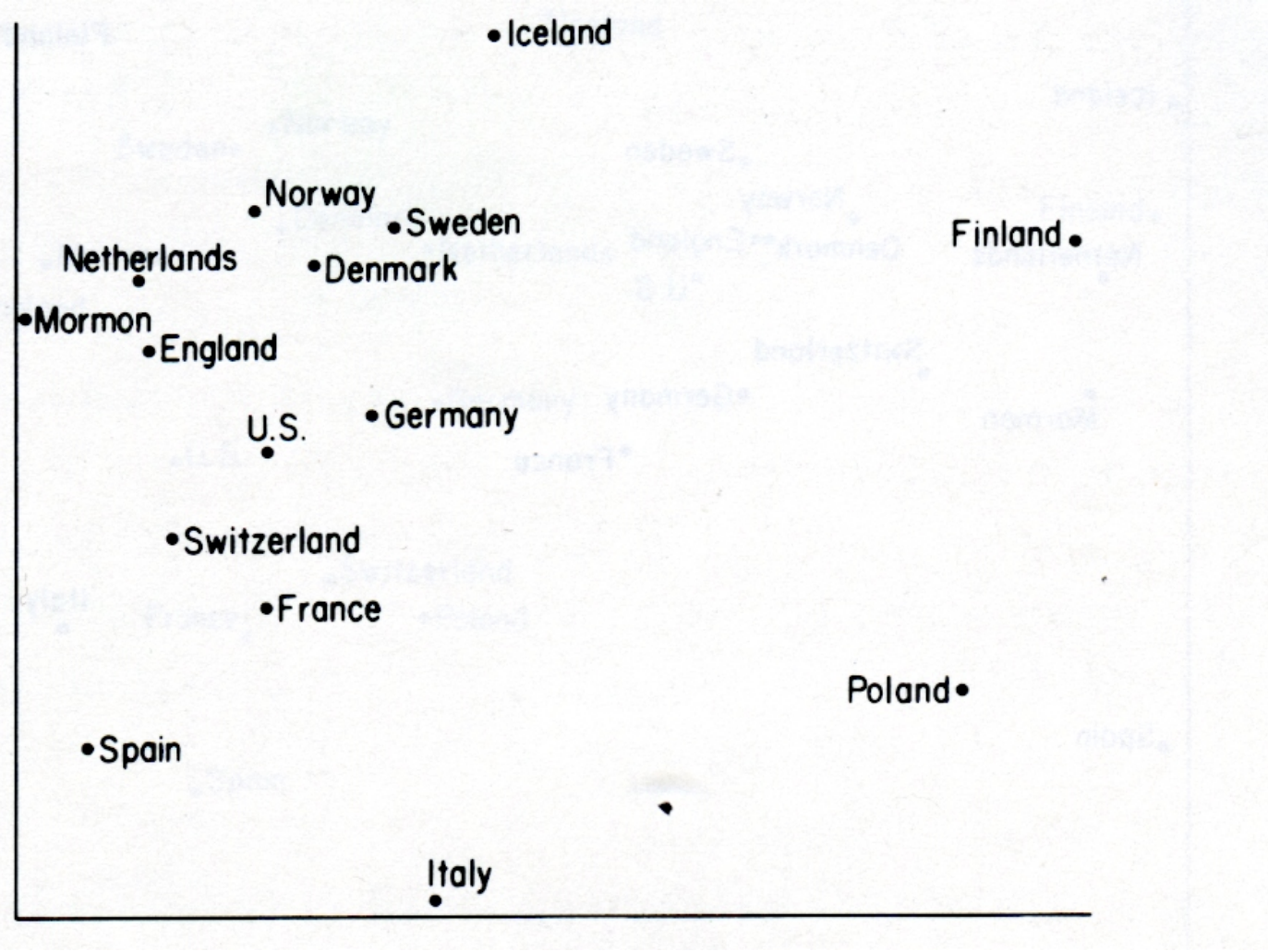


J. Xing et al., 2010, *Genomics*



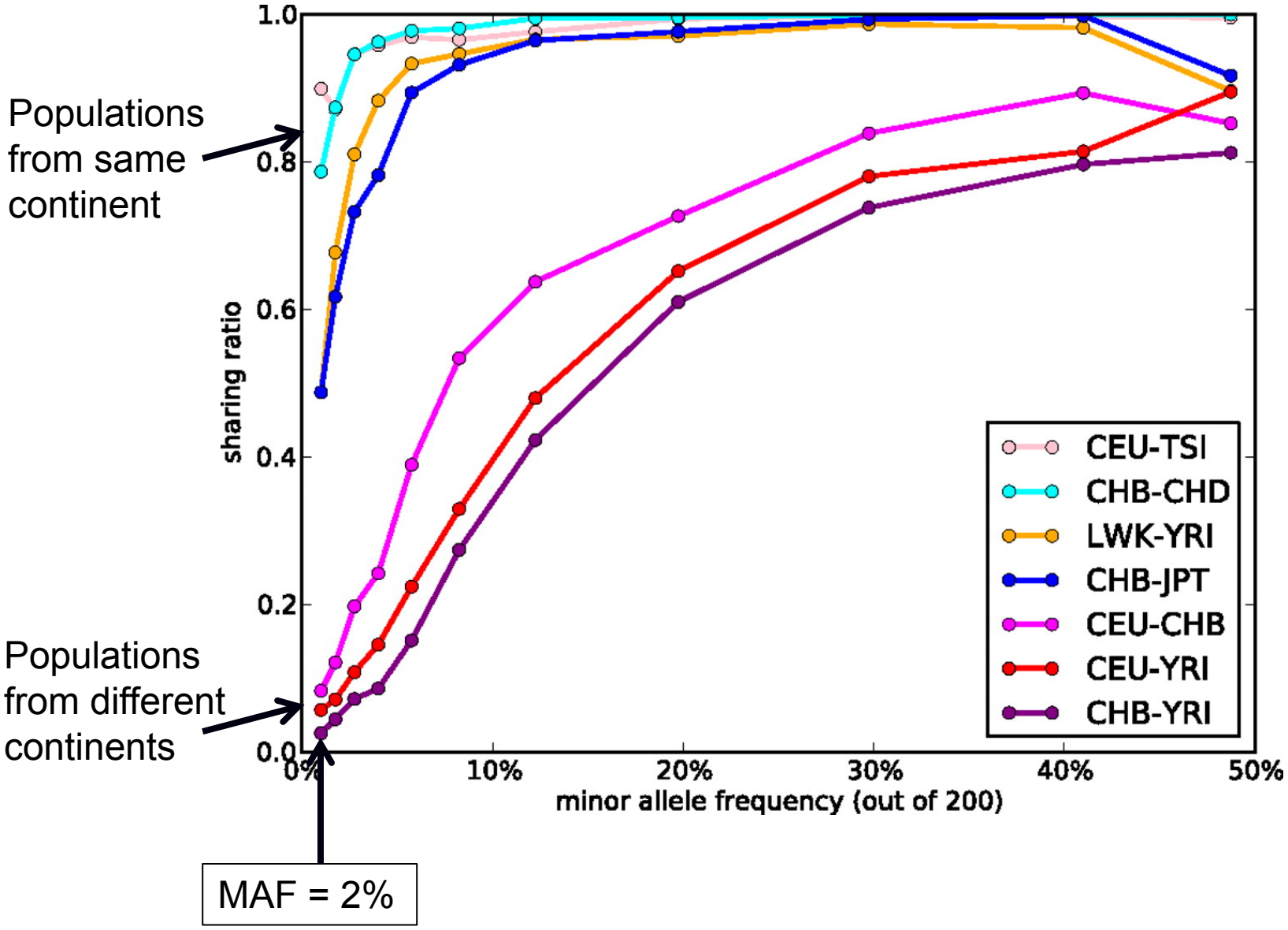
J Novembre *et al.* 2008 *Nature*

Genetic distance analysis: 15 loci



McLellan, Jorde, and Skolnick, 1984, *Am. J. Hum. Genet.*

Proportion of shared alleles between pairs of individuals, relative to a single panmictic population



Gravel S et al. PNAS 2011;108:11983-11988

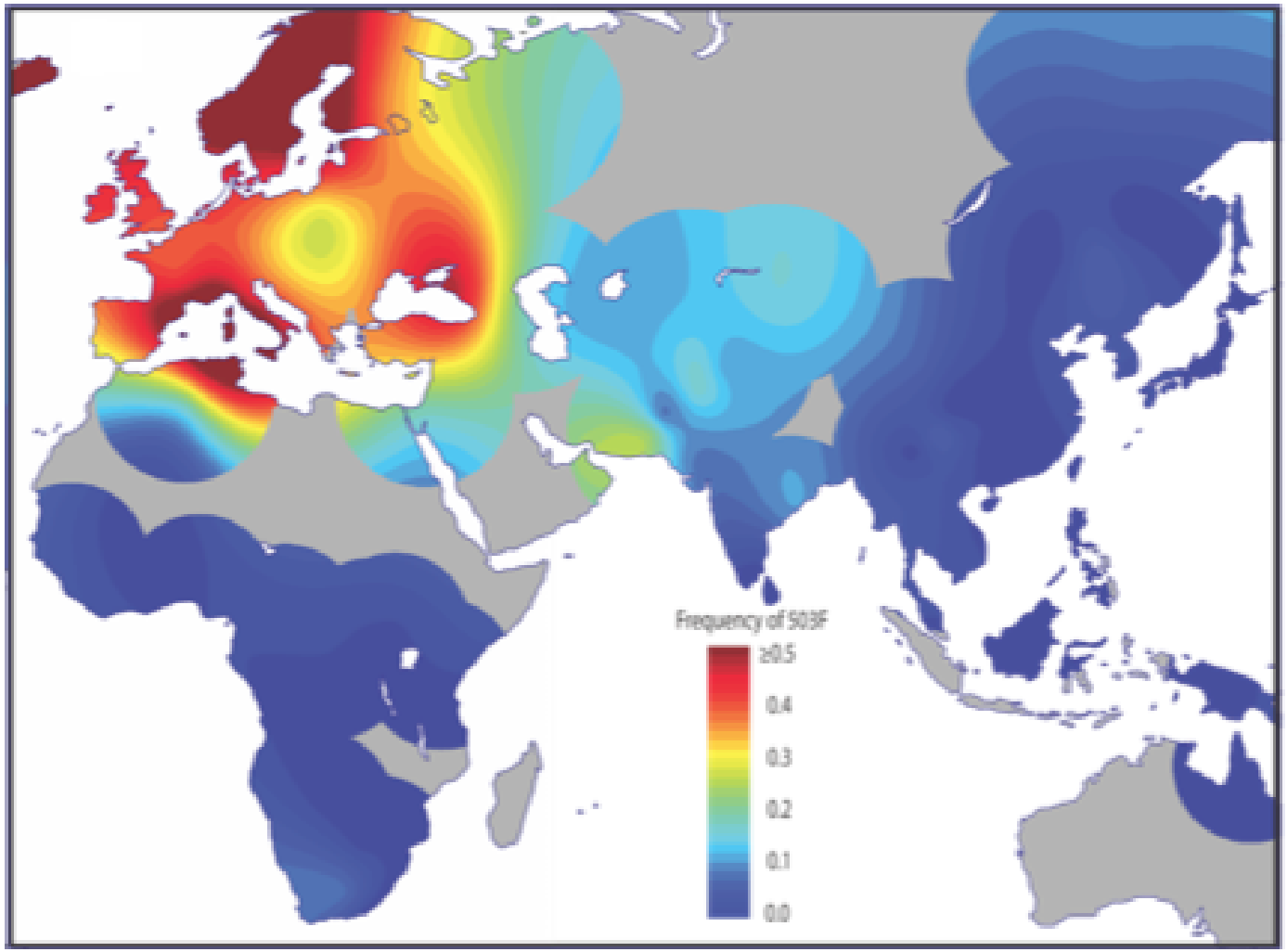
Examples of genes in which elevated LD indicates recent natural selection

Gene	Phenotype
G6PD	Malaria protection
HFE (hemochromatosis)	Iron absorption
CYP3A5	Sodium retention
*LCT (*lactase enhancer)	Lactase persistence
SLC24A5	Skin pigmentation
Alcohol dehydrogenase	Ethanol metabolism
EPAS1, EGLN1	Hypoxia response

Voight et al., 2006, *PLOS Biology* 4: 446-458

Simonson et al., 2010, *Science*

503F Variant of *OCTN1*



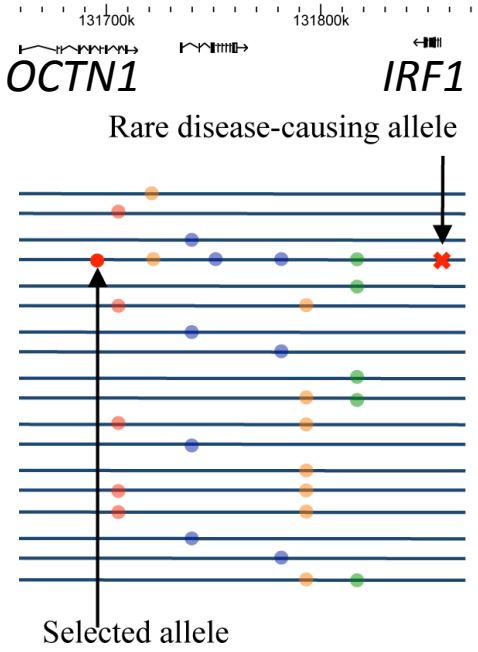
- Arose approximately 12,000 years ago; freq. 30-50% in Europe.
- 503F is a gain-of-function mutation that increases ergothioneine substrate efficiency by 300%.

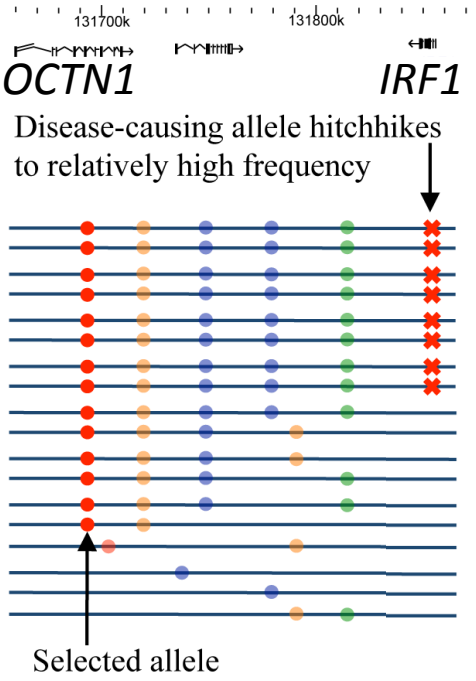
Recent Positive Selection at IBD5

Sample	iHS	p-value
HapMap CEU	-3.1	0.0007
HGDP Russian	-2.75	0.0044
HGDP Sardinian	-2.76	0.0075
HGDP French	-2.64	0.0076
HGDP Basque	-2.37	0.0128

Huff et al., 2011, *Mol. Biol. Evol.*

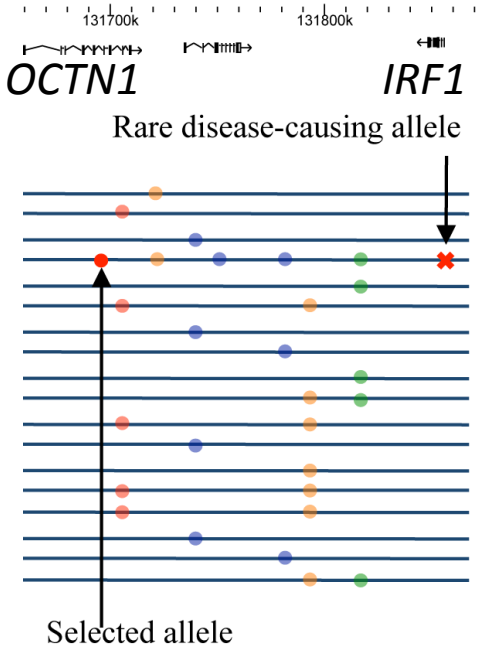
The *OCTN1* association could be explained by genetic hitchhiking

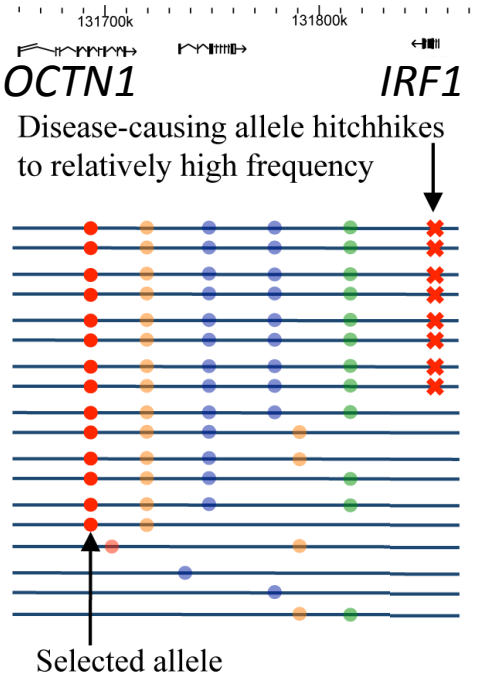




IRF1 is involved in innate immunity and clearance of intracelluar bacteria

Disease association in 1868 cases and 5540 controls





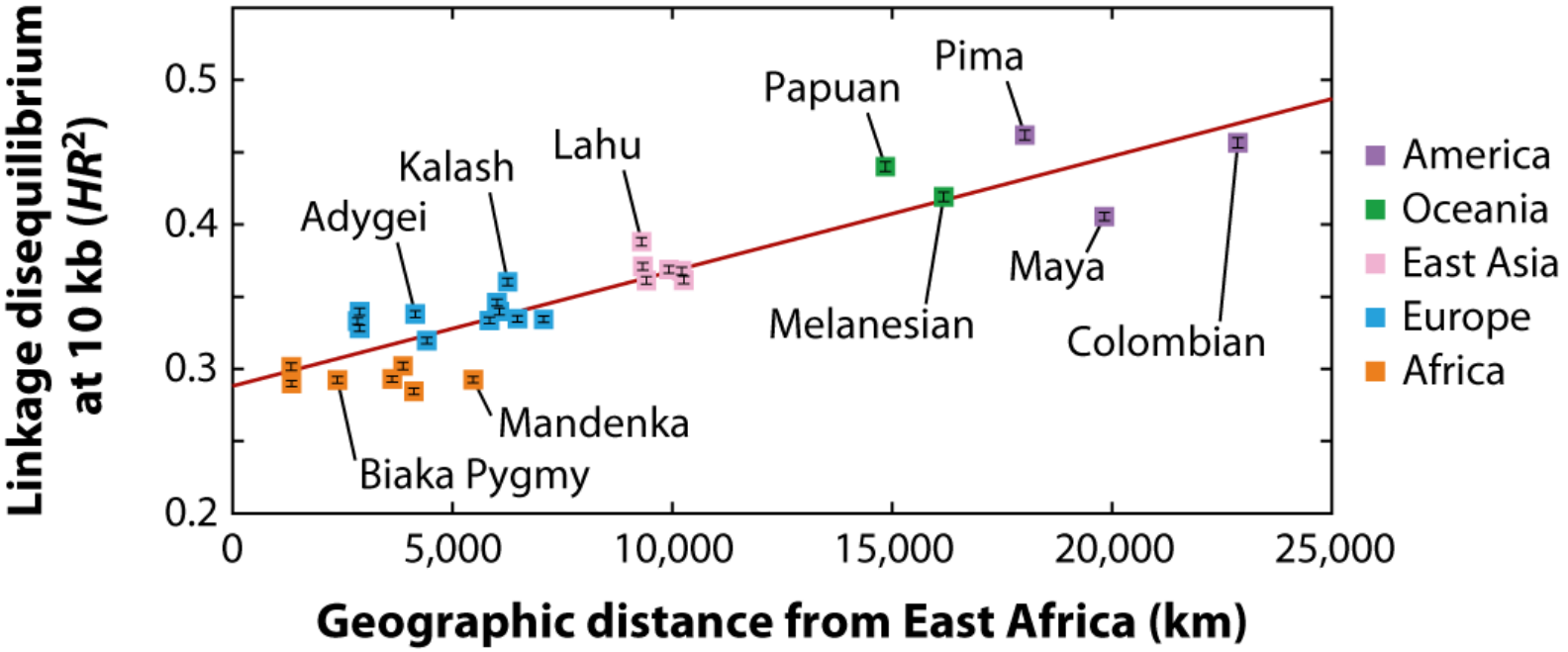
Odds ratio: 1.24
$p=2.6 \times 10^{-8}$
Frequency in controls: 32%
(37% in cases)

Odds ratio: 1.05
p=0.2
Frequency in controls: 10%

IRF1 is expressed 72% more highly in Crohn disease intestinal tissue than in control tissue; no other gene in *IBD5* region shows expression differences

Huff et al., 2011, *Mol. Biol. Evol.*

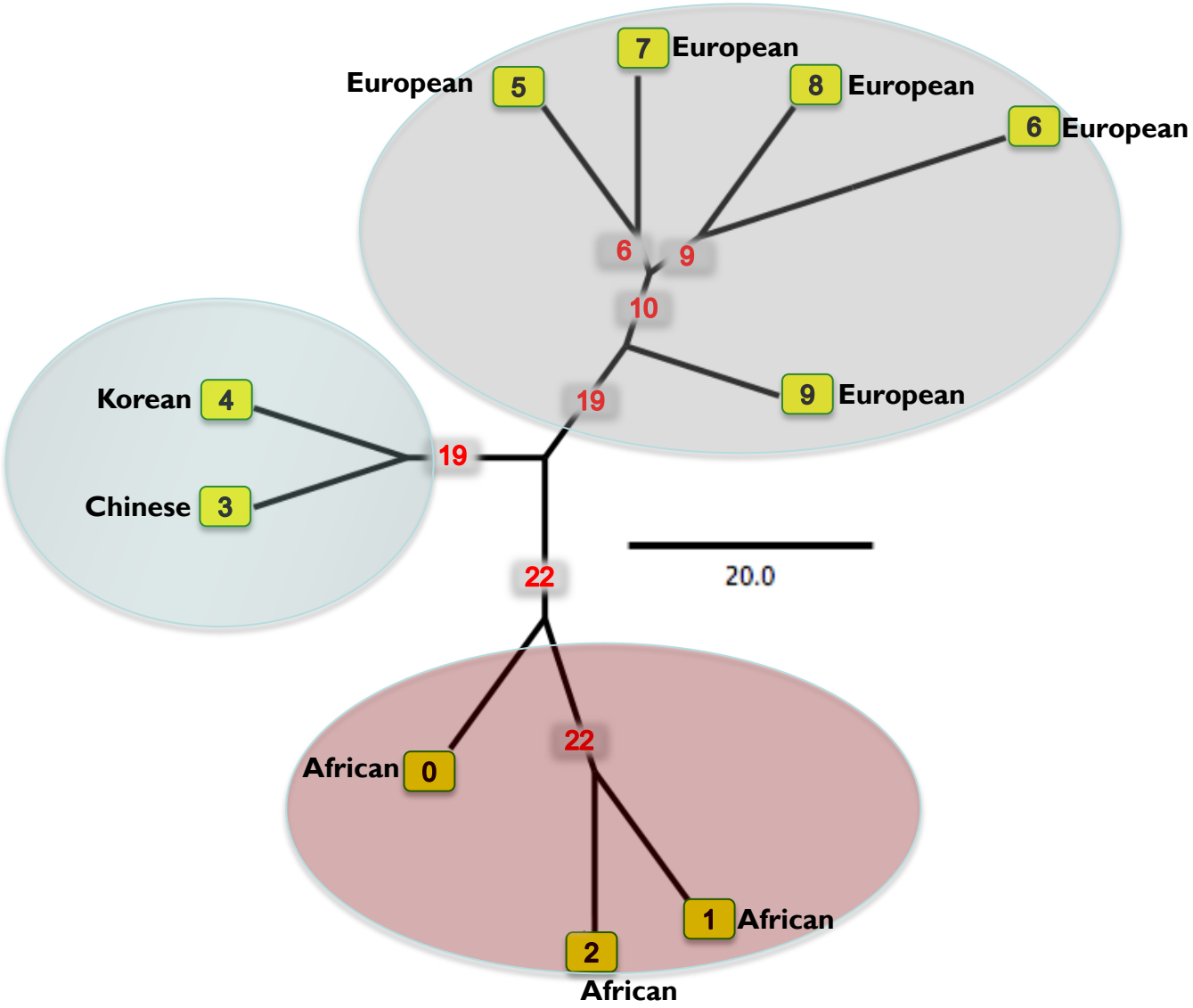
Linkage disequilibrium* increases with distance from Africa



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

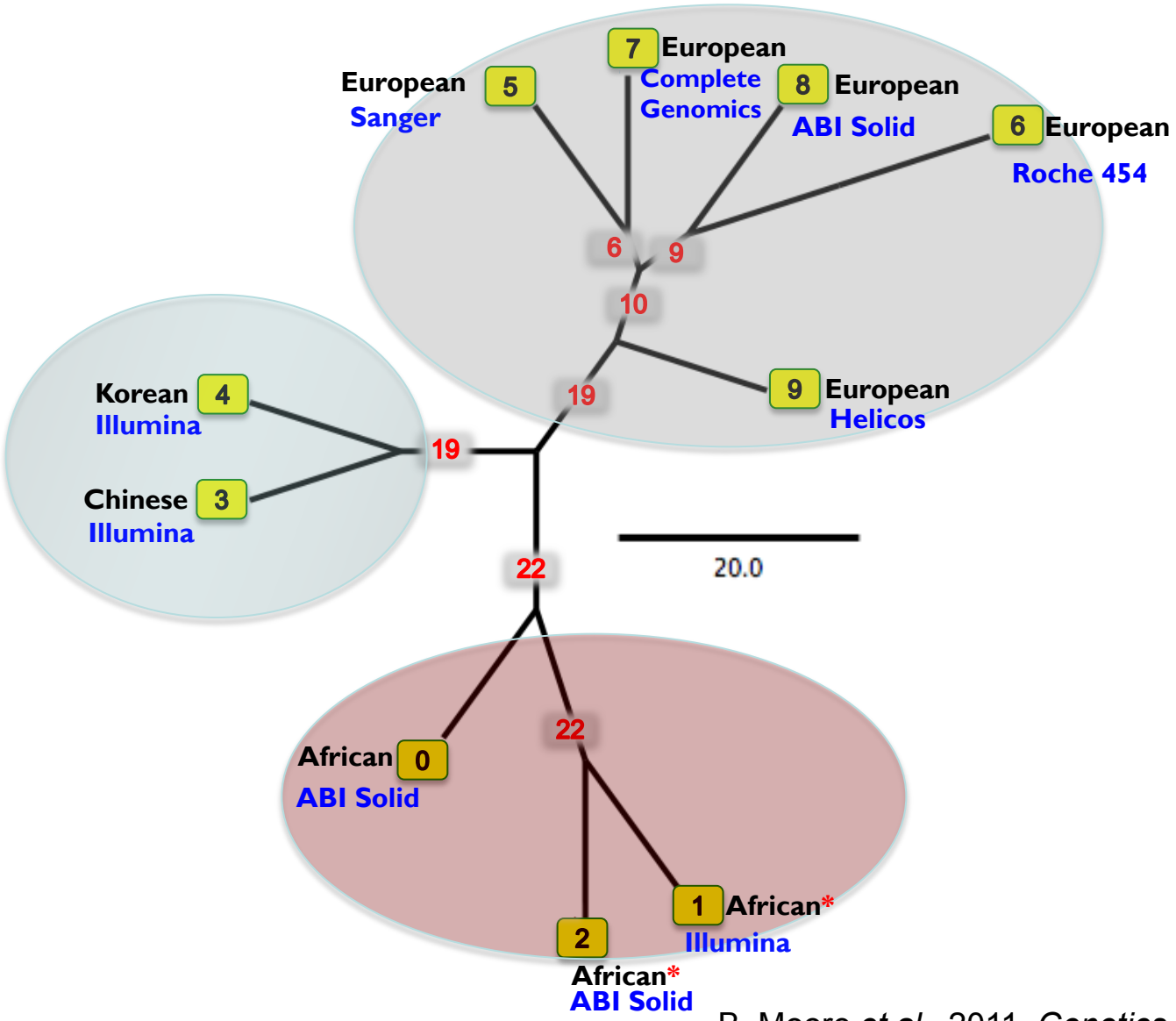
*We can think of linkage disequilibrium as a measure of multi-locus homozygosity.

Whole-genome sequence data give results congruent with array SNPs



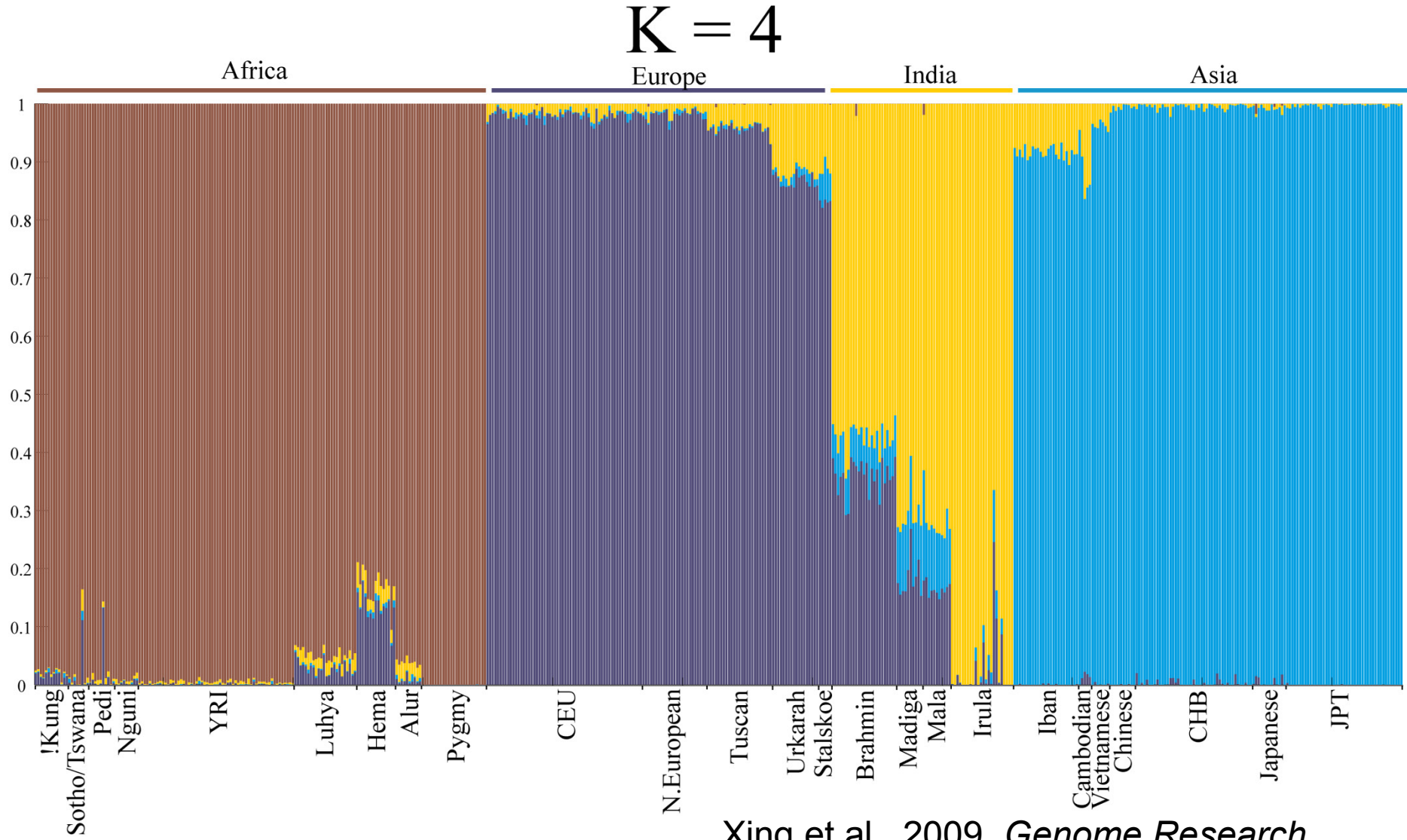
B. Moore *et al*., 2011, *Genetics in Medicine*

But results vary by platform



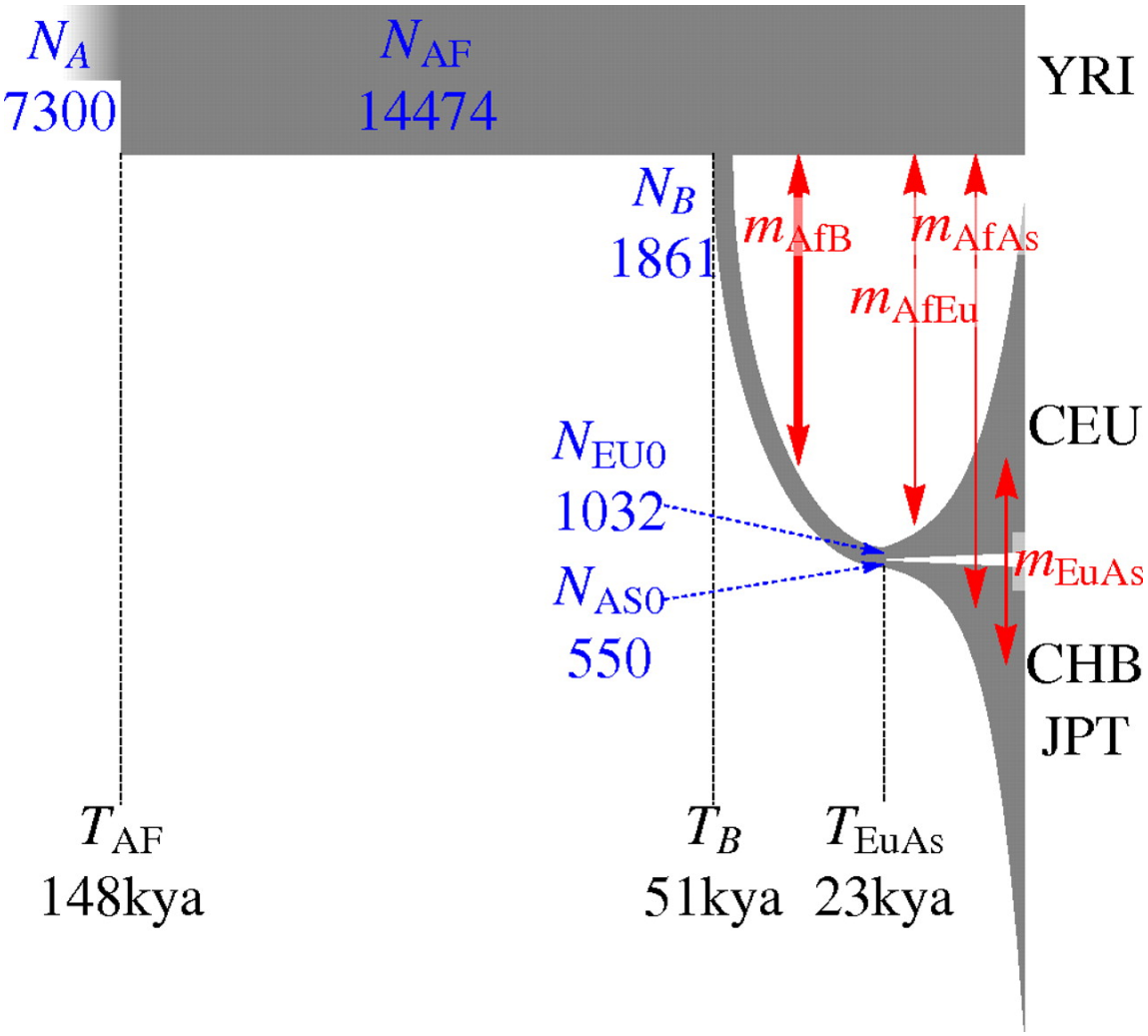
B. Moore *et al.*, 2011, *Genetics in Medicine*

Ancestral profiles: 250K SNPs *Structure* analysis



Xing et al., 2009, *Genome Research*

An inferred demographic model, with line width corresponding to population size and time flowing from left to right (1000 Genomes data)



Gravel S et al. PNAS 2011;108:11983-11988

The age, *t*, of a neutral allele can be estimated by its frequency

$$t = \frac{-4Np}{1-p} \ln(p)$$

(where *N* is effective population size)

Kimura and Ohta, 1973, *Genetics*

Allele sharing and allele frequency for 3,228 *Alu* insertion polymorphisms

Allele frequency, p:	**0 < p < 0.05**	**0.05 < p < 0.10**	**p > 0.10**
Probability of observing *Alu* outside Africa, given ascertainment in Africa	0.09	0.25	0.80
Probability of observing *Alu* in Africa, given ascertainment outside Africa	0.41	0.76	0.97

53 African samples (Bantu and Pygmy)
49 non-African samples (Tuscan and Brahmin)

David Witherspoon, PhD