
Using ancestry estimates as tools to
better understand group or individual
differences in disease risk or disease
outcomes

Jill Barnholtz-Sloan, Ph.D.

Assistant Professor

Case Comprehensive Cancer Center

CWRU School of Medicine

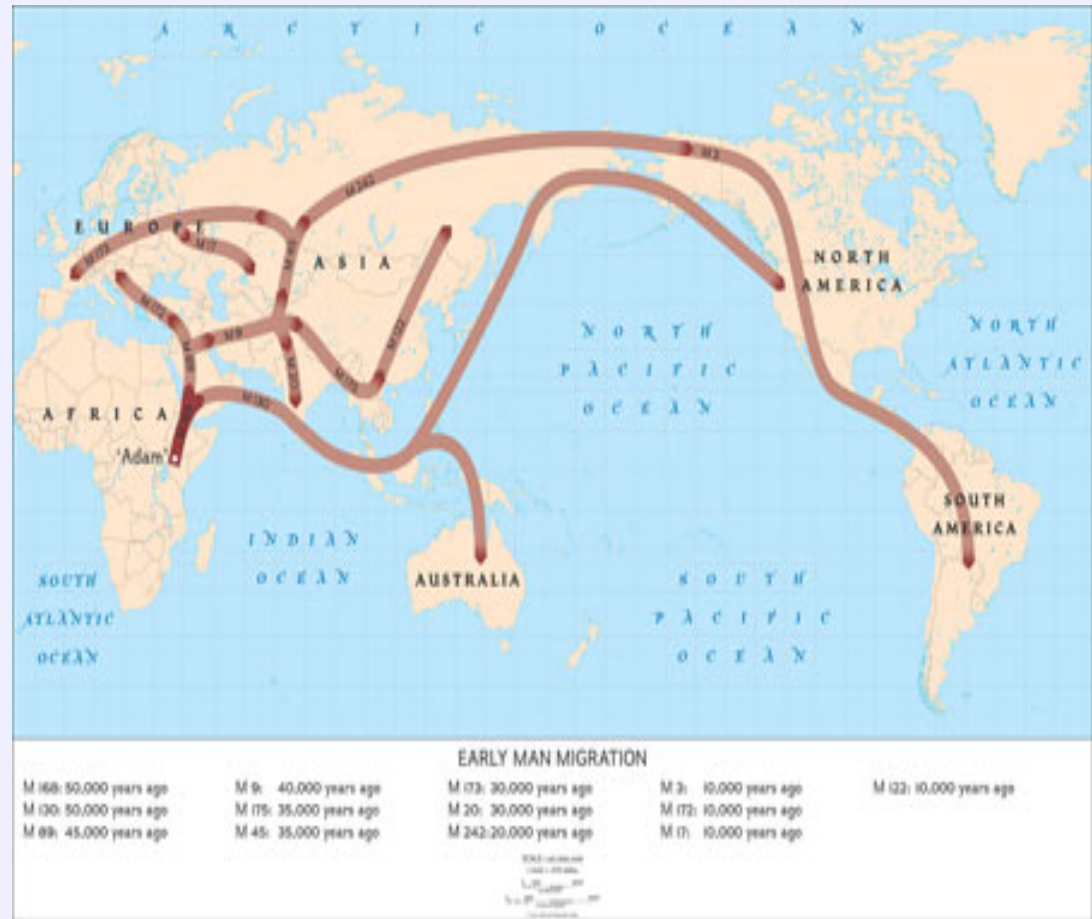
jsb42@case.edu

Human Genetic Variation

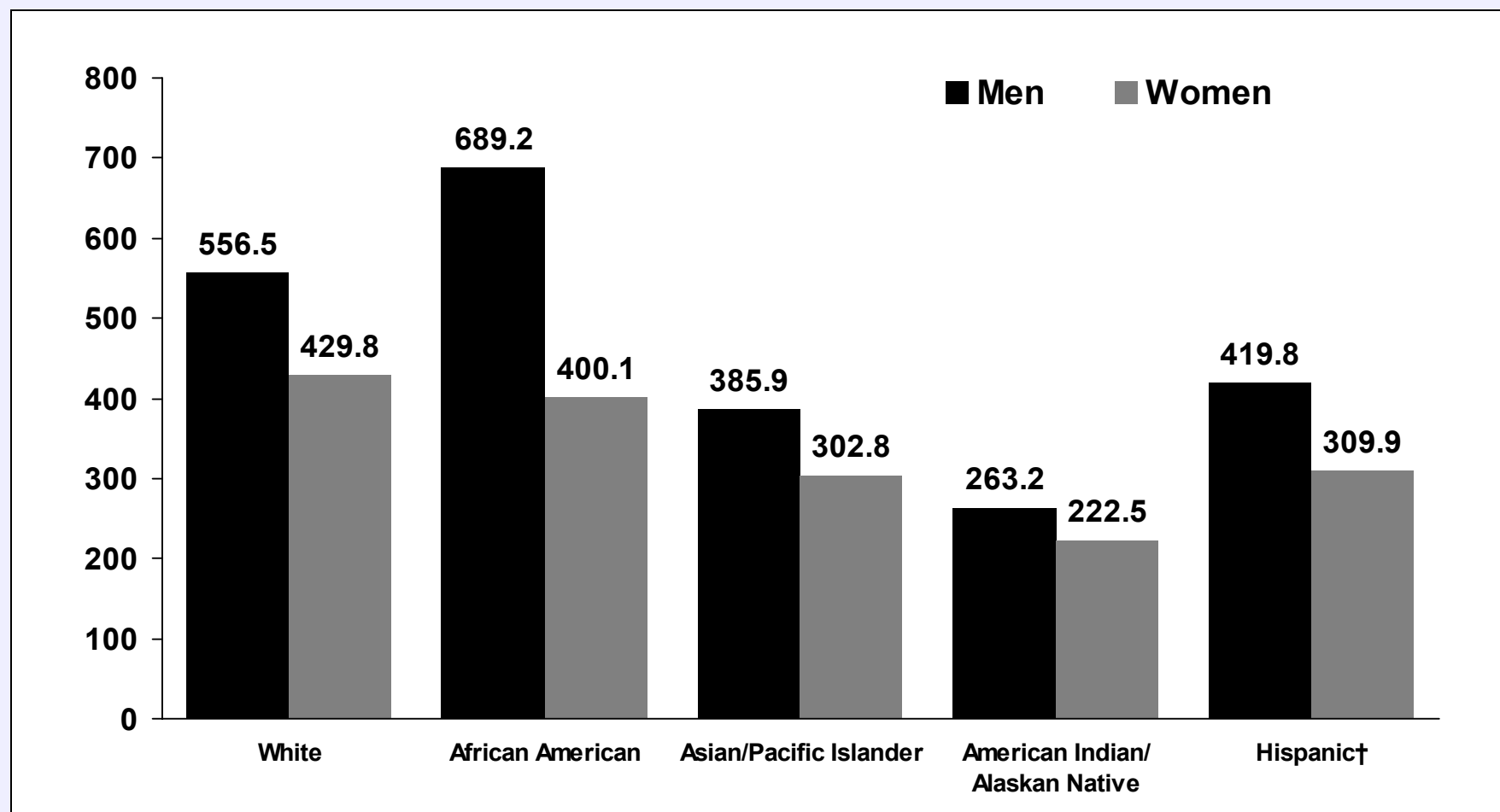
- Human genome = 3 billion nucleotides
 - 99.5-99.8% similar between humans
 - 0.2-0.5% of genome causes the genetic variation in phenotypes
 - 6 to 15 million nucleotides!!
 - Only 5-15% of this variation is between continents
 - Majority of variation is between individuals within the same continent
-
-

Out of Africa?

- Fossil evidence that we all evolved from one population in Africa within the last ~200,000 years



Cancer Incidence Rates* by Race and Ethnicity, 1997-2001



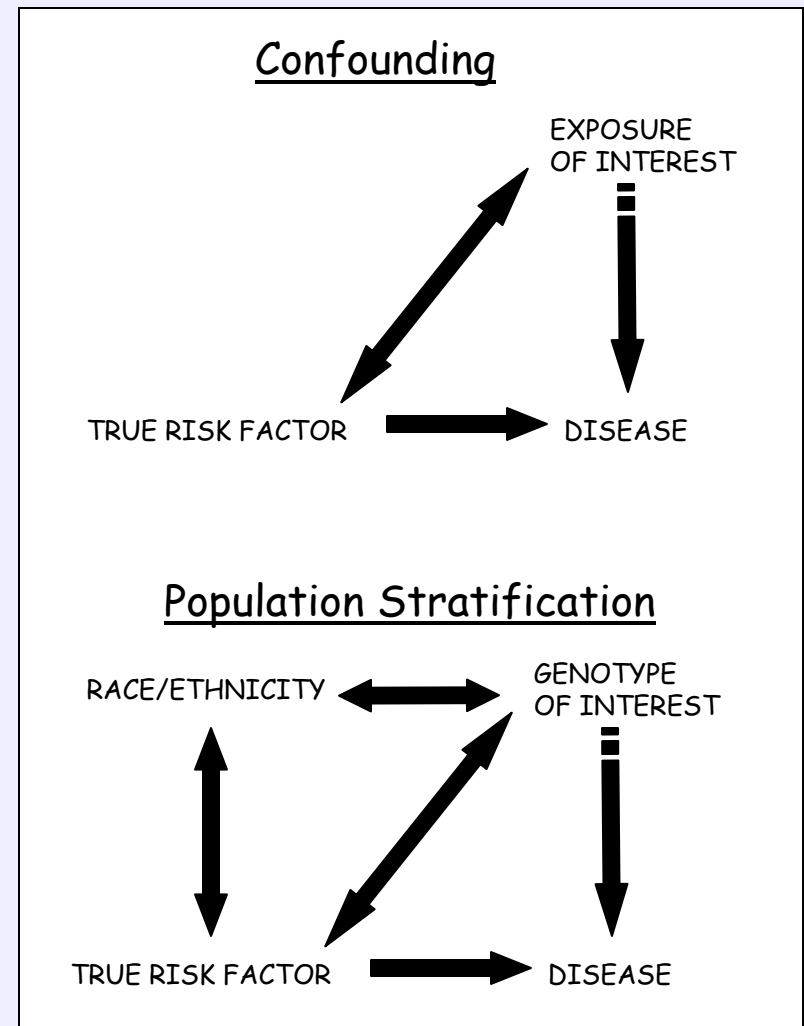
*Age-adjusted to the 2000 US standard population; rates per 100,000.

†Hispanic is not mutually exclusive from whites, African Americans, Asian/Pacific Islanders, and American Indians.

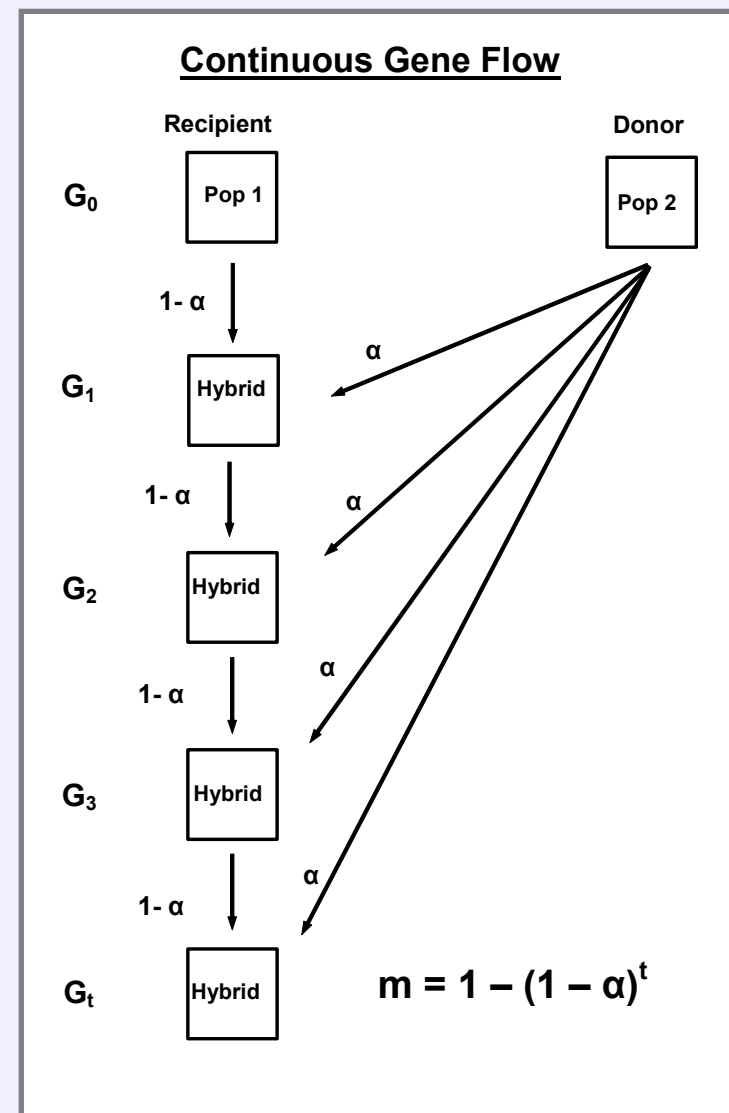
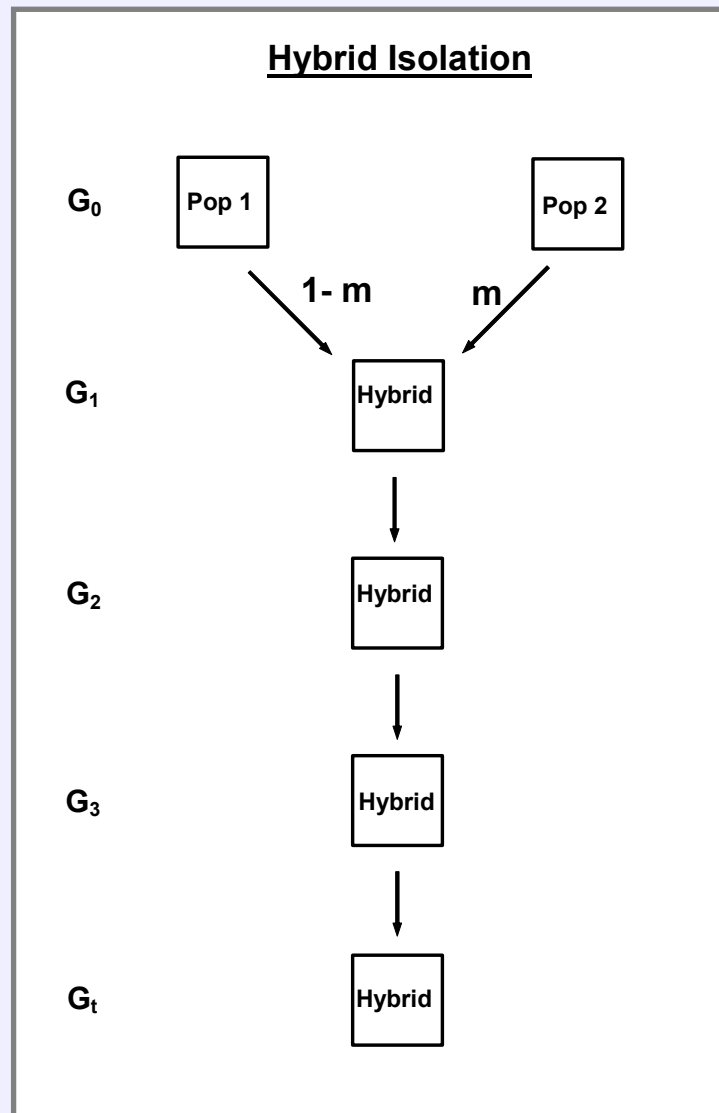
Source: Surveillance, Epidemiology, and End Results Program, 1975-2001, Division of Cancer Control and Population Sciences, National Cancer Institute, 2004.

Population Stratification

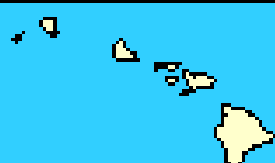
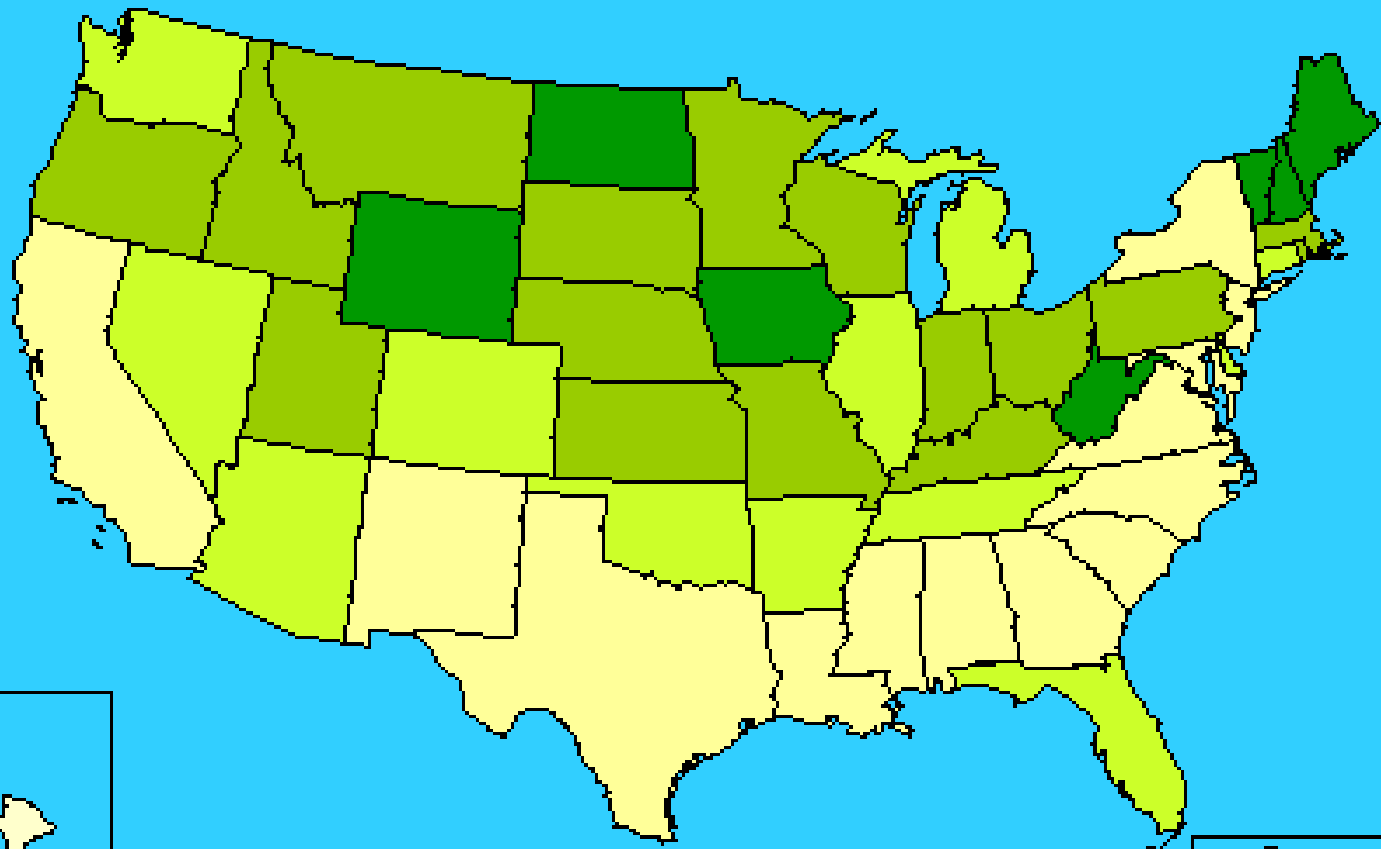
- Frequency of unmeasured "TRUE" risk factor for disease differs by race.
- Race acts as a surrogate for the true risk factor.



Two Models for Genetic/Racial Admixture

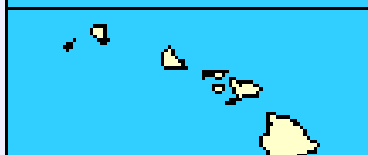
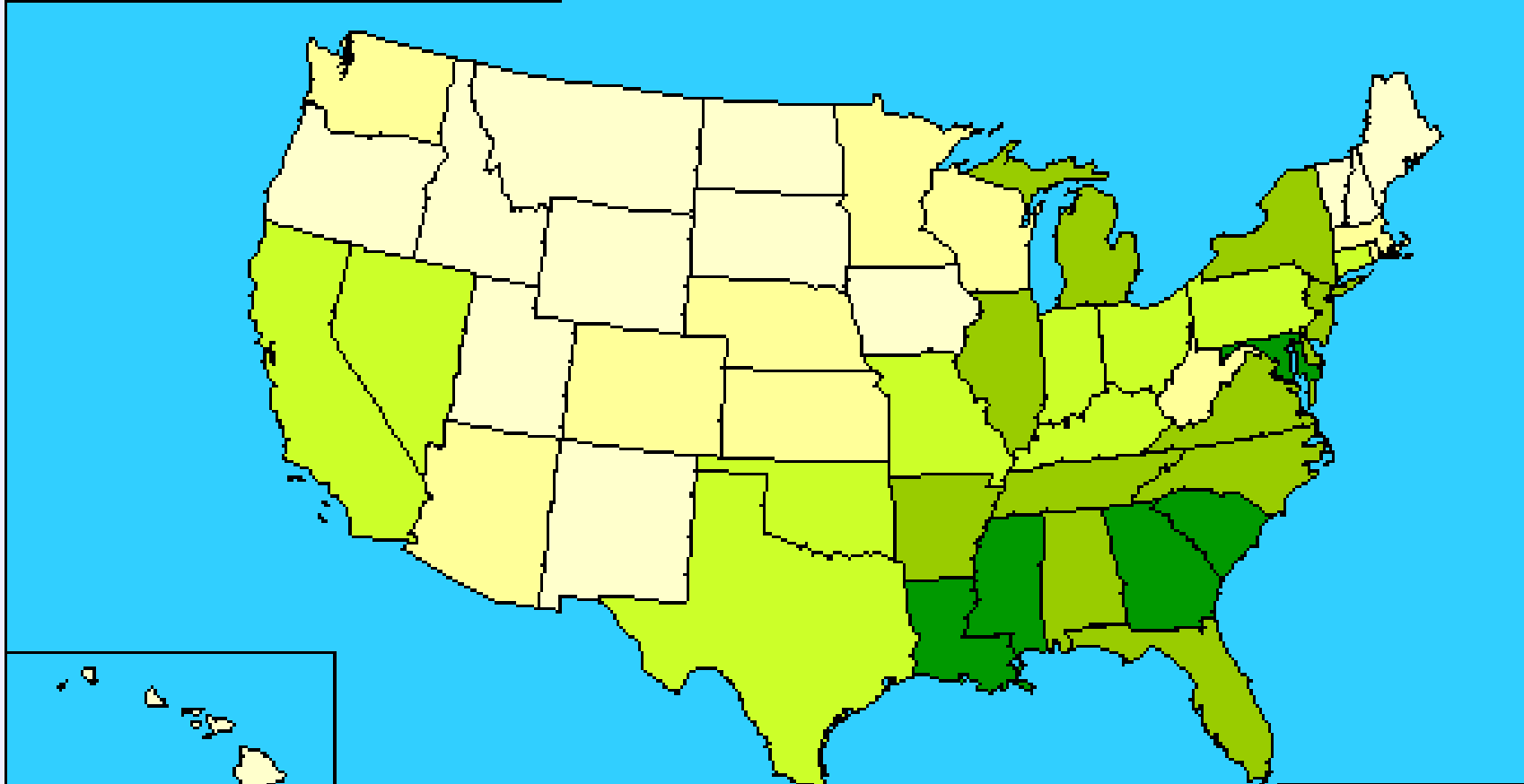


**Percent of Persons Who Are White Alone, United States by State
Census 2000**



Prepared with American FactFinder.

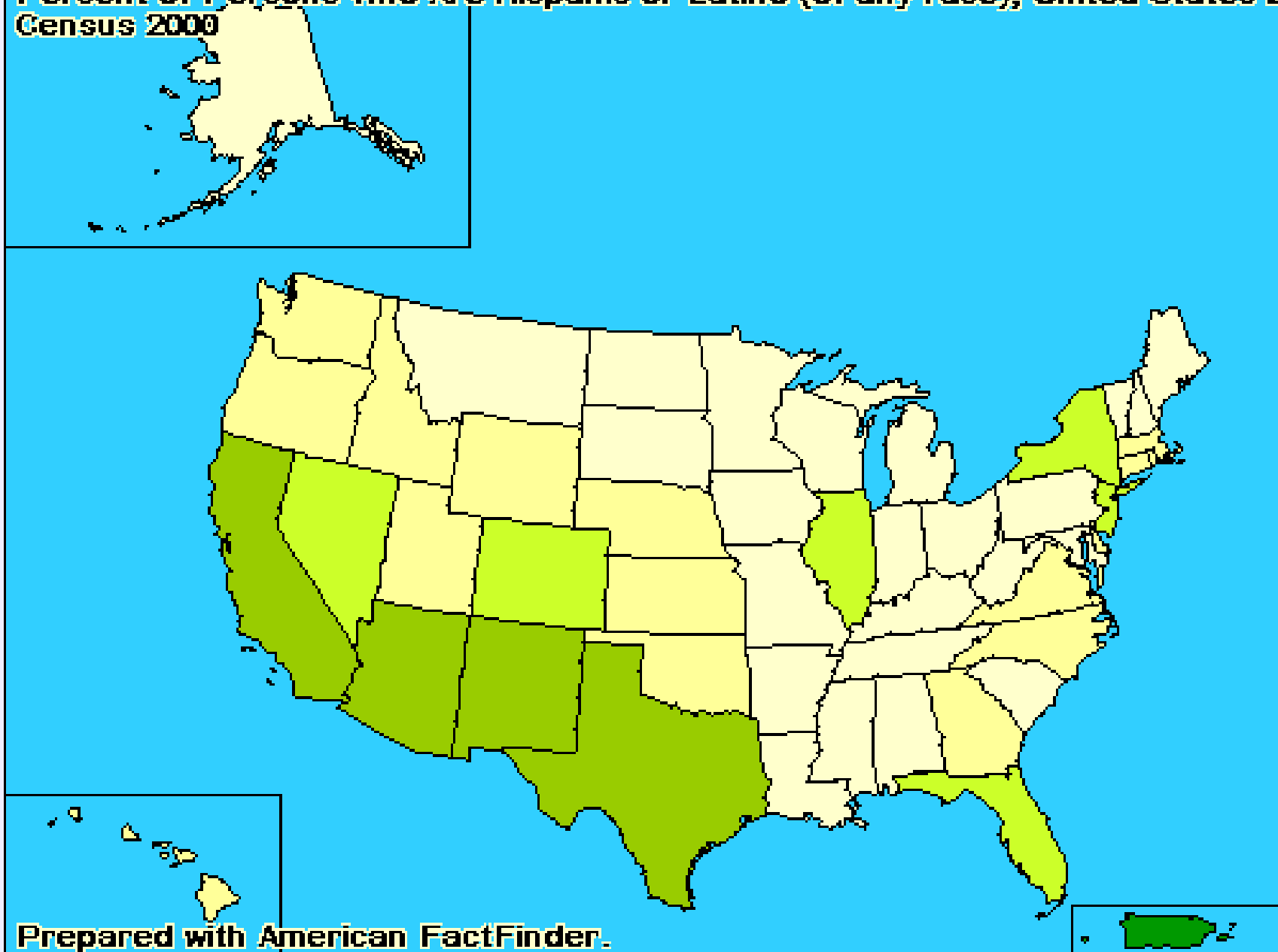
**Percent of Persons Who Are Black or African American Alone, United States
Census 2000**



Prepared with American FactFinder.

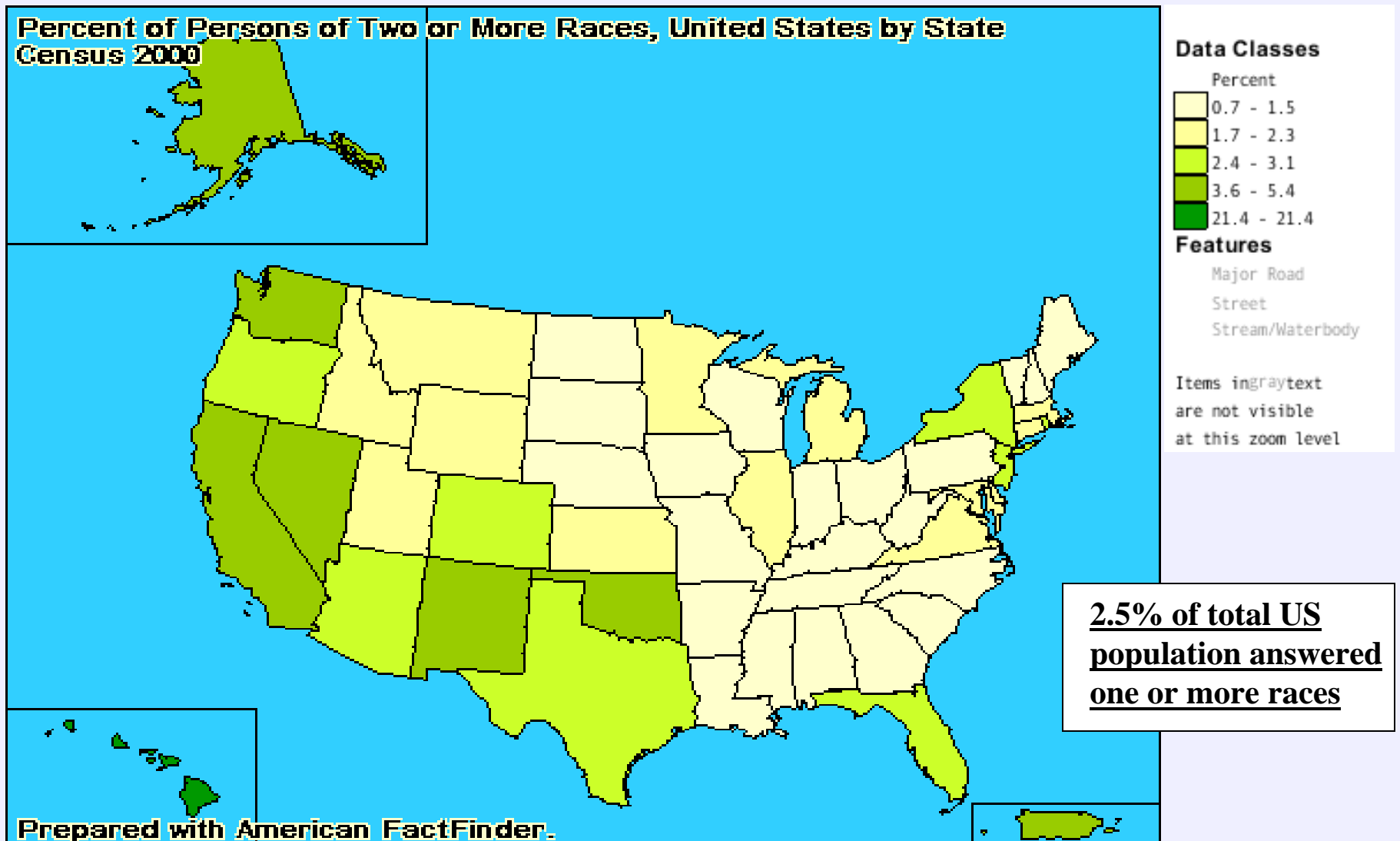


Percent of Persons Who Are Hispanic or Latino (of any race), United States by State, Census 2000



Prepared with American FactFinder.

Racial Admixture via the US 2000 CENSUS?



KEY CONCEPTS

Hardy-Weinberg Equilibrium

Linkage Disequilibrium (LD)

LD or Admixture LD Mapping

Hardy-Weinberg Equilibrium

- Proposed independently by George Hardy, an English mathematician, and Wilhelm Weinberg, a German physician, in 1908.
- Then, if there is random mating in a large population, with no migration, mutation and selection, the frequencies of the three genotypes, AA, Aa, and aa, in the population can be expressed in terms of simple *products of the allele frequencies*:

$$f(AA) = p * p = p^2$$

$$f(Aa) = p * q + p * q = 2pq$$

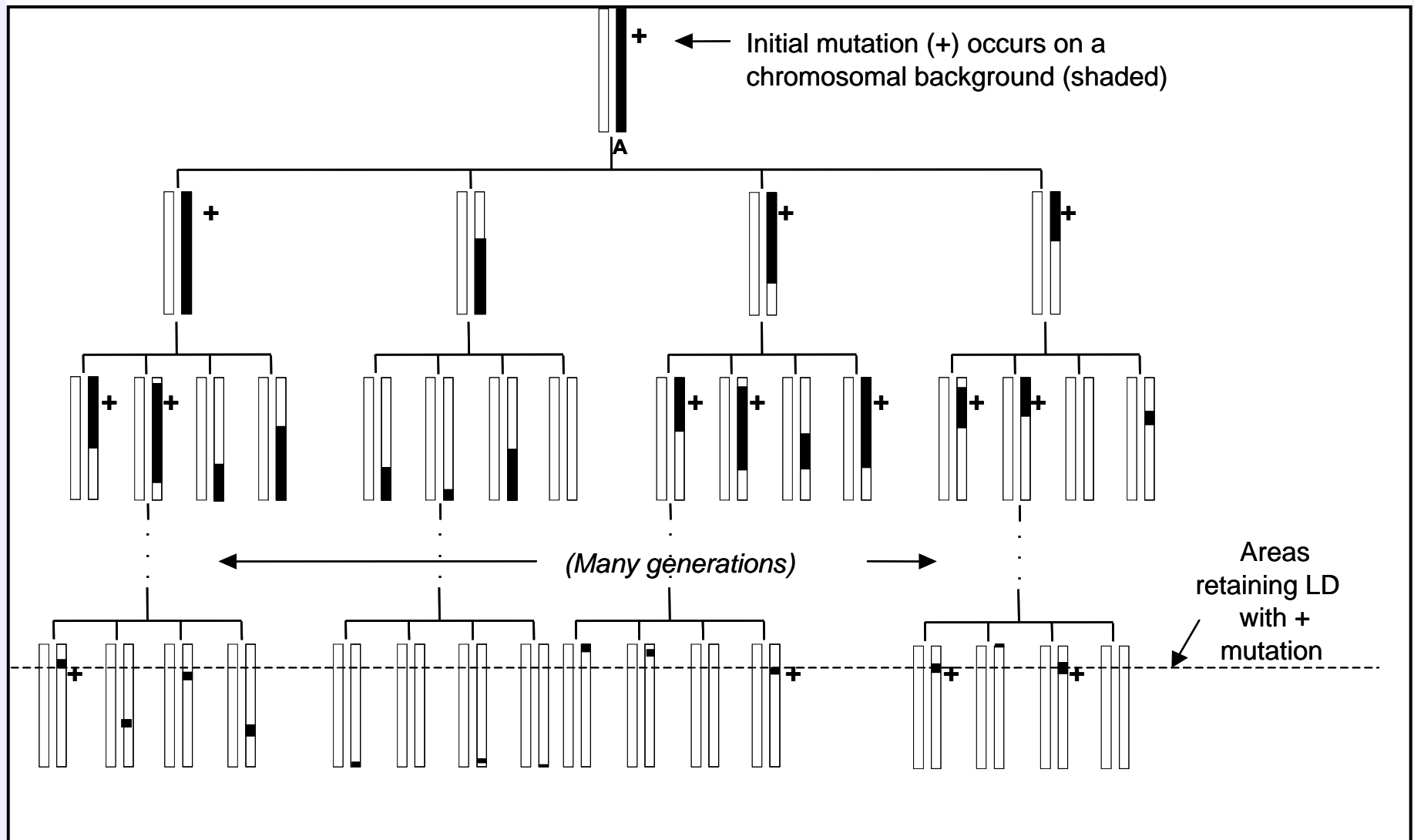
$$f(aa) = q * q = q^2$$

Linkage Disequilibrium (LD)

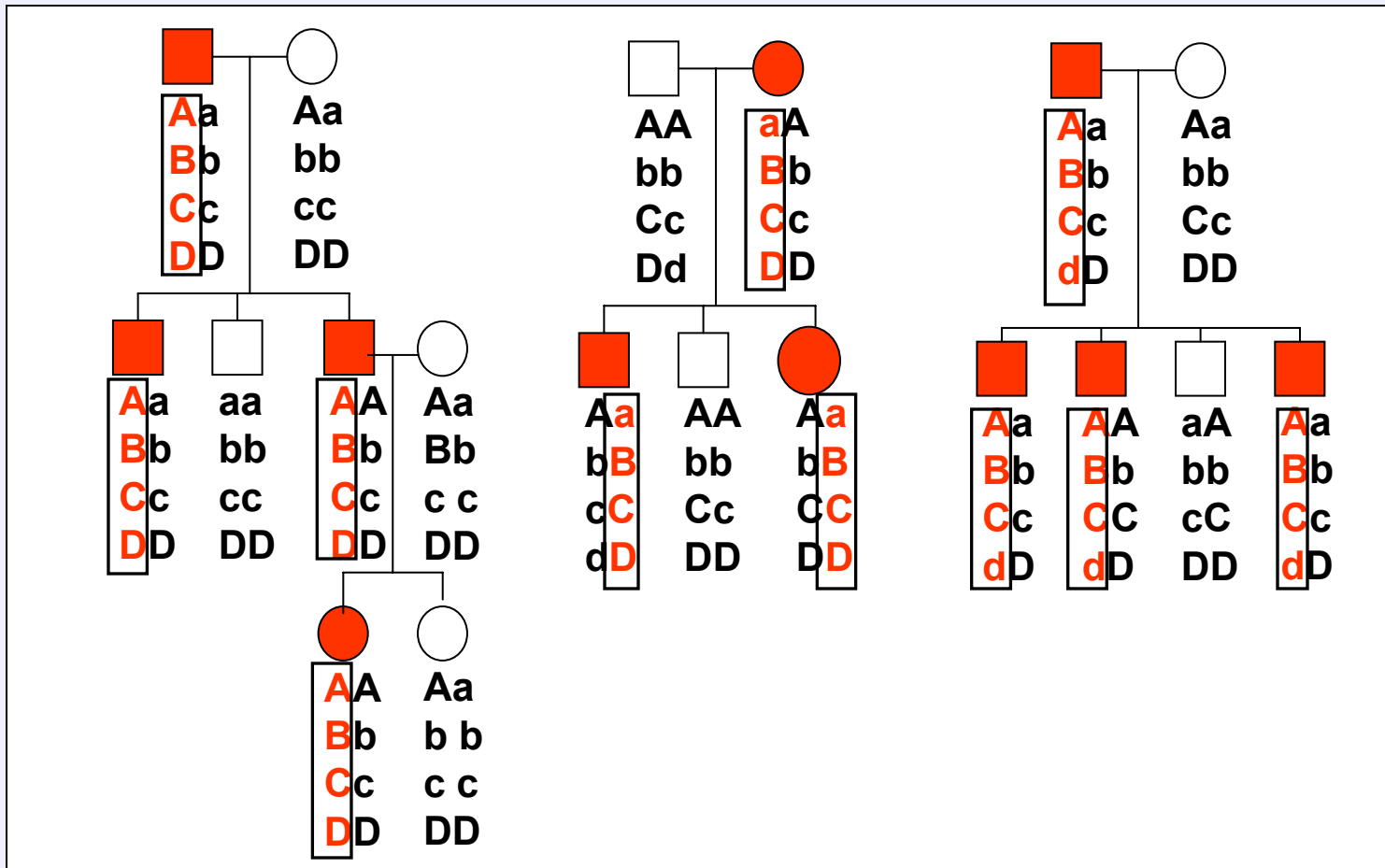
- A state of non-random gametic combinations of alleles of different genes
 - Excess or deficiency of haplotypes as compared to what would be expected by chance.
- Reduction of LD is a function of time and the recombination fraction.

Origin of Linkage Disequilibrium

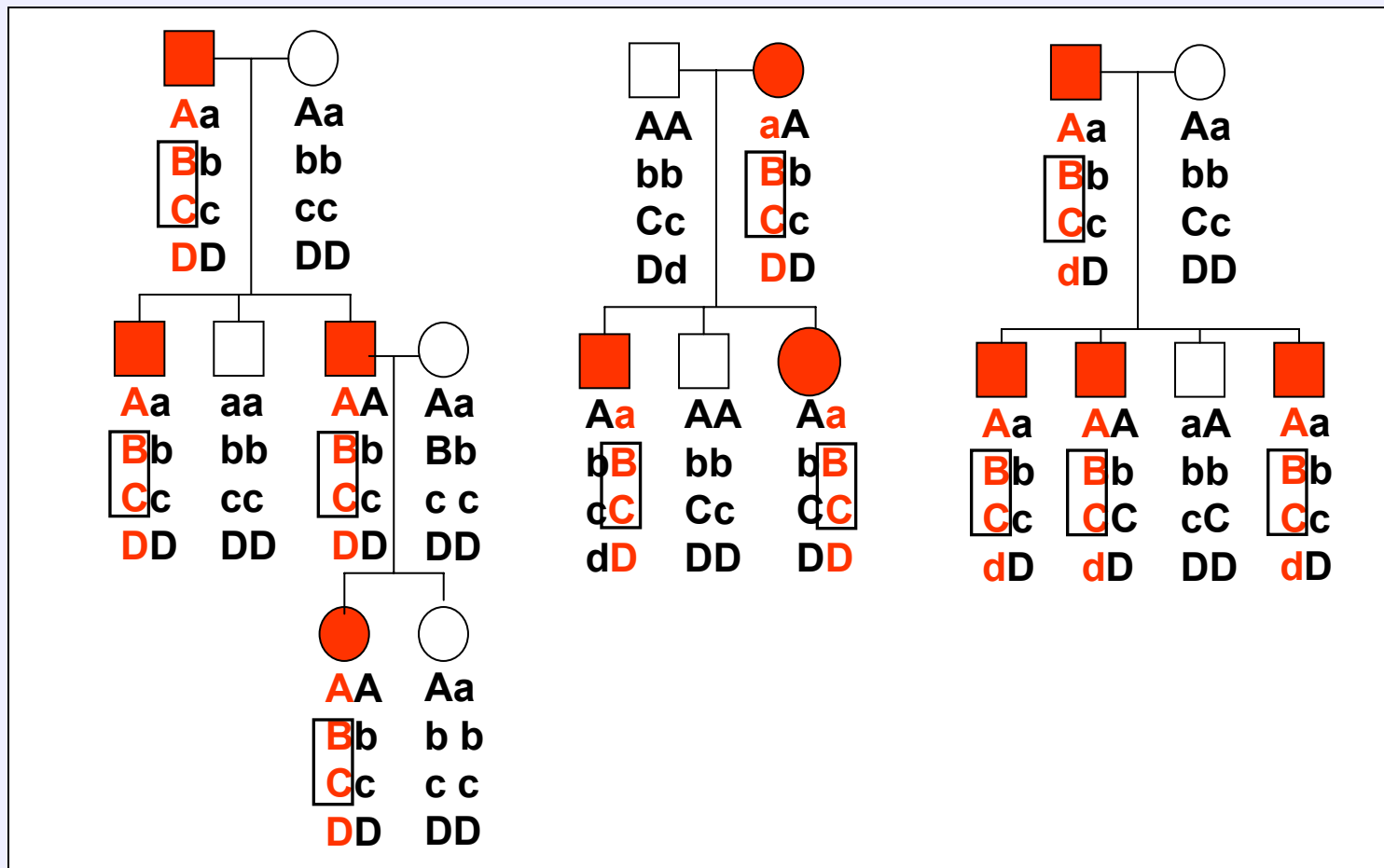
Example of Linkage Disequilibrium through generations



All 4 Loci "Linked" to (Unobserved) Disease Allele WITHIN Each Family



All 4 Loci are Linked to Disease Allele *WITHIN* Each Family, *BUT* Only Alleles 'B' and 'C' are Associated with Disease Allele *ACROSS* Families: Linkage Disequilibrium



LD or ALD Mapping

- Association studies based on LD are more powerful for modest risk of disease across a large range of genotype frequencies
 - ALD mapping detects disease-associated variants that vary significantly in frequency by race/ethnic group.
 - Recent admixture between populations can create extensive LD that will not decay for 10-20 generations.
 - Large chromosomal blocks by ancestry or race/ethnicity are formed
 - Populations in Africa have shorter LD blocks
 - Hap Map Project (www.hapmap.org)
 - Estimates amount of LD or ALD present and then uses this measure to assess association between trait and genotype in cases and controls.
-
-

Methods to assess population stratification in case-control studies

- Association studies can produce false results if cases and controls have differing allele frequencies by racial or ancestral group.

(1) GENOMIC CONTROL (Devlin and Roeder; <http://www.stat.cmu.edu/>)

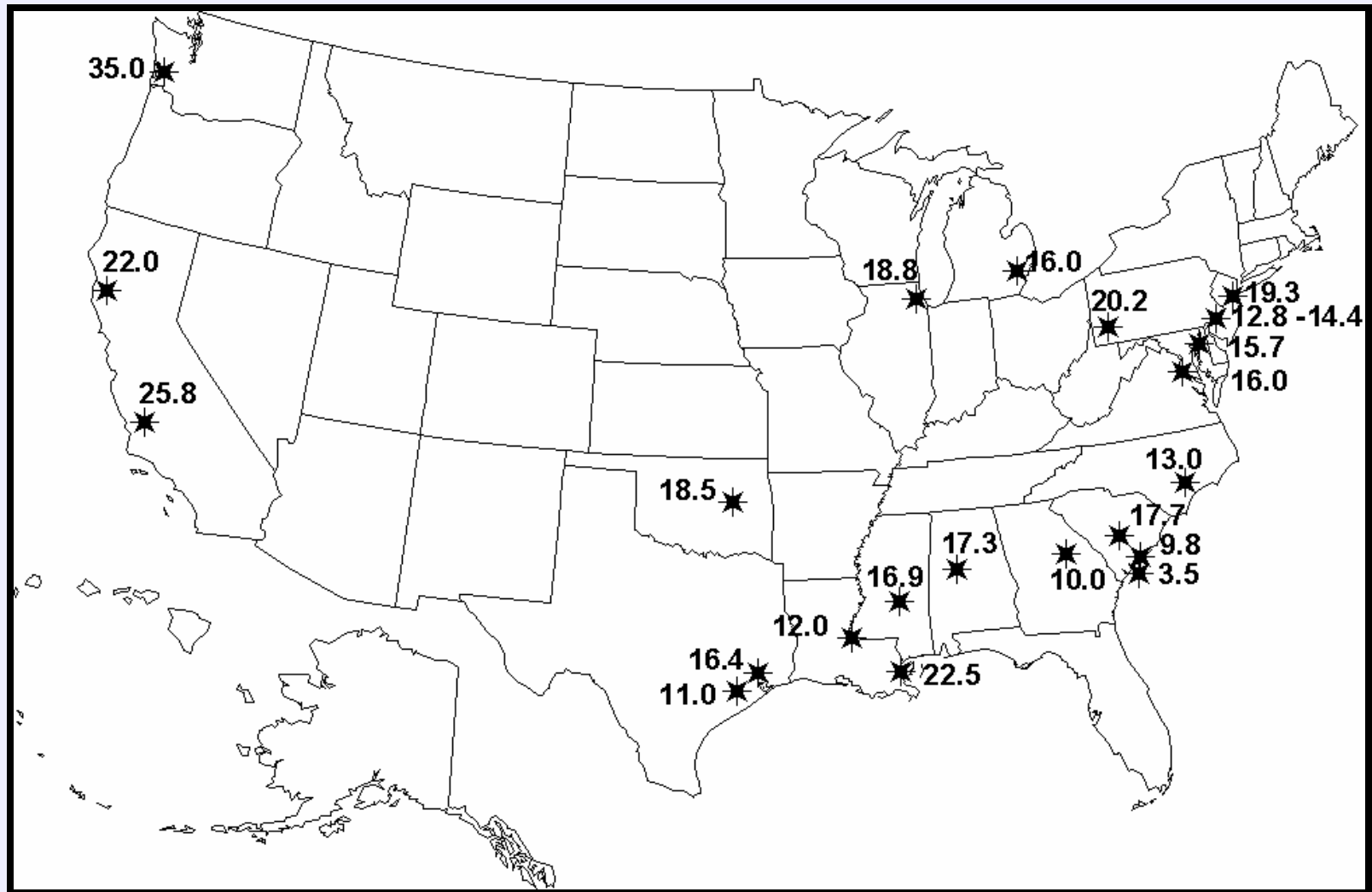
(2) STRUCTURE and STRAT (Pritchard et al.; <http://pritch.bsd.uchicago.edu/software/>)

(3) INDIVIDUAL OR GROUP ANCESTRY ESTIMATION

Individual or Group Ancestry Estimation

- Estimating ancestry and using it in association with a trait is a new technique.
 - Ancestry estimates vary widely in the United States.
 - Examples:
 - (1) African Americans and Mexicans
 - (2) Breast cancer in Latinas
 - (3) LCT gene and height
 - (4) drug metabolizing enzymes and ancestry
 - (5) early-onset lung cancer
 - (Barnholtz-Sloan et al., 2005)
 - Other examples: diabetes (Williams et al, 2000), obesity (Fernandez et al, 2003), insulin-related phenotypes (Gower et al, 2003); asthma (Choudhry et al, 2005); prostate cancer (Kittles et al.)
-
-

European genetic contribution in African-American populations living in different geographical areas of the US.

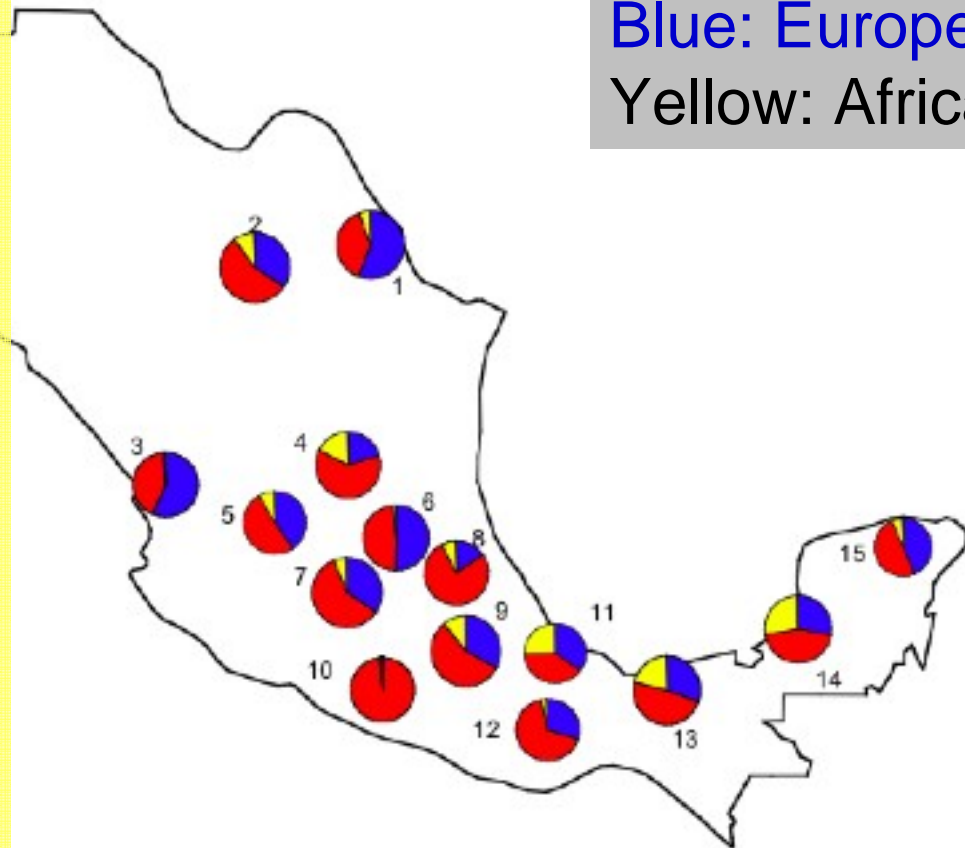


Parra et al. AJHG 1998; Parra et al. AJPA 2002; Kittles et al. unpublished

Heterogeneity in Mexican Admixture: an example of The Metapopulation Fallacy

- 1: Monterrey, Nuevo León
- 2: Saltillo, Coahuila
- 3: Guadalajara, Jalisco
- 4: Cuernalán, Mexico
- 5: León, Guanajuato
- 6: DF1
- 7: DF2
- 8: Tlaxcala, Tlaxcala
- 9: Puebla, Puebla
- 10: Tlapa, Guerrero
- 11: Veracruz, Veracruz
- 12: Oaxaca, Oaxaca
- 13: Paraíso, Tabasco
- 14: El Carmen, Campeche
- 15: Mérida, Yucatán

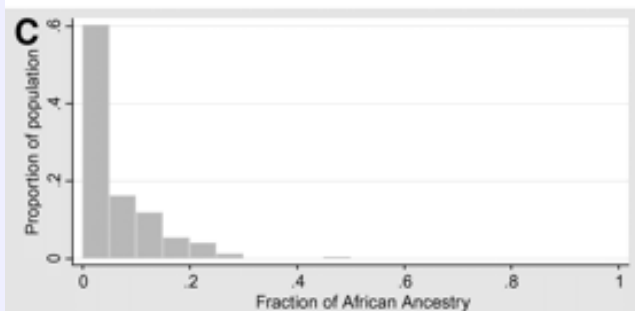
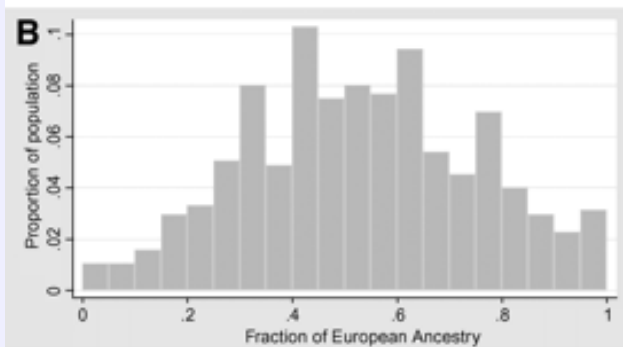
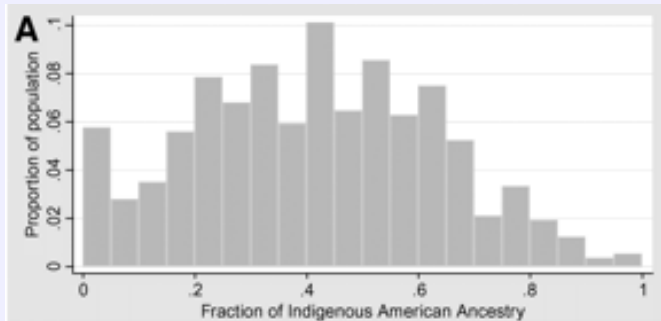
Red: Indigenous
Blue: European
Yellow: African



Summarized in Bonilla *et al.*, 2005 *AJPA*

Breast cancer in Latinas

(Ziv et al., 2006)



- Latina women with invasive breast cancer and age-race/ethnicity matched controls (234 cases and 329 controls)
- 44 ancestry informative markers used to estimate proportions of Native American, European and African ancestry
- Are breast cancer risk factors (BMI, hormone therapy use, parity) associated with ancestry?
- Hormone therapy less common among women with higher Native American ancestry (OR=0.78 (0.63,0.93))
- Higher Native American ancestry associated with being overweight (BMI=25-29.9) or obese (BMI \geq 30) but only among foreign born Latina women (OR=3.44 (1.97,5.99) and OR=1.95 (1.24,3.06))

LCT gene and height

(Campbell et al., 2005)

- Height is a heritable trait that varies significantly across Europe
 - European American individuals discordant for height were studied
 - Using markers highly informative for ancestry showed NO population stratification

 - LCT 13910 C->T polymorphism has variation in allele freqs that follows the variation in average height across Europe
 - T allele associated with tall stature (OR=1.37 (1.22-1.54))
 - Height and LCT polymorphism correlated with grandparental ancestry
 - Not found in Scandinavian or Polish populations
 - Slightly lower association if matched on grandparental ancestry (OR=1.19 (1.05-1.36))

 - Recent data for population substructure in Iceland (Helgason et al., 2005); and between Northern/Southern European (Seldin et al, 2006)
-
-

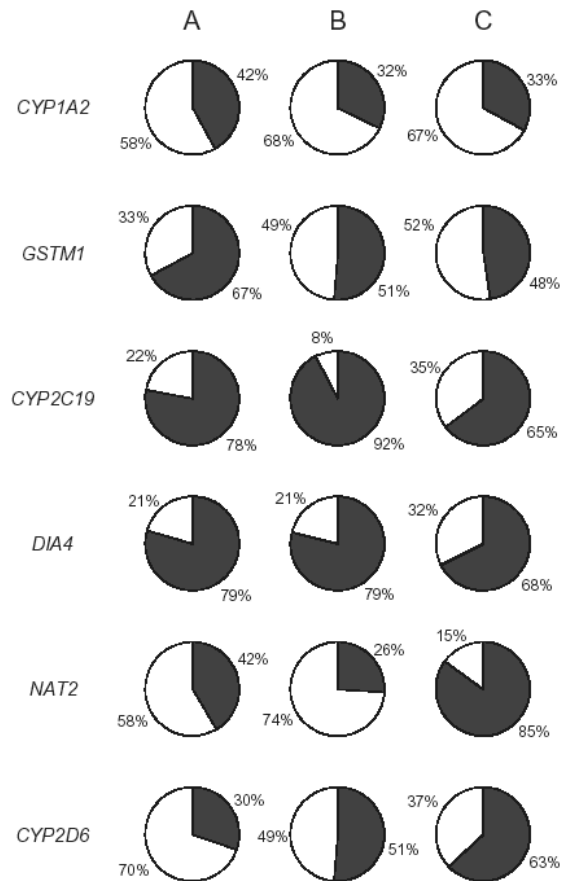
Drug metabolizing enzymes and ancestry

(Wilson et al., 2001)

- Inter-individual drug response best explained by differences in ancestry not in racial/ethnic group

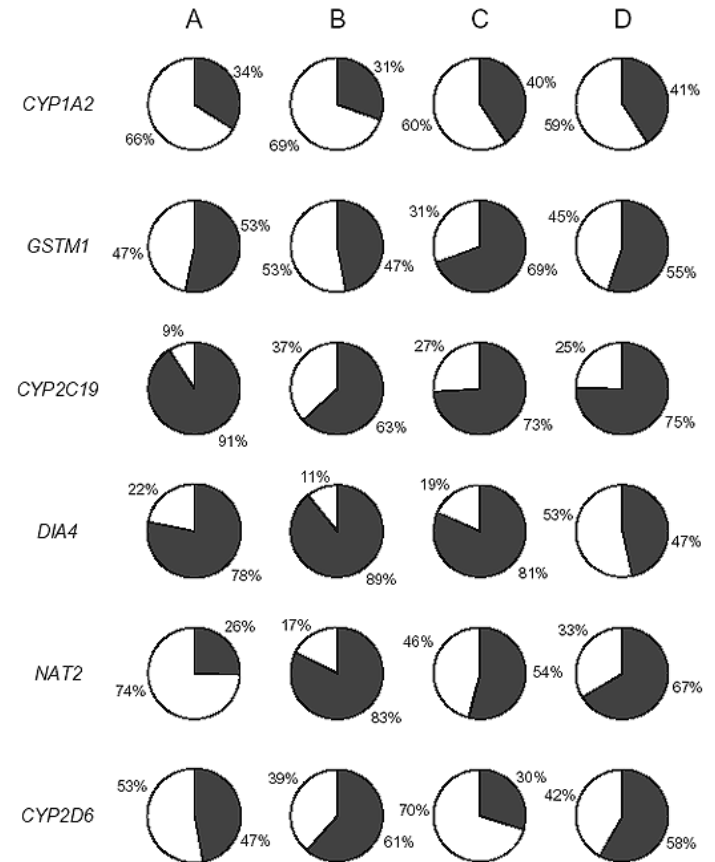
By race/ethnic Group

A - Bantu, Ethiopian, Afro-Carib
 B - Norwegian, Ashkenazi, Armenian
 C - Chinese, New Guinea



By ancestry STRUCTURE cluster

A - European
 B - New Guinea
 C - African
 D - Asian



Early-onset lung cancer

Barnholtz-Sloan, Chakraborty, Sellers, Schwartz. Examining population stratification via individual ancestry estimates versus self-reported race. *CEBP* 2005.

- **Early onset lung cases** (age at diagnosis <50) and population-based controls
 - EUROPEAN AMERICAN (Non-Hispanic): 192 cases and 363 controls
 - AFRICAN AMERICAN: 60 cases and 131 controls
 - **2 different ancestral/parental populations** used to estimate individual ancestry from FBI CODIS 13 STR loci set:
 - European: average of German and Polish*
 - West African: average of Nigerian and Rwandan*
 - All individuals also genotyped for *GSTM1* candidate gene locus (present or null)
-
-

Statistical Analysis

- Calculated composite δ as $\frac{1}{2}$ of the sum across all loci pairs of the allele frequencies for European versus West African
 - Used when there are multiple alleles at a locus.
 - Estimated individual ancestry using MLE and STRUCTURE
 - “West African” ancestry equals one minus the “European” ancestry.
 - Logistic regression modeling of early-onset lung cancer risk for the *GSTM1* null genotype comparing adjustments for self-reported race versus individual ancestry
-
-

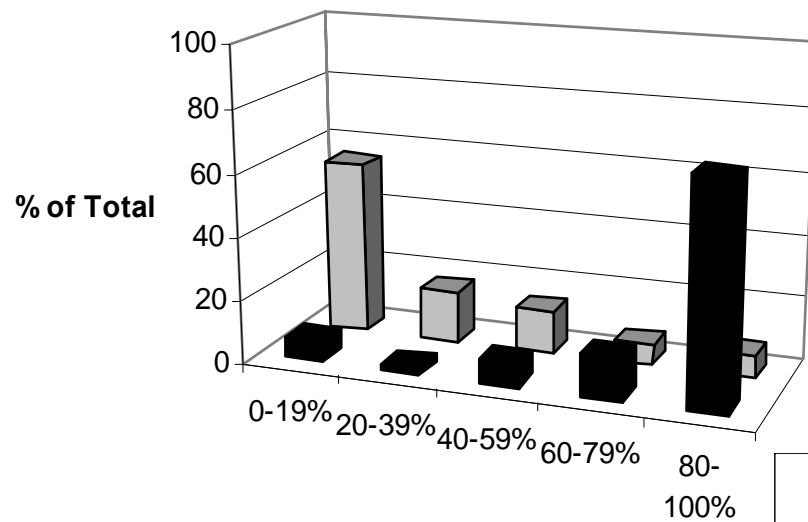
CODIS STR loci characteristics

	Chromosomal location (number of alleles)	Overall composite δ (δ_c)
CODIS loci name		European versus West African
CSF1PO	5q33.3-34 (14)	0.18
D13S317	13q22-q31 (12)	0.29
D16S539	16q22-24 (10)	0.19
D18S51	18q21.3 (20)	0.31
D21S11	21q21.1 (34)	0.25
D3S1358	3p21 (12)	0.16
D5S818	5q21-q31 (12)	0.17
D7S820	7q (18)	0.14
D8S1179	8q24.1-24.2 (11)	0.27
FGA	4q28 (31)	0.30
THO1	11p15-15.5 (10)	0.31
TPOX	2p23-2pter (10)	0.26
vWA	12p12-pter (12)	0.15

Median European Individual Ancestry by Self-reported Race and Case/Control Status

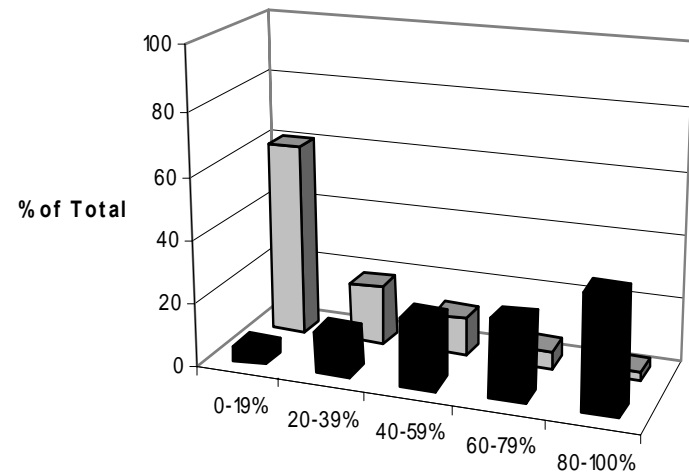
	European American			African American		
	Cases (N=192)	Controls (N=363)	p-value	Cases (N=60)	Controls (N=131)	p-value
Median "European" MLE	0.99	1.0	0.65	0.20	0.16	0.69
Median "European" STRUCTURE (cluster 2)	0.65	0.71	0.31	0.15	0.13	0.71
<i>GSTM1</i> Null (column %)	43.6%	48.3%	0.29	27.1%	28.8%	0.81

Histograms of Individual European ancestry by self-reported race



"European" MLE ancestry

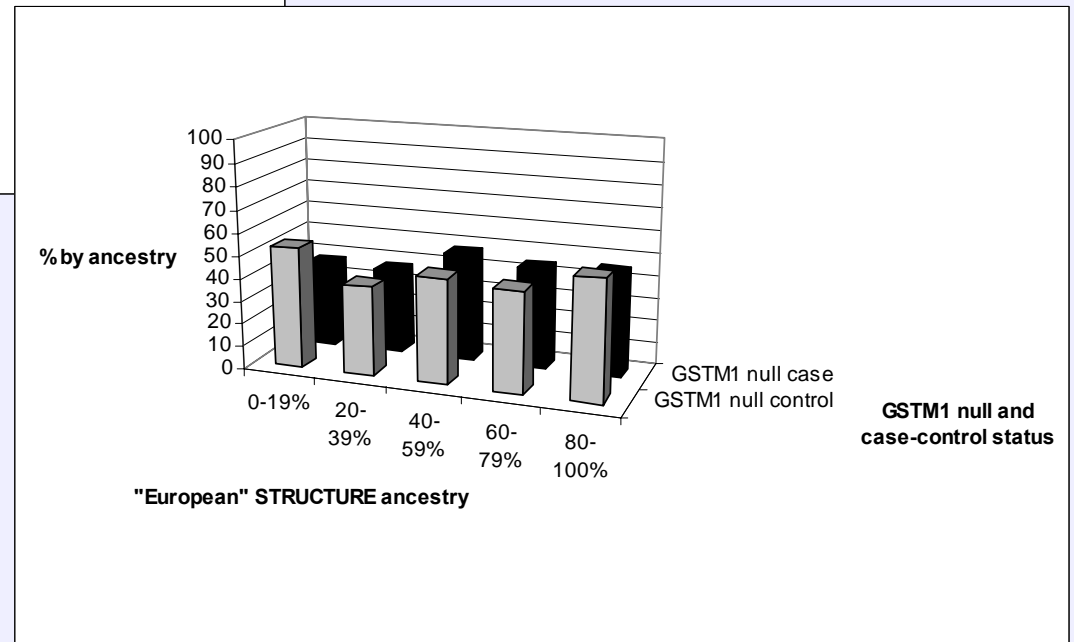
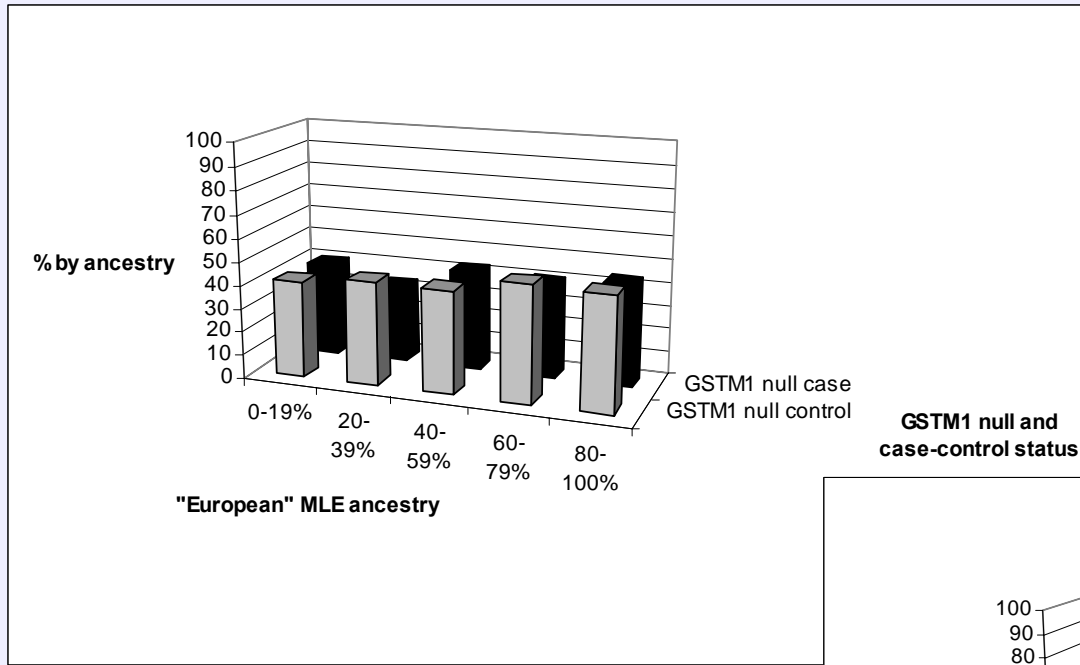
■ European American □ African American



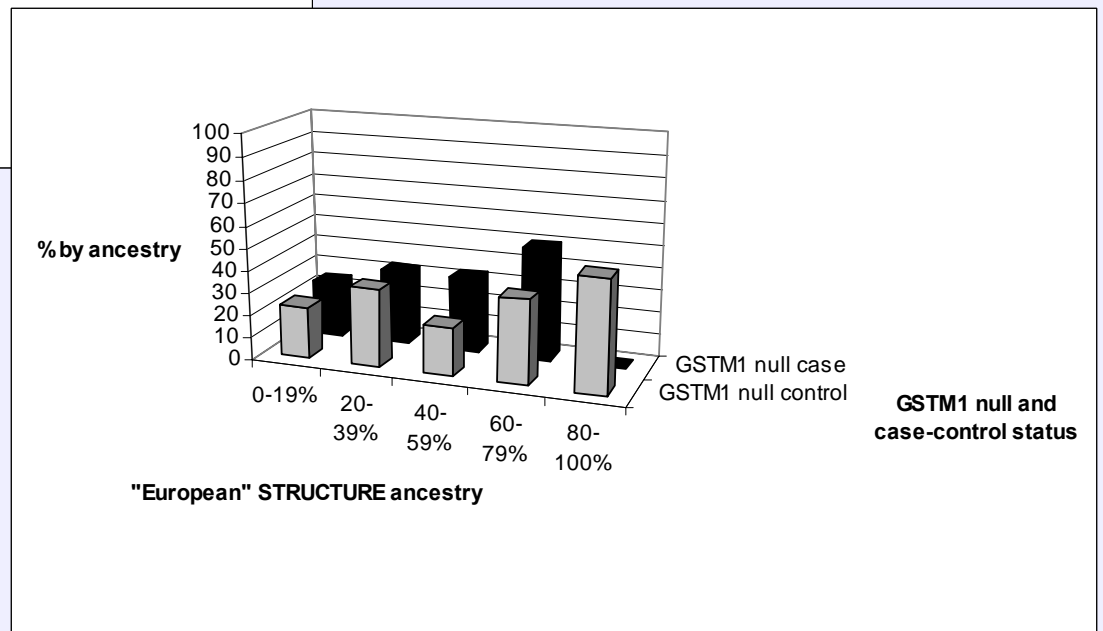
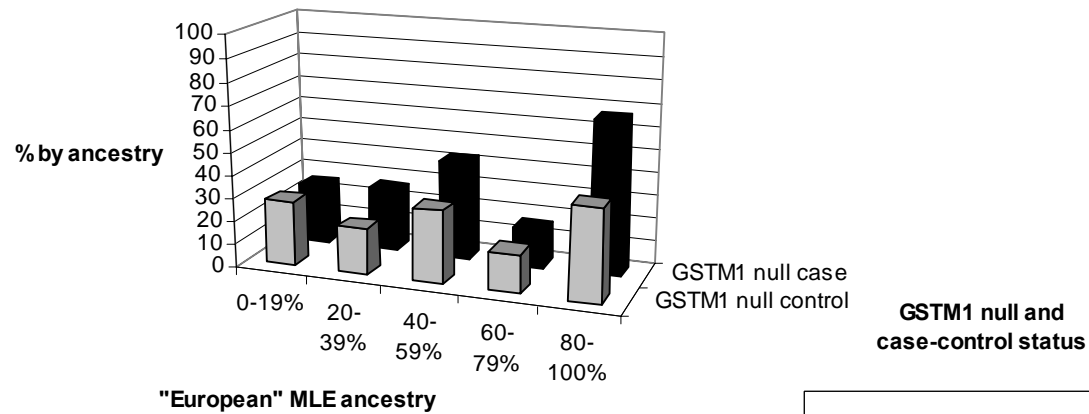
"European" STRUCTURE ancestry

■ European American □ African American

GSTM1 null genotype by case-control status within European ancestry group for European Americans only



GSTM1 null genotype by case-control status within European ancestry group for African Americans only



Logistic regression models comparing adjustments for self-reported race versus individual ancestry

Model	Odds ratio for <i>GSTM1</i> null genotype	95% confidence interval	-2 log likelihood	Number of parameters	LRT chi-square (p-value) ^a	AIC ^b
Base ^c	1.11	(0.77,1.61)	732.58	5	-----	742.58
Base ^c + self-reported race	1.12	(0.77,1.64)	732.47	6	0.11 (0.74)	744.47
Base + MLE ^d	1.13	(0.77,1.65)	722.45	6	10.13 (0.001)	734.45
Base + STRUCTURE ^e	1.26	(0.86,1.84)	704.52	6	28.06 (<0.0001)	716.52

^a LRT=Likelihood Ratio Test comparing all models to the Base model

^b AIC=Akaike's Information Criterion

^c Model is adjusted for age at diagnosis for cases or age at participation in study for controls, packyears of smoking, gender and family history of lung cancer.

^d Model is additionally adjusted for MLE individual "European" ancestry.

^e Model is additionally adjusted for STRUCTURE individual "European" ancestry.

Conclusions

- Significant overlap exists within and between self-reported racial groups by individual ancestry --- significant population structure differences exist that self-reported race alone does not capture.
 - Individual ancestry may confound disease/candidate gene associations and provides a better measure of ancestral background than self-reported race.
-
-

Individual Ancestry Estimation Limitations

1. Choice of markers to use for the ancestry estimation.
 - Need to be informative for ancestry
 - Need to be unlinked to trait
 - Need to be large enough number to make standard error of estimate small
 2. Choice of founding populations to use for analysis - accurate allele frequency sets.
 3. Number of founding population used in analysis.
-
-

Ancestry Informativeness

- Three factors that contribute to the information on ancestry include
 - (1) allele frequency difference between the parental populations (δ)
 - (2) the respective genetic contribution of each founding population to the admixed population (m)
 - (3) parental population allele frequencies, p , (irrespective of δ)

(Pfaff et al, *Genetic Epidemiology* 2004; Barnholtz-Sloan et al, *JFS* 2005)

BOTTOM LINE.....

- Individual ancestry may confound disease/candidate gene associations and provides a better measure of ancestral background than self-reported race.
 - Population stratification and individual ancestry is a controversial subject!
 - Some groups say it does not matter
 - (Wacholder et al, 2002)
 - Others say it does matter and we need to adjust for it
 - (Thomas et al, 2002; Shriver et al, 1998,2000,2003, Kittles et al, 2001,2002,2003, Parra et al, 1997,2001)
 - Can individual ancestry analyses help us gain a better and more complete understanding of groups of people in order to better understand their differences?
-
-

THANK YOU!!!!

