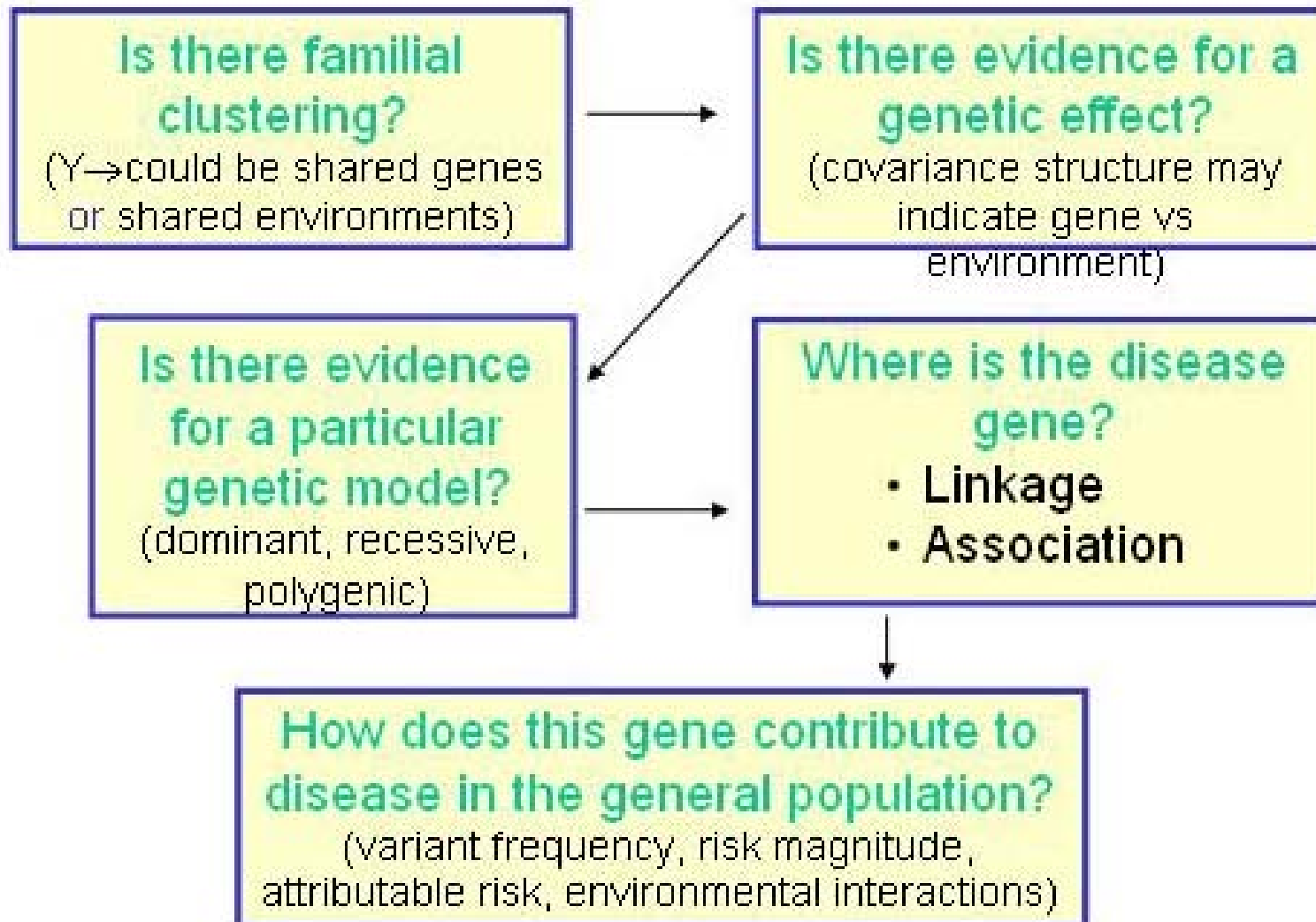


# Population stratification and use of genetic ancestry to assess risk for cancer

Jill Barnholtz-Sloan, PhD  
[jsb42@case.edu](mailto:jsb42@case.edu)

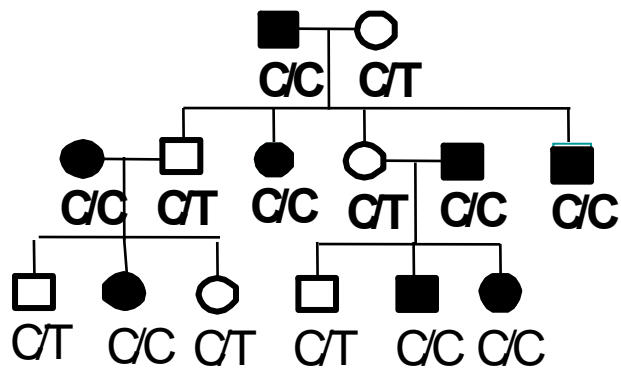
# Genetic Epidemiology Questions



Palmer LJ.  
Webcast

# Human Genetic Analysis

## Families Linkage Studies



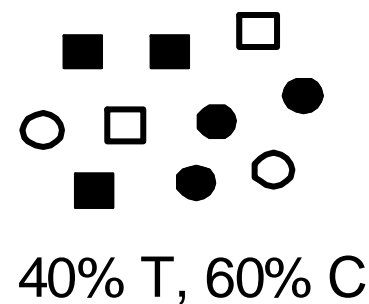
Simple Inheritance (Segregate)

Single Gene with Major Effect

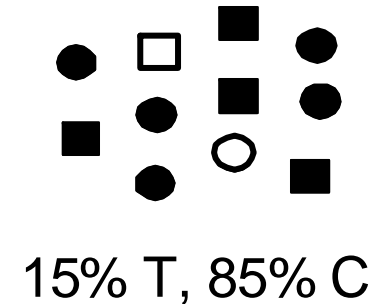
Variant Rare in the Population

~600 Short Tandem Repeat Markers

## Population Association Studies



Cases



Controls

Complex Inheritance (Aggregate)

Multiple Genes with Small Contributions  
and Environmental Contexts

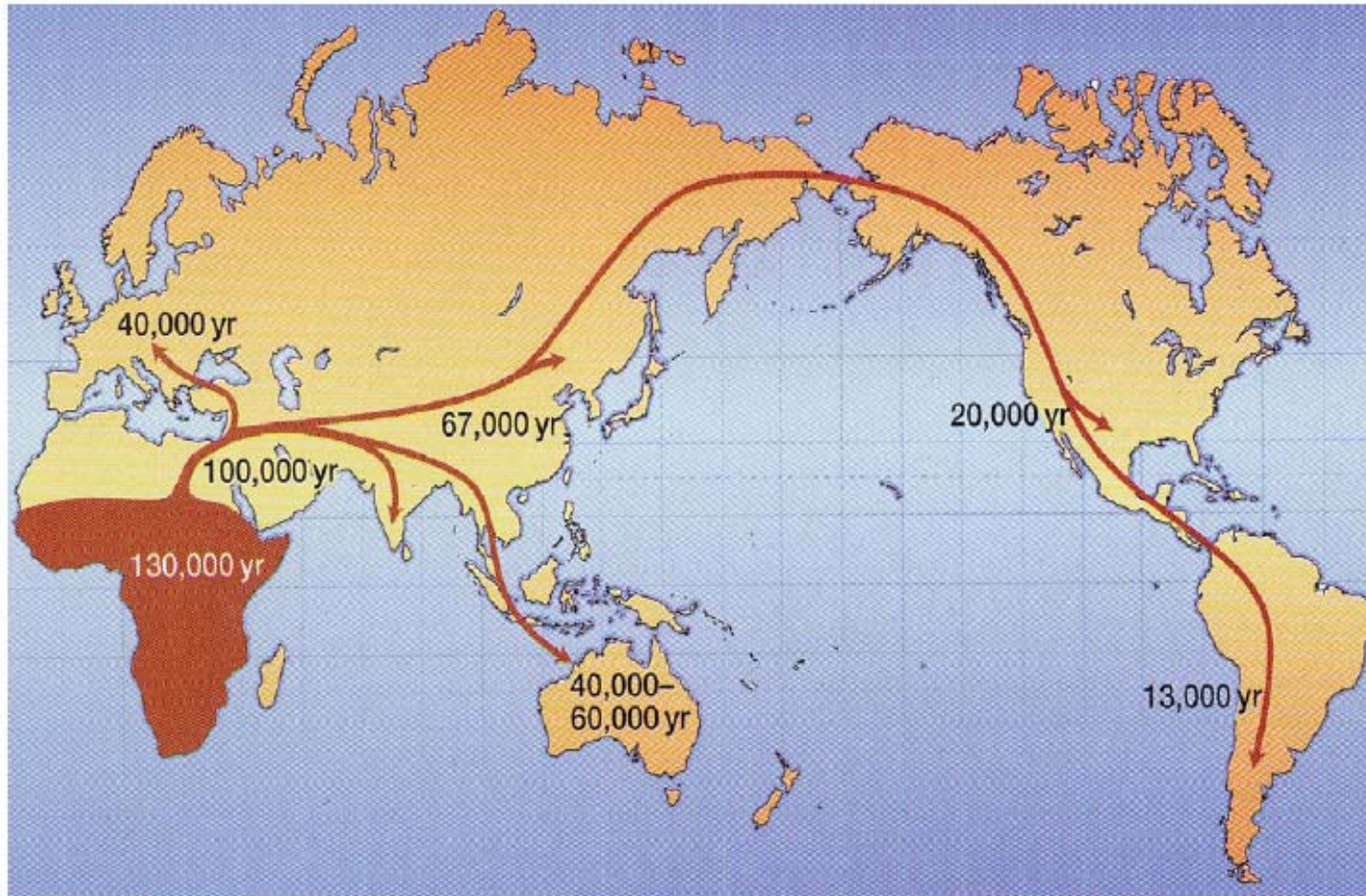
Variant(s) Common in the Population

Polymorphic Markers > 500,000 -1,000,000  
Single Nucleotide Polymorphisms (SNPs)

# 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!!

# Out of Africa?



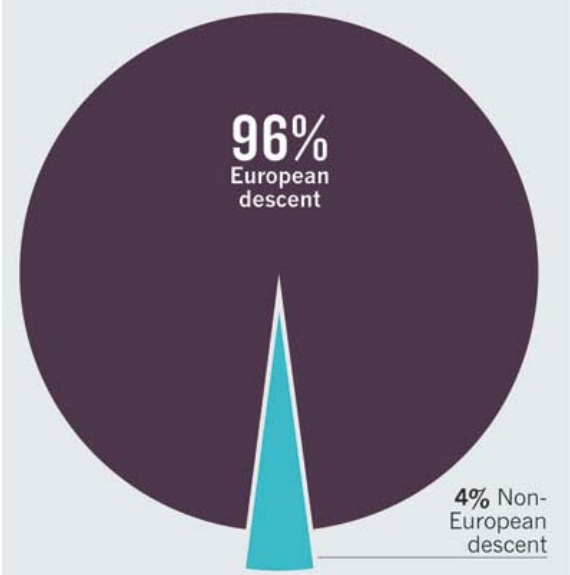
- Modern day humans evolved from one population in Africa

Hedges Nature

# Genomics for the world, Bustamante et al Nature 2011

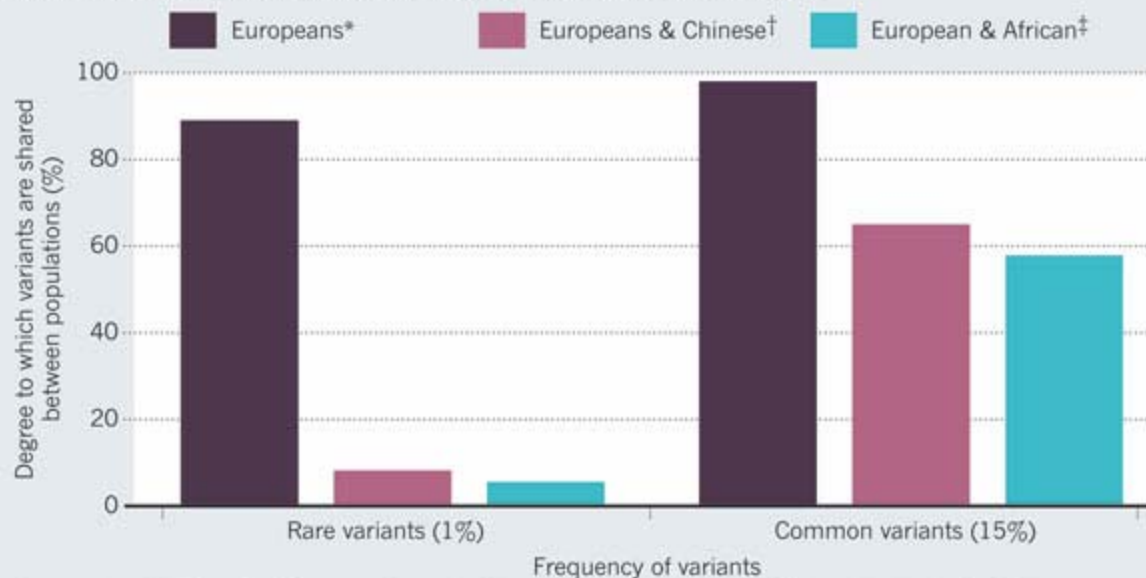
## SAMPLING BIAS

Most genome-wide association studies have been of people of European descent.



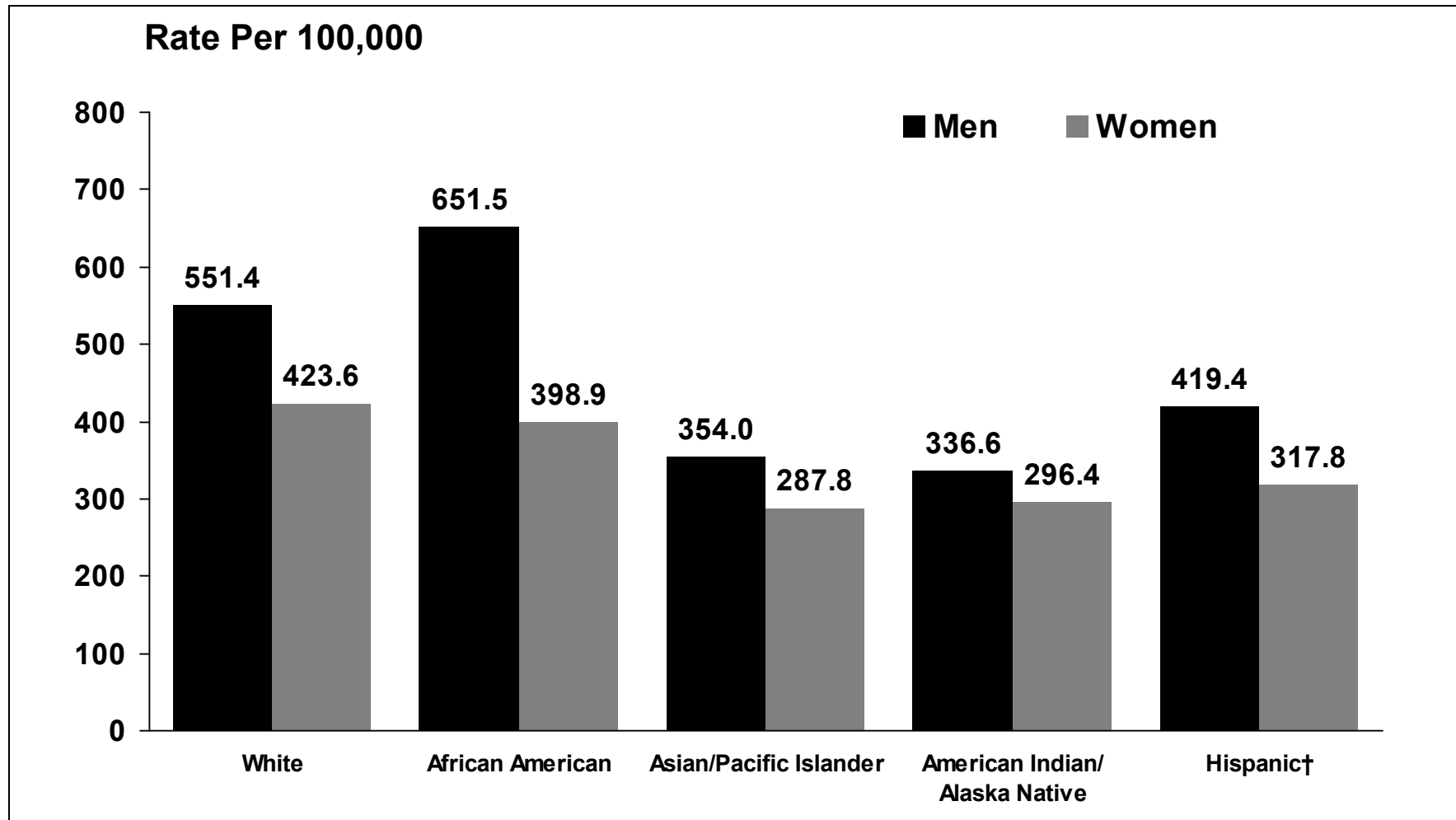
## COMPARING THE UNCOMPARABLE

The rarer a genetic variant is within a population, the less likely it is to be found in all ethnic groups. One hundred people were sampled from each population.



\*Comparison of individuals of European descent in Utah and in Tuscany, Italy. † Han Chinese individuals from Beijing compared with Utah sample ‡ Yoruba individuals from Ibadan, Nigeria, compared with Utah sample.

# US Cancer Incidence Rates\* by Race and Ethnicity, 2001-2005

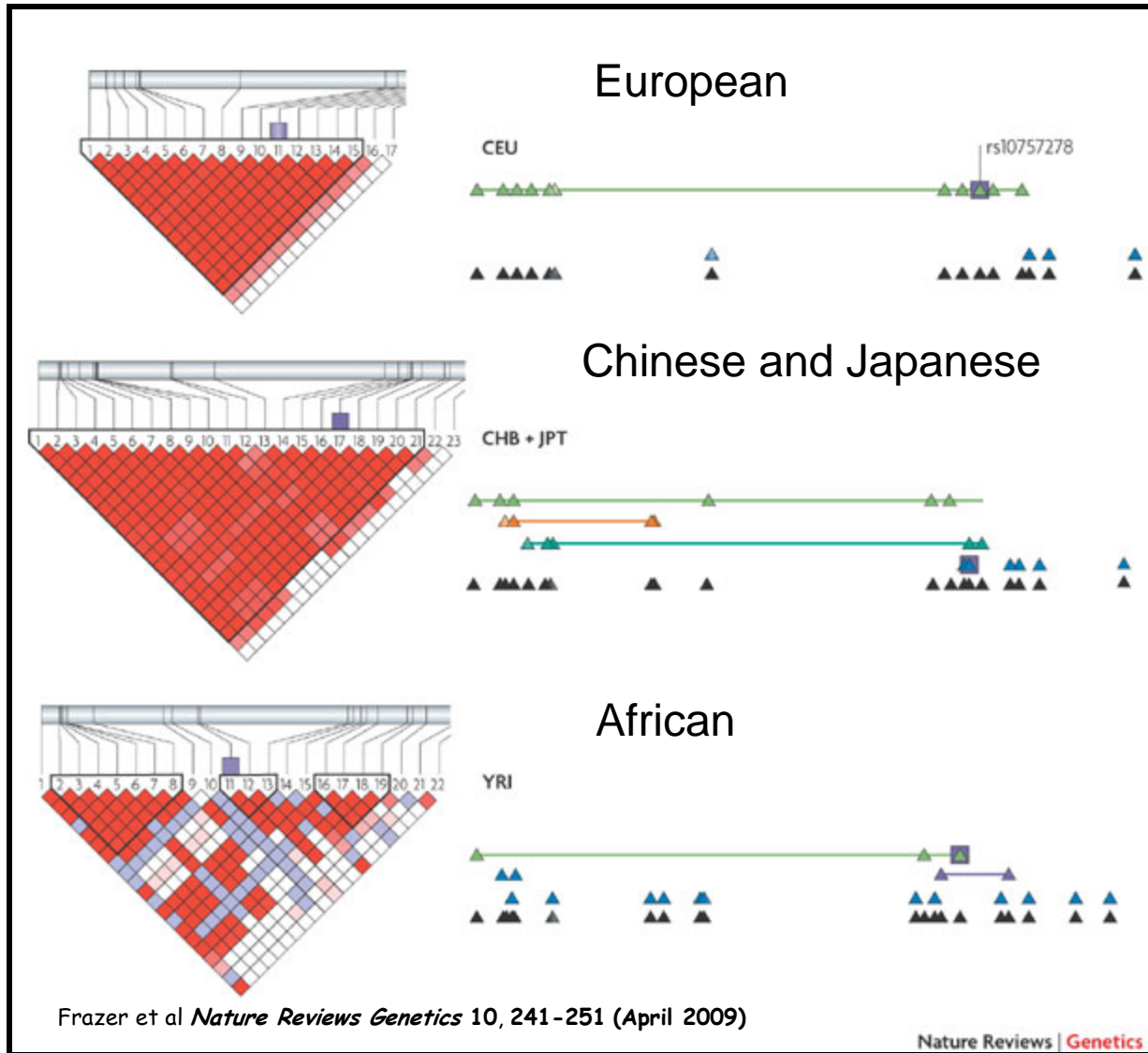


\*Age-adjusted to the 2000 US standard population.

†Person of Hispanic origin may be of any race.

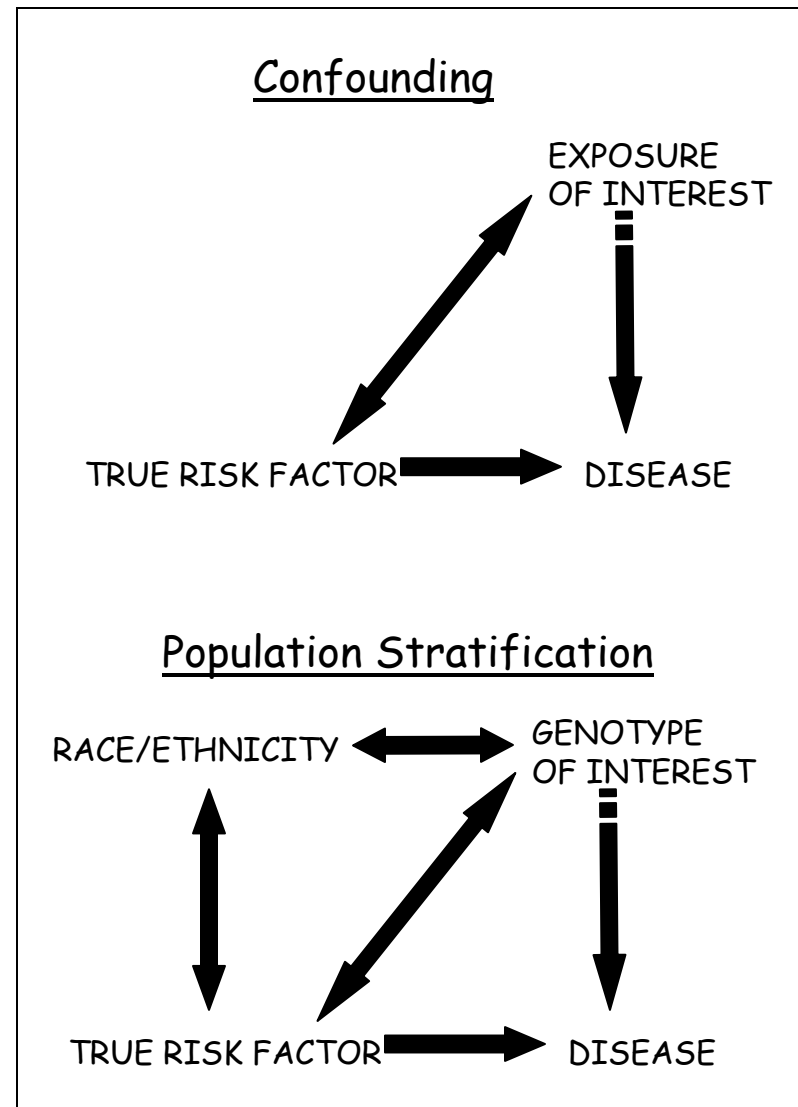
Source: Surveillance, Epidemiology, and End Results Program, 1975-2005, Division of Cancer Control and Population Sciences, National Cancer Institute, 2008.

# LD blocks by ancestral group

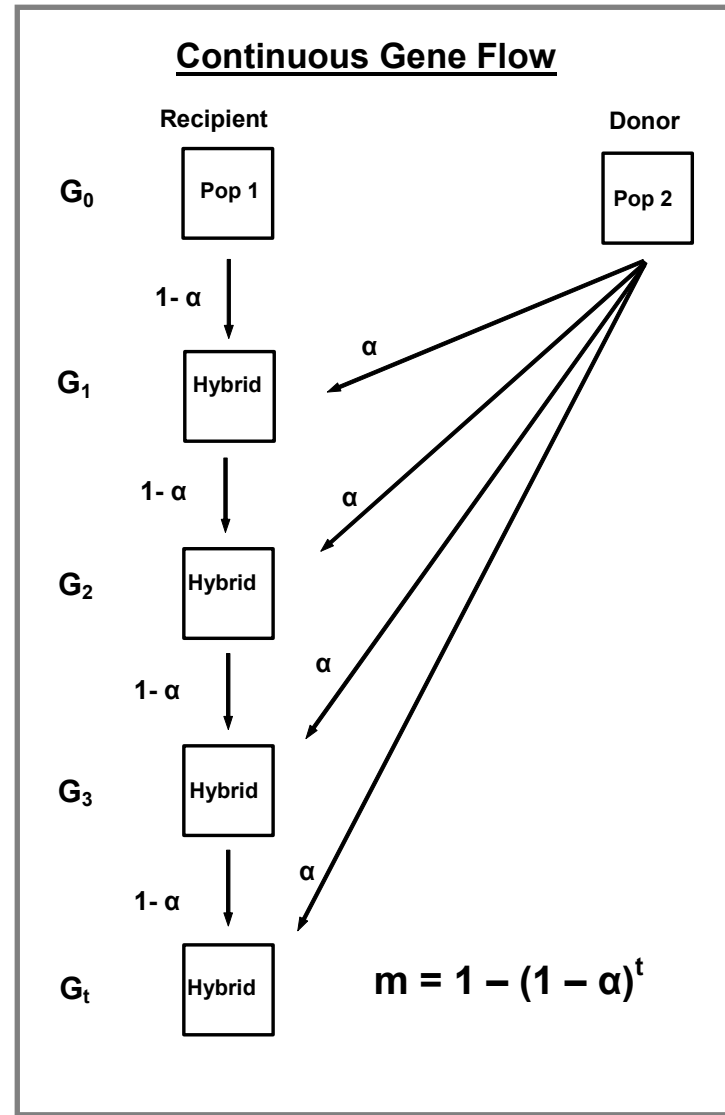
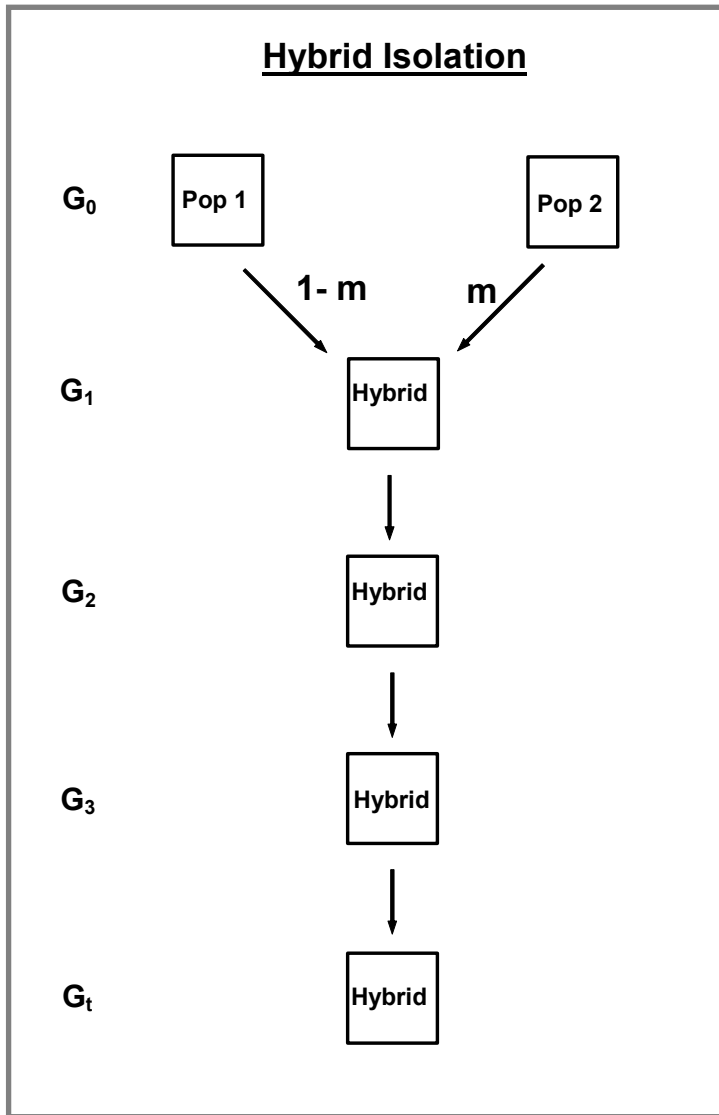


# Population Stratification (PS)

- Frequency of marker genotype varies by race/ethnicity
- Background disease prevalence varies by race/ethnicity
- Race/ethnicity acts as a surrogate for the true risk factor.



# Two Models for Genetic/Racial Admixture





# Consequences of Population Stratification

- Potential for increased allelic associations (Linkage Disequilibrium - LD)
- Potential for deviations from Hardy-Weinberg Equilibrium (HWE)
- Can induce both false+ and false- associations
- *CAUSING INCONSISTENCY ACROSS STUDIES?*

# Types of Markers for assessing Population Stratification

- Random markers across the genome
  - Non-candidate gene markers throughout the genome
  - Unlinked to each other and to disease
- Ancestry informative markers (AIMs)
  - Markers showing large allele frequency differences between ancestral groups,  $\delta$
  - Shriver et al., 1997 first to come up with a panel of ancestry informative markers (~40 SNPs)
  - Now others have published general panels of markers for ancestry analyses in specific populations
    - HapMap Project ([www.hapmap.org](http://www.hapmap.org)); 1000 genomes project
    - African Americans, Hispanic Americans, Europeans

# Ancestry Estimation

## 1. MLE

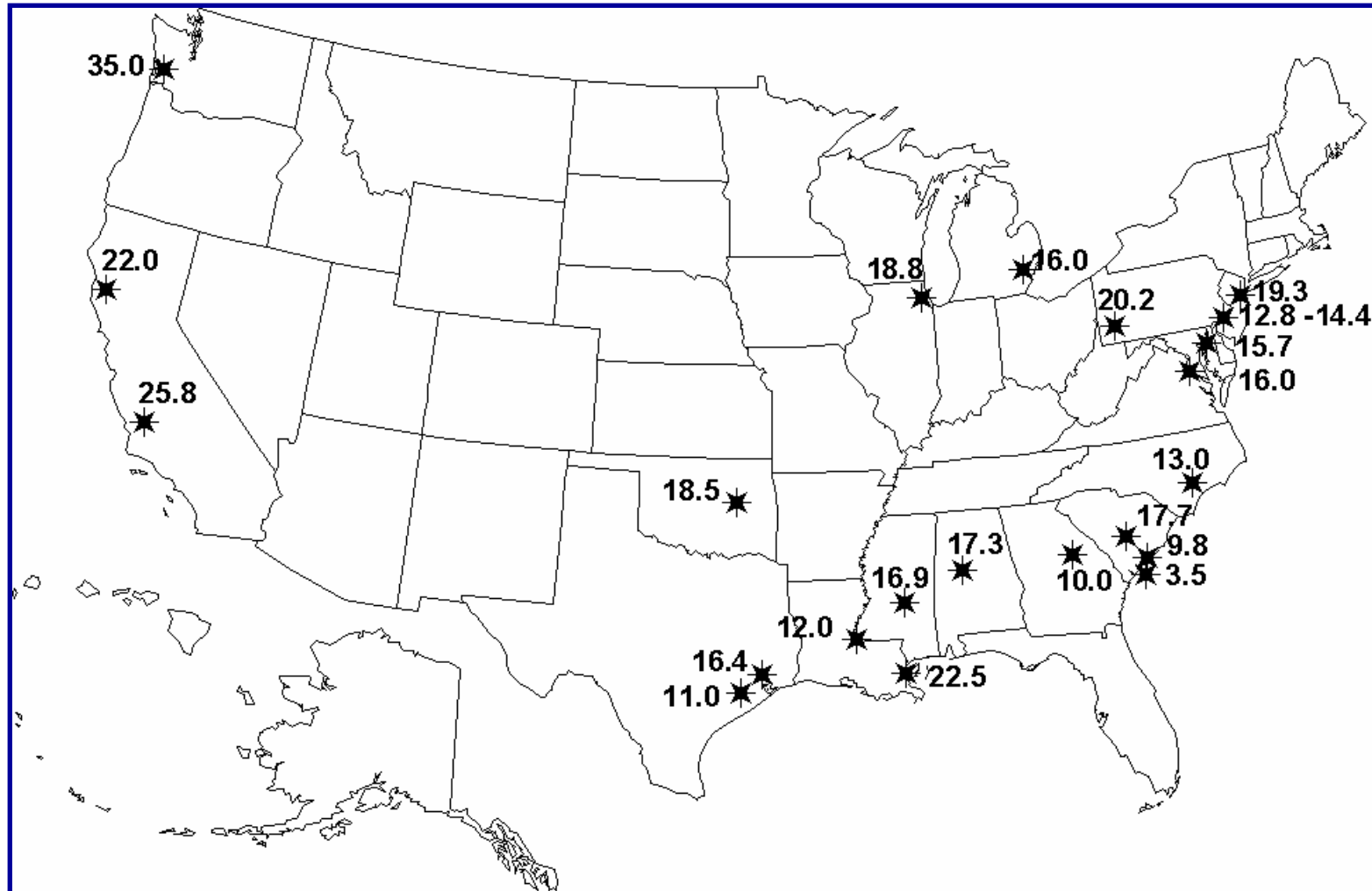
- Need genotype data on individuals
- Need ancestral allele frequencies on all markers

## 2. Bayesian techniques – structured association

- Implemented in STRUCTURE and ADMIXMAP
- Need genotype data on individuals
- DO NOT need marker ancestral allele frequencies
- Need ancestral genotypes

- Comparable in terms of accuracy
- Validity dependent on informativeness of markers used

# European ancestry proportions in African Americans

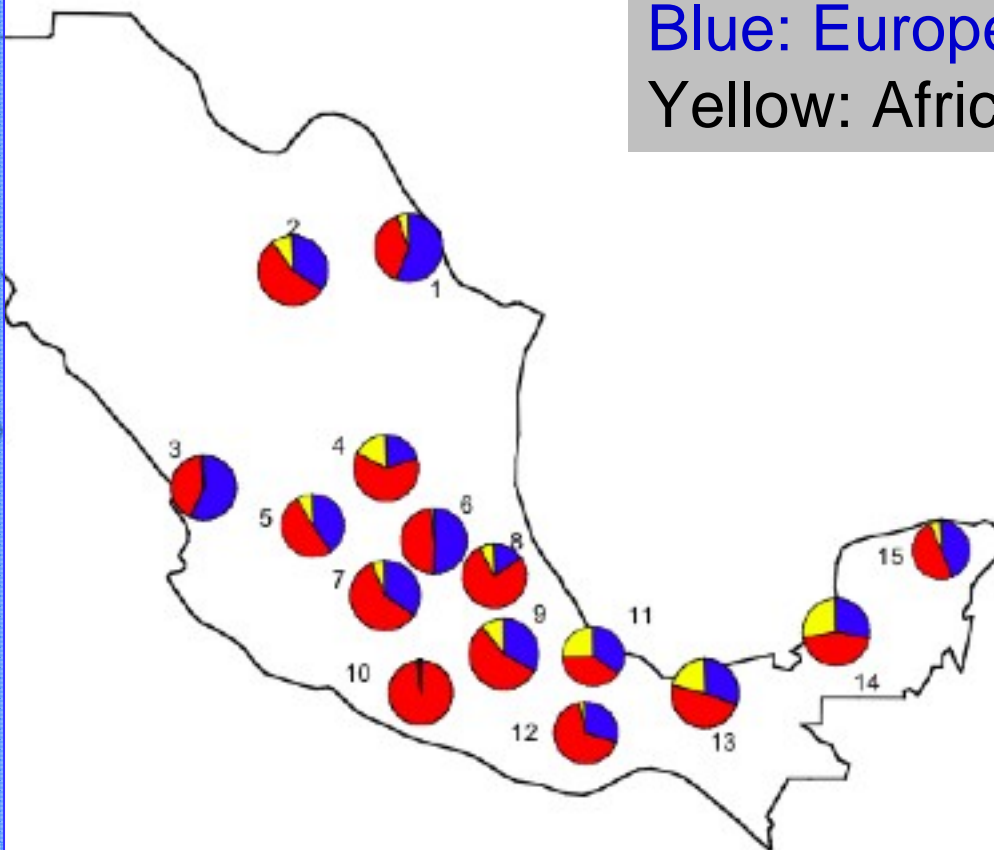


Parra et al. AJHG 1998; Parra et al. AJPA 2002; Kittles et al. unpublished

# Heterogeneity in Mexican Admixture

- 1: Monterrey, Nuevo León
- 2: Saltillo, Coahuila
- 3: Guadalajara, Jalisco
- 4: Cuernavaca, 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., 2008)

- Latina women with breast cancer; 440 cases and 597 controls
- 106 AIMs; MLE individual ancestry estimation
- The OR for a 25% increase in European ancestry was 1.79
  - After adjustment for known risk factors and place of birth the OR was 1.39

**Table 2.** Multivariate logistic regression model of association between genetic ancestry and breast cancer risk ( $n = 975$ )

	OR (95% CI)	$P >  z $
<b>Univariate analysis</b>		
European ancestry*	1.79 (1.28–2.79)	<0.001
<b>Multivariate analysis</b>		
European ancestry	1.39 (1.06–2.11)	0.013
Age at diagnosis	1.02 (1.01–1.04)	0.013
Foreign born	0.73 (0.54–0.99)	0.046
Family history of breast cancer	1.34 (0.88–2.04)	0.160
Benign breast disease	1.12 (0.77–1.59)	0.580
Age at menarche	0.93 (0.86–1.01)	0.074
Hormone replacement therapy use	0.92 (0.68–1.24)	0.570
Daily alcohol intake <sup>†</sup>	1.98 (1.21–3.24)	0.006
Ln daily kilocalorie intake <sup>‡</sup>	1.78 (1.24–2.42)	0.001
Parity	0.86 (0.80–0.94)	<0.001
Breast-feeding per child	0.97 (0.95–1.00)	0.070
Education level	1.11 (0.96–1.28)	0.131

NOTE: Thirty-two cases and 25 controls were excluded from the analysis because of missing data.

\*OR is for every 25% increase in European ancestry.

†Daily intake of >10 versus ≤10 g.

‡Individuals with daily kilocalorie intake of <600 or >5,000 were excluded from the analysis. Daily kilocalorie intake was log transformed for analysis.

# CYP3A4 and prostate cancer

(Walker et al, 1998; Zeigler-Johnson, 2001; Kittles et al, 2002)

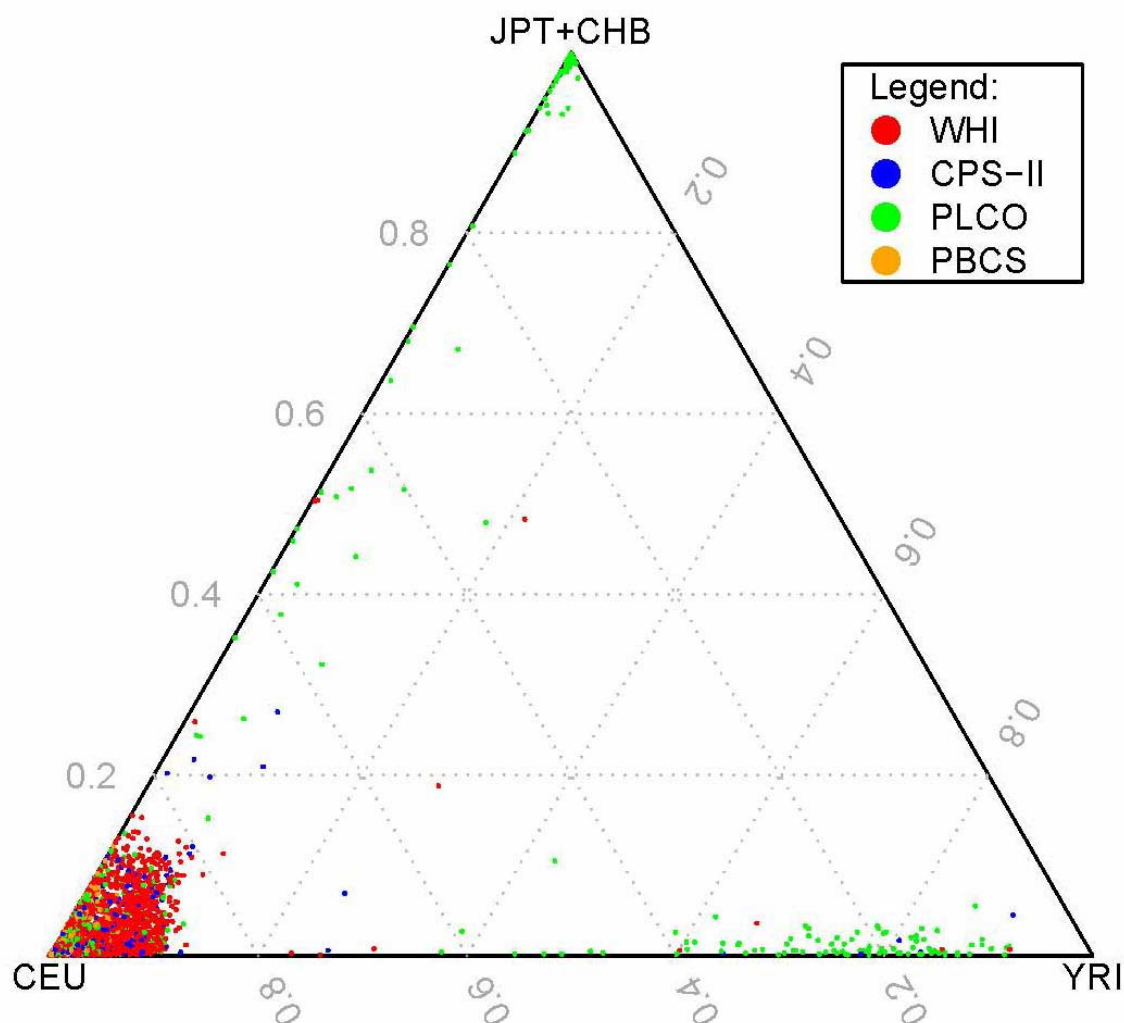
- **CYP3A4-V**, an A to G promoter variant, is associated with Prostate CA in AA
  - metabolism of steroids
- High frequency of this variant in AA compared to Caucasians and Taiwanese.
- Sharp differences in CYP3A4-V frequencies were seen between Nigerian and EA controls (0.87 vs. 0.10), with AA in between (0.66).
- Significant population stratification was found in AA, using unlinked markers to estimate ancestry.
  - Association no longer significant in AA for Prostate Cancer and CYP3A4-V when adjusted for ancestry.

# 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)

# GWAS study of Breast Cancer (Thomas et al, 2009)



• **STRUCTURE**  
program used to  
estimate indiv 3-  
way ancestry

**Removed indivs  
with >20%  
ancestry other  
than European**

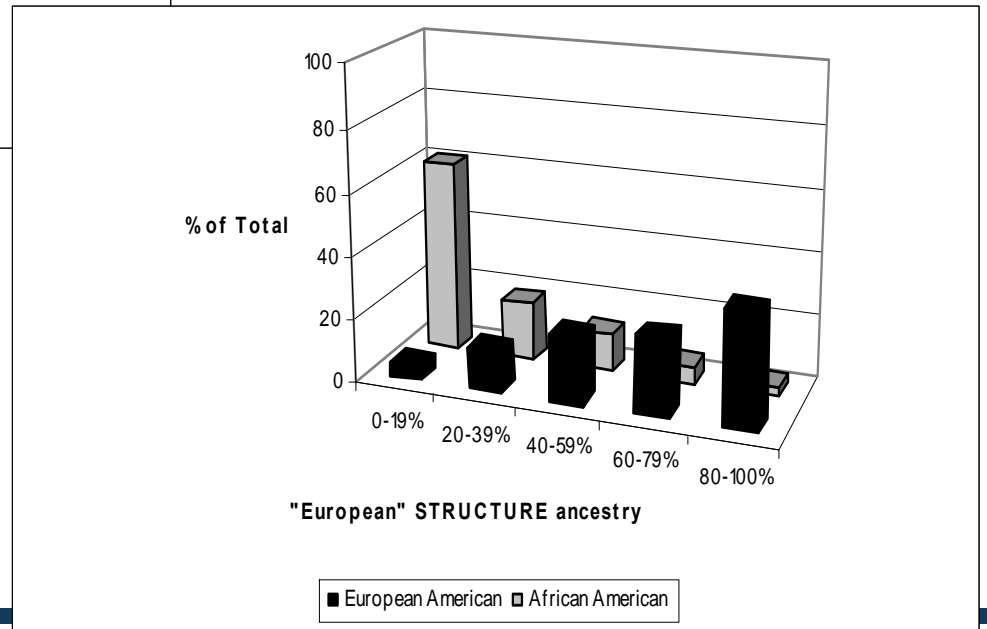
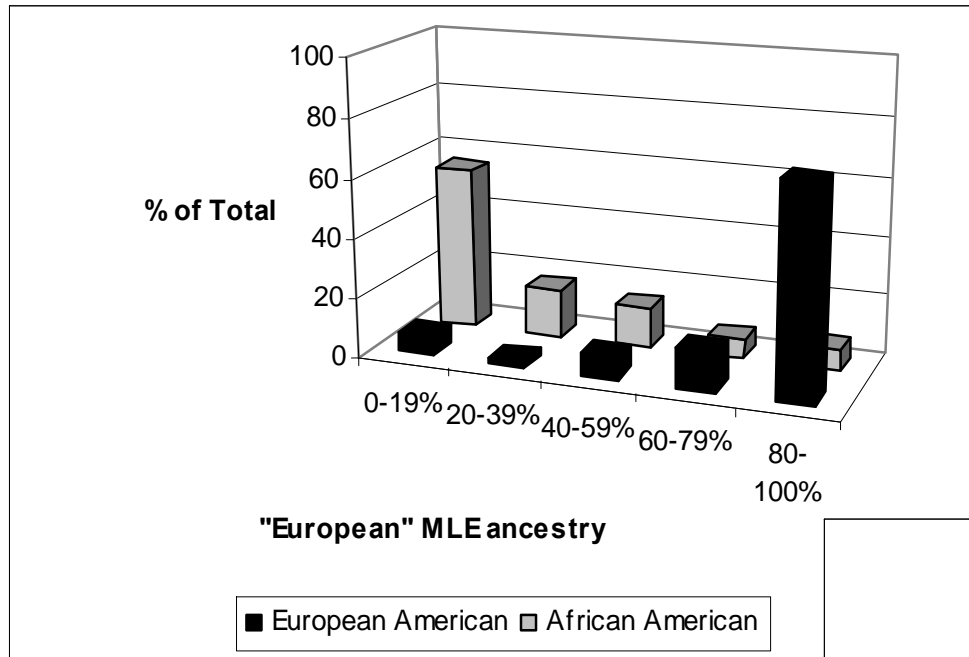
# GWAS study of Breast Cancer (Thomas et al, 2009)

Table viii lists subjects excluded from association analysis due to estimated non-CEU admixture.

**Table viii. Admixture analysis exclusions**

Study	Imputed race	Self-reported race								
		American Indian	Asian	Black	Black, non-Hispanic	Hispanic	Pacific islander	White	White, non-Hispanic	White/Hispanic
CPSII	ADMIXED:CEU									1
	CEU,JPTCHB					4				
	CEU,YRI			4						
	JPTCHB		4							
	YRI			3						
PLCO	ADMIXED:CEU					1				
	ADMIXED:YRI				6					
	CEU,JPTCHB	2	6			7	7			
	CEU,YRI				50					
	JPTCHB		72				2			
	YRI				30	1			1	
WHI	ADMIXED:CEU							1		
	CEU,JPTCHB							3		
	CEU,JPTCHB,YRI							1		
	CEU,YRI							6		
	YRI							2		
<b>Total</b>		2	82	7	86	13	9	13	1	1

# Individual European Ancestry in Early-onset lung cancer cases and controls from Detroit, Michigan (BARNHOLTZ-SLOAN ET AL., 2005)



## Comparing estimates of early-onset lung cancer risk adjusting for self-reported race vs. individual ancestry (BARNHOLTZ-SLOAN ET AL., 2005)

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

<sup>a</sup> LRT=Likelihood Ratio Test comparing all models to the Base model

<sup>b</sup> AIC=Akaike's Information Criterion

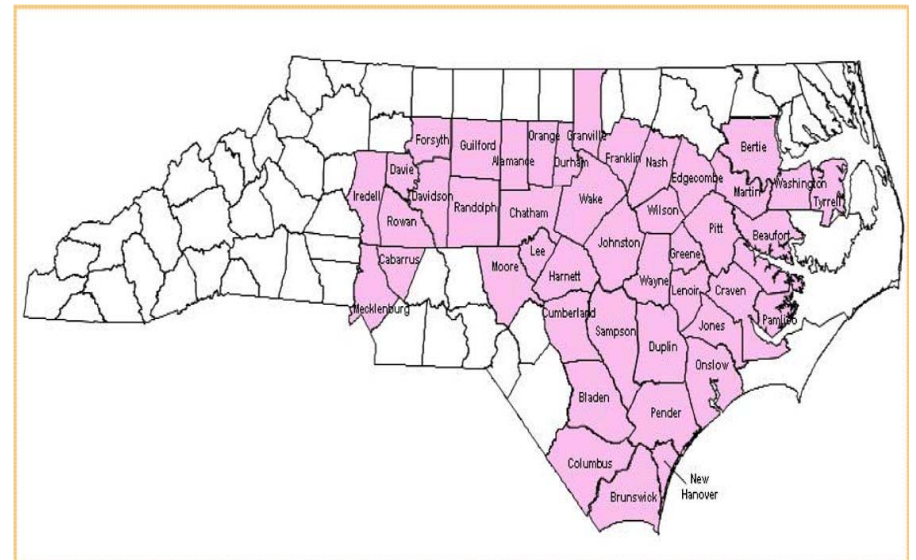
<sup>c</sup> Model is adjusted for age at diagnosis for cases or age at participation in study for controls, packyears of smoking, gender and family history of lung cancer.

<sup>d</sup> Model is additionally adjusted for MLE individual "European" ancestry.

<sup>e</sup> Model is additionally adjusted for STRUCTURE individual "European" ancestry.

# Carolina Breast Cancer Study (CBCS)

- Examining causes of breast cancer in White and Black women in North Carolina
  - Also investigating survival
- Began in 1993
- Cases – identified through the NC Cancer Registry
  - All newly diagnosed women with breast cancer
- Controls – population-based; RDD; frequency matched on age and race

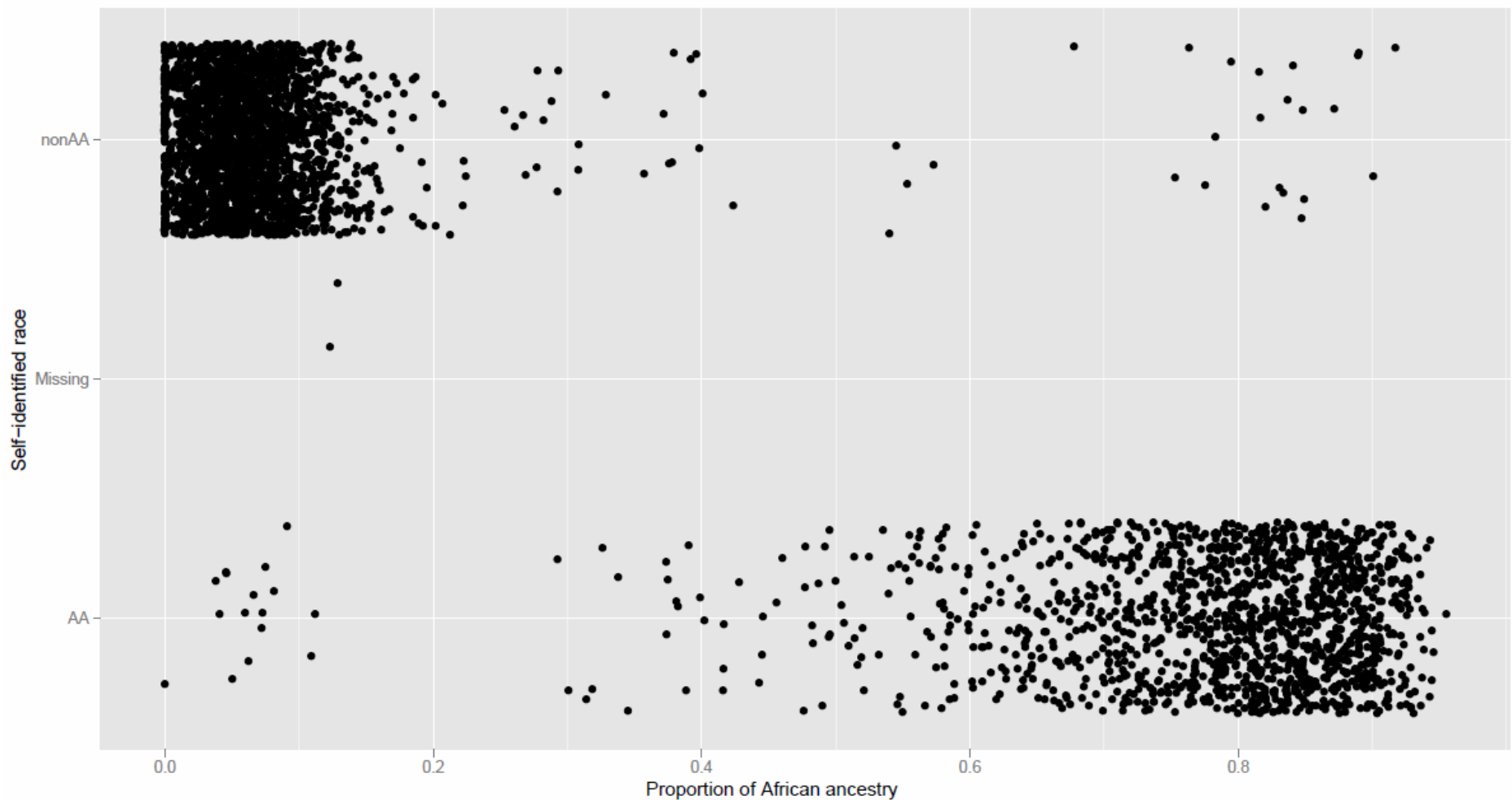


# Summary information about CBCS study

	Cases	Controls
non-Black	1230	1118
White	1204	1089
Indian	8	11
Asian	12	6
Other	6	12
Black	742	658
Average Age	51.97	52.53
Average European	0.66	0.67
Average African	0.34	0.33

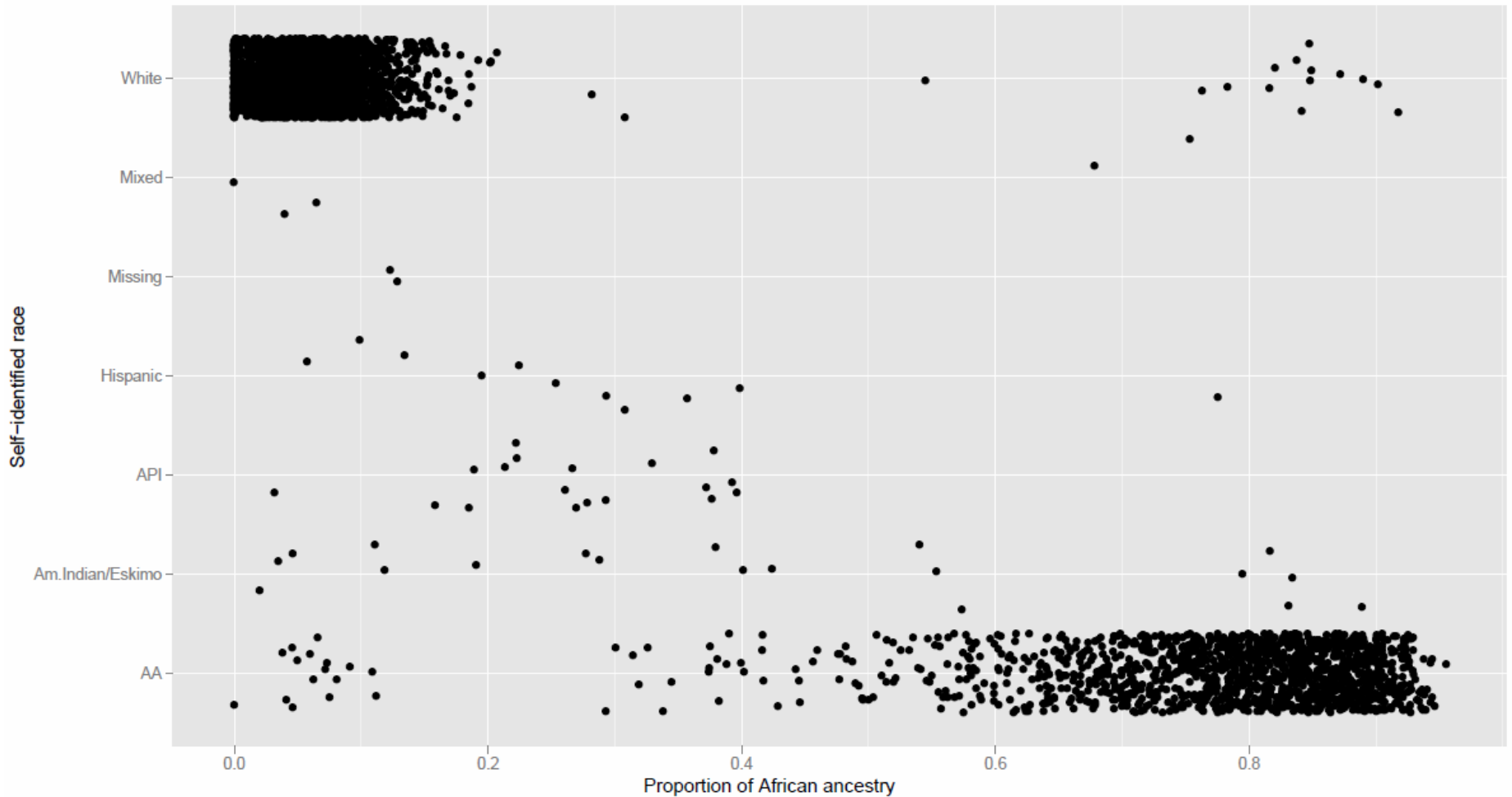
# MLE African Ancestry by Black or non-Black

CBCS: MLE African ancestry by self-identified race



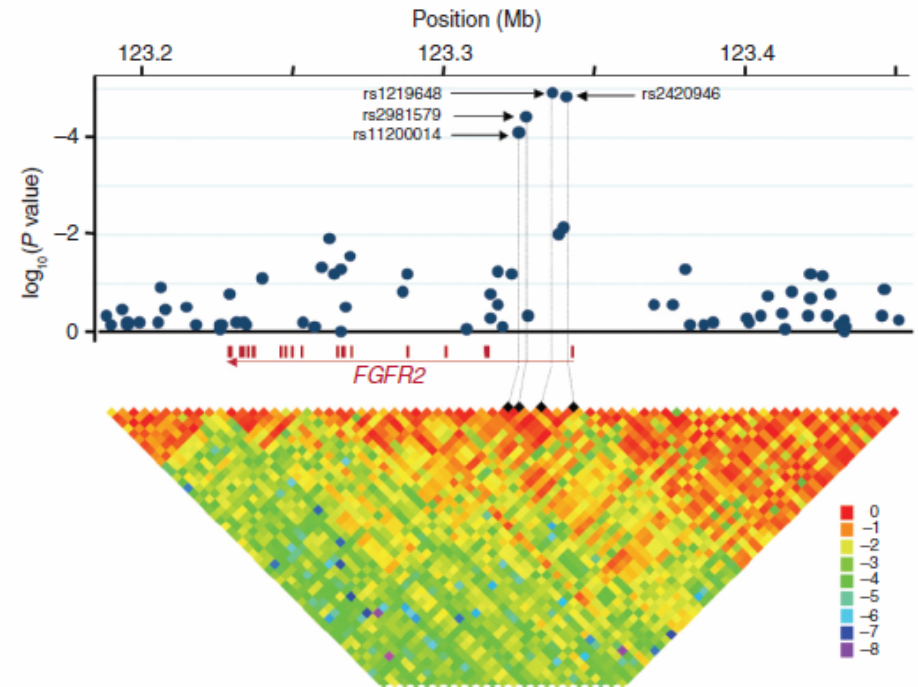
# MLE African ancestry by detailed race

CBCS: MLE African ancestry by self-identified race



# FGFR2 and breast cancer risk

- FGFR2 -- tumor suppressor gene
  - amplified and overexpressed in breast cancer.
- Alternatively spliced variants result in differential signal transduction and transformation of mammary epithelial cell lines.
- Located on Chromosome 10
- SNPs in this gene associated with breast cancer risk



# Hunter et al, Nature Genetics 2007

**Table 2 Haplotypes for four SNPs in intron 2 of *FGFR2* and association with breast cancer risk**

Haplotype <sup>a</sup>	Cases	Controls	OR	95% c.i.	P
Nurses' Health Study					
G-G-A-C	1,195	1,348	1.0		
A-A-G-T	998	842	1.33	1.18-1.50	3 × 10 <sup>-6</sup>
A-A-A-C	40	47	0.96	0.62-1.48	0.85
A-A-G-C	24	19	1.40	0.76-2.57	0.28
Rare <1%	33	28	1.36	0.81-2.29	0.24
Nurses' Health Study 2					
G-G-A-C	295	667	1.0		
A-A-G-T	276	474	1.34	1.09-1.65	0.0061
A-A-A-C	7	21	0.76	0.30-1.90	0.55
A-A-G-C	13	20	1.50	0.70-3.21	0.30
Rare <1%	12	16	1.96	0.90-4.28	0.09
PLCO study					
G-G-A-C	994	1,140	1.0		
A-A-G-T	795	728	1.13	0.99-1.29	0.064
A-A-A-C	32	24	1.44	0.83-2.47	0.19
A-A-G-C	11	24	0.47	0.23-0.97	0.041
Rare <1%	21	34	0.63	0.36-1.10	0.11
ACS CPS-II					
G-G-A-C	583	664	1.0		
A-A-G-T	482	406	1.38	1.16-1.65	0.00040
A-A-A-C	21	22	1.07	0.59-1.94	0.82
A-A-G-C	6	10	0.67	0.23-1.89	0.45
Rare <1%	18	10	1.98	0.91-4.31	0.084
Pooled across studies					
G-G-A-C	3,068	3,718	1.0		
A-A-G-T	2,551	2,450	1.26	1.17-1.35	6 × 10 <sup>-10</sup>
A-A-A-C	102	114	1.09	0.83-1.43	0.55
A-A-G-C	54	74	0.88	0.61-1.27	0.50
Rare <1%	107	151	0.87	0.68-1.12	0.28

<sup>a</sup>For each study, haplotypes are indicated (from top to bottom) for SNPs rs11200014, rs2981579, rs1219648 and rs2420946, respectively.

# CBCS Haplotype frequencies – Overall and by Race (SAS Genetics)

		CEU	YRI
rs11200014	A	0.42	0.58
	G	0.23	0.77
rs2981579	C	0.58	0.42
	T	0.33	0.67
rs1219648	A	0.58	0.42
	G	0.53	0.47
rs2420946	C	0.59	0.41
	T	0.38	0.62

	Haplotype	Frequency
<b>Overall</b>	G-C-A-C (Hunter Reference)	48.2%
	A-T-G-T (Hunter Risk)	30.7%
	G-T-G-T	8.9%
	G-T-A-C	3.0%
	G-T-A-T	3.5%
	A-T-A-C	2.3%
	G-C-A-T	1.1%
	A-T-G-C	1.0%
	G-C-G-T	1.0%
<b>Black</b>	G-C-A-C (Hunter Reference)	35.4%
	A-T-G-T (Hunter Risk)	15.8%
	G-T-G-T	23.1%
	G-T-A-C	7.4%
	G-T-A-T	8.9%
	A-T-A-C	3.2%
	G-C-A-T	2.2%
	A-T-G-C	1.0%
	G-C-G-T	2.4%
<b>White</b>	G-C-A-C (Hunter Reference)	55.7%
	A-T-G-T (Hunter Risk)	39.4%
	G-T-G-T	0.6%
	G-T-A-C	0.2%
	G-T-A-T	0.2%
	A-T-A-C	2.0%
	G-C-A-T	0.4%
	A-T-G-C	1.1%
	G-C-G-T	0.3%

Barnholtz-Sloan et al, 2010

# Haplotype analysis – overall

(haplo.stats)

Gene	Haplotype	Adjusted OR (95% CI) for race and age	p-value (T-statistic) adjusted for race and age	Adjusted OR (95% CI) for race, European ancestry and age	p-value (T-statistic) adjusted for race, European ancestry and age
FGFR2	GCAC	Reference	Reference	Reference	Reference
	ATGT	<b>1.25 (1.13, 1.39)</b>	<b>0.000024</b>	<b>1.25 (1.15, 1.36)</b>	<b>0.0000305</b>
	GTGT	<b>1.37 (1.13, 1.65)</b>	<b>0.00104</b>	<b>1.38 (1.19, 1.57)</b>	<b>0.000775</b>
	GTAC	1.10 (0.82, 1.48)	0.515	1.12 (0.83, 1.42)	0.444
	GTAT	1.28 (0.98, 1.68)	0.073	<b>1.30 (1.03, 1.58)</b>	0.060
	ATAC	1.26 (0.91, 1.73)	0.164	1.26 (0.94, 1.58)	0.156
	GCAT	0.75 (0.46, 1.21)	0.232	0.75 (0.27, 1.23)	0.237
	ATGC	1.25 (0.77, 2.02)	0.368	1.25 (0.77, 1.73)	0.362
	GCGT	1.19 (0.75, 1.91)	0.463	1.21 (0.74, 1.68)	0.432

Barnholtz-Sloan et al, 2010

# Haplotype analysis by race

(haplo.stats)

Dataset	Gene	Haplotype	Adjusted OR (95% CI) for race and age	p-value (T-statistic) adjusted for race and age	Adjusted OR (95% CI) for race, European ancestry and age	p-value (T-statistic) adjusted for race, European ancestry and age
AA	<i>FGFR2</i>	GCAC	Reference	Reference	Reference	Reference
		ATGT	1.11 (0.88, 1.39)	0.38	1.11 (0.88, 1.39)	0.38
		GTGT	<b>1.28 (1.04, 1.57)</b>	<b>0.0191</b>	<b>1.27 (1.04, 1.56)</b>	<b>0.0205</b>
		GTAC	1.08 (0.78, 1.49)	0.65	1.08 (0.78, 1.49)	0.66
		GTAT	1.25 (0.94, 1.67)	0.13	1.25 (0.93, 1.67)	0.14
		ATAC	1.11 (0.68, 1.80)	0.68	1.11 (0.68, 1.80)	0.68
		GCAT	0.77 (0.43, 1.38)	0.38	0.77 (0.43, 1.38)	0.38
		ATGC	.	.	.	.
		GCGT	1.11 (0.65, 1.89)	0.71	1.10 (0.65, 1.89)	0.72
White	<i>FGFR2</i>	GCAC	Reference	Reference	Reference	Reference
		ATGT	<b>1.30 (1.16, 1.47)</b>	<b>0.0000155</b>	<b>1.30 (1.15, 1.46)</b>	<b>0.0000191</b>
		GTGT	.	.	.	.
		GTAC	.	.	.	.
		GTAT	.	.	.	.
		ATAC	1.31 (0.87, 1.99)	0.2	1.32 (0.87, 2.01)	0.19
		GCAT	.	.	.	.
		ATGC	0.99 (0.57, 1.74)	0.98	1.00 (0.57, 1.75)	1.00
		GCGT	.	.	.	.

Barnholtz-Sloan et al, 2010

# Limitations to ancestry/PS analysis

1. Choice of markers to use for the ancestry estimation
  - Need to be informative for ancestry
  - Need to be large enough number to make standard error of estimate small
  - How many?
2. Accurate ancestral allele frequency sets
3. Number of ancestry groups used in analysis
  - Knowledge of migration and immigration history of study population
4. Choice of statistical analysis tool(s)

# Bottom line.....

- PS may confound disease/candidate gene associations in minority and non-minority populations
- Ancestry based analyses becoming more commonly used in case-control studies
- Testing and adjusting for PS now “standard” in most large-scale association studies

# Acknowledgements

## Collaborators

- Karmanos – Ann Schwartz, PhD
- Moffitt – Thomas Sellers, PhD
- U Cincinnati – Ranajit Chakraborty, PhD
- UNC – Robert Millikan, PhD; Sarah Nyante
- U Pennsylvania – Tim Rebbeck, PhD; Klara Stefflova, PhD
- Pennsylvania State U – Mark Shriver, PhD
- U Michigan – Jeff Long, PhD
- CWRU – Priya Shetty, Xiaowei Guan

## Funding – US National Cancer Institute



**Aces**  
African American  
Cancer Epidemiology Study

# Rationale

- African-Americans have lower incidence rates but poorer 5-year survival rates compared to whites
  - Incidence rates:
    - 14.1/100,000 in whites
    - 10.1/100,000 in African-Americans
  - 5-year survival rates:
    - 37% in 1975-77 to 45% in 1996-2003 for whites
    - 43% in 1975-77 to 38% in 1996-2003 for African-Americans
  - Change in incidence rates 1975 - 2005 was larger for whites (-2.7%) than African Americans (-0.4%)
  - African Americans are diagnosed on average three years earlier than whites, 61 versus 64 years

# Specific Aims

- To recruit African-American women with invasive epithelial ovarian cancer and controls.
- To determine risk factors for ovarian cancer in African-American women.
- To evaluate genetic associations for ovarian cancer in African-American women, focusing on results in white women in studies from large consortia or GWAS.
- To study predictors of survival among African-American ovarian cancer patients.

# Study Design

- Population-based, case-control study (enrollment 1/1/2011)
  - 1000 cases through rapid case ascertainment (physician consent)
  - 1000 age-matched controls using random digit dialing
  - Telephone interview
  - Blood draw and paraffin-embedded tumor tissue
  - Multicenter Study in 9 Geographic Regions:
    - North Carolina
    - South Carolina
    - Georgia
    - Alabama
    - Louisiana
    - Ohio
    - Detroit
    - Illinois
    - New Jersey

# Questionnaire

- Socio-Demographics
- Reproductively history
- Contraception and hormone use
- Medical history
- Symptoms
- Medication Use
- Radiation exposure
- Family history
- Sun exposure
- Talc
- Height and weight
- Physical activity
- Insurance, access to care
- Trust in physicians
- Social Support
- Perceived discrimination
- Religiosity and spirituality
- Cultural and folk beliefs
- Smoking
- Food Frequency Questionnaire

# Overall Accrual To Date

October 27, 2011

## Enrollment Summary for 9 States

- 94 Total cases from all sites
- 232 Total Controls from all states

## Ineligible/refusals for cases

- 4 MD Refusals
- 3 unable to contact
- 3 Ineligible (does not self-identify)
- 12 refusals
- 8 Deceased

## Ineligible/refusals for controls

- 12 unable to contact
- 5 Ineligible (does not self-identify)
- 13 refusals

# Ohio RCA Hospital Network

Enrolling cases from 22/30 hospitals with IRB approval

11- Eligible cases to date

3 - Interviewed to date

Ineligible/refusals

1- MD refusal

•Ohio RCA Network will include a total of 30 hospitals from 8 different counties

# Ohio RCA Hospital Network

- **University Hospitals Case Medical Center (2 hospitals)**
  - Investigator: Jill S. Barnholtz-Sloan
  - Co-Investigators: Dr. Steven E. Waggoner, Dr. Peter Rose, Dr. Kimberly E. Resnick
  - **Metro Health Medical Center**
    - Dr. Rose
- **Cleveland Clinic (6 hospitals)**
  - Dr. Rose
- **Summa Health System (8 hospitals)**
  - Dr. Gruenigen
- **The University of Toledo Medical Center**
  - Dr. Geisler
- **Arthur G. James Cancer Hospital and Richard J. Solove Research Institute (OSU)**
  - Dr. Cohn
- **Mount Carmel (3 Hospitals)**
  - Dr. Vaccarello
- **Miami Valley Hospital**
  - Dr. Nahhas
- **Good Samaritan Hospital/Dayton**
  - Dr. Nahhas
- **Christ Hospital**
  - Dr. Bowling
- **Good Samaritan Hospital/Cincinnati**
  - Dr. Pavelka
- **University of Cincinnati Hospital**
  - Dr. Richards

# Overall Investigative Team

- Duke University Medical Center
  - Joellen Schildkraut, PhD
- The Cancer Institute of New Jersey
  - Elisa Bandera, MD
- **Case Western Reserve University**
  - **Jill Barnholtz-Sloan, PhD**
  - **Diana Slone, CTR**
  - **Yingli Wolinsky, PhD, MBA**
- Hollings Cancer Center
  - Anthony Alberg, PhD
- LSU Health Sciences Center
  - Neal Simonsen, PhD
- University of Alabama-Birmingham
  - Ellen Funkhouser, DrPH
- University of Illinois at Chicago
  - Therese Dolecek, PhD
- University of Tennessee (Georgia)
  - Paul Terry, PhD
- Wayne State University
  - Ann Schwartz, PhD

**THANK YOU!!!!**

