Inferring Ancestry from Genetic Data

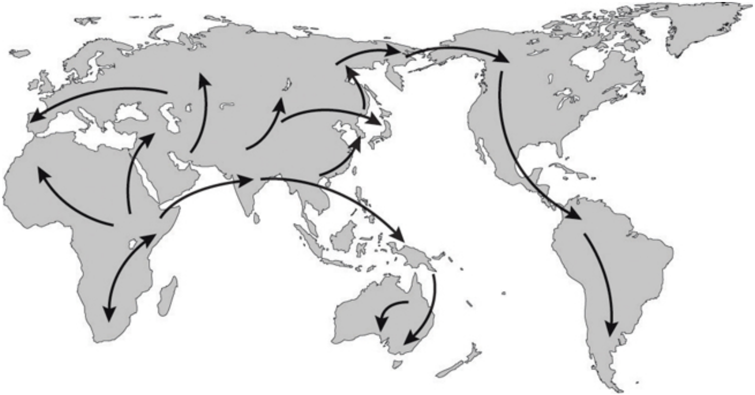
Timothy Thornton

Professor
Department of Biostatistics
University of Washington

Clinical Research Education in Genome Science (CREiGS) 2020

Background: Population Structure

- Humans originally spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between isolated groups.



Population Structure Inference

- Inference on genetic ancestry differences among individuals from different populations, or **population structure**, has been motivated by a variety of applications:
  - population genetics
  - genetic association studies
  - personalized medicine
  - forensics
- Advancements in array-based genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail
- A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.

Inferring Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA applied to genotype data can be used to calculate **principal components** (PCs) that explain differences among the sample individuals in the genetic data
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- Individuals with "similar" values for a particular top principal component are expected to have "similar" ancestry for that axes.

Standard Principal Components Analysis (sPCA)

- sPCA is an unsupervised learning tool for dimension reduction in multivariate analysis.
- Widely used in genetics community to infer population structure from genetic data.
  - Belief that top principal components (PCs) will reflect population structure in the sample.
- Orthogonal linear transformation to a new coordinate system
  - sequentially identifies linear combinations of genetic markers that explain the greatest proportion of variability in the data
  - these define the axes (PCs) of the new coordinate system
  - each individual has a value along each PC
- EIGENSOFT (Price et al., 2006) is a popular implementation of PCA.

Data Structure

- Sample of N individuals, indexed by $i = 1, 2, \ldots, N$.
- Genome screen data on M genetic autosomal markers, indexed by $m = 1, 2, \ldots, M$.
- At each marker, for each individual, we have a genotype value, g_{im}.
- Here we consider bi-allelic markers, so g_{im} takes values 0, 1, or 2, corresponding to the number of reference alleles.
- We center and standardize these genotype values:

$$z_{im} = \frac{g_{im} - 2\hat{p}_m}{\sqrt{2\hat{p}_m(1 - \hat{p}_m)}}$$

where $\hat{p}_m$ is an estimate of the reference allele frequency for marker m.

Genetic Correlation Estimation

- Create an N x M matrix, **Z**, of centered and standardized genotype values, and with **Z** we can obtain an N x N genetic relatedness matrix (GRM) for all possible pairs in the sample:

$$\widehat{\boldsymbol{\Psi}} \;=\; \frac{1}{M}\mathbf{Z}\mathbf{Z}^T$$

- The (i, j)th element of this matrix is

$$\widehat{\boldsymbol{\Psi}}_{ij} = \frac{1}{M}\sum_{m=1}^{M}\frac{(g_{im} - 2\hat{p}_m)\,(g_{jm} - 2\hat{p}_m)}{2\hat{p}_m(1 - \hat{p}_m)},$$

  where $\widehat{\boldsymbol{\Psi}}_{ij}$ can be viewed as an estimate of the genome-wide average genetic correlation between individuals i and j.
- Individuals from the same ancestral population are expected to have genotypic values that are more correlated than individuals from different ancestral populations.

Standard Principal Components Analysis (sPCA)

- PCA is performed by obtaining the eigen-decomposition of $\widehat{\boldsymbol{\Psi}}$; that is, we find **eigenvectors** and **eigenvalues** such that $\widehat{\boldsymbol{\Psi}} = \mathsf{\mathbf{V}}^T\mathsf{\mathbf{L}}\mathsf{\mathbf{V}}$ where
  - $\mathsf{\mathbf{V}} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N]$ is a N x N matrix consisting of N eigenvectors, each of length N
  - $\mathsf{\mathbf{L}}$ is a diagonal matrix of N eigenvalues, $(\lambda_1 > \lambda_2 > \ldots > \lambda_N)$, that are in decreasing order, i.e.,

$$\mathsf{\mathbf{L}} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & \lambda_N \end{bmatrix}$$

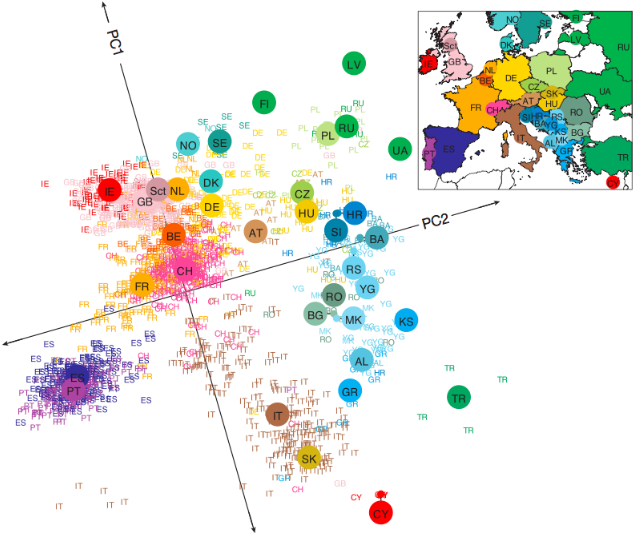
Standard Principal Components Analysis (sPCA)

- The d^{th} principal component (eigenvector) corresponds to eigenvalue λ_d, where λ_d is proportional to the percentage of variability in the genome-screen data that is explained by $\mathbf{V}_d$.
- The top principal components are viewed as continuous axes of variation that reflect genetic variation that best explain genotypic variability amongst the N sample individuals.
- Individuals with "similar" values for a particular top principal component are expected to have "similar" ancestry for that axes.
- As a result, eigenvectors (PCs) are often used as surrogates for ancestry (or population structure).

PCA of Europeans

- In a very influential paper, an application of PCA to genetic data from European samples (Novembre et al., 2008) illustrated that among Europeans for whom all four grandparents originated in the same country, the first two principal components computed using 200,000 SNPs could map their country of origin quite accurately in a plane

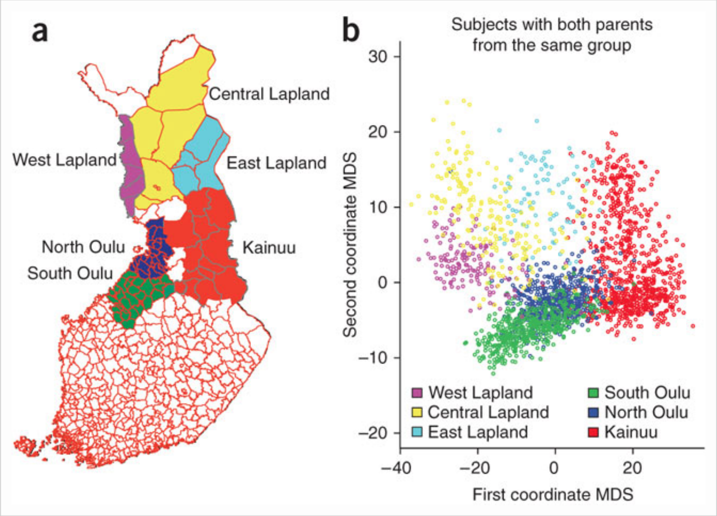
PCA of Europeans



PCA in Finland

- There can be population structure in all populations, even those that appear to be relatively "homogenous"
- An application of principal components to genetic data from Finland samples (Sabatti et al., 2009) identified population structure that corresponded very well to geographic regions in this country.

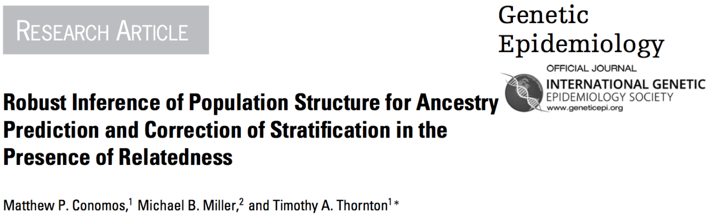
PCA in Finland



Caution: Relatedness Confounds sPCA

- Recall that the elements in the GRM used by sPCA, $\widehat{\mathbf{\Psi}}_{ij}$, are an estimate of the genome-wide average genetic correlation between individuals i and j.
- Conomos et al. (2015) showed that $\mathbf{\Psi}_{ij} = 2\left[\phi_{ij} + (1 - \phi_{ij})A_{ij}\right]$
  - ϕ_{ij}: kinship coefficient - a measure of familial relatedness (more about this later!)
  - A_{ij}: a measure of ancestral similarity
- sPCA is an unsupervised method; in related samples we don't know the correlation structure each eigenvector is actually reflecting
  - If the only genetic correlation structure among individuals is due to ancestry, $\mathbf{\Psi}$ and the top PCs will capture this.
  - If there is relatedness in the sample, the top PCs may reflect this or some combination of ancestry and relatedness.

PCA for Related Samples



RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness

Matthew P. Conomos,[1] Michael B. Miller,[2] and Timothy A. Thornton[1]*

The PC-AiR method was developed for performing a **P**rincipal **C**omponents **A**nalysis **i**n **R**elated samples. The algorithm has the following steps:

Genetic Analysis of Admixed Populations

Timothy Thornton

Professor
Department of Biostatistics
University of Washington

Clinical Research Education in Genome Science (CREiGS) 2020

Recently Admixed Populations

- Multi-ethnic cohorts often include individuals sampled from **admixed populations**: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated.
- Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations.
- Examples of admixed populations include
  - African Americans and Hispanic Americans in the U.S
  - Latinos from throughout Latin America
  - Uyghur population of Central Asia
  - Cape Verdeans
  - South African "Coloured" population

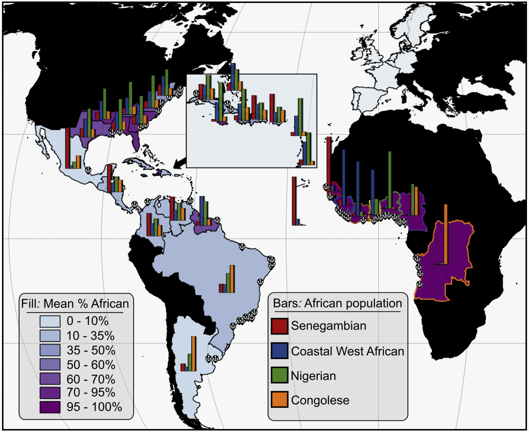
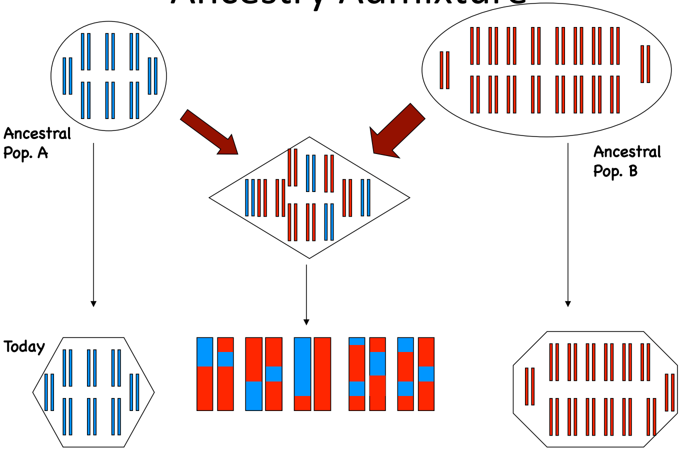


Figure: Estimated Proportions of African Ancestry across Atlantic Africa, the Americas, and Europe

The American Journal of Human Genetics 2020 107265-277DOI: (10.1016/j.ajhg.2020.06.012)

Ancestry Admixture



- The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.

Recently Admixed Populations

- Can be substantial genetic heterogeneity among individuals in admixed populations
- Admixed populations are ancestrally admixed and thus have population structure.
- Statistical method for estimating admixture proportions using genetic data are available

Supervised Learning of Ancestry Admixture

- Methods, such as ADMIXTURE (Alexander et al., 2009), have recently been developed for supervised learning of ancestry proportions for an admixed individuals using high-density SNP data.
- Most use either a hidden Markov model (HMM) or an Expectation-Maximization (EM) algorithm to infer genome-wide or global ancestry
- Other methods, such as RFMix (Maples et al., 2013) have been implemented to infer local ancestry of admixed individuals, i.e., ancestry at specific locations on the genome.

Supervised Learning of Ancestry Admixture

- Example: We are interested in identifying the ancestry proportions for an admixed individual
- Suppose the observed sequence on a chromosome for an admixed individual is:

  ...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...

- Assume that we have a suitable reference panel with diverse ancestries, and a similar sequence is observed in samples from one of the "homogenous" reference populations:

  ...TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

  ...AGATCCTGAACCTAGATTACAGATTACAGATTACAGATACCAATGCTTCGATGGAC...

  ...CGATCCTGAACCTAGATTACAGATTACAGATTTGCGTATACAATGCTTCGATGGAC...

- Can infer the likelihood of the observed sequence in the admixed individuals being derived from each of the reference population samples. This can be performed across the genome.

Example: HapMap ASW and MXL Ancestry Inference

- Genome-screen data on 150,872 autosomal SNPs was used to estimate ancestry
- Estimated genome-wide ancestry proportions of every individual using the ADMIXTURE (Alexander et al., 2009) software
- A supervised analysis was conducted using genotype data from the following reference population samples for three "ancestral" populations
    - HapMap YRI for West African ancestry
    - HapMap CEU samples for northern and western European ancestry
    - HGDP Native American samples for Native American ancestry.

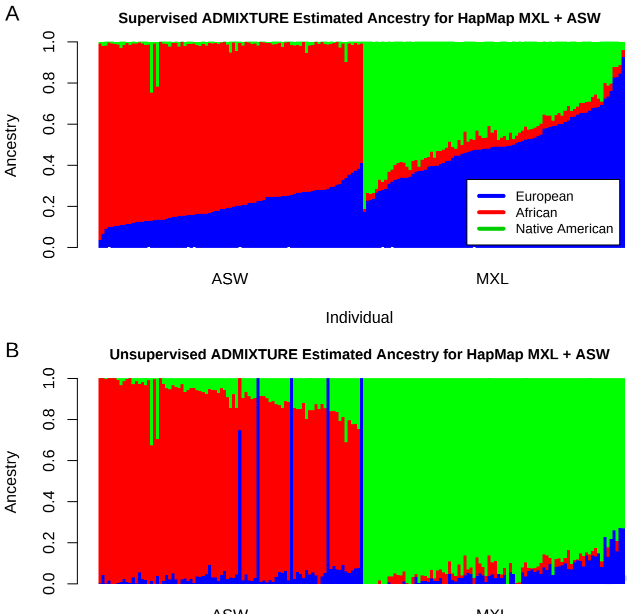
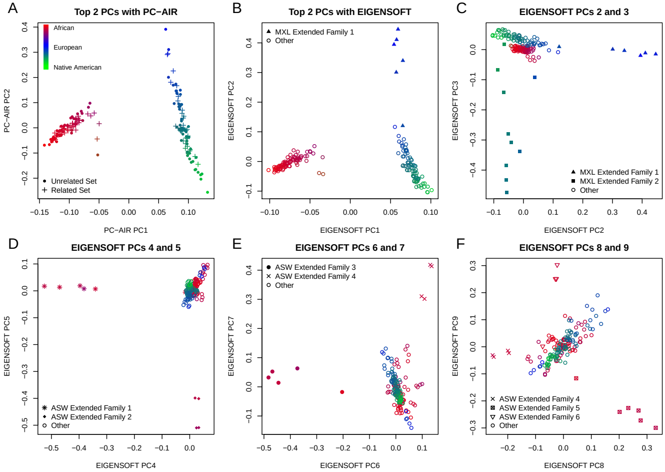

A
Supervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW
Ancestry
0.0
0.2
0.4
0.6
0.8
1.0
European
African
Native American
ASW
MXL
Individual
B
Unsupervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW
Ancestry
0.0
0.2
0.4
0.6
0.8
1.0
ASW
MXL

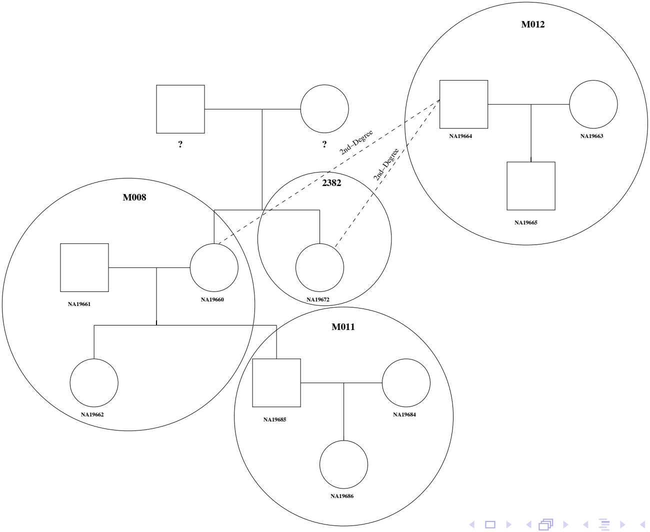
Table: Average Estimated Ancestry Proportions for HapMap African Americans and Mexican Americans

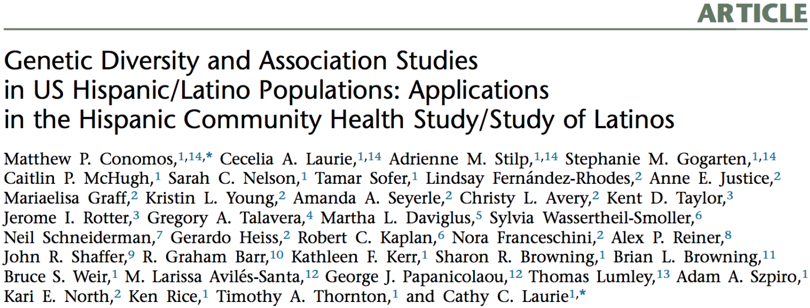
	Estimated Ancestry Proportions (SD)		
Population	European	African	Native American
MXL	49.9% (14.8%)	6%(1.8%)	44.1% (14.8%)
ASW	20.5% (7.9%)	77.5% (8.4%)	1.9% (3.5%)

Figure: **HapMap MXL + ASW Sample**



HapMap MXL: Known and Cryptic Relatedness





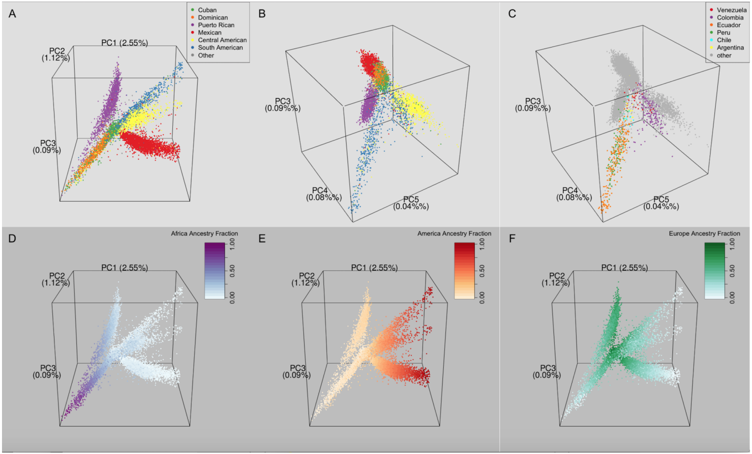
ARTICLE

Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

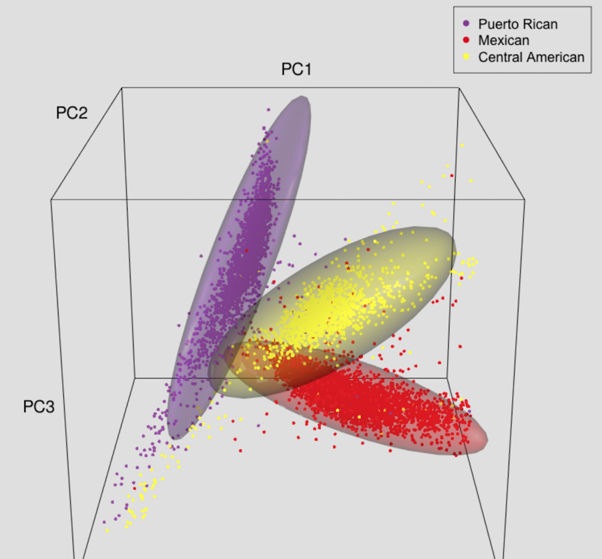
Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent D. Taylor,[3] Jerome I. Rotter,[3] Gregory A. Talavera,[4] Martha L. Daviglus,[5] Sylvia Wassertheil-Smoller,[6] Neil Schneiderman,[7] Gerardo Heiss,[2] Robert C. Kaplan,[6] Nora Franceschini,[2] Alex P. Reiner,[8] John R. Shaffer,[9] R. Graham Barr,[10] Kathleen F. Kerr,[1] Sharon R. Browning,[1] Brian L. Browning,[11] Bruce S. Weir,[1] M. Larissa Avilés-Santa,[12] George J. Papanicolaou,[12] Thomas Lumley,[13] Adam A. Szpiro,[1] Kari E. North,[2] Ken Rice,[1] Timothy A. Thornton,[1] and Cathy C. Laurie[1,*]

- "Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos." (2016) *American Journal of Human Genetics* 98(1), 165-184.

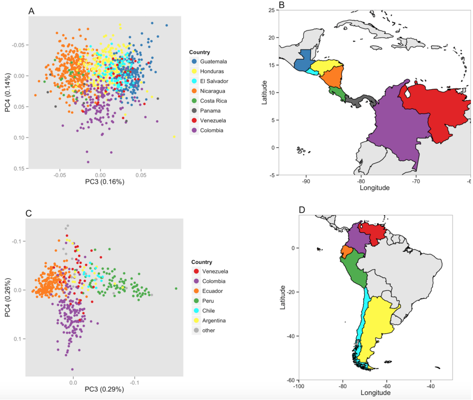
PCA-AiR: Hispanic Community Health Study



PC-AiR: Hispanic Community Health Study



PC-AiR: Hispanic Community Health Study

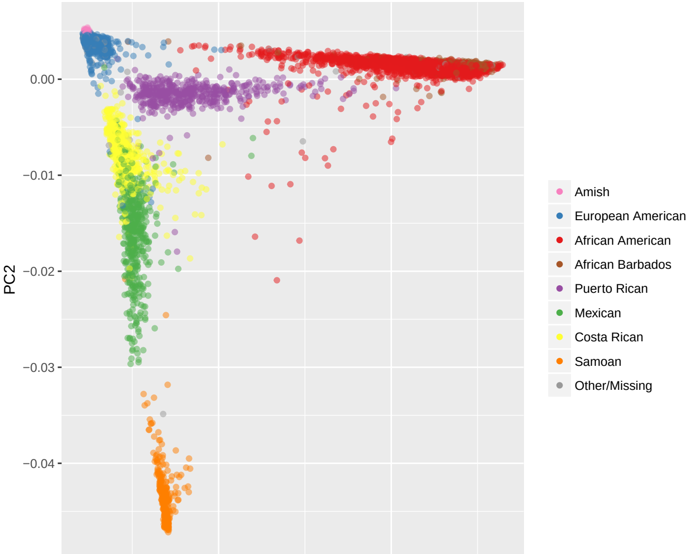


- Genetic differentiation among individuals is associated with the geography of their countries of grandparental origin.

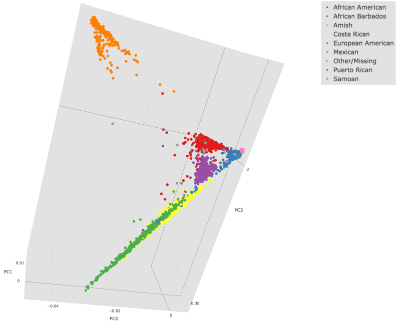
TOPMed Phase I: Population Structure Inference

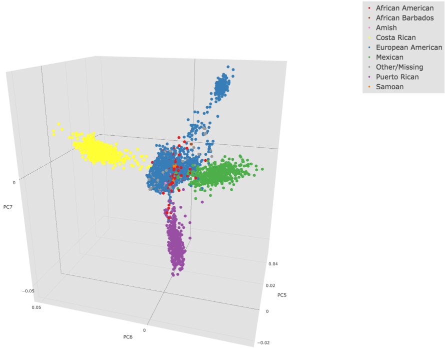
- TOPMed cohorts are multi-ethnic
- Variety of study designs: family-based, case-control, founder populations (Amish).
- PC-AiR algorithm applied to TOPMed Phase I data
- Variants with minor allele frequency $> 1\%$ (common variants) were used for population structure inference

TOPMed Phase I: PC-AiR

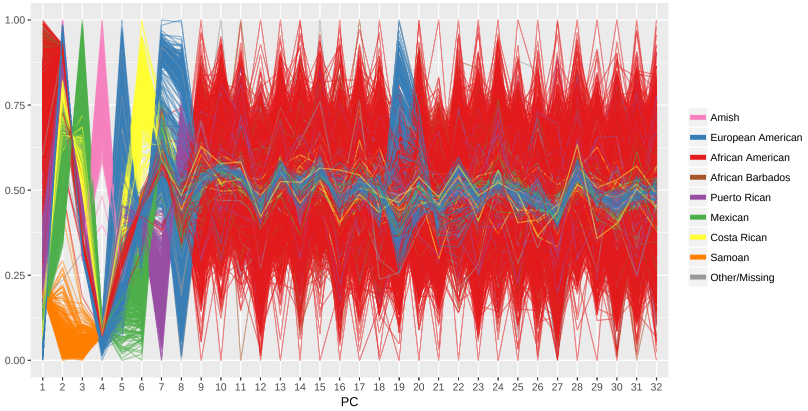


TOPMed Phase I: PC-AiR





TOPMed Phase I: PC-AiR



References

- Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**,1655-1664.
- Conomos MP, Miller M, Thornton T (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* **39**, 276-93
- Maples, B. K., Gravel, S., Kenny, E. E., Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, **93**, 278-288.
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867-2873.

References

- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R. (2008). Genes mirror geography within Europe. *Nature* **456**, 98-101.
- Patterson,N., Price, A.L., Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.
- Sabatti, C., Service, S. K., Hartikainen, A. L., Pouta, A., Ripatti, S., Brodsky, J., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics.*, **41**, 35-46.

Genetic Analysis of Complex Traits in Diverse Populations: Challenges and Opportunities

Timothy Thornton, PhD

Robert W. Day Endowed Professor of Public Health

Department of Biostatistics

University of Washington

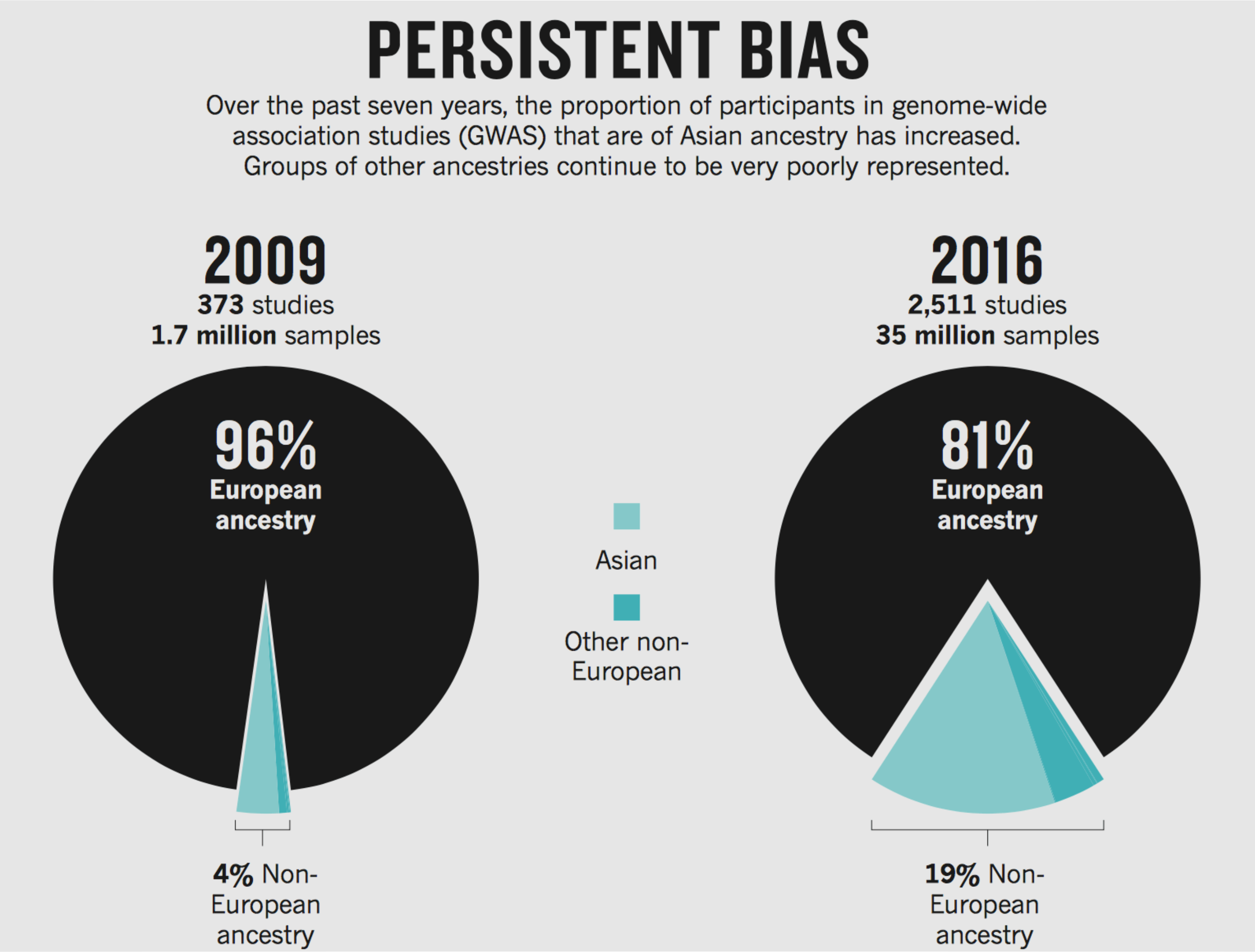
Clinical Research Education in Genome Science (CREiGS) 2020

SCHOOL OF PUBLIC HEALTH
UNIVERSITY *of* WASHINGTON

Introduction

- To date, millions of genomes have been interrogated for the identification of genetic variants influencing complex traits related to human health and diseases
- Whereas genetic studies have primarily examined populations of European ancestry, there has been significant efforts in recent years to increase the diversity study participants
- Genetic complex trait mapping in multi-ethnic cohort studies offer unprecedented opportunities for:
    - Identification of novel population-specific variants underlying phenotypic diversity
    - new insights into human health and health disparities of underrepresented minority populations for many complex diseases

Current State of Affairs

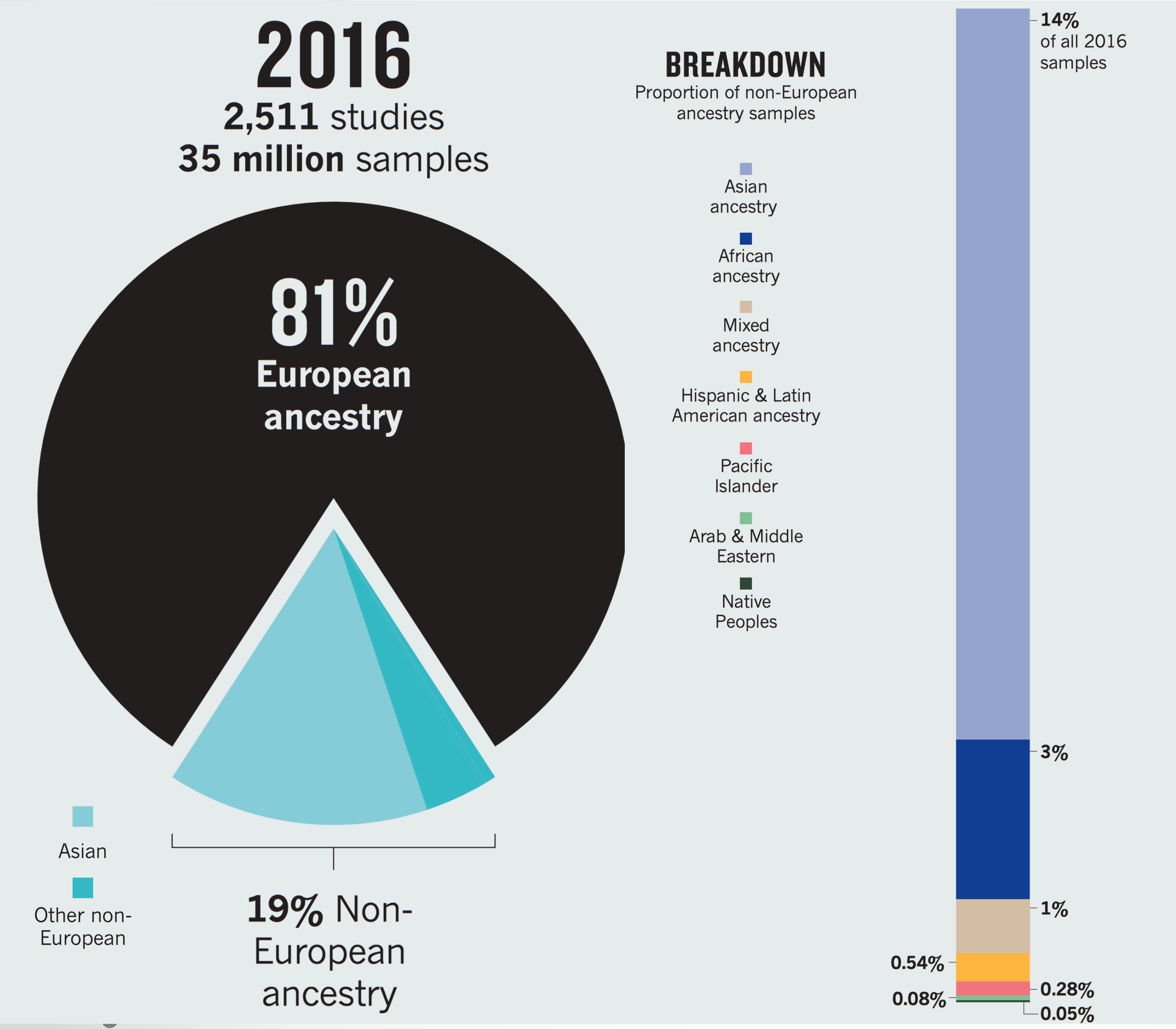


PERSISTENT BIAS

Over the past seven years, the proportion of participants in genome-wide association studies (GWAS) that are of Asian ancestry has increased. Groups of other ancestries continue to be very poorly represented.

Popejoy and Fullerton (Nature, 2016)

Over-representation of European Populations in Genetic Studies



- Biased understanding of which variants are important
- Potential for new health care inequalities

Popejoy and Fullerton. *Nature*, 2016

Need for Genetic Studies in Diverse Populations

- Medical genomics has focused almost entirely on those of European descent.
- Other race and ethnic groups must be studied to ensure that more people benefit

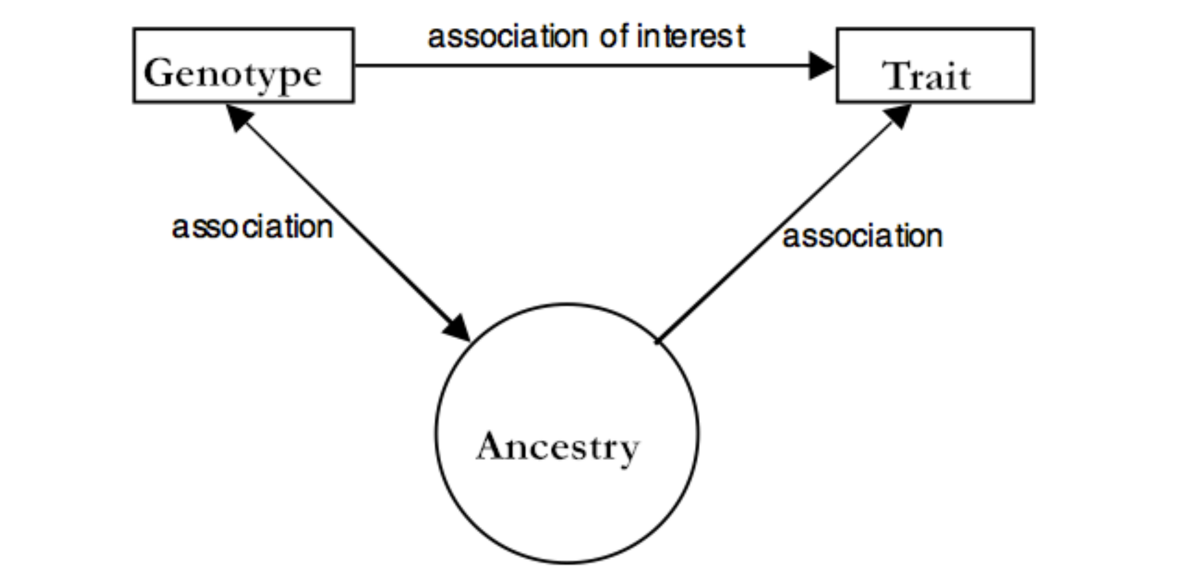
Bustamante et al. (Nature, 2011)

Genetic Studies in Multi-Ethnic Populations

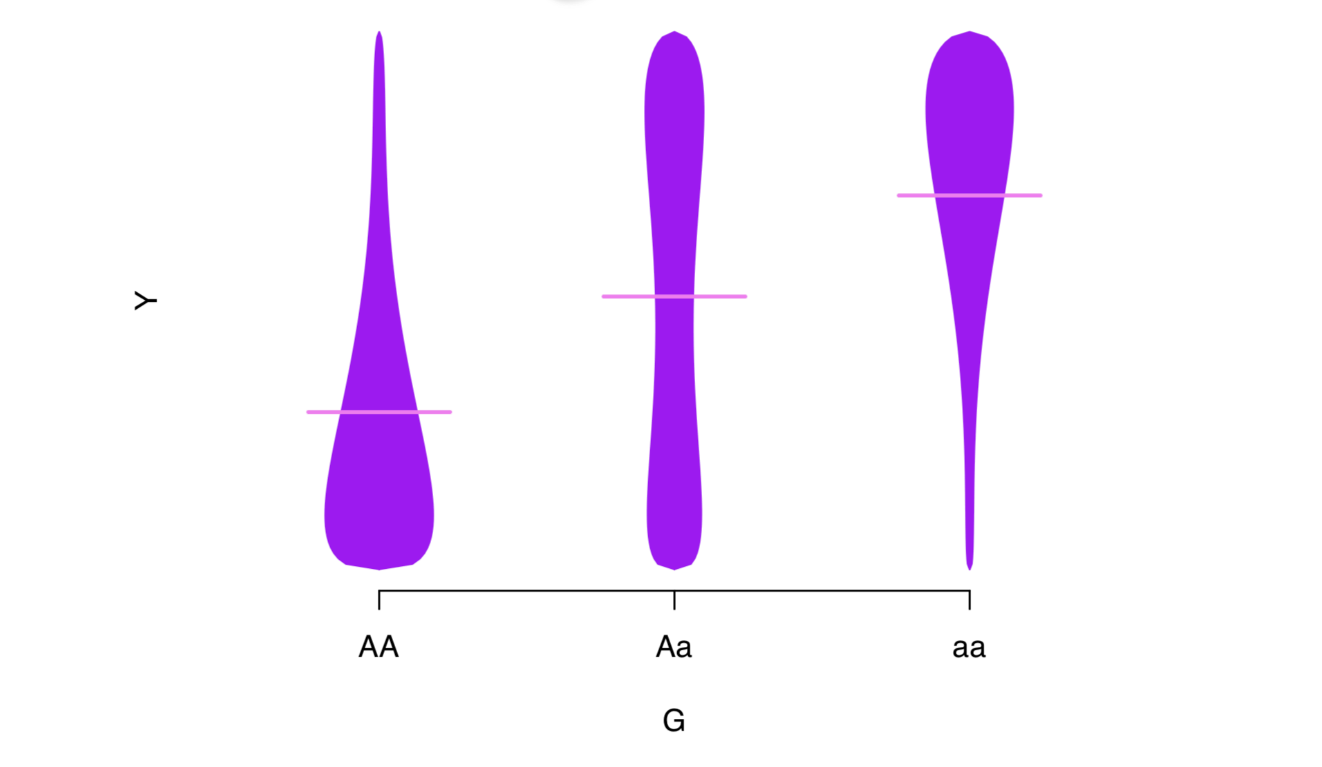
- There remain significant challenges with complex trait mapping in multi-ethnic populations
- Two well know challenges are:
  - **Heterogeneous genetic ancestry and environmental backgrounds among sampled individuals**
  - **Correlated genotype and phenotype data among relatives, known and/or cryptic, in the sample**

Confounding is a serious concerns for genetic studies in multi-ethnic populations

- Ethnic groups (and subgroups) have often share distinct dietary habits and other lifestyle characteristics that result in traits of interest having **different distributions** that are correlated with genetic ancestry and/or ethnicity.



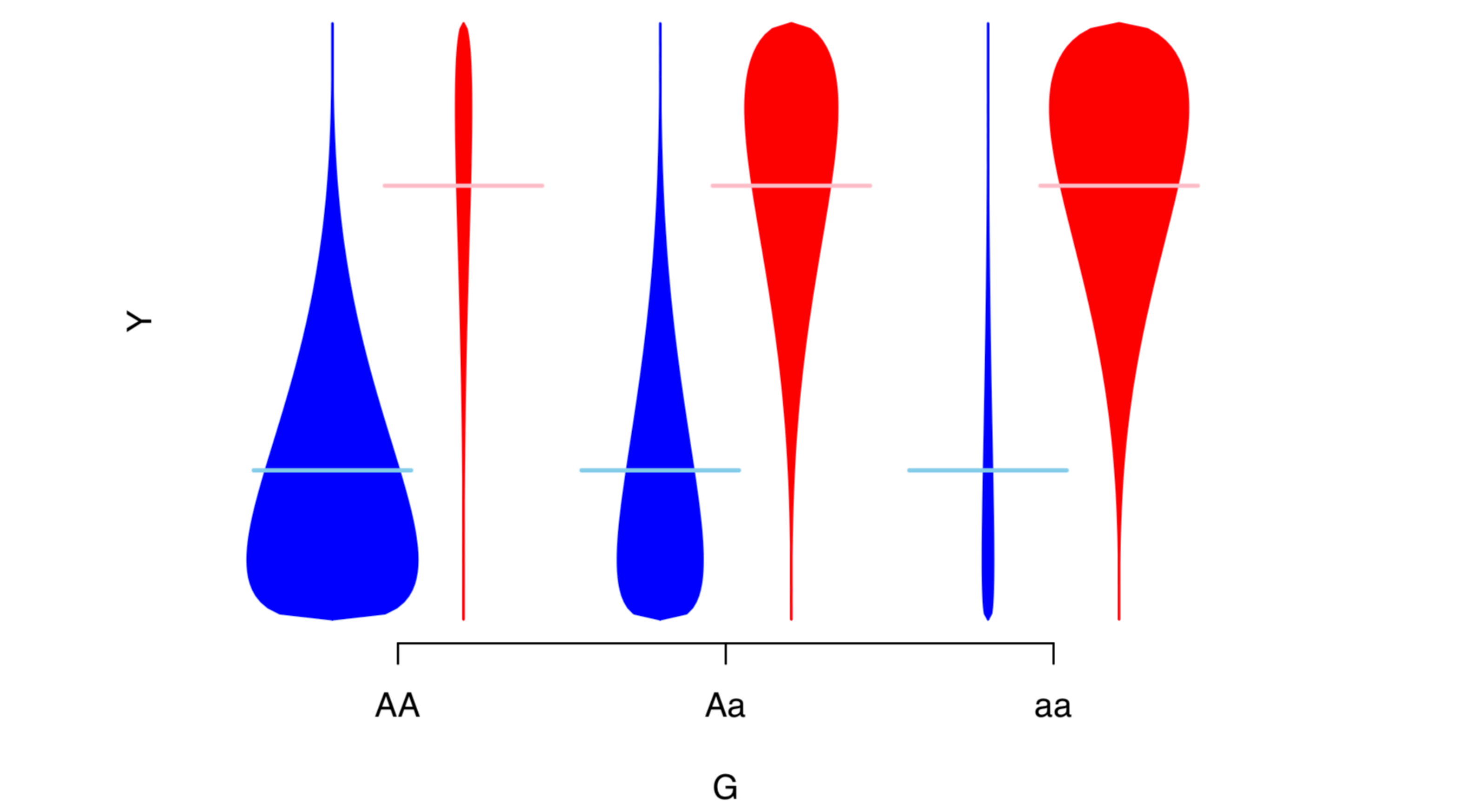
Confounding: multi-ethnic studies



$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{g}_s \beta_1 + \epsilon$$

- $\mathbf{Y}$ is a vector of phenotypes
- $\mathbf{g}_s$ is an additive genotype count vector for a SNP s, where each entry corresponds to the number of reference alleles (A) an individual has, e.g., 0, 1, or 2;
- $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$

Confounding: multi-ethnic studies



- The relationship between phenotype vector (**Y**) and genotype vector (**G**) looks much less interesting when broken down and assessed within ancestry groups;
- Well known that the genetic ancestry is a confounder in multi-ethnic studies that can lead to spurious associations

Genetic Ancestry Inference and Adjustment

- In practice, multi-ethnic cohort studies will not have discrete or a fixed number of ancestry groups.
- In addition, admixed populations have ancestry from multiple ancestral groups
- As previously discussed, principal components analysis (PCA) is widely used to infer population structure from genetic data
- Principal components (PCs) can also be used as surrogates for ancestry (or population structure) to protect against spurious association in genetic association studies

Adjusting for Genetic Ancestry

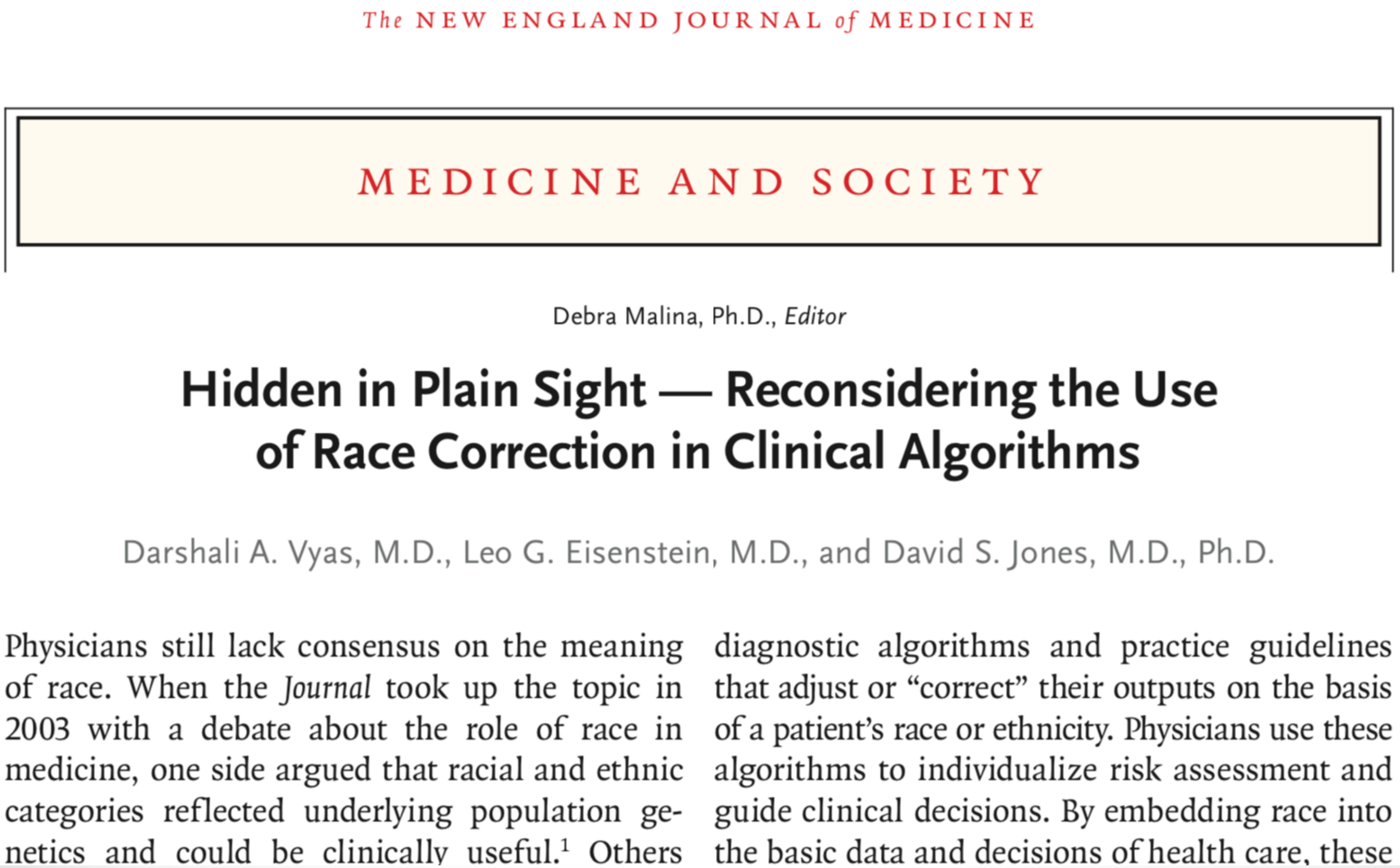
- The top PCs are often included as **fixed effects** in regression models used for assessing genotype/phenotype associations in samples with population structure

$$E\left(\mathbf{Y}\right) = \beta_0 \mathbf{1} + \mathbf{g_s}\beta_1 + \gamma_1 PC_1 + \gamma_2 PC_2 + \gamma_3 PC_3 + \cdots$$

i.e., regression model adjusting for PC1, PC2, PC3 etc. (Logistic, Cox regression can be adjusted similarly)

- Among people with the same ancestry (i.e. the same PCs) β_1 gives us the difference in mean phenotype, per 1-unit difference in $\mathbf{g}_s$
- If the effect of $\mathbf{g}_s$ differs by PCs, β_1 provides a (sensible) average of these genetic effects

Challenges in understanding and using Race/Ethnicity in Medicine



The NEW ENGLAND JOURNAL of MEDICINE

MEDICINE AND SOCIETY

Debra Malina, Ph.D., *Editor*

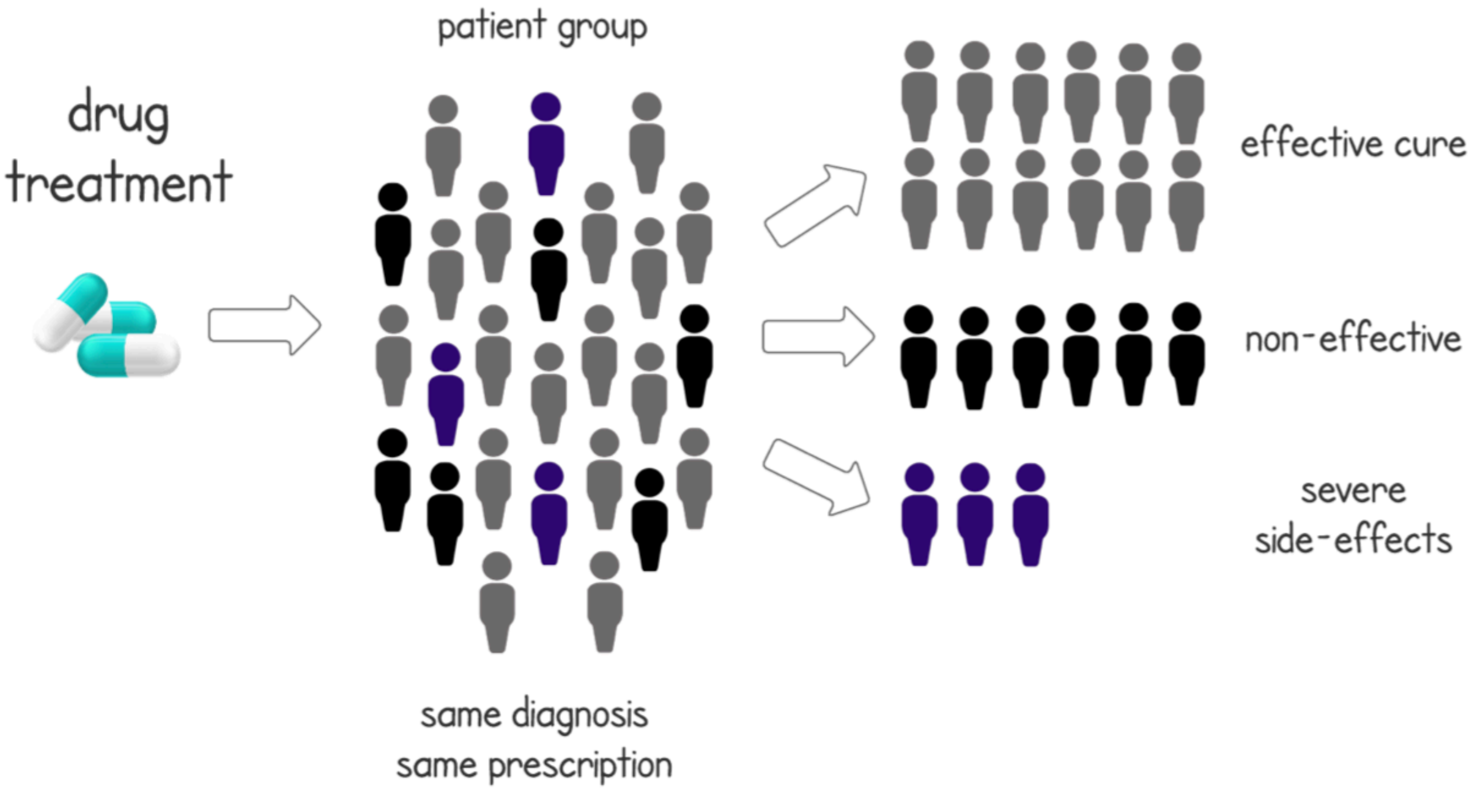
Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms

Darshali A. Vyas, M.D., Leo G. Eisenstein, M.D., and David S. Jones, M.D., Ph.D.

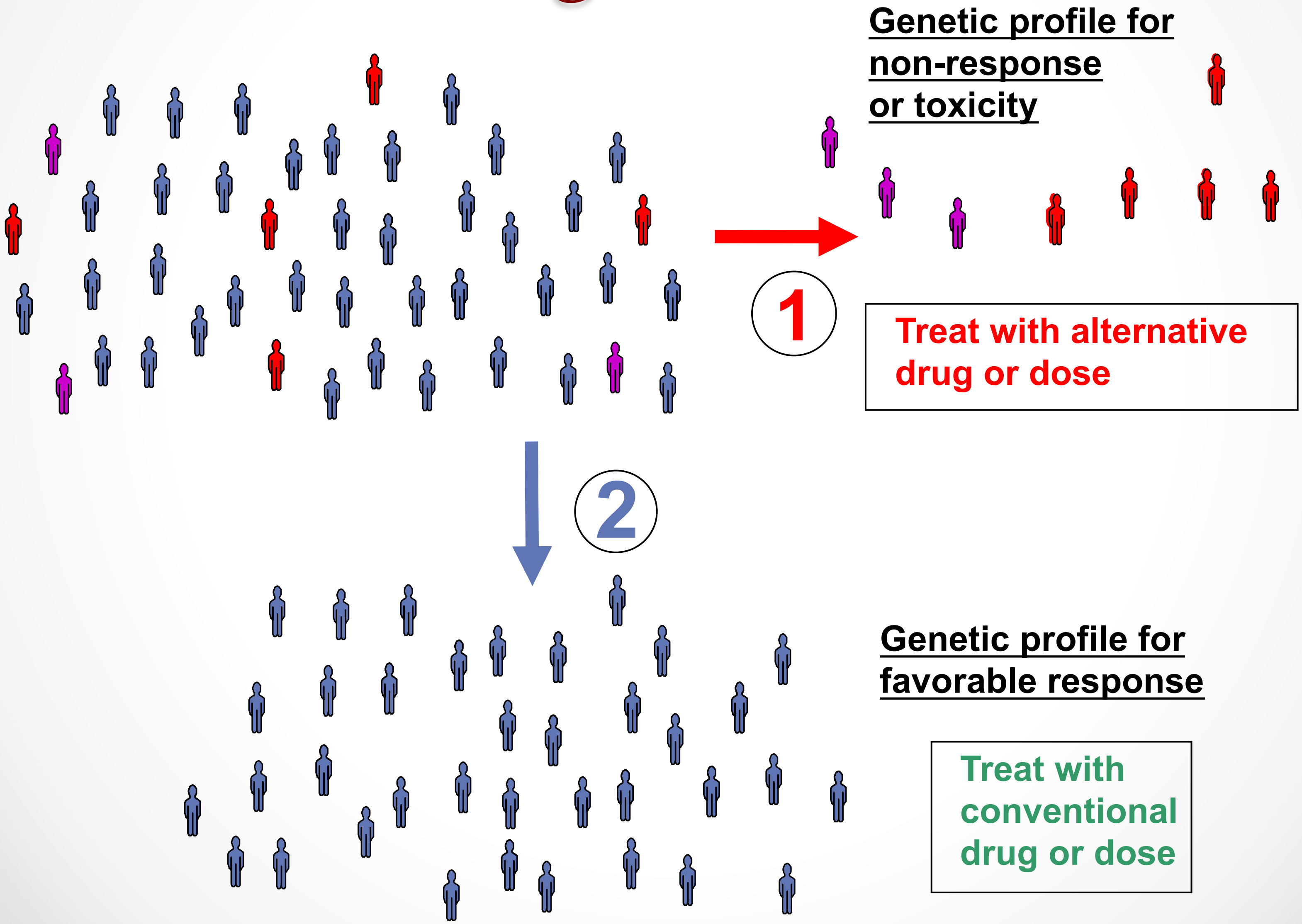
Physicians still lack consensus on the meaning of race. When the *Journal* took up the topic in 2003 with a debate about the role of race in medicine, one side argued that racial and ethnic categories reflected underlying population genetics and could be clinically useful.[1] Others

diagnostic algorithms and practice guidelines that adjust or "correct" their outputs on the basis of a patient's race or ethnicity. Physicians use these algorithms to individualize risk assessment and guide clinical decisions. By embedding race into the basic data and decisions of health care, these

Personalized Medicine: Pharmacogenomics



Substantial Ethnic/Racial Disparities in Pharmacogenomics Research



Genetic profile for non-response or toxicity

Treat with alternative drug or dose

Genetic profile for favorable response

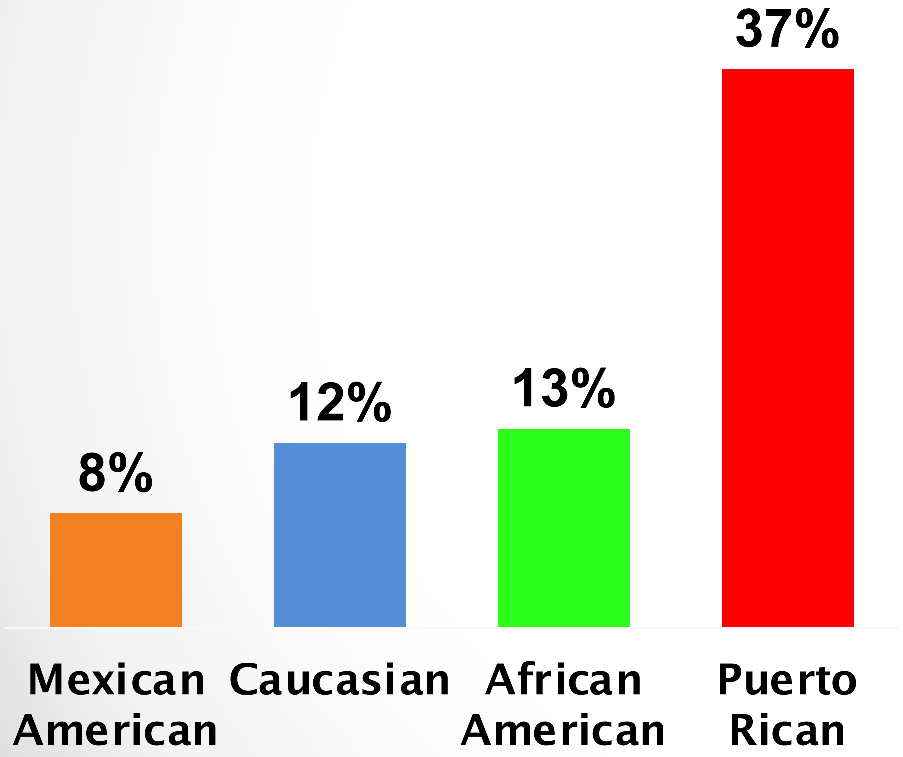
Treat with conventional drug or dose

Example: Asthma Health Disparities

- These disparities extend to asthma mortality, which is 3- to 4-fold higher in Puerto Ricans and African Americans compared to Whites and Mexicans.
- Albuterol is the most commonly prescribed asthma medication in the world.
- Dr. Esteban Burchard (UCSF) and colleagues leading Genetics of Asthma in Latino Americans (GALA) and Study of African Americans, Asthma, Genes, & Environments (SAGE)
    - **Marked differences in drug response to Albuterol between racial and ethnic groups, which contribute to health disparities in asthma morbidity and mortality.**

Example: Asthma Affects ~334M Globally

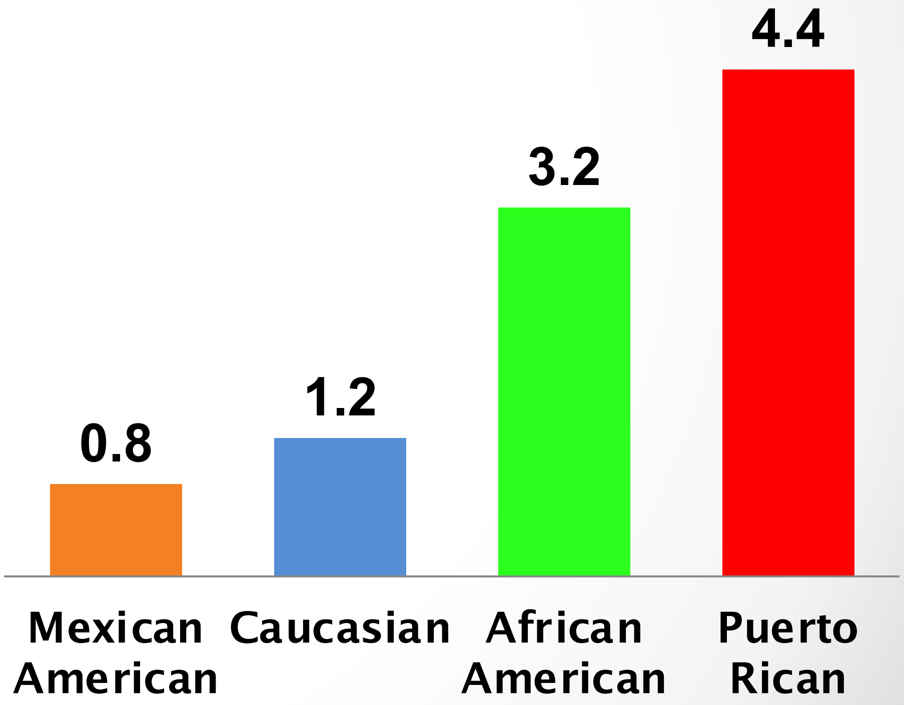
Prevalence



NHLBI Study of Latinos (SOL)

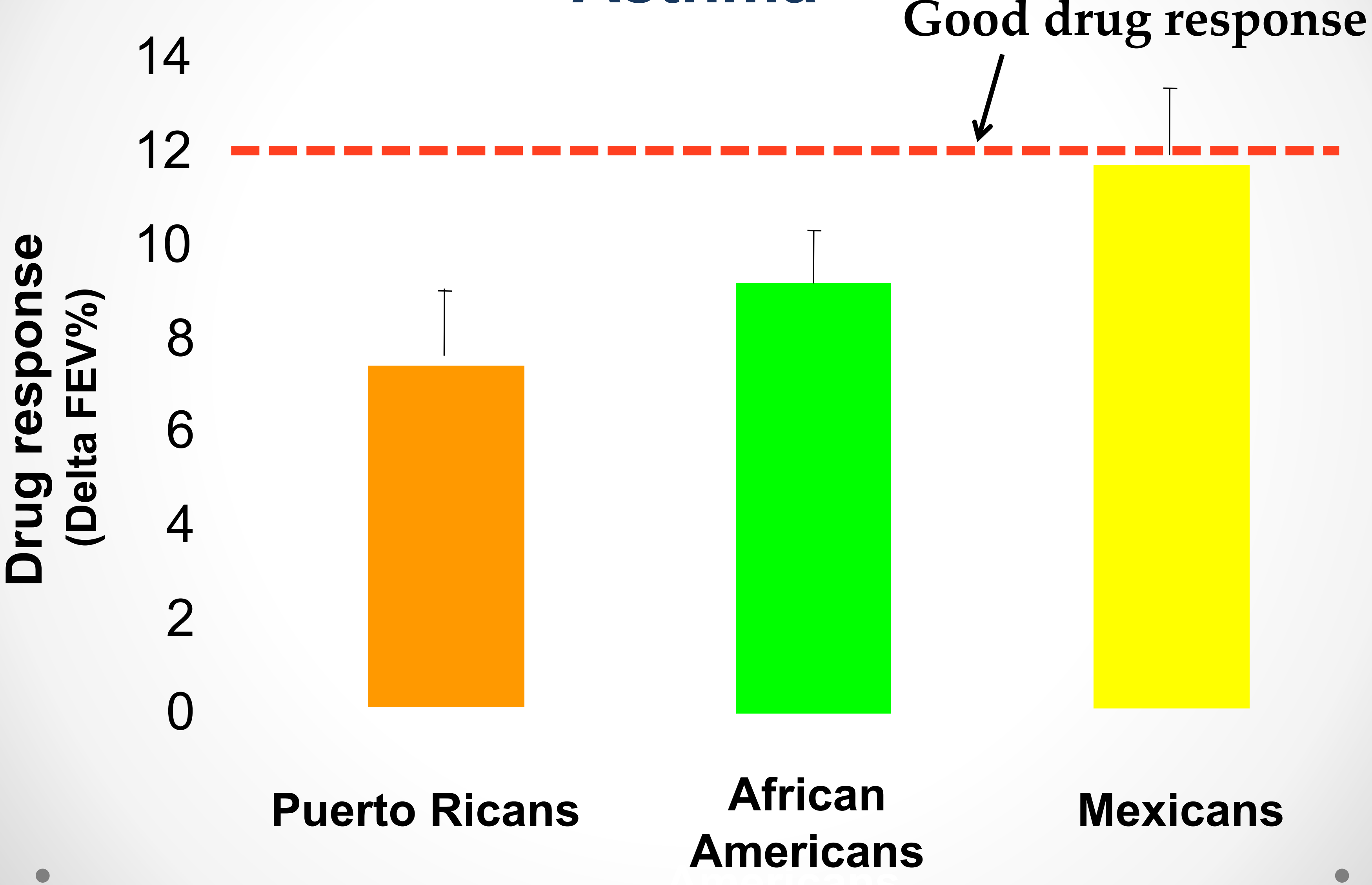
Barr et. ai., *AJRCCM* 2016

Mortality



Akinbami L. CDC/NCHS

GALA: Children with Moderate-to-Severe Asthma



Good drug response

Naqvi, M, *J. of Asthma*, 2007

Courtesy of Dr. Esteban Burchard; UCSF

Salmeterol tiny Black Box Warning

- Salmeterol is used in moderate-to-severe persistent asthma following previous treatment with a short-acting β_2 adrenoreceptor agonist(SABA) such as salbutamol (albuterol).
- However, African Americans, beware!

"In African Americans, asthma-related deaths occurred at a higher rate in patients treated with Salmeterol than those treated with placebo (..relative risk: 7.26..)…"

Pharmacogenomics in Diverse Populations: Cytochrome P450s

- **The Cytochromes P450 (CYPs)** constitute a major drug metabolizing enzyme family that catalyzes the oxidative metabolism of many clinically used compounds in pharmaceutics.
- CYP genes play a major role in inter-individual differences in drug response.
- CYPs have been well studied in European Populations,
- Little is known, however, about CYPs and pharmacogenetic variation in diverse populations

Northwest-Alaska Pharmacogenetic Research Network

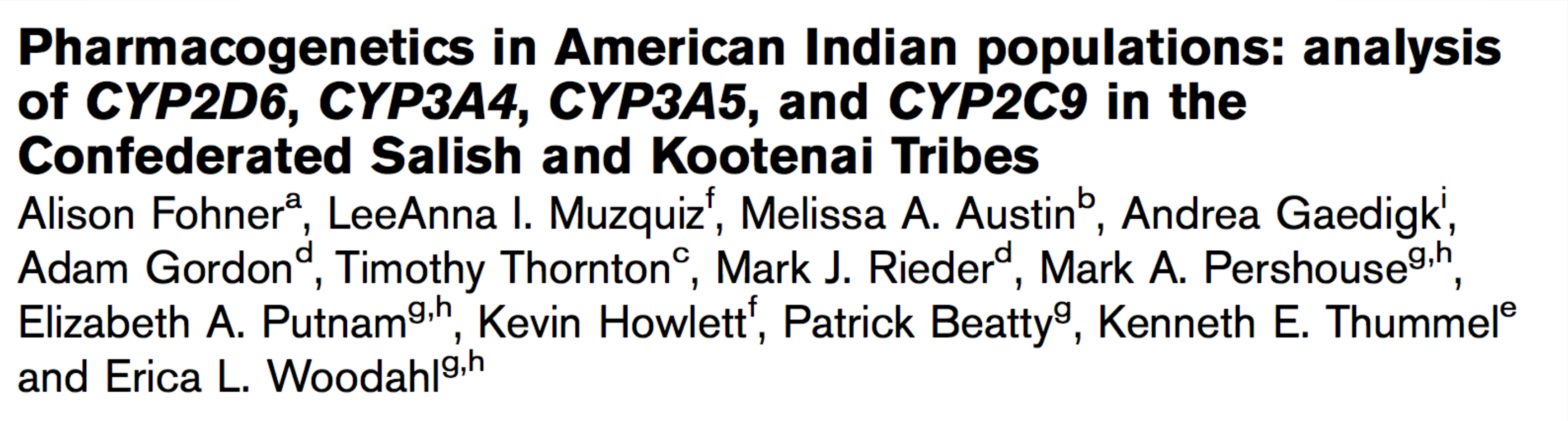
- Support community-university partnerships
- Discover and characterize novel variation among American Indian and Alaska Native (AI/AN) people
- Identify genetic and dietary factors that influence drug response in AI/AN populations

NWA PGRN
Northwest–Alaska Pharmacogenomics Research Network

PGRN

Pharmacogenomics in AI/AN

- Fohner et al. [Pharmacogenet Genomics, 2013]



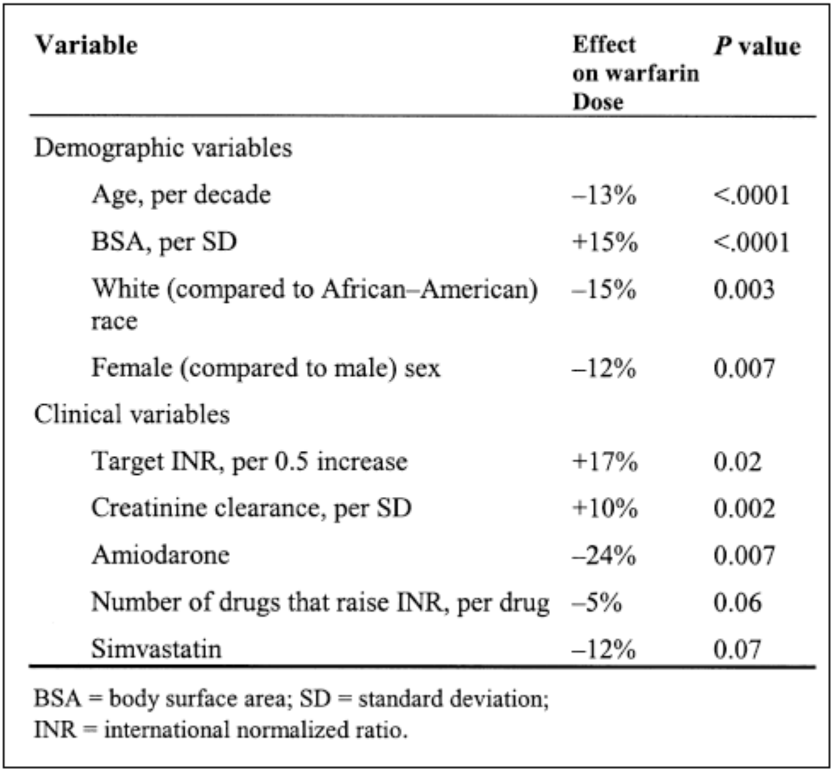
Pharmacogenetics in American Indian populations: analysis of *CYP2D6*, *CYP3A4*, *CYP3A5*, and *CYP2C9* in the Confederated Salish and Kootenai Tribes

Alison Fohner[a], LeeAnna I. Muzquiz[f], Melissa A. Austin[b], Andrea Gaedigk[i], Adam Gordon[d], Timothy Thornton[c], Mark J. Rieder[d], Mark A. Pershouse[g,h], Elizabeth A. Putnam[g,h], Kevin Howlett[f], Patrick Beatty[g], Kenneth E. Thummel[e] and Erica L. Woodahl[g,h]

"The combination of nonfunctional CYP3A5*3 and putative reduced function CYP3A4*1G alleles may predict diminished clearance of CYP3A substrates in American Indian Populations"

Clinical and Demographic Factors Affecting Warfarin Dose

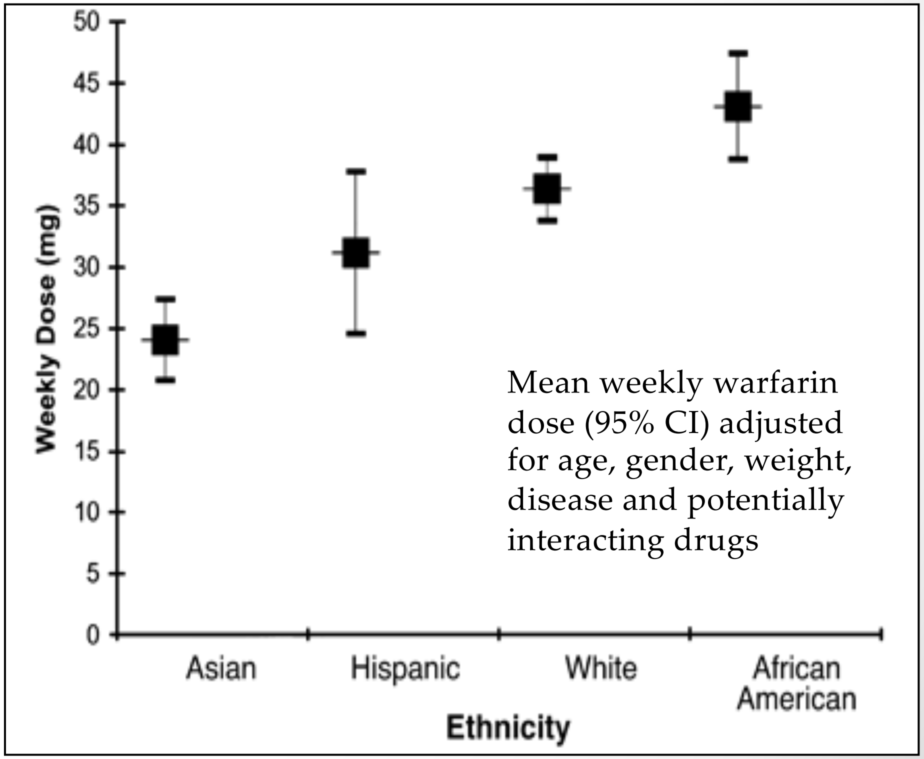
- Wafarin is a medication that is used as an anticoagulant (blood thinner). Commonly used to treat used to treat blood clots such as deep vein thrombosis and pulmonary embolism and to prevent stroke in people who have atrial fibrillation
- Clinical and demographic factors estimated to contribute **20%** to warfarin dose variance
-



Variable	Effect on warfarin Dose	*P* value
Demographic variables		
Age, per decade	–13%	<.0001
BSA, per SD	+15%	<.0001
White (compared to African–American) race	–15%	0.003
Female (compared to male) sex	–12%	0.007
Clinical variables		
Target INR, per 0.5 increase	+17%	0.02
Creatinine clearance, per SD	+10%	0.002
Amiodarone	–24%	0.007
Number of drugs that raise INR, per drug	–5%	0.06
Simvastatin	–12%	0.07

BSA = body surface area; SD = standard deviation; INR = international normalized ratio.

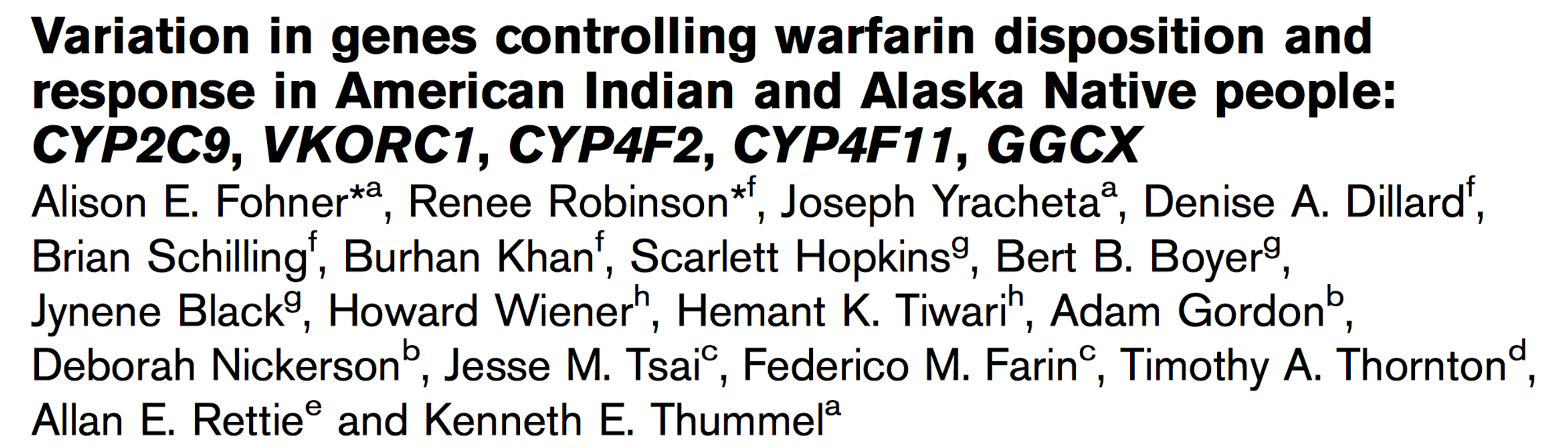
Gage et al., *Thromb Haemost.* (2004)



Mean weekly warfarin dose (95% CI) adjusted for age, gender, weight, disease and potentially interacting drugs

Dang et al., *Ann. Pharmacother.* (2005)

- Fohner et al. [Pharmacogenet Genomics, 2015]

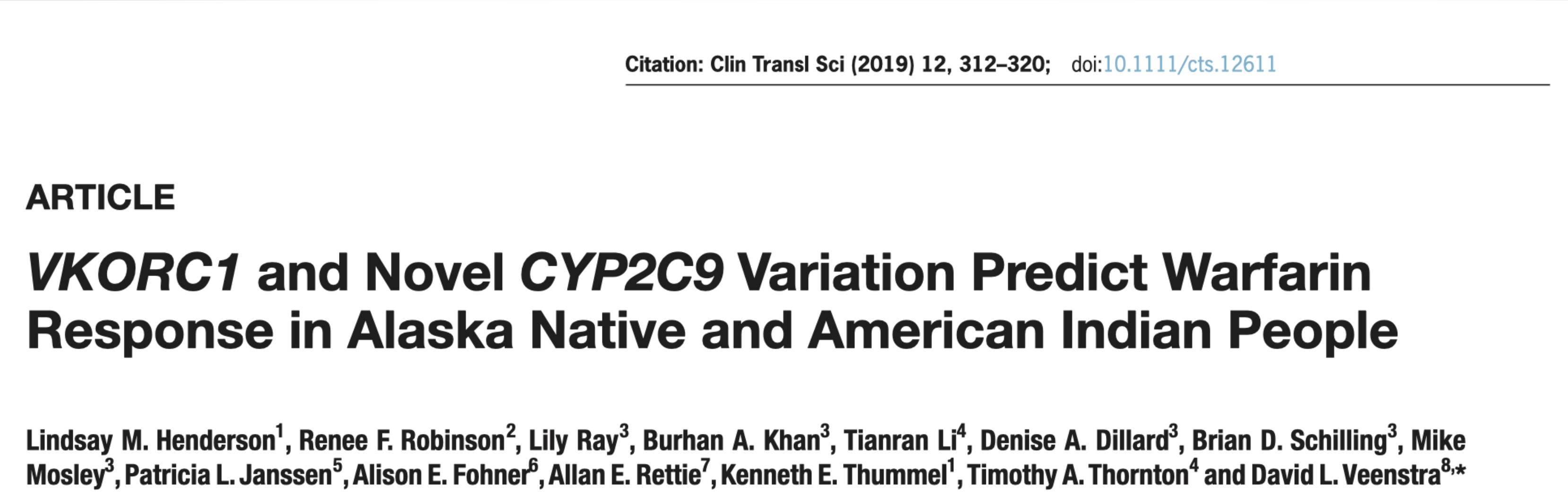


Variation in genes controlling warfarin disposition and response in American Indian and Alaska Native people: *CYP2C9, VKORC1, CYP4F2, CYP4F11, GGCX*

Alison E. Fohner*[a], Renee Robinson*[f], Joseph Yracheta[a], Denise A. Dillard[f], Brian Schilling[f], Burhan Khan[f], Scarlett Hopkins[g], Bert B. Boyer[g], Jynene Black[g], Howard Wiener[h], Hemant K. Tiwari[h], Adam Gordon[b], Deborah Nickerson[b], Jesse M. Tsai[c], Federico M. Farin[c], Timothy A. Thornton[d], Allan E. Rettie[e] and Kenneth E. Thummel[a]

“We identified two relatively common, novel, and potentially function-disrupting variants in CYP2C9 (M1L and N218I)”

***"Overall, we predict a lower average warfarin dose requirement in AI/AN populations in Alaska than that seen in non-AI/AN populations of the US, a finding consistent with clinical experience in Alaska*."**



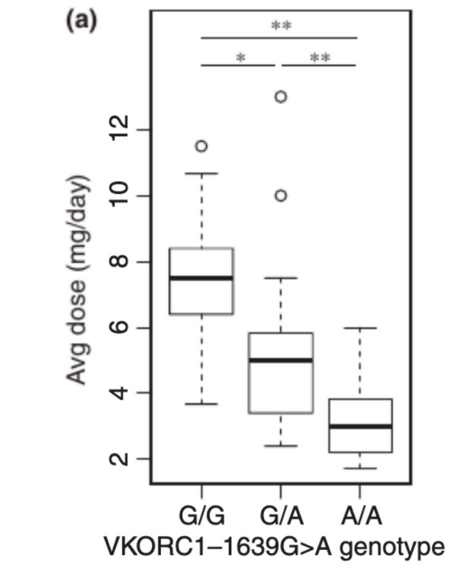
Citation: Clin Transl Sci (2019) 12, 312–320; doi:10.1111/cts.12611

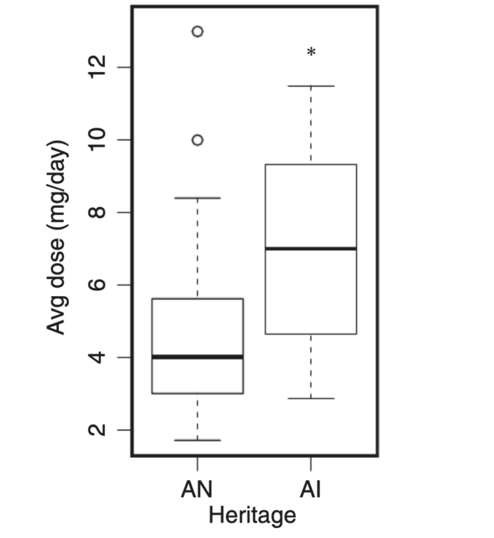
ARTICLE

VKORC1 and Novel *CYP2C9* Variation Predict Warfarin Response in Alaska Native and American Indian People

Lindsay M. Henderson[1], Renee F. Robinson[2], Lily Ray[3], Burhan A. Khan[3], Tianran Li[4], Denise A. Dillard[3], Brian D. Schilling[3], Mike Mosley[3], Patricia L. Janssen[5], Alison E. Fohner[6], Allan E. Rettie[7], Kenneth E. Thummel[1], Timothy A. Thornton[4] and David L. Veenstra[8,*]

- The primary outcome was stable warfarin dose, defined as one dose, and associated international normalized ratio (INR) within the target range, at least 6 months after starting therapy, with two matching doses at least 2 weeks apart.
- VKORC1 genotype explained 34% of dose variability,

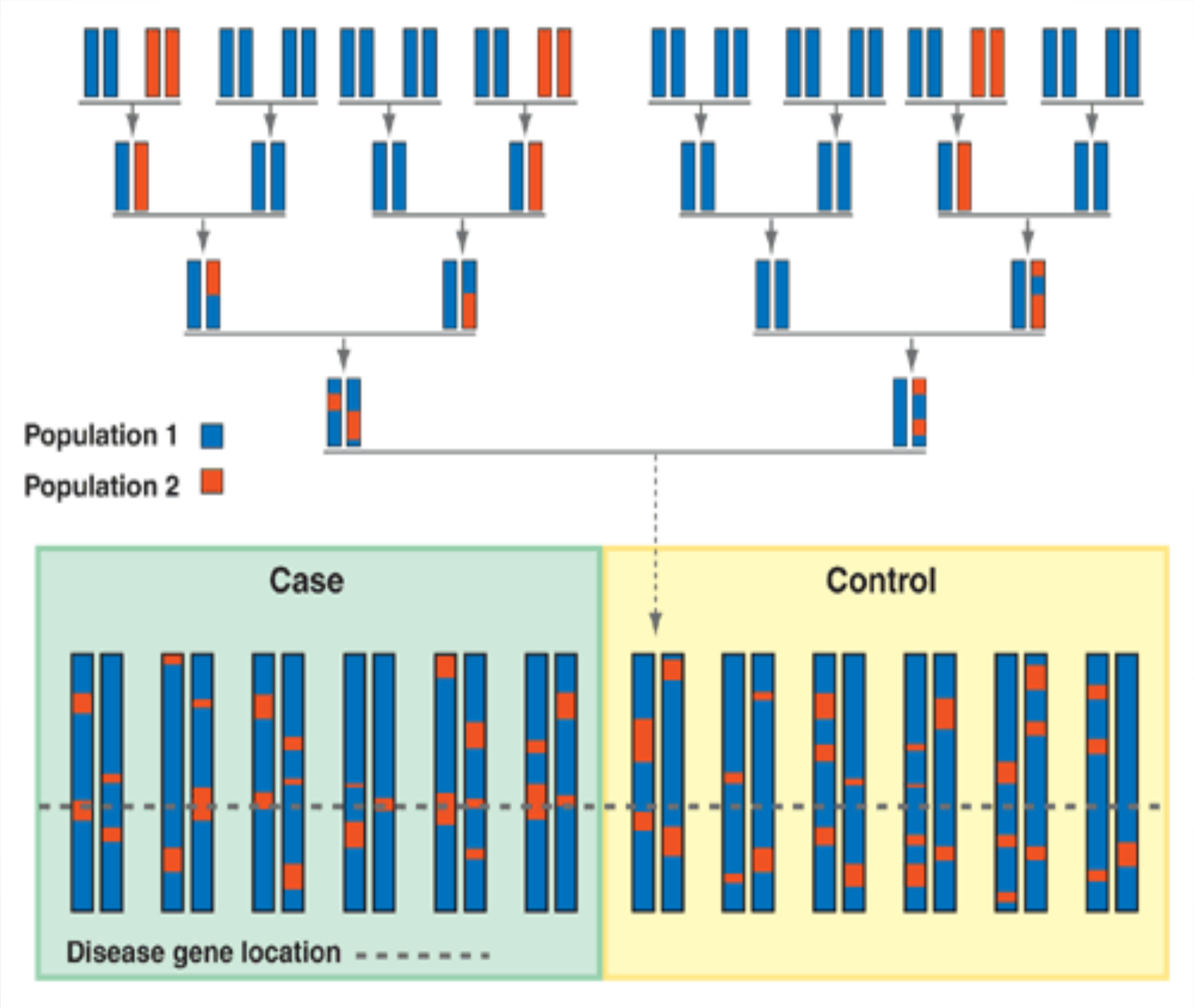




Can genetic ancestry be leveraged for new genetic insights?

- Causal variant occurs more often on segments inherited from the ancestral population with the higher disease variant frequency
- **Admixture mapping**, leverages heterogeneity in genetic ancestry of admixed individuals by identifying loci harboring variants that are
  - Associated with a trait
  - Have differing allele frequencies across the ancestral populations
- In case-control studies, admixture mapping identifies loci with significant differences in ancestry between cases and controls

Admixture Mapping



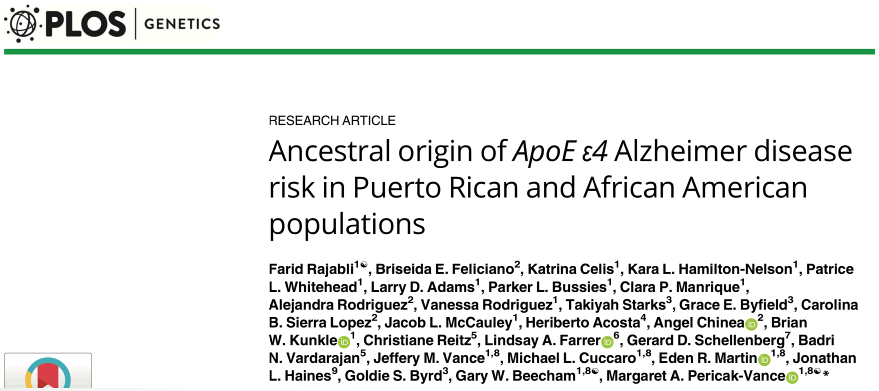
Example: *APOE* and AD Risk in Diverse Populations

- Genetic studies of Alzheimer's disease (AD) have primarily been conducted in European ancestry populations
- *APOE* has long stood at the forefront of common genetic risk factors for AD in European ancestry populations
    - *APOE ε4 and ε2* isoforms are most consistent risk and protective variants for AD in European ancestry populations
- Effects of risk and protective variants for AD identified in European populations, however, are often not transferable across populations

Example: *APOE* and AD Risk in Diverse Populations

- Conflicting evidence of associations between *APOE* and AD in ancestrally diverse populations
  - Hispanic/Latino and African American populations have *APOE* effect sizes that are orders of magnitude smaller than European ancestry populations
- Significant challenges to disentangle and understanding the complex relationships of *APOE*, AD, and ancestry
- Differential effect sizes of *APOE alleles across populations*
- *APOE ε4* allele frequency varies considerably worldwide
  - *More* common in African populations than in European ancestry populations
  - Less common in Latino/Hispanic populations

APOE Ancestral Origin

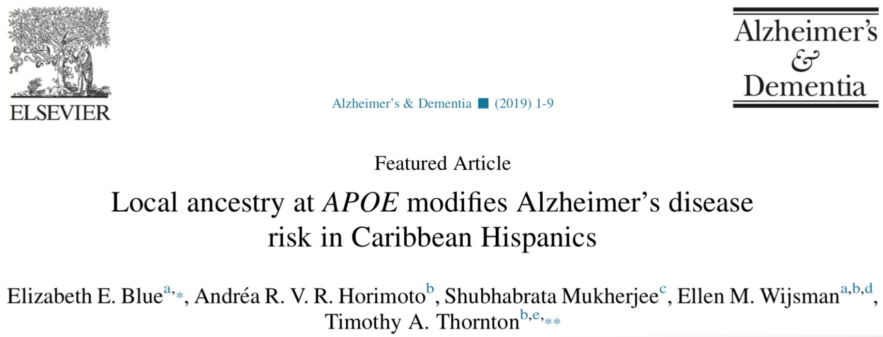


PLOS | GENETICS

RESEARCH ARTICLE

Ancestral origin of *ApoE ε4* Alzheimer disease risk in Puerto Rican and African American populations

Farid Rajabli[1☯], Briseida E. Feliciano[2], Katrina Celis[1], Kara L. Hamilton-Nelson[1], Patrice L. Whitehead[1], Larry D. Adams[1], Parker L. Bussies[1], Clara P. Manrique[1], Alejandra Rodriguez[2], Vanessa Rodriguez[1], Takiyah Starks[3], Grace E. Byfield[3], Carolina B. Sierra Lopez[2], Jacob L. McCauley[1], Heriberto Acosta[4], Angel Chinea[2], Brian W. Kunkle[1], Christiane Reitz[5], Lindsay A. Farrer[6], Gerard D. Schellenberg[7], Badri N. Vardarajan[5], Jeffery M. Vance[1,8], Michael L. Cuccaro[1,8], Eden R. Martin[1,8], Jonathan L. Haines[9], Goldie S. Byrd[3], Gary W. Beecham[1,8☯], Margaret A. Pericak-Vance[1,8☯*]



ELSEVIER

Alzheimer's & Dementia ■ (2019) 1-9

Alzheimer's & Dementia

Featured Article

Local ancestry at *APOE* modifies Alzheimer's disease risk in Caribbean Hispanics

Elizabeth E. Blue[a,*], Andréa R. V. R. Horimoto[b], Shubhabrata Mukherjee[c], Ellen M. Wijsman[a,b,d], Timothy A. Thornton[b,e,**]

- Recent studies by Rajabli et al (2018)and Blue et al (2019) and investigated *APOE* and AD risk in recently admixed populations: Caribbean Hispanics and African Americans
- Focused on ancestral origin and effects at *APOE* locus for European, African, and Native American ancestry
- Will focus on the **Blue et al (2019)** as an example of leveraging local ancestry differences for new insights into *APOE*

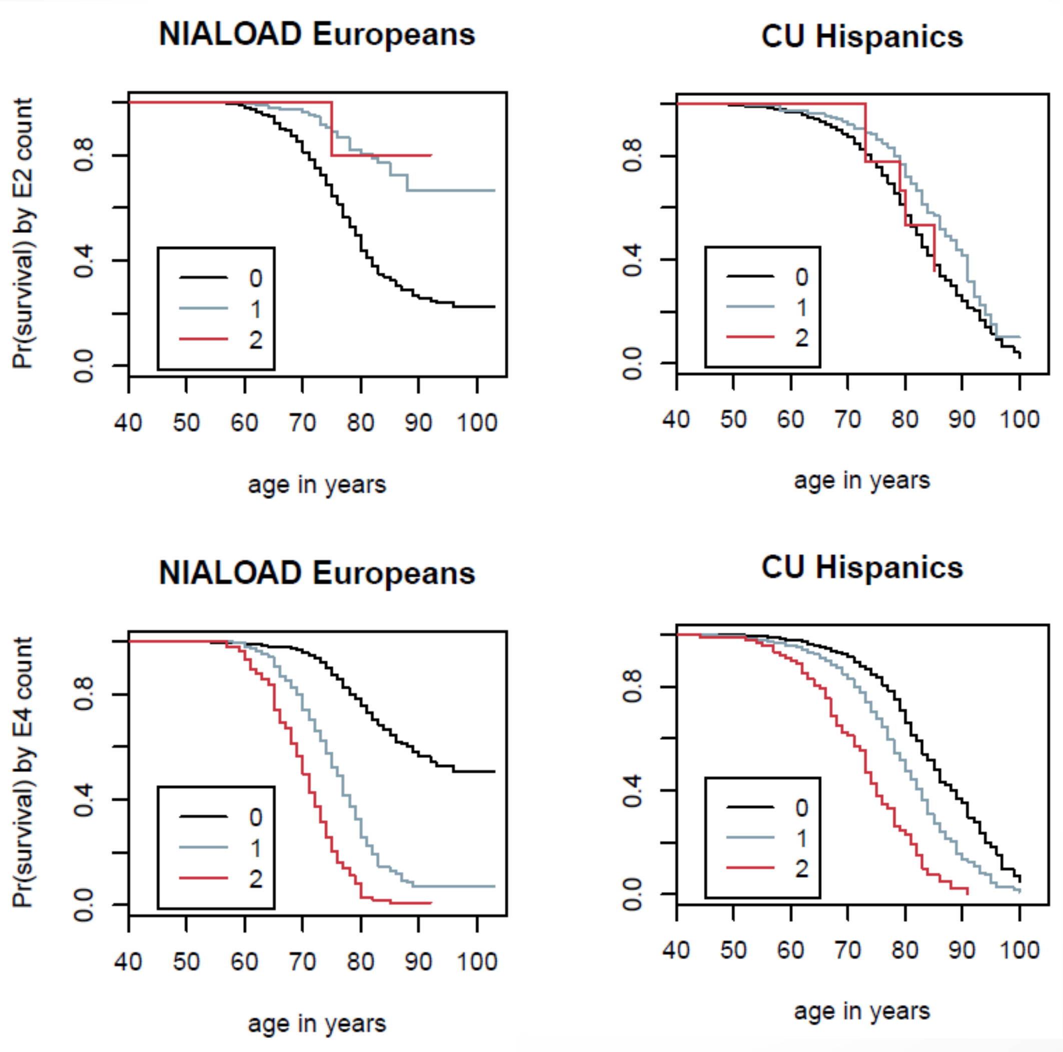
Survival Analysis of AD in Diverse Samples

- Blue et al. (2019) first investigated if protective and risk effects of APOE ε2 and ε4 dosage, respectively, on AD were similar in European Ancestry (EA) vs. Caribbean Hispanic (CH) populations
- Conducted a survival analysis and compared age of onset of AD using Cox Proportional Hazards regression in similarly sized EA and CH cohorts
- The effect of ε2 and ε4 on the hazard of AD were significantly different by multiple orders of magnitude
  - **ε2 Hazard Ratio (95%CI):** 0.28 (0.19-0.40) in Europeans vs. 0.66 (0.54-0.81) in Hispanics
  - **ε4 Hazard Ratio (95%CI):** 3.44 (3.08-3.83) in Europeans vs. 1.98 (1.80-2.19) in Hispanics

Survival analysis of AD in Diverse Samples

Survival curves by allele count

ε2



ε4

[Blue *et al.* (2019) *Alzheimer's & Dementia* **15** (12): 1524-1532]

Risk of AD and Global Ancestry

- Blue et al. (2019) also found differences in APOE ε4 effects on AD risk in EA and CH populations for AD:
  - Europeans: Odds Ratio = 15.89
  - Hispanics: Odds Ratio = 4.59
- However, genome-wide average ancestry proportions couldn't explain differences in AD effects
- They next focused on investigating the role that ancestry of origin of APOE may in effects.
- Identified a subset of CH individuals and compared odds of AD for individuals who were homozygous for either E3 or E4 and had either 100% EUR or 100% AFR ancestry of origin at *APOE*

APOE Ancestry of origin and AD risk

- Local ancestry at the *APOE* locus influences AD risk independently of ε2/ε3/ε4 genotype in CH
  - Those inheriting *APOE* alleles on an African haplotype are associated with **39% lower odds of AD** compared to those inheriting alleles on a European haplotype

Covariate	Odds Ratio (95% CI)
APOE ε4/ε4 genotype	8.5935 (4.49-16.43)
% global AMR ancestry	1.0334 (1.00-1.06)
% global AFR ancestry	1.0042 (0.99-1.01)
local AFR ancestry at *APOE*	**0.6058 (0.38-0.97)**
Age (years)	1.0245 (1.01-1.04)
Logistic regression for AD among subjects homozygous for either African- or European-derived ε3 or ε4 alleles in the Caribbean Hispanics	

Blue *et al.* (2019) *Alzheimer's & Dementia* **15** (12): 1524-1532

APOE Ancestry of origin and AD risk

- Blue et al. (2019) provided significant evidence that there is more to the relationship between *APOE* and AD than the missense variants defining the ε2/ε3/ε4 alleles.
- These missense variants are believed to be the functional variants driving the association between *APOE* and AD risk and age-at-onset.
- However, since local ancestry near *APOE* is also associated with AD risk after adjustment for *APOE* genotype indicates that additional coding and/or non-coding variants which differ in frequency between Europeans and Africans also influence an individual's risk of AD.

Future of Genetic Studies in Diverse Populations

- Many variants previously identified in European ancestry populations have been found in to either (1) not be transferable across multi-ethnic populations or (2) have significantly smaller effects
- Little is known about the biological mechanisms contributing to this phenomenon, and significantly more and larger genetics studies in ancestrally diverse populations are needed
- Understanding the role that genetic ancestry plays, both locally and globally, in health and health disparities will be critical in the effectiveness of precision or personalized medicine in ancestrally diverse populations