New approaches for complex trait genetic mapping in ancestrally diverse populations

Timothy Thornton, PhD
Robert W. Day Endowed Professor of Public Health
Associate Professor
Department of Biostatistics
University of Washington

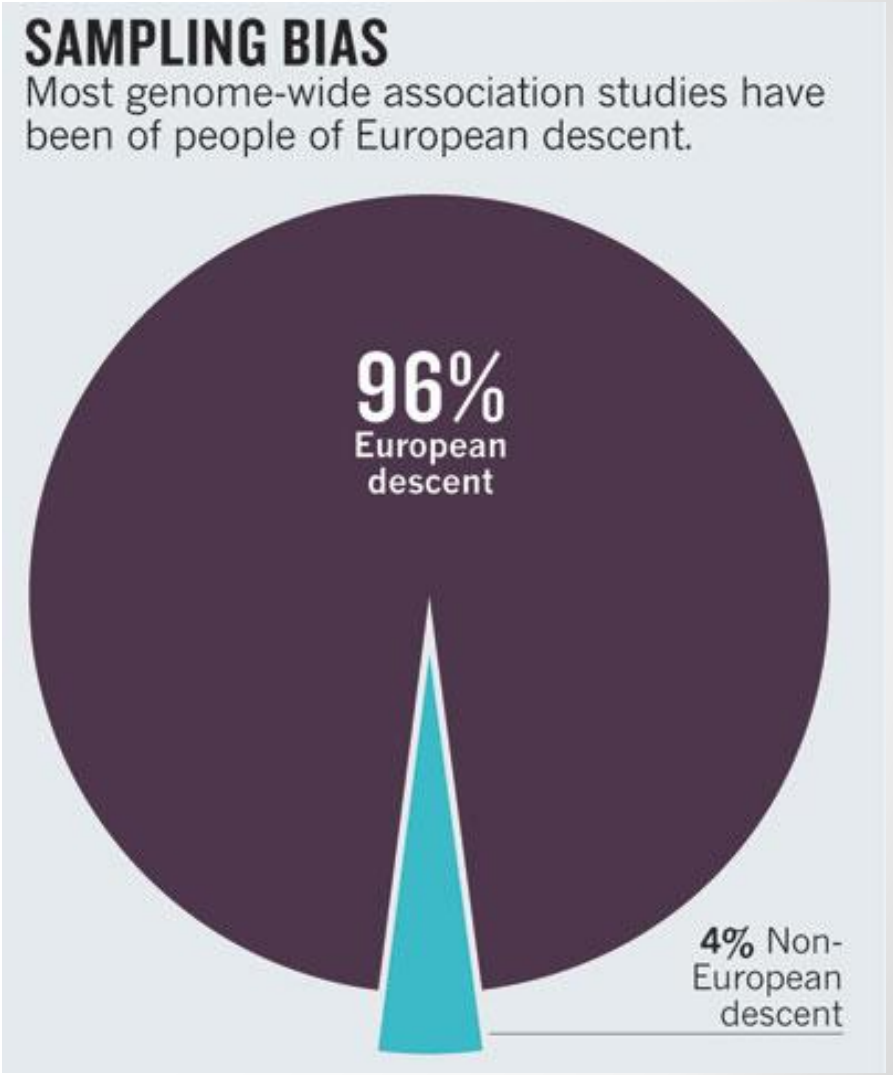
September 29, 2016

SCHOOL OF PUBLIC HEALTH

UNIVERSITY *of* WASHINGTON

Introduction

- To date, more than a million individuals have been included in GWAS and sequencing association studies for the mapping of complex traits.
- The vast majority of these studies, however, have been conducted in populations of European ancestry
- Nature paper (in press) by Dr. Malia Fullerton and Alice Popejoy (Univ. of Wash) show only a 2% increase over the last 5 years.
-



Bustamante et al. (Nature, 2011)

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)

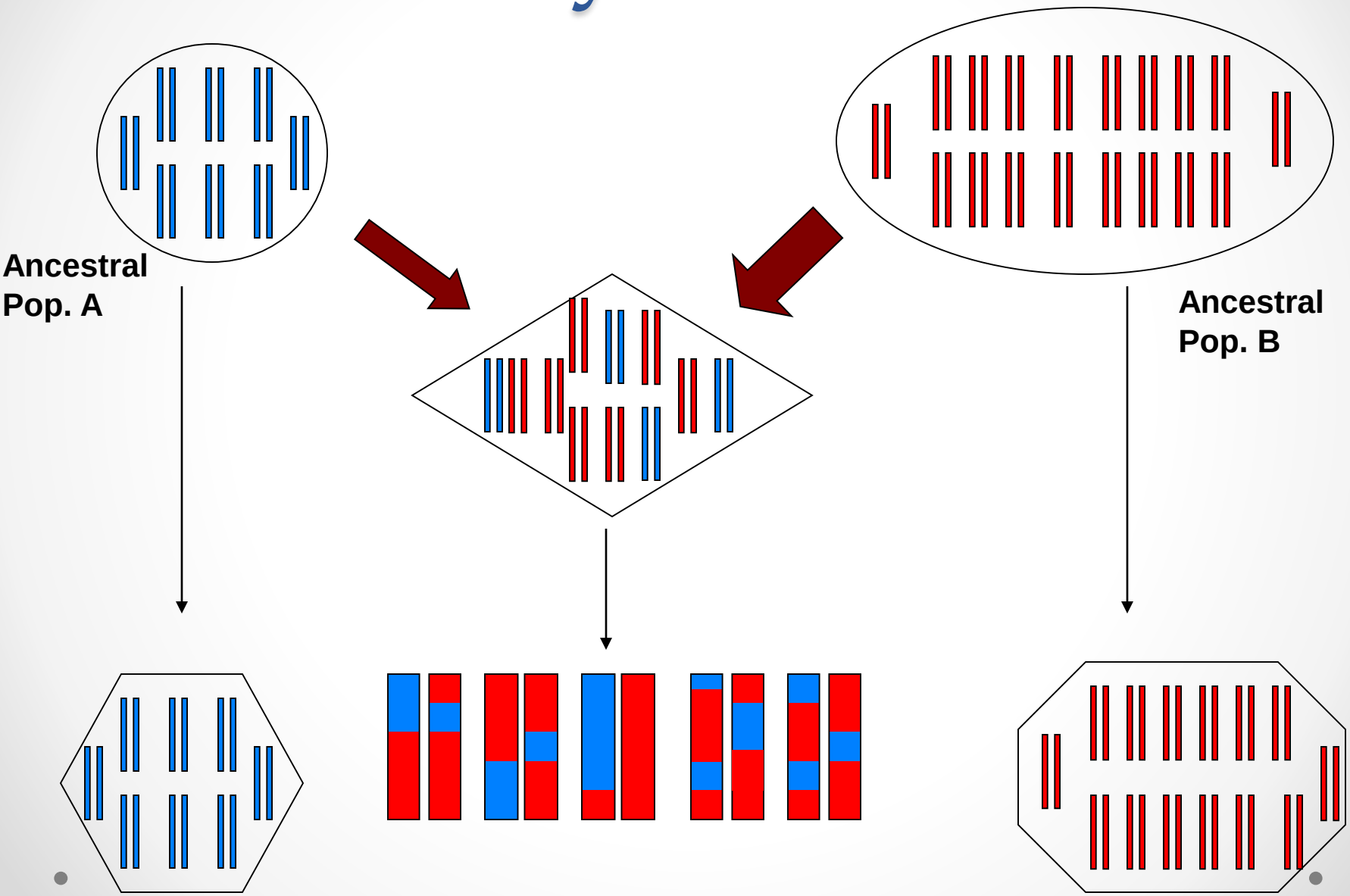
Genetic Studies in Diverse Populations

- Recently, there has been an increased interest in complex trait mapping in diverse populations
- Trans-Omics for Precision Medicine (TOPMed) Program recently funded by NHLBI
  - Whole-genome-sequence (WGS) data currently being generated for over 100,000 individuals
  - Multi-ethnic cohort includes European Americans, African Americans, Hispanics/Latinos, and Samoans.
- NIH launched the Precision Medicine Initiative (PMI) in 2015
  - PMI Cohort Program will build a large research cohort of **one million or more** Americans
  - Goal is to support and advance the targeted prevention and treatment strategies that take an individual's unique characteristics into account, including individual genome sequences, environmental factors and lifestyles.

Admixed Populations

- Populations who have experienced admixing among continentally divided ancestral populations within the past 200 to 500 years.
- Admixed populations have largely arisen 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 Hispanics in the U.S
  - Latinos from throughout Latin America
  - Uyghur population of Central Asia
  - Cape Verdeans
  - South African "Coloured" population

Ancestry Admixture



HCHS/SOL

- The Hispanic Community Heath Study / Study of Latinos (HCHS/SOL) is the largest epidemiological study of U.S. Hispanics/Latinos
- Initiated in 2006; funded by NHLBI & several other NIH institutes
- Recruited 16,415 men and women who self-identify as Hispanic or Latino
- Aged 18-74 years, multiple household members eligible
- Sample households in defined communities in Bronx, Chicago, Miami, and San Diego
- >2,000 subjects each of the following origins: Mexican, Puerto Rican and Dominican, Cuban, Central and South American
- Investigate prevalence and risk factors for (among others): heart, lung and blood disorders kidney and liver function, diabetes, cognitive function, dental and periodontal conditions, hearing disorders, sleep apnea

HCHS/SOL

- Baseline exam at field center lasting ~ 7 hrs
- Recruitment occurred over a 3-year period (2008-2011)
- Second in person visit (6 yr interval) in progress (2014-2017)
- Funding runs through 2019 for event follow-up

- **Events:** Annual phone call to ascertain hospitalizations or other significant clinical events
- Medical records for events are obtained, reviewed and adjudicated

- **Design paper Reference**: Sorlie *Annals of Epidemiology* 2010

Phenotypic Data at Baseline

Questionnaires
Health and Medical History
Family History
Acculturation
Social and Behavioral
Occupational
Health Care Access
24-Hour Dietary Recall and food propensity questionnaire
Smoking
Alcohol Consumption
Physical Activity
Disability
Weight Loss/Gain
Sleep
Medication
Oral/Dental Health
Hearing

Medical Examinations
Blood Pressure
Pulmonary Function
Sleep Assessment
ECG
Anthropometry
Dental
Audiometry
Accelerometry/Physical Activity

Specimen Collection
Fasting Blood
2 hour oral (75g) Glucose Tolerance Test
Spot Urine
Storage of additional blood and urine

Laboratory Measurements
Lipids
Glucose
Insulin
Glycosylated hemoglobin
Iron
Creatinine
Cystatin C
ALT
AST
GGT
Ferritin
CRP
UIBC
CBC (w/ differentials)
Serology for Hep-A,-B,-C
Albumin (urine)
Creatinine (urine)

Demographic and Socioeconomic Characteristics of HCHS/SOL

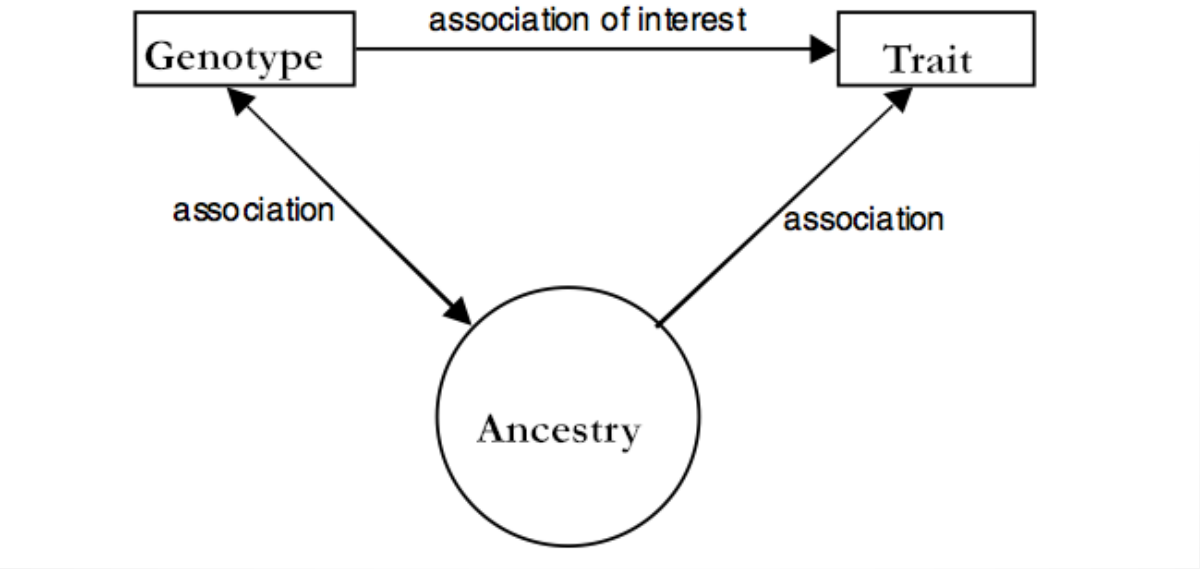
Characteristics (mean or %)	ALL	Cuban	Domin.	Mexico	Puerto Rican	Cent. Am.	So. Am.
Unweighted N	15,079	2,201	1,400	6,232	2,590	1,634	1,022
Age (yrs)	43.2	43.5	43.1	43.0	43.2	43.4	43.2
Men	40.1	46.8	34.6	37.9	41.8	39.4	40.8
US Residence >10 years	69.5	45.0	73.6	73.2	92.7	62.6	53.9
Language Preferred – Spanish	77.5	91.9	80.4	81.4	42.7	89.0	89.9
College degree	15.3	20.2	15.6	12.4	14.5	14.8	22.4
Annual family income >$50K	11.4	8.2	7.2	14.0	14.0	7.2	11.6
Health insurance	50.9	40.0	72.3	44.7	77.3	34.4	41.9

Genetic Studies in Ancestrally Diverse Populations: Opportunities and Challenges

- Opportunities:
  - Identification of novel genetic variants underlying phenotypic diversity and health disparities among populations.
  - Potential to provide new insights for health disparities of minority populations for many complex diseases
- Challenges for complex trait mapping:
  - Heterogeneous genetic background
  - Confounding due to population stratification
  - Familial structure and/or cryptic relatedness

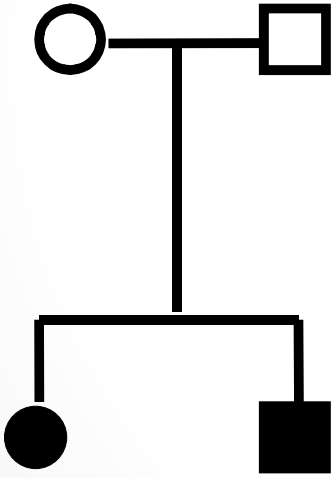
Confounding due to Admixed Ancestry

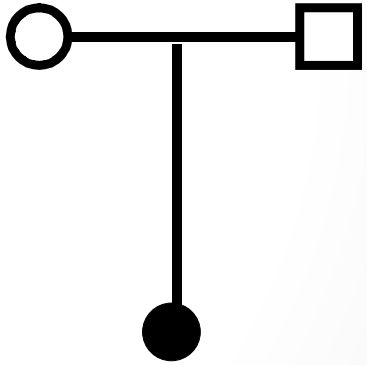
- Ethnic groups (and subgroups) often share distinct dietary habits and other lifestyle characteristics that leads to many traits of interest being correlated with ancestry and/or ethnicity.



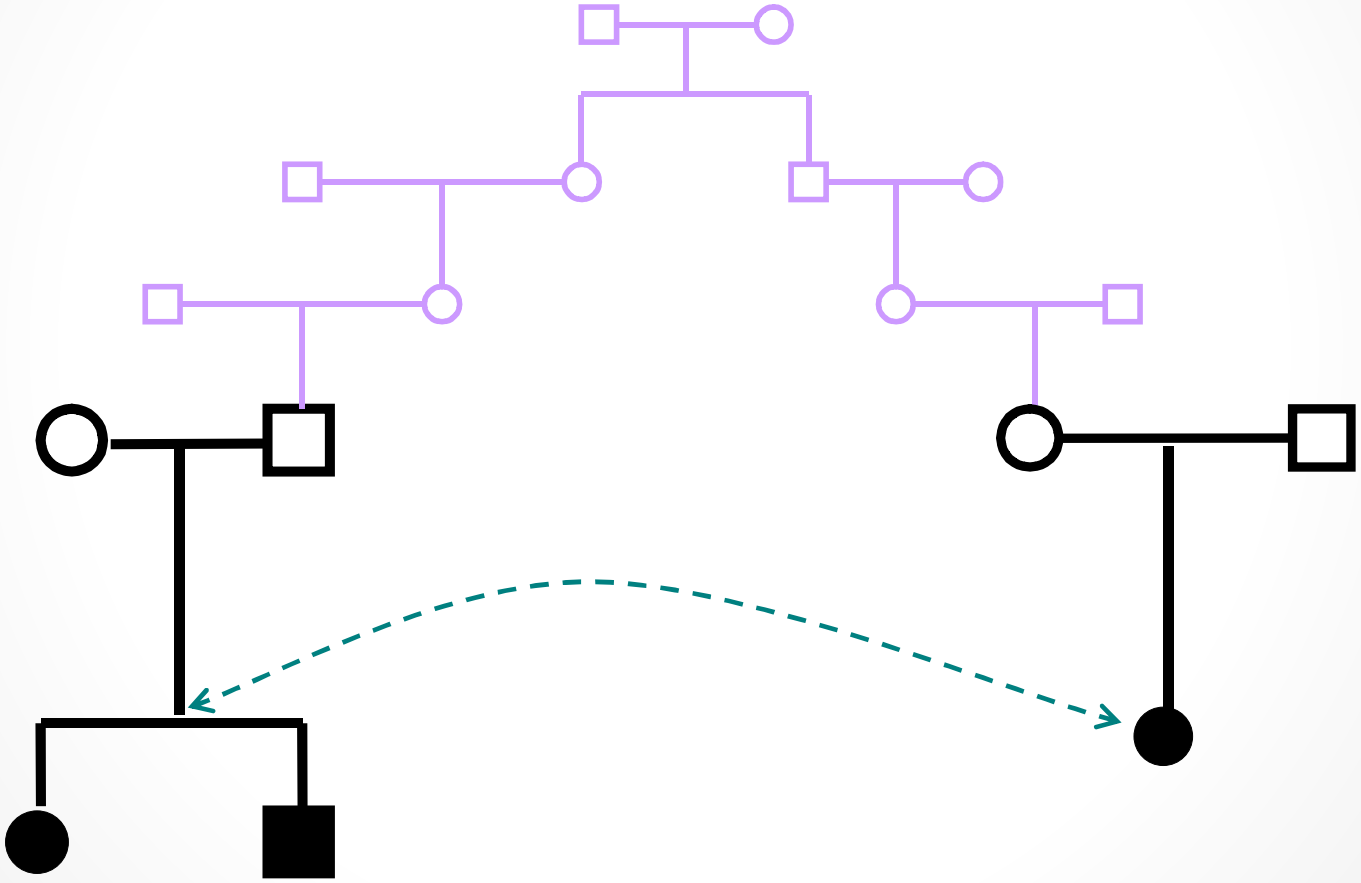
Cryptic Relatedness

- Failure to account for relatedness among sample individuals can lead to spurious association





Cryptic Relatedness

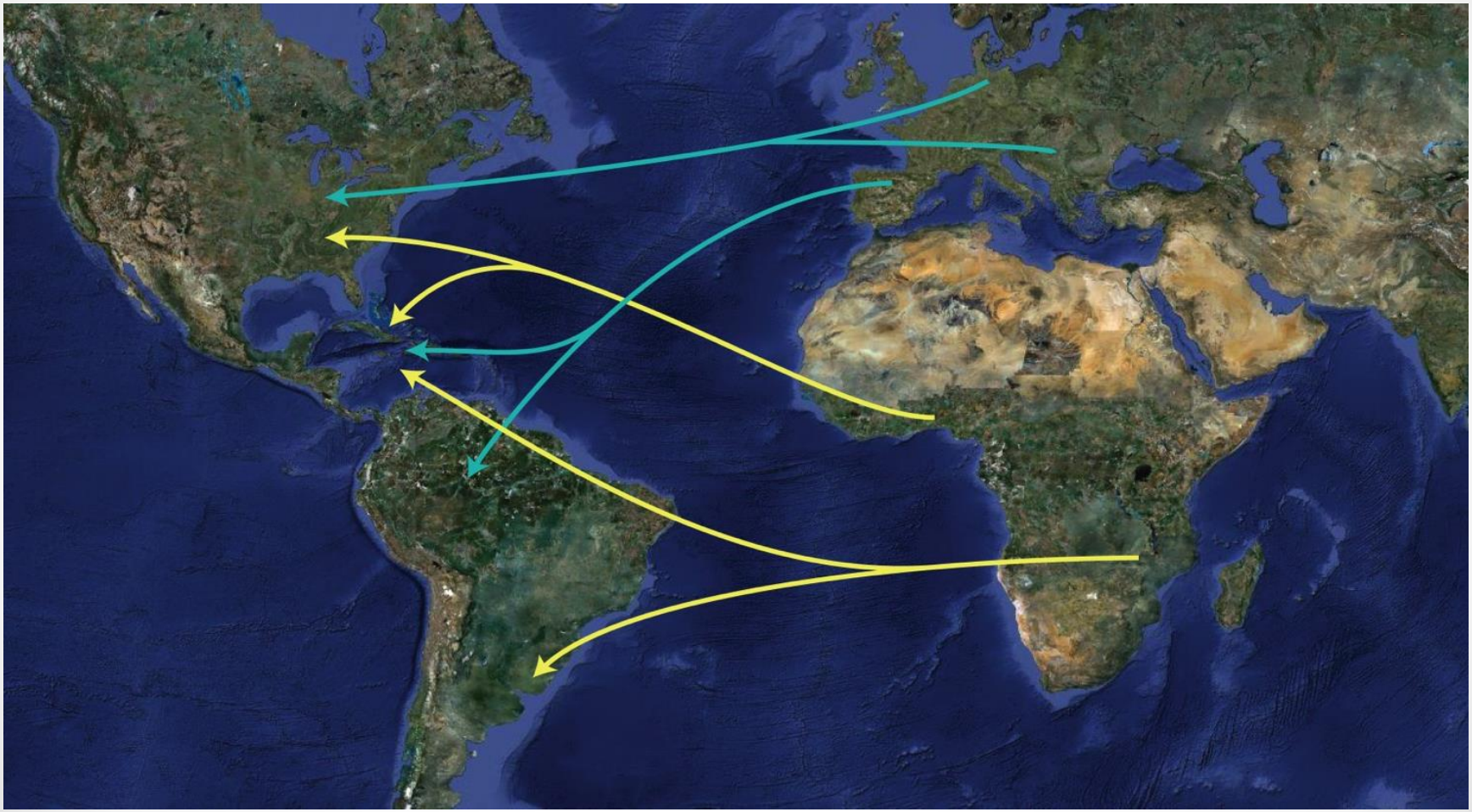


Genetic Relatedness in Admixed Populations

- The genealogy of individuals in a sample consists of:
  - Distant genetic relatedness, such as population structure
  - Recent genetic relatedness: pedigree relationships of close relatives

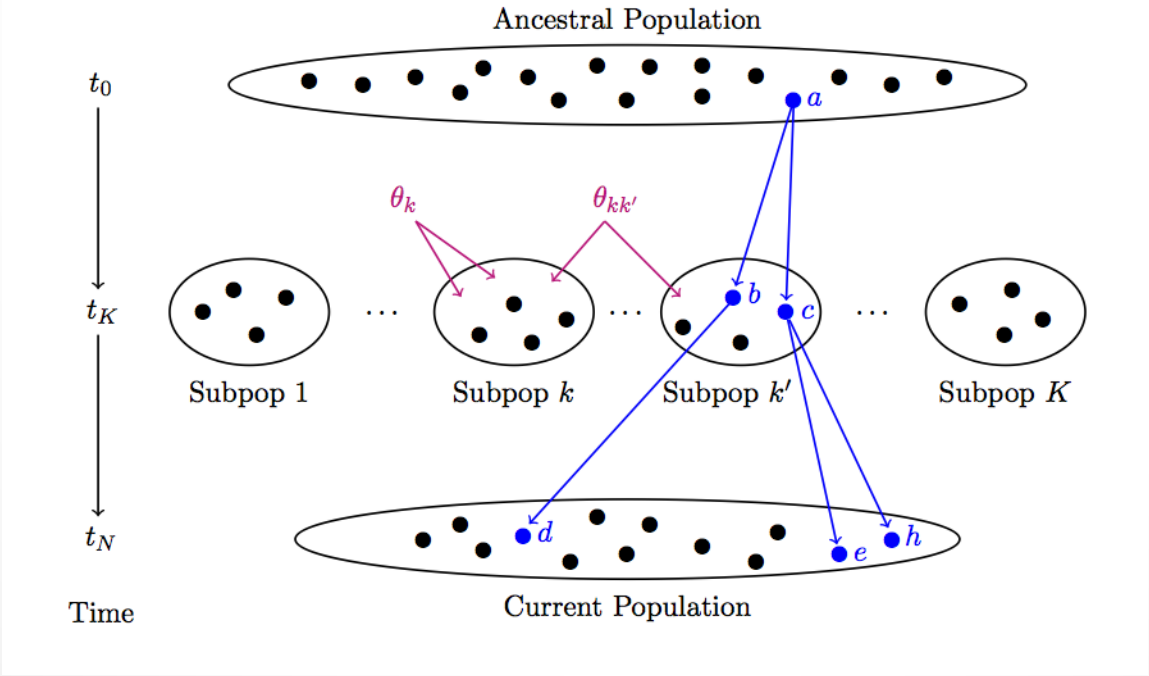
- Samples from admixed populations often have complex genealogy due to ancestry admixture and both recent and distant genetic relatedness

Complex Genealogy of Racially Admixed Populations



Recent versus Distant Genetic Relatedness

- Distinguishing familial relatedness from ancestry using genotype data in diverse populations is difficult, as both manifest as genetic similarity through the sharing of alleles.



Conomos et al. (AJHG, 2016)

Deconvolution of Genetic Relatedness

- Conomos et al. [Am J Hum Genet, 2016]

ARTICLE

Model-free Estimation of Recent Genetic Relatedness

Matthew P. Conomos,[1,*] Alexander P. Reiner,[2,3] Bruce S. Weir,[1] and Timothy A. Thornton[1,*]

- Conomos et al. [Genet Epidemiol, 2015]

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]*

- Thornton et al. [Am J Hum Genet, 2012]

ARTICLE

Estimating Kinship in Admixed Populations

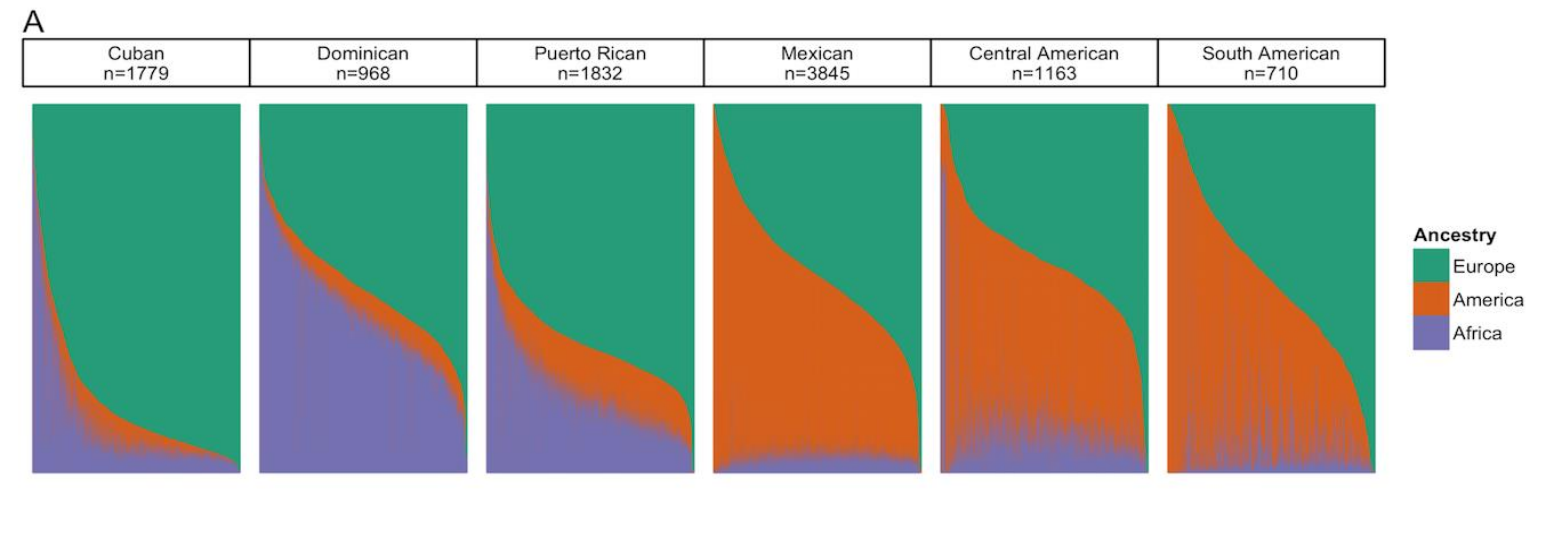
Timothy Thornton,[1,*] Hua Tang,[2] Thomas J. Hoffmann,[3,4] Heather M. Ochs-Balcom,[5] Bette J. Caan,[6] and Neil Risch[3,4,6,*]

Genetic Diversity in HCHS/SOL

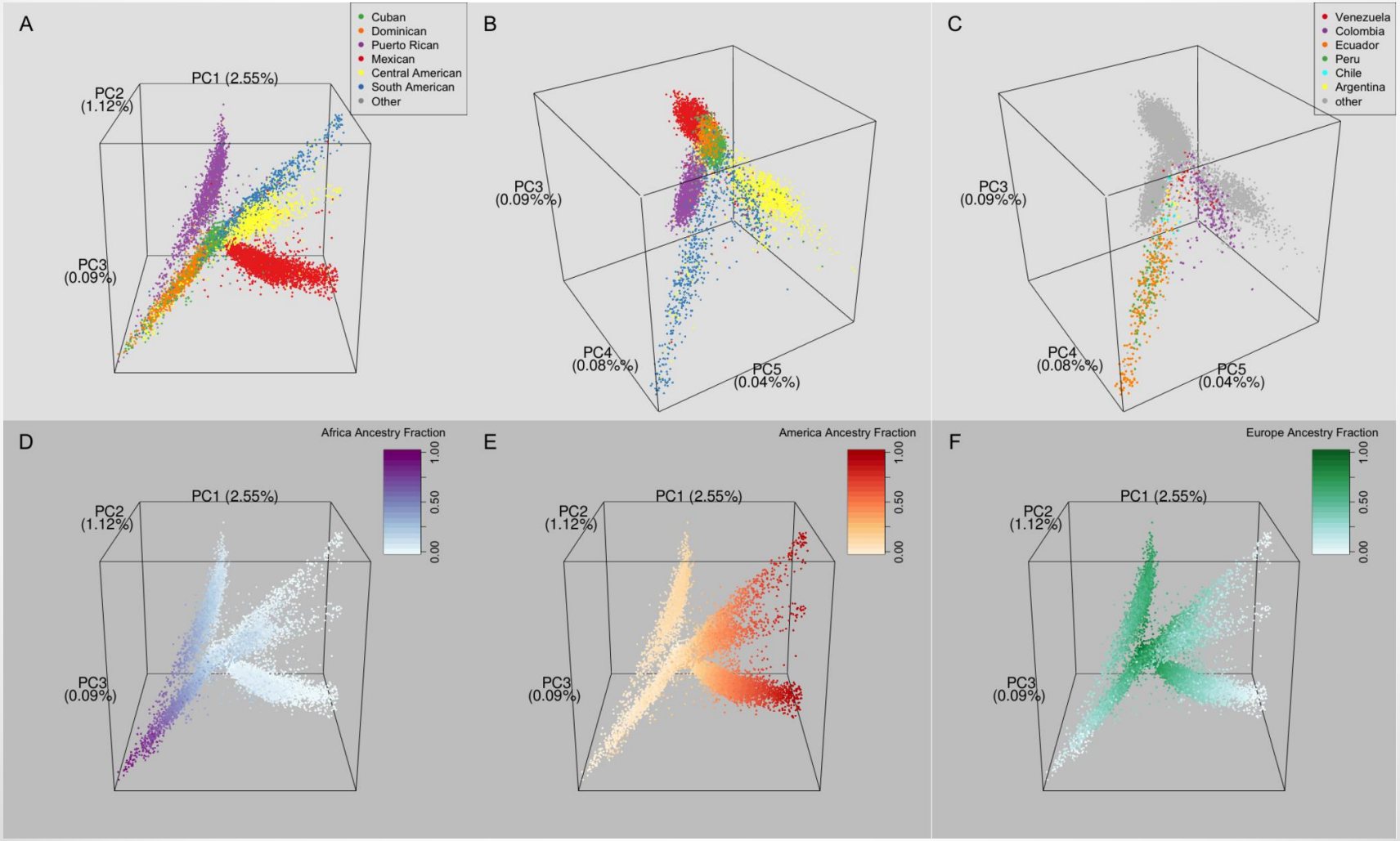
ARTICLE

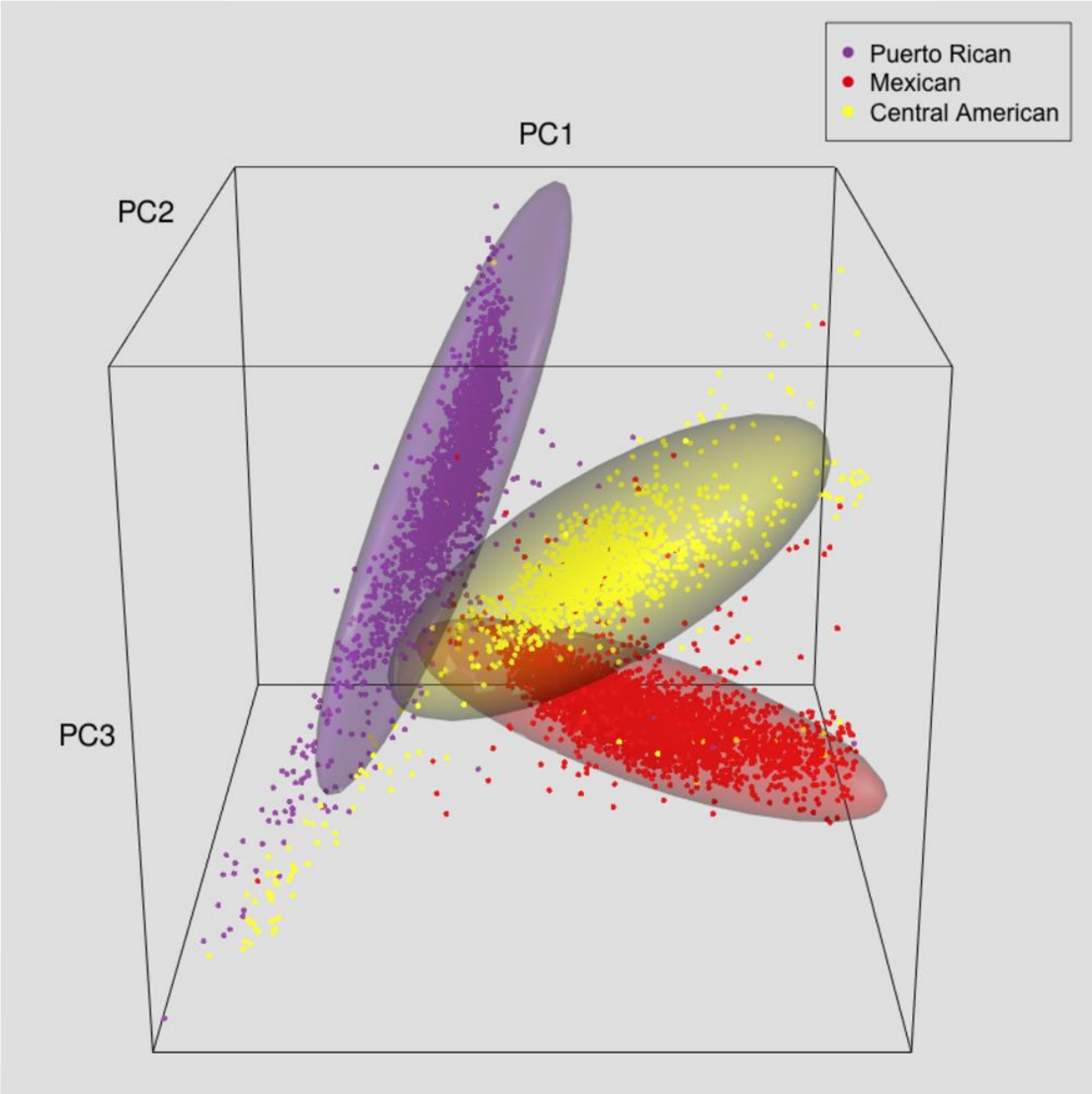
Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

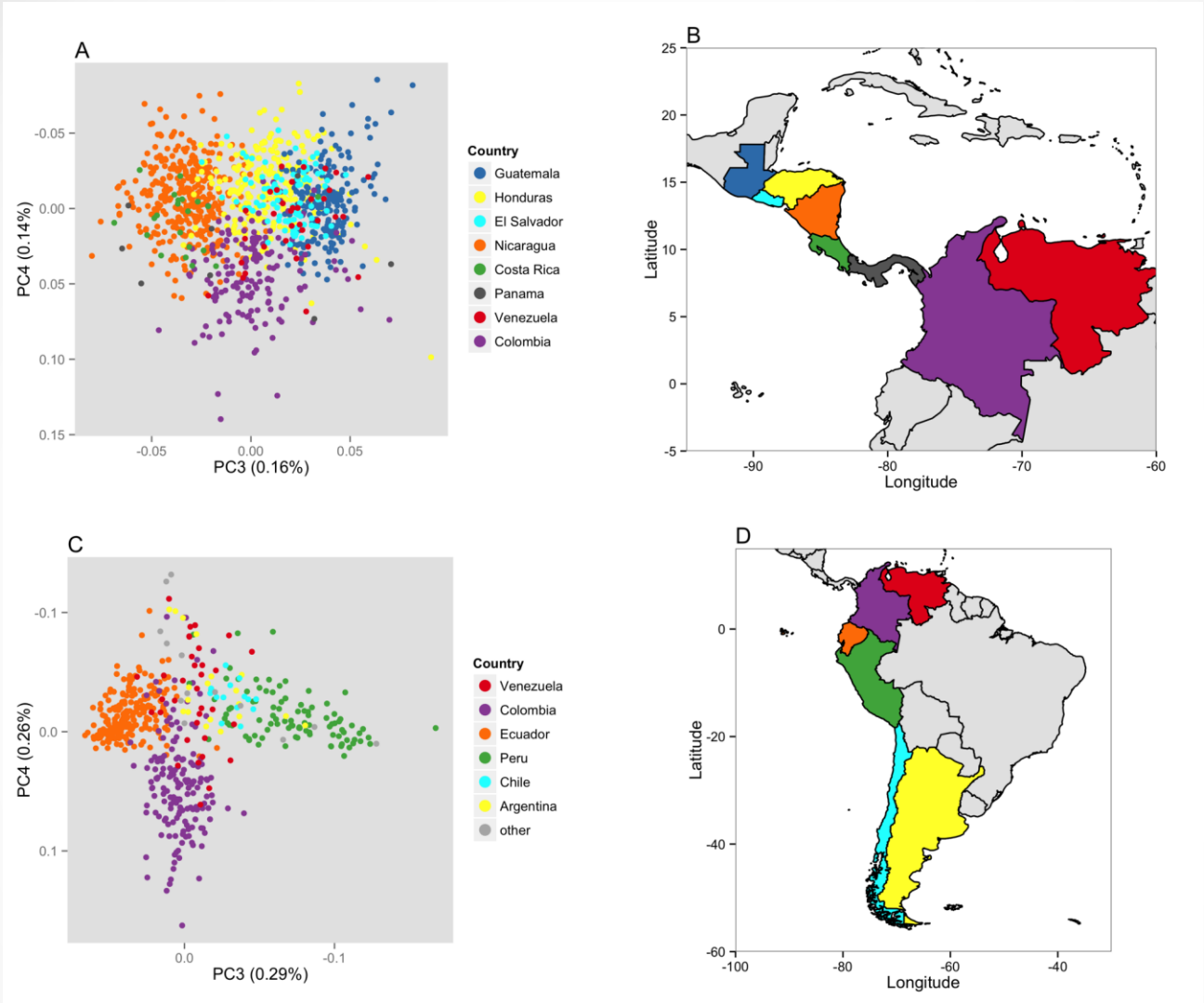
Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent 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,*]



- Conomos et al. (2016) "Genetic Diversity and Association Studies in U.S. Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos"

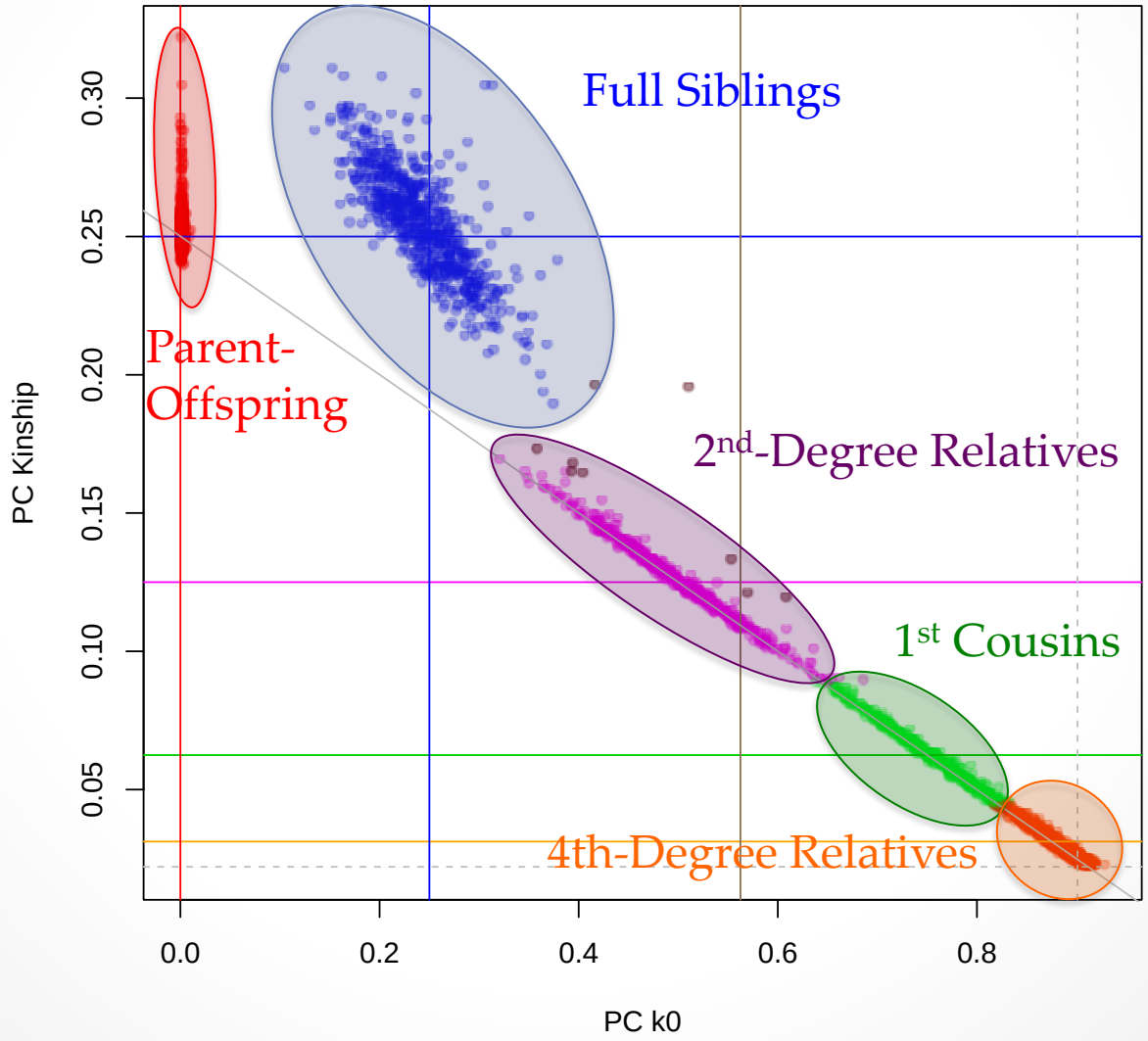



Puerto Rican
Mexican
Central American
PC1
PC2
PC3



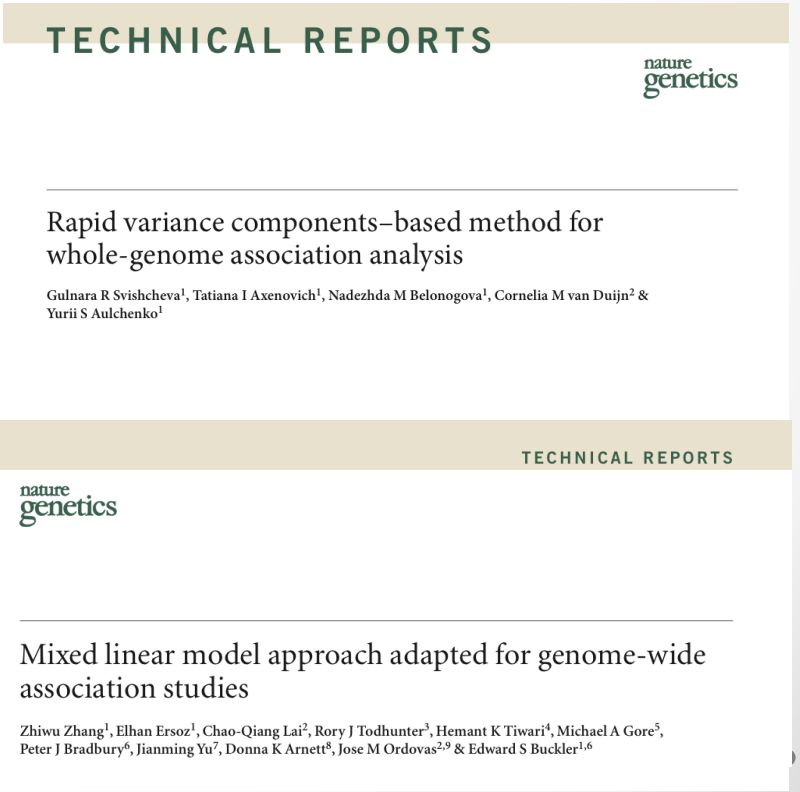
- Genetic differentiation among individuals is associated with the geography of their countries of grandparental origin.
- Plots of PCs from analyses using individuals for whom all four grandparents were born in a specific country in Central or South America show geographic structure

Recent Genetic Relatedness Inference with PC-Relate in HCHS-SOL



Linear Mixed Models for GWAS

- Linear mixed models (LMMs) have emerged as a powerful and effective approach for genetic association testing in the presence of sample structure



Association Mapping in Multi-Ethnic Populations

- We (Conomos, Reiner, McPeek, Thornton) developed the a new linear mixed model method for association mapping in diverse populations
- **LMM-OPS**, linear mixed models with orthogonal partitioned structure
- Appropriately accounts for the complex genealogy of admixed individuals by partitioning sample structure into two orthogonal components:
  1. a component for the sharing of alleles inherited identical by descent (IBD) from recent common ancestors, which represents familial relatedness
  2. and another component for allele sharing due to more distant common ancestry, which represents population structure.

Genomic Control Inflation Evaluation of LMMs

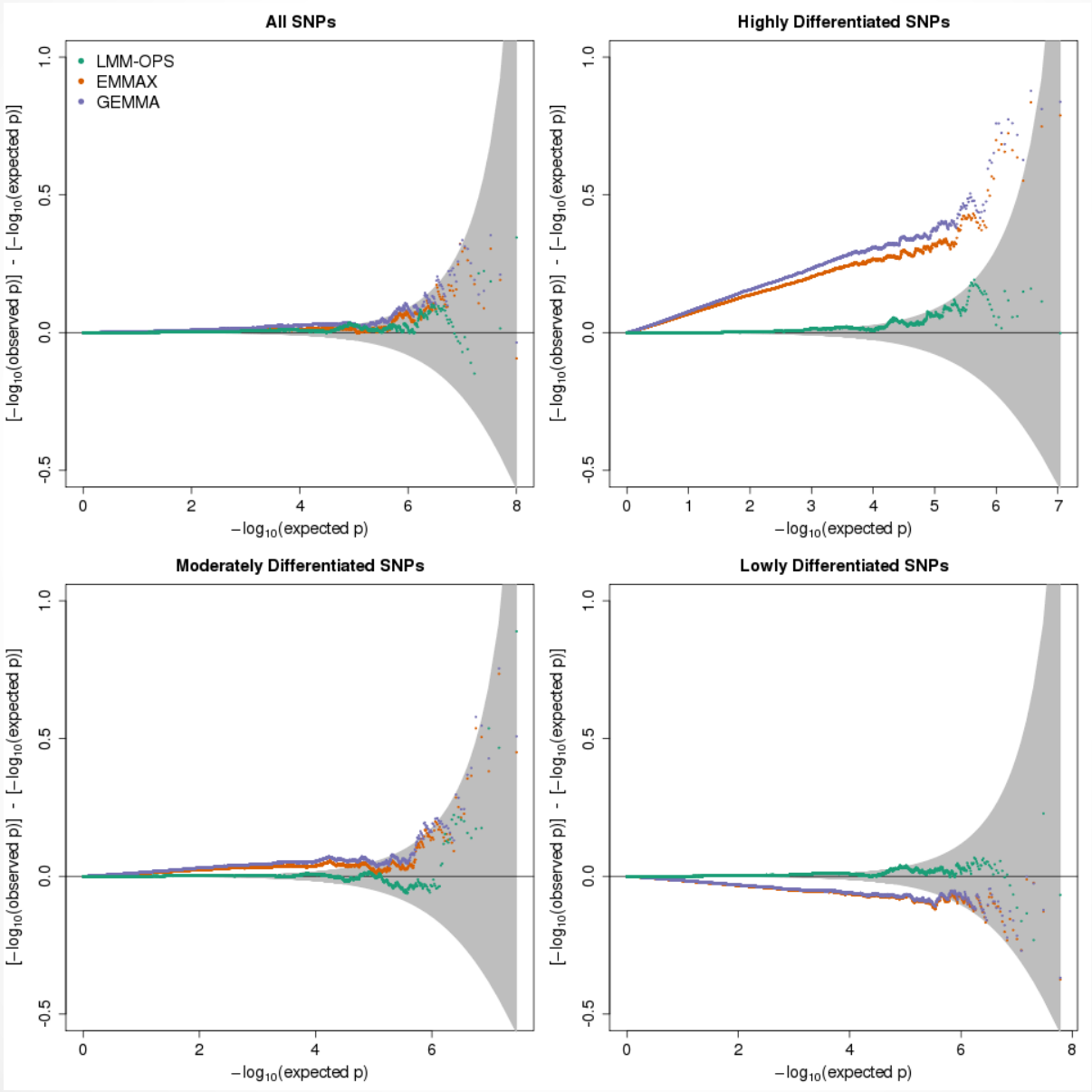
Method	Genome-Wide	Highly[a] Differentiated	Moderately[b] Differentiated	Weakly[c] Differentiated
LMM-OPS	1.000 (0.0002)	0.999 (0.0007)	1.001 (0.0004)	1.001 (0.0003)
EMMAX	1.001 (0.0002)	1.098 (0.0011)	1.016 (0.0004)	0.979 (0.0003)
GEMMA	1.004 (0.0002)	1.110 (0.0011)	1.020 (0.0005)	0.980 (0.0003)
Linear Reg. with PCs	1.026 (0.0006)	1.025 (0.0009)	1.027 (0.0007)	1.026 (0.0006)

[a] Highly differentiated SNPs: $D_s \geq 0.4$ between the two populations

[b] Moderately differentiated SNPs: $0.4 > D_s \geq 0.2$ between the two populations

[c] Weakly differentiated SNPs: $D_s < 0.2$ between the two populations

Type-I Error Evaluation of LMMs



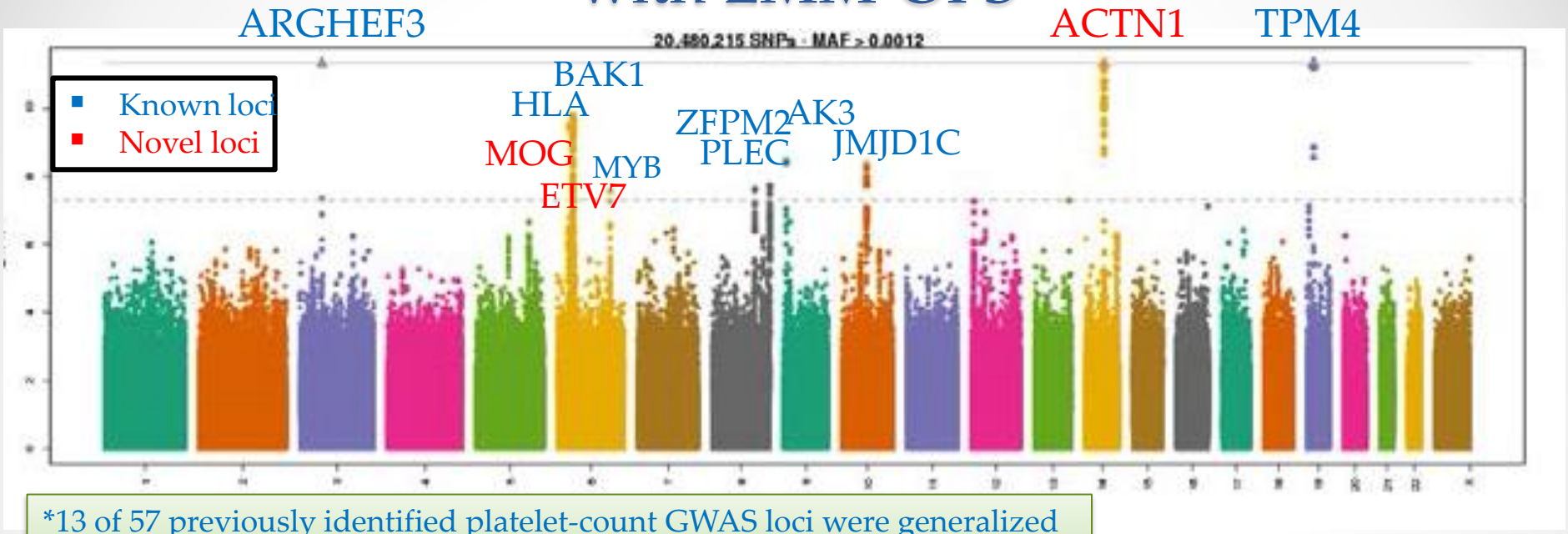
Genome-wide Allele Frequency Differentiation: HapMap Populations

	CEU	TSI	CHD	JPT	LWK	YRI
CEU	-	0.00	0.047	0.048	0.084	0.095
TSI	0.001	-	0.047	0.049	0.080	0.092
CHD	0.208	0.208	-	0.000	0.111	0.121
JPT	0.209	0.209	0.003	-	0.112	0.122
LWK	0.254	0.251	0.261	0.261	-	0.000
YRI	0.262	0.260	0.266	0.267	0.004	-

The upper half of the table gives the proportion of SNPs highly differentiated ($|D_s| \geqslant 0.4$) between the two populations. The lower half of the table gives the proportion of SNPs moderately differentiated ($0.4 > |D_s| \geqslant 0.2$) between the two populations.
CEU: Utah residents with Northern and Western European ancestry from the CEPH collection ($n = 165$)
TSI: Toscans in Italy ($n = 88$)
CHD: Chinese in Metropolitan Denver, Colorado ($n = 85$)
JPT: Japanese in Tokyo, Japan ($n = 86$)
LWK: Luhya in Webuye, Kenya ($n = 90$)
YRI: Yoruba in Ibadan, Nigeria ($n = 172$)

Genetic Association for Platelet Count in HCHS/SOL*

with LMM-OPS



*13 of 57 previously identified platelet-count GWAS loci were generalized to SOL

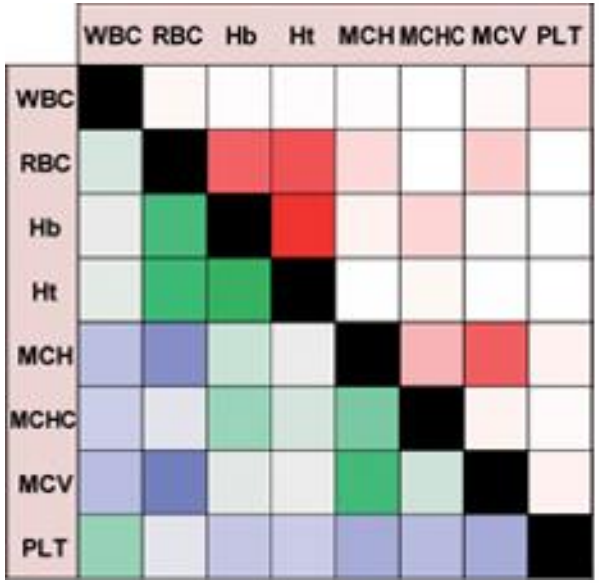
ARTICLE

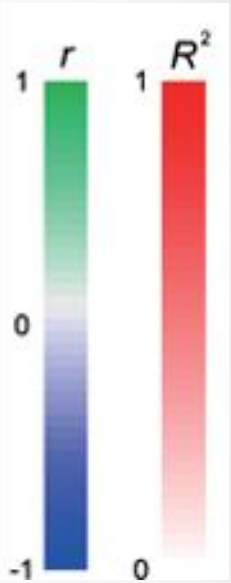
Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans

Ursula M. Schick,[1,2,3,16] Deepti Jain,[4,16] Chani J. Hodonsky,[5,16] Jean V. Morrison,[4] James P. Davis,[6] Lisa Brown,[4] Tamar Sofer,[4] Matthew P. Conomos,[4] Claudia Schurmann,[2,3] Caitlin P. McHugh,[4] Sarah C. Nelson,[4] Swarooparani Vadlamudi,[6] Adrienne Stilp,[4] Anna Plantinga,[4] Leslie Baier,[7] Stephanie A. Bien,[1] Stephanie M. Gogarten,[4] Cecelia A. Laurie,[4] Kent D. Taylor,[8,9] Yongmei Liu,[10] Paul L. Auer,[11] Nora Franceschini,[5] Adam Szpiro,[4] Ken Rice,[4] Kathleen F. Kerr,[4] Jerome I. Rotter,[8] Robert L. Hanson,[7] George Papanicolaou,[12] Stephen S. Rich,[13,14] Ruth J.F. Loos,[2,3,15] Brian L. Browning,[4] Sharon R. Browning,[4] Bruce S. Weir,[4] Cathy C. Laurie,[4] Karen L. Mohlke,[6] Kari E. North,[5,16] Timothy A. Thornton,[4,16] and Alex P. Reiner[1,16,*]

Blood count phenotypes available in HCHS/SOL

- Red blood cell
  - Hemoglobin/Hematocrit
  - Red blood cell indices (MCH, MCHC, MCV), RBC count, red cell distribution width (RDW)
- White blood cell (WBC) count and subtypes
- Platelet count



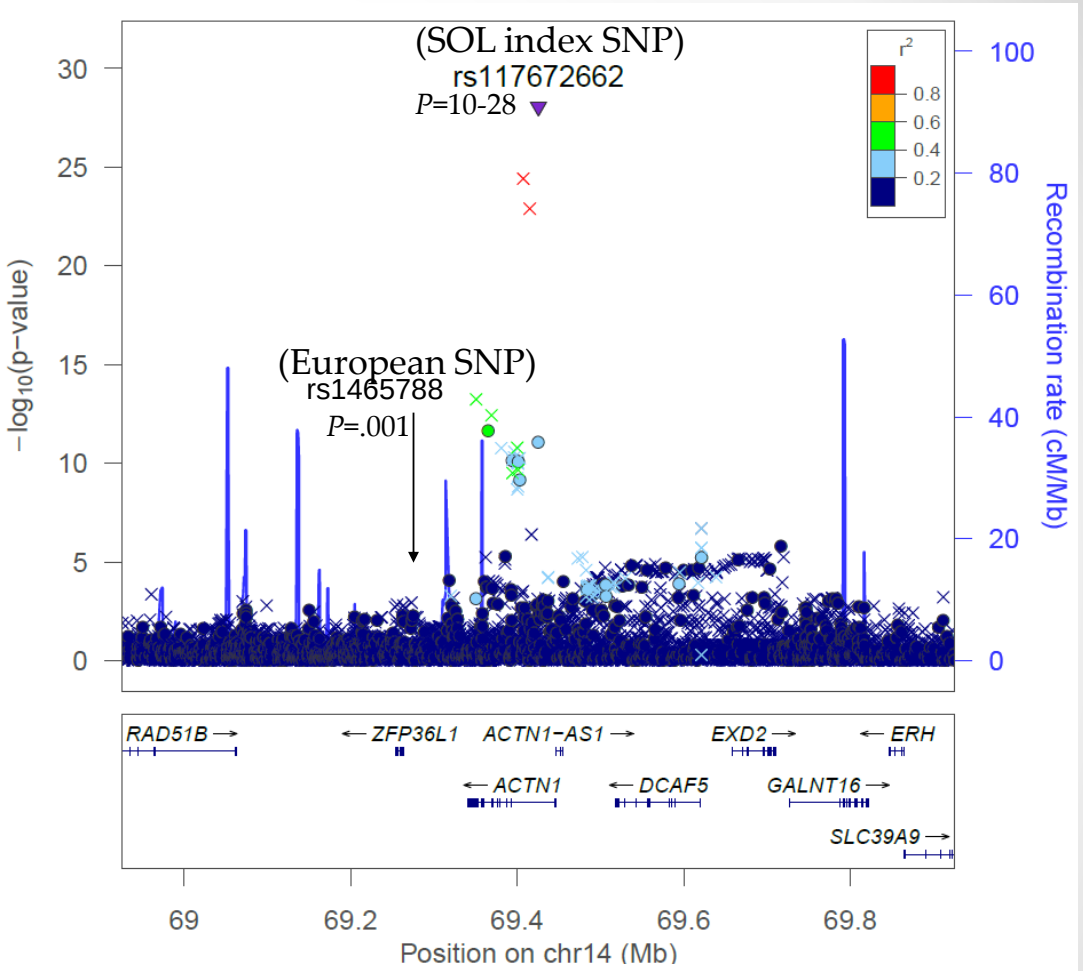


Index	Description	Calculation
MCV	Mean corpuscular volume	Hct / RBC x 10
MCH	Mean corpuscular hemoglobin	Hgb / RBC x 10
MCHC	Mean corpuscular hemoglobin concentration	Hgb / Hct * 100

RBC - red blood cell volume in millions per microliter (10^6 / µL)
Hct - hematocrit (packed cell volume) in percent
Hgb - hemoglobin in grams per deciliter (g/dL)

ACTN1 and platelet count

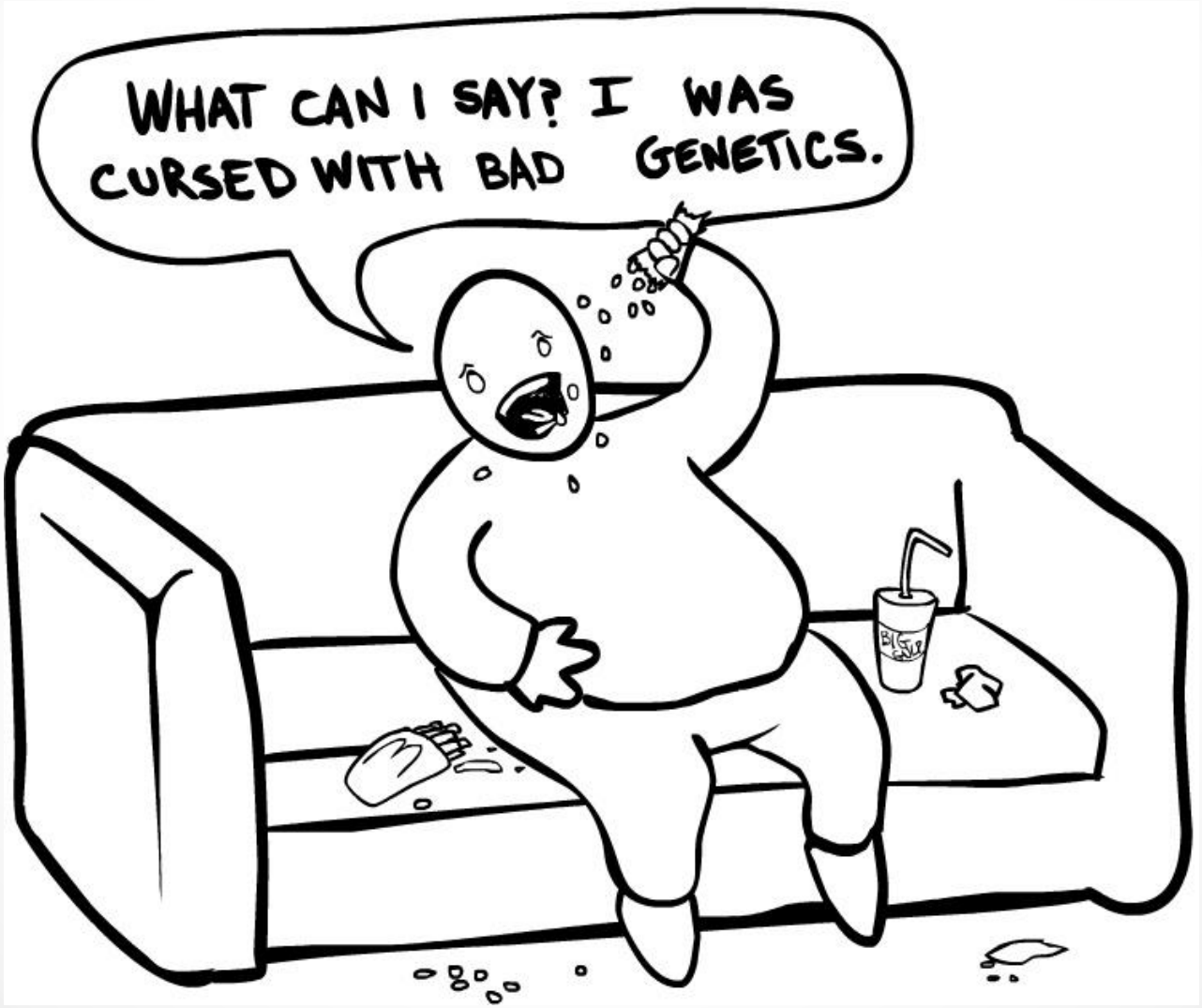
- 1 Mb gene-rich region on chr 14 previously associated with platelet count in whites through GWAS and exome array (*ZFP36L1*)
  - Hispanic index SNP rs117672662 located in an enhancer region located within the first intron of *ACTN1*
- *ZFP36L1-ACTN1* region also contains GWAS signals for fibrinogen and IBD
- *ACTN1* index SNP appears to be distinct from European index SNP in *ZFP36L1*



Frequency of *ACTN1* rs117672662 in 1000G populations

CEU	YRI	ASW	GBR	TSI	CLM	MXL	PUR	CHB	CHS	JPT	IBS	FIN
0	0	0	0	0	0.07	0.06	0.03	0	0	0	0	0

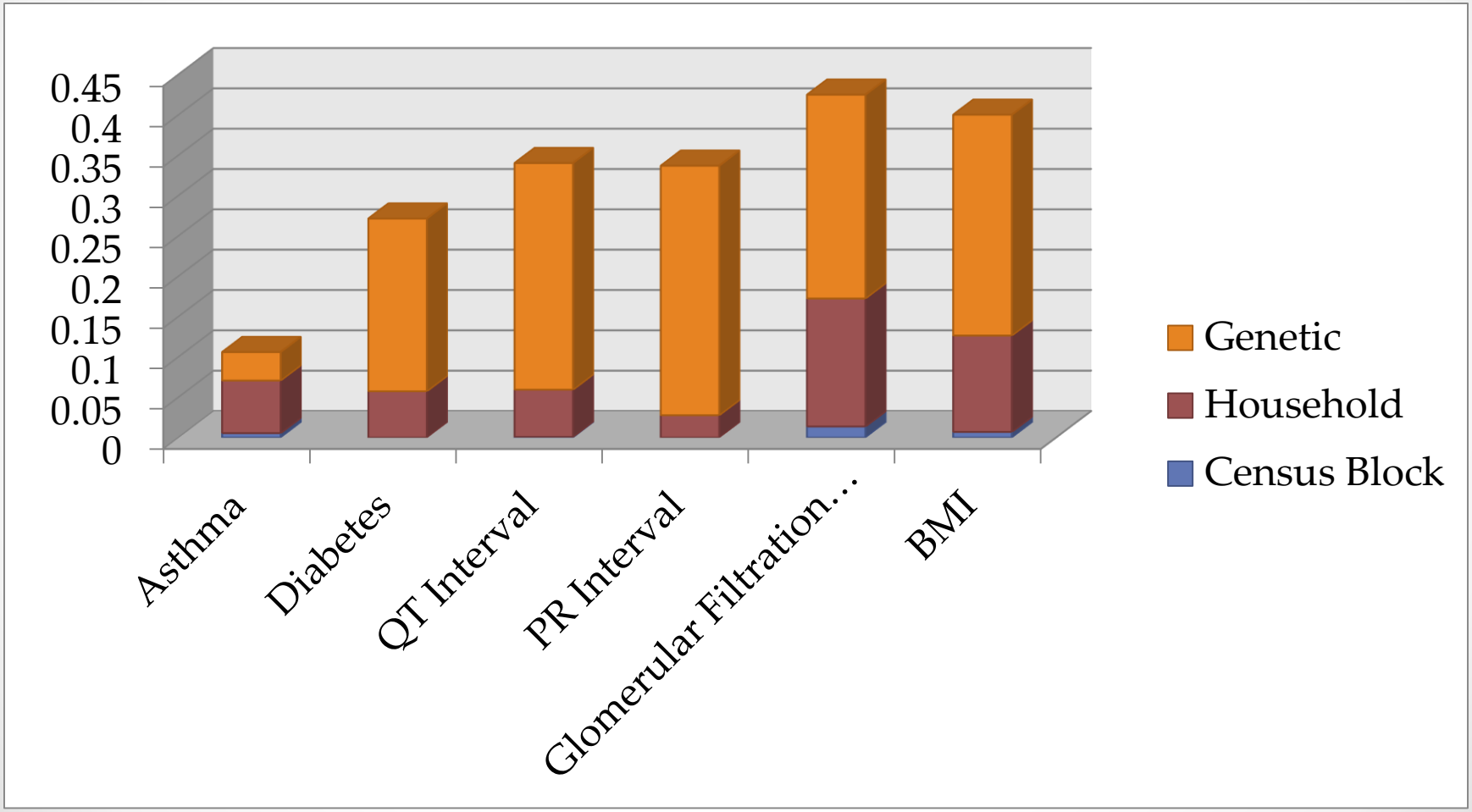
What About Environmental Factors?



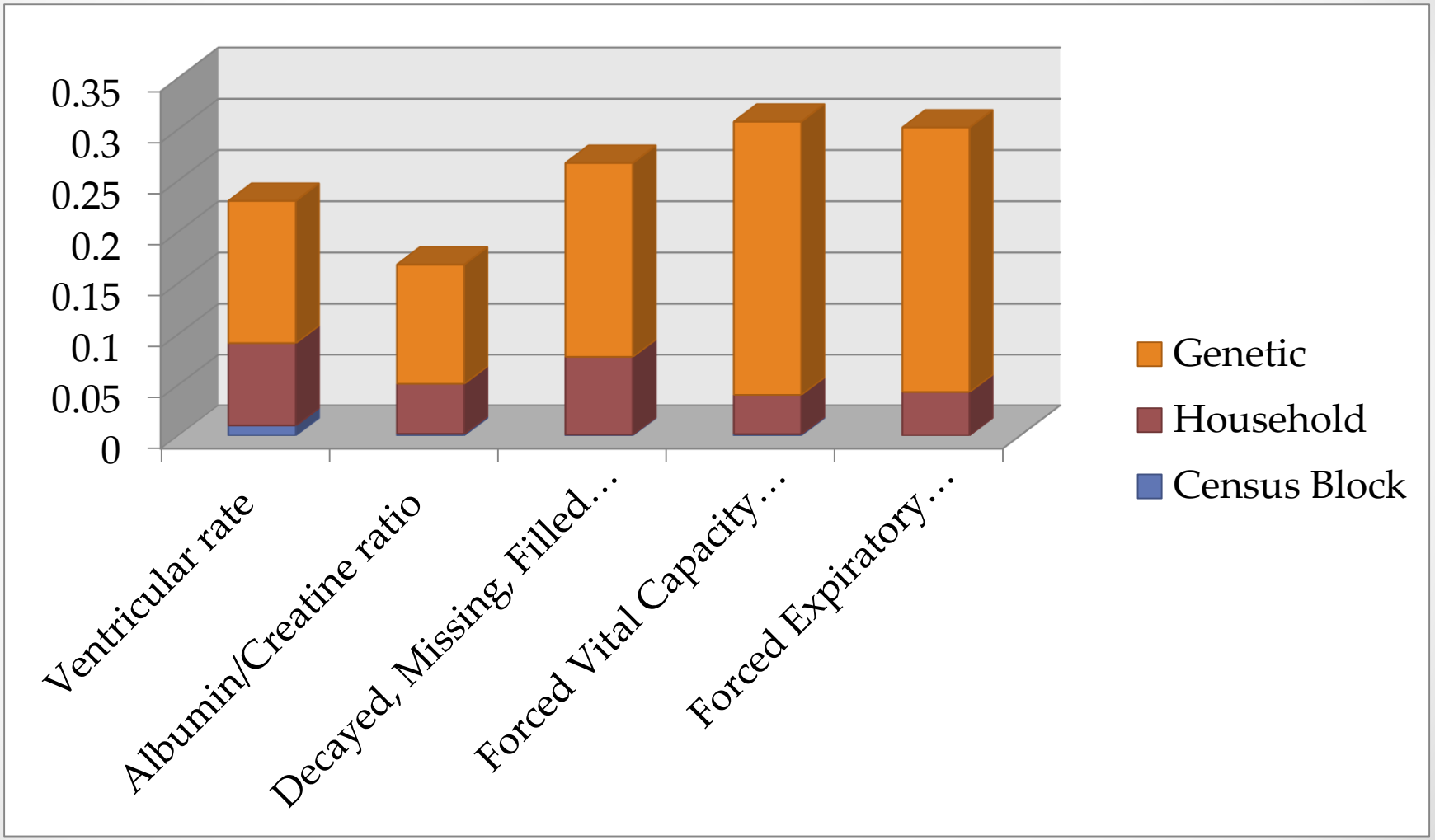
Environmental Contributions to HCHS/SOL Phenotypes

- Complex sampling design
- Have extensive information on HCHS/SOL subjects including household and US Census block group
- Develop a new LMM method to estimate contributions of multiple non-genetic variance components to phenotype variability:
  - block group
  - household
  - polygenic
  - unique environment

Proportional Variance attributed to Household and Genetic effects

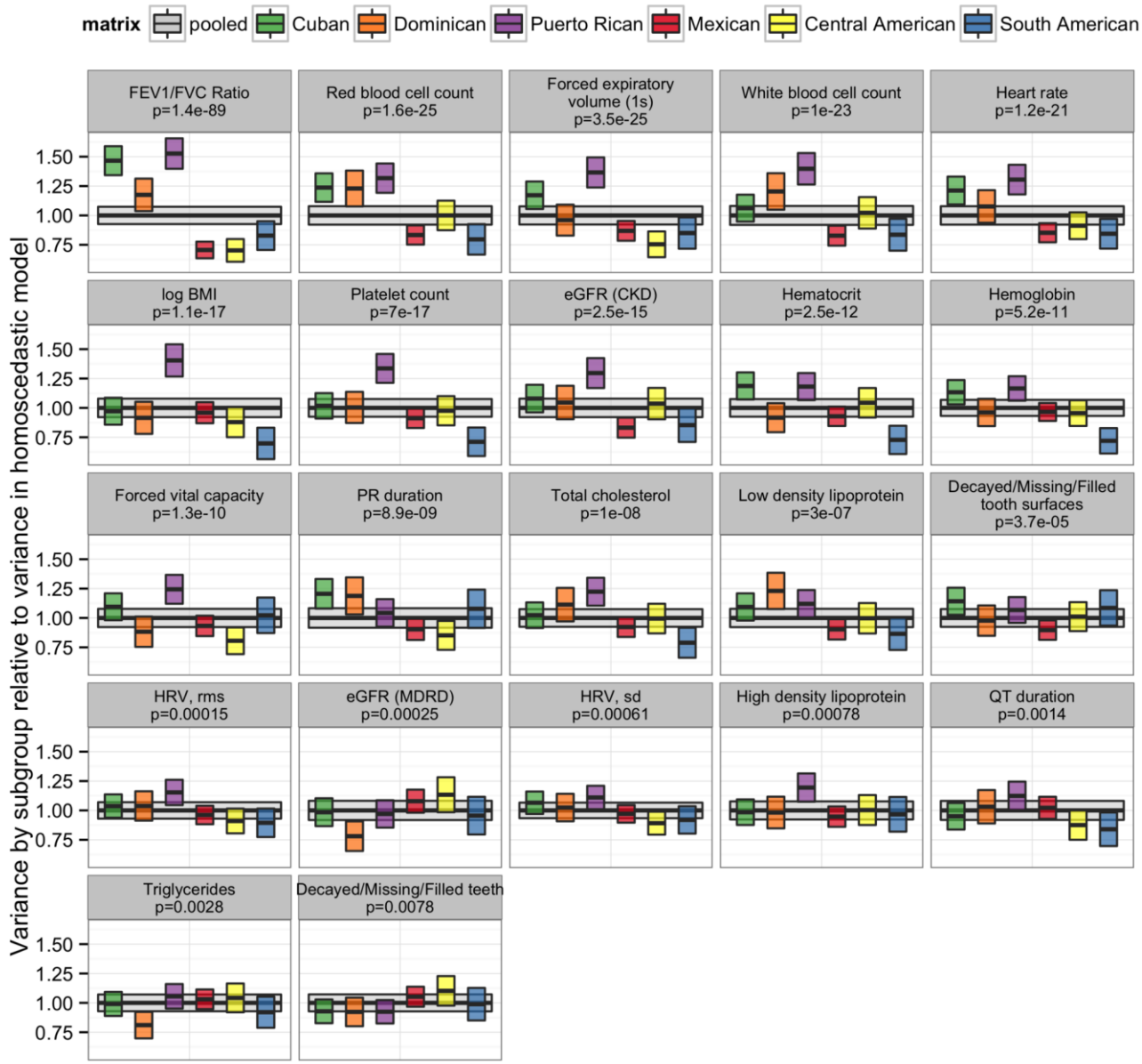


Proportional Variance attributed to Household and Genetic effects



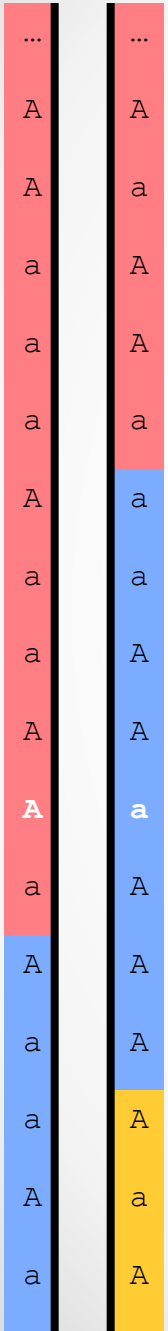
Heterogeneity Among different Hispanic/Latino background groups

- Genomic control (GC) inflation is low for most traits in HCHS/SOL with LMM-OPS
- There are some traits with moderate GC inflation
- We investigated the possibility that heterogeneity in phenotypic variability among different Hispanic/Latino background groups might contribute to the moderate inflation observed for some traits


matrix
pooled
Cuban
Dominican
Puerto Rican
Mexican
Central American
South American
Variance by subgroup relative to variance in homoscedastic model
1.50
1.25
1.00
0.75
FEV1/FVC Ratio
p=1.4e-89
Red blood cell count
p=1.6e-25
Forced expiratory volume (1s)
p=3.5e-25
White blood cell count
p=1e-23
Heart rate
p=1.2e-21
log BMI
p=1.1e-17
Platelet count
p=7e-17
eGFR (CKD)
p=2.5e-15
Hematocrit
p=2.5e-12
Hemoglobin
p=5.2e-11
Forced vital capacity
p=1.3e-10
PR duration
p=8.9e-09
Total cholesterol
p=1e-08
Low density lipoprotein
p=3e-07
Decayed/Missing/Filled tooth surfaces
p=3.7e-05
HRV, rms
p=0.00015
eGFR (MDRD)
p=0.00025
HRV, sd
p=0.00061
High density lipoprotein
p=0.00078
QT duration
p=0.0014
Triglycerides
p=0.0028
Decayed/Missing/Filled teeth
p=0.0078

Leveraging Ancestry Heterogeneity: Locus-specific Ancestry

Full information

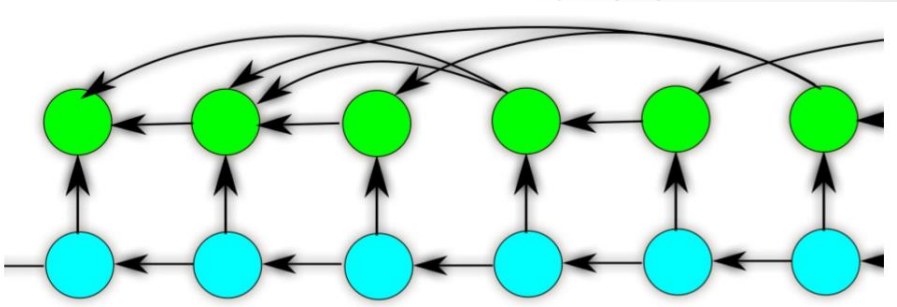


Data structure

… 2 1 1 1 0 1 0 1 2 1 1 2 1 1 1 1 1 2 …

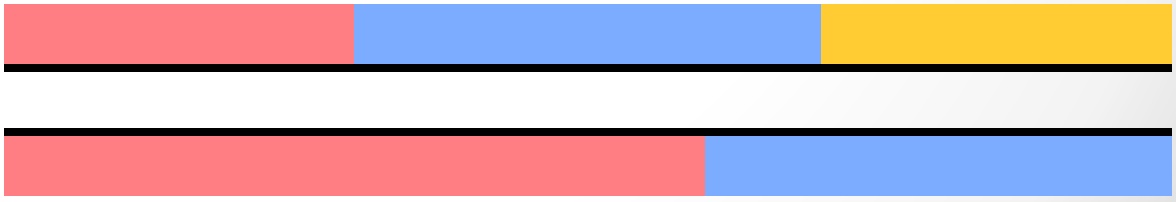
Observed (no phase!)

allele



ancestry

inferred



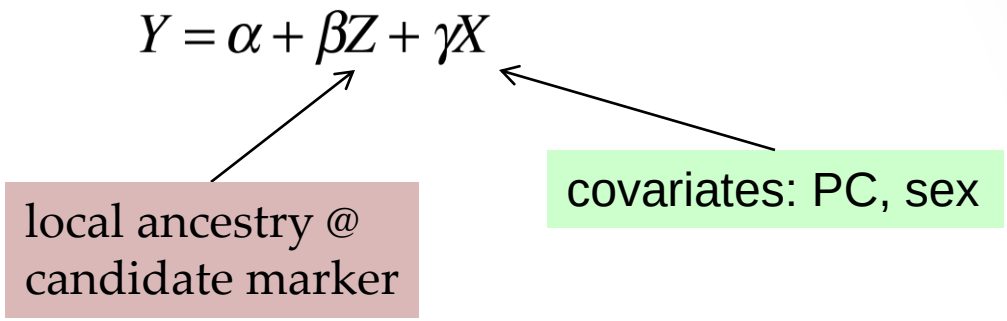
African
European
Native American

Admixture Mapping: Leveraging Heterogeneity

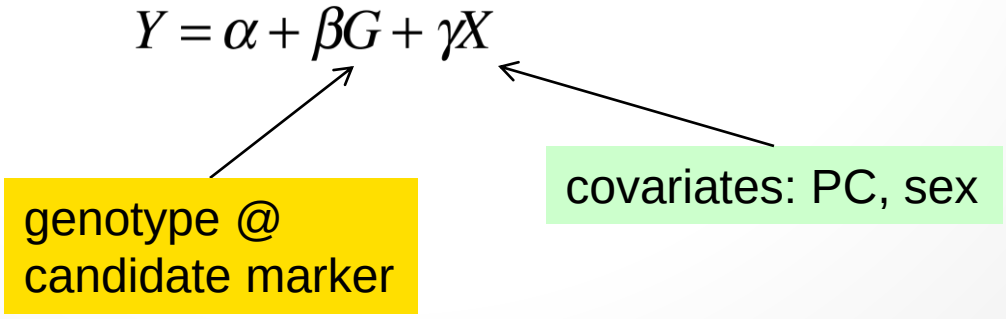
- The heterogeneous genomes of individuals from admixed populations may provide advantages over genetic association analyses in homogeneous populations
- In admixed populations, we can also conduct gene mapping of by using admixture linkage disequilibrium (i.e., admixture mapping)
- For admixture mapping, ancestry is first estimated at specific genomic locations with high-density genotype data.
- Local-ancestry estimates can then be used for complex-trait admixture mapping, for which loci that have unusual deviations of local ancestry and that are significantly associated with a trait are identified.

Admixture Mapping versus Genetic Association models

- Admixture mapping

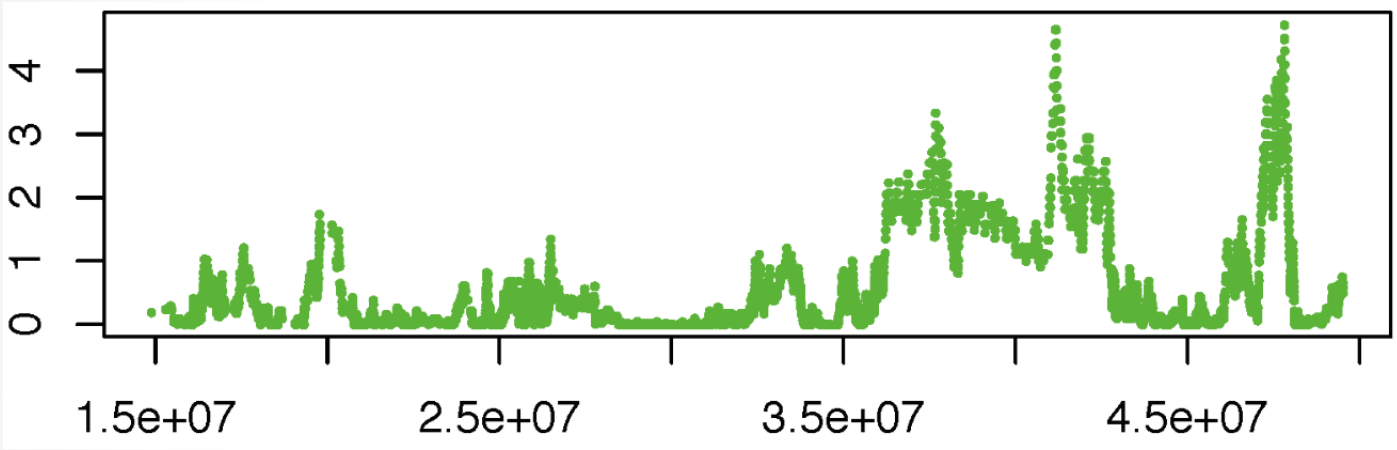


- Genotype association

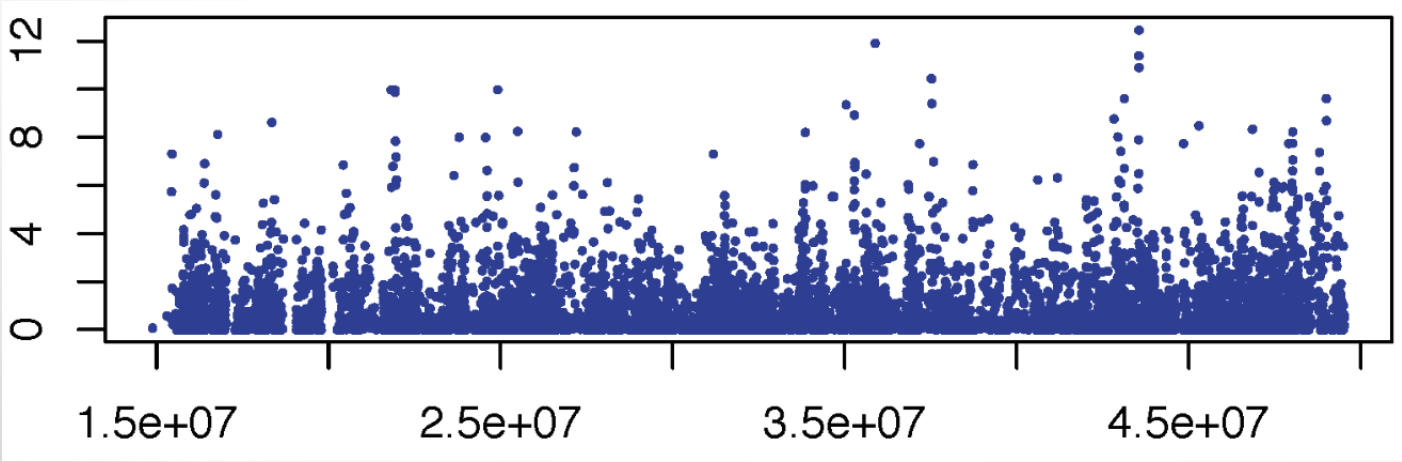


Admixture Mapping versus Genetic Association models

Ancestry test: Genome-wide analysis



Genotype test: Genome-wide Analysis



Albuminuria in Hispanics/Latinos

- Increased urine albumin excretion (albuminuria) is a biomarker of kidney damage
- While the impact of genetic background on albuminuria risk remains elusive, previous studies have found an association between albuminuria and Amerindian ancestry in Hispanic/Latino populations
- Prevalence of Albuminuria is highest in Native Americans (~20%)

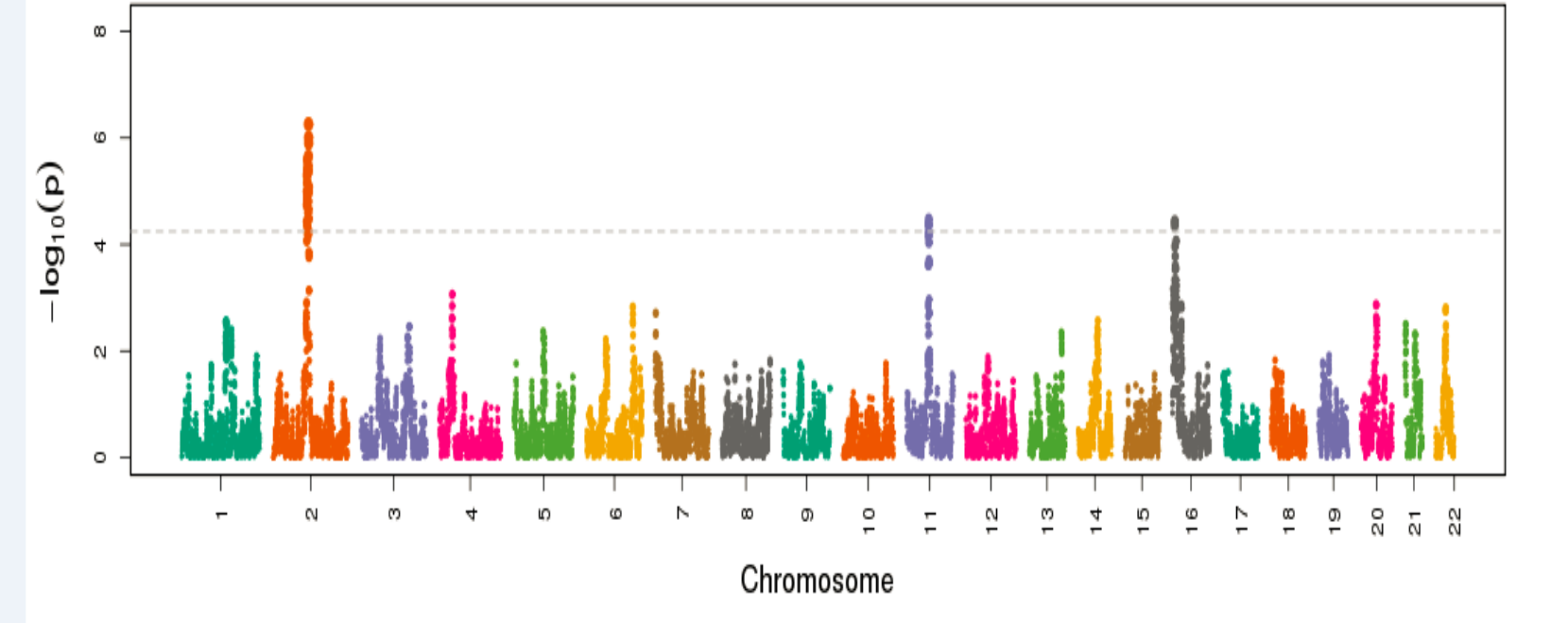
Inferred local ancestry estimates [15]

Local Ancestry Analysis of HCHS/SOL

- RFMix [Maples et al.; AJHG 2013] was implemented for local ancestry inference (LAI)of HCHS/SOL
- BEAGLE (v.4) [Browning & Browning; AJHG 2007] was employed for phasing and imputation of sporadic missing genotypes in the HCHS/SOL and reference panel data sets.
- European, African, and Native American ancestry were inferred with RFMix at 419,645 markers genome-wide

Admixture Mapping of Albuminuria in HCHS/SOL

- Linear mixed model Admixture mapping analysis of albuminuria conducted using 12,212 individuals from HCHS/SOL with



Admixture Mapping of Albuminuria in HCHS/SOL

- Three novel genome-wide significant signals identified at chromosomes 2, 11, and 16.
- The admixture mapping signal identified on chromosome 2, spanning q11.2-14.1, is driven by Amerindian-ancestry.
- Within this locus, the most significant variant is common among Pima Indians (MAF=0.45) but is monomorphic in the 1000 Genomes European and African populations.

SOFTWARE

- **GENESIS**: R software package is available from Bioconductor
- Installation in R:
  - **source("https://bioconductor.org/biocLite.R")**
  - **biocLite("GENESIS")**
- Current release of GENESIS:
  - **PC-AiR**
  - **PC-Relate**
- Recent release includes **LMM-OPS**

Methodology Collaborators

University of Washington

- **Matt Conomos**
- **Lisa Brown**
- **Caitlin McHugh**
- **Jennifer Kirk**
- **Anya Mikhaylova**

- Bruce Weir
- Alex Reiner
- Ken Rice
- Adam Szpiro
- Tamar Sofer

University of Chicago

- Mary Sara McPeek

University of Auckland

- Thomas Lumley

UW Genetic Analysis Center

Department of Biostatistics, University of Washington

Bruce Weir
Ken Rice
Tim Thornton
Sharon Browning
Brian Browning
Katie Kerr
Tamar Sofer

Cathy Laurie
David Levine
Cecelia Laurie
Stephanie Gogarten
Adrienne Stilp
Caitlin McHugh
Quenna Wong

HCHS/SOL

Nora Franceschini
(UNC Chapel Hill)

Alex Reiner
(UW)

Kari North
(UNC Chapel Hill)

Chani Hodonsky
Cathy Laurie
Cecelia Laurie
Jean V Morrison
George Papanicolaou
Alex Reiner
Ursula Schick
Tamar Sofer
Adrienne Stilp
Bruce Weir

ADSP Collaborators

University of Washington

- Ellen Wijsman
- Liz Blue
- Lisa Brown
- Andrew Nato
- Mohamad Saad

Case-Western University

- Jonathon Haines
- William Bush

Columbia University

- Richard Mayeux
- Badri Vardarajan
- Guiseppe Toto

University of Miami

- Margaret Pericak-Vance
- Gary Beecham