Evidence of recent and ongoing admixture in the U.S. and influences on health and disparities

Hannah M. Seagle[1], Jacklyn N. Hellwege[1,2], Brian S. Mautz[2,3], Chun Li[4], Yaomin Xu[5], Siwei Zhang[5], Dan M. Roden[1,3,6,7], Tracy L. McGregor[8], Digna R. Velez Edwards[1,6,9], Todd L. Edwards[1,3]*

[1]Vanderbilt Genetics Institute, [2]Department of Medicine, [5]Department of Biostatistics, [6]Department of Biomedical Informatics, [7]Department of Pharmacology and Biomedical Informatics, [8]Department of Pediatrics, [9]Department of Obstetrics and Gynecology, Vanderbilt University School of Medicine, Nashville, TN 37203, USA

[3]Janssen Pharmaceutical Companies of Johnson & Johnson, Spring House, PA 19477, USA

[4]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA

**Email: todd.l.edwards@vumc.org*

Many researchers in genetics and social science incorporate information about race in their work. However, migrations (historical and forced) and social mobility have brought formerly separated populations of humans together, creating younger generations of individuals who have more complex and diverse ancestry and race profiles than older age groups. Here, we sought to better understand how temporal changes in genetic admixture influence levels of heterozygosity and impact health outcomes. We evaluated variation in genetic ancestry over 100 birth years in a cohort of 35,842 individuals with electronic health record (EHR) information in the Southeastern United States. Using the software STRUCTURE, we analyzed 2,678 ancestrally informative markers relative to three ancestral clusters (African, East Asian, and European) and observed rising levels of admixture for all clinically-defined race groups since 1990. Most race groups also exhibited increases in heterozygosity and long-range linkage disequilibrium over time, further supporting the finding of increasing admixture in young individuals in our cohort. These data are consistent with United States Census information from broader geographic areas and highlight the changing demography of the population. This increased diversity challenges classic approaches to studies of genotype-phenotype relationships which motivated us to explore the relationship between heterozygosity and disease diagnosis. Using a phenome-wide association study approach, we explored the relationship between admixture and disease risk and found that increased admixture resulted in protective associations with female reproductive disorders and increased risk for diseases with links to autoimmune dysfunction. These data suggest that tendencies in the United States population are increasing ancestral complexity over time. Further, these observations imply that, because both prevalence and severity of many diseases vary by race groups, complexity of ancestral origins influences health and disparities.

Keywords: Disparities; Electronic Health Records; Health Outcomes; Admixture.

1. Introduction

Genetic admixture has previously been used to identify geographic variability and historical migration patterns across several human populations[1-11] and to investigate the genetic basis of diseases[12-14]. Two studies have shown temporal increases in heterozygosity due to urbanization, one in a Croatian population[15] and one in a U.S. population of European ancestry[16]. However, these studies of admixture have not connected migratory or urbanization patterns to health outcomes. One study that performed a meta-analysis on populations of individuals with European American and African American ancestry found a positive association between levels of heterozygosity and mortality in humans[17]. However, this study investigated only a single outcome and omitted the impact of temporal trends in admixture and heterozygosity on epidemiological outcomes. Further, these studies have not explored variability in admixture with respect to age or generational trends.

Understanding temporal changes in ancestry and heterozygosity has important implications for individual- and population-level health in humans that remain unexplored. Using human population genetic data to study the connection between ancestry, heterozygosity, and health is ideal due to the substantial number of individuals with genetic data linked to electronic health records (EHR)[18,19]. Further, many diseases and their etiologies recorded in EHRs are known in detail and are well classified, facilitating the estimation of the relationship between heterozygosity and disease risks.

In our cohort of 35,842 individuals from the Southeastern U.S., we investigated temporal changes and variance of admixture by age with de-identified information from the EHR on race, ethnicity, and year of birth linked to genotype data from the Illumina HumanExome array in Vanderbilt University Medical Center's biorepository resource (BioVU)[18]. In addition, we used a phenome-wide association study (PheWAS[20]) to connect genetic data with the clinical phenome capturing clinical disease outcomes in BioVU. This approach allowed us to investigate the relationship between increased ancestral complexity and disease risk. Our study provides important insights into the changing landscape of genetic admixture in a clinical context.

2. Methods

2.1. *Study Population*

Individuals were selected from the BioVU DNA repository which links clinical data from de-identified electronic medical records to DNA samples obtained from patients at Vanderbilt University Medical Center (VUMC)[18]. Each individual's race was designated in the Electronic Health Record (EHR) as either White, Black, Asian, Pacific Islander, American Indian/Alaska Native, or declined/unknown, and an ethnicity of Hispanic/Latino, Not Hispanic/Latino, or declined/unknown. BioVU also contains third-party designated race, which is a good predictor of genetically estimated ancestry in this database[21]. This study of de-identified data was determined to be non-human subject research by the institutional review board (IRB) of Vanderbilt University, Nashville, TN.

2.2. *DNA Extraction and Genotyping*

All DNA samples were isolated from whole blood using the Autopure LS system (QIAGEN Inc., Valencia, CA). Genomic DNA was quantitated via an ND-8000 spectrophotometer and DNA quality was evaluated via gel electrophoresis. Individuals were genotyped using the Illumina Infinium HumanExome Array [12v1-1] (Illumina Inc., San Diego, CA). The data were processed for genotype calling using Illumina's Genome Studio (Illumina Inc., San Diego, CA).

2.3. *Genotyping Quality Control*

Data on 240,117 SNPs and 35,842 individuals (16,289 males and 19,552 females) were available prior to implementation of quality control (QC) measures. No individuals were excluded for low genotyping efficiency (<98%). 6,599 SNPs were excluded for low genotyping efficiency (<98%) and 71,667 SNPs were monomorphic. Twenty-six individuals (14 EHR males and 12 EHR females) were excluded for inconsistent genetic and database sex. After QC, 163,135 SNPs remained for analyses in 35,456 individuals. No SNPs were removed for deviations from Hardy-Weinberg equilibrium.

2.4. *Quantification and Statistical Analyses*

Descriptive statistics on demographic and clinical characteristics were expressed as means with standard deviation or median with interquartile range for continuous covariates and as frequencies or proportions for categorical data using SPSS statistical software (IBM Corporation, Armonk, NY) (Table 1).

Table 1. Summary of demographic characteristics of study individuals

Race*	White	Black	Hispanic/Latino	Asian	Other/Unknown
N (%)	28,723 (80.1)	4,129 (11.5)	550 (1.5)	270 (0.75)	2,170 (2.8)
Male %	46.9%	38.9%	43.6%	42.5%	39.6%
Birth Year					
Mean (SD)	1957 (24.1)	1968 (26.0)	1976 (25.6)	1959 (19.8)	1955 (20.2)
Median (IQR)	1951 (1938-1971)	1964 (1948-1995)	1977 (1957-2000)	1958 (1944-1972)	1953 (1940-1967)
Range	1905-2012	1908-2011	1918-2012	1915-2010	1906-2012

*Non-overlapping categories

A subset of 2,678 ancestry-informative markers (AIMs) were selected for subsequent analysis. We chose AIMs from the ExomeChip selected to have strong differences between African and European ancestry populations as well as between Asian and European ancestry populations. AIMs were used instead of pruned SNP data due to the particular composition of the ExomeChip platform, which was designed with a panel of AIMs to enable evaluation of ancestry.

2.4.1. *Principal component analysis*

EIGENSTRAT v6.0.1 software was used to conduct principal component analysis (PCA) to estimate continuous axes of ancestry from AIMs in all populations together[22]. SPSS was used to create plots of individuals, stratified by birth year which demonstrate trends in changing demography in individuals over time as shown in Fig. 1.

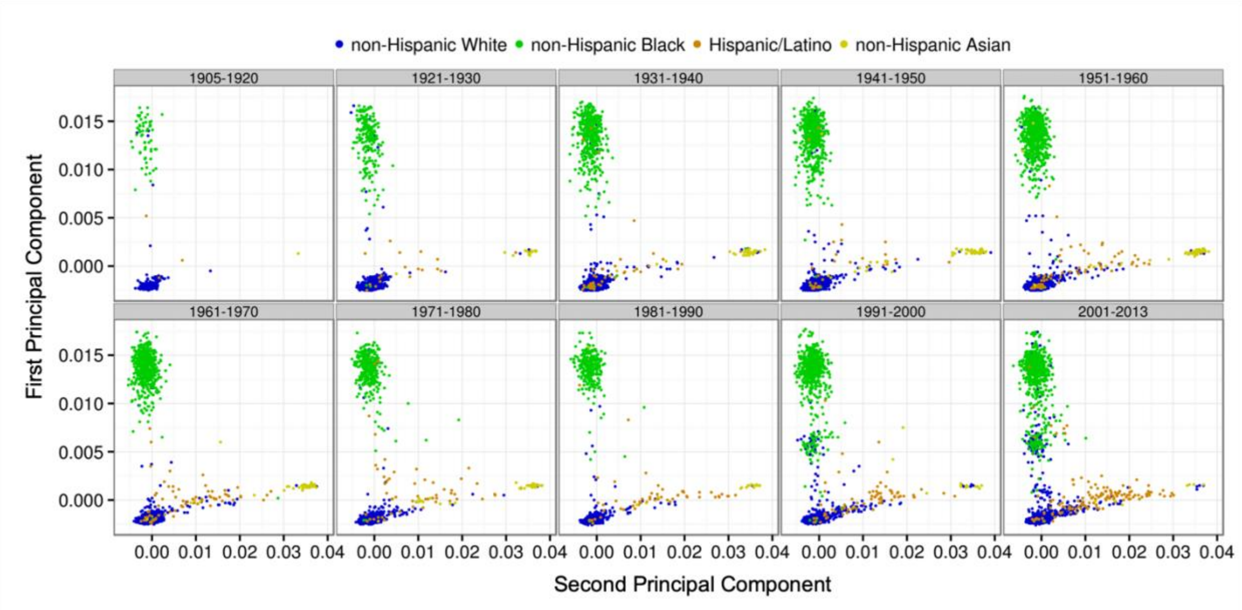

Fig. 1. Principal components plots of study individuals by decade of birth between 1905-2013. Principal components were calculated in EIGENSTRAT and anchored to populations in 1000 Genomes. Sample size information can be found in Table 1.

2.4.2. *STRUCTURE analysis*

STRUCTURE software v2.3.3[23,24] was used to quantify ancestry in combined study and 1000 Genomes Project Phase 3 individuals using the AIMs[22]. We estimated proportions of ancestry assuming ancestral clusters (*K*) ranging from one to 16, where 16 is the number of sub-populations in the 1000 Genomes Phase 3 data plus two. We assumed unlinked SNPs and used 5,000 iterations of burn-in and 10,000 iterations for analysis without providing population information to the software. We observed that the –log-likelihood of the data given *K* did not vary significantly for *K*'s greater than three and observed that *K*'s greater than three primarily subdivided the European populations (data not shown). The three STRUCTURE clusters corresponded to African, Asian, and European ancestry based on comparisons to the 1000 Genomes reference data (data not shown).

Each of the three derived proportions of continental ancestry from STRUCTURE were regressed onto birth year using generalized additive models with integrated smoothness estimation (GAM)[25] implemented in the R package mgcv in all study individuals (Fig. 2), Non-Hispanic Whites, Non-Hispanic Blacks, Hispanics/Latinos, and Non-Hispanic Asians (data not shown). We derived admixture proportion, α, for the i-th individual using the formula $\alpha_i = 1 -$ maximum(%European, %African, %East Asian). We regressed α_i onto birth year using generalized additive models with integrated smoothness estimation for all study individuals, Non-Hispanic White, Non-Hispanic Black, Hispanic/Latino, and Non-Hispanic Asian (Fig. 3).

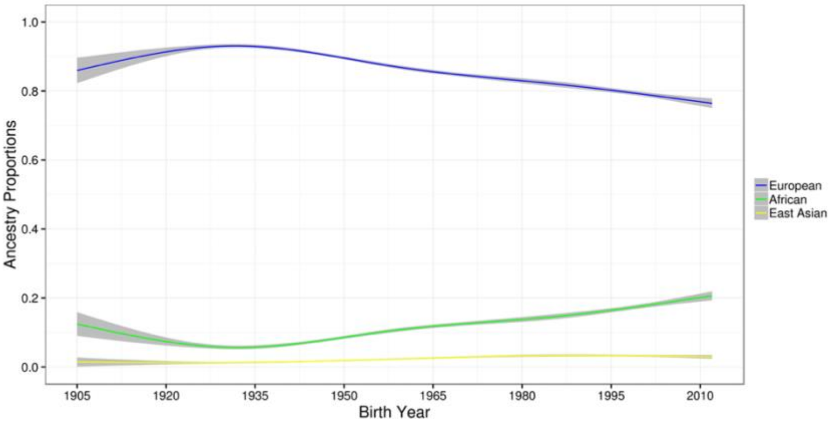

Fig. 2. Proportions of ancestry derived by STRUCTURE analysis of AIMS for study individuals plotted against birth year. Shaded regions represent 95% confidence intervals.

It has been previously shown that when parental populations stop contributing to admixture, that the variance of admixture proportions decreases rapidly, and when parental populations continue to contribute to the admixture, the variance of admixture proportions increases over time[26]. To test the null hypothesis that the observed levels of admixture in our data were not due to ongoing and increasing rates of admixture, we analyzed the association between the variance of α and birth year. We used the software package MVtest and modeled the admixture proportion variance as a log-linear function of birth year and five principal components of ancestry using estimating equations.

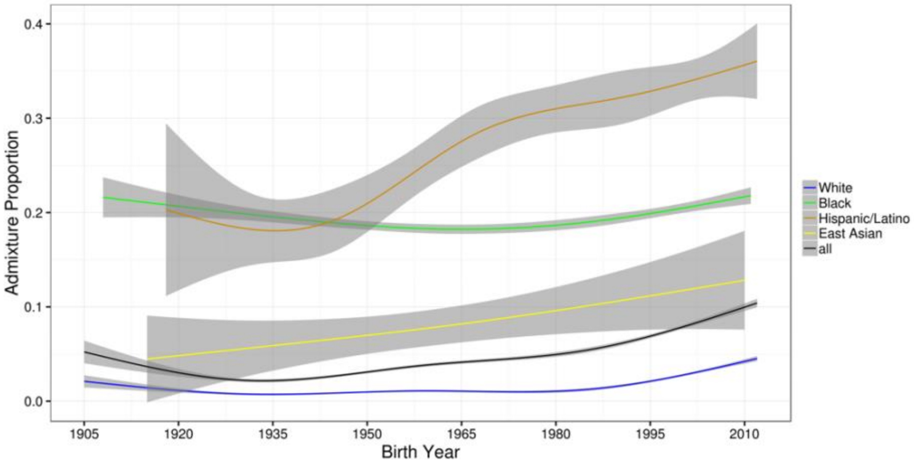

Fig. 3. Admixture proportion plotted against birth year. Admixture proportion is defined as 1 minus the maximum ancestry proportion. The smoothing curves are obtained using the generalized additive model method (gam) with a cubic spline basis implemented in R package mgcv and plotted using R package ggplot2. Sample size information can be found in Table 1. Shaded regions represent 95% confidence intervals. EHR-designated race categories are non-overlapping.

2.5. *Analysis of Admixture Proportion Variance Over Time*

We modeled admixture proportion mean and variance simultaneously as functions of birth year and covariates. Specifically, let m_i be the mean and σ_i^2 be the variance of the trait Y_i for the i-th individual. We model them as

$$m_i = \beta_0 + \beta_g G_i + \sum_{j=1}^{p} \beta_j X_{ij} \quad (1)$$

and

$$\ln(\sigma^2{}_i) = \gamma_0 + \gamma_g G_i + \sum_{j=1}^{p} \gamma_j X_{ij}\ . \quad (2)$$

where G_i is the birth year or variable of interest for the i-th individual and $X_{i1,\ \dots,}\ X_{ip}$ are p covariates. In this model the variance is monotonic with respect to birth year, an assumption that holds in most circumstances. The parameters are estimated simultaneously. This framework allows for testing of the null hypothesis of no effect on mean, variance, or both for any term in the model. These correspond to a mean test with null H_0: $\beta_g = 0$, a variance test with H_0: $\gamma_g = 0$, both having one degree of freedom (DF), and a 2-DF test with H_0: $\beta_g = 0, \gamma_g = 0$.

2.6. *Model Fitting with Estimating Equations*

The parameters are estimated through the estimating equations approach, which does not require a full specification of the outcome distribution, but only a few constraints for the parameters of interest. These constraints are often written as equations, and the parameter estimates can be obtained by solving the equations. The asymptotic distribution for the parameter estimates can be derived[27]. Specifically, suppose the random variable has mean $m_i = \beta_0 + \beta_g G_i + \sum_{j=1}^{p} \beta_j X_{ij}$, and log-variance $\ln(\sigma^2{}_i) = \gamma_0 + \gamma_g G_i + \sum_{j=1}^{p} \gamma_j X_{ij}$. There are $k = 2(p+2)$ parameters, which can be written as a vector, $\theta = (\beta, \gamma)$, where $\beta = (\beta_0, \beta_g, \beta_1, \dots, \beta_p)$ and $\gamma = \gamma_0, \gamma_g, \gamma_1, \dots, \gamma_p)$. Let γ_i and $x_i = (1, g, x_{i1}, \dots, x_{ip})^T$ be the observed values for subject i. If we had assumed normality for the outcome, the log-likelihood for the observation i would have been

$$l_i(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\gamma' x_i - \frac{(y_1 - \beta' x_i)^2}{2\exp(\gamma' x_i)'} \quad (3)$$

for which the partial derivatives with respect to the parameters θ is a k-vector,

$$\Psi_i(\theta) = \begin{pmatrix} \frac{\partial l_i}{\partial \beta} \\ \frac{\partial l_i}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} \frac{y_i - \beta' x_i}{\exp(\gamma' x_i)} \\ \frac{1}{2}\left[\frac{(y_i - \beta' x_i)^2}{\exp(\gamma' x_i)} - 1\right] x_i \end{pmatrix}, \quad (4)$$

and maximum likelihood estimates of the parameters could have been obtained by solving the k equations $\sum_{i=1}^{n} \Psi_i(\theta) = 0$. This motivated us to use these k equations,

$$\sum_{i=1}^{n} \Psi_i(\theta) = 0, \quad (5)$$

as the starting point for our estimating equations approach to obtain parameter estimates $\hat{\theta} = \left(\hat{\beta}, \hat{\gamma}\right)$. If normality holds, then $\hat{\theta}$ are the maximum likelihood estimates. Note that although the estimating equations were motivated by the Gaussian likelihood, one can always start from these equations to obtain $\hat{\beta}$ and $\hat{\gamma}$, whether normality holds or not, and proceed with statistical inference using the M-estimation theory[27]. This is a major advantage for using estimating equations. The partial derivative of $\Psi_i(\theta)$ is a $k \times k$ matrix, denoted as $\psi_i(\theta)$. Using the M-estimation theory, we have

$$\sqrt{n}(\hat{\theta} - \theta)^{d} \rightarrow N(0, V), \quad (6)$$

where the $k \times k$ covariance matrix V can be estimated as $A^{-1}B(A^{-1})^{T}$ with $A = -\frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{\theta})$

$$\text{and } B = \frac{1}{n}\sum_{i=1}^{n} \Psi_i(\hat{\theta})\Psi_i(\hat{\theta})^{T}.$$

If our interest is on the effect of G, the asymptotic result for the joint distribution for the parameter estimates $\hat{\beta}_g$ and $\hat{\gamma}_g$ is

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_g \\ \hat{\gamma}_g\end{pmatrix} - \begin{pmatrix}\beta_g \\ \gamma_g\end{pmatrix}\right] \xrightarrow{d} N(0, V_2), \quad (7)$$

where V_2 is the corresponding 2×2 submatrix of V, with diagonal values denoted as $\widehat{\sigma_{\beta}^{2}}_{g}$ and $\widehat{\sigma_{\gamma}^{2}}_{g}$, respectively. A mean test (H_0: $\beta_g = 0$) can be performed by comparing $\sqrt{n}\beta_g$with N (0, $\widehat{\sigma_{\beta}^{2}}_{g}$), and similarly, a variance test (H_0: $\gamma_g = 0$) by comparing $\sqrt{n}\hat{\gamma}_g$ with N (0, $\widehat{\sigma_{\gamma}^{2}}_{g}$). A 2-DF joint test ($H_0$: $\beta_g = 0, \gamma_g = 0$) can be performed by comparing $n(\hat{\beta}_g, \hat{\gamma}_g)V_2^{-1}\begin{pmatrix}\hat{\beta}_g \\ \hat{\gamma}_g\end{pmatrix}$ with a chi-squared distribution with two degrees of freedom. MVtest software for genetic analysis of SNP data or general analysis of variables is freely available at https://github.com/edwards-lab/MVtest.

2.7. *Heterozygosity Analysis*

Standardized measures of heterozygosity among the AIMs were calculated to evaluate trends in heterozygosity over time relative to expectations. We first estimated the expected number of heterozygous genotypes in an individual in the *k*-th subpopulation as

$$H_k = \sum_i 2p_{ik}q_{ik} \quad (8)$$

where the sum is over all SNPs in our analysis, and p_{ik} and $q_{ik} = 1 - p_{ik}$ are the allele frequencies for the *i*-th SNP. Hardy-Weinberg equilibrium was assumed. Then for every individual *j* in the *k*-th subpopulation, we standardized the observed number of heterozygous genotypes, O_{kj}, by comparing it with the expected number H_k:

$$(O_{kj} - H_k)/H_k \quad (9)$$

Standardized heterozygosity was regressed onto birth year using GAM for all study individuals, and the results were plotted for Non-Hispanic White, Non-Hispanic Black, Hispanic/Latino, and Non-Hispanic Asian.

2.8. *Analysis of Long-Range Linkage Disequilibrium*

To evaluate the presence of admixture long-range linkage disequilibrium (LRLD), pairwise linkage disequilibrium (LD) D' statistics were calculated for all pairs of common (MAF > 0.05) SNPs within 10 megabases (Mb) using Haploview software[28]. D' statistics were regressed onto physical distance between SNPs using generalized additive models with integrated smoothness estimation for distances in the interval from 9-10 Mb for each birth decade (Fig. 4).

2.9. *United States Census Data Analysis*

We downloaded the 1% representative sample of individual-level response to the American Community Survey from the Integrated Public Use Microdata Series (IPUMS) (IPUMS USA, Minneapolis, MN). We regressed the number of major race groups claimed by individuals onto their reported birth year using generalized additive models with integrate d smoothness estimation

and frequency weights provided by IPUMS for TN, the South East Central census region, and the entire U.S. (Fig. 5). For individual groups, such as "White" and "Black or African American", we plotted all individuals who responded affirmatively to those items; thereby, the samples for the individual race group plots in Fig. 5 are not independent and overlap at observation where participants claim two or more race groups.

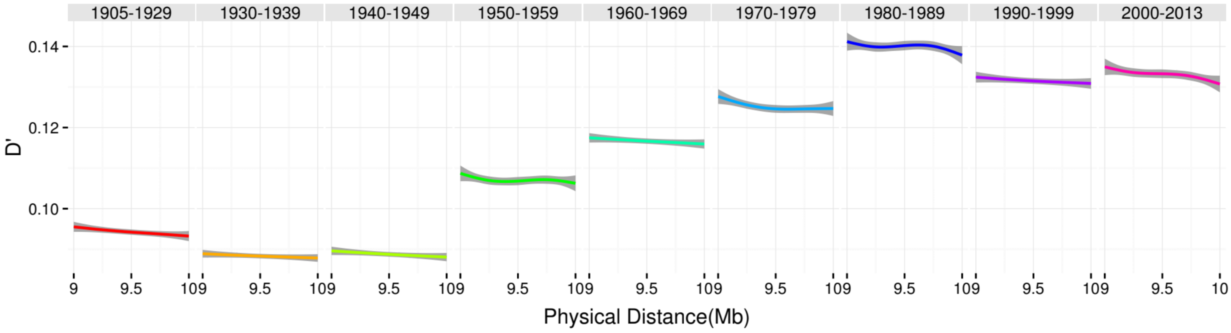

Fig. 4. Pairwise D' for SNPs between 9-10 Mb for all common SNPs on the exome array in all study individuals stratified by intervals of birth decade. Shaded regions represent 95% confidence intervals.

2.10. *Phenotype Classification*

Each individual was classified according to 1,645 phenotypes based on the International Classification of Disease, Ninth Revision, Clinical Modification (ICD9) Codes[20]. Our classification strategy includes all ICD9 features except for procedures. Additionally, the system is hierarchical such that disease subtypes are also classified, such as cardiac arrhythmias are the parent to atrial fibrillation and atrial flutter. Additional phenotypes that are not represented directly in the ICD9 hierarchy are also included, such as inflammatory bowel disease as the parent for Crohn's disease and ulcerative colitis. Diagnoses that were not possible for an individual were set to missing, such as pregnancy for biological males, or prostate disease for biological females. Detailed feature of all phenotype algorithms used are available from: http://phewascatalog.org.

2.11. *Clinical Outcomes*

For each phenotype, we regressed the binary outcome onto the standardized heterozygosity from equation 9 above, adjusted for birth year and the top 5 principal components of ancestry using logistic regression. We limited analysis to outcomes with 40 or more cases and individuals with at least 2 ICD9 codes. We determined the threshold for statistical significance by Bonferroni correction for the number of analyses where the model converged.

3. Results

We evaluated 2,678 ancestry informative markers (AIMs) from genetic data in Vanderbilt University's BioVU. These AIMs were from ExomeChip data in a cohort of 28,723 White, 4,129 Black, 550 Hispanic/Latino, and 270 Asian individuals, based on EHR-third party race designation. The demographic characteristics of study individuals are presented in Table 1.

3.1. *Analysis of Temporal Trends in Genetic Admixture*

After combining our data with the 1000 Genomes as a reference group[29], we calculated principal components to identify patterns of ancestry in each individual. We used the ancestral classifications to test for temporal trends in mean and variance in admixture proportion.

Analysis of temporal trends in genetic admixture showed an increase in ancestral diversity over time. Plots of the first two principal components demonstrated a distinct pattern change in younger generations. To assess the trend of increasing admixture in younger individuals, we calculated the admixture proportion, defined as (1-Predominant fraction of ancestry) for each EHR-designated race (Fig. 3). The admixture proportion consistently increased with younger ages in Asian and Hispanic/Latino groups. In White individuals, the level of admixture remained stable until approximately 1990 when it began to increase notably. Black individuals presented with a decrease in admixture

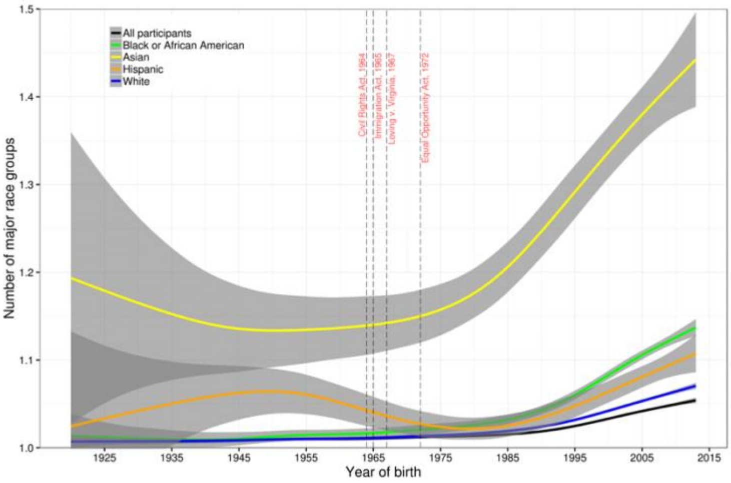

Fig. 5. Plot of individual-level 2013 U.S. Census data for the East South Central Division. Sample size information can be found in Table 1. Shaded regions represent 95% confidence intervals. Race groups are non-overlapping.

proportion in those born in the early- to mid-20th century, but an increase after the 1980's. In all age groups, each recorded race had a small number of individuals who plotted outside of the expected clusters. In later birth years, an increased number of individuals midway between the European and African clusters appear, creating a new cluster representing Black-White biracial children (Fig. 1). Stratifying these plots by EHR-designated race revealed that individuals in this ancestral cluster identify as both White and Black (data not shown). In addition to the clear biracial cluster apparent in the principal component plots, the overall proportions of ancestry across all recorded race groups exhibit increasing admixture in younger individuals. Additionally, a significant increase in the variance of admixture proportions over time was observed (variance coefficient = 0.0193 ± 0.0009 [SE], p-value $< 1.44\text{x}10^{-100}$), indicating that there is a linear increase in variance of the admixture proportion of 0.0193 with every birth year. This finding is consistent with recent and ongoing admixture[26].

We detected similar patterns in three additional sources. First, we compared the rate of heterozygosity to birth year to assess the extent of isolate breaking in our data over time. This occurs when genetically distinct populations reproduce resulting in a temporary excess of heterozygosity. We standardized individual heterozygosity estimates by comparing this estimate with the expected heterozygosity for the EHR-designated race of the individual as described in the methods. Analysis of the standardized heterozygosity by birth year and stated race strongly supports the finding of increasing ancestral diversity in younger individuals (Fig. 6). The timing of inflection for increased standardized heterozygosity varied between race groups, but the data indicated that ancestral diversity has accelerated rapidly in Asian, non-Hispanic Black, and non-Hispanic White cohorts since approximately 1980, while Hispanic/Latino groups have exhibited a relatively steady rate of increasing diversity since the 1940s. This finding reflects the increasing number of children born to biological parents of predominantly different ancestral backgrounds over the past few decades.

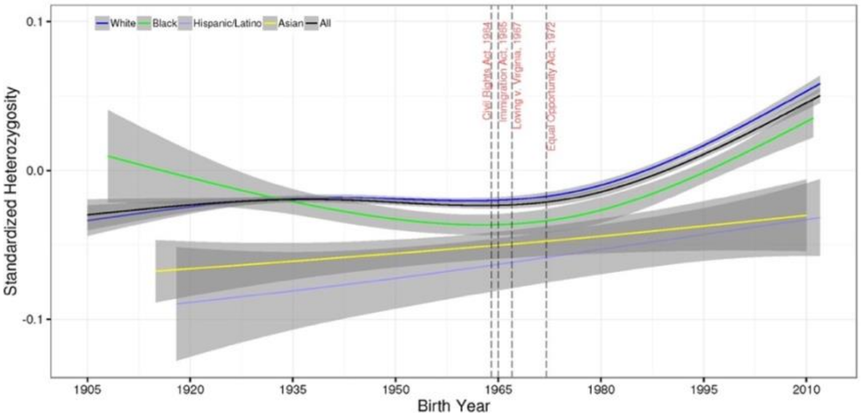

Fig. 6. Standardized heterozygosity plotted against birth year for non-overlapping EHR-designated race groups. Sample size information can be found in Table 1. Shaded regions represent 95% confidence intervals.

Second, because recent admixture leads to increased LRLD, we verified patterns by estimating pairwise LRLD in our dataset. Using common single nucleotide polymorphisms (SNPs) (minor allele frequency [MAF] > 5%) from the genotyping array, we calculated pairwise D' using Haploview[28] and plotted against physical distance for all pairs of SNPs between 9-10 megabases of each other (Fig. 4). These results show a small drop in LRLD in individuals born before the 1950s, followed by significant steady increases in the 1950s through the 1980s and fluctuation at higher levels in the 1990s to 2010s. This finding is consistent with the results of the admixture proportion analysis (Fig. 3), where admixture proportions decreased from the 1910s to the 1940s, and then steadily increased thereafter.

Third, we estimated changes in the number of races indicated in self-reported ancestry in the 2013 American Communities Survey. Respondents were instructed to select Hispanic/Latino/Spanish status and all applicable races for each individual in the household[30]. We analyzed the average number of race categories selected for each individual by birth year and stratified these results within race categories for Tennessee, the East South-Central Region which includes Tennessee, Alabama, Kentucky, and Mississippi (Fig. 5). The results suggest that younger individuals are more likely to indicate multiple races. The inflection point appears to have been earlier in the Asian race category but is demonstrated in all race groups by the mid-1960s in the entire U.S. sample. These data mirror the findings of our cohort genetic analyses.

3.2. ***Changes in Health Diagnoses with Admixture***

To investigate the possible impact of increased ancestral complexity on human health and disparities, we evaluated the association between individual heterozygosity and disease diagnoses using a phenome-wide association study approach (PheWAS[20]). Increasing genetic admixture resulted in fewer diagnoses of female reproductive traits across all data (Table 2). These results remain statistically significant after correction for multiple tests. Phenotype codes for "disorders of menstruation and other abnormal bleeding from female genital tract" and "irregular menstrual cycle/bleeding" were significantly associated with protection by increasing heterozygosity (p-value = $7.21\text{x}10^{-6}$ and $4.37\text{x}10^{-5}$, respectively; Table 2). Other protective findings were also gynecological in nature, including cervical cancer/dysplasia and abnormal Papanicolaou smear results. Significant phenotypes in adults were predominantly detected for biological females. Outside of genitourinary findings, other nominally significant associations (Bonferroni significant < $p \leq 0.05$) show increased risk with genetic admixture and include atopic dermatitis, AV Block, obstructive asthma, and Sicca syndrome.

Table 2. Results from the phenome-wide association study of heterozygosity and clinical outcomes for full sample.

PheCode	Phenotype	P-value	OR (95% Confidence Interval)
626	Disorders of menstruation and other abnormal bleeding from female genital tract	$7.21\text{x}10^{-6}$	0.37 (0.24 – 0.57)
626.1	Irregular menstrual cycle/bleeding	$4.37\text{x}10^{-5}$	0.37 (0.23 – 0.60)
939	Atopic/contact dermatitis due to other or unspecified	$2.86\text{x}10^{-4}$	1.82 (1.32 – 2.52)
180	Cervical cancer and dysplasia	$4.02\text{x}10^{-4}$	0.19 (0.08 – 0.48)
792.1	Papanicolaou smear of cervix or vagina with atypical squamous cells	$5.36\text{x}10^{-4}$	0.21 (0.09 – 0.51)
180.3	Cervical intraepithelial neoplasia [CIN] [Cervical dysplasia]	$6.24\text{x}10^{-4}$	0.15 (0.05 – 0.45)
426.2	Atrioventricular [AV] block	$7.62\text{x}10^{-4}$	2.97 (1.58 – 5.61)
495.11	Chronic obstructive asthma with exacerbation	$8.13\text{x}10^{-4}$	4.63 (1.89 – 11.34)

4. Discussion

Mitigating racial disparities in health is a significant challenge for precision medicine. Some of these population-level health differences may be caused by phenotypic variability associated with ancestral genetic backgrounds. Admixture introduces additional complexity to genetic studies of health disparities and the effects of historical, ongoing, and increasing admixture on population-level health are not well understood. This study evaluates the level of admixture over time and the relationship between ancestral diversity and population health from a clinical perspective.

In our Southeastern United States cohort of 35,842 individuals, we found that for individuals with an EHR race designation of White, the mean proportion of European ancestry decreased from 98% to 92% after the 1990s as the proportion of African ancestry increased to 6%. For individuals designated as Black in the EHR, the mean African proportion decreased by 3% after the 1990s. The European ancestry proportion in the EHR-designated Hispanic/Latino group decreased by 15% after the 1980s (data not shown). Comparing these changes to historical socio-cultural shifts in our cohort's geographic region provides context for these results. In the Southeastern U.S., laws and policies enforced segregation of populations of European and African ancestry. Consistent with these socio-cultural boundaries, there is little change in admixture through the 1960s. Additionally, despite legal rulings and socio-cultural transformation, there remained a very slow increase in admixture and heterozygosity for an additional 20-30 years, followed by a sharp increase over the next few decades.

Our results, qualitatively mirrored in the 2013 American Communities Survey, show that younger individuals are more likely to have greater ancestral diversity than older age groups. The results of our long-range disequilibrium (LRLD) analyses support this notion, with LRLD consistently increasing for individuals born between 1990 and 2010. It is important to note, however, that other possible sources of LRLD (e.g. drift, epistatic selection) cannot be necessarily ruled out, although they seem unlikely given the recent nature of admixture and formation of LRLD. This increase in ancestral heterogeneity of the younger population may also lend itself to more powerful admixture mapping projects in populations not traditionally considered for these types of studies.

Further, we show that changes in population genetic parameters have important consequences for individual and population-level health. Several statistically significant ($p < 5x10^{-5}$) associations of genetic diversity with adult female genitourinary diagnosis codes (e.g. irregular menstrual cycle/bleeding, cervical cancer/dysplasia) were observed. These novel findings linking admixture to protection from menstruation and gynecological abnormalities suggest that ancestral diversity may decrease risk of disorders that could affect reproduction. Further, the changes in reproductive diagnoses were detected predominantly for biological females, suggesting a potential sex-specific population clinical response to changes in admixture. However, the sex-specific response we detected could also be a result of differences in treatment for reproductive health. For example, male reproductive traits, such as sperm quality, may not be routinely checked and reported as with female reproductive parameters.

Other patterns emerge when considering nominally significant ($p < 0.05$) PheWAS results. Several of these diagnoses increase risk with increasing genetic diversity. Importantly, each of these diseases have at least suggestive links to autoimmune dysregulation, including atopic dermatitis[31], AV block[32,33], asthma[34,35], and Sicca/Sjögren syndrome[36]. These patterns suggest a connection between increased heterozygosity and increased activity in the immune system. Because our results show continued increase in genetic admixture over time, it is possible that there will be increases in prevalence of these types of diseases with time as well. Future research should address these immunity-disease relationships with respect to admixture to determine the validity and consistency of these patterns.

The present study has several limitations that warrant consideration. First, the use of EHR data may have high levels of missingness and can introduce inherent selection bias due to patients seeking care at tertiary care centers. Furthermore, given the constraints imposed by our limited sample size and the unavailability of comprehensive reference data for Hispanic/Latino and Native American populations, we were unable to estimate Native American ancestry in this study. Therefore, to provide more robust insights into individuals who identify as Hispanic/Latino and/or Native American, it is necessary to independently validate these results using larger datasets with more diverse reference data.

The concept of race was utilized in this study to reflect demographic dynamics in our cohort's geographic region and to investigate changes to admixture and heterozygosity within these groups. Although the concept of race is a construct with social underpinnings and has limited biological meaning[37], race is often captured in the clinical setting and is the basis for some clinical decision making. It is important to consider the changing implications of classifying individuals by race given the trend of increasing genetic diversity observed in this work and others[38]. As prevalence of many diseases and some drug efficacies vary by race, understanding race-associated factors in patients with complex ancestries may be increasingly important for effective delivery of precision medical care.

5. Acknowledgements

The dataset used in the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. This work was supported by the National Institutes of Health (grant numbers RC2GM092618, U01HG004603, K23HG000001, R25/T32CA160056, R01LM010685). H. Seagle was funded by T32 GM145734-01 (PI: David Samuels). The authors thank Dr. Joshua C. Denny, MD, MS for his contributions to the manuscript.

References

1. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet* **96**, 37-53 (2015).
2. Baharian, S. et al. The Great Migration and African-American Genomic Diversity. *PLOS Genet* **12**, e1006059 (2016).

3. Byrc, K. et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci U S A* **107**, 8954–8961 (2010).
4. Han, E. et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun* **8**, 14238 (2017).
5. Monero-Estrada, A. et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*. **344**, 1280–1285 (2014).
6. Tishkoff, S.A. et al. The genetic structure and history of Africans and African Americans. *Science* **344**, 1035-1044 (2009).
7. Wang, C. et al. Genome-wide analysis of runs of homozygosity identifies new susceptibility regions of lung cancer in Han Chinese. *J Biomed Res* **27**, 208-214 (2013).
8. Lao, O. et al. Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**, 1241-8 (2008).
9. Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309-14 (2015).
10. Novembre, J. et al Genes mirror geography within Europe. *Nature **456**, 98-101 (2008).*
11. Pena, S.D. et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* **6**, e17063 (2011).
12. Franceschini, N. et al. Genome-wide association analysis of blood pressure traits in African-Ancestry Individuals reveals common associated genes in African and Non-African populations. *Am J Hum Genet* **93**, 545-554 (2013).
13. Kato, N. et al. Trans-ancestry genome-wise association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat Genet* **47**, 1282-1293 (2015).
14. Reich, D., & Patterson, N. Will admixture mapping work to find disease genes? *Philo Trans R Soc Lon B Biol Sci* **60**, 1605-1607 (2005).
15. Rudan, I. et al. Quantifying the increase in average human heterozygosity due to urbanization. *Eur J Hum Genet* **16**, 1097-1102 (2008).
16. Nalls, M., A. et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet* **5**, e1000415 (2009).
17. Bihlmeyer, N. A. et al. Genetic diversity is a predictor of mortality in humans. *BMC Genet* **15**, 159 (2014).
18. Roden, D.M. et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84**, 362-369 (2008).
19. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
20. Denny, J.C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1111 (2013).
21. Dumitrescu, L. et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med* **12**, 648-650 (2010).

22. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
23. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-87 (2003).
24. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).
25. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Statist Soc B* **73**, 3-36 (2011).
26. Verdu, P., & Rosenberg, N.A. A general mechanistic model for admixture histories of hybrid populations. *Genetics* **189**, 1413-1426 (2011).
27. Stefanski, L.A. & Boos, D.D. The Calculus of M-Estimation. *Am Stat* **56**, 29-38 (2022).
28. Barrett, J.C., Fry, B., Maller, J., & Dally M.J. Haploview: analysis and visualization of LD maps. *Bioinformatics* **21**, 263-265 (2005).
29. Durbin, R.M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
30. Ruggles, S. Big microdata for population research. *Demography* **51**, 287-297 (2014).
31. Mittermann, I. et al. Autoimmunity and atopic dermatitis. *Curr Opin Allergy Clin Immunol* **4**, 367-371 (2004).
32. Buyon, J.P. et al. Autoimmune-associated congenital heart block: demographics, mortality, morbidity, and recurrence rates obtained from a national neonatal lupus registry. *J Am Coll Cardiol* **31**, 1658-1666 (1998).
33. Villuendas, R. et al. Autoimmunity and atrioventricular block of unknown etiology in adults. *J Am Coll Cardiol* **63**, 1335-1336 (2014).
34. Barnes, P.J. Immunology of asthma and chronic obstructive pulmonary disease. *Nat Rev Immunol* **8**, 183-192 (2008).
35. Tedeschi, A., and Asero, R. Asthma and autoimmunity: a complex but intriguing relation. *Expert Rev Clin Immunol* **4**, 767-776 (2008).
36. Brito-Zerón, P., Izmirly, P.M., Ramos-Casals, M., Buyon, J.P., and Khamashta, M.A. The clinical spectrum of autoimmune congenital heart block. *Nat Rev Rheumatol* **11**, 301-312 (2016).
37. Maglo, K.N., Mersha, T.B., and Martin, L.J. Population genomics and the statistical values of race: an interdisciplinary perspective on the biological classification of human populations and implications for clinical genetic epidemiological research. *Front Genet* **7**, 22 (2016).
38. Wendt, F.R. et al. Modeling the longitudinal changes of ancestry diversity in the Million Veteran Program. *Hum Genomics* **17**, 46 (2023).