

Genetic Diversity Turns a New PAGE in Our Understanding of Complex Traits

Genevieve L Wojcik* (1), Mariaelisa Graff* (2), Katherine K Nishimura* (3), Ran Tao* (4), Jeffrey Haessler* (3), Christopher R Gignoux* (1), Heather M Highland* (2), Yesha M Patel* (5), Elena P Sorokin (1), Christy L Avery (2), Gillian M Belbin (6), Stephanie A Bien (3), Iona Cheng (7), Chani J Hodonsky (2), Laura M Huckins (6), Janina Jeff (6), Anne E Justice (2), Jonathan M Kocarnik (3), Unhee Lim (8), Bridget M Lin (2), Yingchang Lu (6), Sarah C Nelson (9), Sung-Shim L Park (5), Michael H Preuss (6), Melissa A Richard (10), Claudia Schurmann (6), Veronica W Setiawan (5), Karan Vahi (11), Abhishek Vishnu (6), Marie Verbanck (6), Ryan Walker (6), Kristin L Young (2), Niha Zubair (3), Jose Luis Ambite (11), Eric Boerwinkle (12), Erwin Bottinger (6), Carlos D Bustamante (1), Christian Caberto (13), Matthew P Conomos (9), Ewa Deelman (11), Ron Do (6), Kimberly Doheny (14), Lindsay Fernandez -Rhodes (2), Myriam Fornage (10), Gerardo Heiss (2), Lucia A Hindorf (15), Rebecca D Jackson (16), Regina James (17), Cecelia A Laurie (9), Cathy C Laurie (9), Yuqing Li (7), Dan-Yu Lin (2), Girish Nadkarni (6), Loreall C Pooler (5), Alexander P Reiner (9), Jane Romm (14), Chiara Sabati (1), Xin Sheng (5), Eli A Stahl (6), Daniel O Stram (5), Timothy A Thornton (9), Christina L Wassel (18), Lynne R Wilkens (13), Sachi Yoneyama (2), Steven Buyske[‡] (19), Chris Haiman[‡] (5), Charles Kooperberg[‡] (3), Loic Le Marchand[‡] (13), Ruth JF Loos[‡] (6), Tara C Matisse[‡] (19), Kari E North[‡] (2), Ulrike Peters[‡] (3), Eimear E Kenny^{*†} (6), Christopher S Carlson^{*†} (3)

* Shared first authorship

‡ Shared senior authorship

* Corresponding authorship

(1) Stanford University, Stanford CA

(2) University of North Carolina at Chapel Hill, Chapel Hill NC

(3) Fred Hutchinson Cancer Research Center, Seattle WA

(4) Vanderbilt University Medical Center, Nashville TN

(5) Keck School of Medicine, University of Southern California, Los Angeles CA

(6) Icahn School of Medicine at Mount Sinai, New York NY

(7) Cancer Prevention Institute of California, Fremont CA

(8) University of Hawaii Cancer Center, Honolulu HI

(9) University of Washington, Seattle WA

(10) Brown Foundation Institute for Molecular Medicine, The University of Texas Health Science Center, Houston TX

(11) Information Sciences Institute, University of Southern California, Marina del Rey CA

(12) Human Genetics Center, School of Public Health, The University of Texas Health Science Center, Houston TX

(13) University of Hawaii, Honolulu HI

(14) Johns Hopkins University, Baltimore MD

(15) NIH National Human Genome Research Institute, Bethesda MD

(16) Ohio State Medical Center, Columbus OH

(17) NIH National Institute on Minority Health and Health Disparities, Bethesda MD

(18) University of Vermont College of Medicine, Burlington VT

(19) Rutgers University, New Brunswick NJ

Summary/Abstract

Genome-wide association studies (GWAS) have laid the foundation for many downstream investigations, including the biology of complex traits, drug development, and clinical guidelines. However, the dominance of European-ancestry populations in GWAS creates a biased view of human variation and hinders the translation of genetic associations into clinical and public health applications. To demonstrate the benefit of studying underrepresented populations, the Population Architecture using Genomics and Epidemiology (PAGE) study conducted a GWAS of 26 clinical and behavioral phenotypes in 49,839 non-European individuals. Using novel strategies for multi-ethnic analysis of admixed populations, we confirm 574 GWAS catalog variants across these traits, and find 28 novel loci and 42 residual signals in known loci. Our data show strong evidence of effect-size heterogeneity across ancestries for published GWAS associations, which substantially restricts genetically-guided precision medicine. We advocate for new, large genome-wide efforts in diverse populations to reduce health disparities.

1 Introduction

2 A significant European-centric bias has been noted in the field of genome-wide association studies (GWAS), with
3 the vast majority of discovery efforts conducted in populations of European ancestry¹⁻³ while individuals of African or
4 Latin American ancestry account for only 4% of samples analyzed³. (**Extended Data Fig. 1**) Genetic data from ethnically
5 diverse populations will be crucial to powering genome-phenome association studies. Recent publications have reported
6 that some genetic predictors are restricted to certain ancestries, and thus may partially explain risk differences among
7 racial/ethnic groups.⁴⁻⁹ Additionally, as the field shifts its attention towards low frequency variants, which are more likely
8 to be population specific, we can no longer rely on the transferability of findings from one population to another, a
9 complication that has also been observed with some common variants.^{10,11}

10 The lack of representation of diverse populations in genetic research will exacerbate health disparities that exist
11 for many diseases. In the US, minority populations have a disproportionately higher burden of chronic conditions.¹²
12 Globally, developing countries account for 89% of the world's population and 93% of the global disease burden.¹³ By the
13 encouragement of diversity in genomics research, new opportunities for discovery will emerge, and the precision of
14 translational applications will improve. It is imperative that the research community rectifies the imbalance in
15 representation, not only because it is vital for precision medicine and translational research, but also because it is the right
16 thing to do.

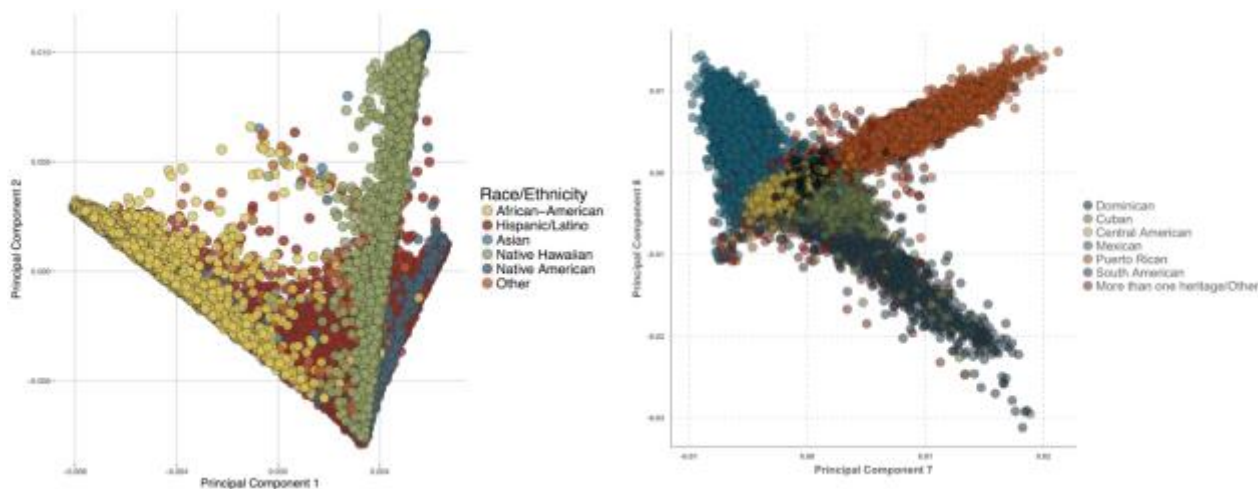
17 Many factors contribute to this bias in genetic research, including the paucity of studies recruiting minorities, lack
18 of information and access to available studies, and complex statistical analyses required for multi-ethnic and admixed
19 study populations.¹⁴ However, recent advancements in statistical analyses and genotyping technologies have lessened
20 many methodological concerns, removing barriers that had previously made researchers reluctant to recruit and analyze
21 heterogeneous samples. The Population Architecture using Genomics and Epidemiology (PAGE) study focuses on
22 exploring the genetics of underrepresented populations.^{15,16} In a study of 49,839 individuals of non-European ancestry, we
23 describe strategies for addressing challenges unique to multi-ethnic studies, investigate population bias in the current
24 GWAS literature, identify numerous new population-specific findings across 26 traits and diseases, consider the
25 implications for clinical genetics, and illustrate the many advantages of genetic inclusion.

26 Unique Methodological Challenges Inherent to Multi-ethnic Studies

27 GWAS in diverse populations have many complexities that must be considered and addressed. PAGE was
28 specifically developed by the National Human Genome Research Institute and the National Institute on Minority Health
29 and Health Disparities to conduct genetic epidemiology research in ancestrally diverse populations, including three major
30 population-based cohorts (HCHS/SOL, WHI, and MEC) and one metropolitan biobank (BioMe). Eligible participants self-
31 identified as Hispanic/Latino (N=22,250), African American (N=17,328), Asian (N=4,696), Native Hawaiian (N=3,944),
32 Native American (N=653), or Other (N=1,056), which includes participants who did not identify with any of the available
33 options and primarily includes those from South Asia or with mixed heritage (**Supplementary Table 1**). Utilizing the
34 detailed phenotype data collected and harmonized across studies, we present genetic association results from 26
35 phenotypes related to inflammation, diabetes, hypertension, kidney function, cardiac electrophysiology, dyslipidemias,
36 anthropometry, and behavior/lifestyle (smoking and coffee consumption).

1 Another major challenge in multi-ethnic studies is the limited availability of genotyping arrays that comparably tag
2 variation in multiple genetic ancestries, especially in those with African ancestry. To address this, a collaboration among
3 PAGE, Illumina, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA)¹⁷, and other
4 academic partners developed the Multi-Ethnic Genotyping Array (MEGA), which includes a GWAS scaffold designed to
5 tag both common and low frequency variants in global populations.¹⁸ (**Extended Data Fig. 2**) Additionally, it contains
6 enhanced tagging in exonic regions, hand-curated content to interrogate clinically relevant variants, and enriched
7 coverage to fine-map known GWAS loci.¹⁹ The principles used to design MEGA are currently being used to create other
8 multi-ethnic genotyping arrays, including the Multi-Ethnic Global Array and the Global Screening Array.

9 Historically, analyses have been stratified by self-identified race/ethnicity to account for confounding by genetic
10 ancestry. In PAGE, we conducted principal component analysis to evaluate population substructure and mapped self-
11 identified racial/ethnic groups onto the estimated principal components (PCs). Most notably in Hispanics/Latinos, but
12 evident to a lesser extent in all populations, genetic ancestry reveals greater demographic complexity compared with
13 culturally assigned labels, appearing as a continuum and demonstrating that genetic ancestry is not categorical in diverse
14 populations that have varying degrees of admixture (**Figure 1**). Stratifying by self-reported race/ethnicity would fail to
15 separate groups with similar patterns of genetic ancestry and therefore would still require adjustment of PCs with reduced
16 statistical power in a smaller sample size. For this reason, we pooled all samples in a single analysis.



18 **Figure 1: Principal Component Analysis of PAGE Populations.**

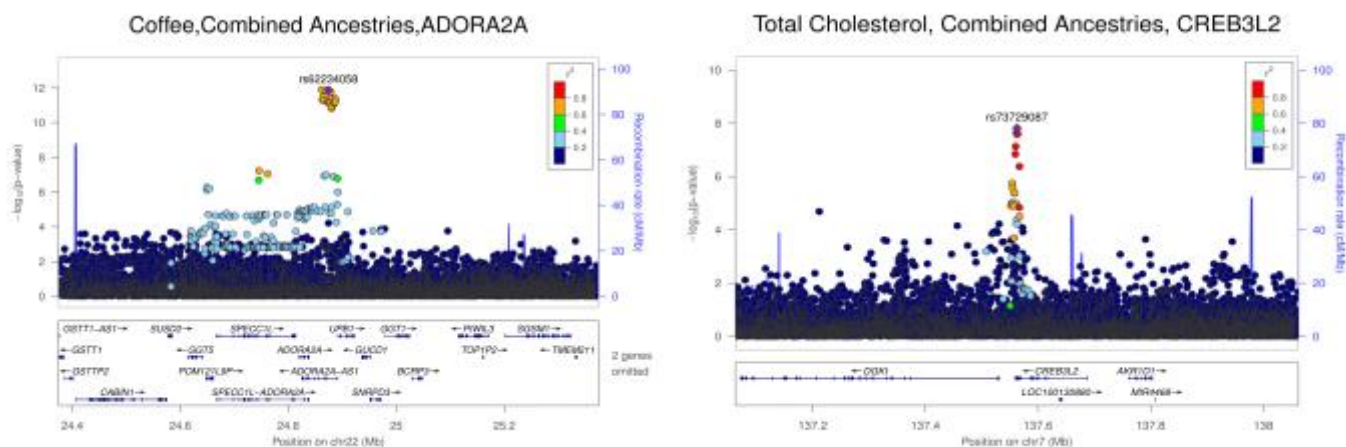
19 *Scatter plot of PCs for PAGE racial/ethnic groups. Each point represents one individual, color-coded by self-identified*
20 *race/ethnicity. (a) Global variation (PC1 vs PC2) (b) Hispanic/Latino variation (PC7 vs PC8).*

21
22
23 Multi-ethnic GWAS also require sophisticated statistical modeling. Known and cryptic relatedness are often
24 concerns for studies recruiting from smaller, more isolated populations. WHI, MEC, and BioMe used population-based
25 recruitment, whereas HCHS/SOL used a household sampling study design, which increased the inclusion of relatives. To
26 account for relatedness within and across studies, we used two recently developed analytical methods for GWAS of
27 related individuals from admixed populations. GENESIS^{20–22} uses a linear mixed model and accounts for the correlation
28 among genetically similar samples through a kinship matrix that estimates the known and cryptic relatedness in the
29 presence of population structure and admixture. SUGEN²³ uses a modified version of generalized estimating equations
30 and creates “extended” families by connecting the households who share first degree relatives. Single-variant association

1 testing was completed in both GENESIS and SUGEN using phenotype-specific models that were adjusted by indicators
 2 for study, self-identified race/ethnicity as a proxy for cultural background, phenotype-specific standard covariates, and the
 3 first 10 PCs. Because at the time of analysis SUGEN could analyze both continuous and binary phenotypes, while
 4 GENESIS could only analyze continuous phenotypes, we present SUGEN results below, and include GENESIS results in
 5 the Supplementary Tables. For comparison against traditional multi-ethnic approaches, we analyzed stratified by self-
 6 identified race/ethnicity, and meta-analyzed to assess heterogeneity by ancestry.

7 28 Novel Loci Found in 26 Phenotypes

8 Since the majority of GWAS have been conducted in European-ancestry populations, we hypothesized that the
 9 examination of underrepresented populations would reveal ancestry-specific associations that European-centric studies
 10 were unable to detect. Across 26 phenotypes, we discovered 28 novel loci at least 1 Mb away from a known locus that
 11 remained genome-wide significant ($P_{\text{cond}} < 5 \times 10^{-8}$) after conditioning on all previously identified variants on that
 12 chromosome (**Table 1, Supplementary Tables 2-3**). We attribute many of these discoveries to MEGA's globally diverse
 13 panel of variants and to a study population that includes ancestries where these variants are more frequent. Here, we
 14 briefly discuss two illustrative examples (**Figure 2, Extended Data Fig. 3**).



15
 16 **Figure 2: Exemplars of Novel Loci Identified within PAGE**

17 *LocusZoom* plots for examples of novel loci are illustrated based on results from the pooled sample, specifically: coffee
 18 (cups/day)^a with lead SNP rs62234058, and total cholesterol(mg/dl)^b with lead SNP rs73729087.

19 a The association model for coffee (cups/day) was adjusted for age, sex, PC1-10, study, study center, and ancestry. Prior to
 20 analyses, a value of 1 was added to coffee intake followed by a log transformation.

21 b The association model for total cholesterol (mg/dl) was adjusted for age, sex, body mass index, PC1-10, study, study center, and
 22 ancestry. Intake of lipid medications was accounted for by adding a constant based on the class of lipid medication. See Methods for details.

23
 24 A novel locus on chromosome 22q11 was associated with coffee intake (cups/day; **Figure 2A**) in *ADORA2A*
 25 ($P=1.33 \times 10^{-12}$, $N=35,902$). The lead variant (rs62234058) is common in African Americans (coding allele frequency
 26 (CAF)=0.22) and Hispanic/Latinos (CAF=0.05), but rare in those of Asian and European ancestry (CAF<0.01). Given the
 27 rarity of the minor allele in Europeans, the discovery of this association was facilitated by our multi-ethnic study design
 28 and driven by PAGE African Americans ($P=3.70 \times 10^{-7}$, $N=11,862$) and Hispanic/Latinos ($P=3.21 \times 10^{-6}$, $N=15,837$). The
 29 *ADORA2A* gene is the main target of caffeine action in the central nervous system, and another SNP in this gene has
 30 previously been associated with caffeine-induced sleep disturbance (rs4822498²⁴). This finding showcases an ancestry-
 31 specific genetic trait which impacts a behavioral phenotype.

1 The second example describes a novel locus in *CREB3L2/7q33* associated with total cholesterol levels
2 (rs73729087: $P=1.52 \times 10^{-8}$, $N=33,185$, $CAF=0.05$) (**Figure 2B**). While rare in European populations ($CAF=0.005$), it is
3 more common in PAGE racial/ethnic groups, including African Americans ($P=1.77 \times 10^{-6}$, $N=10,137$, $CAF=0.11$) and
4 Hispanic/Latinos ($P=2.58 \times 10^{-3}$, $N=17,802$, $CAF=0.02$). This noncoding variant is located in the 3'-UTR, possibly
5 contributing to the regulation of *CREB3L2* expression. These examples represent just two of numerous novel findings that
6 would not have been discovered in a European-descent study population.

7 Genetic Heterogeneity in the GWAS Catalog Reveals Need for Fine-Mapping

8 In general, GWAS identify loci where one or more tagSNPs show significant association with the trait of interest.
9 However, GWAS do not lead directly to the identification of the functional variant (fSNP), which ideally is in strong linkage
10 disequilibrium (LD) with the tagSNP(s) as surrogates. However, LD can vary among populations, so a tagSNP in perfect
11 LD with the fSNP in one population may be in weak LD in a different population. This can lead to inconsistent estimates of
12 the effect sizes among populations (and therefore effect size heterogeneity) if the tagSNP (instead of the causal fSNP) is
13 used for effect size calculations. Because European-descent individuals are overrepresented in GWAS discovery
14 populations and have different LD structures than other racial/ethnic groups, we hypothesized that effect size
15 heterogeneity among populations may exist for many previously reported tagSNP associations.

16 To test this hypothesis, we measured the frequency of effect heterogeneity in PAGE's multi-ethnic study
17 population of tagSNPs, primarily discovered in European populations, reported to the GWAS Catalog. We were able to
18 replicate ($P < 5 \times 10^{-8}$) a total of 574 tagSNPs in 261 distinct genomic regions across 26 traits out of the related 3,322 unique
19 GWAS Catalog variants (**Supplementary Table 4**).²⁵ After Bonferroni correction for 574 tests, 132 tagSNPs (23.0%)
20 showed significant evidence of effect heterogeneity by genetic ancestry (SNP \times PC $P < 8.71 \times 10^{-5}$). Thus, we observe that
21 nearly a quarter of reported GWAS Catalog tagSNPs, the preponderance of which were identified in European-based
22 studies, show evidence of effect heterogeneity upon replication in a multi-ethnic study population. This estimate is
23 conservative, because some of the effects that failed to replicate at $P < 5 \times 10^{-8}$ might have been underpowered to detect
24 heterogeneous effects, especially for the less frequent alleles.

25 While we replicate 261 regions previously implicated in the GWAS Catalog, for most of these regions (77%) the
26 strongest signal was not the previously reported tagSNP from the GWAS Catalog but a different tagSNP. Additionally,
27 heterogeneity was only observed at 6% of these tagSNPs with the strongest associations within all 261 regions. This is
28 consistent with multi-ethnic analyses fine-mapping known association signals at a majority of reported GWAS catalog loci,
29 attributable to differential tagging of the underlying functional variation among populations, rather than that there are truly
30 differential underlying fSNP effect sizes. These results have important implications for precision medicine, as risk
31 prediction models based on heterogeneous GWAS Catalog tagSNPs could have poor accuracy in non-European
32 ancestries.

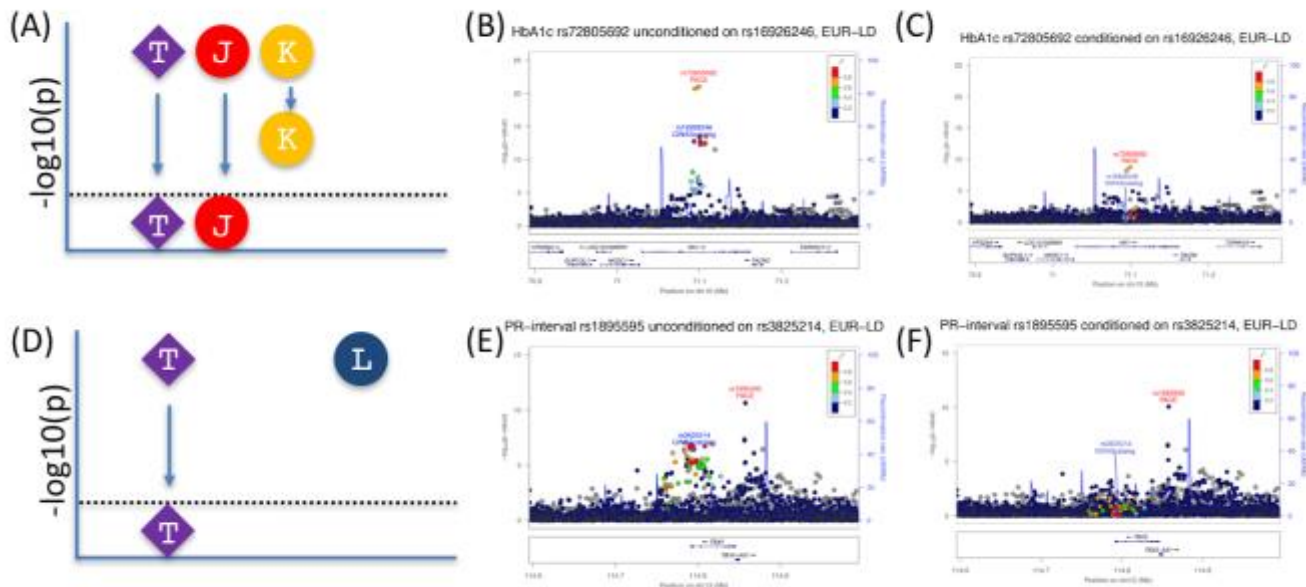
33 42 Residual Signals in Known Loci Found in 26 Phenotypes

34 In addition to refining loci, multi-ethnic analysis affords an opportunity to identify independent signals (secondary
35 variants) within known loci, further enriching our understanding of the genetic architecture of traits. To test for secondary

1 signals, we screened for statistical associations that remained genome-wide significant ($P_{\text{cond}} < 5 \times 10^{-8}$) after adjusting for all
2 known tagSNPs (the “adjusted” model), identifying 42 new variants located within 1 Mb of a previously known variant
3 (**Table 1, Supplementary Tables 2-3**). If the residual signal represents a statistically independent association, then we
4 would expect no net change in the strength of the association between unadjusted and adjusted models and that the
5 known tagSNPs were in weak LD with the residual SNPs. Out of the 42 residual variants, 23 and 25 in Hispanic/Latino
6 and African-descent populations, respectively, show evidence of a secondary association independent (LD $r^2 < 0.2$) of
7 previously known loci. This analysis suggests that approximately half of these known loci, from majority European-based
8 GWAS, contain novel secondary signals in these populations.

9 To further illustrate the difference in mechanism between fine-mapping and secondary independent signals, we
10 highlight two examples (**Figure 3**). The first is a refinement of the association between hexokinase 1 (*HK1*) and HbA1c.
11 The residual signal at rs72805692 ($P_{\text{unadj}} = 9.22 \times 10^{-22}$, $N = 11,178$, $\text{CAF} = 0.061$) is in moderate LD in European ($r^2 = 0.61$) and
12 Hispanic/Latino ($r^2 = 0.63$) populations with the previously implicated SNP (rs16926246) 5.7kb away. Therefore, after
13 adjustment, the signal is greatly diminished but remains statistically significant ($P_{\text{cond}} = 3.05 \times 10^{-9}$). This represents the
14 refinement of a known locus (fine-mapping), as the high LD present in this area results in an attenuated, but still
15 statistically significant, signal, and may represent only one underlying fSNP. In contrast, we found a residual signal for PR
16 interval at rs1895595, upstream of *TBX5* ($P_{\text{unadj}} = 2.16 \times 10^{-11}$, $N = 17,428$, $\text{CAF} = 0.17$). After adjustment for 5 known tagSNPs
17 in this region (rs3825214, rs7312625, rs7135659, rs1895585, rs1896312), the signal remains largely unchanged
18 ($P_{\text{cond}} = 1.99 \times 10^{-11}$). This secondary signal at rs1895595 is independent of all 5 conditioned SNPs, with extremely low LD
19 ($r^2 < 0.03$) across all global populations, and therefore likely represents an independent fSNP. Both fine-mapping of primary
20 findings and knowledge of independent, secondary alleles are important to comprehensively characterize GWAS loci,
21 particularly in diverse populations, thereby improving genetic risk prediction.

22



23

24

Figure 3: Residual signals can represent either refinement of signal or secondary alleles.

(A) Fine-mapping: $-\log_{10} p$ values are plotted against position for a GWAS catalog tagSNP T , as well as two tagged SNPs: J is strongly tagged by T ($r^2=1$) in all populations, and K is variably tagged across populations. After adjustment, signal at T and J is no longer significant, but residual signal at K indicates that the original association has been fine-mapped. Unadjusted (B) and adjusted (C) results for trait HbA1c, showing weakened signal at residual SNP rs72805692 after adjusting for GWAS catalog tagSNP rs16926246, consistent with signal refinement. This tagSNP was first reported from a study of 46,368 Europeans²⁶, so LD with the tagSNP is shown from a European reference panel, illustrating how the set of strongly tagged SNPs (red/orange) is fine-mapped to the two strongest (residual) signals in the multi-ethnic population. (D) Secondary alleles are independent of known loci, so L is not in significant LD with T ($r^2 \sim 0$). After adjustment for T , signal at L is unchanged. Unadjusted (E) and adjusted (F) results for trait PR interval, showing no change in signal at residual SNP rs1895595 after adjusting for GWAS catalog tagSNP rs3825214, consistent with the residual signal being an independent secondary allele. Again, LD shown is from a European population, as the GWAS catalog report²⁷ was from 12,670 Europeans.

Ancestries that Drive PAGE Findings

To tease apart the influence of specific ancestral components on the 28 novel and 42 residual loci, we calculated the correlation between the risk allele and each of the first ten PCs in the full PAGE sample (Figure 4A). These correlations reveal population structure underlying many of our novel and residual findings, in which there are population differences in allele frequencies for the risk alleles. Most notably, the risk allele for a novel finding for cigarettes per day among smokers on chromosome 1 (rs182996728; $P=3.1 \times 10^{-8}$) was found to show significant correlation with PC4, which represents Native Hawaiian/Pacific Islander ancestry. While this variant is monomorphic or rare in most populations, it is found at 17.2% within our Native Hawaiian participants. An additional example is shown with the 5 novel and residual loci highly correlated with PC6 which are related to height and found to be at higher frequencies in 1000 Genomes within a subgroup of populations within East Asia, such as Japanese or Vietnamese. The observed variability in allele frequency for our findings will result in differential impacts across populations and must be considered when building risk prediction models. That our findings exhibit substantial variability in allele frequencies further illustrates a need for the inclusion of diverse populations disproportionately affected by disease.

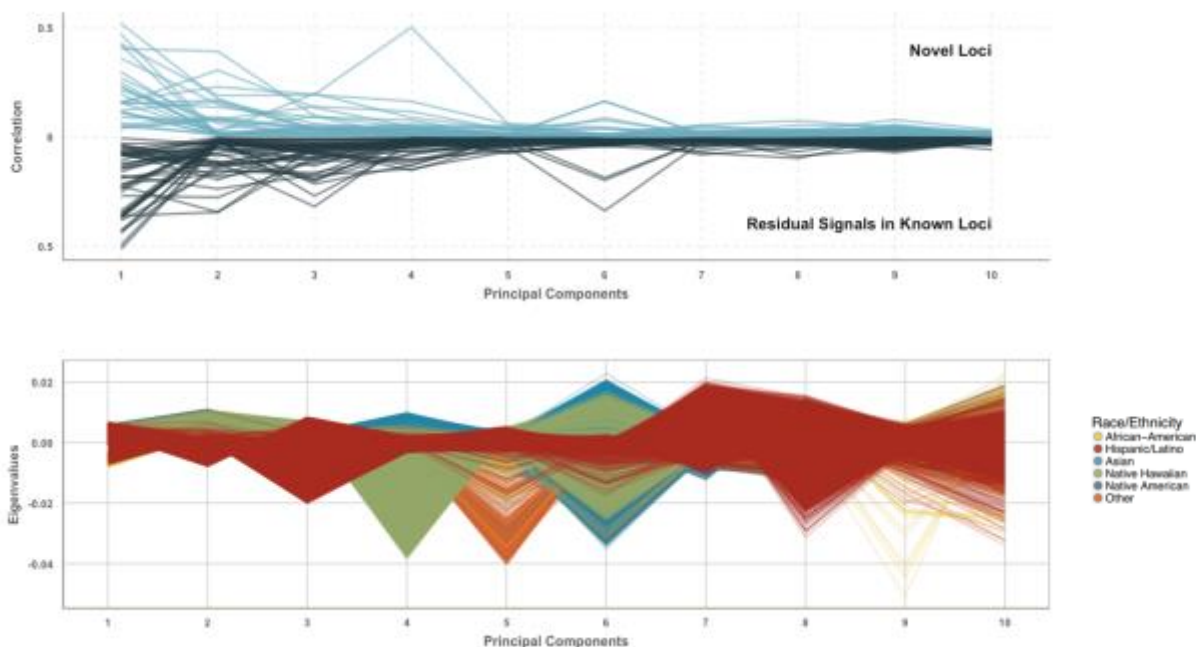


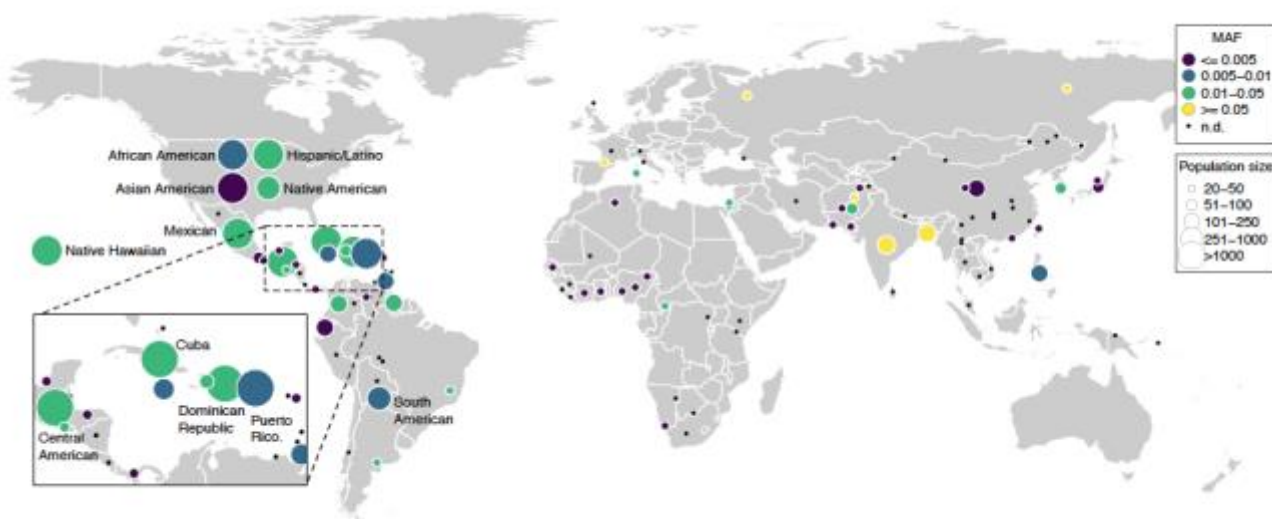
Figure 4: Correlation between SNP genotype and PC1-PC10.

A) The correlation (r^2) for each novel and residual loci calculated by obtaining the individual level data for all PAGE participants, and correlating the SNP genotype with each of the 10 PCs. The correlation for each of the 10 PCs was plotted on the y-axis, with novel loci plotted above the horizon, and residual loci plotted below. B) The individual level data for all PAGE participants were obtained and plotted in a parallel coordinates plot, such that each PAGE individual is represented by a set of line segments connecting their eigenvalues.

Relevance of Multi-ethnic Genetic Variation to Clinical Care

Not only has the genetic diversity of PAGE improved characterization of previously known associations and enabled the discovery of novel genetic associations, but it has also provided population-specific allele frequencies for clinically relevant variants (CRVs) that will have immediate impact on clinical care. MEGA was designed to include CRVs from well-known and frequently used knowledge bases.¹⁹ A finding within our analyses shows an association between *HBB* (rs3334) and HbA1c levels ($P_{\text{cond}}=6.87 \times 10^{-31}$; $N=11,178$), with the majority of the association among Hispanic/Latinos ($P=7.65 \times 10^{-27}$; $N=10,408$; Coded Allele Frequency=0.01), followed by African Americans ($P=5.62 \times 10^{-4}$; $N=559$; CAF=0.06). The lead SNP, rs3334, is a missense variant in *HBB*, which encodes the adult hemoglobin beta chain and is known for its role in sickle cell anemia. Although this association was recently reported in African Americans²⁸, this is the first time it has been reported in Hispanic/Latinos with admixed European, African, and Native American ancestry. Hemoglobin genetic variants are also known to affect the performance of some HbA1c assays^{29–31}, potentially leading practitioners to incorrectly believe that a patient has achieved glucose control. This conclusion leaves the patient more susceptible to type II diabetes (T2D) complications. Alternative long-term measures of glucose control that are not impacted by hemoglobin variants, such as the fructosamine test, should be considered for sickle cell carriers being evaluated for T2D. This result illustrates how ancestry-specific findings may be transferable to other groups that share the same genetic ancestry, such as, in this case, the African ancestry present in both African Americans and some Hispanic/Latinos.

We also investigated the *HLA-B*57:01* haplotype, which interacts with the HIV drug abacavir to trigger a potentially life-threatening immune response in 5–8% of patients.^{32–34} The FDA recommends screening all patients for *HLA-B*57:01*, prior to starting abacavir treatment.³⁵ The rs2395029 variant in *HCP5*, a near perfect tag of *HLA-B*57:01*, is used to screen for abacavir hypersensitivity.³⁶ Using PAGE and Global Reference Panel samples, we show that risk allele (T) frequencies for rs2395029 rise above 5% in multiple large South Asian populations, and rise above 1% within some, but not all, admixed populations with Native American ancestry (**Figure 5**). Thus, the population attributable risk for this variant varies between continental populations and also within sub-continental regions. The allele frequencies from PAGE for clinically relevant variants, particularly polymorphisms with a medical guideline, will be available through several online databases, including ClinGen and dbSNP, to further help researchers and clinicians identify at-risk groups. PAGE allele frequencies can therefore aid in expanding the reach of precision medicine to encompass individuals of diverse ancestry.



1
2 **Figure 5: World map of HLA-B*57:01 frequencies.**

3 *The pharmacogenetic haplotype HLA-B*57:01 interacts with the HIV drug abacavir to stimulate a hypersensitivity*
4 *response. A variant in a nearby gene, HCP5 rs2395029 (G allele), can be used to genotype for the star allele because it*
5 *has been shown to be in linkage disequilibrium with HLA-B*57:01*^{36–38}. This HCP5 SNP segregates within all continental
6 *populations of the PAGE study, providing increased resolution of the global haplotype frequency, particularly within Latin*
7 *America. Above, minor allele (G) frequency is shown. Population size is indicated by the radius of the circle. Black dot*
8 *(n.d.): population has less than twenty individuals or the variant is a singleton in that population.*

9 Discussion

10 Using a multi-ethnic study with the novel MEGA product and methods for analyzing admixed populations, we
11 provide empirical evidence supporting theoretical concerns regarding the European-centric bias in GWAS. To our
12 knowledge, this is the first time effect heterogeneity in the GWAS Catalog has formally been assessed, and the
13 observation that a quarter of GWAS Catalog tagSNPs show evidence of effect heterogeneity by genetic ancestry has
14 profound implications for precision medicine. Furthermore, our results suggest that a majority of GWAS catalog
15 associations are fine-mapped in a multi-ethnic population, consistent with differential LD between tagSNP and functional
16 variant across populations. It is imperative that clinically relevant variants are validated in diverse populations to prevent
17 the use of imprecise genetic tags in clinical applications. Genetic tests are already being used to guide clinical decisions,
18 and efforts to develop polygenic risk prediction models are currently underway. Researchers need to be aware of the
19 limitations of tagSNPs that have not been replicated in non-European populations.¹¹

20 This study also provides evidence that a significant number of novel loci (as well as independent, secondary
21 alleles in known loci) relevant to non-European ancestries remain to be identified, many of which are undiscoverable in
22 European-only study populations due to low allele frequencies in Europeans. Cumulatively, these results expose several
23 shortcomings that arise from an overreliance on European GWAS.

24 The findings from this research demand a reevaluation of how future genetic studies are designed and
25 implemented. As next-generation sequencing, precision medicine, and direct-to-consumer genetic testing become more
26 common, it is critical that the genetics community takes a forward-thinking approach towards research in diverse
27 populations. The increasing ability to identify rare variants further highlights the necessity to study genetically diverse

1 populations, as rare variation is more likely to be ancestry specific. The All of Us Research Program embraces the reality
2 that the success of precision medicine requires precision genomics and therefore emphasizes the recruitment and active
3 participation of underrepresented minorities ³⁹. It is in the best interest of our research community to follow suit and take
4 steps to become more inclusive. As world populations become increasingly diverse ^{40,41}, geneticists and clinicians will be
5 required to evaluate genetic predictors of complex traits in non-Europeans. Our current genomic databases are not
6 representative of populations with the greatest health burden or that will ultimately benefit from this work. This realization,
7 combined with the increased availability of resources for studying diverse populations, means that researchers and
8 funders can no longer afford to ignore non-European populations. This study provides evidence and motivation to make
9 research in diverse populations a priority in the field of genetics.

1 Methods

2 **Studies.** The PAGE consortium includes eligible minority participants from four studies. The Women's Health Initiative
3 (WHI) is a long-term, prospective, multi-center cohort study investigating post-menopausal women's health in the US and
4 recruited women from 1993-1998 at 40 centers across the US. WHI participants of European descent were excluded from
5 this analysis. The Hispanic Community Health Study / Study of Latinos (HCHS/SOL) is a multi-center study of
6 Hispanic/Latinos with the goal of determining the role of acculturation in the prevalence and development of diseases
7 relevant to Hispanic/Latino health. Starting in 2006, household sampling was used to recruit self-identified
8 Hispanic/Latinos from four sites in San Diego, CA, Chicago, IL, Bronx, NY, and Miami, FL. All SOL Hispanic/Latinos were
9 eligible for this study. The Multiethnic Cohort (MEC) is a population-based prospective cohort study recruiting men and
10 women from Hawaii and California, beginning in 1993, and examines lifestyle risk factors and genetic susceptibility to
11 cancer. Only the African American, Japanese American, and Native Hawaiian participants for MEC were included in this
12 study. The BioMe™ BioBank is managed by the Charles Bronfman Institute for Personalized Medicine at Mount Sinai
13 Medical Center (MSMC). Recruitment began in 2007 and continues at 30 clinical care sites throughout New York City.
14 BioMe participants were African American (25%), Hispanic/Latino, primarily of Caribbean origin (36%), Caucasian (30%),
15 and Others who did not identify with any of the available options (9%). Biobank participants who self-identified as
16 Caucasian were excluded from this analysis. The Global Reference Panel (GRP) was created from Stanford-contributed
17 samples to serve as a population reference dataset for global populations. GRP individuals do not have phenotype data
18 and were only used to aid in the evaluation of genetic ancestry in the PAGE samples. Additional information about each
19 participating study can be found in the Supplementary Information.

20
21 **Phenotypes.** The 26 phenotypes included in this study were previously harmonized across the PAGE studies. They
22 include: White Blood Cell (WBC) count, C-Reactive Protein (CRP), Mean Corpuscular Hemoglobin Concentration
23 (MCHC), Platelet Count (PLT), High Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Total Cholesterol (TC),
24 Triglycerides (TG), glycated hemoglobin (HbA1c), Fasting Insulin (FI), Fasting Glucose (FG), Type II Diabetes (T2D),
25 Cigarettes per Day (CPD), Coffee Consumption, QT interval, QRS interval, PR interval, Systolic Blood Pressure (SBP),
26 Diastolic Blood Pressure (DBP), Hypertension (HT), Body Mass Index (BMI), Waist-to-hip ratio (WHR), Height (HT),
27 Chronic Kidney Disease (CKD), End-Stage Renal Disease (ESRD), and Estimated glomerular filtration rate (eGFR) by the
28 CKD-Epi equation. Single variant association testing was completed for all phenotypes using phenotype-specific models,
29 adjusting by indicators for study, self-identified race/ethnicity as a proxy for cultural background, phenotype-specific
30 standard covariates, and the first 10 PCs. Additional information about phenotype-specific cleaning, exclusion criteria, and
31 the model covariates are included in the Supplementary Information.

32
33 **Genotyping.** A total of 53,338 PAGE and GRP samples were genotyped on the MEGA array at the Johns Hopkins Center
34 for Inherited Disease Research (CIDR), with 52,878 samples successfully passing CIDR's QC process. Genotyping data
35 that passed initial quality control at CIDR were released to the Quality Assurance / Quality Control (QA/QC) analysis team
36 at the University of Washington Genetics Coordinating Center (UWGCC). The UWGCC further cleaned the data
37 according to previously described methods⁴², and returned genotypes for 51,520 subjects. A total of 1,705,969 SNPs
38 were genotyped on the MEGA. Quality Control of genotyped variants was completed by filtered through various criteria,
39 including the exclusion of (1) CIDR technical filters, (2) variants with missing call rate $\geq 2\%$, (3) variants with more than 6
40 discordant calls in 988 study duplicates, (4) variants with greater than 1 Mendelian errors in 282 trios and 1439 duos, (5)
41 variants with a Hardy-Weinberg p-value less than 1×10^{-4} , (6) SNPs with sex difference in allele frequency ≥ 0.2 for
42 autosomes/XY, (7) SNPs with sex difference in heterozygosity > 0.3 for autosomes/XY, (8) positional duplicates. Sites
43 were further restricted to chromosomes 1-22, X, or XY, and only variants with available strand information. After SNP QC,
44 a total of 1,402,653 MEGA variants remained for further analyses.

45
46 **Imputation.** In order to increase coverage, and thus improve power for fine-mapping loci, all PAGE individuals who were
47 successfully genotyped on MEGA were subsequently imputed into the 1000 Genomes Phase 3 data release⁴³.
48 Imputation was conducted at the University of Washington Genetic Analysis Center (GAC). Genotype data which passed
49 the above quality control filters was phased with SHAPEIT2⁴⁴ and imputed to 1000 Genomes Phase 3 reference data
50 using IMPUTE version 2.3.2⁴⁵. Segments of the genome which were known to harbor gross chromosomal anomalies
51 were filtered out of the final genotype probabilities files. Imputed sites were excluded if the IMPUTE info score was less
52 than 0.4. A total of 39,723,562 imputed SNPs passed quality control measures. (See Supplemental Methods)

53
54 **Principal Component Analysis.** The selection of unrelated individuals was essential for accurate estimation of the
55 principal components within the global study population. Kinship coefficients were estimated using PC-Relate, as
56 implemented in the R package GENESIS^{20,21}. The *SNPRelate*⁴⁶ package was implemented in R for principal components

1 analysis. The relevant principal components (PCs) were selected using scatter plots. Scatter plots, with various PCs on
2 the x- and y-axes, helped to assess the spread of genetic ancestry within with self-identified racial/ethnic clusters. A
3 parallel coordinate plots for the first 10 PCs was generated, where each PAGE individual is represented by a set of line
4 segments connecting his or her PC values. The amount of variance explained diminished with each subsequent PC, and
5 we estimated that the top 10 PCs provided sufficient information to explain the majority of genetic variation in the PAGE
6 study population.

7
8 **Genome-Wide Association Testing.** All imputed autosomal variants with IMPUTE info score >0.4 ($n=39,723,562$) were
9 eligible for association testing in phenotype-specific models. An effective sample size (effN) was calculated for each SNP
10 in a given phenotype-specific model, where $\text{effN} = 2 \cdot \text{MAF} \cdot (1 - \text{MAF}) \cdot N \cdot \text{info}$, where MAF is the minor allele frequency
11 among the set of individuals included in a phenotype-specific model, N is the total sample size for a given phenotype, and
12 info is the SNP's IMPUTE info score. Variants with an effN less than 30 (continuous phenotypes) or 50 (binary
13 phenotypes), were excluded from the final set of phenotype-specific results. QQ plots and λ_{GC} were used to assess
14 genomic inflation in all phenotypes, for which λ s ranged from 0.98 to 1.15. Single-variant association testing for
15 each phenotype used an additive model that was adjusted by indicators for study, self-identified race/ethnicity, the first 10
16 PCs, and phenotype-specific covariates.

17 Additional information about the phenotype-specific model covariates and transformations are included in the
18 Supplementary Information. Association testing was completed in both SUGEN and GENESIS programs.

19 The GENESIS program ²² is a Bioconductor package made available in R that was developed for large-scale
20 genetic analyses in samples with complex structure including relatedness, population structure, and ancestry admixture.
21 The current version of GENESIS implements both linear and logistic mixed model regression for genome-wide association
22 testing. The software can accommodate continuous and binary phenotypes. The GENESIS package includes the program
23 PC-Relate, which uses a principal component analysis based method to infer genetic relatedness in samples with
24 unspecified and unknown population structure. By using individual-specific allele frequencies estimated from the sample
25 with principal component eigenvectors, it provides robust estimates of kinship coefficients and identity-by-descent (IBD)
26 sharing probabilities in samples with population structure, admixture, and HWE departures. It does not require additional
27 reference population panels or prior specification of the number of ancestral subpopulations.

28 The SUGEN program ²³ is a command-line software program developed for genetic association analysis under
29 complex survey sampling and relatedness patterns. It implements the generalized estimating equation (GEE) method,
30 which does not require modeling the correlation structures of complex pedigrees. It adopts a modified version of the
31 "sandwich" variance estimator, which is accurate for low-frequency SNPs. Association testing in SUGEN requires the
32 formation of "extended" families by connecting the households who share first degree relatives or either first- or second-
33 degree relatives. Trait values are assumed to be correlated within families but independent between families. In our
34 experience in analyzing this dataset, it is sufficient to account for first-degree relatedness. The current version of SUGEN
35 can accommodate continuous, binary, and age-at-onset traits. A comparison of p-values produced by SUGEN and
36 GENESIS for all previously identified known loci are included in Extended Data Fig. 5.

37
38 **Conditional Analyses.** Phenotype-specific lists of previously identified "known loci" were hand-curated for each
39 phenotype and included SNPs indexed in the GWAS Catalog or identified through non-GWAS high-throughput methods
40 (e.g. Metabochip, Exomechip, ImmunoChip, etc.). The full known loci lists for each phenotype are available in the
41 Supplementary Table 5. Conditional analyses were conducted for all phenotypes by conditioning on all previously
42 identified known loci on a given chromosome. P-values estimated in conditional analyses are denoted by " P_{cond} " in the
43 main text, with the SUGEN conditional results for all novel and residual findings in Supplementary Table 3.

44
45 **Effect Heterogeneity by Genetic Ancestry and Self-Identified Race/Ethnicity.** We used two approaches to assess
46 effect heterogeneity within PAGE participants. First, we used interaction analyses with models that included variant by PC
47 (SNPxPC) interaction terms for all 10 PCs. The fit of nested models was compared using the F-statistic, where the
48 associated interaction p-value indicated whether the inclusion of the 10 SNPxPC interaction terms improved the model fit
49 compared to a model that lacked the interaction terms. The overall SNPxPC interaction p-values evaluated whether the
50 additional variance explained by variant x genetic ancestry interactions was statistically significant, and represent effect
51 modification driven by genetic ancestry. Interaction p-values for all novel and residual findings are included in
52 Supplementary Table 3.

53 For comparison against more traditional (stratified) analysis strategies, all analyses were also run stratified by
54 self-identified race/ethnicity. A minor allele count of at least 5 was required for a stratified model to be run within an ethnic
55 group. The stratified analyses were then meta-analyzed using a fixed-effect model implemented in METAL⁴⁷. I^2 and χ^2
56 heterogeneity p-values were estimated for all meta-analyzed results, and represent effect size heterogeneity driven by
57 self-identified race/ethnicity. The race/ethnicity-specific results, I^2 , and χ^2 heterogeneity p-values for all novel and

1 residual findings are included in Supplementary Table 3.

2
3 **Assessing Single-Variant Results.** SUGEN association results were used for the identification of novel and residual
4 findings for all phenotypes. The variant with the smallest p-value in a 1Mb region was considered the “lead SNP”. A lead
5 SNP was considered to be a novel loci if it met the following criteria: 1) the lead SNP was located greater than +/- 500 Kb
6 away from a previously known loci (per the phenotype-specific known loci list); 2) had a SUGEN p-value less than 5×10^{-8} ;
7 3) had a SUGEN conditional p-value less than 5×10^{-8} after adjustment for all previously known loci on the same
8 chromosome; and 4) had 2 or more neighboring SNPs (within +/- 500 Kb) with a p-value less than 1×10^{-5} . A lead SNP was
9 considered to be a residual signal in a previously known loci if it met the following criteria: 1) the lead SNP was located
10 within +/- 500 Kb of a previously known loci; 2) had a SUGEN p-value less than 5×10^{-8} ; and 3) had a SUGEN conditional
11 p-value less than 5×10^{-8} after adjustment for all previously known loci on the same chromosome. Full results for all novel
12 and residual findings are included in Supplementary Table 2-3.

13
14 **GWAS Catalog Heterogeneity.** The full GWAS Catalog database was downloaded on December 31, 2016. The data
15 were filtered to identify results relevant to any of the 26 PAGE phenotypes, producing a subset of 3,322 unique tagSNPs
16 that were genome-wide significant ($p < 5 \times 10^{-8}$) in the GWAS Catalog. The PAGE results for each of the 3,322 GWAS
17 Catalog tagSNPs was examined to first identify the subset of tagSNPs that replicated ($p < 5 \times 10^{-8}$) in PAGE unconditioned
18 models ($n=574$). Pairs of tagSNPs within 500,000 base pairs of each other were merged into loci, yielding 302 unique
19 associated loci. Of the GWAS Catalog tagSNPs that were replicated in PAGE, SNPs that had a Bonferroni corrected
20 SNPxPC interaction heterogeneity p-value ($p < 8.71 \times 10^{-5}$, $0.05/574$) were considered to have evidence of effect size
21 heterogeneity (132/574, 23.0%). Effect heterogeneity was also assessed using PAGE’s multi-ethnic study population by
22 first identifying the “lead SNP” in each locus with the smallest p-value in PAGE, totalling 333 SNPs (302 known loci from
23 the GWAS catalog, plus 31 novel loci discovered in the present analysis). Among the 333 lead SNPs, 24 (7.2%) had a
24 significant Bonferroni corrected SNPxPC interaction heterogeneity p-value ($P < 1.5 \times 10^{-4}$, $0.05/333$).

25
26 **Allele frequency estimation.** Population labels were compiled from self-identified ancestry information from the PAGE-
27 wide sample manifest, as well as self-reported country of origin metadata from the Mount Sinai BioMe cohort. Allele
28 frequencies were calculated in PLINK 1.90, and results were visualized in R using the ggplot2.

1 References

- 2 1. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends*
3 *Genet* **25**, 489–494 (2009).
- 4 2. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011).
- 5 3. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
- 6 4. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*
7 **108**, 11983–11988 (2011).
- 8 5. SIGMA Type 2 Diabetes Consortium *et al.* Association of a low-frequency variant in HNF1A with type 2 diabetes in a
9 Latino population. *JAMA* **311**, 2305–2314 (2014).
- 10 6. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with
11 prostate cancer. *Nat Genet* **44**, 1326–1329 (2012).
- 12 7. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes.
13 *Nature* **512**, 190–193 (2014).
- 14 8. Kenny, E. E. *et al.* Melanesian blond hair is caused by an amino acid change in TYRP1. *Science* **336**, 554 (2012).
- 15 9. Manning, A. *et al.* A Low-Frequency Inactivating Akt2 Variant Enriched in the Finnish Population is Associated With
16 Fasting Insulin Levels and Type 2 Diabetes Risk. *Diabetes* (2017). doi:10.2337/db16-1329
- 17 10. Carlson, C. S. *et al.* Generalization and dilution of association results from European GWAS in populations of non-
18 European ancestry: the PAGE study. *PLoS Biol* **11**, e1001661 (2013).
- 19 11. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J*
20 *Hum Genet* **100**, 635–649 (2017).
- 21 12. Liao, Y. *et al.* Surveillance of health status in minority communities - Racial and Ethnic Approaches to Community
22 Health Across the U.S. (REACH U.S.) Risk Factor Survey, United States, 2009. *MMWR Surveill Summ* **60**, 1–44
23 (2011).
- 24 13. Satcher, D. From the Surgeon General: Eliminating global health disparities. *JAMA* **284**, 2864 (2000).
- 25 14. Oh, S. S. *et al.* Diversity in clinical and biomedical research: A promise yet to be fulfilled. *PLoS Med* **12**, e1001918
26 (2015).
- 27 15. Carlson, C. S. Ethnicity: Diversity is future for genetic analysis. *Nature* **540**, 341 (2016).
- 28 16. Matise, T. C. *et al.* The Next PAGE in understanding complex traits: design for the analysis of Population
29 Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol* **174**, 849–859 (2011).

- 1 17. Johnston, H. R. *et al.* Identifying tagging SNPs for African specific genetic variation from the African Diaspora
2 Genome. *Sci Rep* **7**, 46398 (2017).
- 3 18. Wojcik, G. L. *et al.* Imputation aware tag SNP selection to improve power for multi-ethnic association studies. *bioRxiv*
4 (2017). at <<http://biorxiv.org/content/early/2017/02/03/105551>>
- 5 19. Bien, S. A. *et al.* Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping
6 array. *PLoS ONE* **11**, e0167758 (2016).
- 7 20. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and
8 correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276–293 (2015).
- 9 21. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness.
10 *Am J Hum Genet* **98**, 127–148 (2016).
- 11 22. Conomos, M. P. *et al.* Genetic diversity and association studies in US hispanic/latino populations: applications in the
12 hispanic community health study/study of latinos. *Am J Hum Genet* **98**, 165–184 (2016).
- 13 23. Lin, D.-Y. *et al.* Genetic association analysis under complex survey sampling: the Hispanic Community Health
14 Study/Study of Latinos. *Am J Hum Genet* **95**, 675–688 (2014).
- 15 24. Byrne, E. M. *et al.* A genome-wide association study of caffeine-related sleep disturbance: confirmation of a role for a
16 common variant in the adenosine receptor. *Sleep* **35**, 967–975 (2012).
- 17 25. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).
18 *Nucleic Acids Res* **45**, D896–D901 (2017).
- 19 26. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A_{1c} (C) levels via glyceemic and
20 nonglyceemic pathways. *Diabetes* **59**, 3229–3239 (2010).
- 21 27. Holm, H. *et al.* Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet* **42**, 117–122
22 (2010).
- 23 28. Lacy, M. E. *et al.* Association of sickle cell trait with hemoglobin a1c in african americans. *JAMA* **317**, 507–515
24 (2017).
- 25 29. Lin, C.-N. *et al.* Effects of hemoglobin C, D, E, and S traits on measurements of HbA1c by six methods. *Clin Chim*
26 *Acta* **413**, 819–821 (2012).
- 27 30. Mongia, S. K. *et al.* Effects of hemoglobin C and S traits on the results of 14 commercial glycated hemoglobin
28 assays. *Am J Clin Pathol* **130**, 136–140 (2008).
- 29 31. Roberts, W. L. *et al.* Effects of hemoglobin C and S traits on glycohemoglobin measurements by eleven methods.

- 1 *Clin Chem* **51**, 776–778 (2005).
- 2 32. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med* **358**, 568–579 (2008).
- 3 33. Sousa-Pinto, B. *et al.* Pharmacogenetics of abacavir hypersensitivity: A systematic review and meta-analysis of the
4 association with HLA-B*57:01. *J Allergy Clin Immunol* **136**, 1092–4.e3 (2015).
- 5 34. Hetherington, S. *et al.* Hypersensitivity reactions during therapy with the nucleoside reverse transcriptase inhibitor
6 abacavir. *Clin Ther* **23**, 1603–1614 (2001).
- 7 35. Drug Safety and Availability > Information for Healthcare Professionals: Abacavir (marketed as Ziagen) and
8 Abacavir-Containing Medications. at <<https://www.fda.gov/Drugs/DrugSafety/ucm123927.htm>>
- 9 36. Martin, M. A. *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for HLA-B Genotype and
10 Abacavir Dosing: 2014 update. *Clin Pharmacol Ther* **95**, 499–500 (2014).
- 11 37. Colombo, S. *et al.* The HCP5 single-nucleotide polymorphism: a simple screening tool for prediction of
12 hypersensitivity reaction to abacavir. *J Infect Dis* **198**, 864–867 (2008).
- 13 38. Sanchez-Giron, F. *et al.* Association of the genetic marker for abacavir hypersensitivity HLA-B*5701 with HCP5
14 rs2395029 in Mexican Mestizos. *Pharmacogenomics* **12**, 809–814 (2011).
- 15 39. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793–795 (2015).
- 16 40. - United Nations Population Fund. State of World Population 2016. (2016). at <<http://www.unfpa.org/swop>>
- 17 41. Colby, S. L. & Ortman, J. M. *Projections of the Size and Composition of the U.S. Population: 2014 to 2060.* (United
18 States Census Bureau, 2015).
- 19 42. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies.
20 *Genet Epidemiol* **34**, 591–602 (2010).
- 21 43. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- 22 44. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat*
23 *Methods* **9**, 179–181 (2011).
- 24 45. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation
25 of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
- 26 46. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP
27 data. *Bioinformatics* **28**, 3326–3328 (2012).
- 28 47. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans.
29 *Bioinformatics* **26**, 2190–2191 (2010).

1 **Supplementary Information** is available in the online version of the paper at www.nature.com/nature.

2
3 **Acknowledgements:** The Population Architecture Using Genomics and Epidemiology (PAGE) program is funded by the
4 National Human Genome Research Institute (NHGRI) with co-funding from the National Institute on Minority Health and
5 Health Disparities (NIMHD). The contents of this paper are solely the responsibility of the authors and do not necessarily
6 represent the official views of the NIH. The PAGE consortium thanks the staff and participants of all PAGE studies for
7 their important contributions. We thank Rasheeda Williams and Margaret Ginoza for providing assistance with program
8 coordination. The complete list of PAGE members can be found at <http://www.pagestudy.org>.

9 Assistance with data management, data integration, data dissemination, genotype imputation, ancestry deconvolution,
10 population genetics, analysis pipelines, and general study coordination was provided by the PAGE Coordinating Center
11 (NIH U01HG007419). Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is
12 fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract
13 number HHSN268201200008I. Genotype data quality control and quality assurance services were provided by the
14 Genetic Analysis Center in the Biostatistics Department of the University of Washington, through support provided by the
15 CIDR contract.

16 The data and materials included in this report result from collaboration between the following studies and organizations:

17 **BioMe Biobank:** Samples and data of The Charles Bronfman Institute for Personalized Medicine (IPM) BioMe
18 Biobank used in this study were provided by The Charles Bronfman Institute for Personalized Medicine at the
19 Icahn School of Medicine at Mount Sinai (New York). Phenotype data collection was supported by The Andrea
20 and Charles Bronfman Philanthropies. Funding support for the Population Architecture Using Genomics and
21 Epidemiology (PAGE) IPM BioMe Biobank study was provided through the National Human Genome Research
22 Institute (NIH U01HG007417).

23 **HCHS/SOL:** Primary funding support to Dr. North and colleagues is provided by U01HG007416. Additional
24 support was provided via R01DK101855 and 15GRNT25880008. The Hispanic Community Health Study/Study
25 of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and
26 Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234),
27 Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego
28 State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a
29 transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on
30 Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National
31 Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke,
32 NIH Institution-Office of Dietary Supplements.

33 **MEC:** The Multiethnic Cohort study (MEC) characterization of epidemiological architecture is funded through the
34 NHGRI Population Architecture Using Genomics and Epidemiology (PAGE) program (NIH U01 HG007397). The
35 MEC study is funded through the National Cancer Institute U01 CA164973.

36 **PAGE Global Reference Panel:** The Stanford Global Reference Panel was created by Stanford-contributed
37 samples and comprises multiple datasets from multiple researchers across the world designed to provide a
38 resource for any researchers interested in diverse population data on the Multi-Ethnic Global Array (MEGA),
39 funded by the NHGRI PAGE program (NIH U01HG007419). The authors thank the researchers and research
40 participants who made this dataset available to the community. The specific datasets are:

41 Mexico: Samples of indigenous origin in Oaxaca were kindly provided by Drs. Karla Sandoval Mendoza, Samuel
42 Canizales Quinteros, and Victor Acuña Alonzo. Peru: Individuals from a primarily Quechuan and Aymaran-
43 speaking community in Puno, Peru were kindly provided by Drs. Julie Baker and Carlos Bustamante, with funding
44 support from the Burroughs Welcome Fund. Rapa Nui (Easter Island): Samples were kindly provided by Drs.
45 Karla Sandoval Mendoza and Andres Moreno Estrada with funding from the Charles Rosenkranz Prize for Health
46 Care Research in Developing Countries.

47 South Africa: Samples of KhoeSan individuals from the ̳Khomani and Nama communities were kindly provided
48 by Drs. Brenna Henn and Christopher Gignoux with funding from the Morrison Institute for Population and
49 Resource Studies. Honduras and Colombia: Samples from communities in Honduras and Colombia were kindly
50 provided by Dr. Kathleen Barnes (University of Colorado, Denver), Edwin Herrero-Paz (Universidad Católica de
51 Honduras, San Pedro Sula, Honduras), Alvaro Mayorga (Universidad Católica de Honduras, San Pedro Sula,
52 Honduras), Luis Caraballo (University of Cartagena), Javier Marrugo (university of Cartagena) Additional global
53 samples: The following datasets are open access and available through the lab website of Carlos Bustamante
54 (<https://bustamantelab.stanford.edu/>). The Human Genome Diversity Panel (HGDP-CEPH) is a group of cell lines
55 maintained by the Centre d'Étude du Polymorphisme Humain, Fondation Jean Dausset (Paris, France)
56 comprising 52 diverse populations across the world (Africa, Near East, Europe, South Asia, Central Asia, East
57 Asia, Oceania and the Americas). Additional information on these datasets can be found on the CEPH website

(http://www.cephb.fr/en/hgdp_panel.php), or originally at <http://www.ncbi.nlm.nih.gov/pubmed/11954565> and <http://www.ncbi.nlm.nih.gov/pubmed/12493913>, with numerous subsequent publications. Samples were filtered to include the H952 unrelated individuals as published here: <http://www.ncbi.nlm.nih.gov/pubmed/17044859>. Also available on the Bustamante Lab website is genotype data for the Maasai from Kinyawa, Kenya (MKK) samples maintained by the Coriell Institute for Medical Research (<https://catalog.coriell.org/1/NHGRI/Collections/HapMap-Collections/Maasai-in-Kinyawa-Kenya-MKK>) and genotyped as part of the International HapMap Project Phase 3 (<http://hapmap.ncbi.nlm.nih.gov/>, <http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html>). We have genotyped a subset of unrelated individuals using the filters recommended in <http://www.ncbi.nlm.nih.gov/pubmed/20869033>.

WHI: Funding support for the “Exonic variants and their relation to complex traits in minorities of the WHI” study is provided through the NHGRI PAGE program (NIH U01HG007376). The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A listing of WHI investigators can be found at: <https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf>

Individual Acknowledgements: KKN was supported by the Cancer Prevention Training Grant in Nutrition, Exercise and Genetics R25CA094880 from the National Cancer Institute. CRG was supported by NHGRI training grant T32 HG000044. HMH was supported by NHLBI training grant T32 HL007055. AEJ was supported by NIH 5K99HL130580-02 and NIH L60 MD008384-02. KLY supported by NCATS KL2TR001109. JMK was supported by KL2TR000421. RWW was supported by NIH 5T32HD049311-07. D-YL was supported by R01CA082659, R01GM047845, and P01CA142538. LFR was supported by NICHD training grant T32 HD007168 and P2C HD050924. TAT was supported by P01GM099568.

Author Contributions: Overall project supervision and management: ED, J-LA, LRW, RSJ, LAH, SB, CH, CK, LLM, RJFL, TM, KEN, UP, EEK, CSC. Genotyping and quality control: GLW, JH, CRG, NZ, SB, JMK, EPS, KV, GMB, RWW, CS, MHP, MF, CDB, LCP, JR, KD, MPC, XS, CAL, CCL, RD, GN, EB, SCN, CK, UP, EEK, CSC. Phenotype harmonization: MG, KKN, JH, HMH, YMP, AEJ, CJH, CLW, CLA, KLY, MAR, NZ, SB, JMK, IC, VWS, GMB, CS, AV, MHP, GH, LFR, MF, APR, LRW, YL, S-SLP, CPC, RD, GN, EB, SB, CK, LLM, UP, EEK. Association analyses: GLW, MG, KKN, RT, JH, CRG, HMH, YMP, AEJ, BML, CJH, CLW, CLA, KLY, MAR, SB, JMK, IC, VWS, EPS, GMB, MV, YL, D-YL, TAT, J-LA, DOS, YL, S-SLP, CK, UP, EEK, CSC. Manuscript preparation: GLW, MG, KKN, RT, JH, CRG, HMH, YMP, AEJ, BML, CJH, CLW, CLA, KLY, MAR, JMK, IC, VWS, EPS, RWW, AV, LH, D-YL, GH, APR, TAT, DOS, RSJ, LAH, RD, GN, EAS, SB, CH, CK, LLM, RJFL, TM, KEN, UP, EEK, CSC.

Author Information

Reprints and permissions information is available at www.nature.com/reprints.

Competing financial interests: CDB is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, and Etalon and is a founder of CDB Consulting. CRG owns stock in 23andMe. EEK and CRG are members of the scientific advisory board for Encompass Bioscience. EEK consults for Illumina.

Data Availability: Individual-level phenotype and genotype data are available through dbGaP at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000356. Allele frequency data will be available for all genotyped sites on dbSNP (<https://www.ncbi.nlm.nih.gov/projects/SNP/>) and the University of Chicago Geography of Genetic Variants Browser (<http://popgen.uchicago.edu/ggv/>). Clinically-relevant variant frequency data will also be available through ClinGen.

Correspondence and requests for materials should be addressed to eimear.kenny@mssm.edu and ccarlson@fredhutch.org.

Phenotype	Largest GWAS catalog discovery population ¹				PAGE	GWAS catalog tagSNPs			best PAGE tagSNPs		Novel Loci (count) ⁶	Residual Loci (count) ⁶
	European	East Asian	African	Hispanic/Latino		Unique	P<5x10 ⁻⁸	Het. ⁴	P<5x10 ⁻⁸	Het. ⁵		
Inflammatory Traits												
CRP	66,185	10,112	8,280	3,548	28,537	82	38	7	16	1	0	0
WBC	19,509	33,231	16,388	-	28,534	27	10	5	11	3	1	1
MCHC	62,553	-	16,485	-	19,803	21	9	1	5	0	0	2
Platelet Count	48,666	14,806	7,943	12,491	29,328	92	23	0	28	0	1	1
Lipid Traits												
HDL	99,900	12,545	7,917	4,383	33,063	244	71	8	21	1	2	5
LDL	94,595	12,545	7,861	4,383	32,221	192	46	12	18	0	0	2
TG	96,598	12,545	7,601	4,383	33,096	179	75	29	16	1	1	2
TC	100,184	8,344	6,480	4,383	33,185	166	31	4	20	0	1	3
Lifestyle Traits												
Cigarettes/Day Excluding Nonsmokers	74,035	11,696	32,389	-	15,862	12	0	0	3	0	2	1
Coffee Cups/Day	91,462	-	-	-	35,902	16	3	1	3	0	1	0
Glycemic Traits												
HbA1c	46,368	17,290	-	-	11,178	29	8	1	9	0	2	3
Fasting Insulin	51,750	7,696	1,040	229	21,596	34	0	0	3	0	1	0
Fasting Glucose	58,074	24,740	2,029	4,176	23,963	55	15	3	7	0	2	0
Type II Diabetes ²	12,171/ 56,862	15,463/ 26,183	1,264/ 5,678	3,848/ 4,366	14,075/ 31,752	286	28	2	13	0	0	1
Electrocardiogram Traits												
QT Interval	71,061	6,805	13,105	-	17,348	183	39	1	11	0	0	2
QRS Interval	60,255	6,085	13,031	-	17,052	63	9	3	12	0	1	2
PR Interval	28,517	6,085	13,415	-	17,428	154	19	1	10	0	1	2
Blood Pressure Traits												
Systolic Blood Pressure	74,064	31,516	29,378	-	35,433	74	2	0	4	0	1	1
Diastolic Blood Pressure	74,064	31,516	29,378	-	35,433	81	2	0	4	0	0	0
Hypertension	74,064	31,516	29,378	-	49,158	111	0	0	2	0	1	1
Anthropometric Traits												
Waist-to-hip Ratio ³	142,762	39,869	19,744	3,484	33,904	94	5	0	6	0	1	0
Height	253,288	36,227	20,427	-	49,781	698	99	42	93	18	5	13
Body Mass Index	236,781	82,438	39,144	3,484	49,335	572	41	12	13	0	1	0

Kidney Traits												
eGFR by CKD Epi Equation	133,413	23,536	16,840	16,325	27,900	135	1	0	5	0	3	0
Average	90,953	20,953	14,710	5,570	Total	3356	548	194	333	24	28	42

Table 1: GWAS Catalog heterogeneity by Trait, including number of novel and residual findings.

¹ Only includes studies indexed in the GWAS Catalog on December 31, 2016

² Cases/Controls

³ Includes pooled and sex-stratified studies / results

⁴ $P < 8.71 \times 10^{-5}$ for genotype:PC interaction in PAGE, adjusting for multiple tests (0.05/574)

⁵ $P < 1.50 \times 10^{-4}$ for genotype:PC interaction in PAGE, adjusting for multiple tests (0.05/333)

⁶ Significant loci have $P < 5 \times 10^{-8}$ after conditioning on all known loci from the literature

1
2
3
4
5
6
7
8
9
10