Published in final edited form as:

Pharmacogenet Genomics. 2016 April ; 26(4): 147–153. doi:10.1097/FPC.0000000000000200.

A Comparison of DMET Plus Microarray and Genome-Wide Technologies by Assessing Population Substructure

Jami N. Jackson¹, Kevin M. Long^{3,4}, Yijing He^{4,5,6}, Alison. A. Motsinger-Reif^{1,2,4}, Howard L. McLeod^{3,4,5,6}, and John Jack^{1,2}

¹Department of Statistics, North Carolina State University, Raleigh, NC, USA

²Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA

³Ortelion LLC, Raleigh, NC, USA

⁴Pharmacogenetics for Every Nation Initiative, Raleigh, NC, USA

⁵DeBartolo Family Personalized Medicine Institute, Moffitt Cancer Center, Tampa, FL, USA

⁶Xiangya Hospital, Central South University, Changsha, China

Abstract

Objective—The capacity for the Affymetrix DMET Plus pharmacogenomics genotyping chip to estimate population substructure and cryptic relatedness was evaluated. The results were compared with estimates using genome-wide HapMap data for the same individuals.

Methods—For 301 unrelated Individuals, spanning three continental populations and one admixed population, genotypic data was collected using the Affymetrix DMET Plus microarray. Genome-wide data on these individuals was obtained from HapMap release 3. Population substructure was assessed using Eigenstrat and ADMIXTURE software for both platforms. Cryptic relatedness was explored via inbreeding coefficient estimation. Nonparametric tests were used to determine correlations of the analytical results of the two genotyping platforms.

Results—Principal components analysis identified population substructure for both datasets with 15.8% and 16.6% of the total variance explained in the first two principal components for DMET Plus and HapMap data, respectively. ADMIXTURE results correctly identified four subpopulations within each dataset. Nonparametric rank correlations indicated significant associations between analyses with an average $\rho = .7272$ ($p < 10^{-16}$) across the 3 continental populations and $\rho = .4888$ for the admixed population. Concordance correlation coefficients (average $\rho_c = .9693$ across all four subpopulations) strongly indicate concordance between ADMIXTURE results. Inbreeding coefficients were slightly inflated (16 individuals $>.15$) using DMET Plus data, and no cryptic relatedness was indicated using HapMap data.

Conclusions—The DMET Plus pharmacogenomics genotyping chip is effective in quantifying population substructure across the three continental populations and inferring the presence of an

Correspondence and request for reprints: John Jack, North Carolina State University, Bioinformatics Research Center, Campus Box 7566, 2601 Stinson Drive, Raleigh, NC 27695-7566.

Conflict of Interest and Source of Funding: The authors have no additional conflicts of interests to declare.

admixed population. Based on our results, these microarrays offer sufficient depth for covariate adjustment of population substructure in genomic association studies.

Keywords

pharmacogenomics; population stratification; DMET Plus; metabolism; transporters; affymetrix; principal component analysis; admixture; inbreeding coefficients; cryptic relatedness

Introduction

Genome-wide association studies (GWAS) are typically performed on populations of unrelated individuals to decipher the relationships between genetic variations and complex traits. When study designs involve individuals from diverse ethnic or geographic backgrounds, population substructure or stratification can play a potentially confounding role in the association results. Self-reported race and ethnicity is an unreliable metric for accurately assessing population structure. For best practices, it is crucial to make inferences on genetic ancestry from genotype information and incorporate population substructure in the covariates of genetic association modeling. Historically, population stratification has been detected using microsatellite and biallelic markers [1], genomic control [2], structured association [3], and ancestry informative markers [4]. Given its computational efficiency and proven ability to detect substructure, principal components analysis (PCA) is considered the standard approach in controlling for population stratification.

The International HapMap Project (HapMap), a multi-country effort to identify and catalog genetic similarities and differences in humans, analyzed DNA from populations with African, Asian, and European ancestry [5]. The most recent data from this ambitious undertaking, the HapMap 3 samples, were collected on two genome-wide technology platforms, the Illumina Human1M and the Affymetrix SNP 6.0 [6]. These genotyping platforms have been used in numerous GWAS [7–10] with varying degrees of success. The impressive scope and resolution of the HapMap 3 technology readily allows inference of population substructure for mapping genes across the entire human genome.

Newer and more cost-effective technologies were recently developed to conduct rapid genotyping and specifically capture genetic differences in drug metabolism enzymes and transporters (DMET). One such technology, the Affymetrix DMET Plus microarray, is capable of genotyping a large number of known markers (1,936 variants spanning 225 genes) in 2.5 days [11]. The cost/speed benefits of this platform were capitalized in several recent studies: St. Jude Children's Research Hospital implemented pre-emptive pharmacogenetics for over 1000 patients [12]; a pharmacogenetic study of thalidomide and docetaxel in patients with castration-resistant prostate cancer [13]; and a genetic prediction model created to identify patients with low paclitaxel clearance [14].

Ethnic differences in the allele frequencies of DMET Plus genes have been shown [15]. Moreover, genetic differences among DMET Plus genes have been found in the three continental populations: European, African, and East Asian populations [16]. However, there is no systematic evaluation in the literature which compares the DMET Plus microarray to

genome-wide technologies of HapMap in order to infer genetic ancestry and accurately quantify population substructure for subsequent analytical methods.

For this work, we used data from 301 individuals genotyped with both DMET Plus and HapMap technology platforms. These samples span four self-reported races—U.S. Mexican, Caucasian, East Asian, and African populations. Multiple statistical methods were used to make inferences on genetic ancestry and population substructure and the results from each technology were directly compared. This work supports the hypothesis that the DMET Plus microarray is a sufficient alternative to genome-wide genotyping technologies for a targeted, cost-effective approach to association mapping and pharmacogenomics research.

Methods

DNA Samples

Single nucleotide polymorphism data were obtained for 301 unrelated samples using the DMET Plus microarray. These individuals were selected for DMET Plus genotyping based on two criteria: (i) each sample had genome-wide data available from HapMap and (ii) each sample belonged to one of four self-reported ethnicities. The breakdown of individuals by self-reported ethnicity is the following: 50 Mexican (MEX) individuals from Los Angeles, California, 84 Chinese (CHB) individuals from Beijing, China, 112 Yoruban (YRI) individuals from Ibadan, Nigeria, and 55 Caucasian (CEU) individuals from Utah.

The DMET Plus genotype data have passed a rigorous quality control (QC) pipeline which we will summarize. The raw data contained 1,936 genetic variants across 301 samples [17–18]. These data were filtered based on missingness per individual (5%), missingness per variant (5%), and deviations from Hardy-Weinberg equilibrium stratified for each country and ethnicity combination. The total number of variants which passed quality control was 1,839. Seven of the SNPs were dropped for this analysis because they reported more than two alleles across the global population; as such, they were not amenable to comparison using standard techniques for correcting for population stratification. We did not want to confound our comparison approach by using nonstandard techniques for population stratification analysis.

We obtained the genotype data from the HapMap 3 release for the 301 individuals. The intersection of the SNPs genotyped for HapMap across all four populations was 1,126,226. In summary, after DMET Plus QC and merger of the HapMap data, there were 301 individuals with genetic data on 1,832 variants from the DMET Plus microarray and 1,126,226 SNPs from HapMap genotyping technology. For the remainder of this work, we will refer to the datasets as either DMET Plus or HapMap.

Statistical Analysis

Population structure was calculated with two distinct yet common software packages encompassing alternative statistical methods for identifying population substructure. First, PCA was performed using Eigenstrat via SMARTPCA Version 8000 [19]. PCA has been used to characterize population substructure in genomic studies for decades [20–23]. It is a popular method for population stratification correction in unrelated samples because it is

efficient and computationally tractable with large genomic datasets. PCA has also been proven useful in cases in which genome-wide data is unavailable, such as studies involving smaller numbers of ancestry-informative markers [24–25].

Second, we estimated individual ancestries using ADMIXTURE Version 1.23 [26]. We performed ten-fold cross validation to determine the optimal value for K for each dataset. Next, we ran ADMIXTURE on the DMET Plus and HapMap data. For HapMap, we first ran ADMIXTURE on a random selection of 1,832 markers. We then used PLINK (Version 1.07) [27] to LD-prune the entire HapMap data (window size = 50, step size = 10, $r^2 = .1$), which resulted in a reduced set of 77,437 SNPs in linkage equilibrium. ADMIXTURE analysis was carried out on these remaining SNPs. The LD-pruning was necessary per author discretion [26]. A nonparametric rank correlation was used to determine the significance of association between clusters inferred by ADMIXTURE from DMET Plus and the 77,437 SNPs from HapMap. Lin's concordance correlation coefficients were calculated using the epiR package in the R statistical language.

Lastly, BED, BIM, and FAM files were created using PLINK in order to calculate the inbreeding coefficients using R (Version 3.1.1) and the R packages, SNPRelate (Version 0.9.18) and gdsfmt (Version 1.1.0.1), for each of the individuals genotyped with the DMET Plus chip and the genotyping technology [27–29]. Voight and Pritchard [30] showed that for association studies with no serious sampling biases in large outbred populations, the effect of cryptic relatedness is expected to be negligible, but may have a more serious effect in small, growing populations. Inbreeding coefficients were calculated on the full DMET Plus dataset and the full HapMap datasets using Visscher's estimator described in [31]. Additionally, inbreeding coefficients were calculated on a reduced set of 1,832 markers randomly selected from HapMap. A nonparametric rank correlation was used to determine association significance of the inbreeding coefficients on DMET Plus and the full HapMap data.

Results

Principal Component Analysis

Pairwise comparisons for the first three principal components (PCs) are illustrated in six scatterplots (Fig. 1). Each point represents an individual and the colors were selected to identify self-reported race. Panel A shows the results for the analysis of DMET Plus and panel B shows the results of the HapMap data. The ratio of explained to total variance in the first two PCs was similar for the DMET Plus genotypes (15.8%) and HapMap genotypes (16.6%). Moreover, the PCA plot for each dataset shows a clear separation of the three continental populations (CEU, YRI, and CHB) along the PC 1 and PC 2 axes. For both genotyping platforms, the MEX population clusters closest to the CEU population, and clusters closer to the CHB population than to the YRI population—i.e., the MEX identifies as an admixture of CEU and CHB. For HapMap, there is a clearer separation of the PCs between the CEU and MEX populations than there is with the DMET Plus genotypes.

ADMIXTURE Analysis

The Q values from the ADMIXTURE analysis are visually demonstrated in the two stacked bar plots (Fig. 2). Each bar along the x-axis corresponds to an individual and the y-axis represents the Q value pertaining to each of the four clusters. Panel A shows the results for the analysis of DMET Plus and panel B shows the results of the HapMap data. The individuals are sorted along the x-axis according to their self-reported ethnicities.

From this analysis, it is clear that both DMET Plus and HapMap genotyping platforms are able to infer population structure across the three continental populations. Indeed, the concordance correlation coefficients for each cluster strongly indicate concordance in the population substructure estimated by ADMIXTURE (Cluster 1: $\rho_c = .9718$, Cluster 2: $\rho_c = .9951$, Cluster 3: $\rho_c = .9908$, Cluster 4: $\rho_c = .9194$). The MEX block of individuals highlights the detection of admixture (Q4) in both analyses. A nonparametric rank correlation between each inferred cluster of the DMET Plus and HapMap genotypes from the ADMIXTURE analyses indicates that there is a statistically significant correlation between clusters from the different technologies (Cluster 1: p-value $< 10^{-16}$ and $\rho = .7245$, Cluster 2: p-value $< 10^{-16}$ and $\rho = .7701$, Cluster 3: p-value $< 10^{-16}$ and $\rho = .6870$, and Cluster 4: p-value $< 10^{-16}$ and $\rho = .4888$).

Inbreeding Coefficients Analysis

The inbreeding coefficients for DMET Plus and HapMap are shown as two scatterplots (Fig. 3). The mean inbreeding coefficient for the DMET Plus genotypes is 0.06922 and the mean inbreeding coefficient for the HapMap genotypes is 0.08985. We compared the number of individuals with inbreeding coefficients that were > 0.15 and found that there were 16 individuals from the DMET Plus genotypes that had inbreeding coefficients > 0.15 and there were zero individuals from the HapMap genotypes that had inbreeding coefficients > 0.15 . Spearman's rank correlation between the estimated inbreeding coefficients for DMET Plus and HapMap indicated a statistically significant positive correlation ($p < 10^{-7}$ and $\rho = 0.2999$).

Discussion

A study of European genetic substructure using 300K SNPs found aspects of substructure not revealed in previous studies that used a maximum of 10K SNPs and further showed that for principal components greater than 1, sets of SNPs smaller than 11K had a lower correlation with the full 300K SNP set than with random sets greater than 40K SNPs [32]. The number of markers in the DMET Plus microarray is several orders of magnitude smaller than the technologies used in HapMap 3; hence, as we've shown in this study, the capacity for DMET Plus to differentiate between global populations while exhibiting admixture on MEX between CEU and CHB on this limited number of markers is impressive. Moreover, our results indicate that the Affymetrix DMET Plus microarray is capable of determining population substructure adequately when compared with analyses using genome-wide HapMap 3 data.

The PCA of the DMET Plus and HapMap datasets resulted in a clear separation of the three major continental populations: CEU, CHB, and YRI. While both datasets exhibited clustering of the MEX population, the results from the HapMap data yielded a more definitive separation between the MEX population and the CEU population on the 3rd and 4th PCs. These differences in cluster separation could be due to differences in the number of SNPs between the datasets (1,832 SNPs vs 1,126,226 SNPs). For example, loss of resolution in PCs 1 and 2 due to the number of SNPs has been previously found in a study involving a PCA using European samples with 391 markers [33]. Furthermore, we followed our analyses up with a PCA on set of 1,832 SNPs randomly selected from the HapMap genotypes (results not shown). We found that while the random set of HapMap 3 SNPs provided clearer separation than DMET Plus genotypes, the reduced number of SNPs made the difference between the two datasets less pronounced.

The results from the ADMIXTURE analysis provided definitive identification of the admixed, MEX population, than the PCA. The ADMIXTURE results show a distinct fourth cluster which is predominately represented in the MEX individuals. Moreover, this fourth cluster distinguishes the MEX population from the other three global populations. For DMET Plus and HapMap analyses, the self-reported MEX population predominately identifies with the Caucasian and the admixture clusters, which is expected from Mexican individuals living in North America. Spearman rank correlations were used to test for associations between DMET Plus and HapMap data for inferred genetic ancestry. All four clusters exhibited statistically significant positive correlations between the two technologies. The very high concordance correlation coefficients strongly indicates the ability of DMET Plus microarray data to estimate population substructure on par with genome-wide technologies (e.g., HapMap 3 data).

Even with its limited number of SNPs, the DMET Plus chip is comparable to the genotyping platforms from HapMap in detecting population substructure—in particular, the emergence of a fourth cluster, presumably Native American, from the ADMIXTURE analysis. This finding seems to contrast the results from the PCA, since the DMET Plus chip demonstrated a reduced ability to capture the MEX population as a separate cluster. Thus, the ADMIXTURE analysis may be better suited for identifying population substructure of the genotypes from the DMET Plus chip than PCA because ADMIXTURE assigns probabilities to the genotypes for each individual, while the PCA uses principal components to explain the variation between each population and not each individual. In addition, when standard PCA-based methods are used, closely related subpopulations can appear as a conglomerate and this problem can worsen when some subpopulations are genetically far removed from others in the data set [34]. Thus, the reduced ability of the PCA to capture the MEX population as a distinct cluster could be due to the MEX population's close relationship to the CEU population and its far relationship to the YRI and CHB populations.

On average, the mean inbreeding coefficients are very similar for the DMET Plus genotypes and the HapMap genotypes. For any reasonable choice of alpha in multiple testing correction, there is a significant correlation between the inbreeding coefficients of the DMET Plus genotypes and the HapMap genotypes, based on a nonparametric rank correlation. Notably, there are 16 individuals with inbreeding coefficients > 0.15 from the

DMET Plus chip, while there are zero individuals with inbreeding coefficients > 0.15 , which means that the DMET Plus chip seems to overestimate relatedness. Similar overestimates were seen for DMET Plus with alternative approaches to estimating inbreeding coefficients—e.g., plink and maximum likelihood estimation (results not shown). Hence, the DMET Plus chip could lead to conservative filtering, in which more individuals are flagged and removed during quality control due to the higher inbreeding coefficients. Perhaps the estimation of inbreeding is confounded by the fact that the DMET Plus microarray characterizes functional variants which typically exhibit less variation and, therefore, alter the estimation of inbreeding coefficients.

Estimation of genomic inbreeding coefficients requires accurate estimates of allele frequencies in the base population. When inbreeding coefficients were computed on a random sample of 1,832 SNPs from the HapMap data set, the results were similar to DMET Plus inbreeding coefficients. There were 39 individuals with inbreeding coefficients greater than or equal to 0.15. This suggests that the number of SNPs could be driving the implication of cryptic relatedness from inflated inbreeding coefficients. Given the relatively small sample size and the difference in total variants in DMET Plus compared with HapMap, it is not entirely surprising that DMET Plus was inferior in predicting cryptic relatedness via inbreeding coefficient estimation.

Conclusion

Recent technological advances have led to cost-effective and rapid genotyping that captures common variants in genes encoding for drug metabolism enzymes and drug transporters. For one such chip, the Affymetrix DMET Plus microarray, we have shown that these markers are capable of detecting population substructure across continental populations in a similar manner comparable to genome-wide genotyping technology. That is, the DMET Plus microarray readily allows for accurate population stratification correction across CEU, CHB, and YRI continental populations. For PCA, the DMET Plus exhibited slightly diminished capacity for separating an admixed cluster of MEX from CEU. However, for the ADMIXTURE analysis, the DMET Plus genotypic data were able to clearly identify an admixed genetic ancestry distinct from the CEU group. Moving forward, it is best to use structure-like analysis for a stratified analysis on the DMET Plus technology. Although these data might be ill-suited for a population genetics analyses, as illustrated by overestimation relatedness from DMET Plus data, these microarrays offer sufficient genetic resolution for covariate adjustment of population structure in genomic association studies. These results are important for subsequent analyses that involve detecting and correcting for population substructure among continental and admixed populations in studies using DMET Plus technology.

Acknowledgments

The authors would like to thank the National Institute for General Medicine for their funding of this project. JNJ would also like to thank the National Science Foundation for assistance through a graduate research fellowship and the NCSU Biostatistics Training in the Omics Era training grant.

JNJ has received support from a graduate research fellowship from the National Science Foundation and from the NCSU Biostatistics Training in the Omics Era training grant (5T32GM081057). This work was supported by a grant (T32 GM081057) from the National Institute of General Medicine.

References

1. Pritchard JK, Rosenberg NA. Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. *Am J Hum Genet.* 1999; 65(1):220–228. [PubMed: 10364535]
2. Devlin B, Roeder K, Wasserman L. Genomic Control, a New Approach to Genetic-Based Association Studies. *Theor Popul Biol.* 2001; 60(3):155–166. [PubMed: 11855950]
3. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.* 2000; 155(2):945–959. [PubMed: 10835412]
4. Enoch MA. Using ancestry-informative markers to define populations and detect population stratification. *J Psychopharmacol.* 2006; 20(4 suppl):19–26.
5. The International HapMap Consortium. The International HapMap Project. *Nature.* 2003; 426(6968):789–796. [PubMed: 14685227]
6. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010; 467(7311):52–58. [PubMed: 20811451]
7. Kühnisch J, Thiering E, Heitmüller D, Tiesler CT, Grallert H, Heinrich-Weltzien R, et al. Genome-wide association study (GWAS) for molar–incisor hypomineralization (MIH). *Clin Oral Investig.* 2014; 18(2):677–682.
8. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature.* 2010; 464(7289):713–720. [PubMed: 20360734]
9. Ma D, Salyakina D, Jaworski JM, Konidari I, Whitehead PL, Andersen AN, et al. A Genome-wide Association Study of Autism Reveals a Common Novel Risk Locus at 5p14.1. *Ann Hum Genet.* 2009; 73(3):263–273. [PubMed: 19456320]
10. Rothman N, Garcia-Closa M, Chatterjee N, Malats N, Wu X, Figueroa JD, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet.* 2010; 42(11):978–984. [PubMed: 20972438]
11. Burmester, J, Sedova, M, Shapero, M, Mansfield, E. DMET™ Microarray Technology for Pharmacogenomics-Based Personalized Medicine. In: Chittur, S, editor. *Microarray Methods for Drug Discovery, Methods Mol Biol.* Vol. 632. Humana Press; 2010. 99–124.
12. Hoffman JM, Haidar CE, Wilkinson MR, Crews KR, Baker DK, Kornegay NM, et al. PG4KDS: A model for the clinical implementation of pre-emptive pharmacogenetics. *Am J Med Genet C Semin Med Genet.* 2014; 166(1):45–55.
13. Deeken JF, Cormier T, Price DK, Sissung TM, Steinberg SM, Tran K, et al. A pharmacogenetic study of docetaxel and thalidomide in patients with castration-resistant prostate cancer using the DMET genotyping platform. *Pharmacogenomics J.* 2010; 10(3):191–199. [PubMed: 20038957]
14. de Graan AJ, Elens L, Smid M, Martens JW, Sparreboom A, Nieuweboer AJ, et al. A Pharmacogenetic Predictive Model for Paclitaxel Clearance Based on the DMET Platform. *Clin Cancer Res.* 2013; 19(18):5210–5217. [PubMed: 23918604]
15. Suarez-Kurtz G. Pharmacogenomics in admixed populations. *Trends Pharmacol Sci.* 2005; 26(4): 196–201. [PubMed: 15808344]
16. Man M, Farmen M, Dumauval C, Teng CH, Moser B, Irie S, et al. Genetic Variation in Metabolizing Enzyme and Transporter Genes: Comprehensive Assessment in 3 Major East Asian Subpopulations With Comparison to Caucasians and Africans. *J Clin Pharmacol.* 2010; 50(8):929–940. [PubMed: 20173083]
17. He YJ, Misher AD, Irvin W, Motsinger-Reif A, McLeod HL, Hoskins JM. Assessing the utility of whole genome amplified DNA as a template for DMET Plus array. *Clinical Chemistry and Laboratory Medicine: CCLM/FESCC.* 2012; 50(8):1329–1334.

18. Hertz DL, Roy S, Jack J, Motsinger-Reif AA, Drobish A, Clark LS, et al. Genetic heterogeneity beyond CYP2C8*3 does not explain differential sensitivity to paclitaxel-induced neuropathy. *Breast Cancer Research and Treatment*. 2014; 145(1):245–254. [PubMed: 24706167]
19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8): 904–909. [PubMed: 16862161]
20. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic Maps of Human Gene Frequencies in Europeans. *Science*. 1978; 201(4358):786–792. [PubMed: 356262]
21. Service SK, Teslovich TM, Fuchsberger C, Ramensky V, Yajnik P, Koboldt DC, et al. Re-sequencing Expands Our Understanding of the Phenotypic Impact of Variants at GWAS Loci. *PLoS Genet*. 2014; 10(1)
22. Ma L, Wiggans GR, Wang S, Sonstegard TS, Yang J, Crooker BA, et al. Effect of sample stratification on dairy GWAS results. *BMC Genomics*. 2012; 13(1):536. [PubMed: 23039970]
23. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. *Am J Hum Genet*. 2010; 86(6):860–871. [PubMed: 20560206]
24. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, et al. Discerning the Ancestry of European Americans in Genetic Association Studies. *PLoS Genet*. 2008; 4(1):e236. [PubMed: 18208327]
25. Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Gene*. 2009; 10(1):39.
26. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009; 19(9):1655–1664. [PubMed: 19648217]
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007; 81:559–575. [PubMed: 17701901]
28. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. URL <http://www.R-project.org/>
29. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics*. 2012; 28(24):3326–3328. [PubMed: 23060615]
30. Voight BF, Pritchard JK. Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genet*. 2005; 1(3):e32. [PubMed: 16151517]
31. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010; 42(7):565–569. [PubMed: 20562875]
32. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, et al. Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genet*. 2008; 4(1):e4. [PubMed: 18208329]
33. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Med Genet*. 2008; 16(12):1413–1429.
34. Intarapanich A, Shaw PJ, Assawamakin A, Wangkumhang P, Ngamphiw C, Chaichoompu K, et al. Iterative pruning PCA improves resolution of highly structured populations. *BMC Bioinformatics*. 2009; 10(1):382. [PubMed: 19930644]

Figure 1. Principal Component Analysis

The PCA results for (A) DMET Plus and (B) HapMap are illustrated. For each dataset, three pairwise plots illustrate the top four PCs. Each point represents a sample and self-declared ethnicities colored as follows: blue for CHB, red for YRI, yellow for MEX, and green for CEU.

Figure 2. ADMIXTURE Analysis

The results of an ADMIXTURE analysis of the genotypic data from (A) DMET Plus and (B) HapMap data are shown. The stacked bar plots represent the genetically inferred ethnicities for each individual represented as percentages for Clusters 1–4.

Figure 3. Inbreeding Coefficients

The two plots representing the estimated inbreeding coefficients for (A) DMET Plus and (B) HapMap are shown. The black dashed line in each plot represents the mean of the inbreeding coefficients across all samples. Self-declared ethnicities are displayed by the following colors: green for CEU, blue for CHB, yellow for MEX, and red for YRI.