Cell Genomics, Volume *4*

Supplemental information

Genetics of Latin American Diversity Project: Insights into population genetics and association studies in admixed groups in the Americas

Victor Borda, Douglas P. Loesch, Bing Guo, Roland Laboulaye, Diego Veliz-Otani, Jennifer N. French, Thiago Peixoto Leal, Stephanie M. Gogarten, Sunday Ikpe, Mateus H. Gouveia, Marla Mendes, Gonçalo R. Abecasis, Isabela Alvim, Carlos E. Arboleda-Bustos, Gonzalo Arboleda, Humberto Arboleda, Mauricio L. Barreto, Lucas Barwick, Marcos A. Bezzera, John Blangero, Vanderci Borges, Omar Caceres, Jianwen Cai, Pedro Chana-Cuevas, Zhanghua Chen, Brian Custer, Michael Dean, Carla Dinardo, Igor Domingos, Ravindranath Duggirala, Elena Dieguez, Willian Fernandez, Henrique B. Ferraz, Frank Gilliland, Heinner Guio, Bernardo Horta, Joanne E. Curran, Jill M. Johnsen, Robert C. Kaplan, Shannon Kelly, Eimear E. Kenny, Barbara A. Konkle, Charles Kooperberg, Andres Lescano, M. Fernanda Lima-Costa, Ruth J.F. Loos, Ani Manichaikul, Deborah A. Meyers, Michel S. Naslavsky, Deborah A. Nickerson, Kari E. North, Carlos Padilla, Michael Preuss, Victor Raggio, Alexander P. Reiner, Stephen S. Rich, Carlos R. Rieder, Michiel Rienstra, Jerome I. Rotter, Tatjana Rundek, Ralph L. Sacco, Cesar Sanchez, Vijay G. Sankaran, Bruno Lopes Santos-Lobato, Artur Francisco Schumacher-Schuh, Marilia O. Scliar, Edwin K. Silverman, Tamar Sofer, Jessica Lasky-Su, Vitor Tumas, Scott T. Weiss, Latin American Research Consortium on the Genetics of Parkinson's Disease (LARGE-PD), National Institute of Neurological Disorders and Stroke (NINDS) Stroke Genetics Network (SiGN) Consortium, Trans-Omics for Precision Medicine (TOPMed) Population Genetics Working Group, Ignacio F. Mata, Ryan D. Hernandez, Eduardo Tarazona-Santos, and Timothy D. O'Connor

Data S1 - Study acknowledgments

NHLBI TOPMed - The Genetic Epidemiology of Asthma in Costa Rica: This study was supported by NHLBI grants R37 HL066289 and P01 HL132825. We wish to acknowledge the investigators at the Channing Division of Network Medicine at Brigham and Women's Hospital, the investigators at the Hospital Nacional de Niños in San José, Costa Rica 39 and the study subjects and their extended family members who contributed samples and genotypes to the study, and the NIH/NHLBI for its support in making this project possible.

NHLBI TOPMed - San Antonio Family Heart Study (SAFHS): Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. Sample processing for sequencing was performed in facilities constructed with the support of NIH grant C06 RR020547. ORCID ID: 0000-0001-6250-5723.

NHLBI TOPMed - Women's Health Initiative (WHI): 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 HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

NHLBI TOPMed - Hispanic Community Health Study - Study of Latinos (HCHS/SOL): The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the 47 NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

NHLBI TOPMed - Multi-Ethnic Study of Atherosclerosis: MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, and R01HL105756. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org.

NHLBI TOPMed - Severe Asthma Research Program (SARP): The authors thank the SARP participants, investigators, clinical research staff and data coordinating center. SARP was conducted with the support of the National Institutes of Health (NIH), National Heart, Lung, and Blood Institute (NHLBI) grants R01 HL069116, R01 HL069130, R01 HL069149, R01 HL069155, R01 HL069167, R01 HL069170, R01 HL069174, R01 HL069349, U10 HL109086, U10 HL109146, U10 HL109152, U10 HL109164, U10 HL109168, U10 HL109172, U10 HL109250, and U10 HL109257.

NHLBI TOPMed - Recipient Epidemiology and Donor Evaluation Study-III Brazil Sickle Cell Disease Cohort (REDS-BSCDC): The Recipient Epidemiology and Donor Evaluation Study (REDS-III) International Component - Brazil Sickle Cell Disease Cohort Study was funded by National Heart, Lung, and Blood Institute of NIH (Contract No. HHSN268201100007I) and is a collaboration of Blood Systems Research Institute and Research Triangle Institute International in the USA and the University of Sao Paulo, Institute of Tropic Medicine, Fundacao Pro-Sangue, Institute for the Treatment of Childhood Cancer (ITACI), Hemominas, Hemorio, and Hemope in Brazil.

NHLBI TOPMed - My Life Our Future (MLOF) Research Repository of Patients with Hemophilia A (Factor VIII Deficiency) or Hemophilia B (Factor IX Deficiency): My Life Our Future (MLOF) is a project governed by a Steering

committee representing Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation and Biogen. The project is funded by Biogen. Patients with hemophilia were enrolled at 80 Hemophilia Treatment Centers from across the U.S.

NHLBI TOPMed - Boston-Brazil Sickle Cell Disease (SCD) Cohort: The sickle cell patients included in this cohort were collected by Drs. Igor Domingos, Marcos Andre Bezerra, and Aderson Araujo, and colleagues from their institution, the Hematology and Hemotherapy Foundation of Pernambuco. This study has received partial support from NIH grants R01DK103794 and U01HL117720 (to Vijay G. Sankaran). We are grateful to all study participants for their willingness to be involved in this study.

NHLBI TOPMed: Children's Health Study (CHS) Integrative Genomics and Environmental Research of Asthma (IGERA): The Integrative Genomics and Environmental Research of Asthma (IGERA) Study was supported by the National Heart, Lung and Blood Institute (grant # RC2HL101543 -The Asthma BioRepository for Integrative Genomics Research, PI Gilliland/Raby). The Children's Health Study (CHS) was supported by the Southern California Environmental Health Sciences Center (grant P30ES007048); National Institute of Environmental Health Sciences (grants 5P01ES011627, ES021801, ES023262, P01ES009581, P01ES011627, P01ES022845, R01 ES016535, R03ES014046, P50 CA180905, R01HL061768, R01HL076647, R01HL087680 and RC2HL101651), the Environmental Protection Agency (grants RD83544101, R826708, RD831861, and R831845), and the Hastings Foundation.

NHLBI TOPMed: Children's Health Study (CHS) Effects of Air Pollution on the Development of Obesity in Children (Meta-AIR): The Effects of Air Pollution on the Development of Obesity in Children (Meta-AIR) study was supported by the Southern California Children's Environmental Health Center funded by the National Institute of Environmental Health Sciences (NIEHS) (P01ES022845) and the Environmental Protection Agency (EPA) (RD-83544101–0). The Children's Health Study (CHS) was supported by the Southern California Environmental Health Sciences Center (grant P30ES007048); National Institute of Environmental Health Sciences (grants 5P01ES011627, ES021801, ES023262, P01ES009581, P01ES011627, P01ES022845, R01 ES016535, R03ES014046, P50 CA180905, R01HL061768, R01HL076647, R01HL087680, RC2HL101651 and K99ES027870), the Environmental Protection Agency (grants RD83544101, R826708, RD831861, and R831845), and the Hastings Foundation.

NHLBI TOPMed - NHGRI CCDG: The BioMe Biobank at Mount Sinai: The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

NHLBI TOPMed - Lung Tissue Research Consortium (LTRC): This study utilized biological specimens and data provided by the Lung Tissue Research Consortium (LTRC) supported by the National Heart, Lung, and Blood Institute (NHLBI).

NHLBI TOPMed - Childhood Asthma Management Program (CAMP): We thank all subjects for their ongoing participation in this study. We acknowledge the CAMP investigators and research team, supported by NHLBI, for collection of CAMP Genetic Ancillary Study data. All work on data collected from the CAMP Genetic Ancillary Study was conducted at the Channing Laboratory of the Brigham and Women's Hospital under appropriate CAMP policies and human subject's protections. The CAMP Genetics Ancillary Study is supported by U01 HL075419, U01 HL65899, P01 HL083069, R01 HL 086601, R37 HL066289 and T32 HL07427 from the National Heart, Lung and Blood Institute, National Institutes of Health.

SUPPLEMENTARY FIGURES

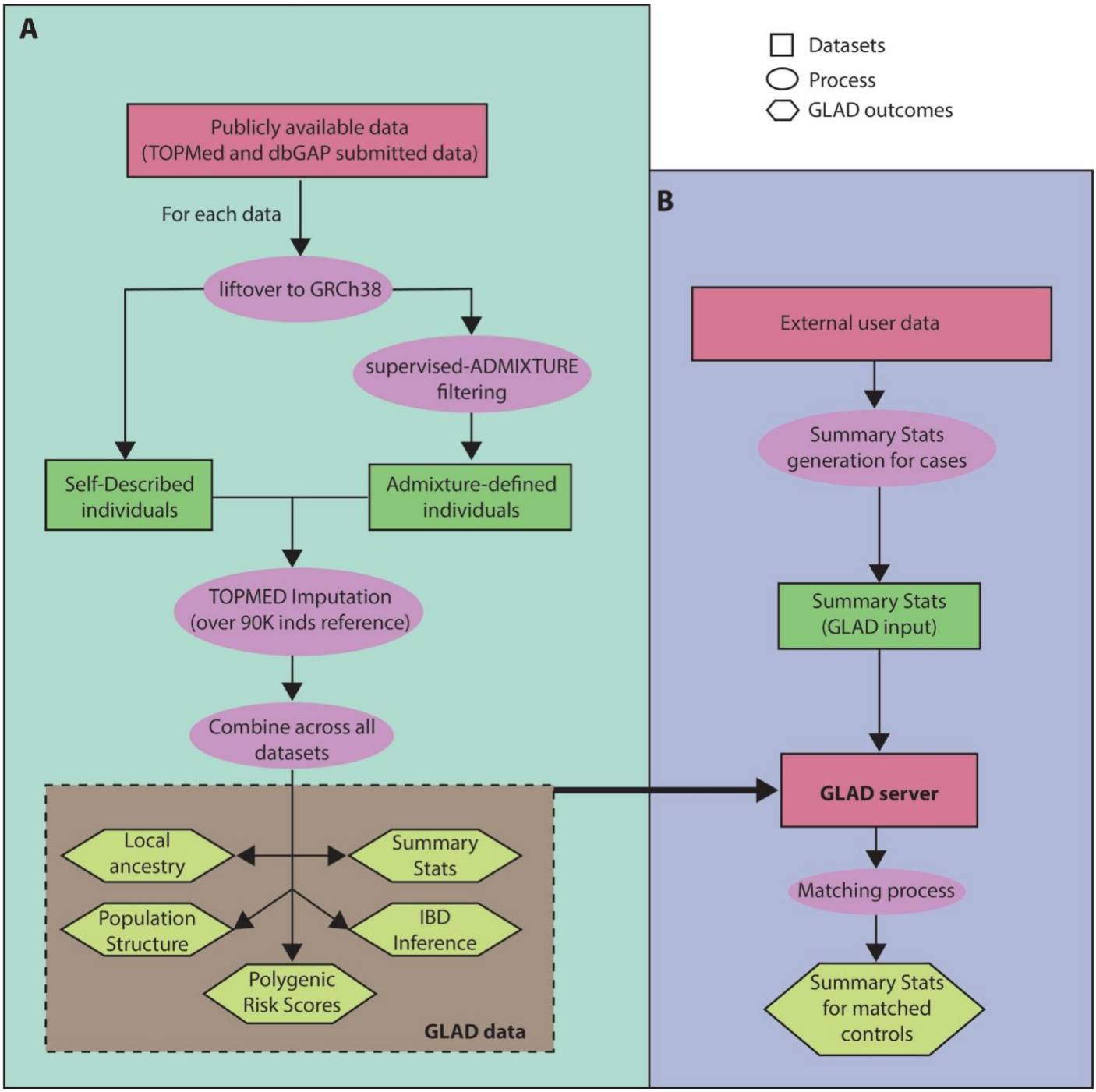

Figure S1. Workflow for the building and the use of the GLAD database, related to STAR Methods. A) For each dbGaP cohort, we extracted and self-described Latino and ADMIXTURE-defined subjects with at least 2% of Indigenous American ancestry. Then, each cohort was imputed in the Michigan Imputation Center using the TOPMED Imputation panel. After imputation, we selected the best-imputed loci (r^2>0.9) and merged the data. We characterized the GLADdb using PCA, IBD, and local ancestry analyses. B) By identifying the GLAD individuals that have genetic patterns similar to those of a query sample, we provide summary statistics of the control subjects from GLADdb.

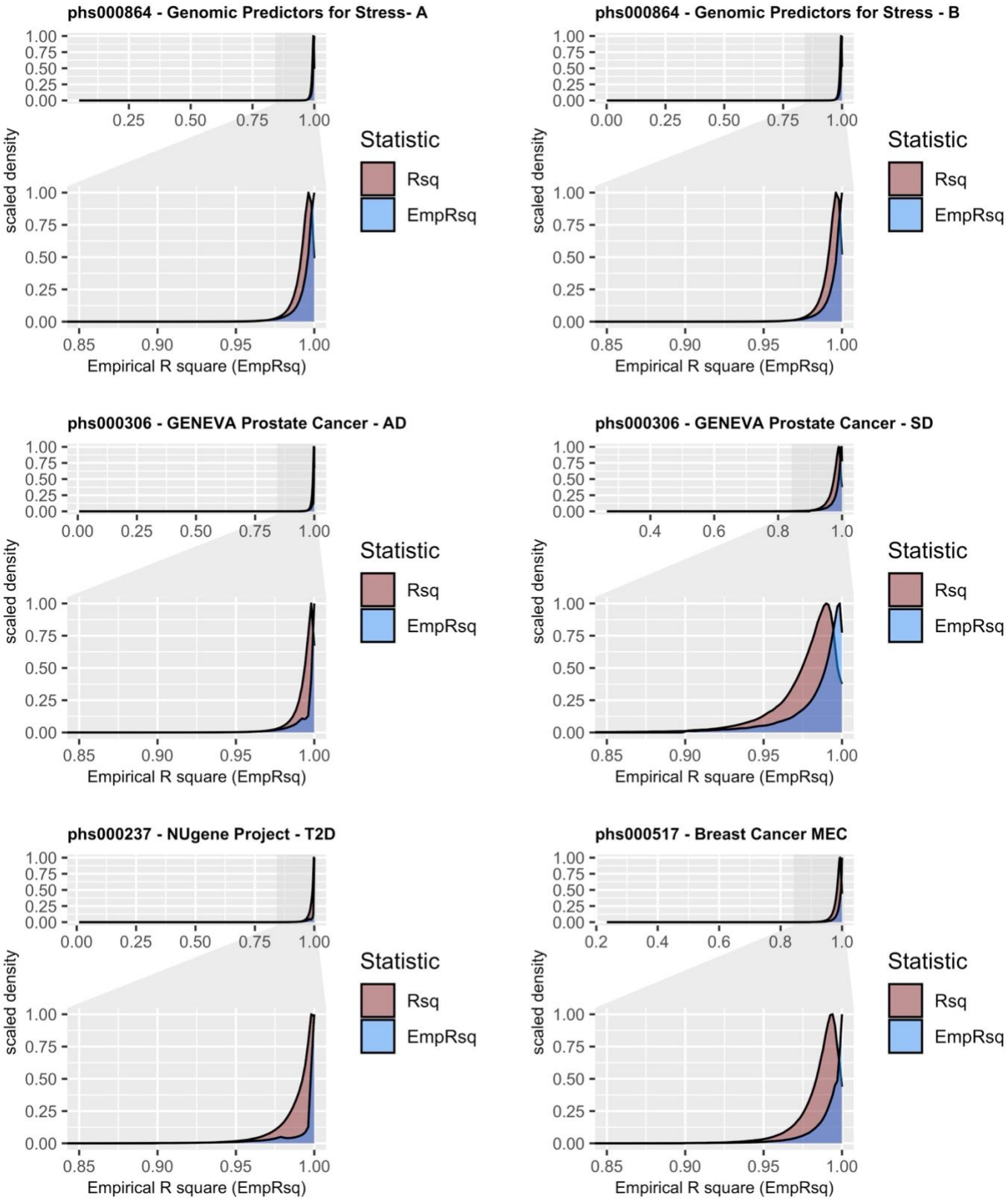

Figure S2. Distribution of Imputation Statistics for variants included in six GLADdb cohorts, related to STAR Methods. To assess imputation quality within each dbGaP cohort, we generated distribution plots for Rsq (blue area) and Empirical Rsq values (red area). This analysis specifically focused on genotyped variants that were incorporated into GLADdb. Empirical Rsq (EmpRsq) reflects the correlation between the true genotyped values and the imputed dosages obtained by hiding all known genotypes for the specific SNP. Higher values of this correlation reflect the higher imputation accuracy.

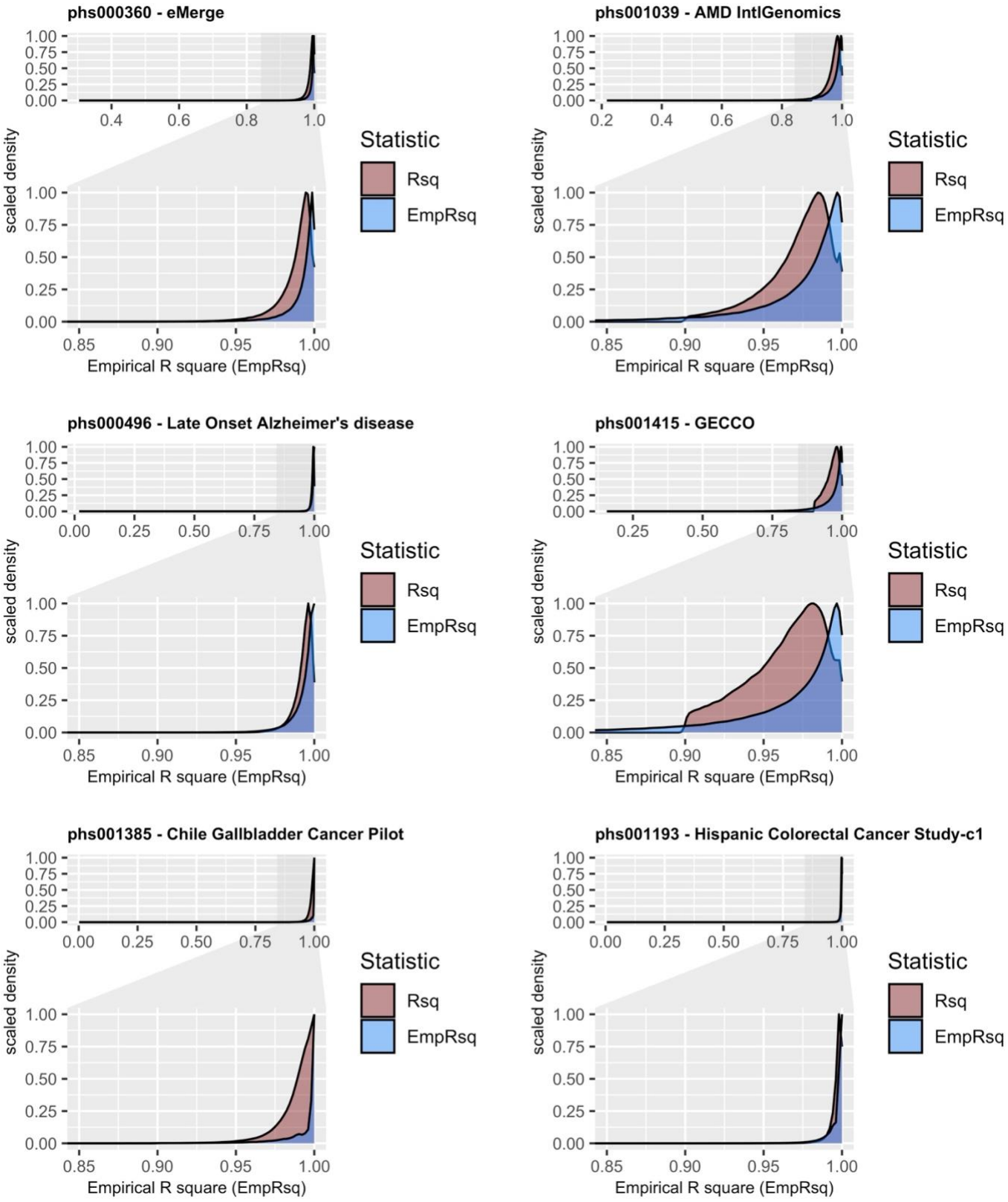

Figure S3. Distribution of Imputation Statistics for variants included in six GLADdb cohorts, related to STAR Methods. To assess imputation quality within each dbGaP cohort, we generated distribution plots for Rsq (blue area) and Empirical Rsq values (red area). This analysis specifically focused on genotyped variants that were incorporated into GLADdb. Empirical Rsq (EmpRsq) reflects the correlation between the true genotyped values and the imputed dosages obtained by hiding all known genotypes for the specific SNP. Higher values of this correlation reflect the higher imputation accuracy.

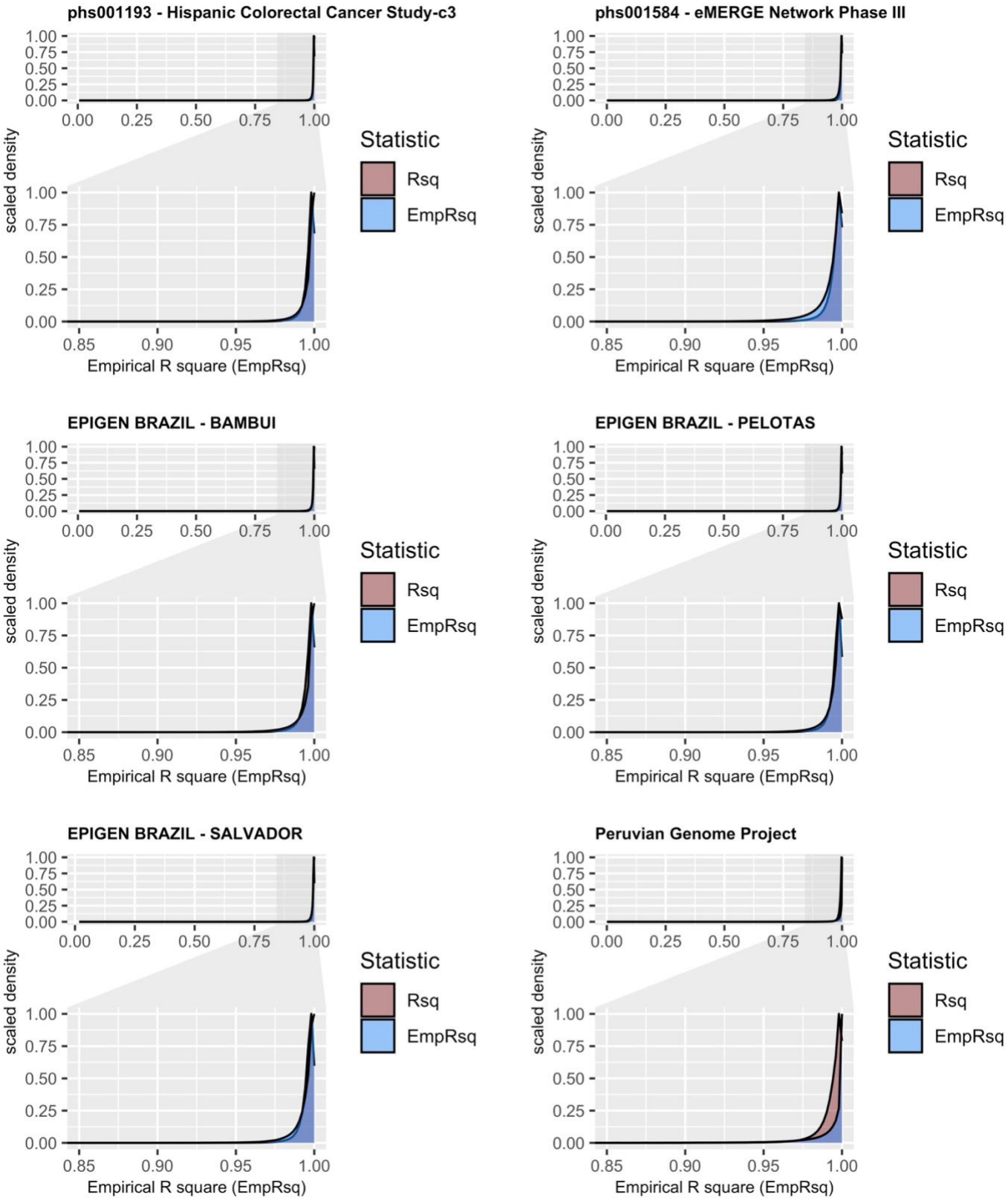

Figure S4. Distribution of Imputation Statistics for variants included in six GLADdb cohorts, related to STAR Methods. To assess imputation quality within each dbGaP cohort, we generated distribution plots for Rsq (blue area) and Empirical Rsq values (red area). This analysis specifically focused on genotyped variants that were incorporated into GLADdb. Empirical Rsq (EmpRsq) reflects the correlation between the true genotyped values and the imputed dosages obtained by hiding all known genotypes for the specific SNP. Higher values of this correlation reflect the higher imputation accuracy.

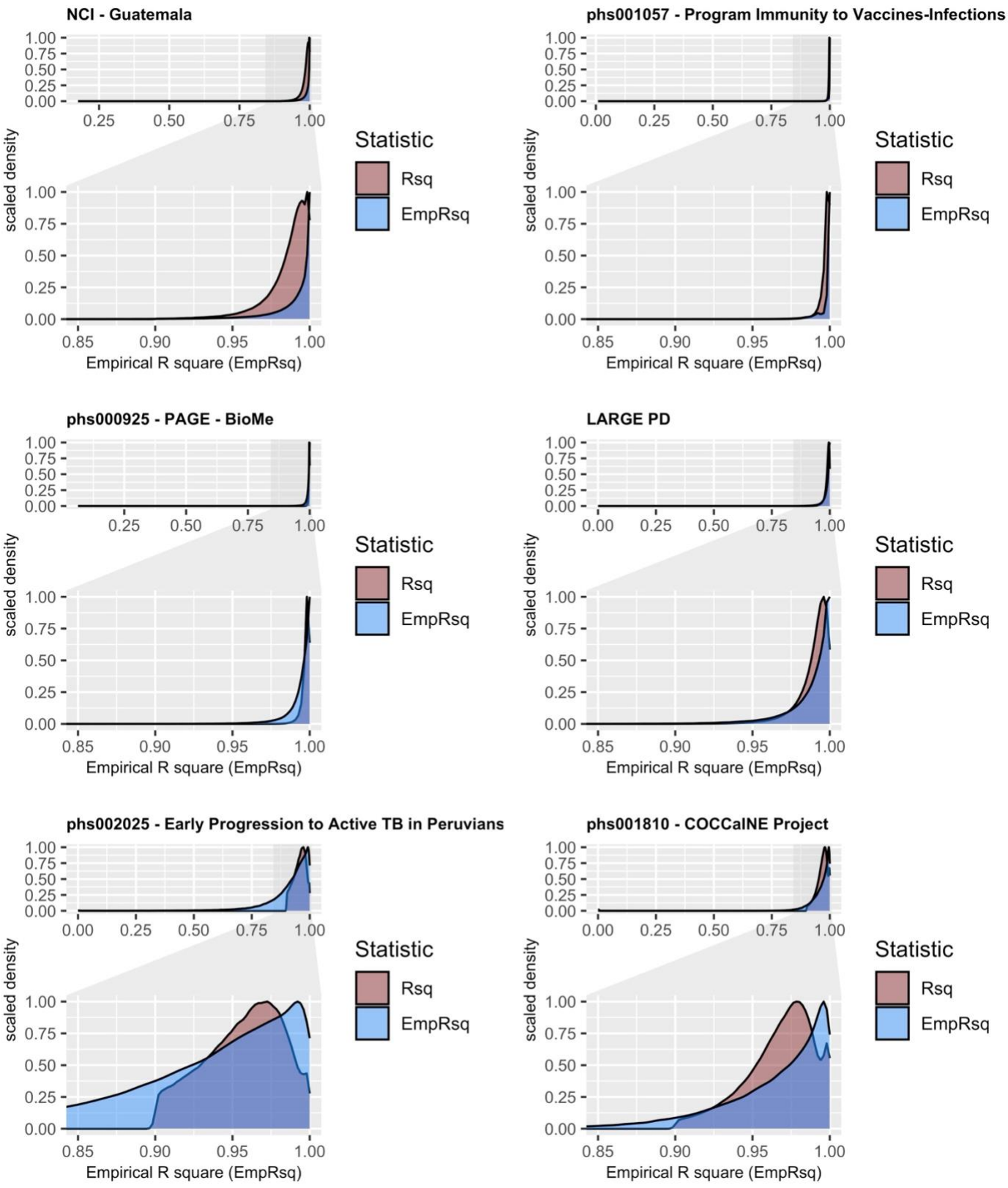

Figure S5. Distribution of Imputation Statistics for variants included in six GLADdb cohorts, related to STAR Methods. To assess imputation quality within each dbGaP cohort, we generated distribution plots for Rsq (blue area) and Empirical Rsq values (red area). This analysis specifically focused on genotyped variants that were incorporated into GLADdb. Empirical Rsq (EmpRsq) reflects the correlation between the true genotyped values and the imputed dosages obtained by hiding all known genotypes for the specific SNP. Higher values of this correlation reflect the higher imputation accuracy.

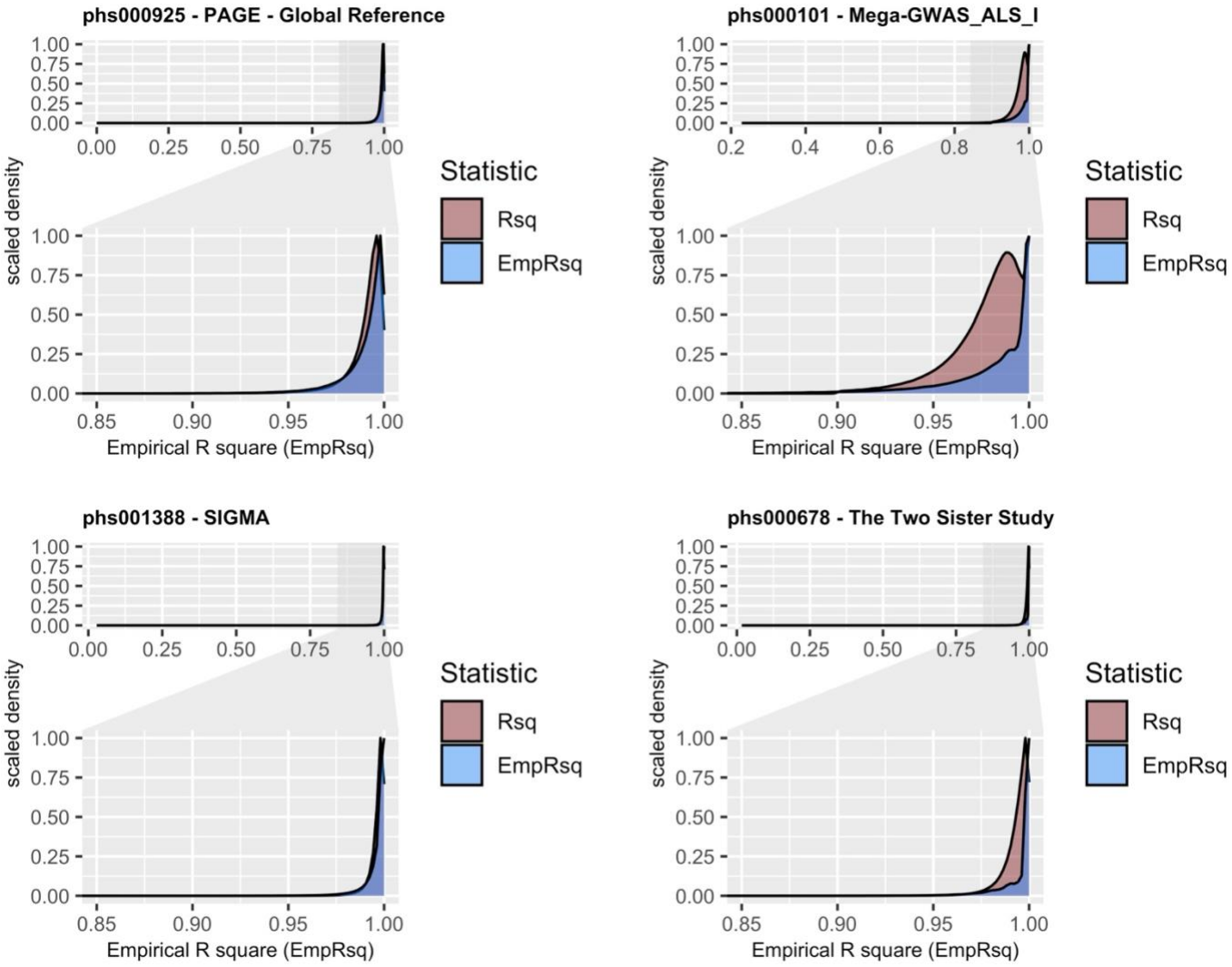

Figure S6. Distribution of Imputation Statistics for variants included in four GLADdb cohorts, related to STAR Methods. To assess imputation quality within each dbGaP cohort, we generated distribution plots for Rsq (blue area) and Empirical Rsq values (red area). This analysis specifically focused on genotyped variants that were incorporated into GLADdb. Empirical Rsq (EmpRsq) reflects the correlation between the true genotyped values and the imputed dosages obtained by hiding all known genotypes for the specific SNP. Higher values of this correlation reflect the higher imputation accuracy.

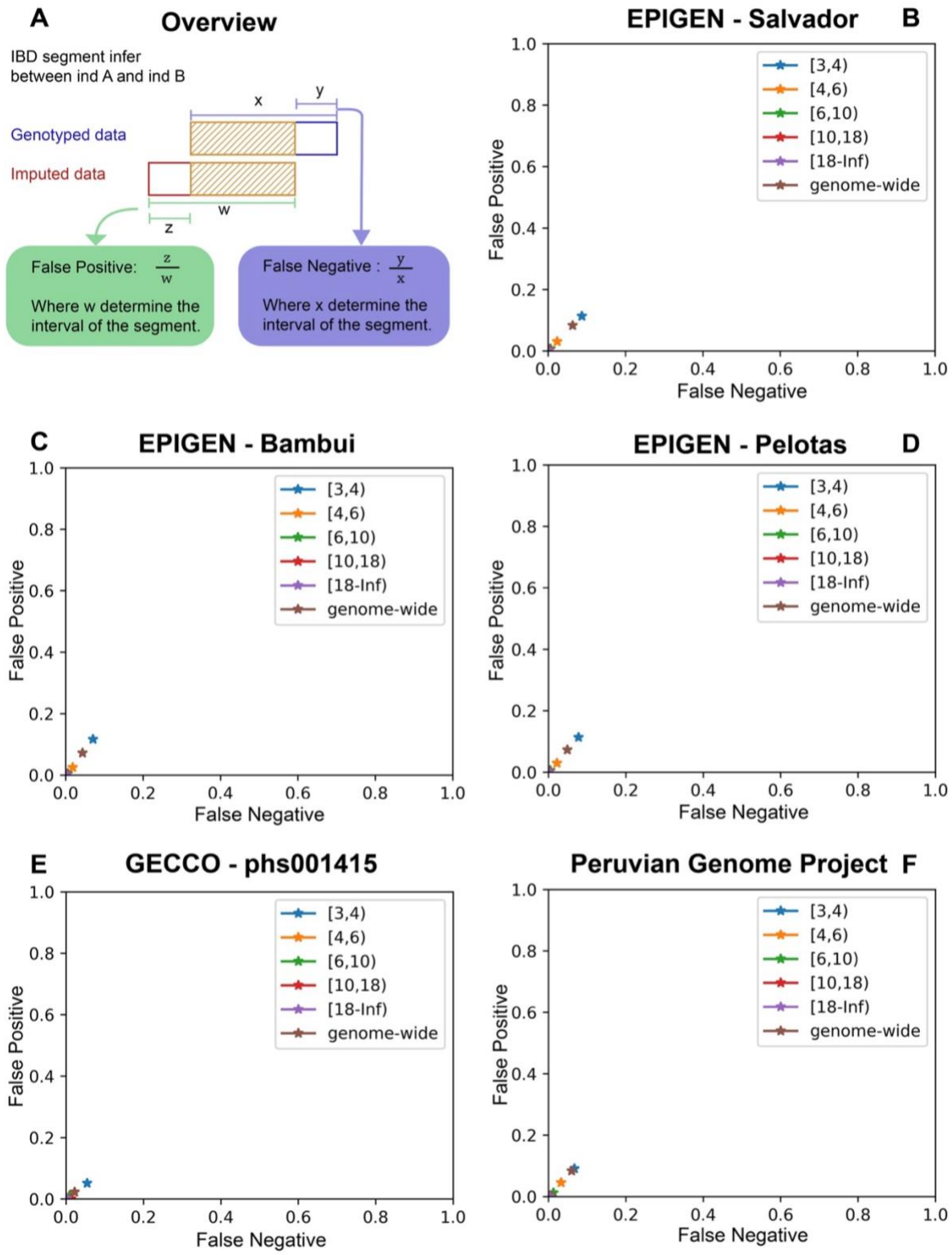

Figure S7.Comparison of IBD Inferred Genotyped and Imputed data through overlapping of individual IBD segments, related to STAR Methods. A) To identify the bias generated from imputed data in IBD inference, we determined the level of overlapping between IBD segments for a pair of haplotypes from different individuals (haplotype-level IBD) inferred using genotyped and imputed data (**See STAR Methods**). Based on the overlapping, we calculated levels of false positive and false negative for each IBD segment and then averaged within a specific length interval or at genome-wide levels. Further, to see the performance of IBD inference in different ancestry compositions, we selected three cohorts with different ancestry backgrounds: EPIGEN **(B**, **C**, and **D**, Predominantly admixed of EUR and AFR ancestries from Brazil), GECCO **(E**, Predominantly EUR and AFR from the US and Puerto Rico), and Peruvian Genome Project (**F**, Predominantly Indigenous from South America).

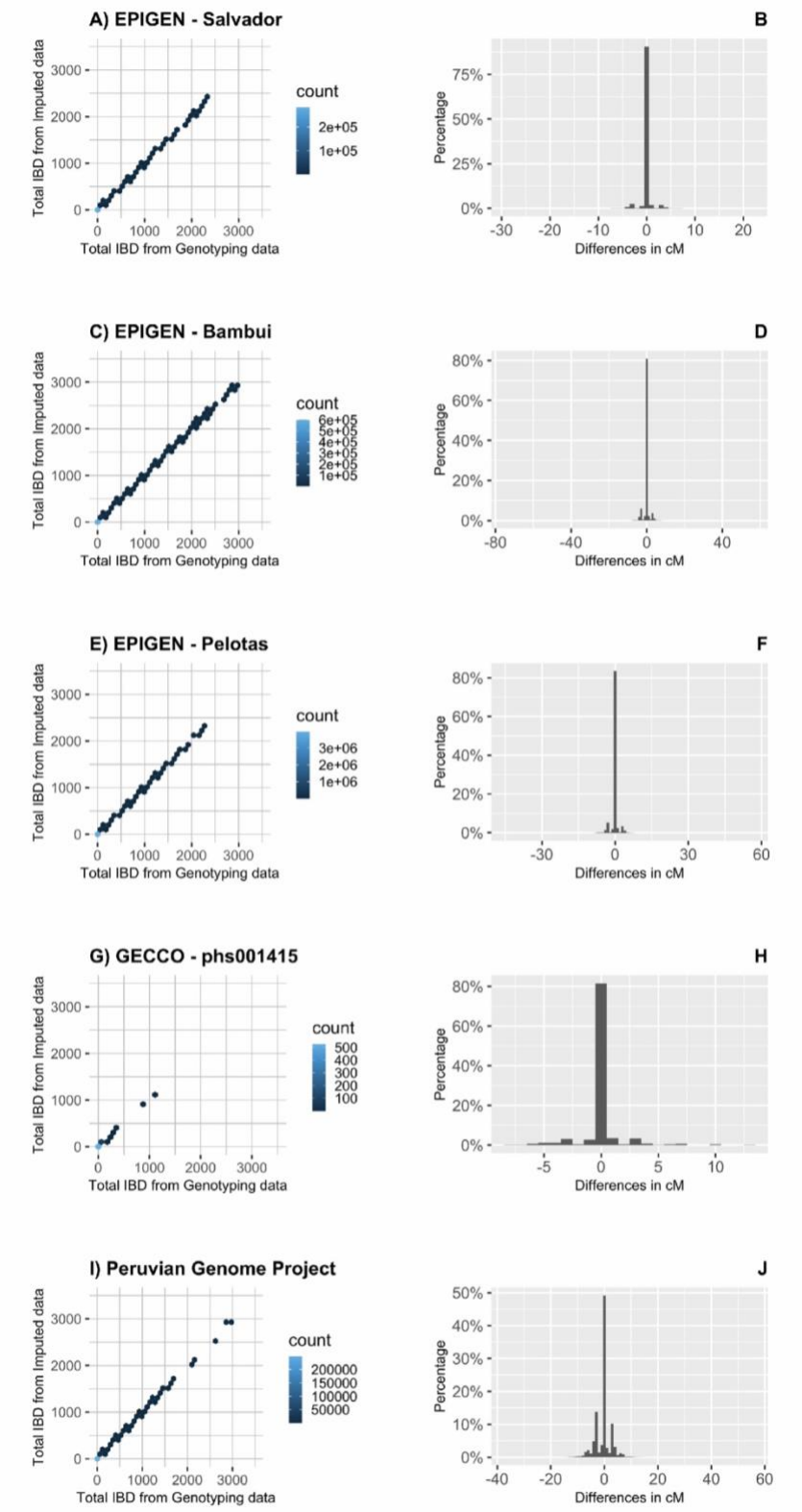

Figure S8. Comparison of IBD inferred Genotyped and Imputed data through individual-pair total IBD sharing, related to STAR Methods. Three cohorts with different ancestry backgrounds (EPIGEN [Predominantly admixed of EUR and AFR ancestries from Brazil], GECCO [Predominantly EUR and AFR from the US and Puerto Rico], and Peruvian Genome Project [Predominantly Indigenous from South America]) were selected for comparison. Total IBD sharing was inferred from genotyped and imputed data only and then compared to determine the level of bias observed in imputed data. In the scatter plots (A, C, E, G, and I), each dot represents the total amount of IBD obtained from genotyped vs imputed data for the same individual pair. Histograms (B, D, F, H, and J) represent how frequent the absolute differences (cM) between the Total IBD amount from genotyped and imputed data are.

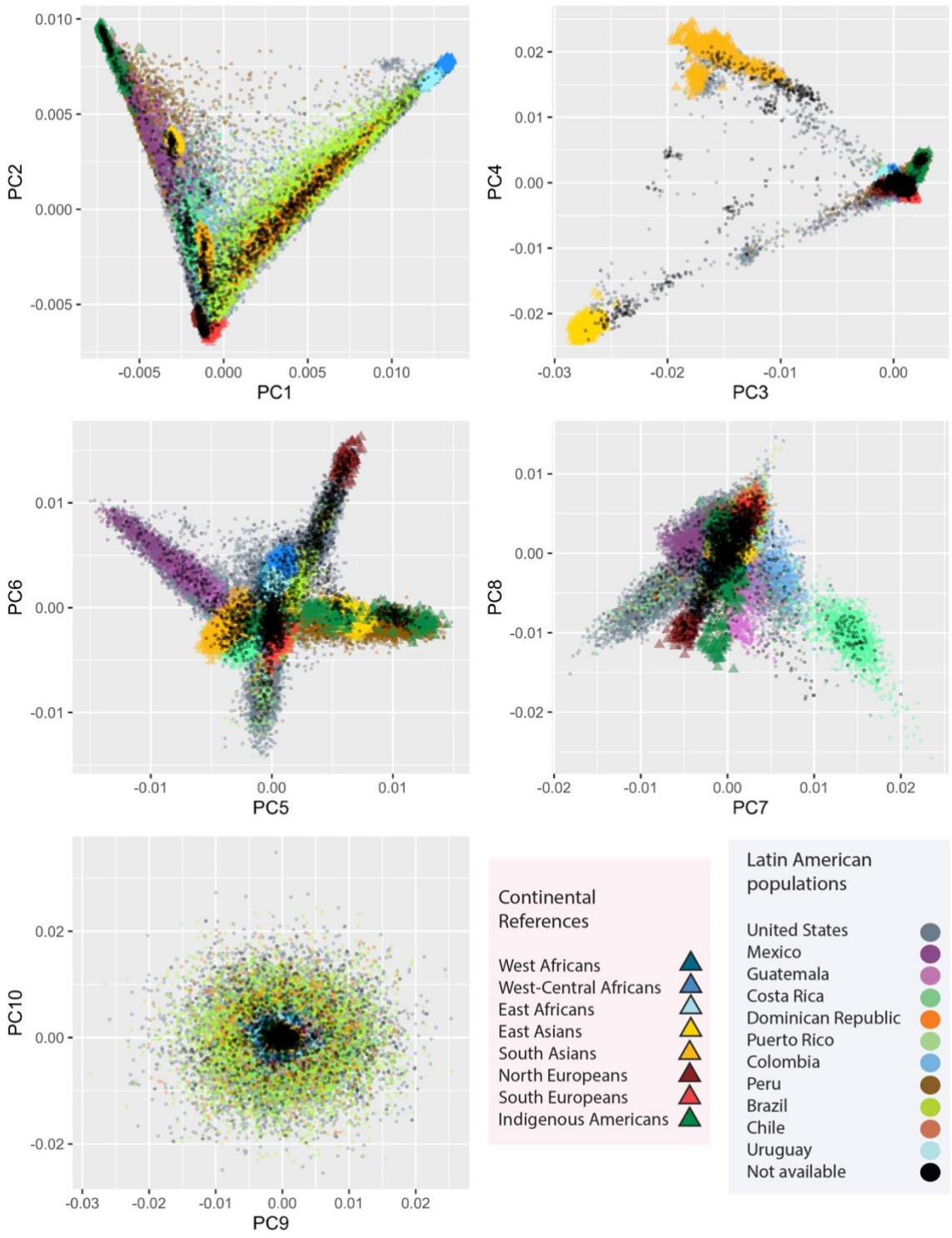

Figure S9. Principal component (PC) analysis GLADdb and ancestral reference groups individuals, related to Figure 1. The first ten PCs include reference groups (triangles) and GLAD individuals (circles).

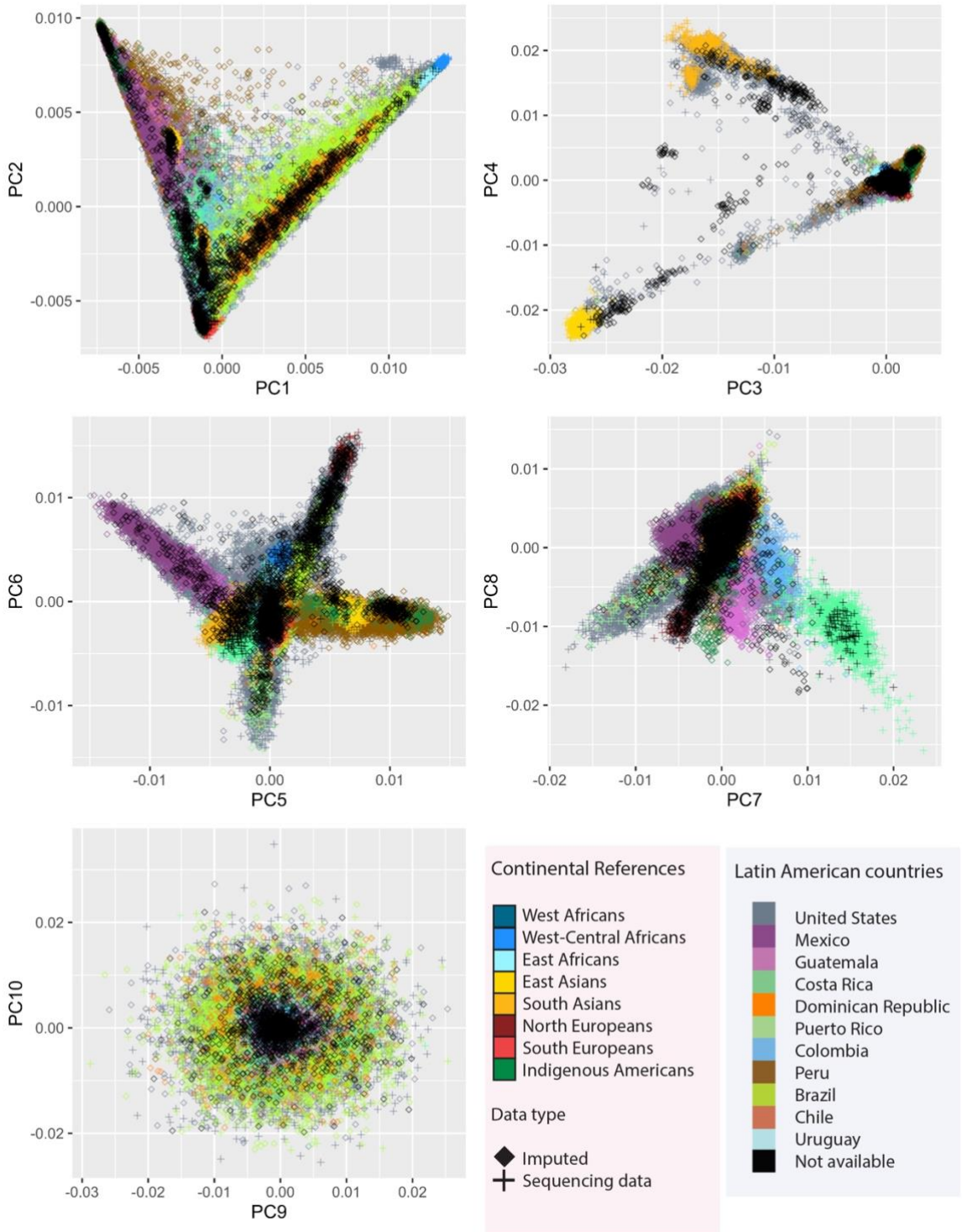

Figure S10. Principal component (PC) analysis GLADdb and ancestral reference groups individuals, related to Figure 1. The plot shows the relationships between GLADdb individuals with different data types: Imputed (diamond) and sequencing data (cross).

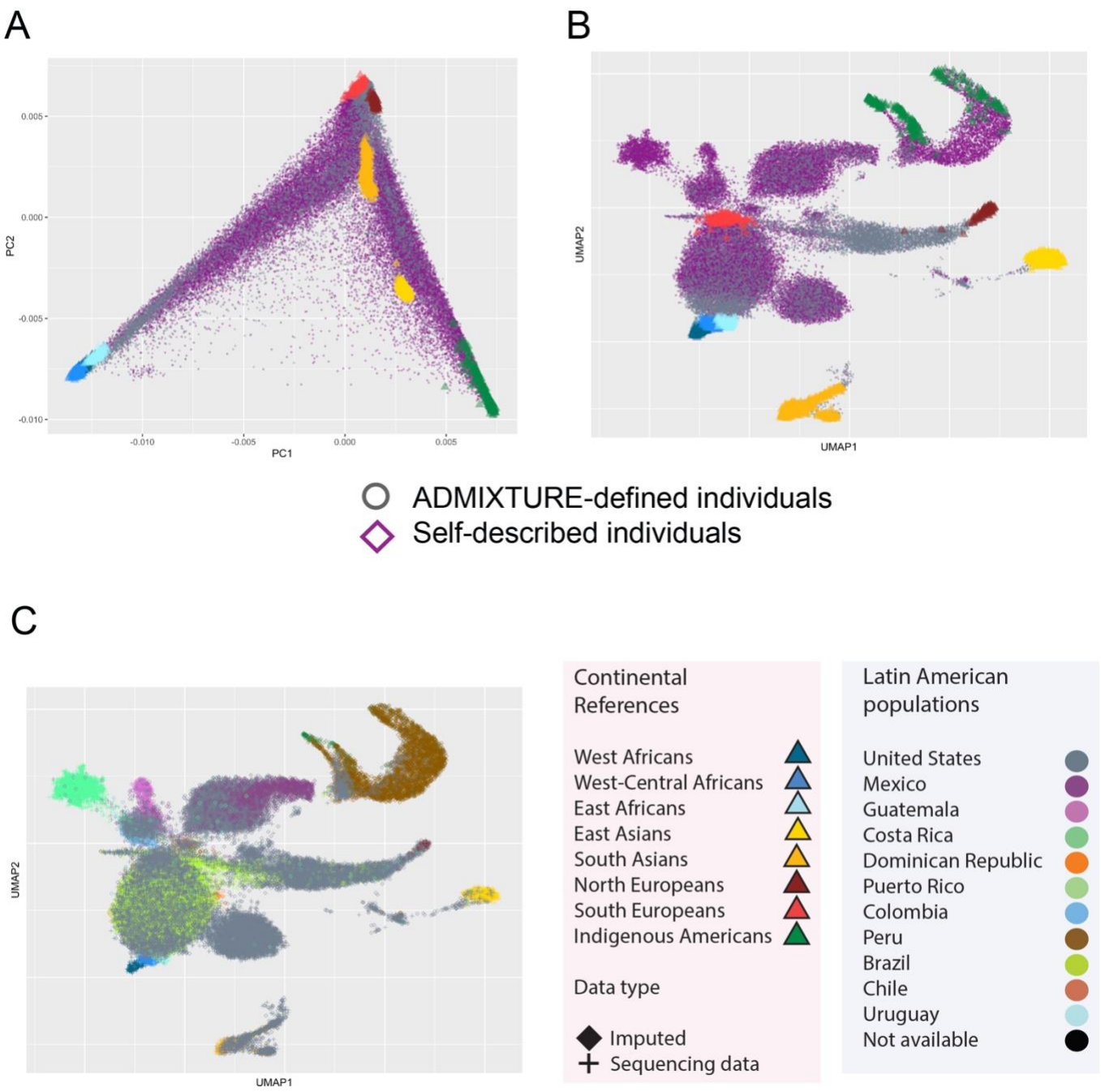

Figure S11. Population structure analysis GLADdb and ancestral reference groups individuals, related to Figure 1. Principal component (A) and UMAP (B) analyses show the relationship between self-described and ADMIXTURE-defined individuals in GLADdb. C) UMAP analysis of GLADdb individuals with different data types: Imputed (diamond) and sequencing data (cross).

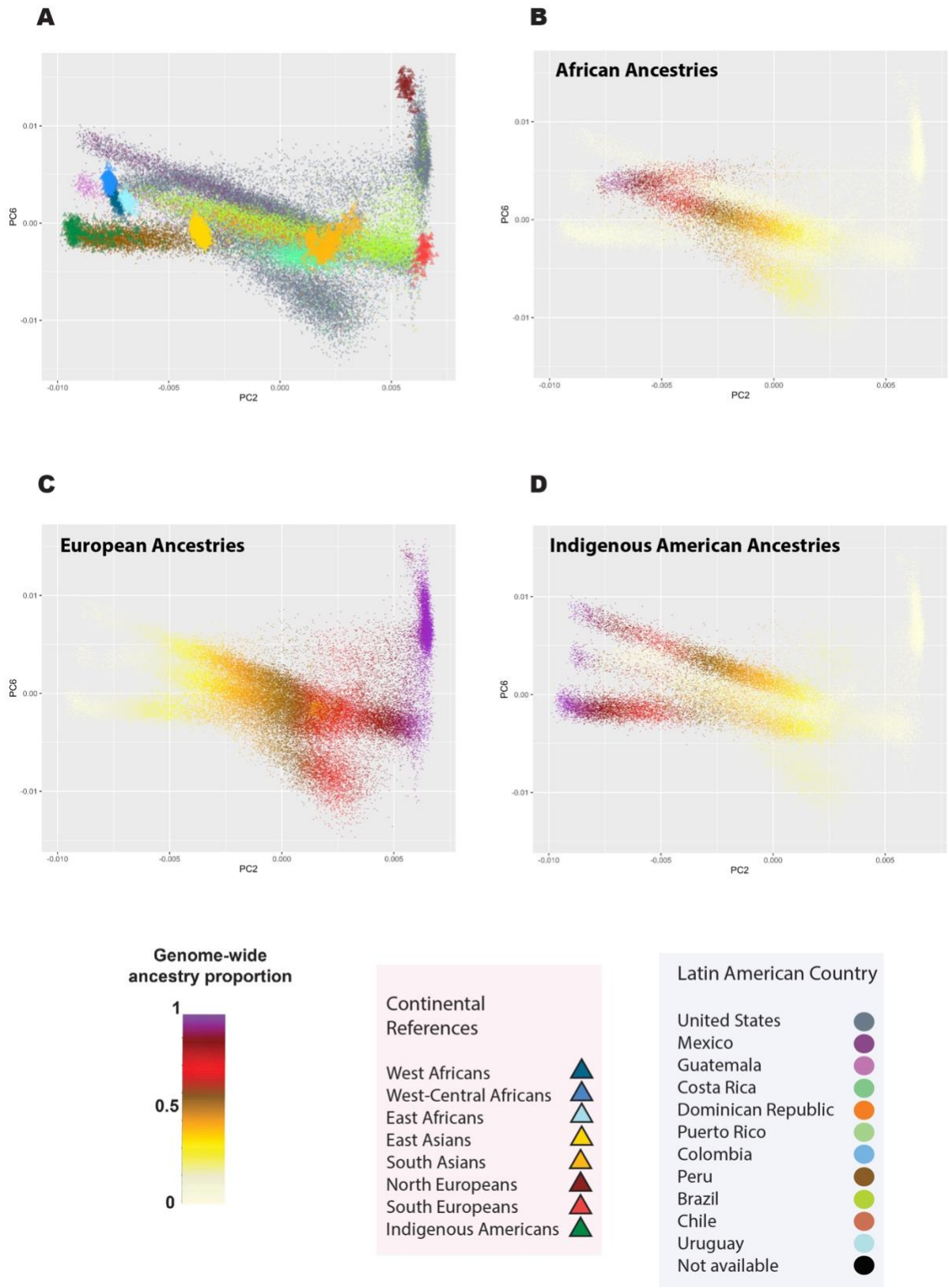

Figure S12. Genome-wide ancestry clines projected on Principal Component Analysis, related to Figure 1. Continental ancestry clines based on ancestry proportions inferred by ADMIXTURE for African (AFR), European (EUR), and Indigenous American (IA) ancestries in GLADdb individuals. (A) Dispersion plot showing the relationship between PC2 and PC6. PC2 explains the distribution of African and European ancestries, and PC6 explains the distribution of Indigenous American ancestries. (C) PC2 vs PC6 with individuals colored based on European ancestry proportion. (B) PC2 vs PC6 with individuals colored based on African ancestry proportion. (D) PC2 vs PC6 with individuals colored based on Indigenous American ancestry proportion.

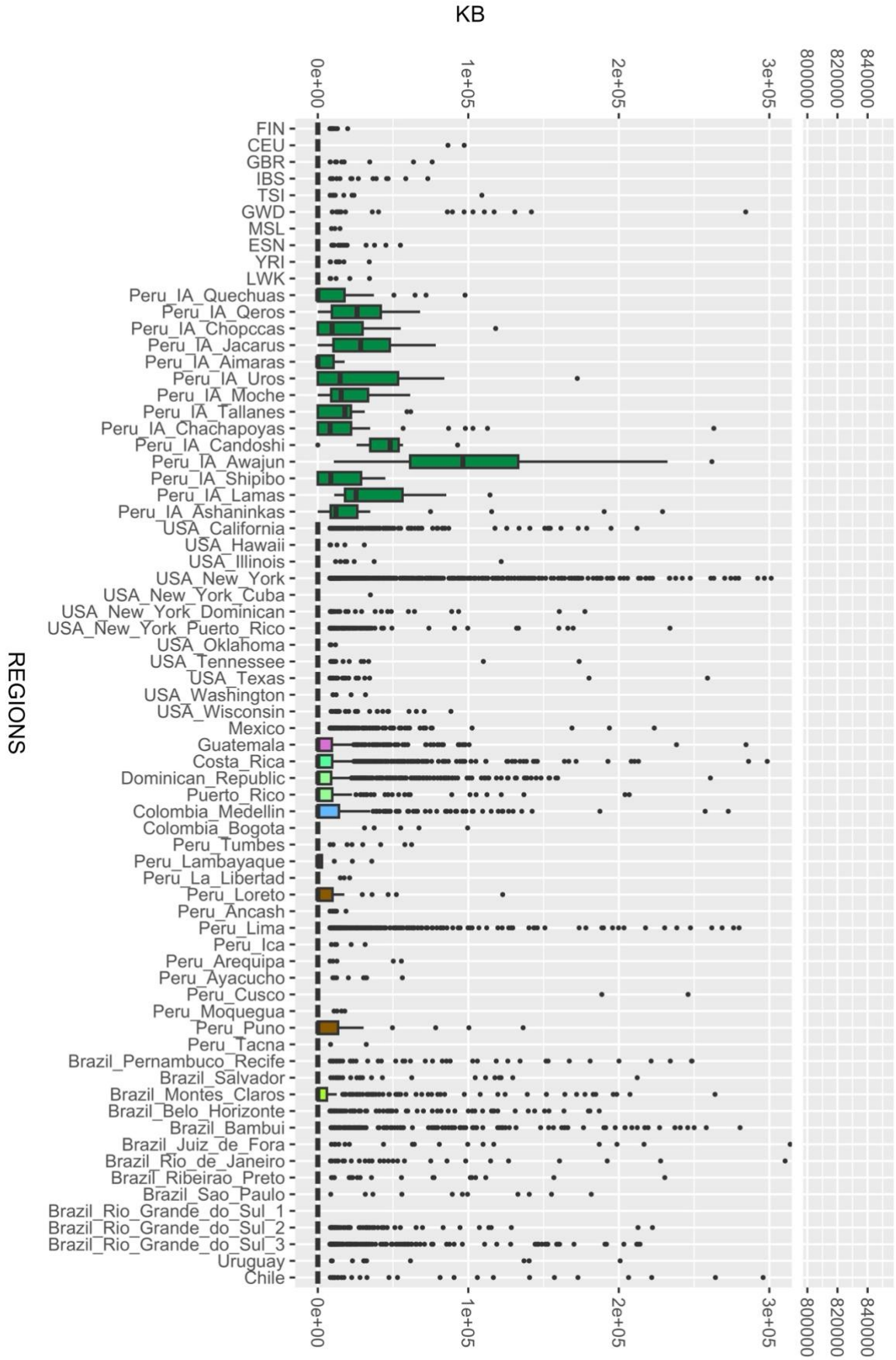

Figure S13. Distribution of Genome-wide amount of Runs of Homozygosity segments greater than 8Mb for Latin American groups and Reference populations included in GLADdb, related to Figure 1.

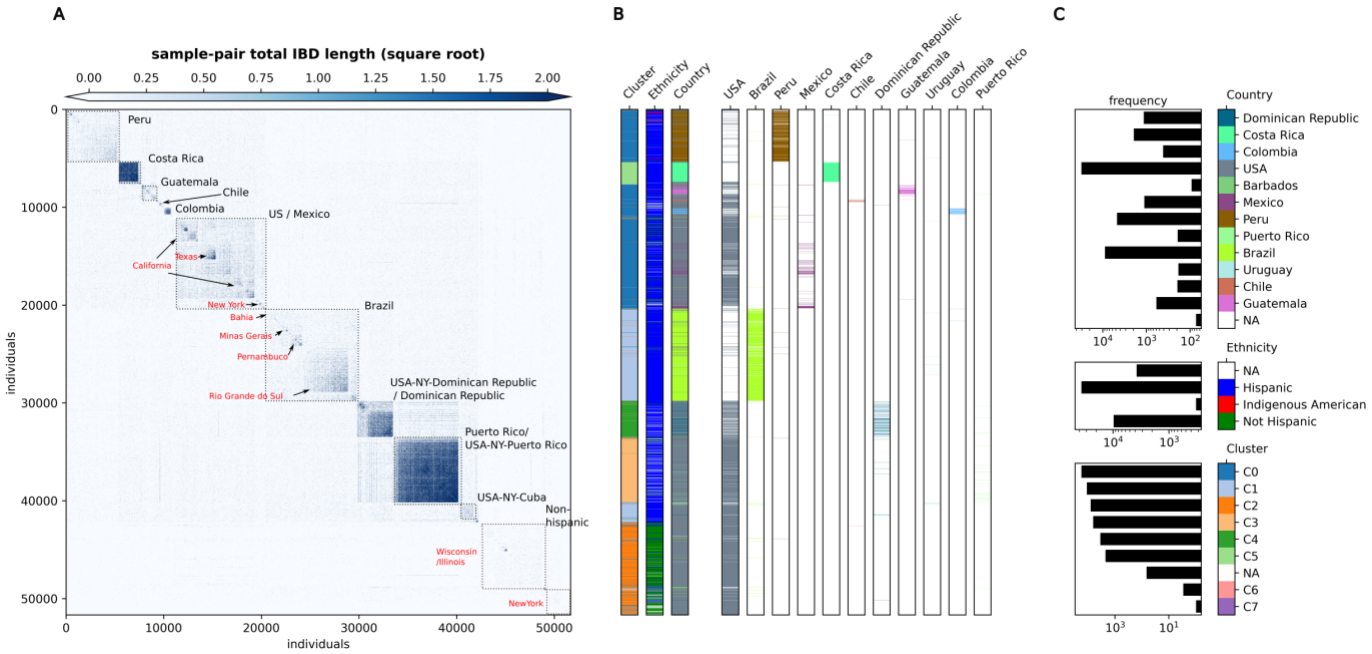

Figure S14. Clustering of total IBD matrix of unrelated individuals from GLADdb, related to Figure 2. A) Heatmap of the square root of sample-pair total IBD shared among unrelated individuals sampled from Latin American countries within GLADdb . Each pixel represents a pair of individuals; x and y axes indicate individual IDs sorted by unsupervised hierarchical clustering. Annotations within the heatmap represent the most enriched demographic labels in the indicated blocks. Labels with “USA-NY-country” correspond to self-described US-Hispanic living in New York with a specific country of origin. B) Individual-level annotations for the heatmap. The annotations include (i) labels of cluster assignment based on the Louvain algorithm via louvain-igraph package in the 1st vertical bar, (ii) self-described ethnicity in the 2nd bar, and (iii) sampling Country (combined indicators in the 3rd bar and country-specific indicators in the 4th to 14th bars). Each row in these bars corresponds to an individual. Note that the row orders in all label bars in B) are shared with that of A). C) Frequency of labels (log scale) and color keys for the Louvain clustering (bottom), self-described ethnicity (middle), and sample country (top), respectively. Note that “NA” label means individuals who are not assigned with any of the non-NA labels.

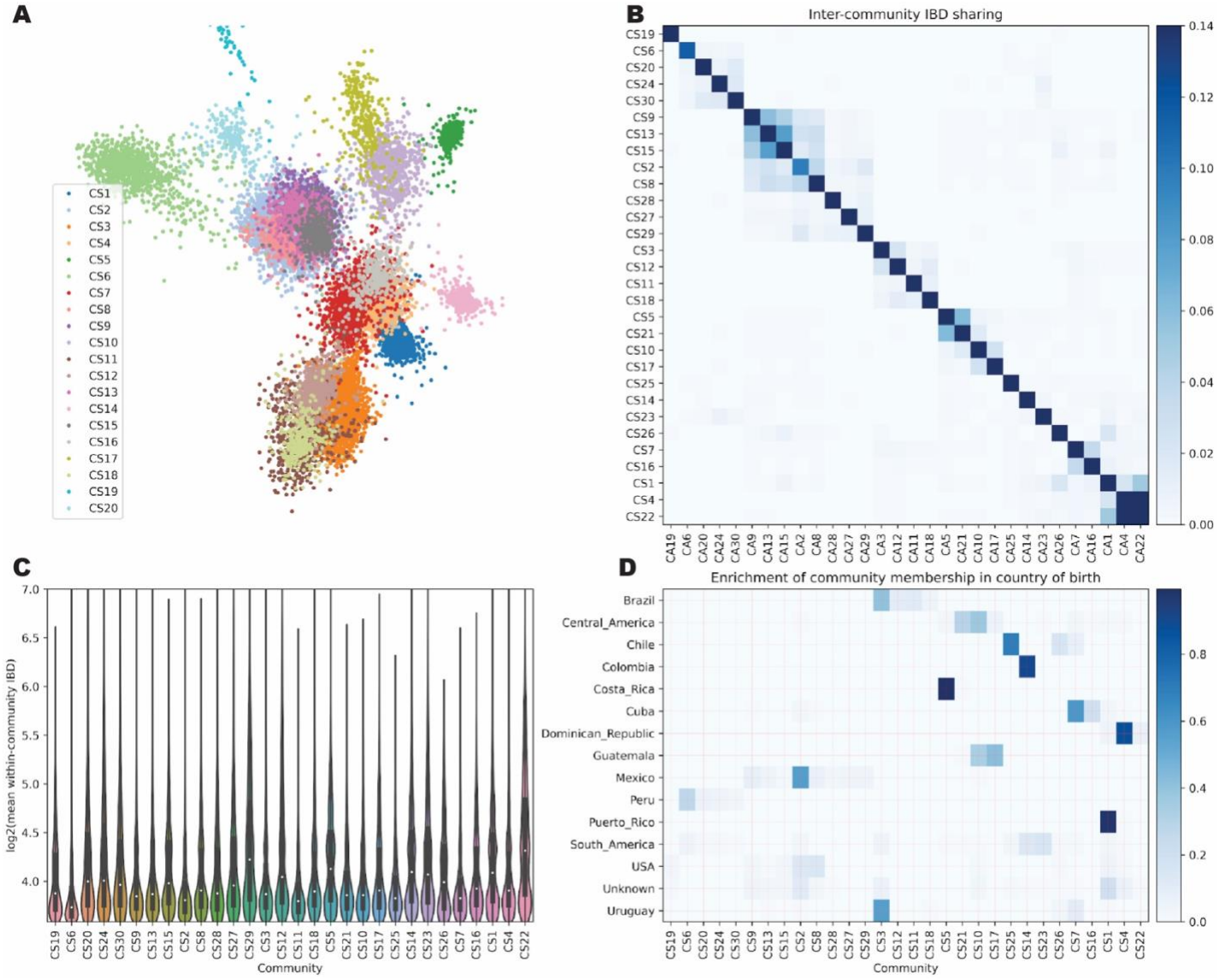

Figure S15. IBD network community detection using IBD segments between 5-9.3cM, related to Figure 3. This interval was selected to explore the population dynamics before colonial times. A) Top 20 IBD network communities visualized using Fruchterman-Reingold layout algorithm. Only individuals with connections > 30 are included in the layout calculation for visualization purposes. The community labels, such as CS1 and CS2, are named according to the IBD version used and the rank of the community sizes, with CS1 representing the largest community when using short IBD segments (5-9.3cM). B) IBD sharing among top 30 inferred communities (ordered by agglomerative clustering; the same order was followed in C and D). C) Distribution of IBD shared among individuals in each community. D) Enrichment of IBD community membership in the country of origin (i.e., proportions of community labels for individuals born in a given country).

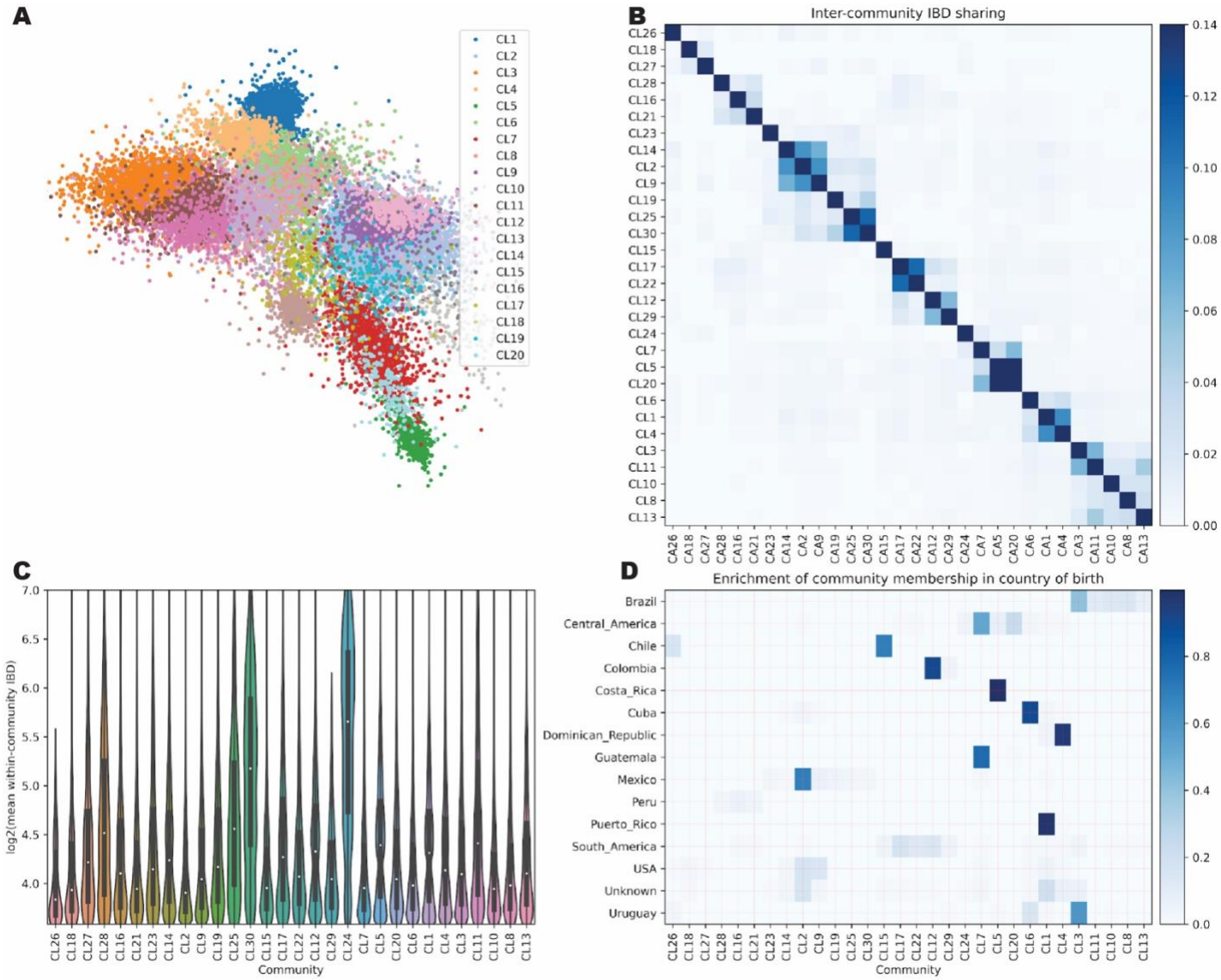

Figure S16. IBD network community detection using IBD segments greater than 9.3cM, related to Figure 3. This interval was selected to explore the population dynamics after the colonial times. A) Top 20 IBD network communities visualized using Fruchterman-Reingold layout algorithm. Only individuals with connections > 30 are included in the layout calculation for visualization purposes. The community labels, such as CL1 and CL2, are named according to the IBD version used and the rank of the community sizes, with CL1 representing the largest community when using large IBD segments (> 9.3cM). B) IBD sharing among top 30 inferred communities (ordered by agglomerative clustering; the same order was followed in C and D). C) Distribution of IBD shared among individuals in each community. D) Enrichment of IBD community membership in the country of origin (i.e., proportions of community labels for individuals born in a given country).

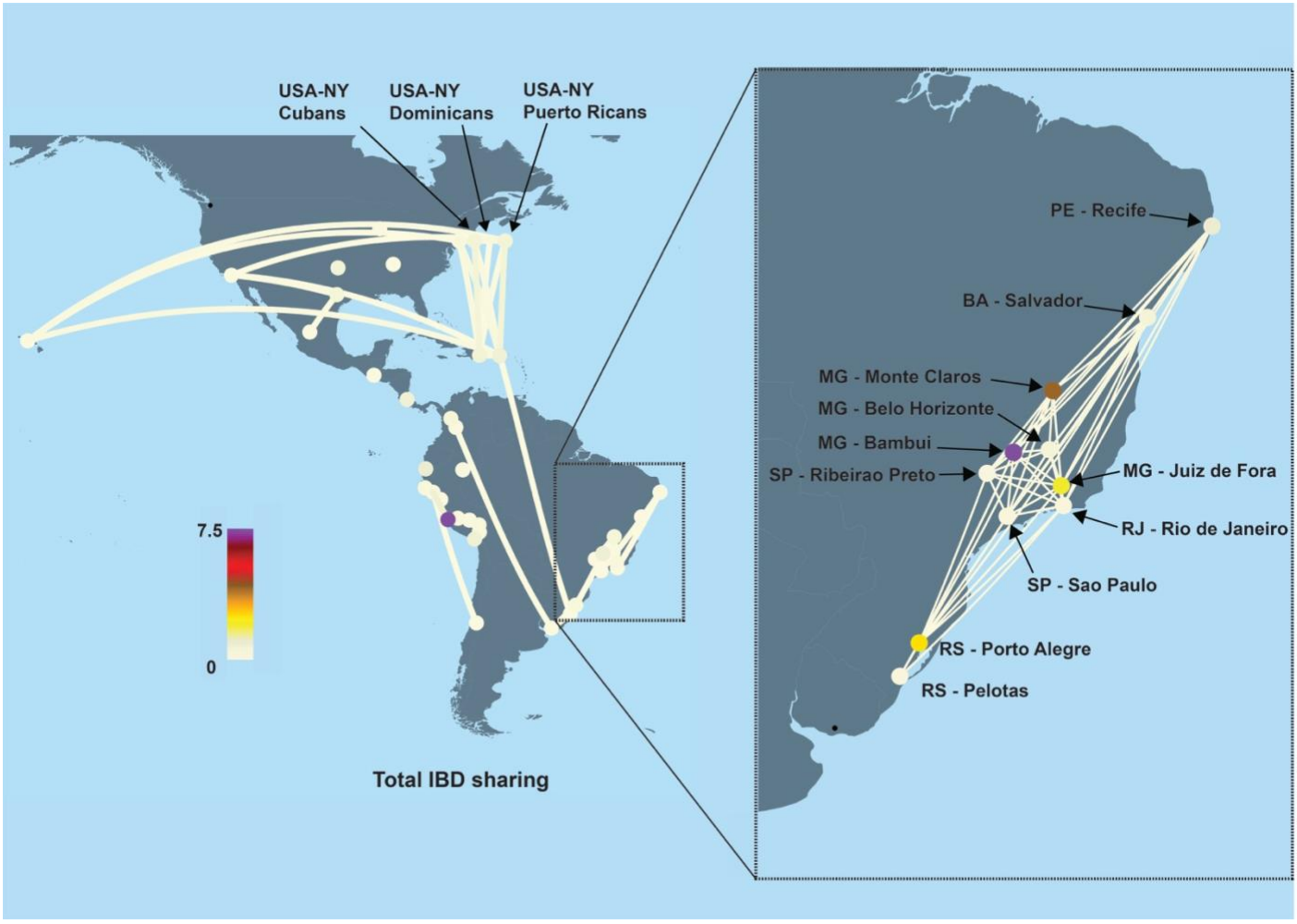

Figure S17. IBD network greater than 21cM, related to Figure 4. We explored the relationship among LAm regions by inferring the average IBD shared among all regions (left plot). Additionally, we conducted a focused analysis of the IBD sharing network, specifically within Brazil, as depicted in the right plot.

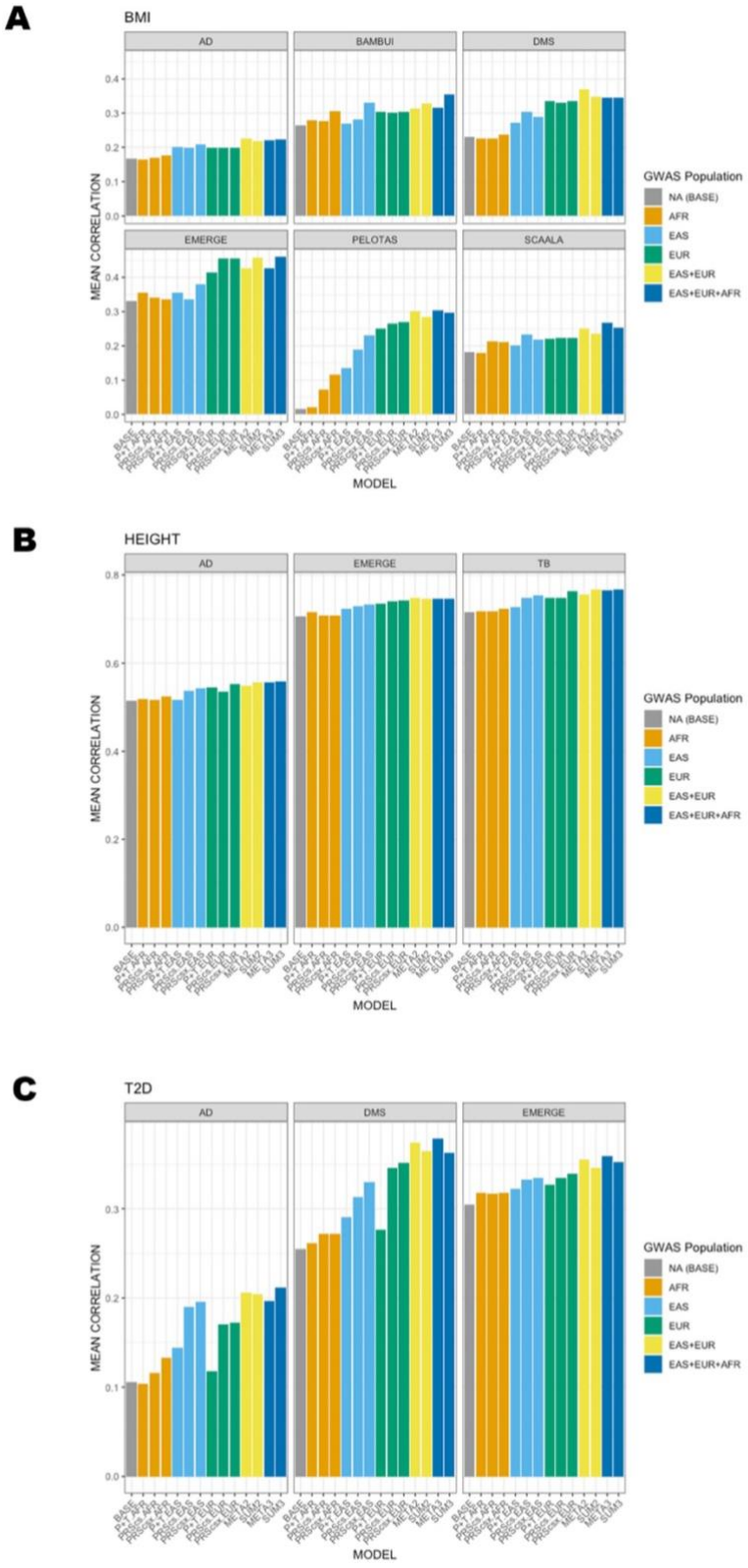

Figure S18. Predictive Performance as measured by the mean correlation of the trait with the prediction, related to Figure 6. A: Predictive performance for BMI. B: Predictive performance for height. C: Predictive performance for T2D. AD: Columbia University Study of Caribbean Hispanics and Late Onset Alzheimer's disease (phs000496), DMS: Slim Initiative in Genomic Medicine for the Americas (SIGMA): Diabetes in Mexico Study (phs001388), EMERGE: eMERGE Network Phase III: HRC Imputed Array Data (phs001584), TB: Early Progression to Active Tuberculosis in Peruvians (phs002025), and EPIGEN-Brasil (Bambui, Pelotas, and SCAALA).

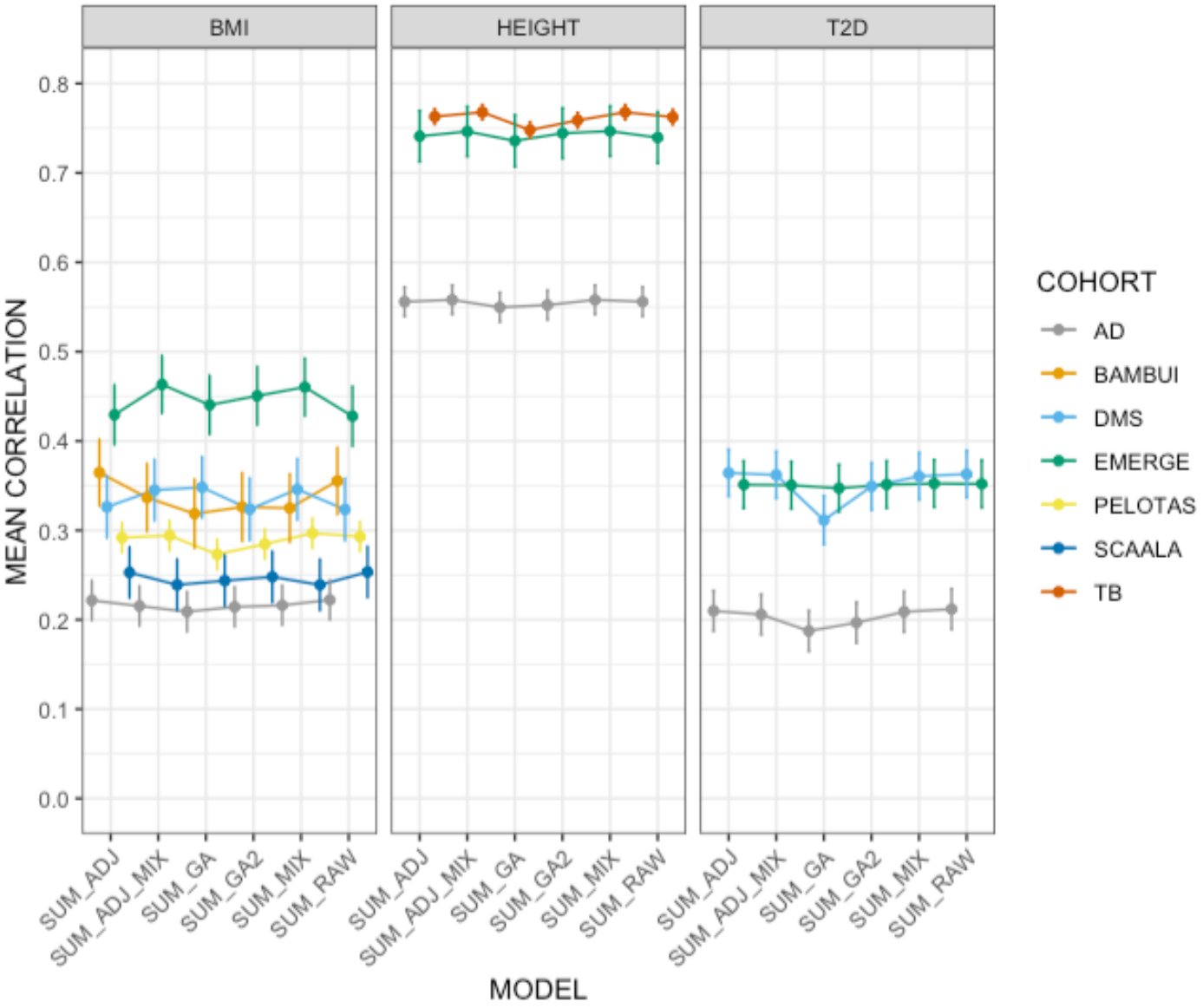

Figure S19. PRS linear combination methods across three traits, related to STAR Methods. Predictive performance of linear combination methods across 3 traits and 7 cohorts. Error bars represent the standard error of the correlation. See STAR Methods for model definitions. AD: Columbia University Study of Caribbean Hispanics and Late Onset Alzheimer's disease (phs000496), DMS: Slim Initiative in Genomic Medicine for the Americas (SIGMA): Diabetes in Mexico Study (phs001388), EMERGE: eMERGE Network Phase III: HRC Imputed Array Data (phs001584), TB: Early Progression to Active Tuberculosis in Peruvians (phs002025), and EPIGEN-Brasil (Bambui, Pelotas, and SCAALA).

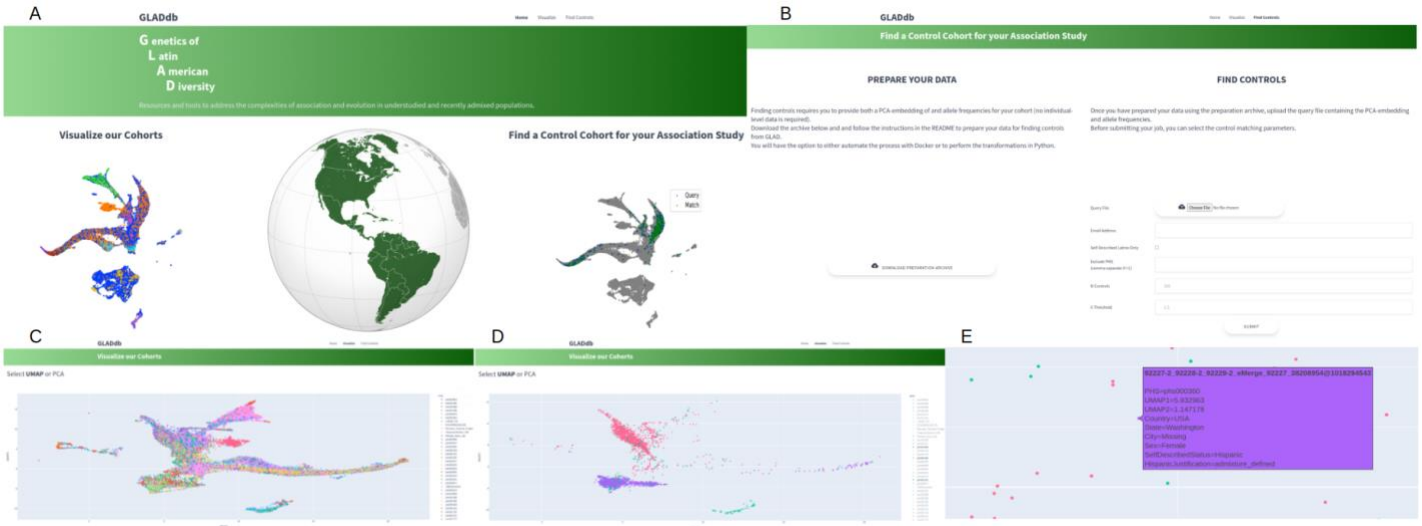

Figure S20. Screenshots from GLADdb website, related to STAR Methods. A) Landing page. B) The control matching page, encompassing data preparation and matching job submission. C-E) Visualization pages, showing all cohorts, selected cohorts, and a zoomed-in view with a highlighted individual.

SUPPLEMENTARY TABLES

Table S3. List of reference populations employed as parental populations for Local ancestry analyses, related to STAR Methods.

Population	Number of individuals	Ancestry label	Country	Source	Reference
IBS	100	European	Spain	1000 Genome Project	[S1]
CEU	50	European	USA	1000 Genome Project	[S1]
TSI	50	European	Italy	1000 Genome Project	[S1]
GBR	50	European	United Kingdom	1000 Genome Project	[S1]
YRI	41	African	Nigeria	1000 Genome Project	[S1]
MSL	48	African	Sierra Leona	1000 Genome Project	[S1]
ESN	32	African	Nigeria	1000 Genome Project	[S1]
GWD	30	African	Gambia	1000 Genome Project	[S1]
LWK	99	African	Kenya	1000 Genome Project	[S1]
JPT	50	East Asian	Japan	1000 Genome Project	[S1]
CHS	50	East Asian	China	1000 Genome Project	[S1]
CHB	50	East Asian	China	1000 Genome Project	[S1]
CDX	50	East Asian	China	1000 Genome Project	[S1]
KHV	50	East Asian	Vietnam	1000 Genome Project	[S1]
Uros	11	Native American	Peru	Peruvian Genome Project	[S2, S3]
Tallanes	25	Native American	Peru	Peruvian Genome Project	[S3]
Shipibo	14	Native American	Peru	Peruvian Genome Project	[S3]
Shimaa	22	Native American	Peru	Peruvian Genome Project	[S3]
Quechuas	2	Native American	Peru	Peruvian Genome Project	[S3]
Qeros	7	Native American	Peru	Peruvian Genome Project	[S2, S3]
Nahua	2	Native American	Peru	Peruvian Genome Project	[S3]
Moche	2	Native American	Peru	Peruvian Genome Project	[S2, S3]
Matsiguenkas	3	Native American	Peru	Peruvian Genome Project	[S2, S3]
Matses	11	Native American	Peru	Peruvian Genome Project	[S2, S3]
Lamas	7	Native American	Peru	Peruvian Genome Project	[S3]
Chopccas	2	Native American	Peru	Peruvian Genome Project	[S2, S3]
Candoshi	15	Native American	Peru	Peruvian Genome Project	[S3]
Awajun	22	Native American	Peru	Peruvian Genome Project	[S3]
Ashaninkas	34	Native American	Peru	Peruvian Genome Project	[S3]
Aimaras	9	Native American	Peru	Peruvian Genome Project	[S3]
Guatemala	44	Native American	Guatemala	NCI Michael Dean Lab	-

Table S5. Cohort descriptions for Polygenic Risk Score analyses, related to STAR Methods.

STUDY	ABBREV	PHS	TRAIT	MEAN (SD) TRAIT	h2 (SE)	N (UNR)	AGE (SD)	SEX RATIO	REFERENCE
Multiethnic GWAS of Prostate Cancer	MEC_PRCA	phs000306	BMI	24.9-29.9 (NA)	0.0 (0.26)	1710 (1251)	70-75 (NA)	NA	[S4]
Columbia University Study of Caribbean Hispanics with Familial and Sporadic Late Onset Alzheimer's disease	AD	phs000496	BMI	26.23 (5.62)	0.51 (0.16)	2802 (2053)	74.5 (9.29)	0.52	[S5]
			HEIGHT	160.10 (10.03)	0.33 (0.15)	2802 (2053)	74.5 (9.29)	0.52	
Slim Initiative in Genomic Medicine for the Americas (SIGMA): Diabetes in Mexico Study	DMS	phs001388	BMI	28.73 (5.27)	0.58 (0.19)	676 (675)	57 (11.1)	0.41	[S6]
eMERGE Network Phase III: HRC Imputed Array Data	EMERGE	phs001584	HEIGHT	163.54 (17.73)	0.0 (0.50)	271 (271)	50.56 (20.17)	0.5	[S7]
			BMI	28.73 (5.27)	0.94 (0.30)	629 (619)	46.80 (18.50)	0.63	
Early Progression to Active Tuberculosis in Peruvians	TB	phs002025	HEIGHT	160.20 (8.83)	0.39 (0.09)	3134 (2620)	34.90 (13.80)	1.34	[S8]
EpiGen BAMBUI	BAMBUI	NA	BMI	25.14 (4.98)	0.99 (0.29)	1396 (574)	68.90 (6.98)	0.65	[S9]
EpiGEN PELOTAS	PELOTAS	NA	BMI	23.70 (4.38)	0.37 (0.09)	3719 (3058)	23.00 (0)	1.01	
EpiGen SCAALA	SCAALA	NA	BMI	15.52 (2.04)	0.0 (0.36)	1300 (1116)	6.87 (1.68)	1.18	
PAGE: IPM BioMe Biobank	BIOME	phs000925	HEIGHT (LAT)	163.62 (9.84)	0.37 (0.07)	4486 (4486)	52.36 (15.87)	0.62	[S10]
			HEIGHT (ALL)	166.15 (10.26)	0.59 (0.03)	12968 (11625)	51.29 (15.41)	0.64	
PAGE: IPM BioMe Biobank	BIOME	phs000925	BMI (LAT)	29.40 (6.51)	0.28 (0.07)	4486 (4486)	52.36 (15.87)	0.62	
			BMI (ALL)	29.47 (7.08)	0.73 (0.03)	12968 (11625)	51.29 (15.41)	0.64	
Multiethnic GWAS of Prostate Cancer	MEC_PRCA	phs000306	PROSTATE CANCER	0.004 (0.24)	ALL	1303 (1255)	70-75 (NA)	NA	[S4]
					CASES	690 (662)	65-70 (NA)	NA	
					CONTROLS	613 (593)	70-75 (NA)	NA	
IPM BioBank GWAS	IPM	phs000388	T2D	0.36 (0.24)	ALL	1276 (906)	63.9 (12.2)	0.67	[S10]
					CASES	665 (514)	65.8 (10.6)	0.692	
					CONTROLS	611 (392)	61.7 (13.3)	0.647	
			CAD	0.23 (0.25)	ALL	1208 (848)	63.5 (12.2)	0.666	
					CASES	415 (354)	67.6 (9.93)	0.878	
					CONTROLS	793 (494)	61.4 (12.7)	0.573	
Columbia University Study of Caribbean Hispanics with Familial and Sporadic Late Onset Alzheimer's disease	AD	phs000496	T2D	0.18 (0.16)	ALL	3133 (2324)	74.9 (9.3)	0.5	[S5]
					CASES	797 (627)	75 (8)	0.436	
					CONTROLS	2336 (1696)	74.8 (9.71)	0.524	
Multiethnic Cohort (MEC)	MEC_BRCA	phs000517	BREAST CANCER	0.0 (0.63)	ALL	459 (452)	65-69 (NA)	NA	[S4]

Breast Cancer Genetics					CASES	301 (295)	65-69 (NA)	NA	
					CONTROLS	158 (157)	65-69 (NA)	NA	
Slim Initiative in Genomic Medicine for the Americas (SIGMA): Diabetes in Mexico Study (DMS)	DMS	phs001388	T2D	0.23 (0.15)	ALL	1147 (1146)	54.4 (10)	0.43	[S6]
					CASES	681 (680)	55.7 (11.2)	0.497	
					CONTROLS	466 (466)	55.7 (11.2)	0.343	
eMERGE Network Phase III: HRC Imputed Array Data	EMERGE	phs001584	T2D	0.21 (0.25)	ALL	1203 (1142)	NA	0.746	[S7]
					CASES	734 (684)	NA	0.735	
					CONTROLS	469 (458)	NA	0.763	

STUDY: study name. ABBREV: abbreviation used in this work PHS: PHS number. TRAIT: trait name. MEAN (SD): mean and standard deviation of trait in study; height is in cm. h2 (SE): additive heritability and standard error of trait using GCTA. SUBSET: subset of cohort, either cases, controls, or all. N (UNR): total sample size and sample size of unrelated subjects. AGE (SD): mean and standard deviation of age in study in years. SEX RATIO: ratio of males:females.

Table S6. GWAS summary statistics. Publicly-available African, European, and East Asian ancestry GWAS data to construct both single-ancestry PRS and multi-ancestry PRS, related to STAR Methods.

TRAIT	POP	N	N CASES	h2 (SE)	h2 liab (SE)	GC lambda	Reference
T2D	AFR	4347	2633	0.2712 (0.1412)	0.2241 (0.1166)	1.0016	
	EAS	191764	36614	0.1083 (0.0068)	0.1629 (0.0103)	1.334	[S11]
	EUR	420529	24348	0.0434 (0.0025)	0.1999 (0.0116)	1.2831	UKBB pan-ancestry https://pan.ukbb.broadinstitute.org/
PrCa	EAS	109347	5408	0.0555 (0.008)	0.1422 (0.0206)	1.0165	[S12]
	EUR	140306	79194	0.3232 (0.0818)	0.1553 (0.0393)	1.2564	[S13]
BrCa	EAS	95283	5552	0.0342 (0.0085)	0.074 (0.0185)	1.0225	[S12]
	EUR	228951	122977	0.1384 (0.0114)	0.0695 (0.0057)	1.3546	[S14]
CAD	EAS	212453	29319	0.0781 (0.0082)	0.1392 (0.0147)	1.2005	[S12]
	EUR	296525	34541	0.0756 (0.0045)	0.1557 (0.0093)	1.3101	[S15]
HEIGHT	AFR	20521	NA	0.1302 (0.0321)	NA	0.768	[S16]
	EAS	159095	NA	0.419 (0.0198)	NA	1.7648	[S17]
	EUR	695648	NA	0.4579 (0.0194)	NA	3.6129	[S18]
BMI	AFR	20521	NA	0.0978 (0.0266)	NA	0.7575	[S16]
	EAS	159195	NA	0.1684 (0.0076)	NA	1.4566	[S19]
	EUR	683365	NA	0.1958 (0.0055)	NA	2.7872	[S18]

TRAIT: trait name. POP: ancestry of GWAS cohort. N: total sample size. N CASES: sample size of cases only. h2 (SE): heritability and standard error of trait estimated using LD score regression. h2 liab (SE): same as h2(SE) but on the liability scale. GC lambda: genomic control lambda as estimated by LD

Table S7. Comparison between pairs of PRS models, related to Figure 6 and STAR Methods.

MODEL	MODEL2	PROP MODEL1	PROP MODEL2	PROP SIGNIF
BASE	EUR3	0	1	1
BASE	EUR2	0	1	1
BASE	META3	0	1	1
BASE	META2	0	1	1
BASE	P+T	0.11	0.89	0.53
BASE	PRS-CS	0.03	0.97	0.69
BASE	PRS-CSx	0	1	0.83
BASE	SUM3	0	1	1
BASE	SUM2	0	1	1
META3	EUR3	0.92	0.083	0.58
META3	META2	0.58	0.42	0.33
META3	SUM3	0.33	0.67	0.17
META2	EUR2	0.75	0.25	0.75
META2	SUM2	0.67	0.33	0.42
P+T	PRS-CS	0.39	0.61	0.39
P+T	PRS-CSx	0.11	0.89	0.42
PRS-CS	PRS-CSx	0.14	0.86	0.39
SUM3	EUR3	1	0	0.58
SUM3	SUM2	0.83	0.17	0.17
SUM2	EUR2	1	0	0.58

MODEL1: first model in comparison. MODEL2: second model in comparison. PROP MODEL1: Proportion of times (across all cohorts and traits) model 1 predictions had a higher correlation with the trait than model 2. PROP MODEL2: Proportion of times model 2 predictions had a higher correlation with the trait than model 1 predictions. PROP SIGNIF: proportion of model comparisons where the comparison was statistically significant (p-value < 0.05) using the Williams Test from the psych R package. For P+T, PRS-CS, and PRS-CSx models, comparisons were made between algorithms when utilizing the same set of GWAS summary statistics. META2 and SUM2 models combined both East Asian and European GWAS data, while META3 and SUM3 models combined East Asian, European, and African GWAS data. EUR2 and EUR3 are PRS-CSx European-ancestry models estimated after leveraging East Asian (EUR2) or East Asian and African GWAS data (EUR3).

SUPPLEMENTAL REFERENCES

[S1]. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

[S2]. Harris, D.N., Song, W., Shetty, A.C., Levano, K.S., Cáceres, O., Padilla, C., Borda, V., Tarazona, D., Trujillo, O., Sanchez, C., et al. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc. Natl. Acad. Sci. *115*, E6526–E6535. https://doi.org/10.1073/pnas.1720798115.

[S3]. Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G.B., Leal, T.P., Furlan, V., Scliar, M.O., Zamudio, R., Zolini, C., et al. (2020). The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proc. Natl. Acad. Sci. *117*, 32557–32565. https://doi.org/10.1073/pnas.2013773117.

[S4]. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M.Y., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E., and Nagamine, F.S. (2000). A Multiethnic Cohort in Hawaii and Los Angeles: Baseline Characteristics. Am. J. Epidemiol. *151*, 346–357. https://doi.org/10.1093/oxfordjournals.aje.a010213.

[S5]. Lee, J.H., Cheng, R., Barral, S., Reitz, C., Medrano, M., Lantigua, R., Jiménez-Velazquez, I.Z., Rogaeva, E., St. George-Hyslop, P.H., and Mayeux, R. (2011). Identification of Novel Loci for Alzheimer Disease and Replication of CLU, PICALM, and BIN1 in Caribbean Hispanic Individuals. Arch. Neurol. *68*. https://doi.org/10.1001/archneurol.2010.292.

[S6]. Estrada, K., Aukrust, I., Bjørkhaug, L., Burtt, N.P., Mercader, J.M., García-Ortiz, H., Huerta-Chagoya, A., Moreno-Macías, H., Walford, G., Flannick, J., et al. (2014). Association of a Low-Frequency Variant in *HNF1A* With Type 2 Diabetes in a Latino Population. JAMA *311*, 2305. https://doi.org/10.1001/jama.2014.6511.

[S7]. Kho, A.N., Hayes, M.G., Rasmussen-Torvik, L., Pacheco, J.A., Thompson, W.K., Armstrong, L.L., Denny, J.C., Peissig, P.L., Miller, A.W., Wei, W.-Q., et al. (2012). Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J. Am. Med. Inform. Assoc. *19*, 212–218. https://doi.org/10.1136/amiajnl-2011-000439.

[S8]. Luo, Y., Suliman, S., Asgari, S., Amariuta, T., Baglaenko, Y., Martínez-Bonet, M., Ishigaki, K., Gutierrez-Arcelus, M., Calderon, R., Lecca, L., et al. (2019). Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. Nat. Commun. *10*, 3765. https://doi.org/10.1038/s41467-019-11664-1.

[S9]. Kehdy, F.S.G., Gouveia, M.H., Machado, M., Magalhães, W.C.S., Horimoto, A.R., Horta, B.L., Moreira, R.G., Leal, T.P., Scliar, M.O., Soares-Souza, G.B., et al. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc. Natl. Acad. Sci. *112*, 8696–8701. https://doi.org/10.1073/pnas.1504447112.

[S10]. Tayo, B.O., Teil, M., Tong, L., Qin, H., Khitrov, G., Zhang, W., Song, Q., Gottesman, O., Zhu, X., Pereira, A.C., et al. (2011). Genetic background of patients from a university medical center in Manhattan: implications for personalized medicine. PloS One *6*, e19166. https://doi.org/10.1371/journal.pone.0019166.

[S11]. Suzuki, K., Akiyama, M., Ishigaki, K., Kanai, M., Hosoe, J., Shojima, N., Hozawa, A., Kadota, A., Kuriki, K., Naito, M., et al. (2019). Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. Nat. Genet. *51*, 379–386. https://doi.org/10.1038/s41588-018-0332-4.

[S12]. Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.-K., Okada, Y., et al. (2020). Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat. Genet. *52*, 669–679. https://doi.org/10.1038/s41588-020-0640-3.

[S13]. The Profile Study, Australian Prostate Cancer BioResource (APCB), The IMPACT Study, Canary PASS Investigators, Breast and Prostate Cancer Cohort Consortium (BPC3), The PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Cancer of the Prostate

in Sweden (CAPS), Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS), The Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium, Schumacher, F.R., et al. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat. Genet. *50*, 928–936. https://doi.org/10.1038/s41588-018-0142-8.

[S14]. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94. https://doi.org/10.1038/nature24284.

[S15]. Van Der Harst, P., and Verweij, N. (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ. Res. *122*, 433–443. https://doi.org/10.1161/CIRCRESAHA.117.312086.

[S16]. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell *179*, 984-1002.e36. https://doi.org/10.1016/j.cell.2019.10.004.

[S17]. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393. https://doi.org/10.1038/s41467-019-12276-5.

[S18]. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649. https://doi.org/10.1093/hmg/ddy271.

[S19]. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467. https://doi.org/10.1038/ng.3951.