Supplementary note

Table of Contents

Evaluation of limitations of phasing approach

We sought to address several limitations of our current phasing analysis. First, 4.7% of variant pairs in the 4,775 trios (that we used for P_{trans} evaluation) were not present in gnomAD and thus not amenable to phasing even using cosmopolitan P_{trans} estimates. To understand how the proportion of variants amenable to phasing changes as a function of gnomAD reference sample size, we performed a subsampling analysis of gnomAD from 121,912 (all of gnomAD v2 after removing overlapping trio samples) down to 1,000, 10,000, or 100,000 samples (**Supplementary Fig. 6a**). We found that subsampling greatly reduces the proportion of variants amenable to phasing, but that accuracy is generally preserved. For example, when subsampling down to 10,000 samples, just 76.4% of variant pairs observed in the trios were amenable to phasing, but phasing accuracy remained high (91.9%) when using cosmopolitan P_{trans} estimates (compared to 93.6% accuracy when using the full gnomAD cohort).

We also assessed variant pairs with intermediate EM scores ($0.02 < P_{trans} < 0.55$) where our approach gives an indeterminate phase estimate. We found that nearly all (99.8%) of the variant pairs with intermediate P_{trans} scores included more common variants (AF ≥ 0.001) (**Supplementary Fig. 6b**). For variant pairs where the more common variant had AF ≥ 0.001, 9.5% of variant pairs had an intermediate P_{trans} score. In contrast, variant pairs where the more common variant with AF < 0.001, just 0.19% of variant pairs had an intermediate P_{trans} score. Intermediate P_{trans} scores can only occur when all four haplotypes are observed. For rare variants, it is less likely that all four haplotype combinations are observed in a population. This can be due to lower likelihood of sampling rare haplotypes and/or because rare variants are younger and have less opportunity for recombination/recurrent mutation to generate all haplotype combinations.

Finally, we investigated the seemingly counterintuitive observation that phasing accuracy was lowest for NFE, where we had the highest number of gnomAD reference samples. We postulated that this apparent lower phasing accuracy in NFEs might be due to the larger number of trios we tested (for example, we tested 2815 NFE trio samples compared to 73 AFR samples), rather than an issue with phasing of NFE samples in the gnomAD reference dataset itself. To test this, we randomly subsampled the NFE trio samples from 2815 trios down to 282 (10%), 563 (20%) or 1408 (50%) trios. Upon subsampling, we found that a smaller proportion of unique variant pairs were in lower AF bins ($< 1 \times 10^{-4}$) where phasing is most challenging (**Supplementary Fig. 6c**), with a corresponding improvement in accuracy upon subsampling of the trio samples (**Supplementary Fig. 6d**). These results suggest that the observation of a lower phasing accuracy in NFE is an artifact of ascertaining and testing a larger number of NFE trio samples. Intuitively, this artifact results from our approach of measuring accuracy using unique variant pairs within a population. With increasing numbers of trios tested in our trio validation set, more common variant pairs where phasing accuracy is higher are observed multiple times yet counted only once. In contrast, with larger numbers of trios tested, we observe a larger number of unique rare variant pairs where accuracy is lower.

Discussion on singleton variant pairs

Pairs of singleton variants pose a unique challenge. When a pair of singleton variants is observed in different individuals in a population, this provides evidence that the variants are on different haplotypes. However, if a pair of singleton variants is observed in the same individual in the population, we cannot readily distinguish whether the variants are on the same haplotype or different haplotypes as we lack information from other individuals in the population for singleton variants. For this reason, we have chosen to not report phasing estimates for singleton/singleton variant pairs that are observed in the same individual in gnomAD. Nonetheless, using our trio data, 93% of these singleton/singleton variant pairs observed in the same individual in gnomAD were in *cis* based on our trio validation data.

Multinucleotide variant analysis

To further examine the effect of recombination, we also analyzed a set of 20,319 multinucleotide variants (MNVs), which are pairs of genetic variants in *cis* that are very close together in physical distance (≤ 2 bp) and thus have minimal opportunity for recombination between them. These variants have previously been accurately phased using physical read data[1,38]. When examining this set of MNVs, we found that the phasing accuracy using our approach was 96.0%, with only 3.5% of MNVs phased incorrectly (the remaining 0.46% had indeterminate phasing estimates).

Discussion on accuracy for rare variant pairs in *cis*

We found that our approach was less accurate for rare variant pairs in *cis*. This lower accuracy for variants in *cis* is intuitive, as for rare variants, a recombination event or germline mutation event is much more likely to disrupt a haplotype comprised of two variants than to bring two rare variants onto the same haplotype. Consistent with this intuition, we found that for variants in *cis,* phasing accuracy diminished linearly with genetic distance as a measure of recombination rates, but that phasing accuracy was maintained across genetic and physical distances between pairs of variants in *trans*. Similarly, the phasing accuracy for variant pairs in *cis* was lower at more mutable sites such as CpG sites that are frequently methylated. Thus, users should exercise caution for rare variants at highly mutable sites where our approach predicts the variants to be *trans*.

Discussion on use of cosmopolitan versus population-specific phasing estimates

In our work, we compared population-specific estimates with phasing estimates derived from samples across all genetic ancestry groups in gnomAD v2 ("cosmopolitan"). While population-specific phasing estimates are more likely to match the haplotypes seen in a given individual, they utilize information from fewer samples in gnomAD. We found that, in general, population-specific estimates were similar in accuracy to using cosmopolitan estimates. For AFR individuals, however, we found that use of cosmopolitan estimates resulted in slightly lower phasing accuracy than the use of AFR-specific estimates for variants in *trans* (**Fig. 3a)**. This is consistent with the observation that there are more unique haplotypes seen in individuals of AFR genetic ancestry and/or older haplotypes in individuals of AFR genetic ancestry for which recombination is more likely to have occurred[39]. Moreover, there are other genetic ancestry groups not currently represented in gnomAD for which we expect this phasing approach to have lower accuracy than in the well-represented genetic ancestry groups. Additionally, we recognize that many individuals are not well represented by a discrete genetic ancestry group, but instead represent admixtures of two or more populations. Future work on phasing will likely benefit from considering ancestry as a continuous variable[40]. For analysis of patients with rare diseases carrying candidate compound heterozygous variants, our data suggests that population-specific estimates, when available, should be used first-line followed by cosmopolitan estimates.

Discussion on tabulation of co-occurring variant pairs in gnomAD

To aid the medical genetics community in interpreting the clinical significance of rare co-occurring variants in the context of recessive disease, we have released gene-wise counts of co-occurring variants across a spectrum of variant consequences (pLoF, missense, and synonymous) and allele frequencies. These counts of co-occurring variants provide a background frequency of compound heterozygous rare damaging variants and can be used to assess the probability that a given variant pair identified in a patient may have occurred by chance. These values are released in the gnomAD browser.

Our ability to identify rare variant pairs in *trans* in gnomAD v2 individuals is limited by the fact the same dataset was used for training. Indeed, in these individuals, our ability to detect variant pairs in *trans* extends largely to variant pairs with AF > 0.5%, as nearly all rarer variant combinations were dominated by indeterminate phase and very few predictions for variant pairs in *trans*. The per-gene variant co-occurrence resource developed and released here is therefore to be considered a first step in this space. We plan to use the predictions from this dataset on newer versions of gnomAD with additional samples, where we can more confidently predict rare variant pairs that are in *trans*.

Funding information for Genome Aggregation Database Consortium authors

Emelia J Benjamin: Framingham Heart Study 75N92019D00031; and R01HL092577

Matthew J. Bown: British Heart Foundation awards CS/14/2/30841 and RG/18/10/33842

Josée Dupuis: National Heart Lung and Blood Institute’s Framingham Heart Study Contract (HHSNI); National Institute for Diabetes and Digestive and Kidney Diseases (NIDDK) R DK

Martti Färkkilä: State funding for university level health research

Laura D. Gauthier: Intel, Illumina

Stephen J. Glatt: U.S. NIMH Grant R MH

Leif Groop: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312063 and 336822), Sigrid Jusélius Foundation; IMI 2 (grant No 115974 and 15881)

Mikko Hiltunen: Academy of Finland (grant 338182) Sigrid Jusélius Foundation the Strategic Neuroscience Funding of the University of Eastern Finland

Chaim Jalas: Bonei Olam

Jaakko Kaprio: Academy of Finland (grants 312073 and 336823)

Jacob McCauley: National Institute of Diabetes and Digestive and Kidney Disease Grant R01DK104844

Yukinori Okada: JSPS KAKENHI (19H01021, 20K21834), AMED (JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217, JP21ek0410075), JST Moonshot R&D (JPMJMS2021)

Michael J. Owen: Medical Research Council UK: Centre Grant No. MR/L010305/1, Program Grant No. G0800509

Aarno Palotie: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant numbers 312074 and 336824) and Sigrid Jusélius Foundation

John D. Rioux: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK; DK062432), from the Canadian Institutes of Health (CIHR GPG 102170), from Genome Canada/Génome Québec (GPH-129341), and a Canada Research Chair (#230625)

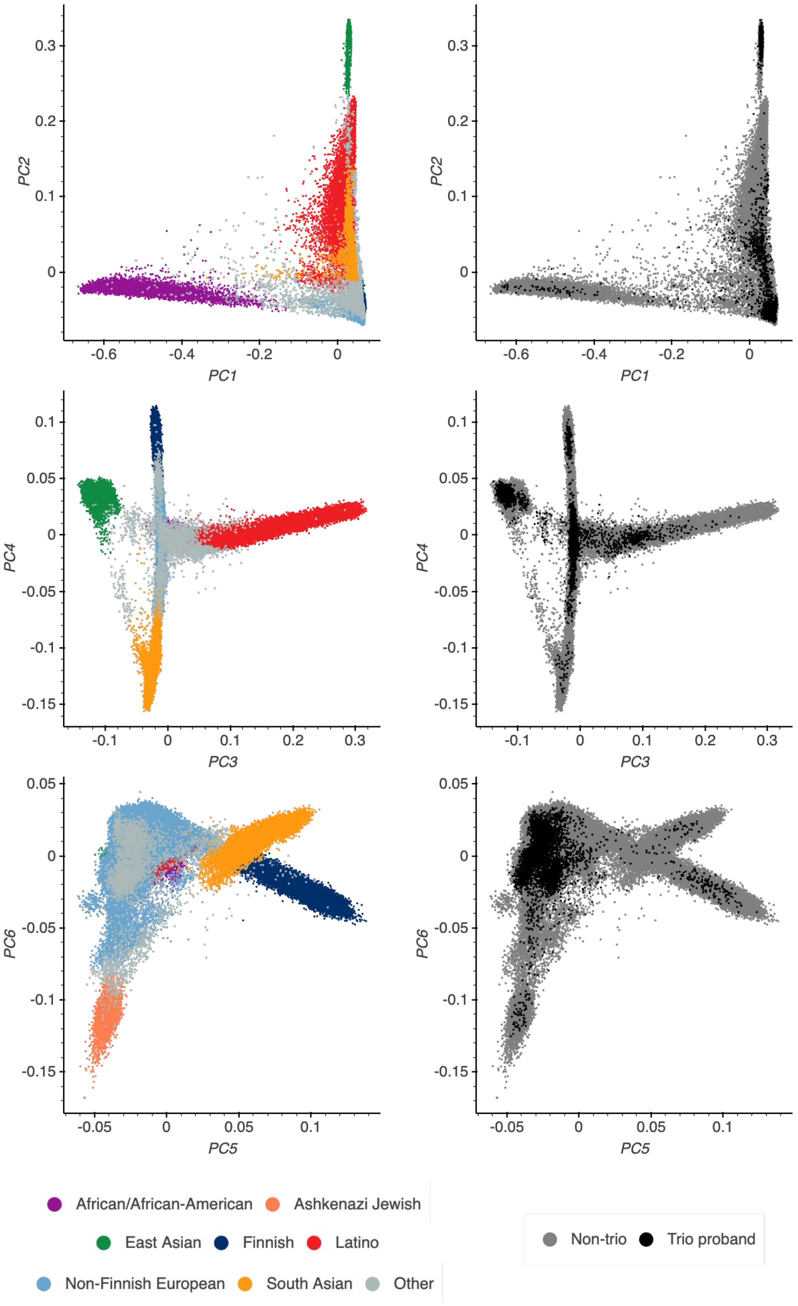
Samuli Ripatti: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant number) Sigrid Jusélius Foundation

Jerome I. Rotter: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is 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, and UL1-TR-001420. 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
Edwin K. Silverman: NIH Grants U01 HL089856 and U01 HL089897
J. Gustav Smith: The Swedish Heart-Lung Foundation (2019-0526), the Swedish Research Council (2017-02554), the European Research Council (ERC-STG-2015-679242), Skåne University Hospital, governmental funding of clinical research within the Swedish National Health Service, a generous donation from the Knut and Alice Wallenberg foundation to the Wallenberg Center for Molecular Medicine in Lund, and funding from the Swedish Research Council (Linnaeus grant Dnr 349-2006-237, Strategic Research Area Exodiab Dnr 2009-1039) and Swedish Foundation for Strategic Research (Dnr IRC15-0067) to the Lund University Diabetes Center
Kent D. Taylor: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is 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, and UL1-TR-001420. 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
Tiinamaija Tuomi: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312072 and 336826), Folkhalsan Research Foundation, Helsinki University Hospital, Ollqvist Foundation, Liv och Halsa foundation; NovoNordisk Foundation
Teresa Tusie-Luna: CONACyT Project 312688
James S. Ware: Sir Jules Thorn Charitable Trust [21JTA], Wellcome Trust [107469/Z/15/Z], Medical Research Council (UK), NIHR Imperial College Biomedical Research Centre
Rinse K. Weersma: The Lifelines Biobank initiative has been made possible by subsidy from the Dutch Ministry of Health Welfare and Sport the Dutch Ministry of Economic Affairs the University Medical Centre Groningen (UMCG the Netherlands) the University of Groningen and the Northern Provinces of the Netherlands

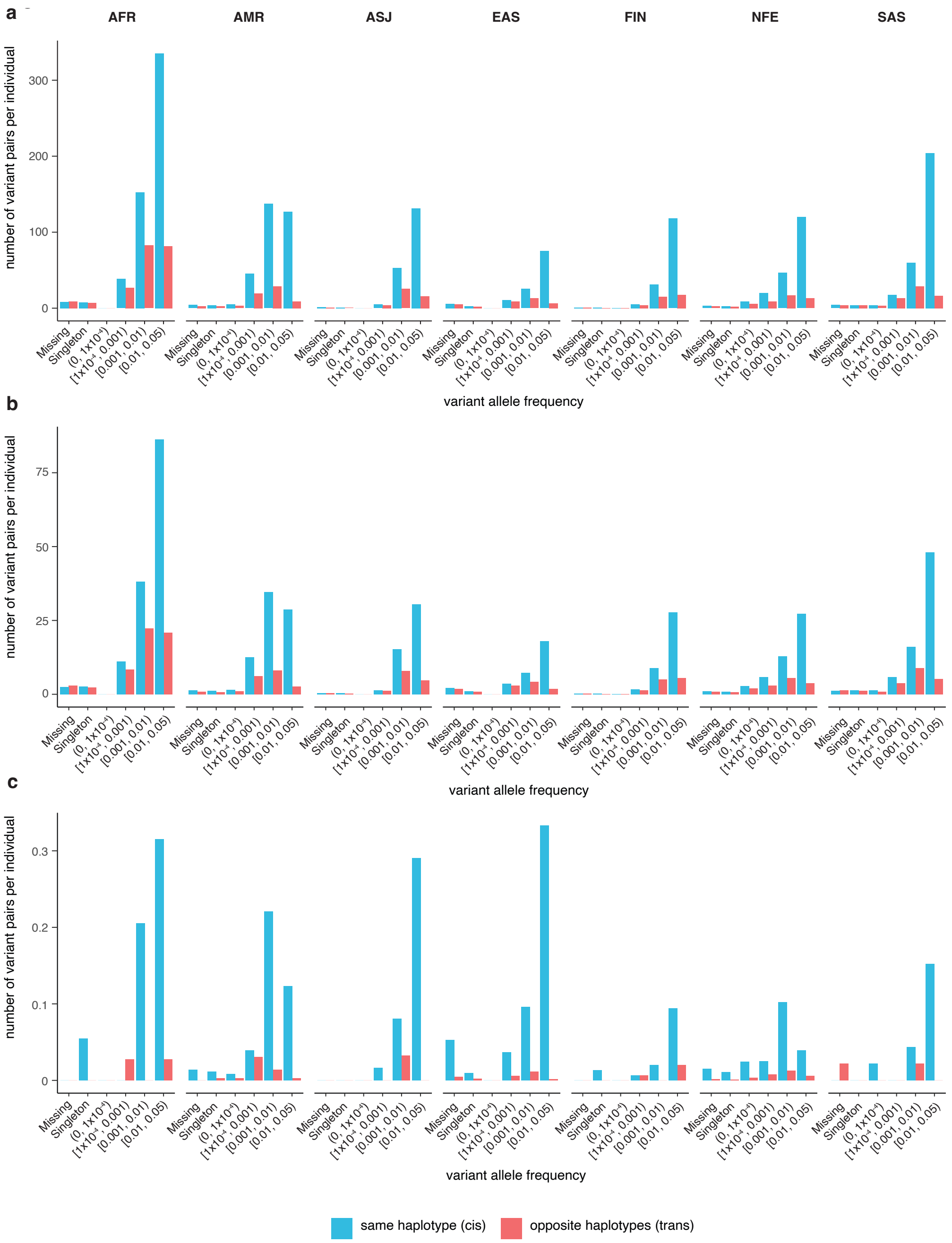
No conflicts of interest to declare

Supplementary Figures



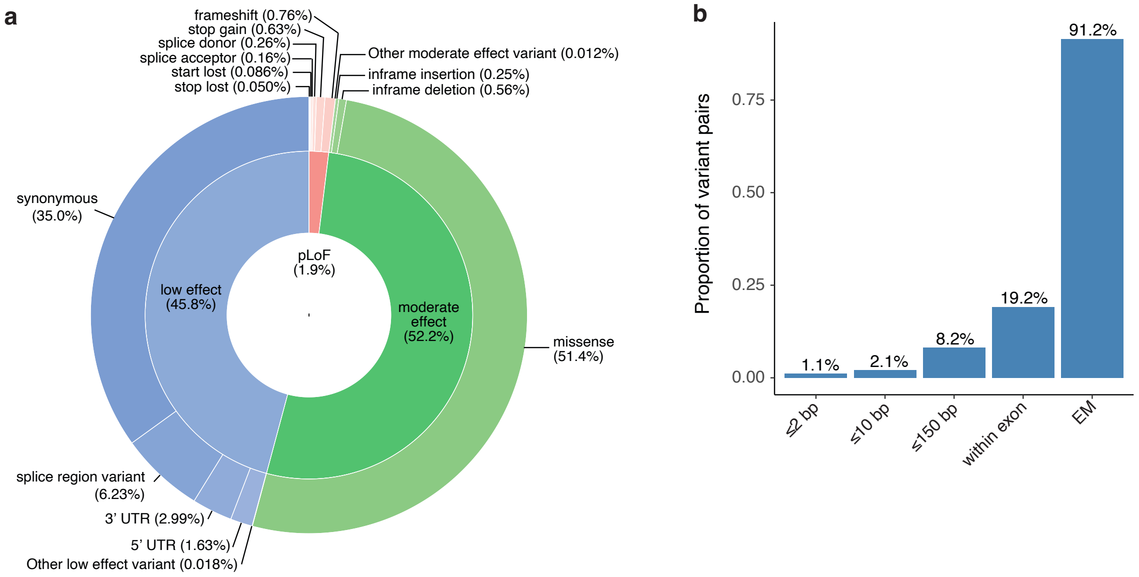
Supplementary Figure 1

Principal component analysis (PCA) plot for the full gnomAD v2 cohort (left) and specifically for the trios (right, trios in black) included in this paper. The top row shows PC1 vs PC2, the middle row shows PC3 vs PC4, and the bottom row shows PC5 vs PC6. Genetic ancestry group labels for the global gnomAD populations were done as described in Karczewski et al. 2020[14].



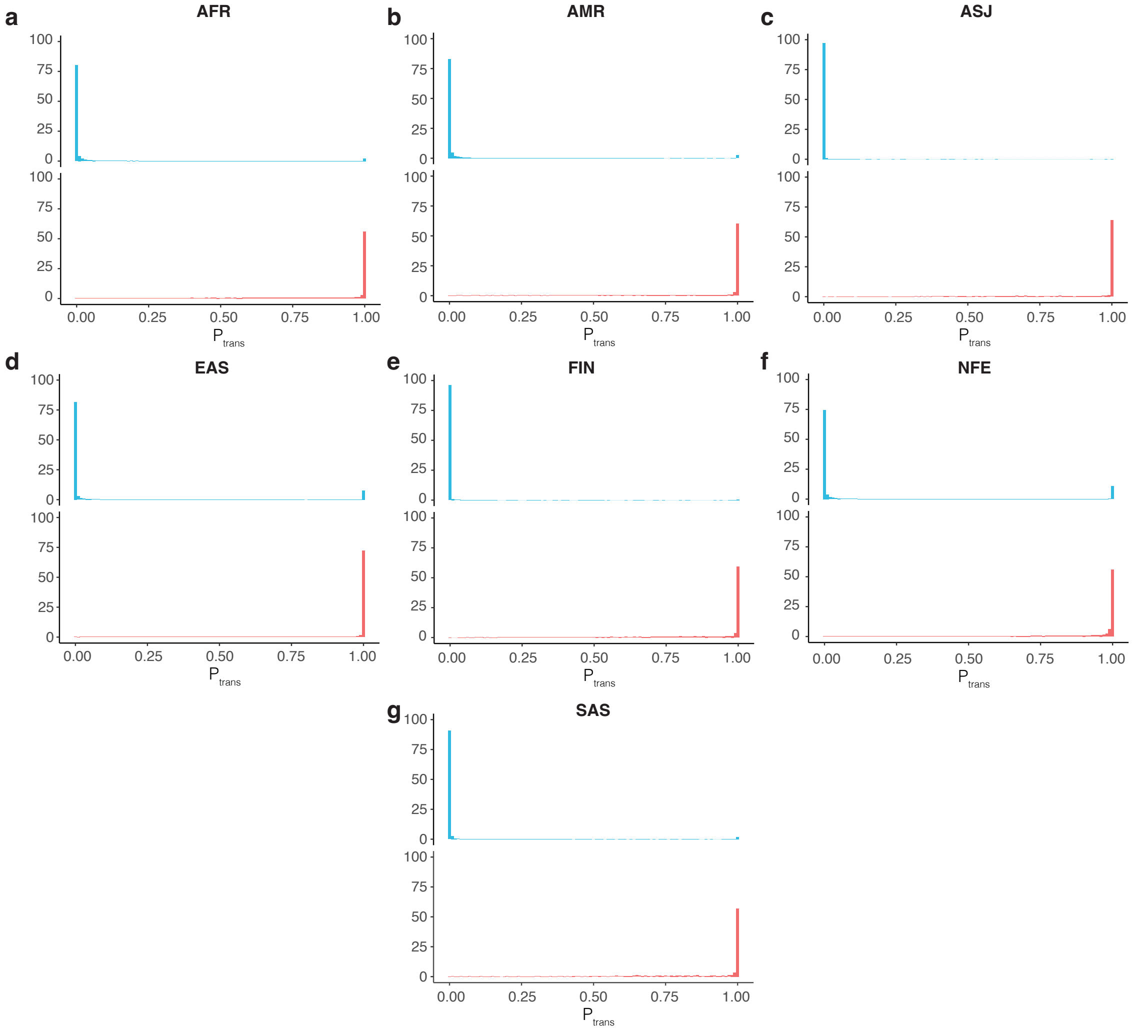
Supplementary Figure 2

Number of variant pairs observed per trio sample as a function of ancestry and AF. All variant pairs are shown in **a.** Variant pairs in which both variants are moderate effect or predicted loss-of-function (pLoF) are shown in **b.** Variant pairs in which both variants are pLoF are shown in **c.** Variant AF is the AF of the less common variant in a given variant pair and is population-specific frequency. AFR = African/African American; AMR = Admixed American/Latino; ASJ = Ashkenazi Jewish; EAS = East Asian; FIN = Finnish; NFE = non-Finnish European; SAS = South Asian.



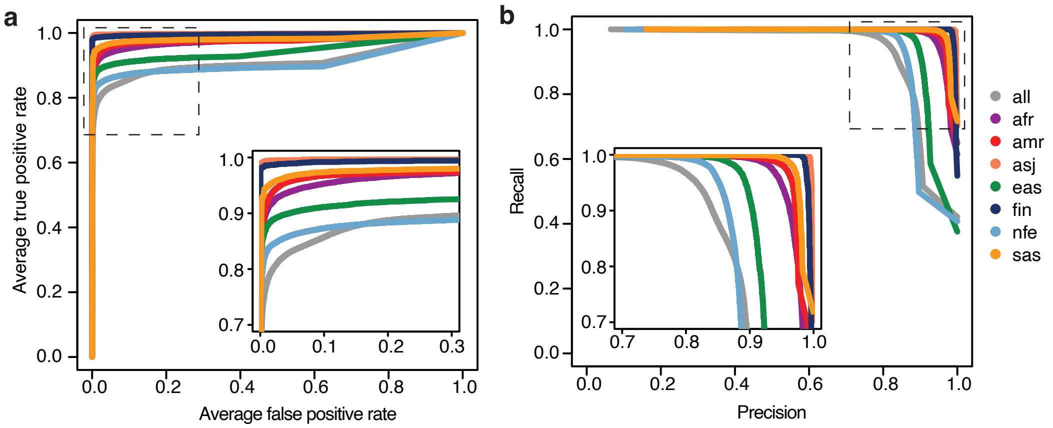
Supplementary Figure 3

a, Pie chart of variant effect annotations in the trio samples**.** Effect predictions are stratified among pLoF, moderate effect, and low effect variants. Percentages are shown in parentheses. **b,** Proportion of variant pairs falling within 2 bp, within 10 bp, within 150 bp, within the same exon, and proportion that can be phased using the EM algorithm applied to the gnomAD resource.



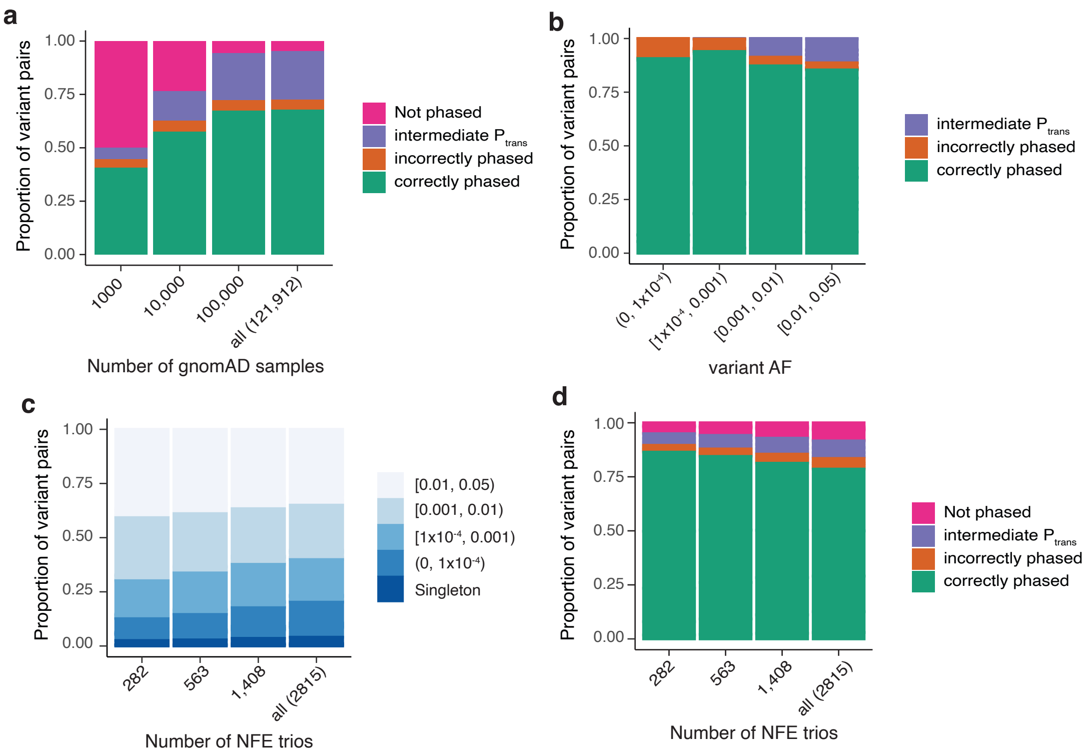
Supplementary Figure 4

a-g, Histogram of P_{trans} scores for variant pairs in *cis* (top, blue) and in *trans* (bottom, red) for each population. P_{trans} scores are population-specific. AFR = African/African American; AMR = Admixed American/Latino; ASJ = Ashkenazi Jewish; EAS = East Asian; FIN = Finnish; NFE = non-Finnish European; SAS = South Asian.



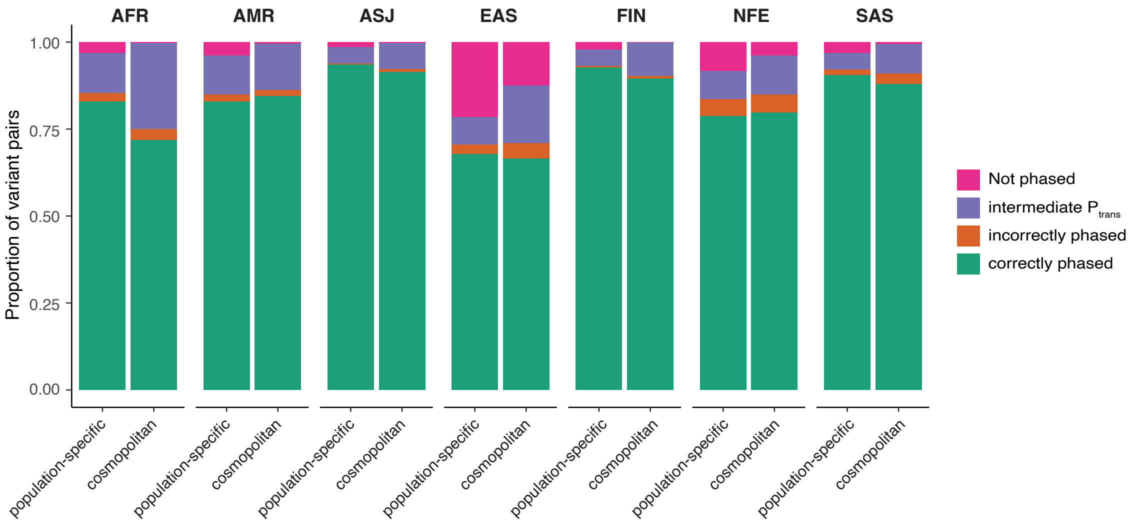
Supplementary Figure 5

Receiver-operator (**a**) and Precision-recall (**b**) curves for use of P_{trans} for distinguishing between variant pairs on same versus opposite haplotypes. Separate lines are shown for each genetic ancestry group. P_{trans} scores are population-specific. AFR = African/African American; AMR = Admixed American/Latino; ASJ = Ashkenazi Jewish; EAS = East Asian; FIN = Finnish; NFE = non-Finnish European; SAS = South Asian.



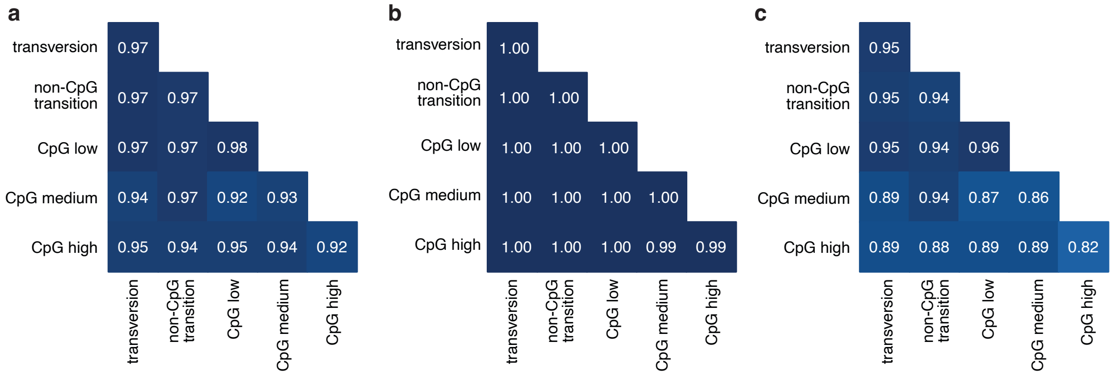
Supplementary Figure 6

a, Phasing performance when subsampling gnomAD to 1000, 10,000, 100,000 or using all samples. Phasing performance is based on cosmopolitan P_{trans} estimates and is calculated across trio samples from all populations. **b,** Phasing performance as a function of variant AF for the more common variant in a variant pair. Phasing performance is based on population-specific P_{trans} estimates and is calculated across trio samples from all populations. **c,** Proportion of variants falling into different AF bins when subsampling NFE gnomAD trios from 2815 trios down to 282, 563, or 1408 trios. Allele frequencies reflect the rarer variant in a variant pair. **d,** Phasing performance when subsampling NFE gnomAD samples as described in **c.**



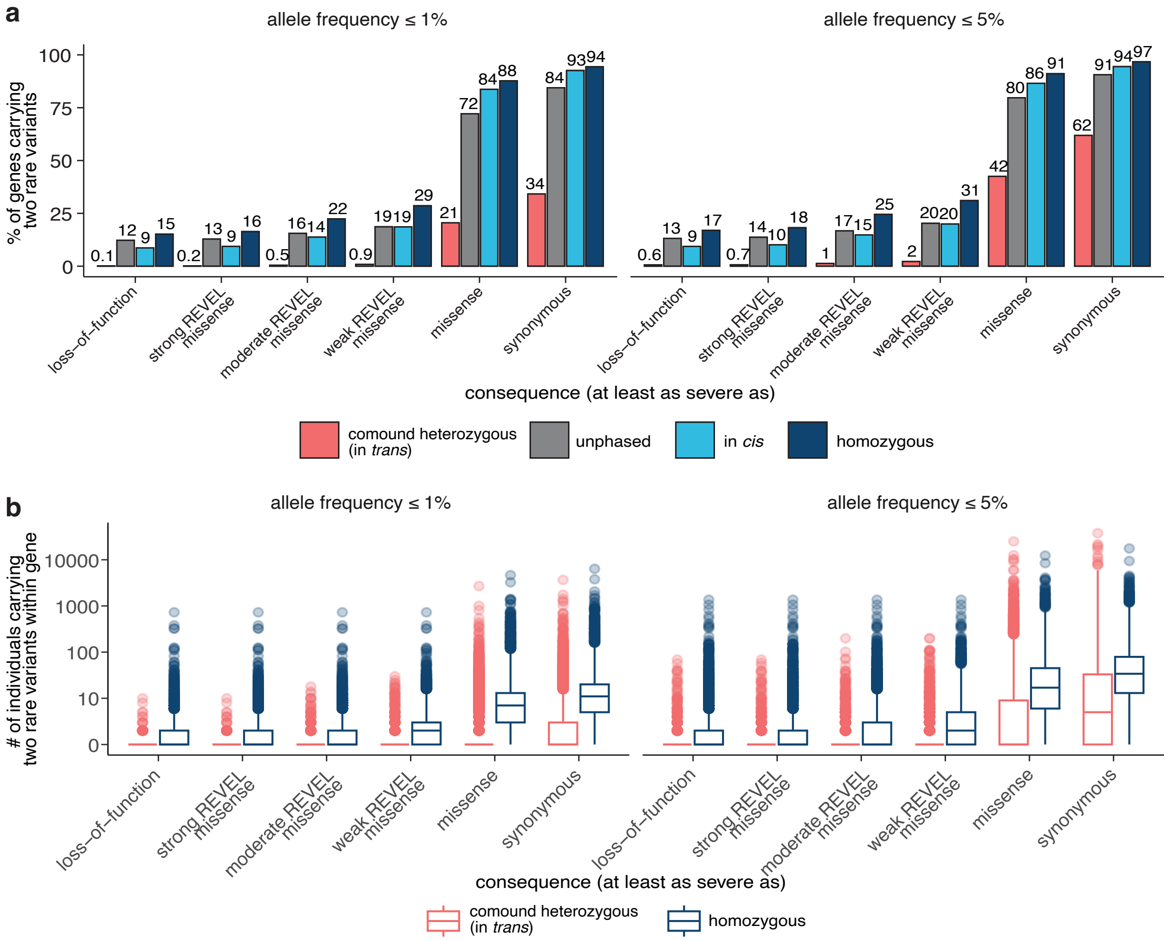
Supplementary Figure 7

Phasing performance for population-specific versus cosmopolitan P_{trans} scores for each population. AFR = African/African American; AMR = Admixed American/Latino; ASJ = Ashkenazi Jewish; EAS = East Asian; FIN = Finnish; NFE = non-Finnish European; SAS = South Asian.



Supplementary Figure 8

Phasing accuracy for transversions, non-CpG transitions, and CpG transitions. CpG transitions are further stratified by degree of DNA methylation (low, medium, or high) as in Karczewski et al[14]. Shading of squares and numbers in each square represents phasing accuracy. Phasing accuracies are based on variant pairs seen in all populations and utilize population-specific P_{trans} estimates. Accuracy is shown for all variants **(a)**, variants in *trans* **(b),** and variants in *cis* **(c).**



Supplementary Figure 9

a, Proportion of genes with one or more individuals in gnomAD carrying two rare variants at ≤ 1% and ≤ 5% AF stratified by predicted functional effect and phase. For compound heterozygous (comp het, in *trans*), unphased, and in *cis*, both variants in the variant pair must be annotated with a consequence at least as severe as the consequence displayed. **b,** Number of individuals in gnomAD (total sample size = 125,748 exomes) carrying two rare variants within gene at ≤ 1% and ≤ 5% AF, stratified by predicted functional effect and phase. For compound heterozygous (in *trans*) both variants in the variant pair must be annotated with a consequence at least as severe as the consequence displayed. In the box plots, the center line is the median, the box limits are the upper and lower quartiles, and the whiskers extend to the 1.5x the interquartile range. Any points shown are outliers.

Supplementary Tables

Supplementary Table 1. CMG diagnostic variants. In this table, we provide details about the presumed bi-allelic causal variants from 293 individuals from the Broad Institute Center for Mendelian Genetics. For each variant pair, we provide the gene symbol (“gene_name”), information about the position and alleles of both variants, whether both of the variants were singletons in gnomAD (“singleton_singleton”) and seen in the same individual or not, the estimated cosmopolitan P_{trans} value, the predicted phase based on the cosmopolitan P_{trans} value (“cosmopolitan_phase_prediction”), the imputed population ancestry of the CMG individual (“imputed_population_ancestry”), the predicted phased based on the population-specific P_{trans} value (“population_specific_phase_prediction”), the known phase from phase by transmission when trio data were available (“phase_by_transmission”), and an explanation for incorrect predictions where applicable (“incorrect_prediction_explanation”).

Supplementary Table 2. Manual curation results for compound heterozygous loss-of-function variants. Here, we provide the variant curation information for the 28 genes that have predicted compound heterozygous loss-of-function variants with AF ≤ 1%. For every predicted compound heterozygous variant pair, we provide the gene symbol, maximum AF in the gnomAD exomes from the two variants (“variant_pair_max_af”), the number of individuals who carry the variant pair (“n_individuals”), information about the position and alleles of both variants, any manual curation flags e.g., mapping error for the variants, and the final loss-of-function curation for both variants as well as the variant pair (“high_confidence_human_knock_out”).

Additional references in Supplementary Note

38. Kaplanis, J. *et al.* Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. *Genome Res.* **29**, 1047–1056 (2019).
39. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10**, e1004528 (2014).
40. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).