

Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits

Received: 10 October 2021

Accepted: 22 March 2023

Published online: 01 May 2023

Brian C. Zhang¹, Arjun Biddanda¹, Árni Freyr Gunnarsson^{1,2}, Fergus Cooper³ & Pier Francesco Palamara^{1,2}  

Genome-wide genealogies compactly represent the evolutionary history of a set of genomes and inferring them from genetic data has the potential to facilitate a wide range of analyses. We introduce a method, ARG-Needle, for accurately inferring biobank-scale genealogies from sequencing or genotyping array data, as well as strategies to utilize genealogies to perform association and other complex trait analyses. We use these methods to build genome-wide genealogies using genotyping data for 337,464 UK Biobank individuals and test for association across seven complex traits. Genealogy-based association detects more rare and ultra-rare signals ($N = 134$, frequency range 0.0007–0.1%) than genotype imputation using ~65,000 sequenced haplotypes ($N = 64$). In a subset of 138,039 exome sequencing samples, these associations strongly tag (average $r = 0.72$) underlying sequencing variants enriched (4.8×) for loss-of-function variation. These results demonstrate that inferred genome-wide genealogies may be leveraged in the analysis of complex traits, complementing approaches that require the availability of large, population-specific sequencing panels.

Modeling genealogical relationships between individuals plays a key role in a wide range of analyses, including the study of natural selection¹ and demographic history², genotype phasing³ and imputation⁴. Due to the very large number of genealogical relationships that may give rise to observed genomic variation, data-driven inference of these relationships is computationally difficult⁵. For this reason, available methods for the inference of genealogies rely on strategies that trade model simplification for computational scalability, such as the use of approximate probabilistic models^{6–11}, scalable heuristics^{12–16} or combinations of the two^{17,18}. Recent advances enabled efficient estimation of the genealogical distance between genomic regions from ascertained genotype data¹¹, rapid genealogical approximations for hundreds of thousands of samples¹⁵ and improved scalability of probabilistic inference¹⁷. However, available methods do not simultaneously offer all these features, so that scalable and accurate genealogical inference

in modern biobanks remains challenging. In addition, these datasets contain extensive phenotypic information, but applications of inferred genealogies have primarily focused on evolutionary analyses. Early work suggested that genealogical data may be used to improve association and fine-mapping^{13,19}, but the connections between genealogical inference and modern methodology for complex trait analysis^{20–22} remain under-explored.

We introduce a new algorithm, ARG-Needle, to accurately infer the ancestral recombination graph²³ (ARG) for large collections of genotyping or sequencing samples. We demonstrate that the ARG of a sample may be used within a linear mixed model (LMM) framework to increase association power, detect association to unobserved genomic variants, infer narrow sense heritability and perform polygenic prediction. Using ARG-Needle, we infer the ARG for 337,464 UK Biobank samples and perform a genealogy-wide association scan for seven complex traits.

¹Department of Statistics, University of Oxford, Oxford, UK. ²Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. ³Department of Computer Science, University of Oxford, Oxford, UK. ✉e-mail: palamara@stats.ox.ac.uk

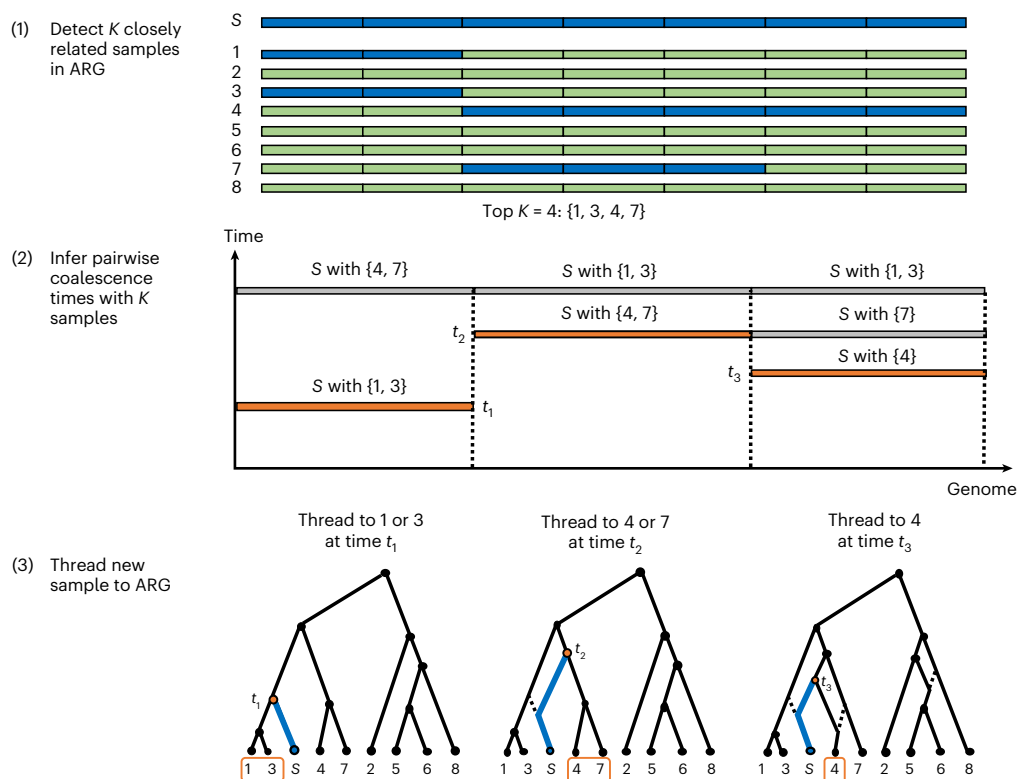


Fig. 1 | Overview of the ARG-Needle algorithm. ARG-Needle adds one haploid individual sample at a time to an existing ARG, each time performing three steps: (1) shortlisting a subset of most related samples already in the ARG through genotype hashing, (2) obtaining pairwise coalescence time estimates with these samples using ASMC¹¹ and (3) using the ASMC output to ‘thread’ the new sample to the ARG. We depict an example of adding sample S to an ARG, focusing on one genomic region. Step 1 divides the genome into ‘words’ and checks for identical matches with sample S . Based on these matches (shown in blue), samples 1, 3, 4 and 7 are output as the $K = 4$ candidate most related samples already in the ARG.

Step 2 computes pairwise coalescence time estimates between sample S and each of the samples 1, 3, 4 and 7. The minimum time for each position is highlighted. Step 3 uses these minimum times and samples to define a ‘threading instruction’ that is performed to add sample S to the ARG. Threading connects the new sample to the ancestral lineage of each chosen sample at the chosen time. Dotted lines indicate previous ARG edges that are inactive due to recombination. When all samples have been threaded, ARG-Needle performs a final postprocessing step called ARG normalization (Methods).

We show that despite being inferred using only array data, the ARG detects more independent associations to rare and ultra-rare variants (minor allele frequency (MAF) $< 0.1\%$) than imputation based on a reference panel of ~65,000 sequenced haplotypes of matched ancestry. We use 138,039 exome sequencing samples to confirm that these signals correspond to unobserved sequencing variants, which are strongly enriched for loss-of-function and other protein-altering variation and overlap with likely causal associations detected using within-cohort exome sequencing imputation. Using the ARG, we detect associations to variants as rare as $MAF \approx 4 \times 10^{-6}$ and independent higher frequency variation that is not captured using imputation.

Results

Overview of the ARG-Needle method

The ARG is a graph in which nodes represent the genomes of individuals or their ancestors and edges represent genealogical connections (see Supplementary Note 1 for additional details). ARG-Needle infers the ARG for large genotyping array or sequencing datasets by iteratively ‘threading’ one haploid sample at a time, as depicted in Fig. 1. Given an existing ARG, initialized to contain a single sample, we randomly select the next sample to be added (or threaded). We then compute a ‘threading instruction’, which at each genomic position provides the index of a sample in the ARG that is most closely related to the target sample, as well as their time to most recent common ancestor (TMRCA). We use this instruction to thread the target sample to the current ARG, and iterate until all samples have been included.

To compute the threading instruction of a sample, ARG-Needle first performs genotype hashing^{24,25} to rapidly detect a subset of candidate closest relatives within the ARG. It then uses the Ascertained Sequentially Markovian Coalescent (ASMC) algorithm¹¹ to estimate the TMRCA between the new sample and each of these individuals at each site, threading to the closest individual. When all samples have been included, ARG-Needle uses a fast postprocessing step, which we call ARG normalization, to refine the estimated node times. ARG-Needle builds the ARG in time approximately linear in sample size (see below).

We also introduce a simple extension of ASMC¹¹, called ASMC-clust, that builds genome-wide genealogies by forming a tree at each site using hierarchical clustering on pairwise TMRCAs output by ASMC. This approach scales quadratically with sample size but yields improved accuracy compared with ARG-Needle in certain simulated scenarios (see below). ARG-Needle and ASMC-clust efficiently represent and store ARGs using a graph data structure, which is an adaptation of the representation used within the ARGON simulator²⁶. Additional details, theoretical guarantees and properties for the ARG-Needle and ASMC-clust algorithms are described in the Methods and Supplementary Note 1.

Accuracy of ARG reconstruction in simulated data

We used extensive simulations to compare the accuracy and scalability of ARG-Needle, ASMC-clust, Relate¹⁷, tsinfer and a variant of tsinfer designed for sparse datasets we refer to as ‘tsinfer-sparse’¹⁵. We considered several metrics to compare ARGs, including: the Robinson–Foulds

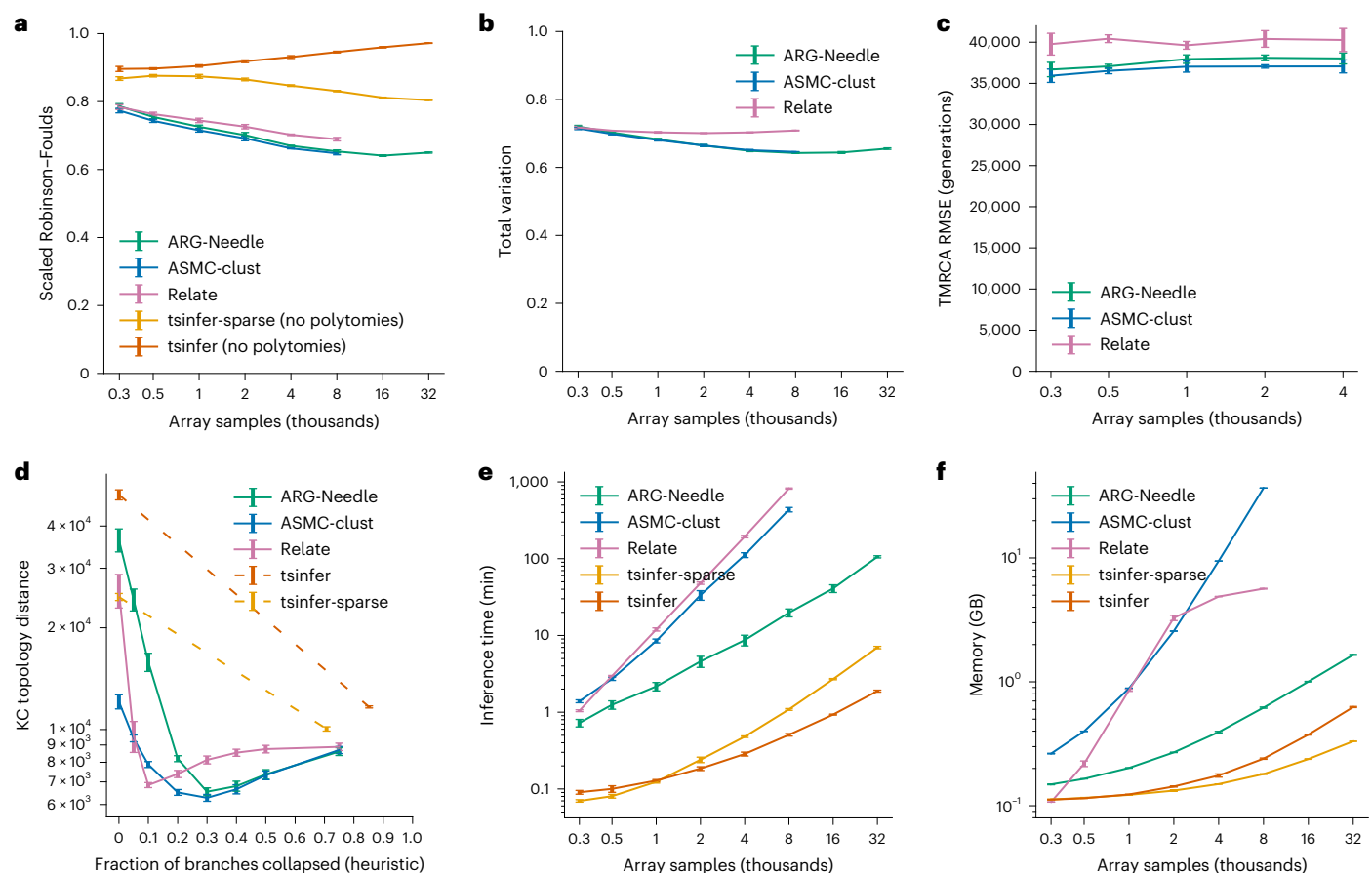


Fig. 2 | Comparison of ARG inference algorithms in simulation. a–f, We benchmark ARG inference performance for ARG-Needle, ASMC-clust, Relate, tsinfer and a variation of tsinfer for sparse data ('tsinfer-sparse') in realistic CEU demography array data simulations across a variety of metrics related to accuracy and computational resources (lower values indicate better performance for all metrics). **a**, The Robinson–Foulds distance (polytomies are randomly resolved). **b**, The ARG total variation distance (Methods). **c**, Pairwise TMRCA RMSE. **d**, The KC topology-only metric. **e**, Runtime. **f**, Peak memory. In **c**, we only run up to $N = 4,000$ haploid samples. In **d**, we fix $N = 4,000$ haploid samples and vary the fraction of branches per marginal tree that are collapsed

to form polytomies, using a heuristic that preferentially collapses branches that are less confidently inferred (Methods). For tsinfer and tsinfer-sparse, we instead rely on the default amount of polytomies in the output, additionally showcasing when polytomies are randomly resolved (dashed lines indicate a linear trend that may not hold). All panels use five random seeds, with ASMC-clust and Relate capped at $N = 8,000$ haploid samples due to runtime or memory constraints. Data are presented as mean values ± 2 s.e.m. Relate's default settings cap the memory for intermediate computations at 5 GB (see **f**). ARG-Needle and ASMC-clust include ARG normalization by default (Methods), while Relate does not. For additional simulations, see Extended Data Figs. 1–4 and Supplementary Figs. 1–6.

distance²⁷, which reflects dissimilarities between the mutations that may be generated by two ARGs; the root mean squared error (RMSE) between true and inferred pairwise TMRCAs, which captures the accuracy in predicting allele sharing between individuals; and the Kendall–Colijn (KC) topology-only distance²⁸. We found that the KC distance is systematically lower for trees containing polytomies (that is, nodes with more than two children), which are not output by Relate, ASMC-clust or ARG-Needle (Extended Data Fig. 1b,c). We therefore applied a heuristic to allow these methods to output polytomies (see the Methods and Supplementary Note 2 for additional discussion). Although these three metrics capture similarity between marginal trees, they are not specifically developed for comparing ARGs. We therefore developed an additional metric, called the ARG total variation distance, which generalizes the Robinson–Foulds distance to better capture the ability of a reconstructed ARG to predict mutation patterns that may be generated by the true underlying ARG (see the Methods and Supplementary Note 2 for further details).

We measured ARG reconstruction accuracy in synthetic array datasets of up to 32,000 haploid samples (Fig. 2 and Methods). We also tested a variety of additional conditions, including different demographic histories, varying recombination rates and genotyping error.

We also examined the effects of ARG normalization, of variations of the KC distance that account for branch lengths and of stratifying the total variation distance by allele frequency (Extended Data Figs. 1 and 2 and Supplementary Figs. 1–4). ARG-Needle tended to achieve best performance across all accuracy metrics in array data, sometimes tied or in close performance with ASMC-clust or Relate. In simulations of sequencing data, ASMC-clust performed best on the ARG total variation and TMRCA RMSE metrics, with ARG-Needle and Relate close in performance, while Relate performed best on the Robinson–Foulds metric (Extended Data Fig. 3). We next measured the speed and memory footprint of these methods. ARG-Needle requires lower computation and memory than Relate and ASMC-clust, which both scale quadratically with sample size (Fig. 2e,f and Extended Data Fig. 1e). It runs slower than tsinfer and tsinfer-sparse but with a similar (approximately linear) scaling (also see the Methods and Supplementary Note 1).

We next examined additional properties of the ARG-Needle and ASMC-clust algorithms. We found that the order used to thread samples into the ARG does not substantially affect accuracy (Supplementary Fig. 5a–d), but that averaging estimates obtained using different random threading orders may produce improved estimates of genealogical relationships and higher similarity to ARGs inferred

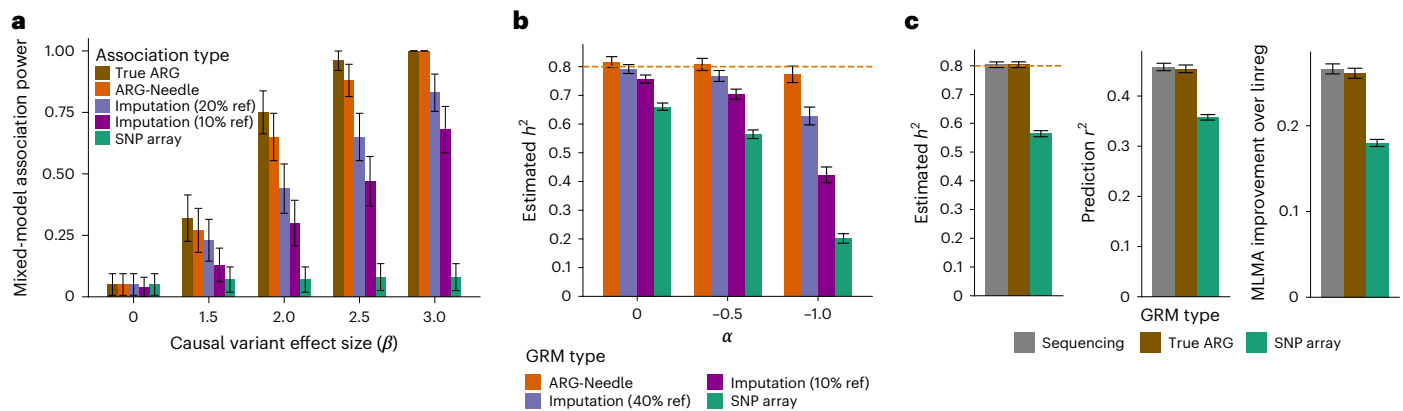


Fig. 3 | ARG-based analysis of simulated complex traits. **a**, Power to detect a rare causal variant (MAF = 0.025%) in simulations of a polygenic phenotype. We compare ARG-MLMA of ground-truth ARGs and ARG-Needle-inferred ARGs with MLMA of imputed and SNP array variants as we vary the effect size β (100 independent simulations of $h^2 = 0.8$, $\alpha = -0.25$, $N = 20,000$ haploid samples and 22 chromosomes of 5 Mb each; Methods). **b**, Heritability estimation using ARG-GRMs from ARG-Needle inference on SNP array data, compared with using imputed or array SNPs (5 simulations of 25 Mb, $N = 5,000$ haploid samples, $h^2 = 0.8$ and varying α). **c**, ARG-GRMs computed using ground-truth ARGs perform equivalently to GRMs computed using sequencing data in heritability estimation, polygenic prediction and mixed-model association

($N = 10,000$ haploid samples, $h^2 = 0.8$ and $\alpha = -0.5$). Heritability and prediction involve 5 simulations of 50 Mb, and association involves 50 simulations of 22 chromosomes of 2.5 Mb each, for a total of 55 Mb. For association, we show the relative improvement in mean $-\log_{10}(P)$ of MLMA compared with linear regression (Methods). “% ref” indicates the size of the reference panel used for imputation as a percentage of the number of haploid samples ($N = 20,000$ in **a**, $N = 5,000$ in **b**). Data are presented as estimates ± 2 s.e.m., where the estimates are from meta-analysis in the case of heritability estimation, represent fractions in **a** and represent means otherwise. Additional results are shown in Extended Data Figs. 6–8 and Supplementary Fig. 8. linreg, linear regression.

using ASMC-clust (Supplementary Fig. 5c–f). We observed that inferred genealogies contain realistic linkage disequilibrium (LD) patterns (Extended Data Fig. 4a,b). ARG-Needle, ASMC-clust and Relate do not guarantee that the variants used to infer genealogies may be mapped to inferred marginal trees, but performed well when we considered the fraction of unobserved variants that could be mapped back to inferred genealogies (Extended Data Fig. 4c; also see ref. 17). Finally, we assessed the similarity of ARGs inferred using different algorithms, observing highest similarities between ASMC-clust and ARG-Needle, as well as between these methods and, in decreasing order, Relate, tsinfer-sparse and tsinfer (Supplementary Fig. 6a,b).

A genealogical approach to LMM analysis

LMMs enable state-of-the-art analysis of polygenic traits^{20,29,30,31}. We developed an approach that uses the ARG of a set of genomes to perform mixed linear model association (MLMA²⁹; Methods). More in detail, we use an ARG built from genotyping array data to infer the presence of unobserved variants and perform MLMA testing of these variants. This increases association power in two ways: the ARG is used to uncover putatively associated variants, while the LMM utilizes estimates of genomic similarity to model polygenicity, relatedness and population stratification²⁹. We refer to association analyses that test variants in the ARG as ‘genealogy-wide association’ scans and, more specifically, to analyses that incorporate mixed linear model testing as ARG-MLMA. Genealogy-wide association complements genotype imputation based on a sequenced reference panel, as it enables capturing rare variants in the sample that may be absent from the panel or cannot be accurately imputed (Extended Data Fig. 5a). It also generalizes rare variant association strategies based on haplotype sharing^{13,19,25,32–35}, as detailed in Supplementary Fig. 7. In simulations, we observed that for low-frequency variants genealogy-wide association may achieve higher association power than testing of variants imputed from a sequenced reference panel (Fig. 3a and Extended Data Fig. 6).

In addition, we developed strategies to leverage the ARG to obtain estimates of genomic similarity across individuals, which are aggregated in a genomic relatedness matrix (GRM; Methods) and are a key element of several mixed-model analyses of complex traits. We refer

to GRMs built using this approach as ARG-GRMs and provide details of their construction and properties in Supplementary Note 3. We used ARG-GRMs to measure the amount of phenotypic variance captured by inferred ARGs (Extended Data Fig. 7a). In simulations, ARG-GRMs built using ARGs inferred by ARG-Needle in array data captured more narrow sense heritability than GRMs built using array data^{30,36,37} (Methods, Fig. 3b and Supplementary Fig. 8). We also performed additional simulations to test whether the modeling of unobserved genomic variation using ARG-GRMs may be leveraged to obtain performance gains in other LMM analyses. Indeed, ARG-GRMs built using true ARGs performed as well as GRMs computed using sequencing data in LMM-based heritability estimation, polygenic prediction and association (Methods, Fig. 3c and Extended Data Fig. 8). Applying these strategies to large-scale inferred ARGs, however, will require improved accuracy and scalability (Discussion).

Overall, these experiments suggest that accurate genealogical inference combined with LMMs improves association power, by testing variants that are not well tagged using available markers while modeling polygenicity. The ARG may also be potentially utilized to obtain improved estimates of genomic similarity and perform additional LMM-based complex trait analyses.

Genealogy-wide association scan of rare and ultra-rare variants in the UK Biobank

We applied ARG inference methods in a subset of the genome using UK Biobank data and observed results consistent with our simulations (Supplementary Fig. 6c,d). We then used ARG-Needle to build the genome-wide ARG from SNP array data for 337,464 individuals in the white British ancestry subset defined by ref. 38 (Methods). We performed ARG-MLMA for height and six molecular traits, comprising alkaline phosphatase, aspartate aminotransferase, low-density lipoprotein (LDL) / high-density lipoprotein (HDL) cholesterol, mean platelet volume and total bilirubin. To achieve the required scalability, we built on a recent MLMA method^{22,39}, implicitly relying on an array-based GRM (Methods and Discussion). We compared ARG-MLMA with standard MLMA testing of variants imputed using the Haplo-type Reference Consortium (HRC) and UK10K reference panels^{38,40,41}

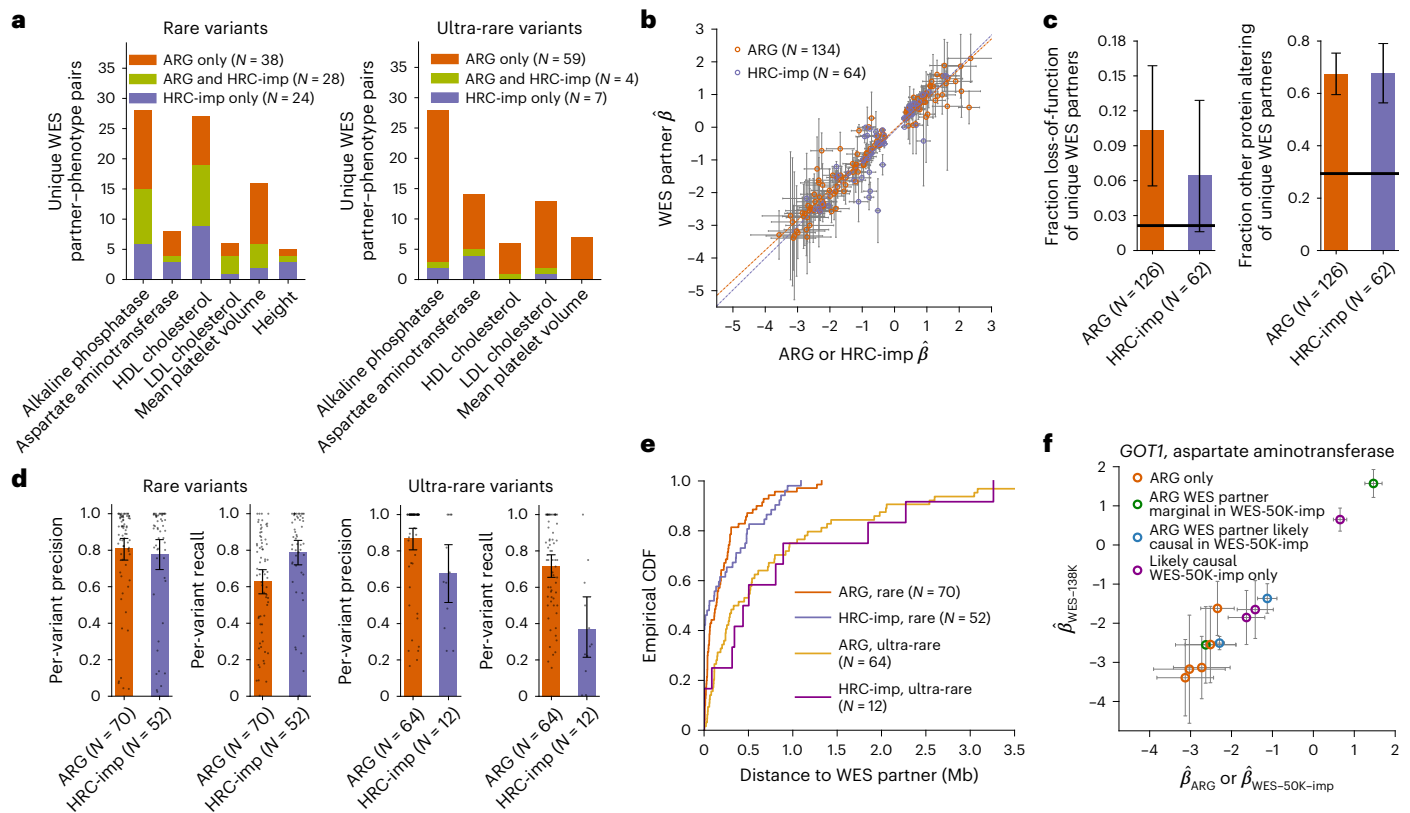


Fig. 4 | Association of ARG-derived and imputed rare and ultra-rare variants with seven quantitative traits in UK Biobank. **a**, Counts of unique WES partners for ARG and HRC + UK10K-imputed ('HRC-imp') independent associations, partitioned by traits and frequency and showing overlap. Total bilirubin was not associated at these frequencies and height was not associated for ultra-rare variants. **b**, Scatter plot of $\hat{\beta}$ (estimated effect) for independent variants (estimated within 337,464 samples) against $\hat{\beta}$ for their WES partners (estimated within 138,039 samples), with linear model fit. **c**, Fraction of loss-of-function and other protein-altering variants for the unique WES partners of independent variants (125 WES partners for ARG and 62 for imputed variants). Horizontal black lines represent averages across exome sequencing variants. **d**, Average per-variant precision and recall of predicting WES carrier status, partitioned by frequency and showing individual value as jittered points (71 rare ARG, 53 rare

imputed, 62 ultra-rare ARG and 12 ultra-rare imputed variants). **e**, Cumulative distribution function (CDF) for the distance between independent variants and their WES partners. **f**, Scatter plot of $\hat{\beta}$ for ARG-derived independent variants associated with aspartate aminotransferase in the *GOT1* gene (estimated within 337,464 samples) against $\hat{\beta}$ for their WES partners (estimated within 138,039 samples). We color points based on whether the WES partner is likely causal in WES-50K-imp (imputation from WES-50K into ~459,000 samples⁴³), not likely causal but marginally significant in WES-50K-imp or not marginally significant in WES-50K-imp ('ARG only'). We also plot the $\hat{\beta}$ for the additional likely causal variants in WES-50K-imp against the $\hat{\beta}$ in WES-138K. Bars represent fractions in **c** and means in **d**. Error bars represent 1.96 s.e.m. in **b** and **f** and represent bootstrap 95% CIs in **c** and **d**. Additional results are shown in Extended Data Fig. 9. HDL, high-density lipoprotein; LDL, low-density lipoprotein.

(hereafter, HRC + UK10K), comprising ~65,000 haplotypes. We focused on rare ($0.01\% \leq \text{MAF} < 0.1\%$) and ultra-rare ($\text{MAF} < 0.01\%$) genomic variants. We used resampling-based testing⁴² to establish genome-wide significance thresholds of $P < 4.8 \times 10^{-11}$ for ARG variants (sampled with mutation rate $\mu = 10^{-5}$) and $P < 1.06 \times 10^{-9}$ for imputed variants (Supplementary Table 1). For each analysis, we performed LD-based filtering to extract a stringent set of approximately independent associations (hereafter, 'independent associations'; Methods). We leveraged a subset of 138,039 individuals with whole-exome sequencing (WES) data (hereafter, WES-138K) to validate these independent associations. For each detected independent variant, we selected the WES variant with the largest correlation, which we call its 'WES partner'.

Applying this approach, we detected 134 independent signals using the ARG and 64 using imputation, jointly implicating 152 unique WES partners (Supplementary Tables 2 and 3). Of these WES variants, 36 were implicated using both approaches (Fig. 4a, and see Extended Data Fig. 9a for region-level results). The fraction of WES partners uniquely identified using the ARG was larger among ultra-rare variants (84%) compared with rare variants (42%), reflecting a scarcity of ultra-rare variants in the HRC + UK10K imputation panel. The phenotypic effects estimated in the 337,464 individuals using ARG-derived

or imputed associations were strongly correlated to those directly estimated for the WES partners in the WES-138K dataset (Fig. 4b), with stronger average correlation (bootstrap $P = 0.003$) for ARG-derived variants ($r^2_{\text{ARG}} = 0.93$) compared with imputed variants ($r^2_{\text{imp}} = 0.80$). Only 74% of the WES partners for ARG-derived rare variant associations were significantly associated ($P < 5 \times 10^{-8}$) in the smaller WES-138K dataset, a proportion that dropped to 59% for ultra-rare variants. Variants detected using genealogy-wide association had a larger average phenotypic effect than those detected via imputation (bootstrap $P < 0.0001$; average $|\hat{\beta}_{\text{ARG}}| = 1.46$; average $|\hat{\beta}_{\text{imp}}| = 0.90$), reflecting lower average frequencies. In addition, WES partners of ARG-derived variants were $\sim 4.8\times$ enriched for loss-of-function variation (bootstrap $P < 0.001$; Fig. 4c), and WES partners implicated by either ARG or imputation were $\sim 2.3\times$ enriched for other protein-altering variation (Methods), supporting their likely causal role.

We also used variant-level precision and recall statistics (Methods) to measure the extent to which carrying an associated ARG-derived or imputed variant is predictive of carrying sequence-level WES partner variants (Fig. 4d). ARG-derived and imputed rare variants had similar levels of variant-level precision, while imputation had higher recall (bootstrap $P = 0.0005$). For ultra-rare variants, ARG-derived signals

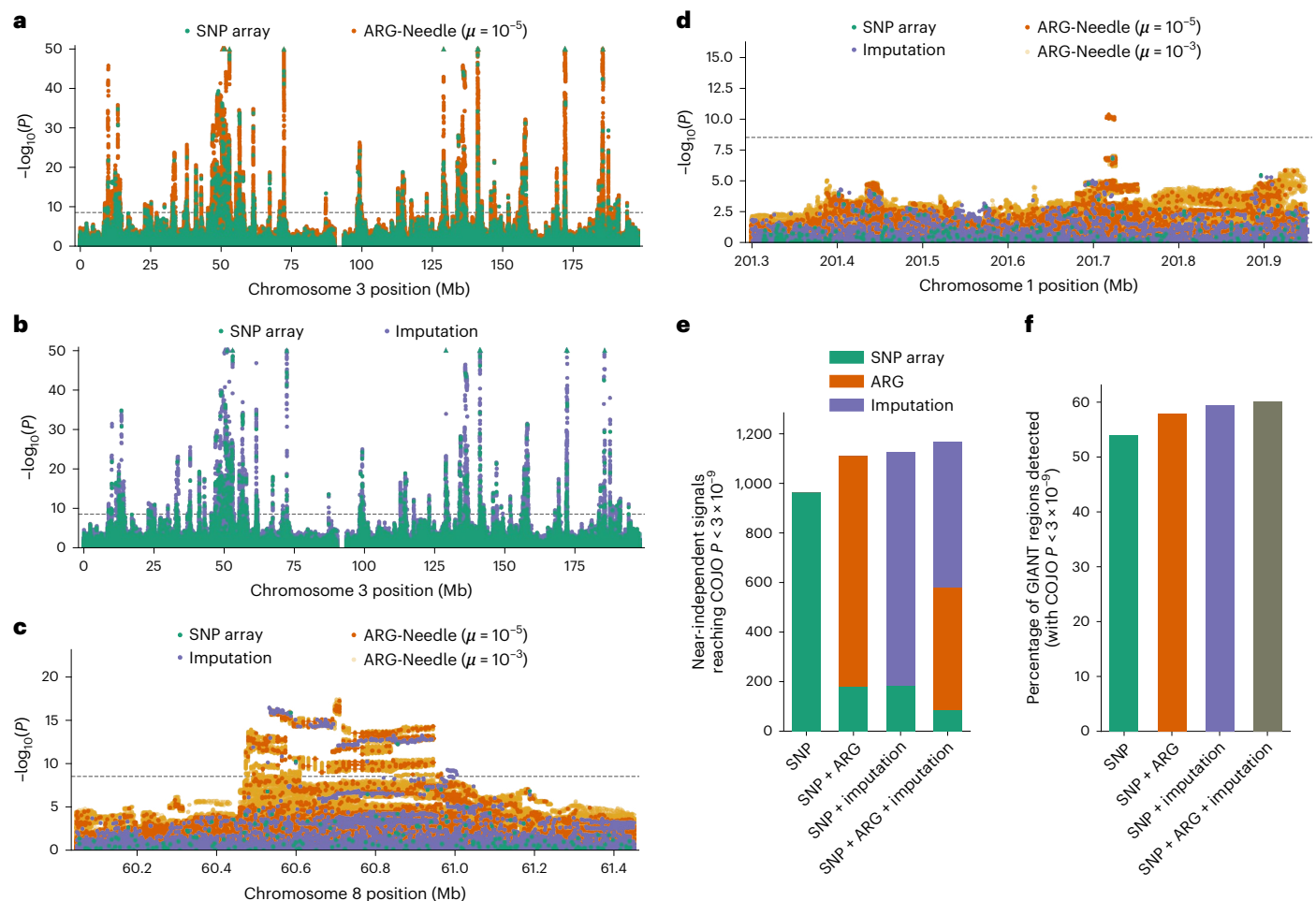


Fig. 5 | Genealogy-wide association of higher frequency variants with height in UK Biobank. a, b, Chromosome 3 Manhattan plots showing MLMA of ARG-Needle on SNP array data versus array SNPs (**a**) and HRC + UK10K-imputed variants versus array SNPs (**b**). **c, d,** Manhattan plots of two loci. **c,** ARG-MLMA detects haplotype structure that is found using imputation, with a different association peak. **d,** An association peak found by ARG-MLMA that was significant ($P < 3 \times 10^{-9}$) in a GIANT consortium meta-analysis of ~700,000 samples. **e, f,** Approximately independent associations (defined as having $\text{COJO } P < 3 \times 10^{-9}$; Methods) when considering array SNPs alone, array SNPs and

ARG-Needle variants, array SNPs and imputed variants, and all three types of variants. **e,** Total number of independent variants found and attribution based on data type. **f,** Percentage of 1-Mb regions containing COJO associations in the GIANT meta-analysis that are detected using each method. For the Manhattan plots, the order of plotting is ARG-Needle with $\mu = 10^{-3}$ (used for follow-up), then ARG-Needle with $\mu = 10^{-5}$ (used for discovery), then imputation, then SNP array variants on top. Dotted lines correspond to $P = 3 \times 10^{-9}$ (Methods) and triangles indicate associations with $P < 10^{-50}$. See also Supplementary Figs. 9 and 10.

performed better than imputed variants for both precision (bootstrap $P = 0.01$) and recall (bootstrap $P = 0.002$). Similarly, ARG-derived and imputed rare variants provided comparable tagging for their WES partners (Extended Data Fig. 9b), while ARG-derived ultra-rare variants provided stronger tagging compared with imputed ultra-rare variants (average $r_{\text{ARG}} = 0.77$, $r_{\text{imp}} = 0.42$, bootstrap $P < 0.001$). Compared with ARG-derived variants, genotype imputation has the advantage that associated variants that are sequenced in the reference panel may be directly localized in the genome. We found that for 21 of 52 rare and 2 of 12 ultra-rare independent imputation signals the WES partner had been imputed, while the remaining signals likely provide indirect tagging for underlying variants. ARG-derived and imputed variants, however, had similar distributions for the distance to their WES partners (Fig. 4e and Extended Data Fig. 9c). This suggests that genealogy-wide associations have the same spatial resolution as associations obtained using genotype imputation in cases where the variant driving the signal cannot be directly imputed.

We compared our results with those of a recent study that leveraged exome sequencing data from a subset of ~50,000 participants (hereafter, WES-50K) to perform genotype imputation for ~459,000

samples⁴³. We found that, among the WES partners implicated using the ARG but not using HRC + UK10K imputation, 14 of 30 partners of rare and 26 of 55 partners of ultra-rare ARG variants were also flagged as likely causal associations ($P < 5 \times 10^{-8}$) in ref. 43 (Supplementary Table 2). The remaining 45 WES partners detected using the ARG but not reported in ref. 43 are often very rare variants (median $\text{MAF} = 3.6 \times 10^{-5}$; Extended Data Fig. 9d) of large effect (median $|\beta| = 1.14$), which are difficult to impute; 21 of 45 such variants were absent or singletons in the WES-50K reference panel or had poor imputation quality score. Associations uniquely detected using the ARG often extended allelic series at known genes. For instance, restricting to loss-of-function or other protein-altering WES partners for independent ARG signals not present or marginally significant in ref. 43, five novel associations with aspartate aminotransferase are mapped to the *GOT1* gene (Fig. 4f) and four with alkaline phosphatase are mapped to *ALPL* (Extended Data Fig. 9e). A subset of strong independent associations uniquely detected by the ARG had weak correlation with their WES partners, possibly due to tagging of structural or regulatory variation absent from the WES-138K dataset (for example, a signal for aspartate aminotransferase with

$P = 7.4 \times 10^{-39}$, $MAF_{ARG} = 0.0005$, WES partner $r = 0.21$, minor allele count (MAC)_{WES-138K} = 6, MAC_{WES-50K} = 1).

In summary, genealogy-wide association using an ARG inferred from common SNPs revealed more rare and ultra-rare signals than genotype imputation based on ~65,000 reference haplotypes, and detected ultra-rare variants that were not associated using within-cohort imputation based on ~50,000 exome-sequenced participants. ARG-derived associations accurately predicted effect sizes for underlying sequencing variants, as well as the subset of carrier individuals.

Genealogy-wide association for low- and high-frequency variants

Lastly, we performed genealogy-wide association for low- ($0.1\% \leq MAF < 1\%$) and high- ($MAF \geq 1\%$) frequency variants, which are more easily imputed using reference panels that are not necessarily large and population-specific. Consistent with this, extending our previous analysis to low-frequency variants yielded a similar number of independent associations for ARG-derived and HRC + UK10K-imputed variants ($N_{ARG} = 103$, $N_{imp} = 100$; Supplementary Tables 4 and 5 and Extended Data Fig. 10a–c). Associations detected using the ARG had slightly larger effects compared with those found using imputation (bootstrap $P = 0.026$; average $|\beta_{ARG}| = 0.32$, $|\beta_{imp}| = 0.27$) but provided lower tagging to WES partners (bootstrap $P < 0.001$; average $r_{ARG} = 0.57$, $r_{imp} = 0.73$), reflecting the large fraction (42 of 100) of imputation WES partners that were directly imputed.

We hypothesized that, although imputation of higher frequency variants is generally more accurate, branches in the marginal trees of the ARG may in some cases complement available markers by providing improved tagging of underlying variation. This may be the case, for instance, for short insertions/deletions or structural variants⁴⁴, which are often underrepresented in reference panels⁴¹, or for variants of moderately high frequency, which may be difficult to impute⁴⁵ (Extended Data Fig. 5a). To test this, we performed MLMA for height using HRC + UK10K-imputed variants, filtered as in ref. 38 ($MAF > 0.1\%$, info score > 0.3 ; Methods), for which we established a resampling-based genome-wide significance threshold of 4.5×10^{-9} (95% confidence interval (95% CI): 2.2×10^{-9} , 9.6×10^{-9}). To facilitate direct comparison, we selected ARG-MLMA parameters ($MAF > 1\%$, $\mu = 10^{-5}$; Methods) corresponding to a higher MAF cutoff but a comparable genome-wide significance threshold of 3.4×10^{-9} (95% CI: 2.4×10^{-9} , 5×10^{-9}) and adopted a threshold of 3×10^{-9} for all downstream analyses.

We first assessed the number of 1-megabase (Mb) regions that contain an association ($P < 3 \times 10^{-9}$) for genotype array, imputed or ARG-derived variants. We found that ARG-MLMA detected 98.9% of regions found by both SNP array and imputation, as well as 71% of regions found by imputation but not array data and an additional 8% of regions not found using either imputation or array data (Extended Data Fig. 10d). A significant fraction (54 of 92, permutation $P < 0.0001$) of regions identified using the ARG but not imputation contained associations ($P < 3 \times 10^{-9}$) in a larger meta-analysis by the Genetic Investigation of ANthropometric Traits (GIANT) consortium⁴⁶ ($N \approx 700,000$) comprising the UK Biobank and additional cohorts. Inspecting associated loci, we observed that ARG-MLMA captures association peaks and haplotype structure found using imputation but not array data (Fig. 5a–c and Supplementary Figs. 9 and 10a–e) as well as association peaks uniquely identified using ARG-MLMA (Fig. 5d and Supplementary Fig. 10f–h).

We sought to further assess the degree of overlap and complementarity of associations detected using SNP array data, imputation and the ARG, by performing LD-based filtering and conditional and joint (COJO⁴⁷) association analyses (Fig. 5e and Methods). When we jointly considered either or both ARG-derived and imputed variants in addition to array markers, we observed an increase in the number of approximately independent COJO associations ($P < 3 \times 10^{-9}$; $N_{SNP} = 964$, $N_{SNP+ARG} = 1,110$, $N_{SNP+imp} = 1,126$, $N_{SNP+ARG+imp} = 1,161$). The fraction of COJO-associated array markers was reduced by the inclusion

of ARG-derived or imputed variants, which resulted in comparable proportions of associations when jointly analyzed (Fig. 5e), suggesting that both ARG and imputation provide good tagging of underlying signal. By considering the set of 1-Mb regions harboring significant COJO associations, we verified that the additional COJO signals detected when including ARG-derived or imputed variants concentrated within regions that also harbor significant ($P < 3 \times 10^{-9}$) COJO signals in the GIANT meta-analysis⁴⁶ (Fig. 5f and Extended Data Fig. 10e).

In summary, genealogy-wide association using the ARG inferred by ARG-Needle from SNP array data was less effective for the analysis of higher frequency variants because these variants could be more accurately imputed compared to rare and ultra-rare variants. However, ARG-derived variants revealed associated peaks and haplotypes that were not found through association of array data alone and in some cases complemented genotype imputation in detecting approximately independent associations. We note that the choices of filtering criteria, such as MAF threshold, imputation info score and ARG mutation rate, all affect the sensitivity and specificity of these analyses. Results for an analysis restricting to association of variants with $MAF > 10\%$ are shown in Supplementary Fig. 11.

Discussion

We developed ARG-Needle, a method for accurately inferring genome-wide genealogies from genomic data that scales to large biobank datasets. We performed extensive simulations, showing that ARG-Needle is both accurate and scalable when applied to genotyping array and sequencing data. We also developed a framework that combines inferred genealogies with LMMs to increase association power, and showed that this strategy may be further leveraged in analyses of heritability and polygenic prediction. We built genome-wide ARGs from genotyping array data for 337,464 UK Biobank individuals and performed a genealogy-wide association scan for seven quantitative phenotypes. Using the inferred ARG, we detected more large-effect associations to rare and ultra-rare variants than using genotype imputation from ~65,000 sequenced haplotypes, down to an allele frequency of $\sim 4 \times 10^{-6}$. We validated these signals using exome sequencing, showing that they tag underlying variants enriched for loss-of-function and other protein-altering variation. Associations detected using the ARG overlap with and extend fine-mapped associations detected using within-cohort genotype imputation. Applied to the analysis of higher frequency variants, the ARG revealed haplotype structure and independent signals complementary to those obtained using imputation.

These results highlight the importance of genealogical modeling in the analysis of complex traits. Genome-wide association analyses rely on the correlation between available markers and underlying variation⁴⁸ and the MLMA framework accounts for polygenicity, relatedness and population structure²⁹. In genealogy-wide association, the signal of LD is amplified by further modeling of past recombination events to infer the presence of hidden genomic variation. Through ARG-GRMs, inferred genealogies may facilitate better modeling of genomic similarity and polygenic effects, leading to improved robustness and increased statistical power.

These analyses also demonstrate that genealogical inference provides a complementary strategy to genotype imputation approaches, which rely on haplotype sharing between the analyzed samples and a sequenced reference panel to extend the set of available markers. Imputation has been successfully applied in several complex trait analyses^{4,36}, but its efficacy for the study of rare variants hinges on the availability of large, population-specific sequencing panels, which are not widely available for all populations. Genealogy-wide association may therefore offer new avenues to better utilize genomic resources for underrepresented groups⁴⁹.

We highlight several limitations and directions of future development. First, although genealogy-wide association enables detecting individuals carrying associated variants, it may implicate large

genomic regions, whereas genotype imputation may associate individual variants if they are sequenced in the reference panel. When sequencing data are available, however, they may be utilized to further localize ARG-derived signals, for instance using WES partners. Second, although we have shown in simulation that ARG-GRMs built from true ARGs may be used to estimate heritability, perform prediction and increase association power, real data applications of this approach will require methodological improvements to increase LMM scalability^{50,51}. Third, although our study was restricted to unrelated samples of homogeneous ancestry, we expect genealogy-wide association to be as susceptible as standard association to issues such as relatedness and population stratification^{29,52,53}, requiring adequate control for these confounders. Fourth, although we have focused on leveraging an ARG inferred from array data alone, ARG-Needle enables building an ARG using a mixture of sequencing and array data. This approach may be used to perform additional analyses such as ARG-based genotype imputation, which is likely to improve upon approaches that do not model the TMRCA between target and reference samples⁵⁴. In simulations we performed, this ARG-based imputation strategy obtained promising results (Supplementary Note 4, Extended Data Fig. 5 and Supplementary Fig. 12). Fifth, our analyses were limited to quantitative traits; support for MLMA of rare case/control traits will require methodological extensions. Sixth, we adopted a computationally intensive resampling-based approach⁴² to establish significance thresholds across filtering parameters; future work may lead to improved strategies to address multiple testing. Seventh, although we relied on several existing and novel metrics to analyze properties of the reconstructed ARGs, further research should develop additional metrics and explore their properties and relationships to downstream analyses. These metrics should be applicable for benchmarking methods that only infer the topology of an ARG as well as methods that focus on estimating branch lengths⁵⁵. Eighth, reconstructing biobank-scale ARGs will likely aid the study of additional evolutionary properties of disease-associated variants, including analyses of natural selection acting on complex traits^{11,56,57} which we have not explored in this work. Finally, our analysis focused on the UK Biobank dataset, which provides an excellent testbed due to the large volumes of high-quality data of different types available for validation. Future applications of our methods will involve analysis of cohorts that are less strongly represented in current sequencing studies. Nevertheless, we believe that the results described in this work represent an advance in large-scale genealogical inference and provide new tools for the analysis of complex traits.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01379-x>.

References

- Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99–110 (2003).
- Beichman, A. C., Huerta-Sanchez, E. & Lohmueller, K. E. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annu. Rev. Ecol. Evol. Syst.* **49**, 433–456 (2018).
- Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
- Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
- McVean, G. A. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 1387–1393 (2005).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
- Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
- Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
- Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
- Palamara, P. F., Terhorst, J., Song, Y. S. & Price, A. L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* **50**, 1311–1317 (2018).
- Lyngsø, R. B., Song, Y. S. & Hein, J. Minimum recombination histories by branch and bound. in *International Workshop on Algorithms in Bioinformatics* (eds Casadio, R. & Myers, G.) 239–250 (Springer, 2005).
- Minichiello, M. J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
- Mirzaei, S. & Wu, Y. RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics* **33**, 1021–1030 (2017).
- Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
- Schaefer, N. K., Shapiro, B. & Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Sci. Adv.* **7**, eabc0776 (2021).
- Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
- Speidel, L. et al. Inferring population histories for ancient genomes using genome-wide genealogies. *Mol. Biol. Evol.* **38**, 3497–3511 (2021).
- Zöllner, S. & Pritchard, J. K. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092 (2005).
- Kang, H. M. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
- Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
- Griffiths, R. C. & Marjoram, P. An ancestral recombination graph. *Inst. Math. Appl.* **87**, 257 (1997).
- Gusev, A. et al. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
- Nait Saada, J. et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).
- Palamara, P. F. ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics* **32**, 3032–3034 (2016).
- Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).

28. Kendall, M. & Colijn, C. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.* **33**, 2735–2743 (2016).
29. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
30. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
31. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
32. Templeton, A. R., Crandall, K. A. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619–633 (1992).
33. Houwen, R. H. et al. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* **8**, 380–386 (1994).
34. Gusev, A. et al. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am. J. Hum. Genet.* **88**, 706–717 (2011).
35. Browning, S. R. & Thompson, E. A. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* **190**, 1521–1531 (2012).
36. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
37. Wainschein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
38. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
39. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
40. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
41. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
42. Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).
43. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
44. Mukamel, R. E. et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
45. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
46. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
47. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
48. Reich, D. E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
49. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
50. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385 (2015).
51. Pazokitoroudi, A. et al. Efficient variance components analysis across millions of genomes. *Nat. Commun.* **11**, 4020 (2020).
52. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
53. Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
54. Si, Y., Vanderwerff, B. & Zöllner, S. Why are rare variants hard to impute? Coalescent models reveal theoretical limits in existing algorithms. *Genetics* **217**, iyab011 (2021).
55. Wohns, A. W. et al. A unified genealogy of modern and ancient genomes. *Science* **375**, eabi8264 (2022).
56. Yasumizu, Y. et al. Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Mol. Biol. Evol.* **37**, 1306–1316 (2020).
57. Stern, A. J., Speidel, L., Zaitlen, N. A. & Nielsen, R. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* **108**, 219–239 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Methods

ARG-Needle and ASMC-clust algorithms

We introduce two algorithms to construct the ARG of a set of samples, called ARG-Needle and ASMC-clust. Both approaches leverage output from the ASMC algorithm¹¹, which takes as input a pair of genotyping array or sequencing samples and outputs a posterior distribution of the TMRCA across the genome. ARG-Needle and ASMC-clust use this pairwise genealogical information to assemble the ARG for all individuals.

ASMC-clust runs ASMC on all pairs of samples and performs hierarchical clustering of TMRCA matrices to obtain an ARG. At every site, we apply the unweighted pair group method with arithmetic mean (UPGMA) clustering algorithm⁵⁸ on the $N \times N$ posterior mean TMRCA matrix to yield a marginal tree. We combine these marginal trees into an ARG, using the midpoints between sites' physical positions to decide when one tree ends and the next begins. Using an $O(N^2)$ implementation of UPGMA^{59,60}, we achieve a runtime and memory complexity of $O(N^2M)$. We also implement an extension that achieves $O(NM)$ memory but increased runtime (Supplementary Note 1).

ARG-Needle starts with an empty ARG and repeats three steps to add additional samples to the ARG: (1) detecting a set of closest genetic relatives via hashing, (2) running ASMC and (3) 'threading' the new sample into the ARG (Fig. 1). Given a new sample, step 1 performs a series of hash table queries to determine the candidate closest samples already in the ARG²⁴. We divide up the sites present in the genetic data into nonoverlapping 'words' of S sites and store hash tables mapping from the possible values of the i th word to the samples that carry that word. We use this approach to rapidly detect samples already in the ARG that share words with the target sample and return the top K samples with the most consecutive matches. A tolerance parameter T controls the number of mismatches allowed in an otherwise consecutive stretch. We also allow the top K samples to vary across the genome due to recombination events, by partitioning the genome into regions of genetic distance L . Assuming this results in R regions, the hashing step outputs a matrix of $R \times K$ sample identities (IDs) containing the predicted top K related samples for each region. We note, however, that the hashing step can look arbitrarily far beyond the boundaries of each region to select the K samples.

The sample IDs output by step 1 inform step 2, in which ASMC is run over pairs of samples. In each of the R regions, ASMC computes the posterior mean and maximum a posteriori TMRCA between the sample being threaded and each of the K candidate most related samples. We add burn-in on either side of the region to provide additional context for the ASMC model, 2.0 centimorgans (cM) for all simulation experiments unless otherwise stated and 1.0 cM in real data inference for greater efficiency.

In step 3, ARG-Needle finds the minimum posterior mean TMRCA among the K candidates at each site of the genome. Note that both the use of a posterior mean estimator with a pairwise demographic prior and the selection of a minimum among K estimated values lead to bias in the final TMRCA estimates (Supplementary Fig. 3h), which we later address using a postprocessing normalization step (see below). The corresponding IDs determine which sample in the ARG to thread to at each site. Because the posterior mean assumes continuous values and changes at each site, we average the posterior mean over neighboring sites where the ID to thread to and the associated maximum a posteriori remain constant. This produces piecewise constant values which determine how high above the sample to thread, with changes corresponding to inferred recombination events. The sample is then efficiently threaded into the existing ARG, utilizing custom data structures and algorithms.

Throughout our analyses we adopted $K = 64$, $T = 1$, $L = 0.5$ cM for array data and $L = 0.1$ cM for sequencing data. We used $S = 16$ in simulations, and in real data analyses we increased S as threading proceeded to reduce computation without a major loss in accuracy. For additional details on all three steps in the ARG-Needle algorithm and our parameter choices, see Supplementary Note 1.

ARG normalization

ARG normalization applies a monotonically increasing mapping from existing node times to transformed node times (similar to quantile normalization), further utilizing the demographic prior provided in input. We compute quantile distributions of node times in the inferred ARG as well as in 1,000 independent trees simulated using the demographic model provided in input under the single-locus coalescent. We match the two quantile distributions and use this to rewrite all nodes in the inferred ARG to new corresponding times (Supplementary Note 1). ARG normalization preserves the time-based ordering of nodes and therefore does not alter the topology of an ARG. It is applied by default to our inferred ARGs and optionally to ARGs inferred by Relate (Extended Data Figs. 2–4 and Supplementary Figs. 1, 3 and 4).

Simulated genetic data

We used the msprime coalescent simulator⁶¹ to benchmark ARG inference algorithms. For each run, we first simulated sequence data with given physical length L for N haploid individuals, with $L = 1$ Mb for sequencing and $L = 5$ Mb for array data experiments. Our primary simulations used a mutation rate of $\mu = 1.65 \times 10^{-8}$ per base pair per generation, a constant recombination rate of $\rho = 1.2 \times 10^{-8}$ per base pair per generation and a demographic model inferred using SMC++ on CEU (Utah residents with ancestry from Northern and Western Europe) 1,000 Genomes samples¹⁰. These simulations also output the simulated genealogies, which we refer to as 'ground-truth ARGs' or 'true ARGs'. To obtain realistic SNP data, we subsampled the simulated sequence sites to match the genotype density and allele frequency spectrum of UK Biobank SNP array markers (chromosome 2, with density defined using 50 evenly spaced MAF bins). When running ASMC, we used decoding quantities precomputed for version 1.1, which were obtained using a European demographic model and UK Biobank SNP array allele frequencies, setting two haploid individuals for pairwise TMRCA inference as 'distinguished' and sampling 298 haploid individuals as 'undistinguished'¹¹. ASMC and the hashing step of ARG-Needle also require a genetic map, which we computed based on the recombination rate used in simulations.

In addition to our primary simulations, we included various additional simulation conditions where we modified one parameter while keeping all others fixed. First, we varied the recombination rate to $\rho \in \{6 \times 10^{-9}, 2.4 \times 10^{-8}\}$ per base pair per generation. Second, we used a constant demographic model of 15,000 diploid individuals, for which we generated new decoding quantities to run ASMC. Third, we inferred ARGs using sequencing data, running ASMC in sequencing mode. Fourth, we introduced genotyping errors into the array data. After sampling the array SNPs, we flipped each haploid genotype per SNP and individual with probability p (Supplementary Fig. 4).

Comparisons of ARG inference methods

We compared ASMC-clust and ARG-Needle with the Relate¹⁷ and tsinfer¹⁵ algorithms. Relate runs a modified Li-and-Stephens algorithm⁶² for each haplotype, using all other haplotypes as reference panel. It then performs hierarchical clustering on the output to estimate the topology of marginal trees at each site. Finally, it estimates the marginal tree branch lengths using a Markov chain Monte Carlo algorithm with a coalescent prior. tsinfer uses a heuristic approach to find a set of haplotypes that will act as ancestors for other haplotypes and to rank them based on their estimated time of origin. It then runs a variation of the Li-and-Stephens algorithm to connect older ancestral haplotypes to their descendants, forming the topology of the ARG. To improve the performance of tsinfer in the analysis of UK Biobank array data, the authors developed an alternative approach where subsets of the analyzed individuals are added as potential ancestors¹⁵. This approach was motivated by the sparsity of the variant sites, so we refer to it as 'tsinfer-sparse', obtaining its code from ref. 63.

We ran Relate with the mutation rate, recombination rate and demographic model used in simulations. We kept Relate's default option which limits the memory used for storing pairwise matrices to 5 GB. Because the branch lengths output by tsinfer and tsinfer-sparse are not calibrated, we omitted these methods in comparisons for metrics involving branch lengths. For each choice of sample size, we generated genetic data using five random seeds (25 random seeds in Extended Data Fig. 3d,e) and applied ARG-Needle, ASMC-clust, Relate, tsinfer and (when dealing with array data) tsinfer-sparse to infer ARGs. Due to scalability differences, we ran ASMC-clust and Relate in up to $N = 8,000$ haploid samples ($N = 4,000$ for sequencing) and ARG-Needle, tsinfer and tsinfer-sparse in up to $N = 32,000$ haploid samples. All analyses used Intel Skylake 2.6 GHz nodes on the Oxford Biomedical Research Computing cluster.

The Robinson–Foulds metric²⁷ counts the number of unique mutations that can be generated by one tree but not the other. Because polytomies can skew this metric, we randomly break polytomies as done in ref. 15. We report a genome-wide average, rescaled between 0 and 1.

We generalized the Robinson–Foulds metric to better capture the accuracy in predicting unobserved variants by incorporating ARG branch lengths. To this end, we consider the probability distribution of mutations that arise from uniform sampling on an ARG, and compare the resulting distributions for the true and inferred ARG using the total variation distance, a metric for comparing probability measures. Polytomies do not need to be broken using this metric, as they simply concentrate the probability mass on fewer predicted mutations. We refer to this metric as the ARG total variation distance, and note that it bears similarities to previous extensions of the Robinson–Foulds metric^{64,65} (see Supplementary Note 2 for further details, including an extension that stratifies by allele frequencies).

We also used the KC topology-only distance averaged over all positions to compare ARG topologies. We observed that for methods that output binary trees (Relate, ASMC-clust and ARG-Needle), performance substantially improved when we selected inferred branches at random and collapsed them to create polytomies (solid lines in Extended Data Figs. 1c and 3g), suggesting that the KC topology-only distance rewards inferred ARGs with polytomies. We further quantified the amount of polytomies output by tsinfer and tsinfer-sparse as the mean fraction of nonleaf branches collapsed per marginal tree, observing that when polytomies were randomly broken¹⁵, performance on the KC topology-only distance deteriorated (dashed lines in Fig. 2d and Extended Data Figs. 1b,c and 3f,g). To account for these observations, we compared all methods both with the restriction of no polytomies and with allowing all methods to output polytomies (Fig. 2d and Extended Data Figs. 1b,d and 3f,h). In the latter case, we formed polytomies in ARGs inferred by Relate, ASMC-clust and ARG-Needle using a heuristic to select and collapse branches that are less confidently inferred. For each marginal binary tree, we ordered the $N - 2$ nonleaf branches by computing the branch length divided by the height of the parent node, and collapsed a fraction f of branches for which this ratio is smallest (for additional details, see Supplementary Note 2).

We used the pairwise TMRCA RMSE metric to measure accuracy of inferred branch lengths. The KC distance may also consider branch lengths²⁸, and we performed evaluations using the branch-length-aware versions of the KC distance with parameter $\lambda = 1$, which compares lengths between pairwise MRCA events and the root, and $\lambda = 0.02$, which combines branch length and topology estimation (Supplementary Fig. 1).

Supplementary Note 2 provides further details on the computation of these metrics and their interpretation in the context of ARG inference and downstream analyses.

ARG-MLMA

We developed an approach to perform MLMA of variants extracted from the ARG, which we refer to as ARG-MLMA. We sampled mutations

from a given ARG using a specified rate μ and applied a mixed model association test to these variants. Note that each mutation occurs on a single branch of marginal trees, so that recurrent mutations are not modeled.

For simulation experiments (Fig. 3a and Extended Data Fig. 6) we tested all possible mutations on a true or inferred ARG, which is equivalent to adopting a large value of μ . We used sequencing variants from chromosomes 2–22 to form a polygenic background and added a single causal sequencing variant on chromosome 1 with effect size β . We varied the value of β and measured power as the fraction of runs (out of 100), detecting a significant association on the ARG for chromosome 1. For ARG-MLMA UK Biobank analyses we adopted $\mu = 10^{-5}$, also adding variants sampled with $\mu = 10^{-3}$ to locus-specific Manhattan plots to gain further insights. For additional details on our ARG-MLMA methods, including the determination of significance thresholds, see Supplementary Note 4.

Construction of ARG-GRMs

Consider N haploid individuals, M sites and genotypes x_{ik} for individual i at site k , where variant k has mean p_k . Under an infinitesimal genetic architecture, the parameter α captures the strength of negative selection^{30,66}, with a trait's genetic component given by $g_i = \sum_k \beta_k x_{ik}$ where $\text{Var}(\beta_k) = (p_k(1 - p_k))^\alpha$. Using available markers, a common estimator for the ij -th entry of the $N \times N$ GRM²¹ is

$$K_\alpha(i, j) = \frac{1}{M} \sum_{k=1}^M \frac{(x_{ik} - p_k)(x_{jk} - p_k)}{[p_k(1 - p_k)]^{-\alpha}}. \quad (1)$$

Given an ARG, we compute the ARG-GRM as the expectation of the marker-based GRM that would be obtained using sequencing data, assuming that mutations are sampled uniformly over the area of the ARG. When sequencing data are unavailable but an ARG can be estimated from an incomplete set of markers, the ARG-GRM may provide a good estimate for the sequence-based GRM. We briefly describe how ARG-GRMs are derived from the ARG for the special case of $\alpha = 0$. We discuss the more general case and provide further derivations in Supplementary Note 3.

Assuming $\alpha = 0$, equation (1) is equivalent (up to invariances described in Supplementary Note 3) to the matrix whose ij -th entry contains the number of genomic sites at which sequences i and j differ (that is, their Hamming distance). This may be expressed as

$$K_H(i, j) = \frac{1}{M} \sum_{k=1}^M x_{ik} \oplus x_{jk},$$

where \oplus refers to the exclusive or (XOR) function. Assume there are L base pairs in the genome and a constant mutation rate per base pair of μ , and let t_{ijk} denote the TMRCA of i and j at base pair k . The ij -th entry of the ARG-GRM is equivalent to the expected number of mutations carried by only one of the two individuals, which is proportional to the sum of the pairwise TMRCAs across the genome (Extended Data Fig. 7a):

$$\begin{aligned} K_{\text{ARG}}(i, j) &= \mathbb{E}[K_H(i, j) | \text{ARG}] \\ &= \sum_{k=1}^L P(\text{Poisson}(2\mu t_{ijk}) > 0) = \sum_{k=1}^L 1 - \exp(-2\mu t_{ijk}) \approx \sum_{k=1}^L 2\mu t_{ijk}. \end{aligned}$$

For increased efficiency, we compute a Monte Carlo ARG-GRM by uniformly sampling new mutations on the ARG with a high mutation rate and apply equation (1) to build the ARG-GRM using these mutations. We used simulations to verify that Monte Carlo ARG-GRMs converge to exactly computed ARG-GRMs for large mutation rates, saturating at $\mu \approx 1.65 \times 10^{-7}$ (Extended Data Fig. 7b,c), the default value we used for ARG-GRM computations. Stratified Monte Carlo ARG-GRMs may also be computed by partitioning the sampled mutations based on allele frequency, LD or allele age^{36,31,67,68} (Supplementary Note 3).

ARG-GRM simulation experiments

We simulated polygenic traits from haploid sequencing samples for various values of h^2 and α . We varied the number of haploid samples N but fixed the ratio L/N throughout experiments, where L is the genetic length of the simulated region. For heritability and polygenic prediction experiments, we adopted $L/N = 5 \times 10^{-3}$ Mb per individual. For association experiments, we simulated a polygenic phenotype from 22 chromosomes, with each chromosome consisting of equal length $L/22$ and $L/N = 5.5 \times 10^{-3}$ Mb per individual. Mixed-model prediction r^2 and association power may be roughly estimated as a function of h^2 and the ratio N/M , where M is the number of markers^{39,69,70}. We thus selected values of M and L such that the N/M ratio is kept close to that of the UK Biobank ($L = 3 \times 10^3$ Mb, $N \approx 6 \times 10^5$).

We computed GRMs using ARGs, SNP data, imputed data (IMPUTE4 (ref. 38) within-cohort imputation) and sequencing data, and performed complex trait analyses using GCTA²¹. Polygenic prediction used cvBLUP⁷¹ leave-one-out prediction within GCTA. ARG-GRM association experiments (Fig. 3c and Extended Data Fig. 8c,f) tested array SNPs on each chromosome while using GRMs built on the other 21 chromosomes to increase power, measured as the relative increase of mean $-\log_{10}(P)$ compared with linear regression. We observed that MAF-stratification for ARG-GRMs of true ARGs enabled robust heritability estimation and polygenic prediction if α is unknown (Extended Data Fig. 8g). In experiments involving inferred ARGs (Fig. 3b and Supplementary Fig. 8), we applied MAF-stratification for ARG-Needle ARGs and imputed data, but not for SNP data, for which GCTA did not converge.

ARG-Needle inference in the UK Biobank

Starting from 488,337 samples and 784,256 available autosomal array variants (including SNPs and short indels), we removed 135 samples (129 withdrawn, 6 due to missingness > 10%) and 57,126 variants (missingness > 10%). We performed reference-free phasing of the remaining variants and samples using Beagle 5.1 (ref. 72) and extracted the unrelated white British ancestry subset defined in ref. 38, yielding 337,464 samples. We built the ARG of these samples using ARG-Needle, using parameters described above. We parallelized the ARG inference by splitting phased genotypes into 749 nonoverlapping ‘chunks’ of approximately equal numbers of variants. We added 50 variants on either side of each chunk to provide additional context for inference and independently applied ARG normalization on each chunk. For our brief comparison of ARG inference methods in real data (Supplementary Fig. 6c,d), we repeatedly sampled independent subsets of $N = 2,000$ and $N = 16,000$ diploid individuals, and inferred the ARG for these individuals using the first chunk in the second half of chromosome 1, which amounted to 7.5 Mb.

Genealogy-wide association scan in the UK Biobank

To process phenotypes (standing height, alkaline phosphatase, aspartate aminotransferase, low-density lipoprotein (LDL)/high-density lipoprotein (HDL) cholesterol, mean platelet volume and total bilirubin) we first stratified by sex and performed quantile normalization. We then regressed out age, age squared, genotyping array, assessment center and the first 20 genetic principal components computed in ref. 38. We built a noninfinite BOLT-LMM mixed model using SNP array variants, then tested HRC + UK10K-imputed data^{38,40,41} and variants inferred using the ARG (ARG-MLMA, described above). For association of imputed data (including SNP array) we restricted to variants with Hardy–Weinberg equilibrium $P > 10^{-12}$, missingness < 0.05 and info score > 0.3 (matching the filtering criteria adopted in ref. 38). For all analyses we did not test variants with an MAC < 5 and used MAF thresholds detailed below.

Association analysis of seven traits

Using the filtering criteria above and no additional MAF cutoff, we obtained resampling-based genome-wide significance thresholds of

$P < 4.8 \times 10^{-11}$ (95% CI: 4.06×10^{-11} , 5.99×10^{-11}) for ARG and $P < 1.06 \times 10^{-9}$ (95% CI: 5.13×10^{-10} , 2.08×10^{-9}) for imputation (Supplementary Table 1 and Supplementary Note 4). After performing genome-wide MLMA for the seven traits, we selected genomic regions harboring low-frequency ($0.1\% \leq \text{MAF} < 1\%$), rare ($0.01\% \leq \text{MAF} < 0.1\%$) or ultra-rare ($\text{MAF} < 0.01\%$) variant associations. We then formed regions by grouping any associated variants within 1 Mb of each other and adding 0.5 Mb on either side of these groups.

We next performed several further filtering and association analyses using a procedure similar to ref. 43 to extract sets of approximately independent signals (‘independent’ for short; Supplementary Tables 2–5 and Supplementary Note 4). Of the seven phenotypes, total bilirubin did not yield any rare or ultra-rare independent signals and height did not yield any independent ultra-rare signals. We leveraged the UK Biobank WES data to validate and localize independent associations. We extracted 138,039 exome-sequenced samples that overlap with the analyzed set of white British ancestry individuals and performed lift-over of exome sequencing positions from genome build hg38 to hg19. We computed pairwise LD between the set of independent associated variants and the set of all WES variants, defining the ‘WES partner’ of an independent variant to be the WES variant with largest r^2 to it. In a few instances, the same WES partner was selected by two ARG variants or two imputation variants (Supplementary Tables 2–5). Additionally, three WES partners were selected by one ultra-rare ARG and one rare imputation variant, and one WES partner was selected by one rare ARG and two ultra-rare imputation variants; these WES partners counted towards the 36 WES partners identified by both methods in rare and ultra-rare analyses, but were not counted as jointly identified when restricting to only rare or ultra-rare categories (as in Fig. 4a). We labeled WES variants by gene and functional status (‘loss-of-function’ and ‘other protein altering’; Supplementary Note 4) based on annotations obtained using the Ensembl Variant Effect Predictor (VEP) tool⁷³.

Association analysis for higher frequency variants and height

For genome-wide association analyses of higher frequency variants and height, we matched filtering criteria used in ref. 38, retaining imputed variants that satisfy the basic filters listed above, as well as $\text{MAF} \geq 0.1\%$. Using these criteria, we estimated a resampling-based genome-wide significance threshold of 4.5×10^{-9} (95% CI: 2.2×10^{-9} , 9.6×10^{-9}); Supplementary Table 1). To facilitate direct comparison, we aimed to select parameters that would result in a comparable significance threshold for the ARG-MLMA analysis. Two sets of parameters satisfied this requirement: 3.4×10^{-9} (95% CI: 2.4×10^{-9} , 5×10^{-9}), obtained for $\mu = 10^{-5}$, $\text{MAF} \geq 1\%$; and 4×10^{-9} (95% CI: 3.1×10^{-9} , 5.3×10^{-9}), obtained for $\mu = 10^{-6}$, $\text{MAF} \geq 0.1\%$. We selected the former set of parameters, as a low sampling rate of $\mu = 10^{-6}$ leads to worse signal-to-noise and lower association power. We thus used a significance threshold of $P < 3 \times 10^{-9}$ for all analyses of higher frequency variants. We used PLINK^{74,75} (v.1.90b6.21 and v.2.00a3LM) and GCTA²¹ (v.1.93.2) to detect approximately independent associations using COJO⁴⁷, retaining results with COJO $P < 3 \times 10^{-9}$ (Fig. 5e,f, Extended Data Fig. 10e, Supplementary Fig. 11 and Supplementary Note 4).

Statistics and reproducibility

For real data analysis in the UK Biobank, we included all 337,464 individuals of white British ancestry (as reported in ref. 38) who did not have genotype missingness > 10% and had not withdrawn from the UK Biobank at the time of our analysis. To further explore our findings, we selected the 138,039 of these individuals who were exome sequenced in the 200,000 UK Biobank whole-exome sequencing release.

Ethics

UK Biobank data were analyzed after approval of UK Biobank proposal no. 43206.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

COJO association signals for higher frequency ARG variants with height are available at <https://doi.org/10.5281/zenodo.7411562>. VEP annotations were generated using the Ensembl VEP tool (v.101.0, output produced February 2021), <https://www.ensembl.org/info/docs/tools/vep/index.html>. UK Biobank data can be accessed by approved researchers through <https://www.ukbiobank.ac.uk/>. Other datasets were downloaded from the following URLs: summary statistics from whole-exome imputation from 50,000 sequences⁴³, https://data.broadinstitute.org/lohlab/UKB_exomeWAS/; likely causal associations from whole-exome imputation from 50,000 sequences⁴³, <https://www.nature.com/articles/s41588-021-00892-1> Supplementary Table 3; GIANT consortium summary statistics in ~700,000 (ref. 46), https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files.

Code availability

The arg-needle and arg-needle-lib software packages, which implement the ARG-Needle and ASMC-clust methods as well as methods for the main analyses in this paper, are available at <https://palamaralab.github.io/software/argneedle/>. Python packages can be downloaded at <https://pypi.org/project/arg-needle/> and <https://pypi.org/project/arg-needle-lib/>; analysis scripts are available at <https://doi.org/10.5281/zenodo.7745745>. External softwares used in the current study were obtained from the following URLs: msprime (v.0.7.4), <https://pypi.org/project/msprime/>; tsinfer (v.0.1.4), <https://pypi.org/project/tsinfer/>; tsinfer scripts for sparse data (accessed January 2022), <https://github.com/mcveanlab/treeseq-inference>; Relate (v.1.0.15), <https://myersgroup.github.io/relate/>; ARGON (v.0.1.160415), <https://github.com/pierpal/ARGON/>; DASH (v.1.1) and GERMLINE (v.1.5.3), <http://www1.cs.columbia.edu/~gusev/dash/>; IMPUTE4 (v.4.1.2), <https://jmarchini.org/software/#impute-4>; Beagle (v.5.1), <https://faculty.washington.edu/browning/beagle/b5.1.html>; PLINK (v.1.90b6.21), <https://www.cog-genomics.org/plink/>; PLINK (v.2.00a3LM), <https://www.cog-genomics.org/plink/2.0/>; GCTA (v.1.93.2), <https://cns.genomics.com/software/gcta/>; BOLT-LMM (v.2.3.2), <https://alkesgroup.broadinstitute.org/BOLT-LMM/downloads/>; LiftOver (used April 2021), <https://genome.ucsc.edu/cgi-bin/hgLiftOver>.

References

58. Sneath, P. H. & Sokal, R. R. *Numerical Taxonomy. The Principles and Practice of Numerical Classification* (W. H. Freeman and Co., 1973).
59. Gronau, I. & Moran, S. Optimal implementations of UPGMA and other common clustering algorithms. *Inf. Process. Lett.* **104**, 205–210 (2007).
60. Müllner, D. fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. *J. Stat. Softw.* **53**, 1–18 (2013).
61. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
62. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
63. Wong, Y., Kelleher, J., Wohns, A. W. & Fadil, C. Evaluating tsinfer. GitHub <https://github.com/mcveanlab/treeseq-inference> (2020).
64. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
65. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).
66. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
67. Lee, S. H. et al. Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
68. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
69. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
70. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008).
71. Mefford, J. et al. Efficient estimation and applications of cross-validated genetic predictions to polygenic risk scores and linear mixed models. *J. Comput. Biol.* **27**, 599–612 (2020).
72. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
73. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
74. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
75. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

Acknowledgements

We thank P.-R. Loh, A. Gusev, S. R. Myers, R. Davies, N. Whiffin, A. Dahl and R. Fournier for discussions and suggestions; and S. Shi, J. Nait Saada, G. Kalantzis and J. Zhu for sharing code used for various parts of the analysis. This work was conducted using the UK Biobank resource (application no. 43206). We thank the participants of the UK Biobank project. This work was supported by the Clarendon Scholarship (to Á.F.G. and B.C.Z.); NIH grant no. R21-HG010748-01 (to P.F.P., F.C. and A.B.); Wellcome Trust ISSF grant no. 204826/Z/16/Z (to P.F.P.); Wellcome Trust grant no. 222336/Z/21/Z (to Á.F.G.); and ERC Starting Grant no. ARGPHENO 850869 (to P.F.P., F.C., A.B. and B.C.Z.). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant no. 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research was funded in whole, or in part, by the Wellcome Trust (204826/Z/16/Z; 222336/Z/21/Z; 203141/Z/16/Z). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

P.F.P. designed the study. B.C.Z. and P.F.P. implemented the ASMC-clust and ARG-Needle algorithms. B.C.Z., Á.F.G. and P.F.P. performed and analyzed simulations. B.C.Z., A.B. and P.F.P. performed

analysis of UK Biobank data. F.C. developed software tools. B.C.Z. and P.F.P. wrote the manuscript.

Competing interests

During the revision of this manuscript, A.B. became an employee of 54gene and Á.F.G. became an employee of deCODE genetics/Amgen. The remaining authors declare no competing interests.

Additional information

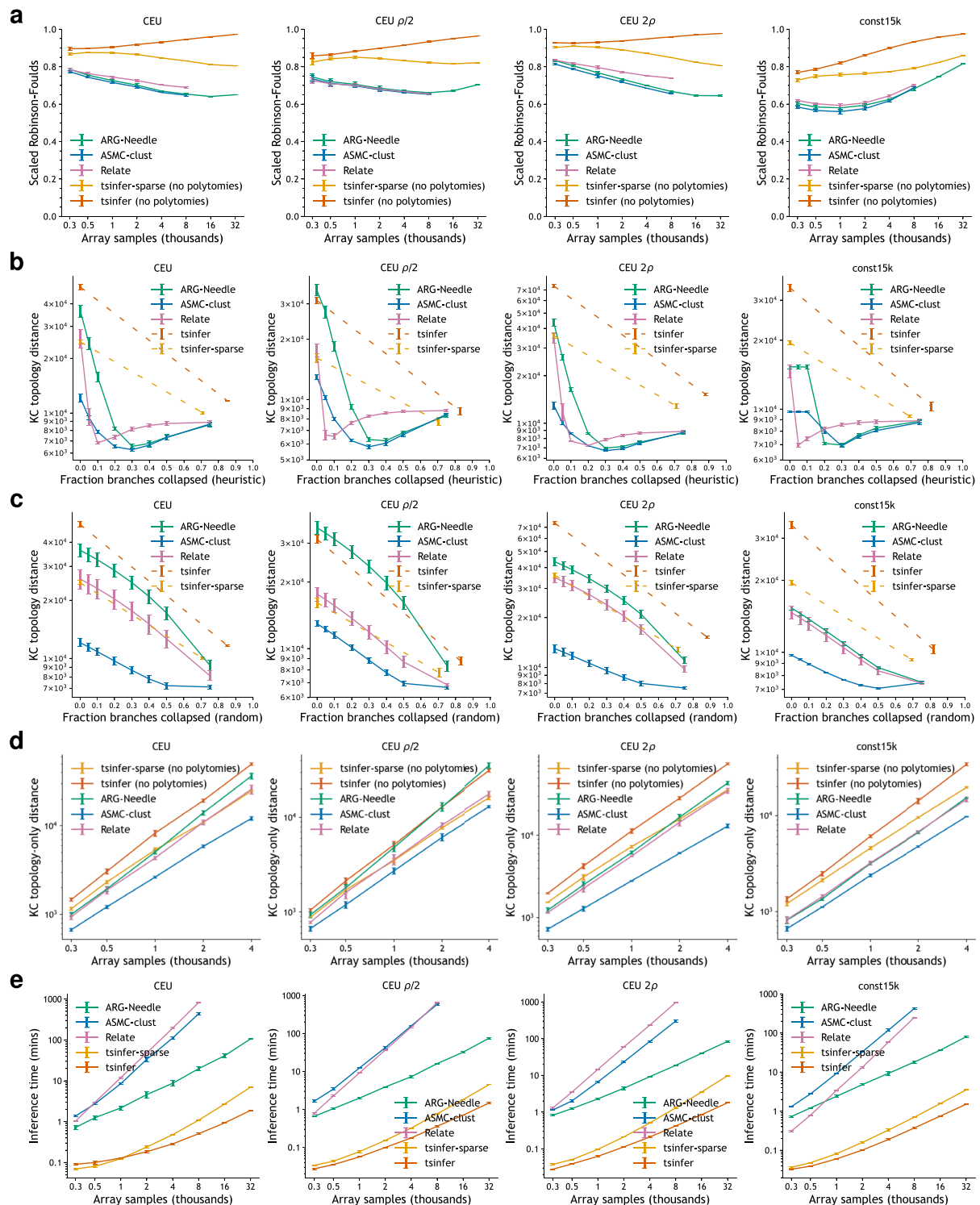
Extended data is available for this paper at <https://doi.org/10.1038/s41588-023-01379-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01379-x>.

Correspondence and requests for materials should be addressed to Pier Francesco Palamara.

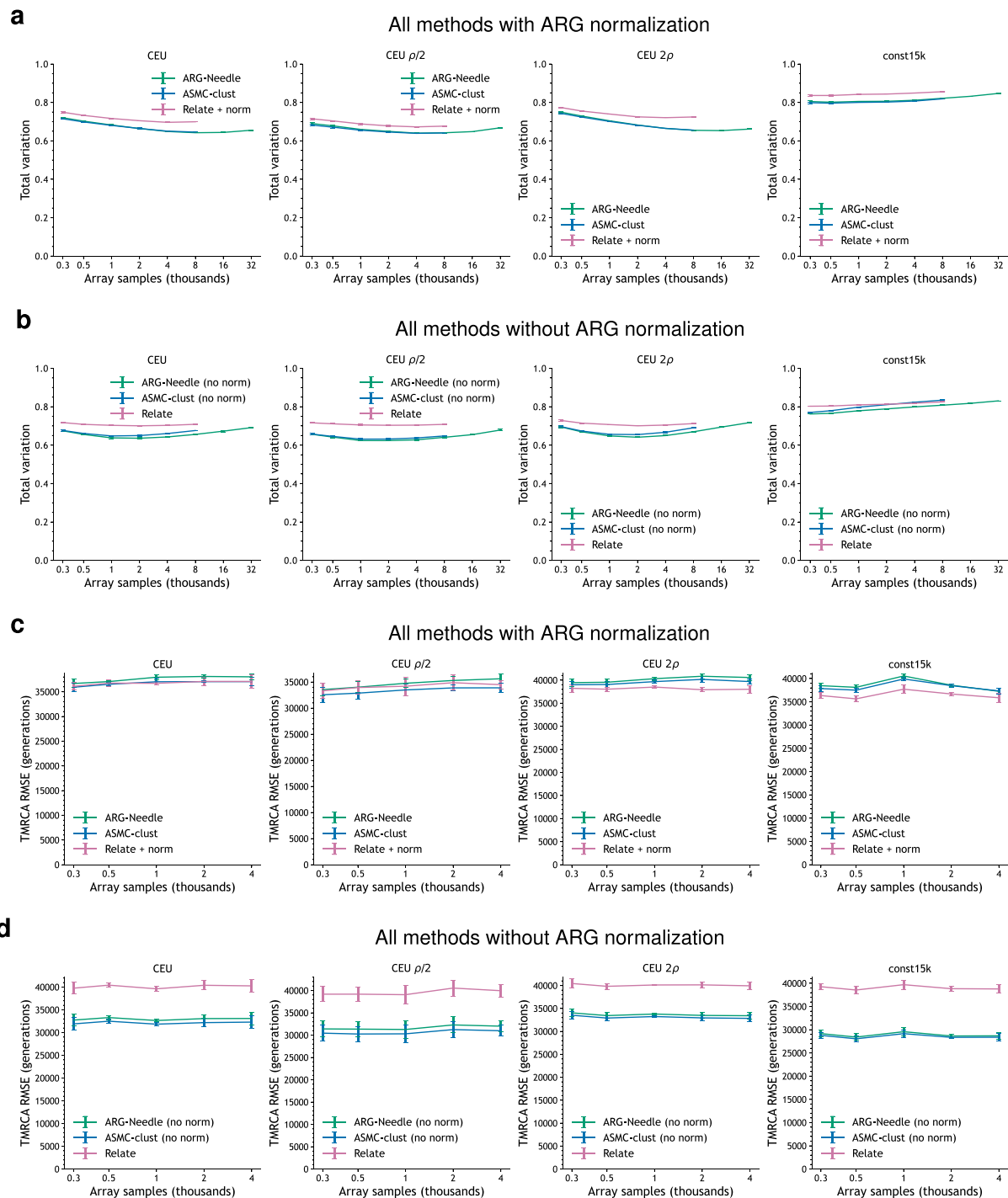
Peer review information *Nature Genetics* thanks Leo Speidel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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



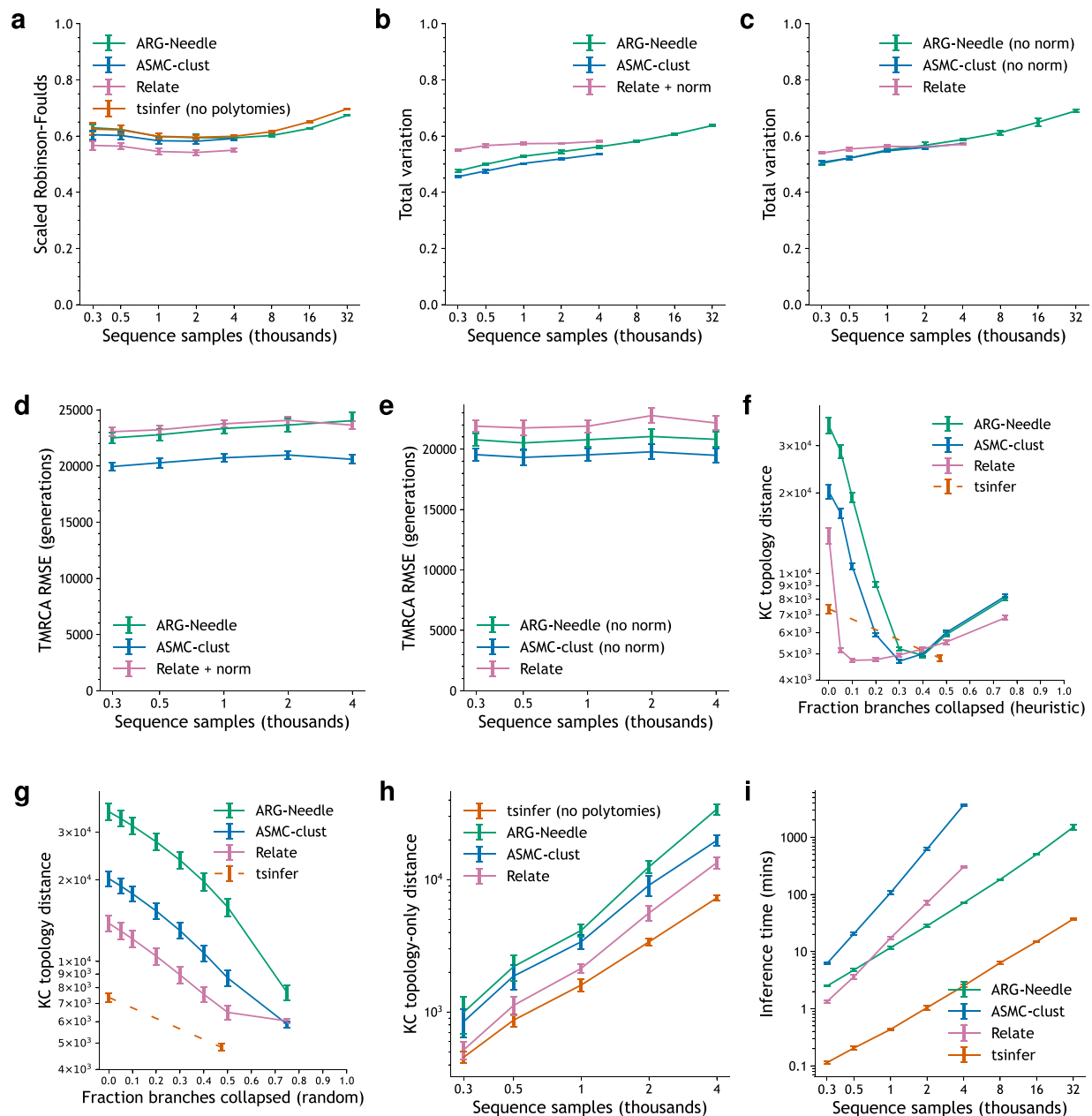
Extended Data Fig. 1 | Additional comparison of ARG inference methods with array data and topology-only metrics. We compare methods on runtime and topology-only metrics, as in Fig. 2 but with additional simulation conditions. All columns are for 5 Mb of CEU demography array data, and individual columns represent standard parameters (see Methods), a factor of 2 smaller recombination rate ($\rho = 6 \times 10^{-9}$), a factor of 2 larger recombination rate ($\rho = 2.4 \times 10^{-8}$), and a constant population size demography of 15,000 individuals. **a.** Robinson-Foulds distance as a function of the number of samples N , where values are scaled to lie between 0 and 1 (polytomies are randomly resolved). **b.** KC topology-only distance for $N = 4,000$ samples, showing performance

as branches in marginal inferred trees are collapsed to form polytomies, using a heuristic to preferentially collapse branches that are least certain (see Methods). For tsinfer and tsinfer-sparse, we instead rely on the default amount of polytomies in the output, additionally showcasing when polytomies are randomly resolved (dashed lines indicate a linear trend may not hold). **c.** The same as **b**, except branches are randomly collapsed to form polytomies. **d.** KC topology-only distance as a function of N , with polytomies randomly resolved. **e.** Inference time as a function of N . All panels use 5 random seeds. Data are presented as means ± 2 s.e.m.



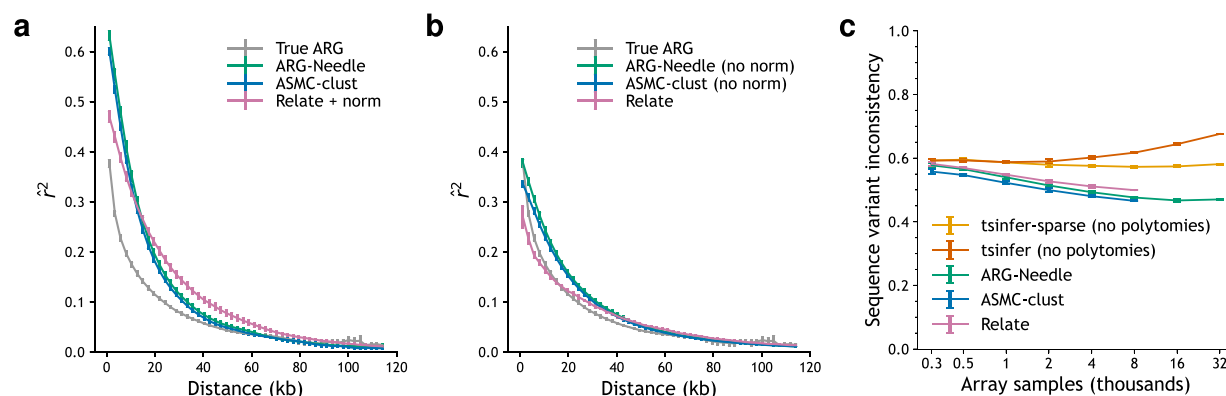
Extended Data Fig. 2 | Additional comparison of ARG inference methods with array data and two branch length-aware metrics. We compare methods as in Fig. 2b, c, but with additional simulation conditions. All columns are for 5 Mb of CEU demography array data, and individual columns represent standard parameters (see Methods), a factor of 2 smaller recombination rate ($\rho = 6 \times 10^{-9}$),

a factor of 2 larger recombination rate ($\rho = 2.4 \times 10^{-8}$), and a constant population size demography of 15,000 individuals. We show results for the ARG total variation distance (a–b) and pairwise TMRCA RMSE (c–d), with (a, c) and without (b, d) ARG normalization, as these metrics are sensitive to branch length. All panels use 5 random seeds. Data are presented as means \pm 2 s.e.m.



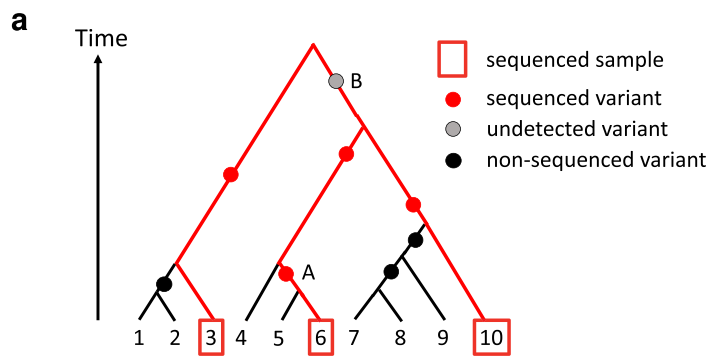
Extended Data Fig. 3 | Comparison of ARG inference methods with sequencing data. Simulations use 1 Mb of CEU sequencing data and otherwise standard parameters (see Methods). Individual panels correspond to rows of Extended Data Figs. 1a, 2a–d, and 1b–e, in that order, with the same metrics used, namely **a**, scaled Robinson-Foulds distance (polytomies are randomly resolved), **b–c**, ARG total variation distance with **(b)** and without **(c)** ARG normalization,

d–e, pairwise TMRCA RMSE with **(d)** and without **(e)** ARG normalization, **f–g**, KC topology-only distance for $N = 4,000$ samples with heuristic **(f)** and random **(g)** collapsing of branches, **h**, KC topology-only distance with polytomies randomly resolved, and **i**, inference time. **d, e** use 25 random seeds, whereas all other panels use 5 random seeds. Data are presented as means ± 2 s.e.m.

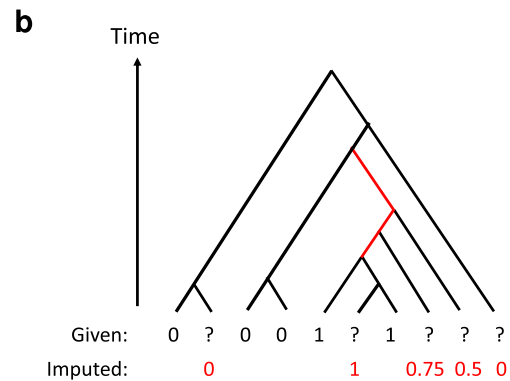


Extended Data Fig. 4 | Consistency of inferred ARGs with underlying linkage patterns and sequence-level variation. a,b. Linkage disequilibrium (LD) decay up to 120 kilobases for ground truth ARGs as well as ARGs inferred by ARG-Needle, ASMC-clust, and Relate. LD was evaluated by placing mutations with a mutation rate of 5×10^{-8} per base pair per generation and filtering to variants with $\text{MAF} > 5\%$. Lines show mean r^2 as a function of distance between variants, averaging across 10 independent simulations. Simulations were of 5 Mb of CEU demography array data with standard simulation parameters (see Methods).

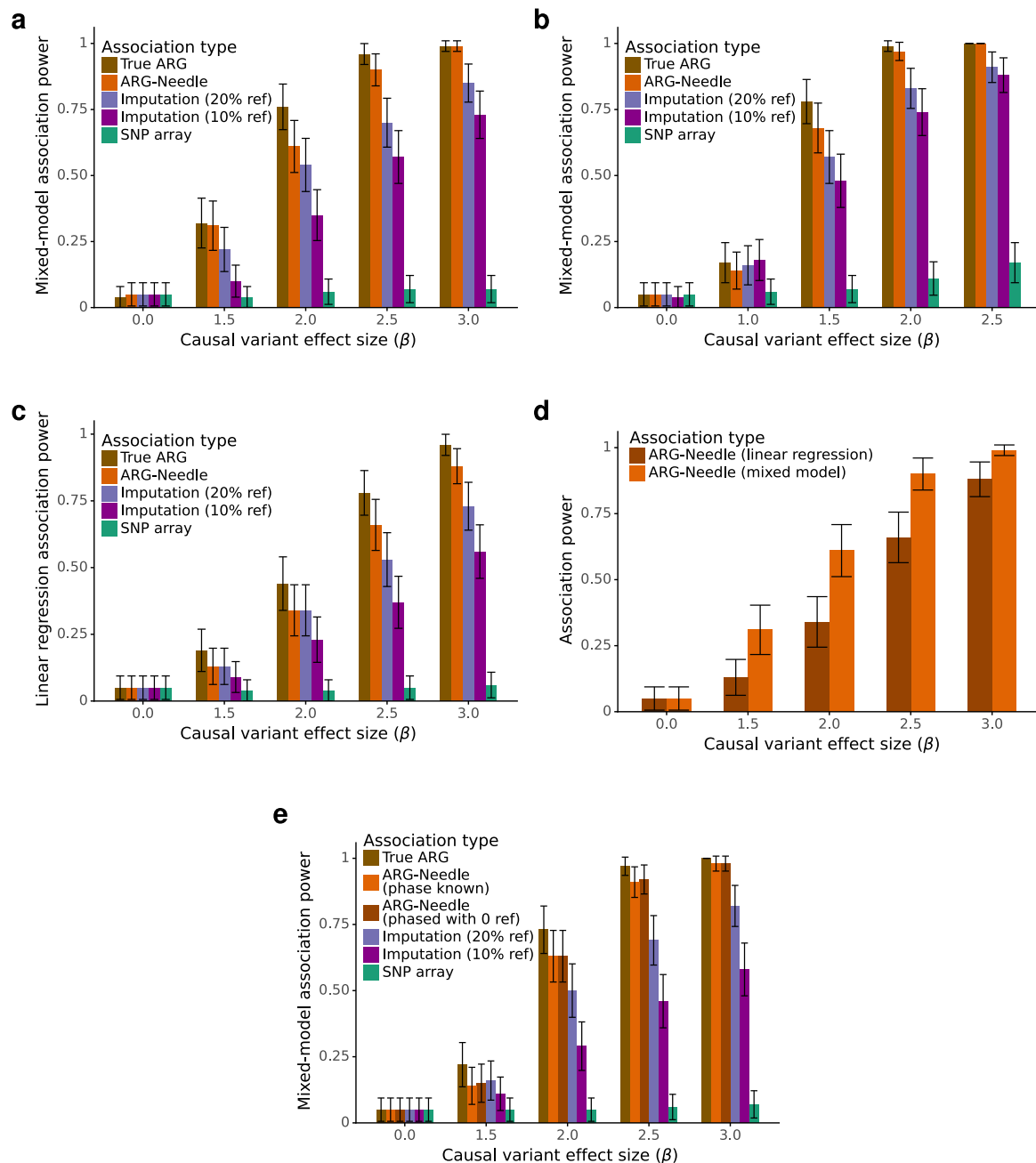
Methods including ARG normalization are shown in **a**, and methods without ARG normalization are shown in **b**, as branch lengths affect the probability for mutations to be sampled. **c.** We compute the fraction of underlying sequencing sites, of which the array variants are a subset, that cannot be mapped to branches of inferred ARGs (lower is better). Inference is on 5 Mb of CEU demography array data simulated with standard parameters (see Methods), averaging over 5 random seeds. no polytomies refers to randomly resolving polytomies of tsinfer and tsinfer-sparse (see Methods). Data are presented as means ± 2 s.e.m.



Extended Data Fig. 5 | A genealogical view of genotype imputation and an algorithm for ARG-based imputation. a. The marginal tree represents the relationships of 10 haploid samples and variant ages at a locus. 3 of the 10 samples are sequenced and used as a reference panel to impute sequenced variants into the remaining samples. An imputation algorithm may recognize sample 6 as the closest relative in the reference panel for samples 4 and 5, but if TMRCAs and variant times are not modeled, it may incorrectly impute variant 'A' into sample 4. Variant 'B' may represent a high frequency variant that is not present in the sequencing panel (for example, an undetected indel or structural variant). Non-sequenced variants cannot be imputed. All variants may be tested for association using branches of an accurately inferred genealogy.

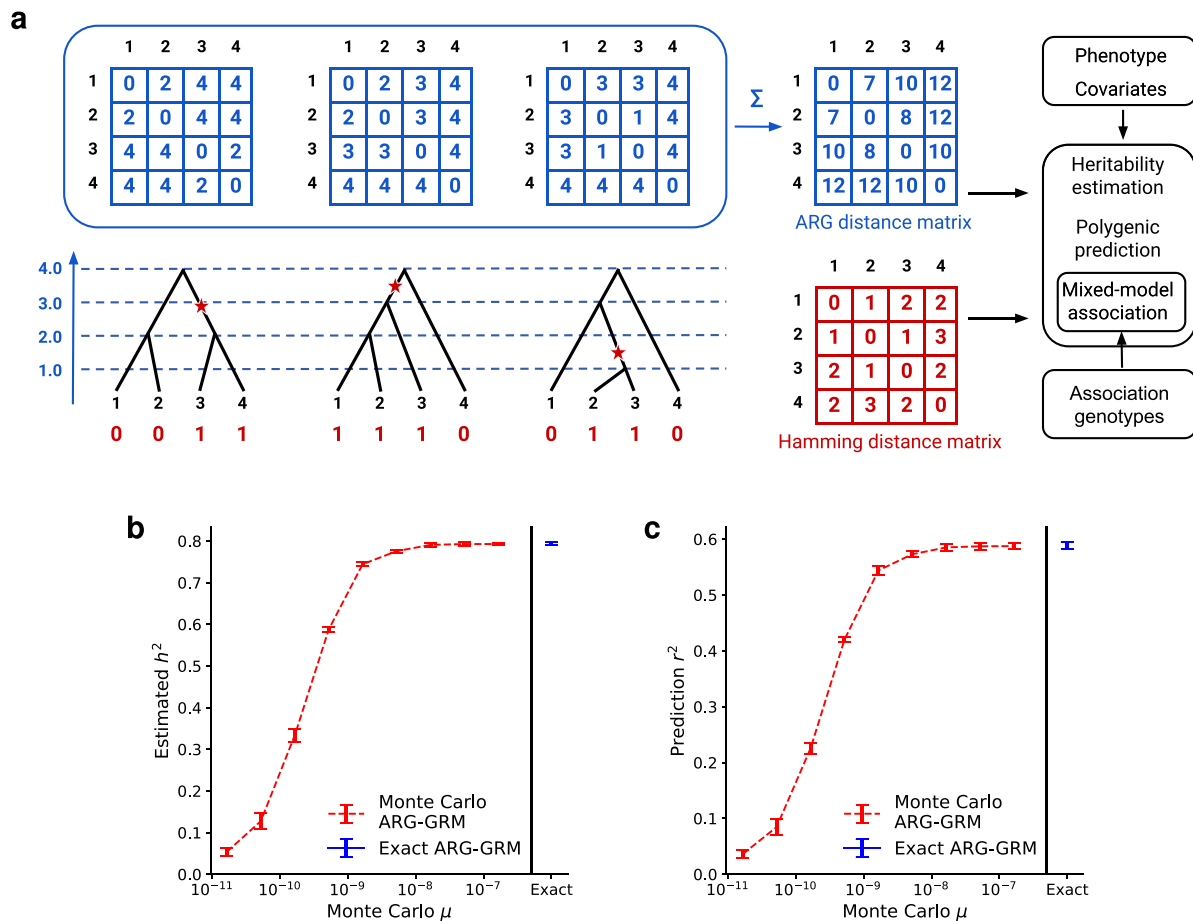


b. Schematic of an ARG-based imputation algorithm (see Supplementary Fig. 12 for exploratory results). Given a polymorphic sequenced site containing sequenced samples, unobserved genotypes for array samples, and a marginal tree relating all samples, we perform genotype imputation as follows. We first identify all branches in the tree for which a mutation on that branch best explains the observed sequencing data in terms of Hamming distance (red branches in the example). Each branch implies genotypes of 0 or 1 for the array samples, and we weight by branch length to produce a weighted predicted dosage for each array sample. In this example, the three branches have lengths in ratio 1:1:2, resulting in the predicted dosages shown in red.



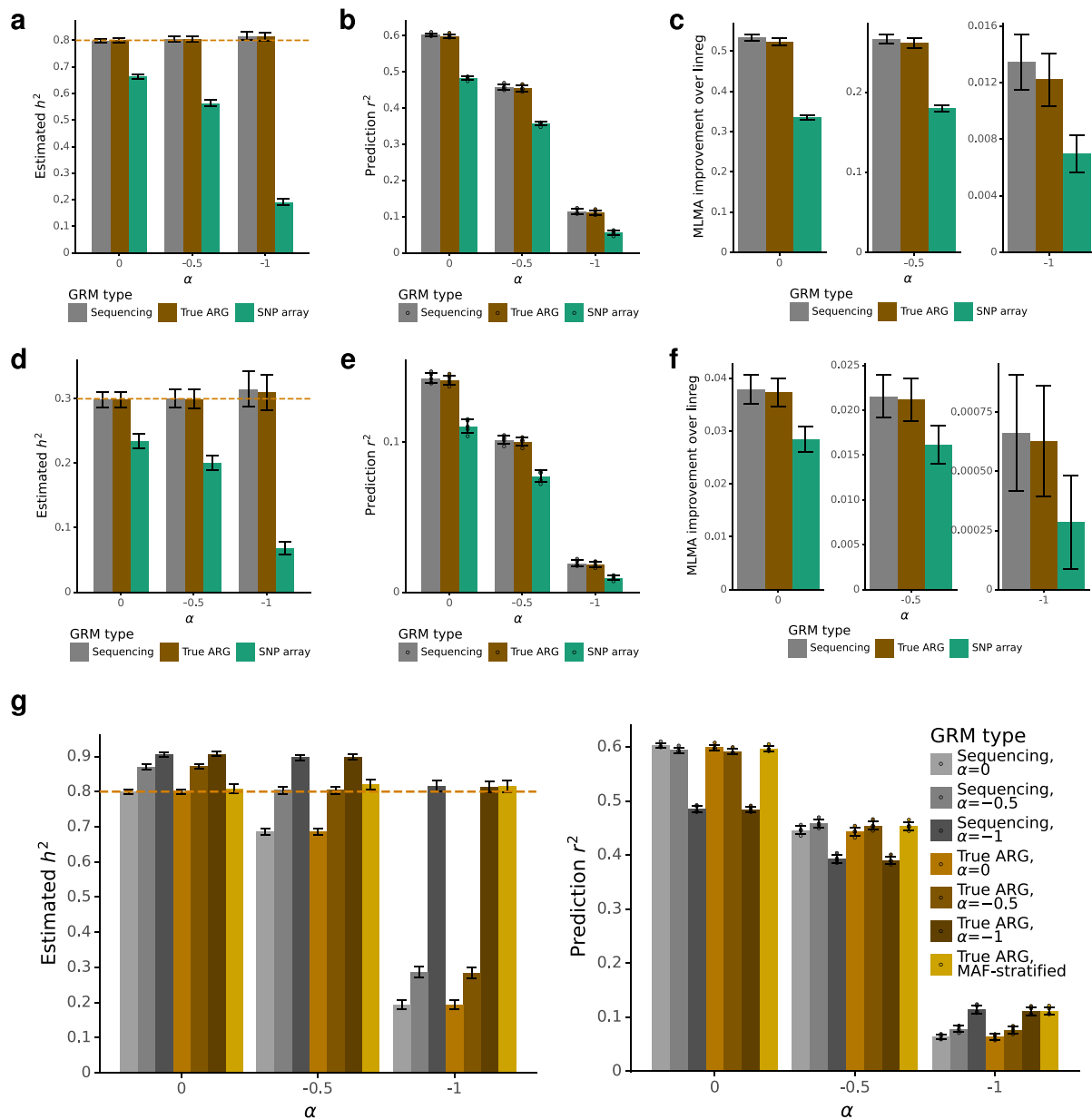
Extended Data Fig. 6 | Additional simulations of ARG-MLMA genealogy-wide association power. **a.** Similar to Fig. 3a, except with a low-frequency causal variant (MAF = 0.05%) and a smaller simulation with $N = 10,000$ haploid samples and 22 chromosomes of 2.5 Mb each. **b.** Similar to **a**, except with the causal variant MAF chosen to be 0.1%. **c.** Similar to **a**, except using linear regression instead of the linear mixed model to test for association. **d.** We combine the association power results of ARG-Needle association from **a** and **c**, highlighting

the improvement of ARG-MLMA compared to directly testing ARG clades using linear regression. **e.** As in **a**, but with $N = 10,000$ diploid instead of $N = 10,000$ haploid individuals. ARG-Needle is run with the true phase known and with reference-free phasing. %ref indicates the size of the reference panel used for imputation as a percentage of the number of haploid samples ($N = 10,000$ in **a-c**, $2N = 20,000$ in **e**). All panels use 100 independent simulations to measure power. Data are presented as fractions ± 2 s.e.m.



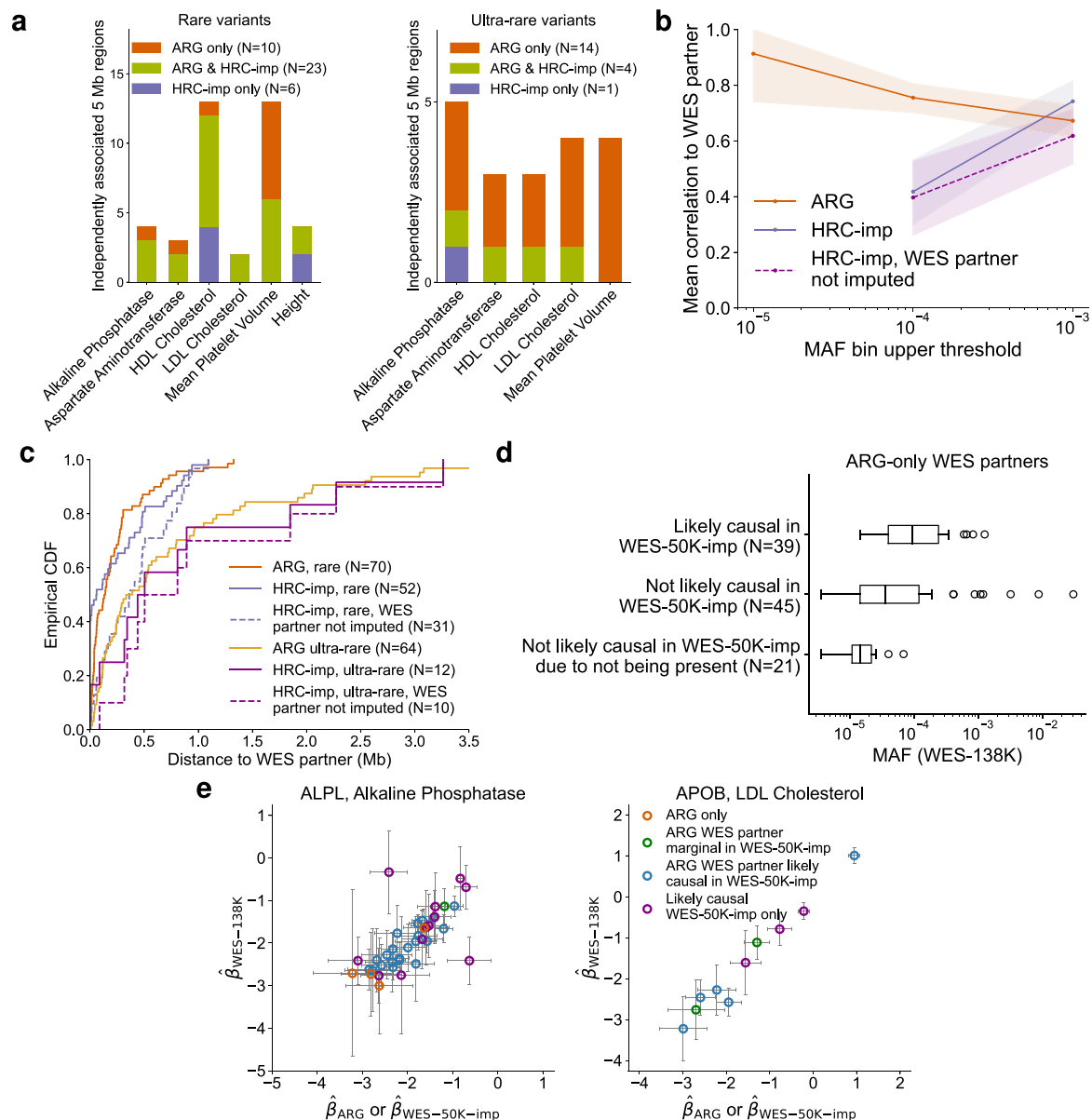
Extended Data Fig. 7 | Overview of ARG-GRM definition and Monte Carlo estimator. a. Schematic of ARG-GRMs. Given an ARG between samples, we can compute the TMRCA matrix at each site and sum this over the genome to obtain the $\alpha = 0$ ARG distance matrix (top, in blue). This equals a scaled version of the expected Hamming distance matrix (bottom, in red), which is formed by counting the number of differences between the genotypes of samples. By applying a series of simple matrix transformations to the ARG distance matrix (see Supplementary Note 3), we obtain the ARG-GRM, which can subsequently be

used in complex trait analysis just like genotype-based GRMs. **b,c.** We compare the use of an exact $\alpha = 0$ ARG-GRM to Monte Carlo $\alpha = 0$ ARG-GRMs for heritability estimation (**b**) and polygenic prediction (**c**). As we increase the mutation rate for the Monte Carlo ARG-GRMs (rightmost value of $\mu = 1.65 \times 10^{-7}$), we approach results from using the exact ARG-GRM. Shown are 5 independent simulations of $N = 2,000$ haploid samples, $h^2 = 0.8$, $\alpha = 0$, 10 Mb. Data are presented as estimates ± 2 s.e.m., where the estimates are from meta-analysis in the case of heritability estimation and represent means otherwise.



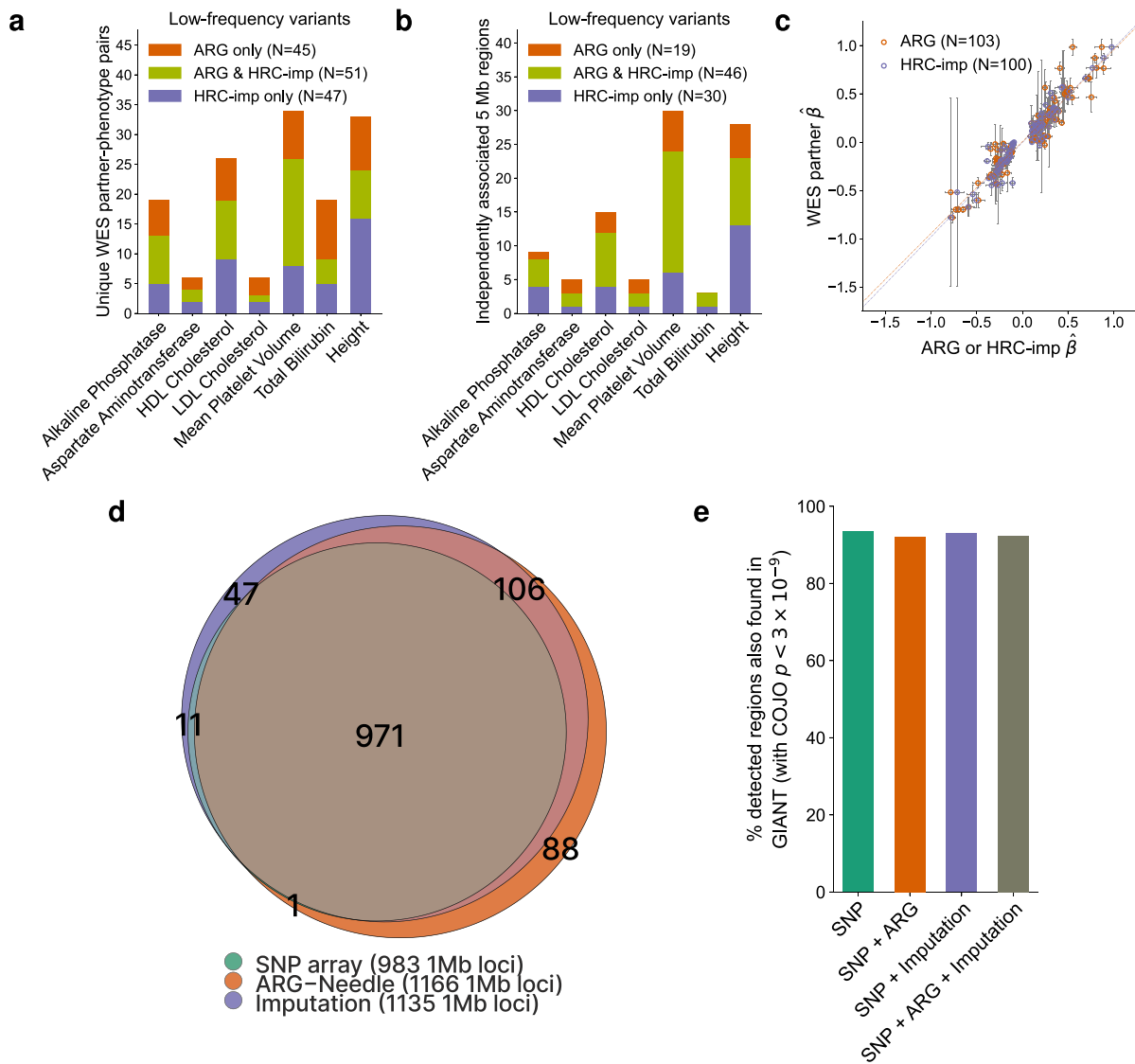
Extended Data Fig. 8 | Additional simulations for ground-truth ARG-GRMs. **a-f.** As in Fig. 3c, with $N = 10,000$ haploid samples, except we vary $h^2 \in \{0.8, 0.3\}$ and $\alpha \in \{0, -0.5, -1\}$. **a, d.** Heritability estimation for a 50 Mb region for $h^2 = 0.8$ (**a**) and $h^2 = 0.3$ (**d**). **b, e.** Polygenic prediction for a 50 Mb region for $h^2 = 0.8$ (**b**) and $h^2 = 0.3$ (**e**). **c, f.** Mixed-model association for 22 chromosomes of 2.5 Mb each for $h^2 = 0.8$ (**c**) and $h^2 = 0.3$ (**f**). **g.** Panels **a-f** assumed it is possible to infer α and used the true α when building genotype-based or ARG-GRMs. If this value of α is misspecified, heritability estimation is biased and prediction r^2 is hampered. This is true both for ARG-GRMs and sequencing GRMs. However, using MAF-stratified

ARG-GRMs provides a robust way to estimate the true heritability when α is unknown, and achieves prediction r^2 comparable to using the true α ($N = 10,000$ haploid samples, 50 Mb, $h^2 = 0.8$). For all panels, heritability and prediction experiments involve 5 simulations per bar, and most association experiments involve 50 simulations per bar, except for the $h^2 = 0.3$, $\alpha = -1$ condition in **f**, which involved 500 simulations. Data are presented as estimates ± 2 s.e.m., where the estimates are from meta-analysis in the case of heritability estimation and represent means otherwise. Prediction r^2 for individual simulations is shown in **b** and **e**.



Extended Data Fig. 9 | Further results for rare and ultra-rare variant associations. **a.** Counts of implicated 5 Mb regions containing ARG and HRC + UK10K imputation ('HRC-imp') independent associations, partitioned by traits and frequency and showing overlap. Total bilirubin was not associated at these frequencies. **b.** Average Pearson correlation between independent variants and their WES partners as a function of frequency, for ARG-derived variants, HRC + UK10K imputed variants, and HRC + UK10K imputed variants for which the WES partner was not the imputed variant. Dots represent the upper end of a frequency range. Central lines represent means and shaded areas represent 95% bootstrap confidence intervals. **c.** Cumulative distribution function for the distance between independent variants and their WES partners, partitioned by frequency. As in Fig. 4b, but also showing HRC + UK10K imputed variants for which the WES partner was not the imputed variant. **d.** Box plots of MAF for WES

partners found by ARG-derived but not HRC + UK10K imputed independent variants (center line, median; box limits, upper and lower quartiles, whiskers, 1.5× interquartile range; points, outliers), stratifying by status in WES-50K-imp (imputation from WES-50K). **e.** Scatter plot of $\hat{\beta}$ (estimated effect) for ARG-derived independent variants (estimated within 337,464 samples) against $\hat{\beta}$ for their WES partners (estimated within 138,039 samples), as in Fig. 4f but for associations with alkaline phosphatase in the *ALPL* gene and with LDL cholesterol in the *APOB* gene. We color points based on whether the WES partner is likely causal in WES-50K-imp, not likely causal but marginally significant in WES-50K-imp, or not marginally significant in WES-50K-imp ('ARG only' in figure). We also plot the $\hat{\beta}$ for the additional likely causal variants in WES-50K-imp against the $\hat{\beta}$ in WES-138K. Error bars represent 1.96 s.e.m.



Extended Data Fig. 10 | Additional results for low ($0.1\% \leq \text{MAF} < 1\%$) and high frequency ($\text{MAF} \geq 1\%$) variant associations. a–c. Association of ARG-derived and imputed low-frequency variants with 7 quantitative traits. **a.** Counts of unique WES partners for ARG and HRC + UK10K imputed ('HRC-imp') independent associations, partitioned by traits and showing overlap. **b.** Counts of implicated 5 Mb regions containing ARG and HRC + UK10K imputation independent associations, partitioned by traits and showing overlap. **c.** Scatter plot of estimated effect ($\hat{\beta}$) for independent variants (estimated within 337,464 samples) against $\hat{\beta}$ for their WES partners (estimated within 138,039 samples), with linear model fit. Error bars represent 1.96 s.e.m. **d, e.** Association of higher

frequency variants with height. **d.** Venn diagram of number of 1 Mb regions containing a significant hit at $P < 3 \times 10^{-9}$ for ARG-Needle ($\text{MAF} \geq 1\%$, $\mu = 10^{-5}$), HRC + UK10K imputed ($\text{MAF} \geq 0.1\%$, info score > 0.3) and SNP array association. ARG-Needle association detected 971 out of 982 (98.9%) 1 Mb regions found by both imputation and array, 108 out of 153 (71%) 1 Mb regions found by imputation but not array and an additional 92 (8% increase upon 1140) 1 Mb regions to those already found by imputation and array. **e.** Percent of 1 Mb regions containing independent associations (defined as having COJO $P < 3 \times 10^{-9}$, see Methods) in association scans of 337,464 UK Biobank individuals that were also present in a GIANT consortium meta-analysis of ~700,000 samples.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input checked="" type="checkbox"/>	<input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input type="checkbox"/>	<input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used in data collection.
Data analysis	The arg-needle and arg-needle-lib software packages, which implement the ARG-Needle and ASMC-clust methods as well as methods for the main analyses in this paper, are available at https://palamaralab.github.io/software/argneedle/ . Python packages can be downloaded at https://pypi.org/project/arg-needle/ and https://pypi.org/project/arg-needle-lib/ ; analysis scripts are available at https://doi.org/10.5281/zenodo.7745745 . External software used in the current study were obtained from the following URLs: msprime (v0.7.4), https://pypi.org/project/msprime/ ; tsinfer (v0.1.4), https://pypi.org/project/tsinfer/ ; tsinfer scripts for sparse data (accessed Jan 2022), https://github.com/mcveanlab/treeseq-inference ; Relate (v1.0.15), https://myersgroup.github.io/relate/ ; ARGON (v0.1.160415), https://github.com/pierpal/ARGON/ ; DASH (v1.1) and GERMLINE (v1.5.3), http://www1.cs.columbia.edu/~gusev/dash/ ; IMPUTE4 (v4.1.2), https://jmarchini.org/software/#impute-4 ; Beagle (v5.1), https://faculty.washington.edu/browning/beagle/b5_1.html ; PLINK (v1.90b6.21), https://www.cog-genomics.org/plink/ ; PLINK (v2.00a3LM), https://www.cog-genomics.org/plink/2.0/ ; GCTA (v1.93.2), https://cnsgenomics.com/software/gcta/ ; BOLT-LMM (v2.3.2), https://alkesgroup.broadinstitute.org/BOLT-LMM/downloads/ ; LiftOver (used April 2021), https://genome.ucsc.edu/cgi-bin/hgLiftOver .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

COJO association signals for higher frequency ARG variants with height are available at <https://doi.org/10.5281/zenodo.7411562>. VEP annotations were generated using the Ensembl VEP tool (v101.0, output produced February 2021), <https://www.ensembl.org/info/docs/tools/vep/index.html>. UK Biobank data can be accessed by approved researchers through <https://www.ukbiobank.ac.uk/>. Other datasets were downloaded from the following URLs: summary statistics from whole exome imputation from 50K sequences, https://data.broadinstitute.org/lohlab/UKB_exomeWAS/; likely causal associations from whole exome imputation from 50K sequences, <https://www.nature.com/articles/s41588-021-00892-1> Supplementary Table 3; GIANT consortium summary statistics in ~700K, https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For real data analysis in the UK Biobank, we included all 337,464 individuals of White British ancestry (as reported in Bycroft et al. Nature 2018) whom did not have genotype missingness > 10% and had not withdrawn from the UK Biobank at the time of our analysis. To further explore our findings using exome sequencing data, we selected the 138,039 of these individuals who were exome sequenced in the 200K UK Biobank whole exome sequencing release.
Data exclusions	We excluded individuals who had withdrawn from the UK Biobank at the time of our analysis and individuals with genotype missingness > 10%.
Replication	The best tagged whole exome sequencing (WES) variants uniquely identified by our rare and ultra-rare ARG associations were validated using the likely-causal WES variants reported in Barton et al. Nature Genetics 2021. We found that 14/30 rare and 28/54 ultra-rare tagged WES variants were also detected as likely-causal associations (at $p < 5 \times 10^{-8}$) in Barton et al. Higher-frequency ARG associations with height were validated using GIANT consortium meta-analysis of 700K individuals comprising the UK Biobank and additional cohorts. A significant fraction (54/92, permutation $p < 0.0001$) of regions identified uniquely using the ARG contained significant associations ($p < 3 \times 10^{-9}$) in this meta-analysis.
Randomization	There was no allocation into experimental groups in this study. When performing our phenotypic association scans, we controlled for covariates by first stratifying by sex and performing quantile normalization. We then regressed out age, age squared, genotyping array, assessment center, and the first 20 genetic principal components computed in Bycroft et al. Nature 2018.
Blinding	Blinding was not applicable to the analyses we performed within the UK Biobank, which analyzed all samples jointly. Data collection was performed previously by the UK Biobank.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging