

Rare variant enriched identity-by-descent enables the detection of distant relatedness and older divergence between populations

Amol C. Shetty^{1,2,3,*}, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium⁴, TOPMed Population Genetics Working Group⁴, Jeffrey O'Connell^{2,3}, Braxton D. Mitchell^{3,5,6}, Timothy D. O'Connor^{1,2,3,*}

¹ Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

² Epidemiology and Human Genetics, University of Maryland School of Medicine, Baltimore, MD

³ Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD

⁴ National Heart, Lung and Blood Institute (NHLBI), Bethesda, MD

⁵ Department of Medicine, University of Maryland School of Medicine, Baltimore, MD

⁶ Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD

First Author: amolcarlshettyphd@gmail.com

* Corresponding Author: amolcarlshettyphd@gmail.com, timothydoconnor@gmail.com

Abstract

Motivation: The global human population has experienced an explosive growth from a few million to roughly 7 billion people in the last 10,000 years. Accompanying this growth has been the accumulation of rare variants that can inform our understanding of human evolutionary history. Common variants have primarily been used to infer the structure of the human population and relatedness between two individuals. However, with the increasing abundance of rare variants observed in large-scale projects, such as Trans-Omics for Precision Medicine (TOPMed), the use of rare variants to decipher cryptic relatedness and fine-scale population structure can be beneficial to the study of population demographics and association studies. Identity-by-descent (IBD) is an important framework used for identifying these relationships. IBD segments are broken down by recombination over time, such that longer shared haplotypes give strong evidence of recent relatedness while shorter shared haplotypes are indicative of more distant relationships. Current methods to identify IBD accurately detect only long segments ($> 2\text{cM}$) found in related individuals. **Algorithm:** We describe a metric that leverages rare-variants shared between individuals to improve the detection of short IBD segments. We computed IBD segments using existing methods implemented in *Refined IBD* where we enrich the signal using our metric that facilitates the detection of short IBD segments ($< 2\text{cM}$) by explicitly incorporating rare variants. **Results:** To test our new metric, we simulated datasets involving populations with varying divergent time-scales. We show that rare-variant IBD identifies shorter segments with greater confidence and enables the detection of older divergence between populations. As an example, we applied our metric to the Old-Order Amish cohort with known genealogies dating 14 generations back to validate its ability to detect genetic relatedness between distant relatives. This analysis shows that our method increases the accuracy of identifying shorter segments that in turn capture distant relationships. **Conclusions:** We describe a method to enrich the detection of short IBD segments using rare-variant sharing within IBD segments. Leveraging rare-variant sharing improves the information content of short IBD segments better than common variants

44 alone. We validated the method in both simulated and empirical datasets. This method can benefit
45 association analyses, IBD mapping analyses, and demographic inferences.
46

Introduction

Relatedness is the estimation of shared ancestry between individuals and is a fundamental concept in genetics that plays an integral role in many fields ranging from animal breeding, human disease gene mapping, and forensic science. For example, studies involving population structure (Pritchard et al. 2000; Novembre et al. 2008; Biswas et al. 2009; Lawson et al. 2012; Elhaik et al. 2014) make inferences about shared ancestry based on estimates of genetic relatedness. Population structure often closely agrees with geography under the assumption that populations can be partitioned into “islands” of individuals with increased intra-island mating and decreased inter-island migrations (Aistle and Balding 2009). Seminal work by Novembre and colleagues (Novembre et al. 2008) showed that genetic markers mirror geography in Europe allowing for accurate inference of geographic origin using an individual’s DNA.

Failure to estimate cryptic relatedness among study subjects in genetic association studies can inflate statistical evidence for association. Relatedness estimates are thus integral to the performance of genetic association studies that have traditionally been estimated from pedigrees and more recently from genomic data in large populations lacking pedigree information (Purcell et al. 2007; Manichaikul et al. 2010; Thornton et al. 2012). Mixed linear models are commonly used in genetic association studies in order to incorporate a genetic relationship matrix (GRM) that help to prevent false-positive associations due to population structure or relatedness (Kang et al. 2010; Price et al. 2010; Zhang et al. 2010). More recently, commercial products such as those provided by AncestryDNA™, use genetic information to identify recent relatedness between its consumers allowing them to discern distant relatives who may not be represented in their documented pedigrees (Han et al. 2017). Similarly, the field of forensic science has utilized large public repositories to identify relatives of forensic samples since the alleged suspect may share

genetic information inherited from a common ancestor with relatives represented in these repositories (Jobling and Gill 2004; Weir et al. 2006; Kayser and de Knijff 2011).

Genetic relatedness between two individuals can be defined as the probability that their alleles were inherited from a common ancestor, in other words the alleles are identical-by-descent (IBD). Traditionally, relatedness was estimated directly from known pedigrees and was limited by the generational depth of the pedigree. However, datasets from modern sampling procedures are often not accompanied with any pedigree information or have incomplete pedigrees (Wellcome Trust Case Control Consortium 2007; International HapMap Consortium et al. 2007; Auton et al. 2009; Henn et al. 2010; Ralph and Coop 2013; Carlson et al. 2018). With the introduction of high-throughput genotyping and sequencing technologies, the relatedness estimates can be made directly from genotype information without prior information about the genealogy of the samples.

Over the past decade, multiple methods have been developed to estimate relatedness from single nucleotide polymorphism (SNP) information. Software tools such *PLINK* (Purcell et al. 2007; Chang et al. 2015), *KING* (Manichaikul et al. 2010), *REAP* (Thornton et al. 2012), and *PC-Relate* (Conomos et al. 2016) use allele frequencies across multiple loci to estimate the probabilities of sharing zero, one, or two alleles at a given genetic locus corresponding to the number of copies that are in IBD. When summed across multiple loci these estimates can be transformed into a kinship coefficient estimate for a pair of individuals. Allele frequency-based kinship estimates utilize genome-wide averages of single-SNP statistics that do not take the lengths of genomic regions shared between two individuals into account. However, software tools like *Germline* (Gusev et al. 2009), *Beagle* (Browning and Browning 2010), *fastIBD* (Browning and Browning 2011), and *Refined IBD* (Browning and Browning 2013b) detect IBD segments in the genome shared between pairs of individuals that have been used to infer the degree of relatedness

between individuals. More recently, the *KING* software has incorporated IBD-based kinship estimates.

Alternatively, other software tools such as *RelateAdmix* (Moltke and Albrechtsen 2014), *ERSA* (Huff et al. 2011; Li et al. 2014), and *HaploScore* (Durand et al. 2014) combine multiple statistics to improve these estimates of relatedness. *RelateAdmix* uses a maximum likelihood estimator that combines allele frequencies with admixture proportions while *ERSA* utilizes a maximum likelihood method to estimate recent shared ancestry from the number and length of IBD segments. *HaploScore* takes a different approach in order to improve the accuracy of segments detected as IBD by incorporating different error statistics. All of the above methods have high accuracy when detecting first-degree to third-degree relationships (Ramstetter et al. 2017) and haplotype-based methods perform better than allele frequency-based methods (Gattepaille and Jakobsson 2012; Gusev et al. 2012; Palamara et al. 2012). However, other software tools such as *PRIMUS* (Staples et al. 2014) and *PADRE* (Staples et al. 2016) were developed to estimate more distant relationships by reconstructing pedigrees from existing relatedness estimates.

The scale of human sequencing studies has grown exponentially in recent years with the initiation of large-scale projects such as DiscovEHR (Dewey et al. 2016), UK Biobank (Sudlow et al. 2015), Precision Medicine Initiative (Collins and Varmus 2015), TOPMed (Taliun et al. 2019), MVP (Gaziano et al. 2016), and the All of Us Research Program that involve hundreds of thousands of samples. These studies include large number of related individuals between whom cryptic relatedness can be uncovered using their genetic data. For example, nearly 30% of the UK Biobank participants were found to be related (3rd order or closer) to another person in the cohort (Bycroft et al. 2018). Estimates of cryptic relatedness would help in building a better GRM and in turn improve analyses that are sensitive to inaccurate estimates of relationships and incorrect

pedigrees in large-scale genetic association cohorts. A feature of large-scale human sequencing projects is the presence of numerous rare variants with minor allele frequencies as low as 1×10^{-5} . Rare variants have typically arisen as de-novo mutations concurrent with the explosion of population size in recent generations (Keinan and Clark 2012). Their recent origins compared to common variants make them a powerful resource for delineating fine-scale population structure (Baye et al. 2011; Keinan and Clark 2012; Tennessen et al. 2012; O'Connor et al. 2015).

IBD segments resulting from shared ancestry are detected mainly as large segments (2cM and larger) with greater accuracy of detection as the length increases (Chapman and Thompson 2003; Moltke et al. 2011; Gusev et al. 2009; Browning and Browning 2013b; Ralph and Coop 2013). Current haplotype sharing methods for estimation of relatedness use common variants (minor allele frequency $> 5\%$) since phasing of variants into haplotypes is critical to IBD detection and it is difficult to phase rare variants. Hence, these methods have the disadvantage of not using information available from the sharing of rare variants between two individuals. Since the underlying principle of IBD detection is based on the sharing of very low frequency haplotypes, rare variants can be highly informative for IBD detection. While common variants can be shared between two individuals by chance without the presence of a common ancestor, sharing of rare variants between two individuals without a common ancestor is less likely. Hence, shared rare variants provide additional evidence of haplotype sharing that will increase the odds of IBD vs identity-by-state (IBS) due to rarity of co-ancestry and in turn improve the accuracy of IBD detection especially for short segments with lengths less than 2cM.

In the *Algorithm and Implementation* section, we describe a novel method for enriching the detection of haplotype sharing between two individuals by leveraging shared rare variants. In the *Results* section, we evaluate the method in simulated datasets as well as an empirical dataset

146 involving a 14-generation founder population of Old Order Amish (OOA) individuals, 1,100 of
147 whom were sequenced as part of the TOPMed initiative (Mitchell et al. 2008; Taliun et al. 2019).
148 We then describe the assessment of the rare-variant IBD metric (rvIBD) in the simulated dataset
149 and evaluate its performance in detecting higher-order relatedness using prior knowledge about
150 the different degrees of relationships ascertained from the OOA pedigree (Agarwala et al. 1998).

Algorithm and Implementation

We present the rare-variant IBD (rvIBD) metric, which identifies short IBD segments by leveraging shared rare variants between individuals from a large sequencing cohort. The implementation of the rvIBD metric involves three stages ([Supplementary Figure S1](#)). The first stage is the detection of IBD segments using common variants and existing methods such as Refined IBD (Browning and Browning 2013b) with relaxed filtering criteria to allow for the detection of short IBD segments (<2cM). The second stage is the identification of rare variants with minor allele frequency < 1%, that are shared between two samples. The third stage utilizes the combination of rare variant genotype and allele frequency within a Bayesian framework to update the odds of distinguishing IBD from IBS.

Rare-variant IBD (rvIBD) Metric

The first step involves the detection of IBD segments between all pairs of samples from large-scale cohorts that include participants who have been sequenced or densely genotyped. We used Refined IBD (Browning and Browning 2013b), an accurate and computationally efficient algorithm, for IBD segment detection. Refined IBD uses a two-step process that first identifies shared haplotypes greater than a user-defined length threshold and then evaluates the evidence for IBD using a hidden Markov model. Using markers with a minimum minor allele frequency of 5%, we identified IBD segments using default parameters except for lowering the length threshold and the log-odds threshold. The second step involves the identification of rare variants shared between two samples. Excluding singletons, bi-allelic markers are first filtered to retain variants with a maximum minor allele frequency of 1%. The lower bound of the allele count for rare variants could be increased to control for genotyping errors depending on the yield of the different filters. For each IBD segment identified between two samples, we extract the genotypes for the rare variants within the endpoints of the segment. Finally, in the third step, we iterate over the set of

rare variants and estimate the rvIBD metric by enriching the original log-odds (LOD) score from Refined IBD using the following equation based on the Bayesian approach for odds ratios,

$$rvIBD_{ij} = Odds(IBM_{ij}|G^i, G^j) \dots \dots (Equation 1)$$

$$= \prod_{rv \in S_{rv}} \frac{P(G_{rv}^i, G_{rv}^j | IBM_{ij} = 1)}{P(G_{rv}^i, G_{rv}^j | IBM_{ij} = 0)} \times \text{original Odds}(IBM)_{\text{Refined IBD}}$$

where,

S_{rv} = set of rare variants within the endpoints of the IBD segment

G_{rv}^i = genotype for rare variant rv in sample i

G_{rv}^j = genotype for rare variant rv in sample j

IBM_{ij} = IBD status of locus between samples i and j

The probabilities for a pair of genotypes $P(G_{rv}^i, G_{rv}^j | IBM_{ij} = 0 \text{ or } 1)$ given IBD or non-IBD status are summarized in [Figure 1](#). In order to incorporate allele error into the rvIBD metric, we replace the observed minor allele frequency f_B with the corrected minor allele frequency $p_B = (f_B - \epsilon)/(1 - 2\epsilon)$ and corrected major allele frequency $p_A = (1 - p_B)$ where ϵ is the allele error at the biallelic marker with major allele A and minor allele B.

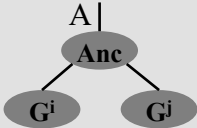
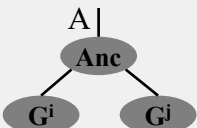
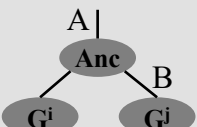
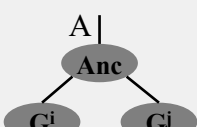
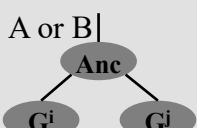
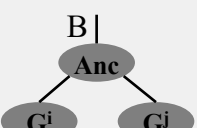
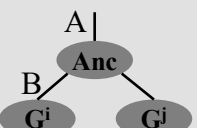
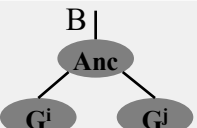
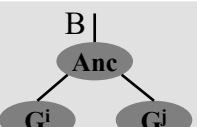
Likely Ancestry Tree	G_{rv}^i	G_{rv}^j	$P(G_{rv}^i, G_{rv}^j IBD_{ij} = 1)$	$P(G_{rv}^i, G_{rv}^j IBD_{ij} = 0)$
	AA	AA	$\lambda_0 p_A^3 + 2\lambda_1 p_A^2 p_B$	f_A^4
	AA	AB	$\lambda_0 p_A^2 p_B + \lambda_1 (p_A p_B + 2p_A^3)$	$2f_A^3 f_B$
	AA	BB	$2g\mu + \lambda_1 p_A p_B$	$f_A^2 f_B^2$
	AB	AA	$\lambda_0 p_A^2 p_B + \lambda_1 (p_A p_B + 2p_A^3)$	$2f_A^3 f_B$
	AB	AB	$\lambda_0 p_A p_B + 4\lambda_1 p_A p_B$	$4f_A^2 f_B^2$
	AB	BB	$\lambda_0 p_A p_B^2 + \lambda_1 (2p_B^3 + p_A p_B)$	$2f_A f_B^3$
	BB	AA	$2g\mu + \lambda_1 p_A p_B$	$f_A^2 f_B^2$
	BB	AB	$\lambda_0 p_A p_B^2 + \lambda_1 (2p_B^3 + p_A p_B)$	$2f_A f_B^3$
	BB	BB	$\lambda_0 p_B^3 + 2\lambda_1 p_A p_B^2$	f_B^4

Figure 1: Probability of rare variant genotype pairs given IBD states

Probabilities for rare-variant genotype pairs with major allele A and minor allele B are computed using corrected allele frequencies given by p_A and p_B respectively and adjusted for allele error ε using error term $\lambda_j = \varepsilon^j (1 - \varepsilon)^{(4-j)}$ for $j \in [0, 1]$ (Browning and Browning 2013a). $2g\mu$ is the probability of mutation since common ancestor g generations ago with a mutation rate μ (Palamara et al. 2015). $g \cong 3/2l$ where l is IBD segment length in Morgans (Baharian et al. 2016).

199

200 We also adjusted the probabilities for genotype pairs for both allele errors and mutations acquired
201 since most recent common ancestor (Palamara et al. 2015; Baharian et al. 2016; Smith et al.
202 2018). Given that rare variants cannot be accurately phased, the rvIBD metric was designed
203 specifically to use the rare variant genotype and the number of rare alleles shared between two
204 individuals to enrich the odds of IBD vs non-IBD. The new rvIBD metric and number of shared
205 rare variants can be used to assess accuracy of IBD detection, especially for short IBD segments.

206

207

Results

In order to assess the performance of the rare-variant IBD metric (rvIBD), we evaluated its ability to detect short IBD segments between populations with older divergence using a simulated dataset. In addition, we applied the rvIBD metric to the Old Order Amish (OOA) cohort to assess detectability of distant relatedness using prior knowledge about the different degrees of relationships ascertained from the OOA pedigree (Agarwala et al. 1998).

Simulated dataset

We simulated a demographic model using the program *msprime* (Kelleher et al. 2016), a coalescent-based simulator. Our estimated demographic model started with an initial population size of 12,300 about 220,000 ya, with an out-of-Africa reduction to 2,100 occurring 140,000 ya (Gutenkunst et al. 2009). This population then diverged into two populations of size 1000 and 510 respectively about $T_{\text{POP2-POP3}}$ generations ago where $T_{\text{POP2-POP3}}$ ranges from 5 generations (recent divergence) to 1000 generations ago (distant divergence) (Supplementary Figure S2). We used a mutation rate of 2×10^{-8} and a recombination rate of 2×10^{-8} (i.e. 1Mb \cong 1cM). For each divergence time, we simulated 20 data sets consisting of 10 Mb segments and 1,500 diploid individuals (500 individuals per population). *msprime* outputs genotypes in VCF format as well as ancestry trees that can be queried to compute time to most recent common ancestor (TMRCA) for any genomic segment between any two individuals. Using markers with a minimum minor allele frequency of 5%, we identified IBD segments using default parameters except for lowering the length threshold to 0.15cM and the log-odds threshold to 1. Each IBD segment detected using *Refined IBD* was leveraged for the presence of shared rare variants (minor allele frequency < 1%) to compute the number of shared rare variants and the enriched rvIBD log-odds (rvLOD) score (see Equation 1). The IBD segments were further segregated in groups either based on the number of rare variants shared or based on the original LOD and enriched rvLOD scores. The

distributions of the segment lengths were compared between groups using the Wilcoxon rank sum test (for 2 groups) and Kruskal-Wallis test (for 3 groups). We also estimated the TMRCA (in generations) for each IBD segment using the ancestry tree generated by msprime for each simulation (Kelleher et al. 2016). The distributions of the TMRCA estimates were also compared between groups based on the LOD and rvLOD scores using the Kruskal-Wallis test.

Shared rare variants capable of tagging short IBD segments

For each simulation based on varying divergence times ranging from 5 to 1000 generations, we first binned the IBD segments into two categories, namely segments with no shared rare variants and segments with at least 1 shared rare variant within the endpoints of the segment. We then compared the distribution of IBD segment length between the two categories ([Figure 2](#); [Supplementary Figure S3](#)).

For all simulations, we see the density of segment lengths skewed towards shorter segments. However, when we focus on short IBD segments ($< 2\text{cM}$), we observe a significant shift in the mode of the distribution wherein IBD segments with shared rare variants are shorter than segments with no presence of shared rare variants. In addition, as we transition from a recent divergence time (i.e. $\sim 750\text{ya}$) to an older divergence time (i.e. $\sim 25,000\text{ya}$), we observe the expected decrease in the range of IBD segment lengths and prominent difference between the two categories. Hence, we conclude that shared rare variants would better tag short IBD segments and allow us to assess older divergence time between populations.

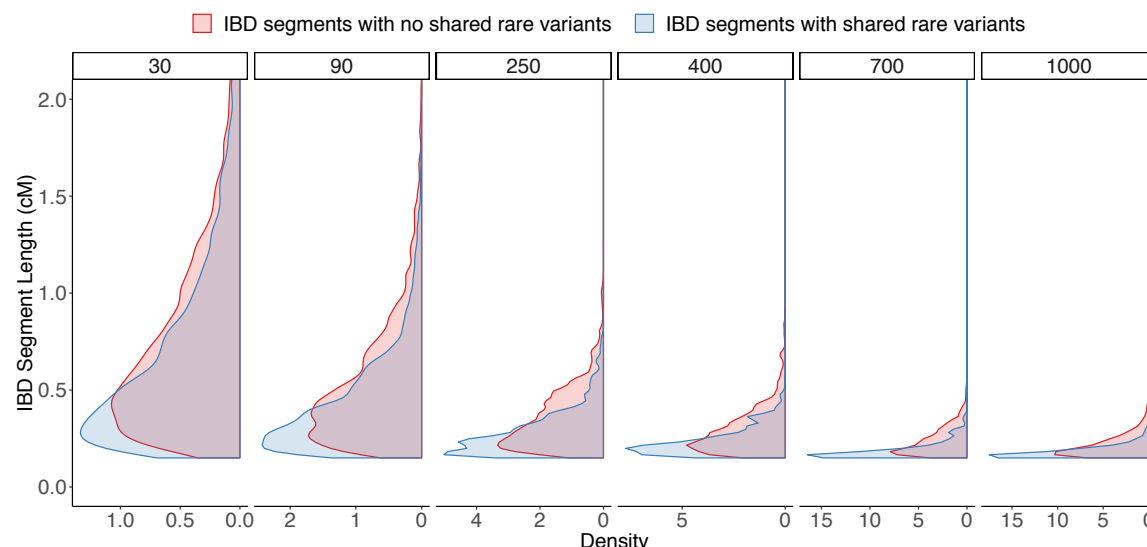


Figure 2: Distribution of IBD segment length grouped by presence/absence of shared rare variants

IBD segment were binned based on presence (blue) or absence (red) of shared rare variants. The distributions of segment length were compared across the two bins. Analysis was repeated for varying divergence times from left to right ranging from recent (30 generations ago) to older (1000 generations ago) divergence between POP₂ and POP₃ (i.e. $T_{\text{POP2-POP3}}$).

Shared rare variants within short IBD segments improve odds of IBD vs IBS

Next, we binned the IBD segments based on the original log-odds (LOD) from Refined IBD and the enriched rvIBD log-odds (rvLOD) into those that satisfy one of $\text{LOD} > 3$ or $\text{rvLOD} > 3$ or both conditions. Since the log-odds is the base 10 logarithm scaled ratio of the likelihood of IBD to the likelihood of non-IBD, a threshold of 3 represents a 1000-fold difference increase in the likelihood of IBD. We then compared the distribution of IBD segment length between the three sets (Figure 3; Supplementary Figure S3). Again, we observe a significant shift (skewed towards shorter segments) in the distribution of segment lengths for segments satisfying the $\text{rvLOD} > 3$ criterion than those that satisfy both or only the $\text{LOD} > 3$ criterion. We also see a decrease in the expected IBD segment lengths as expected and prominent differences between the three categories as we transition from a recent to older divergence time between the two populations.

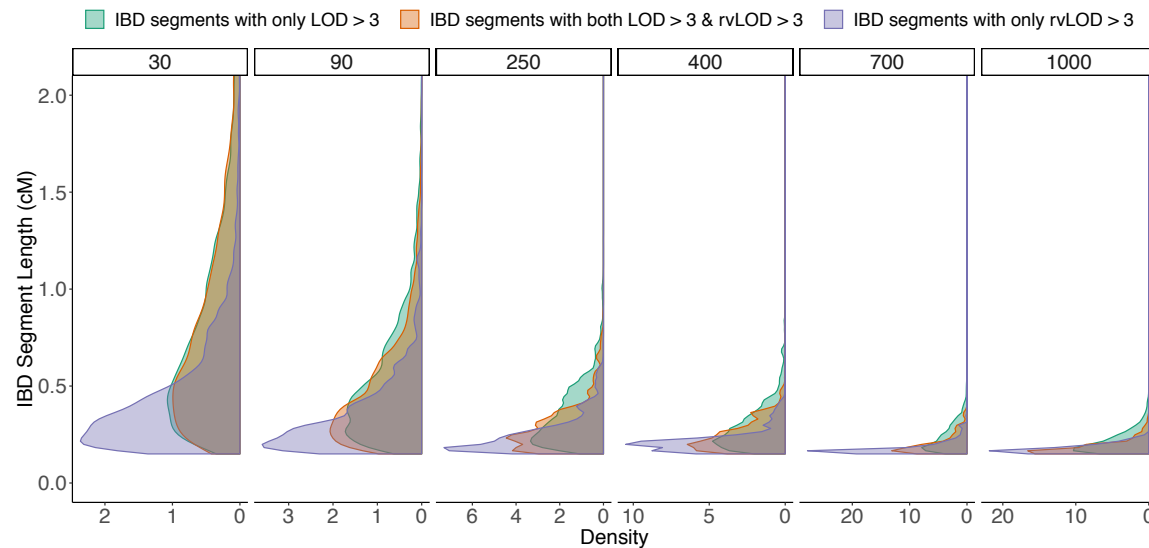


Figure 3: Distribution of IBD segment length categorized by LOD and rvLOD thresholds

IBD segment were binned based on satisfying one or both thresholds of $\text{LOD} > 3$ and/or $\text{rvLOD} > 3$. The distributions of segment length were compared across the three bins. Analysis was repeated for varying divergence times from left to right ranging from recent (30 generations ago) to older (1000 generations ago) divergence between POP_2 and POP_3 (i.e. $T_{\text{POP}_2\text{-POP}_3}$).

Short IBD segments leveraged by shared rare variants identify older relatedness

We compared the distribution of TMRCA estimates for the same three categories segregated based on LOD and rvLOD scores, as described above (Figure 4; Supplementary Figure S3). Since IBD segments are broken down over generations, the expected TMRCA estimates should be older as the IBD segment length decreases.

For each of the simulated datasets, we do observe that the distribution of TMRCA estimates agrees with the divergence time utilized to simulate the dataset. While the distribution of the TMRCA estimates for segments with only $\text{LOD} > 3$ peaks close to the divergence time, we see that the distribution of the TMRCA estimates for segments with $\text{rvLOD} > 3$ extend to estimates

even older than the divergence time detecting inheritance of IBD segments from common ancestors before the separation of the two populations.

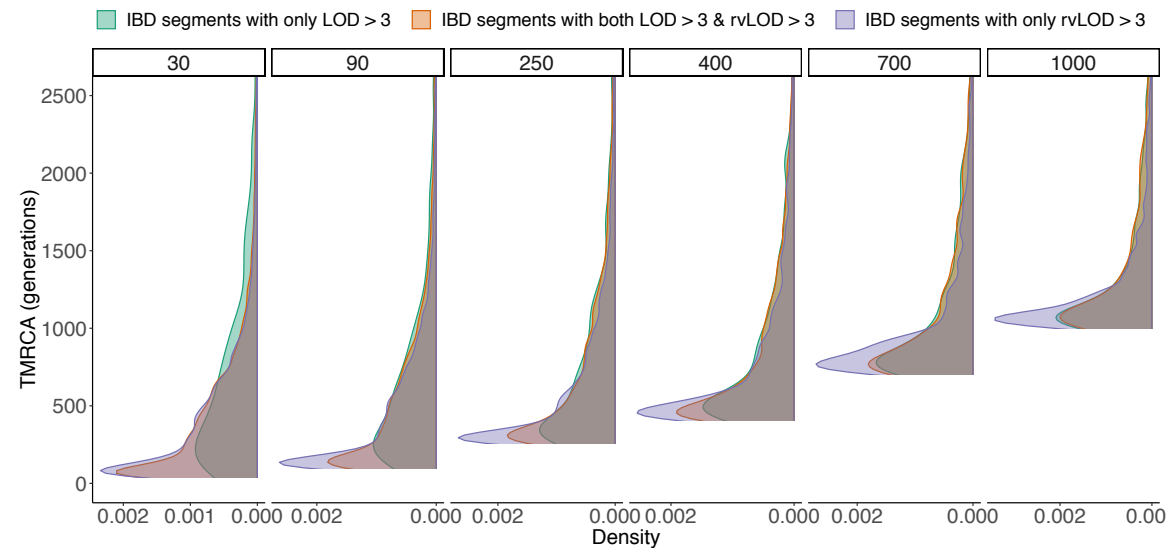


Figure 4: Distribution of TMRCA estimates categorized by LOD and rvLOD thresholds

The time to most recent common ancestor (TMRCA) was estimated for each IBD segment using the ancestry trees generated by *msprime*. IBD segment were binned based on satisfying one or both thresholds of LOD > 3 and/or rvLOD > 3. The distributions of TMRCA estimates were compared across the three bins. Analysis was repeated for varying divergence times from left to right ranging from recent (30 generations ago) to older (1000 generations ago) divergence between POP₂ and POP₃ (i.e. $T_{POP2-POP3}$).

This demonstrates that the new rvIBD metric is not only able to enrich the detection of short IBD segments but also identify IBD segments corresponding to older relatedness as simulated by the different divergence times.

Empirical dataset

In order to assess the application of the rvIBD metric to detect distant relatedness, we used the high-coverage whole genome sequencing data from freeze3 dataset of the Trans-Omics for

Precision Medicine (TOPMed) Project (Taliun et al. 2019) that includes ~1,100 individuals from the Old Order Amish (OOA) cohort in Lancaster County, Pennsylvania. The OOA population immigrated to the Colonies from Western Europe in the early 1700's to escape religious persecution (Cross 1976). There are now over 38,000 Amish individuals in the Lancaster County, nearly all of whom can trace their ancestry back 12-14 generations to approximately 200 founders (McKusick 1978; Beiler 1988; Lee et al. 2010). The complex multi-generational pedigree allows us to assess genetic sharing between 1st and 13th order related individual pairs. The inclusion of both closely-related (less than 3rd order) and higher order relatedness enables us to assess the accuracy of the rvIBD metric for different orders of relatedness.

Using markers with a minimum minor allele frequency of 5%, we identified IBD segments using default parameters except for lowering the length threshold to 0.5cM and the log-odds threshold to 1. For the assessment of the rvIBD metric, we filtered the possible ~600,000 pairs of individuals to retain only individual-pairs with non-consanguineous common ancestors as inferred from the pedigree structure. Each IBD segment detected using *Refined IBD* was leveraged for the presence of shared rare variants (minor allele frequency < 1%) to compute the number of shared rare variants and the enriched rvIBD log-odds (rvLOD) score.

For each pair of individuals, we filtered the set of IBD segments based on the different segment length thresholds, thresholds for LOD score and thresholds for rvLOD scores. The observed IBD proportion for each pair was computed as the ratio of the cumulative length of IBD segments (in cM) retained after the filtering process to the length of the genome (in cM). For each pair of individuals, we also estimated the order of relatedness inferred from the number of generations to their most recent common ancestor in the pedigree. The expected IBD proportion was computed as described by Manichaikul et. al (2010). Further, we estimated the precision of

detection of short IBD segments by comparing IBD segments shared between distant cousins to the IBD segments shared between one of the cousins and the parents/grandparents of the other cousin, when available. The precision was calculated for different segment bins and assessed for different orders of relatedness.

Cryptic genetic relatedness in the Older Order Amish cohort

Overall, in the OOA dataset, we illustrate large estimates of IBD sharing between individuals expected due to the cohort originating from a founder population. We compared the observed IBD proportions to the expected IBD proportions after using different sets of thresholds for IBD segment length, original LOD score and rvLOD score based on the rvIBD algorithm ([Figure 5](#); [Supplementary Figure S4](#)).

When using a minimum LOD score of 3 and minimum segment length of 5cM ([Figure 5A](#)), we observe more than the expected IBD sharing in 90% of pairs for relatedness order ranging from 1st to 9th degree relatives with no IBD sharing estimated in the seemingly unrelated individuals. When we decrease the segment length threshold to 2cM ([Figure 5B](#)), we observe similar results with IBD sharing among some unrelated individuals while a lenient length threshold of 0.5cM ([Figure 5C](#)) results in IBD sharing detected in over 90% pairs across all orders of relatedness as well as over 80% of the seemingly unrelated pairs. This sizeable increase in observed IBD sharing proportions could be caused by the introduction of false positive IBD segments ranging from 0.5cM to 2cM in length.

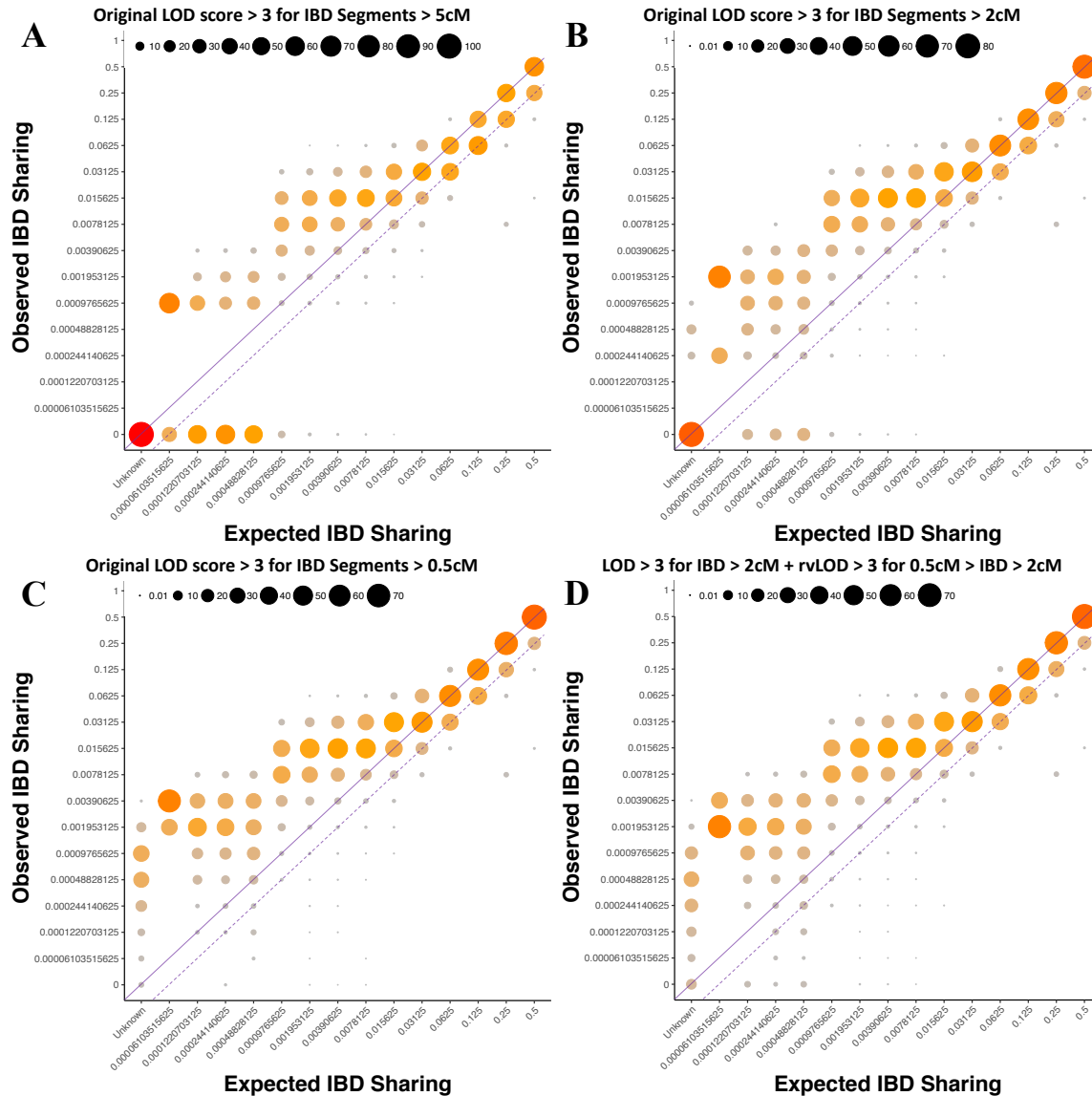


Figure 5: Performance of relatedness estimation using different sets of IBD segment thresholds

Comparisons of observed IBD proportion using genetic data was made to the expected IBD proportion estimated from the pedigree for different degrees of relatedness using (A) IBD segments > 5cM with LOD > 3, (B) IBD segments > 2cM with LOD > 3, (C) IBD segments > 0.5cM with LOD > 3, and (D) IBD segments > 2cM with LOD > 3 and IBD segments < 2cM with rvLOD > 3. Area of the circles indicates the percentage of individual pairs with specific proportions of observed IBD sharing segregated by the estimated degrees of relationship. The solid line represents bins where the observed IBD proportions were comparable to the expected IBD proportions while the dotted line represents a lower threshold used to estimate the proportion of individual pairs with detectable relatedness based on genetic IBD sharing illustrated in Supplementary Figure S4.

In order to reduce these false positives, we implemented a composite threshold of $LOD > 3$ for segment lengths greater than 2cM and $rvLOD > 3$ for segment lengths less than 2cM ([Figure 5D](#)) and observed IBD sharing in 90% of pairs for relatedness order ranging from 1st to 9th degree relatives with greater than 10% increase in the detection rate of 5th cousins (10th order) and more distant relatedness. IBD sharing in the seemingly unrelated individuals rose from a single individual sharing IBD with other individuals. Detection of possible cryptic relatedness among the seemingly unrelated individuals with potential common ancestors older than 14 generations ago cannot be assessed due to limited pedigree information and lack of knowledge regarding the relatedness among founders of the Old-Order Amish community.

Lastly, we assessed the precision of detecting short IBD segments using the original LOD estimates and the $rvLOD$ estimates. For each degree of relationships ranging from 4th to 13th order relatedness, we segregated the IBD segments by length into 7 groups. For a given IBD segment between a pair of cousins, we assessed if an overlapping segment was also detected between one of the cousins and the parents or grandparents of the other cousin under the premise that the genomic region was inherited from one of the two parents or one of the 4 grandparents, when present in the sequenced OOA cohort. We considered the IBD to be true if overlapping IBD segments were detected and false if overlapping segments were not found. The precision of detecting IBD segments of a particular length was computed based on the count of true IBD and false IBD segments. The count of false IBD segments was scaled when full pedigree information was unavailable. We repeated the analysis for the different thresholds defined in [Figure 5](#). We show an improvement in the precision of detecting short segments ($< 2cM$) detected using the $rvIBD$ metric over the original IBD callset ([Supplementary Figure S5](#)). The precision of shorter segments based on $rvIBD$ is comparable to the precision of longer segments using the original IBD callset which are already known to have higher accuracy as demonstrated by previous work

394 in the field (Browning and Browning 2010; Genovese et al. 2010; Browning and Browning 2012;
395 Browning and Browning 2013b).

396

Discussion

Here we have described an approach, r_lIBD, for leveraging the presence of shared rare-variants between two individuals to better assess their genetic relatedness. Genetic relatedness is a fundamental concept in population genetics and genealogical studies and traditionally described as the probability of sharing alleles that are identical-by-descent from common ancestors in a pedigree. Identity-by-descent is the phenomenon whereby two individuals inherit a genomic region from a common ancestor without an intervening detectable recombination event (Weir et al. 2006). With the onset of dense genotyping and next-generation sequencing, genetic relatedness has been computed directly from single nucleotide polymorphism (SNP) data using maximum likelihood estimates (Milligan 2003). Using averages of single SNP-based estimates, matrices of genetic relatedness have been computed and incorporated in to genome wide association studies (GWAS) (Forni et al. 2011; Yang et al. 2011; Yang et al. 2013; Chang et al. 2015). However, single SNP-based estimates do not take into account linkage disequilibrium (LD), discard SNPs based on minor allele frequency thresholds, and ignore the length of shared genomic regions between two individuals.

More recently, methods detecting haplotype sharing have been developed (Gusev et al. 2009; Browning and Browning 2010; Browning and Browning 2013b) and utilize more sophisticated methods involving hidden Markov models to differentiate IBD from IBS, which have been shown to perform better (Ramstetter et al. 2017). We specifically focus on the detection of short IBD segments (<2cM) that have been previously unreliable and are more indicative of older co-ancestry between individuals. We demonstrated through our simulations of populations with varying divergence times, the capability of rare variants to better tag short IBD segments over common variants and subsequently improve the assessment of IBD vs IBS.

In the last decade, there has been an increasing focus on the existence of cryptic relatedness in both isolated and cosmopolitan genetic samples (Voight and Pritchard 2005; Pemberton et al. 2010; Henn et al. 2012; Fedorova et al. 2016; Staples et al. 2018). Cryptic relatedness refers to the presence of relatives in a sample of seemingly unrelated individuals. Voight and Pritchard showed that cryptic relatedness has a significant inflation of association p-values from genetic association studies involving small and isolated populations resulting in false positives. First, Pemberton and colleagues were able to infer previously unidentified relationships in phase III of the HapMap Project (Pemberton et al. 2010). Later, Henn and colleagues detected thousands of higher-order relationships in a heterogeneous European population (Henn et al. 2012). Recently, Staples and colleagues highlighted the prevalence of cryptic relatedness in large healthcare studies (Staples et al. 2018).

Through our analyses of simulated data, we show the ability of detecting short IBD segments by leveraging shared rare variants. We demonstrated that these short segments when enriched with shared rare variants are indicative of older relatedness between individuals from populations that diverged many thousands of years in the past. Furthermore, using the empirical dataset from the multi-generational Older Order Amish cohort, we illustrate the improvement in the detection of higher-order relationships greater than fifth-cousins after the inclusion of the rvIBD metric for short IBD segments. Using the short segments leveraged by shared rare variants, we were also able to identify genetic relatedness between the seemingly unrelated individuals which will need further investigation. If left undetected, these higher-order relationships contribute to the prevalence of confounding due to cryptic relatedness in large healthcare studies.

Recent large-scale human sequencing projects have identified tens of millions of variants (1000 Genomes Project Consortium et al. 2012; 1000 Genomes Project Consortium et al. 2015; Sudlow

et al. 2015; Dewey et al. 2016; Gaziano et al. 2016; Taliun et al. 2019), most of which are rare variants (minor allele frequency < 1%). Common variants are often shared by chance with no underlying IBD segments. Based on the infinite-site model and the low rates of mutation, the possibility of a rare variant shared between two individuals without common ancestry is less likely. Hence, the sharing of rare variants is more indicative of co-ancestry than the sharing of common variants. Recent explosive population growth has resulted in a deluge of rare genetic variants due to the increased mutational capacity of recent human populations (Coventry et al. 2010; Keinan and Clark 2012; Fu et al. 2013). Fu and colleagues showed that a majority of the variants are predicted to have arisen in the last 10,000 to 50,000 years (Fu et al. 2013). Work by O'Connor and colleagues further demonstrated that rare variants have more information content than common variants when making inferences of fine-scale population structure since they can accurately detect recent demographic changes in the last 25,000 years (O'Connor et al. 2015). Other works focused on rare variants also showed the ability to identify distant relatedness within and between cohorts (Hochreiter 2013; Mathieson and McVean 2014; Al-Khudhair et al. 2015).

Overall, we show that leveraging shared rare variants, which were previously ignored, allows for the detection of older co-ancestry. Detection of older co-ancestry will help identify distant relatedness between populations and will provide new insights in to population structure using rare variants. In addition, detection of older co-ancestry will also help identify cryptic relatedness in large-scale studies involving cohorts ranging from ten to hundred thousand individuals which if left unidentified could lead to spurious results in downstream association analyses. Indirectly, the detection of shorter IBD segments would augment other applications for IBD segments involving the estimation of population demographic parameters, inferring fine-scale population structure, detecting migratory events, IBD mapping and forensic science (Palamara et al. 2012; Browning and Browning 2012; Thompson 2013; Palamara and Pe'er 2013; Palamara et al. 2015).

Acknowledgements

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish" (phs000956.v1.p1) was performed at Broad Institute of MIT and Harvard (3R01HL121007-01S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. A full list of TOPMed collaborators can be found at <https://www.nhlbiwgs.org/topmed-banner-authorship>. The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, and R01 AG18728.

References

- 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422), pp. 56–65.
- 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. 2015. A global reference for human genetic variation. *Nature* 526(7571), pp. 68–74.
- Agarwala, R., Biesecker, L.G., Hopkins, K.A., Francomano, C.A. and Schaffer, A.A. 1998. Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. *Genome Research* 8(3), pp. 211–221.
- Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Saha-Mandal, A., Dutta, R., Fedorova, L. and Fedorov, A. 2015. Inference of distant genetic relations in humans using “1000 genomes”. *Genome Biology and Evolution* 7(2), pp. 481–492.
- Astle, W. and Balding, D.J. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* 24(4), pp. 451–471.
- Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N., King, K.S., Nelson, M.R. and Bustamante, C.D. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research* 19(5), pp. 795–803.
- Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C. and Gravel, S. 2016. The Great Migration and African-American Genomic Diversity. *PLoS Genetics* 12(5), p. e1006059.
- Baye, T.M., He, H., Ding, L., Kurowski, B.G., Zhang, X. and Martin, L.J. 2011. Population structure

510 analysis using rare and common functional variants. *BMC Proceedings* 5 Suppl 9, p. S8.

511 Beiler, K. 1988. *Fisher Family History: Descendants and History of Christian Fisher 1757-1838*. Pequea
512 Bruderschaft Library.

513 Biswas, S., Scheinfeldt, L.B. and Akey, J.M. 2009. Genome-wide insights into the patterns and
514 determinants of fine-scale population structure in humans. *American Journal of Human Genetics* 84(5),
515 pp. 641–650.

516 Browning, B.L. and Browning, S.R. 2011. A fast, powerful method for detecting identity by descent.
517 *American Journal of Human Genetics* 88(2), pp. 173–182.

518 Browning, B.L. and Browning, S.R. 2013a. Detecting identity by descent and estimating genotype error
519 rates in sequence data. *American Journal of Human Genetics* 93(5), pp. 840–851.

520 Browning, B.L. and Browning, S.R. 2013b. Improving the accuracy and efficiency of identity-by-descent
521 detection in population data. *Genetics* 194(2), pp. 459–471.

522 Browning, S.R. and Browning, B.L. 2010. High-resolution detection of identity by descent in unrelated
523 individuals. *American Journal of Human Genetics* 86(4), pp. 526–539.

524 Browning, S.R. and Browning, B.L. 2012. Identity by descent between distant relatives: detection and
525 applications. *Annual Review of Genetics* 46, pp. 617–633.

526 Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D.,
527 Delaneau, O., O’Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S.,
528 Allen, N., Donnelly, P. and Marchini, J. 2018. The UK Biobank resource with deep phenotyping and
529 genomic data. *Nature* 562(7726), pp. 203–209.

530 Carlson, J., Weeks, D., Hawley, N.L., Sun, G., Chen, H., Naseri, T., Reupena, M.S., Deka, R., McGarvey,
531 S.T. and Minster, R.L. 2018. Genome-wide association studies in Samoans give insight into the genetic
532 etiology of fasting serum lipid levels. *BioRxiv*.

533 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. 2015. Second-generation
534 PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4, p. 7.

535 Chapman, N.H. and Thompson, E.A. 2003. A model for the length of tracts of identity by descent in finite
536 random mating populations. *Theoretical Population Biology* 64(2), pp. 141–150.

537 Collins, F.S. and Varmus, H. 2015. A new initiative on precision medicine. *The New England Journal of*
538 *Medicine* 372(9), pp. 793–795.

539 Conomos, M.P., Reiner, A.P., Weir, B.S. and Thornton, T.A. 2016. Model-free Estimation of Recent
540 Genetic Relatedness. *American Journal of Human Genetics* 98(1), pp. 127–148.

541 Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J.,
542 Muzny, D.M., Lewis, L.R., Wheeler, D.A., Sabo, A., Lusk, C., Weiss, K.G., Akbar, H., Cree, A.,
543 Hawes, A.C., Newsham, I., Varghese, R.T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan,
544 M., Chang, K., Iv, W.H., Templeton, A.R., Boerwinkle, E., Gibbs, R. and Sing, C.F. 2010. Deep
545 resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature*
546 *Communications* 1, p. 131.

547 Cross, H.E. 1976. Population studies and the Old Order Amish. *Nature* 262(5563), pp. 17–20.

548 Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Carey, D.J., et al. 2016. Distribution and clinical
549 impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*
550 354(6319).

551 Durand, E.Y., Eriksson, N. and McLean, C.Y. 2014. Reducing pervasive false-positive identical-by-descent
552 segments detected by large-scale pedigree analysis. *Molecular Biology and Evolution* 31(8), pp. 2212–
553 2222.

554 Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I.S., Maria Calò, C., De Montis, A., Atzori, M., Marini,
555 M., Tofanelli, S., Francalacci, P., Pagani, L., Tyler-Smith, C., Xue, Y., Cucca, F., Schurr, T.G., Gaieski,
556 J.B., Melendez, C., Vilar, M.G., Owings, A.C., Gómez, R., Fujita, R., Santos, F.R., Comas, D.,

557 Balanovsky, O., Balanovska, E., Zalloua, P., Soodyall, H., Pitchappan, R., Ganeshprasad, A., Hammer,
 558 M., Matisoo-Smith, L., Wells, R.S. and Genographic Consortium 2014. Geographic population
 559 structure analysis of worldwide human populations infers their biogeographical origins. *Nature*
 560 *Communications* 5, p. 3513.

561 Fedorova, L., Qiu, S., Dutta, R. and Fedorov, A. 2016. Atlas of cryptic genetic relatedness among 1000
 562 human genomes. *Genome Biology and Evolution* 8(3), pp. 777–790.

563 Forni, S., Aguilar, I. and Misztal, I. 2011. Different genomic relationship matrices for single-step analysis
 564 using phenotypic, pedigree and genomic information. *Genetics, selection, evolution : GSE* 43, p. 1.

565 Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler,
 566 D., Shendure, J., Nickerson, D.A., Bamshad, M.J., NHLBI Exome Sequencing Project and Akey, J.M.
 567 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.
 568 *Nature* 493(7431), pp. 216–220.

569 Gattepaille, L.M. and Jakobsson, M. 2012. Combining markers into haplotypes can improve population
 570 structure inference. *Genetics* 190(1), pp. 159–174.

571 Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J.,
 572 Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa,
 573 K., Moser, J., Huang, G., Muralidhar, S., Przygodzki, R. and O'Leary, T.J. 2016. Million Veteran
 574 Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical*
 575 *Epidemiology* 70, pp. 214–223.

576 Genovese, G., Leibon, G., Pollak, M.R. and Rockmore, D.N. 2010. Improved IBD detection using
 577 incomplete haplotype information. *BMC Genetics* 11, p. 58.

578 Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M. and Pe'er, I.
 579 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Research* 19(2), pp.
 580 318–326.

581 Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P. and Pe'er, I. 2012. The
582 architecture of long-range haplotypes shared within and across populations. *Molecular Biology and*
583 *Evolution* 29(2), pp. 473–486.

584 Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. and Bustamante, C.D. 2009. Inferring the joint
585 demographic history of multiple populations from multidimensional SNP frequency data. *PLoS*
586 *Genetics* 5(10), p. e1000695.

587 Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermay, A.R., Myres,
588 N.M., Barber, M.J., Rand, K.A., Song, S., Roman, T., Battat, E., Elyashiv, E., Guturu, H., Hong, E.L.,
589 Chahine, K.G. and Ball, C.A. 2017. Clustering of 770,000 genomes reveals post-colonial population
590 structure of North America. *Nature Communications* 8, p. 14238.

591 Henn, B.M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. and Bustamante, C.D. 2010. Fine-scale
592 population structure and the era of next-generation sequencing. *Human Molecular Genetics* 19(R2),
593 pp. R221-6.

594 Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I. and Mountain, J.L. 2012.
595 Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *Plos One* 7(4),
596 p. e34267.

597 Hochreiter, S. 2013. HapFABIA: identification of very short segments of identity by descent characterized
598 by rare variants in large sequencing data. *Nucleic Acids Research* 41(22), p. e202.

599 Huff, C.D., Witherspoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., Zhang, Y., Tuohy, T.M., Neklason,
600 D.W., Burt, R.W., Guthery, S.L., Woodward, S.R. and Jorde, L.B. 2011. Maximum-likelihood
601 estimation of recent shared ancestry (ERSA). *Genome Research* 21(5), pp. 768–774.

602 International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., et al., et al. 2007. A second
603 generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164), pp. 851–861.

604 Jobling, M.A. and Gill, P. 2004. Encoded evidence: DNA in forensic analysis. *Nature Reviews. Genetics*

5(10), pp. 739–751.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C. and Eskin, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42(4), pp. 348–354.

Kayser, M. and de Knijff, P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews. Genetics* 12(3), pp. 179–192.

Keinan, A. and Clark, A.G. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082), pp. 740–743.

Kelleher, J., Etheridge, A.M. and McVean, G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* 12(5), p. e1004842.

Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics* 8(1), p. e1002453.

Lee, W.-J., Pollin, T.I., O’Connell, J.R., Agarwala, R. and Schäffer, A.A. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. *BMC Medical Genetics* 11, p. 68.

Li, H., Glusman, G., Hu, H., Shankaracharya, Caballero, J., Hubley, R., Witherspoon, D., Guthery, S.L., Mauldin, D.E., Jorde, L.B., Hood, L., Roach, J.C. and Huff, C.D. 2014. Relationship estimation from whole-genome sequence data. *PLoS Genetics* 10(1), p. e1004144.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22), pp. 2867–2873.

Mathieson, I. and McVean, G. 2014. Demography and the age of rare variants. *PLoS Genetics* 10(8), p. e1004528.

McKusick, V.A. 1978. *Medical Genetic Studies of the Amish: Selected Papers*. John Hopkins University

628 Press.

629 Milligan, B.G. 2003. Maximum-likelihood estimation of relatedness. *Genetics* 163(3), pp. 1153–1167.

630 Mitchell, B.D., McArdle, P.F., Shen, H., Rampersaud, E., Pollin, T.I., Bielak, L.F., Jaquish, C., Douglas,
631 J.A., Roy-Gagnon, M.-H., Sack, P., Naglieri, R., Hines, S., Horenstein, R.B., Chang, Y.-P.C., Post, W.,
632 Ryan, K.A., Brereton, N.H., Pakyz, R.E., Sorkin, J., Damcott, C.M., O’Connell, J.R., Mangano, C.,
633 Corretti, M., Vogel, R., Herzog, W., Weir, M.R., Peyser, P.A. and Shuldiner, A.R. 2008. The genetic
634 response to short-term interventions affecting cardiovascular function: rationale and design of the
635 Heredity and Phenotype Intervention (HAPI) Heart Study. *American Heart Journal* 155(5), pp. 823–
636 828.

637 Moltke, I. and Albrechtsen, A. 2014. RelateAdmix: a software tool for estimating relatedness between
638 admixed individuals. *Bioinformatics* 30(7), pp. 1027–1028.

639 Moltke, I., Albrechtsen, A., Hansen, T.V.O., Nielsen, F.C. and Nielsen, R. 2011. A method for detecting
640 IBD regions simultaneously in multiple individuals--with applications to disease genetics. *Genome*
641 *Research* 21(7), pp. 1168–1180.

642 Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann,
643 S., Nelson, M.R., Stephens, M. and Bustamante, C.D. 2008. Genes mirror geography within Europe.
644 *Nature* 456(7218), pp. 98–101.

645 O’Connor, T.D., Fu, W., NHLBI GO Exome Sequencing Project, ESP Population Genetics and Statistical
646 Analysis Working Group, Emily Turner, Mychaleckyj, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal,
647 S.M., Smith, J.D., Rieder, M.J., Bamshad, M.J., Nickerson, D.A. and Akey, J.M. 2015. Rare variation
648 facilitates inferences of fine-scale population structure in humans. *Molecular Biology and Evolution*
649 32(3), pp. 653–660.

650 Palamara, P.F., Francioli, L.C., Wilton, P.R., Genovese, G., Gusev, A., Finucane, H.K., Sankararaman, S.,
651 Genome of the Netherlands Consortium, Sunyaev, S.R., de Bakker, P.I.W., Wakeley, J., Pe’er, I. and

652 Price, A.L. 2015. Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion
 653 Rates. *American Journal of Human Genetics* 97(6), pp. 775–789.

654 Palamara, P.F., Lencz, T., Darvasi, A. and Pe'er, I. 2012. Length distributions of identity by descent reveal
 655 fine-scale demographic history. *American Journal of Human Genetics* 91(5), pp. 809–822.

656 Palamara, P.F. and Pe'er, I. 2013. Inference of historical migration rates via haplotype sharing.
 657 *Bioinformatics* 29(13), pp. i180-8.

658 Pemberton, T.J., Wang, C., Li, J.Z. and Rosenberg, N.A. 2010. Inference of unexpected genetic relatedness
 659 among individuals in HapMap Phase III. *American Journal of Human Genetics* 87(4), pp. 457–464.

660 Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. 2010. New approaches to population stratification in
 661 genome-wide association studies. *Nature Reviews. Genetics* 11(7), pp. 459–463.

662 Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus
 663 genotype data. *Genetics* 155(2), pp. 945–959.

664 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de
 665 Bakker, P.I.W., Daly, M.J. and Sham, P.C. 2007. PLINK: a tool set for whole-genome association and
 666 population-based linkage analyses. *American Journal of Human Genetics* 81(3), pp. 559–575.

667 Ralph, P. and Coop, G. 2013. The geography of recent genetic ancestry across Europe. *PLoS Biology* 11(5),
 668 p. e1001555.

669 Ramstetter, M.D., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Mezey, J.G. and
 670 Williams, A.L. 2017. Benchmarking Relatedness Inference Methods with Genome-Wide Data from
 671 Thousands of Relatives. *Genetics* 207(1), pp. 75–82.

672 Smith, J., Coop, G., Stephens, M. and Novembre, J. 2018. Estimating time to the common ancestor for a
 673 beneficial allele. *Molecular Biology and Evolution* 35(4), pp. 1003–1017.

674 Staples, J., Maxwell, E.K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C., Hawes, A., Penn, J., Ulloa, R.,

675 Bai, X., Lopez, A.E., Van Hout, C.V., O'Dushlaine, C., Teslovich, T.M., McCarthy, S.E.,
 676 Balasubramanian, S., Kirchner, H.L., Leader, J.B., Murray, M.F., Ledbetter, D.H., Shuldiner, A.R.,
 677 Yancoupolos, G.D., Dewey, F.E., Carey, D.J., Overton, J.D., Baras, A., Habegger, L. and Reid, J.G.
 678 2018. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. *American*
 679 *Journal of Human Genetics* 102(5), pp. 874–889.

680 Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., University of Washington Center for Mendelian
 681 Genomics, Nickerson, D.A. and Below, J.E. 2014. PRIMUS: rapid reconstruction of pedigrees from
 682 genome-wide estimates of identity by descent. *American Journal of Human Genetics* 95(5), pp. 553–
 683 564.

684 Staples, J., Witherspoon, D.J., Jorde, L.B., Nickerson, D.A., University of Washington Center for
 685 Mendelian Genomics, Below, J.E. and Huff, C.D. 2016. PADRE: Pedigree-Aware Distant-Relationship
 686 Estimation. *American Journal of Human Genetics* 99(1), pp. 154–162.

687 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J.,
 688 Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T.
 689 and Collins, R. 2015. UK biobank: an open access resource for identifying the causes of a wide range
 690 of complex diseases of middle and old age. *PLoS Medicine* 12(3), p. e1001779.

691 Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Abecasis, G.R., et al. 2019. Sequencing of 53,831
 692 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*.

693 Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu,
 694 X., Jun, G., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D.,
 695 Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J., Akey, J.M., Broad
 696 GO, Seattle GO and NHLBI Exome Sequencing Project 2012. Evolution and functional impact of rare
 697 coding variation from deep sequencing of human exomes. *Science* 337(6090), pp. 64–69.

698 Thompson, E.A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations.

699 *Genetics* 194(2), pp. 301–326.

700 Thornton, T., Tang, H., Hoffmann, T.J., Ochs-Balcom, H.M., Caan, B.J. and Risch, N. 2012. Estimating
701 kinship in admixed populations. *American Journal of Human Genetics* 91(1), pp. 122–138.

702 Voight, B.F. and Pritchard, J.K. 2005. Confounding from cryptic relatedness in case-control association
703 studies. *PLoS Genetics* 1(3), p. e32.

704 Weir, B.S., Anderson, A.D. and Hepler, A.B. 2006. Genetic relatedness analysis: modern data and new
705 challenges. *Nature Reviews. Genetics* 7(10), pp. 771–780.

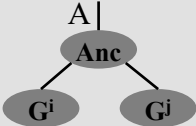
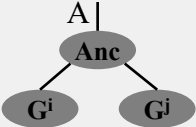
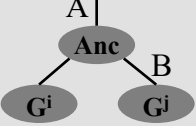
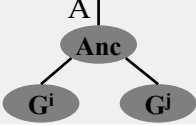
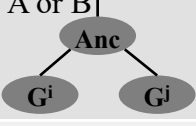
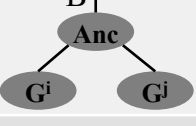
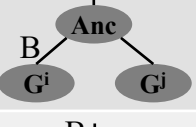
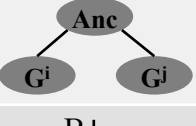
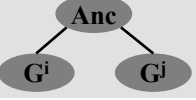
706 Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven
707 common diseases and 3,000 shared controls. *Nature* 447(7145), pp. 661–678.

708 Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. 2011. GCTA: a tool for genome-wide complex trait
709 analysis. *American Journal of Human Genetics* 88(1), pp. 76–82.

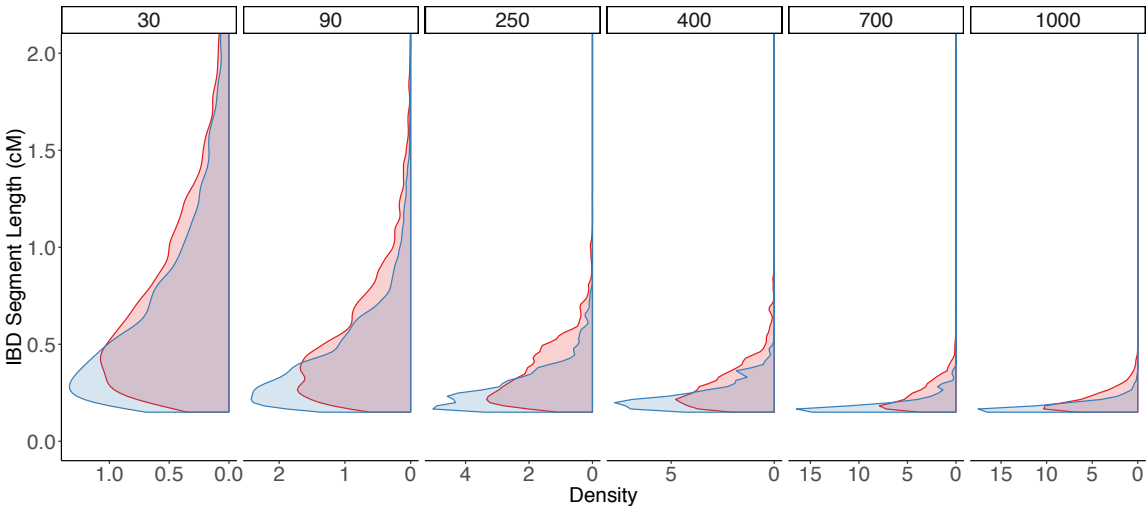
710 Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. 2013. Genome-wide complex trait analysis (GCTA):
711 methods, data analyses, and interpretations. *Methods in Molecular Biology* 1019, pp. 215–236.

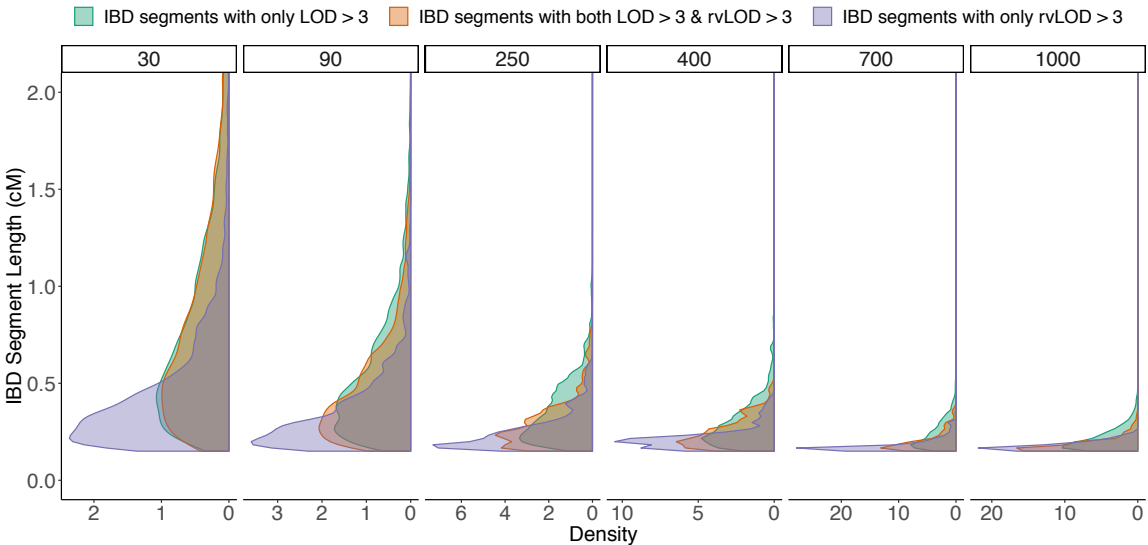
712 Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett,
713 D.K., Ordovas, J.M. and Buckler, E.S. 2010. Mixed linear model approach adapted for genome-wide
714 association studies. *Nature Genetics* 42(4), pp. 355–360.

715

Likely Ancestry Tree	G_{rv}^i	G_{rv}^j	$P(G_{rv}^i, G_{rv}^j IBD_{ij} = 1)$	$P(G_{rv}^i, G_{rv}^j IBD_{ij} = 0)$
	AA	AA	$\lambda_0 p_A^3 + 2\lambda_1 p_A^2 p_B$	f_A^4
	AA	AB	$\lambda_0 p_A^2 p_B + \lambda_1 (p_A p_B + 2p_A^3)$	$2f_A^3 f_B$
	AA	BB	$2g\mu + \lambda_1 p_A p_B$	$f_A^2 f_B^2$
	AB	AA	$\lambda_0 p_A^2 p_B + \lambda_1 (p_A p_B + 2p_A^3)$	$2f_A^3 f_B$
	AB	AB	$\lambda_0 p_A p_B + 4\lambda_1 p_A p_B$	$4f_A^2 f_B^2$
	AB	BB	$\lambda_0 p_A p_B^2 + \lambda_1 (2p_B^3 + p_A p_B)$	$2f_A f_B^3$
	BB	AA	$2g\mu + \lambda_1 p_A p_B$	$f_A^2 f_B^2$
	BB	AB	$\lambda_0 p_A p_B^2 + \lambda_1 (2p_B^3 + p_A p_B)$	$2f_A f_B^3$
	BB	BB	$\lambda_0 p_B^3 + 2\lambda_1 p_A p_B^2$	f_B^4

IBD segments with no shared rare variants IBD segments with shared rare variants





IBD segments with only LOD > 3 IBD segments with both LOD > 3 & rvLOD > 3 IBD segments with only rvLOD > 3

