Attacks on genetic privacy via uploads to genealogical databases

Michael D. Edge & Graham Coop

Center for Population Biology and Department of Evolution and Ecology, University of California, Davis

December 17, 2019

Abstract

Direct-to-consumer (DTC) genetics services are increasingly popular for genetic genealogy, with tens of millions of customers as of 2019. Several DTC genealogy services allow users to upload their own genetic datasets in order to search for genetic relatives. A user and a target person in the database are identified as genetic relatives if the user's uploaded genome shares one or more sufficiently long segments in common with that of the target person—that is, if the two genomes share one or more long regions identical by state (IBS). IBS matches reveal some information about the genotypes of the target person, particularly if the chromosomal locations of IBS matches are shared with the uploader. Here, we describe several methods by which an adversary who wants to learn the genotypes of people in the database can do so by uploading multiple datasets. Depending on the methods used for IBS matching and the information about IBS segments returned to the user, substantial information about users' genotypes can be revealed with a few hundred uploaded datasets. For example, using a method we call IBS tiling, we estimate that an adversary who uploads approximately 900 publicly available genomes could recover at least one allele at SNP sites across up to 82% of the genome of a median person of European ancestries. In databases that detect IBS segments using unphased genotypes, approximately 100 uploads of falsified datasets can reveal enough genetic information to allow accurate genome-wide imputation of every person in the database. Different DTC services use different methods for identifying and reporting IBS segments, leading to differences in vulnerability to the attacks we describe. We provide a proof-of-concept demonstration that the GEDmatch database in particular uses unphased genotypes to detect IBS and is vulnerable to genotypes being revealed by artificial datasets. We suggest simple-to-implement suggestions that will prevent the exploits we describe and discuss our results in light of recent trends in genetic privacy, including the recent use of uploads to DTC genetic genealogy services by law enforcement.

1 Introduction

As genotyping costs have fallen over the last decade, direct-to-consumer (DTC) genetic testing (Hogarth et al., 2008; Hogarth and Saukko, 2017; Khan and Mittelman, 2018) has become a major industry, with over 26 million people enrolled in the databases of the five largest companies (Regalado, 2019). One of the major applications of DTC genetics is genetic genealogy. Customers of companies such as 23andMe and Ancestry, once they are genotyped, can view a list of other customers who are likely to be genetic relatives. These putative relatives' full names are often given, and sometimes contact details are given as well. Such genealogical matching services are of interest to people who want to find distant genetic relatives to extend their family tree, and can be particularly useful to people who otherwise may not have information about their genetic relatives, such as adoptees or the biological children of sperm donors. Several genetic genealogy services—including GEDmatch, MyHeritage, FamilyTreeDNA, and LivingDNA (Table 1)—allow users to upload their own genetic data if they have been genotyped by another company. These entities generally offer some subset of their services at no charge to uploaders, which helps to grow their databases. Upload services have also been used by law enforcement, with the goal of identifying relatives of the source of a crime-scene sample (Erlich et al., 2018; Edge and Coop, 2019), prompting discussion about genetic privacy (Court, 2018; Ram et al., 2018; Kennett, 2019; Scudder et al., 2019).

The genetic signal used to identify likely genealogical relatives is identity by descent (IBD, Browning and Browning 2012; Thompson 2013. We use "IBD" to indicate both "identity by descent" and "identical by descent," depending on context.) Pairs of people who share an ancestor in the recent past can share segments of genetic material from that ancestor. The distribution of IBD sharing as a function of genealogical relatedness is well studied (Donnelly, 1983; Huff et al., 2011; Browning and Browning, 2012; Thompson, 2013; Buffalo et al., 2016; Conomos et al., 2016; Ramstetter et al., 2018), and DTC genetics entities can use information about the number and length of inferred IBD segments between a pair of people to estimate their likely genealogical relationship (Staples et al., 2016; Ramstetter et al., 2017). These shared segments—IBD segments—result in the sharing of a near-identical stretch of chromosome (a shared haplotype). Shared haplotypes can most easily be identified looking for long genomic regions where two people share at least one allele at nearly every locus.

For the rest of the main text, we focus on identical-by-state (IBS) segments, which are genomic runs of (near) identical sequence shared among individuals and can be thought of as a superset of true IBD segments. Very long IBS segments, say over 7 centiMorgans (cM), are likely to be IBD, but shorter IBS segments, say <4 cM, may or may not represent true IBD due to recent sharing—they may instead represent a mosaic of shared ancestry deeper in the past. Many of the algorithms for IBD detection that scale well to large datasets rely principally on detection of long IBS segments, at least as their first step (Gusev et al., 2009; Henn et al., 2012; Huang et al., 2014). We consider the effect on our results of attempting to distinguish IBS and IBD in supplementary material.

Many DTC genetics companies, in addition to sharing a list of putative genealogical relatives, give customers information about their shared IBS with each putative relative, possibly including the number, lengths, and locations of shared genetic segments (Table 1). This IBS report may represent substantial information about one's putative relatives—one already has access to one's own genotype, and so knowing the locations of IBS sharing with putative relatives reveals

Service	Database Size (millions)	Individuals Shown	IBS/IBD Segments Reported
GEDmatch	1.2	3,000 closest matches shown free; Unlimited w/ $10/month license; any two kits can be searched against each other	Yes if longer than user-set threshold. Min. threshold 0.1cM, default 7cM
FamilyTreeDNA	1*	All that share at least one 9cM block or one 7.69cM block and 20 total cM	Yes, down to 1cM, for $19 per kit
MyHeritage	3	All that share at least one 8cM block	Yes, down to 6cM, for $29 per kit or unlimited for $129/year. Customers may opt out
LivingDNA	Unknown	Putative relatives out to ≈ 4th cousin	Only sum length of matching segments reported
DNA.LAND**	0.159	Top 50 matches shown with minimum 3cM segment	Yes

Table 1: Key parameters for several genetic genealogy services that allow user uploads as of July 26th, 2019. *Though Regalado (2019) reports that FamilyTreeDNA has two million users, he also suggests that only about half of these are genotyped at genome-wide autosomal SNPs, which is in line with other estimates (Larkin, 2018). **DNA.LAND has discontinued genealogical matching for uploaded samples as of July 26th, 2019.

information about those relatives' genotypes in those locations (He et al., 2014). Users of genetic genealogy services implicitly or explicitly agree to this kind of genetic information sharing, in which large amounts of genetic information are shared with close biological relatives and small amounts of information are shared with distant relatives.

Here we consider methods by which it may be possible to compromise the genetic privacy of users of genetic genealogy databases. In particular, we show that for services where genotype data can be directly uploaded by users, many users may be at risk of sharing a substantial proportion of their genome-wide genotypes with any party that is able to upload and combine information about several genotypes. We consider two major tools that might be used by an adversary to reveal genotypes in a genetic genealogy database. One tool available to the adversary is to upload real genotype data or segments of real genotype data. When uploading real genotypes, the information gained comes by virtue of observed sharing between the uploaded genotypes and genotypes in the database (an issue also raised by Larkin, 2017). Publicly available genotypes from the 1000Genomes Project (1000 Genomes Project Consortium, 2012), Human Genome Diversity Project (Cann et al., 2002), OpenSNP project (Greshake et al., 2014), or similar initiatives might be uploaded.

A second tool available to the adversary is to upload artificial genetic datasets (Ney et al., 2018). In particular, we consider the use of artificial genetic datasets that are tailored to trick algorithms that use a simple, scalable method for IBS detection, that of identifying long segments in which a pair of genomes contains no incompatible homozygous sites (Henn et al., 2012; Huang et al., 2014). Such artificial datasets can be designed to reveal the genotypes of users at single sites of interest or sufficiently widely spaced sites genome-wide. We describe how a set of a few hundred artificial datasets could be designed to reveal enough genotype information to allow accurate imputation of common genotypes for every user in the database.

Below, we describe these procedures and illustrate them using either publicly available or artificial data. We show that under some circumstances, many users could be at risk of having their genotypes revealed, either at key positions or at many sites genome-wide. In particular, we show that GEDmatch, as of mid-December 2019, was vulnerable to an attack we term IBS baiting that obtains genotype data via artificial data uploads. Our results are largely complementary to the independent work of Ney et al. (2020), which was first posted publicly within a week of the first public posting of this manuscript on bioRxiv. In the discussion, we consider our work in light of other genetic privacy concerns (Erlich and Narayanan, 2014; Naveed et al., 2015) and the work of Ney and colleagues (2020), and we give some suggested practices that DTC genetics services can adopt to prevent privacy breaches by the techniques described here.

2 Results

We describe three general methods for revealing the genotypes of users in genetic genealogy databases that allow uploads. The first, **IBS tiling**, involves uploading many real genotypes in order to identify genotype information from many regions in many people. The second, **IBS probing**, involves uploading a dataset containing a long haplotype that includes an allele of interest, creating matches at this locus. Genotypes at other places in the genome are chosen to be unlikely to generate IBS with any user in the database, so matches with the uploaded dataset are likely to be users who carry the allele of interest. The third method, **IBS baiting**, involves uploading fake datasets with long runs of heterozygosity to induce phase-unaware methods for IBS calling to reveal genotypes.

2.1 IBS tiling

In IBS tiling, the genotype information shared between a target user in the database and each member of a set of comparison genomes is aggregated into potentially substantial information about the target's genotypes. For example, consider a user of European ancestries. She is likely to have some degree of IBS sharing with a large set of people from across Europe (Ralph and Coop, 2013) and beyond. If one knows the user's IBS sharing locations with one random person of European ancestries (and the random person's genotype), then one can learn a little about the user's genotype. But if one can upload many people's genotypes for comparison, then one can uncover small proportions of the target user's genotypes from many of the comparison genotypes, eventually uncovering much of the target user's genome by virtue of a "tiling" of shared IBS with known genotypes (Figure 1A). Similar ideas have been suggested with application to IBD-based genotype imputation (Carmi et al., 2014).

We consider the amount of IBS tiling possible within a set of publicly available genotypes for 872 people of European origin genotyped at 544,139 sites. We phased the sample using Beagle 5.0 (Browning and Browning, 2007) and used Refined IBD software (Browning and Browning, 2013) to identify IBS segments (see Methods). In the main text, we include IBS segments that are not particularly likely to be IBD—these are IBS segments returned by Refined IBD with relatively low

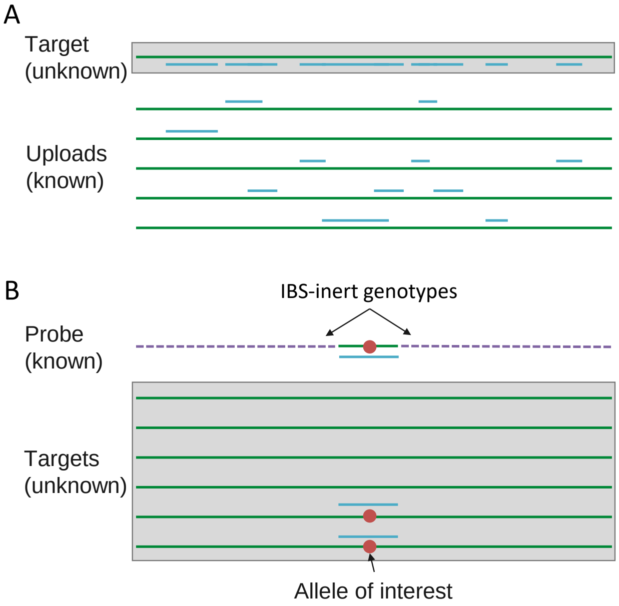

Figure 1: Schematics of the IBS tiling and IBS probing procedures. (A) In IBS tiling, multiple genotypes are uploaded (green lines) and the positions at which they are IBS with the target (represented by blue lines) are recorded. Once enough datasets have been uploaded, the target will eventually have a considerable proportion of their genome "tiled" by IBS with uploads that have known genotypes. (B) In IBS probing, the uploaded probe consists of a haplotype carrying an allele of interest (red dot) surrounded by "IBS-inert" segments (purple dashed lines)—fake genotype data designed to be unlikely to share any IBS regions with anyone in the database. In the event of an IBS match in the database, the matching database entry is likely to carry the allele of interest.

LOD scores for IBD, between 1 and 3. We consider the results obtained after filtering segments likely to be true IBD in Figure S1. True IBD segments reveal more than mere IBS segments about shared genotypes because untyped variants (including rare variants) within an IBD segment are likely to be shared. At the same time, mere IBS is sufficient to infer sharing for SNPs that are genotyped within the segment.

Once we identified IBS segments shared among the 872 people in our sample, we set out to estimate the amount of genotype information that could be identified using IBS tiling. The amount of genotype information obtainable is strongly influenced by two factors: the size of the comparison set used (i.e., the number of people used to identify IBS segments with a target sample), and the restrictiveness of the criteria by which IBS segments are identified. For example, if only long IBS segments are shown to users, then the proportion of a typical person's genotype data obtainable will be smaller than if short IBS segments are also shown. The minimum IBS length reported by several genetic genealogy services as of July 26th, 2019 is shown in Table 1.

Figure 2 shows the median amount of coverage obtainable by IBS tiling as a function of comparison sample size, imposing various restrictions on the minimum segment length in cM. (For similar results, see Figure 2b of Carmi et al. (2014) and Figure 2 of Panoutsopoulou et al. (2014).) Approximately 2.8 Giga base-pairs (Gbp) were covered by IBS segments anywhere in

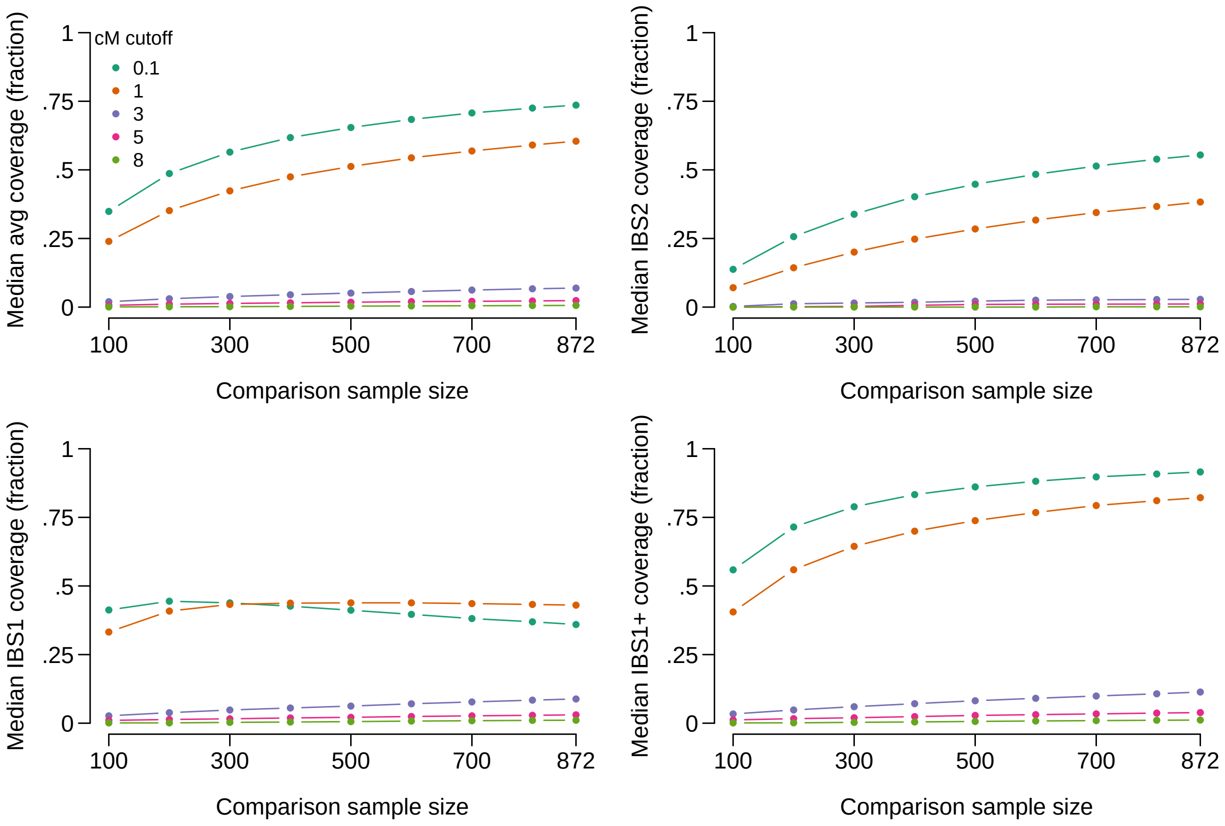

Figure 2: Lengths of genome in Giga base-pairs (Gbp) covered by IBS tiling as a function of minimum required length of IBS segments in centiMorgans (cM) and size of a randomly selected comparison sample for the median person in our dataset. The top-left panel shows the average coverage across each of the person's two haplotypes. The top-right shows IBS2 coverage, the length of genome where both haplotypes are covered by IBS tiles. The bottom-left panel shows IBS1, the length of genome where exactly one haplotype is covered by IBS tiles. (IBS1 coverage can decrease at larger comparison sample sizes because IBS2 coverage increases.) The bottom-right panel shows IBS1+ coverage, the length of genome covered by either IBS1 or IBS2.

the genome; we take this to be approximately the maximum possible genomic length recoverable by IBS with our SNP set. Using the entire sample (871 genotypes, since the target is left out) and including all called IBS segments >1 cM, the median person has an average of 60% of the maximum length of 2.8 Gbp covered by IBS segments (with the average taken across their two chromosomes), and sites across 82% of this length have at least one of two alleles recoverable by IBS tiling. Increasing the cM threshold required for reporting substantially reduces the amount of IBS tiling. With a cutoff of 3 cM, approximately 6.9% of the median person's genotype information is recoverable, including at least one of two alleles at sites in 11% of the genome. When a more stringent cutoff of 8 cM is used, only 1% of the genome has at least one of two alleles recoverable for the median person when using a comparison sample of 871. Our reports for segments longer than 3 cM may be conservative because Refined IBD sometimes splits long IBS segments into multiple shorter segments in the presence of phasing errors (Browning and Browning, 2013; Bjelland et al., 2017).

For some people, the amount of information obtainable by IBS tiling is even larger. In our sample, the top 10% of people have genotypes across 76% of their total genome covered by IBS tiles, including one or more alleles at sites in at least 93% of the 2.8 Gbp covered by IBS tiles anywhere. If only segments longer than 3 cM are reported, the top 10% of people have one or both alleles covered at sites in at least 38% of the total, and if only segments longer than 8 cM are reported, the top 10% have one or both alleles covered at sites in at least 6% of the total.

The coverage obtained by IBS tiling and its informativeness about target genotypes depends on the specific practices used for reporting IBS information (Figures S1-S5). For example, some DTC genealogy services only report matching segments for pairs of people who share at least one long IBS segment (Table 1), but then allow users to see shorter IBS segments (> 1cM) for those pairs of people. Unsurprisingly, we find that this strategy allows a higher level of IBS tiling than if only long segments are revealed (Figure S2), because people who share a long IBS segment may also share shorter segments that are hidden if only long segments are reported.

In this demonstration of IBS tiling, we used haplotype information provided by the Refined IBD software to determine which haplotypes were covered by IBS in each person. Most genetic genealogy services that provide information on the location of IBS matches with putative relatives do not provide haplotype information, making it difficult to distinguish IBS1 (in which one chromosome is covered by an IBS segment) and IBS2 (in which both chromosomes are covered by IBS segments). One tool available to an adversary pursuing IBS tiling is to upload genotype information that is homozygous at all sites using one of two phased haplotypes as a basis, effectively searching for IBS with one chromosome at a time. In the presence of phasing errors, some IBS segments may be missed, and the assumption that phase is known would render the coverage rates in Figure 2 overestimates. At the same time, the decrease in tiling performance is small for short segments, which can be seen by conducting our test of IBS tiling using Germline software with the haploid flag, which causes putative IBS segments to terminate with a single phasing error (Figure S3). It may remain difficult to distinguish some cases—such as distinguishing IBS1 from IBS2 with a run of homozygosity on the database genotype—but there will be no question about which uploaded haplotype is IBS with the database genotype. Thus, at any point where a homozygous upload and a target are IBS, at least one of the target's alleles is known. Further, if the target is IBS with any other uploaded datasets at a genetic locus of interest, it will often be possible to infer the target's full genotype.

IBS tiling rates vary somewhat by population, with Finnish samples showing the highest tiling rates among the 1000Genomes populations included (Figure S4). There also appear to be slight biases for IBS tiles to appear in regions with low SNP density and lower heterozygosity, meaning that the proportion of alleles—and particularly the proportion of minor alleles—recovered by tiling is typically slightly lower than the proportion of the genome length in Mbp covered (Figure S5).

2.2 IBS probing

IBS probing is an application of the same idea underlying IBS tiling. By IBS probing, one could identify people with specific genotypes of interest, such as risk alleles for Alzheimer's disease (Corder et al., 1993), even if the DTC service does not report chromosomal locations of IBS matches. To identify people carrying a particular allele at a locus of interest, one could use haplotypes carrying the allele in publicly available databases. To do so, one would extract a haplotype that surrounds the allele of interest and place it into a false genetic dataset designed to have no long IBS segments with any real genomes (Figure 1B). Thus, any returned putative relatives must match at the allele of interest, revealing that they carry the allele. We call this attack "IBS probing" by analogy with hybridization probes, as the genuine haplotype around the allele of interest acts as a probe. Whereas IBS tiling recovers genetic information from across the genome, IBS probing acts only on a single locus of interest. The advantage is that IBS probing is possible even in databases that do not report the chromosomal locations of IBS segments.

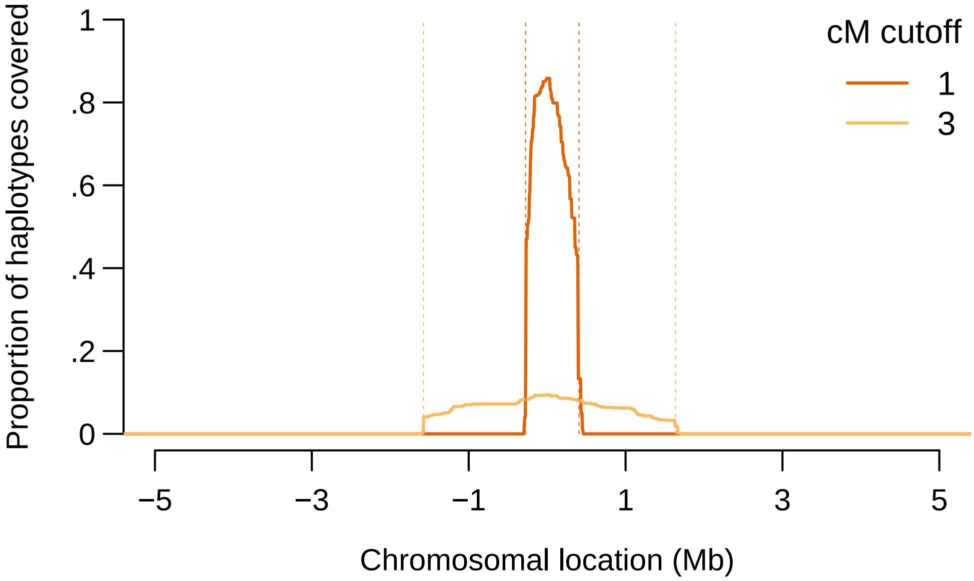

Figure 3: A demonstration of the IBS probing method around position 45411941 on chromosome 19 (GRCh37 coordinates), in the APOE locus. We show the proportion of haplotypes among the 872 Europeans in our sample covered IBS by probes constructed from the sample, as a function of the chromosomal location in a 10-Mb region around the site of interest. In red, we show the coverage using a 1cM threshold for reporting IBS, where the probes are constructed using real data in a 1.9-cM region centered on the site of interest (region boundaries shown in dashed orange). In orange, we show the coverage using a 3cM threshold for reporting IBS, where the probes are constructed using real data in a 5.9-cM region around the site of interest.

There are several ways of generating chromosomes unlikely to have long shared segments with any entries in the database. One simple way is to sample alleles at each locus in proportion to their frequencies. Chromosomes generated in this way are free of linkage disequilibrium (LD) and thus unlike genuine chromosomes. If the database distinguishes between IBS and IBD, then these fake data are unlikely to register as IBD with any genuine haplotypes. However, they may appear as IBS in segments where genetic diversity is low, depending on the length threshold used by the database. Near-zero rates of IBS can be obtained by generating more unusual-looking fake data, such as by sampling alleles from one minus their frequency or by generating a dataset of all minor alleles.

Figure 3 shows a demonstration of IBS probing performance in our set of 872 Europeans in a window around the APOE locus. For a 1cM threshold for reporting IBS, we generated probes by retaining 1.9cM of real data around a site of interest in the APOE locus from all 872 people. Outside that 1.9cM window, we generated data by drawing alleles randomly (see Methods). For a

3cM threshold for reporting IBS, we generated probes by retaining 5.9cM of real data around the site of interest. With 1cM matching, 1497 of 1744 haplotypes (86%) matched one of the probes at the site of interest. (Target haplotypes were not allowed to match probes constructed from the same person that carried the target haplotype.) With 3cM matching, 164 of 1744 haplotypes (9.4%) matched one of the probes at the site of interest. Very few matches occurred outside the region of interest—none with a 3cM threshold and only 0.1% of matches with a 1cM threshold. Moreover, we generated different inert genotypes for all 872 probes, and the great majority of these had no matches with any real sample. An adversary would only need to generate one inert dataset, which can be tested by uploading to the database and confirming that no matches are returned. Probes could then be constructed by stitching real haplotypes at the site of interest into the the same set of inert data. The probes would then be likely to match each other, but the adversary would know those identities and could ignore those matches.

The efficacy of IBS probing will depend on the minimum IBS-match length reported to users, the specific methods used for identifying IBS segments (Figures S6-S7), and whether the genotype of interest is included on the SNP chip. These factors vary in terms of whether they affect the sensitivity of IBS probing—the proportion of people carrying the allele of interest returned by a probe or set of probes—or the precision of IBS probing—the proportion of people returned by a probe who in fact carry the genotype of interest. For example, high thresholds for IBS reporting will mean that uploaded genotypes will need to have long IBS segments with targets at the locus of interest. Long IBS segments are likely to represent relatively close genealogical relatives (i.e., long IBS segments are likely to be IBD segments), and not many targets will be close relatives of the source of any given haplotype of interest, meaning that the sensitivity of IBS probing is reduced by reporting thresholds that require long IBS segments. If the locus of interest, or a highly correlated one, is not included on the chip used to genotype either the uploaded sample or the target sample, then probing may only expected to work well if the upload and the target are truly IBD rather than merely IBS, reducing the precision of IBS probing for variants that are not genotyped. Limiting probing results to likely IBD matches will decrease the number of matches returned, particularly for short cM thresholds (Figure S6).

Another factor that will affect the success of IBS probing is the frequency of the allele of interest. For example, if the allele of interest is very rare, then it is likely to be only somewhat enriched on the haplotypes that tend to carry it, and reported matches may not actually carry the allele, even if they are IBD with an uploaded haplotype that carries it. IBS probing will perhaps be most sensitive and precise when the allele of interest is both common and relatively young, as is the case for founder mutations. In this case, most carriers of the allele will share the same long haplotype around the site of interest, meaning that fewer probes would need to be uploaded in order to learn the identities of the majority of the carriers in the database.

2.3 IBS baiting

IBS tiling and IBS probing take advantage of publicly available genotype data. The idea of both is that an adversary uploads genuine genetic datasets—or, in the case of IBS probing, datasets with genuine segments—to learn about entries in the database that share segments with the uploaded genomes.

In this section, we describe an exploit called IBS baiting. The specific strategy for IBS baiting that we describe is possible if the database identifies putative IBS segments by searching for long regions where a pair of people has no incompatible homozygous sites. An incompatible homozygous site is a site at which one person in the pair is homozygous for one allele, and the other person is homozygous for the other allele. Identifying IBS segments in this way does not require phased genotypes and scales relatively easily to large datasets—we refer to methods in this class as "phase-unaware" and contrast them with phase-aware methods for IBS detection. Phase-unaware methods are robust to phasing errors, which are an issue for long IBD segments (Durand et al., 2014). Major DTC genetics companies have used phase-unaware methods in the past for IBS detection (Henn et al., 2012; Hon et al., 2013), and some state-of-the-art IBD detection and phasing pipelines feature an initial phase-unaware step (Huang et al., 2014; Loh et al., 2016).

The main tool used in IBS baiting is the construction of apparently IBS segments by assigning every uploaded site in the region to be heterozygous. (SNPs with missing data may also included in these regions). These runs of heterozygosity, which are unlikely to occur naturally (unlike runs of homozygosity, McQuillan et al., 2008; Pemberton et al., 2012), will be identified as IBS with every genome in the database using phase-unaware methods: because they contain no homozygous sites at all, they cannot contain homozygous sites incompatible with any person in the database.

Here, we consider a database in which an apparent IBS segment is halted exactly at the places at which the first incompatible homozygous site occurs on each side of the segment. We also assume that the database detects all segments without incompatible homozygous sites that pass the required length threshold. Ney et al. (2020) independently proposed a similar approach in their section VII ‘Genetic Marker Extraction Using Matching Segments,’ showing that GEDmatch was vulnerable to it. Similarly, we demonstrate below that IBS baiting can be implemented against GEDmatch.

2.3.1 Single-site IBS Baiting

The simplest application of IBS baiting is to use it to reveal genotypes at a single site. If IBS is identified by looking for single incompatible homozygous sites and missing data can be ignored, then users’ genotypes at any single biallelic site of interest can be determined by examining their putative IBS with each of two artificial datasets (Figure 4A). In each artificial dataset, the site of interest is flanked by a run of heterozygosity. The combined length of these two runs of heterozygosity must exceed the minimum length of IBS segment reported by the database. The adversary uploads two datasets with these runs of heterozygosity in place. In one dataset, the site of interest is homozygous for the major allele, and in the other, the site of interest is homozygous for the minor allele. If the target user is homozygous at the site of interest, then one of these two uploads will not show a single, uninterrupted IBS segment—IBS will be interrupted at the site of interest (or may not be called at all). If the IBS segment with the dataset homozygous for the major allele is interrupted, then the target user is homozygous for the minor allele. Similarly, if the IBS segment with the dataset homozygous for the minor allele is interrupted, then the target user is homozygous for the major allele. If both uploads show uninterrupted IBS segments with the target, then the target user is heterozygous at the site of interest. Thus, for any genotyped biallelic site of interest, the genotypes of every user shown as a match can be revealed after uploading two artificial datasets. Depending on how possible matches are made accessible to the adversary, the genotypes of every user could be returned. Genotypes of medical interest that

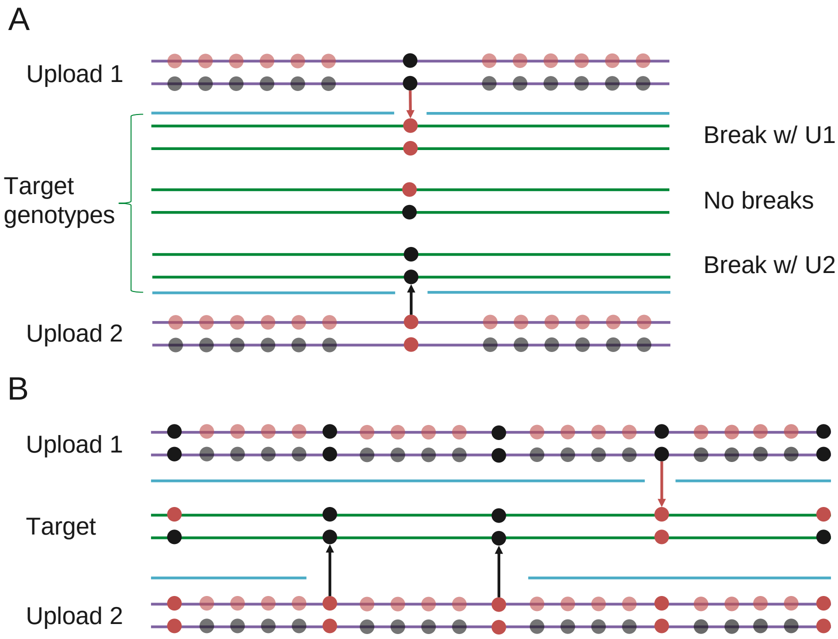

Figure 4: Schematics of the IBS baiting procedure. (A) To perform IBS baiting at a single site, two uploads are required, each with runs of heterozygous genotypes flanking the key site. At the key site, the two uploaded datasets are homozygous for different alleles. The three possible target genotypes at the key site can each be determined by examining their IBS coverage with the uploads. If there is a break in IBS with either upload, then the target is homozygous for the allele not carried by the upload that shows the break in IBS (with the broken IBS segment shown as a cyan line). If there is no break in IBS with either upload, then the target is heterozygous at the key site. (B) Target genotypes at many key sites across the genome can be learned by comparison with two uploaded datasets, as long as key sites are spaced widely enough.

are often included in SNP chips, such as those in the APOE locus (Corder et al., 1993), are potentially vulnerable to single-site IBS baiting.

Here, we have considered a database using the simplest possible version of a phase-unaware method for detecting IBS, that in which an apparent IBS segment is halted exactly at the places at which the first incompatible homozygous site occurs on each side of the segment. In principle, phase-unaware IBS-detection algorithms can be altered to allow for occasional incompatible homozygous sites before halting as an allowance for genotyping error, or the extent of the reported region might be modified to be less than the full range between incompatible homozygous sites. Versions of IBS baiting might be developed to work within such modifications. The key insight is that if two artificial kits differ at exactly one site in a region and they produce two different patterns of called IBS with a target, then the target's genotype is revealed at that site. For example, if a database uses a phase-unaware method for IBS calling that requires two incompatible homozygous sites before a putative IBS segment is halted, then an attacker might modify our scheme by putting in a rare homozygote at a site near the key site. For most target users, the rare homozygote in the uploaded files would be an incompatible homozygous site, implying that a mismatch at the key site will cause a break in a putative IBS region. By using different homozygote genotypes nearby, an attacker might still identify the genotypes of everyone in the database at the key site. As discussed below, such measures do not appear to be necessary to perform IBS baiting in GEDmatch. Further, in GEDmatch, uploading a third bait dataset with a missing genotype at the key site can distinguish targets with missing genotypes from heterozygous targets.

Single-site IBS baiting could also be used if chromosomal locations of matches are not reported. To do so, one would use the the scheme we describe in a large region surrounding the locus of interest and use fake IBS-inert segments to fill in the rest of the dataset.

2.3.2 Parallel IBS Baiting

The second method we consider applies the IBS baiting technique to many sites in parallel (Figure 4B). By parallel application of IBS baiting, users' genotypes at hundreds or thousands of sites across the genome can be identified by comparison with each pair of artificial genotypes. By repeated parallel IBS baiting, eventually enough genotypes can be learned that genotype imputation becomes accurate, and genome-wide genotypes could in principle be imputed for every user in the database. If IBS segments as short as 1cM are reported to the user, then accurate imputation (97-98% accuracy) becomes possible after comparison with only ≈100 uploaded datasets. The procedure starts by designing a single pair of uploaded files as follows:

1. Identify a set of key sites to be revealed by the IBS baiting procedure. For every key site, the sum of the distances in cM to the nearest neighboring key site on each side (or the end of the chromosome, if there is no flanking key site on one side) must be at least the minimum IBS length reported by the database.

2. Produce two artificial genetic datasets. In each, every non-key site is heterozygous. In one, each key site is homozygous for the major allele, in the other, each key site is homozygous for the minor allele.

3. Upload each artificial dataset and compare them to a target user. Key sites that are covered by putative IBS segments between the target and both artificial datasets are heterozygous in the target. The target is homozygous for the major allele at key sites that are covered by putative IBS segments between the target and the major-allele-homozygous dataset only. Similarly, the target is homozygous for the minor allele at key sites that are covered by putative IBS segments between the target and the minor-allele-homozygous dataset only.

Carrying out this procedure reveals the target's genotype at every key site. If IBS segments of length at least t cM are reported, and a chromosome is c cM long, then up to $2c/t - 1$ key sites can be revealed with each pair of uploaded files. (To see this, consider the case where $c = tk$, with k a positive integer, and place key sites at $t/2, t, 3t/2, ..., c - t/2$. This calculation ignores the possibility of missing data at key sites in the target.) This means that with a minimum reported IBS threshold of 1cM, 100 uploaded datasets could reveal approximately 100 genotypes per cM, which is enough to impute genome-wide genotypes at $97 - 98\%$ accuracy (Shi et al., 2018). In principle, the key sites could also be chosen to ensure good LD coverage and higher imputation accuracy. Of course, higher accuracy imputation can be obtained by recovering exact genotypes for more sites, and with several thousand uploads, the genotypes at every genotyped site could be revealed by IBS baiting without the need to impute.

2.3.3 IBS baiting in GEDmatch

We hypothesized that IBS baiting would work in the GEDmatch DTC database. GEDmatch provides no public documentation of the IBS algorithm they use, but IBS segments identified by GEDMatch seem to terminate only on incompatible homozygous sites, as would be expected if they use phase-unaware IBS detection. Specifically, the GEDmatch 1-to-1 match tool identifies the locations of IBS segments between pairs of genetic datasets ("kits" in GEDMatch terminology) and allows the user to specify the minimum genetic length and minimum number of matching SNPs to include in a segment. The 1-to-1 tool also returns a 'full resolution' picture of the chromosome that appears to be a SNP-by-SNP picture of the match between the kits along each chromosome. (These pictures are themselves a major security risk. We alerted GEDmatch to the risk in a July 24th email (posted here: `https://github.com/mdedge/IBS_privacy/blob/master/IBS_baiting_demo/GEDmatch_emails.pdf`) but did not analyze them further. Ney et al. (2020) showed in detail that the images provided by GEDmatch allow an adversary to learn the full genotype of a target person.)

To demonstrate IBS baiting in GEDmatch, we uploaded a small number of artificial genotypes to their database beginning in late November 2019. These kits were designed in accordance with the algorithm discussed above, but with some slight alterations to bypass counter-measures that GEDMatch has put in place since we (and, independently, Ney and colleagues) informed them of the risk of IBS baiting in summer 2019. Before uploading any data to GEDmatch, we first confirmed our planned procedure with the UC Davis IRB and with GEDmatch representatives. We uploaded our kits into the GEDmatch 'research' and not 'public' category to prevent matches to the public database, and only used the 1-to-1 IBS match tool among our own uploaded test kits. In this way, we avoided interacting with any genotype data of real GEDmatch users and did not violate GEDmatch's terms and conditions.

We targeted four random SNPs along chromosome 22 for IBS baiting. We uploaded two bait genotype kits (B1 & B2) that had opposite-homozygote genotypes at each of these key SNPs. Each key SNP was in turn surrounded by a $\sim$ 1cM stretch of SNPs containing genotypes that were either heterozygoous or coded missing. The rest of the genome was specified to be IBS-inert. We then uploaded three target genotype datasets whose genotypes we wanted to determine at the key sites. Two of these target kits (T1 and T3) had opposite-homozygous genotypes at each of the key SNPs, while the third (T2) was heterozygous at each key SNP. (See section 4.4.1 for more details on the kit design.) We then used the GEDmatch 1-to-1 match tool, choosing the parameters so a single opposite-homozygous genotype between a bait and target kit would interrupt a putative IBS segment.

In each case, our two bait kits had the correct IBS patterns with the target kits, allowing correct determination of the target genotypes by IBS baiting. On the left of Figure 5, we show a zoomed-in view of the three targets' matches around one of the key SNP sites. The homozygous targets have IBS matches with only one of the bait kits, whereas the heterozygous target has IBS matches with both bait kits. This pattern is seen across all four target regions (right side of Figure 5, see Section 4.4.2 for more detailed results). The target and bait kits displayed in Figure 5 were uploaded and analyzed on December 15, 2019, showing that GEDmatch has remained vulnerable to IBS-baiting attacks even after its acquisition by Verogen, which was announced on December 9, 2019.

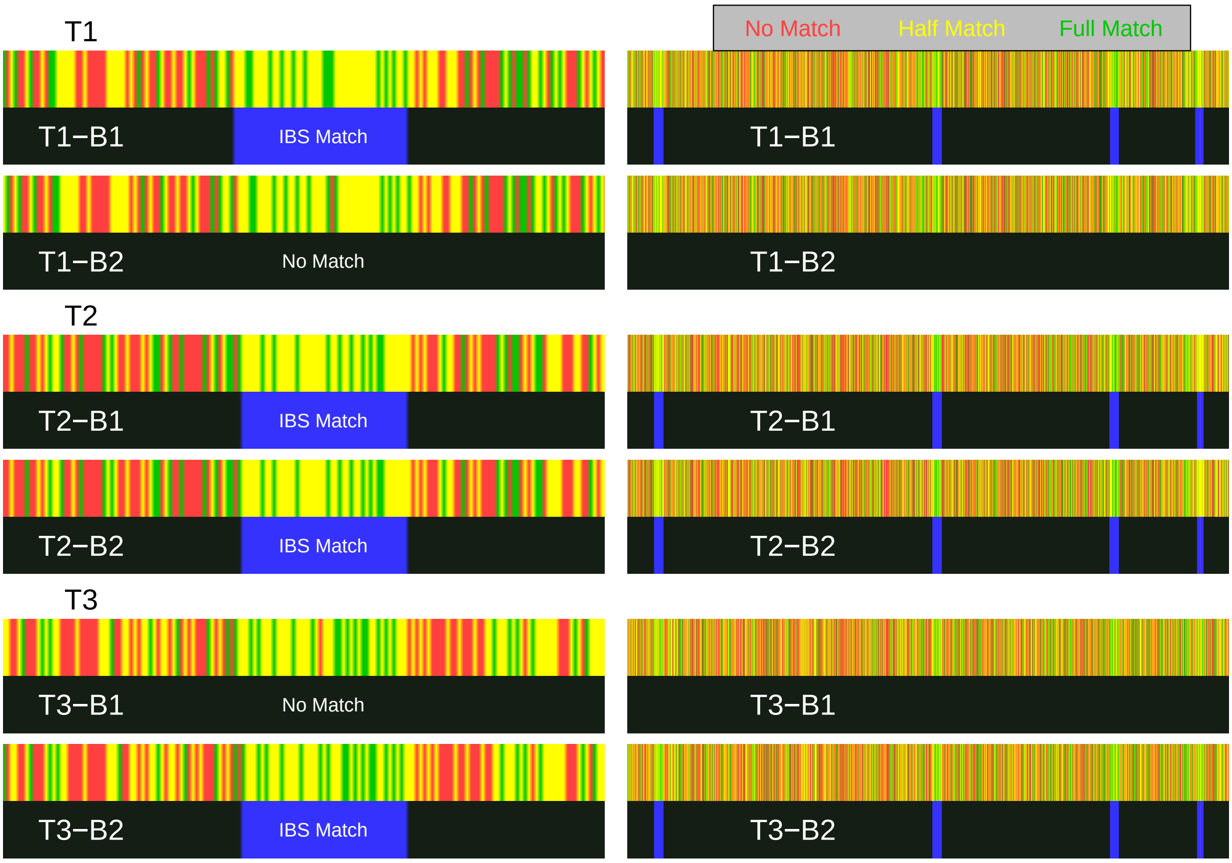

Figure 5: Visualization of IBS baiting using GEDmatch's 1-to-1 chromosome browser. **Left:** Zoomed-in view of the region containing key SNP 1, showing the three target kits (T1-T3) matched to the two bait kits (B1 & B2). **Right:** Zoomed-out views of regions containing all four key SNPs on chromosome 22. For each pair of bait & target kits, the top rectangle (red, yellow, or green) shows the GEDmatch SNP-level pairwise genotype-match image (colored to show no match, half match, or full match) returned by the 1-to-1 GEDmatch tool. The bottom rectangle (black & blue) shows the GEDmatch IBD-track image, black for no putative IBD match, blue regions showing putative IBD segments. The white text on the IBS track is not provided by GEDmatch and was added as a guide to the eye. Opposite-homozygote calls at the key SNP are seen in the left panel as a red line in an otherwise matching region (yellow & green). The spatial positions of SNPs in the match panel appears to have been jittered; e.g. the location of the red line varies slightly in the different plots that should have the same coordinate system (perhaps as a countermeasure against a Ney et al. (2020)-style attack.

3 Discussion

We have suggested several methods by which an adversary might learn the genotypes of people included in a genetic genealogy database that allows uploads. Our methods take advantage of both the population-genetic distributions of IBS segments and of methods used for calling IBS. In particular, IBS tiling works simply because there are background levels of IBS (and IBD) even among distantly related members of a population (e.g. Ralph and Coop, 2013). In our dataset, the median person had the majority of their genetic information susceptible to IBS tiling on the basis of other members of the dataset, depending on the procedures used for reporting IBS. IBS tiling performance will also depend on the ancestries of the target and comparison samples because IBD rates differ within and among populations (Palamara et al., 2012; Carmi et al., 2013; Ralph and Coop, 2013), as well as on the prevalence of close biological relatives in the dataset. IBS tiling performance improves as the size of the comparison sample increases. Thus, if enough genomes are compared with a target for IBS, eventually a substantial amount of the target genome is covered by IBS with one or more of the comparison genomes.

IBS probing combines the principles behind IBS tiling with the idea of "IBS-inert" artificial segments. If the majority of the genome—everywhere except a locus of interest—can be replaced with artificial segments that will not have IBS with any genome in the database, then the adversary knows that any matches identified are in a locus of interest. As such, IBS probing could be used to reveal sensitive genetic information about database participants even if chromosomal locations of matches are not reported to users.

Finally, IBS baiting exploits phase-unaware IBS calling algorithms that use incompatible homozygous sites to delimit putative IBS regions. Although such methods can be useful in genetic genealogy because they scale well to large data, they are vulnerable to fake datasets that include runs of heterozygous sites, which will be identified as IBS with everyone in the database. By inserting homozygous genotypes at key sites and heterozygotes everywhere else, we estimate that approximately 100 well-designed uploads could reveal enough genotypes to impute genome-wide information for any user in a database, provided that the threshold for reporting a matching segment is ≈ 1 cM. Similarly, two uploads could reveal any genotype at a single site of interest, such as rs429358, which reveals whether the user carries an APOE-ϵ4 variant and is associated with risk of late-onset Alzheimer's disease.

There are millions of people enrolled in genetic genealogy databases that allow uploads (Table 1). Genetic genealogy has many applications, and uploads are popular with users who want to find relatives who may be scattered across different databases. Though allowing uploads brings several benefits for both customers and DTC services, it also entails additional privacy risks. Users of DTC genetic genealogy services that allow uploads could be at risk of having their genetic information extracted by the procedures we describe here, depending on the methods that these services use to identify and report IBS. Concerns arising from the methods we report are in addition to standard digital security concerns. The attacks we describe require little special expertise in computing; the adversary only needs to be able to procure or create the appropriate data files and to process and aggregate the information returned from the database.

We have not set out to determine precisely how vulnerable users of each specific DTC service are. We do not know the full details of methods used by each service for matching, nor have we attempted to deanonymize any real users' genotypes. We contacted representatives of each of the organizations listed in Table 1 90 days (July 24th, 2019) before posting this manuscript publicly in order to give them time to repair any security vulnerabilities related to the methods we describe. We have posted our emails to GEDmatch representatives here: `https://github.com/mdedge/IBS_privacy/blob/master/IBS_baiting_demo/GEDmatch_emails.pdf`.

On the basis of our results, we do have serious concerns about the privacy of GEDmatch users. As of this writing, GEDmatch uses length thresholds for displaying matching segments that are too short, allowing for effective IBS tiling attacks, and GEDmatch also appears to use phase-unaware IBD detection methods, allowing for IBS baiting attacks. Additionally, as detailed by Ney et al. (2020), whose work was independent of ours, GEDmatch provides users with high-resolution images comparing the chromosomes of any two users at SNP-level resolution, allowing for reconstruction of a target's genotype using these images. GEDmatch was recently purchased by Verogen, a forensic genetics company, but as of December 15, 2019, GEDmatch has not as yet prevented the attacks we describe. Since our and Ney et al.'s initial communications with GEDmatch in July, GEDmatch has placed a recaptcha on its upload and 1-to-1 tool forms, which may deter bulk bot attacks to harvest large numbers of kit genotypes. However, as we outline below, there are simple steps that could be taken to make IBS attacks much less of a risk.

In our estimation, the other active services listed in Table 1 (MyHeritage, FamilyTreeDNA, and LivingDNA) are likely substantially less vulnerable than GEDmatch to the attacks we describe here. LivingDNA does not provide a chromosome browser, precluding IBS tiling attacks. MyHeritage and FamilyTreeDNA use thresholds for revealing matching segment locations that make IBS tiling much less efficient. (However, FamilyTreeDNA's practice of showing matches as short as 1cM given that two people share at least one long match is still somewhat permissive, see Figure S2.) Representatives of MyHeritage, FamilyTreeDNA, and LivingDNA have confirmed to us that their IBD-calling algorithms rely on phased data, which should preclude IBS baiting. (We have not tested this ourselves.) DTC genetic genealogy is a growing field, and any new entities that begin offering upload services may also face threats of the kind we describe.

Genetic genealogy databases that allow uploads have been in the public eye recently because of their role in long-range familial search strategies recently adopted by law enforcement. In long-range familial search, investigators seek to identify the source of a crime-scene sample by identifying relatives of the sample in a genetic genealogy database that allows uploads. Searching in SNP-based genealogy databases allows the detection of much more distant relationships than does familial searching in traditional forensic microsatellite datasets (Rohlfs et al., 2012), vastly increasing the number of people detectable by familial search (Erlich et al., 2018; Edge and Coop, 2019). At this writing, both GEDmatch and FamilyTreeDNA have been searched in this way. Long-range familial search raises a range of privacy concerns (Court, 2018; Ram et al., 2018; Kennett, 2019; Scudder et al., 2019). One response from advocates of long-range search has been to note that "raw genetic data are not disclosed to law enforcement...Search results display only the length and chromosomal location of shared DNA blocks" (Greytak et al., 2018). However, the methods we describe here illustrate that there are several ways to reveal users' raw genetic data on the basis of the locations of shared DNA blocks. Because companies that work with law enforcement on long-range familial searching—including Parabon Nanolabs and Bode Technology (Kennett, 2019)—now routinely upload tens of datasets to genetic genealogy databases, they may be accidentally accumulating information that would allow them to reconstruct many people's genotypes.

Data breaches via IBS tiling, IBS probing, and IBS baiting are preventable. We have identified a set of strategies that genetic genealogy services could adopt to protect their genotype data from IBS-based attacks. We give a detailed list of these strategies in Appendix A (also summarized in Table 2). Broadly, the suggestions consist of restrictions on they types of datasets that can be uploaded, restrictions on the kinds of information shared with users, and restrictions on classes of methods used for identifying putative IBD segments. For example, to prevent IBS tiling,

Strategy	Prevents IBS tiling	Prevents IBS probing	Prevents IBS baiting
Require cryptographic signature from genotyping service	Yes	Yes	Yes
Restrict reporting of IBS to long segments (e.g. >8 cM)	Partially	Partially	Weakly
Report number and lengths of IBS segments but not locations	Yes	No	Partially
Block homozygous uploads	Partially	No	No
Report small number of matching individuals per kit	Partially	Partially	Partially
Disallow matching between arbitrary kits	Partially	Partially	Partially
Block uploads of publicly available genomes	Partially	No	No
Block uploads with evidence of IBS-inert segments	No	Yes	No
Block uploads with long runs of heterozygosity	No	No	Partially
Use phase-aware methods for IBS detection	No	No	Yes

Table 2: Potential countermeasures against the methods of attack outlined here, with their likely effectiveness against IBS tiling, IBS probing, and IBS baiting.

the simplest measures are either to forgo the use of a chromosome browser feature or only to show users the positions of long IBS segments, such as segments of at least 8cM. To prevent IBS baiting, the most robust countermeasure is to phase data before identifying IBS segments, allowing only relatively few phase switches in any putative segment. Phasing the data and only reporting long segments both decrease the uncertainty of IBD calls and so may improve user experience as well. Finally, we also support the strategy of requiring encrypted signatures on uploaded files, proposed by Erlich et al. (2018), which would allow DTC databases to block any files that do not originate from trusted sources. Some of our suggestions limit the potential uses of genetic genealogy data, and users will vary in the degree to which they value these potential uses and in the degree to which they want to protect their genetic information.

All of these suggestions assume that genealogy services will maintain raw genetic data for people in their database. Another possibility would be for individual people instead to upload an encrypted version of their genetic data, with relative matching performed on the encrypted datasets, as has been suggested previously (He et al., 2014).

Our IBS tiling and IBS probing results focus on users of European ancestries, in part because most users of DTC genetic genealogy services appear to have substantial European ancestries. (DTC genetics companies generally do not release this kind of information on their users, but their research papers suggest that they have access to especially large samples with European ancestries—for example, a 23andMe paper on demography in the United States included almost 150,000 self-described European Americans and less than 10,000 each of self-described African Americans and Latino Americans (Bryc et al., 2015). For a qualitatively similar sample composition in a study from Ancestry, see Han et al. (2017).) One question is how these results would generalize to other populations. Because IBD sharing is generally greater within populations than between populations (e.g., Ralph and Coop, 2013), potential users are more vulnerable if there are more publicly available genomes from people with similar ancestries. If IBD-detection algorithms are not well calibrated to differences in heterozygosity across populations, then spurious IBD calls will be more common in populations with lower heterozygosity, leading to greater risk of IBD tiling. Finally, we show in Figure S4 that in our sample, Finnish samples are more vulnerable to IBS tiling than other populations, which is likely due to Finns tracing substantial ancestry to a founder population that experienced a bottleneck 1̃00 generations ago (Kere, 2001). Members of other groups with similar demographic histories are likely to be at elevated risk of IBS tiling and IBS probing as well.

We have focused on genetic genealogy databases that allow uploads because at this writing, it is straightforward to download publicly available genetic datasets and to produce fake genetic datasets for upload. In principle, however, another way to perform attacks like the ones we describe would be to use biological samples. For example, a group of people willing to share their genetic data with each other could collaborate to perform IBS tiling by sending actual biological samples for genotyping. Even IBS probing and IBS baiting could be performed with biological samples by adversaries who could synthesize the samples. Though synthesizing such samples is technically challenging now, it may become easier in the future. Such methods could present opportunities to attack databases that do not allow uploads, such as the large databases maintained by Ancestry (>14 million) and 23andMe (>9 million) (Regalado, 2019). They would also thwart the countermeasure of requiring uploaded datasets to include an cryptographic signature indicating their source.

The IBS-based privacy attacks we describe here add to a growing set of threats to genetic privacy (Homer et al., 2008; Nyholt et al., 2009; Im et al., 2012; Gymrek et al., 2013; Humbert et al., 2015; Shringarpure and Bustamante, 2015; Edge et al., 2017; Ayday and Humbert, 2017; Kim et al., 2018; Erlich et al., 2018). A person's genotype includes sensitive health information that might be used for discrimination, and people whose genetic information is compromised may be vulnerable to scams involving falsified relatives (Ney et al., 2020). Though there are many emerging threats to privacy, some of the more unsettling of which have nothing to do with genetics, genetic data do have special features that might require special considerations. In particular, genetic privacy concerns not only the person whose genotypes are directly revealed but also their relatives whose genotypes may be revealed indirectly (Humbert et al., 2013), a point highlighted by the use of genetic genealogy for long-range forensic searches (Erlich et al., 2018; Edge and Coop, 2019).

Though many forms of genetic discrimination are prohibited legally, rules vary between countries and states. For example, in the United States, the Genetic Information Nondiscrimination Act (GINA) protects against genetic discrimination in the provision of health insurance but does not explicitly disallow genetic discrimination in the provision of life insurance, disability insurance, or long-term care insurance (Bélisle-Pipon et al., 2019). In addition to measures for protecting genetic privacy in the short term, there is a need for more complete frameworks governing the circumstances under which genetic data can be used (Clayton et al., 2019).

4 Methods

4.1 Data assembly

We performed IBS tiling with publicly available genoytpes from 872 people of European ancestries. Of these 872 genotypes, 503 came from the EUR subset of the 20130502 release of phase 3 of the 1000 Genomes project (1000 Genomes Project Consortium, 2012), downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`. This release set has been pruned to remove close biological relatives. The EUR subset includes the following population codes and numbers of people: CEU (Utah residents with Northern and Western European Ancestry, 99 people), FIN (Finnish in Finland, 99 people), GBR (British in England and Scotland, 91 people), IBS (Iberian Population in Spain, 107 people), TSI (Toscani in Italia, 107 people).

The remaining 369 were selected from samples typed on the Human Origins SNP array (Patterson et al., 2012), including 142 genotypes from the Human Genome Diversity Project (Cann et al., 2002). Specifically, we downloaded the Human Origins data from `https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers`, using the 1240K+HO dataset, version 37.2. The 372 selected people were all contemporary samples chosen according to population labels. We also excluded people from the Human Origins dataset if they appeared in the 1000 Genomes dataset. The populations used for selecting data, along with the number of participants included after excluding 1000 Genomes samples, were as follows: "Adygei" (16), "Albanian" (6), "Basque" (29), "Belarusian" (10), "Bulgarian" (10), "Croatian" (10), "Czech" (10), "English" (0), "Estonian" (10), "Finnish" (0), "French" (61), "Greek" (20), "Hungarian" (20), "Icelandic" (12), "Italian_North" (20), "Italian_South" (4), "Lithuanian" (10), "Maltese" (8), "Mordovian" (10), "Norwegian" (11), "Orcadian" (13), "Romanian" (10), "Russian" (22), "Sardinian" (27), "Scottish" (0), "Sicilian" (11), "Spanish" (0), "Spanish_North" (0), and "Ukrainian" (9). The populations with 0 people included are those for which all the samples in the Human Origins dataset are included in the 1000 Genomes phase 3 panel. Samples with group labels marked "ignore" were excluded, including samples marked as close relatives.

We down-sampled the sequence data from the 1000 Genomes project to include only sites typed by the Human Origins chip. Of the 597,573 SNPs included in the Human Origins dataset, 558,257 sites appeared at the same position in the 1000 Genomes dataset, 557,999 of which appear as biallelic SNPs. For 546,530 of these, both the SNP identifier and position match in 1000 Genomes, and for 544,139 of them, the alleles agreed as well. We merged the dataset at the set of 544,139 SNPs at which SNP identifiers, positions, and alleles matched between the Human Origins and 1000 Genomes datasets.

We used vcftools (Danecek et al., 2011), bcftools (Li, 2011), PLINK (Purcell et al., 2007), and EIGENSOFT Price et al. (2006) to create the merged file. The script used to create it is available at `github.com/mdedge/IBS_privacy/`, and the merged data file is available at `https://doi.org/10.25338/B8X619`.

4.2 Phasing, IBS calling, and IBS tiling

We phased the combined dataset using Beagle 5.0 Browning and Browning (2007) using the default settings and genetic maps for each chromosome. We used linear interpolation to obtain the genetic map position of each SNP on the build GRCh37 LDhat genetic map (International HapMap Consortium, 2007) downloaded from the Beagle website (`http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/`). We used Refined IBD software (Browning and Browning, 2013) to identify IBS segments, retaining segments of at least .1 centiMorgans (cM) with LOD scores >1. We also used Germline (Gusev et al., 2009) to identify IBS segments under alternative parameters, shown in the supplement. The resulting IBS segments were analyzed using the GenomicRanges package (Lawrence et al., 2013) in R (R Core Team, 2013). Scripts used for phasing, IBS calling, and IBS tiling are available at `github.com/mdedge/IBS_privacy/`. website.

4.3 IBS probing

To generate IBS-inert genotypes for IBS probing in Figure 3, we computed allele frequencies within the set of 872 Europeans for chromosome 19. Allele frequencies less than 10% were changed to 10%, and then alleles were sampled at one minus their frequency. This strategy generates genetic data that look quite unlike real data, with the advantage (for the purposes of IBS probing) of being unlikely to return IBS matches anywhere. An adversary attempting IBS probing in a real database would need to tailor the approach to the quality control and IBS calling methods used by the database.

After inert genotypes were produced, we stitched them with real phased genotypes from windows around GRCh position 45411941 on chromosome 19, the site of SNP rs429358. SNP rs429358 is in the APOE locus; if a haplotype has a C at rs429358 and a C at nearby SNP rs7412, then that haplotype is said to harbor the APO-$\epsilon 4$ allele, which confers risk for Alzheimer's disease Corder et al. (1993). rs429358 is not genotyped on the Human Origins chip, but it is included on recent chips used by both Ancestry and 23andMe. To simulate probing with a 1cM threshold for matching, we pulled real data from a region of 1.9cM around the site, and to simulate probing with a 3cM threshold, we pulled real data from a region of 5.9cM around the site. Distances in cM were computed by linear interpolation from a genetic map in GRCh37 coordinates. Scripts used to generate Figure 3 are available at `github.com/mdedge/IBS_privacy/`.

4.4 GEDmatch demonstration

On Novermber 21st, 2019, we first uploaded artificial genetic datasets to GEDmatch's research mode in order to demonstrate the possibility of IBS baiting. GEDMatch has not published details of its IBS detection procedures. However, the options available to users in the 1-to-1 match tool and the description of how those options can be used to ignore single-site matches led us to hypothesize that GEDmatch uses phase-unaware IBS detection and that the 1-to-1 match tool might be vulnerable to IBS baiting.

Description of GEDmatch 1-to-1 tool GEDmatch's 1-to-1 match tool allows the user to compare the IBS matches of any two genetic datasets (or, in GEDmatch parlance, "kits"), as long as the kit numbers are known to the user. Thus, to identify the genotypes of many users an adversary would need access to the kit numbers of many users. The 1-to-many tool in default GEDmatch reports 3000 of the closest genetic relatives of any kit whose number is known to the user, and reports the kit numbers of those match kits (along with names and email addresses). Thus an adversary can iteratively search for all the kit numbers matching a known kit, and so obtain many kit numbers to use in 1-to-1 searches. We alerted GEDmatch to this issue with the 1-to-many tool, as nearly the entire GEDmatch database of kit numbers and genetic relationships could be scraped.

The 1-to-1 match tool allows the user to specify parameters that govern IBS calling. In particular, the user can specify the minimum cM length of the blocks (down to 0.1cM) and the minimum number of SNPs in a block (down to 25 SNPs). GEDmatch also allows the user to specify the ‘mis-match bunch limit,’ which appears to be the minimum number of IBS-compatible SNPs after an opposite-homozygous site that are required in order for a second opposite-homozygous site not to break the IBS segment.

Ethics. In order to comply with GEDmatch’s terms and conditions, we used artificial datasets designed not to match any genuine genetic data uploaded to GEDmatch. The kits were uploaded in "Research" mode, where they are not visible to other users via 1-to-many search. We did not interact with any other users’ data; we ran GEDmatch’s 1-to-1 comparison tools only comparing among our artificial kits. We exercised care not to interact with any other tools and to avoid accidental discoveries. Prior to uploading the artificial datasets, we also consulted with the UC Davis Institutional Review Board to ensure that these uploads do not constitute human subjects research. Upon receiving confirmation from the IRB that our uploads do not constitute human subjects research and before uploading the datasets, we alerted GEDmatch that we would be making the uploads, and we also shared the kit numbers with them after we had completed our analyses.

4.4.1 Construction of artificial datasets

We constructed artificial "target" and "bait" kits using the SNPs included in the 23&Me v4 chip. (The "target" kits are the targets of inference, and the "bait" kits are designed to reveal their genotypes.) We identified the alleles at these SNP positions in the 1000Genomes dataset, along with their frequencies in the EUR subset of 1000Genomes. We assigned as missing (‘- -’) any SNP that we could not match by position in 1000Genomes. We chose four target SNPs at random on chromosome 22. These SNPs were chosen at random from the set of strand-unambiguous polymorphisms, i.e. not A/T and G/C SNPs. These strand-unambiguous sites include the majority of SNPs on the chip, e.g. 89% of the SNPs on the 23&Me chip on chromosome 22.

Target genomes We uploaded three artificial target genome kits (T1-T3). These vary in their genotypes at the target SNPs. T1 and T3 are homozygous for different alleles; T2 is heterozygous. At the rest of the loci, we constructed genotypes by randomly sampling alleles according to their frequencies at each SNP. Thus, there is no LD among loci.

Bait genomes. We uploaded two artificial bait genome kits. These two kits have opposite-homozygote genotypes at each of the target SNPs. The two bait uploads were then set to have identical genotypes in the rest of their autosomes, with their genotypes specified as below.

To create a region around the target that would bait a phase-unaware method into calling IBD, we took SNPs in the 0.6cM on either side of the target SNP, selected at random 22 on each side, and set them to be heterozygous in both bait genomes. The rest of the SNPs within this bait region were set to be missing. We used only 22 heterozygous SNPs on each side and filled in the rest with missing data (rather than making all sites heterozygous) because large numbers of heterozygous sites generated an error on upload, "HTZ string too long" and would not be processed further. Blocking uploads with long runs of heterozygous sites is a countermeasure put in place by GEDmatch after we and Ney et al. initially alerted GEDmatch to the risks of upload-based privacy attacks. However, we found that the countermeasure was not triggered by runs of heterozygous sites with missing sites interspersed, and these runs of heterozygosity interspersed with missingness also effectively baited GEDmatch into calling IBD segments. Additionally, we confirmed with Peter Ney (personal communication) that his previously-uploaded kits including long runs of heterozygosity remain active even though re-uploads of those same kits are blocked as of December 3rd, 2019, suggesting that the block applies only to newly uploaded kits and not to existing data on GEDmatch.

The alleles in target kits at all other autosomal SNPs in the genome were drawn at random with frequency $1-p$, where p is the frequency in the 1000Genomes EUR subsample. This scheme was chosen to ensure hat the bait genomes were unlikely to have spurious IBS matches anywhere with any target genome, so that the only potential IBS was in the target regions.

4.4.2 Detailed results of baiting

We compared each target to both bait genomes using the 1-to-1 GEDmatch tool. We set the minimum block to a length of > 0.7cM and 25 SNPs, with a mismatch cutoff of 25 SNPs. This ensured that we could detect IBS in the key regions, but that a single opposite-homozygous mismatch would be sufficient to prevent the identification of a putative IBS segment in the key region.

The baiting attempt was successful; we observed IBS only where we expected it between bait and target kits (Figure 5). We observed no putative IBD segments on any chromosome except 22, as expected on the basis of our procedure for filling in artificial genotypes in both sets of kits. The details of the matches on chromosome 22 are reported in Table 3. We observed 4 putative IBD segments overlapping our target bait regions in the comparisons with matching homozygote genotypes at the bait site, i.e. in the T1-B1 and T3-B2 comparisons, as well in both heterozygote-homozygote comparisons, i.e. T2-B1 and T2-B2. We observed no putative IBD segments in the pairs with opposite-homozygous mismatches, T1-B1 and T3-B2. Thus the genotypes of the targets are readily discernable from from the putative IBD segments output by GEDmatch. The full results returned by GEDmatch are available as images here (`https://github.com/mdedge/IBS_privacy/tree/master/IBS_baiting_demo`; the kit numbers are redacted to prevent reuse).

Some of the IBS blocks have fewer SNPs than we expect. We believe this to be due to the removal of SNPs during the tokenization stage, during which rare SNPs and SNPs with strand-ambiguous alleles seem to be removed Ney et al. (2020). We did not investigate this further, but multiple uploads could be used to determine the approximate criteria for SNPs to be included, and hence determine where an adversary should set cutoffs.

Our two bait kits could both generate IBS matches to the target because the target genotype is

Matching pairs		Target 1	Target 2	Target 3	Target 4
	target bp	27613130	34024097	37673781	42008068
T1-(B1 B?)					
	IBS L bp	27427698	33771672	37519864	40054428
	IBS R bp	27680780	34328741	37827711	43112674
	IBS cM	1.3	0.8	1.1	1.2
	# SNPs	47	45	42	40
	# SNPs B?	46	44	41	39
T2-(B1 B2 B?)					
	IBS L bp	27433179	33771672	37508507	40357667
	IBS R bp	27680780	34328741	37827711	43112674
	IBS cM	1.3	0.8	1.2	0.9
	# SNPs	45	45	45	32
	# SNPs B?	44	44	44	31
T3-(B3 B?)					
	IBS L bp	27433179	33771672	37519864	40357667
	IBS R bp	27680780	34328741	37827711	43112674
	IBS cM	1.3	0.8	1.1	0.9
	# SNPs	45	45	45	32
	# SNPs B?	44	44	41	31
T?-(All Baits)					
	IBS L bp	27433179	33771672	37519864	40357667
	IBS R bp	27680780	34328741	37827711	43112674
	IBS cM	1.3	0.8	1.1	0.9
	# SNPs	44	44	44	31
	# SNPs B?	44	44	44	31

Table 3: Summary of the SNPs targeted by baiting and the IBS returned by GEDmatch. For each region, we give the position of the key SNP (target bp). Because by design our bait kits are genetically identical outside of the target SNPs, the IBS regions returned by GEDmatch's 1-to-1 tool are identical across bait kits generating a match. For each pairwise comparison, we report the IBS information returned: Left-Right bp of the IBS region, the cM length, the number (#) of SNPs in the IBS region with a non-missing target. We also report the number (#) of SNPs spanned by the region IBS when matched to the missing target Bmiss.

missing rather than heterozygous. To determine whether a genotype was missing, we implemented a trick borrowed from Ney et al. (2020), and uploaded a third bait kit (Bmiss) with the target SNP set to missing (i.e. '- -') and then looked at the number of SNPs an IBS match across the target site spans. In each case, the non-missing baits (B1 and B2) generated an IBS block match with with T1-T3 that was one SNP longer than the IBS block generated by the Bmiss bait (Table 3). Comparing these baits to a new target with a missing genotype at each target site (T?), we see that in each pairwise comparison the IBS blocks are the same number of SNPs long regardless of whether the target SNP bait genotype was missing (Table 3). Therefore, we can distinguish the target being heterozygote or missing by the use of a third bait kit and inspection of the number of SNPs included in a IBS match.

The possibility of IBS-baiting-like procedures also interacts with the vulnerabilities arising from the presentation of SNP-level visualizations explored by Ney et al.. Even if short IBS blocks were not reported to the user explicitly, it is clear from the zoomed-in view that we can see the target mismatches in question (see Figure 5). One measure that GEDmatch appears to have taken against a Ney et al.-style attack is to jitter the positions of SNPs in their visualization slightly. However, an attacker could counter such jittering by embedding key sites in runs of heterozygosity, making it easier to identify them in visualizations after jittering. Thus, the images displayed by GEDmatch still pose additional security risks.

Acknowledgments

We thank Matt Bishop, Elizabeth Joh, Peter Ney, and Mike Sweeney for useful conversations, and we thank Shai Carmi, Yaniv Erlich, Debbie Kennett, Leah Larkin, Magnus Nordborg, Rori Rohlfs, Noah Rosenberg, Ann Turner, Amy Williams, and an anonymous reviewer for helpful comments on the manuscript. Swapan Mallick and David Reich answered questions about the Human Origins dataset, Brian Browning answered questions about Refined IBD, and Alexander Gusev answered questions about Germline software. We acknowledge support from the National Institutes of Health (R01-GM108779 and F32-GM130050).

References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*(7422), 56.

Ayday, E. and M. Humbert (2017). Inference attacks against kin genomic privacy. *IEEE Security & Privacy 15*(5), 29–37.

Bélisle-Pipon, J.-C., E. Vayena, R. C. Green, and I. G. Cohen (2019). Genetic testing, insurance discrimination and medical research: what the united states can learn from peer countries. *Nature Medicine 25*(8), 1198–1204.

Bjelland, D. W., U. Lingala, P. S. Patel, M. Jones, and M. C. Keller (2017). A fast and accurate method for detection of ibd shared haplotypes in genome-wide snp data. *European Journal of Human Genetics 25*(5), 617.

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

Browning, S. R. and B. L. Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics 81*(5), 1084–1097.

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

Bryc, K., E. Durand, J. Macpherson, D. Reich, and J. Mountain (2015). The genetic ancestry of african americans, latinos, and european americans across the united states. *The American Journal of Human Genetics 96*(1), 37 – 53.

Buffalo, V., S. M. Mount, and G. Coop (2016). A genealogical look at shared ancestry on the x chromosome. *Genetics 204*(1), 57–75.

Cann, H. M., C. De Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al. (2002). A human genome diversity cell line panel. *Science 296*(5566), 261–262.

Carmi, S., K. Y. Hui, E. Kochav, X. Liu, J. Xue, F. Grady, S. Guha, K. Upadhyay, D. Ben-Avraham, S. Mukherjee, et al. (2014). Sequencing an ashkenazi reference panel supports population-targeted personal genomics and illuminates jewish and european origins. *Nature Communications 5*, 4835.

Carmi, S., P. F. Palamara, V. Vacic, T. Lencz, A. Darvasi, and I. Pe'er (2013). The variance of identity-by-descent sharing in the wright–fisher model. *Genetics 193*(3), 911–928.

Clayton, E. W., B. J. Evans, J. W. Hazel, and M. A. Rothstein (2019). The law of genetic privacy: applications, implications, and limitations. *Journal of Law and the Biosciences*, 1–36.

Conomos, M., A. Reiner, B. Weir, and T. Thornton (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics 98*(1), 127 – 148.

Corder, E. H., A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. D. Roses, J. Haines, and M. A. Pericak-Vance (1993). Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science 261*(5123), 921–923.

Court, D. S. (2018). Forensic genealogy: Some serious concerns. *Forensic Science International: Genetics 36*, 203–204.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, et al. (2011). The variant call format and vcftools. *Bioinformatics 27*(15), 2156–2158.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology 23*(1), 34–63.

Durand, E. Y., N. Eriksson, and C. Y. McLean (2014, 04). Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution 31*(8), 2212–2222.

Edge, M. D., B. F. B. Algee-Hewitt, T. J. Pemberton, J. Z. Li, and N. A. Rosenberg (2017). Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proceedings of the National Academy of Sciences 114*(22), 5671–5676.

Edge, M. D. and G. Coop (2019). How lucky was the genetic investigation in the golden state killer case? *bioRxiv*.

Erlich, Y. and A. Narayanan (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics 15*(6), 409.

Erlich, Y., T. Shor, I. Pe'er, and S. Carmi (2018). Identity inference of genomic data using long-range familial searches. *Science 362*(6415), 690–694.

Greshake, B., P. E. Bayer, H. Rausch, and J. Reda (2014). Opensnp–a crowdsourced web resource for personal genomics. *PLoS One 9*(3), e89204.

Greytak, E., D. H. Kaye, B. Budowle, C. Moore, and S. Armentrout (2018). Privacy and genetic genealogy data. *Science 361*, 857.

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

Gymrek, M., A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich (2013). Identifying personal genomes by surname inference. *Science 339*(6117), 321–324.

Han, E., P. Carbonetto, R. E. Curtis, Y. Wang, J. M. Granka, J. Byrnes, K. Noto, A. R. Kermany, N. M. Myres, M. J. Barber, et al. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of north america. *Nature Communications 8*, 14238.

He, D., N. A. Furlotte, F. Hormozdiari, J. W. J. Joo, A. Wadia, R. Ostrovsky, A. Sahai, and E. Eskin (2014). Identifying genetic relatives without compromising privacy. *Genome Research 24*(4), 664–672.

Henn, B. M., L. Hon, J. M. Macpherson, N. Eriksson, S. Saxonov, I. Pe'er, and J. L. Mountain (2012, 04). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE 7*(4), 1–13.

Hogarth, S., G. Javitt, and D. Melzer (2008). The current landscape for direct-to-consumer genetic testing: Legal, ethical, and policy issues. *Annual Review of Genomics and Human Genetics 9*(1), 161–182. PMID: 18767961.

Hogarth, S. and P. Saukko (2017). A market in the making: the past, present and future of direct-to-consumer genomics. *New Genetics and Society 36*(3), 197–208.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics 4*(8), e1000167.

Hon, L., S. Saxonov, B. T. Naughton, J. L. Mountain, A. Wojcicki, and L. Avey (2013, June 11). Finding relatives in a database. US Patent 8,463,554.

Huang, L., S. Bercovici, J. M. Rodriguez, and S. Batzoglou (2014). An effective filter for ibd detection in large data sets. *PloS ONE 9*(3), e92713.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins, Y. Zhang, T. M. Tuohy, D. W. Neklason, R. W. Burt, S. L. Guthery, et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome Research 21*(5), 768–774.

Humbert, M., E. Ayday, J.-P. Hubaux, and A. Telenti (2013). Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & Communications Security*, pp. 1141–1152. ACM.

Humbert, M., K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux (2015). De-anonymizing genomic databases using phenotypic traits. *Proceedings on Privacy Enhancing Technologies 2015*(2), 99–114.

Im, H. K., E. R. Gamazon, D. L. Nicolae, and N. J. Cox (2012). On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics 90*(4), 591–598.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million snps. *Nature 449*(7164), 851.

Kennett, D. (2019). Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Science International 301*, 107 – 117.

Kere, J. (2001). Human population genetics: Lessons from finland. *Annual Review of Genomics and Human Genetics 2*(1), 103–128. PMID: 11701645.

Khan, R. and D. Mittelman (2018). Consumer genomics will change your life, whether you get tested or not. *Genome Biology 19*(1), 120.

Kim, J., M. D. Edge, B. F. Algee-Hewitt, J. Z. Li, and N. A. Rosenberg (2018). Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell 175*(3), 848 – 858.e6.

Larkin, L. (2017, Mar). Cystic fibrosis: A case study in genetic privacy.

Larkin, L. (2018, Sept). Database sizes—september 2018 update.

Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology 9*(8), e1003118.

Li, H. (2011, 09). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics 27*(21), 2987–2993.

Loh, P.-R., P. F. Palamara, and A. L. Price (2016). Fast and accurate long-range phasing in a uk biobank cohort. *Nature Genetics 48*(7), 811.

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, et al. (2008). Runs of homozygosity in european populations. *The American Journal of Human Genetics 83*(3), 359–372.

Naveed, M., E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang (2015, August). Privacy in the genomic era. *ACM Computing Surveys 48*(1), 6:1–6:44.

Ney, P., L. Ceze, and T. Kohno (2020). Genotype extraction and false relative attacks: Security risks to third-party genetic genealogy services beyond identity inference. Network and Distributed System Security Symposium (NDSS). Preprint Posted 10/29/19, `https://dnasec.cs.washington.edu/genetic-genealogy/ney_ndss.pdf`.

Ney, P. M., L. Ceze, and T. Kohno (2018). Computer security risks of distant relative matching in consumer genetic databases. *CoRR abs/1810.02895*.

Nyholt, D. R., C.-E. Yu, and P. M. Visscher (2009). On jim watson's apoe status: genetic information is hard to hide. *European Journal of Human Genetics 17*(2), 147.

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

Panoutsopoulou, K., K. Hatzikotoulas, D. K. Xifara, V. Colonna, A.-E. Farmaki, G. R. Ritchie, L. Southam, A. Gilly, I. Tachmazidou, S. Fatumo, et al. (2014). Genetic characterization of greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nature Communications 5*, 5345.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich (2012). Ancient admixture in human history. *Genetics 192*(3), 1065–1093.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li (2012). Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics 91*(2), 275–292.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*(8), 904.

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Ram, N., C. J. Guerrini, and A. L. McGuire (2018). Genealogy databases and the future of criminal investigation. *Science 360*(6393), 1078–1079.

Ramstetter, M. D., T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, J. G. Mezey, and A. L. Williams (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics 207*(1), 75–82.

Ramstetter, M. D., S. A. Shenoy, T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, J. G. Mezey, and A. L. Williams (2018). Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. *The American Journal of Human Genetics 103*(1), 30–44.

Regalado, A. (2019). More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*.

Rohlfs, R. V., S. M. Fullerton, and B. S. Weir (2012, 02). Familial identification: Population structure and relationship distinguishability. *PLOS Genetics 8*(2), 1–13.

Scudder, N., D. McNevin, S. F. Kelty, C. Funk, S. J. Walsh, and J. Robertson (2019). Policy and regulatory implications of the new frontier of forensic genomics: direct-to-consumer genetic data and genealogy records. *Current Issues in Criminal Justice 31*(2), 194–216.

Shi, S., N. Yuan, M. Yang, Z. Du, J. Wang, X. Sheng, J. Wu, and J. Xiao (2018). Comprehensive assessment of genotype imputation performance. *Human Heredity 83*(3), 107–116.

Shringarpure, S. S. and C. D. Bustamante (2015). Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics 97*(5), 631–646.

Staples, J., D. J. Witherspoon, L. B. Jorde, D. A. Nickerson, J. E. Below, C. D. Huff, U. of Washington Center for Mendelian Genomics, et al. (2016). Padre: pedigree-aware distant-relationship estimation. *The American Journal of Human Genetics 99*(1), 154–162.

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics 194*(2), 301–326.

Appendix A Detailed rationale for proposed countermeasures

Here, we detail the rationale and possible advantages and disadvantages of the countermeasures listed in Table 2.

1. **Require uploaded files to include cryptographic signatures identifying their source.** This recommendation was initially made by Erlich et al. (2018). Under this suggestion, DTC genetics services would cryptographically sign the genetic data files they provide to users. Upload services might then check for a signature from an approved DTC service on each uploaded dataset, blocking datasets from upload otherwise. An alternative procedure that would accomplish the same goal would be for the DTC entities to exchange data directly at the user's request (Ney et al., 2018). Such a procedure would allow upload services to know the source of the files they analyze and to disallow uploaded datasets produced by non-approved entities and user-modified datasets. All the methods we describe require the upload of multiple genetic datasets. As such, requiring cryptographic signatures would force the adversary to have multiple biological samples analyzed by a DTC service in order to implement any of our procedures, and IBS probing and IBS baiting would require synthetic samples, which are much harder to produce than fake datasets. Another benefit of this approach is that it would protect research participants against being reidentified using DTC genetic genealogy services (Erlich et al., 2018). A disadvantage of this strategy is that it requires the cooperation of several distinct DTC services.

2. **Restrict reporting of IBS to long segments.** Reporting short IBS segments increases the typical coverage of IBS tiling (Figure 2) and IBS probing (3), as well as the efficiency of IBS baiting. Very short blocks may be of little practical utility for genetic genealogy (Huff et al., 2011). Reporting only segments longer than 8 cM would substantially limit IBS tiling attacks. A partially effective variant of this strategy is to report short segments only for pairs of people who share at least one long segment (Figure S2). One disadvantage is that short segments, though less reliably inferred than longer segments, may still be of interest to genealogists.

3. **Do not report locations of IBS segments.** Another tactic for preventing IBS tiling is not to report chromosomal locations at all. If chromosomal locations are not reported, IBS tiling as we have described it becomes impossible.

4. **Block uploads of genomes with excessive homozygosity.** IBS tiling is especially informative if genotypes that are homozygous for phased haplotypes are uploaded, so blocking genomes with excessive homozygosity presents a barrier to IBS tiling attacks. However, runs of homozygosity occur naturally (Pemberton et al., 2012), and allowing for naturally occurring patterns of homozygosity would leave a loophole for an adversary who could upload many genotypes, using including homozygous regions and using only those for tiling.

5. **Report only a small number of putative relatives per uploaded kit.** Reporting only the closest relatives (say the ≈ 50 closest relatives) of an uploaded kit would decrease the efficiency of all the methods we describe here. Only a small number of people could

have their privacy compromised by each upload. This countermeasure comes with costs to genealogists, who may want to explore as many matches as possible in order to build family trees.

6. **Disallow arbitrary matching between kits.** Some services allow searches for IBS between any pair of individuals in the database. Allowing such searches makes all potential IBS attacks easier. This countermeasure might hamper the investigations of genealogists exploring complex hypotheses about relatedness.

7. **Block uploads of publicly available genomes.** There are now thousands of genomes available for public download, and these publicly available genomes can be used for IBS tiling. Genetic genealogy databases could include publicly available genomes (potentially without allowing them to be returned as IBS matches for typical users) and flag accounts that upload them. This strategy would go some distance toward blocking IBS tiling, but it could be thwarted in several ways, for example by uploading genetic datasets produced by splicing together haplotypes from publicly available genomes.

8. **Block uploads with evidence of IBS-inert segments.** IBS-inert segments—i.e. false genetic segments designed to be unlikely to be IBS with anyone in the database—are key to IBS probing. Some methods for constructing IBS-inert segments are easy to identify, but others may not be. If a database is large enough, genomes with IBS-inert segments could be identified by looking for genomes that have much less apparent IBS with other database members than might be expected.

9. **Block uploads with long runs of heterozygosity.** Long runs of heterozygosity do not arise naturally but are key to the IBS baiting approaches we describe here. Blocking genomes with long runs of heterozygosity—or alternatively, blocking genomes that have much more apparent IBS with a range of other database members than expected—would hamper IBS baiting. However, this countermeasure might be hard to apply to a small-scale IBS baiting attack, where only one or a few short runs of heterozygosity might be necessary. In our sample, the longest run of heterozygosity (in terms of number of SNPs) consisted of 38 SNPs and spanned .06 cM. This suggests that filtering out long runs of heterozygosity might be a promising strategy, though identifying a specific procedure would require more careful consideration of variation in non-European populations and of the composition of commercial SNP chips (including SNP density and allele frequencies).

10. **Use phase-aware methods for IBS detection.** Although calling IBS by looking for long segments without incompatible homozygous genotypes scales well to large datasets, such methods are easy to trick, allowing IBS baiting approaches. In addition to allowing IBS estimation methods that are harder to trick, faked samples may stand out as unusual during the process of phasing, raising more opportunities for quality-control checks.

Supplementary Figures

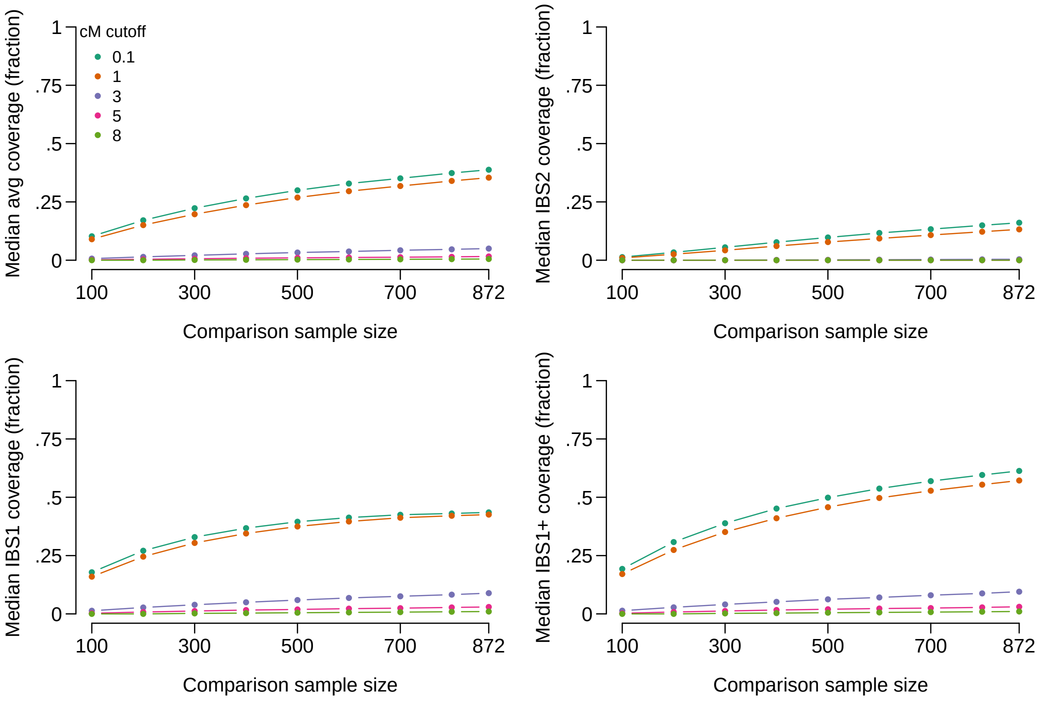

Figure S1: Tiling performance with IBS segments that are unlikely to be IBD filtered out. Conventions are the same as in Figure 2; the difference is that now only IBS segments that represent likely IBD (LOD score > 3) are included. As expected, the amount of tiling possible is reduced when the LOD score threshold is increased, particularly when segments as short as 1 cM are allowed. However, tiling still reveals a substantial amount of information about target genotypes. Using a comparison sample of 871, and including all called IBS segments >1 cM, the median person has an average of 35% of the maximum length of 2.8 Gbp covered by IBD segments with LOD >3, and has at least one chromosome covered for approximately 57% of the genome. If only segments >3 cM are included, then averaging across the two chromosomes, median coverage is 5.0%, and the median proportion for which at least one chromosome is covered is 9.5%. As before, the percentage of the genome recoverable by tiling varies among people, and some people still have large proportions of their genetic data recoverable by tiling. With a LOD score threshold of 3, the top 10% of people have at least 58% of their total genotype information covered by IBD tiles, including one or more alleles at sites in at least 81% of the genome covered by IBD tiles.

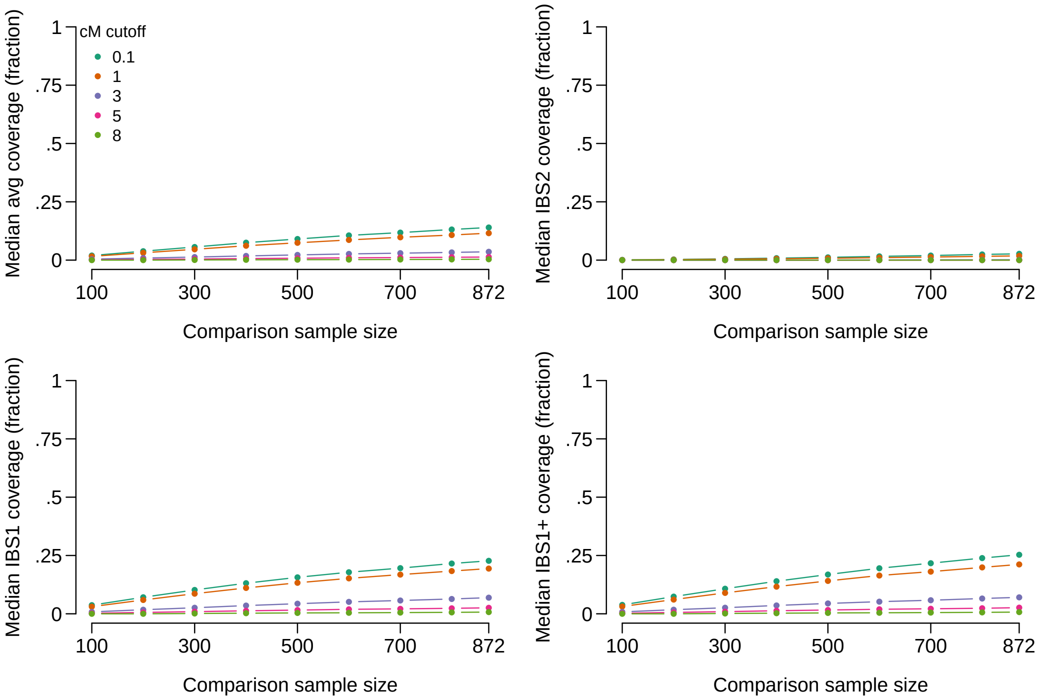

Figure S2: IBS tiling performance, limiting to comparison samples who share at least 1 IBS segment of 8 cM or more with the target. Conventions are the same as in Figure 2. Some DTC genetics companies use a two-step approach for reporting IBS information to users. For example, at this writing, MyHeritage identifies people who are likely matches of a given user as all those who share an apparent IBD segment of at least 8 cM with the user. However, once matches are identified, inferred IBD segments down to a minimum length of 6 cM are reported to the user (see Table 1). Similarly, FamilyTreeDNA only reports matching segments for pairs of people who pass a sharing threshold, and for those pairs of individuals they report all matches down to 1cM. As expected, reporting only IBS segments for pairs of people who share at least one long IBS segment (>8 cM) substantially reduces but does not eliminate the effectiveness of IBS tiling. With 872 comparison samples, the median person has approximately 12% of their genome covered by IBS tiles of 1 cM or more (averaged across both chromosomes) and at least one chromosome covered for 21% of the genome. People in the top 10% of IBS tiling coverage have 44% of their genome length recoverable by tiling (averaging across both chromosomes), with at least one chromosome tiled over more than 67% of the genome. Importantly, the practice of requiring at least one long IBS match in order to report any IBS segments will not reduce the effectiveness of IBS tiling if phase-unaware methods are used for calling IBS. In that case, the attacker could simply insert a long run of heterozygous sites in each of the genomic datasets uploaded, causing an apparent long run of IBS with every user in the database (see section 2.3). After getting "in the door" with a long run of heterozygous sites, the attacker could then use tiling to find out about the rest of the genome.

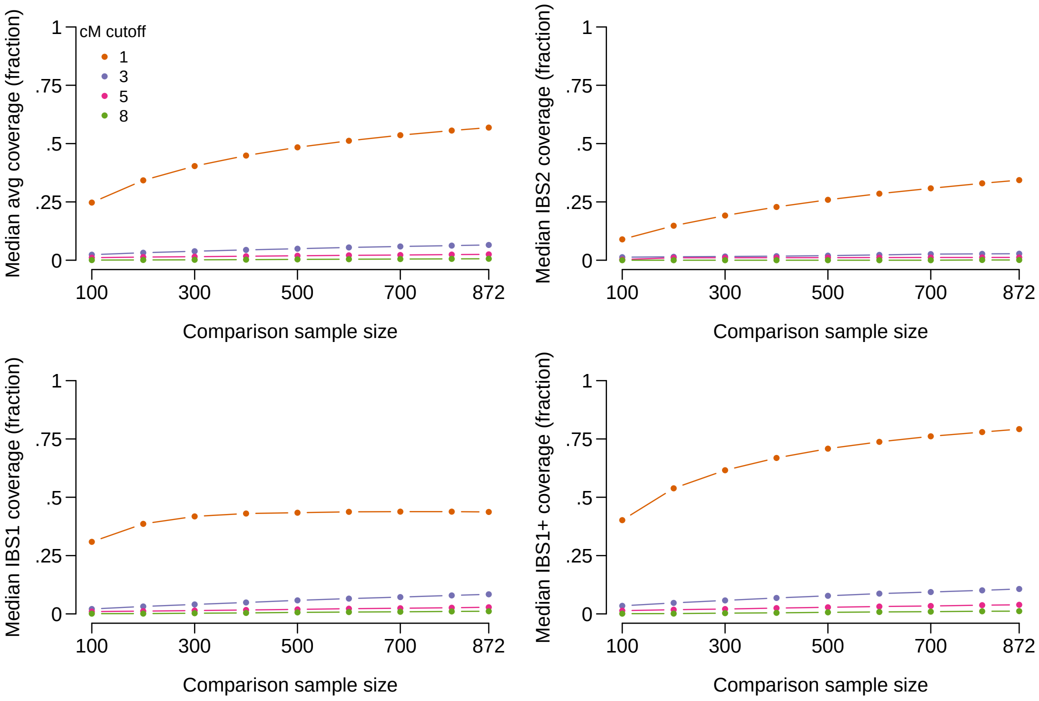

Figure S3: IBS tiling performance when genotype phasing switches are disallowed. Conventions are the same as in the Figure 2. We called IBS segments using Germline (Gusev et al., 2009), using the haploid flag to find IBS segments within the phased chromosomes produced by Beagle. We also set the err_hom argument to zero, set the bits argument to 32 to increase sensitivity for short segments, used the w_extend flag to extend segments beyond the slices produced by Germline, and set the minimum IBS segment length to 1cM. (Setting the minimum segment length shorter than 1cM does not appear to be possible in Germline.) The amount of tiling possible is reduced somewhat when phase switches are disallowed. However, tiling still reveals substantial information about target genotypes. Using a comparison sample of 871, and including all called IBS segments >1 cM, the median person has an average of 57% of the maximum length of 2.8 Gbp covered by IBS segments, and has at least one chromosome covered for approximately 79% of the genome. If only segments >3 cM are included, then averaging across the two chromosomes, median coverage is 6.5%, and the median proportion for which at least one chromosome is covered is 11%. The top 10% of people have at least 73% of their genomes covered by IBS tiles of 1 cM or more, including one or more alleles at sites in at least 91% of the genome covered by IBS tiles.

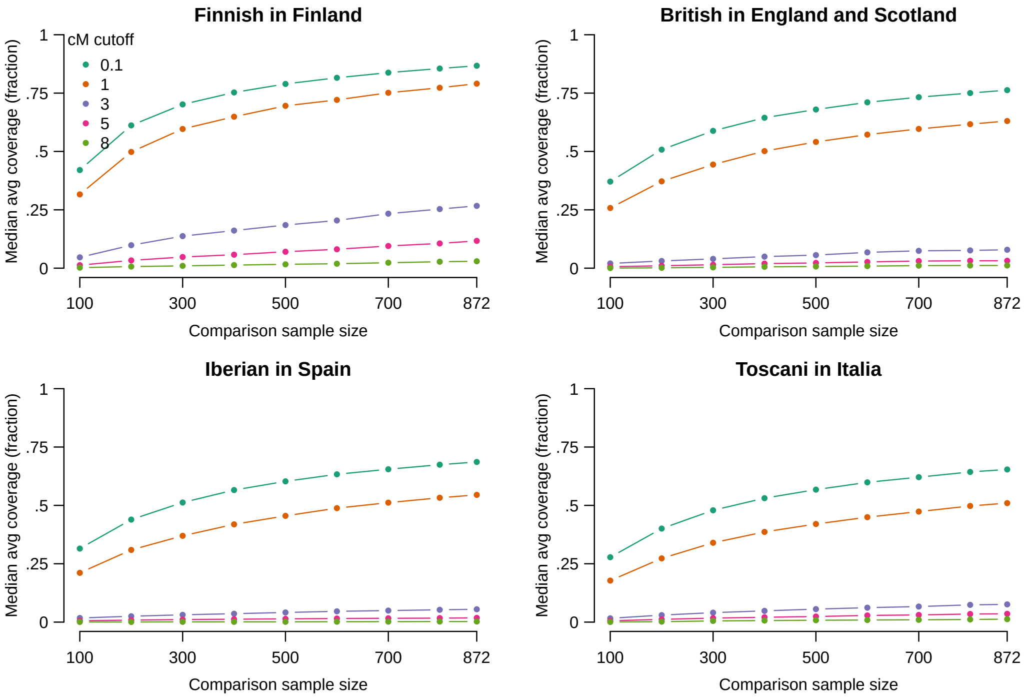

Figure S4: IBS tiling performance in selected populations. We examined IBS tiling performance in four European populations from the 1000Genomes data—Finnish in Finland (FIN, top left, 99 people), British in England and Scotland (GBR, top right, 91 people), Iberian in Spain (IBS, bottom left, 107 people), and Toscani in Italia (TSI, bottom left, 107 people). In all panels, median average IBS tiling coverage is shown on the vertical axis using refinedIBD with LOD>1, as in the top-left panel of Figure 2. Median tiling accuracy varies among populations. For example, using IBS tiles >1cM and with all 871 other individuals used in the comparison sample, the median coverage percentages by population were 79% (FIN), 63% (GBR), 55% (IBS), and 51% (TSI). The most striking population difference is the higher IBS tiling rates attained among Finns, who have long been of interest as a founder population, having experienced a bottleneck approximately 100 generations ago (Kere, 2001). Another factor that likely influences these results is the genetic similarity of members of each population to members of other populations included—for example, the CEU population of 1000Genomes is closely related to the GBR population, and the inclusion of CEU may partially account for the higher tiling rates in GBR than in IBS or TSI.

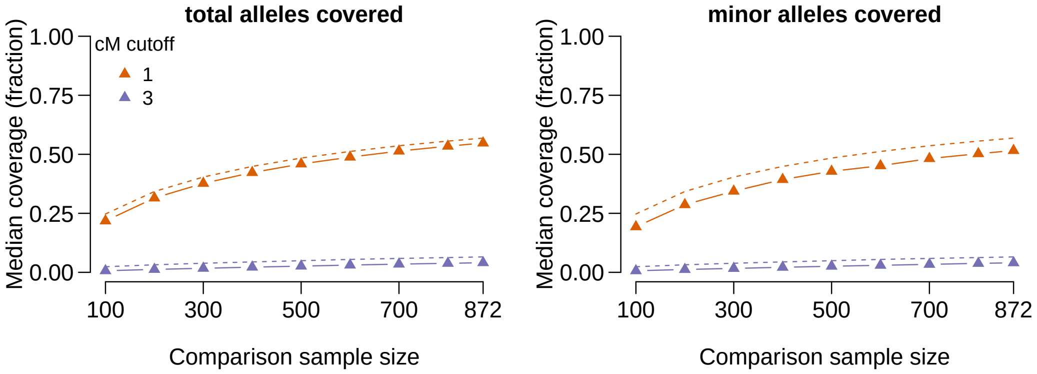

Figure S5: IBS tiling performance in terms of number of total alleles covered (left panel) and number of minor alleles covered (right panel, 18.6% of total alleles were minor alleles). We used Germline in haploid mode (as in Figure S3), as it allows easier identification of which allele is covered by a given IBS segment. Dashed lines show the results in terms of fraction of base pairs covered, whereas the solid lines show results in terms of alleles covered. Results for cM cutoffs <1 are not shown because they cannot be run in Germline, and results for cM cutoffs >3 are not shown because it is difficult to distinguish the dashed and solid lines. As would be expected, there is a slight bias for IBS tiles to fall in regions with lower SNP density, leading to slightly fewer alleles on the chip being covered than would be expected on the basis of total base pairs covered. For example, with a 1cM cutoff and all samples included, the median is 57% of the genome length in base pairs covered by IBS tiles, whereas the median proportion of total alleles covered by IBS tiles is 55%. It also appears that IBS tiles may be more likely to appear in regions that are less genetically diverse, as the proportion of minor alleles tiled is slightly lower than the proportion of total alleles covered. For example, with a 1cM cutoff and all samples included, the proportion of minor alleles covered by IBS tiles is 52%.

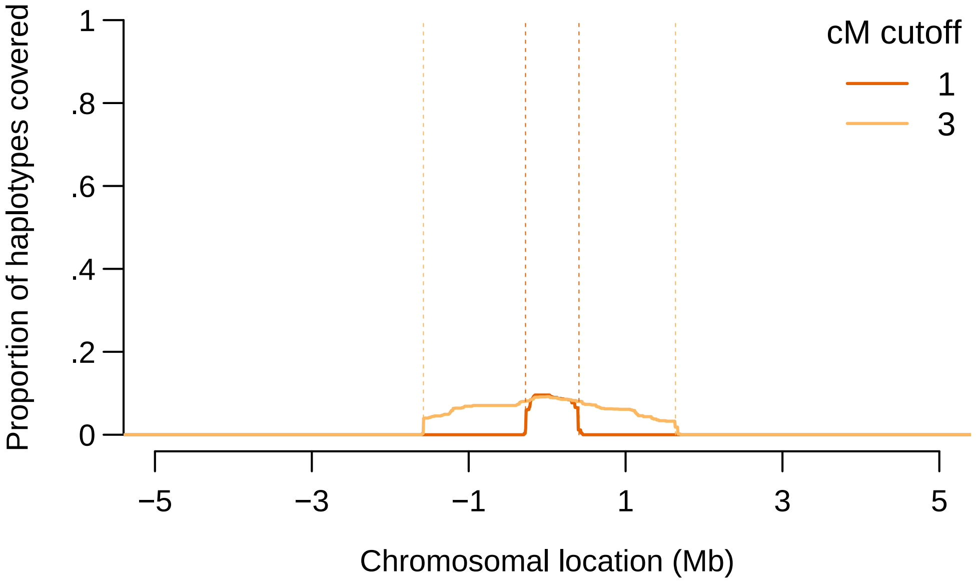

Figure S6: A demonstration of the IBD probing method around position 45411941 on chromosome 19 (GRCh37 coordinates), in the APOE locus. Conventions are the same as in Figure 3; the difference is that only IBS segments with a LOD score >3 for IBD are included. When IBD probing is performed with a 1-cM threshold, 9.6% of haplotypes had a match among the probes constructed from the other 871 people in the dataset. With a 3-cM threshold, 9.2% of haplotypes had a match.

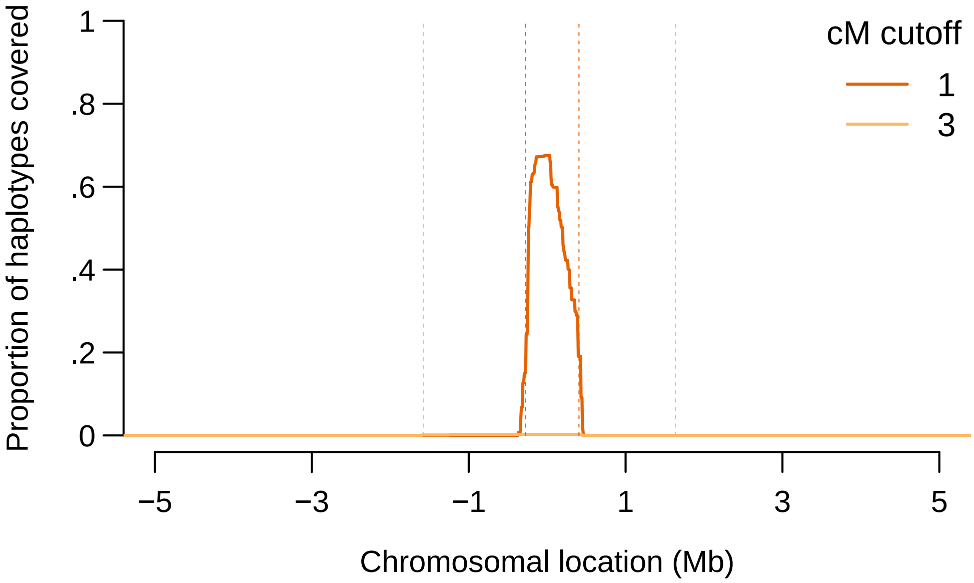

Figure S7: A demonstration of the IBS probing method around position 45411941 on chromosome 19 (GRCh37 coordinates), in the APOE locus. Conventions are the same as in Figure 3; the difference is that IBS calling was performed by Germline (Gusev et al., 2009) in haploid mode, meaning that phasing switches are disallowed. We set the err_hom argument to zero, we used the w_extend flag to extend segments beyond the slices produced by Germline, and we set the minimum IBS segment length to 1cM. All other arguments were kept at their default values. When IBS probing is performed with a 1-cM threshold, 67.5% of haplotypes had a match among the probes constructed from the other 871 people in the dataset. With a 3-cM threshold, 0.2% of haplotypes had a match.