Letters to GEDmatch

Email to GEDmatch July 24^{th}, 2019.

Dear Mr. Rogers and Mr. Olson,

My name is Graham Coop, and I am a population geneticist at UC Davis. I'm writing with Doc Edge, a postdoc who works with me. We have been following news about GEDmatch closely since the initial news about the Golden State Killer case last year. We imagine that it's been a stressful year for both of you, with a lot of weighty decisions. It's clear that you've been giving all of this a great deal of thought.

As we have been thinking about GEDmatch over the past year, two related security issues occurred to us. We will not exploit either of these issues, but we are studying their feasibility. In short, it seems to us that it would be possible for a crawling bot (or for groups such as Parabon, which have uploaded multiple genomes to the site) to reconstruct a large proportion of the genomes of most GEDmatch participants by repeatedly querying the database with publicly available (known) genomes.

To be more explicit, imagine that someone were to download a large set of publicly available genome-wide data, such as those provided by the 1000 Genomes Project (call the publicly available data a "bait set"). One could upload each bait individual to GEDmatch in Ancestry or 23&Me format. Then, one could query GEDmatch for matches between a target of interest (given that their kit number is known) and each bait individual. Because the genotypes of the bait individuals are known, matches between the bait individuals and a target individual reveal information about the target's genotype. Because so many bait individuals are available and because GEDmatch shows matches down to 1 centimorgan in length, it should be possible to reconstruct a substantial percentage of any target individual's genotype—we are calling this approach IBD tiling. We have estimated that at least half of the people of European descent in GEDmatch could have at least a third of their genomes revealed by IBD tiling with publicly available genomes (depending on the parameters of the IBD calling procedure), and we believe that this estimate is conservative (i.e., likely an underestimate of the true proportion).

A related attack is possible as well, focusing instead on identifying people who carry particular alleles, for example risk alleles at disease-associated loci like APOE4 or BRCA. To identify people carrying particular alleles at these loci, one can use haplotypes carrying these alleles in publicly available databases. One could isolate a short segment around the allele of interest, and then construct a genome-wide genotype that is unlikely to be called as IBD with any other user—one way of doing this is to sample alleles from population frequencies, eliminating linkage disequilibrium (LD). Uploading a genotype with a haplotype containing the allele of interest and LD-free genotypes everywhere else will return members of the database who match at the haplotype containing the allele of interest (since matching at other sites has been precluded), thus revealing their genotype at that site. We call this approach IBD probing. (Ideas related to this one have been laid out previously, see for example `https://thednageek.com/cystic-fibrosis-a-case-study-in-genetic-privacy/`)

We will not deanonymize the genotypes of any GEDmatch users in either of these ways, nor have we uploaded any public genomes to GEDmatch. We do not want to compromise anyones genotype nor violate your terms of service. However, we do plan to publish the fact that it is possible to do this, using analyses of data from public datasets as a proof of principle. (That is, we will be masking public genomes and showing that we can use this method to reconstruct their genotype.) One reason for publicizing this issue is that it is also an issue, to varying degrees, for other consumer genomics sites that allow uploading of genotypes. We are writing to you and these other entities today, and we will post our results publicly in 90 days. We are writing to you in advance to give you a chance to communicate with us and respond. Though we do not anticipate this, we also reserve the right to go public sooner in the event that any of the entities to which were writing today publicize results like those we have obtained independently, or otherwise seek to block or publicly undermine our specific results.

It is possible that we will have more recommendations as we advance this work, but at this writing, there are a few actions GEDmatch could take that would reduce the severity of these potential vulnerabilities. In particular, GEDmatch could consider:

1) Raising the minimum threshold for reporting a match to something larger, such as 8cM. (Even at 8cM, some genotypes will be recoverable by IBD tiling, but the proportion of the genome recoverable will be much reduced.) Another approach would be to only allow people to obtain matches where there were two or more large IBD blocks matching between them.

2) Reporting a much smaller number of Many-to-one matches, for example reducing from 3,000 to a few hundred or so. This would prevent people from obtaining so many kit ids and also reduce the yield of IBD probing methods.

3) Allowing only the uploader of a kit to see matches with that kit—i.e. that it is no longer possible to see matches between a pair of matches using only their kit numbers unless the users opt in. The current practice of allowing anyone with a kit number to see the matches means that the entire database may well be obtainable through iterated searches.

4) Ceasing reporting of the genomic locations of matches, instead of reporting just the number of fragments and their sizes.

5) Requiring uploaded data files to contain an encrypted key verifying the source of the genotypes included, as suggested by Erlich et al., 2018, Identity inference of genomic data using long-range familial searches.

6) Including publicly available genomes and being alert to users searching against many of them.

7) Instituting login procedures that block bots and/or limiting queries from one IP. (We are not computer security experts, but we understand that there are ways to get around many of these procedures for motivated actors. We thus do not think that such procedures would be sufficient for solving this problem, but they may stop more casual abuse.)

We note that there are now many ancient genomes (from ancient DNA sequencing) available, thus allowing people to upload genotype files of people they know to be dead, allowing the pursuit of IBD tiling completely within your terms of service.

A separate issue that is specific to GEDmatch is the images you provide of SNP matches between any pair of kit numbers. These pictures seem to be at the SNP scale. This is distinct from the IBD tiling procedure we outline above, but we believe it would be possible to use image processing software to obtain any kit numbers genotype after uploading a dataset that is, for example, homozygous for the human reference allele at every site. Removing these pictures or limiting their resolution substantially (and/or possibly adding some statistical noise to them) would also help to protect the privacy of users genetic information.

Each of these alterations comes with costs to you and to your users, and you are in a better position than we are to consider what might be practical. We are in the process of determining the impact of some of these alterations in limiting security concerns, but we believe that some combination of (1) and (4) is likely to strike a balance of being possible to implement quickly and also quite effective in preventing attacks by IBD tiling. However, some degree of IBD probing will still be possible with (1) and (4) implemented; measures like (5) will be more effective against IBD probing.

Please let us know if you have any questions.

Regards,

Graham Coop and Doc Edge

Email to GEDmatch July 29th, 2019.

(Note in this email we refer to 'IBS flanking' and 'Parallel IBS flanking', which were our earlier names for the IBS baiting and parallel IBS flanking strategies. We also describe 'IBS pathfinding' a technique that we do not describe in the main paper.)

Dear John and Curtis,

Last week, we wrote to you about some potential privacy issues associated with genetic genealogy databases that accept uploads. As we have been working on this project, we have come upon a new category of potential vulnerabilities. These issues are different enough conceptually from the previous set that we wanted to alert you to them separately. As before, we will not attempt to carry out the exploits we describe below in your database at any point, and our anticipated schedule for publication remains the same as the last time we wrote. We want to give you time to address any issues related to these that might arise in your database.

The new exploits we are exploring pertain to methods of identity-by-state / identity-by-descent detection using unphased data that rely on identifying long regions lacking incompatible homozygotes. We do not know the details of your IBS/IBD detection methods, and so we do not know whether these exploits could be possible in your database.

Last time we wrote, we suggested several possible actions that would limit the effectiveness of the exploits we described previously. Most of those actions would also help to prevent the methods we describe now. However, one of the easiest recommendations to implement from last time, that of raising the minimum threshold for reporting a matchs location, would likely do little against these new exploits on its own. In addition, these new techniques lead us to make two more suggestions about methods for preventing security issues:

8) Using phase-aware methods for detecting IBS / IBD. (I.e., not using variants of the opposite-homozygote test.)

9) Testing uploaded datasets for long runs of heterozygosity. Unlike long runs of homozygosity, long runs of heterozygosity are not generated by any population-genetic process that applies to humans. However, artificial datasets with long runs of heterozygosity can be used to trick incompatible homozygous tests for IBS/IBD.

Below, we append some text modified from a rough draft version of our paper that describes the new exploits. The text is not finalized, and the details of the methods described should also not be considered final.

Please let us know if you have any questions.

Regards,

Graham and Doc

Here we consider exploits that become possible if the database identifies putative IBS segments as 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 easily to large datasets. It has been used in the past by major genetic genealogy companies (e.g. Henn et al), and it remains an initial step in some state-of-the-art IBD detection and phasing pipelines (e.g. SpeeDB; Eagle).

Below, we consider some exploits that are possible if a database relies on incompatible homozygous sites to identify IBS regions. A main tool in these exploits is the construction of apparently IBS segments by assigning every uploaded site in the region to be heterozygous. These long runs of heterozygosity, which are unlikely to occur naturally (unlike long runs of homozygosity), will show up as IBS with every genome in the database: because they contain no homozygous sites at all, they cannot contain incompatible homozygous sites with anyone in the database.

Here, we consider a database using the simplest possible version of an incompatible-homozygotes method for detecting IBS, 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, such 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 the techniques reported here could be developed to work within such modifications.

IBS flanking

The first technique, IBS flanking, is the simplest. If IBS is called by looking for single incompatible homozygous sites, then users genotypes at any single genotyped site of interest can be determined by examining their putative IBS with each of two artificial datasets. In each artificial dataset, the site of interest is flanked by a run of heterozygosity. These runs of heterozygosity must be long enough so that together, they are long enough to be reported to the user as an IBS segment. The adversary uploads two dataset with these runs of heterozygosity in place. In the first, the site of interest is homozygous for the major allele, and in the second, 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—it will be interrupted at the site of interest. 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 neither IBS segment is interrupted, then the target user is heterozygous at the site of interest. Thus, the genotype at the site of interest of every user in the database can be revealed after uploading two artificial datasets. This method works for genotyped sites. Sites of medical interest that are often included in SNP chips, such as the APOE locus, are potentially vulnerable to IBS flanking.

Parallel IBS flanking

The second method we consider applies the IBS flanking technique to many sites in parallel. By parallel application of IBS flanking, 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 flanking, 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 for a single pair of uploaded files is as follows:

Identify a set of key sites. 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 threshold reported by the database.

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.

Upload each artificial dataset and compare them to a target user. Key sites that are covered by putative IBS segments between the target and each artificial dataset 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 targets 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 2*c/t-1 key sites can be revealed with each pair of uploaded files. (Consider the case where c=tk, with k a positive integer, and place key sites at t/2, t, 3t/2,...,t-c/2.) This means that with a minimum reported IBS threshold of 1cM, 100 uploaded datasets could reveal approximately 100 genotypes per cM, which is enough for accurate imputation.

IBS pathfinding

Finally, we consider a brute-force approach that relies entirely on serial uploads of artificially generated genetic datasets. Whereas parallel IBS flanking reveals many target users genotypes at several thousand sites, allowing probabilistic imputation at many more, the approach in this section reveals one targets exact genotype at every site in the genome. In the circumstances we consider, it is possible to learn a target users exact genotype at every site on a SNP platform after uploading a few hundred custom datasets.

If IBS is called by looking for incompatible homozygous sites and the rules for reporting the boundaries of IBS segments are understood, then the genotype of the target is revealed at the edge of every reported IBS segment—it is the opposite homozygote of the uploaded genotype. Thus, suppose that an adversary wants to identify the genotypes of a target person in the database, and that any files uploaded by the adversary can be searched for IBS with the target. One procedure is:

Upload a dataset with major-allele homozygotes at every site in the region of interest (which is assumed to be next to an IBS region).

Identify the first site at which IBS between the target genome and the uploaded genome stops. Record the target genomes genotype at this point as a minor-allele homozygote.

Upload a new dataset in which the major-allele homozygous genotype at the previously incompatible site is changed.

Repeat steps 2-3 until every site at which the target genome is a minor-allele homozygote is known.

Repeat a procedure analogous to steps 1-4, but initially uploading a dataset with minor-allele homozygotes in the region of interest and recording the locations of major-allele homozgous sites in the target.

After steps 1-5, every site at which the target is a major-allele homozygote or minor-allele homozygote is known. The target is heterozygous at the remaining sites, and the targets genotype is thus known at every site in the region.

Call this procedure IBS pathfinding, as the targets genotype can be viewed as a path through a graph, which is revealed by testing edges iteratively for obstacles. IBS pathfinding will reveal the full genotype of the target in the region of interest after a number of uploads equal to the number of sites at which the target is homozygous in the region of interest.

If the region of interest can be surrounded by IBS segments on both sides, the number of uploads necessary can be halved, because two opposite-homozygous genotypes that stop the extension of IBS can be identified with each upload—one on each side of the region of interest—until the two flanking IBS segments meet. Further, this procedure can be carried out in parallel for each region of interest, so the total number of uploads required for multiple regions of interest is half the number of homozygous genotypes in the region with the maximum number of homozygous genotypes in the target.

By covering the genome with many regions of interest surrounded by IBS regions, the targets genotype at every site can in principle be revealed after a few hundred uploads. Under the opposite-homozygote method for identifying IBS segments, apparent IBS segments can be generated wherever the uploader desires by constructing a segment where each site is heterozygous. If the database returns IBS segments as short as 1 cM in length and allows queries of each upload against the target genome, then the adversary can begin by uploading a genome with alternating blocks of 1cM of heterozygous genotypes (which will be called as IBS) with 1cM of homozygous genotypes. The genotypes in the initially homozygous blocks are then revealed by IBS pathfinding. Once they are revealed, another set of uploads can begin, this time with the pattern of initially heterozygous and homozygous genotypes reversed. In this way, the targets genotype at every site can be revealed in a number of uploads approximately equal to the maximum number of SNPs in one of the 1cM windows defined initially. (Each half of the genome can be revealed in half this number of steps by working from both sides.) The segments at the ends of chromosomes, if they are of interest, may take more steps to reveal because IBS pathfinding can proceed from only one side. Together, the autosomes are about 3500 cM long, so for a SNP chip with 600,000 autosomal SNPs, a typical 1cM window contains approximately 170 SNPs.

Email to GEDmatch November 20^{th}, 2019.

Dear John and Curtis,

We have been asked by our reviewers to demonstrate that our hypothetical attacks work in principle. Thus we will now attempt a dummy version of the approach on GEDmatch using only a handful of artificial datasets.

To comply by GEDmatchs terms and conditions we will use artificial datasets, designed to not match anyone in the dataset. These kits will be uploaded in Research mode, where they are not visible to other users, and so we will not be interacting with other users. We will not interact with any other users data, just run the databases tools on our fake datasets. Thus we will not gain access to any information about any member of the public, nor will we violate any of the databases terms and conditions. We have also consulted with our University IRB to ensure that this does not constitute human subjects research.

We are still happy to talk about any of the proposed strategies to ensure that the effectiveness of these approaches is minimized in the future.

Best wishes,

Graham and Doc

Email to GEDmatch December 4th, 2019.

Our email describing our first baiting strategy. This was replaced by the smaller baiting regions of 0.6cM containing 22 SNPs either side.

Dear John and Curtis,

We have now demonstrated our IBS baiting technique on GEDmatchs research sandbox. We have kept to the principles laid out in our previous email, thus we have not violated the terms and conditions of GEDMatch nor interacted with data from other users. We have demonstrated that by using 2 bait kits we can use the 1-to-1 match tool to identify the genotype of each of four target SNPs on chromosome 22.

We found that kits with long heterozygous runs are now stopped from tokenization with a message 'HTZ string too long' . Can you confirm whether this block was put in place in response to the reported issues from Ney and ourselves?

By uploading bait kits with multiples of 10 SNPs blocks of heterozygotes we determined that 45 heterozygous SNPs in a row were allowed to proceed. We found that the text implemented by GEDMatch did not test for missing data, and so we created a 2cM window of heterozygosity and missing sites around each of 4 target sites in our bait genomes.

We uploaded three artificial target genomes (T1-T3), constructed from sampling genotypes at random at each SNP. These vary in their genotype at the target SNPs. T1 and T3 are opposite homozygous, T2 is heterozygous.

We then upload two bait genomes (B1 & B2), both of which have 4 runs of heterozygosity and missing sites surrounding each bait site, with the two baits having opposite homozygote genotypes at each target SNP.

We use the one-to-one tool to determine where a short run of IBS between a target genome and bait genome is present surrounding the target region. That allows us to determine whether the bait and target genomes have opposite homozygotes at the target SNPs.

The attached pic shows it in action, note the calling of IBD blocks in each of the four regions, with the right panel being a zoomed-in version of the first target region. We are happy to share any other details that might be useful to you.

As promised the kit numbers are:

Bait genome (B1 and B2) kit numbers:*R*EMOVED

Target genomes (T1-T3): *R*EMOVED

Our simulated attack on privacy on GEDMatch was facilitated by the fact that small IBS blocks can be identified such that baits can be made using small combinations of runs heterozygosity and missing data. Only allowing users to call blocks of IBS of say >5cM using 1-to-1 matching, and blocking kits with long runs of heterozygosity and missing data would have blocked this attack.

As an aside, the high-resolution visualization of the SNP matches along the chromosomes is still also an issue. Even if there were no IBS block calling in the regions, it is clear from the zoomed-in view that we can see the target mismatches in question. It looks like positions in the visualization have been jittered, perhaps as a measure against a Ney et al style approach. However, it would still be easy to computationally extract our target SNP genotypes from these images as the mismatches are clear.

We are happy to discuss how these types of attack could be stopped in the future. Best wishes, Graham and Doc