A New Inference Attack against Kin Genomic Privacy

Fatma Balci[1], Handan Kulan[2], Can Alkan[3], and Erman Ayday[4]
Computer Engineering Department, Bilkent University, 06800, Ankara, Turkey
{[1]fatma.balci, [2]handan.kulan}@bilkent.edu.tr, {[3]calkan, [4]erman}@cs.bilkent.edu.tr

With the constantly decreasing cost of whole genome sequencing, today, it is possible to obtain a list of the mutations in your DNA, learn your predispositions to several human diseases, or create your family tree based on your DNA in a cost effective way [1]. However, these rapid developments in the field of genomics, and especially direct-to-consumer-genomics, also brings some concerns about privacy. As one's genome includes sensitive information not only about him but also about his family members, it is crucial to make sure it is stored and processed in a privacy-preserving way [2].

A recent debate between the family members of deceased Henrietta Lacks and medical researchers (who published her genome on a public Website without the consent of the family members) is a clear example of potential future conflicts on kin-genomic privacy [3]. In CCS 2013, focusing on the single nucleotide polymorphisms (or shortly SNPs, the most common type of genetic variation among humans [5]), Humbert et al. propose an efficient algorithm [4] to infer the genome of a family member (victim) from genomes of other family members, public genomic knowledge, Mendel's Law, and correlations between the nucleotides. On one hand, they show that full reconstruction of the victim's genome is possible if the attacker has enough data about the genomes of other family members.

On the other hand, there is the following limitation in the algorithm proposed in [4]. Assume we are working on a dataset consisting of a trio (a father, a mother, and a child), and we are trying to infer a particular SNP of the father, given the SNPs of the mother and the child. Following the Mendel's Law, if the child is homozygous (carrying two identical nucleotides) in that SNP position, we can easily infer the nucleotide in one strand of the father. However, if both the child and the mother are heterozygous (carrying two different nucleotides) in that SNP position, based on [4], we cannot get any information about the nucleotide that is inherited from the father to the child.

This limitation of [4] can be ameliorated by using the haplotype information. A haplotype is a group of nucleotides on a single chromosome that are closely linked to be inherited usually as a unit. Haplotypes are identical by descent (IBD) if they are identical and inherited from a common ancestor. So, if we can detect the IBD between the father, mother, and the child, we can circumvent the limitation in the aforementioned example. There are several existing methods for detecting IBD. We use Beagle [6] for this detection. Beagle allows SNPs to be in LD by modeling haplotype frequencies.

By using this haplotype information, in this work, we introduce a new inference attack to find one of the parent's SNPs by using the genomes of the other parent and the children. We use the regions that are inherited together and we use the idea that if the child's SNPs included in a haplotype block is not coming from the mother's genome, then it is coming from the father's genome. Then, we deduce that the other haplotype of the child is inherited from the father. We evaluate our approach on CEPH/Utah Pedigree 1463 dataset [7], a Caucasian family that is comprised of 4 grandparents, 2 parents, and 11 children containing partial DNA sequences of all family members. We show that accurate inference (about the father's SNPs) can be done by using less data (i.e., genomic data of fewer family members) compared to [4]. Note that haplotype information can be also integrated to [4] for better and more accurate inference with less data.

References

[1] https://www.23andme.com/.

[2] E. Ayday, E. D. Cristofaro, G. Tsudik, and J.-P. Hubaux. Whole genome sequencing: Revolutionary medicine or privacy nightmare. IEEE Computer Magazine, 48(2), Feb. 2015.

[3] http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html?pagewanted=all.

[4] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks family: Quantification of kin genomic privacy. In CCS, 2013.

[5] http://ghr.nlm.nih.gov/handbook/genomicresearch/snp.

[6] B.L. Browning and S.R. Browning. A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics, 88(2), 173-182, 2011.

[7] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, et al. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. Science, 327(5961):78–81, 2010.