A Privacy-preserving Reference Panel for Genotype Imputation

Mohammad Nabil Ahmed[1,a)] Kana Shimizu[1]

Abstract: Genomic data can be used to infer private and sensitive information about individuals, which prevents it from being shared publicly. Despite the use of data de-anonymization techniques, the release of statistical measures from a genomic database can make it vulnerable to privacy-centric attacks. Genotype imputation, a technique developed from statistical genetics has recently found increasing usage in Genome-wide Association Studies (GWAS), where it is used to increase the coverage of genotype information. The privacy-centric nature of genetic information has led to the adoption of database governance that stonewalls it from public-access, limiting the access to information-rich imputation reference datasets. In this research we propose mechanisms through which privately-held imputation reference panels can be released without invalidating data privacy.

Keywords: Privacy-preserving, Genotype Imputation, Differential Privacy, Data Privacy

1. Introduction

Whole-genome sequencing (WGS) of human genetic reference population [7], [10] have all yielded quality reference datasets that can be used for inferring untyped genotypes in study populations. This technique known as genotype imputation has proven useful in revealing genotype information in an extremely cost-effective manner, leading to its widespread adoption in Genome-wide Association Studies (GWAS). A larger and more importantly, diverse reference panel is highly likely to contain rare variants and thus impute genotypes with higher accuracy. In this research, we consider a scenario where a researcher is unable to to access a high-quality imputation reference panel, stored in *some* private genome data bank. We devise mechanisms through which a data owner may release a privacy-preserving version of this dataset without destroying its utility.

2. Related Works

The growing importance of Genome wide Association Studies (GWAS) has led to the development of commercial SNP arrays with known variants. The coverage of genotyping arrays is increased through a technique known as genotype imputation, of which several methodologies have been proposed in literature [1], [2], [3].

As for privacy preserving mechanisms, differential privacy has emerged as a robust privacy model with mathematically proven guarantee of individual privacy [4]. Kasiviswanathan et al. [5] proposed an exponential based mechanism to derive a synthetic dataset that can answer any query set. Meanwhile, Blaum et al. [6] proposed a similar approach over a discrete domain. Subsequent research works by Dwork et al. [4] suggested the use of boosting to improve the accuracy of synthetic dataset.

To the best of our knowledge, this research work appears to be the only one that proposes a data release mechanism to construct privacy-preserving imputation reference panels.

3. Methodology

The fundamental idea behind the creation of a privacy preserving haplotype reference panel is based on Li and Stephens model of linkage disequilibrium [9]. This basic principle of representing haplotypes as an imperfect mosaic of ancestral haplotypes forms the basis through which we propose an algorithm for deriving privacy preserving panels.

3.1 Shuffling Based Algorithm

Given a reference panel D of $2N$ distinct haplotypes with k distinct markers or SNPs, we propose an algorithm M_s that creates a new panel P_s with $2N*$ distinct haplotypes. We first represent the $2N$ haplotypes from D using a matrix-like data structure of size $k \times 2N$. We then divide this dataset into k/l segments for a series of consecutive markers of fixed length, l and iterate through the dataset. We then randomly select x haplotypes, through the use of a user-defined parameter c , from each k/l^{th} subset of D. The selected haplotypes are then appended to the previous selection at the end of each iteration.

3.2 Differentially-private Based Algorithm

Differentially private algorithm, M_d is designed with the same fundamentals as M_s while introducing a new privacy parameter ε. However, instead of randomly selecting haplotypes, we iterate through each k/l^{th} segment and retrieve a set of unique haplotypes z for and calculate its probability distribution. We then define a utility function $u(D_l, z_i)$ based on exponential mechanism of dif-

[1] Department of Computer Science and Communication Engineering, Waseda University, Japan
[a)] mnabil.ahmed@asagi.waseda.jp

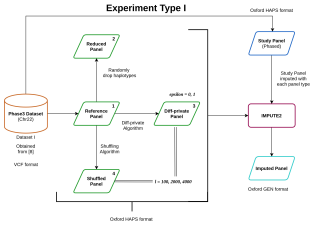

Fig. 1 Experimental Type I: Dataset sourced from [8]

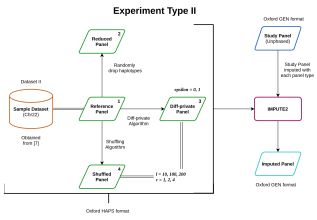

Fig. 2 Experimental Type II: Dataset sourced from [7]

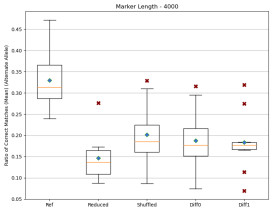

Fig. 3 Mean Overall Accuracy (Alt. Alleles) $l = 4000$

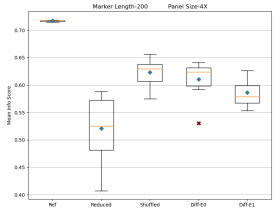

Fig. 4 Mean Info Score $l = 200$, $c = 4$

ferential privacy [4] that selects haplotypes from each segment. The subsequent steps are the same as the shuffling algorithm, M_s.

4. Experiments

We carried out two experiments using datasets from two distinct sources as shown in **Fig. 1** and **Fig. 2**. The dataset used in experiment I was obtained from the phase3 data release of 1000 Genomes Project [8]. While for experiment II, we used a sample test dataset that is packaged along with IMPUTE2 imputation software [7]. In addition to the using datasets from varying sources, the data processing pipeline that we adopted for both these experiments was also slightly different.

For both these experiments, we then derived three kinds of reference panels from the reference panel, D:

- `Reduced panel` D_r
- `Shuffled panel` P_s
- `Differentially-private panel` P_d

We obtained the private P_s and P_d based on varying parameter combinations of l, c and ε as shown in Fig. 1 and Fig. 2. Finally for each experiment, we performed genotype imputation on the study panel, S and repeated this process for a series of 10 randomized trials.

5. Results

In experiment I, we report the result using the accuracy metric. Since the true genotypes for each of the imputed markers is already known, we compute the accuracy through a simple binary comparison. Since IMPUTE2 reports imputed genotypes as a probability triple $(x, y, z) \mid 0 \geq x, y, z \geq 1$, while making the comparison we simply call the $max(x, y, z)$ regardless of its value. Meanwhile for experiment II, the results are reported using info metric [11], obtained from the info file generated by IMPUTE2.

Figure 3 and **Fig. 4** shows the distribution of mean accuracy and info scores across the 10 imputation runs for all the reference panels. In general, we noticed that increasing the value of parameter l, leads to a better distribution of mean accuracy in P_s and P_d. Importantly, this trend is more pronounced in sites where alternate allele is expressed as shown in Fig. 3. Since parameter l is correlated to the proportion of contiguous markers between the reference and private panels, it is intuitive to expect better accuracy at higher values of l.

In Experiment II, we notice a similar trend, that is the mean info score gradually improves as we increase the values of parameter c and l as shown in Fig. 4. By increasing the size of the private panels, P_s and P_d, we are effectively increasing the diversity of information contained withing these panels. Once again, we notice that both private panels P_s and P_d outperform the reduced panel, D_r on this metric. Since info score is a measure of confidence, it is reasonable to expect higher measures of info score at higher values of c and l.

6. Conclusion

We have thus, proposed two algorithms to derive privacy preserving reference panels from haplotype release data. We have also demonstrated that for certain parameter combinations, privacy preserving panels impute genotypes with a higher quality than panels with a reduced set of haplotypes. This is indicative of the fact that privacy-preserving panels can serve as an alternative in cases where there is a barrier to accessing high quality haplotype reference panels. Future work in line with this research should investigate the possibility of designing privacy-preserving mechanisms which takes genetic and evolutionary processes into account as it may lead to increasing the accuracy of imputation.

References

[1] Clark, A.: Inference of haplotypes from PCR amplified samples of diploid populations, *Molecular Biology and Evolution*, Vol.7, No.2, pp.111–122, (1990).
[2] Excoffier, L. and Slatkin, M.: Maximum likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution*, Vol.12, No.5, pp.921–927, (1995).
[3] Durbin, R.: Efficient haplotype matching and storage using the positional burrows-wheeler transform (pbwt) *Bioinformatics*, Vol.30, No.9, pp.1266–1272, (2014).
[4] Roth, A. and Dwork, C.: The Algorithmic Foundations of Differential Privacy, DOI: 10.1561/0400000042 (2014).
[5] Kasiviswanathan, S. P., Rudelson, M., Smith, A., and Ullman, J.: The price of privately releasing contingency tables and the spectra of random matrices with correlated rows, *Proc. STOC '10*, Cambridge, Massachusetts, USA (ACM), pp.775–784 (2010).
[6] Blum, A., Ligett, K. and Roth, A.: A learning theory approach to non-interactive database privacy, *Proc. STOC '08*, Victoria, British Columbia, Canada (ACM), pp.609–618 (2008).
[7] Howie, B. and Marchini, J.: Examples (Online), available from ⟨http://mathgen.stats.ox.ac.uk/impute/impute_v2.html #examples⟩, (accessed 2021-06-01)
[8] The 1000 Genome Project Consortium, Auton, A. et al.: A global reference for human genetic variation, *Nature*, Vol.526, No.7571, pp.68–74, (2015).
[9] Li, N. and Stephens, M.: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics*, Vol.165, No.4, pp.2213–33, (2003).
[10] The International HapMap Consortium.: The international hapmap project, *Nature*, Vol.426, pp.789–796 (2003)
[11] Marchini, J.: Info measures (Online), available from ⟨https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#info_measures⟩, (accessed 2021-06-01)
[12] Howie, B. N., Donnelly, P. and Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLOS Genetics*, Vol.5, No.6, pp. 1–15, DOI: 10.1371/journal.pgen.1000529 (2009).

Appendix

A.1 Privacy-preserving Algorithms

Algorithm 1: Shuffling Algorithm, M_s

Input: Reference panel D
// HAPS data format with k markers and $2N$ haplotypes
Input: c, l
/* User-defined parameters of type *int, where c is the number of haplotypes in P_s as a multiple of D and l is the number of consecutive markers from D that are to be preserved during segmentation* */
Output: Private panel, P_s
// HAPS data format with k markers and $2N^*$ haplotypes

$D \leftarrow$ **ConvertToMatrix**(D) // Converts D into a $D^{k\times 2N}$ matrix
$D_l \leftarrow$ **Segment**(D) // Segments D into k/l chunks
$P \leftarrow$ *empty*
$P_s \leftarrow$ *empty*

for $i \leftarrow 1$ **to** c **do**
  while *not all segments traversed in* D_l **do**
    $Select \leftarrow$ **SelectHaplotypes**(D_l)
    // Chooses $2N$ haplotypes from segment k/l of D without replacement
    if *first segment* (l *equals* 1) **then**
      $P \leftarrow Select$
    else
      $P \leftarrow$ **Append**($Select$) // Selected segments are appended column-wise to P
    end
    go to *segment* $k/l+1$
  end
  if i *equals* 1 **then**
    $P_s \leftarrow P$
  else
    $P_s \leftarrow$ **Stack**(P) *along axis* 0) // P_s is stacked with P along axis 0
  end
  increment i
end
$P_s \leftarrow$ **ConvertToHaps**(P_s) // Convert P_s into HAPS format
return P_s

Algorithm 2: Differentially Private Algorithm, M_d

Input: Reference panel D
// HAPS data format with k markers and $2N$ haplotypes
Input: c, l, ε
/* User-defined parameters of type **int**, where c is the number of haplotypes in P_d as a multiple of D and l is the number of consecutive markers from D that are to be preserved during segmentation. ε is the privacy parameter that controls the degree of randomization */
Output: Private panel, P_d
// HAPS format with k markers and $2N^*$ haplotypes

$D \leftarrow$ **ConvertToMatrix**(D) // Converts D into a $D^{k\times 2N}$ matrix
$D_l \leftarrow$ **Segment**(D) // Segments D into k/l chunks
$P_d \leftarrow$ *empty*

while *not all segments traversed in* D_l **do**
  $Unique \leftarrow$ **Unique**(D_l)
  // Finds a set of unique haplotypes z in D_l
  $Score \leftarrow$ **Score**($Unique$)
  // Assigns a score to each z_i based on utility function $u(D_l, z_i)$
  $Prob \leftarrow$ **Exponential**($Unique, Score, \varepsilon$)
  // Computes the probability of each element in $Unique$ using the exponential mechanism of differential privacy
  $Select \leftarrow$ **SelectHaplotypes**($Unique, Prob, c$)
  // Chooses $c \times 2N$ haplotypes from $Unique$ based on its computed probability

  if *first segment* (l *equals* 1) **then**
    $P_d \leftarrow Select$
  else
    $P_d \leftarrow$ **Append**($Select$)
    // Selected segments are appended column-wise to P_d
  end
  go to *segment* $k/l+1$
end
$P_d \leftarrow$ **ConvertToHaps**(P_d) // Convert P_d into HAPS format
return P_d