IDENTIFYING COREFERENT GENOTYPES IN ONE-WAY CRYPTOGRAPHIC HASH USING HAPLOTYPE INFORMATION

[1]J. C-N. Chang and [2]M. S. Cline
[1]Department of Computer Science, University of California, Sana Cruz, CA
[2]UC Santa Cruz Genomics Institute, Sana Cruz, CA
Contact: [1]*cchan60@ucsc.edu*, [2]*mcline@ucsc.edu*

Introduction: Genome-wide association studies (GWASs) are powered by analyzing and learning from large amounts of data. Due to the sensitive nature of individual-level genotype data, such large datasets are sometimes hard to come by. A common strategy is to merge smaller datasets by identifying and removing duplicate individuals using one-way cryptographic hashing of the genotype [1]. In practice, however, we are often unable to obtain the entire genotype array due to privacy concerns and divergent data collecting purposes. Therefore, we want to find a small subset of SNPs to represent each individual that provides sufficiently low collision rate. In this study, we approach the challenge in selecting SNPs to identify duplicate individuals and explore methods using entropy and Linkage Disequilibrium (LD) [2] information.

Materials and Methods: In this study, we use the 1000 Genome Project dataset [3], with 2504 individuals, each with 5756 SNPs. We filter out all intronic SNPs and genes other than BRCA1 and BRCA2, as our target data sources focus on analyses of the BRCA genes, leaving 525 SNPs. We consider two methods to eliminate SNPs from this set, namely Shannon Entropy and LD information. A low entropy indicates similar allele distribution for a large proportion of patients and hence low distinguishability, so we want to remove SNPs with low entropy. We also use r^2 values in LD information from SNP Annotation and Proxy Search (SNAP) [4] to identify haplotypes, SNPs in which are inherited together and tend to provide similar information. Hence, for SNPs in the same haplotype, we want to only keep a few in our selection. We combine these two methods in two ways: Composite, where we simply apply the two methods sequentially, and Knapsack, where we solve a knapsack problem [5], an optimization problem, with value and weight functions based on entropy and LD.

Results and Discussion: Both methods achieve high identifiability with a relatively small set of SNPs, where identifiability is computed as the fraction of uniquely identified patients. For Composite, we remove SNPs with high r^2 and select the top k SNPs ranked by entropy, varying k. For Knapsack, we empirically determine the maximum weight and vary the selection size k. As a baseline, we simply select SNPs according to their entropy ranking. The results are below. Composite performs best overall. While Knapsack is reasonably good, the baseline surpasses when k is large. We also consider various levels of birth information (not presented). As expected, the granularity in birth information has a positive effect on identifiability.

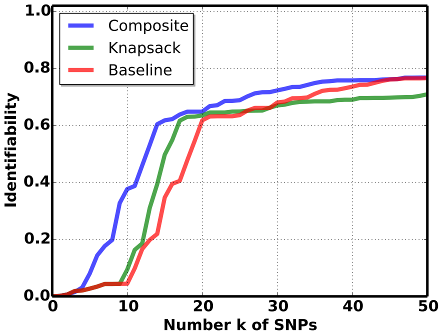

Figure 1: Identifiability versus number of SNPs used. Composite corresponds to results where SNPs with high r^2 are removed, and the top k SNPs, ranked by entropy, are selected; Knapsack corresponds to results where we solve a knapsack problem with a fixed maximum weight; and Baseline corresponds to selecting SNPs based on the entropy ranking only.

Conclusions: We can identify 60% of the individuals with 14 SNPs using the Composite solution. This can be boosted to 93% with 9 SNPs if we also consider the birth year of the individuals.

References:

[1] Turchin, Michael C., and Joel N. Hirschhorn. "Gencrypt: one-way cryptographic hashes to detect overlapping

individuals across samples." *Bioinformatics* 28.6 (2012): 886-888.
[2] Slatkin, Montgomery. "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future." *Nature Reviews Genetics* 9.6 (2008): 477.
[3]1000 Genomes Project Consortium. "A global reference for human genetic variation." *Nature* 526.7571 (2015): 68.
[4] Johnson, Andrew D., et al. "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap." *Bioinformatics* 24.24 (2008): 2938-2939.
[5] Andonov, Rumen, Vincent Poirriez, and Sanjay Rajopadhye. "Unbounded knapsack problem: Dynamic programming revisited." *European Journal of Operational Research* 123.2 (2000): 394-407.