BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 6 2012, pages 886–888
doi:10.1093/bioinformatics/bts045

Genetics and population analysis

Advance Access publication February 1, 2012

Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples

Michael C. Turchin[1,2] and Joel N. Hirschhorn[1,2,3,*]

[1]Divisions of Endocrinology and Genetics, Children's Hospital Boston, Boston, MA 02115, [2]Broad Institute, Cambridge, MA 02140 and [3]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Meta-analysis across genome-wide association studies is a common approach for discovering genetic associations. However, in some meta-analysis efforts, individual-level data cannot be broadly shared by study investigators due to privacy and Institutional Review Board concerns. In such cases, researchers cannot confirm that each study represents a unique group of people, leading to potentially inflated test statistics and false positives. To resolve this problem, we created a software tool, Gencrypt, which utilizes a security protocol known as one-way cryptographic hashes to allow overlapping participants to be identified without sharing individual-level data.

Availability: Gencrypt is freely available under the GNU general public license v3 at http://www.broadinstitute.org/software/gencrypt/

Contact: joelh@broadinstitute.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 27, 2011; revised on January 7, 2012; accepted on January 20, 2012

1 INTRODUCTION

Genome-wide association studies (GWASs) have identified over 1000 genomic loci associated with numerous diseases and traits (Hindorff *et al*., 2009). These associations are often discovered by meta-analysis of summary statistics from multiple, independent cohorts. One key concern of this type of approach is that the same individuals may be present in multiple cohorts, especially if the cohorts recruit from overlapping regions. The presence of duplicate individuals in multiple studies may affect meta-analysis results, producing inflated test statistics and spurious associations (Lin and Sullivan, 2009; Newman *et al.*, 2001). However, sharing individual-level data across research groups is often not permissible due to Institutional Review Board (IRB) restrictions, thus inhibiting researchers from directly assessing whether individuals are present in multiple datasets ('overlapping individuals'). To address this issue, we have created a software suite, Gencrypt, which takes advantage of a security protocol known as one-way cryptographic hashes.

One-way cryptographic hashes are a form of security algorithms that alter input in such a way that the resulting output bears no resemblance to the original content (referred to as 'digesting' the input), and attempt to produce a unique output for each unique input (referred to as a 'collision' when this fails (Burr, 2006)). Cryptographic approaches have already successfully been applied to other fields in biology, such as DNA forensics (Bohannon *et al.*, 2000). The advantages of using such an algorithm in the context of GWASs is that the original genotype specificity is kept, but the privacy that is necessary to maintain when using IRB and consent-form protected data are not compromised. As a result, we can use one-way cryptographic hashes on genetic information, and compare the resulting outputs against one another to securely ascertain identical individuals between datasets used in GWASs.

2 IMPLEMENTATION

Gencrypt uses Perl v2.8.9+ and operates on PLINK (Purcell *et al.*, 2007) formatted files (.ped and .bim). It separates each individual's genetic information into groups of SNPs, and hashes each group one at a time. SNPs are hashed in a random order generated in-code from a user-provided seed value, and the number of SNPs included per hash is also defined by the user. The default, and recommended, one-way cryptographic hash algorithm used is SHA-256 (Eastlake and Hansen, 2006), a NSA developed algorithm that remains unbroken and has an exceedingly low likelihood of producing a collision. Gencrypt also supports WHIRLPOOL and MD5 (Barreto and Rijmen, 2000; Rivest, 1992), alternative one-way cryptographic hash algorithms available at the user's discretion. The output file produced contains rows of hashes representing the same sets of grouped SNPs for every individual. Gencrypt then takes two of these output files and compares their rows of hashes against one another, printing out pairs of individuals who have high percentages of identical hashes. It is then suggested that these pairs may be overlapping individuals. However, it is necessary to confirm that the output files being compared were originally created using the same set of SNPs, the same SNP order, and the same number of SNPs per hash. To accomplish this, Gencrypt constructs a simulated positive control individual who is homozygous for the reference allele at all SNPs used and includes this individual at the beginning of every hash output file. When comparing these hash output files, positive controls created from the same set of SNP parameters should produce 100% identical hashes, thus confirming the same SNPs, SNP order and SNP hash sizes were used.

The number of SNPs used per hash is set by the user, but we recommend a lower bound of 150 to ensure data security. One of the main ways to bypass the security SHA-256 offers is by hashing all possible genotypes that could exist for a given set of

*To whom correspondence should be addressed.

SNPs until a hash output of interest is recreated, thus revealing the original genotypes. The efficiency of this brute force approach scales inversely with the number of SNPs included per hash—the more SNPs there are, the more computationally intractable it becomes to test all possible combinations. To find where this threshold of SNP group size may be, groups of different numbers of SNPs were hashed via SHA-256 1×10^6 times on a virtual machine running CentOS 5.5 ×86_64 with 3 AMD Opteron Processors @ 2.3 GHz and 8 GB ram, using Perl 5.8 with the Perl module Digest::SHA. Doing so took 12, 13 and 14 seconds for SNP group sizes of 50, 100 and 150, respectively. By extension, given the number of possible genotype combinations that SNP groupings of these sizes contain (3^{50}, 3^{100} and 3^{150}), it would take on average this same machine 1.48×10^{11}, 1.05×10^{35} and 7.63×10^{58} years to hash all these SNP genotype combinations. While there may be ways for a malicious script to intelligently cut down the number of genotype combinations that need to be tested, we believe these times are large enough to suggest hashes produced from 150 SNP groupings are sufficiently secure. Therefore, the recommended number of SNPs users should include per hash group is 150.

A potential problem that arises from using these many SNPs per group is the high likelihood of including at least one missing genotype per hash. Missing data are problematic because, given a hash that would otherwise be identical between two overlapping individuals, missing genotypes in either individual will lead to different hash outputs, thus producing false negative comparisons. In an effort to deal with this problem, Gencrypt uses the known alleles at the missing SNP site and recreates the current group of SNP genotypes with all three possible genotypes the missing SNP could have contained (Bohannon *et al.*, 2000). This produces a total of 3^M hashes for a single SNP group, where M is the number of missing genotypes originally in the group of SNPs being hashed. The maximum number of accepted missing genotypes per hash is set by the user, up to a limit of four. When the number of missing genotypes exceeds the user-defined threshold, the current group of SNPs being hashed is exited and a '0' is put in the place of a hash output. It is thus treated as a missing hash, and provides no data in the downstream comparison procedure. The recommended number of missing genotypes users should accept is two. This should produce at most nine different possible genotype groupings given a set of SNPs, which at the recommended SNP group size of 150 SNPs does little to infringe on the specificity required by the program to work correctly, but does allow realistic levels of missingness to be handled.

Given the set of SNPs that overlap the datasets being compared, there are multiple recommendations for choosing which SNPs to use in Gencrypt. SNPs should be chosen such that the minor allele frequency is between 40% and 50%. Using SNPs with the most variability provides Gencrypt with greater power. Gencrypt attempts to take into account strandedness based on the input .bim file, so SNPs whose alleles are either A and T, or G and C, should be avoided due to ambiguous strandedness. Additionally, SNPs should be used that minimize the amount of missing information whenever possible. The recommended total number of SNPs Gencrypt should be run on is 20 000. This number is kept low in order to facilitate users identify SNPs present in the studies being compared, since these studies may be based on different platforms. Given 20 000 SNPs, computational runtime is currently $\sim n2/600$ seconds, where n is the number of samples being analyzed, suggesting that comparing two samples with sizes over 100 000 may be computationally intensive. There is no limit to the number of SNPs being used, although run time increases linearly with the number of SNPs, and there is no reason to go above 20 000. As is shown below, using 20 000 SNPs gives Gencrypt enough data points to successfully identify overlapping individuals between studies. Including more SNPs if available can increase Gencrypt's accuracy, but users should not use fewer than 20 000 SNPs. It should be noted though, as is also shown below, Gencrypt successfully identifies overlapping individuals even without any specific choices regarding SNP selection.

3 RESULTS

We tested Gencrypt on two datasets derived from real genotype data. The first test dataset was based on African American individuals collected from Maywood, IL and genotyped on the Affymetrix 6.0 SNP array (Kang *et al.*, 2010). A total of 743 individuals were broken into two subsets of 421 and 422, with 100 individuals overlapping between the two groups. Twenty thousand SNPs were randomly chosen from a full set of 859 332 SNPs available as a test of the program's robustness to methods of SNP selection, and to show Gencrypt's utility with a small number of SNPs used. A variety of SNP groupings per hash were tested, ranging from 10 SNPs per hash to 150 SNPs per hash. To test the performance of Gencrypt with a fixed rate of missingness, missing genotypes were given a randomly chosen genotype based on the missing SNP's alleles, individuals were included in both 'halves' of the dataset, and then 1% of genotypes were removed randomly across all individuals. For every SNP grouping tested, all 100 duplicate pairs of individuals were identified. Additionally, when fewer than 40 SNPs per group were used per hash, other familial relations were picked up, such as first-degree relatives. However, using $\leqslant$40 SNPs per hash is not recommended due to the potential insecurity hashes based on SNP groupings of these sizes have. When $\geqslant$50 SNPs per group were used, none of the 177 562 non-duplicate pairs of individuals produced a spuriously overlapping hash. For groups of 50, 100 and 150 SNPs, the percentages of identical hashes for duplicate individuals ranged from ∼50% to ∼100% (Supplementary Fig. S1). While including more SNPs per hash does decrease the total number of identical hashes found between two overlapping individuals, using 150 SNPs per hash still produces >50% of total hash comparisons per duplicate pair as being identical.

To test Gencrypt's performance with the addition of genotyping error, a random 0.1, 0.2, 0.5, 1 and 2% of SNPs in both halves of the Maywood data had a single allele altered ('miscalled') in order to simulate a range of genotyping error rates. These datasets were then run through Gencrypt using a hash group size of 150 SNPs. For genotyping error rates of 0.1, 0.2 and 0.5%, all duplicate individuals were still identified (with decreasing amounts of overlapping hashes between identical individuals). A total of 99 out of 100 individuals were identified with 1% genotyping error, and 15 out of 100 individuals were identified with 2% genotyping error (Supplemental Fig. S2). While Gencrypt is robust to low, and realistic, levels of genotyping error, SNPs should still be selected to minimize the effects of genotyping error.

The second dataset used was the publicly available US GoKinD study (Mueller *et al.*, 2006). A group of 1825 individuals from the USA and Canada with long-term type 1 diabetes, these were genotyped on the Affymetrix 5.0 SNP array and five duplicate pairs

Table 1. US GoKinD IBD estimates and identical hashes percentages for known duplicates

Pair of duplicate individuals	IBD estimates	Identical hashes (%)
1	0.9997	97.7
2	0.9997	95.5
3	0.9995	94.0
4	0.9925	27.8
5	0.9748	6.02

Shown are IBD estimates and identical hash percentages for the five known duplicate pairs from the US GoKinD data identified by Gencrypt. Percent identical hashes represent the total number of hash outputs that are shared between the individuals in each pair. IBD estimates were calculated using PLINK (Purcell *et al.*, 2007). Lower IBD estimates and percent identical hashes are the results of higher levels of missingness within those individual pairs.

of individuals, genotyped separately, had been identified in the QC process. In this case, missing data were maintained as is to test the performance of Gencrypt in a real dataset. Twenty thousand SNPs were randomly extracted from the full dataset of 233 100 SNPs, and size groupings of 150 were used. The resulting output was compared against itself, with results between people of the same individual IDs dropped and not considered duplicates. For three of the five known duplicate pairs, >93% of their hash outputs were identical (Table 1). The other two pairs of duplicate individuals were still detected, with 28.1 and 6% of hash outputs being identical. In each of these two pairs, one of the two input genotypes had missingness rates >1% (1.3 and 1.23%). Despite the relatively low proportion of identical hash outputs, these five pairs were the only pairs identified with any duplicate hashes among the 1.7×10^6 pairs compared, demonstrating the high specificity of Gencrypt.

4 SUMMARY

In summary, our program, Gencrypt, successfully secures and compares individual-level data in order to identify overlapping individuals in different genotype datasets. The program maintains the security necessary to handle individual-level data, while also retaining the specificity and sensitivity needed to identify identical individuals. Additionally, while individual IRB opinions may vary, one IRB consulted on this matter agreed this approach would likely be a feasible way for identifying overlapping samples for the purposes of subsequent meta-analyses, in situations where unencrypted individual level data would not be immediately sharable. Therefore, this program allows researchers to share individual-level data without infringing on IRB guidelines and to remove duplicate individuals from their respective studies.

ACKNOWLEDGEMENTS

The authors would like to thank Cameron D. Palmer, Charleston W.K. Chiang, Rany M. Salem, Yingleong Chan, Thutrang T. Nguyen and the rest of the Hirschhorn Lab for their helpful comments and feedback. US GoKinD data was downloaded from the NIH database of Genotypes and Phenotypes resource (dbGaP) http://www.ncbi.nlm.nih.gov/gap. We thank Xiaofeng Zhu and Richard Cooper for sharing individual level genome-wide data from the Maywood dataset.

Funding: March of Dimes (6-FY09-507 to J.N.H.); NIDDK (1R01DK075787 to J.N.H.).

Conflict of Interest: none declared.

REFERENCES

Barreto,P. and Rijmen,V. (2000) The Whirlpool Hashing Function. *First open NESSIE Workshop*. Lueven, Belgium.

Bohannon,P. *et al*. (2000) Cryptographic approaches to privacy in forensic DNA databases. In *Public Key Cryptography '00*. LNCS 1751, pp. 1373–1390.

Burr,W. (2006) Cryptographic hash standards: where do we go from here?. *IEEE Security & Privacy*, **4**, 88–91.

Eastlake,D. and Hansen,T. (2006) US Secure Hash Algorithm (SHA and HMAC-SHA). RFC 4634.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Kang,S.J. *et al.* (2010) Genome-wide association of anthropometric traits in African- and African-derived populations. *Hum. Mol. Genet.*, **19**, 2725–2738.

Lin,D.Y. and Sullivan,P.F. (2009) Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.*, **85**, 862–872.

Mueller,P.W. *et al.* (2006) Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *J. Am. Soc. Nephrol.*, **17**, 1782–1790.

Newman,D.L. *et al.* (2001) The importance of genealogy in determining genetic associations with complex traits. *Am. J. Hum. Genet.*, **69**, 1146–1148.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rivest,R. (1992) The MD5 Message Digest Algorithm. MIT and RSA Data Security, Inc., RFC 1321.