nature genetics

Genomic privacy and limits of individual detection in a pool

Sriram Sankararaman[1,5], Guillaume Obozinski[2,5], Michael I Jordan[1,2] & Eran Halperin[3,4]

Recent studies have demonstrated that statistical methods can be used to detect the presence of a single individual within a study group based on summary data reported from genome-wide association studies (GWAS). We present an analytical and empirical study of the statistical power of such methods. We thereby aim to provide quantitative guidelines for researchers wishing to make a limited number of SNPs available publicly without compromising subjects' privacy.

© 2009 Nature America, Inc. All rights reserved.

A major challenge in the design of a GWAS is that of achieving desired levels of statistical power for detecting weak associations while limiting the rate of false positives. Power of detection can be enhanced by combining data across multiple studies in a meta-analysis or by using replication studies. Such methods require data to be accessible to the scientific community, which may raise concerns over privacy. Until recently, many studies have pooled individual genotype data together while making the allele frequencies of each SNP in the pool publicly available. It has been implicitly assumed that releasing such summary data provides a secure way to share the results of a study without compromising the privacy of the study participants. However, Homer *et al.*[1] recently showed that it is possible, by examining datasets based on high-density SNP arrays, to accurately detect the presence of individual genotypes in a mixture of pooled DNA even when each individual's DNA is present in only small concentrations. Although aimed at applications in forensic science, these findings raised the possibility that the presence of individual genotypes could be inferred from summary data, and this possibility has led to the removal of publicly available summary data from previous studies as a conservative means of protecting the privacy of human subjects[2].

For many applications in genetic analyses[3–5], it is sufficient to have access to the summary SNP data for only a subset of the SNPs ('exposed' SNPs). It is therefore worth investigating whether some appropriately defined level of privacy can be maintained if the number of exposed SNPs is sufficiently small. Establishing privacy guidelines of this kind requires an understanding of how the number of exposed SNPs varies as a function of factors such as the allele frequencies of the SNPs, the number of individuals in the DNA pool and—of particular importance—the method used to detect the individual in the pool. An analysis of this kind was pursued by Homer *et al.*[1], who proposed a particular detection method and estimated the statistical power of detecting an individual genotype in a sample of exposed SNPs using this new method. But although an analysis of any specific detection method provides an estimate of its power of detection, it remains possible that another method could provide increased power and that therefore no guarantee could be provided that its power of detection was below some acceptable level. What is needed is an upper bound on the power achievable by any method.

Here, we present an upper bound on the power of detection, which yields guidelines as to which set of SNPs can be safely exposed for a given pool size with a maximal allowable power β and false-positive level α. We approached this problem through a statistical hypothesis testing formulation, for which the likelihood ratio test (LR test, as defined in the **Supplementary Methods**) attains the maximal power achievable[6]. This provides a guarantee that it will be safe to expose a set of SNPs for which the LR test does not achieve sufficient power. Moreover, our empirical results show that the LR test is more powerful than the method suggested in reference 1, especially when α is small. Finally, our theoretical and empirical results lead to a conclusion that is qualitatively different than that of reference 1 in that we found that the power achieved by considering whole-genome datasets is in fact limited.

We characterize the power of the LR test when m common SNPs in linkage equilibrium are exposed in a pool of n individuals. In this case, we show (see **Supplementary Note**) that the relation between m, n, α and β can be described by

$$z_\alpha + z_{1-\beta} \approx \sqrt{(m/n)} \qquad (1)$$

where z_x is the $100(1-x)$th percentile of the normal distribution. Equation (1) is valid for large pools ($n > 100$) and for common SNPs (minor allele frequency > 0.05). It provides an upper bound on the number of SNPs that can be safely exposed for a particular choice of false-positive rate and power of detection. Note that equation (1) implies that m, the allowed number of exposed SNPs, is linear in n for a fixed α and β, and, importantly, that the power of the test does not depend on the allele frequencies $p_1,\ldots,p_m$, as long as the minor allele frequencies (MAFs) are sufficiently large. The conditions necessary for our analysis to hold suggest the following prefiltering protocol to obtain a set of SNPs that can potentially be exposed: remove all SNPs with MAF ≤ 0.05 and retain a subset of SNPs in linkage equilibrium. We then use the LR test to determine the set of exposed SNPs.

To illustrate this protocol, we simulated a pool of $n = 1{,}000$ individuals, with m exposed SNPs ($m = 1{,}000$, $10{,}000$) and a reference dataset of 2,000 individuals (see **Supplementary Methods**). The LR test is based on the population allele frequencies, which in practice are not known. In our experiments, we estimate these from the allele frequencies in the pool and the reference dataset. We then calculated the LR test under two hypotheses: assuming that the tested individual

[1]Computer Science Division and [2]Department of Statistics, University of California, Berkeley, Berkeley, California, USA. [3]International Computer Science Institute, Berkeley, California, USA. [4]School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel. [5]These authors contributed equally to this work. All correspondence should be addressed to E.H. (heran@icsi.berkeley.edu).

Received 16 October 2008; accepted 27 July 2009; published online 23 August 2009; doi:10.1038/ng.436

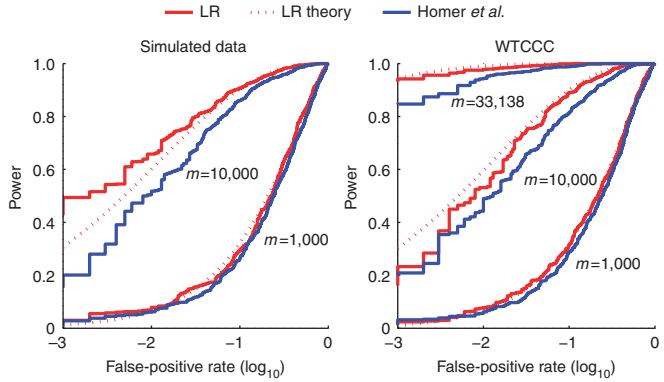

Figure 1 ROC curves comparing the LR test with a plug-in allele frequency estimate, its theoretical power (denoted "LR theory") as computed using a modified version of equation (1) corrected to account for the use of the plug-in estimate, and the statistic proposed by Homer *et al.*[1] on a pool of size $n = 1{,}000$. Left, ROC curves for simulated data with $m = 1{,}000$, 10,000 exposed SNPs. Right, ROC curves on the WTCCC data with $m = 1{,}000$, 10,000 and 33,138 SNPs (the total set of independent SNPs). The LR test performs significantly better ($P = 3.9 \times 10^{-18}$) than the test of Homer *et al.* Nonetheless, the power stays <0.95 for a false-positive level of 10^{-3} even when all the independent SNPs are used. Note the close agreement between the empirical and the theoretical results.

is or is not in the pool. The resulting receiver operating characteristic (ROC) curves (**Fig. 1**) show that the LR test performs better than the test of reference 1, and this difference is particularly prominent for low false-positive rates. We also observe that our theoretical analysis closely matches the empirical evaluation.

We also evaluated the LR test on the 58C and UKBS control group of the Wellcome Trust Case Control Consortium (WTCCC)[7]. This dataset was genotyped on the 500K Affymetrix array and contained 2,937 individuals (see **Supplementary Methods**). We applied the prefiltering described above, asserting independence for P values $< 10^{-5}$. This resulted in 33,138 SNPs that could potentially be exposed.

We created a pool of size 1,000 and a reference dataset consisting of the remaining individuals. Using the protocol described above, it is notable that, at a false-positive level of 10^{-3}, the power stays <0.95 for all methods (**Fig. 1**). At a false-positive level of 10^{-6}, the power is found to be <0.5. The latter contradicts the results of reference 1, which finds that the power to detect individuals is high even with a false-positive level of 10^{-6}. This discrepancy can be attributed to different formulations of the hypothesis testing problem (see **Supplementary Methods** and **Supplementary Figs. 1** and **2**). In particular, the method of reference 1 tests whether an individual is present in the pool or alternatively in the reference dataset, whereas our analysis tests whether an individual is present in the pool or alternatively in the larger underlying population. Although the null hypothesis of reference 1 may be of interest in forensics applications, we argue that our formulation is more relevant to the discussion of privacy issues.

Further analysis of the LR test in modified scenarios (see **Supplementary Methods**) led to several additional conclusions: (i) genotyping errors reduce detection power (see **Supplementary Fig. 3**); (ii) the power to detect a relative in the pool is lower than the power to detect a specific individual (see **Supplementary Fig. 4**); and (iii) our protocol yielded the same pattern of results when applied to the YRI population from the HapMap project, confirming that our analysis is not population specific (see **Fig. 2**). A caveat is that the number of independent SNPs with small MAFs may differ across populations. This would affect the total number of SNPs that can potentially be exposed for a given population.

The analysis presented here provides an upper bound on the power of any method for the detection of an individual in a pool, given the false-positive rate, the size of the pool and the number of exposed SNPs. In using this bound, several issues should be kept in mind. First, our analysis assumes that the exposed SNPs are in linkage equilibrium. When the exposed SNPs are in linkage disequilibrium, the power of the LR test is reduced (see **Supplementary Fig. 5**); nonetheless, under these circumstances, there is a potential risk that one could leverage the linkage disequilibrium in order to get better power from a different test. We thus recommend that dependent SNPs not be exposed until this issue can be studied rigorously. Second, equation (1) is based on the assumptions of common SNPs and large pools (MAF > 0.05 and $n > 100$). The presence of rare SNPs may improve the power of the LR test or other tests and thus jeopardize privacy. The theoretical analysis leading to equation (1) is based on asymptotic estimation of the mean and variance of the LR test under the null and alternative hypotheses, hence the requirement for the pool size. We have studied the effect of pool size empirically using both simulated data and real summary data (see **Supplementary Methods**), and we found that the means and variances converges to the predicted asymptotic values very quickly (see **Supplementary Fig. 6**) and that equation (1) is accurate for $n > 100$ for a population of European descent (see **Supplementary Fig. 7**). However, unless it is clear that the assumptions of common SNPs and large pools are met, we would recommend that equation (1) be used as a rough guide and that final decisions regarding the set of exposed SNPs should be based on an empirical computation of the power of the LR test.

To this end, we have implemented a tool, SecureGenome, that takes as input a genotype dataset (including the individuals' genotypes), a reference dataset and a ranking of the SNPs, removes SNPs that are in linkage disequilibrium, and determines the number of highly ranked SNPs that can be safely exposed. The program outputs

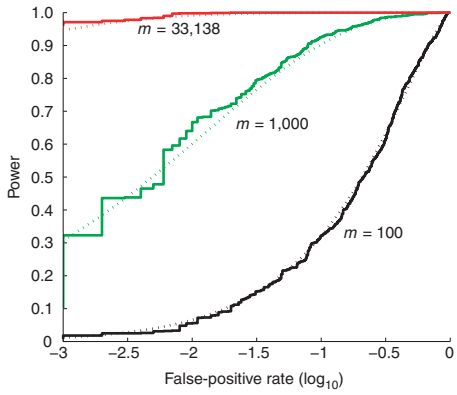

Figure 2 The power attained by the LR test is not population specific. The power of the LR test, with a plug-in allele frequency estimate, computed for $m = 1{,}000$, 10,000 and 33,138 on the HapMap YRI dataset, closely matches its theoretical power, computed using a modified version of equation (1) corrected to account for the use of the plug-in estimate. Note that the theoretical power does not depend on the specific allele frequencies of a population.

© 2009 Nature America, Inc. All rights reserved.

this value along with the power of the LR test evaluated both empirically and theoretically. This tool can serve as a practical guide to allow researchers to develop a consensus that takes into account both privacy and the need to leverage data collected throughout the community.

URL. SecureGenome software: http://securegenome.icsi.berkeley.edu/securegenome/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

E.H. was supported by US National Science Foundation grant IIS-0713254. E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. M.I.J., S.S. and G.O. were supported by NIH/NIGMS R01 grant GM071749. This study makes use of data generated by the Wellcome Trust Case Control Consortium; a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

AUTHOR CONTRIBUTIONS

S.S. and G.O. contributed to the design of the experiments and implemented the experiments; they also developed the theoretical analysis and contributed to writing the paper. M.I.J. contributed to the design of the experiments, the theoretical analysis and the writing of the paper as well as to the funding of the project. E.H. initiated the project and proposed the framework, and he contributed to the design of the experiments, the theoretical analysis, the writing of the paper and the funding of the project.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Homer, N. *et al. PLoS Genet.* **4**, e1000167 (2008).
2. Gilbert, N. *Nature* doi:10.1038/news.2008.1083 (4 September 2008).
3. Barrett, J.C. *et al. Nat. Genet.* **40**, 955–962 (2008).
4. Zeggini, E. *et al. Nat. Genet.* **40**, 638–645 (2008).
5. Cooper, J.D. *et al. Nat. Genet.* **40**, 1399–1401 (2008).
6. Lehmann, E.L. *Testing Statistical Hypotheses* (Springer, New York, 2005).
7. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–683 (2007).

© 2009 Nature America, Inc. All rights reserved.