Accurate liability estimation improves power in ascertained case-control studies

Omer Weissbrod[1], Christoph Lippert[2], Dan Geiger[1] & David Heckerman[2]

Linear mixed models (LMMs) have emerged as the method of choice for confounded genome-wide association studies. However, the performance of LMMs in nonrandomly ascertained case-control studies deteriorates with increasing sample size. We propose a framework called LEAP (liability estimator as a phenotype; https://github.com/omerwe/LEAP) that tests for association with estimated latent values corresponding to severity of phenotype, and we demonstrate that this can lead to a substantial power increase.

In recent years, genome-wide association studies (GWAS) have uncovered thousands of risk variants for genetic traits[1]. However, only a small fraction of disease variance is explained by discovered variants, possibly because contemporary sample sizes are relatively small and causal variants tend to have small effect sizes[2]. To identify such variants, future studies will need to include hundreds of thousands of individuals.

Population structure and family relatedness[3] lead to spurious results and a higher type I error rate. As sample sizes continue to increase, this difficulty becomes even more severe because larger samples are more likely to include individuals with a different genetic ancestry or related individuals.

Recently, LMMs have emerged as the method of choice for GWAS because of their robustness with respect to diverse sources of confounding[3]. LMMs gain resilience to confounding by testing for association conditioned on pairwise kinship coefficients between study subjects. Although designed for continuous phenotypes, LMMs have been used in several large case-control GWAS[4–6] because alternative methods cannot capture diverse sources of confounding[3].

However, LMMs in ascertained case-control studies, wherein cases are oversampled relative to the disease prevalence, lose power with increasing sample size compared to alternative methods[7]. This loss is due to several model violations: dependence between tested causal variants and variants used to estimate kinship, dependence between genetic and environmental effects, and use of a noncontinuous trait (**Supplementary Note 1**). Thus, the use of LMMs resolves the difficulty of sensitivity to confounding but leads to a different difficulty instead.

A possible remedy is to test for associations with a model that directly represents the case-control phenotype and takes the ascertainment scheme into account (**Supplementary Note 1**). Such models assume that observed case-control phenotypes are generated by an unobserved stochastic process with a well-defined distribution. One prominent example is the liability threshold model[8], which associates individuals with a latent, normally distributed variable called the liability, such that cases are individuals whose liability exceeds a given cutoff. Despite their elegance, such models are extremely computationally expensive, rendering whole-genome association tests infeasible in most circumstances.

As an alternative, we propose approximating such models by first estimating latent liability values and model parameters conditional on phenotypes, genotypes and disease prevalence and then testing for association with the estimated liabilities via an LMM (Online Methods). LEAP is motivated by the observation that cases of rare diseases have a sharply peaked liability distribution (**Supplementary Fig. 1**), leading to highly accurate liability estimation (**Supplementary Note 1**). When testing for association in ascertained case-control studies, LEAP yields substantially increased power over naïve LMMs while remaining resilient to confounding because it largely compensates for the violations listed above.

LEAP bears similarities to several recent methods for estimating the portion of the liability that is due to a small set of explanatory variables[9,10]. Unlike these methods, LEAP estimates liabilities using the entire genome (**Supplementary Note 2**). In parallel work, Hayeck *et al.* proposed another framework called LTMLM (liability threshold mixed linear model) for association testing in ascertained case-control studies[11]. Both LTMLM and LEAP first estimate latent liability values and then test for association with these estimates. However, whereas LTMLM tests for association with the posterior mean of the liabilities in a score-test framework, LEAP tests for association with a maximum a posteriori (MAP) estimate, which is often more robust to model violations and can be evaluated at a substantially lower computational cost (Online Methods and **Supplementary Note 1**).

We evaluated the performance of LEAP on synthetic and real data sets, using the following methods for comparison: (i) LEAP, (ii) a standard LMM, (iii) a linear regression test using ten principal-component (PC) covariates[12] (denoted Linreg+PCs) and (iv) a univariate linear regression test (Linreg) without PC covariates, used as a baseline measure. Linreg+PCs and Linreg use the linear link function to prevent evaluation bias due to using a different

[1]Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel. [2]eScience Group, Microsoft Research, Los Angeles, California, USA. Correspondence should be addressed to O.W. (omerw@cs.technion.ac.il) or D.H. (heckerma@microsoft.com).

RECEIVED 5 SEPTEMBER 2014; ACCEPTED 18 DECEMBER 2014; PUBLISHED ONLINE 9 FEBRUARY 2015; DOI:10.1038/NMETH.3285

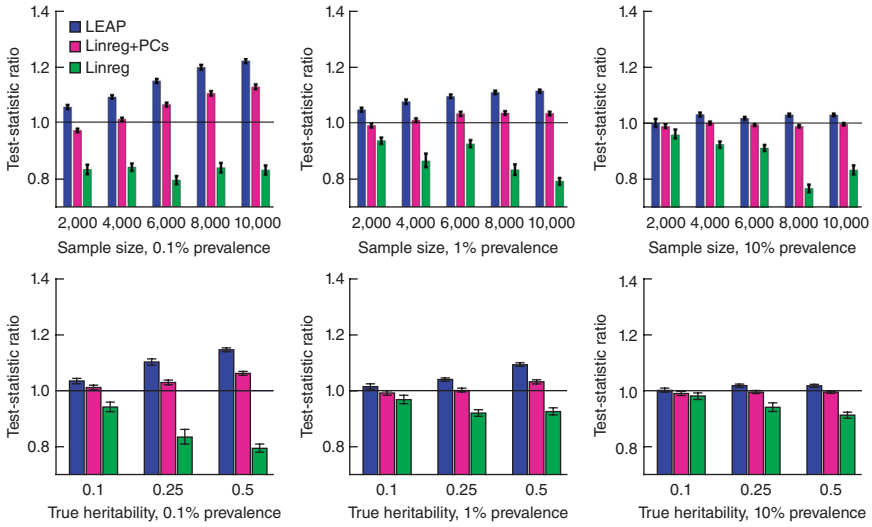

Figure 1 | Synthetic data demonstrating that the power of LEAP increases with sample size and disease heritability. The values shown are the mean ratio of normalized test statistics for causal SNPs between each evaluated method and a standard LMM, under different sample sizes (top row) and heritability levels (bottom row), and 95% confidence intervals. Larger mean ratios indicate higher power. Values above the horizontal line indicate that a method has test statistics that are on average greater than that of a standard LMM.

link function. Experiments using logistic regression yielded results very similar to those with Linreg (data not shown).

Sensitivity to confounding was evaluated by measuring type I error rates for synthetic data sets (**Supplementary Note 2**). We evaluated various combinations of population structure (quantified via the F_{ST} measure[13]) and family relatedness (measured via the fraction of sibling pairs), using F_{ST} levels of 0, 0.01 and 0.05 and sibling pair fractions of 0%, 3% and 30%. Only LEAP and a standard LMM properly controlled for type I error in the presence of confounding (**Supplementary Figs. 2** and **3** and **Supplementary Table 1**).

The power of the methods was evaluated according to the distribution of test statistics of causal variants[9,10]—normalized according to the type I error rate—to prevent Linreg and Linreg+PCs from falsely appearing to be more powerful than other methods owing to inflation of *P* values (**Supplementary Note 2**).

To investigate the effects of sample size and ascertainment on power, we generated ascertained case-control data sets with prevalences of 0.1%, 1% and 10%, and sample sizes in the range 2,000–10,000. The advantage of LEAP over a standard LMM increased with sample size and with decreasing prevalence (**Fig. 1** and **Supplementary Figs. 4** and **5**). In simulated samples with 0.1% prevalence and 10,000 individuals, LEAP gained an average increase of over 20% in test statistics of causal variants (**Fig. 1**) and a power increase of over 5% for significance thresholds smaller than 5×10^{-5} (**Supplementary Fig. 4**). LEAP also outperformed other methods under more complex ascertainment schemes (**Supplementary Note 1**). We further verified that the increased power of LEAP stems from its accurate liability estimation in the presence of ascertainment (**Supplementary Fig. 6** and **Supplementary Note 1**).

Accurate liability estimation depends on the fraction of liability variance that is driven by genetic factors, called the narrow-sense heritability[2]. A higher heritability is expected to improve estimation accuracy because more of the liability signal can be inferred from observed variants. We empirically verified that the advantage of LEAP over other methods increased with heritability, with noticeable power gains for diseases with heritability ≥25% (**Fig. 1** and **Supplementary Fig. 7**). We also performed a series of experiments to demonstrate that LEAP outperforms other methods under diverse levels of population structure, family relatedness, polygenicity and covariate effects (**Supplementary Figs. 8–15** and **Supplementary Note 2**).

To evaluate performance on real data, we analyzed nine disease data sets from the Wellcome Trust Case Control Consortium (WTCCC)[5,14,15]. Measuring power for real data sets is an inherently difficult task because the identities of true causal single-nucleotide polymorphisms (SNPs) are unknown. Evaluating type I error rates for real data is also a difficult task because inflation of *P* values may stem either from sensitivity to confounding or from high polygenicity of the studied trait[16].

As an approximate measure for type I error, we verified that the proportion of SNPs having $P < 0.05$ and $P < 10^{-5}$, and that are not within 2 Mbp of SNPs reported to be associated with the disease in previous studies, is comparable under LEAP and under a standard LMM. As an approximate measure for power, we computed normalized test statistics for SNPs that tag known

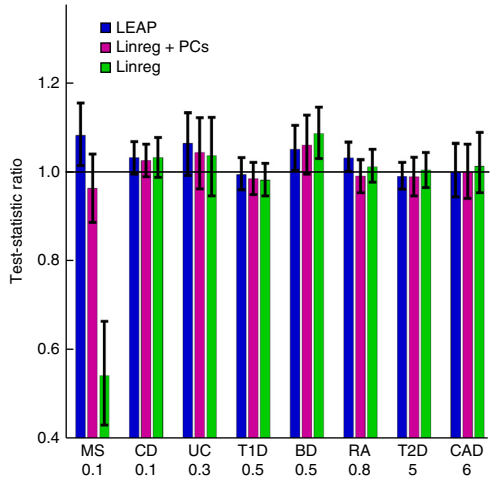

Figure 2 | Analysis of real data sets with LEAP and other methods. The values shown are the mean ratio of normalized test statistics of SNPs tagging known variants between each evaluated method and a standard LMM, with 95% confidence interval. A higher mean ratio indicates higher power. Values above the horizontal line indicate that a method has test statistics that are on average greater than that of a standard LMM. The number of tag SNPs is 44, 38, 20, 11, 19, 35, 17 and 15 for multiple sclerosis (MS), Crohn's disease (CD), ulcerative colitis (UC), type 1 diabetes (T1D), bipolar disorder (BD), rheumatoid arthritis (RA), type 2 diabetes (T2D) and coronary artery disease (CAD), respectively. The prevalence of each disease is shown below its name (in percent units).

risk variants from the US National Human Genome Research Institute catalog[17] as a 'bronze standard'.

LEAP demonstrated robustness to confounding and was significantly more powerful than a standard LMM ($P < 0.05$) in five out of the six rare phenotypes, with prevalence less than 1% (**Fig. 2** and **Supplementary Tables 2** and **3**; results for hypertension are omitted from **Fig. 2** owing to the low number of known risk variants). As expected from the simulations, the advantage of LEAP increased with sample size and with confounding. Thus, only a small advantage was observed in the WTCCC1 data sets, which contain about 4,500 individuals per data set and little population structure or family relatedness, whereas a significantly greater advantage was observed in the larger and more confounded WTCCC2 data sets.

In the highly confounded multiple sclerosis (MS) data set[5], LEAP obtained a mean increase of more than 8% over an LMM in test statistics of tag SNPs, and an even greater advantage over other methods, while demonstrating robustness to confounding. All genome-wide significant loci identified by LEAP and LMM, with $P < 5 \times 10^{-8}$, have previously been reported to be associated with MS in meta-analyses. In contrast, Linreg+PCs and Linreg identified 2 and 508 previously unidentified significant loci, respectively.

There are several avenues for future research. First, liabilities follow a truncated multivariate normal distribution; thus, their likelihood cannot be computed by an LMM without model misspecification, even if they are perfectly estimated. It may be possible to modify the objective function of LEAP so that its estimated liabilities follow a normal distribution[18]. Second, liability estimation may be improved by improving kinship estimation in ascertained studies (**Supplementary Note 1**). Finally, liability estimation may be improved by adopting richer models with a heterogeneous effect-size variance[19,20].

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the Israeli Science Foundation. 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. The MS and ulcerative colitis data sets were filtered by A. Gusev. We thank N. Zaitlen and S. Rosset for helpful discussions.

AUTHOR CONTRIBUTIONS

O.W. and D.H. designed research, conducted experiments, contributed analytic tools, analyzed data and wrote the paper. C.L. and D.G. designed research and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Welter, D. *et al. Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Golan, D., Lander, E.S. & Rosset, S. *Proc. Natl. Acad. Sci. USA* **111**, E5272–E5281 (2014).
3. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. *Nat. Rev. Genet.* **11**, 459–463 (2010).
4. Fakiola, M. *et al. Nat. Genet.* **45**, 208–213 (2013).
5. Sawcer, S. *et al. Nature* **476**, 214–219 (2011).
6. Tsoi, L.C. *et al. Nat. Genet.* **44**, 1341–1348 (2012).
7. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. *Nat. Genet.* **46**, 100–106 (2014).
8. Dempster, E.R. & Lerner, I.M. *Genetics* **35**, 212–236 (1950).
9. Zaitlen, N. *et al. Bioinformatics* **28**, 1729–1737 (2012).
10. Zaitlen, N. *et al. PLoS Genet.* **8**, e1003032 (2012).
11. Hayeck, T. *et al.* Preprint at http://biorxiv.org/content/early/2014/09/04/008755 (2014).
12. Price, A.L. *et al. Nat. Genet.* **38**, 904–909 (2006).
13. Wright, S. *Ann. Eugen.* **15**, 323–354 (1949).
14. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
15. The UK IBD Genetics Consortium & the Wellcome Trust Case Control Consortium 2. *Nat. Genet.* **41**, 1330–1334 (2009).
16. Yang, J. *et al. Eur. J. Hum. Genet.* **19**, 807–812 (2011).
17. Hindorff, L.A. *et al. Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
18. Fusi, N., Lippert, C., Lawrence, N.D. & Stegle, O. *Nat. Commun.* **5**, 4890 (2014).
19. Zhou, X., Carbonetto, P. & Stephens, M. *PLoS Genet.* **9**, e1003264 (2013).
20. Widmer, C. *et al. Sci. Rep.* **4**, 6874 (2014).

ONLINE METHODS

LEAP overview. The LEAP procedure is composed of three parts, which are now briefly overviewed, with detailed explanations below.

1. Heritability estimation: the heritability of a trait quantifies the degree to which it is driven by genetic factors[2,21,22]. Several methods for heritability estimation in case-control studies have been proposed recently[2,22]. We adopt the method of ref. 2, which directly models the ascertainment procedure.
2. Liability estimation: using the heritability estimate, we fit a regularized probit model to estimate the effect size of each genetic variant on the liability. We use the probit model to estimate liabilities for the sample individuals.
3. Association testing: the liability estimate is used as an observed phenotype in a GWAS context. Genetic variants are tested for association with this estimate via a standard LMM. The LMM is fitted using the heritability estimate, as described below.

We first provide a brief overview of the liability threshold model and then derive a corresponding liability estimation model. Detailed derivations are found in **Supplementary Note 3**.

The liability threshold model. According to the liability threshold model, every individual i is associated with a latent normally distributed variable $l_i \sim N(0, 1)$, such that cases are individuals with $l_i > t$, where t is the liability cutoff for a particular trait of interest. Assuming a trait with a population prevalence K, t is given by $\Phi^{-1}(1 - K)$, where $\Phi^{-1}(\bullet)$ is the inverse cumulative probability density of the standard normal distribution. We decompose the liability l_i into $l_i = g_i + e_i$, where g_i and e_i are the genetic and environmental components of the liability, respectively. Under standard assumptions, $e_i \sim N(0, \sigma_e^2)$ and $g_i = X_i^T\beta$, where X_i is an $m \times 1$ vector of m standardized genetic variants carried by individual i, and β is an $m \times 1$ vector of effect sizes, which follows the distribution

$$\beta \sim N\left(0, \frac{\sigma_g^2}{m} I\right)$$

where I is the identity matrix. The genetic and environmental variances σ_g^2 and σ_e^2 are closely related to the narrow-sense heritability of a trait[2], defined as $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$.

Liability estimation. As discussed in **Supplementary Note 3**, testing for associations under the liability threshold model requires integrating the underlying liability vector over its support. Motivated by this observation, we propose approximating such association testing by selecting a liability estimator and treating it as the observed phenotype vector. A good liability estimator has values close to the true, unobserved, underlying liability. Thus, the problem is equivalent to inferring the value of an unknown continuous variable with a known distribution.

We estimate liabilities via a maximum a posteriori (MAP) estimator, which estimates the MAP of the effect sizes of all genetic variants conditional on the phenotypes, genotypes and disease prevalence. The likelihood to maximize can be written as

$$P\left(\beta; 0, \frac{\sigma_g^2}{m} I\right) \prod_{i \in \text{controls}} \Phi\left(t - X_i^T\beta; 0, \sigma_e^2\right) \prod_{i \in \text{cases}} \left(1 - \Phi\left(t - X_i^T\beta; 0, \sigma_e^2\right)\right) \quad (1)$$

Taking the logarithm and using the normal distribution definition, the quantity to maximize is

$$\sum_{i \in \text{controls}} \log \Phi\left(t - X_i^T\beta; 0, \sigma_e^2\right) + \sum_{i \in \text{cases}} \log\left(1 - \Phi\left(t - X_i^T\beta; 0, \sigma_e^2\right)\right) - \frac{1}{2\sigma_g^2/m} \sum_j \beta_j^2 + W \quad (2)$$

where W is a quantity that does not depend on β and can thus be ignored. This problem is equivalent to probit regression[23] with L2 regularization and a prespecified offset term, and can thus be solved using standard techniques (**Supplementary Note 2**). Unlike typical uses of such models, here the regularization parameter is known in advance, given a value for σ_g^2.

The MAP for g_i is given by $\hat{g}_i = X_i^T\hat{\beta}$, where $\hat{\beta}$ is the MAP of β. Given the MAP $\hat{g}_i$, $\hat{l}_i$ is equal to $\hat{g}_i$ if individual i is a control and $\hat{g}_i < t$ or if individual i is a case and $\hat{g}_i > t$, and is equal to t otherwise. This follows because e_i has a zero-mean normal distribution.

Dimensionality reduction. A straightforward solution of the optimization problem presented above is difficult owing to high dimensionality, which is equal to the number of genotyped variants. Fortunately, the problem can be reformulated as a lower-dimensional problem, with dimensionality equal to the number of individuals.

The equivalence stems from the fact that the genotypes matrix X can be represented in terms of the eigenvectors of its covariance matrix alone. To see this, we rewrite equation (2) as follows:

$$f(X\beta) - \frac{1}{2\sigma_g^2/m} \beta^T\beta \quad (3)$$

where $f(X\beta)$ is a function that depends on β only through the product $X\beta$. Consider the singular value decomposition (SVD) of X, given by

$$X = USV^T \quad (4)$$

where U is the matrix of the eigenvectors of XX^T, and V is orthonormal. Denote $Z = US$ and $\beta_Z = V^T\beta$. Owing to the orthonormality of V, the following equations hold:

$$\beta_Z{}^T\beta_Z = \beta^T\beta \quad (5)$$

$$Z\beta_Z = X\beta \quad (6)$$

Therefore, equation (3) can be rewritten as

$$f(Z\beta_Z) - \frac{1}{2\sigma_g^2/m}\beta_Z{}^T\beta_Z \quad (7)$$

Denoting the number of individuals and genotyped variants by n and m, respectively, and assuming $m > n$ and that the columns of Z are ordered according to the magnitude of their respective eigenvalues, then all columns of Z except for the leftmost n ones are equal to 0. Consequently, the vector $Z\beta_Z$ depends only on the top n entries of the vector β_Z, and thus all the other entries can be set to 0.

We conclude that the quantity in equation (2) can be maximized by considering only the nonzero components of the matrix Z and the vector β_Z, which have dimensionalities $n \times n$ and n, respectively. In contrast, the original formulation of the problem uses the matrix X and the vector β, which have dimensionalities $n \times m$ and m, respectively. The original effect sizes are given by $\beta = V\beta_Z$. However, they are not needed in practice, as the liabilities estimator can be computed using β_Z directly.

Finally, we note that when performing GWAS, the matrix Z is typically computed regardless of whether LEAP is employed and is thus available at no further computational cost. This results from the close relation between the SVD of X and the eigendecomposition of the matrix XX^T. Namely, given the eigendecomposition $XX^T = US^2U^T$, the matrix Z is given by $Z = US$, where S is the matrix of the componentwise square roots of the entries of S^2. In GWAS, the eigendecomposition of XX^T is computed both when using an LMM[24] and when performing regression using principal-component covariates[12] and is thus available for use in LEAP at no further computational cost.

Use in GWAS. LEAP uses liability estimates by treating them as observed continuous phenotypes in an LMM. Three difficulties that must be dealt with are accurate fitting of the LMM parameters, avoiding testing SNPs for association with the liability estimator that they helped estimate and dealing with family relatedness. We now describe solutions to these difficulties.

The difficulty of parameter estimation stems from the non-normality of the liability under case-control sampling. This non-normality arises because in rare diseases, the majority of cases share a similar liability close to the cutoff. Parameter estimation can be suboptimal in such settings. The most important parameter that is fitted in LMMs is the variances ratio $\delta = \sigma_e^2/\sigma_g^2$. Given this parameter, all other parameters can be evaluated via closed-form formulas[24]. There is a close connection between this parameter and the narrow-sense heritability, $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, expressed via $\delta = 1/h^2 - 1$. We therefore fit this parameter by estimating the heritability using the method of ref. 2, as described in **Supplementary Note 2**.

A second difficulty arises because SNPs should not be tested for association with a liability estimator that they helped estimate. Otherwise the test statistic for these SNPs will be inflated because they can always account for some of the liability variance. Similarly, SNPs in linkage disequilibrium with a tested SNP should not participate in the liability estimation. To prevent such inflation, we estimate liabilities on a per-chromosome basis. For every chromosome, the liability is estimated using all SNPs except for the ones on the chromosome. The SNPs on the excluded chromosome are then tested for association using this liability estimator. We note that LMM-based GWAS typically compute the eigendecomposition of the covariance matrix on a per-chromosome basis as well in order to prevent a SNP from incorrectly affecting the null likelihood (the phenomenon termed proximal contamination[7,24,25]). LEAP can make use of these available eigendecompositions for dimensionality reduction, thus incurring no computational cost other than the liability estimation procedure itself.

A third difficulty arises when the data are confounded by family relatedness. The presence of related individuals can lead to biased effect-size estimates and, consequently, to a biased liability estimator. We deal with this difficulty by excluding related individuals from the parameter estimation stage of the MAP computation. We employ a greedy algorithm, where at each stage we exclude the individual having the largest number of related individuals with correlation coefficient >0.05. After fitting the model, we estimate liabilities for the excluded individuals as well. We note that population structure does not present similar problems because it is naturally captured by top principal components[3,12,26], which are fitted in the MAP computation.

Data simulation. All experiments reported in this paper are based on a uniform data generation procedure that can simulate different settings via a variety of parameters. In these simulations, each individual carried 60,100 SNPs that do not affect the phenotype, as well as 50–5,000 causal SNPs with normally distributed effect sizes. Population structure was simulated via the Balding-Nichols model[27], which generates populations with genetic divergence measured via Wright's F_{ST}[13]. Family relatedness was simulated by generating various numbers of sibling pairs in one of the two populations, as in ref. 3. To simulate ascertainment, we generated $3{,}000/K$ individuals and a latent liability value for every individual, where K is the disease prevalence. We then determined the $1 - K$ percentile of the liabilities and generated new individuals until 50% of the sample had liabilities exceeding this cutoff[9]. Unless otherwise noted, all simulations use 6,000 individuals, $F_{ST} = 0.01$, and 30% of the individuals in one of the two populations are sibling pairs. In all experiments, ten data sets were generated for each unique combination of settings. A detailed description of the simulation procedure and its default parameters is provided in **Supplementary Note 2**.

Software and code availability. LEAP is available to download from https://github.com/omerwe/LEAP.

LEAP has the same memory requirements as the FaST-LMM package[20] and is computationally efficient. On a 2-GHz CPU, it can accurately estimate liabilities for samples as large as 50,000 individuals in fewer than 5 min.

21. Yang, J. *et al. Nat. Genet.* **42**, 565–569 (2010).
22. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
23. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn. (Springer, 2009).
24. Lippert, C. *et al. Nat. Methods* **8**, 833–835 (2011).
25. Listgarten, J. *et al. Nat. Methods* **9**, 525–526 (2012).
26. Patterson, N., Price, A.L. & Reich, D. *PLoS Genet.* **2**, e190 (2006).
27. Balding, D.J. & Nichols, R.A. *Genetica* **96**, 3–12 (1995).

doi:10.1038/nmeth.3285