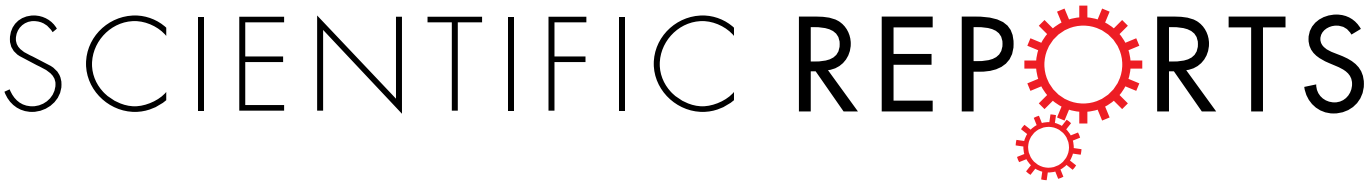

OPEN

Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization —GALLOP algorithm

Received: 14 December 2017
Accepted: 22 March 2018
Published online: 01 May 2018

Karolina Sikorska[1], Emmanuel Lesaffre[2], Patrick J. F. Groenen[3], Fernando Rivadeneira[4] & Paul H. C. Eilers[5]

Genome-wide association studies (GWAS) with longitudinal phenotypes provide opportunities to identify genetic variations associated with changes in human traits over time. Mixed models are used to correct for the correlated nature of longitudinal data. GWA studies are notorious for their computational challenges, which are considerable when mixed models for thousands of individuals are fitted to millions of SNPs. We present a new algorithm that speeds up a genome-wide analysis of longitudinal data by several orders of magnitude. It solves the equivalent penalized least squares problem efficiently, computing variances in an initial step. Factorizations and transformations are used to avoid inversion of large matrices. Because the system of equations is bordered, we can re-use components, which can be precomputed for the mixed model without a SNP. Two SNP effects (main and its interaction with time) are obtained. Our method completes the analysis a thousand times faster than the R package lme4, providing an almost identical solution for the coefficients and p-values. We provide an R implementation of our algorithm.

Genome-wide association studies with longitudinal phenotypes create opportunities and challenges. On the one hand we can identify genetic variants that are associated with development of traits over time. On the other hand statistical analysis gets more complicated, because (linear) mixed models have to be used.

In this paper we discuss the application of the linear mixed model to repeated measures, collected on unrelated individuals. We assume that the number of measurements per person is just a handful, allowing to model only a linear evolution of the trait over time. In a genome-wide analysis the mixed model has to be fitted for every SNP. It contains fixed effects for time, SNP, and their interaction, and possibly other covariates; it has random intercept and slope for change over time. Of main interest is the time x SNP effect, but multiple observations per individual also increase the power to detect a statistically significant main SNP effect.

A mixed model assumes that some model parameters, in the present case intercept and slope per individual, have been drawn from a (normal) distribution with unknown variance. Also unknown is the variance of the observation error. Once these variances are known, it is straightforward to estimate individual slopes and intercepts. The hard work for mixed models is estimating the variances. Common software, like SAS PROC MIXED and **lme4** in R do this efficiently, using special algorithms. It takes approximately 2.0 seconds to fit a mixed model for several thousand individuals. For a single application this is fast, but for GWAS it is far too slow. Fitting one million mixed models, one for each SNP, would take several weeks of non-stop computation. This assumes that the overhead of accessing the SNP data is negligible, which usually is not the case.

We emphasize that analysis of longitudinal data is different from analysis of cross-sectional outcomes where mixed models are used either to estimate heritability[1,2] or to correct for hidden correlation due to population stratification[3,4]. Extensive work has been done on how to speed up computations in the latter case, see e.g.[5,6]. Unfortunately it does not solve our problem; see the Discussion.

In an earlier effort, we proposed the conditional two-step (CTS) approach[7], which summarizes the developmental pattern of a trait as an individual slope, reducing the dimensionality of the data to one pseudo-observation

[1]Department of Biometrics, Netherlands Cancer Institute, Amsterdam, The Netherlands. [2]Leuven Biostatistics and Statistical Bioinformatics Centre, Leuven University, Leuven, Belgium. [3]Erasmus School of Economics, Erasmus University, Rotterdam, The Netherlands. [4]Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands. [5]Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands. Correspondence and requests for materials should be addressed to K.S. (email: karolina_sikorska@hotmail.com)

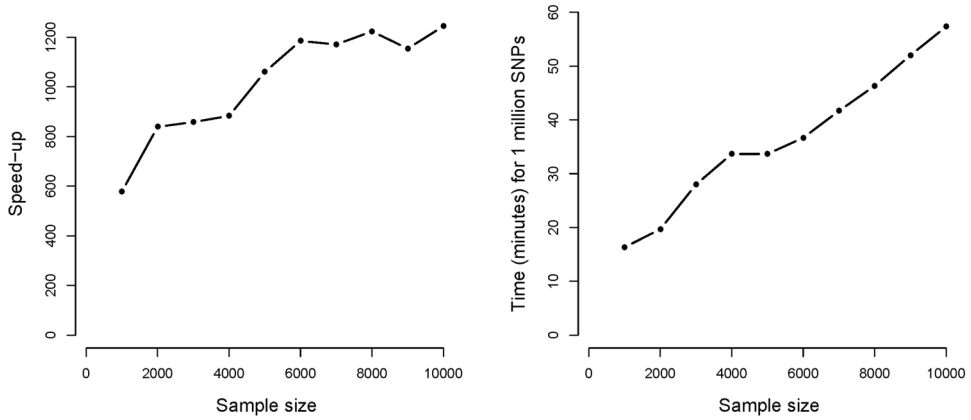

Figure 1. Speed-up compared to the *lmer* function in R. Results based on the simulated data for 1000 SNPs, 4 time points and 3 covariates. Performed on a 64-bit Windows running on a laptop with CPU @ 2.3 GHz and 6 GB RAM.

per individual. This allows the use of our fast GWAS algorithm[8] to obtain an approximate *p*-value for the interaction between SNP and time.

Here, we present a new algorithm for Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization (GALLOP) which swiftly computes coefficients and p-values for cross-sectional and longitudinal SNP effects. To arrive at an almost exact solution we exploit several properties of the model. The effect of a SNP generally is (very) small. We estimate the variances in the mixed model without any SNP and assume that they will not change when a SNP is added. This assumption will lead to conservative p-values in case of non-zero SNP-effects. The magnitude of this imprecision is explored in the Results section. Using the equivalence between a mixed model and penalized least squares, a large system of linear equations can be set up. This system is very sparse (it contains many zeros) and only the last rows and columns change from SNP to SNP. With careful organization of the computations a solution is obtained very quickly. No special programming tricks are needed, our program (about 85 lines) is written in pure R and achieves a speed-up by three orders of magnitude, compared to brute-force application of **lme4**. Thanks to the sparseness of the equations, memory use is modest.

Quick access to SNP data is crucial and we also discuss it. An R implementation of GALLOP algorithm is provided. Simulated and real data are used to illustrate performance.

Results

Two characteristics of our method are of main interest: high speed and accuracy as compared to *lmer* function in the R package **lme4**. We assessed them via a simulation study and using real data.

In the simulation study exploring precision we generated 200 longitudinal data sets on the basis of the mixed model (Equation (3) in the Methods section) using the following settings:

- n = 2000, k = 4, 3 additional covariates
- Measurements occasions ($n \times k$ vector of t_{ij}'s) drawn from a uniform distribution between 0 and 10
- Covariates assumed to be independent, time-varying, and drawn from $\mathcal{N}$ (2, 0.5)
- Coefficient for fixed effects: $\beta_0 = -2.6$, $\beta_1 = -1.9$, β^{COV} independent drawn from $\mathcal{N}(0, 1)$
- SNP effects: β_2 and β_3 independent, 200 equally spaced values between 0 and 1
- Variance-covariance matrix of random effects: $D = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}$ and measurement error $\sigma = 2.5$
- SNPs drawn from a uniform distribution between 0 and 2

Data sets used to evaluate computation times were generated in a similar manner with sample sizes varying from 1k to 10k with increment of 1k. For each sample size 1000 SNPs were analyzed to summarize computation time.

Results of the simulation study assessing computation times are shown in Fig. 1. The speed-up is linear in the number of individuals. For a genome-wide association study with 5000 individuals our methods finishes the analysis a thousand times faster. A genome-wide scan for 1 million SNPs of a phenotype, collected on 5000 individuals measured on 4 occasions, takes about 30 minutes, instead of 3 weeks when using the package **lme4**. The speed-up depends also on *k*. This is mainly attributed to the fact that **lme4** requires expansion of the SNP vector.

Results of the simulations exploring accuracy are summarized graphically. Based on our theoretical derivations described in the Methods section we know that a non-zero main SNP effect affects the approximation of the variance of the random intercept. Similarly, the size of the interaction influences the variance of the random slope. On the other hand, genome-wide association studies typically show only very small SNP effects which barely contribute to the improvement of the goodness of fit. We ran simulations to explore the practical dangers and consequences of using the approximate variances. Despite the difficulties in defining the variance explained in mixed models we used a simple definition quantifying predictive power as the ratio $R^2 = 1 - ||y - \hat{y}||^2 / ||y - \overline{y}||^2$, where $\hat{y}$ stands for the fitted values and $\overline{y}$ for the average of y.

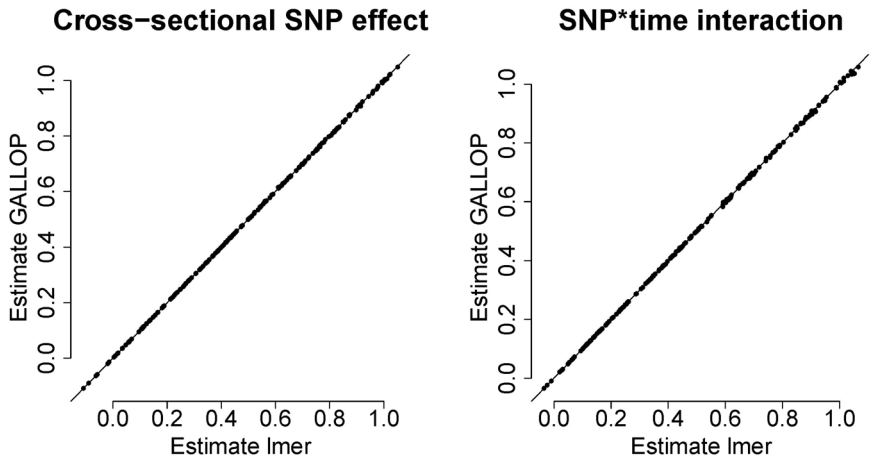

Figure 2. Simulation study. Accuracy of the coefficients computed by GALLOP compared to *lmer*.

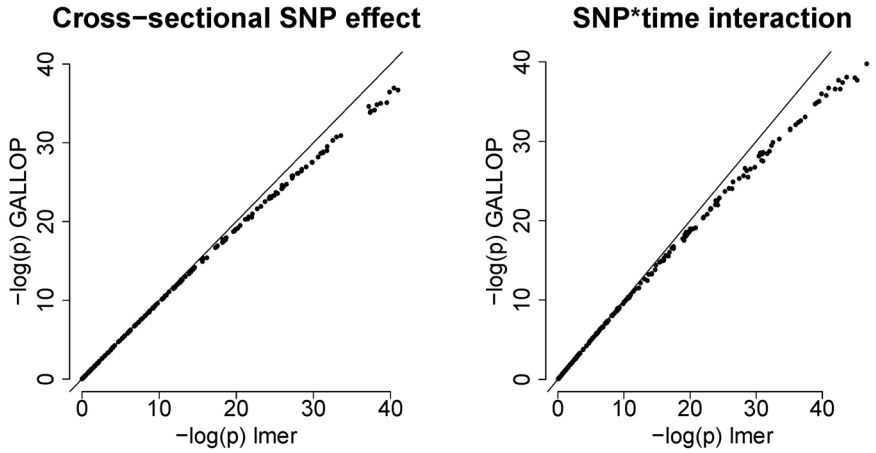

Figure 3. Simulation study. Accuracy of the p-values computed by GALLOP compared to *lmer*.

The estimates are very accurate throughout the entire range of observed values (Fig. 2). The standard errors are somewhat overestimated for the larger values of β, which is expected as variances of random effects are inflated due to omitted SNP effects. However, the main interest in GWAS always lies in *p*-values (Fig. 3). These are almost exact (and never too optimistic) in the common GWA-range ($0 < -\log 10\,(p) < 7$). That eliminates the danger of finding too many false positive results. Due to overestimated standard errors, the $-\log 10(p)$ for larger betas are too pessimistic. Nevertheless, they increase monotonically with larger effect sizes, just with bias downward with respect to the $-\log 10\,(p)$ from *lmer*. This loss of power can be solved by lowering the threshold for "GWAS significance" and repeating the analysis for promising SNPs with the correct model. In our simulation study, to find all SNPs for which $-\log 10\,(p_{lmer}) > 7.3$ we had to use the threshold $-\log 10\,(p_{GALLOP}) > 7.05$. In our simulation study the maximum contribution to R^2 of the SNP effects around 6%.

To confirm the accuracy of GALLOP on real data, we used the BMD data from the Rotterdam Study[9]. Details on the longitudinal BMD data set are provided in ref.[7]. For this analysis we used SNP data imputed according to the 1000 Genomes Project, which were stored per (part of) a chromosome as DatABEL files. To test our algorithm we used one of the files, which contained 97384 SNPs. We performed the association analyses with three methods: GALLOP, CTS, and *lmer* (only for 20 K SNPs). Comparison between p-values is shown in Fig. 4. CTS approach gives a good approximation of the *p*-values for longitudinal SNP effect, which coincide with our previous results on the real and simulated data. However, *p*-values from GALLOP are basically exact for main and longitudinal effect, irrespective of minor allele frequency. The analysis took 3.5 minutes for GALLOP, 40 seconds for CTS and 48 hours (extrapolated time based on the 20 K SNPs) for *lmer*, respectively.

Discussion

We presented a new algorithm for fast genome-wide analysis with longitudinal data. Our method runs a thousand times faster than **lme4**, which is the fastest option in R. This speed-up is achieved by combining an accurate approximation with a careful implementation. We showed that our method provides practically exact results. In case of doubt one can always do a full mixed model analysis for each of the most significant SNPs. Generally this is a small number; in case of BMD data 6 genotypes for any MAF reached threshold of $-\log 10\,(p) > 7$; so the extra computation time is negligible.

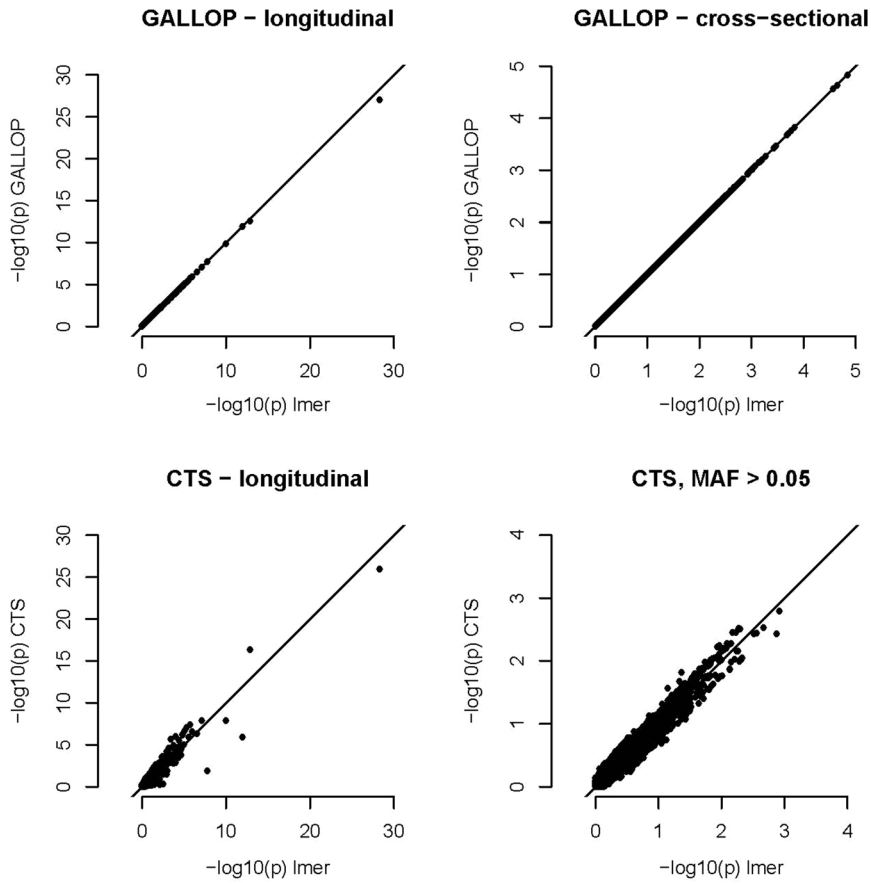

Figure 4. BMD data. Accuracy of the p-values for the GALLOP and CTS.

Our previous approach, conditional two-step (CTS) method combined with semi-parallel regression, computes p-values for the interaction effect about 15 times faster than the GALLOP. However, for CTS, SNP data access is still a bottleneck, 85% of the analysis time is spent on data access (Fig. 5). The genome-wide analysis of the BMD data was completed 5 times faster with CTS than with GALLOP. In case of very massive genome-wide analysis one could consider running CTS to filter out the least significant SNPs and proceed with GALLOP for more precise results.

GALLOP converts a genome-wide analysis with a longitudinal phenotype from a taxing multi-computer task to a job that can be run overnight on a single everyday computer. However, this is only true if access to the SNP data is fast enough. The memory limit in R depends on available RAM, but will usually not be larger than several gigabytes. The size of SNP data, even when split per chromosome, will exceed that size. GALLOP needs quick access to reasonably sized data blocks with multiple SNPs for all individuals. This is possible only when array-oriented binary files are used to store genotypes. We discussed this problem in detail, and proposed solutions in our previous work on fast analyses of cross-sectional outcomes[8].

For correcting population stratification, in cross-sectional GWAS with possibly related individuals, mixed models are well established. Several algorithms have been proposed and implemented performing fast mixed model analysis in this framework. Multiple publications have proposed that this type of mixed models can be tweaked to analyze longitudinal data. Indeed, one may pretend that the repeated outcomes come from different pseudo-individuals and induce the correlation by passing the kinship matrix to the software. A quite extensive discussion on that topic is found in[10]. The author concludes that "the proper" longitudinal data analysis is to be preferred, but that it is too slow. Similarly, in ref.[11]. the authors analyzed longitudinal blood pressure data using EMMA, which tackles cross-sectional outcomes for related individuals. The authors tricked the software by mimicking an autoregressive structure in the kinship matrix. Although both papers study longitudinal data, their results touch only upon the main SNP effect. The interaction between SNP and time is not discussed.

Our algorithm assumes that the individuals are independent. An important extension is to adjust it for longitudinal data collected on related individuals, combining two types of mixed models. One approach to population stratification uses principal components of correlation matrix of the genotypes as covariates. They can be introduced as fixed effects in our model. The overhead is relatively small, because a large mixed model, without SNPs, is fitted once and each SNP is handled as a perturbation as described in the algorithm section.

The preferred approach would be to use multilevel modelling. Two sources of correlation then have to be combined: the temporal correlation between the repeated outcomes and the genetic correlation between the individuals. It would generate an additional random intercept, derived from the kinship matrix, which would destroy

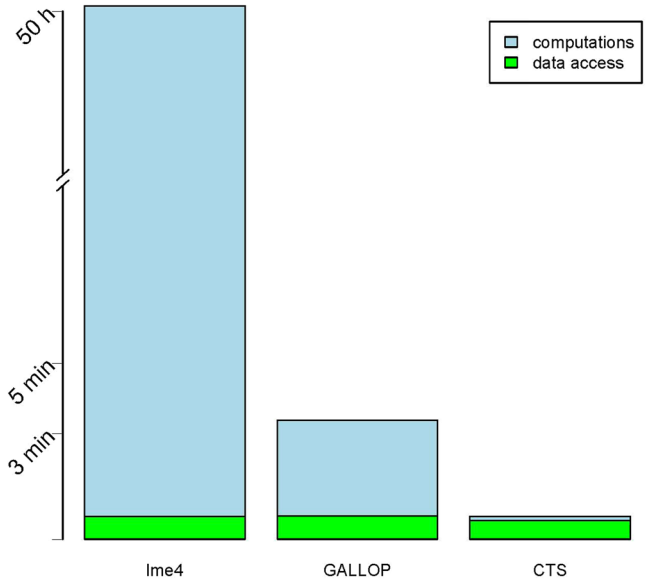

Figure 5. Time of the genome-wide analysis of the BMD data, 97384 SNPs from chromosome 22. Time spent on data access and time spent on computations are separated.

the sparseness of the estimating equations. But still the SNPs can be handled by perturbing a solution obtained without SNPs. This would be an interesting and fruitful topic for the future research.

Methods

A linear mixed model for a longitudinal outcome which assumes random intercepts and slopes has the following hierarchical form[12]:

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad i = 1, \ldots, n \\ b_i \sim N(0, D) \\ \varepsilon_i \sim N(0, \Sigma_i) \\ b_1, \ \ldots, b_n, \varepsilon_1, \ \ldots, \varepsilon_n \quad \text{independent.} \end{cases} \tag{1}$$

In (1) Y_i is k_i dimensional vector of responses for individual i, X_i is $k_i \times p$ matrix with all predictors, Z_i is $k_i \times 2$ dimensional matrix with ones in the first column and t_i in the second column, β is a p-dimensional vector of coefficients identical for all individuals and b_i is a 2-dimensional vector containing the random effects. Measurement error is represented by the k_i-dimensional vector ε_i. Furthermore, D is the 2×2 variance-covariance matrix of random effects and Σ_i is $k_i \times k_i$ the variance-covariance matrix of measurement error. Typically, the unknown parameters, consisting of variances, fixed and random effects, are estimated using for example Newton-Raphson algorithm. However, if the variances are known, the fixed and random effects can be estimated simultaneously by solving a penalized least squares problem given by equations:

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + P \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}. \tag{2}$$

In (2) matrices X, Y, and b are build of X_i's, Y_i's, and b_i's stacked underneath each other. Matrices Z and P are block diagonal with Z_i and P_i on the diagonals, where $P_i = (D/\sigma^2)^{-1}$. System (2) is similar to Henderson's system of equations for mixed models.

A typical linear mixed model in a genome-wide association study will have a form:

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 \mathrm{SNP}_i + \beta_3 t_i \mathrm{SNP}_i + C_i\beta^{\mathrm{COV}} + b_{0i} + b_{1i}t_i + \varepsilon_i, \tag{3}$$

where t_i is a k_i-dimensional vector with measurement occasions, SNP_i is a k_i-dimensional vector with SNP values (constant over time), and C_i is a $k_i \times q$-dimensional matrix with constant or time-varying additional covariates (such as height, weight, age etc.). We call model (3) a full model. Additionally, the reduced model is constructed from (3) omitting the SNP effects, as given in (4)

$$Y_i = \beta_0^* + \beta_1^* t_i + C_i\beta^{\mathrm{COV}^*} + b_{0i}^* + b_{1i}^* t_i + \varepsilon_i^*. \tag{4}$$

The system of equations solving the penalized least squares problem for the reduced model have a special structure. We illustrate it for the case $n = 3$:

$$\left(\begin{array}{c|ccc} X^{*\prime}X^{*} & X_1^{*\prime}Z_1 & X_2^{*\prime}Z_2 & X_3^{*\prime}Z_3 \\ \hline Z_1'X_1^{*} & S_1+P^{*} & 0 & 0 \\ Z_2'X_2^{*} & 0 & S_2+P^{*} & 0 \\ Z_3'X_3^{*} & 0 & 0 & S_3+P^{*} \end{array}\right)\begin{pmatrix}\beta^{*}\\ b_1^{*}\\ b_2^{*}\\ b_3^{*}\end{pmatrix}=\begin{pmatrix}\sum_i r_i\\ r_1\\ r_2\\ r_3\end{pmatrix}, \tag{5}$$

where:

$$S_i=\begin{pmatrix}\sum_j 1 & \sum_j t_{ij}\\ \sum_j t_{ij} & \sum_j t_{ij}^2\end{pmatrix},\quad b_i^{*}=\begin{pmatrix}b_{0i}^{*}\\ b_{1i}^{*}\end{pmatrix}\quad\text{and}\quad r_i=\begin{pmatrix}\sum_j y_{ij}\\ \sum_j t_{ij}y_{ij}\end{pmatrix}.$$

Note that in (5) matrix $X^{*}=[\mathbf{1}\ \ t_i\ \ C_i]$ and the * distinguishes which components of the model are altered (with respect to length and/or values) due to misspecified model (4). The above system has a block structure as divided by the solid lines and can be written as

$$\begin{pmatrix}A_{11} & A_{21}'\\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix}\beta^{*}\\ b^{*}\end{pmatrix}=\begin{pmatrix}q_1\\ q_2\end{pmatrix}, \tag{6}$$

with the explicit solution given by:

$$\beta^{*}=(A_{11}-A_{21}'A_{22}^{-1}A_{21})^{-1}(q_1-A_{21}'A_{22}^{-1}q_2)\quad\text{and}\quad b^{*}=A_{22}^{-1}(q_2-A_{21}\beta^{*}). \tag{7}$$

The P^{*} matrix can be easily obtained by fitting a mixed-effects model excluding SNP in any standard software (for example the R package **lme4**). The software does not explicitly return the P^{*}, but it does return the variance-covariance matrix of the random effects (matrix D) and the variance of measurement error (σ^2). In R matrix P^{*} is obtained by calling solve (D/σ^2).

An additional computational simplification can be obtained by ensuring that A_{22} in (7) is the identity matrix. This goal can be achieved as follows. Any system $Ab=q$ can equivalently be solved by $(KAK')(K^{-1}b)=Kq$. Applied to (7), we get

$$\begin{pmatrix}I & 0\\ 0 & \Phi\end{pmatrix}\begin{pmatrix}A_{11} & A_{21}'\\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & \Phi'\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & \Phi^{-1}\end{pmatrix}\begin{pmatrix}\beta^{*}\\ b^{*}\end{pmatrix}=\begin{pmatrix}I & 0\\ 0 & \Phi\end{pmatrix}\begin{pmatrix}q_1\\ q_2\end{pmatrix} \tag{8}$$

$$\begin{pmatrix}A_{11} & (\Phi A_{21})'\\ \Phi A_{21} & \Phi A_{22}\Phi'\end{pmatrix}\begin{pmatrix}\beta^{*}\\ \Phi^{-1}b^{*}\end{pmatrix}=\begin{pmatrix}q_1\\ \Phi q_2\end{pmatrix} \tag{9}$$

thus, the goal is to choose Φ such that $\Phi A_{22}\Phi'=I$. Fortunately, A_{22} is block diagonal with each 2×2 block being equal to S_i+P^{*}. Consequently, Φ is also block diagonal with 2×2 blocks Φ_i. Then, we need to find Φ_i such that $\Phi_i(S_i+P^{*})\Phi_i'=I$. Let $U_i\Omega_iU_i'$ be the eigendecomposition of S_i+P^{*}, where U_i is the matrix of eigenvectors with $U_i'U_i=I$ and Ω_i is the diagonal matrix of positive eigenvalues. Choose $\Phi_i=\Omega_i^{-1/2}U_i'$ and it is readily verified that

$$\Phi_i(S_i+P^{*})\Phi_i'=\Omega_i^{-1/2}U_i'(S_i+P^{*})U_i\Omega_i^{-1/2}=\Omega_i^{-1/2}U_i'U_i\Omega_iU_i'U_i\Omega_i^{-1/2}=I. \tag{10}$$

The linearly transformed system becomes

$$\left(\begin{array}{c|ccc} X^{*\prime}X^{*} & (\Phi_1Z_1'X_1)' & (\Phi_2Z_2'X_2)' & (\Phi_3Z_3'X_3)' \\ \hline \Phi_1Z_1'X_1 & I & 0 & 0 \\ \Phi_2Z_2'X_2 & 0 & I & 0 \\ \Phi_3Z_3'X_3 & 0 & 0 & I \end{array}\right)\left(\begin{array}{c}\beta^{*}\\ \hline \theta_1\\ \theta_2\\ \theta_3\end{array}\right)=\left(\begin{array}{c}\sum_i r_i\\ \hline \Phi_1r_1\\ \Phi_2r_2\\ \Phi_3r_3\end{array}\right) \tag{11}$$

$$\begin{pmatrix}A_{11} & A_{21}^{\text{tran}\prime}\\ A_{21}^{\text{tran}} & I\end{pmatrix}\begin{pmatrix}\beta^{*}\\ \theta\end{pmatrix}=\begin{pmatrix}q_1\\ q_2^{\text{tran}}\end{pmatrix} \tag{12}$$

with $\theta_i=\Omega_i^{1/2}U_i'b_i$ and solutions

$$\beta^{*}=(A_{11}-A_{21}^{\text{tran}\prime}A_{21}^{\text{tran}})^{-1}(q_1-A_{21}^{\text{tran}\prime}q_2^{\text{tran}})\quad\text{and}\quad\theta=q_2^{\text{tran}}-A_{21}^{\text{tran}}\beta^{*}. \tag{13}$$

Note that the random effects have been transformed, such that $b_i^{*}=U_i\Omega_i^{-1/2}\theta_i$. Usually, the solution for the subject-specific effects is not of interest in GWA analyses. Nevertheless, random intercepts and slope from the reduced model can be easily obtained from the lme4 package.

We add a SNP to the model, creating a border to the previous system of equations. Two effects, cross-sectional and longitudinal are added, so $G=[\text{SNP}\ \ \text{SNP}*t]$ is a $\Sigma_ik_i\times 2$ dimensional matrix, with SNP values repeated k_i times for all individuals in the first column and the SNP values repeated k_i times multiplied time occasions in the second column. Repeating SNP data for each individual k_i times seems like a time consuming step. However, SNP is constant over time and thus $G_i=\text{SNP}_i*Z_i$. In our implementation the vector replicating SNP vector k

times is never created explicitly. Regardless the value of k_i SNP values have to be replicated twice per individual resulting in additional efficiency. The augmented system of equation has the form:

$$\left(\begin{array}{c|c|ccc} G'G & G'X^* & G_1'Z_1\Phi_1' & G_2'Z_2\Phi_2' & G_3'Z_3\Phi_3' \\ \hline X^{*\prime}G & X^{*\prime}X^* & X_1^{*\prime}Z_1\Phi_1' & X_2^{*\prime}Z_2\Phi_2' & X_3^{*\prime}Z_3\Phi_2' \\ \Phi_1Z_1'G_1 & \Phi_1Z_1'X_1^* & I & 0 & 0 \\ \Phi_2Z_2'G_2 & \Phi_2Z_2'X_2^* & 0 & I & 0 \\ \Phi_3Z_3'G_3 & \Phi_3Z_3'X_3^* & 0 & 0 & I \end{array}\right)\left(\begin{array}{c} \beta^{\mathrm{SNP}} \\ \beta^* \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{array}\right) = \left(\begin{array}{c} G'y \\ \hline X^{*\prime}y \\ \Phi_1Z_1'y_1 \\ \Phi_2Z_2'y_2 \\ \Phi_3Z_3'y_3 \end{array}\right), \tag{14}$$

where $\beta^{\mathrm{SNP}} = (\beta_2, \beta_3)'$. Note that system (14) is just Henderson's system for the full model, where $X = [X^*G]$ and the transformations have been used to simplify $Z'Z + P$. The transformation has been done based on P^* and not P, assuming that they are the same. This assumption does not strictly hold, but the approximation is very precise. In another article[13] we showed that D^* of the SNP-model equals

$$\begin{pmatrix} \sigma_0^2 + \beta_2^2 \text{ var (SNP)} & \rho\sigma_1\sigma_2 + \beta_2\beta_3 \text{ var (SNP)} \\ \rho\sigma_1\sigma_2 + \beta_2\beta_3 \text{ var (SNP)} & \sigma_1^2 + \beta_3^2 \text{ var (SNP)} \end{pmatrix}. \tag{15}$$

When a SNP is not important in the model, i.e β_2 and β_3 are practically zero, D^* is essentially equal to D. This is the case for most of the SNPs in GWAS. In the situation when SNP has an effect (cross-sectional and/or longitudinal), the variances in D^* will be inflated. The cross-sectional effect inflates the variance of the random intercept, while the longitudinal effect affects the variance of the random slope. The magnitude of this inflation depends on the β_2 and β_3. The covariance in D^* is influenced only if both SNP effects are non-zero.

We can write the system (14) as

$$\begin{pmatrix} H_{11} & H_{21}' \\ H_{21} & H_{22} \end{pmatrix}\begin{pmatrix} \beta^{\mathrm{SNP}} \\ \psi \end{pmatrix} = \begin{pmatrix} J_{11} \\ J_{21} \end{pmatrix}. \tag{16}$$

Solving system (16) for β^{SNP} gives us

$$\beta^{\mathrm{SNP}} = (H_{11} - H_{21}'H_{22}^{-1}H_{21})^{-1}(J_{11} - H_{21}'H_{22}^{-1}J_{21}). \tag{17}$$

It may seem that, in (17), $H_{22}^{-1}J_{21}$ and $H_{22}^{-1}H_{21}$ are expensive operations, since they involve inverting $(2n + p + 2) \times (2n + p + 2)$ dimensional matrix. However, the inverse of H_{22} is not needed explicitly. Note that $H_{22}^{-1}J_{21}$ is a matrix with two columns, containing solutions of system (13). It can be computed once and stored. The second operation is a solution for the mixed model given in (13) but for a different right hand side, namely H_{21}. Note that in this case the RHS of the system is two-dimensional.

Standard errors. To compute the variance-covariance matrix of the estimated fixed and random effects in a mixed model we need to invert the LHS matrix of system (2). Standard errors are equal to the square roots of the diagonal elements of that matrix. In penalized least squares notation, we need to invert LHS of system (2) and multiply diagonal elements by σ^2. In our case we are interested only in the inference for SNP effects. They are the upper-left part of the expression

$$\hat{\sigma}^2\begin{pmatrix} H_{11} & H_{21}' \\ H_{21} & H_{22} \end{pmatrix}^{-1}. \tag{18}$$

Using the formula for the matrix inverse in block form, the standard errors of β^{SNP} are given by

$$\hat{\sigma}\sqrt{(\mathrm{diag}\,(H_{11} - H_{21}'H_{22}^{-1}H_{21})^{-1})}. \tag{19}$$

Note that this diagonal has already been computed in (17) showing that the computation of the standard errors is trivial.

Missing phenotype. Mixed models handle unbalanced data with ease; all subjects, whatever their number of observations, are taken into the analysis. In this sense the concept of missing data in case of mixed models does not exist. However, our algorithm assumes that the phenotype data for every subject consists of k rows and that some of the values are missing (coded as NA). To properly estimate the solution of a mixed model the weighting matrix has to be introduced

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + P \end{pmatrix}\begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'Wy \\ Z'Wy \end{pmatrix}. \tag{20}$$

Matrix W is a diagonal $nk \times nk$ matrix with 0 or 1 in the diagonal indicating if the observation is valid or not. Note that in practice matrix W does not have to be build, since applying weights is equivalent to replacing rows with missing data by all zeros.

Implementation. GALLOP is implemented in one relatively short R program, provided in the Supplementary Materials. An important computing challenge was to avoid repeating each SNP value k times, to be able to calculate cross products in the border matrices. We achieved this by storing the basis of those matrices, calculated using Kronecker products, in a vector instead of a matrix. This way we can summarize the SNP state directly with two numbers per individual, regardless of k.

Data availability. The BMD data are a part of Rotterdam Study and are confidential. The scripts used to generate the data in the simulation study are available from the corresponding author upon request.

References

1. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation. *The American Journal of Human Genetics* **91**, 1011–1021 (2012).
2. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).
3. Shin, J. & Lee, C. A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics* **105**, 191–196 (2015).
4. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010).
5. Lippert, C. *et al*. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).
6. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).
7. Sikorska, K. *et al*. Fast linear mixed model computations for genome-wide association studies with longitudinal data. *Statistics in Medicine* **32**, 165–180 (2013).
8. Sikorska, K., Lesaffre, E., Groenen, P. F. J. & Eilers, P. H. C. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics* **14**, 166 (2013).
9. Hofman, A. *et al*. The Rotterdam Study: 2010 objectives and design update. *European Journal of Epidemiology* **24**, 553–572 (2009).
10. Eu-Ahsunthornwattana, J. *et al*. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet* **10**, e1004445 (2014).
11. Chung, W. & Zou, F. *Mixed-effects models for GAW18 longitudinal blood pressure data* In *BMC Proceedings* **8**, 1 (2014)
12. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 963–974 (1982).
13. Sikorska, K. *et al*. GWAS with longitudinal phenotypes: performance of approximate procedures. *European Journal of Human Genetics* **23**, 1384–1391 (2015).

Author Contributions

K.S. and P.E. developed the algorithm, implemented it and wrote the first draft of the manuscript. P.G. improved the algorithm. F.R. provided the BMD data. E.L., P.G. and F.R. revised the manuscript. K.S. analyzed the BMD data. P.E. supervised the whole project.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-24578-7.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.