SOFTWARE TOOL ARTICLE

High-Performance Mixed Models Based Genome-Wide Association Analysis with omicABEL software [v1; ref status: indexed, http://f1000r.es/40b]

Diego Fabregat-Traver[1], Sodbo Zh. Sharapov[2,3], Caroline Hayward[4], Igor Rudan[5,6], Harry Campbell[5], Yurii Aulchenko[2,3,5], Paolo Bientinesi[1]

[1]Aachen Institute for Advanced Study in Computational Engineering Science, Aachen, 52062, Germany
[2]Institute of Cytology and Genetics, Siberian Division of the Russian Academy of Sciences, Novosibirsk, 630090, Russian Federation
[3]Novosibirsk State University, Novosibirsk, 630090, Russian Federation
[4]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
[5]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, EH8 9AG, UK
[6]Split University, Split, 21000, Croatia

v1 **First published:** 20 Aug 2014, **3**:200 (doi: 10.12688/f1000research.4867.1)
Latest published: 20 Aug 2014, **3**:200 (doi: 10.12688/f1000research.4867.1)

Abstract

To raise the power of genome-wide association studies (GWAS) and avoid false-positive results in structured populations, one can rely on mixed model based tests. When large samples are used, and when multiple traits are to be studied in the 'omics' context, this approach becomes computationally challenging. Here we consider the problem of mixed-model based GWAS for arbitrary number of traits, and demonstrate that for the analysis of single-trait and multiple-trait scenarios different computational algorithms are optimal. We implement these optimal algorithms in a high-performance computing framework that uses state-of-the-art linear algebra kernels, incorporates optimizations, and avoids redundant computations,
increasing throughput while reducing memory usage and energy consumption. We show that, compared to existing libraries, our algorithms and software achieve considerable speed-ups. The OmicABEL software described in this manuscript is available under the GNU
GPL v. 3 license as part of the GenABEL project for statistical genomics at http: //www.genabel.org/packages/OmicABEL.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees 1	2	3
version 1 published 20 Aug 2014	? report	✓ report	✓ report

1 **Dirk Jan de Koning**, Swedish University of Agricultural Sciences Sweden

2 **Peter Holmans**, Cardiff University UK

3 **Bertram Muller-Myhsok**, Max Planck Institute for Psychiatry Germany

Discuss this article

Comments (0)

Corresponding author: Yurii Aulchenko (yurii.aulchenko@gmail.com)

How to cite this article: Fabregat-Traver D, Sharapov SZ, Hayward C *et al.* **High-Performance Mixed Models Based Genome-Wide Association Analysis with omicABEL software [v1; ref status: indexed, http://f1000r.es/40b]** *F1000Research* 2014, **3**:200 (doi: 10.12688/f1000research.4867.1)

Grant information: PB and DFT gratefully acknowledge the support received from the Deutsche Forschungsgemeinschaft (German Research Association) through grant GSC 111. The VIS study in the Croatian island of Vis was supported through the Core grant from the Medical Research Council UK and a grant of the Ministry of Science, Education, and Sport of the Republic of Croatia (number 108- 1080315-0302). EUROSPAN (European Special Populations Research Network) was supported by the European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947). The work of SZhS was funded by a joint grant from RFBR and the Helmholtz society (Joint Research Groups, 12-04-91322). The work of YA was supported by grant from the Russian Science Foundation (RSCF, grant no. 14-14-00313). The work of IR, CH and HC has received funding from the European Union's Seventh Framework Programme (FP7-Health) under the grant agreement no. 305280 (MIMOmics).

Competing interests: No competing interests were disclosed.

First published: 20 Aug 2014, **3**:200 (doi: 10.12688/f1000research.4867.1)
First indexed: 11 Feb 2015, **3**:200 (doi: 10.12688/f1000research.4867.1)

Introduction

Current biomedical research is experiencing a large boost in the amount of data generated. In particular, investigations involve human cohorts comprising hundreds of thousands of participants as part of nation-wide biobanking initiatives; furthermore, by using both arrays that include hundreds of thousands of single nucleotide polymorphisms (SNPs), and more recently, exome and whole-genome re-sequencing, the genomes of these participants are being characterized at an increasing level of detail, bringing the number of features assessed to tens of millions. At the same time, technologies for high-throughput assessment of different molecular "omics" phenotypes in large study cohorts are becoming more and more affordable. These molecular phenotypes characterize different classes and sub-classes of biological molecules, their functional modifications and relationships. Examples include hundreds of thousands of epigenetic modifications (epigenome[1]), levels of tens of thousands of transcripts (transcriptome[2,3]), metabolites (metabonome or metabolome[4,5]), glycans (glyco(protein)ome[6,7]), and proteins (proteome[8]). The evolution of current molecular techniques expands our capacity to access different components in the omics space, and new prominent omics emerge (e.g. cellomics, interactomics, activomics). The study of the genetic control of different omics brings the promise of new fundamental and applied biological discoveries; however, such analyses pose "big data" challenges.

Genome-Wide Association Studies (GWAS) is an established tool for analyzing the genetic control of complex traits[9]. In GWAS, the association between millions of genetic markers (usually SNPs) and phenotype(s) of interest is studied, with significant associations highlighting the genomic regions harboring the functional variants involved in the control of the trait. While initially GWAS were mostly used to study common diseases, with the rising availability and affordability of omics phenotypes, this methodology is now also applied to investigate the omics space[6,7,10–13], providing important insights into both the mechanisms underlying the genetic regulation of particular biological systems, and the determinants of human health and disease[14–16].

In this work, we address the computational challenges posed by big-data GWAS. These challenges arise when size of sample under analysis is very large or when (potentially hundreds of) thousands of omics phenotypes are studied. We consider analyses facilitated by the use of linear mixed models (LMMs)[17,18], which allow for modeling of correlations between phenotypes of relatives. The LMMs are among the most flexible and powerful methods to account for the genetic (sub)structure that inevitably occurs even in carefully designed large population-based studies. However, the increase in power and precision achieved through the use of mixed models comes with considerable costs in terms of computing time.

Recent advances in GWAS using mixed models[19–25] represent a breakthrough compared to older methods, and allow analyses of a limited number of traits in reasonably sized samples even on personal computers. Still, current algorithms and software may be prohibitively expensive for analysis of large samples when dealing with omics data, since the time needed for a multi-trait analysis is essentially that of a one-trait study multiplied by the number of traits. Under this scenario, the analysis of even relatively small samples sizes leads to extremely long wait times. Therefore at the moment, the LMM-based GWAS analysis of large cohorts (tens to hundreds of thousands of participants), and even small (thousands of participants) studies involving omics measurements, represents a considerable problem. This limitation compromises the analyses availability, the data-to-knowledge turnaround time, and leads to excessive energy spending.

With this work, we aimed to address the aforementioned problems for big-data LMM-based GWAS. To do so, we took advantage of properties specific to the LMM formulation of GWAS, and analyzed a number of possible algorithms applicable to the analysis of large data. By combining sophisticated linear algebra and optimization techniques, we produced a fast and scalable software. Our software facilitates GWAS of tens of thousands of samples and hundreds of thousands of omics phenotypes, without the need for super-computing facilities.

Methods

Linear Mixed Models for GWAS

In a nutshell, LMM models the phenotypes of a group of n studied individuals as a point in an n-dimensional space, which comes from a multivariate Normal distribution. The expected mean is modeled using a standard regression model as $E[Y] = X\beta$, where X is the design matrix which includes the genotypes of interest and other covariates, and β are fixed effects. The variance-covariance matrix is defined as $M = \sigma^2 \cdot (h^2\Phi + (1 - h^2)I)$; here, σ^2 is the total variance of the trait, h^2 is the heritability coefficient, I is the identity matrix, and Φ is the matrix containing the relationship coefficients for all pairs of studied individuals. GWAS are performed by consecutively including SNPs in the analysis model (usually one SNP at a time) and computing the association statistics for the included SNP, thus iteratively applying the model throughout the genome.

The statistical model considered in this work is the same as that outlined in previous works[19–21], and proceeds with analysis in two steps: for each trait considered, we first estimate the matrix of (co)variances between phenotypes, and then we use it when estimating the SNP effects (see Supplementary Note S1 for mathematical details).

Figure 1 illustrates how a multi-trait analysis consists of a series of t separate single-trait analyses, each of which, in turn, consists of a series of m Generalized Least-Squares (GLS) problems. The key to fast analysis algorithms is the realization that such problems are correlated, both along the m and the t direction; in big-data GWAS, any approach that ignores such correlations cannot be feasible in terms of time-to-solution.

Efficient algorithms for single-trait and multi-trait GWAS

Aiming at supporting computational scientists in the design of efficient software, two of the authors recently developed CLAK, an algebraic system that replicates the reasoning of human experts for the automatic discovery of linear algebra solvers[26]. The core idea is to first decompose a target matrix-based problem in terms of library-supported kernels, and then apply algorithmic and algebraic

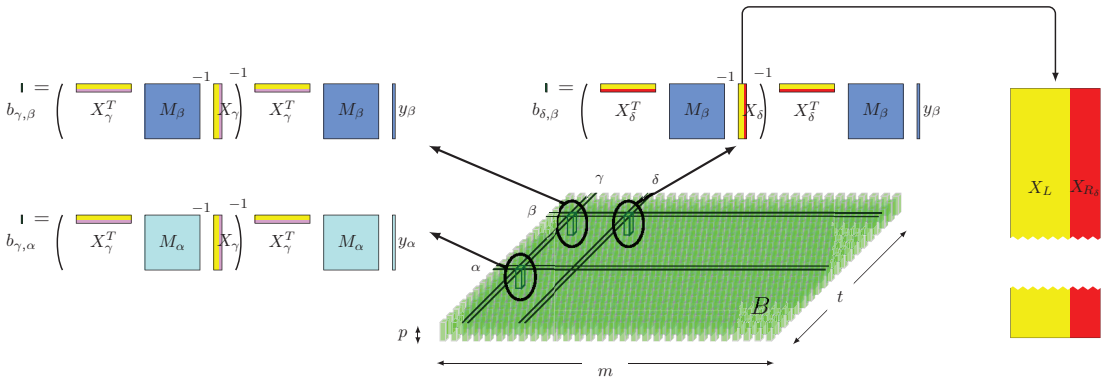

Figure 1. Interpretation of multi-trait GWAS as a two-dimensional grid of generalized least-squares problems ($b := (X^T M^{-1} X)^{-1} X^T M^{-1} y$). GWAS with multiple phenotypes requires the solution of $m \times t$ correlated Generalized Least-Squares (GLS) problems, originating a three-dimensional object B of size $m \times t \times p$. Along the t direction, the variance-covariance matrix M and the phenotype y vary, while the design matrix X does not; conversely, in the m direction, M and y are fixed while X varies. Specifically, X can be viewed as consisting of two parts, X_L and X_R, where the former is constant across the entire grid and the latter changes along m. The figure also captures GWAS with single phenotype, in which case the dimension t reduces to 1.

optimizations. Since the decomposition is not unique, CLAK returns not one but a family of possible solutions, all mathematically equivalent, but exhibiting different space and time complexity.

With the help of CLAK, we generated many solutions to perform the aforementioned GWAS analyses (for a representative list, see Table S2). Each solution was subjected to the analysis of its computational complexity, expressing the cost in terms of the number of samples, markers, and traits in question. Interestingly, depending on the number of traits, the best theoretical performance was attained by two different solutions, which we named CLAK-Chol and CLAK-Eig (described in Table S1), respectively. Figure 2 shows the surfaces representing the time complexity of CLAK-Chol and CLAK-Eig as a function of the number of traits and markers analyzed; the solid curve denotes the crossover between the surfaces. When fewer than four traits are considered, CLAK-Chol attains better theoretical performance; on the contrary, for a higher number of traits, CLAK-Eig is expected to perform better (see also Table 1, and Supplementary Methods).

The idea underlying CLAK-Chol is to linearly transform the input data to de-correlate the observations. To this end, the variance-covariance matrix M is formed explicitly, and its triangular Cholesky factor L is computed ($LL^T = M$); through this factor, each SNP X and each trait y is then linearly transformed, giving raise to a sequence of Ordinary Least Squares problems of the form $b := (X^T X)^{-1} X^T y$. While such problems are solvable with standard techniques, CLAK-Chol takes advantage of the fact that the covariates are fixed for

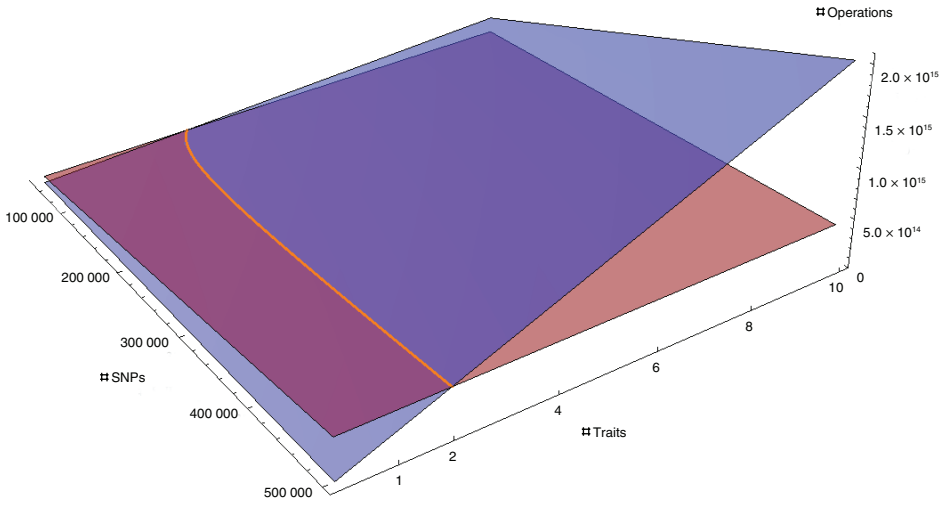

Figure 2. Cost analysis for CLAK-Eig and CLAK-Chol. The brown and blue surfaces indicate the number of operations performed by CLAK-Eig and CLAK-Chol, respectively, for a given number of SNPs and traits. The crossover curve suggests that for analyses with more than just a handful of traits, CLAK-Eig is the fastest algorithm.

Table 1. Computational costs for the solution of single-trait and multiple-trait analyses. The variables *n*, *m* and *t* denote the sample size, the number of genetic markers, and the number of traits, respectively. *v* is the average number of iterations necessary to estimate the model parameters σ^2 and h^2 (see "Time complexity" in Supplementary Note S1).

Algorithm	Estimation of σ^2 and h^2	Single-trait analysis	Multi-trait analysis ($t > n$)	Multi-trait analysis ($n > t$)
CLAK-CHOL	$O(n^3 + tvn)$	**$O(mpn^2)$**	$O(tmn^2)$	$O(tmn^2)$
CLAK-EIG	$O(n^3 + tvn)$	$O(mpn^2)$	**$O(tmn)$**	**$O(mn^2)$**
FaST-LMM	$O(n^3 + tvn)$	$O(mpn^2)$	$O(tmn^2)$	$O(tmn^2)$
GWFGLS	$O(n^3 + tvn)$	$O(mpn^2)$	$O(tmn^2)$	$O(tmn^2)$
EMMAX	$O(n^3 + tvn)$	$O(mpn^2)$	$O(tmn^2)$	$O(tmn^2)$

all SNPs, hence lowering the computational complexity. In single-trait analyses, it is also possible to exploit the fact that the matrix M is symmetric and positive definite. Although asymptotically the Cholesky factorization is equivalent to an eigen-decomposition, in practice it requires 10 times fewer operations. Moreover, instead of using the eigenvectors (a full matrix) to rotate the SNPs, the Cholesky factor (a triangular matrix) allows us to transform them at half the cost. For more details, see the work we have previously published[27] describing the CLAK-CHOL algorithm.

The design of CLAK-EIG is based on three insights: 1) For a given study, the relationship matrix Φ is constant across both SNPs and traits; 2) since the variance-covariance matrix M is built by merely shifting and scaling relationship matrix, its eigenvectors are the same as those of Φ, and its eigenvalues are obtained by shifting and scaling those of Φ; 3) the inverse of M is easily expressed by inverting the diagonal matrix containing its eigenvalues (note that in our solutions we never explicitly invert matrices; we instead factor them, and operate with their factors). Together, these insights suggest that the eigen-decomposition of Φ can be computed once and for all, and most importantly, the eigenvectors of the relationship matrix can be used to rotate all the SNPs and traits only once. After the data is rotated, the computation of the mixed model based GWAS can be carried out by means of a grid of inexpensive Weighted Least Squares problems.

Once the initial eigen-decomposition of Φ is available, the complexity of CLAK-EIG is determined by three operations: the rotation of the SNPs, the rotation of the traits, and the solution of the Weighted Least Squares problems. The dominant term depends on the size of population (n), number of SNPs (m), and number of traits (t). When $n > t$ (or $n > m$), the overall time complexity comes from the rotation of the SNPs (or the traits), and amounts to $O(n^2m)$ (or $O(n^2t)$); if instead both t and m are larger than n, then the dominant term comes from the Least Squares problems, and is linear in population, SNPs and traits: $O(nmt)$ (Table 1). Note that the CLAK-EIG algorithm is a generalization of the eigen-decomposition based algorithms published before (e.g.[22,28]) for a case of multiple trait analysis.

Compared with current state-of-the-art algorithms[19,21,22] in multi-trait analyses, CLAK-EIG achieves a lower computational complexity. As shown in Table 1, there are two scenarios of interest, depending on whether the number of traits is larger than the population size or not. In the first case ($t > n$), which is probably the most typical for current and near-future omics studies, the time complexity of CLAK-EIG is linear on the number of markers, traits, and samples; by contrast, all the other methods have quadratic complexity with the sample size. In the second case ($n > t$), which takes place for smaller 'omics' and also will become more common with the increasing affordability of omics technologies and hence larger sample sizes, the cost of CLAK-EIG is determined by the sample size and the number of SNPs, and its complexity is a factor t lower than other methods.

For both CLAK-EIG and CLAK-CHOL, the space complexity is mainly determined by the square of the sample size; also, a minimum of one trait, one SNP, and the p covariates must reside in memory. In total, our methods only require enough memory to accommodate $n^2+(2+p)n$ entries. If multiple SNPs and/or traits fit in main memory at once, —e.g., dozens or hundreds of them—the computational throughput of our methods improves noticeably. In this case, the space requirement becomes $n^2+(k+p)n$, where k is the number of SNPs and traits resident in memory. As examples, for sample sizes of 10,000, 20,000 and 40,000, the n^2 space requirement translates to 1, 3, and 12 GBs, respectively. More details on space complexity are provided in Supplementary Note S1.

Results

Implementation and comparison

To demonstrate the practical advantages of CLAK-EIG and CLAK-CHOL, we implemented these algorithms in the OMICABEL software package. In doing so, we tailored our implementations to save intermediate results across adjacent problems; we also re-organized the calculations to fully benefit from both the efficiency of highly optimized linear algebra kernels, and the parallelism offered by modern computing platforms.

Since the size of the datasets involved in GWAS is considerably larger than the memory capacity of current processors, input and output data can only be stored in disk devices. Aware that the penalty for accessing information residing on disk is enormous—several orders of magnitude greater than the cost for performing one arithmetic operation—it is imperative to handle these big-data efficiently. By means of asynchronous transfers between memory and disk, our algorithms achieve a perfect overlap of computation and data movement. As long as the relationship matrix fits in the main memory, and regardless of the size of the data sets—both in terms of SNPs and phenotypes—, the processor never idles waiting for a transfer to complete, thus computing at maximum efficiency.

We compared the GWAS run-time of CLAK-CHOL and CLAK-EIG as implemented in OMICABEL with that of several well-established packages: EMMAX[19], FaST-LMM[22] (two-step approximation), and GWFGLS (implementation of the mmscore method of ProbABEL[21] in the MixABEL-package[29]). In the experiments, we considered

three different scenarios, varying one among sample size n, number of SNPs m, and number of traits t, while keeping the other two values constant (Figure 3). A description of the experimental setup is provided in Supplementary Note S2 and Supplementary Note S3.

In the first scenario (single trait and m = 10,000,000 SNPs; the sample size varies from 1,000 to 40,000), all methods exhibit a quadratic behavior, and CLAK-Chol is the only algorithm that completed all tests within 25 hours (Figure 3(a)). For the largest problem considered (sample size n = 40,000), the speed-up over the next-fastest software (FaST-LMM) is 8.3: 205 vs. 25 hours; when n = 1,000, the speed-ups over GWFGLS, FaST-LMM and EMMAX are 24, 32 and 68, respectively: 86, 112 and 240 vs. 3.5 minutes.

The second scenario (single trait and sample size of 10,000; the number of markers varies between 1 and 36 millions) shows a linear dependence on the number of genetic markers for all software packages. Again, CLAK-Chol attains the best timings, outperforming FaST-LMM, GWFGLS and EMMAX by a factor of 11.7, 93.6 and 298, respectively (Figure 3(b)), when the number of analysis SNPs is 36 millions.

Finally, the third scenario illustrates the analysis of multiple phenotypes (sample size of 1,000 and 1,000,000 genetic markers; number of traits varies between 1 and 100,000). The (estimated) time required for these analyses is presented in Figure 3(d). Note that the time scale on this graph is in years. Due to CLAK-Eig's linear time complexity with respect to sample size, SNPs and traits, its advantage becomes most apparent: when thousands of traits are considered, CLAK-Eig outperforms GWFGLS, FaST-LMM and EMMAX by a factor of 1012, 1352, and 2789, respectively (Figure 3(d)), bringing the execution time down from months to hours.

Demonstration of application to real data

We applied the OmicABEL (CLAK-Eig) to study 107,144 metabolomic traits in a sample of 781 people from a genetically isolated population of the Vis island (Croatia). These data are part of the EUROSPAN data set reported in the works of 11, and 12. In short, the data comprise plasma levels of 23 sphingomyelins (SPM), 9 ceramides (CER), 56 phosphatidylcholines (PC), 15 lysophosphatidylcholines (LPC), 27 phosphatidylethanolamines (PE), and 19 PE-based plasmalogens (PLPE). From these data, additional traits were defined by aggregating species into groups with similar characteristics (e.g. unsaturated ceramides), and also by expressing data as molar percentages (instead of absolute concentrations) within classes. Following the standards accepted in genetic analysis of metabolomics data[30], in this work, 328 such measurements served as a base to compute all pair-wise ratios, resulting in 107,584 traits, which were analyzed for association with 266,878 SNPs. More details about the data are provided in Supplementary Note S4.

Previously, it took several weeks to accomplish the original analysis of only few hundreds "original" traits. However, using our OmicABEL software and a computer with 40 cores, we were able to finish the analysis of more than 100,000 traits in only 8 hours.

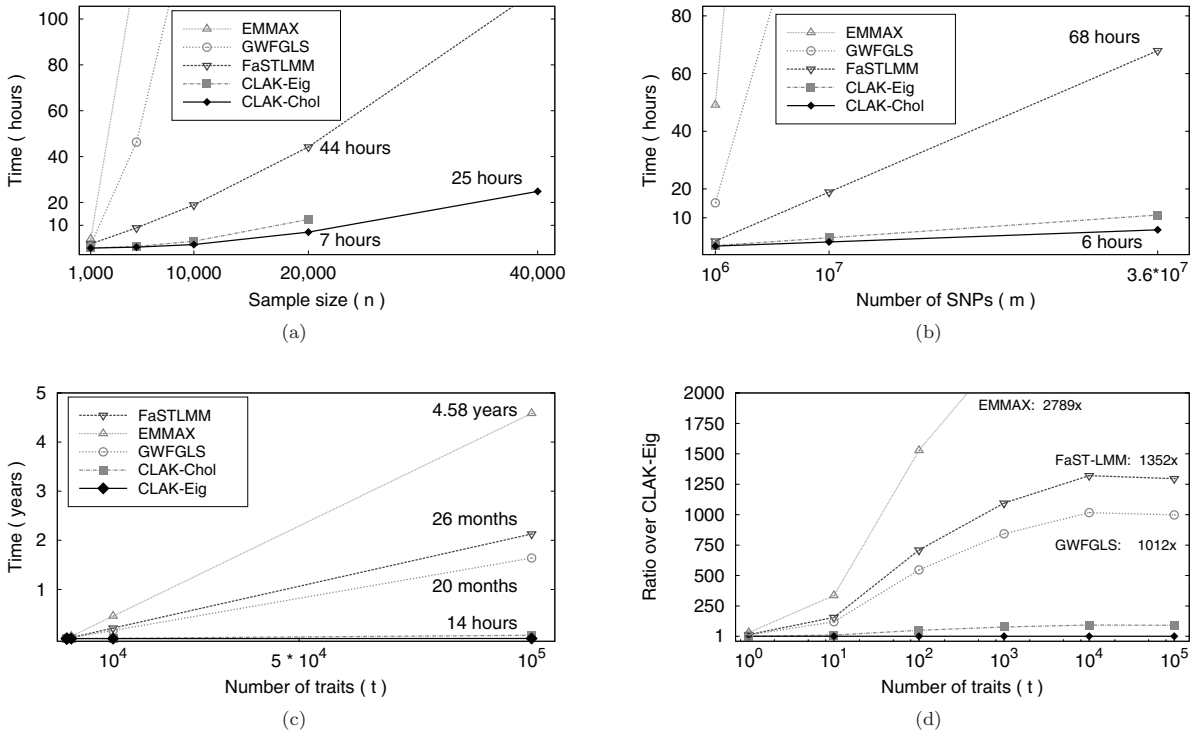

Figure 3. Timings comparison. Panels (**a**) and (**b**) include timings for EMMAX, GWFGLS, FaST-LMM, and our OmicABEL software, for single trait analyses; (**c**) and (**d**) present a comparison of EMMAX, GWFGLS, FaST-LMM, and our OmicABEL software, in the case of multiple traits. In (**a**), the number of SNPs is fixed to $m = 10^7$ and the sample size n ranges from 1,000 to 40,000. In panel (**b**), the sample size is fixed to n = 10,000 and the number of SNPs m ranges between 10^6 and 3.6×10^7. In (**c**) and (**d**), n = 1,000, $m = 10^6$ and t ranges from 1 to 100,000.

Table S4 shows the results for five SNPs previously reported[11] to be significantly associated with levels of circulating sphingolipid concentrations. The best results obtained for these SNPs when using the original and derived traits are reported. The implicated traits were also analyzed by using other approaches: the full likelihood ratio test based LMM[18,22] (as implemented in MixABEL::FMM), Grammar-γ[25] (as implemented in the GenABEL-package[29]), and the two-step approach[19–21] (as implemented in MixABEL::GWFGLS). From Table S4 one can see that our results are consistent with those obtained by other methods. Additionally, Table S3 summarizes the results concerning heritabilities of analyzed traits, while Table S5 lists the Genomic Control inflation factor λ obtained when analyzing the selected traits with different methods. More details of the analysis of this human metabolomics data set are provided in Supplementary Note S4.

Discussion

In contemporary human genomics, methods and tools face the additional challenge posed by the sheer size of the datasets. Big data are produced from the investigation of large human cohorts including hundreds of thousands of participants; these massive samples facilitate the identification of small effects and lead to important biological insights. Large data also come from the field of functional (gen)omics, which aims at establishing the functional roles of genetic variants; hence GWAS are increasingly applied not only to study complex traits in large cohorts, but also to understand the regulation of human and animal transcriptome[15,31,32], metabolome[10–12], glyco(protein)ome[6,7] and other types of omics data. Results of these studies are used to uncover the link between these molecular phenotypes and high-level complex traits, including human diseases[14–16].

In recent years, linear mixed model was accepted as a powerful tool for whole-genome analysis of genetic associations[17,18]. Most current LMM-based methods for GWAS[19–24] exhibit linear dependency of the compute time on the number of genetic polymorphisms and traits studied, but at least quadratic dependency on the sample size. A notable exception from the latter "at-least-quadratic" rule is the GRAMMAR-Gamma method[25] and a method based on low rank approximation of the similarity matrix[22] - with the latter exploiting the ideas similar to EIGENSTRAT approach[33] and the methods assuming the adjustment of the model for top Principal Components of the kinship matrix variation. However, the computational advantage of these methods comes at the cost of mathematical approximation. For example, the GRAMMAR-Gamma method, while extremely fast, and showing excellent results for human studies, is less suited for analyses of samples with uneven genetic structure; adjusting for top principle components (and EIGENSTRAT) is known to provide incomplete correction for stratification in case of complex kinship. Increasing sample sizes and availability of molecular omics phenotypes lead to "big data challenges" and the computational throughput of LMM's starts being more and more of an issue, which at present sometimes cannot be resolved without resorting to supercomputing facilities.

With this work, we address the problem of mixed-model based whole-genome analysis of genetic association for an arbitrary number of traits. We describe the CLAK-Chol and CLAK-Eig algorithms and software tools (OmicABEL package) to address LMM-based GWAS. Specifically, our CLAK-Chol will be useful for investigation of complex traits in very large (tens of thousands of individuals) samples, while CLAK-Eig will be a useful tool for the investigation of genetic control of different omics, potentially including hundreds of thousands or even millions of features.

As for our CLAK-Chol approach, we are not aware of similar, Cholesky-based solutions proposed before. The CLAK-Eig approach behind our solutions bear similarities, and actually reduces to previously suggested methods (e.g.[22,28]) when the number of traits is one. It is also worth mentioning the Matrix eQTL software[34], which, while not implementing the LMM, in many respects exploits the problem-specific properties of multi-trait GWAS in the ways similar to ours.

The key achievement of this research is that it facilitates big-data LMM-based GWAS without supercomputers. For a sample problem with a population of 1,000, three covariates, one million SNPs and 100,000 traits, we estimate that the available methods would require the entire Sequoia supercomputer (equipped with 1.5 million cores)[1] for about 3 minutes; by contrast, using a common 40-core compute node (see Supplementary Methods, Note S2), our method completes within a day and reduces the energy consumption by a factor of 200 (estimated). It should be noted that this impressive speed-up comes at the price of additional assumption of complete data. For many types of omics assays assumption of absence of missing data could be (almost) true, and a small proportion of missing data could be imputed (in the simplest case – replaced with average value) with little negative effect onto statistical properties of the method. However, for the omics assays which produce large proportion of missing data, our CLAK-Eig method in its current formulation and implementation would be inapplicable, unless the missing values could be reliably imputed.

In case of single-trait analysis, our results are somewhat less impressive, and our CLAK-Chol solution outperforms advanced current methods (e.g. FaST-LMM) by about one order of magnitude for large-sample-size problems. It is worth mentioning that the latter speed-up becomes possible because we show that for single-trait GWAS problems our CLAK-Chol algorithm is superior to CLAK-Eig, but other current methods are actually implementing solutions similar to our CLAK-Eig algorithm to address the single-trait GWAS problem.

Further optimizations of our solutions are possible, for instance by exploiting the structure of the kinship matrix. A "compressed MLM" approach was proposed for decreasing the effective sample size of datasets by clustering individuals into groups[20]; similarly, the fast decaying and possibly sparse structure of the kinship matrix can be exploited to lower the number of mathematical operations. Caution must be exercised in the interpretation of findings resulting from GWAS analyses as they may generate false positives if the multiple testing problem is not addressed adequately. A conservative strategy to determine whether an association is statistically significant would be to apply a Bonferronni correction, that is, in our example analysis of 107,144 traits, the conventional genome-wide significance threshold p-value of $5 \cdot 10^{-8}$ should be replaced by $4.7 \cdot 10^{-13}$. This is the common approach applied in the metabolomics studies (see[10,30]). On the other hand, this threshold would probably be too conservative,

given that many of the measurements may be highly correlated. Several methods have been introduced recently[35–37], which may help to overcome this problem; however, this topic lays outside the scope of the current work.

For the analysis of specific omics data, our methods (and software) might require some modifications. For example, in the genetic regulation of human transcriptome, large attention is dedicated to cis-eQTLs, computationally a relatively simple task. In contrast, our implementation is tailored to perform full GWAS for every trait analyzed. While OmicABEL could be used for the identification of trans-eQTLs, one should be aware of the specifics of the analysis of this type of data (e.g. allele-specific expression in RNAseq studies) and a body of methods developed (e.g. methods to account for influences of hidden factors[38,39]).

We foresee that the primary use of our algorithms and software is within the domain of analysis of complex traits in very large samples and for the genetic analysis of "omics" data. However, potentially, there are other uses. The same set of methods and tools can be used for scanning through other omics in e.g. search for biomarkers for a complex trait or in order to determine functional relations between different omics. For example, one may be interested in doing epigenome-wide scans relating the epigenome to a complex trait (or other omics, such as metabolome). Under this use, the genomic inputs would be replaced by epigenomic data. Advanced statistical and machine learning methods, such as penalized regression, can make use of joint analysis of up to several hundreds of thousands of predictors[40]. One of the common scenarios include the construction of millions of features, and their filtering for further joint analysis—a task which can be also effectively addressed by our methods. Finally, our algorithms can be easily extended, e.g. to search for interactions.

Conclusions

We demonstrated that different computational algorithms are optimal for the problems of single- and multi-trait Mixed-Model based GWAS, and implemented these algorithms in a freely available OmicABEL software.

Software availability

Software access

The OmicABEL software implementing computational methods described in this manuscript is available as part of the GenABEL project for statistical genomics at http://www.genabel.org/packages/OmicABEL. The web-page provides the link to OmicABEL tutorial, giving examples of its use.

Latest source code

https://r-forge.r-project.org/scm/viewvc.php/pkg/OmicABEL/?root=genabel

Archived source code as at the time of publication

http://dx.doi.org/10.5281/zenodo.10999[41]

Software license

GNU GPL v. 3

Author contributions

YA and PB jointly designed and supervised the project. DFT and PB developed the methods and algorithms. DFT designed the software, and, together with SZS and YA, analyzed the data. CH, IR and HC provided the data for demonstration analysis. DFT, YA and PB wrote the original version of the paper. All authors contributed to review of the manuscript during its preparation and agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

PB and DFT gratefully acknowledge the support received from the Deutsche Forschungsgemeinschaft (German Research Association) through grant GSC 111. The VIS study in the Croatian island of Vis was supported through the Core grant from the Medical Research Council UK and a grant of the Ministry of Science, Education, and Sport of the Republic of Croatia (number 108-1080315-0302). EUROSPAN (European Special Populations Research Network) was supported by the European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947). The work of SZhS was funded by a joint grant from RFBR and the Helmholtz society (Joint Research Groups, 12-04-91322). The work of YA was supported by grant from the Russian Science Foundation (RSCF, grant no. 14-14-00313). The work of IR, CH and HC has received funding from the European Union's Seventh Framework Programme (FP7-Health) under the grant agreement no. 305280 (MIMOmics).

Acknowledgements

We are grateful to Tatiana Axenovich, Felix Agakov, Xia Shen, and Lars Rönnegård for fruitful discussion of the manuscript, and to Nadezhda Belonogova for help with setting up the analysis. We thank the Center for Computing and Communication at RWTH Aachen for the computing resources.

Supplementary Tables

Table S1. The CLAK-CHOL and CLAK-EIG algorithms for single-trait and multi-trait analyses, respectively.

CLAK-CHOL: Single-trait analysis	CLAK-EIG: Multi-trait analysis
1 $\hat{M} := \hat{\sigma}^2 (\hat{h}^2 \Phi + (1 - \hat{h}^2) I)$	1 $ZWZ^T = \Phi$
2 $LL^T = \hat{M}$	2 $X_L := Z^T X_L$
3 $X_L := L^{-1} X_L$	3 $X_R := Z^T X_R$
4 $X_R := L^{-1} X_R$	4 $Y := Z^T Y$
5 $y := L^{-1} y$	5 **for** j = 1: t
6 $S_{TL} := X_L^T X_L$	6 $D := (\hat{\sigma}_j^2 (\hat{h}_j^2 W + (1 - \hat{h}_j^2) I))^{-1}$
7 $b_T := X_L^T y$	7 $KK^T = D$
8 **for** i = 1:m	8 $Y_j := K^T Y_j$
9 $S := \left(\begin{array}{c\|c} S_{TL} & * \\ \hline X_{R_i}^T X_L & X_{R_i}^T X_{R_i} \end{array} \right)$	9 $W_L := K^T X_L$
10 $b_i := \left(\begin{array}{c} b_T \\ \hline X_{R_i}^T y \end{array} \right)$	10 $S_{TL} := W_L^T W_L$
11 $b_i := S_i^{-1} b_i$	11 $b_T := W_L^T Y_j$
	12 **for** i = 1:m
	13 $W_R := K^T X_{R_i}$
	14 $S := \left(\begin{array}{c\|c} S_{TL} & * \\ \hline W_R^T W_L & W_R^T W_R \end{array} \right)$
	15 $b_{ij} := \left(\begin{array}{c} b_T \\ \hline W_R^T Y_j \end{array} \right)$
	16 $b_{ij} := S^{-1} b_{ij}$

Table S2. Representative list of algorithms for solving a single generalized least-squares problem, as generated by the CLAK expert system.

CLAK-CHOL's for a single GLS problem		CLAK-EIG's for a single GLS problem
1 $\hat{M} := \hat{\sigma}^2 (\hat{h}^2 \Phi + (1 - \hat{h}^2) I)$		1 $ZWZ^T = \Phi$
2 $LL^T = \hat{M}$		2 $D := (\hat{\sigma}^2 (\hat{h}^2 W + (1 - \hat{h}^2) I))^{-1}$
3 $X := L^{-1}X$		3 $KK^T = D$
4 $y := L^{-1}y$		4 $X' := Z^T X$
5 $S := X^T X$		5 $y' := Z^T y$
6 $b := X^T y$		6 $V := K^T X'$
7 $b := S^{-1} b$		7 $v := K^T y'$
		8 $S := V^T V$
		9 $b := V^T v$
		10 $b := S^{-1} b$
Algorithm 7		**Algorithm 13**
1 $\hat{M} := \hat{\sigma}^2 (\hat{h}^2 \Phi + (1 - \hat{h}^2) I)$		1 $ZWZ^T = \Phi$
2 $LL^T = \hat{M}$		2 $D := (\hat{\sigma}^2 (\hat{h}^2 W + (1 - \hat{h}^2) I))^{-1}$
3 $X := L^{-1}X$		3 $KK^T = D$
4 $QR := X$		4 $X' := Z^T X$
5 $y := L^{-1}y$		5 $V := K^T X'$
6 $b := Q^T y$		6 $QR := V$
7 $b := R^{-1} b$		7 $y' := L^{-1}y$
		8 $v := K^T y'$
		9 $b := Q^T v$
		10 $b := R^{-1} b$

Table S3. Regression analysis of heritability estimates in different traits classes. The label "Original" refers to the average heritability of the original traits (intercept of the model); similarly, "cer-cer" refers to the heritability of ratios involving CER, "cer-lpc" to the heritability of ratios involving CER and LPC, and so on.

Trait	Estimate	Std. Error	t value	Pr(> |t|)
Original	0.310493	0.007645	40.613	< 2e-16
cer-cer	0.126283	0.009268	13.625	< 2e-16
cer-lpc	0.009870	0.008183	1.206	0.22777
cer-pc	0.012576	0.007833	1.606	0.10838
cer-pe	-0.016897	0.007982	-2.117	0.03426
cer-pls	-0.008811	0.008176	-1.078	0.28119
cer-spm	0.067942	0.008095	8.393	< 2e-16
lpc-lpc	0.009306	0.008364	1.113	0.26590
lpc-pc	-0.035843	0.007769	-4.614	3.96e-06
lpc-pe	-0.053195	0.007868	-6.761	1.38e-11
lpc-pls	-0.024212	0.008001	-3.026	0.00248
lpc-spm	-0.021573	0.007942	-2.716	0.00660
pc-pc	-0.062170	0.007732	-8.041	9.03e-16
pc-pe	-0.063491	0.007722	-8.222	< 2e-16
pc-pls	-0.059816	0.007767	-7.702	1.36e-14
pc-spm	-0.044365	0.007749	-5.725	1.04e-08
pe-pe	-0.041556	0.007922	-5.246	1.56e-07
pe-pls	-0.056304	0.007864	-7.160	8.13e-13
pe-spm	-0.049549	0.007834	-6.325	2.54e-10
pls-pls	-0.056809	0.008341	-6.811	9.75e-12
pls-spm	-0.039083	0.007941	-4.922	8.60e-07
spm-spm	-0.010163	0.008164	-1.245	0.21319

Table S4. Nominal *p*-values for the analysis of SNPs previously associated with circulating sphingolipid concentrations in Vis data. *Lead trait*: the original trait for which the association signal was most significant in Vis data. *Lead percentage*: the molar percentage, for which the most significant association was obtained. *Lead Ratio*: the ratio, for which the most significant association was obtained. OmicA: OmicABEL, our implementation of CLAK-EIG; FASTA: the two-step approach[19–21] (as implemented in MixABEL::GWFGLS); LRT: the full likelihood ratio test based LMM[18,22] (as implemented in MixABEL::FMM); Gra-γ: Grammar-γ[25] (as implemented in the GenABEL-package[29]).

SNP	Trait	Method			
		OmicA	FASTA	LRT	Gra-γ
	Original traits				
rs10938494	PC 40:6	3.83E-03	3.66E-03	3.71E-03	3.89E-03
rs1000778	PC 38:4	4.07E-06	3.49E-06	3.14E-06	3.39E-06
rs4902242	SPM 14:0	3.42E-13	2.70E-13	1.03E-13	2.75E-13
rs7258249	SPM 18:1	3.93E-08	3.71E-08	2.42E-08	3.97E-08
rs680379	C23:0	9.90E-05	9.53E-05	8.69E-05	9.77E-05
	Molar percentages				
rs10938494	% PC 40:6	8.42E-03	8.25E-03	8.04E-03	8.25E-03
rs1000778	% PC 36:4	2.90E-04	2.92E-04	2.29E-04	2.30E-04
rs4902242	% SPM 14:0	1.32E-13	1.11E-13	4.08E-14	1.19E-13
rs7258249	% SPM 18:1	2.30E-05	2.33E-05	1.92E-05	2.35E-05
rs680379	% C16:0	9.43E-04	8.91E-04	8.61E-04	9.07E-04
	Ratios				
rs10938494	SPM 24:1 / glucosylceramide	2.29E-06	1.72E-06	1.31E-06	1.89E-06
rs1000778	% SPM 23:0 / % PC 36:4	8.83E-09	8.16E-09	6.04E-09	7.64E-09
rs4902242	SPM 14:0 / % SPM 23:0	1.71E-15	1.35E-15	4.44E-16	8.95E-15
rs7258249	SPM 16:1 / SPM 18:1	6.41E-08	5.72E-08	3.74E-08	6.23E-08
rs680379	SPM 16:1-OH / % C16:0	1.56E-07	1.52E-07	1.17E-07	1.60E-07

Table S5. Genomic Control λ estimates for traits analyzed in Table S4 (main text).

Trait	Method			
	OmicA	FASTA	LRT	Gra-γ
Original traits				
PC 40:6	0.995	1.008	0.999	0.989
PC 38:4	0.987	1.000	1.001	1.026
SPM 14:0	0.992	0.999	0.996	0.998
SPM 18:1	0.998	1.005	1.014	0.995
C23:0	0.995	1.004	0.999	0.993
Molar percentages				
% PC 40:6	1.007	1.015	1.006	0.985
% PC 36:4	0.964	0.995	1.006	1.010
% SPM 14:0	0.995	1.005	1.004	0.993
% SPM 18:1	0.998	1.002	1.002	0.994
% C16:0	0.989	1.001	1.008	0.991
Ratios				
SPM 24:1 / glucosylceramide	0.975	1.005	0.998	0.994
% SPM 23:0 / % PC 36:4	1.008	1.007	1.006	1.011
SPM 14:0 / % SPM 23:0	1.006	1.009	0.997	0.994
SPM 16:1 / SPM 18:1	1.008	1.010	0.999	0.994
SPM 16:1-OH / % C16:0	0.997	1.002	1.003	0.993

Supplementary Note S1

Methods and algorithms

Mixed Models for GWAS

The Variance Components model for a quantitative trait can be formulated as

$$Y = X\beta + R,$$

where Y is the vector containing the phenotypes for n individuals, X is the design matrix, and β and R are the vectors of fixed and random effects, respectively. The partitioning $X = [1|L|g]$ indicates that the design matrix is composed of three parts: 1 denotes a column-vector (corresponding to the intercept) containing ones, L is an $n \times p$ matrix corresponding to fixed covariates such as age and sex, and g typically consists of a single column-vector containing genotypes. The vector of random effect R is assumed to be distributed as a Multivariate Normal with mean zero and variance-covariance matrix $M = \sigma^2(h^2\Phi + (1 - h^2)I)$; here, σ^2 is the total variance of the trait, h^2 (in the range [0, 1]) is the heritability coefficient, I is the identity matrix, and Φ is the matrix containing the relationship coefficients for all pairs of studied individuals. The relationship matrix Φ can be estimated from the pedigree or from the genomic data[18].

In GWAS, the quantity of interest is the effect of the genotype, that is, the element(s) of β corresponding to g. Technically, a GWAS with mixed model consists of traversing all measured polymorphic sites in the genome, substituting the corresponding g into X, and fitting the above model; the result is millions of estimates of genetic effect together with their p-values.

One of the most used mixed model-based approaches used in GWAS relies on a two-step analysis methodology[19–22,41]. In the first step, the reduced model (with $X = [1|L]$) is fit to the data, thus obtaining the estimates $\{\hat{\sigma}^2, \hat{h}^2\}$; the variance-covariance matrix corresponding to such estimates is denoted by $\hat{M} = \hat{\sigma}^2(\hat{h}^2\Phi + (1 - \hat{h}^2)I)$. In the second step, for each g_i and corresponding $X_i = [1|L|g_i]$, the estimates of the effects and the variance-covariance matrix are respectively obtained as

$$\hat{\beta}_i = \left(X_i^T \hat{M}^{-1} X_i\right)^{-1} X_i^T \hat{M}^{-1} y, \qquad (1)$$

and

$$Var\left(\hat{\beta}_i\right) = \hat{\sigma}^2 \left(X_i^T \hat{M}^{-1} X_i\right)^{-1},$$

with $i = 1, \ldots, m$, where m is the number of genetic markers considered.

In this work, we consider an extended formulation of this problem to the case of multiple phenotypes, that is, Y is a collection of t vectors, with y_j ($j = 1, \ldots, t$) being a vector corresponding to a specific trait. In this case, trait-specific estimates $\hat{\sigma}_j^2$, $\hat{h}_j^2$ need to be obtained, resulting in t different $\hat{M}_j$s. As the result of the analysis, $m \times t$ vectors of estimates of $\hat{\beta}_{ij}$ and corresponding $Var(\hat{\beta}_{ij})$ are generated. In summary, the problem we are facing is

$$\begin{cases} \hat{\beta}_{ij} = \left(X_i^T \hat{M}_j^{-1} X_i\right)^{-1} X_i^T \hat{M}_j^{-1} y_j \\ Var(\hat{\beta}_{ij}) = \hat{\sigma}_j^2 \left(X_i^T \hat{M}_j^{-1} X_i\right)^{-1} & \text{with } 1 \le i \le m \\ \hat{M}_j = \hat{\sigma}_j^2 \left(\hat{h}_j^2 \Phi + \left(1 - \hat{h}_j^2\right) I\right) & \text{and } 1 \le j \le t; \end{cases} \quad (2)$$

see Figure 1 for a visual description. For each i and j, Equation (2) represents a generalized least-square (GLS) problem. Single-trait and multiple-trait analyses correspond to the solution of $m \times 1$ (1-dimensional) and $m \times t$ (two-dimensional) grids of GLS problems, respectively.

Single-instance algorithms

We provide here an overview of a simplified version of CLAK-Chol and CLAK-Eig to solve a single GLS; the versions tailored for one-dimensional and two-dimensional grids of GLS's are discussed in the Methods section and presented in Table S1. In the following, we use b to indicate $\hat{\beta}$; in both our algorithms, the computation of $Var(b)$ represents an intermediate result towards b.

CLAK-Chol. The approach for CLAK-Chol (Table S2, top-left) is to first reduce the initial GLS $b := (X^T \hat{M}^{-1} X)^{-1} X^T \hat{M}^{-1} y$ to a linear least-squares problem, and then solve this via normal equations. Specifically, the algorithm starts by forming $\hat{M} = \hat{\sigma}^2(\hat{h}^2 \Phi + (1 - \hat{h}^2) I)$, which is known to be symmetric positive definite, and by computing its Cholesky factor L. This leads to the expression $b := (X^T L^{-T} L^{-1} X)^{-1} X^T L^{-T} L^{-1} y$, in which two triangular linear systems can be identified and solved—$\overline{X} := L^{-1} X$ and $\overline{y} := L^{-1} y$ —thus completing the reduction to the standard least-squares problem $b := (\overline{X}^T \overline{X})^{-1} \overline{X}^T \overline{y}$. Numerical considerations allow us to safely rely on the Cholesky factorization of $S := \overline{X}^T \overline{X}$ without incurring instabilities. The algorithm completes by computing $\overline{b} := \overline{X}^T \overline{y}$ and solving the linear system $b := S^{-1} \overline{b}$, for a total cost of $\frac{1}{3} n^3 + O(n^2 p)$.

CLAK-Eig. Instead of forming the matrix $\hat{M}$, Algorithm CLAK-Eig (Table S2, top-right) operates on the matrix Φ: At first, it diagonalizes Φ as ZWZ^T, leading to the expression

$$\hat{M} := \hat{\sigma}^2(\hat{h}^2 ZWZ^T + (1 - \hat{h}^2) I),$$

with diagonal W. By orthogonality of Z, the inverse of $\hat{M}$ can be represented as

$$\hat{M}^{-1} := Z(\hat{\sigma}^2(\hat{h}^2 W + (1 - \hat{h}^2) I))^{-1} Z^T$$

and easily computed with a cost of $O(n)$ operations via

$$D := (\hat{\sigma}^2(\hat{h}^2 W + (1 - \hat{h}^2) I))^{-1};$$

the solution to the GLS is thus given by

$$b := (X^T Z D^{-1} Z^T X)^{-1} X^T Z D^{-1} Z^T y. \quad (3)$$

Moreover, since D is symmetric positive definite, Equation 3 can be rewritten as

$$b := (X^T Z K K^T Z^T X)^{-1} X^T Z K K^T Z^T y,$$

and the algorithm proceeds by computing $V := X^T Z K$ (matrix-matrix multiplication and scaling), obtaining $b := (V V^T)^{-1} V K^T Z^T y$. Similar to CLAK-Chol, b is finally obtained through matrix-vector multiplications and a linear system, for a total cost of $\frac{10}{3} n^3 + O(n^2 p)$.

Effectively solving the grids of Generalized Least-Squares problems

As shown in Table 1, the strength of the algorithms CLAK-Chol and CLAK-Eig becomes apparent in the context of 1D and 2D grids of GLS problems, corresponding to GWAS with single and multiple phenotypes, respectively. The straightforward approach, which is the only alternative provided by current general-purpose numerical libraries, lies in a loop that utilizes the best performing algorithm for a single least-square problem. No matter how optimized the GLS solver, such an approach is prohibitively expensive for either single or multiple phenotypes, due to the unmanageable complexity of $O(mn^3)$ and $O(tmn^3)$, respectively. By contrast, the versions of CLAK-Chol and CLAK-Eig shown in Table S1 are the product of a number of optimizations aimed at saving intermediate results across successive problems, thus avoiding redundant calculations. For instance, the matrix X is logically split as $[X_L | X_R]$, where X_L includes the intercept and the covariates ($[1|L]$), while X_R is the collection of all the genetic markers g_i. Thanks to these savings, both algorithms achieve a lower overall complexity.

Unfortunately, the reduced complexity of algorithms CLAK-Chol and CLAK-Eig is not enough to guarantee high-performance implementations. It is well known that in terms of execution time, the difference between a straightforward translation of an algorithm into code and a carefully assembled routine is of at least one order of magnitude. In other words, the benefits inherent to our new algorithms might go unnoticed unless paired with state-of-the-art implementation techniques. In this following, we detail our strategy to attain high-performance routines.

Both CLAK-Chol and CLAK-Eig are entirely expressed in terms of standard linear algebra operations, such as matrix products and matrix factorizations, provided by the BLAS[42] and LAPACK[43] libraries. Since LAPACK itself is built in terms of BLAS kernels, these are the main responsible for the overall performance of an algorithm. BLAS consists of a relatively small set of highly optimized kernels, organized in three levels (1, 2, and 3), corresponding to vector, matrix-vector, and matrix-matrix operations, respectively.

A common misconception is that all the BLAS kernels, across the three levels, attain a comparable (and high) level of efficiency. Instead, it is only BLAS-3—when operating on large matrices—that fully exploits the processors' potential; as an example, the matrix-vector multiplication, matrix-matrix multiplication on small matrices, and matrix-matrix multiplication on large matrices attain an efficiency of $\approx 5\%$, $\approx 15\%$, and more than 95%, respectively. In this context, the linear systems $X := L^{-1} X$ in CLAK-Chol (Line 3 of the top-left algorithm in Table S2), to be solved for each individual SNP, should ideally be aggregated into a single—very large—linear system $X_R := L^{-1} X_R$, in which X_R is the collection of the genetic

markers of all SNPs (Line 3 of the left Algorithm in Table S1). An analogous comment is valid for CLAK-EIG (see the multiplications at Lines 3 and 4 of the right Algorithm in Table S1), in which the number of markers accessed at once is a user-configurable input parameter.

Since all the current computing platforms include multicore processors, we briefly discuss how to take advantage of this architecture. An effective practice is to invoke a multi-threaded version of the BLAS library. Both in CLAK-CHOL and CLAK-EIG (Table S1) we rely on such a solution for the sections leading up to the outer loop (Lines 1–7 and 1–4, respectively). In the remainder of the algorithms (Lines 8–11 and 5–16), due to the lack of matrix-matrix operations, we instead apply a thread-based parallelization in conjunction with single-threaded BLAS. This mixed use of multi-threading becomes more and more effective as the number of available computing cores increases.

Time complexity

Prior to solving a grid of GLS, the heritability (h^2) and the total variance (σ^2) have to be estimated. For this first step, our algorithms use an approach similar to Software packages such as GenABEL[29], EMMAX, and FaST-LMM: First, the eigen-decomposition of the kinship matrix is performed, for a computational cost of $O(n^3)$. Then, the model parameters are estimated using an iterative procedure based on the Maximum Likelihood (GenABEL, FaST-LMM) or Restricted Maximum Likelihood (EMMAX, CLAK-CHOL, CLAK-EIG) methods, for a cost of $O(tvnp^2)$, where v is the average number of iterations required to reach convergence. In total, the parameter estimation has a complexity of $O(n^3 + tvnp^2)$ operations.

In the second step, a 1D or 2D grid of GLS is solved, corresponding to single and multiple phenotype analysis. For 1D grids, all the considered methods share the same asymptotic time complexity, but the constant factor for CLAK-CHOL is the lowest, yielding a speedup factor of at least 6. In 2D grids, EMMAX, FaST-LMM and GenABEL simply tackle each individual trait independently, one after another, for a total complexity of $O(tmpn^2)$. By exploiting the common structure of the variance-covariance matrix of different phenotypes, CLAK-EIG reduces the complexity by a factor of n, down to $O(tmpn)$. As a result, our algorithm outperforms the other methods by factors higher than 1,000.

Space requirement

CLAK-CHOL. To form and factor the variance matrix, the algorithm uses n^2 memory (Lines 1–2 in the left Algorithm of Table S1). The overall space requirement is determined by the triangular solve (Line 4), which necessitates the full L and a portion of X to reside in memory at the same time. This operation is performed in a streaming fashion—operating on k SNPs at the time—and overwriting X. All the other instructions do not require any extra space. In total, CLAK-CHOL uses about $n^2 + kn$ memory.

CLAK-EIG. The initial eigen-decomposition (Line 1 in the right Algorithm of Table S1) needs $2n^2$ memory. The following matrix-matrix multiplications (Lines 2, 3 and 8) overwrite the input matrices and again are performed in a streaming fashion; in terms of space, the analysis is similar to that for the triangular solve in CLAK-CHOL. The remaining instructions do not affect the overall memory usage. In total, the space requirement is $2n^2 + kn$.

If memory is a limiting factor, one can set k to 1 in either algorithm, possibly at the cost of exposed data movement. Moreover, by using a different (slower) eigensolver, the eigenvectors Z can overwrite the input matrix Φ, effectively saving n^2 memory. These considerations indicate that our algorithms are capable of solving GWAS of any size, as long as the kinship matrix, one SNP, and one trait fit in RAM.

Supplementary Note S2

Computing environment

All the computing tests were run on a symmetric multiprocessor system consisting of four Intel Xeon E7-4850 10-core processors, operating at a frequency of 2.00 GHz. The system is equipped with 256GB of RAM and 4TB of disk as secondary memory[2]. The routines were compiled with the GNU C Compiler (gcc v4.4.5) and linked to Intel's MKL multi-threaded library (v12.1). For double-buffering, our out-of-core routines make use of the AIO library, one of the standard libraries on UNIX systems. The available multi-core parallelism is exploited through MKL's multi-threaded BLAS and OpenMP's pragma directives.

Supplementary Note S3

Simulated data

A data set for GWAS can be characterized by the number of individuals in the sample (n), the number of measured and/or imputed SNPs (m) to be tested, the number of outcomes to be analyzed (t), and the number of covariates (p) to be included in the model. In current GWAS, a typical scenario consists of a few covariates (for example, two, such as sex and age), 10^5–10^7 SNPs, and thousands or tens of thousands individuals. In our experiments, we assumed that the number of covariates is two (p=2), and we varied the three other characteristics (n, m, t) of the data set, leading to these three scenarios.

(A) The number of SNPs was fixed to m = 10,000,000 and one single outcome (t = 1) was studied. As sample size, we used 1,000, 5,000, 10,000, 20,000, and 40,000. The latter test represents a scenario with large number of individuals.

(B) The number of individuals was fixed to n = 10,000 and one single outcome (t = 1) was studied. The number of markers m to be analyzed was set to 1,000,000, 10,000,000, and 36,000,000. The latter test is a scenario that represents a whole genome re-sequencing.

(C) The number of individuals and markers were fixed to n = 1,000 and m = 1,000,000, respectively. The number of outcomes t studied varied (1, 10, 100, 1,000, 10,000, and 100,000), corresponding to an Omics analysis.

For testing purposes, we generated artificial data sets which met pre-specified values of t, m, p, and n.

Supplementary Note S4

Application to human metabolomics data

CROATIA-Vis study

The CROATIA-Vis study includes 1056 unselected Croatians, aged 18–93 years, who were recruited into the study during 2003 and 2004 from the villages of Vis and Komiza on the Dalmatian island of Vis[44,45]. The settlements on Vis island have unique population histories and have preserved their isolation from other villages and from the outside world for centuries. Participants were phenotyped for more than 450 disease-related quantitative traits. Biochemical and physiological measurements were performed, detailed genealogies reconstructed, questionnaire of lifestyle and environmental exposures collected, and blood samples and lymphocytes extracted and stored for further analyses. Samples in all studies were taken in the fasting state. The CROATIA-Vis study was approved by the ethics committee of the medical faculty in Zagreb and the Multi-Centre Research Ethics Committee for Scotland.

Metabolomics measurements

We applied the OmicABEL to study 107,144 metabolomic traits in a sample of 781 people from a genetically isolated population of the Vis island (Croatia). These data are a part of the EUROSPAN data set reported in work of Hicks[11] and Demirkan and colleagues[12]. In short, the data comprise plasma levels of 23 sphingomyelins (SPM), 9 ceramides (CER), 56 phosphatidylcholines (PC), 15 lysophosphatidylcholines (LPC), 27 phosphatidylethanolamines (PE), and 19 PE-based plasmalogens (PLPE). From these data, additional traits were then defined by aggregating species into groups with similar characteristics (e.g. unsaturated ceramides), and also by expressing data as molar percentages (instead of absolute concentrations) within classes. In this work, 328 such measurements served as a base to compute all pair-wise ratios, resulting in 107,584 traits. The traits with >95% of measured values without ties were sex- and age-adjusted and then Gaussianized using the quantile normalization. The resulting 107,144 traits were subject to GWAS with 266,878 SNPs.

Here we report several findings showing the power of this approach to large-scale hypothesis-free analysis; while such an analysis was hardly conceivable before, is now within the reach for most researchers by using our new methods, algorithms and software.

Results

In total, heritability was >0 for 102,985 (96.1%) traits, with an average heritability of 27%, and a maximum of 85%. We investigated if there were systematic differences in heritabilities between different classes of measurements. The average heritability of the original traits was 0.31 (see Table S3). We observed that heritabilities were specific for different classes of ratios. For example, the ratios involving ceramides were more heritable than the "original" traits, especially for the ratios CER:CER ($h^2_{CER:CER} = 0.44$ vs. $h^2_{orginal} = 0.31$, $p < 10^{-15}$) and CER:SPM ($h^2_{CER:SPM} = 0.38$, $p < 10^{-15}$). All ratios involving LPC, PC, PE, PLS, and SPM had significantly lower (all $p < 0.01$) heritabilities, except for the cases when CER's were involved in the ratio or, in the biologically plausible case in which the ratios were derived within the same lipids class (i.e. LPC:LPC, PE:PE, etc.).

To this end, we re-analyzed the associations for five SNPs previously reported[11] to be significantly associated with levels of circulating sphingolipid concentrations. Table S4 reports the best results obtained for these SNPs when using the original and derived traits. The traits implicated were also analyzed by using the full likelihood ratio test based LMM (MixABEL::FMM), Grammar-γ[25], and MixABEL::GWFGLS (which is very similar to OmicABEL) methods. All p-value were corrected for the Genomic Control inflation factor (presented in Table S5). From Table S4 one can see that our results are consistent with these obtained by other methods.

Note that we have analyzed only a subset of the data reported by Hicks *et al.*[11], and therefore our lead concentrations did not always come from the class reported in previous work (e.g. *ATP10D* was reported to be associated with glucosylceramides and *FADS* with SPM).

Notes

[1]As of January of 2014, Sequoia is the third fastest supercomputer in the world: http://www.top500.org/lists/2013/11/.

[2]Recall that our algorithms do not require large amounts of available RAM, as long as it accommodates the kinship matrix, the algorithms will complete.

References

1. Flintoft L: **Human epigenomics: Putting epigenetic variation on the map.** *Nat Rev Genet.* 2009; **10**: 663–663.
Publisher Full Text
2. de Koning DJ, Haley CS: **Genetical genomics in humans and model organisms.** *Trends Genet.* 2005; **21**(7): 377–381.
PubMed Abstract | Publisher Full Text
3. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet.* 2009; **10**(1): 57–63.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Nicholson JK, Lindon JC, Holmes E: **'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data.** *Xenobiotica.* 1999; **29**(11): 1181–1189.
PubMed Abstract | Publisher Full Text
5. Raamsdonk LM, Teusink B, Broadhurst D, *et al.*: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.** *Nat Biotechnol.* 2001; **19**(1): 45–50.
PubMed Abstract | Publisher Full Text
6. Lauc G, Essafi A, Huffman JE, *et al.*: **Genomics meets glycomics: the first GWAS study of human N-glycome identifies HNF1α as a master regulator of plasma protein fucosylation.** *PLoS Genet.* 2010; **6**: e1001256.
Publisher Full Text
7. Lauc G, Huffman JE, Pučić M, *et al.*: **Loci associated with N-glycosylation of human immunoglobulin g show pleiotropy with autoimmune diseases and haematological cancers.** *PLoS Genet.* 2013; **9**(1): e1003225.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Altelaar AF, Munoz J, Heck AJ: **Next-generation proteomics: towards an**

integrative view of proteome dynamics. *Nat Rev Genet.* 2013; **14**(1): 35–48.
PubMed Abstract | Publisher Full Text

9. Hindorff LA, Sethupathy P, Junkins HA, *et al.*: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A.* 2009; **106**(23): 9362–9367.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Gieger C, Geistlinger L, Altmaier E, *et al.*: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum.** *PLoS Genet.* 2008; **4**(11): e1000282.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Hicks AA, Pramstaller PP, Johansson A, *et al.*: **Genetic determinants of circulating sphingolipid concentrations in European populations.** *PLoS Genet.* 2009; **5**(10): e1000672.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Demirkan A, van Duijn CM, Ugocsai P, *et al.*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations.** *PLoS Genet.* 2012; **8**(2): e1002490.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Fu J, Wolfs MG, Deelen P, *et al.*: **Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.** *PLoS Genet.* 2012; **8**(1): e1002431.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Cookson W, Liang L, Abecasis G, *et al.*: **Mapping complex disease traits with global gene expression.** *Nat Rev Genet.* 2009; **10**(3): 184–194.
PubMed Abstract | Publisher Full Text

15. Westra HJ, Peters MJ, Esko T, *et al.*: **Systematic identification of trans eQTLs as putative drivers of known disease associations.** *Nat Genet.* 2013; **45**(10): 1238–1243.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Thanabalasingham G, Huffman JE, Kattla JJ, *et al.*: **Mutations in HNF1A result in marked alterations of plasma glycan profile.** *Diabetes.* 2013; **62**(4): 1329–1337.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Yu J, Pressoir G, Briggs WH, *et al.*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet.* 2006; **38**(2): 203–208.
PubMed Abstract | Publisher Full Text

18. Astle W, Balding DJ: **Population structure and cryptic relatedness in genetic association studies.** *Statist Sci.* 2009; **24**(4): 451–471.
Publisher Full Text

19. Kang HM, Sul JH, Service SK, *et al.*: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet.* 2010; **42**(4): 348–354.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Zhang Z, Ersoz E, Lai CQ, *et al.*: **Mixed linear model approach adapted for genome-wide association studies.** *Nat Genet.* 2010; **42**(4): 355–360.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Aulchenko Y, Struchalin M, van Duijn C: **ProbABEL package for genome-wide association analysis of imputed data.** *BMC Bioinformatics.* 2010; **11**: 134.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Lippert C, Listgarten J, Liu Y, *et al.*: **FaST linear mixed models for genome-wide association studies.** *Nat Methods.* 2011; **8**(10): 833–835.
PubMed Abstract | Publisher Full Text

23. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet.* 2012; **44**(7): 821–824.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Segura V, Vilhjalmsson BJ, Platt A, *et al.*: **An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations.** *Nat Genet.* 2012; **44**(7): 825–830.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Svishcheva GR, Axenovich TI, Belonogova NM, *et al.*: **Rapid variance components-based method for whole-genome association analysis.** *Nat Genet.* 2012; **44**(10): 1166–1170.
PubMed Abstract | Publisher Full Text

26. Fabregat-Traver D, Bientinesi P: **Applicationtailored linear algebra algorithms: A search-based approach.** *Int J High Perform Comput Appl.* 2013; **27**(4): 425–438.
Publisher Full Text

27. Fabregat-Traver D, Aulchenko Y, Bientinesi P: **Solving sequences of generalized least-squares problems on multi-threaded architectures.** *Appl Math Comput.* 2014; **234**: 606–617.
Publisher Full Text

28. Astle W, Balding DJ: **Population structure and cryptic relatedness in genetic association studies.** Ph.D. thesis, The University of London. *Statist Sci.* 2009; **24**(4): 451–471.
Publisher Full Text

29. Aulchenko YS, Ripke S, Isaacs A, *et al.*: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics.* 2007; **23**(10): 1294–6.
PubMed Abstract | Publisher Full Text

30. Suhre K, Shin SY, Petersen AK, *et al.*: **Human metabolic individuality in biomedical and pharmaceutical research.** *Nature.* 2011; **477**(7362): 54–60.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Goring HH, Curran JE, Johnson MP, *et al.*: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes.** *Nat Genet.* 2007; **39**(10): 1208–1216.
PubMed Abstract | Publisher Full Text

32. Lonsdale J, Thomas J, Salvatore M, *et al.*: GTEx Consortium. **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet.* 2013; **45**(6): 580–585.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Price AL, Patterson NJ, Plenge RM, *et al.*: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet.* 2006; **38**(8): 904–909.
PubMed Abstract | Publisher Full Text

34. Shabalin AA: **Matrix eQTL: ultra fast eQTL analysis via large matrix operations.** *Bioinformatics.* 2012; **28**(10): 1353–1358.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Conneely KN, Boehnke M: **So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests.** *Am J Hum Genet.* 2007; **81**(6): 1158–1168.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Li MX, Gui HS, Kwan JH, *et al.*: **GATES: a rapid and powerful gene-based association test using extended Simes procedure.** *Am J Hum Genet.* 2011; **88**(3): 283–293.
PubMed Abstract | Publisher Full Text | Free Full Text

37. van der Sluis S, Posthuma D, Dolan CV: **TATES: Efficient multivariate genotype-phenotype analysis for genome-wide association studies.** *PLoS Genet.* 2013; **9**(1): e1003235.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Fusi N, Stegle O, Lawrence ND: **Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies.** *PLoS Comput Biol.* 2012; **8**(1): e1002330.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Fusi N, Lippert C, Borgwardt K, *et al.*: **Detecting regulatory gene-environment interactions with unmeasured environmental factors.** *Bioinformatics.* 2013; **29**(11): 1382–1389.
PubMed Abstract | Publisher Full Text

40. Shen X, Alam M, Fikse F, *et al.*: **A novel generalized ridge regression method for quantitative genetics.** *Genetics.* 2013; **193**(4): 1255–1268.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Fabregat-TRaver D, Sharapov SZ, Hayward C, *et al.*: **OmicABEL software for genome-wide association studies.** *Zenodo.* 2014.
Data Source

42. Chen W, Abecasis G: **Family-based association tests for genomewide association scans.** *Am J Hum Genet.* 2007; **81**(5): 913–926.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Dongarra J, Croz JD, Hammarling S, *et al.*: **A set of level 3 basic linear algebra subprograms.** *ACM Trans Math Softw.* 1990; **16**(1): 1–17.
Publisher Full Text

44. Anderson E, Bai Z, Bischof C, *et al.*: **LAPACK Users' Guide**. Philadelphia, PA: Society for Industrial and Applied Mathematics, third edition. 1999.
Publisher Full Text

45. Vitart V, Biloglav Z, Hayward C, *et al.*: **3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia.** *Eur J Hum Genet.* 2006; **14**(4): 478–487.
PubMed Abstract | Publisher Full Text

46. Rudan I, Marusić A, Janković S, *et al.*: **"10001 dalmatians:" Croatia launches its national biobank.** *Croat Med J.* 2009; **50**(1): 4–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status:   

Version 1

Referee Report 11 February 2015

doi:10.5256/f1000research.5195.r5915

Bertram Muller-Myhsok
Max Planck Institute for Psychiatry, Munich, Germany

This is a very well written report on a very important topic of very high impact in the actual of omics and multi-level omics data, allowing beyond sheer analysis (which in large data sets is very demanding indeed in terms of runtime and memory) many more questions answered also for methods research.

The methods put forward present a very major step forward for this type of analysis. The methodology is very sound and promising. A tool to actually use is provided, which is also not always the case.

The authors are to be commended for this work.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 05 September 2014

doi:10.5256/f1000research.5195.r5916

Peter Holmans
MRC Center for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK

This article introduces a new software tool (omicABEL) for performing rapid association analyses between SNP data and high-dimensional "omics" data (e.g. gene expression) using mixed models. This is an extremely important development for the following reasons:

1. Identification of genetic variants associated with gene expression (eQTLs), methylation (mQTLs) etc is becoming ever more important in understanding the biology of complex genetic disorders.

2. It is desirable, when performing association analyses, to allow for population stratification and/or cryptic relatedness among members of the sample. Mixed models are an effective way of doing this.

3. The high dimensionality of the "omics" data (tens or even hundreds of thousands of measurements) has made the application of mixed model association approaches computationally

prohibitive up to now. In addition to detailing an application of considerable practical utility, the article is also clearly written and mathematically rigorous.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 29 August 2014

doi:10.5256/f1000research.5195.r5981

Dirk Jan de Koning
Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

The manuscript shows a computational approach that enABELs the use of linear mixed mixed models for large pedigrees, large numbers of markers and large numbers of (omics) traits.

While the work is interesting, several things need to be clarified before the work can be of interest to the wider community.

I assume that the manuscript is aimed at the potential users: researchers in genetics with an interest to analyze their 'big data' in a correct fashion and in a repeatable fashion. The manuscript is compact but possibly too compact in places. I would like the following points to be addressed:

1. Accuracy of estimates and comparison of results. The authors make a big deal about comparing the approaches in terms of computing time. They should also be compared systematically in terms of type I error, accuracy of estimates etc. The computational approximations that are made must come at some cost and as researchers we need to know this cost. There are some token comparisons in the supplementary tables but these deal with heritability and P values. Not with the actual SNP effects which are what we want from these analyses. Some scatter plots between different methods may be helpful.

2. We need more detail about the simulations. You simulate different population sizes but you must clarify the family structures and the family complexity. How many generations? All family sizes equal? etc. Also how you simulate marker data in terms of historical population size, ID structure etc. Likewise for correlation structure among the multiple traits.

3. In many places, variables are used before they are defined. p appears in Table 1 but is only clarified later in the text. The tables are currently not readable on their own and the acronyms and variables should be explained within the tables.

In general, the mathematical notation as well as the language in general needs a bit of work. Some sentences are too long with too many clauses, for example CLAK is not defined at first use. X and y are linearly transformed: should you not introduce new variables for the transformed data?

- Some parts of result should be in the Materials and Methods.
- Figure 3C could have a logarithmic Y-axis.
- Figure 3D could omit line for CLAK-Eig and CLAK-Chol is unlabeled.

- The abbreviation in Table S3 cer-cer lpc, pls spm pc pe etc. makes no sense to me. They are different trait classes but still the table adds very little.

The manuscript has the promise to introduce an interesting new computational approach, but we need more details to make up our mind.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.