Ziyatdinov *et al. BMC Bioinformatics* (2018) 19:68
https://doi.org/10.1186/s12859-018-2057-x

BMC Bioinformatics

SOFTWARE Open Access

lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals

Andrey Ziyatdinov[1*], Miquel Vázquez-Santiago[2,3], Helena Brunel[2], Angel Martinez-Perez[2], Hugues Aschard[1,4†] and Jose Manuel Soria[2†]

Abstract

Background: Quantitative trait locus (QTL) mapping in genetic data often involves analysis of correlated observations, which need to be accounted for to avoid false association signals. This is commonly performed by modeling such correlations as random effects in linear mixed models (LMMs). The R package *lme4* is a well-established tool that implements major LMM features using sparse matrix methods; however, it is not fully adapted for QTL mapping association and linkage studies. In particular, two LMM features are lacking in the base version of *lme4*: the definition of random effects by custom covariance matrices; and parameter constraints, which are essential in advanced QTL models. Apart from applications in linkage studies of related individuals, such functionalities are of high interest for association studies in situations where multiple covariance matrices need to be modeled, a scenario not covered by many genome-wide association study (GWAS) software.

Results: To address the aforementioned limitations, we developed a new R package *lme4qtl* as an extension of *lme4*. First, *lme4qtl* contributes new models for genetic studies within a single tool integrated with *lme4* and its companion packages. Second, *lme4qtl* offers a flexible framework for scenarios with multiple levels of relatedness and becomes efficient when covariance matrices are sparse. We showed the value of our package using real family-based data in the Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT2) project.

Conclusions: Our software *lme4qtl* enables QTL mapping models with a versatile structure of random effects and efficient computation for sparse covariances. *lme4qtl* is available at https://github.com/variani/lme4qtl.

Keywords: Linear mixed models, Covariance, Related individuals, GWAS, lme4

Background

Many genetic study designs induce correlations among observations, including, for example, family or cryptic relatedness, shared environments and repeated measurements. The standard statistical approach used in quantitative trait locus (QTL) mapping is linear mixed models (LMMs), which is able to effectively assess and estimate the contribution of an individual genetic locus in the presence of correlated observations [1–4]. However, LMMs are known to be computationally expensive when applied in large-scale data. Indeed, the LMM approach has the cubic computational complexity on the sample size per test [3]. This is a major barrier in today's genome-wide association studies (GWAS), which consist in performing millions of tests in sample size of tens of thousands or more individuals. Therefore, recent methodological developments have been focused on reduction in computational cost [4].

There has been a notable improvement in computation of LMMs with a single genetic random effect. Both population-based [3, 5, 6] and family-based methods [7] use an initial operation on eigendecomposition of the genetic covariance matrix to rotate the data, thereby removing its correlation structure. The computation time drops down to the quadratic complexity on

*Correspondence: ziyatdinov@hsph.harvard.edu
†Equal contributors
[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America
Full list of author information is available at the end of the article

the sample size per test. When LMMs have multiple random effects, the eigendecomposition trick is not applicable and computational speed up can be achieved by tuning the optimization algorithms, for instance, using sparse matrix methods [8] or incorporating Monte Carlo simulations [9].

However, the decrease in computation time comes at the expense of flexibility. In particular, most efficient LMM methods developed for GWAS assume a single random genetic effect in model specification and support simple study designs, for example, prohibiting the analysis of longitudinal panels. We have developed a new *lme4qtl* R package that unlocks the well-established *lme4* framework for QTL mapping analysis. We demonstrate the computational efficiency and versatility of our package through the analysis of real family-based data from the Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT2) project [10]. More specifically, we first performed a standard GWAS, then showed an advanced model of gene-environment interaction [11], and finally estimated the influence of data sparsity on the computation time.

Implementation

Linear mixed models

Consider the following polygenic linear model that describes an outcome y:

$$y = X\beta + Zu + e$$

where n is the number of individuals, $y_{n\times 1}$ is vector of size n, $X_{n\times p}$ and $Z_{n\times n}$ are incidence matrices, p is the number of fixed effects, $\beta_{p\times 1}$ is a vector of fixed effects, $u_{n\times 1}$ is a vector of a random polygenic effect, and $e_{n\times 1}$ is a vector of the residuals errors. The random vectors u and e are assumed to be mutually uncorrelated and multivariate normally distributed, $\mathcal{N}(0, G_{n\times n})$ and $\mathcal{N}(0, R_{n\times n})$. The covariance matrices are parametrized with a few scalar parameters such as $G_{n\times n} = \sigma_g^2 A_{n\times n}$ and $R_{n\times n} = \sigma_e^2 I_{n\times n}$, where A is a genetic additive relationship matrix and I is the identity matrix. In a general case, the model is extended by adding more random effects, for instance, the dominant genetic or shared-environment components.

R packages for linear mixed models

The first group of R packages implement routines to fit linear mixed models as stand-alone programs, for example, the most recent *Gaston* package [12]. The second group of R packages were developed as extensions of the *lme4* R package, including our *lme4qtl* package. Of the many existing *lme4*-based extensions, the closest to *lme4qtl* is the *pedigreemm* R package [13]. Although this package does support analysis of related individuals, the relationships are coded using pedigree annotations rather than custom covariance matrices. Furthermore, the *pedigreemm* package is not able to fit many advanced models in comparison with *lme4qtl* (Additional file 1: Supplementary Note 1).

Implementation of *lme4qtl*

As an extension of the *lme4* R package, *lme4qtl* adopts its features related to model specification, data representation and computation [14]. Briefly, models are specified by a single formula, where grouping factors defining random effects can be nested, partially or fully crossed. Also, underlying computation relies on sparse matrix methods and formulation of a penalized least squares problem, for which many optimizers with box constraints are available. While *lme4* fits linear and generalized linear mixed models by means of `lmer` and `glmer` functions, *lme4qtl* extends them in `relmatLmer` and `relmatGlmer` functions. The new interface has two main additional arguments: `relmat` for covariance matrices of random effects and `vcControl` for restrictions on variance component model parameters. Since the developed `relmatLmer` and `relmatGlmer` functions return output objects of the same class as `lmer` and `glmer`, these outputs can be further used in complement analyses implemented in companion packages of *lme4*, for example, *RLRsim* [15] and *lmerTest* [16] R packages for inference procedures.

We have implemented three features in *lme4qtl* to adapt the mixed model framework of *lme4* for QTL mapping analysis. First, we introduce the positive-definite covariance matrix G into the random effect structure, as described in [13, 17]. Provided that random effects in *lme4* are specified solely by Z matrices, we represent G by its Cholesky decomposition LL^T and applied a substitution $Z^* = ZL$, which takes the G matrix off from the variance of the vector u

$$Var(u) = ZGZ^T = ZLL^TZ^T = Z^*(Z^*)^T$$

Second, we address situations when G is positive semi-definite, which happen if genetic studies include twin pairs [1]. To define the Z^* substitution in this case, we use the eigendecomposition of G. Although G is not of full rank, we take advantage of *lme4*' special representation of covariance matrix in linear mixed model, which is robust to rank deficiency [14, p. 24-25].

Third, we extend the *lme4* interface with an option to specify restrictions on model parameters. Such functionality is necessary in advanced models, for example, for a trait measured in multiple environments (Additional file 1: Supplementary Note 2).

We note that the later two features are available only in *lme4qtl*, but not in other *lme4*-based extensions such as the *pedigreemm* package [13].

Analysis of the GAIT2 data

The sample from the Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT2) project consisted of 935 individuals from 35 extended families, recruited through a proband with idiopathic thrombophilia [10]. We conducted a genome-wide screening of activated partial thromboplastin time (APTT), which is a clinical test used to screen for coagulation-factor deficiencies [18]. The samples were genotyped with a combination of two chips, that resulted in 395,556 single-nucleotide polymorphisms (SNPs) after merging the data. We performed the same quality control pre-processing steps as in the original study: phenotypic values were log-transformed; two fixed effects, age and gender, and two random effects, genetic additive and shared house-hold, were included in the model; individuals with missing phenotype values were removed and all genotypes with a minimum allele frequency below 1% were filtered out, leaving 263,764 genotyped SNPs in 903 individuals available for GWAS. We compared the performances between our package and SOLAR [2, 19], one of the standard tool in family-based QTL mapping analysis.

Results

We considered three models for the analysis of APTT in the GAIT2 data, namely polygenic, SNP-based association and gene-environment interaction.

Before conducting the analysis, we organized trait, age, gender, individual identifier `id`, house-hold identifier `hhid` variables and SNPs as a table `dat`. The additive genetic relatedness matrix was estimated using the pedigree information and stored in a matrix `mat`. A polygenic model `m1` was fitted to the data by the `relmatLmer` function as follows.

```
m1 <- relmatLmer(aptt ~ age + gender
  + (1|id) + (1|hhid), dat,
  relmat = list(id = mat))
```

The proportion of variance explained by the genetic effect (heritability) was `0.56`, and its 95% confidence interval, estimated by profiling the deviance [14], was `[0.45; 0.84]`.

We further tested whether the genetic effect was statistically significant by simulations of the restricted likelihood ratio statistic, as implemented in the `exactRLRT` function of the *RLRsim* R package [15]. The *p*-value of the test was below 2.2×10^{-16}.

For a single SNP named `rs1`, the `update` function created an association model `m2` from `m1` and the `anova` function then performed the likelihood ratio test.

```
m2 <- update(m1, . ~ . + rs1)
anova(m1, m2)
```

To automate the GWAS analysis, we created an example `assocLmer` function with several options such as different tests of association and parallel computation. By using the `assocLmer` function, we have replicated some loci previously reported for APTT in a larger cohort of 9,240 individuals [18] (Additional file 1: Figure S1) applying the likelihood ratio test and running the analysis in parallel on a desktop computer (2.8GHz quad-core Intel Core i5 processor, 8GB RAM).

The GWAS computation time of the association analysis with two random effects by *lme4qtl* was 7.6 h. We performed the same analyses, using SOLAR, and observed a computation time 3 fold larger (25.1 hours, Additional file 1: Table S1). In additional experiments varying the number of fixed and random effects, the *lme4qtl* package was also several times faster than SOLAR (Additional file 1: Table S1, Additional file 1: Figure S2), owing to the efficient *lme4* implementation of sparse matrix methods. Though, in a special case when a model has a single random effect, SOLAR had a option to apply the eigendecomposition trick and substantially speed up the computation (3.8 h), while this option has not been implemented in *lme4qtl* (6.6 h). When including a widely used *lmekin* function from the *coxme* package [20] in the comparison study, our package *lme4qtl* also showed the lowest computation time (Additional file 1: Figure S3). As comparison with other packages is beyond the scope of this work, we suspect that *lme4qtl* will likely outperform others or show similar results under scenario of sparse covariance matrices. We note that the *lme4qtl* performance substantially declines for dense covariance matrices, as described further below.

If one is interested in more complex models than `m1` and `m2`, our package *lme4qtl* is flexible enough for advanced model specification. For instance, *lme4qtl* allows for extension of the polygenic model `m1` to assess the hypothesis of sex-specificity (a special case of gene-environment interaction) [11].

```
m3 <- relmatLmer(aptt ~ age + gender
  + (0 + gender|id)+(0+dummy(gender)|rid),
  dat, relmat = list(id = mat))
```

The first genetic random effect, denoted as `(0 + gender|id)`, has three parameters σ_{g_1}, σ_{g_2} and ρ_g and its variance is partitioned among three groups of pairs: male-specific $\left(\sigma^2_{g_1}\right.$, the genetic variance captured by males$\left.\right)$, female-specific $\left(\sigma^2_{g_2}\right)$ and male-female pairs $\left(\rho_g \sigma_{g_1} \sigma_{g_2}\right)$. The second random effect, denoted as `(0 + dummy(gender)|rid)`, models the heteroscedasticity in residual variance between the two groups of males and females, where the variable `rid` is a copy of the individual identifier `id` variable. The random effect `(1|hhid)` presented in `m1` is not included for simplicity reasons.

Additional file 1: Supplementary Notes 1 and 2 contain the details on model specification and numerical results obtained on the GAIT2 data.

To assess the null hypothesis of no gene-environment interaction, Blangero proposed the likelihood ratio test when comparing to either of two null models: the correlation coefficient is one ($\rho_g = 1$) or the variances are equal ($\sigma_{g_1} = \sigma_{g_2}$) [11]. We implemented different restrictions on model parameters in *lme4qtl* by means of a special syntax for the `vcControl` parameter, as described in Additional file 1: Supplementary Note 2. The next two (null) models, `m4` and `m5`, were fitted with the parameter restrictions described above for the gene-environment interaction analysis.

```
m4 <- relmatLmer(aptt ~ age + gender +
  (0+ gender|id) + (0 + dummy(gender)|rid),
   dat, relmat = list(id = dkin),
   vcControl = list(rho1 = list(id = 3)))

m5 <- relmatLmer(aptt ~ age + gender +
  (0 + gender|id)+(0 + dummy(gender)|rid),
  dat, relmat = list(id = dkin),
  vcControl=list(vareq=list(id=c(1,2,3))))
```

Numerical results of the likelihood ratio tests in Additional file 1: Supplementary Note 3 showed that the evidence for gene-environment interaction is weak. Otherwise, a new `m3`-based association model can be sought for GWAS, in which a SNP has both marginal and interaction effects with the `gender` variable.

Lastly, we evaluated how the *lme4qtl* computation time depends on the sparse structure of covariance matrices, as the genetic relationship matrices are not necessarily sparse. We used the polygenic model `m1` as an initial model (the random effect `(1|hhid)` was omitted), where the genetic relationship matrix `mat` has a high proportion of zero values (sparsity) equal to 0.98. We then gradually fill zeros in `mat` by small non-zero values, thus reducing the sparsity towards 0, and refitted the model `m1`. We found that the time required to fit the polygenic model increased substantially: it became an order of magnitude greater once the sparsity changed from the GAIT2 level 0.98 to 0.60 (Additional file 1: Figure S4).

Discussion and conclusions

We have extended the *lme4* R package, a well-established tool for linear mixed models, for application to QTL mapping. The new *lme4qtl* R package has adopted the *lme4*'s powerful features and contributes with two key building blocks in QTL mapping analysis, custom covariance matrices and restrictions on model parameters. To our knowledge, the *lme4qtl* R package is the most comprehensive extension of *lme4* to date for QTL mapping analysis.

Our package also has limitations. In particular, introducing covariance matrices in random effects implies that some of the statistical procedures implemented in *lme4* might not be applicable anymore. For instance, bootstrapping in the `update` function from *lme4* cannot be directly used for *lme4qtl* models. Furthermore, the residual errors in *lme4* models are only allowed to be independent and identically distributed, and ad hoc solutions need to be applied in more general cases, as we showed for the gene-environment interaction model. However, this restriction on the form of residual errors may be relaxed in the future *lme4* releases, according to its development plan on the official website [21]. Also, *lme4qtl* cannot compete with tools optimized for particular GWAS models with a single genetic random effect: *lme4qtl* allows for association models with multiple random effects.

In practice, *lme4qtl* is mostly applicable to datasets with sparse covariance matrices. Its use in population-based studies with dense matrices may lead to a considerable overhead in computation time. The typical study designs suitable for *lme4qtl* are family-based studies, longitudinal and similar studies with many sparse grouping factors. Also, *lme4qtl* would be applicable in a 2-step GWAS procedure even in population-based studies: at the first step, the linear mixed model is fitted a single time under the null hypothesis of no association; at the second step, association tests make use of the variance component parameters estimated at the previous step, thus, avoiding fitting the linear mixed model again and speeding up the computation [3, 4]. Of a practical note, *lme4qtl* was able to fit a linear mixed model with many structured random effects, including the dense genetic covariance matrix, on several thouthands of individuals in less than half an hour on the desktop computer (data not shown).

In conclusion, the *lme4qtl* R package enables QTL mapping models with a versatile structure of random effects and efficient computation for sparse covariances.

Additional file

Additional file 1: Supplementary Tables and Figures. Supplementary Note 1: R code to compare lme4qtl and pedigreemm R packages. Supplementary Note 2: Multi-trait and multi-environment linear mixed models. Supplementary Note 3: R code applied to the GAIT2 data. (PDF 1341 kb)

Abbreviations

APTT: Activated partial thromboplastin time; GAIT2: Genetic analysis of idiopathic thrombophilia 2; GWAS: Genome-wide association study; LMM: Linear mixed model; QTL: Quantitative trait locus; SNP: Single nucleotide polymorphism

Acknowledgments

AZ thanks Donald Halstead for reading and providing feedback on early drafts of the manuscript.

Funding
This study was supported by funds from the Instituto de Salud Carlos III Fondo de Investigación Sanitaria PI 14/0582. AZ and HA were supported by NIH grant R21HG007687.

Availability of data and materials
Source code of lme4qtl is available at https://github.com/variani/lme4qtl.

Authors' contributions
AZ, MVS and HB conceived the study; AZ implemented the software; AZ, MVS and HB tested the software; AZ, MVS, HB and AMP analyzed the data; HA and JMS directed the study; AZ and HA drafted the manuscript; all authors read and approved the final version of the manuscript.

Ethics approval and consent to participate
The GAIT2 study was performed according to the Declaration of Helsinki and adult subjects gave written informed consent for themselves and for their minor children. The GAIT2 study was reviewed and approved by the Institutional Review Board of the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.

Consent for publication
Not applicable.

Competing interests
The authors have declared that no competing interests exist.

Publisher's Note

Author details
[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America. [2]Unitat de Genòmica de Malalties Complexes, Institut d'Investigació Biomèdica Sant Pau (IIB-Sant Pau), Barcelona, Spain. [3]Unitat d'Hemostàsia i Trombosi, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain. [4]Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France.

Received: 4 August 2017 Accepted: 13 February 2018
Published online: 27 February 2018

References
1. Lynch M, Walsh B, et al. Genetics and analysis of quantitative traits, vol 1. MA: Sinauer Sunderland; 1998.
2. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Human Genet. 1998;62(5):1198–211.
3. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23.
4. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–6.
5. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8(10):833–7.
6. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.
7. Blangero J, Diego VP, Dyer TD, Almeida M, Peralta J, Kent Jr JW, Williams JT, Almasy L, Göring HHH. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. Adv Genet. 2013;81:1.
8. Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D, et al. ASReml user guide release 3.0. UK: VSN International Ltd, Hemel Hempstead; 2009.
9. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, de Candia TR, Lee SH, Wray NR, Kendler KS, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015;47(12):1385.
10. Martin-Fernandez L, Ziyatdinov A, Carrasco M, Millon JA, Martinez-Perez A, Vilalta N, Brunel H, Font M, Hamsten A, Souto JC, et al. Genetic determinants of thrombin generation and their relation to venous thrombosis: results from the GAIT-2 project". PloS ONE. 2016;11(1):e0146922.
11. Blangero J. Statistical genetic approaches to human adaptability. Hum Biol. 2009;81(5):523–46.
12. Perdry H, Dandine-Roulland C. Gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. 2017. https://CRAN.R-project.org/package=gaston. R package version 1.5.
13. Vazquez AI, Bates DM, Rosa GJM, Gianola D, Weigel KA. Technical note: an r package for fitting generalized linear mixed models in animal breeding. J Anim Sci. 2010;88(2):497–504.
14. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48.
15. Scheipl F, Greven S, Kuechenhoff H. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Comput Stat Data Anal. 2008;52(7):3283–99.
16. Kuznetsova A, Bruun Brockhoff P, Haubo Bojesen Christensen R. lmerTest: Tests in Linear Mixed Effects Models. 2016. https://CRAN.R-project.org/package=lmerTest. R package version 2.0-33.
17. Harville DA, Callanan TP. Computational aspects of likelihood-based inference for variance components. In: Advances in statistical methods for genetic improvement of livestock. Springer; 1990. p. 136–76.
18. Weihong Tang, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, Samani NJ, Basu S, Gögele M, Davies G, et al. Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am J Hum Genet. 2012;91(1):152–62.
19. Andrey Ziyatdinov HelenaBrunel Angel Martinez-Perez. Alexandre Perera, and Jose Manuel Soria. solarius: an R interface to SOLAR for variance component analysis in pedigrees. Bioinformatics. 2016;32(12):1901–2.
20. Therneau TM. coxme: Mixed Effects Cox Models. 2015. https://CRAN.R-project.org/package=coxme. R package version 2.2-5.
21. Github. https://github.com/lme4/lme4. Last accessed 27 Jan 2017.