Human Genetics (2020) 139:61–71
https://doi.org/10.1007/s00439-019-02001-z

ORIGINAL INVESTIGATION

OpenMendel: a cooperative programming project for statistical genetics

Hua Zhou[1] · Janet S. Sinsheimer[2] · Douglas M. Bates[3] · Benjamin B. Chu[4] · Christopher A. German[1] · Sarah S. Ji[1] · Kevin L. Keys[5] · Juhyun Kim[1] · Seyoon Ko[6] · Gordon D. Mosher[7] · Jeanette C. Papp[2] · Eric M. Sobel[2] · Jing Zhai[8] · Jin J. Zhou[8] · Kenneth Lange[4]

Received: 26 October 2018 / Accepted: 15 March 2019 / Published online: 26 March 2019

Abstract

Statistical methods for genome-wide association studies (GWAS) continue to improve. However, the increasing volume and variety of genetic and genomic data make computational speed and ease of data manipulation mandatory in future software. In our view, a collaborative effort of statistical geneticists is required to develop open source software targeted to genetic epidemiology. Our attempt to meet this need is called the OpenMendel project (https://openmendel.github.io). It aims to (1) enable interactive and reproducible analyses with informative intermediate results, (2) scale to big data analytics, (3) embrace parallel and distributed computing, (4) adapt to rapid hardware evolution, (5) allow cloud computing, (6) allow integration of varied genetic data types, and (7) foster easy communication between clinicians, geneticists, statisticians, and computer scientists. This article reviews and makes recommendations to the genetic epidemiology community in the context of the OpenMendel project.

Keywords Statistical genomics · GWAS · Computational statistics · Open source · Collaborative programming

Introduction

Genome-wide association studies (GWAS) query the entire genome to identify genetic variants associated with a trait of interest. GWAS have enjoyed many successes (Visscher et al. 2017) and have uncovered many clues to the genetic etiology of common diseases (Cookson et al. 2009). Case–control tests of association between markers and traits predate GWAS by more than 50 years (Aird et al. 1953).

Hua Zhou and Janet S. Sinsheimer authors contributed equally.

NIH Grants R01-GM53275, R01-HG006139, R01-GM105785, R01-HL135156 and T32-HG002536; NSF grant DMS-1052210; the UCSF Bakar Computational Health Sciences Institute; the UC Berkeley Institute for Data Sciences as part of the Moore-Sloan Data Sciences Environment Initiative; and the 2018 Google Summer of Code.

✉ Hua Zhou
huazhou@ucla.edu

✉ Janet S. Sinsheimer
jsinshei@g.ucla.edu

✉ Kenneth Lange
klange@ucla.edu

[1] Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, USA

[2] Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, USA

[3] Department of Statistics, University of Wisconsin, Madison, USA

[4] Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, USA

[5] Department of Medicine, University of California, San Francisco, USA

[6] Department of Statistics, Seoul National University, Seoul, South Korea

[7] Departments of Statistics and Computer Science, University of California, Riverside, USA

[8] Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, USA

However, association studies were rarely undertaken in the pre-GWAS era unless there were candidate genes with strong prior evidence. The situation changed at the turn of the millennium when dense SNP (single nucleotide polymorphism) maps became available and SNP genotyping costs plummeted. Suddenly, it became possible to exploit linkage disequilibrium (LD) and survey hundreds of thousands to millions of genome-wide SNPs. In the subsequent decade, hundreds of associations found by GWAS were published (Visscher et al. 2012, 2017). Genomics is in the midst of a second technological evolution driven by high-throughput sequencing (Metzker 2010; Pickrell et al. 2012; Kilpinen and Barrett 2013; Van Dijk et al. 2014). Geneticists can now survey both rare and common variants.

This sudden expansion of data leads to enormous challenges in statistical genetics. Many current algorithms and programs are ill-adapted to handle modern data sets with 10^5 cases and 10^7 markers. Ever more types of genetic variation are being observed and catalogued (Visscher et al. 2017). These changes demand more complex data structures and data integration across multiple biological scales. Precision health and predictive medicine raise the stakes even further (Kilpinen and Barrett 2013). Concurrently, the nature of computing is rapidly changing. In addition to new hardware, new programming paradigms and new algorithms must be brought online as quickly as possible to sustain progress in statistical genetics.

The following three studies exemplify the variety and magnitude of genomic data sets being collected today: (a) The Million Veteran Program contains GWAS data (657,459 SNPs) on 359,964 veterans (Gaziano et al. 2016). Simply storing the genotypes in compressed format requires > 100 GB. Obviously, this data set and others like it (Sudlow et al. 2015; Brody et al. 2017) will continue to grow. (b) A recent study (Telenti et al. 2016) obtained whole genome sequence (WGS) data on 10,545 humans at 30–40 × coverage for < $2000 per genome. These researchers identified > 150 million variants, the majority of which are rare or *de novo*. (c) The iPOP (integrative personal omics profile) study (Chen et al. 2012) followed a single individual for 401 days and collected transcriptome, proteome, metabolome, microbiome, epigenome, exposome, and phenome data at 20 time points, along with an extremely high coverage WGS. This type of omics profiling yields a dynamic picture of the heteroallelic changes between healthy and diseased states.

Current analysis pipelines juggle a multitude of computer programs that are implemented in different languages, run on different platforms, and require different input/output formats. This heterogeneity unintentionally creates barriers to communication, data exchange, data visualization, and scientific replication. End-users treat the entire pipeline as a black box and often fail to use their biological insight to inform statistical analysis. Students, post-docs, and researchers spend inordinate amounts of time coding and debugging the low-level languages instead of thinking about the science. In addition to these disadvantages, current software packages are straining under the volume, velocity, variety, and veracity of modern genomics data. Many programs do not even run on multiple threads. Distributed computing across different machines is largely ignored. In our view, the time is ripe to put in place a better paradigm for statistical genetics.

In this review, we first explain why the new Julia language (Bezanson et al. 2017) is an ideal choice for the OpenMendel analysis platform. We then present what we see as some of the most pressing needs in gene mapping and our efforts to advance them through the cooperative OpenMendel effort. Owing to page and time constraints, we do not offer encyclopedic coverage of recent advances in GWAS or sequence analysis. Many promising methods are left unmentioned, for example meta-analysis based on summary statistics (Cantor et al. 2010; Chiu et al. 2016; Fan et al. 2016; Kim et al. 2015; Mancuso et al. 2017; Yang et al. 2012) or estimation of fine-scale population structure (Novembre and Peter 2016). Instead we focus on topics related to projects already underway in OpenMendel. These projects include methods for handing SNP data, genotype imputation, rapid GWAS, iterative hard thresholding, kinship comparison, variance component modeling, and SNP-set analyses.

We want to point out that there are other groups who have made notable strides in making genetic analyses accessible to researchers who work with big data but lack the support available at large genomic centers (Bickerstaffe et al. 2017; Ranaweera et al. 2018). Some of these projects are further along than OpenMendel. A particularly interesting example is the PLATO software project [(Hall et al. 2017) and https://ritchielab.org/software/plato-download], which is designed to provide a single platform for a variety of association analyses. A major difference in our approach and PLATO is language choice. This difference might seem minor but, as we outline below, we believe it is a fundamental difference and is important for our goal of getting user-initiated modules and modifications. Another example is the Ark software (Bickerstaffe et al. 2017), which focuses on data management and is complementary to rather than competitive with OpenMendel.

The importance of the Julia computing language

Many compelling features make Julia (Bezanson et al. 2017) an ideal vehicle for implementing methods for modern statistical genetics. First, it is free, open source, and easy to install. Second, its clear, powerful syntax lends itself to compact, readable code and quick algorithm mock-ups. As

needed, it can easily call Fortran, C, R, and Python functions. Third, because Julia incorporates an excellent just-in-time (JIT) compiler, it achieves the efficiency of low-level languages with minimal programming efforts. Fourth, Julia is built for parallelism at the multicore, graphical processing unit (GPU), and cluster levels. Fifth, Julia employs a modern, easy-to-use package management system. Of particular relevance to statistical genetics, Julia has many statistical and numerical analysis packages ready to use. Finally, end-users can run their analyses via the interactive Jupyter (**Ju**lia, **Py**thon, **R**) Notebook, an attractive interface for data visualization and reproducible research. Together these tools constitute an integrated environment for rapid prototyping of new applications and, with the same code, the analysis of large-scale genetic data.

Traditional high-level languages such as R, Matlab, and Python face the notorious two-language problem. In this scenario, one high-level language is used for prototyping, but a second low-level language is later needed for producing fast code for real world, large data sets. The high-level code is typically more compact, readable, and amenable to change, but much slower to execute. Most of the popular statistical genetics analysis tools or their most demanding subroutines are implemented purely in low-level languages, greatly restricting the community that feels comfortable exploring the code. Most tools are also restricted to certain computer platforms and input formats.

Today, a typical analysis pipeline requires a glue language such as Bash, Perl or Python to chain packages together. Data plotting and display require additional software, typically R or Matlab. Current analysis pipelines are cumbersome, opaque, and error-prone, creating barriers to the development of new statistical methods. Researchers wade through a swamp of low-level code and reinvent statistical genetics wheels instead of focusing on their unique contributions. This can be avoided as Julia has solved the two-language problem through careful design of the programming language itself. Julia is both easy to code and scales to peta-flop computing levels (Claster 2017). We can now use Julia in all phases of our methods development, from prototyping to production software. OpenMendel includes many leading-edge statistical genetics methods written in this fast, high-level language that invites easy contributions from scientists. Using Julia, OpenMendel can become the first highly efficient, open source statistical genetics software that can scale to million-subject studies and is both user- and developer-friendly.

Handling SNP data

The SnpArrays.jl module of OpenMendel provides a convenient bridge between binary SNP data and downstream statistical analysis. The VCFTools.jl module achieves the same end for the richer genetic information distributed in VCF and BCF file formats. In SnpArrays.jl, biallelic genotype data are held in BitArrays, which store four genotypes per byte. As much as possible, compressed storage is also maintained during computation. Julia allows operators such as matrix multiplication to be defined directly on BitArrays without decompression. The design features of Julia make it easy to build high-performance statistical genetics software that is scalable to data sets with millions of subjects and tens of millions of SNPs.

The functionality of SnpArrays.jl includes: (1) reading and writing compressed SNP files, (2) computing summary statistics, (3) filtering data by genotyping success rates and other criteria, (4) copying compressed data into numerically oriented vectors and matrices, (5) computing genetic relationship matrices, (6) computing principal components, and (7) extending matrix and vector operations to compressed SNP data. SnpArrays.jl serves as a data interface to other OpenMendel modules.

Genotype imputation

Genotype imputation involves the inference of unobserved genotypes from observed genotypes. It is possible to base inference on the observed genotypes of surrounding pedigree members (Sobel et al. 1996), but pedigree data are now viewed as poor substitutes for linkage disequilibrium. In particular, pedigree data are incapable of imputing genotypes at completely untyped SNPs in a study. Recent versions of genotype imputation rely on panels of reference genotypes and employ hidden Markov models, with the hidden states being underlying haplotype pairs (Howie et al. 2009; Li et al. 2010; Marchini et al. 2007). These programs are computationally intensive and operate by haplotyping individuals on the typed SNPs in the sample. These partial haplotypes are then compared to the reference panel to impute the full set of genotypes (Howie et al. 2012; Van Leeuwen et al. 2015). We have taken an alternative approach based on the generic data mining technique of matrix completion (Candès and Recht 2009; Chi et al. 2013).

Matrix completion fills in the missing entries of an $m \times n$ matrix $\boldsymbol{X} = (x_{ij})$ whose observed entries are indexed by a subset Ω of $\{1, \ldots, m\} \times \{1, \ldots, n\}$. Imputation involves finding a low-rank matrix $\boldsymbol{Y} = (y_{ij})$ consistent with the observed entries of $\boldsymbol{X} = (x_{ij})$. This is done by minimizing the loss function

$$f(\boldsymbol{Y}) = \sum_{(i,j)\in\Omega} (x_{ij} - y_{ij})^2 \tag{1}$$

over the set of matrices $\boldsymbol{Y}$ of rank r or less. Taking r small is a form of parsimony capturing the hidden structure of

the data. In genotype imputation, $\boldsymbol{X}$ records the observed genotype dosages (0, 1, or 2 counts of the reference allele), with rows corresponding to people and columns to SNPs. Imputation is performed over a narrow genomic window of a few hundred SNPs where linkage disequilibrium prevails. Including reference individuals typed on out-of-sample SNPs is a key part of the strategy.

Because every rank r matrix $\boldsymbol{Y}$ of dimension $m \times n$ can be expressed as a matrix product $\boldsymbol{UV}$, where $\boldsymbol{U}$ is $m \times r$ and $\boldsymbol{V}$ is $r \times n$, matrix completion can be phrased as updating the factors $\boldsymbol{U}$ and $\boldsymbol{V}$ of $\boldsymbol{Y}$ in the loss function (1). Imputation is iterative, and to restore symmetry at iteration m, each missing entry x_{ij} is imputed by its current best guess $(\boldsymbol{U}_m\boldsymbol{V}_m)_{ij}$. New values $\boldsymbol{U}_{m+1}$ and $\boldsymbol{V}_{m+1}$ can be recovered by taking the singular value decomposition (SVD) of $\boldsymbol{Z}_m$, the current completed version of $\boldsymbol{X}$. The MM (majorization/minimization) principle of optimization shows that this procedure drives the loss downhill (Lange 2016).

Chi et al. (2013) compared the matrix completion program Mendel Impute to several popular model-based imputation programs including MaCH and IMPUTE2 using a number of simulated and real datasets. The accuracy of imputation is dependent on the nature of the specific scenarios and so no program was universally most accurate. The least favorable scenario for Mendel Impute in terms of accuracy occurred when imputing genotypes between high density microarray platforms using as a measure of accuracy the mean r^2 between the imputed values and the true genotypes at masked loci. In this case Mendel Impute was slightly worse then MaCH which was slightly worse than IMPUTE2 (Table 1). In other scenarios Mendel Impute was more accurate than MaCH and IMPUTE2 and in still others they were are roughly the same. However, in all the scenarios presented Mendel Impute was at least an order of magnitude faster than MaCH or IMPUTE2.

Alternating least squares provides an alternative to SVD that is potentially much faster (Hastie et al. 2015). The alternating updates

$$\boldsymbol{V}_{m+1} = (\boldsymbol{U}_m^t\boldsymbol{U}_m)^{-1}\boldsymbol{U}_m^t\boldsymbol{Y}_m \quad \text{and}$$
$$\boldsymbol{U}_{m+1} = \boldsymbol{Z}_m\boldsymbol{V}_{m+1}^t(\boldsymbol{V}_{m+1}\boldsymbol{V}_{m+1}^t)^{-1}$$

can achieve extremely high numerical throughput on modern computer architecture such as multicore CPUs and multiple GPUs. Because alternating least squares offers no guarantee of finding the global minimum of the loss, initial values for $\boldsymbol{U}$ and $\boldsymbol{V}$ should be as accurate as possible. Application of a randomized SVD to supply initial values is one possibility (Liberty et al. 2007). In practice we divide the current window into equal thirds and construct a hold-out-set by masking entries in the outer two thirds. We then choose the best rank r based on performance on the hold-out-set. Once we impute missing entries in the middle third, we shift the window to the right and begin again.

Table 1 Comparison of imputation methods

	Mendel impute	MaCH	IMPUTE2
r^2	0.683	0.751	0.802
Relative time	1.00	13.10	7.41

Enhancements to ordinary GWAS

MendelGWAS.jl performs ordinary SNP-by-SNP association testing. To maximize speed in linear, logistic, and Poisson regression, MendelGWAS.jl employs score tests (Amin et al. 2007; Chen and Abecasis 2007; Clark et al. 2016; Zhou et al. 2017). For the most significant SNPs, the score test is supplemented by the slower but more accurate likelihood ratio test (LRT). Principal components can be included as predictors, SNPs and subjects can be filtered by success rates, and a Manhattan plot is provided.

In addition to these standard approaches to GWAS, we are in the process of implementing score tests for generalized linear models (GLMs) (Schaid et al. 2002). GLMs permit trait–genotype relations to be modeled with more exotic response distributions. We are also planning to develop an efficient score test for the challenging Cox survival model (Kawaguchi et al. 2018; Mittal et al. 2014; Suchard et al. 2013). Multinomial regression models for complex categorical phenotypes would be a valuable extension of logistic regression (Morris et al. 2010). Finally, efficient GWAS for ordered discrete phenotypes is becoming increasingly important for the study of complex diseases and traits derived from electronic health records.

Iterative hard thresholding

To avoid the computational complexity of multiple regression and the identifiability issues caused by having more predictors p than sample individuals n (Bühlmann and Van De Geer 2011), GWAS has traditionally focused on the marginal effects of single SNPs. Previously we introduced lasso penalized regression to GWAS to perform subset selection (Wu et al. 2009; Zhou et al. 2010). Our recent paper (Keys et al. 2017) implements a better heuristic, iterative hard thresholding (IHT), to solve this inherently combinatorial problem. We showed that IHT is better for GWAS than lasso or MCP penalties in controlling for false positive and false negative rates, in reducing parameter shrinkage, and in capturing heritability. It achieves these goals with little sacrifice in computational speed.

We now sketch how IHT iterates toward good local optima. To keep the discussion simple, consider the setting of linear regression with design matrix $\boldsymbol{X}$, response vector $\boldsymbol{y}$, and parameter vector $\boldsymbol{\beta}$. The goal is to minimize the loss function $f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to the sparsity condition $\|\boldsymbol{\beta}\|_0 \leq k$. The notation $\|\boldsymbol{\beta}\|_0$ is shorthand for the number of nonzero entries of $\boldsymbol{\beta}$. In GWAS the entry x_{ij} of $\boldsymbol{X}$ denotes the number (0, 1, or 2) of reference alleles carried by individual i at SNP j or the imputed dosage value. The entry y_i of $\boldsymbol{y}$ corresponding to individual i encodes a continuous trait such as height, blood pressure, or an expression level.

At iteration n, the IHT algorithm (Blumensath and Davies 2008) moves in the steepest descent direction $-\nabla f(\boldsymbol{\beta}_n)$ modified by the sparsity constraint. Here the gradient $\nabla f(\boldsymbol{\beta})$ of the objective equals $-\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. The IHT update is explicitly

$$\boldsymbol{\beta}_{n+1} = P_{S_k}\left(\boldsymbol{\beta}_n - s\nabla f(\boldsymbol{\beta}_n)\right), \tag{2}$$

where s is the steplength and $P_{S_k}(\boldsymbol{\beta})$ denotes projection onto the sparsity set $S_k = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq k\}$. The projection operator $P_{S_k}(\boldsymbol{\beta})$ sends to 0 all but the k largest entries of $\boldsymbol{\beta}$ in magnitude. The preferred entries of $\boldsymbol{\beta}$ are untouched. The steplength s is chosen to minimize $f(\boldsymbol{\beta})$ along the ray $s \mapsto \boldsymbol{\beta}_n - s\nabla f(\boldsymbol{\beta}_n)$ prior to projection. This is achieved by taking

$$s = \frac{\|\nabla f(\boldsymbol{\beta}_n)\|^2}{\|\boldsymbol{X}\nabla f(\boldsymbol{\beta}_n)\|^2}.$$

The best value of k can be chosen by cross-validation.

The theory and practice of IHT continues to advance. Shen and Li (2017) show how to relax the restricted isometry property originally invoked to prove convergence (Blumensath and Davies 2009). Yang et al. (2016) suggest group-sparse IHT to promote sparsity on a group-level. Khanna and Kyrillidis (2018) validate the application of momentum acceleration to IHT. Yuan et al. (2017) and Bahmani et al. (2013) adapt IHT to logistic regression. Further extension to generalized linear models is a natural target. MendelIHT.jl brings IHT under the OpenMendel umbrella. Integration of IHT with SnpArrays.jl unifies data handling and leads to faster code with a smaller memory footprint. Finally, we are investigating weighting predictors to accommodate candidate genes and candidate SNPs (Zhou et al. 2011).

Kinship comparison

Kinship coefficients quantify the degree of relationship between two relatives. Two genes are identical by descent (IBD) if one is a copy of the other or they are both copies of the same ancestral gene. The theoretical kinship coefficient ϕ_{ij} is the probability that a randomly sampled gene at some arbitrary locus from individual i is IBD to a randomly sampled gene at the same locus from individual j. For example, if we assume no inbreeding, $\phi_{ij} = \frac{1}{2}$ if $i = j$, and $\phi_{ij} = \frac{1}{4}$ if i and j are first degree relatives. In the former case, the two genes are sampled with replacement. In an accurately constructed pedigree, the full matrix $\boldsymbol{\Phi}$ of kinship coefficients ϕ_{ij} can be calculated from a simple recurrence (Lange 2003). Jacquard's more complex kinship coefficients (Jacquard 1974) are less useful in practice and harder to calculate (Lange and Sinsheimer 1992). In the MendelKinship.jl module of OpenMendel, Jacquard's coefficients are approximated by the Monte Carlo method of gene dropping.

When pedigrees are unknown or suspect, SNP markers can be used to estimate the kinship matrix $\boldsymbol{\Phi}$ empirically. One popular estimate is the genetic relationship matrix (GRM), represented here by $\boldsymbol{S} = (s_{ij})$. If p_k denotes the reference allele frequency of SNP k, x_{ik} counts the number of reference alleles carried by individual i, and K is the number of SNPs, then the elements of $\boldsymbol{S}$ are calculated as

$$s_{ij} = \frac{1}{K}\sum_{k=1}^{K}\frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{4p_k(1-p_k)}.$$

Alternatives to the GRM include a methods of moments estimator MoM (Day-Williams et al. 2011) and a robust GRM (Manichaikul et al. 2010; VanRaden 2008). The latter is

$$\hat{\phi}_{ij} = \frac{1}{\sum_{k=1}^{K} 4p_k(1-p_k)}\sum_{k=1}^{K}(x_{ik} - 2p_k)(x_{jk} - 2p_k).$$

This unbiased estimator generally has smaller variance than the standard estimator $\boldsymbol{S}$, which is sensitive to low minor allele frequencies (Wang et al. 2017). The MendelKinship.jl module calculates the GRM, the robust GRM, and the MoM estimators. All three of these estimators are special cases of general kinship estimators that are unbiased under ideal conditions (Wang et al. 2017). When there is ethnic inhomogeneity and spread in the degrees of relationships, $\boldsymbol{S}$ can exhibit bias because it confounds close relatedness and ancestry differences (Conomos et al. 2016). Ethnic admixture can be accommodated by replacing the allele frequency p_k by an ethnic specific estimate for each individual i (Conomos et al. 2016).

Finding the variances of these estimators has been impossible without simplifying assumptions (Wang et al. 2017). Our own unpublished approximation to the variance

$$\mathrm{E}(\|\boldsymbol{S} - \mathrm{E}(\boldsymbol{S})\|_{\mathrm{F}}^2) \approx \frac{1}{K^2}\|\boldsymbol{R}\|_{\mathrm{F}}^2\left[\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 + \mathrm{tr}(\boldsymbol{\Phi})^2\right] \tag{3}$$

of the GRM matrix $\boldsymbol{S}$ allows for inbreeding, linkage disequilibrium, and closely related relatives. It relies on the

simplifying assumption that the fourth moments of the SNP counts coincide with the fourth moments of similarly distributed Gaussian random variables. In formula (3), $\text{tr}(\boldsymbol{A})$ is the trace of $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\text{F}}$ is the Frobenius norm of $\boldsymbol{A}$, and $\boldsymbol{R}$ is the correlation matrix of the SNPs (LD matrix).

To check suspect pedigrees for hidden relatedness, one can compare theoretical kinships ϕ_{ij} and empiric kinships $\hat{\phi}_{ij}$. It is convenient to put these on a common scale by subjecting them to an approximate variance-stabilizing transformation. RA Fisher considered the simpler problem of comparing an ordinary covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ to a sample covariance matrix $\boldsymbol{S} = (s_{ij})$. Under an assumption of normality, he argued (Fisher 1915; Fisher 1921) that the quantity

$$\tanh^{-1}\left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right) - \tanh^{-1}\left(\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\right)$$

is approximately normal with mean 0 and variance $(K-3)^{-1}$, where K is the sample size, and $\tanh^{-1}$ is the inverse hyperbolic tangent function. By analogy, we subject the GRM matrix or one of its variants to Fisher's transformation and order the discrepancies from least to greatest in absolute value. The OpenMendel MendelKinship.jl tutorial explains in a concrete example how transformation identifies outlier pairs.

Variance component models

Association studies are subject to the effects of unmeasured confounding. The most common confounder is ethnic ancestry (Astle and Balding 2009; Helgason et al. 2005; Knowler et al. 1988), which arises when both trait values and marker allele frequencies differ by region of origin. Ancestry informative markers are particularly prone to show up as false positives in a naïve GWAS (Rosenberg et al. 2003). Currently there are two general adjustments for ethnic ancestry. The first approach uses either a few principal components of the GRM matrix (Patterson et al. 2006; Price et al. 2006; Zhu et al. 2002) or estimated ancestry proportions (Alexander et al. 2009; Pritchard et al. 2000) as fixed effects. The second approach explicitly accounts for the correlation between subjects by including an estimate of the kinship matrix, e.g. the GRM matrix, as a random effect in a variance components model. When reliable pedigrees are available, the second approach is analogous to positing the theoretical kinship matrix as a random effect (Boerwinkle and Sing 1987). Because the theoretical kinship matrix does not capture hidden correlations, inclusion of the one of the SNP based estimates of the kinship matrix is usually preferred.

In any event, the variance components model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sum_{j=1}^{k} \sigma_j^2 \boldsymbol{V}_j)$ figures prominently in genome-wide association testing (Falconer and Mackay 1996; Lange 2003). In this model $\boldsymbol{\beta}$ are the fixed effects of covariates $\boldsymbol{X}$ and σ_j^2 is the variance of the jth random effect. Estimation of the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)^t$ has been the subject of intense study for decades. Most statisticians opt for maximum likelihood or restricted maximum likelihood. In the linear mixed model, the covariance matrices $\boldsymbol{V}_j$ factor as $\boldsymbol{U}_j\boldsymbol{U}_j^t$. The factored form is advantageous if $\boldsymbol{U}_j$ is $n \times r_j$ with r_j small. In the absence of low rank structure, one can take $\boldsymbol{U}_j$ to be the Cholesky factor of $\boldsymbol{V}_j$.

The covariance model $\boldsymbol{W} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{I}$ corresponds to polygenic background (σ_a^2) plus random noise (σ_e^2). The kinship matrix here can be theoretical or empirical. The model is overly simplistic but widely applied due to its computational tractability. It omits dominance effects, shared environment, and parent of origin effects, among other things. Calculation of the inverse and determinant of $\boldsymbol{W}$ is the rate limiting step in estimation. In the simple polygenic model, a good tactic is to first calculate the spectral decomposition $\boldsymbol{ODO}^t$ of $\boldsymbol{\Phi}$, where $\boldsymbol{D}$ is a diagonal matrix. One can then exploit the formulas $\det \boldsymbol{W} = \det(\sigma_a^2\boldsymbol{D} + \sigma_e^2\boldsymbol{I})$ and $\boldsymbol{W}^{-1} = \boldsymbol{O}(\sigma_a^2\boldsymbol{D} + \sigma_e^2\boldsymbol{I})^{-1}\boldsymbol{O}^t$. The indicated determinant and inverse of the diagonal matrix are trivial to compute (Kang et al. 2010; Lippert et al. 2011; Svishcheva et al. 2012).

Our program VarianceComponentModels.jl incorporates this spectral decomposition tactic. It also treats more realistic models with multiple variance components and multivariate traits. For estimation we have compared Fisher scoring and the EM algorithms long familiar to computational statisticians. We have also explored a new MM algorithm that alternates updates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ (Zhou et al. 2018). The normal equation update of $\boldsymbol{\beta}$ in our algorithm is

$$\boldsymbol{\beta}_{n+1} = (\boldsymbol{X}^t\boldsymbol{W}_n^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}_n^{-1}\boldsymbol{y},$$

where $\boldsymbol{W}_n$ is the value of $\sum_{j=1}^{k} \sigma_j^2 \boldsymbol{V}_j$ at the current estimate of $\boldsymbol{\sigma}^2$. The variance component updates are

$$\sigma_{n+1,i}^2 = \sigma_{ni}^2 \sqrt{\frac{(\boldsymbol{y} - \boldsymbol{X\beta}_{n+1})^t\boldsymbol{W}_n^{-1}\boldsymbol{V}_i\boldsymbol{W}_n^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}_{n+1})}{\text{tr}(\boldsymbol{W}_n^{-1}\boldsymbol{V}_i)}}. \quad (4)$$

The MM algorithm converges faster than the standard EM algorithm. Fisher scoring requires fewer iterations to converge but substantially more effort per iteration, particularly in high dimensions. Both the EM and MM algorithms can be accelerated by quasi-Newton extrapolation (Zhou et al. 2011).

The VarianceComponentModels.jl module serves as a convenient vehicle for other genetic applications. One example is Mendelian randomization (MR). Observational studies

often find an association between a biomarker or expression (or methylation) level at a particular locus and a quantitative trait. The goal of MR is to assess the statistical support for this "exposure" as a cause of the trait, as opposed to reverse causality or confounding (Burgess and Thompson 2015). Our Mendelian randomization tutorial for continuous traits demonstrates the value of modularized genetic software such as VarianceComponentModels.jl.

When there are many loci to test, we (Clark et al. 2016; Zhou et al. 2017) and others (Amin et al. 2007; Chen and Abecasis 2007; Kang et al. 2010; Lippert et al. 2011) have employed score tests or their equivalents in variance component models. Score tests are much faster than likelihood ratio tests (LRTs) because score tests require the likelihood to be maximized only under the null hypothesis. In contrast, LRTs require the likelihood be maximized both under the null and alternative hypotheses. When the null hypothesis is the same for all loci tested, this can amount to substantial savings. These score tests are easily extended to include maternal genetic effects and maternal-offspring genetic interaction as fixed effects (Clark et al. 2016). Although most software programs implementing the score test adopt the simple covariance model $2\sigma_a^2\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{I}$, in principle other variance components such as household effects can be included.

Our recent analysis of the GWAS data from the COPDGene study (http://www.copdgene.org) exemplifies the vast performance gain and yet ease of use of a typical OpenMendel workflow in genetic heritability analysis of a realistically large data set (Zhou et al. 2018). The data are available from NIH dbGap under phs000179.v5.p2. The steps are: (1) load the binary genotypes of 6,670 individuals at 630,860 SNPs, (2) compute summary statistics on the SNPs, (3) impute missing genotypes, (4) calculate the empirical kinship matrix, (5) load 13 phenotypes, (6) estimate the heritability of each phenotype, (7) estimate the coheritability of each pair of phenotypes, and (8) fit a joint model to all 13 phenotypes. All these steps are performed in a single interactive Julia environment on a common laptop computer. Typically such an analysis pipeline would require running at least five separate programs on a Linux machine.

In our experience, a pure Julia computation is often faster than the corresponding computation in a low-level language such as C, and much faster than any other high-level language such as R or Python. Figure 1 compares the speed of fitting large-scale variance component models in our VarianceComponentModels.jl module to the two cutting edge programs GCTA (Yang et al. 2011) and GEMMA (Zhou and Stephens 2014), both implemented in C++. In this example, there are two variance components, one for additive genetic effects and one for environmental effects. To make a fair comparison, the genetic relationship matrix $\boldsymbol{S}$ was pre-computed using the GCTA software. There are

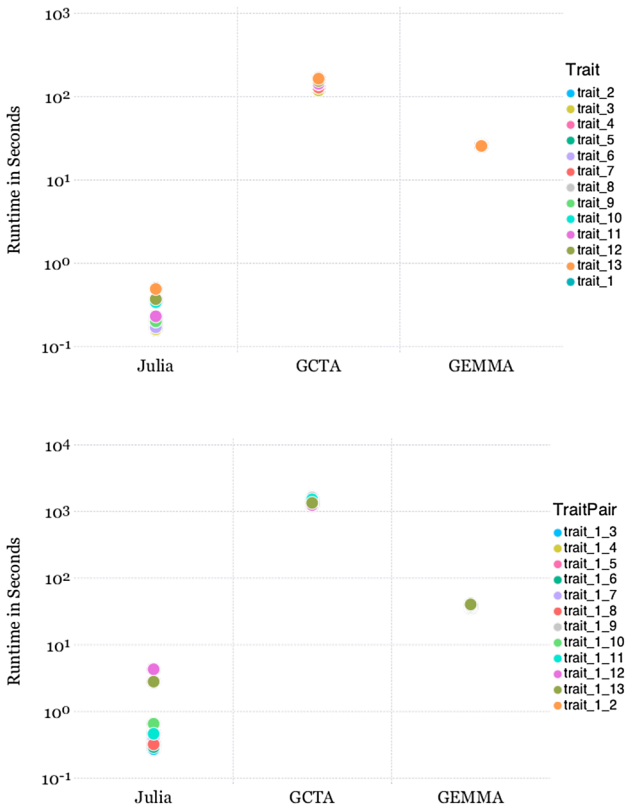

Fig. 1 Comparison of the OpenMendel VarianceComponentModels.jl implementation with GCTA (C++) and GEMMA (C++) for fitting a univariate variance component model $\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_a^2\boldsymbol{S} + \sigma_e^2\boldsymbol{I})$ (top panel) and a bivariate variance component model $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_a \otimes \boldsymbol{S} + \boldsymbol{\Sigma}_e \otimes \boldsymbol{I})$ (bottom panel). GEMMA and OpenMendel runtimes exclude the eigen-decomposition of $\boldsymbol{S}$, which is precomputed

13 continuous phenotypes. For both univariate (top panel) and bi-variate (bottom panel) models, we observe between 5 and 100 fold speedup over GEMMA and even more over GCTA. In all cases, the final log-likelihoods by Julia match those by GCTA and GEMMA to the third digit.

The current versions of GCTA and GEMMA are only available for the x86 64-bit Linux operating system, while Julia, and thus OpenMendel, are available on all common systems. It is remarkable that a cross-platform, interactive, high-level language such as Julia can achieve such excellent computational efficiency.

SNP-set analysis

SNP-set analysis, or pathway-based analysis (Wang et al. 2007), is a powerful, widely-used strategy in sequencing studies. SNPs are grouped into sets to be examined for association with a certain phenotype. This analysis has been shown to have increased power over individual SNP

analysis, especially for identifying rare variant associations (Lee et al. 2014).

Two types of SNP-set analyses are under active development within OpenMendel. The VarianceComponentTests.jl module implements different approaches for testing a set of markers as random effects. Notably the sequence kernel association test (SKAT) (Wu et al. 2011) is the first method to incorporate the generalized linear mixed model in testing the effect of a set of variants on a quantitative or dichotomous trait. Our recent work on exact tests (Zhou et al. 2016) boosts the power of SKAT on small samples.

In contrast to marginal SNP-set analysis, an alternative approach is subset selection in a joint model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sum_{j=1}^{m} \sigma_j^2 \boldsymbol{V}_j + \sigma_0^2 \boldsymbol{I})$, where $\boldsymbol{V}_j$ is the kernel matrix for the jth SNP-set and the σ_j^2 for $j \geq 1$ are the variance components subject to selection. Variance component selection is achieved by minimizing the penalized log-likelihood

$$-L(\boldsymbol{\beta}, \boldsymbol{\sigma^2}) + \sum_{j=1}^{m} P_\lambda(\sigma_j),$$

where $L(\boldsymbol{\beta}, \boldsymbol{\sigma^2})$ is the log-likelihood function and $P_\lambda(\sigma_j)$ is a penalty function. Several penalties, including the ridge, the lasso, the smoothly clipped absolute deviation (SCAD), and the minimax concave penalty (MCP), are implemented. The MM update (4) generalizes to penalized estimation because the variance components σ_j^2 are nicely separated in the surrogate function (Zhou et al. 2018).

Simulation utilities

Simulation is vital in demonstrating the accuracy and power of new statistical methods. It is also important in designing genetic studies, where overly simplistic assumptions can lead to low power. Although there are a number of simulators already available (Liu et al. 2008; Schäffer et al. 2011; Yuan et al. 2012), there is plenty of room for improvement. The unified nature of the OpenMendel environment makes it easy to craft code for simulating traits conditional on genotypes under any generalized linear model (GLM) or generalized linear mixed models (GLMM).

At the time that this article was written, the MendelTraitSimulate.jl option was under development. In its current version, we accommodate study designs involving both unrelateds and multigenerational families. We allow the user to specify both fixed and random effects for simulated univariate or multivariate traits. The simulated traits can be based on arbitrary functions of the provided covariates. By default, the program will use the PLINK format and make appropriate calls to SnpArrays.jl and VCFTools.jl.

Tutorials

Accompanying this article we have prepared a collection of tutorials via Jupyter Notebooks to demonstrate interactive genetic analysis using OpenMendel packages (https://github.com/OpenMendel/Tutorials). These include (1) PLINK binary data input, summary statistics, filtering, and visualization, (2) kinship calculation and comparison, (3) population GWAS, (4) iterative hard thresholding for GWAS, (5) heritability estimation, (6) Mendelian randomization, (7) GWAS based on linear mixed models, (8) SNP-set analysis, and soon to come (9) trait simulation. These tutorials will adapt to and grow with the expanding OpenMendel ecosystem.

Discussion

Readers may be familiar with our existing statistical package Mendel (Lange et al. 2013). Although Mendel possesses many advantages, our goal going forward is not to modernize it, but to create an entirely new open source platform. Although Mendel is free, it is not open source. The Fortran language underlying it is also antiquated. Fortran lacks the supporting libraries of R and Matlab, its graphics functionality is nil, it neglects crucial statistical and linear algebra tools, and its code is needlessly verbose.

For the sake of brevity, we have not discussed many OpenMendel modules. Omitted modules include: (1) discovery of ancestry informative markers, (2) estimation of allele frequencies from pedigree data, (3) testing for transmission disequilibrium by the gamete competition model, (4) random genotype generation by gene dropping, (5) genetic counseling, (6) two-point linkage analysis, (7) location scores for linkage analysis, and (8) function optimization by recursive quadratic programming. Table 2 lists the currently available OpenMendel analysis options and utility packages as well as those soon to be released as part of the OpenMendel project.

OpenMendel is inspired by a vision of genomic analysis that extracts the maximum benefit from the world-wide increase in genetic data and exploits the promise of collaborative, parallel, and distributed computing. We are not alone in this vision. As examples, notable strides have been made by HAIL (https://hail.is/) and TOPMed (https://www.nhlbiwgs.org/awards) in enabling large-scale sequence analysis and the Ark data management system for health and biomedical research (Bickerstaffe et al. 2017; Ranaweera et al. 2018). In our opinion, however, the barriers need to be lowered further to encourage more statistical geneticists and genetic epidemiologists to take part. The Julia language provides the ideal vehicles for this purpose. It is our hope that the OpenMendel project will spark a global effort to build a computing platform equal to the challenges of 21st-century genetic research.

Table 2 Julia packages in the OpenMendel project

OpenMendel option	Description
`MendelAimSelection.jl`	Selects the most informative SNPs for predicting ancestry
`MendelEstimateFrequencies.jl`	Estimates allele frequencies from pedigree data
`MendelGameteCompetition.jl`	Tests for association under the gamete competition model
`MendelGeneticCounseling.jl`	Computes risks in genetic counseling problems
`MendelGWAS.jl`	Tests for association in genome-wide data
`MendelIHT.jl`	GWAS using Iterative Hard Thresholding (forthcoming)
`MendelImpute.jl`	Genotype imputation (forthcoming)
`MendelKinship.jl`	Computes kinship and other identity coefficients
`MendelLocationScores.jl`	Maps a trait via the method of location scores
`MendelTwoPointLinkage.jl`	Implements two-point linkage analysis
`OrdinalGWAS.jl`	Implements GWAS for ordinal categorical phenotypes
`MendelBase.jl`	Base functions for OpenMendel
`MendelGeneDropping.jl`	Simulates genotypes based on pedigrees
`MendelSearch.jl`	Optimization routines
`MendelTraitSimulate.jl`	Trait simulation using GLM and GLMM (forthcoming)
`SnpArrays.jl`	Utilities for handling compressed storage of biallelic SNP data
`VCFTools.jl`	Utilities for handling compressed storage of sequence data
`VarianceComponentModels.jl`	Utilities for fitting and testing variance components models

References

Aird I, Bentall HH, Roberts JF (1953) Relationship between cancer of stomach and the abo blood groups. Br Med J 1(4814):799

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–1664. https://doi.org/10.1101/gr.094052.109

Amin N, Van Duijn CM, Aulchenko YS (2007) A genomic background based method for association analysis in related individuals. PLoS One 2(12):e1274

Astle W, Balding DJ et al (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24(4):451–471

Bahmani S, Raj B, Boufounos PT (2013) Greedy sparsity-constrained optimization. J Mach Learn Res 14(Mar):807–841

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: a fresh approach to numerical computing. SIAM Rev 59(1):65–98. https://doi.org/10.1137/141000671

Bickerstaffe A, Ranaweera T, Endersby T, Ellis C, Maddumarachchi S, Gooden GE, White P, Moses EK, Hewitt AW, Hopper JL (2017) The Ark: a customizable web-based data management tool for health and medical research. Bioinformatics 33(4):624–626. https://doi.org/10.1093/bioinformatics/btw675

Blumensath T, Davies ME (2008) Iterative thresholding for sparse approximations. J Fourier Anal Appl 14(5–6):629–654

Blumensath T, Davies ME (2009) Iterative hard thresholding for compressed sensing. Appl Comput Harmon Anal 27(3):265–274

Boerwinkle E, Sing C (1987) The use of measured genotype information in the analysis of quantitative phenotypes in man. Ann Hum Genet 51(3):211–226

Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, Huffman JE, Ames DC, Carroll A, Conomos MP, Gabriel S et al (2017) Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. Nat Genet 49(11):1560

Bühlmann P, Van De Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, New York

Burgess S, Thompson SG (2015) Mendelian randomization: methods for using genetic variants in causal estimation. Chapman and Hall/CRC, Boca Raton

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. Found Comput Math 9(6):717–772. https://doi.org/10.1007/s10208-009-9045-5

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86(1):6–22

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148(6):1293–1307. https://doi.org/10.1016/j.cell.2012.02.009

Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81(5):913–926

Chi EC, Zhou H, Chen GK, Del Vecchyo DO, Lange K (2013) Genotype imputation via matrix completion. Genome Res 23(3):509–518. https://doi.org/10.1101/gr.145821.112

Chiu Cy, Jung J, Chen W, Weeks DE, Ren H, Boehnke M, Amos CI, Liu A, Mills JL, Ting Lee Ml, Xiong M, Fan R (2016) Meta-analysis of quantitative pleiotropic traits for next-generation sequencing with multivariate functional linear models. European Journal Of Human Genetics 25:350 EP. https://doi.org/10.1038/ejhg.2016.170

Clark MM, Blangero J, Dyer TD, Sobel EM, Sinsheimer JS (2016) The quantitative-MFG test: a linear mixed effect model to detect maternal-offspring gene interactions. Ann Hum Genet 80(1):63–80. https://doi.org/10.1111/ahg.12137

Claster A (2017) Julia joins petaflop club. URL https://juliacomputing.com/press/2017/09/12/julia-joins-petaflop-club.html

Conomos MP, Reiner AP, Weir BS, Thornton TA (2016) Model-free estimation of recent genetic relatedness. Am J Hum Genet 98(1):127–148

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184

Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM (2011) Linkage analysis without defined pedigrees. Genet Epidemiol 35(5):360–370. https://doi.org/10.1002/gepi.20584

Falconer D, Mackay T (1996) C. 1996. Introduction to Quantitative Genetics, pp 82–86

Fan R, Wang Y, Chiu Cy, Chen W, Ren H, Li Y, Boehnke M, Amos CI, Moore JH, Xiong M (2016) Meta-analysis of complex diseases at gene level with generalized functional linear models. Genetics 202(2):457–470. https://doi.org/10.1534/genetics.115.180869. http://www.genetics.org/content/202/2/457

Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10(4):507–521

Fisher RA (1921) On the probable error of a coefficient of correlation deduced from a small sample. Metron 1:3–32

Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, Guarino P, Aslan M, Anderson D, LaFleur R, Hammond T, Schaa K, Moser J, Huang G, Muralidhar S, Przygodzki R, O'Leary TJ (2016) Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 70:214–223. https://doi.org/10.1016/j.jclinepi.2015.09.016

Hall MA, Wallace J, Lucas A, Kim D, Basile AO, Verma SS, McCarty CA, Brilliant MH, Peissig PL, Kitchner TE et al (2017) Plato software provides analytic framework for investigating complexity beyond genome-wide association studies. Nat Commun 8(1):1167

Hastie T, Mazumder R, Lee JD, Zadeh R (2015) Matrix completion and low-rank SVD via fast alternating least squares. J. Mach. Learn. Res. 16(1):3367–3402, http://dl.acm.org/citation.cfm?id=2789272.2912106

Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37(1):90

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44(8):955

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6):e1000,529. https://doi.org/10.1371/journal.pgen.1000529

Jacquard A (1974) The genetic structure of populations, vol 5. Springer Science & Business Media, New York

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42(4):348–354

Kawaguchi ES, Suchard MA, Liu Z, Li G (2018) Scalable sparse Cox regression for large-scale survival data via broken adaptive ridge. arXiv:1712.00561 (in preparation)

Keys KL, Chen GK, Lange K (2017) Iterative hard thresholding for model selection in genome-wide association studies. Genet Epidemiol 41(8):756–768

Khanna R, Kyrillidis A (2018) Iht dies hard: Provable accelerated iterative hard thresholding. In: International Conference on Artificial Intelligence and Statistics, pp 188–198

Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet 29(1):23–30

Kim J, Bai Y, Pan W (2015) An adaptive association test for multiple phenotypes with GWAS summary statistics. Genet Epidemiol 39(8):651–663

Knowler WC, Williams R, Pettitt D, Steinberg AG (1988) Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. Am J Hum Genet 43(4):520

Lange K (2003) Mathematical and statistical methods for genetic analysis. Springer Science & Business Media, New York

Lange K (2016) MM Optimization Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA. https://doi.org/10.1137/1.9781611974409.ch1

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013) Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics 29(12):1568–1570

Lange K, Sinsheimer J (1992) Calculation of genetic identity coefficients. Ann Hum Genet 56(4):339–346

Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95(1):5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816–834. https://doi.org/10.1002/gepi.20533

Liberty E, Woolfe F, Martinsson PG, Rokhlin V, Tygert M (2007) Randomized algorithms for the low-rank approximation of matrices. Proc Natl Acad Sci USA 104(51):20167–20172. https://doi.org/10.1073/pnas.0709640104

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8(10):833–835

Liu Y, Athanasiadis G, Weale ME (2008) A survey of genetic simulation software for population and epidemiological studies. Hum Genom 3(1):79

Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B (2017) Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am J Hum Genet 100(3):473–487

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867–2873

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913. https://doi.org/10.1038/ng2088

Metzker ML (2010) Sequencing technologies-the next generation. Nat Rev Genet 11(1):31

Mittal S, Madigan D, Burd RS, Suchard MA (2014) High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. Biostatistics 15(2):207–221. https://doi.org/10.1093/biostatistics/kxt043

Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI (2010) A powerful approach to sub-phenotype analysis in population-based genetic association studies. Gen Epidemiol 34(4):335–343

Novembre J, Peter BM (2016) Recent advances in the study of fine-scale population structure in humans. Curr Opin Genet Dev 41:98–105

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190

Pickrell WO, Rees MI, Chung SK (2012) Next generation sequencing methodologies-an overview. In: Advances in protein chemistry and structural biology, vol. 89, pp. 1–26. Elsevier

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909. https://doi.org/10.1038/ng1847

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Ranaweera T, Makalic E, Hopper JL, Bickerstaffe A (2018) An open-source, integrated pedigree data management and visualization tool for genetic epidemiology. Int J Epidemiol 47(4):1034–1039. https://doi.org/10.1093/ije/dyy049

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73(6):1402–1422

Schäffer AA, Lemire M, Ott J, Lathrop GM, Weeks DE (2011) Coordinated conditional simulation with slink and sup of many markers linked or associated to a trait in large pedigrees. Hum Hered 71(2):126–134

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70(2):425–434

Shen J, Li P (2017) A tight bound of hard thresholding. J Mach Learn Res 18(1):7650–7691

Sobel E, Lange K, OConnell JR, Weeks DE (1996) Haplotyping algorithms. In: Genetic mapping and DNA sequencing, pp. 89–110. Springer

Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D (2013) Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Transactions on Modeling and Computer Simulation (TOMACS) 23(1):article10:1–17. https://doi.org/10.1145/2414416.2414791

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):e1001,779

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. Nat Genet 44(10):1166

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C, Brewerton SC, Bulsara N, Garner C, Metzker G, Sandoval E, Perkins BA, Och FJ, Turpaz Y, Venter JC (2016) Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci 113(42):11901–11906. https://doi.org/10.1073/pnas.1613365113

Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9):418–426

Van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Abdellaoui A, Hofman A, Schönhuth A, Menelaou A, de Craen AJ, van Schaik BD et al (2015) Population-specific genotype imputations using minimac or impute2. Nat Protocols 10(9):1285

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7–24

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101(1):5–22

Wang B, Sverdlov S, Thompson E (2017) Efficient estimation of realized kinship from SNP genotypes. Genetics 210(2)

Wang K, Li M, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81(6):1278–1283

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 1(89):82–93

Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6):714–721

Yang F, Barber RF, Jain P, Lafferty J (2016) Selective inference for group-sparse linear models. In: Advances in Neural Information Processing Systems, pp 2469–2477

Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ et al (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44(4):369

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yuan X, Miller DJ, Zhang J, Herrington D, Wang Y (2012) An overview of population genetic data simulation. J Comput Biol 19(1):42–54

Yuan XT, Li P, Zhang T (2017) Gradient hard thresholding pursuit. J Mach Learn Res 18:166–221

Zhou H, Alexander D, Lange K (2011) A quasi-newton acceleration for high-dimensional optimization algorithms. Stat Comput 21(2):261–273

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel E, Lange K (2011) Penalized regression for genome-wide association screening of sequence data. In: Biocomputing 2011, pp. 106–117. World Scientific

Zhou H, Blangero J, Dyer TD, Chan KhK, Lange K, Sobel EM (2017) Fast genome-wide QTL association mapping on pedigree and population data. Genet Epidemiol 41(3):174–186. https://doi.org/10.1002/gepi.21988

Zhou H, Hu L, Zhou J, Lange K (2018) MM algorithms for variance components models. J Comput Graph Stat Accept. https://doi.org/10.1080/10618600.2018.1529601

Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26(19):2375–2382

Zhou JJ, Hu T, Qiao D, Cho MH, Zhou H (2016) Boosting gene mapping power and efficiency with efficient exact variance component tests of SNP sets. Genetics 204(3):921–931

Zhou JJ, Sinsheimer JS, Cho MH, Castaldi P, Zhou H (2018) MMVC: An efficient mm algorithm to quantify genetic correlations across large number of phenotypes in giant datasets. manuscript in preparation

Zhou X, Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 11(4):407–409. https://doi.org/10.1038/nmeth.2848

Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23(2):181–196