Journal of Statistical Software

August 2015, Volume 66, Code Snippet 3. *http://www.jstatsoft.org/*

MixMAP: An R Package for Mixed Modeling of Meta-Analysis p Values in Genetic Association Studies

Gregory J. Matthews
University of Massachusetts-Amherst

Andrea S. Foulkes
University of Massachusetts-Amherst

Abstract

Genetic association studies are commonly conducted to identify genes that explain the variability in a measured trait (e.g., disease status or disease progression). Often, results of these studies are summarized in the form of a p value corresponding to a test of association between each single nucleotide polymorphisms (SNPs) and the trait under study. As genes are comprised of multiple SNPs, post hoc approaches are generally applied to determine gene-level association. For example, if any SNP within a gene is significantly associated with the trait at a genome-wide significance level ($p < 5 \times 10^{-8}$), then the corresponding gene is considered significant. A complementary strategy, termed *mix*ed modeling of *m*eta-*a*nalysis p values (MixMAP) was proposed recently to characterize formally the associations between genes (or gene regions) and a trait based on multiple SNP-level p values. Here, the **MixMAP** package is presented as a means for implementing the MixMAP procedure in R.

Keywords: genetic association studies, genotype, meta-analysis p values, mixed effects modeling, phenotype, R package, trait.

1. Introduction

Genetic association studies provide an opportunity to investigate the relationships between complex disease phenotypes and genetic polymorphisms. First stage analysis of data arising from these studies generally involves characterizing the association between each single nucleotide polymorphisms (SNPs) and a measured trait. For example, in an analysis unadjusted for covariates where interest lies in modeling a binary measure of disease status, investigators may perform a Cochran-Armitage trend test. Here each SNP is defined by the number of variant alleles present. The results of genetic association studies are then summarized by a

p value for each SNP as a measure of significance of the association with the trait.

Since interest generally lies in characterizing association between genes (or gene regions) and the trait, where genes are comprised of multiple SNPs, an additional analysis step is required. One simple approach that is commonly applied, is to declare a gene as significant if at least one SNP within that gene reaches "genome-wide significance" defined as a p value less than 5×10^{-8}. This is equivalent to a Bonferroni adjustment based on one million tests.

In a recent manuscript, a complementary strategy was proposed to use summary level data (p values) to investigate a group of SNPs simultaneously in their association with the trait. This approach, termed *mix*ed modeling of *m*eta-*a*nalysis p values (MixMAP; Foulkes, Matthews, Das, Ferguson, Lin, and Reilly 2012), has the advantage of identifying genes comprised of several SNPs that exhibit moderate signals but do not individually reach genome-wide significance. MixMAP involves applying a mixed effects modeling framework to transformed SNP-level p values with random gene level cluster effects. This is a natural framework as SNPs within the same gene tend to have moderate to high linkage disequilibrium (LD), which in turn leads to potentially correlated p values. For each gene, the **MixMAP** package provides functions that return an empirical Bayes estimate of the corresponding random effect and a determination of whether the gene exhibits an association with the trait.

This manuscript describes the **MixMAP** package (Matthews 2015) in R which was developed to implement the MixMAP approach. Section 2 begins with a detailed description of the MixMAP procedure. This is followed in Section 3 by an outline of the **MixMAP** package with a comprehensive data example. Finally, the manuscript concludes with a brief discussion and topics for future work in Section 4.

2. The MixMAP approach

MixMAP is designed as a complementary analytic strategy to characterize the association between a gene (or gene region) comprised of one or more SNPs and a trait. This section provides a brief overview of the MixMAP approach, while a more complete description and extensive simulation studies to characterize the method can be found in Foulkes *et al.* (2012). The primary inputs to the MixMAP procedure are: (1) a set of p values for single SNP tests of association for multiple SNPs within and across genes; (2) the SNP name corresponding to each p value; (3) a mapping of SNPs to genes (or gene regions). For the purpose of this presentation, the term "locus" is used to refer generally to a gene or gene region in which there are moderate to high levels of LD among the SNPs. The input p values can be based on a single cohort study or generated based on a meta-analysis of several genetic association studies, as is common practice in large-scale investigations.

Formally, let $\mathcal{L}_i$ represent the ith locus, $i = 1, 2, \ldots, N$, and let $\mathcal{L}_i = (s_{i1}, \ldots, s_{in_i})$ where s_{ij} is the jth SNP in locus i. In a typical genetic association study, a p value is calculated for each s_{ij}. MixMAP begins by ranking the set of all $M(= \sum_{i=1}^{N} n_i)$ p values and then applying an inverse normal transformation. The result of this transformation is expressed as $y_k = \Phi^{-1}(r_k)$ where $r_k = \frac{k}{M+1}$, k is the rank of the SNP, and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative density function for a standard normal. A mixed effects models is then fit of the form:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \epsilon_i, \tag{1}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is a vector of transformed p values (y_k's) for the ith location and

Step 1: Transformation. p values resulting from univariate analysis in a genetic association study are ranked and used to calculate r_k. In turn, the r_k are transformed to normality to produce y_k.

Step 2: Mixed Modeling. Using the y_k as the dependent variable, a mixed effects model is fitted with a fixed intercept and random intercepts for each locus.

Step 3: Prediction. Empirical Bayes estimates and one-sided prediction intervals are created for each locus with a correction for multiple loci.

Step 4: Locus Selection. Loci are selected as statistically meaningful if the upper limit of the corresponding prediction interval is less than 0.

Figure 1: The MixMAP algorithm.

the matrix $\mathbf{X}_i$ is the corresponding matrix of covariates. $\mathbf{Z}_i$ is a vector of 1's of length n_i and b_i is used to model the latent effect of the ith location. By assumption, we also have that $b_i \sim N(0, \sigma_b^2)$, $\epsilon_i \sim N(0, \mathbf{I}_{n_i}\sigma^2)$ and $b_i \perp \epsilon_i$, where $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identity matrix.

The values of interest here are the collection of estimates of b_i for $i = 1, \ldots, N$, corresponding to the gene-level effects. In this case, the best linear unbiased estimator for b_i is:

$$\mathsf{E}(b_i|\mathbf{y}_i) = \sigma_b^2 J_{n_i}^\top \Sigma_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\beta}), \tag{2}$$

where $\Sigma_i = \mathsf{COV}(\mathbf{y}_i) = \sigma_b^2 J_{n_i} J_{n_i}^\top + \sigma^2 \mathbf{I}_{n_i}$ and J_{n_i} is a vector of 1's of length n_i. The quantity $\widehat{b}_i$, referred to as the empirical Bayes estimator of b_i, is found by replacing σ_b and Σ_i in Equation 2 with corresponding restricted maximum likelihood (REML) estimates. The prediction variance for the ith locus is given by:

$$\mathsf{VAR}(\widehat{b}_i - b_i) = \sigma_b^2 - \sigma_b^4 J_{n_i}^\top \Sigma_i^{-1} \left[\mathbf{I}_{n_i} - \mathbf{X}_i(\mathbf{X}^\top \Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_i^\top \Sigma_i^{-1}\right] J_{n_i} \tag{3}$$

where $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_N^\top)^\top$ and Σ is a block diagonal matrix whose diagonal elements are the matrices $\Sigma_1, \ldots, \Sigma_N$.

When the number of clusters, N, is large $\mathsf{VAR}(\widehat{b}_i - b_i) \approx \mathsf{VAR}(b_i|y_i)$. Since the default variance returned by the `lmer` function in the R package **lme4** (Bates, Maechler, and Bolker 2015) is given by $\mathsf{VAR}(b_i|y_i)$ this can be used in this setting. Here, one is interested in detecting locus effects that are smaller than zero. This is of interest because it is the large negative values of locus effects that correspond to smaller p values. Therefore, one-sided prediction intervals are constructed for each region with the upper limit for the $(1-\alpha)\%$ prediction interval for b_i expressed as:

$$\widehat{b}_i + z_{(1-\alpha)}\sqrt{\mathsf{VAR}(\widehat{b}_i - b_i)}. \tag{4}$$

As multiple prediction intervals are being created, a correction is applied, and α is replaced with $\alpha^* = \alpha/N$ where again N is the number of genes. A summary step-by-step procedure is provided in Figure 1 and the procedure described in this section is implemented in the function `mixmapPI`.

2.1. Testing framework

While prediction intervals are a commonly used framework for estimating random effects, in this context, the width of the prediction interval will approach zero as the number of SNPs within a gene gets large. Therefore, rather than constructing a prediction interval which uses the $\mathsf{VAR}(\widehat{b}_i - b_i)$ as the variance the prediction interval, we also provide an alternative method using a hypothesis testing framework. This method is less sensitive to the number of SNPs within a gene. Formally, the test of interest is $H_0 : \mu_i = 0$ where $b_i \sim N(\mu_i, \sigma_b^2)$ for $i = 1, \ldots, N$. This test is used because $\mathsf{E}(\widehat{b}_i) = \mathsf{E}_y[\mathsf{E}(b_i|y)] = \mathsf{E}(b_i) = \mu_i$, implying that $\widehat{b}_i$ is an unbiased estimator of μ_i. As a result, the test statistic for the ith gene is defined as:

$$U_i = \widehat{b}_i / \sqrt{\mathsf{VAR}(\widehat{b}_i)},$$

where

$$\mathsf{VAR}(\widehat{b}_i) \approx \left(\frac{\frac{n_i \sigma_b^2}{\sigma^2}}{\frac{n_i}{\sigma} + \frac{1}{\sigma_b^2}} \right) = \lambda_i \sigma_b^2.$$

The null hypothesis is then rejected if $U_i > C_{\alpha,i}$. If $C_{\alpha,i}$ is defined to be $C_{\alpha,i} = z_{(1-\frac{\alpha}{N})}$ then this equates to a Bonferroni corrected control of the family-wise error rate. This procedure is implemented in the function `mixmapTest`.

3. Implementation in the MixMAP package

MixMAP is an R (R Core Team 2015) package that is freely available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=MixMAP and that allows for gene-level association analysis based on p values from an association study, and can be installed on any operating system that has R installed. The **MixMAP** package relies on the **lme4** package to implement the mixed models necessary in some of the **MixMAP** functions. The primary functions in the **MixMAP** are `mixmapPI`, which implements the MixMAP algorithm summarized in Section 2, and `mixmapTest`, which implements the procedure described in Section 2.1. In order to use the `mixmapPI` or `mixmapTest` functions, the user must provide a data frame with the following five variables: SNP name, gene name, chromosome number, base pair coordinate, and p value. Both `mixmapTest` and `mixmapPI` output an object of class '`MixMAP`' that has associated methods of `summary` and `plot`.

3.1. Global Lipids Gene Consortium (GLGC) data

The **MixMAP** package is illustrated in an application using meta-analysis results derived from several independent studies of approximately 100,000 individuals in the Global Lipids Gene Consortium (GLGC; Teslovich *et al.* 2010). These data are freely available at http://www.broadinstitute.org/mpg/pubs/lipids2010/. For the purpose of this presentation, we focus on a subset of 31,825 SNPs in 2,960 genes that are also on the ITMAT-Broad-CARe (IBC) 50K SNP array data and can be uniquely mapped to a gene, as described in Foulkes *et al.* (2012). The trait under study is low-density lipoprotein cholesterol (LDL-C), a known causal risk factor for cardiovascular disease. Each p value contained in the data set corresponds to a test of association between LDL-C and a specific SNP, after adjusting for appropriate covariates, as described in Teslovich *et al.* (2010). These p values, along with the mapping of SNPs to genes, serve as the primary input to the `mixmapTest` function as illustrated in the following section.

3.2. Application of MixMAP

The first step after installing the **MixMAP** package is for the user to load the package as follows:

```
R> library("MixMAP")
```

The package contains example data named `MixMAP_example`. This file contains all of the information needed to be run through the `mixmapPI` function, namely SNP name, gene name, chromosome number, base pair coordinate, and p value. Notably, the chromosome number and base pair coordinate are required for the associated `plot` function. If these data are not available, empty columns can be added to the input data frame, and the `mixmapTest` function will still work appropriately. The data are loaded as follows:

```
R> data("MixMAP_example", package = "MixMAP")
```

Inspecting the first ten rows confirms that all of the necessary columns are present for these data to be read directly into the `mixmapTest` function:

```
R> head(MixMAP_example, 10)
```

```
   MarkerName  Gene Chr Coordinate GC.Pvalue
1         rs10  CDK6   7   92221824   0.04027
2  rs10000405 CORIN   4   47411638   0.50130
3  rs10000679 VEGFC   4  177907958   0.11950
4  rs10000850   NMU   4   56195135   0.55470
5   rs1000113  IRGM   5  150220269   0.39400
6   rs1000115  EDG2   9  112834321   0.43150
7  rs10001190  WFS1   4    6335534   0.39640
8  rs10002743  WFS1   4    6327482   0.15830
9   rs1000329  KLF2  19   16310517   0.06940
10 rs10004126 NPY1R   4  164462801   0.60340
```

Note that the variable name for the column containing the SNP name is `"MarkerName"`. This is different than the default (`"SNP"`) and will need to be explicitly stated in the `mixmapTest` function. This is also true of the other variable names in this file. The `mixmapTest` function is applied as follows:

```
R> MixOut <- mixmapTest(MixMAP_example, pval = "GC.Pvalue",
+    snp = "MarkerName", chr = "Chr", coord = "Coordinate", gene = "Gene")
```

The user can specify the α level used in the hypothesis testing framework. This is indicated by `alpha` in the `mixmapTest` function and by default `alpha = 0.05`. The current implementation of `mixmapTest` divides this number by the total number of genes to arrive at the effective level of α for each test.

The `mixmapTest` function returns an object of class '`MixMAP`'. Objects of this class contain 4 slots: `output`, `num.genes.detected`, `detected.genes`, and `lme.out`. The `output` slot contains a data frame with one row for each gene in the input. The columns include the

corresponding empirical Bayes estimates of the random gene-level effects (`postEst`), the estimated posterior variance (`var`), the upper limit of the prediction interval (`predUpper`), and the number of SNPs in the specified gene (`snpCount`), as well as the given chromosome number (`chr`) and base pair location (`coordinate`). The last three columns include the MixMAP p value (`MixMAP_pvalue`), the Bonferroni adjusted MixMAP p value and the false discovery rate adjusted q value (`MixMAP_qvalue`). The first 10 rows of the output slot are displayed below for the GLGC data example:

```
R> MixOut@output[1:10, 1:7]
```

```
     gene     postEst        var predUpper snpCount chr coordinate
1   2-Sep -0.06846933 0.07968001 1.1019438        2   2  241902668
2    A1BG  0.46195614 0.12994785 1.9566390        5  19   63548943
3   A2BP1  0.24109712 0.04844600 1.1537244        1  16    6221176
4     A2M -0.09204013 0.04844600 0.8205872        1  12    9123535
5   AADAT  0.22986746 0.04844600 1.1424948        1   4  171001659
6    AAK1  0.10022170 0.04844600 1.0128490        1   2   69558474
7   AANAT -0.04405343 0.15982268 1.6135617        9  17   71966263
8   ABCA1 -0.53210492 0.21774161 1.4026912      121   9  106690664
9  ABCA12  0.15313878 0.04844600 1.0657661        1   2  215548659
10  ABCA2 -0.49472047 0.12994785 0.9999624        5   9  139031804
```

The `num.genes.detected` slot contains an integer vector of length two with the first element containing the number of genes that were selected as statistically meaningful (`number detected`) and the second containing the total number of genes included in the analysis (`total number of genes`). For the GLGC data, we have:

```
R> MixOut@num.genes.detected
```

```
      number detected total number of genes
                    7                  2960
```

The third slot in a '`MixMAP`' object is called `detected.genes` and contains similar information as the `output` slot, but only for the subset of genes that were selected by the MixMAP algorithm. Along with the information provided in the `output` slot (i.e., empirical Bayes estimates, posterior variances, upper limit of prediction intervals, etc.), this data frame contains the SNP name with the smallest p value within each selected gene, the probit rank-transformed value related to the smallest p value in the selected gene, and several additional summary measures of the p values within the selected gene. Finally, the last slot, `lmer.out`, contains the output from the mixed model fit. For the GLGC data, this is given by:

```
R> MixOut@lmer.out
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: probit.rank.transform ~ 1 + (1 | geneTemp)
   Data: datTemp
REML criterion at convergence: 86755.17
```

```
Random effects:
 Groups   Name        Std.Dev.
 geneTemp (Intercept) 0.4736
 Residual             0.9022
Number of obs: 31825, groups: geneTemp, 2960
Fixed Effects:
(Intercept)
   -0.03679
```

Objects with class 'MixMAP' have methods `summary` and `plot` available for use. The `summary` method displays the number of genes detected and the total number of genes analyzed as well as the top ten MixMAP detected genes based on the upper limit of the prediction interval. For the GLGC data, application of the `summary` method to a 'MixMAP' object yields:

```
R> summary(MixOut)

Number of Genes Detected: 7
Total Number of Genes: 2960

Top Genes:
    gene   postEst snpCount MixMAP_pvalue MixMAP_pvalue_BonferroniAdjusted
2 CELSR2 -2.413384       23  2.085455e-08                       6.172946e-05
4   LDLR -2.022771       28  2.817066e-06                       8.338515e-03
7  SORT1 -1.982453       28  4.309764e-06                       1.275690e-02
1   APOB -1.971705       49  7.984146e-06                       2.363307e-02
3  FADS1 -1.837134        9  2.159791e-06                       6.392981e-03
6  PVRL2 -1.812226       16  1.124855e-05                       3.329572e-02
5 MYBPHL -1.769987       10  6.404506e-06                       1.895734e-02
  MixMAP_qvalue
2  6.172946e-05
4  2.779505e-03
7  3.189225e-03
1  3.938845e-03
3  2.779505e-03
6  4.756532e-03
5  3.791467e-03
```

Finally, the `plot` method associated with a 'MixMAP' object will produce a Manhattan-style plot. Here the x-axis represents the base pair location on the specified chromosome and shading is used to distinguish chromosomes. The y-axis represents the maximum of 0 and the absolute value of the empirical Bayes estimate of the gene level random effect. For the example provided, we have:

```
R> png("ManhattanMixMAP.png", width = 6.83, h = 5, units = "in", res = 300,
+    pointsize = 6)
R> plot(MixOut)
R> dev.off()
```

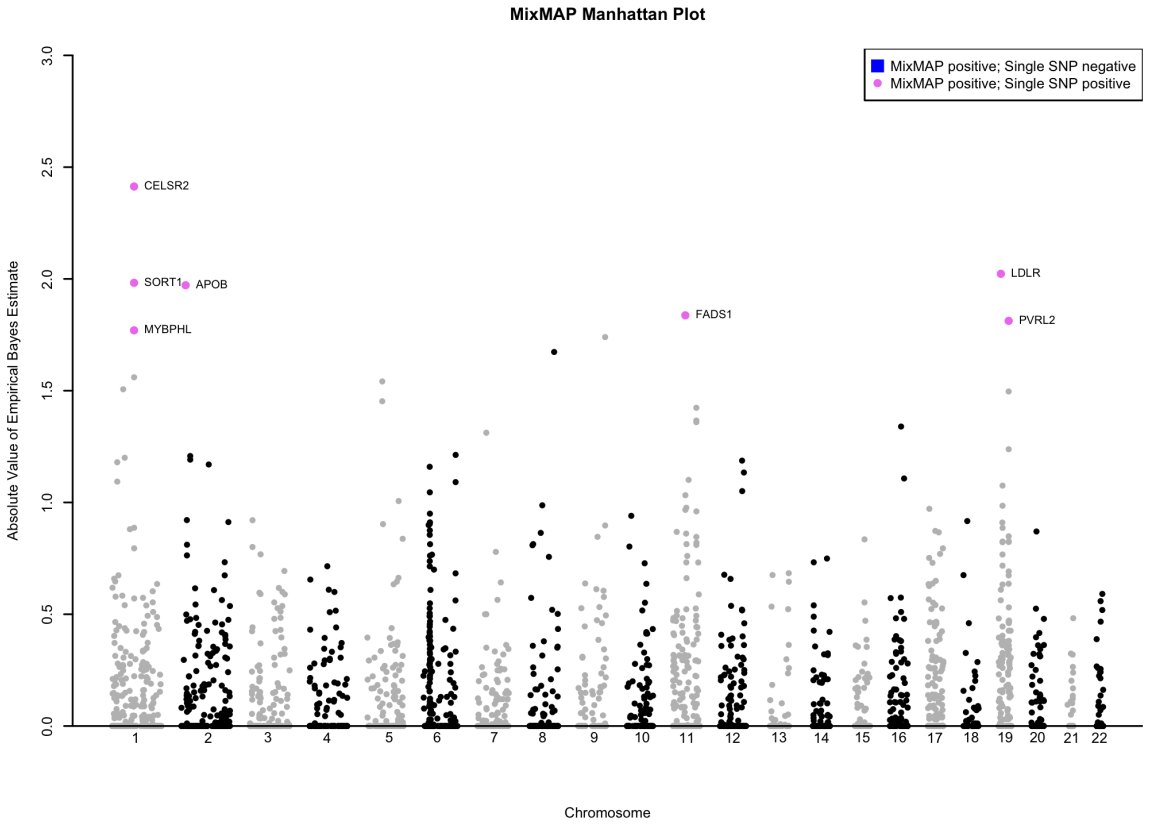

Figure 2: Manhattan style plot for GLGC data.

The resulting plot is illustrated in Figure 2. Colors for the genes in alternating chromosomes can be controlled with the `col.genes` option. The option controlling the color of the detected gene is `col.detected` and `col.text` controls the color of the text that displays the gene name for detected genes. This labeling can be suppressed by specifying `display.text = FALSE`.

4. Discussion

This article presents the R package **MixMAP**, which can be used to implement the algorithm for mixed modeling of meta-analysis p values. The package includes the `mixmapPI` and `mixmapTest` functions, producing output of the class '`MixMAP`', and methods associated with this class include `plot` and `summary`. The `plot` method for objects of class '`MixMAP`' produces a Manhattan-style plot for visualizing the results of the MixMAP procedure. The **MixMAP** package uses S4 methods and classes.

Extensions to the MixMAP procedure will include alternative approaches to defining a significance threshold. In the present application, a Bonferroni correction is applied to account for the number of genes for which a prediction interval is generated. Current research suggests reasonable control of the false discovery rate using this threshold under moderate effect sizes in simulation studies based on the distribution of SNPs and genes in IBC array data (Foulkes *et al.* 2012). Further extensions, however, may be required to apply to data aris-

ing from genome-wide association or deep sequencing studies involving a larger number of genes and/or greater coverage within a gene. Additionally, the **MixMAP** package could be extended to allow for control of potential confounding variables that may lead a gene to be falsely detected.

Further extensions of the **MixMAP** package will also allow users to read in files that are created as output from R functions that perform GWAS analysis directly, such as **GenABEL** (**GenABEL** Project Developers 2013). This would allow users to perform the original testing in R and seamlessly perform the MixMAP algorithm all within the R environment. Finally, extensions of the MixMAP algorithm to the mixture modeling setting are currently being developed. This would allow the model to reflect a more accurate version of the true distribution of the p values, without requiring a first stage ranking, and may result in better levels of detection of genes that are truly associated with the trait of interest.

The **MixMAP** tool presented herein is intended to complement existing approaches to characterizing gene level association. In the example provided, SNPs are grouped into genes (regions) of moderate to high linkage disequilibrium (LD), defined loosely as a measure of correlation between SNPs. Alternatively, a group of SNPs could be defined by a pathway or a gene set, as described in Subramanian *et al.* (2005). To emphasize, the algorithm implemented in this package is flexible with respect to choice of grouping and an alternative clustering variable can be input into the algorithm in place of gene.

Finally, in many scenarios where it is desirable to use the MixMAP procedure, SNPs must be mapped to some grouping, often genes. This in itself is not necessarily a trivial task in and of itself. **ANNOVAR** (Wang, Li, and Hakonarson 2010) is a useful piece of software for annotating SNPs to genes that allows users to annotate genes based on gene definitions (i.e., RefSeq, UCSC, ENSEMBL, GENCODE).

Acknowledgments

Support for this research was provided by NIH/NHLBI R01-HL107196.

References

Bates D, Maechler M, Bolker B (2015). ***lme4****: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-8, URL http://CRAN.R-project.org/package=lme4.

Foulkes AS, Matthews GJ, Das U, Ferguson JF, Lin R, Reilly M (2012). "Mixed Modeling of Meta-Analysis *P*-Values (MixMAP) Suggests Multiple Novel Gene Loci for Low Density Lipoprotein Cholesterol." *PLoS ONE*, **8**(2), e54812.

GenABEL Project Developers (2013). ***GenABEL****: Genome-Wide SNP Association Analysis*. R package version 1.8-0, URL http://CRAN.R-project.org/package=GenABEL.

Matthews GJ (2015). ***MixMAP****: Implements the MixMAP Algorithm*. R package version 1.3.4, URL http://CRAN.R-project.org/package=MixMAP.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirova JP (2005). "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of the United States of America*, **102**(43), 15545–15550.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Cho YS, Go MJ, Kim YJ, Lee JY, Park T, Kim K, Sim X, Ong RTH, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Zhao JH, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RYL, Wright AF, Witteman JCM, Wilson JF, Willemsen G, Wichmann HE, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJG, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruokonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BWJH, Pedersen NL, Pattaro C, Parker AN, Pare G, Oostra BA, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, Meitinger T, McPherson R, McCarthy MI, McArdle W, Masson D, Martin NG, Marroni F, Mangino M, Magnusson PKE, Lucas G, Luben R, Loos RJF, Lokki ML, Lettre G, Langenberg C, Launer LJ, Lakatta EG, Laaksonen R, Kyvik KO, Kronenberg F, Konig IR, Khaw KT, Kaprio J, Kaplan LM, Johansson A, Jarvelin MR, Janssens ACJW, Ingelsson E, Igl W, Hovingh GK, Hottenga JJ, Hofman A, Hicks AA, Hengstenberg C, Heid IM, Hayward C, Havulinna AS, Hastie ND, Harris TB, Haritunians T, Hall AS, Gyllensten U, Guiducci C, Groop LC, Gonzalez E, Gieger C, Freimer NB, Ferrucci L, Erdmann J, Elliott P, Ejebe KG, Doring A, Dominiczak AF, Demissie S, Deloukas P, de Geus EJC, de Faire U, Crawford G, Collins FS, Chen YI, Caulfield MJ, Campbell H, Burtt NP, Bonnycastle LL, Boomsma DI, Boekholdt SM, Bergman RN, Barroso I, Bandinelli S, Ballantyne CM, Assimes TL, Quertermous T, Altshuler D, Seielstad M, Wong TY, Tai ES, Feranil AB, Kuzawa CW, Adair LS, Taylor Jr HA, Borecki IB, Gabriel SB, Wilson JG, Holm H, Thorsteinsdottir U, Gudnason V, Krauss RM, Mohlke KL, Ordovas JM, Munroe PB, Kooner JS, Tall AR, Hegele RA, Kastelein JJP, Schadt EE, Rotter JI, Boerwinkle E, Strachan DP, Mooser V, Stefansson K, Reilly MP, Samani NJ, Schunkert H, Cupples LA, Sandhu MS, Ridker PM, Rader DJ, van Duijn CM, Peltonen L, Abecasis GR, Boehnke M, Kathiresan S (2010). "Biological, Clinical and Population Relevance of 95 Loci for Blood Lipids." *Nature*, **466**(7307), 707–713.

Wang K, Li M, Hakonarson H (2010). "**ANNOVAR**: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data." *Nucleic Acids Research*, **38**(16), e164.

Affiliation:

Gregory J. Matthews, Andrea S. Foulkes
Division of Biostatistics
School of Public Health and Health Sciences
University of Massachusetts Amherst

715 North Pleasant Street
Amherst, MA, United States of America
E-mail: gjm112@gmail.com, foulkes@schoolph.umass.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Code Snippet 3 *Submitted:* 2012-12-10
August 2015 *Accepted:* 2014-10-27