Journal of Statistical Software

July 2016, Volume 71, Code Snippet 2. *doi: 10.18637/jss.v071.c02*

SNPMClust: Bivariate Gaussian Genotype Clustering and Calling for Illumina Microarrays

Stephen W. Erickson
University of Arkansas
for Medical Sciences

Joshua C. Callaway
University of Arkansas
for Medical Sciences

Abstract

SNPMClust is an R package for genotype clustering and calling with Illumina microarrays. It was originally developed for studies using the GoldenGate custom genotyping platform but can be used with other Illumina platforms, including Infinium BeadChip. The algorithm first rescales the fluorescent signal intensity data, adds empirically derived pseudo-data to minor allele genotype clusters, then uses the package **mclust** for bivariate Gaussian model fitting. We compared the accuracy and sensitivity of **SNPMClust** to that of **GenCall**, Illumina's proprietary algorithm, on a data set of 94 whole-genome amplified buccal (cheek swab) DNA samples. These samples were genotyped on a custom panel which included 1064 SNPs for which the true genotype was known with high confidence. **SNPMClust** produced uniformly lower false call rates over a wide range of overall call rates.

Keywords: microarray, genotyping, genomics.

1. Introduction

Genotyping microarrays produce measurements of fluorescent signal intensity corresponding to the alleles of each probed single nucleotide polymorphism (SNP). The fluorescence data are used to infer (“call") genotypes for each subject at each SNP (Figure 1A). Inaccuracy or bias in genotype calling can both decrease the power to detect true genetic associations and yield spurious statistical associations, potentially inflating type I and type II error (Pompanon, Bonin, Bellemain, and Taberlet 2005). Typically, a genotype calling algorithm does not only yield a best call for each subject and SNP but also gives some quantitative measure of the relative certainty associated with each call. This allows an investigator to tune the inevitable trade-off between genotyping sensitivity (a low rate of no-calls) and specificity (a low rate of genotyping errors) according to the requirements of a specific genetic study.

In this code snippet we describe **SNPMClust** (Erickson and Callaway 2013), an R (R Core Team 2016) package for genotype clustering and calling with Illumina microarrays. It builds on **mclust** (Fraley, Raftery, and Scrucca 2013; Fraley and Raftery 2002; Fraley, Raftery, Murphy, and Scrucca 2012), an R package for normal mixture modeling. **SNPMClust** (pronounced *snip-em-clust*) has been used in multiple studies at the Arkansas Center for Birth Defects Research and Prevention on the Illumina GoldenGate custom genotyping platform (Chowdhury *et al.* 2012; Hobbs *et al.* 2014; Li *et al.* 2014a,b; Tang *et al.* 2014, 2015) and is available through the Comprehensive R Archive Network at `https://CRAN.R-project.org/package=SNPMClust`. After describing the package, we compare the performance of **SNPMClust** to that of **GenCall**, Illumina's proprietary algorithm (Illumina Inc. 2005), on a data set of 94 buccal (cheek swab) DNA samples which had undergone whole-genome amplification (WGA) before genotyping on a custom SNP panel using the GoldenGate platform, at 1064 SNPs for which the true genotype was known with high confidence.

2. Description

The four panels of Figure 1 illustrate **SNPMClust**. First, panel A shows a fluorescent intensity plot for a typical SNP in the standard coordinates displayed by **GenomeStudio**, which is the data analysis software provided with Illumina microarrays. Each point represents a different subject and is labeled by its genotype as determined by **GenCall**, the native genotype calling algorithm (Illumina Inc. 2005). Under **GenCall**'s default settings, every subject is assigned a genotype AA, AB, or BB.

To run **SNPMClust**, the fluorescent intensity, genotype, and diagnostic data in **GenomeStudio** must be exported to a tab- or comma-delimited text file and imported to R as a data frame. For each SNP and subject in the data, the following variables are used by **SNPMClust**:

- `X.Raw` and `Y.Raw`, the raw fluorescent intensity values corresponding to the A and B alleles, respectively,
- `X` and `Y`, subject-normalized transformations of `X.Raw` and `Y.Raw`,
- `Theta`, equal to $(2/\pi)\arctan(\texttt{Y}/\texttt{X})$, a measure of relative allele intensity which varies from 0 to 1,
- `R`, equal to $\texttt{X} + \texttt{Y}$, a measure of overall signal intensity,
- `GType`, the genotype call generated by **GenCall**, and
- `Score`, the genotype quality score which varies from 0 (worst) to 1 (best).

The function `prepdata` converts data exported from **GenomeStudio** into the correct format and scaling for **SNPMClust**. Specific to each subject, the normalized vectors `X` and `Y` are linear transformations of `X.Raw` and `Y.Raw`, but with the lowest values of `X` and `Y` collapsed to zero. For these collapsed values, `prepdata` extrapolates positive normalized values by fitting a least-squares line to the non-collapsed normalized vs. raw data, making the normalized log-ratio of B to A allele intensity, $\log(\texttt{Y}/\texttt{X})$, well-defined. As shown in Figure 1B, the normalized log-ratios are, within genotype clusters, much less skewed and heteroskedastic than `Theta`.

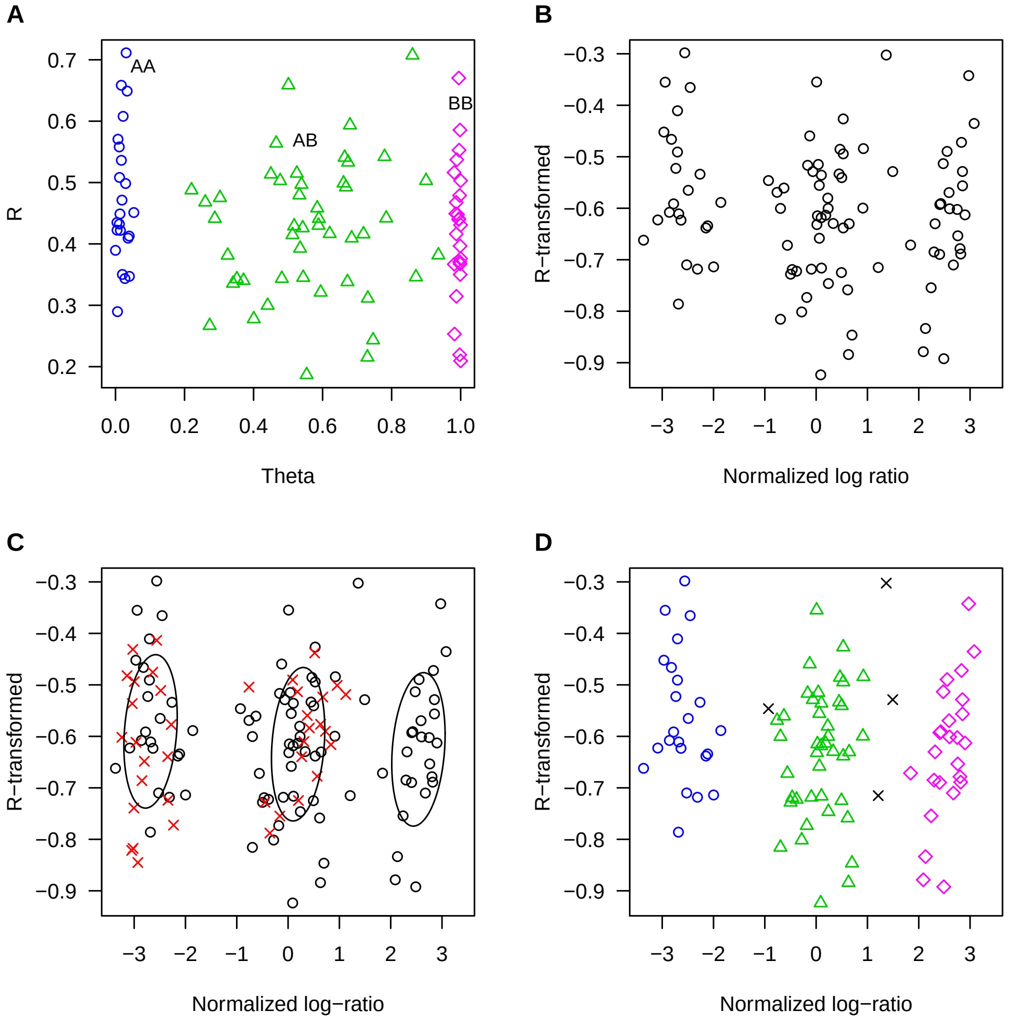

Figure 1: Illustrative example of **SNPMClust**. Panel A shows the intensity plot in **GenomeStudio**'s default scaling (`Theta` and `R`) for the SNP rs2302109 and the buccal data described in Section 3. Color coding indicates **GenCall** genotype assignments under default settings. Panel B shows the same intensity data but in **SNPMClust**'s transformed scaling (normalized log ratio and `R`-transformed). In Panel C, pseudo-data are indicated by red X's and clustering results are shown with 50th-percentile ellipses. The **SNPMClust** genotype calls are shown in panel D, and four no-calls (black X's) are data for which the uncertainty score exceeds the default threshold of 0.01.

There is also right-skewness in `R`, although not obviously so for this particular SNP. The function `prepdata` applies a Box-Cox transformation to `R`, with the parameter λ selected by maximizing the profile likelihood (Venables and Ripley 2010) conditional on the normalized log-ratio. The resulting vector is called `R`-transformed, and Figures 1B, C, and D plot the intensity data rescaled in both dimensions.

The output from `prepdata` is used as input to the function `SNPMClust`. For SNPs with low minor allele frequency (MAF), there might not be enough subjects with minor allele genotypes for these clusters to be identified by **mclust**; this is especially true with lower sample sizes. When this happens, the maximum likelihood solution will use two or even three clusters to model the homozygous major cluster, implicitly defining data from the minor genotypes as outliers. To account for SNPs with low MAF, therefore, `SNPMClust` appends the heterozygous and homozygous minor data with randomly generated pseudo-data whose sample size, location, and variances can either be manually set or, under default settings, automatically selected based on the **GenCall** genotype calls and intensity data (Figure 1C, red X's). The number of pseudo-data points for each minor allele genotype is set to one-fifth of the overall sample size but can be changed via the argument `priorfrac`. SNPs with very low MAF are subject to overfitting and should be used with caution.

Three genotype clusters, corresponding to the genotypes AA, AB, and BB, are defined by three bivariate Gaussian distributions. `SNPMClust` uses the package **mclust** to compute cluster locations and covariances by finding a maximum likelihood solution over the set of all data points. The maximum likelihood solution for the three genotype clusters (shown as ovals in Figure 1C) are computed with **mclust** under two possible covariance models: EEE and EEV. Under EEE, all genotype clusters must share the same covariance matrix, while under EEV, genotype clusters must share the same volume and shape but are allowed to vary in orientation. The final covariance model is chosen based on the Bayesian information criterion (BIC; Schwartz 1978).

Once genotype clusters are specified, the individual likelihood for each point is evaluated under each of the three genotype distributions, and the provisional genotype assignment is based on which model yields the highest likelihood. In addition, **mclust** computes the classification uncertainty for each point, defined as one minus the probability of the most likely genotype (assuming equal genotype probability *a priori*). The default setting for `SNPMClust` is to assign a no-call to any point with uncertainty greater than 0.01, and the cut-off can be changed via the argument `uncertcut`. In Figure 1D, four points falling between cluster centers have uncertainty > 0.01 and are assigned a no-call, indicated by a black X. The plots in Figures 1C and D can be produced by calling `SNPMClust` with the argument `showplots = TRUE`.

Uncertainty is computed conditional on the genotype clusters and therefore does not account for clustering error. One measure of the overall quality of a SNP's clustering is the median (or any other quantile) uncertainty. Setting the argument `qcutoff` to a positive value establishes a lower limit to uncertainty scores for the SNP in question, equal to the specified quantile. This is equivalent to requiring a minimum call rate of `qcutoff` for each SNP, and setting the entire SNP to no-calls if that call rate is not reached. As shown below, using a quantile-based uncertainty threshold can substantially improve genotyping accuracy. The function `SNPMClust` returns a list which includes the genotype calls and uncertainty scores for each subject, as well as the SNP-specific call rate under the arguments `uncertcutoff` and `qcutoff`.

3. Performance comparison

The Arkansas Center for Birth Defects Research and Prevention is a participating center in the National Birth Defects Prevention Study (NBDPS, Yoon *et al.* 2001). As part of an ongoing study of the etiology of non-syndromic congenital heart defects, the Center commissioned a panel of 1536 SNPs in 62 genes in the homocysteine, folate, and transsulfuration pathways using the Illumina GoldenGate custom genotyping platform, as described by Chowdhury *et al.* (2012). The DNA samples used in this study were derived from subject-collected mail-in buccal samples (i.e., cheek swabs) and were subsequently amplified using whole-genome amplification (WGA) because of the relatively low DNA yield and aliquot requirements of the NBDPS. We found that the quality of genotype clustering varied substantially from SNP to SNP, which we attribute to the *in silico* design of the SNP panel based on data from phases I and II of the HapMap project (International HapMap Consortium 2003), without the subsequent quality checks that would be applied to a standard commercial SNP panel.

A subset of Arkansas residents who completed the NBDPS was also recruited for a different study at Arkansas Children's Hospital Research Institute (Hobbs, Cleves, Melnyk, Zhao, and James 2005) and provided both blood and buccal samples. Ninety-six Arkansas NBDPS female participants, for whom both types of samples were available, comprised a pilot study to validate the use of WGA-buccal DNA on the custom genotyping platform. Out of the 96 blood/buccal pairs of DNA samples which were genotyped, 94 pairs exhibited high call rates from both sources of DNA and, furthermore, high concordance with each other. The blood samples, which had not undergone WGA, did indeed yield better overall genotyping performance than the WGA-buccal samples, which has been observed elsewhere (Cunningham *et al.* 2008). For example, the mean GenTrain score, which is Illumina's SNP-wide measure of genotype clustering quality, was 0.786 for blood samples compared with 0.728 for WGA-buccal ($p < 0.001$).

While the original purpose of these paired samples was to validate the use of WGA-buccal DNA, they also gave us the opportunity to compare the performance of **SNPMClust** and **GenCall**. First, a set of reference genotypes was produced solely using the 94 blood DNA samples. To prevent biasing results toward either method, reference genotype calls were required (1) to have concordant calls from each algorithm, (2) to have very good quality scores under both algorithms, and (3) to have 100% call rates for each SNP included in the reference set. The **GenCall** score cutoff was 0.5, resulting in 8993 no-calls (6.2%), while the **SNPMClust** uncertainty cut-off was 10^{-7}, resulting in 7855 no-calls (5.4%). The reference set consisted of 1064 SNPs with 100% concordant call rates under both of these stringent quality score cutoffs.

Genotypes were independently generated from the 94 WGA-buccal samples using both algorithms and were compared to the reference genotypes, with discordant calls considered false calls. By sorting the **SNPMClust** and **GenCall** genotype calls by their respective quality score, a cumulative tally of the number of overall calls and number of false calls was taken, and the false call rate is plotted against overall call rate in Figure 2 (thick black lines). As expected, the false call rate increases with the overall call rate for both algorithms. The **SNPMClust** curve begins at an overall call rate of 66% because this percentage of genotype calls had the lowest possible uncertainty score ($< 2.2 \times 10^{-16}$). At call rates below 97%, the **SNPMClust** curve is uniformly lower than the **GenCall** curve, with both curves rising sharply above 97%.

The performance of **SNPMClust** improves noticeably under a quantile-based uncertainty threshold. Performance improvements for **GenCall** due to quantile thresholding, on the other hand, are moderate because the **GenCall** quality score is already calibrated to a SNP-specific measure of genotyping quality (the GenTrain score). The 95% bootstrap confidence interval for **SNPMClust** under median thresholding, shown in Figure 2, was generated by sampling with replacement from the 1064 SNPs. Over much of the range, all four **GenCall** curves lie above the **SNPMClust** confidence interval.

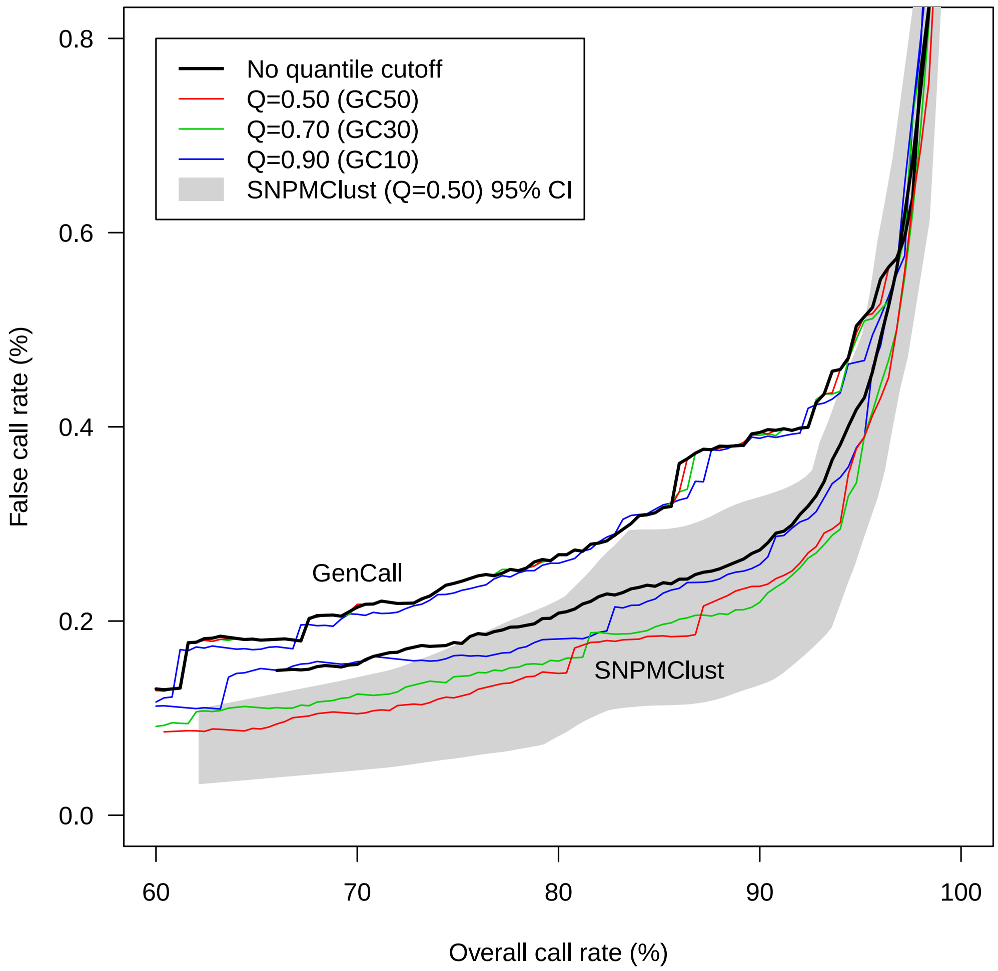

Figure 2: **SNPMClust** and **GenCall** false call rate vs. overall call rate under four different quantile cutoffs. Data are the buccal cell genotypes for 94 subjects and 1064 SNPs. As overall call rate increases (by choosing a more lenient threshold), the rate of false calls also increases. For **SNPMClust**, applying a SNP-specific uncertainty quantile cutoff uniformly improves performance for $Q = 0.50$ (median) and $Q = 0.70$, but not uniformly for $Q = 0.90$. Performance improvement for **GenCall** from quantile thresholding is moderate because the **GenCall** score is already calibrated to GenTrain, which is a SNP-specific measure of genotyping quality. The 95% confidence interval of **SNPMClust** (with median cutoff) is generated by bootstrapping the 1064 SNPs.

4. Discussion

In the above example, **SNPMClust** exhibits uniformly better accuracy than **GenCall** across a wide range of overall call rates. This shows that **SNPMClust** produces a useful metric of genotyping quality for studies in which the quality of intensity clustering is variable.

We note some limitations. **SNPMClust** was specifically developed to analyze noisy intensity data in which genotyping quality varies substantially from SNP to SNP. We did not compare the performance of **SNPMClust** to any of the several third-party algorithms currently available, such as **CRLMM** (Ritchie, Carvalho, Hetrick, Tavare, and Irizarry 2009), **Illuminus** (Teo *et al.* 2007), or **GenoSNP** (Giannoulatou, Yau, Colella, Ragoussis, and Holmes 2008). The package assumes that data are produced from Illumina genotyping microarrays, but could be extended to other sources of data by revising the function `prepdata`.

SNPMClust does not currently incorporate any between-SNP or between-subject information. We envision two extensions that could potentially make the package more robust. First, intensity data across the ensemble of SNPs could be used to specify empirical Bayes priors on cluster locations and variances, using the function `priorControl` as described in Appendix A.4 of Fraley *et al.* (2012). This could be especially helpful for SNPs with low minor allele frequencies. Second, patterns of linkage disequilibrium and/or familial relationships could be used to help resolve ambiguous genotype calls. Such information could be incorporated quite naturally because the function `Mclust` produces a matrix of all genotype probabilities (assuming equal genotype probability *a priori*).

Nevertheless, we have shown in our empirical example that **SNPMClust** significantly outperformed **GenCall** over a wide range of genotyping sensitivities. The accuracy of **SNPMClust** improved substantially under quantile thresholding. Through a combination of uncertainty score and quantile thresholding, an investigator can fine-tune trade-offs between sensitivity and specificity. **SNPMClust** can be used either as the primary genotyping algorithm or as an additional diagnostic for evaluating genotype quality.

Acknowledgments

This research was supported by grants from the National Institutes of Health (5-R01-HD039054-08), the Centers for Disease Control and Prevention (3-U50DD613236-10W1), and the Arkansas Biosciences Institute. We thank Ping Hu and Vanna Gordon for their assistance in troubleshooting the algorithm.

References

Chowdhury S, Hobbs CA, MacLeod SL, Cleves MA, Melnyk S, James SJ, Hu P, Erickson SW (2012). "Associations Between Maternal Genotypes and Metabolites Implicated in Congenital Heart Defects." *Molecular Genetics and Metabolism*, **107**(3), 596–604. doi:10.1016/j.ymgme.2012.09.022.

Cunningham JM, Sellers TA, Schildkraut, M J, Fredericksen ZS, Vierkant RA, Kelemen LE, Gadre M, Phelan CM, Huang Y, Meyer JG, Pankratz VS, Goode EL (2008). "Performance of Amplified DNA in an Illumina GoldenGate BeadArray Assay." *Cancer Epidemiology, Biomarkers & Prevention*, **17**(7), 1781–1789. doi:10.1158/1055-9965.epi-07-2849.

Erickson SW, Callaway J (2013). ***SNPMClust**: Bivariate Gaussian Genotype Clustering and Calling for Illumina Microarrays.* R package version 1.1, URL https://CRAN.R-project.org/package=SNPMClust.

Fraley C, Raftery A, Scrucca L (2013). ***mclust**: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.* R package version 4.1, URL https://CRAN.R-project.org/package=mclust.

Fraley C, Raftery AE (2002). "Model-Based Clustering, Discriminant Analysis and Density Estimation." *Journal of the American Statistical Association*, (97), 611–631. doi:10.1198/016214502760047131.

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012). "**mclust** Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." *Technical Report 597*, Department of Statistics, University of Washington.

Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC (2008). "**GenoSNP**: A Variational Bayes Within-Sample SNP Genotyping Algorithm That Does Not Require a Reference Population." *Bioinformatics*, **24**(19), 2209–2214. doi:10.1093/bioinformatics/btn386.

Hobbs CA, Cleves MA, Macleod SL, Erickson SW, Tang X, Li J, Li M, Nick T, Malik S (2014). "Conotruncal Heart Defects and Common Variants in Maternal and Fetal Genes in Folate, Homocysteine, and Transsulfuration Pathways." *Birth Defects Research Part A: Clinical and Molecular Teratology*, **100**(2), 116–126. doi:10.1002/bdra.23225.

Hobbs CA, Cleves MA, Melnyk S, Zhao W, James SJ (2005). "Congenital Heart Defects and Abnormal Maternal Biomarkers of Methionine and Homocysteine Metabolism." *The American Journal of Clinical Nutrition*, **81**(1), 147–153.

Illumina Inc (2005). "Illumina **GenCall** Data Analysis Software." Technology Spotlight. URL http://www.illumina.com/.

International HapMap Consortium (2003). "The International HapMap Project." *Nature*, **426**(6968), 789–796. doi:10.1038/nature02168.

Li M, Cleves MA, Mallick H, Erickson, W S, Tang X, Nick TG, Macleod SL, Hobbs CA (2014a). "A Genetic Association Study Detects Haplotypes Associated with Obstructive Heart Defects." *Human Genetics*, **133**(9), 1127–1138. doi:10.1007/s00439-014-1453-1.

Li M, Erickson SW, Hobbs CA, Li J, Tang X, Nick TG, Macleod SL, Cleves, A M (2014b). "Detecting Maternal-Fetal Genotype Interactions Associated with Conotruncal Heart Defects: A Haplotype-Based Analysis with Penalized Logistic Regression." *Genetic Epidemiology*, **38**(3), 198–208. doi:10.1002/gepi.21793.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005). "Genotyping Errors: Causes, Consequences and Solutions." *Nature Reviews Genetics*, **6**(11), 847–859. doi:10.1038/nrg1707.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ritchie ME, Carvalho BS, Hetrick KN, Tavare S, Irizarry RA (2009). "R/**Bioconductor** Software for Illumina's Infinium Whole-Genome Genotyping BeadChips." *Bioinformatics*, **25**(19), 2621–2623. doi:10.1093/bioinformatics/btp470.

Schwartz GE (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464. doi:10.1214/aos/1176344136.

Tang X, Cleves MA, Nick TG, Li M, MacLeod SL, Erickson SW, Li J, Shaw, M G, Mosley BS, Hobbs CA (2015). "Obstructive Heart Defects Associated with Candidate Genes, Maternal

Obesity, and Folic Acid Supplementation." *American Journal of Medical Genetics*. doi:10.1002/ajmg.a.36867.

Tang X, Nick TG, Cleves MA, Erickson SW, Li M, Li J, MacLeod SL, Hobbs CA (2014). "Maternal Obesity and Tobacco Use Modify the Impact of Genetic Variants on the Occurrence of Conotruncal Heart Defects." *PLoS ONE*, **9**(9), e108903. doi:10.1371/journal.pone.0108903.

Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, Clark, G T (2007). "A Genotype Calling Algorithm for the Illumina BeadArray Platform." *Bioinformatics*, **23**(20), 2741–2746. doi:10.1093/bioinformatics/btm443.

Venables WN, Ripley BD (2010). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag.

Yoon PW, Rasmussen SA, Lynberg MC, Moore CA, Anderka M, Carmichael SL, Costa P, Druschel C, Hobbs CA, Romitti, A P, Langlois PH, Edmonds LD (2001). "The National Birth Defects Prevention Study." *Public Health Reports*, **116 Suppl 1**, 32–40.

Affiliation:

Stephen W. Erickson
Department of Biostatistics
University of Arkansas for Medical Sciences
currently:
Genetic Epidemiology & Omics Research
RTI International
3040 East Cornwallis Road
Research Triangle Park, North Carolina 27709, United States of America
E-mail: serickson@rti.org

Joshua C. Callaway
Department of Biostatistics
University of Arkansas for Medical Sciences
currently:
PendulumRock Analytics, LLC
E-mail: joshcllw@gmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2016, Volume 71, Code Snippet 2 *Submitted:* 2013-06-05
doi:10.18637/jss.v071.c02 *Accepted:* 2015-06-25