1-2018

Analyzing Metabolomics Data for Association with Genotypes Using Two-Component Gaussian Mixture Distributions

Jason Westra
Iowa State University

Nicholas Hartman
Cornell University

Bethany Lake
Elon University

Gregory Shearer
Pennsylvania State University - Main Campus

Nathan L. Tintle
Dordt College, nathan.tintle@dordt.edu

Analyzing Metabolomics Data for Association with Genotypes Using Two-Component Gaussian Mixture Distributions

Abstract

Standard approaches to evaluate the impact of single nucleotide polymorphisms (SNP) on quantitative phenotypes use linear models. However, these normal-based approaches may not optimally model phenotypes which are better represented by Gaussian mixture distributions (e.g., some metabolomics data). We develop a likelihood ratio test on the mixing proportions of two-component Gaussian mixture distributions and consider more restrictive models to increase power in light of a priori biological knowledge. Data were simulated to validate the improved power of the likelihood ratio test and the restricted likelihood ratio test over a linear model and a log transformed linear model. Then, using real data from the Framingham Heart Study, we analyzed 20,315 SNPs on chromosome 11, demonstrating that the proposed likelihood ratio test identifies SNPs well known to participate in the desaturation of certain fatty acids. Our study both validates the approach of increasing power by using the likelihood ratio test that leverages Gaussian mixture models, and creates a model with improved sensitivity and interpretability.

Keywords

metabolomics, Gaussian mixture distributions, fatty acids

Disciplines

Genetics and Genomics

Analyzing metabolomics data for association with genotypes using two-component Gaussian mixture distributions

Jason Westra
Department of Statistics, Iowa State University
Ames, IA 50011, United States
Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, United States
Email: jwestra@iastate.edu

Nicholas Hartman
Department of Biological Statistics and Computational Biology, Cornell University
Ithaca, NY 14853, United States
Email: ngh32@cornell.edu

Bethany Lake
Department of Mathematics and Statistics, Elon University
Elon, NC 27244, United States
Email: blake@elon.edu

Gregory Shearer
Department of Nutritional Sciences, Pennsylvania State University
University Park, PA 16801, United States
Email: gcs13@psu.edu

Nathan Tintle
Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, United States
Email: Nathan.tintle@dordt.edu

Standard approaches to evaluate the impact of single nucleotide polymorphisms (SNP) on quantitative phenotypes use linear models. However, these normal-based approaches may not optimally model phenotypes which are better represented by Gaussian mixture distributions (e.g., some metabolomics data). We develop a likelihood ratio test on the mixing proportions of two-component Gaussian mixture distributions and consider more restrictive models to increase power in light of *a priori* biological knowledge. Data were simulated to validate the improved power of the likelihood ratio test and the restricted likelihood ratio test over a linear model and a log transformed linear model. Then, using real data from the Framingham Heart Study, we analyzed 20,315 SNPs on chromosome 11, demonstrating that the proposed likelihood ratio test identifies SNPs well known to participate in the desaturation of certain fatty acids. Our study both validates the approach of increasing power by using the likelihood ratio test that leverages Gaussian mixture models, and creates a model with improved sensitivity and interpretability.

Keywords: Metabolomics; Gaussian Mixture Distributions; Fatty Acids

1. Introduction

Genome-wide association studies (GWAS) continue to be viewed as a standard approach to evaluating the genetic component of a variety of diseases and other phenotypes of interest [1]. Standard approaches to the analysis of genotype associations with quantitative phenotypes use linear models.

As suggested in Tintle et al. [2], bimodal distributions are frequently observed in continuous phenotype samples of metabolites, challenging the normality assumption needed in many existing GWAS analysis approaches. For example, red blood cell fatty acid levels have been found to contribute to coronary heart disease [3]. As outlined in Tintle et al. [2], it is biologically reasonable to consider one's fatty acid levels as coming from a mixture of Gaussian distributions, with each of the two or three mean fatty acid levels determined by genetics, and variation around the mean level determined by other factors (e.g., diet; lifestyle). While the standard way of analyzing fatty acids follows the typical GWAS linear model approach, in cases where the distribution does not appear to be normally distributed, a log transformation is sometimes used [4]. However, this log transformation may fail to accurately capture the true distribution of the genotypic and phenotypic data since it ignores the biological reasoning for observing a non-normal distribution. It may be more powerful to directly model the normal mixture distribution and then test for genotype-phenotype association.

Recently, Kim et al. proposed a likelihood ratio test to test for association between copy number polymorphisms (CNP) with quantitative phenotypes and case control outcomes which followed a mixture of Gaussian distributions [5]. The likelihood ratio test evaluates possible differences in the mixing proportions of the Gaussian components by different copy number. Kim et al. showed that the likelihood ratio test was more powerful than a 2 x d chi-squared test with d equal to the number of CNP categories when the underlying data was from a mixture distribution.

We propose adapting the Kim et al. likelihood ratio test to the standard genotype-phenotype testing situation for phenotypes which are distributed as a mixture of Gaussian distributions, like some metabolomics data (e.g., fatty acid levels). We will provide a theoretical framework for the likelihood ratio test, evaluate its performance on simulated data and then apply it to a real set of fatty acid data from the Framingham Heart Study.

2. Methods

2.1. *Notation*

Let X be a quantitative phenotype that follows a two-component Gaussian mixture distribution. Thus, $X \sim \pi N(\mu_1, \sigma^2) + (1-\pi)N(\mu_2, \sigma^2)$ where π is the mixing parameter of the Gaussian components. Let μ_1 and μ_2 be the mean parameters such that $\mu_1 \neq \mu_2$, and we assume a common variance σ^2 for both components. We assume $\pi = p_{01}(n_0/\mathrm{N}) + p_{11}(n_1/\mathrm{N}) + p_{21}(n_2/\mathrm{N})$ where p_{t1} (t = 0, 1, 2) is the proportion of genotype t in the first component of the mixture distribution, n_t (t = 0, 1, 2) is the number of individuals with genotype t, and N is the total number of individuals. We consider the null hypothesis $\mathrm{H}_0 : p_{01} = p_{11} = p_{21}$ and the alternative H_a: at least one is not equal. Let $p_{\phi i} = p_{0i} = p_{1i} = p_{2i}$ $(i = 1,2)$ (see *Figure 1* for a visual representation). Let x_{tb} (b = 1, 2, …

n_t) and (t = 0, 1, 2) be a random variable representing the phenotype for individual b who has genotype t, and let w be a vector of all x_{tb}. Across all the components, the mixing proportion for genotype t must sum to 1 such that $p_{t1} + p_{t2} = 1$ (t = 0 , 1, 2).

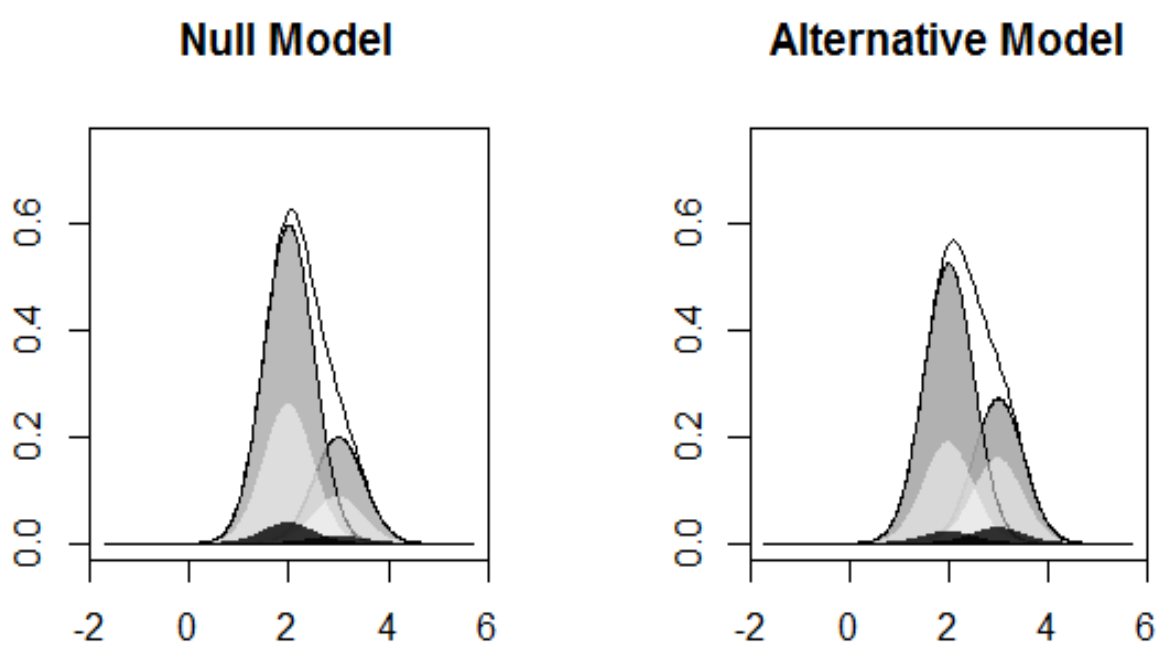

Figure 1 visually illustrates the null and alternative models. The black, light grey, and dark grey two-component mixture distributions are the phenotype distributions for the less common homozygote, the heterozygote and the more common homozygote, respectively. In the null model, 75% of the observations in each genotype are in the component with the smaller mean. In the alternative model, the mixing proportion for the component density with the smaller mean varies across genotypes.

2.2. *Likelihood functions*

2.2.1. *Null and alternative likelihood function*

The likelihood function under the null hypothesis is:

$$L_0 = \prod_{j=1}^{n_0+n_1+n_2} \left(\sum_{i=1}^{2} p_{\phi i} N(w_j | \mu_i, \sigma^2) \right) \quad (1)$$

The likelihood function under the unrestricted alternative hypothesis is:

$$L_1 = \left(\prod_{k=1}^{n_0} \left(\sum_{i=1}^{2} p_{0i} N(x_{0k} | \mu_i, \sigma^2) \right) \right) \left(\prod_{m=1}^{n_1} \left(\sum_{i=1}^{2} p_{1i} N(x_{1m} | \mu_i, \sigma^2) \right) \right) \left(\prod_{h=1}^{n_2} \left(\sum_{i=1}^{2} p_{2i} N(x_{2h} | \mu_i, \sigma^2) \right) \right) \quad (2)$$

2.2.2. *Restricted likelihood function*

When there is a biological understanding of the phenotype-genotype relationship, we recommend restricting the mixing proportions of the test to fit the biological model. We demonstrate two possible models, but our general method easily extends to other models. The first model (LRT_{pro}; Table 1) we consider is that the proportion of change between genotypes 0 and 1 is equal to the change between genotypes 1 and 2. Therefore, we can restrict our parameters of interest to $p_{0i}^* = (p_{01}, 1-p_{01})$, $p_{1i}^* = \big(p_{01}q, 1-(p_{01}q)\big)$, and $p_{2i}^* = \big(p_{01}q^2, 1-(p_{01}q^2)\big)$. The second restricted model (LRT_{add}; Table 2) that we demonstrate describes an equal difference in proportions between groups 0 and 1 and groups 1 and 2. We can restrict our parameters of interest to $p_{0i}^* = (p_{01}, 1-p_{01})$, $p_{1i}^* = \big(p_{01}-q, 1-(p_{01}-q)\big)$, and $p_{2i}^* = \big(p_{01}-2q, 1-(p_{01}-2q)\big)$. Therefore, the likelihood function under these restrictions is:

Table 1. LRT_{pro}

Genotype	Component 1 of Mixture Distribution	Component 2 of Mixture Distribution
0	p_{01}	$1-p_{01}$
1	$p_{01}q$	$1-(p_{01}q)$
2	$p_{01}q^2$	$1-(p_{01}q^2)$

Table 2. LRT_{add}

Genotype	Component 1 of Mixture Distribution	Component 2 of Mixture Distribution
0	p_{01}	$1-p_{01}$
1	$p_{01}-q$	$1-(p_{01}-q)$
2	$p_{01}-2q$	$1-(p_{01}-2q)$

$$L_2 = \left(\prod_{k=1}^{n_0}\left(\sum_{i=1}^{2} p_{0i}^* N(x_{0k}|\mu_i,\sigma^2)\right)\right)\left(\prod_{m=1}^{n_1}\left(\sum_{i=1}^{2} p_{1i}^* N(x_{1m}|\mu_i,\sigma^2)\right)\right)\left(\prod_{h=1}^{n_2}\left(\sum_{i=1}^{2} p_{2i}^* N(x_{2h}|\mu_i,\sigma^2)\right)\right) \quad (3)$$

2.2.3. *Test statistics*

Because $p_{t2} = 1 - p_{t1}$ for all *t*, we can express each likelihood as a function of the parameters μ_1, μ_2, σ^2, and the mixing proportion(s) associated with the $N(\mu_1, \sigma^2)$ distribution. The resulting likelihood ratio test statistics are given by:

$$LRTS = 2\Big(\max_{p_{01}, p_{11}, p_{21}, \mu_1, \mu_2, \sigma^2} \ln(L_1) - \max_{p_{\phi 1}, \mu_1, \mu_2, \sigma^2} \ln(L_0)\Big) \quad (4)$$

$$LRTS_{res} = 2\Big(\max_{p_{01}, q, \mu_1, \mu_2, \sigma^2} \ln(L_2) - \max_{p_{\phi 1}, \mu_1, \mu_2, \sigma^2} \ln(L_0)\Big) \quad (5)$$

Extending the argument provided by Kim et al. the LRTS under the null hypothesis follows a central chi-squared distribution with the degrees of freedom equal to the difference in parameters of the null and alternative models [5]. Therefore, under the null hypothesis, the LRTS has a central chi-squared distribution with 2 degrees of freedom, and the $LRTS_{res}$ follows a central chi-squared distribution with 1 degree of freedom.

2.3. *Simulation*

Using R software, we simulated 1000 datasets with 10,000 individuals per data set. For each, individual, the genotype for a single SNP was generated by assuming Hardy-Weinberg equilibrium and minor allele frequency of either 0.05, 0.10, or 0.25. Trait values for individuals were simulated from two component Gaussian mixture distributions with centers one unit apart and equal variance of the components σ^2 = 0.5 or 0.75. For the mixing proportions of individuals with genotype 0, we used p_{01} = 0.9 or p_{01} = 0.75. We used two different biological models to simulate. In the proportional model we set q equal to 1, 0.9, or 0.75 so that the other mixing proportions were $p_{11} = p_{01}q$ and $p_{21} = p_{01}q^2$. In the additive model we set q equal to 0.1 or 0.2 so that the mixing proportions were $p_{11} = p_{01} - q$ and $p_{21} = p_{01} - 2q$. Simulations were performed on all combinations of the parameters.

2.4. *Statistical analysis*

To evaluate the performance of these tests in direct comparison to the standard procedure of linear and log-linear models, all tests were run on each simulated SNP and phenotype. Each test produced a *p*-value, test statistic and parameter estimates. Type I error rates and power estimates were calculated by dividing the number of observations less than a significance level (Type I error 0.01, power 0.0001) by the total number of simulations. We used an Expectation Maximization (EM) algorithm to find the global maximums of equations (4) and (5). One hundred random start points (RSP) were used for the null likelihood, and 50 RSP and one start point from the maximum of the null were used in the alternative [5]. The EM algorithm ran until a tolerance of 10^{-5} was reached or until 600 and 300 iterations were performed for the null and alternative models respectively.

2.5. *Real data application*

We analyzed 20315 SNPs on chromosome 11 for 5936 individuals from the Framingham Heart Study using the proposed LRT_{pro} test. We looked exclusively at members in the offspring and generation 3 cohorts, all of whom are of European descent. Detailed descriptions of the sample are available elsewhere [6]–[9]. We looked at the red blood cell fatty acid level ratio of arachidonic acid (AA) to dihomo-gamma-linoleneic acid (DGLA). These fatty acid levels were analyzed by gas chromatography as previously described [6]. The desaturation of AA to DGLA occurs primarily via enzymatic activity in the FADS gene complex on chromosome 11. We will use a Bonferroni correction to control the probability of type I errors at 2.47×10^{-6} (0.05/20315).

3. Results

3.1. *Verifying the null distribution and type I error rate*

To confirm that the null distribution of the unrestricted model is a chi-square distribution with two degrees of freedom and that the null distribution of the restricted model is a chi-square distribution with one degree of freedom, we examined simulations when $q = 1$. In addition to examining the novel tests proposed here (LRT_{pro}, LRT_{add}) we also explored the type I error rates of the linear model, log-linear model, and LRT across these same simulations. As shown in Table 3 the type I error rate was controlled by all tests.

Table 3. Type I Error Estimates

		Nominal Significance Level			
	SD	0.05	0.01	0.001	Kolmogorov-Smirnov test p-value[1]
LRT_{pro}	0.5	0.0497	0.011	0.0012	0.6846
	0.75	0.0515	0.0097	0.0010	0.8832
LRT_{add}	0.5	0.0472	0.0108	0.0012	0.7277
	0.75	0.0495	0.0085	0.0008	0.7091
LRT	0.5	0.0557	0.0108	0.0012	0.2269
	0.75	0.0478	0.0078	0.0013	0.7435
Linear Model	0.5	0.0538	0.0107	0.0007	
	0.75	0.0458	0.0070	0.0005	
Log Linear Model	0.5	0.0523	0.0108	0.0007	
	0.75	0.0460	0.0083	0.0007	

[1]As compared to a chi-square distribution.

3.2. *Power estimates*

There were 48 simulations where the alternative hypothesis was true. As summarized in Table 4 (full detailed results are in Supplemental Table 1), the LRT_{pro} has empirical power equal to or greater than all the other tests in all situations. LRT_{add} was the second most powerful test in all 48 simulations. When comparing a linear model to the unconstrained LRT test directly there were 21 simulations where they had different power. In two-thirds of these cases (14 out of 21), LRT had higher power than the linear model. The log-linear model never had an empirical power higher than any other test.

Table 4 Power Estimates

model	q	maf	p_{01}	Linear Model	Log Linear Model	LRT_{pro}	LRT_{add}	LRT
add	0.1	0.05	0.75	0.343	0.26	0.403	0.39	0.295
			0.9	0.44	0.316	0.631	0.624	0.508
		0.1	0.75	0.824	0.736	0.879	0.871	0.798
			0.9	0.898	0.793	0.967	0.966	0.938
		0.25	0.75	0.999	0.997	0.999	0.999	0.999
			0.9	0.999	0.999	1	1	1
pro	0.9	0.05	0.75	0.12	0.095	0.156	0.153	0.105
			0.9	0.325	0.212	0.478	0.467	0.362
		0.1	0.75	0.388	0.31	0.46	0.451	0.342
			0.9	0.75	0.622	0.891	0.887	0.831
		0.25	0.75	0.904	0.844	0.936	0.932	0.892
			0.9	0.998	0.975	1	1	1

Power estimates for standard deviation of .75 for alpha = 0.0001

The choice of 0.0001 as a cutoff for our power estimates is arbitrary as Figure 2 demonstrates. The LRT_{pro} tends to have a smaller *p*-value than the linear model for all thresholds since almost all of the points are above the gray line.

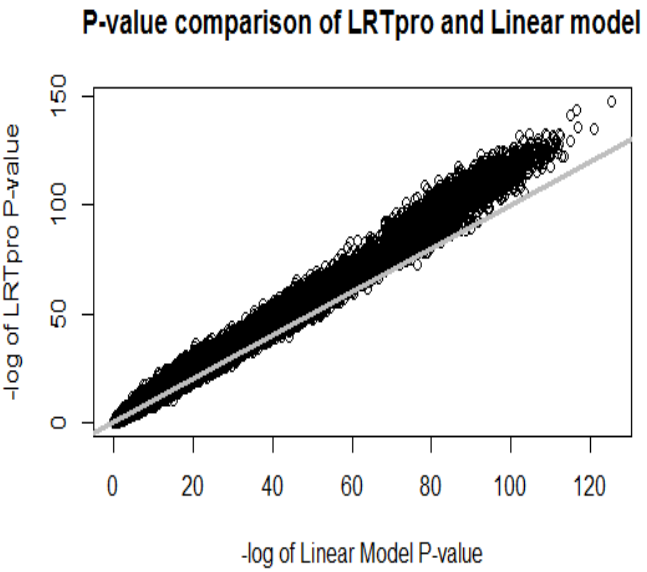

Figure 2. P-value comparison between LRT_{pro} and the linear model.

3.3. *Robustness of model selection*

Since choosing a restriction based on prior knowledge as is done in both LRT_{pro} and LRT_{add} may not be possible in every circumstance, it may not be necessary to choose the exact model. Table 4 shows that LRT_{pro} and LRT_{add} were the most powerful tests even when the other model was simulated. These two restrictions are of similar patterns, but the increase of power is substantial. Therefore, choosing a model at least similar to the true model can increase the power of the test.

3.4. *Parameter estimation*

In order to conduct the LRT, estimates of the underlying parameters of the two-component distribution are obtained. Table 5 illustrates the accuracy and precision of the resulting estimates across a range of simulation settings for the LRT_{pro} approach, with full results for all tests in supplemental tables 2 and 3. In general, LRT_{pro} and LRT_{add} yielded unbiased and accurate estimates across settings. In Table 5, one can see that LRT_{pro} accurately predicted the means of the components both across a wide range of settings and with low variation of the estimate. LRT_{pro} estimated well even when the data was simulated from the additive model. Similar results are obtained when estimating the mixing proportion (see Table 6) and the standard deviation of the components (see supplemental table 4).

Table 5. Estimates of Means for LRT_{pro}

True model	True p_{01}	q	μ_1	Standard deviation of μ_1	μ_2	Standard deviation of μ_2
Add	0.75	0.1	0.0005	0.02936	1.0022	0.0401
	0.9	0.1	-0.0021	0.0240	1.0036	0.0740
	0.75	0.2	-0.0007	0.0240	1.0011	0.0349
	0.9	0.2	-0.0020	0.0206	1.0005	0.0547
Pro	0.75	0.75	0.0005	0.02678	1.0005	0.0356
	0.9	0.75	-0.0014	0.0110	1.0008	0.0518
	0.75	0.9	0.0002	0.0293	1.0030	0.0400
	0.9	0.9	-0.0025	0.02496	1.0028	0.0781

Estimates aggregated across all settings with these parameters and all simulations within each setting, with the true value of $\mu_1 = 0$ and $\mu_2 = 1$.

Table 6 Estimates of Mixing Proportions for LRT_{pro}

model	True p_{01}	q	p_{01}	sd	True p_{11}	p_{11}	sd	True p_{21}	p_{21}	sd
Add	0.75	0.1	0.7509	0.0280	0.65	0.5197	0.1537	0.55	0.5290	0.0625
	0.9	0.1	0.8973	0.0265	0.8	0.6134	0.2786	0.7	0.6059	0.1627
	0.75	0.2	0.7496	0.0247	0.55	0.5097	0.0582	0.35	0.5072	0.1731
	0.9	0.2	0.8985	0.0220	0.7	0.6559	0.1220	0.5	0.5523	0.0744
Pro	0.75	0.75	0.7504	0.0248	0.5625	0.5390	0.0597	0.4219	0.4707	0.1188
	0.9	0.75	0.8983	0.0205	0.6750	0.6627	0.0691	0.5063	0.5139	0.0677
	0.75	0.9	0.7507	0.0284	0.6750	0.5385	0.1754	0.6075	0.5358	0.1037
	0.9	0.9	0.8966	0.0278	0.8100	0.6290	0.2823	0.729	0.6170	0.1814

Estimates aggregated across all settings with these parameters and all simulations within each setting.

3.5. *Real data results*

After analyzing 20321 SNPs on Chromosome 11 in relation to the AA/DGLA ratio, the LRT_{pro} test identified 28 SNPs as significantly associated after applying a Bonferonni multiple testing correction. These 28 SNPs came from 5 different regions on chromosome 11, all of which validated previous GWAS findings. Nineteen significant SNPs are in the well documented [10]–[12]FADS region (bp = 61622896- 61978819). Genes in this region that contain significant SNPs include DAGLA, MYRF, FADS1, FADS2, FADS3, and RAB3IL1 all of which have strong biological basis for desaturation activity [10].

Table 7. Most significant SNPs in each region

rs#	# of SNPs	MAF	Pos	Gene	LRT_{pro} p-value	p_{01}	p_{11}	p_{21}	μ_1	μ_2	σ
rs10751124	1	0.346	85432084	DLG2	$2.50\text{x}10^{-8}$	0.062	0.114	0.162	0.174	0.100	0.023
rs11220658	1	0.350	99618283	CNTN5	$4.52\text{x}10^{-7}$	0.110	0.075	0.051	0.179	0.101	0.024
rs7129015	5	0.198	110772485		$1.86\text{x}10^{-7}$	0.105	0.059	0.034	0.179	0.101	0.024
rs11217753	1	0.167	120181415		$2.94\text{x}10^{-9}$	0.108	0.052	0.025	0.180	0.101	0.024
rs174549	19	0.290	61803910	FADS1	$5.32\text{x}10^{-312}$	0.036	0.183	0.937	0.160	0.097	0.024

As an example interpretation of the results in Table 7, we first note that the significant tests all show similar estimates of the two components of the AA/DGLA ratio (mean of component one between 0.16 and 0.18; mean of component two between 0.097 and 0.101; SD of each component between 0.023 and 0.024). When an individual is genotyped and is the common homozygote at rs174549, they have a 3.6% chance of having their AA/DGLA ratio in the first component. However, if the individual has one less common allele, his chance increases to 18.3%, and with a second copy of the minor allele, it will increase to 93.7%.

4. Discussion

GWAS typically utilize linear models, thus making an assumption about the underlying normality of the data. When data is not normal, a Gaussian mixture distribution may represent a statistically justified and biologically interpretable model of the data. We proposed a constrained likelihood ratio test, which across many simulation settings, was more powerful than the standard linear model and gave accurate parameter estimates. When applied to a real dataset, the method identified biologically relevant SNPs in the well understood FADS region, along with parameter estimates to aid in biological interpretability of the impact of the SNP.

The general LRT framework proposed here shows reasonably good performance compared to the additive linear model, but can be improved upon by further constraining the model and 'saving' a degree of freedom. Our simulations suggest relatively robust performance of the constrained methods (LRT_{pro} and LRT_{add}) to misspecification of the true model though additional simulations across a wider range of misspecifications are needed.

We note that, due to the use of the EM algorithm to generate parameter estimates for use in the LRT, computational time for our proposed methods (3 minutes per test on a single processor with a sample size of 10,000) are much greater than that of the traditional linear model. Nevertheless, with the increasing computational power and the limited number of high minor allele frequency SNPs, it is plausible to run GWAS with this method and is a reasonable option for candidate gene approaches. Further work is necessary to investigate potential areas of computational improvement.

Numerous areas of future work and extension are possible. First, extensions of this work are needed to incorporate covariates and family structure into the method. Standard methods (e.g., first modeling the phenotype by covariates and/or family structure and then modeling the residuals) make normality assumptions and, so, may not be optimal candidates for extension in this Gaussian mixture modeling framework. Imputed data often provides dosages instead of discrete genotypes. Work is needed to extend this framework to allow for dosages in this testing framework. Further applications to genome wide data is necessary to fully understand the impact of this new method. Finally, extensions for multiple-marker testing and relaxing the equal variance assumption are also targets for further exploration.

We have developed a likelihood ratio test that analyzes the differences in mixing proportions between genotypes. The method and null distribution were validated through simulation. There was notable power increase over the more commonly used linear model, especially when we further increased power by restricting the model to incorporate prior biological belief. We have shown that this method is able to accurately predict model parameters. The model was applied to real data, and it replicated many previous findings while also providing more interpretable results. Further work is necessary to apply the model to a wider range of real metabolomics data and to investigate extensions of the model to handle covariates and imputed genotypes.

Supplemental files and R code

All supplemental material can be found at http://homepages.dordt.edu/ntintle/mixture_test.zip.

Acknowledgements

This research was funded by National Institutes of Health (NIH) 2R15HG006915. We thank and acknowledge Hope College and Dordt College for access to the computing clusters.

References

[1] P. M. Visscher *et al.*, "10 Years of GWAS Discovery: Biology, Function, and Translation.," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, Jul. 2017.

[2] N. Tintle, J. W. Newman, and G. C. Shearer, "A novel approach to identify optimal metabotypes of elongase and desaturase activities in prevention of acute coronary syndrome," *Metabolomics*, vol. 11, no. 5, pp. 1327–1337, 2015.

[3] G. C. Shearer, J. V Pottala, J. A. Spertus, and W. S. Harris, "Red blood cell fatty acid patterns and acute coronary syndrome.," *PLoS One*, vol. 4, no. 5, p. e5444, 2009.

[4] D. C. Schwenke, J. P. Foreyt, E. R. 3rd Miller, R. S. Reeves, and M. Z. Vitolins, "Plasma concentrations of trans fatty acids in persons with type 2 diabetes between September 2002 and April 2004.," *Am. J. Clin. Nutr.*, vol. 97, no. 4, pp. 862–871, Apr. 2013.

[5] W. Kim, D. Gordon, J. Sebat, K. Q. Ye, and S. J. Finch, "Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test.," *PLoS One*, vol. 3, no. 10, p. e3475, 2008.

[6] W. S. Harris, J. V Pottala, S. M. Lacey, R. S. Vasan, M. G. Larson, and S. J. Robins, "Clinical correlates and heritability of erythrocyte eicosapentaenoic and docosahexaenoic acid content in the Framingham Heart Study.," *Atherosclerosis*, vol. 225, no. 2, pp. 425–431, Dec. 2012.

[7] B. M. Psaty *et al.*, "Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts.," *Circ. Cardiovasc. Genet.*, vol. 2, no. 1, pp. 73–80, Feb. 2009.

[8] D. R. Govindaraju *et al.*, "Genetics of the Framingham Heart Study Population," *Adv. Genet.*, vol. 62, pp. 33–65, 2008.

[9] L. A. Cupples *et al.*, "The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports.," *BMC Med. Genet.*, vol. 8 Suppl 1, p. S1, Jan. 2007.

[10] N. Tintle *et al.*, "A genome wide assocaition study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart offspring study," *Prostaglandins. Leukot. Essent. Fatty Acids*, vol. 94, pp. 65–72, Mar. 2015.

[11] R. N. Lemaitre *et al.*, "Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium.," *PLoS Genet.*, vol. 7, no. 7, p. e1002193, Jul. 2011.

[12] K. Suhre *et al.*, "Human metabolic individuality in biomedical and pharmaceutical research.," *Nature*, vol. 477, no. 7362, pp. 54–60, Sep. 2011.