Chapter 15
Detection of Genome-wide Selection Signatures

1. Sewall Wright's F Statistics

Wrights (1951) introduced the *F* statistics to describe the evolutionary behaviors of structured populations. The three F statistics are F_{IT} , F_{ST} and F_{IS} , where F_{IT} is called the correlation coefficient of the two alleles within individuals, F_{ST} is the correlation coefficient between two alleles from different individuals within a population and F_{IS} is the correlation of the two alleles within individuals within a population (Weir and Cockerham 1984). It is hard to understand the definitions before we use a hierarchical ANOVA model to describe them.

Malecot (1948) defined the correlation coefficient between two alleles as the probability that the two alleles are "identical by descent" (IBD). Such a definition is also called the fixation index. Under this definition, F_{IT} is the probability that the two alleles from the same individuals are IBD, with the "base population" defined as the ancestors of all individuals in all populations. F_{IS} is the probability that the two alleles from the same individuals are IBD, with the "base population" defined as the population when isolation just started. F_{ST} is the probability that a random allele from one population is IBD with a random allele from another population.

When the *F* statistics are defined as fixation indices, 1 – *F* becomes a heterozygosity reduction. The relationship of the three F statistics in terms of herterozygosity reduction is depicted in Figure 1. Let H_0 be the heterozygosity (proportion of heterozygotes) in the beginning (at time t_0). The heterozygosity at the point where the population split into two populations (at time t_1) is $H_1 = (1 - F_{ST})H_0$. The heterozygosity in the end (at time t_2) is $H_2 = (1 - F_{IS})H_1 = (1 - F_{IS})(1 - F_{ST})H_0$. Since we also know that $H_2 = (1 - F_{IT})H_0$. Therefore,

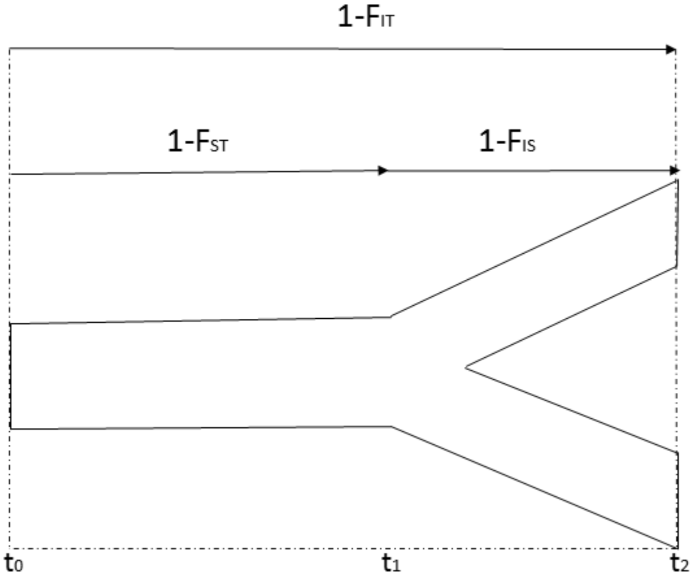
$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \tag{1}$$

This is the very famous equation in population genetics and evolution. These F-statistics can be estimated from molecular data using Cockerham's (1969) hierarchical analysis of variances (ANOVA). Cockerham (1969) used an (f, F, θ) notation, where $f = F_{IS}$, $F = F_{IT}$ and $\theta = F_{ST}$.

The current molecular technology allows us to sequence the entire genome of many species. The variation of DNA sequence data is represented by single nucleotide polymorphism (SNP) markers. Using SNP markers, we can estimate the *F* statistics for every locus of the entire genome. If the F statistics of some loci behave differently from the majority of the loci, some evolution forces, e.g., selection, may have caused the deviation of these loci from the expected population differentiation. Regions of the genome covering these loci are called selection signatures. In this chapter, we will discuss the F statistics and methods for detection of selection signatures.

Figure 1

Relationship of Wright's F statistics: $1-F_{ST}$ represent heterozygosity reduction from t_0 to t_1, $1-F_{IS}$ represents heterozygosity reduction from t_1 to t_2, $1-F_{IT}$ represents heterozygosity reduction from t_0 to t_2. Therefore, $(1-F_{IT})=(1-F_{IS})(1-F_{ST})$.



2. Hierarchical linear model for alleles

Let $i=1,...,p$ indexes population, $j=1,...,n_i$ indexes individual within the *i*th population and $k=1,2$ indexes allele within the same individual. Let

$$y_{ijk}=\begin{cases}1 & \text{for } A_1\\ 0 & \text{for } A_2\end{cases} \tag{2}$$

be the allelic indicator for the *k*th allele of the *j*th individual within population *i*. Cockerham (1969) used the following linear model to describe this allelic indicator variable,

$$y_{ijk}=\mu+\alpha_i+\beta_{(i)j}+\gamma_{(ij)k} \tag{3}$$

where μ is the population mean (frequency of A_1 in the whole population), α_i is the mean value of *y* in the *i*th population expressed as deviation from the mean of the whole population, $\beta_{(i)j}$ is the mean value of *y* for the *j*th individual within the *i*th population expressed as a deviation from the mean of this population and $\gamma_{(ij)k}$ is the residual (the allelic indicator expressed as deviation from all previous terms). Let us assume that all terms in the model except μ are random so that $\mathrm{var}(\alpha_i) = \sigma_\alpha^2$, $\mathrm{var}(\beta_{(i)j}) = \sigma_\beta^2$ and $\mathrm{var}(\gamma_{(ij)k}) = \sigma_\gamma^2$. Therefore, the total variance of *y* is

$$\mathrm{var}(y_{ijk}) = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 \tag{4}$$

Let us look at the definition of the F statistics in terms of correlation coefficients. F_{IT} is the correlation coefficient between the two alleles with an individual. Statistically, this correlation is defined as

$$r_{y_{ijk}y_{ijk'}} = \frac{\mathrm{cov}(y_{ijk}, y_{ijk'})}{\sqrt{\mathrm{var}(y_{ijk})\,\mathrm{var}(y_{ijk'})}} \tag{5}$$

where $k' \neq k$ and both *y*'s have the same subscript *j* (meaning the same individual). The denominator is simply $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2$. The numerator is

$$\begin{aligned}\mathrm{cov}(y_{ijk}, y_{ijk'}) &= \mathrm{cov}(\mu + \alpha_i + \beta_{(i)j} + \gamma_{(ij)k}, \mu + \alpha_i + \beta_{(i)j} + \gamma_{(ij)k'}) \\ &= \mathrm{cov}(\alpha_i, \alpha_i) + \mathrm{cov}(\beta_{(i)j}, \beta_{(i)j}) \\ &= \mathrm{var}(\alpha_i) + \mathrm{var}(\beta_{(i)j}) \\ &= \sigma_\alpha^2 + \sigma_\beta^2\end{aligned} \tag{6}$$

Therefore,

$$F_{IT} = r_{y_{ijk}y_{ijk'}} = \frac{\sigma_\alpha^2 + \sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} \tag{7}$$

The F_{ST} parameter is defined as the correlation coefficient between two alleles from two different individuals within the same population. This correlation is

$$r_{y_{ijk}y_{ij'k'}} = \frac{\mathrm{cov}(y_{ijk}, y_{ij'k'})}{\sqrt{\mathrm{var}(y_{ijk})\,\mathrm{var}(y_{ij'k'})}} \tag{8}$$

The variances in the denominator are all the same, i.e., $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2$. The covariance in the numerator is

$$\mathrm{cov}(y_{ijk}, y_{ij'k'}) = \mathrm{cov}(\mu + \alpha_i + \beta_{(i)j} + \gamma_{(ij)k}, \mu + \alpha_i + \beta_{(i)j'} + \gamma_{(ij)k'}) = \mathrm{cov}(\alpha_i, \alpha_i) = \sigma_\alpha^2 \tag{9}$$

Therefore,

$$F_{ST} = r_{y_{ijk}y_{ij'k'}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} \tag{10}$$

The F_{IS} parameter is defined as the correlation coefficient between two alleles from the same individual within the same population. This means you only consider one population and the parameter is the average of the parameter across all populations. This time we revise the model by focusing on one population only and thus drop subscript *i*,

$$y_{jk} = \mu + \beta_j + \gamma_{(j)k} \tag{11}$$

The total variance here is the variance among all individuals with the same population, which is

$$\mathrm{var}(y_{jk}) = \mathrm{var}(\beta_j) + \mathrm{var}(\gamma_{(j)k}) = \sigma_\beta^2 + \sigma_\gamma^2 \tag{12}$$

The correlation coefficient between the two alleles from the same individual is

$$r_{y_{jk}y_{jk'}} = \frac{\mathrm{cov}(y_{jk}, y_{jk'})}{\sqrt{\mathrm{var}(y_{jk})\,\mathrm{var}(y_{jk'})}} \tag{13}$$

The variances in the denominator are all the same, i.e., $\sigma_\beta^2 + \sigma_\gamma^2$. The covariance in the numerator is

$$\mathrm{cov}(y_{jk}, y_{jk'}) = \mathrm{cov}(\mu + \beta_j + \gamma_{(j)k}, \mu + \beta_j + \gamma_{(j)k'}) = \mathrm{cov}(\beta_j, \beta_j) = \sigma_\beta^2 \tag{14}$$

Therefore,

$$F_{IS} = r_{y_{jk}y_{jk'}} = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\gamma^2} \tag{15}$$

We can rearrange the following equation

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \tag{16}$$

into

$$F_{IS} = \frac{F_{IT} - F_{ST}}{1 - F_{ST}} \tag{17}$$

and substitute the F statistics in the right hand side by the variance ratios. This manipulation will help you verify the equation, as shown below.

$$\begin{aligned} F_{IS} &= \frac{F_{IT} - F_{ST}}{1 - F_{ST}} \\ &= \frac{\dfrac{\sigma_\alpha^2 + \sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} - \dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}}{1 - \dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}} \\ &= \frac{\dfrac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}}{\dfrac{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} - \dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}} \\ &= \frac{\dfrac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}}{\dfrac{\sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}} \\ &= \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\gamma^2} \end{aligned} \tag{18}$$

3. Analysis of variances of alleles

Given the linear model described above, we can sample allelic data from individuals of populations. Again, each data point is an allelic state, a binary variable taking either 0 or 1. In SNP data, there are only two alleles (multiple alleles are very rare) and the "reference allele" is coded 1. Which allele is the reference allele in entirely arbitrary, depending on the investigator's preference. The ANOVA table is shown in Table 1.

Table 1

The ANOVA table for the hierarchical model of a structured whole population.

Variation	df	SS	MS	E(MS)
Between populations	$df_\alpha = p-1$	SS_α	$MS_\alpha = SS_\alpha / df_\alpha$	$\sigma_\gamma^2 + 2\sigma_\beta^2 + 2k_0\sigma_\alpha^2$
Between individuals within populations	$df_\beta = \sum_{i=1}^{p}(n_i - 1)$	SS_β	$MS_\beta = SS_\beta / df_\beta$	$\sigma_\gamma^2 + 2\sigma_\beta^2$
Between alleles within individuals	$df_\gamma = \sum_{i=1}^{p} n_i$	SS_γ	$MS_\gamma = SS_\gamma / df_\gamma$	σ_γ^2

In the above ANAVO table, when the number of individuals within a population is different across different population, the data are called unbalanced. In fact, in population differentiation analysis, population are always unbalanced. The "average" number of individuals of a population is calculated differently from the usual definition of average. It is calculated using

$$k_0 = \frac{1}{p-1}\left(n. - \frac{1}{n.}\sum_{i=1}^{p} n_i^2 \right) \tag{19}$$

where $n. = \sum_{i=1}^{p} n_i$ is the total number of individuals. The three variance components are then estimated using

$$\begin{aligned} \hat{\sigma}_\gamma^2 &= MS_\gamma \\ \hat{\sigma}_\beta^2 &= \frac{1}{2}(MS_\beta - MS_\gamma) \\ \hat{\sigma}_\alpha^2 &= \frac{1}{2k_0}(MS_\alpha - MS_\beta) \end{aligned} \tag{20}$$

The three estimated variance components are used to infer the F statistics. The three sum of squares in the ANOVA are calculated using

$$\begin{aligned} SS_\alpha &= 2\sum_{i=1}^{p} n_i (\bar{y}_{i..} - \bar{y}_{...})^2 = 2\left(\sum_{i=1}^{p} n_i \bar{y}_{i..}^2 - n.\bar{y}_{...}^2 \right) \\ SS_\beta &= 2\sum_{i=1}^{p}\sum_{j=1}^{n_i} (\bar{y}_{ij.} - \bar{y}_{i..})^2 = 2\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} \bar{y}_{ij.}^2 - \sum_{i=1}^{p} n_i \bar{y}_{i..}^2 \right) \\ SS_\gamma &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\sum_{k=1}^{2} (y_{ijk} - \bar{y}_{ij.})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{n_i} \bar{y}_{ij.}^2 \end{aligned} \tag{21}$$

where the bar notations in the SS indicate various means and they are defined

$$
\begin{aligned}
\bar{y}_{...} &= \frac{1}{2n.}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk} \\
\bar{y}_{i..} &= \frac{1}{2n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{2} y_{ijk}, \forall i = 1,..., p \\
\bar{y}_{ij.} &= \frac{1}{2}\sum_{k=1}^{2} y_{ijk}, \forall i = 1,..., p \text{ and } j = 1,..., n_i
\end{aligned}
\tag{22}
$$

In reality, you do not need to know the formulas to perform the ANOVA. The SAS package contains several procedures for the hierarchical ANOVA. There is even a procedure called PROC ALLELE particularly designed for estimation of F statistics.

4. Multiple levels of hierarchy

We now use an experimental mouse population to describe the hierarchy of the alleles. The wheal running experiments in house mice was conducted for 70 generations. We collected DNA samples from 80 female mice at generation 61, 10 mice from each replicated lines. Lines 1, 2, 4 and 5 were the control lines (C) and lines 3, 6, 7 and 8 were the high runner (HR) selection lines. Eight mice were eliminated from the analysis because of low quality SNP callings (1 from line 2, 1 from line 5, 2 from line 3, 1 from line 6 and 3 from line 8). Of the 77808 SNPs, 52476 SNPs were deleted due to missing values or monomorphism across the samples. Therefore, the data set subject to analysis has 72 female mice with 25332 SNPs. These SNPs were evenly distributed across 19 autosomes and the X chromosome. These SNPs also included one from mitochondria and 13 from P elements. The SNP alleles were numerically coded as 1 for the reference allele and 0 for the alternative allele. As a result, there were $72\times 2=144$ observations (one per allele) for each locus analyzed.

Let y_{ijkl} be the indicator variable (0 or 1) for the *l*th allele of the *k*th individual from the *j*th subpopulation within the *i*th population, where $l=1,2$ for the two alleles of each individual, $k=1,...,10$ for the 10 individuals within each subpopulation, $j=1,2,3,4$ for the four subpopulations within each population and $i=1,2$ for the two populations (control and selected populations). Let A_1 be the “reference” allele and A_2 be the alternative allele of a locus under consideration. Denote the whole population frequency of A_1 by p and the frequency of A_2 by $q=1-p$. The allelic indicator variable for reference allele A_1 is

$$
y_{ijkl} = \begin{cases} 1 & \text{for } A_1 \\ 0 & \text{for } A_2 \end{cases}
\tag{23}
$$

which is a Bernoulli variable and thus the expectation is identical to the frequency of the reference allele. We now use Cockerham’s (1969) linear model to describe y_{ijkl},

$$
y_{ijkl} = \mu + \alpha_i + \beta_{(i)j} + \gamma_{(ij)k} + \varepsilon_{(ijk)l}
\tag{24}
$$

where $\mu = p$ is the overall mean (frequency of A_1 for the whole population) $\alpha_i = p_i - p$ is the allele frequency of population *i* expressed as deviation from that of the whole population, $\beta_{(i)j} = p_{ij} - p_i$ is the allele frequency of the *j*th subpopulation expressed as a deviation from the *i*th population, $\gamma_{(ij)k} = p_{ijk} - p_{ij}$ is the allele frequency of the *k*th individual expressed as a deviation from the *j*th subpopulation within the *i*th population, and $\varepsilon_{(ijk)l} = y_{ijkl} - p_{ijk}$ is the residual error. Note that the allele frequency of an individual is defined as $p_{ijk} = (y_{ijk1} + y_{ijk2})/2$, which only takes three possible values, 0, 0.5 and 1. The two populations were not randomly sampled and they were designed by the investigators prior to the experiment. Therefore α_i should be treated as fixed effect. However, the Cockerham's model is random and thus we will take the random model approach as review of the background of population differentiation. The model contains only one fixed effect (μ) and thus it is called the random model. All other effects are random with mean zero and different variances. The variances are denoted by σ_α^2 for effect α_i, σ_β^2 for effect $\beta_{(i)j}$, σ_γ^2 for effect $\gamma_{(ij)k}$ and σ_ε^2 for residual $\varepsilon_{(ijk)l}$. The expectation of y_{ijkl} is $E(y_{ijkl}) = \mu$ and the variance of y_{ijkl} is

$$\mathrm{var}(y_{ijkl}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \tag{25}$$

Cockerham (1969) defined three Wright's F-statistics (Wright 1951) based on these variance components. For the three-level hierarchical model, there are four F-statistics, which are defined as (Yang 1998),

$$F_{IT} = \frac{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{26}$$

$$F_{POP} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{27}$$

$$F_{SUB} = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{28}$$

$$F_{IS} = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{29}$$

These F-statistics are different from the F-statistics developed by Weir and Cockerham (Weir and Cockerham 1984) but they have a nice property of

$$(1 - F_{IT}) = (1 - F_{POP})(1 - F_{SUB})(1 - F_{IS}) \tag{30}$$

If we ignore the populations by treating all subpopulations as populations, we have

$$(1 - F_{ST}) = (1 - F_{POP})(1 - F_{SUB})$$

which leads to

$$F_{ST} = \frac{\sigma_\alpha^2 + \sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{31}$$

This is the typical F_{ST} in the two-level hierarchical population subdivision model, where all subpopulations are promoted to populations and $\sigma_\alpha^2 + \sigma_\beta^2$ represents the variance of the promoted populations. The VARCOMP and MIXED procedures in SAS can estimate variance components, from which various F statistics can be computed. In the PROC VARCOMP and PROC MIXED statements, the method = option should be either TYPE1 or MIVQUE0. The default method is REML and it is not appropriate to use REML because the response variable is not normally distributed.

5. Detection of selection signatures

The experimental populations (treatment and control) of mice presented early were not random populations. Because of this, it is more appropriate to treat α_i as a fixed effect. Therefore, the model defined in equation (24) is a mixed model. Under the mixed model, the expectation of y_{ijkl} is

$$E(y_{ijkl}) = \mu + \alpha_i \quad (32)$$

and the variance of y_{ijkl} is

$$\text{var}(y_{ijl}) = \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \quad (33)$$

Our purpose of the population differentiation study is to test the null hypothesis

$$H_0 : \alpha_1 = \alpha_2 \quad (34)$$

which does not require the F-statistics but we do need the original variance components to facilitate the hypothesis test. We proposed to use the MIVQUE method of Rao (1971b) to estimate the variance components because normal distributions of the random effects and the residual errors are not required with MIVQUE.

It is much more convenient to use a matrix notation to derive the MIVQUE of variance components, as shown below

$$y = X_\mu \mu + X_\alpha \alpha + Z_\beta \beta + Z_\gamma \gamma + \varepsilon \quad (35)$$

where X_μ is an $n \times 1$ vector of unity, X_α is an $n \times 1$ vector whose elements are 1 for individuals in the selected population and -1 for individuals in the control population, $\alpha = \alpha_1 - \alpha_2$ is the difference of allele frequencies between the control and the selected populations, Z_β is an $n \times 8$ incidence matrix representing the 8 lines, β is a 8×1 vector of allele frequencies for the 8 lines, Z_γ is an $n \times 72$ incidence matrix for the 72 mice (38 from the control lines and 34 from the selected lines), γ is an 72×1 vector for individual effects and ε is an 144×1 vector of residuals. All random effects have expectations of zero and a variance σ_β^2 for β, a variance σ_γ^2 for γ and a variance σ_ε^2 for ε.

The expectation and variance of the model are

$$E(y) = X_\mu \mu + X_\alpha \alpha \quad (36)$$

and

$$\text{var}(y) = V = Z_\beta Z_\beta^T \sigma_\beta^2 + Z_\gamma Z_\gamma^T \sigma_\gamma^2 + I\sigma_\varepsilon^2 \quad (37)$$

The MIVQUE of the three variance components $\theta = \left\{\sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2\right\}$ are obtained using the following linear equation system $H_{3\times3}\theta_{3\times1} = Q_{3\times1}$, the details of which are

$$\begin{bmatrix} H_{\beta\beta} & H_{\beta\gamma} & H_{\beta\varepsilon} \\ H_{\gamma\beta} & H_{\gamma\gamma} & H_{\gamma\varepsilon} \\ H_{\varepsilon\beta} & H_{\varepsilon\gamma} & H_{\varepsilon\varepsilon} \end{bmatrix} \begin{bmatrix} \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\varepsilon^2 \end{bmatrix} = \begin{bmatrix} Q_\beta \\ Q_\gamma \\ Q_\varepsilon \end{bmatrix} \tag{38}$$

where the right hand sides of the equations are various quadratic forms of *y* and the left hand sides are the expectations of the quadratic forms. Let us define $X = [X_\mu \| X_\alpha]$ as vertical concatenation of the two matrices in the brackets and $\eta = \begin{bmatrix} \mu & \alpha \end{bmatrix}^T$ as the fixed effects. Further define $P = I - X(X^T X)^- X^T$, $V_\beta = PZ_\beta$, $V_\gamma = PZ_\gamma$ and $V_\varepsilon = PI = P$. The six unique elements of the *H* matrix are

$$H_{\beta\beta} = \mathrm{tr}(V_\beta V_\beta^T V_\beta V_\beta^T)$$
$$H_{\beta\gamma} = \mathrm{tr}(V_\beta V_\beta^T V_\gamma V_\gamma^T)$$
$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T)$$
$$H_{\gamma\gamma} = \mathrm{tr}(V_\gamma V_\gamma^T V_\gamma V_\gamma^T)$$
$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T)$$
$$H_{\varepsilon\varepsilon} = \mathrm{tr}(V_\varepsilon V_\varepsilon^T V_\varepsilon V_\varepsilon^T) = n - 1$$

The remaining three elements of *H* take the three corresponding elements with flipping subscripts because the matrix is symmetrical. The three elements of the *Q* matrix are

$$Q_\beta = y^T V_\beta V_\beta^T y$$
$$Q_\gamma = y^T V_\gamma V_\gamma^T y$$
$$Q_\varepsilon = y^T V_\varepsilon V_\varepsilon^T y = y^T P y$$

The MIVQUE estimate of the parameter vector θ is $\hat{\theta} = H^{-1} Q$. Note that the MIVQUE estimate of a variance component can be negative because of the unbiased nature of the estimate. If that happens, it is simply set to zero.

The estimated variance components, denoted by $\hat{\theta} = \left\{ \hat{\sigma}_\beta^2, \hat{\sigma}_\gamma^2, \hat{\sigma}_\varepsilon^2 \right\}$, are then used to estimate the fixed effects and perform hypothesis tests. The estimated variance matrix of *y* is

$$\mathrm{var}(y) = \hat{V} = Z_\beta Z_\beta^T \hat{\sigma}_\beta^2 + Z_\gamma Z_\gamma^T \hat{\sigma}_\gamma^2 + I \hat{\sigma}_\varepsilon^2 \tag{39}$$

The best linear unbiased estimate (BLUE) of the fixed effect is

$$\hat{\eta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y \tag{40}$$

and the variance matrix of this estimate is

$$\mathrm{var}(\hat{\eta}) = V_\eta = (X^T \hat{V}^{-1} X)^{-1} \tag{41}$$

Note that

$$\hat{\eta} = \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} \text{ and } V_\eta = \begin{bmatrix} \mathrm{var}(\hat{\mu}) & \mathrm{cov}(\hat{\mu}, \hat{\alpha}) \\ \mathrm{cov}(\hat{\alpha}, \hat{\mu}) & \mathrm{var}(\hat{\alpha}) \end{bmatrix}$$

The *F* test for $H_0 : \alpha = 0$ is

$$F = \frac{\hat{\alpha}^2}{\mathrm{var}(\hat{\alpha})} \tag{42}$$

with degrees of freedom 1 (numerator) and 6 (denominator). The p-value is calculated using

$$p = 1 - \Pr(f_{1,6} < F) \tag{43}$$

where $f_{1,6}$ is a random variable of *F* distribution with 1 and 6 degrees of freedom. The *p* value is then converted into –log$_{10}$(*p*), which is used the Manhattan plots. Although the F test appears to be defined the same as the Wald test, it does not follow a Chi-square distribution. Therefore, the p value must be calculated from the F distribution.

6. PROC MIXED and PROC ALLELE

6.1. Example of single locus analysis

We now use an example to demonstrate the MIVQUE0 method described above. The locus demonstrated is located on chromosome 12 at position 90165856bp with a marker identification number UNC2173488. Upon deleting eight mice with low quality SNP callings, there were 72 mice left in the population. The input file for the allelic model has $72 \times 2 = 144$ rows (two rows per mouse) and seven columns. This file contains information about the populations, subpopulations, mouse identifications and the nucleotide type of the allele along with numerically coded allelic values (*y*). The file is given in “mouse1.xlsx”.

First, let us use

The model is

$$y = X_\mu \mu + X_\alpha \alpha + Z_\beta \beta + Z_\gamma \gamma + \varepsilon \tag{44}$$

where X_μ is an $n \times 1 = 144 \times 1$ vector of unity, X_α is an 144×1 vector whose elements are 1 for individuals in the selected population and -1 for individuals in the control population, $\alpha = \alpha_1 - \alpha_2$ is the difference of allele frequencies between the control and the selected populations, Z_β is an 144×8 incidence matrix representing the 8 lines (subpopulations), β is a 8×1 vector of allele frequencies for the 8 lines, Z_γ is an 144×72 incidence matrix for the 72 mice (38 from the control lines and 34 from the selected lines), γ is an 72×1 vector for individual effects and ε is an 144×1 vector of residuals. All random effects have expectations of zero and a variance σ_β^2 for β, a variance σ_γ^2 for γ and a variance σ_ε^2 for ε. The parameters in the model include fixed effects $\eta = \{\mu, \alpha\}$ and variance components $\theta = \{\sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2\}$. The expectation and variance of the model are

$$E(y) = X_\mu \mu + X_\alpha \alpha \tag{45}$$

and

$$\mathrm{var}(y) = V = Z_\beta Z_\beta^T \sigma_\beta^2 + Z_\gamma Z_\gamma^T \sigma_\gamma^2 + I \sigma_\varepsilon^2 \tag{46}$$

The MIVQUE of the three variance components $\theta = \left\{\sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2\right\}$ are obtained using the following linear equation system $H_{3\times3}\theta_{3\times1} = Q_{3\times1}$, the details of which are

$$\begin{bmatrix} H_{\beta\beta} & H_{\beta\gamma} & H_{\beta\varepsilon} \\ H_{\gamma\beta} & H_{\gamma\gamma} & H_{\gamma\varepsilon} \\ H_{\varepsilon\beta} & H_{\varepsilon\gamma} & H_{\varepsilon\varepsilon} \end{bmatrix} \begin{bmatrix} \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\varepsilon^2 \end{bmatrix} = \begin{bmatrix} Q_\beta \\ Q_\gamma \\ Q_\varepsilon \end{bmatrix} \tag{47}$$

where the right hand sides of the equations are various quadratic terms of *y* and the left hand sides are the expectations of the quadratic terms. What we need here is to find all quantities in the above equation set. First, we need to define $X = [X_\mu \,\|\, X_\alpha]$ as horizontal concatenation of the two matrices in the brackets (an 144×2 matrix) and $\eta = \left[\mu \quad \alpha\right]^T$ as the fixed effects. Further define $P = I - X(X^T X)^- X^T$, $V_\beta = PZ_\beta$, $V_\gamma = PZ_\gamma$ and $V_\varepsilon = PI = P$. These matrices are all of dimension 144×144. The six unique elements of the *H* matrix from the example are

$$H_{\beta\beta} = \mathrm{tr}(V_\beta V_\beta^T V_\beta V_\beta^T) = 1950.0893$$

$$H_{\beta\gamma} = \mathrm{tr}(V_\beta V_\beta^T V_\gamma V_\gamma^T) = 215.3065$$

$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T) = 107.6533$$

$$H_{\gamma\gamma} = \mathrm{tr}(V_\gamma V_\gamma^T V_\gamma V_\gamma^T) = 280.0000$$

$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T) = 140.0000$$

$$H_{\varepsilon\varepsilon} = \mathrm{tr}(V_\varepsilon V_\varepsilon^T V_\varepsilon V_\varepsilon^T) = 143.0000$$

The remaining three elements of *H* take the three corresponding elements with flipping subscripts because the matrix is symmetrical. The three elements of the *Q* matrix are

$$Q_\beta = y^T V_\beta V_\beta^T y = 37.19247$$

$$Q_\gamma = y^T V_\gamma V_\gamma^T y = 15.34520$$

$$Q_\varepsilon = y^T V_\varepsilon V_\varepsilon^T y = y^T P y = 13.17260$$

The MIVQUE estimate of the parameter vector θ is $\hat{\theta} = H^{-1}Q$, which is called the MIVQUE(0) estimate. The estimated variance components for this particular example are

$$\begin{bmatrix} \hat{\sigma}_\beta^2 \\ \hat{\sigma}_\gamma^2 \\ \hat{\sigma}_\varepsilon^2 \end{bmatrix} = \begin{bmatrix} 1950.089 & 215.3065 & 107.6533 \\ 215.3065 & 280 & 140 \\ 107.6533 & 140 & 143 \end{bmatrix}^{-1} \begin{bmatrix} 37.19247 \\ 15.3452 \\ 13.1726 \end{bmatrix} = \begin{bmatrix} 0.014229 \\ 0.006191 \\ 0.075342 \end{bmatrix}$$

The estimated variance components are then used to estimate the fixed effects and perform hypothesis tests. The estimated variance matrix of *y* is

$$\mathrm{var}(y) = \hat{V} = Z_\beta Z_\beta^T \hat{\sigma}_\beta^2 + Z_\gamma Z_\gamma^T \hat{\sigma}_\gamma^2 + I\hat{\sigma}_\varepsilon^2$$

which is an 144×144 matrix. The best linear unbiased estimate (BLUE) of the fixed effect is

$$\hat{\eta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y$$

For this particular example,

$$\begin{bmatrix}\hat{\mu}\\ \hat{\alpha}\end{bmatrix}=\begin{bmatrix}417.7194 & 205.5873\\ 205.5873 & 205.5873\end{bmatrix}^{-1}\begin{bmatrix}193.46297\\ 12.20972\end{bmatrix}=\begin{bmatrix}0.8544359\\ \text{-}0.7950464\end{bmatrix}$$

and the variance matrix of this estimate is

$$\operatorname{var}\begin{bmatrix}\hat{\mu}\\ \hat{\alpha}\end{bmatrix}=\begin{bmatrix}417.7194 & 205.5873\\ 205.5873 & 205.5873\end{bmatrix}^{-1}=\begin{bmatrix}0.004714044 & \text{-}0.004714044\\ \text{-}0.004714044 & 0.009578158\end{bmatrix}$$

The *F* test for $H_0:\alpha=0$ is

$$F=\frac{\hat{\alpha}^2}{\operatorname{var}(\hat{\alpha})}=\frac{(\text{-}0.7950464)^2}{0.009578158}=65.99377$$

with degrees of freedom 1 (numerator) and 6 (denominator). The *p*-value is calculated using

$$p=1-\Pr(f_{1,6}<F)=0.0001868454$$

which translates into

$$-\log 10(p)=3.728518$$

Several SAS procedures can generate the same result as described above. PROC MIXED is the most advanced procedure to perform such an analysis.
is the simplest procedure of all to do the analysis. First, let us use PROC MIXED to test the difference between the treatment and the control. Part of the input data is shown in Table 2.

Table 2

Part of the mouse data in a format that is required by PROC MIXED

obs	pop	sub	mouse	allele	type	y
1	0	1	64358	1	A	1
2	0	1	64358	2	A	1
3	0	1	64419	1	A	1
4	0	1	64419	2	A	1
5	0	1	64423	1	G	0
6	0	1	64423	2	A	1
7	0	1	64426	1	G	0
8	0	1	64426	2	G	0
9	0	1	64431	1	G	0
10	0	1	64431	2	A	1
11	0	1	64435	1	A	1
12	0	1	64435	2	A	1
13	0	1	64442	1	G	0
14	0	1	64442	2	A	1
15	0	1	64450	1	A	1
16	0	1	64450	2	A	1
17	0	1	64724	1	A	1
18	0	1	64724	2	A	1
19	0	1	64737	1	A	1
20	0	1	64737	2	A	1

The following SAS code creates the data and estimate variance components.

```
%let dir=C:\Users\STAT231B\Text\Chapter 23;
filename aa "&dir\data\mouse1.xlsx";
proc import datafile=aa dbms=xlsx out=mydata replace;
run;

proc mixed data=mydata method=mivque0;
   class pop sub mouse;
   model y = pop /solution;
   random sub mouse(sub);
run;
```

The outputs are represented by the following tables (Tables 3 and 4):

Table 3
Estimated variance components from PROC MIXED with the METHDOD = MIVQUE0 option

Covariance Parameter Estimates	
Cov Parm	**Estimate**
sub	0.01423
mouse(sub)	0.005668
Residual	0.07639

Table 4
Hypothesis test for the difference between the treatment and control populations

Solution for Fixed Effects

Effect	pop	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.05939	0.06974	6	0.85	0.4271
pop	0	0.7950	0.09787	72	8.12	<.0001
pop	1	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
pop	1	72	65.99	<.0001

The three estimated variances from RPOC MIXED are slightly different from the estimates presented early using the linear equation system, but are close enough to be claimed the same. They are compared in the following table,

Table 5
Comparison of estimated variances from PROC MIXED and MIVQUE equation

Cov Parm	PROC MIXED	MIVQUE EQUATION
Sub $\hat{\sigma}_{\beta}^{2}$	0.01423	0.014229
Mouse(Sub) $\hat{\sigma}_{\gamma}^{2}$	0.005668	0.006191
Residual $\hat{\sigma}_{\varepsilon}^{2}$	0.07639	0.075342

The estimated population differences of the two approaches (PROC MIXED and MIVQUE EQUATION) are the same ($\hat{\alpha} = 0.7950$) and the F test statistic is also the same (65.99). However, the p-values are different because PROC MIXED uses 72 as the denominator degrees of freedom while the actual denominator degrees of freedom should be 6.

To compare PROC MIXED and PROC ALLELE, we now ignore the two populations and treat all eight subpopulations as the populations. This will keep the hierarchy of populations, individuals within populations and alleles within individuals so that PROC ALLELE can handle this level of hierarchy. Both PROC MIXED and PROC VARCOMP take the same set of data and generate the same results of estimated variance components. The SAS codes for the two procedures are

```
proc mixed data=mydata method=type1;
   class sub mouse;
   model y = /solution;
   random sub mouse(sub);
run;

proc varcomp data=mydata method=type1;
   class sub mouse;
   model y = sub mouse(sub);
run;
```

The output from PRCO VARCOMP are shown in Tables 6 and 7.

Table 6
ANOVA table from PROC VARCOMP

Type 1 Analysis of Variance				
Source	**DF**	**SS**	**MS**	**Expected Mean Square**
sub	7	24.758929	3.536990	Var(Error) + 2 Var(mouse(sub)) + 17.968 Var(sub)
mouse(sub)	64	5.678571	0.088728	Var(Error) + 2 Var(mouse(sub))
error	72	5.500000	0.076389	Var(Error)
total	143	35.937500		

Table 7
Estimated variance components from PROC VARCOMP

Type 1 Estimates	
Variance Component	**Estimate**
Var(sub)	0.19191
Var(mouse(sub))	0.0061694
Var(Error)	0.07639

The three variance components are $\hat{\sigma}_\beta^2 = 0.19191$, $\hat{\sigma}_\gamma^2 = 0.0061694$ and $\hat{\sigma}_\varepsilon^2 = 0.07639$. The three F statistics from the variance components are

$$F_{IS} = f = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2} = \frac{0.0061694}{0.0061694 + 0.07639} = 0.074727$$

$$F_{IT} = F = \frac{\sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} = \frac{0.19191 + 0.0061694}{0.19191 + 0.0061694 + 0.07639} = 0.721681$$

$$F_{ST} = \theta = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} = \frac{0.19191}{0.19191 + 0.0061694 + 0.07639} = 0.699204$$

PROC ALLELE requires a different input data format: the two alleles within a locus must occupy two columns. In other words, for *m* loci, the data must have 2m columns, 2 per locus.

Table 8
Part of the mouse data in a format that is required by PROC ALLELE

obs	pop	sub	mouse	allele1	allele2
1	0	1	64358	A	A
2	0	1	64419	A	A
3	0	1	64423	G	A
4	0	1	64426	G	G
5	0	1	64431	G	A
6	0	1	64435	A	A
7	0	1	64442	G	A
8	0	1	64450	A	A
9	0	1	64724	A	A
10	0	1	64737	A	A
11	0	2	64020	A	A
12	0	2	64024	G	A
13	0	2	64365	A	A
14	0	2	64371	G	A
15	0	2	64378	G	A
16	0	2	64382	A	A
17	0	2	64388	G	G
18	0	2	64400	A	A
19	0	2	64681	G	A
20	0	4	64015	A	A

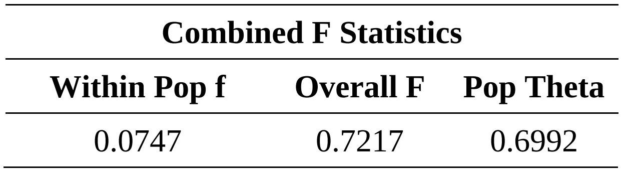
The SAS code for PROC ALLELE is

```
filename bb "&dir\data\mouse2.xlsx";
proc import datafile=bb dbms=xlsx out=mydata2 replace;
run;

proc allele data=mydata2;
    var allele1 allele2;
    pop sub / indivloci;
run;
```

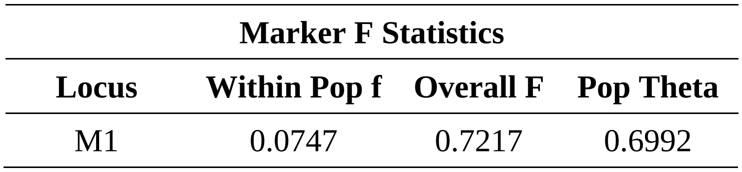
The outputs are represented by two tables (Table 8 and Table 9); the top one are the combined F statistics across all loci and the bottom one are the F statistics for individual loci.

Table 8
Combined F statistics from PROC ALLELE



Combined F Statistics		
Within Pop f	**Overall F**	**Pop Theta**
0.0747	0.7217	0.6992

Table 9
Estimated F statistics from PROC ALLELE



Marker F Statistics			
Locus	**Within Pop f**	**Overall F**	**Pop Theta**
M1	0.0747	0.7217	0.6992

The combined and individual locus F statistics in this case are the same because only one locus was involved. These F statistics are identical to the estimated ones from PROC VARCOMP.

6.2. Example of multiple loci

This data contain 24 SNPs collected from 1397 individuals (human) from 8 populations. Part of the data in a format required by PRCO ALLELE are illustrated in Table 10 (next page).

Table 10
Part of the human data of 24 SNPs with 1397 individuals from 8 populations

OBS	POPID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8
FA-1801	FA	T/T	G/G	G/G	C/C	T/C	T/T	C/T	G/T
FA-1802	FA	T/T	A/A	T/T	C/C	C/C	A/A	C/C	T/T
FA-1803	FA	T/T	A/A	T/T	C/C	C/C	A/A	C/C	T/T
FA-1804	FA	T/T	A/G	G/T	C/C	T/C	T/A	C/C	G/T
FA-1805	FA	T/T	A/G	G/T	C/C	T/C	T/A	C/C	T/T
FA-1806	FA	T/T	A/G		C/C	C/C	T/A	C/T	T/T
FA-1807	FA	T/T	G/G	G/G	C/C	C/C	T/T	C/T	T/T
FA-1808	FA	T/C	G/G		C/C	T/C	T/T	C/T	G/T
FA-1809	FA	T/T	A/G	G/T	C/C	C/C	A/A	C/C	T/T
FA-1810	FA	T/T	A/A		C/T	C/C	T/A	C/C	T/T
FA-1811	FA	T/T	A/A		C/C	C/C	A/A	C/C	T/T
FA-1812	FA	T/T	A/A	T/T	C/C	C/C	A/A	C/C	T/T
FA-1813	FA	T/C	A/G	G/T	C/C	C/C	T/A	C/T	T/T
FA-1814	FA	T/T	A/G		C/C	T/C	T/T	C/C	G/T
FA-1815	FA	T/T	A/G		C/C	C/C	T/A	C/T	T/T
FA-1816	FA	T/T	A/A	T/T	C/C	C/C	A/A	C/C	T/T
FA-1817	FA	T/T	A/G		C/C	T/C	T/A	C/C	T/T
FA-1818	FA	T/T	G/G		C/C	T/T	T/T	C/C	T/T
FA-1819	FA	T/T	A/G	G/T	C/C	C/C	T/A	C/T	T/T
FA-1820	FA	T/T	A/G	G/T	C/C	C/C	T/A	C/T	T/T

The SAS code is shown below,

```
%let dir=C:\Users\SHXU\STAT231B\Text\Chapter 23;
filename aa "&dir\data\human.xlsx";
proc import datafile=aa dbms=xlsx out=mydata replace;
run;

proc allele data=mydata genocol delimiter="/";
    var SNP1-SNP24;
    pop POPID/fperms=1000 indivloci;
    ods output MarkerFStats=fst;
run;
```

The format of the input data is different from the previous example. By default, PROC ALLELE takes a data with each locus occupying to columns. In this format, the locus is coded by genotype rather than by alleles. Therefore, you need the `genocol delimiter="/"` options in the `PROC ALLELE` statement. In addition, the `fperms=1000` option in the `POP` statement randomly shuffles the data to generate an empirical probability statement for the estimated F statistics.

Table 11 shows the combined F statistics from 24 SNP loci. Table 12 lists the F statistics for all individual loci.

Table 11

Combined F statistics from 24 SNP loci estimated by PROC ALLELE for the human data

Combined F Statistics					
Within Pop f	**Pr >Within Pop f**	**Overall F**	**Pr > Overall F**	**Pop Theta**	**Pr > Pop Theta**
0.0630	<.0001	0.0651	<.0001	0.0023	0.0190

All the three F statistics (combined across all loci) are significantly different from zero because the p-values are all smaller than 0.05. Seven of the 24 SNPs shows F_{ST} significant different from zero (Table 12).

Table 12

F statistics of 24 SNPs estimated with PROC ALLELE for the human data

Marker F Statistics						
Locus	**Within Pop f**	**Pr > Within Pop f**	**Overall F**	**Pr > Overall F**	**Pop Theta**	**Pr > Pop Theta**
SNP1	0.1327	<.0001	0.1409	<.0001	0.0094	<.0001
SNP2	0.0378	0.1800	0.0383	0.1860	0.0006	0.3000
SNP3	0.0191	0.5850	0.0202	0.5270	0.0011	0.2920
SNP4	0.0367	0.2300	0.0401	0.0300	0.0036	0.0370
SNP5	0.1428	<.0001	0.1469	<.0001	0.0049	0.0270
SNP6	0.0970	<.0001	0.0999	0.0010	0.0032	0.0640
SNP7	0.0309	0.2880	0.0317	0.2670	0.0009	0.2670
SNP8	0.1938	<.0001	0.1999	<.0001	0.0075	0.0020
SNP9	0.1906	<.0001	0.1965	<.0001	0.0072	0.0030
SNP10	0.0178	0.7110	0.0187	0.6220	0.0009	0.3030
SNP11	0.0390	0.1800	0.0398	0.1450	0.0008	0.2830
SNP12	0.0323	0.3790	0.0356	0.3380	0.0035	0.1400
SNP13	0.0097	0.7930	0.0099	0.6870	0.0002	0.4030
SNP14	0.0588	0.1130	0.0563	0.1260	-0.0026	1.0000
SNP15	-0.0172	1.0000	-0.0109	0.9540	0.0062	0.0230
SNP16	0.0928	0.0010	0.0933	<.0001	0.0006	0.3220
SNP17	0.0116	0.7260	0.0119	0.6800	0.0003	0.3430
SNP18	0.1452	<.0001	0.1574	<.0001	0.0143	<.0001
SNP19	0.0400	0.2870	0.0404	0.2400	0.0004	0.3970
SNP20	-0.0437	0.2510	-0.0429	0.2250	0.0008	0.3220
SNP21	-0.0072	0.8640	-0.0065	0.8230	0.0007	0.3310
SNP22	0.0600	0.0390	0.0599	0.0250	-0.0000	1.0000
SNP23	-0.0094	0.8280	-0.0119	0.7140	-0.0025	1.0000
SNP24	0.0335	0.3750	0.0354	0.3130	0.0020	0.2260

6.3. Detection loci responsible for wheel running behavior of mice

The experiment was described early in this chapter. Here we describe it again to refresh your mind. The wheal running experiments in house mice was conducted for 70 generations. We collected DNA samples from 80 female mice at generation 61, 10 mice from each replicated lines. Lines 1, 2, 4 and 5 were the control lines (C) and lines 3, 6, 7 and 8 were the high runner (HR) selection lines. Eight mice were eliminated from the analysis because of low quality SNP callings (1 from line 2, 1 from line 5, 2 from line 3, 1 from line 6 and 3 from line 8). Of the 77808 SNPs, 52476 SNPs were deleted due to missing values or monomorphism across the samples. Therefore, the data set subject to analysis has 72 female mice with 25332 SNPs. These SNPs were evenly distributed across 19 autosomes and the X chromosome. These SNPs also included one from mitochondria and 13 from P elements. The SNP alleles were numerically coded as 1 for the reference allele and 0 for the alternative allele. As a result, there were $72 \times 2 = 144$ observations (one per allele) for each locus analyzed. The purpose of the study is to detect loci with significant difference in allele frequency between the HR selected population and the control population. Since the number of loci is large, looping over PROC MIXED will take long time to complete the genome scanning. Therefore, we wrote an R program to perform the MIVQUE analysis.

We compared four models: (1) Allelic model: the mixed model described in this chapter where an allele is coded as 1 for the reference allele and 0 otherwise; (2) Genotypic model: the average allelic value (genotypic value) is used as the original data point for the MIVQUE method; (3) Regularized t test (Baldwin-Brown et al. 2014); (4) Regularized regression method. Manhattan plots of the four models are illustrated in Figure 2. The allelic model and the genotypic model generated the same results and both detected a selection signature on chromosome 9. If a less stringent criterion drawn from permutation analysis were used, the two models would detect many more loci responsible for selection. However, the regularized t test and the regression model did not detect any loci, no matter what critical values were used.

QQ-plots (Figure 3) show that the allelic and genotypic models behave as expected. The regularized t test and regression analysis have all points fall around the diagonal line (low power). The ROC curve from a simulated data experiment shows that the allelic model is more power than the regularized test (Figure 4)

Figure 2

Manhattan plots of the mice experiment analyzed from four models

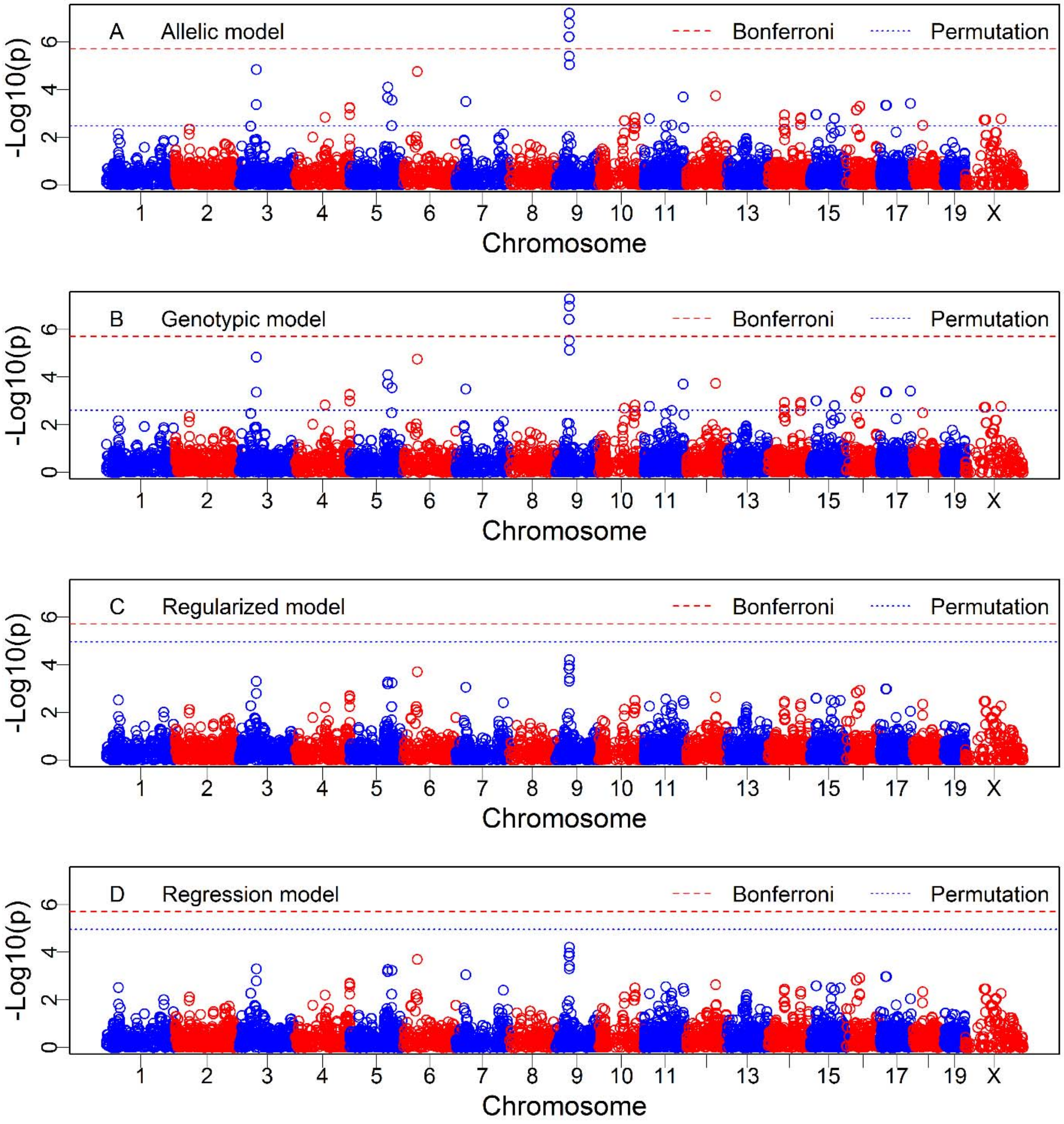


Figure 3

QQ-plots of the mice data from four analytical models

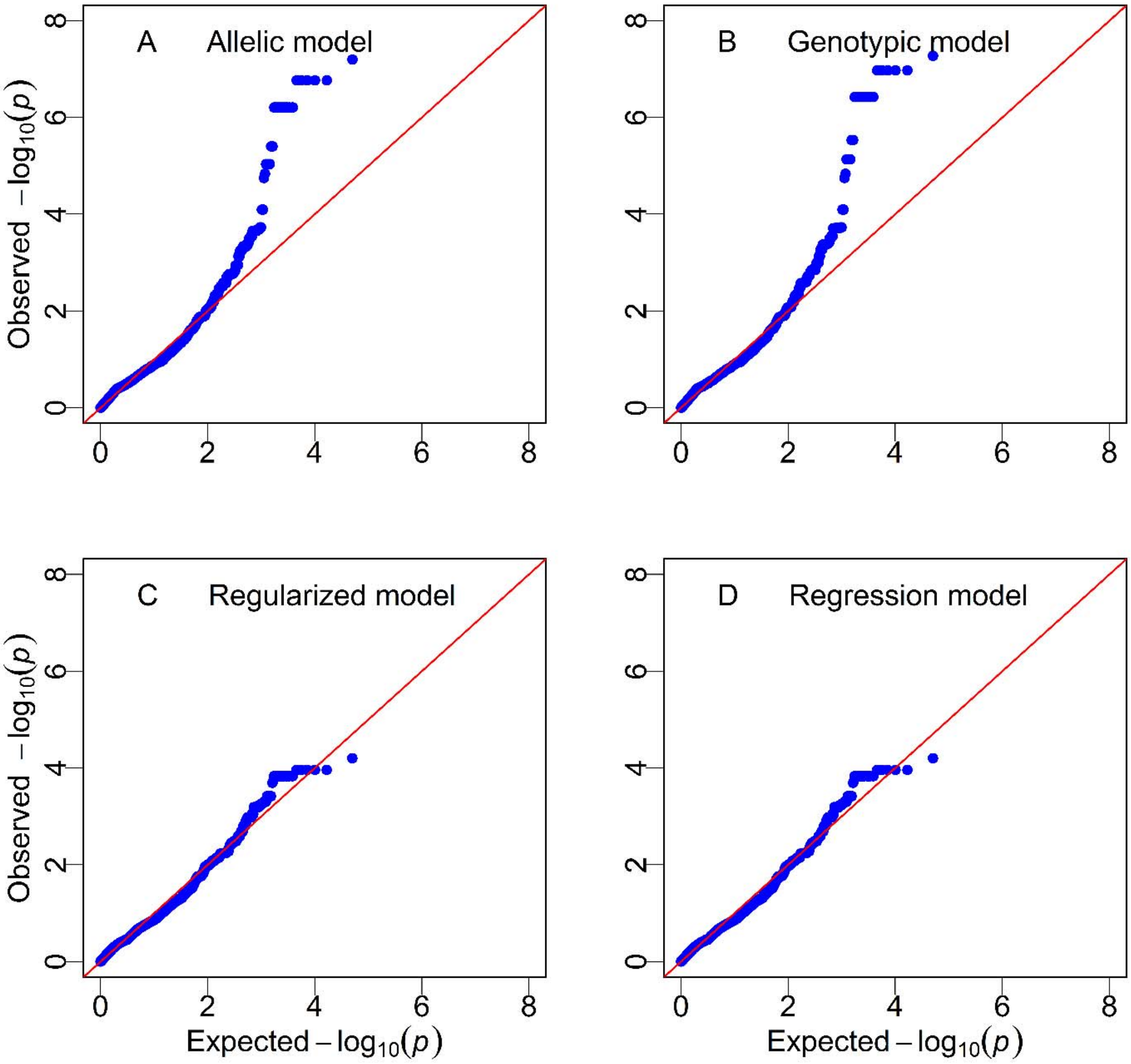
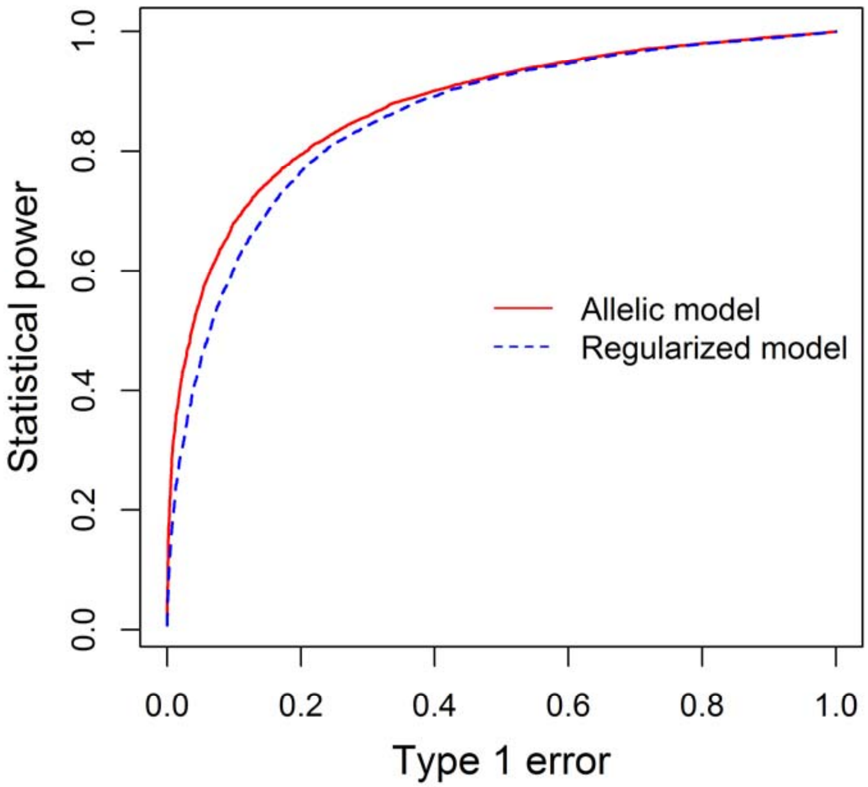


Figure 4

Statistical power plotted against Type 1 error for the allelic model and regularized t test for detection selection signature



Cockerham CC. 1969. Variance of Gene Frequencies. *Evolution* **23**: 72-84.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the analysis of population structure. *Evolution* **38**: 1358-1370.

Wright S. 1951. The genetic structure of populations. *Ann Eugen* **15**: 323-354.

Yang R-C. 1998. Estimating hierarchical F-statistics. *Evolution* **52**: 950-956.

A Mixed Model Approach to Genome-wide Association Studies for Selection Signatures, with Application to Mice Bred for Voluntary Exercise Behavior

Shizhong Xu[*] and Theodore Garland, Jr[†]

[*]Department of Botany and Plant Sciences, University of California, Riverside, CA 92621
[†]Department of Biology, University of California, Riverside, CA 92621

Running title: Mixed Models for GWAS

Key words: Behavior, Experimental evolution, Exercise, F-statistics, Population differentiation

Corresponding author:
Theodore Garland, Jr.
Department of Biology
University of California, Riverside
Riverside, CA 92621
U.S.A.

Phone: 951-827-3524
Fax: 951-827-4286
E-mail: tgarland@ucr.edu

Abstract

Selection experiments and experimental evolution (EE) provide unique opportunities to study the genetics of adaptation because the target and intensity of selection are known relatively precisely. In contrast to natural selection, where populations are never strictly "replicated," EE routinely includes replicate lines so that selection signatures – genomic regions showing excessive differentiation between treatments – can be separated from possible founder effects, genetic drift, and multiple adaptive solutions. We developed a mouse model with four lines within a high runner (HR) selection treatment and four non-selected controls (C). At generation 61, we sampled 10 mice of each line and used the Mega Mouse Universal Genotyping Array to obtain single nucleotide polymorphism (SNP) data for 25,318 SNPs for each individual. Using an advanced mixed model procedure developed in this study, we identified 152 markers that were significantly different in frequency between the two selection treatments. They occurred on all chromosomes except 1, 2, 8, 13, and 19, and showed a variety of patterns in terms of fixation (or the lack thereof) in the four HR and four C lines. Importantly, none were fixed for alternate alleles between the two selection treatments. The current state-of-the-art regularized F test applied after pooling DNA samples for each line failed to detect any markers. We conclude that when SNP or sequence data are available from individuals, the mixed model methodology is recommended for selection signature detection. As sequencing at the individual level becomes increasingly feasible, the new methodology may be routinely applied for detection of selection.

Introduction

Complex traits, such as most behaviors, are affected by alleles segregating at multiple loci. Mapping quantitative trait loci (QTL) for such traits can be difficult, often requiring a large sample. In general, two approaches are used to map QTL, involving use of a designed line cross experiment (LANDER AND BOTSTEIN 1989) or selectively bred populations (WURSCHUM 2012; CUI *et al.* 2015). Use of a line cross experiment requires a large sample size to avoid the Beavis effect (BEAVIS 1994; XU 2003), in which reported QTL effects are often biased and the amount of bias is inversely proportional to the sample size. Moreover, the inference space of QTL parameters is narrow, only applicable to the lines initiating the cross, and the result cannot be extended to crosses derived from other lines (XU 1996). Using selectively bred populations for QTL mapping takes advantage of existing resources with no need to create a line cross. It is also possible to use selected lines to make a mapping cross. QTL detected from a set of selected populations can be directly applied to the same populations to further improve breeding efficiency (WURSCHUM 2012), and results can also be applied to the original starting (base) population from which the selected lines were derived. Another advantage of using selected populations for QTL mapping is that the sample size does not have to be very large because allelic data are used instead of the phenotypic values of a selected trait (CUI *et al.* 2015). The reason for this is that mapping QTL in selected populations takes advantage of the shifts of allele frequencies away from expected Mendelian ratios, i.e., equivalent to detection of segregation distortion, which does not require large sample sizes (LUO AND XU 2003; LUO *et al.* 2005).

Statistical methods for QTL mapping in selected populations often involve the Chi-square test. When replicated lines of a selection experiment are available, a t-test or F-test can be used to detect QTL via comparison of the allele frequencies of the selected population with expected allele frequencies, an approach called detection for segregation distortion (VOGL AND XU 2000; LUO AND XU 2003; LUO *et al.* 2005). If multiple selected populations are involved, then allele frequencies among the populations may be compared, which is called population differential analysis (WEIR AND COCKERHAM 1984; BALDING AND NICHOLS 1995).

Population differential analysis is an important area in population genetics and molecular evolution. Wright proposed three *F* statistics particularly to describe population differentiation, where a whole population is subdivide into two or more subpopulations (WRIGHT 1950). The *F* statistics (not the *F* tests) describe the correlation of alleles at different levels in the population. For example, the correlation coefficient between the two alleles from the sample individual is called the inbreeding coefficient (F_{IT}), the correlation coefficient between two alleles from different individuals in the same subpopulation is called F_{ST}, and the correlation coefficient between the two alleles from the same individual within the same subpopulation is called F_{IS}. In Cockerham and Weir's (COCKERHAM 1969;

WEIR AND COCKERHAM 1984) notation, $F_{IT} = F$, $F_{ST} = \theta$ and $F_{IS} = f$. The three *F* statistics are related by $1 - F = (1-\theta)(1-f)$. The key parameter in population differential analysis is θ. Wright only proposed the concept of *F* statistics and did not address how to estimate them from samples. It was Weir and Cockerham (1984) who developed a systematic approach to estimate these *F* statistics – analysis of variances (ANOVA) by treating the binary indicator (0 or 1) of a reference allele as the response variable. Prior to Weir and Cockerham (1984), much confusion surrounded the relationship between the *F* statistics and correlation coefficients of alleles at different levels of the population hierarchy. Cockerham (1969) discovered that these *F* statistics can actually be expressed as various intra-class correlations (variance ratios) from the analysis of variance. More importantly, one can perform a statistical test for the significance of θ using a non-parametric method, such as the Jackknife, the bootstrap or the permutation method (WEIR AND COCKERHAM 1984). An estimated θ significantly different from that expected from neutrality means that the population differentiation may be caused by some sort of evolutionary forces beyond random genetic drift, e.g., selection. When markers of the entire genome are tested this way, selection signatures can be detected, where a selection signature is defined as a genomic region subject to selection (BROOKFIELD 2001). Conventionally, detection of selection signatures is conducted via population differentiation analysis, and rarely have these applications included replicated lines within the differentiated populations. Without replications, it is difficult to separate selection from drift, and thus false positives may be high.

An alternative and more effective way to investigate selection acting at particular loci is through experimental evolution (GARLAND AND ROSE 2009; BALDWIN-BROWN *et al.* 2014; SCHLOTTERER *et al.* 2015), in which a replicated bi-directional selection experiment or a uni-directional selection experiment with a non-selected control(s) is conducted. In experimental evolution, each treatment population often has multiple replicated lines, which allow separation of selection effects from genetic drift, and thus reduce false positive detections of selection. Although the *F* statistics approach can be applied to selection signature detection from experimental populations, each population represents a treatment level purposely chosen by the investigator and is not a randomly selected level out of a large pool of populations. As a result, the *F* statistics that are based on random selection of populations may not be appropriate. Instead, a mixed model approach may be more appropriate by treating the selection (treatment) effects as fixed (e.g. high-selected, control, low-selected) and effects of replicated lines (subpopulations) as random. Such a mixed model analysis may be more powerful than the F-statistics that are based the assumption of populations being randomly selected. The purpose of the present study is to develop such a mixed model for detection of selection signatures using genome-wide markers in selected and control populations with multiple replicated lines within each population.

To distinguish population differentiation analysis from selection signature detection in experimental evolution, we now use “selection treatments” to represent “populations” and use “replicated lines” within each treatment to represent “subpopulations.” When only two levels of selection treatments (selected treatment and control treatment) are available for comparison, Baldwin-Brown et al (2014) proposed a regularized t-test to compare their allele frequencies. Because this approach requires pooled DNA sequences, it was also called "evolve and resequence" (E&R), initially by Turner et al (2011) and then by Baldwin-Brown et al. (2014). The method depends on replicated lines within each selection treatment to correct allele frequency variation caused by random genetic drift (or possibly founder effects). The idea was very simple – using the allele frequency of each replicated line as the original observed data point to test the mean difference in allele frequency between the two levels of selection treatments. Their main contribution was the addition of a regularization factor to the test to prevent some unexpected behavior of the test from happening (see Discussion). The regularized factor is particularly useful when the number of replicated lines within each selection treatment is small because, by chance, the variances of allele frequency among replicates may be extremely small, leading to false detection of small difference in allele frequency between selection treatments. Many other methods are also available for detecting selection signatures, as reviewed by Schlotterer et al (2015), but the regularized t-test is the state-of-art method for replicated selection experiments. If DNA sequences are available at the individual level, then using pooled allele frequency data may lead to loss of essential information and reduced power of detecting causally related single nucleotide polymorphisms (SNPs). Information on the allelic complement of individual organisms in the population hierarchy may be very important in boosting the statistical power, and incorporation of such information into the detection model is the main goal of the present study. Although the *F* statistics (WEIR AND COCKERHAM 1984) mentioned above already deal with genes at the level of individual organisms, a mixed model approach to detecting selection signatures in artificially manipulated populations may be more appropriate. In this study, we propose to use the minimum variance quadratic unbiased estimation (MIVQUE) procedure (RAO 1971) for mixed models to estimate variance components and test differentiation among selection treatments that contain replicate lines.

To validate the efficacy of the mixed model methodology, we used mouse populations under long-term artificial selection for high amounts of voluntary wheel-running behavior (SWALLOW *et al.* 1998; CAREAU *et al.* 2013). The selection experiment includes two treatments, each with four replicate lines (eight lines in total): four lines bred for high running (HR) and four serving as unselected control lines (C). These lines were developed as a model system to study correlated evolution and coadaptation of behavior with (exercise) physiology (WALLACE AND GARLAND JR. 2016). They are also viewed as relevant to human voluntary exercise behavior, which is very important in human health (GARLAND JR. *et al.* 2011b). DNA samples were collected from 80 mice from generation 61 of the selection experiment, 10 from each line. Detected selection signatures from this study will indicate that these genomic regions harbor genes responsible for voluntary wheel running. In subsequent reports, the biological functions of the identified genomic regions will be considered in detail, but that is beyond the scope of the present study.

While preparing for this manuscript, we found a very similar study in rats to detect selection signatures for alcohol preference (Lo *et al.* 2016). That experiment included bi-directional selection for high and low alcohol preference, with each treatment replicated twice (four lines in total; no non-selected control lines). They collected 10 mice from each line at generation 60, where the first 30 generations were continuously selected and the last 30 generations were relaxed (no selection applied). Although their sample size was only 40 rats, they were able to detect many regions harboring genes that may be causally related to alcohol preference. Lo et al. (2016) directly estimated the θ (F_{ST}) parameters under the random model methodology and used a permutation test to detect θ that significantly deviated from the null model. Results from our mouse selection experiment are expected to be more powerful because of the larger number of lines (8), larger sample size (80), and the use of the mixed model methodology that we propose.

Material and Methods

Experimental material

As described in the original publication (Swallow *et al.* 1998; Swallow *et al.* 2009), replicated within-family selection for increased voluntary wheel running in outbred laboratory house mice (*Mus domesticus*; Hsd:ICR strain: base population was 112 males and 112 females) was applied with four high-selected (High Runner or HR lines) and four non-selected control lines (10 families/line were carried forward each generation, with average litter size at weaning of approximately 10 pups). As young adults, mice were housed individually with access to activity wheels for a period of 6 days, and selection was based on the mean number of revolutions run on days 5 and 6. Animal model analyses indicated that at least three of the four HR lines reached plateaus between generations 17 and 27 of the experiment, depending on sex and line. At the apparent selection limits, mice from the HR lines ran approximately three-fold more than did those from the control lines (Careau *et al.* 2013). Various correlated response to selection have been observed, including reduced body mass and body length, decreased body fat as a percentage of total mass, increased endurance can maximal aerobic capacity, and various alterations related to neurobiology, motivation, and the brain reward system, as reviewed in (Rhodes and Kawecki 2009; Swallow *et al.* 2009; Garland Jr. *et al.* 2011a; Wallace and Garland Jr. 2016)

The selection experiment has been ongoing for almost 80 generations. For the present analyses, we collected DNA samples from 80 female mice at generation 61, 10 mice from each replicate line. Lines 1, 2, 4 and 5 were the non-selected control lines (C) and lines 3, 6, 7 and 8 were the high runner (HR) selection lines. Given that the HR lines had been at selection limits (Careau *et al.* 2013) for many generations at the time of sampling, random genetic drift is likely to have caused further differentiation that may have obscured many SNPs affected by the selection protocol. In the future, we plan to analyze earlier generations by use of historical tissue samples, as described in (Didion *et al.* 2016). Thus, the present data should be viewed as an exemplar to illustrate the utility of the proposed new statistical methods, not definitive with respect to signatures of selection in this particular selection experiment.

We used the Mega Mouse Universal Genotyping Array (MegaMUGA), which provides up to 77,800 single nucleotide polymorphism (SNP) markers and is built on the Illumina® Infinium platform (Morgan *et al.* 2016). The SNP markers are distributed throughout the mouse genome (average spacing of 33 Kb) and with a slight excess of probes in the telomeric regions of each autosome to facilitate detection of recombination events throughout the chromosome. Eight mice were eliminated from the analysis because of low quality SNP callings (1 from line 2, 1 from line 5, 2 from line 3, 1 from line 6 and 3 from line 8). Of the 77,808 SNPs in the panel, 52,490 SNPs were deleted due to missing values or monomorphism across the samples. Therefore, after this quality control, the data set subject to analysis has 72 female mice with 25,318 SNPs. In contrast to GWAS, population differentiation analysis does not use minor allele frequency (MAF) and Hardy-Weinberg disequilibrium as criteria for quality control. The 25,318 selected SNPs in the analysis were evenly distributed across 19 autosomes and the X chromosome. The SNP alleles were numerically coded as 1 for the reference allele and 0 for the alternative allele. As a result, there were $72 \times 2 = 144$ observations (one per allele) for each locus analyzed.

Mixed model analysis

The allelic model: We first introduce the random model methodology for the F-statistic (WEIR AND COCKERHAM 1984). Note that the F-statistics are population differentiation parameters, not the F-tests. As the response variable is the allelic value represented by a binary variable, the maximum likelihood method is not appropriate, unless a generalized linear mixed model is used (discussed earlier). Instead, we used the minimum variance quadratic unbiased estimation (MIVQUE) for variance component estimation (RAO 1971). The basic idea is to construct a hierarchical (nested) model to perform analysis of variances (ANOVA) using allelic indicator (0 or 1) as the response variable and the hierarchical structures of selection treatments and replicate lines within treatments as the design matrices, where the hierarchical structure is represented by alleles within individuals, individuals with replicate lines, and lines within selection treatments. We now consider two selection treatments only, one being the control treatment (unselected) and the other the artificially selected treatment. In the HR mouse experiment, the number of treatments was two (control and selection), the number of replicate lines within each treatment was four (four control lines and four HR selected lines), the number of individuals within each line was 10 (but varied after deletion of 8 mice with low quality SNP callings), and the number of alleles within each individual was two (diploid organism).

Let y_{ijkl} be the indicator variable (0 or 1) for the *l*th allele of the *k*th individual from the *j*th line within the *i*th treatment, where $l = 1, 2$ for the two alleles of each individual, $k = 1, ..., 10$ for the 10 individuals within each line, $j = 1, 2, 3, 4$ for the four lines within each treatment and $i = 1, 2$ for the two treatments (control and selection). Let A_1 be the “reference” allele and A_2 be the alternative allele of a locus under consideration. Denote the whole population frequency of A_1 by p and the frequency of A_2 by $q = 1 - p$. The allelic indicator variable for reference allele A_1 is

$$y_{ijkl} = \begin{cases} 1 & \text{for } A_1 \\ 0 & \text{for } A_2 \end{cases} \tag{1}$$

which is a Bernoulli variable and thus the expectation is identical to the frequency of the reference allele. We now use Cockerham's (1969) linear model to describe y_{ijkl},

$$y_{ijkl} = \mu + \alpha_i + \beta_{(i)j} + \gamma_{(ij)k} + \varepsilon_{(ijk)l} \quad (2)$$

where $\mu = p$ is the overall mean (frequency of A_1 for the whole experimental population) $\alpha_i = p_i - p$ is the allele frequency of treatment *i* expressed as deviation from that of the whole population, $\beta_{(i)j} = p_{ij} - p_i$ is the allele frequency of the *j*th line expressed as a deviation from the *i*th treatment, $\gamma_{(ij)k} = p_{ijk} - p_{ij}$ is the allele frequency of the *k*th individual expressed as a deviation from the *j*th line within the *i*th treatment, and $\varepsilon_{(ijk)l} = y_{ijkl} - p_{ijk}$ is the residual error. Note that the allele frequency of an individual is defined as $p_{ijk} = (y_{ijk1} + y_{ijk2})/2$, which only takes three possible values, 0, 0.5 and 1. The two selection treatments were not randomly sampled and they were designed by the investigators prior to the experiment. Therefore α_i should be treated as fixed effect. However, the Cockerham's model is random and thus we will take the random model approach as review of the background of population differentiation. The model contains only one fixed effect (μ) and thus it is called the random model. All other effects are random with mean zero and different variances. The variances are denoted by σ_α^2 for effect α_i, σ_β^2 for effect $\beta_{(i)j}$, σ_γ^2 for effect $\gamma_{(ij)k}$ and σ_ε^2 for residual $\varepsilon_{(ijk)l}$. The expectation of y_{ijkl} is $E(y_{ijkl}) = \mu$ and the variance of y_{ijkl} is

$$\mathrm{var}(y_{ijkl}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \quad (3)$$

Cockerham (1969) defined three Wright's F-statistics (WRIGHT 1951) based on these variance components. For the four-level hierarchical model, there are four F-statistics, which are defined as (YANG 1998),

$$F_{IT} = \frac{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \quad (4)$$

$$F_{TRT} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \quad (5)$$

$$F_{LINE} = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \quad (6)$$

$$F_{IS} = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2} \quad (7)$$

These F-statistics are different from the F-statistics developed by Weir and Cockerham (1984) but they have a nice property of

$$(1-F_{IT}) = (1-F_{TRT})(1-F_{LINE})(1-F_{IS}) \quad (8)$$

If we ignore the treatments by treating all lines as "populations," then we have

$$(1-F_{ST}) = (1-F_{TRT})(1-F_{LINE})$$

which leads to

$$F_{ST} = \frac{\sigma_\alpha^2 + \sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \tag{9}$$

This is the typical F_{ST} in the three-level hierarchical population subdivision model, where all lines are promoted to populations and $\sigma_\alpha^2 + \sigma_\beta^2$ represents the variance of the promoted populations.

As the two selection treatments (control and selection) presented in this study were not randomly sampled from a universe of all possible selection treatments, it is more appropriate to treat α_i as a fixed effect. Therefore, the model defined in equation (2) is a mixed model. Under the mixed model, the expectation of y_{ijkl} is

$$E(y_{ijkl}) = \mu + \alpha_i \tag{10}$$

and the variance of y_{ijkl} is

$$\mathrm{var}(y_{ijl}) = \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2 \tag{11}$$

Our purpose of detecting selection signals is to test the null hypothesis

$$H_0 : \alpha_1 = \alpha_2 \tag{12}$$

which does not require the F-statistics but we do need the original variance components to facilitate the hypothesis test. We proposed to use the MIVQUE method of Rao (1971) to estimate the variance components because normal distributions of the random effects and the residual errors are not required with MIVQUE.

It is much more convenient to use a matrix notation to derive the MIVQUE of variance components, as shown below

$$y = X_\mu \mu + X_\alpha \alpha + Z_\beta \beta + Z_\gamma \gamma + \varepsilon \tag{13}$$

where X_μ is an $n \times 1$ vector of unity, X_α is an $n \times 1$ vector whose elements are 1 for individuals in the selected treatment and -1 for individuals in the control treatment, $\alpha = \alpha_1 - \alpha_2$ is the difference of allele frequencies between the control and the selected populations, Z_β is an $n \times 8$ incidence matrix representing the 8 lines, β is a 8×1 vector of allele frequencies for the 8 lines, Z_γ is an $n \times 72$ incidence matrix for the 72 mice (38 from the control lines and 34 from the selected lines), γ is an 72×1 vector for individual effects and ε is an 144×1 vector of residuals. All random effects have expectations of zero and a variance σ_β^2 for β, a variance σ_γ^2 for γ and a variance σ_ε^2 for ε.

The expectation and variance of the model are

$$E(y) = X_\mu \mu + X_\alpha \alpha \tag{14}$$

and

$$\mathrm{var}(y) = V = Z_\beta Z_\beta^T \sigma_\beta^2 + Z_\gamma Z_\gamma^T \sigma_\gamma^2 + I\sigma_\varepsilon^2 \tag{15}$$

The MIVQUE of the three variance components $\theta = \left\{\sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2\right\}$ are obtained using the following linear equation system $H_{3\times3}\theta_{3\times1} = Q_{3\times1}$, the details of which are

$$\begin{bmatrix} H_{\beta\beta} & H_{\beta\gamma} & H_{\beta\varepsilon} \\ H_{\gamma\beta} & H_{\gamma\gamma} & H_{\gamma\varepsilon} \\ H_{\varepsilon\beta} & H_{\varepsilon\gamma} & H_{\varepsilon\varepsilon} \end{bmatrix} \begin{bmatrix} \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\varepsilon^2 \end{bmatrix} = \begin{bmatrix} Q_\beta \\ Q_\gamma \\ Q_\varepsilon \end{bmatrix} \qquad (16)$$

where the right hand sides of the equations are various quadratic forms of *y* and the left hand sides are the expectations of the quadratic forms. Let us define $X = [X_\mu \,\|\, X_\alpha]$ as vertical concatenation of the two matrices in the brackets and $\eta = \begin{bmatrix}\mu & \alpha\end{bmatrix}^T$ as the fixed effects. Further define $P = I - X(X^TX)^- X^T$, $V_\beta = PZ_\beta$, $V_\gamma = PZ_\gamma$ and $V_\varepsilon = PI = P$. The six unique elements of the *H* matrix are

$$H_{\beta\beta} = \mathrm{tr}(V_\beta V_\beta^T V_\beta V_\beta^T)$$

$$H_{\beta\gamma} = \mathrm{tr}(V_\beta V_\beta^T V_\gamma V_\gamma^T)$$

$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T)$$

$$H_{\gamma\gamma} = \mathrm{tr}(V_\gamma V_\gamma^T V_\gamma V_\gamma^T)$$

$$H_{\gamma\varepsilon} = \mathrm{tr}(V_\gamma V_\gamma^T V_\varepsilon V_\varepsilon^T) = \mathrm{tr}(V_\gamma V_\gamma^T)$$

$$H_{\varepsilon\varepsilon} = \mathrm{tr}(V_\varepsilon V_\varepsilon^T V_\varepsilon V_\varepsilon^T) = n-1$$

The remaining three elements of *H* take the three corresponding elements with flipping subscripts because the matrix is symmetrical. The three elements of the *Q* matrix are

$$Q_\beta = y^T V_\beta V_\beta^T y$$
$$Q_\gamma = y^T V_\gamma V_\gamma^T y$$
$$Q_\varepsilon = y^T V_\varepsilon V_\varepsilon^T y = y^T P y$$

The MIVQUE estimate of the parameter vector θ is $\hat\theta = H^{-1}Q$. Note that the MIVQUE estimate of a variance component can be negative because of the unbiased nature of the estimate. If that happens, it is simply set to zero.

The estimated variance components, denoted by $\hat\theta = \left\{\hat\sigma_\beta^2, \hat\sigma_\gamma^2, \hat\sigma_\varepsilon^2\right\}$, are then used to estimate the fixed effects and perform hypothesis tests. The estimated variance matrix of *y* is

$$\mathrm{var}(y) = \hat V = Z_\beta Z_\beta^T \hat\sigma_\beta^2 + Z_\gamma Z_\gamma^T \hat\sigma_\gamma^2 + I\hat\sigma_\varepsilon^2 \qquad (17)$$

The best linear unbiased estimate (BLUE) of the fixed effect is

$$\hat\eta = (X^T \hat V^{-1} X)^{-1} X^T \hat V^{-1} y \qquad (18)$$

and the variance matrix of this estimate is

$$\mathrm{var}(\hat\eta) = V_\eta = (X^T \hat V^{-1} X)^{-1} \qquad (19)$$

Note that

$$\hat\eta = \begin{bmatrix} \hat\mu \\ \hat\alpha \end{bmatrix} \text{ and } V_\eta = \begin{bmatrix} \mathrm{var}(\hat\mu) & \mathrm{cov}(\hat\mu, \hat\alpha) \\ \mathrm{cov}(\hat\alpha, \hat\mu) & \mathrm{var}(\hat\alpha) \end{bmatrix}$$

The *F* test for $H_0 : \alpha = 0$ is

$$F = \frac{\hat{\alpha}^2}{\operatorname{var}(\hat{\alpha})} \tag{20}$$

with degrees of freedom 1 (numerator) and 6 (denominator). The p-value is calculated using

$$p = 1 - \Pr(f_{1,6} < F) \tag{21}$$

where $f_{1,6}$ is a random variable of *F* distribution with 1 and 6 degrees of freedom. The *p* value is then converted into $-\log_{10}(p)$, which is used the Manhattan plots.

The genotypic model: Our interest here is not to estimate the F-statistics; rather, we are interested in a statistical test for the difference between the HR selected lines and the C lines. Therefore, we can use a model that takes individual genotypes as input data. Such a model is called the genotypic model, in which the response variable for each individual mouse is the average of the two allelic values (assuming the entire population only includes two alleles at each locus). If there are more than two alleles in the experimental population, then the bi-allelic model still applies by treating all none-reference alleles as the “other” allele, as suggested by Weir (Weir 1996). Let y_{ijk} be the numerically coded genotypic value for the *k*th individual within the *j*th line within the *i*th treatment and it is defined as

$$y_{ijk} = \begin{cases} 0 & \text{for} \quad A_2A_2 \\ 0.5 & \text{for} \quad A_1A_2 \\ 1 & \text{for} \quad A_1A_1 \end{cases} \tag{22}$$

The genotypic model is

$$y_{ijk} = \mu + \alpha_i + \beta_{(i)j} + e_{(ij)k} \tag{23}$$

where $e_{(ij)k} = \gamma_{(ij)k} + \bar{\varepsilon}_{(ijk)}$ is the residual effect with variance $\sigma_e^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2 / 2$, where σ_γ^2 and σ_ε^2 are variances defined in the allelic model. Under the mixed model, the expectation of y_{ijk} is

$$E(y_{ijk}) = \mu + \alpha_i \tag{24}$$

and the variance of y_{ijk} is

$$\operatorname{var}(y_{ijl}) = \sigma_\beta^2 + \sigma_e^2 \tag{25}$$

This genotypic model has reduced the model size by half and only involves two variance components. Therefore, it is computationally much more efficient than the allelic model. Parameter estimation and significance test are the same as the allelic model, except that the sample size has been reduced by half.

The gene frequency model

Baldwin-Brown, Long and Thornton’s regularized F test: Baldwin-Brown et al (2014) recently developed a regularized *t* test for detecting loci responsible for the phenotypic response to artificial selection or in experimentally evolved populations. The square of the regularized *t* test is the regularized *F* test. The test uses arcsine square root transformed allele frequency data. The test statistic is defined as

$$F = \frac{(x_1 - x_2)^2 r}{(1-\omega)(v_1 + v_2) + 2\omega\bar{v}} \qquad (26)$$

where

$\omega = 0.1$ is a coefficient of regularization set by the investigator (0.1 is the default value),

$r = 4$ is the number of lines within each treatment,

$x_1 = \hat{p}_1 = \bar{y}_{1...} = \frac{1}{80}\sum_{j=1}^{4}\sum_{k=1}^{10}\sum_{l=1}^{2} y_{1jkl}$ is the allele frequency of the selected lines,

$x_2 = \hat{p}_2 = \bar{y}_{2...} = \frac{1}{80}\sum_{j=1}^{4}\sum_{k=1}^{10}\sum_{l=1}^{2} y_{2jkl}$ is the allele frequency of the control lines,

$v_1 = \frac{1}{4-1}\sum_{j=1}^{4}(\bar{y}_{1j..} - \bar{y}_{1...})^2$ is the variance of the allele frequencies over the four selected lines,

$v_2 = \frac{1}{4-1}\sum_{j=1}^{4}(\bar{y}_{2j..} - \bar{y}_{2...})^2$ is the variance of the allele frequencies over the four control lines,

$\bar{v} = \frac{1}{2m}\sum_{s=1}^{m}(v_{1s} + v_{2s})$ is the average within treatment variance in allele frequency averaged over the two treatments and over all *m* loci.

When $\omega = 0$ is set, the method is the usual *F* test without regularization. The second term in the denominator of the test, $2\omega\bar{v}/r$, borrows information from all loci under investigation. Baldwin-Brown et al. (2014) interpreted $\bar{v}$ as an empirically motivated Bayesian prior on allowable variances in allele frequencies and has the effect of stabilizing the denominator of the *F* test. They claimed that such a regularization is important in experimental evolution studies in which a SNP could differentially fix in the experimental versus control replicates purely due to drift alone, and thus be associated with a traditional *F* test of infinity. Under the null model, the regularized *F* test follows an *F* distribution of 1 and $2(r-1) = 6$ degrees of freedom.

Regularized F test using linear regression: The regularized *F* test can be achieved using a general linear model (regression analysis). The general linear model has an advantage of being able to handle multiple treatments. For example, if there are three selection treatments and multiple replicated lines are available within each treatment, then the regularized *F* test cannot test the difference among the three selection treatments. In the present study, we extend the regularized *F* test using a general linear model approach. The response variable (*y*) is the arc-sine square root transformed allele frequency with eight observations for the mouse data. The linear model is

$$y = X_0\beta_0 + X_1\beta_1 + e \tag{27}$$

where X_0 is an 8×1 vector of unity, β_0 is the intercept, X_1 an 8×1 vector coded as -1 for the control population and 1 for the HR selected population, β_1 is the regression coefficient representing the difference in allele frequencies between the two selection treatments and e is an 8×1 vector of residual errors with an unknown variance σ^2. Let $\beta = \left[\beta_0 // \beta_1\right]$ and $X = \left[X_0 \parallel X_1\right]$. The estimated parameters are

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{28}$$

and

$$\hat{\sigma}^2 = \frac{1}{8-2}(y - X\hat{\beta})^T (y - X\hat{\beta}) \tag{29}$$

Incorporating the regularized parameter, the variance matrix of $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = (X^T X)^{-1}\left[(1-\omega)\hat{\sigma}^2 + \omega\bar{v}\right] \tag{30}$$

where $\omega = 0.1$ and $\bar{v}$ is the average estimated σ^2 across all loci in the neighborhood of the current locus or in the entire genome. The variance $\text{var}(\hat{\beta})$ is a 2×2 matrix with elements defined as

$$\text{var}(\hat{\beta}) = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{var}(\hat{\beta}_1) \end{bmatrix} \tag{31}$$

The regularized *F* test from this regression analysis is

$$F = \frac{\hat{\beta}_1^2}{\text{var}(\hat{\beta}_1)} \tag{32}$$

One can verify that β_1 is the difference of the allele frequencies between the two selection treatments and $\text{var}(\hat{\beta}_1)$ is identical to the denominator of equation (26) if $\hat{\sigma}^2$ is replaced by $(v_1 + v_2)/2$, the average within-population variance of the current locus.

Permutation test

As the response variable in the mixed model analysis is an indicator of the reference allele (a binary variable), the *F* test statistic does not follow the expected *F* distribution. In addition, multiple tests were involved in the analysis and the nominal 0.05 criterion of Type 1 error for the *p* value cannot be used. To control the genome-wide Type 1 error at 0.05, we used the permutation test (CHURCHILL AND DOERGE 1994) by randomly shuffling the mouse identification numbers so that any association of a locus with the treatment label would be a false positive. For each permuted data set, all 25,318 SNPs were analyzed, and the single largest *F* statistic was recorded. The permutation was replicated 1,000 times and then the 95 percentile of the empirical distribution of *F* statistics from permuted data was compared with the 25,318 real *F* statistics to determine the significance for each SNP. Any SNP for which the *F* statistic was greater than the 95th percentile of the empirical *F* distribution from the permuted data was considered significant at $p < 0.05$. This procedure thus controls the genome-wide Type 1 error rate at 5%. In the Manhattan plot, we presented the –log(*p*) test statistics of all loci against the genome positions. The empirical critical value of the *F* statistic was converted into an empirical critical value of *p* using degrees of freedom of 1 and 6, which reflects the experimental design with one fixed effect (selection) and four replicated lines (random effects (nested within linetype). That empirical critical value in *p* was further converted into the empirical critical value in the –log(*p*) scale. This critical value is sample specific, and thus is more appropriate than the Bonferroni correction, which is often too conservative (GAO *et al.* 2010).

In summary, we have presented four methods for detection of selection signatures. The mixed model approach under the allelic model (ALLELIC MODEL), the mixed model under the genotypic model (GENOTYPIC MODEL), the regularized *F* test using allele frequency (REGULARIZED F TEST), and the regularized F test using regression (REGRESSION F TEST). Except for the REGULARIZED F TEST, all other models can handle more than two treatment levels. All four methods were used to analyze the SNP data from the High Runner mouse selection experiment. A worked example is provided in Supplementary Note S1, using data presented in Supplementary Data S7, S8, and S9. The R code for each method is provided in Supplementary Note S2. Users familiar with SAS programs can directly call PROC MIXED with the Method = MIVQUE0 option to perform the mixed model analysis. However, if the number of markers is large, looping over all markers in SAS can take an extremely long time.

Results

Mouse data analysis

The genetic map of 25,318 markers and information about the mouse populations are provided in Supplementary Data S1 and S2, respectively. The SNP data coded as binary allelic states are provided in Supplementary Data S3. The corresponding SNP data coded as genotypic values are provided in Supplementary Data S4. Each SNP dataset has 25,318 rows (one row per marker), but the allelic dataset has 144 columns (one column per allele) and the genotypic dataset has 72 columns (one column per mouse). The data have no missing values and the number of individuals per line varied due to deletion of eight mice with low quality SNP callings. The mice in the population information file and the mice in the allelic and genotypic data files are arranged in the same order. The allele frequency data taken by the regularized F-test are given in Supplementary Data S5 with 25,318 rows and eight columns (one column per line).

All four approaches described in the Methods section (allelic model, genotypic model, regularized model, and regression model) were used for the data analysis. The first two methods are mixed-model based methods (new methods), while the last two are based on gene frequencies (existing methods). The Manhattan plots of the –log(p) test statistics are shown in Figure 1 for all four methods. The critical value of $-\log_{10}(p)$ from 1,000 permutation analyses is 2.4644 for the MIVQUE allelic model, 2.6405 for the MIVQUE genotypic model, and 4.95 for the two methods using gene frequency data. These critical values are shown in Figure 1 as the horizontal lines (dashed blue). The allelic model and the genotypic model under the mixed model analysis are identical (Figures 1A and 1B). The regularized F test and the regression F test are also identical (Figures 1C and 1D). Compared to the permutation-generated thresholds, the allelic model identified 152 markers, but the regularized F test failed to identify any markers. The 152 loci and their test statistics are listed in Supplementary Data S6, where the column with header "Mixed" shows the significant loci identified by the permutation test of the mixed model procedure. The more stringent threshold calculated from Bonferroni correction is –log(0.05/25318) = 5.70. If we had used this threshold, the allelic and genotypic models would still detect 21 markers in the middle of chromosome 9. These observations imply that the mixed model approach based on allelic data is more powerful than the regularized F test based on gene frequency data (see result of simulation studies). Figure 2 shows the qq-plots of the four methods, where a qq-plot is the plot of the observed test statistics against the expected test statistic calculated under the null model. The allelic and genotypic models (both are mixed models) behave as expected – the majority of markers fall on the diagonal lines and some markers deviate from the diagonal (Figures 2A and 2B). The regularized and regression models (both use frequency data) show that all markers are around the diagonal lines (Figures 2C and 2D).

From one permuted sample, we generated the Manhattan plots (Supplementary Figure S1) for the four methods. None of the markers shows any extreme values of the test statistic for the mixed models, but many markers show very large test statistics for the frequency models. This explains why the permutation-generated critical value for the frequency models are high. For the same permuted sample, we drew qq-plots (Supplementary Figure S2) and observed that the test statistics of the mixed model approaches do not fall on the expected diagonal lines, whereas the frequency models behave as expected. The F tests from the mixed models do not follow the expected F distribution; therefore, if one relied on the standard F distribution the tests would be too conservative. However, the F tests of the arc-sine square root transformed frequency data do follow the expected F distribution.

Although the regularized F test failed to identify any markers, the $-\log_{10}(p)$ test statistic is highly correlated with that of the mixed model ($r_{xy} = 0.96$), as illustrated in Figure 3A, which shows that the test statistic of the mixed model is higher than that of the frequency model. From a single permuted sample, the correlation is reduced to 0.41 (Figure 3B) and the frequency model has a higher statistic than the mixed model (i.e., the behavior is opposite to the real data analysis). We then selected the top 152 markers from the regularized F test to see how many of them overlap with the 152 detected marker from the mixed model analysis. We assume that the top 152 markers from the regularized F test are “significant.” We found that 118 markers overlapped (detected by both methods) and 34 markers were uniquely identified by one of the two methods. The 152 + 34 = 186 markers detected by both methods are listed in Supplementary Data S6 along with the test statistics and allele frequencies for each of the eight lines. Except chromosomes 2, 8, 13 and 19, each chromosome (including chromosome X) caries at least one significant marker.

The 152 significant markers occurred on all chromosomes except 2, 8, and 19, and show a variety of patterns in terms of fixation (or the lack thereof) in the four HR and four C lines (Supplementary Data S6). Although a number of alleles were fixed within lines, none were fixed between the two selection treatments. For example, marker UNC10025993 on chromosome 5 had frequencies of 1, 1, 0.55, and 1 in HR lines 3, 6, 7, and 8, respectively, versus zero in all four C lines. In contrast, marker UNC12559756 on chromosome 7 had frequencies of zero in all HR lines versus 0.45, 0.667, 0.65, and 0.389 in the C lines. Others showed intermediate frequencies, such as UNC24564099 on chromosome 14, with frequencies of 0.1875 0.2222, 0, and 0.2857 in the HR lines versus 0.65, 0.5556, 0.75, and 1 in the C lines.

Intuitively, if this region is under selection due to the artificial selection protocol, then populations with such small sizes after 61 generations of selection should have some loci that are completely fixed in all four HR lines. The lack of this pattern for this region may be in part related to the within-family selection scheme (Swallow et al. 1998), which is known to slow the fixation process (FALCONER AND MACKAY 1996).

Interestingly, a total of 21 loci on chromosome 9 were detected even using the most stringent criterion (Bonferroni correction). These loci are within a 901 kb region on chromosome 9. The p-values from the allelic model and the allele frequencies of the eight lines are given in Table 1. Three lines in C and one line in HR were completely fixed in allele frequency for all loci in this region. There appeared to be two recombination breakpoints taking place within this region.

Power analysis from a simple simulation study

We mimicked the mouse experiment with eight lines and 10 mice in each line to examine the statistical power of the methods. We simulated 10 independently segregating loci to investigate the powers using 10 independent neutral loci to control the Type 1 error. We used two Beta distributions to simulate the allele frequencies of the eight lines. For the four control lines, the Beta distribution was

$\mathrm{Beta}(\alpha_0, \beta_0)$, where $\alpha_0 = 20$ and $\beta_0 = 30$, leading to an average allele frequency of $\alpha_0 / (\alpha_0 + \beta_0) = 0.4$. For the four HR lines, the allele frequencies were generated from $\mathrm{Beta}(\alpha_1, \beta_1)$, where $\alpha_1 = 30$ and $\beta_1 = 20$, leading to an average allele frequency of $\alpha_1 / (\alpha_1 + \beta_1) = 0.6$. Therefore, the average difference in allele frequency between the HR and C populations was 0.2. Once the allele frequencies were simulated for all lines, we then simulated the allele of each line from a Bernoulli distribution with the simulated allele frequency as the parameter. The actual count data (allele presences) for each line were drawn from a Beta-Binomial distribution. Such a simulation was replicated 1,000 times. The number of loci detected over the total number of loci simulated was the empirical power for the methods compared. The criterion of a locus being detected was determined from another 1,000 simulated samples under the null model where the allele frequencies of all lines from the C and HR selection treatments were generated from $\mathrm{Beta}(\alpha, \beta)$, where $\alpha = \beta = 25$. The critical value of the $-\log_{10}(p)$ test statistics under the Type 1 error of 0.05 from the 1,000 null samples was 0.83 for the mixed model and 1.23 for the regularized F test. Based on these critical values, the empirical power was 0.5541 for the mixed model method and 0.4465 for the regularized F test. The new method indeed was more powerful than the current regularized F test (see Figure 4). We then changed the Type 1 error and monitored the change of the empirical statistical power from the 2×1000 simulated samples to perform a sensitivity analysis. The receiver operating characteristic (ROC) curves of the two methods are shown in Figure 4. The curve of the mixed model is consistently higher than that of the regularized F test method, indicating that the power of the former is always higher than or equivalent to the power of the latter for all levels of Type 1 error.

Discussion

When the two selection treatments (HR and C) are treated as random effects (the random model approach), there are four variance components for each locus, σ_α^2 for treatments (TRT), σ_β^2 for lines (LINE) within treatments, σ_γ^2 for individuals within lines within treatments, and σ_ε^2 for residuals. We estimated these variance components for all loci and took the ratios to obtain the F statistics for each of the 25,332 loci. We then pooled the variance components over loci and obtained overall F statistics (over all loci) using the following equations (WEIR 1996),

$$F_{IT} = \frac{\sum_{k=1}^{m} (\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2)}{\sum_{k=1}^{m} (\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)} \tag{33}$$

$$F_{TRT} = \frac{\sum_{k=1}^{m} \sigma_\alpha^2}{\sum_{k=1}^{m} (\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)} \tag{34}$$

$$F_{LINE} = \frac{\sum_{k=1}^{m} \sigma_{\beta}^{2}}{\sum_{k=1}^{m} (\sigma_{\beta}^{2} + \sigma_{\gamma}^{2} + \sigma_{\varepsilon}^{2})} \quad (35)$$

$$F_{IS} = \frac{\sum_{k=1}^{m} \sigma_{\gamma}^{2}}{\sum_{k=1}^{m} (\sigma_{\gamma}^{2} + \sigma_{\varepsilon}^{2})} \quad (36)$$

The four genome-wide F statistics for the mouse populations are $F_{IT} = 0.6314$, $F_{TRT} = 0.0058$, $F_{LINE} = 0.6406$ and $F_{IS} = 0.0316$. Thus, the two selection treatments were not differentiated, but the eight lines were significantly differentiated, which can be attributed to random genetic drift and possibly also to different adaptive responses, called multiple solutions (GARLAND JR. *et al.* 2011a), in the HR lines. The average inbreeding coefficient within lines (0.0316) was very small due to the use of the within-family selection scheme.

The regularized F test proposed by Baldwin-Brown et al (2014) is the state-of-the-art method for detection of selection signatures in selection experiments with multiple replicated lines. The method is extremely simple, yet performs very well based on their simulation studies. The key issues addressed in that study are (i) replications and (ii) regularization. (i) Replications mean that there must be replicated lines within each selection treatment in order to separate the effect of selection from genetic drift (and multiple solutions). However, replications per se rarely happen in natural populations and thus detection of selection signatures from natural populations is more difficult because of the confounding effect between selection, possible multiple solutions, and drift (MUIR 1986). (ii) Regularization refers to a process in which a small positive number is added to the denominator of the F test statistic. Regularization is an intelligent way to deal with a special case where the within-population variances of allele frequencies are extremely small (e.g., due to drift) so that the F test is severely inflated even if the difference in allele frequency between selection treatments is small. The regularized F test borrows the average within-treatment variance from other loci and incorporates it into the within-treatment variance of the current locus to smooth the test statistics and thus prevents such an inflation in test statistics. The regularization procedure can also prevent reckless changes in test statistic between consecutive loci.

Of the 25,318 loci analyzed in the mouse data, 69 have allele frequency of exactly 0.5 for each of the eight lines. The usual F test (without regularization) statistic is not defined for these loci because the denominator is zero. The fact that the numerator of the test for these loci is also zero means that the test statistics should be zero (the two selection treatments are not different in allele frequency). The regularized F test correctly gives a zero test statistic value for all the 69 loci. Another example comes from marker UNC30702889 on chromosome X. The allele frequencies of the four C lines are 0.45, 0.4444, 0.45 and 0.4444, while the allele frequencies of the four HR lines are all 0.5. Although the difference in allele frequency between C and HR is very small (~0.05), the plain F test is 1,075.95 with a p-value of 5.36E-08 and a –$\log_{10}$(p) = 7.2712, which is the highest test value across the entire genome. This test statistic is severely inflated due to the extremely small variance within treatments. The regularized F test, however, gives a test statistic of 0.2012 with a p-value of 0.6695 and a –$\log_{10}$(p) of 0.1743. Thus, as desired, the regularization factor has corrected such an inflation.

The most obvious advantage of the regularized F test is that it takes pooled DNA samples as input data. Each pooled DNA sample represents a replicate line within a given selection treatment. For the eight replicate lines in the mouse selection experiment discussed here, only eight pooled samples are required to perform tests. This represents a tremendous cost saving. Unfortunately, such an advantage can turn into a disadvantage if DNAs are sequenced at the individual level because this F test cannot handle allelic data. Clearly, if all individuals are sequenced, and individual variation within lines exists, then pooling the DNA samples will lead to information loss. This is the very reason for us to develop the mixed model approach when DNAs from multiple individuals are separately sequenced in a selection experiment.

The regularized F test in the current form can only test the difference in allele frequency between two treatment levels because it is a squared t test and a t test is only suitable for comparing two groups. We have extended this method to handle multiple treatment levels using a general linear model approach (regression method). When applied to two treatments, the regularized regression method and the regularized t square method generate identical results (see Figure 1C and 1D). The regression method has an option to incorporate the sample size information of replicated lines into the model. For example, the sample sizes (*n*) were 10, 9, 10, 9, 8, 9, 10 and 7, respectively, for lines 1, 2, 4, 5 (C) and 3, 6, 7, 8 (HR). Such information can be easily incorporated into the regression model through a weight variable that is defined as the total number of alleles (2 times the sample size) of that line. The exact weight value for each line should be the inverse of *pq*/(2*n*). However, when $\hat{p}=0$ or $\hat{p}=1$, the weight is infinity. Therefore, simply using $2n$ as the weight is justifiable. The regularized regression analysis conducted here is not the weighted method because we wanted to demonstrate the equivalence of this method to the regularized t square test.

Current DNA sequencing technology is sufficiently inexpensive so that sequencing can be easily conducted at the individual level. When individuals are sequenced, pooling DNA sequences of all individuals within a line, i.e., using allele frequencies within lines, may represent a tremendous information loss. Therefore, we proposed a mixed model approach to detecting genome-wide selection signatures. The differences between or among selection treatments (e.g., selected versus control groups) are treated as fixed effects, and effects of replicate lines within treatments are treated as random. There are two versions of the mixed models: the allelic model and the genotypic model. The allelic model is the classical model of Weir and Cockerham (1984) where each entry of the response variable is an allele. The hierarchical structure of the alleles is preserved and such a test captures maximum information from the populations. The genotypic model simply takes the “allele frequencies of individuals” as the response variable. Given that every diploid individual only carries two alleles, the “allele frequency” of an individual only takes three possible values, 0, 0.5 and 1. No information is lost by pooling the two alleles of each individual together. Therefore, the genotypic model generates identical results as the allelic model (see Figure 1A and 1D). The genotypic model is computationally much more effective than the allelic model because the number of entries has been reduced to half. Hence, the genotypic model is recommended for genome-wide association studies for selection signatures.

An interesting feature of the mixed model approach (both the allelic and genotypic models) is that no regularization is required in the test. For example, the SNP named UNC30702889 on chromosome X discussed early in this section requires regularization for the F test because the within-treatment variance is too small. However, the allelic model without any regularization gives a test statistic of 0.21345, a p-value of 0.66035 and a $-\log_{10}(p)$ of 0.18022, which are comparable to the regularized F test.

The fact that the response variable of the mixed model analysis is the allelic state (binary) may challenge the validity of the mixed model methodology and lead someone to think that a generalized linear mixed model may be more appropriate. However, there are two justifications for the current mixed model methodology. (1) When the response variable is the allelic state (binary variable), different variance components and variance ratios have special biological meanings – covariance and correlations between alleles at different levels of the hierarchy. Such a treatment also preserves the original natures of Wright’s *F* statistics. (2) The mixed model analysis with the allelic state as the response variable is computationally more effective than the generalized linear mixed model analysis, which requires iterations and often faces convergence issues. If the purpose of the analysis is just to test the difference between two populations, the generalized linear mixed model may be considered if computational complexity is not a concern. The GLIMMIX procedure in SAS (2009) is particularly designed for this. However, looping over ten of thousands of markers for PROC GLIMMIX may take an exceptionally long time and so the gain in power from the GLIMMIX analysis may not justify the computational effort.

The mixed model analysis of genome-wide association studies for detection of selection signatures is similar to the GWAS for quantitative trait analysis (HIRSCHHORN AND DALY 2005; YU *et al.* 2006), except that there is no specific trait associated with the genetic analysis. Therefore, this method is also called GWAS without traits (LO *et al.* 2016). Unlike the regular quantitative trait GWAS, where we can control the polygenic background by incorporating a marker-inferred kinship matrix into the covariance structure, GWAS for selection signature detection does not have an obvious way to control the “polygenic background.” Therefore, the Type 1 error may not be controlled properly. To mimic GWAS in quantitative trait analysis, we may treat the population structure as the response variable and the allelic state as an independent variable. This treatment may be easily modified to incorporate the “polygenic effect” into the model, just like the regular mixed model GWAS (YU *et al.* 2006). It is straightforward to do so if there are only two populations, where the response variable is binary. For multiple populations, a multinomial response may be used to indicate the population entries. However, hierarchical population structures may not be easily handled this way. GWAS and QTL mapping for selection signatures is a relatively new area, with large room for improvement. The present study is one of the first attempts to merge studies of selection and quantitative genetics in the genomic era. We have adopted the mixed model in our selection signature detection but have not yet incorporated the “kinship” matrix into the selection model. A complete unification of GWAS and selection is possible but still in the future. In future studies, it will also be important to identify the presumably smaller number of haplotypes that contain the statistically significantly differentiated SNPs analyzed herein, but doing so accurately will likely require whole-genome sequence data.

Acknowledgements

This work was supported by the National Science Foundation Collaborative Research Grant (DBI-1458515) to SX and NSF grant IOS-1121273 to TG. We thank Fernando Pardo-Manuel de Villena, Daniel Pomp, and Liran Yadgary for providing the data analyzed here and for many useful discussions. We also thank Layla Hiramatsu and David Hillis for helpful discussions.

References

Balding, D. J., and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96**:** 3-12.

Baldwin-Brown, J. G., A. D. Long and K. R. Thornton, 2014 The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. Molecular Biology and Evolution 31**:** 1040-1055.

Beavis, W. D., 1994 The power and deceit of QTL experiments: Lessons from comparitive QTL studies, pp. 250-266 in *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*. American Seed Trade Association, Washington, D.C.

Brookfield, J. F. Y., 2001 Population genetics: The signature of selection. Current Biology 11**:** R388-R390.

Careau, V., M. E. Wolak, P. A. Carter and T. Garland Jr., 2013 Limits to behavioral evolution: the quantitative genetics of a complex trait under directional selection. Evolution 67**:** 3102–3119.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138**:** 963-971.

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution 23**:** 72-84.

Cui, Y., F. Zhang, J. Xu, Z. Li and S. Xu, 2015 Mapping quantitative trait loci in selected breeding populations: A segregation distortion approach. Heredity 115.

Didion, J. P., A. P. Morgan, L. Yadgary, T. A. Bell, R. C. McMullan *et al.*, 2016 R2d2 drives selfish sweeps in the house mouse. Molecular Biology and Evolution 33**:** 1381-1395.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Pearson, Harlow, England.

Gao, X., L. C. Becker, D. M. Becker, J. D. Starmer and M. A. Province, 2010 Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34**:** 100-105.

Garland Jr., T., S. A. Kelly, J. L. Malisch, E. M. Kolb, R. M. Hannon *et al.*, 2011a How to run far: multiple solutions and sex-specific responses to selective breeding for high voluntary activity levels. Proceedings of the Royal Society B: Biological Sciences 278**:** 574–581.

Garland Jr., T., H. Schutz, M. A. Chappell, B. K. Keeney, T. H. Meek *et al.*, 2011b The biological control of voluntary exercise, spontaneous physical activity and daily energy expenditure in relation to obesity: human and rodent perspectives. Journal of Experimental Biology 214**:** 206–229.

Garland, T., and M. R. Rose, 2009 *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*. University of California Press, Berkeley, California.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6**:** 95-108.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121**:** 185-199.

Lo, C. L., A. C. Lossie, T. Liang, Y. Liu, X. Xuei *et al.*, 2016 High resolution genomic scans reveal genetic architecture controlling alcohol preference in bidirectionally selected rat model. PLoS Genet 12**:** e1006178.

Luo, L., and S. Xu, 2003 Mapping viability loci using molecular markers. Heredity 90**:** 459-467.

Luo, L., Y. M. Zhang and S. Xu, 2005 A quantitative genetics model for viability selection. Heredity 94**:** 347-355.

Morgan, A. P., C. P. Fu, K. C. Y., C. E. Welsh, J. P. Didion *et al.*, 2016 The mouse universal genotyping array: from substrains to subspecies. G3: Genes|Genomes|Genetics 6**:** 263–279.

Muir, W. M., 1986 Estimation of response to selection and utilization of control populations for additional information and accuracy. Biometrics 42**:** 381-191.

Rao, C. R., 1971 Minimum variance quadratic unbiased estimation of variance components. Journal of Multivariate Analysis 1**:** 445-456.

Rhodes, J. S., and T. J. Kawecki, 2009 Behavior and neurobiology, pp. 263-300 in *Experimental evolution: concepts, methods, and applications of selection experiments*, edited by T. Garland Jr. and M. R. Rose. University of California Press, Berkeley, California.

SAS Institute Inc, 2009 *SAS/STAT: Users' Guide, Version 9.3*. SAS Institute Inc., Cary, NC.

Schlotterer, C., R. Kofler, E. Versace, R. Tobler and S. U. Franssen, 2015 Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity 114**:** 431-440.

Swallow, J. G., P. A. Carter and T. Garland Jr., 1998 Artificial selection for increased wheel-running behavior in housemice. Behav. Genet. 28**:** 227-237.

Swallow, J. G., J. P. Hayes, P. Koteja and T. Garland Jr. , 2009 Selection experiments and experimental evolution of performance and physiology, pp. 301-351 in *Experimental evolution: concepts, methods, and applications of selection experiments*, edited by T. Garland Jr. and M. R. Rose. University of California Press, Berkeley, California.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genetics 7**:** e1001336.

Vogl, C., and S. Z. Xu, 2000 Multipoint mapping of viability and segregation distorting loci using molecular markers. Genetics 155**:** 1439-1447.

Wallace, I. J., and T. Garland Jr., 2016 Mobility as an emergent property of biological organization: insights from experimental evolution. Evolutionary Anthropology 25**:** 98–104.

Weir, B. S., 1996 *Genetic Data Analysis II - Methods for Discrete Population Genetic Data*. Sinauer Associates, Inc. Publishers, Synderland, Massachusetts.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38**:** 1358-1370.

Wright, S., 1950 Genetical structure of populations. Nature 166**:** 247-249.

Wright, S., 1951 The genetic structure of populations. Ann. Eugen. 15**:** 323-354.

Wurschum, T., 2012 Mapping QTL for agronomic traits in breeding populations. Theor Appl Genet 125**:** 201-210.

Xu, S., 1996 Mapping quantitative trait loci using four-way crosses. Genetical Research 68**:** 175-181.

Xu, S., 2003 Theoretical basis of the Beavis effect. Genetics 165**:** 259-2268.

Yang, R.-C., 1998 Estimating hierarchical F-statistics. Evolution 52**:** 950-956.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38**:** 203-208.

Table 1. Markers detected on Chromosome 9 of the mouse genome that show significant differentiation between the control (C) and the HR selected lines.

Marker (Chr #9)	Position (bp)	p-value	C[a]	C	C	C	HR	HR	HR	HR
			p1	p2	p4	p5	p3	p6	p7	p8
UNC16231229	41246129	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
JAX00170437	41266019	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16231874	41301221	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16232212	41326208	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16232585	41353991	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16232919	41381162	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
JAX00691456	41473757	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16235286	41547967	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
JAX00170461	41592916	6.24E-07	0	0	0	0	0.8125	1	0.9	0.857143
UNC16236699	41636184	1.73E-07	0	0	0	0	0.875	1	0.9	0.857143
UNC16237066	41656313	1.73E-07	0	0	0	0	0.875	1	0.9	0.857143
UNC16237562	41689627	1.73E-07	0	0	0	0	0.875	1	0.9	0.857143
UNC16238010	41729317	3.98E-06	0	0	0	0.166667	0.875	1	0.9	0.857143
UNC16238418	41767394	1.73E-07	0	0	0	0	0.875	1	0.9	0.857143
UNC16240425	41877786	1.73E-07	0	0	0	0	0.875	1	0.9	0.857143
UNC16241644	41948973	3.98E-06	0	0	0	0.166667	0.875	1	0.9	0.857143
UNC16242398	41992897	9.28E-06	1	1	1	0.777778	0	0	0	0.142857
UNC16242829	42013727	9.28E-06	1	1	1	0.777778	0	0	0	0.142857
UNC090061659	42067067	9.28E-06	1	1	1	0.777778	0	0	0	0.142857
UNC16243882	42070360	9.28E-06	1	1	1	0.777778	0	0	0	0.142857
UNC16244740	42147771	6.38E-08	0	0	0	0	1	1	1	0.857143

[a] The last eight columns are the allele frequencies of the eight lines (four from C and four from HR).

Figure Legends

Figure 1. Manhattan plots of genome-wide selection signals from the mouse selection experiment (Swallow et al., 1998) at generation 61 using four different methods. The top two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The bottom two panels show the plot from the regularized F test (Baldwin-Brown et al., 2014: panel C) and the regression model (panel D). The dashed horizontal line (blue) is the empirical threshold obtained from analysis of 1,000 permuted samples.

Figure 2. QQ-plots of genome-wide loci of the mouse selection experiment using four different methods. In each qq-plot, the y-axis is the observed test statistic and the x-axis is the expected test statistic under the null model. The upper two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The lower two panels show the plots from the regularized F test (panel C) and the regression model (panel D). Both of the mixed model approaches show more data points deviating from the diagonal lines than the other approaches, thus indicating higher statistical power.

Figure 3. Comparison of the $-\log_{10}(p)$ test statistics of the allelic model (*y*) with the regularized F test (*x*) from the real data analysis (panel A) and from the analysis of a permuted sample (panel B). The Pearson correlation coefficients between the test statistics of the two methods are represented by r_{xy}. These plots demonstrate that the test statistic of the mixed model are highly correlated with the test statistic of the regularized F test in the real data analysis, while the correlation is significantly reduced in the permuted data analysis (null model).

Figure 4. Comparison of the receiver operating characteristic (ROC) curves of the mixed model method (allelic model) and the regularized F test. The *x*-axis is the Type 1 error and the *y*-axis is the statistical power. The curve for the mixed model is consistently higher than that of the regularized F test method, indicating that power of the former is always higher than or equivalent to the power of the latter for all levels of Type I error. Distance (0.5528 - 04495 = 0.1033) between the two points on the plot represents the gain in statistical power of the mixed model (0.5528) over the regularized F test (0.4495) when the Type 1 error is set at 0.05.

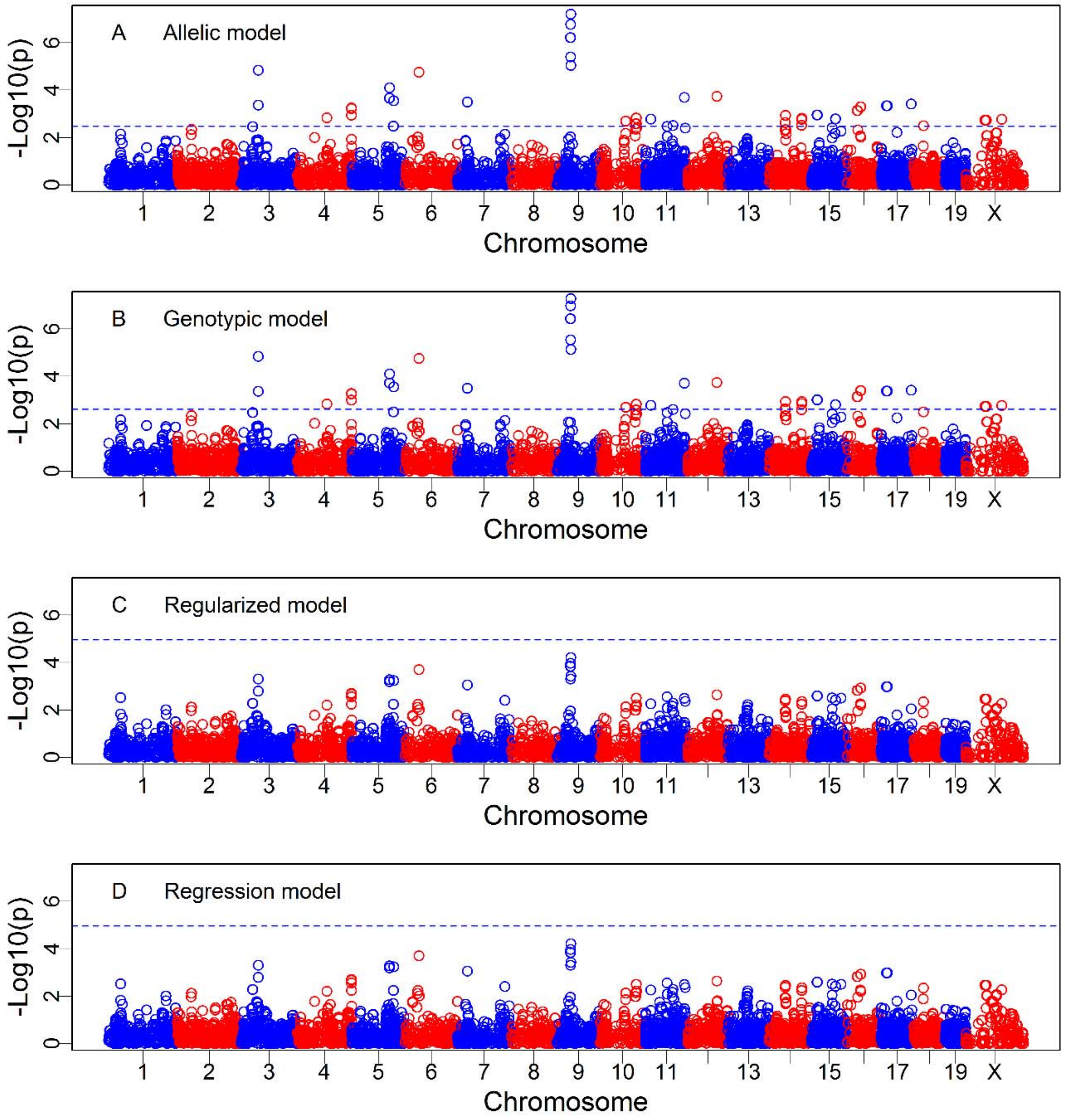


Figure 1. Manhattan plots of genome-wide selection signals from the mouse selection experiment (Swallow et al., 1998) at generation 61 using four different methods. The top two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The bottom two panels show the plot from the regularized F test (Baldwin-Brown et al., 2014: panel C) and the regression model (panel D). The dashed horizontal line (blue) is the empirical threshold obtained from analysis of 1,000 permuted samples.

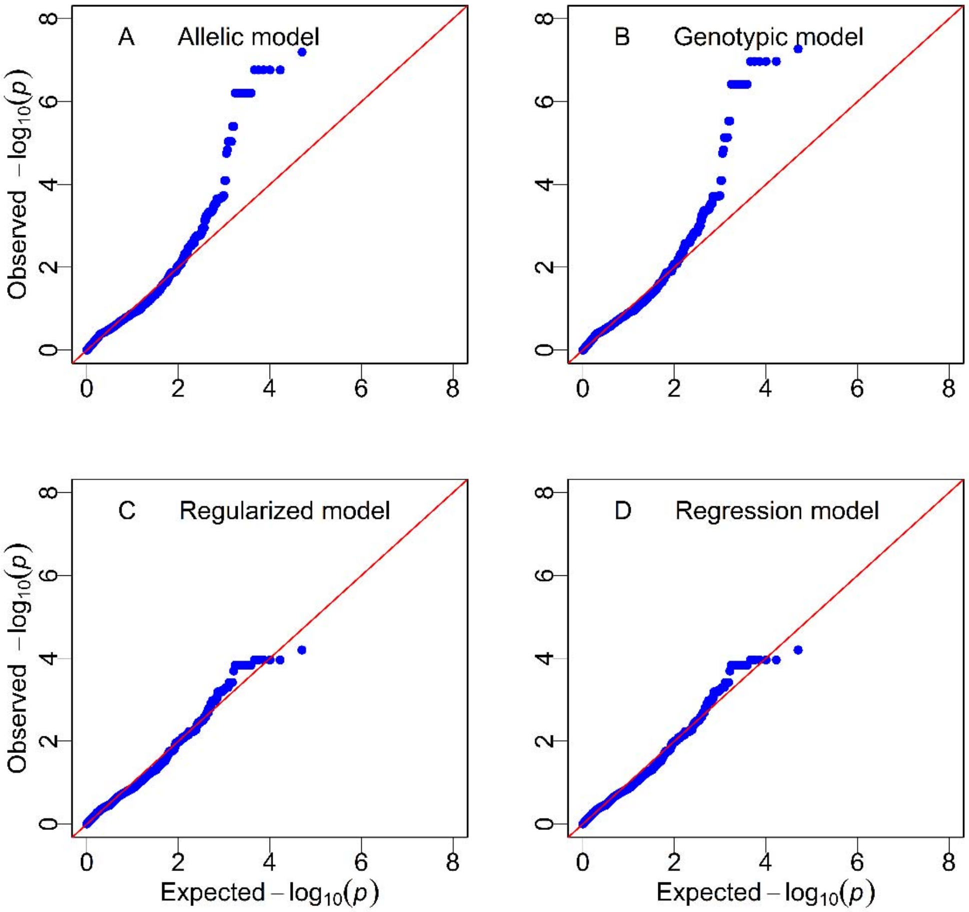


Figure 2. QQ-plots of genome-wide loci of the mouse selection experiment using four different methods. In each qq-plot, the y-axis is the observed test statistic and the x-axis is the expected test statistic under the null model. The upper two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The lower two panels show the plots from the regularized F test (panel C) and the regression model (panel D). Both of the mixed model approaches show more data points deviating from the diagonal lines than the other approaches, thus indicating higher statistical power.

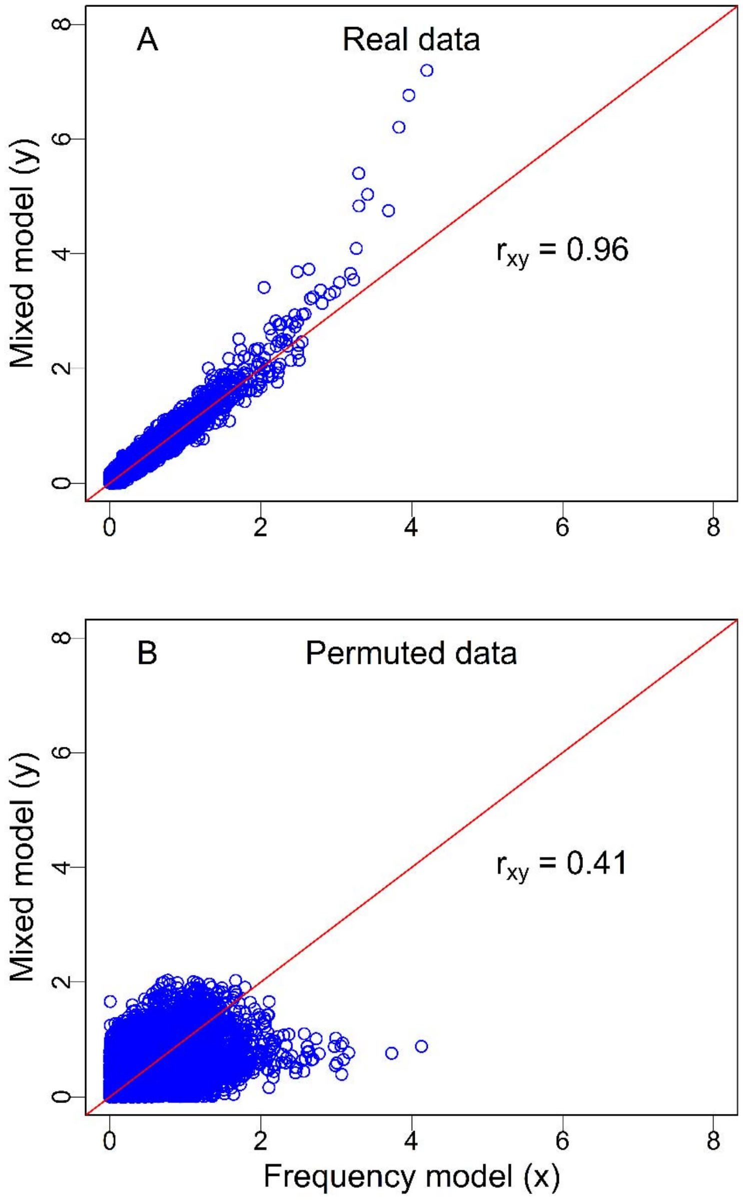


Figure 3. Comparison of the $–\log_{10}(p)$ test statistics of the allelic model (*y*) with the regularized F test (*x*) from the real data analysis (panel A) and from the analysis of a permuted sample (panel B). The Pearson correlation coefficients between the test statistics of the two methods are represented by r_{xy}. These plots demonstrate that the test statistic of the mixed model are highly correlated with the test statistic of the regularized F test in the real data analysis, while the correlation is significantly reduced in the permuted data analysis (null model).

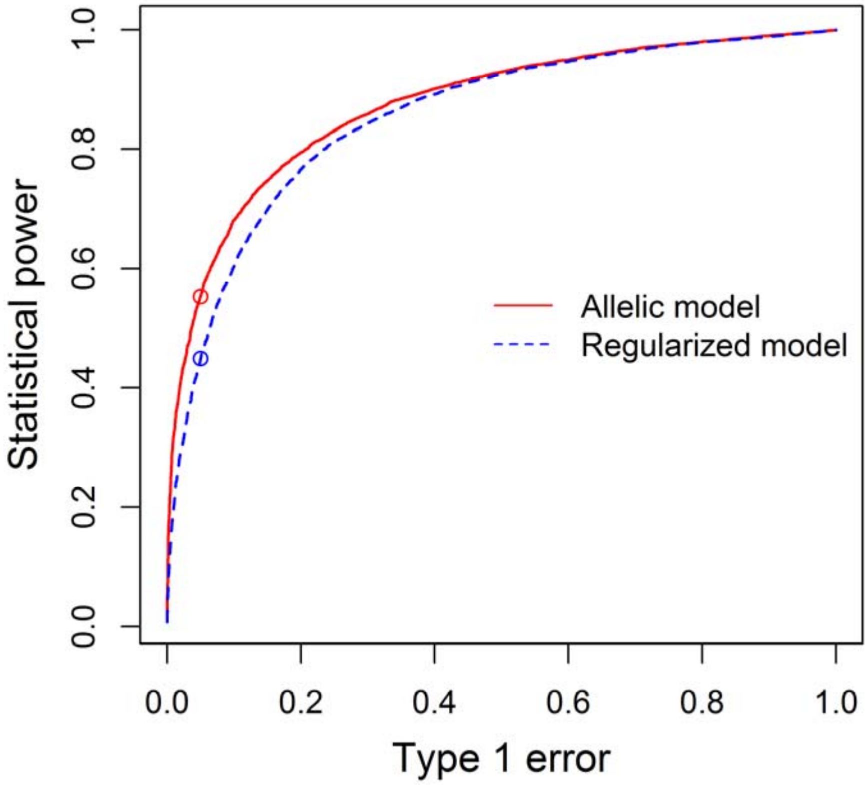


Figure 4. Comparison of the receiver operating characteristic (ROC) curves of the mixed model method (allelic model) and the regularized F test. The *x*-axis is the Type 1 error and the *y*-axis is the statistical power. The curve for the mixed model is consistently higher than that of the regularized F test method, indicating that power of the former is always higher than or equivalent to the power of the latter for all levels of Type I error. Distance (0.5528 - 04495 = 0.1033) between the two points on the plot represents the gain in statistical power of the mixed model (0.5528) over the regularized F test (0.4495) when the Type 1 error is set at 0.05.

Data Availability

The raw SNP data will be made available online once the manuscript has been accepted for publication, but with a one-year embargo.

Data S1.csv Marker map of the 25,332 SNPs used in the mouse data analysis.
757 KM

Data S2.csv Mouse population information including treatments (0 and 1), lines (1, 2, 4, 5, 3, 6, 7 and 8), mouse ID (1, 2, …, 72) and allele (1 and 2).
4 kb

Data S3.csv Allelic data of 144 alleles from 72 mice for 25,332 SNP loci, where 1 and 0 represent presence and absence of the reference allele.
7,878 KB

Data S4.csv Genotypic data of 72 mice for 25,332 SNP loci, where each genotypic values takes one of the three values, 0, 0.5 and 1.
4,814 KB

Data S5.csv Gene frequencies of eight lines (p_1, p_2, …, p_8) of the mouse population for 25,332 SNP loci, where y_i is the count of the reference alleles and n_i is the total number of alleles for the ith line.
2,624 KB

Supplementary Material

Supplementary Material_v2.docx

Figure S1. Manhattan plots from a permuted sample for genome-wide selection signals from the mouse selection experiment using four different methods. The top two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The bottom two panels show the plot from the regularized F test (panel C) and the regression model (panel D). The dashed horizontal line (blue) is the empirical threshold obtained from analysis of 1,000 permuted samples.

Figure S2. QQ-plots from a permuted sample for genome-wide loci of the mouse selection experiment using four different methods. The upper two panels are the mixed model approach under the allelic model (panel A) and the genotypic model (panel B). The lower two panels show the plots from the regularized F test (panel C) and the regression model (panel D).

Note S1: A Working Example

Note S2: R Codes and Brief User Instruction

Data S6.xlsx Test statistics for 186 loci detected by the mixed model and the regularized F test (Sheet 1) and allele frequencies of the 186 loci (Sheet 2).
59 KB

Data S7.xlsx Allelic information for SNP UNC2173488 used in the working example.
14 KB

Data S8.xlsx Genotypic information for SNP UNC2173488 used in the working example.
12 KB

Data S9.xlsx Gene frequencies of the eight lines for SNP UNC2173488 used in the working example.
9 KB