Derivation of the formula

The ANOVA model we are using is basically a linear regression model. For example, we have 7 strains with phenotypic data (v1, v2, …, v7) and haplotypic grouping 1231233. Then the  vector in the regression model is just (v1, v2, …, v7) and the predictor matrix  is just:



Here, , , the 1st column of X is a column of ones, representing the intercept and the 2nd column is just the indictor of whether the corresponding strain is of haplotype 2 or not. And the 3rd column has similar meaning but it is designated for haplotype 3.

Using theory on linear regression, we know that the model is . The estimated coefficient is  and therefore, the residuals of the regression model are just , where  is the identity matrix and . Please note that a)  and b) , and therefore . Finally, the residual sum of squares (RSS) is:



Next, we want to show that if  is multiplied by a non-singular matrix  to the right,  remains unchanged. To show this, it suffices to show that  remains unchanged. This is straightforward as . This identity plays a key role in the method, as it allows us to make  orthonormal without changing the corresponding property of .

Suppose that we have used the Gram-Schmidt process to make column vectors of  have unit norm and be orthogonal to each other. This process is equivalent to multiply  by a non-singular matrix to the right. In our example, this new  is:



Here  represents the i-th column vector of the new . Since these ’s are orthogonal to each other and all have length 1, we see that , where  is a  identity matrix. So  and . (The final equality is because  is indeed a scalar.) Please note that in our example that the 1st column of untransformed  is a vector of one,  and therefore , where . And 

To obtain the p value for the ANOVA model, we need to figure out the F-statistic for the model first. The F-statistic is defined as:. Here  is just derived and  is just the total variation: . So we can rewrite  as:

. It is now clear that after we permute ,  remains unchanged, ’s also remain unchanged and we only need to re-compute .

Computational method used

Step 1: preparation. A) Compute  and B) apply the Gram-Schmidt process to transform . Remember that the first column of the transformed  should be 

Step 2: for each permuted sample , evaluate  and then compute  using the computed  and the stored .

Step 3: convert  to  by comparing  to an F-distribution with df = (p-1, n-p).

~~One more simplification~~

~~ and  are strictly monotonically related, i.e.  if and only if . Our way to calculate the permutation p value is just to count the number of  that is smaller or equal to the observed p value . This is equivalent to count the number of  that is larger or equal to the observed F-statistic . Therefore, the step 3 in the above procedure can be totally omitted and the permutation p value is calculated based on the  obtained in step 2.~~ (This only applies to the case for one single block. For the case when a gene has multiple blocks with different # of haplotypes, this doesn’t apply as the df for F-distribution could be different. Therefore, step 3 is necessary in general case.)