Investigating SNPs in the Talana Sample with Discordant $\mathbf{R_M}$ and $\mathbf{R_\chi}$ Results

General form of the $\mathbf{R_M}$ and $\mathbf{R_\chi}$ statistics

The ROADTRIPS association test statistics $\boldsymbol{R_\chi}$ and $\boldsymbol{R_M}$ can be viewed as having the common form

$$\frac{(\hat{p}_{null} - \hat{p}_{test})^2}{Var_0\,(\hat{p}_{null} - \hat{p}_{test})},$$

where $\hat{p}_{null}$ is an estimator of allele frequency calculated under the assumption of no association, $\hat{p}_{test}$ is a contrasting estimator of allele frequency that should have a different expectation from $\hat{p}_{null}$ when there is association, and $Var_0\,(\cdot)$ denotes variance calculated under the assumption that the null hypothesis of no association is true. So association between a genetic marker and a trait are essentially assessed for both $\boldsymbol{R_\chi}$ and $\boldsymbol{R_M}$ by comparing the allele frequency estimators $\hat{p}_{null}$ and $\hat{p}_{test}$.

$\hat{\boldsymbol{p}}_{\boldsymbol{null}}$ versus $\hat{\boldsymbol{p}}_{\boldsymbol{test}}$

Consider the problem of testing for association between a trait and a genetic marker in a case-control design. For simplicity, assume that the marker to be tested for association with the trait is a SNP, with alleles labelled "0" and "1". Let N be the number of individuals who are genotyped at the SNP, and let $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ be the genotype vector where $Y_i = 1/2 \times$ (the number of alleles of type 1 in individual i). So the value of Y_i is 0, 1/2, or 1.

The difference between $\boldsymbol{R_\chi}$ and $\boldsymbol{R_M}$ is in how $\hat{p}_{null}$ and $\hat{p}_{test}$ are calculated. The most general form of $\boldsymbol{R_M}$ can incorporate into the test statistic additional phenotype information for individuals who have missing genotype data at a SNP, provided that those individuals have a sampled relative who is genotyped at the SNP. For simplicity, we will consider the case where all phenotyped individuals also have genotype data at the SNP. We now give $\hat{p}_{null}$ and $\hat{p}_{test}$ for $\boldsymbol{R_\chi}$ and $\boldsymbol{R_M}$.

1. For $\boldsymbol{R_\chi}$ we have

• $\hat{p}_{test} = \frac{1}{n_c} \sum_{i \in cases} Y_i$

where n_c is the number of cases in the sample. So, $\hat{p}_{test}$ for $\boldsymbol{R_\chi}$ is just the sample mean based on cases.

For $\boldsymbol{R_\chi}$, $\hat{p}_{null}$ is the sample mean based on the entire sample:

• $\hat{p}_{null} = \frac{1}{N} \sum_i^N Y_i$

2. For $\boldsymbol{R_M}$ we have

• $\hat{p}_{test} = (\boldsymbol{A}^T\mathbf{1})^{-1}(\boldsymbol{A}^T\boldsymbol{Y})$

where $\mathbf{1}$ is a vector of 1's of length N and $\boldsymbol{A} = (A_1, \ldots, A_N)$ is a phenotype vector of length N where $A_i = 1$ if i is affected, $A_i = \frac{-k}{1-k}$ if i is unaffected, and $A_i = 0$ if i is of unknown phenotype, where $0 < k < 1$ is a constant that represents an external estimate of the population prevalence of the trait from a suitable reference population.

• $\hat{p}_{null} = \left(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\boldsymbol{Y}\right)$

where $\boldsymbol{\Phi}$ is the $N \times N$ kinship coefficient matrix for the sample individuals. Under the null hypothesis of no association, $\hat{p}_{null}$ for $\boldsymbol{R_M}$ is the best linear unbiased estimator for the frequency of allele 1 at the SNP.

Comparing $\mathbf{R_M}$ and $\mathbf{R}_\chi$

It should be noted that the $\hat{p}_{test}$ estimators used in the calculation of $\boldsymbol{R}_\chi$ and $\boldsymbol{R_M}$ are actually quite similar. For the calculation of $\hat{p}_{test}$ in $\boldsymbol{R}_{\chi,}$ cases are essentially given a weight of 1 while all controls (unaffected and unknown phenotype controls) are given a weight of 0. For $\hat{p}_{test}$ used in $\boldsymbol{R_M}$, a weight of 1 is given to the cases, a weight of 0 is given to individuals with unknown phenotype, and a weight of $(-k)/(1-k)$ is given to unaffected individuals. For traits that are rare, i.e., $k \approx 0$, $\hat{p}_{test}$ for the two test statistics will be almost identical.

In samples with unrelated individuals, the $\hat{p}_{null}$ estimators for $\boldsymbol{R}_\chi$ and $\boldsymbol{R_M}$ are actually equivalent. When there are related individuals included in a sample, however, $\hat{p}_{null}$ for $\boldsymbol{R}_\chi$ and $\boldsymbol{R_M}$ can give very different values. For samples from founder populations, where individuals are likely to be inbred and can be related through multiple lines of descent, the difference between the BLUE ($\hat{p}_{null}$ for $\boldsymbol{R_M}$) and the sample average allele frequency estimator ($\hat{p}_{null}$ for $\boldsymbol{R}_\chi$) can be substantial. In the next subsection, we show that for SNPs in the sample from the Talana founder population for which $\boldsymbol{R}_\chi$ and $\boldsymbol{R_M}$ do not give similar results, the $\hat{p}_{null}$ estimates used in the calculation of the two test statistics are usually substantially different.

Results for SNPs with Discordant $\mathbf{R_M}$ and $\mathbf{R}_\chi$ Values in Talana Sample

We investigated SNPs in the Talana sample for which $\boldsymbol{R_M}$ and $\boldsymbol{R}_\chi$ give discordant results. The table on page 4 gives $\hat{p}_{test}$ and $\hat{p}_{null}$ for 43 SNPs with the most discordant p-values for $\boldsymbol{R_M}$ and $\boldsymbol{R}_\chi$. As can be seen from the table, there is very little difference between the $\hat{p}_{test}$ values for $\boldsymbol{R_M}$ and $\boldsymbol{R}_\chi$ for these SNPs, as we expected. There is, however, a large difference between the $\hat{p}_{null}$ values for the two test statistics for most of the SNPs. Figure 1 is a histogram of the BLUE weights ($\hat{p}_{null}$ for $\boldsymbol{R_M}$) for the 842 individuals in the Talana sample. By comparing the broad range of weights given for the BLUE to the uniform weights given to all individuals for $\hat{p}_{null}$ in the $\boldsymbol{R}_\chi$ test, one can see how the results can be quite different for the two statistics.

We conjecture that the large difference observed for the $\hat{p}_{null}$ values for $\boldsymbol{R_M}$ and $\boldsymbol{R}_\chi$ is due to the small number of founders and the large amount of relatedness in this sample. Based on the kinship and inbreeding coefficients calculated from the known genealogical information for the 842 sample individuals in our study, when comparing the allele frequency variance of the BLUE for this sample to the number of independent (i.e., unrelated non-inbred) individuals that would give the same variance, we estimate the number of independent alleles in the sample to be equivalent to having approximately 61 founders in the sample, i.e., 61 independent individuals.

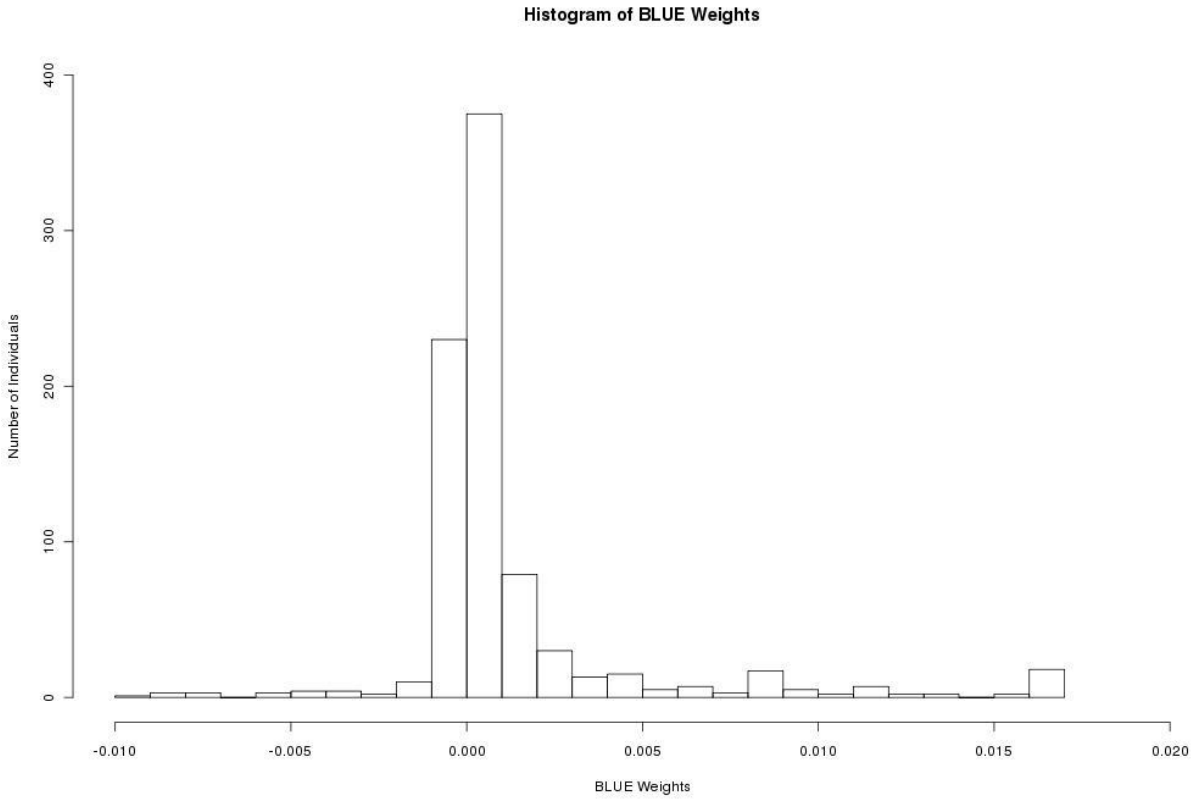
Discussion

For a small subset of SNPs, R_M and R_χ have extreme discordant p-values in the Talana sample. The difference in the p-values appears to be largely driven by the very different $\hat{p}_{null}$ estimates used in the calculation of R_χ and R_M for these SNPs. The different $\hat{p}_{null}$ estimates for the statistics are a result of the complex pedigree structure in the sample as well as the small number of founders. The BLUE, which is used to calculate R_M, adjusts for known relatedness, while the allele frequency estimate used in R_χ does not take into account pedigree information and is just the sample average of the entire sample. We also found that there is relatively little difference in the $\hat{p}_{test}$ estimators used in the two tests statistics for these 43 SNPs. We should point out that the phenotype vector, the weight vectors for both $\hat{p}_{test}$ and $\hat{p}_{null}$, and the empirical covariance matrix will jointly have an impact on the R_χ and R_M values, and it may be possible for R_M and R_χ to give different results even when the $\hat{p}_{null}$ estimates for the two statistics are similar.

SNP	$-log_{10}(pval_M)$	$-log_{10}(pval_\chi)$	$\hat{p}_{test_{RM}}$	$\hat{p}_{test_{R\chi}}$	$\hat{p}_{null_{RM}}$	$\hat{p}_{null_{R\chi}}$	*BLUE MAF*				*naive MAF*			
							cases	*controls*	*unknown controls*	*all*	*cases*	*controls*	*unknown controls*	*all*
1	0.49	4.36	0.2041	0.2055	0.1516	0.1005	0.20	0.09	0.12	0.15	0.21	0.10	0.09	0.10
2	1.77	2.20	0.1351	0.1507	0.2981	0.2512	0.17	0.31	0.32	0.30	0.16	0.34	0.25	0.25
3	1.19	4.40	0.1963	0.1918	0.1099	0.0914	0.12	0.07	0.11	0.11	0.20	0.07	0.08	0.09
4	1.18	4.34	0.1910	0.1875	0.1064	0.0883	0.12	0.07	0.11	0.11	0.20	0.07	0.08	0.09
5	1.05	4.35	0.1913	0.1901	0.1115	0.0888	0.12	0.05	0.11	0.11	0.20	0.06	0.08	0.09
6	1.11	4.13	0.1963	0.1918	0.1129	0.0939	0.12	0.07	0.11	0.11	0.20	0.07	0.09	0.10
7	0.96	4.22	0.1963	0.1918	0.1189	0.0930	0.12	0.07	0.12	0.12	0.20	0.07	0.09	0.09
8	2.92	0.16	0.1314	0.1319	0.3641	0.1437	0.20	0.18	0.37	0.36	0.13	0.13	0.15	0.14
9	3.30	1.12	0.1052	0.1164	0.3534	0.1734	0.18	0.23	0.35	0.35	0.11	0.16	0.18	0.17
10	2.11	1.05	0.1608	0.1849	0.3508	0.2473	0.30	0.32	0.33	0.35	0.19	0.26	0.25	0.25
11	4.01	0.98	0.0590	0.0685	0.3339	0.1120	0.12	0.12	0.33	0.33	0.07	0.10	0.12	0.11
12	1.17	4.45	0.1427	0.1357	0.0726	0.0549	0.11	0.06	0.08	0.07	0.13	0.06	0.05	0.05
13	2.47	0.17	0.0645	0.0685	0.2557	0.0780	0.14	0.12	0.25	0.26	0.08	0.08	0.08	0.08
14	1.21	4.35	0.5716	0.5548	0.4334	0.3863	0.54	0.37	0.42	0.43	0.55	0.41	0.36	0.39
15	1.10	4.23	0.5693	0.5548	0.4391	0.3886	0.54	0.39	0.42	0.44	0.55	0.42	0.36	0.39
16	2.22	0.03	0.0759	0.0822	0.2543	0.0840	0.18	0.18	0.20	0.25	0.08	0.10	0.08	0.08
17	2.06	0.07	0.0662	0.0764	0.2303	0.0806	0.16	0.17	0.18	0.23	0.08	0.10	0.08	0.08
18	2.46	0.06	0.0759	0.0822	0.2702	0.0863	0.18	0.18	0.22	0.27	0.08	0.10	0.08	0.09
19	2.48	0.10	0.0759	0.0822	0.2709	0.0884	0.18	0.18	0.22	0.27	0.08	0.10	0.09	0.09
20	2.48	0.10	0.0758	0.0822	0.2711	0.0886	0.18	0.18	0.22	0.27	0.08	0.10	0.09	0.09
21	2.48	0.10	0.0759	0.0822	0.2709	0.0883	0.18	0.18	0.22	0.27	0.08	0.10	0.09	0.09
22	2.48	0.10	0.0759	0.0822	0.2709	0.0883	0.18	0.18	0.22	0.27	0.08	0.10	0.09	0.09
23	2.48	0.08	0.0759	0.0822	0.2711	0.0875	0.18	0.18	0.22	0.27	0.08	0.10	0.09	0.09
24	1.25	4.28	0.3227	0.3151	0.2071	0.1826	0.32	0.22	0.19	0.21	0.32	0.20	0.17	0.18
25	3.15	1.06	0.0850	0.0822	0.3215	0.1312	0.17	0.19	0.28	0.32	0.09	0.13	0.14	0.13
26	3.69	0.61	0.1389	0.1507	0.4123	0.1892	0.22	0.19	0.41	0.41	0.16	0.18	0.19	0.19
27	3.53	0.62	0.1340	0.1479	0.3979	0.1872	0.21	0.19	0.40	0.40	0.15	0.18	0.19	0.19
28	1.34	2.96	0.3035	0.2808	0.1872	0.1755	0.30	0.11	0.16	0.19	0.28	0.10	0.17	0.18
29	1.17	2.90	0.3034	0.2808	0.1951	0.1765	0.30	0.11	0.17	0.20	0.28	0.10	0.18	0.18
30	2.29	0.65	0.1026	0.0959	0.2934	0.1306	0.20	0.19	0.25	0.29	0.09	0.12	0.14	0.13
31	1.13	1.53	0.2237	0.2260	0.3512	0.3113	0.29	0.39	0.34	0.35	0.23	0.44	0.30	0.31
32	1.20	1.53	0.2238	0.2260	0.3570	0.3114	0.29	0.39	0.35	0.36	0.23	0.44	0.30	0.31
33	0.56	4.28	0.2118	0.2055	0.1532	0.1018	0.18	0.08	0.14	0.15	0.20	0.07	0.10	0.10
34	0.98	4.32	0.3079	0.3438	0.2188	0.1826	0.35	0.13	0.21	0.22	0.34	0.16	0.17	0.18
35	0.76	4.44	0.2118	0.1986	0.1410	0.0956	0.22	0.10	0.13	0.14	0.20	0.10	0.08	0.10
36	0.76	4.36	0.2120	0.1986	0.1410	0.0964	0.22	0.10	0.13	0.14	0.20	0.10	0.08	0.10
37	1.02	4.80	0.5381	0.5274	0.4157	0.3524	0.51	0.42	0.39	0.42	0.53	0.43	0.32	0.35
38	2.59	0.26	0.0755	0.0809	0.2729	0.0962	0.12	0.13	0.27	0.27	0.08	0.07	0.10	0.10
39	2.54	0.24	0.0811	0.0822	0.2816	0.0962	0.14	0.18	0.26	0.28	0.09	0.13	0.09	0.10
40	2.54	0.25	0.0811	0.0822	0.2816	0.0967	0.14	0.18	0.26	0.28	0.09	0.13	0.10	0.10
41	5.58	0.77	0.3735	0.3767	0.6965	0.4346	0.51	0.51	0.66	0.70	0.37	0.42	0.44	0.43
42	0.09	2.00	0.5535	0.5274	0.5367	0.4198	0.60	0.36	0.50	0.54	0.53	0.31	0.42	0.42
43	1.10	4.35	0.2395	0.2329	0.1468	0.1203	0.20	0.09	0.12	0.15	0.24	0.12	0.11	0.12

In the table, $-log_{10}(pval_M)$ and $-log_{10}(pval_\chi)$ are the -log base 10 p-values for the R_M and R_χ statistics, respectively. The covariance matrix used in the calculation of R_M and R_χ in this table were based on a kinship coefficient matrix that was calculated using known genealogical information. $\hat{p}_{test\ RM}$ and $\hat{p}_{null_{RM}}$ are $\hat{p}_{test}$ and $\hat{p}_{null}$, respectively for R_M. Similarly, $\hat{p}_{test\ R\chi}$ and $\hat{p}_{null_{R\chi}}$ are $\hat{p}_{test}$ and $\hat{p}_{null}$, respectively for R_χ.

Figure 1


Histogram of BLUE Weights
Number of Individuals
0
100
200
300
400
-0.010
-0.005
0.000
0.005
0.010
0.015
0.020
BLUE Weights