Supplemental materials for “A versatile omnibus test for detecting mean and variance heterogeneity"

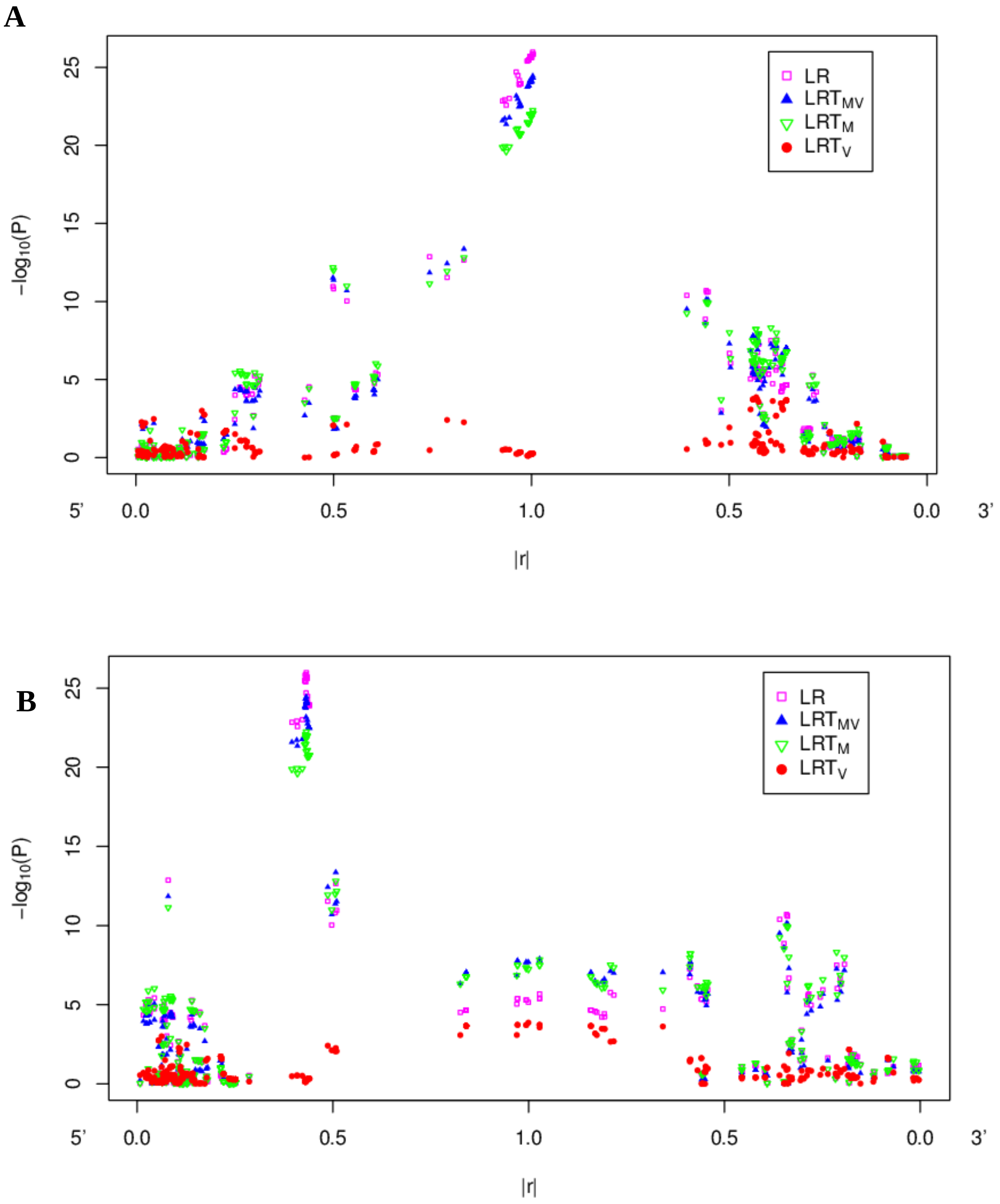
In this supplemental text, we first use real data-based simulation to demonstrate and then analytically show that variance heterogeneity can be induced due to linkage disequilibrium (LD) with a functional SNP with mean heterogeneity. We provide additional simulation results for (1) a lower significance level α=0.01 and (2) methods comparison for causal SNPs of lower MAF (MAF=0.2 and 0.1). We also show that the versatile omnibus test can detect mean and variance heterogeneity using additive genetic model. We demonstrate that, for a given non-normally distributed quantitative trait, the null distribution of the $\mathrm{LRT}_{\mathrm{MV}}$ test statistic does not depend on the SNP to be tested. We also present a simulation study to demonstrate that variance test by $\mathrm{LRT}_{\mathrm{V}}$ following rejecting the global null hypothesis by $\mathrm{LRT}_{\mathrm{MV}}$ can control the family-wise error rate (FWER) at the nominal level.

1. Real data-based simulation study for LD-induced variance heterogeneity

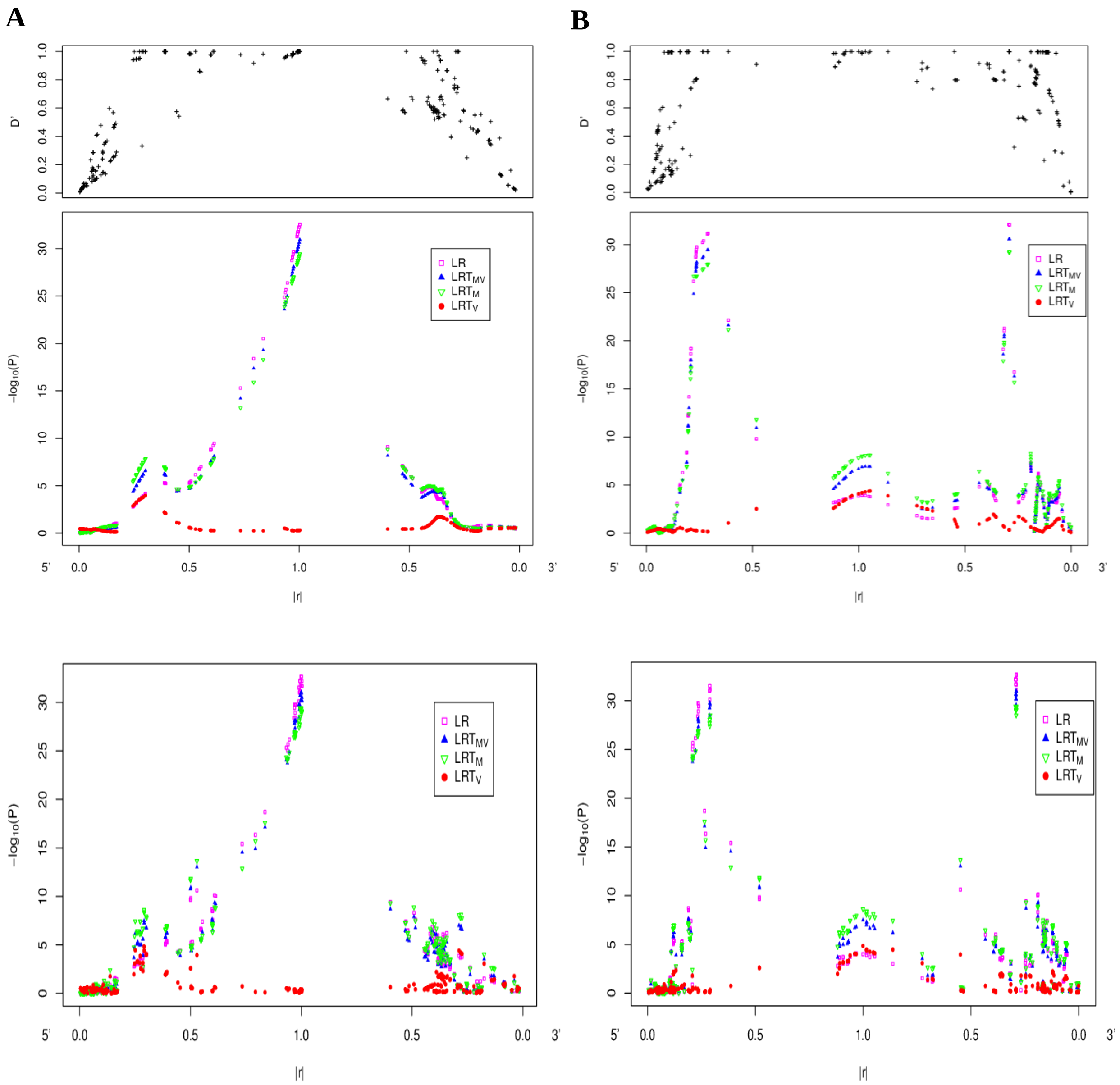
To confirm the observed pattern of variance heterogeneity due to LD with a functional SNP with mean heterogeneity (Figure 1 and Supplemental Figure 1), we performed simulation study using the *MMP3* genetic data with 394 subjects in all. We simulated a quantitative trait on the common variant rs679620 (MAF 0.48) and another quantitative trait on the uncommon variant rs1034375 (MAF 0.08). The quantitative traits were generated from $N(0,1)$, $N(1,1)$, and $N(2,1)$ corresponding to major allele homozygous, heterozygous, and minor allele homozygous. We tested SNPs within 100 kb of rs679620 for association with the simulated quantitative trait on rs679620 using $\mathrm{LRT}_{\mathrm{MV}}$, $\mathrm{LRT}_{\mathrm{M}}$, $\mathrm{LRT}_{\mathrm{V}}$, and LR (Supplemental Figure 2). Similarly, SNPs within 100 kb of rs1034375 were tested for association with the simulated quantitative trait on rs1034375 (Supplemental Figure 3). Consistent with the real data, the variance heterogeneity due to LD also peaks in a short interval where |r| is less than 0.5 ($r^2 < 0.25$) for the quantitative traits simulated on the common variant rs679620 and on the uncommon variant rs1034375.

In addition, we observed a distinct relationship between variance heterogeneity and LD measurement D’, compared to the relationship between variance heterogeneity and |r|. From Supplemental Figure 2A and 3A, we can see that variance heterogeneity peaks correspond to high D’, but low |r| with the functional SNP. For the quantitative trait simulated on the common variant rs679620, the SNP showing largest variance heterogeneity has MAF of 0.08, D’ of 0.999

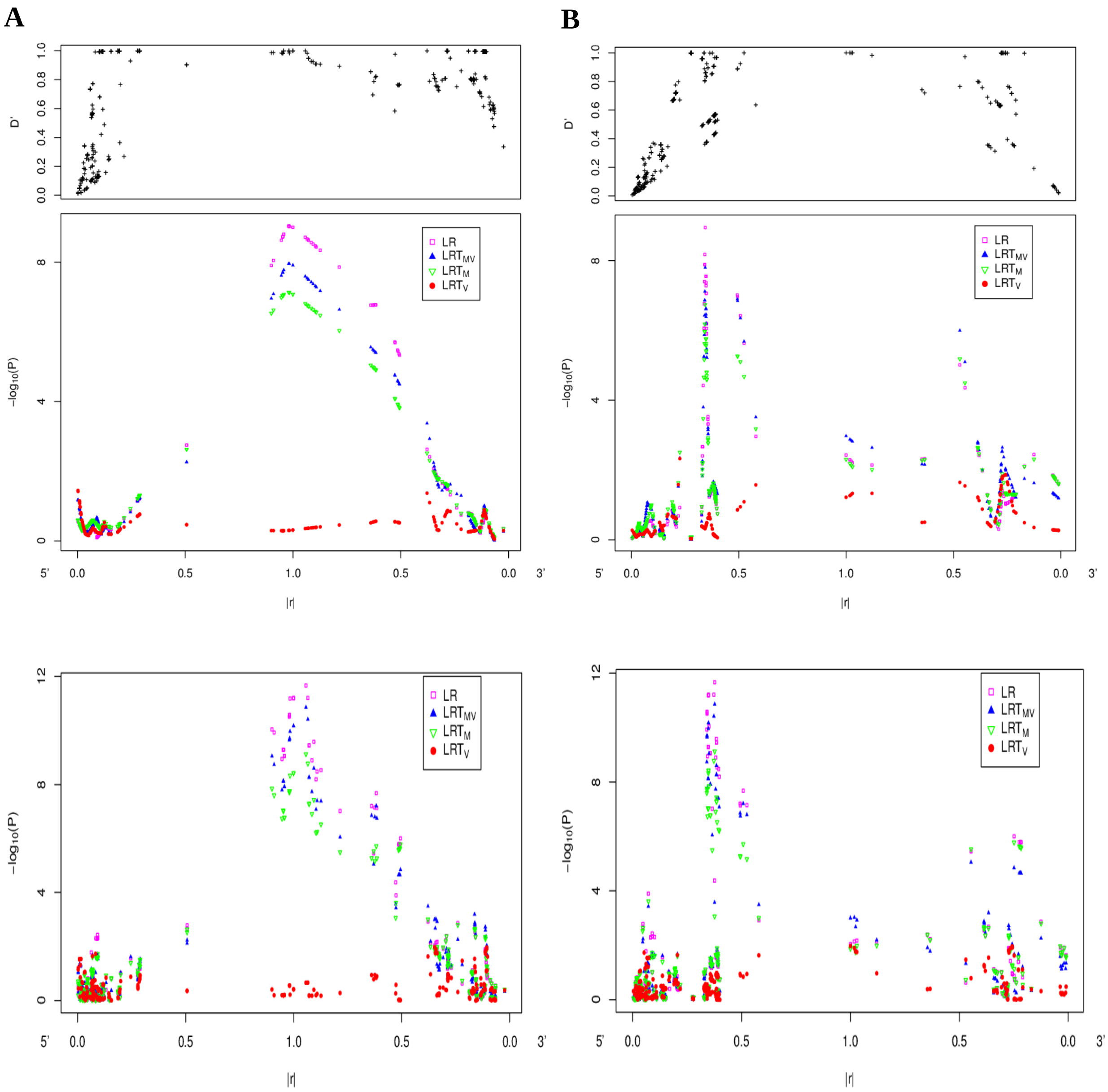
and |r| of 0.29 with rs679620. For the quantitative trait simulated on the uncommon variant rs1034375, the SNP showing largest variance heterogeneity has MAF of 0.32, D’ of 0.84 and |r| of 0.35 with rs1034375.



Supplemental Figure 1: Test statistics (-log$_{10}$(P values)) for association between SNPs within 100kb of functional SNP rs679620 and MMP3 protein levels in Cerebralspinal Fluid. **A**. Test statistics (-log$_{10}$(P values)) against LD (|r|) with the functional SNP rs679620, from 5’ and 3’ separately. **B**. Test statistics (-log$_{10}$(P values)) against LD (|r|) with the SNP having the smallest p-value of LRT$_V$, from 5’ and 3’ separately.



Supplemental Figure 2: Test statistics (-log$_{10}$(*P* values)) for association between SNPs within 100kb of rs679620 and a simulated quantitative trait on the common variant rs679620. **A**. Test statistics (-log$_{10}$(*P* values)) against LD with rs679620, from 5' and 3' separately. **B**. Test statistics (-log$_{10}$(*P* values)) against LD with the SNP having the smallest p-value of LRT$_{V}$, from 5' and 3' separately. **Top panel**: Plot of D' against |r|. **Middle panel**: Lowess of test statistics (-log$_{10}$(*P* values)) against |r|. **Bottom panel**: Unsmoothed plot of test statistics (-log$_{10}$(*P* values)) against |r|.



Supplemental Figure 3: Test statistics (-$\log_{10}$(P values)) for association between SNPs within 100kb of rs1034375 and a simulated quantitative trait on the uncommon variant rs1034375. **A**. Test statistics (-$\log_{10}$(P values)) against LD with rs1034375, from 5’ and 3’ separately. **B**. Test statistics (-$\log_{10}$(P values)) against LD with the SNP having the smallest p-value of LRT_V, from 5’ and 3’ separately. **Top panel**: Plot of D’ against |r|. **Middle panel**: Lowess of test statistics (-$\log_{10}$(P values)) against |r|. **Bottom panel**: Unsmoothed plot of test statistics (-$\log_{10}$(P values)) against |r|.

2. Tests comparison when the significance level α was set at 0.01

Different tests were compared when the significance level α was set at 0.01 (Supplemental Table 1). This was based on simulation studies for a SNP with MAF of 0.4 and 1000 replicates. The relative performance of different tests followed what was observed from Table I.

Supplemental Table 1: A comparison of empirical Type I error/power of different tests in four simulated scenarios when the significance level α was set at 0.01 for a SNP with MAF of 0.4. PB:Parametric Bootstrap; LR:Linear Regression; KW:Kruskal-Wallis; FK:Fligner-Killeen.

A. Simulated normally distributed quantitative traits.

simulated	joint tests			mean tests				variance tests			
effects	LRT_{MV}	LRT_{MV}(PB)	Lepage	LRT_{M}	LR	KW	$DGLM_{M}$	LRT_{V}	Levene	FK	$DGLM_{V}$
no effect	0.007	0.007	0.010	0.004	0.005	0.007	0.004	0.010	0.015	0.015	0.010
mean	0.593	0.588	0.656	0.710	0.713	0.676	0.710	0.013	0.014	0.015	0.013
variance	0.557	0.550	0.576	0.009	0.015	0.009	0.009	0.647	0.589	0.578	0.647
mean & var	0.912	0.910	0.931	0.619	0.692	0.639	0.619	0.662	0.585	0.570	0.662

B. Simulated non-normally distributed quantitative traits.

simulated	joint tests		mean tests			variance tests		
effects	LRT_{MV}(PB)	Lepage	LRT_{M}	LR	KW	LRT_{V}(PB)	Levene	FK
no effect	0.011	0.009	0.007	0.007	0.010	0.012	0.007	0.006
mean	0.236	0.797	0.695	0.702	0.817	0.014	0.009	0.008
variance	0.619	0.216	0.007	0.023	0.008	0.630	0.427	0.214
mean & var	0.889	0.934	0.598	0.680	0.806	0.648	0.461	0.249

C. Tests that cannot control Type I error for non-normally distributed quantitative traits.

simulated effects	LRT_{MV}	LRT_{V}	$DGLM_{V}$	$DGLM_{V}$ Box-Cox transformation
no effect	0.140	0.186	0.186	0.069

3. Tests comparison when the SNP to be tested had lower MAF

To further compare different tests when the SNP to be tested had a lower MAF, we performed simulation studies for a SNP with MAF of 0.2 (Supplemental Table 2) and a SNP with MAF of 0.1 (supplemental Table 3). All the parameters used for simulation studies of SNPs with MAF of 0.2 and 0.1 were the same as those in the simulation studies for the SNP with MAF of 0.4 (see Materials and Methods in the main text). Empirical power/Type I error was calculated as the proportion of replicates at the significance threshold level of 0.05. The advantages and disadvantages of different tests still hold when MAF of the SNP decreases. Given the total sample size of 1000 in our simulation studies, the powers of all the tests decreased as the MAF of a SNP decreased. This is expected because the number of heterozygotes and rare homozygotes decreases as the MAF decreases. In addition, type I errors of parametric tests ($LRT_{MV,}$ LRT_M and LRT_V) were not well controlled for normally distributed quantitative traits when MAF of the simulated SNP is 0.1 due to small sample size of rare homozygotes. Parametric bootstrap can alleviate the type I error inflation (supplemental Table 3). The LRT_{MV} test was more powerful when both mean and variance heterogeneities existed. LRT_V test remained the most powerful test for variance heterogeneity.

Supplemental Table 2: Comparison of empirical Type I error/power of different tests in four simulated scenarios for a SNP with MAF of 0.2.
PB:Parametric Bootstrap; LR:Linear Regression; KW:Kruskal-Wallis; FK:Fligner-Killeen.

A. Simulated normally distributed quantitative traits.

simulated	joint tests			mean tests				variance tests			
effects	LRT$_{MV}$	LRT$_{MV}$(PB)	Lepage	LRT$_{M}$	LR	KW	DGLM$_{M}$	LRT$_{V}$	Levene	FK	DGLM$_{V}$
no effect	0.048	0.044	0.051	0.050	0.044	0.042	0.050	0.047	0.045	0.044	0.047
mean	0.482	0.482	0.565	0.601	0.618	0.586	0.601	0.055	0.036	0.039	0.055
variance	0.478	0.468	0.571	0.041	0.087	0.062	0.041	0.589	0.527	0.516	0.589
mean & var	0.806	0.792	0.850	0.544	0.619	0.559	0.544	0.543	0.505	0.504	0.543

B. Simulated non-normally distributed quantitative traits.

simulated	joint tests		mean tests			variance tests		
effects	LRT$_{MV}$(PB)	Lepage	LRT$_{M}$	LR	KW	LRT$_{V}$(PB)	Levene	FK
no effect	0.061	0.058	0.048	0.050	0.047	0.062	0.060	0.060
mean	0.245	0.693	0.613	0.609	0.701	0.057	0.039	0.045
variance	0.633	0.307	0.041	0.121	0.051	0.646	0.473	0.298
mean & var	0.864	0.853	0.573	0.672	0.713	0.700	0.525	0.323

C. Tests that cannot control Type I error for non-normally distributed quantitative traits.

simulated effects	LRT$_{MV}$	LRT$_{V}$	DGLM$_{V}$	DGLM$_{V}$ Box-Cox transformation
no effect	0.255	0.323	0.323	0.184

Supplemental Table 3: Comparison of empirical Type I error/power of different tests in four simulated scenarios for a SNP with MAF of 0.1.
PB:Parametric Bootstrap; LR:Linear Regression; KW:Kruskal-Wallis; FK:Fligner-Killeen.

A. Simulated normally distributed quantitative traits.

simulated	joint tests			mean tests					variance tests				
effects	LRT$_{MV}$	LRT$_{MV}$(PB)	Lepage	LRT$_{M}$	LRT$_{M}$(PB)	LR	KW	DGLM$_{M}$	LRT$_{V}$	LRT$_{V}$(PB)	Levene	FK	DGLM$_{V}$
no effect	0.066	0.055	0.056	0.064	0.050	0.044	0.043	0.064	0.076	0.055	0.059	0.058	0.076
mean	0.256	0.229	0.289	0.324	0.285	0.325	0.285	0.324	0.067	0.052	0.052	0.050	0.067
variance	0.287	0.258	0.291	0.081	0.065	0.120	0.095	0.081	0.297	0.246	0.273	0.253	0.297
mean & var	0.495	0.460	0.563	0.325	0.283	0.405	0.356	0.325	0.284	0.242	0.266	0.256	0.284

B. Simulated non-normally distributed quantitative traits.

simulated	joint tests		mean tests				variance tests		
effects	LRT$_{MV}$(PB)	Lepage	LRT$_{M}$	LRT$_{M}$(PB)	LR	KW	LRT$_{V}$(PB)	Levene	FK
no effect	0.056	0.056	0.064	0.052	0.061	0.050	0.054	0.061	0.065
mean	0.127	0.381	0.365	0.334	0.335	0.406	0.045	0.042	0.043
variance	0.492	0.237	0.055	0.040	0.161	0.062	0.482	0.350	0.226
mean & var	0.628	0.580	0.358	0.313	0.478	0.431	0.472	0.316	0.194

C. Tests that cannot control Type I error for non-normally distributed quantitative traits.

simulated effects	LRT$_{MV}$	LRT$_{V}$	DGLM$_{V}$	DGLM$_{V}$ Box-Cox transformation
no effect	0.240	0.310	0.310	0.177

4. Detect mean and variance heterogeneity using additive genetic model

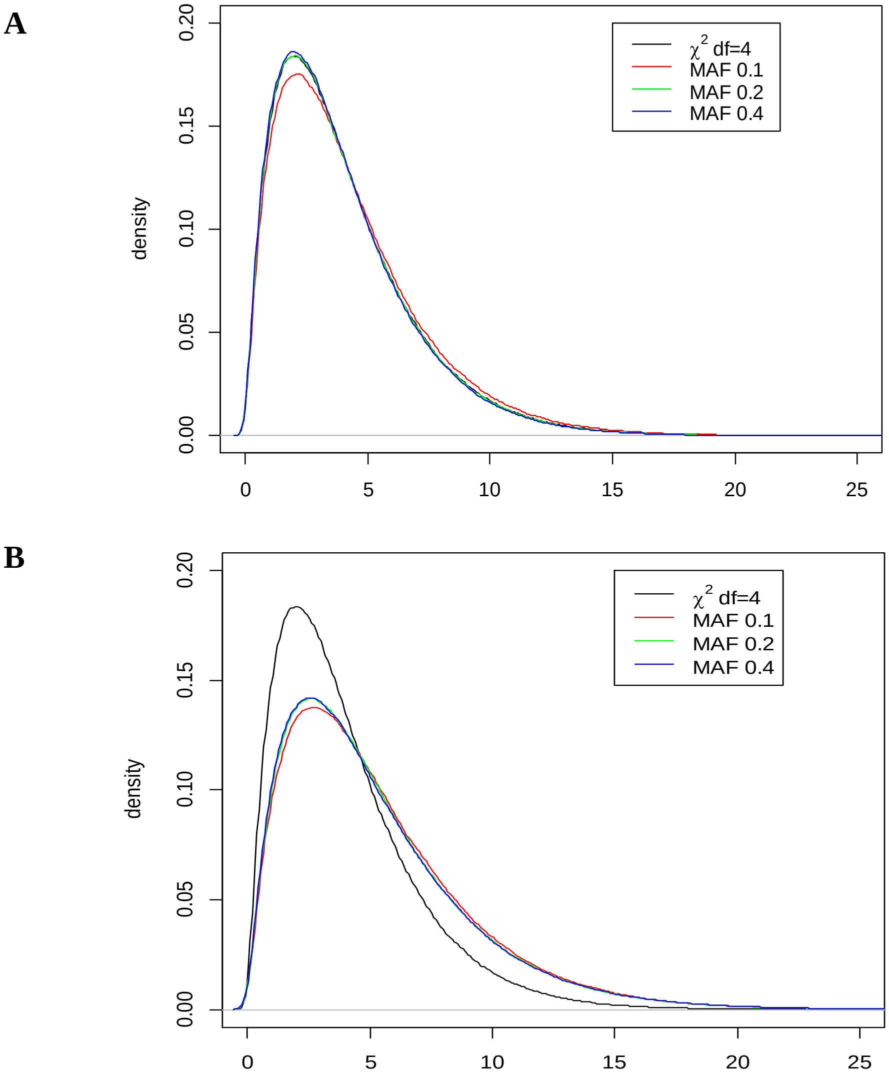
In addition to modeling means and variances as genotypes, the omnibus test can also detect mean and variance heterogeneity using additive genetic model. We performed simulation study using an additive genetic model. A common SNP with MAF of 0.4 was simulated. We considered four scenarios: 1. Genotypes have no effects on quantitative traits; 2. Genotypes have additive effects on means of quantitative traits; 3. Genotypes have additive effects on variances of quantitative traits; and 4. Genotypes have additive effects on both means and variances of quantitative traits. The quantitative traits (y_i) were generated using the model: $y_i = X_i\beta + \varepsilon_i$, where X_i is the number of minor alleles the ith subjects carries. Without mean effects, $\beta = 0$; when genotypes affect means, $\beta = -0.03$. We simulated ε_i from $N(0,\ \ 0.05)$ for scenarios without variance effects. For scenarios with variance effects, ε_i was generated from $N(0,\ \ 0.05)$, $N(0,\ \ 0.065)$, and $N(0,\ \ 0.08)$ corresponding to major allele homozygous, heterozygous, and minor allele homozygous, respectively. For each scenario, 1000 replicates were simulated with sample size of 1000 in each replicate. Empirical power/Type I error was calculated as the proportion of replicates with statistically significant effects at the threshold level of 0.05 (Supplemental Table 4). Type I errors were well controlled for all tests.

Supplemental Table 4: Simulation study using additive genetic model.

simulated	tests		
effects	LRT$_{MV}$	LRT$_{M}$	LRT$_{V}$
no effect	0.047	0.055	0.049
mean	0.642	0.746	0.040
variance	0.924	0.046	0.956
mean & var	0.990	0.758	0.957

5. Construction of null distribution of $\mathrm{LRT_{MV}}$ test statistic for a non-normally distributed quantitative trait

For a given non-normally distributed quantitative trait, the null distribution of $\mathrm{LRT_{MV}}$ test statistic does not depend on the SNP to be tested. We simulated a normally distributed quantitative trait and a quantitative trait from t-distribution (df=5) with sample size of 1000, and three SNPs with MAF of 0.4, 0.2 and 0.1, respectively. We constructed null distribution of $\mathrm{LRT_{MV}}$ test statistic for each simulated quantitative trait with respect to each SNP to be tested by repeating parametric bootstrap 10^6 times (Supplemental Figure 4). For the normally distributed quantitative trait, the null distributions match the χ^2distribution with 4 degrees of freedom. For the non-normally distributed quantitative trait, the null distributions of $\mathrm{LRT_{MV}}$ test statistic are different from χ^2distribution (df=4), but does not change with respect to the SNP to be tested. Due to small sample size of rare homozygotes for the SNP with MAF of 0.1 (only 10 rare homozygotes on average), the null distribution constructed using parametric bootstrap shows subtle difference for both normally and non-normally distributed quantitative traits. Based on these, we suggest constructing a null distribution of $\mathrm{LRT_{MV}}$ test statistic for each non-normally distributed quantitative trait using parametric bootstrap, which can be used to test for any SNPs with respect to the quantitative trait. This is much more computationally efficient than doing parametric bootstrap for each SNP. In addition, depending on the sample size, SNPs with MAF below certain threshold may be excluded from vQTL analysis due to low power and unreliable results.



Supplemental Figure 4: Null distributions of $\mathrm{LRT}_{\mathrm{MV}}$ test statistic for simulated quantitative traits constructed using parametric bootstrap for SNPs with MAF of 0.4, 0.2, and 0.1, respectively. **A**. A normally distributed quantitative trait. **B**. A non-normally distributed quantitative trait.

6 Analytical derivation for LD-induced variance heterogeneity

In this section, we first analytically show that variance heterogeneity can be induced due to LD with a functional locus with only mean effect. We will then demonstrate this using simulations.

6.1 Derivation of LD-induced variance heterogeneity

Assume two SNPs in LD denoted by G_1 and G_2, where G_1 is a functional SNP with mean and/or variance effect on the quantitative phenotype. We denote the major and minor alleles of G_1 and G_2 by A, a and B, b, respectively. We use the following marginal and conditional probabilities to model the LD between G_1 and G_2. Specifically, we have

$$\begin{aligned} p(G_1 = A) &= p_A, \\ p(G_1 = a) &= p_a = 1 - p_A, \\ p(G_2 = B|G_1 = A) &= p_{B|A}, \\ p(G_2 = B|G_1 = a) &= p_{B|a}. \end{aligned}$$

If $p_{B|A} = p_{B|a}$, G_1 and G_2 are in linkage equilibrium; otherwise, G_1 and G_2 are in LD. It also follows that $p(G_2 = B) = p(G_2 = B|G_1 = A)p(G_1 = A) + p(G_2 = B|G_1 = a)p(G_1 = a) = p_{B|A} \times p_A + p_{B|a} \times (1 - p_A)$. The probabilities of the four possible haplotypes, i.e., AB, Ab, aB, ab, are

$$\begin{aligned} p_{AB} &= p_{B|A} \times p_A, \\ p_{Ab} &= (1 - p_{B|A}) \times p_A, \\ p_{aB} &= p_{B|a} \times (1 - p_A), \\ p_{ab} &= (1 - p_{B|a}) \times (1 - p_A). \end{aligned}$$

Based on the above haplotype probabilities, we can easily calculate LD measures, such as D' and r^2.

For simplicity of exposition, we first assume a haploid model and then extend to a diploid model in the simulation study. Assume the quantitative trait Y depends on the functional SNP G_1 via:

$$\begin{aligned} (y|G_1 = A) &\sim N(\mu_A, \sigma_A^2), \\ (y|G_1 = a) &\sim N(\mu_a, \sigma_a^2). \end{aligned}$$

We now derive the mean and variance differences at G_2, i.e., $E(y|G_2 = B) - E(y|G_2 = b)$ and $Var(y|G_2 = B) - Var(y|G_2 = b)$. We need the following four conditional probabilities in the derivation:

$$\begin{aligned}
p_{a|B} &= p(G_1 = a|G_2 = B) \\
&= \frac{p(G_2 = B|G_1 = a)p(G_1 = a)}{p(G_2 = B)} \\
&= \frac{p_{B|a} - p_A \times p_{B|a}}{p_{B|A} \times p_A + p_{B|a} \times (1 - p_A)}, \\
p_{A|B} &= p(G_1 = A|G_2 = B) \\
&= 1 - p_{a|B}, \\
p_{a|b} &= p(G_1 = a|G_2 = b) \\
&= \frac{1 - p_A - p_{B|a} + p_A \times p_{B|a}}{1 - p_{B|A} \times p_A - p_{B|a} + p_A \times p_{B|a}}, \\
p_{A|b} &= p(G_1 = A|G_2 = b) \\
&= 1 - p_{a|b}.
\end{aligned}$$

We have the conditional mean of the phenotype given $G_2 = b$,

$$\begin{aligned}
E(y|G_2 = b) &= \int y f(y|G_2 = b)dy \\
&= \int y \frac{f(y, G_2 = b)}{p(G_2 = b)} dy \\
&= \int y \frac{f(y, G_2 = b|G_1 = a)p(G_1 = a) + f(y, G_2 = b|G_1 = A)p(G_1 = A)}{p(G_2 = b)} dy \\
\text{(since } y \perp G_2|G_1) &= \int y \frac{f(y|G_1 = a)p(G_2 = b|G_1 = a)p(G_1 = a)}{p(G_2 = b)} dy \\
&\quad + \int y \frac{f(y|G_1 = A)p(G_2 = b|G_1 = A)p(G_1 = A)}{p(G_2 = b)} dy \\
&= \int y f(y|G_1 = a)p(G_1 = a|G_2 = b)dy + \int y f(y|G_1 = A)p(G_1 = A|G_2 = b)dy \\
&= p(G_1 = a|G_2 = b)E(y|G_1 = a) + p(G_1 = A|G_2 = b)E(y|G_1 = A) \\
&= p_{a|b} \times \mu_a + p_{A|b} \times \mu_A.
\end{aligned}$$

Similarly, we can derive,

$$E(y|G_2 = B) = p_{a|B} \times \mu_a + p_{A|B} \times \mu_A.$$

Thus, the conditional mean difference is,

$$\begin{aligned}
E(y|G_2 = B) - E(y|G_2 = b) &= (p_{a|B} - p_{a|b})\mu_a + (p_{A|B} - p_{A|b})\mu_A \\
&= \left[\frac{p_A(p_{B|a} - p_{B|A})(1 - p_A)}{(p_{B|A}p_A + p_{B|a} - p_A p_{B|a})(1 - p_{B|A}p_A - p_{B|a} + p_A p_{B|a})}\right]\mu_a \\
&\quad + \left[\frac{p_A(p_{B|A} - p_{B|a})(1 - p_A)}{(p_{B|A}p_A + p_{B|a} - p_A p_{B|a})(1 - p_{B|A}p_A - p_{B|a} + p_A p_{B|a})}\right]\mu_A.
\end{aligned}$$

When $p_{B|A} = p_{B|a}$, i.e., G_1 and G_2 are in linkage equilibrium, the mean difference is 0; when $p_{B|A} \neq p_{B|a}$, we observe mean difference at G_2 induced by G_1 due to LD. These results are consistent with those in association study of mean difference in the phenotype, i.e., QTL study.

For conditional variance at G_2, we have

$$\begin{aligned}
Var(y|G_2 = b) &= E(y^2|G_2 = b) - [E(y|G_2 = b)]^2 \\
&= \big[p(G_1 = a|G_2 = b)E(y^2|G_1 = a) + p(G_1 = A|G_2 = b)E(y^2|G_1 = A\big] \\
&\quad - [p(G_1 = a|G_2 = b)E(y|G_1 = a) + p(G_1 = A|G_2 = b)E(y|G_1 = A)]^2 \\
&= p_{a|b} \times (\sigma_a^2 + \mu_a^2) + p_{A|b} \times (\sigma_A^2 + \mu_A^2) - (p_{a|b} \times \mu_a + p_{A|b} \times \mu_A)^2.
\end{aligned}$$

Similarly,

$$Var(y|G_2 = B) = p_{a|B} \times (\sigma_a^2 + \mu_a^2) + p_{A|B} \times (\sigma_A^2 + \mu_A^2) - (p_{a|B} \times \mu_a + p_{A|B} \times \mu_A)^2$$

Therefore,

$$\begin{aligned}
Var(y|G_2 = B) - Var(y|G_2 = b) &= (p_{a|B} - p_{a|b})(1 - p_{a|B} - p_{a|b})\mu_a^2 + (p_{A|B} - p_{A|b})(1 - p_{A|B} - p_{A|b})\mu_A^2 \\
&\quad -2(p_{a|B}p_{A|B} - p_{a|b}p_{A|b})\mu_a\mu_A + (p_{a|B} - p_{a|b})\sigma_a^2 + (p_{A|B} - p_{A|b})\sigma_A^2
\end{aligned}$$

Some remarks,

1. When $p_{a|B} = p_{a|b}$, and thus, $p_{A|B} = p_{A|b}$, G_1 and G_2 are in linkage equilibrium, and the conditional variance difference at G_2 is 0.

2. Even if $\sigma_a^2 = \sigma_A^2$, i.e., only mean effect at the functional SNP G_1, when G_1 and G_2 are in LD, the conditional variance difference at G_2 is not 0, as confirmed by Supplemental Figures 5 & 6.

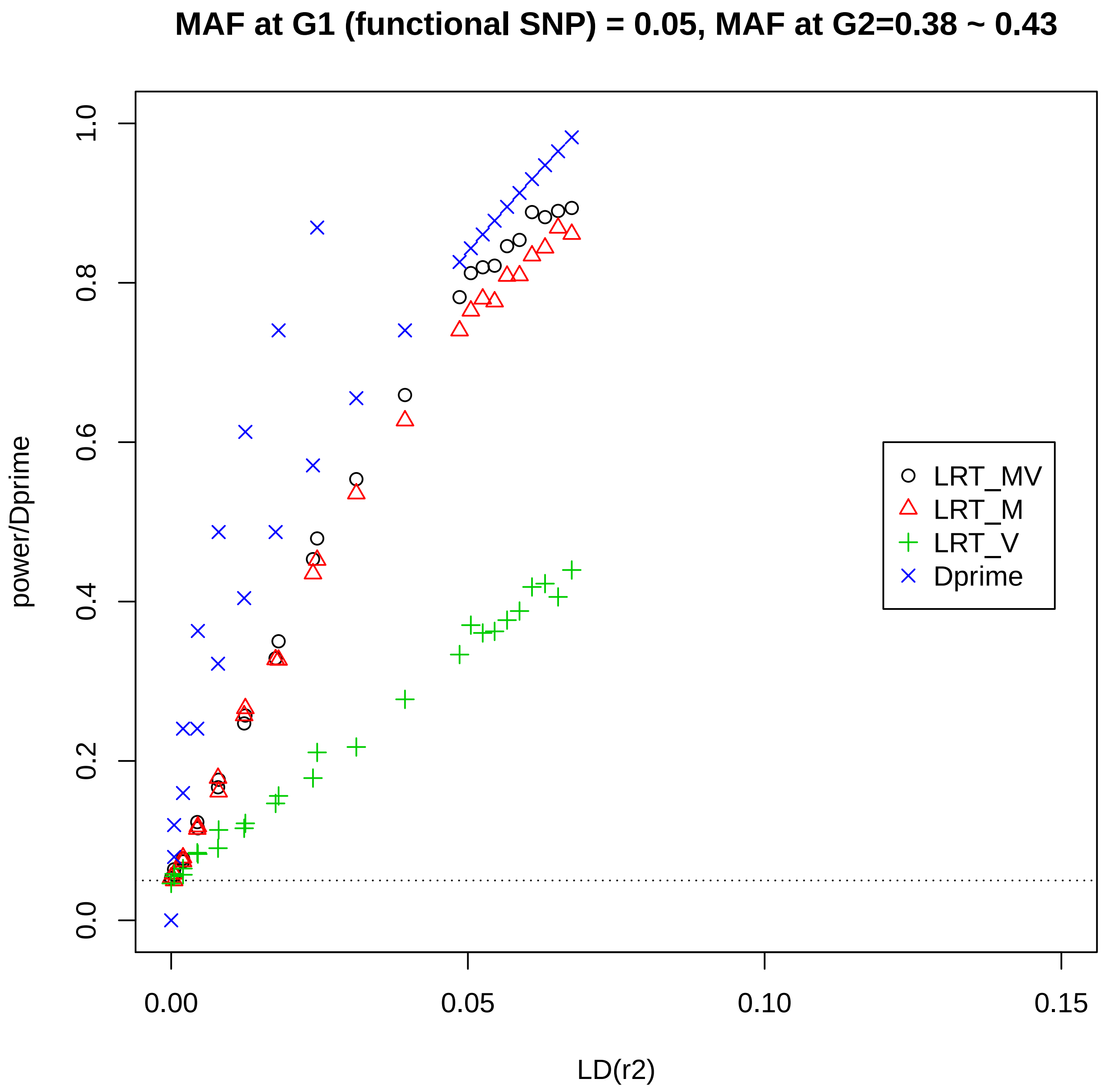
3. The derivations here do not depend on the normality assumption.

6.2 Simulation study to verify the analytical results of LD-induced variance heterogeneity

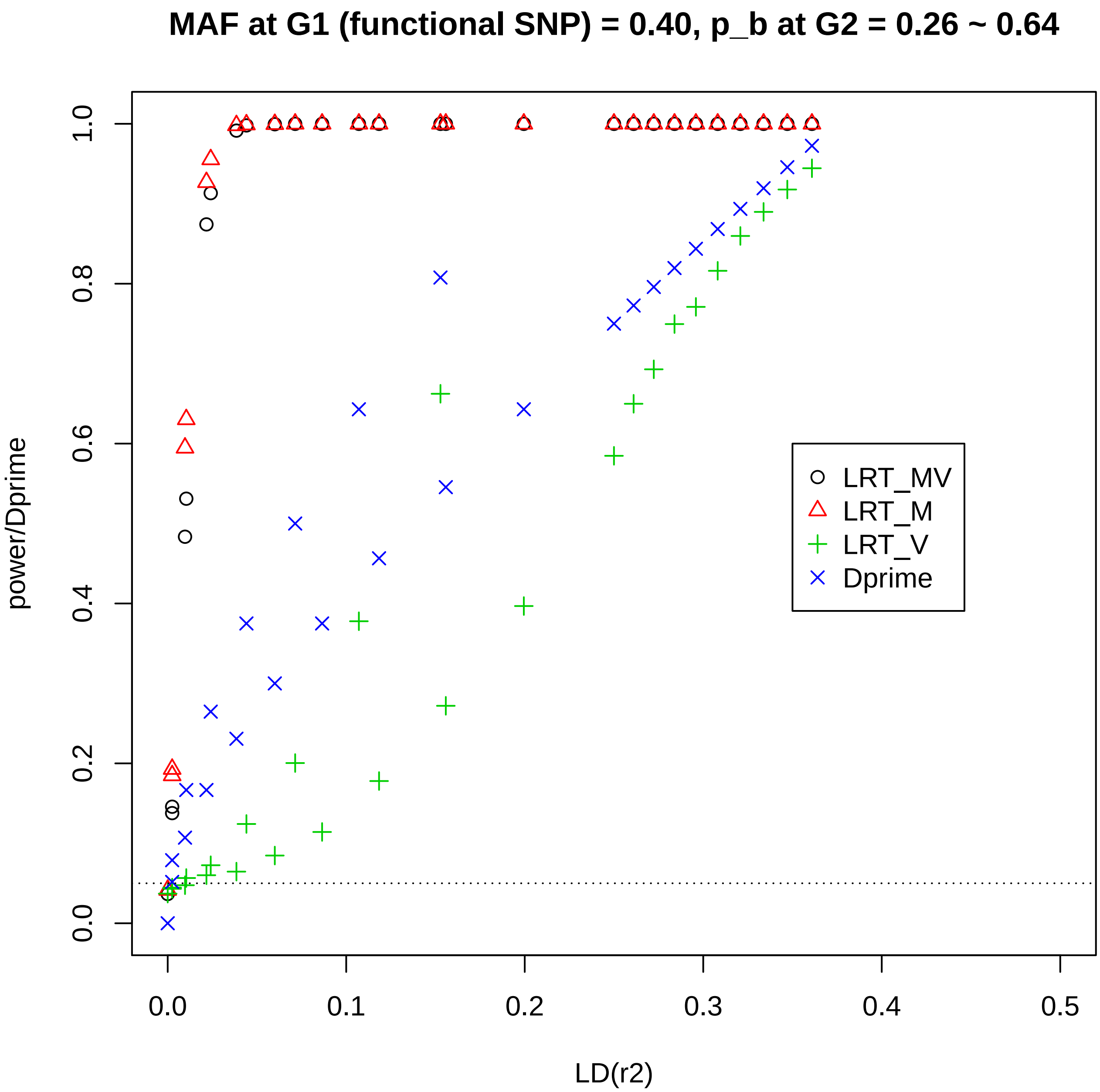
We performed simulation study to verify and illustrate the above analytical results. We simulated 2000 replicated datasets of sample size $n = 2000$ with each subject's diplotype formed by two haplotypes drawn independently from AB, Ab, aB, ab according to their probabilities. We assume $\mu_A = 0$, $\mu_a = 1$ and an additive genetic model so that the mean phenotype increase is 1 per minor allele at the functional SNP G_1, while $\sigma_A^2 = \sigma_a^2 = 1$, i.e., the variance is constant across G_1's three genotypes. We considered two scenarios: low frequency and common functional SNP G_1.

Scenario 1: Low frequency functional SNP MAF of G_1 is $p_a = 0.05$, $p_{b|a} = 0.99, 0.98, \ldots, 0.90, 0.85, \ldots, 0.05$ and $p_{b|A} = 0.4$. The resulting MAF of G_2 ranges from 0.38 to 0.43. The maximum possible r^2 between G_1 and G_2 is around 0.085, due to the result of Wray (2005), i.e., the maximum value for r^2 is the smaller of $\frac{p_A(1-p_B)}{(1-p_A)p_B}$ and its inverse. As demonstrated in Supplemental Figure 5, a functional SNP with mean effect only can induce variance heterogeneity at a SNP in LD with it. Due to the MAF difference between the two SNPs, large D' increases the variance heterogeneity, while r^2 remains quite small (less that 0.1).

Scenario 2: Common functional SNP Minor allele frequency (MAF) of G_1 is $p_a = 1 - p_A = 0.4$, $p_{b|a} = 0.99, 0.98, \ldots, 0.90, 0.85, \ldots, 0.05$, and $p_{b|A} = 0.4$. $p_{b|a}$ closer to 1 indicates that the minor allele of G_2 is more likely to be present with the minor allele of G_1, leading to higher LD between G_1 and G_2. On the other hand, $p_{b|a}$ closer to 0 indicates that the minor allele of G_2 is more likely to be present with the major allele of G_1, again resulting in higher LD between G_1 and G_2. The resulting alternative allele frequency of G_2, i.e., p_b, ranges from 0.26 to 0.64. The maximum possible r^2 between G_1 and G_2 is around 0.53. As demonstrated in Supplemental Figure 6, the results are similar to those in Scenario 1 (low frequency functional SNP): large D' increases the variance heterogeneity and r^2 remains moderate, though the latter is larger than that in Scenario 1 due to both SNPs being common.



Supplemental Figure 5: Low frequency functional SNP: MAF of causal SNP G_1: $p_a = 0.05$, MAF of G_2: $p_b = 0.38 \sim 0.43$, maximum possible r^2 between G_1 and G_2 is 0.085; $\mu_A = 0, \mu_a = 1, \sigma_a^2 = \sigma_A^2 = 1$, sample size = 2000, replications = 2000. The horizontal dotted line is $\alpha = 0.05$.



Supplemental Figure 6: Common functional SNP: MAF of causal SNP G_1: $p_a = 0.40$, alternative allele frequency of G_2: $p_b = 0.26 \sim 0.64$, maximum possible r^2 between G_1 and G_2 is 0.53; $\mu_A = 0, \mu_a = 1, \sigma_a^2 = \sigma_A^2 = 1$, sample size = 2000, replications = 2000. The horizontal dotted line is $\alpha = 0.05$.

7 Variance tests after rejecting the global null hypothesis

In this section, we performed a simulation study to demonstrate that variance test by LRT_V following rejecting the global null hypothesis by LRT_{MV} can control the family-wise error rate (FWER) at the nominal level.

7.1 Simulation setup

Define three hypotheses:

- Mean test: $H_{0M} : \mu_0 = \mu_1 = \mu_2$ versus H_{1M} : at least one " = " not hold
- Variance test: $H_{0V} : \sigma_0^2 = \sigma_1^2 = \sigma_2^2$ versus H_{1V} : at least one " = " not hold
- Global test: $H_{0MV} : \mu_0 = \mu_1 = \mu_2$ and $\sigma_0^2 = \sigma_1^2 = \sigma_2^2$ versus H_{1MV} : at least one of H_{1M} and H_{1V} holds

We simulated 100 SNPs assuming HWE: $X_{ij} \sim Binomial(2, p = 0.4)$ for $i = 1, \dots, n$ and $j = 1, \dots, 100$. We simulated the phenotype in the following two scenarios.

Scenario 1: all 100 SNPs are null The phenotype y_i was simulated from $N(0, 1)$. Thus, $H_{0MV}^{(j)}$, $H_{0M}^{(j)}$ and $H_{0V}^{(j)}$ are true for $j = 1, 2, \dots, 100$.

Scenario 2: 1 SNP with mean effect only and 99 null SNPs The phenotype $y_i = \beta X_{i1} + \epsilon_i$, where $\beta = 0.1$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. Thus, $H_{1MV}^{(j)}$ and $H_{1M}^{(j)}$ are true for $j = 1$, while $H_{0V}^{(j)}$ are true for $j = 1, \dots, 100$ and $H_{0MV}^{(j)}$ and $H_{0M}^{(j)}$ are true for $j = 2, \dots, 100$.

The empirical type I error was based on a sample size of $n = 2000$ and 2000 simulation replications. The proposed test procedure is as follows:

1. Test each of 100 SNPs by LRT_{MV} and use significance threshold $0.05/100 = 5 \times 10^{-4}$.
2. For m significant global tests, we use LRT_V to test variance heterogeneity at the Bonferroni correction level $0.05/m$.

We are interested in controlling the FWER, defined as

$$Pr(\text{at least one true } H_0^{(j)} \text{ among } J \text{ tests is falsely rejected}).$$

The commonly Bonferroni correction is aimed at controlling the FWER at, say, $\alpha = 0.05$ among J independent tests. For the global test, we look at all the 100 SNPs in Scenario 1 and the last 99 SNPs in Scenario 2, while we look at all 100 SNPs for the variance test in Scenarios 1 & 2. The empirical FWER is summarized in Supplemental Table 5. In both scenarios, the global test (LRT_{MV}) remained the nominal FWER at 0.05, while the variance test (LRT_V) following the global test had a bit inflated FWER in Scenario 2. This is because the alternative hypothesis of the global test H_{1MV} is true for the first SNP and the LRT_{MV} is no longer protecting the Type I error of the nested variance test. Nevertheless, the Bonferroni correction for the variance test appeared to work well in both scenarios.

	Tests	FWER
Scenario 1:	LRT_{MV}	0.046
100 null SNPs	LRT_V	0.038
Scenario 2:	LRT_{MV}	0.051
1 mean SNP + 99 null SNPs	LRT_V	0.059

Supplemental Table 5: Empirical FWER: sample size = 2000 based on 2000 replications

REFERENCES

[1] Wray NR. 2005. Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Research and Human Genetics, 8(2):87-94.