Supplementary I: theoretical results

The theoretical mean of T can be decomposed as $\mu = \mathrm{E}\,(L) + \mathrm{E}\,(M) + \mathrm{E}\,(R)$. Because M is an indicator random variable, $\mathrm{E}\,(M) = \Pr(M = 1)$ and $\mathrm{Var}(M) = \Pr(M)[1 - \Pr(M)]$. If we let X_k be the unordered genotype of SNP k, then it is natural to calculate $\mathrm{E}\,(L)$ as $\mathrm{E}\,[\mathrm{E}\,(L \mid X_0)]$. Since X_0 takes only three possible values, the outer expectation in $\mathrm{E}\,[\mathrm{E}\,(L \mid X_0)]$ is trivial to compute. The case $X_0 = 1/2$ is easiest of all because $L = 0$ when $X_0 = 1/2$ and $M = 0$. Similar comments apply to $\mathrm{E}\,(R)$. The most natural route to calculating the variance σ^2 follows the formula

$$\mathrm{Var}(T) \;=\; \mathrm{Var}(L) + \mathrm{Var}(M) + \mathrm{Var}(R) + 2\,\mathrm{Cov}(L, M) + 2\,\mathrm{Cov}(L, R) + 2\,\mathrm{Cov}(M, R).$$

Again it is productive to condition on X_0. For instance,

$$\begin{aligned} \mathrm{Var}(L) &= \mathrm{Var}[\mathrm{E}\,(L \mid X_0)] + \mathrm{E}\,[\mathrm{Var}(L \mid X_0)], \\ \mathrm{Var}(R) &= \mathrm{Var}[\mathrm{E}\,(R \mid X_0)] + \mathrm{E}\,[\mathrm{Var}(R \mid X_0)], \end{aligned}$$

and, assuming L and R are independent given X_0,

$$\mathrm{Cov}(L, R) \;=\; \mathrm{Cov}[\mathrm{E}\,(L \mid X_0), \mathrm{E}\,(R \mid X_0)] + \mathrm{E}\,[\mathrm{Cov}(L, R \mid X_0)] \;=\; \mathrm{Cov}[\mathrm{E}\,(L \mid X_0), \mathrm{E}\,(R \mid X_0)].$$

It is also worth pointing out that $\mathrm{E}\,(LM) = \mathrm{E}(L)$ and $\mathrm{E}\,(RM) = \mathrm{E}(R)$, since L and R equal 0 when M does, and when $M = 1$, LM equals L and RM equals R. Thus, one has

$$\begin{aligned} \mathrm{Cov}(L, M) &= \mathrm{E}\,(LM) - \mathrm{E}\,(L)\,\mathrm{E}\,(M) = \mathrm{E}\,(L)[1 - \mathrm{E}\,(M)], \\ \mathrm{Cov}(R, M) &= \mathrm{E}\,(RM) - \mathrm{E}\,(R)\,\mathrm{E}\,(M) = \mathrm{E}\,(R)[1 - \mathrm{E}\,(M)]. \end{aligned}$$

These considerations emphasize the importance of finding the distributions of L and R conditional on $X_0 = 1/1$ and $X_0 = 2/2$. The next few sections tackle this issue.

To compute the conditional means and variances of L and R numerically, we recommend the right-tail sums

$$\mathrm{E}\,(Y) \;=\; \sum_{j=1}^{\infty} \Pr(Y \geq j), \qquad \mathrm{E}\,(Y^2) \;=\; \sum_{j=1}^{\infty} (2j - 1) \Pr(Y \geq j), \tag{1}$$

valid for any nonnegative random variable Y with integer values; see Section XI.1 of (Feller, 1968). The sums defining $\mathrm{E}\,(Y)$ and $\mathrm{E}\,(Y^2)$ can be truncated as soon as they stabilize.

If we define $h_{i_0,\ldots,i_r}$ to be the population frequency of the haplotype $(i_0,\ldots,i_r)$ extending from SNP 0 to SNP r, then the formula

$$\Pr(R \geq r \mid X_0 = i_0/i_0) \;=\; \frac{1}{p_{0,i_0}^2} \sum_{i_1=1}^{2} \cdots \sum_{i_r=2}^{2} h_{i_0,\ldots,i_r}^2$$

delivers the required right-tail probabilities. Here p_{0,i_0} is the frequency of allele i_0 of the core SNP 0. When all conceivable haplotypes are possible, there are 2^r terms in the multiple sum, and the formula as it stands is cumbersome. On the other hand, if only a few haplotypes are possible, then the sum is straightforward to evaluate. The moment formulas (1) are still applicable. The haplotype frequencies $h_{i_0,\ldots,i_r}$ can be estimated from haplotype data.