APPENDIX A: Proof of the unbiasness of the r_S^2 estimate for a two population structured sample when loci are unlinked.

Let S denote the random Bernoulli variable which equals 1 if the sampled individual comes from the first population and 0 otherwise. Let t denote the probability of $S = 1$ then,

$$Esp(S) = t \qquad \text{(A1)}$$

$$Var(S) = Esp(S) - (Esp(S))^2 = t(1-t) \qquad \text{(A2)}$$

Let X^l denote the dummy random variable equaling 1 when an individual carries the A allele at locus l and 0 otherwise, then

$$\begin{aligned} Esp(X^l) &= Esp(S=0)Esp(X^l \mid S=0) + Esp(S=1)Esp(X^l \mid S=1) \\ &= (1-t)Esp(X^l \mid S=0) + tEsp(X^l \mid S=1) \end{aligned} \qquad \text{(A3)}$$

Using equations (A1) and (A3), we get

$$\begin{aligned} Cov(X^l, S) &= Esp(X^l S) - Esp(X^l)Esp(S) \\ &= Esp(S=1)Esp(X^l \mid S=1) - t(1-t)Esp(X^l \mid S=0) - t^2 Esp(X^l \mid S=1) \\ &= t(1-t)(Esp(X^l \mid S=0) - Esp(X^l \mid S=1)) \end{aligned} \qquad \text{(A4)}$$

Using equations (A2) and (A4) we obtain

$$\begin{aligned} Cov(X^l, S)Var(S)Cov(S, X^m) = t(1-t)(Esp(X^l \mid S=0) - Esp(X^l \mid S=1)) \\ (Esp(X^m \mid S=0) - Esp(X^m \mid S=1)) \end{aligned} \qquad \text{(A5)}$$

On an other hand, theoretical developments of $Cov(X^l, X^m)$ give

$$\begin{aligned} Cov(X^l, X^m) &= Esp(X^l X^m) - Esp(X^l)Esp(X^m) \\ &= (1-t)Esp(X^l X^m \mid S=0) + tEsp(X^l X^m \mid S=1) - Esp(X^l)Esp(X^m) \\ &= (1-t)Cov(X^l, X^m \mid S=0) + tCov(X^l, X^m \mid S=1) - Esp(X^l)Esp(X^m) \\ &\quad + (1-t)Esp(X^l \mid S=0)Esp(X^m \mid S=0) + tEsp(X^l \mid S=1)Esp(X^m \mid S=1) \\ &= (1-t)Cov(X^l, X^m \mid S=0) + tCov(X^l, X^m \mid S=1) \\ &\quad + t(1-t)(Esp(X^l \mid S=0) - Esp(X^l \mid S=1))(Esp(X^m \mid S=0) - Esp(X^m \mid S=1)) \end{aligned} \qquad \text{(A6)}$$

Thus, using equations (A5) and (A6) we obtain that

$$\begin{aligned} Cov(X^l \mid S, X^m \mid S) &= Cov(X^l, X^m) - Cov(X^l, S)Var(S)Cov(S, X^m) \\ &= (1-t)Cov(X^l, X^m \mid S = 0) + tCov(X^l, X^m \mid S = 1) \end{aligned}$$

which equals 0 when the loci are unlinked.

APPENDIX B: Proof that the r_S^2 measure is the proportional factor to apply to sample size in order to achieve the same power of structure corrected association test at a SNP locus in linkage disequilibrium with the causal locus, as at the causal locus itself.

Let a trait, observed on a sample of size N, be explained by a causal locus l in the following linear model

$$Y = 1\!\!1_N \mu + \mathrm{S}\beta + \mathrm{X}^l \theta^l + \varepsilon$$

where $Y = (y_1, \ldots, y_i, \ldots, y_N)^T$ is the vector of observed trait values, ε is the residual vector that is assumed to have expectation 0 and variance σ^2, and (μ, β, θ^l) are the parameters for the mean, the structure effect and the causal locus effect, respectively.

The association t-test is equal to

$$t^l = \frac{\hat{\theta}^l}{\sqrt{\mathrm{V}ar(\hat{\theta}^l)}}$$

with

$$\mathrm{V}ar(\hat{\theta}^l) = \hat{\sigma}^2 ([\tilde{\mathrm{S}}, \tilde{\mathrm{X}}^l]^T [\tilde{\mathrm{S}}, \tilde{\mathrm{X}}^l])^{-1}_{2,2}$$

where $\hat{\sigma}^2$ is the estimate of σ^2, $\tilde{\mathrm{S}}$ and $\tilde{\mathrm{X}}^l$ are the centered S and X^l matrices, respectively, and $_{2,2}$ denotes the second diagonal block of the matrix.

By definition of the sample variance-covariance matrix, we get

$$[\tilde{\mathrm{S}}, \tilde{\mathrm{X}}^l]^T [\tilde{\mathrm{S}}, \tilde{\mathrm{X}}^l] = (N-1) \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,X^l} \\ \Sigma_{X^l,S} & \Sigma_{X^l,X^l} \end{pmatrix}$$

The inversion of the block matrix gives

$$\begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,X^l} \\ \Sigma_{X^l,S} & \Sigma_{X^l,X^l} \end{pmatrix}^{-1}_{2,2} = \frac{1}{(N-1)(\Sigma_{X^l,X^l} - \Sigma_{X^l,S} \Sigma^{-1}_{S,S} \Sigma_{S,X^l})}$$

So, asymptotically we get that t^l is Gaussian with variance 1 and expectation equal to $\sqrt{(N-1)(\Sigma_{X^l,X^l} - \Sigma_{X^l,S}\Sigma_{S,S}^{1}\Sigma_{S,X^l})}\ \theta^l/\sigma$.

The t-test at the SNP locus m is $t^m = \dfrac{\hat{\theta}^m}{\sqrt{Var(\hat{\theta}^m)}}$ with $Var(\hat{\theta}^m) = \hat{\sigma}^2([\tilde{S}, \tilde{X}^m]^T[\tilde{S}, \tilde{X}^m])^{-1}_{2,2}$.

To find the expectation of t^m under the causal model, which is the correct one for the expectation of the data Y , it is necessary to calculate the second block of

$$([\tilde{S}, \tilde{X}^m]^T)^{-1}[\tilde{S}, \tilde{X}^m][\tilde{S}, \tilde{X}^m]^T Esp(Y) = \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,X^m} \\ \Sigma_{X^m,S} & \Sigma_{X^m,X^m} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,X^l} \\ \Sigma_{X^m,S} & \Sigma_{X^m,X^l} \end{pmatrix} \begin{pmatrix} s \\ \theta^l \end{pmatrix}$$

This second block is equal to $(\Sigma_{X^m,X^m} - \Sigma_{X^m,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^m})^{-1}(\Sigma_{X^m,X^l} - \Sigma_{X^m,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^l})\theta^l$

We find, asymptotically,

$$Esp(t^m) \simeq \sqrt{N-1}\,\frac{\Sigma_{X^m,X^l} - \Sigma_{X^m,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^l}}{\sqrt{\Sigma_{X^m,X^m} - \Sigma_{X^m,S}\Sigma_{S,S}^{-1}\Sigma_{S,X^m}}}\,\theta^l/\sigma = \sqrt{\hat{r}_S^2(l,m)}Esp(t^l)$$

Thus, t^m is asymptotically Gaussian with variance 1 and expectation equal to $\sqrt{\hat{r}_S^2(l,m)}Esp(t^l)$. This finishes the first part of the proof showing that $\hat{r}_S^2(l,m)$ is the reducing power factor between the causal locus and a SNP locus in linkage disequilibrium.

Now, let us suppose that we get a N^m sample at the SNP locus and a N^l sample at the causal locus. Using that asymptotically $\hat{r}_S^2(l,m) \simeq r_S^2(l,m)$ and that $\Sigma_{X^l,X^l} - \Sigma_{X^l,S}\Sigma_{S,S}^{1}\Sigma_{S,X^l} \simeq Var(X^l \mid S)$, we get

$$Esp(t^m) \simeq \sqrt{r_S^2(l,m)}\sqrt{(N^m-1)Var(X^l \mid S)}\theta^l/\sigma = \sqrt{r_S^2(l,m)}\sqrt{\frac{N^m-1}{N^l-1}}Esp(t^l)$$

Then, we obtain that the sample size has to be increased by a factor equal to $1/r_S^2(l,m)$ to achieve the same power at the SNP locus, compared to the power at the causal locus.