Methods for d-ldsc: Dominance LD-score regression

Duncan Palmer[1,2,3*], Wei Zhou[1,2,3], Liam Abbott[1],
Nik Baya[1], Claire Churchhouse[1,2,3], Cotton Seed[1,3], Tim Poterba[1,3],
Daniel King[1,3], Masahiro Kanai[1,2,3], Alex Bloemendal[1] and Benjamin Neale[1,2,3]

[1] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.,
[2]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.,
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[*] Correspondence to: duncan.stuart.palmer@gmail.com.

Methods: LD-score dominance extension summary

In order to estimate the contribution of within locus non-additive effects across the genome to phenotypic variation, we extend the infinitesimal model underlying LD-score regression to incorporate a non-additive effect at each site which is uncorrelated from the additive contribution at that site. We then ask: what is the variance of this additional contribution? Letting X^A denote the n samples $\times$ m sites matrix of genotypes (after rescaling to enforce a mean of 0 and a variance of 1), we define X^D to be a re-coding of the genotypes whose columns (X_j^D) are orthonormal to the columns of X^A (X_j^A) under Hardy-Weinberg equilibrium. This extension to the additive infinitesimal model is:

$$y_i = \sum_{j=1}^{m} X_{i,j}^A \beta_{A_j} + \sum_{j=1}^{m} X_{i,j}^D \beta_{D_j} + \varepsilon_i, \tag{1}$$

where y_i is a continuous phenotype measured in individual i, β_{A_j} and β_{A_j} are the additive and dominance effect sizes at site j respectively, and ε_i is a noise term. We mean center the genotypes and construct an orthonormal basis using the Gram-Schmidt procedure. The resultant additive and dominance encodings are

$$\frac{1}{\sqrt{2pq}} \begin{bmatrix} -2p \\ 1-2p \\ 2-2p \end{bmatrix} \text{ and } \begin{bmatrix} -p/q \\ 1 \\ -q/p \end{bmatrix}, \tag{2}$$

respectively. We wish to estimate $h_D^2 := \text{Var}\left(\sum j = 1^m Xi, j^D \beta_{D_j}\right)$. By defining additive summary statistics as the marginal effect sizes obtained by regressing y on X_j^A as usual; $\widehat{\beta}_{A_j} = \frac{1}{n}\left(X_j^A\right)^\top y$, and introducing dominance summary statistics as the analogous marginal associations in the dominance encoding; $\widehat{\beta}_{D_j} = \frac{1}{n}\left(X_j^D\right)^\top y$, we may then proceed to derive a dominance LD-score equation relating dominance summary statistics to dominance LD-scores:

$$\mathbb{E}\left[\chi_{D_j}^2\right] = \frac{n h_D^2}{m} l_j^D + 1 \tag{3}$$

where $l_j^D = \sum_{j=1}^m (r_{j,k}^D)^2$ is the sum of the squared correlations between SNP j and all other SNPs under the dominance encoding, and $\chi_{D_j}^2$ the chi-squared statistic for SNP j under the dominance encoding, $\chi_{D_j}^2 = n\beta_{D_j}^2$.

Detailed derivations of additive LD-score regression and dominance LD-score regression

Additive LD score regression

We first consider the standard LD score regression in which we assume the infinitesimal model: the effect sizes of the genotypes $\beta_j;\ j = 1, 2, \ldots, m$ are independent with mean 0 and variance $\frac{h^2}{m}$. This is formulated as

$$y = X\beta + \varepsilon. \tag{4}$$

That is,

$$y_i = \sum_{j=1}^{m} X_{i,j}\beta_j + \varepsilon_i, \tag{5}$$

where y is a vector of standardised phenotypes, and X is an $n \times m$ matrix of n standardised genotypes both taken from m independent samples (unscaled genotypes $X'_{i,j} \in \{0, 1, 2\}$, subject to rescaling: $X_{i,j} = \left(X'_{i,j} - 2p_j\right) / \sqrt{2p_j(1-p_j)}$, where p_j is the prevalence of the genotype at site j, in the population). Finally we assume that there is are independent uncorrelated noise terms which absorb the effect of the environment.

$$\varepsilon \sim \mathcal{N}(0, 1 - h^2), \tag{6}$$

Note that $\mathrm{Var}(\varepsilon) = 1 - h^2$ since $\mathrm{Var}\left(X\beta\right) = h^2$ by definition of the narrow sense heritability $\left(\mathrm{Var}\left(\sum_{j=1}^{m} X_{i,j}\beta_j\right) = \mathrm{Var}\left(\beta_j\right)\mathrm{Var}\left(\sum_{j=1}^{m} X_{i,j}\right) = \frac{h^2}{m}\sum_{j=1}^{m}\mathrm{Var}\left(X_{i,j}\right) = h^2\right)$ and $\mathrm{Var}(y) = 1$ by construction. Under this model, all of the genetic heritability is additive.

We estimate effect sizes

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}y_i = \frac{1}{n}X_j^\top y \tag{7}$$

where X_j here are the rows of X. This is the marginal effect of SNP j in the sample, and is simply the estimate of the effect size of SNP j using a linear regression of genotype X_j against the phenotype y $\left(\widehat{\beta}_j = \frac{\sum_{i=1}^{n}\left(X_{i,j} - \overline{X_j}\right)(y_i - \overline{y})}{\sum_{i=1}^{n} X_{i,j} - \overline{X_j}} = \frac{\sum_{i=1}^{n} X_{i,j}y_i}{n}\right)$ Let us define χ_1^2 coefficients for each j as $\chi_j^2 = n\widehat{\beta}_j^2$. Note that $n\widehat{\beta}_j^2$ is χ_1^2 distributed under the null due to the central limit theorem:

Recall that if X_i are IID random variables with mean μ and variance σ^2, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right) \to \mathcal{N}(0, \sigma^2) \tag{8}$$

in distribution. We have $\mathbb{E}[X_{i,j}y_i] = 0$, $\mathrm{Var}(X_{i,j}y_i) = 1$. Thus $\sqrt{n}\left(\sum_{i=1}^{m} X_{i,j}y_i\right) \to \mathcal{N}(0,1)$, so $\widehat{\beta_j} \sim \frac{1}{\sqrt{n}}\mathcal{N}(0,1)$. Now recall that if $Y = \sum_{i=1}^{j} X_i^2$ where $X_i \sim \mathcal{N}(0,1)$, then $Y \sim \chi_j^2$. Thus, $n\beta_j^2 \sim \chi_1^2$ (a χ^2 distribution with 1 degree of freedom).

Given our effect size estimate and our model in Equation (5), we can substitute into Equation (7):

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\left(\sum_{k=1}^{m} X_{i,k}\beta_k + \varepsilon_i\right) \tag{9}$$

$$= \sum_{k=1}^{m} \beta_k \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}X_{i,k}\right) + \widetilde{\varepsilon}_j \tag{10}$$

$$= \sum_{k=1}^{m} \widehat{r}_{j,k}\beta_k + \widetilde{\varepsilon}_j \tag{11}$$

where $\widetilde{\varepsilon}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\varepsilon_i$. We now determine the expectation of our χ^2 coefficients to obtain the key formula.

$$\mathbb{E}[\chi_j^2] = \mathbb{E}\left[n\widehat{\beta}_j^2\right] \tag{12}$$

$$= \mathbb{E}\left[n\left(\sum_{k=1}^{m} \widehat{r}_{j,k}\beta_k + \widetilde{\varepsilon}_j\right)^2\right] \tag{13}$$

$$= \mathbb{E}\left[n\sum_{k=1}^{m}\sum_{l=1}^{m} \widehat{r}_{j,k}\widehat{r}_{j,l}\beta_k\beta_l\right] + \underbrace{\mathbb{E}\left[2n\sum_{k=1}^{m} \widehat{r}_{j,k}\widetilde{\varepsilon}_j\right]}_{=0} + \mathbb{E}\left[n\widetilde{\varepsilon}_j^2\right] \tag{14}$$

$$= \underbrace{n\mathbb{E}\left[\sum_{k\neq l} \widehat{r}_{j,k}\widehat{r}_{j,l}\beta_k\beta_l\right]}_{=0} + \mathbb{E}\left[n\sum_{k=1}^{m} \widehat{r}_{j,k}^2\beta_k^2\right] + n\mathbb{E}\left[\widetilde{\varepsilon}_j^2\right] \tag{15}$$

$$= n\left(\sum_{k=1}^{m} \mathbb{E}\left[\widehat{r}_{j,k}^2\right]\mathbb{E}\left[\beta_k^2\right]\right) + n\mathbb{E}\left[\widetilde{\varepsilon}_j^2\right] \tag{16}$$

by independence and linearity of expectations. Since β_k is standardised, $\mathrm{Var}\left(\beta_k\right) = \mathbb{E}[\beta_k^2] = \frac{h^2}{m}$. Also, in an unstructured sample, we can make the approximation $\mathbb{E}\left[\widehat{r}_{j,k}^2\right] \approx r_{j,k}^2 + \frac{1}{n}$. This approximation comes from noting that $\widehat{r}$ is unbiased as we know the population variances for $X_j;\ j \in \{1, 2, \ldots, m\}$ are 1. Therefore, setting $X = X_j$, $Y = X_k$ to make notation simpler

and using subscripts i and l to reiterate that summation is over samples,

$$\mathbb{E}\left[\widehat{r}^2_{X,Y}\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)^2\right] \tag{17}$$

$$= \frac{1}{n^2}\mathbb{E}\left[\sum_{l=1}^{n}\sum_{i=1}^{n} X_i Y_i X_l Y_l\right] \tag{18}$$

$$= \frac{1}{n^2}\sum_{i,l:\ i\neq l}^{n} \mathbb{E}\left[X_i Y_i\right]\mathbb{E}\left[X_l Y_l\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 Y_i^2\right] \tag{19}$$

$$= \frac{1}{n^2}\left(n^2 - n\right) r^2_{X,Y} + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 Y_i^2\right]. \tag{20}$$

since $\mathbb{E}\left[X_i Y_i\right] = r_{X,Y}$ and the genotypes of individuals i and l are independent in an unstructured population. Now, the majority of X and Y will be far apart and so approximately independent. Therefore,

$$\mathbb{E}\left[\widehat{r}^2_{X,Y}\right] \approx \left(1 - \frac{1}{n}\right) r^2_{X,Y} + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]\mathbb{E}\left[Y_i^2\right] = \left(1 - \frac{1}{n}\right) r^2_{X,Y} + \frac{1}{n} \tag{21}$$

$$\approx r^2_{X,Y} + \frac{1}{n}. \tag{22}$$

Substituting this approximation into Equation (16), we have

$$\mathbb{E}\left[\chi_j^2\right] \approx \frac{nh^2}{m}\sum_{k=1}^{m}\left(r^2_{j,k} + \frac{1}{n}\right) + \mathbb{E}\left[\widetilde{\varepsilon}_j^2\right]. \tag{23}$$

Since $X_{i,j}$ and ε_i are independent, the $X_{i,j}$ are standardised, and $\mathbb{E}[\widetilde{\varepsilon}_j] = \mathbb{E}[\varepsilon_i] = 0$,

$$\mathbb{E}\left[\widetilde{\varepsilon}_j^2\right] = \mathrm{Var}\left(\widetilde{\varepsilon}_j\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\varepsilon_i\right) \tag{24}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_{i,j}\right)\mathrm{Var}\left(\varepsilon_i\right) \tag{25}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\varepsilon_i\right) \tag{26}$$

$$= \frac{1}{n}\left(1 - h^2\right). \tag{27}$$

Thus,

$$\mathbb{E}\left[\chi_j^2\right] \approx \frac{nh^2}{m}\sum_{k=1}^{m} r^2_{j,k} + h^2 + 1 - h^2. \tag{28}$$

Finally, letting $l_j := \sum_{k=1}^{m} r_{j,k}^2$; dubbed the LD score of SNP j, we obtain the LD score regression formula:

$$\mathbb{E}\left[\chi_j^2\right] = \frac{nh^2}{m} l_j + 1. \tag{29}$$

LD score regression with dominance

For a more general model, we now allow a dominance term to affect the phenotype y.

$$y = X\beta + \varepsilon \tag{30}$$

$$= X^A \beta_A + X^D \beta_D + \varepsilon \tag{31}$$

$$y_i = \sum_{j=1}^{m} X_{i,j}^A \beta_{Aj} \sum_{j=1}^{m} X_{i,j}^D \beta_{Dj} + \varepsilon_i \tag{32}$$

where now X is the concatenation of two matrices; X^A and X^D (similarly, $\beta = (\beta_A, \beta_D)$). X^A encodes the genotype (in exactly the same was as X does in Equation (5)), and X^D encodes a dominance contribution to the phenotype y. We will pick an zero mean-centered orthonormal encoding for X^A and X^D. This encoding isolates the pure dominance and additive contributions of the genotypes. We start with the basis:

$$X^{0\prime} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \; X^{A\prime} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \; X^{D\prime} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \tag{33}$$

where the first, second and third entries are the contributions of homozygous reference, heterozygous, and homozygous variant respectively. We then apply the Gram-Schmidt process (a method to orthonormalise a set of vectors in an inner product space), where our inner product is defined as

$$\langle X, Y \rangle = q^2 X_0 Y_0 + 2pq X_1 Y_1 + p^2 X_2 Y_2. \tag{34}$$

The subscript here denotes the genotype of the entry: (hom ref, het, hom var) for (0,1,2). We use this inner product as we wish to weight the contributions of the genotypes according to their prevalence in the population, assuming Hardy-Weinberg equilibrium. The inclusion of $[1,1,1]^\top$ will ensure that our resultant orthonormal basis has mean 0. Proceeding with Gram-Schmidt:

$$X^0 = \frac{1}{\left\| X^{0\prime} \right\|} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{35}$$

$$X^A = \frac{X^{A\prime} - \langle X^{A\prime}, X^0\rangle X^0}{\left\|X^{A\prime} - \langle X^{A\prime}, X^0\rangle X^0\right\|} = \frac{1}{\sqrt{2pq}}\left(\begin{bmatrix}0\\1\\2\end{bmatrix} - \left(2pq + 2q^2\right)\begin{bmatrix}1\\1\\1\end{bmatrix}\right) \tag{36}$$

$$= \frac{1}{\sqrt{2pq}}\begin{bmatrix}-2p\\1-2p\\2-2p\end{bmatrix} \tag{37}$$

$$= \frac{1}{\sqrt{2pq}}\begin{bmatrix}-2p\\q-p\\2q\end{bmatrix}, \tag{38}$$

$$X^D = \frac{X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A}{\left\|X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A\right\|} = \frac{\begin{bmatrix}0\\1\\0\end{bmatrix} - 2pq\begin{bmatrix}1\\1\\1\end{bmatrix} - (q-p)\begin{bmatrix}-2p\\q-p\\2q\end{bmatrix}}{\left\|X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A\right\|} \tag{39}$$

$$= \frac{\begin{bmatrix}-2p^2\\2pq\\-2q^2\end{bmatrix}}{\left\|X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A\right\|} \tag{40}$$

$$= \frac{1}{\sqrt{q^2p^4 + 2p^3q^3 + p^2q^4}}\begin{bmatrix}-p^2\\pq\\-q^2\end{bmatrix} \tag{41}$$

$$= \frac{1}{pq}\begin{bmatrix}-p^2\\pq\\-q^2\end{bmatrix}. \tag{42}$$

Each site has additive and dominance effect sizes (β_A and β_D) which act along X^A and X^D respectively. β_A and β_D are drawn from distributions with mean 0 and variances $\frac{h^2_A}{m}$, $\frac{h^2_D}{m}$ respectively. We assume $\beta_j = (\beta_{A_j}, \beta_{D_j})$ are independent across sites, but may have dependencies at a given locus. Let's determine the variance of $X_i\beta$ ($= X_i^A\beta_A + X_i^D\beta_D$), where here, X_i denotes

the i^{th} row of the matrix X.

$$\text{Var}\left(X_i\beta\right) = \text{Var}\left(\sum_{j=1}^{m} X_{i,j}\beta_j\right) \tag{43}$$

$$= \mathbb{E}\left[\sum_{j=1}^{m}\sum_{l=1}^{m} X_{i,j}X_{i,l}\beta_j\beta_l\right] \tag{44}$$

$$= \mathbb{E}\left[\sum_{j=1}^{m}\sum_{l=1}^{m} X^A_{i,j}X^A_{i,l}\beta_{Aj}\beta_{Al} + X^A_{i,j}X^D_{i,l}\beta_{Aj}\beta_{Dl} + X^D_{i,j}X^A_{i,l}\beta_{Dj}\beta_{Al} + X^D_{i,j}X^D_{i,l}\beta_{Dj}\beta_{Dl}\right]. \tag{45}$$

We may remove all cross terms over sites as we assume that β_{Aj}, β_{Dj} are independent across sites, and also independent of the genetic data encoded in X^A and X^D. Thus,

$$\mathbb{E}\left[X_i\beta\right] = \mathbb{E}\left[\sum_{j=1}^{m} {X^A_{i,j}}^2\beta^2_{Aj} + X^A_{i,j}X^D_{i,j}\beta_{Aj}\beta_{Dj} + X^D_{i,j}X^A_{i,j}\beta_{Dj}\beta_{Aj} + {X^D_{i,j}}^2\beta^2_{Dj}\right]. \tag{46}$$

Now we may use that $\mathbb{E}\left[X^A_{i,j}X^D_{i,j}\right] = \mathbb{E}\left[X^A_{i,j}\right]\mathbb{E}\left[X^D_{i,j}\right] = 0$, as X^A and X^D are orthogonal by construction. Thus, we are left with

$$\text{Var}\left(X_i\beta\right) = \mathbb{E}\left[\sum_{j=1}^{m} {X^A_{i,j}}^2\beta^2_{Aj}\right] + \mathbb{E}\left[\sum_{j=1}^{m} {X^D_{i,j}}^2\beta^2_{Dj}\right] \tag{47}$$

$$= \sum_{j=1}^{m}\mathbb{E}\left[{X^A_{i,j}}^2\right]\mathbb{E}\left[\beta^2_{Aj}\right] + \sum_{j=1}^{m}\mathbb{E}\left[{X^D_{i,j}}^2\right]\mathbb{E}\left[\beta^2_{Dj}\right] \tag{48}$$

$$= \frac{h^2_A}{m}\sum_{j=1}^{m} 1 + \frac{h^2_D}{m}\sum_{j=1}^{m} 1 = h^2_A + h^2_D. \tag{49}$$

So $\text{Var}\left(X\beta\right) = h^2_A + h^2_D = h^2$. As in the standard infinitesimal model, the noise term ε has mean 0 and variance $1 - h^2$. $h^2_A \leq h^2_A + h^2_D = h^2$. In particular, equalities hold when the dominance terms don't contribute to the variance.

We wish to determine the extent to which $h^2 > h^2_A$. To achieve this we can simply compare the sums of the first order terms of two LD score regression models. The first under the standard infinitesimal model as outlined above, and the second according the model introduced in Equation (30). It remains to derive the analogue of Equation (29), which can be obtained in a similar manner to the original LD score regression model.

Again, we estimate effect sizes but this time regress our dominance encoded SNPs against the phenotypes as well:

$$\widehat{\beta}_{A_j} = \frac{1}{n}\sum_{i=1}^{n} X^A_{i,j}y_i = \frac{1}{n}{X^A_j}^{\top}y; \qquad \widehat{\beta}_{D_j} = \frac{1}{n}\sum_{i=1}^{n} X^D_{i,j}y_i = \frac{1}{n}{X^D_j}^{\top}y. \tag{50}$$

This time, we substitute y_i using Equation (30):

$$\widehat{\beta}_{A_j} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}\left(\sum_{k=1}^{m} X_{i,k}^{A}\beta_{A_k} + \sum_{k=1}^{m} X_{i,k}^{D}\beta_{D_k} + \varepsilon_i\right) \tag{51}$$

$$= \sum_{k=1}^{m}\beta_{A_k}\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}X_{i,k}^{A} + \sum_{k=1}^{m}\beta_{D_k}\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}X_{i,k}^{D} + \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}\varepsilon_i \tag{52}$$

$$= \sum_{k=1}^{m}\widehat{r}_{j,k}^{AA}\beta_{A_k} + \sum_{k=1}^{m}\widehat{r}_{j,k}^{AD}\beta_{D_k} + \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}\varepsilon_i. \tag{53}$$

Similarly,

$$\widehat{\beta}_{D_j} = \sum_{k=1}^{m}\widehat{r}_{j,k}^{DA}\beta_{A_k} + \sum_{k=1}^{m}\widehat{r}_{j,k}^{DD}\beta_{D_k} + \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{D}\varepsilon_k. \tag{54}$$

Setting $\widetilde{\varepsilon}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}^{A}\varepsilon_i$, we now determine the expectation of the χ^2 statistic,

$$\chi_{A_j}^2 := n\widehat{\beta}_{A_j}. \tag{55}$$

(Obtaining the expectation of $\chi_{D_j}^2 := n\widehat{\beta}_{D_j}$ is analogous).

$$\mathbb{E}\left[\chi_{A_j}^2\right] = \mathbb{E}\left[n\beta_{A_j}^2\right] = n\mathbb{E}\left[\left(\sum_{k=1}^{m}\widehat{r}_{j,k}^{AA}\beta_{A_k} + \sum_{k=1}^{m}\widehat{r}_{j,k}^{AD}\beta_{D_k} + \widetilde{\varepsilon}_j\right)^2\right] \tag{56}$$

$$= n\mathbb{E}\left[\sum_{k=1}^{m}\sum_{l=1}^{m}\left(\widehat{r}_{j,k}^{AA}\beta_{A_k}\widehat{r}_{l,k}^{AA}\beta_{A_l} + \widehat{r}_{j,k}^{AA}\beta_{A_k}\widehat{r}_{l,k}^{AD}\beta_{D_l} + \right.\right. \tag{57}$$

$$\left.\left.\widehat{r}_{j,k}^{AD}\beta_{D_k}\widehat{r}_{l,k}^{AA}\beta_{A_l} + \widehat{r}_{j,k}^{AD}\beta_{D_k}\widehat{r}_{l,k}^{AD}\beta_{D_l}\right) + \underbrace{\widetilde{\varepsilon}_j\left(\sum_{k=1}^{m}\widehat{r}_{j,k}^{AA}\beta_{A_k} + \widehat{r}_{j,k}^{AD}\beta_{D_k}\right)}_{=0} + \widetilde{\varepsilon}_j^2\right]$$

Again, we may remove all cross terms over sites due to the independence of the βs.

$$= n\mathbb{E}\left[\sum_{k=1}^{m}\widehat{r}_{j,k}^{AA^2}\beta_{A_k}^2\right] + 2n\mathbb{E}\left[\sum_{k=1}^{m}\widehat{r}_{j,k}^{AA}\beta_{A_k}\widehat{r}_{j,k}^{AD}\beta_{D_k}\right] + \tag{58}$$

$$n\mathbb{E}\left[\sum_{k=1}^{m}\widehat{r}_{j,k}^{AD^2}\beta_{D_k}^2\right] + n\mathbb{E}\left[\widetilde{\varepsilon}_j^2\right]$$

We may replace the $\widehat{r}_{j,k}^{AA^2}$ and $\widehat{r}_{j,k}^{AD^2}$ as before, but we also have a collection of cross terms (i.e. $\widehat{r}_{j,k}^{AA}\beta_{A_k}\widehat{r}_{j,k}^{AD}\beta_{D_k}$) to consider. Let's determine the expectation of this summation of cross terms.

$$\mathbb{E}\left[\sum_{k=1}^{m}\beta_{A_k}\beta_{D_k}\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right] = \sum_{k=1}^{m}\mathbb{E}\left[\beta_{A_k}\beta_{D_k}\right]\mathbb{E}\left[\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right]. \tag{59}$$

There may be dependency between the additive a dominance effect sizes a locus. So let's turn our attention to $\mathbb{E}\left[\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right]$:

$$\mathbb{E}\left[\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}X_{i,j}^{A}X_{i,k}^{A}X_{l,j}^{A}X_{l,k}^{D}\right] \tag{60}$$

$$= \frac{1}{n^2}\sum_{i,l:\ i\neq l}\mathbb{E}\left[X_{i,j}^{A}X_{i,k}^{A}X_{l,j}^{A}X_{l,k}^{D}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[X_{i,j}^{A^2}X_{i,k}^{A}X_{i,k}^{D}\right] \tag{61}$$

$$= \frac{1}{n^2}\left(n^2-n\right)r_{j,k}^{AA}r_{j,k}^{AD} + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[X_{i,j}^{A^2}X_{i,k}^{A}X_{i,k}^{D}\right]. \tag{62}$$

Since, j and k are far apart for most SNPs, we make the approximation

$$\mathbb{E}\left[\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right] \approx \left(1-\frac{1}{n}\right)r_{j,k}^{AA}r_{j,k}^{AD} + \sum_{i=1}^{n}\mathbb{E}\left[X_{i,j}^{A^2}\right]\mathbb{E}\left[X_{i,k}^{A}X_{i,k}^{D}\right]. \tag{63}$$

Note that $X_{i,j}^{D}$ and $X_{i,j}^{A}$ are independent for all j by construction. So,

$$\mathbb{E}\left[\widehat{r}_{j,k}^{AA}\widehat{r}_{j,k}^{AD}\right] \approx r_{j,k}^{AA}r_{j,k}^{AD}. \tag{64}$$

Plugging in this approximation, and the approximation in Equation (22) into Equation (58), we have

$$\mathbb{E}\left[\chi_{A_j}^2\right] \approx n\sum_{k=1}^{m}\underbrace{\left(r_{j,k}^{AA^2}+\frac{1}{n}\right)}_{\text{by Equation (22)}}\mathbb{E}\left[\beta_{A_k}^2\right] + n\sum_{k=1}^{m}\underbrace{\left(r_{j,k}^{AD^2}+\frac{1}{n}\right)}_{\text{by Equation (22)}}\mathbb{E}\left[\beta_{D_k}^2\right] + \tag{65}$$

$$\sum_{k=1}^{m}\mathbb{E}\left[\beta_{A_k}\beta_{D_k}\right]\underbrace{r_{j,k}^{AA}r_{j,k}^{AD}}_{\text{by Equation (64)}} + n\mathbb{E}\left[\widetilde{\varepsilon}_j^2\right]$$

$$= \frac{nh_A^2}{m}\sum_{k=1}^{m}r_{j,k}^{AA^2} + \frac{nh_D^2}{m}\sum_{k=1}^{m}r_{j,k}^{AD^2} + h_A^2 + h_D^2 + \sum_{k=1}^{m}\mathbb{E}\left[\beta_{A_k}\beta_{D_k}\right]r_{j,k}^{AA}r_{j,k}^{AD} + \frac{n(1-h^2)}{n} \tag{66}$$

$$= \frac{nh_A^2}{m}\sum_{k=1}^{m}r_{j,k}^{AA^2} + \frac{nh_D^2}{m}\sum_{k=1}^{m}r_{j,k}^{AD^2} + \sum_{k=1}^{m}\mathbb{E}\left[\beta_{A_k}\beta_{D_k}\right]r_{j,k}^{AA}r_{j,k}^{AD} + 1. \tag{67}$$

j/k	0	1
0	$1-p_j-p_k$	$p_j-p_{j,k}$
1	$p_k-p_{j,k}$	$p_{j,k}$

Table S1

Similarly,

$$\mathbb{E}\left[\chi^2_{D_j}\right] = \frac{nh^2_A}{m}\sum_{k=1}^{m} r^{DA^2}_{j,k} + \frac{nh^2_D}{m}\sum_{k=1}^{m} r^{DD^2}_{j,k} + \sum_{k=1}^{m}\mathbb{E}\left[\beta_{A_k}\beta_{D_k}\right] r^{DD}_{j,k} r^{AD}_{j,k} + 1. \tag{68}$$

We can investigate the contribution of each of the correlation terms analytically. Let p_j and p_k denote the probability of observing the alternate genotype at sites j and k respectively, and $p_{j,k}$ denote the probability of observing the alternate genotype at both sites j and k.

This then leads to the probability mass function of linked genotypes in offspring (in the absence of inbreeding) by multiplying the relevant entries of Table S1. We may then determine the correlation coefficients (r^{AA}, r^{AD}, r^{DA}, r^{DD}):

$$r^{QR} = \mathbb{E}\left[X_Q^\top X_R\right] = \sum_{g_j,g_k\in\{0,1,2\}} P(g_j,g_k) X^Q_j(g_j) X^R_k(g_k); \quad Q,R\in\{A,D\}. \tag{69}$$

where

$$X^A_j = \frac{1}{\sqrt{2p_jq_j}}\begin{bmatrix}-2p_j\\ q_j-p_j\\ 2q_j\end{bmatrix},\ X^D_j = \frac{1}{p_jq_j}\begin{bmatrix}-p_j^2\\ p_j-q_j\\ -q_j^2\end{bmatrix}. \tag{70}$$

Here, using an abuse of notation, we mean the encoding of the genotypes (homozygous reference, heterozygous, homozgous variant) at site j.

After cancelling, we obtain:

$$r^{AD}=0, \quad r^{DA}=0, \tag{71}$$

$$r^{AA} = \frac{p_{j,k}-p_jp_k}{\sqrt{p_jp_k\left(1-p_j\right)\left(1-p_k\right)}}, \quad r^{DD} = r^{AA^2} = \frac{\left(p_{j,k}-p_jp_k\right)^2}{p_jp_k\left(1-p_j\right)\left(1-p_k\right)}. \tag{72}$$

Equations (67) and (68) then simplify to

$$\mathbb{E}\left[\chi^2_{A_j}\right] = \frac{nh^2_A}{m} l^A_j + 1; \tag{73}$$

$$\mathbb{E}\left[\chi^2_{D_j}\right] = \frac{nh^2_D}{m} l^D_j + 1, \tag{74}$$

where $l^A_j = \sum_{j=1}^m r^{AA^2}_{j,k}$ and $l^D_j = \sum_{j=1}^m r^{DD^2}_{j,k}$.