# Appendix

## Finite locus models

We follow the notation and derivation of [4] as closely as possible. We assume that the trait is genetically the same in males and females and allow for differential gene action in females due to dosage compensation. We assume that assortative mating is solely on the phenotype so that the correlation of alleles between mates is the same within and between loci. We also assume that all loci are unlinked. We used the following notations.

| Parameter | Definition | Autosomes | X-chromosome |
|---|---|---|---|
| $n$ | haploid number of loci | $n_A$ | $n_X$ |
| $f$ | correlation in value of alleles at the same locus | $f_A$ | $f_X$ |
| $k$ | correlation of alleles at different loci in the same gamete | $k_A$ | $k_X$ |
| $l$ | correlation of alleles at different loci in different gametes | $l_A$ | $l_X$ |
| $m$ | correlation of alleles in mates | $m_A$ | $m_X$ |
| $\rho$ | phenotypic correlation between mates | | |
| $a$ | allelic effect size | $a$ | $a_M$, $a_F$ |
| $p$ | frequency of increasing allele | | |

Table 1: Summary of parameters (notations) used for the derivation, under finite locus models, of the equilibrium genetic variance reached after multiple generations of AM.

$Y_F^{(t)}$ and $Y_M^{(t)}$ are the phenotypes of female and male mates in generation $t$, respectively.

### Crow and Kimura derivation for equal effect sizes and allele frequencies: autosomal loci only (Chapter 4.7)

This derivation assumes that all effect sizes and allele frequencies are the same and that all phenotypic variance is genetic. For $n$ unlinked autosomal genes, Crow and Kimura showed that

$$\mathrm{var}(Y_F^{(t)}) = \mathrm{var}(Y_M^{(t)}) = 2np(1-p)a^2\left[1 + f_t + (n-1)(k_t + l_t)\right] \tag{34}$$

$$\mathrm{cov}(Y_F^{(t)}, Y_M^{(t)}) = 4n^2p(1-p)m_t a^2 = \rho\sqrt{\mathrm{var}(Y_F^{(t)})\mathrm{var}(Y_M^{(t)})} = \rho\mathrm{var}(Y_M^{(t)}) \tag{35}$$

For autosomal genes, the recurrence relationship from generation t to t+1 are,

$$
\begin{aligned}
f_{t+1} &= m_t, \\
l_{t+1} &= m_t, \\
k_{t+1} &= \frac{1}{2}k_t + \frac{1}{2}l_t = \frac{1}{2}k_t + \frac{1}{2}f_t
\end{aligned}
$$

Using these and eq. (34) and (35) gives,

$$f_{t+1} = \frac{\rho}{2n}\left[1 + nf_t + (n-1)k_t\right],$$

which at equilibrium yields the known relation ([19]; [4])

$$f_{eq} = \frac{\rho}{\rho + 2n(1-\rho)} \tag{36}$$

and a ratio of equilibrium variance to base population variance of,

$$\frac{\mathrm{var}(Y^{(eq)})}{\mathrm{var}(Y^{(0)})} = \frac{1}{1-\rho[1-1/(2n)]}, \text{ ([19])}. \tag{37}$$

[4] later extended their model to cover cases where causal variants have different allele frequencies ($p_j$) and different effect sizes ($a_j$). In that case, it is shown that equation (36) still holds with $n$ being replaced by $n_e$ defined below as

$$n_e = \frac{\left(\sum_{j=1}^{n}\sqrt{2p_j(1-p_j)a_j^2}\right)^2}{\sum_{j=1}^{n}2p_j(1-p_j)a_j^2}. \tag{38}$$

In all cases, when $n$ is large, the genetic variance is changed by a factor of $1/(1-\rho)$.

**X-chromosome genes**

We now consider that all loci that influence the phenotype are on the X-chromosome. We allow the effect sizes to be different between males and females. The variance in males is due to genic variances at $n_X$ loci

and $n_X(n_X-1)$ covariances between alleles on their single X-chromosome,

$$\mathrm{var}(Y_M^{(t)}) = n_X p(1-p)a_M^2[1+(n_X-1)k_{X(t)}]. \tag{39}$$

The variance in females is analogous to that shown in the first section,

$$\mathrm{var}(Y_F^{(t)}) = 2n_X p(1-p)a_F^2[1+f_{X(t)}+(n_X-1)(k_{X(t)}+l_{X(t)})]. \tag{40}$$

And the covariance

$$\mathrm{cov}(Y_F^{(t)}, Y_M^{(t)}) = 2n_X^2 p(1-p)a_M a_F m_{X(t)}. \tag{41}$$

The covariance at generation $t$ is also equal to

$$\begin{aligned}
\mathrm{cov}(Y_F^{(t)}, Y_M^{(t)}) &= \rho\sqrt{\mathrm{var}(Y_M^{(t)})\mathrm{var}(Y_F^{(t)})} \\
&= \rho\sqrt{n_X p(1-p)a_M^2[1+(n_X-1)k_{X(t)}]} \\
&\quad \times\sqrt{2n_X p(1-p)a_F^2[1+f_{X(t)}+(n_X-1)(k_{X(t)}+l_{X(t)})]} \\
&= \rho n_X p(1-p)a_M a_F\sqrt{2[1+(n_X-1)k_{X(t)}][1+f_{X(t)}+(n_X-1)(k_{X(t)}+l_{X(t)})]}
\end{aligned} \tag{42}$$

For X-chromosome genes, the recurrence relationship from generation t to t+1 are,

$$f_{X(t+1)} = m_{X(t)}, \tag{43}$$
$$l_{X(t+1)} = m_{X(t)}, \tag{44}$$
$$\tag{45}$$

For correlations between alleles in the same gamete, the contribution from males in the previous generation is only from within their single gamete whereas the contribution from females is from within and between

gametes. Therefore,

$$k_{X(t+1)} = \frac{1}{2}\left[k_{X(t)} + \frac{1}{2}\left(k_{X(t)} + l_{X(t)}\right)\right] = \frac{3}{4}k_{X(t)} + \frac{1}{4}l_{X(t)} = \frac{3}{4}k_{X(t)} + \frac{1}{4}f_{X(t)}. \tag{46}$$

It follows that

$$f_{X(t+1)} = \frac{\rho}{n_X\sqrt{2}}\sqrt{[1 + (n_X - 1)k_{X(t)}][1 + f_{X(t)} + 2(n_X - 1)f_{X(t)}]} \tag{47}$$

At equilibrium, all correlations are the same. Solving equation (47) for $f_{X(eq)}$ leads to an equilibrium correlation of

$$f_{X(eq)} = \frac{(3n_X - 2)\rho^2 + \rho n_X\sqrt{8 + \rho^2}}{4n_X^2 - 2\rho^2(2n_X - 1)(n_X - 1)} \tag{48}$$

As for the autosomes, if $\rho = 1$ then $f_{X(eq)} = 1$. The ratio of equilibrium to base population variances are, for males and females, respectively,

$$R_M = \frac{\mathrm{var}(Y_M^{(eq)})}{\mathrm{var}(Y_M^{(0)})} = 1 + (n_X - 1)f_{X(eq)} \tag{49}$$

$$R_F = \frac{\mathrm{var}(Y_F^{(eq)})}{\mathrm{var}(Y_F^{(0)})} = 1 + (2n_X - 1)f_{X(eq)} \tag{50}$$

For large $n_X$, we can approximate $n_X f_{X(eq)}$ as

$$n_X f_{X(eq)} \approx \frac{\rho\left(3\rho + \sqrt{8 + \rho^2}\right)}{4(1 - \rho^2)}$$

Therefore, equations (49) and (50) can be approximated as

$$R_M \approx 1 + \frac{\rho\left(3\rho + \sqrt{8+\rho^2}\right)}{4(1-\rho^2)} \text{ and } R_F \approx 1 + \frac{2\rho\left(3\rho + \sqrt{8+\rho^2}\right)}{4(1-\rho^2)}, \tag{51}$$

which is exactely what found under normal distribution theory (equation 24).

Hence, for a large number of loci, the amount of disequilibrium variance created by AM in females is twice that in males. Note that although we have allowed for different effect sizes in males and females (and different variances), they dont influence the results because all variation is genetic.

An extension of our derivations to different effect sizes and different allele frequencies can be achieved similarly to the autosome case by replacing $n_X$ with the expression given in equation (38).

## Simulation of assortative mating

To simulate AM we proceed iteratively as follows. We start with $N$ simulated individuals ($N/2$ males and $N/2$ females) from a base population under random mating (RM) then sample mates pairs to simulate the next generations. We describe below our sampling strategy.

### The problem

Let us assume that we have observed $Y = (Y_1, \ldots, Y_N)$, the phenotypic values the $N$ individuals of the current generation. We now want to draw pairs of individuals $j$ and $k$ with probability $\theta_{jk} = \theta(Y_j, Y_k)$ such as the correlation between $Y_j$ and $Y_k$ equals $\rho$, the desired phenotypic correlation between mates. Overall, $Y_j$ is sampled with probability

$$\theta_j = \sum_{k=1}^{N} \theta_{jk}$$

Therefore, the sample mean under the desired sampling probability distribution is

$$\mu = \sum_{j=1}^{N} \theta_j Y_j = \sum_{j,k} \theta_{jk} Y_j.$$

We therefore need to define $\theta_{jk}$ (the joint sampling probability) such as

$$\sum_{j,k} \theta_{jk} Y_j Y_k - \left( \sum_{j,k} \theta_{jk} Y_j \right)^2 = \rho \left[ \sum_{j,k} \theta_{jk} Y_j^2 - \left( \sum_{j,k} \theta_{jk} Y_j \right)^2 \right] \tag{52}$$

under the constraints that

$$0 \leq \theta_{jk} \leq 1,\ \theta_{jk} = \theta_{kj},\ \theta_{jj} = 0 \text{ (i.e. no selfing allowed) and } \sum_{j,k} \theta_{jk} = 1.$$

Equation (52) can be rewritten as

$$\sum_{j,k} \left( Y_j Y_k - \rho Y_j^2 \right) \theta_{jk} = (1-\rho) \left( \sum_{j,k} \theta_{jk} Y_j \right)^2 \tag{53}$$

**Asymptotic solution - importance sampling**

Without loss of generality we assume that the $Y_j$ initially observed where independently drawn from a standard Gaussian distribution with 0 mean and variance 1. The assumption on the Gaussian distribution derives from that we assume a large number of causal variants. We also assume that $N$ is large. Equation (53) can be written as

$$\mathbb{E}[\left( Y_j Y_k - \rho Y_j^2 \right) \theta \left( Y_j, Y_k \right)] = (1-\rho) \mathbb{E}[\theta \left( Y_j, Y_k \right) Y_j]^2 \tag{54}$$

Written in terms of integral, this gives that

$$\int_{x,y} \left( xy - \rho x^2 \right) \theta(x,y) \phi(x) \phi(y) \mathrm{d}x \mathrm{d}y = (1-\rho) \left( \int_{x,y} x \theta(x,y) \phi(x) \phi(y) \mathrm{d}x \mathrm{d}y \right)^2 \tag{55}$$

where $\phi(.)$ is the probability density function of a standard Gaussian distribution. It therefore follows that a sensible choice for $\theta(x,y)$ is $p_\rho(x,y)/[\phi(x)\phi(y)]$ where $p_\rho(x,y)$ is the probability density function of the target distribution. In our case, our target distribution is a bivariate normal distribution with means 0 and variances 1 but with a correlation equal to $\rho$:

$$\theta(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] / \frac{1}{2\pi} \exp\left[-\frac{x^2+y^2}{2}\right]$$

or simply

$$\theta(x,y) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2x^2 - 2\rho xy + \rho^2y^2}{2(1-\rho^2)}\right) \tag{56}$$

We can therefore use that function $\theta(.,.)$ to sample the pairs. We still have to set that $\theta(x,x) = 0$ as we don't allow selfing.
For each $(N,\rho)$ we generated 100 samples.

## Simulation of offspring

We describe below how we simulate genotypes of offspring from that of their parents. All derivations presented below only consider bi-allelic variants such as single nucleotide polymorphisms (SNP).

### Autosomal variants

We consider diploid individuals and therefore represent genotypes as pairs of Bernoulli distributed variables indicative of the presence of the causal allele. We consider a trio consisting of a mother $(m)$, a father $(f)$ and one of their children $(o)$. Let us denote $\left(x_m^{(m)}, x_f^{(m)}\right)$, $\left(x_m^{(f)}, x_f^{(f)}\right)$ and $\left(x_m^{(o)}, x_f^{(o)}\right)$ respectively as the genotypes of $(m)$, $(f)$ and $(o)$ at a given locus. The subscript $m$ or $f$ indicates whether the causal allele was inherited from the mother or the father. For example, $x_m^{(f)}$ is the number of causal allele that $(f)$ inherited from his own mother.

We first consider the case of unlinked loci as each of them can be treated independently. Each parent contributes with one allele to the genotype of their offspring, and that contribution depends on the number

of causal alleles possessed by each parent. We can therefore express it as

$$x_m^{(o)}|x_m^{(m)}, x_f^{(m)} \sim \mathcal{B}\left(\frac{1}{2}\left[x_m^{(m)} + x_f^{(m)}\right]\right), \tag{57}$$

and

$$x_f^{(o)}|x_m^{(f)}, x_f^{(f)} \sim \mathcal{B}\left(\frac{1}{2}\left[x_m^{(f)} + x_f^{(f)}\right]\right), \tag{58}$$

where $\mathcal{B}(\pi)$ is the notation used to designate Bernoulli distributed variables of probability $\pi$. Equations (57) and (58) apply independently for each locus where unlinked markers are simulated.

The case of linked loci is slightly more complicated as offspring genotypes at different loci cannot, as for unlinked loci, be simulated independently from one another. We now denote $\mathbf{x}_{m,k}^i = \left(x_{j_1,m}^i, \ldots, x_{j_k,m}^i\right)$ as the vector of indicators of causal alleles across a chromosomal segment $k$ in individual $i = (m), (f)$ or $(o)$. As before, the substrict $m$ indicates that the segment was inherited from individual $i$'s mother. We can therefore similarly define $\mathbf{x}_{f,k}^i = \left(x_{j_1,f}^i, \ldots, x_{j_k,f}^i\right)$ as the vector of indicators of causal alleles inherited from individual $i$'s the father. To simulate the recombined chromosomes that $(o)$ has inherited from $(m)$, we used a 3-steps approach:

1. The first step consists of sampling the number $N_B$ of recombination breakpoints on the chromosome. If $L$ is the length of the chromosome in Morgan, then $N_B$ can be simulated using a Poisson distribution of parameter $L$: $N_B \sim \mathcal{P}(L)$.

2. The second step consists of sampling the location of recombination breakpoints on the chromosome. For that we used a uniform distribution of recombination events along the chromosome.

3. The last step consists of assembling parental recombined chromosomes to be transmitted. We first simulate a random sequence $(Z_1, \ldots, Z_{N_B+1})$ of $(N_B + 1)$ Bernoulli distributed variable with probability 1/2. The $Z_k$'s correspond to the $(N_B + 1)$ chromosome segments defined by the recombination breaks points and indicate whether the $k$-th transmitted segment is inherited from the mother $(Z_k = 1)$ or the father $(Z_k = 0)$. This can also we written as

$$\mathbf{x}_{m,k}^{(o)} = Z_k \mathbf{x}_{m,k}^{(m)} + (1 - Z_k)\mathbf{x}_{f,k}^{(m)}. \tag{59}$$

The same 3-steps approach is then repeated to simulate recombined chromosomes that ($o$) has inherited from the father ($f$).

**X-chromosome variants**

We use the same procedure as above to simulate genotypes on the X-chromosomes of females offspring. Given that male offspring only have one copy of the X-chromosome, we simulated that unique copy using equations (57) and (58) when considering unlinked loci and equation (59) when considering linked causal alleles. In our simulations, we used $L = 1.5$ Morgan for the X-chromosome and $L = 33$ Morgans for the autosome.

## Impact of linkage

We consider a simplified model with $M$ causal variants. We denote $a_j$ as the additive fixed effect of the $j$-th trait increasing allele (therefore $a_j > 0$) and $x_j^{(m)}$ and $x_j^{(f)}$ as the number trait-increasing alleles transmited by the mother and father respectively. At equilibrium positive assortative mating creates a positive correlation $\alpha$ between $x_j^{(m)}$ and $x_j^{(f)}$. We denote $r_{jk}$ as the correlation (linkage disequilibrium) between trait increasing alleles at SNP $j$ and SNP $k$ in the base population. $r_{jk}$ can be positive or negative a priori. For the sake of simplicity we assume that all alleles have the same frequency $p$ and we denote $q = 1-p$.

Useful formulas are ([20])

$$\text{var}[x_j^{(m)}] = \text{var}[x_j^{(f)}] = pq,\ \text{cov}[x_j^{(m)}, x_j^{(f)}] = \alpha pq \text{ and } \text{cov}[x_j^{(m)}, x_k^{(m)}] = \text{cov}[x_j^{(f)}, x_k^{(f)}] = D_{jk} = (r_{jk}+\alpha)pq. \tag{60}$$

Equation (60), in particular the last formula ($D_{jk} = (r_{jk} + \alpha)pq$), translates that the AM adds to the existing covariance between trait increasing alleles in the base population. For unlinked markers, $r_{jk}$ would be equal to 0.

Let us now derive the equilibrium genetic variance $V_A^{(eq)}$:

$$
\begin{aligned}
V_A^{(eq)} &= \text{var}\left[\sum_{j=1}^{M} a_j \left(x_j^{(m)} + x_j^{(f)}\right)\right] \\
&= \sum_{j=1}^{M} \text{var}\left[a_j \left(x_j^{(m)} + x_j^{(f)}\right)\right] + \sum_{j\neq k} \text{cov}\left[a_j \left(x_j^{(m)} + x_j^{(f)}\right), a_k \left(x_k^{(m)} + x_k^{(f)}\right)\right] \\
&= 2pq(1+\alpha)\left[\sum_{j=1}^{M} a_j^2\right] + \sum_{j\neq k} a_j a_k \left(2r_{jk}pq + 4\alpha pq\right) \\
&= \underbrace{2pq\left[\sum_{j=1}^{M} a_j^2\right] + 2pq\left(\sum_{j\neq k} a_j a_k r_{jk}\right)}_{\text{Varince in the base population}} + \underbrace{2pq\alpha\left[\sum_{j=1}^{M} a_j^2 + 2\sum_{j\neq k} a_j a_k\right]}_{\text{Inflation due to assortative mating}} \qquad (61)
\end{aligned}
$$

As previously reported in [2], equation (61) shows that the inflation in genetic variance due to assortative mating (last term in the equation) is independent of linkage. It is important to recall that this result holds because we assumed no selection and therefore the distribution of alleles inherited from the same parent will be unaffected by assortative mating.