Understand half-sibling misspecification

In this vignette, we investigate the theoretical behaviour of Pv3Rs when it is misspecified because data are from half-sibling parasites. For simplicity, we consider only the case of a single recurrent episode and we assume throughout that the prior distribution for recrudescence, relapse, and reinfection is uniform.

Before proceeding, we recall an observations about the posterior odds of relapse to reinfection, and relapse to recrudescence documented elsewhere. Let $\boldsymbol{\mathcal{G}}_{\mathrm{C}}, \boldsymbol{\mathcal{G}}_{\mathrm{L}}$, and $\boldsymbol{\mathcal{G}}_{\mathrm{I}}$ denote subsets of the graph space $\boldsymbol{\mathcal{G}}$, containing the relationship graphs compatible with recrudescence, relapse, and reinfection respectively. Given the prior on the three recurrent states for the single recurrent infection is uniform, the posterior odds of relapse to reinfection is given by

$$o_{\mathrm{L:I}} \coloneqq \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{L})}{\mathbb{P}(\boldsymbol{y}|\mathrm{I})} = \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\mathbb{P}(\boldsymbol{g}|\mathrm{L})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\mathbb{P}(\boldsymbol{g}|\mathrm{I})} = \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|} \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})} = \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|} \left(1 + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{I}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}\right), \tag{1}$$

where $\boldsymbol{\mathcal{G}}_{\mathrm{L}} \setminus \boldsymbol{\mathcal{G}}_{\mathrm{I}}$ is the subset of graphs compatible with relapse but not reinfection (graphs that have at least one non-stranger inter-episode edge). Similarly, the posterior odds of relapse to recrudescence is given by

$$o_{\mathrm{L:C}} \coloneqq \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{L})}{\mathbb{P}(\boldsymbol{y}|\mathrm{C})} = \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|} \left(1 + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{C}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}} \mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}\right). \tag{2}$$

It follows from these results that $o_{\mathrm{L:I}} \geq |\boldsymbol{\mathcal{G}}_{\mathrm{I}}|/|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|$ and $o_{\mathrm{L:C}} \geq |\boldsymbol{\mathcal{G}}_{\mathrm{C}}|/|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|$. In later sections, these bounds dictate the limiting behaviour of posterior probabilities as the number of markers increase. They define a feasible set of posterior probabilities, without any information about the genetic data. In the case of 2 genotypes in the initial episode and 1 genotype in the recurrent episode, we have $o_{\mathrm{L:I}} \geq 2/9$ and $o_{\mathrm{L:C}} \geq 4/9$. The resulting feasible set of posterior probabilities is shown in Figure 1.

Number of genotypes in initial episode	Number of genotypes in recurrent episode				
	1	2	3	4	5
1	0.3333	0.2222	0.1667	0.1339	0.1123
2	0.2222	0.1026	0.0581	0.0375	0.0263
3	0.1667	0.0581	0.0257	0.0135	0.0080
4	0.1339	0.0375	0.0135	0.0059	0.0030
5	0.1123	0.0263	0.0080	0.0030	0.0013

Table 1: Values for $|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|/|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|$ (lower bound of $o_{\mathrm{L:I}}$) for various graph sizes.

Half siblings

Half siblings share one parental genotype and draw collectively from three distinct parental genotypes. Given data on half siblings, Pv3Rs potentially underestimates the posterior probability of relapse because under the Pv3Rs model we assume an identity-by-descent (IBD) partition over siblings can have at most two cells; otherwise stated, siblings inherit from at most two parents. This, combined with the fact that we do not model genotype errors, means that the likelihood of a graph with a sibling component over three or more half siblings is zero as soon as the half siblings inherit three distinct alleles.

We seek to describe scenarios where the odds $o_{\mathrm{L:I}}$ are close to the lower bound $|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|/|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|$ for a simple example where there are 2 genotypes in the initial infection and 1 genotype in the recurrent infection. In what follows, we use the relationship graph labels shown alongside the five IBD partitions in Figure 2. To reduce the analytical computation, we assume that there is some marker j' such that there are three distinct alleles observed at marker j' across

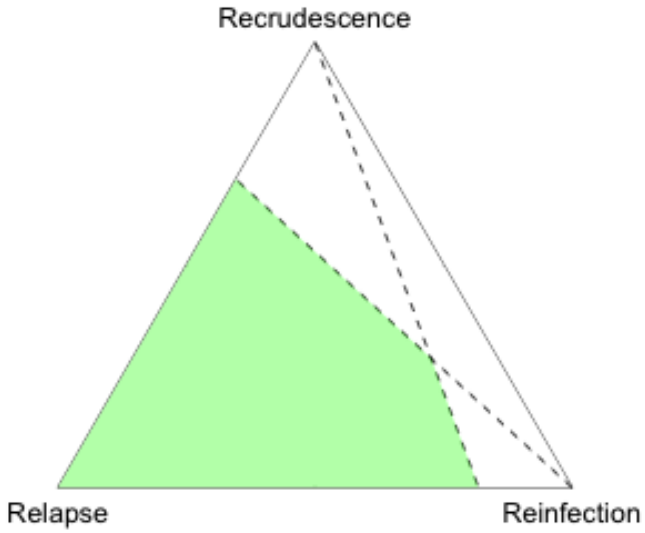

Figure 1: Feasible set of posterior probabilities in the case of two genotypes in the initial episode and one genotype in the recurrent episode. Dashed lines intersect the left and bottom edges of the simplex at $\frac{4}{9}$ and $\frac{2}{9}$, respectively.

the two episodes. This allows us to rule out the possibility of recrudescence, and also the graphs $\boldsymbol{g}_{\mathrm{II}}, \boldsymbol{g}_{\mathrm{III}}, \boldsymbol{g}_{\mathrm{VII}}, \boldsymbol{g}_{\mathrm{VIII}}$, and $\boldsymbol{g}_{\mathrm{IX}}$. In particular, we note that these three genotypes cannot be full siblings, which makes our subsequent analysis more specific to the scenario of half siblings. Under this scenario, the formula in (1) simplifies to

$$o_{\mathrm{L:I}} = \frac{2}{9}\left(1 + \frac{\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}}) + \mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{VI}})}{\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{I}}) + \mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})}\right) = \frac{2}{9}\left(1 + \frac{\prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) + \prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{VI}})}{\prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) + \prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}})}\right), \tag{3}$$

where $\boldsymbol{y}_{\cdot j}$ denotes the alleles observed at marker j. Note that the products account for all the phasing possibilities. To compute (3), we provide expressions for $\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_r)$ up to some proportionality constant that is the same for each $r = \mathrm{I}, \mathrm{IV}, \mathrm{V}, \mathrm{VI}$, but can vary across j. One further simplifying step is to note that $\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) = \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{VI}})$ due to symmetry upon accounting for all possible allele assignments.

Case 1: All differ: three distinct alleles are observed across the two episodes.
In this case, the common term $f(\alpha_j)f(\beta_j)f(\gamma_j)$ cancels and thus we have

$$\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) \propto 1, \quad \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}}) = \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) \propto 1/2.$$

Case 2: All match: the same allele is observed for all three genotypes.
In this case, the common term $f(\alpha_j)^2$ cancels and thus we have

$$\begin{aligned} \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) &\propto f(\alpha_j)^3 \quad & \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}}) = \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) &\propto (f(\alpha_j)^2 + f(\alpha_j)^3)/2 \\ &\propto f(\alpha_j), & &\propto (1 + f(\alpha_j))/2. \end{aligned}$$

Case 3: Intra-match: one allele is observed for the initial episode; a different allele for the recurrent episode.
In this case, the common term $f(\alpha_j)f(\beta_j)$ cancels and thus we have

$$\begin{aligned} \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) &\propto f(\alpha_j)^2 f(\beta_j) \quad & \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}}) &\propto (f(\alpha_j) + f(\alpha_j)^2)f(\beta_j)/2 \quad & \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) &\propto f(\alpha_j)^2 f(\beta_j)/2 \\ &\propto f(\alpha_j), & &\propto (1 + f(\alpha_j))/2, & &\propto f(\alpha_j)/2. \end{aligned}$$

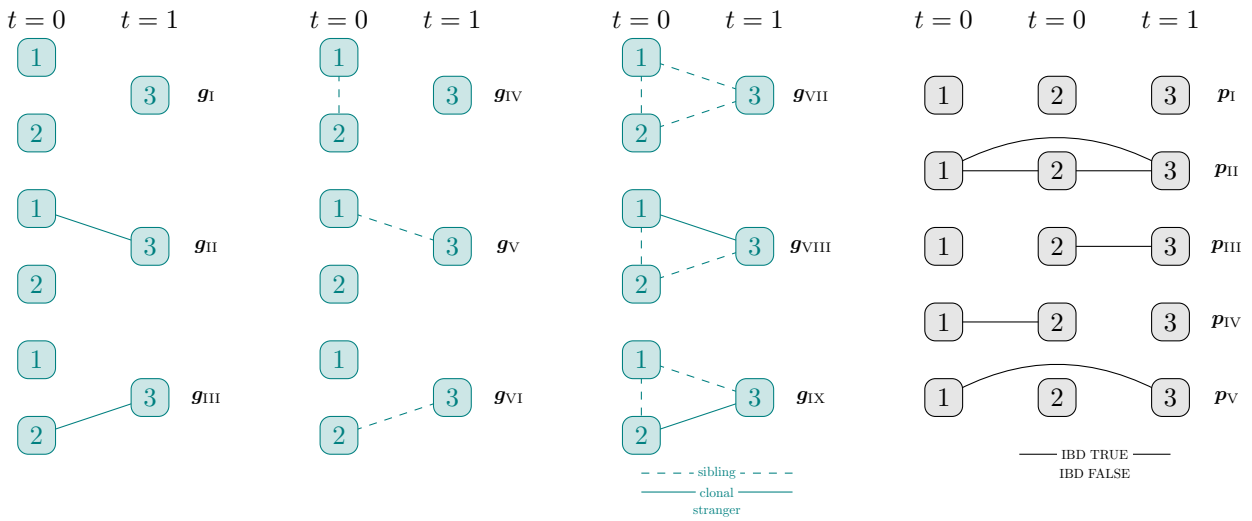

Figure 2: Relationsiph graphs ($\boldsymbol{g}_{\mathrm{I}}$ to $\boldsymbol{g}_{\mathrm{IX}}$) and identity-by-descent (IBD) partitions ($\boldsymbol{p}_{\mathrm{I}}$ to $\boldsymbol{p}_{\mathrm{V}}$) for the case of two genotypes in the $t = 0$ initial episode and one genotype in a $t = 1$ recurrent episode.

Case 4: Inter-match: two alleles are observed for the initial episode, one of which reappears at recurrence. In this case, the common term $f(\alpha_j)f(\beta_j)$ cancels and thus we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) &\propto f(\alpha_j)^2 f(\beta_j) \\ &\propto f(\alpha_j),\end{aligned} \qquad \begin{aligned}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}}) &\propto f(\alpha_j)^2 f(\beta_j)/2 \\ &\propto f(\alpha_j)/2.\end{aligned} \qquad \begin{aligned}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) &\propto 1/2((f(\alpha_j) + f(\alpha_j)^2)/2 + f(\alpha_j)^2/2)f(\beta_j) \\ &\propto (1 + 2f(\alpha_j))/4.\end{aligned}$$

Note that the computation for $\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}})$ is more complex as the two allele-to-vertex assignments result in different likelihoods.

These cases are hereafter referred to as observation cases. To illustrate how these observation cases can be combined to compute $o_{\mathrm{L:I}}$, consider an example where we have $m = 3$ markers, and the alleles observed for marker j follow observation case j for $j = 1, 2, 3$. Recalling that $\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}}) = \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{VI}})$, we compute the odds to be

$$\begin{aligned} o_{\mathrm{L:I}}|\text{marker } j \text{ follows case } j \text{ for } j = 1,2,3 &= \frac{2}{9}\left(1 + \frac{2\prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{V}})}{\prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{I}}) + \prod_{j=1}^{M}\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\mathrm{IV}})}\right) \\ &= \frac{2}{9}\left(1 + \frac{2 \cdot 1/2 \cdot (1 + f(\alpha_2))/2 \cdot f(\alpha_3)/2}{1 \cdot f(\alpha_2) \cdot f(\alpha_3) + 1/2 \cdot (1 + f(\alpha_2))/2 \cdot (1 + f(\alpha_3))/2}\right) \\ &= \frac{2}{9}\left(1 + \frac{(1 + f(\alpha_2))f(\alpha_3)/4}{f(\alpha_2)f(\alpha_3) + (1 + f(\alpha_2))(1 + f(\alpha_3))/8}\right). \end{aligned}$$

As $f(\alpha_3) \to 0$, $o_{\mathrm{L:I}}$ approaches $2/9$, which is the minimum possible value of $o_{\mathrm{L:I}}$. We also have

$$\begin{aligned} o_{\mathrm{L:I}}|\text{marker } j \text{ follows case } j \text{ for } j = 1,2,3 &= \frac{2}{9}\left(1 + \frac{(1 + f(\alpha_2))f(\alpha_3)/4}{f(\alpha_2)f(\alpha_3) + (1 + f(\alpha_2))(1 + f(\alpha_3))/8}\right) \\ &\leq \frac{2}{9}\left(1 + \frac{(1 + f(\alpha_2))f(\alpha_3)/4}{(1 + f(\alpha_2))(1 + f(\alpha_3))/8}\right) \\ &= \frac{2}{9}\left(3 - \frac{2}{1 + f(\alpha_3)}\right) \\ &\leq \frac{4}{9} \qquad \text{since } f(a_3) \leq 1, \end{aligned}$$

where equality holds when $f(\alpha_2) = 0$ and $f(\alpha_3) = 1$.

Special case: three equifrequent alleles per marker

We now turn our attention to the case where every possible allele (3 per marker) is assumed to have a frequency of $1/3$. For each $c = 1, \ldots, 4$, let m_c denote the number of markers where the alleles observed correspond to observation

case c (e.g., $m_2 = 3$ means all alleles match at three markers). Note that we have $m = m_1 + m_2 + m_3 + m_4$, and the assumption that there is some marker j' such that there are three distinct alleles observed at marker j' can be expressed as $m_1 \geq 1$. Substituting $f(\alpha_j) = 1/3$ into (3) for each $j = 1, \ldots, 4$ gives (4), where the second term in the parentheses is expressed in base two for interpretability as follows.

$$\begin{aligned} o_{\text{L:I}} &= \frac{2}{9}\left(1 + \frac{2 \cdot (1/2)^{m_1} \cdot (2/3)^{m_2} \cdot (1/6)^{m_3} \cdot (5/12)^{m_4}}{1^{m_1} \cdot (1/3)^{m_2} \cdot (1/3)^{m_3} \cdot (1/3)^{m_4} + (1/2)^{m_1} \cdot (2/3)^{m_2} \cdot (2/3)^{m_3} \cdot (1/6)^{m_4}}\right) \\ &= \frac{2}{9}\left(1 + \frac{2 \cdot (5/2)^{m_4}}{2^{m_1} \cdot (1/2)^{m_2} \cdot 2^{m_3} \cdot 2^{m_4} + 4^{m_3}}\right) \\ &= \frac{2}{9}\left(1 + \frac{2^{\log_2(5/2)m_4+1}}{2^{m-2m_2} + 2^{2m_3}}\right). \end{aligned} \tag{4}$$

Next, suppose that the three genotypes are offspring genotypes generated under the following sampling scheme:

1. Sample three parental genotypes, with alleles drawn randomly according to their frequency.
2. For each pair of parental genotypes, produce an offspring genotype, with alleles drawn uniformly at random from the two parents.

Under this sampling scheme, and the assumption of equifrequent alleles, the first step can result in one of 27 equally likely outcomes. For each of these outcomes, there are 8 possible ways for the offspring to draw alleles from the parents. By grouping these $27 \times 8 = 216$ possibilities, we find that observation cases $1, 2, 3, 4$ are expected to occur for $1/18, 5/18, 2/9, 4/9$ of the markers respectively. This means that for large m, we expect that $m_4, m - 2m_2, 2m_3$ should all be 'close' to $4/9 \cdot m$. The extra constant of $\log_2(5/2)$ in the exponent of the numerator in (4) implies that the odds diverge to ∞ as $m \to \infty$ because for large m, $o_{\text{L:I}} \sim \frac{2}{9}\left(1 + 2^{\frac{4}{9}m\log_2(5/4)}\right)$. However, in the case of finite m, a small perturbation to the ratios between $m_4, m - 2m_2, 2m_3$ can lead to a large deviation in the odds. Let $\bar{m}_c$ denote the expected number of markers that correspond to the observation case c for $c = 1, 2, 3, 4$. Consider the case where $(m_1, m_2, m_3, m_4) = (\bar{m}_1, \bar{m}_2, \bar{m}_3 + 0.08m, \bar{m}_4 - 0.08m)$, i.e., where there is a slight over-representation of intra- versus inter-matches. We have

$$o_{\text{L:I}} = \frac{2}{9}\left(1 + \frac{2^{0.4818m+1}}{2^{0.4444m} + 2^{0.6044m}}\right) < \frac{2}{9}\left(1 + 2^{-0.123m+1}\right),$$

which quickly converges to $2/9$ as m increases. From (4), we expect the posterior probability to concentrate on relapse (reinfection) when the intra-to-inter match ratio m_3/m_4 is much smaller (larger) than $1/2\log_2(5/2)$, as long as the term 2^{m-2m_2} is relatively negligible to compared to 2^{2m_3}. However, it is unlikely that these results will hold under other scenarios, e.g. non-equifrequent alleles, bigger graph size, different number of possible alleles for each marker, or a sampling scheme that does not follow the allele frequencies.

Beyond equifrequent alleles

Our analysis thus far demonstrates the value of studying the relative sizes of graph likelihoods, i.e. likelihood ratios between different relationship graphs, for investigating misclassifying (inter-episode) half sibling relapse scenarios as reinfections. Some of these likelihood ratios are sensitive to the frequency of the repeat allele, especially when the frequency is small; see Table 2. In particular, when $\boldsymbol{g}_{\text{IV}}$ (intra-sib) or $\boldsymbol{g}_{\text{V}}$ (inter-sib) is comparatively favoured, the likelihood ratio depends on the frequency f of the repeat allele. Note that the expressions coloured light blue and green diverge to ∞ as $f \to 0$.

However, the likelihood ratios shown in Table 2 are computed for one marker only. The 'full' likelihoods $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) = \prod_{j=1}^{m} \mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g})$ depend also on the relative frequency of each observation case. As an illustration, consider a scenario where rare alleles (alleles with very low frequency) are over-represented in the observed data, i.e. we often encounter observation cases 2 (All match), 3 (Intra-match), 4 (Inter-match) with small f. We claim that under this scenario, the full likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{I}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{IV}})$ (all strangers to intra-sib) should be quite small: $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{I}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{IV}})$ is the product of the likelihood ratios $\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\text{I}})/\mathbb{P}(\boldsymbol{y}_{\cdot j}|\boldsymbol{g}_{\text{IV}})$ (first column of Table 2) over each marker j. Since f is small, the effect of the small likelihood ratios under observation cases 2 (All match) and 3 (Intra-match) would dominate observation case 4 (Inter-match), leading to a small full likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{I}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{IV}})$. Similar reasoning applies to $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{I}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{V}})$ (all strangers to inter-sib). On the other hand, the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\text{V}})$ (intra-sib to inter-sib) is more sensitive to how small f is for each marker, and on the relative

Observation case	Likelihood ratio		
	$\frac{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{I}})}{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{IV}})}$ $\left(\frac{\text{all stranger}}{\text{intra-sib}}\right)$	$\frac{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{I}})}{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{V}})}$ $\left(\frac{\text{all stranger}}{\text{inter-sib}}\right)$	$\frac{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{IV}})}{\mathbb{P}(\boldsymbol{y}_{\cdot j}\mid \boldsymbol{g}_{\mathrm{V}})}$ $\left(\frac{\text{intra-sib}}{\text{inter-sib}}\right)$
1 (All differ)	2	2	1
2 (All match)	$1\Big/\left(\frac{1}{2}+\frac{1}{2f}\right)$	$1\Big/\left(\frac{1}{2}+\frac{1}{2f}\right)$	1
3 (Intra-match)	$1\Big/\left(\frac{1}{2}+\frac{1}{2f}\right)$	2	$\left(1+\frac{1}{f}\right)$
4 (Inter-match)	2	$\frac{4f}{1+2f}$	$1\Big/\left(1+\frac{1}{2f}\right)$

Table 2: Likelihood ratios (one marker) between different relationship graphs for observation cases 1, 2, 3, 4, where f is the frequency of the repeat allele. Colours indicate which graph ($\boldsymbol{g}_{\mathrm{I}}, \boldsymbol{g}_{\mathrm{IV}}, \boldsymbol{g}_{\mathrm{V}}$) is comparatively favoured given an observation case. A ratio in black indicates that neither graph is clearly favoured, either because the ratio is exactly 1, or because the ratio can be greater or less than 1 depending on f.

frequencies of observation cases 3 and 4 (last column of Table 2). All else being equal, observation case 4 (Inter-match) should occur twice as often as observation case 3 (Intra-match) due to symmetry. Since

$$\frac{1+\frac{1}{f}}{(1+\frac{1}{2f})^2} = \frac{f^2+f}{f^2+f+\frac{1}{4}} \to 0 \qquad \text{as } f \to 0,$$

we expect that the full likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$ (intra-sib to inter-sib) would be close to zero. By substituting these findings into (3), we expect the odds $o_{\mathrm{L:I}}$ to be large under the large m, small f limit.

Now suppose we modify the offspring sampling scheme from Section such that there are d possible alleles for each marker, with allele frequencies $f_1, f_2, \ldots, f_d$. The probabilities of the observation cases are given by:

$$\begin{aligned} \mathbb{P}(\text{All differ, case 1}) &= \frac{3}{2}\sum_{i<j<k} f_i f_j f_k, \qquad \mathbb{P}(\text{All match, case 2}) = \sum_i f_i^3 + \frac{3}{4}\sum_{i\neq j} f_i^2 f_j, \\ \mathbb{P}(\text{Intra-match, case 3}) &= \frac{1}{4}\left(1-\sum_i f_i^3\right), \quad \mathbb{P}(\text{Inter-match, case 4}) = \frac{1}{2}\left(1-\sum_i f_i^3\right), \end{aligned} \tag{5}$$

where all summation indices are bounded between 1 and d. The working is omitted here, but (5) can be derived by introducing latent binary variables corresponding to the specific parents each offspring inherits from, and marginalising out these latent variables according to each observation case.

Now suppose that the parent alleles are sampled from

$$(f_1, f_2, \ldots, f_d) \sim \mathrm{Dir}(\lambda, \lambda, \ldots, \lambda) \tag{6}$$

for some concentration parameter $\lambda > 0$. The equifrequent assumption corresponds to setting $\lambda \to \infty$. Under (6), we can compute that

$$\begin{aligned} \mathbb{E}[\mathbb{P}(\text{All differ, case 1})] &= \frac{(d-1)(d-2)\lambda^2}{4(d\lambda+1)(d\lambda+2)}, \qquad \mathbb{E}[\mathbb{P}(\text{All match, case 2})] = \frac{((3d+1)\lambda+8)(\lambda+1)}{4(d\lambda+1)(d\lambda+2)}, \\ \mathbb{E}[\mathbb{P}(\text{Intra-match, case 3})] &= \frac{1}{2}\,\mathbb{E}[\mathbb{P}(\text{Inter-match, case 4})] = \frac{(d-1)((d+1)\lambda+3)\lambda}{4(d\lambda+1)(d\lambda+2)}. \end{aligned} \tag{7}$$

Consider the case where $\lambda \to 0$, i.e. one of the allele frequencies dominates the distribution. By inspecting the leading order terms (lowest degree with respect to λ) in (7), we note that when $\lambda \to 0$, case 1 (All differ) would be the least common and case 2 (All match) would be the most common.

Limiting distributions give rise to erratic behaviour

In the case of $d = 3$ possible alleles per marker, large number of markers m, and $\lambda \to \infty$, simulations show that the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$ (intra-sib to inter-sib) is not always close to 0 (see online article "Understand posterior estimates"). The same also occurs when $\lambda \to 0$ (these simulations are not documented). We can corroborate these results analytically. Let c_j denote the observation case number for marker j. Using the results from

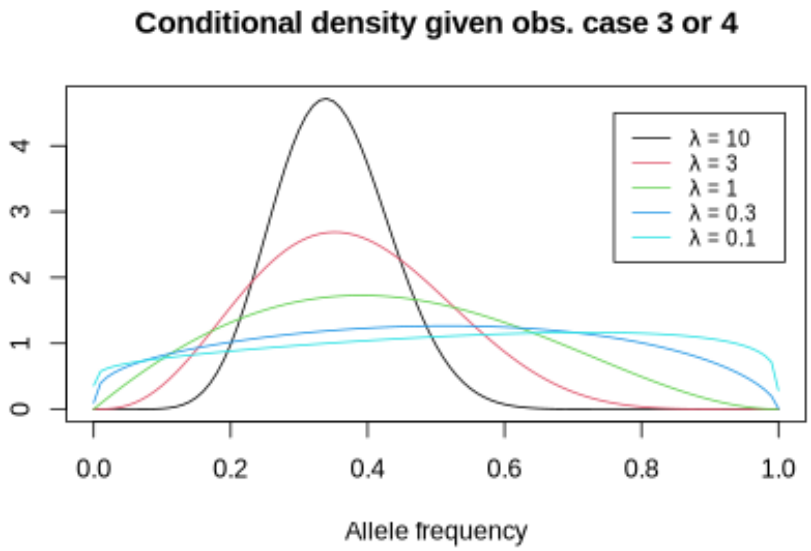

Figure 3: Conditional distribution of the frequency of a repeat allele given observation case 3 or 4.

Table 2, we obtain

$$\frac{\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})}{\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})} = \frac{\displaystyle\prod_{j:\, c_j=3}\left(1+\frac{1}{f(\alpha_j)}\right)}{\displaystyle\prod_{j:\, c_j=4}\left(1+\frac{1}{2f(\alpha_j)}\right)}. \tag{8}$$

However, the distribution of $f(\alpha_j)$, which we write below as f for brevity, does not simply follow (6) as we have conditioned on observation case 3 or 4. Instead, the conditional density of the frequency f of the repeat allele is given by (working not shown):

$$p(f|\text{observation case 3}) = p(f|\text{observation case 4}) = \frac{3\lambda+2}{4\lambda+3}\frac{(1+f)f^{\lambda}(1-f)^{2\lambda}}{\mathrm{B}(\lambda+1,2\lambda+1)}, \tag{9}$$

where $\mathrm{B}(\cdot,\cdot)$ denotes the beta function. The result is obtained by combining a beta prior for f (marginal distribution of (6)) and a likelihood derived in a similar fashion to the results from (5). A plot of the conditional density (9) for various values of λ is shown in Figure 3. Notice that as λ decreases, the conditional mean of the allele frequency f increases, and the tails of the conditional density become heavier. This gives the following implications for the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$, which is inversely related to the posterior probability of relapse (see (3)):

- When λ is very large, f is close to $1/3$. From the calculations in Section 2.1 (for equifrequent alleles), the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$ is close to 0. However, minor perturbations to the ratio m_3/m_4 can cause this likelihood ratio to be larger than expected.

- As λ decreases, the conditional mean of f increases. The factors $1 + 1/f$ and $1 + 1/2f$ from (8) are thus smaller on average, diminishing the effect of perturbing the ratio m_3/m_4. This makes the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$ more likely to be close to 0.

- As λ gets close to 0, the tails of the conditional density of f become heavier. In particular, the factors $1 + 1/f$ and $1 + 1/2f$ from (8) can be quite large for small values of f. Moreover, observation cases 3 and 4 occur less frequently, thus the ratio m_3/m_4 is subject to greater perturbation due to stochasticity. Altogether, the variance of the likelihood ratio $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{IV}})/\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}_{\mathrm{V}})$ increases, and thus the likelihood ratio itself is no longer reliably close to 0.