Simulation study for the genetic model

In this script, simulated genetic data are generated and plotted, and the results of the genetic model applied to those data are loaded and plotted. The simulated data are analysed under the model in a separate script called Generate_GenSimResults.R. This script relies on BuildSimData.R, a function that for a set of input parameters generates data for a graph over an initial episode and single recurrence. The graph contains a single between-episode edge representing a parasite haploid genotype with specified relationship in stranger, sibling or clone, which is among otherwise unrelated 'stranger' parasites if the input complexity of infections (COIs) exceed one. We chose this type of graph, in which the 'noisy' parasite haploid genotypes are unrelated since it is the most diverse thus most computationally challenging. In other words, it is a extreme scenario which is used to bound the performance of the model. Indeed, some of the problems we encounter in this extreme scenario do not apply to the real data. Complex infections in the real data most likely derive from either relapsing or co-inoculated parasites (there is little opportunity for superinfection due to active follow up), and thus are liable to be interrelated, therefore less diverse.

Outline of this script:

- For each "job", simulate data for N individuals, with M = 3 to 12 markers for two episodes, the second episode including a single clonal, sibling or stranger parasite. We consider several jobs: for each of the stranger, sibling and clonal scenario, we simulated data for an initial and recurrent infection with respective COIs: 1 and 1, with and without error; 2 and 1, with and without error; 1 and 2, with and without error; 3 and 1, without error. For the non-erroneous episodes with COIs of 1 and 2, we explored cardinalities of 4 (the minimum effective cardinality of any marker in the MS data that feature in the main text) and 13 (the mean effective cardinality of the MS data that feature in the main text). For the erroneous data and for the episodes with COIs of 3 and 1 cardinality was 13.
- Summarize the simulated data with a series of plots.
- Compute resulting recurrence state estimates (this is currently done in a separate file).
- Plot resulting recurrence state estimates as a function M.

Fraction of markers at which evidence of IBS is detected

Figures 1 and 1 show the fraction of markers at which evidence of IBS is detected for non-erroneous and erroneous data, respectively. When the recurrent episode contains a clone of a parasite haploid genotype in the initial episode and the data are non-erroneous (as they are in Figure 1), the fraction of markers at which evidence of IBS is detected is always one. We thus do not plot the clonal scenario for non-erroneous data. On average, the mean fractions (vertical colored bars) range from approximately 0.1 under the stranger scenario with high cardinality (erroneous and not), to above 0.6 under the sibling scenario with lower cardinality (non-erroneous) and above 0.6 under the clonal scenario with higher cardinality (erroneous).

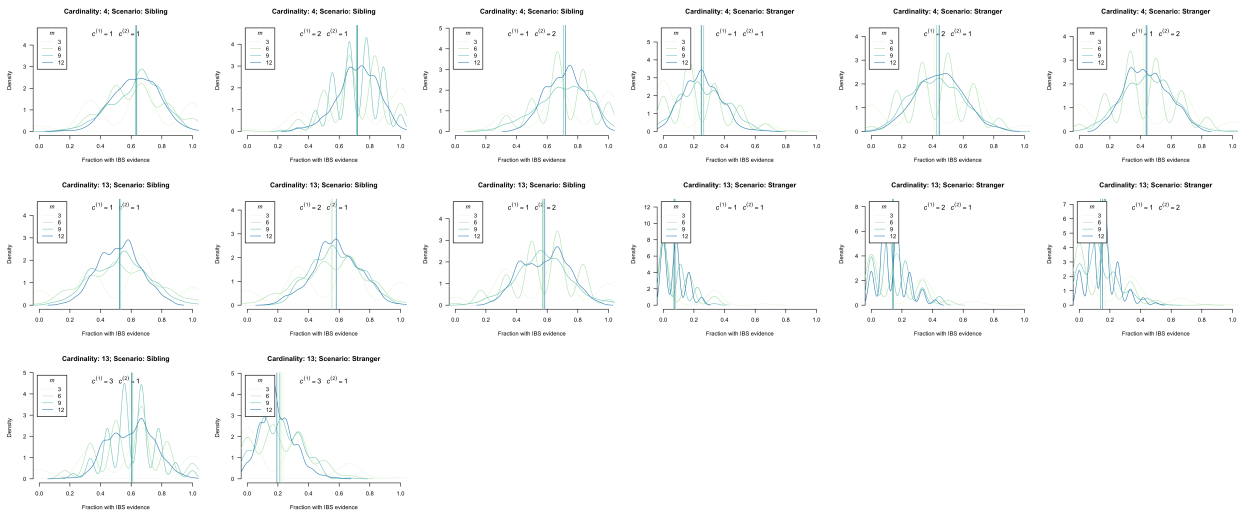

Figure 1: The fraction of markers at which evidence of IBS is detected when the recurrent episode is compared with the initial episode: non-erroneous data. Vertical coloured bars denote mean fractions. Different colours represent results for different numbers of markers, m.

Results

Inference as a function of the number of markers typed discounting error

In Figures 3 to 8, each plot corresponds to a different relationship simulation scenario where data are simulated without error. Colored bars show the median posterior probabilities with error bars extending $\pm$ one standard deviation. The effective cardinality of each marker was set equal to either 4 or 13 in these simulations. The complexities of infection were either one in both first and second infections, two in the first and one in the second, or one in the first and two in the second. The prior probability of each recurrence state was 1/3. As the number of microsatellites genotyped is increased, recurrent state probabilities converge as expected:

- under the sibling scenario, the probabilities converge to one for relapse and zero otherwise;
- under the stranger scenario, the probability of reinfection converges to a probability greater than the prior, meanwhile the probability of relapse converges to the complement of reinfection and probability of recrudescence converges to zero. Note that this is true even when the fraction of markers with evidence of IBS is close to 0.5 (e.g. compare left-top plot of Figure 1 and middle plot of Figure 5).
- under the clonal scenario, the probability of recrudescence converges a probability greater than the prior, meanwhile the probability of relapse converges to the complement of recrudescence and the probability of reinfection converges to zero.

Convergence happens most slowly under the sibling scenario. As such, the sibling scenario dictates marker requirements in general. When the effective cardinality is 4, more than 12 markers are required for full convergence. When the effective cardinality is 13, 9 or more markers are required.

The genetic model relies on data that list alleles detected at genotyped microsatellite markers (i.e. alleles are either detected or not). The model does not account for error in the alleles detected, nor incorporate weighted evidence of majority versus minority alleles. First, let's consider the failure to detect minority clones, second let's consider the impact of error.

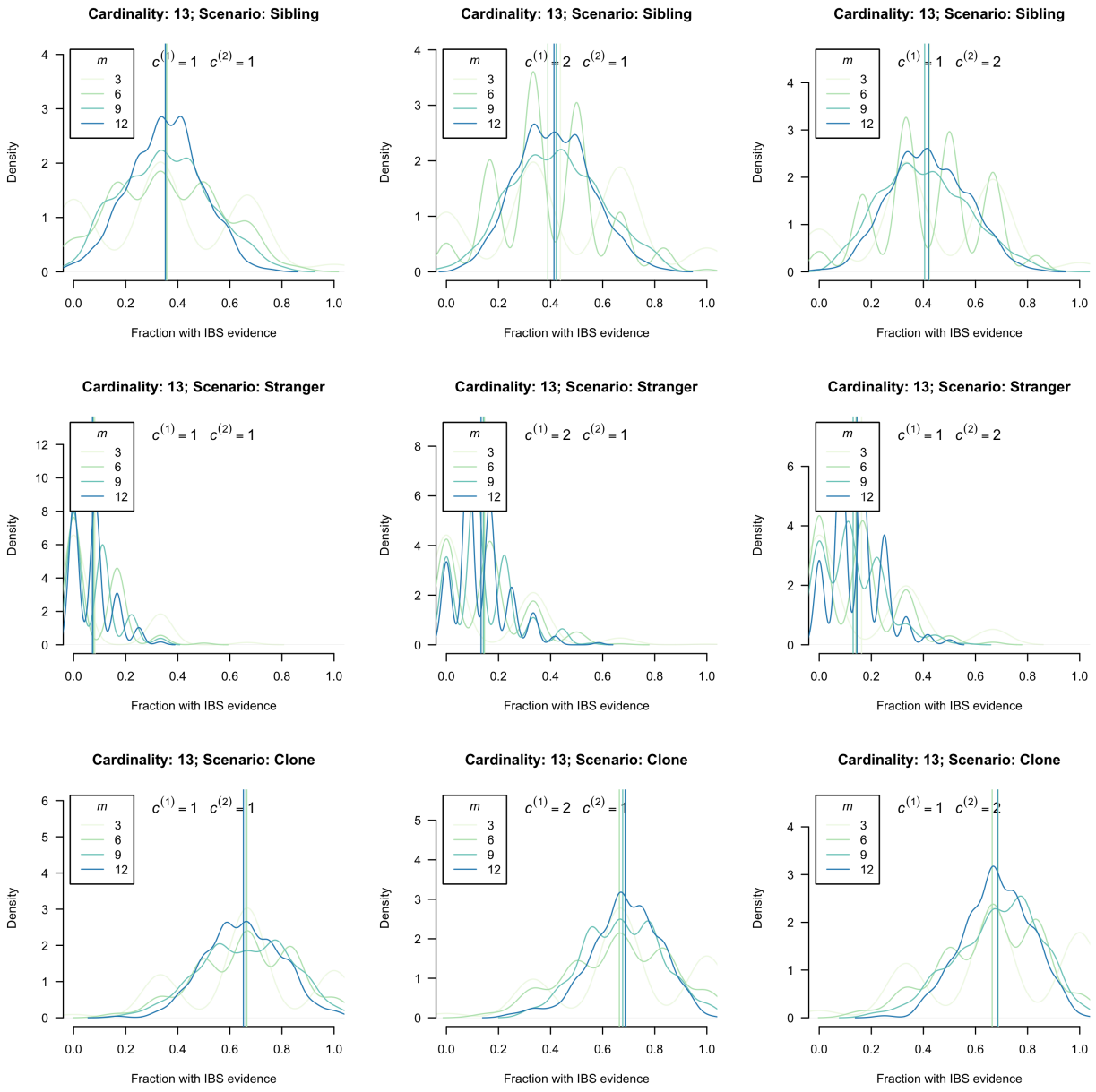

Figure 2: The fraction of markers at which evidence of IBS is detected when the recurrent episode is compared with the initial episode: erroneous data. Vertical coloured bars denote mean fractions. Different colours represent results for different numbers of markers, m.

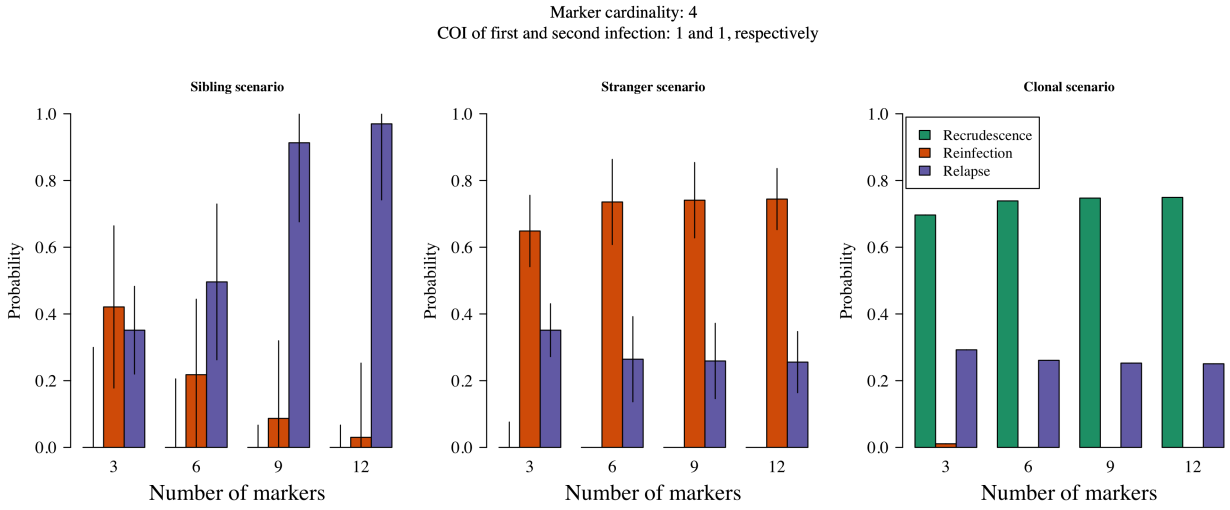

Figure 3: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

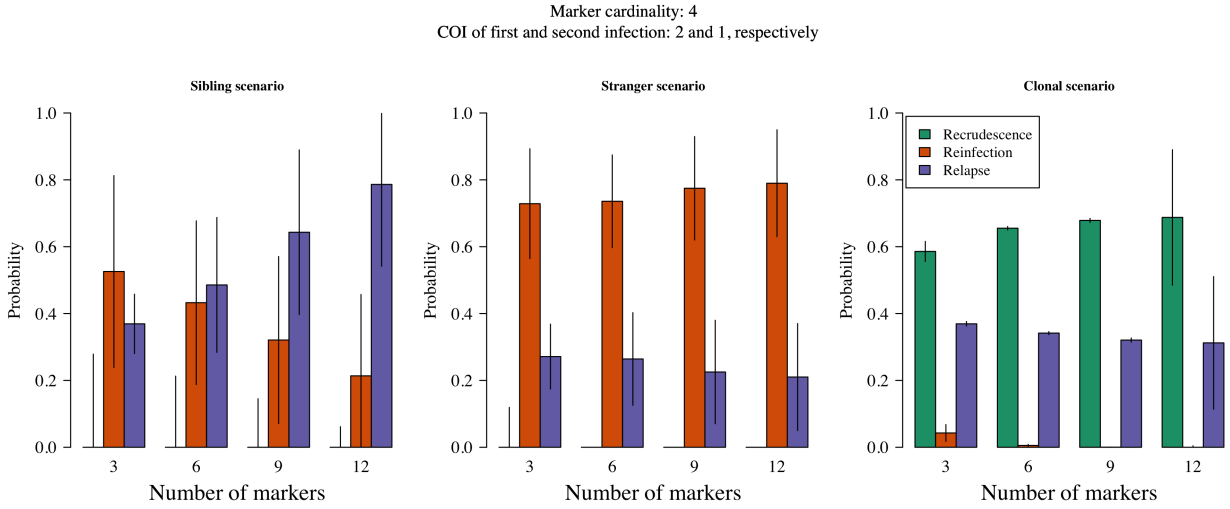

Figure 4: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

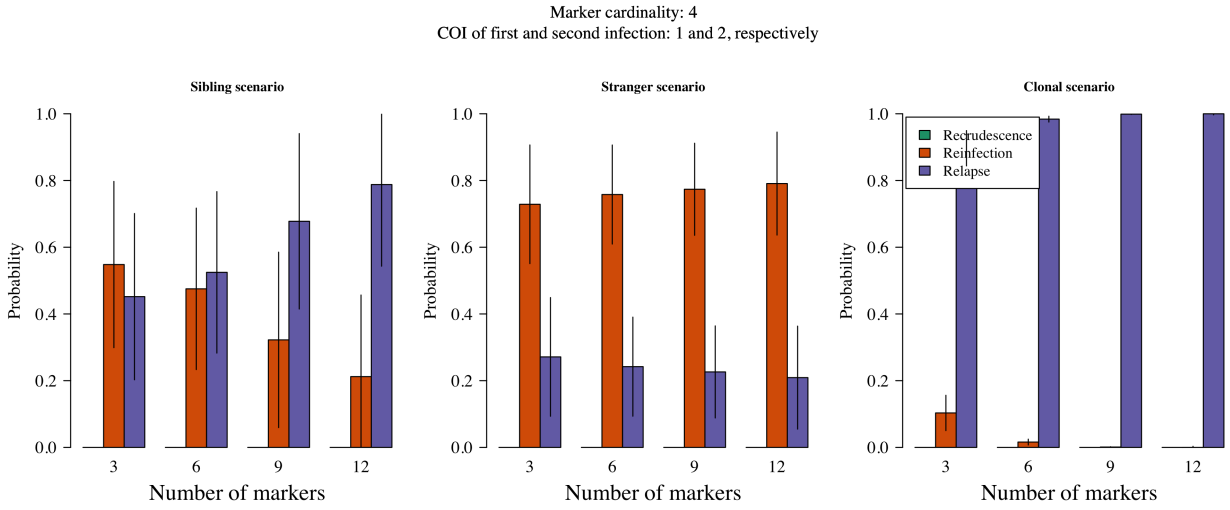

Figure 5: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

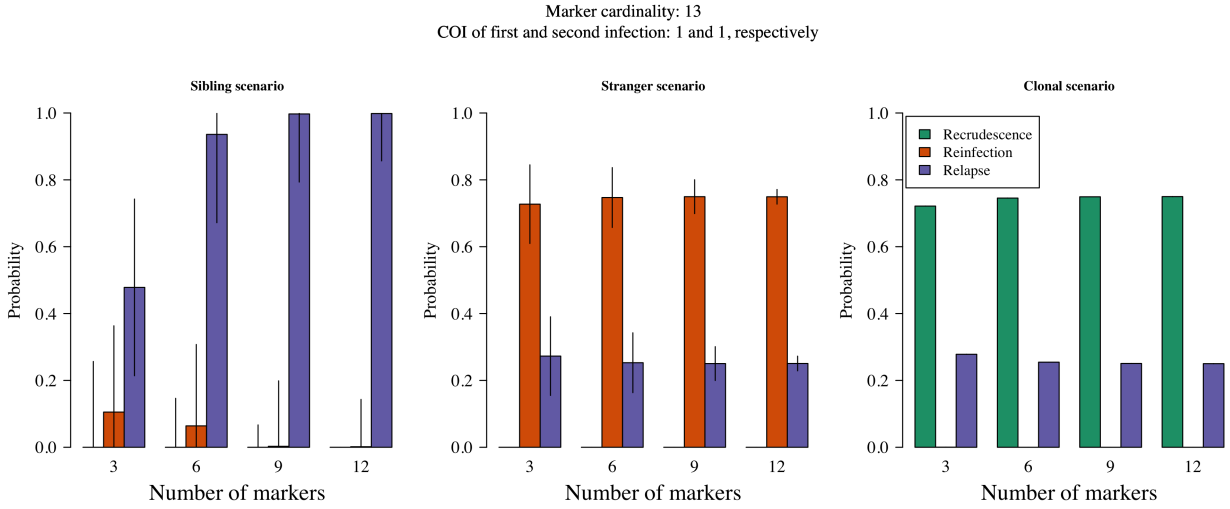

Figure 6: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

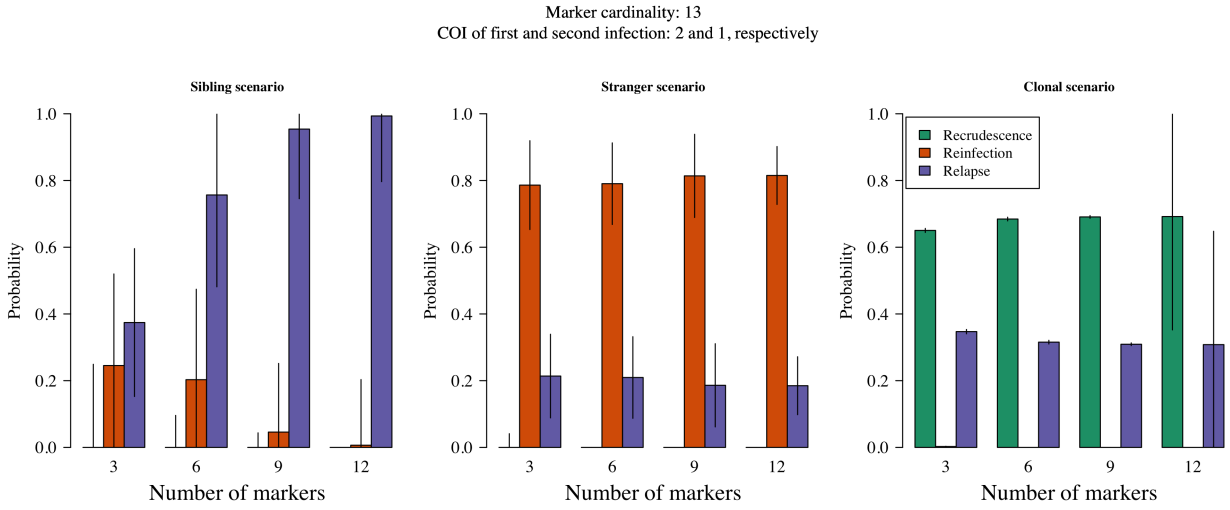

Figure 7: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

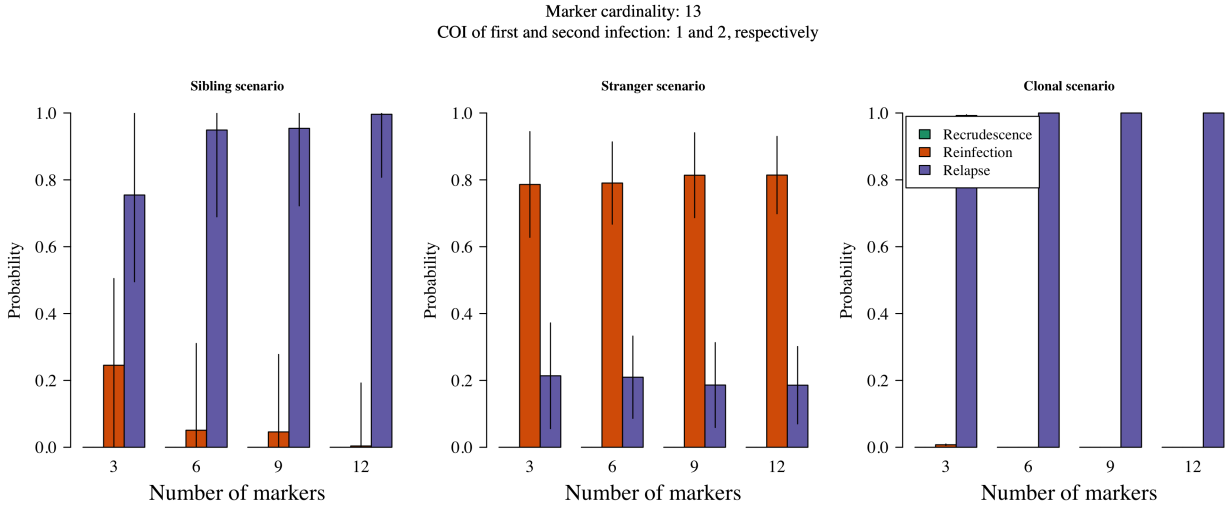

Figure 8: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario without error.

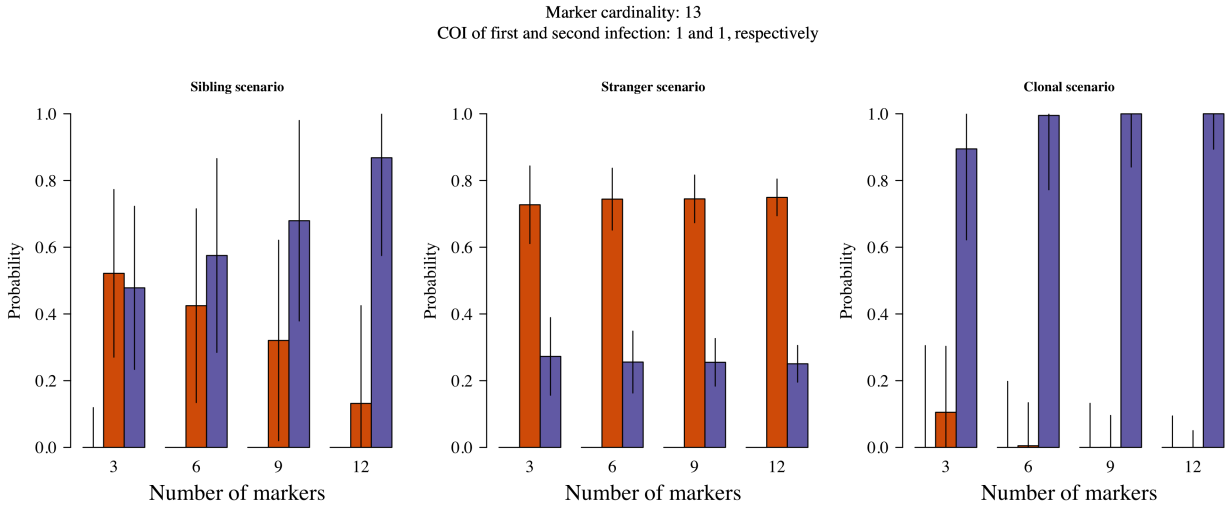

Figure 9: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario assuming an extremely high probability of error.

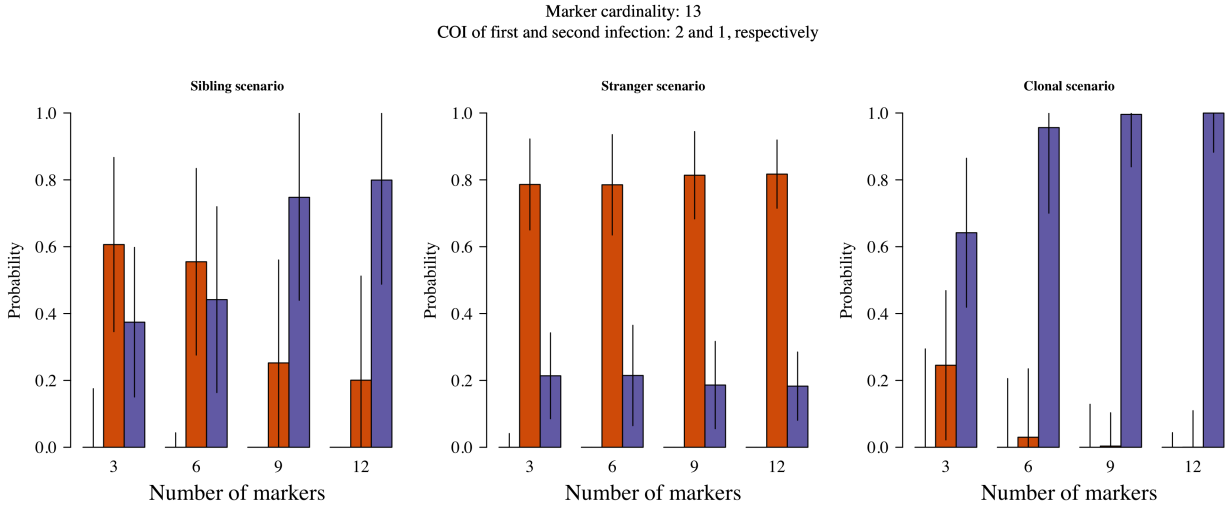

Figure 10: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario assuming an extremely high probability of error.

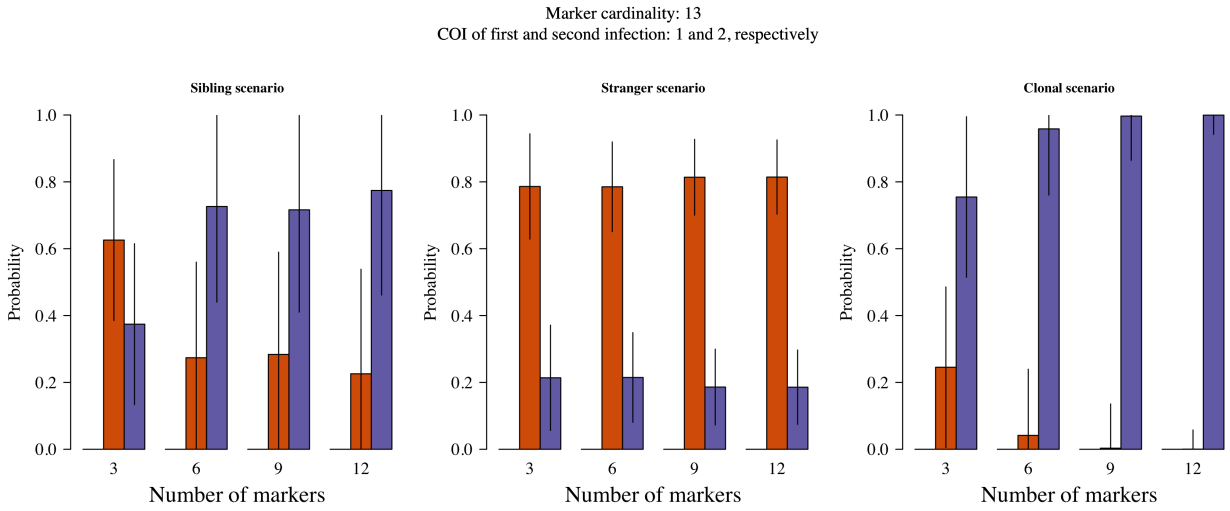

Figure 11: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario assuming an extremely high probability of error.

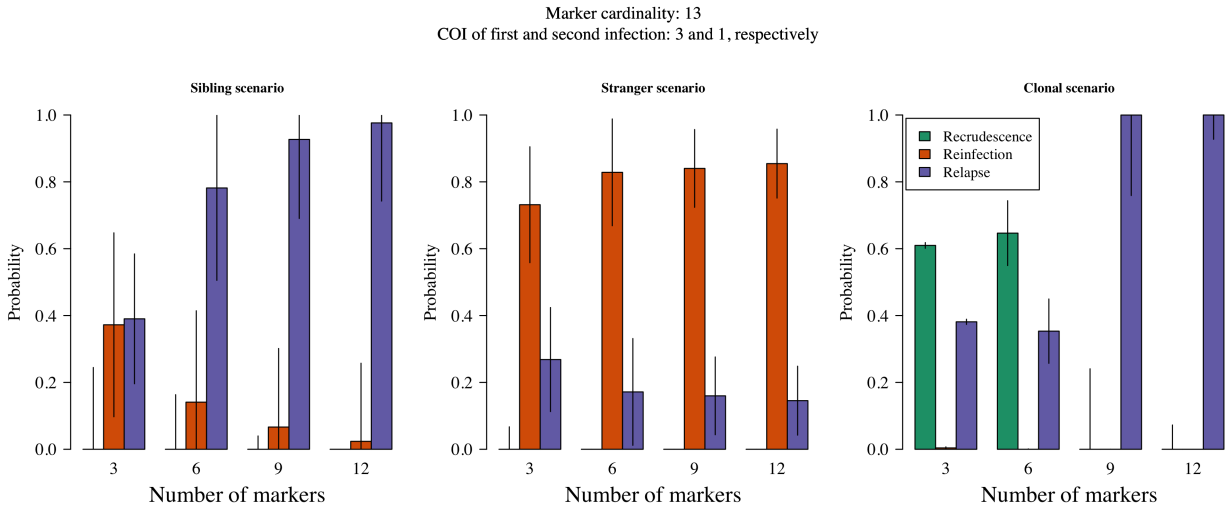

Figure 12: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario when the COI of the initial infection is three and the COI of the recurrent infection is one.

Undetected parasite haploid genotypes

Failure to detect data from a minority parasite haploid genotype will have different consequences depending on the relationship of the minority parasite haploid genotype in relation to other parasite haploid genotypes across episodes. For example, referring to the plots in Figures 6 and 7 as illustrative scenarios where 'COI x y' denotes a COI of x in the first infection and a COI of y in the second infection,

- in the Sibling COI 2 1 case (left plot, Figure 7) failure to detect the stranger parasite will result in the Sibling COI 1 1 case (left plot, Figure 6), thereby increasing the probability of relapse; meanwhile, failure to detect the sibling parasite will result in the Stranger COI 1 1 case (middle plot, Figure 6), thereby decreasing the probability of relapse, but not erasing it. Note that the case in which the noisy parasite is unrelated demonstrates the most severe possible outcome: if the noisy parasite were related, failure to detect it would result in a Sibling COI 1 1 case thereby maintaining the probability of relapse.
- in the Stranger COI 2 1 case (middle plot, Figure 7) failure to detect either stranger parasite will result in the Stranger COI 1 1 case (middle plot, Figure 6), maintaining the probability of reinfection and relapse.
- In the Clone COI 2 1 case (right plot, Figure 7) failure to detect the stranger parasite will result in the Clone COI 1 1 case (right plot, Figure 7), thereby maintaining the probability of recrudescence and relapse; meanwhile, failure to detect the clonal parasite will result in the Stranger COI 1 1 case (middle plot, Figure 7), thereby replacing the probability of recrudescence with reinfection.

The examples above illustrate the robust versus frail nature of relapse inference versus recrudescence inference, respectively. As we shall see in the next section, relapse inference is also robust in the presence of error, whereas recrudescence is not.

Erroneous data

Figures 9 to 11 shows inference in the presence of unmodelled error. The probability of error, 0.2, was set extremely high to clearly illustrate model behavior. Realistic error rates will have much less impact. Error largely impacts impacts inference of recrudescence: in the Clonal scenario clonal parasites are interpreted as sibling parasites and the probability of relapse tends towards one.

Highly complex data

A major limitation of the genetic model has to do with computational complexity. One aspect relevant to the present simulation study is described below. When samples are complex and highly diverse, e.g. when they contain majority unrelated parasites, unconverged probabilistic phasing is liable to miss clonally compatible combinations among the vast number of combinations that are possible. Recall that clonally compatible combinations are the only ones compatible with recrudescence. The total number of possible combinations grows exponentially with the number of markers genotyped, making the probability that probabilistic phasing captures the few clonally compatible combinations increasing slim and rendering probability estimates increasingly frail for recrudescence (Figure 12). Such highly complex scenarios are extreme. They are helpful for illustrating the problem (Figure 12), but not representative of the VHX and BPD data: all those analysed using probabilistic phasing converged. Otherwise stated, inconsistency is not a problem VHX and BPD data analysed under the model, but could be for future data sets.

Conclusion

The current genetic model does not account for error in alleles detected, nor incorporate weighted evidence of majority versus minority alleles. These omissions render inference of recrudescence under the current model brittle, but have little impact on inference of reinfection versus relapse. As such, analyses of data from the Thailand Myanmar border, where evidence of resistant *P. vivax* is lacking, are likely robust to the above omissions. However, the model merits extension before application to data from a region where *P. vivax* resistance is suspected.