Simulation study to test vivax genetic relatedness model

- Most relationship graphs graphs compatible with relapse contain sibling edges.
- Reviewer's example: "a recurrence with MOI of 3, containing a clone and two unrelated strains, but with overall pairwise relatedness close to 0.5."
- Add some plots of erroneous and higher MOI data.

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 anlaysed under the model in a separate script called Generate_GenSimResults.R. This script also 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 COIs exceed one. We chose this type of graph, in which the`noisy' parasite haploid geneotypes are unrelated since it...

Outline of this script:

- For each "job", simulate data for N individuals, with M markers for two episodes, the second episode including a single clonal, sibling or stranger parasite.
- Summarise 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

In the first simulation, we want to assess recurrence state inference as a function of the number of markers typed, adding an extra noisy parasite into an infection when its COI exceeds one. We consider two effective cardinalities: 4 (the minimum effective cardinality of any marker in the MS data that feature in the main text) and 13 (the mean effective cardinalty of the MS data that feature in the main text).

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

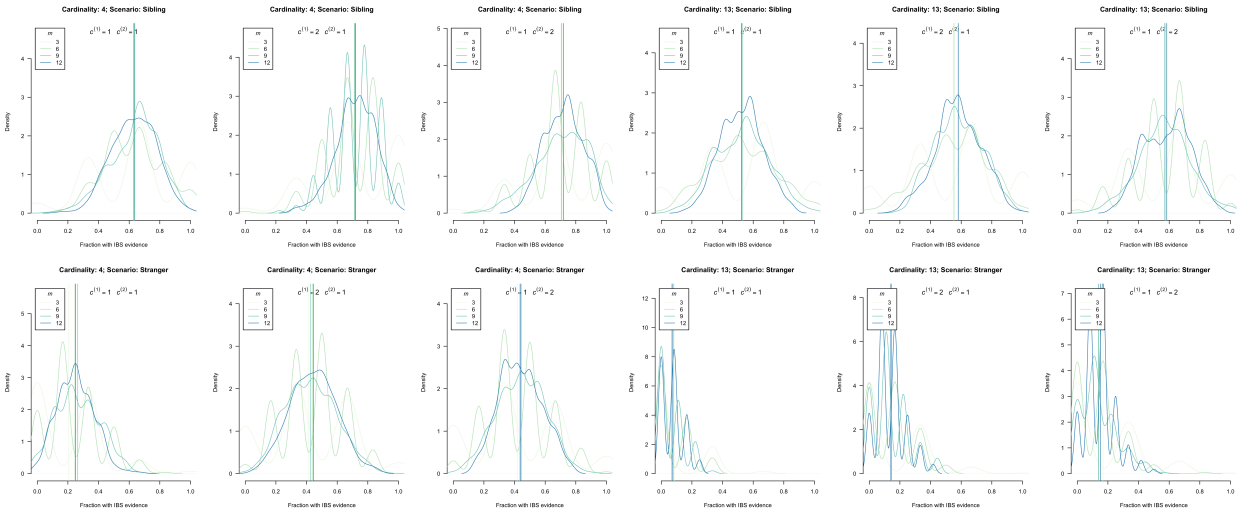
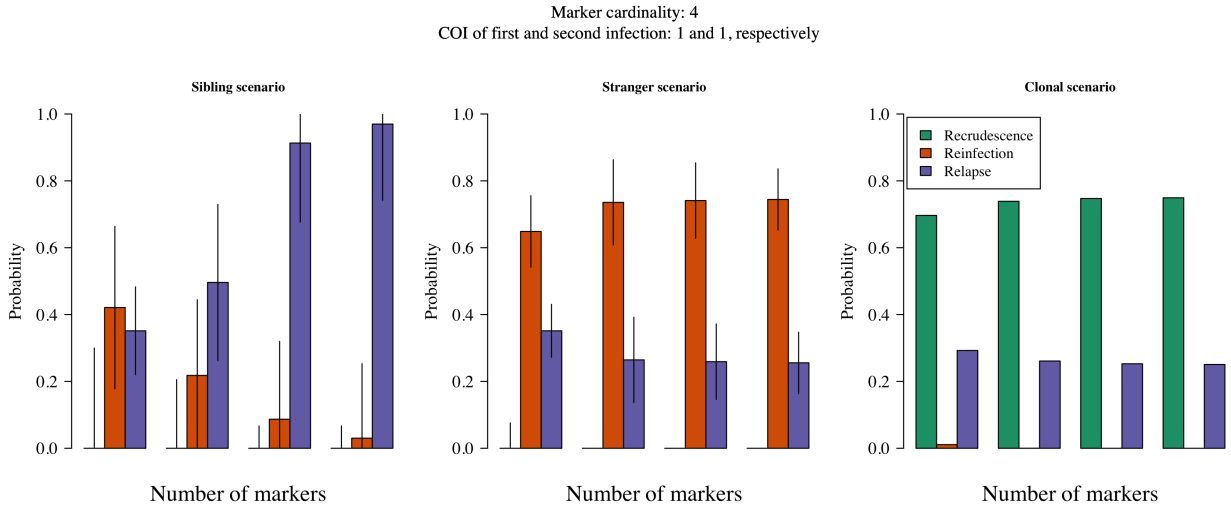
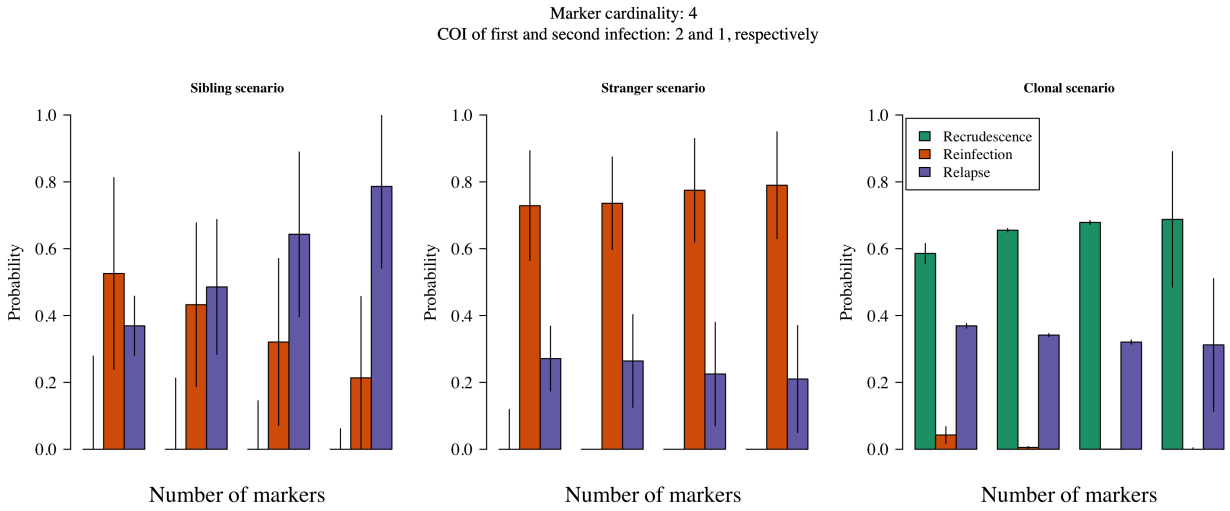
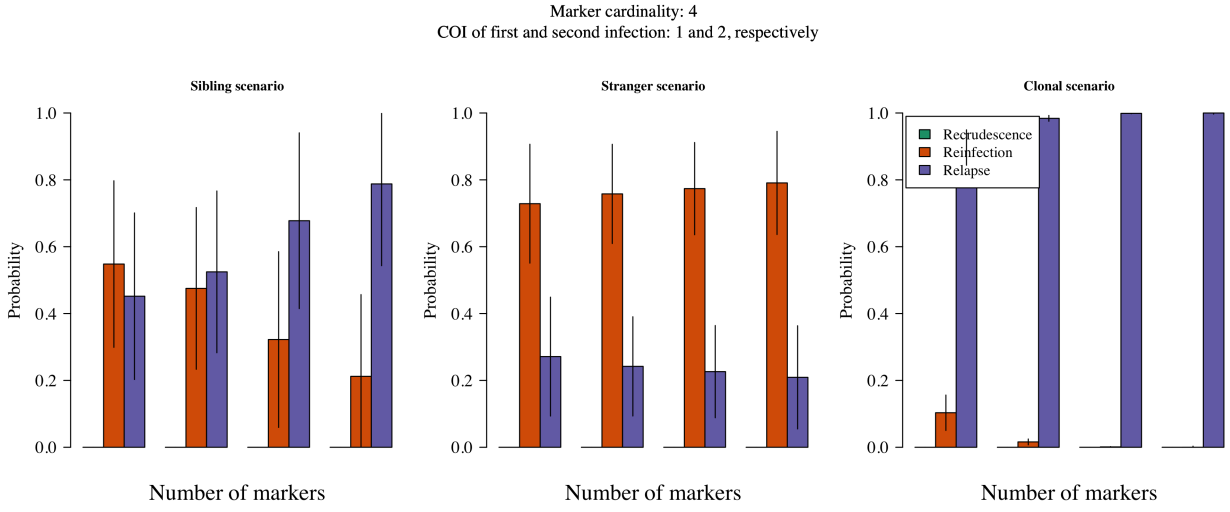


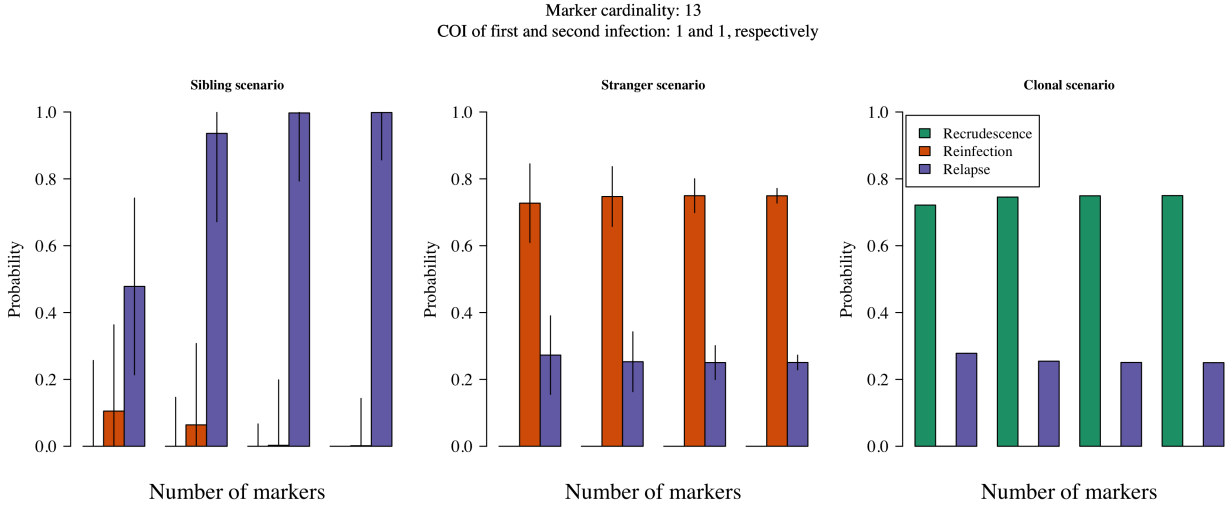
Figure 1: The fraction of markers at which evidence of IBS is detected (one or more alleles identical) when the recurrent episode is compared with the initial episode. Vertical coloured bars denote mean fractions. Different colours represent results for different numbers of markers, m.

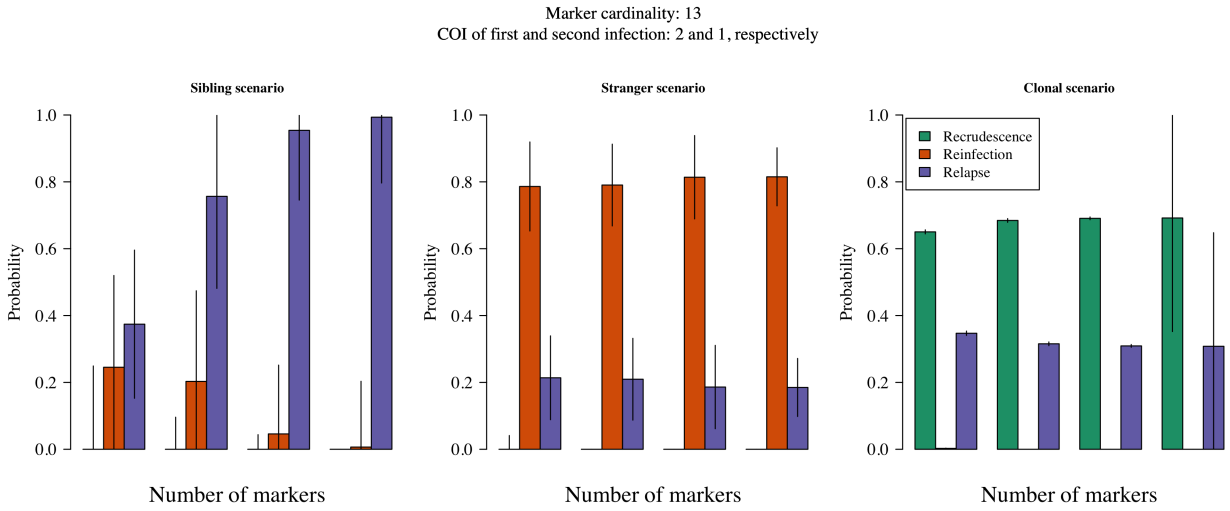
Results

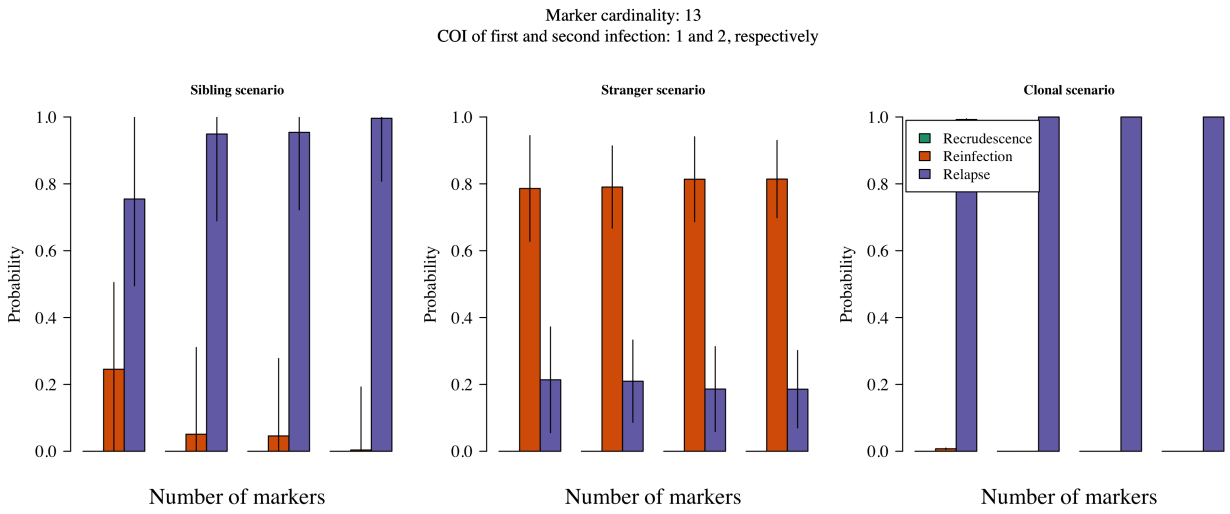



Marker cardinality: 4
COI of first and second infection: 2 and 1, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers

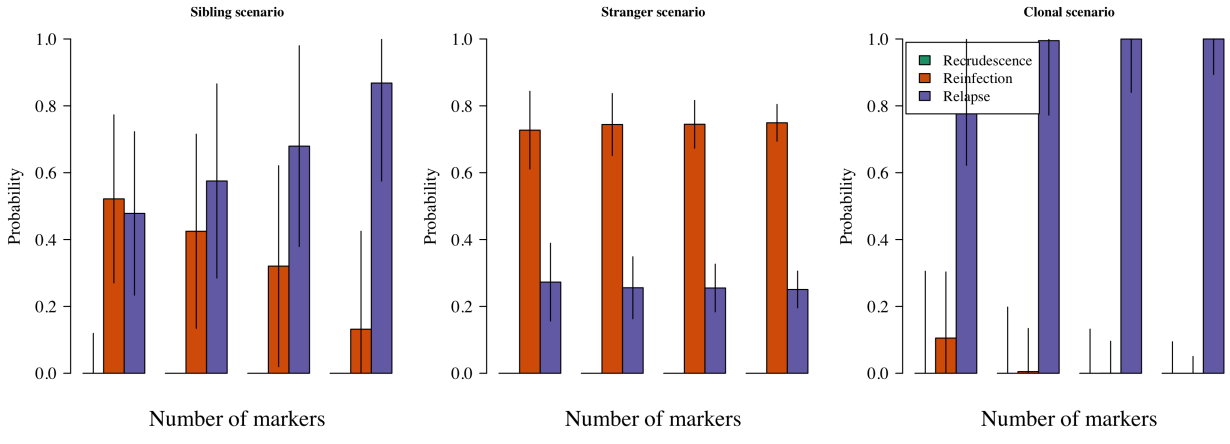

Marker cardinality: 4
COI of first and second infection: 1 and 2, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers

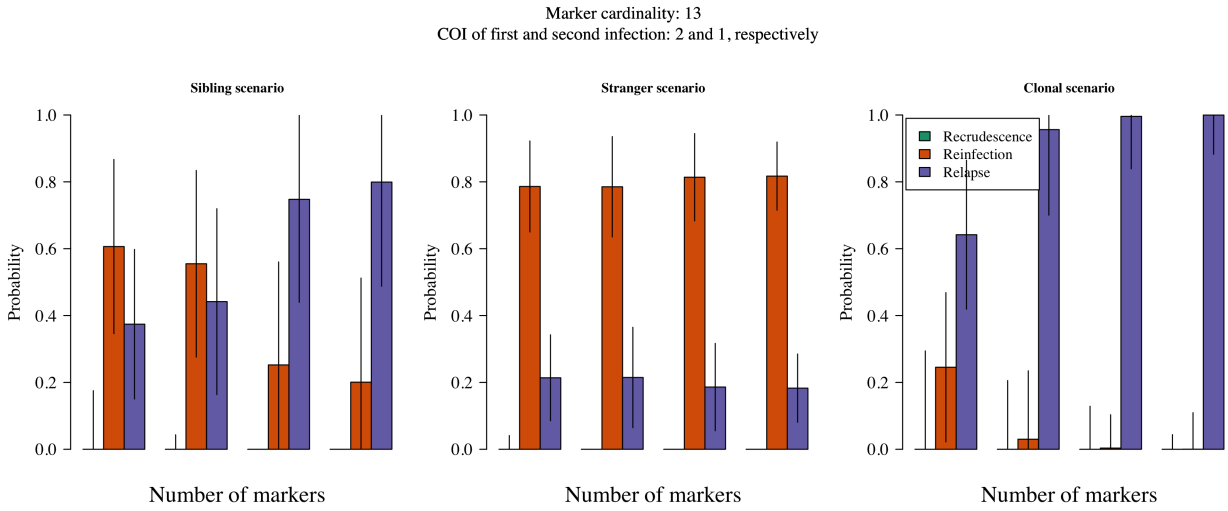

Marker cardinality: 13
COI of first and second infection: 1 and 1, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers

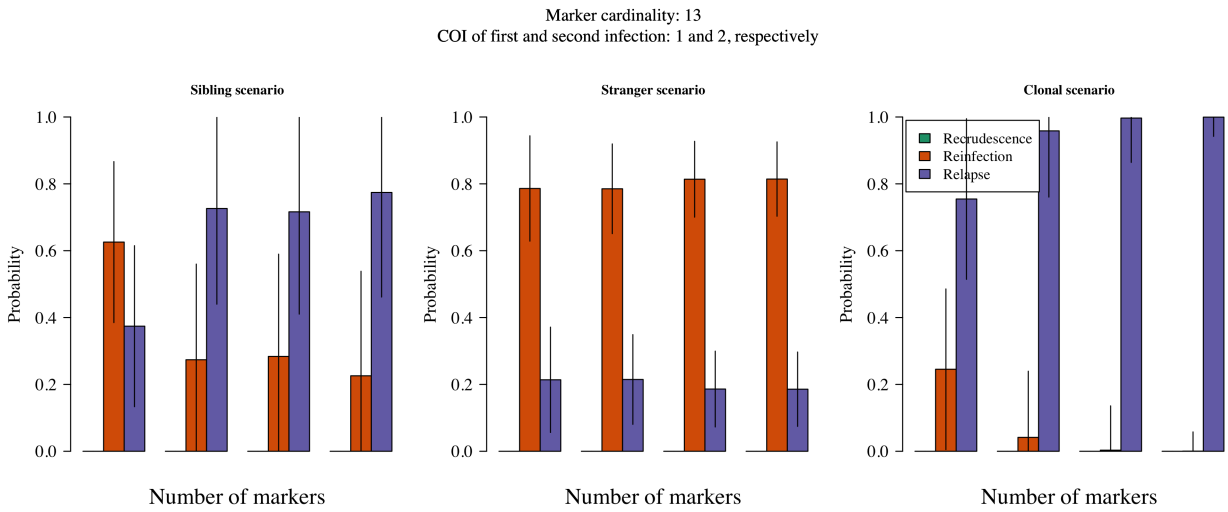


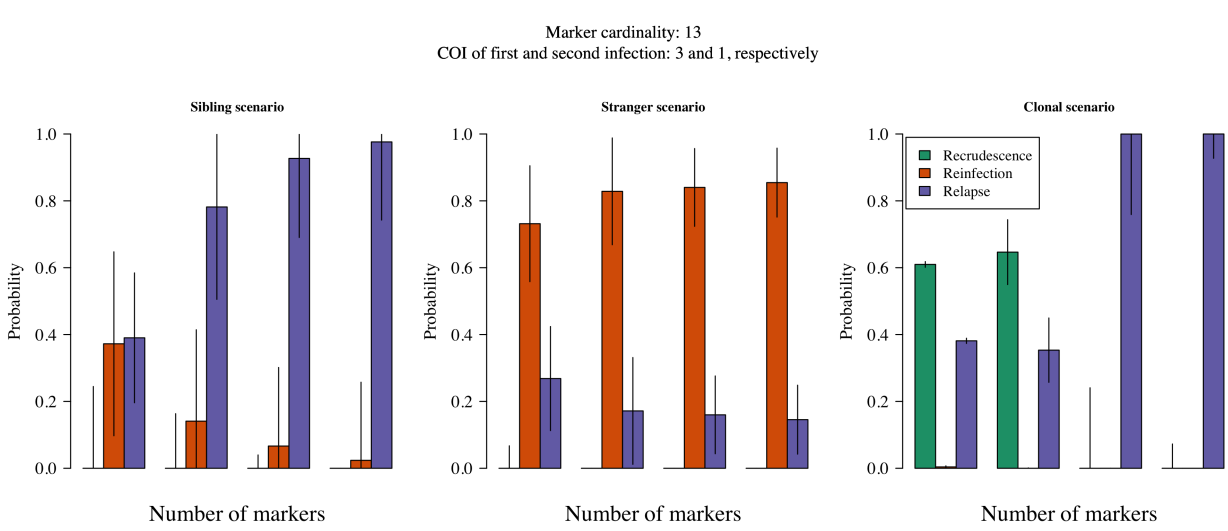



```
## 6.001 sec elapsed
```




Marker cardinality: 13
COI of first and second infection: 2 and 1, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers


Marker cardinality: 13
COI of first and second infection: 1 and 2, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers


Marker cardinality: 13
COI of first and second infection: 3 and 1, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers

Inference as a function of the number of markers typed

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, say. First, let's consider the failure to detect minority clones, second let's consider the impact of error.

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 with others across episodes. For example, referring to each plot in Figure XXX as an illustrative scenario where COI I II denotes a COI of I in the first infection and a COI of II in the second infection,

- in the Sibling COI 2 1 case, failure to detect the stranger parasite will result in the Sibling COI 1 1 case, thereby increasing probability relapsing; meanwhile, failure to detect the sibling parasite will result in the Stranger COI 1 1 case, thereby decreasing probability of relpase, 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 probability of relapse.
- failure to detect either stranger parasite in the Clone COI 2 1 case will result in the Stranger COI 1 1 case, maintaining probability of reinfection and relapse.
- In the Clone COI 2 1 case, failure to detect the stranger parasite will result in the Clone COI 1 1 case, thereby maintaining probability of recrudescence and relapse; meanwhile, failure to detect the clonal parasite will result in the Stranger COI 1 1 case, thereby erasing probability of recrudescence and replacing it with probability of reinfection.

The examples above illustrate the robust versus frail nature of relapse versus recrudescene inference under the model. Relapse infernce is also robust in the prescence of error, whereas recrudescence is not.

Erroneous data

Figure XXX shows inference in the presence of unmodelled error. The probability of error, 0.2, was set extremely high to clearly illustrate model behaviour. Realistic error rates, XXX-XXX, 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 of which is described below. When samples are highly complex, e.g. when they contain majority unrelated parasites, unconverged probablistic phasing is liable to miss clonally compatible combinations among the vast number of combinations that are possible. The number of possible combinations grows exponentially with the number of markers genotyped, rendering probability estimates inconsistent for recrudescence (Figure **??**). Such highly complex scenarios are extreme. They are helpful for illustrating the problem (Figure **??**), but not representative of the VHX and BPD data: all those analysed probablistic phasing converged. Otherwise stated, inconsistency is not a problem VHX and BPD data anlysed under the model.

Nine individuals from the VHX and BPD datasets were deemed to have data too complex to analyse under the model. They received drugs XXX. All nine individuls appear to have had multoiple relapses, based on data vizualisation, which can be used to rapidly identify clonally compatible phases, where computational methods fail (Figure XXX).

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 ommissions. Before application to data from a region where *P. vivax* resistance is suspected, the model merits extension.