Simulation study to test vivax genetic relatedness model

In this scr

Most relationship graphs most relation graphs compatible with relapse contain sibling. A single parasite haploid among majority unrelated

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."

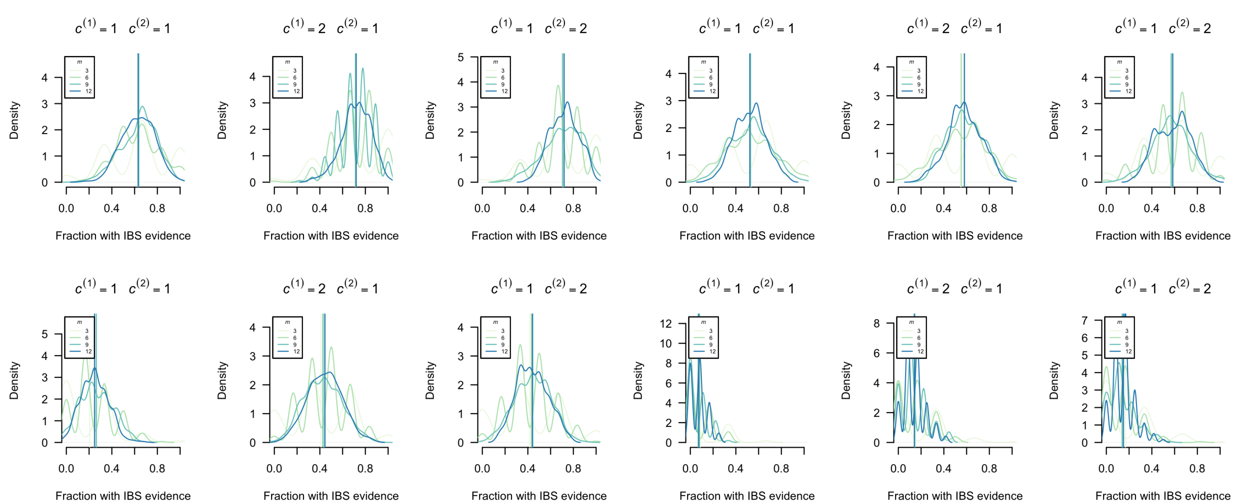
Simulation 1: Effective Complexity of Infection

We want to assess recurrence state inference as a function of the number of markers typed, adding extra noisy parasites into an infections with COI > 1. Outline of simulation is as follows: for each "job",

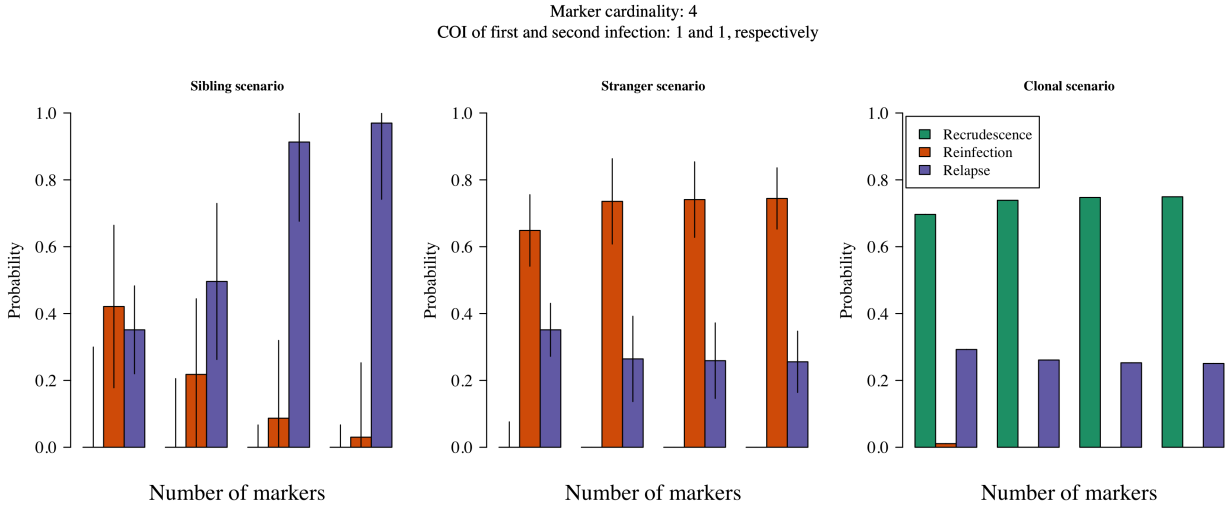
- Simulate data for N individuals, with M markers for two episodes, the second including a clonal, sibling or stranger parasite.
- Summarise the simulated data with a series of plots.
- Compute resulting recurrence state estimates (this is currently done in a separate file)
- Plot resulting recurrence state estimates as a function m

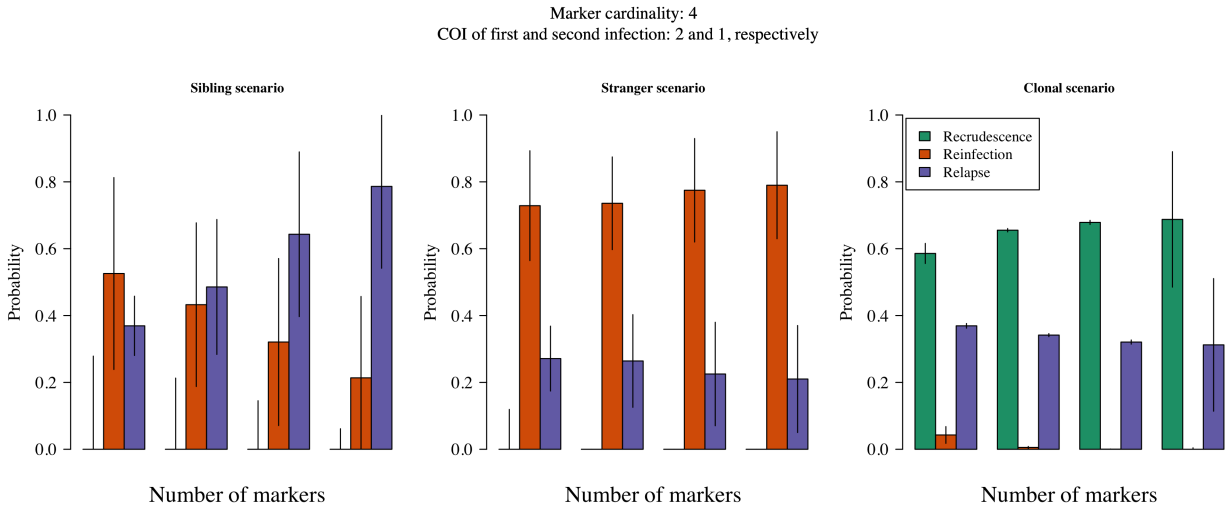
Note that when we previously specified a cut off for the number of heterolallelic calls (K_poly_markers < M), we were unwittingly amplifing evidence for relapse because when `M <= K_poly_markers` the noisy parasite will be a stranger in relation to the other parasites in the same infection; when `K_poly_markers approx 0.5*M` the noisy parasite will be more like a sibling of the other parasites in the same infection; when `K_poly_markers << M`, the noisy parasite will approach a clone of the other parasites in the same infection, but will be considered a sibling under the model.

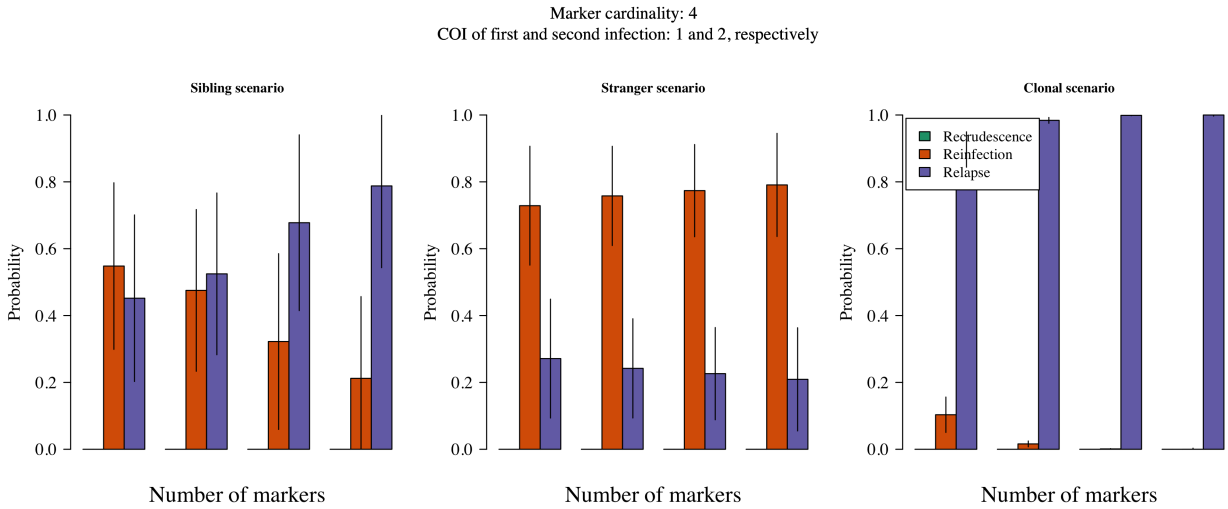
```
## [1] 1
```

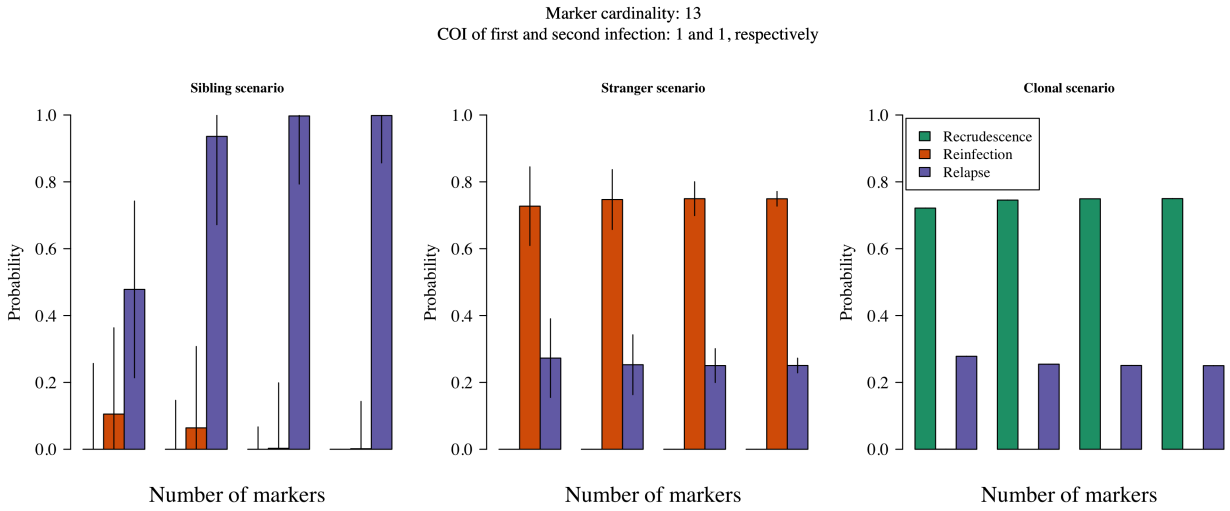


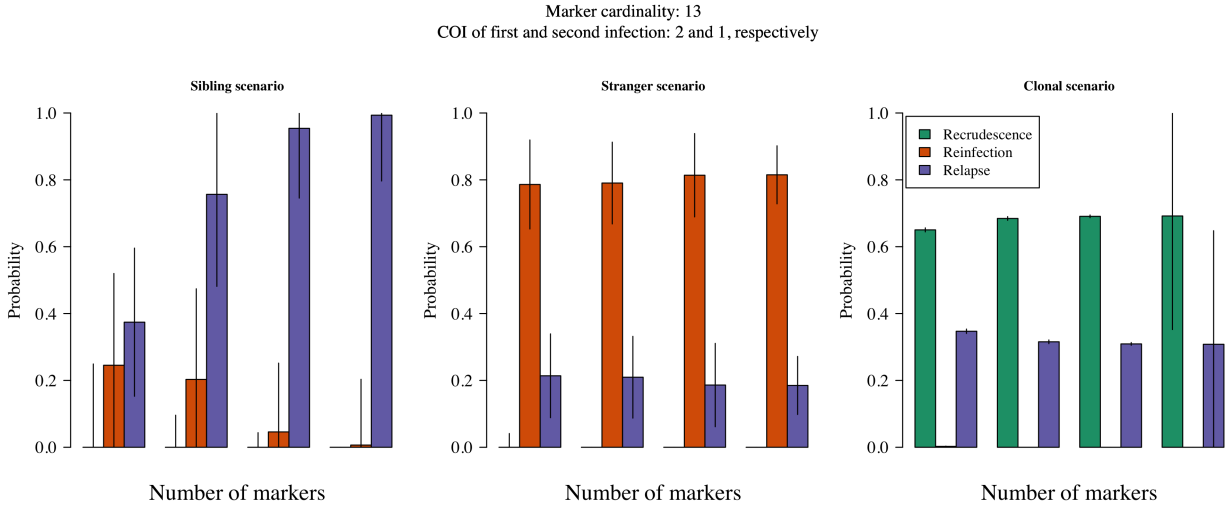
Results

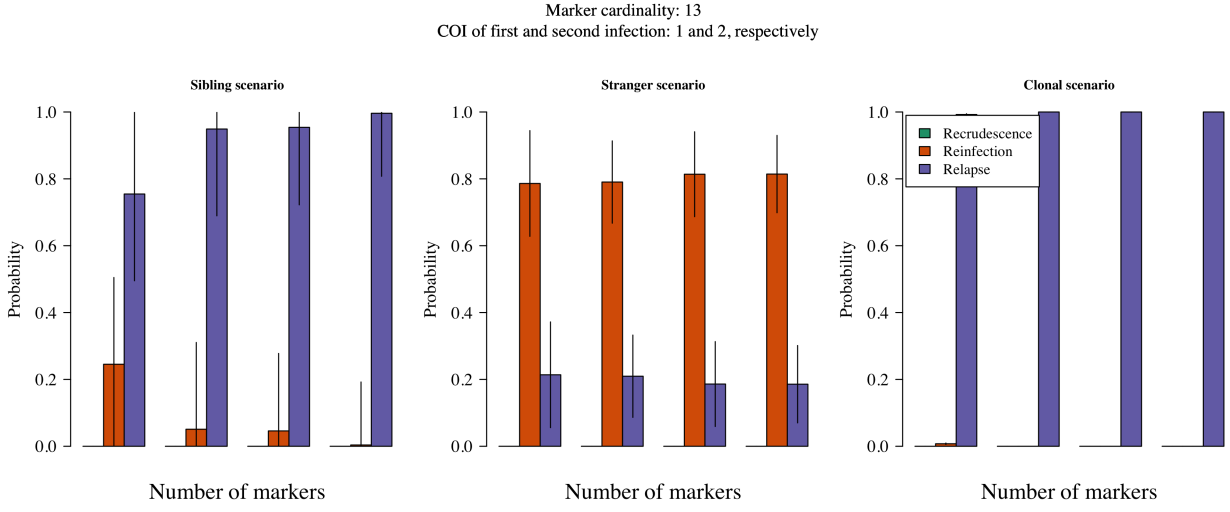

Marker cardinality: 4
COI of first and second infection: 1 and 1, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers


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

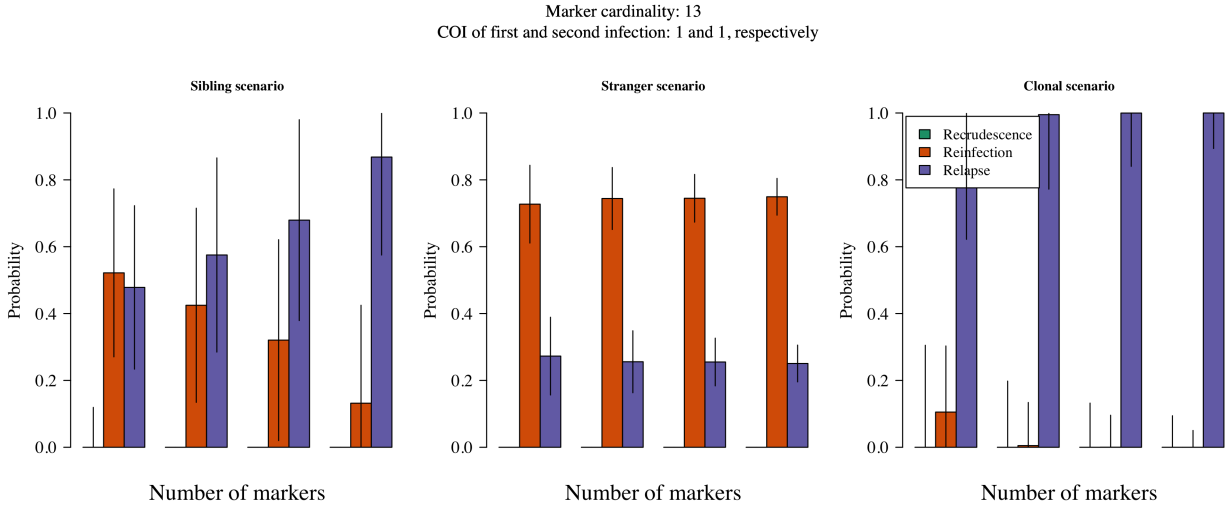

Marker cardinality: 4
COI of first and second infection: 1 and 2, respectively
Sibling scenario
Stranger scenario
Clonal scenario
Recrudescence
Reinfection
Relapse
Probability
Number of markers

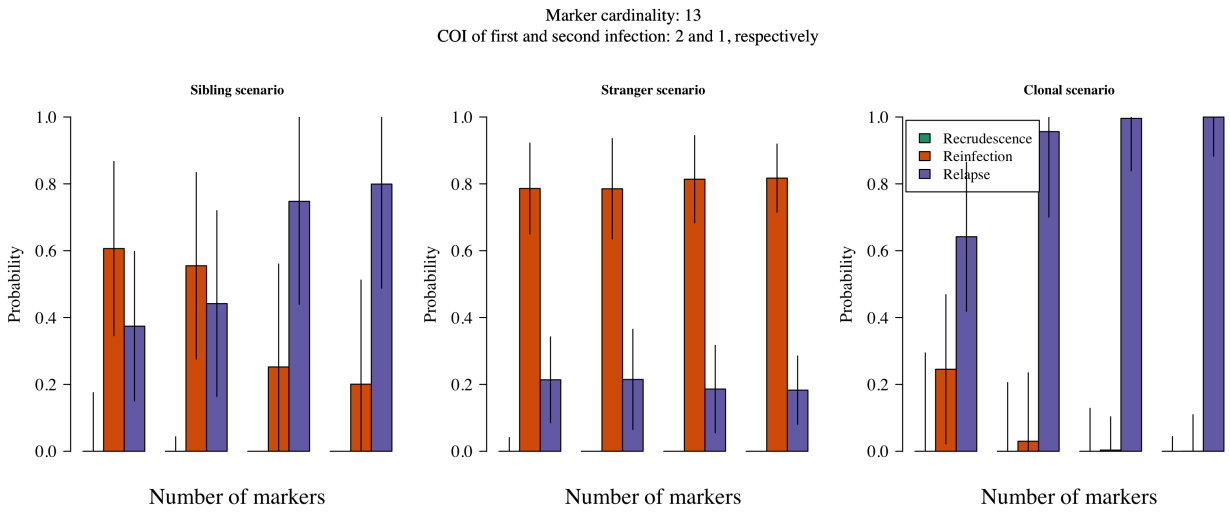


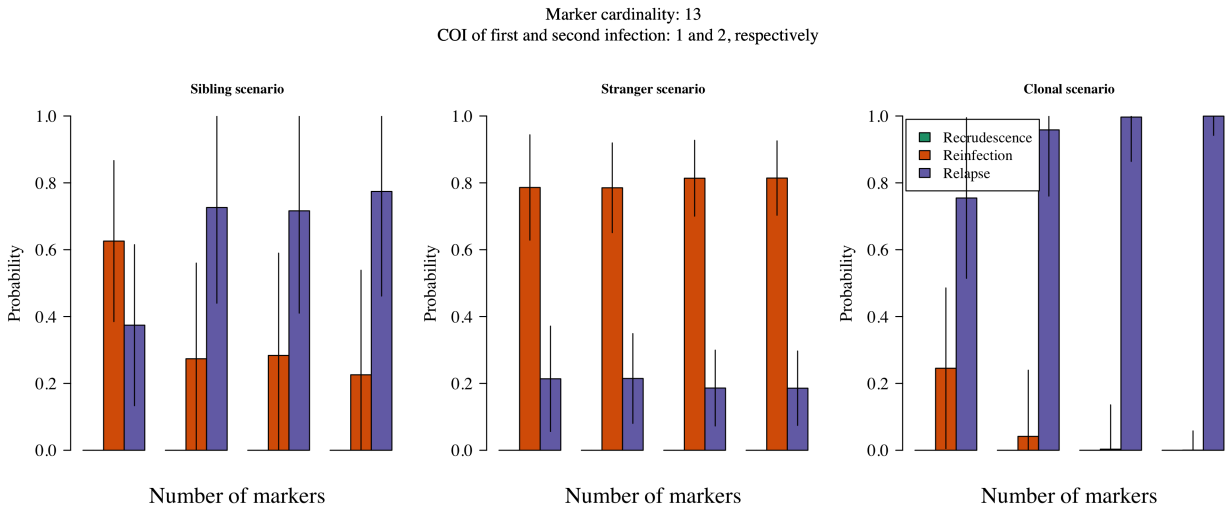





```
## 1.42 sec elapsed
```








```
## 0.815 sec elapsed
```

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

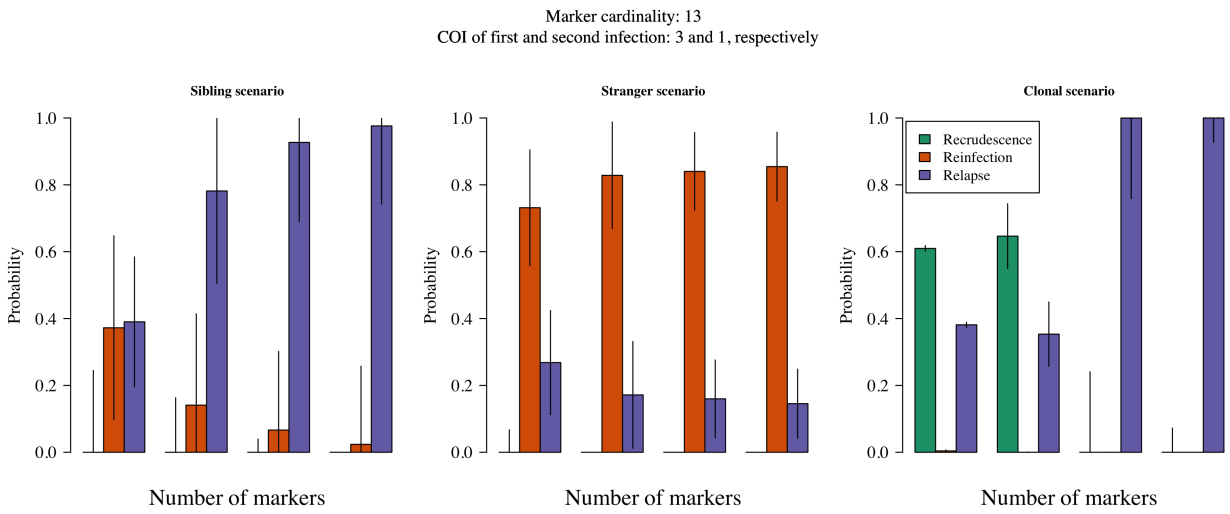


Figure 1: The probability of recurrent states as a function of the number of markers typed in a sibling, stranger and clonal scenario when the COI of the initial infection is three and the COI of the recurrent infection is one.

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. 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.