## **COI distribution in PolySimIBD model**

Bob Verity, 14 Aug 2022

The PolySimIBD model assumes that the number of circulating infections is independent per individual and follows a zero-truncated distribution. We can justify this from more detailed epidemiological modelling if needed, i.e. with a constant force of infection and constant clearance rate *per infection* we would obtain a Poisson distribution at equilibrium.

However, the number of circulating infections is not the same as the complexity of infection (COI), defined as the number of unique circulating genotypes, because multiple infections can introduce the same genotype. As an extreme example, imagine that an infected individual at time is monoclonal, and that an infected individual at time is created by drawing 5 times from this source host. Even though there are 5 distinct infections, the COI is 1 by definition because all genotypes are the same.

This begs the question – what is the equilibrium distribution of COI under the PolySimIBD model? This is a tricky problem, because strictly speaking the equilibrium distribution for any finite population in the absence of mutation is COI=1. Each generation there is a finite (albeit small) chance that all individuals will end up monoclonal for the same genotype. This is true from any genotypic configuration, and once in this state there is zero possibility of returning to any other state. This proves that this is an absorbing state, and hence this is where we will end up after an infinite number of generations.

But there is also a “quasi-equilibrium” between the diversity-generating force of recombination and the diversity-draining force of coalescence at the individual level. The quasi-equilibrium will be reached quickly and will remain approximately stable for a long period. We can derive this level by assuming an infinite population size.

In the PolySimIBD model we conceive of a main source host at time for the infections at time . The probability of an infection coming from the source host is , and the probability of coming from a non-source host is . The number of infections at time is given by . Actually, strictly speaking we assume a zero-truncated Poisson distribution as we are only focusing on infected individuals, but here I will assume the full Poisson distribution and then filter out COI=0 individuals at the end. Using the properties of Poisson and Bernoulli random variables, we can split this into the infections from the source and from non-source as follows:

We will assume an infinite population size, meaning any infection from non-source will be a previously unseen genotype. The question is, how many of are unique genotypes? Recall that these are all descended from the source host at time , which we assume has COI equal to . Each infection has two parents within this host. We assume that if parents are different then this represents a unique strain. This is true even if two infections have the same set of parents; for example, if one infection has parents A and B and a second infection also has parents A and B then these will result in different genotypes because recombination will combine material in different ways. So, the only way of being identical genotypes is if they are passed on clonally from the source. The chance of being clonal (choosing the same parent twice) is . This allows us to *further* split our Poisson probabilities as follows:

Next, we need to know the number of unique genotypes that result from the clonal infections. Each of the infections will be a categorical draw from one of the genotypes with equal probability. The total distribution of infections will therefore be multinomial, and what we are really interested in is the number of non-zero classes in a multinomial draw. This is given by the following theory:

When items are taken with replacement from a sample of size , the probability of obtaining unique original items is given by:

where is the Stirling number of the second kind, and is the falling factorial, defined as

Note that when no items are taken () the probability distribution has mass one when , and mass zero otherwise, as we would expect.

We can use this to write down the probability of seeing distinct clonal genotypes when making draws from the set of options. We can write this probability (in rather cumbersome notation):

Finally, we can start summing up to get the final distribution we want. Summing over the Poisson distribution on :

We need to convolve this with the distribution of all other genetically distinct infections, which follows:

Let this probability distribution be written . Then the final COI distribution at time can be obtained by convolution:

We can use this to derive a probability matrix of transitioning from any to any . At this point, we want to drop any probability of transitioning to , and normalise the remaining probabilities as needed, to ensure we are focusing on a constant population of infected individuals only. Once we have this matrix, we can multiply any COI distribution by this matrix to see how we expect it to change from one generation to the next. We can also compute Eigenvalues to establish the equilibrium distribution.