

Review : Inference of the worldwide invasion routes of the pinewood nematode *Bursaphelenchus xylophilus* using approximate Bayesian computation analysis by Mallez et al.

The invasion history of the pinewood nematode in Europe and Asia remaining unresolved through the use of descriptive population genetics analyses, Mallez et al. undertook approximate Bayesian computation analyses to clarify it. To this aim, the authors analysed samples from the USA representing the native area, and for the invaded area from Japan, China and Portugal. Most of the samples analysed in this paper had already been analysed in a previous paper from the same authors, which had the same objective (Mallez et al. 2015): of the 48 sites (consisting of 1080 individuals), 14 locations were new representing 310 individuals (< 1/3). It should be stated that the study was to complete the sampling of a previous study, and to use a methodology more adapted here. Since you add new locations, it is a pity not to have samples from Canada and western USA, especially as it seems the origin of the invasive populations has not been captured here.

Mallez et al. identified multiple introductions into Japan from USA and an unknown origin, and into China from USA and Japan. Authors claim that the Portuguese populations originated from America. Maybe this is due to misinterpretations/mistakes from my part but I do have issues/interrogations on the ABC analyses, the steps of the procedure and the choice of some models, which I will detail below. The authors also observed discrepancies between the different methods they used, and a lack of power. They hypothesized it could be due to the extremely low level of genetic diversity within the pinewood nematode populations. This can be tested using simulations, as they mention it lines 516-522. I agree and think it should be included in the study to strengthen the results, and justify the paper. Implementing such simulations to test the effect of a low level of genetic diversity on ABC model choice analyses would be a very good addition to the paper to increase value to it.

Material & methods:

Sampling & genotyping: I found the sampling and genotyping section too succinct, although I could find some answers in the previous paper by Mallez et al. 2015.

First, the previously analysed and new sites should be indicated more precisely. It would be nice to have a map of the sampling locations, to get an idea of the sampling scales within each area.

I could not find the years of the samplings. Have the samples been collected at the same or different time? Could you give the information, for example in the table 1 if there are different dates of collection or in the text?

The sampling scheme should be a little bit more detailed, or at least should refer to Mallez et al. (2015; however the access to the paper does not seem to be free, so I would suggest to give details here). If I understood correctly from Mallez et al. 2015, one site sample corresponds to individuals collected from one tree. Has anyone already studied the pinewood nematode population genetics at a local scale using a hierarchical sampling to see how much of the diversity is captured in a single tree relatively to a local scale?

Lines 208-214: I understand from the results section that the influence of the number of generations per year has been tested for the analysis on the origin of the European invasion (with the Japan unit 1) only. If true, it should be indicated here.

Evolutionary relationships between the different site samples:

Both F_{ST} and D_{EST} are computed to estimate population differentiation. If you use the two measures, the explanation on why and the interpretation of both measures should be given. Otherwise one cannot understand why both measures are considered.

Structure analyses: I would first talk about the global analysis, with all the samples considered, and then present the analyses at the country scale. I think it is easier to visualize like this. What about an analysis that would include all the invasive populations to explore relationships between those populations? Is there any specific reason for which it has not been performed? Or eventually, the analysis of the American samples could be used to select a few sites representatives of the America to include in an analysis with the invasive samples?

Lombaert et al. (2014) recommend using several clustering methods to define the genetic clusters in a sampling. Especially given the HW disequilibrium observed in many populations, don't you think that it would be a good idea to use at least one complementary clustering method?

ABC analyses: I am not sure to understand the step-by-step ABC analysis, and the choice of some of the models. The choice of the scenarios to test should be more detailed and explained. Based on the historical records, the first analysis should focus on Japan, and then the Chinese genetic units should be included in the model, and then the Portuguese one. Why China is first treated independently (a detail here: Japan is the first introduction so Fig. S1 should be the figure of the Japan scenarios and Fig. S2 that of the China ones...)? Then, I don't understand the different scenarios considered in Fig.S3 and S4 (why those scenarios, and why in that stepwise manner)... For an easier reading, it could be helpful to have a more precise description of the scenarios tested. There is a hypothesis for admixture for the Japanese genetic unit 2 in Fig.S4, which comes out of the blue: unless I missed it, it has never been mentioned anywhere. I guess it comes from the results of the Structure analyses. However, if we rely on historical records, which is done here and allow the stepwise procedure, Japan was first invaded, then China. I may have missed something and/or be wrong but could it be the Japanese unit that originated from an admixture event (e.g. between a native population and a Japanese unit close to Kosa –from Fig.3D in Mallez et al. 2015), and then the Chinese genetic unit 2 originated from that Japanese unit...? Why only two Japanese genetic units are considered, and what are these two units? The figures 2 and 3 in this paper and figures 2B and S1B seem to indicate at least 3 genetic units (Jap212, 120 and 308 / Kosa, Kasuming 2 and 5 / Kasuming3), don't they? From the analysis of the whole dataset (Fig.3 and lines 318-327), there is one cluster grouping together Portuguese samples, the Japanese sample Jap212 and Laoshan. Although Jap212 and Laoshan seem to belong to the same genetic cluster, and therefore the scenario 4 from Fig.1 could be a test of a Japan origin from the Iwate area, don't you think it could be worth testing a Japanese origin considering Jap212 for the Japanese sample? Also, there is no hypothesis of the Portuguese samples originating from an admixture event. Don't you think it could be possible?

To come back to the methods, lines 229-231: the F_{ST} and $d\mu^2$ distances between pairs of populations have not been considered in the ABC analyses but in the quality checking is that right? If it is, why so?

Results:

Table S2: This is a very large table, difficult to read. Could you use a gradient colour to fill the cells according to their value?

Even if it has been done in the previous analyses, I think it would be nice to still add a figure or at least mention the results of the structure analysis of the Japanese samples. I would not insist too much on the results of Structure. I am not convinced by the consistency between runs.

Figure 2: would it be possible to add to the colour code the results of the genetic clustering using Structure, i.e. to use two colours for the American samples, 2 for the Chinese samples, and 2 if 2 it is for the Japanese ones?

Lines 343-349: maybe I missed it but I did not see a clear description of the genetic units considered in the ABC analyses. It could help to add the list of samples each genetic unit is the representative of. You could also highlight the samples used in the ABC analyses as representatives of their genetic cluster in the figures 2 and/or 3, for example using an asterisk or an arrow. That would help the understanding.

I will not discuss the results of ABC analyses too much. Do you have an explanation for the different results you obtained according to the prior set for the analyses between China 2 and Japan 2? Are the dataset correctly simulated using both prior sets?

Tables 2/3 and others: for the "Analysis and scenario" column, highlight in bold the selected scenario rather than the prior set/nb of generations per year considered. I think you could reorganize the tables for clarity/easiest visualization. For example for the table 2, you could re-organize the results according to the prior set by columns and not lines, and include the mention of the number of generations in the legend rather than in the table for each analysis since there is only one option considered here. There is no need to indicate in the legend the method for the posterior probabilities, it's written in the material & methods section.

Lines 382: all invaded areas, except Japan2.

Discussion:

Even the closest native samples appeared to be quite differentiated from the invaded ones. Don't you think that it could also be a reason of the disagreement between the analyses and of the "mixed" results?