Dear Editor,

We have now revised our manuscript according to the latest suggestions and comments of the reviewer and recommender.

We would like to thank the reviewer and recommender for theses last comments which helped further clarifying the message and reinforced the results.

We hope this version of the manuscript will be suitable for recommendation.

Sincerely

Round #2

Author's Reply:

Decision

by Nicolas Galtier, 2019-06-25 13:38
Manuscript: https://doi.org/10.1101/362129 version May 05, 2019.

Decision on: "Parallel adaptations...", by Koutsovoulos et al.

I concur with the reviewer that the manuscript has been substantially improved. The scope of the study has broadened, and I find the overall message clear and compelling. The analysis of coverage, the distinction between so-called "heterozygous" and "homozygous" variation and the linkage disequilibrium analysis are important, informative additions. The title was appropriately amended and reflects, I think, the more ambitious nature of the study.

We thank the recommender for his encouraging comments and we are glad this new version of the manuscript is now indeed more ambitious and comprehensive thanks to the different comments and suggestions of the recommender and reviewers.

The reviewer has a couple of comments, which deserve to be considered.
First, the reviewer suggests analysing the variation between homeologous regions within a sample (major comments 1 to 4), when the text currently focuses on the between-samples ("haploid") variation. This would be a really nice addition, if possible. I am not sure, however, that separating true variants (between homeologs) from spurious variants (due to assembly/duplication issues) based on coverage is easy to do in this case - figure S1 suggests that the coverage distributions of the two categories of variants overlap quite a bit. Please let us know what you think is doable here.

This is an interesting comment. We would like to take this occasion to clarify what is our current understanding of the genome structure of *M. incognita*. In (Blanc-Matthieu et al. 2017, PLoS Genetics) we showed that the genome is most likely triploid with three copies A, B and C equally highly diverged one another. Most of the homeologous A, B and C copies have been correctly separated during the assembly because of their high divergence. This is supported by the genome assembly size matching the estimated genome size by flow cytometry and the fact that genes in 3 copies represent the highest category of genes. According to per base-coverage (like in figure S2), a little proportion of the genome shows a peak at twice the coverage of the rest. We interpreted this as cases where two of the three A, B an C copies had lower divergence and were merged during the assembly.
However, the distribution of single and double coverages partially overlap and the overlap level varies from one isolate to a single isolate (figures S1-S2). Furthermore, the peaks themselves are not exactly at the same coverage values for each isolate. Because of this, it is particularly tricky (if not impossible) to define an absolute threshold to differentiate accurately putative collapsed regions from the rest of the genome. We think it is safer to discard all the markers that vary within at least one isolate to make sure we do not include artifactual variants due to the collapsed regions. We understand doing so, some analyses cannot be conducted and we now have clarified this in the new version of the manuscript.
We would like to clarify that our goal was not investigate the intra-isolate variations analysis and we would like not to push too far this kind of analysis due to uncertainty on 0/1 SNVs. Our main objective in this paper was rather to study inter-isolate variations in relation to several biological traits and only keep the safest markers for these analyses. The intra-isolate analyses, although interesting are a bit out of the scope of this paper and should be kept for a future version of the genome hopefully fully resolving the few collapsed regions.

The other important comment made by the reviewer (major comments 5-6) is that the population genetic analyses have been done in an unusual way, i.e., by comparing each sample to the reference. This has an unclear meaning, which depends on how the reference was generated (single individual? pool of individuals? from which origin?).
The reviewer rather suggests analysing multiple alignments across the newly sequenced strains, which indeed should provide more reliable estimates of, particularly, piN, piS and their ratio. This is clearly a sensible recommendation, which I think should be followed.
We agree with these suggestions. To clarify, as explained in the introduction, the reference genome comes from Mexico and it was sequenced form a pool of individuals reared from the egg mass (offspring) of one single female. This is exactly the same kind of rearing and pooling than for the 11 isolates from Brazil analyzed in this paper.
We followed the recommendation of the reviewer and now performed the piN, piS and ratio analyses based on a multiple sequence alignment of coding sequences. We have updated the manuscript.

I have a related, minor comment: for the same reason, I find the "homozygous SNP" vs "heterozygous SNP" terminology quite misleading. A SNP is normally a position in a genome at which between-individual variation has been detected - i.e., a variable colon in a within-species

alignment, or a vector of genotypes. Such a vector should not be qualified as homozygous or heterozygous.
We agree that the adjectives homozygous and heterozygous used to describe our SNP dataset could be misleading. We thus clarified and explained we have only worked on variants that were fixed within an isolate (= what we called homozygous). Variations within an isolate (=what we called heterozygous) were ignored because it is almost impossible to differentiate true variations between individuals within an isolate from artefacts due to collapsed homeologous regions of lower divergence (see also answer to the first comment). We have removed the adjectives homozygous and heterozygous SNP and clarified we have worked on SNV fixed per isolate.

Furthermore, because the M. incognita genome is haploid, one would not expect to find any heterozygous genotype at all.
Even though the *M. incognita* genome is effectively haploid, there can be some variations between individuals within an isolate. Indeed, we have to keep in mind that we sequenced pools of >10.E6 individuals. Although these individuals originate from the offspring of one single female, they underwent several cycles of reproduction and thus, at the end there can be some variations between the collected individuals. These variations between individuals within an isolate can be responsible for part of the 0/1 SNPs (the other part being due to the collapsed homeologous regions of lower divergence). Again, because it was particularly tricky if not impossible to safely differentiate 0/1 SNP representing inter-individual variations from collapsed homeologs we discarded them all from the analysis.

Yet, because the authors have applied a variant calling method that assumes diploidy, and because of assembly/duplication errors or partial di/triploidy, a large number of apparently heterozygous variants were called. I would suggest refraining from calling a SNP "homozygous" or "heterozygous", and using "variant" rather than "SNP" when referring to a genotype predicted by the variant caller. The word "SNP" should be restricted to the new analysis suggested by the reviewer, when sequences from the distinct strains have been multiply aligned.
We followed this recommendation and used SNV and removed the terminology homozygous / heterozygous.

I would suggest following these very last suggestions, which I think should help improve this excellent manuscript even further.
We followed these interesting suggestions which indeed further clarifies and reinforced the message and hope the manuscript can now be recommended.

Reviews

Reviewed by anonymous reviewer, 2019-06-13 21:59
In this revised version the authors have done several additional analyses and have extensively rewritten the manuscript. The new results strengthen the manuscript and broaden its interest. However, some problems remain about data analysis, especially for polymorphism analysis, which were not performed properly if my understanding is correct. They should be easily corrected but some results may change.

Major comments

- The title of the first paragraph of the results is "…the genome is mostly haploid…". If the species is triploid due to hybridization it means that two sets of chromosomes should pair whereas the third one should be alone. If I understood correctly, 80% of SNPs were heterozygotes. Does it mean that they correspond to the diploid pair and to the Meselson effect between the two chromosomes?

Yes, the genome is triploid as a result of hybridization events and is effectively haploid. So far, the most likely hypothesis is two rounds of hybridization, the first one having given an intermediate homoploid hybrid (A,B) and the second round a triploid one (A,B,C). Indeed, most of the genes (>90%) are in multiple copies with a peak at 3 copies (Blanc-Mathieu et al. 2017, PLoS Genetics). This observation in combination to a genome assembly size that matches the estimated genome size via flow cytometry shows that most of the 3 homeologous genome copies have been correctly separated during the assembly. This correct separation is mainly because of the equally high nucleotide divergence (8% on average) of the A, B and C copies and the duplication-aware assembly procedure we used. However, as shown in figure S2, a minor proportion of the genome shows a peak of per-base coverage at twice the coverage of the rest of the genome. These regions of double coverage most likely represent rare cases where divergence between two of the three A, B and C copies was sufficiently low to be collapsed during the assembly. Then because reads from the two copies have been mapped to a single collapsed regions, artefactual 0/1 SNPs where called there.

If the triploidy was due to an unreduced gamete (A,A') hybridized with a haploid gamete from another species (B), then we would expect most of the genes in two copies (one copy being the results of the two collapsed highly similar (A,A') pairs of homologous chromosomes from the unreduced gamete and the other copy (B) coming from the other parent of the hybridization). Because we do observe this peak at 3 copies and because the average divergence between the three corresponding genome copies averages 8% whatever the comparison (A vs. B, A vs. C or B vs. C) (cf. Blanc-Matthieu et al. 2017), we think two successive hybridization resulting in 3 genome copies with equally high divergence is more likely.

Hence, the 0/1 heterozygous variants probably not represent accumulating divergence between the former A,A' homologous parental chromosomes, due to Meselson effect. Instead, they most

likely represent regions of lower divergence between the A, B and C homeologous chromosomes from the founders of the hybridization events.

An interesting question for the future would be to investigate how much of the 8% divergence was due to pre-existing divergence between the parental donor species and how much of this was acquired by drift and divergence after the hybridization events. Unfortunately, in the absence of the parental genomes this is so far not yet possible to investigate.

- The fact that most heterozygotes could be duplicates because of the doubling of the coverage is convincing. But in that case, the two parts of the genome should be split based on coverage. This would allow to analyse separately the haploid and the diploid genome. This would clearly help to better understand the reproductive system of this species.

We agree that in theory, that would be ideal. However, in practice this is difficult and challenging because the peaks of single and double coverage substantially overlap. Furthermore the coverage values and level of overlaps between single and double peaks vary from one isolate to the other (figures S1-S2). It is thus not possible to define an exclusive threshold to distinguish easily regions of double coverage from regions of single coverage in all the isolates. For safety reasons, and to make sure we do not include in our analysis artifactual heterozygosity due to the few collapsed homeologous regions, we took the decision to exclude all the heterozygous variants.

Investing in long-read sequencing in the future would probably be a more effective way towards separating the few collapsed homeologous regions because long-range linkage information will probably allow to separate these regions.

- The choice of excluding heterozygote sites can be justified so it indeed prevents to compute Fis. However, it would be important to know how behaves the diploid genome. Is there an excess of heterozygotes or not. If not, or it varies along the genome it could be informative of the kind of asexuality. Modification of meiosis could also be a possibility instead of mitotic reproduction (for example see (Lenormand et al., 2016, Engelstadter, 2017)). We could imagine a form of automixis with one set of haploid chromosome transmitted as a block without segregation.

We would like to remind that at a cytological point of view, meiosis has never been observed in *M. incognita* ovaries, including at the place where meiosis is easily observed in the “meiotic” relatives like *M. hapla* (Triantaphyllou 1981; Triantaphyllou 1985).

Furthermore, as stated in response to the first comment, it is not likely that there is part of the genome that is diploid (A,A’) and a third more distant copy (B). The three copies (A,B,C) are equally distant one to another and as explained in more detail in Blanc-Mathieu et al. 2017, the

genome is most probably an allotriploid due to successive hyrbidization events. Hence I am not sure it makes a lot of sense to try to investigate excess of heteorzygosity between A, B and C.

Moreover, our main goal was to study the level of variation between isolates in relation to various different biological traits and not to study within isolate variability. By keeping only the "homozygous" fixed variants, we still have plenty of markers (>66,000) and we think we remain on the safe side of the dataset to study these inter-isolate variations. We think these questions, although interesting, are a bit out of the scope of the current manuscript, will not change the main conclusions we presented and will probably be better addressed in a future version of the genome more approaching chromosome-scale resolution.

- Related to this important question, and still if my understanding is correct, it's important to note that the lack of recombination is shown (nicely, see below) for the haploid genome (which is expected) but not for the diploid genome. Absence of recombination for the diploid genome (+ Fis <0) would be a strong argument for mitotic recombination (but still some for of automixis can lead to a similar pattern)..

We now clarified in the manuscript that the genome is triploid with A, B, C copies equally diverged. There are no almost identical diploid A, A' copies on the one hand and a more distant B genome on the other hand. The few collapsed regions are most probably caused by regions of lower divergence between the A, B and C copies that would not be separated during the assembly.

What we showed is that for a same genome copy (A, B or C), there is no evidence for recombination across the isolates. This result and others presented in our paper also collectively suggest the lack of genetic exchange between individuals.

Indeed, we did not look for evidence (or lack thereof) of recombination between A and B or A and C or B and C copies within isolates. I am not sure this study would make a lot of sense because A, B and C are not homologous chromosomes but homeologous ones resulting from hybridization and displaying a high divergence (8% on average). As a result, the genes in A, B and C are not alleles but highly diverged gene copies (actually orthologs coming from the parental species). Furthermore, the vast majority of genome reads uniquely mapped to either A, B and C copies, so variations between the A, B and C cannot be called SNPs and whether they result from pre-existing divergence prior hybridization or accumulation of mutations after hybridization remains difficult to assess in the absence of the parental genomes.

Analyzing recombination within the few collapsed regions might be possible, although it would require prior phasing of the haplotypes. Because these regions are minority, however, and the low divergence might as well be due to gene conversion, we think too many possible confounding effects preclude this kind of analysis at the moment.

- If I understood correctly, SNPs are defined as variant compared to the reference. Then polymorphism is computed for each strain as the % of SNP. But this is not a measure of

polymorphism of a strain but of the genetic distance between the strain and the reference. If only homozygote variants are kept, piS and piN cannot be computed for a strain. Here what should be done is to compute piS and piN for the whole species and also interestingly for the three clusters detected by the PCA. And when computing these statistics the reference genome should not be considered, except if the reference strain is added as a data point.

Yes, indeed, SNPs were defined as variants compared to the annotated reference genome, because all the reads have been aligned to this reference genome. We agree that although these measures are informative for the PCA and tree/network classification, it is certainly less meaningful for piS, piN and ratio.

We thus followed the recommendations of the reviewer and re-calculated all the piS, piN and ratio values based on a reconstructed multiple alignment of the coding sequences, excluding the reference genome.

- If the order of magnitude of piS and piN/piS is still valid after correction, it is interesting to note that, although piN/piS is three times higher than in outcrossing nematodes, it is still lower than many other species, including human (around 0.2).

The new piN/piS value is 0.129 and is still ~3 times higher than for the two compared outcrossing Caenorhabditis species (0.051 and 0.041). Indeed, although much higher in *M. incognita* than in outcrossing nematodes, all these piN/piS values remain relatively low compared to some animal species, including human. However, given the differences in life cycles, effective population size, number of offspring per generation etc… that all affect these values, we doubt the values between humans and nematodes can be easily compared.

- The PCA and phylogenetic tree do not support clustering by host. This could be used as an opportunity to try to identify the few SNPs (if any) that could be associated with hosts. It is not possible for all host but for those that can be found in different clusters such as soybean, cotton and tobacco.

Actually this was exactly one of our initial goals. We wanted to identify short-scale variations that could be associated with the host (and host race status as well). Finding such variants associated to compatibility with hosts could lead to the development of molecular markers to easily diagnose the host race status. Unfortunately, all we found is 0 SNV specific to isolates parasitizing cotton, only 1 synonymous SNP for those parasitizing tobacco and only 1 synonymous for soybean. This result has been added in this new version of the manuscript.

Minor comments - The test of absence of recombination is a nice addition and the idea of comparing to a recombining species is a nice control. This part could be move earlier (second or third part of results for example).

We totally agree and moved this section to the second part of the results. This allows interpreting the rest of the results in the light of this important confirmation of absence of recombination.

- It was not one of my previous comments but I'm not really convinced by the argument stating that an ancient polyphageous strain is not very likely. Here the number of strains is too low for ancestral state reconstruction. I don't mean that this idea is wrong but that the alternative proposed by the other reviewer could also be discussed and that additional data would be needed to test it properly.
Yes, we also agree that the two hypotheses at this stage are equally likely and addition of other isolates characterized for their host compatibilities might allow to favor one or the other in the future. We changed the manuscript accordingly.

References: - Engelstadter, J. 2017. Asexual but Not Clonal: Evolutionary Processes in Automictic Populations. Genetics. - Lenormand, T., Engelstadter, J., Johnston, S. E., Wijnker, E. & Haag, C. R. 2016. Evolutionary mysteries in meiosis. Philos Trans R Soc Lond B Biol Sci 371.
Thanks for these interesting references which open new perspectives to be investigated in M. incognita in the future.