The 17th of December 2021,

Dear Recommender,

Please find below our answer to the two reviewers for the manuscript deposited on BioRxiv and now entitled: “*Chromosomal rearrangements with stable repertoires of genes and transposable elements in an invasive forest-pathogenic fungus*”.

We hope that the present version is now correct for recommendation in PCI Genomics.

Sincerely yours,

Cyril Dutech

Corresponding author

Decision for round #1 : *Revision needed*

Review by anonymous reviewer, 2021-05-18 13:57

> Chromosomal rearrangements are major contributors to genome architecture and evolution but remain overlooked in fungal genomes. This is notably due to the lack of high-quality reference genomes for comparative genomics. In this study, the authors report *de novo* long-read assembly for the chestnut blight pathogen C. parasitica. The sequenced strain was sampled in the species center of origin (East Asia). The authors produced a high-quality genome assembly almost reaching the chromosome level. If the overall gene and transposable element contents are stable in comparison to the reference genome of a North-American strain, the two strains exhibit extensive chromosomal rearrangements.
> This genome report manuscript in itself is solid, the analyses support the results and interpretations.

> I have several concerns/comments to address:
> 1- Why the authors show the assembly metrics of the three strains EP155, YVO003 and ESM015 but do not include YVO003 in the comparative genomic analyses? Including the three strains could already provide some inputs regarding the emergence of asexual reproduction lineages in invasive C. parasitica populations.

> > We agree with this comment. However, the YVO003 genome assembly obtained in the previous Demené et al. (2019) study is significantly shorter than the reference EP155 genome. It was likely due to the difficulty to assemble regions of low complexity, and it precludes a rigorous comparison of TEs and gene repertoires. This point is now better underlined in the introduction of the revised manuscript (L219). We agree that having more complete assemblies would help to investigate the causes of emergence of asexual lineages. However, we need to have more assemblies than YVO003 only, and we hope that the methodology developed in this study should help to deserve this question as mentioned L1339.

2- I do not understand why the authors included failed assemblies in the main Table1 (e.g. Miniasm) and do not provide the final assembly metrics. How is it possible to reach 46Mb assembly and have 0.5% BUSCO, is it an error? A common statement is to consider assembly 'good' is to have at least 95% of completeness with BUSCO. Regarding the HybridSPAdes assembly in Table 1, the minimal (contig/scaffold) length is of 129pb which is really small and I assume prior to curation. Can the authors provide the final (after curation) metrics for the chosen assembly in the main Table1? I would suggest to remove the "working" assemblies from Table 1 (more useful for a supplemental table). That will help the reader to catch the final comparison between the three assemblies of the reference strain EP155, YVO003 and the new strain ESM015.

> We agree that the presentation of the different attempts to assembly the genome using different methodology was confusing. We removed these results from Table 1. They are now presented in supplementary material in Table S1. The final assembly of ESM015 after curation is now presented in Table 1. The assembly obtained with Miniasm with only 0.5% of BUSCO genes is an assembly without polishing with the Illumina reads. The sequencing errors associated to the Nanopore sequencing may explain this low recovery rate of BUSCO genes. This recovery rate after polishing is significantly better (Table S1).

3- Regarding RIP signatures found in the TEs, do the two strains carry the "RIP-machinery" genes such as DNA methyletransferases (e.g. N. crassa DIM2 gene, or RID gene)? This would strongly support the potential ongoing RIP activity suggested by the RIP-indexes estimations.

This points was added L1097 : “Furthermore, a homologous protein of the RID gene of *Neurosporora. Crassa* (GenBank: AAM27408.1), the only gene experimentally proven to be necessary for RIP activity (Freitag et al. 2002), was identified in *C. parasitica* genome (blastp e-value 8e-96, g9609.t1 with the same PFAM motif PF00145 known to be involved in RIP activity; Amselem et al. 2015).”

4- To improve the manuscript, I think that overall authors could emphasise the interesting findings and questions addressed in the study instead of focusing on 'negative results'. Actually, there are not really "negative results" but the first half of the results of the comparison between the two reference genomes are focusing on: "no two speed genome", 'no variation in effectors or TE repertoires' sound quite negative and kind of mask the main result: stable gene content, extensive rearrangements.

The absence of (or low) compartmentalization, or the absence of variation in effectors and TEs are not, for us, negative results, but important tests to investigate general trends in fungal genomic evolution. However, for avoiding these results are considered negative, we modify the title, the subtitles and the abstract of the manuscript in agreement with this remark.

Minor comments

5- Maybe the authors would like to add Badet et al 2020 (https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-020-0744-3) as a reference to illustrate how high quality genome assemblies of a variety of strains can unravel structural variation within species.

We added this new reference in introduction (L176)

6- L168-L192: I think this entire paragraph should be removed, I do not think it is needed to define the three generations of sequencing technologies nowadays. I would just mention that high quality assemblies are needed to explore genome architecture, and that's now possible using long-read sequencing. I would just keep the sentence L186-192, which links quite well with the sentence L162-166.

Done

7- L77-78: Is it really necessary to state that some fungi are easy to manipulate in lab conditions? I think it is mostly untrue and that's exactly why genomics are so interesting to perform on such species.

We disagree. References given in the following sentences showed that the interest of many fungi is their easy manipulation for phenotyping, and offering the possibility to link their phenotype and their genomic structure. We keep this assertion, adding "*to characterize their phenotype*" to be more accurate relative to this assertion.

8- L245-246: the types of structural variation were already listed L1.

We removed this part.

9- L352: The common usage is to describe the trimming parameters used (quality and length). What is the reference of the prinseq software (the download website is not a reference)?

The description and the reference are now described L323-331.

Review by Benjamin Schwessinger, 2021-05-28 04:36

I am reviewing this manuscript as part of the PCI Genomics initiative. I do apologize being late. I think the overall conclusion of the manuscript are justified by the presented data that there are some structural variants between the assemblies and some changes in gene content.

I am confused about how candidate effectors were predicted and how CSEPs were analyzed as outlined below.

Major comments:

· I would suggest to tone down the claims on high quality DNA extraction for fungi and nanopore sequencing. The mean read length of 8Kb and N50 15.46kb are good but not outstanding for Nanopore, even for fungi. Also the 260/230 ratio of 1.45 does suggest residual polysaccharides in the DNA prep as this should be closer to 1.8. Plus there are several good fungal DNA extraction protocols available e.g. https://www.protocols.io/workspaces/high-molecular-weight-dna-extraction-from-all-kingdoms/publications and I would suggest the authors to add their protocol to the list.

We agree with this remark. We removed all the parts in relation with the idea of high quality DNA extraction, and we focused on the interest of the method for obtaining DNA of high molecular weight. The protocol is now on the suggested site (dx.doi.org/10.17504/protocols.io.bzbdp2i6) , and this is indicated in the manuscript (L296).

· This study would benefit with comparison to the recent Stauber et al. work https://elifesciences.org/articles/56279. I presume the current strain would fall into clade CL3 of that paper. This other paper also compared strains against the Crouch et al. 2020 reference but with short reads only.

We already mentioned in the previous version, the study of Stauber et al. (2020) where information on variation of gene repertoires among different genotypes was already available. The new 2021 paper was also added.

· Please explain the following: “EffectorP identified 1,117 models with a putative signal peptide,…pp.”. I don’t think effectorP predicts signal pepdites please clarify. This sections needs corrections as it is the wrong usage of EffectorP. EffectorP should only be applied on the secretome as mentioned in both original papers. This sections says there are 88 high confidence CSEPs. Latter analysis talks about effectors e.g S5 how are all these related?

This was indeed unclear. We clarified this point L541-555 by dissociating the potential secreted protein using the prediction of signal peptide done by SignalP (secretion only) and an upper level of potential effectors using the prediction done by EffectorP, defining precisely what the criteria used by effector P were (secretion and function). We then introduce a highly confident set of potential candidate secreted effector protein (CSEP) by intersecting the results of the two.

· What is the pre-RIP index? I could not follow the following section:

o “Pre-RIP index was significantly higher than the estimated baseline frequency (1.28 in both the genomes, Figure 3B), suggesting the absence of a RIP signature. By contrast, estimates of the two post-RIP indexes for the four classes of TEs were either significantly higher or lower than the two estimated baseline frequencies (0.67 and 0.82 for TpA/ApT and CpG/GpC respectively, Figure 3B), suggesting a RIP activity for both the two genomes.”

o These two sentences appear to be contradictory to each other.

The two indices do not estimate the same RIP signature. As indicated in M&M, they measure either the target of the RIP mutation (the Pre-RIP index) or the result of the RIP mutation (the two post-RIP index). They bring a different information, and can yield opposite results, likely due to their power to detect the signal. This point is quickly discussed (L1089-1097). We, however, modify the presentation of M&M (L492-499) and results (L911-919) to try to clarify this point.

o I also did not follow the rational for using CpG/GpC for RIP analysis. Please explain.

This point is explained in M&M L505: “*We also estimated a third index (CpG/GpC), specifically designed for C. parasitica, and measuring the depletion of the RIP on the basis of the preferential nucleotide targets in this species (Clutterbuck, 2011)*”.

Minor comments:

· Some of the citations in the literature in the introduction should be updated. E.g. there were several recent pan-genome papers e.g. tomato and soybean who looked at structural variation at the population level.

The list of references is already large, and we already cited a general recent paper (Wellenreuther et al. 2019) in the introduction. We however added a more recent paper (on tomato) showing the interest to focused on structural variations within species (Alonge et al. 2020).

· The ‘two-speed genome hypothesis’ is only applicable to a subset of oomycete and fungal crop pathogens. There are many crop pathogens that do not show this compartmentalization including rust fungi and others. It would be good to see this reflected int the introduction.

We totally agree. In introduction, we already presented in the previous version this debate in fungi (now in the present version L146-160)

The DNA extraction protocol and especially the polysaccharide cleanup step is really interesting I would encourage the authors to post a detailed protocol at protocols.io https://www.protocols.io/workspaces/high-molecular-weight-dna-extraction-from-all-kingdoms/publications

Done (see above)

· The genome assembly process seems a bit un-orthodox as they lack really good assemblers like Canu, Flye, or Masurca. I am also not convinced the Spades assembly is really the best as the differences in Illumina mapping rates, BUSCOs and Kmers is relatively small and would likely disappear with bootstrapping for these assemblies. Plus the high number of small might be a drawback of this assembly. Have the authors performed some quality control like BlobTools or such to see if these are spurious assemblies of bacterial contaminants? Just a minor comment. Also npScarf is really meant for simpler genomes like bacteria and not eukaryotes.

We tried different assemblers to deal with our Nanopore data. At the time of analysis and our knowledge, there was no standard for assembling these data. We agree that the final choice of the methodology may be not the best one. We modify our conclusions to indicate this test of different assemblers “*was very useful to assess a valuable strategy to obtain a near-chromosome assembly without generating chimeric scaffolds*” (L1030).

Concerning the contamination, we cannot totally ruled out this possibility. However, there are several results which suggested this possibility is small. First, we discarded in the final assembly all the scaffold smaller than 10 000 kb (L641). Second, we performed a manual curation of this assembly, by re-mapping the Illumina and Nanopore reads allowing to identify possible chimeric assemblies, especially too large coverage that may be a signal of bacterial contaminations. Third, during the genome submission to NCBI, no bacterial contamination was detected. We think, the risk of contamination is therefore limited.

Finally npscarf was very useful for us for scaffolfding. Although largely used for Bacteria, the method was originally and successfully also tested on one eukaryote (*Saccharomyces “cerevisae”*) (Cao et al. 2017).

- Overall nicely curated assembly at the end.

Thank you,, we did our best!