Review:

This is a nice paper tackling the assembly of an enigmatic species/ecotype pairs displaying a huge number of chromosomes. The genome will be usefull for people in speciation research, to study the role of SV maintaining difference among the two ecotype as well as for the evo-devo community.

These are nice assemblies of the two haplotypes for each species pairs with good quality. While I am slightly disapointed by the busco scores, QV and k-mer completeness, these remains a huge leap forwards for these species where reference genomes were lacking.

Below I detail some minor issues that will easily be adressed.

Introduction :

Line 76: could cite Rougemont et al. 2017 here whose results suggest that they would be more like ecotype.

Line 81-86: These seem to be the first genome assemblies for these species. Maybe explicitly state this in the introduction ?

Method:

Line 105: it is said that DNA was isolated taken from L. flvuiatilis's blood, but line 121 it is said that two libraries from muscle tissue were prepared. I am not sure to understand the difference. Was there a reason to extract DNA from blood for one species and from muscle for the other ?

Line 150-151: were default options used for hifiasm or a specific set of options ? Please clarify as this can substantially affect the assembly length.

Line 161: how many contaminated sequences were removed ? In my own assemblies, I found better results when removing contaminated sequences from the raw read, before assembly, rather than afterward. It is not clear if this step was done prior assembly or after.

Line 190-208: Please clarify wether some specific sets of parameters were used or only default value with the different set of tools used for annotation (galba, etc).

What e-value are used within diamond ? is a TE database necessary for Red masking ? Or is it only de-novo masking ? Please clarify

This workflow for genome annotation is nice but not really straightforward as compared to a simple "BRAKER3" annotation, have the author tried this first ?

Line 210: When the author state "we ran Flagger to detect possible mis-assemblies". I understand that to do so, the author used winnowmap, Secphase and deepVariant. Please, use some form of logical link between steps, otherwise the reader is left to guess what is done and why.

Also, can you be more accurate about what parameters were used if this is applicable ?

Line 217: Only biallelic SNPs are kept. I am not familiar with DeepVariant, but usually, we perform some more SNP filtering, based on various quality criteria, as in GATK or bcftools. How accurate are these SNPs calls ? Have the author check for the quality of the SNPs?

Results:

What proportion of the genome was masked with TE ? What is the distribution of the different class of TE ?

Do we have data on centromere and telomere motif ? Are they known and could they be identified ?

Line 311: “It was not possible to get an alignment toward the sea lamprey based on the settings we used, it was too divergent.”

I am unsure about the statement “it was too divergent” because I actually managed to align the genome of *Petromyzom marinus and L. fluviatilis* (version ena_PRJEB77117_sequence.fasta.gz) as well as *L. planeri on P. marinus* with minimap2

(minimap2 -cx asm10 --cs genome1 genome2 |gzip > out.aln.gz).

The rough estimate of per-base sequence divergence is indeed huge:

I found the median divergence (“de” column in paf output) to be ~ 0.06 between *P. marinus and L. planeri*, (same for *P. marinus* compared to *L. fluviatilis*) whereas the divergence between L. planeri and L. fluviatilis is ~0.0165 (and ~0.0159 for *L. fluviatilis and L. fluviatilis* UK).

Please consider using minimap2 as well if you wish to compare to *P. marinus.*

If you wish, you can then call the variants with paftools (command like so:

sort -k6,6 -k8,8n cs_aln.fluviatilis_planeri_paf |paftools.js call -f ena_PRJEB77190_sequence.fasta -L10000 -l1000 - > out.vcf ,

and probably look at number of SNP, insertions and deletions...

Although I’d say it is not necessarily usefull as the focuss is on comparing L. fluviatilis and L. planeri and the provided tables are already sufficient to me.

Discussion:

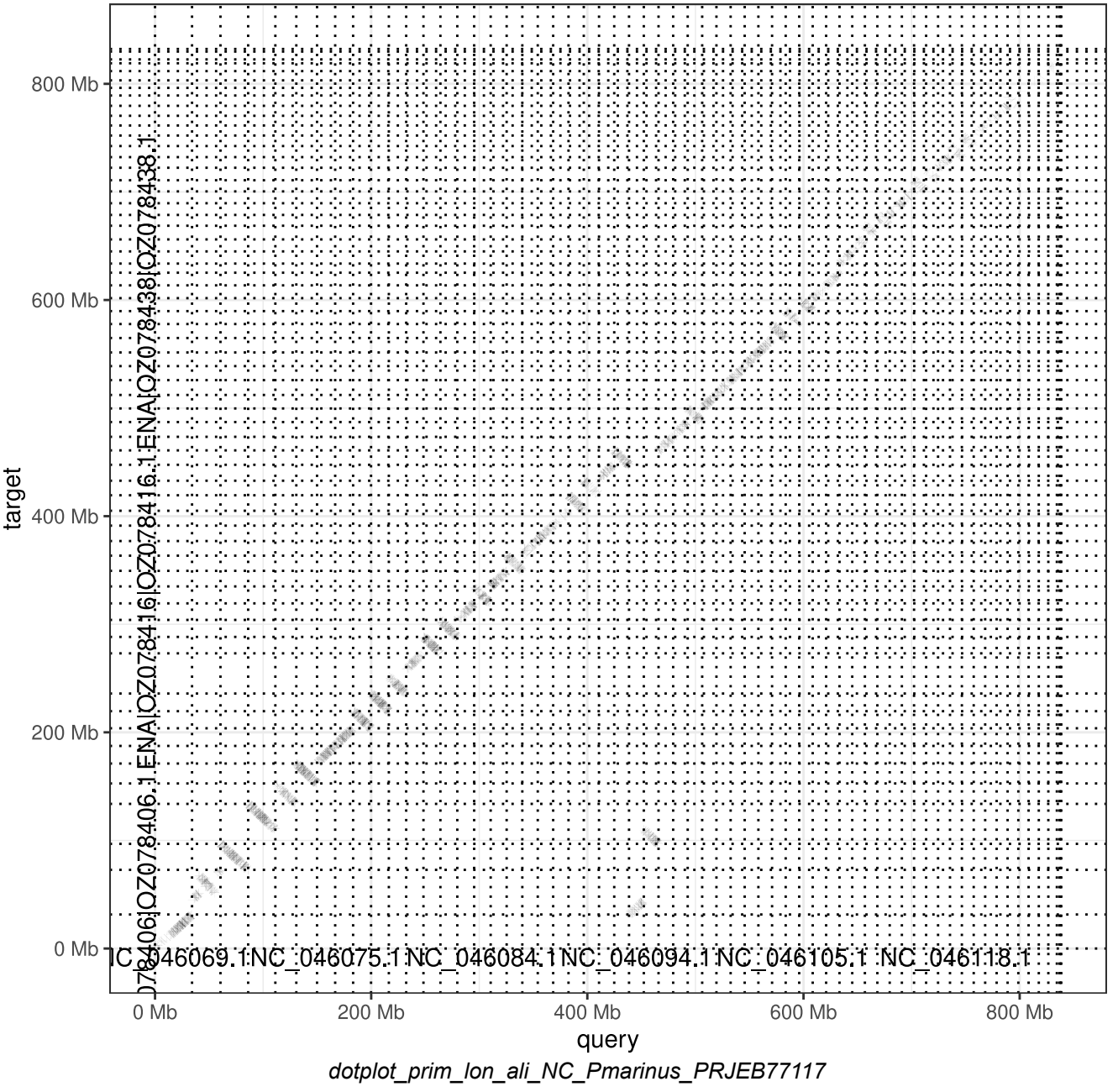
It is very usefull to point to the limits of BUSCO when using a small number of genes.

The discussion pave the way for further analyses of the (non)-speciation history of lamprey. Results are in line with those observed from RADseq data (Rougemont et al. 2017 for instance) where difference between the two ecotypes in sympatry is very weak. I am looking for whole genome sequencing studies taking advantages of these genome, paving the way for pangenome-graph studies in this system.

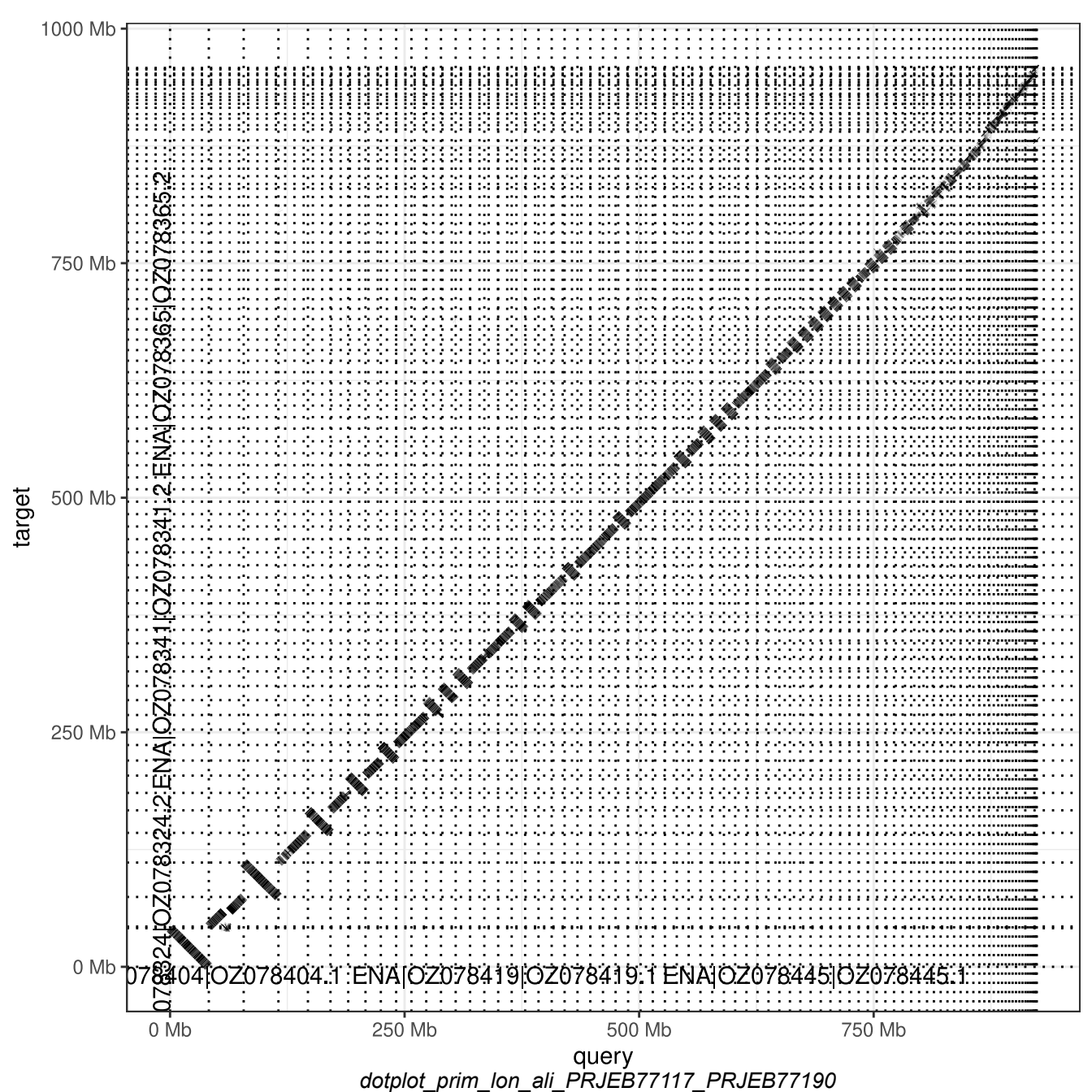
Very minor:

Supp. Fig 3-4 don’t seem to be mentionned in the text.

Supp Fig 5 mentionned after Supp Fig 6 to 10. Please reorder.



dotplot_prim_lon_ali_NC_Pmarinus_PRJEB77117


1000 Mb
750 Mb
500 Mb
250 Mb
0 Mb
target
OZ078324.2 ENA|OZ078324|OZ078324.2
OZ078341.2 ENA|OZ078341|OZ078341.2
OZ078365.2 ENA|OZ078365|OZ078365.2
OZ078404.1 ENA|OZ078404|OZ078404.1
OZ078419.1 ENA|OZ078419|OZ078419.1
OZ078445.1 ENA|OZ078445|OZ078445.1
0 Mb
250 Mb
500 Mb
750 Mb
query
dotplot_prim_lon_ali_PRJEB77117_PRJEB77190