Dear Dr. Holliday,

Please find below our answers to the comments and queries of the two reviewers.

Martin Lascoux,

on behalf of the authors.

Round #1

Decision

by Jason Holliday, 2018-10-26 20:10
Manuscript: https://doi.org/10.1101/402016

Minor revisions requested for "Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway Spruce"

Dear Dr. Lascoux,

Thank-you for your submission to PCI Evol Biol and apologies for the delay in getting back to you. I now have reviews from two experts in the field and invite you to revise your manuscript according to their recommendations. One reviewer mainly asked for clarification regarding how the data were partitioned for the various analyses, and also about the population groupings and their display (in addition to some suggestions for textual changes/clarification). The other reviewer was focused mainly on how the data were processed, the description of associated parameters, and the impact of using an incomplete draft genome on accurately disentangling paralogs.

Once these issues are addressed, I think your paper will make a nice addition to our understanding of the demographic history of trees, and of spruce in particular.

Sincerely,

Jason Holliday

Answer: We have addressed all the comments of the two reviewers and edited the text accordingly.

Reviews

Reviewed by anonymous reviewer, 2018-10-09 07:55

Review of manuscript “Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway Spruce” by Chen et al. 2018.

In this paper, Chen and co-authors describe a complex demographic history in Norway Spruce, Siberian spruce and Serbian spruce (P. omorika). Population structure in Norway spruce has been shaped by complex migration events, admixture events, and bottlenecks. They also estimated divergence times between the three species, which seem to have occurred earlier than reported in other studies. In general, I found the manuscript interesting and well-written, however I have some concerns in relation to the experimental design and other minor comments.

Major comments

My main concern is with the experimental design of this work, particularly about the type of markers used. Exome capture requires a priori knowledge of target sequences, therefore the sequences obtained are not "random". Also, there is generally a heavy emphasis on coding regions, if the SNPs are meant to be used for GWAS or linkage mapping. Of course, non-coding regions can also be targeted but that is not frequent. Alternatively, non-coding regions can be linked to coding regions in targeted sequencing, however these again would not be random. I understand that the 40,018 probes used in this study were designed to cover only coding regions. Then, my question is where did the non-coding SNPs come from? I am mainly worried about ascertainment bias and the absence of markers that could be used as null models in all admixture, SFS and population genomic analyses. Also, it is clear that the SNPs chosen were not "random" therefore may not accurately represent genome-wide patterns. A paragraph about potential limitations and sources of error in this study should be incorporated.

Answer: 1) For the "randomness" of SNP selection, there are in total ~26K coding genes with high annotation confidence (combined from Augustus and Eugene predictions with more than 70% homology with supporting evidence) in P. abies genome database. Our probes cover 77% of these gene models with on average two probes for each, though for technological and financial reasons we could not target all exons of the genes. In this sense, the genes and SNPs selection could be considered random.

Second, Chen *et al.* (2017) calculated π_0/π_4 ratios at coding sequences across a range of species. In trees, the proportion of mutations that are putatively under weak purifying selection is non negligible, especially for conifer trees. In the present study π_0/π_4 ratio is ~0.4 but Tajima's D values tend to be small suggesting that purifying selection did not strongly affected the site frequency spectrum. Furthermore, linkage disequilibrium in spruce decays very quickly within genes (within ~200 bp)(Heuertz *et al.* 2006; Chen *et al.* 2012a; Chen *et al.* 2014) so linked positive selection is also not likely to have affected nearby SNPs through hitchhiking or selective sweeps.

2) The probes can cover part of the exons and part of noncoding regions (introns, promoters, UTR or intergenic) as well.

3) There is no ascertainment bias, nor absence of markers as the data were generated from genome sequencing technology instead of genotyping. The genetic diversity Pi as well as SFS are also similar compared to previous studies based on Sanger sequencing.

4) We added a paragraph in the discussion about the two main sources of errors: the completeness of the spruce genome, and the potential paralogs for errors in SNP-filtering (also suggested by Reviewer2) P. 11 L15 -62.

How much of the exome capture design and SNP calling (lines 39-69) was done for this study and how much was previously done (and already reported) in Vidalis et al. 2018; Bernhardsson et al. 2018; and Baison et al. 2018? Please make this distinction in the paper.

Answer: As we have already reported (P4 Lines 47-55) that same number of probes (40,018) and P. abies gene models (26,219) were sequenced as the final set filtered in Vidalis et al 2018. Bernhardsson et al 2018 used 21,000 SNPs to generate parameters for variant quality score recalibration (VQSR). The VQSR protocols were also applied for the current dataset and resulted in 2,406,289 SNPs in total.

From what I understood, only 399k SNPs were used in the admixture and population genomic analyses. If this is the case then "1M SNPs" should be replaced by 399k or other number of SNPs used in the analyses. In the Results section (line 8, page 6), it is mentioned that only 30,000 SNPs were used, so this is confusing. It would be good to make clear the number of SNPs used in each of the analyses.

Answer: we have changed it to 400K in Abstract. There are about 1M SNPs after hard filtering. We used a subset of 400K unlinked noncoding SNPs for the admixture/clustering analyses and demographic inferences. The 30,000 SNPs (line 92, p5) correspond to the average number of SNPs for each population at coding sites (0-fold and 4-fold) that were used to calculate summary statistics such as genetic diversity (π_0 and π_4) and Tajima's D.

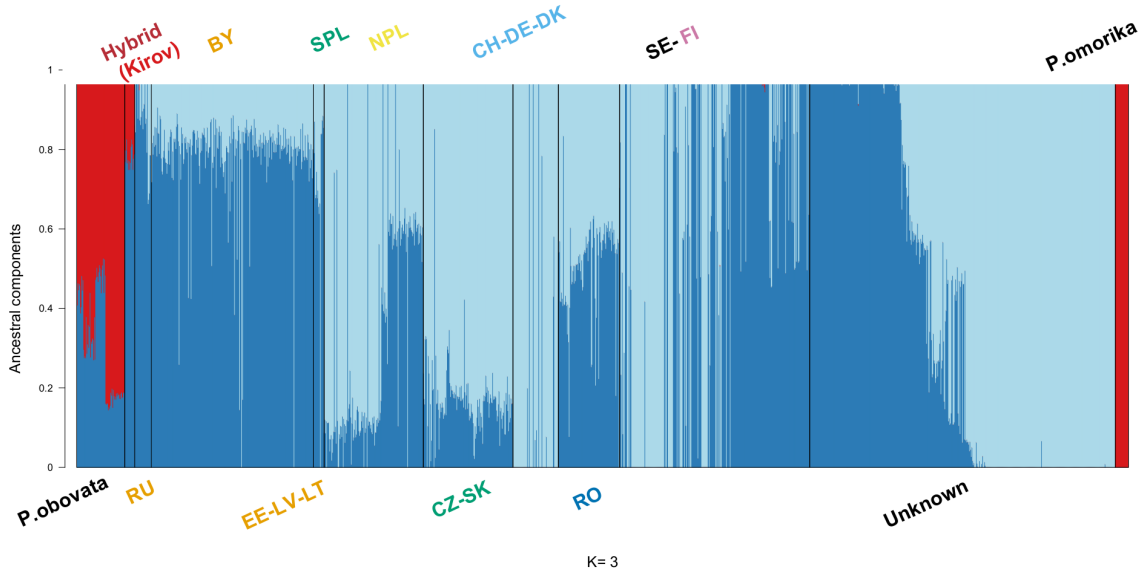
The admixture plot in Figure 2 is confusing. I wonder if it would look better with K=3 (abies, omorika, obovata). Including both plots (K=5, and K=3) would be more informative, I think. I was surprised to see very little signs of admixture between abies and obovata in the Admixture plot, where other analyses in this paper prove otherwise (lines 52-54, page 8). Any comments on that?

Answer: We feel that the reviewer comments stem from a too literal comparison between he ADMIXTURE plot and the TreeMix tree. These two analyses are not equivalent:

1) Both are consistent with migration from P. obovata being limited to hybrids, eastern European populations but not reaching southern populations (Alpine, Romanian). This confirms the finding of Tsuda et al 2016 Mol Ecol., which was based on a more extensive sampling.

2) The percentages in the Admixture plot do not correspond to the migration in TreeMix. Admixture shows the ancestral components for each population but does not imply directionality. The ancestral components could be traced from the shared common ancestors occurring at any level in the divergent history. In contrast, TreeMix mainly tests the significance of migration events in a specific direction and at a specific divergence time. It is in principle more powerful at detecting ancestral migration events that cannot be revealed by ADMIXTURE if the divergence between subpopulations is recent.

Below we show the result for K=3. At K=3 *P. obovata* consists of two components, one from from *P. omorika* and the other from the Fennoscandian population. Since K=5 is the best value and the interpretation of the ADMIXTURE results should only be based on the most likely value of K, we choose to present the result for K=5 in the main text. This also indicates that ADMIXTURE results are not a direct reflection of migration events, something for which TreeMix is better.



Page 9, line 17- Smaller mutation rates have been reported in more recent literature.

Answer: All mutation rates were estimated from synonymous substitution rates between P. abies and related species. It is difficult to make comparisons between different studies. Therefore, we chose the estimate from the P. abies genome paper (Nystedt et al 2013) for convenience, which was estimated from orthology comparisons of multiple species with different distances. The estimate of mutation rate does not change any results in this study but will alter the scaling of the divergence time and effective population size by a given factor (e.g. smaller mu means longer divergence time and bigger Ne).

Page 11, line 51-52.- There is convincing evidence that came from several recent papers that trees have lower and not higher mutation rates than annual plants (Lanfear et al. 2013; Bromham et al. 2015; De La Torre at al. 2017).

Answer: Yes, for mutation rate per year. But for mutation rate per generation we discussed here trees have larger mutation rate because of much longer generation time.

Page 5, Line 48- eliminate “of” after population sizes. Table 1- Kirov population is not in footnote. Figure 1- sample points are not easy to differentiate from each other, and from other group points. I suggest changing shape and/or color. Figure 1- EUFOGEN should read EUFORGEN Lines 15-20, page 7-Although levels of admixture in Kirov and Indigo are mentioned in these lines, these populations are not present in the Admixture plot in Figure 2. Table 3- Please explain what the different parameters names mean?

Answer: Kirov in Table 1 is the full name of the population and hence we did not add a footnote. Indigo does not show any difference in admixture and therefore has been grouped together with other populations of P. obovata. Kirov has shown half-half admixture of P. abies and P. obovata and is classified as a single group “Hybrid”. We have now added Kirov to Fig 2.

We have presented the meaning of all parameters in Table 3 at the left column as “Effective sizes” (N), “Divergence times” (T), “Admixture times” (Tadm), and “bottleneck times” (Tbot). The subscripts mean different domains.

Reviewed by anonymous reviewer, 2018-10-17 06:31

Chen et al present an extensive set of population genetic analyses of European spruce trees based on genotype data from over 1 million SNP loci, but the manuscript and supplementary material provide little detail on how the sequencing data were filtered and how quality control analyses were conducted on the genotype data, and no discussion of whether the overall conclusions would be different if different filtering and QC thresholds were imposed. A comparable analysis could be conducted with a much smaller genotype dataset, and the reader is left to speculate whether the outcomes would have been different if different criteria were used in the process of calling SNP genotypes and filtering to remove low-quality data.

Conifer genomes are large (typically >15 Gb) and full of repetitive sequences, including not only various classes of mobile elements but also processed pseudogenes that are very similar to existing functional genes in the genome. The use of exome capture sequencing as a method for genotyping therefore requires considerable care in filtering the sequencing data to avoid confounding reads derived from paralogous sequences with those derived from the intended target exons. The authors filtered the sequence data from their samples to minimize the likelihood of detecting paralogous sequences, but some additional information could be provided to allow readers to better judge the degree of rigor used. Some caution is also warranted due to the incomplete nature of the current genome assembly, which limits the ability of the authors to detect (and therefore to exclude) sequence reads derived from multi-copy paralogous sequences. An appropriate strategy to deal with this limitation is to filter the resulting SNP loci carefully to exclude those likely to represent confounded data from paralogous sequences rather than true genotype calls from single-copy loci.

The sequencing reads were aligned to the entire v1.0 draft genome assembly of Norway spruce rather than just the sequences of the target exons, which is good - this allows reads from diverged paralogous sequences that are represented in the draft assembly the opportunity to align to the copy of the sequence to which the read is most similar. The Chen et al manuscript does not point out, however, that the v1.0 draft genome assembly is estimated to include only 60% of the Norway spruce genome, although this fact is highlighted in the abstract of the cited reference by Bernhardsson et al (https://doi.org/10.1101/292151). The sequence reads are derived from the genomes of many individual trees (which may be different from each other as well as from the Z4006 reference individual used for genome sequencing), and are not limited to the subset of the genome represented in the v1.0 assembly. It is llikely, therefore, that paralogous sequences exist in the genome that are not represented in the assembly, and so the genotype data used as the basis for the rest of the manuscript may be a mixture, consisting of true genotypes derived from single-copy sequences and also confounded data derived from paralogous sequences. The relative proportion of this mixture cannot be determined from the data presented in this manuscript, nor even in the original complete dataset, given the fragmented and incomplete state of the v1.0 Norway spruce genome assembly.

One criterion commonly used to address this question is testing for genotype frequencies consistent with expectations based on the assumption of Hardy-Weinberg equilibrium - an excess of heterozygotes can be due to the presence of reads derived from paralogous sequences in the filtered sequence dataset used for SNP genotype callling. There is no mention of such a test in the manuscript, and no discussion of why such a test would (or would not) be suitable for filtering the genotype data prior to conducting the analyses described in the rest of the manuscript. There is also no discussion of the possible impacts on the conclusions drawn if confounded data are present in the genotype calls.

In the spirit of reproducible research, the authors could include the parameters used for read alignment by BWA and extraction of "uniquely-aligned pairs", as these steps are critical to the process of excluding reads from paralogous sequences. The supplementary material provided consists of material supporting the conclusions drawn by the authors regarding the genetic hypotheses, but does not include any details about the process of generating the genotype data on which those hypothesis tests are based. The authors could also provide additional information in supplementary material about steps taken to filter the putative SNP loci used for genotype calling, including the results of testing for consistency with HWE expectations for all SNP loci used in the analyses. Summary data could be included for all candidate loci, including the nature of the deviation from

HWE expectations for those that exceed a reasonable threshold, recognizing that some correction for multiple testing will be required due to the large number of loci to be tested. The effects on the number of genotyped loci of imposing different filtering thresholds could also be summarized - such thresholds would include both the depth of read coverage per allele called (the authors used a minimum of two), the minimum fraction of individuals genotyped (the authors used 50%), but could also include deviation from HWE expectations at different FDR thresholds.

Answer: 1) For identifying paralogs, it is very difficult to perform in P. abies genome as the genome sequences and the annotations are both far from perfection. As an alternative, we chose to align the reads to the whole genome and only kept the properly paired reads uniquely mapped to one position.

The paralogs indeed will lead to abnormal values of heterozygosity, coverage, or any other mapping statistics. Those statistics, such as p-value for HW test, have been included in GATK haplotypeCaller steps. The p-values of HW test (reported as ExcessHet in VCF) in final dataset were all insignificant. However, we did not choose to filter SNPs based on arbitrary cutoffs. Instead we used the variant quality score recalibration (VQSR) to do the filtering. The VQSR was trained based on sets of true SNPs gained from pedigree study (see details in Baison et al 2018 BioRxiv) and used various scores for filtering.

2) We have now provided a more detailed description of the alignment and filtering.

3) We have now added a paragraph in the Discussion about the two main sources of sequencing errors as paralogs and genome incompleteness.

4) We have now added quality plots in supplementary for 1M SNP after filtering.