Peer Community In Evolutionary Biology

Disentangling the recent and ancient demographic history of European spruce species

Jason Holliday *based on peer reviews by 1 anonymous reviewer*

Jun Chen, Lili Li, Pascal Milesi, Gunnar Jansson, Mats Berlin, Bo Karlsson, Jelena Aleksic, Giovanni G Vendramin, Martin Lascoux (2019) Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway spruce. bioRxiv, ver. 1, peer-reviewed and recommended by Peer Community in Evolutionary Biology. 10.1101/402016

Submitted: 29 August 2018, Recommended: 10 January 2019

Cite this recommendation as:
Holliday, J. (2019) Disentangling the recent and ancient demographic history of European spruce species. *Peer Community in Evolutionary Biology*, 100064. 10.24072/pci.evolbiol.100064

Published: 10 January 2019

Genetic diversity in temperate and boreal forests tree species has been strongly affected by late Pleistocene climate oscillations [2,3,5], but also by anthropogenic forces. Particularly in Europe, where a long history of human intervention has re-distributed species and populations, it can be difficult to know if a given forest arose through natural regeneration and gene flow or through some combination of natural and human-mediated processes. This uncertainty can confound inferences of the causes and consequences of standing genetic variation, which may impact our interpretation of demographic events that shaped species before humans became dominant on the landscape. In their paper entitled "Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway spruce", Chen *et al.* [1] used a genome-wide dataset of 400k SNPs to infer the demographic history of *Picea abies* (Norway spruce), the most widespread and abundant spruce species in Europe, and to understand its evolutionary relationship with two other spruces (*Picea obovata* [Siberian spruce] and *P. omorika* [Serbian spruce]). Three major Norway spruce clusters were identified, corresponding to central Europe, Russia and the Baltics, and Scandinavia, which agrees with previous studies. The density of the SNP data in the present paper enabled inference of previously uncharacterized admixture between these groups, which corresponds to the timing of postglacial recolonization following the last glacial maximum (LGM). This suggests that multiple migration routes gave rise to the extant distribution of the species, and may explain why Chen *et al.*'s estimates of divergence times among these major Norway spruce groups were older (15mya) than those of previous studies (5-6mya) – those previous studies may have unknowingly included admixed material [4]. Treemix analysis also revealed extensive

admixture between Norway and Siberian spruce over the last ~100k years, while the geographically-restricted Serbian spruce was both isolated from introgression and had a dramatically smaller effective population size (*Ne*) than either of the other two species. This small *Ne* resulted from a bottleneck associated with the onset of the iron age ~3000 years ago, which suggests that anthropogenic depletion of forest resources has severely impacted this species. Finally, ancestry of Norway spruce samples collected in Sweden and Denmark suggest their recent transfer from more southern areas of the species range. This northward movement of genotypes likely occurred because the trees performed well relative to local provenances, which is a common observation when trees from the south are planted in more northern locations (although at the potential cost of frost damage due to inappropriate phenology). While not the reason for the transfer, the incorporation of southern seed sources into the Swedish breeding and reforestation program may lead to more resilient forests under climate change. Taken together, the data and analysis presented in this paper allowed inference of the intra- and interspecific demographic histories of a tree species group at a very high resolution, and suggest caveats regarding sampling and interpretation of data from areas with a long history of occupancy by humans.

References:

[1] Chen, J., Milesi, P., Jansson, G., Berlin, M., Karlsson, B., Aleksić, J. M., Vendramin, G. G., Lascoux, M. (2018). Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway spruce. BioRxiv, 402016. ver. 3 peer-reviewed and recommended by PCI Evol Biol. doi: [10.1101/402016](https://dx.doi.org/10.1101/402016)

[2] Holliday, J. A., Yuen, M., Ritland, K., & Aitken, S. N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. Molecular Ecology, 19(18), 3857–3864. doi: [10.1111/j.1365-294X.2010.04767.x](https://dx.doi.org/10.1111/j.1365-294X.2010.04767.x)

[3] Ingvarsson, P. K. (2008). Multilocus patterns of nucleotide polymorphism and the demographic history of Populus tremula. Genetics, 180, 329-340. doi: [10.1534/genetics.108.090431](https://dx.doi.org/10.1534/genetics.108.090431)

[4] Lockwood, J. D., Aleksić, J. M., Zou, J., Wang, J., Liu, J., & Renner, S. S. (2013). A new phylogeny for the genus Picea from plastid, mitochondrial, and nuclear sequences. Molecular Phylogenetics and Evolution, 69(3), 717–727. doi: [10.1016/j.ympev.2013.07.004](https://dx.doi.org/10.1016/j.ympev.2013.07.004)

[5] Pyhäjärvi, T., Garcia-Gil, M. R., Knürr, T., Mikkonen, M., Wachowiak, W., & Savolainen, O. (2007). Demographic history has influenced nucleotide diversity in European Pinus sylvestris populations. Genetics, 177(3), 1713–1724. doi: [10.1534/genetics.107.077099](https://dx.doi.org/10.1534/genetics.107.077099) ”

Reviews

Evaluation round #2

Reviewed by anonymous reviewer 2, 26 November 2018

The revisions made by the authors have added sufficient clarity to the manuscript to address the concerns about how SNPs were identified and filtered. The work described represents a significant technical achievement in application of genomic technologies to conifer population genetics.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/402016
Version of the preprint: 1

Authors' reply, 23 November 2018

Download author's reply
Download tracked changes file

Decision by Jason Holliday, posted 23 November 2018

Revise

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

Reviewed by anonymous reviewer 2, 17 October 2018

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.