Pauline Garnier-Géré
Bordeaux-Aquitaine center
INRAE, UMR Biodiversity Genes & Communities
Team "Evolutionary genetics of species with complex life cycles: Data and Modelling Integration"
69, route d'Arcachon, 33612 Cestas cedex, France
pauline.garnier-gere@inrae.fr

Manuscript: "High-quality SNPs from genic regions highlight introgression patterns among European white oaks (*Quercus petraea* and *Q. robur*)." by Lang *et al.*

Dear Dr Dreyer,

We are grateful for your time in searching for referees, and to you and the referees for the positive, detailed and constructive comments on our work. A fully revised version is now available in BioRxiv at https://www.biorxiv.org/content/10.1101/388447v3, with new or completed supporting information. This version accounts for all comments and issues raised, in particular the issue about clarifying the results content and objectives of the manuscript, in order to distinguish it from a data paper.

Please see below our point-by-point replies to the referees' comments and questions, and to the new analyses suggested, as well as about changes that we included in the manuscript. Line numbers are indicated in the new paper for each correction or change. See also the correction mode versions uploaded for tracking changes in both the new manuscript and the new Appendix S1, if needed.

Table 1 and Figure 3 were completed following comments from referee 1, and a new Figure (S8, supporting information) was produced to better answer another comment (see **reply 1.3**). Appendix S1 was largely extended to simplify the main text in the method part, and to explain new analyses that were suggested by referee 3.

As requested all initial and produced data, along with original Bioperl and R scripts are available in different repositories detailed in the Data accessibility section and throughout the text, and all web links have been double checked.

Looking forward to hear from you regarding this new version of our manuscript,
Yours sincerely,

Pauline Garnier-Géré, on behalf of the authors,

Recommender's report

Minor revision of the preprint before final and positive recommendation.

Dear Pauline,

I do apologize for the long time it took to reach this decision. We had trouble finding suitable referees and had to sollicit a large number of colleagues to do the job. This is linked to the fact that the concept of PCI is still not fully integrated by the potential referees and our research community; we still have a large communication effort to do.

The reviewers produced very positive assessments of the manuscript and praised the precision and details of the description of the data and the usefulness of the attached data based. One of the reviewers stated that the preprint sounded like a data paper, but added that the text provided many useful and precise discussions of the data that go beyond what is expected from a data paper.

The different reviewers suggested some minor changes in the text that I am not going to detail again here, and that you will find in their comments. I guess from what I read that the required changes are very minor and should not be too difficult to introduce.

Once these changes are made and the new version of the preprint is deposited in the archive, I will be able to provide a very positive recommendation of the preprint, which would be the first (or possibly the second one) for the PCI Forest&Wood Science.

Thank you anyway for your patience.

I am looking forward to consider the revised version and write the final recommendation.

With best regards

Erwin Dreyer

INRAE, UMR Silva (Université de Lorraine, AgroParisTech & INRAE), Nancy.

Recommender for this preprint

Reviews

Referee 1

Reviewed by Oliver Brendel, 2020-09-28 07:28

Review of "High-quality SNPs from genic regions highlight introgression patterns among 2 European white oaks (Quercus petraea and Q. robur). " from Lang et al.

This review is based on a version downloaded from the BioRxiv server: https://www.biorxiv.org/content/10.1101/388447v2

My expertise is in forest ecophysiology, especially oaks, with some knowledge on the genetics of these two oak species. So I will not be able to comment the sequencing techniques that were used in this manuscript as well as the finer points of the genetic implications of the results.

Overall, this manuscript describes a large data base of nucleotide polymorphisms of overall 24 trees (13 Q. robur,11 Q. petraea) distributed over the geographical range and over 4 of the 5 cpDNA lineages along white oaks recolonization routes, with 2 to 4 individuals per location. The large number of detected nucleotide polymorphisms are then described in terms of their quality, coverage, introgression patterns, exclusive polymorphisms, polymorphism typology and relation to gene ontology . The presented catalogue of nucleotide diversity will not only be very useful for future research in landscape ecology of these two and other genetically close oak species, but has been analysed in terms of differences between the two species. These confirm earlier results, but using a large covering of the geographic range and a high number of nucleotide polymorphisms.

The abstract is very detailed and descriptive on the experimental set-up and the results in terms of detected polymorphisms, numbers of SNPs, the quality of the data provided for future research. There is only one short allusion on the results of differences in diversity between the two species where "these patterns are discussed in the context". This is unfortunate, as the discussion broaches interesting subjects such as the introgressions detected and their relation to leaf morphology. Also two hypotheses are discussed for the detected higher diversity in Q. petraea (as compared to the one mentioned in the abstract). This gives the reader rather the impression of a data paper, which is less the case once the whole manuscript is read.

Reply 1.1: Thank you, you are right that we provide estimates of nucleotide diversity across hundreds of genic regions for the first time in European oaks. The maximum number of genes from previous cited studies for this type of estimates to our knowledge was below 10, and most large datasets in *Quercus* species focus on SNP diversity due to the difficulties of recovering haplotype data of good quality. In that sense, this manuscript was always going to be more than a data paper, with a detailed description of diversity patterns across many genes, while more work is needed for inferring evolutionary scenarios using those data (see also reply 2). However, we felt that is was important to recognize the large amount of work involved for producing the genomic resources, and also the fact that they have been used in several studies during their development and before being fully developed. Following your advice, we have now further summarized the result content of the paper and possible interpretations for the patterns observed (lines 50-59 of the new manuscript, hereafter noted "new lines…").

The introduction gives a quite complete context of establishing such a catalogue in terms of sequencing techniques, available genomic resources in oaks, knowledge on species differentiation and species introgression pattern. Their main objective was to provide a detailed characterization of sequence variation in Q. petraea and Q. robur. Using these data, they investigated the species differences and suggest future research possibilities.

Given my expertise, I have only few comments on the techniques described in the material and methods section as well as the very detailed genetic results presented in that section. The representativity of the 24 trees in terms of covering the geographic range as well as the genetic diversity is nicely presented in this section. It would however be helpful if the affiliation to the different postglacial migration routes could be included into table 1.

Reply 1.2: Thank you for this suggestion; we have now added in Table 1 the information for the main lineages, haplotypes and putative refugia, based on Petit et al. (2002a, 2002b). Using the information compiled in the GD^2 Database of the Quercus Portal (http://gd2.pierroton.inra.fr/), we report the cpDNA haplotypes from the trees located within a 50 km radius around the trees sampled in this study (last 2 columns in Table 1, new lines 149-153 below Table 1 and 158-163 in the main text for information on putative refugia, and one new reference Petit et al. 2002b).

A large part of the discussion describes again the results on the detected nucleotide polymorphisms, in terms of availability in other species, the representativity of the experimental setup, the sequencing method used, the quality of the data, polymorphism typology, and the content and formats of the catalogue, by comparing with already available data in the literature, for oaks or other forest tree species. They mention that the chosen locations cover 4 of the 5 cpDNA lineages along white oaks recolonization routes. As it is argued that 8 gametes were sufficient to represent most polymorphisms. Therefore, given the 40 gametes across both species (line 386), I wonder if it would not be possible to do a comparative analysis of the diversity in the four cpDNA lineages. This could add an interesting discussion of differences between these lineages.

Reply 1.3: Questions related to a comparative analysis of nuclear gene diversity across different lineages representing past colonization routes are not easily answered, given the multiplicity of evolutionary factors and the geographic scale involved. However, based on white oaks' life history traits (high fecundity and dispersal rates), large population sizes, knowledge of their past recolonization history throughout Europe from putative refuge areas, the large adaptive differentiation in contrast to a very low genetic differentiation among distant populations, and also previous comparisons between cpDNA and nuclear or phenotypic traits divergence across populations (e.g. Kremer et al. 2002, Kremer et al. 2010; Petit et al. 2002a, 2002b; Guichoux et al. 2013, Saenz Romero et al. 2017), predictions are that we should observe both

a large and comparable diversity within groups of individuals belonging to different cpDNA main lineages, and an absence or a small divergence among those lineages at nuclear genes.

Following your suggestion, we tested whether nucleotide diversity levels were different between the main lineages and estimated also their differentiation in both species across genes. After checking cpDNA haplotypes of trees located within a 50 km radius of the studied trees (see reply 1.2), only the 3 main European lineages A, B and C are actually represented in our discovery panel (although geographical areas including haplotypes A, C and E largely overlap). Due to a mixture of putative haplotypes from lineages A and C in the Netherlands and Germany, we could not attribute the samples to either one or the other. This was consistent with historical knowledge showing that Italian and Balkan refugial areas shared both A and C lineages and were probably not fully isolated during the last glaciation period. It was thus possible to consider one group of individuals from lineages A and C, and another group for lineage B individuals, still excluding the most introgressed individuals in both groups within each species. As expected, we found no significant differences in nucleotide diversity among lineages, and a very low differentiation overall (new lines 435-442, see also the new Figure S8), and those results are discussed in the context of previous knowledge on these species past demographic and adaptive history (new lines 727-746).

The authors also discuss in more detail the observed introgression in relation to the morphological species attribution, and they confirm on their small data set, that morphological *Q. petraea* were more likely to be introgressed individuals than morphological Q. robur. Further, they observed a higher diversity of *Q. petraea* for genes classed as “abiotic stress” in gene ontology terms. The authors then discuss two non-exclusive hypotheses proposed in the literature, one pertaining to the difference in life-history strategies for colonizing new stands, the other pertaining to higher selective constraints in Q. robur. The discussion finishes with an outlook for a future usage of the presented nucleotide diversity catalogue.

This manuscript not only provides a solid foundation for future landscape ecology studies but also presents interesting insights into the genetic difference between two closely related, hybridizing oak species. I therefore recommend strongly the publication of this manuscript.

Detailed comments

Abstract : Probably not all readers will know what Sanger sequencing is, similarly for the “phi-pi” (line 51).

Reply 1.4: We removed the “Sanger” reference from the abstract but added its definition in the introduction with the original reference (i.e. “Sequence data from the classical Sanger’ chain-terminating dideoxynucleotides method…”, new lines 125-126). For theta-pi, we linked it to its definition in the sentence just before in the abstract (new line 45).

Introduction : line 88 "species pairs" : This refers probably to the four species cited above and their capability to hybridize ? Also it is not clear what "scenarios" refers to, scenarios for the species distribution ? This section might need some editing to clarify these points.

Reply 1.5: Both points have been clarified in the introduction (new lines 91-92).

Material and Methods line 144 : It would help the reader at this point if it could be mentioned that the à priori morphological species assignments will be supplemented with genetic assignments and an analysis of introgression further down in the manuscript.

Reply 1.6: This has been added to the description of Table 1 (new lines 152-155).

Lines162-163 : It is not clear to me what these 146 individuals from 3 French regions relate to. Were these the basis for the EST ? This seems to be a much larger choice than the 25 individuals described at the beginning of the M&M section.

Reply 1.7: The tissues of these 146 individuals were sampled to produce the different libraries used for the previous development of EST databases that we initially assembled (published in Ueno et al. 2010). This

assembly was then used for choosing fragments to be re-sequenced in this study in a smaller sample of 25 individuals (our Discovery panel). In order to avoid confusion between both samples, and also to reduce the length of the methods part (see also reply 2), the part referring to the 146 individuals has now been moved to Appendix S1 in supporting information ("*Original assembly description…*" part, lines 12-18, and 35-37).

Results : In Figure 3, the limits of the attribution of the individual to one or the other species should be visible. It is not clear which 4 individuals were excluded from the following analyses.

Reply 1.8: We added on Figure 3 two horizontal lines corresponding to the 0.125 and 0.875 values for assignment probabilities, which can be considered as biologically meaningful threshold for belong to a genetic cluster (see Guichoux et al. 2013, and new lines 353-356). Mean *Q*-values for 3 individuals fall within those limits and were excluded for species nucleotide diversity estimation due to their introgression levels. A fourth individual was also considered more introgressed than those belonging to genetic clusters, due to its large Bayesian confidence intervals across many runs of the STRUCTURE analyses. This is now clarified in Figure 3 legend, and in the text (new lines 365-368, see also new line 414 in the footnotes of table 3).

Referee 2

Reviewed by Ricardo Alia, 2020-07-17 07:10

The paper is a complete characterization of SNPs in oak species, and the application of these resources to two important topics of research: characterization of sequence variation in Quercus petraea and Quercus robur, and the analysis of introgression asymmetry. The paper is very well written, with a detailed description of the methods applied and a complete discussion of the results. The resources developed are important for the study of the species, and the only concern is that the different objectives of the paper (description of the new resources, characterization, and application to the analysis of the introgression) are quite different. Then, the paper is quite long and in some cases difficult to follow, but as the information is relevant, I do not have any significant recommendation to the authors.

Reply 2: thank you, we understand and agree with your critic here about the different objectives. The main goal was to make oak SNP resources (and corresponding transferable sequence primers/targets) readily available to the community, and these resources have already largely been used in the last 10 years. However, since these SNPs come from high-quality resequencing, we felt that it was important to provide a detailed characterization of nucleotide diversity and differentiation patterns across a large number of genes, which is rarely provided, and also difficult to perform in oaks without considering introgression patterns. More applications could be performed and will follow, but we thought that developing a few would help potential users understand how these data could be useful. We have now provided more information about the study results in the abstract to clarify this (reply 1.1). In order to make it easier for readers, we reduced text length by moving details on bioinformatics analyses to Appendix S1 (lines 35-65, corresponding to new lines 172-182), and added homogeneity to some bioinformatics terms throughout both texts.

Referee 3

Reviewed by Komlan Avia, 2020-07-17 14:46

Review of the article entitled "High-quality SNPs from genic regions highlight introgression patterns among European white oaks (*Quercus petraea* and *Q. robur*)"
Authored by Lang et al.

Lang et al produced Sanger sequences from over 800 gene fragments (including a set of genes representing broad functional categories potentially involved in species ecological preferences as well as a random set of genes) across the genic portion in 25 individuals of 3 European oak species. They set up a pipeline to clean up and characterize these gene fragments giving over 14500 polymorphisms that were used to provide various summary statistics within and among species. The authors observed patterns of significantly higher diversity in *Q. petraea* vs *Q. robur* and a heterogeneous landscape of both diversity and divergence. The authors highlighted the usefulness of the generated data in medium scale landscape and molecular ecology projects.

Comment to authors

The manuscript is very well written, the provided data are sound and well used to support the drawn conclusions and the discussion section was well constructed. The generated resources will be valuable for the community. I particularly liked the fact that answers to questions coming up while reading the manuscript could be found in the discussion part. For example, one immediate question was the usefulness of such a data covering only 529Kb of genic regions, which corresponds to barely 0.072% of the Q. robur genome or 1% of the gene space length, while methods such a GBS are very popular nowadays with decreasing costs.
Even though NGS methods are now the preferred ones for genomic studies, Sanger sequences still valuable resources especially regarding their lower error rates; and as indicated by the authors, this dataset will be useful as control for future NGS sequences.

- Although the authors acknowledged a possible ascertainment bias depending on which materials their produced resources will be used on, it would be useful that they discuss the possibility that the produced resourced might be skewed towards more conserved genes in *Quercus* and we know that the more transferable primers are, the more conserved the targeted genomic regions are. For instance, what would be the outcome if primers specific to each species were used (pairs of primers amplifying fragments in one species and failing in the other and vice versa, hence targeting more divergent loci), in terms of calculated summary statistics). Based on the current data, could the authors specify whether some primer pairs were actually successful in only a specific species? If so, would it be possible to produce within species summary statistics for those amplicons and compare them to common ones?

Reply 3.1: This is an interesting point that we had tried to partly address during the initial choice of regions/genes to resequence from the EST data available, using both BLASTX results onto model plant species (some distantly-related) databases, and in the primer design strategy:

- **We verified that there was no significant differences in polymorphism patterns between contigs whose consensus had very low E-values in BlastX searches ($<10^{-80}$) and those with relatively higher E-values. This was indicated already in Fig. S2-B -steps 1 and 3, but we have added more details now in Appendix S1 (lines 56-62).**
- **Also since we had access to ESTs from a large number of individuals in both two species, and thus displaying potential variants in contigs' assembly, we designed primers within regions showing no or very low diversity among ESTs *a priori* in both species, these regions often being very small compared to other contig parts, and often framed by regions with many potential variants. Thus primers were actually designed to avoid targeting conserved sequence regions across both species (see also step 3 of Fig. S2-B, and details added in lines 50-55 of Appendix S1).**
- **In contrast, all steps aimed at targeting polymorphic genic regions, which is consistent with the results obtained (e.g. large number of detected variants, low percentage of fragments with no detected variants, significant portion of variation due to Indels and SSRs, Table 2), and tends to demonstrate that the strategy followed did not focus on genes that were particularly conserved.**

One possible bias however is for genes belonging to large multigene families or genes with fairly conserved paralogs since these were had to be excluded for optimizing sequence data recovery with the technique used.

What would have been the outcome if species-specific primers had been used is difficult to predict, and this question needs to be examined in the particular context of very closely related species at the genomic level, since they share at minimum half of their variants, and show a low overall differentiation for most of the genic regions analyzed here (Fig. 4-C). Within many gene fragments, both exclusive and shared variants were often separated by less than 100 bp due the overall very low linkage disequilibrium (see Tables S3 and S4, supporting information). In this particular context, assuming that we had known about some regions showing more divergence initially, we might even wonder if they would not have shown less polymorphism, not because of purifying selection but due to possible divergent selection among species, although this might be very dependent on the particular genes analyzed.

Following your suggestion, we examined further the initial data obtained from the 2000 designed amplicons (we could finally obtain data on 1968 of them). At this stage (step 1 in Fig. 1-A), we only had data on 2 diploid individuals, one from each species, and chose the best 1000 amplicons based on a strong quality filter, favoring those where data had been obtained for both individuals. Additional analyses considering all initial 1968 fragments are now detailed in Appendix S1 (lines 66-97), and referred to in the methods (new lines 188-189) and further in the discussion on possible ascertainment bias (new lines 565-576). They show in summary that:

- **More than 85% of the fragments amplified in individuals of both species with good enough quality, consistently with their design targeting both species initially (step 1 in Fig. 1-A but the best ~1000 (50%) in terms of quality were kept due to budget for the next steps).**
- **For more than 150 independent fragments (~100 kb) amplifying in only one individual of one species (fragments A), the number of detected heterozygotes was on average twice smaller compared to fragments B amplifying in both individuals (covering ~500 kb, whether comparing with the Q. robur/11P individual or with the Q. petraea/Qs21 individual). This could be due to lower diversity across the ~100 kb represented by those fragments A, but since data on one individual per species was available only, it would need to be verified on more individuals per species to conclude. Also, the quality filtering may have masked polymorphic sites located within heterozygote indels fragments' parts, consistently with their overall lesser apparent quality. However, bioinformatic treatments for fragments A and B were the same, and if some polymorphisms were missed they do not represent the majority of variants included in the comparisons (see Appendix S1 lines 84-98 for more details).**

Overall, given the strategy followed for choosing the fragments and designing the primers, given the preliminary results on the 1968 amplicons, given the diversity patterns observed, we can conclude that the targeted regions were not particularly conserved, many of them were actually very polymorphic (and included intron sequences that were not in LD with exons), although genomic sequences with different characteristics such as a greater distance to genic regions are likely even more polymorphic.

- What species were used to produce the first 103.000 Sanger sequences?

Reply 3.2: Both species were used in a balanced manner in the different libraries. This is added in the methods (new line 175), and further detailed now in Appendix S1 (lines 12-15).

- Line 740 – : I don't quite agree with the authors. Since it is possible to multiplex hundreds of samples for methods such as RAD-seq with a reasonable cost, such methods could capture even larger genome portions than the one obtained here, to address questions such as those addressed in the manuscript. If there is a low overall differentiation as mentioned, it is even more likely that enzyme digestion produces similar generated fragments for sequencing. Of course, simulation using the published Q. robur genome could tell what might be the proportion of genic regions that would be sequenced if only genic regions are of interest. And to me, developing SNP arrays is another question that should be separated from methods such as RAD-seq.

Reply 3.3: thank you for your comment, this part should be clarified indeed. We agree that RADseq and related methods can cover/target much larger portions (and also random parts) of the genome than what is reported in this study and they can also allow multiplexing for larger numbers of individuals. Thus such data could have been useful for developing many SNPs covering the genome and capture a large part of the genomic diversity. When we related to research questions "above", we actually meant the future questions just mentioned in the paragraph, that would require more accurate estimates of nucleotide diversity or differentiation, linkage disequilibrium and haplotype diversity, which are more difficult to obtain with RAD-seq derived data (see references cited in new lines 796-800 and 810-818), i.e. we did not mean the questions we addressed partly in this study with SNP data, so the text has been clarified.

Since RADseq methods have been preferentially used to develop large arrays of SNPs and thus SNP diversity and Fst estimates, we wanted to point out that due to all the filtering steps for limiting potential biases, error rates and paralogs assembly in many species, accurate nucleotide diversity estimates and inferences

are difficult to obtain (Andrews et al. 2016, Nature Reviews Genetics 17: 81-92), especially for species harboring a large heterogeneity of diversity such as in the oaks studied here. Methods advance rapidly and new bioinformatic pipelines could however surely allow, with new data, overcoming some of the RADseq limitations, especially given a reference genome sequence which is now available in *Quercus*. Still, given the significant amount of complex polymorphisms described here, it's difficult to make predictions without comparing methods on real datasets, and bioinformatic strategies actually followed. We modified the text by referring to more comprehensive reviews on these issues with RADseq than can be proposed here, and separated the discussion on SNP arrays as you suggested (new lines 800-810).

Referee 4
Reviewed by Hilke Schröder, 2020-07-01 10:08

I am impressed by the amount of data, analyses and interpretation presented in this paper. The authors not only made an exhaustive effort with Sanger resequencing of a lot of interesting genes, they also gave a kind of review for already existing data, interpreted them and made recommendations for further studies. The data sets will provide scientist working with oaks with unprecedented possibilities for future projects in a broad range. This is a very valuable paper as well how it is organized as because of the extensive range it is covering. So, I can only suggest to recommend it.

Only one small request: Already in the introduction („across a large part of both species geographic range") and also in the discussion, the authors stated that they used *Q. robur* and *Q. petraea* populations „across a large part of their geographic range". I would like to have this statement qualified because the most Eastern population used by the authors is in Hungary. The further South-Eastern distribution is missing. Maybe it can be stated as „Western and Central Europe" as already mentioned in the chapter „sample collection"

Reply 4.1: thank you for your nice comments. You are correct about the geographic range needs to be better described. We detailed it further in the abstract (new line 37), introduction (new line 157), and discussion (new line 485). We also cited a recent application of these SNP resources, for populations located outside the range of our discovery panel in the south-eastern margins for *Q. robur*, which showed a high rate of genotyping success (new lines 543-547 in the discussion), thus illustrating the representativity of the panel studied for a wider geographic range.