UNITÉ EVOLUTION, ECOLOGIE ET PALÉONTOLOGIE - UMR 8198
C.N.R.S. - UNIVERSITÉ DE LILLE, SCIENCES ET TECHNOLOGIES

Lille, November 11, 2021

Dear authors and editor,

In the manuscript "*Conserved genomic landscapes of differentiation across Populus speciation continuum*" by Shang *et al.* the authors study the evolution of genomic patterns of polymorphism and divergence along a divergence gradient. Personally, I will speak of a divergence gradient and not of a speciation continuum, as I am not yet convinced that there can be a continuum of a process that is itself continuous. For this, the authors analyze a still rare and very high quality dataset composed of 201 whole poplar genomes sequenced entirely from 8 different species, with an average coverage varying from 21X to 32X. This yields >30 million SNPs for a 500Mb genome, and thus, allows for robust quantitative analyses.
Overall, I am impressed with the data acquisition work and the underlying analyses. Despite the revolution in sequencing methods that began more than 10 years ago, biological models studied in genomics in such detail can be counted on the fingers of one hand. First, the study is an important step in advancing our knowledge of the *Populus* model, with results that are consistent with each other. It is impressive how the observed genomic patterns (pi, Tajima's D) are suggested by the SMC++ analysis, making the interpretations consistent. The rarity of introgression is however very surprising with here only two lineages connected by gene flow, while the IBD analysis could suggest more exchanges between species. This point is in my opinion little discussed by the authors and would deserve more development.

This study is also interesting beyond the *Populus* model as it is one of the pioneering studies describing the effects of the divergence process on genome-wide molecular patterns (Figure 6). I find these patterns important to empirically illustrate expectations and to aid future interpretations. However, it is not clear to me how such patterns will be used, especially to discriminate different demo-genomic scenarios (figure 1). Can't we imagine, in a future analysis, model comparisons to interpret these patterns more finely? [Background selection *versus* selective sweep versus background + sweep] x [isolation *versus* heterogeneous migration] and then quantify the parameters of the different forces?
The authors have everything to perform such an analysis (recombination map + patterns along a divergence gradient), and a future study on this would be impactful in speciation genomics.

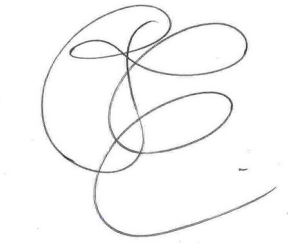
My various comments are mainly about presentation, especially in the introduction. There is a whole literature already discussing the distinction between background selection and selective sweep to explain genomic variation in molecular diversity. It is thus surprising to disconnect interspecific analyses as performed here from what we have learned with intraspecific studies. In particular, the recent studies of Peter Keightley finely quantifying the

Université de Lille
cnrs dépasser les frontières
Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

relative contributions of these forces in genomic patterns of (intra) diversity. So I think some of the points in the introduction might consider a bit more the conclusions drawn from intraspecific diversity analyses.
Still on the introduction, it seems to me to be written too much for population genomists and may seem obscure to people outside the discipline. I thus note in my remarks below various points that could benefit from a little pedagogical effort.

In conclusion, the authors provide a detailed picture of divergence in poplar, with a pioneering analysis of such a process along a divergence gradient. The comparison of scenarios to explain the described patterns is so far still verbal, but could be the main subject of a future statistical study. The work done so far is impressive and I can only recommend this paper as a reference for future analyses of divergence in a given clade at such a fine scale.

Camille Roux



Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

Introduction

The introduction relies, in my opinion, too much on recent empirical studies. In particular, the first two paragraphs provide a review of the genomic patterns of differentiation based solely on the empirical literature. It is a pity not to mention theoretical expectations on such variation, which have been previously established.
This gives the impression that the discipline (evolutionary genomics) draws generalizations from one-off observations on a small number of organisms, whereas all these observations are the result of experiments whose results were anticipated as early as the 1970s.

- Lewontin, Richard C., and Jesse Krakauer. "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms." Genetics 74.1 (1973): 175-195.
- Barton, Nick, and Bengt Olle Bengtsson. "The barrier to genetic exchange between hybridizing populations." Heredity 57.3 (1986): 357-376.

The same is true for the empirical literature, which is composed only of recent papers, whereas they are almost incremens of seminal papers. In 1989, for example, a negative relationship was already highlighted in Drosophila Melanogaster between heterozygosity and linkage disequilibrium.

- Aguade, Montserrat, Naohiko Miyashita, and Charles H. Langley. "Reduced variation in the yellow-achaete-scute region in natural populations of Drosophila melanogaster." Genetics 122.3 (1989): 607-615.

As well as 3 years later with other molecular markers:

- Begun, David J., and Charles F. Aquadro. "Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster." Nature 356.6369 (1992): 519-520.

The interpretations at the time were still from the perspective of molecular hitchhiking, but the observation is not new and should not be overshadowed by many more recent but incremental references.

Line 96: I think we should provide some explanatory detail on the mechanism linking balanced selection to interspecific divergence here. It is really not obvious to the reader how selection pressure maintaining transpecific polymorphism over large evolutionary periods can at the same time locally increase interspecies divergence.

Figure 1:
I don't understand the title of the "scenarios" column whose associated graphs look more like molecular patterns. On the other hand, the underlying demo-genetic scenario is not visible on the figure which would be worth showing this information or, at the very least, indicating the name of the scenario.
For each statistic displayed, a genomic peak or trough is shown, but without information in

Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

the figure (or legend) about the target under selection.
Scenario 1: Is the diversity dip the result of a recent sweep (before returning to mutation-drift equilibrium)? Or is it the result of recurrent selection against maladapted migrant alleles at the target under selection, with an increase in diversity
by introgression?

Regarding the negative correlation between D_{xy} and pi, that doesn't seem intuitive to me either. I still have in mind the expectation of the HKA test which is a positive correlation between diversity (neutral) and divergence (neutral). So what is the cause here of a negative correlation?
Scenario 2: The way in which the lack of relationship between D_{xy} and rho/pi/F_{ST} is represented is unusual and may confuse some readers. It suggests a single divergence value in the genome. While it is well understood that this is the average expected value, the associated variance is gigantic, unless the lineage sort is complete and the common ancestor has a minimum size.

On B's panel, I don't understand why the correlation between recombination and D_{xy} is 0.5. Same remark for the correlation between recombination and pi.

In the legend of the same figure 1, it would be important to indicate the origin of such expectations. Analytical expectations? Simulations? Verbal expectations?

Line 130: *D_{xy} is sensitive to ancestral polymorphism, its values are expected to remain relatively stable under this model*.
I suggest a slight clarification. In scenario 2, D_{xy} does indeed depend on the age of the population split as well as the time of coalescence in the ancestral population. But the genomic variance of D_{xy} is impacted by the ancestral size squared. I think the figure would be clearer by removing all genomic variation in D_{xy} in Figure 1 (panel A, square 2, red line) altogether by specifying somewhere that this is the expected coalescence time. Otherwise, the reader may wonder what would be the source of the variation in D_{xy} shown here, having in mind the quadratic relationship between effective population size and variance in coalescence times.

Line 136-137: I understand that for reasons of visibility, the authors represent on the same scale the effects of directional selection and balancing selection on surrounding diversity, but it is important to recall at some point in the text that the genomic impacts are not the same, at all. Finally, the genomic impact of balancing selections (overdominance, sex determinism, negative frequency dependent selection) is on regions under selection (especially if associated with suppression of recombination), but with patterns that rapidly fade away as one moves away from the directly selected target.

- Hudson, Richard R., and Norman L. Kaplan. "The coalescent process in models with selection and recombination." Genetics 120.3 (1988): 831-840.
- Kirkpatrick, Mark, Rafael F. Guerrero, and Samuel V. Scarpino. "Patterns of neutral genetic variation on recombining sex chromosomes." Genetics 184.4 (2010): 1141-

Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

1152.

- Roux, Camille, et al. "Recent and ancient signature of balancing selection around the S-locus in Arabidopsis halleri and A. lyrata." Molecular biology and evolution 30.2 (2012): 435-447.

Lines 142-144: maybe I'm wrong, but it seemed to me that if background selection would be the force that best explains the sharing of differentiation patterns between different pairs of species/populations of the same genus, it was notably because recombination maps are globally conserved at this time scale. I don't understand how the frequency of occurrence of deleterious mutations would explain the sharing of genomic patterns of differentiation, if this is the case then the authors should detail this relationship a bit better.

While it is true that new mutations are mainly deleterious (especially if organisms are at phenotypic optimum), it does not seem to me to diminish the importance of selective sweeps on the grounds that backgroud selection is a recurrent process whereas sweeps would be evolutionarily one-off events. Nevertheless, there is ample empirical evidence that the trough of diversity around non-synonymous substitutions is consistent with sweeps but not with background selection (in *Drosphila simulans* and in *Capsella grandiflora*).
Sweeps are also very common in the genomes of organisms with small population sizes such as humans (Enard et al, 2014).
Some authors would therefore not appreciate the fact that the role of selective sweeps in genomic variation in diversity, and thus, in differentiation, is so obscured (Elyashiv et al, 2016 for instance).

- Sattath, Shmuel, et al. "Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans." PLoS genetics 7.2 (2011): e1001302.
- Enard, David, Philipp W. Messer, and Dmitri A. Petrov. "Genome-wide signals of positive selection in human evolution." Genome research 24.6 (2014): 885-895.
- Elyashiv, Eyal, et al. "A genomic map of the effects of linked selection in Drosophila." PLoS genetics 12.8 (2016): e1006130.

Line 144: A reference is missing here to explain the conservation of differentiation landscapes as a consequence of polygenic adaptation.

Lines 144-147: The most powerful tool for understanding the relative importance of evolutionary forces on genomic variation in diversity is ultimately the estimates of fitness effect distributions (DFE). Such a distribution of the effects of deleterious and advantageous mutations allows one to generate theoretical expectations, and thus test, the relative effects of these forces.
It might be more relevant to cite work along these lines (Booker and Keightley in *Mus musculus castaneus* for example) before mentioning Burri, 2017a and Stankowski, et al. 2019, without seeking to reduce even the major contributions of these two studies.

- Booker, Tom R., and Peter D. Keightley. "Understanding the factors that shape

Université de Lille
cnrs
Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

patterns of nucleotide diversity in the house mouse genome." Molecular biology and evolution 35.12 (2018): 2971-2988.

Results and Discussions

A geographical map with samples and suspected ranges for the 8 species would have been welcome before Figure 2 to accompany readers who do not follow the literature around *Populus*.

All figures :
For more readability I propose to replace all labels **pgra, palb, pade, prot, ptma, pdav, ptmu** by **gra, alb, ade, rot, tma, dav, tmu**.

Line 201: two-three lines of reminder about the analyses performed with the KING toolset would be welcome (**1.** Type of data needed; **2.** Test performed).

Line 230 + Figure 2-D: I do not understand what is mentioned here by IBD. Regarding the 10 inter-specific clusters, are these segments inferred to be from introgression by a given method? Is the length mentioned an average length? The total sum?

Line 238: "*Future investigations on the plant grey zone based on a large number of taxa are needed to get more general insights about plant diversification*". This would indeed be a project that deserves ambitious funding.

Figures 2 and 3:
I find that figure 2-D is buried in clustering analyses, which do not need this figure to be convincing. The analyses performed are convincing in themselves.
Perhaps Figure 2 could be limited to 2-A + 2-C.
And that a new figure 3, focused on the topology of genealogy + inferences on introgression could be proposed with 2-B + 3-A + 2-D.

Figure 3-b:
What do the vertical dotted lines in the figure MSC++ correspond to?
A new figure 4 with 3-B + 3-C + 4-C would make sense to me because Tajima's D as well as intra-specific diversity (pi) are consistent with MSC++ inferences. It's nice, so you might as well put them side by side.

To validate the inference made by MSC++, I suggest simulating the (intra-specific) history of the ~30 million SNPs under the inferred model to see if the authors can reproduce the 8 Tajima's D distributions shown in Figure 3-C.

Line 316:
"The correlations of F_{ST} between independent species pairs become stronger with advancing differentiation. This may be the case that the effect of linked selection accumulates with differentiation advance."

Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr

Couldn't an increase in correlation simply be explained by a reduction in genomic variance? If the genome is homogeneous for an F_{ST}=1 (or an F_{ST}=0) then shouldn't a strong correlation be expected? With partial decorrelations during the differentiation process, and this, in a purely neutral model?

Figure 5:
I propose to replace **pdavprot, ptmaprot** with **dav-rot, tma-rot** for more readability.

Figure 6:
As for the article Stankowski, Sean, et al. "Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers." PLoS biology 17.7 (2019): e3000391. , I suggest putting all axis-y between -1 and 1 for ease of understanding.

Pannel-C: I don't understand figure 6-C. How could the "polymorphism - recombination rate" relationship, that is, the relationship between two intraspecific variables, be impacted by the level of divergence with another lineage? Finding an effect of divergence seems to me more of a concern than a result. On this point, I await clearer explanations from the authors on the expectations of such a relationship as well as on the mechanisms that could justify it.

Bâtiment SN2, Cité Scientifique
F 59655 Villeneuve d'Ascq Cedex – FRANCE
TEL : +33 6 30 71 19 02
E-mail : camille.roux@univ-lille.fr
http://eep.univ-lille.fr