Dear Recommender,

Thanks you and the two reviewers for the comments on our manuscript. You'll find hereafter our point by point response to all the points raised as well as indications of modifications in the text when applicable. These modifications are also visible in the revised version of the manuscript. We hope this version of the manuscript will meet your expectations, as well as those of the reviewers.

Best regards,

Vincent Segura on behalf of all the co-authors.

****

Review by anonymous reviewer 1, 14 Dec 2024 03:51

In their work, the authors analyze the relationship between variants associated with the chalcone isomerase gene and the diversity existing among different /Populus nigra/ accessions and its effect on diameter growth. Their system biology analysis approach considered integrating phenotypic and transcriptomic data, GWAS, and population genetics analyses. The subject and the approach developed by the authors are very interesting, given the importance of /P.nigra/, and the need to characterize the mechanisms that control growth traits. Overall, the manuscript is well-structured and well-written. However, the authors must improve some methodological aspects to complement the information provided and the interpretation of the results obtained. My main concerns deal with how the analyses handled the genetic structure of the association population, the lack of correction of false positives in the analyses, and the very high phenotypic variation explained by individual SNPs.

Thanks for your comments on our manuscript. Please see our responses hereafter for your specific questions about how analyses were handled.

Specific comments

Title:

-The concept of "growth" should be more specific. Radial stem growth was studied.

-Replace "between" with "among".

Thanks for these suggestions, we have changed the title: "Natural variation in chalcone isomerase defines a major locus controlling radial stem growth variation among Populus nigra populations." (lines 1-5)

Introduction:

-This section should be complemented with information about the candidate genes on which the analyses were done.

The introduction highlights the importance of using integrative approaches in the context of genomic studies. No previous study specifically highlights the CHI gene, which is why we believe that this information is more appropriate as currently in the discussion section of the manuscript. We have thus included extra information about the gene in the discussion (lines 470-476, 478-483).

-The hypothesis that guided the study should also be made explicit.

The hypothesis that guided the study is that omic data allow to dissect the genetic architecture of complex traits such as secondary growth. This was already presented at the end of the introduction (lines 115-118). We have added some the text to further clarify this point (lines 119-123).

Materials and Methods:

-L131-132. Please include the geographical coordinates for ORL and SAV here (It should be optional to search for this information in other articles).

Geographical coordinates have been added in the text (lines 134-135).

-L135-136. Briefly explain the criteria used to select this subset of genotypes.

We have added more details on the selection of the subset of genotypes (lines 139-144).

-L140. Indicate why the climatic data from that particular range was included.

We apologize, there was a mistake with the version of WorldClim used. The database used for our analysis was the version 2.1 of WorldClim (1970-2000). We used this database because WorldClim is a reference for this type of data (more than 9k citations) in offering standardized mean climate data for a specific period. We have corrected this information in the text (148-149).

-L142-143. Indicate how much of the total variation was explained by PC1 and PC2.

The percentage of variance explained was 42% for PC1 and 22% for the PC2. We have added this information (lines 150-151).

-L146. About the concept of “circumference,” please explain here what specifically is referred to: perimeter, radius, or diameter. Usually, in forestry studies, the latter is used. Please also indicate how it was measured.

The circumference corresponds to the perimeter of the stem measured at 1 m with a measuring tape. We have added this information (lines 157-158).

-L146. Please check or explain the concept of “Infraden”.

We have replaced this abbreviation by the term “wood basic density” throughout the text.

-L156. Clarify if the square root transformation was used for circumference and basic density.

Thanks for spotting this, the transformation was only required for circumference, not for basic density. We have added this information to the text (lines 167-168).

-L167. Indicate the source/platform where this genomic sequence was available. Phytozome?

This information has been added (lines 178, 189).

L178-179. Describe which specific callers were used.

We have added this information (lines 190-192).

Include a description of the variance partitioning models mentioned in Figure 3.

The model is already presented in line 205. However, we agree that some information is missing. We made use of the relationship between multi-trait and multi-environmental models (Itoh and Yamada, 1990) to compute the variance components. This information has been added to the text (lines 213-215).

L195-L196. Describe how these variance and covariance components were obtained.

We have added the R package used to fit the model (line 204).

L196. Explain the type of genetic relationship that was considered for obtaining the kinship matrices. IBD, IBS, other?.

The kinship matrix used is an average allelic correlation matrix weighted by the linkage disequilibrium between SNPs as proposed by Speed et al. (2012).

L198. Clarify why two kindship matrices were averaged if the association population was only one (the same set of accessions was established in two common garden trials).

The association population is structured in subpopulations (Faivre-Rampant et al. 2016). So to fit the model presented in line 205, the global kinship matrix was split into a between-population genomic relationship matrix consisting of an average kinship coefficient per population and a within-population relationship matrix with kinship coefficients different from zero within populations and equal to zero otherwise. This description was already present in the manuscript (lines 210-212).

-L200-L203. Please add a Hardy-Weinberg test for the SNPs analyzed. It is important to know whether these markers are in gene equilibrium and to determine if they are neutral or reflect some adaptive process.

We did not perform the Hardy-Weinberg equilibrium test, because the panel of 241 genotypes under study consists of several subpopulations (some being isolated from the others). Thus, it is not expected to be at panmixia, which is a requirement for Hardy-Weinberg equilibrium.

-L205. "...GWAS was performed for circumference in each site with genotypic adjusted means..." Explain why basic density was not considered in these analyses.

Basic density was not considered in these analyses because the focus of the paper is stem radial growth. Basic density was in fact considered as a secondary trait to complement the findings on circumference.

-L205-L213. Indicate if the association models included any factors to model population structure (Q-matrix). If not, explain why.

Population structure confounding was accounted for in the GWAS model by the random polygenic term (with a covariance determined with the genomic relationship matrix). We did not include population structure as a fixed effect for several reasons. First, it is usually estimated with the same genomic data as the kinship matrix and consequently, such information appears to be somehow redundant leading to a potential loss of power. In addition, the phenotype under study (Circumference) being particularly differentiated between populations, using a fixed effect of population structure in the model would clean the association. We already recognized in the discussion that such a situation is difficult to handle, which is why we have decided to include some kind of validation for instance in another population (controlled crosses).

Describe how the percentage of phenotypic variation that accounted for the significant SNPs was calculated.

The previously reported R2 did not account for population structure confounding and thus were overestimated and probably misleading. We have now re-estimated the R2 in the mixed-model following (Sun et al., 2010 https://doi.org/10.1038/hdy.2010.11). The corresponding values have been modified in the manuscript (lines 262, 265).

From what was described, a test to control for false positives (e.g., FDR, Bonferroni) was not applied. It is essential to do this test to validate the character-SNP associations identified as significant. Please include it.

We indeed used the Bonferroni correction to account for multiple testing. This information has been added in the method section (lines 228-229).

L214-215. "...GWAS were also carried out using transcriptomic data (eQTL analysis) but focusing only on two genes of particular interest in this work..." Explain how these two genes were chosen or from which analysis they came.

These 2 genes included significant SNPs in the GWAS. This information has been included in the text (line 232).

L220. "...parental population of P. nigra...". Explain if there is any relationship with the accessions used for the GWAS.

There is no relationship between this parental population and the genotypes used for the GWAS. This parental population was used as a validation (included in the text, line 237).

L223. "...using a simple linear model...". Describe what this model consisted of.

This model simply tested for association between the phenotype and each of these SNPs separately, without any other term. The model has been added to the text (line 242).

Results:

L227-229. At this point, the specific density should be mentioned.

We have added an introductory paragraph to the results section to recall the general strategy and thus explain why the focus of the work is tree circumference and not any other trait such as specific density (lines 244-252).

L230-231. "...Potri.010G212900, annotated as a Beta-Hexosaminidase 1 (Hexo1)...". Check this annotation. It seems that is not the one currently in Phytozome13 for Ptricho3.0.

Thanks for spotting this. We initially found this annotation (Hexo1) in McKown et al., 2014 ; but you're right: it rather seems to be a protein of unknown function (PUF) in Phytozome 13. We have thus corrected this all along the manuscript.

-L234-236. More than 20% or 50% of the phenotypic variance is very high for an individual SNP, considering that girth is a continuous/metric/complex trait. Please elaborate on "without accounting for population structure" and its relation to such percentages.

We agree and have thus re-estimated the R2 from the GWAS model (see our previous response).

- L235. “...Although not significant in the ...”. Indicar el umbral de significancia estadística (alfa).

We have included the corresponding p-value to the text (line 264).

-Incorporate a reference to the Hexo1 gene.

Following your previous comment, we have renamed the gene as PUF.

Figure 1:

- Indicate what “r2” means at the top of panel b).

We have included the information in the legend of the figure (lines 279-281).

- In panel c), indicate the molecular genotype associated with codes “0”, “1” and “2”.

We have included this information in the legend of the figure (lines 284-286).

- Replace the coefficient of determination “R2” with the correlation coefficient “r” to be consistent with the concept of correlation used in the text. Also, add the p-value for each coefficient.

We have removed the R2 from the boxplots of this figure.

-L261-269. Rewrite using the correlation value “r” instead of “R2”, which, although associated, represents the coefficient of determination.

Otherwise, the level of relationship between both variables is underestimated.

Thanks for this suggestion, we have modified the figure and the text to have r instead of R2, with the corresponding p-values (lines 292-293, 295-297, 301-302).

Figure 2:

-Panels a) and b). Indicate the correlation coefficient instead of the determination coefficient. Add the p-value for each coefficient. Add the title and unit to the x-axis. For clarity, replace gene model names with gene codes (CHI or Hexo1).

We have made the modifications. Note that there is no unit for the gene expression because it was estimated from a linear mixed-model model (BLUP value).

-Panels c) and d). For clarity, replace gene model names with gene codes (CHI or Hexo1).

We have included the gene codes together with the gene model names (CHI and PUF).

- In the description of the figure, and concerning “...The expression level of transcripts have been standardized with a genetic analysis...”, briefly mention what type of analysis you are referring to.

We have added this information in the figure legend (lines 318-319).

L293-294. “...Interestingly, when the top SNP was included as a fixed effect in this variance partitioning model, it explained up to 24% of the total phenotypic variance...”. Usually, when a factor is defined as fixed within a model, its variance is a constant (a specific numerical

value), and it has no associated variance component (this is only estimable for random factors). Please explain how this contribution to the total variance was estimated.

You are right that no variance component was associated with the SNP because it is a fixed effect. Yet, it is possible to estimate the percentage of variance explained by fixed effect in mixed models, as usually done in association studies (Sun et al., 2010).

Figure 3:

What is described at the bottom of the figure should be mentioned in detail in the Materials and Methods section.

The corresponding information has now been included in the materials and methods (lines 213-215).

Figure 4:

Add the units for the Y-axis variables for panels a), b), and c).

We have increased the size of the already existing text and included more information in each of the plots of the panel c).

Figure 5:

For panels a), b) and c), indicate the correlation coefficient instead of the coefficient of determination.

We have replaced the R2 by the correlation coefficients r with their corresponding p-values.

Discussion:

-L357. "...We made use of growth data collected...". specify that growth data are referred to stem diameter/circumference. Please include a supplementary table with the range of values (untransformed) and averages for the accessions evaluated (grouped by origin/populations).

We have made the suggested table (Table S2).

-L371-381. Include a mention of the fact that circumference measurements and transcriptomic analyses were performed on trees of different ages.

We have included this information in the text (line 422).

-L382-388. Explain why models with a factor for population structure (Q matrix) were not considered.

A sentence has been added to explain this point (lines 436-438).

-L393. Indicate the meaning of the acronym "GRM".

The corresponding meaning has replaced the acronym (line 435).

-L409-414. Please include in this part of the discussion the fact that some of the top SNPs (Table S1) are part of nucleotide triplets that correspond to stop codons (Chr10-20120172) or involve nucleotides very close to them (Chr10-20120195). When using Phytozome tools to explore the CHI gene (Potri.010G213000), it can be seen that there are two alternative transcripts for that gene: Potri.010G213000.3 and Potri.010G213000.2. The latter is shorter

(the last exon would be missing), most likely due to the presence of SNPs related to stop codons, which would imply that a variant is associated with a truncated and probably non-functional protein.

Thanks for this suggestion, that has been added to the text (lines 452-457).

-L417-422. The results discussed here should also have been mentioned in the Results section.

We have moved the text to the Results section (lines 390-396).

****

Review by Gancho Slavov , 02 Dec 2024 21:41

This is a well written manuscript based on what looks like a carefully collected and extensive data set. The crux of the paper are the GWAS and eQTL analyses, and I have a few technical questions, the answers to which will largely determine my overall assessment:

1) As the authors acknowledge in the Introduction, there is strong population structure in black poplar. What exactly was done to control for that, as well as for confounding caused by the presence of close relatives, in the GWAS analysis? I am assuming the authors used standard mixed linear model methodology, including a kinship matrix and principal components, but this needs to be spelled out.

As previously mentioned in one of our responses to reviewer 1, we used only a random polygenic term to account for population structure confounding. Because the target trait is heavily structured in our panel (Qst between 0.3 and 0.5), if we were using a fixed effect of population structure in the model the signal would vanish. Also please note that in such a standard Q+K model, the information used to build the corresponding matrices is the same (genotypic data) leading to some redundancy and potential loss of power. Yet, we recognize that this situation is not ideal, and discuss this point (lines 424-438). Also, we would like to point out that other analyses are carried out and our conclusions do not simply rely on the GWAS analyses.

2) How much confounding was still left after applying the methodology referenced in 1)? As a minimum, QQ plots and/or lambda (mean genomic control test statistic) need to be reported.

We now provide QQ plots for all the GWAS (Figs S1, S2, S3, S4).

3) Similar questions about the eQTL analysis…

The exact same analyses were carried out, and QQ plots have been added to the supplementary.

4) The Fst analysis and correlations to climate/geography do very little to reassure me that this association is not an artifact of population structure. There are likely to be thousands of SNPs with similar properties across the black poplar genome (https://www.biorxiv.org/content/10.1101/2024.10.11.617670v1). Would the association hold in a within-population analysis? Is the PVE realistic given what is known about Beavis effects and complex traits?

We do agree with the reviewer that the situation is not so reassuring because the phenotype under study is highly structured. Regarding within-population analyses, we have shown that the association holds within a *P. nigra* pedigree. For the R2 we have made corrections to report more realistic values from the mixed-model (lines 262, 265).