Comparison of the population genetics results produced by Li et al. and the Untwist project

Asis Hallab, group Björn Usadel,
Institute of Bio- and Geosciences (IBG), Bioinformatics (IBG-4)

July 23rd 2024

1 Introduction

Li et al. [1] carried out population genetics and phylogenetics analyses similar to the analyses done within the Untwist project. In this document we explore differences in the methods and discuss the results, especially in the light of scientific robustness.

Additionally, in the progress of the population genetics analysis, parameters were refined and optimized to yield biologically meaningful results and ensure maximum statistical confidence. A previous version of figure 1 (see below) shows a slightly different dendrogram as the latest result. This differences are discussed here, too.

2 Filtering of SNP marker (sites)

Identified genetic polymorphic sites (marker) require filtering in order to obtain biologically meaningful results with maximum statistical robustness and confidence. In this, both Li et al. as we applied an almost identical filtering pipeline based on `vcftools` and `bcftools`. In the following the differences in both pipelines are briefly discussed.

- We use a minor allele frequency cutoff of 0.05, while Li et al. use a slightly more stringent one of 0.1.
- We test for linkage disequilibrium in a window of 10kb, while Li et al. do not provide any information on this. However, they measure the distance at which the maximum Rˆ2 is halved to be 593 kb. Whether they used this value as a window for the LD correlation filtering remains unclear. The value seems a bit large to have been used in the filtering, as it has been assessed within the context of the Quantitative Trait Loci (QTL) analysis and is used in it as a parameter.
- We still allow a correlation between markers <= 0.9, while Li et al. set the cutoff at 0.4.
- We and Li et al. filter markers by heterozygosity <= 0.5. However, Li et al. do this with bcftools and we do it with a Java program called `vcffilterjs`. However both declare to implement the same filtering.

2.1 Adjustment of filtering parameters

As stated in the introduction, in order to obtain more robust and biologically meaningful results the filtering of SNP sites was adjusted and made more stringent. Especially the a-priori knowledge about the untwist lines sampling the European genetic diversity of Camelina populations well was used as a measure of confidence.

The iterative adjustment of these filtering parameters produced intermediate results and plots. Especially, two different ADMIXTURE [2] and clustering genetic similarities results differ due to one having been produced *before* and the other *after* filtering out markers (sites) with high heterozygosity.

2.2 Comparison of intermediate and final ADMIXTURE and hierarchical clustering results

A slight differences between the ADMIXTURE [2] and genetic similarity dendrogram generated *before* and *after* filtering out markers (sites) with high heterozygosity. See figures one and two.

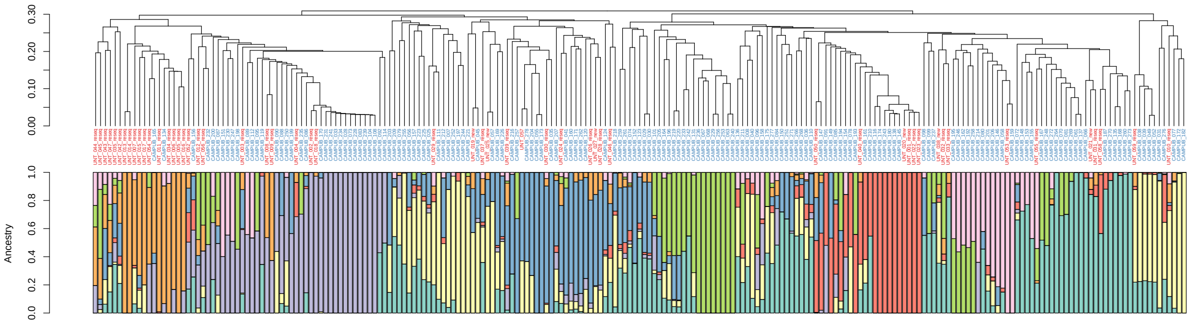

Figure 1: ADMIXTURE and genetic similarity dendrogram plot *before* filtering out sites with high heterozygosity (>= 0.5)

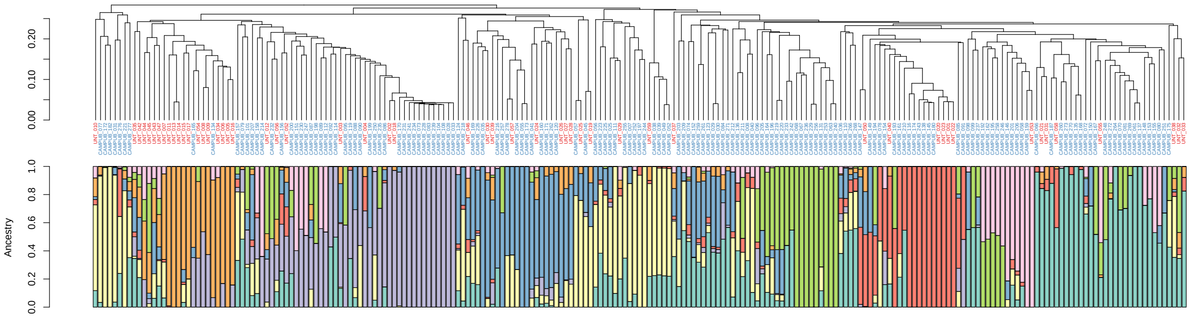

Figure 2: ADMIXTURE and genetic similarity dendrogram plot *after* filtering out sites with high heterozygosity (>= 0.5)

In general the resulting grouping (dendrogram topology) is highly similar for the hierarchical clustering done *before* (figure 1) and *after* (figure 2) filtering out sites with high heterozygosity.

For example, the accession `UNT_010` groups in both dendrograms with `CAMPUP_077`, `CAMPUB_172`, and `CAMPUB_182`. However, *after* filtering out highly heterozygous sites, the sister group comprises `CAMPUB_031`, `CAMPUB_276`, `CAMPUB_221`, and `CAMPUB_277`, while before heterozygosity filtering it only contained `CAMPUB_031` and `CAMPUB_276`.

Another example is visualized in figure 3, where for six Untwist lines the general genetic relationship is reproduced after filtering out highly heterozygous sites, while some siblings changed their position.

3 Number of clusters in the population genetics analysis

Li et al. ran ADMIXTURE [2] with different numbers of assumed populations k = 2,3,...,30+ evaluating the cross-validation error as a quality measure of the fit of that number k for the genetic marker data provided

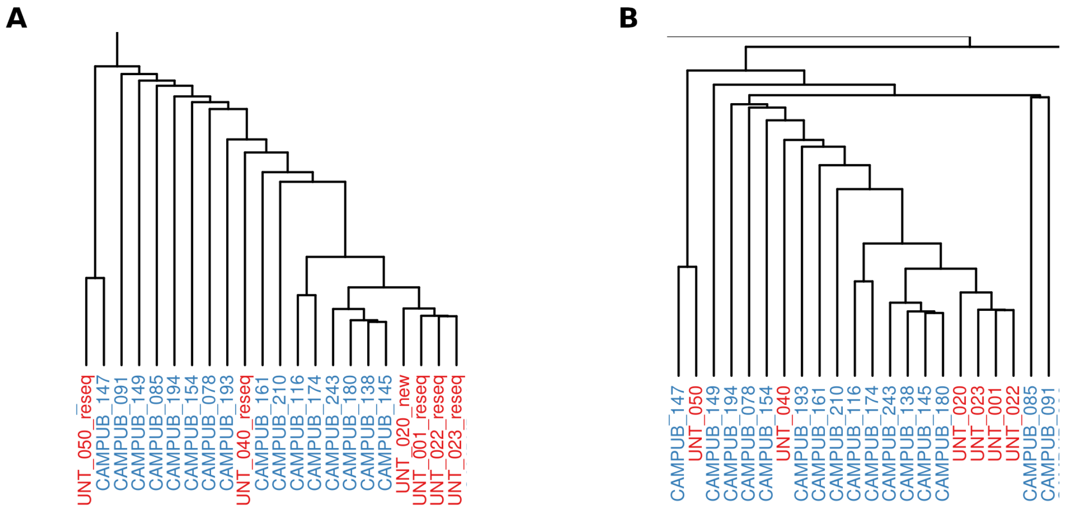

Figure 3: Comparison of a cluster around six Untwist lines before (A) and after (B) filtering out highly heterozygous sites. Most of the grouping topology is retained, while a few public *Camelina* lines changed their genetically closest siblings.

(figure 4). We applied the same methodology evaluating the cross-validation error for $k = 3,4,\ldots,12$ different populations (figure 5).

In order to compare our ADMIXTURE results with the ones published by Li et al. compare figures 6 and 7.

4 Methods, Data, and scientific plot availability

All source code used to generate the population genetics and related results can be found on GitHub: github.com/usadellab/untwist

All scientific visualizations discussed here are also shared in the joint Untwist Google Drive

5 References

[1] Li, H., Hu, X., Lovell, J. T., Grabowski, P. P., Mamidi, S., Chen, C., Amirebrahimi, M., Kahanda, I., Mumey, B., Barry, K., Kudrna, D., Schmutz, J., Lachowiec, J., & Lu, C. (2021). Genetic dissection of natural variation in oilseed traits of camelina by whole-genome resequencing and QTL mapping. The Plant Genome, 14(2), e20110. https://doi.org/10.1002/tpg2.20110

[2] Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

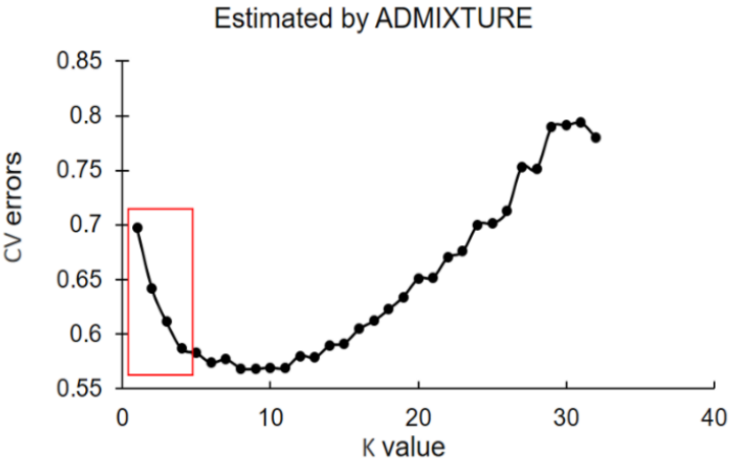

Figure 4: Li et al. ADMIXTURE cross-validation errors for k=2,3,. . . ,30+ assumed populations. The scatterplot shows the lowest cross-validation error for k=8. None the less Li et al. selected k=4 populations in order to represent the geographic sampling.

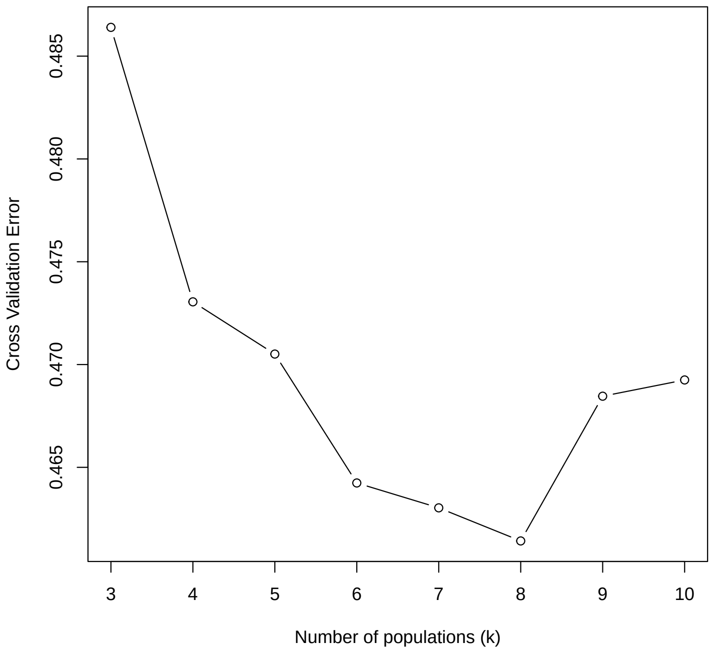

Figure 5: Untwist ADMIXTURE cross-validation errors for k=3,4,. . . ,12 assumed populations. The scatterplot shows the lowest error for k=8 populations. For this reason, k=8 was selected as the best fitting number of populations to explain the genetic diversity of *Camelina* lines in the Untwist project

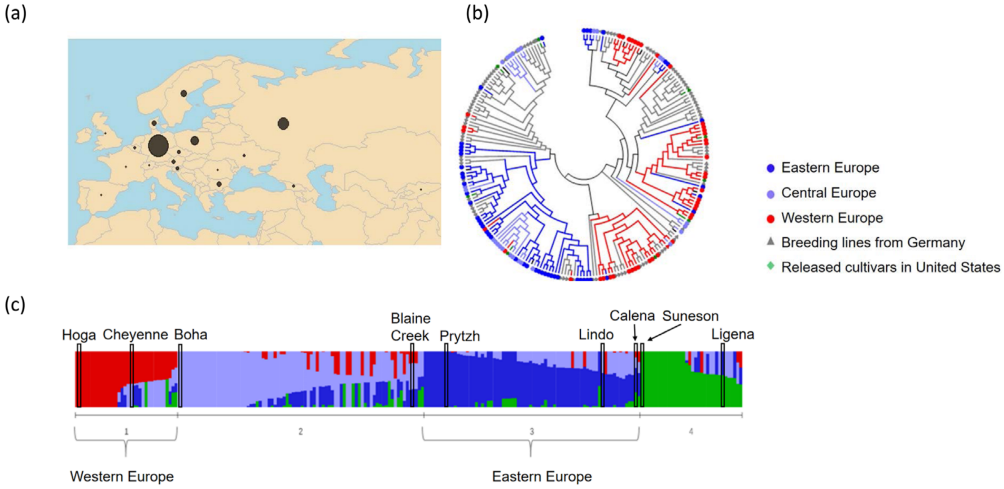

Figure 6: Original "figure 1" from the article of Li et al. [1] showing the ADMIXTURE and phylogeny of European *Camelina* lines for k=4 assumed populations

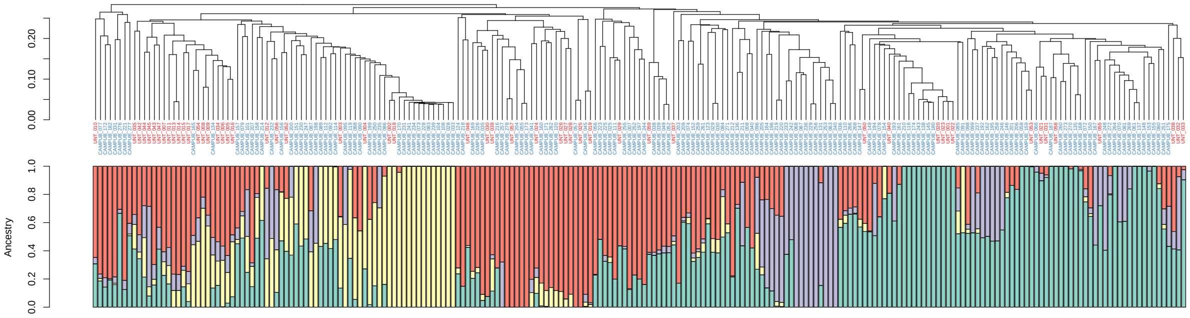

Figure 7: ADMIXTURE and hierarchical clustering of genetic marker based similarity for k=4 assumed populations as produced within the Untwist project