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

Asis Hallab, Institute of Bio- and Geosciences (IBG),
Bioinformatics (IBG-4), Forschungszentrum Jülich

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 R^2 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.

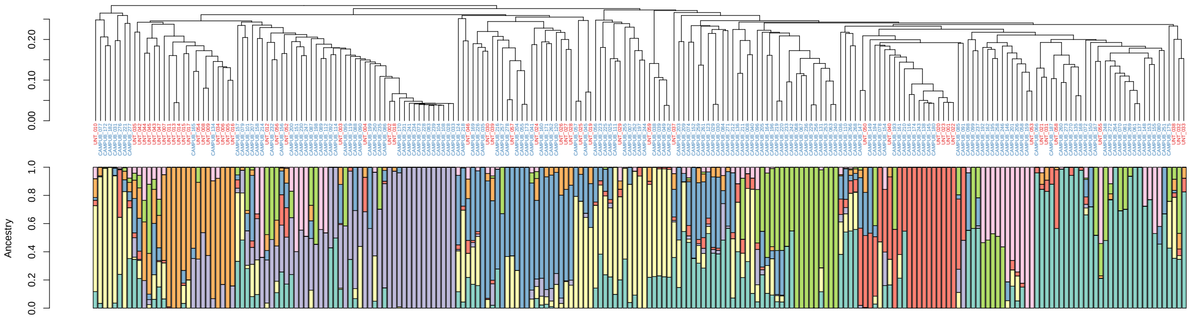

Figure 1: Final ADMIXTURE and genetic clustering dendrogram

2.2 Comparison of intermediate and final ADMIXTURE and hierarchical clustering results

3 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

4 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