# Supplementary methods

Alignment and SNP calling are often challenging in conifers as due to the repetitive nature of their genomes (Wegrzyn et al., 2014), leading to errors in read alignment. Alignments of individual samples visualized with IGV (Thorvaldsdottir, James, & Jill, 2012) revealed that large proportion of target area contained not only reads with probably correct alignment (few or no mismatches to reference genome) but in addition reads which were likely incorrectly aligned to the area (multiple mismatches and gaps). This suggests that the off-target, paralogous sequence had been captured along with the target sequence.

Areas containing paralogous sequence may lead to spurious SNP calls inflating the number of SNPs and distorting the shape of the allele frequency spectrum. Identification of such areas is possible since overlap of correctly and incorrectly aligned reads may be detected as heterozygous genotype calls, which are not expected with haploid DNA that was used for sequencing. To alleviate the incorrect alignment issue the SNP calling step was performed twice. The first SNP call to detect spurious heterozygous areas was performed with freebayes (Garrison & Marth, 2012) using parameter --ploidy 2 to detect heterozygous variant calls. Other parameters were set as -T 0.01 --min-coverage 5 -u -X. Areas containing heterozygous genotypes and nearby areas with radius of 25 bases were then marked to a bed file for later filtering as paralogous areas may also contain spurious homozygous SNP calls which should be removed. Visualizations of alignments indicated that paralogous areas were often saturated with heterozygous SNP calls hence a filter radius of 25 bases was deemed sufficiently large to filter out possible spurious homozygous calls. A bed file defining non-paralogous areas was then created by generating a bed file for the whole *P. taeda* reference genome and removing the paralogous areas with bedtools (Quinlan & Hall, 2010) subtract command.

Variant calling was redone to achieve the final set of SNPs with the same parameters as in the first call, but --ploidy 1 argument was used instead and the bed file defining non-paralogous areas was provided as target area with -t parameter. As some downstream analysis requires information on the length of the sequence available for analysis (henceforth referenced as ‘available genome’) a VCF file containing genotype calls including monomorphic sites was generated by including --report-monomorphic parameter to freebayes command line. As calling monomorphic genotypes for large number of samples requires excessive amount of RAM and a long runtime, the SNP calling was performed only for scaffolds longer than 1000 bases. This removes only 14.3 % of total sequence but the total number of scaffolds is reduced from over 14 million to one million greatly reducing the computational runtime of downstream analysis. No exome capture baits had been designed for the omitted areas. The genotype calling was parallelized by performing the run separately for each sample in its own thread. The produced VCF files were then combined with vcflib vcfcombine tool.

## References

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing, 1–20. doi:arXiv:1207.3907 [q-bio.GN]

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Thorvaldsdottir, H., James, T., & Jill, P. (2012). Integrative Genomics Viewer ( IGV ): high-performance genomics data visualization and exploration, *14*(2), 178–192. doi:10.1093/bib/bbs017

Wegrzyn, J. L., Liechty, J. D., Stevens, K. a, Wu, L.-S., Loopstra, C. a, Vasquez-Gross, H. a, … Neale, D. B. (2014). Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. *Genetics*, *196*(3), 891–909. doi:10.1534/genetics.113.159996