Strong Purifying Selection in Haploid Tissue–Specific Genes of Scots Pine Supports the Masking Theory

Sandra Cervantes [1,2] Robert Kesälahti [1] Timo A. Kumpula [2,3] Tiina M. Mattila [4] Heikki Helanterä [1] and Tanja Pyhäjärvi *[,5]

[1]Department of Ecology and Genetics, University of Oulu, Oulu, Finland
[2]Biocenter Oulu, University of Oulu, Oulu, Finland
[3]Laboratory of Cancer Genetics and Tumor Biology, Research Unit of Translational Medicine, University of Oulu, Oulu, Finland
[4]Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden
[5]Department of Forest Sciences, University of Helsinki, Helsinki, Finland

***Corresponding author:** E-mail: tanja.pyhajarvi@helsinki.fi.

Associate editor: John Parsch

Abstract

The masking theory states that genes expressed in a haploid stage will be under more efficient selection. In contrast, selection will be less efficient in genes expressed in a diploid stage, where the fitness effects of recessive deleterious or beneficial mutations can be hidden from selection in heterozygous form. This difference can influence several evolutionary processes such as the maintenance of genetic variation, adaptation rate, and genetic load. Masking theory expectations have been confirmed in single-cell haploid and diploid organisms. However, in multicellular organisms, such as plants, the effects of haploid selection are not clear-cut. In plants, the great majority of studies indicating haploid selection have been carried out using male haploid tissues in angiosperms. Hence, evidence in these systems is confounded with the effects of sexual selection and intraspecific competition. Evidence from other plant groups is scarce, and results show no support for the masking theory. Here, we have used a gymnosperm Scots pine megagametophyte, a maternally derived seed haploid tissue, and four diploid tissues to test the strength of purifying selection on a set of genes with tissue-specific expression. By using targeted resequencing data of those genes, we obtained estimates of genetic diversity, the site frequency spectrum of 0-fold and 4-fold sites, and inferred the distribution of fitness effects of new mutations in haploid and diploid tissue–specific genes. Our results show that purifying selection is stronger for tissue-specific genes expressed in the haploid megagametophyte tissue and that this signal of strong selection is not an artifact driven by high expression levels.

Key words: **haploid selection, masking theory, purifying selection, gymnosperms, *Pinus sylvestris*, DFE.**

Introduction

The masking theory predicts that the efficacy of selection is stronger in haploid genomes (Kondrashov and Crow 1991) in comparison with diploid, because the number of chromosomal copies in a genome directly affects the efficacy of selection. For genes expressed on diploid genomes, any level of dominance (h) other than 0.5 (additivity) implies that the fitness effect of one allele will be partially or totally masked by the other allele. Consequently, both deleterious and beneficial mutations can be less affected by selection and hide in a heterozygous state. In contrast, genes expressed in haploid genomes will be readily exposed to selection due to the lack of masking (Crow and Kimura 1965; Kondrashov Alexey and Crow James 1991). This difference has important evolutionary consequences as selection acting differently in haploid and diploid genomes will affect the spread and fixation of new mutations and influence genetic load, genetic variation, and adaptation rate (Szövényi et al. 2013; Immler and Otto 2018).

Most of the empirical evidence supporting haploid selection has been obtained through yeast experimental evolution studies (Otto and Gerstein 2008; Gerstein et al. 2011), where haploid organisms adapt faster and have better fitness when exposed to environmental changes, for example, nutrient limitation (Otto and Gerstein 2008 and references therein; Mable and Otto 1998). However, expectations of masking theory can be extended to multicellular organisms with alternation of generations between haploid and diploid phases, such as plants (Immler 2019; Beaudry et al. 2020). In these cases, selection in the haploid stage can help, for example, to reduce the burden of recessive deleterious mutations carried in the diploid stage, or it can lead to the evolution of

Open Access

Article

heteromorphic life cycles such as ecological niche differentiation at haploid and diploid stages, which would maximize resource exploitation (Mable and Otto 1998).

Research on haploid selection in plants has been mostly conducted in angiosperms, particularly in genera within the Brassicaceae family, for example, *Arabidopsis* and *Capsella* (Arunkumar et al. 2013; Gossmann et al. 2014; Gutiérrez-Valencia et al. 2022), where the female gametophytic stage is very reduced and dissection of female structures can be technically difficult (Beaudry et al. 2020, but see Gossmann et al. 2014 and Gutierrez-Valencia et al. 2022). Therefore, most of these studies have compared male gametophytes, typically pollen-expressed genes, to genes expressed in sporophytic tissues. Results of these studies have shown that selection on pollen (haploid)–expressed genes is stronger relative to selection on sporophyte (diploid)–expressed genes (Arunkumar et al. 2013; Gossmann et al. 2014). Studies in other angiosperm systems have shown similar results. In *Silene latifolia* and *Rumex* species, increased efficacy of purifying selection at the haploid stage allows the purging of deleterious alleles of Y-linked genes with expression at the male gametophytic stage, which in turn slows down the degeneration of their Y chromosome (Chibalina and Filatov 2011; Sandler et al. 2018).

However, it is argued that in angiosperms, the increased selective pressure observed in haploid stage–expressed genes carries the confounding effects of pollen competition (Moore and Pannell 2011). The general assumption is that selection has a stronger effect on male haploid stages than on female stages, in part because in plants, the pollen is released to an exterior heterogeneous environment, which implies exposure to varying degrees of environmental selective pressures. In comparison, plant female reproductive structures, gametophyte and eggs, remain sheltered inside the sporophyte (Immler and Otto 2018; Sandler et al. 2018). In addition, most of the female eggs will be fertilized, while only a small number of male gametes carried by pollen will be able to participate in the fertilization (Immler 2019). For example, Arunkumar et al. (2013) acknowledged that it is difficult to disentangle the signal of haploid selection from the signal of sexual selection in pollen of *Capsella grandiflora*.

Moreover, current knowledge on haploid selection mostly represents the evolutionary dynamic in angiosperms and does not represent the variety of length of the haploid stage in other plant groups such as mosses or gymnosperms. To our knowledge, there is only one study in a nonangiosperm species (*Funaria hygrometrica*) where explicit testing of haploid selection has been approached (Szövényi et al. 2013). In *F. hygrometrica*, a moss with an extended haploid phase and short diploid stage, Szövényi et al. (2013) found that haploid-expressed genes have higher sequence variation due to relaxed selection. This finding contradicts the expectations of the masking theory. Szövényi et al. (2013) argue that confounding effects of the evolutionary dynamics of gene expression per se could be the reason for inefficient purifying selection. Level and breadth of expression are both known determinants of protein evolutionary rate, and genes with expression on a higher number of tissues (broad breadth of expression) and genes with high level of expression are under tighter selective constraints (e.g., stronger purifying selection) (Duret and Mouchiroud 2000; Zhang and Yang 2015). Therefore, the low level and narrow breadth of expression of tissue-specific genes relax the selective pressure over them, rendering haploid selection insufficient to purge putative deleterious variation. Similar results have been observed in sperm-expressed genes of *C. grandiflora* and *Arabidopsis thaliana*, where low level of expression has been suggested as the explanation for relaxed selection (Arunkumar et al. 2013; Gossmann et al. 2014).

Here, we used *Pinus sylvestris*, a gymnosperm with high realized outcrossing level, large population size, and low level of genetic structure, as a study model (Pyhäjärvi et al. 2019; Tyrmi et al. 2020; Hall et al. 2021). *Pinus sylvestris* has a haploid megagametophyte stage of approximately 2 years. The megagametophyte is functionally homologous to the endosperm in the seed of angiosperms. However, the origin of the megagametophytic tissue is completely different in comparison with angiosperms. This multicellular structure originates from the meiosis of the megaspore mother cell, which develops in the megasporangia of the ovuliferous scales on the female strobilus. Hence, unlike the endosperm, the megagametophytic tissue does not undergo fertilization and only represents the maternal genotype (Williams 2009). Beside this, the megagametophyte is a metabolically and transcriptionally active tissue (Vuosku et al. 2009; Cervantes et al. 2021) and does not carry the confounding effects of haploid male tissue competition observed in angiosperms.

Here, we have looked at the effect of purifying selection on the genetic diversity (π), the folded site frequency spectrum (fSFS), and the distribution of fitness effects (DFEs) of genes expressed in haploid (megagametophyte) and diploid (embryo, vegetative bud, needle, and phloem) tissues of *P. sylvestris*. We chose to use tissue-specific genes as pleiotropic constraints arising from broad tissue expression could cause a confounding signal of purifying selection (Huber et al. 2017). Also, the use of tissue-specific genes allows a more reliable comparison across gene categories within species. We focus our analysis on single sampling location to minimize the demographic and population structure effects on the SFS. *Pinus sylvestris* has a very subtle population structure with low levels of genetic differentiation across its distribution range (Pyhäjärvi et al. 2007; Tyrmi et al. 2020; Hall et al. 2021). Thus, this single sampling location represents well selection patterns for the whole species because it is very unlikely that haploid selection varies geographically. Here, we expect that genes specific to the haploid stage in predominately diploid organisms should display a similar response to selection as laid out by theoretical and empirical expectations of haploid selection in single-cell organisms. Hence, we expect to observe lower levels of genetic diversity and lower values of the 4-fold to 0-fold pairwise nucleotide diversity ratio (π_0/π_4) for genes

Table 1. Summary of Genetic Diversity Estimates, Tajima's *D*, and GC-Content.

Tissue (*N*)	$\pi_{0\text{-fold}}$ (SE)	$D_{0\text{-fold}}$	$\pi_{4\text{-fold}}$ (SE)	$D_{4\text{-fold}}$	$\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ (SE)	%GC
Megagametophyte (332)	**0.0009 (0.0001)**	**−0.0104**	**0.0032 (0.0004)**	**−0.0318**	**0.25 (0.06)**	**42.03**
Bud (378)	**0.0009 (0.0001)**	**−0.0118**	**0.0029 (0.0004)**	**−0.0069**	**0.35 (0.05)**	**40.98**
Embryo (284)	**0.0014 (0.0002)**	**−0.0291**	**0.0037 (0.0005)**	**−0.0123**	**0.39 (0.07)**	**42.06**
Needle (579)	**0.0005 ($6.72e^{-05}$)**	**−0.0143**	**0.0021 (0.0003)**	**0**	**0.21 (0.04)**	**41.20**
Phloem (387)	**0.0007 ($9.98e^{-05}$)**	**−0.0253**	**0.0025 (0.0004)**	**−0.0073**	**0.32 (0.07)**	**42.01**
All-genes (4,814)	**0.0008 ($2.88e^{-05}$)**	**−0.0077**	**0.0028 (0.0001)**	**−0.0213**	**0.28 (0.01)**	**41.16**

NOTE.—Estimates are provided for tissue specifically expressed genes (τ index 0.8–1) across the five tissues. The all-genes category includes all genes present in the data set regardless of their tissue specificity. Estimates of pairwise diversity (π) represent mean values across genes. SE, standard errors; *N*, number of genes used for estimation. Additional information used for Tajima's *D* estimates is shown in supplementary table S1, Supplementary Material online.

expressed in the haploid stage due to more effective purifying selection (Charlesworth et al. 1993; Charlesworth 1994). Additionally, when efficient, purifying selection draws deleterious mutations to low frequencies. Thus, we also expect to observe a skew towards rare alleles in fSFS of haploid tissue–specific genes and a strong signal of purifying selection in haploid tissue–specific genes based on estimates of the DFE.

Results

We base our study on genetic diversity and expression patterns of tissue specifically expressed genes in one haploid and four diploid tissues. To assess the influence of purifying selection on the nucleotide diversity of the tissue specifically expressed genes, we genotyped 20 megagametophytes from unrelated trees of a single population using exome capture (Kesälahti R, unpublished data). We identified tissue-specific genes based on a previous study (Cervantes et al. 2021). Analyses were done in five data sets, one for each tissue-specific set of genes, plus two reference data sets (all-genes and all-sites, see Materials and Methods).

Genetic Diversity Level of Genes with Tissue-Specific Expression Patterns

To identify the effects of purifying selection across genes with varying expression patterns, we first estimated nucleotide diversity at 0-fold and 4-fold sites ($\pi_{0\text{-fold}}$, $\pi_{4\text{-fold}}$) and their ratios per gene. As purifying selection reduces molecular genetic diversity due to background selection around the selected sites (Charlesworth et al. 1993; Charlesworth 1994) and reduces the $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio, we expected reduced estimates of haploid tissue–specific genes. Although there are significant differences between some estimates of diversity among tissue-specific genes, there is no consistent trend of differences between haploid and diploid tissue–specific genes (table 1; supplementary data 1, Supplementary Material online). Overall, the observed π and π_0/π_4 ratio are well within the range of previously observed values across populations in the entire distribution range for *P. sylvestris* (Grivet et al. 2017; Pyhäjärvi et al. 2019; Tyrmi et al. 2020). Based on a Kruskal–Wallis test followed by a Dunn test, the $\pi_{0\text{-fold}}/\pi_{4\text{-fold}}$ ratio varied significantly among tissues, but again, the differences were not particularly strong between haploid- and diploid-expressed genes (table 1; supplementary data 1, Supplementary Material online). Tajima's *D* (Tajima 1989), a summary statistic of SFS, was negative across the whole data set, as is typical for the species (Pyhäjärvi et al. 2007; Kujala and Savolainen 2012), but again, no extreme values were observed in haploid-specific 0-fold sites (table 1). It is noteworthy that the most negative Tajima's *D* values are observed in 4-fold haploid-specific sites. However, Tajima's *D* exact value is dependent on the total amount of segregating sites and, similar to the other summary statistics, does not capture the whole information of SFS.

DFEs of New Mutations

Estimates of π, their ratios, and other summary statistics of SFS are not optimal signals of purifying selection because they do not fully consider the effect of purifying selection on different classes of allele frequencies. There are, however, more sensitive methods to specifically estimate the strength of purifying selection based on SFSs that rely on the differences of observed and expected allele frequencies.

The DFE-alpha is a method for estimating the expected fitness effects of new mutations entering the population. It is based on the concept that the fitness effect of a mutation is one of the determinants of the frequency at which it will be found in the population (Eyre-Walker and Keightley 2007; Keightley and Eyre-Walker 2007, 2010). Briefly, the DFE-alpha uses the observed amount of putative neutral diversity (4-fold) to estimate the population mutation rate and to account for the effects of demographic history. Using this information and the number of sites available for mutations to occur, the program can infer the amount of amino acid–changing mutations that could have entered the population and compares it to the observed number of 0-fold sites. Considering all these components, DFE-alpha allows the inference of the strength of purifying selection.

To estimate the DFE of different tissue specifically expressed genes and all-sites (see Materials and Methods for definition), we obtained fSFSs for 0-fold and 4-fold positions (fig. 1). A visual inspection of the fSFS shows that in the megagametophyte specifically expressed genes, 4-fold sites had proportionately more singletons than 0-fold sites in contrast to all other site categories and all other tissue-

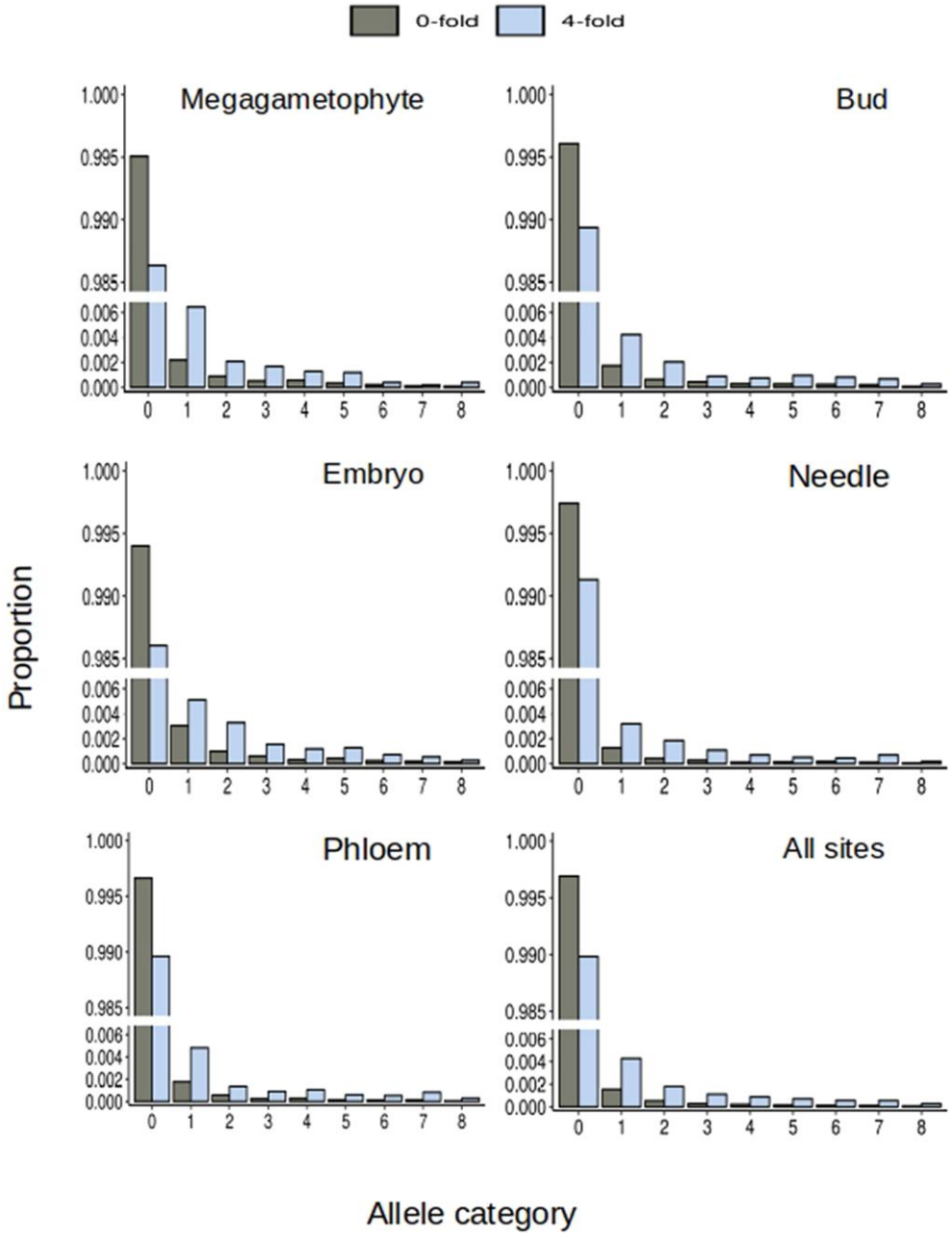

FIG. 1. fSFS for 0-fold and 4-fold sites at tissue-specific genes. The fSFSs were obtained from a downsampling of the observed data to 16 alleles to account for missing data. The same data were also used as the input fSFSs to estimate the DFE. Amount of genes used to obtain the fSFS for each tissue is presented in supplementary table S2, Supplementary Material online.

specific genes. Additionally, the difference between the proportion of 4-fold sites to 0-fold sites at the invariant category is larger in the megagametophyte compared with other tissue-specific expressed genes.

To obtain the DFE of 0-fold sites of the five tissue-specific gene sets and at all-sites (see Materials and Methods), we used DFE-alpha 2.16 with a two-epoch model (Keightley and Eyre-Walker 2007). Further, all nucleotide diversity data came from a single population to avoid confounding effects due to population structure. Based on visual inspection of the DFEs, tissue-specific genes had a distinct DFE from the all-sites data set that represents the genome-wide DFE (fig. 2). However, the tissue-specific genes did not have a consistent deviation from the genome-wide pattern. For example, the bud tissue–specific genes have a high proportion of mutations in the category $N_e s > 100$, which is an indicator of a stronger purifying selection, whereas phloem has fewer mutations in this category in comparison with the genome-wide average (fig. 2). Other differences among tissues, in comparison with all-sites, are evident in all DFE classes as shown in figure 2.

Interestingly, megagametophyte tissue–specific genes have the highest proportion of sites in the class $N_e s > 100$, indicating a large proportion of sites under strong

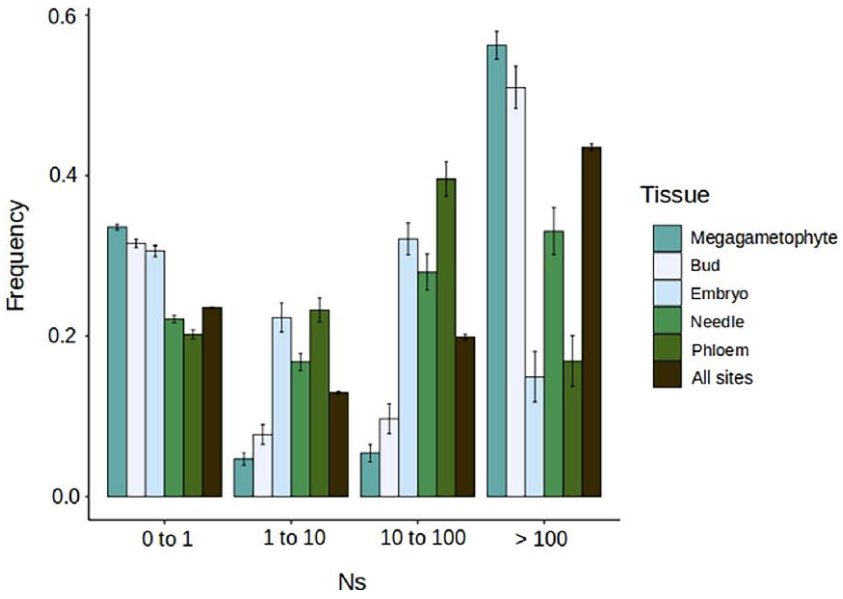

FIG. 2. DFEs of tissue-specific genes (τ index ≥ 0.8) for the five tissues and all-sites data sets. Inferences are based in 200 resamplings of the fSFS. Bars represent standard errors of the mean.

purifying selection, but also display the higher proportion of neutral or nearly neutral mutations ($0 < N_e s < 1$ category). Overall, our results show that each tissue-specific set of genes has a distinct DFE, which is not surprising considering that different tissues are under different selective pressures according to their functionality and in some cases depending on the developmental stage of the organ or tissue where they are expressed.

The DFE visualization is usually divided into four different classes according to scaled strength of selection (fig. 2). Classes are inferred from the β parameter (shape parameter of the gamma distribution) and the mean selective effect of a new mutation (Es) estimate (scale parameter of the gamma distribution). Low values of the β parameter ($\beta \rightarrow 0$) are indicative of a highly leptokurtic gamma distribution (L-shaped gamma) with most of the mutations being either nearly neutral (low fitness effect) or strongly deleterious (high fitness effect) (Keightley and Eyre-Walker 2007; Brevet and Lartillot 2021). To estimate how different the DFE of the megagametophyte tissue–specific genes is compared with the other tissues, we inspected the distribution of the β parameter, instead of calculating confidence intervals over the different classes of strength of selection ($N_e s$) as is usually done. A distribution of the β parameter values obtained from 200 permuted DFE inferences is shown in figure 3, where megagametophyte tissue–specific genes have the lowest values of the β parameter, and their distribution only overlaps with that of the bud tissue–specific genes, with the overlap accounting for approximately 7% of the values.

Evolutionary expectations over gene expression breadth establish that narrowly expressed genes are under more relaxed selection. As haploid tissue–specific genes are narrowly expressed, this contrasting effect could conceal the patterns arising from haploid selection. Also, genes with higher levels of expression are under stronger selective constraints, which can confound the signal of purifying selection observed in megagametophyte tissue–specific genes. Hence, we looked at the extreme values of tau and the level of expression to see if our results were robust to these confounding factors. First, we further restricted the DFE estimations to genes with tau values above the median. Our results show (fig. 4) that contrary to expectations over breadth of expression, highly tissue-specific genes in the megagametophyte show an even clearer signal of purifying selection compared with the other tissue-specific genes. Second, we evaluate if the strong signal of purifying selection we observed in megagametophyte tissue–specific genes was mainly driven by highly expressed genes or was independent of the level of expression. Hence, we subset the tissue-specific genes according to their level of expression, and we run a DFE inference on genes that range in expression from the lowest value up to the median value of expression. Our results show that the signal of purifying selection in haploid-specific genes is not driven by megagametophyte specifically expressed genes with higher levels of expression, but that also genes with low levels of expression show a strong signal of purifying selection (fig. 5).

Discussion

Here, for the first time, we present evidence of haploid selection on tissue-specific genes of a long-lived gymnosperm tree. Our findings demonstrate that purifying selection is stronger on genes with tissue-specific expression in haploid female–derived megagametophytic tissue compared with diploid tissue–specific genes. In comparison with other studies where the presence of haploid

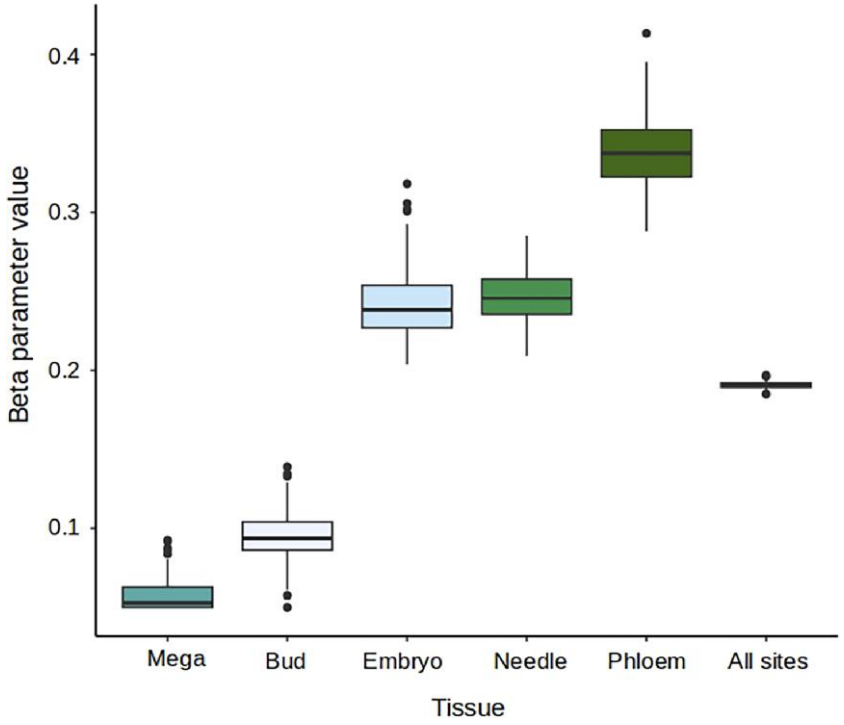

FIG. 3. Distribution of the β parameter values obtained from 200 inferences of the DFE of 0-fold mutations for tissue-specific genes (τ index ≥ 0.8) showing the extreme distribution of the values observed in haploid megagametophyte genes. The β parameter determines the shape of the gamma distribution from which new mutations are drawn during the estimation of the DFE. When values of β get closer to zero, the gamma distribution becomes highly leptokurtic, indicative of a high proportion of new mutations with $N_e s > 100$.

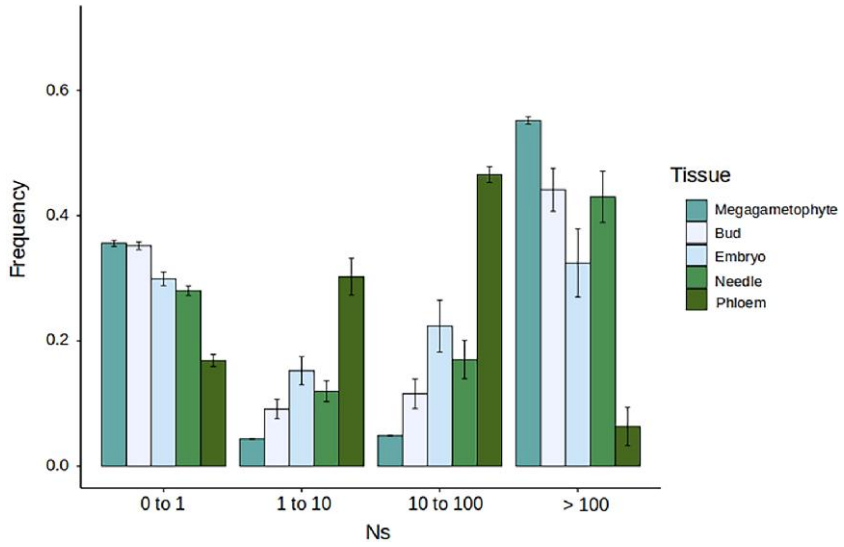

FIG. 4. DFE for highly tissue-specific genes (above median). Results show the DFE across 200 resamplings of the fSFS. Bars represent the standard error of the mean.

selection has been demonstrated (Chibalina and Filatov 2011; Arunkumar et al. 2013; Gutiérrez-Valencia et al. 2022), our results are free of the confounding effects of sexual selection. Szövényi et al. (2013) suggested that in *F. hygrometrica*, a nonangiosperm system also free of confounding effects of sexual selection, the low level and narrow breadth of expression of tissue-specific genes could explain the lack of significant differences in the amount of selective constraint between diploid and haploid tissue–specific expressed genes. Here, our results show that the strength of purifying selection varies across tissues with some diploid tissues, for example, bud and needle, also having strong signals of selection. Evidence of selective constraint variation between genes expressed in different tissues has been reported earlier, for example, in a comparative study of 22 different tissue types

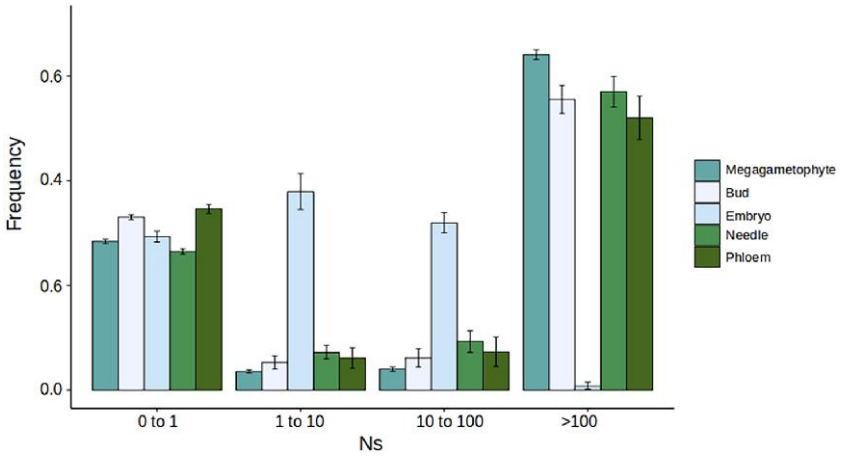

Fig. 5. DFE of tissue-specific genes with expression from the low to the median value of gene expression. Results show the DFE across 200 resamplings of the fSFS. Bars represent the standard error of the mean.

(including reproductive and nonreproductive organs) of mouse (Kryuchkova-Mostacci and Robinson-Rechavi 2015). Similarly in plants, selective constraint differs between genes expressed in different female and male reproductive tissues, reflected as differences in the DFE (Gutiérrez-Valencia et al. 2022). However, we found that unlike diploid tissues, the signal of purifying selection of haploid tissue–specific genes in *P. sylvestris* was robust to further narrowing of the breadth of expression and to low level of expression, two known determinants of evolutionary rate (Duret and Mouchiroud 2000). The signal of strong purifying selection in megagametophyte is in line with both the theoretical expectation of strong selection at haploid stage (masking theory) and the known important function of megagametophyte as a nutrition-providing tissue for a germinating embryo (Vuosku et al. 2009).

The method used for the estimation of the DFE (Keightley and Eyre-Walker 2007) assumes additivity, where the effect on an allele does not depend on the other allele, although this assumption is often violated and ignored. For the purpose of comparative studies seeking qualitative differences in DFE among, for example, gene groups, the assumption of additivity is not a problem when there is no reason to believe that the distribution of dominance is different among the gene groups compared. In our study, dominance (h) has an essential role, as the masking is most efficient for new mutations with low dominance. For haploid-expressed genes, the assumption of additivity is insignificant, as all the alleles are exposed to selection independent of the other alleles, and dominance level does not matter. However, for all diploid-expressed genes, dominance will have an effect on the DFE estimation. Higher dominance leads to more efficient purifying selection and vice versa. Thus, the differences we observe between haploid- and diploid-expressed gene DFE are not only reflecting their Ns but also Nhs (Huber et al. 2018). Thus, some proportion of DFE differences are probably caused by masking of new mutations with h lower than 0.5, moving the Ns estimate closer to neutral/nearly neutral class ($Ns < 1$) when expressed in the diploid stage.

Unlike the clear signal of purifying selection in haploid tissue–specific genes, we did not observe a striking difference in the levels of genetic diversity among tissue-specific genes expressed in the haploid and diploid stages. However, in effective outcrossers with low levels of linkage disequilibrium (such as *P. sylvestris*), the signal of background selection can be less evident due to the effects of recombination (Charlesworth et al. 1995). We neither observed a skew towards rare alleles in the fSFS of haploid tissue–specific genes but instead observed an excess of completely invariant sites, as would be expected under very strong selection. In diploid genomes, recessive deleterious alleles can remain segregating in low frequency because they are rarely exposed to selection as homozygotes. This can be reflected in the SFS as a skewness toward rare alleles. However, in haploid-expressed genes, even rare recessive alleles are exposed to selection, which may explain why polymorphisms of 0-fold sites are not enriched among rare allele classes in megagametophyte tissue–specific genes.

Haploid life stages are common in nature, with green plants displaying great variation in the spatial and temporal extent of the haploid stage, from dominantly haplontic liverworts to diplontic angiosperms (e.g., Mable and Otto 1998). Nevertheless, studies on the effects of selection tend to concentrate on the sporophyte stage, especially in species with a less conspicuous haploid stage. In this study, we observed a strong signal of purifying selection in genes expressed in a reproductively important and relatively long-lived tissue: the gymnosperm female gametophyte. Our results, as well as results in other systems such as *Rumex* (Sandler et al. 2018), show that genes expressed in haploid stages can be exposed to effective selection. Selection over haploid stages and female tissues deserves more attention because it can also affect the evolutionary dynamics of the sporophyte stage, depending on the amount of pleiotropy across sporophyte and gametophytes.

Materials and Methods

Biological Material and DNA Extraction

We obtained seeds for nucleotide diversity–based analysis from 20 randomly selected open-pollinated *P. sylvestris* trees from a single naturally regenerated population at the Punkaharju Intensive Study Site Finland (https://www.evoltree.eu/resources/intensive-study-sites/sites/site/punkaharju) managed by Natural Resources Institute Finland (Luke). The same tree population has been previously used to obtain RNA sequence information for five different tissues using six different genotypes (Ojeda et al. 2019; Cervantes et al. 2021). Three of those six genotypes have been included in this study (supplementary table S3, Supplementary Material online). We obtained DNA from megagametophytes dissected from germinating seeds, which had been incubated overnight in Petri dishes with wet paper at room temperature in the dark. We extracted DNA with the E.Z.N.A Plant DNA DS Kit (Omega BIO-TEK), quantified the concentration with NanoDrop ND-1000 (Thermo Fisher Scientific), and fragmented it by sonication with a Bioruptor UCD-200 (Diagenode) using two periods of 15 min and one period of 13 min, with periods consisting of cycles of 30 s on and 90 s off. We did a double-side size selection with AMPure XP beads (Beckman Coulter) targeting fragments between 300 and 350 bp. We then confirmed the fragment distribution on a 2100 Bioanalyzer (Agilent) using the Agilent DNA High Sensitivity chips.

Library Preparation and Exome Capture

We prepared DNA libraries using the Kapa HyperPrep (Roche) library kit according to the manufacturer's protocol, and libraries were indexed with Kapa Single-Indexed Adapter kits A and B. Indexed libraries were enriched with three to four polimerase chain reaction (PCR) cycles (depending on library concentration). We verified the quality and the length distribution of the libraries with 2100 Bioanalyzer and High Sensitivity chips (Agilent) and their concentration with Qubit HS dsDNA kit (Thermo Fisher). To obtain a reduced representation of the genome, we used the PiSy UOULU exome capture design (Roche) (Kesälahti R, unpublished data). The use of exome capture for the study of purifying selection has some limitations; for example, it does not provide information about noncoding regions, which can harbor regulatory regions affected by selection. However, we opted for it as the *P. sylvestris* genome size (~20 Gb), the high amount of repetitive content on it (De La Torre et al. 2014), and the lack of a reference genome make working with whole genome sequencing extremely challenging. Besides, exome capture provides a sufficient amount of data for comparing diversity patterns of tissue specifically expressed genes.

To optimize the target capture, we set two hybridization reactions each containing species-specific c0t-1 DNA, the exome capture baits, and ten samples pooled equimolarly to a total amount of 1,000 ng of DNA (Kesälahti R, unpublished data). Each hybridization reaction was incubated for 18 h, followed by 14 PCR cycles for enrichment. The final pools were quantified using the Kapa Library Quant kit (Roche) according to the manufacturer's protocol in a LightCycler 480 (Roche). After quantification, we pooled equimolarly the two hybridization reactions in a single sample. The sequencing was done in an Illumina NextSeq550 with 150 bp paired-end reads at the Biocenter Oulu Sequencing Centre.

Mapping and Variant Call

Demultiplexed and adapter-removed reads of the 20 samples obtained from the sequencing facility were mapped to the *Pinus taeda* reference v2.01 (https://treegenesdb.org/FTP/Genomes/Pita/v2.01/) (Zimin et al. 2017) using BWA (Li and Durbin 2009) with default parameters. SAM files were converted to BAM and sorted using Picard tools 2.21.4. We filled coordinates information in the bam files with samtools 1.9 (Li et al. 2009) fixmate, sorted the files by leftmost coordinates, and removed duplicates with samtools markdup. We added read groups with Picard tools and indexed the final bam files with samtools index.

To identify paralog regions, we implemented the approach described in Tyrmi et al. (2020), which consists of a double variant call at different ploidy levels over the same data set. We did the two variant calls with Freebayes v 1.3.1 (Garrison and Marth 2012). The first call (hereafter diploid call) was done with default parameters and a ploidy level of two. As our samples originated from haploid tissue, we did not expect to observe any real heterozygosity. Hence, we used any observed heterozygosity as a proxy of different paralog copies mapping to the same genomic region. We used BCFtools (Danecek et al. 2021) to identify single nucleotide polymorphism (SNP) positions with two or more heterozygous calls. We then identified the position surrounding the SNP in a 150-bp window with a custom R script (https://github.com/GenTree-h2020-eu/GenTree/blob/master/kastally/paralog_window_filtering/paralog_window_filtering.R) and removed them as putative paralogous regions.

For the second variant call (hereafter haploid call), we used Freebayes with the option "population model" and a theta value of 0.005, ploidy level one, and report of invariant sites. We removed complex variants and sites with more than two alleles with BCFtools and indels with VCFtools (Danecek et al. 2011). We then removed the putative paralog regions identified in the diploid call with VCFtools. We filtered variant and invariant positions at genotype level for genotype quality (GQ) > 20, depth (DP) > 10, and maximum of 20% missing data with VCFtools. Then we used the vcffixup command from vcflib (Garrison et al. 2021) to update the allele number (AN) and allele count (AC) fields to reflect the final genotype counts on the VCFs.

Identification of Genes with Tissue-Specific Expression

To identify the genes targeted on the bait design and correlate them to their pattern and levels of expression, we used the gene expression data of Cervantes et al. (2021), which had been mapped to *P. sylvestris* reference transcriptome (Ojeda et al. 2019; BioProject PRJNA531617). We linked the variant call data mapped to *P. taeda* to the gene expression data indicating the most likely homologous region between *P. sylvestris* transcriptome and *P. taeda* reference (version 2.01) based on Blast (Kesälahti R, unpublished data).

Next, we generated a bed file for all polymorphic and monomorphic positions in the vcf file with the function vcf2bed from BedOps v 2.4.38 (Neph et al. 2012). Then, we used BedTools (Quinlan and Hall 2010) intersect with the options -wa, -wb, and -loj, to link the positions in the vcf file to the information in Kesälahti R, unpublished data (supplementary data 2 and 3, Supplementary Material online). We only retained positions that had a unique match in the *P. sylvestris* transcriptome. For each gene identified, we obtained their tau (τ) index as a measure of tissue specificity (Yanai et al. 2005) and gene expression level information from the TMM matrix reported in Cervantes et al. (2021). We used the information to identify the positions and scaffolds containing the genomic information of tissue-specific genes for five different tissues (megagametophyte, haploid, and four diploid tissues vegetative bud, embryo, needle, and phloem). All targeted genes identified regardless of their pattern of expression (unspecific or specific) were included in the all-genes category. We only considered genes with a τ index ≥ 0.8 as tissue specific (Yanai et al. 2005). To evaluate the confounding effects of selection due to codon usage, we estimated the percentage of GC on the transcriptome sequences corresponding to each one of the data set (Li et al. 2015; De Oliveira et al. 2021; Morton 2022). Code for estimating the GC content can be found at https://github.com/cervantesarango/Haploid_selection_Pinus_sylvestris.

Structural Annotation of Variant and Invariant Positions

We used the NewAnnotateRef.py script (https://github.com/fabbyrob/science/blob/master/pileup_analyzers/NewAnnotateRef.py) and the *P. taeda* gtf file of v2.01 genome (https://treegenesdb.org/FTP/Genomes/Pita/v2.01/annotation/) to obtain the 0-fold and 4-fold positions of the *P. taeda* scaffolds where we had mapping information. After this, we used VCFtools with the –positions option to obtain one VCF file with monomorphic and polymorphic positions at 0-fold and 4-fold sites (supplementary data 4 and 5, Supplementary Material online).

Genetic Diversity Estimates

We calculated π at 0-fold and 4-fold sites for tissue-specific genes (tau ≥ 0.8) with pixy (Korunes and Samuk 2021). We used the all-genes data set as an input for intervals having a corresponding unique *P. sylvestris* transcript and --bypass_invariant_check as "no" to include invariant sites. We kept only genes for which we have both 0-fold and 4-fold estimates available (supplementary data 6, Supplementary Material online). Additionally, we retained only genes where the total number of sites (0-fold plus 4-fold) used for the estimates of diversity was 50 bp or larger. To calculate the ratio of 0-fold/4-fold sites, we kept only genes with diversity estimate > 0.

Site Frequency Spectrum

As *P. sylvestris* lacks a reference genome and good representation of the outgroup species, we decided against unfolding the SFS due to the uncertainty of assigning the ancestral state. To obtain the fSFSs, we used the bait positions for each set of tissue-specific genes and generated two VCF files per each data set, one for 0-fold and another for 4-fold. Since the amount of missing data varies across sites, we downsampled without replacement the fSFS to different sample sizes (see section below) using an R-script (https://github.com/GenTree-h2020-eu/GenTree/tree/master/kastally/sfs_resampling). To account for positions monomorphic among *P. sylvestris* samples, but carrying a different allele than the *P. taeda* reference, we set all the alternate count (AC) position to zero for sites where AN = AC. Then, we used the fSFS downsampled at 16 alleles to obtain an estimate of the theta parameter and Tajima's *D* (Watterson 1975; Tajima 1989).

Distribution of Fitness Effects

We calculated the DFEs of 0-fold sites using dfe-alpha 2.16 (Keightley and Eyre-Walker 2007). We used the downsampled fSFS of the 0- and 4-fold sites as input. To maximize the amount of information available for the inference, for each data set, we generated five different sample sizes of the fSFS by downsampling without replacement 200× over the data to 16, 17, 18, 19, and 20 alleles representing from 20% to 0% missing data. Additionally, we generated an all-sites data set that included all 0-fold and 4-fold sites identified (target and off-target) in the variant call as a reference point for genome-wide DFE. We used a two-epoch model of inference. For the neutral (4-fold) sites, we searched for the best N_2 (population size after the change of N_e), t_2 (duration of the epoch of population size change), and an initial t_2 of 50 generations. For the selected sites (0-fold), mean effect of deleterious mutations (s) had an initial value of −0.1, and beta had an initial value of 0.5.

Overlap of DFE Inferences

To quantify the amount of overlap between the β parameter values of the bud tissue–specific genes and the megagametophyte tissue–specific genes, we generated a density distribution plot for the β parameter values obtained from the 200 inferences of the DFE (supplementary fig. S1, Supplementary Material online). Then we calculated the

percentage of area overlapping between the two density distributions. Details and R code for the estimation can be found at https://github.com/cervantesarango/Haploid_selection_Pinus_sylvestris.

All figures were done using R ggplot2 package (Wickham et al. 2019), and colors for all figures included were obtain from R package Pacific North West Colors (Lawlor 2020). Breaks on graphs axes were introduced with ggbreak (Xu et al. 2021).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online and at figshare https://doi.org/10.6084/m9.figshare.c.6442517.v1.

Acknowledgments

This work was supported by the Academy of Finland grants 287431, 293819, and 319313 to T.P., Biocenter Oulu (S.C.), and Emil Aaltonen Foundation (R.K.). The authors acknowledge the CSC—IT Center for Science, Finland, for computational resources. Open access funded by Helsinki University Library.

Author Contributions

The study was designed by T.P. All lab work was done by T.A.K. Initial bioinformatic work was done by S.C., R.K., and T.M.M. Annotation and all analyses were done by S.C. Writing of draft manuscript was done by S.C. T.P. and H.H. contributed substantial comments to the manuscript and provided guidance on all analyses. All authors reviewed and commented on the manuscript.

Data Availability

Raw read for the exome capture is deposited at NCBI BioProject PRJNA937910. The following is the temporary link to the metadata at NCBI https://dataview.ncbi.nlm.nih.gov/object/PRJNA937910? reviewer = uhbblccfd09dr7ejffipve4fso.

References

Arunkumar R, Josephs EB, Williamson RJ, Wright SI. 2013. Pollen-specific, but not sperm-specific, genes show stronger purifying selection and higher rates of positive selection than sporophytic genes in *Capsella grandiflora*. *Mol Biol Evol*. **30**(11):2475–2486.

Beaudry FEG, Rifkin JL, Barrett SCH, Wright SI. 2020. Evolutionary genomics of plant gametophytic selection. *Plant Commun*. **1**(6):100115.

Brevet M, Lartillot N. 2021. Reconstructing the history of variation in effective population size along phylogenies. *Genome Biol Evol*. **13**(8):1–16.

Cervantes S, Vuosku J, Pyhäjärvi T. 2021. Atlas of tissue-specific and tissue-preferential gene expression in ecologically and economically significant conifer *Pinus sylvestris*. *PeerJ* **9**:1–21.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. **63**(3):213–227.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**(4):1289–1303.

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* **141**(4):1619–1632.

Chibalina MV, Filatov DA. 2011. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol*. **21**(17):1475–1479.

Crow JF, Kimura M. 1965. Evolution in sexual and asexual populations. *Am Nat*. **99**(909):439–450.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15):2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2):1–4.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K, *et al.* 2014. Insights into conifer giga-genomes. *Plant Physiol*. **166**(4):1724–1732.

De Oliveira JL, Morales AC, Hurst LD, Urrutia AO, Thompson CRL, Wolf JB. 2021. Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. *Mol Biol Evol*. **38**(8):3247–3266.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. **17**(1):68–070.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. **8**(8):610–618.

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. 2021. Vcflib and tools for processing the VCF variant call format. *bioRxiv*:2021.05.21.445151.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *bioRxiv*:1207.3907.

Gerstein AC, Cleathero LA, Mandegar MA, Otto SP. 2011. Haploids adapt faster than diploids across a range of environments. *J Evol Biol*. **24**(3):531–540.

Gossmann TI, Schmid MW, Grossniklaus U, Schmid KJ. 2014. Selection-driven evolution of sex-biased genes is consistent with sexual selection in *Arabidopsis thaliana*. *Mol Biol Evol*. **31**(3):574–583.

Grivet D, Avia K, Vaattovaara A, Eckert AJ, Neale DB, Savolainen O, González-Martínez SC. 2017. High rate of adaptive evolution in two widespread European pines. *Mol Ecol*. **26**(24):6857–6870.

Gutiérrez-Valencia J, Fracassetti M, Horvath R, Laenen B, Désamore A, Drouzas AD, Friberg M, Kolář F, Slotte T. 2022. Genomic signatures of sexual selection on sollen-expressed genes in *Arabis alpina*. *Mol Biol Evol*. **39**:1.

Hall D, Olsson J, Zhao W, Kroon J, Wennström U, Wang XR. 2021. Divergent patterns between phenotypic and genetic variation in Scots pine. *Plant Commun*. **2**(1):100139.

Huber CD, Durvasula A, Hancock AM, Lohmueller KE. 2018. Gene expression drives the evolution of dominance. *Nat Commun* **9**(1):333.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci U S A*. **114**(17):4465–4470.

Immler S. 2019. Haploid selection in "diploid" organisms. *Annu Rev Ecol Evol Syst*. **50**:9.1–9.18.

Immler S, Otto SP. 2018. The evolutionary consequences of selection at the haploid gametic stage. *Am Nat*. **192**(2):241–249.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**(4):2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philos Trans R Soc Lond B Biol Sci*. **365**(1544):1187–1193.
Kondrashov Alexey S, Crow James F. 1991. Haploidy or diploidy: which is better? *Nature* **351**:314–315.
Korunes KL, Samuk K. 2021. pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour*. **21**(4):1359–1368.
Kryuchkova-Mostacci N, Robinson-Rechavi M. 2015. Tissue-specific evolution of protein coding genes in human and mouse. *PLoS One* **10**(6):e0131673.
Kujala ST, Savolainen O. 2012. Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genet Genomes*. **8**(6):1451–1467.
Lawlor J. 2020. PNWColors: color palettes inspired by nature in the US Pacific Northwest. R package version 0.1.0. doi: doi.org/10.5281/zenodo.3971033.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14):1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16):2078–2079.
Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-content of synonymous codons profoundly influences amino acid usage. *G3 Genes Genomes Genet.* **5**(10):2027–2036.
Mable BK, Otto SP. 1998. The evolution of life cycles. *BioEssays* **20**(6):453–462.
Moore JC, Pannell JR. 2011. Sexual selection in plants. *Curr Biol.* **21**(5):R176–R182.
Morton BR. 2022. Context-dependent mutation dynamics, not selection, explains the codon usage bias of most angiosperm chloroplast genes. *J Mol Evol*. **90**(1):17–29.
Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, *et al.* 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**(14):1919–1920.
Ojeda DI, Mattila TM, Ruttink T, Kujala ST, Kärkkäinen K, Verta JP, Pyhäjärvi T. 2019. Utilization of tissue ploidy level variation in de novo transcriptome assembly of *Pinus sylvestris*. *G3 Genes, Genomes, Genet*. **9**(10):3409–3421.
Otto SP, Gerstein AC. 2008. The evolution of haploidy and diploidy. *Curr Biol.* **18**(24):R1121–R1124.
Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O. 2007. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* **177**(3):1713–1724.
Pyhäjärvi T, Kujala ST, Savolainen O. 2019. 275 years of forestry meets genomics in *Pinus sylvestris*. *Evol Appl*. **13**(1):11–30.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6):841–842.
Sandler G, Beaudry FEG, Barrett SCH, Wright SI. 2018. The effects of haploid selection on Y chromosome evolution in two closely related dioecious plants. *Evol Lett.* **2**(4):368–377.
Szövényi P, Ricca M, Hock Z, Shaw JA, Shimizu KK, Wagner A. 2013. Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Mol Biol Evol*. **30**(8):1929–1939.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3):585–595.
Tyrmi JS, Vuosku J, Acosta JJ, Li Z, Sterck L, Cervera MT, Savolainen O, Pyhäjärvi T. 2020. Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. *G3 Genes Genomes Genet*. **10**(8):2683–2696.
Vuosku J, Sarjala T, Jokela A, Sutela S, Sääskilahti M, Suorsa M, Läärä E, Häggman H. 2009. One tissue, two fates: different roles of megagametophyte cells during Scots pine embryogenesis. *J Exp Bot*. **60**(4):1375–1386.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* **276**(7):256–276.
Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, *et al.* 2019. Welcome to the tidyverse. *J Open Source Softw*. **4**(43):1686.
Williams CG. 2009. *Conifer reproductive biology*. Dordrecht: Springer.
Xu S, Chen M, Feng T, Zhan L, Zhou L, Yu G. 2021. Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers. *Front Genet.* **12**:1–7.
Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, *et al.* 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**(5):650–659.
Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet.* **16**(7):409–420.
Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg SL. 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience* **6**(1):giw016.