Received: 22 October 2021 | Revised: 7 August 2023 | Accepted: 9 August 2023

DOI: 10.1111/mec.17106

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Quantifying adaptive evolution and the effects of natural selection across the Norway spruce genome

Xi Wang[1] | Pär K. Ingvarsson[2]

[1]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[2]Linnean Centre for Plant Biology, Department of Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden

Correspondence
Pär K. Ingvarsson, Linnean Centre for Plant Biology, Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Science, Uppsala, Sweden.
Email: par.ingvarsson@slu.se

Present address
Xi Wang, Department of Biology, University of Copenhagen, Copenhagen Biocenter, Copenhagen N, Denmark

Funding information
Knut och Alice Wallenbergs Stiftelse; Stiftelsen för Strategisk Forskning, Grant/Award Number: RBP14-0040

Handling Editor: Luke Browne

Abstract

Detecting natural selection is one of the major goals of evolutionary genomics. Here, we sequenced the whole genome of 25 *Picea abies* individuals and quantified the amount of selection across the genome. Using an estimate of the distribution of fitness effects, we showed that both negative selection and the rate of positively selected substitutions are very limited in coding regions. We found a positive correlation between the rate of adaptive substitutions and recombination rate and a negative correlation between the rate of adaptive substitutions and gene density, suggesting a widespread influence from Hill–Robertson interference on the efficiency of protein adaptation in *P. abies*. Finally, the distinct population statistics between genomic regions under either positive or balancing selection with that under neutral regions indicated the impact of natural selection on the genomic architecture of Norway spruce. Further gene ontology enrichment analysis for genes located in regions identified as undergoing either positive or long-term balancing selection also highlighted the specific molecular functions and biological processes that appear to be targets of selection in Norway spruce.

KEYWORDS
balancing selection, negative selection, *Picea abies*, positive selection, whole-genome resequencing

1 | INTRODUCTION

Natural selection leaves detectable signatures in the genome of a species, and characterizing and quantifying such signatures at the molecular level is one of the major goals of evolutionary genomics. Detecting genes, or genomic regions, that have been targeted by natural selection is significant not only because they illustrate the action of evolutionary processes and shed light on species histories but also because they could represent biologically meaningful variation that may provide important functional information (Nielsen, 2005; Vitti et al., 2013).

The nearly neutral theory (Ohta, 1973) was proposed as an extension to the neutral theory model (Kimura, 1983) to overcome some of the shortcomings of the neutral model which inadequately explained emerging molecular data, in particular, the constancy of the molecular clock (Chen et al., 2020). Compared with the neutral theory, where mutations are assumed to be either neutral or strongly deleterious, the nearly neutral theory also considers a class of mutations that are weakly selected and effectively neutral and the fraction of mutations affected by selection hence depends on the effective population size (Ohta, 1973, 1992; Ohta & Gillespie, 1996). This weak selection model was described by Kreitman (1996) as 'the

slightly deleterious model' and primarily considers slightly deleterious mutations. Both the strict neutral theory and the nearly neutral theory assume that 'only a minute fraction of DNA changes in evolution are adaptive in nature' (Kimura, 1983; Ohta, 1973) and thus suggest that mutations influenced by positive selection are too rare to have any 'statistically' significant effect on the rate of evolution in most organisms (Ingvarsson, 2009). While we do not expect advantageous mutations to play much role in patterns of polymorphism, however, we do expect them to pay a contribution to substitutions occurring over evolutionary time, referred to as 'the rate of evolution'. Although we should carefully treat the situation, e.g., if the population is mutation rate limited, the adaptive evolution rate will be influenced by population size, since there will be a larger pool of potential adaptive mutations with larger effective population size (Lanfear et al., 2014). Alternative models allocating a greater role to positive selection thus have been proposed, and these suggest that the rate of evolution should be determined by both the beneficial mutation rate and how quickly these mutations can spread and ultimately fix in a species (Gillespie, 1991, 2000, 2001). Under such models, clarifying and quantifying the relative contribution of neutral, beneficial and deleterious mutations to rates of evolution across the genome is one of the outstanding problems in evolutionary genetics.

Keightley and Eyre-Walker (2007) developed a site frequency spectrum (SFS) based on maximum-likelihood approach that combines within-species nucleotide polymorphism data and parameters of a demographic model that allows a population size change at some time in the past, to estimate the distribution of fitness effects (DFE) of newly arisen mutations. This method was subsequently extended to also estimate the efficiency of adaptive molecular evolution by inferring the rate and fitness effects of advantageous mutations by accounting for the contribution of slightly deleterious mutations to polymorphism and divergence (Eyre-Walker & Keightley, 2009). This method, thus, makes it possible to quantify the adaptive rate in species while still regarding the nearly neutral model as the de facto null model.

Current evidence suggests that both positive and negative selection are common in coding and some noncoding regions in several model systems, e.g. *Drosophila* (Andolfatto, 2005; Fraïsse et al., 2019; Haddrill et al., 2008), humans (Arbiza et al., 2013; Lindblad-Toh et al., 2011; Torgerson et al., 2009) and mice (Halligan et al., 2010; Kousathanas et al., 2011). It has been suggested that the majority of adaptive evolution may occur in noncoding, regulatory regions because new mutations that occur in these regions may have fewer deleterious pleiotropic effects (Carroll, 2005; Wray, 2007). Halligan et al. (2013) showed that there have been many adaptive substitutions in noncoding DNA than in coding regions in house mice, although adaptive substitutions in coding regions may experience stronger positive selection. However, our understanding of the action of natural selection remains relatively limited in plants, particularly in noncoding regions. Studies in *Capsella grandiflora* found widespread positive and negative selection in both coding and regulatory regions, but also suggested that both positive and negative selection on plant noncoding sequences are considerably rarer than in animal genomes (Williamson et al., 2014).

Accumulated evidence over the last decade shows that the rates of adaptive protein evolution are high in some species. For instance, 45% of all amino acid substitutions are thought to have been fixed by positive selection in *Drosophila* (*D. simulans* and *D. yakuba*, Smith & Eyre-Walker, 2002), more than 50% in enteric bacteria (*E. coli* and *S. enterica*, Charlesworth & Eyre-Walker, 2006), 57% in wild mice (*Mus musculus castaneus*, Halligan et al., 2010), 40% in *C. grandiflora* (Slotte et al., 2010), 30% in *Populus tremula* (Ingvarsson, 2009) and between 10% and 20% of substitutions differentiating humans and chimpanzees (Boyko et al., 2008; Gojobori et al., 2007). Estimating the rates of adaptive evolution in protein-coding sequences and the findings that the rates of adaptive evolution differ across species opens up possibilities to investigate the factors affecting the efficiency of natural selection. Hill–Robertson interference (HRi) is expected to reduce the overall efficiency of natural selection when there is linkage between sites occurring on haplotype under selection (Castellano et al., 2016; Comeron et al., 2008; Felsenstein, 1974; Hill & Robertson, 1966). Specifically, when a newly arisen advantageous mutation is linked to other beneficial mutations, the probability of fixation will be reduced because of competition among the different adaptive mutations. Similarly, when an advantageous mutation arises in linkage disequilibrium with deleterious mutations, its fixation probability will also decrease if it cannot recombine away from the deleterious background (Castellano et al., 2016; Comeron et al., 2008). The magnitude of this linkage effect depends on recombination rates and the strength of selection. We, therefore, expect a positive correlation between recombination rate and the rate of adaptive evolution, as the influence of linkage is expected to be stronger and will hence result in stronger HRi in regions of low recombination compared to high recombination (Corbett-Detig et al., 2015; Cutter & Payseur, 2013; Ellegren & Galtier, 2016; Wang et al., 2016). Similarly, genes embedded in gene-rich regions should show stronger HRi due to stronger linkage effect than genes located in gene poor regions because the densities of selected sites are thought to be higher in gene-rich regions (Cutter & Payseur, 2013; Flowers et al., 2012; Wang et al., 2016). The net result is an expected negative correlation between gene density and rates of adaptive evolution.

The factors that maintain genetic and phenotypic variation within natural populations have long been an important topic in evolutionary biology (Delph & Kelly, 2014). Positive selection raises the frequency of adaptive mutations over time in a population, while negative selection decreases the frequency of alleles that impair fitness, and both processes act to reduce genetic diversity (Dutheil, 2020). In contrast, balancing selection maintains multiple advantageous polymorphisms in populations, leading to an increased genetic diversity in regions surrounding a balanced polymorphism (de Filippo et al., 2016). The importance of balancing selection shaping genetic diversity has been investigated in many species. Koenig et al. (2019) pointed to long-term balancing selection as an important factor shaping the genetics of immune systems in plants and as the predominant driver of genomic variability after a population bottleneck

in the genus *Capsella*. Wang, Street, et al. (2020) and provided evidence that apart from background selection, both recent positive selection and long-term balancing selection have also been crucial components in shaping patterns of genome-wide variation during the speciation process among the three aspen species (*P. tremula*, *P. davidiana* and *P. tremuloides*) by inferring the genealogical relationships and estimating the extent of ancient introgression across the genome.

Conifers are the most widely distributed group of gymnosperms and are estimated to cover ~39% of the world's forests (De La Torre et al., 2014). Conifer genomes are large (typically 20–40 Gb) and highly repetitive but nevertheless show a largely conserved synteny even over long evolutionary timescales (Nystedt et al., 2013; Pavy et al., 2012). The large size of most conifer genomes has made them inaccessible to genome-wide studies, but the recent publication of a draft reference genome for Norway spruce (Nystedt et al., 2013), one of the most important conifer species from both an ecological and economical perspective, has opened up possibilities for whole-genome resequencing in this species and thus also to assess how patterns of natural selection vary across the genome (Wang, Bernhardsson, et al., 2020). In this paper, we use available genome resources for Norway spruce together with whole-genome resequencing data generated from samples of trees spanning the distribution range of *Picea abies* to quantify both positive and negative selection acting on coding and noncoding regions. In addition, we use genome scans across the Norway spruce genome to identify genomic regions that have been targeted by either recent positive selection or long-term balancing selection to assess how natural selection affects patterns of variation and to also understand the functions of gene located in those regions.

2 | METHODS

2.1 | Sampling, sequencing and variant calling

We used whole-genome resequencing data from 34 individuals of Norway spruce (*P. abies*) previously described in Wang, Street, et al. (2020). The samples were collected to span the natural distribution of the species, extending from Finland in the east to Sweden and Norway in the west and Belarus, Poland and Romania in the south (Table S1; Figure S1). All samples were collected from newly emerged needles or dormant buds and were stored at −80°C until DNA extraction using a Qiagen plant mini kit following the manufacturer's instructions. All sequencing were performed at the National Genomics Initiative platform at the SciLifeLab facilities in Stockholm, Sweden, using paired-end libraries with an insert size of 500 bp. We additionally downloaded whole-genome resequencing data from one sample of *Picea glauca* from NCBI (BioSample: SAMN02736787) as the outgroup in this study.

The bioinformatics pipeline used to handle all sequencing data has previously been described in detail in Bernhardsson et al. (2020) and Wang, Street, et al. (2020). Briefly, all raw sequence reads were first mapped to the complete *P. abies* reference genome v.1.0 (Nystedt et al., 2013) using BWA-MEM v0.7.15 (http://bio-bwa.sourceforge.net/bwa.shtml, Li, 2013) with default settings. To reduce the computational complexity of the subsequent SNP calling (due to the large genome size, high repetitive content and fragmented genome assembly), we reduced the data by only considering genomic scaffolds greater than 1 kb in size and then subdivided the BAM files containing the mapped reads into 20 smaller data subsets to enable us to curate the data in parallel and to enable existing software tools (e.g. GATK) to handle the entire data set. In order to eliminate artefacts introduced due to DNA amplification by PCR, which could potentially lead to excessively high read depth in some regions, PCR duplicates were marked in all data subsets using MarkDuplicates in Picard v2.0.1 (http://broadinstitute.github.io/picard/). Local realignment was further performed to minimize mismatching bases occurring in regions with insertions and/or deletions (indels) during the mapping step by first flagging suspected intervals using RealignerTargetCreator, followed by realignment of those intervals using IndelRealigner, both implemented in GATK v3.7 (DePristo et al., 2011). Finally, we performed variant calling using GATK HaplotypeCaller to generate intermediate genomic VCFs (gVCFs) and then carried out joint calling on all gVCF files for the 34 samples using the GenotypeGVCFs module in GATK. Information including original sampling location, platform used and estimated coverage from raw sequencing reads and of BAM files after mapping are given for all individuals in Table S1. In addition, we downloaded raw sequences from one individual of White spruce (*P. glauca*) to use as an outgroup from the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra?LinkName=bioproject_sra_all&from_uid=242552) and performed all steps of variant calling described above except for the joint calling step.

2.2 | Filtering to maintain high-quality SNPs

In order to retain high-quality SNPs, we performed the following filtering steps for all called variants to reduce the number of false-positive SNPs. We only included biallelic SNPs positioned >5 bp away from an indel and where the SNP quality parameters fulfilled GATK recommendations for hard filtering.

(https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set). We also recoded genotype calls with a depth outside the range 6–30 and a GQ <15 to missing data and filtered each SNP for being variable with an overall average depth in the range of 8–20 and a 'maximum missing' value of 0.8 (max 20% missing data). Finally, as SNPs were called in collapsed regions in the assembly, likely containing non-unique regions in the genome, should show excess heterozygosities as they are based on reads that are derived from different genomic regions, we removed all SNPs that displayed a *p*-value for excess of heterozygosity <.05. SNPs that passed all the different hard filtering criteria were used in the downstream analyses. For

more detailed information on the genotype hard filtering criteria used, please refer to Bernhardsson et al. (2020) and Wang, Street, et al. (2020).

2.3 | Population structure

The population structure of the 34 individuals has previously been investigated in Wang, Street, et al. (2020). Wang, Street, et al. (2020) identified three main populations where samples from Belarus, Poland and Romania grouped into a cluster defined as 'Central-Europe', three Finish samples clustered into one group defined as 'Finland', and the remaining 25 individuals from Norway and northern Sweden clustered into one group defined as the 'Sweden–Norway' population. To avoid underlying confounding effects of population structure, we performed all downstream analyses in this paper using only individuals derived from the 'Sweden–Norway' population.

2.4 | Functional sites and genomic features estimates: Nucleotide diversity, Tajima's D, divergence, recombination rate, gene density, GC density and repeat density

BED files for different genomic contexts (fourfold synonymous sites, zerofold nonsynonymous sites, intronic sites and intergenic sites) were generated from the genome annotation for *P. abies* v1.0 (available from ftp://plantgenie.org/Data/PlantGenIE/Picea_abies/v1.0/gff) using a custom-made python script (https://github.com/parkingvarsson/Degeneracy/). We also generated a BED file from the ultra-conserved intergenic regions identified between *P. abies* and *P. sylvestris* (ftp://plantgenie.org/Data/PlantGenIE/Picea_abies/v1.0/gff/Conserved_sequences/Psylvestris.gff3.gz). Nystedt et al. (2013) sequenced a single *P. sylvestris* individual to 12.5× coverage to enable comparative analyses with *P. abies*. Ultra-conserved regions between the two species (150MY divergence time) were identified using two approaches: (i) aligning *P. sylvestris* sequencing reads against the *P. abies* genome using BWA with alignment lengths of 100 bps and allowing up to 10 mismatches and one gap opening, hence requiring a 90% identity, and (ii) sequences not as highly conserved as required by the BWA approach were identified using BLAST. *P. sylvestris* genome contigs were aligned against the *P. abies* genome using BLAST with a minimum percentage of identity at 60% and an e-value requirement of e-10. These strategies were combined by intersecting regions identified with regions covered by either genes or repetitive elements. All conserved regions not corresponding to genes and repeats were considered to be ultra-conserved regions between *P. sylvestris* and *P. abies*. Separate VCF files were generated for the different site categories from the original VCF files based on the genomic BED files using vcftools.

We used ANGSD v0.921 (Korneliussen et al., 2014) to estimate pairwise nucleotide diversity and Tajima's D by calculating the site allele frequency likelihood based on normalized phred-scaled likelihoods of the possible genotypes (PL tag in the VCF file). Divergence was calculated between *P. abies* and the outgroup species *P. glauca* at fourfold, zerofold, intronic and intergenic sites by measuring the number of fixed differences per scaffold. The population-scaled recombination rate ($4N_e r$) was estimated per scaffold for each population using a Bayesian reversible-jump Markov Chain Monte Carlo scheme under the crossing-over model as implemented in LDhat v2.2 (McVean et al., 2004). We performed 1,000,000 Markov Chain Monte Carlo iterations with sampling every 2000 iterations and set up a block penalty parameter of 5 using a data set consisting of only scaffolds longer than 5 kb because shorter scaffolds generally did not produce stable estimates. The first 100,000 iterations of the reversible-jump Markov Chain Monte Carlo scheme were discarded as a burn-in. We measured gene density per scaffold as the ratio of sites falling within a gene model on the scaffold to the overall length of the scaffold. The same method was used to estimate repeat density using information on repeat content per scaffold (ftp://plantgenie.org/Data/ConGenIE/Picea_abies/v1.0/GFF3/Repeats/). GC density was calculated at the scaffold level as the fraction of bases where the reference sequence (*P. abies* genome v1.0) was a G or a C using BEDtools.

2.5 | Estimating the distribution of deleterious fitness effects and the fraction of adaptive substitutions for both genic and nongenic regions

The distribution of fitness effects of new mutations (DFE) specifies the probability of a new mutation having a given fitness effect (Keightley & Eyre-Walker, 2007). The software DFE-alpha (Keightley & Eyre-Walker, 2007) was employed to estimate the fraction of sites under negative selection with different effective strengths by incorporating the expected allele frequency distribution generated by transition matrix methods and simultaneously a demographic model that includes a step population size change. This method was based on the 'slightly deleterious model' which assumes that the fitness effects of new mutations at putatively neutral sites are zero and that mutations are unconditionally deleterious at selected sites as advantageous mutations are assumed to be too rare to contribute to polymorphism (Keightley & Eyre-Walker, 2007). In order to make the results more robust, we generated the folded SFS for each category of selected sites (zerofold nonsynonymous, intronic, conserved, promoters, intergenic sites) and a class of putatively neutral reference sites (fourfold synonymous sites) from SNP data using ANGSD and reported the proportion of mutations falling into different effective strengths of selection ($N_e s$, where N_e is the effective population size and s is the selection coefficient) range: 0–1, 1–10, 10–100 and >100, which correspond to effectively neutral, mildly deleterious, deleterious and strongly deleterious, respectively.

Based on the estimated distribution of fitness effects of new deleterious mutations from the polymorphism data, DFE-alpha further predicts the numbers of substitutions originating from

neutral and slightly deleterious mutations between two species by using divergence data (Eyre-Walker & Keightley, 2009). If the observed number of substitutions is greater than this expectation, the proportion of adaptive substitutions (α) and the rate of adaptive substitutions expressed relative to the neutral substitution rate (ω) are estimated from the difference between observed and expected substitutions (Eyre-Walker & Keightley, 2009). We used *P. glauga* as outgroup to infer parameters under positive selection at zerofold, intronic, conserved, promoters and intergenic sites of Norway spruce genome. Jukes–Cantor multiple hits correction was applied to the divergence estimates (Jukes & Cantor, 1969). To calculate 95% of confidence intervals (CI) range for each parameter, we first generated 100 bootstrap replicates by resampling randomly across all scaffolds for each site class (zerofold, introns, conserved, promoters and intergenic region), together with the fourfold (neutral) sites simultaneously using R (R Core Team, 2017). We then calculated the SFS for both functional and neutral sites for each 100 bootstraps and used them as input to DFE-alpha. CI range of 95% for each parameter ($N_e s$, α and ω) was finally represented by excluding the top and bottom 2.5% of the estimated parameter values.

2.6 | Gene bins and the factors affecting amino acid adaptive evolution

To understand the factors influencing the efficiency of natural selection, we further assessed correlations between the rate of adaptive evolution and population genetic statistics within genic regions. To avoid the noisy and large sampling variances arising from estimates of single genes due to limited numbers of segregating or divergent sites for some site classes (Castellano et al., 2016; Moutinho et al., 2020; Stoletzki & Eyre-Walker, 2011), we grouped scaffolds with genes into bins according to their estimated rate of recombination, gene density and mutation rates. The rank of values for all these bins can be referred in Tables S4 and S5. Assuming that fourfold synonymous sites are free of the effects of selection and thus serve as our baseline, we first detected the proportion of adaptive substitutions (α) (zerofold nonsynonymous sites) in each bin by running DFE-alpha (Eyre-Walker & Keightley, 2009). From these estimates, the rate of adaptive evolution (Ka+) can be estimated using the expression: $Ka+ = \alpha Ka$, where Ka represents the number of nonsynonymous substitutions per nonsynonymous site. As a comparison, we also calculated the ratio of nonsynonymous to synonymous substitutions (K_a/K_s) for all bins separately and estimate the correlation between the K_a/K_s ratio and the genomic features of interest.

All statistical analyses were performed using the R statistical package (R Core Team, 2017). Linear regressions were carried out with the R function 'lm', and nonlinear regressions were run using the R function 'nls'. To compare the linear and nonlinear model fits, we calculated Akaike's Information Criterion (AIC) using the R functions 'AIC' and 'BIC'. Pairwise correlations between the variables of interest were calculated using Spearman's rank correlations using the basic R function 'cor.test'.

2.7 | Genome-wide scan for regions under positive selection and balancing selection

Selective sweeps leave distinct signatures in the genome of an organism (Stephan, 2019). To identify the loci and/or regions that have undergone recent positive selection, we scanned the whole genome of Norway spruce, using RAiSD (Raised Accuracy in Sweep Detection), which use multiple signatures of a selective sweep via the enumeration of SNP vectors (Alachiotis & Pavlidis, 2018). This programme introduced the μ statistic, a composite evaluation test that scores genomic regions by quantifying changes in the SFS, the levels of LD and the amount of genetic diversity along a chromosome, achieving high sensitivity and accuracy while reducing the computational complexity, which allows for faster processing with limited memory requirements (Alachiotis & Pavlidis, 2018). We ran RAiSD using default settings on a subset of the complete data, consisting of individuals from the Sweden–Norway population (25 individuals), to remove possible influences due to population structure.

In order to detect regions under long-term balancing selection, we scanned the whole genome of Norway spruce (again using the Sweden–Norway population to limit the effects of population structure) using the β (beta) score summary statistic which detects clusters of alleles at similar frequencies (Siewert & Voight, 2017). This statistic was proposed based on simulations showing that new mutations which arise in close proximity to a site targeted by balancing selection accumulate at frequencies nearly identical to that of the balanced polymorphism (Siewert & Voight, 2017). Compared to existing summary statistics, the β score has improved power to detect balancing selection, is reasonably powered under nonequilibrium demographic models and across a range of recombination and mutation rates, and is computationally efficient and applicable to species that lack appropriate outgroup sequences (Siewert & Voight, 2017). We first converted our VCF file to a betascan input file by running two scripts, 'vcfm2acf' and 'acf2betascan', from the glactools package (https://github.com/grenaud/glactools). We calculated the folded β statistic in 1 kbp windows since the signals of long-term balancing selection are usually localized to very narrow genomic regions (Gao et al., 2015; Wang, Street, et al., 2020). To prevent false positives, we filtered out SNPs with a folded frequency lower than 20% and defined significant SNPs as those SNPs with extreme β scores in the top 0.1% of sites from the genome-wide β score distribution.

To assess the effects of positive and balancing selection on patterns of genome-wide variation, we compared outlier windows identified in two methods, representing regions under positive or balancing selection, with the remaining genomic regions using a variety of population genetic summary statistics, including pairwise nucleotide diversity, Tajima's D, the population-scaled recombination

rate, gene density, GC density and repeat density for Sweden–Norway population (25 individuals).

2.8 | GO enrichment

To determine whether any functional categories were overrepresented among genes in the regions that we identified as being under positive or balancing selection, we performed functional enrichment analysis of GO categories using Fisher's exact test (http://conge nie.org/enrichment). *p*-values for Fisher's exact test were further corrected for multiple testing using the Benjamini–Hochberg FDR method (Benjamini & Hochberg, 1995). GO terms with an FDR corrected *p*-value <.1 were considered to be significantly enriched.

3 | RESULTS AND DISCUSSION

Whole-genome resequencing data were generated for 25 individuals sampled to span the natural distribution of *P. abies* using Illumina HiSeq 2000 or HiSeq X with a mean sequence coverage of 18.1× per individual (Table S1). All raw sequence reads were mapped to the complete *P. abies* reference genome v.1.0, which contains ~10 million scaffolds covering 12.6 Gb out of the estimated genome size of ~20 Gb (Nystedt et al., 2013). To reduce the computational complexity of SNP calling, BAM files were subsetted to only include scaffolds longer than 1 kb and then subdivided into 20 genomic subsets with ~100,000 scaffolds in each. After PCR duplication removal, local realignment, variant calling and several steps for hard filtering, 293.9 million high-quality SNPs remained for all downstream analyses. The number of sites in each functional category is reported in Table S2.

3.1 | Genome-wide measures of purifying selection

Purifying selection in a genomic region is usually considered to be a clear sign that the region has some functional importance and therefore shows evolutionary conservation. To quantify the amount and strength of purifying selection acting across the genome of Norway spruce, we used the methods of Keightley and Eyre-Walker (2007) to infer the percentage of deleterious mutations falling into different categories based on the effective strength of negative selection (in terms of $N_e s$). We assessed this for different categories of sites, including zerofold nonsynonymous, intronic, conserved, promoters and intergenic sites using fourfold synonymous sites as a putatively neutral baseline (Figure 1a; Table 1). For zerofold nonsynonymous sites, most of the mutations fell in the range representing weakly deleterious mutations that behave as effectively neutral ($0 < N_e s < 1$), making up 68.7% (95% CI: 62.3%–77.1%) of all sites. Meanwhile, 13.7% (95% CI: 4.30%–21.5%) of amino acid mutations are considered strongly deleterious ($N_e s > 100$), suggesting that they are under strong purifying selection. The remaining nonsynonymous mutations are under moderate level of negative selection, of which 8.40% (95% CI: 7.60%–9.20%) are classified as mildly deleterious ($1 < N_e s < 10$) and 9.30% (95% CI: 8.40%–9.60%) are deleterious mutations ($10 < N_e s < 100$). As a comparison, for intronic, conserved, promoters and intergenic sites, the vast majority mutations, approaching 100%, are considered nearly neutral with $N_e s$ falling in the range between 0 and 1.

By analysing more than 2400 loci with an average length of ~280 nucleotides from 11 plant species, Gossmann et al. (2010) estimated the distribution of fitness effects of new mutations (DFE) in protein-coding sequences and found that in all species, the largest proportion of mutations are strongly deleterious with $N_e s > 100$ and for most

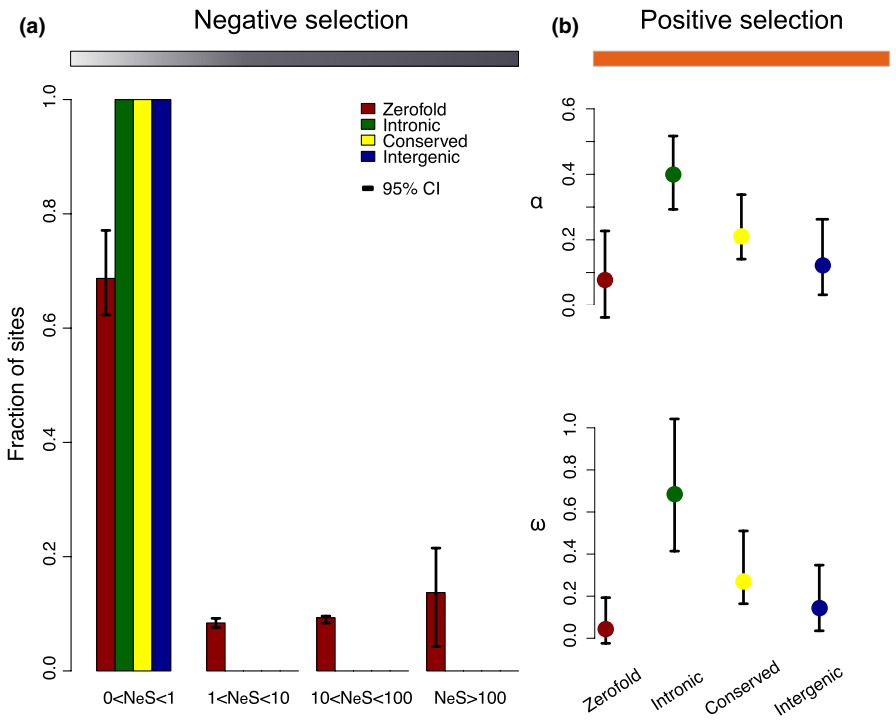


FIGURE 1 Estimates of negative and positive selection on coding and noncoding sites in *P. abies*. (a) The proportion of sites found in each bin of purifying selection strength, separated by site type. (b) The proportion of divergent sites fixed by positive selection (α), and (c) the rate of adaptive substitution relative to neutral divergence (ω). Error bars represent 95% bootstrap confidence intervals.

TABLE 1 Estimates of the distribution of fitness effects of new mutations at zerofold nonsynonymous sites, intronic sites, conserved region, promoters and intergenic sites falling in different $N_e s$ ranges, and proportion of divergence driven to fixation by positive selection (α) and the rate of adaptive substitution relative to neutral divergence (ω) in *P. abies*.

P. abies	Negative selection				Positive selection	
					P. Glauga (outgroup)	
Category	$N_e s$ (0–1)	$N_e s$ (1–10)	$N_e s$ (10–100)	$N_e s$ (>100)	α	ω
Zerofold	0.687 (0.623–0.771)	0.084 (0.076–0.092)	0.093 (0.084–0.096)	0.137 (0.043–0.215)	.096 (−.037–.227)	.071 (−.024–.193)
Intronic	1.000 (1.000–1.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	.417 (.293–.517)	.716 (.414–1.07)
Conserved[a]	1.000 (1.000–1.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	.228 (.141–.338)	.295 (.164–.510)
Intergenic	1.000 (1.000–1.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	0.000 (0.000–0.000)	.139 (.032–.256)	.161 (.033–.345)

Note: Ninety-five percent bootstrap confidence intervals are shown in parentheses.

[a]Conserved: conserved region between *P. abies* and *P. sylvestris* (Scots pine).

species less than 25% of amino acid-changing mutations behave as effectively neutral ($0 < N_e s < 1$). However, three species (*Boechera stricta*, *Populus balsamifera* and *Schiedea globosa*) showed an excess of neutral mutations (>25%) and a decrease of strongly deleterious mutations (<55%). These results suggest a clear impact of the species-wide effective population size (N_e) on the DFE that species with small N_e tend to have a relatively large proportion of mutations that are effectively neutral. Gossmann et al. (2010) additionally estimated the DFE from both wild and domesticated populations of rice and found that domesticated varieties of *Oryza japonica* showed a higher proportion of effectively neutral mutations than the wild species *O. rufipogon*, which may reflect a lower effective population size associated with domesticated varieties. Wang, Street, et al. (2020) inferred the demographic history of Norway spruce and found that it is characterized by several reoccurring bottlenecks corresponding to drastic climate fluctuations during the Quaternary with concomitant decreases in effective population size, even though the species has a widespread current geographic distribution. Historical population size fluctuations and reoccurring bottlenecks have sharply reduced the effective population size in Norway spruce and are one of the likely reasons for the large proportion of new amino acid mutations that we classify as effectively neutral. Human-mediated selection and the use of limited seed sources for reforestation could also have contributed to a reduction in the effective population size. Chen et al. (2019) showed that a large proportion of the 1499 individuals stemming from the Norway spruce breeding programme in southern Sweden correspond to recent introductions from mainland Europe. This fact suggests that humans have affected the demography of Norway spruce through breeding, although the extent of such processes is unclear and is worthy of further study. Moreover, when estimating negative selection, the DFE accounts for only deleterious mutations rather than both deleterious and adaptive mutations (Keightley & Eyre-Walker, 2007). Putative mutations under positive selection will be added to the category of effectively neutral mutations, leading to a higher proportion of mutations falling to range representing effectively neutral ($0 < N_e s < 1$), although the biases due to these processes vary with N_e (Chen, Glémin, & Lascoux, 2017; Chen, Zhang, et al., 2017). Finally, the presence of strong and extensive codon usage bias might violate the presumed neutrality of synonymous sites. Several analyses of codon usage bias from population genetic data suggest the action of selection on synonymous sites in plant species (Ingvarsson, 2010; Lawrie et al., 2013; Machado et al., 2020; Qiu et al., 2011). De La Torre et al. (2015) found high levels of codon bias, measured as F_{op}, in *P. abies* and *P. glauca* when analysing genome-wide levels of gene expression (>50,000 expressed genes) data. Thus, selection on at least a fraction of the synonymous sites in Norway spruce due to strong codon usage bias cannot be ruled out and may therefore contribute to a downwardly biased estimate of the proportion of strongly deleterious amino acid mutations. An alternative would be to use, for example, sites from short introns that have shown to be better approximations for neutrally evolving sites in, e.g., *Drosophila* (Haddrill et al., 2005; Halligan et al., 2004). Once the contiguity of the current Norway spruce genome assembly has been improved, it would be valuable to assess to what extent intron sites in introns of varying lengths are conserved.

Compared to the variable proportions of different classes of purifying selection observed at nonsynonymous sites, negative selection appears to largely be absent in both intronic and intergenic regions. This is similar to the observations by Williamson et al. (2014) in whole-genome sequences from 13 *Capsella grandiflora* individuals. Their data showed that the proportion of intergenic sites that are nearly neutral approached 100% and that approximately 70% of intronic sites were behaving as effectively neutral. Furthermore, after bootstrapping the latter estimate, it was deemed not significantly different from 100% of intronic sites being effectively neutral. Similar patterns were also found by Lin et al. (2018) which showed that negative selection was rare or absent in noncoding regions for two aspen species, with 72.7% of intronic sites in *Populus tremula*, 74.8% of intronic sites in *P. tremuloides* and 100% in intergenic region for both species falling in the effectively neutral category ($0 < N_e s < 1$).

However, the DFE approach might not be sensitive enough when purifying selection is acting on only a very small number of sites. In an effort to circumvent this problem, we quantified negative selection in two more restricted classes of intergenic sites where we have reason to believe that purifying selection is stronger – gene promoters which are regulatory regions located close to genes and regions previously shown to have high sequence conservation between *P. abies* and *P. sylvestris* (Nystedt et al., 2013). However, we fail to detect the evidence of purifying selection also in these smaller compartments of noncoding regions. Although these regions were intended to identify noncoding regions that are evolving under greater functional constraint than a randomly selected region of intergenic space, the actual sites that are the target of selective constraint probably make up an ever smaller subset of these sites. Unfortunately, the poor annotation of the Norway spruce genome does not allow for more fine-grained analyses. Our failure to detect purifying selection does, therefore, not exclude the possibility that some sites nevertheless are evolving under functional constraint, as these categories are rather crudely defined.

The overall neutrality of noncoding regions we detected for Norway spruce is in stark contrast with results from *Drosophila* and from humans, where a relatively large fraction of sites in noncoding regions are under selection, e.g., only 30%–70% of intronic and intergenic regions are classified as nearly neutral in *Drosophila* (Andolfatto, 2005; Hough et al., 2013; Sella et al., 2009), and approximately 30% of amino acid-changing mutations behave as nearly neutral in humans, although there are quite a conspicuous differences between different data sets, e.g., the Environmental Genome Project (EGP) and Program for Genomic Applications PGA data sets (Keightley & Eyre-Walker, 2007).

3.2 | Genome-wide measures of positive selection

The extent to which positive selection contributes to molecular evolution has been a long-standing question in evolutionary genetics (Booker et al., 2017). By accounting for the effect of slightly deleterious mutations, we employed an extension of the McDonald–Kreitman test (Eyre-Walker & Keightley, 2009), to estimate the proportion of adaptive substitutions (α) and the rate of adaptive substitutions expressed relative to the neutral substitution rate (ω). For these analyses, we used *P. glauca* as an outgroup species and focused on estimating rates of adaptive evolution at zerofold sites and sites in intronic, conserved, promoters and intergenic regions across the Norway spruce genome (Figure 1b,c; Table 1). The results suggest that noncoding regions show relatively high proportions adaptive substitutions, with intronic sites having the highest proportion of adaptive substitutions (α=.417) and the highest estimate of the rate of adaptive substitutions (ω=.716), followed by promoters (α=.403, ω=.675) and conserved region (α=.228, ω=.295). Intergenic sites have substantially lower proportions of adaptive substitutions (α=.139) as well as rate of adaptive substitutions (ω=.161). Bootstrap analysis for both the proportion of adaptive substitutions and adaptive rate shows that these estimates are significantly greater than 0 for intronic, conserved, promoters and intergenic sites, suggesting the action of widespread positive selection in noncoding regions in Norway spruce. As a comparison, we found the lowest proportion of adaptive mutations (α=.096) and lowest rate of adaptive evolution (ω=.071) at zerofold nonsynonymous sites.

The large genome size of *P. abies* might be explained by the slow and constant accumulation of a heterogeneous set of LTR-RTs that are not counterbalanced by efficient removal mechanisms (Nystedt et al., 2013). Of 41 genes, 8 (two are transposon-related genes, three genes were unknown and three genes are related to the small RNA biogenesis pathway) genes show significant differential expression patterns related to embryo environment. These results are in line with earlier studies that have shown that temperature during embryogenesis alters the embryo maturation programme and the regulatory elements that affect gene expression, which in turn enable seedlings of Norway spruce to preserve a memory of an environmental influence and therefore fine-tune the regulation of adaptive performance across many years (Yakovlev et al., 2011). As a comparison, Williamson et al. (2014) showed that there is little evidence for a difference in the strength of positive selection on substitutions in coding regions compared to noncoding regions in the flowering plant *C. grandiflora*. This observation is consistent with earlier suggestions that, unlike in animals, flowering plant genomes may contain relatively few noncoding regulatory sequences that are subject to selection, possibly because gene expression can be modified through frequent gene duplication and functional divergence rather than through the evolution of novel regulatory elements (Lockton & Gaut, 2005). The very large proportion of adaptive substitutions and the high adaptive rate of evolution we observe in noncoding regions in Norway spruce genome therefore either are an artefact or require a novel explanation. Polyploidy is a common mode of speciation and evolution in angiosperms (Leitch & Bennett, 1997); however, there is little evidence for recent whole-genome duplications playing significant role in the gymnosperm lineage (Nystedt et al., 2013). Thus, gene duplications might be rare in the Norway spruce genome possibly suggesting more important roles for novel regulatory elements located in noncoding regions for modifying gene expression. Moreover, conifer genomes contain an abundance of repeat-rich content, mostly in the form of transposable elements (De La Torre et al., 2014). Genomic repeats, and in particular transposable elements, have been a rich source of material for the assembly and tinkering of eukaryotic gene regulatory systems (Feschotte, 2008). The very large repetitive fraction (>70%) of Norway spruce genome, and specifically the fraction of long terminal repeat-retrotransposons (LTR-RTs) comprising the Ty3/Gypsy superfamily and Ty1/Copia superfamily, makes contributions to large proportion of putative regulatory elements that ultimately could result in higher rates of adaptive evolution in noncoding regions compared to coding regions. Finally, poor annotation of Norway spruce assembly may explain the very high effect of positive selection in noncoding region. The very large repetitive

fraction of the Norway spruce genome makes it difficult to assemble and annotate, resulting in >10 million scaffolds in the current publicly available genome assembly (Bernhardsson et al., 2019; De La Torre et al., 2014; Nystedt et al., 2013; Wang, Bernhardsson, & Ingvarsson, 2020).

3.3 | Determinants of the efficacy of amino acid adaptive evolution

Compared to the high rates of adaptive evolution we observe in noncoding regions, the fraction of adaptive substitutions fixed by positive selection and the scaled rate of adaptive evolution at zerofold nonsynonymous sites are low and not significantly different from zero as judged by the bootstrap estimates (α 95% CI: −0.037 to 0.227 and ω 95% CI: −0.024 to 0.193, Figure 1b,c; Table 1). This suggests that the proportion and rate of adaptive evolution in coding regions in Norway spruce are limited. Using DNA sequence data from 167 orthologous nuclear gene fragments, Eckert et al. (2013) found little evidence for long-term adaptive nonsynonymous evolution in 11 species of soft pines (subgenus *Strobus*), with only one of the α estimates being significantly different from zero. It is reasonable to expect that estimates of α will vary among plants with differing life histories because life history characteristics of plants are known to affect standing levels of genetic diversity (Eckert et al., 2013). Soft pines are important components of coniferous forests distributed throughout the Northern Hemisphere, and they share many life history characteristics with Norway spruce, such as long generation time and a predominantly outcrossing mating system that rely almost exclusively on wind pollination (Neale & Wheeler, 2019), and it is, thus, nor surprising that we observe similar patterns of adaptive evolution in soft pines and Norway spruce. As a contrast, some angiosperm plants have reported relatively large proportions of amino acid substitutions being fixed by positive selection, for example, 40% in *C. grandiflora* (Slotte et al., 2010) and 75% in sunflowers (Strasburg et al., 2009). Even angiosperm species that have similar life history traits as gymnosperms may still show much higher rates of adaptive evolution. Estimates in species from the genus *Populus* that are long-lived outcrossing forest trees show proportions of amino acid substitutions being fixed by positive selection in excess of 30% in *P. tremula* (Ingvarsson, 2009) and 33.8% in *P. tremuloides* (Lin et al., 2018), suggesting life history traits may not play a large role in determining adaptive evolution but rather other factors. An open question that thus remains is why we observe such a large disparity in the rates of adaptive evolution at nonsynonymous sites among different plant species and what factors are important for explaining variation in the efficacy of selection.

Hill–Robertson interference (HRi) is expected to reduce the overall efficiency of natural selection when a newly arisen advantageous mutation occurs in linkage disequilibrium with either other beneficial mutations or with deleterious mutations (Castellano et al., 2016; Comeron et al., 2008; Felsenstein, 1974; Hill & Robertson, 1966). The magnitude of this linkage effect depends on local recombination rate and selection intensities. We, therefore, first assessed the relationship between the rate of adaptive evolution (Ka+) and recombination rate (scaffold length above 10 kb), expecting to see a positive correlation because the influence of linkage is expected to be stronger, hence resulting in stronger HRi, in regions of low compared to high recombination rates. To estimate the rate of adaptive evolution, it is necessary to combine data from several genes because estimates tend to be error prone and sometimes undefined for individual genes (Castellano et al., 2016). We grouped genes into 10 bins based on recombination rates estimated from the Sweden–Norway population, with each bin on average containing 3079 scaffolds and 3710 genes (Table S4). We found a positive correlation between the rate of adaptive evolution and the recombination rate (Spearman's rank correlation coefficient $r = .164$, Figure 2b). This correlation is not significant, likely due to the limited number of bins used and possibly also because estimated population-scaled recombination rates from data on linkage disequilibrium (LD) are unstable in such a fragmented genome assembly as in Norway spruce. We further test whether a curvilinear relationship fits the data better than a linear model by fitting the function $y = a^{bx}$ to our data and comparing it to the fit of a linear model (Table S3). Both Akaike's Information Criterion (AIC) and Schwarz's Bayesian criterion (BIC) showed that the curvilinear model is favoured, in which 17.4% of the variation in Ka+ can be explained by variation in the recombination rate (Figure 2b; Table S3). A positive correlation between recombination rate and Ka+ could be due to mutagenic recombination. Wang, Bernhardsson, and Ingvarsson (2020) found a positive correlation between nucleotide diversity and recombination rate but no correlation between divergence and recombination rate, suggesting that recombination is not mutagenic in Norway spruce. Similarly, genes embedded in gene-rich regions are expected to show stronger HRi because more sites are under selection in these genes compared to genes located in gene poor regions, resulting in an expected negative correlation between gene density and the rate of adaptive evolution. We again grouped genes into 10 bins but now based on gene density, with each bin including 5472 scaffolds and 6591 genes on average (Table S5). We observe a negative relationship between Ka+ and gene density with a Spearman's rank correlation equal to −0.345 (Figure 2d). Again, a curvilinear model ($y = a^{bx}$) provided a better fit to the data compared to a linear model, showing that gene density explains 9.90% of the variation in Ka+ (Figure 2d; Table S4). We also calculated correlations between the ratio of nonsynonymous to synonymous substitution rates (K_a/K_s) and recombination rates or gene density. As expected, K_a/K_s was positively correlated with recombination rate (a simple linear model was favoured), while K_a/K_s was negatively correlated with gene density (a curvilinear model was favoured). All these correlations, thus, suggest a widespread influence from HRi on the efficiency of adaptive evolution in the coding regions of Norway spruce. One concern that needs to be addressed in the future is that neither of the correlations between Ka+ and recombination rate or gene density are significant, reflecting the large noisy data sets that

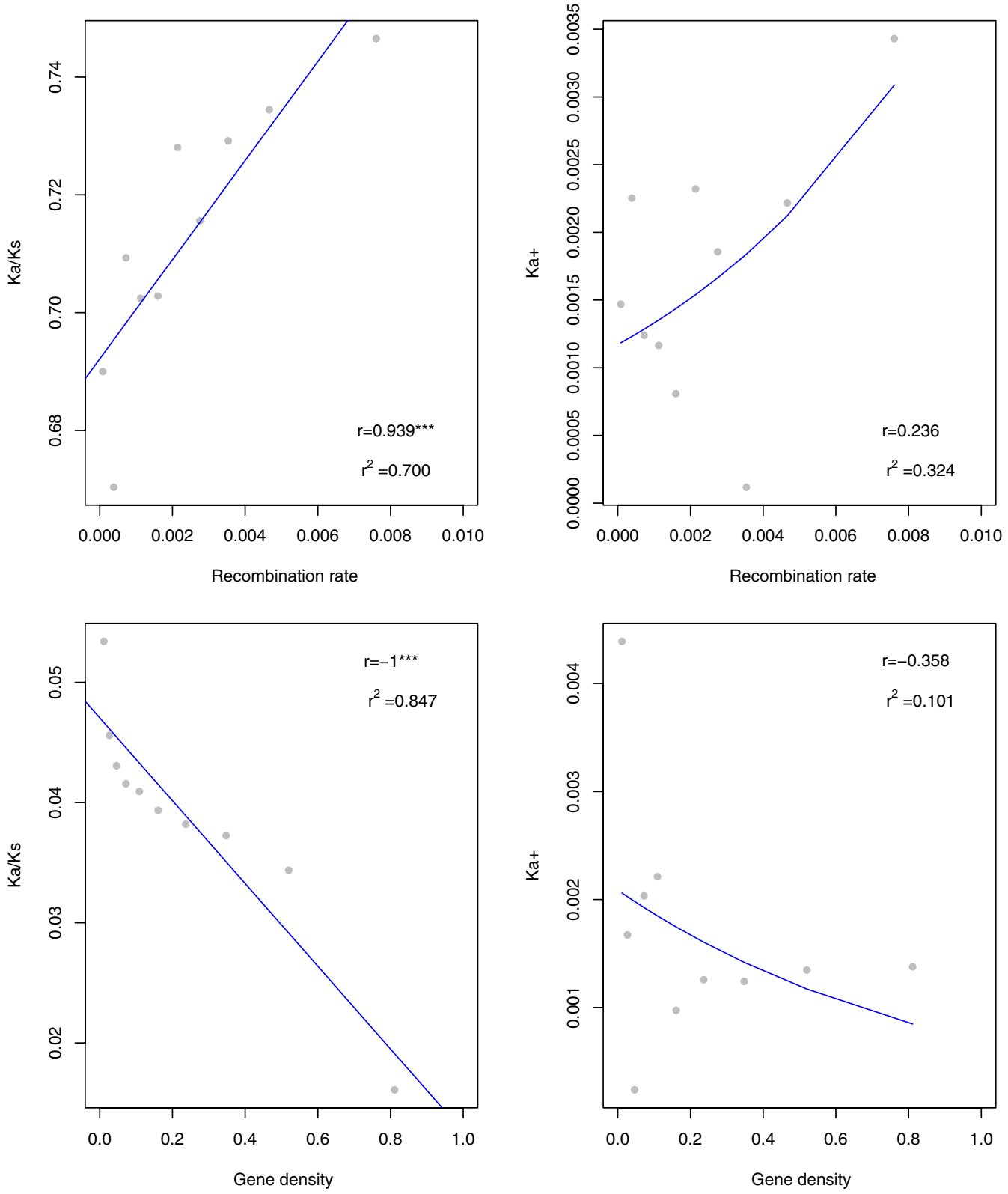


FIGURE 2 Relations between recombination rate in the *x* axis and (a) the ratio of nonsynonymous to synonymous substitution rates (K_a/K_s) in the *y* axis, (b) the rate of adaptive amino acid substitutions (Ka+). Relations between gene density and (c) K_a/K_s, (d) Ka+. Each data point has been estimated binning genes. Parameters for each bin can be consulted in Tables S4 and S5. The Spearman correlation coefficient (*r*) and linear/nonlinear regression (r^2) are shown in each plot. Best fitting regression lines are depicted in blue.

we currently have access to for this study. As genome assemblies are refined and become more contiguous in conifers (Niu et al., 2022; Shalev et al., 2022), the resources for performing the types of analyses we have presented here should improve substantially.

Our results are in accordance with those of Castellano et al. (2016) who found that the rate of adaptive amino acid substitution at a given position of the genome is positively correlated to both the rate of recombination and the mutation rate and negatively correlated with the gene density of the region in *Drosophila*. Castellano et al. (2016), therefore, concluded that HRi hampers the rate of adaptive evolution in *Drosophila* and that the variation in recombination, gene density and mutation along the genome affects the impact of HRi. We did not detect any correlations between Ka+ and mutation rate in Norway spruce, and thus there is no support for the hypothesis that genes with high mutation rates adapt at a higher rate than those with low mutation rates. However, this can largely be attributed to the limited outgroup data we have access to which makes it difficult to estimate mutation rates accurately in our study. Moreover, there is evidence that genes with specific functions, such as immune system in *Drosophila* (Obbard et al., 2009), undergo higher rates of adaptive evolution than other genes. The lack of such information in plants suggests that there is a need for more detailed understanding of how gene functions influence the efficiency of adaptive evolution in plants and especially in forest trees.

Variation in effective population size (N_e) among species also drives variation in the efficacy of selection (Kimura, 1983; Ohta, 1973, 1992). In species with low N_e, the impact of positive selection will decrease, leading to lower fixation rates of adaptive mutations and also longer waiting time for beneficial mutations to arise. In contrast, in species with large populations, selection is more effective resulting in higher rates of fixation of adaptive mutations. The end result is that we expect a positive relationship between population size and the rate of adaptive evolution (Gossmann et al., 2010). Compared to the relatively large estimates of N_e seen in some flowering plants, e.g., ~100,000 in *P. tremula* (Ingvarsson, 2009), ~500,000 in *C. grandiflora* (Foxe et al., 2009), ~832,154 in *H. annuus* and ~733,133 in *H. petiolaris* (Gossmann et al., 2010), Wang, Street, et al. (2020) inferred low effective population size in all of three populations of Norway spruce (~1633 in Finland population, ~4428 in Sweden–Norway population and ~311,241 in Central-Europe population), ultimately resulting in the low rates of adaptive evolution we observe in coding regions. Finally, using the folded SFS is expected to yield a greater underestimation of α than using the unfolded SFS (Charlesworth & Eyre-Walker, 2008). Using data from a single outgroup individual, as we have done in this study, makes it difficult to infer ancestral states accurately, and use of the folded SFS in our calculations can thus be a possible factor contributing to the low estimates for the efficiency of positive selection we observe in Norway spruce.

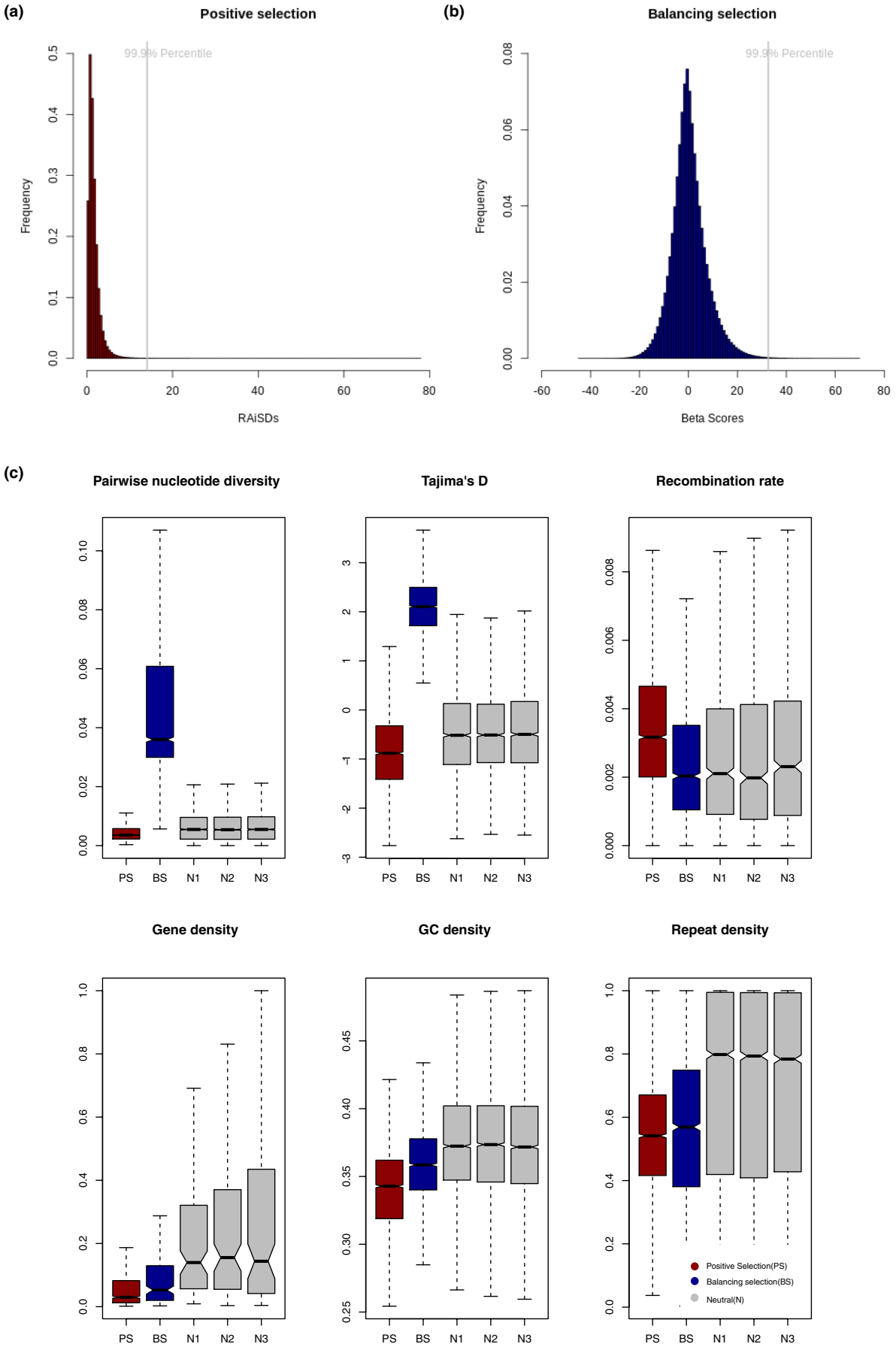
3.4 | Genome-wide scan for regions under positive and balancing selection

Positive selection changes not only the frequency of an advantageous variant but also neighbouring polymorphic sites, leaving distinct patterns in the levels of polymorphism and linkage disequilibrium across the genome (Sabeti et al., 2002). Similarly, balancing selection occurs when multiple alleles are maintained at intermediate frequencies in a population, which can result in their preservation over long evolutionary time periods, leading to an excess number of intermediate frequency polymorphisms near a balanced variant (Siewert & Voight, 2017). Identifying such genomic signatures through genome-wide scans can help identifying genes or genomic regions that are evolving under the influence of natural selection. This will further help us to better understand the role of natural selection in the evolutionary history of a species and aid in interpreting results for regions previously associated with phenotypes of interest.

We used RAiSD to scan the whole genome for signals of selective sweeps in Norway spruce. The RAiSD algorithm accounts for the expected reduction of variation in the region of a putative selective sweep, the shift in the SFS towards low- and high-frequency-derived variants and the emergence of a localized LD pattern characterized by high LD on each side of a beneficial mutation and low LD between loci that are located on different sides of the beneficial allele. We defined the top 0.1% windows of the genome-wide distribution RAiSD values as putative outliers, resulting in a total of 61,756 outlier windows across the whole genome of Norway spruce. These 61,756 windows tag 4208 unique genomic scaffolds and cover a total of 1,594,244 SNPs (Figure 3a; Table S6). We further used a test based on summary statistics, β, to search for signals of long-term balancing selection across the genome for Norway spruce (Siewert & Voight, 2017). We identified a total of 61,451 outlier windows putatively under balancing selection in Norway spruce covering a total of 492,769 variants with those windows (Figure 3b; Table S6).

To understand the impact of positive and balancing selection on genetic variation, we used SNPs within outlier windows that were identified as being under either positive or balancing selection to calculate a variety of population genetic statistics that were then compared to the remaining genomic regions. We also randomly sampled putatively 'neutral' scaffolds, creating pseudo-data sets with similar numbers of scaffolds as we observed in the sets of scaffolds under positive and balancing selection to avoid imbalances in the data sets used to calculate population statistics. The fixation of advantageous mutations increases fitness and skews the SFS towards an excess of low- and/or high-frequency-derived alleles and reduces genetic diversity in the vicinity of the selected site (Dutheil, 2020), while balancing selection maintains advantageous polymorphisms in

FIGURE 3 Genome-wide scan to detect outliers under positive selection and balancing selection for Sweden–Norway population (25 individuals). (a) Histogram distribution of RAiSDs to identify outliers under positive selection. (b) Histogram distribution of Beta Scores to identify outliers under balancing selection. (c) Comparisons between outliers identified representing regions under positive (PS) or balancing (BS) selection, with the remaining neutral genomic regions (N1, N2, N3) using a variety of population genetic summary statistics.



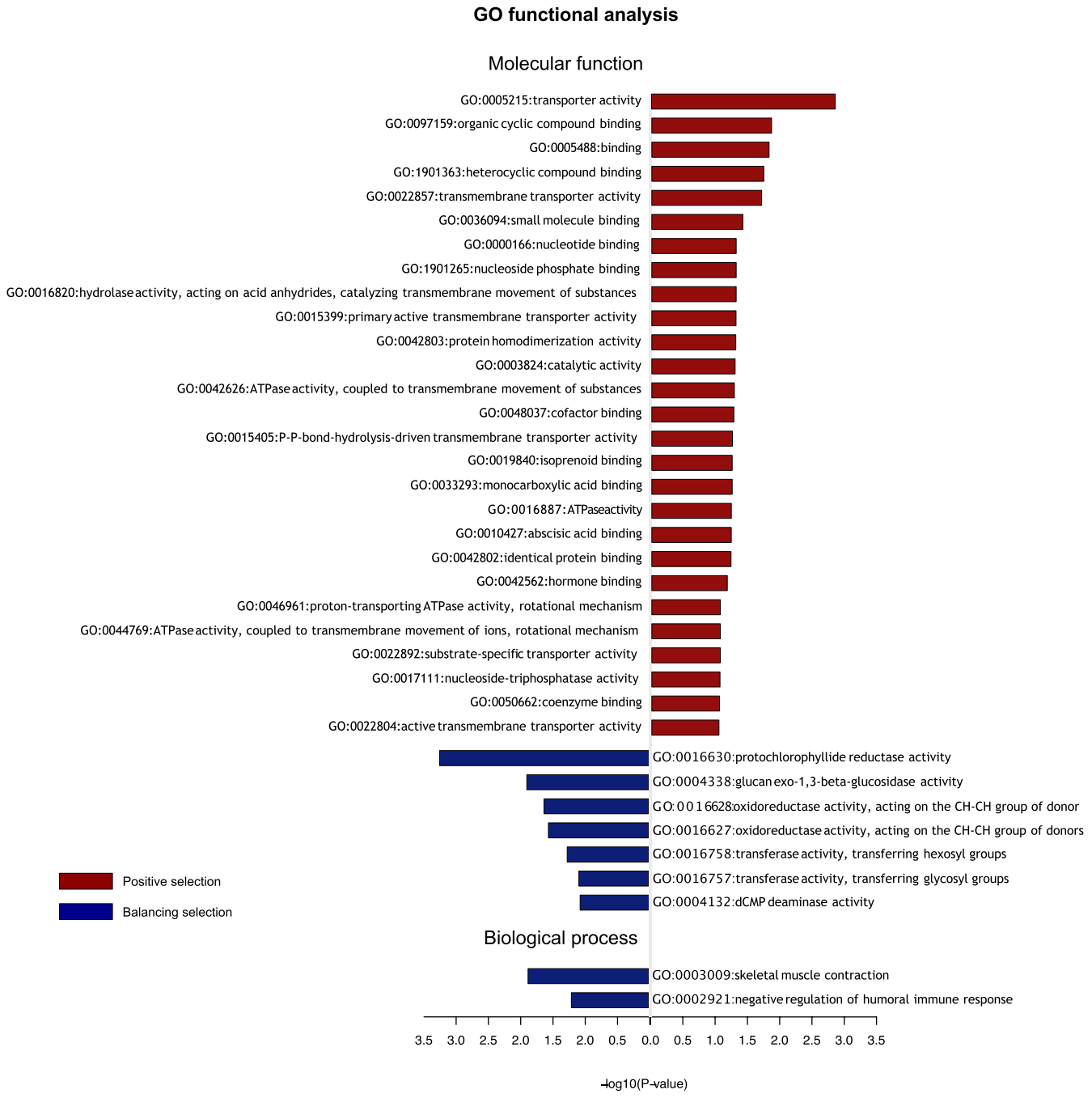


FIGURE 4 Enriched GO categories for genes under positive selection and balancing selection.

populations, leading to an excess number of intermediate frequency polymorphisms and an increased genetic diversity near the balanced variant (de Filippo et al., 2016). In line with predictions from population genetics theory, we observe lower pairwise nucleotide diversity (π=0.005) at SNPs that were classified as evolving under positive selection. Similarly, variants identified as evolving under balancing selection have higher nucleotide diversity (π=0.053) compared to the genome-wide average and as randomly sampled 'neutral' regions (π=0.006–0.007) (Figure 3c; Figure S2; Table S7). These deviations are also mirrored in a summary statistic based on the SFS, Tajima's D, where values are generally positive (2.11) for variants located in regions under balancing selection, suggesting an excess of alleles at intermediate frequencies in those regions as expected from population genetics theory. In contrast, both neutral regions and sites under positive selection showed predominantly negative Tajima's D values. The overall negative values (−0.423 to −0.445) at 'neutral' regions are indicative of a demographic history in Norway spruce characterized by a population expansion following a recent bottleneck (Wang, Bernhardsson, & Ingvarsson, 2020). For sites in regions under positive selection, the Tajima's D values are even more negative (−0.849), which suggest an even greater abundance of rare alleles compared to that in neutral regions which matches with expectations from

regions having experiences recent selective sweeps (Tajima, 1989; Zeng et al., 2006). We found that regions under positive selection had higher recombination rates and lower gene densities compared to neutral regions. These results are in accordance with our previous results that we confirmed a positive correlation between the adaptive rate (Ka+) and recombination rate and the negative correlation between adaptive rate and gene density, both of which are mediated by Hill–Robertson interference (Figure 2). However, compared with neutral regions, we observed lower recombination rates in regions under balancing selection, possibly because the identification of a locus with an old balanced polymorphism is facilitated by low recombination rates that result in the genealogical histories of adjacent SNPs to be more strongly correlated. In these cases, the ability to pinpoint the actual target of selection will also be reduced because larger segments of the genome will be affected (Tian et al., 2002). Moreover, we found that both GC density and repeat density were lower in regions under positive and balancing selection compared to neutral regions. Apuli et al. (2020) found a significantly positive correlation between gene density and GC contents and a significantly negative correlation between repeat density and LD-based recombination rates in European aspen (*P. tremula*). A similar pattern in Norway spruce would explain the low GC content and repeat density for sites under positive selection and the low GC content for sites under balancing selection that we observed in Norway spruce. However, although we detect low repeat densities around sites under balancing selection, there is evidence of balancing selection on transposable elements (TEs). For instance, Chen, Zhang, et al. (2017) and Chen, Glémin, and Lascoux (2017) showed that the heat-shock protein *Hsp*90 is found only in a heterozygote state and seems to display latitudinal variation (Bourgeois & Boissinot, 2019). The role of repeats in regions undergoing balancing selection in Norway spruce is thus worth investigating in greater detail in future studies.

3.5 | Genes under positive and balancing selection

To assess whether there were any specific biological functions that were significantly overrepresented among the genes located in regions identified as undergoing either positive (197 genes) or long-term balancing selection (13 genes), we performed gene ontology (GO) enrichment analysis. We identified 27 significantly enriched GO categories for genes under positive selection, which all belong to different molecular functions (Figure 4, red bars; Table S7). These GO clusters were primarily associated with transporter activity (transmembrane and substrate-specific), binding process (heterocyclic and organic cyclic compound, small molecule, nucleotide, nucleoside phosphate, cofactor, isoprenoid, monocarboxylic acid, identical protein, hormone and coenzyme), ATPase activity, hydrolase activity, protein homodimerization activity, catalytic activity and nucleoside-triphosphatase activity. We only detected one significant GO terms for the candidate genes under long-term balancing selection with FDR corrected $p < .1$, associated with molecular function of protochlorophyllide reductase activity (Figure 4, blue bars). In addition, other GO categories for candidate genes under long-term balancing selection that were significant before false-discovery correction include both molecular function of glucan exo-1,3-beta-glucosidase activity, oxidoreductase activity transferase activity, dCMP deaminase activity and biological process involved in skeletal muscle contraction and negative regulation of humoral immune response (Table S7).

4 | CONCLUSION

In this population genomic survey, we use whole-genome resequencing data from 34 individuals of Norway spruce, spanning most of the natural distribution of the species. We first evaluate the efficacy of both purifying and positive selection across the whole genome of *P. abies*. Our results show that negative selection is limited to coding regions as we fail to detect any evidence of purifying selection in intronic and intergenic regions as well as in two more restricted sets of intergenic sites, ultra-conserved regions between *P. abies* and *Pinus sylvestris* and gene promoters.

We further analysed which factors are important for determining the efficacy of protein adaptation in Norway spruce. We observe a positive correlation between the rate of adaptive evolution and recombination rates and a negative correlation between the rate of adaptive evolution and gene density which both suggest a widespread influence from Hill–Robertson interference. We finally scanned the Norway spruce genome to identify potential outlier regions evolving under either positive or balancing selection and compared population statistics for those outliers with neutral regions. Gene ontology enrichment analysis for genes located in regions identified as undergoing either positive or long-term balancing selection also highlighted specific molecular functions and biological processes in that appear to be targets of selection in Norway spruce. This study constitutes one of the first to find convincing evidence of natural selection within both coding and noncoding genomes for a conifer species. Future studies should aim to use even more broadly sampled populations of the species to further extend the analyses on the importance of adaptive evolution throughout the evolutionary history of Norway spruce.

AUTHOR CONTRIBUTIONS

This study was conceived and designed by PI. Population genomic data analysis was done by XW under the supervision of PI. Interpretation of the results was undertaken by XW and PI. The manuscript was drafted by XW with help from PI.

ACKNOWLEDGEMENTS

The research has been funded by grants from the Knut and Alice Wallenberg Foundation (Norway spruce genome project) and the Swedish Foundation for Strategic Research (SSF, grant no. RBP14-0040). Data generation was supported by Science for Life Laboratory and the National Genomics Infrastructure (NGI) which provided access to massive parallel sequencing. All analyses were performed

on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX) under the projects b2012141, SNIC 2017/1-438, SNIC 2018/3-529, SNIC 2919/3-555 and uppstore2017066. X.W. was supported by a scholarship from the Chinese Scholarship Council (CSC).

CONFLICT OF INTEREST STATEMENT

The authors of this article declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

All sequencing data used in this article are publicly available through the European Nucleotide Archive under study number PRJEB34927. Scripts used to generate all analyses and plots can be found at the Github page https://github.com/xiqtcacf/Norway-Spruce-Scripts.

ORCID

Pär K. Ingvarsson https://orcid.org/0000-0001-9225-7521

REFERENCES

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, *1*(1), 79.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, *437*(7062), 1149–1152.

Apuli, R. P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., Street, N. R., & Ingvarsson, P. K. (2020). Inferring the genomic landscape of recombination rate variation in European aspen (*Populus tremula*). *G3: Genes, Genomes, Genetics*, *10*(1), 299–309.

Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., & Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. *Nature Genetics*, *45*(7), 723–729.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Methodological*, *57*(1), 289–300.

Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D. G., Schiffthaler, B., Baison, J., Street, N. R., García-Gil, M. R., & Ingvarsson, P. K. (2019). An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway Spruce (*Picea abies*). *G3: Genes, Genomes, Genetics*, *9*, 1623–1632.

Bernhardsson, C., Wang, X., Eklöf, H., & Ingvarsson, P. K. (2020). Variant calling using whole genome resequencing and sequence capture for population and evolutionary genomic inferences in Norway spruce (*Picea abies*). In I. Porth, & A. de la Torre (Eds.), *The Spruce genome* (pp. 9–36). SpringerNature.

Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, *15*, 1–10.

Bourgeois, Y., & Boissinot, S. (2019). On the population dynamics of junk: A review on the population genomics of transposable elements. *Genes*, *10*(6), 419.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, *4*(5), e1000083.

Carroll, S. B. (2005). Evolution at two levels: On genes and form. *PLoS Biology*, *3*(7), e245.

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. (2016). Adaptive evolution is substantially impeded by Hill–Robertson interference in drosophila. *Molecular Biology and Evolution*, *33*(2), 442–455.

Charlesworth, J., & Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution*, *23*(7), 1348–1356.

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald–Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*, *25*(6), 1007–1015.

Chen, B., Zhang, B., Xu, L., Li, Q., Jiang, F., Yang, P., Xu, Y., & Kang, L. (2017). Transposable element-mediated balancing selection at Hsp90 underlies embryo developmental variation. *Molecular Biology and Evolution*, *34*(5), 1127–1139.

Chen, J., Glémin, S., & Lascoux, M. (2017). Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, *34*(6), 1417–1428.

Chen, J., Glémin, S., & Lascoux, M. (2020). From drift to draft: How much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? *Genetics*, *214*(4), 1005–1018.

Chen, J., Li, L., Milesi, P., Jansson, G., Berlin, M., Karlsson, B., Aleksic, J., Vendramin, G. G., & Lascoux, M. (2019). Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce. *Evolutionary Applications*, *12*(8), 1539–1551.

Comeron, J. M., Williford, A., & Kliman, R. M. (2008). The Hill–Robertson effect: Evolutionary consequences of weak selection and linkage in finite populations. *Heredity*, *100*(1), 19–31.

Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology*, *13*(4), e1002112.

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nature Reviews Genetics*, *14*(4), 262–274.

de Filippo, C., Key, F. M., Ghirotto, S., Benazzo, A., Meneu, J. R., Weihmann, A., NISC Comparative Sequence Program, Parra, G., Green, E. D., & Andrés, A. M. (2016). Recent selection changes in human genes under long-term balancing selection. *Molecular Biology and Evolution*, *33*(6), 1435–1447.

De La Torre, A. R., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J., Keeling, C. I., MacKay, J., Nilsson, O., Ritland, K., Street, N., Yanchuk, A., Zerbe, P., & Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, *166*(4), 1724–1732.

De La Torre, A. R., Lin, Y. C., Van de Peer, Y., & Ingvarsson, P. K. (2015). Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in picea gene families. *Genome Biology and Evolution*, *7*(4), 1002–1015.

Delph, L. F., & Kelly, J. K. (2014). On the importance of balancing selection in plants. *New Phytologist*, *201*(1), 45–56.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498.

Dutheil, J. Y. (2020). *Statistical population genomics* (p. 468). Springer Nature.

Eckert, A. J., Bower, A. D., Jermstad, K. D., Wegrzyn, J. L., Knaus, B. J., Syring, J. V., & Neale, D. B. (2013). Multilocus analyses reveal little evidence for lineage-wide adaptive evolution within major clades of soft pines (Pinus subgenus S trobus). *Molecular Ecology*, *22*(22), 5635–5650.

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422–433.

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, *26*(9), 2097–2108.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*(2), 737–756.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405.

Flowers, J. M., Molina, J., Rubinstein, S., Huang, P., Schaal, B. A., & Purugganan, M. D. (2012). Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular Biology and Evolution*, *29*(2), 675–687.

Foxe, J. P., Slotte, T., Stahl, E. A., Neuffer, B., Hurka, H., & Wright, S. I. (2009). Recent speciation associated with the evolution of selfing in *Capsella*. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(13), 5241–5245.

Fraïsse, C., Puixeu Sala, G., & Vicoso, B. (2019). Pleiotropy modulates the efficacy of selection in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *36*(3), 500–515.

Gao, Z., Przeworski, M., & Sella, G. (2015). Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution*, *69*(2), 431–446.

Gillespie, J. H. (1991). *The causes of molecular evolution* (Vol. *2*). Oxford University Press On Demand.

Gillespie, J. H. (2000). Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, *155*(2), 909–919.

Gillespie, J. H. (2001). Is the population size of a species relevant to its evolution? *Evolution*, *55*(11), 2161–2169.

Gojobori, J., Tang, H., Akey, J. M., & Wu, C. I. (2007). Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(10), 3907–3912.

Gossmann, T. I., Song, B. H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., & Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, *27*(8), 1822–1832.

Haddrill, P. R., Bachtrog, D., & Andolfatto, P. (2008). Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology and Evolution*, *25*(9), 1825–1834.

Haddrill, P. R., Charlesworth, B., Halligan, D. L., & Andolfatto, P. (2005). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, *6*, 1–8.

Halligan, D. L., Eyre-Walker, A., Andolfatto, P., & Keightley, P. D. (2004). Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Research*, *14*(2), 273–279.

Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eöry, L., Keane, T. M., Adams, D. J., & Keightley, P. D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genetics*, *9*(12), e1003995.

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., & Keightley, P. D. (2010). Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics*, *6*(1), e1000825.

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*, 269–294.

Hough, J., Williamson, R. J., & Wright, S. I. (2013). Patterns of selection in plant genomes. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 31–49.

Ingvarsson, P. K. (2009). Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Molecular Biology and Evolution*, *27*(3), 650–660.

Ingvarsson, P. K. (2010). Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Molecular Biology and Evolution*, *27*, 650–660.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–132). Academic Press.

Keightley, P. D., & 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.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Neuffer, B., Wright, S. I., & Weigel, D. (2019). Long-term balancing selection drives evolution of immunity genes in *Capsella*. *eLife*, *8*, e43606.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 1–13.

Kousathanas, A., Oliver, F., Halligan, D. L., & Keightley, P. D. (2011). Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Molecular Biology and Evolution*, *28*(3), 1183–1191.

Kreitman, M. (1996). The neutral theory is dead. Long live the neutral theory. *Bioessays*, *18*(8), 678–683.

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution*, *29*(1), 33–41.

Lawrie, D. S., Messer, P. W., Hershberg, R., & Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics*, *9*(5), e1003527.

Leitch, I. J., & Bennett, M. D. (1997). Polyploidy in angiosperms. *Trends in Plant Science*, *2*(12), 470–476.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997

Lin, Y. C., Wang, J., Delhomme, N., Schiffthaler, B., Sundström, G., Zuccolo, A., Nystedt, B., Hvidsten, T. R., de la Torre, A., Cossu, R. M., Hoeppner, M. P., Lantz, H., Scofield, D. G., Zamani, N., Johansson, A., Mannapperuma, C., Robinson, K. M., Mähler, N., Leitch, I. J., … Street, N. R. (2018). Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(46), E10970–E10978.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., … Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–482.

Lockton, S., & Gaut, B. S. (2005). Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics*, *21*(1), 60–65.

Machado, H. E., Lawrie, D. S., & Petrov, D. A. (2020). Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. *Genetics*, *214*(2), 511–528.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*(5670), 581–584.

Moutinho, A. F., Bataillon, T., & Dutheil, J. Y. (2020). Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, *34*(3), 315–338.

Neale, D. B., & Wheeler, N. C. (Eds.). (2019). *The conifers: Genomes, variation and evolution* (pp. 1–21). Springer International Publishing.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics*, *39*, 197–218.

Niu, S., Li, J., Bo, W., Yang, W., Zuccolo, A., Giacomello, S., Chen, X., Han, F., Yang, J., Song, Y., Nie, Y., Zhou, B., Wang, P., Zuo, Q., Zhang, H., Ma, J., Wang, J., Wang, L., Zhu, Q., … Wu, H. X. (2022). The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell*, *185*(1), 204–217.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A.,

Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., … Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*(7451), 579–584.

Obbard, D. J., Welch, J. J., Kim, K. W., & Jiggins, F. M. (2009). Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genetics*, *5*(10), e1000698.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, *246*, 96–98.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, *23*(1), 263–286.

Ohta, T., & Gillespie, J. H. (1996). Development of neutral and nearly neutral theories. *Theoretical Population Biology*, *49*(2), 128–142.

Pavy, N., Pelgas, B., Laroche, J., Rigault, P., Isabel, N., & Bousquet, J. (2012). A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biology*, *10*(1), 1–19.

Qiu, S., Zeng, K., Slotte, T., Wright, S., & Charlesworth, D. (2011). Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and *Capsella* species. *Genome Biology and Evolution*, *3*, 868–880.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Sella, G., Petrov, D. A., Przeworski, M., & Andolfatto, P. (2009). Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics*, *5*(6), e1000495.

Shalev, T. J., Gamal El-Dien, O., Yuen, M. M. S., Shengqiang, S., Jackman, S. D., Warren, R. L., Coombe, L., van der Merwe, L., Stewart, A., Boston, L. B., Plott, C., Jenkins, J., He, G., Yan, J., Yan, M., Guo, J., Breinholt, J. W., Neves, L. G., Grimwood, J., … Bohlmann, J. (2022). The western redcedar genome reveals low genetic diversity in a self-compatible conifer. *Genome Research*, *32*(10), 1952–1964.

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005.

Slotte, T., Foxe, J. P., Hazzouri, K. M., & Wright, S. I. (2010). Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution*, *27*(8), 1813–1821.

Smith, N. G., & Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature*, *415*(6875), 1022–1024.

Stephan, W. (2019). Selective sweeps. *Genetics*, *211*(1), 5–13.

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, *28*(1), 63–70.

Strasburg, J. L., Scotti-Saintagne, C., Scotti, I., Lai, Z., & Rieseberg, L. H. (2009). Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. *Molecular Biology and Evolution*, *26*(6), 1341–1355.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Tian, D., Araki, H., Stahl, E., Bergelson, J., & Kreitman, M. (2002). Signature of balancing selection in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11525–11530.

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., Sninsky, J. J., Cargill, M., Adams, M. D., Bustamante, C. D., & Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics*, *5*(8), e1000592.

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, *47*, 97–120.

Wang, J., Street, N. R., Park, E. J., Liu, J., & Ingvarsson, P. K. (2020). Evidence for widespread selection in shaping the genomic landscape during speciation of *Populus*. *Molecular Ecology*, *29*(6), 1120–1136.

Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016). Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, *202*(3), 1185–1200.

Wang, X., Bernhardsson, C., & Ingvarsson, P. K. (2020). Demography and natural selection have shaped genetic variation in the widely distributed conifer Norway spruce (*Picea abies*). *Genome Biology and Evolution*, *12*(2), 3803–3817.

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., & Wright, S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genetics*, *10*(9), e1004622.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216.

Yakovlev, I. A., Asante, D. K., Fossdal, C. G., Junttila, O., & Johnsen, Ø. (2011). Differential gene expression related to an epigenetic memory affecting climatic adaptation in Norway spruce. *Plant Science*, *180*(1), 132–139.

Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, *174*(3), 1431–1439.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang, X., & Ingvarsson, P. K. (2023). Quantifying adaptive evolution and the effects of natural selection across the Norway spruce genome. *Molecular Ecology*, *32*, 5288–5304. https://doi.org/10.1111/mec.17106