A GBAS marker system for assessing genetic diversity and population differentiation in Norway Spruce

Thapasya Vijayan

thapasya.vijayan@boku.ac.at

BOKU University, Institute of Integrative Nature Conservation Research

Christina Rupprecht

BOKU University, Institute of Integrative Nature Conservation Research

Manuel Curto

BOKU University, Institute of Integrative Nature Conservation Research

Bernhard Splechtna

BOKU University, Institute of Integrative Nature Conservation Research

Sebastian Sonnenberg

BOKU University, Institute of Integrative Nature Conservation Research

Harald Meimberg

BOKU University, Institute of Integrative Nature Conservation Research

---

Research Article

Keywords: Norway Spruce, Genotyping by Amplicon Sequencing, Forest genetics, Genetic monitoring, Microsatellite markers, EPIC markers

Posted Date: October 17th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7722464/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Tree Genetics & Genomes on February 12th, 2026. See the published version at https://doi.org/10.1007/s11295-026-01731-2.

Abstract

Effective management of *Picea abies* (Norway spruce), a widely planted conifer species in Europe, requires high-resolution, scalable genotyping tools capable of capturing genetic variation across complex genomes. The primary aim of this study was to establish a cost-effective marker toolkit for *Picea abies* suitable for population genetics, provenance studies, and long-term monitoring. We employed a Genotyping by Amplicon Sequencing (GBAS) approach to investigate sequence-level variation in Norway spruce, using 124 highly multiplexed markers. The markers included 43 novel microsatellite markers from anonymous genomic regions, 33 elongated established microsatellite markers and 48 EPIC (Exon-Primed Intron-Crossing) markers targeting introns. As a proof of concept, the markers were applied to 85 individuals from five populations in Austria. Genotyping was performed using the Genotyping by Amplicon Sequencing (GBAS) pipeline, enabling the extraction of both traditional length-based genotypes and whole amplicon information. Diversity metrics and clustering analyses revealed spatial genetic structure among populations, likely reflecting neutral demographic processes such as elevation-related isolation.

Introduction

Norway spruce (*Picea abies*) is an important species in European forestry, both ecologically and economically. Its broad distribution across much of Europe underscores its adaptability to diverse environmental conditions. However, challenges such as climate change, habitat fragmentation, and emerging pests and diseases affect its genetic integrity and long-term viability (Seidl et al. 2017). These pressures necessitate comprehensive genetic studies to inform forest management strategies and ensure the resilience of spruce populations in changing environments.

A fundamental aspect of forest management is the provenance principle, which asserts that the geographic origin of planting material profoundly influences growth and survival as described for Norway spruce in long-term provenance trials (Cieslar 1899; Günzl 1979; Liesebach et al. 2010). This principle underpins the selection of locally adapted planting material or populations experimentally identified to perform well under specific conditions. Such practices help ensure optimal resource performance while preserving natural genetic diversity that may predate the last glacial maximum, as evidenced by *Picea abies* populations in Norway (Tiret et al. 2023).

Genetic monitoring is an essential tool for assessing the long-term viability of forest tree populations (Aravanopoulos 2011). By assessing genetic diversity and population structure, it enables the detection of changes over time, guiding conservation efforts and sustainable management (Allendorf and Luikart 2007). The field has made significant advances with the establishment of standardized methods for genotyping, which have facilitated accurate assessment of genetic diversity and informed conservation strategies (Neale and Ingvarsson 2008). Despite its ecological significance, genomic research in Norway spruce presents considerable challenges. Its genome, estimated at 20 gigabases, is among the largest and most complex sequenced to date, with a high prevalence of repetitive sequences, duplicated

regions, and extensive non-coding DNA (Nystedt et al. 2013). These features complicate even the application of advanced techniques, such as exome capturing and genotyping-by-sequencing, as the presence of paralogous sequences can lead to ambiguous read mapping and variant calling, thereby affecting the accuracy of genotyping data.

Historically, forest genetic research has relied on various marker systems, including microsatellites, expressed sequence tag-derived markers (ESTs), and allozymes (Thomson et al. 2010; Hong and Lu 2004; Vendramin et al. 2000; Neophytou et al. 2018). In *P.abies*, microsatellites, particularly those derived from ESTs, have been widely used to assess genetic diversity and population structure (Pfeiffer et al. 1997; Rajora et al. 2001; Maghuly et al. 2006; Meloni et al. 2007; Rajora et al. 2014; Rajora and Mann 2021). However, detecting fine scale genetic structure remains challenging in this species due to high gene flow, large population sizes, and extensive historical seed transfers, which maintain low levels of differentiation among populations. Moreover, the complexity of spruce genomes, characterized by paralogous sequences and high homoplasy, adds to the difficulty in marker development and interpretation (Nystedt et al. 2013). To address these limitations and expand available marker systems, we developed a novel group of codominant markers specifically for Genotyping by Amplicon Sequencing (GBAS) in *P.abies*. Most existing markers developed for *Picea abies* have been derived from expressed sequence tags (ESTs) and are thus limited to transcribed regions of the genome. Our markers were designed from anonymous genomic DNA regions, allowing them to target microsatellite motifs outside annotated gene regions and complementing existing resources focused on coding sequences. Intron-flanking markers such as EPICs (Exon-Primed Intron-Crossing) may offer complementary information by targeting additional genic-adjacent regions not captured by EST-SSRs. (Li et al. 2010; Curto et al. 2012). Beyond their genomic position, EPIC markers also present technical advantages over conventional SSRs. Since they target intronic regions, these markers are less prone to genotyping errors caused by stutter bands and size homoplasy, issues commonly associated with short SSR repeat motifs (Schlötterer 2000; Guichoux et al. 2011). Additionally, EPIC markers tend to exhibit lower frequencies of null alleles, as primer binding sites are placed in conserved exonic regions rather than potentially variable flanking sequences (Thomson et al. 2010). Furthermore, because variation in the introns will be linked with functional variation they might be useful in flagging for local adaptation or selection, warranting further investigation into non-neutral evolutionary forces (Jo and Choi 2015). These properties make EPIC markers a practical addition to existing marker sets for routine population genetic studies in Norway spruce.

The objective of our study is to develop and evaluate a set of sequence-based microsatellite and EPIC markers as a cost-effective toolkit for population genetic analysis in *Picea abies*.. Traditional SSR genotyping relies solely on length polymorphisms of the repeat motifs, often overlooking sequence-level variation within the amplicon. Amplicon-based sequencing provides allele calling through complete sequence information, including SNPs and indels within microsatellite motifs and their flanking regions, thereby increasing resolution. We present the first EPIC markers developed for spruce, complementing microsatellite-based markers by enhancing the toolkit with improved resolution and reliability for genetic diversity analyses. In addition, we include previously published EST-derived markers since they are

known to be informative in describing population genetic patterns of *P. abies*. These were specifically restructured to produce longer amplicons (450-550bp) to capture additional flanking variation. Although sequencing itself addresses size homoplasy (Estoup et al. 2002), elongation increases the likelihood of detecting polymorphisms in the flanking regions (Mogg et al. 2002). We present a case where 124 markers can be pooled into as few as 8 multiplex-mixes, demonstrating their application in a dataset of 85 trees sampled from 4 regions across Austria, representing the east, west, and southern parts of the country. We demonstrate the utility of this marker set for resolving genetic diversity and structure in *Picea abies*.

Methods

Screening population and DNA extraction

A total of 85 individuals were collected from five populations across Austria: Rothwald (n = 35), Higher Muhr (n = 15), Lower Muhr (n = 15), Flurwald (n = 10), and Silbertal (n = 10). Rothwald is a primeval forest with minimal human intervention (mainly hunting) (Splechtna and Splechtna 2016), and the other old growth populations sampled are also presumed to have undergone natural regeneration. In Muhr (Salzburg), samples were collected from both valley forests (~ 1100m) and high elevation (~ 1680m) to explore potential altitude-driven population structure. In Vorarlberg, Flurwald and Silbertal represent separate naturally regenerated populations. Within each population, trees were selected maintaining a minimum distance of 20-30m between sampled individuals wherever possible. All sampling was conducted in areas with no visible signs of recent planting or intensive forest management. The collected samples were then processed at the Molecular Biodiversity laboratory at the Institute for Integrative Nature Conservation Research at Boku University, Vienna. The DNA isolation protocol is explained in the Supplementary material. For marker development, three samples with high-quality DNA and minimal fragmentation were sent for library preparation and whole-genome shotgun sequencing on the Illumina MiSeq paired-end (PE) 300 platform at the Genomics Service Unit, Ludwig Maximillian University of Munich, Germany.

Marker development

The approaches used for developing the three marker types are detailed in the following sections. For consistency, we refer to them throughout the manuscript as SSRs (derived from genomic regions), ESTs (elongated versions of previously published EST-SSR markers), and EPIC markers (derived from exon-intron junctions), according to their genomic origins.

To develop SSR markers, the genomic sequences obtained from the MiSeq whole-genome shotgun sequencing were used to identify regions with microsatellite motifs. The raw data underwent quality control (Phred < 20) using Cutadapt 3.0 (Martin 2011) to trim adapters and low-quality bases, ensuring high-quality sequence data. Post quality control, the forward and reverse reads were merged using PEAR v0.9.4 with default parameters (Zhang et al. 2014). The merged reads were then screened for microsatellite containing sequences using the SSR pipeline program (Miller et al. 2013). The settings

used for microsatellite identification included a minimum of 40 bp of flanking sequence on both sides of the repeat region, with minimum thresholds of 6 repeats for tetra- and pentanucleotide motifs, 8 repeats for trinucleotide motifs, and 10 repeats for dinucleotide motifs. The resulting sequences containing microsatellites were subsequently used for primer development using Primer3 (Untergasser et al. 2012) implemented in Geneious v. 8.0.5 (https://www.geneious.com). We specified primer lengths between 18 and 22 bp, an optimal melting temperature (Tm) of 55°C, and a Tm difference of less than 5°C between forward and reverse primers. The target PCR product sizes were set between 400 and 550 bp. Additionally, the GC content was set between 40% and 60%. The designed primers were then mapped back to the Spruce genome (Nystedt et al. 2013) to evaluate their specificity. A total of 93 primer pairs were designed that met all the requirements in the process. The primers were then concatenated with a recognition sequence that corresponded to the Illumina adapter regions as described in (Curto et al. 2019). In addition to developing new markers, existing EST-based SSR markers (Table S1) were included. A total of 47 markers, originally producing amplicons of around 200–300 bp, were selected for elongation. The elongation process to develop EST markers involved mapping the original primers onto the low coverage genome sequences and the original genome to locate the precise genomic regions they amplified. New primers were designed using Geneious v.8.0.5 flanking these regions to elongate the product size to approximately 450–550 bp following other parameters as explained above.

The primer designing strategy for EPIC markers closely followed Curto et al. (2012). We selected a set of functional genes that encompass relevant biological functions in spruce, including genes associated with growth, stress response and immune pathways. Additionally, we included both highly conserved genes, to ensure comparability across populations and species, and less conserved genes, to capture variation that may be specific to local adaptations or ecological pressures. Publicly available mRNA sequences were compiled and used as queries in NCBI BLAST against the spruce genome to identify the corresponding genomic regions. To ensure specificity, only those that resulted in a single hit were selected for further analysis. The regions matching the query mRNA sequences were extracted with 2000 bp flanking sequences on both ends to ensure that the intronic regions were adequately covered. These regions were then imported into Geneious for primer design. Using Primer3 in Geneious, primers were designed to anneal to the exon regions while amplifying the variable introns. Primer design specifications, including length, Tm, GC content, and product size, follow the criteria described in the previous section. To further ensure the cross-species utility of these primers, we mapped them against the *Picea glauca* genome as well. Only primers that produced a single match in both *P.abies* and *P.glauca* genomes were retained. In total, we designed 65 primer pairs. Amplification and sequencing

Primer performance was first evaluated individually on three spruce samples in 10 µl PCR reactions (5 µl QIAGEN Multiplex PCR Master Mix, 2 µl primers, 1 µl DNA) under standard cycling conditions (95°C for 15 min; 35 cycles of 95°C for 30 s, 55°C for 1 min, 72°C for 1 min; final extension 72°C for 10 min). Successful amplification was confirmed by agarose gel electrophoresis, yielding 96 microsatellite-based primers and 52 EPIC primers for multiplex testing. PrimerPooler (Brown et al. 2017) was used to assign primers to eight pools. Multiplex PCRs were performed in 5 µl reactions (384-well format) with the same proportions and cycling conditions, and amplification success was confirmed by randomly checking

products on agarose gels. PCR products from each pool were combined, purified with Agencourt AMPure XP beads (Beckman Coulter Inc.) following Curto et al. (2019), and subjected to a second PCR to add dual indices (10 µl reactions, 10 cycles at 58°C annealing). Indexed products were pooled and sequenced (2 × 250 bp, Illumina Novaseq SP 500, Vienna BioCenter Core Facilities).

Sequence analysis and genotyping

The raw fastq files from Illumina were processed using the SSR-GBAS pipeline (version published in 2019: github.com/mcurto/SSR-GBS-pipeline) which performed quality control, adapter trimming, merging of paired reads, and demultiplexing based on primer sequences. The SSR-GBAS pipeline incorporates widely recognized tools, such as Trimmomatic (Bolger et al. 2014) and USEARCH (Edgar 2010), and has been validated in several studies (Curto et al. 2019; Tibihika et al. 2019; Kwikiriza et al. 2023; Forgiarini et al. 2024; Sinigaglia et al. 2024; Baptista et al. 2024; Tibihika et al. 2024). The pipeline employs a two-step allele-calling process. First, alleles are called based on amplicon length (AL), replicating traditional SSR length genotyping. This is followed by capturing SNP variation within length-based alleles, and the recovery of whole amplicon information (WAI), which provides higher resolution. Each distinct sequence variant is treated as a separate allele and is given an allele number, capturing all polymorphisms within the amplified region, which includes the repeat motif and flanking areas. The output is a standard codominant matrix with integer-coded alleles, where each integer corresponds to an allele per locus. This matrix was converted to the required input files for DAPC and STRUCTURE using the ‘Export’ function of GenAlex 6.503 (Peakall and Smouse 2012). Adaptations to the pipeline have expanded its applicability to other amplicon types, such as EPIC markers.

Marker mapping and annotation

Although primer sequences were initially mapped to verify specificity during marker development, we used allele sequences obtained from the SSR-GBAS pipeline for more accurate classification. To map all the microsatellite allele sequences onto the genome, we aligned them with the *P.abies* reference genome Paab v1.0 from TreeGenes database (Falk et al. 2018; Wegrzyn et al. 2019) using BLASTn (BLAST + v2.6.0) (Camacho et al. 2009). A local BLAST database was created from the scaffold-level genome assembly. Initial searches were conducted with stringent parameters (e-value ≤ 1e-20, minimum 95% identity, minimum 90% query coverage), followed by relaxed searches (e-value ≤ 1e-2, minimum 85% identity) to recover additional potential hits. For each marker, only the highest-scoring alignment (based on bit score and alignment length) was retained. Using Bedtools v2.26.0 (Quinlan and Hall 2010), we classified markers as genic (overlapping exons, CDS, or genes), intronic, or intergenic. For markers with no overlapping annotation the distance to the closest one was calculated using the ‘closest’ function of Bedtools.

Duplication detection

Given the high duplication level of *P. abies* genome, it is possible that a portion of the newly developed markers amplify paralog regions. We tested if we could use the amplicon length histograms to

disentangle the signal from duplicated loci using comparisons with the reference genome as guidance. We expected many duplicated loci to produce amplification products in different size ranges across all samples. If these ranges do not overlap the duplicated signal can be disentangled by calling alleles to these independently.

In the first step, the SSR-GBAS pipeline produces amplicon length histograms (marker plots) for each marker across individuals, where the X-axis represents fragment lengths and the Y-axis indicates their relative frequency. Although these plots are not explicitly designed to detect duplication, would expect multiple, distinct length peaks if duplication was present. They could thus be used as a proxy to detect and disentangle the signal of co-amplification of more than one genomic locus, which we test. To guide this, we first identified markers that produced multiple high-confidence BLAST hits to the *P.abies* genome. For these markers, we compare the reported alignment lengths with the peaks observed in the marker plots. If two non-overlapping peaks were present and matched the fragment sizes, we manually defined length intervals corresponding to each peak (S1). We then repeated the allele calling step, restricting it to these specified intervals. Final allele matrices were generated by merging results across all retained intervals and markers.

Marker performance and population genetic analysis

To assess marker performance, we calculated various diversity metrics, including observed heterozygosity (Ho), expected heterozygosity (He), polymorphic information content (PIC), number of alleles (Na), and Hardy-Weinberg Equilibirum tests (HWE tests) using GenAlEx 6.503 (Peakall and Smouse 2012). FreeNA (Chapuis and Estoup 2006) was used to estimate the proportion of null alleles. All these analyses were conducted using both length-based (AL) and sequence-based (WAI) genotyping data, separately for each marker category allowing us to compare the two genotyping approaches as well.

The obtained marker sets may have different degrees of neutrality. To identify potential outlier loci showing signals of selection, we performed an F_{ST} versus heterozygosity analysis following the method described by Beaumont and Nichols (1996). We calculated pairwise F_{ST} values for each locus, averaged across populations using GenAlEx 6.503. Expected heterozygosity was also calculated using the same dataset. The relationship between F_{ST} and He was visualized in a scatter plot, to infer potential outliers from the main F_{ST} -He distribution. To flag putative outliers, we used an empirical threshold, designating loci with F_{ST} outside the 2.5th -97.5th percentiles of the observed distribution as outliers.

The ability of the marker sets in assessing provenance of *P. abies* was primarily evaluated using population structure and differentiation analyses conducted using WAI data, both separately for each marker category and on the combined dataset. We conducted principal coordinates analysis (PCoA) using GenAlEx 6.503. and used the R package adegenet 2.1.10 (Jombart 2008) for the discriminant analysis of principal components (DAPC) and utilized the optim.a.score function to determine the optimal number of principal components axes to retain through cross-validation. This ensured the

selection of axes that enhance discriminatory power while minimizing noise. Missing data were imputed using the scaleGen function from the package, applying the mean allele frequency method (NA.method = “mean”), which replaces missing values with the mean frequency of observed alleles at each locus (Jombart 2008). After imputation, we further removed any loci or individuals that still contained missing values to ensure that the dataset used for DAPC was fully complete. Further, we used STRUCTURE 2.3.4 (Hubisz et al. 2009) for Bayesian clustering. STRUCTURE was run using 10 independent replicates for 100,000 generations after a burn-in period of 10,000. We tested the K-values between 1 and 5 and evaluated the best K using STRUCTURE SELECTOR (Li and Liu 2018). We tested for isolation by distance using the Mantel test in GenAlEx with 9,999 permutations, which evaluates the correlation between pairwise genetic and geographic distance matrices. The geographic distance between sampling sites were computed using great-circle (haversine) distances from latitude and longitude (R = 6371 km) using the R package ‘geosphere’ (Hijmans 2010). We used this as the X matrix and the pairwise F_{ST} values computed before as the Y matrix for the Mantel test.

Results

SSR search and microsatellite mining

The SSR search identified 189,029 microsatellite-containing sequences out of the 10,056,984 reads produced for marker development. The distribution of SSR motifs revealed a predominance of dinucleotide repeats (89,456), mononucleotide repeats (38,814), followed by trinucleotide repeats (26,348), tetranucleotide repeats (9,753), pentanucleotide repeats (5,201), and hexanucleotide repeats (279). Figure 1a illustrates the most abundant motifs in decreasing order of frequency. Markers in the SSR category were mostly designed to cover pentanucleotide repeats (31 loci), followed by trinucleotides (12) and tetranucleotides (11). This was a deliberate design choice, as longer motifs are less prone to stutter artifacts and therefore increase genotyping reliability. This was not possible for EST category since these are enriched for trinucleotides (26 loci), with fewer dinucleotides (6), tetranucleotides (2), and 1 pentanucleotide.

Sequence analysis and marker retention

After individual primer testing, a total of 76 SSR, 47 EST and 55 EPIC markers were included in the multiplex PCR (S1-S3). The Novaseq sequencing runs yielded a total of 11,714,950 paired reads. Subsequent quality control and merging steps resulted in 10,056,984 high-quality reads suitable for further analysis. All primers were designed for 2 × 300 bp Illumina sequencing runs; therefore, the 2 × 250 bp sequencing configuration did not yield overlapping reads for 20 microsatellite-based markers and 4 EPIC markers. Additionally, markers that failed to produce data for at least 60% of the samples were filtered out (marked in Tables S1-S3). After these filtering steps, a set of 43 SSR, 33 EST, and 48 EPIC markers were retained.

Genomic localization of SSR markers

The Bedtools results showed that 40 out of 43 SSR markers did not overlap with any annotated element. We therefore classified them as having no overlap and potentially intergenic. The three remaining markers overlapped with annotations of exons or CDS regions. In addition, proximity analysis revealed that three other markers were located within 1 kb of an annotated feature and one within 15kb (Table 1). For consistency, we also mapped the EST markers. Although they are expected to localize within expressed regions, we verified their genomic positions. Five of these markers did not overlap with any annotated features, while the rest mapped within exons or introns or showed overlapping classifications. Full results for the EST markers are provided in Table S4. These classifications are based on intersections with the TreeGenes annotation (Paab v1.0), which includes gene, CDS, exon, and intronic features. As this annotation does not cover UTRs or distant regulatory elements, it is likely that some markers classified as intergenic may lie near or within functional, yet annotated, genomic regions.

Markers with potential duplication

We observed potential duplication in six markers based on the presence of multiple mapping hits and distinct, non-overlapping, peaks in their amplicon length distributions. We defined intervals based on the marker plots for all six markers. Three markers, S24_AGA, S93_CATGC, and S95_ATGGT and their intervals met the read threshold of 20 and were included in the final dataset. We named the intervals S24dup_AGA, S93dup_CATGC, and S95dup_ATGGT respectively and included them for the allele call step. Markers S1_ACAAA, S112_ATCAT and S94_ATAAA had read count lesser than 20 in one of the intervals and hence we proceeded the allele calling in the intervals that met the read count. We have provided the mapping results, the defined length intervals and examples of marker plots in the Supplementary material (S5).

Evaluation of AL and WAI genotyping methods in microsatellite-based markers

To evaluate the relative performance of length-based genotyping (Amplicon Length, AL) and sequence-based genotyping (Whole Amplicon Information, WAI), we assessed four key metrics across all SSR and EST loci: observed heterozygosity (Ho), expected heterozygosity (He), polymorphic information content (PIC), and the number of alleles (Na). The mean Ho increased from 0.24 in AL to 0.49 in WAI, and the mean He increased from 0.31 in AL to 0.69 in WAI. The mean PIC was 0.28 for AL and 0.65 for WAI, with the PIC range extending from 0.006–0.79 in AL and 0.08–0.97 in WAI. Furthermore, WAI identified 49 loci with PIC ≥ 0.5, 12 loci with PIC between 0.25 and 0.5, and only 4 loci below 0.25, whereas AL yielded 15 loci with PIC ≥ 0.5, 21 loci between 0.25 and 0.5, and 29 loci below 0.25. These differences highlight the increased resolution of WAI, which detects sequence-level polymorphisms that remain hidden under length-based scoring, thereby inflating both allele counts and heterozygosity estimates. Diversity metrics for EPIC markers are presented separately in the ‘Marker performance’ section (Fig. 1b).

F_{ST} -Heterozygosity outlier analysis

The F_{ST} vs. heterozygosity analysis identified eight outlier loci based on the 95% confidence interval (CI). The green dashed lines indicate the empirical 2.5% and 97.5% quantile thresholds of the F_{ST} distribution,

representing the lower and upper bounds of the 95% range. Four loci fell above the upper bound, while four were below the lower bound (highlighted in red circles in Fig. 2b). The loci above the upper bound, indicating elevated levels of differentiation, were all EST markers: S47_GAG, S87_AG, S88_GA, and S89_AG. These were associated with the following annotations: protein phosphatase 2A regulatory B subunit family protein; MBD9 methyl-CPG-binding domain 9, DME HhH-GPD base excision DNA repair family protein. Conversely, the loci below the lower bound included one SSR marker (S115_AGCCA) and three EPIC markers: 2-SprEPIC_, 31-SprEPIC_, and 32-SprEPIC. The SSR marker below the lower bound was found to overlap with a region annotated as the *peroxidase 31-like* gene MA_102713g0010 (Sundell 2015), including an exon and contiguous intron (Table 1). The EPIC markers below the lower bound correspond to genes with functional annotations, including disease resistance, homeodomain-leucine zipper transcription factor HB-3 (HB-3), and PEPC-1b phosphoenolpyruvate carboxylase, respectively.

Performance across marker categories

Having compared genotyping approaches (length-based vs. whole-amplicon) in the previous section, we assessed how diversity metrics differ across marker categories – SSR, EST, and EPIC- based on WAI data (Fig. 1b). The results indicate that SSR markers captured more allelic diversity and heterozygosity than the other two categories, while EPIC markers, although not based on SSR repeat variation, showed comparable levels of diversity to the EST markers. SSR markers yielded a mean He of 0.79 and Ho of 0.56, PIC of 0.76, and Na of 12.9. EST markers showed a mean He of 0.60, Ho of 0.43, PIC of 0.55, and Na of 7.5. EPIC markers showed a mean He of 0.606, Ho of 0.587, PIC of 0.572, and Na of 9.02.

Population genetic analysis

After removing the outliers from the F_{ST}-heterozygosity analysis and one marker that turned out to be monomorphic across all populations (S55_CAG), we conducted population genetic analyses on the WAI datasets from each type of marker and merged all of them for a combined analysis as well.

SSR markers

Using the SSR markers, we observed that expected heterozygosity (He) consistently exceeded observed heterozygosity (Ho) across all populations, ranging from 0.61 to 0.73 (Table 2). Na was highest in the Rothwald population at 15.6, indicative of substantial allelic richness. No global deviations from Hardy-Weinberg equilibrium (HWE) were detected, although five markers showed significant deviations in at most three populations. All HWE p-values are provided in Table S6. Additionally, none of the markers exhibited a null allele frequency exceeding 0.2 across all populations, as indicated by the FreeNA analysis. As shown in Fig. 2a, pairwise F_{ST} values were lowest between Rothwald and Higher Muhr, while the highest was between Silbertal and Lower Muhr. The DAPC (Fig. 3a) revealed that Rothwald and Flurwald formed distinct clusters, clearly separated from a group comprising Higher Muhr, Lower Muhr, and Silbertal, Vorarlberg. Some individuals from Rothwald overlapped with the Muhr-Silbertal cluster. STRUCTURE analysis at K = 2 (Fig. 3e), identified as the optimal number of clusters by the delta K method from STRUCTURE SELECTOR, showed a primary division with Higher and Lower Muhr and

Rothwald in one cluster. At K = 3, additional substructure emerged, with Rothwald, Flurwald and Silbertal showing increased differentiation. Silbertal shows mixed ancestry across all three clusters despite the geographical closeness with Flurwald.

EST markers

Using only EST markers, we observed a higher Na value in the Rothwald population with a mean of 9.8 (Table 2). Like SSR markers, He consistently surpassed Ho across all populations, ranging from 0.47 to 0.53. No significant deviation from HWE was found and FreeNA analysis indicated no null allele frequency above 0.2. Pairwise F_{ST} values (Fig. 2a) showed the lowest genetic distance between Higher Muhr and Lower Muhr (0.052), while the highest was between Rothwald and Flurwald. The DAPC (Fig. 3b) identified Rothwald separating with few individuals overlapping with Muhr-Flurwald cluster. In contrast to the findings observed with SSR markers, the Flurwald and Silbertal populations exhibited considerable overlap and formed a single cluster. The STRUCTURE analysis (Fig. 3f) corroborated the cluster formation identified in DAPC, with Rothwald demonstrating differentiation at K = 2, while the remaining populations displayed mixed assignments. At K = 3, the clustering pattern remained largely unchanged.

EPIC markers

Using only EPIC markers, we observed that the highest mean Na was observed at 9.5 in the Rothwald population, a pattern consistent with findings from the other marker types. In contrast to the results obtained with SSR and EST markers, He did not exceed Ho across populations. No global deviation from HWE was indicated. Pairwise F_{ST} values (Fig. 2a) indicated the lowest genetic differentiation between Higher Muhr and Silbertal, while the highest was observed between Flurwald and Lower Muhr. The DAPC analysis, retaining 25 axes, resolved into three clusters (Fig. 3c). Unlike the patterns observed with SSR and EST markers, Higher Muhr and Lower Muhr did not cluster together and exhibited considerable overlap with Rothwald. Flurwald, however, separated completely, consistent with its isolation as observed with SSR markers. The STRUCTURE analysis (Fig. 3g) showed at K = 2, Higher Muhr, Lower Muhr and Rothwald showing overlap in clustering, while Flurwald as differentiated, and Silbertal showing mixed ancestry. At K = 3, the result provided limited additional insight with the differentiation of Flurwald persisting, however there is no further resolution.

Combined analysis

Population genetic analysis integrating SSR, EST, and EPIC markers revealed the highest mean Na at 12.4 in Rothwald, with He consistently exceeding Ho across all populations, ranging from 0.54 to 0.61. The lowest pairwise F_{ST} value was observed between Silbertal and Rothwald, while the highest was between Flurwald and Lower Muhr. The DAPC (Fig. 3d), retaining 25 axes, demonstrated that Rothwald formed a distinct cluster, clearly separated from the other populations. The Muhr populations clustered

together and were distinctly apart from the remaining groups, while the Vorarlberg populations (Flurwald and Silbertal) exhibited some overlap yet remained clearly separated from the other clusters. The STRUCTURE analysis (Fig. 3h) corroborated the patterns observed in the DAPC, reinforcing the distinct clustering of Rothwald and the partial overlap among Vorarlberg populations. Using the combined dataset, we tested for isolation by distance using a Mantel test (Figure S7). The test revealed no significant correlation between genetic and geographic distances between our sampling locations ($r = 0.01$, $p = 0.38$). We also conducted a PCoA (Figure S8) for all the four categories. The results were congruent with the clusters observed in DAPC.

Table 1
Genomic localization of SSR markers in the *P.abies* genome

Marker	Feature contig	Start position	End position	Annotation feature	Distance
S1_ACAAA	MA_791733	1152	1651	No overlap	-
S105_TTCTT	MA_101094	16463	16807	No overlap	454
S106_TTTTG	MA_7851150	746	1154	No overlap	-
S107_AGATA	MA_10187907	1562	1882	No overlap	-
S109_TTATC	MA_169129	38362	38762	No overlap	-
S11_TTA	MA_474718	2630	2977	No overlap	-
S12_ATG	MA_10231278	1077	1415	No overlap	-
S110_AATAC	MA_9342	17732	18140	No overlap	14923
S113_AAAAG	MA_139673	16731	17145	No overlap	-
S114_TTTTA	MA_219245	1888	2286	No overlap	-
S115_AGCAA	MA_102713	16604	16855	Exon	
		16856	17773	Intron	
S116_AGATA	MA_10056486	6556	6960	No overlap	-
S118_TTTTC	MA_149572	27924	28324	No overlap	-
S121_ATTTT	MA_909778	9009	9456	No overlap	-
S123_CTTCT	MA_615504	9305	9702	No overlap	-
S17_TCTTT	MA_10282294	326	765	No overlap	-
S18_TTTCC	MA_62292	9362	9834	No overlap	-
S19_CATCA	MA_182539	18364	18841	No overlap	-
S20_ATTT	MA_6123658	161	457	No overlap	
S23_ATG	MA_10335377	2785	3217	No overlap	-
S25_ATT	MA_9191613	1155	1609	No overlap	920
S27_ATA	MA_75794	16674	16993	No overlap	-
S31_AGA	MA_151401	430	874	No overlap	-
S32_AAT	MA_8775	11090	11518	No overlap	946
S33_TTC	MA_48140	2812	3137	No overlap	-
S35_TAAAA	MA_683520	892	1333	No overlap	-

Marker	Feature contig	Start position	End position	Annotation feature	Distance
S38_GATA	MA_58938	7529	7949	No overlap	-
S39_CATG	MA_541813	788	1209	No overlap	-
S4_AATAA	MA_13035	5308	5836	No overlap	
S42_TTAT	MA_363612	3540	3968	No overlap	
S43_TTTA	MA_123109	24438	24687	No overlap	
S94_ATAAA	MA_613225	8607	9002	No overlap	-
S95_ATGGT	MA_92201	4904	5315	CDS	
S96_TGAAT	MA_196748	7554	7680	Intron	
		7681	7742	Exon	
		7743	7887	Intron	
		7888	7948	Exon	
S97_GGAAA	MA_711665	1780	2149	No overlap	-
S98_AAAAG	MA_58289	1	390	No overlap	-
S99_TTGGC	MA_9313675	1760	1933	No overlap	-
S13_CAT	MA_24700	3639	4106	No overlap	-
S2_AAAAT	MA_94131	4255	4771	No overlap	-
S48_AGG	MA_10432	9878	10361	No overlap	
S61_AGG	MA_884	39102	39531	Exon	
S71_ATT	MA_50379	12708	13188	Exon	

Table 2
Diversity metrics (mean ± SE), including mean number of different alleles (N_a), mean number of effective alleles (N_e), mean observed heterozygosity (H_o), and mean expected heterozygosity (H_e), for each population using only the SSR markers, EST markers, EPIC markers and all markers combined.

Population	N	N_a	Ne	H_o	H_e
SSR					
Higher Muhr	10.67 ± 0.12	10.77 ± 0.7	4.9 ± 0.577	0.517 ± 0.008	0.652 ± 0.006
Lower Muhr	9.25 ± 0.108	9.5 ± 0.622	4.6 ± 0.551	0.465 ± 0.008	0.613 ± 0.007
Rothwald	18.83 ± 1.29	15.61 ± 1.03	6.498 ± 0.78	0.525 ± 0.006	0.735 ± 0.004
Flurwald	6.93 ± 0.066	8.5 ± 0.46	4.25 ± 0.58	0.534 ± 0.007	0.647 ± 0.006
Silbertal	6.953 ± 0.068	8.43 ± 0.47	4.403 ± 0.12	0.49 ± 0.008	0.626 ± 0.007
EST					
Higher Muhr	9.5 ± 0.343	7.0 ± 0.68	2.618 ± 0.109	0.408 ± 0.018	0.473 ± 0.016
Lower Muhr	7.937 ± 0.265	6.4 ± 0.55	2.474 ± 0.080	0.454 ± 0.02	0.485 ± 0.015
Rothwald	17.5 ± 0.444	10.4 ± 0.99	3.611 ± 0.145	0.554 ± 0.016	0.597 ± 0.015
Flurwald	6.625 ± 0.164	6.26 ± 0.53	2.687 ± 0.108	0.454 ± 0.019	0.489 ± 0.017
Silbertal	7.062 ± 0.143	6.13 ± 0.52	2.841 ± 0.121	0.403 ± 0.019	0.486 ± 0.018
EPIC					
Higher Muhr	11.75 ± 0.111	7.21 ± 0.41	2.804 ± 0.040	0.574 ± 0.006	0.515 ± 0.005
Lower Muhr	10.2 ± 0.098	6.51 ± 0.37	2.72 ± 0.038	0.557 ± 0.006	0.505 ± 0.005
Rothwald	16.9 ± 0.189	9.55 ± 0.58	3.142 ± 0.054	0.534 ± 0.007	0.515 ± 0.005
Flurwald	7.08 ± 0.068	5.59 ± 0.34	2.442 ± 0.033	0.467 ± 0.006	0.465 ± 0.005
Silbertal	7.041 ± 0.062	6.19 ± 0.33	2.558 ± 0.030	0.541 ± 0.006	0.509 ± 0.005
Combined					
Higher Muhr	10.9 ± 0.52	8.84 ± 0.38	3.6 ± 0.28	0.53 ± 0.02	0.56 ± 0.027
Lower Muhr	9.48 ± 0.451	7.91 ± 0.32	3.44 ± 0.26	0.50 ± 0.034	0.546 ± 0.029
Rothwald	17.87 ± 0.82	12.4 ± 0.56	4.56 ± 0.393	0.534 ± 0.030	0.616 ± 0.026
Flurwald	6.953 ± 0.029	7.21 ± 0.28	3.21 ± 0.213	0.493 ± 0.003	0.542 ± 0.002
Silbertal	7.009 ± 0.027	7.21 ± 0.027	3.342 ± 0.226	0.5 ± 0.031	0.552 ± 0.028

Discussion

We provide a genotyping framework for Norway spruce based on a comprehensive Genotyping by Amplicon Sequencing (GBAS) approach, incorporating multiple marker types. Despite the challenges posed by the large and highly repetitive genome, most primers successfully amplified target loci. This system incorporates SSR markers containing microsatellite motifs and EPIC markers, while also extending the flanking regions of previously published EST-SSR loci to increase its polymorphic content. This approach provides a scalable toolkit for studying genetic diversity of species with large and repetitive genomes and enabling long-term genetic monitoring.

While whole-genome sequencing offers the most comprehensive overview of genetic variation, its application to species with very large genomes such as Norway spruce remains constrained by cost and computational demands (Ellegren 2014). Reduced-representation approaches such as RADseq and GBS provide alternatives but suffer from uneven genome coverage and missing data (Lowry et al. 2017). Targeted sequencing methods like exon capture, though highly specific, require prior genomic knowledge, limiting their use in non-model species (Jones and Good 2016). GBAS represents an intermediate solution, balancing marker density, genome-wide representation, and cost efficiency while ensuring reproducible amplification across a large set of samples (Curto et al. 2019). Its successful application in a wide range of taxa, including hedgehogs, tilapiines, and other systems (Curto et al. 2019; Tibihika et al. 2019; Kendlbacher et al. 2025; Sinigaglia et al. 2024; Baptista et al. 2024), highlights its versatility for population-level studies in complex genomes.

Unlike traditional EST-based SSRs, the newly developed markers in this study were designed from anonymous genomic sequences, leveraging advancements in sequencing technologies to create highly polymorphic markers exhibiting varying degrees of neutrality. For instance, the SSR markers exhibited an average PIC of 0.76, Na of 12.9, significantly surpassing EST markers (PIC: 0.54, Na: 8). These findings align with previous studies demonstrating that anonymous genomic SSRs capture more neutral diversity due to less selective constraints unless linked to a region under selection (Morgante et al. 2002; Ellis and Burke 2007; Simko 2009). We included tetranucleotide and pentanucleotide loci in our marker set to increase the range of SSR motif types. Longer motifs are generally less prone to stutter and homoplasy than di- and tri nucleotide repeats, which can improve allele calling accuracy and reduce genotyping error (Schlötterer 2000; Guichoux et al. 2011). The inclusion of these longer motifs provides a more comprehensive survey of the spruce genome enabling examination of previously underrepresented loci. The elongation of EST-SSR markers through newly designed primers by extending the length of the PCR amplicons allows for the capture of additional sequence-level variation, including SNPs and indels, increasing informativeness beyond simple repeat number variation.

Despite efforts to ensure careful primer design, we suspect that a small proportion of the targeted loci correspond to duplicated regions, as indicated by distinct non-overlapping peaks in amplicon length distributions. This happened despite the implementation of a quality control step during marker development that enriched single copy markers through comparisons with available genome sequences.

This likely stems from the Norway spruce genome being at the contig level, where fragmented assemblies can lead to redundant regions being mistaken for unique loci (Schatz et al. 2012). This raises challenges in resolving gene copy status during marker development, especially in large and repetitive genomes. The same challenge applies to allele calls where again amplicons from paralogs may be interpreted as homologous variants (Mastretta-Yanes et al. 2014), highlighting the need for alternative approaches to deal with duplicated loci. Our test confirmed that loci identified as duplicated in genome mapping followed the expected amplicon length distribution, thereby validating the recent developments in the GBAS pipeline that define length intervals for paralogous amplicons. This allows allele calls to be restricted to homologous variants within each interval (publication in preparation, https://github.com/mcurto/SSR-GBS-pipeline). Although further validation is required, this exploratory approach highlights a promising direction for handling duplicated regions.

The Exon-Primed Intron-Crossing (EPIC) markers designed in this study represent the first set for spruce which targets intronic regions. The main goal of the application of the EPIC markers was to evaluate population genetic patterns. However, by specifically targeting functional loci, such as those associated with disease resistance and stress tolerance, we open the possibility to flag loci associated with local adaptation for further investigation. Functional annotation of the genes containing these markers indicated involvement in processes such as developmental regulation, lignin biosynthesis, transcription factors, and responses to environmental cues such as dehydration response etc (Table S3). The markers showed high amplification success in our multiplex PCRs, exceeding that of the SSR and EST markers. Three EPIC markers emerged as potential outliers in the F_{ST}-He analysis, with functions related to disease resistance, transcriptional regulation of homeodomain-leucine zipper *HB-3*, and phosphoenolpyruvate carboxylase *PEPC-1b*. HD-Zip transcription factors such as *HB-3* are plant-specific regulators involved in developmental processes and abiotic stress responses, including drought and ABA signaling pathways (Li et al. 2022). Phosphoenolpyruvate carboxylase functions in photosynthetic carbon fixation and general metabolic reactions (O'Leary et al. 2009). The identification of these loci as outliers hints at their potential adaptive relevance and makes them potential candidates for follow-up functional or environmental association studies. In addition to the EPIC outliers, five microsatellite-based markers were identified outside the empirical 95% F_{ST} range. Four of these were EST, while one was an SSR marker. Genome mapping revealed that this locus targeted by the SSR marker S115_AGCCA is located within the gene MA_102713g0010 in the genome, indicating that this locus resides within a coding region and is therefore less likely to behave as neutral. The four EST markers corresponded to genes that are involved in signal transduction, epigenetic regulation, and DNA repair, all of which can influence plant developmental and stress responses. Like the EPIC outliers, these outliers could also be subjected to further investigation. Since this study focuses on population genetics patterns resulting mostly from demographic processes, we excluded all potential non-neutral loci from all marker categories, thereby reducing the likelihood that the final marker set included loci under selection and increasing confidence that it reflected neutral variation predominantly.

The transition from length-based genotyping (AL) to whole amplicon information (WAI) provided a significant improvement in extracting genetic variation from microsatellite loci. More specifically, diversity metrics such as Ho, He, and PIC increased when WAI was used, reflecting an increase in the number of alleles. By sequencing the entire amplicon, WAI captured additional sequence-level variation, including SNPs and indels, which are often missed in traditional length-based approaches. This enhanced resolution has been demonstrated in several taxa, where WAI outperformed AL in resolving population structure and uncovering finer-scale genetic diversity (Curto et al. 2019; Tibihika et al. 2019; Kwikiriza et al. 2023; Sinigaglia et al. 2024; Baptista et al. 2024; Forgiarini et al. 2024). The inclusion of flanking regions disentangles some of the homoplasy typical to microsatellite markers decreasing noise in population genetics analysis (Curto et al 2018).

Population genetic patterns in Norway spruce

We sampled 85 individuals across Norway spruce populations located in eastern, western, and southern regions of Austria. Vorarlberg, Austria's westernmost region bordering Switzerland, is geographically distant from the other sites, whereas Rothwald is an old-growth primary forest and thus a useful point of comparison to locations that may include planted or seed-introduced material.

We observed that the combined panel integrated signals detected independently by multiple marker categories, whereas individual marker categories emphasized different facets of the same underlying pattern. SSR markers and EPIC markers captured similar overall differentiation (mean pairwise F_{ST} = 0.146 and 0.143, respectively), yet their clustering outcomes diverged. With EPIC, DAPC resolved a within-region split between Higher Muhr and Lower Muhr (altitudinally distinct populations), placed Rothwald closer to Muhr and Silbertal, and kept Flurwald distinct. The SSR panel, by contrast, emphasized inter-regional contrasts with Rothwald and Vorarlberg against the rest and did not separate Higher Muhr and Lower Muhr as clearly, while a few Rothwald individuals overlapped with Muhr and Silbertal. The EST markers showed the lowest mean F_{ST} (0.072) and the least distinct clusters comparably. Because all analyses used WAI and hence size homoplasy is not the explanation, a more plausible reason could be that these loci carried fewer distinct haplotypes as shown by the PIC and Na values and are more constrained than SSR markers. That yields smaller among-population frequency contrasts and consequently weaker signal. Thus, it is surprising that the Higher and Lower Muhr separation was not detected by the SSRA markers that contained higher information content. An explanation can be the lower likelihood of EPIC markers in accumulating amplification-based artifacts. The EPIC dataset had an exceeding Ho and hence lesser heterozygote deficit, which points to fewer null allele issues. With fewer such artefacts, small allele frequency differences between nearby or altitudinally separated populations like the Muhr populations could have been picked by EPIC markers and not the remaining marker sets. An alternative explanation could be selective differences between higher and lower altitude populations. However, after removing potential candidate non-neutral loci across categories this is unlikely.

Geographical separation among the sample populations, combined with differences in altitude, could be expected to produce some degree of spatial genetic differentiation. However, the absence of a statistically significant Mantel correlation indicates that isolation by distance is not the primary driver of the observed structure. In spruce, pollen and seed movement is typically local on the order of meters to tens of meters and occasional long-distance dispersal reaching up to a kilometer (Burczyk et al. 2004; Shimono et al. 2011; O'Connell et al. 2007). In Europe, Norway spruce has also been widely planted using mixed or non-local seed sources (Jansen et al. 2017). Long-distance gene flow together with historical seed transfers can weaken the effects of isolation by distance. Historical processes add a further layer: Different glacial refugia and subsequent migration routes have shaped the genetic landscape of Austrian Norway spruce populations, resulting in varying degrees of admixture and differentiation (Maghuly et al. 2006).

These historical and anthropogenic factors could have contributed to the patterns we see in this study, reflecting the complex interplay between natural recolonization processes and human activities. Our sampling represents a modest subset of the species' distribution in Austria. To assess how general these patterns are, future work should apply the markers to larger sample sizes, denser spatial coverage, and additional regions. Broader geographic context and increased replication would improve power to detect weak isolation-by-distance, clarify the roles of planting and postglacial history, and test whether the population partitions we observe here persist on a larger scale.

As a practical recommendation, we suggest analysing the combined WAI dataset after neutrality filtering as a conservative baseline for demographic inference that integrates evidence across marker categories. At the same time, single-panel results could be examined alongside the pooled analysis, because panel-specific contrasts provide complementary signals that are valuable for follow-up. This two-tier approach reduces reliance on any one marker type while retaining biologically informative patterns that may be specific to a subset of loci.

Conclusion

This study establishes the first comprehensive GBAS framework for *Picea abies*, integrating newly developed SSRs and the first EPIC markers designed for this species. By combining markers from different genomic sources with a standardized sequencing-based genotyping strategy, the approach overcomes limitations of traditional length-based SSR scoring and improves reproducibility across studies. The deliberate inclusion of longer repeat motifs and extended amplicons increases informativeness while addressing common sources of error such as homoplasy and stutter. Together, these features provide a scalable and cost-effective toolkit for population genetic research in spruce and related conifers. Beyond serving as a proof of concept, the framework offers a robust basis for long-term genetic monitoring and the conservation management of forest resources in species with large and complex genomes.

Declarations

Supplementary Information

The online version contains supplementary material.

Acknowledgements

We are grateful to Eva Dornstauder-Schrammel for her skilled technical assistance. Finally, we would like to thank the anonymous reviewers for their comments and suggestions that helped us improve this manuscript. We are grateful to the Bundesministerium für Landwirtschaft, Regionen und Tourismus for the funding.

Funding Information

This project was supported by the the Bundesministerium für Landwirtschaft, Regionen und Tourismus (Forschungsprojeckt Nr. 101660, GZ BMLRT 2021-0.370.496).

Conflict of interest

The authors declare that they have no conflict of interest.

Data archiving statement

The Illumina raw reads generated in this study are available on NCBI under the BioProject ID: PRJNA1260055

References

1. Allendorf F, Luikart G (2007) Conservation and genetics of populations. Blackwell Publishing, Malden, MA.
2. Aravanopoulos FA (2011) Genetic monitoring in natural perennial plant populations. Botany 89(2):75–81. 10.1139/b10-087
3. Baptista L, Curto M, Waeschenbach A, Berning B, Santos A, Ávila S, Meimberg H (2024) Population genetic structure and ecological differentiation in the bryozoan genus Reteporella across the Azores Archipelago (central North Atlantic). Heliyon 10:e38765. 10.1016/j.heliyon.2024.e38765
4. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. 10.1093/bioinformatics/btu170
5. Brown SS, Chen Y-W, Wang M, Clipson A, Ochoa E, Du M-Q (2017) PrimerPooler: automated primer pooling to prepare library for targeted sequencing. Biology Methods Protocols 2(1).

10.1093/biomethods/bpx006

6. Burczyk J, Lewandowski A, Chałupka W (2004) Local pollen dispersal and distant gene flow in Norway spruce (Picea abies [L.] Karst. For Ecol Manag 197:39–48. 10.1016/J.FORECO.2004.05.003
7. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10(1):421. 10.1186/1471-2105-10-421
8. Chapuis M-P, Estoup A (2006) Microsatellite Null Alleles and Estimation of Population Differentiation. Mol Biol Evol 24(3):621–631. 10.1093/molbev/msl191
9. Cieslar A (1899) Neues aus dem Gebiete der forstlichen Zuchtwahl. Ein wissenschaftlicher Beitrag zum Waldbau und zum Forstkulturwesen insbesondere. CbI f d g Fw 25:49–74Separatabdruck 44 S.
10. Curto M, Winter S, Seiter A, Schmid L, Scheicher K, Barthel LMF, Plass J, Meimberg H (2019) Application of a SSR-GBS marker system on investigation of European Hedgehog species and their hybrid zone dynamics. Ecol Evol 9(5):2814–2832. 10.1002/ece3.4960
11. Curto MP, Ferreira P, Meimberg DNM H (2012) Development of phylogenetic markers from single-copy nuclear genes for multi locus, species level analyses in the mint family (Lamiaceae). Mol Phylogenet Evol 63:758–767
12. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461. 10.1093/bioinformatics/btq461
13. Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol 29(1):51–63. 10.1016/j.tree.2013.09.008
14. Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99(2):125–132. 10.1038/sj.hdy.6801001
15. Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11(9):1591–1604. 10.1046/j.1365-294X.2002.01576.x
16. Falk T, Herndon N, Grau E, Buehler S, Richter P, Zaman S, Baker EM, Ramnath R, Ficklin S, Staton M, Feltus FA, Jung S, Main D, Wegrzyn JL (2018) Growing and cultivating the forest genomics database, TreeGenes. Database 2018. 10.1093/database/bay084
17. Forgiarini C, Meimberg H, Curto M, Stiehl-Alves EM, Vijayan T, Engl PT, Bräuchler C, Kollmann J, de Souza-Chies TT (2024) Low genetic differentiation despite high habitat fragmentation in an endemic and endangered species of Iridaceae from South America: implications for conservation. Bot J Linn Soc. 10.1093/botlinnean/boae036
18. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11(4):591–611. 10.1111/j.1755-0998.2011.03014.x
19. Günzl L (1979) Internationale Fichten-Provenienzversuche der IUFRO 1938 und 1964/68 sowie Versuche mit österreichischen Herkünften. Allgemeine Forstzeitung (Wien) 90(7):182–190
20. Hijmans RJ (2010) geosphere: Spherical Trigonometry. R package version 1.5–20

21. Hong L, Lu M-Z (2004) Expressed Sequence Tag (EST) and Its Application in Forest Research. Forest research
22. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9(5):1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x
23. Jansen S, Konrad H, Geburek T (2017) The extent of historic translocation of Norway spruce forest reproductive material in Europe. Annals of Forest Science 74
24. Jo BS, Choi SS (2015) Introns: The Functional Benefits of Introns in Genomes. Genomics Inf 13(4):112–118. 10.5808/gi.2015.13.4.112
25. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24(11):1403–1405. 10.1093/bioinformatics/btn129
26. Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. Mol Ecol 25(1):185–202. 10.1111/mec.13304
27. Kendlbacher V, Krenn YP, Vijayan T, Rupprecht C, Spießberger M, Curto M, Landler L, Meimberg H (2025) Genotyping by Amplicon Sequencing (GBAS) With Newly Developed SSR and EPIC Markers Reveals Structure in Populations of the Green Toad (Bufotes viridis) Across Rural and Urban Environments. Ecol Evol 15(7):e71652. https://doi.org/10.1002/ece3.71652
28. Kwikiriza G, Vijayan T, Tibihika PD, Curto M, Winkler G, Nattabi J, Kariuki J, Meimberg H (2023) Introgressive hybridization levels of Tilapiine species in Lake Victoria basin, Kenya inferred from microsatellite and mitochondrial DNA genotyping based on next-generation sequencing. Conserv Genet 25:1–14. 10.1007/s10592-023-01570-x
29. Li C, Riethoven JJ, Ma L (2010) Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. BMC Evol Biol 10:90. 10.1186/1471-2148-10-90
30. Li Y, Yang Z, Zhang Y, Guo J, Liu L, Wang C, Wang B, Han G (2022) The roles of HD-ZIP proteins in plant abiotic stress tolerance. Front Plant Sci Volume 13–2022. 10.3389/fpls.2022.1027071
31. Li YL, Liu JX (2018) StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. Mol Ecol Resour 18(1):176–177. 10.1111/1755-0998.12719
32. Liesebach M, Rau HM (2010) König AO Fichtenherkunftsversuch von 1962 und IUFRO-Fichtenherkunftsversuch von 1972. In
33. Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A (2017) Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour 17(2):142–152. 10.1111/1755-0998.12635
34. Maghuly F, Pinsker W, Praznik W, Fluch S (2006) Genetic diversity in managed subpopulations of Norway spruce *Picea abies* (L.) Karst. For Ecol Manag 222(1–3):266–271. 10.1016/j.foreco.2005.10.025
35. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17 (1):3. 10.14806/ej.17.1.200

36. Mastretta-Yanes A, Zamudio S, Jorgensen TH, Arrigo N, Alvarez N, Piñero D, Emerson BC (2014) Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. Genome Biol Evol 6(10):2611–2624. 10.1093/gbe/evu205
37. Meloni M, Perini D, Binelli G (2007) The distribution of genetic variation in Norway spruce (*Picea abies* Karst.) populations in the western Alps. J Biogeogr 34(6):929–938. 10.1111/j.1365-2699.2006.01668.x
38. Miller MP, Knaus BJ, Mullins TD, Haig SM (2013) SSR_pipeline—Computer software for the identification of microsatellite sequences from paired-end Illumina High-Throughput DNA sequence data. US Geological Survey Data Series 778
39. Mogg R, Batley J, Hanley S, Edwards D, O'Sullivan H, Edwards K (2002) Characterization of the flanking regions of Zea mays microsatellites reveals a large number of useful sequence polymorphisms. Theor Appl Genet 105(4):532–543. 10.1007/s00122-002-0897-1
40. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30(2):194–200. 10.1038/ng822
41. Neale D, Ingvarsson P (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11 2:149–155. 10.1016/j.pbi.2007.12.004
42. Neophytou C, Torutaeva E, Winter S, Meimberg H, Hasenauer H, Curto M (2018) Analysis of microsatellite loci in tree of heaven (Ailanthus altissima (Mill.) Swingle) using SSR-GBS. Tree Genet Genomes 14. 10.1007/s11295-018-1295-4
43. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451):579–584. 10.1038/nature12211
44. O'Connell L, Mosseler A, Rajora O (2007) Extensive long-distance pollen dispersal in a fragmented landscape maintains genetic diversity in white spruce. J heredity 98 7:640–645. 10.1093/JHERED/ESM089
45. O'Leary B, Rao SK, Kim J, Plaxton WC (2009) Bacterial-type phosphoenolpyruvate carboxylase (PEPC) functions as a catalytic and regulatory subunit of the novel class-2 PEPC complex of vascular plants. J Biol Chem 284(37):24797–24805. 10.1074/jbc.M109.022863
46. Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics 28(19):2537–2539. 10.1093/bioinformatics/bts460
47. Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (Picea abies K). Genome 40(4):411–419

48. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. 10.1093/bioinformatics/btq033
49. Rajora O, Forneris N, Shi Y (2014) Highly Informative Single-Copy Nuclear Microsatellite DNA Markers Developed Using an AFLP-SSR Approach in Black Spruce (Picea mariana) and Red Spruce (P. rubens). PLoS ONE 9. 10.1371/journal.pone.0103789
50. Rajora OP, Mann IK (2021) Development and characterization of Novel EST-based single-copy genic microsatellite DNA markers in white spruce and black spruce. Mol Biol Rep 48(3):2963–2971. 10.1007/s11033-021-06231-1
51. Rajora OP, Rahman MH, Dayanandan S, Mosseler A (2001) Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. Mol Gen Genet 264(6):871–882
52. Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13(4):243. 10.1186/gb-2012-13-4-243
53. Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109(6):365–371. 10.1007/s004120000089
54. Seidl R, Thom D, Kautz M, Martin-Benito D, Peltoniemi M, Vacchiano G, Wild J, Ascoli D, Petr M, Honkaniemi J, Lexer MJ, Trotsiuk V, Mairota P, Svoboda M, Fabrika M, Nagel TA, Reyer CPO (2017) Forest disturbances under climate change. Nat Clim Change 7(6):395–402. 10.1038/nclimate3303
55. Shimono A, Wang XR, Torimaru T, Lindgren D, Karlsson B (2011) Spatial variation in local pollen flow and mating success in a *Picea abies* clone archive and their implications for a novel breeding without breeding strategy. Tree Genet Genomes 7(3):499–509. 10.1007/s11295-010-0351-5
56. Simko I (2009) Development of EST-SSR Markers for the Study of Population Structure in Lettuce (Lactuca sativa L). J Hered 100(2):256–262. 10.1093/jhered/esn072
57. Sinigaglia L, Baptista L, Alves C, Feldmann F, Sacchetti C, Rupprecht C, Vijayan T, González E, Ávila S, Santos A, Curto M, Meimberg H (2024) Oceanic islands act as drivers for the genetic diversity of marine species: Cardita calyculata (Linnaeus, 1758) in the NE Atlantic as a case-study. BMC Ecol Evol 24. 10.1186/s12862-024-02322-2
58. Splechtna BE, Splechtna K (2016) Rothschild's Wilderness. Rachel Carson Center for Environment and Society. http://www.environmentandsociety.org/node/7420
59. Sundell DM, Chanaka; Netotea S, Delhomme N, Lin Y-C, Sjödin A, Van de Peer Y, Jansson S, Hvidsten, Torgeir R (2015) The Plant Genome Integrative Explorer Resource: PlantGenIE.org. New Phytologist. 10.1111/nph.13557
60. Thomson RC, Wang IJ, Johnson JR (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol Ecol 19(11):2184–2195. 10.1111/j.1365-294X.2010.04650.x
61. Tibihika PD, Aruho C, Namulawa V, Ddungu R, Atukunda G, Aanyu M, Nkambo M, Vijayan T, Kwikiriza G, Curto M, Meimberg H (2024) Unlocking Nile tilapia (Oreochromis niloticus Linn., 1758) selective

breeding programmes in Uganda through geographical genetic structure mapping. Aquaculture Fish Fisheries 4(4):e197. https://doi.org/10.1002/aff2.197

62. Tibihika PD, Curto M, Dornstauder-Schrammel E, Winter S, Alemayehu E, Waidbacher H, Meimberg H (2019) Application of microsatellite genotyping by sequencing (SSR-GBS) to measure genetic diversity of the East African Oreochromis niloticus. Conserv Genet 20(2):357–372. 10.1007/s10592-018-1136-x
63. Tiret M, Olsson L, Grahn T, Karlsson B, Milesi P, Lascoux M, Lundqvist SO, García-Gil MR (2023) Divergent selection predating the Last Glacial Maximum mainly acted on macro-phenotypes in Norway spruce. Evol Appl 16(1):163–172. 10.1111/eva.13519
64. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3–new capabilities and interfaces. Nucleic Acids Res 40(15):e115. 10.1093/nar/gks596
65. Vendramin GG, Anzidei M, Madaghiele A, Sperisen C, Bucci G (2000) Chloroplast microsatellite analysis reveals the presence of population subdivision in Norway spruce (*Picea abies* K). Genome 43(1):68–78. 10.1139/gen-43-1-68
66. Wegrzyn JL, Staton MA, Street NR, Main D, Grau E, Herndon N, Buehler S, Falk T, Zaman S, Ramnath R, Richter P, Sun L, Condon B, Almsaeed A, Chen M, Mannapperuma C, Jung S, Ficklin S (2019) Cyberinfrastructure to Improve Forest Health and Productivity: The Role of Tree Databases in Connecting Genomes, Phenomes, and the Environment. Front Plant Sci 10–2019. 10.3389/fpls.2019.00813
67. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30(5):614–620. 10.1093/bioinformatics/btt593

Figures

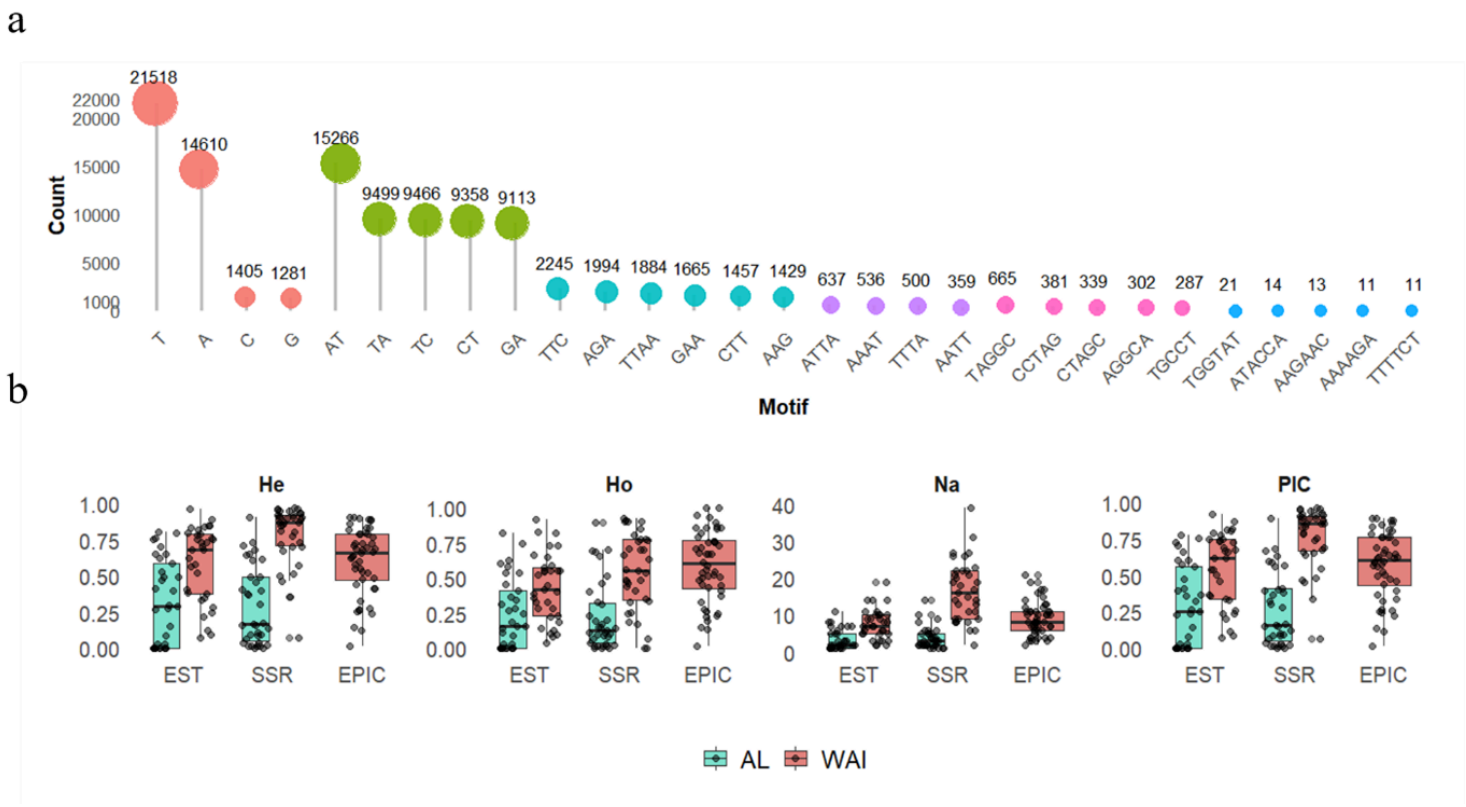

Figure 1

(a) Distribution of identified microsatellite motifs across the Norway spruce genome. Counts categorized by repeat unit size, ranging from mononucleotides to hexanucleotide repeats. (b) Genetic diversity and polymorphism measures for length genotyping (AL) and whole amplicon information (WAI) genotyping across EST and SSR markers compared with EPIC markers through four key metrics: Expected and observed heterozygosity (He and Ho), number of alleles (Na), and polymorphism information content (PIC)

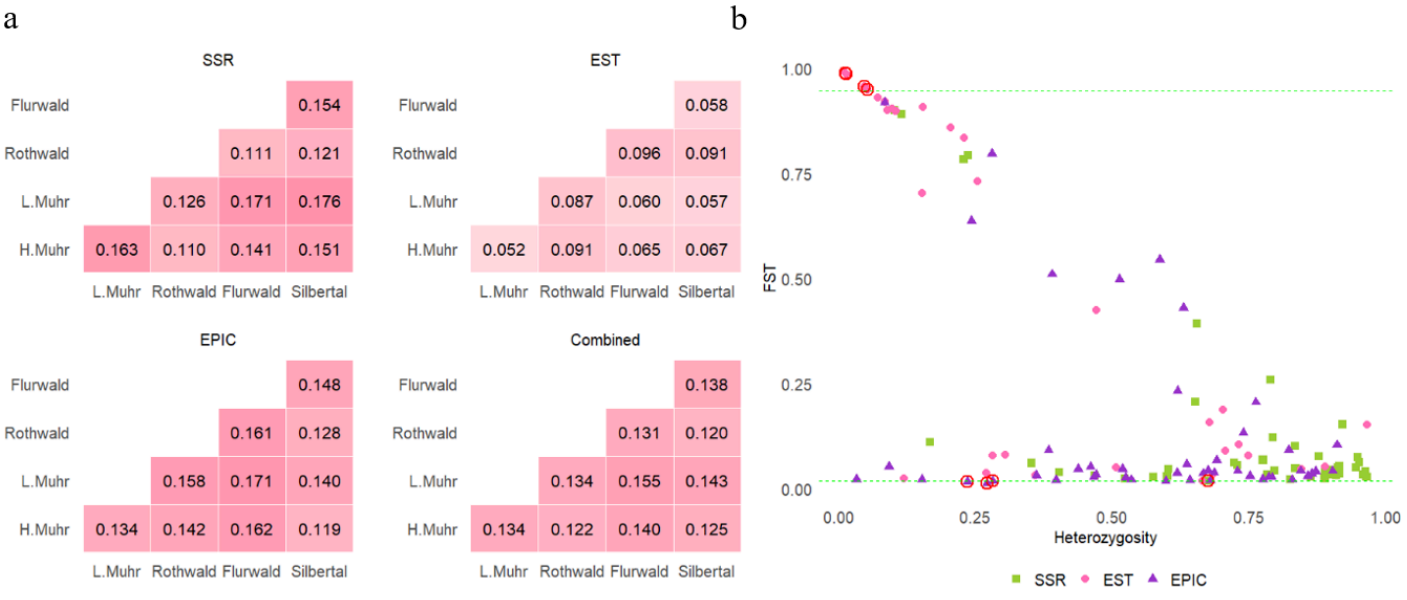

Figure 2

(a) Pairwise F_{ST} values for population differentiation based on SSR, EST, EPIC and all combined (b) F_{ST} - heterozygosity outlier graph for all markers. The green dashed lines represent the upper and lower F_{ST} thresholds, and the red circle highlights the outliers

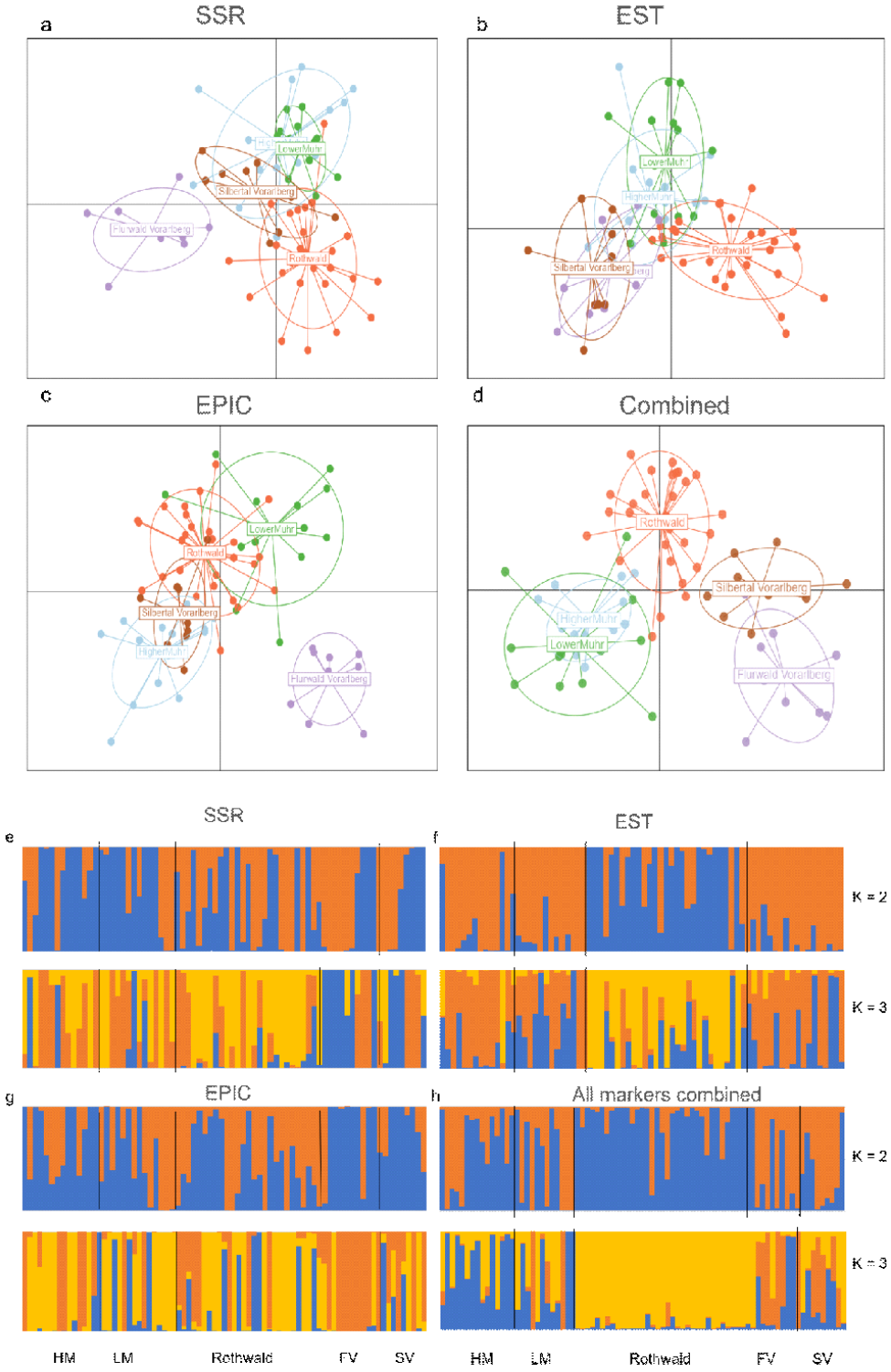

Figure 3

(a-d) Discriminant Analysis of Principal Components (DAPC) plots based on (a) SSR (b) EST (c) EPIC and (d) Combined. The axes correspond to the first two discriminant functions derived from the DAPC, which maximize among-group variation relative to within-group variation (e-h) STRUCTURE bar plots based on (e) SSR, (f) EST, (g) EPIC, and (h) All markers combined. Population labels: HM (Higher Muhr), LM (Lower Muhr), FV (Flurwald, Vorarlberg), SV (Silbertal, Vorarlberg)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarymaterialVijayanetal.docx
- PiabAllelematrixVijayanetal2025.xlsx