

MISS AUDE ROGIVUE (Orcid ID : 0000-0002-6864-0587)
MRS. RIMJHIM ROY CHOUDHURY (Orcid ID : 0000-0002-0499-4124)
DR. STÉPHANE JOOST (Orcid ID : 0000-0002-1184-7501)
PROF. CHRISTIAN PARISOD (Orcid ID : 0000-0001-8798-0897)

Article type : Resource Article

**Genome-wide variation in nucleotides and retrotransposons in alpine populations
of *Arabis alpina* (Brassicaceae)**

Aude Rogivue^{1,+}, Rimjhim R. Choudhury^{2,3,+}, Stefan Zoller⁴, Stéphane Joost⁵, François Felber^{2,6}, Michel Kasser⁷, Christian Parisod^{3,*}, Felix Gugerli^{1,*}

¹ WSL Swiss Federal Research Institute, Zürcherstrasse 111, CH–8903 Birmensdorf, Switzerland

² University of Neuchâtel, Rue Emile-Argand 11, CH–2000 Neuchâtel

³ Institute of Plant Sciences, University of Berne, Altenbergrain 21, CH–3013 Bern, Switzerland

⁴ Genetic Diversity Centre, ETH Zürich, CH–8092 Zürich, Switzerland

⁵ Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

⁶ Musée et Jardins botaniques cantonaux, Avenue de Cour 14bis, CH–1007 Lausanne, Switzerland

⁷ HEIG-VD, Rte de Cheseaux 1, CH–1400 Yverdon-les-Bains, Switzerland

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12991

This article is protected by copyright. All rights reserved.

⁺ AR and RRC equally contributed to this article as joint first authors

^{*} CP and FG are joint senior authors of this article

*Correspondence: felix.gugerli@wsl.ch, christian.parisod@ips.unibe.ch

Abstract

Advances in high-throughput sequencing have promoted the collection of reference genomes and genome-wide diversity. However, the assessment of genomic variation among populations has hitherto mainly been surveyed through single-nucleotide polymorphisms (SNPs) and largely ignored the often major fraction of genomes represented by transposable elements (TEs). Despite accumulating evidence supporting the evolutionary significance of TEs, comprehensive surveys remain scarce. Here, we sequenced the full genomes of 304 individuals of *Arabidopsis thaliana* sampled from four nearby natural populations to genotype SNPs as well as polymorphic long terminal repeat retrotransposons (polymorphic TEs; i.e. presence/absence of TE insertions at specific loci). We identified 291,396 SNPs and 20,548 polymorphic TEs, comparing their contributions to genomic diversity and divergence across populations. Few SNPs were shared among populations and overall showed high population-specific variation, whereas most polymorphic TEs segregated among populations. The genomic context of these two classes of variants further highlighted candidate adaptive loci having a putative impact on functional genes. In particular, 4.96% of the SNPs were identified as non-synonymous or affecting start/stop codons. In contrast, 43% of the polymorphic TEs were present next to *Arabidopsis* genes enriched in functional categories related to the regulation of reproduction and responses to biotic as well as abiotic stresses. This unprecedented dataset, mapping variation gained from SNPs and complementary

polymorphic TEs within and among populations, will serve as a rich resource for addressing microevolutionary processes shaping genome variation.

Keywords: *Arabis alpina*, Genome-wide diversity, Linkage disequilibrium, Single-nucleotide polymorphisms, Transposable elements.

Introduction

The drivers of genetic variation in natural populations across diverse environments is a fundamental issue in evolutionary biology. Such knowledge is important to understand the impact of climate change on adaptive responses of extant populations. Advances in high-throughput sequencing technologies have increasingly facilitated studies dissecting the genotypic basis of phenotypic variation. Such methodological progress and the availability of high-quality reference genomes has led to the increasing use of single-nucleotide polymorphisms (SNPs) in population genetic studies, efficiently applied to a wide range of questions and organisms including non-model species.

Parallel sequencing of DNA fragments covering the whole genome of multiple individuals has been used to develop large catalogues of genomic variation. Accordingly, whole-genome sequencing has generated large datasets comprising more than 100 individual plants, such as 948 inbred accessions of *Arabidopsis thaliana* from the entire species range (Hancock et al., 2011), 419 accessions of upland cotton from diverse locations (Ma et al., 2018), or 302 wild and cultivated accessions of *Glycine max* (Zhou et al., 2015). The application of whole-genome sequencing to surveys of genomic variation within populations of Chinese *Gossypium arboreum* (230 accessions; Du et al., 2018), the Rice Genome Project (3,010 accessions; Wang et al., 2018) or outcrossing *Capsella grandiflora* (188 individuals;

Josephs, Lee, Stinchcombe, & Wright, 2015) provided new insights regarding their evolution. However, few studies so far investigated variation among large numbers of individuals from natural populations of non-model species with whole-genome sequencing. Such studies at a local scale (i.e. within less than 10km distance) are expected to offer important resources to address interactions between genomic and environmental variation, shedding new light on the neutral and adaptive responses to changing environments at the population level. The distribution of genetic variation at hundreds of thousands of loci across the genome of individuals from various environments or populations indeed offers unprecedented insights to understand ecological and evolutionary processes.

The reliability and power of whole-genome SNP data to investigate natural populations is well established (Morin, Luikart, & Wayne, 2004) and the detection of SNPs has nowadays become routine (Garvin, Saitoh, & Gharrett, 2010; Seeb et al., 2011). Such data have been regularly used to infer the genetic basis of adaptive traits (Atwell et al., 2010; Exposito-Alonso et al., 2018) or, in association with environmental features, to detect gene variants underlying local adaptation (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015). SNP data can further accommodate sophisticated modelling approaches to infer the drivers of genomic variation through neutral and adaptive processes. In particular, the whole-genome sequencing of 38 individuals of *A. alpina* from populations with divergent levels of outcrossing/selfing allowed to evaluate the effect of mating system on the purging of deleterious alleles (Laenen et al., 2018).

Individual whole-genome sequencing can detect a wide range of genome-wide molecular variants to possibly complement SNPs to address the demographic and adaptive components of genome evolution. In particular, transposable elements (TEs) correspond to a

major fraction of many plant genomes (Bennetzen, 2005; Lisch, 2013), but only a small fraction of population genomics studies have provided detailed datasets with the presence/absence of TE insertions at specific loci (polymorphic TEs) in natural plant populations. Although first described as “controlling elements” to illustrate the capacity of TEs to regulate phenotypes in maize (McClintock, 1956), recent studies increasingly support TEs as central to the evolution of gene regulation in plants and other genome rearrangement such as gene duplication (Doolittle, 2013).

TEs can be classified into copy-and-paste-based retrotransposons (Class I), which proliferate via RNA intermediates, and cut-and-paste-based DNA transposons (Class II), which proliferate via excising from one location and inserting into another in the genome (Wicker et al., 2007). In plants, the most abundant TEs are the long terminal repeat (LTR) retrotransposons, representing from 5.6% (8.8Mb) of the compact *Arabidopsis thaliana* genome (Pereira, 2004) to 75% (1.5Gb) of the complex maize genome (Baucom et al., 2009) or 48% (9.4Gb) of the *Picea abies* genome (Nystedt et al., 2013). Such TEs thus represent a major source of raw evolutionary material not only when inserting into new genomic locations, but also as interspersed targets of ectopic recombination and other mechanisms of chromosome rearrangement (Gray, 2000). Several TEs further contain regulatory sequences that can affect the structure and expression of nearby genes and thereby have an impact on phenotypes (Chuong, Elde, & Feschotte, 2016; Elbarbary, Lucas, & Maquat, 2016). Similarly, it has been shown that epigenetic marks repressing inserted TEs can regulate the expression of nearby genes (Fedoroff, 2012; Hollister et al., 2011; Lisch & Bennetzen, 2011). Finally, specific TE families (or related inserted copies having evolved as quasi-species within a given genome; Casacuberta, Vernhettes, Audeon, & Grandbastien, 1997) appear activated in association to hybridization (e.g. Senerchia, Felber, & Parisod, 2015) and various

abiotic stresses (e.g. Cavrak et al., 2014; Grandbastien et al., 2005; Kalendar, Tanskanen, Immonen, Nevo, & Schulman, 2000), supporting genome changes and possibly adaptation in plants facing challenging situations (Rey, Danchin, Mirouze, Loot, & Blanchet, 2016). Given their possible consequences on the evolution of host genomes, TEs would ideally be included in population surveys aiming at shedding light on the underpinnings of adaptation in Eukaryotes (Bonchev & Parisod, 2013).

Variation associated with TEs is often ignored because TE-rich genomic regions are difficult to assemble and annotate in genome drafts (Hoban et al., 2016; Treangen & Salzberg, 2012). The genotyping of polymorphic TEs (i.e. presence/absence of TE insertions at specific loci) benefits from the availability of high-quality TE annotation and, thus, allows for the assessment of their significance (Choudhury & Parisod, 2017). In contrast to the detection of SNPs, the development of appropriate computational tools to detect polymorphic TEs from current sequence data is still in its infancy. However, challenges can be overcome, and appropriate software packages properly calling reference TE polymorphisms, accurately estimating breakpoint intervals and treating TE polymorphisms as codominant loci should be favored (Bergman, 2012; Rishishwar, Mariño-Ramírez, & Jordan, 2016). Accurate genotyping of polymorphic TEs among large numbers of samples is further improved by considering evidence of TE variation from related samples such as implemented in specific approaches designed for population studies (e.g. TEPID; Stuart et al. 2016). Keeping such caveats in mind, studies that included TEs highlighted their important role for adaptive evolution in various organisms (Bennetzen & Wang, 2014; Casacuberta & González, 2013).

Ecologists and evolutionary biologists have recently turned their attention to species related to the model plant *Arabidopsis thaliana*. In particular, *Arabis alpina* has become an established model plant for highlighting candidate loci with molecular function of possible adaptive significance (Wang et al. 2009; Tor ng et al., 2015; de Villemereuil, Mouterde, Gaggiotti, & Till-Bottraud, 2018). This short-lived perennial plant shows variation in the mating system from nearly strict outcrossing to almost full selfing (Tedder et al., 2011; Laenen et al., 2018) and further grows in diverse ecological niches, overcoming some shortcomings of *A. thaliana* as a model species while benefiting from the broad and diverse knowledge on the molecular and developmental biology of the latter species. Accordingly, a high-quality assembly of the 370Mb genome of *A. alpina* has recently been released (Jiao et al., 2017; Willing et al., 2015), making it possible to address challenging evolutionary questions difficult to pursue in other species (Woetzel et al., *submitted*).

Here, we investigated the distribution of SNPs and polymorphic insertions of LTR retrotransposons (hereafter referred to as polymorphic TEs) that may be important for deciphering patterns of demographic processes and of local adaptation in *A. alpina*. Benefitting from the latest high-quality reference genome of a Spanish accession (Pajares; Jiao et al., 2017), we assembled a dataset with whole-genome sequencing data from 304 individuals collected in four nearby alpine regions to provide in-depth annotation of their SNPs and TE content. Our aim was to establish a genomic resource and describe nucleotide and LTR retrotransposon variation and their putative impact on genes in natural populations. SNPs as well as polymorphic TEs close to or within genes suggest that both types of genomic variation may have a key impact on adaptive variation in *A. alpina*.

Material and methods

Study species and sampling

The alpine rock cress *Arabis alpina* L. is a perennial arctic-alpine herb of the Brassicaceae family. It grows mainly on calcareous bedrock along a wide elevational range and occupies various habitat types (Buehler et al., 2013). We chose four study regions in the western Swiss Alps, which are descending from the same original population that recolonized the Alps from the East after the last glaciation (Rogivue, Graf, Parisod, Holderegger, & Gugerli, 2018). Rosette leaves of 306 individuals were sampled between 2016m and 2457m a.s.l.: 70 in La Para (1), 70 in Pierredar (2), 70 in Les Essets (3), and 96 in Les Martinets (4; Figure S1 in Supplementary Information) and stored in silica gel until DNA extraction. The four sampled regions cover an average area of 0.42 km², within which plants were sampled, with few exceptions, at least 1m apart. Despite their common post-glacial ancestry (Rogivue et al., 2018), we treated plants from each region as different populations, as they are situated in topographically separated Alpine valleys presumably with limited gene flow among them (Buehler, Graf, Holderegger, & Gugerli, 2012). Alpine populations of *A. alpina* sexually reproduce mainly by selfing (Ansell, Grundmann, Russell, Schneider, & Vogel, 2008; Buehler et al., 2012) or, rarely, asexually via stoloniferous above-ground growth, but mixed-mating and predominant outcrossing is also found within the European range (Tedder et al.; 2011; Laenen et al., 2018).

DNA extraction, sequencing, and mapping

DNA was extracted from silica gel-dried leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), and quality was checked on agarose gels as well as with the UV-Vis Spectrometer (NanoDrop 1000, Thermo Scientific Wilmington, DE, USA), while quantity was measured with the Quantus™ Fluorometer (Promega Corporation, Madison,

WI, USA). Library preparation (NEBNext® Ultra™, New England Biolabs, Ipswich, MA, USA) and sequencing with Illumina HiSeq2500 (ATLAS Biolabs GmbH, Berlin, Germany; 125-bp paired-end reads) were performed by the Functional Genomics Centre Zürich (Zürich, Switzerland). Firstly, whole-genome sequencing of 96 individuals from Les Martinets was performed on four lanes in order to evaluate output and sequence quality. Subsequently, genomes of individuals of the three remaining populations were fully sequenced on eight lanes and samples were randomly placed among the lanes to avoid artefacts.

After quality control using FastQC

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), adapters were trimmed with Trimmomatic v.0.35 (Bolger, Lohse, & Usadel, 2014), polymerase chain reaction (PCR) duplicates were removed, and leading and trailing bases were removed below the quality of 5. The sequenced reads were filtered with a sliding window of 4 and an average Phred score of 15 within the window, and reads shorter 50bp were removed. The remaining sequences were used for mapping against the reference genome V5.1 (Jiao et al., 2017) with BWA v.0.7.12 (Li & Durbin, 2010); more details are found in Appendix S1 in Supplementary Information.

Annotation of TEs in the *Arabis* genome assembly

LTR retrotransposons were *de novo* re-annotated in the reference sequence of *A. alpina* V5.1 based on Choudhury, Neuhaus, & Parisod (2017), and hereafter referred to as reference TEs. More details are found in Appendix S1 in Supplementary Information. This new annotation of LTR retrotransposons in the improved V5.1 assembly of the genome of *A. alpina* identified a higher number of copies compared to the same approach on a prior genome version (Choudhury et al., 2017). In particular, eight additional monophyletic lineages of

LTR-RT sequences shared among Brassicaceae species (i.e. Tribes) have been here identified (Table S1 in Supplementary Information). Solo-LTRs were further annotated.

SNP calling

The SNP calling was done with FreeBayes v.1.0.2 (Garrison & Marth, 2012) with the following options: ploidy 2, a minimum alternate fraction of 0.2, base quality at the site > 10 and coverage > 4. We strongly filtered SNPs according to the following settings: We required that a SNP (i) was biallelic (FreeBayes v.1.0.2; Garrison & Marth, 2012) and (ii) had a quality/depth > 0.25 (vcffilter from vcflib v.1.0.1; Garrison 2012, <https://github.com/vcflib/vcflib#vcflib>, accessed May 2017). The following filtering steps were performed with VCFtools v.0.1.14 (Danecek et al., 2011). We required: (iii) a minimum depth of 8 and a maximum of 100, and a minimum count of alternative alleles of 2 and a maximum count of 100 (to avoid incorrect SNPs due to the mapping of repeat variants). Finally, (iv) we filtered SNPs with a minor allele frequency (MAF) < 0.025 and (v) SNPs with more than 10% missing data. Because Ribeiro et al. (2015) showed an enrichment of false positive SNPs in TE sequences due to the inherent difficulties for accurate mapping in repetitive sequences, we further classified the SNPs into two categories, namely those outside and those within annotated TE sequences. Accordingly, we refer to these datasets as “non-TE SNPs” as opposed to “SNPs” for the full dataset that also contains SNPs within TE sequences (Table 1, Table S2 in Supplementary Information). Our analyses are based on both SNP sets for comparative purposes, because standard approaches used in population genomic studies rarely account for this difference. For each SNP we calculated the MAF based on the frequency of the SNP using option “--freq” of VCFtools (Danecek et al., 2011).

Polymorphic TE calling

Polymorphic TEs among the sequenced genomes of individuals of *A. alpina* were identified and genotyped using TEPID v.0.6 (Stuart et al., 2016). This approach described in Figure S2 employs split and discordant read mapping information, read mapping quality, sequencing breakpoints and local variation in sequencing coverage to infer absence of reference TEs (i.e. present in the reference assembly but absent in the genotyped individual; hereafter referred to as ‘TE-absence’) as well as the presence of non-reference TE insertions (i.e. absent in the reference assembly but present in the genotyped individual; hereafter referred to as ‘TE-presence’). Quality filtered FASTQ files of sampled individuals were mapped to the reference genome assembly using the ‘tepid-map’ algorithm, which identifies the reads split between two genomic mapping coordinates. The ‘tepid-discover’ algorithm then identified TE-absence and TE-presence with respect to the reference genome. Only the best mapping region was considered and a series of stringent filters was applied to accurately identify variants with respect to the reference genome. In particular, TE-absence was only called when at least 80% of the TE sequence was spanned by split or discordant reads, with the annotated TE region covered by less than 10% of the sequencing depth of 2kb flanking sequences. The identification of TE-presence relies exclusively on split reads spanning less than 5kb, with mapping quality of at least 5 and at least two reads coverage to remove possible false candidates. Calls required at least 80% overlap between TE and split/discordant read mapping coordinates as well as few independent discordant read pairs in opposite orientation at the insertion sites.

Calls of identified variants were further refined using the ‘tepid-refine’ algorithm, which reduces false negatives by examining the nearby genome region for corresponding non-reference alleles being identified in other individuals of the four populations.

Accordingly, this refinement step looks for evidence supporting the non-reference allele in the focal individual using lower thresholds as compared to the 'tepid-discover' step and thus takes polymorphism in the population into account. In case of insufficient sequencing coverage in a sample for a given locus, missing information is called as 'NA' to account for uncertainty. Such a procedure adequately accommodates individuals with a low sequencing depth across the genome with about 80% true positives at coverage above 10x (Stuart et al., 2016).

TE-absence and TE-presence were integrated across all individuals using the merge_insertion.py module from TEPID to generate the final call sets of TE presence by merging calls with the same TE family when their coordinates are within 100bp. The merge_deletion.py was used to simply merge TE absence per locus within the four populations. The genotype.py module was used to generate a coherent output file with all polymorphic TEs recorded as either present or absent in each individual across the four populations. It also identified individuals for which 'NA' was called at that locus by the refinement step.

As SNPs are possibly miscalled in TE-rich regions, inference of the zygosity at each identified TE locus was not based on SNPs within TEs but on the coverage at breakpoints of polymorphic TEs. The number of reads supporting presence (supportive reads) as well as those not supporting presence (non-supportive reads) at the borders of a given TE insertion were retrieved in all individuals. Accordingly, the number of properly mapped, discordant and soft clipped reads were counted in the region surrounding the predicted border of each locus. A custom R script then determined the probability that a TE locus was heterozygous or homozygous by using the ratio between the number of supportive and non-supportive reads.

For an allele to be considered heterozygous, the expected 2:1 ratio of supportive paired-end reads (which cover both borders of TE) to the non-supportive paired-end reads (Jiang, Chen, Huang, Liu, & Verdier, 2015) was tested through binomial tests ($p < 0.05$). Zygosity was estimated as unknown and coded as 'NA' if in total there were less than five uniquely mapped reads surrounding the borders. Only polymorphic TEs with a MAF > 0.025 and less than 10% NAs were finally retained to focus on the most confident, heritable polymorphic TEs for downstream analyses. Accordingly, possible somatic transposition events (e.g. Treiber & Waddell, 2017; Baillie et al., 2011) were not considered here.

Distribution of SNPs and polymorphic TEs within/among populations

For each polymorphism set (SNPs, non-TE SNPs, TEs), we estimated observed heterozygosity (H_o), expected heterozygosity (H_e), population-wise inbreeding coefficients (F_{IS}) and pairwise genetic differentiation (F_{ST}) among populations with HIERFSTAT v. 0.4.28 (Goudet, 2005). For mean F_{IS} and F_{ST} , 95% confidence intervals were obtained by bootstrapping over loci ($n = 100$).

To compare population genomic results with those obtained from alternative, traditional molecular markers, we genotyped 110 samples (20 each from Essets, Martinets and Pierredar, and 50 from Para) with 19 nuclear microsatellite markers (Buehler, Graf, Holderegger, & Gugerli, 2011) following Rogivue et al. (2018). We performed a one-way analysis of variance (ANOVA) with Tukey's multiple comparison test to check if estimates of H_o and F_{IS} based on the different marker types significantly differed. Accordingly, we applied a Mantel test implemented in ecodist v.2.0.1 (Goslee & Urban, 2007) with 1000 permutations to test for correlations between respective values of pairwise F_{ST} among populations.

Genomic variation and linkage disequilibrium along chromosomes

The genomic context of both SNPs and polymorphic TEs was described using the annotation of genes from the *A. alpina* reference genome V5.1 and our *de novo* TE annotation. The catalogue of SNPs and polymorphic TEs was screened for overlaps with introns, exons, 2kb upstream and downstream regions and intergenic regions. SnpEff (Cingolani et al., 2012) was used to annotate the SNPs, whereas polymorphic TEs were annotated with Bedtools intersect v.2.25 (Quinlan & Hall, 2010). The observed counts of non-TE SNPs and polymorphic TEs for each of the functional features were compared to their expectation computed as the total number of polymorphisms multiplied by the fraction of the genome occupied by each feature. The significance of the differences in the observed versus expected counts for each feature were calculated using Fisher's exact tests. Note that the calculation for non-TE SNPs was based on the number of annotated SNPs (452,257), as SnpEff can annotate a SNP several times given the position of the SNPs according to neighbouring genes.

For SNP and non-TE SNP sets, we estimated the extent of linkage disequilibrium (LD) with the function “pairwise LD measures for multiple SNPs (genome wide)” of Plink (Purcell et al., 2007), which calculates the squared correlation coefficient between two loci (r^2); for each SNP we considered adjacent SNPs that were less than 1,000bp apart, separately for each chromosome. Knowing that on average LD decays within 10kb in *A. thaliana* (Kim et al., 2007), we estimated LD in different window sizes (between 250kb and 1500kb; Table S3), which enabled us to have a considerable amount of data for a good estimation, while giving a reasonable number of pairwise comparisons to compute for each chromosome. To show the decline of LD with physical distance, we calculated the decay of r^2 according to Remington et al. (2001), with the function implemented in R by Marroni et al. (2011), yielding half-decay distance as the distance at which LD is half of its maximum value. This

value is commonly used to compare an estimation of LD among chromosomes and populations (Vos et al., 2017) like in *A. thaliana* (Kim et al., 2007).

We estimated gene density (Jiao et al., 2017), TE density, SNP density and polymorphic TE density in 250kb windows along chromosomes. To obtain the measure of LD along the chromosomes, we averaged r^2 for a window of 250kb.

Functional analysis

Using high-quality annotations of genes from the *A. alpina* reference, we tested for the enrichment of GO biological process terms in genes with at least one non-synonymous SNP or genes with at least one polymorphic TE located within it or in the 2kb surrounding region using topGO v.2.28.9 (Alexa, Rahnenführer, & Lengauer, 2006). Significance of terms was determined using Fisher's exact tests with the "weight01" algorithm in topGO as it was shown to improve the explanatory power of GO group scoring by taking the hierarchical relationships between terms into account and eliminating local dependencies between GO terms (Alexa et al., 2006). This approach computes p values of a GO term conditioned on the neighboring terms to reduce the impact of non-independent comparisons and were thus considered as corrected for multiple testing. We further reduced statistical artefacts by relying on a stringent p value cut-off of 0.01 and by only considering terms supported by five or more annotated genes. REViGO (<http://revigo.irb.hr>; Supek, Bošnjak, Škunca, & Šmuc, 2011), which removes redundant and similar terms from long Gene Ontology lists by semantic clustering, was applied to visualize the enrichment results.

We tested whether significantly enriched GO term categories related to non-synonymous SNPs include genes with larger coding regions than the overall gene set. Exon lengths per gene were extracted using the annotations of genes from the *A. alpina* reference (Jiao et al., 2017). We assessed the difference in length of coding region in each enriched GO term with the coding region of all genes using the Bonferroni-corrected one-sided Wilcoxon rank-sum test. For polymorphic TEs, we tested whether the overall length of the gene and the 2kb surrounding regions were longer for enriched GO categories than the overall gene set. The difference in length was assessed using the Bonferroni-corrected one-sided Wilcoxon rank-sum test.

Results

Whole-genome sequencing

The sequencing of 306 genomes yielded 10,081,497,170 filtered reads (average 32,946,069 per sample; Table S4 in Supplementary Information) for mapping. After exclusion of chloroplast and mitochondrial sequences, 6,873,004,145 reads were properly mapped to the reference genome V5.1 (average 22,460,798 per sample), corresponding to an average coverage of 11.77x. After excluding two samples due to low coverage (Pa9 = 1.2x and Pi95 = 0.02x; Table S4 in Supplementary Information), we remained with a dataset of 304 individuals with mean coverage between 5x and 43x (median 8.79x: range of medians per individual 3.30x–33.81x).

SNP and TE variation

From an initial number of 78,859,801 called SNPs, our stringent filtering steps left us with 439,670 SNPs, of which 291,396 were non-TE SNPs (Table 1, Figure S3). The two data sets comprised 10.3% (SNPs) and 6% (non-TE SNPs) segregating sites that were shared among

the four populations (Figure 1a, Figure S4 in Supplementary Information). We observed a below-average SNP density at the centromere region and at the end of the chromosomes for both SNP sets (Figure 2, Figure S5 in Supplementary Information). By filtering SNPs within TEs, we lost SNPs mainly around the centromere region, and peaks of high and low density were less extreme, as exemplified by one peak in chromosome 6 highlighted in Figure S5 in Supplementary Information. On average, we observed one SNP every 742bp for the SNP set and every 1151bp for the non-TE SNP set.

Following *de novo* annotation, the reference genome V5.1 of *A. alpina* consisted of 36.5% LTR retrotransposons with 244,486 copies of length greater than 80bp. We further identified 38,768 loci with presence of non-reference TE insertions and 31,143 loci with absence of reference TE insertions among the 304 sequenced genomes. The number of TE-presence and TE-absence variants identified here were positively correlated with sequencing depth (Figure S6), indicating that both types of calls benefit from high coverage. However, individuals with low coverage gained more TE-presence and TE-absence calls during the TEPID refinement step (Figure S6), indicating effective reduction of false negatives by incorporating TE variant information from the whole population. After a filtering step to remove polymorphic TEs with $MAF < 0.025$ and with $> 10\%$ missing data, the studied individuals of *A. alpina* presented 20,548 polymorphic TEs (Figure S3). The majority of the polymorphic TEs (89.24%) were shared among the four populations (Figure 1b). In contrast to the genomic distribution of reference TEs, which showed higher density at the pericentromeric region and lower density close to genes, the density of polymorphic TEs suggests that they are evenly distributed along chromosomes (Figure 2).

Population structure

Observed heterozygosity H_o of both SNP sets and the nuclear microsatellites had a value of 0.09 for the 304 individuals, but polymorphic TEs showed H_o of 0.03 (Table 2). Population-wise values ranged between 0.04 and 0.18 for the SNPs, non-TE SNPs and nuclear microsatellites, but were as low as 0.03–0.04 for polymorphic TEs. Accordingly, F_{IS} varied between 0.18–0.28 for SNPs, 0.22–0.33 for non-TE SNPs, 0.83–0.86 for polymorphic TEs, and 0.49–0.69 for nuclear microsatellites (Table 2). ANOVA revealed that population-specific estimates were significantly different among the four marker sets for F_{IS} ($p < 0.0001$; except SNPs vs non-TE SNPs, $p < 0.68$), but not for H_o ($p = 0.22$; Table S5 in Supplementary Information).

The overall value of F_{ST} was 0.13 for both SNP datasets, 0.07 for polymorphic TEs, and 0.16 for nuclear microsatellites. The lowest pairwise F_{ST} value for SNPs, non-TE SNPs and nuclear microsatellites was between Essets and Para for both SNP sets (0.09) and for nuclear microsatellites (0.14), whereas Pierredar and Essets showed the lowest value for the polymorphic TEs (0.03, Table 3). The largest values occurred between Pierredar and Martinets: 0.18 for both SNP sets and 0.09 for polymorphic TEs. Nuclear microsatellites showed the largest values between Para and Martinets and between Pierredar and Martinets (0.25). There were no significant correlations among the matrices of pairwise F_{ST} except between SNPs and non-TE SNPs ($r_{MT} = 0.99$, $p = 0.035$, Table S5 in Supplementary Information).

Genomic context of non-TE SNPs and TEs

The annotation showed that 32,972 non-TE SNPs (7.29%) occurred in introns, which is significantly higher than expected based on the genomic fraction occupied by them

(expected: 28,881, $p < 0.001$; Table 4), and 48,046 (10.62%) occurred in exons (expected: 47,460, $p < 0.05$). Thirty-five percent of the non-TE SNPs were located within 2kb upstream (78,517, 17.36%) or downstream (81,771, 18.08%) of gene regions, which appeared significantly lower than expected (expected: 91,742, $p < 0.001$ and 91724, $p < 0.001$, respectively), whereas the remaining SNPs occurred in intergenic regions (210,951, 46.64%). Of the putatively high-impact non-TE SNPs (likely to cause a loss of function in a protein), 22,603 (4.96%) were identified as non-synonymous variants and 532 (0.11%) were involved in start or stop codon mutations (gain or loss of function).

Polymorphic TEs showed a significant depletion inside both exons and introns with 1113 loci (5.42%) having interrupted exons (expected: 2156, $p < 0.001$, Table 4) and 957 loci (4.66%) were located within introns (expected: 1312, $p < 0.001$). However, they were significantly enriched within 2 kb of gene regions with 5241 loci (25.51%) upstream (expected: 4168, $p < 0.001$) and 4664 loci (22.70%) downstream of genes (expected: 4167, $p < 0.001$). A total of 2203 loci showed evidence consistent with transposition into regions encompassing more than a single genic feature (i.e. either exons and introns, and/or 2kb surrounding regions), suggesting that corresponding TEs did not only interrupt, but were further involved in rearrangements of genic regions. Out of the 34,218 protein coding genes annotated in the *A. alpina* V5.1 reference genome, 21.85% were interrupted by or located within 2kb of 8905 polymorphic TEs. In other words, 43.34% of the polymorphic TEs identified among individuals were associated with annotated genes ($p < 0.05$, observed: 8905, expected: 8663).

Whole-genome LD decay was estimated as the half-decay distance (Table S3 in Supplementary Information), showing that LD decayed to an $r^2 < 0.1$ within 40.09kb for the SNPs and within 30.98kb for the non-TE SNPs. In general, the difference between the two SNP sets was minor, except for chromosome 8, for which the SNP dataset yielded a value of 101.99kb, substantially higher than what we observed for all other chromosomes (19.68-46.83kb). The r^2 values of the non-TE SNPs strongly varied along the chromosomes, especially in the peri-centromeric regions and at the end of the chromosomes (Figure 2).

Functional annotation and enrichment analysis

Gene Ontology (GO) term analysis of the non-TE SNPs revealed that 6,943 of the genes presenting non-synonymous variation (i.e. non-synonymous SNPs or SNPs involved in start or stop codons; 23,135 SNPs) were significantly overrepresented in 25 GO terms ($p < 0.01$; Figure 3 and Table S6 in Supplementary Information). Among these genes, 941 were responsive to stress (plant-type hypersensitive response, defence response, defence response signalling pathway). Within a majority of enriched GO categories (i.e. 19/25), genes with non-synonymous SNPs also had larger coding regions on average (Bonferroni-corrected Wilcoxon rank sum test, Table S6). For the TE dataset, only 4511 of the genes next to polymorphic TEs were annotated by Willing et al. (2015) and assigned to GO terms. Among those, 850 genes were classified as response to stress, with 73 genes classified as response to heat (Table S7 in Supplementary Information) and 124 genes as response to cold (Table S7 in Supplementary Information). We identified 23 GO terms with significant enrichment ($p < 0.01$; Figure 3 and Table S7 in Supplementary Information), including 326 genes with nearby polymorphic TEs involved in response to wounding, 597 in response to oxidative stress, or 193 in relation to aging. Wilcoxon rank sum tests showed that only few of the enriched GO

categories related to polymorphic TEs (i.e. 5/23) presented significantly longer insertion targets than the overall set of loci (Table S7).

Discussion

Large-scale genotyping of SNPs and TEs

Assessing genomic variation is a key step when studying natural populations, as patterns of genome-wide variation are pertinent to identify signatures imprinted by demographic or adaptive processes. Accordingly, genotyping approaches based on whole-genome sequencing have been developed to generate hundreds of thousands of SNPs (e.g. Alonso-Blanco et al., 2016; Branca et al., 2011; Martin et al., 2017; Qiu et al., 2015), but SNPs are representative of only one type of molecular variation. In particular, the last decade has offered a deeper understanding of the diversity, abundance and global significance of TEs on the evolution of functional host genomes (Biemont & Vieira, 2006; Kidwell & Lisch, 2001), and surveys of their distribution in natural populations are needed (Bonchev & Parisod, 2013). Although the vast majority of TE insertions within genomes is arguably neutral, the few studies having assessed TE dynamics within species suggested an adaptive impact (Casacuberta & González, 2013) and divergent TE arrangements matching the eco-geographical distribution of gene pools (González, Karasov, Messer, & Petrov, 2010; Kalendar, Tanskanen, Immonen, Nevo, & Schulman, 2000). The dataset produced here through whole-genome sequencing of 304 individuals from four closely related alpine *Arabis alpina* populations served to compare the widely used SNPs with genomic variation represented by polymorphic TEs. The comparison of genome-wide diversity and differentiation in natural populations and insights into the putative functional impact of candidate adaptive loci ideally complement resources from the single-individual Spanish Pajares reference genome (Jiao et al., 2017).

A key step towards reliable genotyping of polymorphic TEs using TEPID was the determination of loci suffering from low coverage based on the ‘tepid-refine’ step that considers genotypes from the population. In the case of *A. thaliana*, Stuart et al. (2016) reported limited gain in TE-absence calls, but on average 4% more TE-presence calls for each accession following this refinement step. Here, both TE-presence and TE-absence gained considerably more calls following refinement. This may be due to our medium coverage per individual that offered multiple opportunities for refinement but could also be related to the longer reads used in this study that likely promoted higher recovery of informative split/discordant reads to confidently call polymorphic TEs in at least 90% genotyped individuals.

Distribution of SNPs and polymorphic TEs within/among populations

As expected in such selfing populations, heterozygosity (H_o) was generally low. Congruently, F_{IS} estimates were significantly higher than zero and showed limited variation among populations. The estimations of F_{IS} based on TEs and nuclear microsatellites are coherent with values reported from multiple populations of this region based on nuclear microsatellite data (average 0.68 (SD = 0.18); Rogivue et al., 2018) and with the predominantly selfing regime of the species in the western Alps (Laenen et al., 2018). On the contrary, F_{IS} values for both the full vs non-TE SNP sets were substantially lower. We attribute this difference to the respective characteristics of these contrasting types of loci, as also demonstrated in other studies (e.g. Fischer et al. 2017). These authors pointed out that SNPs were better suited than microsatellites for genetic diversity estimations, more reliably reflecting whole-genome patterns. Estimates of expected heterozygosity, H_e , are sensitive to the number of sampled alleles and may thus account for the higher F_{IS} values provided by multi-allelic nuclear microsatellites as compared to bi-allelic SNPs. Notably, also bi-allelic TE polymorphisms

yielded values of F_{IS} above those from microsatellites. This difference may result from our conservative inference of zygosity based on coverage.

The estimates of genetic differentiation among populations, F_{ST} , were consistently low among different types of loci. Both SNP sets presented a slightly higher degree of genetic structure than the polymorphic TEs, reflecting the pattern of shared variants among populations. The consistent genetic structure (F_{IS} and F_{ST}) inferred from both SNP sets appears remarkable provided that they differ by 33% of total SNPs comprised in TE sequences. The present dataset suggests that current mapping software coupled with high-quality genome references may cope with a possible bias in the mapping in repetitive sequences (Ribeiro et al., 2015), even though marginal differences between the outcomes using all vs only non-TE SNPs remained in our study.

As compared to SNPs that showed large numbers of private alleles within closely located populations, the total number of polymorphic TEs varied little among the four populations with a limited amount of private insertions. Such levels of shared polymorphic TEs contrasts with surveys at global scales in *A. thaliana* (Quadrana et al., 2016), *Brachypodium distachyon* (Stritt, Gordon, Wicker, Vogel, & Roulin, 2018) or *Drosophila melanogaster* (Kofler, Nolte, & Schlötterer, 2015) that rather reported considerable contributions of TEs to genome variation among populations. Although our conservative exclusion of low-frequency polymorphic TEs may have increased the number of insertions shared among our study populations, it seems rather coherent with a reduced transposition rate (i.e. TE quiescence) in selfing species (Ågren, 2014) or strong purifying selection against new insertions. Accordingly, polymorphic TEs identified here certainly represent standing genetic variation segregating among regions. Such standing genetic variation, characterizing

divergence from the Spanish Pajares reference genome, offers ample opportunities to gain insights into the impact of selection vs drift in driving specific loci to fixation (Barrett & Schluter, 2008).

Genome-wide variation

Both SNP sets presented densities matching gene density along chromosomes of *A. alpina* that show typically large peri-centromeric regions with a considerable amount of TE copies and relatively few genes (Willing et al., 2015). Non-TE SNPs were significantly more present in genes than expected and affected introns significantly more than exons. Although such a pattern is consistent with the removal of SNPs in exons by purifying selection, a quarter of the non-TE SNPs reported within genes were detected as putative high-impact SNPs (non-synonymous SNPs or SNPs involved in start and stop codon gain/loss) that likely affected genes involved in different biological processes such as stress responses. Such genomic variation may thus support traits related to individual fitness, which remains to be confirmed.

Other than non-TE SNPs, the even distribution of polymorphic TEs along chromosomes of *A. alpina* contrasted with the overall density of genes and TEs. Lower efficiency of TE calls in TE-rich regions may inflate such a pattern that is otherwise coherent with random insertion sites (Brookfield, 2005). The common distribution of polymorphic TEs across gene-rich regions further indicates that several functional loci may be influenced by nearby TE insertions. Nearly half (43.3%) of the polymorphic TEs indeed appeared as interrupting or flanking as much as 21.9% of the genes annotated in *A. alpina*. This fairly even distribution of polymorphic TEs suggests a multifactorial balance between insertion and deletion along chromosomes. Baucom et al. (2009) showed that LTR retrotransposon accumulation is non-random in maize with low-copy-number families primarily inserted into

or near genes, whereas high-copy-number families were inserted into other high-copy-number TEs. Although the distribution of reference TEs showed high abundance near pericentromeres of *A. alpina*, consistent with preferential insertion in heterochromatin, the even distribution of polymorphic TEs detected here favors increased removal from gene-rich euchromatin. This pattern could be due to lower power in detecting polymorphisms in TE-dense regions, but strongly contrasts with expectations under differential insertion of TEs among genome fractions.

In this dataset, polymorphic TEs belonging to tribes having proliferated in a relatively distant past in *A. alpina* (Choudhury et al., 2017) appeared to outnumber those from TEs having recently proliferated. Such a pattern is consistent with predominant removal of ancient TE copies from the genome, whereas TEs having recently proliferated are under tight control in this selfing species and may not currently show massive transposition (Bennetzen & Park, 2018). As expected if TEs were excluded from functionally essential regions by purifying selection, the observed number of polymorphic TEs within genes was significantly lower than expected. However, unlike non-TE SNPs that appeared under-represented around genes, polymorphic TEs were significantly enriched in the 2kb surrounding genes. Such deficit in polymorphic TEs in, but enrichment near genes was also reported in rice, maize and sorghum (Wei et al., 2016) as well as in *B. distachyon* (Stritt et al., 2018), supporting the hypothesis that TEs influence flanking genes and may not always be functionally inert (Feschotte, Jiang, & Wessler, 2002; Fedoroff, 2012; Hollister & Gaut, 2009; Sigman & Slotkin, 2016). Furthermore, polymorphic TEs in alpine populations of *A. alpina* were associated with genes mostly involved in environmental stimuli and may hence affect fitness in natural populations. However, to what extent such candidate adaptive TEs may account for differential gene

expression, phenotypic variation and the eco-geographic distribution of species/individuals is beyond the scope of the present study.

Possible adaptive impact of the variants

Variation of both SNPs and polymorphic TEs is potentially adaptive as shown by the gene ontology of both types of polymorphisms highlighting functional categories of broad diversity. As sampled individuals of *A. alpina* are enduring harsh, highly heterogeneous alpine environments, we expected to find genes related to abiotic conditions such as responses to cold and heat (Wingler, Juvany, Cuthbert, & Munné-Bosch, 2015). Such genes were found next to polymorphic TEs (response to wounding, oxidative stress, hypoxia, cold and heat). Similarly, genes involved in important metabolic pathways related to low-temperature tolerance (e.g. carbohydrate biosynthesis and lipid metabolic process; Kaplan et al., 2004) were also associated with SNPs and polymorphic TEs. Genes with non-synonymous SNPs in enriched GO categories also appeared longer and may thus represent larger mutational targets. In contrast, only very few of the enriched GO categories with genes presenting polymorphic TEs showed such a pattern and thus appear consistent with functional, possibly selective, consequences of polymorphic TEs. To disentangle the possible involvement of such variants in adaptation, it is promising to investigate the frequency of candidate adaptive SNPs and TEs in populations (Luikart, England, Tallmon, Jordan, & Taberlet, 2003) and to test for their association with individual microhabitat conditions (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015).

Among the four studied populations, SNPs were mostly restricted to nearby individuals and showed significant variation within the genes, whereas polymorphic TEs were largely shared among individuals, and enriched within 2kb regions of coding loci.

Patterns of genetic diversity in SNPs and polymorphic TEs were expectedly similar under a drift–migration equilibrium, and to what extent mutation and selection drive the frequency of alleles at such different types of loci deserves further attention. The high relatedness and spatial proximity of the four populations investigated here, considered to be predominantly selfing (Rogivue et al., 2018), supports the assumption that both substitution and transposition rates are negligible, arguing for drift or selection as the main drivers of observed patterns for SNPs vs TEs. However, it remains elusive to what extent polymorphic TEs modify the local recombination rate (He & Dooner, 2009; Zamudio et al., 2015; Kent, Uzunović, & Wright, 2017) and may thus influence the fixation of SNPs and polymorphic TEs. The unprecedented resolution of SNP and TE diversity across natural landscapes presented may serve future studies assessing the evolutionary forces driving genome-wide variation within and among populations, at best complemented by functional studies on the fitness relevance of the genomic variation observed.

Acknowledgements

We thank Jessica Joaquim and Lolita Ammann for the sampling, René Graf and Noémie Chevret for the lab work, Christian Rellstab and Katalin Csilléry for their support for the analyses, Wen-Biao Jiao and George Coupland for early access to genomic data and Benjamin Dauphin for generating figures of the sampling sites. We also thank Christian Rellstab and Rolf Holderegger for critical discussions and comments to the manuscript, and four anonymous reviewers who helped further improving the article. Financial support was provided for AR, RRC and CP through the Swiss National Science Foundation (GeneScale, CR32I3_149741) to FG, FF, SJ and MK; RRC was further supported by a Swiss Government Excellence Scholarships for Foreign Students (fellowship 2014.0821) to RRC.

References

- Ågren, J. A. (2014). Evolutionary transitions in individuality: insights from transposable elements. *Trends in Ecology & Evolution*, *29*, 90–96.
- Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, *22*, 1600–1607.
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., . . . Zhou, X. F. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, *166*, 481–491.
- Ansell, S. W., Grundmann, M., Russell, S. J., Schneider, H., & Vogel, J. C. (2008). Genetic discontinuity, breeding-system change and population history of *Arabis alpina* in the Italian Peninsula and adjacent Alps. *Molecular Ecology*, *17*, 2245–2257.
- Atwell, S., Huang, Y. S., Vilhjalmsón, B. J., Willems, G., Horton, M., Li, Y., . . . Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, *465*, 627.
- Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapiro, F., . . . Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, *479*, 534–537.
- Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*, 38–44.
- Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., . . . Bennetzen, J. L. (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics*, *5*, e1000732.
- Bennetzen, J. L. (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics & Development*, *15*, 621–627.
- Bennetzen, J. L., & Park, M. (2018) Distinguishing friends, foes and freeloaders in giant genomes. *Current Opinion in Genetics & Development*, *49*, 49–55
- Bennetzen, J. L., & Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, *65*, 505–530.
- Bergman, C. M. (2012). A proposal for the reference-based annotation of de novo transposable element insertions. *Mobile Genetic Elements*, *2*, 51–54.
- Biemont, C., & Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. *Nature*, *443*, 521–524.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.
- Bonchev, G., & Parisod, C. (2013). Transposable elements and microevolutionary changes in natural populations. *Molecular Ecology Resources*, *13*, 765–775.
- Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., . . . Tiffin, P. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, E864–E870.
- Brookfield, J. F. Y. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nature Reviews. Genetics*, *6*, 128–136.
- Buehler, D., Graf, R., Holderegger, R., & Gugerli, F. (2011). Using the 454 pyrosequencing-based technique in the development of nuclear microsatellite loci in the alpine plant *Arabis alpina* (Brassicaceae). *American Journal of Botany*, *98*, e103–e105.

- Buehler, D., Graf, R., Holderegger, R., & Gugerli, F. (2012). Contemporary gene flow and mating system of *Arabidopsis alpina* in a Central European alpine landscape. *Annals of Botany*, *109*, 1359–1367.
- Buehler, D., Poncet, B. N., Holderegger, R., Manel, S., Taberlet, P., & Gugerli, F. (2013). An outlier locus relevant in habitat-mediated selection in an alpine plant across independent regional replicates. *Evolutionary Ecology*, *27*, 285–300.
- Casacuberta, E., & González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*, *22*, 1503–17.
- Casacuberta, J. M., Vernhettes, S., Audeon, C., & Grandbastien, M. A. (1997). Quasispecies in retrotransposons: a role for sequence variability in Tnt1 evolution. *Genetica*, *100*, 109–117.
- Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M., & Scheid, O. M. (2014). How a retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genetics*, *10*, e1004115.
- Choudhury, R. R., & Parisod, C. (2017). Jumping genes: Genomic ballast or powerhouse of biological diversification. *Molecular Ecology*, *26*, 4587–4590.
- Choudhury, R. R., Neuhaus, J.-M., & Parisod, C. (2017). Resolving fine-grained dynamics of retrotransposons: comparative analysis of inferential methods and genomic resources. *Plant Journal*, *90*, 979–993.
- Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, *351*, 1083–1087.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*, 80–92.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.
- de Villemereuil, P., Mouterde, M., Gaggiotti, O. E., & Till-Bottraud, I. (2018). Patterns of phenotypic plasticity and local adaptation in the wide elevation range of the alpine plant *Arabidopsis alpina*. *Journal of Ecology*, *106*, 1952–1971.
- Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 5294–5300.
- Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., ... Li, F. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nature Genetics*, *50*, 796–802.
- Elbarbary, R. A., Lucas, B. A., & Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science*, *351*, aac7247.
- Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H. A., & Weigel, D. (2018). Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. *Nature Ecology & Evolution*, *2*, 352.
- Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, *338*, 758–767.
- Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, *3*, 329–341.
- Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., ... Widmer, A. (2017). Estimating genomic diversity and population differentiation—an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics*, *18*, 69.
- Garrison, E. (2012). Vcflib: A C++ library for parsing and manipulating VCF files. *GitHub* and <https://github.com/ekg/vcflib> (accessed July 21, 2015).

- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907*.
- Garvin, M., Saitoh, K., & Gharrett, A. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, *10*, 915–934.
- Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*, 1–19.
- Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Resources*, *5*, 184–186.
- Grandbastien, M.-A., Audeon, C., Bonnivard, E., Casacuberta, J., Chalhoub, B., Costa, A.-P., . . . Mhiri, C. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic and genome research*, *110*, 229–241.
- Gray, Y. H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics*, *16*, 461–468.
- González, J., Karasov, T. L., Messer, P. W., & Petrov, D. a. (2010). Genome-wide patterns of adaptation to temperate environments associated with transposable elements in *Drosophila*. *PLoS Genetics*, *6*, e1000905.
- Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., . . . Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, *334*, 83–86.
- He, L., & Dooner, H. K. (2009). Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 8410–8416.
- Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., . . . Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *American Naturalist*, *188*, 379–397.
- Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, *19*, 1419–1428.
- Hollister, J. D., Smith, L. M., Guo, Y.-L., Ott, F., Weigel, D., & Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2322–2327.
- Jiang, C., Chen, C., Huang, Z., Liu, R., & Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics*, *16*, 72.
- Jiao, W.-B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., . . . Schneeberger, K. (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome research*, *27*, 778–786.
- Josephs, E. B., Lee, Y. W., Stinchcombe, J. R., & Wright, S. I. (2015). Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 15390–5.
- Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., & Schulman, A. H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 6603–6607.

- Kaplan, F., Kopka, J., Haskell, D. W., Zhao, W., Schiller, K. C., Gatzke, N., ... Guy, C. L. (2004). Exploring the temperature-stress metabolome of *Arabidopsis*. *Plant Physiology*, *136*, 4159–4168.
- Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160458.
- Kidwell, M. G., & Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, *55*, 1–24.
- Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., ... Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, *39*, 1151–1155.
- Kofler, R., Nolte, V., & Schlötterer, C. (2015). Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genetics*, *11*, e1005406.
- Laenen, B., Tedder, A., Nowak, M. D., Toräng, P., Wunder, J., Wötzel, S., ... Slotte, T. (2018). Demography and mating system shape the genome-wide impact of purifying selection in *Arabis alpina*. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 816–821.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*, 589–595.
- Lisch, D., & Bennetzen, J. L. (2011). Transposable element origins of epigenetic gene regulation. *Current Opinion in Plant Biology*, *14*, 156–161.
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*, 49–61.
- Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, *4*, 981–994.
- Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*, 345–352.
- Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., ... Du, X. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nature Genetics*, *50*, 803–813.
- Marroni, F., Pinosio, S., Zaina, G., Fogolari, F., Felice, N., Cattonaro, F., & Morgante, M. (2011). Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics & Genomes*, *7*, 1011–1023.
- Martin, H. C., Batty, E. M., Hussin, J., Westall, P., Daish, T., Kolomyjec, S., ... Donnelly, P. (2017). Insights into platypus population structure and history from whole-genome sequencing. *Molecular Biology and Evolution*, *35*, 1238–1252.
- McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harbor Laboratory Press*. 21,197–216.
- Morin, P. A., Luikart, G., & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*, 208–216.
- Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*, 579–584.
- Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biology*, *5*, R79.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575.

- Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., . . . Liu, J. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nature Communications*, *6*, 10283.
- Quadrana, L., Silveira, A. B., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., & Colot, V. (2016). The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLIFE*, *5*, e15716.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842.
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*, 4348–4370.
- Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., . . . Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 11479–11484.
- Rey, O., Danchin, E., Mirouze, M., Loot, C., & Blanchet, S. (2016). Adaptation to global change: a transposable element–epigenetics perspective. *Trends in Ecology & Evolution*, *31*, 514–526.
- Ribeiro, A., Golicz, A., Hackett, C. A., Milne, I., Stephen, G., Marshall, D., . . . Bayer, M. (2015). An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC Bioinformatics*, *16*, 382.
- Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2016). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, *18*, 908–918.
- Rogivue, A., Graf, R., Parisod, C., Holderegger, R., & Gugerli, F. (2018). The phylogeographic structure of *Arabis alpina* in the Alps shows consistent patterns across different types of molecular markers and geographic scales. *Alpine Botany*, *128*, 35–45.
- Seeb, J., Carvalho, G., Hauser, L., Naish, K., Roberts, S., & Seeb, L. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, *11*, 1–8.
- Senerchia, N., Felber, F., & Parisod, C. (2015). Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation. *Proceedings of the Royal Society of London B: Biological Sciences*, *282*, 20142874.
- Sigman, M. J., & Slotkin, R. K. (2016). The first rule of plant transposable element silencing: location, location, location. *Plant Cell*, *28*, 304–313.
- Stritt, C., Gordon, S. P., Wicker, T., Vogel, J. P., & Roulin, A. C. (2018). Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the mediterranean grass *Brachypodium distachyon*. *Genome Biology and Evolution*, *10*, 304–318.
- Stuart, T., Eichten, S. R., Cahn, J., Karpievitch, Y. V., Borevitz, J. O., & Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLIFE*, *5*, e20777.
- Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*, e21800.
- Tedder, A., Ansell, S. W., Lao, X., Vogel, J. C., & Mable, B. K. (2011). Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*. *Annals of Botany*, *108*, 699–713.
- Toräng, P., Wunder, J., Obeso, J. R., Herzog, M., Coupland, G., & Ågren, J. (2015). Large-scale adaptive differentiation in the alpine perennial herb *Arabis alpina*. *New Phytologist*, *206*, 459–470.

- Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*, 36–46.
- Treiber, C. D., & Waddell, S. (2017). Resolving the prevalence of somatic transposition in *Drosophila*. *eLIFE*, *6*, e28297.
- Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, *130*, 123–135.
- Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., . . . Albani, M. C. (2009). PEP1 regulates perennial flowering in *Arabidopsis thaliana*. *Nature*, *459*, 423–427.
- Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., . . . Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*, 43–49.
- Wei, B., Liu, H., Liu, X., Xiao, Q., Wang, Y., Zhang, J., . . . Huang, Y. (2016). Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. *BMC Genomics*, *17*, 536.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*, 973–982.
- Willing, E.-M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J., . . . Schneeberger, K. (2015). Genome expansion of *Arabidopsis thaliana* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants*, *1*, 14023.
- Wingler, A., Juvany, M., Cuthbert, C., & Munné-Bosch, S. (2015). Adaptation to altitude affects the senescence response to chilling in the perennial plant *Arabidopsis thaliana*. *Journal of Experimental Botany*, *66*, 355–367.
- Woetzel, S., Kemi, U., Wunder, J., Andrello, M., Rogivue, A., Gugerli, F., Coupland, G. (submitted). *Arabidopsis thaliana* and relatives, model system for ecological genetics and life-history evolution. *New Phytologist*.
- Zamudio, N., Barau, J., Teissandier, A., Walter, M., Borsos, M., Servant, N., & Bourc'his, D. (2015). DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes & Development*, *29*, 1256–1270.
- Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., . . . Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*, 408–414.

Supplementary Information

Additional Supplementary Information may be found in the online version of this article.

Supplementary material:

Appendix S1 Mapping and TE annotation.

Table S1 Number of annotated TEs per tribe in V5.1 of *Arabidopsis* reference genome.

Table S2 List of the SNPs present in TE sequences.

Table S3 Results of linkage disequilibrium estimation for the SNPs and the non-TE SNPs.

Table S4 Number of reads obtained for each sample after whole-genome sequencing and filtering.

Table S5 Comparison among population genomics parameters estimated for four types of loci.

Table S6 Results of Gene Ontology analysis for non-TE SNPs.

Table S7 Results of Gene Ontology analysis for TEs.

Figure S1 Location and aerial photos of four studied populations.

Figure S2 Flowchart of steps involved in TEPIID.

Figure S3 Site frequency spectra of non-TE SNPs and polymorphic TEs.

Figure S4 Venn diagram of shared and private SNPs.

Figure S5 Plot of the density of both SNP sets along the chromosomes.

Figure S6 Efficiency of 'tepid-refine' algorithm.

Availability of data and materials

All genomic data are publicly available through respective data repositories: raw reads are available on NCBI (BioProject ID: PRJNA489364), SNPs and TE polymorphisms are available on DRYAD (doi:10.5061/dryad.58g217k). Custom scripts are available upon request.

Authors' contributions

Designed research: AR, RRC, SZ, FF, MK, SJ, CP, FG

Performed field and lab work: AR

Analyzed SNPs and microsatellite datasets: AR, SZ

Analyzed TE dataset: RRC

Wrote manuscript: AR, RRC, CP and FG, with contributions from all co-authors

All authors read, commented and approved the final manuscript.

Tables and Figures

TABLE 1 Number of single-nucleotide polymorphisms (SNPs) after filtering along steps i to vi. Non-TE SNPs exclude SNPs within annotated transposable element (TE).

| Filtering steps | Number of SNPs |
|---|----------------|
| All SNPs | 78,859,801 |
| i Biallelic SNPs | 60,909,867 |
| ii Quality/Depth >0.25 | 1,493,089 |
| iii Depth: min 8, max 100 and alternative alleles counts min 2, max 100 | 610,027 |
| iv Minor allele frequency <0.025 | 443,801 |
| v SNPs with less than 10 % missing data | 439,670 |
| vi Non-TE SNPs with less than 10 % missing data | 291,396 |

TABLE 2 Population genomics parameters inferred for all (ALL) individuals (N = 304) and separately for each of the four study populations of *Arabis alpina*, based on single-nucleotide polymorphisms (SNPs), SNPs excluding transposable element sequences (non-TE SNPs) and polymorphic transposable elements (TEs). A subsample of 110 individuals was genotyped at 19 nuclear microsatellite loci (Microsat). Observed and expected heterozygosity, H_o and H_e , and inbreeding coefficients, F_{IS} , were estimated..

| Samples | H_o | | | | H_e | | | | F_{IS} | | | |
|-----------|-------|-------------|------|----------|-------|-------------|------|----------|----------|-------------|------|----------|
| | SNPs | Non-TE SNPs | TEs | Microsat | SNPs | Non-TE SNPs | TEs | Microsat | SNPs | Non-TE SNPs | TEs | Microsat |
| All | 0.09 | 0.09 | 0.03 | 0.09 | 0.13 | 0.12 | 0.23 | 0.30 | - | - | - | - |
| Essets | 0.07 | 0.06 | 0.03 | 0.09 | 0.10 | 0.10 | 0.22 | 0.36 | 0.28 | 0.33 | 0.86 | 0.69 |
| Martinets | 0.05 | 0.05 | 0.03 | 0.04 | 0.07 | 0.06 | 0.23 | 0.18 | 0.18 | 0.22 | 0.86 | 0.49 |
| Pierredar | 0.17 | 0.16 | 0.04 | 0.18 | 0.22 | 0.21 | 0.22 | 0.24 | 0.22 | 0.26 | 0.83 | 0.55 |
| Para | 0.08 | 0.08 | 0.03 | 0.06 | 0.12 | 0.12 | 0.23 | 0.46 | 0.26 | 0.30 | 0.85 | 0.59 |

TABLE 3 Matrix of pairwise F_{ST} values (lower diagonal) of four populations of *Arabis alpina* for single-nucleotide polymorphisms (SNPs), SNPs excluding transposable element sequences (non-TE SNPs), polymorphic transposable elements (TEs) and nuclear microsatellites, with their 95% confidence intervals obtained by 100 bootstrappings over loci presented in italics (upper diagonal).

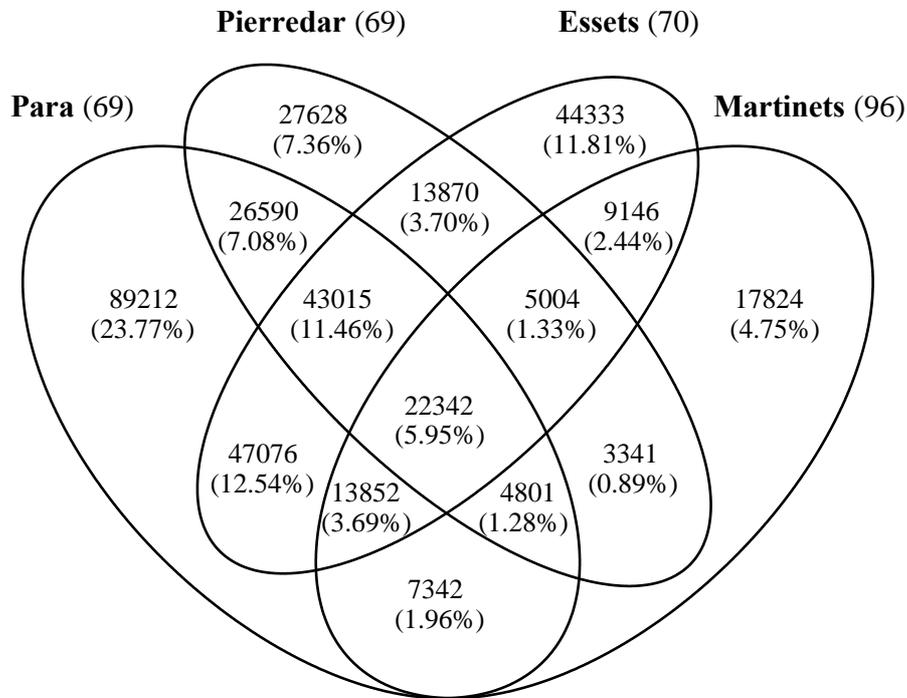
| Markers | | Essets | Martinets | Pierredar | Para |
|-----------------|-----------|--------|------------------|------------------|------------------|
| SNPs | Essets | - | <i>0.13-0.13</i> | <i>0.14-0.14</i> | <i>0.08-0.09</i> |
| | Martinets | 0.13 | - | <i>0.18-0.18</i> | <i>0.09-0.09</i> |
| | Pierredar | 0.14 | 0.18 | - | <i>0.11-0.11</i> |
| | Para | 0.09 | 0.09 | 0.11 | - |
| Non-TE SNPs | Essets | - | <i>0.13-0.13</i> | <i>0.13-0.13</i> | <i>0.08-0.09</i> |
| | Martinets | 0.13 | - | <i>0.18-0.18</i> | <i>0.10-0.10</i> |
| | Pierredar | 0.14 | 0.18 | - | <i>0.11-0.11</i> |
| | Para | 0.09 | 0.10 | 0.11 | - |
| TEs | Essets | - | <i>0.06-0.07</i> | <i>0.03-0.03</i> | <i>0.03-0.04</i> |
| | Martinets | 0.07 | - | <i>0.09-0.09</i> | <i>0.09-0.09</i> |
| | Pierredar | 0.03 | 0.09 | - | <i>0.05-0.06</i> |
| | Para | 0.04 | 0.09 | 0.06 | - |
| Microsatellites | Essets | - | <i>0.09-0.24</i> | <i>0.12-0.28</i> | <i>0.08-0.19</i> |
| | Martinets | 0.18 | - | <i>0.12-0.36</i> | <i>0.17-0.31</i> |
| | Pierredar | 0.20 | 0.25 | - | <i>0.10-0.23</i> |
| | Para | 0.14 | 0.25 | 0.17 | - |

TABLE 4 Counts of observed single-nucleotide polymorphisms outside transposable element sequences (non-TE SNPs) and of polymorphic transposable elements (TEs) for each functional feature compared to their expectation across the assembled genome. Significant differences following Fisher's exact test are reported as *** $p < 0.001$, * $p < 0.05$.

| Marker | Functional features | Region length (Mb) | Proportion in assembly | | Level of significance | |
|-----------------|---------------------|--------------------|------------------------|----------|-----------------------|-----|
| | | | Expected | Observed | | |
| Non-TE SNPs | Genes | 56.90 | 0.17 | 76418.8 | 81018 | *** |
| | Exon | 35.34 | 0.10 | 47460.2 | 48046 | * |
| | Intron | 21.50 | 0.06 | 28881.4 | 32972 | *** |
| | Upstream 2kb | 68.30 | 0.20 | 91741.8 | 78517 | *** |
| | Downstream 2kb | 68.29 | 0.20 | 91724.3 | 81771 | *** |
| Polymorphic TEs | Genes | 56.90 | 0.17 | 3472.0 | 1588 | *** |
| | Exon | 35.34 | 0.11 | 2156.3 | 1113 | *** |
| | Intron | 21.50 | 0.06 | 1312.2 | 957 | *** |
| | Upstream 2kb | 68.30 | 0.20 | 4168.2 | 5241 | *** |
| | Downstream 2kb | 68.29 | 0.20 | 4167.4 | 4664 | *** |

FIGURE 1 Venn diagram showing the number of shared and private variants (with percentage of loci) in each region for (a) single-nucleotide polymorphisms excluding transposable element sequences (non-TE SNPs) and (b) polymorphic transposable elements (TEs) among individuals of *Arabidopsis thaliana* sampled in the four study populations (with number of sampled individuals).

a)



b)

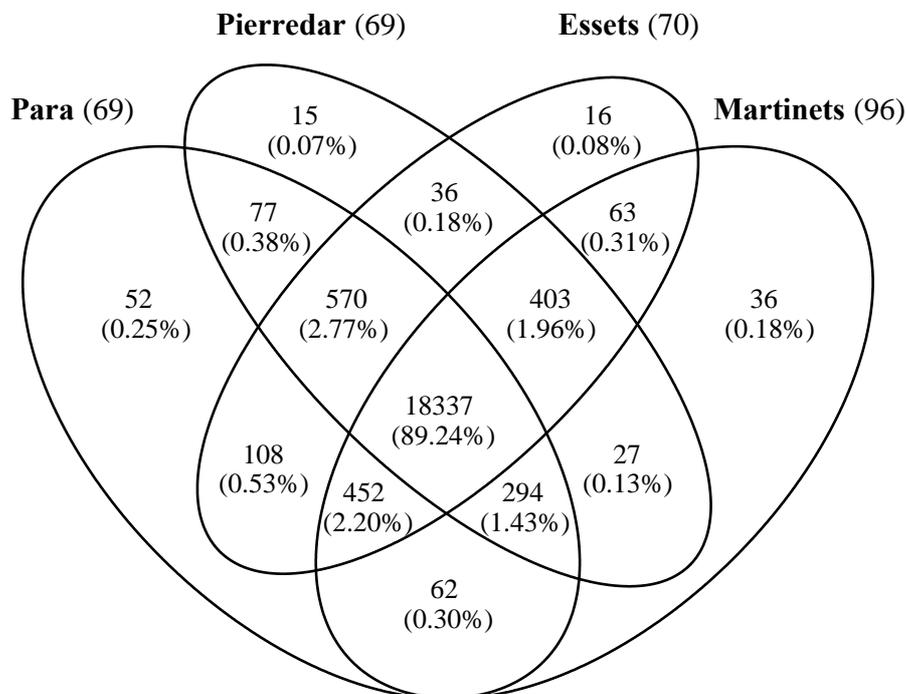


FIGURE 2 Representation of genomic features in 250kb windows along chromosomes of *Arabidopsis thaliana* in four alpine populations. Gene density is from Jiao et al. (2017), complemented by densities of transposable elements (TE) along with single-nucleotide polymorphisms excluding TE sequences (non-TE SNPs) and polymorphic TEs. Linkage disequilibrium (LD) is estimated as the half-decay distance of r^2 . Each coloured line represents a LOESS smooth of corresponding feature along each chromosome (numbered below).

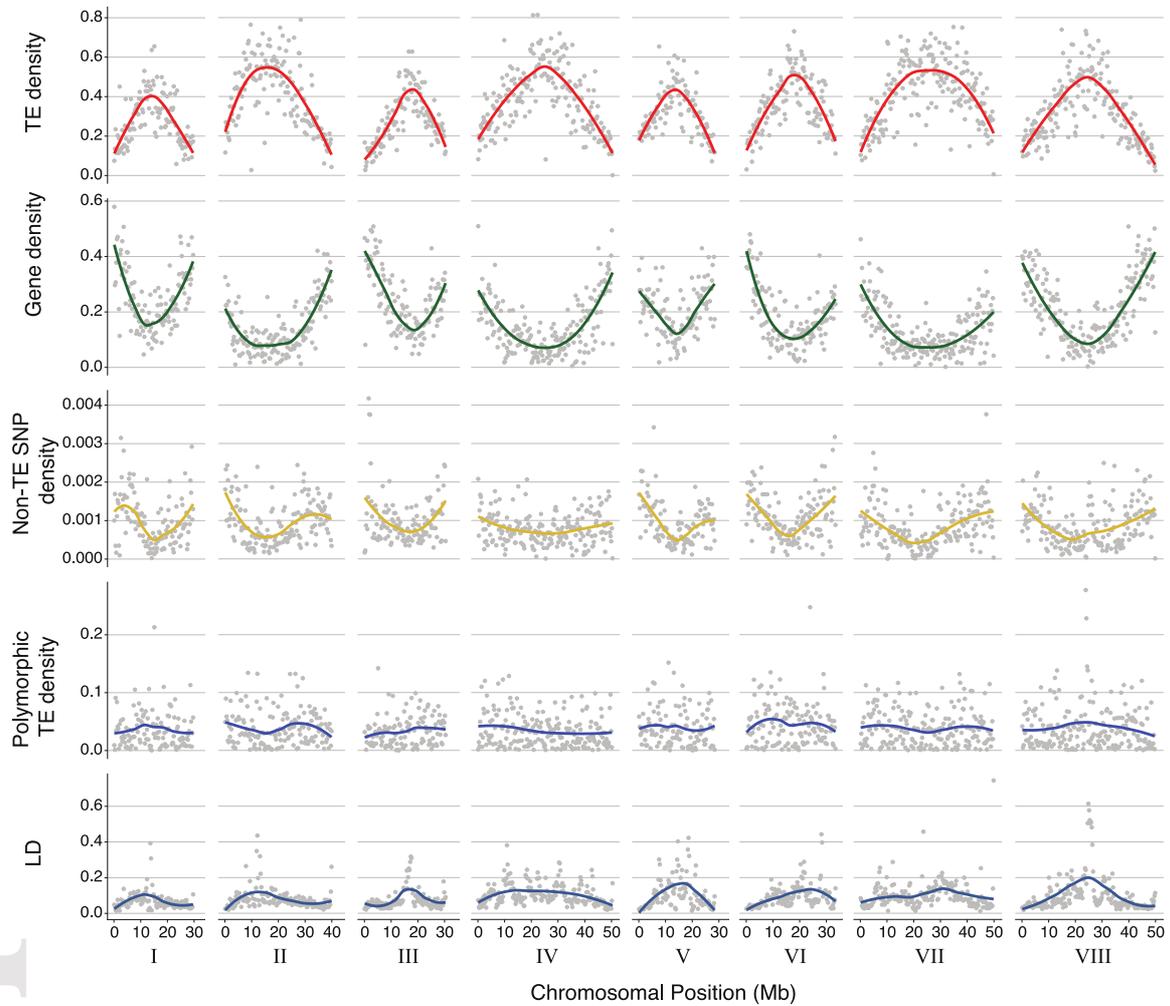
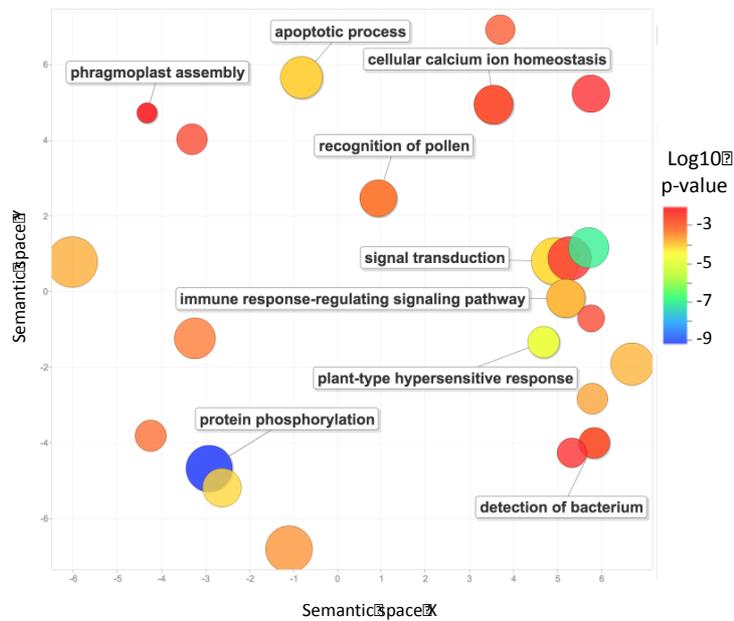


FIGURE 3 Significant enrichment of Gene Ontology (GO) terms ($p < 0.01$) among loci with (a) non-synonymous single-nucleotide polymorphisms excluding transposable element sequences (non-TE SNPs) and (b) polymorphic transposable elements (TEs) next to genes in alpine populations of *Arabidopsis thaliana*. GO terms are coloured according to significance of terms based on the topGO ranking and circle size is representative of the percentage of genes annotated with the corresponding term.

a)



b)

