universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

The genetic basis of parallel edaphic adaptation during the rapid radiation of New Caledonian Diospyros

verfasst von | submitted by

Florian Robert Schmidt BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Genomics and Systems Biology
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Ing. Dr. Ovidiu Paun

Acknowledgements

I would like to thank my Master's thesis supervisor Ovidiu Paun for capturing my interest into evolutionary genomics/adaptive radiations and convincing me to further pursue it. Your invaluable support in terms of feedback or general expertise helped me to always continue and go forward. Also, I would like to thank you and the whole Plant Ecological Genomics group for the warm welcome as well as integrating me in scientific and non-scientific discussions/activities. In particular I would like to show my gratitude towards Teerna and Amin who always answered my endless questions so dearly towards the study system. Additionally, I would like to thank the remaining group including Florian x2, Luiz and past group members for their discussions with me inside or outside group meetings to propel my project or my future endeavors.

This project would not have been possible without the hard work done previously by the people who have collected, curated and studied these plants. Therefore a special appreciation goes to: Jerome Munzinger (IRD Montpellier, France), David Bury (IRD Noumea, New Caledonia), Barbara Turner (BOKU Wien), Rose Samuel (University of Vienna).

Personally, I would like to thank my friends and family who have supported me through all different means throughout these years. I would like therefore to thank my parents Michaela and Robert for always believing and supporting me. And of course I would like to show my gratitude to my partner Isabella, who supported me whenever needed and without whom my motivation would have depleted.

Funding

The research conducted in this thesis was funded by the Austrian Science Fund (FWF) [grant DOI 10.55776/P35275].

Note

I took all measures I could to identify the copyright holders and obtain their consent to use the images in this work. However, if a copyright infringement becomes known, please contact me.

Abstract

Adaptive radiations present exceptional opportunities to study ecological diversification. While considerable attention has been given to adaptive radiations in animals to understand the interplay between ecology and adaptation, research on plant systems remains insufficient. Oceanic islands, with their high concentration of adaptive radiations, serve as natural laboratories for studying natural selection. Among them, the Pacific island group of New Caledonia is particularly renowned for its biodiversity. The island's distinctive geography, geology, and species richness have fostered an explosive radiation of the plant genus *Diospyros* (Ebenaceae), resulting in 24 endemic species during the last 9 My. Multiple species of the radiation have independently adapted to similar ecological niches, most notably to ultramafic soils. Identifying the genetic basis of such parallel adaptations can provide valuable insights into the mechanisms driving diversity in adaptive radiations. To investigate selection in ultramafic-adapted species, this thesis employed whole-genome scans and population genetic tests, with a focus on genes repeatedly adapted to this challenging soil type. Our findings reveal extensive adaptation in functions related to heavy-metal tolerance. One exception was *D. rufotomentosa*, which may either be a generalist or may have adapted via pleiotropic genes. Moreover, several candidate genes associated with ultramafic adaptation were identified as repeatedly adapted, although these findings are limited to non-independent comparisons, rendering further research necessary.

Kurzfassung

Adaptive Radiationen bieten einzigartige Möglichkeiten zur Untersuchung ökologischer Diversifikation. Während adaptive Radiationen bereits bei Tieren erforscht wurden, um die Interaktionen zwischen Ökologie und Adaption besser zu verstehen, mangelt es an umfassend vergleichbaren Studien bei Pflanzen. Ozeanische Inseln, die eine Vielzahl von adaptiven Radiationen beherbergen, können als natürliche Labore für die Erforschung natürlicher Selektion dienen. Besonders die Pazifik-Inselgruppe Neukaledonien ist für ihre Biodiversität bekannt. Eine ausgeprägte Geographie, Geologie sowie Artenvielfalt haben entscheidend zur explosiven Radiation der Pflanzen-Gattung *Diospyros* (Ebenaceae) beigetragen, wobei in den letzten 9 Millionen Jahren 24 endemische Spezies entstanden sind. Mehrere Spezies dieser Radiation haben sich unabhängig in ähnlichen ökologischen Nischen adaptiert, insbesondere auf ultramafischem Gestein. Die Aufklärung der genetischen Grundlagen solcher parallelen Adaptationen kann wertvolle Einblicke in die Mechanismen liefern, die die Diversität in adaptiven Radiationen vorantreiben. Diese Studie beschäftigt sich mit Ganzgenomanalysen und populationsgenetischen Tests, um Selektionen in ultramafisch-adaptierten Spezies zu identifizieren. Besonderes Augenmerk lag auf den wiederholt adaptierten Genen der Spezies in ultramafischem Gestein. Unsere Ergebnisse zeigen umfassende Adaptionen hinsichtlich Schwermetall-Toleranzfunktionen. Eine Ausnahme stellt *D. rufotomentosa* dar, die entweder als Generalist fungiert oder sich durch pleiotrope Gene angepasst hat. Zudem wurden mehrere Genkandidaten in Verbindung mit ultramafischen Adaptionen als mehrfach adaptiert identifiziert. Obwohl die Ergebnisse auf nicht unabhängigen Vergleichen basieren, ist eine weitergehende Forschung unerlässlich.

Contents

List of Figures

1. Introduction

Adaptive radiations offer unique opportunities to study the effects of ecological diversification and natural selection, as they often represent evolution at its greatest pace (Glaubrecht and Schneider, 2010). Since Darwin's foundational work, studies of adaptive radiations have been a cornerstone of evolutionary research - whether to understand mechanisms of speciation (Petren et al., 2005) or the intricate interplay between ecology and adaptation observed for example in African cichlids (Seehausen, 2006). In contrast to the attention given to animal adaptive radiations, similarly specious clades in the plant kingdom remain understudied, although they comprise a substantial proportion to all known radiations (Cerca et al., 2023a).

A high number of adaptive radiations reside on oceanic islands (Cerca et al., 2023a), which serve as excellent natural laboratories to study evolution. These environments often present a mosaic of geological and ecological conditions. Two major factors contribute to their richness in radiations: Firstly, the geological history of islands, which are often relatively young and consist of various rock types promoting habitat heterogeneity (García-Verdugo et al., 2014; Stroud and Losos, 2016). Secondly, the geographical features of islands aiding isolation and unique population dynamics, including genetic drift, allopatric speciation and hybridization. For example, the explosive radiation of African cichlids and other species was largely driven by admixture and ecological differentiation (Seehausen, 2004; Meier et al., 2018).

One particularly interesting region is New Caledonia (NC), which is known for its rich biodiversity and numerous endemic species, in particular plants. The archipelago is located in the southwestern Pacific, approximately 1,220 km east of Australia and about 400 km from Vanuatu. The formation of the New Caledonian main island, Grand Terre, resulted from continental drift, consequently accumulating ultramafic deposits on its lithosphere (Brothers and Lillie, 1988). The aggregation of soils enriched with heavy-metal (e.g., Ni, Cu) and the diverse geography resulted in the complex ecological mosaic seen today in NC (Fig. 1.1A). Apart from a high concentration of heavy metals, ultramafic soils are typically characterized by a low retention of water and a reduced content in macro- and micronutrients, and are generally regarded as stressful for plants. Grand Terre emerged around 37 Million years ago (Mya), and is characterized by a mountain chain up to 1,600 m in height. The mountainous geography created a unique island climate, trapping precipitation on the eastern part of the island, while keeping the other side arid. These geological and climatic features created a unique environment with isolated, diverse habitats in which endemic species have adapted to specific ecological conditions (Grandcolas et al., 2008; Wulff et al., 2013). The relatively recent emergence of this island

cluster suggests that it was only recently colonized by the lineages that gave rise to its current biodiversity (Grandcolas et al., 2008).

The genus *Diospyros* L. (Ebenaceae), comprising over 700 species predominantly found in the tropics, includes agro-economically important species, such as those producing ebony wood or persimmon fruits (Samuel et al., 2019). *Diospyros* has repeatedly colonized the islands of NC during the last 25 My, establishing four extant lineages. One of these lineages resulted from an explosive diversification into nearly all ecological niches in NC except mangroves. The species in this adaptive radiation are widely distributed across the entire archipelago and adapted to varied soil and climate conditions (Fig. 1.1B). Furthermore, in contrast to the non-radiating species of New Caledonian *Diospyros* spp., radiating species exhibit notable divergence in both ecological preference and morphological traits (Fig. 1.1C) (Duangjai et al., 2009; Samuel et al., 2019).

The last decade saw an increased interest in the patterns and processes behind the NC radiated *Diospyros* (Fig. 1.1B). Chloroplast sequencing indicated that ancestral populations to the radiation colonized from Pacific islands 9 Mya (Turner et al., 2013a). Subsequent phylogenetic analyses generated plastid trees with polyphyletic placement of some species, a phenomenon attributed to chloroplast capture, geographic and ecological clustering (Turner et al., 2013b, 2016). Although more clarity has been gained through genomic data (Paun et al., 2016)(Khastgir et al., in prep.), the underlying mechanisms of divergence remain incompletely understood. Notably, dioecy and gene flow via bird-dispersed seeds and fly-pollinated flowers suggests mechanisms beyond simple allopatry (Carpenter et al., 2003; Tassin et al., 2010). Indeed, whole-genome sequencing of the radiation revealed that hybridization is omnipresent and may have contributed significantly to adaptation and speciation in the group. Finally, the demographic history of selected species of the radiated clade indicated constant effective population sizes over time, with a sharp decline in recent times (Brück, 2024; Zechmeister, 2024).

Early efforts using sequencing of a reduced representation of the genome have identified candidate genes under selection related to adaptation in the group. Notably, genes involved in metal ion transport and tolerance were found to differ significantly between species with a preference for heavy-metal rich soils and non-ultramafic relatives (Paun et al., 2016).

To understand what distinguishes radiating species from their non-radiating counterparts, parallel-adapted phenotypes or traits may offer crucial insight into their underlying genetic basis. Convergent and parallel evolution can arise from divergent lineages or from lineages sharing a common ancestor, respectively (Stern, 2013; Lee and Coop, 2017). Directional selection is recognized as the primary evolutionary mechanism driving both parallel adaptation ((Louis et al., 2021; Tan et al., 2020); reviewed in (Cerca, 2023; James et al., 2023)) and the evolution of diverged traits (Whiting et al., 2024).

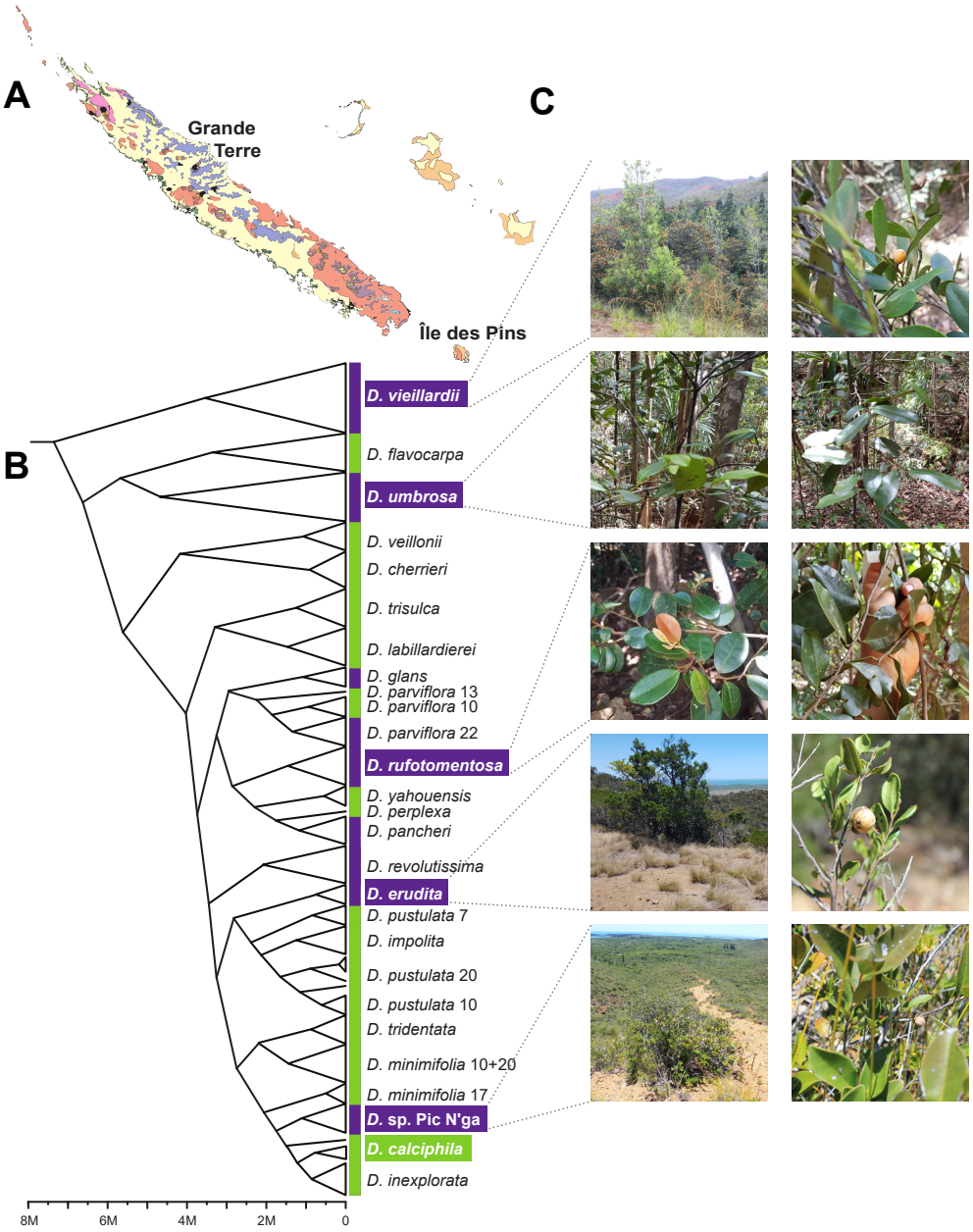

Figure 1.1.: The adaptive radiation of *Diosypros* in NC. (A) The geological/ecological mosaic of NC. Orange and purple regions are maquis and humid forests on ultramafic rock respectively. Map modified from (Paun et al., 2016). (B) Phylogenetic tree of NC *Diospyros* radiation. Purple and lime blocks highlight ultramafic versus non-ultramafic soil preference, respectively. Highlighted names are species used for detailed analyses in this thesis. Image modified from (Paun et al., 2016). (C) Pictures of the studied ultramafic species and their typical habitats. Photos courtesy of Paun O.

Three key factors have been suggested to provide the substrate for directional selection in parallel evolution: (*i*) de novo mutations (which are instrumental in novel trait emergence), (*ii*) Standing Genetic Variation (SGV), and (*iii*) gene flow (Cerca, 2023). These drivers can vary in their influence on adaptive processes. For example, mutations at loci of small effect can significantly shape variation in complex traits following environmental changes (Höllinger et al., 2019, 2023). Nevertheless, given that ancestral populations of adaptive radiations typically have small effective population size (N_e), the contribution of mutations to polygenic adaptation is expected to be minimal during early stages of radiation (Höllinger et al., 2023). As time passes, SGV and gene flow, derived from mutational polymorphisms, may be under directional selection and are not mutually exclusive. For instance, an adaptive radiation in spiders developed parallel ecomorphs through SGV and adaptive introgression (Cerca et al., 2023b). While progress has been made in understanding the genetics of complex traits (Liu et al., 2019; Barghi et al., 2020), a unified framework integrating repeated evolution across systems is still lacking (Lee and Coop, 2017, 2019). It has been proposed that evolution tends to favor specific molecular pathways for trait evolution over functionally redundant alternatives. For instance, convergent climate-related traits in plants have recurrently been selected for, and appear typically associated with the same orthologous gene groups (Whiting et al., 2024). Despite this findings, the underlying genetic basis of parallel adaptation during adaptive radiations remains elusive.

This Master's thesis aims to identify the genetic basis underlying adaptation to ultramafic soils in the *Diospyros* adaptive radiation in NC. The study leverages whole genome sequencing (WGS) data from 93 re-sequenced individuals from six focal species selected for in-depth analysis. Five of these species are adapted to heavy metal containing ultramafic soils, while one species *D. calciphila* is not. This design allows the study of five species pairs along a divergence gradient (Fig. 1.1B), allowing us to test multiple hypotheses regarding the genomic signatures of adaptation. Specifically, this study investigates:

1. which population genomic test (see Chapter 2) can detect adaptation across varying divergence times;
2. how gene reuse or functional convergence is distributed across species pairs adapted to ultramafic versus non-ultramafic soils along a divergence gradient - providing insights into the predictability of evolution;
3. the likely contribution of SGV and gene flow to parallel evolution.

2. Methods

To investigate the degree of parallelisms within the New Caledonian *Diospyros* adaptive radiation, we applied various genomic scans and population genetic statistics focusing on five pairs of species adapted to ultramafic versus non-ultramafic soils along a divergence gradient. Chosen species were: *D. sp PicN'ga* hereafter SPN, *D. erudita* as ERU, *D. rufotomentosa* as RUF, *D. umbrosa* as UMB, and *D. vieillardii* as VIE. Alongside, a species residing in a non-ultramafic soils (i.e., *D. calciphila*; hereafter in abbreviations as "c") was chosen to be paired with the other five species (hereafter called species-pair(s), e.g., cSPN for species-pair *D. calciphila* - *D. sp PicN'ga*).

2.1. Datasets

This study builds upon whole-genome resequencing data of New Caledonian *Diospyros* species, previously generated by the Plant Ecological Genomics group at the University of Vienna (Khastgir et al., in prep.). The dataset includes 93 accessions from the six focal species. The provided, starting Variant Call Format (VCF) file was pre-filtered to exclude - among others - any Single Nucleotide Polymorphism (SNP) within known Transposable Element (TE) regions.

For mapping and variant calling, *Diospyros vieillardii* was used as the reference genome, as it represents the earliest diverging species within the radiation. The scaffold-level assembly of the *D. vieillardii* genome (v1.0; Paun et al., unpublished) and its accompanying gene annotation (Emelianova et al., unpublished) were used to build a custom SnpEff (Cingolani et al., 2012) database for variant annotation (see below). This reference database enabled functional annotations of SNPs, and was subsequently used for gene identification and downstream analysis.

2.2. Data pre-pocessing

The analysis of New Caledonian *Diosypros* spp. focused on contrasting species-pairs along a divergence gradient. To accommodate the different methods (see below) under one unified analysis pipeline, a snakemake workflow was developed. For the execution of the workflow snakemake v9.1.7 (Mölder et al., 2021) was used. This was done to create a reproducible and scalable analysis pipeline, so that multiple species-pairs could be

tested and comparatively interrogated. A repository containing the code and other script resources to run the analysis can be found online (Supp. Data 1).

2.2.1. VCF filtering

To enhance variant quality and reduce biases from low-confidence sites, the vcf file was filtered with vcftools v0.1.16 (Danecek et al., 2011) based on missingness and Minor Allele Frequency (MAF) thresholds. Missing sites were excluded in vcftools with the `--max-missing` flag and a value of `0.1`, which removes sites with a missing genotype rate above 10%. A MAF value was dynamically chosen for each species-pair, to retain variants with alleles present in at least four chromosomal complements. The calculated value was then supplied to the `--maf` flag in vcftools.

2.2.2. Protein prediction

To create unbiased protein sequences for variant annotations, Coding Sequences (CDS) were predicted with Transdecoder v5.7.1 (Haas, BJ; `https://github.com/TransDecoder/TransDecoder`). Transcripts were generated *in silico* based on exon regions from the genome annotation file using AGAT v1.4.2 (Dainat, J; `https://www.doi.org/10.5281/zenodo.3552717`). The transcripts served as input for `TransDecoder.LongOrfs`, with the flag `--complete_orfs_only` to retain only complete open reading frames (ORFs). Finally, `Transdecoder.Predict` was run with the flag `--single_best_only` to retrieve one CDS per transcript and translate it.

2.2.3. Variant annotation

Amino-acid changes of SNPs in the VCF file were annotated with SnpEff v5.2 (Cingolani et al., 2012). A custom SnpEff database was built for the *D. vieillardii* reference genome using the reference FASTA, GFF annotation, and protein FASTA files combined within a new database folder named after the species. A SnpEff config file included the required name and version of each database entry. The database was built with the `build` command of SnpEff and default arguments with `-gff3`, `-noCheckCds` (since the protein FASTA was provided) and `-dataDir`. Annotated VCFs were then produced using the custom database with flags: `-c` (config-file), `-dataDir` (database location), and `-stats` (summary statistics output).

2.2.4. Filtering and allele frequencies

To reduce computation time, SnpEff-annotated VCFs were further subsetted to retain only SNPs causing "missense" or "synonymous" amino acid changes, using bcftools v1.21 (Li, 2011) with the inclusion flag (`-i`) and filtering with following expression: `'ANN ~ "missense" || ANN ~ "synonymous"'`. Further, allele frequencies for each species and SNP in each pair were calculated with vcftools v0.1.16 (Danecek et al., 2011) and the

`--freq` flag, and were filtered with the `--keep` flag to select individuals. The filtered species-pair VCFs were taken from the prior step, to match their length and content.

2.3. Genomic scans for selection

Since genetic drift can cause false positives - especially in long-diverged, early branched sister species - we combined two genomic scan methods for signatures of selection, pcadapt (Privé et al., 2020) and McDonald-Kreitman Test (MKT) (McDonald and Kreitman, 1991) (hereafter genomic scans), together with five tests for neutrality including F_{ST}, nucleotide diversity within- (π_X) and between-species (π_{XY}), Tajima's D within- and between-species (hereafter as neutrality tests) (Weir and Cockerham, 1984; Tajima, 1989). The MKT was used to detect genes under positive selection between species-pairs on the far end of the divergence gradient, whereas the Principal Component Analysis (PCA)-based genome scan of pcadapt for identifying selection signals mainly within recently diverged pairs.

2.3.1. McDonald-Kreitman Test (MKT)

The MKT compared every target, ultramafic adapted species against *D. calciphila* as an outgroup. For this purpose the method described by Blankers T (`https://github.com/thomasblankers/popgen/blob/master/MKTtest`) was adapted so that a single mapped reference (see Section 2.2.3) could be used both for the out- and target-species. Divergent sites were calculated unidirectionally, considering sites divergent only if they were fixed (i.e., reference allele frequency or alternative allele of target equals 1) in the target-species. Polymorphisms were sites where the target allele frequency between 0 and 1. Divergent and polymorphic sites were summarized per-gene. Finally, dN/dS and pN/pS ratios were calculated from these summaries and tested using Fisher's exact test as in (McDonald and Kreitman, 1991).

2.3.2. Outlier scans for recent divergence

Signatures of recent selection were searched with the R package pcadapt v4.4.0 (Privé et al., 2020). To satisfy requirements of pcadapt, species-pair VCF files were converted to PLINK format (.bed) with plink v1.90b6.21 (Purcell et al., 2007), using flags `--double-id`,`--allow-extra-chr`, and `--make-bed`. Pcadapt analyses were performed with K=1, as reruns with values up to K=20 did not improve species separation (see Fig. 3.3B). The pcadapt output reported p-values per SNP but lacks position information. Positions from the PLINK .bim file were combined with pcadapt results in R to produce a combined dataframe and a tab-delimited file. Adjusted p-values and False Discovery Rate (FDR) corrections were added.

2.3.3. Population genetic summary statistics

To provide context for the genomic scans results, three population genetic summary statistics (i.e., F_{ST}, π, and Tajima's D) were calculated using vcftools v0.1.16 (Danecek et al., 2011). Window size was set to 10kb non-overlapping windows (see Section 2.4). The fixation index (F_{ST}) was calculated using the method proposed by (Weir and Cockerham, 1984) via vcftools flags `--weir-fst-pop` and `--fst-window-size`. Tajima's D and π were calculated for both between- and within-species comparisons, with between-species calculations focusing on ultramafic vs. non-ultramafic species pairs, whereas within-species estimates restricted to ultramafic-adapted species. Filtering for within-species comparisons and the respective species was performed with the `--keep` flag in vcftools. Between- and within-species estimates were computed through the `--TajimaD` and `--window-pi` for the Tajima's D and π analyses respectively.

2.4. Window-based analyses and data integration

Bacause outputs from MKT, pcadapt and population genetic summary statistics were heterogeneous and not directly comparable, all results were converted to non-overlapping genomic windows. A 10 kb window size was selected based on linkage disequilibrium (LD)-decay analyses (Khastgir et al., in prep.). Results for summary statistics required no conversion as vcftools output was already in 10 kb windows. Furthermore, all genomic scans results were converted into Browser Extensible Format (BED)-format to facilitate integration.

Vcftools outputs were transformed into individual BED files containing chromosome, start-, end-position, and the statistic values (weighted F_{ST}, Tajima's D, π) plus the variant count. MKT results, summarized per gene but lacking physical position, were first converted to GFF format using the genome annotation of *D. vieillardii*. Awk was used to find positions in the genome annotation containing only "gene" entries. Genomic windows were produced with bedtools v2.31.1 (Quinlan and Hall, 2010) using the genome fasta index and the `makewindows` command. Together with the converted gff-file the genomic-windows were intersected with the bedtools `intersect` command and the `-wao` flag. Intersections lacking a match (indicated by "-1") were excluded. Output columns for chromosome, start, end, as well as the dN/dS, pN/pS, Fisher's exact test p-value, and adjusted p-value were saved as a new BED-file.

Pcadapt site-level results were converted to BED files by duplicating SNP positions as start- and end-position, with end incremented by one. Due to large file size, the BED file was split into equal chunks using bedtools `split -a simple`. Each chunk was then intersected with the created *D. vieillardii* BED-file genomic windows using bedtools' `intersect -wo` flag. Multiple p- and p-adjusted values within one window were transformed as $-\log 10(\text{pvalue})$ and $-\log 10(\text{padj})$ - respectively - using awk. Finally, aggregation of p-values was achieved with bedtools' `groupby` command, where transformed p- and p-adjusted values where smoothed by calculating their mean with `-o mean` flag.

2.5. Functional annotation and enrichment

To support further downstream gene enrichment and ortholog analysis, proteins predicted by TransDecoder were functionally annotated. Two programs were used to assign Gene Ontology (GO) terms. EggNog-mapper (emapper) v2.1.12 (Cantalapiedra et al., 2021) searched the eggNOG database (Huerta-Cepas et al., 2019). Parameters were chosen to find a medium hit rate (ca. 50% of genes) and included following flags: `--pident 60`, `--query_cover 60` and `--subject_cover`, using the default diamond search option for computational speed. In addition, InterProScan v5.73-104.0-11.0.4 (Jones et al., 2014) was used to search the InterPro database (Blum et al., 2025) for functional annotations which were not found through eggNOG. The program was run with following flags: `-goterms` to report GO-terms, `-pa` to also include pathways, `-iprlookup` to include InterPro annotations and `-t p` to indicate the molecule type of the query (protein). A summary file of the InterProScan output was created by choosing the longest name of the ortholog/domain and concatenating GO-terms per gene. Furthermore, the gene name, gene ortholog description, GO-terms, accession number, suggested short name for the gene and the name of the database the ortholog was found in (InterPro/EggNOG) were extracted from both outputs, concatenated and duplicates (i.e. InterPro, because of lower information content) were removed with awk. Awk was used in a similar fashion as for the MKT (see Section 2.4) where a gff-file with genomic positions of the annotated and missing annotated genes was produced. Genomic windows of functional annotations were generated with bedtools as described above.

2.5.1. Functional annotation of selected genes

To reveal the functional implications of putative selected regions, intersected and non-intersected pcadapt and MKT outlier BED files were intersected with a gene annotation BED file using bedtools' `intersect`. Resulting gene lists served as the foreground for GO enrichment, with the full gene annotation set (EggNOG and InterPro) as background. For each species-pair duplicate windows (i.e., candidate genes appearing into multiple windows) were retained only once. (GO) enrichment analysis was performed with the R package topGO v3.22 (Adrian Alexa, 2017) using the conservative "elim" algorithm and Fisher's test on biological process terms. GO terms represented by fewer than one gene or more than 1,000 were excluded to avoid overly narrow or broad functional categories. GO terms with p-value < 0.05 were considered significant.

2.6. Comparison of genomic scans for selection

To compare the effectiveness of the different genomic scans, BED-files for F_{ST}, π_X, π_{XY}, Tajima's D (within and between species), MKT, and pcadapt were filtered with awk. Windows with ≤ 3 sites were excluded. Furthermore, population summary statistics outliers were filtered with awk as the top 5% quantile for all statistics except π_X, where

the bottom 5% quantile was taken. Moreover, pcadapt outliers were determined with awk by including all windows whith $-\log 10(\text{padj}) < -\log 10(0.05)$. MKT outliers were windows/genes under positive selection (dN/dS > pN/pS) and p-value < 0.05.

Outlier windows were intersected individually per species-pair with the pcadapt and MKT candidates using bedtools' *intersect* command with the `-b` flag for summary statistics (F_{ST}, π_X, π_{XY}, Tajima's D) and `-wo` and `-names` for further grouping with bedtools' `groupby`. The count of each name occurrence (`-o count` were piped into awk to add a column to identify the species. To the produced summary file another column was added with awk to indicate the number of total outlier windows of pcadapt. Finally results of the intersection summary table were merged into one larger file from all species-pairs. The same procedure was done also for MKT and both summary tables were finally merged. Distribution plots of summary statistics were created in R using compiled data from all species pairs with the R package ComplexHeatmaps (Gu et al., 2016).

2.7. Parallelism and repeated selection

2.7.1. Testing for parallelism

To assess deterministic gene reuse across species pairs, a Fisher's exact test was employed with the R package SuperExactTest v1.1.0 (Wang et al., 2015) to test departures from stochastic expectations, and to visualize parallelism of genes and enriched functions. Pcadapt and F_{ST} intersected outlier windows with gene annotations (as described in Section 2.5.1) were tested for their gene re-usage and their functional parallelisms of species-pairs (i.e. between pairs, duplicates, etc.). Furthermore, BED-files and a gene table from topGO as obtained in Section 2.5.1 were used as gene-sets and functional-sets respectively. The background population size (n) of the `supertest` function was chosen as the total amount of genes (27,000) and the total number of found GO-terms (5,483).

2.7.2. Identifying parallel windows with Picmin

To detect genome-wide repeated adaptation between distinct lineages, the R package PicMin v0.0.0.9 (Booker et al., 2023) was used following the provided vignette. Since the available data represent a divergence gradient rather than distinct lineages, the method was adapted accordingly. Filtered pcadapt BED files (see Section 2.4 and Section 2.6) were used as input. Parameters `large_i_small_p = TRUE` and `numReps = 100000` ensured precision given the transformed $-\log(\text{padj})$ values. The analysis was repeated three times to analyze species-pairs windows missing 0, 1 or 2 times, as recommended. P-values were FDR corrected and calculated for each missingness-level and for the pooled output (i.e., all missingness-levels together). Outputs were saved as BED-files, annotated, and gene enrichment analyses were performed as explained in Section 2.5.1.

2.8. Data analysis environment and code availability

Genomic and statistical analyses were conducted in R v4.4.1. The entire analysis was performed using the Snakemake workflow management software (Mölder et al., 2021) on the Life Science Compute Cluster (LiSC) of the University of Vienna. All code of the snakemake-workflow and the associated scripts are publicly available on GitHub at `https://github.com/flotji-potji/dioresel`.

3. Results

3.1. Data set variant filtering and annotation

The raw dataset included 93 accessions, with between 10 and 19 accessions per species. Initially, it contained over 10 million SNPs, but after filtering, between 1.2 million and 2.4 million SNPs were retained in the individual datasets for the five different species-pairs (see Table 3.1). Variant annotation with SnpEff showed that the great majority of SNPs did not affect functionality of the respective regions between the species-pairs Table 3.1. Although exonic regions were least affected by SNPs, the majority of functional effects were reported as non-synonymous (missense) substitutions.

Table 3.1.: Summary statistics of VCF filtering and variant annotation. Down- and upstream SNPs (SnpEff default) indicated polymorphism within 5kb of genes.

	cSPN	cERU	cRUF	cUMB	cVIE
Number of accessions	27	29	36	34	35
of which *D. calciphila*			17		
Number of SNPs					
Before VCF filtering			10,813,276		
After VCF filtering	1,230,299	1,390,818	1,579,854	2,399,024	2,123,956
SNP annotation					
% Intergenic	38.7	38.4	38.1	37.9	37.4
% Intronic	19.8	19.9	20.2	20.4	20.8
% Exonic	4.4	4.4	4.5	4.6	4.7
of which missense	2.4	2.4	2.4	2.6	2.6
of which synonymous	2.0	2.0	2.0	2.0	2.1
% Downstream	19.3	19.4	19.3	19.1	19.3
% Upstream	17.5	17.6	17.4	17.5	17.5

3.2. Genomic scans along a divergence gradient

The number of putative selected windows identified by neutrality tests was consistent across tests and species-pairs (e.g., Fig. 3.1). However, outlier windows detected by genome scans tended to increase in number with increasing divergence between species pairs, in particular for MKT (i.e., cSPN $<$ cERU, also see Supp. Fig. A.1). MKT identified between 0 (for cSPN and cRUF) and 72 genes (for cVIE) with signals of positive selection (see Fig. 3.2). Pcadapt detected between 59 (for cSPN) and 9,581 (for cUMB) outlier windows, whereas the increase in divergence levels was not strictly followed (see for example Fig. 3.3). Notably, no MKT outlier windows were detected in cSPN (Fig. 3.4A) and cRUF (Fig. 3.4D), in agreement with the lower divergence between *D. calciphila* and these species, whereas cERU showed greater differentiation than cRUF (see Fig. 3.4D).

To evaluate the results of genome scans and neutrality test quantitatively, we generated heatmaps representing all possible intersections of outlier windows for each species-pair (Supp. Fig. A.1A-E). Since "hot spots" of intersections seemed to be consistent between species-pairs, an averaged heatmap Fig. 3.4B revealed general trends in test interactions. The majority of outlier windows from genome scans overlapped with F_{ST} and π_{XY} top outlier windows. Interestingly, tests with higher proportions of outliers located within gene regions (Fig. 3.4C) showed more frequent intersections among themselves (Fig. 3.4B), while tests with fewer genic outliers exhibited the opposite trend.

Since we aimed to uncover the genetic basis of parallel adaptation, we were interested in which tests can detect already known genes with signals of positive selection related to edaphic adaptation. Paun et al. (2016) reported 50 Resctriction-enzyme associated DNA (RAD) regions with significant fixation index differences between ultramafic and non-ultramafic species within the NC *Diospyros* radiation. Of these, 39 genes were located in the outlier windows uncovered by the present study in the *D. vieillardii* genome (Fig. 3.4A). We observed an excess of pcadapt outlier windows in species-pairs cUMB and cVIE. However, MKT did not overlap with any of the previoulsy published candidate gene regions. This discrepancy might be due to the principle of detection used by the different methods and to the approximately 100-fold difference in the number of outlier windows detected by pcadapt compared to MKT (Fig. 3.4). Finally, it is important to note that RAD data covers only a very small percentage of the genome, potentially missing a notable proportion of the genomic regions of adaptive relevance, and Paun et al. (2016) analyzed their data in the absence of a reference genome.

Although overall patterns of outlier intersections between species-pairs changed little, we examined how intersection proportions varied between species-pairs. As already seen in Fig. 3.4B, π_X and Tajima's D outlier windows exhibited minor overlaps with pcadapt and none with MKT (Supp. Fig. A.2B and C). Tajima's D distributions between species-pairs typically ranged between -2 and 2, consistent with neutrality (Tajima, 1989). However, the distributions seemed to shift asymmetrically towards values indicative of purifying selection (> 2). In contrast, π_{XY} and F_{ST} produced intersections with both genome scan methods across all species-pairs (Supp. Fig. A.2A and Fig. 3.4D respectively).

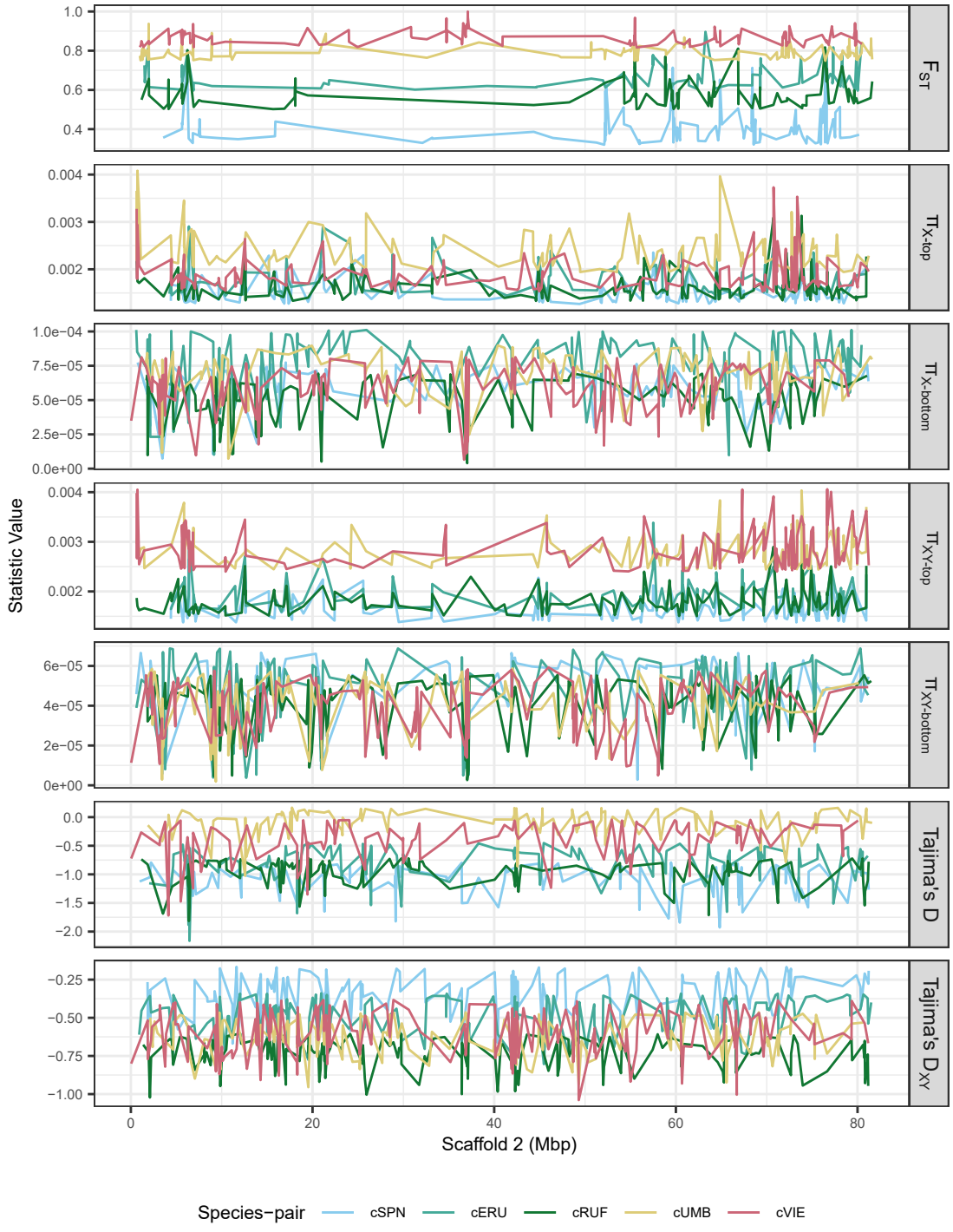

Figure 3.1.: Example of results of population genetic summary statistics on scaffold 2 (ptg000002l). Each panel shows smoothed, window-based estimates of the respective summary statistic (see Section 2.4.

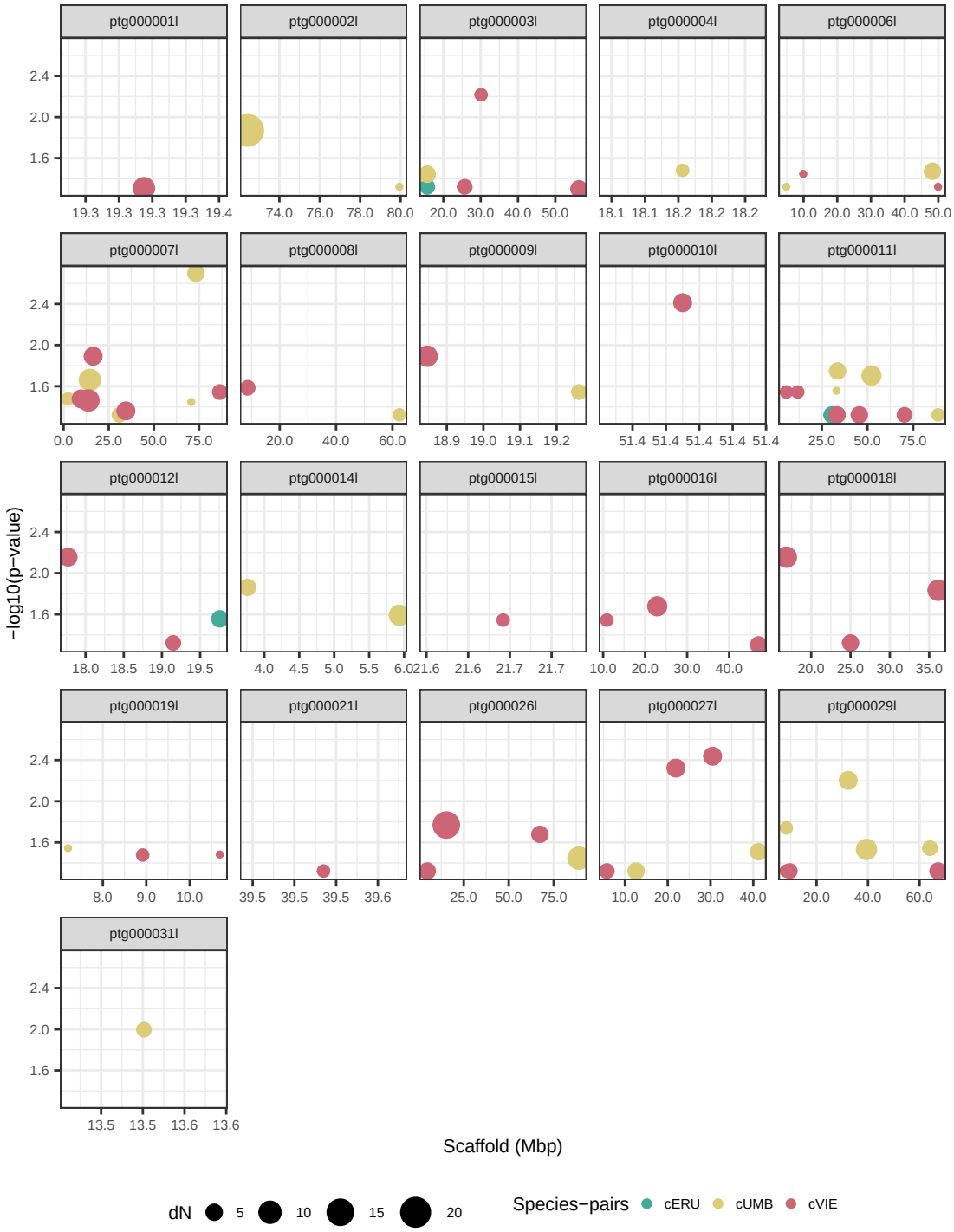

Figure 3.2.: Genome-wide MKT results. Bubbles show genes under significant positive selection, with their color indicating the species pair according to legend and their size indicating the number of divergent non-synonymous substitutions.

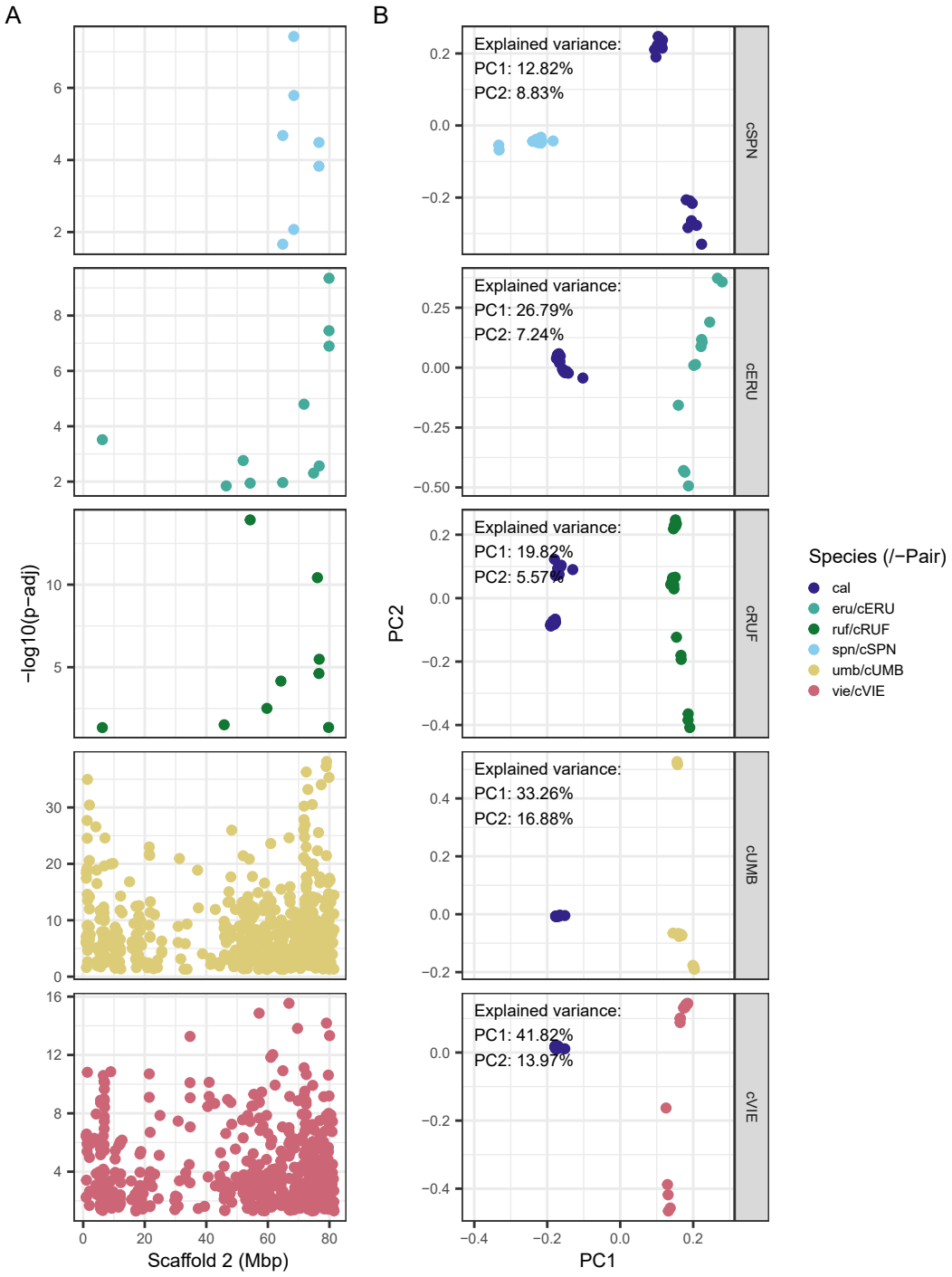

Figure 3.3.: Pcadapt and PCA results. (A-B) Colors in legend and right-hand side labels in (B) correspond to both (A) and (B). (A) Example of pcadapt results on scaffold 2 (ptg000002l). Indicated are the outlier windows with significant FDR < 0.05. (B) PCA of the two major PCs as reported by pcadapt.

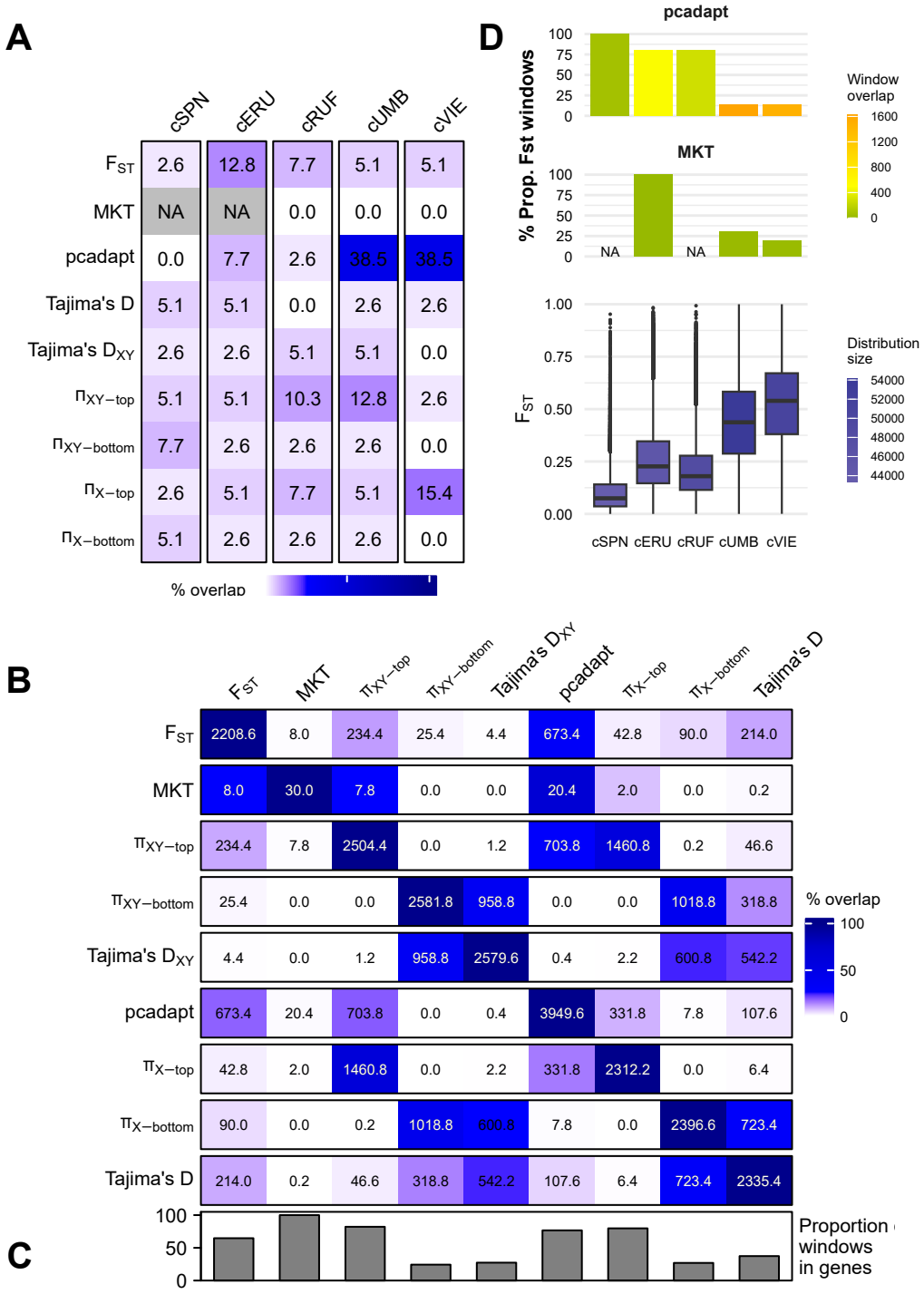

Figure 3.4.: Comparisons of neutrality tests and genomic scans, hereafter referred to as "tests". (A) Detection of previously reported parallel-adapted genes.
Continued on next page

Figure 3.4.: Test outliers intersected with candidate genes identified by (Paun et al., 2016) are shown as proportions of candidate genes (numbers in the heatmap are percentages). Proportions were calculated based on the 39 candidates genes identified in the reference genome (see Supp. Data 2. Horizontally separated columns indicate differences in test proportions for each species-pair. (B) Overlap of outlier windows among tests. The number of window intersections was averaged across species-pairs (see Supp. Fig. A.1). Numbers in the heatmap represent average intersections (i.e., floating point values). Heatmap colors are scaled to the maximum value within each row (i.e. color scales differ for symmetric intersections). (C) Proportion of outlier windows located within annotated gene regions (`gff-type = gene`), averaged across all species pairs (see Supp. Fig. A.1). (D) Proportions of F_{ST} outlier windows that overlap with outlier windows from pcadapt and MKT for each species pair (top two bar plots). Bar plot colors indicate the number of overlapping windows according to legend. The bottom box plot shows the distribution of F_{ST} values across all species pairs, with colors indicating the number of windows in the respective distribution.

When intersected with π_{XY}, pcadapt and MKT outliers behaved similarly, but the number of overlapping genes decreased significantly compared to the original outliers. This effect was less pronounced for F_{ST} intersections in cSPN, cERU and cRUF (Fig. 3.4D). However, F_{ST} distributions shifted notably in cUMB and cVIE, potentially influencing outlier detection.

3.3. Genes and functions under selection

To explore strategies of repeated adaptions to ultramafic soils, we investigated candidate genes detected by genome scans and their functions. As already seen in Section 3.2, pcadapt, F_{ST} and π_{XY-top} produced the most outlier windows in genic regions across species-pairs (Supp. Fig. A.1). Their proportion did not differ between the species pairs, although pcadapt in cSPN was an exception (Supp. Fig. A.1A).

Gene enrichment analysis of pcadapt outliers revealed significant functional enrichments after FDR-correction in cUMB and cVIE, but not in cSPN, cERU and cRUF. This was also the case for MKT, and pcadapt intersections with F_{ST}. Consequently, GO terms with p-values < 0.05 were considered enriched for further interpretation. To highlight biological processes associated to ultramafic adaptation, a list of GO terms of interest was created. Terms found in previous works which were linked to edaphic adaptation (Paun et al., 2016; Meier et al., 2018) or GO terms connected to metal detoxification-processes were chosen. The latter terms were chosen based on metals found in abundance in New Caledonias ultramafic soils (Khastghir et al., in prep.) (Grandcolas et al., 2008; Wulff et al., 2013); also see Supp. Data 3 for chosen GO terms).

Enriched biological processes from pcadapt outliers predominantly related to gene expression/regulation and cellular signaling (Fig. 3.5, for a full list see Supp. Data 3). Most species-pairs, except cRUF showed GO terms potentially related to ultramafic adaptation. Notably, cSPN (Fig. 3.5A) and cERU (Fig. 3.5B) showed larger proportions of GO terms and genes associated to those terms. Additional terms such as "xenobiotic transmembrane transport" (Fig. 3.5C) or "vacuole organization" (Fig. 3.5D) may be linked to ultramafic adaptation.

Beyond gene regulation associated GO terms, a minority of biological processes involved development and morphogenesis, such as "leaf development" (Fig. 3.5E) or "flower development" (Fig. 3.5C). Finally, the trend that the majority of GO terms were related to cellular signaling or regulation is reinforced by a gene enrichment for the collective gene-sets from all species-pairs (Fig. 3.5F).

Since MKT yielded no outliers for cSPN and cRUF, gene enrichment was performed only for the remaining species-pairs (Fig. 3.6). In stark contrast to pcadapt outliers (Fig. 3.5), MKT-enriched GO terms related primarily to plant developmental and morphogenic processes (Fig. 3.6), often with greater functional specificity. Interestingly, additional GO terms appeared for MKT genes from combined cERU, cUMB and cVIE sets (Fig. 3.6 all), including "response to nematode" or "response to fungicide", each supported by >4 genes. The "response to nematode" term corresponded with adaptation of the metal-nicotianamine transporter gene YSL9 in cUMB, potentially linked to metal accumulation.

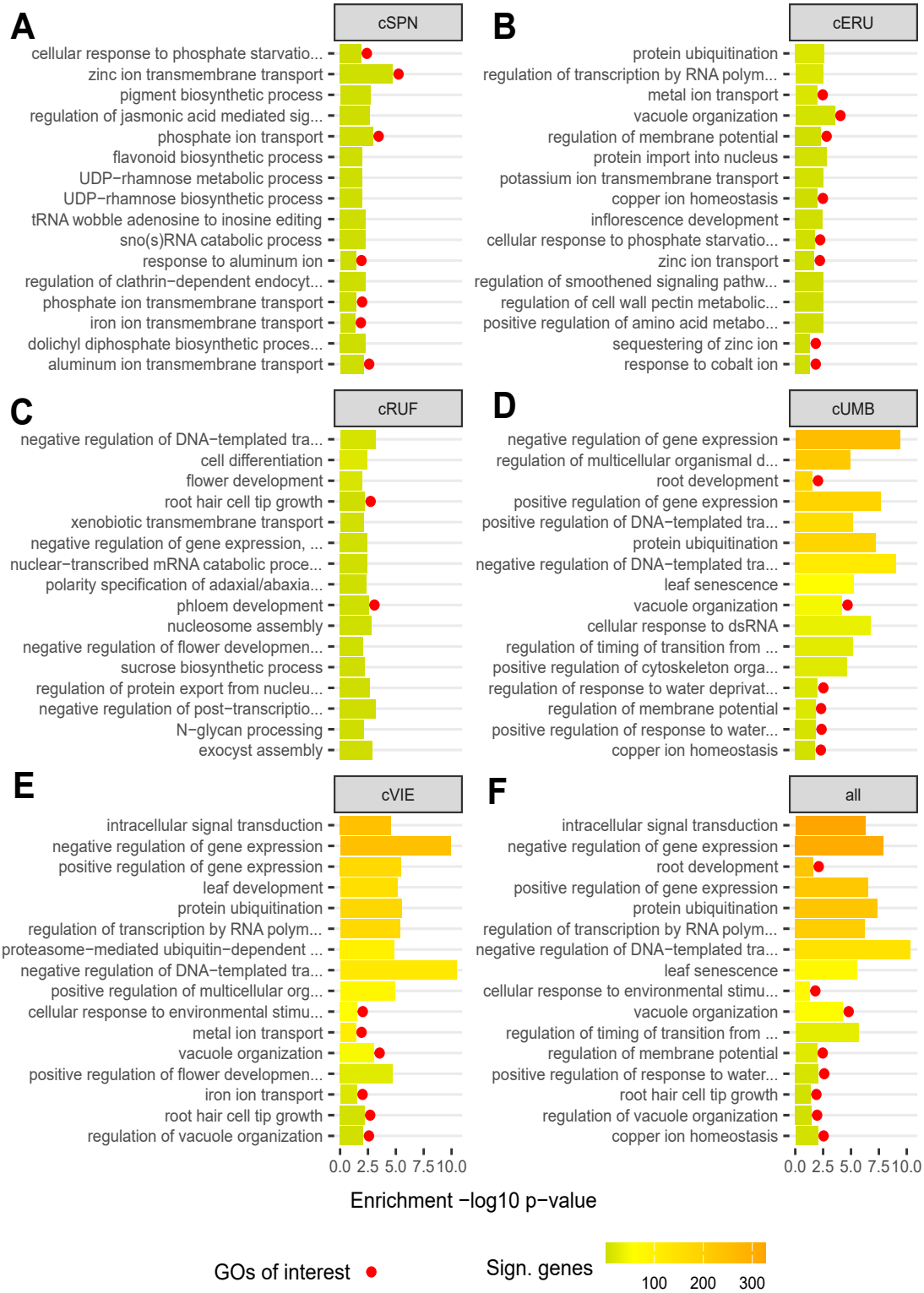

Figure 3.5.: Gene enrichment results of pcadapt outliers. For detailed description see Fig. 3.6.

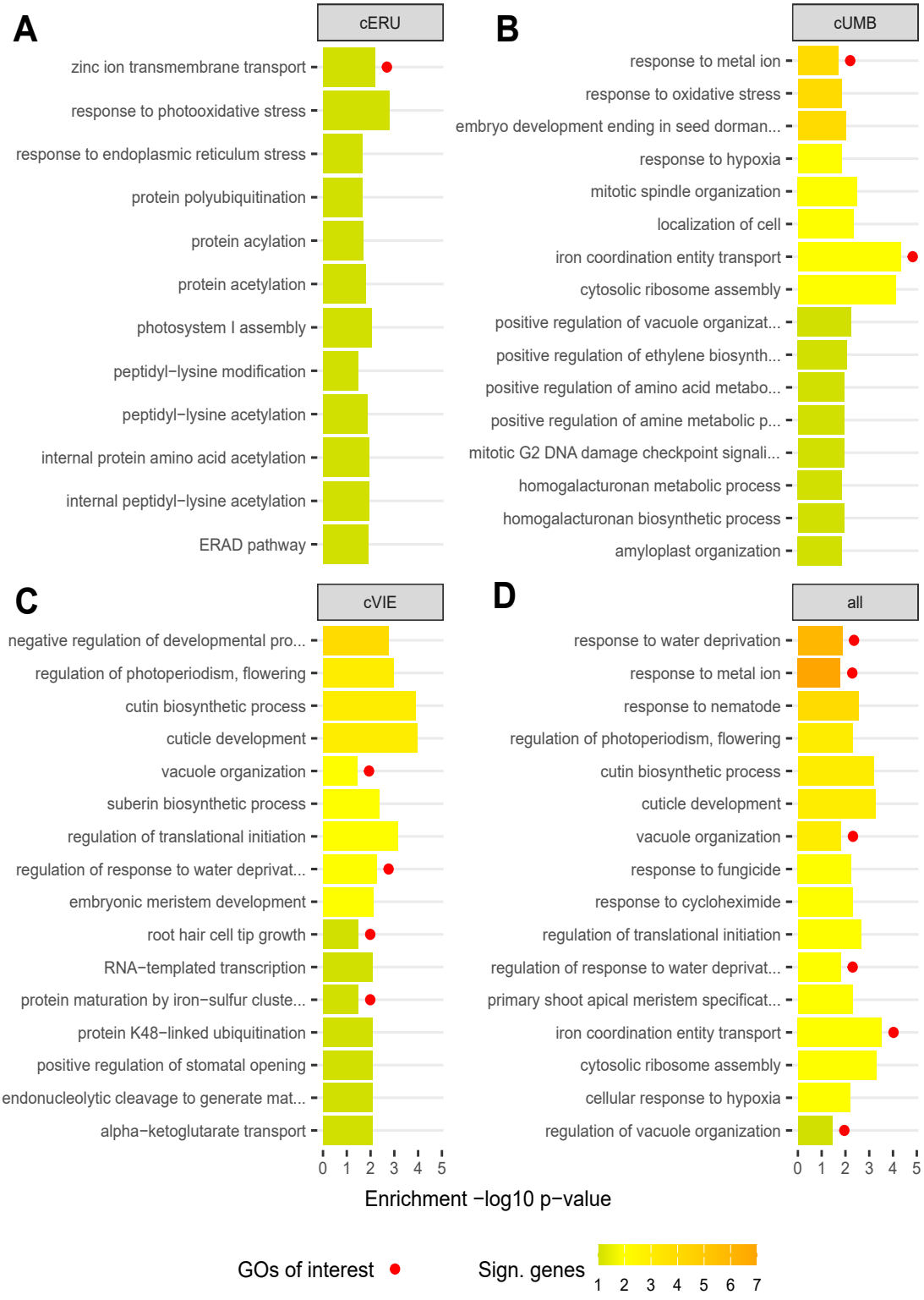

Figure 3.6.: Gene enrichment results of MKT outliers.
Continued on next page.

Figure 3.6.: (A-D) Top 16 significant (p-value < 0.05) GO terms taken from Supp. Data 3 are shown. Here, the shown term consist either of general significant terms or those which were present in the list of GO terms associated to plant heavy metal soil adaptation (Supp. Data 3). Bar height indicates p-values obtained from the gene enrichment's Fisher's exact test (used algorithm: "elim"). Bar color depicts the number of genes which were present in the enrichment gene set. (D) An enrichment Enrichment results for genes identified as outliers across all species pairs only once in (A-E).

Although gene enrichment revealed significant functional evolution within individual species-pairs, it offered limited insights into a shared ultramafic adaptation history (Fig. 3.5F or Fig. 3.6F). To address this, we refined pcadapt outlier genes by intersecting them with F_{ST} outliers, balancing the large discrepancy observed in outlier window counts (59 in cSPN versus 9,261 in cVIE) (see Fig. 3.4D). This produced enrichment results in cSPN, cERU and cRUF comparable to unfiltered sets (Fig. 3.7A-C), likely due to large window retention (Fig. 3.4D). Conversely, the intersection removed GO terms of interest in cVIE (Fig. 3.7E), while cUMB showed an increase in terms which might be related to ultramafic adaptation (Fig. 3.7). Lastly, a combined gene enrichment of all species-pairs intersected outliers increased the number of significant GO terms of interest (Fig. 3.7F), despite low number of associated genes to the terms and borderline p-values. Thus, filtering outlier windows in early-diverging species-pairs (cUMB and cVIE) enhanced detection of shared ultramafic adapation signatures.

Finally, intersected F_{ST} and pcadapt outliers from cSPN, cERU and cRUF, together with MKT outliers from cUMB and cVIE, were combined for downstream analyses (Fig. 3.8). This collective gene enrichment revealed more genes supporting GO terms of interest (Fig. 3.8F). To test which species-pairs contribute to the gene enrichment, overlaps of genes and GO terms showed that cERU and cRUF contributed most shared genes (Fig. 3.9A), while cVIE added a significant portion of functions (Fig. 3.9B).

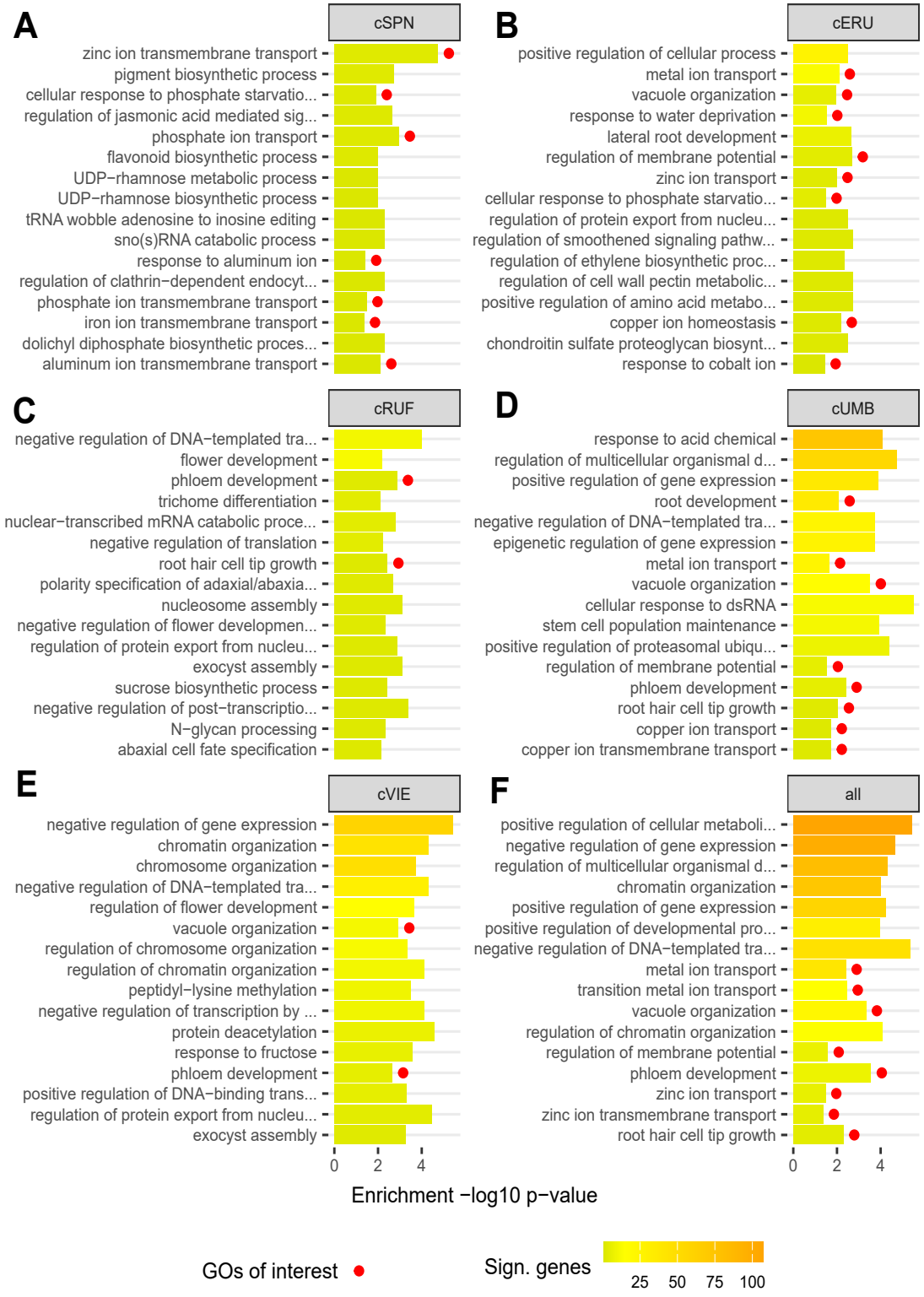

Figure 3.7.: Gene enrichment results of F_{ST} intersected with pcadapt. For detailed description see Fig. 3.6.

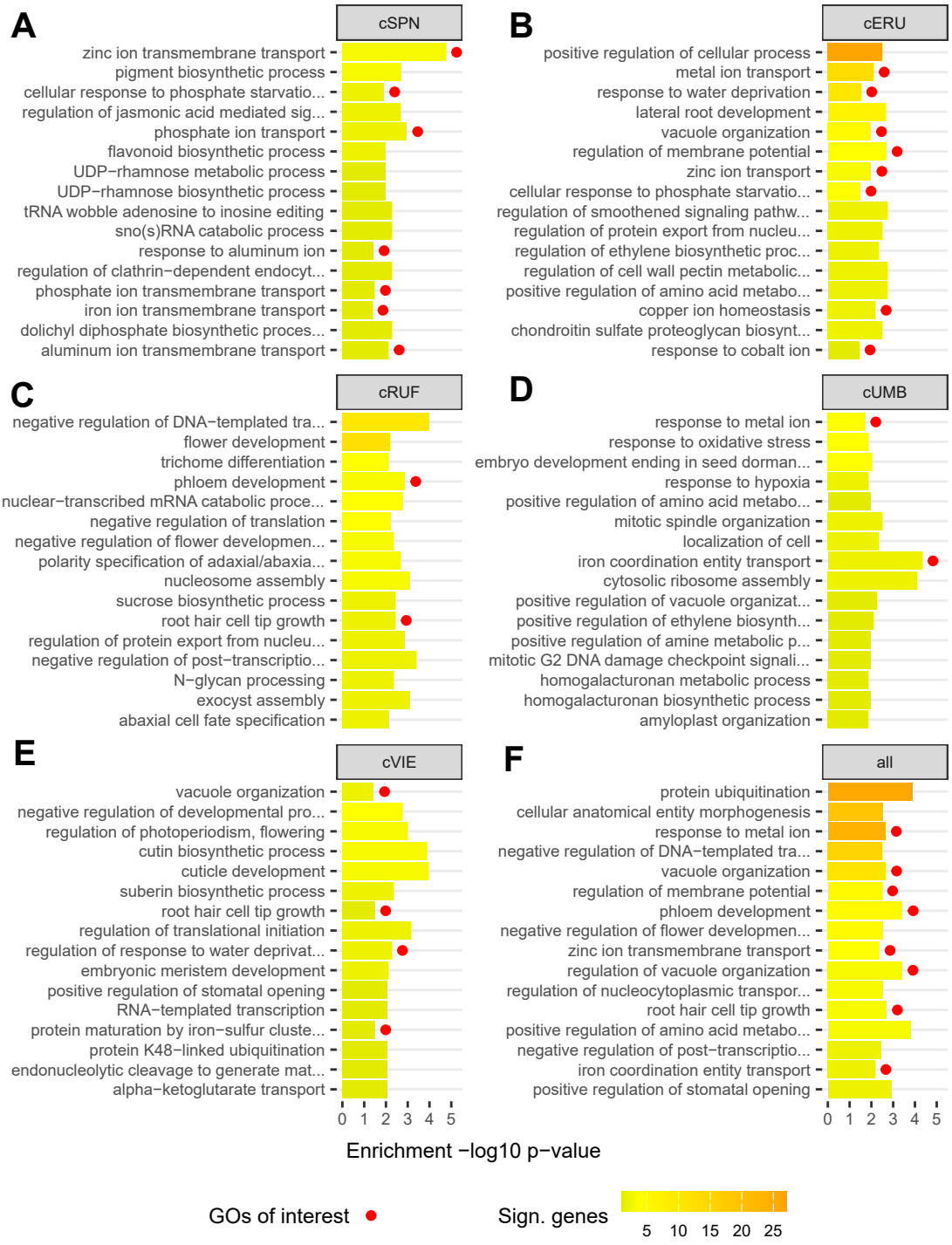

Figure 3.8.: Gene enrichment of pcadapt and MKT outliers. For detailed description see Fig. 3.6.

3.4. Repeated selection in ultramafic edaphic species

3.4.1. Gene reuse and functional overlaps

After filtering, the collective gene enrichment showed an increased number of GO terms of interest. To explore how species-pairs share enriched functions versus individual genes, we performed multi-set intersections of gene- and GO term sets across all species-pairs. We tested whether the observed overlaps exceeded expectations given set sizes. Similar methodologies have shown that gene-reusage decreases with increasing divergence between lineages, whereas functional parallelism seemed to be independent of divergence (Bohutínská et al., 2021). Here, divergence was measured as F_{ST} between ultramafic species (e.g., between *D. sp PicN'ga* and *D. erudita*). The focus of the analysis relied primarily on the gene enrichment results which were presented in Section 3.3.

Disproportionate gene-set sizes from pcadapt led to high proportion of significant multi-set intersections (Supp. Fig. A.3), particularly combinations involving cUMB and cVIE. Here, cUMB-cVIE or cERU-cUMB shared a greater proportion of gene and function intersections than newly diverged species-pair combinations such as cERU-cRUF. Of 26 possible combinations, 22 showed significant gene overlaps (p-value < 0.05). Functional multi-set intersections (Supp. Fig. A.3B) were fewer (15 combinations), but most were significant. Other than intersection and set size difference between the gene and functional multi-set intersections, there were more similarities than distinctions. Especially, the three top combinations shared the majority of the total intersection size and were in both cases ordered similarly (Supp. Fig. A.3A and B).

Intersecting F_{ST} and pcadapt outlier windows (same as in Fig. 3.7) reduced gene set size disparities, lowering significant gene intersections combinations from 22 to 17, with 2 combination losing significance (Supp. Fig. A.3C). The number of combinations and how many of them were significant did not change between unintersected (Supp. Fig. A.3B) and intersected functions (Supp. Fig. A.3D).

Further investigation how combination of species-pairs share the combined pcadapt and MKT results showed inverse results to the ones above. From the study of Bohutinska et al., we would expect with increased divergence less gene reusage (Bohutínská et al., 2021). This was shown in Fig. 3.9 where combination of species-pairs with lower divergence exhibited a higher amount of overlap. Gene and functional intersections resulted in 8 combinations for both of the total possible 26. Additionally, just 2 and 3 significant intersection-combinations had a higher than expected number of overlaps for genes and functions respectively. Interestingly, gene-set sizes were dominated by species-pairs cERU and cRUF (Fig. 3.9A), whereas functional-set sizes were more evenly distributed (Fig. 3.9B).

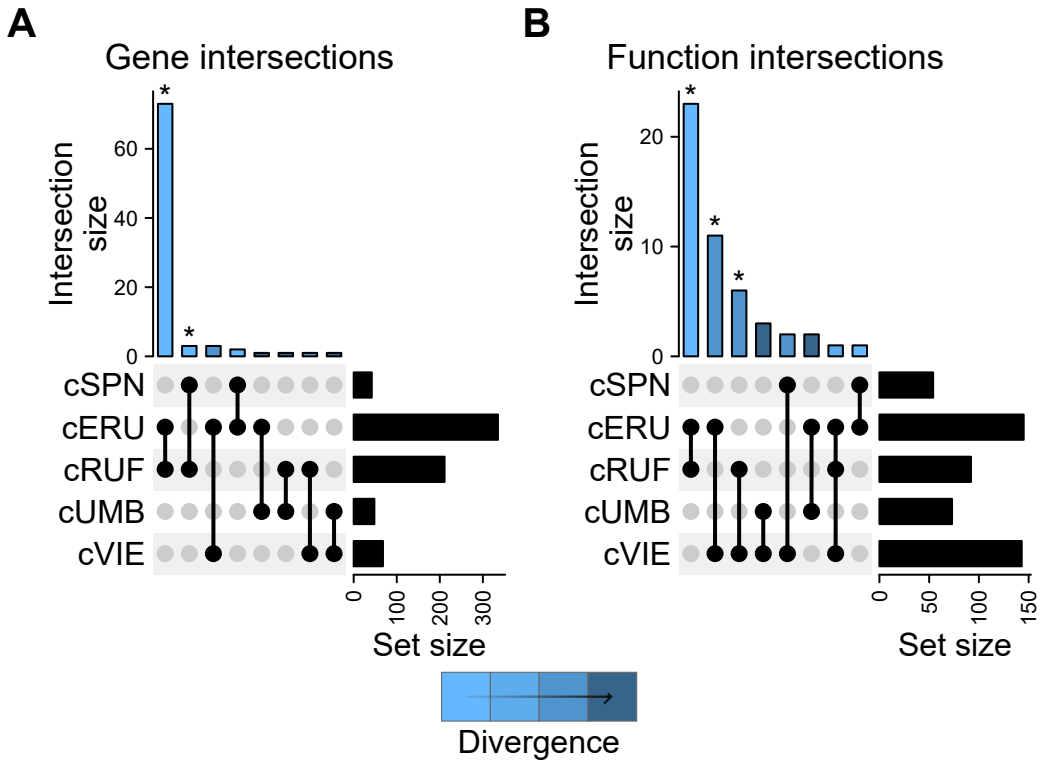

Figure 3.9.: Share of repeated genes and functions, between F_{ST} intersected pcadapt outliers of cSPN, cERU and cRUF and MKT outliers of cUMB and cVIE. (A-B) showing upset plots of species-pairs gene or function set intersections. Bar color indicates the amount of differentiation F_{ST} between two or more species (i.e. cSPN-cERU: F_{ST} *D. sp PicN'ga* and *D. erudita*), and was grouped into four levels (bottom legend). Top bar-anchored asterisk indicates if multi-set intersection passed signifance threshold (p-value < 0.05). Shown intersections were all existing possibilities (singletons not shown). (A) Unique genes from pcadapt (cSPN, cERU, cRUF) and MKT (cUMB, cVIE) were used for multi-intersections. (B) Whereas gene enrichment results as shown in Fig. 3.8 were taken.

3.4.2. The genetic basis of parallel adapted traits

While Section 3.4.1 assessed repeated selection on a per combination-set basis, this approach could not pinpoint individual genes under parallel selection. To address this, we applied PicMin (Booker et al., 2023), a method that can detect repeated adaptation at the genomic window level. PicMin was used here on the genome-wide scan pcadapt results from all species-pairs.

The analysis with PicMin found first hand evidence that ultramafic species repeatedly used the same gene for adapting to heavy metal containing soil (Fig. 3.10). In total, 149 windows were found to be significantly (FDR < 0.05) diverged during the repeated adaptation between the studied species-pairs (Supp. Fig. A.4). These windows varied in their presence across species-pairs (e.g., the gene HMA3 in Fig. 3.10B windows present in all species-pairs, but was classified as significantly repeated adapted by PicMin only in cERU, cUMB and cVIE). In accordance, gene enrichments of gene-sets stratified by missingness revealed that most GO terms were associated to general cell regulatory or developmental processes (Supp. Fig. A.5). However, windows present in all species-pairs were enriched for functions related to metal ion transport and detoxification (Supp. Fig. A.5B, Fig. 3.10), consistent with ultramafic adaptation. These terms were present as well when combining all levels of missingness (Supp. Fig. A.5D).

Repeatedly adapted genes and functions broadly fell into abiotic and biotic responses categories (Fig. 3.10). While the majority of genes and function were related to biotic responses, a few abiotic processes were related to ultramafic adaptation, such as "response to cobalt ion" (gene HMA3) or "vacuolar acidification" (V-ATPase). Both genes contained non-synonymous substitutions in varying amounts across species-pairs, with a maximum of 9 for HMA3 in cRUF and 5 for V-ATPase in cERU.

Next to these abiotic factors, other biological mechanisms such as those associated to the circadian rhythm with genes such as the "Carbon catabolite repressor protein 4" (Fig. 3.10B) were found to be repeatedly adapted. As already mentioned, genes and functions associated to biotic mechanisms included developmental processes. One gene which was significantly (by PicMin) repeatedly adapted in all species-pairs was a Homeodomain containing gene (Supp. Fig. A.4). Other genes such as HASTY 1, ETR3 or VIP2 which could also be related to developmental processes were found to be present in four out of the total 5 species-pairs (Supp. Fig. A.4, Fig. 3.10B). Their associated enrichments included biological processes such as "leaf morphogenesis" or "xylem and phloem pattern formation" (Fig. 3.10A). A "transcriptional regulator" gene was also found to be highly repeatedly adapted (4 out of 5 species-pairs) which exhibited enriched GO terms to "maintenance of DNA repeat elements" (Fig. 3.10A). Gene enrichments of PicMin included a greater share of developmental GO terms (Supp. Fig. A.5) whereas the pcadapt results had a greater share of processes related to cellular regulation (Fig. 3.5).

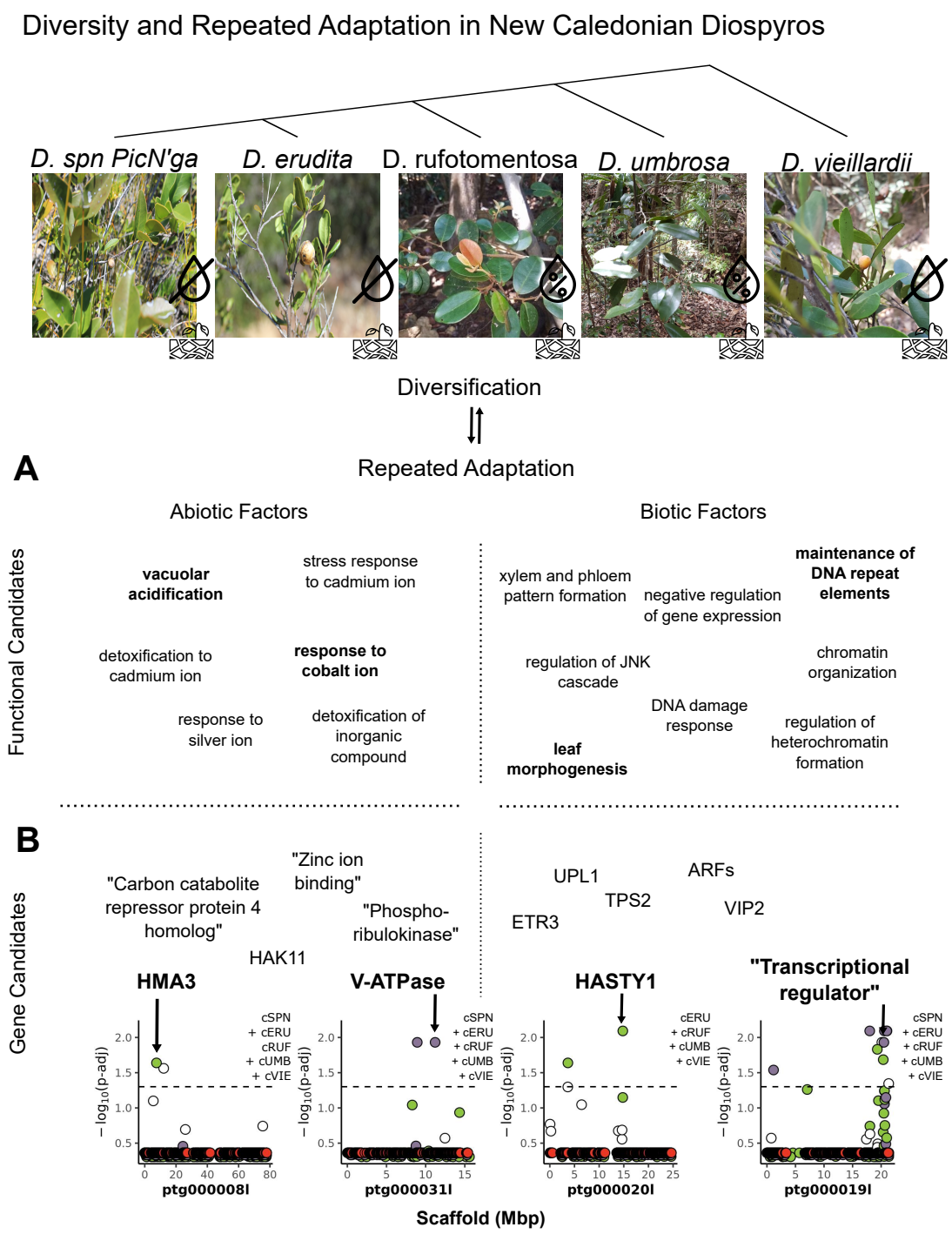

Figure 3.10.: The interplay between diversification and repeated adaptation. The top panel shows the ultramafic species of the New Caledonian *Diospyros* adaptive radiation in their presumed phylogeny. The images highlight how the species have adapted different morphologies.
Continued on next page

Figure 3.10.: Pictograms show the difference in preference to dry (crossed out water droplet) and humid (percentage sign in water droplet). Although the species have diversified in other traits they have repeatedly adapted to the same soil type. This was leveraged to find regions in the genome of the species which were presumed to have been repeatedly used for adaptation. (A-B) presents processed results found with PicMin, a method to determine regions under repeated selection. Of 149 PicMin outlier windows (FDR < 0.05), 133 remained inside gene regions (B). Outlier gene windows were used for gene enrichment analysis (A). Most interesting genes and functions were shown here (for a full list of genes and functions see Supp. Data 4, a manhattan plot of PicMin results Supp. Fig. A.4, and highlighted gene enrichments Supp. Fig. A.5). Left side of the lower panel refers to genes and functions which can mostly be associated to a abiotic mechanism, while the right side refers to biotic factors respectively. (A) enriched GO terms of repeatedly adapted functions. (B) parallel gene candidates which highlight the position on the *D. vieillardii* scaffold. Species-pair names next to scaffold indicate if windows were present in species-pair pcadapt result, whereby a "+"-sign indicates if PicMin found window to be significant. For color of dots please refer to Supp. Fig. A.4.

4. Discussion

This study demonstrates that species-pairs inhabiting the same soil-types have adapted to edaphic conditions via similar genes. Previous findings of selection signatures based on RAD-sequencing were expanded using Whole Genome Sequencing (WGS), enabling corroboration and further testing across various genomic scans and population genomic assays (Paun et al., 2016). The combination of diverse tests revealed quantitative effects correlated with species pairs with different divergence times. While gene enrichment analyses confirmed previously found adapted functions (Paun et al., 2016), it was observed that these functions were largely contingent upon the number of outlier windows incorporated into each analysis. The functions suggested a potential selection trend across all species-pairs. A greater disparity was highlighted with respect to gene and functional sharedness, which shifted between divergence times (Bohutínská et al., 2021). Using the PicMin approach, we were able to find gene and functional candidates which are tentatively regarded as repeatedly adapted (Booker et al., 2023). This thesis thus provides a framework for future investigations applying the developed pipeline to different pairings, including independent lineages rather than only species-pairs.

The different genomic scans and neutrality tests exhibited an increase in outlier windows with increased differentiation towards *D. calciphila*. This was particularly evident for MKT, where a distinct cut-off threshold was observed for genes exhibiting significant positive adaptation. This phenomenon may be attributed to the fact that cSPN and cRUF were not as differentiated from *D. calciphila* as the other species, limiting the number of divergent sites required for statistical significance (McDonald and Kreitman, 1991). This highlights the importance of selecting outliers based on pair-wise differentiation rather than purely on phylogenetic distance (e.g., *D. erudita* appears closer to *D. calciphila* phylogenetically [see Fig. 1.1B], despite greater divergence). The agreement among within-neutrality tests likely reflects the influence of purifying selection on many outliers, supported by shifts in Tajima's D values towards positive values, with most outliers above 2 (Tajima, 1989).

Nonetheless, neither within- nor between-neutrality tests capture the entire range of outlier windows generated by MKT or pcadapt. Each method has different temporal sensitivity for detecting selection. MKT outliers, for example, were rarely detected by neutrality tests, reflecting its suitability for long-diverged species pairs. One notable exception was observed in Supp. Fig. A.2C where cUMB exhibited a window identified both by MKT and Tajima's D, potentially reflecting recent selection or drift (since no Tajima's D fell below -2). Although between-neutrality tests retained more information compared to within-tests, the shape of their statistical distributions differed substantially.

Drift or other non-selective processes likely influenced F_{ST} outlier in cUMB and cVIE, as indicated by heterozygous SNPs of neutral origin scattered throughout the genome (Weir and Cockerham, 1984). Similarly, pcadapt outliers may be affected by neutral processes, particularly in cUMB and cVIE, where approximately one-sixth of windows were presumed adaptive, but this may be inflated.

The intersection of previously found adapted genes and genome scans with neutrality tests might further corroborate the notion of drift in cUMB and cVIE. The predominant intersections of pcadapt outliers in cUMB and cVIE (Fig. 3.4A) likely result from the large amount of outliers generated by pcadapt. Additionally, although drift could have contributed to F_{ST} outliers, it is notable that only a small number was found, despite the gene candidates being originally identified through F_{ST}. This may be related to the limited sample size and the filtering of F_{ST} values from the previous RAD-seq study (Paun et al., 2016), whereas the outliers in this thesis were identified with more stringent criteria (see Section 2.6). Consequently, to address the initial hypothesis, all tests employed were able to find throughout all species-pairs (except for MKT in cSPN and cRUF) putative signatures of adaptation (see Section 2.6 for definition). Admittedly, the usage of tests depends very much on the proposed question as MKT being optimal for pairs with long diverged pairs, neutrality tests for recent adaptation, and pcadapt occupying an intermediary position. Therefore, no single method can consistently identify a uniform number of outliers across all species-pairs.

Gene enrichment analysis revealed that a majority of GO-terms were enriched in biotic processes with relatively few related to abiotic adaptation. The pcadapt results were heavily influenced by the cUMB and cVIE species-pairs, due to their high numbers of outliers, leading to generalized enriched functions. The categorical enrichment algorithm "elim" employed in this study may be overly specialized when the objective is to identify genes with minimal annotated GO terms (Adrian Alexa, 2017). For instance, multiple genes displayed heavy metal ATPase (HMA) family motifs primarily associated with "metal ion binding." A more conservative enrichment algorithm, such as over-representation analysis, might better detect relevant gene functions. Notably, the species-pair cRUF was an exception, showing neither enriched functions nor directly linked genes for heavy-metal adaptation (see Supp. Data 3). This may reflect a generalist ancestor that speciated through admixture with *D. rufotomentosa* (Paun et al., 2016; Brück, 2024). These prior findings imply a lack of initial heavy-metal genetic adaptation due to plasticity. Alternatively, recent adaptation to soil type might not be detectable by pcadapt. F_{ST} with pcadapt outliers for gene enrichments enhanced the identification of enriched GO terms of interest. However, the number of significant genes supporting GO terms remained notably higher in cUMB and cVIE compared to the other species-pairs. Despite the potential to detect signatures of drift with F_{ST} and pcadapt in cUMB and cVIE, the intersection of these tests still provided the most satisfactory results, featuring a higher number of terms of interest and fewer significant genes. Ultimately, the study may indicate the presence of functions related to ultramafic adaptation, suggesting that the studied species have selected such genes. An exception to these findings is the cRUF species-pair, which may

have undergone an alternative evolutionary trajectory.

Multi-set intersections of pcadapt outliers and combined pcadapt/MKT produced major differences between both analysis. The large proportion of shared genes/functions and high amount of significant overlaps between species-pairs might be indicative of lost statistical power. Although the observed overlaps are significantly higher than expected by chance (Wang et al., 2015), the aforementioned high number of cUMB and cVIE pcadapt outlier windows might inflate statistical results and could lead to a higher number of false-positives. The demonstrative decrease of significant overlaps when applying a F_{ST} "filter" to pcadapt outliers might be testimonial here. Combining pcadapt outlier results with MKT might be contra-intuitive as each genome scan has the highest power in detecting adaptation at different time horizons. However, gene and functional multi-set intersections with or between cUMB and cVIE species-pairs presented the majority of intersections. Despite gene intersections with cUMB/cVIE exhibiting non-significant overlaps they thereby confirmed that gene reuse might decrease with increasing divergence (Bohutínská et al., 2021). To conclude, multi-set intersections identified that gene and functional reuse between species-pairs might be prevalent yet heavily dependent on the gene-set context in question.

The PicMin analysis could reveal putatively repeated adapted genes and functions across the studied species-pairs. Genes and functions identified by PicMin to be significantly repeatedly adapted were again mostly found to be of biotic process origin (as compared to the pcadapt gene enrichment analysis). On the one hand this can be explained due to the usage of pcadapt results for the PicMin analysis, therefore there cannot arise any "surprises" which have not previously been found (Booker et al., 2023). However, on the other hand one would expect a greater number of parallel adapted functional candidates related to heavy-metal soil adaptation. The polygenic nature of this trait (hypothesized due to the continuous range of metal concentrations in soils, (Paun et al., 2016)) combined with the high overlap of functions found in the pcadapt multi-set intersection would suggest to expect more convergent functional candidates. Finally, the PicMin results indicate first possible parallel genes and function adaptation candidates. Nevertheless, a correct analysis of the same ultramafic species as independent lineages might provide results with increased specificity.

Overall, the results presented in this thesis refine our understanding of selective signatures in the NC *Diospyros* radiation and may be relevant to plant adaptive radiations more broadly. Since the MKT can detect adapted genes over long divergence times; it is therefore especially useful to understand early adaptation of specialist ultramafic species (Paun et al., 2016). Thus, it is important for a MKT analysis to take differentiation of the out-group into consideration. Pcadapt in turn is well-suited for more recent events. Yet, pcadapt outliers of highly diverged species pairs should be used carefully and validated by neutrality tests. Furthermore, multi-set intersections of genes and functions and PicMin analyses should use independent lineages to maximize statistical power and reduce bias when using a genome scan such as pcadapt (Bohutínská and Peichel, 2024). In the case of PicMin and the *Diospyros* radiation sister-pairs must be chosen carefully for *D. umbrosa*

and *D. vieillardii* (Fig. 1.1B) to keep both species as independent from each other and to retain enough lineages that PicMin's statistical power does not suffer (Booker et al., 2023). Applying such methodological refinements could help resolve outstanding questions, such as whether *D. rufotomentosa* lacks adapted genes related to metal transport or has evolved via pleiotropic genes affecting both soil adaptation and other traits such as "flower development". Pleiotropic genes in relation to soil adaptation could have played a pivotal role for the adaptation process (Rennison and Peichel, 2022) if *D. rufotomentosa* is in fact a specialist. Pleiotropy has been linked in recent years to parallel adaptation (Lai et al., 2025; Schlötterer, 2023), and may be a key driver in adaptive radiations.

Bibliography

Adrian Alexa, J. R. (2017). topGO.

Barghi, N., Hermisson, J., and Schlötterer, C. (2020). Polygenic adaptation: a unifying framework to understand positive selection. *Nature Reviews Genetics*, 21(12):769–781.

Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., Pinto, B. L., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G. A., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I., Llinares-López, F., Marchler-Bauer, A., Meng-Papaxanthos, L., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Piovesan, D., Rivoire, C., Sigrist, C. J. A., Thanki, N., Thibaud-Nissen, F., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., and Bateman, A. (2025). InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456.

Bohutínská, M. and Peichel, C. L. (2024). Divergence time shapes gene reuse during repeated adaptation. *Trends in Ecology & Evolution*, 39(4):396–407.

Bohutínská, M., Vlček, J., Yair, S., Laenen, B., Konečná, V., Fracassetti, M., Slotte, T., and Kolář, F. (2021). Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its relatives. *Proceedings of the National Academy of Sciences*, 118(21):e2022713118.

Booker, T. R., Yeaman, S., and Whitlock, M. C. (2023). Using genome scans to identify genes used repeatedly for adaptation. *Evolution*, 77(3):801–811.

Brothers, R. N. and Lillie, A. R. (1988). Regional Geology of New Caledonia. In Nairn, A. E. M., Stehli, F. G., and Uyeda, S., editors, *The Ocean Basins and Margins*, pages 325–374. Springer US, Boston, MA.

Brück, F. L. (2024). Population genomics of a selected clade within the New Caledonian Diospyros radiation. Publisher: :none.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829.

Carpenter, R. J., Read, J., and Jaffré, T. (2003). Reproductive traits of tropical rain-forest trees in New Caledonia. *Journal of Tropical Ecology*, 19(4):351–365.

Cerca, J. (2023). Understanding natural selection and similarity: Convergent, parallel and repeated evolution. *Molecular Ecology*, 32(20):5451–5462.

Cerca, J., Cotoras, D. D., Bieker, V. C., De-Kayne, R., Vargas, P., Fernández-Mazuecos, M., López-Delgado, J., White, O., Stervander, M., Geneva, A. J., Guevara Andino, J. E., Meier, J. I., Roeble, L., Brée, B., Patiño, J., Guayasamin, J. M., Torres, M. D. L., Valdebenito, H., Castañeda, M. D. R., Chaves, J. A., Díaz, P. J., Valente, L., Knope, M. L., Price, J. P., Rieseberg, L. H., Baldwin, B. G., Emerson, B. C., Rivas-Torres, G., Gillespie, R., and Martin, M. D. (2023a). Evolutionary genomics of oceanic island radiations. *Trends in Ecology & Evolution*, 38(7):631–642.

Cerca, J., Cotoras, D. D., Santander, C. G., Bieker, V. C., Hutchins, L., Morin-Lagos, J., Prada, C. F., Kennedy, S., Krehenwinkel, H., Rominger, A. J., Meier, J., Dimitrov, D., Struck, T. H., and Gillespie, R. G. (2023b). Multiple paths toward repeated phenotypic evolution in the spiny-leg adaptive radiation (*Tetragnatha* ; Hawai'i). *Molecular Ecology*, 32(18):4971–4985.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and 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 w^{1118} ; iso-2; iso-3. *Fly*, 6(2):80–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Duangjai, S., Samuel, R., Munzinger, J., Forest, F., Wallnöfer, B., Barfuss, M. H., Fischer, G., and Chase, M. W. (2009). A multi-locus plastid phylogenetic analysis of the pantropical genus Diospyros (Ebenaceae), with an emphasis on the radiation and biogeographic origins of the New Caledonian endemic species. *Molecular Phylogenetics and Evolution*, 52(3):602–620.

García-Verdugo, C., Baldwin, B. G., Fay, M. F., and Caujapé-Castells, J. (2014). Life history traits and patterns of diversification in oceanic archipelagos: a meta-analysis: Plant Traits and Speciation on Islands. *Botanical Journal of the Linnean Society*, 174(3):334–348.

Glaubrecht, M. and Schneider, H. (2010). *Evolution in action: case studies in adaptive radiation, speciation and the origin of biodiversity*. Springer, Heidelberg New York.

Grandcolas, P., Murienne, J., Robillard, T., Desutter-Grandcolas, L., Jourdan, H., Guilbert, E., and Deharveng, L. (2008). New Caledonia: a very old Darwinian island? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1508):3309–3317.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L., von Mering, C., and Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314.

Höllinger, I., Pennings, P. S., and Hermisson, J. (2019). Polygenic adaptation: From sweeps to subtle frequency shifts. *PLOS Genetics*, 15(3):e1008035.

Höllinger, I., Wölfl, B., and Hermisson, J. (2023). A theory of oligogenic adaptation of a quantitative trait. *GENETICS*, 225(2):iyad139.

James, M. E., Brodribb, T., Wright, I. J., Rieseberg, L. H., and Ortiz-Barrientos, D. (2023). Replicated Evolution in Plants. *Annual Review of Plant Biology*, 74(1):697–725.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9):1236–1240.

Lai, W.-Y., Hsu, S.-K., Futschik, A., and Schlötterer, C. (2025). Pleiotropy increases parallel selection signatures during adaptation from standing genetic variation. *eLife*, 13:RP102321.

Lee, K. M. and Coop, G. (2017). Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. *Genetics*, 207(4):1591–1619.

Lee, K. M. and Coop, G. (2019). Population genomics perspectives on convergent adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777):20180236.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21):2987–2993.

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4):1022–1034.e6.

Louis, M., Galimberti, M., Archer, F., Berrow, S., Brownlow, A., Fallon, R., Nykänen, M., O'Brien, J., Roberston, K. M., Rosel, P. E., Simon-Bouhet, B., Wegmann, D., Fontaine, M. C., Foote, A. D., and Gaggiotti, O. E. (2021). Selection on ancestral genetic variation fuels repeated ecotype formation in bottlenose dolphins. *Science Advances*, 7(44):eabg1245.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654.

Meier, S. K., Adams, N., Wolf, M., Balkwill, K., Muasya, A. M., Gehring, C. A., Bishop, J. M., and Ingle, R. A. (2018). Comparative RNA-seq analysis of nickel hyperaccumulating and non-accumulating populations of *Senecio coronatus* (Asteraceae). *The Plant Journal*, 95(6):1023–1038. Publisher: Wiley.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, 10:33.

Paun, O., Turner, B., Trucchi, E., Munzinger, J., Chase, M. W., and Samuel, R. (2016). Processes Driving the Adaptive Radiation of a Tropical Tree (*Diospyros* , Ebenaceae) in New Caledonia, a Biodiversity Hotspot. *Systematic Biology*, 65(2):212–227.

Petren, K., Grant, P. R., Grant, B. R., and Keller, L. F. (2005). Comparative landscape genetics and the adaptive radiation of Darwin's finches: the role of peripheral isolation. *Molecular Ecology*, 14(10):2943–2957.

Privé, F., Luu, K., Vilhjálmsson, B. J., and Blum, M. G. B. (2020). Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4. *Molecular Biology and Evolution*, 37(7):2153–2154.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rennison, D. J. and Peichel, C. L. (2022). Pleiotropy facilitates parallel adaptation in sticklebacks. *Molecular Ecology*, 31(5):1476–1486.

Samuel, R., Turner, B., Duangjai, S., Munzinger, J., Paun, O., Barfuss, M. H. J., and Chase, M. W. (2019). Systematics and evolution of the Old World Ebenaceae, a review with emphasis on the large genus *Diospyros* and its radiation in New Caledonia. *Botanical Journal of the Linnean Society*, 189(2):99–114.

Schlötterer, C. (2023). How predictable is adaptation from standing genetic variation? Experimental evolution in *Drosophila* highlights the central role of redundancy and linkage disequilibrium. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1877):20220046.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, 19(4):198–207.

Seehausen, O. (2006). African cichlid fish: a model system in adaptive radiation research. *Proceedings of the Royal Society B: Biological Sciences*, 273(1597):1987–1998.

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, 14(11):751–764.

Stroud, J. T. and Losos, J. B. (2016). Ecological Opportunity and Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, 47(1):507–532.

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

Tan, Y., Barnbrook, M., Wilson, Y., Molnár, A., Bukys, A., and Hudson, A. (2020). Shared Mutations in a Novel Glutaredoxin Repressor of Multicellular Trichome Fate Underlie Parallel Evolution of Antirrhinum Species. *Current Biology*, 30(8):1357–1366.e4.

Tassin, J., Boissenin, M., and Barré, N. (2010). Can *Ptilinopus greyii* (Columbidae) Disperse Seeds in New Caledonia's Dry Forests? *Pacific Science*, 64(4):527–532.

Turner, B., Munzinger, J., Duangjai, S., Temsch, E. M., Stockenhuber, R., Barfuss, M. H., Chase, M. W., and Samuel, R. (2013a). Molecular phylogenetics of New Caledonian Diospyros (Ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution*, 69(3):740–763.

Turner, B., Paun, O., Munzinger, J., Chase, M. W., and Samuel, R. (2016). Sequencing of whole plastid genomes and nuclear ribosomal DNA of *Diospyros* species (Ebenaceae) endemic to New Caledonia: many species, little divergence. *Annals of Botany*, 117(7):1175–1185.

Turner, B., Paun, O., Munzinger, J., Duangjai, S., Chase, M. W., and Samuel, R. (2013b). Analyses of amplified fragment length polymorphisms (AFLP) indicate rapid radiation of Diospyros species (Ebenaceae) endemic to New Caledonia. *BMC Evolutionary Biology*, 13(1):269.

Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports*, 5(1). Publisher: Springer Science and Business Media LLC.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358.

Whiting, J. R., Booker, T. R., Rougeux, C., Lind, B. M., Singh, P., Lu, M., Huang, K., Whitlock, M. C., Aitken, S. N., Andrew, R. L., Borevitz, J. O., Bruhl, J. J., Collins, T. L., Fischer, M. C., Hodgins, K. A., Holliday, J. A., Ingvarsson, P. K., Janes, J. K.,

Khandaker, M., Koenig, D., Kreiner, J. M., Kremer, A., Lascoux, M., Leroy, T., Milesi, P., Murray, K. D., Pyhäjärvi, T., Rellstab, C., Rieseberg, L. H., Roux, F., Stinchcombe, J. R., Telford, I. R. H., Todesco, M., Tyrmi, J. S., Wang, B., Weigel, D., Willi, Y., Wright, S. I., Zhou, L., and Yeaman, S. (2024). The genetic architecture of repeated local adaptation to climate in distantly related plants. *Nature Ecology & Evolution*, 8(10):1933–1947.

Wulff, A. S., Hollingsworth, P. M., Ahrends, A., Jaffré, T., Veillon, J.-M., L'Huillier, L., and Fogliani, B. (2013). Conservation Priorities in a Biodiversity Hotspot: Analysis of Narrow Endemic Plant Species in New Caledonia. *PLoS ONE*, 8(9):e73371.

Zechmeister, J. (2024). Whole genome re-sequencing unravels complex speciation processes within the RRc clade of the radiated Diospyros in New Caledonia. Publisher: :none.

Acronyms

π_{XY} pi between. 9, 10

π_{X} pi within. 9, 10

BED Browser Extensible Format. 8–10

CDS Coding Sequences. 6

cERU species-pair *D. calciphila* - *D. erudita*. 13, 14, 19, 20

cRUF species-pair *D. calciphila* - *D. rufotomentosa*. 13, 14, 19, 20

cSPN species-pair *D. calciphila* - *D. sp. PicN'ga*. 5, 13, 14, 19, 20

cUMB species-pair *D. calciphila* - *D. umbrosa*. 13, 14, 19, 20

cVIE species-pair *D. calciphila* - *D. vieillardii*. 13, 14, 19, 20

ETR3 Ethylene receptor 3. 28

FDR False Discovery Rate. 7, 10, 19, 28, 30

GFF General Feature Format. 8

GO Gene Ontology. 9, 10, 19, 20, 23, 26, 28, 30, 32

HASTY 1 Protein HASTY 1. 28

HMA heavy metal ATPase. 32

HMA3 Cadmium/zinc-transporting ATPase HMA3. 28

MAF Minor Allele Frequency. 6

MKT McDonald-Kreitman Test. vii, ix, 7–10, 14, 16, 19, 20, 23, 25–27, 31–33

Mya Million years ago. 1, 2

NC New Caledonia. 1–4, 14

Acronyms

PCA Principal Component Analysis. 7

RAD Resctriction-enzyme associated DNA. 14, 31, 32

SGV Standing Genetic Variation. 4

SNP Single Nucleotide Polymorphism. 5–7, 13

spp. species pluralis. 2

TE Transposable Element. 5

VCF Variant Call Format. 5–7

VIP2 Inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase VIP2. 28

WGS Whole Genome Sequencing. 31

YSL9 Probable metal-nicotianamine transporter YSL9. 20

A. Appendix

A.1. Supplementary Data

Supplementary Data 1. *Snakemake workflow and scripts used in this thesis are available on* https://github.com/flotji-potji/dioresel

Supplementary Data 2. *BED-file of annotated gene accessions identified by Paun et al. (2016). Available on* https://github.com/flotji-potji/dioresel/blob/df862dea1a12c326e8c187f2476235c648cd6b94/supp_data/genes_of_interest.bed

Supplementary Data 3. *Gene enrichments from Section 3.3. Available on* https://github.com/flotji-potji/dioresel/blob/df862dea1a12c326e8c187f2476235c648cd6b94/supp_data/enrichedGOterms.xlsx

Supplementary Data 4. *Gene and enriched functions of the Picmin analysis. Available on* https://github.com/flotji-potji/dioresel/blob/df862dea1a12c326e8c187f2476235c648cd6b94/supp_data/picminGenesFunctions.xlsx

A.2. Supplementary Figures

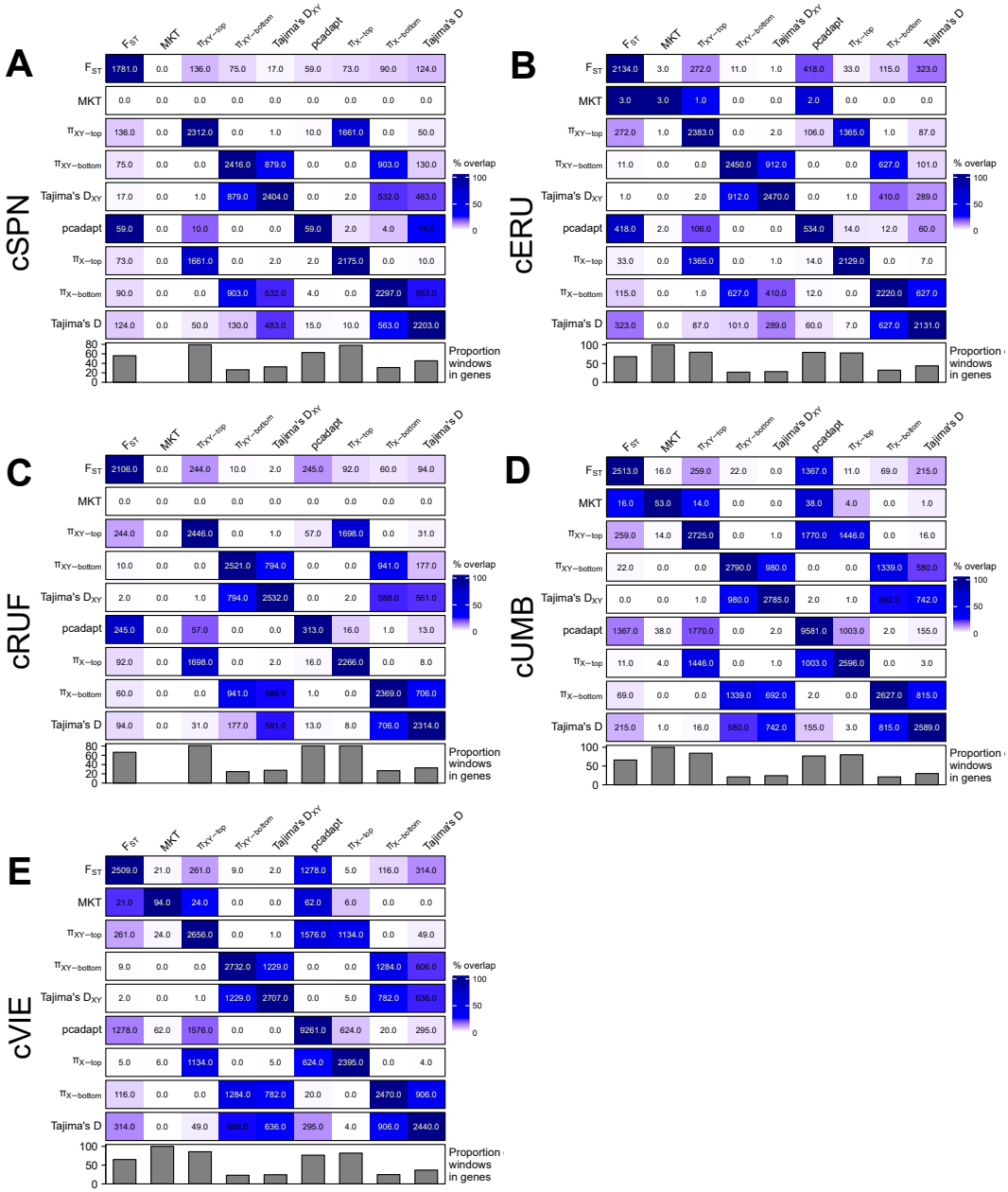

Figure A.1.: Test intersections of species-pairs (A) cSPN (B) cERU (C) cRUF (D) cUMB and (E) cVIE. For detailed description see Fig. 3.4B.

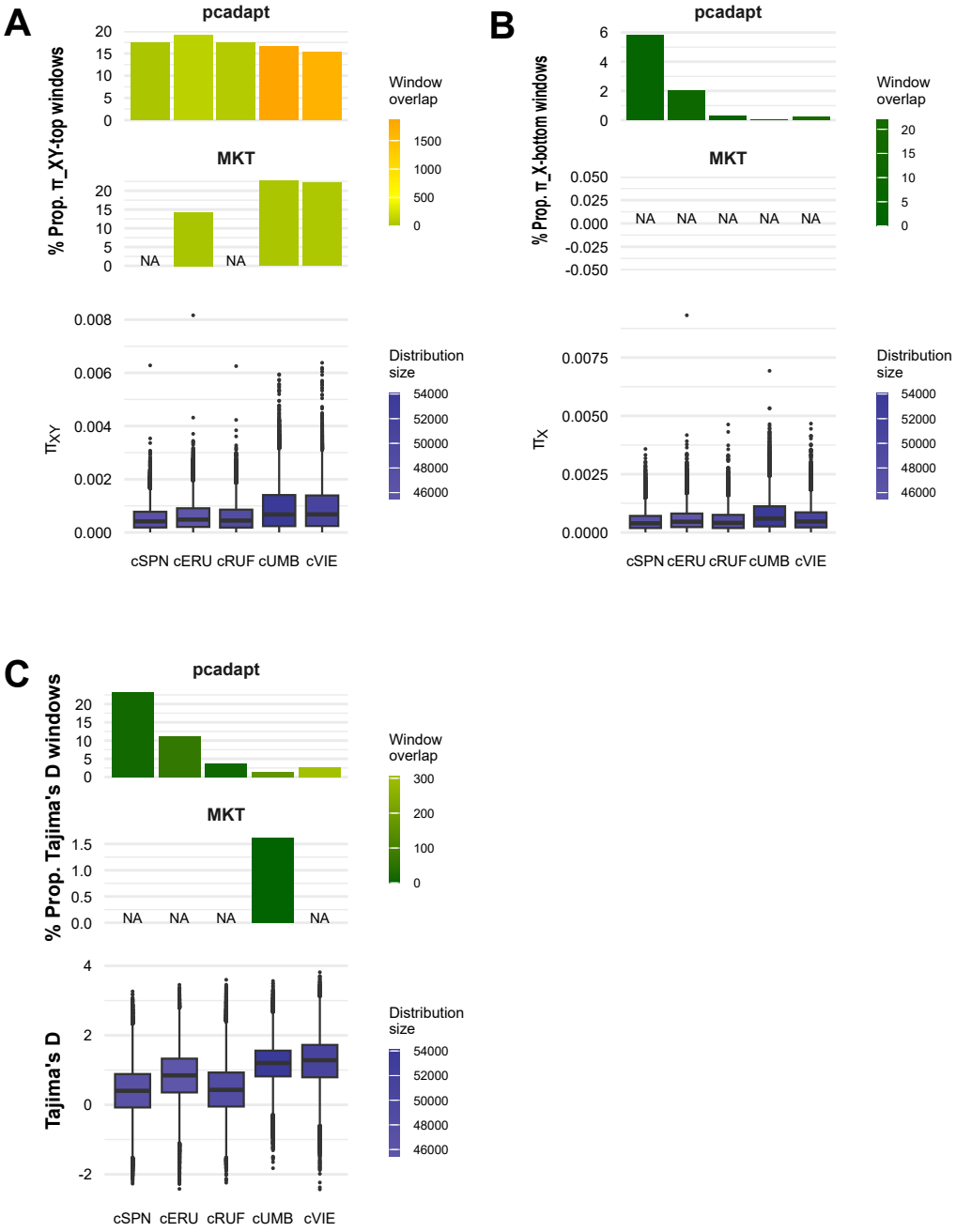

Figure A.2.: Intersections of pcadapt/MKT with test distribution of (A) π_{XY} (B) π_X and (C) Tajima's D. for detailed description see Fig. 3.4D.

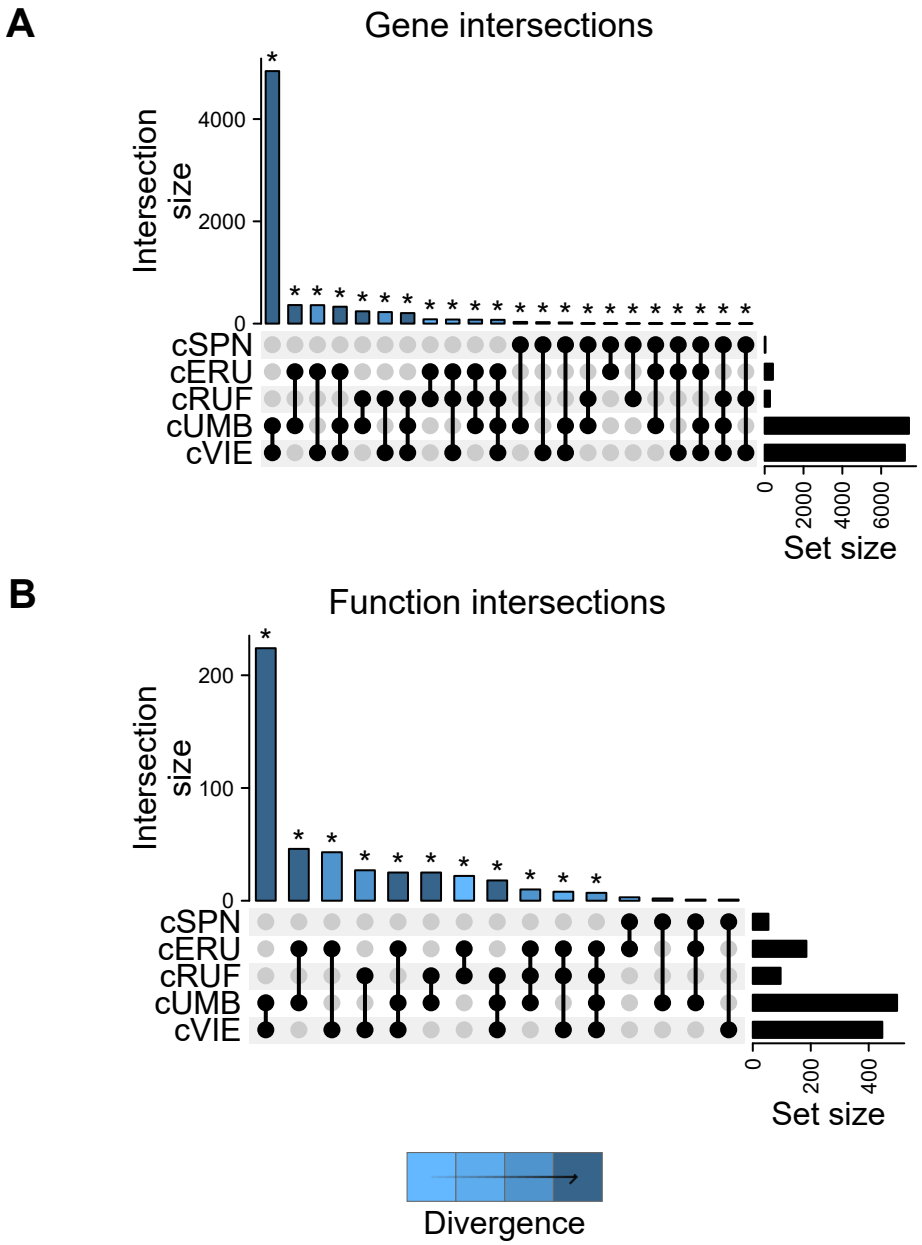

Figure A.3.: Gene and function multi-set intersections of (A-B) pcadapt outliers and (C-D) F_{ST} intersected with pcadapt. For detailed description see Fig. 3.9.
Continued figure on next page.

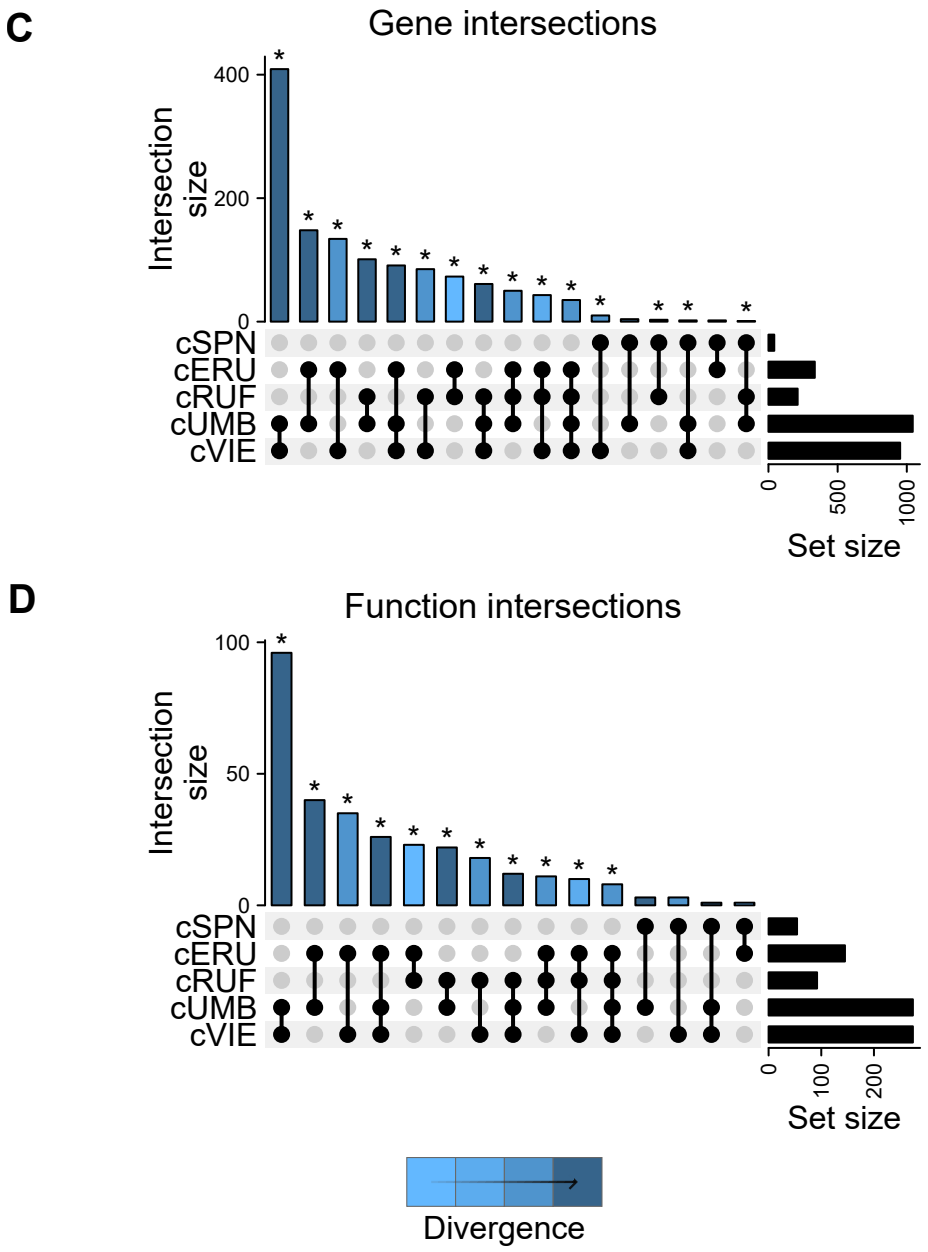

Figure A.3.: Multi-set intersection of pcadapt and Fst/pcadapt

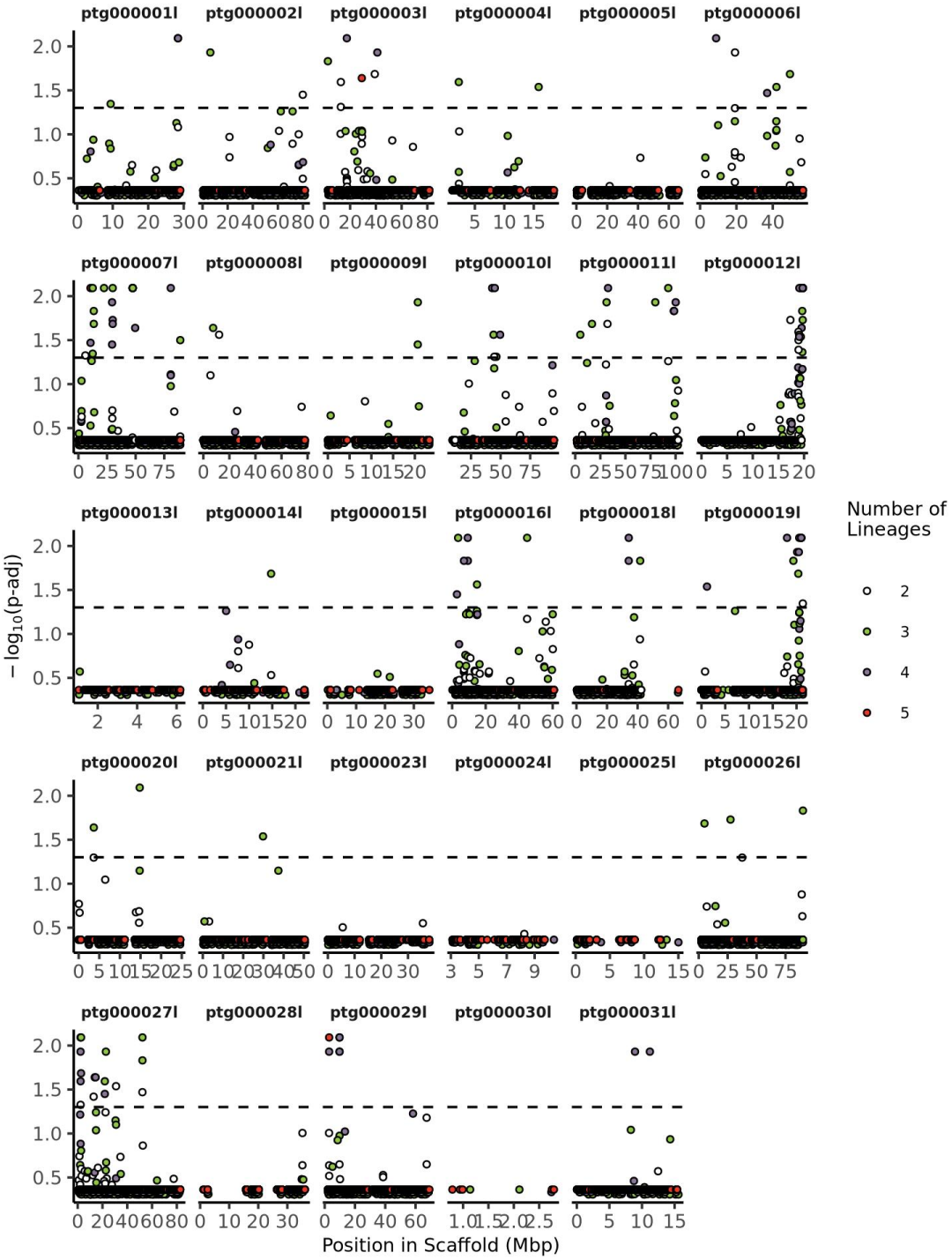

Figure A.4.: Picmin window results separated into each scaffold. Number of lineages means in how many lineages was this particular window significantly - classified by Picmin - repeatedly adapted.

Figure A.5.: Gene enrichment results of the Picmin analysis. For detailed description see Fig. 3.8.