Annotation of complex genomes for comparative genomics

by

Kristina Kirilova Gagalova

B.Sc, University of Bologna, 2010
M.Sc, University of Bologna, 2013
M.Sc, Vrije University of Amsterdam, 2015

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in

The Faculty of Graduate and Postdoctoral Studies
(Bioinformatics)
The University of British Columbia
(Vancouver)

April 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Annotation of complex genomes for comparative genomics

submitted by Kristina Kirilova Gagalova in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Bioinformatics

Examining Committee:

Prof. Inanc Birol, Medical Genetics, University of British Columbia
Supervisor

Prof. Steven J.M. Jones, Medical Genetics, University of British Columbia
Supervisory Committee Member

Dr. Mathias Schuetz, Botany, University of British Columbia
Supervisory Committee Member

Prof. Naomi M. Fast, Botany, University of British Columbia
University Examiner

Prof. Michael Deyholos, Biology, University of British Columbia
University Examiner

Prof. Dave Edwards, School of Biological Sciences, The University of Western Australia
External Examiner

Additional Supervisory Committee Members:

Prof. Sean Graham, Botany, University of British Columbia
Supervisory Committee Member

Prog. Joerg Bohlmann, Michael Smith Laboratories, University of British Columbia
Supervisory Committee Member

Abstract

Advancements in whole-genome sequencing technologies have opened the use of genomic approaches to study a variety of organisms and allowed studies at the whole-genome scale in non-model organisms. In these studies, genome annotation is a fundamental step to extract diverse biological information from sequences that are otherwise strings of characters incomprehensible to humans.

Here I assembled and annotated genomes of plant and insect species of applied interest. A common theme in my thesis is comparative and evolutionary genomics of the described organisms. The sequenced species I studied have complex genomic features, including large genome sizes and high repeat contents, which I described in detail.

In Chapter 2, I investigate the protein-coding genes of four spruces (*Picea*, Pinaceae) native to North America. Comparison to other annotated conifers highlights changes in selection in gene families. Several gene families have a significantly expanded number of genes. Some genes are under positive selection: previous studies in spruce highlighted the same proteins as genetic markers for local adaptation. In Chapter 3, I characterize the genome of *Pissodes strobi*, a naturally occurring pest of the spruces described in Chapter 2. The genome of *P. strobi* is larger and more repetitive than other sequenced species in the same family (Curculionidae). In Chapter 4, I assemble and annotate the genome of a proprietary *Cannabis sativa* strain, and study the flavonoid/anthocyanin metabolic pathway, uncovering the upregulation of key metabolic genes involved in the regulation of leaf pigmentation.

The presented genome annotations and comparative analyses provide insights into the biology and evolution of the described species. Comparative genome studies are important for generating hypotheses and open avenues of inquiry in future studies in population genomics. In the case of *Picea* gen. and *P. strobi*, such studies will enable us to understand the local adaptation of species and the genetic basis of regulatory processes, such as biotic stress mitigation and pest resistance.

Lay summary

With recent technical advances, it is possible to sequence the genomes of a variety of species, including those with large and complex genomes. The increasingly large number of sequenced genomes will help us understand the genomes' evolution and function. Here I extract biological information from six newly assembled plant and insect genomes. I use a comparative approach to study their evolution to identify and characterize the genes involved in biological processes and pathways. I performed a genomic comparison of four North American spruces and their natural pest, the spruce weevil, looking for genomic features that help understand their adaptation to the environment. I also characterized the genome of a cannabis variety to study its pigmentation process and find the genes associated with dark color in some strains.

Preface

I performed the work presented within this thesis, except for **Chapter 4**, at Canada's Michael Smith Genome Sciences Centre, as part of the interdisciplinary Ph.D. program in Bioinformatics, at the University of British Columbia. I performed the work presented in **Chapter 4** at Willow Biosciences Inc., in collaboration with Canada's Michael Smith Genome Sciences Centre and the Bioinformatics Technology Lab. **Chapters 2** and **3** contain content from publications in which I am a first or co-first author, and they are published or currently under revision with peer-reviewed journals. **Chapters 1** and **5** (introduction and discussion chapters of the thesis) are my work and are not published. Details of my involvement and contribution to each research-based chapter are provided below.

Chapter 2: This chapter is a version of a submitted manuscript: "Spruce giga-genomes: structurally similar despite rapidly evolving genes and positive selection that offers clues into local adaptation" by myself, Kristina K. Gagalova (first author), René L. Warren (RLW) (Genome Sciences Centre), Lauren Coombe (LC) (Genome Sciences Centre), Johnathan Wong (JW) (Genome Sciences Centre), Ka Ming Nip (KMN) (Genome Sciences Centre), Macaire Man Saint Yuen (MMSY) (UBC, Michael Smith Laboratories), Justin G.A. Whitehill (JGAW) (NC State University, Department of Forestry and Environmental Resources), Jose M. Celedon (JMC) (UBC, Michael Smith Laboratories), Carol Ritland (CR) (UBC, Michael Smith Laboratories), Greg A. Taylor (GAT) (Genome Sciences Centre), Dean Cheng (DC) (Genome Sciences Centre), Patrick Plettner (PP) (Genome Sciences Centre), S. Austin Hammond (SAH) (Genome Sciences Centre), Hamid Mohamadi (HM) (Genome Sciences Centre), Yongjun Zhao (YZ) (Genome Sciences Centre), Richard A. Moore (RAM) (Genome Sciences Centre), Andrew J. Mungall (AJM) (Genome Sciences Centre), Brian Boyle (BB) (Université Laval, Institute for Systems and Integrative Biology), Jérôme Laroche (JL) (Université Laval, Institute for Systems and Integrative Biology), Joan Cottrell (JC) (Forest Research, U.K. Forestry Commission), John J. Mackay (JJM) (University of Oxford, Department of Plant Sciences), Manuel Lamothe (ML) (Laurentian Forestry Centre, Natural Resource Canada), Nathalie Isabel (NI) (Laurentian Forestry Centre, Natural Resource

Canada), Nathalie Pavy (NP) (Université Laval, Canada Research Chair in Forest Genomics), Steven J.M. Jones (SJMJ) (Genome Sciences Centre), Joerg Bohlmann (JoB) (UBC, Michael Smith Laboratories), Jean Bousquet (JeB) (Université Laval, Canada Research Chair in Forest Genomics), and Inanc Birol (IB) (Genome Sciences Centre). LC and SAH did the repeat annotation and its downstream analysis. LC performed mitochondrial and chloroplast phylogeny. Additional data curation was performed by GAT, DC, PP, YZ, RAM, AJM, BB, JL, and JC. IB, JeB, and JoB revised the manuscript. HM performed the analysis of spruce genome sizes based on k-mer histograms.

I drafted the manuscript together with RLW and NP and collaborated in rounds of editing and revision. My other specific contributions to the work include the following:

- Performed the genome annotations
- *De novo* assembled RNAseq data used for the genome annotations and coordinated with KMN, JAT, MMSY, JGAW, JMC, ML, NI, and CR on genome annotations and resource curation
- Performed the phylogeny of protein-coding genes
- Analyzed gene families, gene gain-loss, and positive selection in orthologs
- Updated phylogeny described in the manuscript to provide resampling information and to include an additional phylogenetic tree in the analysis based on single-copy genes and Astral-III
- Updated gene family expansion/contraction analysis based on the phylogeny I derived from Astral-III

Chapter 3: The work described in this chapter is part of a manuscript accepted for publication in *G3 Genes|Genomes|Genetics* (doi: 10.1093/g3journal/jkac038): "The genome of the forest insect pest, *Pissodes strobi,* reveals genome expansion and evidence of a *Wolbachia* endosymbiont" by Kristina K. Gagalova (co-first author with JGAW), Justin G.A. Whitehill (JGAW) (NC State University, Department of Forestry and Environmental Resources), Luka Culibrk (LCu) (Genome Sciences Centre), Diana Lin (DL) (Genome Sciences Centre), Véronique Lévesque-Tremblay (VLT) (Laurentian Forestry Centre, Natural Resource Canada), Christopher I. Keeling (CIK) (Laurentian Forestry Centre, Natural Resource Canada), Lauren Coombe (LCo) (Genome Sciences Centre), Macaire M.S. Yuen (MMSY) (UBC,

Michael Smith Laboratories), Inanc Birol (IB) (Genome Sciences Centre), Joerg Bohlmann (JoB) (UBC, Michael Smith Laboratories), and Steven J.M. Jones (SJMJ) (Genome Sciences Centre). CIK and VLT performed the flow cytometry for the genome size determination. DL performed the mitochondrial genome assembly and annotation.

The publication is a collaboration between me and JGAW, under the supervision of SJMJ and JoB. I wrote the first draft of the manuscript and revised it. My other specific contributions to the work include the following:

- Performed the genome assembly in collaboration with LCu and LCo. I optimized the SuperNova and added the Purge Haplotigs stage to the assembly
- Performed *in silico* genome complexity estimate
- Performed genome annotations
- *De novo* assembled RNAseq used for the genome annotations, with help from MMSY
- Inferred phylogeny
- Characterized repeats

Chapter 4: The work in this chapter is planned for publication. I led the work in this chapter on concept formulation and data analysis. I led the writing with Dr. Mathias Schuetz (Adjunct Professor, Department of Botany, UBC, and Vice President of Plant Science at Willow Biosciences Inc. to August 2021). Dr. Matt Workentine (Willow Biosciences Inc.) supervised genome assembly and genome annotations. Dr. Shumin Wang (Willow Biosciences Inc.) performed the *C. sativa* sample preparation for RNAseq quantification and metabolomics profiling. Mrs. Yifan Yan (UBC, Wine Research Centre) and Associate Prof. Simone Castellarin (UBC, Wine Research Centre) performed the anthocyanin and flavonoid metabolite analysis.

My other specific contributions to the work include the following:

- Performed the genome assembly of Willow-alpha in collaboration with Dr. Matt Workentine. I performed the Racon, wtdbg2, ntJoin, and Sealer assembly steps
- Performed genome annotations

- Performed phylogeny
- Recovered phenylpropanoid/flavonoid/anthocyanin pathway genes
- Performed differential gene expression

Table of contents

Abstract…….. iii

Lay summary……...iv

Preface……...v

Table of contents ……ix

List of Tables ……xiv

List of Figures...……xvii

Abbreviations ……xix

Glossary ……xxi

Acknowledgments ...……xxiv

Dedication ……xxvi

Chapter 1: Introduction……..1

1.1 Motivation……..……1

1.2 Genome complexity in genomes of non-model species ……2

1.2.1 Gene structure and gene families…….2

1.2.2 Genome repetitiveness…….4

1.2.3 Genome size……5

1.3 Genome annotation.……...6

1.3.1 Structural genome annotation.……7

1.3.1.1 Protein-coding gene annotation.……7

1.3.1.2 Repeat annotation.……..9

1.3.2 Functional gene annotation.……...……..9

1.3.3 Benchmarking gene annotations.……...……10

1.4 Research objectives.……...……12

Chapter 2: Comparative genome annotation of four North American spruces (*Picea*, Pinaceae)….13

2.1 Author summary……..…13

2.2 Introduction…….13

2.3 Methods……...17

2.3.1 Sample collection, sequencing, and genome assembly……..17

2.3.2 Estimate of genome sizes……...……17

2.3.3 Annotation of protein-coding genes……...……18

2.3.4 Comparative genomics: phylogeny of gen. *Picea*……...……..20

2.3.5 Gene families and gene gain/loss ……..……...…….22

2.3.6 Positive selection acting on orthologs……..……..23

2.3.7 GO term enrichment analysis...……..……25

2.4 Results……..……...25

2.4.1 Genome assemblies……...……..25

2.4.2 Genome annotations……..…….25

2.4.3 Phylogenomics of gen. *Picea*……..……...26

2.4.4 Gene gain/loss in gene families……..……...28

2.4.5 Positive selection in orthologs……..……....31

2.4.6 GO term enrichment analysis……..…….....32

2.5 Discussion……..…….....39

2.5.1 Genome annotation of spruce giga-genomes …….39

2.5.2 Phylogeny of gen. *Picea*……...…….41

2.5.3 Protein domains of expanded gene families and genes under positive selection……..42

2.5.4 Future directions……43

Chapter 3: Genome assembly and annotations of *Pissodes strobi*, a North American forest insect…... pest……...45

3.1 Author summary……..45

3.2 Introduction…….45

3.3 Methods……...47

3.3.1 Sample collection and sequencing…….47

3.3.2 *In silico* genome complexity estimate……47

3.3.3 Experimental genome size estimate by flow cytometry……48

3.3.4 Genome assembly……..48

3.3.5 Annotation of protein-coding genes……...49

3.3.6 Annotation and quantification of repeat elements…….50

3.3.7 Comparative genomics: Curculionidae phylogenomics……51

3.4 Results…….52

3.4.1 Genome complexity and estimated genome size of *P. strobi*……52

3.4.2 Genome assembly……..53

3.4.3 Protein-coding gene annotation…….54

3.4.4 Inference of Curculionidae phylogeny from genomic data..……..55

3.4.5 Genomic repeats annotation……...57

3.4.5.1 Repeat annotation and quantification……...57

3.4.5.2 Repeats turnover during weevil (Curculionidae) evolution…….59

3.5 Discussion……61

3.5.1 Genome complexity of *P. strobi* ……...…..61

3.5.2 Annotation of *P. strobi* repeats from genome assembly and unassembled reads……..61

3.5.3 Phylogeny and host preference of Curculionidae……..62

3.5.4 *Wolbachia* putative endosymbiont of *P. strobi*……..63

3.5.5 The case of the *Elaeidobius kamerunicus* genome ……...…..63

3.5.6 Future directions……..……..64

Chapter 4: Genome assembly and annotation of Willow-alpha, a *Cannabis sativa* variety, with a focus on anthocyanin biosynthesis……..……..65

4.1 Author summary……..……65

4.2 Introduction……..……...65

4.2.1 Evolution, domestication, and chemotype classification of *Cannabis sativa*…...…….65

4.2.2 *Cannabis sativa* secondary metabolites: flavonoids...……...……...…….66

4.2.3 *Cannabis sativa* genome complexity and available genome assemblies……...……69

4.3 Methods……...……70

4.3.1 Genome assembly……...……...……70

4.3.2 Genome annotation……...…….……72

4.3.3 Comparative genomics: phylogenomics of *C. sativa* varieties.…….74

4.3.4 Correlative transcriptome and metabolomic of Willow-alpha and three other varieties…...74

4.3.5 Identification of genes involved in the general phenylpropanoid, flavonoid, anthocyanin, and catechin pathways……76

4.4 Results…..……...……..…77

4.4.1 Genome assembly…..……...……...…77

4.4.2 Genome annotation ..…..……...……..80

4.4.3 Estimation of phylogenetic relationship among *C. sativa* varieties……...81

4.4.4 *Cannabis sativa* leaf pigmentation…….82

4.4.5 General phenylpropanoid, flavonoids, and anthocyanin biosynthesis genes in *C. sativa*….83

4.4.6 Flavonoid/anthocyanin gene expression in Willow-alpha and three *C. sativa* varieties…...85

4.4.7 Flavonoid/anthocyanin metabolite profiling in Willow-alpha and three *C. sativa* varieties.87

4.5 Discussion……89

4.5.1 Genome assembly and annotation of Willow-alpha *C. sativa* variety……...89

4.5.2 General phenylpropanoids and flavonoids biosynthesis genes in *C. sativa*……..90

4.5.3 Future directions……92

Chapter 5: Conclusion……...94

5.1 Comparative study of annotated genomes……95

5.2 Pangenomes: joint analysis of gene annotations for comparative genomics……97

5.3 Integration of genome, transcriptome, and metabolome for *C. sativa* breeding……..……..98

5.4 Limitations in quality benchmarking tools…….....……..99

5.5 Comparison of current pipelines for genome annotation and opportunities in the field……...99

5.6 Inferring function of hypothetical genes..……101

5.7 Impact of the annotated genomes and future directions……..……102

Bibliography……104

Appendices……141

Appendix A Chapter 2 - Supplemental Material……..141

Appendix B Chapter 3 - Supplemental Material……..169

Appendix C Chapter 4 - Supplemental Material……..179

List of Tables

Table 2.1 Spruce genome annotation statistics: number of annotated genes and transcripts, median mRNA, exon and intron length, and corresponding annotation completeness...……...……..…...26

Table 2.2 Top GO molecular functions and biological processes, enriched in rapidly evolving gene families in the North American spruce species……...35

Table 2.3 Top GO molecular functions and biological processes, enriched in orthologs with positive selection in the North American spruce species……38

Table 3.1 *Pissodes strobi* genome assembly statistics for each assembly step……54

Table 3.2 Gene annotation statistics for "total annotated" and "high confidence" datasets in *P. strobi*……55

Table 4.1 Willow-alpha assembly statistics at each assembly step…….....78

Table 4.2 Number of annotated genes and transcript, and gene length statistics for Willow-alpha, comparing MAKER and BRAKER pipelines…….81

Table A.1 Geographical origin of the collected representative spruce accessions used for DNA extraction……..……..…141

Table A.2 Sequencing reads and corresponding fold coverage (millions of generated reads) for each genome assembly……...142

Table A.3 RNAseq samples used for genome annotation and scaffoldsfiltering…….…….144

Table A.4 Conifer genomes used for phylogenomics comparison…….....…145

Table A.5 Genome assembly statistics and gene completeness for *P. engelmannii*, *P. sitchensis*, *P. glauca,* and interior spruce…….……..…....147

Table A.6 Genome assembly statistics and gene completeness for the conifer species used in the comparative study with the North American spruces: *P. abies*, *P. lambertiana*, and *P. taeda*…….....148

Table A.7 Genome annotation completeness for the *P. abies*, *P. lambertiana* and *P. taeda*…….....149

Table A.8 Number of genes in orthogroups (OGs) and number of orthogroups per species……..….150

Table A.9 CAFE analysis for gene families contraction and expansion for each genotype in the analysis……...152

Table A.10 Top significant GO biological process terms of expanded gene families in *P. engelmannii*…153

Table A.11 Top significant GO biological process terms of expanded gene families in *P. sitchensis*….155

Table A.12 Top significant GO biological process terms of expanded gene families in *P. glauca*…….156

Table A.13 Top significant GO molecular function terms of expanded gene families in interior spruce…157

Table A.14 Top significant GO molecular function terms of expanded gene families in *P. engelmannii*……..……..……..……..……..……..……..……..……..…….....158

Table A.15 Top significant GO molecular function terms of expanded gene families in *P. sitchensis*.....159

Table A.16 Top significant GO molecular function terms of expanded gene families in *P. glauca*…….....160

Table A.17 Top significant GO biological process terms of expanded gene families in interior spruce….161

Table A.18 Top significant GO biological process terms of genes under positive selection in *P. engelmannii*……..…….....……..……..……..……..……..……..……..……..……..…..….162

Table A.19 Top significant GO biological process terms of genes under positive selection in *P. sitchensis*……...……...……...……...……...……...……...……...……...……...……...…..….163

Table A.20 Top significant GO biological process terms of genes under positive selection in *P. glauca*……..……...……...……...……...……...……...……...……...……...……...……...…….164

Table A.21 Top significant GO biological process terms of genes under positive selection in interior spruce……..……...……...……...……...……...……...……...……...……...……...……...……165

Table A.22 Top significant GO molecular function terms of genes under positive selection in *P. engelmannii*……...……...……...……...……...……...……...……...……...……...……...……...166

Table A.23 Top significant GO molecular function terms of genes under positive selection in *P. sitchensis*……...……...……...……...……...……...……...……...……...……...……...……...167

Table A.24 Top significant GO molecular function terms of genes under positive selection in *P. glauca*……...……...……...……...……...……...……...……...……...……...……...……....168

Table B.1 Supporting evidence in MAKER protein-coding gene annotation in *Pissodes strobi*.....169

Table B.2 Curculionidae genomes and short reads accession IDs used in the comparative analysis.....170

Table B.3 Number of genes supporting each topology node in the Curculionidae species tree.....173

Table B.4 Total percent and breakdown of genomic repeats as annotated by EDTA repeat pipeline in Curculionidae genomes…...…..……...……..…...…….....174

Table B.5 RepeatExplorer comparative analysis results for Curculionidae species…...…….....175

Table C.1 Chromosome scale *C. sativa* genomes and hops outgroup used for phylogenomic inference…180

Table C.2 Genome assembly statistics and gene completeness of Willow-alpha, Cannbio-2, CBDRx cs10, Purple Kush, JL wild accession, Finola and Cascade hops genomes……..181

Table C.3 Phenylpropanoid/flavonoid/anthocyanin biosynthetic genes identified in Willow-alpha and their annotation in the TAIR *A. thaliana* database…….....…...….....182

Table C.4 - Differential gene expression (DGE) of general phenylpropanoid, flavonoid, anthocyanins, and catechins biosynthetic pathway genes….....…...….....183

Table C.5 - Anthocyanin quantification (average and standard deviation for biological samples in triplicate) for Willow-alpha and the three high pigmentation strains, as shown in Figure 4.7.....184

Table C.6 - Flavonoid quantification (average and standard deviation for biological samples in triplicate) for Willow-alpha, two high and one medium intensity pigmented variety….....186

Table C.7 - Anthocyanin and flavonoid mass-spec data from Trap and QTOF methods, together with the published exact mass.....187

List of Figures

Figure 2.1 (a) Geographical distributions of Engelmann (*P. engelmannii*), Sitka (*P. sitchensis*), and white spruce (*P. glauca*) and the location of the sampled trees. (b) Spruce dendrometric attributes……16

Figure 2.2 Genome-based inference of spruce phylogenies from (a) nuclear data based on a coalescent-based analysis of single-copy genes, (b) k-mer based estimate using mitochondrial genome, (c) k-mer based estimate using plastid, and (d) maximum-likelihood estimate using plastid genome…..…28

Figure 2.3 Gene family gain-loss computed by CAFE analysis. (a) ultrametric species tree (chronogram) for North American spruces and spruce - pines outgroup, and rapidly evolving gene families on tree nodes. (b) Heatmap of rapidly evolving gene families (y-axis) showing the number of genes in each gene family……...31

Figure 2.4 Histogram of ω ratios and dN, dS scores calculated for orthologs in the four North American spruces.…….33

Figure 2.5 Significant Pfam domains in rapidly evolving gene families……37

Figure 2.6 Significant Pfam domains in orthologous genes with positive selection…….40

Figure 3.1 Ploidy level, k-mer profiles, and genomic features of *P. strobi* genome (a) Smudge plot profiles for heterozygous k-mers in *P. strobi* (b) k-mer histograms and genomic features of *P. strobi*…….…….53

Figure 3.2 Curculionidae species tree inferred from BUSCO single-copy genes……...56

Figure 3.3 Comparative repeat analysis of unassembled genomic reads of eight Curculionidae accessions and *Tribolium castaneum*…….…58

Figure 3.4 Sequence divergence distribution for transposable elements and corresponding genome sizes for Curculionidae……..60

Figure 4.1 Classification and examples of flavonoid metabolites and their chemical structure……...….67

Figure 4.2 Jupiter plot of Willow-alpha genome assembly and cs10 reference genome……..79

Figure 4.3 Species tree derived from phylogenomic inference of *C. sativa* strains and the Cascade variety of hops (*Humulus lupulus*).....82

Figure 4.4 Rachis, petiole, and leaf pigmentation phenotypes of Willow-alpha (green phenotype), CA19210 and CK1926 (dark purple phenotype), and Cali Kush (intermediate pigmentation phenotype)…83

Figure 4.5 General phenylpropanoid, flavonoid, anthocyanin, and catechin biosynthesis pathways, showing the synthesized compounds in squares and the corresponding enzymes on the arrows…….85

Figure 4.6 Gene expression of phenylpropanoid, flavonoid, and anthocyanin enzymes in leaf for Willow-alpha, CA19210, and CK19206 (dark purple phenotype), and Cali Kush (medium pigmentation) varieties.....87

Figure 4.7 Anthocyanin profiling of leaf samples from Willow-alpha and three anthocyanin-producing varieties….....89

Figure A.1 Scaffold filtering and genome annotation pipeline.....143

Figure A.2 K-mer histograms and genome sizes estimates for interior, *P. glauca*, *P. sitchensis* and *P. engelmannii.*.....146

Figure A.3 Orthogroups overlap between spruce and pine taxa.....151

Figure B.1 GenomeScope2.0 estimates for *P. strobi* for k-mer sizes 21, 23, 25, 27, and 29 bp.....171

Figure B.2 DiscoVista quartet relative frequencies for Curculionidae phylogeny tree……..172

Figure B.3 Repeat composition of clusters generated by RepeatExplorer run for *P. strobi* as individual species……176

Figure B.4 Repeat composition of clusters generated by RepeatExplorer comparative analysis for Curcilionidae species…..……177

Figure B.5 GenomeScope2.0 genomic features estimates for *E. kamerunicus*…….….178

Figure C.1 Willow-alpha genome assembly steps and read datasets.……...179

Figure C.2 Liquid chromatography ultraviolet (LC-UV) chromatogram peaks for the four analyzed varieties..…….…….……..185

Abbreviations

aa: Amino acids

AED: Annotation edit distance

eAED: Exon annotation edit distance

bp: Base pairs

cDNA: Complementary DNA

Gb: Billions of base pairs

BP: Before present

BUSCO: Benchmarking universal single-copy orthologs

CAFE: Computational analysis of gene family evolution

CBDA: Cannabidiolic acid

CDS: Coding sequence

GHMM: Generalized hidden Markov model

ILS: Incomplete lineage sorting

HPLC-MS: High-performance liquid chromatography - Mass Spectrometry

kb: Thousands of base pairs

LTR: Long terminal repeat

LRT: Likelihood-ratio test

Mb: Millions of base pairs

ML: Maximum likelihood

MSA: Multiple sequence alignment

MYA: Millions of years ago

NCBI: National center for biotechnology information

RBH: Reciprocal best bit

ROS: Reactive oxidative species

PE: Paired-end reads

PK: Purple Kush

SNV: Single Nucleotide Variant

TE: Transposable element

TR: Tandem repeats

THCA: Tetrahydrocannabinolic acid

UTR: Untranslated region

UV: Ultraviolet

Glossary

***ab initio* and homology-based annotations:** genome annotation approaches; *ab initio* performs the prediction based on statistical models using the DNA sequence motifs, while the homology-based relies on the sequence similarity from closely related species

AED and eAED: measurements on how well the annotations agree with the provided supporting evidence, with 1 denoting total lack of evidence support. eAED score, differently from AED, takes the protein reading frame into account

Annotation: the process of identifying gene location in the genome (structural annotation) and assigning information that describes the biological function (functional annotation)

Assembly: a computational representation of a genome or transcriptome sequence; also the process of putting nucleotide sequences in the correct order

BUSCO: a tool to assess genome, transcriptome, and proteome completeness. It searches for single-copy orthologs from OrthoDB that are expected to be highly conserved among related species

Class I and class II repeats: TE class of repeats; class I of repeats replicates via an mRNA intermediate that is reverse-transcribed to DNA. Each replication cycle produces an extra copy of the repeat (copy-and-paste mechanism). The class II type of repeats moves through a DNA type of intermediate, integrating the repeat sequence into another position of the genome (cut-and-paste mechanism)

Chemotype: chemical types; *Cannabis sativa* varieties are defined by the content and composition of cannabinoids and terpenoids

Contig: a set of overlapping DNA segments representing a consensus region. In the assembly, contigs refer to overlapping read data

dN/dS: Ratio of number of synonymous substitutions per synonymous site

Embryophyta: major grouping of plants, known as “land plants.” The clade includes both the non-vascular and vascular land plants

Gene: in the context of genome annotation, the smallest unit that encodes for a product. The term gene is here both used for proteins and repeats

Holometabolous vs. hemimetabolous: respectively complete vs. incomplete metamorphosis in insects. Coleopterans are holometabolous species, and the insect develops through all four stages: embryo, larva, pupa, and imago

k-mer: substrings of length k contained within a biological sequence such as sequencing reads, used as parameters in several genome assembly tools and k-mer frequency histograms as in GenomeScope

Monoecious vs. dioecious: here used for plants; monoecious organisms have both male (or more accurately pollen-producing) and female (or more accurately ovule producing) reproductive organs, whereas dioecious organisms consist of male and female organs in separate individuals. Gen. *Picea* is monoecious, and *Cannabis sativa* is dioecious

N50: assembly metric; the length of the shortest longer contig length that is part of the set of contigs that covers 50% of the total assembly

NG50: genome assembly metric; same as N50 except that the length of the shortest longer contig is part of the set of contigs that covers 50% of the total genome size. NG50 allows for more accurate comparisons between assemblies because it accounts for different reconstruction sizes

Orthogroup: used to define gene families; a set of genes from multiple species descending from a single gene in the last common ancestor (LCA). Orthogroups are identified through homology by sequence similarity

Pfam: Protein family; database of protein domain families commonly used for functional annotation

Reconstruction size: genome size in bp of the assembled genome, calculated for contigs/scaffolds >1 kb. Reconstruction size is expected to be the same as experimentally estimated genome size (1C) and describes genome assembly completeness

RNAseq: a sequencing technique that generates sequencing reads and reveals the presence and quantity of RNA at a given time and tissue. Common RNAseq protocols select for RNA molecules with a poly-A tail, such as the mRNA that codes for proteins

Scaffold: a later stage of assembly; two or more connected contigs, typically joined by gaps

Secondary metabolites: small organic molecules produced by an organism and not directly involved in the organism's growth, development, and reproduction. Cannabinoids or flavonoids are two examples of secondary metabolites

Speciation: an evolutionary process by which populations evolve to become distinct species. The isolating mechanisms that lead to speciation are allopatry, peripatry, parapatry, and sympatry. In allopatric speciation, a species population becomes separated by a geographic barrier. Peripatric/parapatric speciation is speciation in which a new species is formed from an isolated peripheral population. Sympatric speciation occurs when populations of a species that share the same habitat become reproductively isolated from each other

ω: ratio of nonsynonymous to synonymous substitution rates used to quantify selective pressure on amino-acid sequences

λ: rate of gene gain/loss, per gene, per millions of years. The birth and death rate parameter describes the probability that any gene will be gained or lost

Acknowledgments

I would like to acknowledge my advisor Prof. Inanc Birol for guidance and support. I would also like to thank my supervisory committee, Prof. Steven Jones, Prof. Sean Graham, Dr. Mathias Schuetz, and Prof. Joerg Bohlmann, for their insightful comments and valuable feedback during my committee meetings. I am grateful to Prof. Sean Graham for his thesis editing and revisions, his dedication to supervising the phylogenomic analyses in my projects, and for providing exciting scientific discussions regarding plant evolution: he provided fundamental support to this thesis work.

I want to acknowledge the financial support of the UBC Bioinformatics program, the UBC Ph.D. four-year fellowship, the ECOSCOPE NSERC Create program, and the MITACS accelerate fellowship.

Chapter 2: This work was funded by Genome Canada, Genome Quebec, Genome prairies, and Genome British Columbia. I thank the SpruceUp project for the genome assemblies of spruce that I used in my analysis.

I want to especially thank the Bioinformatics Technology Lab spruce team, Mr. Rene Warren and Mrs. Lauren Coombe, for supporting my involvement in the project and helping with the genome annotations. Their invaluable help made the project's success possible and allowed me to add this work to my thesis. I want to thank Dr. Carol Ritland for the scientific discussions about spruce genomics and Mr. Mack Yuen for the help with navigating the spruce transcriptomes. I also want to thank Mr. Ka Ming Nip for giving me insights and help with the transcriptome assembly of non-model organisms. I also want to thank Dr. Nathalie Pavy for supervising the gene family's expansion and sharing fruitful ideas for my thesis. I want to especially thank Prof. Loren Rieseberg for the opportunity to share my work with his lab and receive precious feedback about plant evolutionary genomics.

Chapter 3: Genomics data from this work are part of the Canadian Foundation for Innovation and Canada’s Genomic Enterprise (CGEn) CanSeq150 program.

I want to thank Prof. Justin Whitehill and Prof. Steven Jones for their support and supervision, which allowed me to add this chapter to this thesis. They have followed all the stages of my research and have always been excited about my work. I also thank Dr. Christopher Keeling for his scientific discussions about insect genomics. I acknowledge Mr. Luka Culibrk and Mrs. Lauren Coombe for their scientific discussions and help with the genome assemblies. I want to thank Dr. Petr Novak and Dr. Pavel Neumann, authors of RepeatExplorer, for the help with running the tool. I also want to thank Dr. Kamil S. Jaron for sharing his knowledge on k-mer histograms and genome ploidy.

Chapter 4: This work was part of an eight-month industrial internship. I thank Willow Biosciences Inc. and the MITACS accelerate program for making this work possible. I thank Willow Biosciences Inc. for giving me the chance to work on Willow-alpha.

I want to especially thank Dr. Mathias Schuetz for always being excited about science and providing continuous guidance in the project. Dr. Matt Workentine's help was fundamental for the genome assembly and annotations; furthermore, thanks to his support, I have learned a lot about annotations. I want to thank the Willow team for their help, especially Dr. Till Matzat and Dr. Shumin Wang. Furthermore, I want to thank Prof. Simone Castellarin and Mrs. Yifan Yan for their collaboration and insightful discussions about plant metabolomics.

I would also like to mention the Evidence to Innovation team at the BC Children's Hospital, Prof. Matthias Gorges and Prof. Elodie Portales-Casamar, for having the opportunity to work with their team and learn more about scientific writing.

I thank Mr. Luca Intini for assisting with the thesis figures and for editorial help with the thesis.

And finally, I want to thank myself for not giving up. All this work will not be possible without me believing in it.

Dedication

"The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day. Never lose a holy curiosity."

A. Einstein

To Luka for his endless love; he is the place where I feel at home every minute of my day. My infinite love goes to him for all the battles that we have fought together and for all the routes that we have undertaken together.

To my grandmother and mother. My grandmother has taught me to laugh in bad and good situations, and my mother has taught me to fight and never give up.

To the people that have supported me, fed my natural curiosity, and been my journey mates in science. Every bit of experience I have absorbed from them is like a Roman mosaic, where all the pieces are combined into a beautiful picture that can only be appreciated if looking from a further distance.

To the ECOSCOPE microcosmos and the remarkable group of entrepreneurs where I felt welcome and part of the "unity." To my teammates at Willow Biosciences Inc., who have helped me achieve my thesis and taught me to look at science with different eyes.

To all the small stones that I have collected along my way and keeping on my shelf, remembering the hard times and the victories.

To all the problems in evolutionary genomics that do not have an answer yet. To the lonely orphan genes and the outlying species. To all the arthropods and their incredible evolutionary history. To the Cambrian explosion.

To my curiosity and knowledge thirst, to all my dreams including (but not only):

- Joining an Antarctic scientific expedition as in "Encounters at the end of the word" and sequencing the genome of a marine species that none has discovered yet
- Discovering an extremophile bacterial species from a remote Vulcano, being part of the expedition, and the adventure
- Sequencing extinct plants and ancient amphibians, studying paleogenomics

Science and research are my oxygen and make everything exciting and extraordinary.

I dedicate this work to Douglas Hofstadter, Sir. Karl Popper, Thomas Kuhn, Gottfried Wilhelm Leibniz, Maurits Cornelis Escher, Rene Magritte, Salvador Dalí, Italo Svevo, and Werner Herzog for the inspiration I received when working on my projects.

To those about to rock.

To every challenging situation that comes to an end. It is only being patient and strong forges the spirit of the adventurers.

Always remember: *"Everything will be okay in the end. If it's not okay, it's not the end."*

John Lennon

Chapter 1: Introduction

1.1 Motivation

The fast decrease in DNA sequencing costs and the development of more sophisticated bioinformatics tools has broadened opportunities for applying genomic approaches to the study of non-model organisms. Genome assembly is now being applied to a broad range of eukaryotic species. Studies using genome assemblies can answer questions about genome evolution and characterize genetic variation, which can be used for species conservation and selective breeding. The number of sequenced non-model organisms with complex genomes, often defined as large and/or highly repetitive, is growing fast, and the number of assembled genomes is also accumulating rapidly (Tagu, Colbourne, and Nègre 2014; Ellegren 2014).

Several large-scale initiatives are currently generating a catalog of genome assemblies from eukaryotic species, such as the Earth BioGenome Project (Lewin et al. 2018), i5K (specializing in insects) (i5K Consortium 2013), and other collaborations that have undertaken the effort to sequence individual species of economic interest (Scott et al. 2020; Badouin et al. 2017). The assembled genomes are of less value to the scientific community if not combined with careful annotation. Genome annotation identifies putative functional elements in the genome, which provides fundamental information relevant to understanding the biology of the sequenced organism. Genome annotations have become essential for addressing an extensive range of downstream analyses on the species genome evolution in the context of related species.

Several Canadian initiatives have provided high-quality genomes of non-model organisms that are important for the national economy. Examples include the Spruce-Up (Spruce-Up consortium; Depardieu et al. 2021; Girardin et al. 2021) and the CGen-funded (the national platform for sequencing and analysis) CanSeq150 initiative (Canada's Genomics Enterprise CanSeq). The Spruce-Up project delivers knowledge relevant to breeding spruce trees, gen. *Picea* (Pinaceae), that are among Canada's most significant forest resources (Natural resources Canada 2020). The CanSeq150 project supports biodiversity and conservation across Canadian animal and plant species. The two projects have led to

genome sequencing and assembly of (a) two spruce taxa endemic to Canada: *Picea engelmannii* (Engelmann spruce), *Picea sitchensis* (Sitka spruce), and to the re-assembly of two taxa, *Picea glauca* (white spruce), and *Picea engelmannii* x *glauca* x *sitchensis* (interior spruce), together with the sequencing and assembly of (b) the natural insect pest of spruce, the spruce weevil or *Pissodes strobi* (Curculionidae).

Several sequencing projects have been accomplished by Canadian industrial sectors, such as those aiming to improve the cultivation and breeding of *Cannabis sativa*, of increasing economic interest following medical legalization in 2001. Canadian businesses are currently sequencing the genomes of *C. sativa* varieties in combination with the study of their gene expression and metabolic profiles. Combining "omics" data and genome annotations, like the ones generated in my thesis work, is an efficient approach to studying *C. sativa* biosynthetic pathways, allowing us to better understand its biology (Hurgobin et al. 2021).

The primary goal of my thesis is to study the genomes of the species that are important for the forestry and *C. sativa* industries in Canada. The described genomes have complex genomic features, such as large genome sizes and/or high repetitiveness, which make their annotation and comparison especially challenging. The availability of genomes and annotations from closely related species allows for comparative genome analysis.

1.2 Genome complexity in genomes of non-model species

1.2.1 Gene structure and gene families

Protein-coding genes in eukaryotic genomes contain stretches of coding and non-coding regions. Protein-coding genes start with a transcribed and untranslated sequence called the 5'-untranslated region (5'-UTR). The first exon in multi-exonic genes contains this UTR region. The terminal exon contains another untranslated region called the 3'-UTR. The portion of the gene that codes for protein is called the coding sequence (CDS) and is delimited by translation start and stop codons. Regions that are noncoding

stretches of DNA internally to the CDS sequence are defined as introns and form a significant portion of a gene, covering up to 95% of its length (Mattick and Gagen 2001). Gene lengths in eukaryotes, including exons, introns, and UTRs, can reach several kbs and have been shown to correlate with genome size (Xu et al. 2006). The identification of genes is complicated by substantial variation in exon lengths and exon numbers: individual exonic regions can be as short as a few bps. Introns can also be quite variable in size, reaching up to several kbs in length in some eukaryotes (Zhu et al. 2009). Eukaryotic genomes can have non-coding intergenic regions, reaching 98.8% of the total genome in some species (Hou and Lin 2009). The placement of gene positions does not follow a regular pattern in genomes, with some genes being close together or even overlapping and others dispersed more sparsely around the genome.

Eukaryotes are also known to have a higher rate of gene duplication than prokaryotes (Lynch and Conery 2003). Eukaryotic genomes, on average, contain 20,000 annotated proteins per species, according to databases curated by the national center for biotechnology information (NCBI) (National Center for Biotechnology Information 2021a), and double that in some land plants, which can have >40,000 annotated genes. Genes can be organized into evolutionarily related groups defined as gene families. Genes in the same gene family share a similar protein sequence through common decent (homology) and often cover the same biological function. By definition, gene families contain homologous genes, divided into orthologous and paralogous s (Kristensen et al. 2011). Orthologous genes arise by speciation at their most recent point of origin, and paralogous genes occur by duplication from their ancestor (Demuth and Hahn 2009). Comparative genomics of both paralogs and orthologs assumes that every gene belongs to a gene family (Ohno 2013), even if it contains only one member. Every gene descends by duplication from an existing gene derived from an ancient gene family. Although newly duplicated genes lose their function in most cases in one of the copies (non-functionalization), occasionally, they gain new functions, creating gene variability. In neo-functionalization scenarios, the new copy gains a novel function different from the original copy (Force et al. 1999); alternatively, mutations occur in all the gene copies that specialize in performing complementary functions as in the sub-functionalization scenario (Lynch and

Conery 2000; Lynch and Force 2003). The study of gene families is one of the most important analyses of gene sequences. It is applied to studying the evolutionary relationship between the genes and the species they have evolved from. Gene duplication is a leading mechanism in generating functional diversity in eukaryotes (Ponting 2008). It produces new functions that can serve, for example, as a mechanism for adaptation (Kondrashov 2012) and for reducing the risk of species extinction (Crow and Wagner 2005).

1.2.2 Genome repetitiveness

Repetitive sequences in eukaryotic genomes are formed from homologous DNA fragments repeated in high copy numbers. Two major classes of genomic repeats are transposable elements (TE) and tandem repeats (TR). TR has sequential repeat sequences in proximity and is represented by two major classes: satellites and rDNA repeats. Satellites are classified according to the length of the repeated unit into microsatellites (1–5 bp), minisatellites (6–100 bp), and satellites of long repeated groups that are several Mbs in length (Mehrotra 2014). Satellites compose the essential structures of the centromeres and ensure the stability of the chromosomes (Lower et al. 2018; Garrido-Ramos 2017). rDNA repeats code for the RNA-based components of the ribosomal machinery and harbor fundamental housekeeping genes. rDNA gene repeats encompass polycistronic units of 18S, 5.8S, and 25S rDNA genes, in that order, separated by internal transcribed spacer (ITS) and external transcribed spacer (ETS) regions. They also have the 5S rDNA gene, which comprises a separate repeat unit.

Transposable elements are divided into two major classes based on their transposition mechanisms, and each class is subdivided based on its specific mechanism of genome integration (Kapitonov and Jurka 2008). Class I TEs, also known as retrotransposons, mobilize with a copy-and-paste mechanism. An RNA intermediate is reverse-transcribed to complementary DNA (cDNA) and integrates into chromosomes through cleavage and strand-transfer reaction catalyzed by an integrase like retroviruses (Brown et al., 1997). Two repeat subclasses from the class I TE are the long interspersed terminal repeats (LINE) and the short interspersed terminal repeats (SINE), commonly identified as

non-LTRs, that use target-primed reverse transcription as an integration mechanism (Luan et al., 1993). Class II TEs are known as DNA transposons and mobilize through a cut-and-paste replicative mechanism without using an amplification step. Subclasses of TE are further divided into subgroups conserved across various organisms (see (Craig 2020) for more detail).

1.2.3 Genome size

While the number of genes remains similar across the genomes of all species, ranging between 20,000 and 40,000 on average (National Center for Biotechnology Information 2021a), genome size can vary significantly across species and depends primarily on the length of intergenic and intronic regions. Genome size varies by five orders of magnitude in eukaryotic genomes (Oliver et al. 2007) and varies to a similar degree in plants, where it reaches a variation interval between 10 Mb and 149 Gb (Blommaert 2020).

In eukaryotes, two major forces contribute to genome size evolution. Whole genome duplication (WGD), or doubling of the genome size, is a prominent duplication mechanism that occurs in multiple rounds of polyploidization and is especially common in plant lineages (Carretero-Paulet and Van de Peer 2020). WGD induces genomic alterations that change the gene content over the evolutionary time, where the newly created genes diversify in their function due to neo and sub-functionalization (Adams and Wendel 2005).

The second mechanism of genome size evolution is through the accumulation/removal of TEs (Kazazian 2004). TE activity introduces significant variation in the genomes even between closely related species (Gregory 2005). Several studies have shown that the rate at which transposition occurs is an important driver of genome complexity (Bennetzen and Wang 2014). Indeed, variation in genome size correlates with the amount of TEs, as reported for plants and insect genomes (Michael 2014; Wu and Lu 2019). An abundance of copy-and-paste class I repeats correlates with increased genome complexity and genome size, observed in plants (Ibarra-Laclette et al. 2013). Although class I repeats are more likely to

contribute to the genome size (Lee and Kim 2014), class II acts through several mechanisms that can modify the number of genome copies, such as TE excision events that are repaired by gene conversion, using a sister chromatid as a template (Plasterk and van Luenen 2002). Genome size expansion driven by repeats is counterbalanced by genome purging phenomena, including the elimination of repeat sequences by recombination. Repeat expansion is also regulated by a strict gene expression regulation that blocks the replication of repeated sequences by compacting the chromatin structure (Grover and Wendel 2010). Although genomes may have a large abundance of TEs, most mobile elements are inactivated by gene regulation or loss of function (Raskina et al. 2008; Martienssen 2008). Different amounts of TEs in the genomes of related species serve as evidence of repeat rearrangements in past times.

Other mechanisms that alter the genome size include unequal crossover (Smith 1976), especially in TR, where mutations in the sequence and random homology-dependent unequal crossover events can alter the number of TR copies. Another mechanism that modifies the genome size is the non-allelic homologous recombination (NAHR), a driving force for the replication through segmental duplication (SD) in genomic segments containing high sequence similarity. SD is considered to play an important role in gene gain/loss in gene families (Marques-Bonet, Girirajan, and Eichler 2009; Duda and Palumbi 1999).

1.3 Genome annotation

Genome annotation provides putative biological information to the sequenced genome, identifying the genes present and determining their function. The most important genome databases such as Ensembl (Aken at al. 2016) and NCBI GenBank/RefSeq (Pruitt et al. 2012) perform largely automated annotations on submitted genomes. Genome databases, however, only select contiguous and complete assemblies for annotation. Thus, they only annotate a small fraction of the submitted genomes. According to the number of genomes submitted to databases and the number of available annotations, only ~10% of the eukaryotic genomes are annotated this way (National Center for Biotechnology Information 2021a; National Center

for Biotechnology Information 2021b). Because of the long processing times for genome annotations by NCBI, often several years, labs that assemble genomes usually prefer also to annotate them.

The annotation process has two interrelated stages: structural and functional genome annotation. Each stage requires several software tools to generate the final set of annotated genes. I will outline the two annotation stages and describe the most common annotation methods used for eukaryotes, including those I use elsewhere in this thesis. I will also describe common benchmarking methods used to evaluate the quality and completeness of annotations.

1.3.1 Structural genome annotation

Structural annotation is the discovery of regions in a genome that encode genomic features. Structural annotation uses *ab initio* and homology-based approaches, which predict gene signals from the DNA sequence using statistical models and gene homology based on the sequence similarity to other genes, respectively. The *ab initio* approach requires the gene models to be trained for species-specific gene properties, such as intron length and intron-exon organization. This approach can also be applied to more distantly related species. Homology-based methods instead require genes from evolutionarily closely related species to generate generally more accurate gene predictions. Several genome annotation tools are built around one of the two approaches or use a combination of both, reviewed in more detail elsewhere (Campbell and Yandell 2015; Ejigu and Jung 2020).

Structural annotation focuses on identifying and characterizing various genomic elements, with the protein-coding genes being of particular interest for downstream data analysis. Repeat annotation is one of the most common types of annotation performed in non-model genomes, and downstream analysis of repeats also provides information about the evolutionary genomics of the organism as for the “age” of repeats, providing an estimate of the repeat class amplification.

1.3.1.1 Protein-coding gene annotation

Annotation algorithms for identifying protein-coding genes in eukaryotic genomes have remained unchanged for two decades. General hidden Markov models (GHMMs) are methods in use for genome annotation implemented first in AUGUSTUS (Stanke et al. 2006) and GeneMark (Lomsadze et al. 2005). The annotation algorithms, also defined *ab initio*, are based on the genome sequence alone and recognize promoter regions, UTRs, exons-intron boundaries, and non-coding regions with high accuracy. Given the high amount of available transcriptomics data, a significant improvement in the field has come from biological "hints" in guiding the *ab initio* algorithms (Stanke et al. 2008; Lomsadze, Burns, and Borodovsky 2014). These hints are provided by aligning biological sequences from other reference organisms to the genome, defining gene features that can be further used in the annotation, including start and stop codons and exon-intron boundaries.

The most common computational pipelines for genome annotation are MAKER (Holt and Yandell 2011) and the more recently developed BRAKER (Brůna et al. 2021) tool suites. Both pipelines use a combination of software run in a stepwise manner. Common *ab initio* tools, such as AUGUSTUS and GeneMark, are part of these pipelines as statistical gene prediction tools. The user can provide biological hints or evidence of the location of gene elements through transcriptomic or protein-based data. The annotated genes are scored according to their concordance with the evidence, for example, as reported by the annotation edit score (AED), exon annotation edit score (eAED), and quality information tag (QI) implemented in MAKER. Low AED/eAED scores lower than one indicate the degree of concordance between the supporting evidence and the predicted genes. The QI tag provides information about the fraction of supported splice sites and exons (Campbell et al. 2014), given the supporting evidence used in the annotation.

A combination of protein and transcriptome sequences is often preferred because this produces a more robust gene prediction (Keilwagen et al. 2018). Each type of evidence has its strengths and limitations that need to be considered when annotating a newly assembled genome. Protein evidence is

more evolutionarily conserved across organisms than transcriptome-based evidence, allowing the proteins to be used from more distantly related species if data from a close species is unavailable. Transcriptome sequences are generally more helpful if from closely related taxa (preferably from the same species or individual), given the higher rate of evolution of DNA sequences compared to amino-acid sequences. A major caveat of using transcriptomic data is that genes with low gene expression may be missed and fail to provide annotation evidence. Transcriptomic data remains one of the most important annotation datasets because gene expression implies that the predicted gene may produce a biologically viable product. In contrast to protein sequence data, transcriptomic data sets, such as RNAseq or cDNA libraries, are relatively easily obtained from non-model organisms and can provide supporting information for annotating UTR regions in genes.

1.3.1.2 Repeat annotation

Unlike the protein-coding gene annotation, repeat annotation exclusively relies on the repetitive sequence search without the necessity of orthogonal data. Furthermore, the annotation of repeats benefits from the high number of repeated sequences, making their discovery more straightforward than protein-coding genes, whose number is variable across the gene families. The annotation of repeats in non-model organisms includes the initial generation of a repeat library, a non-redundant collection of repeat sequences found in that particular genome. Repeat libraries are generated using both *de novo* (Flynn et al. 2020) and homology-based strategies (Ellinghaus, Kurtz, and Willhoeft 2008; Neumann et al. 2019), and several tools provide a combination of both for a more comprehensive annotation. There is a wide variety of annotation pipelines for different classes of repeats, especially in the case of TE (Ou et al. 2019). The second step of repeat annotation is the identification of repeated sequences in a genome, using the repeat library as a reference. The generation of repeat libraries is actively being researched (Ou et al. 2019; Girgis 2015), but methods for identifying repeats in a genome have not undergone significant changes for a decade (Smit, Hubley, and Green 2008). Repeat annotation is run before the annotation of

protein-coding genes, which helps avoid repeated sequences being erroneously annotated as part of the coding sequences of genes.

1.3.2 Functional gene annotation

Functional gene annotation of protein-coding genes characterizes the putative biochemical or biological function of newly identified genes. Automated functional annotation is preferred due to the large number of genes involved. The first method for functional annotation is to assign a function based on inferred homology to genes in existing databases (Friedberg 2006). The function is automatically assigned based on previously annotated genes with high sequence similarity to the newly annotated genes. To achieve high efficiency of similarity-based functional annotation, it is necessary to have a good representation of genes from closely related species. A complementary approach for functional annotation uses intrinsic sequence functionality extracted directly from protein sequences (Griesemer et al. 2018). Two examples of the latter are based on protein families (Pfam) (Mistry et al. 2021) and the Gene Ontology (GO) used to infer molecular function, biological process, and cellular component (Hill et al. 2008).

1.3.3 Benchmarking gene annotations

A valuable estimate of the annotation quality is the number of annotated genes and transcripts and statistics that includes the length of genes, mRNAs, exons, and introns (Yandell and Ence 2012). Such counts and lengths are especially helpful when compared to already annotated genome assemblies, but they do not estimate the completeness of the annotated genes. Gene annotation completeness is evaluated through an overall completeness score according to a set of genes or protein domains that are universally distributed. A broadly used tool for benchmarking annotation completeness is benchmarking universal single-copy orthologs (BUSCO) (Simão et al. 2015), run on the annotated protein sequences; the software evaluates the presence of predicted single-copy proteins. The number of “complete – single copy,” “complete – duplicated,” “fragmented,” and “missing” BUSCO genes are reported as a percentage of the

total expected proteins in lineages from the OrthoDB database (Kriventseva et al. 2017). Domain-based transcriptome or proteome quality assessment, or DOGMA (Kemena, Dohmen, and Bornberg-Bauer 2019), uses a similar principle that evaluates the presence of conserved protein family (Pfam) domains in annotated proteins. The tool looks for Pfam domains (Ekman, Björklund, and Elofsson 2007) and reports the Pfam completeness based on precomputed conserved domain arrangements (CDA) obtained from related species. The final score considers the presence of single Pfam domains and the consecutive order of domains. Benchmarking methods, such as the ones mentioned before, rely on conserved genes common to multiple species in a lineage, which makes the number of elements for comparison relatively low, as in the case of BUSCO Embryophyta lineage, where only 1,614 orthogroups are available for benchmarking. The number of orthogroups for benchmarking increases for specific lineages and is less abundant for lineages, including more taxa. The OrthoDB database (Kriventseva et al. 2019) used in BUSCO is under constant development, yet it lacks characterization for most plant lineages.

BUSCO and DOGMA provide single overall scores for the completeness of annotated genes. The AED/eAED scores and the QI tag reported by the MAKER pipeline described earlier provide individual scores for each annotated gene. In the latter case, the use of individual gene scores helps select high-quality genes and filter less reliable annotations.

1.4 Research objectives

Recent Canadian sequencing research projects have generated a set of genomes from non-model species, providing an opportunity to study their biology and evolution. In my thesis, I assembled and collated the genomes of several species and produced high-quality annotations for use in comparative genomics. For each sequenced and annotated genome, I answered different biological questions regarding the unique biology of each species. My research objectives are as follows:

1. **Chapters 2, 3, and 4:** Infer phylogenetic relationships between the newly sequenced species and previously sequenced organisms. I focused on using single-copy genes obtained from the annotations.
2. **Chapters 2 and 3:** Whole-genome annotation and comparative genomics
    1. **Chapter 2:** Genome annotation of four North American spruces (*Picea*) and identification of unique to each species protein-coding molecular features. I also compared my annotations to annotations from other conifers (spruce and pines) in this chapter.
    2. **Chapter 3:** Genome assembly and annotation of the spruce weevil, *Pissodes strobi*. Discovery and timing of the origin of genomic repeats in the genome of *P. strobi*. This chapter also compares the genomic features of *P. strobi* to several species in the same family (Curculionidae).
3. **Chapter 4:** Genome assembly and annotation of Willow-alpha *Cannabis sativa* variety. I characterized flavonoid/anthocyanin biosynthetic pathways in *C. sativa* by using the Willow-alpha reference. The chapter focuses on studying the two pathways using transcriptome and metabolome data. I specifically compared four *C. sativa* varieties with different leaf pigmentation and identified differentially expressed genes that correlate with metabolites from the flavonoid/anthocyanin metabolic pathways.

Chapter 2: Comparative genome annotation of four North American spruces (*Picea*, Pinaceae)

2.1 Author summary

Spruces (*Picea*) are keystone species of many boreal and mountain ecosystems, with considerable economic significance for countries at higher northern latitudes, such as Canada. Genomics has been key to understanding the evolutionary biology of spruces and characterizing their genetic diversity. The spruce genomes have a large genome size and high repetitiveness, limiting the generation of high-quality genome assemblies and gene annotations. Here, I perform and compare the genome annotations of four spruce taxa: *Picea engelmannii* (Engelmann spruce), *Picea sitchensis* (Sitka spruce), *Picea glauc*a (white spruce), and a naturally occurring introgress of these three species, interior spruce. I investigated the phylogenetic relationships among these taxa and detected patterns of expansion and contraction of gene families. I further identified genes under positive selection. These analyses strengthen our understanding of conifer genome evolution as their comparison offers clues into the genetic basis of adaptation and ecology of conifers. This chapter addresses objectives 1 and 2.1 in section 1.4.

2.2 Introduction

Spruces (*Picea*) are conifers and a widespread genus of gymnosperms in the northern hemisphere. Spruces are keystone species of many ecosystems because of their abundance and play a major role in forestry. More than 300 million spruce seedlings are planted in Canada alone every year (National Forestry Database–Canada 2021), and mature spruces are an essential source of lumber and wood fiber, contributing a substantial portion of the national domestic product or GDP (Natural Resources Canada, 2020).

The sequencing and assembly of conifer genomes require an unprecedented computational effort because of their large genome size and high repetitive sequence content. The genome size of conifers

ranges between 7 and 37 Gb (Ahuja and Neale 2005), with most members of Pinaceae, the pine family, above 20 Gb (Pellicer and Leitch 2020). The genomes of the largest seed plant group, angiosperms, have been reshaped by recent whole genome duplication (WGD) events (Zimmer et al. 2007) that account for the substantial diversity of genome sizes in this group (Landis et al. 2018; Ren et al. 2018); in contrast, WGD is absent in extant conifer lineages. Conifers experienced WGD in their ancestry, observed in the Cupressaceae + Taxaceae and Pinaceae lineages ~210–275 and 200–342 million years ago, respectively (Li et al. 2015). Evidence of hybridization events that are common in angiosperms, such as autopolyploidy and allopolyploidy, are mostly absent in conifer lineages: the coast redwood (*Sequoia sempervirens*, Cupressaceae) is one of the two described conifers with a polyploid genome structure (Neale et al. 2022; Scott et al. 2016). Given the lack of evidence of recent WGD and polyploidy, the primary driving force in the expansion of conifer genomes appears to be their tendency to accumulate transposable elements (TE) (Prunier, Verta, and MacKay 2016). Repeat elements in conifers are largely consistent across published conifer genomes, with ~70% of their genome covered by repeats (Stevens et al. 2016; Nystedt et al. 2013). Long terminal repeats (LTR) subdivided into LTR-Gypsy, LTR-Copia, and unclassified LTR have a large abundance in the conifer assemblies, representing ~67% of repeats in sugar pine (*Pinus lambertiana*) (Stevens et al. 2016) and ~50% in Norway spruce (*Picea abies*) (Nystedt et al. 2013).

Recently, there has been a rapid accumulation of genomics resources for several spruces, including spruce genomes from the North American continent. The genomes of two spruces have been newly sequenced and assembled: *Picea engelmannii* (Engelmann spruce [genotype Se404-851]) and *Picea sitchensis* (Sitka spruce [genotype Q903]). Along with those, the improved version of the genome assembly is available for the draft genomes of *Picea glauca* (white spruce [genotype WS77111] (Warren et al. 2015)) and interior spruce (*P. glauca* × *engelmannii* × *sitchensis* [genotype PG29] (Birol et al. 2013)), a naturally occurring introgress of the former three. The interior spruce PG29 genotype shares an

asymmetric contribution from the three other genotypes, with a significant contribution from *P. glauca* (Hamilton, De la Torre, and Aitken 2015; Haselhorst and Buerkle 2013).

The different North American spruce taxa, long-lived conifers, have a broad range of geographic distributions, spanning from the east to the west coast of North America. *Picea engelmannii*, which can live for up to 300 years, has a scattered distribution in western North America, confined to the east of the Coastal Mountains of the Rocky Mountains (Figure 2.1a). *Picea sitchensis* is one of the typical trees of the Pacific coastal forests, with a natural range from Northern California to Alaska and a life span reaching 700 to 800 years (Farrar 1995). Its mature trees are the largest among the species represented here, reaching up to 55 m in height and 200 cm in diameter (Figure 2.1b). *Picea sitchensis* is well adapted to the temperate rainforest climatic conditions of the Pacific Northwest with its abundant rain precipitation. Cold tolerant *P. glauca* has a vast continental range that spreads across the North American boreal forests, and it can reach 200 years of age. Their geographic region of intersection is the interior of British Columbia, where the introgressed line of interior spruce is located. Of these four species, interior spruce, the most cold-tolerant phenotype, is widely used in managed forests in western Canada and the United States, and it corresponds to the large area of sympatry between the previous three species.

To date, limited genome annotations are available for spruce. I present the first comprehensive annotations from spruce genomes here. I produced the annotation with a common methodology across the species, enabling direct downstream comparisons. I then use the annotations to characterize selection patterns among orthologous genes and to document expanding gene families, representing some of the molecular differences between the examined spruces.

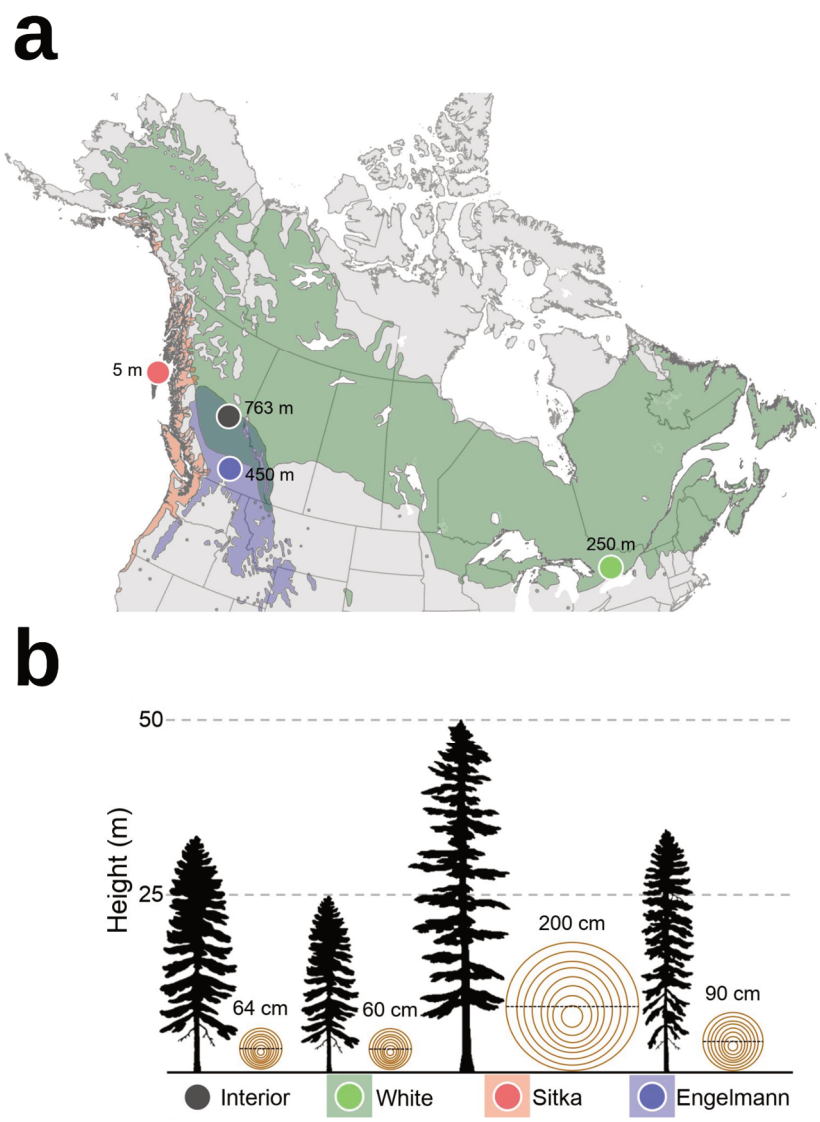

Figure 2.1 (a) Geographical distributions of Engelmann (*P. engelmannii*), Sitka (*P. sitchensis*), and white spruce (*P. glauca*) and the location of the sampled trees (interior, white, Sitka, and Engelmann spruce). The sampling location of interior spruce (*P. glauca* × *engelmannii* × *sitchensis*) is on the overlap between Engelmann and white spruce geographic distributions. Colored dots indicate the locations of the specific *P. engelmannii* (blue), *P. sitchensis* (red), *P. glauca* (green), and interior spruce (gray) genotypes sampled for genome sequencing, numbers indicating the location elevations in meters. **(b) Spruce dendrometric attributes.** Maximum height and diameter of spruce taxa.

2.3 Methods

2.3.1 Sample collection, sequencing, and genome assembly

Apical shoot tissues were collected from a single individual used as a representative genotype for each taxon (collection locations noted in Appendix Table A.1). The genotypes sequenced for *P. engelmannii*, *P. sitchensis*, and *P. glauca* are allopatric populations distant from the area of sympatry where interior spruce is found. Genomic DNA was extracted and used to build sequence libraries with a range of sequencing platforms and protocols for each spruce genotype, as described in Appendix Table A.2. The sequenced reads with the highest genome coverage are from short and linked-read protocols. Long-read technology was sequenced at 2–4-fold coverage for *P. engelmannii* and *P. sitchensis* and was used for scaffolding their draft genomes, thus not used in the primary genome assembly. At the time of this study, Chromium 10x linked reads and Oxford Nanopore long reads were the most recent sequencing platforms, and genomic reads from these platforms were generated starting in 2017 (Appendix Table A.2). The sequenced reads are assembled with algorithms that consider short and linked reads, such as Abyss (Jackman et al. 2017), Tigmint (Jackman et al. 2018), and ARCS (Yeo et al. 2018), and were scaffolded with mate-pair short reads (MPET) and long reads when available. Genome assembly completeness was assessed using BUSCO v5.1.4 (Simão et al. 2015) and the odb10 Embryophyta (land plants clade) core genes (n=1,614) (Kriventseva et al. 2019).

2.3.2 Estimate of genome sizes

I estimated genome sizes using the k-mer frequency histograms computed by ntCard v1.1.0 (Mohamadi, Khan, and Birol 2017). The software was run on the complete set of short reads from each species. After excluding the effect of erroneous k-mers from the histogram, the homozygous k-mer (common in both parental alleles) is identified, usually the maximum peak in the histogram. Although estimating abundances can be refined using distribution mixture models, this first-order approximation works well for the range of experiments analyzed here. The genome size estimate was then performed by integrating

the error-free k-mer frequency histogram curve. I estimated the final value of the genome size by averaging the values for the range of k-mer lengths of 30, 40, 50, 60, 70, 80, and 90 bp. More details about the process can be found in Birol, Mohamadi, and Chu (2018).

2.3.3 Annotation of protein-coding genes

High-quality genome annotations can be produced using gene expression and proteome information from the same or close species commonly referred to as supporting genes evidence. The four spruce genomes were annotated based on protein-coding sequences from full-length cDNAs from *P. glauca*, interior, and *P. sitchensis* (Rigault et al. 2011; Ralph et al., 2008) and plant proteins from SwissProt ("UniProt: The Universal Protein Knowledgebase in 2021" 2021). I used additional RNAseq libraries for interior spruce and *P. sitchensis* (Appendix Table A.3), assembling their short reads with pooled assembly approach (*-pool* option) in RNA-Bloom v0.9.8 (Nip et al. 2020). The resulting transcripts were screened for biological contamination. Briefly, a series of Bloom filters (Chu et al. 2014) were created with k-mers from viruses, bacteria, fungi, and aphids; transcripts without a match in any Bloom filter (default settings) were kept for annotation. Only transcripts with putative protein-coding sequence (CDS), selected through EvidentialGene v2017.12.21 (Gilbert 2013), were used as supporting evidence.

I subsampled genomes for scaffolds containing putative genes to decrease the annotation effort and speed up the annotation pipeline. Some scaffolds were first removed based on their large fraction of repeat content and gaps, so I annotated only those with at least 1 kb of non-repetitive and unambiguous nucleotides. I further filtered the resulting contigs to select scaffolds with at least one complete transcript from the spruce transcriptome resources (RNAseq or full-length cDNAs) aligned with ≥95% sequence coverage and identity. The selected scaffolds were divided into bins containing 1,000 scaffolds and annotated using the MAKER v2.31.0 pipeline (Holt and Yandell 2011) in two iterations. The first iteration, run with *prot2genome=1* and *est2genome=1*, included full-length cDNAs, proteome, and species-specific evidence from the RNAseq, where available (Appendix Table A.3). The second iteration

ran with the *prot2genome=0* and *est2genome=0* options using combined evidence from the four genotypes, with the full-length cDNAs and the proteome. The annotation steps and supporting evidence are provided in Appendix Figure A.1.

Gene parameters for the *ab initio* gene prediction are not available for conifers, making it unlikely to predict accurate gene annotations in spruce. To address this lack, I generated spruce meta parameters for AUGUSTUS v2.5.5 (Stanke et al. 2008) by retraining the gene models with BUSCO v3.1 *-long* option and odb9 Embryophyta core gene set. AUGUSTUS and SNAP v2013-11-29 (Korf 2004) gene predictors were trained with a preliminary MAKER run. The genes used in the training were selected from the genome annotation, with filtering for exon annotation edit (eAED) score ≥0.5 and quality index (QI) tag. I used the QI tag to select genes with 80% transcript or protein evidence overlap with predicted exons and genes with 80% of the splice-sites confirmed by transcript evidence. GeneMark v2.3c (Lomsadze et al. 2005) was self-trained as GeneMark-ES with an unsupervised procedure where the algorithm defines the annotation parameters automatically.

The final annotated genes were selected if being compliant with all the following conditions: (a) having an annotated Pfam domain (Mistry et al. 2021) or BLASTP (Camacho et al. 2009) hit to SwissProt ("UniProt: The Universal Protein Knowledgebase in 2021" 2021) with an e-value <1e-5, (b) having an eAED Score <1 by MAKER, (c) having gene length >1 kb, (d) polyexonic genes have a minimum intron length of 10 bp, (e) having a complete CDS (start and stop codons), (f) their start and stop being more than 500 bp from the scaffold ends, and (g) translated proteins lacking Pfam match with LTR repeat elements, such as GAG, Env or Pol (viral LTR domains).

I functionally annotated the final gene set using InterProScan v5.30-69 (Jones et al. 2014) and Pfam v31 (Mistry et al. 2021); gene name assignment was done by BLASTP top hit to SwissProt.

Masking repetitive DNA elements before performing gene predictions helps minimize spurious predictions. So a custom repeat library was built for each spruce genome by combining *de novo*-identified elements for each genome and curated elements from RepBase (Bao, Kojima, and Kohany 2015). Long

terminal repeat (LTR) elements were identified using LTR_retriever v1.3 (Ou and Jiang 2018) with candidate sequences provided by LTRharvest v1.5.9 (Ellinghaus, Kurtz, and Willhoeft 2008) *GenomeTools,* using the arguments *-similar90 -vic10 -seed20 -minlenltr100 -maxlenltr7000 -mintsd4 -maxtsd6 -motifmis1,* both with and without *-motifTGCA*, and LTR_FINDER v1.06 (Xu and Wang 2007) with the arguments *-D15000 -d1000 -L7000 -l100 -p20 -M0.9*. Redundant elements identified by LTR_retriever were removed with CD-HIT-EST v4.6.6 (Fu et al. 2012) using the arguments *-c0.8 -G0.8 -s0.9 -aL0.9 -aS0.9 -M0.* Additional *de novo* repeat elements predicted by RepeatModeler v1.0.8 (Hubley 2021) were combined with repeats from RepBase v22.08 (Bao, Kojima, and Kohany 2015) to yield a final custom library of repeat elements for each taxon.

I assessed the overall genome annotation quality using DOGMA v3.4 (Kemena, Dohmen, and Bornberg-Bauer 2019), evaluating conserved plant Pfam domains utilizing 918 single-domain Conserved Domain Arrangements (CDAs) and 563 multiple-domain CDAs from the plant kingdom. I also assessed annotations using BUSCO v5.1.4 with *-m protein* options for the odb10 Embryophyta core genes. A haphazardly selected set of ~20 annotations was checked in integrative genomics viewer (IGV) (Robinson et al. 2011) to visually inspect the quality of the generated gene annotations in the context of the supporting gene evidence. Additionally, I manually inspected genes that resulted significant in the comparative analysis performed in this chapter.

2.3.4 Comparative genomics: phylogeny of gen. *Picea*

The species tree was estimated for spruces *P. engelmannii*, *P. sitchensis*, *P. glauca*, interior spruce (genotype PG29), and *P. abies* (Norway spruce) (Nystedt et al. 2013), and two pines, *Pinus taeda* (loblolly pine) (Wegrzyn et al. 2014), and *Pinus lambertiana* (sugar pine) (Stevens et al. 2016), with the latter three designated as outgroup taxa (Appendix Table A.4).

Genome annotations are available for all taxa. I grouped the sequences in gene families using OrthoFinder v2.3.1 (Emms and Kelly 2019) and inferred gene families from sequence similarity in

orthogroups (OGs), as reported by the tool. I selected a set of 246 single-copy genes to infer the species tree. Proteins from genes in each orthogroup were aligned with Mafft v7.453 (Nakamura et al. 2018) (using the argument *-auto*), which were then used to infer the gene trees with RAxML v8.2.12 (Stamatakis 2014) (PROTGAMMAAUTO and 100 bootstraps options). The species tree topology was inferred with ASTRAL-III v5.6.3 (C. Zhang et al. 2018). The resulting tree was scored with the quartet score (*-q* option in ASTRAL-III), and scores were visualized with DiscoVista v1.0 (Sayyari, Whitfield, and Mirarab 2018). Because ASTRAL-III assigns branch lengths only for internal nodes, I estimated the branch length for all the branches with IQ-TREE v2.1.4 (Minh et al. 2020) under the fixed topology estimated by ASTRAL-III (*-te* option). The option ProtTest in IQ-TREE determined the JTT+F+R5 substitution model as the best fitting. The final species tree was midpoint rooted, and the estimated concordance was evaluated with *phyparts* (Smith et al. 2015) to summarize the percentage of concordance and conflict of individual gene trees with the species topology.

The mitochondrial genomes (mt) are available for all the North American spruces (Jackman et al. 2016, 2020) (Appendix Table A.4), except for *P. glauca* and *P. engelmannii*. The genomes of the latest were assembled with ABySS 2.0.1 (Jackman et al. 2017) (run with k-mer size *k=112* and k-mer minimum coverage multiplicity cutoff *kc=3*) using 404,326,342 and 272,610,382 HiSeq paired-end (PE) reads (2x250 bp) for *P. glauca* and *P. engelmannii*, respectively. The mt genome sequences were selected from these initial assemblies by further aligning assemblies to the interior spruce mt genome assembly using BWA-MEM (Li 2013). The genome assemblies of *P. glauca* and *P. engelmannii* resulted fragmentary, thus containing a higher number of assembled scaffolds than the other genomes in the phylogeny. Given the different contiguity of the mitochondrial genomes, I used MashTree v1.0 (Katz et al. 2019) to infer the sequence dendrogram based on the mt genome. The tool generates sketches of the genome with genomic k-mers. The species dendrogram is inferred through k-mers overlap and the Neighbor-Joining algorithm. Confidence intervals were added from 100 jackknife trees through the mashtree_jackknife.pl script.

Full plastid (pt) genomes are available for all the species (Lin et al. 2019a, 2019b; Coombe et al. 2016; Asaf et al. 2018; Cronn et al. 2008) (Appendix Table A.4), and the entire genome sequence was used to infer phylogeny. I estimated the pt-based phylogeny using two different methods, one based on standard methods and the other on based k-mers. I inferred a phylogenetic tree with standard methods using Mafft v7.453 (Nakamura et al. 2018) (using the argument *-auto*) for multiple sequence alignment and RAxML v8.2.12 (Stamatakis 2014) to infer a pt-based phylogenetic tree (GTRGAMMA model for partitioned analysis and bootstrap analysis (Felsenstein 1985) with 100 repetitions). I used the k-mer based approach to generate a plastid dendrogram with MashTree (Katz et al. 2019), same as for the mt genome, to see how this method performs compared to the RAxML standards.

2.3.5 Gene families and gene gain/loss

I scanned annotated proteins from *P. engelmannii*, *P. sitchensis*, *P. glauca*, interior – PG29, *P. abies*, and pines, *P. taeda,* and *P. lambertiana* to select the longest isoform, which I took as representative for each gene. Analysis of gene family expansion or contraction (gene turnover rate) was run on orthogroups (OG) assigned by OrthoFinder and represented by at least five taxa. I assessed whether gene families are statistically expanded or contracted using CAFE v4.2.1 (Han et al. 2013), based on a random birth and death model, along the taxa of the phylogenetic tree. The software first estimates the gene turnover rate, λ (gene duplication per gene, per million years), from the user-provided gene families and compares the turnover rate against the random birth and death model estimated from the phylogeny. I used the tree derived from the 246 single-copy genes to model the random birth and death in CAFE. Because CAFE requires an ultrametric tree, I converted the tree format with the OrthoFinder make_ultrametric.py script, which rescales the branch lengths by calibrating the divergence between pine and spruces at 116 MYA, as reported by Wang, Tank, and Sang (2000). I run CAFE with a single λ modeling rate across the phylogeny. I estimated the λ rate on less numerous gene families (those with at most 100 total genes) and applied this λ value to the larger gene families (those with >100 total genes) to avoid non-informative parameter estimates. Because errors in the assembly and annotation may lead to biased λ estimates, I

applied error correction models implemented in CAFE to recover less biased estimates. Rapidly evolving gene families were reported for Viterbi P <0.001 and annotated as “expanding” or “contracting” with respect to the phylogenetically closest taxon.

For plotting purposes, significantly expanding/contracting gene families were displayed as a heatmap using R *heatmap.2*. The number of genes in each significant gene family was converted to a Z-score with the following:

$$Z = \frac{x - \mu}{\sigma}$$

Where x is the number of genes in each species, μ is the average number of genes per gene family, and σ is the standard deviation of the gene family. The scores were clustered based on complete linkage for hierarchical clustering using the Manhattan distance (Sørenson 1948).

A modified version of the original gene families dataset (OGF) in which proteins with a sequence length of at least 50% of the length of the longest protein from the same species were filtered from each gene family, as described by Casola and Koralewski (2018), which we refer to here as the F50 dataset. This dataset estimates the potential bias in gene turnover rates due to erroneous annotations and gene cleavage on short scaffolds, observed in taxa with large genomes and fragmented genome assemblies (Denton et al. 2014).

2.3.6 Positive selection acting on orthologs

To assess if any genes experienced positive selection for each of the annotated spruce genomes, I used the ratio of non-synonymous and synonymous substitutions (ω=dN/dS) as a measure of evolutionary pressure on protein-coding genes, considering ω-values significantly more than 1.0 as evidence of evolutionary pressure.

I selected 1,930 OrthoFinder genes clustered as a single copy in all the North American spruces and used these as single-copy orthologs for comparison in all the taxa. I aligned the protein sequences with Mafft v7.453 (Nakamura et al. 2018) (using the argument *-auto*) and used the resulting alignment to guide the CDS alignment with PAL2NAL (Suyama, Torrents, and Bork 2006). Gene families with alignment gaps covering more than 5% of the CDS alignment were removed, and the remaining sites with any ambiguous nucleotides were ignored in the analysis. I calculated the ω ratios with F3x4 codon frequency and the branch-site model that Yang and Nielsen (2002) implemented in PAML v4 (Yang 2007). I used the tree topology inferred from single-copy genes (Figure 2.2a) as the underlying tree.

I calculated the average ω for each homolog with the model M0 (*model=0* and *NSsites=0*). To test whether a particular gene is under positive selection in one of the four spruces, I used the likelihood ratio test (LRT) to compare the goodness of fit of the likelihood scores from the PAML models MA0 (null hypothesis) and MA1 (alternative hypothesis). I tested each species as the foreground against the other species assigned to the background. MA0 assumes that the protein sites evolve neutrally or under purifying selection (*model=2*, *NSsites=2*, *fix_omega=1*, and *omega=1*) and calculates the ω as an average for the entire protein; MA1 assumes that protein sites are under positive selection in the foreground (*model=2, NSsites=2, fix_omega=0*). The LRT compares the likelihood scores from the two models as in:

$$LR = 2x(\ln(MA0) - \ln(MA1))$$

To determine whether the LR score is statistically significant, the difference is tested using the $\chi 2$ statistic with one degree of freedom (Yang and Reis 2011). I corrected for multiple testing by adjusting for the total number of tests with a false discovery rate (FDR) <0.01 (Benjamini and Hochberg 1995). Extreme ratios higher than 40 were removed from the average ω estimate but considered for the LRT significance testing.

2.3.7 GO term enrichment analysis

Gene families were annotated according to the most frequent Pfam domain. The significant gene families and homologous from positive selection were used for domain enrichment analysis in dcGO (Fang and Gough 2013) and the GoSlim ontology. The hierarchy of GO terms was inspected in QuickGO (Binns et al. 2009).

2.4 Results

2.4.1 Genome assemblies

The four annotated genomes have NG50 contiguity ranging from 38 kb in *P. sitchensis* to 355 kb in *P. engelmannii* (Appendix Table A.5). The total reconstructed genome assembly size was ~21 Gb for each spruce taxon, closely matching their estimated genome sizes (Appendix Figure A.2). The four genome assemblies have similar completeness as measured by the number of reconstructed "complete – single copy" BUSCO orthologs. The gene space completeness ranges from 29.9 to 41.1% "complete – single copy" BUSCO across the North American spruces (Appendix Table A.5). Other published conifers have a comparable BUSCO completeness, such as 28.9 and 48.7% BUSCO "complete – single copy" in *P. abies* and *P. lambertiana*, respectively (Appendix Table A.6).

2.4.2 Genome annotations

The genomes of *P. engelmannii*, *P. sitchensis*, *P. glauca*, and interior spruce have similar annotated genes (34,365, 30,324, 30,410, and 28,944, respectively; Table 2.1), with an average of two transcript isoforms per annotated gene. The median predicted mRNA length ranges from 1,140 bp in *P. sitchensis* to 1,038 bp in interior spruce. Large parts of the annotated transcripts lack untranslated regions (UTR), possibly due to fragmentary assembly, in which each gene is not fully contained in the assembled scaffold. I estimated the content and completeness of functional domains and the presence of putative single-copy genes with reference to BUSCO "complete" genes and DOGMA Conserved Domain Arrangements (CDA). The

percent of reconstructed CDA ranges between 32.01% and 33.56%, with the highest percent of Pfam CDA in *P. sitchensis*. BUSCO completeness ranges between 17.6% (*P. engelmannii*) and 18.2% (*P. sitchensis* and interior spruce). BUSCO completeness and Pfam CDA metrics follow a similar pattern for the annotated genes: *P. engelmannii* has the lowest percentage of annotated Pfams and BUSCO complete. *Picea sitchensis* has the lowest degree of genomic contiguity but has comparable annotation completeness to other genomes, based on Pfam and complete BUSCO genes.

The gene annotations have a similar degree of completeness for species with similar contiguity (Appendix Table A.6 and Table A.7). *Picea abies* and *Pinus lambertiana* have comparable Pfam completeness and the North American spruces. In contrast, *P. taeda*, which has a more contiguous genome assembly when compared to the other genotypes, has a more complete genome annotation than the others according to the BUSCO "complete" and total CDA completeness scores (Appendix Table A.7).

Table 2.1 Spruce genome annotation statistics: number of annotated genes and transcripts, median mRNA, exon and intron length, and corresponding annotation completeness. The length is shown as the median value. The gene median described in the table contains introns, while the mRNA median excludes the intron lengths. The annotation completeness is shown as BUSCO "complete" ("single copy" and "duplicated") core set genes and percent Pfam Conserved Domain Arrangements (CDA).

Taxon	Total genes	Total mRNA	Median gene length (bp)	Median mRNA length (bp)	Median exon length (bp)	Median intron length (bp)	BUSCO complete (%)	Total CDA completeness (%)
P. engelmannii	34,365	60,224	2,455	1,116	162	187	17.6	32.01
P. sitchensis	30,324	58,175	2,757	1,140	158	201	18.2	33.56
P. glauca	30,410	56,535	2,569	1,086	153	189	18.0	32.14
Interior spruce	28,944	62,397	2,718	1,038	161	195	18.2	33.36

2.4.3 Phylogenomics of gen. *Picea*

I compared phylogenetic relationships among the spruce species based on the whole plastid (pt) and mitochondrion (mt) genomes (Figures 2.2b,c,d) and on a coalescent-based analysis of single-copy nuclear genes (nc) (Figures 2.2a). The mt (Figure 2.2b) and pt-based phylogenies (Figure 2.2d) were generated with k-mers, and the pt-based phylogeny was generated with a standard maximum-likelihood method (Figure 2.2c). The plastid-based phylogenetic estimates differed from the nuclear and mitochondrial estimates concerning the relative positions of *P. abies* and *P. sitchensis* as respective sister groups of the remaining *Picea* taxa.

Picea engelmannii was inferred to be the sister group of *P. glauca* and interior spruce in the phylogenies inferred from all three genomes (Figures 2.2a,b,c). *Picea sitchensis* appeared as a sister group to *P. engelmannii* and *P. glauca*. Most relationships inferred by the nuclear data displayed substantial discordance among gene trees (at least 50% of gene trees discordant; pie charts in Figure 2.2a).

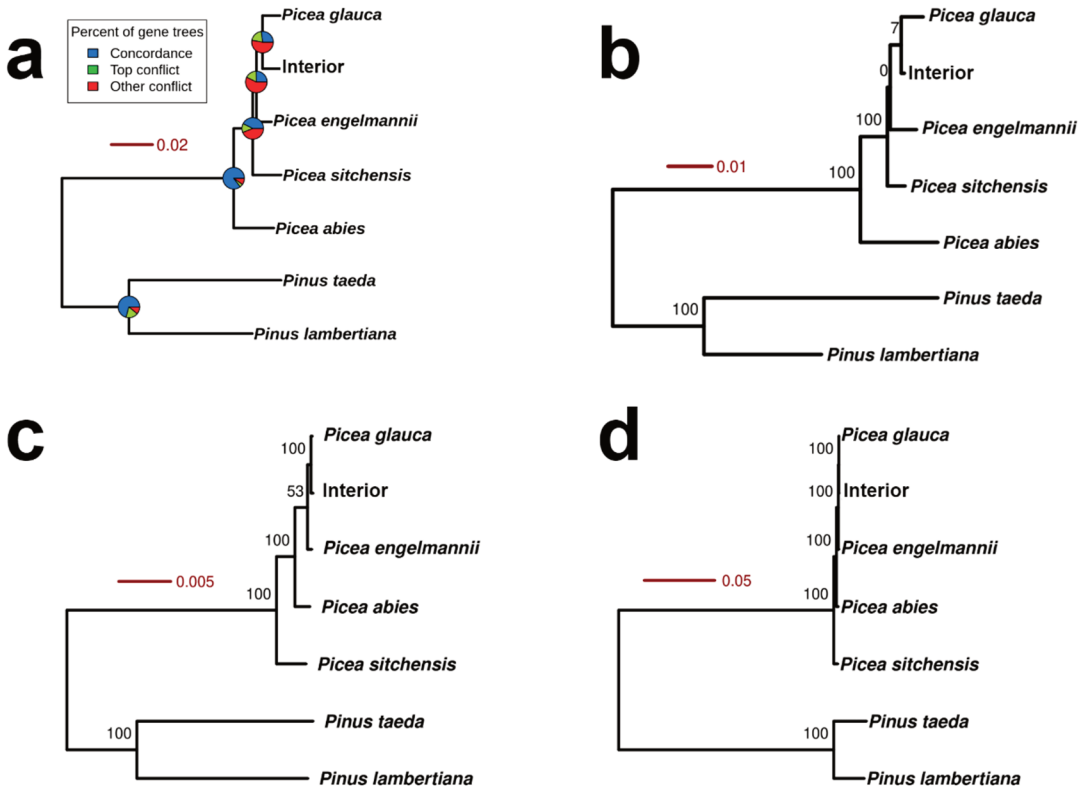

Figure 2.2 Genome-based inference of spruce phylogenies from (a) nuclear data based on a coalescent-based analysis of single-copy genes; pie charts show the proportion of gene trees supporting the species tree, estimated by *phyparts* and the percent of genes in conflict (blue = agreement between the gene and species trees; green = most common gene conflicting topology; red = all the other gene conflicting topologies). **(b) k-mer based estimate using mitochondrial genome** with jackknife estimates of branch support. **(c) k-mer based estimate using plastid genome** with jackknife estimates of branch support; **(d) maximum-likelihood estimate using plastid genome** with bootstrap estimates of branch support. Scales show the number of substitutions per site.

2.4.4 Gene gain/loss in gene families

Proteins derived from predicted genes are clustered in gene families for comparative analysis across spruce and pine taxa. A large fraction (88–90%) of genes in the four North American spruces is included in a gene family (Appendix Table A.8), with ~14,000 orthogroups being found per taxon. Out of a total of

22,397 orthogroups, 3,165 were shared between the different species, 907 were shared between all spruces and 1,215 between the North American spruces (Appendix Figure A.3).

I considered 9,464 gene families for the gene family expansion/contraction study. These were chosen because they contain representative genes from at least five species. I investigated the gene family expansion/contraction rate using the model implemented in CAFE (Han et al. 2013). Models that estimate the gene gain/loss may overestimate both types of changes in draft genome assemblies; together with the original gene families dataset (OGF) used in the study, I also created a filtered dataset (F50) to check the degree of overestimated gene families in the draft genomes compared in this chapter. The filtering strategy removed 21,713 putative misannotations from all the examined taxa, equaling ~11% of the total genes. The estimated average rate of change λ is +0.0085 for OGF after the error correction of 18.44% applied by CAFE. In contrast, the filtered dataset F50 yielded a slightly lower λ score of +0.0083, with a CAFE error rate correction of 10.95%. Because the unfiltered and filtered estimates are similar, this supports that the effect of misannotations on the estimate of λ is marginal. CAFE estimated global error rate implies that 10–18% of the gene families have an incorrect estimate of the number of component genes.

I identified gene families evolving at significantly different rates between parent and child nodes in the phylogeny compared to the stochastic gene birth/death in CAFE (Figure 2.3a; Appendix Table A.9). The software estimated that a total of 856 gene families are significant or rapidly evolving across pine and spruce species, with 517 gene families having a gene expansion and 339 gene contraction. The North American spruces, *P. sitchensis*, *P. engelmanni*, *P. glauca*, and the interior spruce have 463 rapidly evolving gene families with a gene-family expansion/contraction in at least one branch and 334 unique gene families. I observed a substantial loss of genes in gene families along the stem leading to all *Picea* except *P. abies*. A significant loss of 85 gene families occurred in the branch leading to *P. abies*. North American spruces have a large number of expanding gene families. In particular, *P. engelmannii* and *P. sitchensis* experienced a large number of expanding gene families, +110 and +52, respectively.

The species with the largest total number of rapidly evolving gene families is the interior spruce with 153 (+72/-81), followed by *P. engelmannii*, *P. glauca*, and *P. sitchensis* with 130 (+110/-20), 115 (+59/-56), and 65 (+52/-13) gene families, respectively.

I used the number of genes in rapidly evolving gene families to cluster the seven conifer taxa and define common gene family gain/loss events (Figure 2.3b). *Picea glauca*, *P. sitchensis*, and *P. engelmannii* clustered together, with the interior spruce on the outside of this cluster of North American spruces. Four gene family clusters (highlighted in Figure 2.3b) show distinctive gene family expansion in, respectively, (1) interior spruce, (2) *P. sitchensis*, (3) *P. glauca*, and (4) *P. engelmannii*. *Picea abies* and *P. lambertiana* had a similar gene gain/loss profile. The species with the most divergent profile of gene gain/loss is *Pinus taeda*, likely due to an artifact in its annotated genes, which differ from its sister species *P. lambertiana* (47,602 and 31,253, respectively): the number of annotated genes in *P. taeda* is not reported with the high-quality tag as the other species (Appendix Table A.7), which inflates the number of genes used in the comparison.

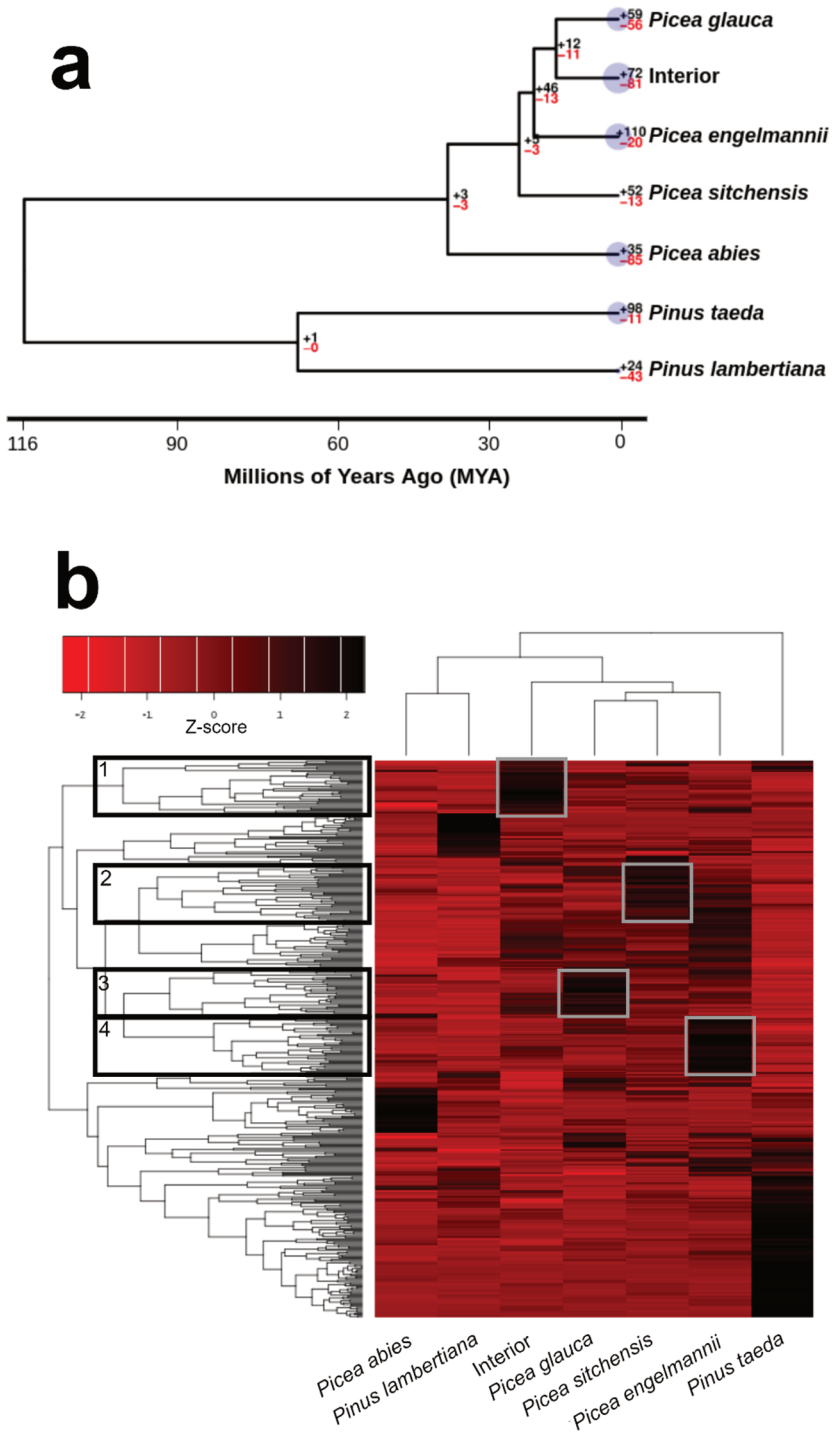

Figure 2.3 Gene family gain/loss computed by CAFE analysis. (a) ultrametric species tree (chronogram) for North American and Norway spruces (*P. abies*) + pines outgroup, and rapidly evolving gene families on tree

nodes. The numbers on the tree nodes indicate the number of rapidly evolving gene families with gene expansion (+, black) and contraction (-, red) compared to the parent node. The blue circle size on the tree tips is proportional to rapidly evolving gene families (expansion and contraction). **(b) Heatmap of rapidly evolving gene families (y-axis) showing the number of genes in each gene family.** The conifer taxa (x-axis) are clustered based on the number of genes in the rapidly evolving gene families. Four gene family clusters are highlighted for the North American spruces, showing expanding gene families (black) in (1) interior spruce, (2) *P. sitchensis*, (3) *P. glauca*, and (4) *P. engelmannii.* The expanded gene families are highlighted in the heatmap (gray squares) corresponding to the gene family clusters on the left (1–4).

2.4.5 Positive selection in orthologs

The average dN/dS ratios (ω) were estimated for 726 ortholog genes across the four North American spruces. The ω values (Figure 2.4) range between 0.0001 and 4.83, with a median value of 0.399, indicating a high degree of purifying selection acting on most codons. The dN and dS values showed a median score of 0.00660 and 0.01665, respectively.

I tested the ω-values in orthologs for significance, considering each foreground species branch in turn, using LRTs. The test compares the goodness of fit of the null hypothesis MA0 PAML model that assumes neutral or purifying selection against the alternative hypothesis MA1 PAML model that assumes positive selection. A total of 32, 62, 50, and 46 orthologs were significant (P <0.001) for *P. engelmannii*, *P. sitchensis*, *P. glauca,* and interior spruce, respectively.

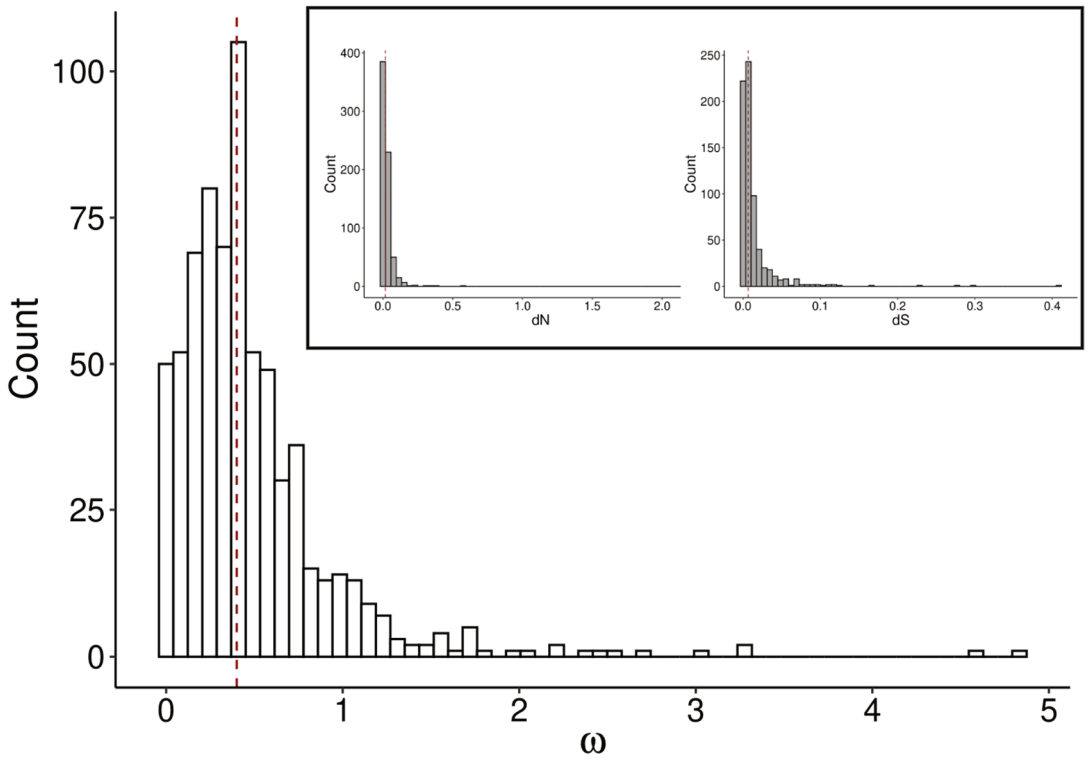

Figure 2.4 Histogram of ω ratios and dN, dS scores calculated for orthologs in the four North American spruces. The ratios for orthologous genes are used for the positive selection study. The dashed line shows the median values of ω, dN, and dS.

2.4.6 GO term enrichment analysis

I analyzed 334 unique gene families that experienced significant rapid gene turnover in the four North American spruces. I focused on the gene families with expanding genes for each taxon which is the most common mechanism of gene turnover in the analyzed taxa (Appendix Table A.9). Additionally, as shown in Figure 2.3b, the expanding gene families represent the strongest clustering pattern for the North American spruces.

The most common biological processes in expanded gene families are as follows (summarized in Table 2.2): 1) response to endogenous stimulus, 2) response to abiotic stimulus, 3) response to organic substance, 4) reproduction, and 5) multi-organism process. *Picea engelmannii* and *P. glauca* genotypes show a unique set of biological processes such as “catabolic process” and “response to stress” for the

former and “macromolecule localization” for the latter. *P. sitchensis* is the unique taxon with “response to external stimulus.”

I analyzed the significantly enriched molecular processes and looked at the Pfam domains linked to GO terms (Appendix Tables A.10–A.13 and A.14–A.17). The terms unique to *P. engelmannii* describe post-transcriptional gene silencing (“regulatory RNA binding” and “single-stranded RNA binding”) and contain Pfam domains such as Argonaute linker 1 and 2, PAZ (Piwi, Argonaut, and Zwille), and Piwi domains. The GO term of “ubiquitin-protein transferase activity” is again enriched uniquely in the top *P. engelmannii* molecular functions and describes protein degradation; Pfam domains linked to protein degradation are BTB and C-terminal Kelch domain, Zinc finger C3HC4 amino acid motif, RING finger domain, and kinetochore protein required for cell cycle progression (Skp1) dimerization and tetramerization domains. The unique molecular processes in *P. glauca* are related to DNA repair. Pfam domains found in this class are the OB-fold domains responsible for DNA replication, DNA recombination, and DNA repair, and the Bromodomain, which is essential for chromatin binding. Other terms, such as carbohydrate derivative binding and protein-containing complex binding, contain diverse Pfam domains that describe the interaction of ATP or nucleotides with large protein complexes, such as Hsp70 heat-shock proteins, alpha/beta-hydroxylase family, and oxidoreductases P450.

Table 2.2 - Top GO molecular functions and biological processes enriched in rapidly evolving gene families in the North American spruce species. The ranking is based on *P*-values calculated by GO terms enrichment analysis, displaying the ranked top common GO terms for each genotype. The colors highlight the common among multiple taxa GO-terms; the highlighted GO-terms are unique to the specific taxon. Terms that show a relationship with other terms in the table are marked with →, listed together with relation type and related terms.

GO domain	*P. engelmannii*	*P. sitchensis*	*P. glauca*	Interior
Biological process	1. response to organic substance	1. response to endogenous stimulus	1. reproduction	1. response to endogenous stimulus
	2. response to stress	2. response to organic substance	2. response to abiotic stimulus	2. response to organic substance
	3. catabolic process	3. response to abiotic stimulus	3. response to endogenous stimulus	3. response to abiotic stimulus
	4. response to endogenous stimulus	4. multi-organism process	**4. macromolecule localization**	4. reproduction
	5. reproduction	**5. response to external stimulus**	5. multi-organism process	5. multi-organism process
Molecular function	**1. regulatory RNA binding** *→is an RNA binding* *→is a nucleic acid binding*	1. small molecule binding	**1. single-stranded DNA binding** *→is a DNA binding* *→is a nucleic acid binding*	1. small molecule binding
	2. single-stranded RNA binding *→is an RNA binding* *→is a nucleic acid binding*	2. nucleoside phosphate binding	**2. damaged DNA binding** *→is a DNA binding* *→is a nucleic acid binding*	2. nucleoside phosphate binding
	3. ubiquitin-protein transferase activity	**3. molecular transducer activity**	**3. ribonucleoprotein complex binding** *→is a protein-containing complex binding*	3. carbohydrate derivative binding
	4. catalytic activity, acting on RNA	4. carbohydrate derivative binding	4. protein-containing complex binding	4. protein-containing complex binding
	5. protein-containing complex binding	**5. ion binding**	**5. nucleic acid binding**	**5. enzyme binding**

Figure 2.5 shows the significant Pfam domains divided into major functional classes with respect to their presence in the expanding gene families in the four spruce taxa. Transcriptional factors such as Apetala 2 are key regulators of several abiotic and multiple hormone responses in plants. As AP2 covers a large set of abiotic type responses (Lata and Prasad 2011; Mizoi, Shinozaki, and Yamaguchi-Shinozaki 2012; Licausi, Ohme-Takagi, and Perata 2013; Phukan et al. 2017), including cold, drought, and heat, it plays an important molecular role under all types of environmental conditions, which can explain why it is in expanding gene families in the majority of spruce genotypes. The other transcription factors, containing Homeobox and the TCP domains, are found in expanding gene families in *P. sitchensis* and *P. glauca*, and *P. glauca*, respectively.

The Hsp70 is an important molecular chaperone that interacts with newly translated proteins for correct protein folding (Mayer and Bukau 2005); expanded gene families described as Hsp70 are uniquely found in *P. sitchensis* and interior spruce, while the Hsp20/alpha-crystallin is found in *P. engelmannii* and interior spruce. Other Pfam domains found in multiple biological processes that are uniquely expanded in *P. engelmannii* and *P. glauca* include Cytochrome P450, which is involved in terpenoid biosynthesis and the oxidation of a variety of compounds.

The protein domain of ubiquitin family and Leucine-rich repeats are significantly expanded in gene families across all the spruce taxa, similar to the pentatricopeptide repeat (PPR) that is expanded in all four taxa except interior spruce. Two zinc-finger domains, C3HC4 and C2H2, cover several fundamental functions. Specifically, the RING finger C3HC4 mediates the ubiquitination of proteins (Joazeiro and Weissman 2000) are exclusively found in expanded gene families in the *P. engelmannii*, while the zinc finger C2H2, which is known to regulate the abiotic stress in plants (Han et al. 2020), is found in to be expanded in *P. glauca*.

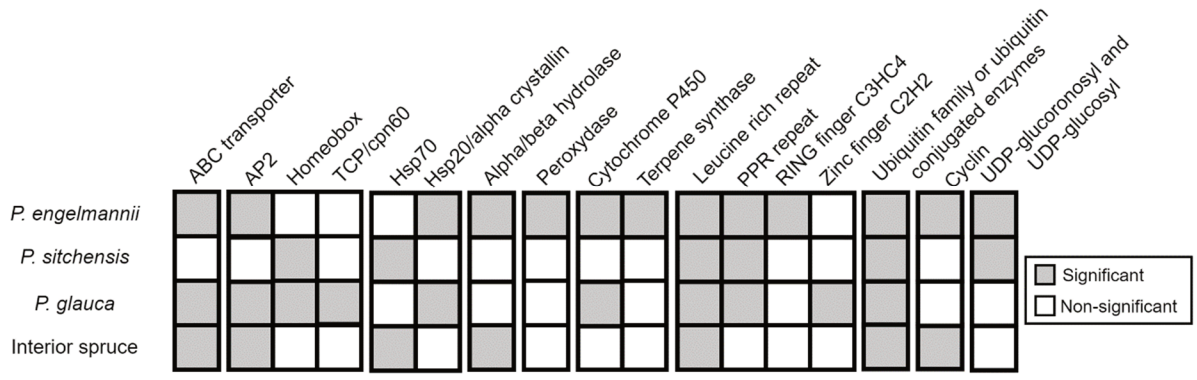

Figure 2.5 Significant Pfam domains in rapidly evolving gene families. Relevant Pfam domains are divided based on their major functional class (left to right): transporters, transcription factors, heat shock proteins, hydrolases, peroxidases, terpenoids, protein repeated domains, protein degradation, cell cycle, and small molecules glycosylation. Gray boxes indicate significant Pfam domains.

A total of 190 single-copy orthologs showed significant positive selection in at least one North American spruce. These were used for GO term enrichment analysis based on Pfam domains (Appendix Tables A.18–A.21 and A.22–A.24). The GO terms that are the most common among the biological process (Table 2.3) are 1) response to endogenous stimulus, 2) response to abiotic stimulus, 3) reproduction, 4) response to organic substance, and 5) embryo development which in part recapitulates the GO terms found for the expanding gene families (Table 2.2). The GO term "response to stress" is unique in *P. engelmannii*, "regulation of biological quality" and "response to abiotic stimulus" in *P. sitchensis*, regulation of biosynthetic process" in *P. glauca,* and "negative regulation of metabolic process" in interior spruce.

Table 2.3 - Top GO molecular functions and biological processes, enriched in orthologs with positive selection in the North American spruce species. The ranking is based on *P*-values calculated by GO terms enrichment analysis, displaying the top common GO terms for each genotype. The colors highlight the common among multiple taxa GO-terms; the highlighted GO-terms are unique to the specific taxon. Terms that show a relationship with other terms in the table are marked with →, listed together with relation type and related terms.

GO domain	*P. engelmannii*	*P. sitchensis*	*P. glauca*	Interior
Biological process	1. response to endogenous stimulus	1. response to endogenous stimulus	1. response to abiotic stimulus	1. reproduction
	2. reproduction	2. embryo development	**2. regulation of biosynthetic process**	2. response to abiotic stimulus
	3. response to abiotic stimulus	**3. regulation of biological quality**	3. reproduction	3. response to endogenous stimulus
	4. response to stress	4. reproduction	4. response to endogenous stimulus	**4. negative regulation of metabolic process**
	5. response to organic substance	5. response to abiotic stimulus	5. response to organic substance	5. embryo development
Molecular function	1. enzyme binding	1. enzyme binding	**1. transcription regulator activity**	-
	2. transcription regulation activity	**2. protein serine/threonine/ tyrosine kinase activity**	**2. kinase regulator activity**	-
	3. calmodulin binding	**3. kinase binding** *→is an enzyme binding*	**3. enzyme activator activity**	-
	-	-	**4. methyltransferase activity**	-
	-	-	**5. nucleic acid binding**	-

Figure 2.6 shows the major Pfam domains found as significant from the positive selection analysis and their significance in the four annotated spruce taxa. A significant finding of positive selection is the enrichment in transcriptional factor (TF) domains in the "response to abiotic stimulus" that are known to be involved with plants' abiotic response, such as cold, drought, salt, heat, and freezing. The Auxin/Indole-3-acetic acid (AUX/IAA) protein domains are encoded by the auxin gene family, which is involved in the plant's growth and development. It is under positive selection in all the described genotypes, except for *P. sitchensis*. A set of TFs that regulate growth and development is the Myeloblastosis-like (Myb-like) and Homeobox domains, significant in all the described genotypes. I also found protein domains involved in the chromatin remodeling and modulation of the DNA transcription accessibility; among those, the most common in all the genotypes are the PWWP DNA methyltransferase (named after its central amino acid core motif Pro-Trp-Trp-Pro), high-mobility group (HMG) proteins and helicases.

One TFs domain found exclusively under positive selection in *P. engelmannii* is the basic leucine-zipper (bZIP) domain, which is a large class of TFs involved in several abiotic responses such as cold, drought (Kang et al. 2002), waterlogging, osmotic, and salt stress (Hartmann et al. 2015). *Picea sitchensis*, a species that grows in locations with high rain annual precipitation shows selection for Ankyrin repeats (ANK) proteins, known to confer drought tolerance in *Arabidopsis* (Zhao et al. 2020).

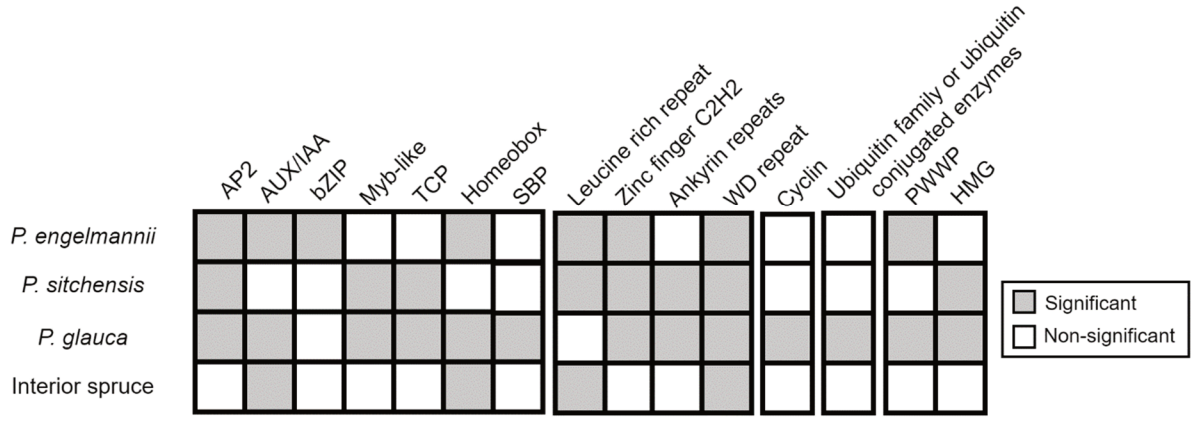

Figure 2.6 Significant Pfam domains in orthologous genes with positive selection. Relevant Pfam domains are divided based on their major functional class (left to right): transcription factors, protein repeat domains, cell cycle, protein degradation, and chromatin remodeling. Gray boxes indicate significant Pfam domains.

2.5 Discussion

2.5.1 Genome annotation of spruce giga-genomes

The task of genome annotation (identifying and describing regions in the genome that are of putative biological interest) is challenging. In the case of spruce giga-genomes, the task is even more challenging given their large genome size and generally high number of genomic scaffolds. To address this major issue, I designed and tested an improved version of the MAKER pipeline for large and fragmented genomes to make the annotations faster. When a large number of scaffolds are being processed independently, this can substantially increase the run time of MAKER, taking up to several months to annotate a single spruce genome. Spruce genomes are rich in repeat sequences, and software such as RepeatMasker (Smit, Hubley, and Green 2008) embedded in MAKER, which annotates repeat regions using a user-provided library, required a significant run time in spruce. Another challenging genomic feature of conifers is the presence of numerous gene deserts or the large fraction of intergenic regions that were estimated to be dominant in the annotated spruce genomes. I tested the ability of scaffold filtering to speed up the annotation process, using only scaffolds with low repeat content. Another filtering strategy

aims to remove gene deserts and excludes scaffolds lacking complete genes, which did not appear to result in loss of gene annotations. Upstream data filtering, together with an efficient data binning of the remaining scaffolds, brings the computational time down to about 120 hours per bin on a single high-performance computing (HPC) cluster node (Slurm Workload Manager, -mem=350G, -ntasks=24). I recommend the genome-filtering strategies employed here as a preparation step for conifer annotations. This step in the annotation can greatly prepare for a targeted annotation of selected genomic scaffold to avoid (a) misannotated genes due to gene cleavage and (b) long run time for annotation pipelines.

Pre-trained gene models in the common annotation pipelines have several limitations in conifers. Conifer genomes are unique in their gene structure. A well-known gene feature of repetitive genomes such as those of conifers is their large intron length, which can reach several kb in length and even up to Mb, as recently discovered in the genome of giant sequoia (*Sequoiadendron giganteum*) (Scott et al. 2020). Based on an analysis of the gene structures (exon-intron organization) of 35 genes in *P. glauca* (Stival Sena et al. 2014), intron length is a significant driver for genome structure reorganization in *P. glauca* and accounts for the large variability in ortholog genes. Using pre-trained gene models is not optimal for annotating all the gene structures as those with long introns, as shown previously for spruce (Warren et al. 2015). I addressed this limitation by performing a pre-training step using the Embryophyta core gene dataset in odb9 as a pre-run and then running the MAKER annotation pipeline with an iterative approach (Appendix Figure A.1). I used results from a preliminary MAKER run to train the gene models in the second stage of annotation, which generated spruce species-specific gene models. The advantage of this training is observed in annotations of the most contiguous genomes, as in the case of *P. engelmannii*, where I predicted longer intron sizes (>250 kbp) for specific genes typical of conifers.

Annotation of large and fragmented genomes, such as the spruce genomes studied in this study, remains challenging. While the fragmented nature of the draft genomes is the most challenging issue, genome assembly errors and scaffolding gaps are also problematic. I considered those challenges in the study of spruce by filtering fragmented and possibly erroneous annotations; I added several filtering

criteria that consider both the integrity of the gene structure and the functional annotation evidence. Although this filtering approach is likely to miss some gene annotations, it specifically selects genes with strong support from a gene sequence or high similarity with known genes. I only considered for comparison orthologs common to all the species examined, for example, in positive selection analysis. Additionally, I performed error quantification in gene family expansions to estimate the degree of misannotations. Overall, I am confident that the careful gene selection strategy applied is appropriate for controlling the significant annotation biases known to be problematic in draft genome assemblies (Florea et al. 2011; Han et al. 2013).

2.5.2 Phylogeny of gen. *Picea*

Genome annotations are fundamental for discovering single-copy genes for comparative genomics and the study of spruce evolution. Those annotated here include a set of single-copy nuclear genes, the largest dataset of nuclear genes (n=246) used to determine the evolutionary relationship between spruces. The species tree was estimated with ASTRAL-III under the coalescent-based method. Genomic conflict emerging from the nuclear phylogeny in Figure 2.2a contains information about the fragmented speciation events that may have been influenced by incomplete lineage sorting (ILS) or introgression. The species-tree phylogeny by ASTRAL-III corrects for such ILS events (Zhang et al. 2018) and handles introgression (Meleshko et al. 2021; Schley et al. 2020) as in the interior spruce – PG29 genotype. *Picea engelmannii* was inferred to be the sister group of *P. glauca* and interior spruces in the phylogenies inferred from plastid, mitochondrion, and single-copy genes. The large number of conflicting single-copy genes from the nuclear phylogeny in Figure 2.2a indicates the close phylogenomic proximity of *P. engelmannii* and *P. glauca*, which naturally hybridize in their large area of sympatry.

The organellar genomes presented in the spruce study showed a mostly consistent topology with the earlier reports that used few genes from plastid (He et al. 2012) and mitochondrion (Lockwood et al. 2013). One possible scenario that justifies the discordant topology of the plastid when compared to the

other phylogenies is that the paternally inherited plastid in this genus (Adams 2019) may lead to topologies different from nuclear data due to ancient reticulation events, likely affected by long-distance gene flow driven by pollen dispersal (Bouillé, Senneville and Bousquet 2011). Sequencing the genomes of the other spruces endemic to Canada, such as *P. mariana* (black spruce) and *P. rubens* (red spruce), which have overlapping geographical ranges with the spruces in this study, is needed to better characterize interspecies relationships in the genus.

2.5.3 Protein domains of expanded gene families and genes under positive selection

The GO-term enrichment analyses for gene families experiencing expansion and orthologs with positive selection indicate that a dynamic set of biological processes may contribute to these phenomena, including responses to abiotic and endogenous stimuli. It is helpful to compare our results with existing spruce association studies that are performed on spruce populations (Hornoy et al. 2015; Yeaman et al. 2016; Holliday, Ritland, and Aitken 2010; Prunier et al. 2011; Prunier, Verta, and MacKay 2016; Hamilton, De la Torre, and Aitken 2015). Among the molecular markers that emerged from these studies and indicate local adaptation, I annotated protein domains such as AP2, AUX/IAA, Myb-like, and Homeobox, involved in overall plant development, including seed development and plant growth (Hornoy et al. 2015; Yeaman et al. 2016; Prunier et al. 2011). AP2 and AUX/IAA protein domains, in particular, are considered to be enriched in studies of local adaptation in several conifer species, such as *Picaa glauca*, *Picea mariana*, and *Pinus contorta* (Hornoy et al. 2015; Yeaman et al. 2016; Prunier et al. 2011). Ubiquitin-related proteins are also enriched in studies of local adaptation and have signatures of positive selection in spruce population studies (Pavy et al. 2013; Prunier et al. 2011); this gene family regulates protein degradation and turnover and is significantly expanded in all the taxa and under positive selection in *P. glauca*. Other significant protein domains cover the genetic and epigenetic transcription control functions, as for the core histone H2A H2B H3 H4 (Pavy et al. 2013; De La Torre et al. 2015;

Hornoy et al. 2015), PWWP methyltransferase, and HMG domains. The latter two are under positive selection in *P. glauca* and *P. sitchensis*.

The Leucine-rich repeat (LRR) gene family is expanded ubiquitously across the North American spruces. LRR are associated with plant resistance and adaptative selection (Richter and Ronald 2000); LRR gene clusters are known to evolve more rapidly than other regions in the genome in part due to their association with TE. The annotation of resistance genes is tightly connected to repeat masking approaches, as reported by Bayer, Edwards, and Batley (2018), and the number of annotated genes can be influenced by the repeat library used for annotation. It is advised to re-annotate the LRR genes with repeat free annotation strategies to corroborate the gene family expansion of LRR genes and evaluate the findings considering eventual repeat masking bias.

2.5.4 Future directions

The genomes and annotations generated in this study provide curated reference genomes and annotations for population-based studies for natural conifer populations. Sequencing additional individuals from allopatric and introgressed spruce individuals will be helpful to tease apart information on intraspecies variability (not addressed here) and local adaptation. In particular, whole-genome sequencing of multiple additional individuals at low coverage can genotype wild-type trees and identify single nucleotide variants (SNV). Additionally, population studies can study the genomic traits involved in a local adaptation by combining the identified SNVs with information on abiotic conditions, such as temperature, precipitation, and soil salinity. In addition, whole-genome sequencing of interior spruce individuals should be a powerful tool for better characterizing the evolution of naturally occurring sympatric populations of spruce in British Columbia, which constitute an elite tree in managed forests. For example, the search for genome-scale SNVs in the population of interior spruce characterized by different degrees of cold temperature tolerance may help select genomic markers related to determining elite genomic makeup.

Chapter 3: Genome assembly and annotations of *Pissodes strobi*, a North American forest insect pest

3.1 Author summary

The highly diverse insect family of true weevils, Curculionidae, includes many agricultural and forest pests. *Pissodes strobi* is one of the major pests of spruce and pine forests in North America. Here I present the draft genome assembly of *P. strobi* and its gene and repeat annotations. I compare the genome of *P. strobi* to eight Curculionidae genomes that have been recently sequenced. The *P. strobi* genome has a substantial increase in genome size compared to other Curculionidae, likely driven by a recent expansion of transposable elements (TE). This chapter addresses objectives 1 and 2.2 in section 1.4.

3.2 Introduction

Beetles (Coleoptera) are the largest order of insects, representing more than 400,000 extant species (Stork et al. 2015). The beetle family Curculionidae (true weevils) is a heterogeneous taxon with more than 60,000 species that are primarily phytophagous or feeding on plants and include some of the world's most disruptive forest and agricultural pests (McKenna et al. 2009; Oberprieler, Marvaldi, and Anderson 2007). Curculionidae evolved initially in close associations with gymnosperms, including conifers, during the Jurassic (200–145 MYA) and later coevolved with angiosperms in the Cretaceous (145–66 MYA). A key innovation of the Curculionidae colonization (McKenna et al. 2009; S.-Q. Zhang et al. 2018) is the development of an extended apical rostrum used to excavate oviposition cavities directly in plant tissues.

Pissodes strobi (Curculionidae) is a destructive pest of conifers in North America. Its annual life cycle is divided into two major phases, exophase and endophase (Alfaro 1994; Whitehill and Bohlmann 2019). During exophase, adult weevils live outside the tree and feed on its bark without causing substantial damage to the host. The endophase takes place after the female deposits its eggs into oviposition holes carved with the insect's apical rostrum; the egg, larvae, and pupal stages occur inside

the host tree. Larvae feed basipetally inside the tree, through the cortex, phloem, cambium, and outer xylem of the apical shoot. This disrupts the flow of water and nutrients and leads to apical shoot mortality, resulting in stunting of trees and deformed growth. The weevil is most destructive during endophase, which continues until the pupation and emergence of adult insects from the tree. Repeated infestation over multiple years can result in tree death (Gara and Wood 1989).

We have sequenced the genome of *P. strobi* and present it here as a reference genome for future population and evolutionary studies. *Pissodes strobi* has a broad geographic and infestation range, with several distinct populations that have long been considered separate species, until their recognition as a single species based on multiple lines of genetic evidence (Phillips and Lanier 2000; D. W. Langor and Sperling 1997; Laffin, Langor, and Sperling 2004; David W. Langor and Sperling 1995; Smith and Sugden 1969). Common names for *P. strobi* differ in geographic origin and host association (Laffin, Langor, and Sperling 2004). For instance, in eastern North America, *P. strobi* is commonly known as the white pine weevil as its primary host in that geographic location is the eastern white pine (*Pinus strobus*). In western North America, it is referred to as spruce weevil as its primary hosts there are Sitka (*Picea sitchensis*), white (*Picea glauca*), Engelmann (*Picea engelmannii*) spruce, as well as interior spruce (*P. glauca x engelmannii x sitchensis*), the genomes of which I presented in Chapter 2.

Sequenced Curculionidae genomes are fast-growing, with eight new reference genomes from forest and agricultural pests released in 2020 and 2021. Available genomes include the coffee borer beetle, *Hypothenemus hampei* (Navarro-Escalante et al. 2021), the Argentine stem weevil, *Listronotus bonariensis* (Harrop et al. 2020), the red palm weevil, *Rhynchophorus ferrugineus* (Hazzouri et al. 2020), the oil palm pollinating weevil, *Elaeidobius kamerunicus* (Apriyanto and Tambunan 2021), the mountain pine beetle, *Dendroctonus ponderosae* (Keeling et al. 2013), the Easter egg weevil, *Pachyrhynchus sulphureomaculatus* (Van Dam et al. 2021), the Eurasian spruce bark beetle, *Ips typographus* (Powell et al. 2020), and the rice weevil, *Sitophilus oryzae* (Parisot et al., 2021). Here, I

report the *P. strobi* genome, its annotations, and phylogenetic comparison to other sequenced Curculionidae pest species.

3.3 Methods

3.3.1 Sample collection and sequencing

Pissodes strobi is challenging to rear from eggs in the laboratory. We isolated fourth instar larvae from the apical shoot tip of the interior spruce tree (*P. glauca* x *engelmannii* x *sitchensis*) on 6 May 2013 (Whitehill et al. 2016). The tree is located at the BC Ministry of Forests' Kalamalka Research Station, Vernon, British Columbia, Canada, 50 24'N: -119 28'W. The larvae were reared on a semi-artificial diet containing a biostatic (methylparaben) and an antifungal (sorbic acid) to reduce potential surface contaminants (Whitehill et al. 2016) for two weeks until they entered the pupal stage. As Coleoptera larvae void their gut before pupation, this diet minimizes gut-associated microbial sequences (Keeling et al. 2013). An individual pupa was selected for genome sequencing. The sex of the pupa could not be determined prior to DNA isolation. We isolated high molecular weight (HMW) DNA and prepared and sequenced it following Taylor et al. (2018). Specifically, in 2018, we sequenced a library comprising Chromium 10x linked reads on an Illumina HiSeqX sequencer, using the paired-end (PE) protocol. The library produced 831 million PE reads with a length of 150 bp.

3.3.2 *In silico* genome complexity estimate

I characterized the ploidy level of *P. strobi* from genomic reads with Smudgeplot v0.2.1 (Ranallo-Benavidez, Jaron, and Schatz 2020); there were no prior studies of ploidy for the organism. After extracting the barcode information, I trimmed the reads to remove bases with Phred quality score <=30 using Cutadapt v3.4 (Martin 2011). 21-mers were counted with KMC v3.1.1 (Kokot, Dlugosz, and Deorowicz 2017), with the argument *-k21 -ci1 -cs10000*. I used the *smudgeplot cutoff hetkmers* function to extract heterozygous k-mers between two cutoffs (lower, L, and upper, U, coverages, of L=20 and

U=790) and represented k-mer smudges using a *smudgeplot plot* to determine its ploidy level. I inferred genome features (genome size, heterozygosity, and repetitiveness) using GenomeScope v2.0 (Ranallo-Benavidez, Jaron, and Schatz 2020), assuming the ploidy determined by smudgeplot. I computed k-mer frequency histograms for all the genomic reads with ntCard v1.1.0 (Mohamadi, Khan, and Birol 2017) for k-mer lengths of 21, 23, 25, 27, and 29 bp.

Genome size was estimated for *Elaeidobius kamerunicus* (Apriyanto and Tambunan 2021), which we identified as the closest Curculionidae genome to *P. strobi*. The reads from *E. kamerunicus* are downloaded from the accessions SRR12726955–SRR12726958. K-mer frequency histogram is computed with ntCard (Mohamadi, Khan, and Birol 2017) for k-mer length 21 bp. All the genomic reads are included in the estimate, having an expected genome coverage of 170-fold based on genome assembly reconstruction. The genome size is estimated with GenomeScope2.0 (Ranallo-Benavidez, Jaron, and Schatz 2020), assuming a diploid genome.

3.3.3 Experimental genome size estimate by flow cytometry

We estimated the genome size of *P. strobi* by quantifying the amount of DNA contained in haploid nuclei (C-value) using flow cytometry (Johnston, Bernardini, and Hjelmen 2019). We used DNA standards estimated for fluorescence values at 2C and 4C of the *D. melanogaster* genome size. The standard includes six *D. melanogaster* biological replicates, with three biological replicates for male and three female standards. We converted the genome size to base pairs using the conversion factor of 978 Mb/pg DNA (Dolezel et al. 2003).

3.3.4 Genome assembly

The national center for biotechnology information (NCBI) contamination screening revealed a substantial number of sequences from *Wolbachia* bacterial spp. To remove these sequences from the genome assembly of the *P. strobi*, we filtered reads using a k-mer-based approach. *Wolbachia* spp. whole

genomes were accessed from NCBI GenBank as of December 16, 2019, and the sequences were loaded into a Bloom filter using BioBloomTools v.2.3.2 (Chu et al. 2014). Reads without matches to the filter were used in the *P. strobi* genome assembly using Supernova™ v2.1.1 (Weisenfeld et al. 2017) at different genome coverages (from 22-fold to 53-fold), considering the maximum number of reads that provided the highest N50 value. A haploid assembly was generated using the *-pseudohap* argument *mkoutput* in Supernova. To remove heterozygosity-induced duplicated scaffolds, I ran the Purge Haplotigs pipeline v1.0.4 (Roach, Schmidt, and Borneman 2018) with the argument *-l 1 -m 60 -h 200 -a 70*. We then used Tigmint v.1.1.2 (Jackman et al. 2018) with the options *-s 20 -w 1000* to identify possible misassemblies to break the draft assembly in regions with poor linked-read support. The genome was scaffolded with ARKS v1.0.1 (Coombe et al. 2018), using the argument *-c 5 -k 30 -j 0.55 -l 0 -d 0 -e 30000 -r 0.05*. Finally, we performed an iterative gap filling in Sealer v2.2.3 (Paulino et al. 2015) with k-mer sizes of 90, 100, 110, and 120 bp.

3.3.5 Annotation of protein-coding genes

I annotated the genome of *Pissodes strobi* with supporting evidence from cDNAs and transcriptome assemblies from Endopterygota species (taxon id: 33392), downloaded from NCBI, and with proteins from *Drosophila melanogaster* (Appendix Table B.1). The redundancy of the downloaded transcripts was removed through CD-HIT-EST v4.8.1 (Fu et al. 2012) run with the *-c 0.98* and *-n 10* parameters. Additionally, I assembled short RNAseq reads from thirty *Dendroctonus ponderosae* libraries (SRR1702878–SRR1703019) with a pooled assembly approach in RNA-Bloom v1.0.0 (Nip et al. 2020). I screened the assembled transcripts for biological contamination, and only transcripts with putative CDS were included in the annotation, selected through EvidentialGene v2017.12.21 (Gilbert 2013).

I used the MAKER v2.31.10 annotation pipeline (Holt and Yandell 2011), with annotation limited to genomic scaffolds longer than 1 kbp. I generated meta parameters for AUGUSTUS v2.5.5 (Stanke et al. 2008) by retraining the gene models with BUSCO v3.1 with *-long* option and the

Endopterygota core gene set odb9. I trained SNAP v2013-11-29 (Korf 2004) with high-quality gene models generated by a preliminary MAKER run with cDNA and *D. melanogaster* proteome evidence. I re-trained GeneMark v2.3c (Lomsadze et al. 2005) as GeneMark-ES, where the algorithm meta parameters are trained automatically by the software itself. Repetitive elements were identified with the repeat library described below and used as a customized library during the annotation process in MAKER. I used the EnTAP v0.92 (Hart et al. 2020) package to functionally annotate the proteins predicted by MAKER. enTAP was provided with two databases; NCBI RefSeq 99 (Pruitt et al. 2012) and Swiss-Prot/TrEMBL ("UniProt: The Universal Protein Knowledgebase in 2021" 2021), downloaded in April 2020. The annotated genes were filtered based on all the following criteria: (a) a minimum eAED score of 0.5, (b) having a minimum fraction of splice sites as shown by the quality index tag (QI) of 0.5, and (c) being assigned with functional annotation, either a Pfam domain or a BLAST hit, in enTAP. This annotation dataset is further referred to as "total annotated." Annotated genes are included in the "high confidence" genes if (a) have splice sites supported by a canonical splicing motif (GT–AG, GC–AG, AT–AC at the donor and acceptor splice sites), (b) polyexonic genes have a minimum intron length of 10 bp, and c) having a complete CDS (start and stop codons).

3.3.6 Annotation and quantification of repeat elements

I built a repeat library from the assembled genome of each Curculionidae accession using the EDTA v1.8.4 pipeline (Ou et al. 2019) with *sensitive=1* and *annotate=1* options. The first option calls RepeatModeller v2.0.0 (Flynn et al. 2020) for *de novo* annotating repeat sequences that are not recovered by the other annotation software in the pipeline. I combined the repeat library with insect repeats from RepBase v22.08 (Bao, Kojima, and Kohany 2015) to yield a final custom library. The second option annotates the repeats in the genome with RepeatMasker v4.0.9 (Smit, Hubley, and Green 2008). Using the annotated repeats, I estimated the Kimura 2-parameter (K2P) sequence divergence within each repeat family using the calcDivergenceFromAlign.pl script from RepeatMasker.

I used RepeatExplorer v2 (Novák et al. 2013) to quantify the genomics repeats independently from the reference genomes. The pipeline compares the repeats represented by raw genomic reads. I used this pipeline because different assembly qualities characterize the presented genomes, and repeat sequences are notoriously difficult to assemble with short reads. Draft quality genomes may result in a lower fraction of complete repeats when compared to genomes with higher contiguity (Ou, Chen, and Jiang 2018). *Ips typographus* was not included in the RepeatExplorer analysis because of the lack of short-read libraries (Appendix Table B.2). I trimmed each read set with Cutadapt v3.4 (Martin 2011) and the options *-minimum-length 80*, *-maximum-length 80 -l 80* and *-q 10*. I excluded singletons (unpaired reads) that remained after trimming and randomly sampled non-overlapped PE reads equivalent to 0.3-fold genome coverage. I ran RepeatExplorer with two protocols: a comparative analysis (*-f* option) to annotate and quantify the clusters of repeats across Curculionidae and an individual sample protocol for *P. strobi*. I used the PE mode in both the runs with a minimal default overlap of 66 bp and annotated only read clusters with at least 0.01% of the input reads. I used the repeat library from EDTA (Ou et al. 2019), RepeatModeller v2.0.0 (Flynn et al. 2020), and RepBase v22.08 (Bao, Kojima, and Kohany 2015) as repeat databases in RepeatExplorer to improve the sensitivity of repeat classification, together with Metazoan repeat taxonomy v2.0 from REXdb (Neumann et al. 2019), and the database from TAREAN (TAndem REpeat ANalyzer) (Novák et al. 2017). I assigned a repeat cluster to a class if more than 33% of the reads were uniquely assigned. I manually inspected the resulting repeat annotations to check whether all the classes were correctly inferred. Clusters with reads that do not reach 33% in any class are assigned to "unknown" repeats.

3.3.7 Comparative genomics: Curculionidae phylogenomics

I inferred phylogenetic relationships based on single-copy orthologs from the "odb10" (Kriventseva et al. 2019) Endopterygota lineage, annotated with BUSCO v5.2.1 (Simão et al. 2015). I aligned protein sequences in each orthogroup using Mafft v7.453 (argument *-auto*) (Nakamura et al. 2018) and then used

RAxML v8.2.12 (Stamatakis 2014) to infer gene trees (PROTGAMMAAUTO and 100 bootstraps). Fragmentary genes, defined as sequences with gaps in more than 67% of the sequence length in the multiple-sequence alignment, were excluded from the phylogeny. I estimated a coalescence-based species tree using ASTRAL-III v5.6.3 (C. Zhang et al. 2018). The results were visually inspected with the relative frequency analysis in DiscoVista v1.0 (Sayyari, Whitfield, and Mirarab 2018). I re-rooted gene trees using *T. castaneum* as an outgroup with Newick Utilities (argument *nw_reroot*) (Junier and Zdobnov 2010) and compared them to the species tree. I evaluated the percentage of gene-tree concordance at each tree node with *phyparts* (Smith et al., 2015).

3.4 Results

3.4.1 Genome complexity and estimated genome size of *P. strobi*

The Smudgeplot results support the diploid genome (Figure 3.1a), with the most abundant heterozygous k-mer pair being AB (64%). The genome is highly heterozygous according to GenomeScope, with the first peak in the k-mer frequency distribution being higher than the second peak (Figure 3.1b). GenomeScope estimates an average of 2.56 heterozygous bases every 100 bp. Using diploid k-mers modeling, the estimated average genome size was 1.762 Gb over the k-mers range of 21–29 bp (Appendix Figure B.1). The flow cytometry estimate of haploid genome size is 2.07 ± 0.05 pg (mean ± SD), or 2.02 ± 0.05 Gb.

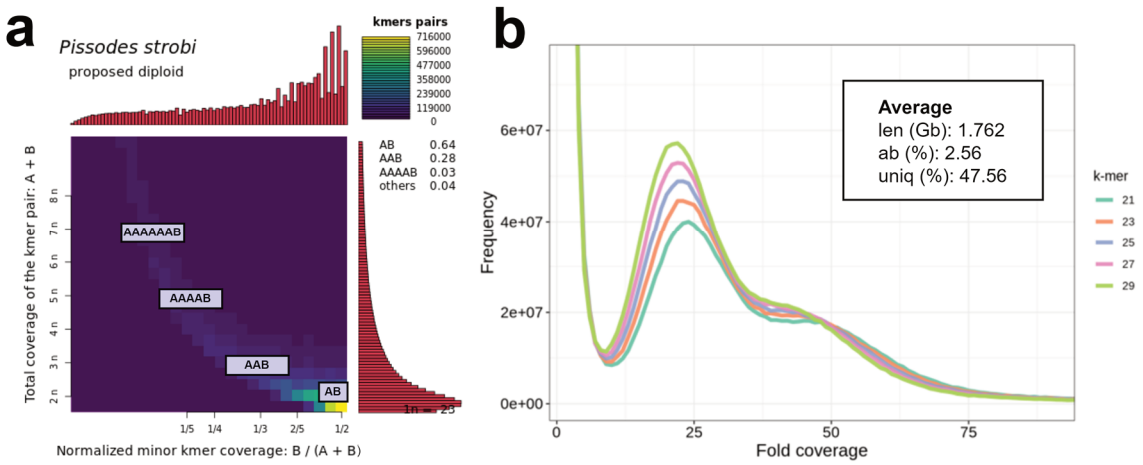

Figure 3.1 Ploidy level, k-mer profiles, and genomic features of *P. strobi* genome. (a) Smudgeplot profile for heterozygous k-mers in *P. strobi.* (b) k-mer histograms and genomic features of *P. strobi*. The k-mer histograms show the frequency of k-mers on the y-axis and the fold coverage on the x-axis for five chosen k-mer sizes. The curve shows the sequencing error k-mers in the fold coverage range between 0 and 3, a maximum peak representing the heterozygous k-mers around 25, and a minor last peak (around 50) representing the homozygous k-mers (common to both parental alleles); highly repetitive k-mers are found in the region after 50-fold coverage. The average genome size (Gb), percent heterozygosity, and percent unique sequences are averaged among the k-mer sizes 21–29 bp, highlighted in the plot ("len" = genome size; "ab" = percent heterozygosity, "uniq" = percentage of unique k-mers).

3.4.2 Genome assembly

The genome of *P. strobi* is sequenced at 53-fold coverage from short reads. The assembled draft genome has a reconstructed genome size of 1.83 Gb (Table 3.1), close to the estimated genome size of 1.762–2.00 Gb, and an NG50 of 87.7 kb. The high heterozygosity of the *P. strobi* genome (Figure 3.1b) impeded complete haploid genome reconstruction and caused some over-assembly. To compensate, 80,527 heterozygosity-induced duplicated scaffolds are removed at the Purge Haplotigs stage (Table 3.1), equivalent to 0.4 Gb. After removing heterozygosity-induced scaffolds, a decrease in the number of

BUSCO "duplicated" genes was observed from 12.2% to 8.2%, which likely indicates the removal of duplicated sequences. The number of BUSCO "complete – single copy" decreased after splitting the misassembled scaffolds with Tigmint and increased after ARKS scaffolding, with 70.7% and 71.0% BUSCO "complete – single-copy" genes. At the final assembly stage (highlighted in Table 3.1), Sealer closed 9,207 scaffold gaps, corresponding to 13.41% of the total gaps. The final genome assembly has genic space completeness of 71.0%.

Table 3.1. *Pissodes strobi* genome assembly statistics for each assembly step. Starting with the first assembly (Supernova) on top and ending with the final assembly step (Sealer) on the bottom. The statistics for the final genome assembly is shown in bold in the bottom line. Values are calculated for scaffolds longer than 1 kb. The estimated genome size for computing NG50 is 1.83 Gb. The Endopterygota BUSCO core gene set (n=2,124) was used to evaluate the gene completeness: BUSCO is reported as "complete – single copy" and "complete – duplicated."

Assembly stage	No. of scaffolds	Longest scaffold (kb)	NG50 (kb)	Reconstructed size (Gb)	BUSCO single copy (%)	BUSCO duplicated (%)
Supernova	163,521	2,374.58	79.50	2.23	66.7	12.2
Purge Haplotigs	82,994	2,374.58	79.34	1.83	71.0	8.2
Tigmint	84,653	2,139.77	74.90	1.83	70.7	8.6
ARKS	82,897	2,209.50	87.59	1.83	71.0	8.6
Sealer	**82,896**	**2,210.97**	**87.74**	**1.83**	**71.0**	**8.6**

3.4.3 Protein-coding gene annotation

I annotated 19,484 genes and 19,532 mRNAs (Table 3.2). The annotation pipeline for this dataset predicts 58.6% (1,244) of the total BUSCO genes as "complete" ("single copy" and "duplicated"). The quality selection criteria used to identify a "high confidence" set of genes reduced the number of genes to 11,382, annotating 42.9% of the total BUSCO "complete." Altogether, the total gene bases of the annotated genes

covered 1.1% and 0.7% of the total *P. strobi* genome for the "total annotated" and "high confidence" gene sets, respectively.

Table 3.2. Gene annotation statistics for "total annotated" and "high confidence" datasets in *P. strobi*. The total annotated genes are shown as the total gene bases (coding and non-coding gene regions) and their corresponding percent from the genome. The BUSCO completeness, shown as the sum of BUSCO "complete – single copy" and "complete – duplicated," is used to estimate the annotation completeness.

	Total genes	Total mRNA	Total gene bases (Mb)	BUSCO complete (%)
Total annotated	19,484	19,532	22.69 (1.1%)	58.5
High confidence	11,382	11,405	14.51 (0.7%)	42.9

3.4.4 Inference of Curculionidae phylogeny from genomic data

My phylogenomic analysis based on BUSCO single-copy genes (Simão et al. 2015) included nine species of Curculionidae and *Tribolium castaneum* as an outgroup in Coleopterans (Appendix Table B.2), comparing a total of 2,080 BUSCO orthologs and 16,418 total genes included in the phylogeny inference. A large percentage of orthologous genes (63%) was obtained from at least eight out of ten species. The poorly overlapping orthologs (reconstructed in less than six species) represented only 3% of the total orthogroups. Only a small fraction of the genes used for phylogeny were fragmentary, containing more than 67% of sequence gaps following multiple sequence alignment; these were removed from the final analysis (i.e., 43 of 16,461 total genes). The inferred species tree has local posterior probabilities (LPP) of 100% for all branches (Figure 3.2), with most branches supported by at least 50% of gene trees. Several recognized taxa are inferred to be monophyletic at the current taxon sampling, including the bark beetle subfamily Scolytinae (represented here by *Dendroctonus ponderosae*, *Ips typographus,* and *Hypothenemus hampei*) and the pantropical weevil subfamily Dryophthorinae (represented here by

Sitophilus oryzae and *Rhynchophorus ferrugineus*). *Pissodes strobi* (subfamily Molytinae) is inferred to be the sister group of *Elaeidobius kamerunicus* (the flower weevil subfamily, Curculioninae). However, considering only genes with signal, substantial gene trees conflict with this relationship than support it (i.e., discordant trees: 446, concordant trees: 187; Figure 3.2; Appendix Table B.3). Another branch with a high degree of conflict between underlying gene trees, considering the relative quartet frequencies calculated by ASTRAL-III and DiscoVista (Appendix Figure B.2), is the one representing a possible sister-group relationship between *I. typographus* and *H. hampei*).

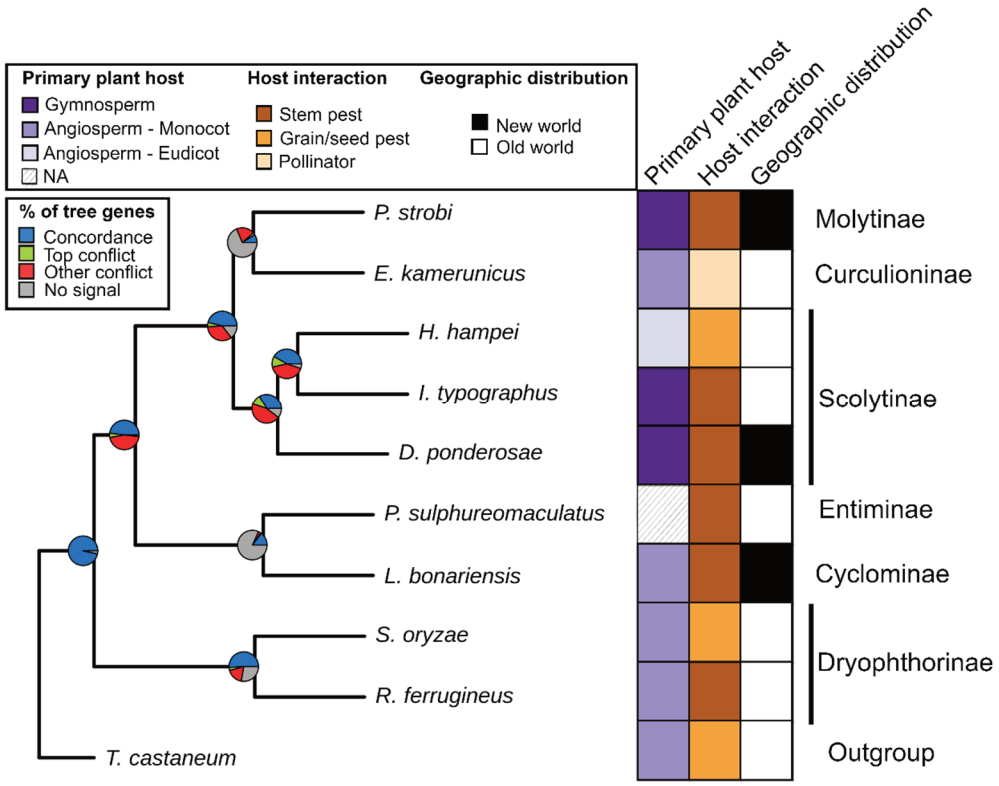

Figure 3.2. Curculionidae species tree inferred from BUSCO single-copy genes. The pie charts summarize the gene support on specific branches under the multispecies coalescent model. The supporting genes at each internal node are shown in blue ("concordance"); green and red represent gene trees that are conflicting with the species tree, respectively supporting the main alternative ("top conflict") or other alternatives ("other conflicts"); gray represents gene trees that are missing for that relationship. The block graph classifies Curculionidae based on plant host, host interaction, and geographic origin; the Curculionidae subfamily is noted for each species.

3.4.5 Genomic repeats annotation

3.4.5.1 Repeat annotation and quantification

The genome assembly of *P. strobi* is annotated with a high fraction of repeats, equaling 53% of the genome (Appendix Table B.4). K-mer profiling based on unassembled reads annotates 52% of the genome as repetitive (Figure 3.1b). The most abundant class of annotated repeats are class II DNA transposon elements (TE), representing ~30% of the genome (Appendix Table B.4). The class I TE or LTRs represent ~16% of the total genome. Repeat annotation based on unassembled reads annotated ~48% of 7,953,905 randomly sampled reads as genomic repeats (Appendix Figure B.3).

The RepeatExplorer cumulative analysis for *P. strobi*, seven Curculionidae, and *Tribolium castaneum* (outgroup) compares the repeat diversity among genomes. RepeatExplorer selected 4,486,313 total reads from all the species, proportionally to each genome size. The software grouped the reads in 529,599 clusters based on their sequence similarity (Appendix Figure B.4). I use the top 443 clusters, covering ≥0.1% of the analyzed reads, to compare the genomic repeats from each species. The clusters show distinctive patterns of repeats (Figure 3.3), particularly for the four most repeat-rich genomes: *Pissodes strobi*, *Pachyrhynchus sulphureomaculatus*, *Listronotus bonariensis,* and *Sitophilus oryzae*. *Pissodes strobi* has the highest number of repeat clusters (x-axis in Figure 3.3) with 204 annotated clusters (Appendix Table B.5), followed by *Pachyrhynchus sulphureomaculatus* with 167 clusters of repeats. The 106 clusters in *P. strobi* are classified as LTR or class I type of repeats, and 39 are classified with DNA or class II type of repeats.

The two annotation methods, based on the genome assembly and the unassembled reads, yielded different types of repeat classification. The most common type of annotated repeats in the assembly is the DNA class II TE type; on the contrary, the most common class of repeats detected from unassembled short reads was the LTR class I TE type.

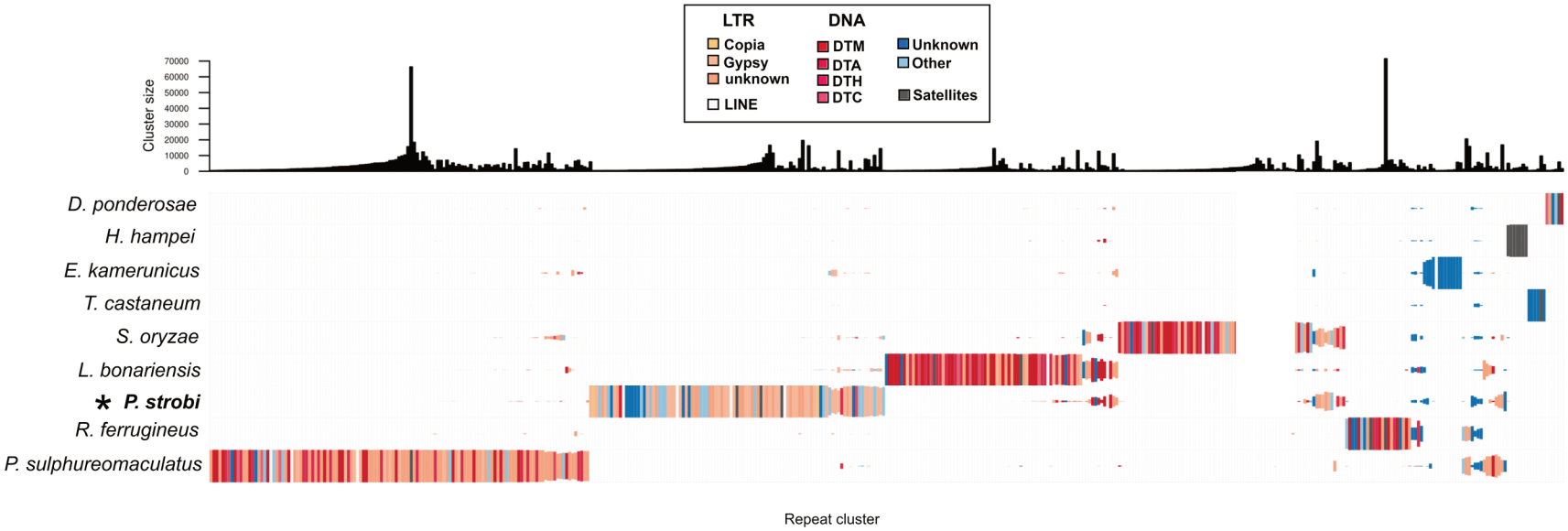

Figure 3.3 Comparative repeat analysis of unassembled genomic reads of eight Curculionidae accessions and *Tribolium castaneum*. The top read clusters are shown on the x-axis. Top figure – barplot: cluster size expressed as number reads (y-axis). The bottom section represents the repeat annotation of each cluster in each accession (*P. strobi* is marked with an asterisk). The height of each rectangle represents the proportion of reads classified in the repeat class (most abundant repeat). The annotated classes of repeats are 1) LTR or long terminal repeats class I type (pink and salmon), sub-divided into Copia, Gypsy, and unknown subclasses, 2) DNA class II type (red), sub-divided in DTM, DTA, DTH and DTC subclasses, 3) LINE or long interspersed nuclear elements (white), 4) Unknown repeats (dark blue), 4) "Other" repeats (light blue) include Mavericks, Penelope, DIRs, TIRs, and Helitrons, and 5) satellites (gray) (Craig 2020).

3.4.5.2 Repeats turnover during weevil (Curculionidae) evolution

The repeat landscape based on Kimura 2-parameter (K2P) distances indicates repeat amplification and extinction, represented by peaks and valleys in the repeat landscape. The Kimura repeats landscape highlights that there has been a recent repeat expansion in *P. strobi*. *Pissodes strobi* displays a large spike of proliferation for the DNA, LTR, and unknown repeats in the Kimura distance between 3 and 6 units (Figure 3.4). *Pachyrhynchus sulphureomaculatus* has a comparable genome size and a similar peak of proliferation that may indicate an analogous repeat evolution event. The genome of *L. bonariensis*, the closest species to *P. sulphureomaculatus* (Figure 3.2), is also characterized by large genome size and displays a peak (repeat expansion) in the recent K2P distances (5 units) and a more ancient peak around 20 units.

Most species characterized by large genome sizes, excluding *I. typographus*, display a recent burst of repeats at around five units of K2P divergence. *Rhynchophorus ferrugineus* and *S. oryzae* have a comparable genome size of 0.7 Gb, but their repeat expansion follows a different K2P profile, with peaks and valleys at different units. While *S. oryzae* has a recent repeat amplification, as in the genomes of *Pissodes strobi* and *Pachyrhynchus sulphureomaculatus*, *R. ferrugineus* has a more ancient repeat amplification at around 30 units, followed by a recent repeat extinction at 20 and a recent amplification at 15 units.

The repeat annotation of *Elaeidobius kamerunicus* (Appendix Table B.4), the closest Curculionidae genome to *P. strobi* (Figure 3.2) with a genome size of 0.2 Gb, indicates a large abundance of mostly “Unknown” type of repeats, uncommon to the other Curculionidae genomes. The different from the other species’ repeat profile required further analysis of the genome: the k-mer histogram built from *E. kamerunicus* reads (Appendix Figure B.5) shows the absence of between heterozygous and homozygous peaks, as shown for the k-mer histogram of *P. strobi* (Appendix Figure B.1).

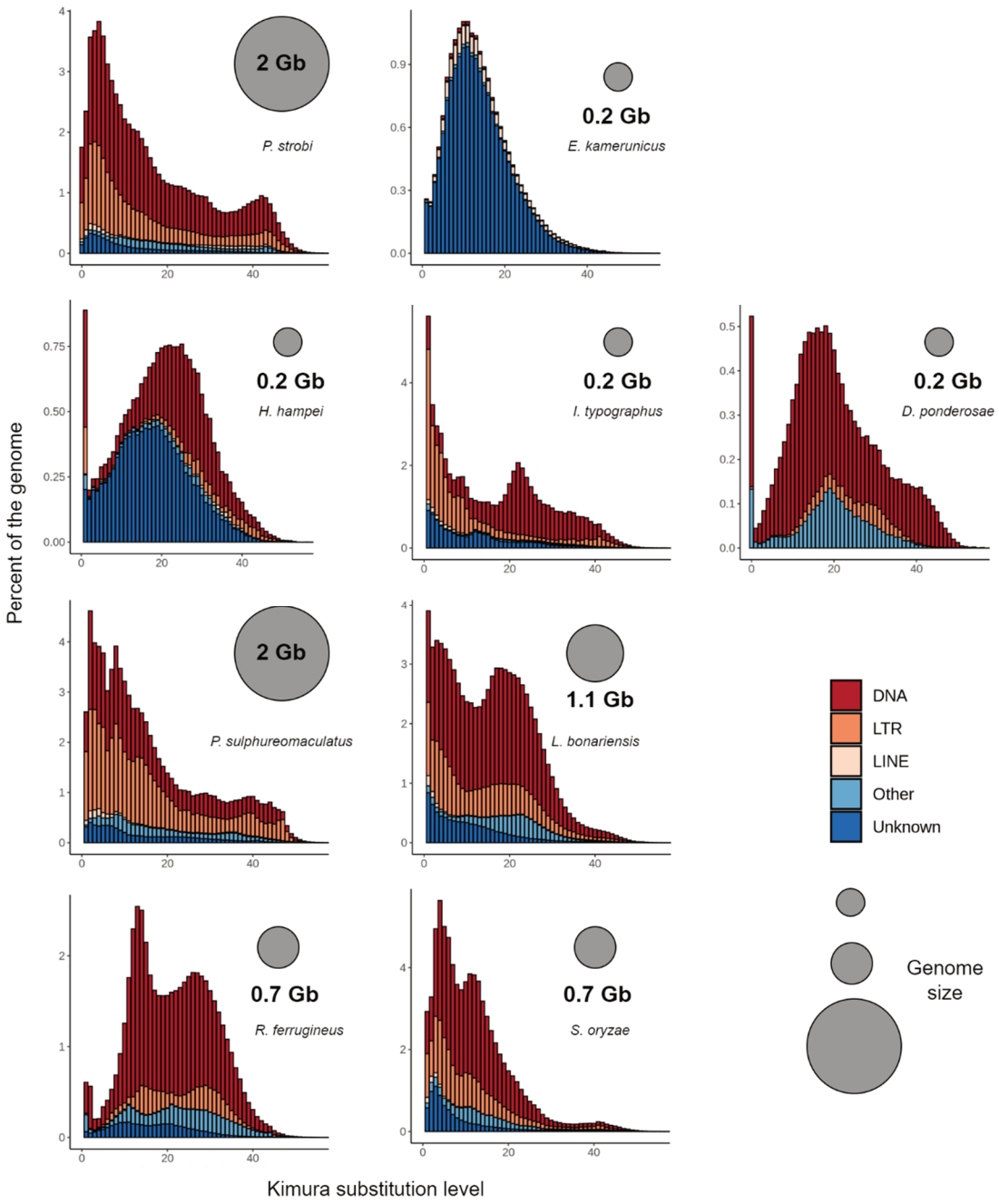

Figure 3.4 Sequence divergence distribution for transposable elements and corresponding genome sizes for Curculionidae. The x-axis shows the Kimura-2 parameter (K2P) sequence divergence between TE copies and the reference repeat library. High values of substitution indicate a more divergent sequence distance than low values. The y-axis shows the percent of the genome that is annotated with TE. Repeats are grouped into five major classes: DNA, LTR, LINE, Unknown, and Other. The last class includes Mavericks, Penelope, DIRs, TIRs, and Helitrons (Craig 2020).

3.5 Discussion

3.5.1 Genome complexity of *P. strobi*

We sequenced and assembled the genome of *Pissodes strobi*, which is a devastating forest pest in North America (Ebata 1991; Alfaro 1994). The genome of *P. strobi* is complex, repeat-rich, and larger than most other Curculionidae genomes (Appendix Table B.2; Appendix Table B.4). Experimental analyses support a nuclear genome size of ~2.0 Gb, comparable in size in the family only to *Pachyrhynchus sulphureomaculatus*. The latter also has an assembled genome size of ~2.0 Gb.

Our genome annotation highlights the large abundance of TEs, estimated to be around 48–53% of the genome, depending on the method used. The repeat annotation of the assembled genome of *P. strobi* found the class II type of TEs to be the dominant type (Appendix Table B.4). In contrast, an analysis of the unassembled short reads annotated a larger fraction of LTR repeat clusters belonging to the class I type of TE (Figure 3.3).

Timing analysis of repeat turnover (Figure 3.4) suggests a recent repeat reshaping in *P. strobi*. The Kimura 2-parameter (K2P) distance landscape highlights a recent repeat expansion and lack of TE elimination that may have played a role in the genome size expansion of *P. strobi*. Class II type of TE (DNA) and class I of TE (LTR) show a contemporary divergence peak in *P. strobi*. These two types of repeats show a synchronous expansion, which indicates a related repeat expansion event. The "Unknown" type of TEs, covering ~3% of the total genome (Appendix Table B.4), have a steady expansion, culminating with a synchronous peak with the DNA and LTR types of repeats.

3.5.2 Annotation of *P. strobi* repeats from genome assembly and unassembled reads

Repeats from non-model organisms are under-represented in repeat databases such as RepBase (Bao, Kojima, and Kohany 2015). Given this knowledge gap, it is important to characterize transposable elements (TEs) in an unbiased manner across newly sequenced organisms. Here, I applied one of the most recent repeat annotation pipelines, EDTA (Ou et al. 2019), which uses a combination of tools for TE annotation. The software annotates subclasses of LTRs, TIRs, and Helitrons and generates a

non-redundant library that contains species-specific repeats. I independently run the repeat annotation pipeline on the genomes of *P. strobi* and eight Curculionidae genomes to allow a direct and unbiased comparison.

Along with the annotation of repeats on the assembled Curculionidae genomes, I also performed a repeat annotation and comparative analysis on the unassembled short reads of *P. strobi*, seven Curculionidae, and *T. castaneum*, using the RepeatExplorer pipeline (Novák et al. 2013). The pipeline subsamples genomic reads to determine clusters of repeats without the use of known repeat elements. RepeatExplorer assigns repeat clusters to know classes of repeats in a later annotation stage. I included the repeat annotation of unassembled reads because it is a broadly used approach for repeat quantification in non-model organisms (Heitkam et al. 2021; Novák et al. 2020) and is independent of the quality of the reference genome.

The two pipelines, EDTA and RepeatExplorer, annotated a similar percent of total repeats; however, the classes of repeats differed between the two methods. The lower fraction of annotated LTRs in the genome assembly, the major difference between the two pipelines, is possibly due to the lack of complete LTR elements in the assembly. This type of repeats is difficult to assemble because its length ranges between 4 kb and 20 kb, significantly longer than the short reads used in the assembly. Frequent indels in the repeat structure make their assembly challenging (Ou, Chen, and Jiang 2018; Ou and Jiang 2018). The RepeatExplorer method based on unassembled reads suggests that LTR elements may be abundant in the genome of *P. strobi*.

3.5.3 Phylogeny and host preference of Curculionidae

The phylogeny of the family Curculionidae has been previously reconstructed using >500 single-copy orthologs in Shin et al. (2018). I employed 2,080 single-copy orthologs from the nuclear genome, resulting in 16,418 genes: to my knowledge, this is the largest number of gene markers used to infer the phylogeny of Curculionidae. I compared *P. strobi* to eight Curculionidae species, representing five different subfamilies. *Pissodes strobi* is one of only three species considered here that infest gymnosperm

trees (Figure 3.2) but is more closely related to *Elaeidobius kamerunicus,* a "pollinator" weevil that feeds on angiosperms, rather than to other bark beetles sampled here (*Dendroctonus ponderosae* and *Ips typographus*, some of the most destructive forest pests known). These points suggest a substantial capacity for host plant switching in the family (Cognato, Smith, and Jordal 2021). Similar host disparity is evident elsewhere in the phylogeny (Figure 3.2), as host preference also differs within Scolytinae: the bark beetle *Hypothenemus hampei* feeds on angiosperm, in contrast to its closest species, *D. ponderosae,* and the *I. typographus*, which are conifer pests.

3.5.4 *Wolbachia* putative endosymbiont of *P. strobi*

We screened the sequenced genomic library of *P. strobi* for contaminants and for accompanying exogenous DNA. This flagged the presence of reads with sequence similarity to *Wolbachia* spp. *Wolbachia* spp. are widespread endocellular α-proteobacteria recognized as reproductive parasites. Given their close contact with host reproductive tissues, the presence of *Wolbachia* plays an important role in host development and reproduction (Werren, Baldo, and Clark 2008). The presence of bacterial endosymbionts such as *Wolbachia* has been extensively described in *Sitophilus oryzae,* where specialized host cells called bacteriocytes have been found to contain *Wolbachia* (Heddi et al. 1999). Detailed studies of bacteriocyte tissues of *P. strobi* have not yet been performed. However, the structures that house bacteriocytes (i.e., bacteriomes) have been previously documented and described in *P. strobi* (Whitehill et al. 2016). Further study of *Wolbachia* infection in *P. strobi* is warranted.

3.5.5 The case of the *Elaeidobius kamerunicus* genome

The genome of the *E. kamerunicus* (Apriyanto and Tambunan 2021) used in the comparative genome analysis showed a low number of genomic repeats (Appendix Table B4). Annotated repeats covered 11.32% of the total genome and were classified for the most part as "Unknown" (9.83%). The k-mer coverage histogram showing k-mers from all the reads used in the genome assembly (Appendix Figure B.5) likely indicates a low k-mer coverage, correlating with low coverage of sequenced reads, which does

not completely represent the sequenced genome. Based on all those observations, it is possible that the genome of *E. kamerunicus* is larger than the reconstructed genome size of 0.26 Gb. The genome assembly is likely under-assembled based on its reconstructed genome size, meaning that only a fraction of the genome is represented by the assembly.

3.5.6 Future directions

The genome and annotations generated in this chapter provide a reference for future population-based studies of *P. strobi* natural populations. The insect pest populations successfully evolved to exploit new host species and inhabit forests with different climates, such as those characterized by different annual precipitation and different temperatures. Studies based on random polymorphic DNA and mitochondrial genome markers (Laffin, Langor, and Sperling 2004; Lewis et al. 2000; Lewis et al. 2001) highlighted the existence of three distinct *P. strobi* populations in Canada: one from the South Coast of British Columbia and Vancouver Island, one from the North Coast of British Columbia, and one including populations for interior British Columbia and east of the continental divide. Having such a diverse geographic distribution, studies of *P. strobi* individuals in different natural populations could reveal clues related to its host and environment adaptation.

A recent study of *Dendroctonus ponderosae* natural populations (Keeling et al. 2013), a Curculionidae forest pest compared to *P. strobi* in this chapter, highlights a large intra-species variability between two distinct populations (Keeling et al. 2021). The genome of *D. ponderosae* displays high variability in the chromosome region associated with the X sex-chromosome (neo-X) (Keeling et al., 2021). The chromosome originates from the fusion of an autosomal chromosome and an ancient X sex-chromosome (Bracewell et al. 2017). This observation suggests that the formation of the neo-X chromosome limits the gene flow or the transfer of genetic material between populations and that large genomic rearrangements can play an important role in establishing insect populations.
Future studies of *P. strobi* and its sex determination can help understand if that can play a role in the local adaptation and evolution of insect pest populations, as demonstrated for *D. ponderosae*.

Chapter 4: Genome assembly and annotation of Willow-alpha, a *Cannabis sativa* variety, with a focus on anthocyanin biosynthesis

4.1 Author summary

More than 400 different chemical compounds are described in *Cannabis sativa*, yet most of the gene pathways involved in their biosynthesis are not well characterized. In this chapter, I assemble the genome of Willow-alpha *C. sativa* variety with a combination of short and long reads and test two standard gene annotation pipelines to find the best performing method. The gene annotation of Willow-alpha, in combination with transcriptome quantification and metabolites profiling, is used to characterize the gene pathway regulating the flavonoid and anthocyanin production in *C. sativa*. I present a comparative study of Willow-alpha and other *C. sativa* varieties to characterize the anthocyanin accumulation phenotypes. This chapter addresses objectives 1 and 3 in section 1.4.

4.2 Introduction

4.2.1 Evolution, domestication, and chemotype classification of *Cannabis sativa*

Cannabis sativa is one of the oldest cultivated plants and has been used for food, fiber, and as an intoxicant (Small 2015). The medical benefits of *C. sativa* for the treatment of specific ailments and symptoms of diseases are becoming clear and medical use of *C. sativa* has been approved in several countries, including Canada, since 2001 (Spurgeon 2001). Moreover, *C. sativa* use for recreational purposes became legal in Canada in 2018 (Fischer, Russell, and Boyd 2020; Cox 2018).

The Cannabaceae family includes *C. sativa*, *Humulus* spp. (hops) and eight other genera (Yang et al. 2013); *C. sativa* has diverged from hops around 21 MYA based on plastid gene markers (Zerega et al. 2005). Given the close association between *C. sativa* and humans who domesticated the plant around ~12,000 years before the present (BP) (Ren et al. 2021), *C. sativa* has been strongly modified by human migrations, breeding, and selections.

During the recent history of *C. sativa* prohibition (Collins 2020), clandestine plant breeding and the resulting frequent gene flow between plant varieties have made the taxonomic classification of *C. sativa* challenging, with several models for its classification proposed (Small and Cronquist 1976; Hillig 2005). Combining evidence from single nucleotide polymorphisms (SNPs) and chemotype classification in *C. sativa* varieties identified two major groups (Sawler et al. 2015). The first group is a narcotic-medicinal-type cultivated for its very high tetrahydrocannabinolic acid (THCA) content. The second group is the hemp-type, which is phenotypically and chemically distinct from narcotic-medical-type *C. sativa* and has been grown historically for fiber and food. Hemp-type *C. sativa* has traditionally been associated with the accumulation of modest amounts of non-intoxicating cannabidiolic acid (CBDA), with little to no THCA accumulation. However, recent interbreeding between narcotic-medicinal and hemp types has resulted in varieties of hemp-like *C. sativa* that can accumulate much higher amounts of THCA and CBDA than has historically been found in them. Landrace *C. sativa* varieties from Central and East Asia, described as the origin of *C. sativa*, are still found in nature and represent naturally occurring feral plants. Such varieties are studied to understand better the recent history of *C. sativa* and its domestication (Soorni et al. 2017; de Meijer and van Soest 1992; McPartland, Hegman, and Long 2019; Ren et al. 2021).

4.2.2 *Cannabis sativa* secondary metabolites: flavonoids

Cannabis sativa produces a plethora of specialized metabolites, with more than 400 different compounds described so far (Elsohly and Slade 2005). The most well-known ones are the cannabinoids THCA and CBDA, which are well studied due to their psychoactive and medical properties. The intoxicating aspects of THCA are associated with the medical and recreational use of *C. sativa*. In contrast, CBDA is a non-intoxicating cannabinoid that has recently been approved as a therapeutic compound for the treatment of epilepsy (Epidiolex 2018: Abu-Sawwa, Scutt, and Park 2020).

Additional classes of specialized metabolites have also been identified in *C. sativa*. Terpenoids such as myrcene and b-caryophyllene are among the most notable of these other metabolites, and they

contribute to the distinct aromas found in different *C. sativa* varieties. The presence of terpenoids is hypothesized to modulate the intoxicating aspects of THCA in a phenomenon termed the "entourage effect" (Russo 2011). Phenylpropanoids comprise a large family of specialized metabolites that are derived from the amino acid phenylalanine via the function of phenylalanine lyase (PAL). The flavonoid-specific branch of phenylpropanoid biosynthesis commences via chalcone synthase (CHS), which shuttles carbon toward various flavonoids such as the highly pigmented anthocyanins. Phenylpropanoids are well studied in other plants, but their structural diversity and biosynthesis have thus far not been well characterized in *C. sativa*. Flavonoids are constructed from two benzene rings (A and B) connected by a 3-carbon linking chain (C), as shown in Figure 4.1. The chemical structure defining the major classes of flavonoids depends on the functional groups on the B benzene ring's 3', 4', and 5' carbons and the degree of oxidation and unsaturation of C (Figure 4.1). Flavonoids can exist as primary structures (compounds in Figure 4.1) or can be further modified into a highly diverse repertoire of metabolites via methylation, acetylation, prenylation, or through conjugation to glycosides.

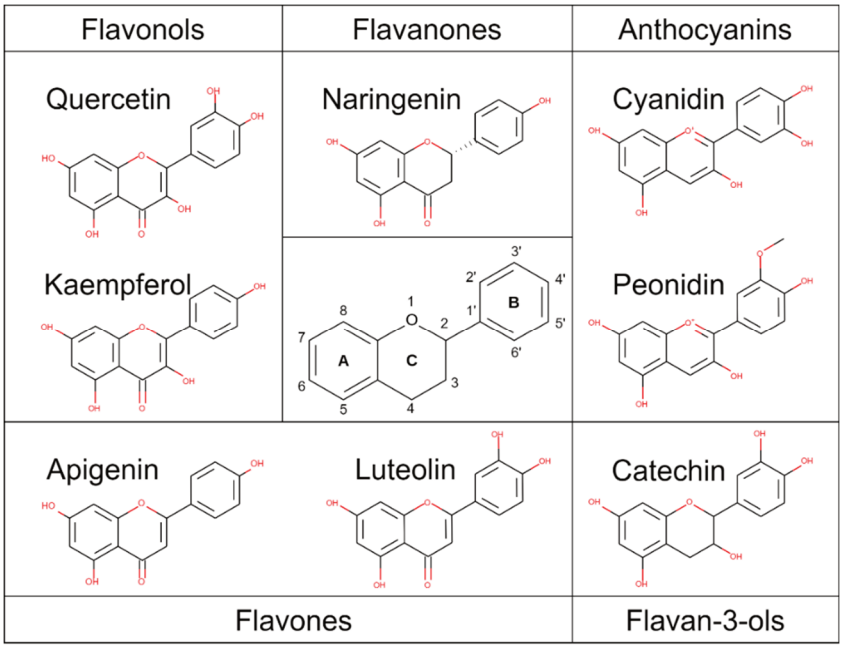

Figure 4.1. Classification and examples of flavonoid metabolites and their chemical structure. The structure of the flavonoids (middle panel) is a 15-carbon skeleton, with two benzene rings (A and B) connected by a carbon

linking chain (C). Flavonoids are classified into six groups: flavonol, flavanone, isoflavone (not shown), flavone, flavan-3-ols, and anthocyanin. The chemical structure for each class was downloaded from ChEMBL (Gaulton et al. 2012) and combined with online diagram software – draw.io (Jgraph 2021).

Flavonoids have several important functions in plants (Mierziak, Kostyn, and Kulma 2014), including response to ultraviolet (UV) radiation and response to infection, pigmentation, and various aspects of plant development and reproduction. Studies in several plant species, such as *Vitis vinifera* and *Arabidopsis thaliana* (Berli et al. 2010; Schenke, Böttcher, and Scheel 2011), have shown UV light induces the production of flavonoids. Flavonoid biosynthesis plays a vital role in countering the accumulation of reactive oxygen species (ROS) generated by UV radiation (Verdan et al. 2011). Indeed, the potent radical scavenging and antioxidant properties of flavonoids make them important components of human nutrition (Kumar and Pandey 2013). Flavonoids provide taste and color to the fruit and flowers of many plants, such as blueberries and red roses, respectively.

The observable phenotypes of individual *C. sativa* plants are highly variable (Aardema and DeSalle 2021). One of the most noticeable phenotypes is the color of *C. sativa* leaves and flowers. Some *C. sativa* varieties are light green and almost yellow, whereas others are brightly purple or even black in appearance. The red pigments associated with the purple and black phenotypes are likely anthocyanins, a type of flavonoid that accumulates in plant cell vacuoles. Chemical characterization of the flavonoids in the context of their biosynthesis pathway and genomics is lacking in *C. sativa*. Considering the wide range of impacts that flavonoids have on plant physiology and their effect on the visual appeal of commercial products, further study of the genes behind their biosynthesis pathways is warranted. Finally, identifying the genes involved in the biosynthesis pathways in *C. sativa* will aid marker-assisted breeding strategies for plant varieties with pigmentation colors, whose phenotype is potentially of added value to *C. sativa* producers.

4.2.3 *Cannabis sativa* genome complexity and available genome assemblies

Cannabis sativa is generally dioecious, although some varieties are monecious. It has nine pairs of autosomal chromosomes and a pair of sex chromosomes, heterogametic in staminate plants (males) (XY) and homogametic in ovulate plants (females) (XX). Its genome size varies between 810 Mb and 840 Mb for female vs. male plants, respectively (Sakamoto et al. 1998; Muyle, Shearn, and Marais 2017). Its genome has undergone several whole genome duplication (WGD) events: a study of chromosome block synteny with *Trema orientale*, part of the Cannabaceae family, and *Ziziphus jujuba* and *Morus notabilis* as an outgroup, uncovered three recent WGD events in *C. sativa* in the last 35 MYA (Gao et al. 2020). The same study shows that two additional ancient WDG occurred before the divergence with *T. orientale*. The diploid structure of the *C. sativa* genome suggests that the species has undergone repeated cycles of genome diploidization after each cycle of recent WGD (Gao et al. 2020). It has relatively high genome complexity due to abundant repeats, representing ~65% of its genome (Pisupati, Vergara, and Kane 2018). The most common transposable element (TE) repeat type is the LTR-Gypsy and LTR-Copia, covering ~40% of the genome; there was a recent burst of these two classes around 1 MYA (Gao et al. 2020).

An initial draft genome of *C. sativa* was published in 2011 based on the Purple Kush (PK) variety (van Bakel et al. 2011). The first genome assembly used a set of short reads, and the genome assembly represented an incomplete reconstruction due to the large portion of repeats. Since then, several other genome assemblies of *C. sativa* have been publicly available using long reads (Pollard et al. 2018) and chromosome conformation mapping (Burton et al. 2013) for scaffolding. These include an improved genome assembly of PK, and a newly assembled genome for Finola hemp (Laverty et al. 2019), CBDRx cs10 (Grassa et al. 2021), the JL landrace variety (Gao et al. 2020), and Cannbio-2 (Braich et al. 2020), all of which have high-quality assemblies with chromosome-level genome contiguity. PK is a narcotic-medical-type variety that produces THCA predominantly; CBDRx-cs10 and Cannbio-2 are also narcotic-medical-type varieties but have a balanced production of both THCA and CBDA. Finola is a

hemp-type, and the JL is a naturally occurring feral plant found growing in Tibet, which has an uncharacterized chemotype.

Here I produce the genome assembly and annotation for the proprietary Willow-alpha *C. sativa* variety. Moreover, I collected quantitative transcriptome and metabolome data derived from leaf tissue in Willow-alpha and three other *C. Sativa* varieties. The transcriptomic and metabolite data, combined with a well-annotated genome of Willow-alpha, provides a powerful system for further characterizing secondary metabolite biosynthesis in *C. sativa*. I identify the flavonoid/anthocyanins biosynthesis genes in the *C. sativa* genome and correlate their gene expression with flavonoid and anthocyanin accumulation phenotypes in Willow-alpha and three other *C. sativa* varieties.

4.3 Methods

4.3.1 Genome assembly

Willow-alpha is a narcotic-medicinal variety of *C. sativa* that produces ~3.8% THCA (fresh weight) or ~19% THCA (dry weight) in mature flower tissue. We extracted the genomic DNA from a single female individual. The overall sequencing approach was to sequence the genome using a combination of short and long reads and genome-wide chromatin conformation Hi-C Phase Genomics capture for scaffolding. The assembly pipeline is shown in Appendix Figure C.1, with the corresponding datasets and tool references. The genomic DNA was submitted for sequencing in the summer of 2018. The long reads were generated from 16 Pacific Biosciences (PacBio) RS II lanes, providing 100-fold genome coverage. The reads were removed from their adapter sequences and selected for subreads. Subread extraction in PacBio for genome *de novo* assembly commonly involves selecting a single pass of polymerase reaction on a single strand, which I did here. Short paired-end (PE) reads (2x150 bp) were produced using Illumina TruSeq PCR-free protocol, with a total genome coverage of 160-fold. Additionally, Hi-C short chromosome conformation reads were generated for scaffolding.

A primary genome was assembled using the PacBio long reads and the hierarchical genome assembly Process (HGAP) available in Falcon (Chin et al. 2016). Briefly, reads longer than 1 kb were selected as "seeds" for error correction, and all the shorter reads from the long-read dataset were aligned to these reads to generate a consensus sequence using the argument *-min-coverage 2* and *-max-coverage 120*. The resulting assembly was polished with the PacBio subreads to create a consensus. The Falcon-Unzip module was then applied, in which the raw reads were used to phase the genome according to diploid single nucleotide variants (SNVs). Using Purge_dups v1.0.0 (Guan et al. 2020) with default parameters, I removed heterozygous and duplicated scaffolds from the Falcon primary assembly. The tool identifies heterozygous scaffolds based on the excess of heterozygous k-mers from the first peak in the k-mer histogram.

I scaffolded the resulting assembly with Salsa v2.2 (Ghurye et al. 2017), run with default settings and chromosome conformation reads produced using Hi-C PhaseGenomics short reads (Belaghzal, Dekker, and Gibcus 2017). To reduce the error rate from the PacBio long reads, the genome was polished with Illumina PE short reads in Racon v1.4.13 (Vaser et al. 2017). A genome assembly of PacBio long reads was generated with wtdbg2 (Ruan and Li 2020) to scaffold the polished genome. The primary assembly (target) was scaffolded with the wtdbg2 assembly (reference) using in ntJoin v1.0.6 (Coombe et al. 2020) and the argument *-w 500 -k 28 -n 2*. Gap filling was done with Sealer v2.2.3 (Paulino et al. 2015), with k-mer sizes of 90, 100, 110, and 120 bp. The scaffolds from the gap-filling stage were grouped and oriented based on the cs10 *C. sativa* reference genome (Grassa et al. 2021). Scaffolding was done with ntJoin software using the argument *-w 300 -k34 -n2 -G100000*, considering cs10 as a reference and Willow-alpha as the target.

To inspect the quality of the genome assembly, I plot its synteny to cs10 in Circos Assembly Consistency plot or Jupiter plot (Chu 2017) executed with the parameters *ng=250*, *maxGap=1000000*, and *minBundleSize=800000*. I evaluated the genome completeness with BUSCO v4.1.4 and the odb10 Embryophyta (land plants clade) core gene set (Kriventseva et al. 2019).

4.3.2 Genome annotation

I inferred high-quality genome annotations using a set of 14 RNAseq samples as supporting evidence, which were sequenced from four Willow-alpha tissues (leaf, preflower, flower, and isolated trichomes) collected from clonal plants. Willow Biosciences Inc. generated RNAseq data using Illumina NovaSeq 6000 PE protocol (2x100 bp) at Genome Quebec. 25 million reads were generated per sample and screened for possible contamination. Briefly, I created a series of Bloom filters (Chu et al. 2014) based on viruses, bacteria, fungi, aphids (superfamily Aphidoidea), mites (subclass Acari), and thrips (order Thysanoptera), and Univec build #10 representing cDNA library vectors, downloaded in September 2020. Reads without a match to the listed Bloom filters were used for annotation. I provided additional evidence for gene structures from a cloned cDNA library: the dataset was derived from *C. sativa* trichome RNA, which Willow Biosciences sequenced using a PacBio RS II platform and which resulted in 170,416 Willow-alpha unique cDNA sequences. I also considered protein-coding sequences from the cs10 reference (Grassa et al. 2021).

For the genome annotation, I compared two annotation methods, namely the widely adopted MAKER v3.01.03 pipeline (Holt and Yandell 2011) and the newer BRAKER v2.1.5 pipeline (Brůna et al. 2021). I ran two annotation pipelines because I wanted to evaluate the optimal approach for Willow-alpha; the two resulting annotations were assessed using DOGMA v3.4 (Kemena, Dohmen, and Bornberg-Bauer 2019), evaluating conserved Pfam domains in plants, and BUSCO v4.1.4 (Simão et al. 2015) with the argument *-m protein*, to evaluate single-copy genes from the odb10 Embryophyta core genes set.

Both annotation strategies include AUGUSTUS (Stanke et al. 2008) and GeneMark (Lomsadze et al. 2005) in their pipelines. BRAKER has the advantage that it allows a fully automated training of the gene prediction models. I performed the AUGUSTUS (Stanke et al. 2008) meta parameters retraining with BUSCO v4.1.4 executed with the *-long* option (Simão et al. 2015) using the Embryophyta core gene set.

I run MAKER in two iterations: the first iteration, using the options *prot2genome=1* and *est2genome=1*, included the RNAseq supporting evidence, with transcriptome assembly done using StringTie v2.1.4 (Pertea et al. 2015), together with cs10 proteins and trichome cDNAs. AUGUSTUS v3.4.0 (Stanke et al. 2008) and SNAP v2006-07-28 (Korf 2004) predictors were trained with the predictions from the first iteration of MAKER. The second iteration of MAKER was run with *prot2genome=0* and *est2genome=0* options, using the output from the first iteration as supporting evidence (eAED <1), together with the cs10 proteome and the trichome cDNAs.

The supporting evidence for BRAKER included transcriptomic sequences mapped with hisat2 (Kim, Langmead, and Salzberg 2014) for the RNAseq reads and minimap2 (Li 2016) for the cDNA data with the protein sequences from cs10. BRAKER trains the gene models by calling GeneMark-ETP (*-etpmode* option). The UTRs of the predicted gene models were annotated with an additional run of BRAKER with the GUSHR tool using the argument *-addUTR*. Some UTRs are annotated several kb apart from the predicted CDS regions because of long introns in the first and last exons due to a bug in the BRAKER v2.1.5 release. These extreme UTRs were filtered out if annotated 373 bp and 283 bp further than the start or end of the CDS, respectively. These distances represent the values of the third quartile distance from the start or stop codon, respectively, as estimated from the cs10 gene annotations. The final annotated genes were selected if matching all the following criteria: (a) having a translated protein length >20 aa, (b) polyexonic genes have a minimum intron length of 10 bp, (c) having a complete CDS (containing both start and stop codons), (d) the translated protein does not contain known misannotations in protein domains (no match with AntiFam (Eberhardt et al. 2012)), and (e) the translated protein does not have a Pfam match with LTR repeat elements, such as GAG, Env or Pol (viral LTR domains). I functionally annotated the final set of genes using the enTAP tool suite v0.10.7 (Hart et al. 2020), including NCBI RefSeq 99 (Pruitt et al. 2012) and Swiss-Prot/TrEMBL ("UniProt: The Universal Protein Knowledgebase in 2021" 2021) databases and EggNOG v5 (Huerta-Cepas et al. 2019) and Pfam v33 (Mistry et al. 2021), downloaded in March 2021.

Given the highly repetitive nature typical of plant genomes (Mehrotra and Goyal 2014; Sahebi et al. 2018), I built a repeat library using a transposable elements annotator pipeline EDTA v1.8.4 (Ou et al. 2019) to predict LTR, TIR, and Helitron repeats and used RepeatModeller v2.0.0 for *de novo* identification of repeats (Flynn et al. 2020). The predicted repeat elements were combined with plant repeats from RepBase v22.08 (Bao, Kojima, and Kohany 2015) and used to mask the repetitive regions of the genome for more accurate gene annotation.

4.3.3 Comparative genomics: phylogenomics of *C. sativa* varieties

I used complete single-copy orthologs annotated with BUSCO v4.1.4 (Simão et al. 2015) and the odb10 Embryophyta lineage core gene set to infer phylogenetic relationship among Willow-alpha, other *C. sativa* varieties, and *Humulus lupulus* (outgroup) (Appendix Table C.1). I aligned protein sequences recovered using each orthologous group with Mafft v7.453 (using the argument *-auto*) (Nakamura et al. 218) and then used RAxML v8.2.12 (Stamatakis 2014) to build gene trees (PROTGAMMAAUTO and 100 bootstraps). I excluded fragmented genes from downstream analysis, which I defined as sequences with gaps in more than 67% of the sequence length in the multiple-sequence alignment; a species tree was estimated with ASTRAL-III v5.6.3 (C. Zhang et al. 2018), and relationships in the tree were evaluated with the relative frequency analysis in DiscoVista v1.0 (Sayyari, Whitfield, and Mirarab 2018). I rooted the final species tree using *H. lupulus* as an outgroup.

4.3.4 Correlative transcriptome and metabolome of Willow-alpha and three other varieties

Willow Biosciences Inc. performed chemical profiling, and I performed transcriptome analysis on three *C. sativa* varieties with high and medium leaf pigmentation (based on purple/red coloration) to compare to Willow-alpha, which lacks any obvious purple/red leaf pigmentation. CA19210 and CK19206 are two varieties with high leaf pigmentation, and Cali Kush is a variety with an intermediate intensity of leaf pigmentation as judged visually. All *C. sativa* plant material was obtained from Willow Biosciences Inc.

and was grown under identical conditions. Leaf tissues from these three varieties and Willow-alpha were harvested and frozen for subsequent metabolite and RNA extraction, as follows.

Willow Biosciences Inc. extracted total phenolics from ground frozen leaf tissues using acidified methanol (v/v, methanol:water:formic acid, 49.5:49.5:1), which was analyzed using high-performance liquid chromatography and mass spectrometry (HPLC-MS) in collaboration with the wine research center at the University of British Columbia, as described in Yan et al. (2020). All standards were purchased from Extrasynthese (Genay, France). Quantification of anthocyanins was based on a standard curve of cyanidin 3-O-glucoside (Cat. # 0915S) with malvidin 3,5-diglucoside (Cat. #0930S) as an internal standard. Quantification of flavonols was based on a standard curve of quercetin 3-O-glucoside (Cat. #1099) with baicalein (Cat. #1400S) as an internal standard. The two most abundant anthocyanins were Cyanidin-3-O-rutinoside and Peonidin-3-O-rutinoside, verified using LC-QTOF (Agilent) analysis against authentic standards (Cat #0914S and 0945, respectively). Three biological replicates per genotype were extracted for metabolite analysis.

Willow Biosciences isolated RNA from frozen leaf tissues using Invitrogen PureLink Plant RNA reagent, according to the manufacturer's instructions. Three biological replicates per genotype were extracted for RNA analysis. RNA sequencing was performed by Genome Quebec using Illumina NovaSeq 6000 PE protocol (2x100 bp). A total of 25 million reads were generated per sample and screened for possible contamination with Bloom filters as described for Willow-alpha RNAseq in section 4.3.2. All the reads without a match in any Bloom filter (default settings) were kept for quantification. Willow-alpha annotated transcripts were used as a reference, and quantification was conducted with Kallisto v0.46 (Bray et al. 2016). The transcript abundances estimated by Kallisto were imported in DESeq2 v1.32.0 (Love, Anders, and Huber 2014) and normalized based on the library sample size. Differentially expressed genes are calculated between the Willow-alpha reference and each of the other cultivars (Willow-alpha vs. CA210; Willow-alpha vs. CK206; Willow-alpha vs. Cali Kush). *P*-value statistical significance is adjusted with Benjamini–Hochberg for multiple-testing correction at FDR <0.01

(Benjamini and Hochberg 1995). A gene was considered differentially expressed if it has an absolute fold-change higher than 1.5 and *P*-value <0.05.

4.3.5 Identification of genes involved in the general phenylpropanoid, flavonoid, anthocyanin, and catechin pathways

I downloaded genes involved in the phenylpropanoid biosynthesis pathway, particularly those involved in the synthesis of flavonoids, anthocyanins, and catechins (Gutierrez et al. 2017; Saito et al. 2013) from the Arabidopsis information resource database (TAIR) (Reiser et al. 2016). I recovered the genes of the before-mentioned biosynthetic pathways in Willow-alpha by aligning the total proteins from Willow-alpha to TAIR by using a reciprocal best hit (RBH) approach. RBH is shown to have a low false-positive rate and is still broadly used in comparative genomics (Moreno-Hagelsieb and Latimer 2008; Hernández-Salmerón and Moreno-Hagelsieb 2020). Willow-alpha proteins with a reciprocal best match to the characterized flavonoid, anthocyanin, and catechin biosynthesis genes from *A. thaliana* were used for further comparative analysis and RNAseq quantification in *C. sativa*. Additional genes were identified through gene expression correlation because RBH homologs assignment is not fully effective and can miss paralogous genes. I looked for a correlation of gene expression between the flavonoid/anthocyanins key-pathway genes, *CHS*, *DFR*, *LDOX/ANS*, and other Willow-alpha annotated genes. I selected genes with a high positive Spearman correlation with the key-pathway genes, with an arbitrary threshold set at 0.7. Genes with high correlation are likely involved in the same process and included as genes potentially functioning in the biosynthetic pathway (Stuart et al. 2003; Serin et al. 2016).

4.4 Results

4.4.1 Genome assembly

We sequenced the genome of Willow-alpha using various short- and long-read-based approaches, including chromosome conformation capture, and assembled them into 131 final scaffolds with an NG50 of 80.2 Mb (Table 4.1). The primary Falcon assembly produced a portion of heterozygosity-induced overrepresented sequences, as indicated by a reconstructed genome size of ~1 Gb and a high percentage of BUSCO "complete – duplicated" genes (37.7%). To compensate, heterozygosity-induced scaffolds equaling 300 Mb were removed, and the reconstructed genome size was brought down to 731 Mb. A total of 527 scaffold gaps were closed, equaling 19% of the total gaps. I scaffolded the final assembly with the cs10 reference genome, with the longest scaffolds representing the ten known *C. sativa* chromosomes (Divashuk et al. 2014) with high synteny (Figure 4.2). In total, 90.8% of the Embryophyta core gene sets were recovered and counted as BUSCO "complete – single copy" in the final assembly (highlighted in the bottom row of Figure 4.2). The final assembly of Willow-alpha has the highest percentage of BUSCO "complete – single-copy" genes (Appendix Table C.2) when compared to the other chromosome-scale genome assemblies of *C. sativa* compared in this study.

Table 4.1. Willow-alpha assembly statistics at each assembly step. Starting with the first assembly (Falcon Unzip) on the top and ending with the final assembly step (ntJoin–Cs10) on the bottom. The final genome assembly is shown in bold in the bottom line. The assembly statistics for each step are listed; values are calculated for scaffolds longer than 1 kb. The estimated genome size for female plants (i.e., 810 Mb) was used to calculate the NG50 metric. The reconstruction size is calculated over the non-ambiguous genome characters and excludes scaffolding N gaps; the final reconstruction size, including Ns, is 831 Mb. The Embryophyta BUSCO core gene set (n=1,614) was used to estimate gene completeness with the BUSCO "complete – single copy" and "complete – duplicated" metrics.

Assembly stage	No. of Scaffolds	Largest scaffold (Mb)	NG50 (Mb)	Reconstructed size (Mb)	BUSCO single copy (%)	BUSCO duplicated (%)
Falcon Unzip	4,386	5.100	0.530	1051	61.2	37.7
Purge dups	2,685	5.100	0.432	732	90.6	7.1
SALSA	1,521	7.631	0.837	732	90.8	6.8
Racon	1,521	7.630	0.837	731	90.9	6.7
wtdbg2	19,454	4.629	0.115	919	87.9	5.6
ntJoin – wtdbg2	1,387	7.630	0.917	731	91.0	6.6
Sealer	1,387	7.630	0.917	731	90.8	6.8
ntJoin – cs10	**131**	**92.11**	**80.2**	**731**	**90.8**	**6.8**

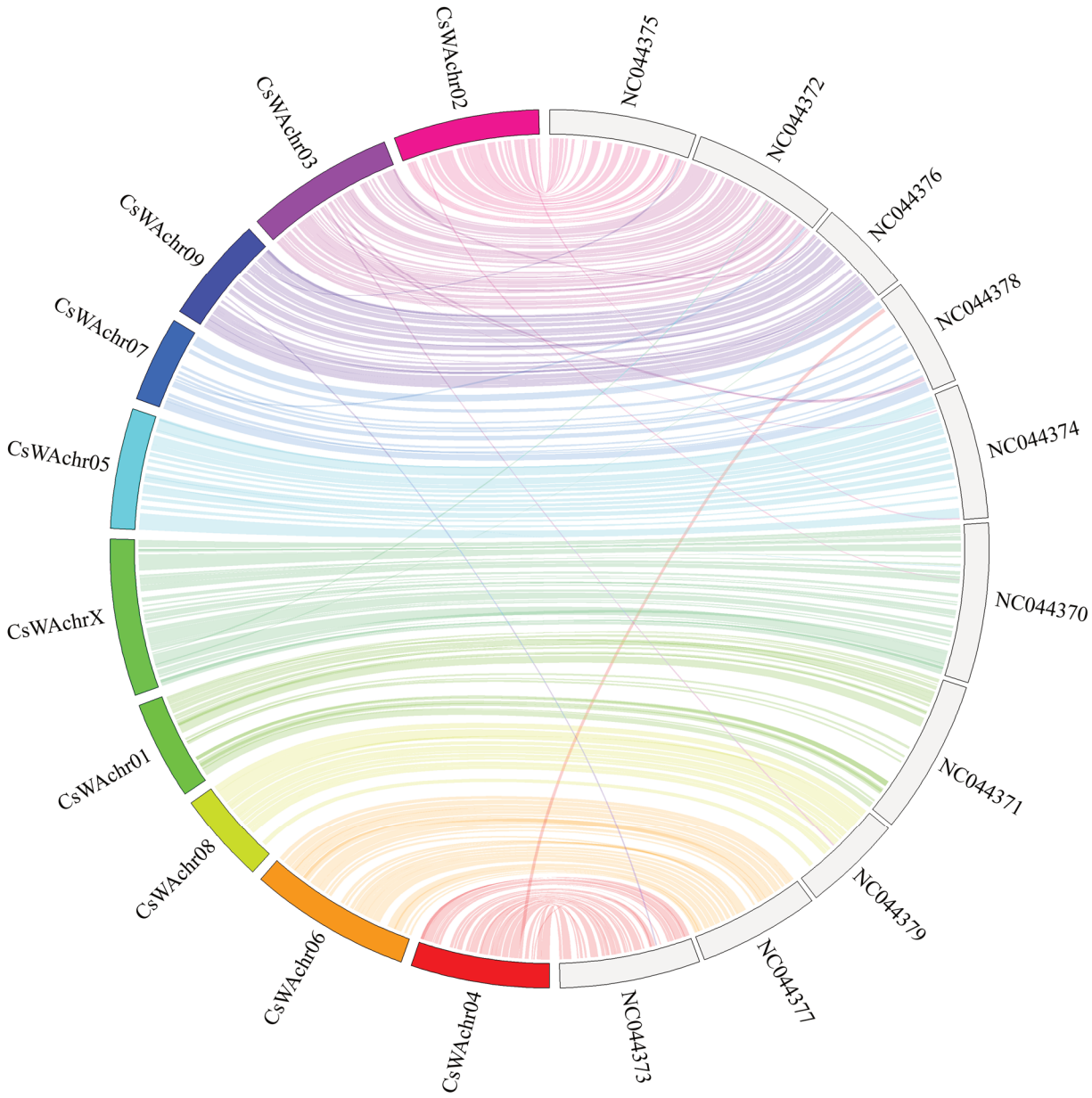

Figure 4.2 Jupiter plot of Willow-alpha genome assembly and cs10 reference genome. The plot shows the ten longest scaffolds, corresponding to the nine autosomal chromosomes and a sex chromosome shown in green (CsWAchrX in Willow-alpha and NC044370 in cs10). Each line is a synteny block representing a region with high sequence similarity between the two genomes, with Willow-alpha on the left (color) and cs10 on the right (gray).

4.4.2 Genome annotation

To arrive at the optimal genome annotation, I compared two state-of-the-art pipelines, MAKER and BRAKER. I evaluated the annotation quality based on the expected single-copy genes with BUSCO "complete," which is the sum of the "single copy" and "duplicated" and expected plant Pfam domains evaluated as conserved domain arrangements (CDA) (Table 4.2).

In the MAKER iterative annotation using genes predicted at the first iteration as supporting evidence in the second iteration, the total mRNAs decreased at the second iteration, from 41,027 to 35,474. Even though the number of mRNAs is lower at the second iteration than the first, the percentage of BUSCO "complete" and total CDA completeness increased from 86.7% to 90.7% and 88% to 93%, respectively, indicating more accurate predictions with the additional iteration.

Compared to the MAKER second iteration, which identified 35,474 mRNAs, BRAKER identified 41,912 mRNAs in a single step. Even though the two pipelines predicted a similar number of total mRNAs (Table 4.2), the number of annotated genes differed between the two tools. BRAKER median transcript length was shorter than for MAKER: 843 and 1,621, respectively. Overall, BRAKER attained a higher annotation accuracy, with 94.4% of the BUSCO annotated genes as "complete" and 96.8% total CDA completeness. MAKER prediction accuracy was lower for the BUSCO complete genes and total CDA, with 90.7% and 93.1%, respectively.

BRAKER annotation was chosen to be the better reference annotation because it has a higher annotation accuracy when compared to MAKER, as shown by the BUSCO "complete" metric and total Pfam CDA. This final gene dataset was filtered to remove incomplete and spurious Pfam domains (misannotations or LTRs). The final set of annotations contains 38,753 mRNAs and 36,097 protein-coding genes.

Table 4.2 - Number of annotated genes and transcripts, and gene length statistics for Willow-alpha, comparing MAKER and BRAKER pipelines. The annotation completeness is shown as BUSCO "complete" ("single copy" and "duplicated") and percent Pfam conserved domain arrangements (CDA).

Annotation	Total genes	Total mRNA	Median gene length (bp)	Median mRNA length (bp)	Median exon length (bp)	Median intron length (bp)	BUSCO complete (%)	Total CDA completeness (%)
MAKER first it.	23,373	41,027	3,160	1,676	138	156	86.6	88.6
MAKER second it.	22,517	35,474	3,015	1,621	140	153	90.7	93.1
BRAKER	38,559	41,912	2,436	843	123	152	94.9	96.8

4.4.3 Estimation of phylogenetic relationship among *C. sativa* varieties

A comprehensive phylogenomic analysis was performed based on BUSCO "complete – single copy" genes (Simão et al. 2015) with five publicly available chromosome-scale *C. sativa* genomes, Willow-alpha and *Humulus lupulus* (hops) in the outgroup (Appendix Table C.1). A total of 8,307 genes from 953 single-copy orthogroups were used for phylogeny. The orthogroups are selected for phylogeny because they are reconstructed in at least six out of the seven genotypes. The genome with the lowest number of reconstructed genes is the *H. lupulus* outgroup with the most fragmented genome, possibly due to its draft genome assembly and the large genome size and higher repeat content than *C. sativa*. Willow-alpha is the *C. sativa* genome with the highest number of reconstructed genes, 90.8% BUSCO "complete – single copy" (Appendix Table C.2). I filtered only a small fraction of genes for the data set used for phylogenetic inference (21 of 8,328 initial genes). The tree inferred from these genes (Figure 4.3) groups together the THCA–CBDA hybrid chemotype varieties, Cannbio-2 and cs10, with Willow-alpha being the closest genotype to these two. PK-formed a cluster with these three varieties. The JL landrace, which lacks chemotype characterization, groups with the previous four narcotic-medical genotypes. Its long terminal branch length indicates a substantial distance from the other *C. sativa* genotypes (Figure 4.3). The hemp-type *C. sativa* (Finola) forms a group with the other narcotic-medical genotypes.

Branches in the tree topology (Figure 4.3) are supported by high local posterior probabilities (LPPs) around 1.0 for all the nodes (except 0.98 for the branch below the Willow-alpha, cs10, Cannbio-2 cluster). The average across branches normalized quartet score in ASTRAL-III is 0.425, possibly indicating strong incomplete lineage sorting (ILS) and/or introgression among the *C. sativa* varieties.

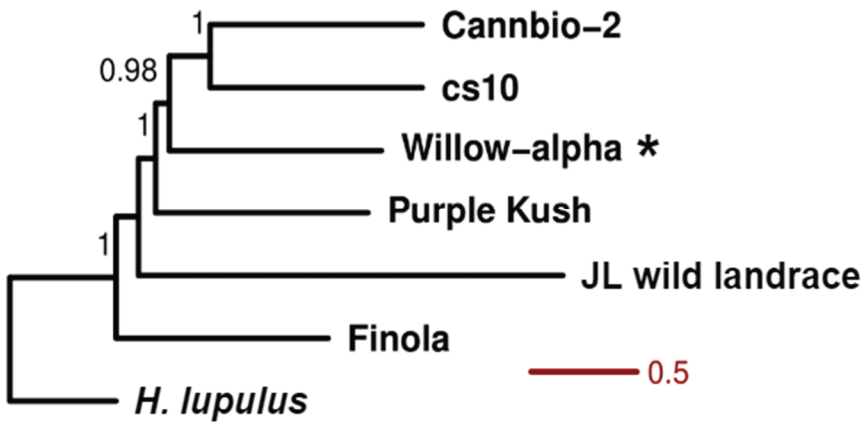

Figure 4.3 – Species tree derived from phylogenomic inference of *C. sativa* strains and the Cascade cultivar of hops (*Humulus lupulus*). The numbers on the internal nodes show the local posterior probabilities (LPP). The Willow-alpha strain is shown with an asterisk. The scale indicates the number of substitutions per site.

4.4.4 *Cannabis sativa* leaf pigmentation

Willow-alpha has a bright green leaf phenotype (Figure 4.4), likely reflecting a low concentration of anthocyanins in this variety compared to other varieties such as the previously reported PK. Seeds or cuttings of the PK variety were not available. Therefore, we chose three additional *C. sativa* genotypes, CA19210, CK19206, and Cali Kush, compared to Willow-alpha because they have variable degrees of leaf pigmentation. Willow-alpha has bright green leaf blades and unpigmented leaf mid ribs, whereas CK19206 leaf blades were darker green with dark purple/black midribs and petioles (Figure 4.4).

CA19210 and Cali Kush displayed an intermediate phenotype with bright green leaf blades but purple-colored midribs and petioles (Figure 4.4).

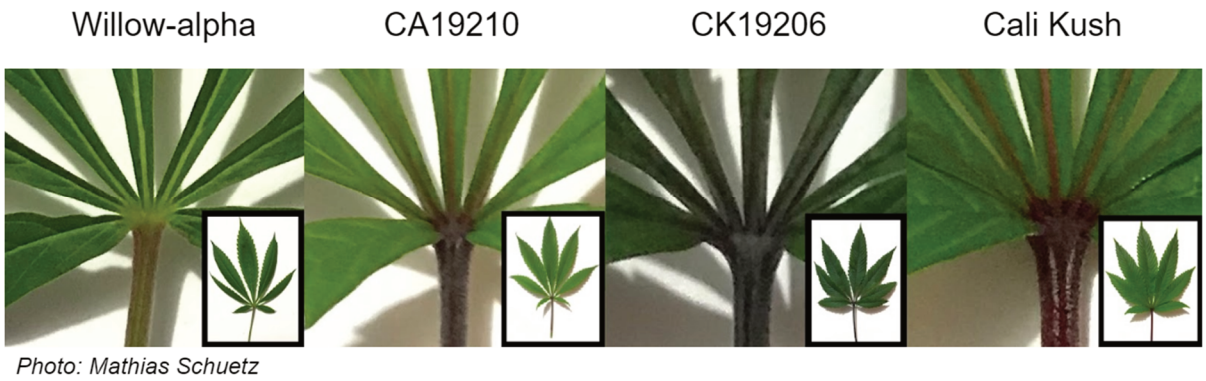

Figure 4.4 - Rachis, petiole, and leaf pigmentation phenotypes of Willow-alpha (green phenotype), CA19210 and CK1926 (dark purple phenotypes), and Cali Kush (intermediate pigmentation phenotype). The images show a highlighted region of the plant leaf and the overall leaf morphology in the bottom right rectangle.

4.4.5 General phenylpropanoid, flavonoids, and anthocyanin biosynthesis genes in *C. sativa*

I recovered a total of seventeen flavonoid biosynthetic gene homologs in Willow-alpha using a reciprocal best hit (RBH) approach. The relevant genes assigned to Willow-alpha, and compounds produced at each pathway stage, are summarized in Figure 4.5 and Appendix Table C.3. The general phenylpropanoid biosynthesis pathway includes three enzymes, phenylalanine lyase (PAL), cinnamate 4-hydroxylase (C4H), and 4-coumarate lyase (4CL), and their protein sequence are highly conserved between *A. thaliana* and Willow-alpha. PAL is the entry point for the general phenylpropanoid pathway via the determination of phenylalanine (Vogt 2010). Chalcone synthase (CHS) is the first commitment step toward the flavonoid biosynthesis branch of the phenylpropanoid biosynthesis pathway; it converts *p*-coumaroyl CoA to naringenin chalcone. Two different chalcone isomerase genes were identified and included: *CHI* and a putative isoform, *CHI-L1*. Flavanone 3 (*F3H*) and, flavanone 3' hydroxylase (*F3'H*), flavonol synthase (*FLS*) *A. thaliana* genes have a reciprocal best hit (RBH) with Willow-alpha. We

observed the production of vitexin and luteolin metabolites that are the products of flavone synthase (FNS). The enzyme is absent in *A. thaliana* (Martens and Mithöfer 2005), and I did not recover it in Willow-alpha using the current approach. Flavonols are synthesized from dihydroflavonols (dihydrokaempferol and dihydroquercetin) through the action of flavonol synthase (FLS), which gene sequence recovered in this study. Subsequently, UGT73C6 glycosyltransferase facilitates the downstream glycosylation of kaempferol and quercetin. The genes for these two enzymes were also recovered in this study.

Entry into the anthocyanin-specific branch of the flavonoid biosynthesis pathway is regulated mainly by dihydroflavonol 4-reductase (DFR) and leucoanthocyanidin dioxygenase (LDOX), which is also described as anthocyanin synthase (ANS) in other species (Falcone Ferreyra, Rius, and Casati 2012). LDOX enzyme converts leucocyanidin to cyanidin, an anthocyanin compound found in many plant species (Abrahams et al. 2003). All these genes are found in this study. Cyanidins are typically glycosylated on the hydroxyl group at C3 and C5 positions and then sequestered into plant vacuoles, resulting in red/purple pigmented cells. The two main UDP-glycosyltransferases involved in cyanidin glycosylation in *A. thaliana* are UGT78D2 and UGT75C1 (Tohge et al. 2005), and homologs for both exist in the Willow-alpha genome. A diversity of other sugars can be joined to cyanidin, and therefore other UGT genes may be involved in this process. I also recovered genes coding for anthocyanidin reductase (ANR) and laccase (LAC15) (Rani et al. 2011), which are involved in catechins production.

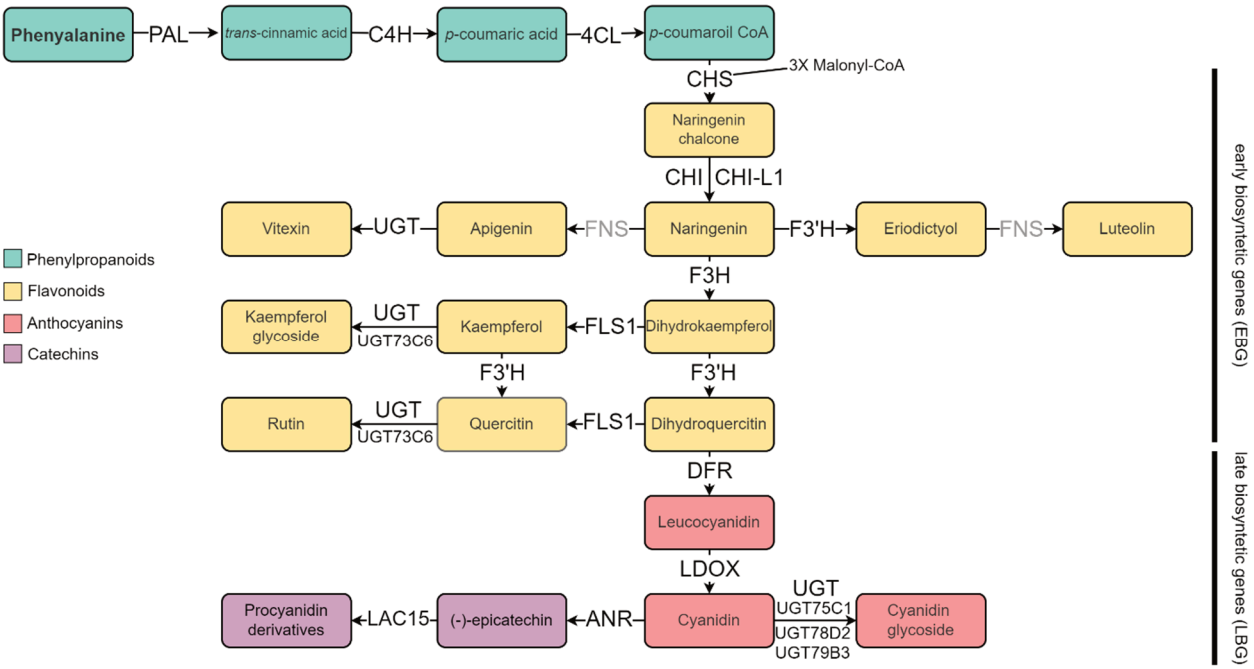

Figure 4.5 - General phenylpropanoid, flavonoid, anthocyanin, and catechin biosynthesis pathways, showing the synthesized compounds in squares and the corresponding enzymes on the arrows. The first three enzymes show the general phenylpropanoid pathway (blue); all the flavonoids, including Chalcone Synthase (CHS), are part of the early biosynthetic genes (EBGs), the enzymes, including DFR and enzymes regulating downstream reactions, are part of the late biosynthetic genes (LBG). We observed the production of vitexin and luteolin, produced by reactions from FNS – gray (not recovered through the current computational approach).

4.4.6 Flavonoid/anthocyanin gene expression in Willow-alpha and three *C. sativa* varieties

I evaluated expression levels of genes in the flavonoid/anthocyanin biosynthesis pathway for Willow-alpha and three other *C. sativa* varieties, CA19210, CK19206, and Cali Kush (Figure 4.6; Appendix Table C.4). The gene expression for *C4H*, *PAL*, and *4CL* remained similar for all four varieties. On the other hand, *CHS* gene expression was highest in varieties with more pronounced leaf pigmentation. *CHS* had a log2 fold-change of 1.90 higher expression in CA19210 than Willow-alpha. *CHS* is also highly expressed in the CK19206, with a log2 fold-change of 2.71, but not significantly

different in Cali Kush Willow-alpha. The early flavonoid biosynthetic genes, *F3H*, *F3'H*, and *FLS1*, are upregulated in all the three varieties compared to Willow-alpha, with *F3H* having the highest log2 fold-change of 4.43 in CA19210. As with *CHS* expression, expression of *DFR* and *LDOX/ANS* genes (producing leucocyanidins and cyanidins from precursory flavonols; Figure 4.5) had a log2 fold-change of 1.98 in CA19210 genotype and were also expressed at much higher levels in CK19206, with log2 fold-change of 2.70 compared to Willow-alpha. Among the UGTs involved in the glycosylation of cyanidin, (2) *UGT78D2* is the only gene with a significantly different (higher) gene expression in the higher pigmentation varieties compared to Willow-alpha. In contrast, gene expression for (1) *UGT73C6*, (3) *UGT76C1*, and (4) *UGT79B3* were not differentially expressed in all the samples in the comparison. The gene expression level of the catechin-producing enzymes, LAC15, was not differentially expressed across any of the varieties. *ANR* gene expression was nearly absent in all the replicas and genotypes. In summary, the genes with significantly upregulated gene expression in *C. sativa* varieties with visually more pigmented leaves than Willow-alpha are *CHS*, *F3H, F3'H, FLS1, DFR, LDOX/ANS*, and (2) *UGT78D2*.

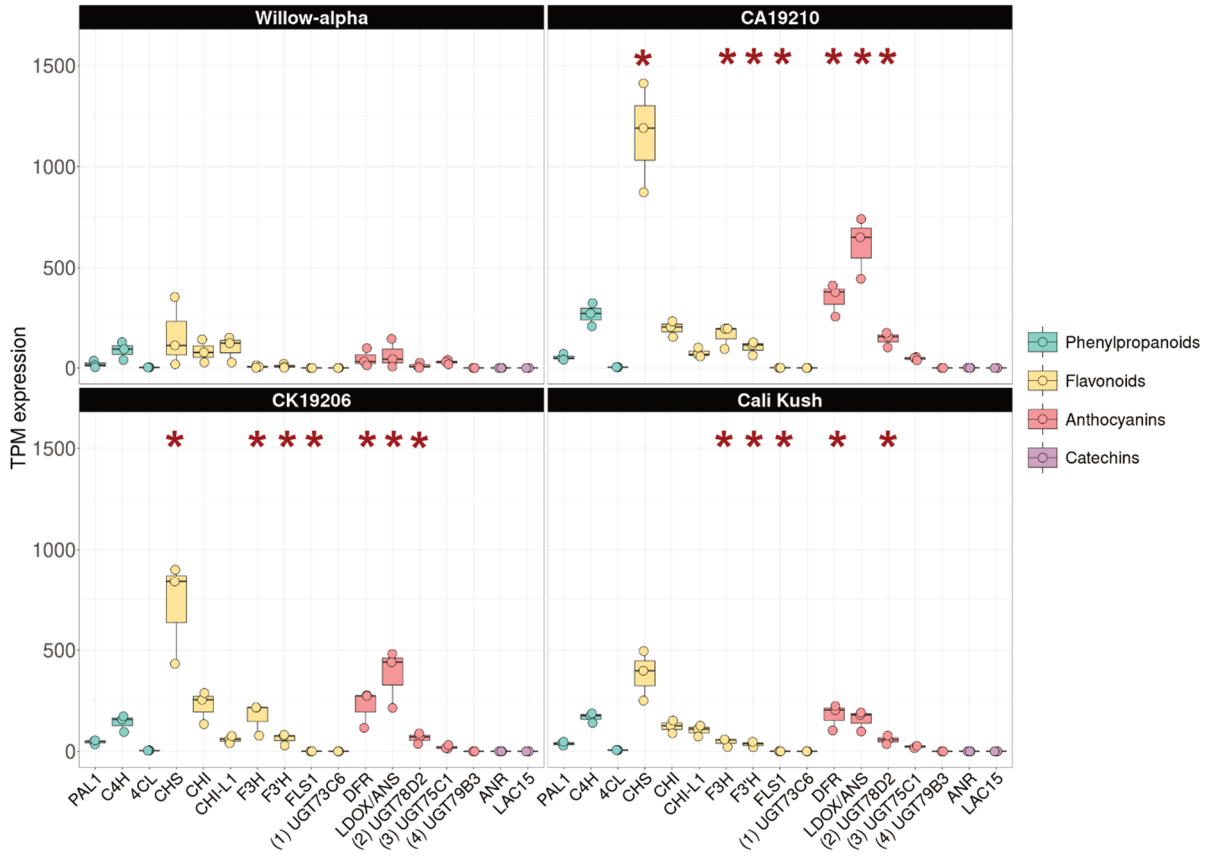

Figure 4.6 - Gene expression of phenylpropanoid, flavonoid, and anthocyanin enzymes in leaf for Willow-alpha, CA19210 and CK19206 (dark purple phenotype), and Cali Kush (medium pigmentation) varieties. The gene quantification for three biological replicates is shown as transcript per million (TPM). Asterisks indicate differential expression for the dark purple and medium pigmentation varieties (treatment) vs. Willow-alpha (control).

4.4.7 Flavonoid/anthocyanin metabolite profiling in Willow-alpha and three *C. sativa* varieties

We identified six anthocyanins and nine flavonoids in the *C. sativa* varieties analyzed (Appendix Table C.5, Appendix Table C.6, Appendix Table C.7 for more details about the mass spec quantification). Anthocyanin quantification is shown in Figure 4.7, including cyanidin and peonidin, all identified as O-glycosides. The most abundant anthocyanin found in all *C. sativa* varieties is cyanidin

rutinoside. Moreover, cyanidin rutinoside was much more abundant in the *C. sativa* varieties, showing a more pigmented leaf phenotype such as CA19210, CK19206, and Cali Kush.

Peonidin rutinoside, an O-methylated derivative of cyanidin rutinoside, is the second most abundant anthocyanin found in CA19210, CK19206, and Cali Kush but was nearly undetectable in Willow-alpha. CA19210 is the only *C. sativa* variety that accumulated cyanidin sophoroside, resulting in that variety having the highest total amount of anthocyanins (Figure 4.7). Anthocyanins have absorption maxima around 520nm (Vivar-Quintana, Santos-Buelga, and Rivas-Gonzalo 2002). Chromatogram peaks for the interval of 512–528 nm show that the anthocyanin absorbances in the four varieties analyzed are consistent with the mass spectrometry-based anthocyanin quantification (Appendix Figure C.2). We also detected several previously identified flavones and flavonols (Jin et al. 2020; Izzo et al. 2020; Clark and Bohm 1979), including two vitexin glycosylated at the 3-O and 7-O positions, isovitexin, apigenin 3-O-glucuronide, kaempferol glucuronide, quercetin glucoside, and glucuronide, and rutin (Appendix Table C.5, Appendix Table C.6, Appendix Table C.7 for more details about the mass spec quantification). The pattern of anthocyanin accumulation in each variety matches well with the corresponding gene expression patterns for the key genes (*CHS*, *DFR*, and *LDOX/ANS*) in the flavonol and anthocyanin branches of the phenylpropanoid biosynthesis pathway in each variety (Figure 4.5).

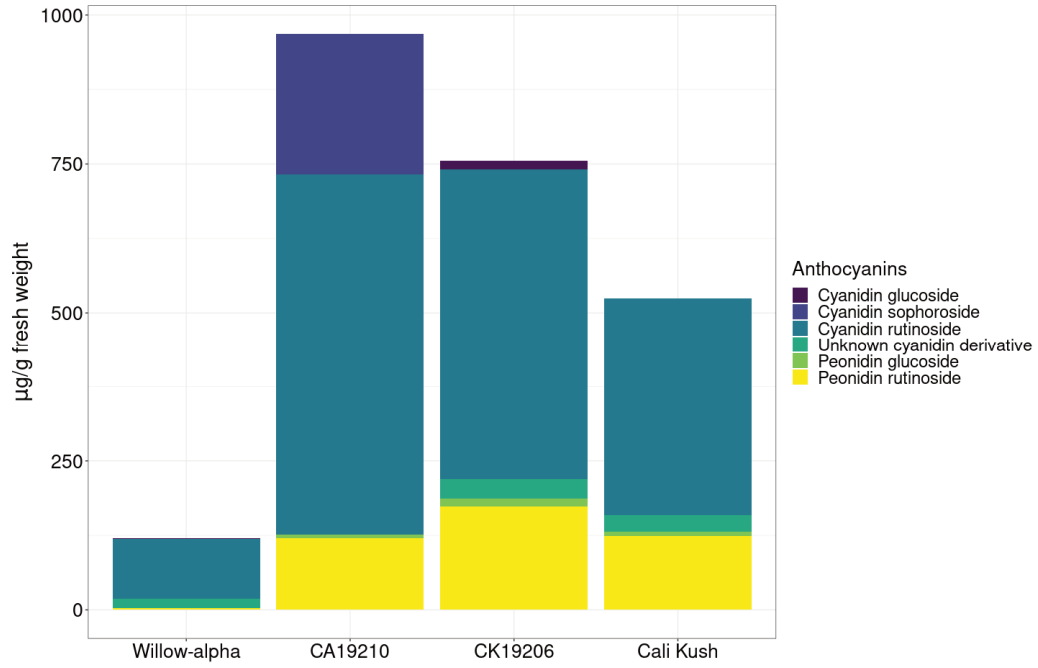

Figure 4.7 - Anthocyanin profiling of leaf samples from Willow-alpha and three anthocyanin-producing varieties.

4.5 Discussion

4.5.1 Genome assembly and annotation of Willow-alpha *C. sativa* variety

Cannabis sativa is a valuable plant given its economic importance as an established food and fiber crop and as a medical and recreational drug due to the large number of specialized metabolites that it can produce (Elsohly and Slade 2005). Here I present the genome assembly of Willow-alpha, a narcotic-medicinal-type *C. sativa* variety. I assembled the genome with short and long reads, scaffolded with a chromatin interaction approach from Hi-C PhaseGenomics (Belaghzal, Dekker, and Gibcus 2017). I scaffolded the assembly with the cs10 reference genome (Grassa et al. 2021), producing a chromosome-scale assembly with an NG50 of 80.2 Mb. The BUSCO "complete – single copy" score, which quantifies the reconstruction rate in the genic space, is 90.8%, higher than the other published

chromosome-scale *C. sativa* genomes (van Bakel et al. 2011; Braich et al. 2020; Grassa et al. 2021; Gao et al. 2020; Laverty et al. 2019). A phylogenomic analysis of Willow-alpha with other *C. sativa* varieties with publicly available genomes shows its similarity with the narcotic-medical varieties such as Purple Kush (PK), cs10, and Cannbio-2. PK has a dominant production of THCA (van Bakel et al. 2011), whereas cs10 and Cannbio-2 are the last two to have a balanced production of both THCA and CBDA (Grassa et al. 2021; Braich et al. 2020).

We need robust and accurate genome annotations to study *C. sativa* physiology and secondary metabolite biosynthesis. I used RNAseq evidence from four different plant tissues derived from Willow-alpha as supporting evidence. I explored two different genome annotation pipelines for Willow-alpha, MAKER (Holt and Yandell 2011) and BRAKER (Brůna et al. 2021), to identify which gives a more accurate annotation for highly complex and repeat-rich genome structures than the ones found in *C. sativa* (Mehrotra and Goyal 2014; Sahebi et al. 2018). I examined which pipelines provide a higher number of conserved genes based on reference sets of single-copy genes and Pfam domains in plants. Both pipelines reconstructed a large fraction of expected single-copy genes and Pfam domains. However, the BRAKER pipeline had a higher reconstruction of BUSCO "complete" genes (94.4% BUSCO "complete," including "single copy" and "duplicated") and higher Pfam domain reconstruction (96.8%). The BRAKER pipeline showed an improved annotation compared to MAKER and generated a final set of 38,599 genes. After gene filtering, the final gene annotation presented 36,097 protein-coding genes.

4.5.2 General phenylpropanoids and flavonoids biosynthesis genes in *C. sativa*

To further investigate the utility of the Willow-alpha genome annotation, I examined flavonoid biosynthesis genes identified in this assembly using the reciprocal best hit (RBH) approach with the well-characterized *A. thaliana* model system. The core flavonoid biosynthesis pathway is conserved among different plant species (Wen, Alseekh, and Fernie 2020). I identified the general phenylpropanoid biosynthetic genes in the pathway as well as the later flavonoid and anthocyanin biosynthesis specific

genes in this *C. sativa* variety. I hypothesized that the *C. sativa* leaf pigmentation phenotypes we observed in different varieties result from the differential accumulation of anthocyanins, which would, in turn, be regulated by differential gene expression of the biosynthesis genes. Previous studies have characterized some of the flavonoids found in *C. sativa*, such as flavones (luteolin, apigenin, and vitexin) and flavonols (quercetin and kaempferol (Izzo et al. 2020)). Those groups exist as primary structures or conjugated O-glycosides or C-glycosides (Flores-Sanchez and Verpoorte 2008; Pellati et al. 2018; Ross et al. 2005) and are found in flowers, leaves, stems, and pollen (Vanhoenacker et al. 2002; Ross et al. 2005). In contrast, the anthocyanin branch of the flavonoid biosynthesis has not been characterized thus far in *C. sativa*.

I identified six different anthocyanins and found that cyanidin rutinoside was abundant across all four different *C. sativa* accessions analyzed. Peonidin rutinoside was the second most abundant anthocyanin found but was absent in the Willow-alpha variety, whereas the CA19210 variety uniquely accumulated substantial amounts of cyanidin sophoroside. These three anthocyanins account for the bulk of all anthocyanins identified and matched well with the leaf pigmentation phenotypes observed among the different *C. sativa* varieties analyzed here (Figures 4.4, 4.6).

I hypothesized that flavonoid biosynthesis, specifically the anthocyanin biosynthesis branch, should have higher gene expression in the varieties that contain higher amounts of anthocyanins. I found that three genes in the early stage of the flavonoid pathway, *F3H*, *F3'H*, and *FLS1*, are significantly overexpressed in CA19210, CK19206, and Cali Kush compared to Willow-alpha. Moreover, the genes coding for anthocyanin-specific pathway enzymes, namely *DFR*, *LDOX/ANS*, and *UGT78D2*, were also significantly upregulated in these genotypes compared to Willow-alpha (Figure 4.6, Table C.4). Similar correlative transcriptome and metabolome profiling have been performed in several other plant species. For example, a study of anthocyanins in poplar (Zhuang et al. 2019) demonstrated that a colored-leaf variety had increased expression of *DFR*, *F3'H*, *LDOX/ANS*, and *UFGT* genes compared to green-leaf varieties. Similar findings for purple leaf pigmentation phenotypes were reported in *Brassica napus* (Mushtaq et al. 2016; He et al. 2021), with upregulation of *F3H*, *F3'H*, and *LDOX/ANS*, in *B. napus*

varieties which have highly pigmented leaves compared to unpigmented varieties. The same study in *B. napus* highlights that *DFR*, *UGT75C1*, and *UGT79B1* are co-expressed with genes involved in regulating anthocyanins. Here I identified *CHS* as the most obviously upregulated gene in varieties with the highest anthocyanin accumulation (Figure 4.6, Table C.4). I found that the level of *CHS* gene expression is highly correlated with the anthocyanin content in the *C. sativa* accessions that we analyzed. *CHS* is an important gatekeeper for carbon flowing into the flavonoid biosynthesis pathway because it directs carbon flux from the general phenylpropanoid biosynthesis pathway to the flavonoid biosynthesis pathway (Zhang et al. 2017; Shirley et al. 1995). This gene is highly expressed in other taxa, such as varieties of *A. thaliana* with purple/red leaf color (Clark and Verwoerd 2011) and red-colored leaf poplar varieties (Zhuang et al. 2019). Thus, *CHS* expression similarly indicates anthocyanin accumulation and leaf pigmentation in *C. sativa*.

4.5.3 Future directions

Several transcription factor groups modulate the gene expression of flavonoid biosynthetic genes in *A. thaliana* at the early and late biosynthetic stages of the flavonoid pathway. In *A. thaliana,* the gene expression of early flavonoid biosynthetic genes such as *CHS*, *CHI*, *F3H*, and *FLS1* is regulated by MYB transcription factors: MYB11, MYB12, and MYB111 (Mehrtens et al. 2005; Stracke et al. 2007). Moreover, *A. thaliana* PRODUCTION OF ANTHOCYANIN PIGMENT 1 (MYB75/PAP1) overexpression in a wide range of different plant species is sufficient to induce anthocyanin accumulation and red/purple phenotypes in these plants (Li et al. 2010; Zhou et al. 2008; Skaliter et al. 2019; Gatica-Arias et al. 2012). These studies demonstrate that the differential expression of key transcription factors can induce the overexpression of flavonoid/anthocyanin biosynthesis genes resulting in the accumulation of these compounds. Here I have characterized four *C. sativa* varieties with varying flavonoid accumulation phenotypes, and I have found that elevated expression of key biosynthesis genes correlates well with those phenotypes. The genome of Willow-alpha that I constructed and annotated in this study provides a common reference to finding genes that are differentially expressed in other

C. sativa varieties and can also be used to identify the regulatory genes which are putatively involved in regulating those processes. The potential of comparative genomics between these and other *C. sativa* varieties is the next step in understanding how the genomic structure impacts the expression of genes of interest. Future studies using whole-genome assemblies for these or other *C. sativa* varieties could study the physical organization of genetic loci, their 5'-untranslated promoter regions, and proximity to repeat regions that may influence gene expression. The current project metabolomics and gene expression profiling represent the starting point for deeper analysis for future genomic comparative studies.

Chapter 5: Conclusion

Next-generation sequencing technologies have led to the assembly of thousands of non-model organism genomes, and many others are soon to come. Among the drivers that influence the choice of sequenced genomes, arguably, the two most important factors are the novelty of characterizing new species in each taxonomic family and their economic importance (Vallée, Muñoz, and Sankoff 2016; Marks et al. 2021). Newly sequenced genomes can be important in thriving economic sectors, such as Canada's forestry and *Cannabis sativa* industries. The main application of sequenced genomes is to create a reference with *de novo* assembly, which is the process of putting together the sequenced reads and generating a representation of the genome. However, genome assembly alone does not provide sufficient information to study a species' biology; annotating its assembled genome is important for obtaining biological insights from sequences. Genome annotation is the process of assigning putative functional roles to motifs or stretches of sequence in a genome. The core feature of genome annotations is the protein-coding genes; other functional elements are annotated in relation to the surrounding protein-coding genes. Genome annotation also enables studies of comparative genomics, where annotations from a genome of interest are compared to annotations from other evolutionarily related species. A growing number of available genomes with annotations provide an incredible opportunity to study whole-genome features that otherwise are studied one sequence at a time with low-throughput methods. The annotation provides a community resource for others to study species' biology. Hence, my thesis focuses on the assembly and annotation of non-model organisms of evolutionary and economic importance, their comparative study with other related species, and the discovery of unique biological features. Depending on the species, I have used different methods to assemble, annotate and compare the sequenced genomes to answer various questions related to the species' biology.

5.1 Comparative study of annotated genomes

Comparative analysis is a fundamental part of my thesis work. In chapter two of my thesis, I present the genome annotations of four North American spruces (gen. *Picea*). The sequencing of four spruce genomes provides an unprecedented dataset to study their evolution and adaptation to different environments. The spruce genomes are sequenced with a combination of short and long-read technology. The first is the dominant technology for assembly; the long reads technology was limited to only two species sequenced at low coverage, thus not used in the primary assembly. The large genome size and repetitiveness of spruce posed a significant challenge in assembling a contiguous genome with short reads. Based on benchmarking on expected single-copy genes, the genome assemblies lacked expected protein-coding genes, likely because of the fragmented nature of the assemblies. Naive comparisons of genome annotations from draft-quality genomes can easily lead to erroneous conclusions. To overcome this issue, I applied a common genome annotation strategy among the four genotypes and employed combined evidence to recover as many genes as possible in all the taxa. The genome annotations showed comparable completeness among the spruces and similar completeness to other annotations in pines and spruce. As part of my research objectives, I found several molecular markers unique to each genotype highlighted in conifer population studies, including local adaptation, abiotic stress response, and plant development.

As part of my thesis work in chapter three, I have also assembled, annotated, and analyzed the genome of *P. strobi*, a spruce insect pest whose genomic features were not described before my work. The genome size was validated by flow cytometry, considered the experimental standard for a genome size estimate. Both the *in silico* and experimental results showed the expanded genome size of *P. strobi* compared to species from the same family (Curculionidae). Furthermore, the studied genome has a recent repeat expansion, possibly contributing to the large genome size. I also compared the repeat composition of *P. strobi* to eight Curculionidae species. In this chapter, I used several computational methods for

repeat annotation and quantification, involving the repeat annotation of the genome assembly, the annotation of unassembled read clusters, and k-mer representations.

Although I have successfully designed and performed a comparable annotation strategy, the genome annotations described in chapters two and three are still at the draft stage. Because the underlying genome assemblies are highly fragmented, I still would not claim comprehensive annotation of all elements in the genome of spruces and *P. strobi*. As evident from the annotation evaluation scores by BUSCO (Simão et al. 2015) and DOGMA (Kemena, Dohmen, and Bornberg-Bauer 2019), annotations of protein-coding genes are not complete in draft genomes, even when using supporting evidence from multiple gene expression libraries and reference protein sequences. This issue is more common for genomes assembled with short reads, which generate fragmented genomes, especially for complex genomes, such as those studied in this thesis. The resulting scaffolds consist of hundreds of thousands of disconnected pieces that do not contain full-length genes. The state-of-the-art annotation pipelines cannot put gene fragments across multiple scaffolds together. The use of more recent long-read sequencing technologies, such as those from Pacific Biosciences or Oxford Nanopore Technologies, usually yields more contiguous genome assemblies when used in the primary steps of assembly (Amarasinghe et al. 2020). In principle, if a read length is longer than the length of a given repeat sequence, uniquely connecting the sequences flanking the repeat, it can resolve the repeat, improving the contiguity of the assembly. I recognize the importance of generating an improved and more contiguous genome assembly using long reads, which can resolve longer repeats and improve the overall assembly contiguity. The assembly of long reads sequenced at high coverage will provide the opportunity to generate higher quality reference genomes and annotate a large number of genes. Long-read technology applied to transcriptome sequencing produces full-length transcripts (Hardwick et al. 2019) that, coupled with a more contiguous genome assembly, additionally improve the annotation of genes.

Considering the trade-off between the quality of assembled genomes and the number of sequenced/assembled genomes per taxonomic family, a sensible choice for comparative genomics is to use long-read sequencing for a representative species (model genome), combined with a series of lower

quality genomes with short reads to increase the number of species per taxonomic family, as advocated by the 10 KP consortium (Cheng et al. 2018). A model genome in each taxonomic family can then be used to generate high-quality genome annotations by providing matching RNAseq samples with the sequenced genome. The annotations can then be transferred from one species to another, taking advantage of the extensive work done on the model genome. Several newly developed tools, such as LiftOff (Shumate and Salzberg 2020) and nf-LO (Talenti and Prendergast 2021), take advantage of available high-quality genome annotations and transfer annotations between species. Newly assembled species are then more efficiently annotated using high-quality annotation from closely related species rather than *de novo* annotating each genome.

5.2 Pangenomes: joint analysis of gene annotations for comparative genomics

Genome annotations from closely related species as those of gen. *Picea* contain common genes that can be presented jointly or used as a reference through the pangenome representation (Computational Pan-Genomics Consortium 2018). The use of pangenomes provides an efficient representation of variable genes in individual species as already described for crops like wheat (Montenegro et al. 2017) and gen. *Brassica* (Bayer et al. 2018; Dolatabadian et al. 2019).

The homology-based strategy for building pangenomes (Hu, Wei, and Li 2020) involves the *de novo* assembly and annotation of independent genomes: the gene annotations are clustered into gene families, and shared genes across all the species are classified as “core genes.” The work presented in my thesis is a step towards a pangenome representation for the gen. *Picea* based on OrthoFinder protein clustering (Table A.8, Figure A.3). The homology-based strategy strongly relies on the genome assembly and annotation completeness: improved and more contiguous genome assemblies will produce an improved pangenome representation of the gen. *Picea*. The “map-to-pan” is another possible pangenome generation, where the newly assembled and annotated species are mapped against a high-quality reference

genome (Hu, Wei, and Li 2020). The "map-to-pan" strategy is warranted for gen. *Picea* in case a chromosome-scale and high-quality genome assembly is available for one of the presented species.

5.3 Integration of genome, transcriptome, and metabolome for *C. sativa* breeding

As part of my thesis work, I generated a chromosome-scale genome assembly of Willow-alpha, a *C. sativa* proprietary variety. I assembled the genome of Willow-alpha with a combination of short and long reads and scaffolded it with chromatin conformation Hi-C capture. The assembly pipeline ultimately gave high genome contiguity and gene completeness, providing a solid base for the genome annotation; I performed the annotation of the genome of Willow-alpha. The annotation, which resulted in high completeness, was used as a reference to quantify the differential gene expression between Willow-alpha, a variety producing a low leaf pigmentation, and three other cannabis varieties that produce a high and medium leaf pigmentation. I first recovered the genes involved in the flavonoid and anthocyanin biosynthetic pathways. I identified the differentially expressed genes, which I hypothesize regulate the production of anthocyanin metabolites in varieties with high leaf pigmentation. In addition, the use of metabolomics profiling coupled with the visually observable pigmentation phenotypes of red/purple pigmentation provided additional insight into the molecules produced by *C. sativa* varieties.

This study reveals the importance of integrating transcriptome and metabolome profiling to develop new strategies for selective breeding in *C. sativa*. The combination of these two "omics" methods has the potential to discover key genes that affect plant physiological processes, such as the production of anthocyanins. The list of differentially expressed genes obtained from Chapter 4 can be manipulated with genomic techniques (Rönspies et al. 2021) or be selected in selective breeding projects (Wang et al. 2020) to change the color of the plant varieties.

5.4 Limitations in quality benchmarking tools

Although benchmarking is fundamental for assessing the quality of the annotations in comparative studies, the most common tool for gene benchmarking, BUSCO (Simão et al. 2015), has several limitations, as described by the review in Jauhal and Newcomb (2021). The first limitation is that BUSCO only considers single-copy genes for evaluation; therefore, it does not apply to all genes in the annotation. Even if BUSCO evaluates the degree of gene duplication through the "complete – duplicated" metric, it does not consider multi-copy or paralogous genes for benchmarking. This type of gene is the most common in eukaryotes, and currently, there are no benchmarking tools that perform this task. Furthermore, BUSCO summarizes the completeness of the gene annotations through a unique percent score. While it still provides a good start in comparing annotations from related species and assigns a baseline quality in comparing different annotation methods, more than a single value is necessary to describe the quality of all the annotated genes in the genome.

I warrant the implementation of a new score that evaluates the quality of possibly all of the genes in the annotation. Sequencing a larger number of species per taxonomic family makes it possible to align genomic regions together, identify conserved high-synteny regions (Exposito-Alonso et al. 2020), and compare their genes. The score should consider both protein sequence from the annotated genes and the genomic regions around the gene to evaluate the conservation among species in taxonomic families (Exposito-Alonso et al. 2020). Benchmarking against conserved annotations will allow to score annotations in newly assembled genomes, and using such a scoring system will make it possible to expand the single completeness score of BUSCO to multi-copy genes.

5.5 Comparison of current pipelines for genome annotation and opportunities in the field

The high contiguity of the Willow-alpha genome, whose primary assembly is performed with long reads, and the availability of RNAseq libraries allowed the exploration of more advanced annotation strategies.

Chapter four compares two common pipelines for genome annotation: BRAKER (Brůna et al. 2021), a new pipeline for genome annotation, and MAKER (Holt and Yandell 2011), a well-established annotation pipeline. In my study, BRAKER showed improved annotation completeness when compared to the MAKER pipeline. The improved annotation is likely due to two major innovations implemented by BRAKER. The first breakthrough innovation is the exploitation of multiple libraries of RNAseq reads, used as read alignments and without the necessity to assemble them into transcripts as required by MAKER. The second innovation is a self-training strategy for the gene model in AUGUSTUS (Stanke et al. 2008). The difficulty in selecting genes for model training and the time-consuming and laborious model training process are the major obstacles to creating a customized gene model. Users with limited expertise in the annotation field often skip this stage and prefer to use gene models from model organisms provided by the AUGUSTUS software (Stanke et al. 2008); this choice impacts the genome annotation because species-specific gene features are unlikely to be annotated. BRAKER uses the RNAseq libraries to select genomic regions with gene features for training and executes the training process through the GeneMark algorithm (Lomsadze et al. 2005). The new BRAKER pipeline is preferred for the annotation of eukaryotic genomes, given the lower number of annotation steps that do not require training the gene models and assembling the reads from the RNAseq libraries.

With many studies generating RNAseq reads to accompany the whole genome sequencing data, genome annotations can be improved in completeness by using multiple RNAseq libraries. As demonstrated in this study, the BRAKER pipeline is one of the first to combine various evidence from RNAseq reads and has implemented several steps to pre-process the supporting evidence (Banerjee et al. 2021; Venturini et al. 2018). I expect the use of software that processes and prioritizes transcriptomic datasets to gain increasing importance in genome annotations. Additionally, using long-read RNAseq from Pacific Biosciences or Oxford Nanopore Technologies can improve the accuracy of the genome annotation by providing a more precise mapping of the reads than short reads, as already demonstrated in the long-read annotation (LoReAn) pipeline (Cook et al. 2019). The development of bioinformatics

software that combines supporting evidence from different RNAseq libraries will improve the current gene annotation pipelines.

5.6 Inferring function of hypothetical proteins

Inference of protein function is fundamental to understanding the biological role of annotated genes. By sequencing genomes of non-model organisms, it is now possible to annotate a plethora of putative proteins (Heck and Neely 2020). However, the proteins generated by genome annotation often remain hypothetical, as in the genome annotation presented in this thesis work, where ~40% of the genes lack assigned function. Considering the growing gap between the generated sequences and their inferred function, developing additional methods to infer protein function beyond protein sequence identity or similarity is necessary. A promising method that can allow functional annotation of annotated genes is the use of protein structural information. The protein structure is the arrangement in the three-dimensional space of the protein sequence (Alberts 2008), and it is determined by the protein bonds that form along the protein chain. It is well established that protein structure is more conserved than protein sequence (Illergård, Ardell, and Elofsson 2009), and protein structures provide greater insights into protein function than just the sequence. Protein structures are experimentally solved for only ~0.2% of the known protein sequences (Kc 2017). Several computational approaches have been developed in the past years to bridge this gap and compute the protein structure directly from the protein sequence. AlphaFold (Jumper et al. 2021), an artificial intelligence system based on deep neural networks, demonstrated a high accuracy in protein structures prediction, which is an unprecedented achievement. The AlphaFold system, already tested on uncharacterized proteins from the human genome annotation (Tunyasuvunakool et al. 2021), provides an additional tool to infer the protein function and can be applied to annotated proteins as in the case of non-model organisms.

Protein functional annotation is often accompanied by additional high-throughput experiments that provide further insight into the gene function (Zhou et al. 2019; Schnoes et al. 2013). I consider two

experimental methods to be able to improve the annotation of hypothetical proteins. Differential gene expression quantifies transcriptome differences between two conditions (control vs. treatment) and is often used to understand biological processes (Arick and Hsu 2018). Gene co-expression analysis can infer functional information for genes taking part in the same biological process as well-characterized genes (Emamjomeh et al. 2017). Even though the correlation of gene expression does not directly imply causation, it suggests that co-expressed genes are enriched for genes from the same pathway and may cover the same or similar function as known proteins. Gene function can be additionally tested by gene silencing technologies or RNA interference (RNAi), which can specifically silence gene expression; it is possible to identify the gene function by observing a phenotype that gives direct evidence of gene function. RNAi has been successfully used in functional studies on plants (Abdurakhmonov et al. 2016; Matchett-Oates, Spangenberg, and Cogan 2021) and insects (Kyre, Bentz, and Rieske 2020; Pampolini et al. 2020) and can also be applied to studies improving functional gene annotation.

5.7 Impact of the annotated genomes and future directions

To summarize, the genome assemblies and annotations that I performed during my Ph.D. projects provided new and detailed biological information about the genome evolution of four conifer species (gen. *Picea*) and the genome size evolution of *Pissodes strobi*, and a secondary metabolite pathway in *Cannabis sativa*. The annotations from the three projects are expected to be used as a reference for future studies in the forestry and plant genomics fields. In particular, the gene annotations of four spruce genomes are the most comprehensive dataset of *Picea* genes so far and represent a significant advancement for studying the genus.

I believe that the generated annotation and comparative analysis will serve to design new studies in the field and will be applied as a foundation for population studies, particularly for the gen. *Picea* and *P. strobi.* As described in section 2.5.4 for gen. *Picea*, the spruce genomes, and annotations represent a template for studying the intra-species variability among individuals from populations. The population

distributions of the spruces are widespread across the North American continent. The annotations of spruces, generated with the same computational approach, provide, to date, the largest gene reference for short nuclear variants (SNV) genotyping that can be used to explore the intra-species variability. Additionally, the comparative analysis of the gene annotations highlighted a set of proteins and functional domains specific to each taxon and may have an active role in the local adaptation. By investigating the population intra-variability of the listed proteins, it will be possible to gain further insight into the molecular processes driving local adaptation.

As described in section 3.5.6, the study of *P. strobi* populations can provide us with a better understanding of the evolution of the insect pest genome. As described for spruce, the genotyping of individuals from distinct *P. strobi* populations can highlight genomic regions that differentiate each population. The large distribution of *P. strobi* across different areas of North America and its adaptation to other host trees raises several questions about the molecular mechanisms of local adaptation and host pathogenicity that remain broadly unresolved.

Bibliography

Aardema, Matthew, and Rob DeSalle. 2021. “Can Public Online Databases Serve as a Source of Phenotypic Information for Cannabis Genetic Association Studies?” *PlosOne* 16 (2): e0247607.

Abdurakhmonov, Ibrokhim Y., Mirzakamol S. Ayubov, Khurshida A. Ubaydullaeva, Zabardast T. Buriev, Shukhrat E. Shermatov, Haydarali S. Ruziboev, Umid M. Shapulatov, et al. 2016. “RNA Interference for Functional Genomics and Improvement of Cotton (Gossypium Sp.).” *Frontiers in Plant Science* 7: 202.

Abrahams, Sharon, Elizabeth Lee, Amanda R. Walker, Gregory J. Tanner, Philip J. Larkin, and Anthony R. Ashton. 2003. “The Arabidopsis TDS4 Gene Encodes Leucoanthocyanidin Dioxygenase (LDOX) and Is Essential for Proanthocyanidin Synthesis and Vacuole Development.” *The Plant Journal: For Cell and Molecular Biology* 35 (5): 624–636.

Abu-Sawwa, Renad, Brielle Scutt, and Yong Park. 2020. “Emerging Use of Epidiolex (Cannabidiol) in Epilepsy.” *The Journal of Pediatric Pharmacology and Therapeutics: JPPT: The Official Journal of PPAG* 25 (6): 485–499.

Adams, Keith L., and Jonathan F. Wendel. 2005. “Polyploidy and Genome Evolution in Plants.” *Current Opinion in Plant Biology* 8 (2): 135–141.

Adams, Robert P. 2019. “Inheritance of Chloroplasts and Mitochondria in Conifers: A Review of Paternal, Maternal, Leakage and Facultative Inheritance.” *Phytologia* 101 (2): 134–138.

Ahuja, M. Raj, and David B. Neale. 2005. “Evolution of Genome Size in Conifers.” *Silvae Genetica* 54 (1–6): 126–37.

Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet et al. 2016. “The Ensembl Gene Annotation System.” *Database* 2016: baw093.

Alberts, Bruce. 2008. “Molecular Biology of the Cell 5 Edition.”

Alfaro, Rene I. 1994. “White Pine Weevil in British Columbia: Biology and Damage.” *FRDA Report*.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 1–16.

Apriyanto, Ardha, and Van Basten Tambunan. 2021. “Draft Genome Sequence, Annotation, and SSR Mining Data of Elaeidobius Kamerunicus Faust., an Essential Oil Palm Pollinating Weevil.” *Data in Brief* 34: 106745.

Arick, Mark, 2nd, and Chuan-Yu Hsu. 2018. “Differential Gene Expression Analysis of Plants.” *Methods in Molecular Biology* 1783: 279–298.

Asaf, Sajjad, Abdul Latif Khan, Muhammad Aaqil Khan, Raheem Shahzad, Lubna, Sang Mo Kang, Ahmed Al-Harrasi, Ahmed Al-Rawahi, and In-Jung Lee. 2018. “Complete Chloroplast Genome Sequence and Comparative Analysis of Loblolly Pine (Pinus Taeda L.) with Related Species.” *PloS One* 13 (3): e0192966.

Badouin, Hélène, Jérôme Gouzy, Christopher J. Grassa, Florent Murat, S. Evan Staton, Ludovic Cottret, Christine Lelandais-Brière, et al. 2017. “The Sunflower Genome Provides Insights into Oil Metabolism, Flowering and Asterid Evolution.” *Nature* 546 (7656): 148–152.

Bakel, Harm van, Jake M. Stout, Atina G. Cote, Carling M. Tallon, Andrew G. Sharpe, Timothy R. Hughes, and Jonathan E. Page. 2011. “The Draft Genome and Transcriptome of Cannabis Sativa.” *Genome Biology* 12 (10): R102.

Banerjee, Sagnik, Priyanka Bhandary, Margaret Woodhouse, Taner Z. Sen, Roger P. Wise, and Carson M. Andorf. 2021. “FINDER: An Automated Software Package to Annotate Eukaryotic Genes from RNA-Seq Data and Associated Protein Sequences.” *BMC Bioinformatics* 22 (1): 1–26.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. “Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes.” *Mobile DNA* 6 (1): 1–6.

Bayer, Philipp E., David Edwards, and Jacqueline Batley. 2018. “Bias in Resistance Gene Prediction due to Repeat Masking.” *Nature Plants* 4 (10): 762–765.

Bayer, Philipp E., Agnieszka A. Golicz, Soodeh Tirnaz, Chon-Kit Kenneth Chan, David Edwards, and Jacqueline Batley. 2019. “Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome.” *Plant Biotechnology Journal* 17 (4): 789–800.

Belaghzal, Houda, Job Dekker, and Johan H. Gibcus. 2017. “Hi-C 2.0: An Optimized Hi-C Procedure for High-Resolution Genome-Wide Mapping of Chromosome Conformation.” *Methods* 123 (July): 56–65.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bennetzen, Jeffrey L., and Hao Wang. 2014. “The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes.” *Annual Review of Plant Biology* 65: 505–530.

Berli, Federico J., Daniela Moreno, Patricia Piccoli, Leandro Hespanhol-Viana, M. Fernanda Silva, Ricardo Bressan-Smith, J. Bruno Cavagnaro, and Rubén Bottini. 2010. “Abscisic Acid Is Involved in the Response of Grape (Vitis Vinifera L.) Cv. Malbec Leaf Tissues to Ultraviolet-B Radiation by Enhancing Ultraviolet-Absorbing Compounds, Antioxidant Enzymes and Membrane Sterols.” *Plant, Cell & Environment* 33 (1): 1–10.

Binns, David, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'donovan, and Rolf Apweiler. 2009. “QuickGO: a web-based tool for Gene Ontology searching.” *Bioinformatics* 25(22): 3045–3046.

Birol, Inanc, Hamid Mohamadi, and Justin Chu. 2018. “ntPack: A Software Package for Big Data in Genomics.” In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, 41–50.

Birol, Inanc, Anthony Raymond, Shaun D. Jackman, Stephen Pleasance, Robin Coope, Greg A. Taylor, Macaire Man Saint Yuen, et al. 2013. “Assembling the 20 Gb White Spruce (Picea Glauca) Genome from Whole-Genome Shotgun Sequencing Data.” *Bioinformatics* 29 (12): 1492–1497.

Blommaert, Julie. 2020. “Genome Size Evolution: Towards New Model Systems for Old Questions.” *Proceedings. Biological Sciences / The Royal Society* 287 (1933): 20201441.

Bouillé, Marie, Sauphie Senneville, and Jean Bousquet. 2011. “Discordant mtDNA and cpDNA Phylogenies Indicate Geographic Speciation and Reticulation as Driving Factors for the Diversification of the Genus Picea.” *Tree Genetics & Genomes* 7 (3): 469–484.

Bracewell, Ryan R., Barbara J. Bentz, Brian T. Sullivan, and Jeffrey M. Good. 2017. “Rapid Neo-Sex Chromosome Evolution and Incipient Speciation in a Major Forest Pest.” *Nature Communications* 8 (1): 1–14.

Braich, Shivraj, Rebecca C. Baillie, German C. Spangenberg, and Noel O. I. Cogan. 2020. “A New and Improved Genome Sequence of *Cannabis Sativa*.” *Gigabyte* 2020: 1–13.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–527.

Brown, Patrick O., Bruce Bowerman, Harold E. Varmus, and J. Michael Bishop. 1987. “Correct integration of retroviral DNA in vitro.” *Cell* 49 (3): 347–356.

Brůna, Tomáš, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. 2021. “BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP and AUGUSTUS Supported by a Protein Database.” *NAR Genomics and Bioinformatics* 3 (1): lqaa108.

Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. “Chromosome-Scale Scaffolding of de Novo Genome Assemblies Based on Chromatin Interactions.” *Nature Biotechnology* 31 (12): 1119–1125.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 1–9.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell. 2014. “Genome Annotation and Curation Using MAKER and MAKER-P.” *Current Protocols in Bioinformatics* 48 (1): 4.11.1–39.

Campbell, Michael S., and Mark Yandell. 2015. “An Introduction to Genome Annotation.” *Current Protocols in Bioinformatics* 52: 4.1.1–4.1.17.

Canada’s Genomics Enterprise CanSeq. “CanSeq150.” Accessed November 7, 2021. https://www.cgen.ca/canseq150-overview.

Carretero-Paulet, Lorenzo, and Yves Van de Peer. 2020. “The Evolutionary Conundrum of Whole-Genome Duplication.” *American Journal of Botany* 107 (8): 1101–1105.

Casola, Claudio, and Tomasz E. Koralewski. 2018. “Pinaceae Show Elevated Rates of Gene Turnover That Are Robust to Incomplete Gene Annotation.” *The Plant Journal: For Cell and Molecular Biology*, 95 (5): 862–876.

Cheng, Shifeng, Michael Melkonian, Stephen A. Smith, Samuel Brockington, John M. Archibald, Pierre-Marc Delaux, Fay-Wei Li, et al. 2018. “10KP: A Phylodiverse Genome Sequencing Plan.” *GigaScience* 7 (3): 1–9.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. “Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.” *Nature Methods* 13 (12): 1050–1054.

Chu, Justin. 2017. *Circos Assembly Consistency (Jupiter) Plot*. Available: github.com/JustinChu/JupiterPlot.

Chu, Justin, Sara Sadeghi, Anthony Raymond, Shaun D. Jackman, Ka Ming Nip, Richard Mar, Hamid Mohamadi, Yaron S. Butterfield, A. Gordon Robertson, and Inanç Birol. 2014. “BioBloom Tools: Fast, Accurate and Memory-Efficient Host Species Sequence Screening Using Bloom Filters.” *Bioinformatics* 30 (23): 3402–3404.

Clark, M. N., and B. A. Bohm. 1979. “Flavonoid Variation in Cannabis L.” *Botanical Journal of the Linnean Society. Linnean Society of London* 79 (3): 249–257.

Clark, Sangaalofa T., and Wynand S. Verwoerd. 2011. “A Systems Approach to Identifying Correlated Gene Targets for the Loss of Colour Pigmentation in Plants.” *BMC Bioinformatics* 12: 1–12.

Cognato, Anthony I., Sarah M. Smith, and Bjarte H. Jordal. 2021. “Patterns of Host Tree Use within a Lineage of Saproxlic Snout-Less Weevils (Coleoptera: Curculionidae: Scolytinae: Scolytini).” *Molecular Phylogenetics and Evolution* 159: 107107.

Collins, John. 2020. “A Brief History of Cannabis and the Drug Conventions.” *The American Journal of International Law* 114: 279–284.

Computational Pan-Genomics Consortium. 2018. “Computational pan-genomics: status, promises and challenges.” *Briefings in bioinformatics* 19 (1): 118–135.

Cook, David E., Jose Espejo Valle-Inclan, Alice Pajoro, Hanna Rovenich, Bart P. H. J. Thomma, and Luigi Faino. 2019. “Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing.” *Plant Physiology* 179 (1): 38–54.

Coombe, Lauren, Vladimir Nikolić, Justin Chu, Inanc Birol, and René L. Warren. 2020. “ntJoin: Fast and Lightweight Assembly-Guided Scaffolding Using Minimizer Graphs.” *Bioinformatics* 36 (12): 3885–3887.

Coombe, Lauren, René L. Warren, Shaun D. Jackman, Chen Yang, Benjamin P. Vandervalk, Richard A. Moore, Stephen Pleasance, et al. 2016. “Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics’ GemCode Sequencing Data.” *PloS One* 11 (9): e0163059.

Coombe, Lauren, Jessica Zhang, Benjamin P. Vandervalk, Justin Chu, Shaun D. Jackman, Inanc Birol, and René L. Warren. 2018. “ARKS: Chromosome-Scale Scaffolding of Human Genome Drafts with Linked Read Kmers.” *BMC Bioinformatics* 19 (1): 1–10.

Cox, Chelsea. 2018. “The Canadian Cannabis Act Legalizes and Regulates Recreational Cannabis Use in 2018.” *Health Policy* 122 (3): 205–209.

Craig, Nancy L. 2020. "Mobile DNA III."

Cronn, Richard, Aaron Liston, Matthew Parks, David S. Gernandt, Rongkun Shen, and Todd Mockler. 2008. “Multiplex Sequencing of Plant Chloroplast Genomes Using Solexa Sequencing-by-Synthesis Technology.” *Nucleic Acids Research* 36 (19): e122.

Crow, Karen D., and Günter P. Wagner. 2005. “What is the Role of Genome Duplication in the Evolution of Complexity and Diversity?” *Molecular biology and evolution* 23 (5): 887–892.

De La Torre, Amanda R., Yao-Cheng Lin, Yves Van de Peer, and Pär K. Ingvarsson. 2015. “Genome-Wide Analysis Reveals Diverged Patterns of Codon Bias, Gene Expression, and Rates of Sequence Evolution in Picea Gene Families.” *Genome Biology and Evolution* 7 (4): 1002–1015.

Demuth, J.P. and Hahn, M.W., 2009. “The life and death of gene families.” *Bioessays*, 31(1), 29–39.

Denton, James F., Jose Lugo-Martinez, Abraham E. Tucker, Daniel R. Schrider, Wesley C. Warren, and Matthew W. Hahn. 2014. “Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies.” *PLoS Computational Biology* 10 (12): e1003998.

Depardieu, Claire, Sébastien Gérardi, Simon Nadeau, Geneviève J. Parent, John Mackay, Patrick Lenz, Manuel Lamothe, Martin P. Girardin, Jean Bousquet, and Nathalie Isabel. 2021. “Connecting Tree-Ring Phenotypes, Genetic Associations and Transcriptomics to Decipher the Genomic Architecture of Drought Adaptation in a Widespread Conifer.” *Molecular Ecology* 30 (16): 3898–3917.

Divashuk, Mikhail G., Oleg S. Alexandrov, Olga V. Razumova, Ilya V. Kirov, and Gennady I. Karlov. 2014. “Molecular Cytogenetic Characterization of the Dioecious Cannabis Sativa with an XY Chromosome Sex Determination System.” *PLoS ONE* 9 (1): e85118.

Dolatabadian, Aria, Philipp E. Bayer, Soodeh Tirnaz, Bhavna Hurgobin, David Edwards, and Jacqueline Batley. 2020. “Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation.” *Plant biotechnology journal* 18 (4): 969–982.

Dolezel, J., J. Bartos, H. Voglmayr, and J. Greilhuber. 2003. “Nuclear DNA Content and Genome Size of Trout and Human.” *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* 51: 127–128.

Duda, T. F., Jr, and S. R. Palumbi. 1999. “Molecular Genetics of Ecological Diversification: Duplication and Rapid Evolution of Toxin Genes of the Venomous Gastropod Conus.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (12): 6820–6823.

Ebata, T. 1991. “Summary Report of Two Spruce Weevil Surveys in Twelve Plantations in the Kitimat Valley. BC Ministry of Forests.” *Victoria Internal Report, PM-PB-69*.

Eberhardt, Ruth Y., Daniel H. Haft, Marco Punta, Maria Martin, Claire O’Donovan, and Alex Bateman. 2012. “AntiFam: A Tool to Help Identify Spurious ORFs in Protein Annotation.” *Database: The Journal of Biological Databases and Curation* 2012 (March): bas003.

Ejigu, Girum Fitihamlak, and Jaehee Jung. 2020. “Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing.” *Biology* 9 (9): 295.

Ekman, Diana, Asa K. Björklund, and Arne Elofsson. 2007. “Quantification of the Elevated Rate of Domain Rearrangements in Metazoa.” *Journal of Molecular Biology* 372 (5): 1337–1348.

Ellegren, Hans. 2014. “Genome Sequencing and Population Genomics in Non-Model Organisms.” *Trends in Ecology & Evolution* 29 (1): 51–63.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. “LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons.” *BMC Bioinformatics* 9 (1): 1–14.

Elsohly, Mahmoud A., and Desmond Slade. 2005. “Chemical Constituents of Marijuana: The Complex Mixture of Natural Cannabinoids.” *Life Sciences* 78 (5): 539–548.

Emamjomeh, Abbasali, Elham Saboori Robat, Javad Zahiri, Mahmood Solouki, and Pegah Khosravi. 2017. “Gene Co-Expression Network Reconstruction: A Review on Computational Methods for Inferring Functional Information from Plant-Based Expression Data.” *Plant Biotechnology Reports* 11 (2): 71–86.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 1–14.

Exposito-Alonso, Moises, Hajk-Georg Drost, Hernán A. Burbano, and Detlef Weigel. 2020. “The Earth BioGenome Project: Opportunities and Challenges for Plant Genomics and Conservation.” *The Plant Journal: For Cell and Molecular Biology* 102 (2): 222–229.

Falcone Ferreyra, María L., Sebastián P. Rius, and Paula Casati. 2012. “Flavonoids: Biosynthesis, Biological Functions, and Biotechnological Applications.” *Frontiers in Plant Science* 3: 222.

Fang, Hai, and Julian Gough. 2013. “DcGO: Database of Domain-Centric Ontologies on Functions, Phenotypes, Diseases and More.” *Nucleic Acids Research* 41 (D1): D536–544.

Farrar, John Laird. 1995. “Trees in Canada.”

Felsenstein, Joseph. 1985. “Confidence Limits on Phylogenies: an Approach Using the Bootstrap.” *Evolution; International Journal of Organic Evolution* 39 (4): 783–791.

Fischer, Benedikt, Cayley Russell, and Neil Boyd. 2020. “A Century of Cannabis Control in Canada: A Brief Overview of History, Context and Policy Frameworks from Prohibition to Legalization.” *Legalizing Cannabis*, 89–115.

Florea, Liliana, Alexander Souvorov, Theodore S. Kalbfleisch, and Steven L. Salzberg. 2011. “Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies.” *PloS One* 6 (6): e21400.

Flores-Sanchez, Isvett Josefina, and Robert Verpoorte. 2008. “PKS Activities and Biosynthesis of Cannabinoids and Flavonoids in Cannabis Sativa L. Plants.” *Plant & Cell Physiology* 49 (12): 1767–1782.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” *Proceedings of the National Academy of Sciences* 117 (17): 9451–9457.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–1545.

Friedberg, Iddo. 2006. “Automated Protein Function Prediction—the Genomic Challenge.” *Briefings in Bioinformatics* 7 (3): 225–242.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–3152.

Gao, Shan, Baishi Wang, Shanshan Xie, Xiaoyu Xu, Jin Zhang, Li Pei, Yongyi Yu, Weifei Yang, and Ying Zhang. 2020. “A High-Quality Reference Genome of Wild Cannabis Sativa.” *Horticulture Research* 7 (1): 73.

Gara, R. I., and J. O. Wood. 1989. “Termination of Reproductive Diapause in the Sitka Spruce Weevil, Pissodes Strobi (Peck) (Col., Curculionidae) in Western Washington1.” *Journal of Applied Entomology, Zeitschrift Fur Angewandte Entomologie* 108 (1-5): 156–63.

Gatica-Arias, A., M. A. Farag, M. Stanke, J. Matoušek, L. Wessjohann, and G. Weber. 2012. “Flavonoid Production in Transgenic Hop (Humulus Lupulus L.) Altered by PAP1/MYB75 from Arabidopsis Thaliana L.” *Plant Cell Reports* 31 (1): 111–119.

Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, et al. 2012. “ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery.” *Nucleic Acids Research* 40 (Database issue): D1100–1107.

Ghurye, Jay, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. 2017. “Scaffolding of Long Read Assemblies Using Long Range Contact Information.” *BMC Genomics* 18 (1): 1–11.

Gilbert, Don. 2013. “Gene-Omes Built from mRNA-Seq Not Genome DNA.” Available: https://scholarworks.iu.edu/dspace/bitstream/handle/2022/22617/Evigene_RNA2013.pdf?sequence=1&isAllowed=y.

Girardin, Martin P., Nathalie Isabel, Xiao Jing Guo, Manuel Lamothe, Isabelle Duchesne, and Patrick Lenz. 2021. “Annual Aboveground Carbon Uptake Enhancements from Assisted Gene Flow in Boreal Black Spruce Forests Are Not Long-Lasting.” *Nature Communications* 12 (1): 1169.

Girgis, Hani Z. 2015. “Red: An Intelligent, Rapid, Accurate Tool for Detecting Repeats de-Novo on the Genomic Scale.” *BMC Bioinformatics* 16 (1): 1–19.

Grassa, Christopher J., George D. Weiblen, Jonathan P. Wenger, Clemon Dabney, Shane G. Poplawski, S. Timothy Motley, Todd P. Michael, and C. J. Schwartz. 2021. “A New Cannabis Genome Assembly Associates Elevated Cannabidiol (CBD) with Hemp Introgressed into Marijuana.” *The New Phytologist* 230 (4): 1665–1679.

Gregory, T. Ryan. 2005. “Synergy between Sequence and Size in Large-Scale Genomics.” *Nature Reviews. Genetics* 6 (9): 699–708.

Griesemer, Marc, Jeffrey A. Kimbrel, Carol E. Zhou, Ali Navid, and Patrik D'haeseleer. 2018. "Combining Multiple Functional Annotation Tools Increases Coverage of Metabolic Annotation." *BMC Genomics* 19 (1): 1–11.

Garrido-Ramos, Manuel A. 2017. "Satellite DNA: an evolving topic." *Genes* 8 (9): 230.

Grover, Corrinne E., and Jonathan F. Wendel. 2010. "Recent Insights into Mechanisms of Genome Size Change in Plants." *Journal of Botany* 2010.

Guan, Dengfeng, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. 2020. "Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies." *Bioinformatics* 36 (9): 2896–2898.

Gutierrez, Enrique, Ana García-Villaraco Velasco, Jose Antonio Lucas, F. Javier Gutierrez-Mañero, and Beatriz Ramos-Solano. 2017. "The Flavonol-Anthocyanin Pathway in Blackberry and Arabidopsis: State of the Art." *Flavonoids—from Biosynthesis to Human Health, InTech*, 129–150.

Hamilton, Jill A., Amanda R. De la Torre, and Sally N. Aitken. 2015. "Fine-Scale Environmental Variation Contributes to Introgression in a Three-Species Spruce Hybrid Complex." *Tree Genetics & Genomes* 11 (1): 1–14.

Han, Guoliang, Chaoxia Lu, Jianrong Guo, Ziqi Qiao, Na Sui, Nianwei Qiu, and Baoshan Wang. 2020. "C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants." *Frontiers in Plant Science* 11: 115.

Han, Mira V., Gregg W. C. Thomas, Jose Lugo-Martinez, and Matthew W. Hahn. 2013. "Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3." *Molecular Biology and Evolution* 30 (8): 1987–1997.

Hardwick, Simon A., Anoushka Joglekar, Paul Flicek, Adam Frankish, and Hagen U. Tilgner. 2009. "Getting the Entire Message: Progress in Isoform Sequencing." *Frontiers in genetics* 709.

Harrop, Thomas W. R., Marissa F. Le Lec, Ruy Jauregui, Shannon E. Taylor, Sarah N. Inwood, Tracey van Stijn, Hannah Henry, et al. 2020. “Genetic Diversity in Invasive Populations of Argentine Stem Weevil Associated with Adaptation to Biocontrol.” *Insects* 11 (7): 441.

Hart, Alexander J., Samuel Ginzburg, Muyang Sam Xu, Cera R. Fisher, Nasim Rahmatpour, Jeffry B. Mitton, Robin Paul, and Jill L. Wegrzyn. 2020. “EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes.” *Molecular Ecology Resources* 20 (2): 591–604.

Hartmann, Laura, Lorenzo Pedrotti, Christoph Weiste, Agnes Fekete, Jasper Schierstaedt, Jasmin Göttler, Stefan Kempa, et al. 2015. “Crosstalk between Two bZIP Signaling Pathways Orchestrates Salt-Induced Metabolic Reprogramming in Arabidopsis Roots.” *The Plant Cell* 27 (8): 2244–2260.

Haselhorst, Monia S. H., and C. Alex Buerkle. 2013. “Population Genetic Structure of Picea Engelmannii, P. Glauca and Their Previously Unrecognized Hybrids in the Central Rocky Mountains.” *Tree Genetics & Genomes* 9 (3): 669–681.

Hazzouri, Khaled Michel, Naganeeswaran Sudalaimuthuasari, Biduth Kundu, David Nelson, Mohammad Ali Al-Deeb, Alain Le Mansour, Johnston J. Spencer, Claude Desplan, and Khaled M. A. Amiri. 2020. “The Genome of Pest Rhynchophorus Ferrugineus Reveals Gene Families Important at the Plant-Beetle Interface.” *Communications Biology* 3 (1): 1–14.

Heck, Michelle, and Benjamin A. Neely. 2020. “Proteomics in Non-Model Organisms: A New Analytical Frontier.” *Journal of Proteome Research* 19 (9): 3595–3606.

He, Dan, Dawei Zhang, Ting Li, Lili Liu, Dinggang Zhou, Jinfeng Wu, Lei Kang, Zhongsong Liu, and Mingli Yan. 2021 “Whole Genome Identification and Comparative Expression Analysis of Anthocyanin Biosynthetic Genes in Brassica Napus.” *Frontiers in Genetics* 2114.

Heddi, A., A. M. Grenier, C. Khatchadourian, H. Charles, and P. Nardon. 1999. “Four Intracellular Genomes Direct Weevil Biology: Nuclear, Mitochondrial, Principal Endosymbiont, and Wolbachia.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (12): 6814–6819.

Heitkam, Tony, Luise Schulte, Beatrice Weber, Susan Liedtke, Sarah Breitenbach, Anja Kögler, Kristin Morgenstern, et al. 2021. “Comparative Repeat Profiling of Two Closely Related Conifers (Larix Decidua and Larix Kaempferi) Reveals High Genome Similarity With Only Few Fast-Evolving Satellite DNAs.” *Frontiers in Genetics* 12: 683668.

Hernández-Salmerón, Julie E., and Gabriel Moreno-Hagelsieb. 2020. “Progress in Quickly Finding Orthologs as Reciprocal Best Hits: Comparing Blast, Last, Diamond and MMseqs2.” *BMC Genomics* 21 (1): 1–9.

He, Tianhua, Juli G. Pausas, Claire M. Belcher, Dylan W. Schwilk, and Byron B. Lamont. 2012. “Fire-Adapted Traits of Pinus Arose in the Fiery Cretaceous.” *The New Phytologist* 194 (3): 751–759.

Hill, David P., Barry Smith, Monica S. McAndrews-Hill, and Judith A. Blake. 2008. “Gene Ontology Annotations: What They Mean and Where They Come from.” *BMC Bioinformatics* 9 (5): 1–9.

Hillig, Karl W. 2005. “Genetic Evidence for Speciation in Cannabis (Cannabaceae).” *Genetic Resources and Crop Evolution* 52 (2): 161–180.

Holliday, Jason A., Kermit Ritland, and Sally N. Aitken. 2010. “Widespread, Ecologically Relevant Genetic Markers Developed from Association Mapping of Climate-Related Traits in Sitka Spruce (Picea Sitchensis).” *The New Phytologist* 188 (2): 501–514.

Holt, Carson, and Mark Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12 (1): 1–14.

Hornoy, Benjamin, Nathalie Pavy, Sébastien Gérardi, Jean Beaulieu, and Jean Bousquet. 2015. “Genetic Adaptation to Climate in White Spruce Involves Small to Moderate Allele Frequency Shifts in Functionally Diverse Genes.” *Genome Biology and Evolution* 7 (12): 3269–3285.

Hou, Yubo, and Senjie Lin. 2009. “Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes.” *PloS One* 4 (9): e6978.

Hu, Zhiqiang, Chaochun Wei, and Zhikang Li. 2020. “Computational strategies for eukaryotic pangenome analyses.” *The Pangenome*: 293–307.

Hubley R., Smit A. “RepeatModeller v1.” Accessed November 21, 2021. Available: www.repeatmasker.org/RepeatModeler.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–314.

Hurgobin, Bhavna, Muluneh Tamiru-Oli, Matthew T. Welling, Monika S. Doblin, Antony Bacic, James Whelan, and Mathew G. Lewsey. 2021. “Recent Advances in Cannabis Sativa Genomics Research.” *The New Phytologist* 230 (1): 73–89.

i5K Consortium. 2013. “The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.” *The Journal of Heredity* 104 (5): 595–600.

Ibarra-Laclette, Enrique, Eric Lyons, Gustavo Hernández-Guzmán, Claudia Anahí Pérez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, et al. 2013. “Architecture and Evolution of a Minute Plant Genome.” *Nature* 498 (7452): 94–98.

Illergård, Kristoffer, David H. Ardell, and Arne Elofsson. 2009. “Structure Is Three to Ten Times More Conserved than Sequence–a Study of Structural Response in Protein Cores.” *Proteins* 77 (3): 499–508.

Izzo, Luana, Luigi Castaldo, Alfonso Narváez, Giulia Graziani, Anna Gaspari, Yelko Rodríguez-Carrasco, and Alberto Ritieni. 2020. “Analysis of Phenolic Compounds in Commercial Cannabis Sativa L. Inflorescences Using UHPLC-Q-Orbitrap HRMS.” *Molecules* 25 (3): 631.

Jackman, Shaun D., Lauren Coombe, Justin Chu, Rene L. Warren, Benjamin P. Vandervalk, Sarah Yeo, Zhuyi Xue, et al. 2018. “Tigmint: Correcting Assembly Errors Using Linked Reads from Large Molecules.” *BMC Bioinformatics* 19 (1): 1–10.

Jackman, Shaun D., Lauren Coombe, René L. Warren, Heather Kirk, Eva Trinh, Tina MacLeod, Stephen Pleasance et al. 2020. “Complete mitochondrial genome of a gymnosperm, Sitka spruce (Picea

sitchensis), indicates a complex physical structure." *Genome biology and evolution* 12 (7): 1174–1179.

Jackman, Shaun D., Benjamin P. Vandervalk, Hamid Mohamadi, Justin Chu, Sarah Yeo, S. Austin Hammond, Golnaz Jahesh, et al. 2017. "ABySS 2.0: Resource-Efficient Assembly of Large Genomes Using a Bloom Filter." *Genome Research* 27 (5): 768–777.

Jackman, Shaun D., René L. Warren, Ewan A. Gibb, Benjamin P. Vandervalk, Hamid Mohamadi, Justin Chu, Anthony Raymond, et al. 2016. "Organellar Genomes of White Spruce (Picea Glauca): Assembly and Annotation." *Genome Biology and Evolution* 8 (1): 29–41.

Jauhal, April A., and Richard D. Newcomb. 2021. "Assessing Genome Assembly Quality prior to Downstream Analysis: N50 versus BUSCO." *Molecular Ecology Resources* 21 (5): 1416–1421.

Jrgaph. 2021. "Configurable diagramming/whiteboarding visualization application". Available: github.com/jgraph/drawio

Jin, Dan, Kaiping Dai, Zhen Xie, and Jie Chen. 2020. "Secondary Metabolites Profiled in Cannabis Inflorescences, Leaves, Stem Barks, and Roots for Medicinal Purposes." *Scientific Reports* 10 (1): 1–14.

Joazeiro, C. A., and A. M. Weissman. 2000. "RING Finger Proteins: Mediators of Ubiquitin Ligase Activity." *Cell* 102 (5): 549–552.

Johnston, J. Spencer, Angelina Bernardini, and Carl E. Hjelmen. 2019. "Genome Size Estimation and Quantitative Cytogenetics in Insects." *Methods in Molecular Biology* 1858: 15–26.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–1240.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–589.

Junier, Thomas, and Evgeny M. Zdobnov. 2010. "The Newick Utilities: High-Throughput Phylogenetic Tree Processing in the UNIX Shell." *Bioinformatics* 26 (13): 1669–1670.

Kang, Joung-Youn, Hyung-In Choi, Min-Young Im, and Soo Young Kim. 2002. "Arabidopsis Basic Leucine Zipper Proteins That Mediate Stress-Responsive Abscisic Acid Signaling." *The Plant Cell* 14 (2): 343–457.

Kapitonov, Vladimir V., and Jerzy Jurka. 2008. "A universal classification of eukaryotic transposable elements implemented in Repbase." *Nature Reviews Genetics* 9 (5): 411–412.

Katz, Lee S., Taylor Griswold, Shatavia S. Morrison, Jason A. Caravas, Shaokang Zhang, Henk C. den Bakker, Xiangyu Deng, and Heather A. Carleton. 2019. "Mashtree: a Rapid Comparison of Whole Genome Sequence Files." *Journal of Open Source Software* 4 (44): 1762.

Kazazian, Haig H., Jr. 2004. "Mobile Elements: Drivers of Genome Evolution." *Science* 303 (5664): 1626–32.

Kc, Dukka B. 2017. "Recent Advances in Sequence-Based Protein Structure Prediction." *Briefings in Bioinformatics* 18 (6): 1021–1032.

Keeling, Christopher I., Erin O. Campbell, Philip D. Batista, Victor A. Shegelski, Stephen A. L. Trevoy, Dezene P. W. Huber, Jasmine K. Janes, and Felix A. H. Sperling. 2021. "Chromosome-Level Genome Assembly Reveals Genomic Architecture of Northern Range Expansion in the Mountain Pine Beetle, Dendroctonus Ponderosae Hopkins (Coleoptera: Curculionidae)." *Molecular Ecology Resources*, October.

Keeling, Christopher I., Macaire M. S. Yuen, Nancy Y. Liao, T. Roderick Docking, Simon K. Chan, Greg A. Taylor, Diana L. Palmquist, et al. 2013. "Draft Genome of the Mountain Pine Beetle, Dendroctonus Ponderosae Hopkins, a Major Forest Pest." *Genome Biology* 14 (3): R27.

Keilwagen, Jens, Frank Hartung, Michael Paulini, Sven O. Twardziok, and Jan Grau. 2018. "Combining RNA-Seq Data and Homology-Based Gene Prediction for Plants, Animals and Fungi." *BMC Bioinformatics* 19 (1): 1–12.

Kemena, Carsten, Elias Dohmen, and Erich Bornberg-Bauer. 2019. "DOGMA: A Web Server for Proteome and Transcriptome Quality Assessment." *Nucleic Acids Research* 47 (W1): W507–510.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: a Fast Spliced Aligner with Low Memory Requirements." *Nature methods* 12 (4): 357–360.

Kokot, Marek, Maciej Dlugosz, and Sebastian Deorowicz. 2017. "KMC 3: Counting and Manipulating K-Mer Statistics." *Bioinformatics* 33 (17): 2759–2761.

Kondrashov, Fyodor A. 2012. "Gene Duplication as a Mechanism of Genomic Adaptation to a Changing Environment." *Proceedings of the Royal Society B: Biological Sciences* 279 (1749): 5048–5057.

Korf, Ian. 2004. "Gene Finding in Novel Genomes." *BMC Bioinformatics* 5 (2): 1–9.

Kristensen, David M., Yuri I. Wolf, Arcady R. Mushegian, and Eugene V. Koonin. 2011. "Computational Methods for Gene Orthology Inference." *Briefings in Bioinformatics* 12 (5): 379–391.

Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. "OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs." *Nucleic acids research* 47 (D1): D807–D811.

Kumar, Shashank, and Abhay K. Pandey. 2013. "Chemistry and Biological Activities of Flavonoids: An Overview." *The Scientific World Journal* 2013: 162750.

Kyre, Bethany R., Barbara J. Bentz, and Lynne K. Rieske. 2020. "Susceptibility of Mountain Pine Beetle (Dendroctonus Ponderosae Hopkins) to Gene Silencing through RNAi Provides Potential as a Novel Management Tool." *Forest Ecology and Management* 473: 118322.

Laffin, R. D., D. W. Langor, and F. A. H. Sperling. 2004. "Population Structure and Gene Flow in the White Pine Weevil, Pissodes Strobi (Coleoptera: Curculionidae)." *Annals of the Entomological Society of America* 97 (5): 949–956.

Landis, Jacob B., Douglas E. Soltis, Zheng Li, Hannah E. Marx, Michael S. Barker, David C. Tank, and Pamela S. Soltis. 2018. "Impact of Whole-Genome Duplication Events on Diversification Rates in Angiosperms." *American Journal of Botany* 105 (3): 348–363.

Langor, David W., and Felix A. H. Sperling. 1995. “Mitochondrial DNA Variation and Identification of Bark Weevils in the Pissodes strobi Species Group in Western Canada (Coleoptera: Curculionidae).” *The Canadian Entomologist* 127 (6): 895–911.

Langor, D. W., and F. A. Sperling. 1997. “Mitochondrial DNA Sequence Divergence in Weevils of the Pissodes Strobi Species Complex (Coleoptera:Curculionidae).” *Insect Molecular Biology* 6 (3): 255–265.

Lata, Charu, and Manoj Prasad. 2011. “Role of DREBs in Regulation of Abiotic Stress Responses in Plants.” *Journal of Experimental Botany* 62 (14): 4731–4748.

Laverty, Kaitlin U., Jake M. Stout, Mitchell J. Sullivan, Hardik Shah, Navdeep Gill, Larry Holbrook, Gintaras Deikus, et al. 2019. “A Physical and Genetic Map of Cannabis Sativa Identifies Extensive Rearrangements at the THC/CBD Acid Synthase Loci.” *Genome Research* 29 (1): 146–156.

Lee, Sung-Il, and Nam-Soo Kim. 2014. “Transposable Elements and Genome Size Variations in Plants.” *Genomics & Informatics* 12 (3): 87–97.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–3433.

Lewis, K. G., Y. A. El-Kassaby, R. I. Alfaro, and S. Barnes. 2000. “Population genetic structure of Pissodes strobi (Coleoptera: Curculionidae) in British Columbia, Canada.” *Annals of the Entomological Society of America* 93 (4): 807–818.

Lewis, Kornelia G., Kermit Ritland, Yousry A. El-Kassaby, John A. McLean, Jeffry Glaubitz, and John E. Carlson. 2001. “Randomly Amplified Polymorphic DNA Reveals Fine-Scale Genetic Structure in Pissodes Strobi (Coleoptera: Curculionidae).” *The Canadian Entomologist* 133 (2): 229–238.

Licausi, Francesco, Masaru Ohme-Takagi, and Pierdomenico Perata. 2013. “APETALA2/Ethylene Responsive Factor (AP2/ERF) Transcription Factors: Mediators of Stress Responses and Developmental Programs.” *New Phytologist* 199 (3): 639–649.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–2110.

Li, Heng. 2013. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.” *arXiv preprint arXiv:*1303.3997.

Li, Xiang, Ming-Jun Gao, Hong-Yu Pan, De-Jun Cui, and Margaret Y. Gruber. 2010. “Purple Canola: Arabidopsis PAP1 Increases Antioxidants and Phenolics in Brassica Napus Leaves.” *Journal of Agricultural and Food Chemistry* 58 (3): 1639–1645.

Li, Zheng, Anthony E. Baniaga, Emily B. Sessa, Moira Scascitelli, Sean W. Graham, Loren H. Rieseberg, and Michael S. Barker. 2015. “Early Genome Duplications in Conifers and Other Seed Plants.” *Science Advances* 1 (10): e1501084.

Lin, Diana, Lauren Coombe, Shaun D. Jackman, Kristina K. Gagalova, René L. Warren, S. Austin Hammond, Helen McDonald, et al. 2019. “Complete Chloroplast Genome Sequence of an Engelmann Spruce (Picea Engelmannii, Genotype Se404-851) from Western Canada.” *Microbiology Resource Announcements* 8 (24): e00382-19.

Lin, Diana, Lauren Coombe, Shaun D. Jackman, Kristina K. Gagalova, René L. Warren, S. Austin Hammond, Heather Kirk et al. 2019. "Complete chloroplast genome sequence of a white spruce (Picea glauca, Genotype WS77111) from Eastern Canada." *Microbiology Resource Announcements* 8 (23): e00381-19.

Lockwood, Jared D., Jelena M. Aleksić, Jiabin Zou, Jing Wang, Jianquan Liu, and Susanne S. Renner. 2013. “A New Phylogeny for the Genus Picea from Plastid, Mitochondrial, and Nuclear Sequences.” *Molecular Phylogenetics and Evolution* 69 (3): 717–727.

Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. 2014. “Integration of Mapped RNA-Seq Reads into Automatic Training of Eukaryotic Gene Finding Algorithm.” *Nucleic Acids Research* 42 (15): e119.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. “Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm.” *Nucleic Acids Research* 33 (20): 6494–6506.

Love, Michael, Simon Anders, and Wolfgang Huber. 2014. “Differential Analysis of Count Data–the DESeq2 Package.” *Genome Biology* 15 (550): 10–1186.

Lower, Sarah Sander, Michael P. McGurk, Andrew G. Clark, and Daniel A. Barbash. 2018. “Satellite DNA Evolution: Old Ideas, New Approaches.” *Current opinion in genetics & development* 49: 70–78.

Luan, Dongmei D., Malka H. Korman, John L. Jakubczak, and Thomas H. Eickbush. 1993. “Reverse Transcription of R2Bm RNA is Primed by a Nick at the Chromosomal Target Site: a Mechanism for Non-LTR Retrotransposition.” *Cell* 72 (4): 595–605.

Lynch, Michael, and John S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science*. 290 (5494): 1151–1155.

Lynch, Michael, and Allan Force. 2000. “The Probability of Duplicate Gene Preservation by Subfunctionalization.” *Genetics* 154 (1): 459–473.

Lynch, Michael, and John S. Conery. 2003. “The Origins of Genome Complexity.” *Science* 302 (5649): 1401–1404.

Marks, Rose A., Scott Hotaling, Paul B. Frandsen, and Robert VanBuren. 2021. “Lessons from 20 Years of Plant Genome Sequencing: An Unprecedented Resource in Need of More Diverse Representation.” *bioRxiv*.

Marques-Bonet, Tomas, Santhosh Girirajan, and Evan E. Eichler. 2009. “The Origins and Impact of Primate Segmental Duplications.” *Trends in Genetics: TIG* 25 (10): 443–454.

Martens, Stefan, and Axel Mithöfer. 2005. “Flavones and Flavone Synthases.” *Phytochemistry* 66 (20): 2399–2407.

Martienssen, Robert. 2008. “Great Leap Forward? Transposable Elements, Small Interfering RNA and Adaptive Lamarckian Evolution.” *The New Phytologist* 2008: 570–572.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

Matchett-Oates, Lennon, German C. Spangenberg, and Noel O. I. Cogan. 2021. “Manipulation of Cannabinoid Biosynthesis via Transient RNAi Expression.” *Frontiers in Plant Science* 12: 773474.

Mattick, John S., and Michael J. Gagen. 2001. “The Evolution of Controlled Multitasked Gene Networks: the Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms.” *Molecular biology and evolution* 18 (9): 1611–1630.

Mayer, M. P., and B. Bukau. 2005. “Hsp70 Chaperones: Cellular Functions and Molecular Mechanism.” *Cellular and Molecular Life Sciences* 62 (6): 670–684.

McKenna, Duane D., Andrea S. Sequeira, Adriana E. Marvaldi, and Brian D. Farrell. 2009. “Temporal Lags and Overlap in the Diversification of Weevils and Flowering Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (17): 7083–7088.

McPartland, John M., William Hegman, and Tengwen Long. 2019. “Cannabis in Asia: Its Center of Origin and Early Cultivation, Based on a Synthesis of Subfossil Pollen and Archaeobotanical Studies.” *Vegetation History and Archaeobotany* 28 (6): 691–702.

Mehrotra, Shweta, and Vinod Goyal. 2014. “Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function.” *Genomics, Proteomics & Bioinformatics* 12 (4): 164–171.

Mehrtens, Frank, Harald Kranz, Pawel Bednarek, and Bernd Weisshaar. 2005. “The Arabidopsis Transcription Factor MYB12 Is a Flavonol-Specific Regulator of Phenylpropanoid Biosynthesis.” *Plant Physiology* 138 (2): 1083–1096.

Meijer, E. P. M. de, and L. J. M. van Soest. 1992. “The CPRO Cannabis Germplasm Collection.” *Euphytica/ Netherlands Journal of Plant Breeding* 62 (3): 201–211.

Meleshko, Olena, Michael D. Martin, Thorfinn Sand Korneliussen, Christian Schröck, Paul Lamkowski, Jeremy Schmutz, Adam Healey, et al. 2021. “Extensive Genome-Wide Phylogenetic Discordance Is Due to Incomplete Lineage Sorting and Not Ongoing Introgression in a Rapidly Radiated Bryophyte Genus.” *Molecular Biology and Evolution* 38 (7): 2750–2766.

Michael, Todd P. 2014. “Plant Genome Size Variation: Bloating and Purging DNA.” *Briefings in Functional Genomics* 13 (4): 308–317.

Mierziak, Justyna, Kamil Kostyn, and Anna Kulma. 2014. “Flavonoids as Important Molecules of Plant Interactions with the Environment.” *Molecules* 19 (10): 16240–16265.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–1534.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–419.

Mizoi, Junya, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. 2012. “AP2/ERF Family Transcription Factors in Plant Abiotic Stress Responses.” *Biochimica et Biophysica Acta* 1819 (2): 86–96.

Mohamadi, Hamid, Hamza Khan, and Inanc Birol. 2017. “ntCard: A Streaming Algorithm for Cardinality Estimation in Genomics Data.” *Bioinformatics* 33 (9): 1324–1330.

Montenegro, Juan D., Agnieszka A. Golicz, Philipp E. Bayer, Bhavna Hurgobin, HueyTyng Lee, Chon-Kit Kenneth Chan, Paul Visendi et al. 2017. “The Pangenome of Hexaploid Bread Wheat.” *The Plant Journal* 90 (5): 1007–1013.

Moreno-Hagelsieb, Gabriel, and Kristen Latimer. 2008. “Choosing BLAST Options for Better Detection of Orthologs as Reciprocal Best Hits.” *Bioinformatics* 24 (3): 319–324.

Mushtaq, Muhammad A., Qi Pan, Daozong Chen, Qinghua Zhang, Xianhong Ge, and Zaiyun Li. 2016. “Comparative Leaves Transcriptome Analysis Emphasizing on Accumulation of Anthocyanins in Brassica: Molecular Regulation and Potential Interaction with Photosynthesis.” *Frontiers in Plant Science* 7: 311.

Muyle, Aline, Rylan Shearn, and Gabriel Ab Marais. 2017. “The Evolution of Sex Chromosomes and Dosage Compensation in Plants.” *Genome Biology and Evolution* 9 (3): 627–645.

Nakamura, Tsukasa, Kazunori D. Yamada, Kentaro Tomii, and Kazutaka Katoh. 2018. “Parallelization of MAFFT for Large-Scale Multiple Sequence Alignments.” *Bioinformatics* 34 (14): 2490–2492.

National Center for Biotechnology Information. 2021a. “Genome Assembly Reports.” November 5, 2021. Available: ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/eukaryotes.txt.

National Center for Biotechnology Information. 2021b. “Eukaryotic Genome Annotation at NCBI”. November 7, 2021. Available: www.ncbi.nlm.nih.gov/genome/annotation_euk.

National Forestry Database–Canada. 2021. “Area Artificially Regenerated and Number of Seedlings Planted.” July 12, 2021. Available: http://nfdp.ccfm.org/en/data/regeneration.php.

Natural Resources Canada. 2020. “The State of Canada’s Forests Annual Report 2020 and National Inventory Report.” Ottawa

Navarro-Escalante, Lucio, Erick M. Hernandez-Hernandez, Jonathan Nuñez, Flor E. Acevedo, Alejandro Berrio, Luis M. Constantino, Beatriz E. Padilla-Hurtado et al. 2021. “A Coffee Berry Borer (Hypothenemus hampei) Genome Assembly Reveals a Reduced Chemosensory Receptor Gene Repertoire and Male-specific Genome Sequences.” *Scientific reports* 11 (1): 1–17.

Neale, David B., Aleksey V. Zimin, Sumaira Zaman, Alison D. Scott, Bikash Shrestha, Rachael E. Workman, Daniela Puiu et al. 2022. "Assembled and Annotated 26.5 Gbp Coast Redwood Genome: a Resource for Estimating Evolutionary Adaptive Potential and Investigating Hexaploid Origin." *G3* 12 (1) (2022): jkab380.

Neumann, Pavel, Petr Novák, Nina Hoštáková, and Jiří Macas. 2019. "Systematic Survey of Plant LTR-Retrotransposons Elucidates Phylogenetic Relationships of Their Polyprotein Domains and Provides a Reference for Element Classification." *Mobile DNA* 10 (1): 1–17.

Nip, Ka Ming, Readman Chiu, Chen Yang, Justin Chu, Hamid Mohamadi, René L. Warren, and Inanc Birol. 2020. "RNA-Bloom Enables Reference-Free and Reference-Guided Sequence Assembly for Single-Cell Transcriptomes." *Genome Research* 30 (8): 1191–1200.

Novák, Petr, Laura Ávila Robledillo, Andrea Koblížková, Iva Vrbová, Pavel Neumann, and Jirí Macas. 2017. "TAREAN: A Computational Tool for Identification and Characterization of Satellite DNA from Unassembled Short Reads." *Nucleic Acids Research* 45 (12): e111.

Novák, Petr, Maïté S. Guignard, Pavel Neumann, Laura J. Kelly, Jelena Mlinarec, Andrea Koblížková, Steven Dodsworth, et al. 2020. "Repeat-Sequence Turnover Shifts Fundamentally in Species with Large Genomes." *Nature Plants* 6 (11): 1325–1329.

Novák, Petr, Pavel Neumann, Jiří Pech, Jaroslav Steinhaisl, and Jiří Macas. 2013. "RepeatExplorer: A Galaxy-Based Web Server for Genome-Wide Characterization of Eukaryotic Repetitive Elements from next-Generation Sequence Reads." *Bioinformatics* 29 (6): 792–793.

Nystedt, Björn, Nathaniel R. Street, Anna Wetterbom, Andrea Zuccolo, Yao-Cheng Lin, Douglas G. Scofield, Francesco Vezzi, et al. 2013. "The Norway Spruce Genome Sequence and Conifer Genome Evolution." *Nature* 497 (7451): 579–584.

Oberprieler, Rolf G., Adriana E. Marvaldi, and Robert S. Anderson. 2007. "Weevils, Weevils, Weevils Everywhere." *Zootaxa* 1668 (1): 491–520.

Oliver, Matthew J., Dmitri Petrov, David Ackerly, Paul Falkowski, and Oscar M. Schofield. 2007. "The Mode and Tempo of Genome Size Evolution in Eukaryotes." *Genome research* 17 (5): 594–601.

Ohno, Susumu. 2013. Evolution by Gene Duplication. *Springer Science & Business Media*.

Ou, Shujun, Jinfeng Chen, and Ning Jiang. 2018. "Assessing Genome Assembly Quality Using the LTR Assembly Index (LAI)." *Nucleic Acids Research* 46 (21): e126.

Ou, Shujun, and Ning Jiang. 2018. “LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons.” *Plant Physiology* 176 (2): 1410–1422.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 1–18.

Padgitt-Cobb, Lillian K., Sarah B. Kingan, Jackson Wells, Justin Elser, Brent Kronmiller, Daniel Moore, Gregory Concepcion, et al. 2021. “A Draft Phased Assembly of the Diploid Cascade Hop (Humulus Lupulus) Genome.” *Plant Genome* 14 (1): e20072.

Pampolini, Flavia, Thais B. Rodrigues, Ramya S. Leelesh, Tomokazu Kawashima, and Lynne K. Rieske. 2020. “Confocal Microscopy Provides Visual Evidence and Confirms the Feasibility of dsRNA Delivery to Emerald Ash Borer through Plant Tissues.” *Journal of Pest Science* 93 (4): 1143–1153.

Parisot, Nicolas, Carlos Vargas-Chávez, Clément Goubert, Patrice Baa-Puyoulet, Séverine Balmand, Louis Beranger, Caroline Blanc et al. 2021. “The Transposable Element-rich Genome of the Cereal Pest Sitophilus oryzae.” *BMC biology* 19 (1): 1–28.

Paulino, Daniel, René L. Warren, Benjamin P. Vandervalk, Anthony Raymond, Shaun D. Jackman, and Inanç Birol. 2015. “Sealer: A Scalable Gap-Closing Application for Finishing Draft Genomes.” *BMC Bioinformatics* 16: 230.

Pavy, Nathalie, Astrid Deschênes, Sylvie Blais, Patricia Lavigne, Jean Beaulieu, Nathalie Isabel, John Mackay, and Jean Bousquet. 2013. “The Landscape of Nucleotide Polymorphism among 13,500 Genes of the Conifer Picea Glauca, Relationships with Functions, and Comparison with Medicago Truncatula.” *Genome Biology and Evolution* 5 (10): 1910–1925.

Pellati, Federica, Virginia Brighenti, Johanna Sperlea, Lucia Marchetti, Davide Bertelli, and Stefania Benvenuti. 2018. “New Methods for the Comprehensive Analysis of Bioactive Compounds in Cannabis Sativa L. (hemp).” *Molecules* 23 (10): 2639.

Pellicer, Jaume, and Ilia J. Leitch. 2020. “The Plant DNA C-Values Database (release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *The New Phytologist* 226 (2): 301–305.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. “StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads.” *Nature Biotechnology* 33 (3): 290–295.

Phillips, Thomas W., and Gerald N. Lanier. 2000. “Host Specificity in Pissodes strobi (Coleoptera: Curculionidae): Roles of Geography, Genetics, and Behavior.” *The Canadian Entomologist* 132 (6): 811–823.

Phukan, Ujjal J., Gajendra S. Jeena, Vineeta Tripathi, and Rakesh K. Shukla. 2017. “Regulation of Apetala2/Ethylene Response Factors in Plants.” *Frontiers in Plant Science* 8: 150.

Pisupati, Rahul, Daniela Vergara, and Nolan C. Kane. 2018. “Diversity and Evolution of the Repetitive Genomic Content in Cannabis Sativa.” *BMC Genomics* 19 (1): 156.

Plasterk, Ronald H. A., and Henri Gam van Luenen. 2002. “The Tc1/mariner Family of Transposable Elements.” In *Mobile DNA II*, 519–532. American Society of Microbiology.

Ponting, Chris P. 2008. “The Functional Repertoires of Metazoan Genomes.” *Nature Reviews Genetics* 9 (9): 689–698.

Pollard, Martin O., Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, and Manjinder S. Sandhu. 2018. “Long Reads: Their Purpose and Place.” *Human Molecular Genetics* 27 (R2): R234–241.

Powell, Daniel, Ewald Groβe-Wilde, Paal Krokene, Amit Roy, Amrita Chakraborty, Christer Löfstedt, Heiko Vogel, Martin N. Andersson, and Fredrik Schlyter. 2020. “A Highly Contiguous Genome Assembly of a Major Forest Pest, the Eurasian Spruce Bark Beetle Ips Typographus.” *Cold Spring Harbor Laboratory*.

Pruitt, Kim D., Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. 2012. “NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy.” *Nucleic Acids Research* 40 (Database issue): D130–135.

Prunier, Julien, Jérôme Laroche, Jean Beaulieu, and Jean Bousquet. 2011. “Scanning the Genome for Gene SNPs Related to Climate Adaptation and Estimating Selection at the Molecular Level in Boreal Black Spruce.” *Molecular Ecology* 20 (8): 1702–1716.

Prunier, Julien, Jukka-Pekka Verta, and John J. MacKay. 2016. “Conifer Genomics and Adaptation: At the Crossroads of Genetic Diversity and Genome Function.” *The New Phytologist* 209 (1): 44–62.

Ralph, Steven G., Hye Jung E. Chun, Natalia Kolosova, Dawn Cooper, Claire Oddy, Carol E. Ritland, Robert Kirkpatrick et al. 2008. “A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis).” *BMC genomics* 9 (1): 1–17.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. 2020. “GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes.” *Nature Communications* 11 (1): 1432.

Rani, A., S. Vats, M. Sharma, and S. Kumar. 2011. “Catechin Promotes Growth of Arabidopsis Thaliana with Concomitant Changes in Vascular System, Photosynthesis and Hormone Content.” *Biologia Plantarum* 55 (4): 779–782.

Raskina, O., J. C. Barber, E. Nevo, and A. Belyayev. 2008. “Repetitive DNA and Chromosomal Rearrangements: Speciation-Related Events in Plant Genomes.” *Cytogenetic and Genome Research* 120 (3-4): 351–357.

Richter, Todd E., and Pamela C. Ronald. 2000. “The Evolution of Disease Resistance Genes.” *Plant Molecular Evolution* (2000): 195–204.

Reiser, Leonore, Tanya Z. Berardini, Donghui Li, Robert Muller, Emily M. Strait, Qian Li, Yarik Mezheritsky, Andrey Vetushko, and Eva Huala. 2016. “Sustainable Funding for Biocuration: The Arabidopsis Information Resource (TAIR) as a Case Study of a Subscription-Based Funding Model.” *Database: The Journal of Biological Databases and Curation* 2016.

Ren, Guangpeng, Xu Zhang, Ying Li, Kate Ridout, Martha L. Serrano-Serrano, Yongzhi Yang, Ai Liu, et al. 2021. “Large-Scale Whole-Genome Resequencing Unravels the Domestication History of Cannabis Sativa.” *Science Advances* 7 (29): eabg2286.

Ren, Ren, Haifeng Wang, Chunce Guo, Ning Zhang, Liping Zeng, Yamao Chen, Hong Ma, and Ji Qi. 2018. “Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms.” *Molecular Plant* 11 (3): 414–428.

Rigault, Philippe, Brian Boyle, Pierre Lepage, Janice E. K. Cooke, Jean Bousquet, and John J. MacKay. 2011. “A White Spruce Gene Catalog for Conifer Genome Analyses.” *Plant Physiology* 157 (1): 14–28.

Roach, Michael J., Simon A. Schmidt, and Anthony R. Borneman. 2018. “Purge Haplotigs: Allelic Contig Reassignment for Third-Gen Diploid Genome Assemblies.” *BMC Bioinformatics* 19 (1): 460.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.

Rönspies, Michelle, Annika Dorn, Patrick Schindele, and Holger Puchta. 2021. “CRISPR–Cas-Mediated Chromosome Engineering for Crop Improvement and Synthetic Biology.” *Nature Plants* 7 (5): 566–573.

Ross, Samir A., Mahmoud A. ElSohly, Gazi N. N. Sultana, Zlatko Mehmedic, Chowdhury F. Hossain, and Suman Chandra. 2005. “Flavonoid Glycosides and Cannabinoids from the Pollen of Cannabis Sativa L.” *Phytochemical Analysis: PCA* 16 (1): 45–48.

Ruan, Jue, and Heng Li. 2020. “Fast and Accurate Long-Read Assembly with wtdbg2.” *Nature Methods* 17 (2): 155–158.

Russo, Ethan B. 2011. “Taming THC: Potential Cannabis Synergy and Phytocannabinoid-Terpenoid Entourage Effects.” *British Journal of Pharmacology* 163 (7): 1344–1364.

Sahebi, Mahbod, Mohamed M. Hanafi, Andre J. van Wijnen, David Rice, M. Y. Rafii, Parisa Azizi, Mohamad Osman, et al. 2018. "Contribution of Transposable Elements in the Plant's Genome." *Gene* 665 (July): 155–166.

Saito, Kazuki, Keiko Yonekura-Sakakibara, Ryo Nakabayashi, Yasuhiro Higashi, Mami Yamazaki, Takayuki Tohge, and Alisdair R. Fernie. 2013. "The Flavonoid Biosynthetic Pathway in Arabidopsis: Structural and Genetic Diversity." *Plant Physiology and Biochemistry: PPB / Societe Francaise de Physiologie Vegetale* 72: 21–34.

Sakamoto, Koichi, Yukio Akiyama, Kiichi Fukui, Hiroshi Kamada, and Shinobu Satoh. 1998. "Characterization; Genome Sizes and Morphology of Sex Chromosomes in Hemp (Cannabis Sativa L.)." *Cytologia* 63 (4): 459–464.

Sawler, Jason, Jake M. Stout, Kyle M. Gardner, Darryl Hudson, John Vidmar, Laura Butler, Jonathan E. Page, and Sean Myles. 2015. "The Genetic Structure of Marijuana and Hemp." *PloS One* 10 (8): e0133292.

Sayyari, Erfan, James B. Whitfield, and Siavash Mirarab. 2018. "DiscoVista: Interpretable Visualizations of Gene Tree Discordance." *Molecular Phylogenetics and Evolution* 122: 110–115.

Schenke, Dirk, Christoph Böttcher, and Dierk Scheel. 2011. "Crosstalk between Abiotic Ultraviolet-B Stress and Biotic (flg22) Stress Signalling in Arabidopsis Prevents Flavonol Accumulation in Favor of Pathogen Defence Compound Production." *Plant, Cell & Environment* 34 (11): 1849–1864.

Schley, Rowan J., R. Toby Pennington, Oscar Alejandro Pérez-Escobar, Andrew J. Helmstetter, Manuel de la Estrella, Isabel Larridon, Izai Alberto Bruno Sabino Kikuchi, Timothy G. Barraclough, Félix Forest, and Bente Klitgård. 2020. "Introgression across Evolutionary Scales Suggests Reticulation Contributes to Amazonian Tree Diversity." *Molecular Ecology* 29 (21): 4170–4185.

Schnoes, Alexandra, David Ream, Alexander Thorman, Patricia Babbitt, and Iddo Friedberg. 2013. "Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space." *PLOS Computational Biology* 9 (5): e1003063.

Scott, Alison D., Noah W. M. Stenz, Pär K. Ingvarsson, and David A. Baum. 2016. “Whole Genome Duplication in Coast Redwood (Sequoia Sempervirens) and Its Implications for Explaining the Rarity of Polyploidy in Conifers.” *The New Phytologist* 211 (1): 186–193.

Scott, Alison D., Aleksey V. Zimin, Daniela Puiu, Rachael Workman, Monica Britton, Sumaira Zaman, Madison Caballero, et al. 2020. “A Reference Genome Sequence for Giant Sequoia.” *G3* 10 (11): 3907–3019.

Serin, Elise A. R., Harm Nijveen, Henk W. M. Hilhorst, and Wilco Ligterink. 2016. “Learning from Co-Expression Networks: Possibilities and Challenges.” *Frontiers in Plant Science* 7: 444.

Shin, Seunggwan, Dave J. Clarke, Alan R. Lemmon, Emily Moriarty Lemmon, Alexander L. Aitken, Stephanie Haddad, Brian D. Farrell, Adriana E. Marvaldi, Rolf G. Oberprieler, and Duane D. McKenna. 2018. “Phylogenomic Data Yield New and Robust Insights into the Phylogeny and Evolution of Weevils.” *Molecular Biology and Evolution* 35 (4): 823–836.

Shirley, B. W., W. L. Kubasek, G. Storz, E. Bruggemann, M. Koornneef, F. M. Ausubel, and H. M. Goodman. 1995. “Analysis of Arabidopsis Mutants Deficient in Flavonoid Biosynthesis.” *The Plant Journal: For Cell and Molecular Biology* 8 (5): 659–671.

Shumate, Alaina, and Steven L. Salzberg. 2020. “Liftoff: Accurate Mapping of Gene Annotations.” *Bioinformatics* 37 (12): 1639–1643.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–3212.

Skaliter, Oded, Jasmin Ravid, Elena Shklarman, Nadav Ketrarou, Noam Shpayer, Julius Ben Ari, Gony Dvir, Moran Farhi, Yuling Yue, and Alexander Vainstein. 2019. “Ectopic Expression of PAP1 Leads to Anthocyanin Accumulation and Novel Floral Color in Genetically Engineered Goldenrod (Solidago Canadensis L.).” *Frontiers in Plant Science* 10: 1561.

Small, Ernest. 2015. “Evolution and Classification of Cannabis Sativa (marijuana, Hemp) in Relation to Human Utilization.” *The Botanical Review; Interpreting Botanical Progress* 81 (3): 189–294.

Small, Ernest, and Arthur Cronquist. 1976. “A Practical and Natural Taxonomy for Ccannabis.” *TAXON*: 405–435.

Smit, Arian F.A., Robert Hubley, and Green Philip. 2008. “Repeat Masker.” RepeatMasker Open-4.0. 2008. Available: http://www.repeatmasker.org.

Smith, George P. 1976. “Evolution of Repeated DNA Sequences by Unequal Crossover.” *Science* 191 (4227): 528–535.

Smith, S. G., and B. A. Sugden. 1969. “Host Trees and Breeding Sites of Native North American Pissodes Bark Weevils, with a Note on Synonymy.” *Annals of the Entomological Society of America* 62 (1): 146–148.

Smith, Stephen A., Michael J. Moore, Joseph W. Brown, and Ya Yang. 2015. “Analysis of Phylogenomic Datasets Reveals Conflict, Concordance, and Gene Duplications with Examples from Animals and Plants.” *BMC Evolutionary Biology* 15 (1): 1–15.

Soorni, Aboozar, Reza Fatahi, David C. Haak, Seyed Alireza Salami, and Aureliano Bombarely. 2017. “Assessment of Genetic Diversity and Population Structure in Iranian Cannabis Germplasm.” *Scientific Reports* 7 (1): 15668.

Sørenson, Th A. 1948. “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.” *Biol. Skar.* 5: 1–34.

Spruce-Up consortium. “Spruce-Up.” Accessed November 7, 2021. https://spruce-up.ca.

Spurgeon, D. 2001. “Canada Legalises the Medical Use of Cannabis.” *BMJ* 323 (7304): 68–68.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–1313.

Stanke, Mario, Mark Diekhans, Robert Baertsch, and David Haussler. 2008. “Using Native and Syntenically Mapped cDNA Alignments to Improve de Novo Gene Finding.” *Bioinformatics* 24 (5): 637–44.

Stanke, Mario, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack. 2006. “Gene Prediction in Eukaryotes with a Generalized Hidden Markov Model That Uses Hints from External Sources.” *BMC Bioinformatics* 7 (1): 1–11.

Stevens, Kristian A., Jill L. Wegrzyn, Aleksey Zimin, Daniela Puiu, Marc Crepeau, Charis Cardeno, Robin Paul, et al. 2016. “Sequence of the Sugar Pine Megagenome.” *Genetics* 204 (4): 1613–1626.

Stival Sena, Juliana, Isabelle Giguère, Brian Boyle, Philippe Rigault, Inanc Birol, Andrea Zuccolo, Kermit Ritland et al. 2014. “Evolution of gene structure in the conifer Picea glauca: a comparative analysis of the impact of intron size.” *BMC plant biology* 14 (1): 1–16.

Stork, Nigel E., James McBroom, Claire Gely, and Andrew J. Hamilton. 2015. “New Approaches Narrow Global Species Estimates for Beetles, Insects, and Terrestrial Arthropods.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (24): 7519–7523.

Stracke, Ralf, Hirofumi Ishihara, Gunnar Huep, Aiko Barsch, Frank Mehrtens, Karsten Niehaus, and Bernd Weisshaar. 2007. “Differential Regulation of Closely Related R2R3-MYB Transcription Factors Controls Flavonol Accumulation in Different Parts of the Arabidopsis Thaliana Seedling.” *The Plant Journal* 50 (4): 660–677.

Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. 2003. “A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.” *Science* 302 (5643): 249–255.

Suyama, Mikita, David Torrents, and Peer Bork. 2006. “PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments.” *Nucleic Acids Research* 34: W609–612.

Tagu, Denis, John K. Colbourne, and Nicolas Nègre. 2014. “Genomic Data Integration for Ecological and Evolutionary Traits in Non-Model Organisms.” *BMC genomics* 15 (1): 1–16.

Talenti, Andrea, and James Prendergast. 2021. “Nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over.” *Genome Biology and Evolution* 13 (9): evab183.

Taylor, Gregory A., Heather Kirk, Lauren Coombe, Shaun D. Jackman, Justin Chu, Kane Tse, Dean Cheng, et al. 2018. “The Genome of the North American Brown Bear or Grizzly: Ursus Arctos Ssp. Horribilis.” *Genes* 9 (12): 598.

Tohge, Takayuki, Yasutaka Nishiyama, Masami Yokota Hirai, Mitsuru Yano, Jun-Ichiro Nakajima, Motoko Awazuhara, Eri Inoue, et al. 2005. “Functional Genomics by Integrated Analysis of Metabolome and Transcriptome of Arabidopsis Plants over-Expressing an MYB Transcription Factor.” *The Plant Journal: For Cell and Molecular Biology* 42 (2): 218–235.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. “Highly Accurate Protein Structure Prediction for the Human Proteome.” *Nature* 596 (7873): 590–596.

“UniProt: The Universal Protein Knowledgebase in 2021.” 2021. *Nucleic Acids Research* 49 (D1): D480–489.

Vallée, Geneviève C., Daniella Santos Muñoz, and David Sankoff. 2016. “Economic Importance, Taxonomic Representation and Scientific Priority as Drivers of Genome Sequencing Projects.” *BMC Genomics* 17 (10): 125–133.

Van Dam, Matthew H., Analyn Anzano Cabras, James B. Henderson, Andrew J. Rominger, Cynthia Pérez Estrada, Arina D. Omer, Olga Dudchenko, Erez Lieberman Aiden, and Athena W. Lam. 2021. “The Easter Egg Weevil (Pachyrhynchus) Genome Reveals Syntenic Patterns in Coleoptera across 200 Million Years of Evolution.” *PLoS Genetics* 17 (8): e1009745.

Vanhoenacker, Gerd, Philippe Van Rompaey, Denis De Keukeleire, and Pat Sandra. 2002. “Chemotaxonomic Features Associated with Flavonoids of Cannabinoid-Free Cannabis (Cannabis Sativa Subsp. Sativa L.) in Relation to Hops (Humulus Lupulus L.).” *Natural Product Letters* 16 (1): 57–63.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–746.

Venturini, Luca, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck. 2018. “Leveraging Multiple Transcriptome Assembly Methods for Improved Gene Structure Annotation.” *GigaScience* 7 (8): giy093.

Verdan, Andrea M., Hsiao C. Wang, Carla R. García, William P. Henry, and Julia L. Brumaghim. 2011. “Iron Binding of 3-Hydroxychromone, 5-Hydroxychromone, and Sulfonated Morin: Implications for the Antioxidant Activity of Flavonols with Competing Metal Binding Sites.” *Journal of Inorganic Biochemistry* 105 (10): 1314–1322.

Vivar-Quintana, A. M., C. Santos-Buelga, and J. C. Rivas-Gonzalo. 2002. “Anthocyanin-Derived Pigments and Colour of Red Wines.” *Analytica Chimica Acta* 458 (1): 147–155.

Vogt, Thomas. 2010. “Phenylpropanoid Biosynthesis.” *Molecular Plant* 3 (1): 2–20.

Wang, Wenle, Jinfan Xu, Huiyong Fang, Zhijun Li, and Minhui Li. 2020. “Advances and Challenges in Medicinal Plant Breeding.” *Plant Science: An International Journal of Experimental Plant Biology* 298 (September): 110573.

Wang, X. Q., D. C. Tank, and T. Sang. 2000. “Phylogeny and Divergence Times in Pinaceae: Evidence from Three Genomes.” *Molecular Biology and Evolution* 17 (5): 773–781.

Warren, René L., Christopher I. Keeling, Macaire Man Saint Yuen, Anthony Raymond, Greg A. Taylor, Benjamin P. Vandervalk, Hamid Mohamadi, et al. 2015. “Improved White Spruce (Picea Glauca) Genome Assemblies and Annotation of Large Gene Families of Conifer Terpenoid and Phenolic Defense Metabolism.” *The Plant Journal: For Cell and Molecular Biology* 83 (2): 189–212.

Wegrzyn, Jill L., John D. Liechty, Kristian A. Stevens, Le-Shin Wu, Carol A. Loopstra, Hans A. Vasquez-Gross, William M. Dougherty et al. 2014. “Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation.” *Genetics* 196 (3): 891–909.

Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. “Direct Determination of Diploid Genome Sequences.” *Genome Research* 27 (5): 757–767.

Wen, Weiwei, Saleh Alseekh, and Alisdair R. Fernie. 2020. “Conservation and Diversification of Flavonoid Metabolism in the Plant Kingdom.” *Current Opinion in Plant Biology* 55: 100–108.

Werren, John H., Laura Baldo, and Michael E. Clark. 2008. “Wolbachia: Master Manipulators of Invertebrate Biology.” *Nature Reviews. Microbiology* 6 (10): 741–751.

Whitehill, Justin G. A., and Jörg Bohlmann. 2019. “A Molecular and Genomic Reference System for Conifer Defence against Insects.” *Plant, Cell & Environment* 42 (10): 2844–2859.

Whitehill, Justin G. A., Hannah Henderson, Mathias Schuetz, Oleksandr Skyba, Macaire Man Saint Yuen, John King, A. Lacey Samuels, Shawn D. Mansfield, and Jörg Bohlmann. 2016. “Histology and Cell Wall Biochemistry of Stone Cells in the Physical Defence of Conifers against Insects.” *Plant, Cell & Environment* 39 (8): 1646–1661.

Wu, Changcheng, and Jian Lu. 2019. “Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution.” *Genes* 10 (5): 338.

Xu, Lin, Hong Chen, Xiaohua Hu, Rongmei Zhang, Ze Zhang, and Z. W. Luo. 2006. "Average Gene Length is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms." *Molecular biology and evolution* 23 (6): 1107–1108.

Xu, Zhao, and Hao Wang. 2007. “LTR_FINDER: An Efficient Tool for the Prediction of Full-Length LTR Retrotransposons.” *Nucleic Acids Research* 35 (2): W265–268.

Yandell, Mark, and Daniel Ence. 2012. “A Beginner’s Guide to Eukaryotic Genome Annotation.” *Nature Reviews. Genetics* 13 (5): 329–342.

Yang, Mei-Qing, Robin van Velzen, Freek T. Bakker, Ali Sattarian, De-Zhu Li, and Ting-Shuang Yi. 2013. “Molecular Phylogenetics and Character Evolution of Cannabaceae.” *Taxon* 62 (3): 473–485.

Yang, Ziheng. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” *Molecular Biology and Evolution* 24 (8): 1586–1591.

Yang, Ziheng, and Rasmus Nielsen. 2002. “Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites along Specific Lineages.” *Molecular Biology and Evolution* 19 (6): 908–917.

Yang, Ziheng, and Mario dos Reis. 2011. “Statistical Properties of the Branch-Site Test of Positive Selection.” *Molecular Biology and Evolution* 28 (3): 1217–1228.

Yan, Yifan, Changzheng Song, Luigi Falginella, and Simone D. Castellarin. 2020. “Day Temperature Has a Stronger Effect Than Night Temperature on Anthocyanin and Flavonol Accumulation in ‘Merlot’ (Vitis Vinifera L.) Grapes During Ripening.” *Frontiers in Plant Science* :1095.

Yeaman, Sam, Kathryn A. Hodgins, Katie E. Lotterhos, Haktan Suren, Simon Nadeau, Jon C. Degner, Kristin A. Nurkowski, et al. 2016. “Convergent Local Adaptation to Climate in Distantly Related Conifers.” *Science* 353 (6306): 1431–1433.

Yeo, Sarah, Lauren Coombe, René L. Warren, Justin Chu, and Inanç Birol. 2018. “ARCS: Scaffolding Genome Drafts with Linked Reads.” *Bioinformatics* 34 (5): 725–731.

Zerega, Nyree J. C., Wendy L. Clement, Shannon L. Datwyler, and George D. Weiblen. 2005. “Biogeography and Divergence Times in the Mulberry Family (Moraceae).” *Molecular Phylogenetics and Evolution* 37 (2): 402–416.

Zhang, Chao, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. “ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees.” *BMC Bioinformatics* 19 (6): 15–30.

Zhang, Shao-Qian, Li-Heng Che, Yun Li, Dan Liang, Hong Pang, Adam Ślipiński, and Peng Zhang. 2018. “Evolutionary History of Coleoptera Revealed by Extensive Sampling of Genes and Species.” *Nature Communications* 9 (1): 205.

Zhang, Xuebin, Carolina Abrahan, Thomas A. Colquhoun, and Chang-Jun Liu. 2017. “A Proteolytic Regulator Controlling Chalcone Synthase Stability and Flavonoid Biosynthesis in Arabidopsis.” *The Plant Cell* 29 (5): 1157–1174.

Zhao, Juan-Ying, Zhi-Wei Lu, Yue Sun, Zheng-Wu Fang, Jun Chen, Yong-Bin Zhou, Ming Chen, You-Zhi Ma, Zhao-Shi Xu, and Dong-Hong Min. 2020. “The Ankyrin-Repeat Gene GmANK114 Confers Drought and Salt Tolerance in Arabidopsis and Soybean.” *Frontiers in Plant Science* 11: 584167.

Zhou, Li-Li, Hai-Nian Zeng, Ming-Zhu Shi, and De-Yu Xie. 2008. “Development of Tobacco Callus Cultures over Expressing Arabidopsis PAP1/MYB75 Transcription Factor and Characterization of Anthocyanin Biosynthesis.” *Planta* 229 (1): 37–51.

Zhou, Naihui, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, et al. 2019. “The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens.” *Genome Biology* 20 (1): 244.

Zhu, Liucun, Ying Zhang, Wen Zhang, Sihai Yang, Jian-Qun Chen, and Dacheng Tian. 2009. “Patterns of Exon-Intron Architecture Variation of Genes in Eukaryotic Genomes.” *BMC genomics* 10 (1): 1–12.

Zhuang, Weibing, Hongxue Wang, Tianyu Liu, Tao Wang, Fengjiao Zhang, Xiaochun Shu, Henghua Zhai, and Zhong Wang. 2019. “Integrated Physiological and Genomic Analysis Reveals Structural Variations and Expression Patterns of Candidate Genes for Colored- and Green-Leaf Poplar.” *Scientific Reports* 9 (1): 1–12.

Zimmer, Andreas, Daniel Lang, Sandra Richardt, Wolfgang Frank, Ralf Reski, and Stefan A. Rensing. 2007. “Dating the Early Evolution of Plants: Detection and Molecular Clock Analyses of Orthologs.” *Molecular Genetics and Genomics* 278 (4): 393–402.

Appendix

Appendix A Chapter 2 - Supplemental material

Table A.1 Geographical origin and local climate of the collected representative spruce accessions used for DNA extraction. The seed origin is the same as the geographic location for *P. glauca*, *P. sitchensis*, and interior spruces. *P. engelmannii* seeds were collected from New Mexico (United States), and trees were raised in a provenance test in British Columbia (Canada). The average temperature and average total precipitation are shown for the following stations and year ranges: interior spruce - PG29: PRINCE GEORGE STP (2009-2018); *P. sitchensis* - Q903: MASSET A (2010-2016); *P. glauca* - WS77111: BALDWIN (2009-2018); *P. engelmannii* - Se404-851; VERNON NORTH (2009-2018), PALO NM, US (2011-2018); **Sources:** climate.weather.gc.ca, www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets.

Genotype	Location	Geographical origin	Average Temperature (C°)	Average annual precipitation (mm)
P. engelmannii Se404-851	Raised: Kalamalka Research Station of the British Columbia Ministry of Forests, Lands, Natural Resource Operations and Rural Development, Vernon, British Columbia, Canada	**Raised**: 36.300, -105.400 **Seeds**: 50.244, -119.278	9.37, 6.35	357.55, 514.51
P. sitchensis Q903	Raised: University of British Columbia, Vancouver, British Columbia, Canada Origin: Haida Gwaii island, British Columbia, Canada	**Raised**: 49.261, -123.252 **Origin**: 53.917, -132.083	8.13	1404.07
P. glauca WS77111	Valcartier Experimental Station of Natural Resources Canada near Quebec City, Quebec, Canada	44.330, -78.150	7.62	629.79
Interior spruce PG29	Kalamalka Research Station of the British Columbia Ministry of Forests, Lands, Natural Resource Operations and Rural Development, Vernon, British Columbia, Canada	53.867, -122.250	6.18	396.97

Table A.2 Sequencing reads and corresponding fold coverage (millions of generated reads) for each genome assembly. The fold coverage is estimated for each platform, and the read format is the sum of sequenced bases normalized by the estimated genome size of 21 Gb. Each read format is annotated with the date of data generation. **Chromium linked reads**: Linked reads sequencing on 10x Genomics Chromium platform, **GemCode linked reads**: Linked reads sequencing on 10x Genomics GemCode platform, **HiSeq, and MiSeq**: Illumina short reads, **MPET**: Mate-pair reads with large fragment size; **ONT**: Oxford Nanopore technology long reads.

Platform	Read format	*P. engelmannii* Se404-851	*P. sitchensis* Q903	*P. glauca* WS77111	Interior spruce PG29
Chromium linked reads	128bp-151bp	45.8 (6,892M) Oct 2018	58.3 (8,774M) Feb 2017	37.9 (5,711M) Jan 2017	28.9 (4,349M) May 2018
GemCode linked reads	116bp-126bp	-	5.8 (1,000M) Oct 2015	-	-
HiSeq	2x150 bp	-	-	34.0 (4,754M) Dec 2013	60.5 (8,464M) 2012
HiSeq	2x250 bp	29.6 (2,486M) Oct 2018	13 (1,097M) Mar 2020	19.6 (1,644M) May 2016	-
MiSeq	2x300 bp	-	-	3.6 (249M) Mar 2017	2.6 (182M) 2012
MiSeq	2x500 bp	-	-	-	2.5 (107M) 2012
MPET	2x100 bp	-	-	10.7 (2247M) Apr 2019	23.9 (5,022M) 2012
MPET	2x150 bp	11.2 (1,561M) Apr 2019	-	-	-
ONT	Variable, read N50 as indicated	2.6 (5.86M) N50=14,150bp Jan–Mar 2019	4.9 (9.59M) N50=14,866bp Nov 2017–Sep 2018	-	-
Total cov.		**89.2**	**82.0**	**109.5**	**118.4**

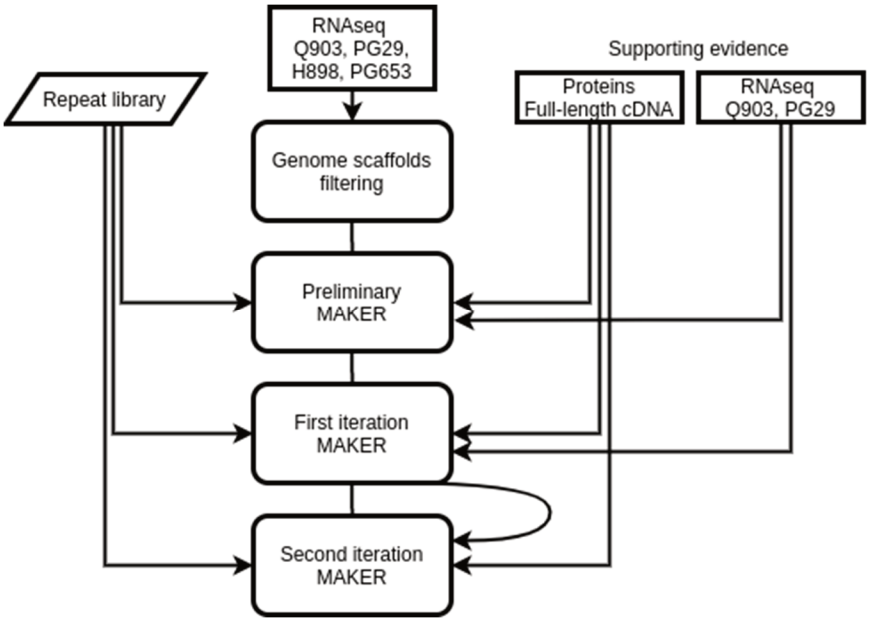

Figure A.1 Scaffold filtering and genome annotation pipeline. The genome was filtered for scaffolds missing complete genes and containing a high fraction of repeats. A preliminary MAKER run was used to train the gene prediction models in AUGUSTUS (Stanke et al. 2008) and SNAP (Korf 2004). Species-specific RNAseq libraries were used in the preliminary and the first MAKER iterations. The second MAKER iteration used combined evidence from all four genotypes generated in the first MAKER iteration. All MAKER prediction steps used supporting evidence from full-length cDNAs and proteins. A custom repeat library was included in the annotation at each MAKER step.

Table A.3 RNAseq samples used for genome annotation and scaffolds filtering. Short read RNAseq samples were assembled and filtered for contaminants. Long reads from the *P. sitchensis* mixed tissues sample (PRJNA304257) were error-corrected before the annotation.

Taxon	Genotype	Number of samples	BioProject	Application
Interior spruce	PG29	8	PRJNA210511	Filtering, Gene prediction
P. glauca	PG653	30	Submission in progress	Filtering
P. glauca	PG653	14	PRJNA290034	Filtering
P. glauca	PG653	16	PRJNA309861	Filtering
P. sitchensis	Q903	6	Submission in progress	Filtering, Gene prediction
P. sitchensis	Q903	12	PRJNA398042	Filtering, Gene prediction
P. sitchensis	Q903	1	PRJNA304257	Filtering, Gene prediction
P. sitchensis	H898	6	Submission in progress	Filtering
P. sitchensis	H898	12	PRJNA398042	Filtering

Table A.4 Conifer genomes used for phylogenomic comparison. The genomes are downloaded from their reference IDs and links to the published nuclear and organellar genomes.

Reference	Version	Nuclear genome	Mitochondrial genome	Chloroplast genome
P. glauca	2	GCA_000966675	https://www.bcgsc.ca/downloads/btl/Spruce/Pglauca_WS77111/mitochondria_assembly/WS77111_mt.fa	MK174379
Interior spruce	5	GCA_000411955	LKAM01000001–LKAM01000036	KT634228
P. engelmannii	1	GCA_009831015	https://www.bcgsc.ca/downloads/btl/Spruce/Engelmann_Se404-851/mitochondria_assembly/Se404-851_mt.fa	MK241981
P. sitchensis	1	GCA_010110895	MK697696–MK697708	KU215903
P. abies	1	GCA_900067695	GCA_900067695 - sequenced labeled as mitochondrial	NC_021456
P. taeda	1	GCA_000404065	https://treegenesdb.org/FTP/Genomes/Pita/mito/ptaeda.mito.scafSeq.gz	KY964286
P. lambertiana	1.5	GCA_001447015	ftp://ccb.jhu.edu/pub/data/Sugar_pine/Assembly/mito/mito.fa	EU998743

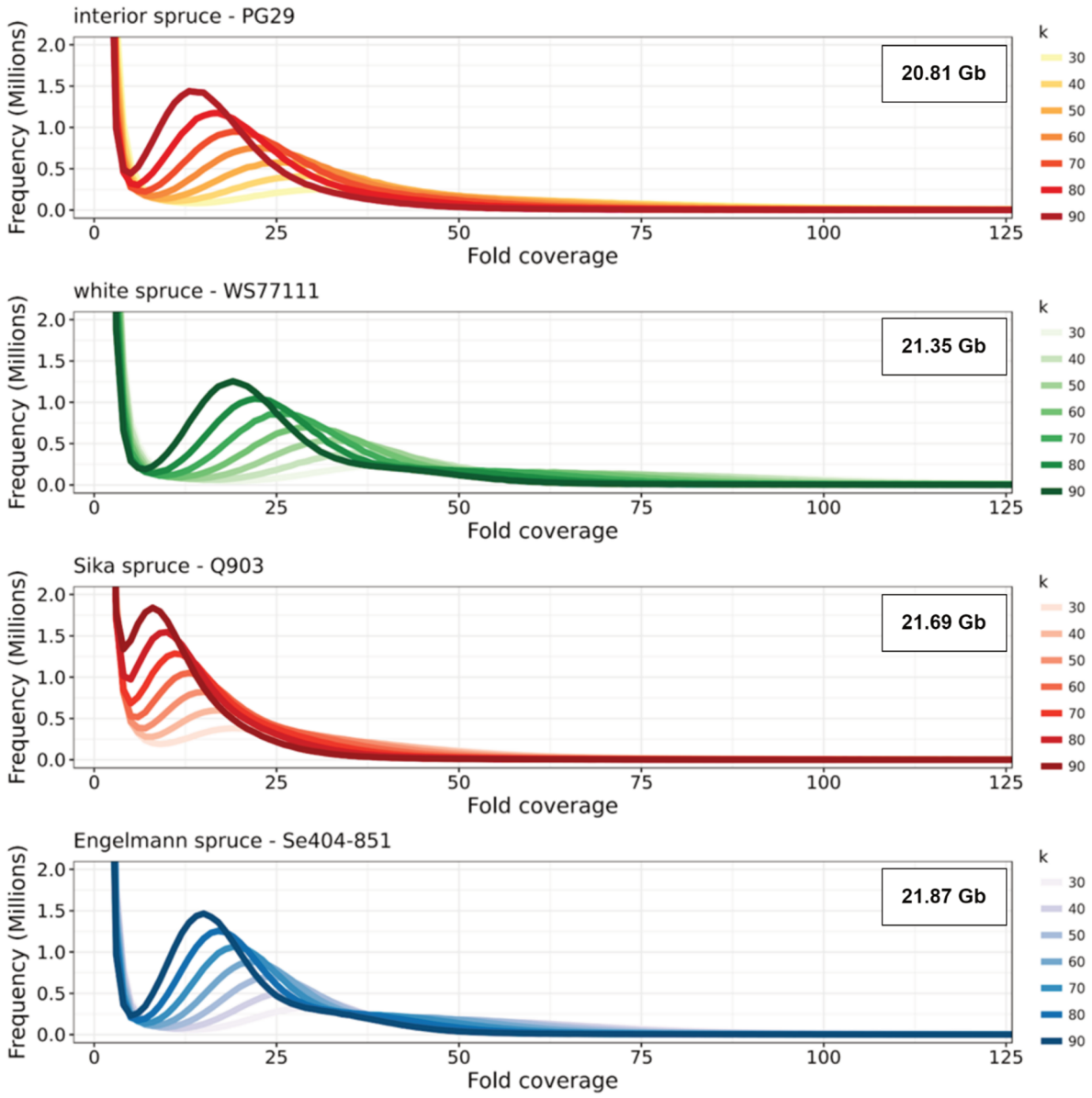

Figure A.2 K-mer histograms and genome sizes estimate for interior, *P. glauca*, *P. sitchensis,* and *P. engelmannii.* Genome sizes are estimated for the k-mer lengths 30, 40, 50, 60, 70, 80, and 90 bp.

Table A.5 Genome assembly statistics and gene completeness for *Picea engelmannii*, *P. sitchensis*, *P. glauca,* and interior spruce. The assembly statistics shown are the number of scaffolds ≥1 kb, NG50 is calculated on 21 Gb genome size. BUSCO "complete," represented by "single copy" and "duplicated," is a metric of the gene space reconstruction using the Embryophyta core gene set.

Taxon	No. of scaffolds	Longest scaffold	Scaffold NG50	Reconstructed size (Gb)	BUSCO single copy (%)	BUSCO duplicated (%)
P. engelmannii	946,053	6,646,027	355,449	20.75	40.3	8.2
P. sitchensis	1,770,974	1,973,130	38,458	18.22	29.5	7.4
P. glauca	2,443,500	4,209,077	131,339	21.58	39.9	8.0
Interior spruce	2,064,648	3,588,992	121,714	20.14	41.1	7.8

Table A.6 Genome assembly statistics and gene completeness for the conifer species used in the comparative study with the North American spruces: *Picea abies*, *Pinus taeda*, *and Pinus lambertiana*. The assembly statistics shown are the number of scaffolds ≥ 1 kb, NG50 is calculated on 21 Gb genome size. BUSCO "complete," represented by "single copy" and "duplicated," is a metric of the gene space reconstruction using the Embryophyta core gene set.

Taxon	No. of scaffolds	Longest scaffold	Scaffold NG50	Reconstruction size (Gb)	BUSCO single copy (%)	BUSCO duplicated (%)
P. abies	1,963,820	194,057	1,000	9.16	28.9	6.0
P. taeda	16,610	2,237,000	2,540,398	19.82	54.6	7.1
P. lambertiana	991,362	8,214,401	78,650	16.78	48.7	7.6

Table A.7 Genome annotation completeness for the *Picea abies*, *Pinus taeda*, and *Pinus lambertiana*.

The statistics include the number of annotated genes and transcripts, mRNA, exon, and intron median lengths, and corresponding annotation completeness. The "median gene length" described in the table contains introns, while the "median mRNA length" lacks introns. The "Median exon length" in *P. lambertiana* and *P. taeda* represents the CDS median length; the authors' annotations lacked "exon" features. The annotation completeness is shown as percent Pfam Conserved Domain Arrangements (CDA) in plants and BUSCO "complete" core set genes.

Taxon	Total genes	Total mRNA	Median gene length (bp)	Median mRNA length (bp)	Median exon length (bp)	Median intron length (bp)	BUSCO complete (%)	Total CDA completeness (%)
P. abies	26,437	26,437	1,366	714	173	173	26.3	35.85
P. taeda	47,602	47,602	2,252	700	166*	184	59.5	69.28
P. lambertiana	31,253	38,518	4,330	978	143*	271	17.5	28.90

Table A.8 - Number of genes in orthogroups (OGs) and number of orthogroups per species. The genes annotated in North American spruces are classified together with *Picea abies*, *Pinus taeda*, and *Pinus lambertiana* in gene families with the OrthoFinder tool suite. The genes in gene families ranged between ~88 and 90% for the four annotated spruces with ~14,000 orthogroups per genotype.

Taxon	Number of genes	Genes in OGs	Total OGs
Picea engelmannii	31,753	30,832 (89.7%)	15,022
Picea sitchensis	30,312	26,923 (88.8%)	14,635
Picea glauca	30,409	27,493 (90.4%)	14,209
Interior spruce	28,943	25,998 (89.8%)	14,237
Picea abies	26,437	21,353 (80.8%)	12,711
Pinus taeda	47,602	28,576 (60.0%)	9,470
Pinus lambertiana	31,253	23,955 (76.6%)	12,166

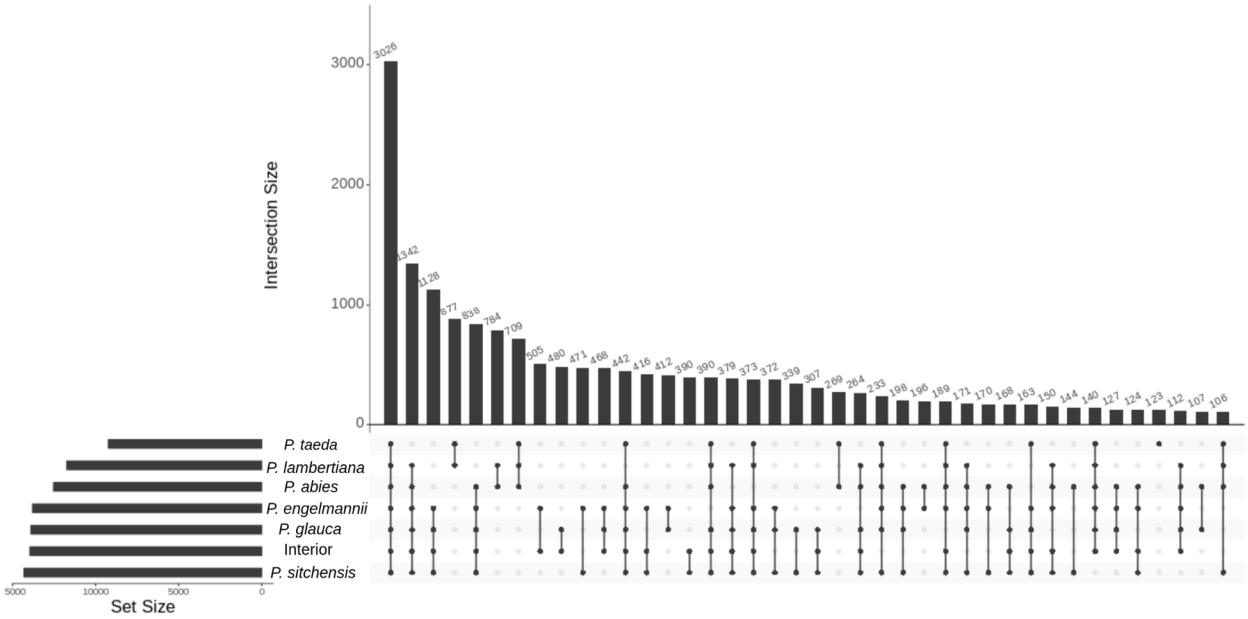

Figure A.3 Orthogroups overlap between spruces and pine taxa. Based on the overlap, a common core set of 3,026 gene orthogroups was found in all the analyzed conifers, 838 in all spruces, and 1,128 in the four North American spruces.

Table A.9 CAFE analysis for gene family expansion/contraction for each genotype. The "Expanded" and "Contracted" columns in the table show the total gene families with expansion or contraction, respectively; the number in brackets is the significant gene families based on the Viterbi algorithm and $P \leq 0.01$, as reported in Figure 2.3a in the main text. "Genes gained" and "genes lost" are the total number of genes gained or lost in the gene families. The "Avg. expansion" shows the average gene family expansion or contraction per species. Most species have an overall gene expansion, except for *Picea abies*, *Pinus lambertiana,* and interior spruce.

Genotype	Expanded	Contracted	Genes gained	Genes lost	Avg. expansion
Picea glauca	2,221 (59)	1,403 (56)	3,405	1,855	0.1638
Interior spruce	1,392 (72)	1,924 (81)	2,313	2,625	-0.03298
Picea engelmannii	2,762 (110)	1,351 (20)	4,846	1,653	0.3375
Picea sitchensis	1,750 (52)	1,647 (13)	2,905	1,984	0.0973
Picea abies	1,302 (35)	2,839 (85)	2,301	4,266	-0.2077
Pinus taeda	1,970 (98)	3,532 (11)	6,594	4,137	0.25975
Pinus lambertiana	1,603 (24)	2,126 (43)	3,093	3,694	-0.0635

Table A.10 Top significant GO biological process terms of expanded gene families in *P. engelmannii.*

GO term	*P* corr	PFAM description
response to organic substance	8.42e-10	1) 2OG-Fe(II) oxygenase superfamily; 2) ABC transporter; 3) Acetyltransferase (GNAT) family; 4) alpha/beta hydrolase fold; 5) AP2 domain; 6) Argonaute linker 1 domain; 7) Ataxin-2 C-terminal region; 8) ATPase family associated with various cellular activities (AAA); 9) BTB/POZ domain; 10) Cyclin, C-terminal domain; 11) Cyclin, N-terminal domain; 12) Cytochrome P450; 13) Endonuclease/Exonuclease/phosphatase family; 14) Eukaryotic-type carbonic anhydrase; 15) Galactose oxidase, central domain; 16) Hsp20/alpha crystallin family; 17) Legume lectin domain; 18) Leucine Rich repeat; 19) Leucine rich repeat; 20) Leucine rich repeat N-terminal domain; 21) non-haem dioxygenase in morphine synthesis N-terminal; 22) PA domain; 23) Peroxidase; 24) Phosphofructokinase; 25) Piwi domain; 26) PPR repeat; 27) Protein kinase domain; 28) Protein phosphatase 2C; 29) Protein tyrosine kinase; 30) Receptor family ligand binding region; 31) Ring finger domain; 32) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 33) STAS domain; 34) Terpene synthase family, metal binding domain; 35) Terpene synthase, N-terminal domain; 36) Ubiquitin-conjugating enzyme; 37) Ubiquitin family; 38) UDP-glucoronosyl and UDP-glucosyl transferase; 39) Zinc finger, C3HC4 type (RING finger); 40) Zinc knuckle; 41) P-type ATPase; 42) PAZ domain; 43) CDC48 N-terminal domain; 44) CDC48 C-terminal domain; 45) Vps4 C terminal oligomerisation domain; 46) N-terminal domain of argonaute
response to stress	1.68e-09	1) 2OG-Fe(II) oxygenase superfamily; 2) ABC transporter; 3) Aminotransferase class I and II; 4) AP2 domain; 5) Argonaute linker 1 domain; 6) Ataxin-2 C-terminal region; 7) ATPase family associated with various cellular activities (AAA); 8) BTB/POZ domain; 9) Cellulose synthase; 10) Cyclin, C-terminal domain; 11) Cyclin, N-terminal domain; 12) Cytochrome P450; 13) Endonuclease/Exonuclease/phosphatase family; 14) FKBP-type peptidyl-prolyl cis-trans isomerase; 15) GDSL-like Lipase/Acylhydrolase; 16) Glycosyl transferase family; 17) Glycosyl transferase family 8; 18) Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase; 19) Hsp20/alpha crystallin family; 20) Legume lectin domain; 21) Leucine Rich Repeat; 22) Leucine Rich repeat; 23) Leucine rich repeat; 24) Leucine rich repeat N-terminal domain; 25) Lipase (class 3); 26) Matrixin; 27) Multicopper oxidase; 28) NAF domain; 29) NB-ARC domain; 30) PA domain; 31) Peroxidase; 32) Piwi domain; 33) Protein kinase domain; 34) Protein phosphatase 2C; 35) Protein tyrosine kinase; 36) Putative peptidoglycan binding domain; 37) Radical SAM superfamily; 38) Receptor family ligand binding region; 39) Replication factor-A C terminal domain; 40) Replication factor-A protein 1, N-terminal domain; 41) Replication protein A OB domain; 42) Ring finger domain; 43) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 44) STAS domain; 45) Terpene synthase family, metal binding domain; 46) Terpene synthase, N-terminal domain; 47) Thaumatin family; 48) Ubiquitin-conjugating enzyme; 49) Ubiquitin family; 50) Zinc finger, C3HC4 type (RING finger); 51) Zinc knuckle; 52) P-type ATPase; 53) PAZ domain; 54) CDC48 N-terminal domain; 55) CDC48 N-terminal domain; 56) Vps4 C terminal oligomerisation domain; 57) N-terminal domain of argonaute; 58) Argonaute linker 2 domain
catabolic process	3.86e-09	1) 2OG-Fe(II) oxygenase superfamily; 2) ABC transporter; 3) Acetyltransferase (GNAT) family; 4) Alcohol dehydrogenase GroES-like domain; 5) alpha/beta hydrolase fold; 6) Aminotransferase class I and II; 7) Amino-transferase class IV; 8) Argonaute linker 1 domain; 9) Argonaute linker 2 domain; 10) ATPase family associated with various cellular activities (AAA); 11) BTB And C-terminal Kelch; 12) BTB/POZ domain; 13) CDC48 C-terminal domain; 14) CDC48 N-terminal domain; 15) Creatinase/Prolidase N-terminal domain; 16) Cyclin, C-terminal domain; 17) Cyclin, N-terminal domain; 18) Cys/Met metabolism PLP-dependent enzyme; 19) Cytochrome P450; 20) Endonuclease/Exonuclease/phosphatase family; 21) F-box domain; 22) Kelch motif; 23) Leucine Rich repeat; 24) Leucine rich repeat; 25) Lipase (class 3); 26) Matrixin; 27) NADH; 28) non-haem dioxygenase in morphine synthesis N-terminal; 29) N-terminal domain of argonaute; 30) PA domain; 31) PAZ domain; 32) Peptidase inhibitor I9; 33) Phosphofructokinase; 34) Piwi domain; 35) Protein kinase domain; 36) Putative peptidoglycan binding domain; 37) Ring finger domain; 38) Skp1 family, dimerisation domain; 39) Skp1 family, tetramerisation domain; 40) Subtilase family; 41) Ubiquitin-conjugating enzyme; 42) Ubiquitin family; 43) Vps4 C terminal oligomerisation domain; 44) Zinc finger, C3HC4 type (RING finger); 45) Zinc knuckle;
response to endogenous stimulus	5.09e-07	1) 2OG-Fe(II) oxygenase superfamily; 2) ABC transporter; 3) Acetyltransferase (GNAT) family; 4) alpha/beta hydrolase fold; 5) AP2 domain; 6) BTB/POZ domain; 7) Cyclin, C-terminal domain; 8) Cyclin, N-terminal domain; 9) Endonuclease/Exonuclease/phosphatase family; 10) Eukaryotic-type carbonic anhydrase; 11) Galactose oxidase, central domain; 12) Legume lectin domain; 13) Leucine Rich repeat; 14) Leucine rich repeat; 15) Leucine rich repeat N-terminal domain; 16) Peroxidase; 17) PPR repeat; 18) Protein kinase domain; 19) Protein phosphatase 2C; 20) Protein tyrosine kinase; 21) P-type ATPase; 22) Receptor family ligand binding region; 23) Ring finger domain; 24) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 25) STAS domain; 26) Terpene synthase family, metal binding domain; 27) Terpene synthase, N-terminal domain; 28) Ubiquitin family; 29) UDP-glucoronosyl and UDP-glucosyl transferase

reproduction	5.09e-07	1) ABC transporter; 2) AP2 domain; 3) Argonaute linker 1 domain; 4) Argonaute linker 2 domain; 5) ATPase family associated with various cellular activities (AAA); 6) BTB/POZ domain; 7) CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B; 8) Chromo (CHRromatin Organisation MOdifier) domain; 9) Cyclin, C-terminal domain; 10) Cyclin, N-terminal domain; 11) FKBP-type peptidyl-prolyl cis-trans isomerase; 12) Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase; 12) Leucine Rich Repeat; 13) Leucine rich repeat; 14) Leucine rich repeat N-terminal domain; 15) Matrixin; 16) Myb/SANT-like DNA-binding domain; 17) N-terminal domain of argonaute; 18) PAZ domain; 19) Piwi domain; 20) PPR repeat; 21) PPR repeat family; 22) Protein kinase domain; 23) Protein tyrosine kinase; 24) Putative peptidoglycan binding domain; 25) Replication factor-A C terminal domain; 26) Replication protein A OB domain; 27) Ring finger domain; 28) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 29) STAS domain; 29) Triose-phosphate Transporter family; 30) Ubiquitin-conjugating enzyme; 31) Ubiquitin family; 32) Zinc finger, C3HC4 type (RING finger)

Table A.11 Top significant GO biological process terms of expanded gene families in *P. sitchensis*.

GO term	*P* corr	PFAM description
response to endogenous stimulus	2.00e-09	1) Endonuclease/Exonuclease/phosphatase family; 2) Gibberellin regulated protein; 3) Homeobox associated leucine zipper; 4) Homeobox domain; 5) Hpt domain; 6) Leucine rich repeat; 7) Leucine rich repeats (6 copies); 8) Phosphatidylinositol 3- and 4-kinase; 9) PPR repeat; 10) Protein kinase domain; 11) Protein tyrosine kinase; 12) Ras family; 13) Ring finger domain; 14) Salt stress response/antifungal; 15) short chain dehydrogenase; 16) TIR domain; 17) Ubiquitin family; 18) UDP-glucoronosyl and UDP-glucosyl transferase
response to organic substance	1.18e-07	1) Endonuclease/Exonuclease/phosphatase family; 2) Gibberellin regulated protein; 3) Homeobox associated leucine zipper; 4) Homeobox domain; 5) Hpt domain; 6) Hsp70 protein; 7) Leucine rich repeat; 8) Leucine rich repeats (6 copies); 9) Oxidoreductase NAD-binding domain; 10) Phosphatidylinositol 3- and 4-kinase; 11) PPR repeat; 12) Protein kinase domain; 13) Protein tyrosine kinase; 14) Ras family; 15) Ring finger domain; 16) Salt stress response/antifungal; 17) short chain dehydrogenase; 18) TIR domain; 19) Ubiquitin family; 20) UDP-glucoronosyl and UDP-glucosyl transferase; 21) Zinc knuckle
response to abiotic stimulus	4.21e-07	1) Aldehyde dehydrogenase family; 2) ATP synthase alpha/beta family, beta-barrel domain; 3) ATP synthase alpha/beta family, nucleotide-binding domain; 4) Endonuclease/Exonuclease/phosphatase family; 5) Homeobox associated leucine zipper; 6) Homeobox domain; 7) Hsp70 protein; 8) Leucine rich repeat; 9) Oxidoreductase NAD-binding domain; 10) Phosphatidylinositol 3- and 4-kinase; 11) Protein kinase domain; 12) Protein tyrosine kinase; 13) Ras family; 14) Ring finger domain; 15) TIR domain; 16) Ubiquitin family; 17) Zinc knuckle
multi-organism process	9.90e-07	1) 1,3-beta-glucan synthase component; 2) Core histone H2A/H2B/H3/H4; 3) Endonuclease/Exonuclease/phosphatase family; 4) Hsp70 protein; 5) Leucine Rich Repeat; 6) Leucine rich repeat; 7) Leucine rich repeats (6 copies); 8) NB-ARC domain; 9) Oxidoreductase NAD-binding domain; 10) Phosphatidylinositol 3- and 4-kinase; 11) Protein kinase domain; 12) Protein tyrosine kinase; 13) Ras family; 14) Reticulon; 15) Salt stress response/antifungal; 16) short chain dehydrogenase; 17) TIR domain; 18) Ubiquitin family; 19) UDP-glucoronosyl and UDP-glucosyl transferase
response to external stimulus	7.95e-06	1) Core histone H2A/H2B/H3/H4; 2) Homeobox domain; 3) Hpt domain; 4) Hsp70 protein; 5) Leucine Rich Repeat; 6) Leucine rich repeat; 7) NB-ARC domain; 8) Oxidoreductase NAD-binding domain; 9) Phosphatidylinositol 3- and 4-kinase; 10) Protein kinase domain; 11) Protein tyrosine kinase; 12) Ras family; 13) Reticulon; 14) Salt stress response/antifungal; 15) TIR domain; 16) Ubiquitin family; 17) UDP-glucoronosyl and UDP-glucosyl transferase; 18) Zinc knuckle

Table A.12 Top significant GO biological process terms of expanded gene families in *P. glauca.*

GO term	*P* corr	PFAM description
Reproduction	2.34e-09	1) Homeobox domain; 2) Protein kinase domain; 3) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 4) TCP-1/cpn60 chaperonin family; 5) Core histone H2A/H2B/H3/H4; 6) Ubiquitin-conjugating enzyme; 7) Cupin; 8) Ubiquitin family; 9) DEAD/DEAH box helicase; 10) Helicase conserved C-terminal domain; 11) Bromodomain; 12) Armadillo/beta-catenin-like repeat; 13) Leucine Rich Repeat; 14) ABC transporter transmembrane region; 15) AP2 domain; 16) Cullin family; 17) PPR repeat; 18) Elongation factor Tu C-terminal domain; 19) CPSF A subunit region; 20) Leucine rich repeat N-terminal domain; 21) Replication factor-A C terminal domain; 22) PPR repeat family; 23) Leucine rich repeat; 24) C2H2-type zinc finger; 25) Replication protein A OB domain; 26) Elongation factor Tu domain 2; 27) Cullin; 28) Atypical arm repeat; 29) Serine-threonine protein phosphatase N-terminal domain
response to stimulus	5.27e-05	1) ABC transporter; 2) ATP synthase alpha/beta family, nucleotide-binding domain; 3) Hsp20/alpha crystallin family; 4) Homeobox domain; 5) Cytochrome P450; 6) Protein kinase domain; 7) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 8) TCP-1/cpn60 chaperonin family; 9) Aldehyde dehydrogenase family; 10) Ubiquitin family; 11) DEAD/DEAH box helicase; 12) Helicase conserved C-terminal domain; 13) Bromodomain; 14) ABC transporter transmembrane region; 15) AP2 domain; 16) OB-fold nucleic acid binding domain; 17) Homeobox associated leucine zipper; 18) ATP synthase alpha/beta family, beta-barrel domain; 19) Endonuclease/Exonuclease/phosphatase family; 20) NAF domain; 21) Leucine rich repeat
response to endogenous stimulus	7.59e-05	1) ABC transporter; 2) Homeobox domain; 3) Protein kinase domain; 4) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 5) Cupin; 6) Ubiquitin family; 7) DEAD/DEAH box helicase; 8) Helicase conserved C-terminal domain; 9) Armadillo/beta-catenin-like repeat; 10) ABC transporter transmembrane region; 11) AP2 domain; 12) Transmembrane amino acid transporter protein; 13) PPR repeat; 14) Homeobox associated leucine zipper; 15) Elongation factor Tu C-terminal domain; 16) Endonuclease/Exonuclease/phosphatase family; 17) Leucine rich repeat N-terminal domain; 18) Leucine rich repeat; 19) Elongation factor Tu domain 2
macromolecule localization	7.59e-05	1) ABC transporter; 2) Hsp20/alpha crystallin family; 3) Homeobox domain; 4) Protein kinase domain RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 5) TCP-1/cpn60 chaperonin family; 6) Core histone H2A/H2B/H3/H4; 7) Ubiquitin family; 8) DEAD/DEAH box helicase; 9) Helicase conserved C-terminal domain; 10) Armadillo/beta-catenin-like repeat; 11) ABC transporter transmembrane region; 12) OB-fold nucleic acid binding domain; 13) DJ-1/PfpI family; 14) PA domain MIF4G domain; 15) Endonuclease/Exonuclease/phosphatase family; 16) Replication factor-A protein 1, N-terminal domain; 17) Replication factor-A C terminal domain; 18) Leucine rich repeat; 19) C-terminus of histone H2A; 20) Replication protein A OB domain
multi-organism process	7.59e-05	1) ABC transporter; 2) Cytochrome P450; 3) Protein kinase domain; 4) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 5) TCP-1/cpn60 chaperonin family; 6) Core histone H2A/H2B/H3/H4; 7) Ubiquitin-conjugating enzyme; 8) Ubiquitin family; 9) DEAD/DEAH box helicase; 10) Helicase conserved C-terminal domain; 11) Thaumatin family; 12) Bromodomain; 13) Armadillo/beta-catenin-like repeat; 14) Leucine Rich Repeat; 15) ABC transporter transmembrane region; 16) AP2 domain; 17) Cullin family; 18) Late embryogenesis abundant protein; 19) Endonuclease/Exonuclease/phosphatase family; 20) Leucine rich repeat N-terminal domain; 21) Leucine rich repeat; 22) C-terminus of histone H2A

Table A.13 Top significant GO molecular function terms of expanded gene families in interior spruce.

GO term	*P* corr	PFAM description
response to endogenous stimulus	6.48e-06	1) ABC transporter; 2) Ras family; 3) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 4) Sugar (and other) transporter; 5) Cyclin, N-terminal domain; 6) C2 domain; 7) Cupin; 8) Lipoxygenase; 9) Phosphatidylinositol 3- and 4-kinase; 10) Armadillo/beta-catenin-like repeat; 11) alpha/beta hydrolase fold; 12) BTB/POZ domain; 13) AP2 domain; 14) TIR domain; 15) Cyclin; C-terminal domain; 16) Endonuclease/Exonuclease/phosphatase family; 17) GRAS domain family; 18) Protein tyrosine kinase; 19) Leucine rich repeat N-terminal domain; 20) Leucine rich repeats (6 copies); 21) Leucine Rich repeat; 22) EF-hand domain pair; 23) Leucine rich repeat
response to abiotic stimulus	6.48e-06	1) ABC transporter; 2) Hsp70 protein; 3) KH domain; 4) Ras family; 5) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 6) Sugar (and other) transporter; 7) Zinc knuckle; 8) Cyclin, N-terminal domain; 9) C2 domain; 10) Oxidoreductase NAD-binding domain; 11) Ubiquitin-conjugating enzyme; 12) Cupin; 13) Lipoxygenase; 14) Phosphofructokinase; 15) Phosphatidylinositol 3- and 4-kinase; 16) Armadillo/beta-catenin-like repeat; 17) alpha/beta hydrolase fold; 18) BTB/POZ domain; 19) AP2 domain; 20) SET domain; 21) TIR domain; 22) DJ-1/PfpI family; 23) Cyclin, C-terminal domain; 24) Endonuclease/Exonuclease/phosphatase family; 25) GRAS domain family; 26) Protein tyrosine kinase; 27) Leucine rich repeat N-terminal domain; 28) Leucine rich repeats (6 copies); 29) Leucine Rich repeat; 30) EF-hand domain pair; 31) Leucine rich repeat
response to abiotic stimulus	7.39e-06	1) ABC transporter; 2) Hsp70 protein; 3) KH domain; 4) Ras family; 5) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 6) Sugar (and other) transporter; 7)Zinc knuckle; 8) Cyclin, N-terminal domain; 9) Calcineurin-like phosphoesterase; 10) Oxidoreductase NAD-binding domain; 11) SNF2 family N-terminal domain; 12) Phosphatidylinositol 3- and 4-kinase; 13) BTB/POZ domain; 14) AP2 domain; 15) IBR domain, a half RING-finger domain; 16) Phosphatidylinositol-4-phosphate 5-Kinase; 17) TIR domain; 18) Cyclin, C-terminal domain; 19) Endonuclease/Exonuclease/phosphatase family; 20) Cellulose synthase; 21) Protein tyrosine kinase; 22) AAA domain (Cdc48 subfamily); 23) FAE1/Type III polyketide synthase-like protein; 24) EF-hand domain pair; 25) Leucine rich repeat
reproduction	7.39e-06	1) ABC transporter; 2) KH domain; 3) Ras family; 4) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 6) Core histone H2A/H2B/H3/H4; 7) Cyclin, N-terminal domain; 8) C2 domain; 9) Cytochrome b5-like Heme/Steroid binding domain; 10) SNF2 family N-terminal domain; 11) Ubiquitin-conjugating enzyme; 12) Cupin; 13) Chromo (CHRromatin Organisation MOdifier) domain; 14) Phosphatidylinositol 3- and 4-kinase; 15) Armadillo/beta-catenin-like repeat; 16) Leucine Rich Repeat; 17) BTB/POZ domain; 18) AP2 domain; 19) SET domain; 20) CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B; 21) Cyclin, C-terminal domain; 22) Triose-phosphate Transporter family; 23) Protein tyrosine kinase; 24) Leucine rich repeat N-terminal domain; 25) PPR repeat family; 26) Leucine rich repeats (6 copies); 27) Leucine rich repeat
multi-organism process	9.03e-06	1) ABC transporter; 2) Hsp70 protein; 3) KH domain; 4) Ras family; 5) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 6) Core histone H2A/H2B/H3/H4; 7) Cyclin, N-terminal domain; 8) C2 domain; 9) Oxidoreductase NAD-binding domain; 10) SNF2 family N-terminal domain; 11) Ubiquitin-conjugating enzyme; 12) Lipoxygenase; 13) Chromo (CHRromatin Organisation MOdifier) domain; 14) Phosphatidylinositol 3- and 4-kinase; 15) Armadillo/beta-catenin-like repeat; 16) Leucine Rich Repeat; 17) F-box domain; 18) BTB/POZ domain; 19) AP2 domain; 20) POT family; 21) NB-ARC domain; 22) TIR domain; 23) Cyclin, C-terminal domain; 24) Endonuclease/Exonuclease/phosphatase family; 25) Protein tyrosine kinase; 26) Leucine rich repeat N-terminal domain; 27) Leucine rich repeats (6 copies); 28) Leucine Rich repeat; 29) Leucine rich repeat

Table A.14 Top significant GO molecular function terms of expanded gene families in *P. engelmannii.*

GO term	*P* corr	PFAM description
regulatory RNA binding	2.50e-06	1) Argonaute linker 1 domain; 2) Argonaute linker 2 domain; 3) Leucine Rich repeat; 4) Leucine rich repeat; 5) N-terminal domain of argonaute; 6) PAZ domain; 7) Piwi domain; 8) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 9) Zinc knuckle
single-stranded RNA binding	7.71e-06	1) Argonaute linker 1 domain; 2) Argonaute linker 2 domain; 3) Leucine Rich repeat; 4) Leucine rich repeat; 5) N-terminal domain of argonaute; 6) PAZ domain; 7) Piwi domain; 8) Zinc knuckle
ubiquitin-protein transferase activity	1.69e-04	1) BTB and C-terminal Kelch; 2) BTB/POZ domain; 3) F-box domain; 4) Kelch motif; 5) Leucine Rich repeat; 6) PA domain; 7) Ring finger domain; 8) Skp1 family, dimerisation domain; 9) Skp1 family, tetramerisation domain; 10) Ubiquitin-conjugating enzyme; 11) Zinc finger, C3HC4 type (RING finger)
catalytic activity, acting on RNA	5.77e-04	1) Argonaute linker 1 domain; 2) Argonaute linker 2 domain; 3) DNA-dependent RNA polymerase; 4) DNA-directed RNA polymerase N-terminal; 5) Endonuclease/Exonuclease/phosphatase family; 6) GAG-pre-integrase domain; 7) Leucine rich repeat; 8) N-terminal domain of argonaute; 9) PAZ domain; 10) Piwi domain; 11) RNA polymerase Rpb2, domain 3; 12) RNA polymerase Rpb2, domain 6; 13) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 14) Zinc knuckle; 15) RNA polymerase Rpb1, domain 2; 16) RNA polymerase Rpb2, domain 7; 17) RNA polymerase Rpb2, domain 2; 18) RNA polymerase Rpb1, domain 1
protein-containing complex binding	7.09e-03	1) ABC transporter; 2) Chromo (CHRromatin Organisation MOdifier) domain; 3) Leucine rich repeat; 4) Matrixin; 5) Pentatricopeptide repeat domain; 6) Peptidase inhibitor I9; 7) Protein kinase domain; 8) Protein tyrosine kinase; 9) Putative peptidoglycan binding domain; 10) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 11) Subtilase family; 12) Ubiquitin family; 13) Zinc finger, C3HC4 type (RING finger); 14) Zinc knuckle

Table A.15 Top significant GO molecular function terms of expanded gene families in *P. sitchensis*.

GO term	*P* corr	PFAM description
small molecule binding	1.83e-06	1) AIG1 family; 2) Aldehyde dehydrogenase family; 3) ATP synthase alpha/beta family, beta-barrel domain; 4) ATP synthase alpha/beta family, nucleotide-binding domain; 5) Beta-eliminating lyase; 6) Carboxyl transferase domain; 7) haloacid dehalogenase-like hydrolase; 8) Hsp70 protein; 9) NB-ARC domain; 10) Oxidoreductase NAD-binding domain; 11) Protein kinase domain; 12) Protein tyrosine kinase; 13) Ras family; 14) short chain dehydrogenase; 15) UDP-glucoronosyl and UDP-glucosyl transferase
nucleoside phosphate binding	2.44e-06	1) ATP synthase alpha/beta family, nucleotide-binding domain; 2) Hsp70 protein; 3) Protein kinase domain; 4) Ras family; 5) short chain dehydrogenase; 6) Aldehyde dehydrogenase family; 7) Oxidoreductase NAD-binding domain; 8) haloacid dehalogenase-like hydrolase; 9) NB-ARC domain; 10) Carboxyl transferase domain; 11) ATP synthase alpha/beta family, beta-barrel domain; 12) AIG1 family; 13) Protein tyrosine kinase
molecular transducer activity	1.29e-05	1) Endonuclease/Exonuclease/phosphatase family; 2) Hpt domain' 3) Leucine Rich Repeat; 4) Leucine rich repeat; 5) Leucine rich repeats (6 copies); 6) Protein kinase domain; 7) Protein tyrosine kinase; 8) short chain dehydrogenase; 9) TIR domain; 10) Homeobox domain
carbohydrate derivative binding	1.65e-05	1) AIG1 family; 2) Aldehyde dehydrogenase family; 3) ATP synthase alpha/beta family, beta-barrel domain; 4) ATP synthase alpha/beta family, nucleotide-binding domain; 5) Carboxyl transferase domain; 6) haloacid dehalogenase-like hydrolase; 7) Hsp70 protein; 8) Leucine rich repeat; 9) Oxidoreductase NAD-binding domain; 10) Protein kinase domain; 11) Protein tyrosine kinase; 12) Ras family
ion binding	2.91e-04	1) AIG1 family; 2) Aldehyde dehydrogenase family; 3) ATP synthase alpha/beta family, beta-barrel domain; 4) ATP synthase alpha/beta family, nucleotide-binding domain; 5) Beta-eliminating lyase; 6) Carboxyl transferase domain; 7) Endonuclease/Exonuclease/phosphatase family; 8) haloacid dehalogenase-like hydrolase; 9) Hsp70 protein; 10) Leucine rich repeat; 11) Oxidoreductase NAD-binding domain; 12) Protein kinase domain; 13) Protein tyrosine kinase; 14) Ras family; 15) Ribulose-phosphate 3 epimerase family; 16) short chain dehydrogenase; 17) UDP-glucoronosyl and UDP-glucosyl transferase

Table A.16 Top significant GO molecular function terms of expanded gene families in *P. glauca.*

GO term	*P* corr	PFAM description
single-stranded DNA binding	5.38e-05	1) DEAD/DEAH box helicase; 2) Endonuclease/Exonuclease/phosphatase family; 3) OB-fold nucleic acid binding domain; 4) Replication factor-A C terminal domain; 5) Replication factor-A protein 1, N-terminal domain; 6) Replication protein A OB domain; 7) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
damaged DNA binding	6.22e-05	1) Bromodomain; 2) Endonuclease/Exonuclease/phosphatase family; 3) OB-fold nucleic acid binding domain; 4) Replication factor-A C terminal domain; 5) Replication factor-A protein 1, N-terminal domain; 6) Replication protein A OB domain
ribonucleoprotein complex binding	6.22e-05	1) ABC transporter; 2) DEAD/DEAH box helicase; 3) Elongation factor Tu domain 2; 4) Helicase conserved C-terminal domain; 5) Pentatricopeptide repeat domain; 6) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 7) Ubiquitin family
protein-containing complex binding	7.18e-05	1) ABC transporter; 2) Core histone H2A/H2B/H3/H4; 3) DEAD/DEAH box helicase; 4) Elongation factor Tu domain 2; 5) GRAM domain; 6) Helicase conserved C-terminal domain; 7) Leucine rich repeat; 8) Leucine-rich repeat; 9) Pentatricopeptide repeat domain; 10) Peptidase inhibitor I9; 11) Protein kinase domain; 12) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 13) Subtilase family; 14) Ubiquitin family
nucleic acid binding	8.82e-05	1) ABC transporter; 2) AP2 domain; 3) Bromodomain; 4) C2H2-type zinc finger; 5) Core histone H2A/H2B/H3/H4; 6) CPSF A subunit region; 7) C-terminus of histone H2A; 8) DEAD/DEAH box helicase; 9) DJ-1/PfpI family; 10) Domain of unknown function (DUF4217); 11) Elongation factor Tu C-terminal domain; 12) Elongation factor Tu domain 2; 13) Endonuclease/Exonuclease/phosphatase family; 14) Helicase conserved C-terminal domain; 15) Homeobox associated leucine zipper; 16) Homeobox domain; 17) Leucine Rich Repeat; 18) Leucine rich repeat; 19) MIF4G domain; 20) OB-fold nucleic acid binding domain; 21) Pentatricopeptide repeat domain; 22) PPR repeat family; 23) Replication factor-A C terminal domain; 24) Replication factor-A protein 1, N-terminal domain; 25) Replication factor C C-terminal domain; 26) Replication protein A OB domain; 27) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 28) Ubiquitin family; 29) Poly-adenylate binding protein, unique domain

Table A.17 Top significant GO biological process terms of expanded gene families in interior spruce.

GO term	*P* corr	PFAM description
small molecule binding	7.21e-04	1) AAA domain (Cdc48 subfamily); 2) ABC transporter; 3) ADP-ribosylation factor family; 4) C2 domain; 5) Carbamoyl-phosphate synthase L chain, ATP binding domain; 6) Chromo (CHRromatin Organisation MOdifier) domain; 7) Hsp70 protein; 8) Leucine Rich repeat; 9) NAD dependent epimerase/dehydratase family; 10) NB-ARC domain; 11) Oxidoreductase NAD-binding domain; 12) Phosphofructokinase; 13) Protein tyrosine kinase; 14) Ras family; 15) SNF2 family N-terminal domain SRP54-type protein, GTPase domain; 16) Sugar (and other) transporter; 17) Transketolase, pyrimidine binding domain; 18) MGS-like domain; 19) SRP54-type protein, helical bundle domain; 20) Signal peptide binding domain
nucleoside phosphate binding	9.79e-04	1) AAA domain (Cdc48 subfamily); 2) ABC transporter; 3) ADP-ribosylation factor family; 4) Carbamoyl-phosphate synthase L chain, ATP binding domain; 5) Chromo (CHRromatin Organisation MOdifier) domain; 6) Hsp70 protein; 7) MGS-like domain; 8) NAD dependent epimerase/dehydratase family; 9) NB-ARC domain; 10) Oxidoreductase NAD-binding domain; 11) Phosphofructokinase; 12) Protein tyrosine kinase; 13) Ras family; 14) Signal peptide binding domain; 15) SNF2 family N-terminal domain; 16) SRP54-type protein, GTPase domain; 17) SRP54-type protein, helical bundle domain
carbohydrate derivative binding	9.79e-04	1) AAA domain (Cdc48 subfamily); 2) ABC transporter; 3) ADP-ribosylation factor family; 4) Calcineurin-like phosphoesterase; 5) Carbamoyl-phosphate synthase L chain, ATP binding domain; 6) Chromo (CHRromatin Organisation MOdifier) domain; 7) Hsp70 protein; 8) Leucine rich repeat; 9) MGS-like domain; 10) Oxidoreductase NAD-binding domain; 11) Phosphofructokinase; 12) Protein tyrosine kinase; 13) Ras family; 14) Signal peptide binding domain; 15) SNF2 family N-terminal domain; 16) SRP54-type protein, GTPase domain; 17) SRP54-type protein, helical bundle domain
protein-containing complex binding	3.73e-03	1) ABC transporter; 2) Chromo (CHRromatin Organisation MOdifier) domain; 3) Core histone H2A/H2B/H3/H4; 4) KH domain; 5) Leucine rich repeat; 6) Phosphatidylinositol 3- and 4-kinase; 7) Protein tyrosine kinase; 8) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 9) SNF2 family N-terminal domain; 10) SRP54-type protein, GTPase domain; 11) SRP54-type protein, helical bundle domain; 12) Zinc knuckle
enzyme binding	6.71e-03	1) C2 domain; 2) Calcineurin-like phosphoesterase; 3) Core histone H2A/H2B/H3/H4; 4) Cyclin, C-terminal domain; 5) Cyclin, N-terminal domain; 6) Eukaryotic initiation factor 4E; 7) Hsp70 protein; 8) IBR domain, a half RING-finger domain; 9) Leucine Rich repeat; 10) Leucine rich repeat; 11) Phosphatidylinositol 3- and 4-kinase; 12) Protein tyrosine kinase; 13) Ras family; 14) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 15) SET domain; 16) Ubiquitin-conjugating enzyme

Table A.18 Top significant GO biological process terms of genes under positive selection in *P. engelmannii.*

GO term	*P* corr	PFAM description
response to endogenous stimulus	6.14E-07	1) Aldo/keto reductase family; 2) AP2 domain; 3) AUX/IAA family; 4) bZIP transcription factor; 5) E1-E2 ATPase; 6) Glycosyl hydrolase family 1; 7) Helix-loop-helix DNA-binding domain; 8) Homeobox domain' 9) Leucine Rich repeat; 10) PPR repeat; 11) Protein kinase domain; 12) Protein phosphatase 2C; 13) Protein tyrosine kinase; 14) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 15) Zinc finger, C2H2 type; 16) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
reproduction	2.18E-06	1))AP2 domain; 2) AUX/IAA family; 3) bZIP transcription factor; 4) CPSF A subunit region; 5) CRAL/TRIO domain; 6) ERCC4 domain; 7) Formin Homology 2 Domain; 8) Helix-loop-helix DNA-binding domain; 9) Homeobox domain; 10) IQ calmodulin-binding motif; 11) PPR repeat; 12) Protein kinase domain; 13) Protein tyrosine kinase; 14) Ribonuclease III domain; 15) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 16) WD domain, G-beta repeat; 17) Zinc finger, C2H2 type
response to abiotic stimulus	1.58E-05	1) AP2 domain; 2) AUX/IAA family; 3) bZIP transcription factor; 4) E1-E2 ATPase; 5) Glycosyl hydrolase family 1; 6) Helix-loop-helix DNA-binding domain; 7) Homeobox domain; 8) Protein kinase domain; 9) Protein phosphatase 2C; 10) Protein tyrosine kinase; 11) PWWP domain; 12) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 13) SAP domain; 14) WD domain, G-beta repeat; 15) Zinc finger, C2H2 type
response to stress	1.65E-05	1) Aldo/keto reductase family; 2) AP2 domain; 3) bZIP transcription factor; 4) CPSF A subunit region; 5) E1-E2 ATPase; 6) ERCC4 domain; 7) Formin Homology 2 Domain; 8) Glutaredoxin; 9) Glycosyl hydrolase family 1; 10) Glycosyltransferase family 8; 11) Helix-loop-helix DNA-binding domain; 12) Leucine Rich repeat; 13) Multicopper oxidase; 14) Protein kinase domain; 15) Protein phosphatase 2C; 16) Protein tyrosine kinase; 17) PWWP domain; 18) Ribonuclease III domain; 19) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 20) SAP domain; 21) WD domain, G-beta repeat; 22) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
response to organic substance	3.50E-05	1) Aldo/keto reductase family; 2) AP2 domain; 3) AUX/IAA family; 4) bZIP transcription factor; 5) E1-E2 ATPase; 6) Glycosyl hydrolase family 1; 7) Helix-loop-helix DNA-binding domain; 8) Homeobox domain; 9) Leucine Rich repeat; 10) PPR repeat; 11) Protein kinase domain; 12) Protein phosphatase 2C; 13) Protein tyrosine kinase; 14) Ribonuclease III domain; 15) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 16) Zinc finger, C2H2 type; 17) Zinc finger C-x8-C-x5-C-x3-H type (and similar)

Table A.19 Top significant GO biological process terms of genes under positive selection in *P. sitchensis.*

GO term	*P* corr	PFAM description
response to endogenous stimulus	3.50e-08	1) 2OG-Fe(II) oxygenase superfamily; 2) alpha/beta hydrolase fold; 3) Ankyrin repeats (3 copies); 4) AP2 domain; 5) E1-E2 ATPase; 6) Elongation factor Tu GTP binding domain; 7) Helix-loop-helix DNA-binding domain; 8) HMG (high mobility group) box; 9) Leucine Rich repeat; 10) Leucine rich repeat; 11) Myb-like DNA-binding domain; 12) PB1 domain; 13) PPR repeat; 14) Protein tyrosine kinase; 15) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 16) Zinc finger, C2H2 type; 17) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
embryo development	3.09e-07	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) C2H2-type zinc finger; 4) E1-E2 ATPase; 5) GDP-fucose protein O-fucosyltransferase; 6) Helix-loop-helix DNA-binding domain; 7) HMG (high mobility group) box; 8) Leucine rich repeat; 9) PB1 domain; 10) PPR repeat; 11) Protein tyrosine kinase; 12) Putative peptidoglycan binding domain; 13) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 14) WD domain, G-beta repeat; 15) Zinc finger, C2H2 type
regulation of biological quality	1.17e-06	1) 2OG-Fe(II) oxygenase superfamily; 2) alpha/beta hydrolase fold; 3) Ankyrin repeats (3 copies); 4) Cation efflux family; 5) Dual specificity phosphatase, catalytic domain; 6) E1-E2 ATPase; 7) Helix-loop-helix DNA-binding domain; 8) HMG (high mobility group) box; 9) Leucine Rich repeat; 10) Leucine rich repeat; 11) MSP (Major sperm protein) domain; 12) Multicopper oxidase; 13) Myb-like DNA-binding domain; 14) PB1 domain; 15) Protein tyrosine kinase; 16) Pumilio-family RNA binding repeat; 17) Putative peptidoglycan binding domain; 18) Retinal pigment epithelial membrane protein; 19) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 20) Sec7 domain; 21) WD domain, G-beta repeat; 22) Zinc finger, C2H2 type; 23) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
reproduction	2.97e-06	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) C2H2-type zinc finger; 4) Dual specificity phosphatase, catalytic domain; 5) Elongation factor Tu GTP binding domain; 6) Helix-loop-helix DNA-binding domain; 7) HMG (high mobility group) box; 8) Leucine rich repeat; 9) Myb-like DNA-binding domain; 10) PPR repeat; 11) Protein tyrosine kinase; 12) Pumilio-family RNA binding repeat; 13) Putative peptidoglycan binding domain; 14) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 15) WD domain, G-beta repeat; 16) Zinc finger, C2H2 type
response to abiotic stimulus	7.87e-06	1) 2OG-Fe(II) oxygenase superfamily; 2) Ankyrin repeats (3 copies); 3) AP2 domain; 4) E1-E2 ATPase; 5) Helix-loop-helix DNA-binding domain; 6) HMG (high mobility group) box; 7) Leucine rich repeat; 8) Myb-like DNA-binding domain; 9) NLI interacting factor-like phosphatase; 10) Piezo non-specific cation channel, R-Ras-binding domain; 11) Protein tyrosine kinase; 12) Putative peptidoglycan binding domain; 13) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 14) WD domain, G-beta repeat; 15) Zinc finger, C2H2 type

Table A.20 Top significant GO biological process terms of genes under positive selection in *P. glauca.*

GO term	*P* corr	PFAM description
response to abiotic stimulus	1.20E-07	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) AUX/IAA family; 4) Bromodomain; 5) Cyclin, N-terminal domain; 6) Cytochrome P450; 7) E1-E2 ATPase; 8) HMG (high mobility group) box; 9) Homeobox domain; 10) Protein kinase domain; 11) Protein phosphatase 2C; 12) Putative peptidoglycan binding domain; 13) PWWP domain; 14) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 15) SAP domain; 16) Thioredoxin; 17) Ubiquitin carboxyl-terminal hydrolase; 18) WD domain, G-beta repeat
regulation of biosynthetic process	5.94E-06	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) AUX/IAA family; 4) Bromodomain; 5) C2H2-type zinc finger; 6) Cyclin, N-terminal domain; 7) Cytochrome P450; 8) Elongation factor Tu GTP binding domain; 9) HMG (high mobility group) box; 10) Homeobox domain; 11) Kelch motif; 12) Myb-like DNA-binding domain; 13) Protein kinase domain; 14) Protein phosphatase 2C; 15) PWWP domain; 16) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 17) SAP domain; 18) SBP domain; 19) TCP family transcription factor; 20) Transcriptional repressor, ovate; 21) Ubiquitin carboxyl-terminal hydrolase; 22) WD domain, G-beta repeat
Reproduction	2.40E-05	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) AUX/IAA family; 4) Bromodomain; 5) C2H2-type zinc finger; 6) Cyclin, N-terminal domain; 7) Elongation factor Tu GTP binding domain; 8) HMG (high mobility group) box; 9) Homeobox domain; 10) Myb-like DNA-binding domain; 11) Protein kinase domain; 12) Putative peptidoglycan binding domain; 13) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 14) SBP domain; 15) WD domain, G-beta repeat
response to endogenous stimulus	2.41E-05	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) AUX/IAA family; 4) Cyclin, N-terminal domain; 5) E1-E2 ATPase; 6) Elongation factor Tu GTP binding domain; 7) HMG (high mobility group) box; 8) Homeobox domain; 9) Myb-like DNA-binding domain; 10) Protein kinase domain; 11) Protein phosphatase 2C; 12) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 13) Ubiquitin carboxyl-terminal hydrolase
response to organic substance	2.41E-05	1) Ankyrin repeats (3 copies); 2) AP2 domain; 3) AUX/IAA family; 4) C2H2-type zinc finger; 5) Cyclin, N-terminal domain; 6) Cytochrome P450; 7) E1-E2 ATPase; 8) Elongation factor Tu GTP binding domain; 9) Ferric reductase like transmembrane component; 10) HMG (high mobility group) box; 11) Homeobox domain; 12) Myb-like DNA-binding domain; 13) Protein kinase domain; 14) Protein phosphatase 2C; 15) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 16) Thioredoxin; 17) Ubiquitin carboxyl-terminal hydrolase

Table A.21 Top significant GO biological process terms of genes under positive selection in interior spruce.

GO term	*P* corr	PFAM description
reproduction	3.47E-04	1) ATPase family associated with various cellular activities (AAA); 2) AUX/IAA family; 3) Helicase conserved C-terminal domain; 4) Homeobox domain; 5) Initiation factor 2 subunit family; 6) IQ calmodulin-binding motif; 7) Leucine rich repeat; 8) No apical meristem (NAM) protein; 9) PPR repeat Pumilio-family RNA binding repeat; 10) SET domain; 11) Triose-phosphate Transporter family; 12) WD domain, G-beta repeat
response to abiotic stimulus	8.79E-04	1) ATPase family associated with various cellular activities (AAA); 2) AUX/IAA family; 3) Helicase conserved C-terminal domain; 4) Homeobox domain; 5) Initiation factor 2 subunit family; 6) Leucine rich repeat; 7) NLI interacting factor-like phosphatase; 8) No apical meristem (NAM) protein; 9) Piezo non-specific cation channel, R-Ras-binding domain; 10) Protein phosphatase 2C; 11) Sodium/calcium exchanger protein; 12) WD domain, G-beta repeat
response to endogenous stimulus	1.53E-03	1) AUX/IAA family; 2) Helicase conserved C-terminal domain; 3) Homeobox domain; 4) Initiation factor 2 subunit family; 5) Leucine Rich repeat; 6) Leucine rich repeat; 7) No apical meristem (NAM) protein; 8) PPR repeat; 9) Protein phosphatase 2C; 10) UDP-glucoronosyl and UDP-glucosyl transferase
negative regulation of metabolic process	2.34E-03	1) Bicoid-interacting protein 3 (Bin3); 2) Helicase conserved C-terminal domain; 3) Homeobox domain; 4) IQ calmodulin-binding motif; 5) Leucine Rich repeat; 6) Leucine rich repeat; 7) Protein phosphatase 2C; 8) Pumilio-family RNA binding repeat; 9) SET domain; 10) Sin3 binding region of histone deacetylase complex subunit SAP; 11) UDP-glucoronosyl and UDP-glucosyl transferase; 12) WD domain, G-beta repeat
embryo development	2.68E-03	1) FAD binding domain; 2) GDP-fucose protein O-fucosyltransferase; 3) Helicase conserved C-terminal domain; 4) Homeobox domain; 5) IQ calmodulin-binding motif; 6) Leucine rich repeat; 7) No apical meristem (NAM) protein; 8) PPR repeat; 9) WD domain, G-beta repeat

Table A.22 Top significant GO molecular function terms of genes under positive selection in *P. engelmannii.*

GO term	*P* corr	PFAM description
enzyme binding	5.64E-04	1) Aldo/keto reductase family; 2) Formin Homology 2 Domain; 3) Helix-loop-helix DNA-binding domain; 4) Importin-beta N-terminal domain; 5) IQ calmodulin-binding motif; 6) Leucine Rich repeat; 7) Protein kinase domain; 8) Protein phosphatase 2C; 9) Protein tyrosine kinase; 10) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 11) WD domain, G-beta repeat; 12) Zinc finger, C2H2 type Zinc finger C-x8-C-x5-C-x3-H type (and similar)
transcription regulator activity	2.65E-03	1) AP2 domain; 2) AUX/IAA family; 3) bZIP transcription factor; 4) Helix-loop-helix DNA-binding domain; 5) Homeobox domain; 6) Protein kinase domain; 7) PWWP domain; 8) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 9) SAP domain; 10) WD domain, G-beta repeat; 11) Zinc finger, C2H2 type; 12) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
calmodulin binding	9.23E-03	1) E1-E2 ATPase; 2) IQ calmodulin-binding motif; 3) Protein kinase domain; 4) Protein tyrosine kinase

Table A.23 Top significant GO molecular function terms of genes under positive selection in *P. sitchensis.*

GO term	*P* corr	PFAM description
enzyme binding	5.33e-04	1) Ankyrin repeats (3 copies); 2) Dual specificity phosphatase, catalytic domain; 3) Helix-loop-helix DNA-binding domain; 4) Leucine Rich repeat; 5) Leucine rich repeat; 6) PB1 domain; 7) Protein tyrosine kinase; 8) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 9) Sec7 domain; 10) WD domain, G-beta repeat; 11) Zinc finger, C2H2 type; 12) Zinc finger C-x8-C-x5-C-x3-H type (and similar)
protein serine/threonine/tyrosine kinase activity	5.33e-04	1) Ankyrin repeats (3 copies); 2) PB1 domain; 3) Protein tyrosine kinase
kinase binding	7.07e-03	1) Ankyrin repeats (3 copies); 2) Dual specificity phosphatase, catalytic domain; 3) PB1 domain; 4) Protein tyrosine kinase; 5) WD domain, G-beta repeat

Table A.24 Top significant GO molecular function terms of genes under positive selection in *P. glauca.*

GO term	*P* corr	PFAM description
transcription regulator activity	8.21E-07	1) AP2 domain; 2) AUX/IAA family; 3) Bromodomain; 4) C2H2-type zinc finger; 5) Cyclin, N-terminal domain; 6) HMG (high mobility group) box; 7) Homeobox domain; 8) Myb-like DNA-binding domain; 9) Protein kinase domain; 10) PWWP domain; 11) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 12) SAP domain; 13) SBP domain; 14) TCP family transcription factor; 15) Ubiquitin carboxyl-terminal hydrolase; 16) WD domain, G-beta repeat
kinase regulator activity	6.32E-04	1) Ankyrin repeats (3 copies); 2) Cyclin, N-terminal domain; 3) Homeobox domain; 4) Protein kinase domain; 5) WD domain, G-beta repeat
enzyme activator activity	1.54E-03	1) Cyclin, N-terminal domain; 2) Homeobox domain; 3) Protein kinase domain; 4) Putative peptidoglycan binding domain; 5) Thioredoxin; 6) WD domain, G-beta repeat
methyltransferase activity	1.54E-03	1) Ankyrin repeats (3 copies); 2) Bromodomain; 3) Pterin binding enzyme; 4) PWWP domain; 5) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 6) WD domain, G-beta repeat
nucleic acid binding	9.85E-03	1) AP2 domain; 2) AUX/IAA family; 3) Bromodomain; 4) C2H2-type zinc finger; 5) Cyclin, N-terminal domain; 6) Elongation factor Tu GTP binding domain; 7) HMG (high mobility group) box; 8) Homeobox domain; 9) Myb-like DNA-binding domain; 10) PWWP domain; 11) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain); 12) SAP domain; 13) SBP domain; 14) WD domain, G-beta repeat

Appendix B Chapter 3 - Supplemental material

Table B.1 Supporting evidence in MAKER protein-coding gene annotation in *Pissodes strobi*.

cDNA, transcriptome shotgun assembly (TSA), and RefSeq transcripts are used as transcript evidence, and proteins downloaded from SwissProt from the *Drosophila* are used as protein evidence.

Evidence	Taxon	Access IDs	Number of sequences
cDNA	*Pissodes strobi*	GT285068.1– GT296156.1, KC464331.1, U63328.1	11,091
cDNA	*Anoplophora glabripennis*	DQ067275.1, DQ067276.1, AY185203.1, KY039580.1, DR108748.1–DR109303.1, EF583868.1–EF583870.1, KX660670.1–KX890113.1, KY062564.1–KY287666.1,	596
cDNA	*Dendroctonus ponderosae*	BT126413.1–BT128693.1, EZ114957.1–EZ116155.1, GO484341.1–GO495894.1, GT316901.1– GT492003.1, JQ855638.1–JQ855707.1, KC113410.1, KC113439.1, KP736107.1– KP736166.1, KF444677.1	189,078
cDNA	*Tribolium castaneum*	AB360761.1–AB918727.1, AF017415.1–AF506022.1, AJ005083.1–AJ973445.1, AM269505.1–M922512.1, AY008296.1–AY887136.1, BK005734.1–BK008731.1, CB334789.1–CB337245.1, CF968201.1–CF968207.1, CO049327.1–CO049345.1, DN643532.1–DN652253.1, DQ138189.1–DQ855506.1, DR753940.1–DR753993.1, DT769880.1–DT805528.1, EB748715.1–EB754265.1, EC009091.1–EC011169.1, EF117815.1–EF688530.1, ES544600.1– ES554556.1, EU008544.1–EU937812.1, EX149741.1–EX149815.1, FJ158649.1–FJ917289.1, FN295953.1–FN824497.1, GQ202020.1–GQ368184.1, GU111762.1–GU727869.1, HM234671.1–HM622134.1, HQ110094.1–HQ824707.1, JF682841.1–JQ922422.1, JX099777.1–JX569831.1, KC161573.1–KC688266.1, KF192693.1–KF951599.1, KJ405472.1–KJ500311.1, KM216386.1–KM925014.1, KP120763.1–KP843191.1, KX553973.1–KX812753.1, KY368366.1–KY971527.1, LC154964.1–LC191269.1, LS991960.1–LS991974.1, LT908025.1–LT908033.1, MF467204.1–MF467212.1, MG011448.1–MG913606.1, MH664125.1–H664127.1, Z69735.1–Z69743.1, CN779602.1, DQ054783.1, DQ060238.1, BN001258.1, FM163173.1, HE608844.1, KT778599.1	65,269
TSA	*Pissodes strobi*	GAEO01000001.1–GAEO01004939.1	4,940
TSA	*Dendroctonus ponderosae*	GABX01000001.1–GABX01000059.1	60
TSA	*Dendroctonus ponderosae*	SRR1702878–SRR1703019	197,866
RefSeq transcripts	*Endopterygota*	sequence prefixes: XM, NM	1,389,102
SwissProt – proteins	*Drosophila*	DROME, DROPS, DROVI, DROYA, DROSI, DROER, DROSE, DROAN, DROPE, DROGR (topmost frequent species)	5,915

Table B.2 Curculionidae genomes and short reads accession IDs used in the comparative analysis.

The assemblies are used for phylogeny and repeat divergence. Short reads are used as input in RepeatExplorer and GenomeScope2.0. The genome size is reported as genome assembly reconstructed size in billions of bases.

Species	Assembly ID	Short reads SRA ID	Reconstructed size (Gb)
Pissodes strobi	GCA_016904865	SRR10590615	1.82
Elaeidobius kamerunicus	GCA_014849505	SRR12726955, SRR12726956, SRR12726957, SRR12726958,	0.205
Hypothenemus hampei	GCA_013372445	SRR11579639	0.147
Ips typographus	GCA_016097725	NA	0.236
Dendroctonus ponderosae	GCA_00035565	SRR546176, SRR546178, SRR546179, SRR546180, SRR546181, SRR546182, SRR546183, SRR546185, SRR546186, SRR54618, SRR546188, SRR546189, SRR546191	0.201
Pachyrhynchus sulpureomaculatus	GCA_019049505	SRR15032041	2.05
Listronotus bonariensis	GCA_014170235	SRR12134605, SRR12134606, SRR12134607, SRR12134608, SRR12134609, SRR12134610, SRR12134611, SRR12134612, SRR12134613, SRR12134614	1.1
Sitophilus oryzae	GCA_00293848	SRR6649886	0.757
Rhynchophorus ferrugineus	GCA_012979105	SRR8617827, SRR8645265, SRR8645265	0.741
Tribolium castaneum	GCA_000002335	SRR5992056	0.152

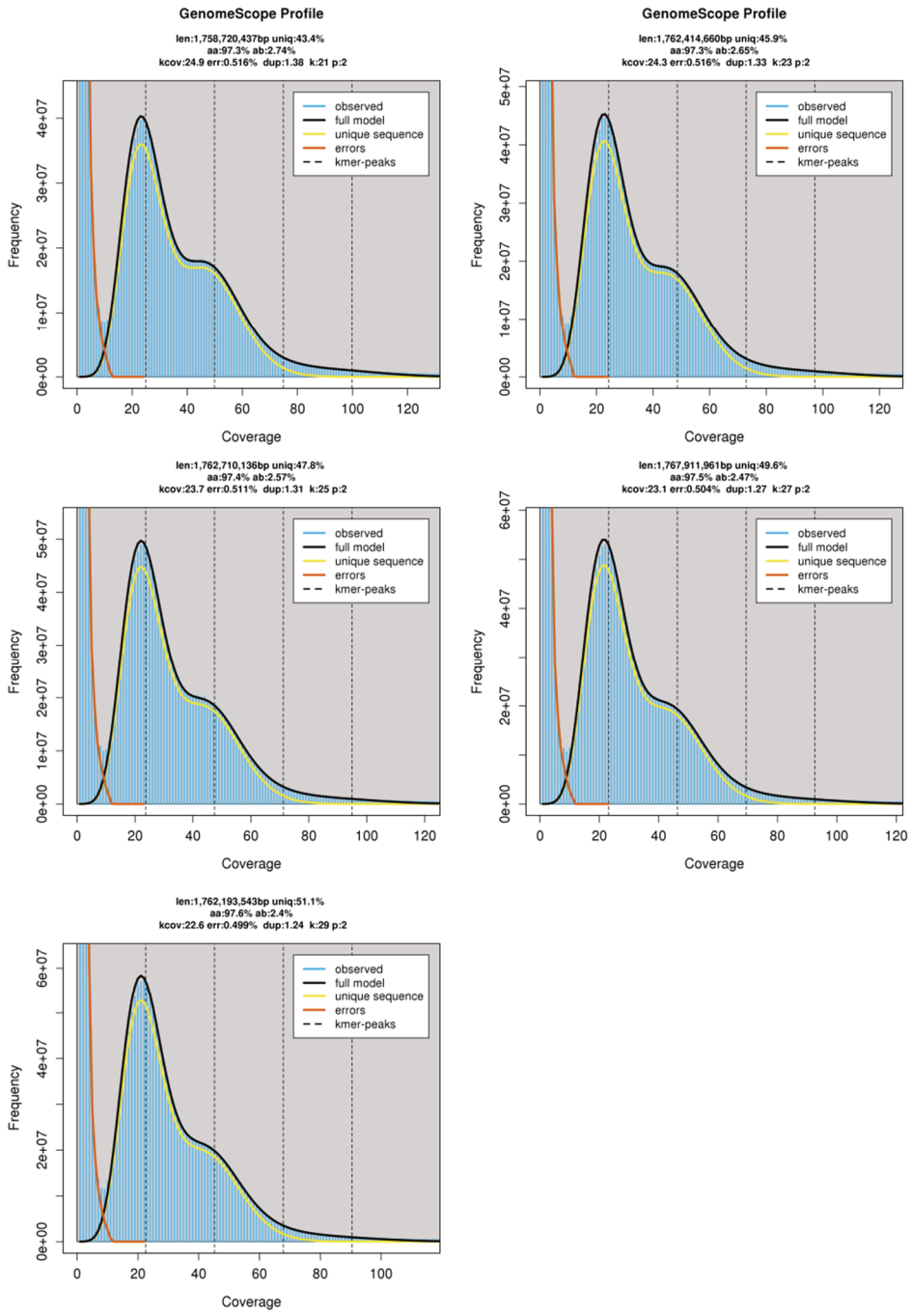

Figure B.1 GenomeScope2.0 estimates for *P. strobi* for k-mer sizes 21, 23, 25, 27, and 29 bp. The values calculated by the k-mers profiles are the genome length, % heterozygosity, and % unique sequences. Each plot shows the GenomeScope model fit on the *P. strobi* k-mer histogram. Genome estimates from the software are shown on the top of each histogram.

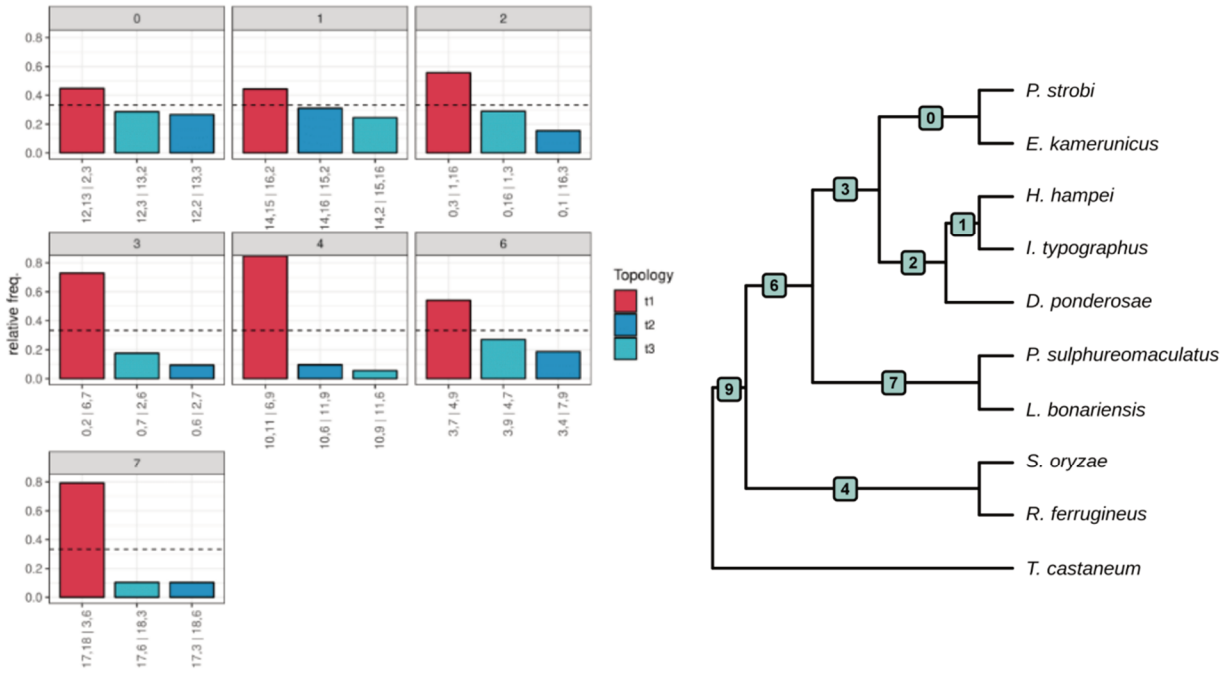

Figure B.2 DiscoVista quartet relative frequencies for Curculionidae phylogeny tree. Each panel in the figure on the left shows the relative frequency of quartets for the internal nodes. The bars display the frequency of the main selected quartet (t1 – red) and the alternative quartet topologies (t2 and t3 – blue). A tree topology is assigned when the relative frequency is more than 33% of all the possible topologies (dashed line in the figure). The panels on the left provide the quartet relative frequency for the numbered branches shown in the tree on the right.

Table B.3 Number of genes supporting each topology node in the Curculionidae species tree. The values are calculated with *phyparts*, showing the supporting genes from the species tree, as shown in the pie charts in Figure 3.2. The node number refers to the phylogeny Figure in Appendix Figure B.2 (i.e., **Node 0**: *P. strobi – E. kamerunicus*; **Node 1***: I. typographus – H. hampei*).

Node	Concordant	Discordant
0	187	446
1	692	1171
2	869	1106
3	961	814
4	307	71
6	1001	1044
7	1060	445
9 - Out	2009	0

Table B.4 Total percent and breakdown of genomic repeats, as annotated by EDTA repeat pipeline in Curculionidae genomes. Each class of repeats is divided into subclasses. Curculionidae species are presented as follows – 1) *Pissodes strobi* (highlighted), 2) *Elaeidobius kamerunicus,* 3) *Dendroctonus ponderosae*, 4) *Hypothenemus hampei,* 5) *Ips typographus,* 6) *Pachyrhynchus sulphureomaculatus, 7) Listronotus bonariensis, 8) Rhynchophorus ferrugineus, 9) Sitophilus oryzae.* "Other" repeats is the sum of the following repeat classes (a) **TIR (**Kolobok, Novosib, Merlin, PIF Harbinger, PiggyBac, Tc1 Mariner, Sola1, Sola2, MuDRMutator, hAT), (b) **DIR** and (c) **Helitron**.

Class	Subclass	1	2	3	4	5	6	7	8	9
DNA	DTA	15.24	0.00	1.09	0.31	3.52	10.08	11.52	1.77	17.39
	DTC	3.31	0.00	1.48	1.10	2.76	3.22	4.50	9.94	2.92
	DTH	1.06	0.00	0.41	0.34	0.38	0.67	2.37	1.37	3.08
	DTM	9.29	0.00	2.73	2.86	9.21	9.47	17.98	13.5	12.57
	DTT	0.41	0.00	0.20	0.02	0.47	0.40	0.41	0.41	0.70
	Helitron	2.40	0.00	1.64	0.46	0.71	4.03	4.35	3.75	4.74
LINE	LINE	1.31	1.06	0.62	0.21	1.24	1.47	0.79	0.01	0.86
LTR	Copia	1.05	0.00	0.00	0.03	0.76	0.00	0.23	0.22	0.20
	Gypsy	8.58	0.00	0.00	0.24	9.68	2.33	4.28	0.04	4.38
	Unknown	6.14	0.08	0.80	0.52	5.82	29.35	10.99	4.76	12.09
MITE	DTA	0.15	0.00	0.17	0.10	0.08	0.07	1.23	0.03	0.53
	DTC	0.08	0.00	0.04	0.04	0.03	0.03	0.18	0.03	0.10
	DTH	0.07	0.00	0.03	0.08	0.06	0.03	0.29	0.00	0.03
	DTM	0.55	0.00	0.14	0.50	0.60	0.18	0.84	0.12	0.54
	DTT	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.00	0.00
Other	Other	0.4	0.35	0.62	0.65	0.57	0.98	0.62	0.28	0.57
Unknown	Unknown	3.44	9.83	4.54	8.82	7.68	5.45	6.80	2.69	6.92
Mixture	Mixture	0.05	0.00	0.00	0.00	0.00	0.06	0.02	0.00	0.00
Total		**53.53**	**11.32**	**14.51**	**16.29**	**43.57**	**67.83**	**67.42**	**38.92**	**67.62**

Table B.5 RepeatExplorer comparative analysis results for Curculionidae species. Read pairs in each sample are subset at 0.3-fold coverage, based on the genome size obtained from the reconstructed genome assembly size. The reads are assembled in superclusters, and their count estimates their genome frequency. DNA and LTR clusters show the number of repeat clusters from class II and class I type of repeats, respectively. *P. strobi* is highlighted in the first line.

	Read pairs	Total read counts	Total clusters	DNA clusters	LTR clusters
Pissodes strobi	**3,750,000**	**1,144,238**	**204**	**39**	**106**
Elaeidobius kamerunicus	750,000	228,048	62	8	26
Dendroctonus ponderosae	375,000	114,616	42	6	23
Hypothenemus hampei	275,625	83,410	46	11	16
Pachyrhynchus sulphureomaculatus	3,750,000	1,144,092	167	50	90
Listronotus bonariensis	2,062,500	627,190	154	61	64
Rhynchophorus ferrugineus	1,387,500	423,634	52	13	29
Sitophilus oryzae	1,419,375	434,052	151	60	62
Tribolium castaneum	285,000	87,032	20	1	7

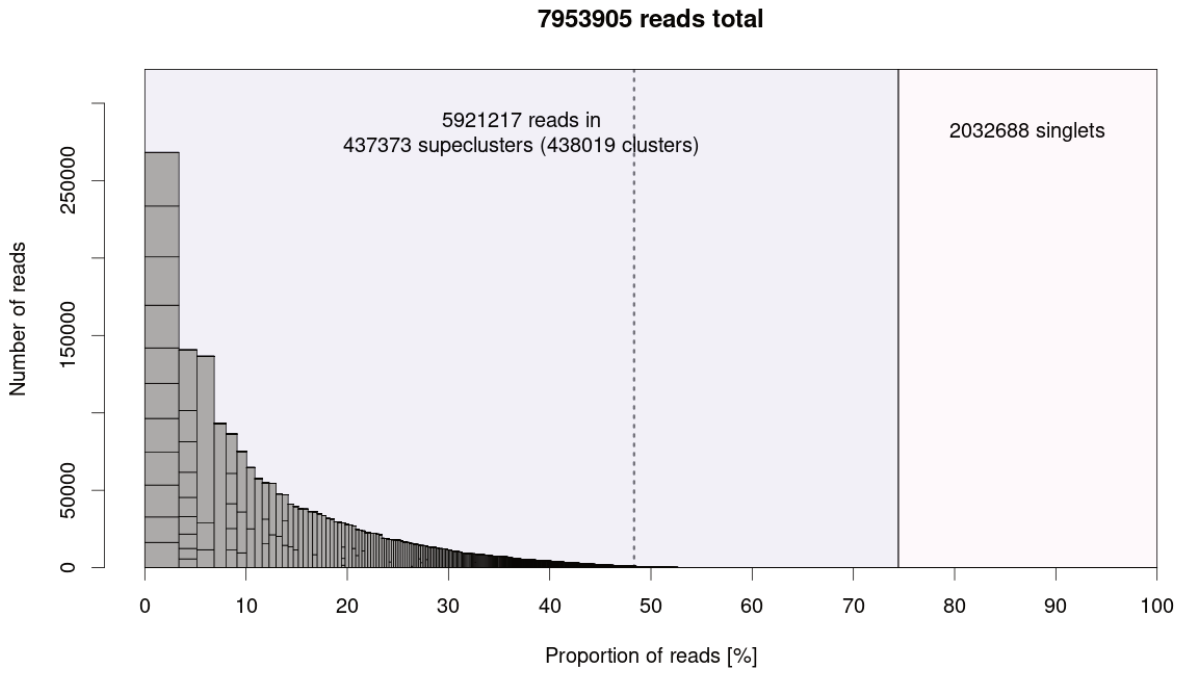

Figure B.3 Repeat composition of clusters generated by RepeatExplorer run for *P. strobi* as individual species. X-axis: cumulative proportion of clustered reads. Y-axis: numbers of reads. Each bar represents a reads cluster identified by RepeatExplorer, with repeat clusters containing a higher number of reads. The software selected a total set of 7,953,905 reads, and 5,921,217 were arranged in 437,373 repeat clusters (gray area in the plot). The dashed line shows the clusters with high copy repeats, defined as ≥0.01% of the sampled reads. Singlets (white area) contain unique reads that are classified as non-repetitive.

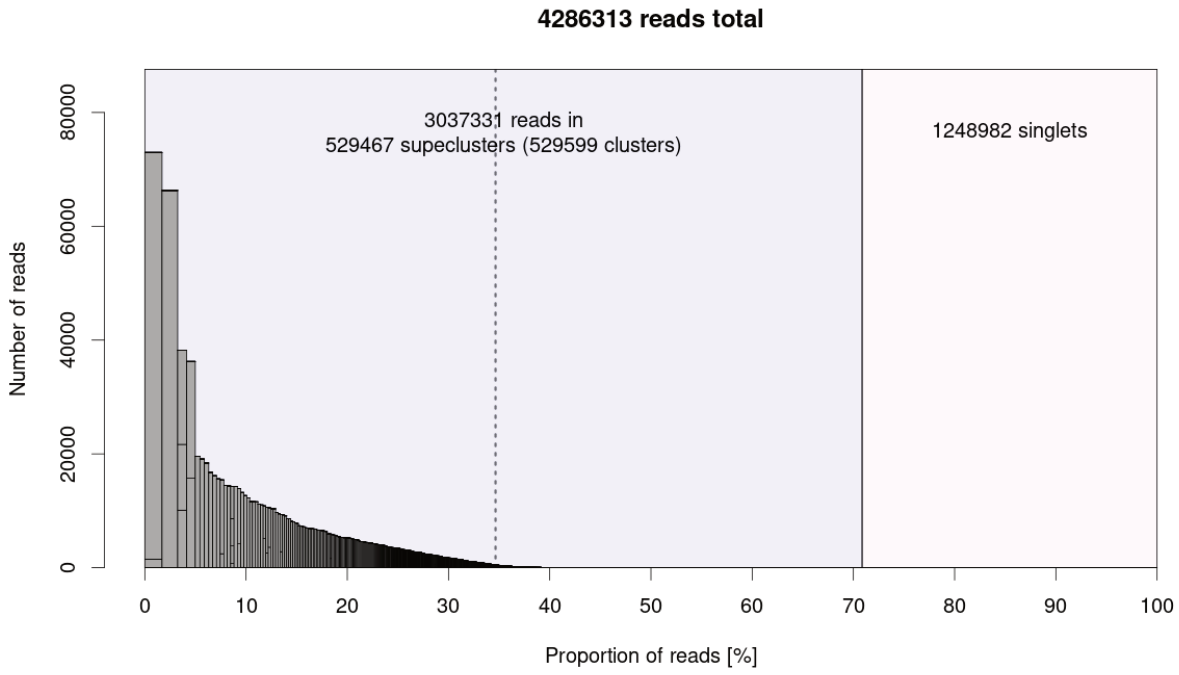

Figure B.4 Repeat composition of clusters generated by RepeatExplorer comparative analysis for Curculionidae species. X-axis: cumulative proportion of clustered reads. Y-axis: numbers of reads. Each bar represents a reads cluster identified by RepeatExplorer, with repeat clusters containing a higher number of reads. The software selected 4,286,313 reads, and 3,037,331 were arranged in 529,599 repeat clusters (gray area in the plot). The dashed line shows the clusters with high copy repeats, defined as ≥0.01% of the sampled reads. Singlets (white area) contain unique reads that are classified as non-repetitive.

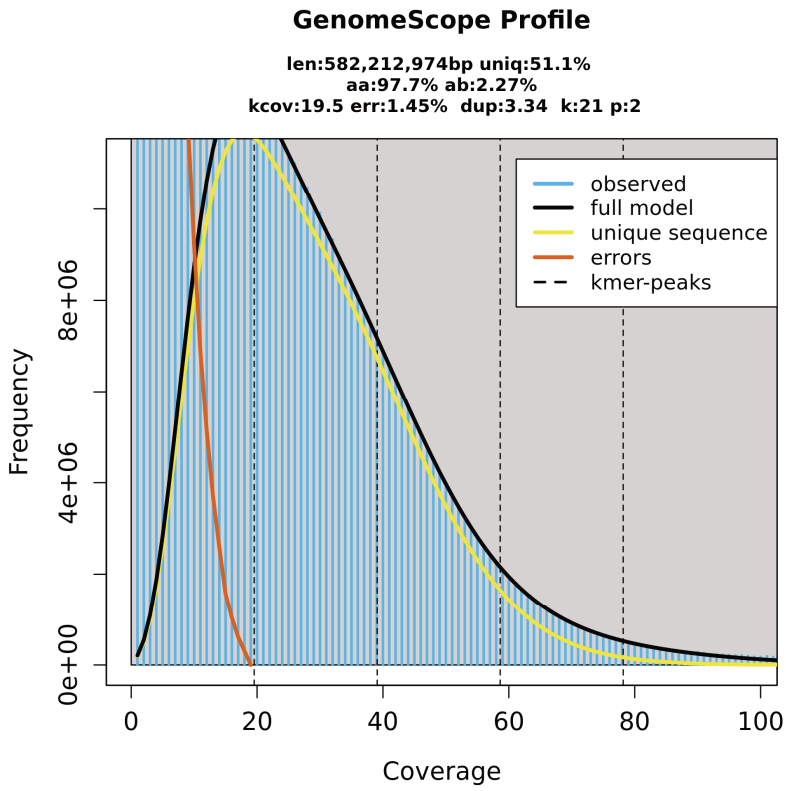

Figure B.5 GenomeScope2.0 genomic features estimates for *E. kamerunicus*. K-mer histogram based on unassembled reads and k-mer 21, assuming a diploid genome. The lack of two peaks in the histogram, typical of the k-mer histograms generated by short reads at high coverage, likely indicates a low genome sequencing coverage.

Appendix C Chapter 4 - Supplemental material

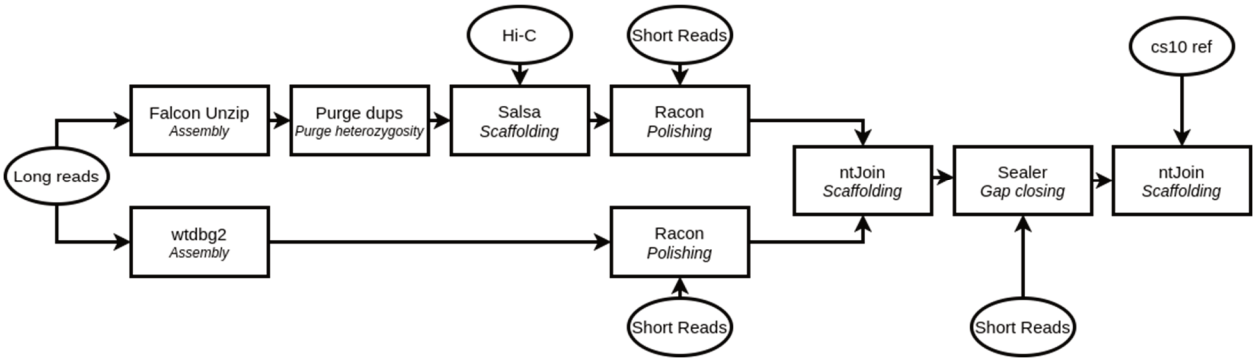

Figure C.1 Willow-alpha genome assembly steps and read datasets. The pipeline shows the tools used for the genome assembly and the corresponding read datasets. Long reads are assembled with Falcon Unzip and wtdbg2; the primary genome assembly was purged from heterozygous scaffolds with Purge_dups, scaffolded with Hi-C chromatin conformation map and polished with short reads and Pilon software. The final assembly was scaffolded with ntJoin, closed for gaps with Sealer, and the final scaffolds were oriented and grouped according to the cs10 reference genome using ntJoin.

Table C.1 Chromosome-scale *C. sativa* genomes and hops outgroup used for phylogenomic inference.

Genome	Species	Chemotype	Assembly ID	Reference
Cannbio-2	*C. sativa*	narcotic-medicinal	GCA_016165845	Braich et al. 2020
CBDRx cs10	*C. sativa*	narcotic-medicinal	GCA_900626175	Grassa et al. 2021
Purple Kush	*C. sativa*	narcotic-medicinal	GCA_000230575	Van Bakel et al. 2011
JL wild accession	*C. sativa*	NA	GCA_013030365	Gao et al. 2020
Finola	*C. sativa*	hemp	GCA_003417725	Laverty et al. 2019
Cascade	*H. lupulus*	outgroup	http://hopbase.org/	Padgitt-Cobb et al. 2021

Table C.2 Genome assembly statistics and gene completeness of Willow-alpha, Cannbio-2, CBDRx cs10, Purple Kush, JL wild accession, Finola and Cascade hops genomes. The assembly statistics shown are the number of scaffolds ≥ 1 kb. BUSCO "complete," represented by "single copy" and "duplicated," is a metric of the gene space reconstruction using the Embryophyta core gene set.

Variety	No. of scaffolds	Longest scaffold (Mb)	Scaffold N50 (Mb)	Reconstructed size (Mb)	BUSCO single copy (%)	BUSCO duplicated (%)
Willow-alpha	**131**	**92.11**	**80.20**	**731.70**	**90.8**	**6.8**
Cannbio-2	147	106.20	91.40	913.50	62.6	35.1
CBDRx cs10	220	86.09	76.97	860.90	89.3	5.6
Purple Kush	6,523	79.17	60.91	891.30	68.3	26.1
JL wild accession	483	92.97	82.97	811.80	80.9	12.9
Finola	2,362	100.60	77.11	1,009.00	79.5	17.5
Cascade - hops	8,661	8,249.94	0.67	3,712.00	60.4	36.8

Table C.3 Phenylpropanoid/flavonoid/anthocyanin biosynthetic genes identified in Willow-alpha and their annotation in the TAIR *A. thaliana* database.

Gene	TAIR annotation	Name	Description
WAG0020687	AT2G37040	PAL1	Phenylalanine ammonia-lyase
WAG0021027	AT2G30490	C4H	Cinnamic acid 4-hydroxylase
WAG0016652	AT1G51680	4CL	4-coumarate:CoA ligase 1
WAG0027265	AT5G13930	CHS	Chalcone synthase
WAG0030806	AT3G55120	CHI	Chalcone isomerase
WAG0004736	AT5G05270	CHI-L1	Chalcone isomerase
WAG0000733	AT3G51240	F3H	Flavanone 3-hydroxylase
WAG0022045	AT5G07990	F3'H	Flavonoid 3'-hydroxylase
WAG0031118	AT5G08640	FLS1	Flavonol synthase
WAG0021021	AT2G36790	UGT73C6	Flavonol 7-O-glucosyltransferase
WAG0020281	AT5G42800	DFR	Dihydroflavonol reductase
WAG0031705	AT4G22880	LDOX-ANS	Leucoanthocyanidin dioxygenase/anthocyanidin synthase
WAG0030150	AT5G17050	UGT78D2	Flavonoid 3-O-glucosyltransferase
WAG0027154	AT4G14090	UGT75C1	UDP-glycosyltransferase 75C1
WAG0033498	AT4G27570	UGT79B3	UDP-glycosyltransferase 79B3
WAG0007954	AT1G61720	ANR	Anthocyanidin reductase
WAG0002077	AT5G48100	LAC15	Laccase

Table C.4 Differential gene expression (DGE) of general phenylpropanoid, flavonoid, anthocyanins, and catechins biosynthetic pathway genes. Willow-alpha leaf gene expression is compared to 1) CA19210, 2) CK10206, and 3) Cali Kush. The rows show the log2FC and adjusted *P*-value for each biosynthetic gene in the pathway; the highlighted entries are significant with *P*-value <0.05 and absolute log2 fold-change >1.5. ANR gene expression is not detectable.

Name	1) Willow-alpha vs. CA19210	2) Willow-alpha vs. CK19206	3) Willow-alpha vs. Cali Kush
PAL1	log2FC -1.133; padj 8.28E-03	log2FC -1.404; pdj 7.27E-04	log2FC -0.777; padj 8.87E-02
C4H	log2FC -0.455: padj 6.99E-02	log2FC -1.421; padj 1.60E-10	log2FC -0.623; padj 1.01E-02
4CL	log2FC -0.127; padj 2.61E-01	log2FC -0.171; padj 1.04E-01	log2FC -0.656; padj 1.03E-12
CHS	**log2FC -1.969; pagj 1.01E-04**	**log2FC -2.717; padj 3.41E-08**	log2FC -0.966; padj 8.40E-02
CHI	log2FC -1.217; padj 1.76E-04	log2FC -1.077; padj 9.68E-04	log2FC -0.263; padj 5.21E-01
CHI-L1	log2FC 1.013; padj 4.46E-03	log2FC 0.582; padj 1.23E-01	log2FC 0.260; padj 5.59E-01
F3H	**log2FC -4.433; padj 1.98E-19**	**log2FC -4.428; padj 2.23E-19**	**log2FC -2.470; padj 1.48E-06**
F3'H	**log2FC -2.526; padj 3.16E-06**	**log2FC -3.304; padj 4.90E-10**	**log2FC -1.635; padj 4.26E-03**
FLS1	**log2FC -2.523; padj 4.98E-05**	**log2FC -3.819; pdj 4.31E-11**	**log2FC -2.021; padj 2.42E-03**
UGT73C6	log2FC 1.556; padj 2.35E-01	log2FC 0.480; padj 7.42E-01	log2FC 2.524; padj 5.24E-02
DFR	**log2FC -1.987; padj 2.79E-05**	**log2FC -2.703; padj 5.09E-09**	**log2FC -1.587; padj 1.19E-03**
LDOX/ANS	**log2FC -2.337; padj 2.24E-05**	**log2FC -3.101; padj 8.59E-09**	log2FC -0.988; padj 1.11E-01
UGT78D2	**log2FC -2.329; padj 7.28E-04**	**log2FC -3.524; padj 1.06E-07**	**log2FC -2.049; padj 3.73E-03**
UGT76C1	log2FC 0.759; padj 7.81E-03	log2FC -0.457; padj 1.28E-01	log2FC 0.724; padj 1.25E-02
UGT79B3	log2FC 1.416; padj 1.35E-01	log2FC 0.659; padj 4.90E-01	log2FC 2.466; padj 2.11E-02
ANR	-	-	-
LAC15	log2FC 2.994; padj 5.55E-02	log2FC 0.331; pdj 7.78E-01	log2FC -0.372; padj 7.48E-01

Table C.5 Anthocyanin quantification (average and standard deviation for biological samples in triplicate) for Willow-alpha and the three high pigmentation varieties, shown in Figure 4.7. The anthocyanins concentration is calculated on cyanidin glucoside equivalence (μg/g fresh weight).

		Willow-alpha		CA19210		CK19206		Cali Kush	
	Anthocyanin	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev
AN1	Cyanidin glucoside	1.6	0.6	0.0	0.0	13.9	4.6	0.0	0.0
AN2	Cyanidin sophoroside	0.0	0.0	235.9	30.4	0.0	0.0	0.0	0.0
AN3	Cyanidin rutinoside	100.4	20.7	606.1	329.8	521.2	135.7	365.2	81.3
AN4	Cyanidin derivative (cyanidin glucoside glucuronide)	15.9	6.4	0.0	0.0	33.2	12.6	27.6	2.4
AN5	Peonidin glucoside	0.0	0.0	5.8	1.4	13.4	4.5	7.9	2.2
AN6	Peonidin rutinoside	2.9	1.4	120.2	37.1	173.3	60.5	123.8	28.8

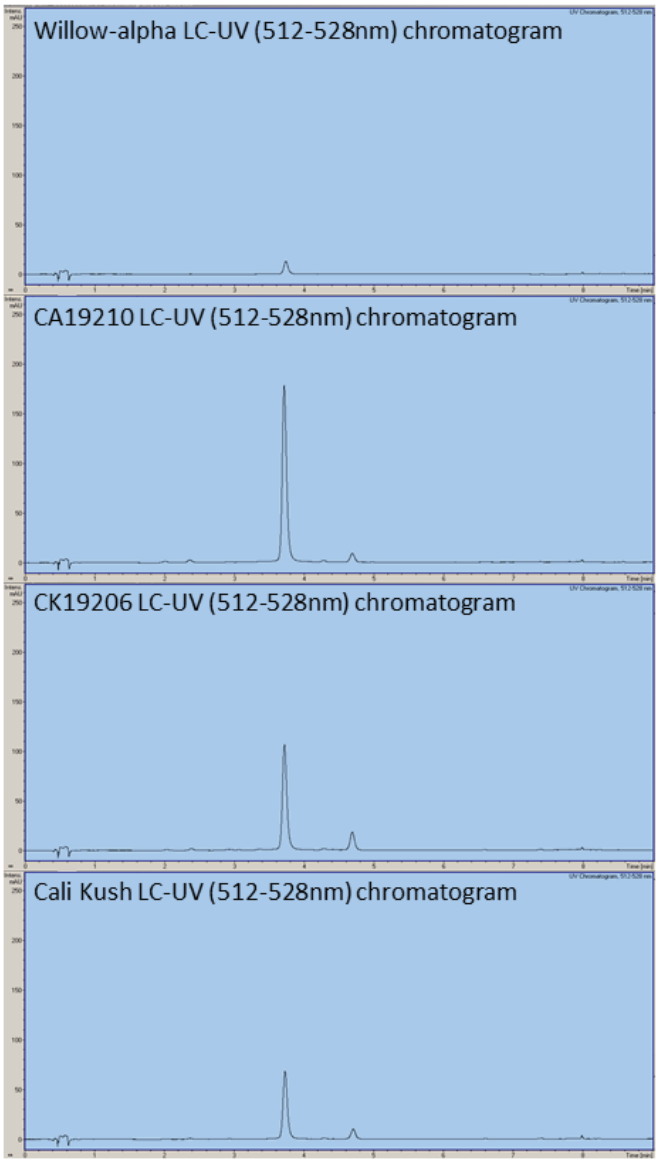

Figure C.2 Liquid chromatography ultraviolet (LC-UV) chromatogram peaks for the four analyzed varieties. Anthocyanins strongly absorb light between 460 and 550 nm (long blue, cyan, and green light), and the peak of maximum absorbance in the four varieties is between 512 and 528 nm.

Table C.6 Flavonoid quantification (average and standard deviation for biological samples in triplicate) for Willow-alpha, two high and one medium pigmented variety. The Flavonoid concentration is calculated on quercetin glucoside equivalence (μg/g fresh weight).

		Willow-alpha		CA19210		CK19206		Cali Kush	
	Flavonoid	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev
FL1	Quercetin glucoside	0.0	0.0	612.8	121.1	0.0	0.0	509.6	48.2
FL2	Quercetin glucuronide	65.6	8.5	0.0	0.0	367.5	54.8	159.0	24.0
FL3	Rutin	0.0	0.0	451.1	25.0	3.7	1.4	298.8	28.0
FL4	Kaempferol glucuronide	42.7	7.5	3.7	0.2	0.0	0.0	27.5	1.9
FL5	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	0.0	0.0	364.1	22.9	0.0	0.0	285.8	19.6
FL6	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	0.0	0.0	328.7	26.0	0.0	0.0	0.0	0.0
FL7	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	1036.3	134.6	0.0	0.0	1801.1	138.7	979.5	92.4
FL8	Apigenin 7-O-glucuronide	789.7	109.2	0.0	0.0	886.5	129.5	264.8	39.6
FL9	Luteolin 7-O-glucuronide	61.9	8.8	36.6	4.5	4.8	217.1	217.1	0.5

Table C.7 Anthocyanin and flavonoid mass-spec data from Trap and QTOF methods, together with the published exact mass.

Compound	Identified MS1 (Trap)	Identified MS2 (Trap)	RT (Trap)	Identified MS1 (QTOF)	Identified MS2 (QTOF)	RT (QTOF)	Published exact mass
Cyanidin glucoside	449	287	3.3	449.1085	287.0570	5.938	449.1084
Cyanidin rutinoside	595	287	3.6	595.1662	287.0548	6.393	595.1662
Peonidin glucoside	463	301	4.3	463.1240	301.0701	7.418	463.1240
Cyanidin sophoroside	611	287	4.5	611.1599	449.1080	7.890	611.1612
Peonidin rutinoside	609	301	4.6	609.1803	301.0683	7.646	609.1819
Cyanidin glucuronide	463	287	4.8	463.0874	287.0567	9.239	463.0875
Quercetin glucoside	465	303	4.4	465.1038	303.0507	7.369	465.1033
Quercetin glucuronide	479	303	6.9	479.0835	303.0514	11.012	479.0825
Rutin	611	303	5.2	611.1601	303.0504	8.768	611.1612
Kaempferol glucuronide	463	287	5.6	463.0868	287.0564	9.239	463.0875
Vitexin glucoside/Isovitexin glucoside	595	433	4.7	433.1105	595.1671	8.020	595.1662
Vitexin glucoside/Isovitexin glucoside	595	433	5.0	595.1659	433.1125	8.475	595.1662
Vitexin glucoside/Isovitexin glucoside	595	433	5.5	595.1668	433.1120	8.751	595.1662
Apigenin 7-O-glucuronide	447	271	6.5	447.0934	271.0626	10.427	447.0926
Luteolin 7-O-glucuronide	463	287	8.0	463.0870	287.0556	12.622	463.0872