Comparative Genomics in Cultivated and Wild Brassicaceae Species

By

JOHN THOMPSON DAVIS

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Julin N. Maloof

Jennifer R. Gremer

Daniel J. Kliebenstein

Committee in Charge

2024

Table of Contents

Acknowledgements: ..... **iii**

Abstract: ..... **iv**

Chapter 1: Genome Report: Whole genome sequence of synthetically derived Brassica napus inbred cultivar Da-Ae **1**

Chapter 2: Genome Report: A chromosome-level genome assembly of the varied leaved jewelflower, Streptanthus diversifolius, reveals a recent whole genome duplication **29**

Chapter 3: Evolutionary dynamics of gene expression associated with germination and climate in California Jewelflowers **57**

Appendix: **83**

Literature Cited: **88**

Supplemental Materials: **102**

Acknowledgements

I would like to thank my parents for supporting me throughout my academic career. From paying my rent in undergrad to paying my phone bill in graduate school you have always helped to relieve my stresses so I can focus on my studies. Of course, I must thank my wife who has been by my side throughout this whole process. Thank you keeping me on track and reminding me what I have been working towards. Lastly, I would like to thank my mentors that have provided me guidance throughout my academic career ever since I was a lost 22-year-old clutching a bachelor's degree and no clue what I was going to do. Thank you for putting me on a path and guiding me to where I am today.

Abstract

This study explores the genetic complexities and evolutionary dynamics of plant adaptation across species within the Brassicaceae family, focusing on *Brassica napus*, *Streptanthus diversifolius*, and exploring gene expression related to seed germination and climate adaptation in the Streptanthoid complex. For *Brassica napus*, an important oilseed crop and an allotetraploid hybrid of *Brassica rapa* and *Brassica oleracea*, we utilized third generation sequencing and assembly technologies to generate a new high quality genome assembly of the synthetic cultivar Da-Ae. This work identifies homoeologous exchange hotspots, illuminating the genetic rearrangements essential for hybrid viability. In parallel, we present a chromosome-level genome assembly for *Streptanthus diversifolius*, also utilizing third generation sequencing technologies. This assembly sheds light on the species' ability to adapt to diverse Californian environments, potentially facilitated by a tribe-specific whole genome duplication event, enhancing its capacity to thrive in inhospitable conditions like serpentine soils. Further, through germination assays and RNA sequencing across multiple California Jewelflower species, we construct gene co-expression networks to correlate with germination and climate adaptation traits. Our findings reveal distinct gene expression patterns driven by climate variations and posit that positive selection has shaped these networks, optimizing germination timing for adaptive success. Collectively, these studies enhance our understanding of the genetic underpinnings of plant adaptability and evolution, showcasing the intricate relationship between genomic architecture, environmental adaptation, and evolutionary innovation in the Brassicaceae family

Genome Report: Whole genome sequence of synthetically derived *Brassica napus* inbred cultivar Da-Ae

Abstract

Brassica napus, a globally important oilseed crop, is an allotetraploid hybrid species with two subgenomes originating from *B. rapa* and *B. oleracea.* The presence of two highly similar subgenomes has made the assembly of a complete draft genome challenging and has also resulted in natural homoeologous exchanges between the genomes, resulting in variations in gene copy number, which further complicates assigning sequences to correct chromosomes. Despite these challenges, high quality draft genomes of this species have been released. Using third generation sequencing and assembly technologies, we generated a new genome assembly for the synthetic *Brassica napus* cultivar Da-Ae. Through the use of long reads, linked-reads, and Hi-C proximity data, we assembled a new draft genome that provides a high quality reference genome of a synthetic *Brassica napus*. In addition, we identified potential hotspots of homoeologous exchange between subgenomes within Da-Ae, based on their presence in other independently-derived lines. The occurrence of these hotspots may provide insight into the genetic rearrangements required for *B. napus* to be viable following the hybridization of *B. rapa* and *B. oleracea.*

Introduction

Brassica napus, commonly known as rapeseed, is the second most widely cultivated oilseed crop in the world (USDA, n.d.). Historically, rapeseed oil was used primarily in the production of lubricants due to its high erucic acid content. In the late 1970s, new, edible, low erucic acid cultivars were created, enabling rapeseed oil to become a major component of most commercial vegetable oil products (Oplinger et al., 1989). The demand for rapeseed oil has caused global production to more than triple in the last few decades, with China and Canada being the world's largest producers ("PSD Online 2020"). Numerous attempts are being made to understand the biology of *B. napus* with the goal of increasing production to keep up with demand.

The genetics of *B. napus* is challenging to untangle due to its genomic complexity. *B. napus* is an outcrossing species that originated from the hybridization of two different diploid parents, *B. rapa* and *B. oleracea* (Nagaharu, 1935). Both *B. rapa* and *B. oleracea* are widely cultivated as human food crops such as cabbage, bok choy, and broccoli. It is believed that *B. napus* first appeared approximately 7,500 years ago when *B. rapa* hybridized with *B. oleracea* and underwent a chromosome doubling event, resulting in an allotetraploid (Chalhoub et al., 2014). *B. napus* (AACC) contains the diploid genomes of both *B. rapa* (AA) and *B. oleracea* (CC). While polyploidy has been hypothesized to provide plants with advantages, such as favorability in domestication (Bertioli et al., 2019), it also has genetic consequences that can cause several analytical challenges. In the case of *B. napus*, the A and C subgenomes are so similar that there can be homoeologous exchange of genetic information between the two subgenomes. Such exchanges range in size from a few base pairs (gene conversion) to larger chromosomal regions

(Chalhoub et al., 2014). The rate and specifics of homoeologous exchange varies between *B. napus* populations. Homoeologous exchange and aneuploidy occur more often in populations that have a newly synthesized *B. napus* as a parent (Ferreira de Carvalho et al., 2021; Higgins et al., 2018; Udall et al., 2005; Xiong et al., 2021, 2011) and loci affecting the homoeologous exchange rate have been identified (Higgins et al., 2021). Homoeologous exchange is thought to be a driving factor in the large amount of diversity found within *B. napus* (Gaeta et al., 2007; Higgins et al., 2018; Hurgobin et al., 2018; Lloyd et al., 2018; Raman et al., 2022; Stein et al., 2017). Consequently, it is important to have genome assemblies from multiple different *B. napus* varieties as an aid to building a pan-genome for this species.

In 2014, a genomic reference assembly for *B. napus* was released to the public (Chalhoub et al., 2014). This assembly, herein referred to as Darmor-bzh, was generated using short read sequencing data. Due to challenges associated with assembling and scaffolding short reads and the high similarity between the two subgenomes, a significant portion of the genome could not be confidently anchored in the assembly and was left unscaffolded (Table 1.1). Since the release of the Darmor-bzh assembly, new sequencing and assembly strategies, including long reads, linked-reads, and proximity data, have become available and fiscally feasible. Recently, new *B. napus* genomes using these technologies have been released to the public (Lee et al., 2020; Rousseau-Gueutin et al., 2020; Song et al., 2020). Concurrently, we generated a genomic assembly for a synthetic *B. napus* that similar to other recently released assemblies includes multiple previously unscaffolded sequences relative to the Darmor-bzh v4.1 assembly. In addition, this new assembly reveals shared and unique homoeologous exchange events compared to different lines of *B. napus*.

Table 1.1. *Length of the 19 pseudomolecules, N50 of the pseudomolecules, total assembly N50, total pseudomolecule length, total number of scaffolds, and total assembly length.*

The statistics of anchored chromosome length of 11 *B. napus* assemblies											
Chromosome	**DaAe**	**Darmor-Bzh_V4.1**	**Darmor-Bzh_V10**	**Quinta**	**No2127**	**Westar**	**Tapidor**	**Gangan**	**Shengli**	**Zheyou7**	**ZS11**
A01	30,963,416	23,267,856	32,958,928	34,049,429	34,418,524	34,072,108	33,385,296	36,624,584	36,358,838	34,646,657	38,004,428
A02	29,581,582	24,793,737	33,432,960	31,833,343	35,184,156	35,445,817	33,310,329	34,912,047	36,815,669	38,454,144	35,943,954
A03	38,724,999	29,767,490	39,685,748	44,229,005	44,024,636	45,938,915	45,374,774	42,633,929	40,627,569	46,180,332	44,868,710
A04	22,079,791	19,151,660	23,101,715	24,995,428	21,150,623	24,941,254	22,742,991	22,175,112	17,977,365	18,425,376	25,679,024
A05	29,228,566	23,067,598	42,112,164	40,081,539	42,988,262	43,898,138	41,389,975	37,886,159	42,946,112	44,454,297	45,991,561
A06	28,937,740	24,396,386	45,146,386	46,545,620	47,152,155	46,607,869	50,976,723	43,921,364	47,226,390	45,572,147	48,704,706
A07	28,277,616	24,006,521	29,390,523	29,069,541	29,386,911	32,489,665	31,880,213	37,845,779	32,247,771	28,254,355	32,302,721
A08	23,154,485	18,961,941	26,309,499	28,540,401	27,050,200	27,113,445	27,789,865	31,072,007	30,634,953	26,738,762	28,329,074
A09	45,044,935	33,865,340	53,549,826	69,118,063	68,282,462	67,849,243	63,820,574	69,520,783	69,970,591	64,447,344	65,862,748
A10	19,559,996	17,398,227	20,778,245	24,204,787	21,441,692	26,564,201	25,274,410	24,379,196	20,220,739	24,607,451	26,592,803
C01	51,431,623	38,829,317	48,239,358	49,844,256	48,500,495	55,568,513	50,976,155	56,937,391	53,237,480	50,906,239	57,880,920
C02	58,167,434	46,221,804	62,297,340	66,774,423	63,747,864	65,831,886	60,608,700	62,611,365	59,402,795	52,853,804	65,293,782
C03	74,222,928	60,573,394	73,669,886	79,398,332	71,696,257	72,844,319	77,070,395	79,770,199	73,374,125	75,927,486	79,061,710
C04	62,924,550	48,930,237	65,837,619	62,271,373	68,372,639	71,968,714	67,448,363	64,192,264	72,788,439	67,164,483	71,179,181
C05	56,537,224	43,185,227	56,382,805	57,344,646	59,684,040	59,594,324	55,324,214	59,804,570	55,526,125	60,477,792	59,550,008
C06	48,209,797	37,225,952	50,218,839	52,702,447	52,092,031	52,822,278	52,406,678	46,598,512	52,241,597	52,102,359	52,512,057
C07	54,958,258	44,770,477	55,656,957	59,821,549	59,911,058	55,243,557	61,376,810	60,504,736	57,612,675	60,142,069	60,986,212
C08	49,204,614	38,477,087	41,681,856	55,907,613	53,895,350	51,700,037	53,000,817	46,677,179	52,530,734	50,112,630	53,660,391
C09	64,956,572	48,508,220	66,465,249	63,235,258	61,068,053	64,409,414	66,565,279	63,602,739	56,715,283	65,173,479	68,416,614
Pseudochromosome N50	51,431,623	38,829,317	53,549,826	55,907,613	53,895,350	55,243,557	53,000,817	56,937,391	53,237,480	52,102,359	57,880,920
Total assembly N50	48,209,797	38,829,317	50,218,839	55,907,613	53,895,350	55,243,557	52,406,678	46,677,179	52,530,734	50,906,239	57,880,920
Pseudochromosome	816,166,126	645,398,471	866,915,903	919,967,053	910,047,408	934,903,697	920,722,561	921,669,915	908,455,250	906,641,206	960,820,604
Total scaffolds	3,164	41	237	3,722	3,733	3,458	3,566	4,930	3,802	4,990	3,332
Total assembly	1,001,499,700	923,795,763	923,795,763	1,004,262,373	1,012,393,425	1,008,283,116	1,014,411,283	1,034,272,646	1,002,553,391	1,016,238,606	1,010,887,456

Methods and Materials

Creation of Synthetic Brassica napus, Da-Ae

The synthetic *B. napus* cultivar Da-Ae (AACC, Korea patent number: 10-1432278-0000, 2014.08.13) used in this study was developed at FnPCo (South Korea) by crossing an inbred *B. rapa (*AA) Chinese cabbage (WC720) with an inbred *B. oleracea (*CC) red cabbage (BW716). After hybridization, the F_1 underwent spontaneous chromosome doubling, producing a naturally occurring allotetraploid *B. napus (*AACC). The hybrid was self-fertilized, and seven seeds were obtained and planted. Only three of the seven plants germinated and flowered, with only one producing seeds. Progeny from this plant were then self-fertilized for six generations with the final generation being designated Da-Ae.

Plant Materials, DNA Extraction, and Library Preparation

Three plant lines were sequenced in this study: the highly inbred Da-Ae, the male parent *B. rapa* (AA, WC720), and the female parent *B. oleracea* (CC, BW716). For each line, 100 seeds

from a single plant were germinated and grown for 8 to 10 days. The resulting seedlings were pooled separately for each line and high molecular weight genomic DNA was extracted by Amplicon Express (Amplicon Express Inc., Pullman, WA, US). The quality of the DNA collected from these three samples was assessed using a Bioanalyzer (Agilent Technologies, Inc. Santa Clara, CA, US). A 10X Genomics library was prepared by the University of California, Davis (UCD) Genome Center. The resulting libraries were sequenced on an Illumina HiSeq X10 by Novogene (Novogene Corporation Inc., Sacramento, CA, US) as 150 bp paired-end reads, producing ~451 million, ~380 million, and ~380 million reads for Da-Ae, the male parent, and the female parent, respectively. An additional 10X Genomics library for Da-Ae was constructed by the UCD Genome Center using a library prep involving sonication, in contrast to the 10X Genomics library prep without sonication. This library was then sequenced on a HiSeq 4000 at the UCD Genome Center, producing ~347 million 151 bp paired-end reads. For Pacific Biosciences (PacBio) sequencing, 32.9 µg high molecular weight DNA from Da-Ae was used for library construction and 19 SMRTcells were sequenced on a PacBio Sequel system (Pacific Biosciences, Menlo Park, CA, US) at the UCD Genome Center, producing ~6.6 million subreads with an average length of ~11.2 kb. An additional 100 seeds from the same Da-Ae plant were grown to produce 4.5 g young leaf tissue, which was sent to Dovetail Genomics (Dovetail Genomics, Scotts Valley, CA, US) for Hi-C library construction. The Hi-C library was then sequenced at the UCD Genome Center on an Illumina HiSeq 4000, producing ~374 million 150-bp paired-end reads.

Generation of 10X Genomics Assemblies

Initial assemblies of *B. napus* were generated using the default Supernova v1.1.5 pipeline (Weisenfeld et al., 2017) with an estimated genome size of 1.12 Gb. The 10X Genomics Da-Ae reads sequenced at the UCD Genome Center and Novogene (hereafter referred to as Da-Ae 10X Davis and Da-Ae 10X Novogene) were both independently assembled. The Da-Ae 10X Davis reads and the Da-Ae 10X Novogene reads resulted in near identical assemblies. As a result, only the Da-Ae 10X Davis reads were used in downstream Supernova assemblies. Upon the release of Supernova-2.0.0, the *B. rapa* 10X*, B. oleracea* 10X, and Da-Ae 10X Davis reads were each individually assembled using the newer software package. The number of reads required for 56X coverage was calculated using the formula genome size x 56 / read length. The expected genome sizes used for *B. rapa, B. oleracea,* and *B. napus* were 530 Mb, 630 Mb, and 1.12 Gb, respectively. These values were then input to Supernova-2.0.0 using the --maxreads parameter. Scaffolds from these three new Supernova assemblies were later used to assess mis-assemblies in Dovetail scaffolding-based assemblies.

Generation of Pac-Bio Assemblies

The PacBio reads were assembled using Canu version 1.6 (Koren et al., 2017) from Maryland Bioinformatics. Canu was configured for the 1.12 Gb genome size of *B. napus* and the reference suggestions for high coverage and polyploid organisms of corrected ErrorRate=0.040 and corOutCoverage=200. The Canu pipeline consisted of three separate steps: correction, trimming, and assembly.

Polishing of Pac-Bio Assemblies

Polishing was performed to improve the quality of the Canu Da-Ae assembly. Polishing was completed using the 10X Da-Ae Davis reads and the Broad Institute's program Pilon v.1.22 (Walker *et al.* 2014). Following the guidelines from 10X Genomics, 23 bp at the start of read 1 and the first base pair of read 2 were removed using Trimmomatic v.0.33 (Bolger et al., 2014a) in order to remove the 10X barcodes and the initial base of read 2 that is often low-quality. The trimmed reads were then mapped to the Canu Da-Ae assembly using bwa version 0.7.16a (Li and Durbin, 2009). The assembly and the mapped read files were fed into Pilon. After polishing, the assembly had approximately the same size and N_{50} as its unpolished counterpart.

Hi-C Scaffolding of Pac-Bio Assemblies

The Canu Da-Ae assembly and the Hi-C reads sequenced at the UCD Genome Center were sent to Dovetail Genomics for scaffolding. The assembly and the Hi-C reads were run through Dovetail's proprietary HiRise pipeline, where the individual contigs were scaffolded to create chromosome scale scaffolds.

Analysis of Hi-C Results

The N_{50}, assembly size, and BUSCO scores of the HiRise scaffolded assembly was measured. Next, all scaffolds from the HiRise generated assembly were compared to the chromosomes of the publicly available Darmor-bzh v4.1 genome hosted by the Brassica database (BRAD) (Cheng et al., 2011). The scaffolds from the HiRise generated assembly were independently aligned to the chromosomes of Darmor-bzh v4.1 using Nucmer with the parameters --maxmatch -l 100 -c 500. The alignments were filtered for quality and all scaffolds 1 Mbp or greater were plotted (See Figure S1.1). If a scaffold aligned best to one reference

chromosome, it was assigned a name based on its alignment. All remaining scaffolds in the assembly were not renamed and retained their HiRise designated sequence IDs. A Hi-C contact map was also generated to access the quality of the assembly. The Da-Ae Hi-C reads were mapped to the Canu Da-Ae assembly using the Arima mapping pipeline developed by Arima Genomics (https://github.com/ArimaGenomics/mapping_pipeline). The BAM file was then converted to a 4dn style pairs file using the *bam2pairs* utility script which is part of the Pairix program suite (https://github.com/4dn-dcic/pairix). The resulting pairs file was then input into Juicer pre program to generate a .hic file (Durand et al., 2016). Visualization of the .hic file was then completed using Juicebox (Robinson et al., 2018).

Assessing Discrepancies Between the Canu Da-Ae Assembly and the Public Reference Assembly

The 21 largest scaffolds in the assembly were independently compared to their corresponding Darmor-bzh v4.1 chromosomes (Darmor-bzh v10 was not available at this time). Regions of discrepancy between the assembly and the reference assembly were identified. The validity of each discrepancy was then tested by aligning PacBio reads and 10X ancestral parent scaffolds to the Canu Da-Ae assembly. The PacBio reads were aligned using BLASR (Chaisson and Tesler, 2012) with a minimum subread length of 10 kb. The 10X ancestral parent scaffolds were aligned using nucmer from the MUMmer software suite (Marçais et al., 2018a). If the region of discrepancy in the assembly had substantial support from the mapped reads and scaffolds, substantial meaning the mapped reads and/or scaffolds spanned the region of the discrepancy and aligned with the Da-Ae assembly, the discrepancy was considered a true difference between our assembly and the Darmor-bzh v4.1 assembly; thus, it was retained. If there was no support, or the mapped reads and scaffolds disagreed with the Canu Da-Ae

assembly, the region of discrepancy was considered a likely error and altered to match Darmor-bzh v4.1. All alterations performed were simple sequence flips to fix assembly inversions. All inversions, except one, were almost exactly encapsulated by the contig boundaries of a scaffold (See Table S1.1). After all identified discrepancies had been addressed, the assembly was considered final. After Darmor-bzh v10 was released we compared Darmor-bzh v4.1 and v10 and found them to be essentially co-linear. The one exception was an inversion on C07, a region where Da-Ae also showed a supported inversion relative to Darmor-bzh v4.1 (that we had not changed). A Hi-C contact map (See Figure S1.2) generated by juicer (Durand et al., 2016) and juicebox (Robinson et al., 2018) supports our assembly.

Transcriptome Assembly and Structural Annotation of Novel Transcripts

RNA-seq reads from thirteen RNA sequencing libraries generated from five tissues (young leaf, flower, bolting tissue, 1 cm silique, and 5 cm silique) of Da-Ae (Li *et al.* 2018) were used for transcriptome assembly and annotation. The raw sequencing data were preprocessed and mapped to the published genome sequence of Darmor-bzh (*Brassica napus* genome v4.1) as described in Li et al., (2018). The mapped reads were then assembled to transcripts using Cufflinks v2.2.1 (Trapnell et al., 2010) with the help of reference annotations. The output GTF files generated by Cufflinks were fed to Cuffmerge and then compared to the annotations from the reference assembly using Cuffcompare. From the output file, transcripts with code “u” were considered novel. Redundant isoforms among these novel transcripts were removed using CAP3 (Huang and Madan 1999), and only transcripts with open reading frames detected using TransDecoder (Haas et al., 2013) were retained for the next step. For *de novo* assembly, post-processed high-quality reads were pooled together and assembled using Trinity (Grabherr et

al., 2011) set to default parameters. The abundance of transcripts was estimated using the Kallisto (Bray et al., 2016) method implemented in the Trinity pipeline, and those with less than one transcript per kilobase million were removed. Transcripts with detected open reading frames were aligned to the Darmor-bzh coding sequences (CDS) using BLASTN (AltschuP et al., n.d.) with an E-value cutoff of 1e-6, and those with high identity (≥ 95%) to Darmor-bzh CDS were filtered. An additional BLASTX search was conducted against NCBI non-redundant protein database using E-value 1e-6 to remove transcripts with no homology to known plant genes. The resulting assembly from reference-based and *de novo* methods were combined for structural annotation using DAMMIT (Scott, 2016) with default parameters to generate the final GFF3 file. BUSCO scores for the final assembly were calculated to assess transcriptome completeness (Cantarel *et al.* 2008; Campbell *et al.* 2014).

Annotation Using MAKER

Annotation was performed using MAKER v.3.01.02-beta (Campbell et al., 2014a; Cantarel et al., 2008a). Prior to running the MAKER pipeline, a custom repeat library was constructed using the MAKER-P Repeat Library Construction-Advanced (Campbell et al., 2014c) (See Table S1.2). MAKER was run with the following parameters: the CDS transcripts from the Darmor-bzh v4.1 assembly (Chalhoub et al., 2014), Darmor-bzh v10 assembly (Rousseau-Gueutin et al., 2020), and the eight *B. napus* assemblies (ZS11, Westar, No2127, Zheyou7, Gangan, Shengli, Tapidor, and Quinta) from Song et. al 2020; the previously identified novel transcripts were used as expressed sequence tag (EST) evidence. The peptide sequences from each *B. napus* assembly mentioned above as well as *B. oleracea* HDEM*, B. rapa* Z1 v2 downloaded from genoscope.cns.fr, and the *A. thaliana* Araport11 peptides downloaded from

the TAIR Project (Berardini et al., 2015a) were used as evidence for protein homology. MAKER parameters that were modified included the following: A custom Augustus gene prediction species model of Da-Ae created using BUSCO v3.0.2 with the long parameter was used as the model species for Augustus; repeat library was set to the custom repeat library we constructed using the MAKER-P Repeat Library Construction-Advanced protocol; est2genome was set to 1; protein2genome was set to 1. All other parameters not listed above were left as the MAKER defaults. Due to an unresolved bioinformatics issue, 10 kb of chrC01 sequence starting at 47,446,387 had to be masked with N before MAKER would run to completion.

Once annotation of each chromosome was completed, the MAKER proteins were compared to the Uniref90 protein set using BLASTP. Protein domains were then identified using InterProScan on the MAKER predicted proteins. Using accessory scripts provided with MAKER, the MAKER genes were then renamed with the prefix "Bna," the suffix "Da-Ae," and the BLASTP and InterProScan results were integrated into the GFF annotation files. Finally, the annotations were filtered to remove any annotation that contained an Annotation Edit Distance (AED) score greater than 0.5. The cutoff of 0.5 was selected based on the recommendation listed in Campbell *et al*. (2014a).

Analysis of Homoeologous Exchange Between Subgenomes

We examined homoeologous exchange using two methods: synteny analysis and read coverage. For synteny analysis we first identified "trusted" syntenic regions of the A and C subgenomes by performing a nucmer (Marçais et al., 2018a) alignment of *B. rapa* (A chromosomes) and *B. oleracea* (C chromosomes) assemblies; hits were filtered to require > 85%

identity for > 1000bp. We next used nucmer to align each *B. napus* assembly in our comparison (Da-Ae, Darmor-BZH_V10, GanganF73, No2127, QuintaA, Shengli3, Tapidor, Westar, Zheyou73, and ZS11) to an *in silico B. napus* genome constructed by combining both the *B. rapa* (Istace et al., 2021) and the *B. oleracea* (Belser et al., 2018) chromosomes (hereafter referred to as "ancestral"). Candidate homoeologous exchange regions were defined as those where the sequence was from an "A" subgenome in the *B. napus* assembly but had its highest hit to a "C" region in the "ancestral" assembly or *vice versa*. Candidate homoeologous exchange regions were filtered to retain only those with greater than 90% identity and a length of greater than 100bp and that overlapped with "trusted" syntenic regions in the ancestral assembly (to eliminate false positives that could be caused by assembly gaps in *B. rapa* or *B. oleracea).* All nucmer runs used version 4.0.0 with default parameters; all nucmer results were filtered to retain the one best alignment using the delta-filter program with option "-1".

When sequence reads from one genome are mapped to an assembly from a different genome, differences in homoeologous exchange between the two genomes will lead to decreased or increased read coverage. To utilize this kind of information, we performed a coverage analysis using the 10X Da-Ae Davis reads. All reads were trimmed for quality using Trimmomatic and the adapter sequences were removed with the parameters ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 before being mapped with BWA to the *in silico* "ancestral" *B. napus* genome described above. To find possible sites of homoeologous exchange, we first filtered the 10X Da-Ae Davis reads to retain those that could reliably be described as coming from either the A or C subgenome (i.e., those with unique and trustworthy mapping locations). To do so, the alignment file was filtered

to only contain alignments that had a MAPQ of five or greater, were properly paired, had no supplementary alignments, and were primary alignments. Reads that passed these filters were then mapped to nine *B. napus* genomes. The coverageBed function from bedtools2 v2.29.2 (Quinlan and Hall, 2010) was then used to calculate the coverage across the genomes and the coverage of the individual potential genes previously identified. The alternate mapping sites were also captured using the "XA" tag from the bwa output. Using edit distance as a filtering parameter, alternate mapping sites that had an edit distance equal to or less than the primary alignment's edit distance were added to the coverage calculation. To calculate coverage across the genomes, median coverage in a window size of 100 kb with a step size of 20 kb was used. The calculated coverages were standardized based on the genome-wide average using R (R Core Team 2020). Prior to standardization, regions that contained ≥10X mean coverage of their chromosome were removed from further analysis. The coverages were then plotted to identify regions across the genome with higher or lower than average coverage. Coverages were plotted using ggplot2 (Wickham, 2009) in R. Plots combining the coverage and synteny analyses were plotted using a modified version of the plotsr program (Goel and Schneeberger, 2022).

Annotation of genes in shared homoeologous exchange regions was done by using BLASTP to query an Arabidopsis ARAPORT11 protein database with *Brassica rapa* or *Brassica oleracea* protein sequences at phytozome (D M Goodstein et al., 2012), keeping the one best hit, and downloading ARAPORT11 annotations from phytozome.

Results

To develop a high-quality assembly of this new, synthetic *B. napus* we took advantage of contemporary technologies by using a combination of 10X Genomics, Pacific Biosciences, and Dovetail / Hi-C methods (Figure 1.1). The application of each is described in turn below, followed by the results of the annotation and homoeologous exchange analysis.

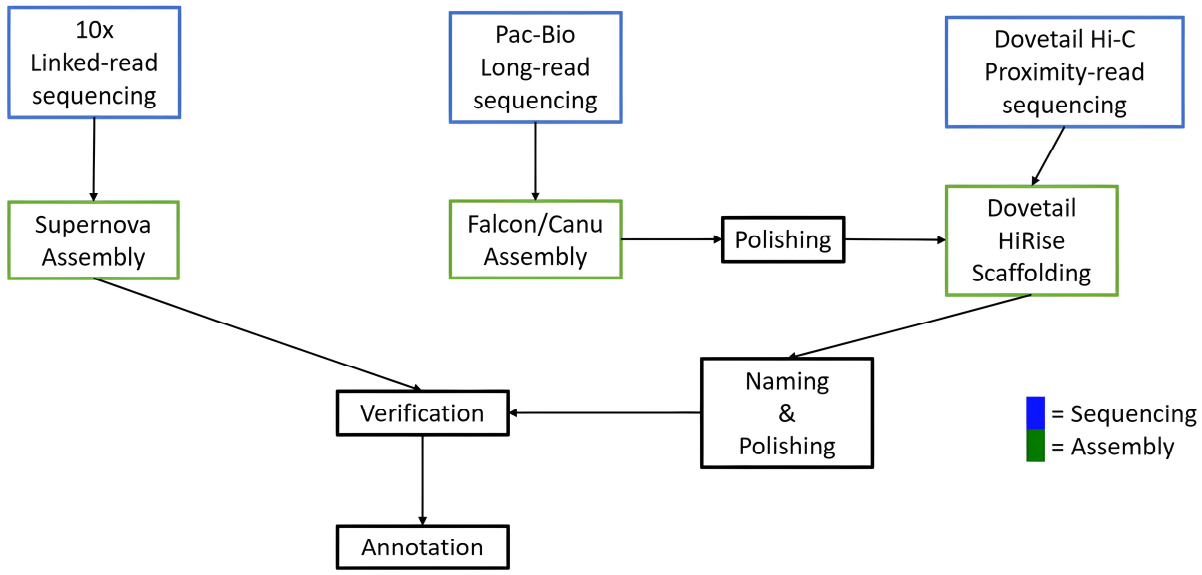

Figure 1.1. *Genome assembly and annotation strategy.*

Supernova Assemblies

The Da-Ae 10X Davis reads were assembled with Supernova v2.0.0. The assembly had a length of 918 Mb and an N_{50} of 1.5 Mb. Notably, the BUSCO scores of this new assembly approached the scores of the Darmor-bzh v4.1. The 10X reads for both *B. rapa* and *B. oleracea* assembled using Supernova v2.0.0 also showed promising results. Both assemblies had N_{50} values over 2 Mb and consisted of less than 20,000 scaffolds. Although all assemblies were smaller than the expected genome sizes, they were all on par with the sizes of the public

references. The assembly metrics and BUSCO scores supported the use of the assembly scaffolds in the manual curation of the subsequent *B. napus* Da-Ae assembly.

PacBio Assembly and Dovetail Scaffolding

Da-Ae PacBio reads were assembled with Canu and polished with Pilon. This Pilon-polished Canu Da-Ae assembly was then scaffolded using the HiRise pipeline by Dovetail Genomics. After HiRise scaffolding, the Canu Da-Ae assembly showed a large increase in N_{50} from 1.59 Mb to 42.79 Mb, and had 3,190 scaffolds. Twenty-three of the scaffolds were greater than 1 Mb, with the largest being 74.2 Mb (See Table S1.3). Regarding BUSCO scores, the scaffolding caused the single to duplicate ratio to increase in the Canu Da-Ae assembly while the percentage of complete BUSCOs, 98.6%, did not change in the Canu Da-Ae assembly (See Table S1.3).

Assigning Scaffolds to Chromosomes

To assign the scaffolds to the established chromosomes, the assembly was aligned to the Darmor-bzh v4.1 assembly using Nucmer. The 19 Darmor-bzh v4.1 chromosomes were covered by the 21 largest Canu scaffolds; 17 spanned the full length of their sister Darmor-bzh scaffold, while the remaining four scaffolds had to be concatenated into pairs to span ChrC06 and ChrC07 (See Figure S1.1). Names were then assigned to the scaffolds based on which Darmor-bzh chromosome they aligned to.

Assembly Discrepancies

Comparison of the Canu Da-Ae assembly to the Darmor-bzh v4.1 assembly revealed 24 assembly discrepancies (See Table S1.1). These discrepancies included inversions, lack of contiguity, and introduction of new sequence. To assess the validity of these discrepancies, both the parental 10X scaffolds and the PacBio reads were mapped to the Canu Da-Ae assembly. In 15 of the 24 discrepancies, the Canu Da-Ae assembly was supported by either read mapping or scaffold evidence. In ChrC06 and ChrC07, two scaffolds spanned the whole reference chromosome but failed to be scaffolded together. These scaffolds were joined with 100 Ns to signify a scaffolding gap and were then able to span the entire Darmor-bzh v4.1 chromosome as one scaffold. In six cases, the Canu Da-Ae assembly had unsupported inversions with four of the inversions spanning from one scaffold gap to another scaffold gap. For each case, the sequence was inverted to match the Darmor-bzh v4.1 assembly. The most prominent discrepancy occurred on ChrA05. Alignment to Darmor-bzh v4.1 suggested that both chromosome arms were inverted at their junction with the centromere. As there was no read or scaffolding evidence to support this, both chromosome arms were inverted to match Darmor-bzh. Although our ChrA05 now agrees with the Darmor-bzh v4.1 assembly, the orientation and centromeric region remains questionable. After all discrepancies were addressed, the assembly was deemed final and annotation began. Darmor-bzh v10 was released after our assembly was finalized. Darmor-bzh v4.1 and v10 are nearly co-linear. In the one place where they are not, C07, Da-Ae matches v10, so there was no need to update our assembly. Figure 1.2 provides a synteny plot of the final Da-Ae assembly against Darmor-bzh v10.

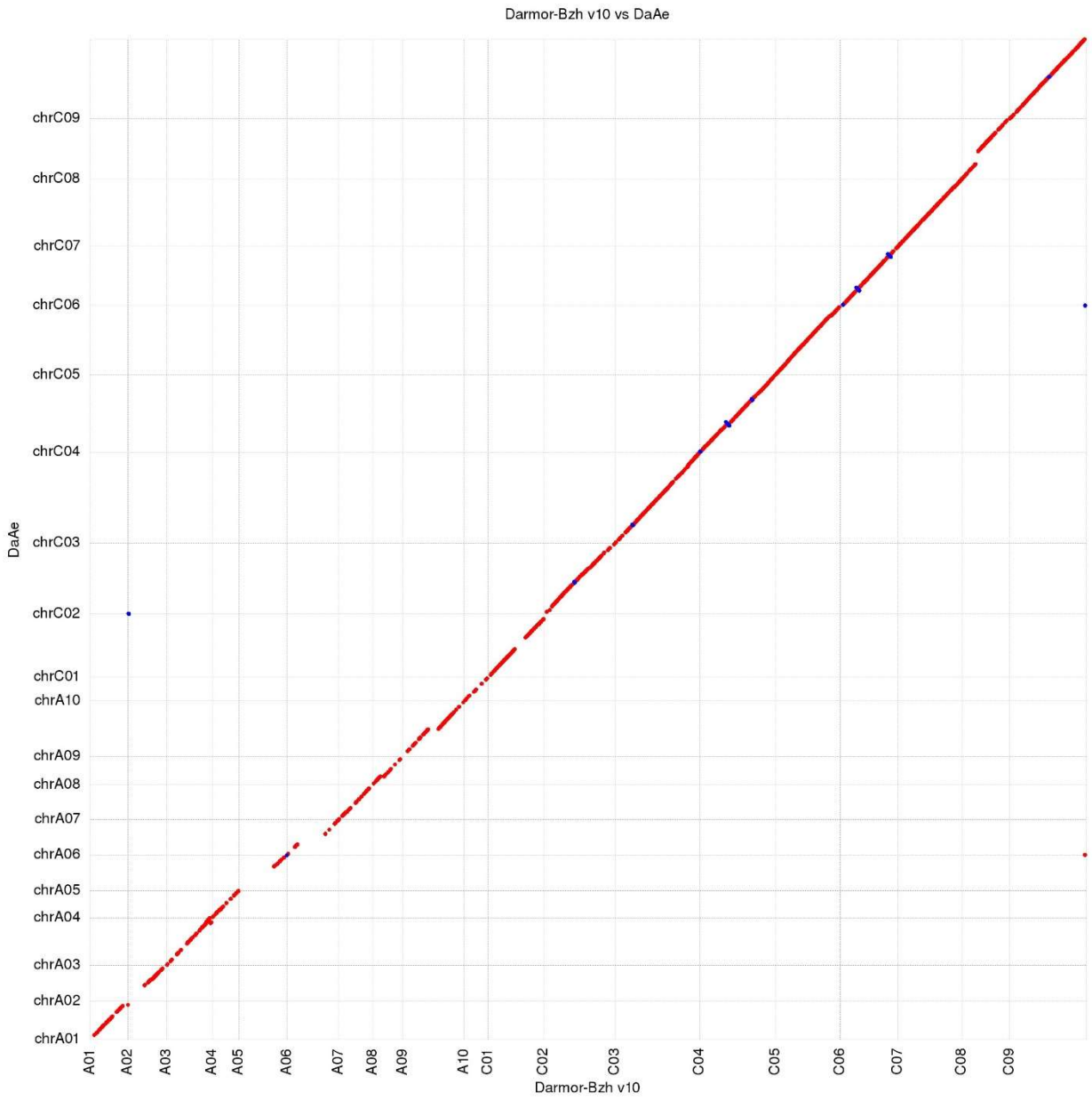

Figure 1.2. *Nucmer plot of the final Da-Ae assembly aligned to the Darmor-bzh v10 reference. A total of 19 final assembly pseudomolecules are aligned to 19 reference pseudomolecules. Red indicates an alignment in the forward direction and blue indicates an alignment in the reverse direction*

Annotation

MAKER analysis of the Da-Ae assembly predicted 125,439 protein coding genes after filtering, compared to the 101,400 and 108,190 genes annotated in the Darmor-bzh v4.1 and v10 assemblies. To explore these differences, we determined the location of the predicted

genes in their respective assemblies. Da-Ae contains more gene models than Darmor-bzh v4.1 and v10, with 123,488 of the Da-Ae gene models being present on its 19 pseudomolecules compared to Darmor-bzh v4.1 and v10 which contain 80,927 and 106,885 gene models on their 19 pseudomolecules, respectively. These discrepancies could be due to the differences in length of time since polyploidization. Since Da-Ae is a new synthetic it has had much less time for gene loss after the polyploidization event.

Final Assembly Comparison

The final Da-Ae assembly improves upon the Darmor-bzh v4.1 assembly by a number of criteria (Tables 1.1 and 1.2). Comparing the full assemblies and the pseudomolecule assemblies, respectively, the N50 is 24% to 32% longer, and there are 36% to 47% more unambiguous bases incorporated into the Da-Ae assembly (Table 1.1). When compared to the Darmor-bzh v10 assembly, the full Da-Ae assembly and the pseudomolecule assembly each have 4% shorter N50s. However, the full Da-Ae assembly has 12% more unambiguous bases than Darmor-bzh v10, while the pseudomolecule Da-Ae assembly has 4% fewer unambiguous bases than Darmor-bzh v10 (Table 1.1). When comparing BUSCO scores using the brassicales_odb10 dataset, both the Da-Ae assembly and the Darmor-bzh v10 assemblies had BUSCO complete scores of 98.5%, while Darmor-bzh v4.1 had a slightly lower score of 98.2%. Both Darmor-bzh assemblies had a higher percentage of complete single copy-BUSCOs, whereas the Da-Ae assembly had a higher percentage of duplicated BUSCOs.

Table 1.2. *Percentages of BUSCO scores.*

The BUSCO (%) statistics of 11 *B. napus* assemblies						
Assembly	Complete BUSCO scores	Complete single-copy BUSCO scores	Complete duplicated BUSCO scores	Fragmented BUSCO scores	Missing BUSCO scores	# BUSCO scores
DaAe	98.5	18.0	80.5	0.2	1.3	4,596
Darmor-Bzh_V4.1	98.2	20.6	77.6	0.2	1.6	4,596
Darmor-Bzh_V10	98.5	19.6	78.9	0.1	1.4	4,596
Quinta	98.7	20.1	78.6	0.0	1.3	4,596
No2127	98.4	25.1	73.3	0.2	1.4	4,596
Westar	98.7	19.9	78.8	0.0	1.3	4,596
Tapidor	98.5	21.7	76.8	0.1	1.4	4,596
Gangan	98.5	21.7	76.8	0.1	1.4	4,596
Shengli	98.4	22.1	76.3	0.1	1.5	4,596
Zheyou	98.5	21.6	76.9	0.0	1.5	4,596
ZS11	98.5	19.9	78.6	0.1	1.4	4,596

BUSCO score percentages were calculated using the brassicales_odb10 data set, which contains 4,596 BUSCO scores.

Genome Completeness Analysis

Genome completeness of Da-Ae and Darmor-bzh v10 was analyzed using the public unigene set of 133,127 *Brassica* sequences. Of the 133,127 sequences, 116,897 (87.81%) were present in the pseudomolecules of both genomes. Overall Darmor-bzh v10 contained the most unigene sequences, 118,199, with Da-Ae a close second with 118,193 unigene sequences. A total of 13,632 (10.24%) were missing from both genomes. To determine if there were classes of genes that were deleted/missing in these genomes, we looked for enriched GO terms among the set of genes missing from the two genomes. Among the enriched categories, enrichment for genes involved in responses to biotic and abiotic stressors was particularly noticeable, as seen in the pink box in the left of Figure 1.3 (Supek et al., 2011a). We also looked for unigenes present in Da-Ae but not in Darmor-bzh v10 and vice versa. Here, among the enriched categories, we noted an enrichment for genes involved in very long chain fatty acid metabolism, perhaps reflecting different breeding selection targets for these oil-seed crops, as seen in the teal box in the bottom middle of the treemap (Figure 1.4).

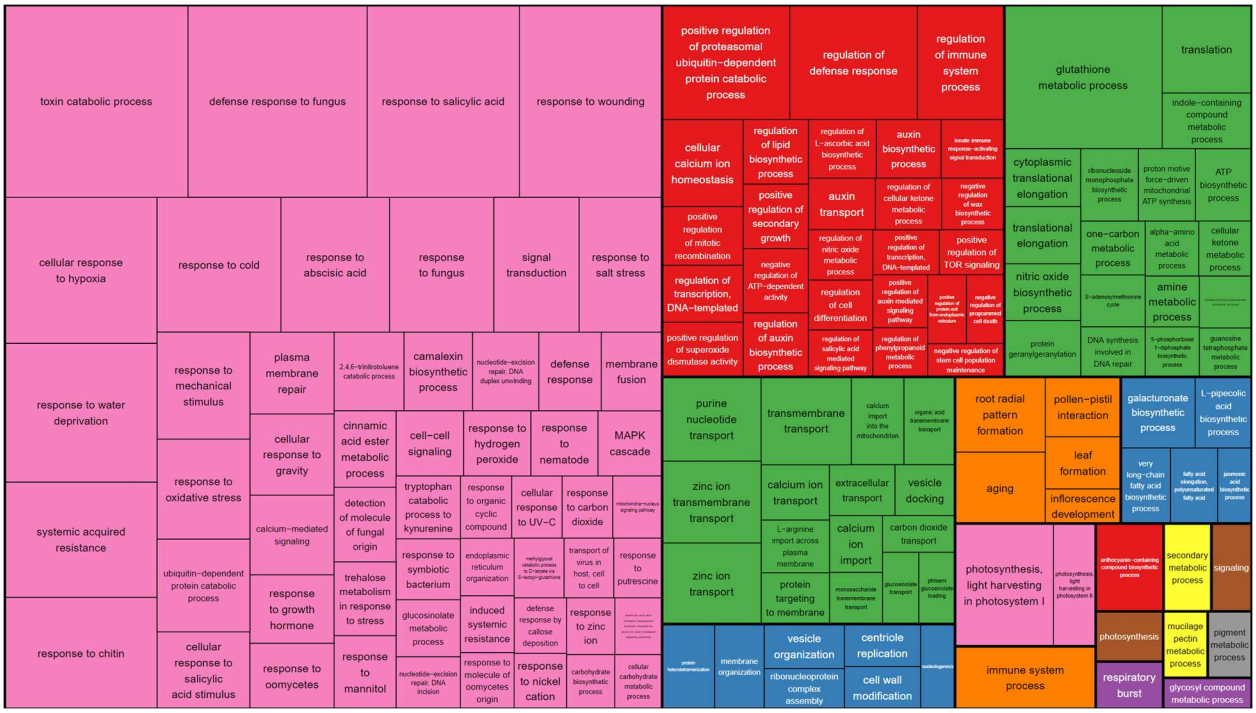

Figure 1.3. *Tree map displaying over-represented BP:GO terms in the set of unigene sequences not found in Da-Ae or Darmor-bzhv10.*

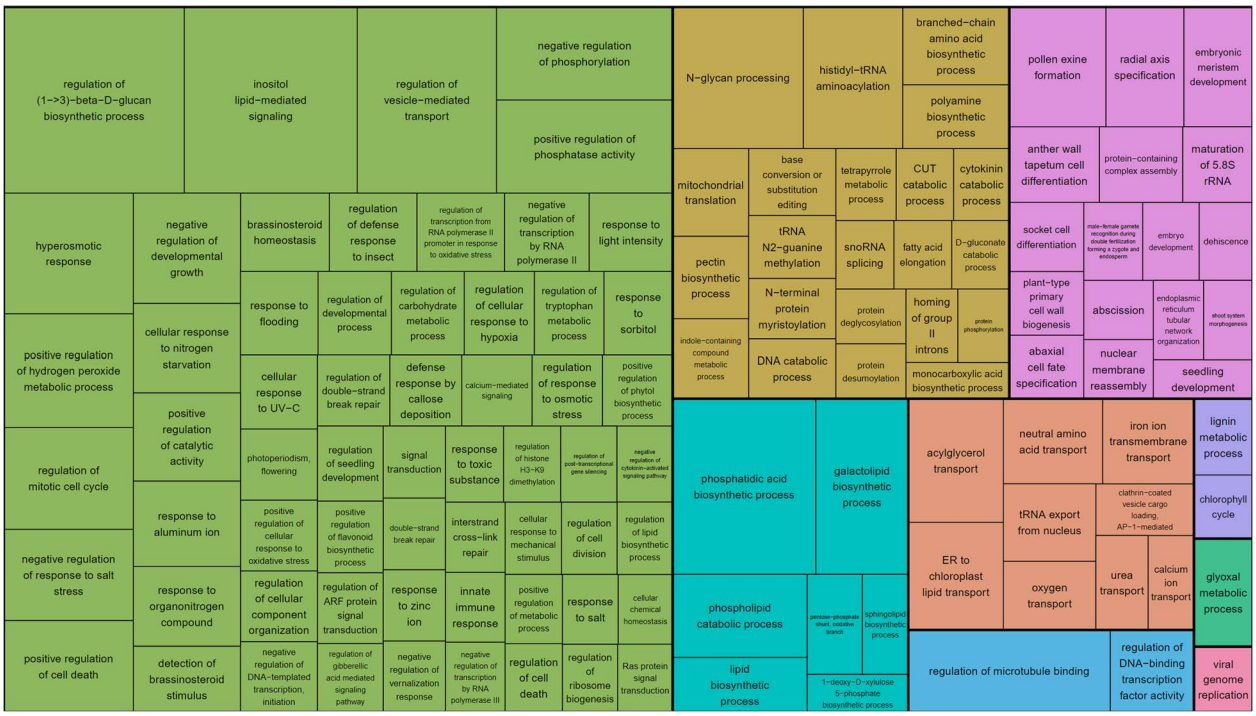

Figure 1.4. *Tree map displaying over-represented BP:GO terms in the set of unigene sequences found in Da-Ae but not Darmor-bzh v10 and vice versa.*

Homoeologous Exchange

Homoeologous exchange is the exchange of genetic material from one subgenome to the other. This could result in the conversion of an A subgenome gene to a C subgenome gene or vice versa. Because *B. napus* is an allotetraploid containing two diploid subgenomes, A and C, homoeologous exchange can result in homoeolog ratios of 2:2, 3:1, or 4:0, corresponding to reciprocal, partial, or complete conversions, respectively. We used two criteria to identify and characterize candidate homoeologous exchange regions: 1) a synteny analysis in which the *B. napus* genome assemblies were aligned to concatenated *B. rapa* and *B. oleracea* genomes, as proxies for the ancestral A and C subgenomes; 2) a coverage analysis in which we looked for regions of decreased or increased coverage that would result when two *B. napus* genomes had different homoeologous exchange (see material and methods).

For the coverage analysis we examined read coverage of Da-Ae when mapped to itself, a pseudo "ancestral" genome of concatenated *B. rapa* and *B. oleracea,* and nine existing *B. napus* assemblies (See Figures S1.3 to S1.21). Note that in these plots, average coverage is normalized to "1"; a partial conversion (one but not both homologs) will result in readings of ~0.5 and ~1.5, whereas a complete conversion will give coverages of ~0 and ~2 on this scale. Each chromosome shows a region of coverage elevated to 4X or higher, likely representing centromeric regions where the repeats are collapsed in the assembly. In addition, we see numerous regions with coverage in the 0, 0.5, 1.5, or 2X range.

To determine if any of regions with increased or decreased coverage might result from homoeologous exchange versus the aneuploidy that is common in nascent synthetic lines

(Ferreira de Carvalho et al., 2021; Xiong et al., 2011), we plotted the coverage and synteny analysis together (See Figures S1.22 to S1.32). In many cases the change in coverage is due to homoeologous exchange. For example, examining ZS11 and Da-Ae reveals that ZS11 had a reciprocal exchange between the right-hand sides of A01 and C01, whereas in Da-Ae this region of C01 was converted to A01 (Figure 1.5A). As a consequence, there is low Da-Ae coverage at the end of ZS11 A01 and high coverage at the end of ZS11 C01 (since that region corresponds to ancestral A01 and Da-Ae has two homoeologs matching A01 in this region). Comparing Zheyou73 and Da-Ae A02 and C02 reveals that both ends of Da-Ae C02 have been converted to A02 and a region in the middle of Zheyou73 A02 has been converted to C02, with read coverage changing as expected (Figure 1.5B). Other regions with increased or decreased coverage but no evidence of homoeologous exchange could result from insertion/deletion differences between the genomes, aneuploidy, or incomplete genome assemblies.

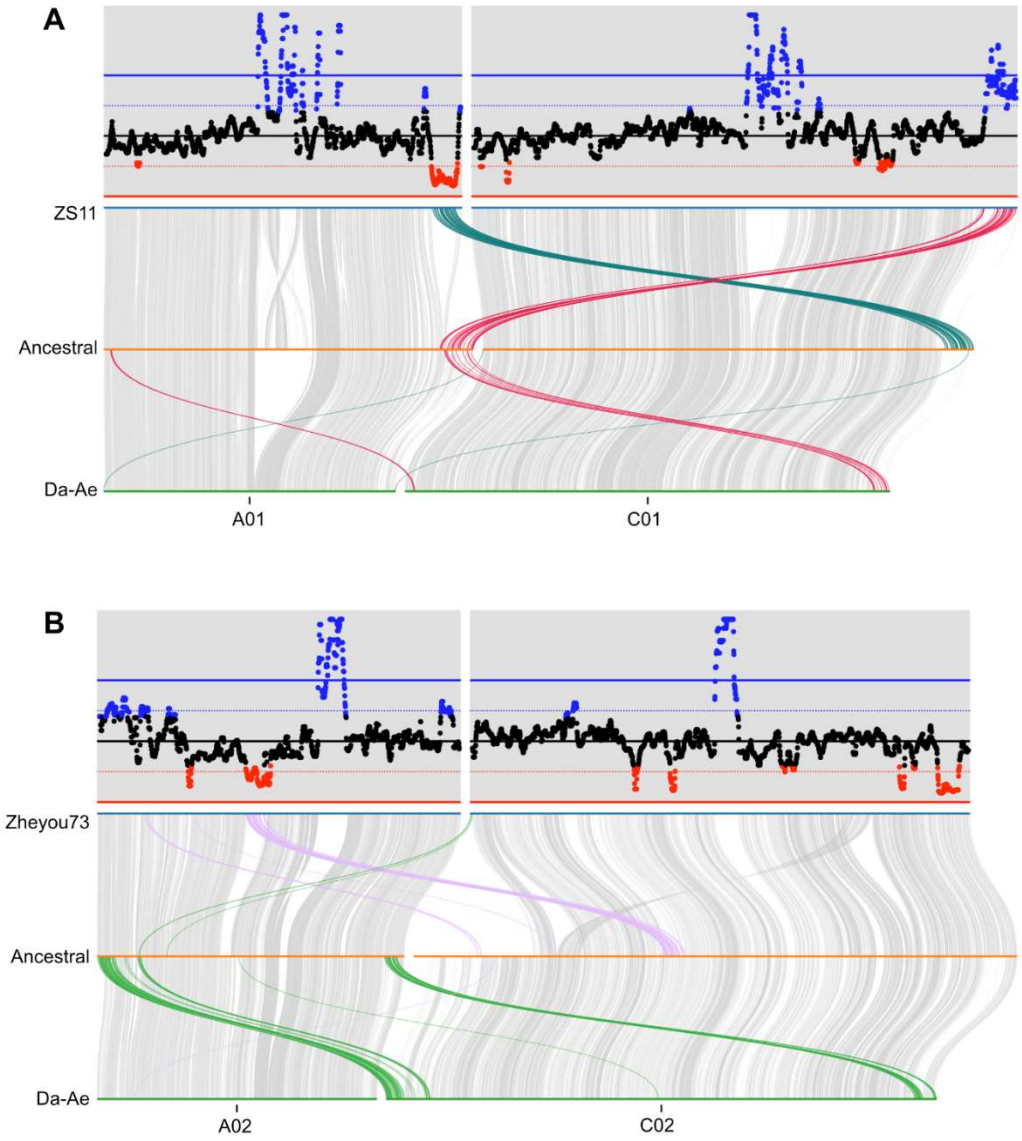

Figure 1.5. *Examples of homoeologous exchange and coverage of Da-Ae reads when mapped to other genomes. The upper parts of each panel show the coverage of Da-Ae reads when mapped to ZS11 (a) or Zheyo73 (b), while the lower parts show syntenic and homoeologous exchange regions. a) ZS11 has a reciprocal exchange between the right ends of A01 and C01, whereas DaAe has a replacement of C01 with A01 in this region. b) Zheyou73 and DaAe show several non-overlapping, non-reciprocal exchanges between A02 and C02.*

The synteny plots (See Figures S1.22 to S1.32) reveal that there are numerous regions where homoeologous exchange has occurred in the same place in different genomes. Since Da-Ae and No2127 are independent synthetic *B. napus* lines, this suggests that there are hotspots

of homoeologous exchange. Figure 1.6 shows the similarity in homoeologous exchange regions across the varieties; as expected, the two synthetic varieties, Da-Ae and No2127 are the most dissimilar from the other varieties. We next asked if there were any homoeologous exchange regions shared among all varieties. We found 31 homoeologous exchange regions encompassing a total of 39kb that were common across all varieties. This is more overlap than predicted by chance; based on the proportion of each genome involved in homoeologous exchange we would expect zero bases to be common across all varieties. There are a total of 16 genes in the conserved exchange regions, 14 of which had strong homologs in the Arabidopsis genome (See Table S1.4). Of these fourteen, one, *BolC8t52214H*, is a nucleotide binding site leucine-rich repeat protein whose closest Arabidopsis homolog is *AT1G12210* or *RPS5-LIKE 1*, a close paralog of the defense R gene *RPS5*. Two other genes have leucine-rich repeats although their relationship to plant immunity is less clear.

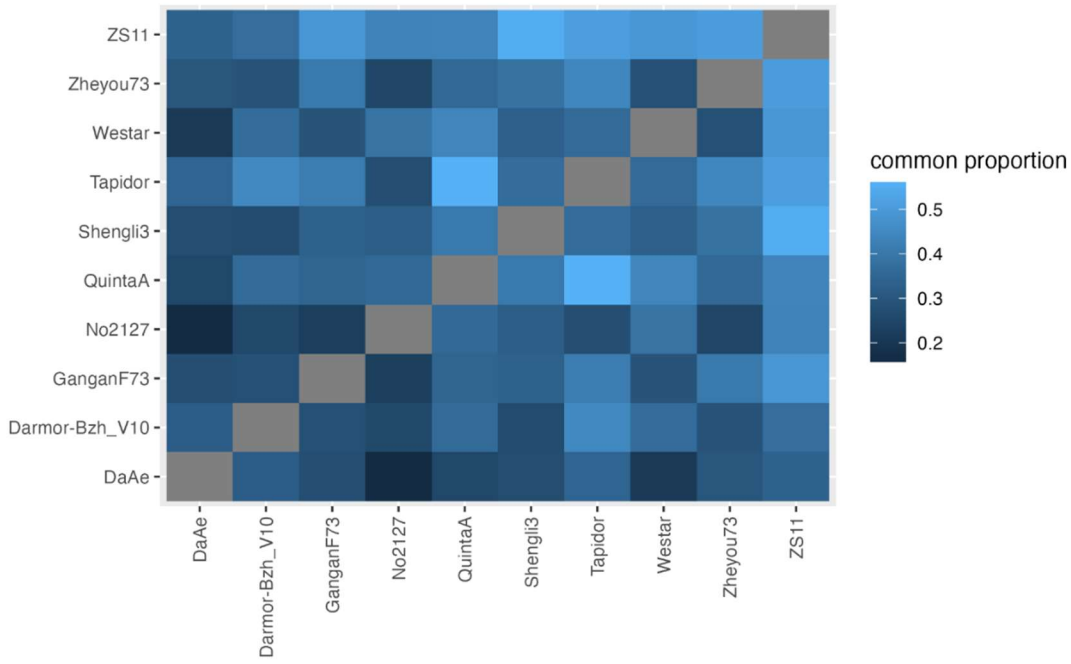

Figure 1.6. *Proportion of common homoeologous exchange regions between genomes. Each tile shows the amount of shared homoeologous exchange regions between the genomes, proportional to the pairwise minimum homoeologous shared amount.*

Discussion

Since the release of the first reference genome (Chalhoub et al., 2014), multiple research groups have released genome assemblies of different *B. napus* cultivars, analyzed homoeologous exchange, and identified quantitative trait loci (QTLs) related to key agricultural traits (Bayer et al., 2017; Boideau et al., 2022; Rousseau-Gueutin et al., 2020; Samans et al., 2017; Song et al., 2020; Stein et al., 2017; Wang et al., 2015). These efforts all contribute to untangling the genome biology of *B. napus* that will one day be combined to create a species-wide pangenome.

The first *B. napus* reference was assembled and released during a time when sequencing technologies from PacBio, 10X Genomics, and Dovetail Genomics were in their infancy and/or not fiscally feasible for most research groups. As a result, the first release of the *B. napus*

genome was not able to benefit from the analytical power of these technologies. This is reflected in the assembly size of the Darmor-bzh V4.1 genome (Chalhoub et al., 2014). Although the expected size of the *B. napus* genome is over 1 Gb, the Darmor-bzh V4.1 genome assembly is only approximately 850 Mb, of which 650 Mb is contained in 19 chromosome-scale pseudomolecule scaffolds. By using a recently created synthetic *B. napus*, *Da-Ae*, along with long-read, linked-read, and proximity ligation technologies, we were able to generate a new synthetic *B. napus* genome that exceeded the first high-quality reference genome by several metrics and is on par with more contemporary assemblies. Our assembly of Da-Ae is over 1 Gb, with more than 800 Mb contained within 19 chromosome-scale pseudomolecule scaffolds. While our assembly is larger compared to both the Darmor-bzh V4.1 and v10 assemblies, it still maintains a high level of sequence collinearity with the two Darmor-bzh assemblies. On a gene level, the Darmor-bzh v4.1 and v10 references have fewer annotated genes than our assembly. The differences in the high-quality assemblies may reflect differences in genome content due to the synthetic Da-Ae having had fewer generations in which to "purge" extra material (resulting in, for example, larger number of bases and more duplicated BUSCOs) or could reflect differences resulting from the assembly process. It is not possible to distinguish between these causes with our available data. The improved assembly enabled by third generation sequencing technologies will serve as an excellent resource for *B. napus* geneticists and scientists aiming to identify genes underlying agronomic traits.

Homoeologous exchange is a biological process observed in allopolyploids, like *B. napus*, where highly similar yet different regions of the two diploid subgenomes exchange genetic material with one another. The result is new chromosome structures that, while being primarily

composed of one ancestral genome, now also contain regions belonging to a different ancestral genome. To investigate the occurrence of homoeologous exchange in Da-Ae, we investigated both genome coverage and synteny across the genomes of *B. napus* Da-Ae, and nine other cultivars. Our results indicate that homoeologous exchange has occurred in both small and large regions throughout the whole genome. Each cultivar of *B. napus* had many unique homoeologous exchange events. More surprising was that there are multiple regions of homoeologous exchange that are shared among the *B. napus* cultivars. It is possible that these homoeologous exchange regions are shared among multiple varieties because specific combinations of homoeologous genes affect plant fitness or agroeconomic traits. Alternatively, these sites could be shared because sequence homology and chromosome topology favors recombination at these sites. These findings further build upon the previous work done to identify hotspot regions (Higgins *et al.* 2018).

In conclusion, using several sequencing technologies, we created a genome assembly similar in quality to other recently published assemblies that used third generation sequencing, allowing for an improvement upon the original Darmor-bzh v4.1 published assembly. We also identified potential hotspots of homoeologous exchange along with single-copy BUSCOs that are shared among different cultivars of *B. napus*. Our assembly and analysis of Da-Ae is another step forward toward the realization of a pan-genome for *B. napus*.

Contributing Authors

Work on this project was also completed by Ruijuan Li, Seungmo Kim, Richard Michelmore, Shinje Kim, and Julin N. Maloof

Data availability

All raw reads and the nuclear genome assembly and annotation are available at NCBI under BioProject PRJNA627442. Analysis code is available at https://github.com/MaloofLab/Davis_B_napus_assembly_2023.

Acknowledgments

I would like to thank the members of the Michelmore Lab (UC Davis), especially Kyle Fletcher, Will Palmer, and Sebastian Reyes Chin Wo, for countless hours of advice and support throughout this project.

Genome Report: A chromosome-level genome assembly of the varied leaved jewelflower, Streptanthus diversifolius, reveals a recent whole genome duplication

Abstract

Streptanthus diversifolius, a varied leaved jewelflower, is a member of the Brassicaceae family. More specifically, it is a member of the 'Streptanthoid Complex' which is a collection comprised primarily of the *Streptanthus* and *Caulanthus* genera located in the Thelypodieae tribe. The Streptanthoid Complex spans the full range of the California Floristic Province including desert, foothill, and mountain environments. The ability of these related species to radiate into dramatically different environments makes them a desirable study subject for exploring how plant species expand their ranges and adapt to new environments over time. Ecological and evolutionary studies for this complex have revealed fascinating variation in serpentine soil adaptation, defense compounds, germination, flowering, and life history strategies. Currently, a lack of available genomic resources has hindered the ability to relate these phenotypic observations to their underlying genetic and molecular mechanisms. To help remedy this situation we present here a chromosome-level genome assembly of *S. diversifolius,* a member of the Brassicaceae family, developed using Illumina, Hi-C, and HiFi sequencing technologies. Construction of this assembly also provides further evidence to support the previously reported recent whole genome duplication unique to the Thelypodieae tribe. This whole genome duplication may have provided individuals in the Streptanthoid Complex the genetic arsenal to rapidly radiate throughout the California Floristic Province and to occupy commonly inhospitable environments including serpentine soils.

Introduction

The California Floristic Province (CFP) is a region comprising much of the state of California west of the Sierra Nevada Mountains, and includes portions of Oregon and Baja California (Howell, 1957). It is a hotspot for biodiversity due to its Mediterranean-type climate with warm dry summers and cool wet winters. The CFP is home to more than 700 genera and more than 4000 vascular plants native to California. Among this collection is the 'Streptanthoid Complex' which is a set of ca. 60 taxa belonging to the Thelypodieae tribe (Burrell, 2010; Burrell et al., 2011; Burrell and Pepper, 2006). This complex is comprised primarily of the *Streptanthus* and *Caulanthus* genera, and while the members are genetically closely related, they occupy a large range of environments including the deserts, mountains, and foothills of the CFP (Al-Shehbaz, 2010).

Streptanthus is well known for extreme edaphic specialization, particularly to infertile serpentine soils. Serpentine is a unique soil known for its extraordinarily low Ca to Mg ratio, low levels of micronutrients, elevated heavy metal concentrations, and poor water retention (Brady et al., 2005; Kruckeberg, 2006). Serpentine soils have a profound effect on plant ecology and evolution, often supporting unique plant communities of serpentine tolerant species, as well as serpentine endemics that occur on no other soil type. Serpentine soil usage has four or five independent origins in the Streptanthoid Complex (Cacho and Strauss, 2014) and has led to more than half of the ~35 known species in the complex to be endemic to serpentine soils (Al-Shehbaz, 2010; Baldwin et al., 2012; Safford et al., 2005). The genus has been the subject of several classic studies on the evolution of edaphic specialization and is an emerging model system for understanding the evolution of edaphic specialization in herbaceous plants.

Streptanthus has also been the subject of molecular phylogenetic investigations, providing a robust phylogenetic context in which to examine the evolution of specific traits during diversification (Cacho et al., 2021; Cacho and Strauss, 2014; Weber et al., 2018).

To improve genomic resources for research in *Streptanthus* and its allies, we carried out whole-genome shotgun sequencing, Hi-C sequencing, and PacBio HiFi sequencing on the nuclear genome of the varied leaved jewelflower, *S. diversifolius.* While *S. diversifolius* is not found on serpentine soils, it has the smallest genome known in the genus (0.36 GB), and is a known diploid (2n=28) (Cacho et al., 2021). A reference genome for *S. diversifolius* will allow for gene discovery, genetic mapping, and re-sequencing of other *Streptanthus* species, all in support of studies that are ongoing. Being a member of the Brassicaceae means that the reference genome sequence described here will allow mechanistic knowledge from Arabidopsis and Brassica to be leveraged for understanding the molecular basis of serpentine specialization, heavy metal tolerance and hyperaccumulation, and climate adaptation. Such insights may lead to improvements of closely related crops in the Brassicaceae family, as well as industrial applications in phytoremediation and phytomining. Additionally, a greater understanding of the *Streptanthus* clade will allow us to build upon previously work including the evolution of glucosinolate defense (Cacho et al., 2015) and the seasonal germination niche observed across an elevational gradient (Gremer et al., 2020a). Combined these analyses will assist in the conservation and management of these species, many of whom are currently endangered or threatened.

Methods & Materials

Plant collection, DNA isolation, and Sequencing

Plant collection, DNA isolation, and sequencing occurred in three different rounds. In the first round, several whole, flowering plants of *S. diversifolius* were collected in April 2013 from a naturally occurring population on Table Mountain, Butte County, California (D. O. Burge 1389; voucher deposited at DAV). Tissues of these plants were dried on silica gel desiccant at room temperature before being returned to the lab at the University of British Columbia. Total genomic DNA was extracted from the leaves, stems, and unopened flower buds of a single plant using the Qiagen (Limburg, Netherlands) DNeasy Kit according to the manufacturer's instructions. A total of 24 extractions were performed. Pooled extractions were then concentrated by adding 1/10 volume of 3M sodium acetate, pH 5.2, and 2 volumes of -20°C 95% EtOH. After centrifugation, the DNA pellet was dried and resuspended in pure, nuclease-free water. Aliquots of the same DNA preparation were then used to construct four sequencing libraries according to the manufacturer's protocols (Illumina, http://www.illumina.com) at the Innovation Centre, Genome Quebec. Library insert sizes were chosen to be compatible with the Allpaths-LG genome assembler (Gnerre et al., 2011). We prepared two libraries designed to overlap when sequenced as paired ends: a 180 bp library sequenced on an Illumina HiSeq with 2x100bp reads, and a 450 bp library sequenced on an Illumina MiSeq with 2x300bp reads. We also prepared two mate-pair libraries with insert sized of approximately 4700 bp and 8200 bp.

In the second round, seeds from *S. diversifolius* were collected in 2016 from a naturally occurring population on Table Mountain, Butte County, California. It should be noted that this is a separate seed collection from that described above. Eight cones containing a mixture of 50% Ron's Mix and 50% sand were saturated with nutrient water. A small divot was then made in

each cone where 3-4 seeds were placed before being covered with a small amount of the soil mixture. The cones were then placed in a rack and covered with plastic wrap to prevent the top layer of the soil from drying out. The covered rack was then placed in a growth chamber set to 22*C with a light/dark cycle of 12/12. Two weeks after being placed in the growth chamber, the plastic wrap was removed, and the recently germinated seedlings were exposed to the air. The seedlings remained in the growth chamber with nutrient water being provided every other day. Once the seedlings on average had attained approximately 8-10 true leaves, two plants were randomly selected for Hi-C sequencing. Young leaves from each plant were collected separately until approximately 0.5 grams were obtained. The leaves from each plant were than processed separately using Phase Genomics' Proximo Hi-C Kit version 3.0 following the standard plant protocol. The libraries where then sent to the University of California, Davis Genome Center where 150bp paired-end sequencing was performed using an Illumina NovaSeq 6000. A total of ~240 million read pairs were sequenced resulting in ~200X coverage of the genome.

In the third round, leaf tissue from three different *S. diversifolius* individuals located at Table Mountain, Butte Country, CA were harvested individually in May 2023. The three samples weighing 0.66g, 0.59g, and 0.52g were placed on ice and transported to UC Davis where they were flash frozen in liquid nitrogen upon arrival. All three samples were then sent to the UC Davis genome center for HMW DNA extraction. DNA was able to be successfully extracted from the 0.66g sample and was used for sequencing and the remaining two leaf samples were kept for backup. Sequencing was also performed at the UC Davis genome center on PacBio's Revio sequencing platform. The reads were then processed using SMRT Link version 12.0.0.177059

resulting in 4,689,053 HiFi reads with a mean read length of 11,052 base pairs and mean read quality of 30. The HiFi yield of 51.8 Gb represents a genome coverage of ~144X.

Genome Assembly

Starting with the Illumina short reads from round one, raw sequence reads were filtered and trimmed of their adapter sequences using Trimmomatic (Bolger et al., 2014). Trimmed and filtered reads were assembled using Allpaths-LG with the haploidify=T parameter. Biological contaminants in the resulting assembly were identified against the NCBI NT database using blastn megablast (Sayers et al., 2022). Seven scaffolds with a match of at least 85% and at least 300 bp to a database sequence belonging to non-plant taxa were removed from the Allpaths-LG assembly. Artificial contaminants were identified using NCBI's VecScreen protocol. Contaminants at scaffold ends were removed; artifacts within scaffolds were masked.

To improve the contiguity of the assembly from round one, the Hi-C sequencing data from round two was added to the assembly. The Hi-C reads were trimmed using Trimmomatic version 0.39 in paired end mode with the parameters ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 resulting in ~180 million surviving read pairs. The reads were then mapped to round one scaffolds using the Arima-HiC Mapping Pipeline (Arima Genomics, n.d.). Following mapping, the alignment file was sorted by read name using Samtools (Li et al., 2009) sort. The sorted alignment file along with the scaffolds were input into the Hi-C scaffolding program YaHS (Zhou et al., 2022).

The Hi-C scaffolded assembly showed signs of a possible whole genome duplication along with highly probable mis-joins of the contigs. To investigate these two occurrences, a new

genome assembly was created using the HiFi sequencing data from round three. The HiFi reads were converted from BAM format to FASTQ format for assembly. The FASTQ reads were assembled using the program HiFiasm version 0.19.5-r593 (Cheng et al., 2021) using default parameters with the –primary flag selected. The genome assembly created using HiFiasm was used for future analyses presented here.

Repeat and gene annotation

The assembly was annotated using the MAKER pipeline (Cantarel et al., 2008; Holt and Yandell, 2011). First, a custom repeat library for *S. diversifolius* was made using the Maker-P pipeline (Campbell et al., 2014b, 2014a). Augustus (Keller et al., 2011) retraining parameters were also calculated using BUSCO v4.1.4 (Manni et al., 2021) in long and genome mode and the brassicales_odb10 lineage dataset. The assembly was input into Maker v3.01.04 along with the repeat library, Augustus retraining parameters, and transcripts from six *Streptanthus* clade species. A second round of Maker was performed using same inputs except this time a newly made *S. diversifolius* hmm file made using Snap (Korf, 2004) v2006-07-28 and the GFF file created in the first round of Maker were included. In the second round of Maker, the parameter of AED=0.5 was also set, previously AED=1, to remove transcripts which had weak evidence support. A total of 40,606 gene models were created. These gene models were post processed following the MAKER Tutorial for WGS Assembly and Annotation Winter School 2018 (https://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018) where they were aligned to the UniProt database using blastp and processed using interproscan (Jones et al., 2014). These results were then incorporated into the final annotations.

Alignment to A. thaliana

To assess the contiguity of the Hi-C and HiFi draft assemblies, the *S. diversifolius* genomes were aligned to a reference *A. thaliana* genome (TAIR 10)(Berardini et al., 2015b). The genomes were aligned in protein space using the promer alignment tool contained within the mummer software package (Marçais et al., 2018). Promer was run using the `maxmatch` flag and the default setting for the other parameters. Following alignment, the delta files were then filtered using delta-filter using minimum alignment length and sequence identity cutoffs of 1000 bp and 85% respectively. The filtered delta files were then plotted using mummerplot. Points in the plots were colored based on the ancestral crucifer karyotype (ACK) blocks (Lysak et al., 2016) defined within *A. thaliana.*

Genomic data collection

To investigate the possibility of a whole genome duplication, comparative genomic analyses were performed. The genome data and annotation files of *Caulanthus amplexicaulis*, *Stanleya pinnata*, *Arabidopsis thaliana* (Cheng et al., 2017), *Arabidopsis lyrata* (Hu et al., 2011) were downloaded from Phytozome (https://phytozome-next.jgi.doe.gov/) (Goodstein et al., 2012). *Cardamine hirsute* genome (Gan et al., 2016) was obtained from Cardamine hirsuta Genetic and genomic (http://chi.mpipz.mpg.de/index.html). *Euclidium syriacum* was downloaded from NCBI under accession number (GCA_900116095.1). *Draba nivalis* genome (Nowak et al., 2021) data was downloaded from Dryad (https://datadryad.org). *Arabis alpina* genome (Jiao et al., 2017)was downloaded from Genomic resources for Arabis alpina (http://www.arabis-alpina.org/index.html). *Thlaspi arvense* genome (Geng et al., 2021) was

downloaded from Pennycress Home (http://pennycress.umn.edu/). *Schrenkiella parvula* genome (Dassanayake et al., 2011) was downloaded from thellungiella.org. *Brassica oleracea* genome was downloaded from Genoscope (http://www.genoscope.cns.fr/externe/plants/chromosomes.html). *Raphanus raphanistrum* ssp. Raphanistrum genome (Zhang et al., 2021) was downloaded from Genome Warehouse (https://ngdc.cncb.ac.cn/gwh/#) under accession number PRJCA003033. *Aethionema arabicum* genome (Fernandez-Pozo et al., 2021) was downloaded from Ae. arabicum DB (https://plantcode.cup.uni-freiburg.de/aetar_db/downloads.php)

Species tree construction

To construct a species tree, we identified single-copy orthogroups from the proteins of 14 genomes, including *S. diversifolius*, using OrthoFinder (Emms and Kelly, 2019). This identified 118 single-copy orthogroups that were then aligned using MAFFT (Katoh and Standley, 2013). Pal2nal was used to perform codon alignment (Suyama et al., 2006). TrimAl was used to trim poorly aligned segments from the alignments (Capella-Gutiérrez et al., 2009). Based on high-quality alignments, the phylogenetic trees for the 118 single-copy orthogroups were constructed using the IQ-TREE (Minh et al., 2020). Finally, the species tree was inferred using ASTRAL-III (Zhang et al., 2018).

Identification of whole genome duplication

To identify whole-genome duplication events in *Streptanthus*, we compared the *S. diversifolius* genome with 13 other genomes. Initially, the DupGene_Finder pipeline (Qiao et al., 2019) was used to identify WGD genes across the 14 genomes, including *S. diversifolius*, while

removing tandem genes. Subsequently, the established calculate_Ka_Ks_pipeline (Qiao et al., 2019) was utilized to calculate the number of substitutions per synonymous site (Ks) values for WGD pairs. We made some modifications to the previous Ks fitting process (identify_Ks_peaks_by_fitting_GMM) to allow fitting based on the Ks values of gene pairs. To place the WGD events, based on the phylogenetic relationships of the 14 species, we identified orthologs between *S. diversifolius* and *C. amplexicaulis* and orthologs between *S. diversifolius* and *S. pinnata*, calculated the Ks values for these orthologs, and plotted the Ks distribution. Additionally, we used MCscan (Tang et al., 2008) (python version, https://github.com/tanghaibao/jcvi) to create dot-plots for *S. diversifolius* vs *S. diversifolius* and *S. diversifolius* vs *A. thaliana* to further confirm the recent WGD events. The timing of the WGD events was estimated to be between 11.54 to 14.06 million years ago (MYA), based on the formula T=Ks/2r (r=7.1 × 10−9 ± 0.7 × 10−9) (Ossowski et al., 2010). To look for evidence of allotetraploidization, SubPhaser (Jia et al., 2022) was run with default parameters (k = 15, q = 200).

Gene family evolution

To identify expanded and contracted gene families, the CAFE5 (Mendes et al., 2021) software was utilized. Following the CAFE5 documentation, orthogroups previously identified by OrthoFinder with more than 100 gene copies in one or more species were excluded in downstream analysis. The previously described species tree was transformed into an ultrametric tree, using MCMCtree which is incorporated in PAML v4.10.7 (Yang, 2007), utilizing the age of the most recent common ancestor of *A. arabaicum* and the remaining 13 other species as 35.2 million years ago (MYA), based on a previous study (Nowak et al., 2021). The

ultrametric tree and the filtered orthogroups were then used as input for CAFE5. Testing was performed on different -k parameters to determine the optimal K value based on the Model Gamma Final Likelihood value, where the best -k value was found to be 2. Finally, significantly expanded and contracted families were extracted from the result files.

Positive selection tests

The identification of positively selected genes (PSGs) in Streptanthus was conducted using the branch-site model in codeML implemented in PAML v4.10.7 (Yang, 2007). Due to a relatively small number of single-copy orthogroup (only 118 single-copy orthogroups), low-copy genes were used to identify PSGs. Initially, orthogroups with less than three copies across all species were identified based on OrthoFinder results. Codon alignments were then performed using the OMM_MACSE pipeline within MACSE v11.05 (Ranwez et al., 2018), which employs MAFFT (Katoh and Standley, 2013) and PRANK (Löytynoja, 2014) for codon alignment. Based on codon alignments from the OMM_MACSE pipeline, GWideCodeML v1.1 (Macías et al., 2020) was used for genome-wide PSG identifications. Briefly, using the species tree, Streptanthus was set as the foreground branch. Sites showing significant positive selection were identified by comparing the likelihood ratio test (LRT) values between the alternative model and the null model. GWideCodeML automates these steps and the comparison, ultimately identifying significantly selected sites and genes. A gene is considered a PSG only if it has an p-value less than 0.05 and also has significantly selected sites.

Results

The construction of this new *S. diversifolius* whole genome assembly took place over multiple rounds, with each round producing an improved assembly compared to its predecessor. Starting with an Illumina short read assembly and progressing to a HiFi assembly, the results of each iteration are described below. Following the assembly are the results of analyses investigating the presence of a suspected whole genome duplication and positive selection of gene families.

Round 1: Illumina Assembly

The four Illumina sequencing libraries, HiSeq with 2x100bp reads, MiSeq with 2x300bp reads, and mate-pair libraries with insert sizes of approximately 4700 bp and 8200 bp were used for the initial assembly. Assembly was performed using the whole-genome shotgun assembler ALLPATHS-LG and subsequently postprocessed using NCBI's VecScreen protocol. The resulting assembly had a total length of 314 Mb, a scaffold N50 of 470 Kb, and was comprised of 4,627 scaffolds (Table 2.1). Genome completeness of the assembly using BUSCO version 5.5.0 and the embryophyta odb10 database found the assembly to have a complete BUSCO percentage of 98.7% with an approximate even split between single-copy and duplicate BUSCOs (Table 2.2).

Round 2: Hi-C Assembly

Looking to improve the contiguity of the Illumina assembly, two Hi-C sequencing libraries were prepared using tissue from two separate *S. diversifolius* plants whose seeds came from a population in the same geographic location as the original plant samples. Both sequencing libraries were aligned to Illumina assembly using the Aria HiC Mapping Pipeline and the

resulting alignment files were input into the Hi-C scaffolding program YaHS. The scaffolded assembly showed signs of improvement compared to the original Illumina assembly. The scaffolding resulted in the overall size of the assembly increasing slightly due to the introduction of scaffolding gaps. The main improvements were a decrease in the number of scaffolds from 4,627 to 3,920 scaffolds and an increase in the scaffold N50 from 470 Kb to 21Mb. Additionally, 17 scaffolds were greater than 1Mb in length and encompassed 89% of the total assembly length (Table 2.1). Scaffolding did not have a significant effect on BUSCO composition and the majority of the BUSCOs were found to be contained in the 17 previously mentioned scaffolds (Table 2.2).

To examine the quality and contiguity of the scaffolded assembly, the assembly was aligned to a reference *A. thaliana* genome in protein space using promer from the mummer bioinformatic tool suite. The resulting plot displayed a mirroring pattern whereby scaffolds had apparent end-to-end joining of duplicate chromosomes (Figure S2.1). This mirroring pattern suggested the possibility of two different phenomena occurring. The first was a potential whole genome duplication due to the frequency and size of the mirrored regions and the second was the presence of mis-joins in the assembly given the proximity of the mirrored regions to their counterparts. While the latter was deemed to be most likely a scaffolding artifact, the former was found to be biologically plausible given the frequency of recent whole genome duplications found across the different tribes of the Brassicaceae family (Kagale et al., 2014; Mandáková et al., 2017).

Round 3: HiFi Assembly

To further investigate the possibility of a whole genome duplication and correct mis-joins created in the Hi-C scaffolding process, a third round of sequencing was performed. Leaf tissue was collected from individuals located in the same geographic region as the first two rounds and extracted high molecular weight DNA was prepped and sequenced on the PacBio Revio sequencing platform. Following postprocessing, the HiFi reads were assembled using the long read assembler HiFiasm. The HiFi assembly was an improvement over the assemblies created in both rounds one and two. Total sequence length increased to 402 Mb with 86.0% of the total sequence length contained in the 18 scaffolds whose length was greater than 1 Mb. Along with the increase in assembly size, the total number of scaffolds dropped to 1,193 and the scaffold N50 increased to 23.5 Mb (Table 2.1). The HiFi assembly also saw an improvement in complete BUSCOs and now showed slightly more duplicated than single-copy BUSCOs. All BUSCOs were also contained within the previously mentioned 18 scaffolds (Table 2.2). When compared to the round one assembly, most of the sequence is shared between the two assemblies, with the main difference being the contiguity of scaffolds (Figure 2.1).

Table 2.1. *Assembly statistics for each round of assembly.*

Assembly	Number of Scaffolds	Number of Contigs	Total Length (bp)	Percent gaps	Scaffold N50	Contigs N50
Round 1	4,627	20,342	314,016,962	10.98 %	470 Kb	49 Kb
Round 2 Full	3,920	20,571	314,205,162	11.04 %	21 Mb	49 Kb
Round 2 1Mb	17	13,149	280,150,441	8.78 %	22 Mb	55 Kb
Round 3 Full	1,193	1,193	402,145,164	0 %	23.5 Mb	23.5 Mb
Round 3 1Mb	18	18	345,697,404	0 %	24 Mb	24 Mb
Arabidopsis (TAIR10)	7	99	119,668,634	0.16%	23 Mb	11 Mb

Table 2.2. *BUSCO summary statistics. Analysis completed using BUSCO version 5.5.0 in genome mode and the embryophyta odb10 dataset.*

Assembly	Complete %	Single-Copy %	Duplicated %	Fragmented %	Missing %	# BUSCOS
Stage 1	98.7	50.6	48.1	0.7	0.6	1614
Stage 2 Full	98.8	50.8	48	0.7	0.5	1614
Stage 2 1Mb	98	51	47	0.7	1.3	1614
Stage 3 Full	99.6	47.1	52.5	0.1	0.3	1614
Stage 3 1Mb	99.6	47.1	52.5	0.1	0.3	1614
Arabidopsis (TAIR10)	99.3	98.6	0.7	0.2	0.5	1614

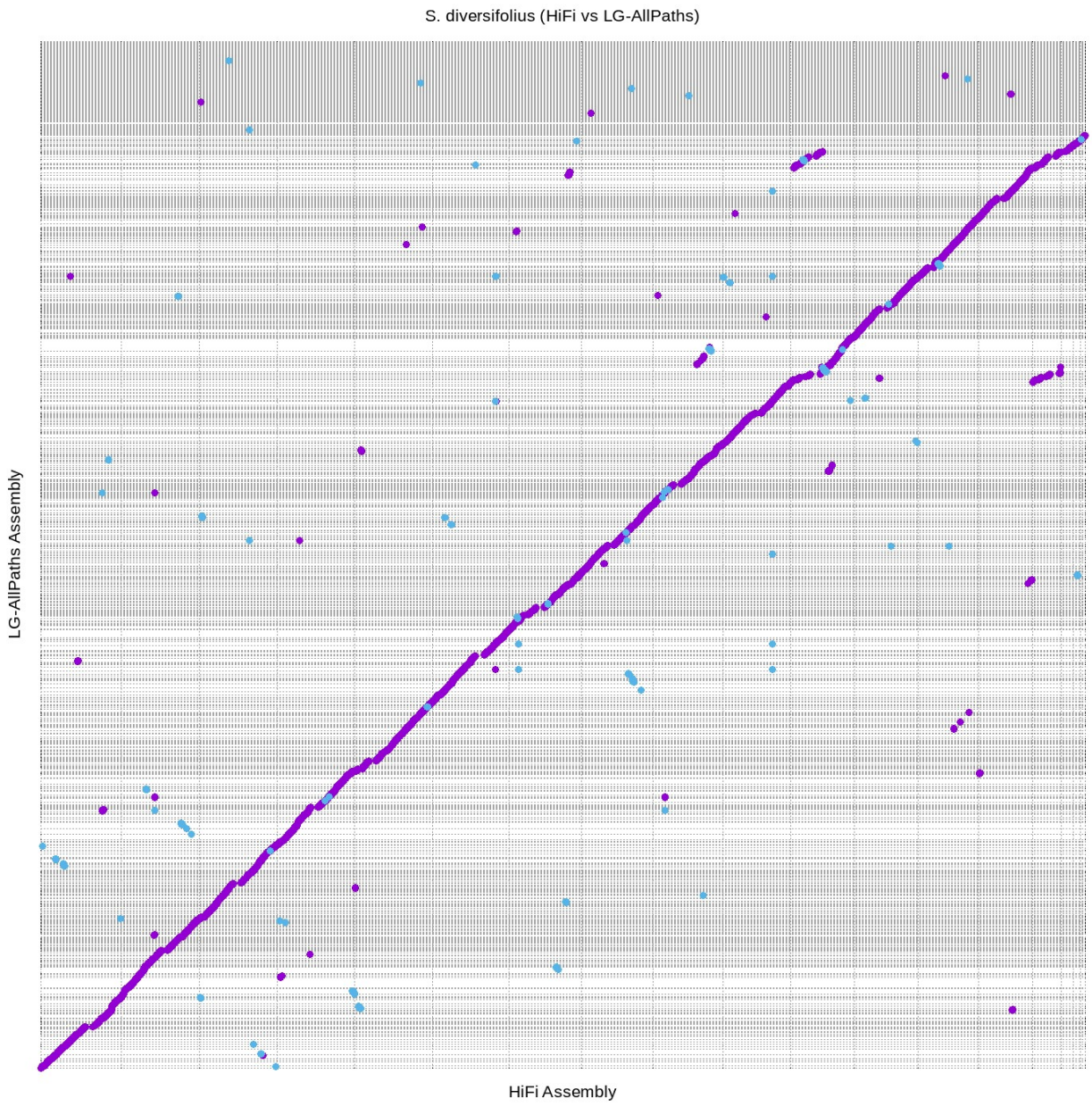

Figure 2.1. *All scaffolds greater than 1 Mb in the HiFiasm assembly aligned to all scaffolds in the LG-AllPaths assembly. Alignment was performed using Nucmer and plotted using Mummerplot, both programs included in the Mummer bioinformatic toolkit. Purple lines indicate regions of alignment in the forward direction and blue lines indicate regions of alignment in the reverse direction. Gray dashed lines indicate separate scaffold in each assembly.*

Annotation was completed using the Maker-P pipeline. Maker analysis of the HiFi assembly from round 3 predicted 40,606 protein-coding genes after filtering. Of these 40,606 protein-coding genes, 34,866 are found in the 18 largest scaffolds.

Comparison of Round 2 and 3:

The HiFi assembly was created to correct mis-joins created in the Hi-C scaffolding process and validate the existence of a recent whole genome duplication. When aligned against one another, the Hi-C scaffolded and HiFi assembly were mostly congruent. (Figure 2.2). The HiFi assembly was aligned to *A. thaliana* in the same way as the Hi-C scaffolded assembly. The mirroring pattern observed in the Hi-C scaffolded assembly was no longer present. Evidence of a recent whole genome duplication was present as seen by the presence of duplicate ancestral genomic regions spread throughout the genome (Figure 2.3).

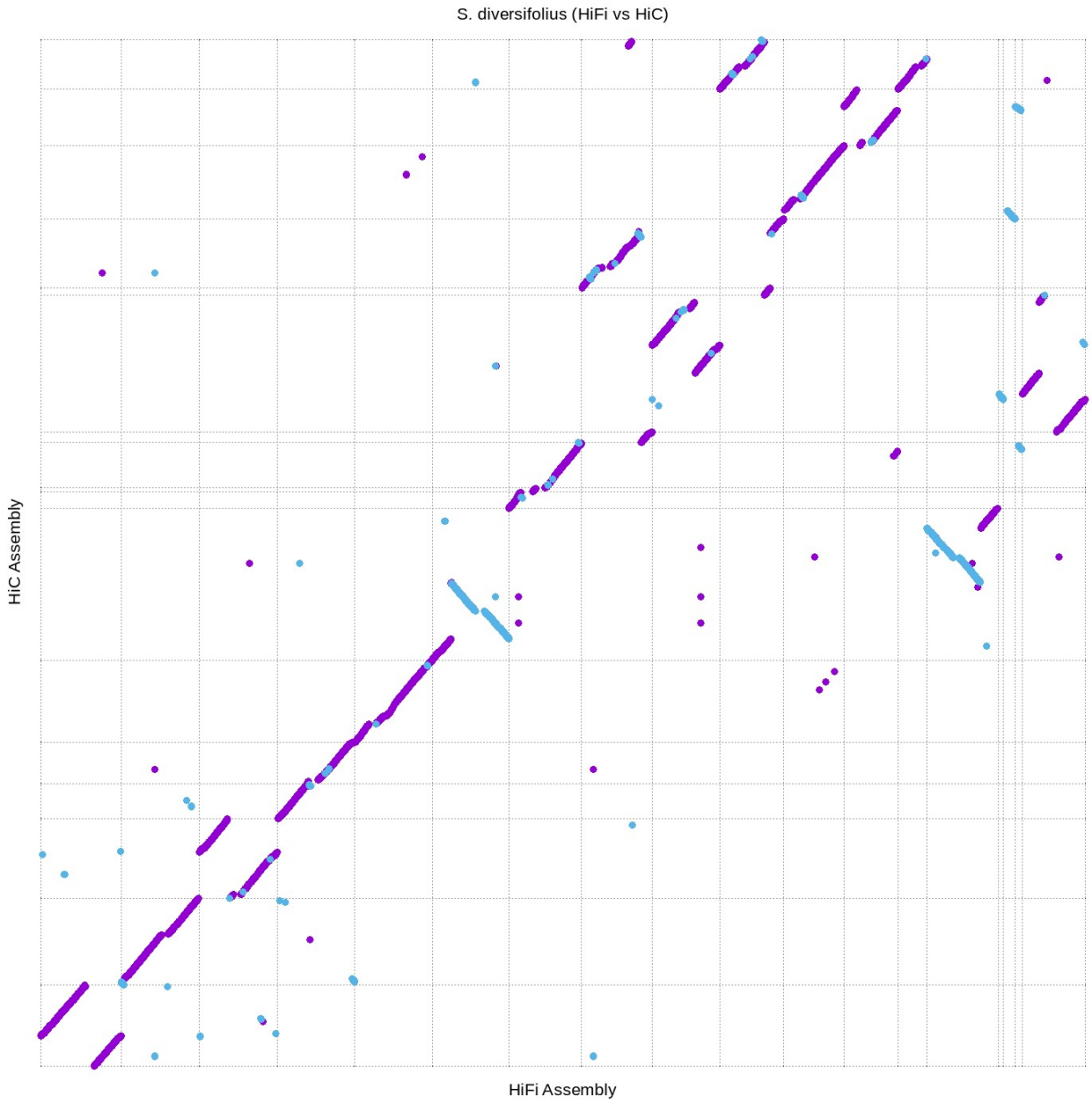

Figure 2.2. *All scaffolds greater than 1 Mb in the HiFiasm assembly aligned to all scaffolds greater than 1 Mb in the HiC assembly. Alignment was performed using Nucmer and plotted using Mummerplot, both programs included in the Mummer bioinformatic toolkit. Purple lines indicate regions of alignment in the forward direction and blue lines indicate regions of alignment in the reverse direction. Gray dashed lines indicate separate scaffold in each assembly.*

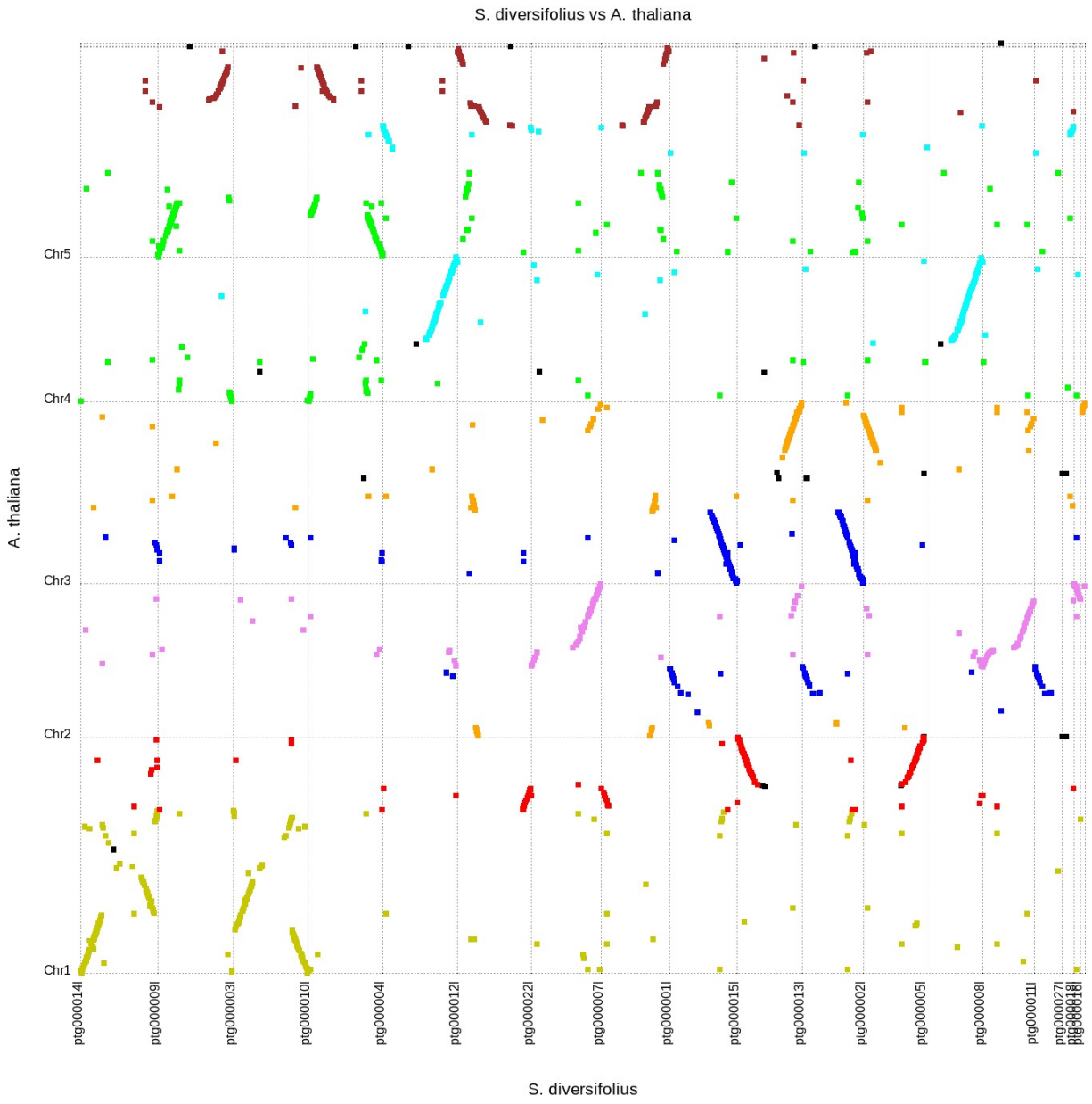

Figure 2.3. *All scaffolds greater than 1 Mb in the HiFiasm assembly aligned to the TAIR 10 A. thaliana genome assembly. Alignment was performed using Promer and plotted using Mummerplot, both programs included in the Mummer bioinformatic toolkit. The color of each line corresponds to the 8 ancestral crucifer karyotype (ACK) blocks of A. thaliana with the exception of black which corresponds to regions which do not belong to an ACK block. ACK block boundaries are based on gene locations summarized in Lysak et. Al 2016. Gray dashed lines indicate separate scaffold in each assembly.*

Whole Genome Duplication

Our genome alignment plot provides additional evidence of a whole genome duplication shared among members of the Thelypodieae tribe. To investigate this further, we examined the

distribution of the synonymous substitution rate (Ks) among paralogs. In the absence of whole genome duplication events, the Ks distribution is expected to show a rapid decrease over time (Lynch and Conery, 2000). Additional peaks of in the Ks distribution indicate whole genome duplication events and can be used to estimate their age (Maere et al., 2005; Ohno, 1970). Analysis of the Ks peaks for internal WGD gene pairs in *S. diversifolius* and the previously mentioned genomes revealed that *S. diversifolius* has a relatively new WGD with a Ks of ~0.18, along with two older peaks at 0.92 and 2.03, likely corresponding to the Brassicaceae α WGD and the Eudicots γ WGT, respectively (Figure 2.4). Our constructed phylogenetic tree (Figure 2.5) which agrees with other phylogenetic analyses (Ivalú Cacho et al., 2014; Kagale et al., 2014; Mandáková et al., 2017) suggests *S. diversifolius, C. amplexicaulis*, and *S. pinnata* are closely related and both *C. amplexicaulis* and *S. pinnata* also exhibit a Ks peak around 0.18. No other species in our analysis shared the Ks peak around 0.18. To determine the species divergence time and ascertain if it is the same WGD we calculated the Ks values between *S. diversifolius* and *S. pinnata* and between *C. amplexicaulis* and *S. diversifolius*. These species show a divergence Ks peak of about 0.08, which is later than the WGD event. Therefore, it appears this WGD, likely occurring around 15 MYA, is shared by *S. diversifolius* and other members of the Thelypodieae.

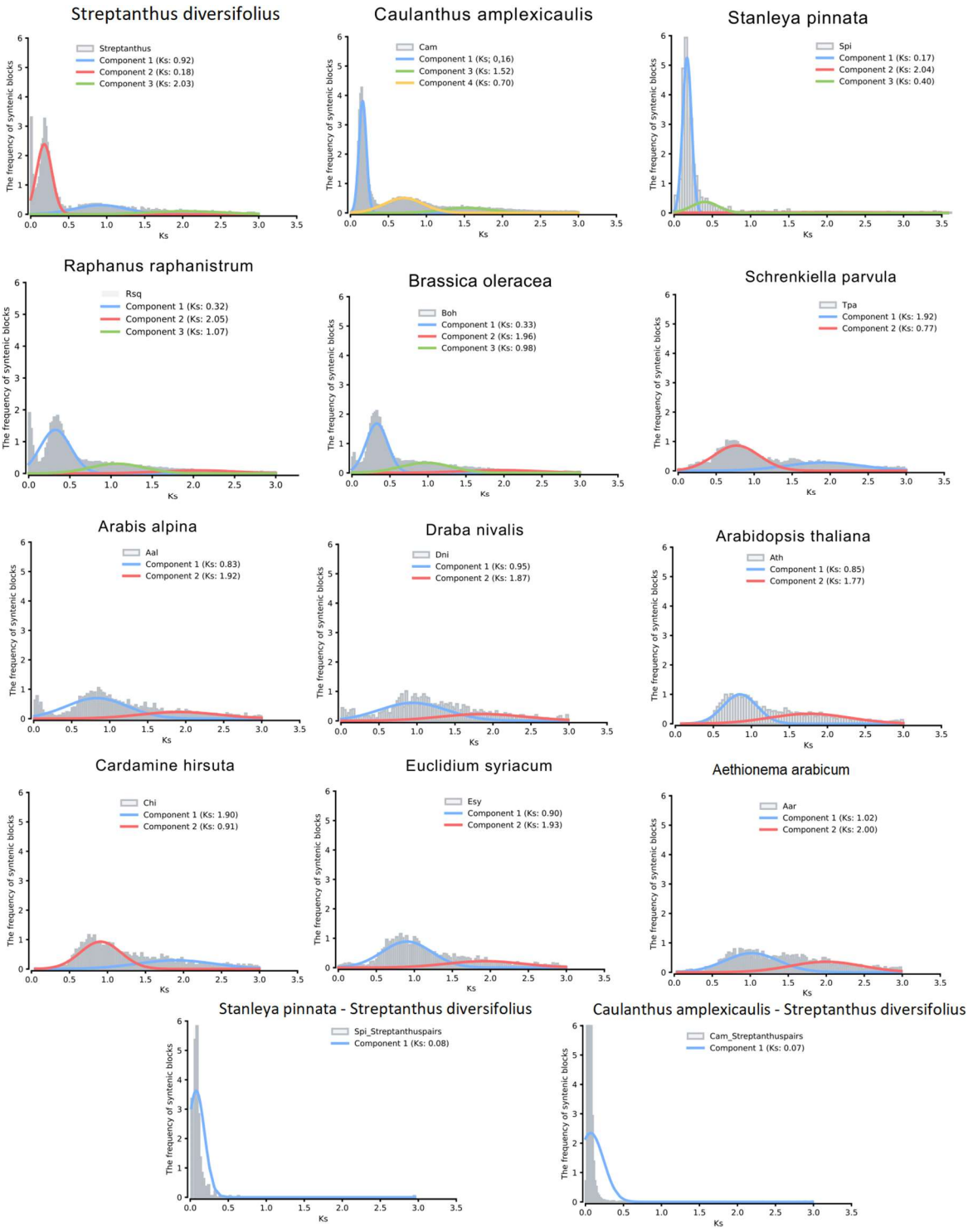

Figure 2.4. *Top plots are intragenomic analysis of Ks for the analyzed species (excluding T. arvense and A. Lyrate). Colored lines indicated distinct Ks peaks in each plot. Bottom plots are intergenomic analysis of Ks between S. diversifolius and either S. pinnata or C. amplexicaulis*

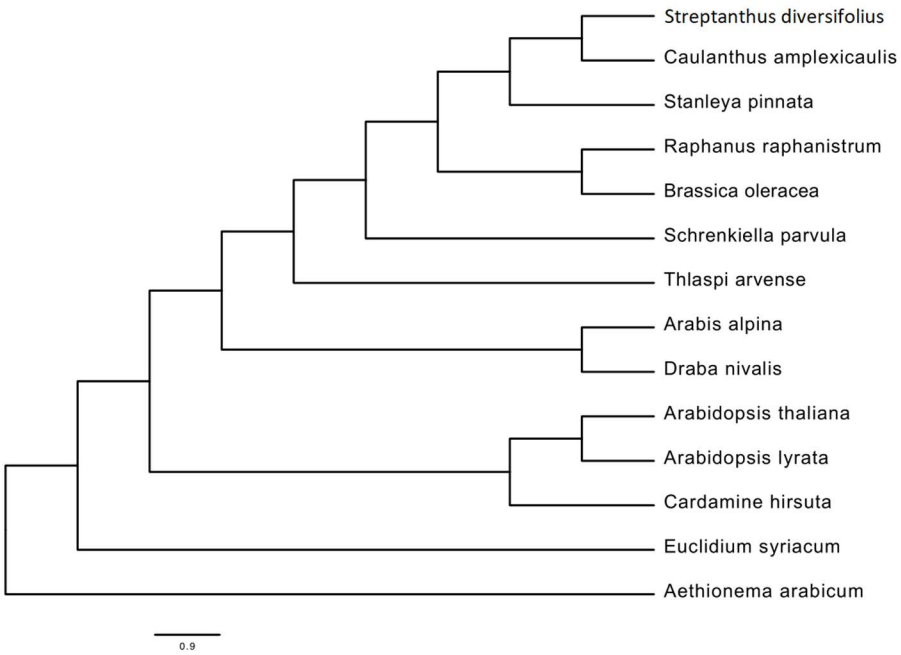

Figure 2.5. *Phylogenetic species tree. Tree was inferred using phylogenetic trees from the 118 single-copy orthogroups identified between our 14 species of interest.*

Whole genome duplication can occur by diploidization of either auto-tetraploids formed by non-disjunction or allo-tetraploids resulting from hybridization of closely related species. These two possibilities can be distinguished by comparing repeat content of homoeologous chromosomes. In the case of autopolyploids the repeats should be similar on homoeologous pairs. However, for allotetraploids it is expected that the repeat sequences will have diverged over evolutionary time in the two progenitors. To analyze which of these scenarios is most likely for the WGD described here, we used SubPhaser (Jia et al., 2022) to look for kmers that could distinguish between homeologs. Of 4,425,565 kmers identified, only 88 were unique to one homeolog among a pair, and there was no significant enrichment of unique kmers on any homeolog. This suggests that the whole genome duplication results from an autopolyploid event (or from hybridization of two very closely related species).

Gene family expansion and duplicate gene retention

Confident in the presence of a WGD we further looked at the composition of the genome to assess if there were gene families that showed significant expansion, or gene categories with preferential retention of both duplicates. CAFE5 analysis reported a total of 19 expanded gene families containing a total of 295 genes (Figure 2.6). GO and KEGG enrichment analysis were performed on this set of genes. GO analysis revealed an enrichment for several biological processes related to ATP. KEGG analysis also showed an enrichment for ATP related processes with the largest most significant category being F-type H+/Na+-transporting ATPase subunit alpha-EC:7.1.2.2 7.2.2.1. The enrichment of ATP related terms is of interest given serpentine soil habitats occupied by many *Streptanthus* species are characterized as having higher levels of magnesium along with elevated levels of other heavy metal which are toxic to most plants (Brady et al., 2005). Chelation of nucleotides by magnesium (Mg) is an essential feature of cell metabolism with adenylates being the most abundant (Kleczkowski and Igamberdiev, 2021). The excess Mg of the serpentine soil needs to be managed to maintain a $[Mg^{2+}]$ that allows the plant to perform metabolic tasks without damage from Mg^{2+} toxicity. One way that plants avoid Mg^{2+} toxicity is by utilizing the CBL-CIPK network which allows cells to sequester excess Mg^{2+} into vacuoles (Tang et al., 2015). Along with enrichment among gene families, GO and KEGG enrichment analysis were also performed on WGD-derived gene pairs identified through the DupGen_finder pipeline. Of note, KEGG analysis identified 3 categories significantly enriched including calcium-dependent protein kinase EC:2.7.11.1 (Figure 2.7). Enrichment of these genes may be benefiting the plant as it negotiates a Mg rich environment

potentially by being involved in the CBL-CIPK network or improving signaling to maintain appropriate [Mg^{2+}] in the cell.

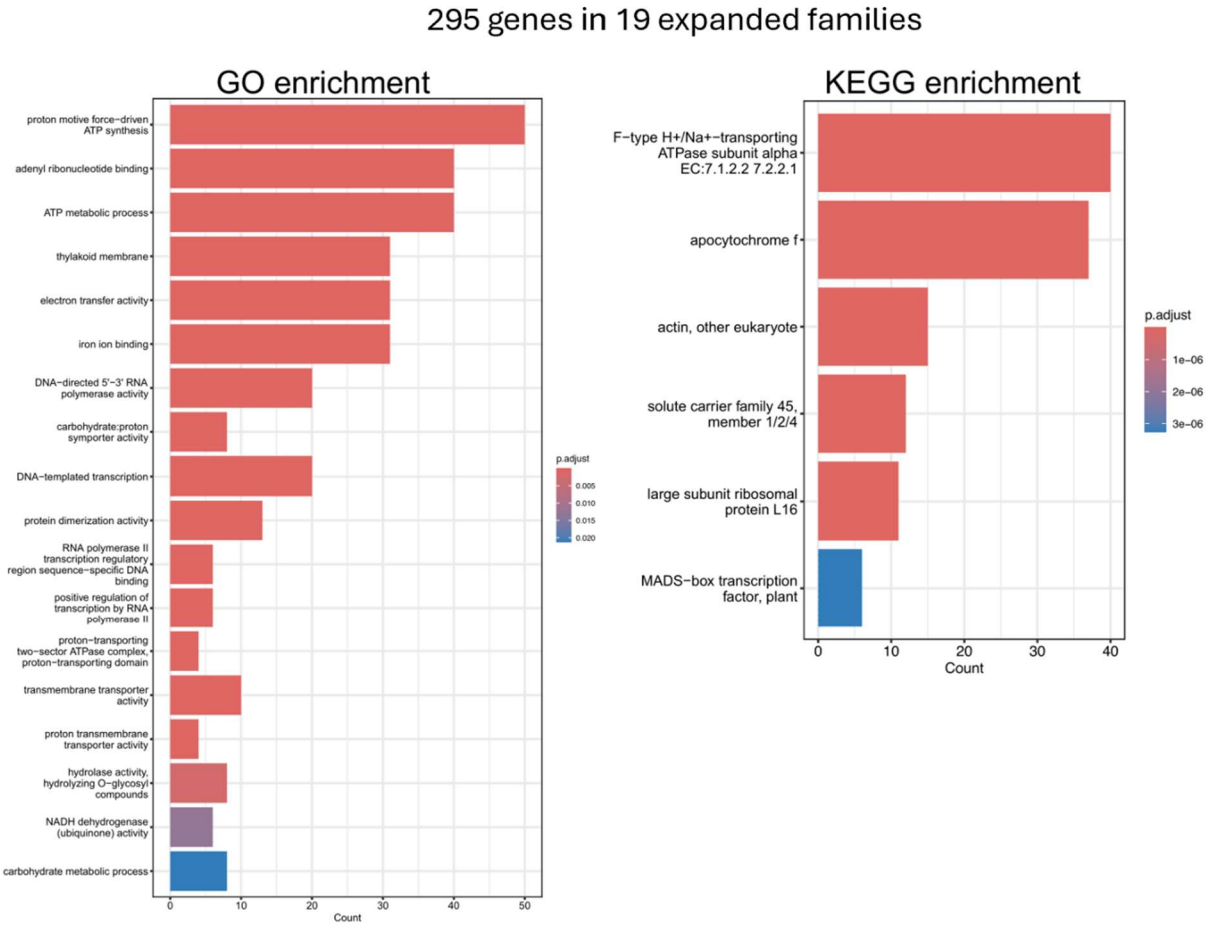

Figure 2.6. *GO and KEGG enrichment analysis of the 295 genes in the 19 expanded gene families.*

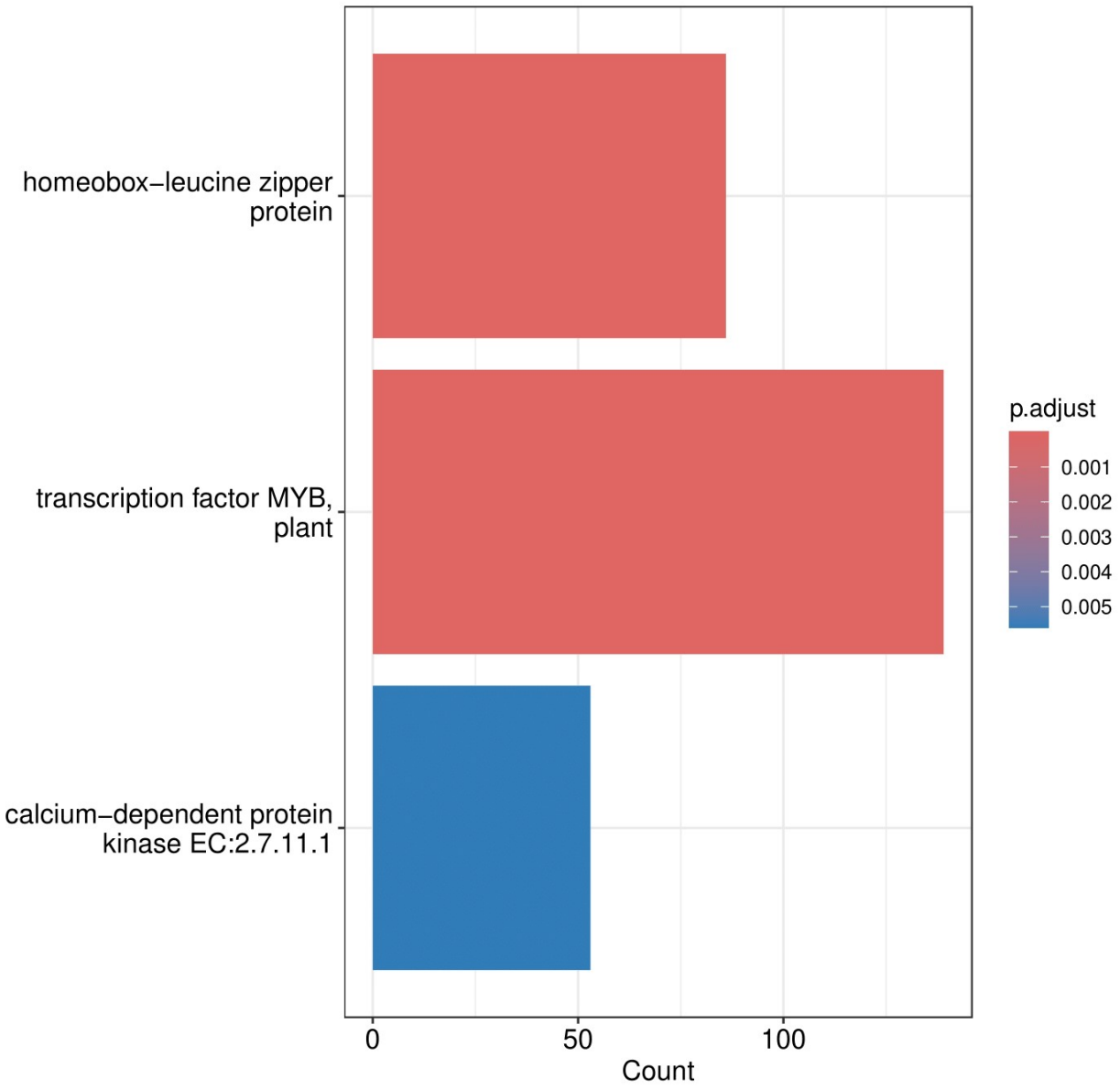

Figure 2.7. *KEGG enrichment analysis of WGD-derived gene pairs. WGD-derived genes identified through GenDup_finder pipeline*

Positive selection of duplicated genes

Analysis of the whole genome duplicated genes found a total of 123 Streptanthus genes to be under positive selection (Table S2.2). However, there was no significant KEGG or GO enrichment after false-discovery rate correction.

Discussion

The Streptanthoid Complex is a collection of plants found throughout the California Floristic Province. Individuals in this collection can be found in environments ranging from arid deserts all the way to snow packed mountain tops (Baldwin et al., 2012b). Despite living in drastically different environments, these flowering plants share a highly connected genetic background. Phylogenetically close relationships along with the expansive range and diverse morphology make this collection of plants an ideal study subject for exploring the rapid radiation and adaptation of flowering plants in the California Floristic Province. Active research is currently being conducted on this complex but is restricted to variation in traits due to a lack of genomic resources. This assembly offers a starting point in expanding the realm of genomic analyses related to these species. Future studies will now have another tool to help bridge the gap between the observed phenotypes and the underlying genetics.

The assembly presented here provides further evidence of a recent whole genome duplication shared throughout the Thelypodieae tribe which contains the Streptanthoid Complex. Whole genome duplications have previously been described in other members of the tribe, including *S. farnthworthianus* (Mandáková et al., 2017), *C. amplexicaulis* (Burrell et al., 2011), *P. antiscorbutica* and *Stanleya pinnata* (Kagale et al., 2014). Prior studies used transcriptome or molecular marker analysis. We build on these studies by providing a whole

genome assembly that can be used to study genome evolution after the whole genome duplication, and through our cross-species Ks analysis that indicates there was a single, shared WGD event among these species.

Gene duplication is a critical component of evolution, allowing rapid evolution of new functions (Ohno, 1970). Following a whole genome duplication event, there is typically a rapid extensive process of genome rearrangement and consolidation as diploidization occurs (Lynch and Conery, 2000; Qiao et al., 2019). During this time, genes and their functions can be lost, gained, and changed through processes including subfunctionalization and neofunctionalization (Almeida-Silva and Van de Peer, 2023; Lynch and Conery, 2000; Qiao et al., 2019). It has been reported that the At-β whole genome duplication event shared among the order Brassicales expanded their ability to produce glucosinolates (Barco and Clay, 2019). It is possible that the recent whole genome duplication described here played a key role in rapid radiation of the Streptanthoid Complex throughout California. *Streptanthus* species may have been able to rapidly adapt to diverse and harsh environments such as serpentine soils through the increased genetic arsenal following the whole genome duplication (Ohno, 1970; Qiao et al., 2019). For example, it is possible that the expansion of genes associated with ATP related processes and calcium signaling has developed a secondary regulatory network able to handle excess external Mg. This would have allowed *Streptanthus* to occupy previously unexploited habitats including heavy metal rich serpentine soils. Future research should be directed towards identifying genes that allowed *Streptanthus* to expand its geographic and climatic range. Identification of these genes could have major impacts on the agricultural field given the close genetic relationship between *Streptanthus* and it economically important sister clade *Brassica.*

Data Availability

Genome sequencing data and assembly data can be found on NCBI under Bioproject PRJNA283414.

Acknowledgments

I thank the Department of Energy Joint Genome Institute and collaborators for prepublication access to the *Caulanthus amplexicaulis* and *Stanleya pinnata* genome sequences. The work (proposal: 10.46936/10.25585/60000980) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. Also, the Division of Environmental Biology under the National Science Foundation for their funding under award 1831913. I would also like to thank Dylan Burge and Chris Grassa for their work on the initial short read assembly. Additionally, I would like to thank Qionghou Li for their work on the Ks and gene family expansion work.

Evolutionary dynamics of gene expression associated with germination and climate in California Jewelflowers

Abstract

This study investigates the evolutionary dynamics of gene expression related to seed germination and adaptation to climate in a clade of California Jewelflowers (Streptanthoid complex, Brassicaceae). We explore the genetic regulation of dormancy and germination. Through extensive germination assays and RNA sequencing of seeds from ten species across thirteen California populations, we constructed gene co-expression networks to assess their correlation with germination traits and climate adaptation. Our results demonstrate distinct gene expression patterns associated with adaptation to local climate conditions and provide evidence that positive selection has acted on gene expression modules associated with climate and germination. These findings revealed genetic networks used to optimize germination timing, offering insights into plant adaptation strategies in diverse climates. This study contributes to our understanding of plant evolutionary biology and the intricate mechanisms underlying seed germination and climate adaptation.

Introduction

The life cycle of flowering plants is composed of numerous developmental phases as they transition from seeds to mature flowering plants. While each developmental phase is important, some phases and their timing have a greater impact on the reproductive success than others. Flowering and germination are among the most critically important phases and the timing of their initiation has significant impacts on fecundity and survival of the plant (Roux et

al., 2006; Sajeev et al., 2024). Numerous studies have been conducted to explore how the timing of flowering affects reproductive success (Dieringer, 1991; Ollerton and Lack, 1998; Rodríguez-Pérez and Traveset, 2016; Schmitt, 1983). Timing of germination, on the other hand, has received less attention than its showier counterpart. Germination, the first stage in a plant's life cycle, has lasting effects on reproductive success, especially in annual plants which only have one growing season to complete their life cycles. For this reason, more work needs to be done to identify the complex network of genes which determine the timing of germination.

Germination is an irreversible transition from a seed to a seedling, which sets the stage for what the plant will experience both biotically and abiotically (Steinbrecher and Leubner-Metzger, 2017). If a seed germinates early in a growing season, it may be able to outcompete its neighbors along with gathering more resources leading to a larger size at flowering and higher fecundity (ten Brink et al., 2020). However, early germination in Mediterranean climates, may expose the fragile seedling to a cold snap in mountainous environments or drought in desert environments which will end the life of the seedling prematurely. Conversely, if a seed germinates later, it may be able to avoid an early season cold snap or arid conditions following a spurious rain shower (Mondoni et al., 2012; Venable and Lawlor, 1980). However, germinating later comes with the cost of a shorter growing period and increased competition from neighbors, leading to a smaller flowering plant with reduced fecundity. Given the costs and benefits of germination timing, determining the optimum time to germinate is key for reproductive success and plants have developed methods to estimate the optimum timing for germination.

The timing of germination is influenced by multiple cues including temperature, precipitation, and exposure to chemicals. These required cues and their magnitudes are not universal and vary both among species and between populations of the same species (Gremer et al., 2020c; Postma and Ågren, 2016). For example, some seeds will not germinate unless exposed to smoke which typically occurs during a forest fire (Flematti et al., 2004) while others require a minimum amount time exposed to winter cold (stratification) (Schütz and Rave, 1999). These cues may signal that favorable conditions for germination and establishment are approaching, such as after a fire or spring conditions. Further, requirements for these cues can prevent germination during unfavorable conditions. Cues related to climate are of particular interest due to their strong influence on germination timing. Overall it has been seen that species which experience greater variability in precipitation are likely to have a higher levels of dormancy while populations which experienced consistent patterns of precipitation had lower levels of dormancy (Freas and Kemp, 1983; Gremer et al., 2020b; Postma and Ågren, 2016; Torres-Martínez et al., 2017). Higher levels of dormancy was also observed in high elevation populations which experienced colder temperatures compared to their low elevation relatives (Gremer et al., 2020b). These adaptive changes to long term environmental conditions and cues may be reflected in gene expression profiles.

The gene expression profile of a seed provides a glimpse into the past climate conditions experienced by the mother plant and the generations proceeding it. We hypothesize that plants which have experienced high levels of climatic variability to have higher levels of dormancy which can observed through higher expression of genes promoting dormancy and inhibiting germination. Conversely, we hypothesize that plants that have experienced consistent climatic

conditions will have lower levels of dormancy which will be reflected in reduced expression of dormancy maintaining genes and increased expression of germination inducing genes. This interplay between dormancy and germination is controlled by a complex gene regulatory network that is being actively studied (Sajeev et al., 2024).

Abscisic acid (ABA) and gibberellins (GA) are two hormones that play key roles in regulating dormancy and germination and have been the topic of numerous studies (Bewley et al., 2013; Graeber et al., 2012; Shu et al., 2016). ABA inhibits germination by maintaining the seed in a dormant state while GA promotes germination by overcoming ABA induced dormancy and inducing cell growth related pathways. Numerous genes have been identified as playing a role in dormancy and ABA signaling including *CYP707A1, CYP707A2* (Okamoto et al., 2006)*, ABI3, ABI4, ABI5* (Giraudat et al., 1992; Koornneef et al., 1984; Skubacz et al., 2016; Wind et al., 2013)*, DELLAs* (Murase et al., 2008), and the *DOG* family (Bentsink et al., 2010, 2006). Of the multiple DOG genes reported, *DOG1* has been identified as having a critical role in dormancy and the ABA signaling pathway where the amount of DOG1 protein, which has been found to vary among populations of the same species, determines the level of dormancy in the seed. (Footitt et al., 2020). While these genes have been extensively studied, there is undoubtedly a suite of genes also influencing dormancy and germination yet to be discovered. By analyzing gene expression profiles, it is possible to discover additional genes playing pivotal roles in the germination process and how these genes how adapted to the cues of their home environments.

As plants adapt and change to the cues of their local environments, we expect the expression of genes related to germination to also adapt and change in order respond to these

cues. To explore this hypothesis, we analyzed the transcriptomic profiles of closely related wildflower species from the Streptanthoid complex (Brassicaceae) that have radiated throughout the California Floristic Province. These species have adapted to wide range of climates from desert environments with limited and variable rainfall to more typical Mediterranean climates that have wet, cool winter and warm, dry summers (Baldwin et al., 2012; Pearse et al., 2022, 2020; Worthy et al. in revision). Gene expression of seeds in both dry and imbibed conditions were measured, compared, and clustered to infer under the genetic pathways involved with germination and how they have evolved across the clade.

Methods

Overview

Seeds were collected from wild populations throughout California in 2018 and then grown out to produce seeds from plants grown in a common environment. Plants were grown in a screenhouse, where they experienced ambient seasonal temperatures at UC Davis 2019. Siliques from each individual maternal plant were collected and stored in envelopes at room temperature allowing for after-ripening. In 2021, the stored seeds were used in both germination and gene expression experiments to explore the germination compacities of the different populations. Gene expression profiles were then calculated to compare gene expression patterns with historical climate conditions.

Seed collection

Seeds from a collection of 10 *Streptanthus and Caulanthus* species across 13 populations throughout California were collected in 2018 (Figure 3.1 & Table 3.1). These seeds

were then bulked in a screenhouse at UC Davis in 2019. Siliques from each individual mother plant were collected, stored in brown envelopes, and stored inside at room temperature.

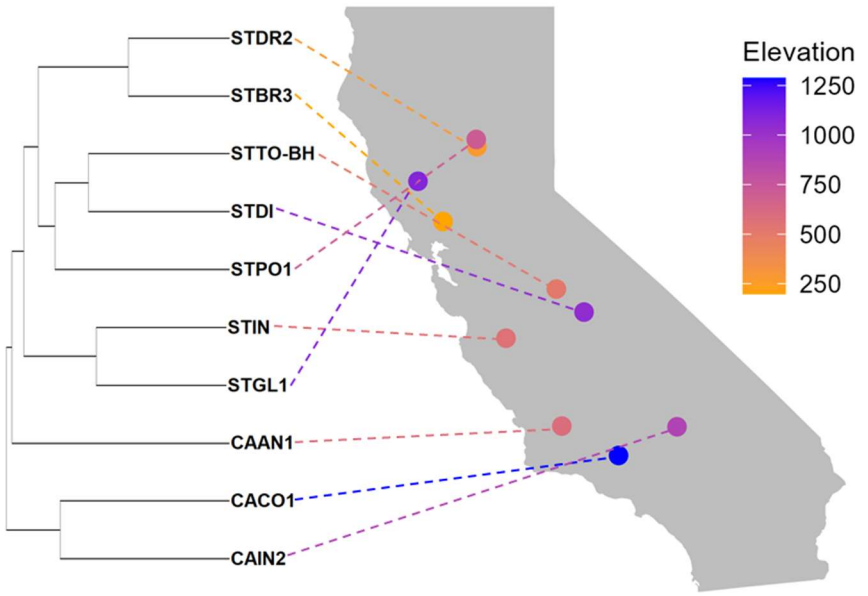

Figure 3.1. *Left: Phylogenetic tree of the 10 species in this study. Tree adapted from (Cacho et al. 2014). Right: Geographic distribution of populations across California. Color of point indicates elevation of the population.*

Table 3.1. *Locations of the 10 species in this study. Weather data extracted from the BCMv8 climate model*

Population	Latitude	Longitude	Elevation	Mean Annual PPT (mm)	Coefficient of Variation (PPT)	Mean Annual Minimum Temperature	Mean Annual Maximum Temperature
CAAN1	35.20	-119.85	588	20.90	1.56	8.06	23.96
CACO1	34.73	-118.71	1288	31.52	1.81	8.20	20.78
CAIN2	35.19	-117.53	881	10.48	1.86	9.85	25.30
STBR3	38.49	-122.24	198	64.46	1.46	8.51	22.95
STDI	37.04	-119.40	1038	63.15	1.36	8.11	21.62
STDR2	39.69	-121.56	278	92.76	1.29	9.98	23.06
STGL1	39.14	-122.75	1087	98.81	1.38	8.22	19.23
STIN	36.61	-120.97	561	26.31	1.47	7.81	23.18
STPO1	39.81	-121.57	712	125.62	1.29	9.04	21.36
STTO-BH	37.40	-119.96	511	48.80	1.38	8.71	23.53

Climate data

Climate data was collected from the Basin Climate Model version 8 (BCMv8) dataset (Flint et al., 2021). BCMv8 contains a down sampled time-series of gridded monthly climate measurements from weather stations throughout California. For each population, precipitation, climatic water deficit, minimum temperature, and maximum temperature were gathered. Since the study populations do not lie directly where the measurements were taken, for each population measurements were obtained from the nearest geographical measurement location in the BCMv8 dataset. To look at the historical effect of climate, the mean average values of these variables across the last 25 years, 1994-2018, was calculated and used in correlation analyses.

Germination assays and analysis

Germination experiments were conducted to quantify germination proportion (Figure 3.2), the number of seeds that germinated out of all possible seeds, in response to temperature for each species. Ten seeds for each population were placed on top of filter paper in a randomly chosen well of a 24 well-plate and were then imbibed with 3 mL of a 0.2% PPE water solution to prevent mold growth (Plant Preservation Mixture, Caisson Laboratories, UT, USA). Five replicate plates were placed at random locations in each of eight temperature-controlled chambers (E7/2 growth chambers, Conviron, Winnipeg, Manitoba, Canada). Each chamber was set at a constant temperature (5, 10, 15, 20, 25, 30, 35, 40 °C) with 12-h daylight cycles. Germination, determined by the emergence of the radicle, was surveyed daily for the first two weeks and every other day for the second two weeks for a total of 28 days. Additional water with plant preservation solution (1-2 mL) was added to the wells as needed to maintain moisture in the plates.

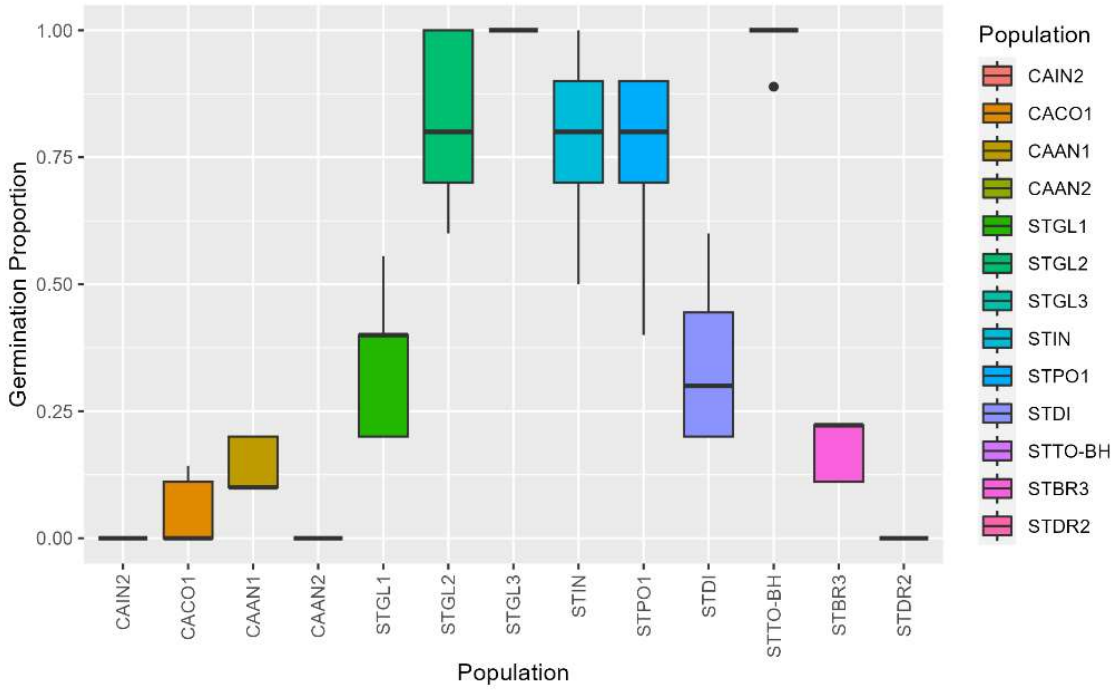

Figure 3.2. *Total germination proportion across the 13 populations at 20 °C, 5 replicates per a population. Populations along x-axis are ordered based on phylogenetic tree position in Figure 3.1.*

Measurements from the thermal germination experiments were used to calculate the optimum temperature, critical max/min temperature, thermal safety margin, and breadth for each germinating population. These values were calculated using the rTPC and nls.multistart packages following the pipeline outlined in (Padfield et al., 2021). Three distributions were tested for best fit of the data: gaussian, quadratic, and weibull. AICc was used to determine the best distribution and if distributions did not differ from each other, weighted model averaging was performed (Worthy, unpublished).

Experimental design for RNA sequencing

In order to compare gene expression between dry and imbibed seeds, we collected RNA from dry and imbibed seeds using seeds from the same seed pools. For each population, seeds were pooled with equal amounts of seeds from 8-10 mother plants, depending on seed

availability. 250mg of seeds from each pooled population were evenly divided among 10 filter paper lined petri dishes. Half the petri dishes from each population were then imbibed with 3 mL of a 0.2% PPE water solution. All dishes were then placed in a growth chamber set to 20 °C and continuous light. After 24 hours, all samples were removed from the growth chamber and flash frozen in liquid nitrogen.

RNA sequencing, mapping, and gene counts

The frozen seeds were sent to Amaryllis Nucleics for RNA extraction and library prep. Paired-end 150bp sequencing was performed at the UC Berkeley / QB3 Vincent J. Coates Genomics Sequencing Lab on Illumina's NovaSeq sequencing platform. Sequencing reads were quality checked using FastQC (Babraham Bioinformatics, n.d.) and were trimmed using Trimmomatic version 0.39 (Bolger et al., 2014b) in paired end mode with the parameters ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. The trimmed reads were mapped to a reference *Streptanthus diversifolius* genome using STAR (Dobin et al., 2013). Following mapping, gene counts were extracted from the alignment files using HTSeq-count (Anders et al., 2015). These gene counts were then used in downstream analysis using R.

Filtering of count data

To reduce noise in the analysis multiple filtering steps were performed. Using R (R Core Team, 2021), genes which did not have at least 10 read counts in at least 3 different samples were removed from the analysis. The gene counts were then normalized using edgeR (Robinson et al., 2010) using the trimmed mean of M-values method (Robinson and Oshlack, 2010). The

normalized counts were plotted, and outlier samples were manually removed from analysis. Following filtering, a total of 28,485 genes across 105 samples remained.

Differential gene expression

EdgeR was used to explore how the different samples clustered based on treatment and population. MDS plots were created using edgeR's plotMDS function. Afterwards a design matrix was created to test each population by treatment interaction in the dataset. The dispersion of the count data was then calculated using the estimateGLMdisp functions. The full interaction model was fit using the glmQLFit function and significant differentially expressed genes were determined using the glmQLFTest and topTags functions with an FDR cutoff of 0.05.

Generation of gene networks

The log base 2 counts per million for each gene were calculated using edgeR's cpm() function. These values were used to create separate gene networks for both the imbibed and not imbibed samples. Starting with the not imbibed, henceforth referred to as the "dry" samples, the top 40% of genes with the highest coefficient of variation were selected for network building. Gene network creation and module identification was performed using the R package WGCNA (Langfelder and Horvath, 2008) following the standard protocol substituting a signed hybrid instead of unsigned for network type. Modules with similar expression profiles were then merged using a cut height of 0.25, corresponding to a correlation of 0.75. This process was then repeated for the imbibed samples.

Correlation of gene modules to traits

The collection of gene modules was then correlated to traits including cumulative germination, optimum germination temperature, and elevation. The mean and standard deviation of annual measurements of maximum and minimum temperature along with the mean and coefficient of variation for annual precipitation across the last 25 years were also correlated to the gene modules. Gene modules were correlated to response variables using the R package Phylolm (Ho and Ané, 2014; Tung Ho and Ané, 2014) and a Brownian motion model of phylogenetic covariance among traits. This analysis identifies relationships where the association between traits is stronger than expected under a Brownian motion model of evolution. Relationships with a p-value less than 0.05 were considered significant. The most positively and negatively correlated gene module of each response variable was used for further analysis.

Shift in optimum

To look at possible selection on expression levels of these gene modules, reversible-jump Markov chain Monte Carlo was used to explore heterogeneity across the branches. Analysis was completed in R using the rjMCMC function from the Bayou (Uyeda and Harmon, 2014) package. Only modules which were found to be significantly correlated with one of our traits of interest were analyzed. On average most modules were found to have approximately two regime shifts across the phylogeny corresponding to three optima.

GO Term analysis

Each significantly correlated module was also analyzed to look for enrichment of GO terms. Genes in each module were compared to all the genes used in the WGCNA network

using the GOseq (Young et al., 2010) package in R to look for enriched GO terms in the biological processes ontology. Significantly enriched GO terms and their p-values were then input into Revigo (Supek et al., 2011b)to build TreeMaps of biological processes. The TreeMaps were then manually inspected to identify relevant GO terms.

Results

Analysis of differentially expressed genes

A total of 123 RNAseq libraries were sequenced across a collection of 10 species comprised of 13 distinct populations. These libraries were then mapped a to reference *S. diversifolius* genome containing 40,594 gene models, resulting in 34,190 gene models with mapped reads. Upon filtering for a minimum of 10 reads across at least 3 samples, the total number of genes being analyzed was reduced to 28,485. The gene counts were then normalized and visually inspected to remove outlier samples following normalization. A total of 18 samples, distributed among the different treatment and populations, were deemed to be outliers that could not be normalized and were removed from analysis resulting in 105 samples being used in downstream analysis.

Visualization of sample clustering in 2D space was performed using the plotMDS function of edgeR. The first two principal components of the data explained 19% of the variation with PC1 accounting for 11% and PC2 accounting for 8%. Treatment had the largest effect on PC1 with imbibed samples clustering on the lefthand side and dry samples clustering on the righthand side. Along PC2, Caulanthus populations clustered at the top of the plot while

Streptanthus populations were either at the same height or below the Caulanthus populations (Figure 3.3).

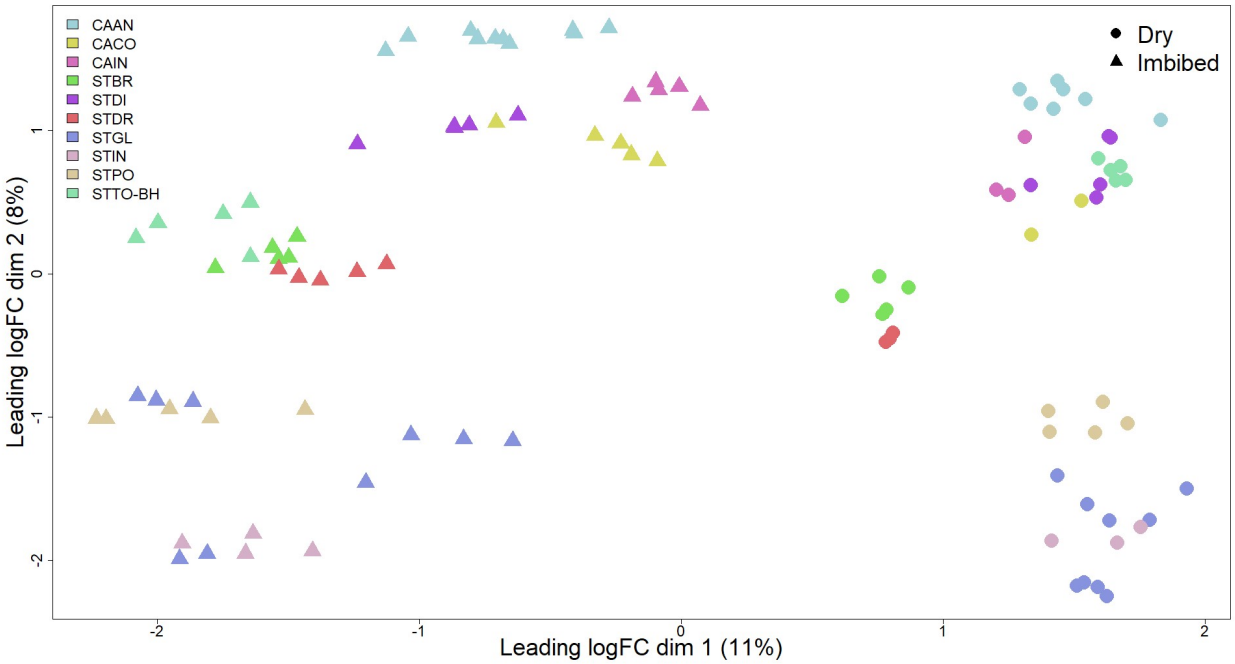

Figure 3.3. *Principal coordinate plot of MDS scaled values. Dry samples are circles while imbibed samples are triangles. Points are colored according to species. Clustering of species along PC2 appears to be mostly in agreement with the phylogenetic tree of Figure 3.1.*

Together these observations indicate that both species and treatment influence gene expression among these samples. To further test this hypothesis, a full interaction model between treatment and population was fit and tested using the glmQLFit and glmQLFTest functions of edgeR. Using an FDR cutoff of 0.05, treatment had a significant influence on 23,134 genes. Population was found to be significant for the expression level of 27,434 genes while the interaction between treatment and population significantly affected the expression of 17,480 (Table 3.2). This suggests that while treatment has a significant effect on gene expression, most of the differential gene expression is being driven by population. The effect of population can be driven by multiple factors including the historical climate of the region occupied by the

population and the unique genetics of the population. To further explore how the genes and their expression have evolved, genetic networks were built using WGCNA.

Table 3.2. *Number of significant and not significantly differentially expressed genes for each coefficient. Significance was determined using an FDR cutoff of 0.05*

Coefficient	Significant	Not Significant
Treatment+Interaction	23,134	5,351
Population+Interaction	27,434	1,051
Interaction	17,480	11,005

Construction of gene co-expression modules

For simplicity, gene networks for both dry and imbibed samples were evaluated independently using only one representative population per a species. Starting with the dry samples, the top 40% of genes with the highest coefficient of variation (11,394 genes) were selected for network building. WGCNA network construction organized these genes into 48 distinct modules with unique color names used to distinguish between the different modules. The grey module contained genes which could not be placed into one of the other 47 modules. The median number of genes per module was 127 genes. Comparison of expression profiles across the different modules and populations showed patterns shared among the populations and patterns unique to individual populations (Figure 3.4). For example, the branch of the module tree containing the tan and ivory modules were found to have higher expression among the Caulanthus populations compared to their sister Streptanthus populations. Populations including STPO1 and STGL1 were also found to have specific modules with greater expression compared to the other population as seen in the darkorange2 and violet modules, respectively.

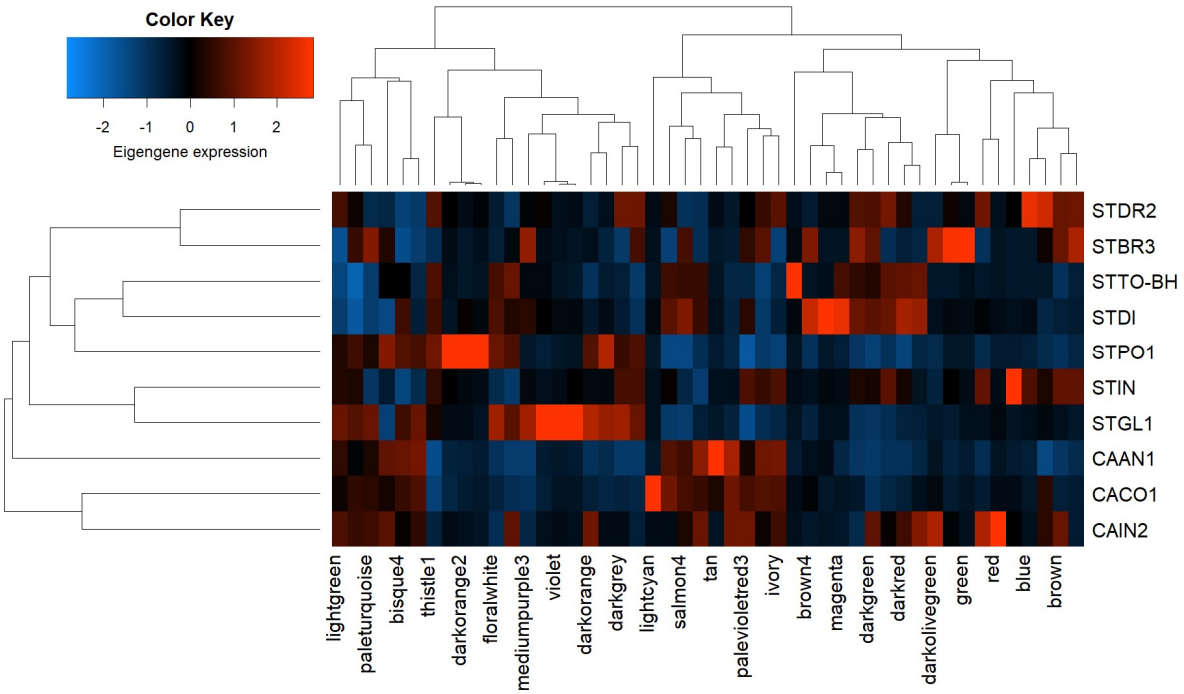

Figure 3.4. *Heatmap of eigengene expression of the 48 dry sample modules. Order of rows is based on previously described phylogeny and columns are organized by correlation of modules*

A different set of 11,394 top varying genes was used to create a WGCNA network using the imbibed samples. Following initial network building and merging of modules, 47 distinct modules were created with the grey module containing genes which could not be placed into one of the other 46 modules. The median number of genes per a module was 156 genes which is slightly higher than its dry counterpart. Similar to the dry network, some populations including STTO-BH and STDI shared similar expression patterns across most of the gene modules. Conversely, STBR3 for example, had an expression profile distinct from the other populations with expression of salmon4 and its closely related modules significantly higher than all other populations (Figure 3.5).

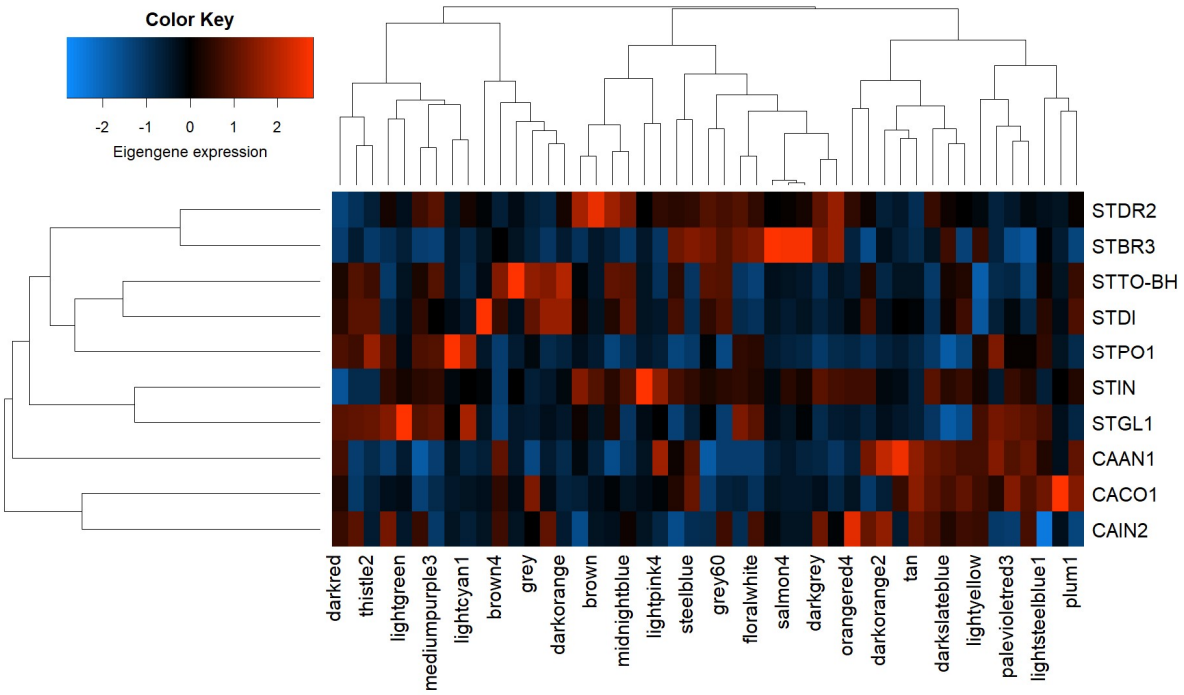

Figure 3.2. *Heatmap of eigengene expression of the 46 imbibed sample modules. Order of rows is based on previously described phylogeny and columns are organized by correlation of modules*

Gene modules correlated with germination and climate show evidence of positive selection

To ask if any of the coexpression modules might be related to climate adaptation or germination behavior, we performed a phylogenetic regression between module eigengenes and germination or climate variables. Within the dry samples, 28 unique gene modules were found to be significantly correlated with traits of interest by phylogenetic regression; the most significant positive and negative correlated modules are discussed here (See Table S3.1). The most significant module was the saddlebrown module which was positively correlated with variation in precipitation (Figure 3.6).

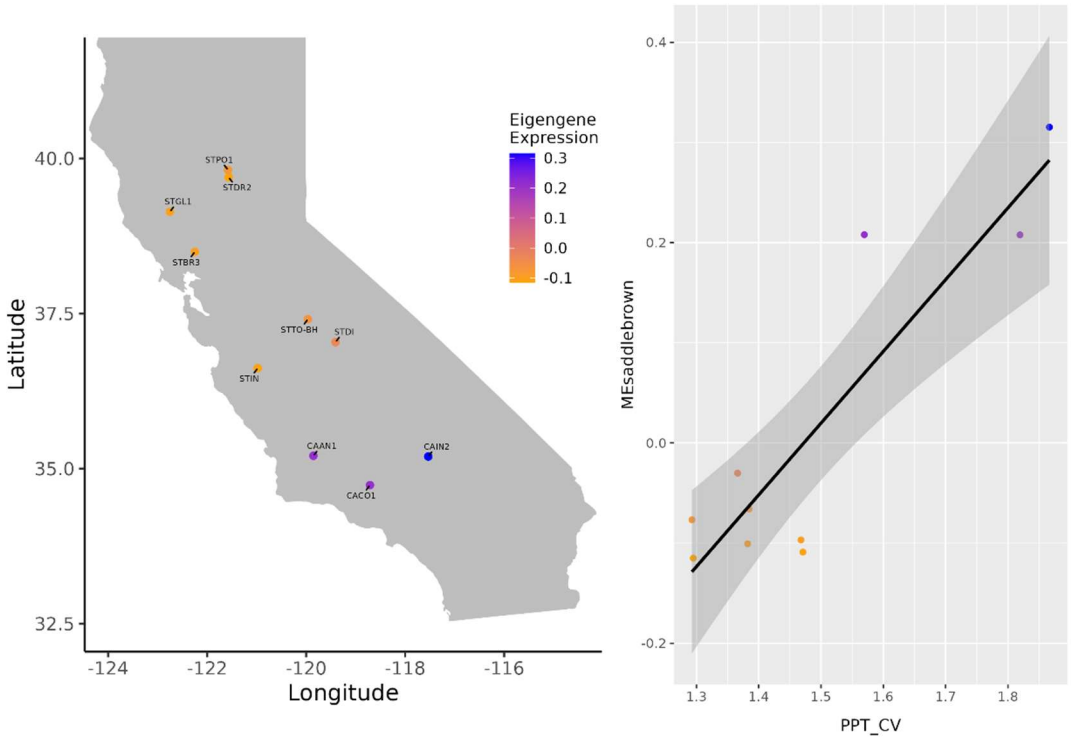

Figure 3.6. *Eigengene expression and population distributions for the saddlebrown module. Left: Geographic distribution of the 10 populations throughout California. Right: Scatterplot of coefficient of variation in precipitation on the x-axis and eigengene expression on the y-axis. Color of the points is based on the mean eigengene expression of the saddlebrown module for each population*

To further probe the changes in expression, we tested whether there was evidence for shifts in the optimal expression level of this module over evolutionary time. Three different expression regimes were predicted among the species being examined (Figure 3.7). The first regime belonged to the *Caulanthus coulteri* and *Caulanthus inflatus* species which are geographically the most southeast populations in the study. The next regime belonged to the *Caulanthus anceps* population which is geographically west of the two previous populations. The final regime was occupied by all the *Streptanthus* species which occupy habitats north of the 36 north parallel. Species in the southern half of California were found to have greater expression of genes in the saddlebrown module along with a greater amount of variation in

annual precipitation. The inverse was found for the northern species which had reduced gene expression and experienced less variation in annual precipitation.

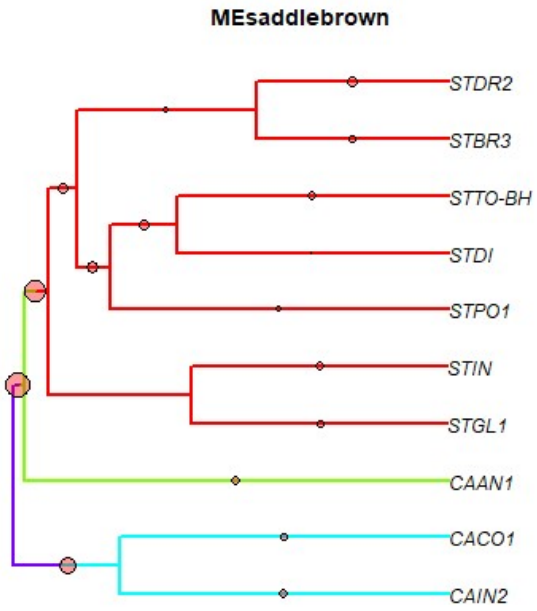

Figure 3.7. *Phylogeny showing the results of a bayou rjMCMC analysis of the multi-regime OU model on the saddlebrown module. Posterior probability of a regime shift is indicated by the size of the circle on the corresponding edge. Each color indicates a unique regime optimum.*

The saddlebrown module was found to be enriched for GO terms related to positive regulation of signaling and transmembrane transport of the ABA storage conjugate Abscisic acid glucosyl ester (ABA-GE).

Another module of interest is the yellowgreen module. It was found to be significantly negatively associated with cumulative germination proportion (Figure 3.8). Estimations of shifts in optimum expression predicted the same three expression regimes as the saddlebrown module (Figure 3.9). The first regime belonged once again to the *Caulanthus coulteri* and *Caulanthus inflatus* species which are geographically the most southeast populations in the

study. The second regime was comprised solely of Caulanthus anceps, and all the remaining *Streptanthus* species belong to the third regime. Genes in the yellowgreen module were also found to be significantly positively correlated with optimum germination temperature where a majority of those with higher gene expression were in desert regions. The yellowgreen module was found to be enriched for GO terms including cellular response to abscisic acid stimulus, seed maturation, and regulation of cellular response to heat.

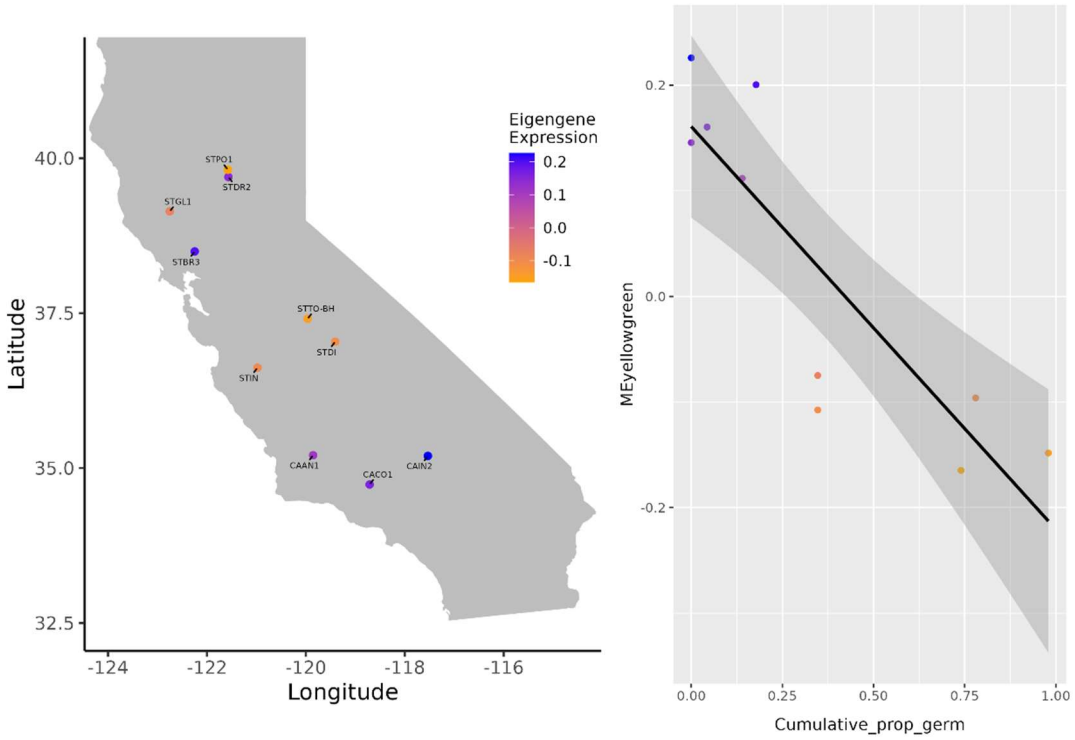

Figure 3.8. *Eigengene expression and population distributions for the yellowgreen module. Left: Geographic distribution of the 10 populations throughout California. Right: Scatterplot of coefficient of variation in precipitation on the x-axis and eigengene expression on the y-axis. Color of the points is based on the mean eigengene expression of the yellowgreen module for each population*

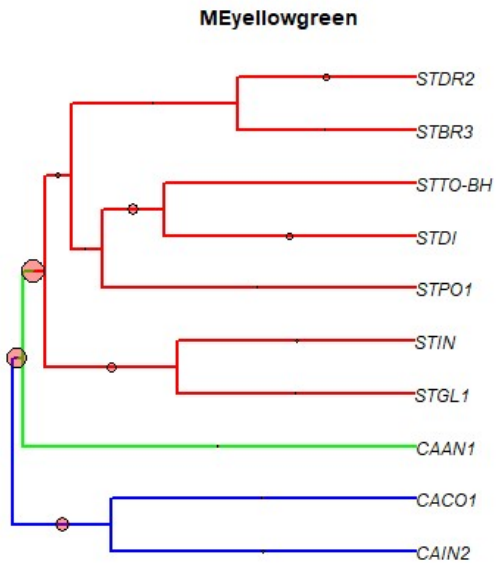

Figure 3.9. *Phylogeny showing the results of a bayou rjMCMC analysis of the multi-regime OU model on the yellowgreen module. Posterior probability of a regime shift is indicated by the size of the circle on the corresponding edge. Each color indicates a unique regime optimum.*

The imbibed samples had a total of 21 unique gene modules significantly correlated with our traits of interest by phylogenetic regression (See Table S3.2). The most significant module was the tan module which was positively correlated with variation in annual precipitation and negatively correlated with latitude (Figure 3.10).

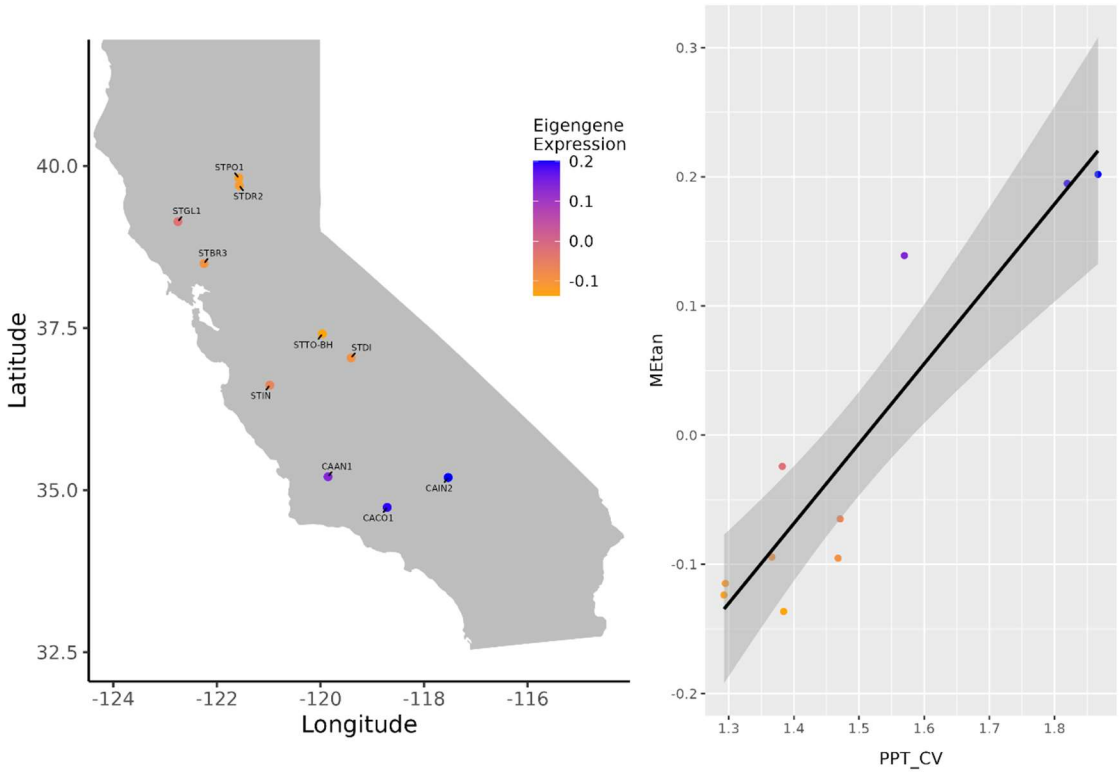

Figure 3.10. *Eigengene expression and population distributions for the tan module. Left: Geographic distribution of the 10 populations throughout California. Right: Scatterplot of coefficient of variation in precipitation on the x-axis and eigengene expression on the y-axis. Color of the points is based on the mean eigengene expression of the tan module for each population*

Estimation of shifts in optimum expression of this module predicted the same three regimes as the saddlebrown module in the dry samples. Similarly, to the saddlebrown module, species in southern California had greater gene expression and increased variation in annual precipitation and the opposite pattern for the northern species. The tan module was found to be enriched for multiple GO terms related to biotic and abiotic stimuli and calcium ion transport.

The blue module was found to be positively associated with cumulative germination proportion, negatively associated with variation in annual precipitation, and positively associated with breadth of the germination temperature window (Figure 3.11). Estimations of shifts in optimum expression predicted three different expression regimes among the populations which were the same as the tan module. All *Streptanthus* populations in the

northern half of California had greater expression than the three *Caulanthus* populations located in southern California. The blue module was found to be enriched for GO terms including developmental growth and positive regulation of gene expression.

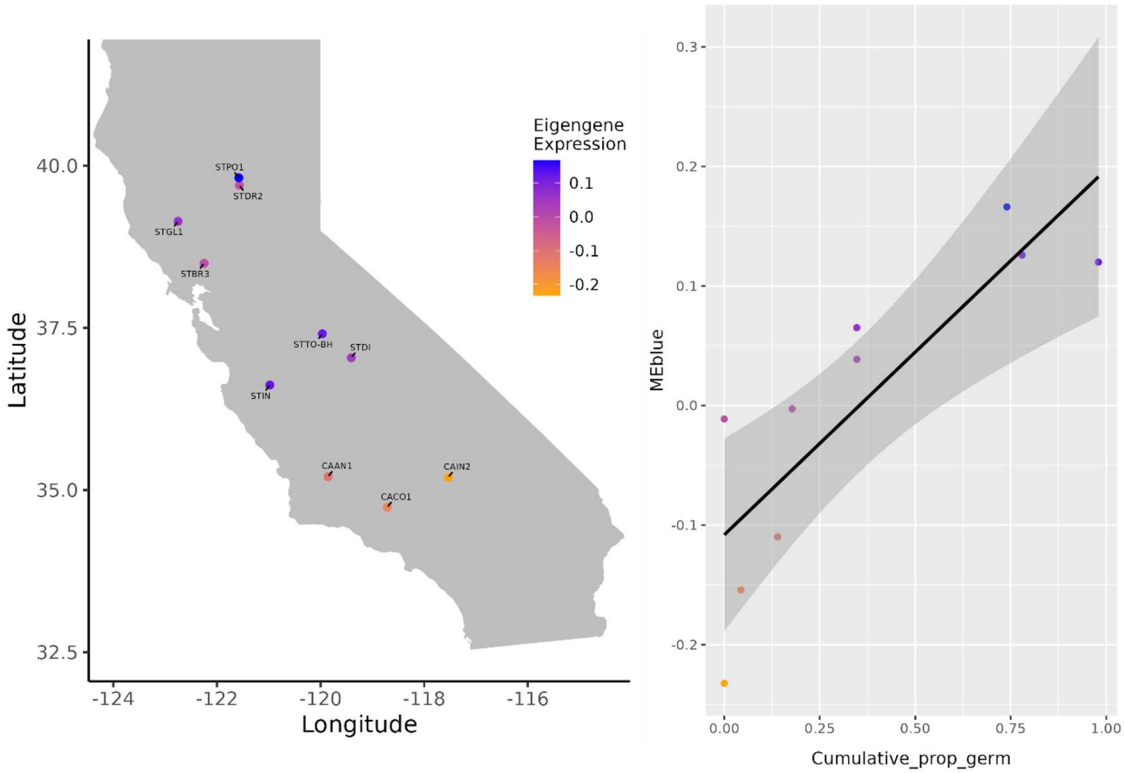

Figure 3.11. *Eigengene expression and population distributions for the blue module. Left: Geographic distribution of the 10 populations throughout California. Right: Scatterplot of coefficient of variation in precipitation on the x-axis and eigengene expression on the y-axis. Color of the points is based on the mean eigengene expression of the blue module for each population*

Discussion

Timing of germination plays a critical role in the life cycle of plants. Sensing the environment and timing germination with optimal conditions is an important strategy for optimizing fitness. Sensing of these environmental cues is ultimately controlled by the underlying genetic networks of the plants. Multiple genes including *CYP707A1, CYP707A2* (Okamoto et al., 2006), *ABI3, ABI4, ABI5* (Giraudat et al., 1992; Koornneef et al., 1984; Skubacz et al., 2016; Wind et al., 2013), *DELLAs* (Murase et al., 2008), and the *DOG* (Bentsink et al.,

2010, 2006) family have been identified as influencing dormancy and germination. While these genes have been well documented, other influential genes and their interactions remain undiscovered. Through the use of genetic networks, we may be able to elucidate these unknown genes and the roles they play in dormancy and germination.

In our experiment we asked whether the expression of genes and their associated gene networks could be related to historical climate and the possible effects that such expression changes could have on germination among the different species and populations. Overall, we found that *Caulanthus* populations germinated at a lower rate than their *Streptanthus* relatives (Figure 3.2). Consistent with this, *Caulanthus* populations historically experience a larger magnitude of variation in precipitation than *Streptanthus*. Variability in precipitation can have dramatic effects on germination and the survival of a seedling (Postma and Ågren, 2016; Tielbörger et al., 2012; Torres-Martínez et al., 2017). In the desert regions where our *Caulanthus* populations are located and where variability in precipitation is greater, timing of germination is key. Indeed, studies have shown that dormancy is common in annual plant species that experience high variability in precipitation (Cuello et al., 2019; Tielbörger et al., 2012; Venable, 2007). These cues can also prevent them from germinating at the wrong time such as after a brief summer shower which may cause the seedling to experience extended periods of drought reducing survivorship and fecundity. To account for this variable climate, the populations may employ genetic networks that increase dormancy and alter the response to typical germination cues in order to prevent germination at the wrong time.

Possible modules in the dry seed network that may be related to this are the saddlebrown module whose gene expression is positively correlated with variation in

precipitation (Figure 3.6). This module is significantly enriched for GO terms including positive regulation of signaling and Abscisic acid glucosyl ester (ABA-GE) transmembrane transport. Deconjugation of ABA-GE by the endoplasmic reticulum and vacuolar β-glucosidases allows the rapid formation of free ABA in response to abiotic stress conditions such as dehydration and salt stress (Han et al., 2020). Higher expression of genes in this module in the variable climates of the *Caulanthus* populations may explain their lower germination proportions and how these populations have adapted to a more variable climate through increased levels of dormancy. This result aligns with other demographic studies which have observed increased dormancy in population which experience higher levels of climatic variability (Freas and Kemp, 1983; Gremer et al., 2020b; Torres-Martínez et al., 2017). In a similar vein, the yellowgreen module's gene expression is negatively correlated with germination proportion and gene expression is higher in the *Caulanthus* populations (Figure 3.8). Genes in this module are significantly enriched for GO terms including cellular response to abscisic acid stimulus, seed maturation, and regulation of cellular response to heat. Expression of these genes may explain how *Caulanthus* populations have adapted to their variable desert climates through methods of increased dormancy.

Examining the imbibed seed network can provide information on differences in how plants respond to rainfall cues. We found that the tan module's gene expression is positively correlated with variation in precipitation (Figure 3.10). This module is enriched for GO terms related to biotic and abiotic stimuli and calcium ion transport. Harsh environments, such as those with high climatic variable, have been found to affect the germination niche by creating narrower germination breadths (Fernández-Pascual et al., 2017), which may limit germination

to more favorable conditions. Higher expression of these genes may improve the individual's ability to detect a narrower range of optimum germination conditions while navigating a more variable climate. Expression of this module is increased in the variable climates of the *Caulanthus* populations and may prevent mistimed germination events. While the first three modules have been associated with increased ABA signaling and reduced germination, the blue module's gene expression is positively correlated with germination proportion. The blue module is enriched for GO terms including developmental growth and positive regulation of gene expression. Expression of genes in the blue module is greater in the *Streptanthus* species compared to the *Caulanthus* species (Figure 3.11) and aligns with their differences in germination proportions (Figure 3.2). The blue module also has a significant negative correlation to annual variability in precipitation. Higher expression of these genes may trigger rapid germination when the seed is imbibed. This would align with past studies that have found germination to be increased in populations which experience more predictable precipitation (Freas and Kemp, 1983; Gremer et al., 2020b; Torres-Martínez et al., 2017). The difference in expression of this module maybe due historical climate as the more variable desert climate of the *Caulanthus* populations has led to decreased expression of this module to prevent undesirable germination. Oppositely, the less variable climates of the *Streptanthus* populations may have promoted expression of genes in this module allowing germination across a larger window of climate conditions.

Conclusion

Timing of germination is critical for seedling establishment and successful reproduction. Dormancy and germination are primarily controlled by the balance between the two hormones

abscisic acid and gibberellins with the former maintaining dormancy and the later suppressing dormancy and promoting germination. The balance between these two hormones is influenced by many factors including the environmental conditions of the mother plant, after-ripening, stratification, and the seed's local environment. In this study we have identified gene networks in both dry and imbibed seeds that are significantly associated with the historical climate conditions of jewelflowers spread through the California Floristic Province. While most studies look at germination and dormancy across multiple populations of the same species, we analyzed a collection of closely related species occupying significantly different environments. Additionally, we focused on overall expression patterns and the networks of interacting genes rather than solely on previously reported focal genes. Future analysis of these gene networks may reveal additional genes controlling germination and may help to further expand our understanding of this complex process and how it has and continues to different to changing climate.

Data Availability

RNA sequencing data can be found on NCBI under Bioproject PRJNA755996.

Acknowledgments

I thank the Division of Environmental Biology under the National Science Foundation for their funding under award 1831913. I would also like to thank Megan Bontrager, Arquel Miller, and all the dedicated undergraduate researchers in the Gremer, Schmitt, and Strauss labs for their help in collecting the RNA samples used in this project. Lastly, I would like to thank Erick Quintana Vazquez for his exploratory work on module discovery.

Appendix

Transcriptomes of Six Streptanthoid Complex Species

Purpose

To increase the number of genomic resources available for species of Streptanthoid Complex, we created transcriptomes for six jewelflower species. Focal species from six branches were selected to cover a majority of the phylogenetic tree. The goal is for these transcriptome sequences to aide future studies which look to explore the evolution and adaptation of the Streptanthoid Complex as it radiated throughout the California Floristic Province.

Availability

Data is freely available on Dryad (DOI: 10.5061/dryad.t1g1jwt99)

Methods

Seed Source

The six species chosen for transcriptome sequencing were *Caulanthus anceps, Caulanthus amplexicaulis, Caulanthus inflatus, Streptanthus breweri, Streptanthus glandulosus, and Streptanthus tortuosus.* Seeds for these species were produced in a screenhouse at UC Davis in 2019 using field collected seeds from 2018 and stored in brown envelopes at room temperature. In 2021, these seeds were used to extract RNA under 10 different tissue and treatment combinations (Table A.1).

A.3.Tissue and treatment combinations

Combination	Tissue	Treatment	Timepoints
1	Seed	Dry 20 °C	1
2	Seed	Chilled, Imbibed, 4 °C	1
3	Leaf	Normal, 20 °C	4
4	Leaf	Cold, 4 °C	4
5	Leaf	Hot, 40 °C	4
6	Leaf	Dark, 20 °C	4
7	Leaf	Drought, 20 °C	4
8	Root	Normal, 20 °C	1
9	Flower	Normal, 20 °C	1
10	Silique	Normal, 20 °C	1

Tissue Collection

For combination one, dry seeds were removed from their envelopes and immediately frozen in liquid nitrogen. For combination two, dry seeds were removed from their envelopes, placed on top of germination paper in two-inch petri dishes, and imbibed with 3 mL of water. The petri dishes were then placed in a 4 °C chamber with constant light for six hours. After six hours the seeds were dried to remove excess water and frozen in liquid nitrogen.

Combinations three through ten were collected from young plants. Randomized cones containing a mixture of 50% Ron's Mix and 50% sand were saturated with nutrient water. A small divot was then made in each cone where 3-4 seeds were placed before being covered with a small amount of the soil mixture. The cones were then placed in a rack and covered with plastic wrap to prevent the top layer of the soil from drying out. The covered rack was then placed in a growth chamber set to 20 °C with a light/dark cycle of 12/12. Two weeks after being placed in the growth chamber, the plastic wrap was removed, and the recently germinated seedlings were exposed to the air. The seedlings remained in the growth chamber with nutrient water being provided every other day. Once the seedlings on average had attained approximately 8-10 true leaves, the seedlings were moved to their treatment conditions and/or

had their tissue collected. Tissue across multiple replicates of the same tissue/treatment combination and different timepoints were pooled and frozen.

Combination 3 was collected 0, 6, 12, and 18 hours after lights on. Combination 4 was collected 3, 6, 12, and 24 hours after being moved to a 4 °C with a light/dark cycle of 12/12. Combination 5 was collected .25, .5, 1, and 3 hours after being moved to a 40 °C with a light/dark cycle of 12/12. Combination 6 was subjected to a full day without light and the following day tissue was collected in the dark 0, 6, 12, and 18 hours after usual lights on. Combination 7 was subjected to five days without watering and on the sixth day tissue was collected in the 0, 6, 12, and 18 hours after lights on. Plants for combination 8 were removed from their pots, had excess soil washed from their roots, and cleaned root tissue collected. Combinations 9 and 10 had their flowers and siliques collected once an adequate amount had formed.

RNA Extraction and sequencing

Total RNA from 50 mg of each sample was extracted using New England Biolab's Monarch Total RNA Miniprep Kit. Total RNA was then quantified, and quality checked on the Qubit 3. For each species total RNA from each tissue/treatment was equally pooled and 300 ng of total RNA was used with the SMRTbell Express Template Prep Kit 2.0 to create a total of 6 SMRTbell libraries. The 6 SMRTbell libraries were then sent to the University of California, Davis Genome Center for sequencing on the Sequel II. Following sequencing, the unaligned BAM files were processed using PacBio's IsoSeq pipeline ("PacificBiosciences/pbbioconda) to create high quality full length non-concatemer reads.

Transcripts, annotation, and orthogroups

Transcript identification for each species was completed using the isON transcript analysis pipeline (Petri and Sahlin, 2023; Sahlin and Medvedev, 2020). The pipeline works on PacBio sequence data in two steps, first by clustering reads based on sequence similarity and then by generating isoforms out of clustered long reads. Following isoform identification, the transcripts were then aligned to the UniProt database (The UniProt Consortium, 2023) using blastx (Sayers et al., 2022). Annotations were added to the transcripts based on their best match. To aide in future investigations, orthogroups were generated using orthofinder (Emms and Kelly, 2019).

Results

A total of 475,604 isoform sequences were generated across the six species with an average of 67,943 isoform sequences per species. These isoform sequences represent a total of 131,618 gene clusters with an average of 21,936 gene clusters per species. Although the number of isoforms varied between the different species, the total number of gene clusters among the six species were approximately 22,000. BUSCO analysis found each transcriptome to have a complete BUSCO percentage of 90% or better (Table A.2). A total of 57,093 unique orthogroups were created using Orthofinder with default settings. Of the 57,093 orthogroups, 20,654 orthogroups contained at least one isoform from each of the six species. Lastly, 631 orthogroups were created that have a single-copy isoform from each species (Table A.3).

A.2 Number of genes, isoforms, and BUSCO summary statistics for each transcriptome. Analysis completed using BUSCO version 5.5.0 in transcriptome mode and the embryophyta odb10 dataset

Species	Genes	Isoforms	Complete BUSCOs	Single Copy	Duplicated	Fragmented	Missing	Total
C. amplexicaulis	22,629	84,786	91.9%	15.9%	76.0%	1.9%	6.2%	1614
C. anceps	21,822	90,802	91.8%	16.7%	75.1%	1.7%	6.5%	1614
C. inflatus	21,459	79,422	94.3%	22.4%	71.9%	1.2%	4.5%	1614
S. breweri	22,475	65,787	94.7%	25.9%	68.8%	1.4%	3.9%	1614
S. glandulosus	21,612	79,150	93.4%	19.2%	74.2%	1.7%	4.9%	1614
S. tortuosus	21,621	75,657	94.3%	21.6%	72.7%	1.7%	4.0%	1614

A.3 Overall statistics from Orthofinder

Number of isoforms	475,614
Number of isoforms in orthogroups	450,720
Number of orthogroups	57,093
Mean orthogroup size	7.9
Median orthogroup size	6
Number of orthogroups with all species present	20,654
Number of single-copy orthogroups	631

Literature Cited

Almeida-Silva, F., Van de Peer, Y., 2023. Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms. Mol. Biol. Evol. 40, msad141. https://doi.org/10.1093/molbev/msad141

Al-Shehbaz, 2010. Brassicaceae in Flora of North America @ efloras.org [WWW Document]. URL http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=10120 (accessed 2.8.24).

AltschuP, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., n.d. Basic Local Alignment Search Tool 8.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. https://doi.org/10.1093/bioinformatics/btu638

Arima Genomics, n.d. Arima-HiC Mapping Pipeline.

Babraham Bioinformatics, n.d. FastQC.

Baldwin, B.G., Goldman, D., Keil, D.J., Patterson, R., Rosatti, T.J., Wilken, D. (Eds.), 2012a. The Jepson Manual: Vascular Plants of California, Thoroughly Revised and Expanded, 2nd ed.

Baldwin, B.G., Goldman, D.H., Vorobik, L.A., 2012b. The Jepson manual: vascular plants of California, 2nd ed.. ed. University of California Press, Berkeley, Calif.

Barco, B., Clay, N.K., 2019. Evolution of Glucosinolate Diversity via Whole-Genome Duplications, Gene Rearrangements, and Substrate Promiscuity. Annu. Rev. Plant Biol. 70, 585–604. https://doi.org/10.1146/annurev-arplant-050718-100152

Bayer, P.E., Hurgobin, B., Golicz, A.A., Chan, C.-K.K., Yuan, Y., Lee, H., Renton, M., Meng, J., Li, R., Long, Y., Zou, J., Bancroft, I., Chalhoub, B., King, G.J., Batley, J., Edwards, D., 2017. Assembly and comparison of two closely related Brassica napus genomes. Plant Biotechnol. J. 15, 1602–1610. https://doi.org/10.1111/pbi.12742

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., Delourme, R., Deniot, G., Denoeud, F., Duffé, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., Vekemans, X., D'Hont, A., Rousseau-Gueutin, M., Barbe, V., Cruaud, C., Wincker, P., Aury, J.-M., 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat. Plants 4, 879–887. https://doi.org/10.1038/s41477-018-0289-4

Bentsink, L., Hanson, J., Hanhart, C.J., Blankestijn-de Vries, H., Coltrane, C., Keizer, P., El-Lithy, M., Alonso-Blanco, C., de Andrés, M.T., Reymond, M., van Eeuwijk, F., Smeekens, S., Koornneef, M., 2010. Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. Proc. Natl. Acad. Sci. 107, 4264–4269. https://doi.org/10.1073/pnas.1000410107

Bentsink, L., Jowett, J., Hanhart, C.J., Koornneef, M., 2006. Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. Proc. Natl. Acad. Sci. 103, 17042–17047. https://doi.org/10.1073/pnas.0607877103

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., Huala, E., 2015a. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. genesis 53, 474–485. https://doi.org/10.1002/dvg.22877

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., Huala, E., 2015b. The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. Genes. N. Y. N 2000 53, 474–485. https://doi.org/10.1002/dvg.22877

Bertioli, D.J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., Leal-Bertioli, S.C.M., Ren, L., Farmer, A.D., Pandey, M.K., Samoluk, S.S., Abernathy, B., Agarwal, G., Ballén-Taborda, C., Cameron, C., Campbell, J., Chavarro, C., Chitikineni, A., Chu, Y., Dash, S., El Baidouri, M., Guo, B., Huang, W., Kim, K.D., Korani, W., Lanciano, S., Lui, C.G., Mirouze, M., Moretzsohn, M.C., Pham, M., Shin, J.H., Shirasawa, K., Sinharoy, S., Sreedasyam, A., Weeks, N.T., Zhang, X., Zheng, Z., Sun, Z., Froenicke, L., Aiden, E.L., Michelmore, R., Varshney, R.K., Holbrook, C.C., Cannon, E.K.S., Scheffler, B.E., Grimwood, J., Ozias-Akins, P., Cannon, S.B., Jackson, S.A., Schmutz, J., 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat. Genet. 51, 877–884. https://doi.org/10.1038/s41588-019-0405-z

Bewley, J.D., Bradford, K.J., Hilhorst, H.W.M., Nonogaki, H., 2013. Seeds: Physiology of Development, Germination and Dormancy, 3rd Edition. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-4693-4

Boideau, F., Richard, G., Coriton, O., Huteau, V., Belser, C., Deniot, G., Eber, F., Falentin, C., Ferreira de Carvalho, J., Gilet, M., Lodé-Taburel, M., Maillet, L., Morice, J., Trotoux, G., Aury, J.-M., Chèvre, A.-M., Rousseau-Gueutin, M., 2022. Epigenomic and structural events preclude recombination in Brassica napus. New Phytol. 234, 545–559. https://doi.org/10.1111/nph.18004

Bolger, A.M., Lohse, M., Usadel, B., 2014a. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolger, A.M., Lohse, M., Usadel, B., 2014b. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brady, K.U., Kruckeberg, A.R., Bradshaw Jr., H.D., 2005. Evolutionary Ecology of Plant Adaptation to Serpentine Soils. Annu. Rev. Ecol. Evol. Syst. 36, 243–266. https://doi.org/10.1146/annurev.ecolsys.35.021103.105730

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. https://doi.org/10.1038/nbt.3519

Burrell, A.M., 2010. Molecular and Genetic Analysis of Adaptive Evolution in the Rare Serpentine Endemic, Caulanthus amplexicaulis var. barbarae(J. Howell). Texas A&M University.

Burrell, A.M., Pepper, A.E., 2006. Primers for 10 polymorphic microsatellites from Caulanthus amplexicaulis var. barbarae, and cross-amplification in other species within the Streptanthoid Complex (Brassicaceae). Mol. Ecol. Notes 6, 770–772. https://doi.org/10.1111/j.1471-8286.2006.01337.x

Burrell, A.M., Taylor, K.G., Williams, R.J., Cantrell, R.T., Menz, M.A., Pepper, A.E., 2011. A comparative genomic map for Caulanthus amplexicaulis and related species (Brassicaceae). Mol. Ecol. 20, 784–798. https://doi.org/10.1111/j.1365-294X.2010.04981.x

Cacho, N.I., Kliebenstein, D.J., Strauss, S.Y., 2015. Macroevolutionary patterns of glucosinolate defense and tests of defense-escalation and resource availability hypotheses. New Phytol. 208, 915–927. https://doi.org/10.1111/nph.13561

Cacho, N.I., McIntyre, P.J., Kliebenstein, D.J., Strauss, S.Y., 2021. Genome size evolution is associated with climate seasonality and glucosinolates, but not life history, soil nutrients or range size, across a clade of mustards. Ann. Bot. 127, 887–902. https://doi.org/10.1093/aob/mcab028

Cacho, N.I., Strauss, S.Y., 2014. Occupation of bare habitats, an evolutionary precursor to soil specialization in plants. Proc. Natl. Acad. Sci. 111, 15132–15137. https://doi.org/10.1073/pnas.1409242111

Campbell, M.S., Holt, C., Moore, B., Yandell, M., 2014a. Genome Annotation and Curation Using MAKER and MAKER-P: Genome Annotation and Curation Using MAKER and MAKER-P, in: Bateman, A., Pearson, W.R., Stein, L.D., Stormo, G.D., Yates, J.R. (Eds.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA, p. 4.11.1-4.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Campbell, M.S., Holt, C., Moore, B., Yandell, M., 2014b. Genome Annotation and Curation Using MAKER and MAKER-P. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al 48, 4.11.1-4.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.-H., Childs, K.L., Sun, Y., Jiang, N., Yandell, M., 2014c. MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. Plant Physiol. 164, 513–524. https://doi.org/10.1104/pp.113.230144

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.-H., Childs, K.L., Sun, Y., Jiang, N., Yandell, M., 2014d. MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations1[W][OPEN]. Plant Physiol. 164, 513–524. https://doi.org/10.1104/pp.113.230144

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., Yandell, M., 2008a. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196. https://doi.org/10.1101/gr.6743907

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., Yandell, M., 2008b. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196. https://doi.org/10.1101/gr.6743907

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238. https://doi.org/10.1186/1471-2105-13-238

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Correa, M., Da Silva, C., Just, J., Falentin, C., Koh, C.S., Le Clainche, I., Bernard, M., Bento, P.,

Noel, B., Labadie, K., Alberti, A., Charles, M., Arnaud, D., Guo, H., Daviaud, C., Alamery, S., Jabbari, K., Zhao, M., Edger, P.P., Chelaifa, H., Tack, D., Lassalle, G., Mestiri, I., Schnel, N., Le Paslier, M.-C., Fan, G., Renault, V., Bayer, P.E., Golicz, A.A., Manoli, S., Lee, T.-H., Thi, V.H.D., Chalabi, S., Hu, Q., Fan, C., Tollenaere, R., Lu, Y., Battail, C., Shen, J., Sidebottom, C.H.D., Wang, X., Canaguier, A., Chauveau, A., Berard, A., Deniot, G., Guan, M., Liu, Z., Sun, F., Lim, Y.P., Lyons, E., Town, C.D., Bancroft, I., Wang, X., Meng, J., Ma, J., Pires, J.C., King, G.J., Brunel, D., Delourme, R., Renard, M., Aury, J.-M., Adams, K.L., Batley, J., Snowdon, R.J., Tost, J., Edwards, D., Zhou, Y., Hua, W., Sharpe, A.G., Paterson, A.H., Guan, C., Wincker, P., 2014. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 345, 950–953. https://doi.org/10.1126/science.1253435

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., Town, C.D., 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 89, 789–804. https://doi.org/10.1111/tpj.13415

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., Wang, X., 2011. BRAD, the genetics and genomics database for Brassica plants. BMC Plant Biol. 11, 136. https://doi.org/10.1186/1471-2229-11-136

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5

Cuello, W.S., Gremer, J.R., Trimmer, P.C., Sih, A., Schreiber, S.J., 2019. Predicting evolutionarily stable strategies from functional responses of Sonoran Desert annuals to precipitation. Proc. R. Soc. B Biol. Sci. 286, 20182613. https://doi.org/10.1098/rspb.2018.2613

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R.A., Zhu, J.-K., Bohnert, H.J., Cheeseman, J.M., 2011. The genome of the extremophile crucifer Thellungiella parvula. Nat. Genet. 43, 913–918. https://doi.org/10.1038/ng.889

Dieringer, G., 1991. Variation in Individual Flowering Time and Reproductive Success of Agalinis Strictifolia (scrophulariaceae). Am. J. Bot. 78, 497–503. https://doi.org/10.1002/j.1537-2197.1991.tb15216.x

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 3, 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Fernández-Pascual, E., Pérez-Arcoiza, A., Prieto, J.A., Díaz, T.E., 2017. Environmental filtering drives the shape and breadth of the seed germination niche in coastal plant communities. Ann. Bot. 119, 1169–1177. https://doi.org/10.1093/aob/mcx005

Fernandez-Pozo, N., Metz, T., Chandler, J.O., Gramzow, L., Mérai, Z., Maumus, F., Mittelsten Scheid, O., Theißen, G., Schranz, M.E., Leubner-Metzger, G., Rensing, S.A., 2021. Aethionema arabicum genome annotation using PacBio full-length transcripts provides a valuable resource for seed dormancy and Brassicaceae evolution research. Plant J. 106, 275–293. https://doi.org/10.1111/tpj.15161

Ferreira de Carvalho, J., Stoeckel, S., Eber, F., Lodé-Taburel, M., Gilet, M.-M., Trotoux, G., Morice, J., Falentin, C., Chèvre, A.-M., Rousseau-Gueutin, M., 2021. Untangling structural factors driving genome stabilization in nascent Brassica napus allopolyploids. New Phytol. 230, 2072–2084. https://doi.org/10.1111/nph.17308

Flematti, G.R., Ghisalberti, E.L., Dixon, K.W., Trengove, R.D., 2004. A Compound from Smoke That Promotes Seed Germination. Science 305, 977–977. https://doi.org/10.1126/science.1099944

Flint, L.E., Flint, A.L., Stern, M.A., 2021. The basin characterization model—A regional water balance software package (No. 6-H1), Techniques and Methods. U.S. Geological Survey. https://doi.org/10.3133/tm6H1

Footitt, S., Walley, P.G., Lynn, J.R., Hambidge, A.J., Penfield, S., Finch-Savage, W.E., 2020. Trait analysis reveals DOG1 determines initial depth of seed dormancy, but not changes during dormancy cycling that result in seedling emergence timing. New Phytol. 225, 2035–2047. https://doi.org/10.1111/nph.16081

Freas, K.E., Kemp, P.R., 1983. Some Relationships Between Environmental Reliability and Seed Dormancy in Desert Annual Plants. J. Ecol. 71, 211–217. https://doi.org/10.2307/2259973

Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E., Osborn, T.C., 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19, 3403–3417. https://doi.org/10.1105/tpc.107.054346

Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R.D., Hofhuis, H., Pieper, B., Cartolano, M., Neumann, U., Nikolov, L.A., Song, B., Hajheidari, M., Briskine, R., Kougioumoutzi, E., Vlad, D., Broholm, S., Hein, J., Meksem, K., Lightfoot, D., Shimizu, K.K., Shimizu-Inatsugi, R., Imprialou, M., Kudrna, D., Wing, R., Sato, S., Huijser, P., Filatov, D., Mayer, K.F.X., Mott, R., Tsiantis, M., 2016. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nat. Plants 2, 1–7. https://doi.org/10.1038/nplants.2016.167

Geng, Y., Guan, Y., Qiong, L., Lu, S., An, M., Crabbe, M.J.C., Qi, J., Zhao, F., Qiao, Q., Zhang, T., 2021. Genomic analysis of field pennycress (Thlaspi arvense) provides insights into mechanisms of adaptation to high elevation. BMC Biol. 19, 143. https://doi.org/10.1186/s12915-021-01079-0

Giraudat, J., Hauge, B.M., Valon, C., Smalle, J., Parcy, F., Goodman, H.M., 1992. Isolation of the Arabidopsis ABI3 gene by positional cloning. Plant Cell 4, 1251–1261. https://doi.org/10.1105/tpc.4.10.1251

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 108, 1513–1518. https://doi.org/10.1073/pnas.1017351108

Goel, M., Schneeberger, K., 2022. plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics 38, 2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Goodstein, D M, Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Goodstein, David M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Graeber, K., Nakabayashi, K., Miatton, E., Leubner-Metzger, G., Soppe, W.J.J., 2012. Molecular mechanisms of seed dormancy. Plant Cell Environ. 35, 1769–1786. https://doi.org/10.1111/j.1365-3040.2012.02542.x

Gremer, J.R., Chiono, A., Suglia, E., Bontrager, M., Okafor, L., Schmitt, J., 2020a. Variation in the seasonal germination niche across an elevational gradient: the role of germination cueing in current and future climates. Am. J. Bot. 107, 350–363. https://doi.org/10.1002/ajb2.1425

Gremer, J.R., Chiono, A., Suglia, E., Bontrager, M., Okafor, L., Schmitt, J., 2020b. Variation in the seasonal germination niche across an elevational gradient: the role of germination cueing in current and future climates. Am. J. Bot. 107, 350–363. https://doi.org/10.1002/ajb2.1425

Gremer, J.R., Wilcox, C.J., Chiono, A., Suglia, E., Schmitt, J., 2020c. Germination timing and chilling exposure create contingency in life history and influence fitness in the native wildflower Streptanthus tortuosus. J. Ecol. 108, 239–255. https://doi.org/10.1111/1365-2745.13241

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8. https://doi.org/10.1038/nprot.2013.084

Han, Y., Watanabe, S., Shimada, H., Sakamoto, A., 2020. Dynamics of the leaf endoplasmic reticulum modulate β-glucosidase-mediated stress-activated ABA production from its glucosyl ester. J. Exp. Bot. 71, 2058–2071. https://doi.org/10.1093/jxb/erz528

Higgins, E.E., Clarke, W.E., Howell, E.C., Armstrong, S.J., Parkin, I.A.P., 2018. Detecting de Novo Homoeologous Recombination Events in Cultivated Brassica napus Using a Genome-Wide SNP Array. G3 Bethesda Md 8, 2673–2683. https://doi.org/10.1534/g3.118.200118

Higgins, E.E., Howell, E.C., Armstrong, S.J., Parkin, I.A.P., 2021. A major quantitative trait locus on chromosome A9, BnaPh1, controls homoeologous recombination in Brassica napus. New Phytol. 229, 3281–3293. https://doi.org/10.1111/nph.16986

Ho, L.S.T., Ané, C., 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. Methods Ecol. Evol. 5, 1133–1146. https://doi.org/10.1111/2041-210X.12285

Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491. https://doi.org/10.1186/1471-2105-12-491

Howell, J., 1957. The California flora and its province. Leafl. West. Bot. 133–138.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F.X., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D., Guo, Y.-L., 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43, 476–481. https://doi.org/10.1038/ng.807

Huang, X., Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9, 868–877.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.-K.K., Tirnaz, S., Dolatabadian, A., Schiessl, S.V., Samans, B., Montenegro, J.D., Parkin, I.A.P., Pires, J.C., Chalhoub, B., King, G.J., Snowdon, R., Batley, J., Edwards, D., 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol. J. 16, 1265–1274. https://doi.org/10.1111/pbi.12867

Istace, B., Belser, C., Falentin, C., Labadie, K., Boideau, F., Deniot, G., Maillet, L., Cruaud, C., Bertrand, L., Chèvre, A.-M., Wincker, P., Rousseau-Gueutin, M., Aury, J.-M., 2021. Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case. Biology 10, 732. https://doi.org/10.3390/biology10080732

Ivalú Cacho, N., Millie Burrell, A., Pepper, A.E., Strauss, S.Y., 2014. Novel nuclear markers inform the systematics and the evolution of serpentine use in Streptanthus and allies (Thelypodieae, Brassicaceae). Mol. Phylogenet. Evol. 72, 71–81. https://doi.org/10.1016/j.ympev.2013.11.018

Jia, K.-H., Wang, Z.-X., Wang, L., Li, G.-Y., Zhang, W., Wang, X.-L., Xu, F.-J., Jiao, S.-Q., Zhou, S.-S., Liu, H., Ma, Y., Bi, G., Zhao, W., El-Kassaby, Y.A., Porth, I., Li, G., Zhang, R.-G., Mao, J.-F., 2022. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. New Phytol. 235, 801–809. https://doi.org/10.1111/nph.18173

Jiao, W.-B., Accinelli, G.G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.-M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M.A., Swan, D., Clavijo, B., Coupland, G., Schneeberger, K., 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. 27, 778–786. https://doi.org/10.1101/gr.213652.116

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter,

S., 2014. InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kagale, S., Robinson, S.J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W.E., Edger, P.P., Links, M.G., Sharpe, A.G., Parkin, I.A.P., 2014. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. Plant Cell 26, 2777–2791. https://doi.org/10.1105/tpc.114.126391

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Keller, O., Kollmar, M., Stanke, M., Waack, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics 27, 757–763. https://doi.org/10.1093/bioinformatics/btr010

Kleczkowski, L.A., Igamberdiev, A.U., 2021. Magnesium Signaling in Plants. Int. J. Mol. Sci. 22, 1159. https://doi.org/10.3390/ijms22031159

Koornneef, M., Reuling, G., Karssen, C.M., 1984. The isolation and characterization of abscisic acid-insensitive mutants of Arabidopsis thaliana. Physiol. Plant. 61, 377–383. https://doi.org/10.1111/j.1399-3054.1984.tb06343.x

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. 27, 722–736. https://doi.org/10.1101/gr.215087.116

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinformatics 9.

Kruckeberg, A.R., 2006. Introduction to California Soils and Plants.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. https://doi.org/10.1186/1471-2105-9-559

Lee, H., Chawla, H.S., Obermeier, C., Dreyer, F., Abbadi, A., Snowdon, R., 2020. Chromosome-Scale Assembly of Winter Oilseed Rape Brassica napus. Front. Plant Sci. 11.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Jeong, K., Davis, J.T., Kim, Seungmo, Lee, S., Michelmore, R.W., Kim, Shinje, Maloof, J.N., 2018. Integrated QTL and eQTL Mapping Provides Insights and Candidate Genes for Fatty Acid Composition, Flowering Time, and Growth Traits in a F2 Population of a Novel Synthetic Allopolyploid Brassica napus. Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.01632

Lloyd, A., Blary, A., Charif, D., Charpentier, C., Tran, J., Balzergue, S., Delannoy, E., Rigaill, G., Jenczewski, E., 2018. Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. New Phytol. 217, 367–377. https://doi.org/10.1111/nph.14836

Löytynoja, A., 2014. Phylogeny-aware alignment with PRANK. Methods Mol. Biol. Clifton NJ 1079, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10

Lynch, M., Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290, 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lysak, M.A., Mandáková, T., Schranz, M.E., 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr. Opin. Plant Biol. 30, 108–115. https://doi.org/10.1016/j.pbi.2016.02.001

Macías, L.G., Barrio, E., Toft, C., 2020. GWideCodeML: A Python Package for Testing Evolutionary Hypotheses at the Genome-Wide Level. G3 GenesGenomesGenetics 10, 4369–4372. https://doi.org/10.1534/g3.120.401874

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., Van de Peer, Y., 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. 102, 5454–5459. https://doi.org/10.1073/pnas.0501102102

Mandáková, T., Li, Z., Barker, M.S., Lysak, M.A., 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91, 3–21. https://doi.org/10.1111/tpj.13553

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018a. MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018b. MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Mendes, F.K., Vanderpool, D., Fulton, B., Hahn, M.W., 2021. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36, 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mondoni, A., Rossi, G., Orsenigo, S., Probert, R.J., 2012. Climate warming could shift the timing of seed germination in alpine plants. Ann. Bot. 110, 155–164. https://doi.org/10.1093/aob/mcs097

Murase, K., Hirano, Y., Sun, T., Hakoshima, T., 2008. Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. Nature 456, 459–463. https://doi.org/10.1038/nature07519

Nagaharu, U., 1935. Genome Analysis in Brassica with Special Reference to the Experimental Formation of B. Napus and Peculiar Mode of Fertilization. Jpn. J. Bot. 389–452.

Nowak, M.D., Birkeland, S., Mandáková, T., Roy Choudhury, R., Guo, X., Gustafsson, A.L.S., Gizaw, A., Schrøder-Nielsen, A., Fracassetti, M., Brysting, A.K., Rieseberg, L., Slotte, T., Parisod, C., Lysak, M.A., Brochmann, C., 2021. The genome of Draba nivalis shows signatures of adaptation to the extreme environmental stresses of the Arctic. Mol. Ecol. Resour. 21, 661–676. https://doi.org/10.1111/1755-0998.13280

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-86659-3

Okamoto, M., Kuwahara, A., Seo, M., Kushiro, T., Asami, T., Hirai, N., Kamiya, Y., Koshiba, T., Nambara, E., 2006. CYP707A1 and CYP707A2, Which Encode Abscisic Acid 8′-Hydroxylases, Are Indispensable for Proper Control of Seed Dormancy and Germination in Arabidopsis. Plant Physiol. 141, 97–107. https://doi.org/10.1104/pp.106.079475

Ollerton, J., Lack, A., 1998. Relationships between flowering phenology, plant size and reproductive success in shape Lotus corniculatus (Fabaceae). Plant Ecol. 139, 35–47. https://doi.org/10.1023/A:1009798320049

Oplinger, E.S., Hardman, L.L., Gritton, E.T., Doll, J.D., Kelling, K., 1989. Canola (Rapeseed): Alternative Field Crops Manual. Collect. Altern. Field Crops Man.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., Lynch, M., 2010. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science 327, 92–94. https://doi.org/10.1126/science.1180677

PacificBiosciences/pbbioconda: PacBio Secondary Analysis Tools on Bioconda. Contains list of PacBio packages available via conda. [WWW Document], n.d. URL https://github.com/PacificBiosciences/pbbioconda (accessed 3.13.24).

Padfield, D., O'Sullivan, H., Pawar, S., 2021. rTPC and nls.multstart: A new pipeline to fit thermal performance curves in r. Methods Ecol. Evol. 12, 1138–1143. https://doi.org/10.1111/2041-210X.13585

Pearse, I.S., Aguilar, J.M., Strauss, S.Y., 2020. Life-History Plasticity and Water-Use Trade-Offs Associated with Drought Resistance in a Clade of California Jewelflowers. Am. Nat. 195, 691–704. https://doi.org/10.1086/707371

Pearse, I.S., McIntyre, P., Cacho, N.I., Strauss, S.Y., 2022. Fitness homeostasis across an experimental water gradient predicts species' geographic range and climatic breadth. Ecology 103, e3827. https://doi.org/10.1002/ecy.3827

Petri, A.J., Sahlin, K., 2023. isONform: reference-free transcriptome reconstruction from Oxford Nanopore data. Bioinformatics 39, i222–i231. https://doi.org/10.1093/bioinformatics/btad264

Postma, F.M., Ågren, J., 2016. Early life stages contribute strongly to local adaptation in Arabidopsis thaliana. Proc. Natl. Acad. Sci. 113, 7590–7595. https://doi.org/10.1073/pnas.1606303113

PSD Online [WWW Document], n.d. URL https://apps.fas.usda.gov/psdonline/app/index.html#/app/downloads (accessed 1.24.19).

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., Paterson, A.H., 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 20, 38. https://doi.org/10.1186/s13059-019-1650-2

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team, 2021. R: A Language and Environment for Statistical Computing.

R Core Team, 2013. R: A language and environment for statistical computing.

Raman, H., Raman, R., Pirathiban, R., McVittie, B., Sharma, N., Liu, S., Qiu, Y., Zhu, A., Kilian, A., Cullis, B., Farquhar, G.D., Stuart-Williams, H., White, R., Tabah, D., Easton, A., Zhang, Y., 2022. Multienvironment QTL analysis delineates a major locus associated with homoeologous exchanges for water-use efficiency and seed yield in canola. Plant Cell Environ. 45, 2019–2036. https://doi.org/10.1111/pce.14337

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N., Delsuc, F., 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol. Biol. Evol. 35, 2582–2584. https://doi.org/10.1093/molbev/msy159

Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P., Aiden, E.L., 2018. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. 6, 256-258.e1. https://doi.org/10.1016/j.cels.2018.01.001

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rodríguez-Pérez, J., Traveset, A., 2016. Effects of flowering phenology and synchrony on the reproductive success of a long-flowering shrub. AoB PLANTS 8, plw007. https://doi.org/10.1093/aobpla/plw007

Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C., Falentin, C., Boideau, F., Boutte, J., Delourme, R., Deniot, G., Engelen, S., de Carvalho, J.F., Lemainque, A., Maillet, L., Morice, J., Wincker, P., Denoeud, F., Chèvre, A.-M., Aury, J.-M., 2020. Long-read assembly of the Brassica napus reference genome Darmor-bzh. GigaScience 9, giaa137. https://doi.org/10.1093/gigascience/giaa137

Roux, F., Touzet, P., Cuguen, J., Le Corre, V., 2006. How to be early flowering: an evolutionary perspective. Trends Plant Sci. 11, 375–381. https://doi.org/10.1016/j.tplants.2006.06.006

Safford, H.D., Viers, J.H., Harrison, S.P., 2005. SERPENTINE ENDEMISM IN THE CALIFORNIA FLORA: A DATABASE OF SERPENTINE AFFINITY. Madroño 52, 222–257. https://doi.org/10.3120/0024-9637(2005)52[222:SEITCF]2.0.CO;2

Sahlin, K., Medvedev, P., 2020. De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm. J. Comput. Biol. 27, 472–484. https://doi.org/10.1089/cmb.2019.0299

Sajeev, N., Koornneef, M., Bentsink, L., 2024. A commitment for life: Decades of unraveling the molecular mechanisms behind seed dormancy and germination. Plant Cell koad328. https://doi.org/10.1093/plcell/koad328

Samans, B., Chalhoub, B., Snowdon, R.J., 2017. Surviving a Genome Collision: Genomic Signatures of Allopolyploidization in the Recent Crop Species Brassica napus. Plant Genome 10, plantgenome2017.02.0013. https://doi.org/10.3835/plantgenome2017.02.0013

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., Sherry, S.T., 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res. 50, D20–D26. https://doi.org/10.1093/nar/gkab1112

Schmitt, J., 1983. Individual flowering phenology, plant size, and reproductive success in Linanthus androsaceus, a California annual. Oecologia 59, 135–140. https://doi.org/10.1007/BF00388084

Schütz, W., Rave, G., 1999. The effect of cold stratification and light on the seed germination of temperate sedges (Carex) from various habitats and implications for regenerative strategies. Plant Ecol. 144, 215–230. https://doi.org/10.1023/A:1009892004730

Scott, C., 2016. dammit: an open and accessible de novo transcriptome annotator. Prep.

Shu, K., Liu, X., Xie, Q., He, Z., 2016. Two Faces of One Seed: Hormonal Regulation of Dormancy and Germination. Mol. Plant, Plant Hormones 9, 34–45. https://doi.org/10.1016/j.molp.2015.08.010

Skubacz, A., Daszkowska-Golec, A., Szarejko, I., 2016. The Role and Regulation of ABI5 (ABA-Insensitive 5) in Plant Development, Abiotic Stress Responses and Phytohormone Crosstalk. Front. Plant Sci. 7. https://doi.org/10.3389/fpls.2016.01884

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., Guo, L., 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6, 34–45. https://doi.org/10.1038/s41477-019-0577-7

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S.V., Obermeier, C., Parkin, I.A.P., Chèvre, A.-M., Snowdon, R.J., 2017. Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus. Plant Biotechnol. J. 15, 1478–1489. https://doi.org/10.1111/pbi.12732

Steinbrecher, T., Leubner-Metzger, G., 2017. The biomechanics of seed germination. J. Exp. Bot. 68, 765–783. https://doi.org/10.1093/jxb/erw428

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011a. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE 6, e21800. https://doi.org/10.1371/journal.pone.0021800

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011b. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE 6, e21800. https://doi.org/10.1371/journal.pone.0021800

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–W612. https://doi.org/10.1093/nar/gkl315

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., Paterson, A.H., 2008. Synteny and Collinearity in Plant Genomes. Science 320, 486–488. https://doi.org/10.1126/science.1153917

Tang, R.-J., Zhao, F.-G., Garcia, V.J., Kleist, T.J., Yang, L., Zhang, H.-X., Luan, S., 2015. Tonoplast CBL–CIPK calcium signaling network regulates magnesium homeostasis in Arabidopsis. Proc. Natl. Acad. Sci. 112, 3134–3139. https://doi.org/10.1073/pnas.1420944112

ten Brink, H., Gremer, J.R., Kokko, H., 2020. Optimal germination timing in unpredictable environments: the importance of dormancy for both among- and within-season variation. Ecol. Lett. 23, 620–630. https://doi.org/10.1111/ele.13461

The UniProt Consortium, 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531. https://doi.org/10.1093/nar/gkac1052

Tielbörger, K., Petrů, M., Lampei, C., 2012. Bet-hedging germination in annual plants: a sound empirical test of the theoretical foundations. Oikos 121, 1860–1868.

Torres-Martínez, L., Weldy, P., Levy, M., Emery, N.C., 2017. Spatiotemporal heterogeneity in precipitation patterns explain population-level germination strategies in an edaphic specialist. Ann. Bot. 119, 253–265. https://doi.org/10.1093/aob/mcw161

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. https://doi.org/10.1038/nbt.1621

Tung Ho, L. si, Ané, C., 2014. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. Syst. Biol. 63, 397–408. https://doi.org/10.1093/sysbio/syu005

Udall, J.A., Quijada, P.A., Osborn, T.C., 2005. Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of Brassica napus L. Genetics 169, 967–979. https://doi.org/10.1534/genetics.104.033209

USDA, F.A.S., n.d. Publication | Oilseeds: World Markets and Trade | ID: tx31qh68h | USDA Economics, Statistics and Market Information System [WWW Document]. URL https://usda.library.cornell.edu/concern/publications/tx31qh68h?locale=en (accessed 8.24.22).

Uyeda, J.C., Harmon, L.J., 2014. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. Syst. Biol. 63, 902–918. https://doi.org/10.1093/sysbio/syu057

Venable, D.L., 2007. Bet Hedging in a Guild of Desert Annuals. Ecology 88, 1086–1090. https://doi.org/10.1890/06-1495

Venable, D.L., Lawlor, L., 1980. Delayed germination and dispersal in desert annuals: Escape in space and time. Oecologia 46, 272–282. https://doi.org/10.1007/BF00540137

Wang, X., Yu, K., Li, H., Peng, Q., Chen, F., Zhang, W., Chen, S., Hu, M., Zhang, J., 2015. High-Density SNP Map Construction and QTL Identification for the Apetalous Character in Brassica napus L. Front. Plant Sci. 6. https://doi.org/10.3389/fpls.2015.01164

Weber, M.G., Cacho, N.I., Phan, M.J.Q., Disbrow, C., Ramírez, S.R., Strauss, S.Y., 2018. The evolution of floral signals in relation to range overlap in a clade of California Jewelflowers (Streptanthus s.l.). Evolution 72, 798–807. https://doi.org/10.1111/evo.13456

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B., 2017. Direct determination of diploid genome sequences. Genome Res. 27, 757–767. https://doi.org/10.1101/gr.214874.116

Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York.

Wind, J.J., Peviani, A., Snel, B., Hanson, J., Smeekens, S.C., 2013. ABI4: versatile activator and repressor. Trends Plant Sci. 18, 125–132. https://doi.org/10.1016/j.tplants.2012.10.004

Worthy, S.J., Miller, A., Ashlock, S.R., Ceviker, E., Maloof, J.N., Strauss, S.Y., Schmitt, J., Gremer, J.R., 2023. Germination responses to changing rainfall timing reveal potential climate vulnerability in a clade of wildflowers. https://doi.org/10.1101/2023.03.22.533835

Xiong, Z., Gaeta, R.T., Edger, P.P., Cao, Y., Zhao, K., Zhang, S., Pires, J.C., 2021. Chromosome inheritance and meiotic stability in allopolyploid Brassica napus. G3 Bethesda Md 11, jkaa011. https://doi.org/10.1093/g3journal/jkaa011

Xiong, Z., Gaeta, R.T., Pires, J.C., 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc. Natl. Acad. Sci. 108, 7908–7913. https://doi.org/10.1073/pnas.1014138108

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R14. https://doi.org/10.1186/gb-2010-11-2-r14

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, X., Liu, T., Wang, J., Wang, P., Qiu, Y., Zhao, W., Pang, S., Li, Xiaoman, Wang, H., Song, J., Zhang, W., Yang, W., Sun, Y., Li, Xixiang, 2021. Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild, and weedy radishes. Mol. Plant 14, 2032–2055. https://doi.org/10.1016/j.molp.2021.08.005

Zhou, C., McCarthy, S.A., Durbin, R., 2022. YaHS: yet another Hi-C scaffolding tool. Bioinformatics btac808. https://doi.org/10.1093/bioinformatics/btac808

Supplemental Materials

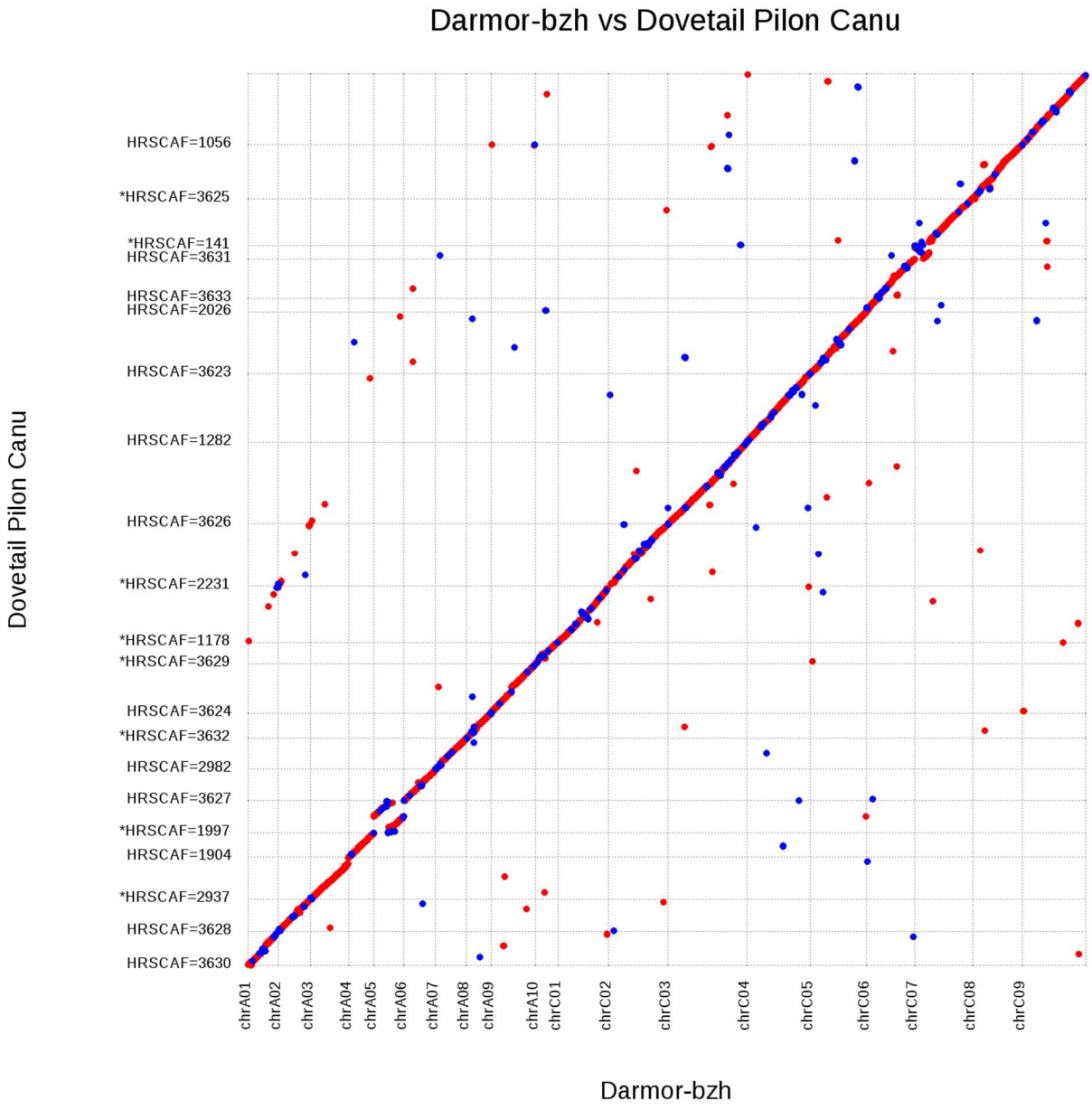

Figure S1.1*. Nucmer plot of Dovetail_Pilon_Canu aligned to Darmor-bzh v 4.1 chromosomes. All sequences aligned are 1 Mbp or greater. A total of 21 Dovetail_Pilon_Canu scaffolds are aligned to 19 reference chromosomes. Red indicates an alignment in the forward direction and blue indicates an alignment in the reverse direction.*

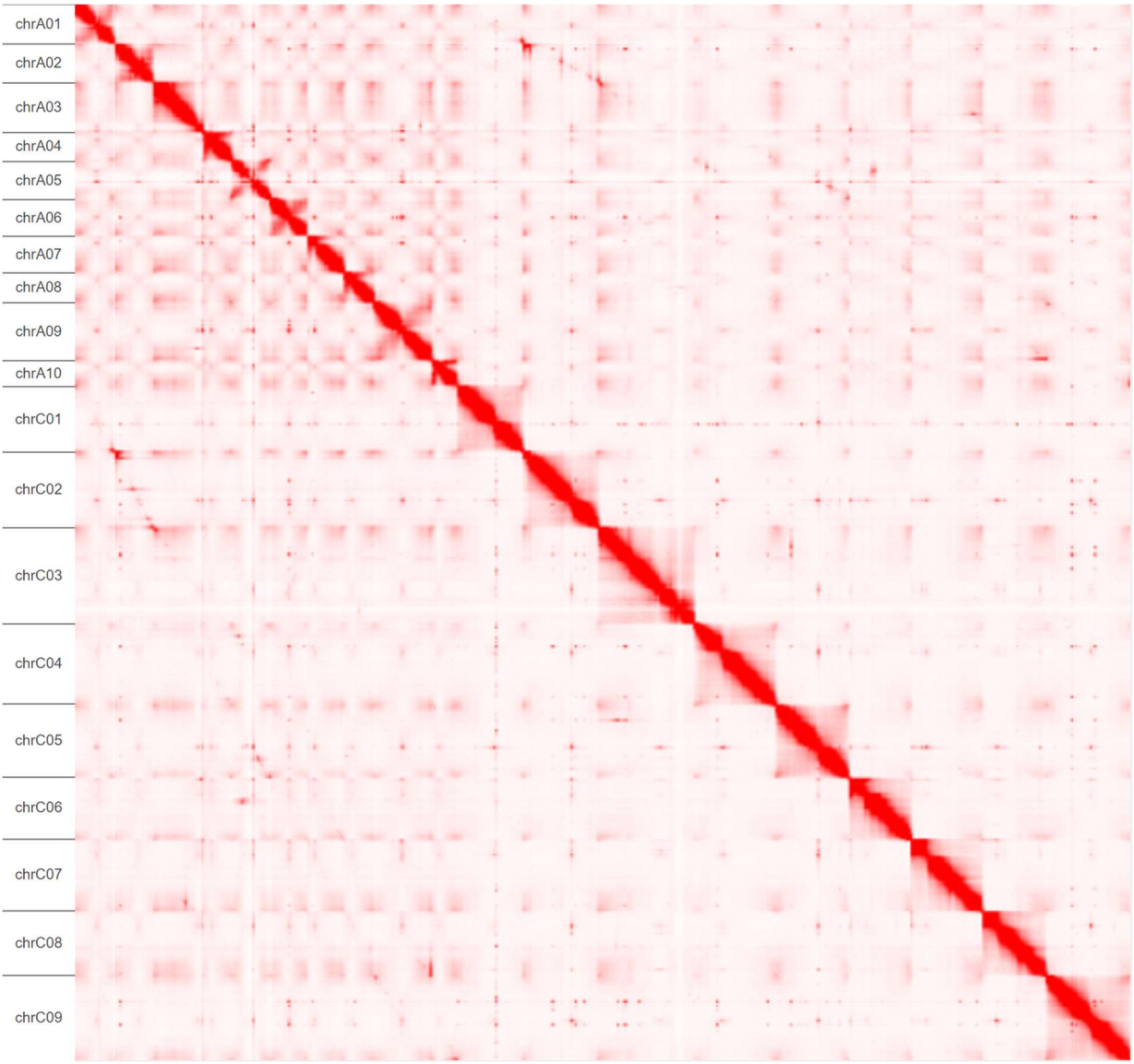

Figure S1.2. *Hi-C Contact map.*

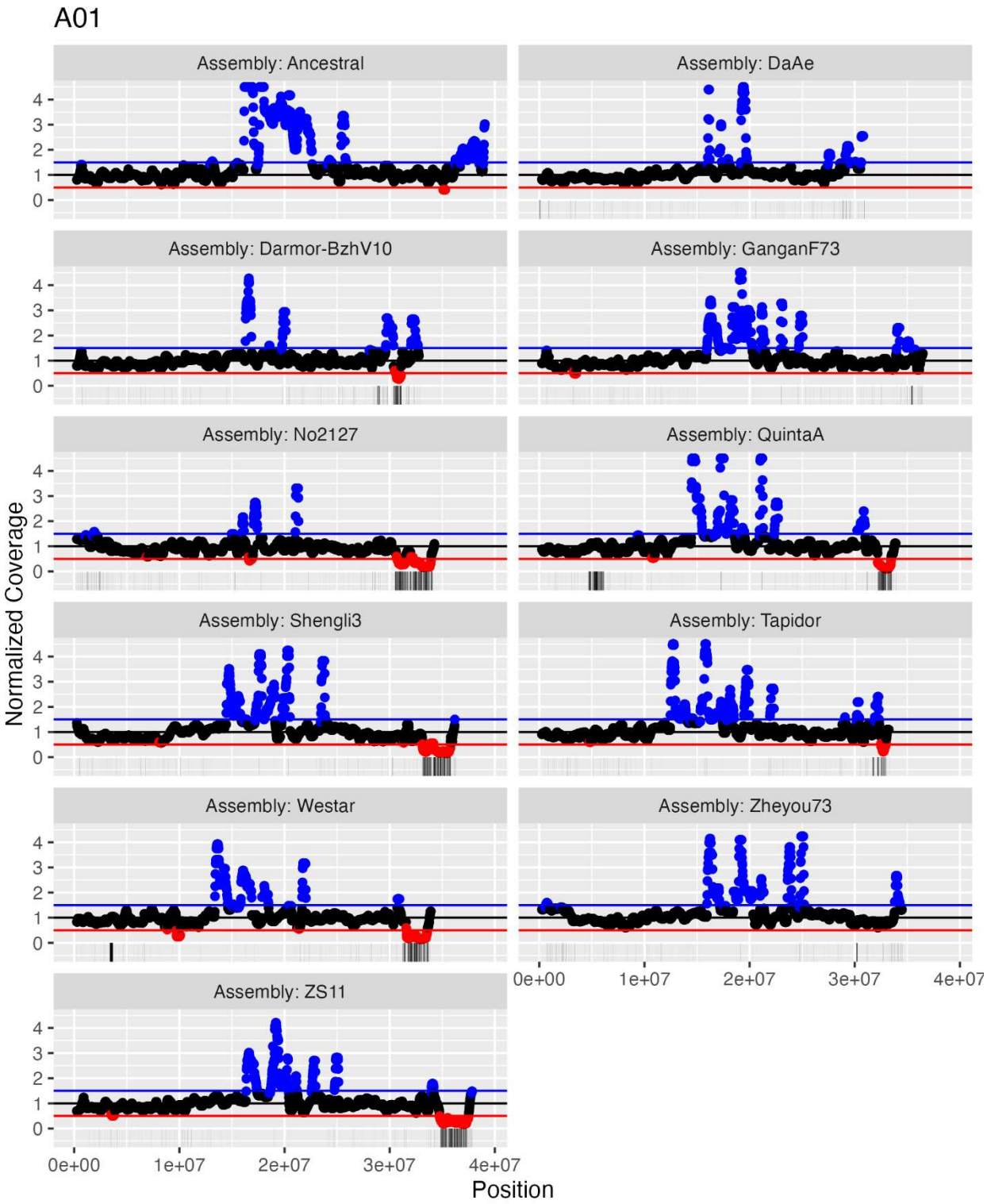

Figure S1.3. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

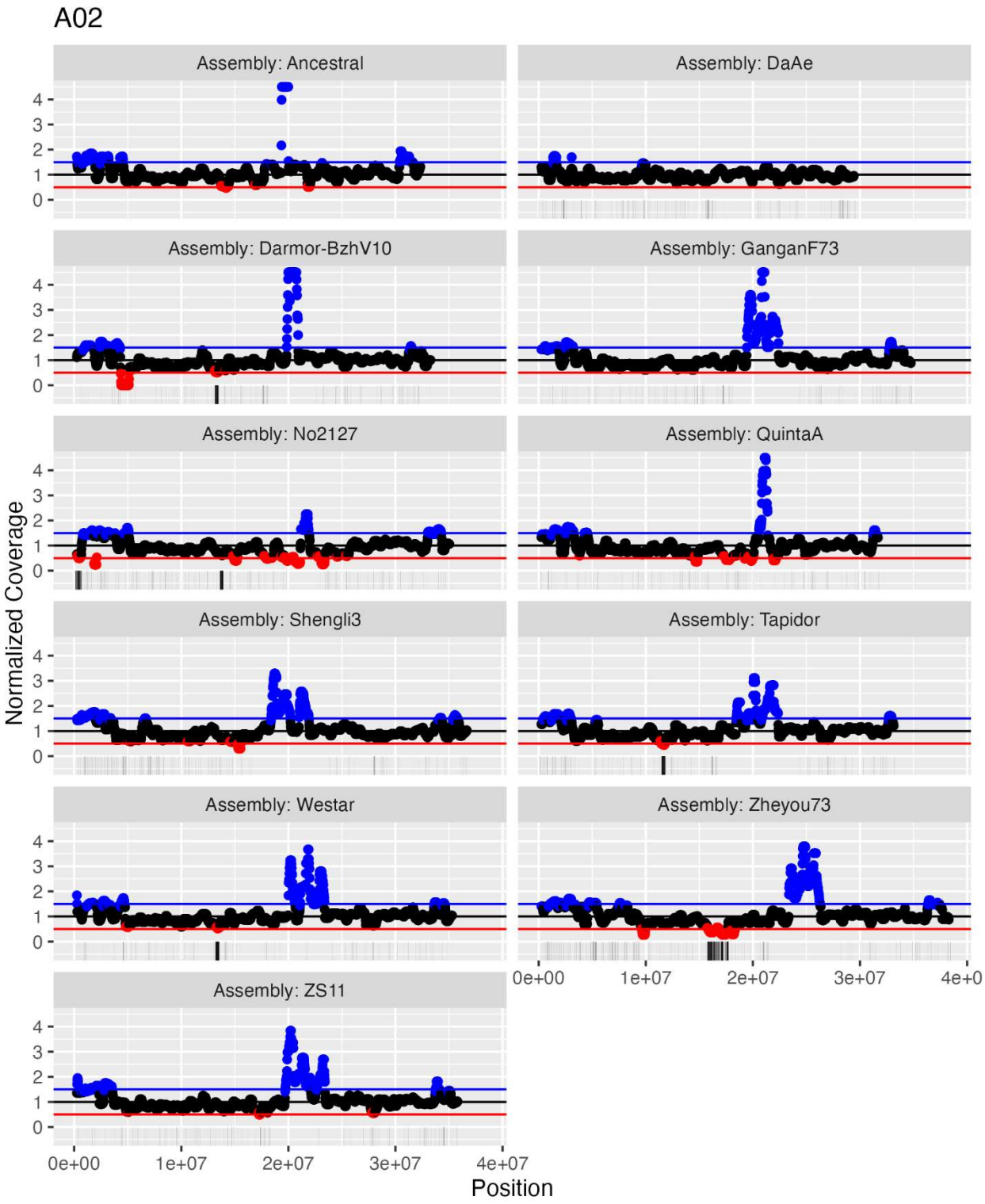

Figure S1.4. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

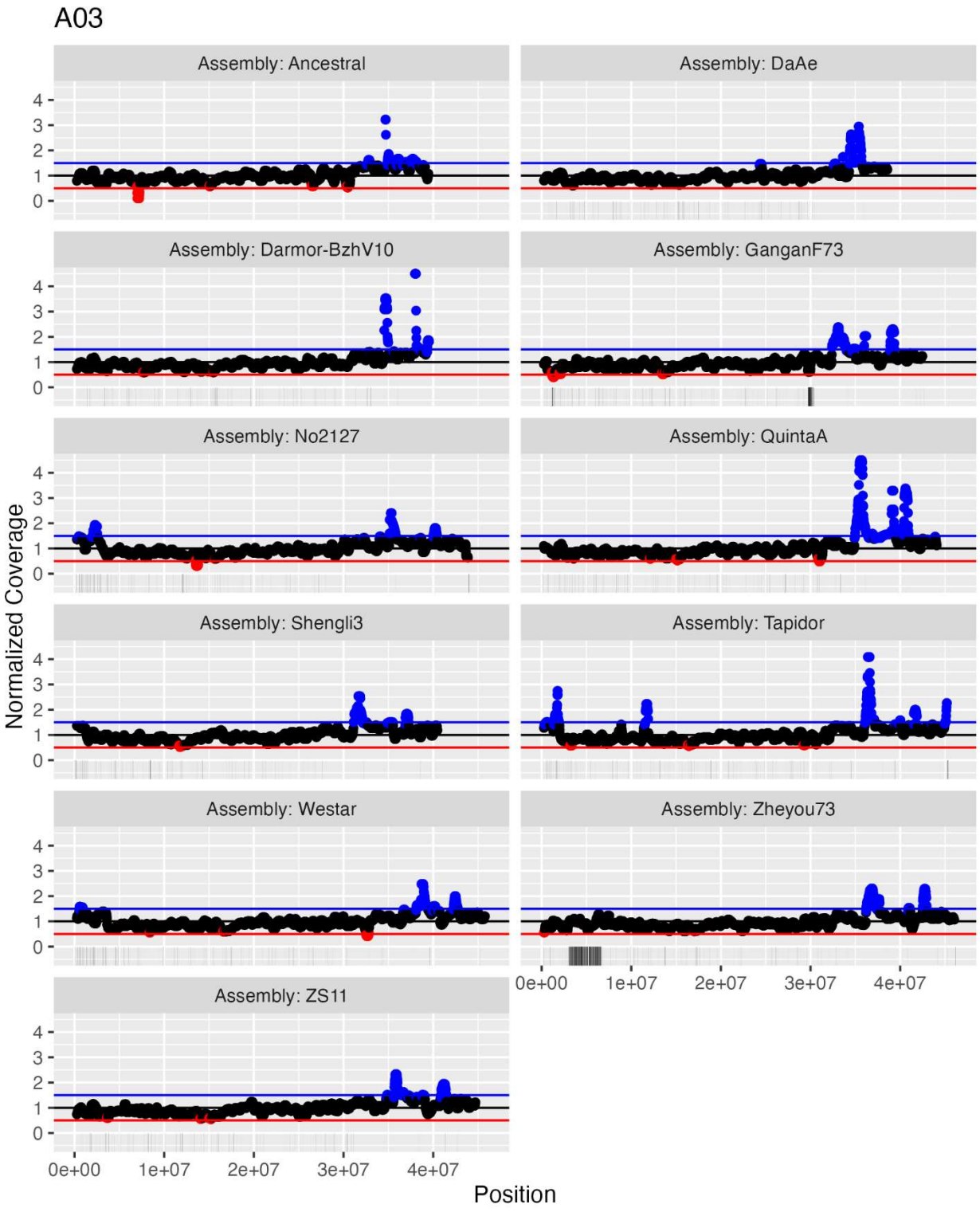

Figure S1.5. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

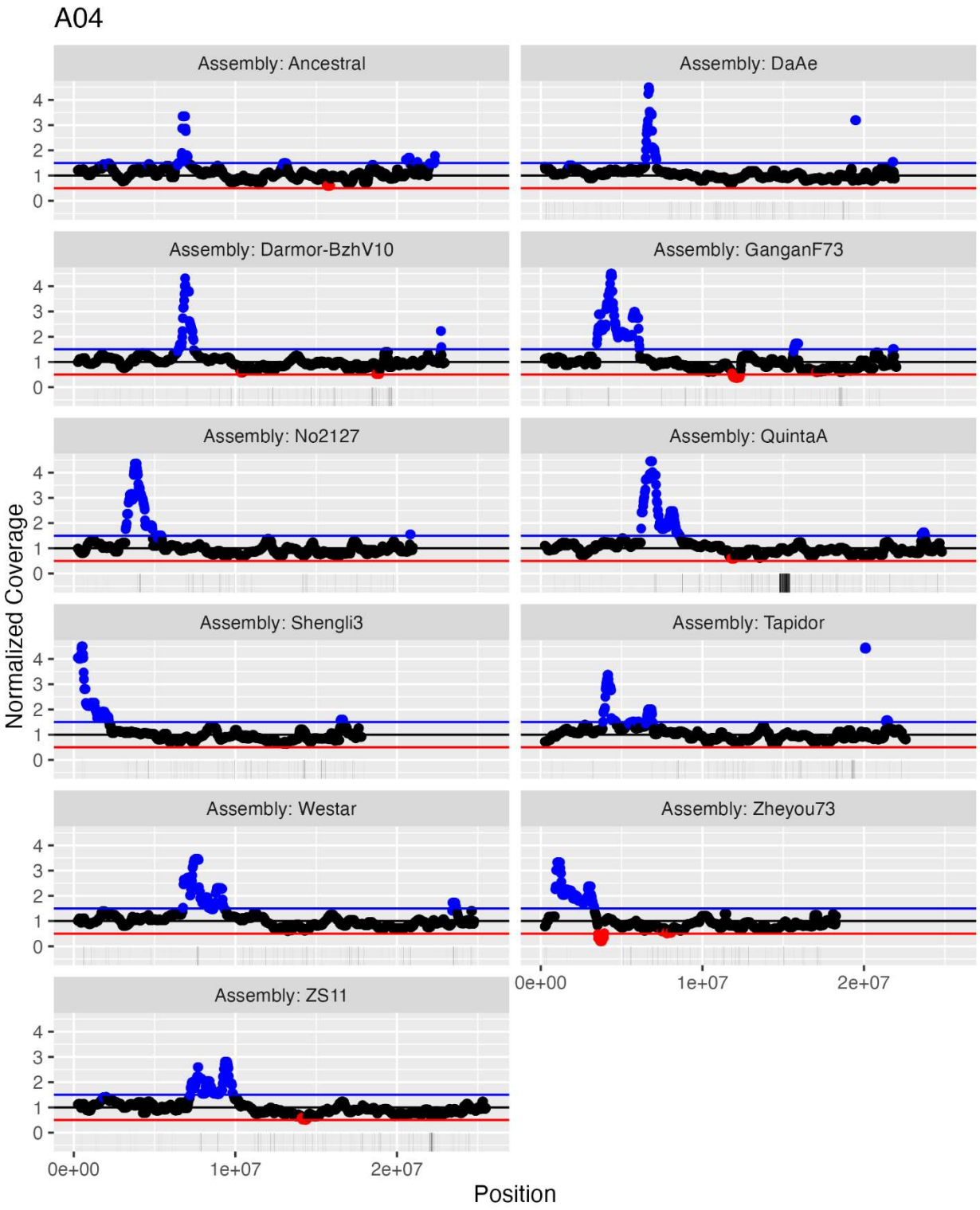

Figure S1.6. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

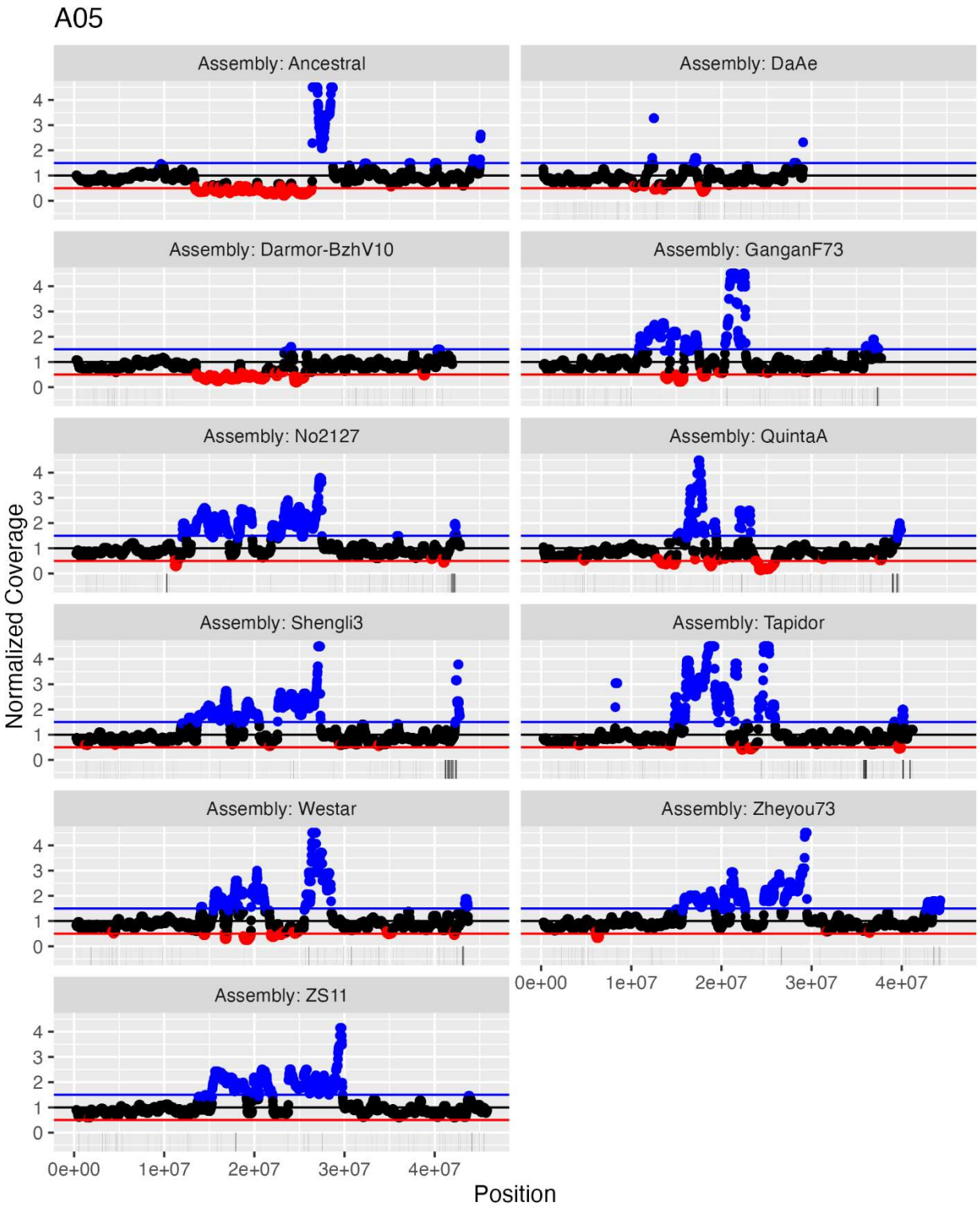

Figure S1.7. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

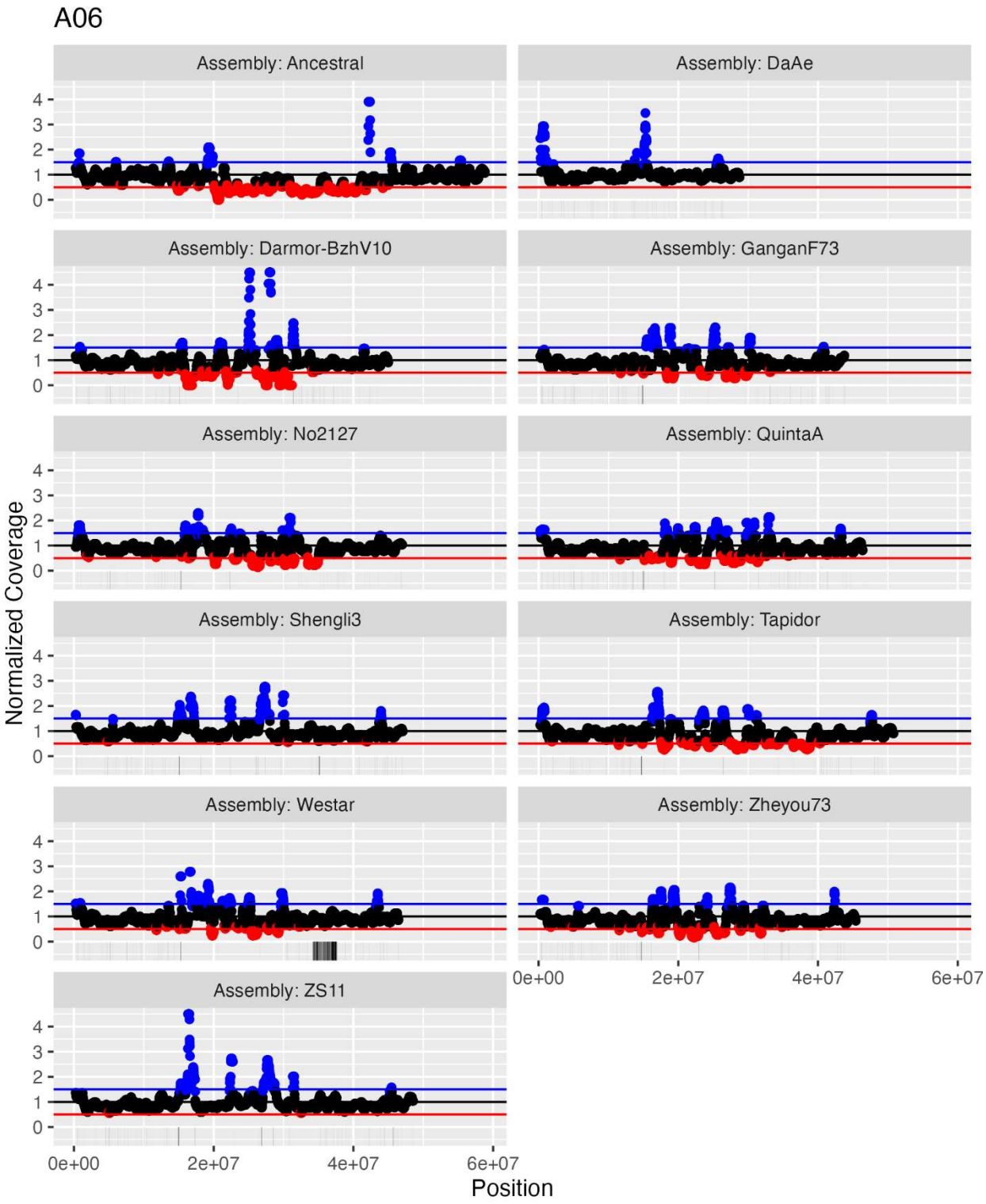

Figure S1.8. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

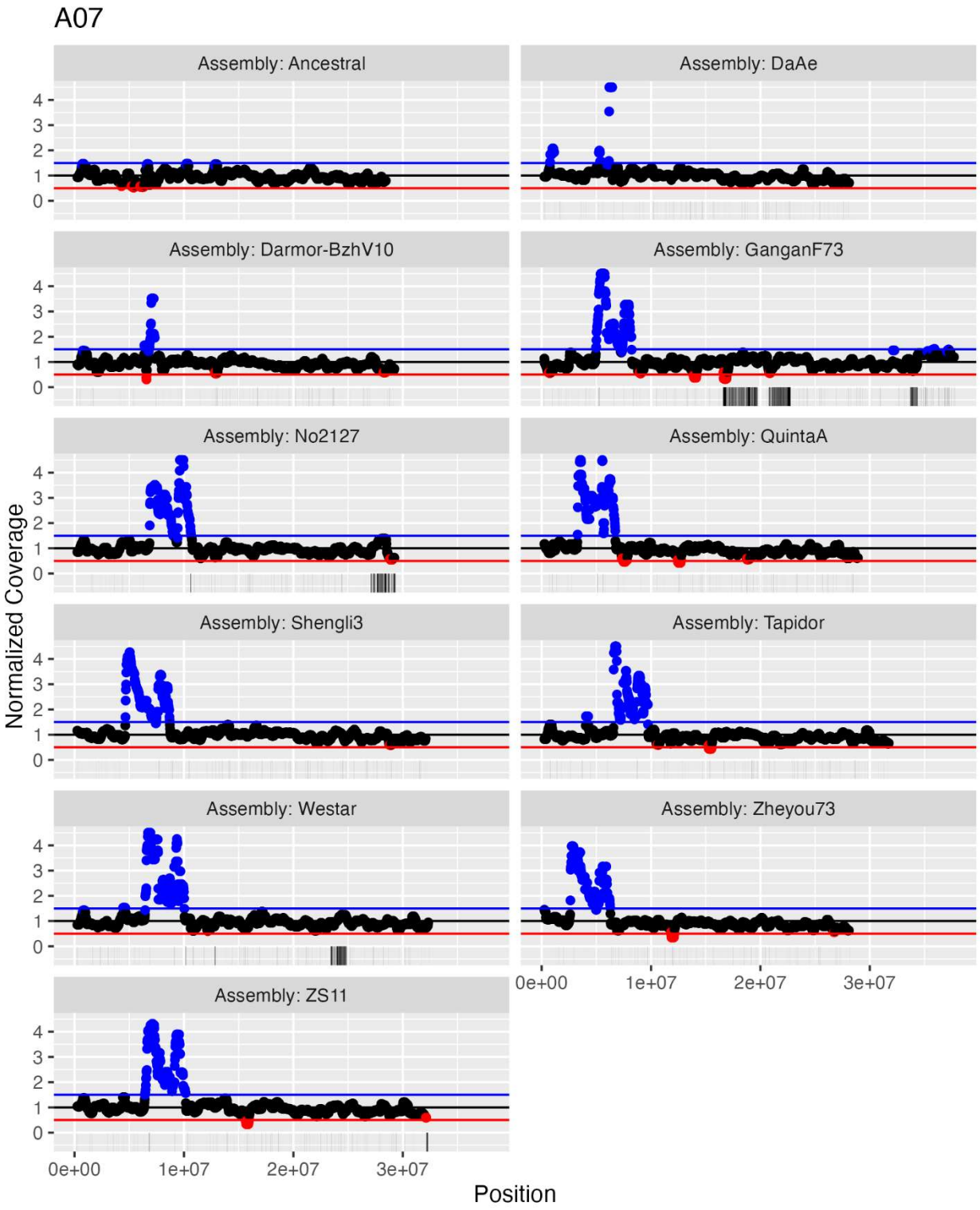

Figure S1.9. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

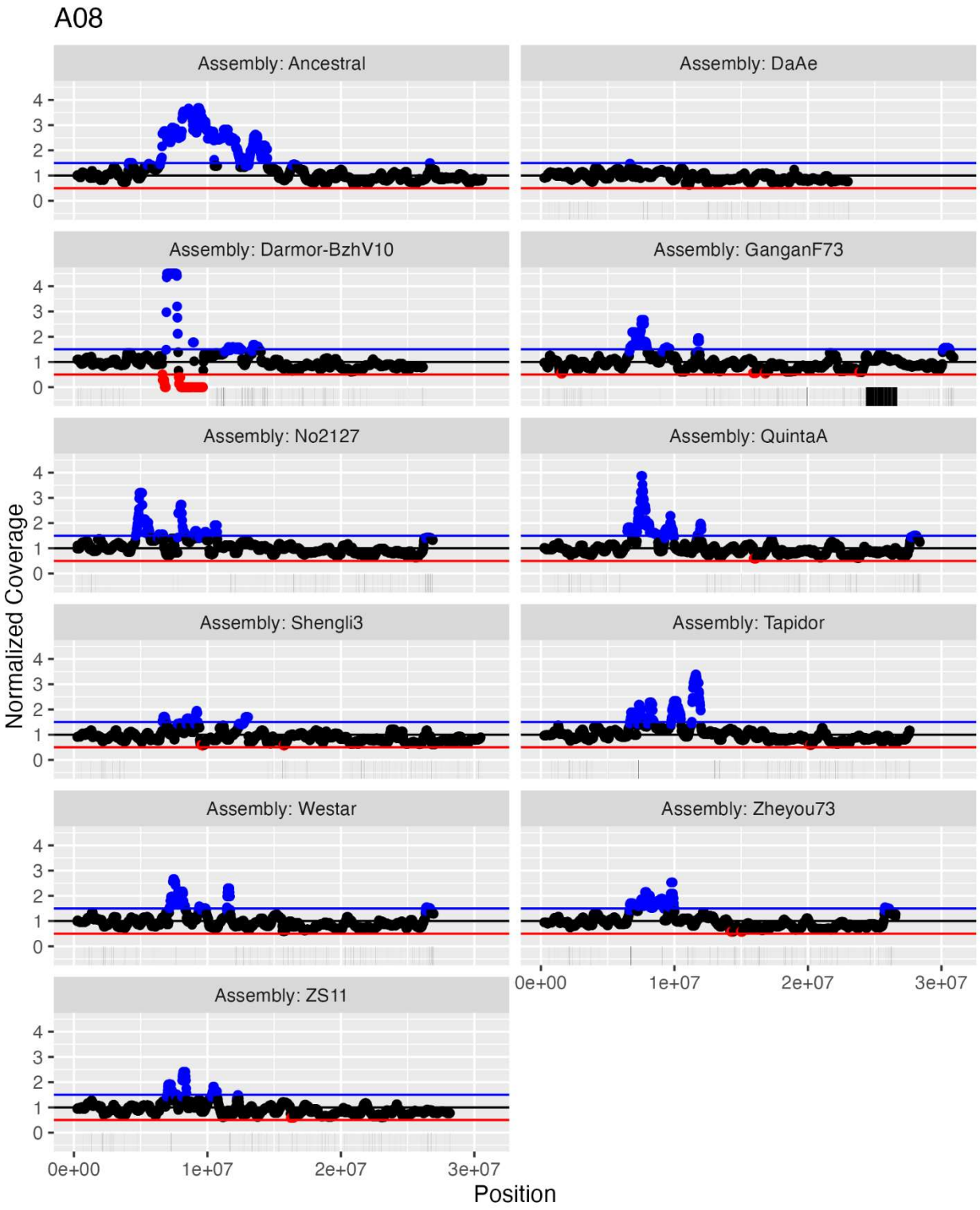

Figure S1.10. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

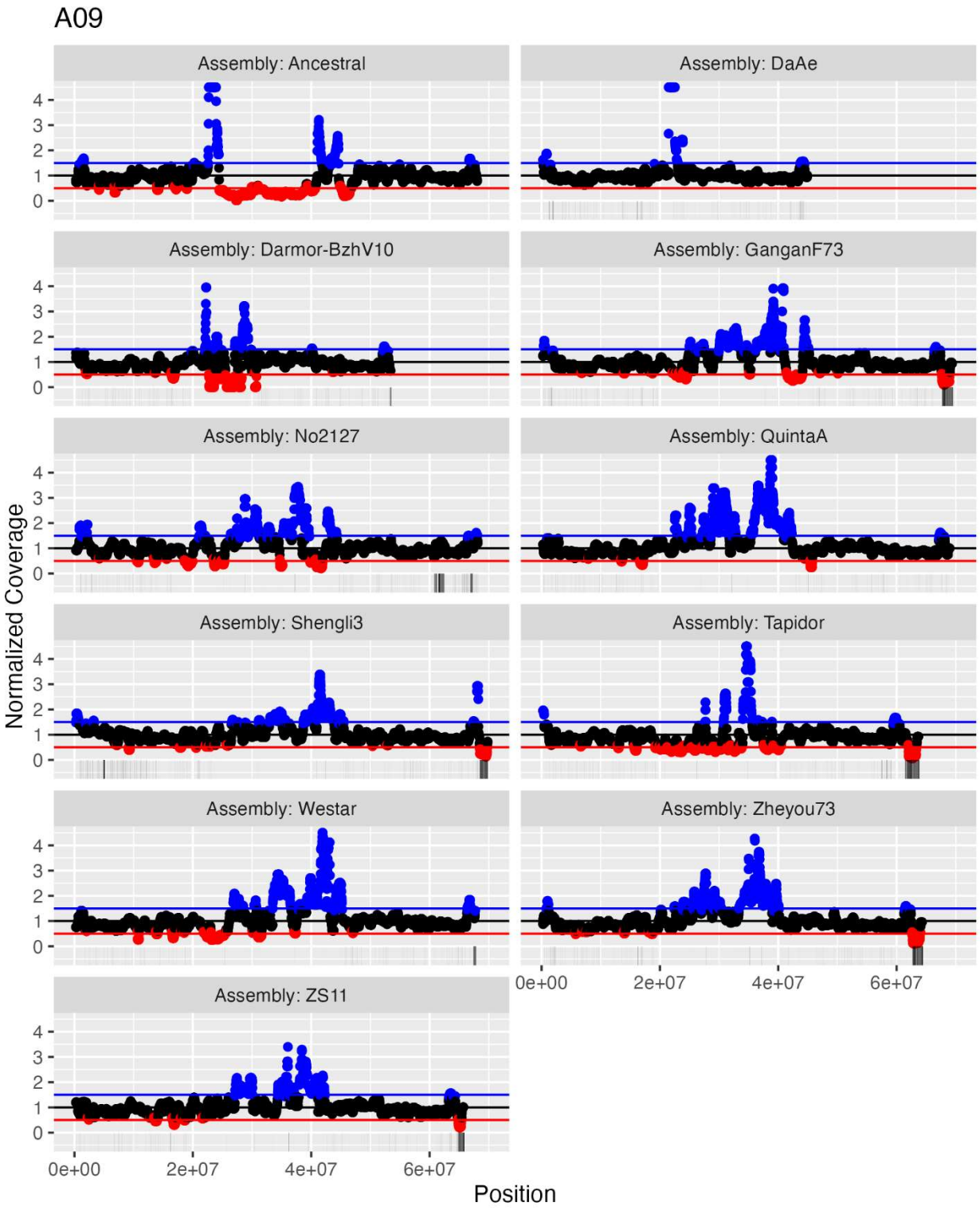

Figure S1.11. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

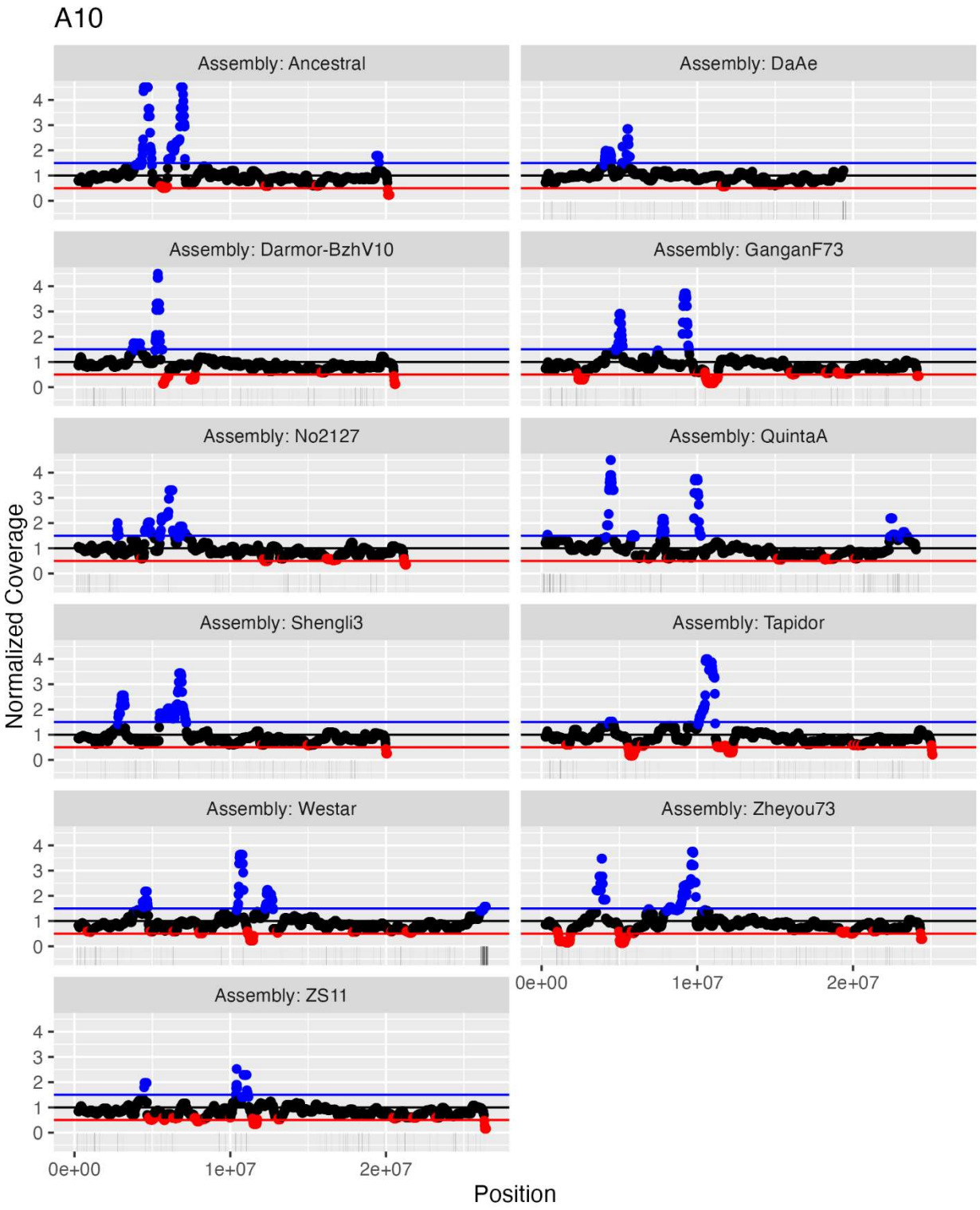

Figure S1.12. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

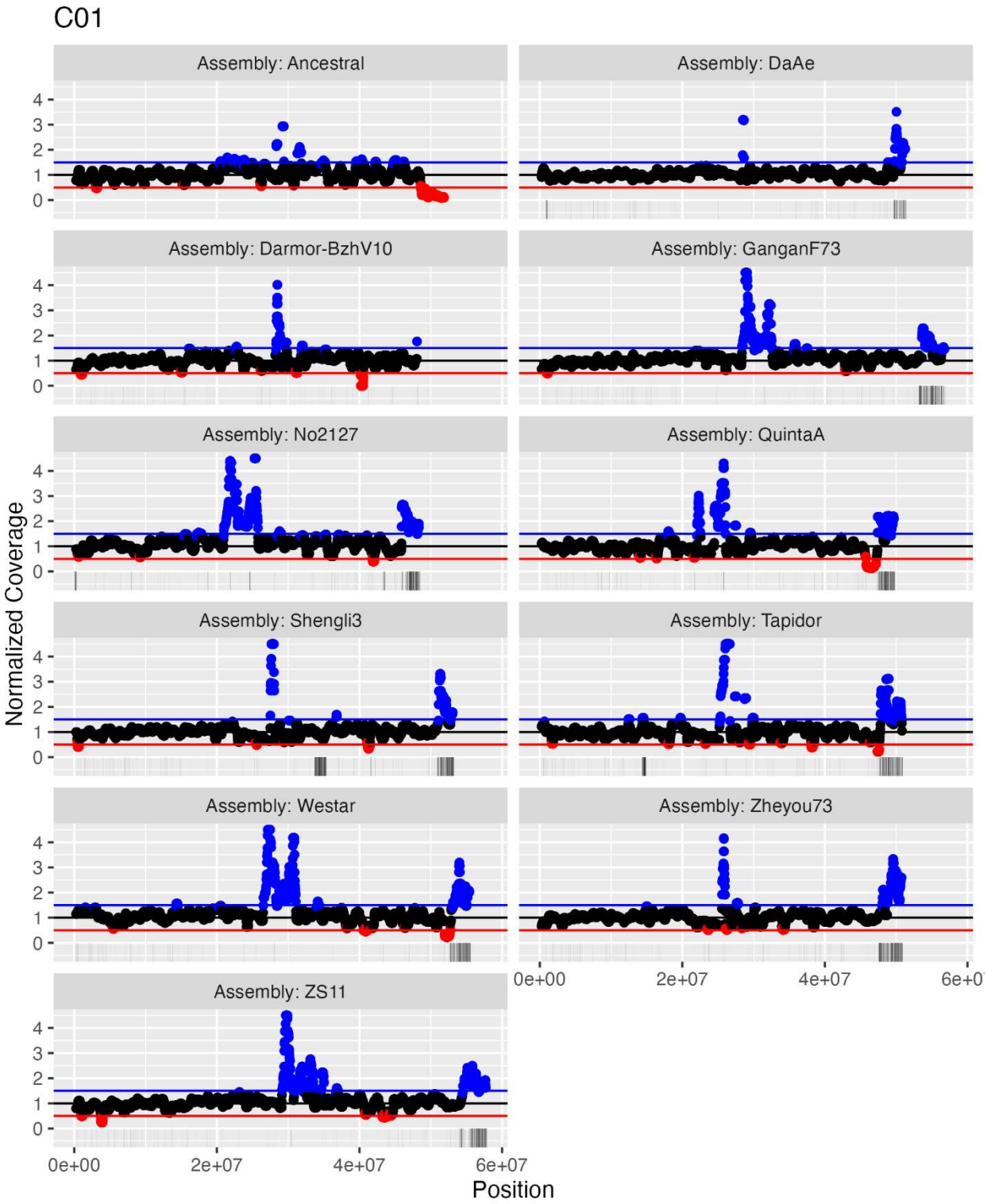

Figure S1.13. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

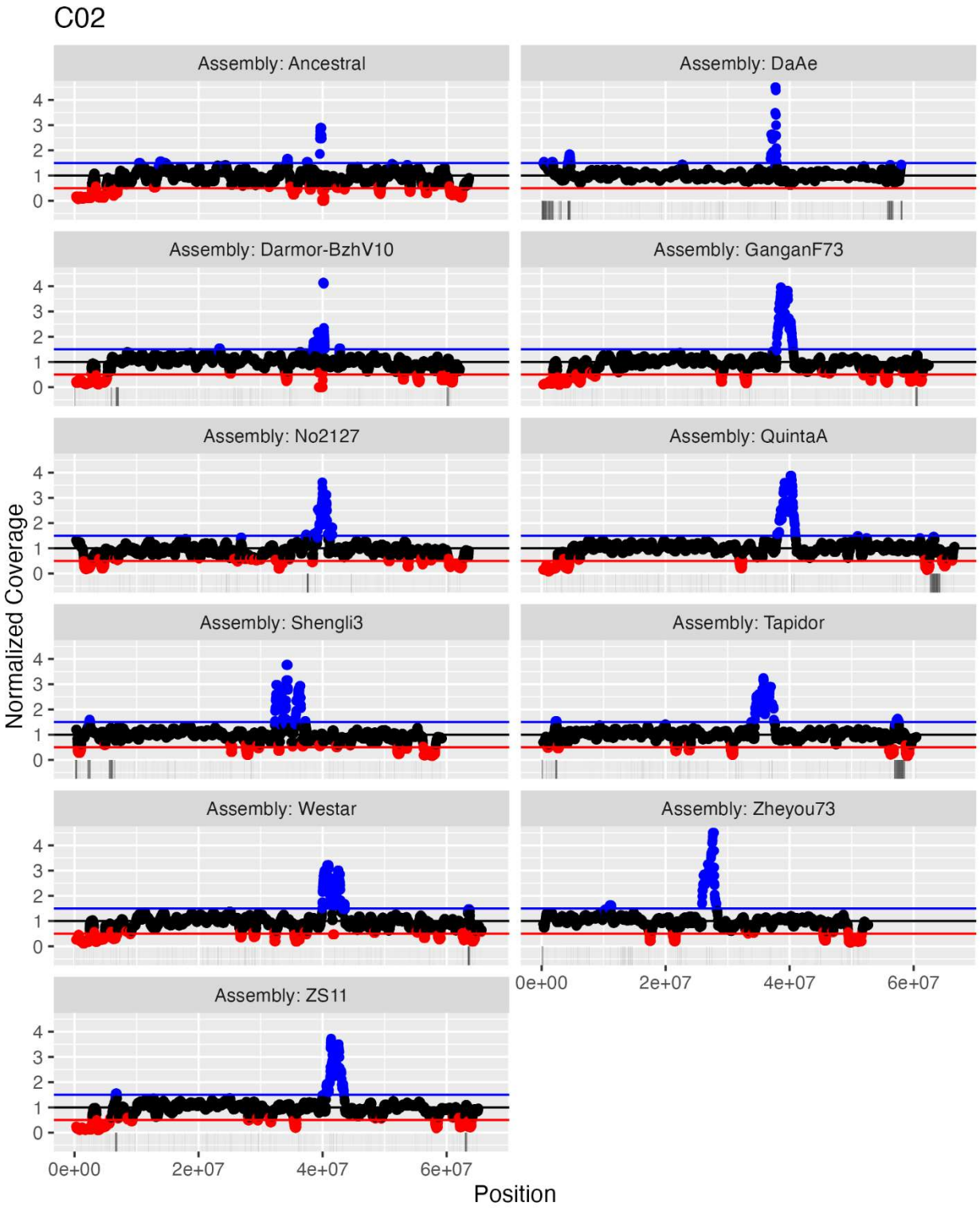

Figure S1.14. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

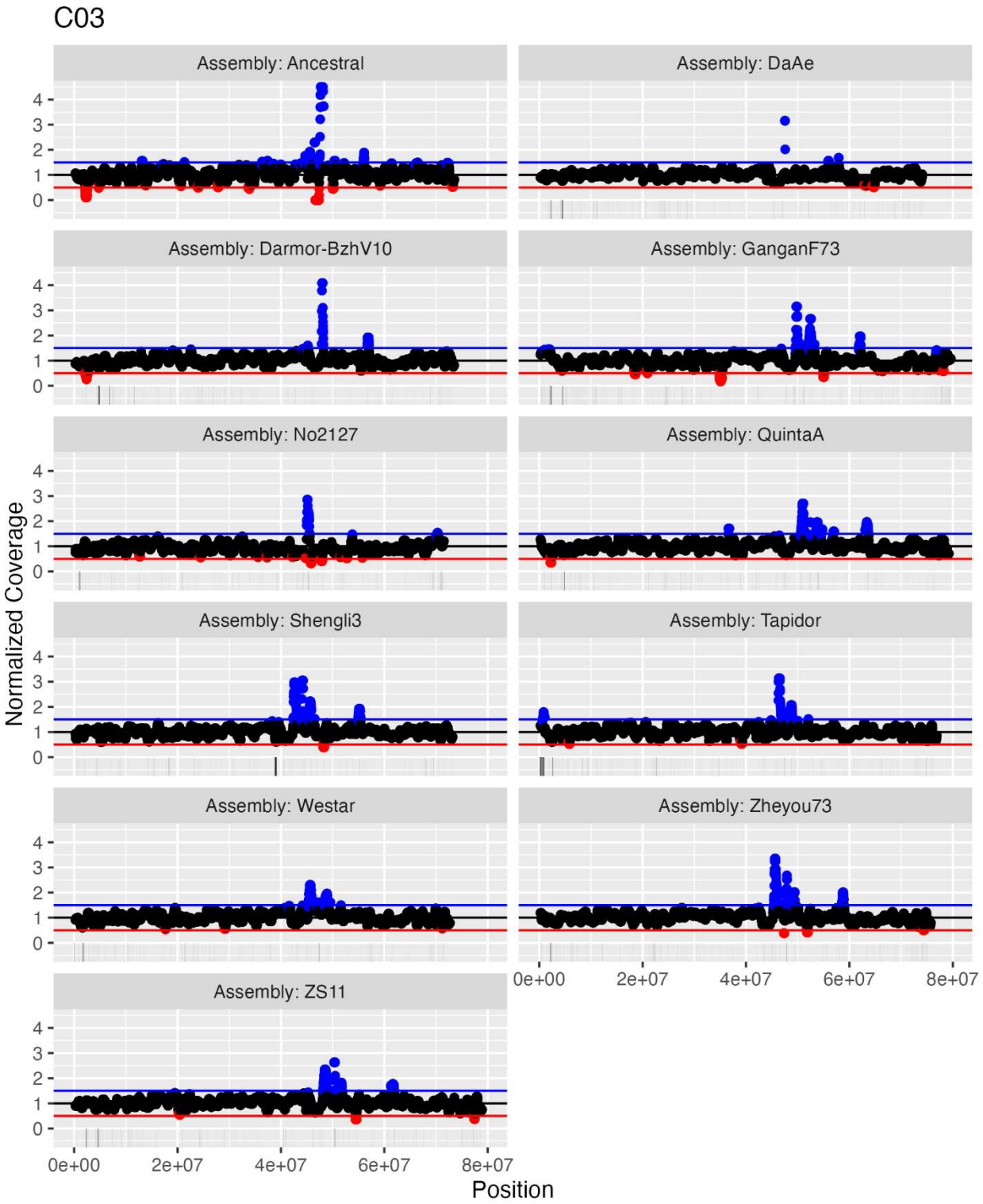

Figure S1.15. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

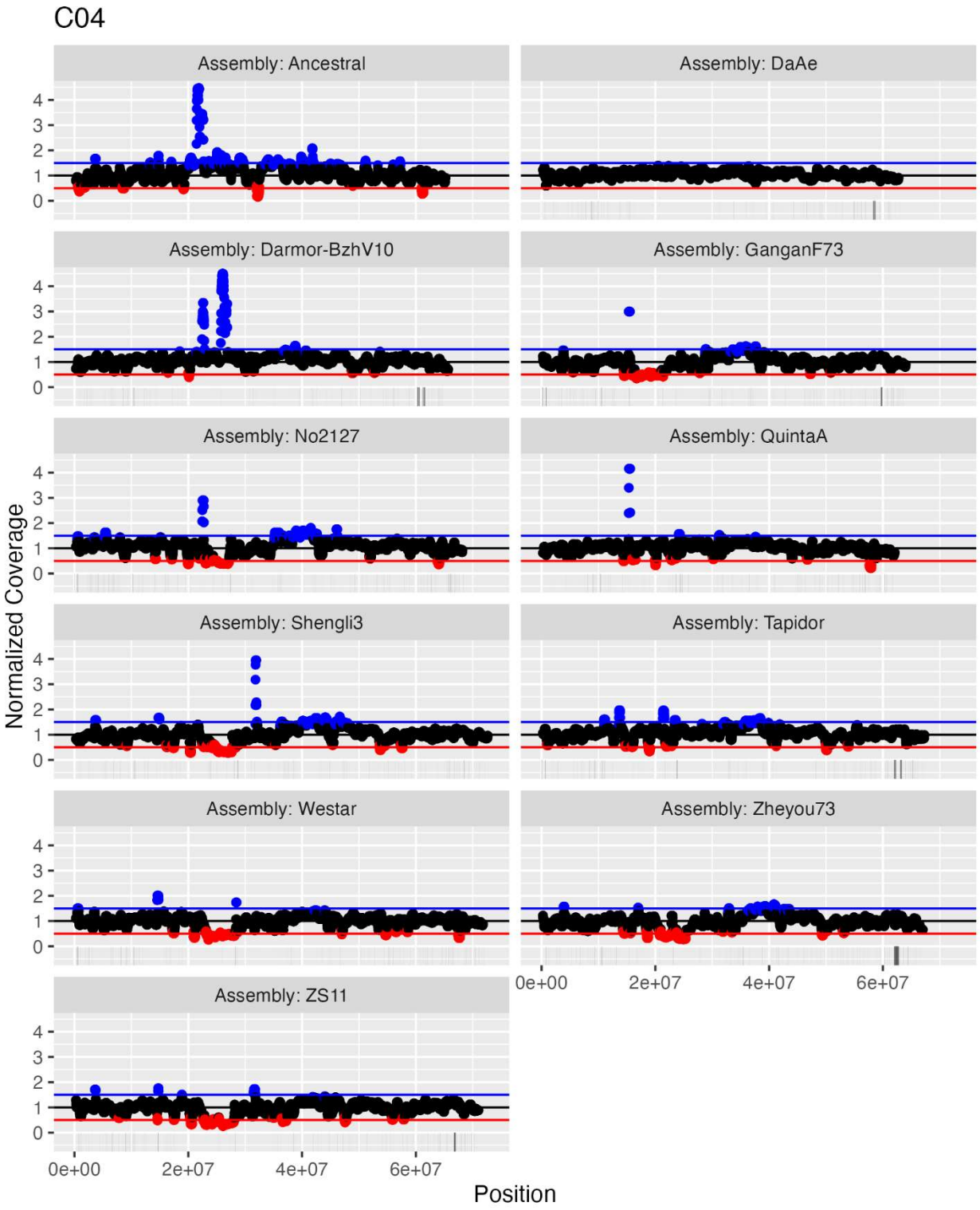

Figure S1.16. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

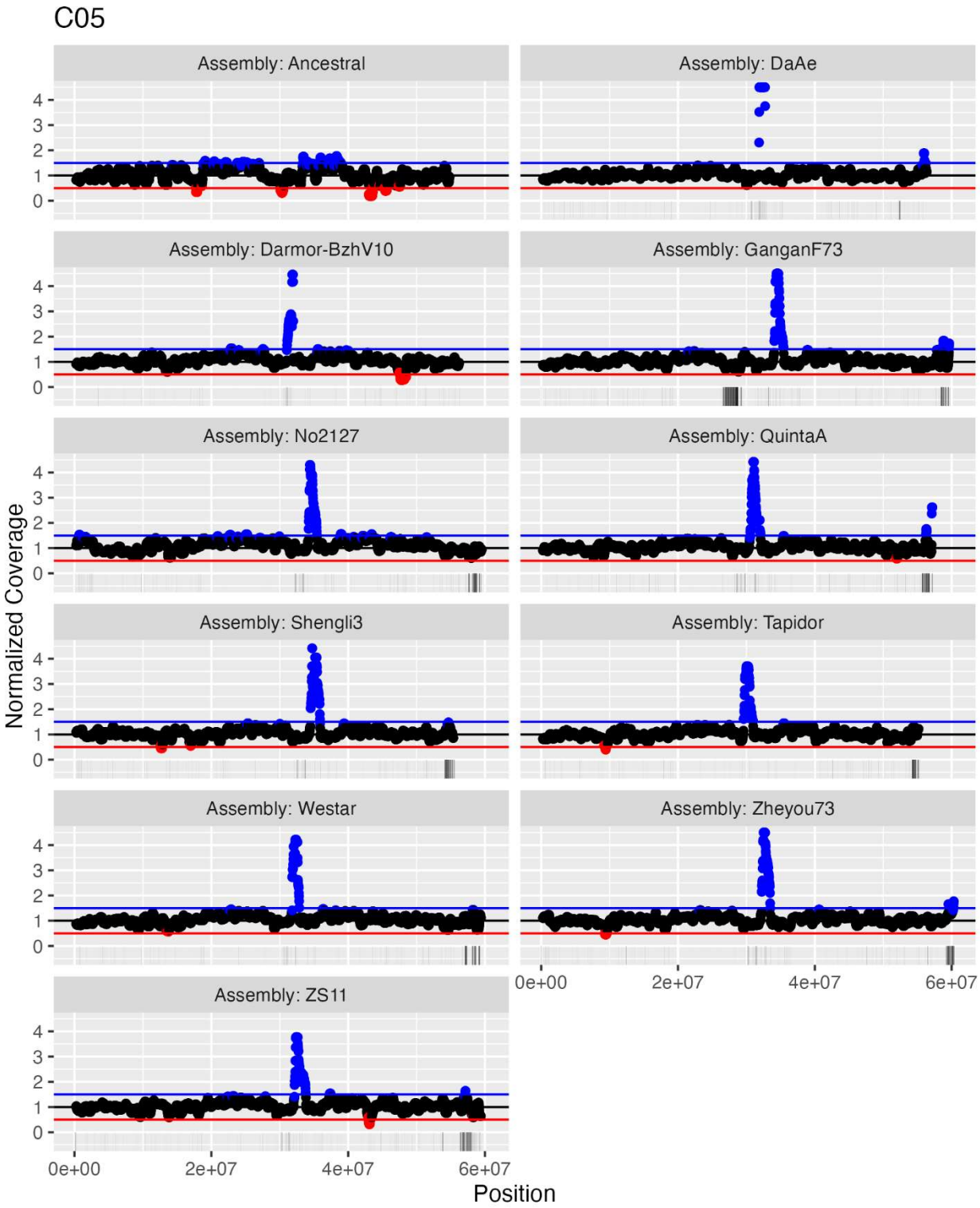

Figure S1.17. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

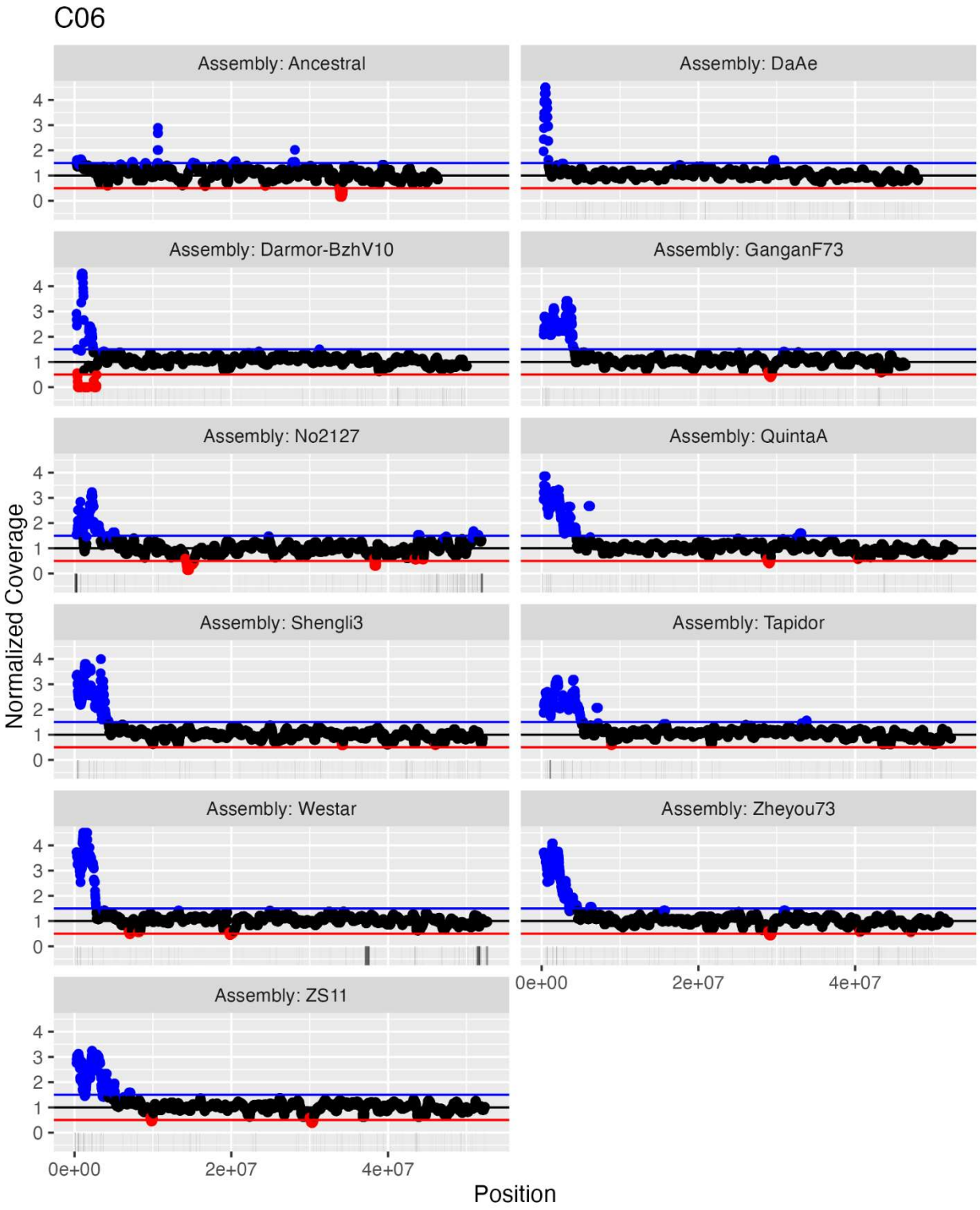

Figure S1.18. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

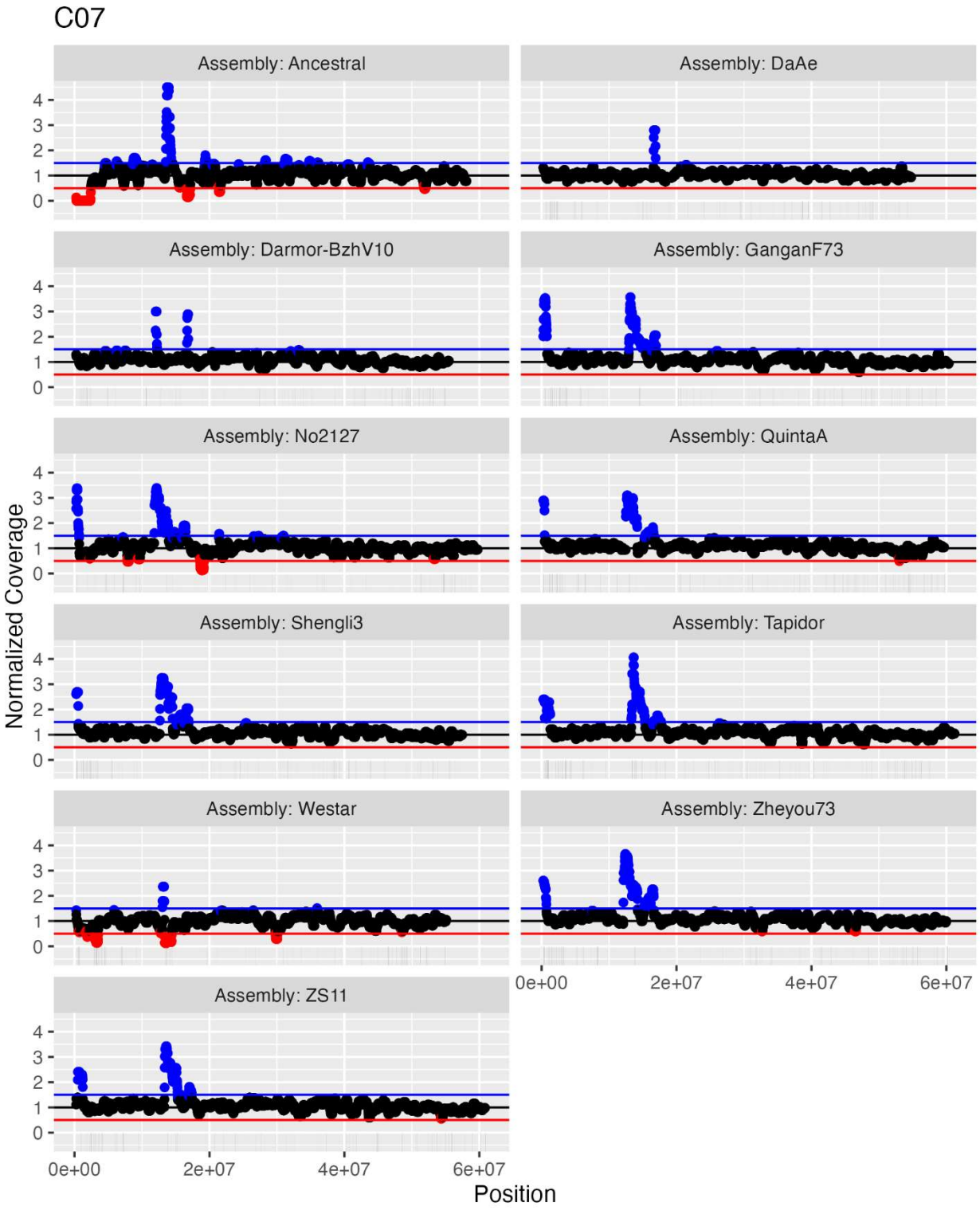

Figure S.19. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

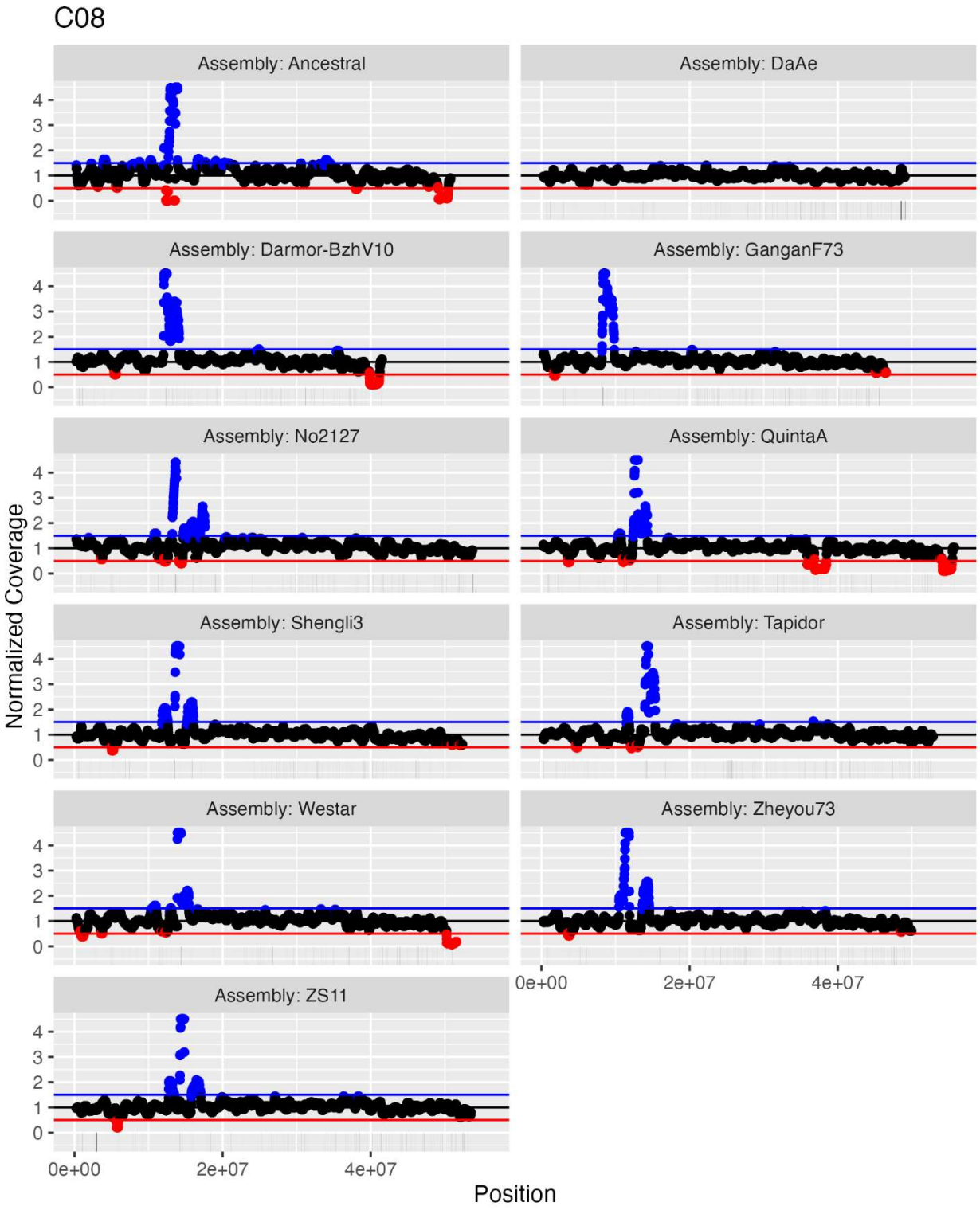

Figure S1.20. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

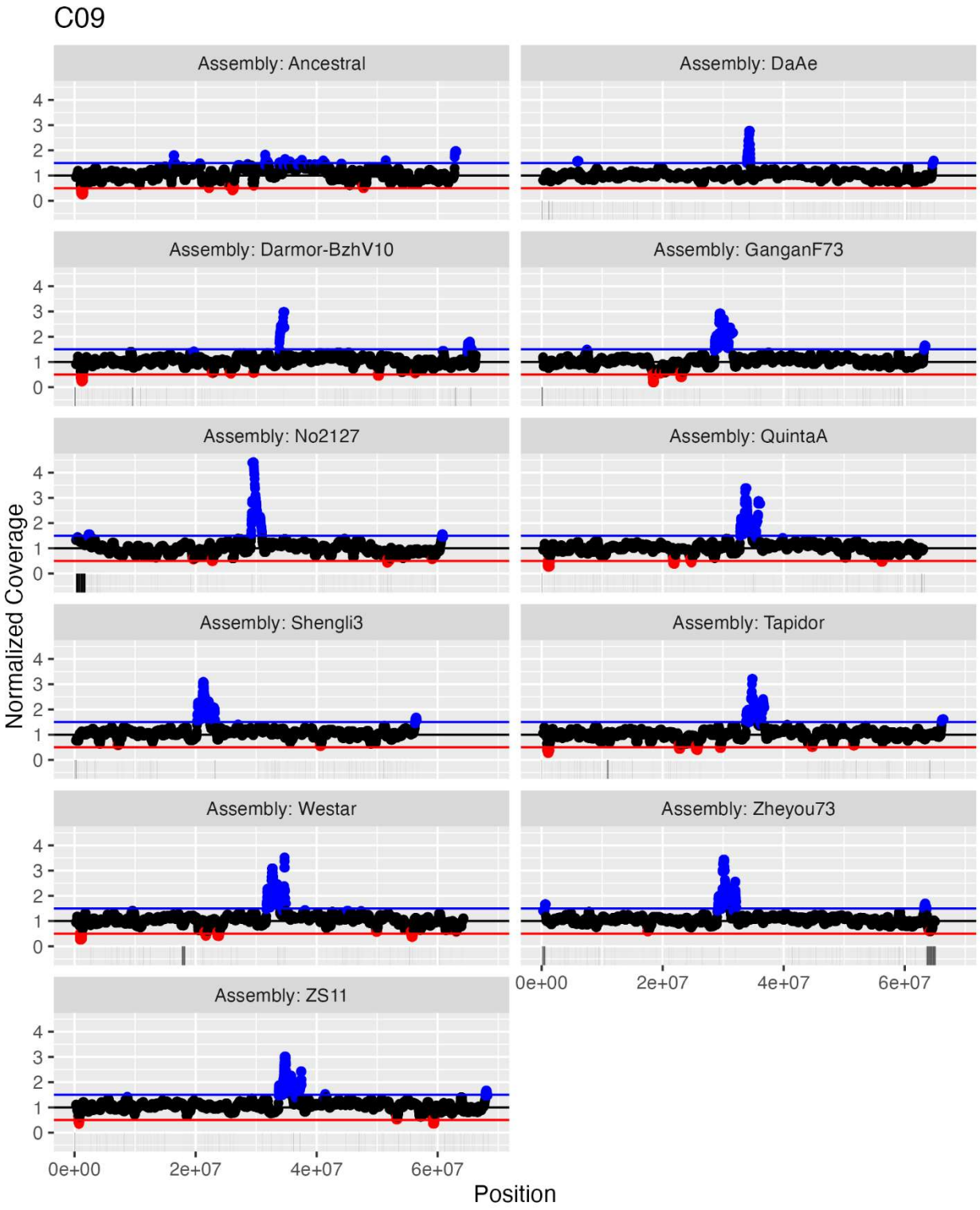

Figure S1.21. *Coverage of Da-Ae reads mapped to each genome, normalized to the genome-wide median. Areas with coverage less than 0.6x are colored red and those greater than 1.4x are colored blue. Horizontal red, black, and blue lines indicate 0.5x, 1.0x, and 1.5x coverage. Vertical lines below 0 indicate regions of potential homoeologous exchanges based on synteny analysis. Max coverages is capped at 4.5x for readability.*

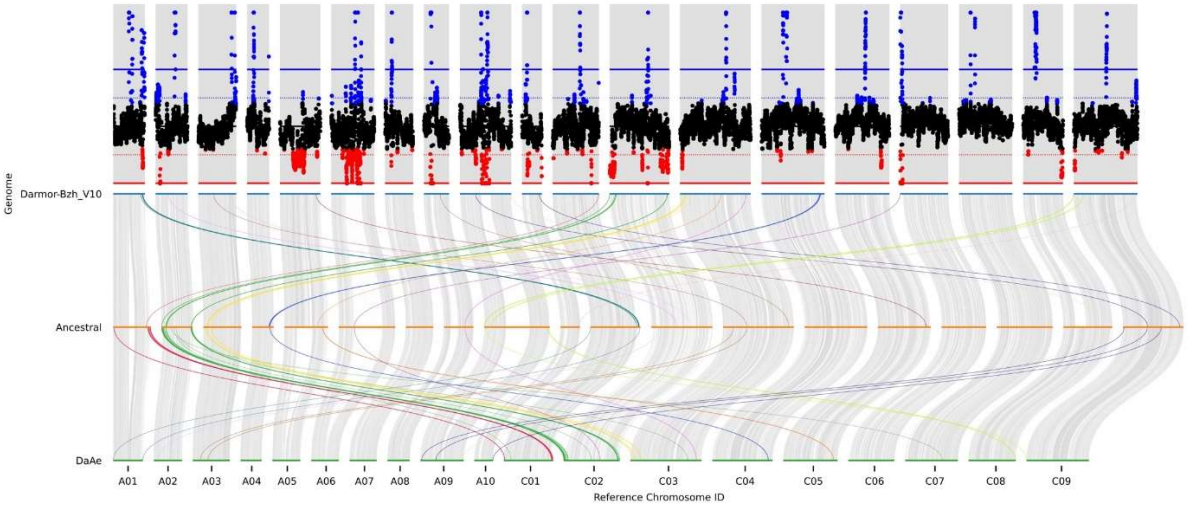

Figure S1.22. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Darmor Bzh V10; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

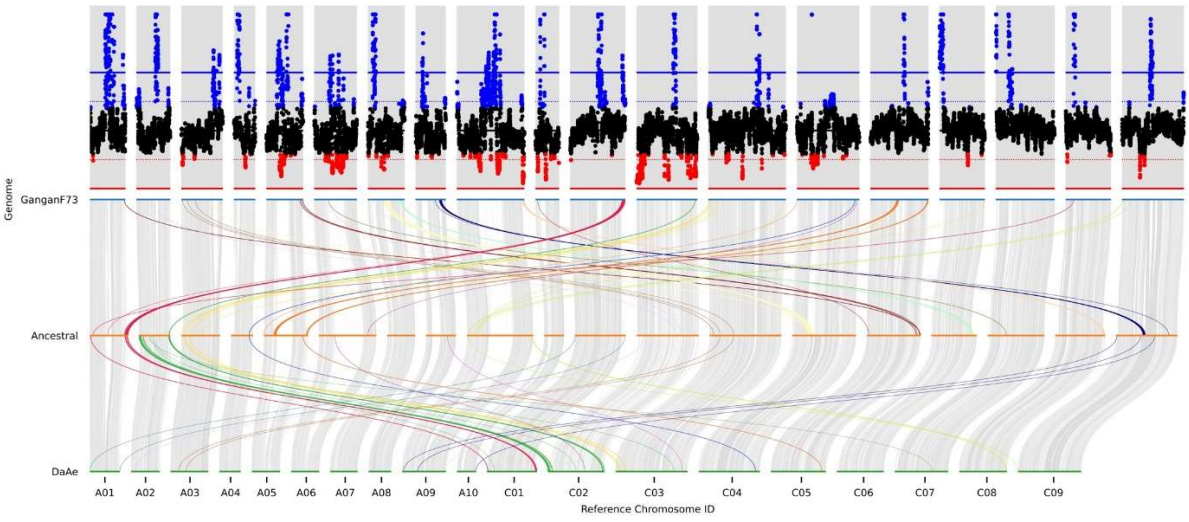

Figure S1.23. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to GanganF73 replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

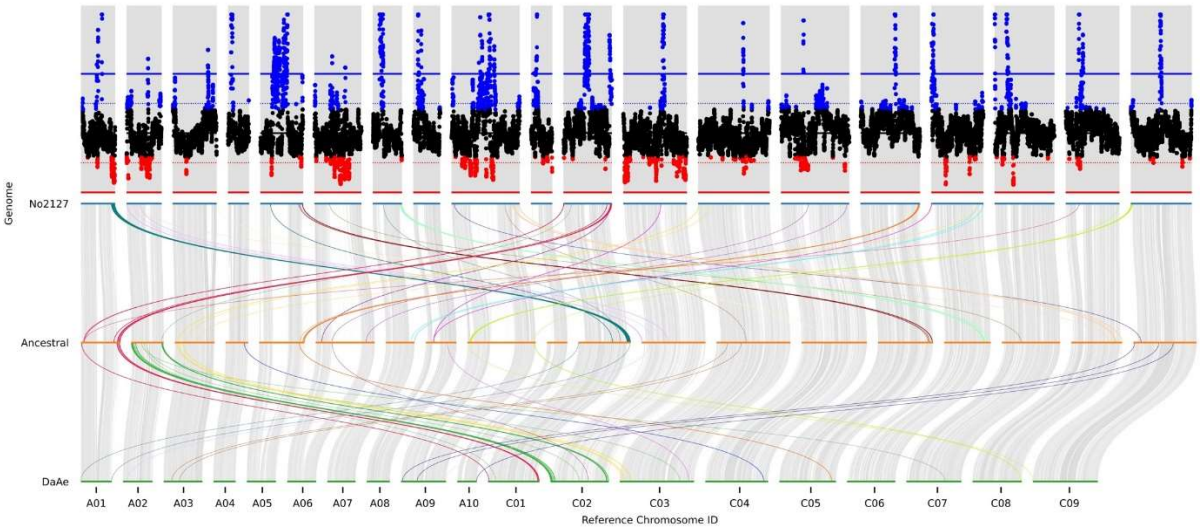

Figure S1.24. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to No2127; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

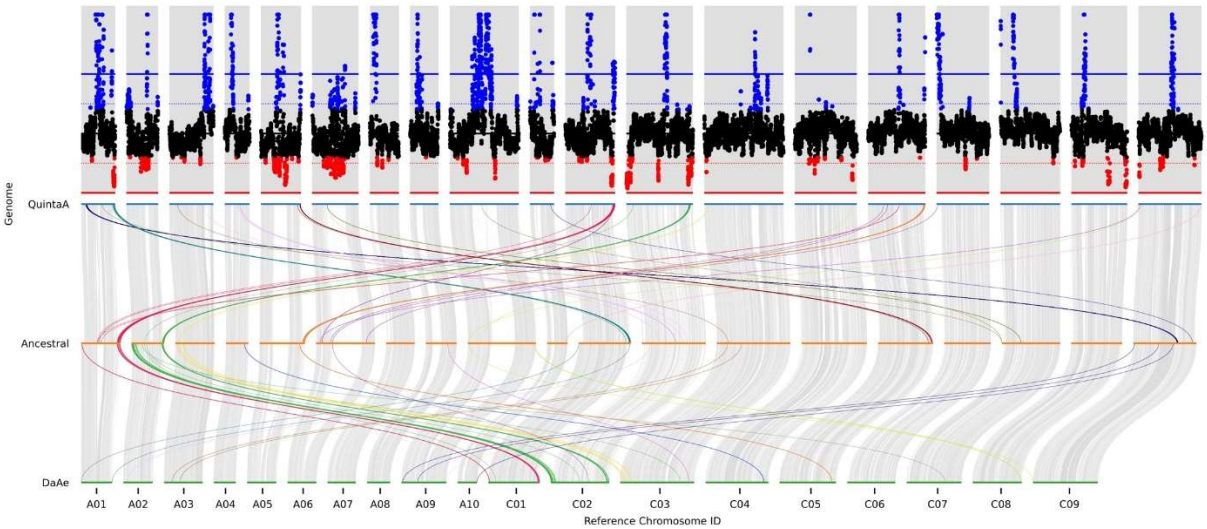

Figure S1.25. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to QunitaA; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

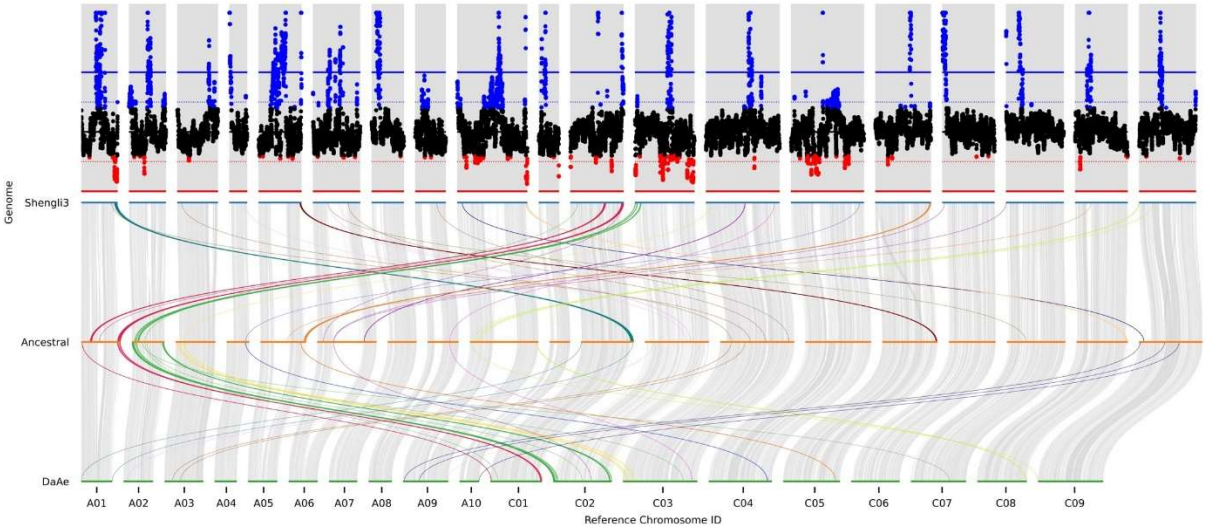

Figure S1.26. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Shengli3; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

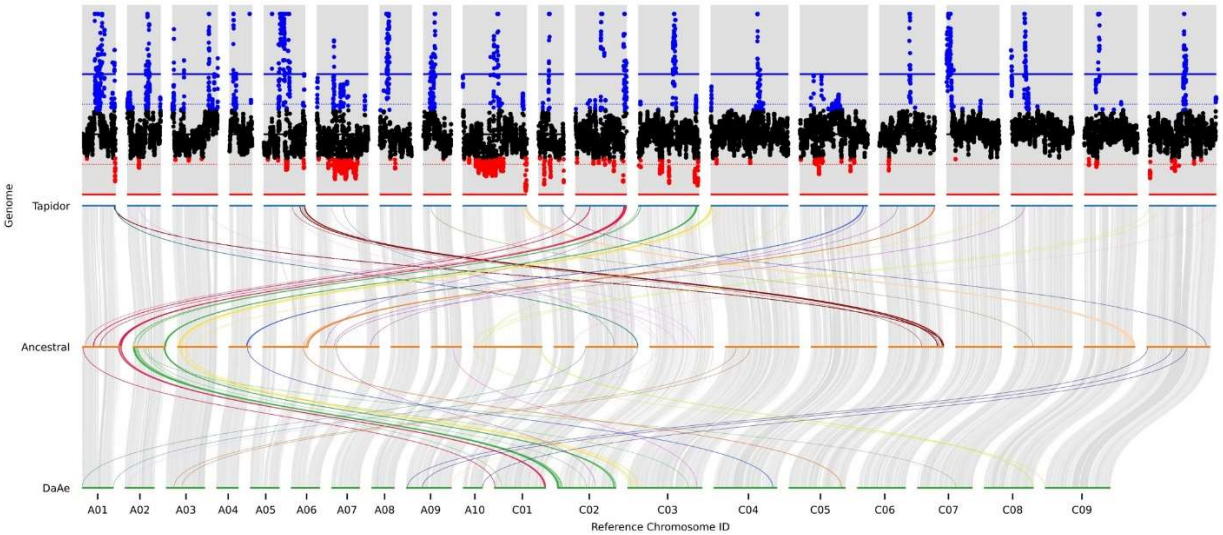

Figure S1.27. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Tapidor; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

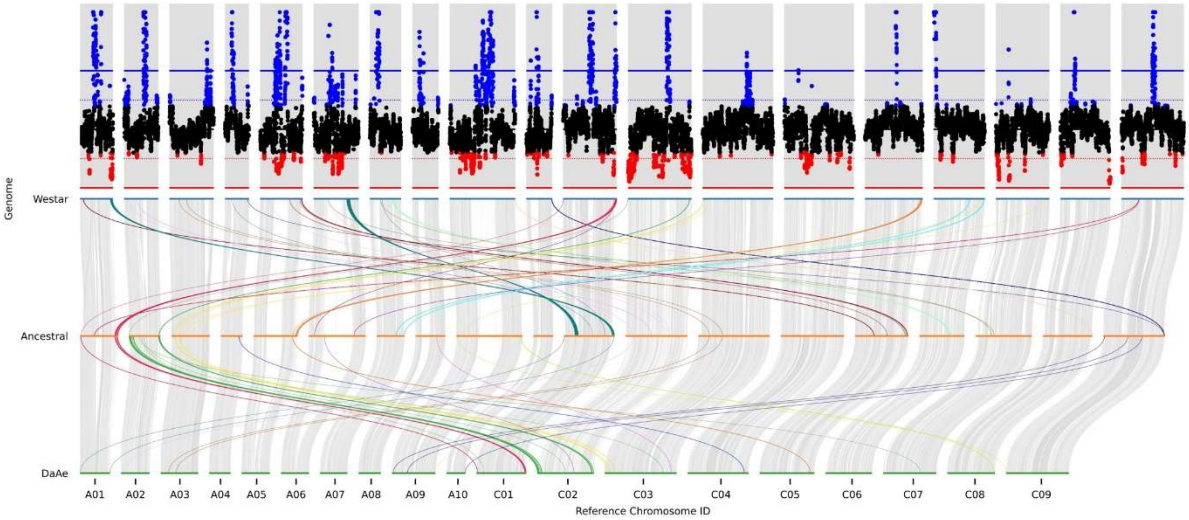

Figure S1.28. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Westar; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

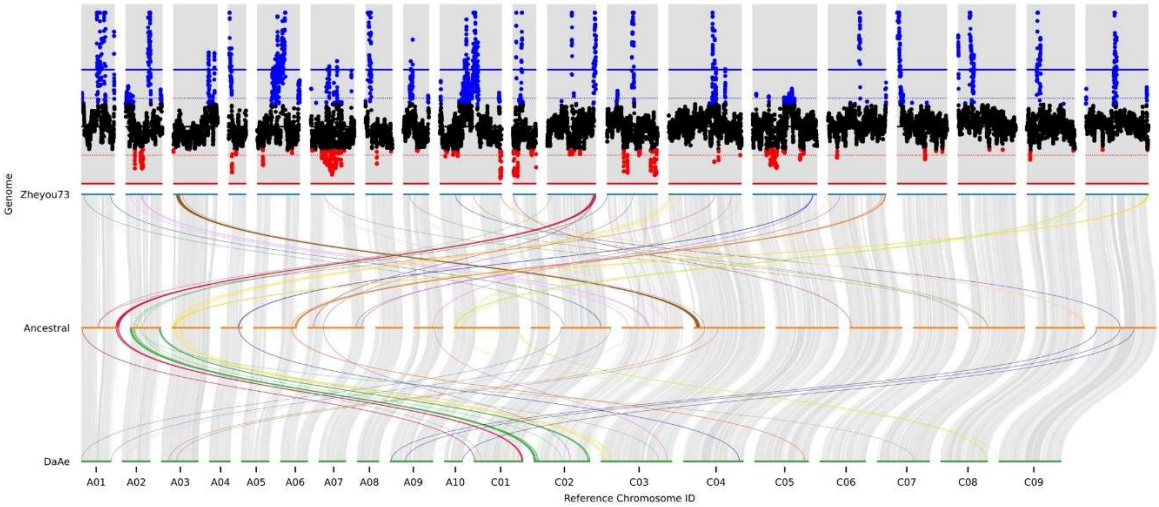

Figure S1.29. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Zheyou73; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

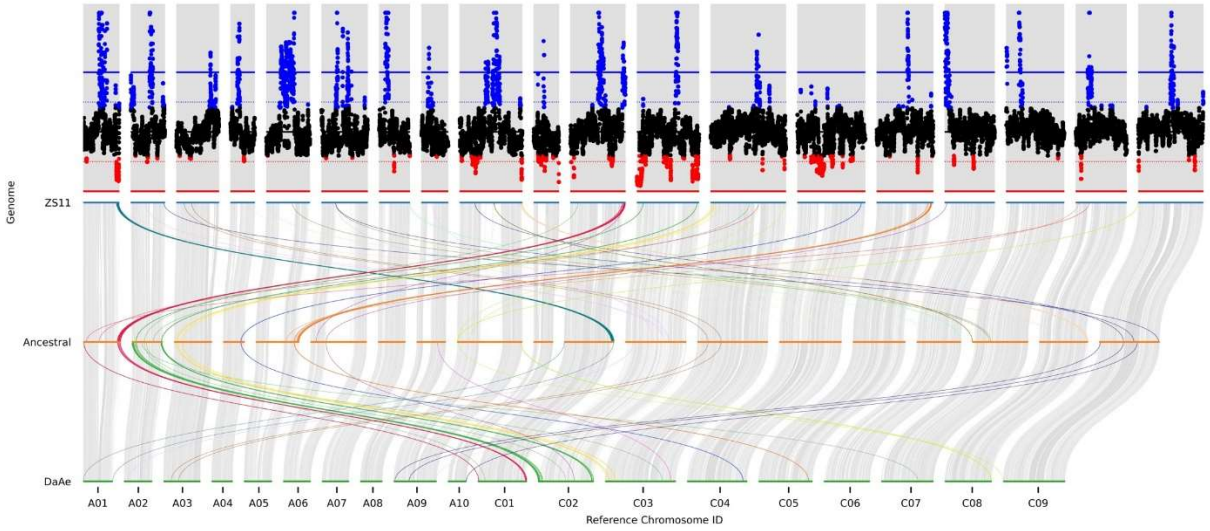

Figure S1.30. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to ZS11; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and the two B. napus varieties. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

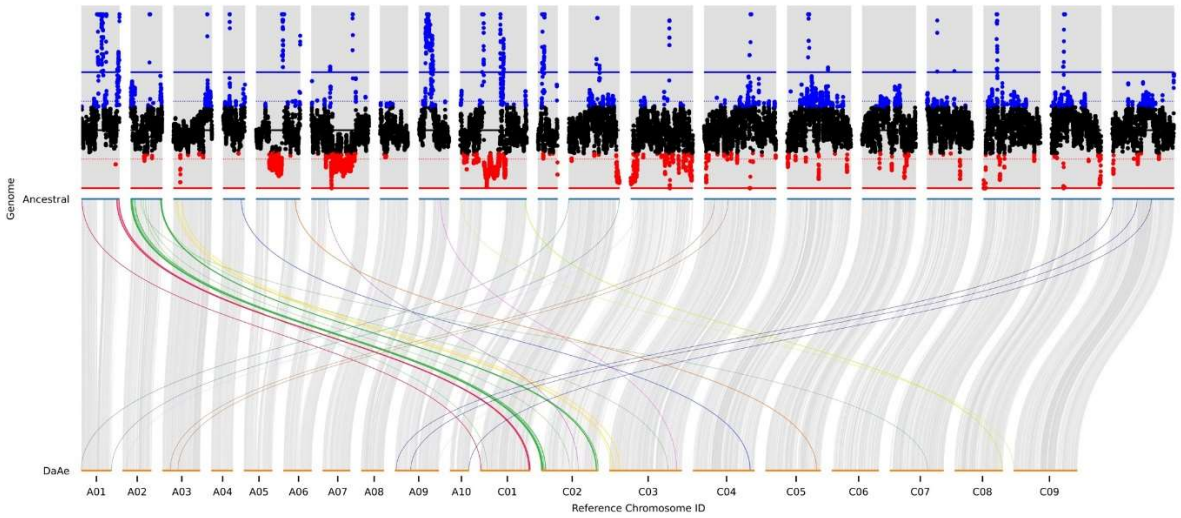

Figure S1.31. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to the "Ancestral" reference; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and Da-Ae. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

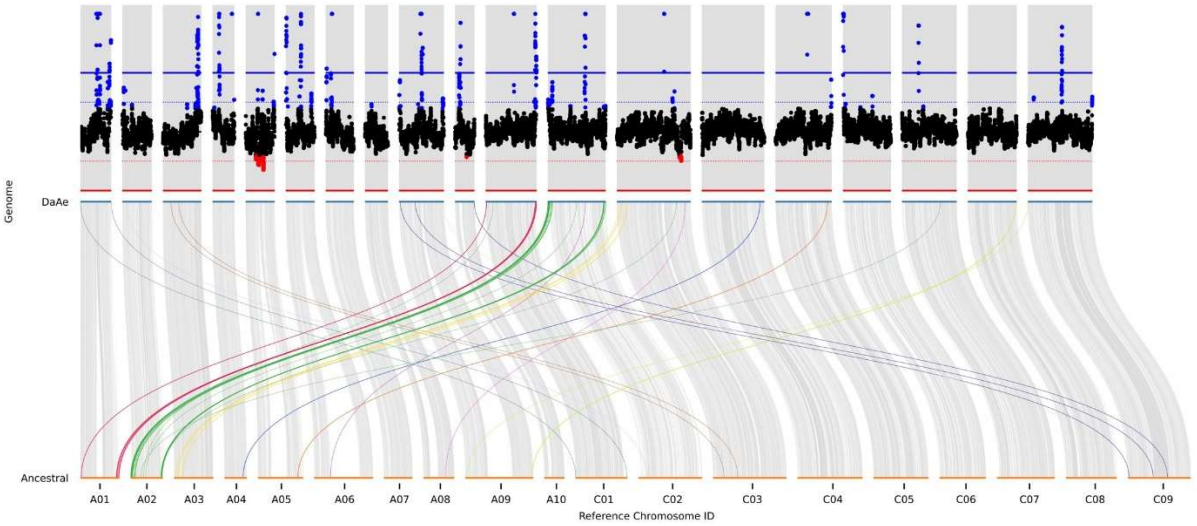

Figure S1.32. *Coverage and homoeologous exchange plot. Top panel: Coverage of Da-Ae reads mapped to Da-Ae; replotted from figures S3 – S21. Vertical lines indicate 0, 0.5x, 1x, 1.5x, and 2x coverage. Bottom panel: homoeologous exchange. Grey lines show homologous regions between the ancestral chromosomes and Da-Ae. Colored lines indicate homoeologous exchange; the color of the line corresponds to the ancestral chromosome.*

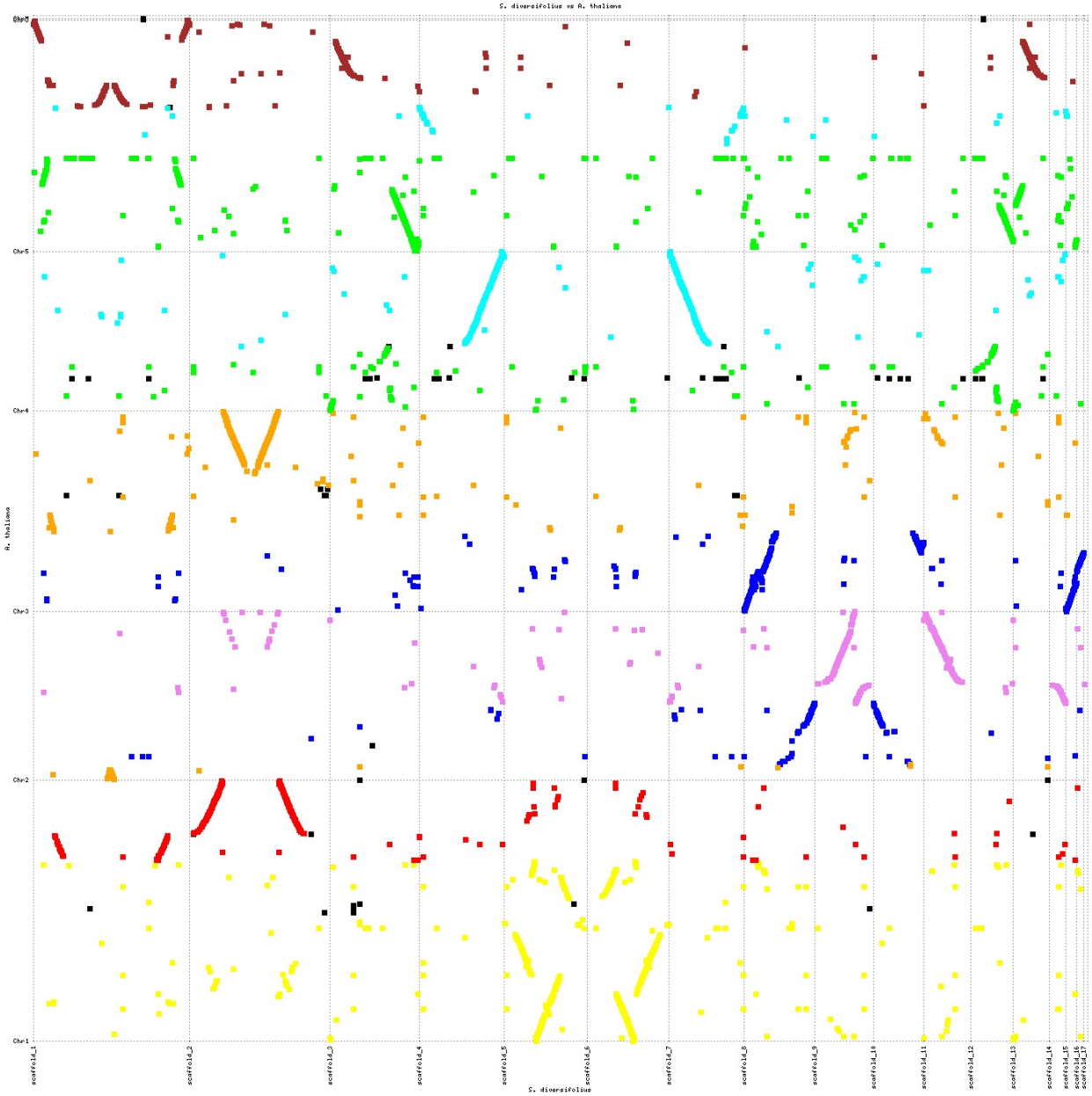

Figure S2.1. *All scaffolds greater than 1 Mb in the HiC assembly aligned to the TAIR10 A. thaliana genome assembly. Alignment was performed using Promer and plotted using Mummerplot, both programs included in the Mummer bioinformatic toolkit. The color of each line corresponds to the 8 ancestral crucifer karyotype (ACK) blocks of A. thaliana with the exception of black which corresponds to regions which do not belong to an ACK block. ACK block boundaries are based on gene locations summarized in Lysak et. Al 2016. Gray dashed lines indicate separate scaffold in each assembly.*

Table S1.1. *Discrepancies between Da-Ae and Darmor-bzh assemblies. Discrepancy number, chromosome discrepancy is located on, type of discrepancy, data able to support Da-Ae's composition, and action taken to resolve discrepancy.*

Discrepancy	Chromosome	Type	Supported	Action taken
1	A01	Inversion	Yes	None
2	A02	Inversion	No	Flipped
3	A02	Duplication	Yes	None
4	A03	Inversion	No	Flipped
5	A04	Inversion	Yes	None
6	A05	Inversion	No	Flipped
7	A06	Inversion	Yes	None
8	A07	Inversion	Yes	None
9	A08	Inversion	Yes	None
10	A09	Gap	Yes	None
11	A10	Inversion	Yes	None
12	C01	Inversion	No	Flipped
13	C02	Inversion	Yes	None
14	C03	Inversion	No	Flipped
15	C04	Inversion	Yes	None
16	C05	Inversion	No	Flipped
17	C05	Inversion	Yes	None
18	C06	Inversion and Gap	No	Flipped and Joined
19	C06	Inversion	No	Flipped
20	C06	Inversion	Yes	None
21	C07	Gap	No	Joined
22	C07	Inversion	Yes	None
23	C08	Duplication	Yes	None
24	C09	Inversion	Yes	None

Table S1.2. *Transposable element content of three B. napus assemblies calculated using default parameters of repeatmasker v4.1.2-p1 and lib file bnapus.TE-families.fa located at (http://cbi.hzau.edu.cn/rape/download_ext/).*

Comparison of the transposable elements among 3 B. napus genomes						
Classification	**DaAe**		**Darmor-Bzh_V10**		**ZS11**	
	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)
Class I: Retrotransposon	279891636	27.94725111	234603345	25.39558573	301651364	29.84025197
SINE	375932	0.037536906	67012	0.007253984	132092	0.013066934
LINE	29216324	2.917257389	27714254	3.000041255	28783458	2.847345452
LTR-Retrotransposon	250299380	24.99245681	206822079	22.38829049	272735814	26.97983958
Copia	82726692	8.260281256	65303382	7.069028092	82083435	8.119938032
Gypsy	104171553	10.40155609	92951519	10.06191225	103423820	10.23099252
Class II: DNA Transposon	41357988	4.12960563	38608863	4.179372167	41077363	4.063495175
hAT	2047996	0.204492922	1999669	0.21646224	2077757	0.205537915
Harbinger	4906031	0.489868444	4484716	0.48546618	4830866	0.477883663
Unclassified	218661515	21.83340794	204308974	22.1162493	219971502	21.76023658
Total Content	539911139	53.91026468	477521182	51.6912072	562700229	55.66398373

Table S1.3. *Assembly statistics of intermediate assemblies along with BUSCOs percentages. BUSCOs percentages were calculated using the brassicales_odb10 dataset, which contains 4,596 BUSCOS*

Assembly	N50 (Mbp)	Sequences	Total Length (Mbp)	Total Unambiguous Length (Mbp)	Complete BUSCOs	Complete single-copy BUSCOs	Complete duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	# BUSCOs
DaAe Canu	1.59	4,008	1,004	1,004	98.6	21.0	77.6	0.1	1.3	4596
DaAe Pilon Canu	1.59	4,008	1,004	1,004	98.6	18.7	79.9	0.1	1.3	4596
DaAe Dovetail Pilon Canu	42.79	3,190	1,004	1,004	98.6	19.9	78.7	0.1	1.3	4596
DaAe Final	48.21	3,164	1,002	1,001	98.5	18.0	80.5	0.2	1.3	4596

Table S1.4. *Conserved homoeologous exchange genes. These genes were found in homoeologous exchange regions in all B. napus genome analyzed. The table also shows the best Arabidopsis thaliana hit (based on blastp) along with the annotation associated with that Arabidopsis gene.*

Brassica gene ID	Best Arabidopsis Match	E-value	Description
A01p50810.2_BraZ1	AT3G01180	3.52E-142	PTHR12526:SF337 - STARCH SYNTHASE 2, CHLOROPLASTIC/AMYLOPLASTIC
A02p19770.2_BraZ1	AT5G67150	0	PTHR31896//PTHR31896:SF12 - FAMILY NOT NAMED // ANTHRANILATE N-HYDROXYCINNAMOYL/BENZOYLTRANSFERASE-LIKE PROTEIN-RELATED
A02p35520.2_BraZ1	AT1G22930	3.78E-66	PTHR12832:SF11 - PROTEIN M05D6.2
A02p35570.2_BraZ1	AT5G44300	1.41E-71	PF05564 - Dormancy/auxin associated protein
A03p20960.2_BraZ1	AT3G51770	0	PTHR23083:SF420 - ETO1-LIKE PROTEIN 2-RELATED
A05p37510.2_BraZ1	AT3G18210	0	1.14.11.4 - Procollagen-lysine 5-dioxygenase / Procollagen-lysine,2-oxoglutarate 5-dioxygenase
A06p55480.2_BraZ1	AT5G44010	3.27E-32	PTHR14449:SF2 - FANCONI ANEMIA GROUP F PROTEIN
A08p02280.2_BraZ1	AT5G50220	6.63E-66	PTHR31790//PTHR31790:SF10 - FAMILY NOT NAMED // F-BOX ASSOCIATED UBIQUITINATION EFFECTOR FAMILY PROTEIN-RELATED
A08p37180.2_BraZ1	AT1G05080	2.66E-108	PTHR32212//PTHR32212:SF112 - FAMILY NOT NAMED // FBD / LEUCINE RICH REPEAT DOMAINS CONTAINING PROTEIN-RELATED
BolC3t19314H	AT3G46340	0	PF07714//PF12819//PF13855 - Protein tyrosine kinase // Carbohydrate-binding protein of the ER // Leucine rich repeat
BolC4t26698H	AT5G44220	2.14E-65	PTHR31790//PTHR31790:SF10 - FAMILY NOT NAMED // F-BOX ASSOCIATED UBIQUITINATION EFFECTOR FAMILY PROTEIN-RELATED
BolC5t30633H	AT1G20720	0	K15362 - fanconi anemia group J protein
BolC5t33508H	AT3G18150	1.76E-157	PTHR32153//PTHR32153:SF2 - FAMILY NOT NAMED // F-BOX/LRR-REPEAT PROTEIN 25-RELATED
BolC8t52214H	AT1G12210	0	PTHR23155//PTHR23155:SF639 - LEUCINE-RICH REPEAT-CONTAINING PROTEIN // DISEASE RESISTANCE PROTEIN RFL1-RELATED

Table S2.1. *Repeat element content of HiFi assembly. Analysis completed using RepeatMasker version 4.1.5 in sensitive mode. Query species was assumed to be Brassicales.*

			Number of elements	Length occupied (bp)	percentage of sequence (%)
Retroelements			74,066	75,569,968	18.79
	SINEs:		270	39,770	0.01
	Penelope:		0	0	0
	LINEs:		4,782	3,064,392	0.76
		CRE/SLACS	0	0	0
		L2/CR1/Rex	0	0	0
		R1/LOA/Jockey	0	0	0
		R2/R4/NeSL	0	0	0
		RTE/Bov-B	0	0	0
		L1/CIN4	4,758	3,061,849	0.76
	LTR elements		69,014	72,465,806	18.02
		BEL/Pao	0	0	0
		Ty1/Copia	32,174	42,611,770	10.6
		Gypsy/DIRS1	33,841	28,849,738	7.17
DNA transposons			20,796	6,705,897	1.67
	hobo-Activator		4,574	1,292,104	0.32
	Tc1-IS630-Pogo		1,476	311,487	0.08
	En-Spm		0	0	0
	MULE-MuDR		6,750	2,499,600	0.62
	PiggyBac		0	0	0
	Tourist/Harbinger		1,260	438,400	0.11
	Other (Mirage, P-element, Transib)		0	0	0
Rolling-circles			994	159,073	0.04
Unclassified:			23	4,153	0
Total interspersed repeats:				82,280,018	20.46
Small RNA			7,919	12,239,420	3.04
Satellites:			58	10,415	0
Simple repeats:			120,687	16,567,216	4.12
Low complexity:			26,984	1,340,333	0.33

Table S2.2. *Whole genome duplicated genes identified as being positively selected.*

Gene ID	dn/ds(foreground w)
Sdiv_ptg000013l_0766	32.17004
Sdiv_ptg000010l_1595	22.20926
Sdiv_ptg000008l_0478	999.0
Sdiv_ptg000001l_1124	31.15026
Sdiv_ptg000014l_1436	105.08251
Sdiv_ptg000001l_1062	31.04736
Sdiv_ptg000007l_1453	73.39811
Sdiv_ptg000010l_0169	350.50941
Sdiv_ptg000014l_2063	221.62612
Sdiv_ptg000014l_1045	12.20527
Sdiv_ptg000015l_1073	999.0
Sdiv_ptg000014l_0024	998.99998
Sdiv_ptg000015l_0865	370.64822
Sdiv_ptg000014l_1947	39.11128
Sdiv_ptg000008l_0928	10.56043
Sdiv_ptg000001l_1298	15.28252
Sdiv_ptg000002l_1596	999.0
Sdiv_ptg000027l_0100	24.50149
Sdiv_ptg000003l_0332	6.19285
Sdiv_ptg000003l_2400	80.22779
Sdiv_ptg000001l_0314	44.22287
Sdiv_ptg000001l_1639	999.0
Sdiv_ptg000014l_2827	183.55386
Sdiv_ptg000015l_0518	97.6169
Sdiv_ptg000008l_0378	23.16575
Sdiv_ptg000011l_0389	92.43532
Sdiv_ptg000014l_2331	17.89055
Sdiv_ptg000022l_0861	10.53962
Sdiv_ptg000004l_2357	30.55106
Sdiv_ptg000011l_0160	16.03093
Sdiv_ptg000007l_2090	16.07821
Sdiv_ptg000015l_0943	999.0
Sdiv_ptg000009l_0137	998.99942
Sdiv_ptg000010l_0005	281.63416
Sdiv_ptg000010l_1603	998.99995
Sdiv_ptg000010l_1902	31.40687
Sdiv_ptg000015l_1275	774.21537
Sdiv_ptg000012l_1036	31.44084
Sdiv_ptg000015l_0898	39.16785
Sdiv_ptg000003l_0834	10.66564
Sdiv_ptg000005l_0609	322.34551
Sdiv_ptg000001l_0806	15.28986
Sdiv_ptg000008l_0432	28.34799
Sdiv_ptg000003l_1516	208.0823
Sdiv_ptg000010l_0344	62.03208
Sdiv_ptg000003l_2703	998.9999
Sdiv_ptg000004l_0159	998.99842
Sdiv_ptg000011l_0100	172.13979
Sdiv_ptg000012l_1611	17.35101
Sdiv_ptg000001l_0552	41.28307
Sdiv_ptg000004l_1328	27.92853
Sdiv_ptg000001l_1046	32.31583
Sdiv_ptg000001l_2258	571.97905

Sdiv_ptg000009l_2521	999.0
Sdiv_ptg000012l_0627	9.23852
Sdiv_ptg000004l_1636	37.79185
Sdiv_ptg000015l_0977	999.0
Sdiv_ptg000014l_1645	35.00693
Sdiv_ptg000002l_0438	127.49068
Sdiv_ptg000008l_0987	93.05138
Sdiv_ptg000004l_1630	31.30369
Sdiv_ptg000008l_1040	10.71572
Sdiv_ptg000014l_0520	999.0
Sdiv_ptg000009l_2439	36.45915
Sdiv_ptg000012l_1728	998.99978
Sdiv_ptg000009l_2235	139.67016
Sdiv_ptg000013l_0327	470.97597
Sdiv_ptg000013l_0829	998.99998
Sdiv_ptg000005l_1645	998.99956
Sdiv_ptg000009l_1893	18.07336
Sdiv_ptg000002l_1419	20.0717
Sdiv_ptg000009l_2762	47.04125
Sdiv_ptg000012l_1500	999.0
Sdiv_ptg000015l_0816	41.87034
Sdiv_ptg000004l_1146	38.38675
Sdiv_ptg000009l_1565	51.71906
Sdiv_ptg000015l_0238	101.00009
Sdiv_ptg000001l_1889	85.17439
Sdiv_ptg000013l_0961	11.96356
Sdiv_ptg000009l_2455	998.99997
Sdiv_ptg000014l_2134	17.69411
Sdiv_ptg000003l_1230	999.0
Sdiv_ptg000003l_1063	998.99969
Sdiv_ptg000012l_0819	630.4726
Sdiv_ptg000001l_1516	14.07566
Sdiv_ptg000009l_1341	57.18135
Sdiv_ptg000015l_2229	142.66961
Sdiv_ptg000007l_1591	7.46057
Sdiv_ptg000002l_0944	998.99988
Sdiv_ptg000008l_0089	999.0
Sdiv_ptg000022l_0945	27.93827
Sdiv_ptg000013l_0270	73.73746
Sdiv_ptg000008l_1165	999.0
Sdiv_ptg000001l_1237	167.64986
Sdiv_ptg000004l_0310	80.72848
Sdiv_ptg000012l_1139	52.82233
Sdiv_ptg000007l_1788	21.26733
Sdiv_ptg000015l_1609	12.37023
Sdiv_ptg000005l_0365	3.23131
Sdiv_ptg000010l_2755	25.97968
Sdiv_ptg000009l_2710	16.21791
Sdiv_ptg000004l_2084	39.69234
Sdiv_ptg000002l_0127	9.95842
Sdiv_ptg000008l_0743	998.99984
Sdiv_ptg000001l_0525	32.48021
Sdiv_ptg000005l_1120	94.29046
Sdiv_ptg000009l_0110	38.90653
Sdiv_ptg000013l_1112	999.0
Sdiv_ptg000002l_0362	32.27228

Sdiv_ptg000009l_1264	15.2677
Sdiv_ptg000014l_1560	999.0
Sdiv_ptg000002l_1891	411.51758
Sdiv_ptg000005l_1682	38.18411
Sdiv_ptg000003l_2775	27.56863
Sdiv_ptg000004l_0557	99.36945
Sdiv_ptg000010l_2889	48.33756
Sdiv_ptg000001l_1087	11.45972
Sdiv_ptg000008l_0959	74.62397
Sdiv_ptg000010l_2481	170.11009
Sdiv_ptg000002l_1480	999.0
Sdiv_ptg000005l_1987	15.82047
Sdiv_ptg000007l_2082	42.42607
Sdiv_ptg000005l_0465	202.5924

Table S3.3. *Dry gene modules significantly correlated with traits of interest*

Response	Predictor	Estimate	p.value	R2	Adj.R2
Breadth	MEorangered4	21.39278	0.014207	0.549059	0.492692
Breadth	MEdarkgreen	17.94169	0.03945	0.430245	0.359025
CWD	MEmediumpurple3	-118.021	0.011847	0.567772	0.513743
CWD	MEivory	127.5367	0.022372	0.499133	0.436525
CWD	MEsaddlebrown	135.5732	0.026831	0.477856	0.412589
CWD	MEgreen	-109.432	0.048601	0.403068	0.328451
CWD	MEskyblue3	127.8113	0.048782	0.402574	0.327896
Critical temperature max	MEgrey60	41.87337	0.037703	0.436013	0.365514
Critical temperature min	MEblue	-78.3922	0.003861	0.668472	0.627031
Critical temperature min	MEthistle2	-473.54	0.009701	0.58755	0.535994
Cumulative proportion germination	MEyellowgreen	-1.98637	0.008385	0.601458	0.551641
Cumulative proportion germination	MEorangered4	1.796466	0.018537	0.520362	0.460408
Cumulative proportion germination	MEbisque4	-1.74795	0.023738	0.492276	0.428811
Elevation	MEpalevioletred3	1697.235	0.016855	0.530804	0.472155
Lat	MEivory	-9.06293	0.011603	0.569868	0.516102
Lat	MEsaddlebrown	-9.59628	0.015287	0.54131	0.483974
Long	MEivory	7.750518	0.009383	0.590772	0.539618
Long	MEsaddlebrown	8.389562	0.009811	0.586462	0.53477
Long	MEfloralwhite	-8.22103	0.016545	0.532818	0.474421
Long	MEmediumpurple3	-6.05503	0.026584	0.47896	0.41383
Long	MEplum1	7.149737	0.041735	0.423004	0.350879
Long	MEtan	5.451832	0.049424	0.400835	0.32594
PPT	MEbrown4	147.3191	0.024665	0.487802	0.423777
PPT	MEivory	-175.229	0.035616	0.443204	0.373605
PPT	MEdarkslateblue	386.3788	0.040535	0.426765	0.35511
PPT CV	MEsaddlebrown	1.068578	0.002727	0.694965	0.656836

PPT CV	MEskyblue3	0.945086	0.019402	0.515288	0.4547
PPT CV	MEtan	0.673417	0.034629	0.446723	0.377563
PPT CV	MEthistle1	-0.74653	0.045998	0.410335	0.336626
PPT CV	MEyellowgreen	0.808042	0.048442	0.403501	0.328939
Temperature optimum	MEthistle2	-143.111	0.007326	0.613957	0.565701
Temperature optimum	MEblue	-19.2283	0.031072	0.460129	0.392645
Temperature optimum	MEyellowgreen	19.52893	0.047381	0.406432	0.332236
Thermal safety margin	MEgrey60	39.23925	0.011618	0.569746	0.515964
Temperature max	MEbrown4	-7.97617	0.011843	0.567804	0.51378
Temperature max	MEmagenta	-12.2868	0.038216	0.434296	0.363583
Temperature max	MEskyblue	-5.05084	0.047166	0.407031	0.33291
Temperature max SD	MEviolet	-2.49344	0.044391	0.414991	0.341865
Temperature min	MEdarkred	-12.829	0.033536	0.450719	0.382059
Temperature min	MEorange	2.390463	0.039534	0.429971	0.358717
Temperature min	MEplum2	4.484348	0.048133	0.404348	0.329892
Temperature min SD	MEfloralwhite	-4.02432	0.017114	0.529142	0.470284
Temperature min SD	MElightcyan1	-3.64889	0.018441	0.52094	0.461058
Temperature min SD	MEsteelblue	-3.51	0.026301	0.480228	0.415257
Temperature min SD	MEtan	2.868139	0.031163	0.459772	0.392243
Temperature min SD	MEviolet	-3.8882	0.034643	0.446674	0.377508
Temperature min SD	MEgreenyellow	-2.62508	0.042567	0.42045	0.348006

Table S3.4. *Imbibed gene modules significantly correlated with traits of interest*

Response	Predictor	Estimate	p.value	R2	Adj.R2
Breadth	MEblue	26.30203	0.015525	0.53966	0.482117
Breadth	MEdarkgreen	-20.8477	0.03785	0.435521	0.364961
CWD	MElavenderblush3	154.4692	0.005324	0.642085	0.597346
CWD	MEdarkturquoise	-159.442	0.014056	0.550177	0.49395
CWD	MEsienna3	145.1602	0.029761	0.465388	0.398561
CWD	MEdarkorange2	143.9432	0.032978	0.452799	0.384399
CWD	MEbisque4	-224.981	0.042228	0.421487	0.349173
CWD	MEblue	-149.814	0.047002	0.40749	0.333426
Critical temperature max	MEskyblue	48.61722	0.038622	0.432951	0.36207
Critical temperature min	MEbrown	-74.2286	0.005252	0.643239	0.598644
Critical temperature min	MEthistle1	-438.271	0.024433	0.488913	0.425027
Cumulative prop germ	MEblue	2.496268	0.004658	0.653292	0.609953
Cumulative prop germ	MEdarkolivegreen	1.9759	0.012456	0.562683	0.508018
Cumulative prop germ	MEgrey60	1.948182	0.028252	0.471677	0.405637
Cumulative prop germ	MEthistle1	10.12483	0.040458	0.42701	0.355386
Cumulative prop germ	MEdarkgreen	-1.75482	0.044487	0.414708	0.341546

Elevation	MEsienna3	1860.111	0.032408	0.454956	0.386826
Elevation	MEdarkslateblue	1801.338	0.044571	0.414463	0.341271
Lat	MEtan	-11.2335	0.022337	0.499315	0.43673
Lat	MEsienna3	-9.72162	0.028191	0.471941	0.405934
Lat	MEpink	-10.2094	0.029552	0.466246	0.399527
Long	MEpink	10.326	0.003503	0.676085	0.635596
Long	MEsaddlebrown	-8.28632	0.028638	0.470046	0.403802
PPT CV	MEtan	1.346908	0.001335	0.74324	0.711145
PPT CV	MEblue	-1.17708	0.009569	0.588881	0.537491
PPT CV	MEsalmon	0.923866	0.024215	0.48996	0.426205
Temperature optimum	MEbrown	-19.031	0.025532	0.483741	0.419209
Temperature optimum	MEdarkolivegreen	-20.2678	0.044765	0.413896	0.340633
Thermal safety margin	MEskyblue	45.41627	0.012513	0.562209	0.507485
Thermal safety margin	MEbrown4	55.85414	0.037568	0.43647	0.366028
Temperature max	MEmagenta	-8.2627	0.026253	0.480448	0.415504
Temperature max SD	MElightgreen	-1.82668	0.039594	0.429777	0.358499
Temperature min	MEthistle1	-26.8737	0.016437	0.533527	0.475218
Temperature min	MEdarkmagenta	3.301184	0.034967	0.445508	0.376197
Temperature min SD	MEsalmon	3.85557	0.025441	0.484164	0.419684
Temperature min SD	MEsaddlebrown	-3.98292	0.033711	0.45007	0.381329
Temperature min SD	MElightgreen	-2.70025	0.044151	0.415698	0.34266