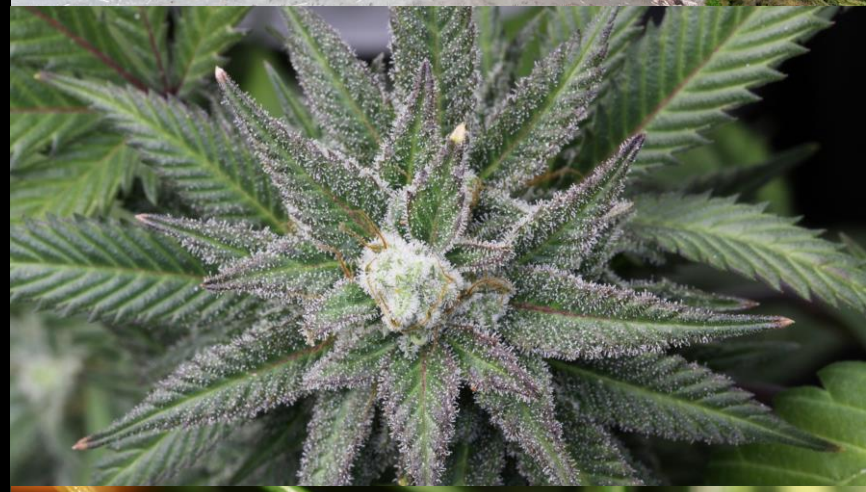
Annotation of complex genomes for comparative genomics

Ph.D. thesis defence

Kristina Gagalova, M.Sc

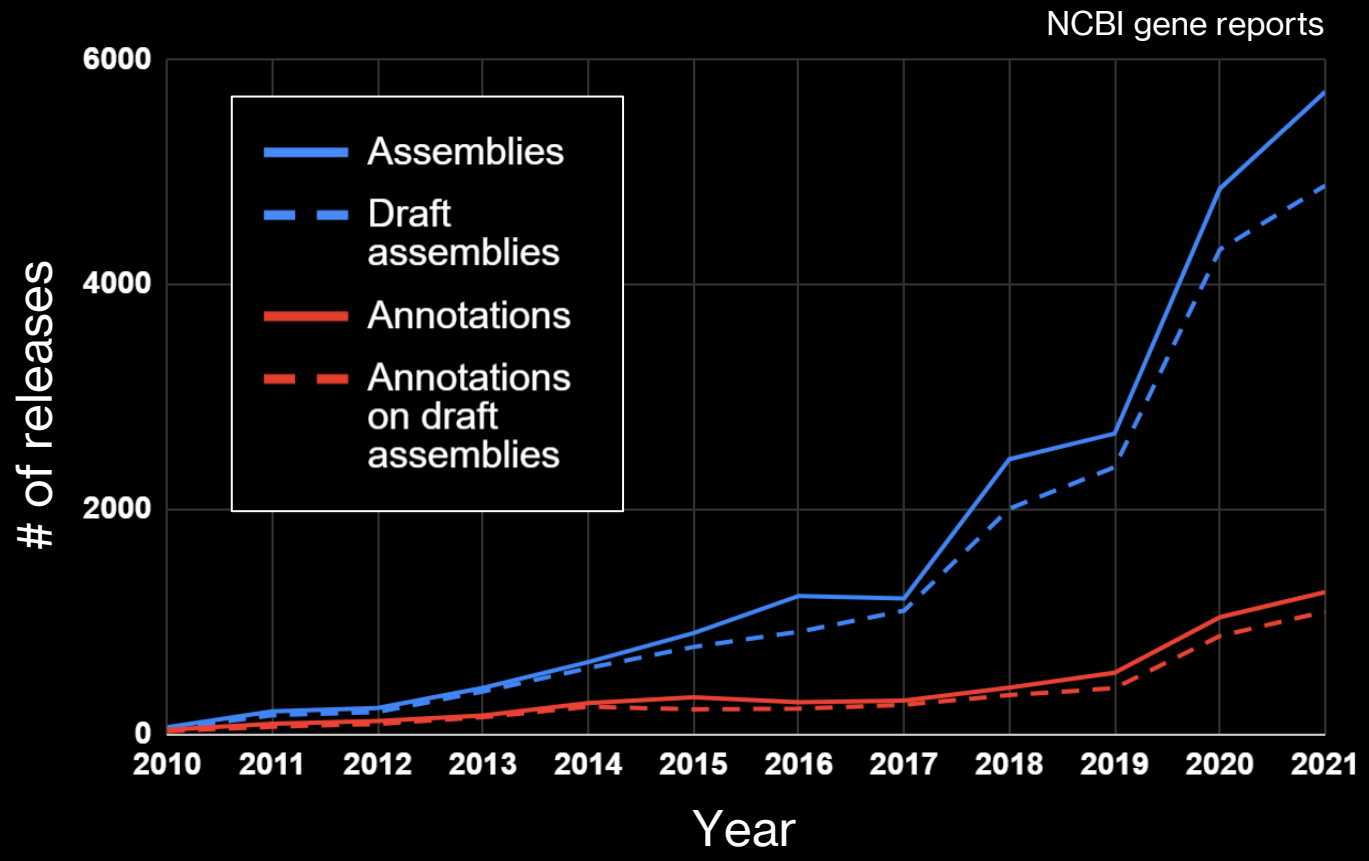


New frontiers in genomics: sequencing of eukaryotes

EARTH BIOGENOME PROJECT
sequencing life for the future of life

JGI Phytozome 13
THE PLANT GENOMICS RESOURCE

i5K 10KP



Research objectives

Annotation of complex genomes for comparative genomics

Research objectives

Annotation of complex genomes for comparative genomics

Growing availability of **non-model** genome assemblies

Many “draft” genomes - large genome size and repeat content

Research objectives

Annotation of complex genomes for comparative genomics

Growing availability of **non-model** genome assemblies
Many “draft” genomes - large genome size and repeat content

Annotation assigns functional information thus biological meaning
Genome wide annotation for protein coding and non-coding regions

Research objectives

Annotation of complex genomes for comparative genomics

Growing availability of **non-model** genome assemblies
Many “draft” genomes - large genome size and repeat content

Annotation assigns functional information thus biological meaning
Genome wide annotation for protein coding and non-coding regions

Comparison to evolutionarily related species
Identify unique species features

Thesis layout

1. Introduction
2. Comparative genome annotation of four North American spruces (*Picea*, Pinaceae)
3. Genome assembly and annotations of *Pissodes strobi*, a North American forest insect pest
4. Genome assembly and annotation of Willow-alpha, a *Cannabis sativa* variety, with a focus on anthocyanin biosynthesis
5. Conclusion

Thesis layout

Contributions

- **Gene annotation**
- **Phylogeny – sc**
- **Gene families**
- **Positive selection**

- **Genome assembly**
- **Genome annotations**
- **Phylogeny**
- **Repeats**

- **Genome assembly**
- **Genome annotations**
- **Phylogeny**
- **Flavonoid genes**
- **Gene expression**

1. Introduction
2. Comparative genome annotation of four North American spruces (*Picea*, Pinaceae)
3. Genome assembly and annotations of *Pissodes strobi*, a North American forest insect pest
4. Genome assembly and annotation of Willow-alpha, a *Cannabis sativa* variety, with a focus on anthocyanin biosynthesis
5. Conclusion

Thesis layout

Publication track

Manuscript under review

10.1093/g3journal/jkac038

Manuscript in preparation

1. Introduction
2. Comparative genome annotation of four North American spruces (*Picea*, Pinaceae)
3. Genome assembly and annotations of *Pissodes strobi*, a North American forest insect pest
4. Genome assembly and annotation of Willow-alpha, a *Cannabis sativa* variety, with a focus on anthocyanin biosynthesis
5. Conclusion

Kristina Gagalova

Genome Sciences Centre
Verified email at bcgsc.ca - Homepage

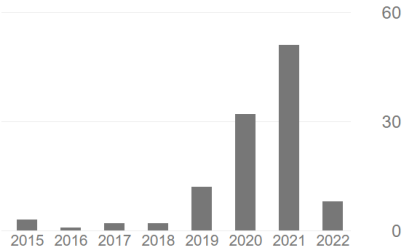
genomics evolutionary genomics

First or co-first author

TITLE	CITED BY	YEAR
The genome of the forest insect pest Pissodes strobi reveals genome expansion and evidence of a Wolbachia endosymbiont KK Gagalova, JGA Whitehill, L Culibrk, D Lin, V Lévesque-Tremblay, ... G3 Genes| Genomes| Genetics		2022
RNA-Scoop: interactive visualization of transcripts in single-cell transcriptomes M Stephenson, KM Nip, S HafezQorani, KK Gagalova, C Yang, ... NAR genomics and bioinformatics 3 (4), lqab105		2021
Pilot Implementation of a Clinical Research Data Warehouse Linking Intra-Operative Physiological Data With Post-Operative Outcomes M Teng, K Gagalova, E Portales-Casamar, M Gorges ANESTHESIA AND ANALGESIA 132, 54-55		2021
What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions KK Gagalova, MAL Elizalde, E Portales-Casamar, M Görges JMIR formative research 4 (8), e17687	7	2020
RNA-Seq in 296 phased trios provides a high-resolution map of genomic imprinting B Jadhav, R Monajemi, KK Gagalova, D Ho, HHM Draisma, ... BMC biology 17 (1), 1-20	14	2019
Induction of interferon-stimulated genes and cellular stress pathways by morpholinos in zebrafish JKH Lai, KK Gagalova, C Kuenne, MA El-Brolosy, DYR Stainier Developmental biology 454 (1), 21-28	21	2019
Complete Chloroplast Genome Sequence of an Engelmann Spruce (*Picea engelmannii*, Genotype Se404-851) from Western Canada D Lin, L Coombe, SD Jackman, KK Gagalova, RL Warren, SA Hammond, ... Microbiology Resource Announcements 8 (24), e00382-19	3	2019
Complete chloroplast genome sequence of a white spruce (Picea glauca, Genotype WS77111) from Eastern Canada D Lin, L Coombe, SD Jackman, KK Gagalova, RL Warren, SA Hammond, ... Microbiology Resource Announcements 8 (23), e00381-19	4	2019
Skewed X-inactivation is common in the general female population E Shvetsova, A Sofronova, R Monajemi, K Gagalova, HHM Draisma, ... European Journal of Human Genetics 27 (3), 455-465	45	2019

Cited by VIEW ALL

	All	Since 2017
Citations	115	107
h-index	6	5
i10-index	3	3



Public access VIEW ALL

0 articles 9 articles

not available available

Based on funding mandates

Co-authors


Szymon M. Kiełbasa
Biomedical Data Sciences


Harmen Draisma
Research Associate, Section of …

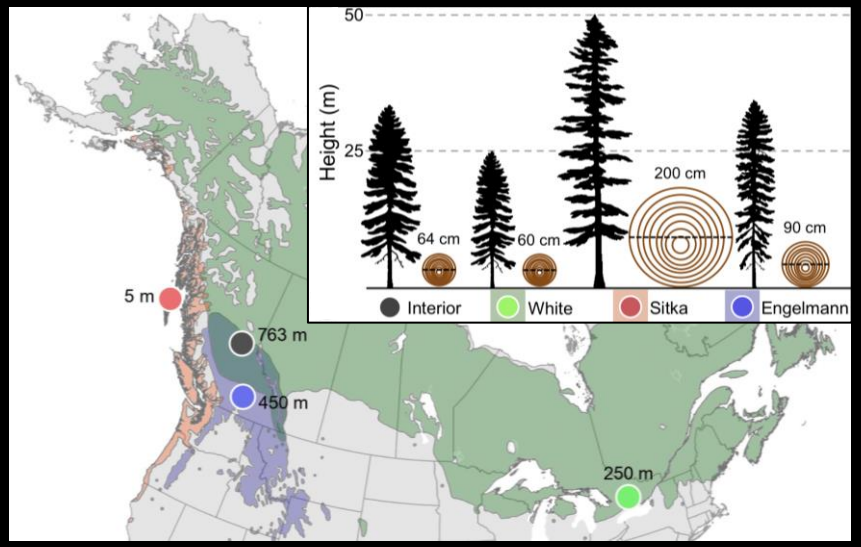

Lude Franke
Professor of Functional Genomic…


Bas Heijmans
Leiden University Medical Center…


Rick Jansen
Assistant Professor, Department …

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

Photo: LukaDrone

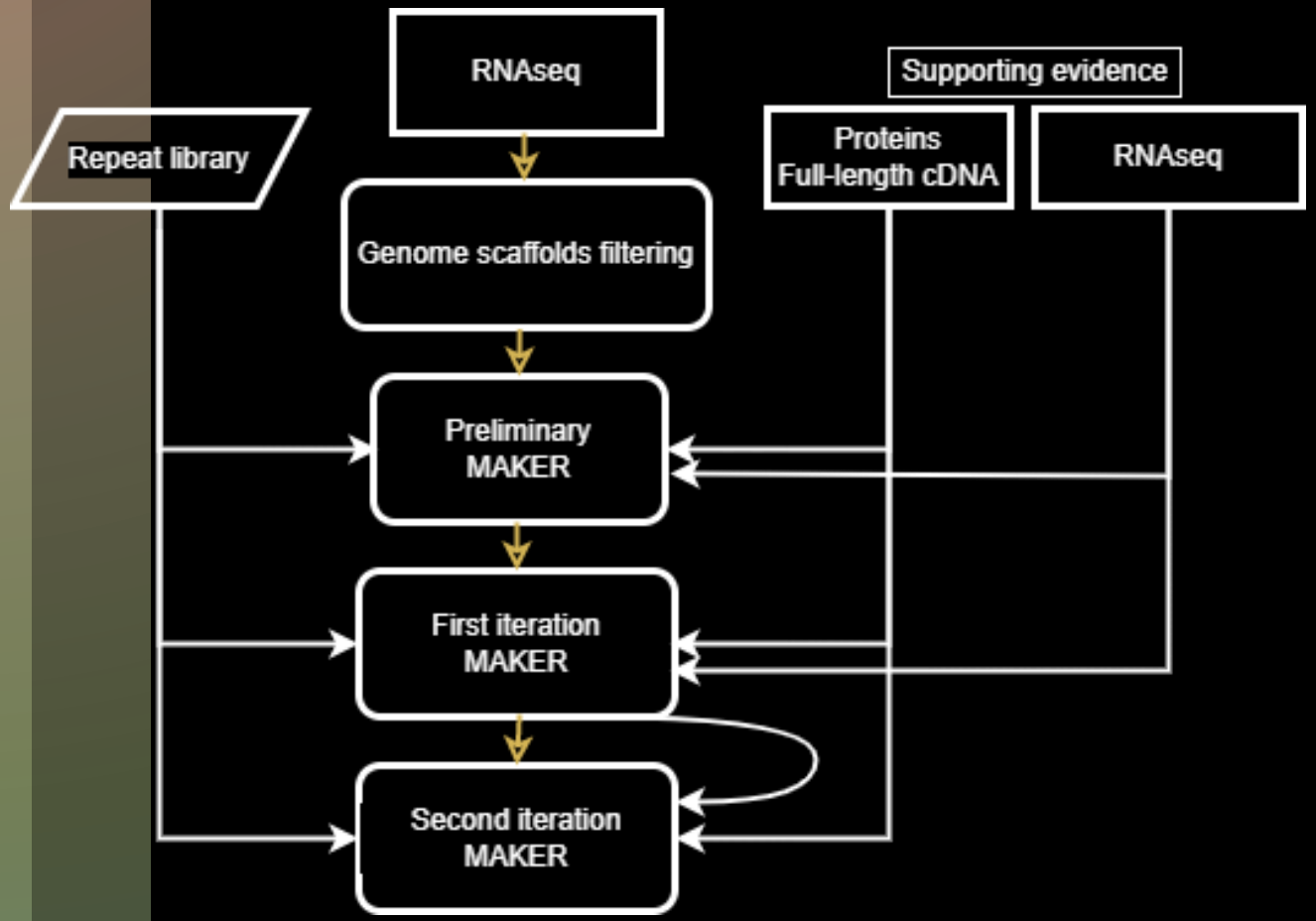


Draft genome assemblies with similar contiguity

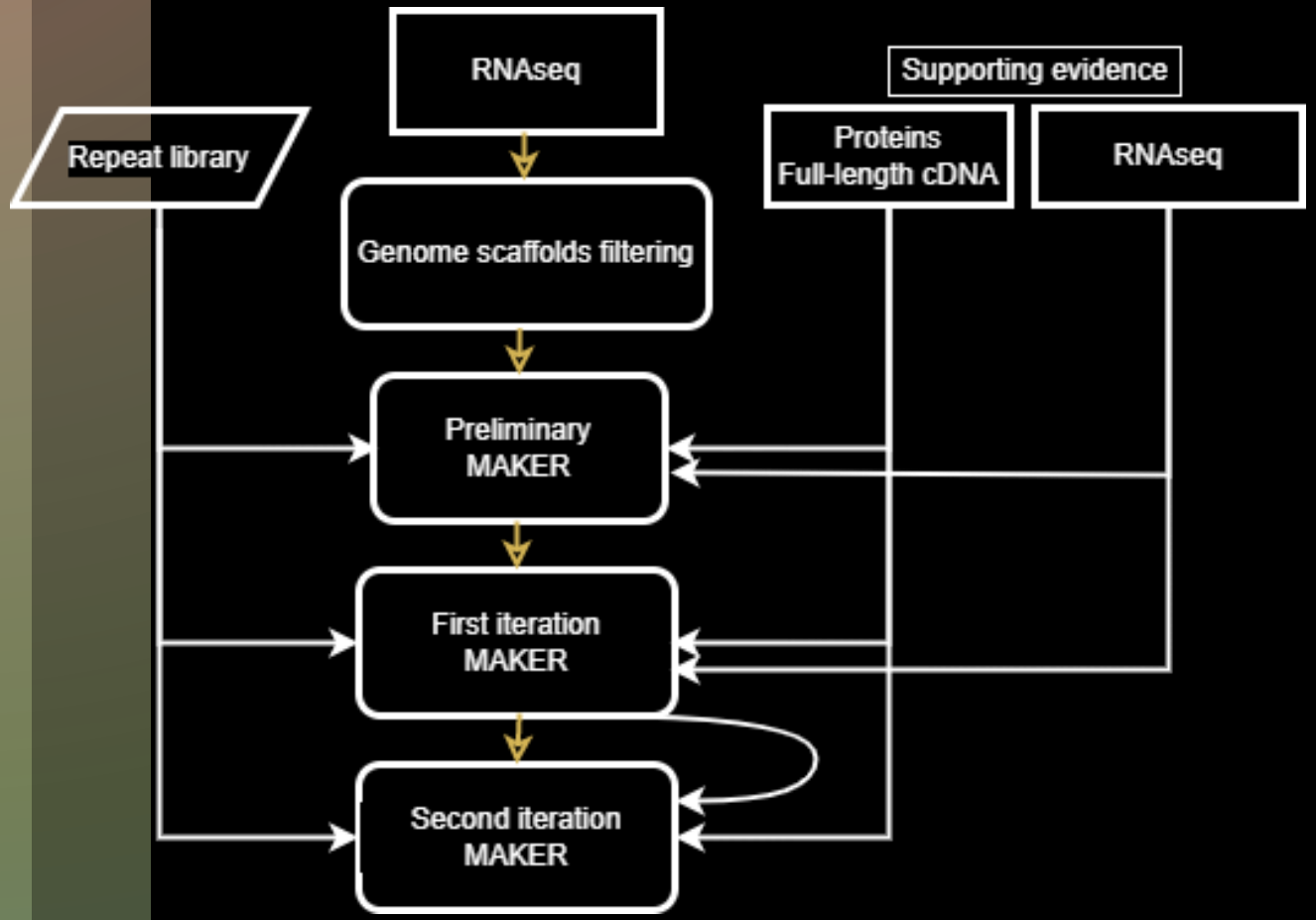
Comprehensive protein coding gene annotation of spruces

Common annotation methodology for comparative genomics

Genome annotation to gain insights into interspecies variability of spruces



Genome annotation to gain insights into interspecies variability of spruces



Genome assembly statistics

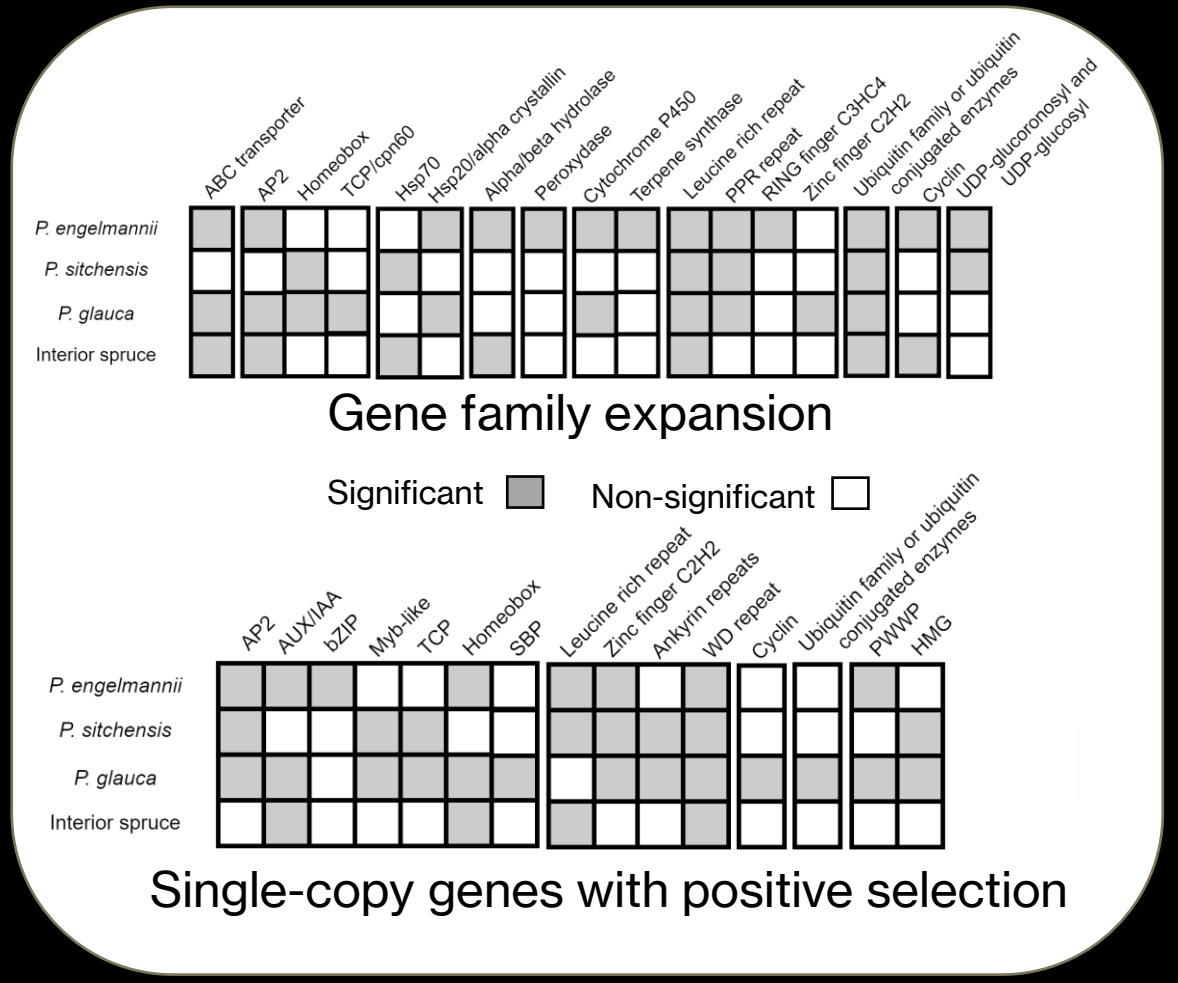
Taxon	NG50 (bp)	Reconstr. size (Gb)	BUSCO single-copy (%)	BUSCO duplicated (%)
P. engelmannii	355,449	20.75	**40.3**	**8.2**
P. sitchensis	38,458	18.22	**29.5**	**7.4**
P. glauca	131,339	21.58	**39.9**	**8.0**
Interior spruce	121,714	20.14	**41.1**	**7.8**

Annotation statistics

Taxon	Total genes	Total mRNA	BUSCO complete (%)	Pfam complete (%)
P. engelmannii	34,365	60,224	**17.6**	**32.01**
P. sitchensis	30,324	58,175	**18.2**	**33.56**
P. glauca	30,410	56,535	**18.0**	**32.14**
Interior spruce	28,944	62,397	**18.2**	**33.36**

Embryophyte BUSCO n=1,616
Pfam plants 918 single and 563 multiple CDA

Genome annotation to gain insights into interspecies variability of spruces



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

Photo: Justin Whitehill

Draft genome assembly

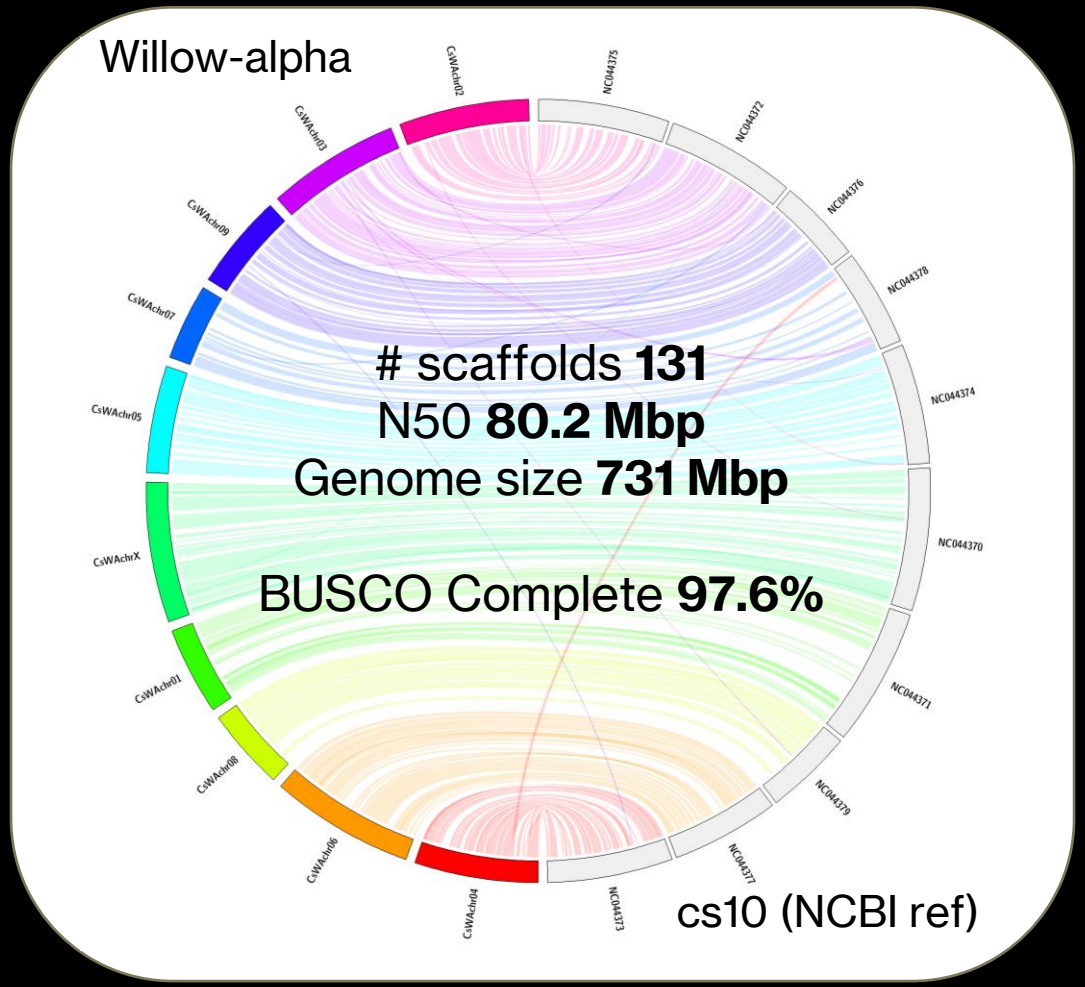
Genome annotations

- Protein coding genes
- Repeats - transposable elements (TE)

Expanded genome size in the species family (Curculionidae)

Large fraction of TE elements in the genome, likely driving the genome expansion

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



Chromosome scale genome assembly

High completeness genome annotation

Complete BUSCO (%)	**94.9**
Complete Pfam (%)	**96.8**
# of transcripts	41,912
# of genes	38,559

Embryophyte BUSCO n=1,616
Pfam plants 918 single and 563 multiple CDA

Flavonoid biosynthesis in *C. sativa*

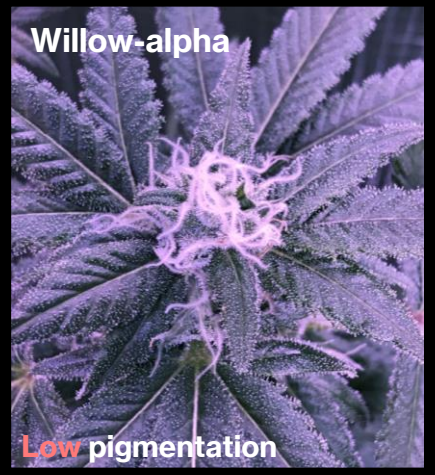
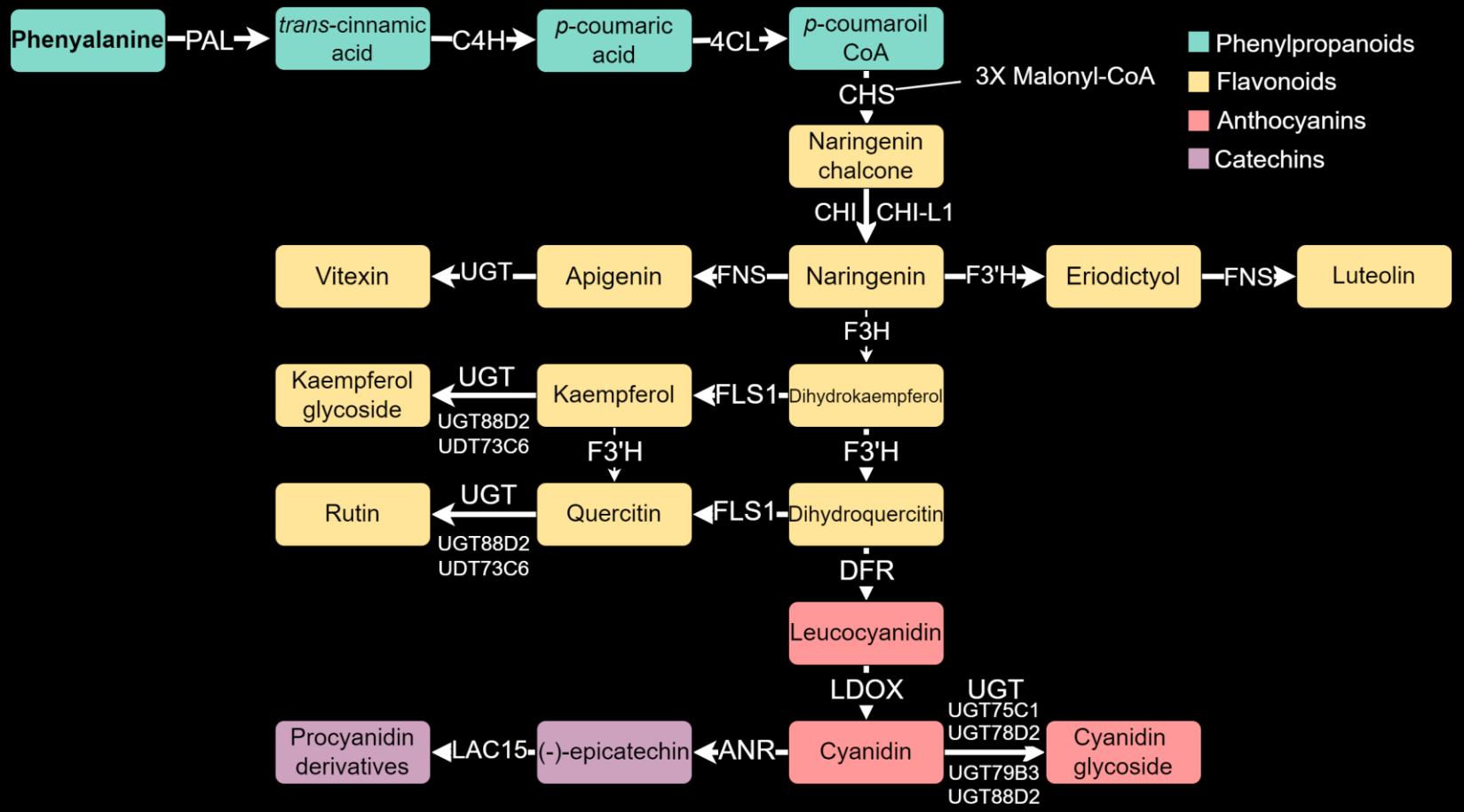


Photo: Mathias Schuetz



Flavonoid biosynthesis in *C. sativa*

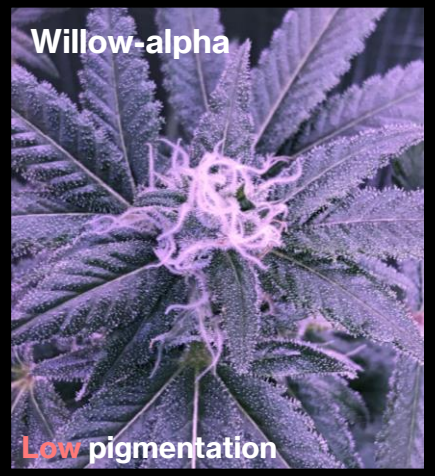
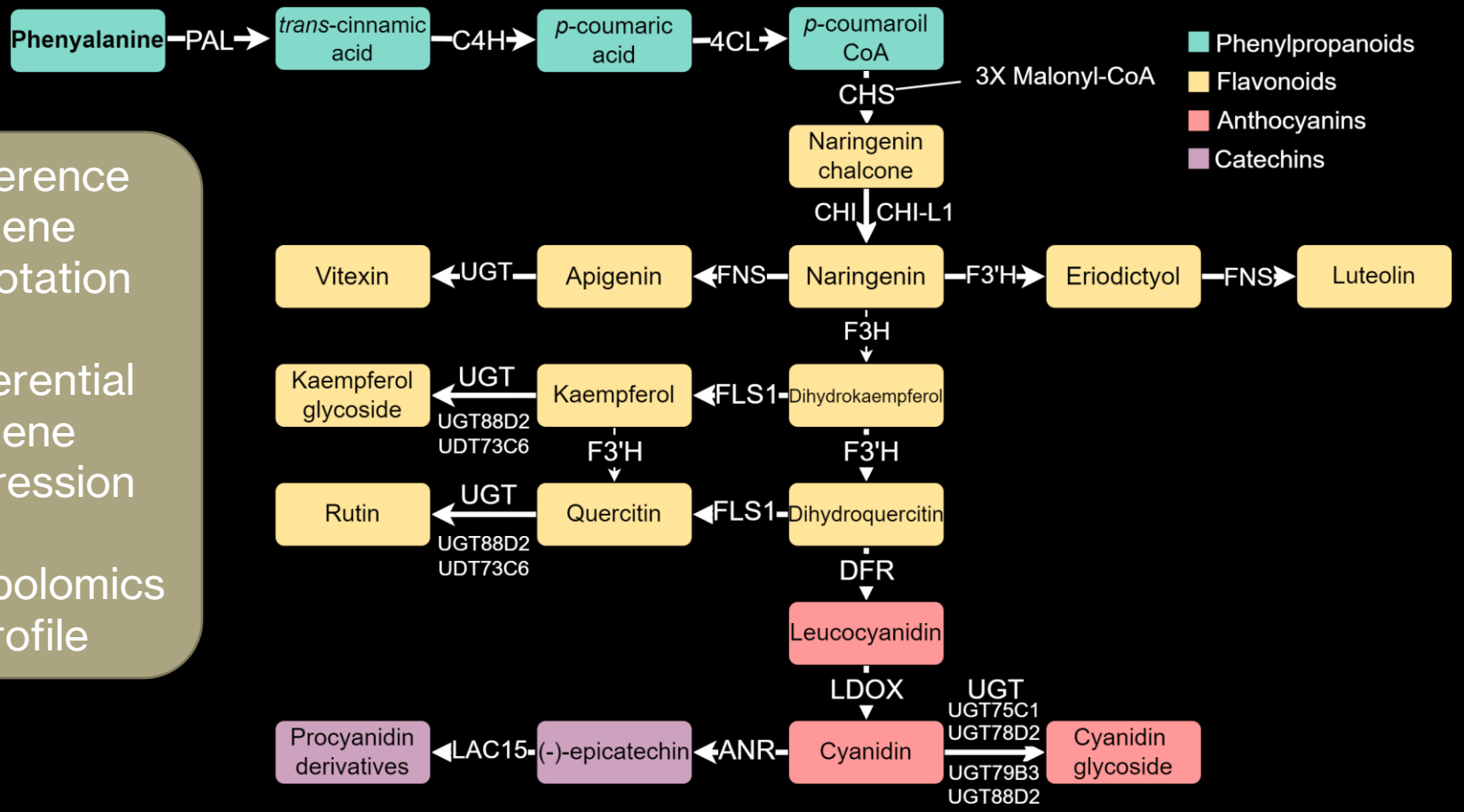


Photo: Mathias Schuetz

Reference gene annotation

Differential gene expression

Metabolomics profile



Flavonoid biosynthesis in *C. sativa*

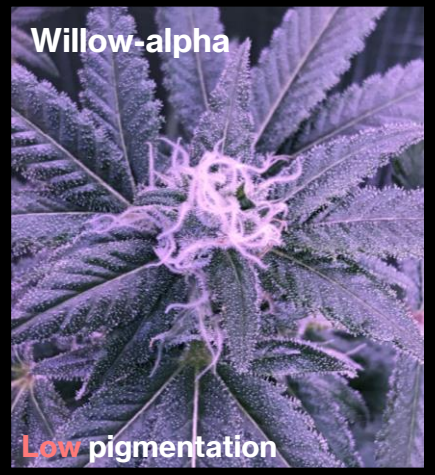
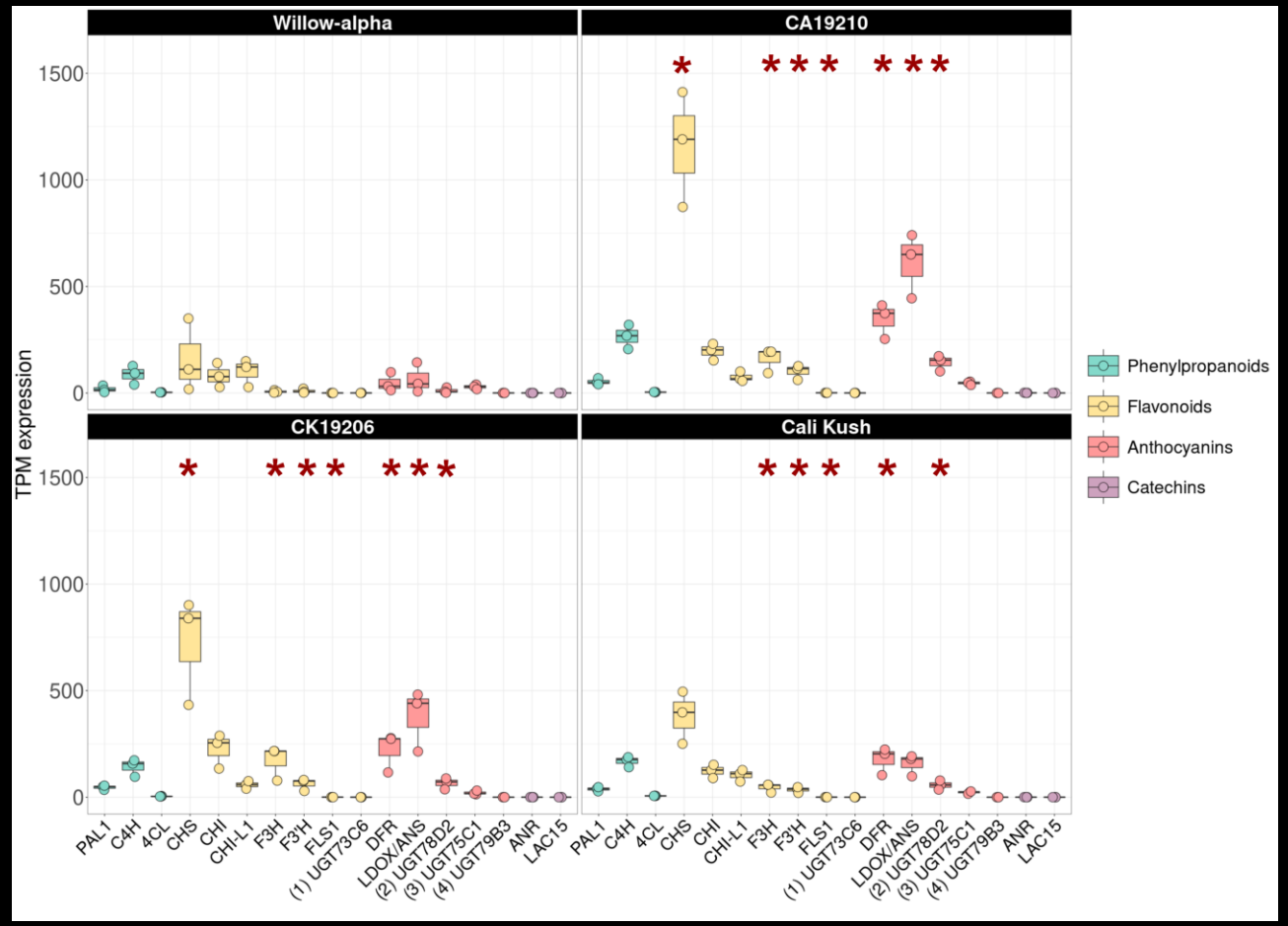


Photo: Mathias Schuetz

Reference gene annotation

Differential gene expression

Metabolomics profile



Flavonoid biosynthesis in *C. sativa*

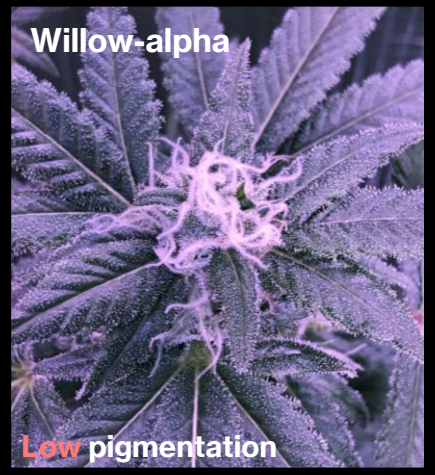
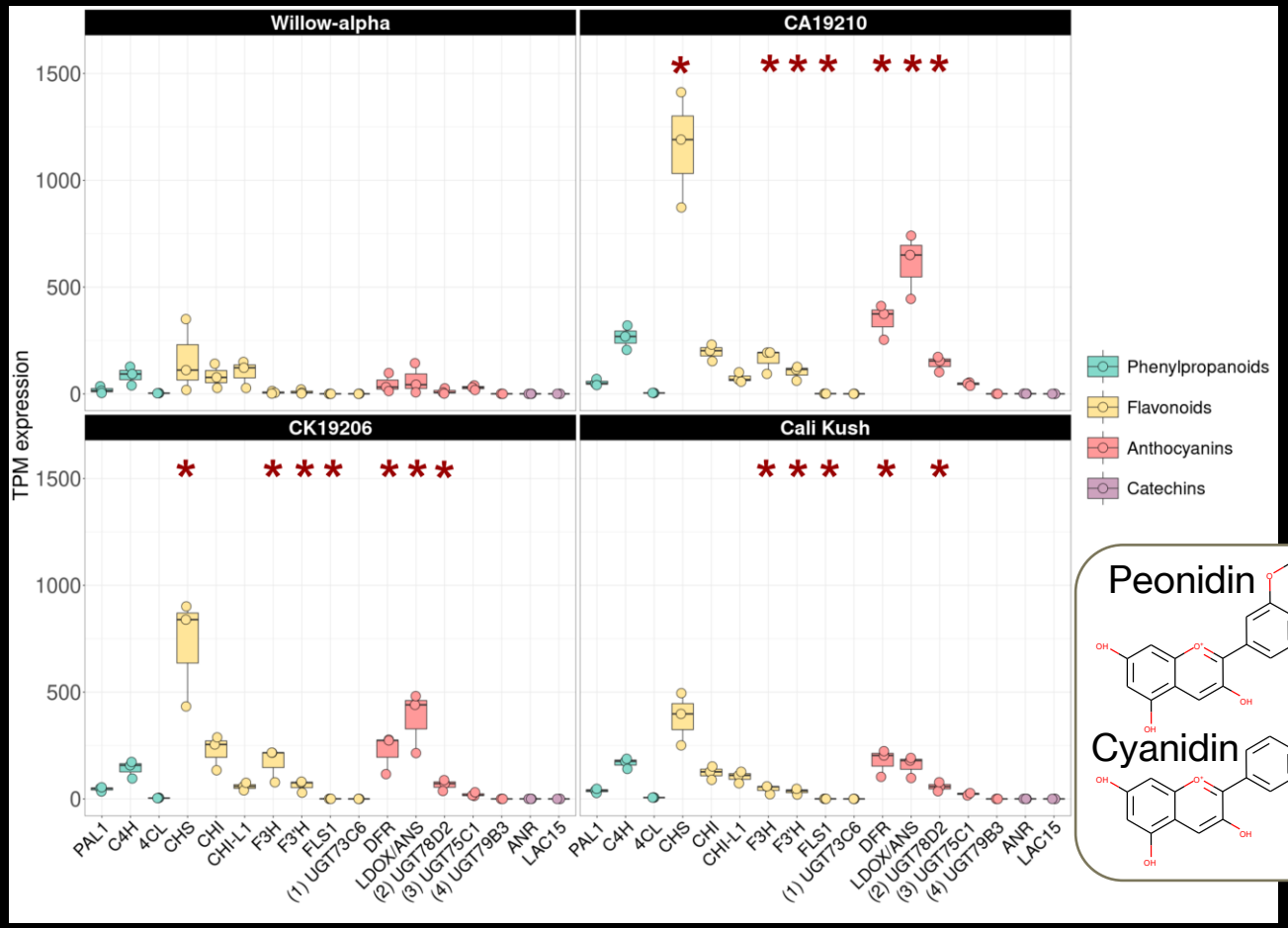


Photo: Mathias Schuetz

Reference gene annotation

Differential gene expression

Metabolomics profile



Impact of the work and future directions

Annotation resources for the research community

Whole-genome comparative analysis

Phylogenomic and genome evolution analysis

Impact of the work and future directions

Annotation resources for the research community

Whole-genome comparative analysis

Phylogenomic and genome evolution analysis

Intra-species variability of gen. *Picea* in North America

Molecular features of local adaptation

Genome characterization of *Pissodes strobi*

Characterization of flavonoid genes in *C. sativa*

Gene candidates regulating the leaf pigmentation

Impact of the work and future directions

Annotation resources for the research community

Whole-genome comparative analysis

Phylogenomic and genome evolution analysis

Intra-species variability of gen. *Picea* in North America

Molecular features of local adaptation

Genome characterization of *Pissodes strobi*

Characterization of flavonoid genes in *C. sativa*

Gene candidates regulating the leaf pigmentation

Population studies for inter-species diversity in *Picea* and *Pissodes strobi*

Selective breeding in *C. sativa*

Supervisor

Prof. Inanc Birol

Committee

Prof. Sean Graham

Prof. Mathias Schuetz

Prof. Steven Jones

Prof. Joerg Bohlmann

René Warren

Lauren Coombe

Ka Ming Nip

Macaire Man Saint Yuen

Dr. Nathalie Pavy

Dr. Carol Ritland

Prof. Loren Rieseberg

Prof. Justin Whitehill

Luka Culibrk

Dr. Jahanshah Ashkani

Dr. Christopher Keeling

Dr. Matt Workentine

Dr. Till Matzat

Dr. Shumin Wang

Prof. Simone Castellarin

Yifan Yan

Annotation of complex genomes for comparative genomics

Kristina Gagalova

Supervisor

Prof. Inanc Birol

Committee

Prof. Sean Graham

Prof. Mathias Schuetz

Prof. Steven Jones

Prof. Joerg Bohlmann

René Warren

Lauren Coombe

Ka Ming Nip

Macaire Man Saint Yuen

Dr. Nathalie Pavy

Dr. Carol Ritland

Prof. Loren Rieseberg

Prof. Justin Whitehill

Luka Culibrk

Dr. Jahanshah Ashkani

Dr. Christopher Keeling

Dr. Matt Workentine

Dr. Till Matzat

Dr. Shumin Wang

Prof. Simone Castellarin

Yifan Yan

Annotation of complex genomes for comparative genomics

Kristina Gagalova

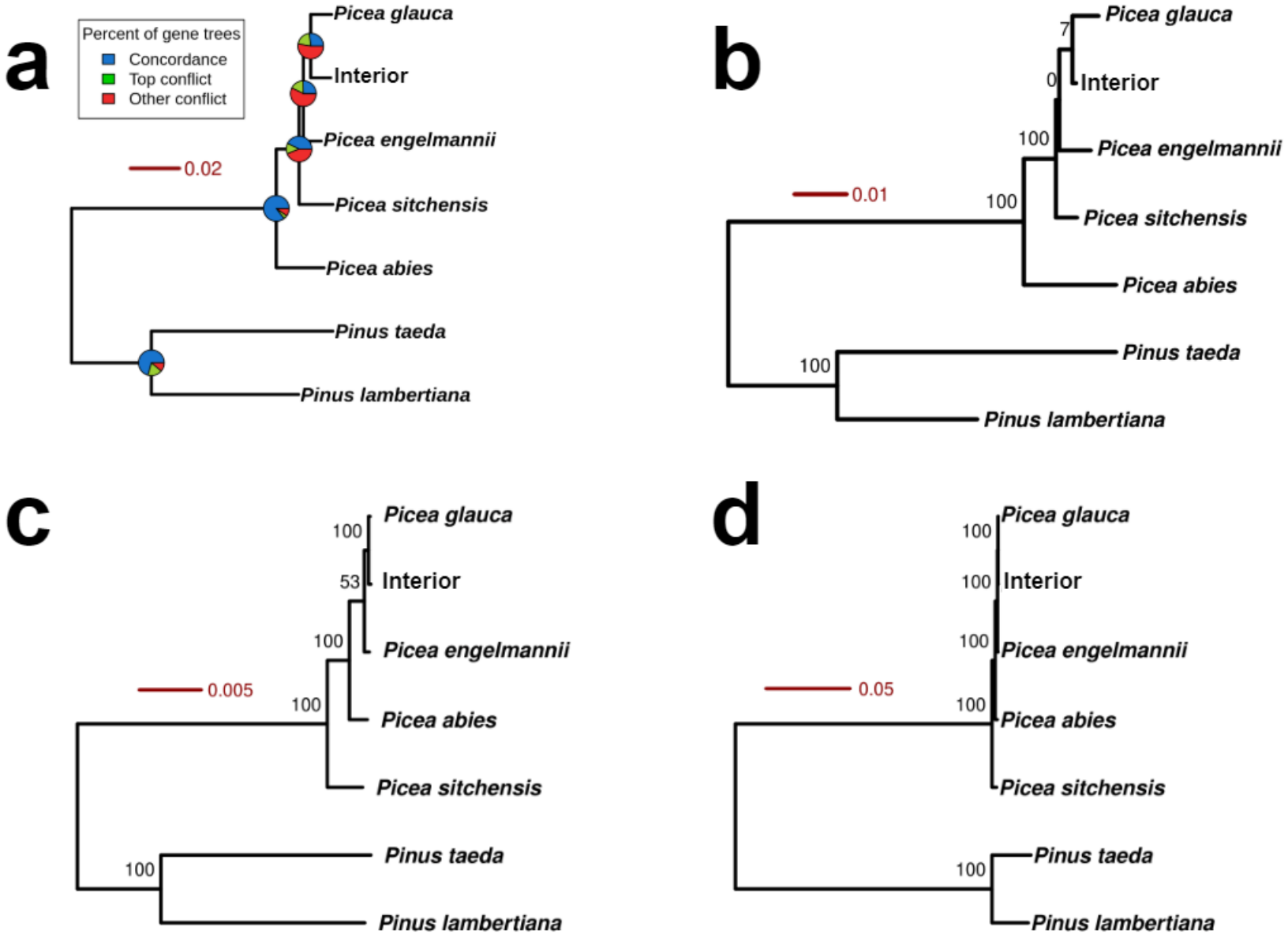

a
Percent of gene trees
Concordance
Top conflict
Other conflict
0.02
Picea glauca
Interior
Picea engelmannii
Picea sitchensis
Picea abies
Pinus taeda
Pinus lambertiana
b
0.01
7
0
100
100
100
Picea glauca
Interior
Picea engelmannii
Picea sitchensis
Picea abies
Pinus taeda
Pinus lambertiana
c
0.005
100
53
100
100
100
Picea glauca
Interior
Picea engelmannii
Picea abies
Picea sitchensis
Pinus taeda
Pinus lambertiana
d
0.05
100
100
100
100
100
Picea glauca
Interior
Picea engelmannii
Picea abies
Picea sitchensis
Pinus taeda
Pinus lambertiana

Figure 2.2

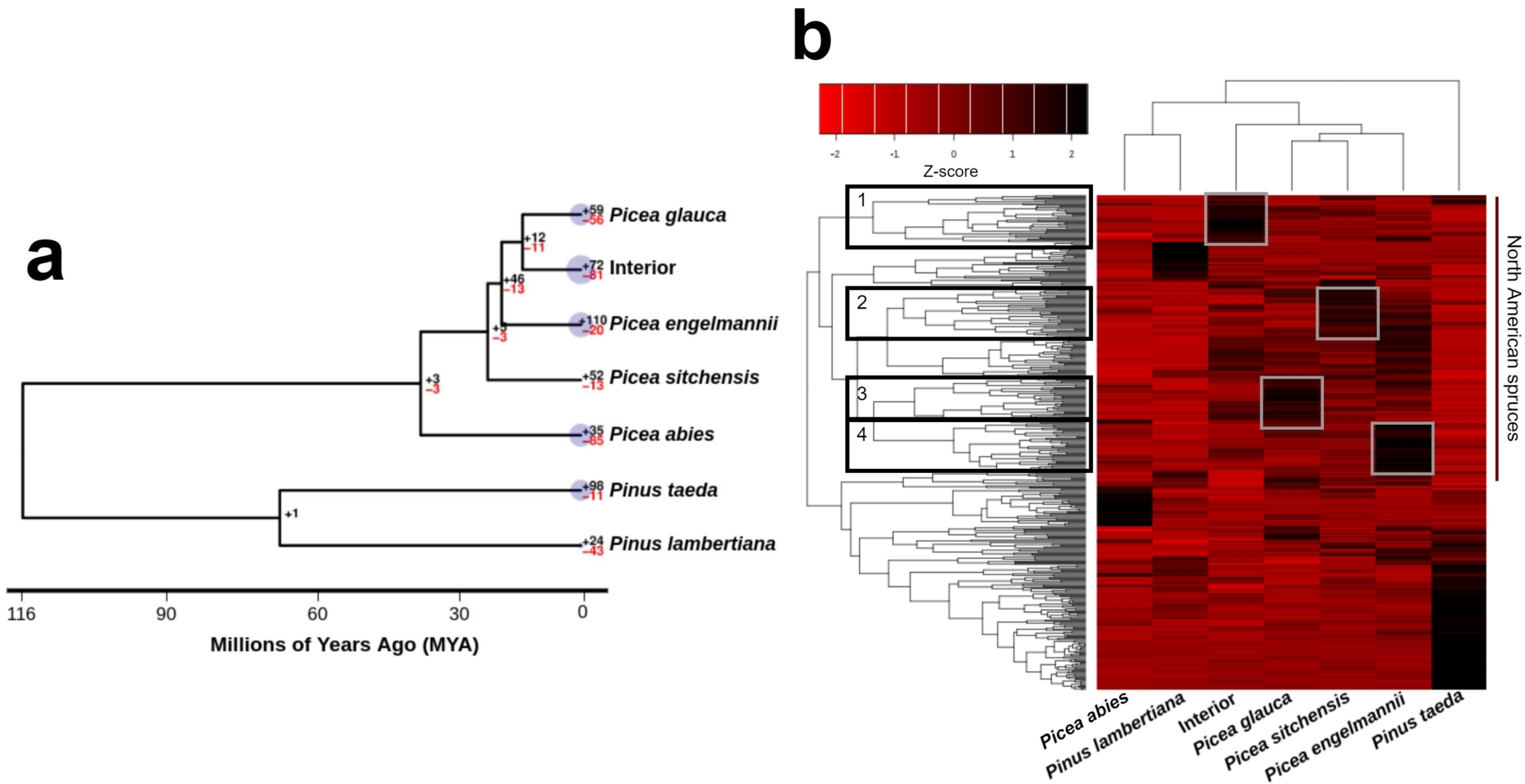

a
+59
−56
Picea glauca
+12
−11
+72
−81
Interior
+46
−13
+110
−20
Picea engelmannii
+3
−3
+52
−13
Picea sitchensis
+3
−3
+35
−85
Picea abies
+98
−11
Pinus taeda
+1
+24
−43
Pinus lambertiana
116
90
60
30
0
Millions of Years Ago (MYA)
b
-2
-1
0
1
2
Z-score
1
2
3
4
North American spruces
Picea abies
Pinus lambertiana
Interior
Picea glauca
Picea sitchensis
Picea engelmannii
Pinus taeda

Figure 2.3

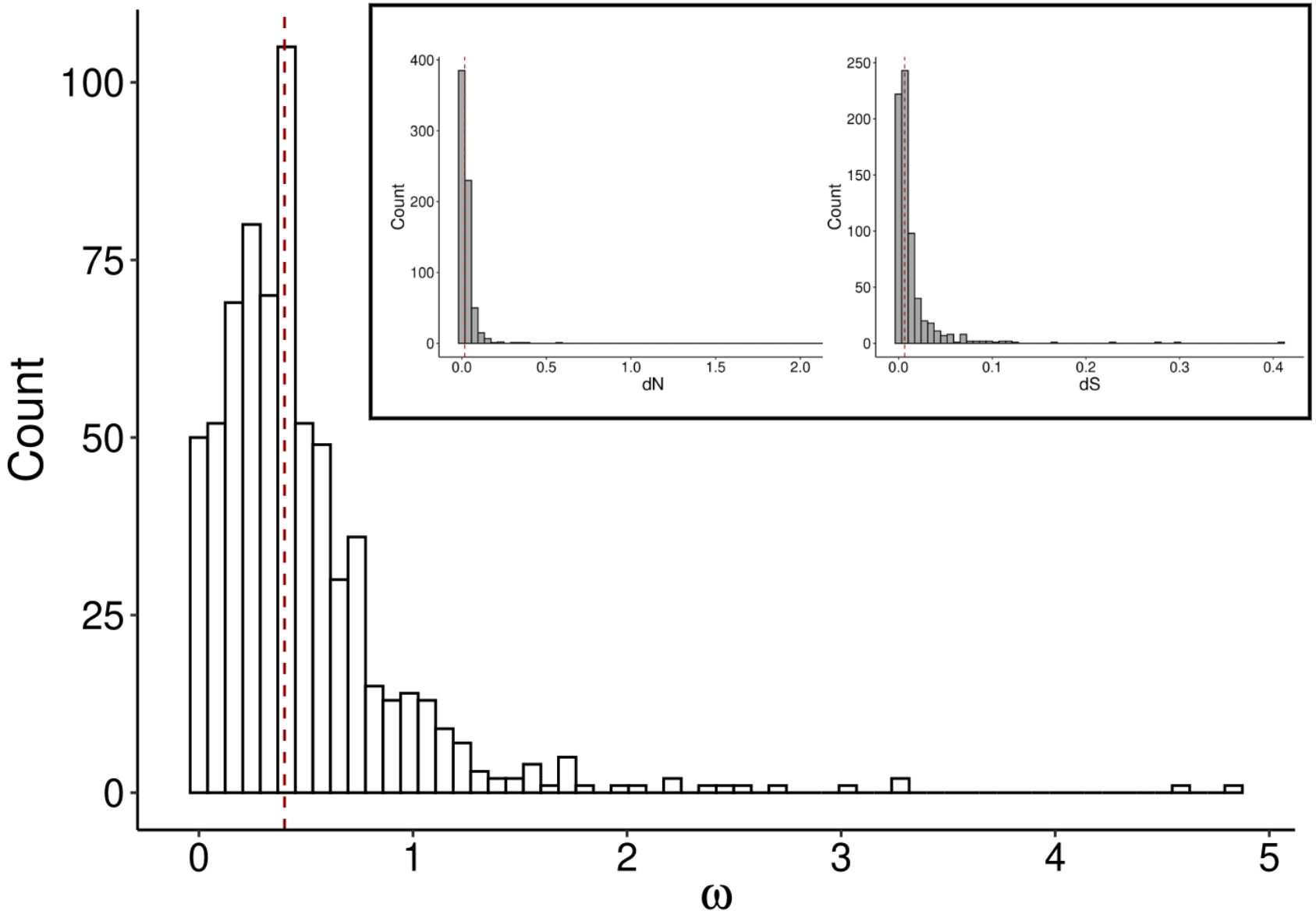

Count
100
75
50
25
0
0
1
2
3
4
5
ω
Count
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
dN
Count
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
dS

Figure 2.4

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

Table 2.1

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

Table 2.2

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

Table 2.3

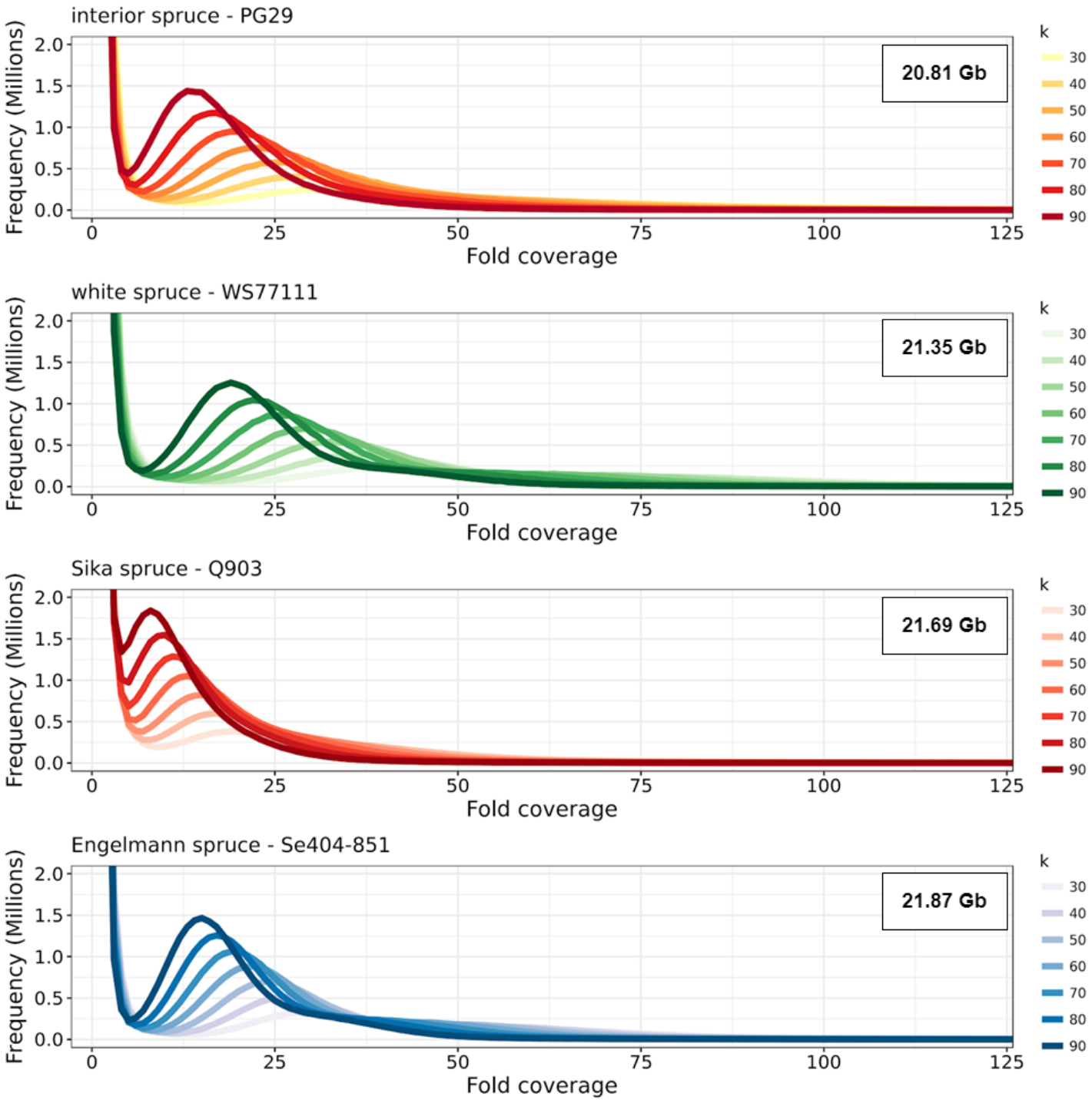

interior spruce - PG29
20.81 Gb
white spruce - WS77111
21.35 Gb
Sika spruce - Q903
21.69 Gb
Engelmann spruce - Se404-851
21.87 Gb
Frequency (Millions)
Fold coverage
k
30
40
50
60
70
80
90

Figure A.2

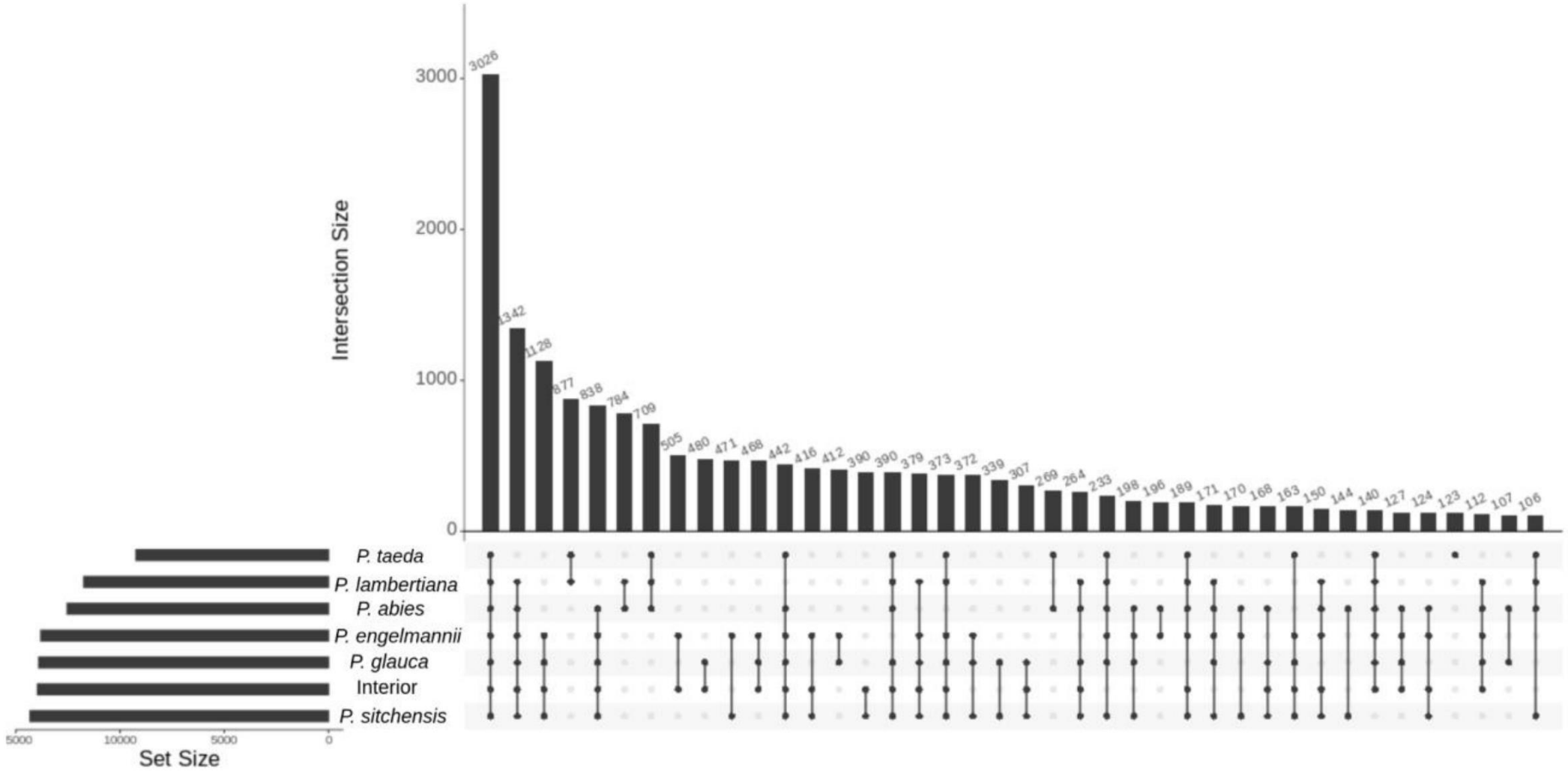

Intersection Size
3000
2000
1000
0
3026
1342
1128
877
838
784
709
505
480
471
468
442
416
412
390
390
379
373
372
339
307
269
264
233
198
196
189
171
170
168
163
150
144
140
127
124
123
112
107
106
P. taeda
P. lambertiana
P. abies
P. engelmannii
P. glauca
Interior
P. sitchensis
5000
10000
5000
0
Set Size

Figure A.3

Platform	Read format	*P. engelmannii* Se404-851	*P. sitchensis* Q903	*P. glauca* WS77111	Interior spruce PG29
Chromium linked reads	128bp-151bp	45.8 (6,892M)	58.3 (8,774M)	37.9 (5,711M)	28.9 (4,349M)
GemCode linked reads	116bp-126bp	-	5.8 (1,000M)	-	-
HiSeq	2x150 bp	-	-	34.0 (4,754M)	60.5 (8,464M)
HiSeq	2x250 bp	29.6 (2,486M)	13 (1,097M)	19.6 (1,644M)	-
MiSeq	2x300 bp	-	-	3.6 (249M)	2.6 (182M)
MiSeq	2x500 bp	-	-	-	2.5 (107M)
MPET	2x100 bp	-	-	10.7 (2247M)	23.9 (5,022M)
MPET	2x150 bp	11.2 (1,561M)	-	-	-
ONT	Variable, read N50 as indicated	2.6 (5.86M) N50=14,150bp	4.9 (9.59M) N50=14,866bp	-	-
Total cov.		**89.2**	**82.0**	**109.5**	**118.4**

Table A.2

Taxon	Genotype	Number of samples	BioProject	Application
Interior spruce	PG29	8	PRJNA210511	Filtering, Gene prediction
P. glauca	PG653	30	Submission in progress	Filtering
P. glauca	PG653	14	PRJNA290034	Filtering
P. glauca	PG653	16	PRJNA309861	Filtering
P. sitchensis	Q903	6	Submission in progress	Filtering, Gene prediction
P. sitchensis	Q903	12	PRJNA398042	Filtering, Gene prediction
P. sitchensis	Q903	1	PRJNA304257	Filtering, Gene prediction
P. sitchensis	H898	6	Submission in progress	Filtering
P. sitchensis	H898	12	PRJNA398042	Filtering

Table A.3

Genome	No. of scaffolds	Longest scaffold	Scaffold NG50	Reconstruction size (Gb)	BUSCO single-copy (%)	BUSCO duplicated (%)
P. engelmannii	946,053	6,646,027	355,449	20.75	40.3	8.2
P. sitchensis	1,770,974	1,973,130	38,458	18.22	29.5	7.4
P. glauca	2,443,500	4,209,077	131,339	21.58	39.9	8.0
Interior spruce	2,064,648	3,588,992	121,714	20.14	41.1	7.8

Table A.5

Species	No. of scaffolds	Longest scaffold	Scaffold NG50	Reconstructed size (Gb)	BUSCO single-copy (%)	BUSCO duplicated (%)
P. abies	1,963,820	194,057	1,000	9.16	28.9	6.0
P. lambertiana	16,610	2,237,000	2,540,398	19.82	54.6	7.1
P. taeda	991,362	8,214,401	78,650	16.78	48.7	7.6

Table A.6

Species	Total genes	Total mRNA	Median gene length (bp)	Medianm RNA length (bp)	Median exon length (bp)	Median intron length (bp)	BUSCO complete (%)	Total CDA completeness (%)
P. abies	26,437	26,437	1,366	714	173	173	26.3	35.85
P. lambertiana	38,518	38,518	4,330	978	143*	271	59.5	69.28
P. taeda	47,602	47,602	2,252	700	166*	184	17.5	28.90

Table A.7

Genotype	Expanded	Contracted	Genes gained	Genes lost	Avg. expansion
Picea glauca	2,221 (59)	1,403 (56)	3,405	1,855	0.1638
Interior spruce	1,392 (72)	1,924 (81)	2,313	2,625	-0.03298
Picea engelmannii	2,762 (110)	1,351 (20)	4,846	1,653	0.3375
Picea sitchensis	1,750 (52)	1,647 (13)	2,905	1,984	0.0973
Picea abies	1,302 (35)	2,839 (85)	2,301	4,266	-0.2077
Pinus taeda	1,970 (98)	3,532 (11)	6,594	4,137	0.25975
Pinus lambertiana	1,603 (24)	2,126 (43)	3,093	3,694	-0.0635

Table A.9

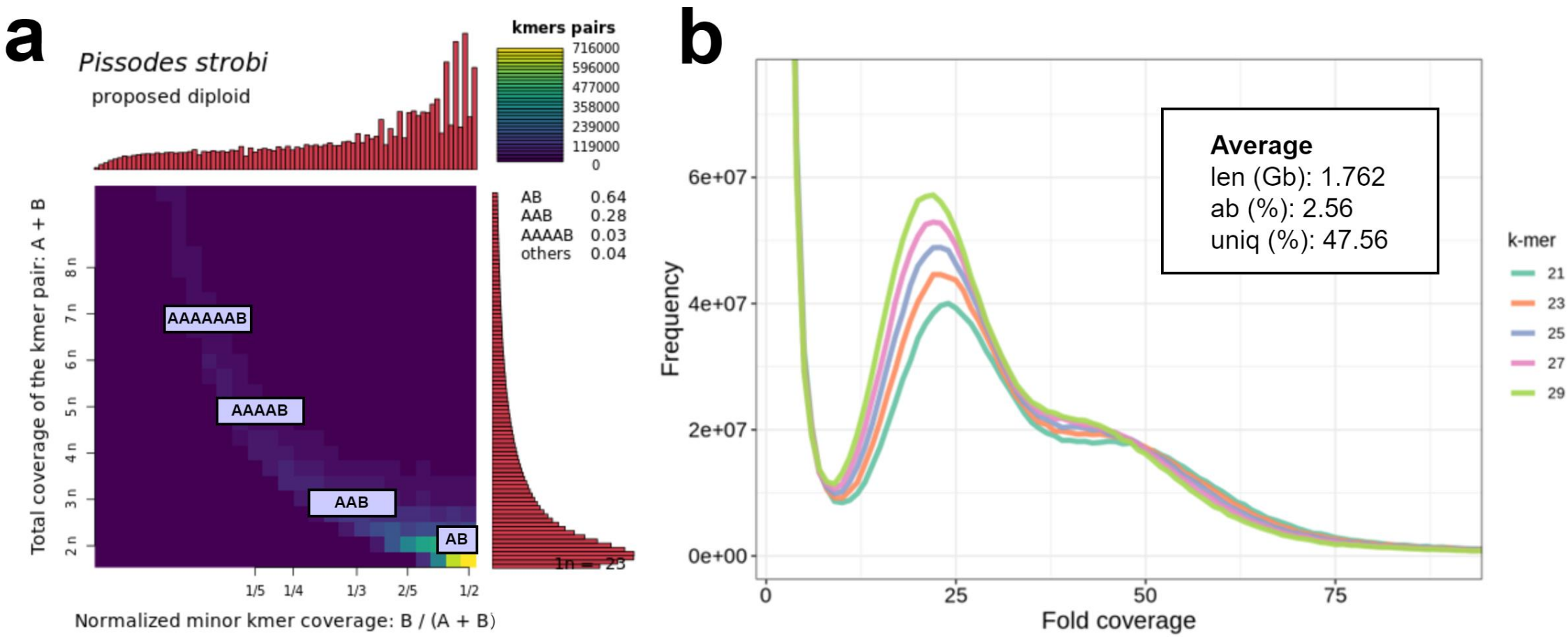

a
Pissodes strobi
proposed diploid
kmers pairs
716000
596000
477000
358000
239000
119000
0
AB 0.64
AAB 0.28
AAAAB 0.03
others 0.04
AAAAAAB
AAAAB
AAB
AB
1n = 23
Total coverage of the kmer pair: A + B
2 n
3 n
4 n
5 n
6 n
7 n
8 n
1/5
1/4
1/3
2/5
1/2
Normalized minor kmer coverage: B / (A + B)
b
Average
len (Gb): 1.762
ab (%): 2.56
uniq (%): 47.56
Frequency
6e+07
4e+07
2e+07
0e+00
0
25
50
75
Fold coverage
k-mer
21
23
25
27
29

Figure 3.1

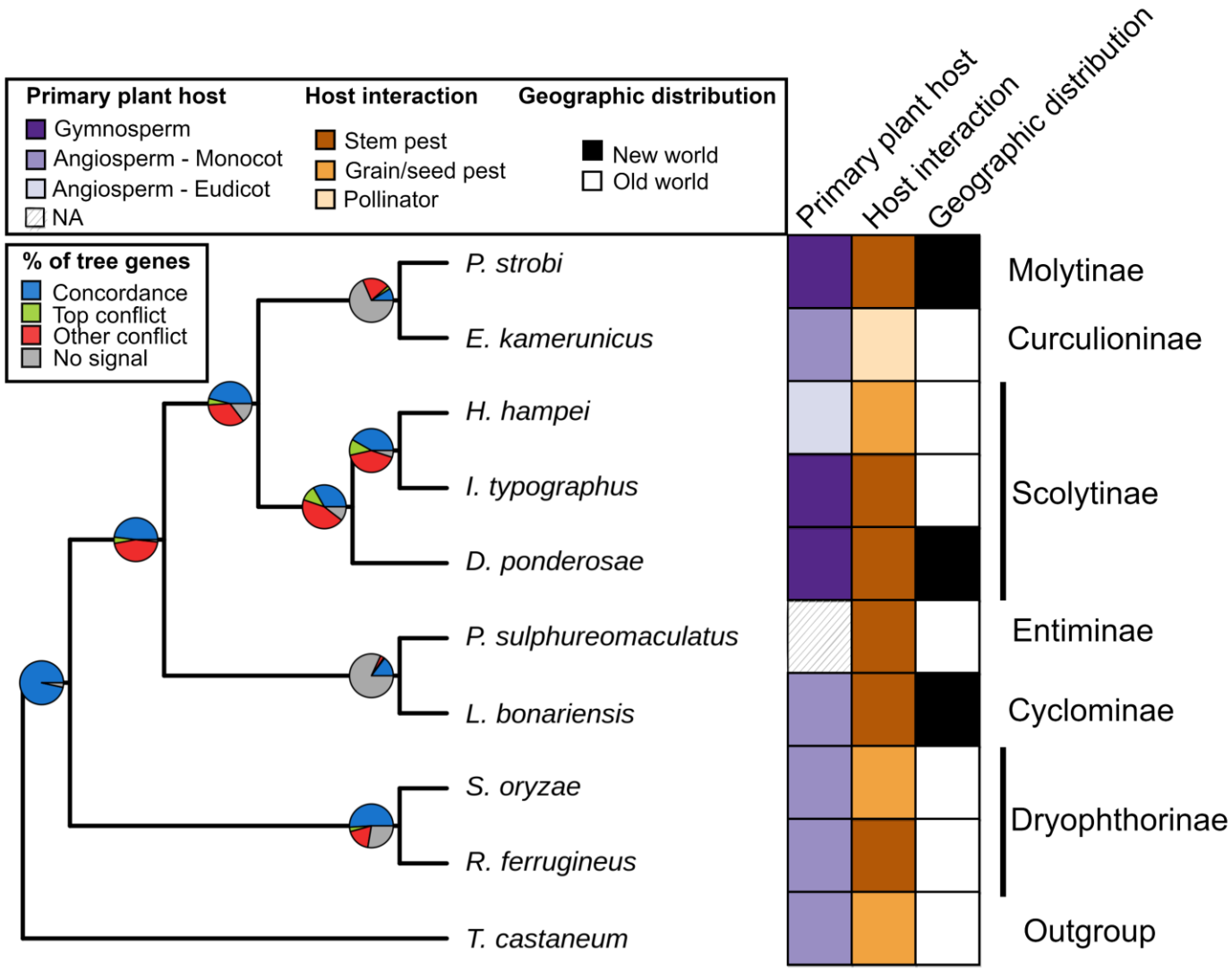

Primary plant host
Gymnosperm
Angiosperm - Monocot
Angiosperm - Eudicot
NA
Host interaction
Stem pest
Grain/seed pest
Pollinator
Geographic distribution
New world
Old world
% of tree genes
Concordance
Top conflict
Other conflict
No signal
P. strobi
E. kamerunicus
H. hampei
I. typographus
D. ponderosae
P. sulphureomaculatus
L. bonariensis
S. oryzae
R. ferrugineus
T. castaneum
Primary plant host
Host interaction
Geographic distribution
Molytinae
Curculioninae
Scolytinae
Entiminae
Cyclominae
Dryophthorinae
Outgroup

Figure 3.2

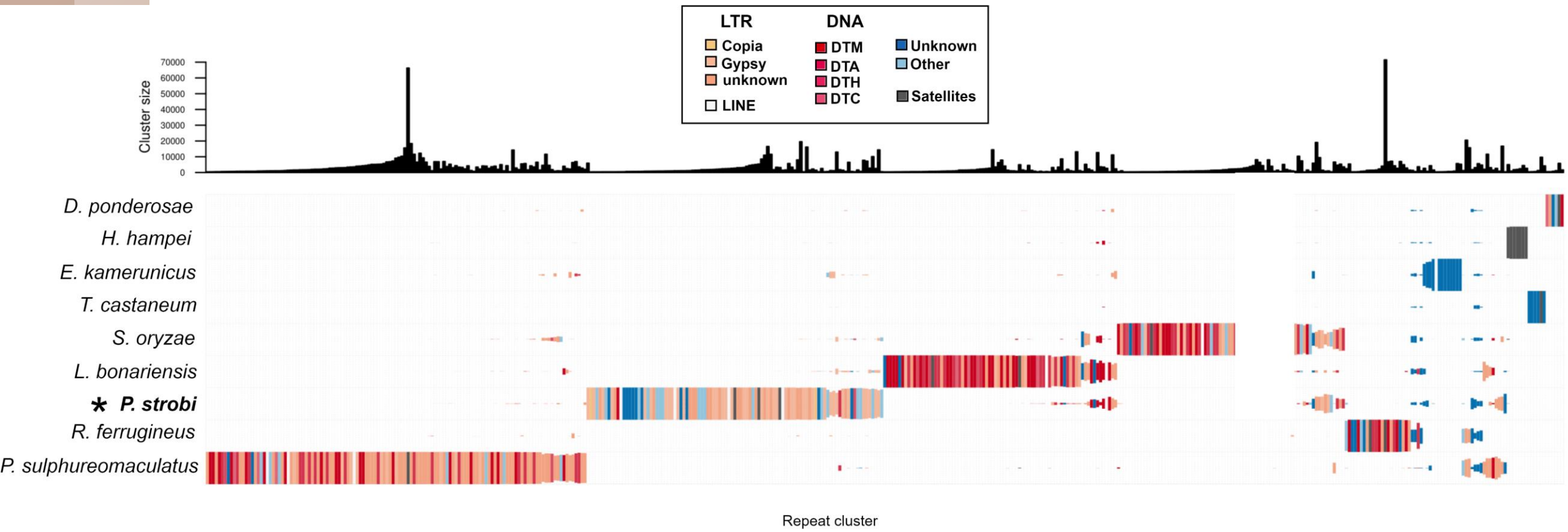

LTR
Copia
Gypsy
unknown
LINE
DNA
DTM
DTA
DTH
DTC
Unknown
Other
Satellites
Cluster size
70000
60000
50000
40000
30000
20000
10000
0
D. ponderosae
H. hampei
E. kamerunicus
T. castaneum
S. oryzae
L. bonariensis
* P. strobi
R. ferrugineus
P. sulphureomaculatus
Repeat cluster

Figure 3.3

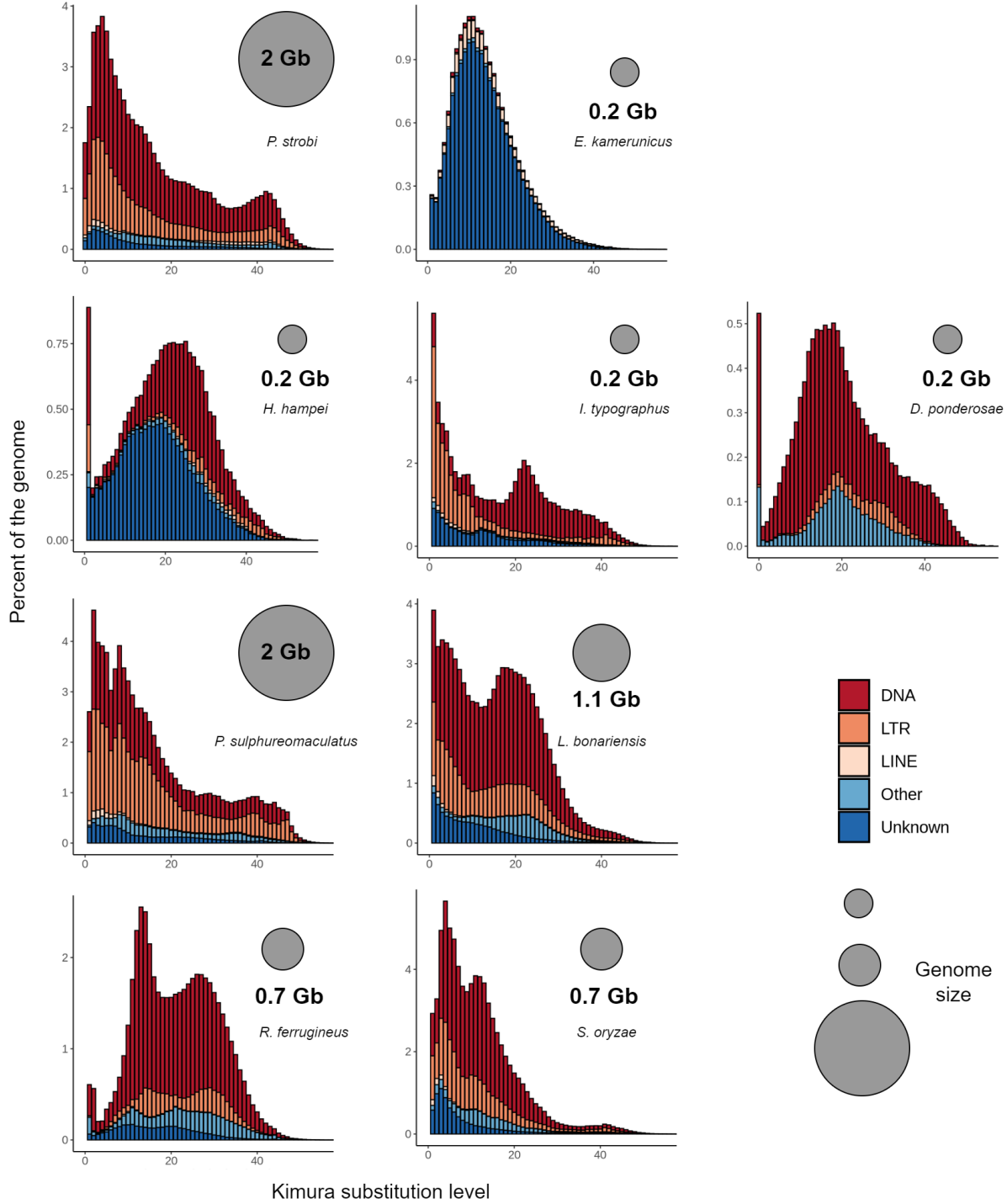

2 Gb
P. strobi
0.2 Gb
E. kamerunicus
0.2 Gb
H. hampei
0.2 Gb
I. typographus
0.2 Gb
D. ponderosae
2 Gb
P. sulphureomaculatus
1.1 Gb
L. bonariensis
0.7 Gb
R. ferrugineus
0.7 Gb
S. oryzae
Percent of the genome
Kimura substitution level
DNA
LTR
LINE
Other
Unknown
Genome size

Figure 3.4

	No. of scaffolds	Longest scaffold (kb)	NG50	Reconstructed size (Gb)	BUSCO single-copy (%)	BUSCO duplicated (%)
Supernova	163,521	2,374.58	79,498	2.23	66.7	12.2
Purge Haplotigs	82,994	2,374.58	79,343	1.83	71.0	8.2
Tigmint	84,653	2,139.77	74,900	1.83	70.7	8.6
ARKS	82,897	2,209.50	87,586	1.83	71.0	8.6
Sealer	**82,896**	**2,210.97**	**87,740**	**1.83**	**71.0**	**8.6**

Table 3.1

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

Table 3.2

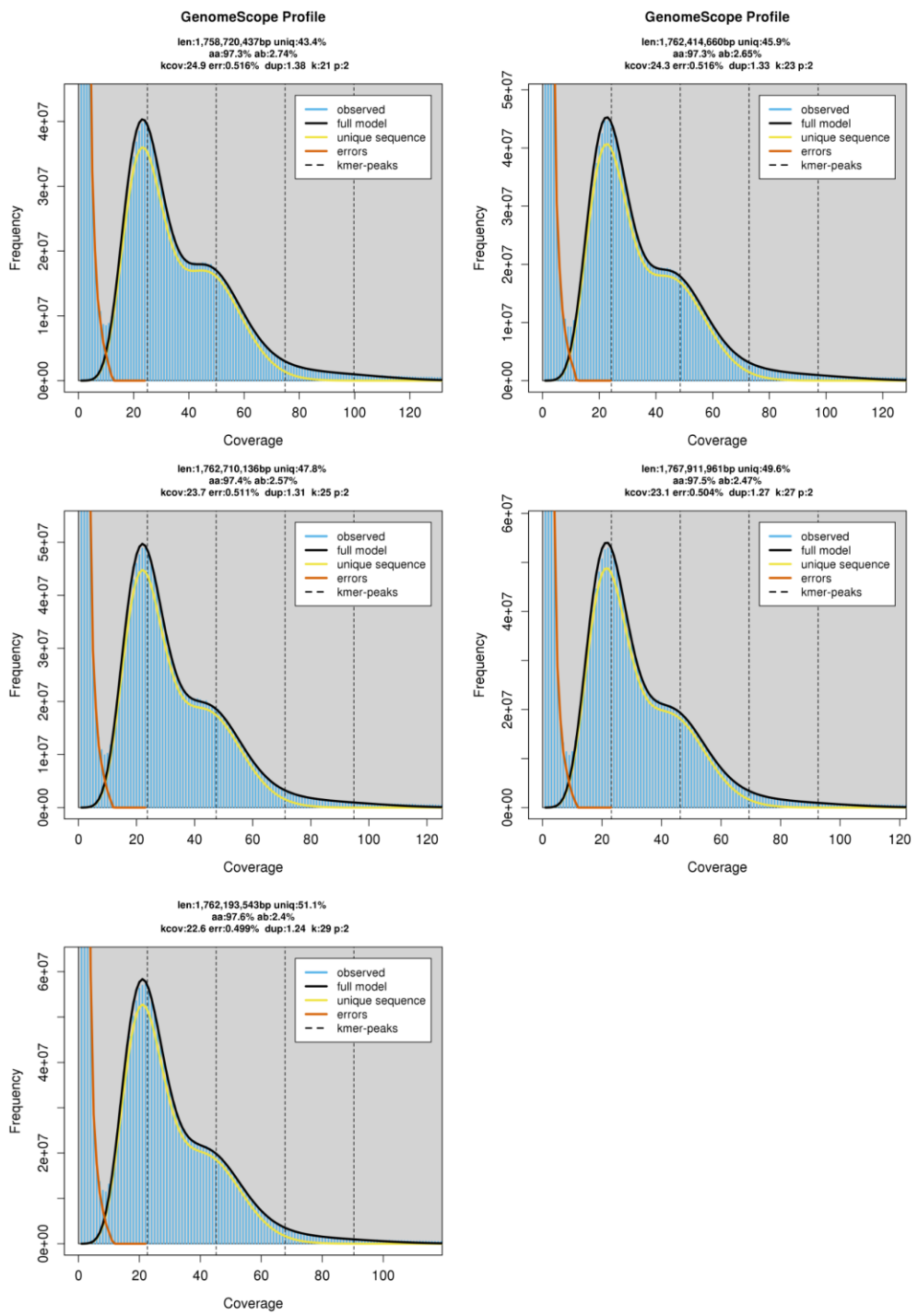
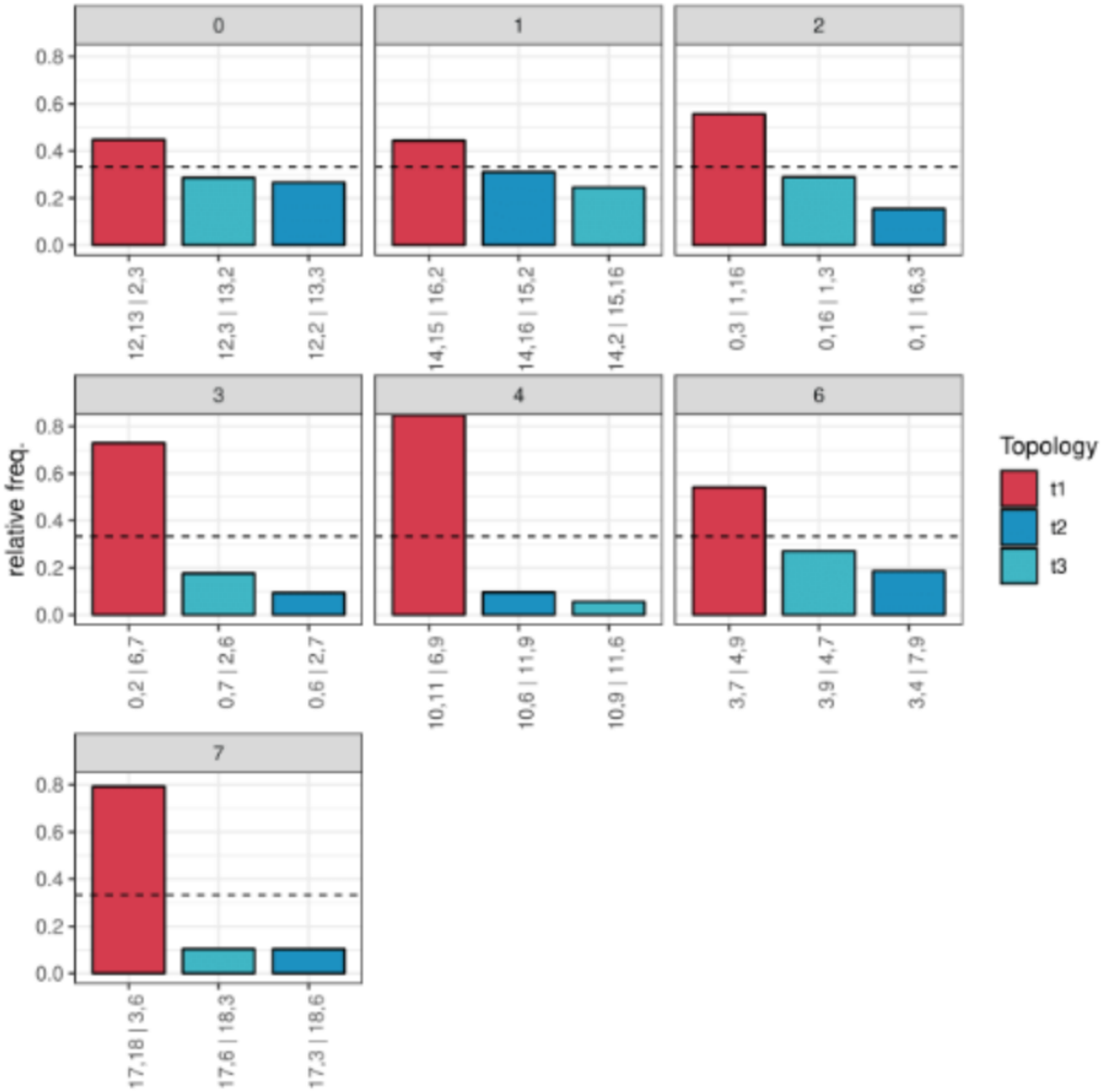

GenomeScope Profile
len:1,758,720,437bp uniq:43.4%
aa:97.3% ab:2.74%
kcov:24.9 err:0.516% dup:1.38 k:21 p:2
GenomeScope Profile
len:1,762,414,660bp uniq:45.9%
aa:97.3% ab:2.65%
kcov:24.3 err:0.516% dup:1.33 k:23 p:2
len:1,762,710,136bp uniq:47.8%
aa:97.4% ab:2.57%
kcov:23.7 err:0.511% dup:1.31 k:25 p:2
len:1,767,911,961bp uniq:49.6%
aa:97.5% ab:2.47%
kcov:23.1 err:0.504% dup:1.27 k:27 p:2
len:1,762,193,543bp uniq:51.1%
aa:97.6% ab:2.4%
kcov:22.6 err:0.499% dup:1.24 k:29 p:2
observed
full model
unique sequence
errors
kmer-peaks
Frequency
Coverage

Figure B.1


0
1
2
3
4
6
7
relative freq.
0.0
0.2
0.4
0.6
0.8
12,13 | 2,3
12,3 | 13,2
12,2 | 13,3
14,15 | 16,2
14,16 | 15,2
14,2 | 15,16
0,3 | 1,16
0,16 | 1,3
0,1 | 16,3
0,2 | 6,7
0,7 | 2,6
0,6 | 2,7
10,11 | 6,9
10,6 | 11,9
10,9 | 11,6
3,7 | 4,9
3,9 | 4,7
3,4 | 7,9
17,18 | 3,6
17,6 | 18,3
17,3 | 18,6
Topology
t1
t2
t3

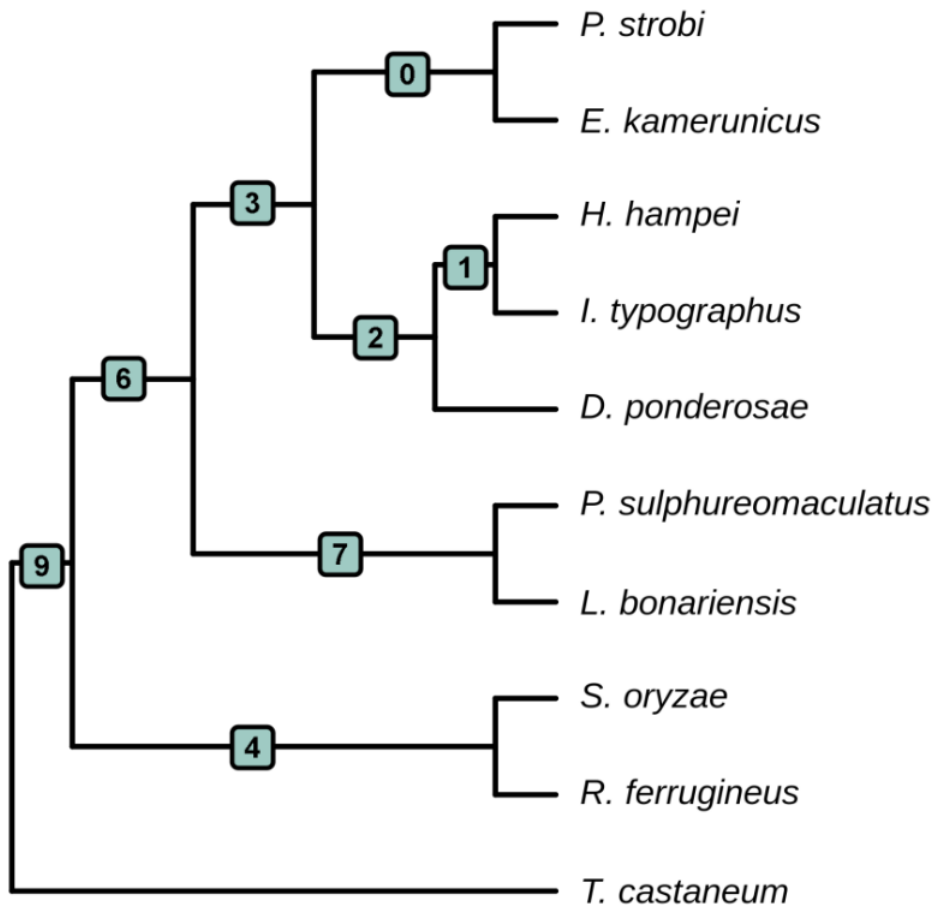

P. strobi
E. kamerunicus
H. hampei
I. typographus
D. ponderosae
P. sulphureomaculatus
L. bonariensis
S. oryzae
R. ferrugineus
T. castaneum
0
1
2
3
4
6
7
9

Figure B.2

Figure B.3

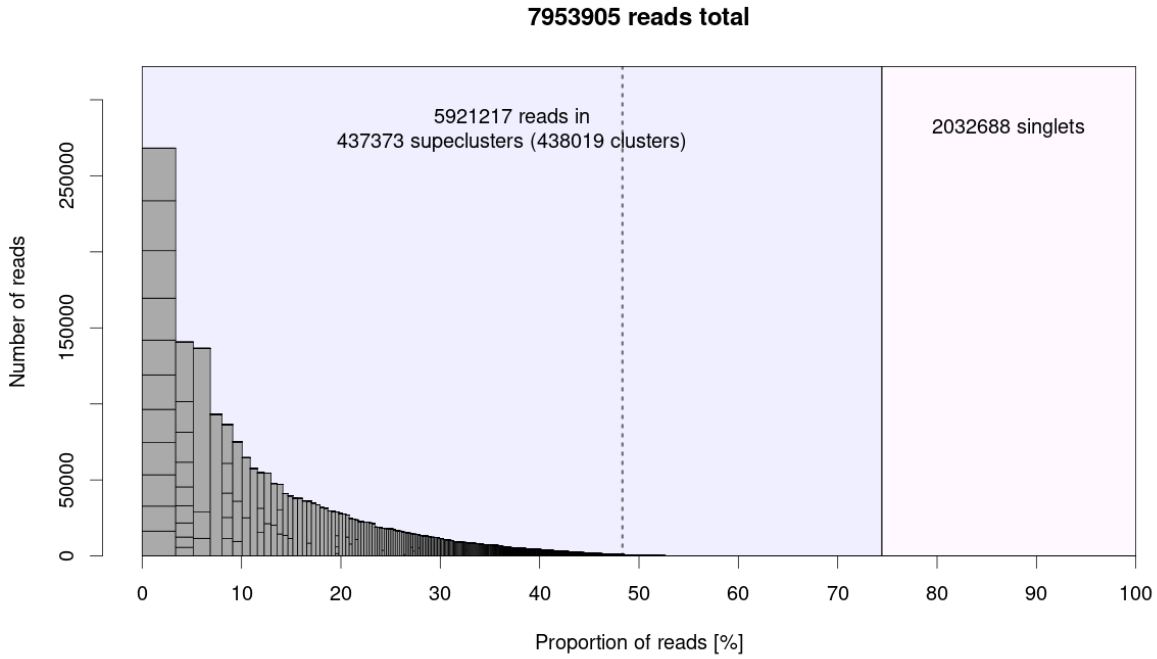
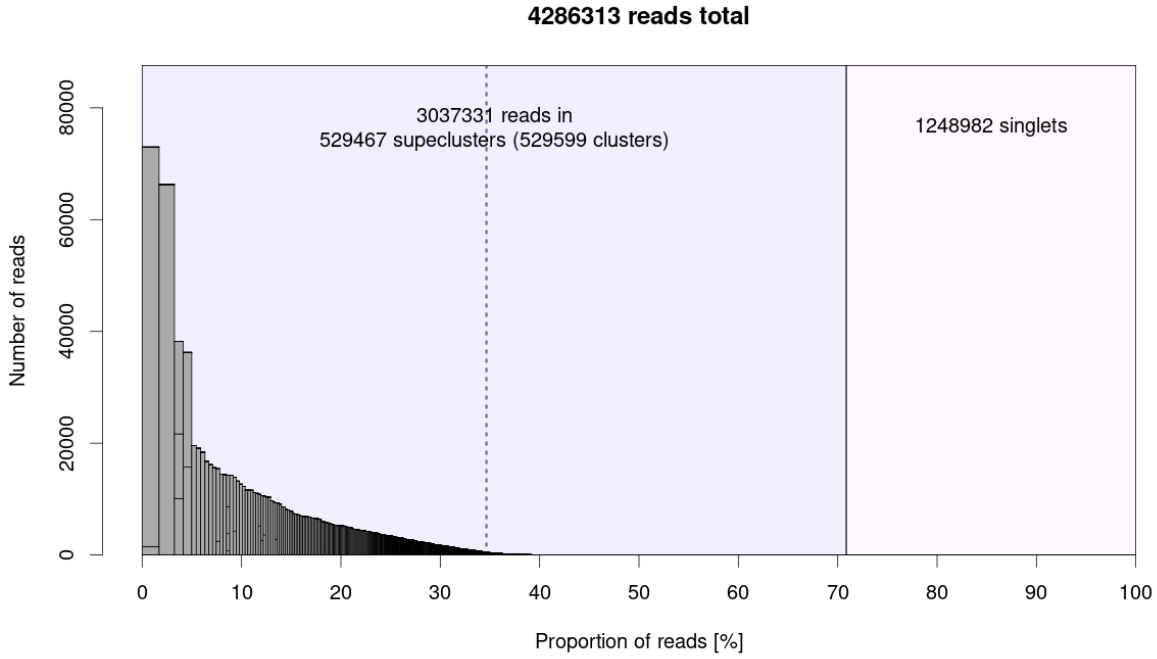


Figure B.4



Node	Concordant	Discordant
0	187	446
1	692	1171
2	869	1106
3	961	814
4	307	71
6	1001	1044
7	1060	445
9 - Out	2009	0

Table B.3

Class	Sub-class	1	2	3	4	5	6	7	8	9
DNA	DTA	15.24	0.00	1.09	0.31	3.52	10.08	11.52	1.77	17.39
	DTC	3.31	0.00	1.48	1.10	2.76	3.22	4.50	9.94	2.92
	DTH	1.06	0.00	0.41	0.34	0.38	0.67	2.37	1.37	3.08
	DTM	9.29	0.00	2.73	2.86	9.21	9.47	17.98	13.5	12.57
	DTT	0.41	0.00	0.20	0.02	0.47	0.40	0.41	0.41	0.70
	Helitron	2.40	0.00	1.64	0.46	0.71	4.03	4.35	3.75	4.74
LINE	LINE	1.31	1.06	0.62	0.21	1.24	1.47	0.79	0.01	0.86
LTR	Copia	1.05	0.00	0.00	0.03	0.76	0.00	0.23	0.22	0.20
	Gypsy	8.58	0.00	0.00	0.24	9.68	2.33	4.28	0.04	4.38
	Unknown	6.14	0.08	0.80	0.52	5.82	29.35	10.99	4.76	12.09
MITE	DTA	0.15	0.00	0.17	0.10	0.08	0.07	1.23	0.03	0.53
	DTC	0.08	0.00	0.04	0.04	0.03	0.03	0.18	0.03	0.10
	DTH	0.07	0.00	0.03	0.08	0.06	0.03	0.29	0.00	0.03
	DTM	0.55	0.00	0.14	0.50	0.60	0.18	0.84	0.12	0.54
	DTT	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.00	0.00
Other	Other	0.4	0.35	0.62	0.65	0.57	0.98	0.62	0.28	0.57
Unknown	Unknown	3.44	9.83	4.54	8.82	7.68	5.45	6.80	2.69	6.92
Mixture	Mixture	0.05	0.00	0.00	0.00	0.00	0.06	0.02	0.00	0.00
Total		**53.53**	**11.32**	**14.51**	**16.29**	**43.57**	**67.83**	**67.42**	**38.92**	**67.62**

1) Pissodes strobi
2) Elaeidobius kamerunicus
3) Dendroctonus ponderosae
4) Hypothenemus hampei
5) Ips typographus
6) Pachyrhynchus sulphureomaculatus
7) Listronotus bonariensis
8) Rhynchophorus ferrugineus
9) Sitophilus oryzae

Table B.4

	Read pairs	Total read counts	Total clusters	DNA clusters	LTR clusters
Pissodes strobi	**3,750,000**	**1,144,238**	**204**	**39**	**106**
Elaeidobius kamerunicus	750,000	228,048	62	8	26
Dendroctonus ponderosae	375,000	114,616	42	6	23
Hypothenemus hampei	275,625	83,410	46	11	16
Pachyrhynchus sulphureomaculatus	3,750,000	1,144,092	167	50	90
Listronotus bonariensis	2,062,500	627,190	154	61	64
Rhynchophorus ferrugineus	1,387,500	423,634	52	13	29
Sitophilus oryzae	1,419,375	434,052	151	60	62
Tribolium castaneum	285,000	87,032	20	1	7

Table B.5

	Read pairs	Total read counts	Total clusters	DNA clusters	LTR clusters
Pissodes strobi	**3,750,000**	**1,144,238**	**204**	**39**	**106**
Elaeidobius kamerunicus	750,000	228,048	62	8	26
Dendroctonus ponderosae	375,000	114,616	42	6	23
Hypothenemus hampei	275,625	83,410	46	11	16
Pachyrhynchus sulphureomaculatus	3,750,000	1,144,092	167	50	90
Listronotus bonariensis	2,062,500	627,190	154	61	64
Rhynchophorus ferrugineus	1,387,500	423,634	52	13	29
Sitophilus oryzae	1,419,375	434,052	151	60	62
Tribolium castaneum	285,000	87,032	20	1	7

Table B.5

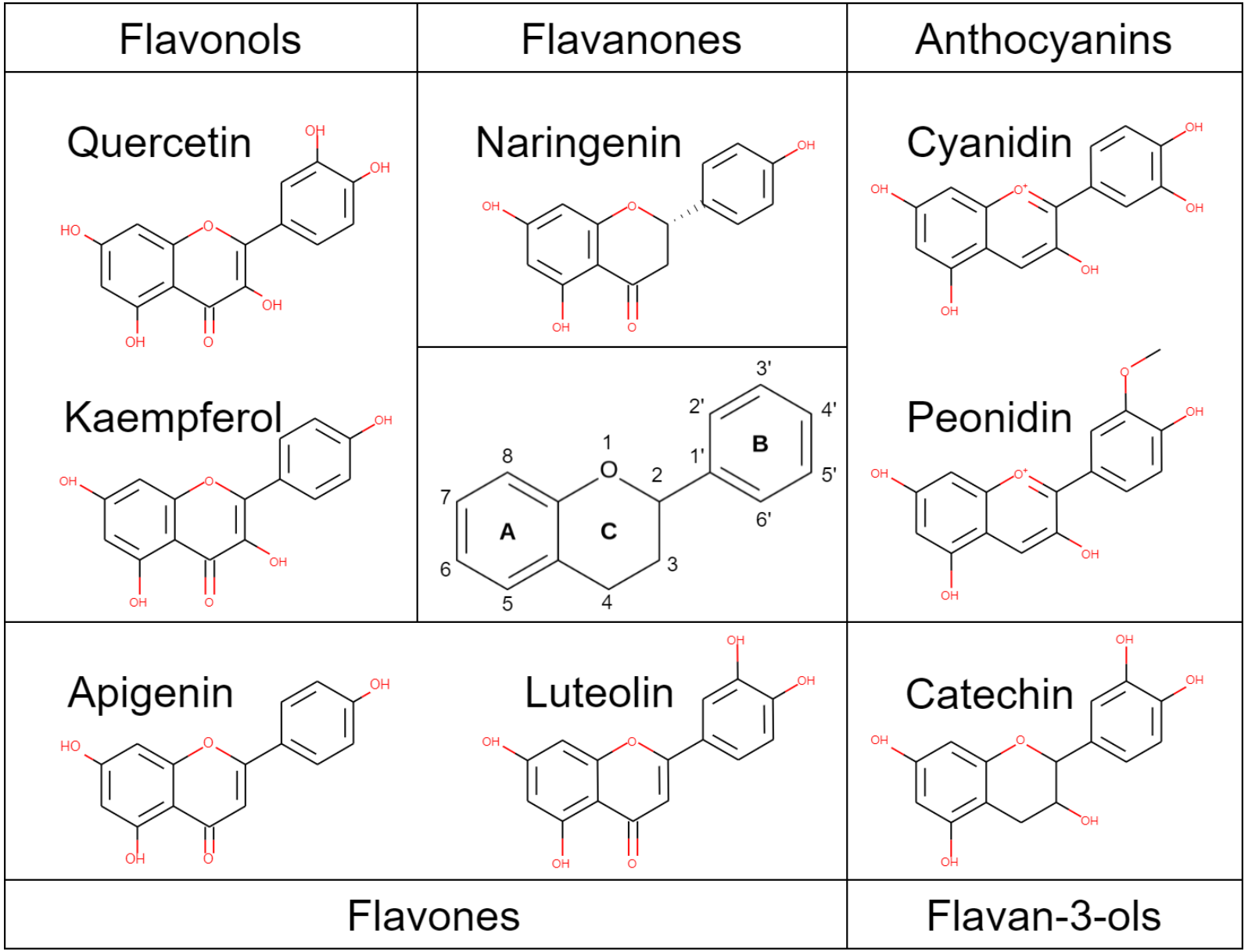

Flavonols
Flavanones
Anthocyanins
Quercetin
Naringenin
Cyanidin
Kaempferol
Peonidin
Apigenin
Luteolin
Catechin
Flavones
Flavan-3-ols

Figure 4.1

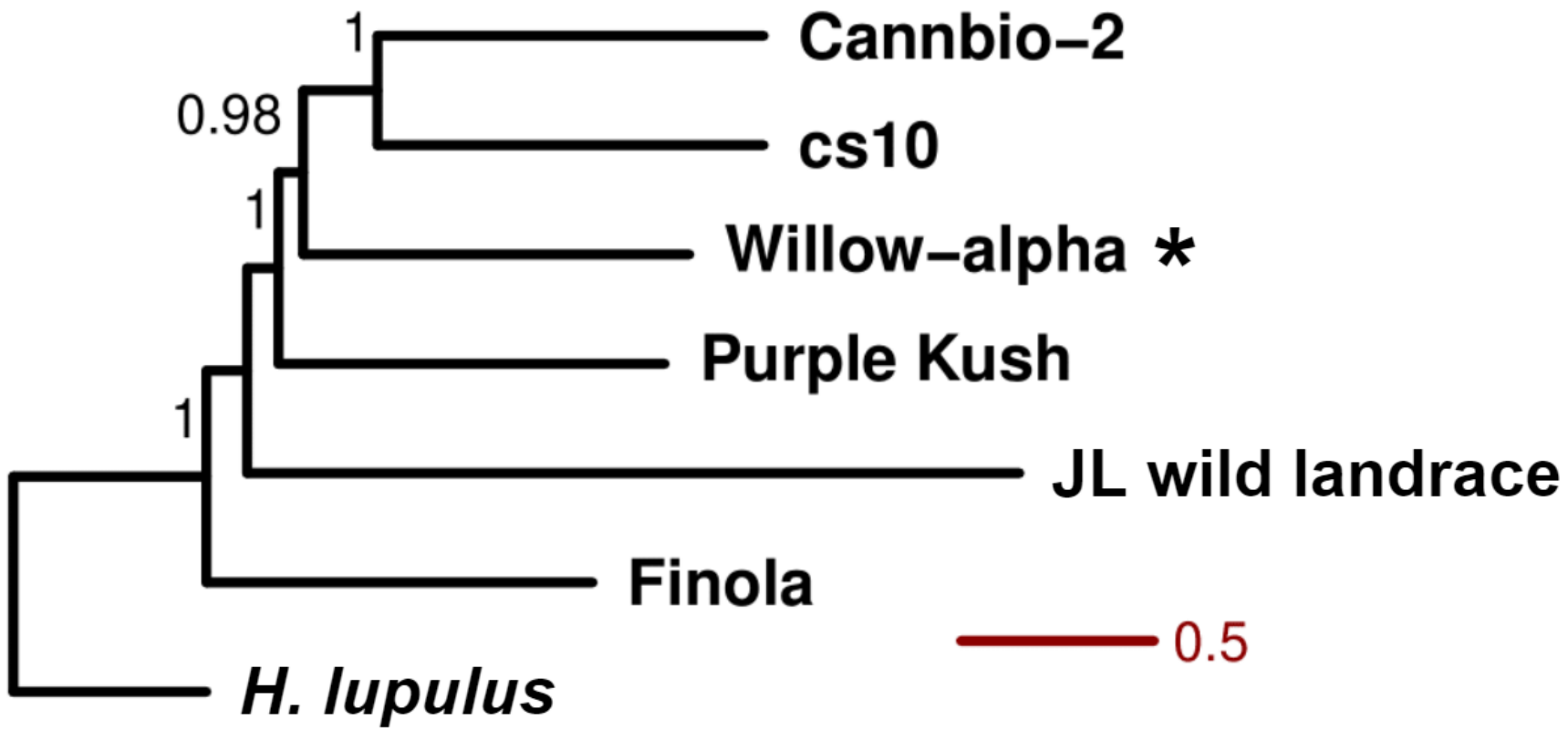

1
0.98
1
1
Cannbio–2
cs10
Willow–alpha *
Purple Kush
JL wild landrace
Finola
H. lupulus
0.5

Figure 4.3

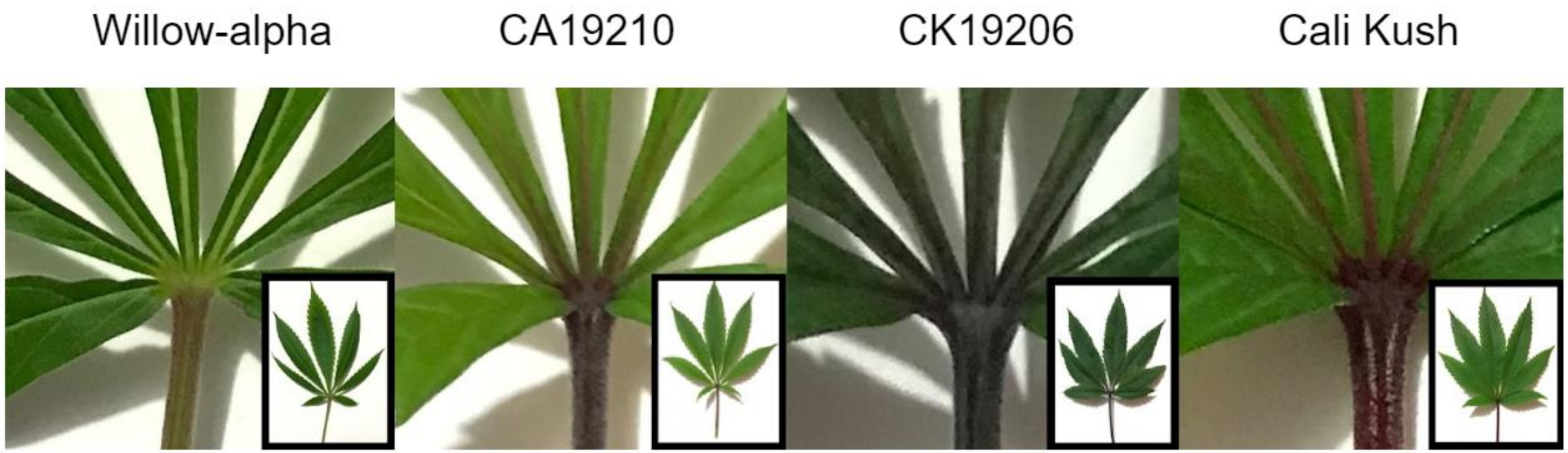


Figure 4.4

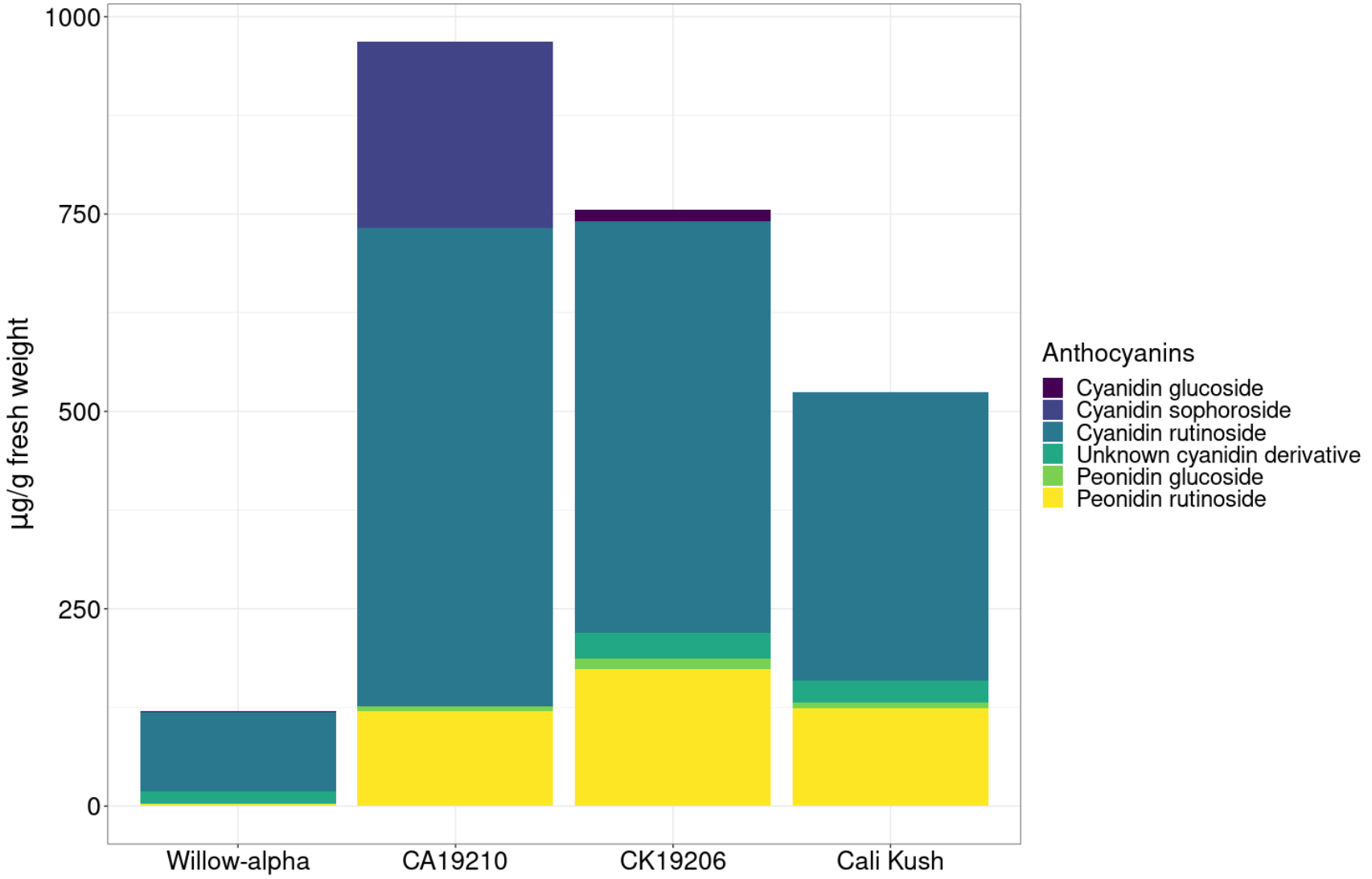

1000
750
500
250
0
µg/g fresh weight
Willow-alpha
CA19210
CK19206
Cali Kush
Anthocyanins
Cyanidin glucoside
Cyanidin sophoroside
Cyanidin rutinoside
Unknown cyanidin derivative
Peonidin glucoside
Peonidin rutinoside

Figure 4.7

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

Table 4.1

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

Table 4.2

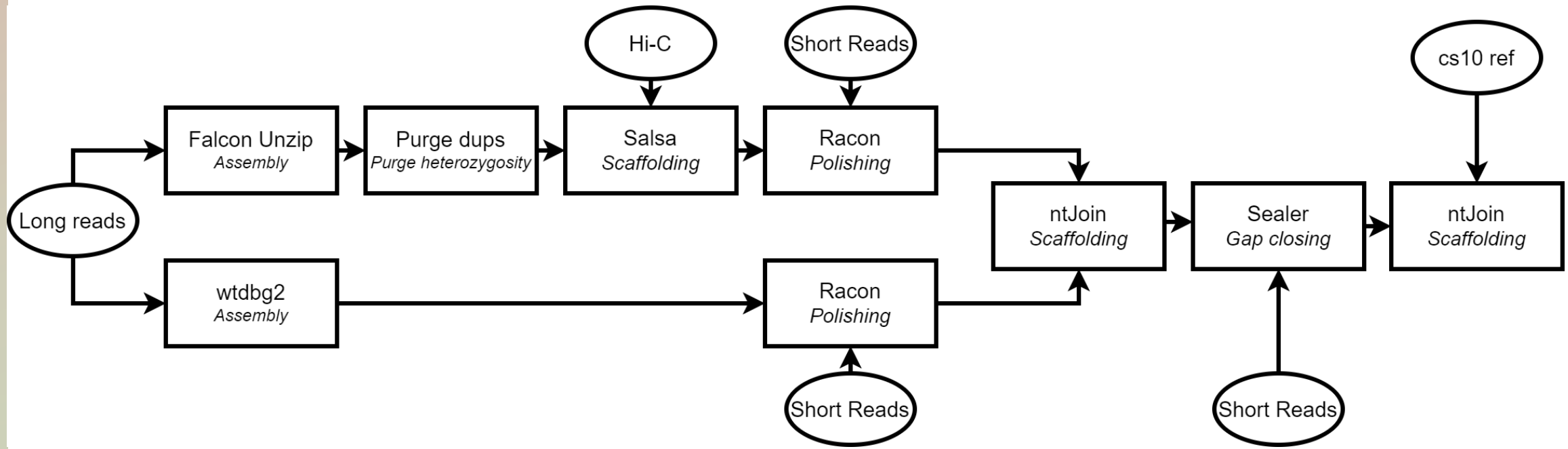

Hi-C
Short Reads
cs10 ref
Falcon Unzip
Assembly
Purge dups
Purge heterozygosity
Salsa
Scaffolding
Racon
Polishing
Long reads
ntJoin
Scaffolding
Sealer
Gap closing
ntJoin
Scaffolding
wtdbg2
Assembly
Racon
Polishing
Short Reads
Short Reads

Figure C.1

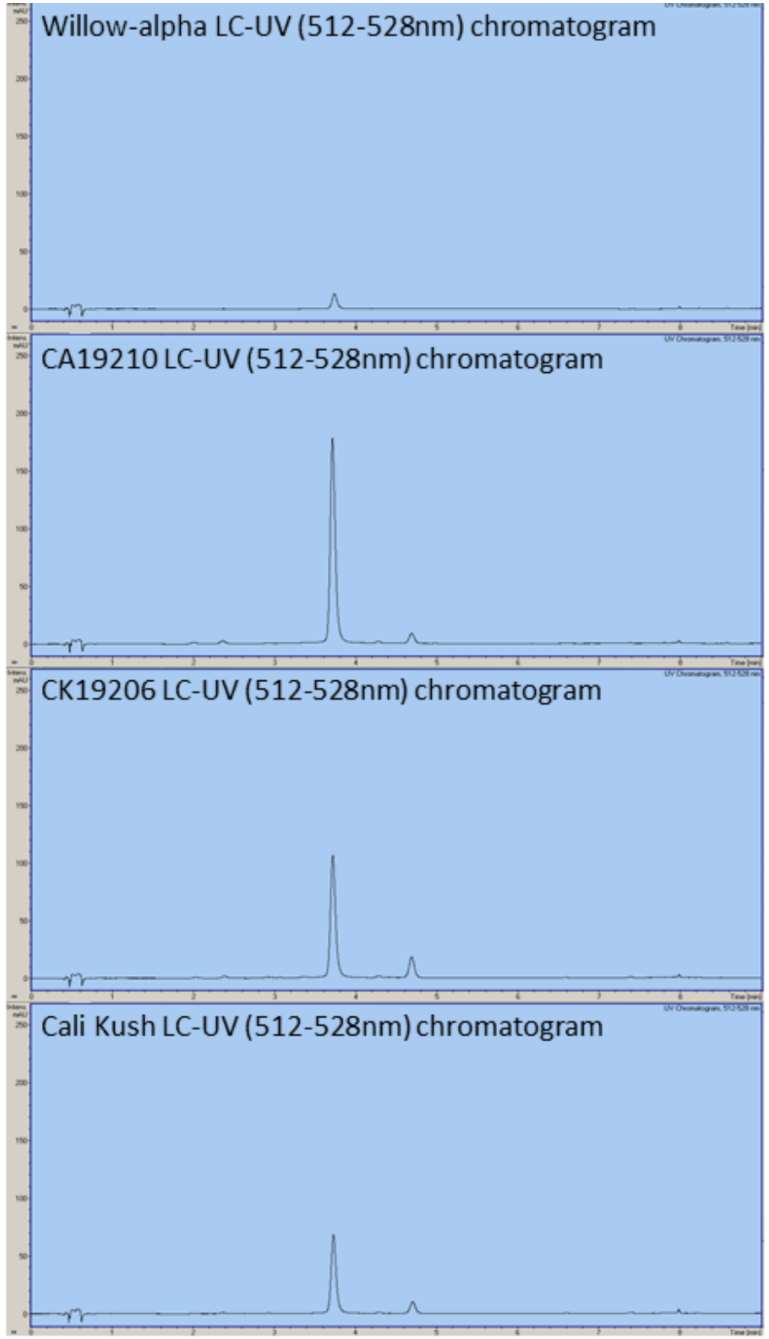

Willow-alpha LC-UV (512-528nm) chromatogram
CA19210 LC-UV (512-528nm) chromatogram
CK19206 LC-UV (512-528nm) chromatogram
Cali Kush LC-UV (512-528nm) chromatogram

Figure C.2

Name	1) Willow-alpha - CA19210	2) Willow-alpha - CK19206	3) Willow-alpha - Cali Kush
PAL1	log2FC -1.133; padj 8.28E-03	log2FC -1.404; pdj 7.27E-04	log2FC -0.777; padj 8.87E-02
C4H	log2FC -0.455: padj 6.99E-02	log2FC -1.421; padj 1.60E-10	log2FC -0.623; padj 1.01E-02
4CL	log2FC -0.127; padj 2.61E-01	log2FC -0.171; padj 1.04E-01	log2FC -0.656; padj 1.03E-12
CHS	**log2FC -1.969; pagj 1.01E-04**	**log2FC -2.717; padj 3.41E-08**	log2FC -0.966; padj 8.40E-02
CHI	log2FC -1.217; padj 1.76E-04	log2FC -1.077; padj 9.68E-04	log2FC -0.263; padj 5.21E-01
CHI-L1	log2FC 1.013; padj 4.46E-03	log2FC 0.582; padj 1.23E-01	log2FC 0.260; padj 5.59E-01
F3H	**log2FC -4.433; padj 1.98E-19**	**log2FC -4.428; padj 2.23E-19**	**log2FC -2.470; padj 1.48E-06**
F3'H	**log2FC -2.526; padj 3.16E-06**	**log2FC -3.304; padj 4.90E-10**	**log2FC -1.635; padj 4.26E-03**
FLS1	**log2FC -2.523; padj 4.98E-05**	**log2FC -3.819; pdj 4.31E-11**	**log2FC -2.021; padj 2.42E-03**
UGT73C6	log2FC 1.556; padj 2.35E-01	log2FC 0.480; padj 7.42E-01	log2FC 2.524; padj 5.24E-02
DFR	**log2FC -1.987; padj 2.79E-05**	**log2FC -2.703; padj 5.09E-09**	**log2FC -1.587; padj 1.19E-03**
LDOX/ANS	**log2FC -2.337; padj 2.24E-05**	**log2FC -3.101; padj 8.59E-09**	log2FC -0.988; padj 1.11E-01
UGT78D2	**log2FC -2.329; padj 7.28E-04**	**log2FC -3.524; padj 1.06E-07**	**log2FC -2.049; padj 3.73E-03**
UGT76C1	log2FC 0.759; padj 7.81E-03	log2FC -0.457; padj 1.28E-01	log2FC 0.724; padj 1.25E-02
UGT79B3	log2FC 1.416; padj 1.35E-01	log2FC 0.659; padj 4.90E-01	log2FC 2.466; padj 2.11E-02
ANR	-	-	-
LAC15	log2FC 2.994; padj 5.55E-02	log2FC 0.331; pdj 7.78E-01	log2FC -0.372; padj 7.48E-01

Figure C.3

		Willow-alpha		CA19210		CK19206		Cali Kush	
	Anthocyanin	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev
AN1	Cyanidin glucoside	1.6	0.6	0.0	0.0	13.9	4.6	0.0	0.0
AN2	Cyanidin sophoroside	0.0	0.0	235.9	30.4	0.0	0.0	0.0	0.0
AN3	Cyanidin rutinoside	100.4	20.7	606.1	329.8	521.2	135.7	365.2	81.3
AN4	Cyanidin derivative (cyanidin glucoside glucuronide)	15.9	6.4	0.0	0.0	33.2	12.6	27.6	2.4
AN5	Peonidin glucoside	0.0	0.0	5.8	1.4	13.4	4.5	7.9	2.2
AN6	Peonidin rutinoside	2.9	1.4	120.2	37.1	173.3	60.5	123.8	28.8

Figure C.4

		Willow-alpha		CA19210		CK19206		Cali Kush	
	Flavonoid*	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev	Avg.	Stdev
FL1	Quercetin glucoside	0.0	0.0	612.8	121.1	0.0	0.0	509.6	48.2
FL2	Quercetin glucuronide	65.6	8.5	0.0	0.0	367.5	54.8	159.0	24.0
FL3	Rutin	0.0	0.0	451.1	25.0	3.7	1.4	298.8	28.0
FL4	Kaempferol glucuronide	42.7	7.5	3.7	0.2	0.0	0.0	27.5	1.9
FL5	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	0.0	0.0	364.1	22.9	0.0	0.0	285.8	19.6
FL6	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	0.0	0.0	328.7	26.0	0.0	0.0	0.0	0.0
FL7	Vitexin 3-O-glucoside Vitexin 7-O-glucoside Isovitexin-O-glucoside	1036.3	134.6	0.0	0.0	1801.1	138.7	979.5	92.4
FL8	Apigenin 7-O-glucuronide	789.7	109.2	0.0	0.0	886.5	129.5	264.8	39.6
FL9	Luteolin 7-O-glucuronide	61.9	8.8	36.6	4.5	4.8	217.1	217.1	0.5

Figure C.5

Compound	Identified MS1 (Trap)	Identified MS2 (Trap)	RT (Trap)	Identified MS1 (QTOF)	Identified MS2 (QTOF)	RT (QTOF)	Published exact mass
Cyanidin glucoside	449	287	3.3	449.1085	287.0570	5.938	449.1084
Cyanidin rutinoside	595	287	3.6	595.1662	287.0548	6.393	595.1662
Peonidin glucoside	463	301	4.3	463.1240	301.0701	7.418	463.1240
Cyanidin sophoroside	611	287	4.5	611.1599	449.1080	7.890	611.1612
Peonidin rutinoside	609	301	4.6	609.1803	301.0683	7.646	609.1819
Cyanidin glucuronide	463	287	4.8	463.0874	287.0567	9.239	463.0875
Quercetin glucoside	465	303	4.4	465.1038	303.0507	7.369	465.1033
Quercetin glucuronide	479	303	6.9	479.0835	303.0514	11.012	479.0825
Rutin	611	303	5.2	611.1601	303.0504	8.768	611.1612
Kaempferol glucuronide	463	287	5.6	463.0868	287.0564	9.239	463.0875
Vitexin glucoside/Isovitexin glucoside	595	433	4.7	433.1105	595.1671	8.020	595.1662
Vitexin glucoside/Isovitexin glucoside	595	433	5.0	595.1659	433.1125	8.475	595.1662
Vitexin glucoside/Isovitexin glucoside	595	433	5.5	595.1668	433.1120	8.751	595.1662
Apigenin 7-O-glucuronide	447	271	6.5	447.0934	271.0626	10.427	447.0926
Luteolin 7-O-glucuronide	463	287	8.0	463.0870	287.0556	12.622	463.0872

Figure C.6