Identifying Mitochondrial Genomes in Draft Whole-Genome Shotgun Assemblies of Six Gymnosperm Species

Yrin Eldfjell

Bachelor's Thesis in Computer Science
at Stockholm University, Sweden, 2018

Stockholm University

Identifying Mitochondrial Genomes in Draft Whole-Genome Shotgun Assemblies of Six Gymnosperm Species

Yrin Eldfjell

Bachelor's Thesis in Computer Science (15 ECTS credits)
Single Subject Course
Stockholm University year 2018
Supervisor at the Department of Mathematics was Lars Arvestad
Examiner was Jens Lagergren, KTH EECS

Department of Mathematics
Stockholm University
SE-106 91 Stockholm, Sweden

Identifying Mitochondrial Genomes in Draft Whole-Genome Shotgun Assemblies of Six Gymnosperm Species

Abstract

Sequencing efforts for gymnosperm genomes typically focus on nuclear and chloroplast DNA, with only three complete mitochondrial genomes published as of 2017. The availability of additional mitochondrial genomes would aid biological and evolutionary understanding of gymnosperms. Identifying mtDNA from existing whole genome sequencing (WGS) data (i.e. contigs) negates the need for additional experimental work but previous classification methods show limitations in sensitivity or accuracy, particularly in difficult cases. In this thesis I present a classification pipeline based on (1) kmer probability scoring and (2) SVM classification applied to the available contigs. Using this pipeline the mitochondrial genomes of six gymnosperm specias were obtained: *Abies sibirica, Gnetum gnemon, Juniperus communis, Picea abies, Pinus sylvestris* and *Taxus baccata*. Cross-validation experiments showed a satisfying and for some species excellent degree of accuracy.

Identifiering av mitokondriers arvsmassa från preliminära versioner av arvsmassan för sex gymnospermer

Sammanfattning

Vid sekvensering av gymnospermers arvsmassa har fokus oftast lagts på kärn- och kloroplast-DNA. Bara tre fullständigt mitokondriegenom har publicerats hittills (2017). Fler mitokondriegenom skulle kunna leda till nya kunskaper om gymnospermers biologi och evolution. Då mitokondriernas arvsmassa identifieras från tillgängliga sekvenser för hela organismen (så kallade "contiger") behövs inget ytterligare laboratoriearbete, men detta förfarande har visat sig leda till bristfällig känslighet och korrekthet, särskilt i svåra fall. I denna avhandling presenterar jag en metod baserad på (1) kmer-sannolikheter och (2) SVM-klassificering applicerad på de tillgängliga contigerna. Med denna metod togs arvsmassan för mitokondrien hos sex gymnospermer fram: *Abies sibirica, Gnetum gnemon, Juniperus communis, Picea abies, Pinus sylvestris* och *Taxus baccata*. Korsvalideringsexperiment visade en tillfredställande och för vissa arter utmärkt precision.

Acknowledgements

Thanks to Lars Arvestad for advice, directions and good discussions, Anastasia Atucha for a fun collaboration during the pilot project, Mattias Frånberg and Kristoffer Sahlin for discussions about probability models, Lukas Käll for advice, Jens Lagergren and Caroline Nordquist for generic patience, Fabian Nordenskjöld for advice about what plastid reference genomes to use, Douglas Scofield for an interesting discussion on conifer mitochondrial genomes, Jonas Nørskov Søndergaard for substantial abstract advice and Marcel Tarbier for concrete abstract advice.

Any inaccuracies remaining in the report is fully the responsibility of the author.

I would also like to thank the many people[1] at the SciLifeLab gamma-4 and gamma-6 floors that have made my time during this project much more endurable, interesting and rewarding.

The majority of computations for this project were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). GNU Parallel was used for several computational steps (Tange 2011). Matplotlib was used to generate many of the plots (Hunter 2007). Finally, the project has generated much evidence in favour of the equation

$$\lim_{t \to \infty} P_{LA} > 0,$$

where t denotes time and P_{LA} denotes the patience of my advisor Lars Arvestad. I much appreciate that.

[1] Too many to name, really. If you doubt whether you should feel included in this list, I'd advice you to err on the side of inclusion. :-)

Contents

Glossary

assembler Software that takes **reads** and *assembles* them into longer **contigs** using overlapping subsequences.

base-pair, bp A (Comp. sci.) *letter* in a (typically) four-letter (Comp. sci.) *alphabet*. [syn: nucleotide (nt), base]

CG Can refer to nucleotides being *either* C (cytosine) or G (guanine), or to C-G base-pairs (i.e. a C on one DNA strand an a G in the corresponding position of the other).

chloroplast Organelle (cellular subunit) where *photosynthesis* occurs.

contig A (longer) DNA sequence **assembled** from (typically) multiple reads.

coverage Sequence **coverage** referens to the number of reads that "cover" (align to) a given sequence position.

FDR False discovery rate, i.e. the proportion of false positives among all samples classified positive.

feature (Machine learning.) A measurable property of an object. A great feature has a high degree of independence vs. other features (which makes the model simpler) and the values it takes on are highly dependent on the target class of the object.

feature-space plot A plot showing the coordinates for a set of contigs in the 2D plane formed by a pair of features. For more details, see section 6.1.

kmer Subsequence of fixed length k, a generalization of the *-mer* concept (with e.g. a *trimer* having k=3). [syn: k-mer]

mitochondrion Organelle (cellular subunit) performing key functions in oxygen-dependent energy metabolism.

N Nucleotide wild-card used to denote "any of" ACGT (DNA) or ACGU (RNA). Example: In the sequence `ATNNT`, the two N's are unknown.

nuclear DNA The main genetic material of a cell, as opposed to the smaller amounts of DNA contained in the **organelles**.

read A short sequence of DNA from a sequencing machine.

sequence (Comp. sci.) *string*.

SVM Support vector machine, a supervised machine learning method where labeled examples are used to train a classifier to achieve the widest possible separation of two (or more) classes of objects in space.

WGS Whole genome sequencing, the process of reading (sequencing) the entire genome of an organism, including organellar DNA.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Plant cells typically contain two types of organelles: *mitochondria* and *plastids*. The most well-known type of plastid is the *chloroplast*.[1]

Obtaining both organellar genomes is of interest for at least three reasons: (1) it helps gaining more knowledge about the organelles themselves, (2) the gender-specific inheritance patterns of organellar DNA allow insights into species evolution and (3) it makes it possible to "clean" the nuclear genome of organellar contamination.

As of 2017, only three complete gymnosperm mito-genomes have been published[2], whereas there are multiple complete gymnosperm chloroplast genomes available.

In the effort to sequence, assemble and annotate the gymnosperm Norway Spruce (*Picea abies*), a large amount of low coverage whole genome shotgun (WGS) data was generated. This includes low-coverage WGS data of five other gymnosperms (*Pinus sylvestris, Abies sibirica, Juniperus communis, Taxus baccata* and *Gnetum gnemon*) intended for comparative analysis (Nystedt et al. 2013). The chloroplast genomes of these six species were assembled[3], while the mitochondrial assemblies were never completed (Lars Arvestad 2017, personal communication, 21 March).

1.2 Project scope

In the summer of 2014 Lars Arvestad tasked fellow student Anastasia Atucha and myself with a small project to produce the assemblies of the mitochondrial genomes of the six[4] gymnosperms. That was effectively a pilot project for this one.

After some time this project became my thesis project. The assembly step was then removed as early experiments indicated this would be too time consuming.[5]

[1] Since all plastids have the same genome (Cooper 2000), the terms *chloroplast* and *plastid* will be used interchangably in this thesis.

[2] *Cycas taitungensis*, *Ginkgo biloba* and *Welwitschia mirabilis*

[3] The completed assemblies are unpublished as of March 2017 but made available to me for this project.

[4] At the time, we only worked with five of the species.

[5] Plant mitochondrial genomes are large and can have complex, repeat-heavy physical structures. They generally have a low gene density, all this making them potentially difficult to assemble (Gualberto et al. 2014, p. 107).

The main objective was now to correctly identify as many mitochondrial contigs as possible.

1.3 The key challenge and choice of strategy

Generally, the total size of all contigs (at least 500 bp long) for a species measured in the low *Gbp* range while the extracted mitochondrial contigs were found to be in the low *Mbp* range (see table 2.1). This meant that even a good classifier picking only one false positive per every hundred or so contigs would swamp the delivered "mitochondrial" contigs with false positives.

We found three main ways of addressing this problem:

1. discard all short contigs and only pick the long and "obvious" ones, or
2. rely heavily on alignments to related species for classification, or
3. combine several (sequence level) classifiers that together achieve the necessary separation.

We opted for strategy (3) as we wanted to try to extract as large fractions of the mitochondrial genomes as possible.

1.4 This report

In broad terms, this report aims to answer the following questions (in this order):

1. What was the goal of the project?
2. What data was available?
3. How had earlier projects used this type of data?
4. How did we use the data and why?
5. What were our findings?
6. Did our method of analysis work well?
7. What conclusions can be drawn?

Chapter 2

Input data

Two groups of genomic sequences were used as input data for the project: (1) the six large sets of contigs from which the mitochondrial contigs were to be extracted (one for each species) and (2) assembled organellar (including nuclear) genomes of other plants used as reference.

2.1 WGS assembly contigs for the studied species

For P. abies the published genome was used (Umeå Plant Science Centre 2013).[1]

For the other five species, sets of contigs assembled by the CLC bio assembler (Nystedt et al. 2013) — currently unpublished — were made available to me.

These source contigs have been through various types of filtering. The full details about this process are unknown to me, but is believed to include contaminant screening (Lars Arvestad 2017, by email 10 feb). All contigs had been screened specificially for chloroplast contamination. However, it became apparent during the project that some low quality, low coverage contigs had slipped through this filter. These remaining chloroplast contigs were a challenge to classify correctly using our pipeline, as we had lost the distinctively high coverage as a signal.

Raw reads for all six assemblies have been deposited at the European Nucleotide Archive. For accession numbers, see Nystedt et al. 2013, supp. materials section 6.1. See table 2.1 for some basic statistics about the reads and assemblies.

For the remainder of this report, these unclassified input contigs will be refered to as the *source contigs*.

2.2 Reference genomes

A number of reference genomes have been used for classifier training and as targets when trying to identify organelle-specific contigs in our unlabeled data. See table 2.2 for a complete list. Included are the plastid genomes of all six studied species, enabling us to remove plastid "look-alikes" from contigs that were classified as likely mitochondrial.

[1] Technically, an internal project file (that *should* be the same version as the one referenced) was used.

Table 2.1: *Statistics about the source contigs.*

Statistics are from the project wiki (Talavera-López 2014) and from direct computations on the filtered source contigs. See also Nystedt et al. 2013.

Statistic	A. sibirica	G. gnemon	J. comm.	P. abies	P. sylv.	T. baccata
#Contigs $\geq$ 500 bp	1.2 M	926 k	861 k	3.2 M	3.2 M	1.7 M
Size of contigs $\geq$ 500 bp	1.2 Gbp	1.5 Gbp	736 Mbp	10 Gbp	3.1 Gbp	1.7 Gbp
#Contigs $\geq$ 500 bp and with cov $\geq$ 100	6.5 k	1.1 k	1.9 k	49.5 k	5.1 k	0.7 k
Size of contigs $\geq$ 500 bp and with cov $\geq$ 100	7.9 Mbp	2.0 Mbp	3.0 Mbp	132 Mbp	5.6 Mbp	1.1 Mbp
Est. whole genome cov.	3.8	5.5	4.5	(very high)	12.5	4.0

Table 2.2: *Reference genomes used.*

Species	Organelle	Accession	Reference
Abies sibirica	Plastid	N/A	(Lars Arvestad 2015, by email 7 sept)
Amborella trichopoda	Mito.	KF754803.1 KF754802.1 KF754801.1 KF754800.1 KF754799.1	Rice et al. 2013
	Plastid	NC_005086.1	Goremykin et al. 2003
Arabidopsis thaliana	Mito.	NC_001284.2	Unseld et al. 1997
	Plastid	NC_000932.1	Sato et al. 1999
	Nuclear	NC_003070.9	Theologis et al. 2000
		NC_003071.7	Lin et al. 1999
		NC_003074.8	Salanoubat et al. 2000
		NC_003075.7	Mayer et al. 1999
		NC_003076.8	Tabata et al. 2000
Cycas taitungensis	Mito.	NC_010303.1	Chaw et al. 2008
	Plastid	NC_009618.1	C.-S. Wu et al. 2007
Gnetum gnemon	Plastid	N/A	(Lars Arvestad 2015, by email 7 sept)
Juniperus communis	Plastid	N/A	(Lars Arvestad 2015, by email 7 sept)
Picea abies	Plastid	NC_021456.1	Nystedt et al. 2013
Pinus sylvestris	Plastid	N/A	(Lars Arvestad 2015, by email 7 sept)
Populus trichocarpa	Mito.	KM091932.1	(unpublished)
	Plastid	NC_009143.1	Tuskan et al. 2006
	Nuclear	N/A (Phytozome v10.0)	Tuskan et al. 2006
Ricinus communis	Mito.	NC_015141.1	Rivarola et al. 2011
	Plastid	NC_016736.1	Rivarola et al. 2011
Silene vulgaris	Mito.	NC_016406.1 NC_016170.1 NC_016402.1 NC_016415.1	Sloan, Alverson, Chuckalovcak, et al. 2012
	Plastid	NC_016727	Sloan, Alverson, M. Wu, et al. 2012
Spirodela polyrhiza	Mito.	NC_017840.1	(unpublished)
	Plastid	NC_015891.1	Wang and Messing 2011
Taxus baccata	Plastid	N/A	(Lars Arvestad 2015, by email 7 sept)
Triticum aestivum	Mito.	GU985444.1	Liu et al. 2011
	Plastid	NC_002762	Ogihara et al. 2002

Chapter 3

Related work

During the course of the project, two related efforts were known to us: a draft *P. Abies* mitochondrial genome and the White spruce mitochondrial genome.

3.1 Draft mitochondrial genome from the Norway spruce project

A draft assembly of the mitochondrial genome of *P. abies* was published with the original Norway spruce paper (Nystedt et al. 2013).

They identified putative mitochondrial contigs using the following criteria (2013, supp. material, p. 12):

1. contig length $>$ 1 kbp,
2. contig coverage $>$ 20 times the average (of the nuclear genome),
3. contig CG content $>$ 40%.

3.2 White spruce mitochondrial genome

With a methodology similar to the one just mentioned, Jackman et al. (2015) prepared the mitochondrial genome of white spruce (*Picea glauca*) by selecting mitochondrial contigs from a much larger set of mixed organellar contigs. The key features used were (my bold font):

> Putative mitochondrial sequences were separated from nuclear sequences by their **length**, depth of **coverage** and **GC** content using k-means clustering in R. — Jackman et al. 2015, p. 31

This paper contains a (in the termonology of this report) *feature-space plot (see Glossary)* of coverage versus GC%, showing identified mitochondrial contigs (Jackman et al. 2015, Figure S2, supp. material). They achieve a reasonable separation just using these two features. See section 7.3 for our corresponding results.

After the assembly Jackman et al. performed additional analysis, such as BLAST alignments (vs. NCBI nucleotide) to verify the mitochondrial classification (Jackman et al. 2015).

Chapter 4

Initial project

This chapter describes earlier versions of the project and the basis for the current approach.

4.1 Initial approach for our project

At the start of this project, project supervisor Lars Arvestad (also co-author of the Nystedt paper) suggested the "Nystedt-method" (see section 3.2 as the initial approach for the classification.

We extended this approach and added similar "basic" sequence-level features, namely "cpg", "N%" and "GCpN" (unpublished internal report 2014). We also added a measure for SNPs (single-nucleotide polymorphisms).

4.2 Formulating the basis of the current approach

To understand why a more powerful approach is necessary, one really has to look at the consequences of *not* using one, see section 7.3. For now, let's just postulate that it *is* necessary and take a quick look at some of the constraints we had (self-imposed or not) when selecting additional features:

- It was determined that developing a Hidden Markov Model (HMM) for sequence classification would be too large of an undertaking for this project.
- The features we use (e.g. GC%, blast hits, coverage) tend to become more stable as the contig length increase, making classification easier (see appendix A). However, due to the "complex" nature of plant mitochondrial structure, it's reasonable to assume that a large fraction of contigs are or could be relatively short. Thus we wanted to set the length cutoff as short as possible without introducing too many false positives.
- The fact that no alignments to any related mitochondrial genomes can be found does not prove that it is not mitochondrial. We felt it was important to include all contigs, even those without BLAST hits.
- A support vector machine (SVM) should be used to do the final classification. This is to make the classification decisions less arbitrary and also hopefully pick up more subtle signals in the features.

Chapter 5

Classification pipeline

This chapter describes the final version of the classification pipeline used to identify the mitochondrial contigs, step by step.

5.1 Pre-processing of source contigs

5.1.1 Filtering out short contigs

Assembled contigs (length $\geq$ 200 nt) were used as input, see section 2.1 for details. First contigs shorter than 500 nt were discarded. This limit was set as a compromise between noise in the features versus skipping too many potential mitochondrial contigs.

5.1.2 Filtering out repeats

Next, the sequences were run through RepeatMasker, a tool that detects repetitive and low complexity DNA regions (Smit et al. 2013-2015). Detected repeats were masked out.

The initial rationale for this was to remove high coverage contigs where the coverage value was effectively an assembly artifact due to repeative regions. We never tested whether repeatmasking helps in this regard, although it is possible. However, another benefit is that repeatmasking is likely to have improved the quality of our BLAST alignments.

Arguments used: `RepeatMasker -pa 1 -x -norna -gccalc -q -species` (species). Version: open-4.0.1.

5.1.3 Filtering out low-coverage contigs

Earlier versions of the pipeline did not have a hard cutoff on contig coverage as it was thought of as more elegant to let the SVM classifier treat it as any other feature. However, for computational reasons it became unmanageable to run `tblastx` on all source contigs. Therefore I decided to put in place a hard coverage cutoff to limit the number of contigs to use for `tblastx` alignments. The threshold value of 100 was based on the draft classifications available at the time, and intended to have some margin of error. There are however no guarantees that all mitochondrial contigs have such a high coverage.

5.2 Pre-processing of reference genomes

A number of reference genomes were used for BLAST queries and also directly as training material for the kmer classifier, see 2.2 for a detailed list.

Upon importing these genomes into the project, the first step was to remove nuclear chromosome 2 of Arabidopsis thaliana due to a large mitochondrial insert on that chromosome, see section B.1 for details.

For cross-validation purposes, partitioned and fragmented copies of the A. thaliana and P. trichocarpa genomes were created. The three (non-overlapping) partitions each contained fragments of about 50 kbp size. (The full genomes were retained as well.)

5.3 Mapping the source contigs to the reference genomes

All source contigs that passed the coverage (≥ 100) and length (≥ 500) filters were aligned to all reference genomes using `blastn` (nucleotide-nucleotide alignment) and `tblastx` ("translation product"-"translation product" alignment where both query and reference nucleotide sequences have been translated into protein sequences corresponding to all six possible reading frames).

BLAST version `2.2.29+` was used with default parameters except for an E-value cutoff of `1e-20`.

5.4 Preliminary contig classification using BLAST

The alignments were used to assign a preliminary organelle label (when possible). The `blastn` and `tblastx` alignments are processed independently. This is done accordingly for each contig:

1. Make one list for each organelle type containing all hits ordered by bitscore.
2. Identify the target organelle of the hit with the highest bitscore.
3. Compare its bitscore with that of the next-best target (if any). If the quotient is less than 1.2, discard the contig.
4. If the best bitscore is less than 100, discard the contig.
5. If the alignment span of the best hit less than 150 bp (`blastn`) or 100 residues (`tblastx`, note that this corresponds to 300 bp), discard the contig.
6. Else, classify the contig with the target organelle type.

The sequences and names of these classified contigs are from now on called the *BLAST extracted [sequences]*.[1]

The constants used for this classification have been determined after investigation of various plots and data, but haven't been subjected to a rigorous evaluation procedure.

[1] Note that there are two sets of contigs, based on `blastn` and `tblastx` respectively.

5.5 Feature extraction

The following features were used by the kmer classifier (see section 5.6), the SVM classifier (see section 5.7) and/or used to filter contig sets or determine e.g. confidence levels.

5.5.1 Coverage

The contig coverage was read from preexisting self-mappings of reads to the assembled contigs that had been made available to me.[2] Note that for the SVM classification, the natural logarithm of the coverage was used as the actual feature.

5.5.2 N%

The fraction N-nucleotides (effectively meaning: non-ACGT) was calculated directly from the contigs.

5.5.3 GC%

The fraction GC (i.e. base C or base G) was calculated as

$$frac_{GC} = \frac{count_{GC}}{count_{ACGT}}.$$

5.5.4 "Blast extracted" (preliminary contig classification)

Target class or "none", directly based on the selection of contigs described in section 5.4.

5.5.5 Number of distinct target reference genomes matched

This feature summarizes the number of matching references genomes of each organelle type. Counting logic for each contig and BLAST type (`blastn` and `tblastx`):

1. For each target organelle, look for hits with a bitscore of at least 100 and a length of at least 100 (`blastn`) or 33 (`tblastx`).
2. If any, add the class of this organelle ("mt", "nu" or "cp") to the list of matches for this contig.

Bug in the script

However, there is unfortunately a bug in the counting script that wasn't spotted before writing this report. The bug causes cross-contamination of the two BLAST types, so that hits of one type may show up in the organelle match list for the other. For a contamination to occur, there need to exist a preliminary classification (other than "none") for the BLAST type in question, and the hits of the contaminating BLAST hit type must pass the minimum-length criteria (33 or 100). This means that it's much easier for `tblastx` match lists to be contaminated by `blastn` hits than the other way around, as a very short 33 bp `blastn` hit will do, whereas for a `blastn` match list to be contaminated there need to exist a 100 residue hit.

[2] Most likely the mappings were done using BWA.

Note that all hits still need to pass the bitscore requirement.
For a discussion of the consequences of this bug, see section 8.3.4.

5.5.6 Score from the kmer classifier

See below. Note that the kmer classifier is in turn trained using some of the other BLAST-based features.

5.6 The kmer classifier

The basic idea for this classifier is to calculate probability tables for occuring kmers (in our case with k=7, see below) based on positive and negative training sequences and use this to classify sequences by determining a candidate score for each class accordingly:

$$s_c = \prod_{k=1}^{n} p_{c,kmer_k},$$

where c is the class (positive or negative), n is the number of kmers in the contig, $p_{c,kmer_k}$ is the probability of kmer k occuring in a sequence of class c.

All contigs received a so called *pseudo count* of 1. This simply means that the table used to count occuring kmers is initialized to 1 in all cells instead of 0. It is a common technique to prevent a model from automatically disqualifying any (in this case) sequence having a least one (in this case) kmer absent from all training sequences. With our long contigs this would of course make the model completely useless in practice.

The resulting class (i.e. score) of each tested contig was defined as:

$$s = \frac{\log s_{pos} - \log s_{neg}}{L},$$

where L is the contig length and s_c is defined as above.

When scoring a contig, the reverse (not the reverse complement) of each kmer was considered as well.

I implemented this classifier as a C program for this project.

5.6.1 Training data for the main classification

As positive (mitochondrial) training data, all *blast extracted sequences* (see section 5.4) (both BLAST types) for the species were used.

As negative training data, the known full chloroplast genome for the species plus the *blast extracted* nuclear sequences were used.

5.6.2 Determining the optimal kmer length

Based on a simple set of tests based on test/training data from A. thaliana, a kmer length of 7 was selected based on visual inspection of the receiver operating characteristic (ROC) curves (see figure 5.1).

5.7 Contig classification using SVM

The final classification was done solely based on the SVM classification score. However, BLAST-based features were used to identify a subset of "high confidence" contigs.

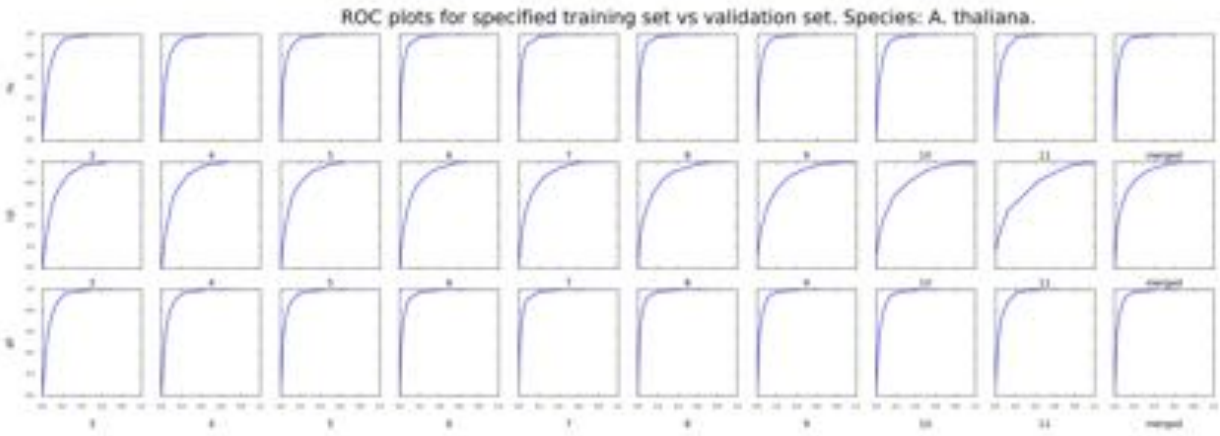

Figure 5.1: *Receiver operating characteristic curves for the kmer classifier. Receiver operating characteristic (ROC) plots for various kmer lengths. The three rows (nu, cp, all = nu+cp) refers to the training set used. Training and test sets are from A. thaliana. The x-axis of each subplot is the FPR (false positive-rate) and the y-axis is the TRP (true positive rate). A steeply increasing ROC-curve thus corresponds to a classifier that quickly "ramps up" true positives without getting too many false positives along the way.*

5.7.1 Training contigs

Contigs used for training the classifier were selected using these criteria:

1. The contig must have a "blast extracted" classification (see section 5.4) for at least one BLAST-type.
2. If both `blastn` and `tblastx` classifications exist they must be identical.
3. At least two different reference species must have BLAST hits (either `blastn` or `tblastx` is fine). Note that this rule is subject to the bug described in section 5.5.5, i.e. it is possible that a contig matching only one species passes this test.

5.7.2 Features

The following "direct" features were used: GC%, N%, ln(*coverage*) (see section 5.5). In addition the kmer score was used (see section 5.6.1).

The features were standardized:

$$n_k = \frac{x_k - \mu_k}{\sigma_k},$$

where x_k denotes the raw feature value, n_k the standardized value, μ_k the feature mean and σ_k the feature standard deviation for feature k. The means and standard deviations were calculated on the training set and then used in both the training and classification steps.

When calculating the cross-validation statistics, in some cases $\sigma_k = 0$. For these cases n_k was set to x_k.

5.7.3 SVM configuration

SVMlight was chosen (Joachims 1999) based on a recommendation from Lars Arvestad.

A gaussian (radial basis function) kernel[3] was selected already during the pilot project. There was no apparent need to change this.

When using this kernel there are two hyperparameters to consider: C and γ. We used $C = 1$ and $\gamma = 1$. It is not clear how these parameters were set. The cross validation runner script has a commented-out section suggesting that a parameter search was carried out, but I can find no trace of the results. (During the pilot project a parameter search was conducted, but the results are irrelevant due to e.g. a different set of features.) Tweaking the C parameter up and down a few orders of magnitude did however not result in any improvements, suggesting that the parameters were set reasonably to begin with.

5.8 Cross-validation statistics

The cross-validations were performed using the same features, settings and parameters as the main classification unless otherwise stated.

5.8.1 Trial design

One-hundred cross-validation series were run, each consisting of nine trials with a different training fraction size (varying from 10% to 90%). Due to the way the random sets were prepared, the proportions are only approximate.

Training and classification

Out of the training candiate set, "trusted" contigs were selected the same way as before (see section 5.7.1). Non-trusted contigs from the training candidate set were discarded.

A new set of kmer scores were then calculated for the training contigs and the test set.

The remaining steps in the SVM training and classification pipeline were then performed the same way as before.

Calculation of statistics

A set of "trusted" test contigs was selected with the same rules as for the SVM training (see section 5.7.1). The SVM scores were noted and statistics about true/false positives/negatives were gathered.

If no true positives were found, the trial was ignored (but not re-run). Otherwise the recall was calulated as: $recall = \frac{TP}{TP+FN}$.

The FDR was calculated as $FDR = \frac{FP}{TP+FP}$. If no positives (true or false) were found, it was set to 0 (by definition).

After running all trials, average recall and FDR were calculated for each species. These statistics were calculated twice: based on the number of contigs and also the number of base-pairs in the contigs.

5.9 Delivering the final contig classification

By necessity (as the number of non-mitochondrial contigs is much larger than the number of mitochondrial contigs), the analysis is designed to be selective.

[3] SVMlight `-t` parameter value `2`.

However, some use-cases may warrant even stricter selection criteria. For that reason we have prepared four contig sets (from now on known as *confidence classes*) with different inclusion criteria for mitochondrial contigs:

- `raw_svm_classified`: Contigs with a SVM score > 0.
- `all`: Contigs with a SVM score > 0 with additional BLAST cleaning from nuclear and and chloroplast hits (see section 5.4).
- `medium`: Same as criteria as "all" but also requires ambiguity in the "BLAST extracted" classification.
- `high`: Same criteria as "all" but requires an *unambiguous* "BLAST extracted" classification of mt (mitochondrial).

Note the following two relations.

1. $\texttt{medium} \cap \texttt{high} = \emptyset$
2. $(\texttt{medium} \cup \texttt{high}) = \texttt{all} \subseteq \texttt{raw\_svm\_classified}$

The "raw_svm_classified" category was a last-minute addition as I discovered that some "good looking" contigs were not included in the `all` list above. See section 8.1.3 for details.

Chapter 6

Plot guide

In this thesis two key types of plots are used to show the contig set for a species. The "feature space" plots visualize the placement of contigs in selected feature dimensions whereas the "contig plots" show information about alignments to reference species.

6.1 Feature-space plots

The feature space plot type shows all contigs (length $\geq$ 500 and coverage $\geq$ 100) for a species displayed in a 2D "feature-space" plane, illustrating clustering and separation of contigs. To be able to label these unknown contigs, the "blast extracted" (see section 5.4) classification has been used. See figure 6.1.

6.2 Contig plots

The contig plot type aims to visualize sets of contigs by displaying the location, span and strength (i.e. the *number* of different species matching) of BLAST alignments and also whether the contig was selected as mitochondrial by the classification pipeline. See figure 6.2 for how to read it.

Technical details

- In this report all "contig plots" are based on the same length $\geq$ 500 and coverage $\geq$ 100 cutoffs as used for the main pipeline.
- The same BLAST alignments as for the pipeline was used, see section 5.3 for details.
- Only alignments with a bitscore at least 20 are shown.
- Each contig is represented as a number of bins of 500 bp. A bin is considered occupied if any part of a sequence or alignment falls within that 500 bp block.

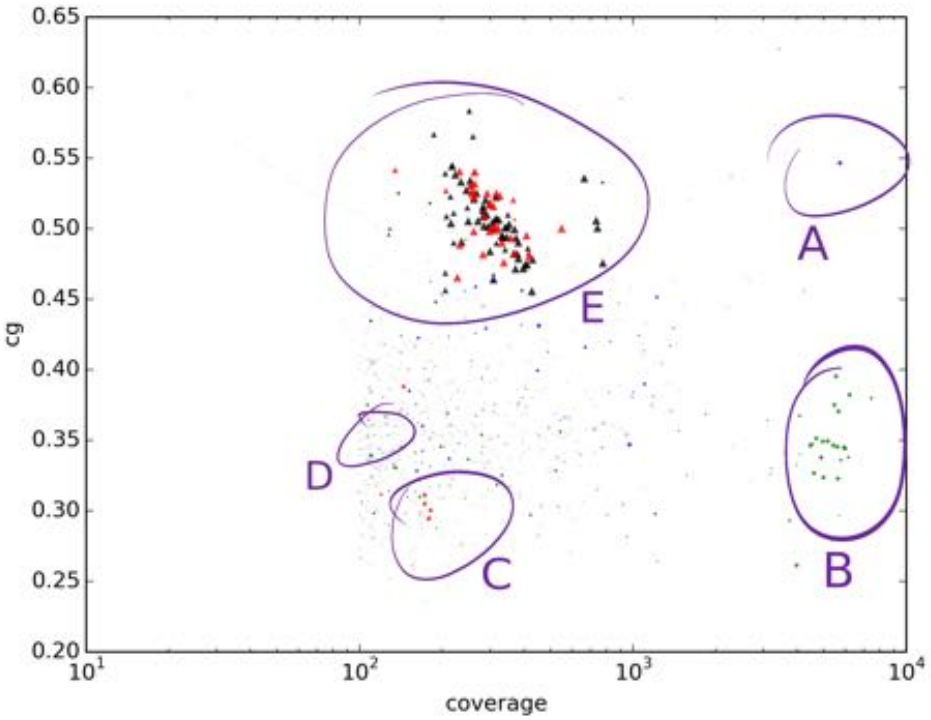

Figure 6.1: *How to read "feature-space" figures.*
Overall structure: *Each marker respresents a contig. The size of the marker is roughly proportional to the contig length. By default contigs are shown as circles. Contigs classified as mitochondrial (using the* `raw_svm_classified` *category, see section 5.9) are shown as triangles.* ***(A)*** *Contigs with an unambiguous "blast extracted classification" (see section 5.4) of* ***nuclear*** *are shown in blue.* ***(B)*** *Likewise,* ***chloroplast*** *contigs are shown in green.* ***(C)*** *Here we see four red (****mitochondrial****) contigs (and some other contigs as the drawing isn't perfect).* ***(D)*** *This circle shows mostly grey contigs. Grey and black contigs have an undetermined label. Yellow contigs (not shown in this plot) have conflicting "blast extracted" classifications.* ***(E)*** *This is the cluster of contigs classified as mitochondrial. Note that seemingly about half have been "detected" by BLAST as well.*

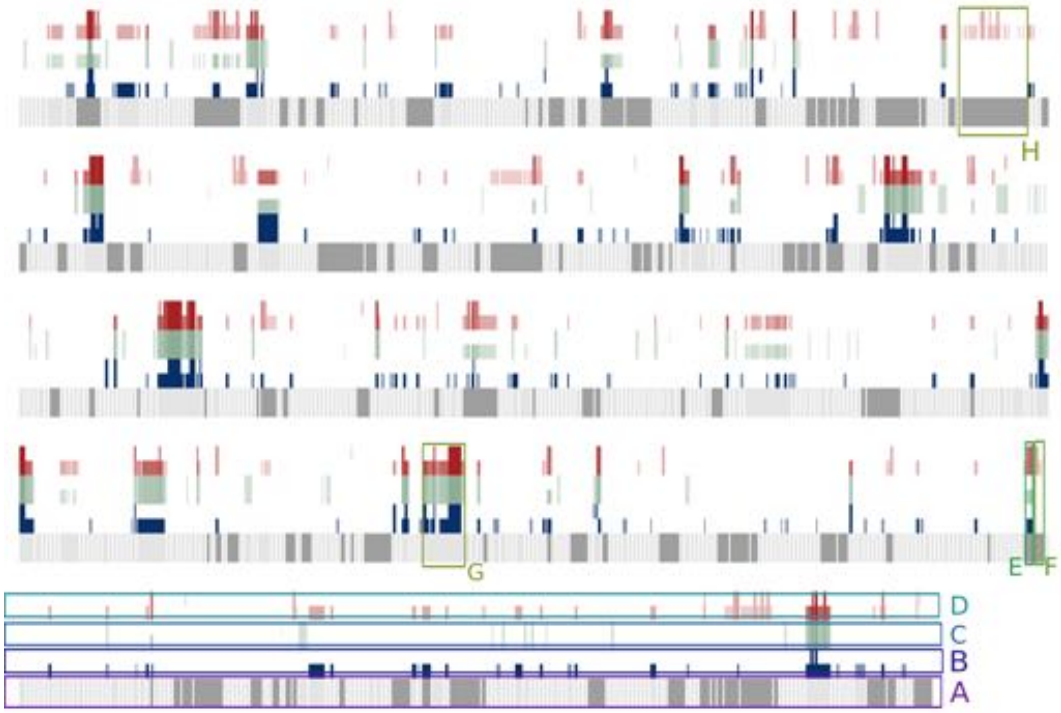

Figure 6.2: *How to read "contig plot" figures.*
See also section 6.2. ***Overall structure****: four main bands of equal height (A-D). Each band is essentially a line that's been stretched out vertically for readability (just like a bar-code). Each contig is thus represented by a grey rectangle, with a width that is proportional to the contig length. Contigs are separated by a tiny gap. To make the plot more compact, several rows (five in this case) of these four-band structures are displayed in one figure.* ***(A)*** *The bottom band has two functions: (1) darker-grey sections indicate a mitochondrial classification (the* `all` *category was used, see section 5.9) and (2) marks the presence of a contig.* ***(B)*** *The blue band represents BLAST-alignments to* ***nuclear*** *sequences of related species, with where the upper half shows* `blastn` *alignments and the lower half shows* `tblastx alignments`*. The color intensity is proportional to the fraction of reference genomes that mapped. Note that alignment quality (e.g. E-value) is* ***not*** *shown.* ***(C)*** *Same as (B) but for* ***chloroplast*** *alignments.* ***(D)*** *Same as (B) but for* ***mitochondrial*** *alignments.* ***(E)*** *This is an example of a contig with* `tblastx` *alignments to all three organelle types but only* `blastn` *alignments to mitochondrial references.* ***(F)*** *This example contig only has mitochondrial alignments.* ***(G)*** *This contig has both types of BLAST alignments to all three reference types, but only the chloroplast alignments cover the whole contig (or close to it).* ***(H)*** *This contig map only to michondria, mostly with* `tblastx` *alignments.*

Chapter 7

Results

7.1 Key results

After the first filtering step of removing contigs shorter than 500 bp or with a coverage less than 100, the amount of sequence material left was generally in the range of 1–8 Mbp, with the exception of *P. abies*, having a full 132 Mbp. These sequences were classified and the sizes of the identified mitochondrial genomes were in the range of 0.5–4.2 Mbp, with *T. baccata* and *G. gnemon* being around 0.5 Mbp, *A. sibirica*, *J. communis* and *P. sylvestris* being around 1–1.7 Mbp and again *P. abies* being the largest at 4.2 Mbp. The number of identified contigs ranged from 66 to 594. No strong relationship between the number of input contigs and classified positive (i.e. mitochondrial) contigs could be seen. See table 7.1 for details.

7.2 Cross-validation statistics

As can be seen in table 7.2, the false discoverate rate (FDR) is generally very good except for *P. abies* and *P. sylvestris*. Furthermore, species with many contigs generally show worse classification performance. There are exceptions however, with *A. sibirica* showing excellent FDR and recall despite being relatively large. This can be explained by the low number of mitochondrial contigs, indicating a higher assembly quality. Another exception is *T. baccata* with a surprisingly low recall of 88%. The FDR performance of *P. abies* is, as expected, the worst of the studied species. However it still represents a situation where 99.4% of the input contigs have been discarded while maintaining a 92% recall.

Figure 7.1 shows plots of recall and FDR as function of the training set size (positive only). The FDR is generally quite low when the amount of training data used is low (except for *P. abies* and *P. sylvestris*). For most species, the recall rapidly increases to high levels with still only a small fraction of the training data used, suggesting that the model does in fact work by capturing general features of mitochondrial contigs rather than effectively over-training on known kmers. Note that the increase in variance seen as the training set size increases past about 80% corresponds to the test set used to calculate the values getting increasingly small. Also, the plots do not properly show that a large fraction of the data points indeed sit at recall 100%.

See figure 7.2 for an alternative version of the plot, with the same statistics but now as a function of the size of both positive *and* negative training data.

Table 7.1: *Key classification statistics.*

Note: (`medium` ∪ `high`) = `all` *(see section 5.9 for more details).*

Statistic	*A. sibirica*	*G. gnemon*	*J. comm.*	*P. abies*	*P. sylv.*	*T. baccata*
Number of input contigs ≥ 500 bp and cov ≥ 100	6.5 k	1.1 k	1.9 k	49.5 k	5.1 k	0.7 k
Length of input contigs ≥ 500 bp and cov ≥ 100	7.9 Mbp	2.0 Mbp	3.0 Mbp	132 Mbp	5.6 Mbp	1.1 Mbp
Pos. SVM train. data	754 kbp	244 kbp	369 kbp	2.00 Mbp	203 kbp	226 kbp
Neg. SVM train. data	641 kbp	284 kbp	279 kbp	51.1 Mbp	370 kbp	191 kbp
Number of contigs classified as mitochondrial for each confidence group:						
`raw_svm_classified`	66	149	160	301	594	112
`all` = (`medium` ∪ `high`)	66	146	160	286	586	112
`medium`	29	93	116	196	508	77
`high`	37	53	44	90	78	35
Length of sequences classified as mitochondrial for each confidence group:						
`raw_svm_classified`	960 kbp	530 kbp	1.08 Mbp	4.69 Mbp	1.71 Mbp	465 kbp
`all` = (`medium` ∪ `high`)	960 kbp	525 kbp	1.08 Mbp	4.20 Mbp	1.65 Mbp	465 kbp
`medium`	197 kbp	255 kbp	651 kbp	1.94 Mbp	1.40 Mbp	265 kbp
`high`	764 kbp	270 kbp	433 kbp	2.25 Mbp	258 kbp	200 kbp

Table 7.2: *Key cross-validation statistics.*

Explanation: Average recall/FDR for various test/train fractions (SVM cross-validation): *The recall statistic indicates the fraction of all positive sequences that were deteteced. The FDR statistic similarly indicates the sequence-fraction of false positives among the identified contigs. For details about how the cross-validation trials were performed and evaluated, see section 5.8.*

Statistic	*A. sibirica*	*G. gnemon*	*J. comm.*	*P. abies*	*P. sylv.*	*T. baccata*
Average recall for various test/train fractions (SVM cross-validation):						
10% train	0.53	0.66	0.76	0.53	0.33	0.65
50% train	0.99	0.97	0.98	0.88	0.79	0.85
90% train	0.98	0.98	0.97	0.92	0.89	0.88
Average FDR for various test/train fractions (SVM cross-validation):						
10% train	0.00	0.05	0.00	0.13	0.09	0.01
50% train	0.00	0.02	0.00	0.20	0.18	0.00
90% train	0.00	0.02	0.00	0.22	0.11	0.00

See section 8.2 for further discussion and section 5.8 for details on how the cross validation statistics were calculated.

7.3 Contig clustering in feature-space

Considering that GC% and coverage (combined with a length filter) were the key features used in both related projects, it was natural to start with those when looking at our data. Looking at the GC%-coverage feature-plane I've identifed three groups of species, here examplified with one species each. (Additional types of feature plots (for all species) are available in appendix C.)

- Classification is easy: *G. gnemon*, *J. juniperus* and *T. baccata*. See figure 7.3.
- Classification is managable: *A. sibirica* and *P. sylvestris*. See figure 7.4.
- Classification is not practical using only GC% and coverage: *P. abies*. See figure 7.5.

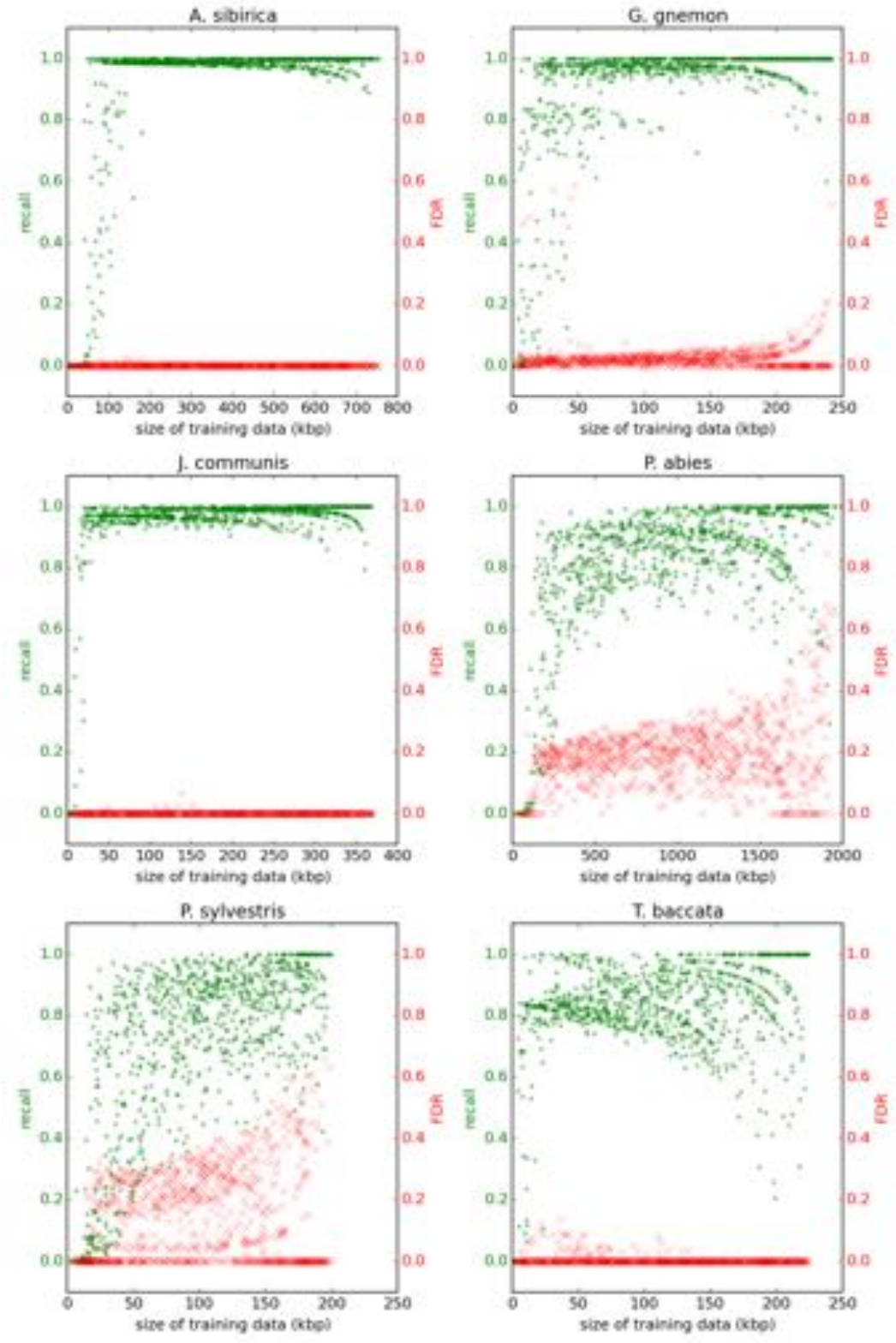

Figure 7.1: *Cross-validation statistics for all species (X-axis: size of positive training data only).*
These plots show recall and FDR as function of the size of the positive training data used. Based on the same cross-validation trials as figure 7.2.

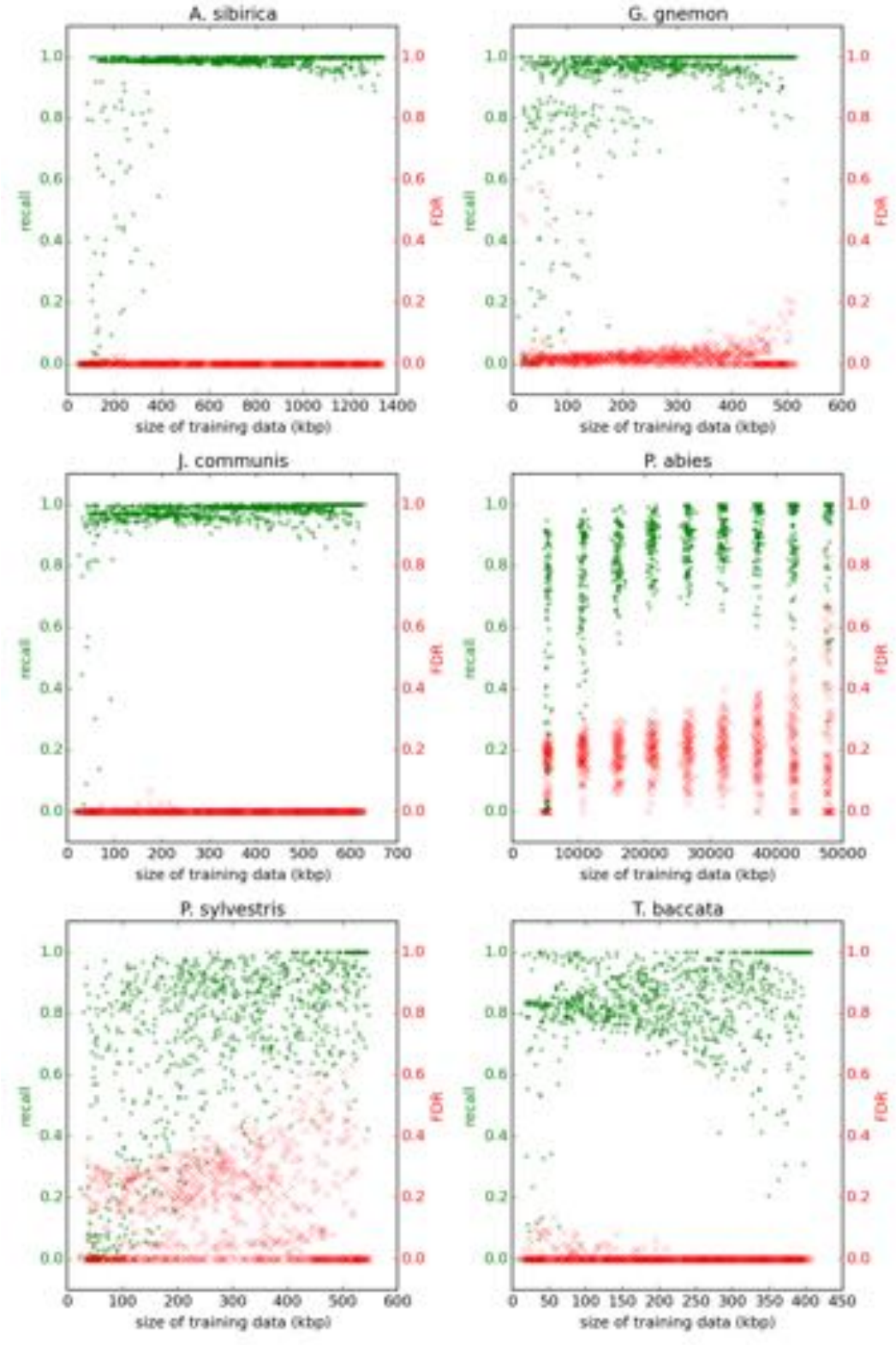

Figure 7.2: *Cross-validation statistics for all species (X-axis: size of all training data).*
These plots show recall and FDR as function of the size of the training data used. For each species, 900 trials were run based on 100 random partitions of the labeled contigs. See section 7.2.

In the "easy" group, it seems that this feature pair is almost entirely sufficient on its own. In the case of *P. abies* however, one can barely tell there is any separation at all.

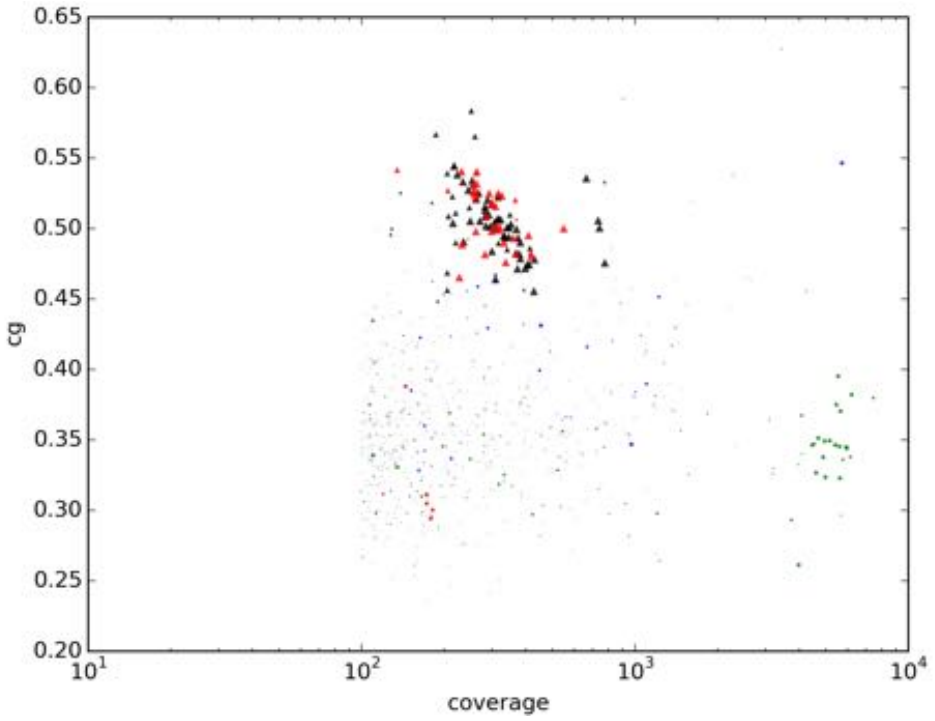

Figure 7.3: *T. baccata: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

Focusing our attention to the *P. abies* case, we see that the kmer classifier scores are key to achieving separation of the mitochondrial contigs. Comparing figure 7.5 to figure 7.6 illustrates the improvement when using the kmer score over the GC% feature. Interestingly, while the kmer score was originally thought of as a kind of generalization of GC% (with kmers of length seven rather than one), it is in instead when these two features are combined that the best separation is achieved for *P. abies*. See figure 7.7.

7.4 Results per-species

As we have seen in the previous section, the "GC% vs. k7" feature-space plots seem to show the best separation. To put the species in relation to another, we here show this feature-space plot for all species.

In the case of *Abies sibirica*, the putative mitochondrial contigs appear to be nicely separated. See figure 7.8. For *Gnetum gnemon* we see a nice separation. The only potential worry is the number of selected contigs with non-mt BLAST classifications (I count at least seven). See figure 7.9. Also for *Juniper communis* we see a very nice separation. See figure 7.10. The separation is good in the GC/coverage feature-plane as well. In the trickier case of *Picea abies* we see an acceptable separation. It's not as clean as the other species due to the large number of contigs involved. See figure 7.7. For *Pinus sylvestris* there are a lot of contigs selected, but the separation looks

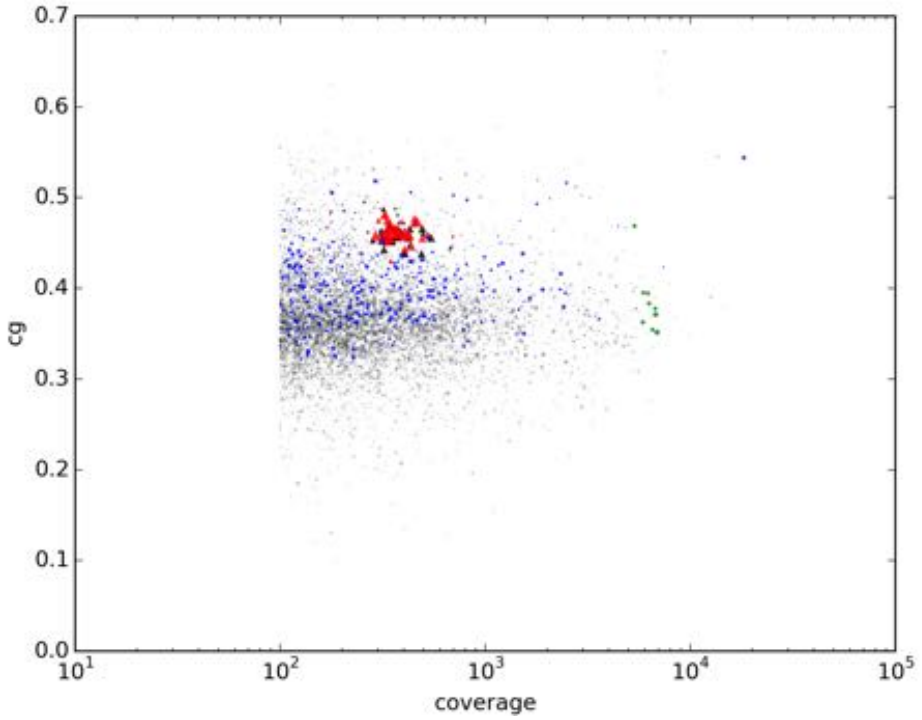

Figure 7.4: *A. sibirica: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

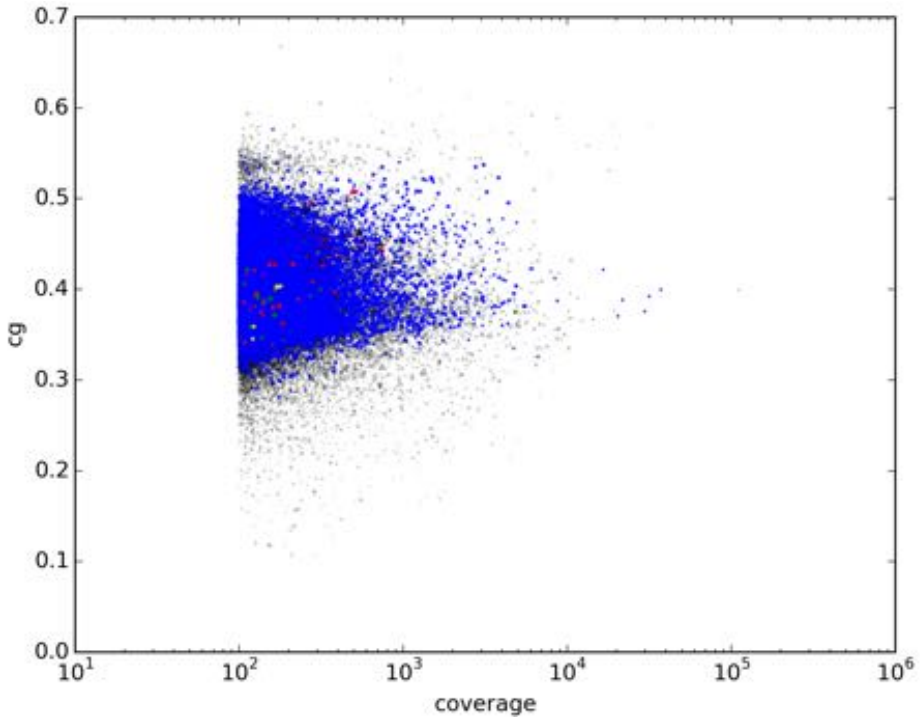

Figure 7.5: *P. abies: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

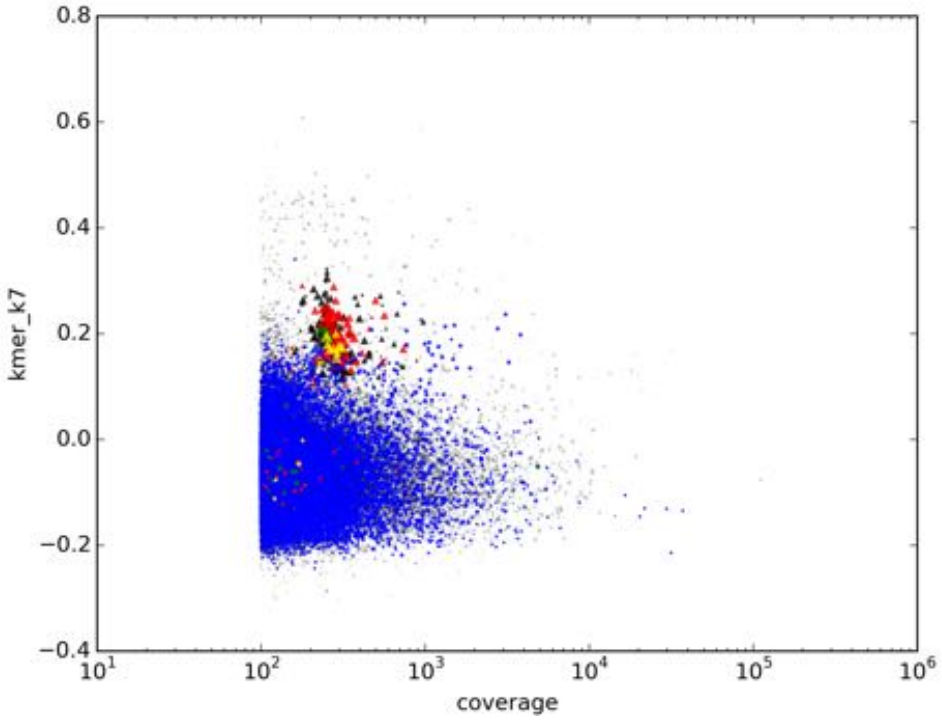

Figure 7.6: *P. abies: coverage vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

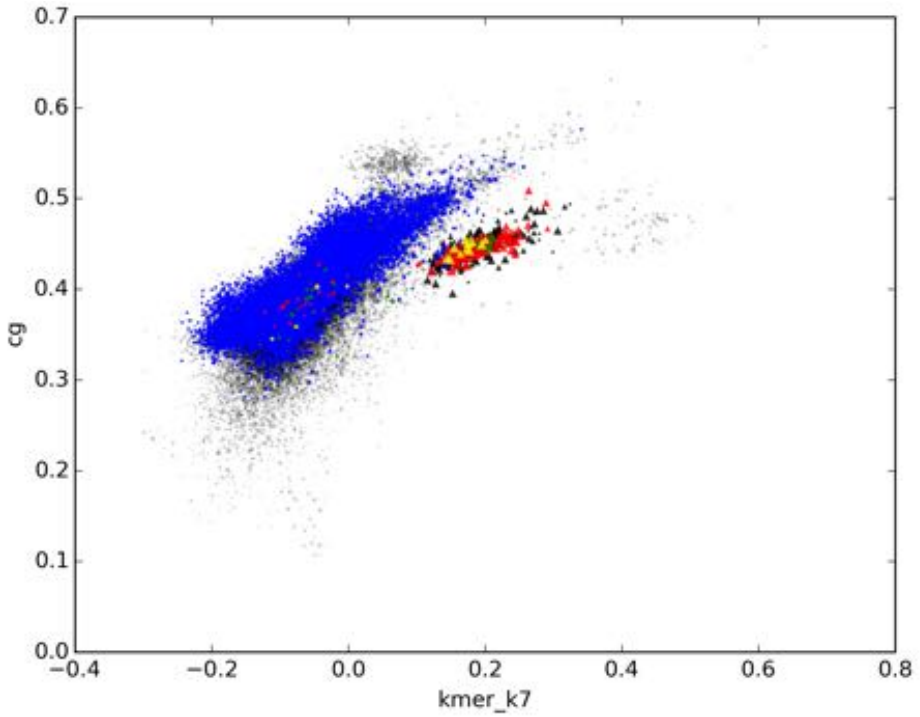

Figure 7.7: *P. abies: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

good. There are issues with non-mt BLAST classifications of selected contigs here as well (see G. gnemon above). See figure 7.11. Finally for *Taxus baccata* the separation is very nice. See figure 7.12.

See also appendix C for more feature-space plots.

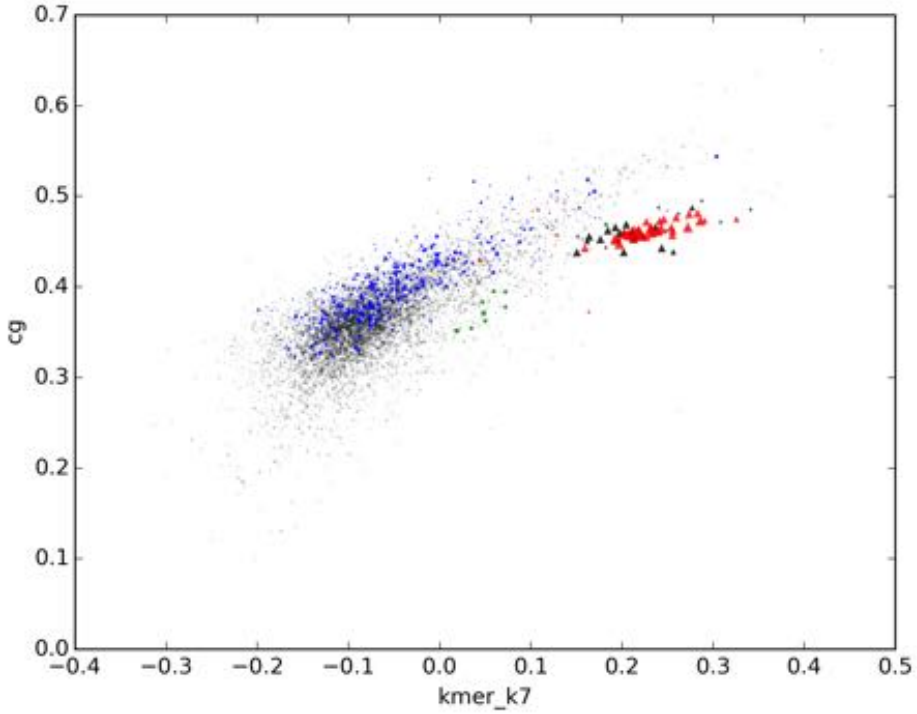

Figure 7.8: *A. sibirica: GC% vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.*

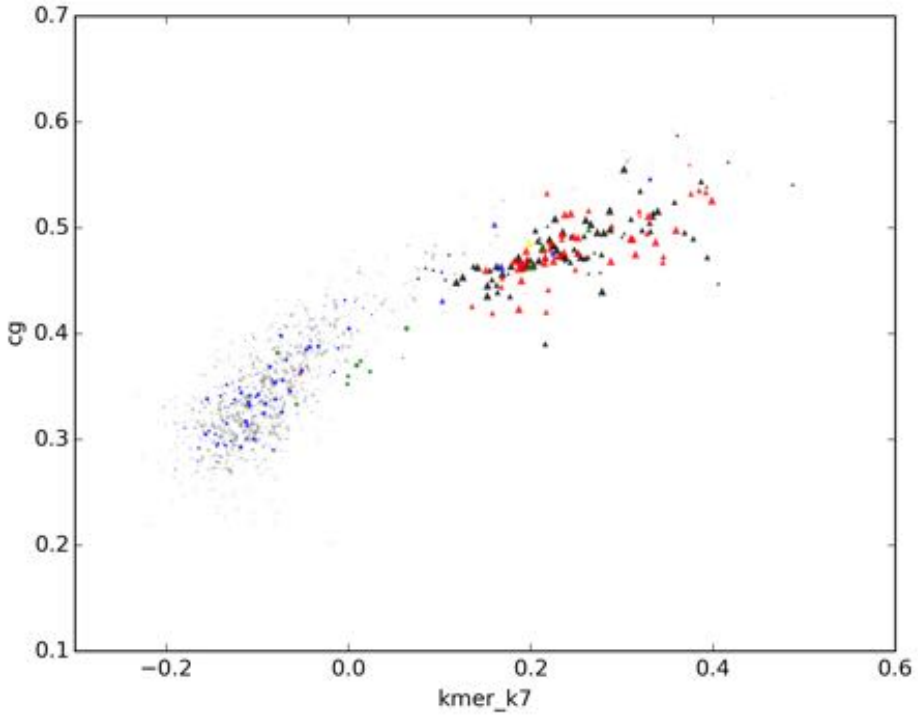

Figure 7.9: *G. gnemon: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

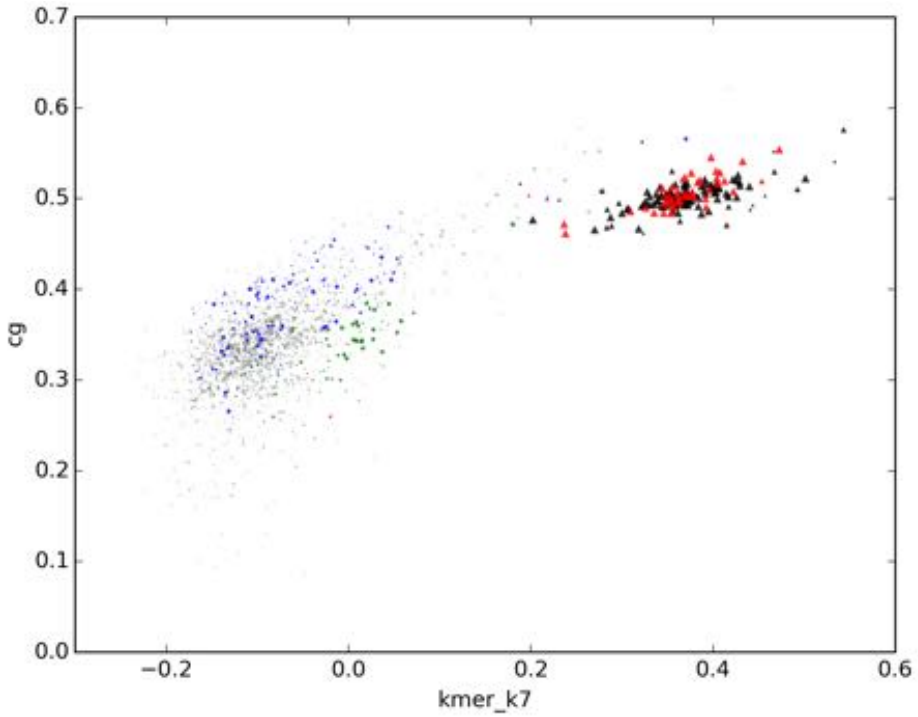

Figure 7.10: *J. communis: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

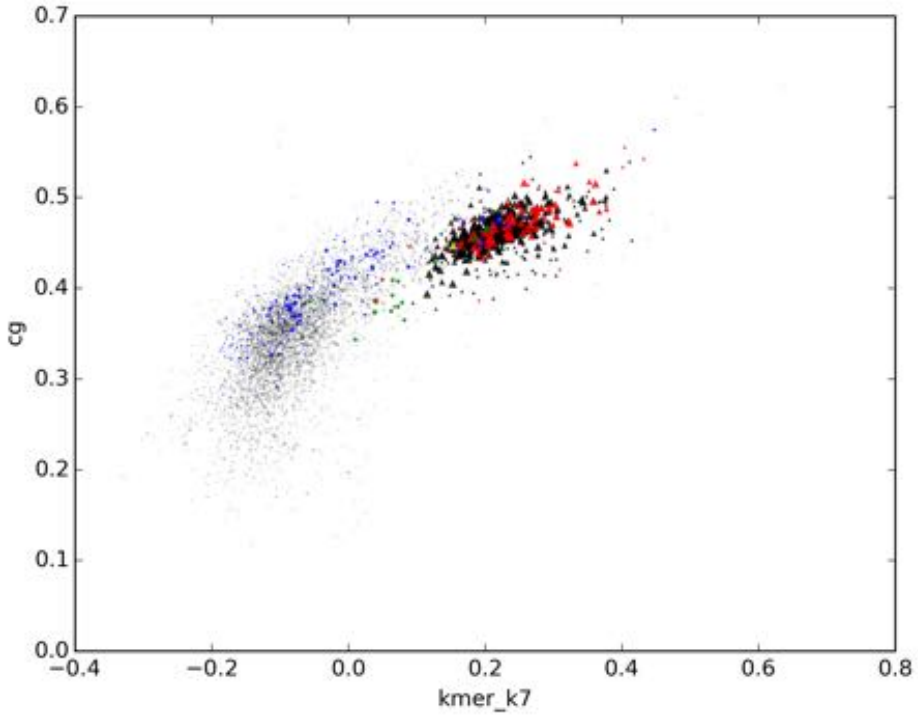

Figure 7.11: *P. sylvestris: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

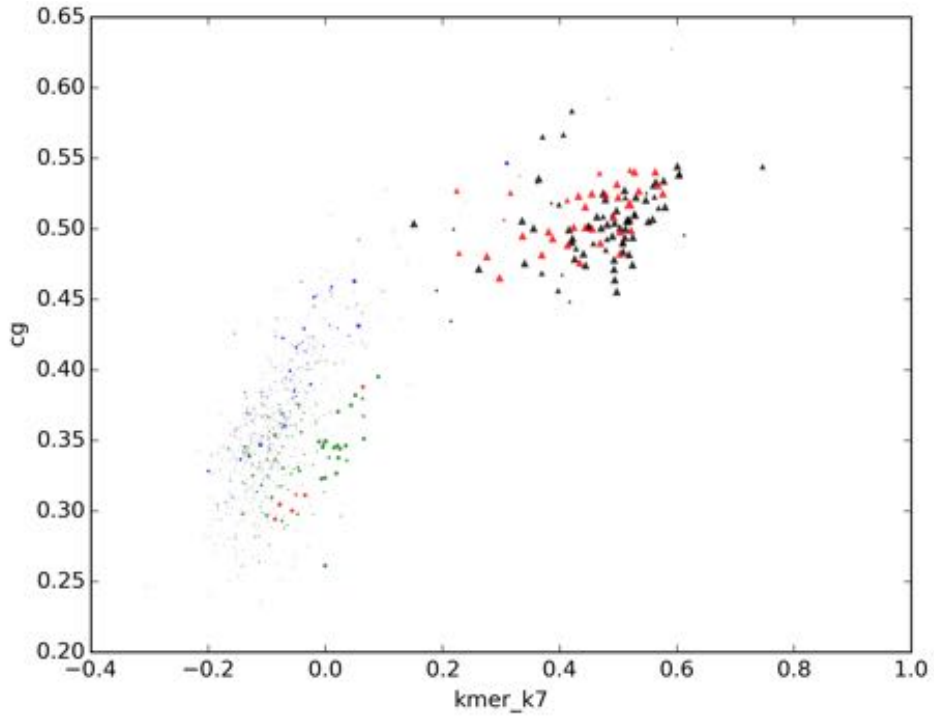

Figure 7.12: *T. baccata: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

Chapter 8

Discussion

In this chapter we will discuss three categories of potential quality issues: the accuracy of the classification, the reliability of the cross-validation and other error sources. Finally we will briefly compare this project to a related one.

8.1 Classification accuracy

8.1.1 Gene-rich DNA bias

Due to the strong weight given to kmer-scores in this classification model, it can be hypothesized that the model has a strong bias for correctly identifiying gene rich contigs. These are of course generally the most biologically interesting ones. No attempt has been made during the project to prove this hypothesis.

8.1.2 Lack of gold standard evaluation method

A persistent issue in this project has been the lack of an independent "gold standard" measure of the accuracy of the classification process (measured e.g. by recall and FDR). As explained in previous chapters, we resorted to using BLAST alignments with strict cutoffs as an "objective" means of classification. However, this is not an independent measure as we also use the BLAST alignments to pick the contigs to use for training.

The only "gold standard" measure I can think of is selecting a random sample of contigs and manually classifying them by carefully studying alignments, ORFs, coverage etc. This would however have been very time consuming.

With that said, I think that the level of accuracy (in terms of recall and FDR) is generally satisfying. Careful study of the feature plots also gives support to the reasonableness of the cross validation statistics.

8.1.3 Post-SVM filtering

The classification pipeline has a post-processing step removing contigs having "better" BLAST hits for nuclear or chloroplast genomes, see section 5.9 for the exact criteria.

When studying the contig plots I found one *P. abies* contig that wasn't selected by SVM but by inspection of the contig plot should have been. See figure 8.1. It was classified as chloroplast based on `tblastx` alignments. However, it clearly has great alignment coverage to mitochondrial genome(s) (about

98%, compared to only about 1.5% chloroplast alignment coverage), making it highly likely to be mitochondrial.

Solution to the over-aggressive filtering

As a last-minute fix to make sure that such good long contigs are not completely excluded from the classification, the `raw_svm_classified` confidence level was created. It simply bypasses the BLAST-based filtering of the SVM classification.

Also, coverage-fraction should have been tested as a feature for the SVM classification.

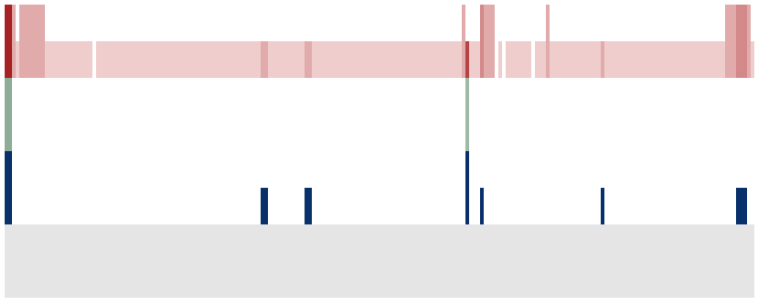

Figure 8.1: *Incorrectly classified Picea abies contig.*
Contig plot for P. abies contig `MA_10431364` *(length=102337). Note the almost complete mitochondrial* `tblastx`*-coverage. It was likely misclassified as non-mitochondrial. See figure 6.2 for how to read this plot type.*

8.2 Cross-validation reliability

8.2.1 Narrow selection criteria for labeled contigs

For both classification and cross-validation, the criteria for selecting a contig as a "known" example of a particular class is a non-conflicting classification between the `blastn` and `tblastx` "best alignment picks" and that the contig maps to at least two reference species. Since we only have two nuclear reference genomes, a contig that maps to only one of them will not be used to train the classifier or to calculate e.g. FDR.

The down-side of this narrow selection process is that there could potentially be a lot of misclassified contigs slipping past the cross-validation net.

Thus, it's important to note that the recall and FDR values given in table 7.2 are *estimates*. For all species except *A. sibirica* about half or more of the contigs classified as positive are "unknown" and have thus not been included in the statistics. This means that in principle *J. communis*, for example, could have a 50% FDR. In reality I do not think it is anywhere near that bad, see the discussion section.

8.2.2 Comments to the cross-validation plots

Some remarks about figure 7.2:

- As the training set increases in size, the test set decreases, which is likely to account for much of the increased variance that can be seen with the larger training sets.
- The partitioning algorithm and the fact that in total nine different sizes of training sets (from approximately 10% to 90%) were used accounts for the banding that can be seen in the *P. abies* plot; in fact this species shows similar characteristics to *P. sylvestris*.
- The *P. sylvestris* data set has almost three times as many contigs to classify as *G. gnemon*, but about the same amount of labeled data, which could explain why *P. sylvestris* is showing worse outcomes.
- *T. baccata* has the smallest negative training set but among the best FDR statistics.

8.2.3 Missing analysis: false seed test

As the kmer classifier ultimately only can be as accurate as its training data, it would have been interesting to study how resilient it (as well as the whole pipeline) is to incorrectly labeled training data. One could also have studied what effect kmer length has on resilience.[1] Unfortunately this wasn't done.

8.3 Error sources

8.3.1 Cross-organellar duplications

A preliminary investigation found over half of the *Picea glauca* mitochondrial genome to be respresented in the nuclear genome (with an average sequence divergence of 4%) (Jackman et al. 2015). They also found 98% of the chloroplast genome duplicated. Looking at the contig plots (see appendix D) of our six studied species, one can suspect similar duplications among our reference species and/or the studied species. This is a complicating factor for using sequence alignments for classification.

8.3.2 Possible mitochondrial contamination of *Populus trichocarpa* nuclear genome

It is "definitely possible" that the *P. trichocarpa* nuclear reference genome used contains mitochondrial contamination. (Nathaniel Street 2015, by email to Lars Arvestad 9 feb). This could potentially skew the classifier towards false negatives.

8.3.3 Bacterial contamination

E. coli and fosmid vector contaminants have been removed for *P. abies* (Lars Arvestad 2017, by correspondence). I haven't performed any screening myself though. The other five species have also been subjected to some contaminant

[1] Generally one of course would expect the classifier to be more susceptible to overtraining with longer kmers, but at what kmer length does this begin to be a problem?

filtering (Lars Arvestad 2017, by correspondence). It should still be noted as a possible error source. We do have something of a safe-guard in that very high coverage contigs should have been excluded by the SVM.

8.3.4 Bug in the code that counts the number of distinct reference genomes matched

The code used to count the number of distinct reference genomes matched contains a bug (described in section 5.5.5).

Since no control experiment has been run[2], the exact consequences are unknown. But an obvious risk is the selection of training contigs based on low quality alignments, biasing the classifier.

8.4 Comparison to related works

Both related projects (*Picea abies* draft mitochondrial genome and *Picea glauca* mitochondrial genome) use only GC%, coverange and length as key features for a first classification step (Nystedt et al. 2013; Jackman et al. 2015).

As can be seen from our feature-space plots (in particular the plots having length as one dimension), the mitochondrial contig selection problem can in many cases be quite managable using only these "basic" features. However, when going for short contigs (as low as 500 bp) and working with species with massive amounts of contigs (*P. abies*), the "basic" features of GC%, coverage and length no longer gives satisying results (see e.g. 7.5).

[2] Fixing this bug and re-running the entire analysis is outside of the scope of writing this report.

Chapter 9

Conclusions

Mixed data-sets containing material from multiple species or organelles are not uncommon in genomics. In this paper we have presented an additional tool for facilitating the classification of contigs: the kmer classifier. Some or all of our data-sets are likely difficult from a classification point-of-view as comparable species are known to contain extensive cross-organellar duplications of high similarity. Despite this, the kmer classifier performance on our data-sets have ranged from acceptable to exceptional.

It merits further study to investigate the capability envelope of this tool: how much training data does it need, how much diversity within genomes can it handle and how different do genomes have to be? Of particular interest would be to study its applicability to metagenomics.

Finally, the project has delivered on its promise to identify the mitochondrial genomes of the six studied species (to a reasonable degree of accuracy).

Bibliography

Chaw, Shu-Miaw, Arthur Chun-Chieh Shih, Daryi Wang, Yu-Wei Wu, Shu-Mei Liu, and The-Yuan Chou (2008). The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. *Molecular biology and evolution* 25 (3), pp. 603–615. ISSN: 1537-1719. DOI: 10.1093/molbev/msn009.

Cooper, GM (2000). *The Cell: A Molecular Approach*. 2nd ed. Sunderland (MA): Sinauer Associates. URL: https://www.ncbi.nlm.nih.gov/books/NBK9905/.

Gan, Xiangchao et al. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* 477 (7365), pp. 419–423. ISSN: 1476-4687. DOI: 10.1038/nature10414.

Goremykin, Vadim V, Karen I Hirsch-Ernst, Stefan Wolfl, and Frank H Hellwig (2003). Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm. *Molecular biology and evolution* 20 (9), pp. 1499–1505. ISSN: 0737-4038. DOI: 10.1093/molbev/msg159.

Gualberto, José M, Daria Mileshina, Clémentine Wallet, Adnan Khan Niazi, Frédérique Weber-Lotfi, and André Dietrich (2014). The plant mitochondrial genome: dynamics and maintenance. *Biochimie* 100, pp. 107–120. ISSN: 1638-6183. DOI: 10.1016/j.biochi.2013.09.016.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Jackman, Shaun D et al. (2015). Organellar Genomes of White Spruce (Picea glauca): Assembly and Annotation. *Genome biology and evolution* 8 (1), pp. 29–41. ISSN: 1759-6653. DOI: 10.1093/gbe/evv244.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*. Ed. by B. Schölkopf, C. Burges, and A. Smola. Cambridge, MA: MIT Press. Chap. 11, pp. 169–184.

Lin, X et al. (1999). Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature* 402 (6763), pp. 761–768. ISSN: 0028-0836. DOI: 10.1038/45471.

Liu, Huitao et al. (2011). Comparative analysis of mitochondrial genomes between a wheat K-type cytoplasmic male sterility (CMS) line and its maintainer line. *BMC genomics* 12, p. 163. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-163.

Mayer, K et al. (1999). Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. *Nature* 402 (6763), pp. 769–777. ISSN: 0028-0836. DOI: 10.1038/47134.

Nystedt, Björn et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497 (7451), pp. 579–584. ISSN: 1476-4687. DOI: 10.1038/nature12211.

Ogihara, Y et al. (2002). Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molecular genetics and genomics : MGG* 266 (5), pp. 740–746. ISSN: 1617-4615. DOI: 10.1007/s00438-001-0606-9.

Rice, Danny W et al. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. *Science (New York, N.Y.)* 342 (6165), pp. 1468–1473. ISSN: 1095-9203. DOI: 10.1126/science.1246275.

Rivarola, Maximo et al. (2011). Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PloS one* 6 (7), e21743. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021743.

Salanoubat, M et al. (2000). Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. *Nature* 408 (6814), pp. 820–822. ISSN: 0028-0836. DOI: 10.1038/35048706.

Sato, S, Y Nakamura, T Kaneko, E Asamizu, and S Tabata (1999). Complete structure of the chloroplast genome of Arabidopsis thaliana. *DNA research : an international journal for rapid publication of reports on genes and genomes* 6 (5), pp. 283–290. ISSN: 1340-2838.

Sloan, Daniel B, Andrew J Alverson, John P Chuckalovcak, Martin Wu, David E McCauley, Jeffrey D Palmer, and Douglas R Taylor (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS biology* 10 (1), e1001241. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001241.

Sloan, Daniel B, Andrew J Alverson, Martin Wu, Jeffrey D Palmer, and Douglas R Taylor (2012). Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus Silene. *Genome biology and evolution* 4 (3), pp. 294–306. ISSN: 1759-6653. DOI: 10.1093/gbe/evs006.

Smit, AFA, R Hubley, and P Green (2013-2015). *RepeatMasker Open-4.0*. URL: http://www.repeatmasker.org.

Tabata, S et al. (2000). Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. *Nature* 408 (6814), pp. 823–826. ISSN: 0028-0836. DOI: 10.1038/35048507.

Talavera-López, Carlos (2014). *Statistics of Raw Data for Comparative Analysis*. [Accessed 27 March 2017]. Internal project wiki.

Tange, O. (2011). GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* 36.1, pp. 42–47. DOI: http://dx.doi.org/10.5281/zenodo.16303. URL: http://www.gnu.org/s/parallel.

Theologis, A et al. (2000). Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. *Nature* 408 (6814), pp. 816–820. ISSN: 0028-0836. DOI: 10.1038/35048500.

Tuskan, G A et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)* 313 (5793), pp. 1596–1604. ISSN: 1095-9203. DOI: 10.1126/science.1128691.

Umeå Plant Science Centre (2013). *Whole genome sequencing of the 20Gbp Norway spruce (Picea abies) genome*. European Nucleotide Archive. URL: http://www.ebi.ac.uk/ena/data/view/PRJEB1822.

Unseld, M, J R Marienfeld, P Brandt, and A Brennicke (1997). The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. *Nature genetics* 15 (1), pp. 57–61. ISSN: 1061-4036. DOI: 10.1038/ng0197-57.

Wang, Wenqin and Joachim Messing (2011). High-throughput sequencing of three Lemnoideae (duckweeds) chloroplast genomes from total DNA. *PloS one* 6 (9), e24670. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0024670.

Wu, Chung-Shien, Ya-Nan Wang, Shu-Mei Liu, and Shu-Miaw Chaw (2007). Chloroplast genome (cpDNA) of Cycas taitungensis and 56 cp protein-coding genes of Gnetum parvifolium: insights into cpDNA evolution and phylogeny of extant seed plants. *Molecular biology and evolution* 24 (6), pp. 1366–1379. ISSN: 0737-4038. DOI: 10.1093/molbev/msm059.

Appendix A

Investigation of feature noise as function of contig length

To build confidence in the approach used, an artificial classification problem of classifying known *A. thaliana* contigs was set up. The main learning outcome from this experiment was, arguably, knowledge about the stability of feature values as a function of contig length.

A.1 Experiment using a reference species

A.1.1 Data used

The full genome and aligned reads of one *Arabidopsis thaliana* variant (Bur-0) from the *19 genomes project* was used (Gan et al. 2011).

A.1.2 Method

First, the `bur_0_A.bam` file was downloaded from the 19 genomes web site. (It appears to be currently unavailable. For reference, the file size of my local copy is 4074261728 bytes). Also the reference genome `bur_0.v7.PR_in_lowercase.fas` was downloaded and split into consecutive fragments of 330, 1000 and 3300 bp, respectively.

Next, FASTQ reads was extracted from the BAM file using `bamUtil` and aligned to the fragmented copies of the reference genome using BWA (version 0.6.2, using the `index, aln` and `sampe/samse` subcommands). Both single-end and paired-end reads were aligned. Coverage statistics was gathered using Samtools (version 1.1).

The plots were generated using a custom script. The marker color was set using the known (true) target contig from the alignment.

A.1.3 Findings

The plots show that longer contigs are significantly less noisy, giving a much better separation. Note that the CpG feature was not used in the final classification pipeline. It clearly hold true for the GC feature though. See figure A.1.

It's worth noting that there were quite a few outliers in terms of coverage (see figure A.2, contig length shown: 1000 bp). The effect was most pronounced with the shortest contigs but true for the longest as well (not shown in the figure). Repeats are thought to be the cause, from the project web site:

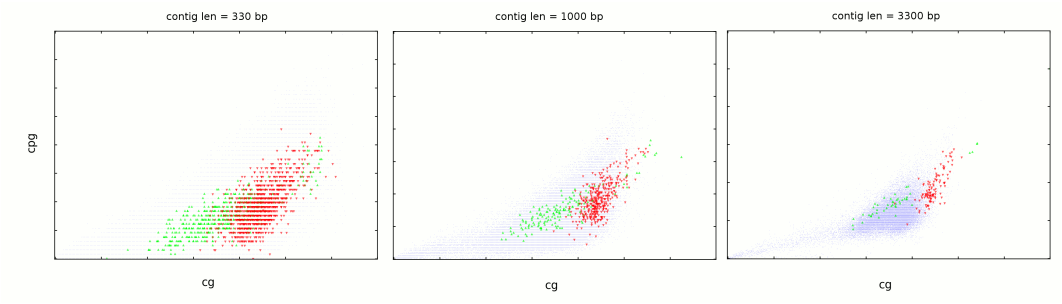

Figure A.1: *A. thaliana control experiment: CG-CpG features. The plots show the CG and CpG features for labeled contigs of lengths 330, 1000 and 3300 bp, respectively, from one sample of A. thaliana. Color code: red = mitochondrial, green = chloroplast, blue = nuclear. Note that for the short contigs there is unfortunately a rasterization effect due to the small number of possible values of the CpG fraction feature.*

> Lower confidence "uncovered regions" are [...] generally repetitive regions that were uncovered when the reads were re-mapped to the final assemblies. These regions therefore may be deleted or may correspond to places where reads could map to more than one locus. — `http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/`

Thus we have found an interesting indication as to how much the coverage is expected to vary in actual mitochondrial contigs. This plot (A.2) clearly illustrates how difficult it is to rely mainly on GC% and coverage to classify the contigs. It also helped inform the decision to set the contig length cutoff at 500 bp.

A.1.4 Investigation of feature-space plots from the results

Studying how the variation in a feature (e.g. GC%) decreases with increasing length gives a nice intuition as to how "reliable" it is when used for contigs of a particular length.

See appendix D for these plots.

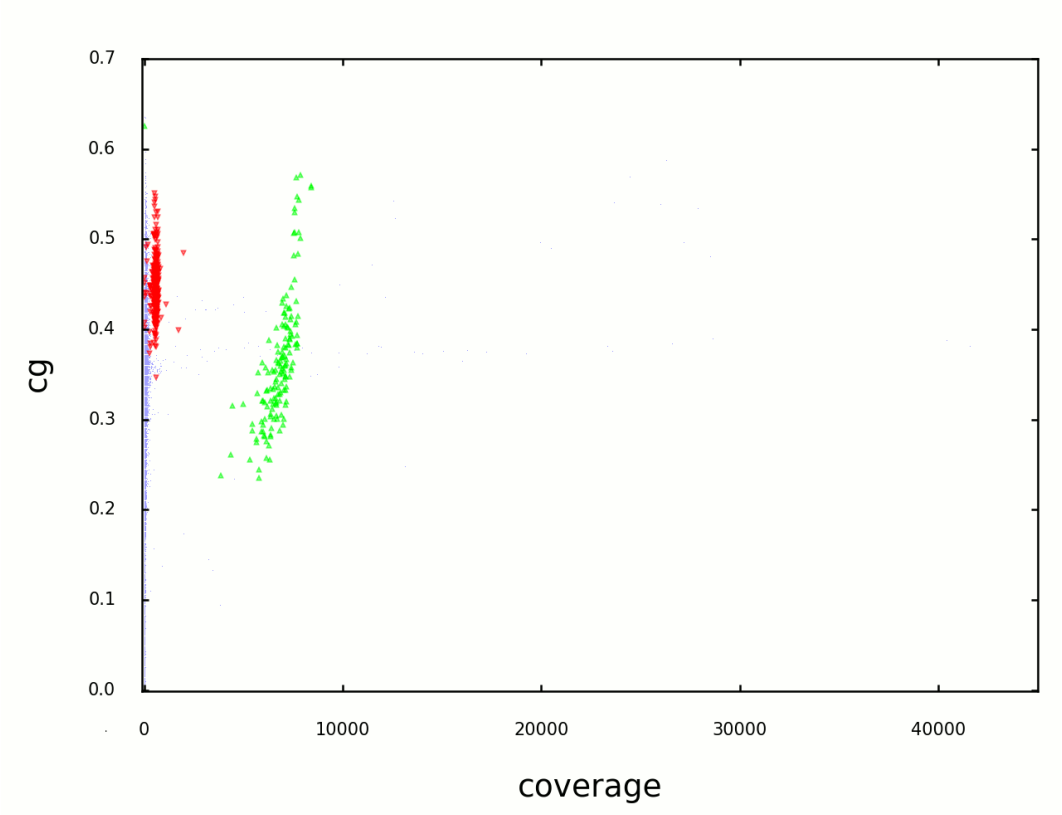

Figure A.2: *A. thaliana control experiment: Coverage variation. The plots show the coverage variation of labeled contigs (1000 bp long) from one sample of A. thaliana. Color code: red = mitochondrial, green = chloroplast, blue = nuclear. The dots representing the nuclear contigs are unfortunately very small. Note e.g. a narrow band of nuclear contigs at* $cg \approx 0.37$ *with widely varying coverage.*

Appendix B

Dead ends and paths not chosen

Here we describe some of the unfruitful attempts and misunderstandings. Hopefully this is of some use or (at least) entertainment for the reader.

B.1 A mitochondrion in disguise

At one stage in the project I encountered major problems with ambiguous classifications of contigs. The problem was eventually tracked down to a large duplication of the *A. thaliana* mitochondrion into the nuclear chromosome 2, see figure B.1. Initially I was very confused about the findings as I wasn't aware of the (practical) possibility of such a large and essentially identical duplication. I find it somewhat comforting that I wasn't alone in being surprised by this:

> More unexpected is what appears to be a recent insertion of a continuous stretch of 75% of the mitochondrial genome into chromosome 2. [...] This insertion is much larger than any of the previously reported organelle-nuclear transfers, and is 99% identical to the mitochondrial genome [...] (Lin et al. 1999).

Since we had four other nuclear chromosomes as well as the nuclear genome of P. trichocarpa, it wasn't deemed worth the effort to "dissect out" the non-duplicated part of chromosome 2. Instead we just ignored the whole chromosome in our analysis.

B.2 ORFs

Attempts were made to detect open reading frames (ORFs) in the source contigs, with the intent of using the ORFs as a feature. These attempts were never followed through as this line of investigation was dropped when other features (mainly the kmer classifier) started to show more promise.

B.3 SNPs

Early on in the project, we tried using single-nucleotide polymorphisms (SNPs) as a feature. It wasn't useful as a feature, see figure B.2.

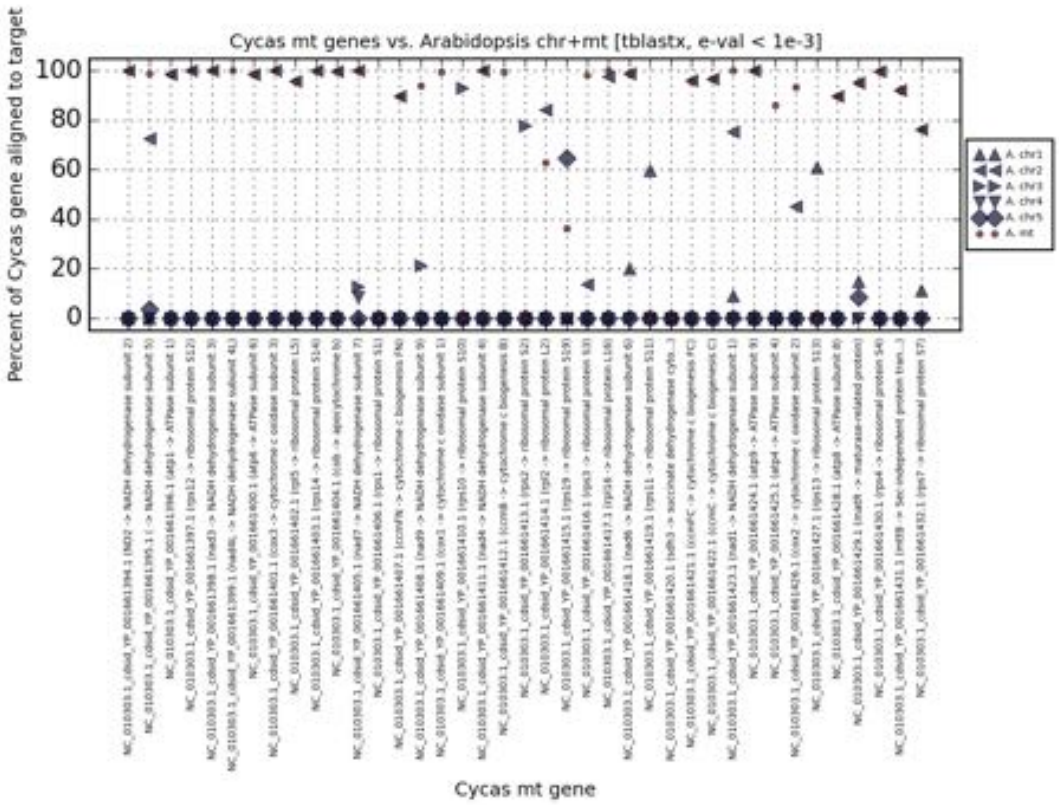

Figure B.1: *A. thaliana mitochondrial duplication investigation. A number of Cycas taitungensis mitochondrial genes TBLASTX-ed to various A. thaliana chromosomes. The plot shows the percentage of the gene that was aligned to the target. The key finding was that a large number of genes map equally well to the nuclear chromosome 2 and the mitochondrion.*

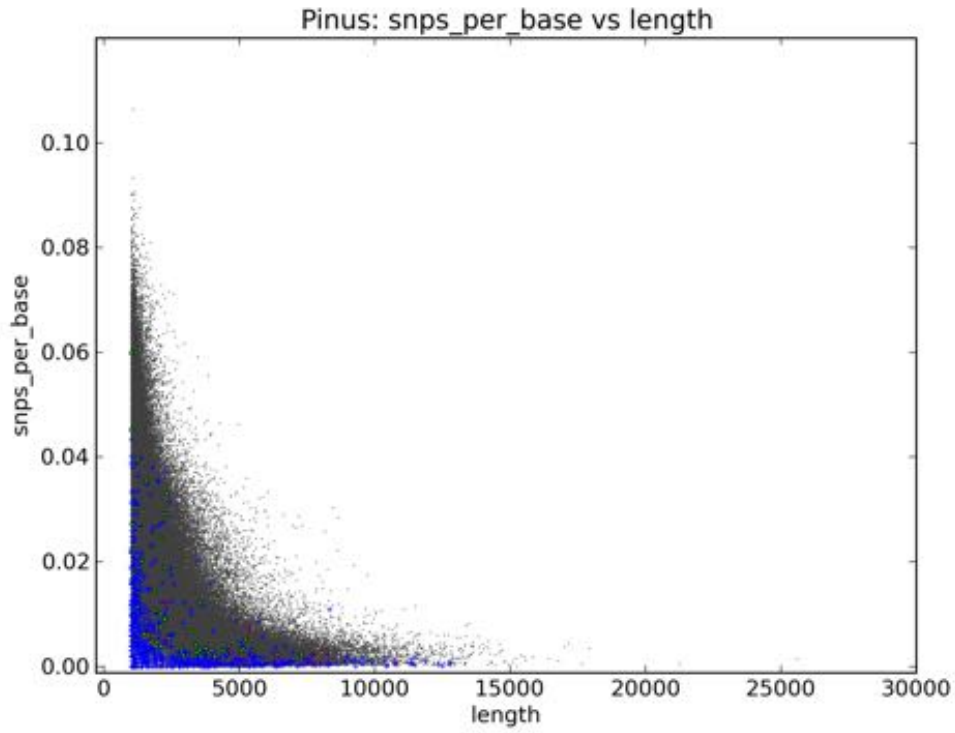

Figure B.2: *SNPs as a feature: an example feature-space plot. SNPs vs length for P. sylvestris. This feature was not used in the final classification. Color code: gray = unclassified contigs, blue = nuclear, red = mitochondrial, green = chloroplast. Classifications are most likely based on BLASTN searches.*

B.4 CpG% (instead of GC%)

Early on in the project CpG-percentage was used as feature. (CpG is simply a C followed by a G in a DNA sequence.) Instead we ended up using the related but distinct feature GC-percentage (i.e. simply what fraction of a sequence that is a C or a G). This was based on concerns that the CpG feature was biased towards gene-rich regions in general, which could cause false positives. However early tests did show some separation from using this feature, so it might have been premature to discard it. See figure B.3 for an example plot.

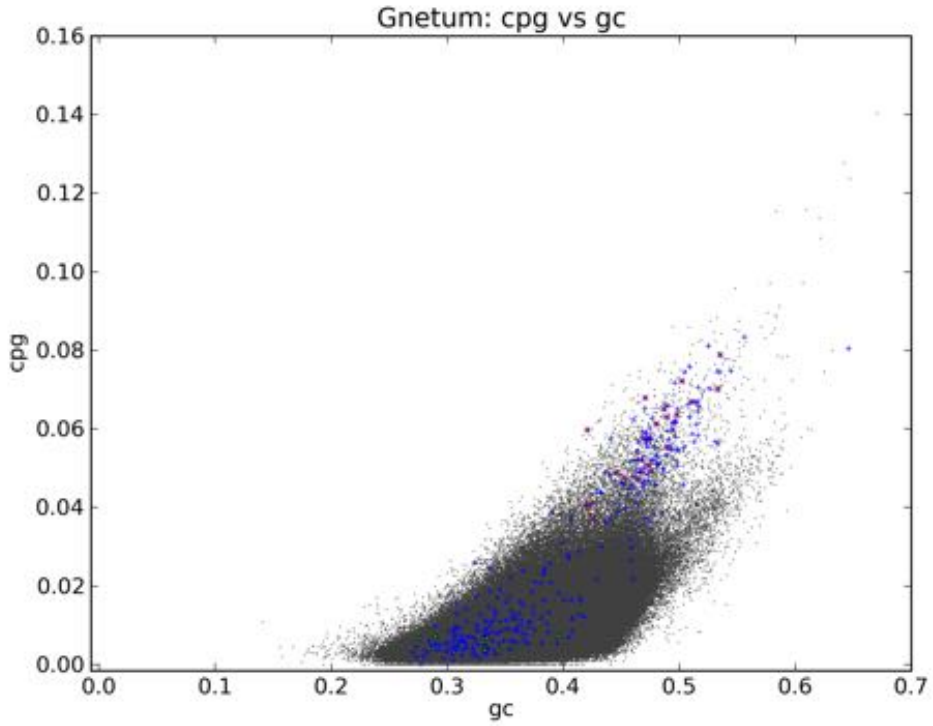

Figure B.3: *Pairwise feature plot of CpG% vs GC% for G. gnemon. This an early version of the feature-space plot type. Color code: gray = unclassified contigs, blue = nuclear, red = mitochondrial, green = chloroplast. Classifications are most likely based on BLASTN searches.*

B.5 Less important features: N% and masked%

The N% feature is the percentage of ambiguous nucleotides in a contig. The masked% is similarly the percentage masked out by RepeatMasker.

As can be seen from the feature-space plots of N% vs masked% (see figure B.4 through B.9), neither feature is particularly useful, with the possible exception of *T. baccata*. It's also worth noting the possibility that some repeat-heavy *P. sylvestris* contigs *may* have been misclassified as mitochondrial (the black rectangle (i.e. selected) contigs with 80% or more masked).

B.5.1 Extra comments on "N%"

"N%" appears to be a noisy feature. Speculation: this could be because there are in fact two underlying signals here: (1) mitochondrial contigs containing

N's because of the underlying complexity of the mitochondrial genome and (2) non-mitochondrial contigs having high coverage (i.e. our main source of false positives) are probably likely to be repeat heavy (this is what's giving them their high coverage), and the difficulty of assembling such contigs give rise to the N's.

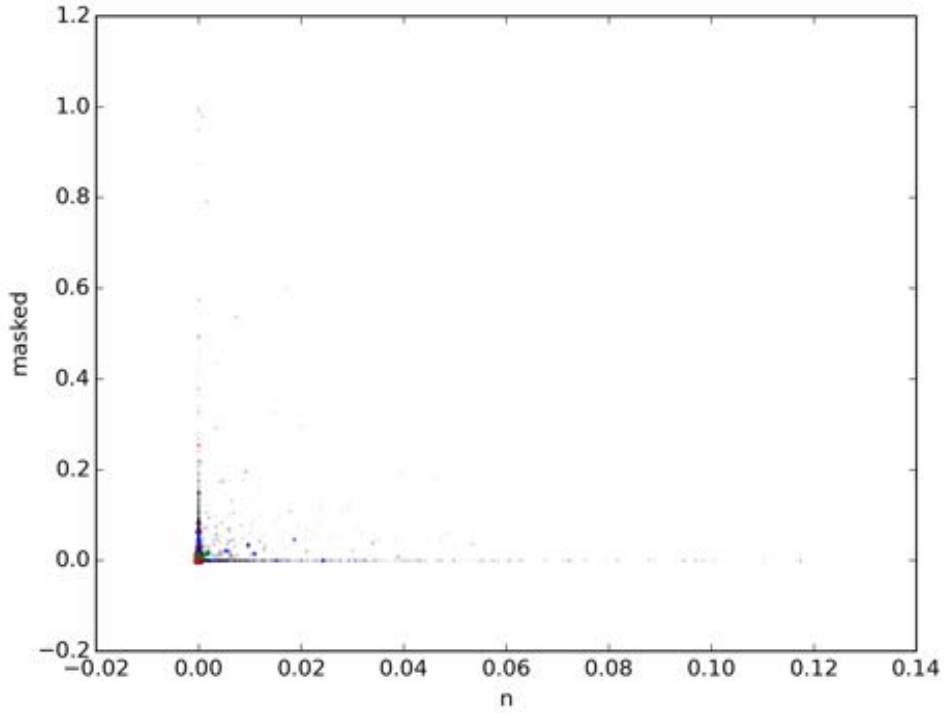

Figure B.4: *A. sibirica: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

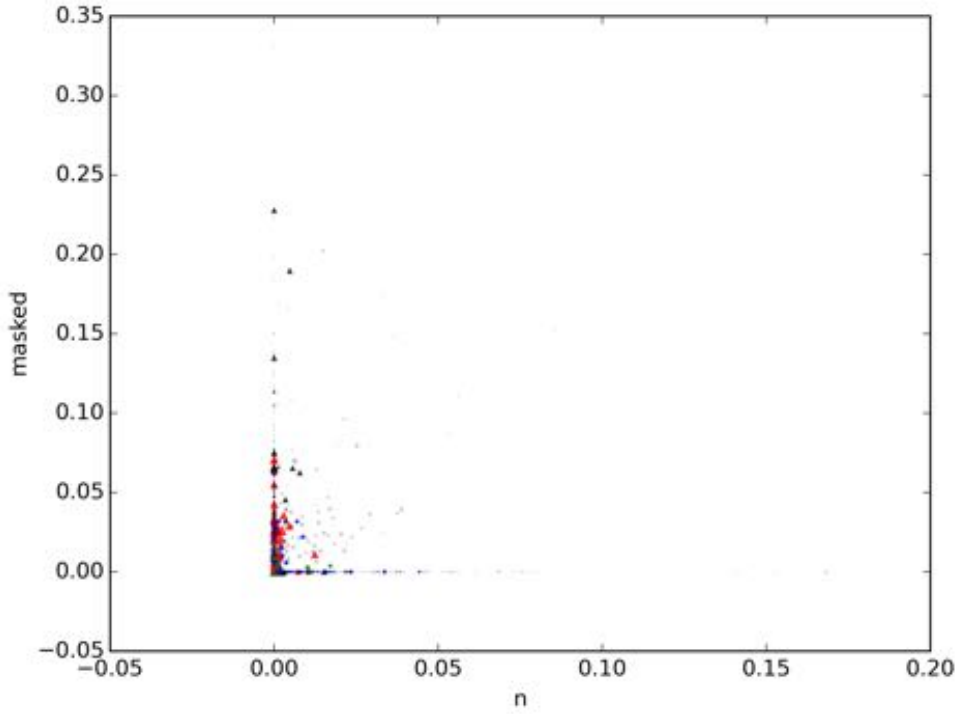

Figure B.5: *G. gnemon: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

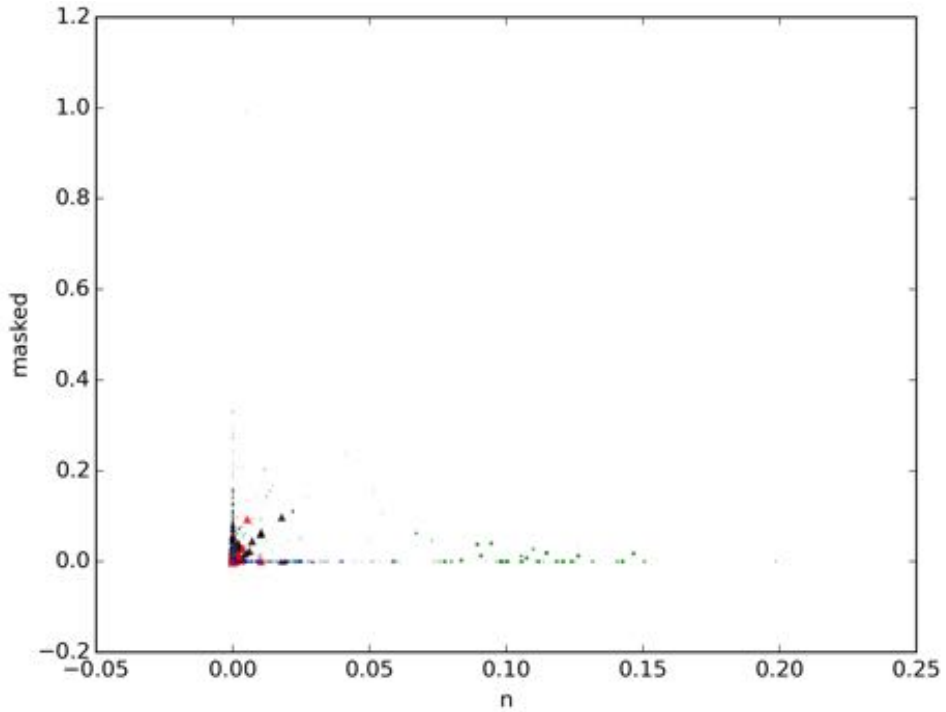

Figure B.6: *J. communis: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

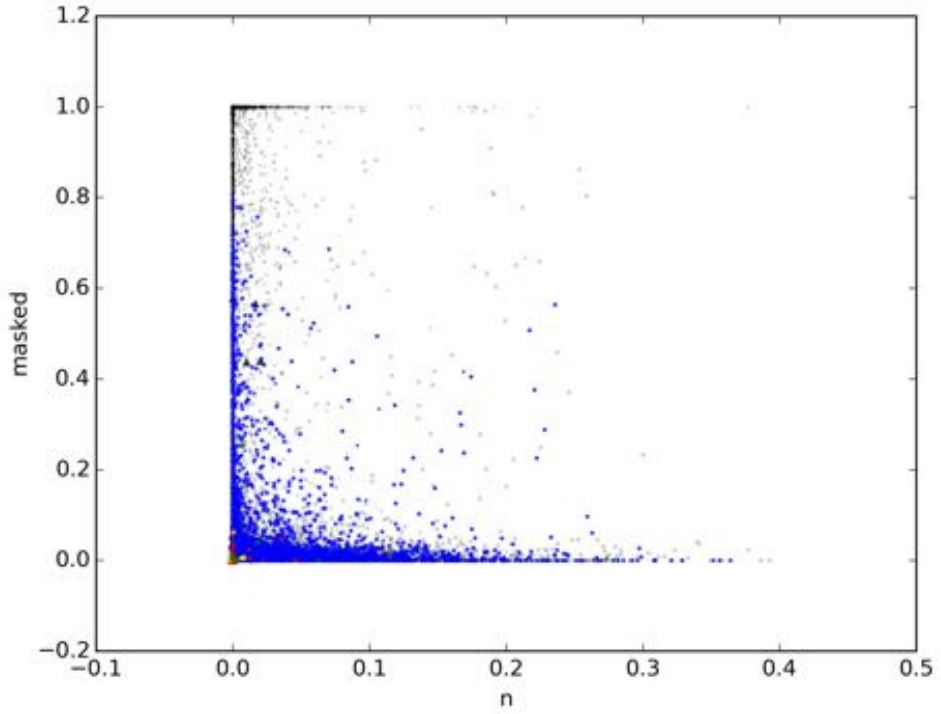

Figure B.7: *P. abies: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

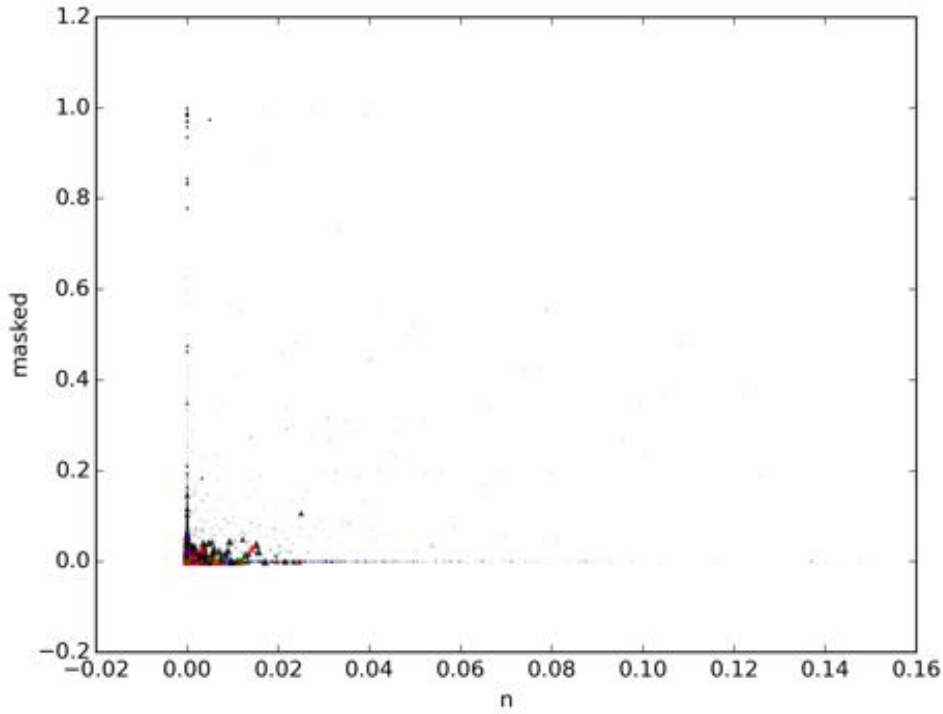

Figure B.8: *P. sylvestris: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

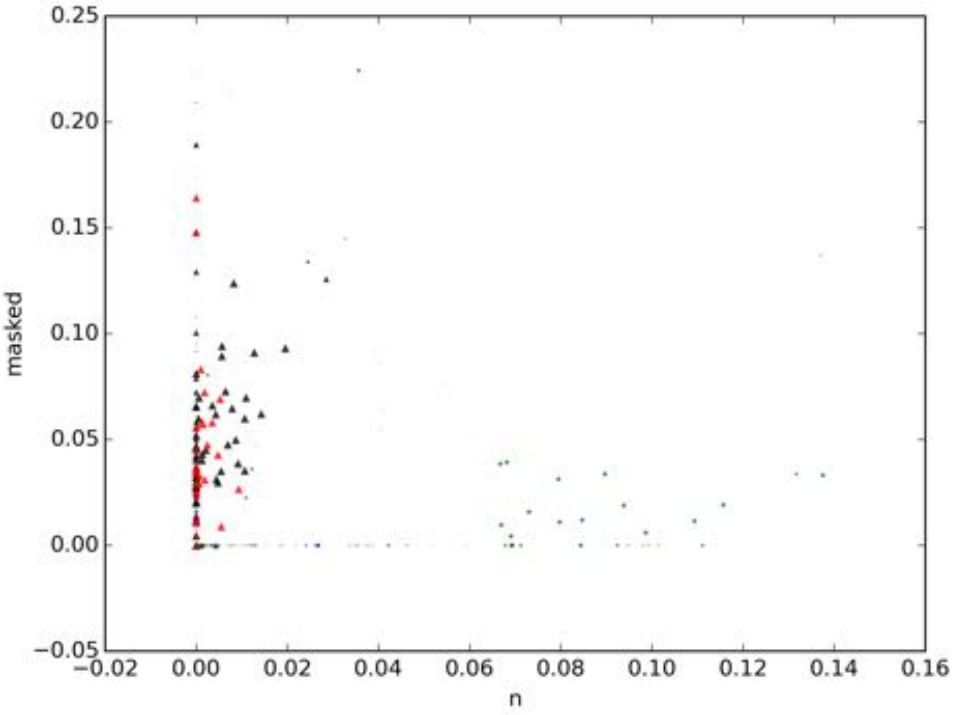

Figure B.9: *T. baccata: masked% vs N%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1.

Appendix C

Reference feature-space plots for all species

This appendix contains a reference of feature-space plots for all species. For an explaination of how to read these plots, see section 6.1

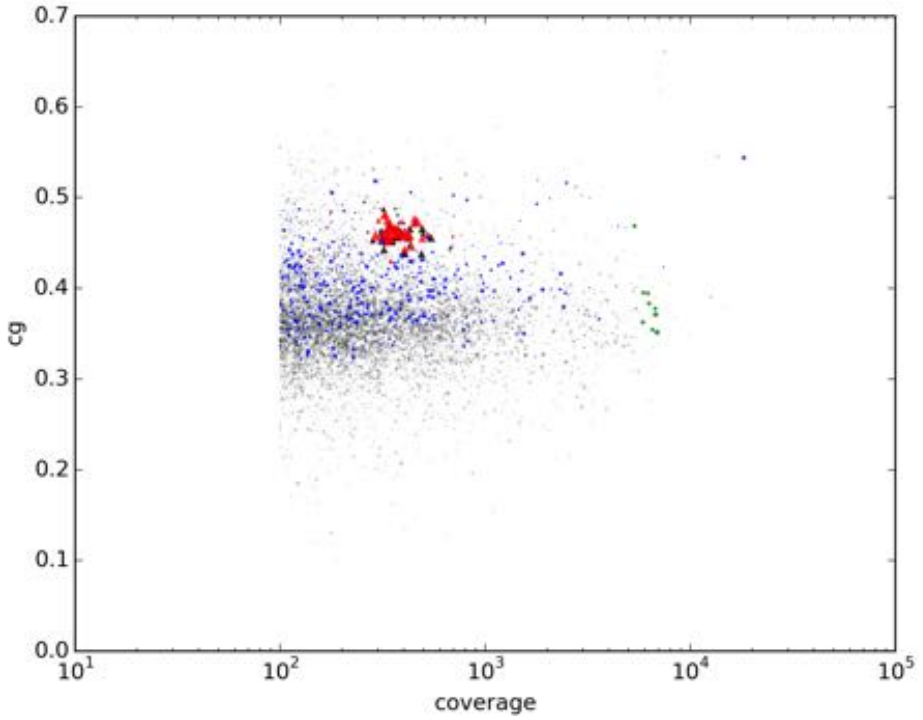

Figure C.1: *Abies sibirica: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

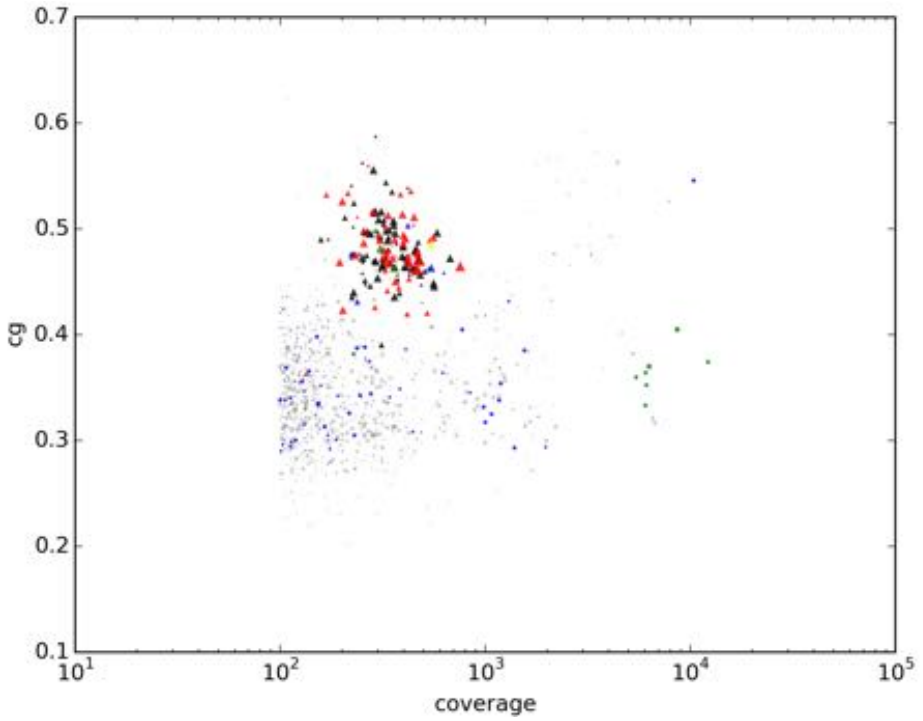

Figure C.2: *Gnetum gnemon: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

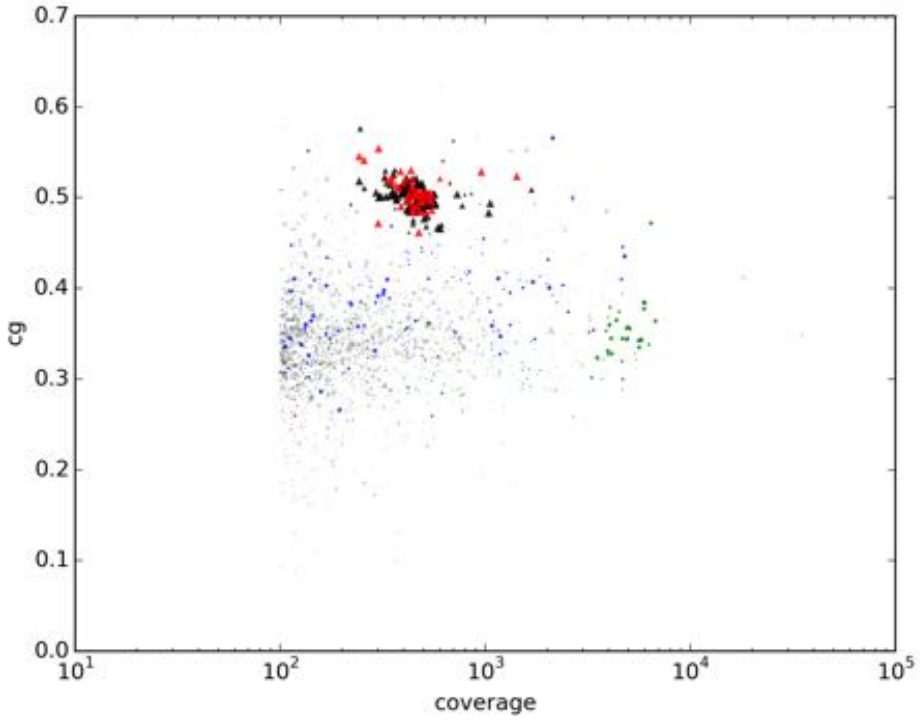

Figure C.3: *Juniperus communis: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

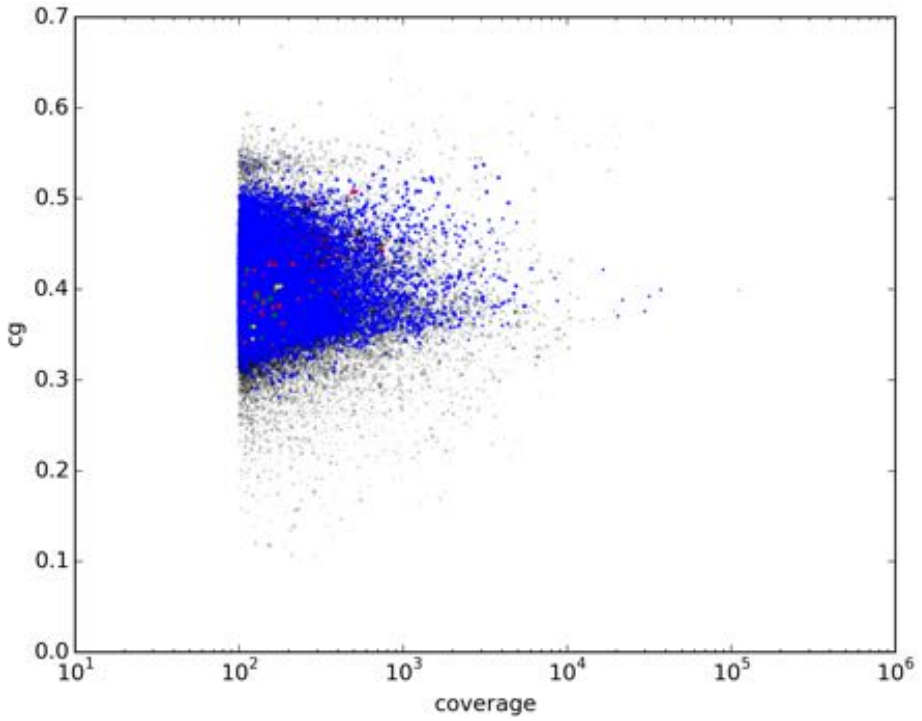

Figure C.4: *Picea abies: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

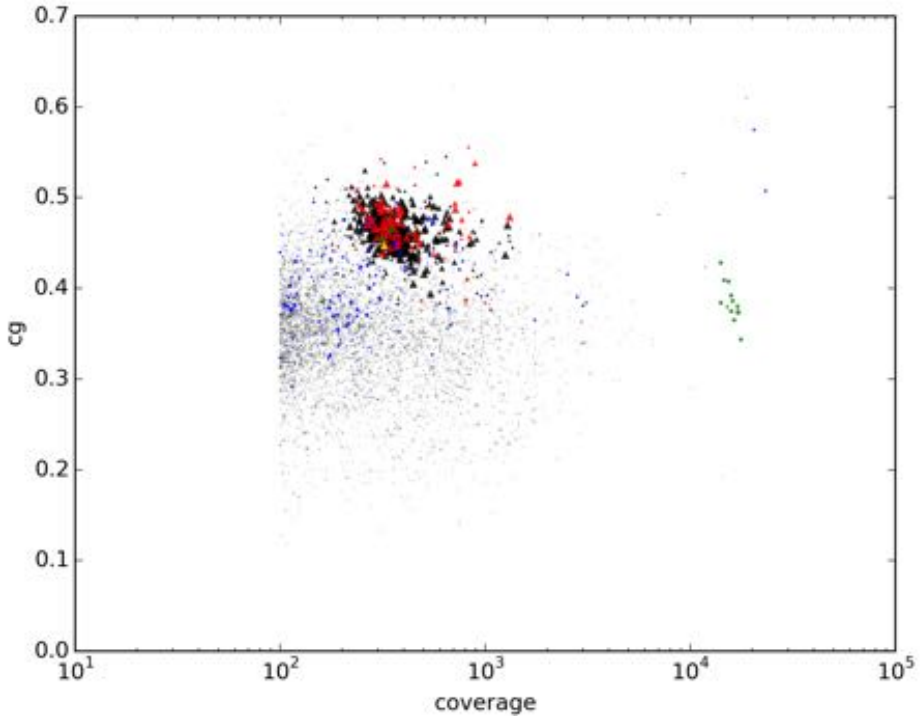

Figure C.5: *Pinus sylvestris: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

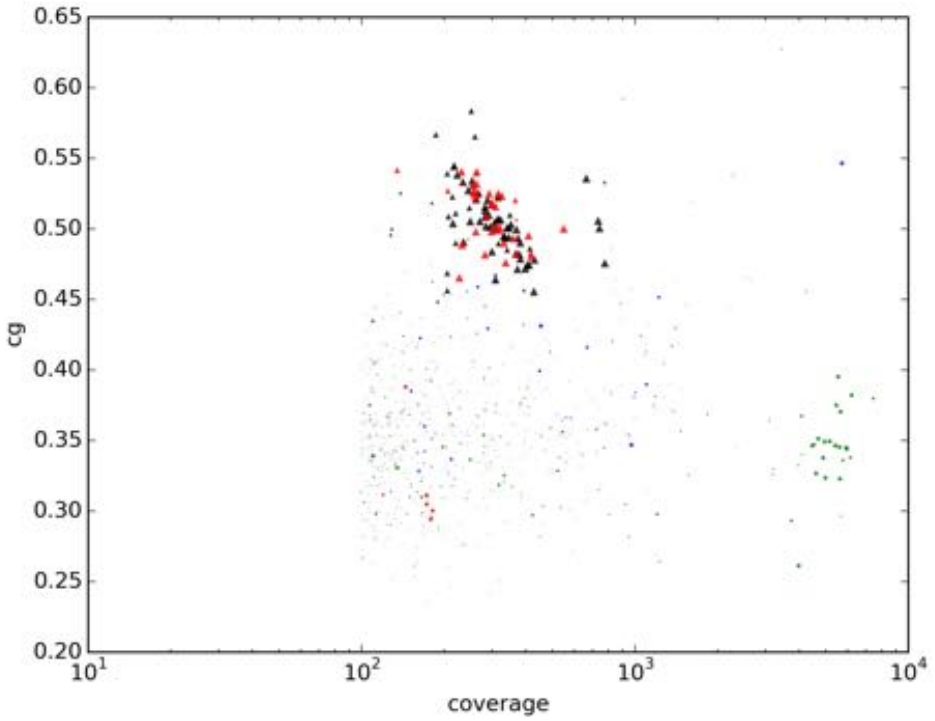

Figure C.6: *Taxus baccata: GC% vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

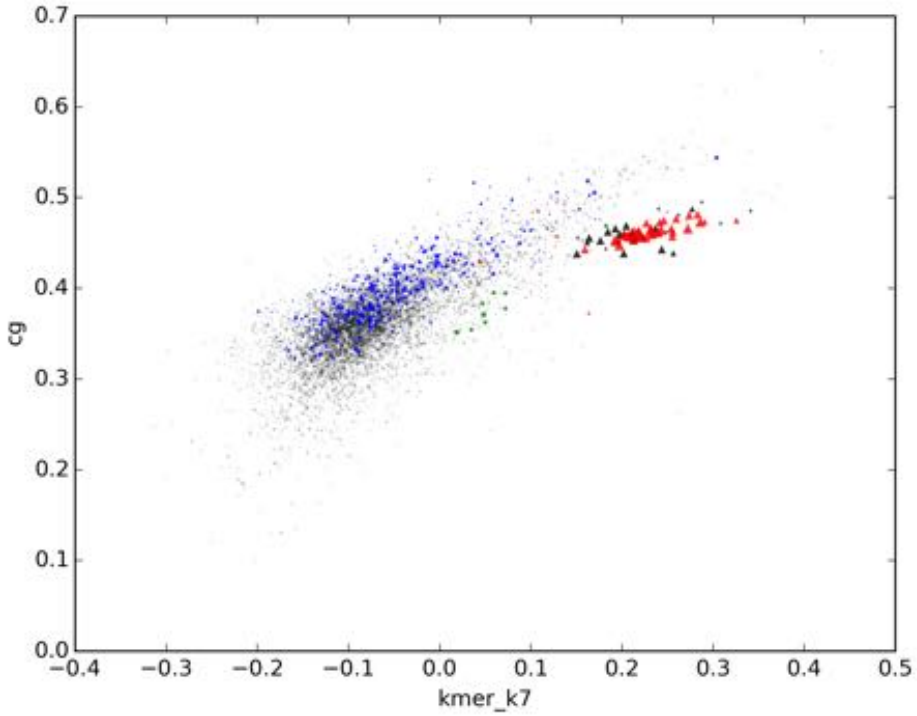

Figure C.7: *Abies sibirica: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

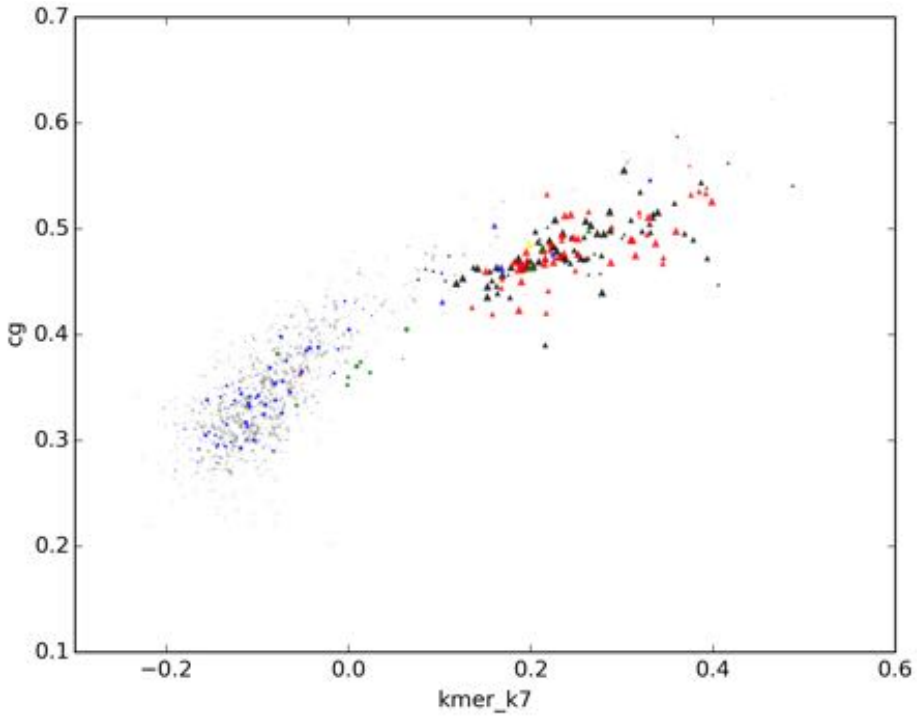

Figure C.8: *Gnetum gnemon: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

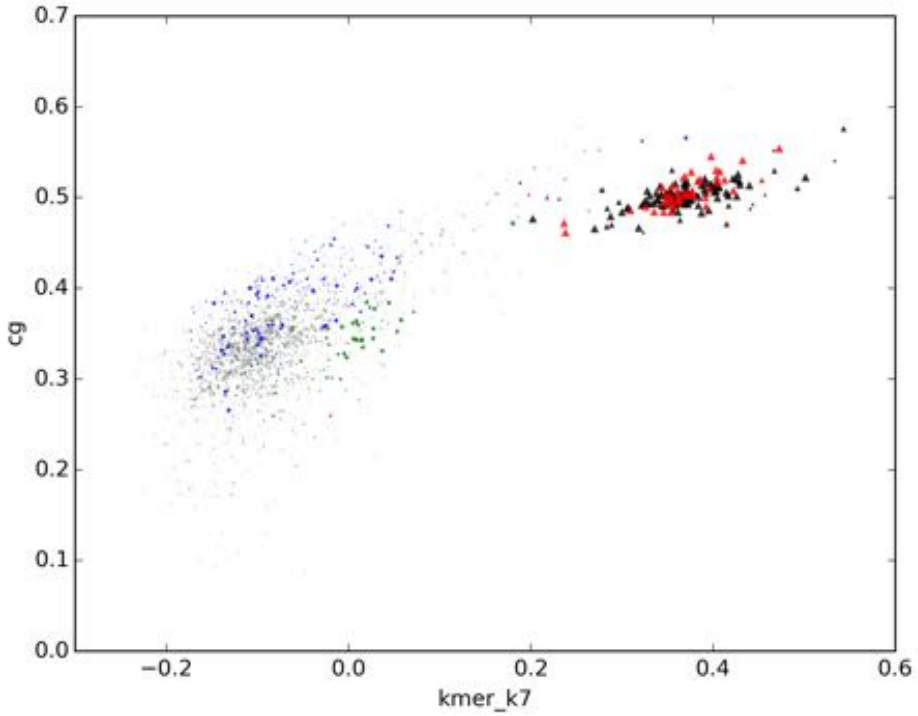

Figure C.9: *Juniperus communis: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

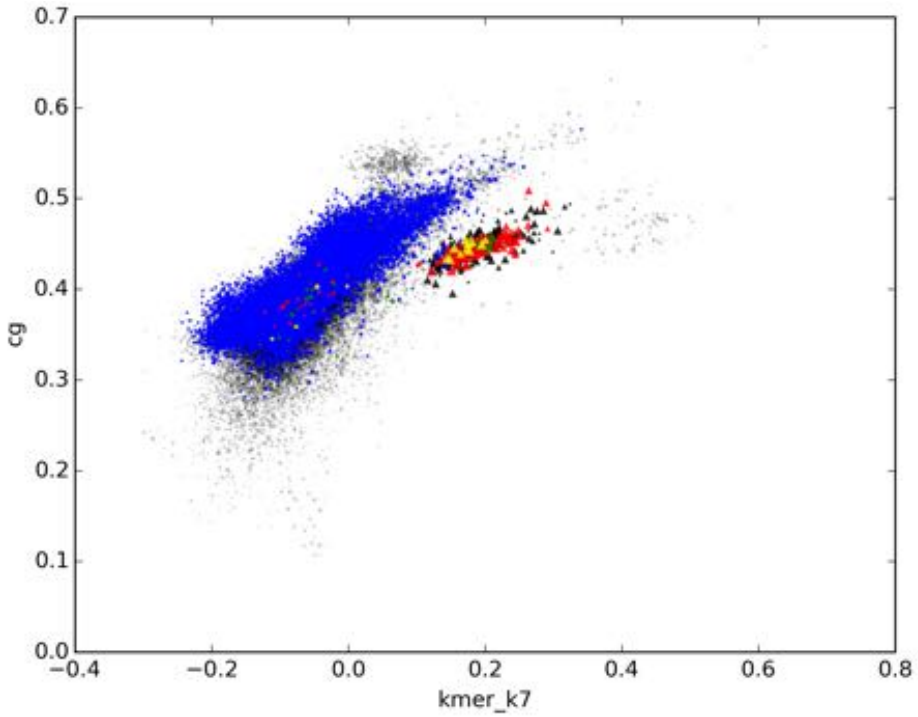

Figure C.10: *Picea abies: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

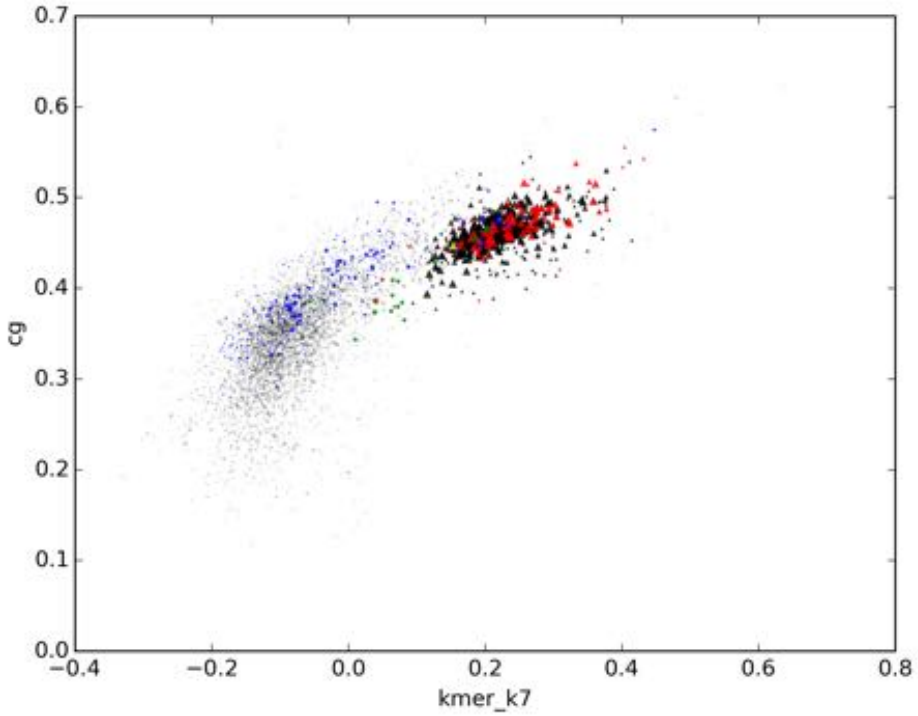

Figure C.11: *Pinus sylvestris: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

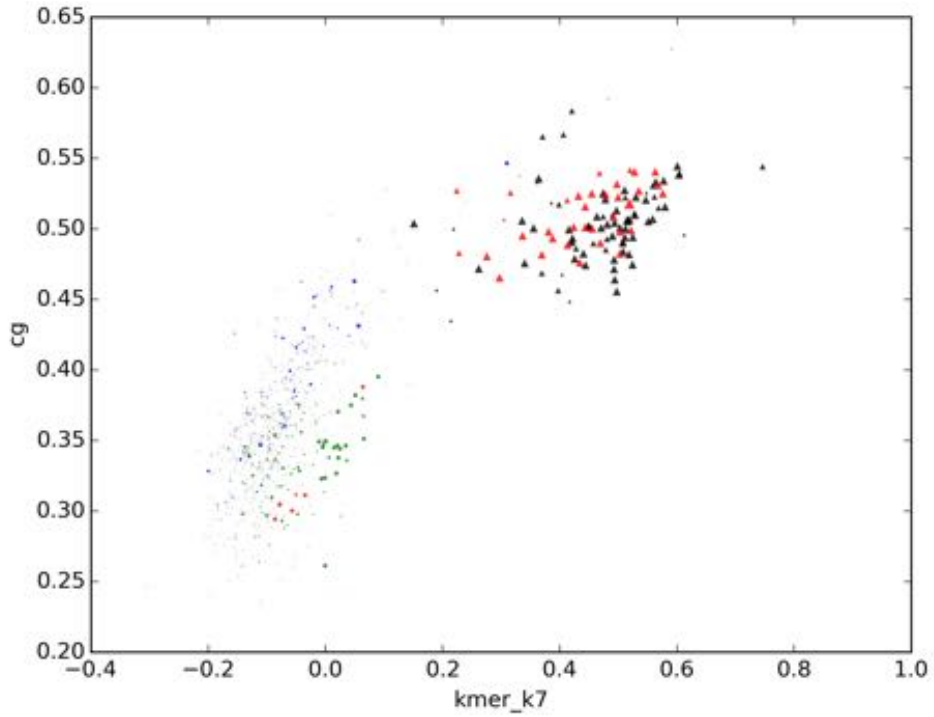

Figure C.12: *Taxus baccata: GC% vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

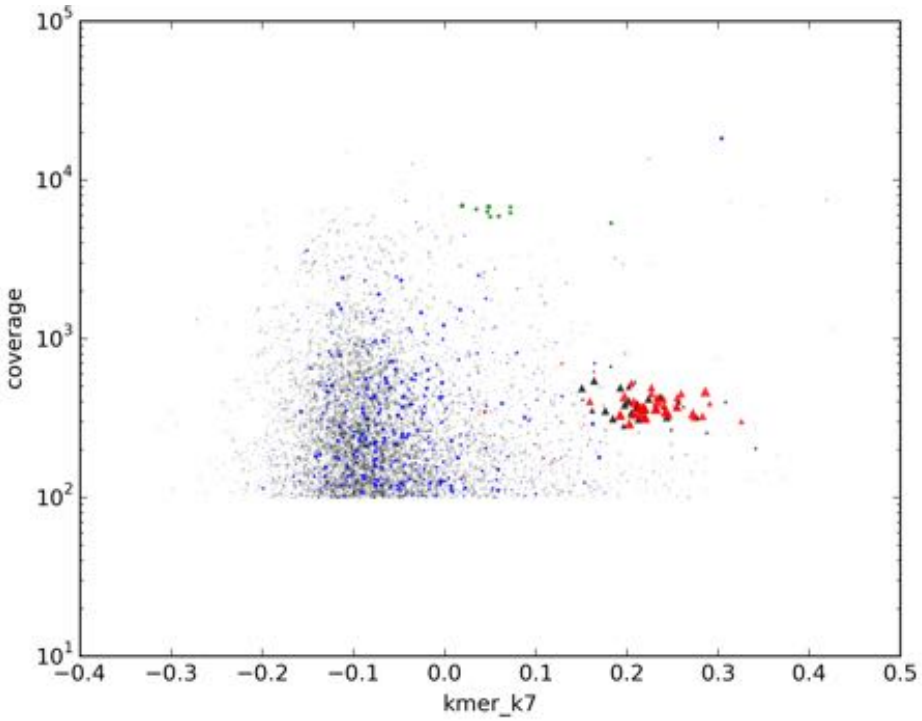

Figure C.13: *Abies sibirica: Coverage vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

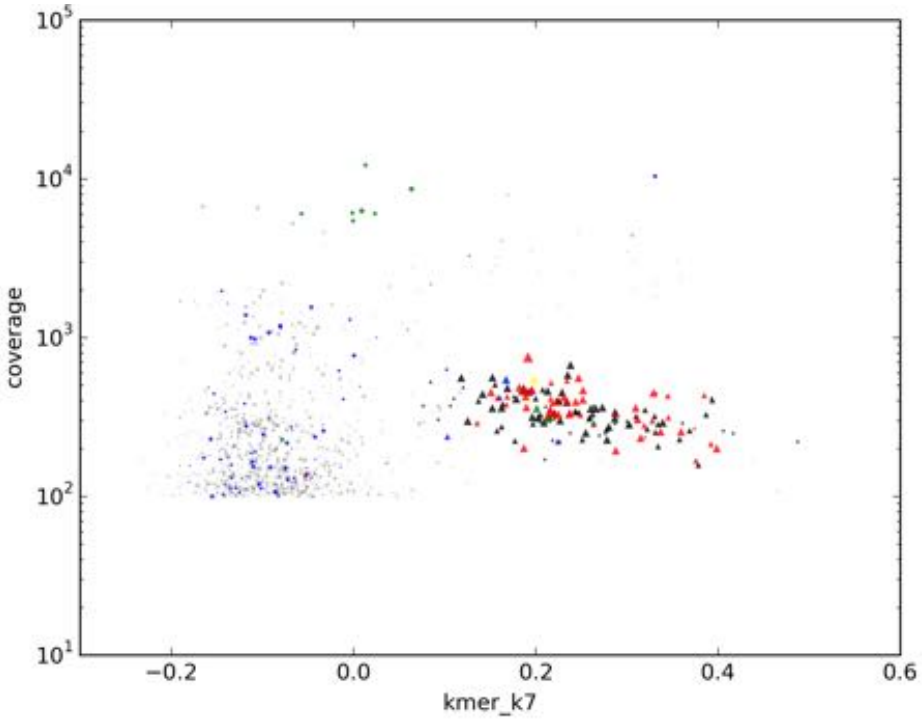

Figure C.14: *Gnetum gnemon: Coverage vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

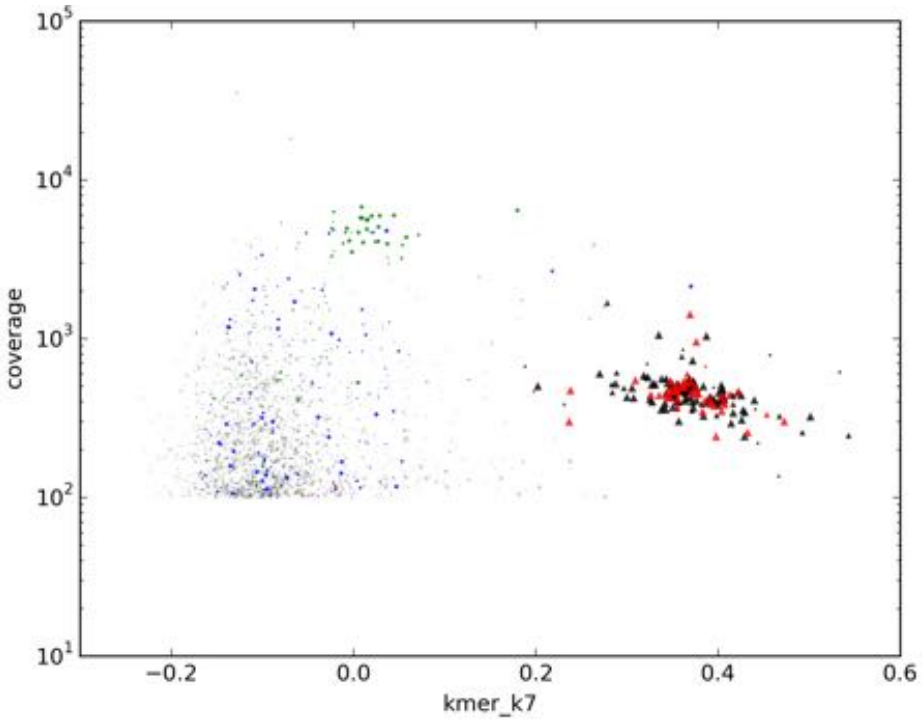

Figure C.15: *Juniperus communis: Coverage vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

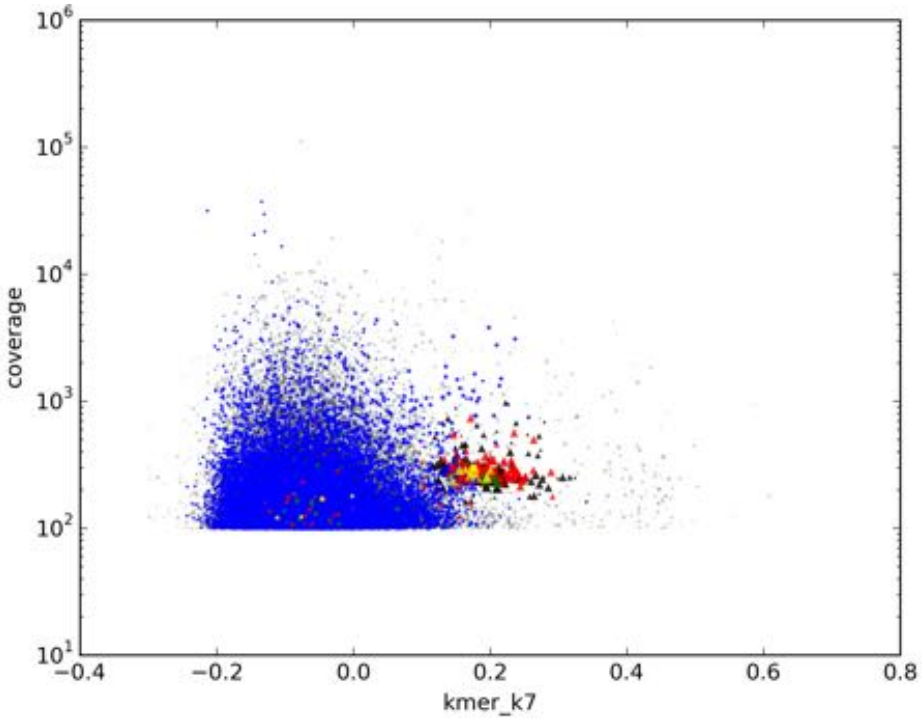

Figure C.16: *Picea abies: Coverage vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

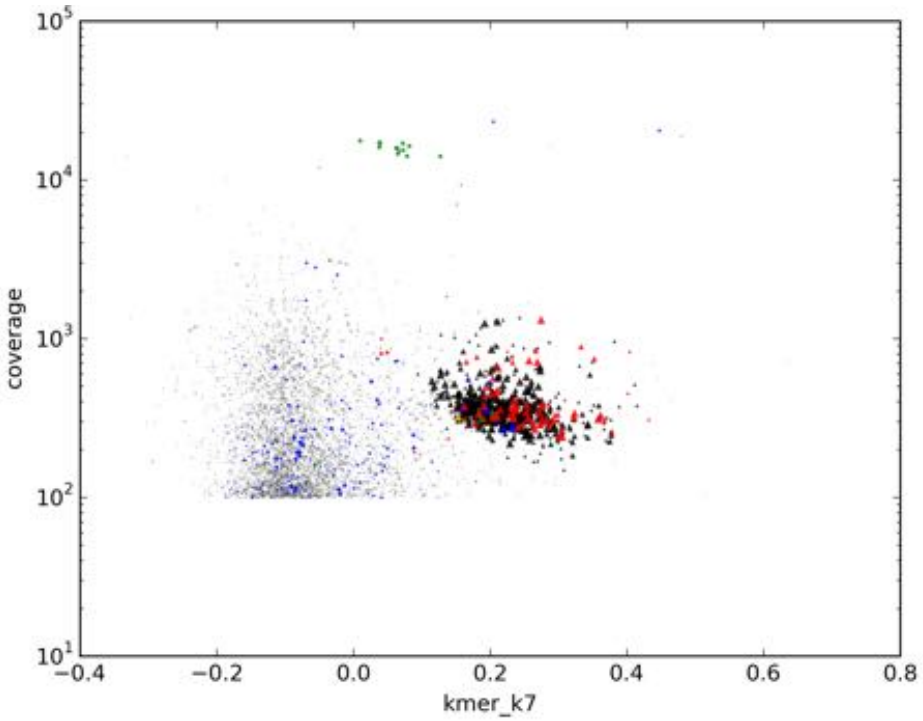

Figure C.17: *Pinus sylvestris: Coverage vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

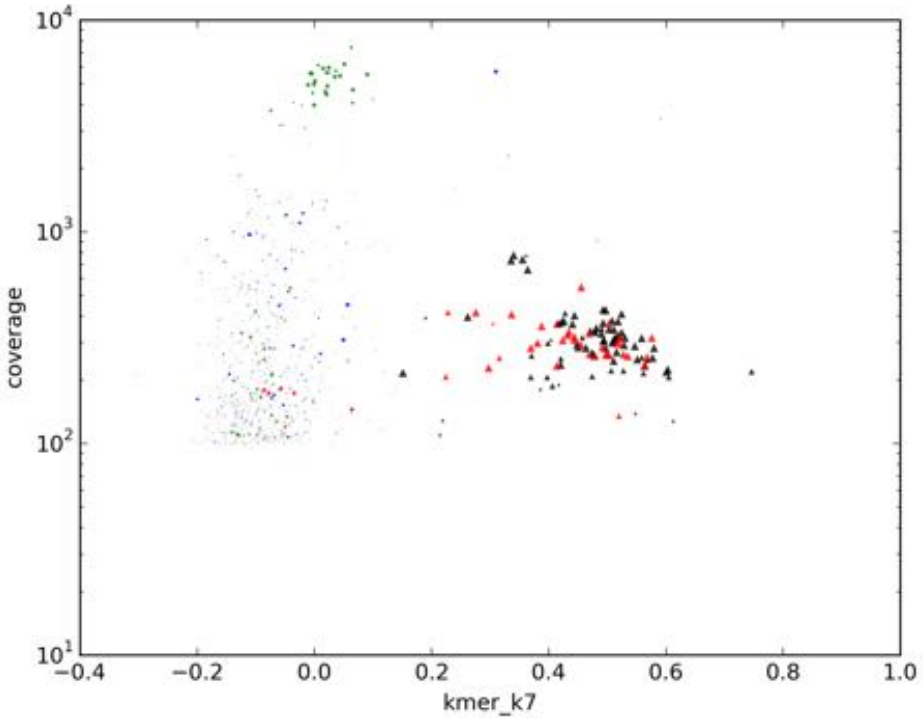

Figure C.18: *Taxus baccata: Coverage vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

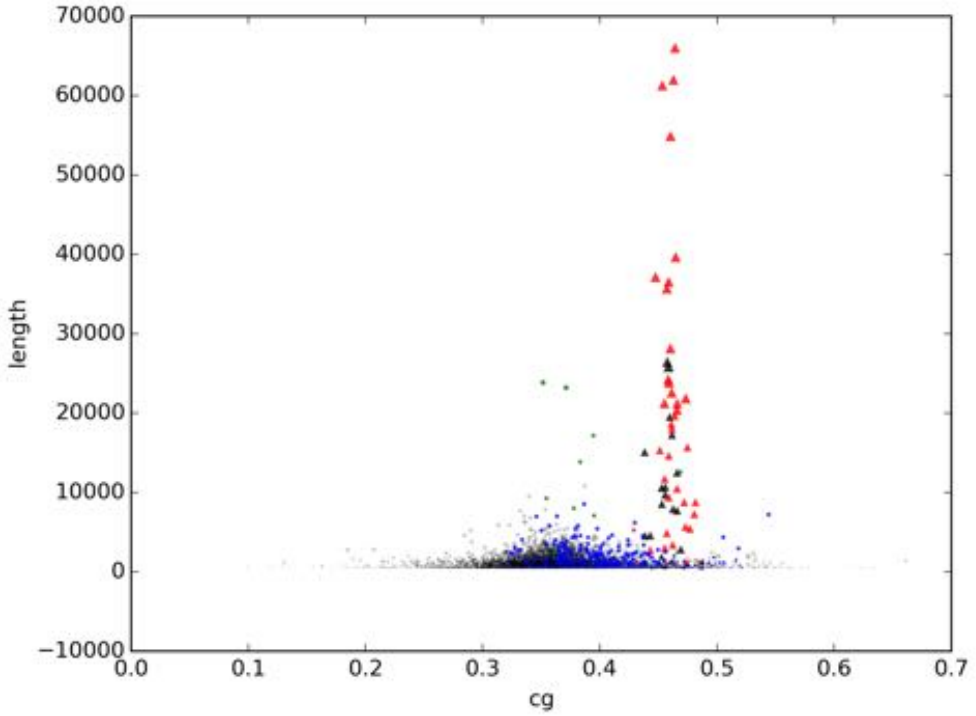

Figure C.19: *Abies sibirica: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

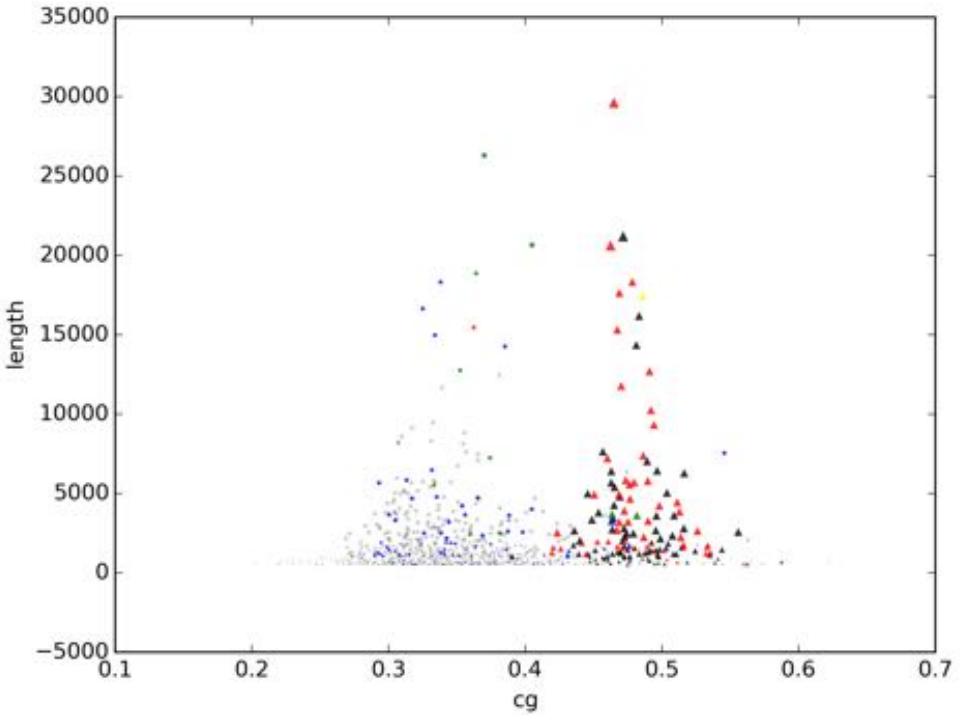

Figure C.20: *Gnetum gnemon: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

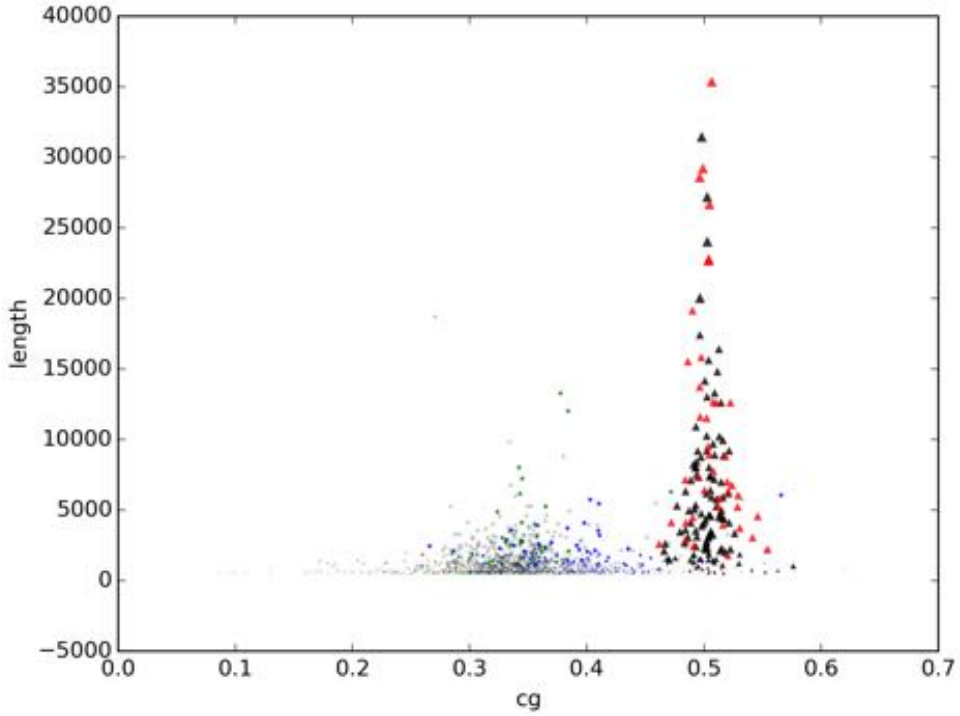

Figure C.21: *Juniperus communis: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

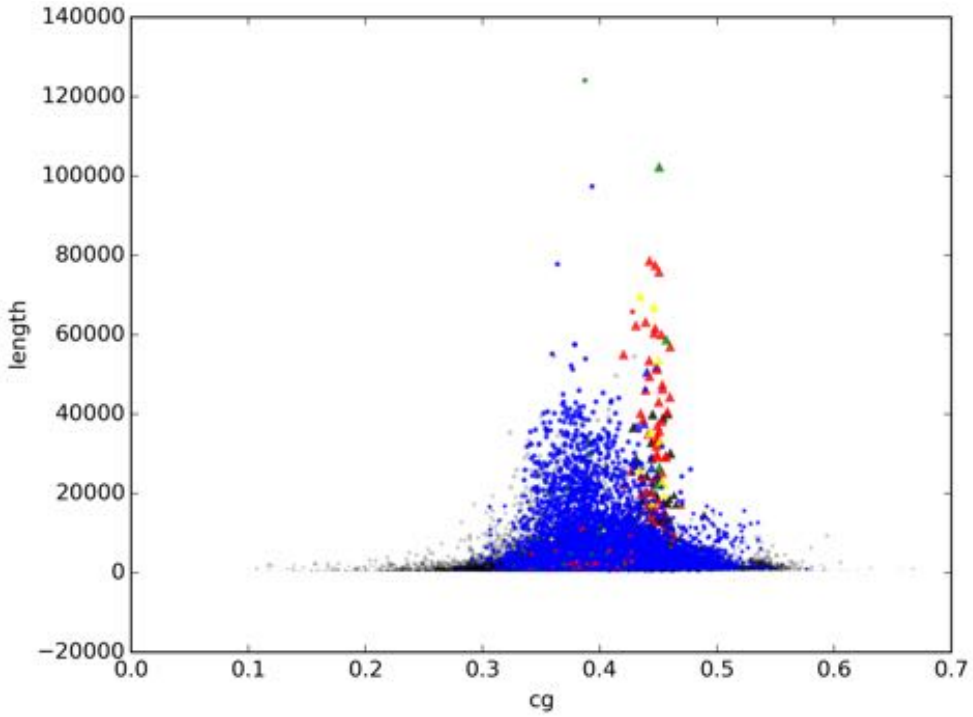

Figure C.22: *Picea abies: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

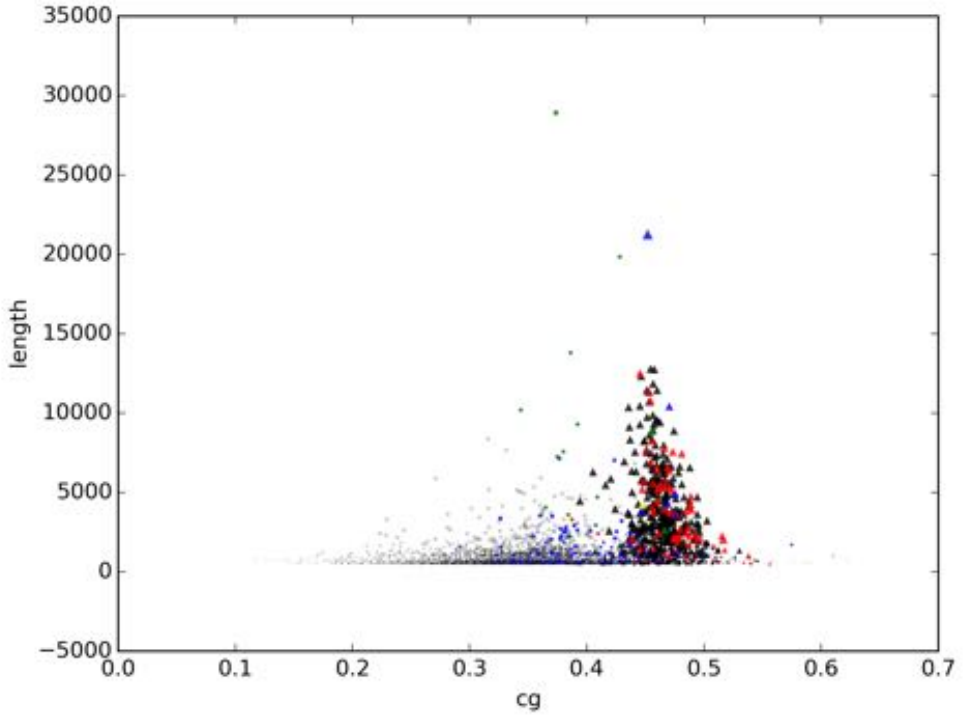

Figure C.23: *Pinus sylvestris: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

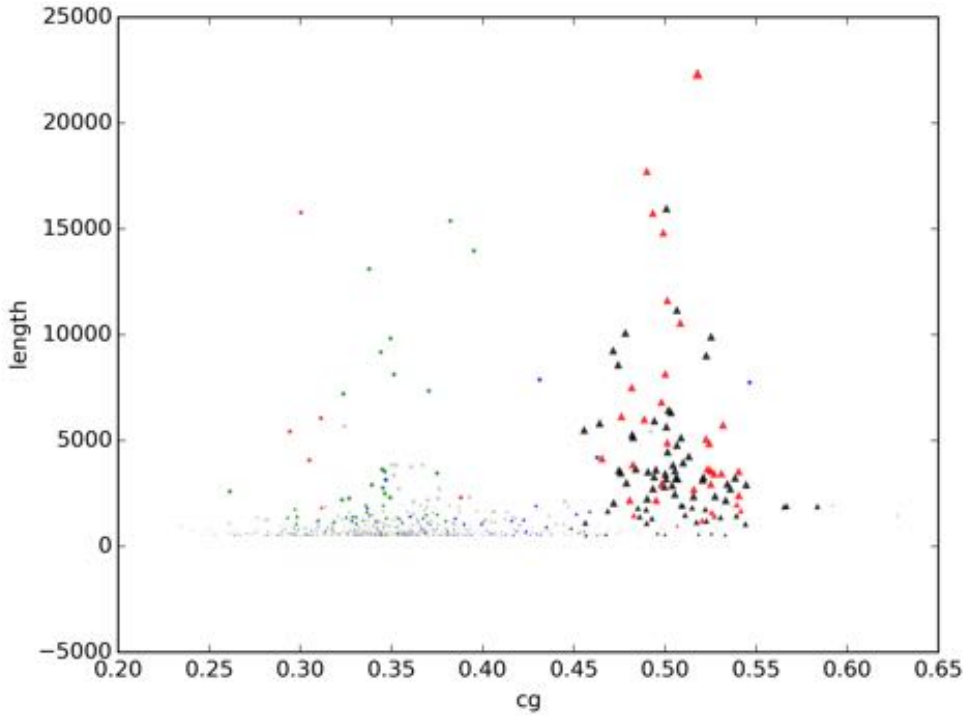

Figure C.24: *Taxus baccata: Length vs GC%.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

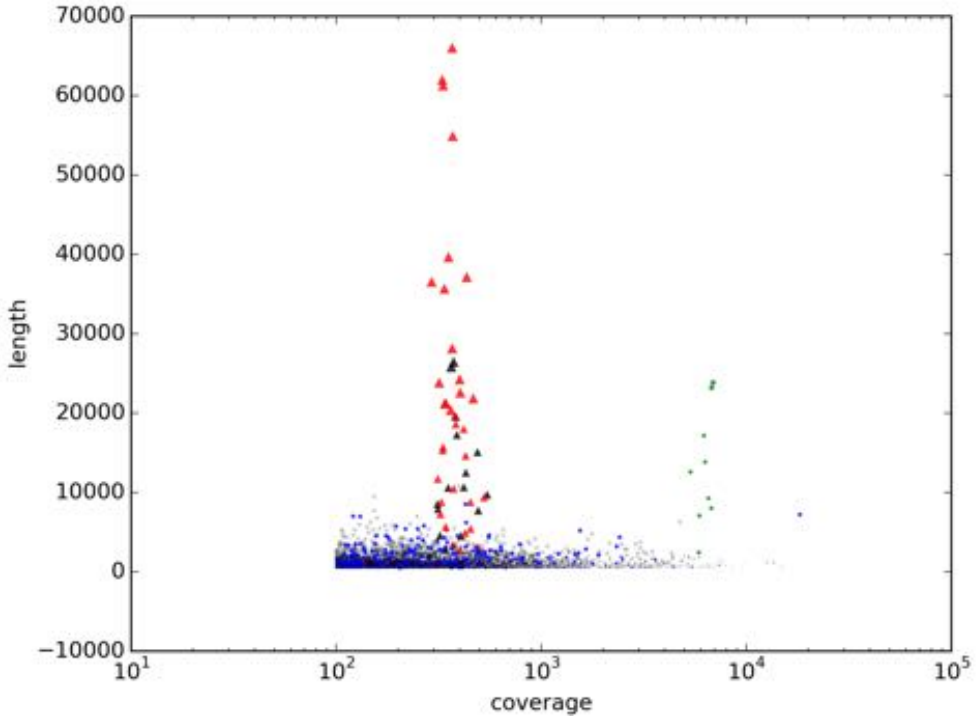

Figure C.25: *Abies sibirica: Length vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

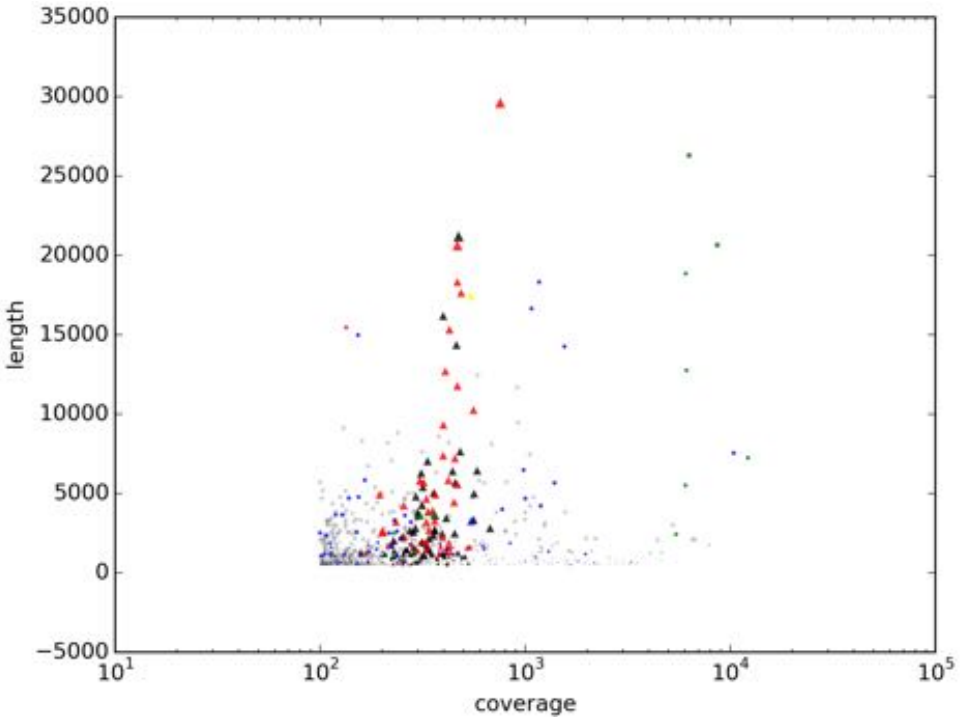

Figure C.26: *Gnetum gnemon: Length vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

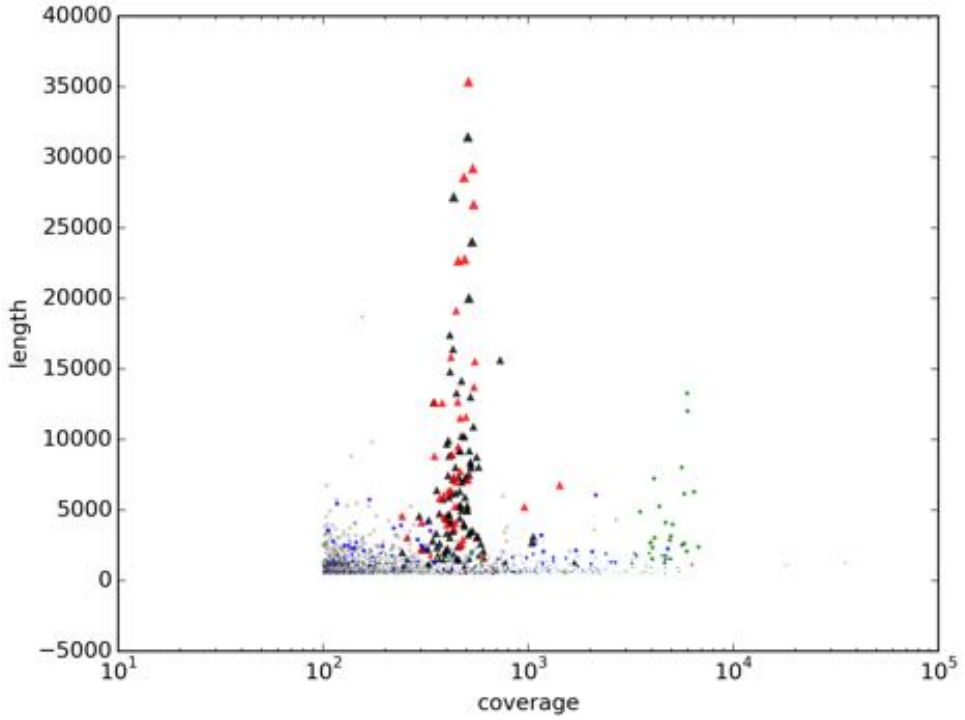

Figure C.27: *Juniperus communis: Length vs coverage. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

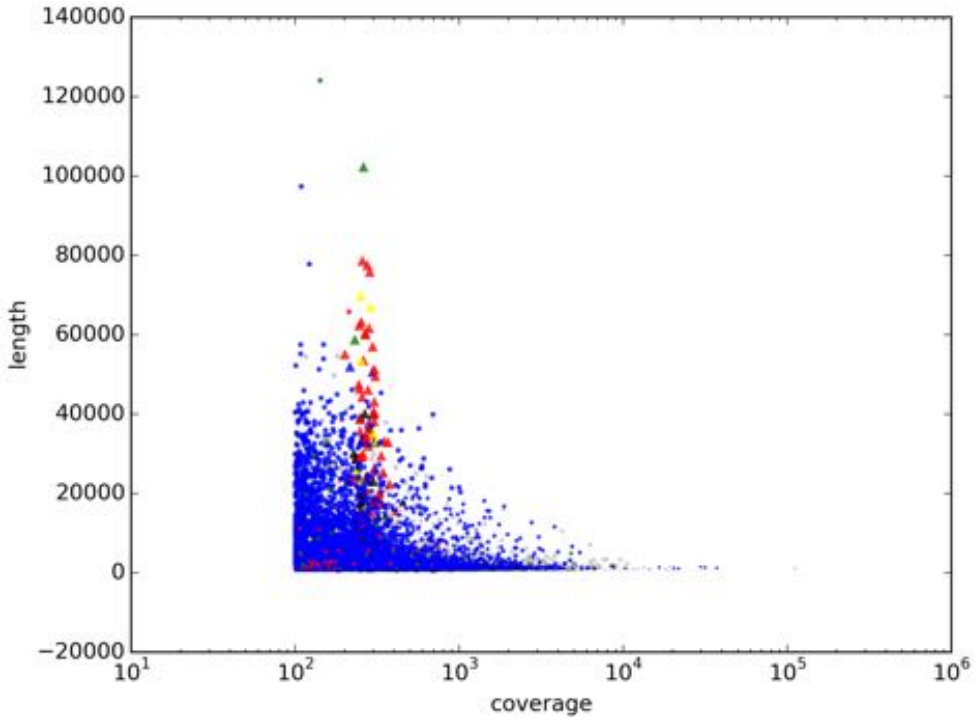

Figure C.28: *Picea abies: Length vs coverage. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

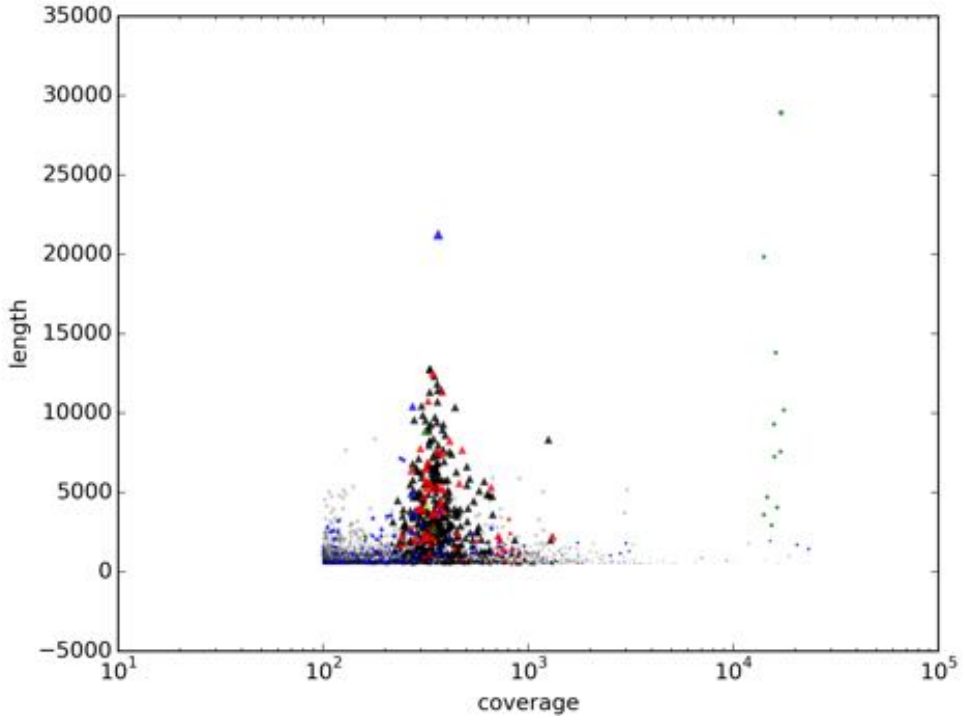

Figure C.29: *Pinus sylvestris: Length vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

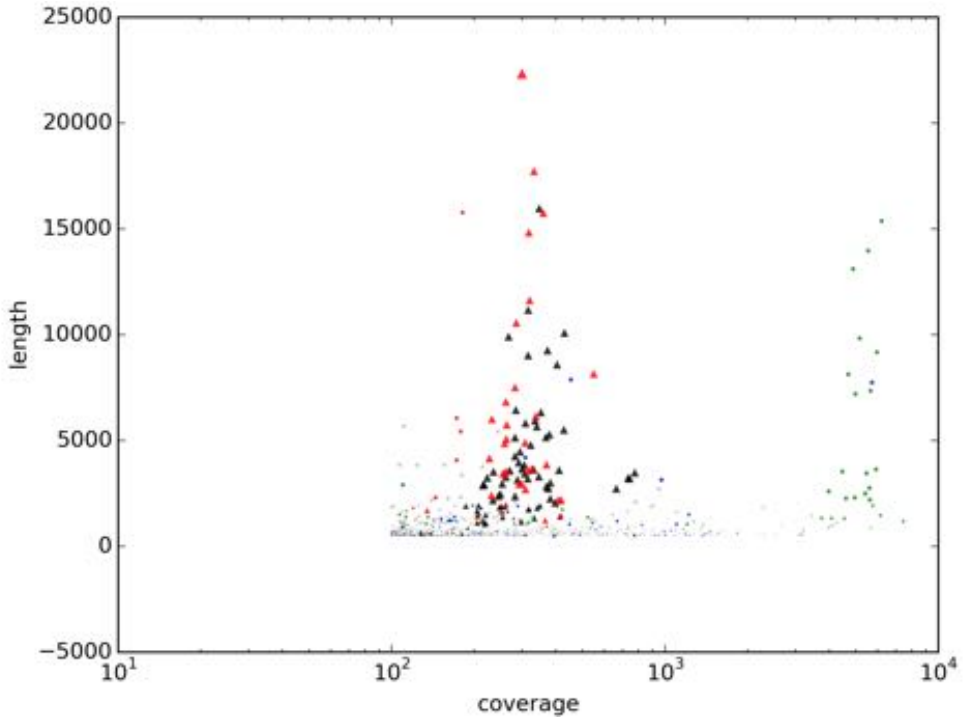

Figure C.30: *Taxus baccata: Length vs coverage.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

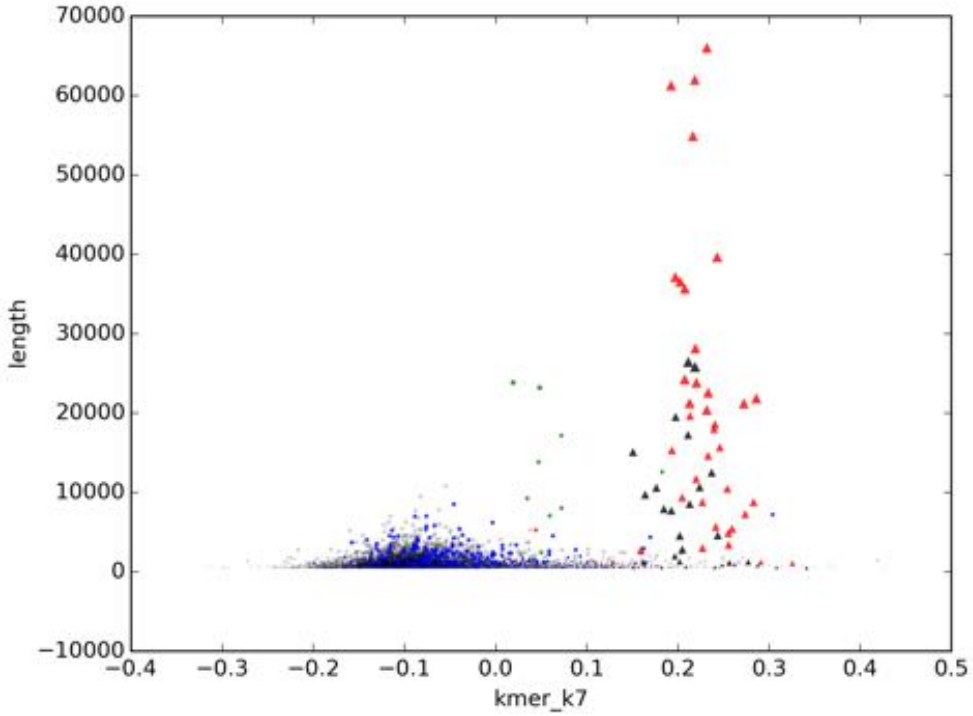

Figure C.31: *Abies sibirica: Length vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

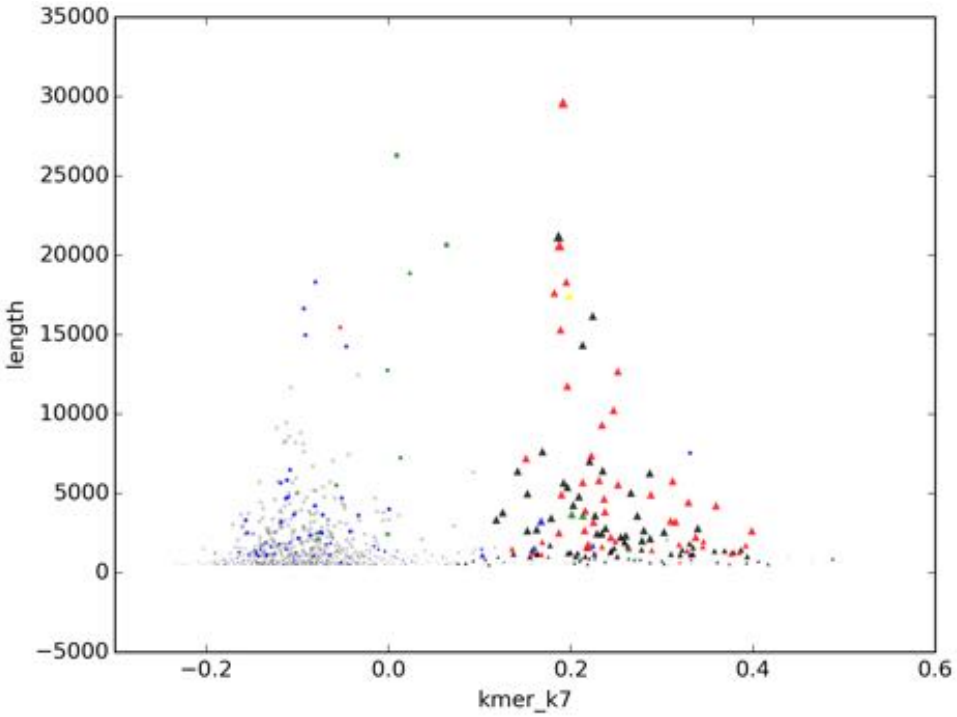

Figure C.32: *Gnetum gnemon: Length vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

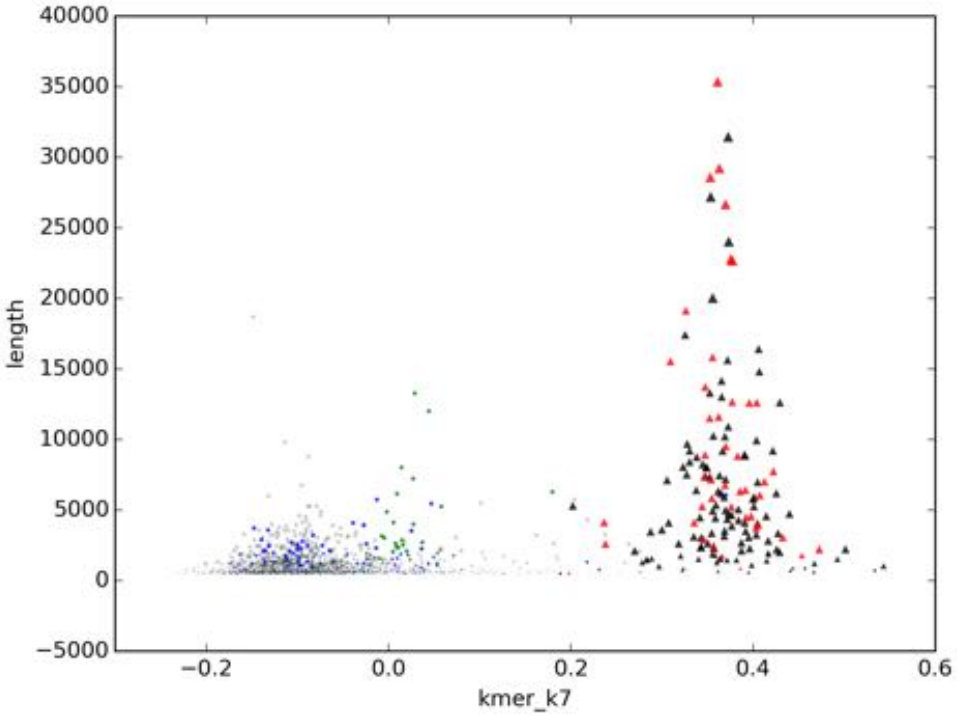

Figure C.33: *Juniperus communis: Length vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

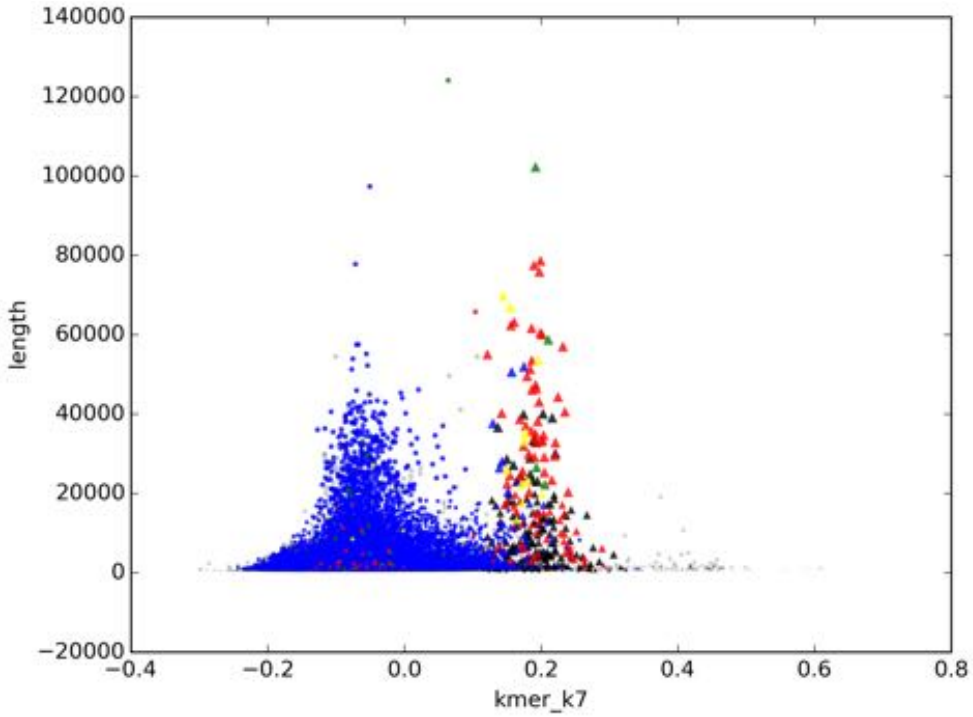

Figure C.34: *Picea abies: Length vs kmer score. Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1*

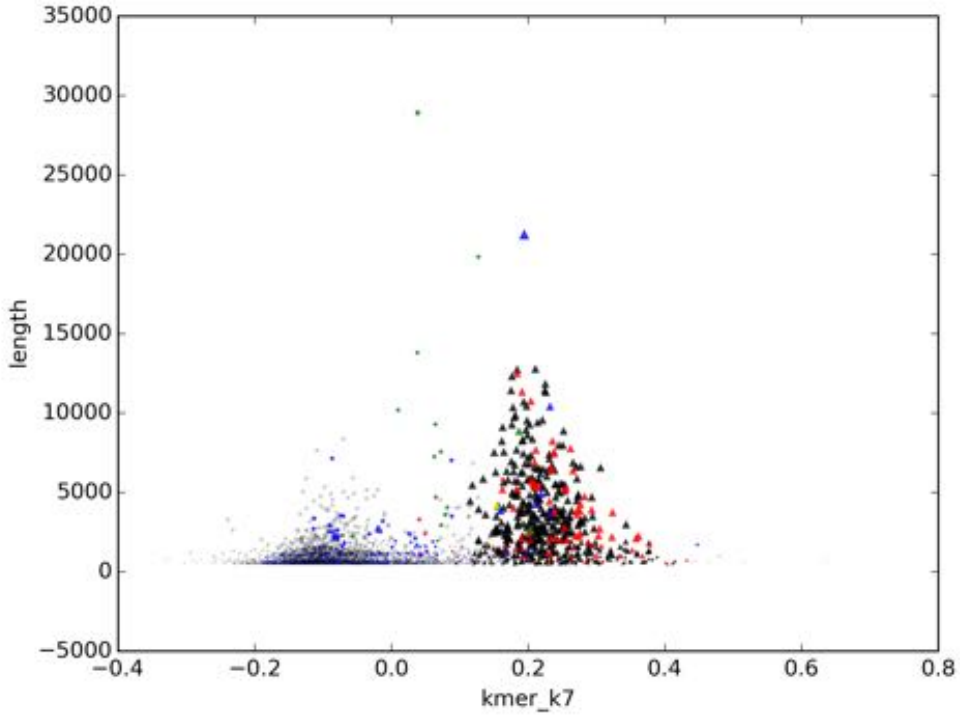

Figure C.35: *Pinus sylvestris: Length vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

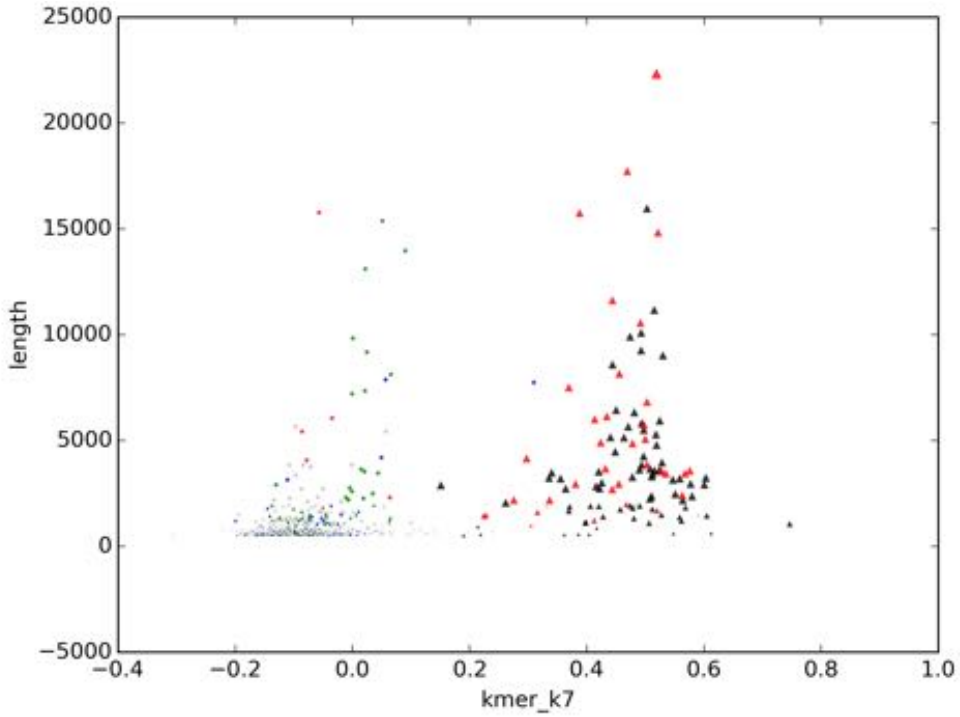

Figure C.36: *Taxus baccata: Length vs kmer score.*
Each marker is a contig. Triangles: mitochondrial (from classification pipeline). Red: mitochondrial (from BLAST alignments). Blue: nuclear. Green: chloroplast. For a more detailed explaination, see figure 6.1

Appendix D

Reference contig plots for all species

This appendix contains a reference of "contig plots" (contigs plus BLAST alignments to reference species) for all species. For an explaination of how to read these plots, see figure 6.2 and section 6.2.

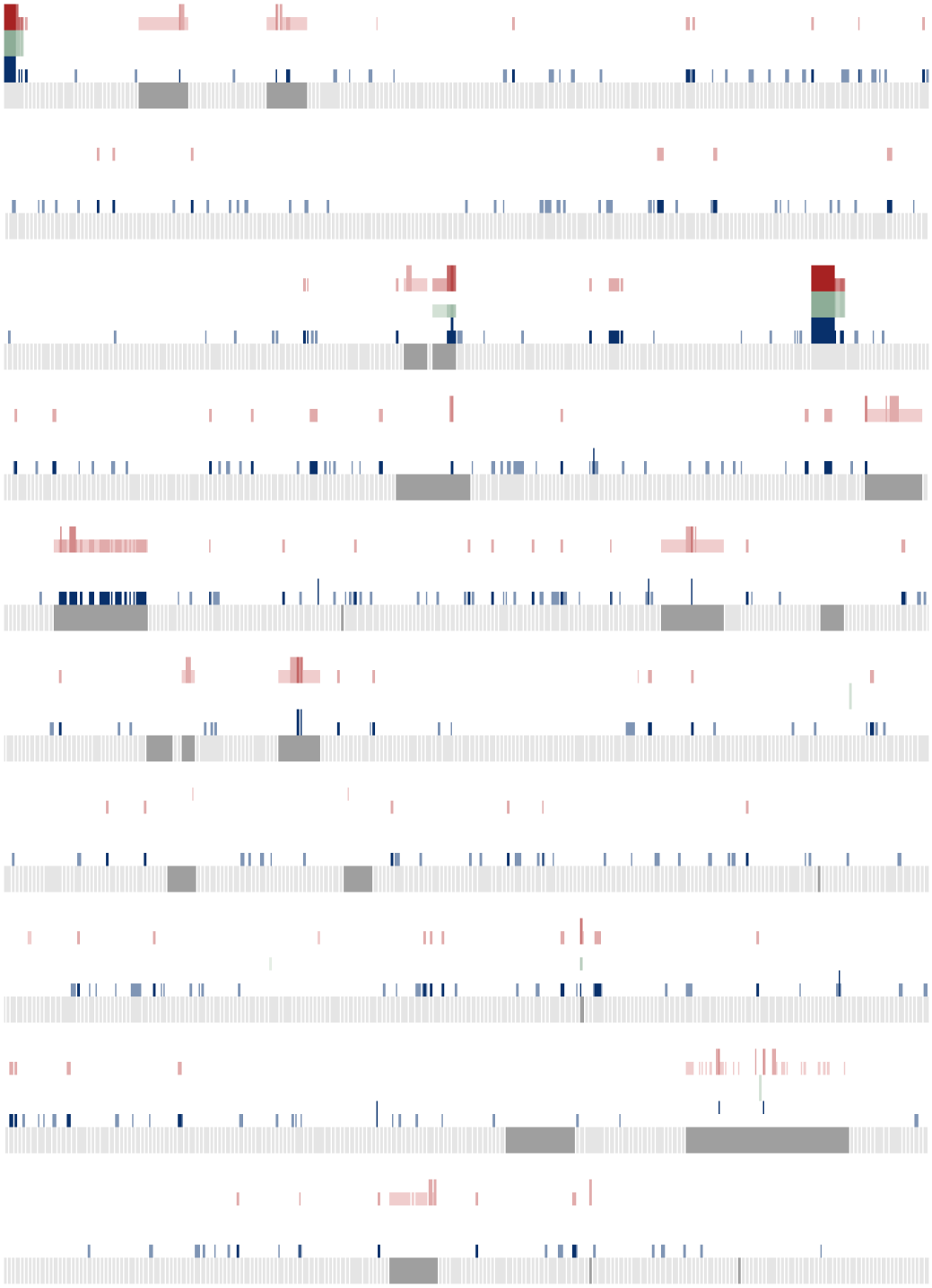

Figure D.1: *Abies sibirica contig plot #1.*

Figure D.2: *Abies sibirica contig plot #2.*

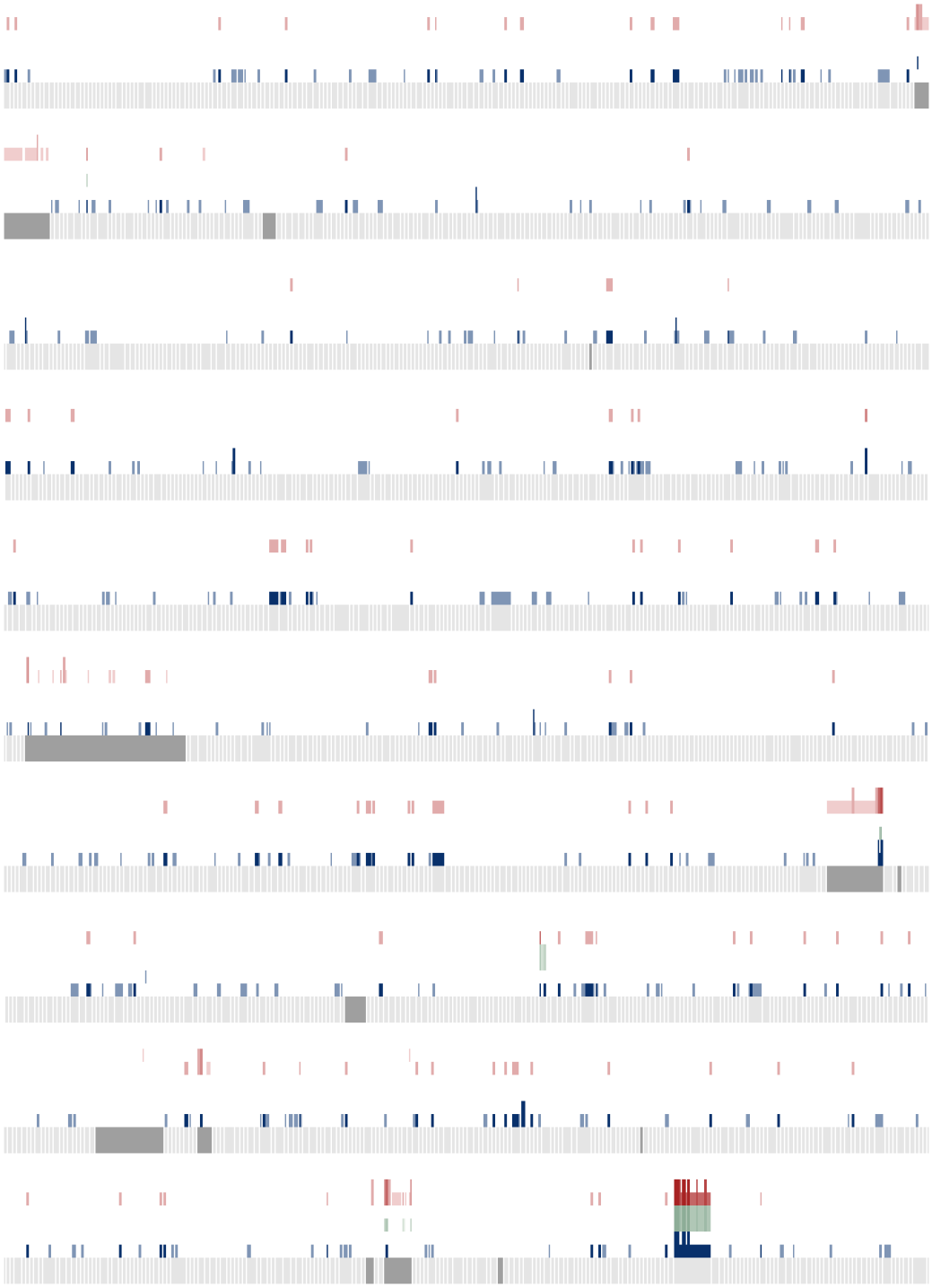

Figure D.3: *Abies sibirica contig plot #3.*

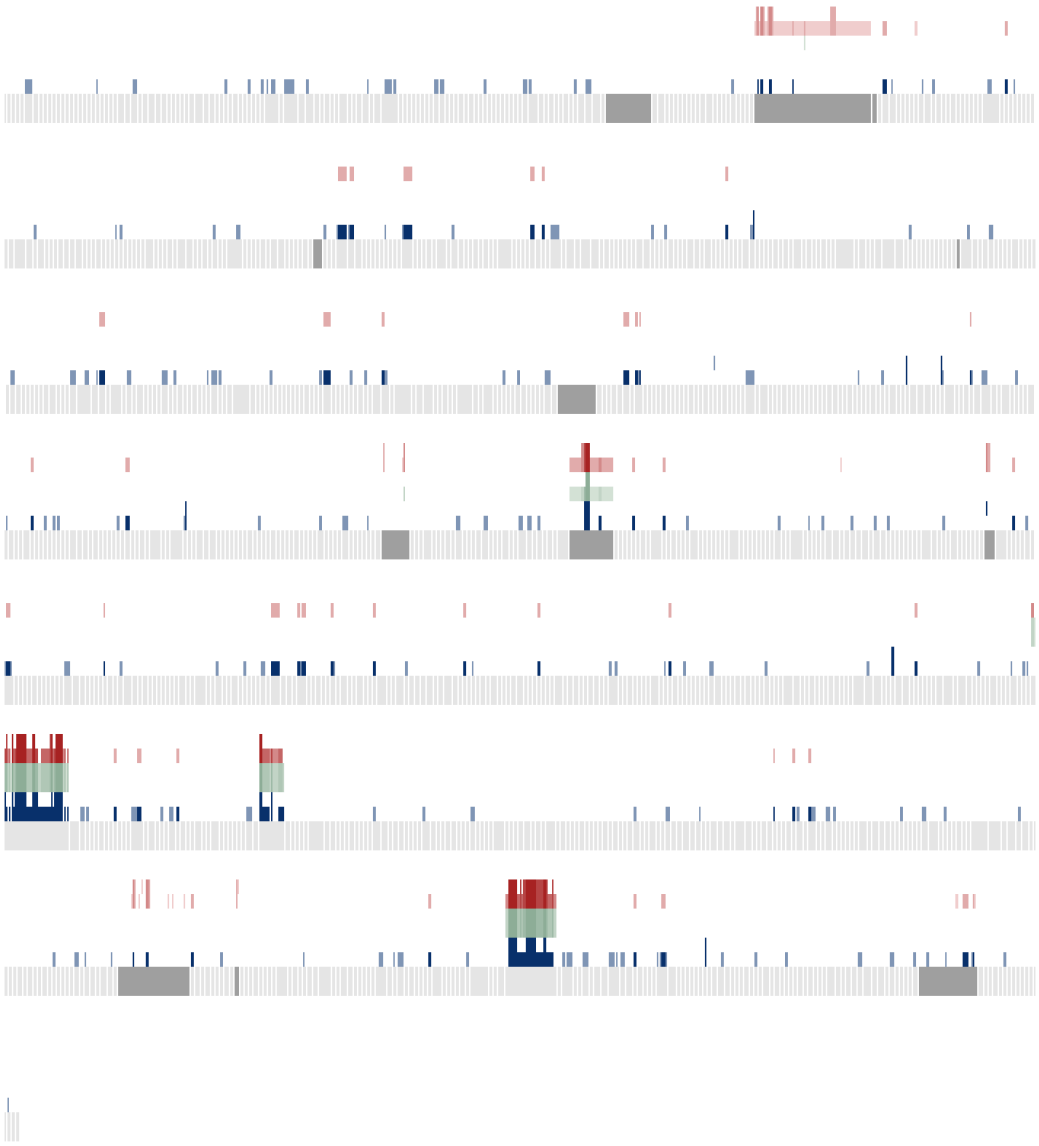

Figure D.4: *Abies sibirica contig plot #4.*

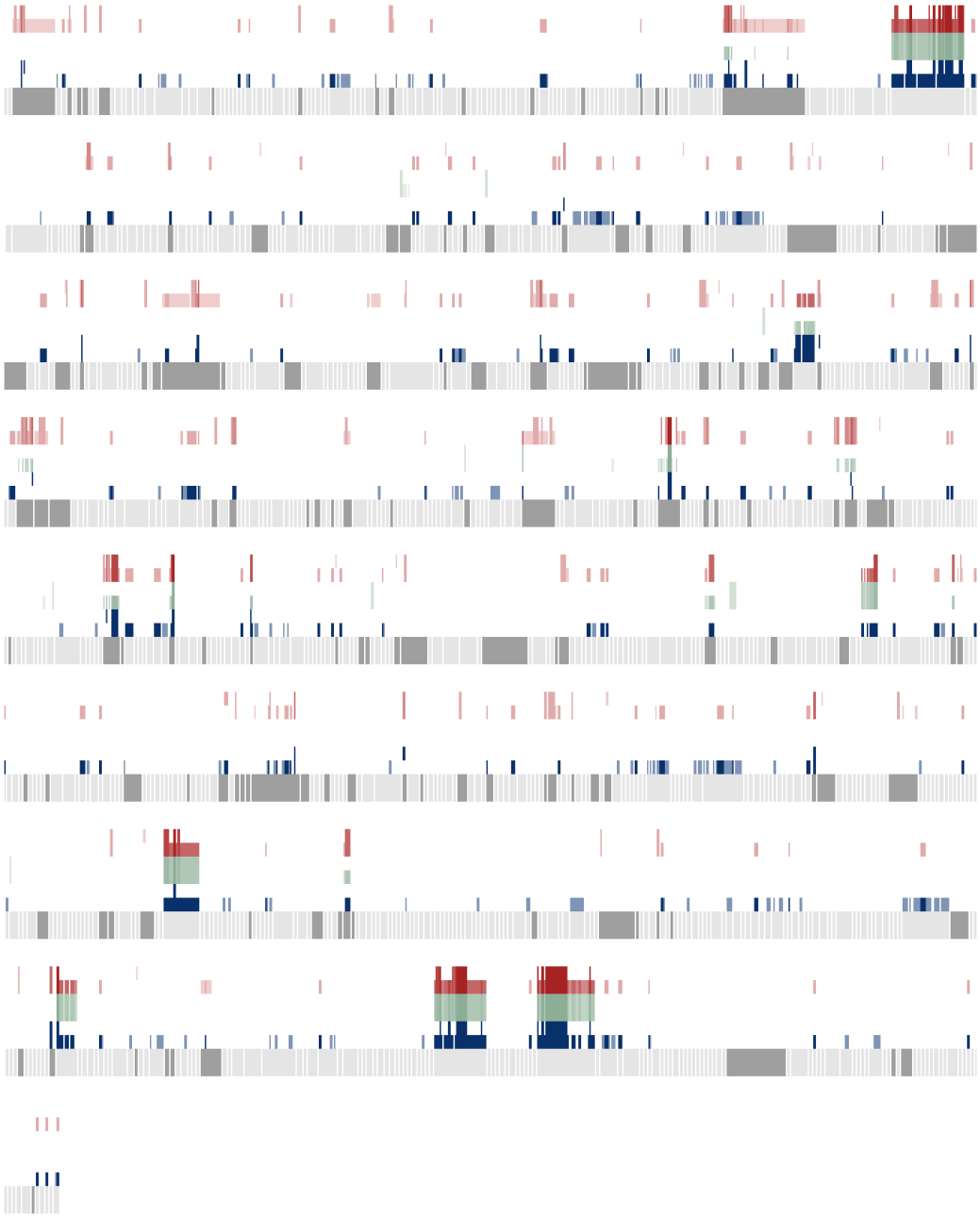

Figure D.5: *Gnetum gnemon contig plot #1.*

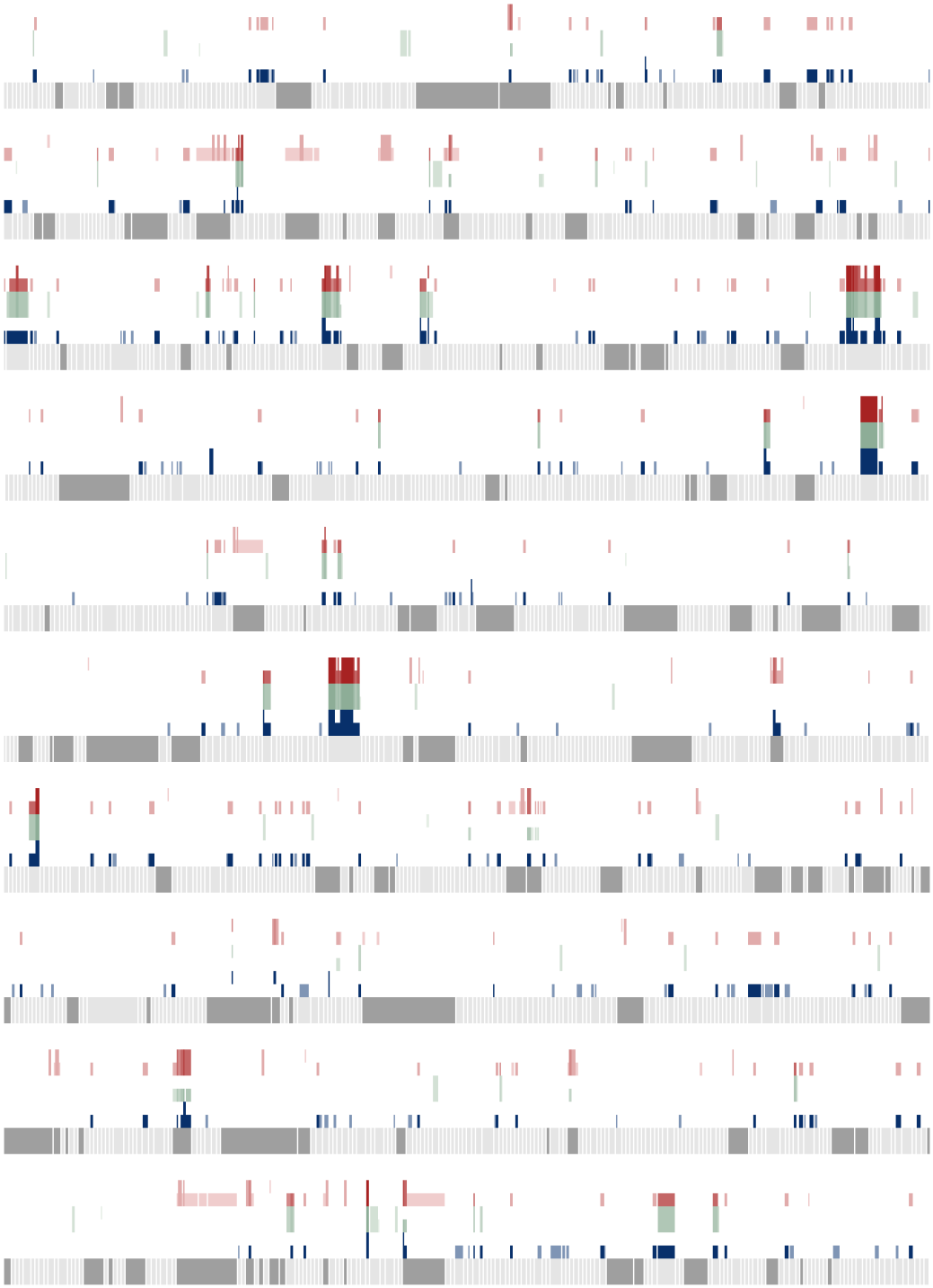

Figure D.6: *Juniperus communis contig plot #1.*

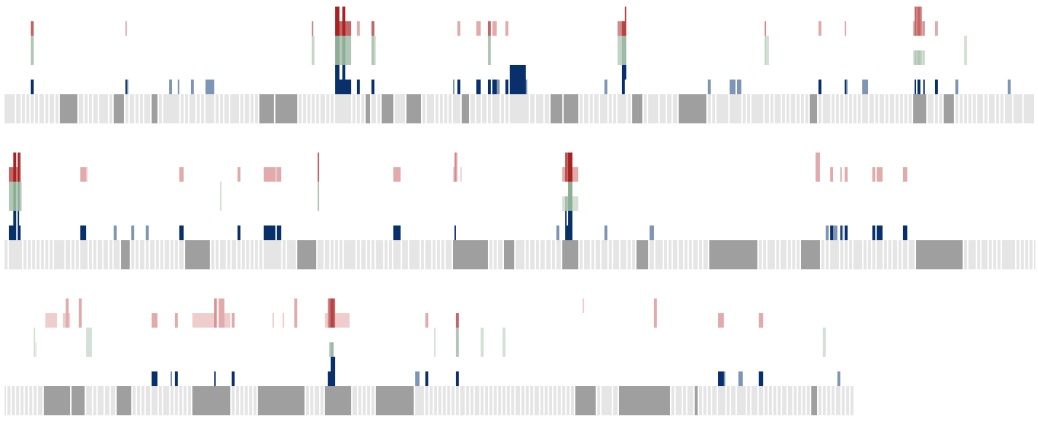

Figure D.7: *Juniperus communis contig plot #2.*

Figure D.8: *Picea abies contig plot #1.*

Figure D.9: *Picea abies contig plot #2.*

Figure D.10: *Picea abies contig plot #3.*

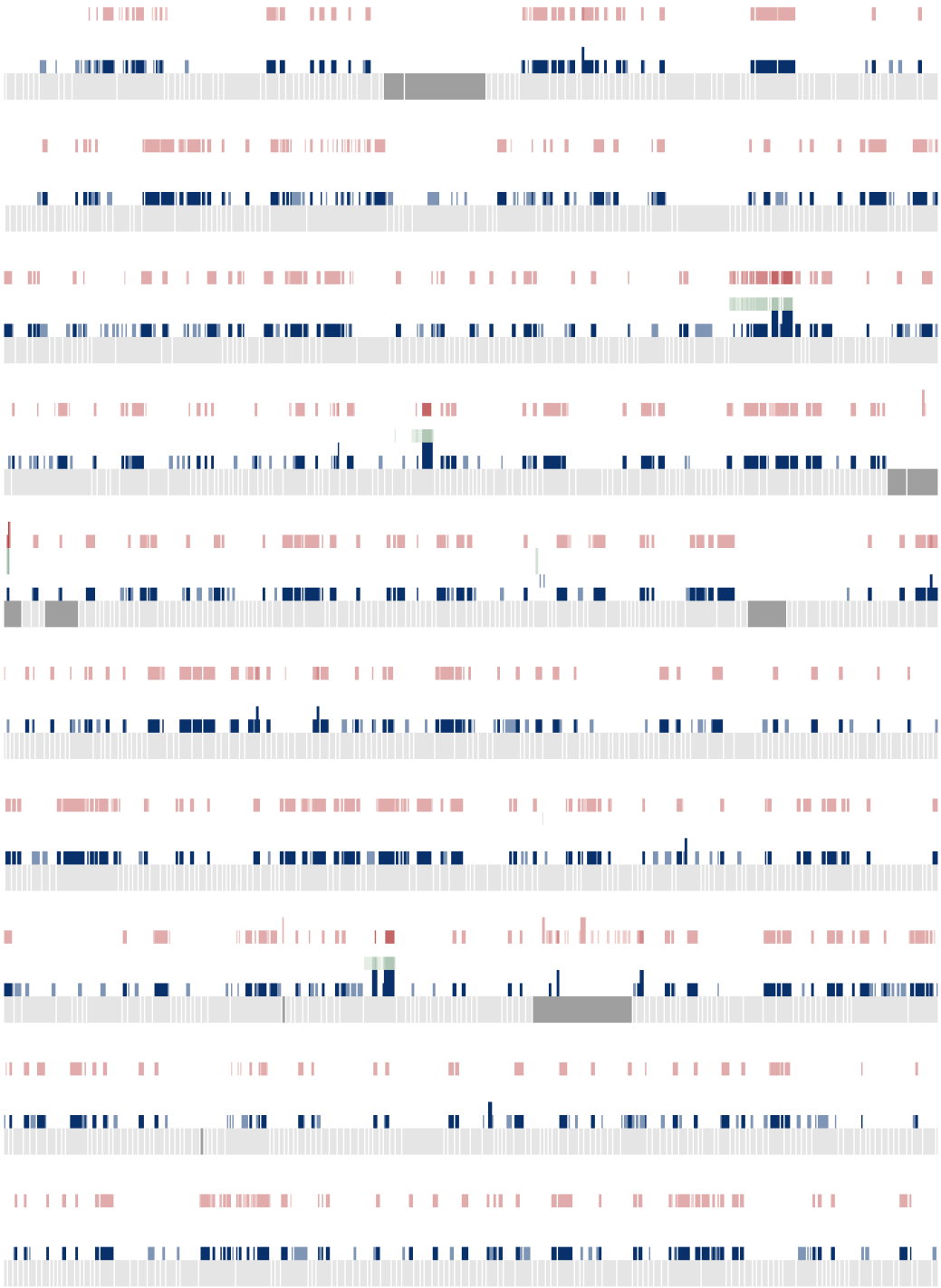

Figure D.11: *Picea abies contig plot #4.*

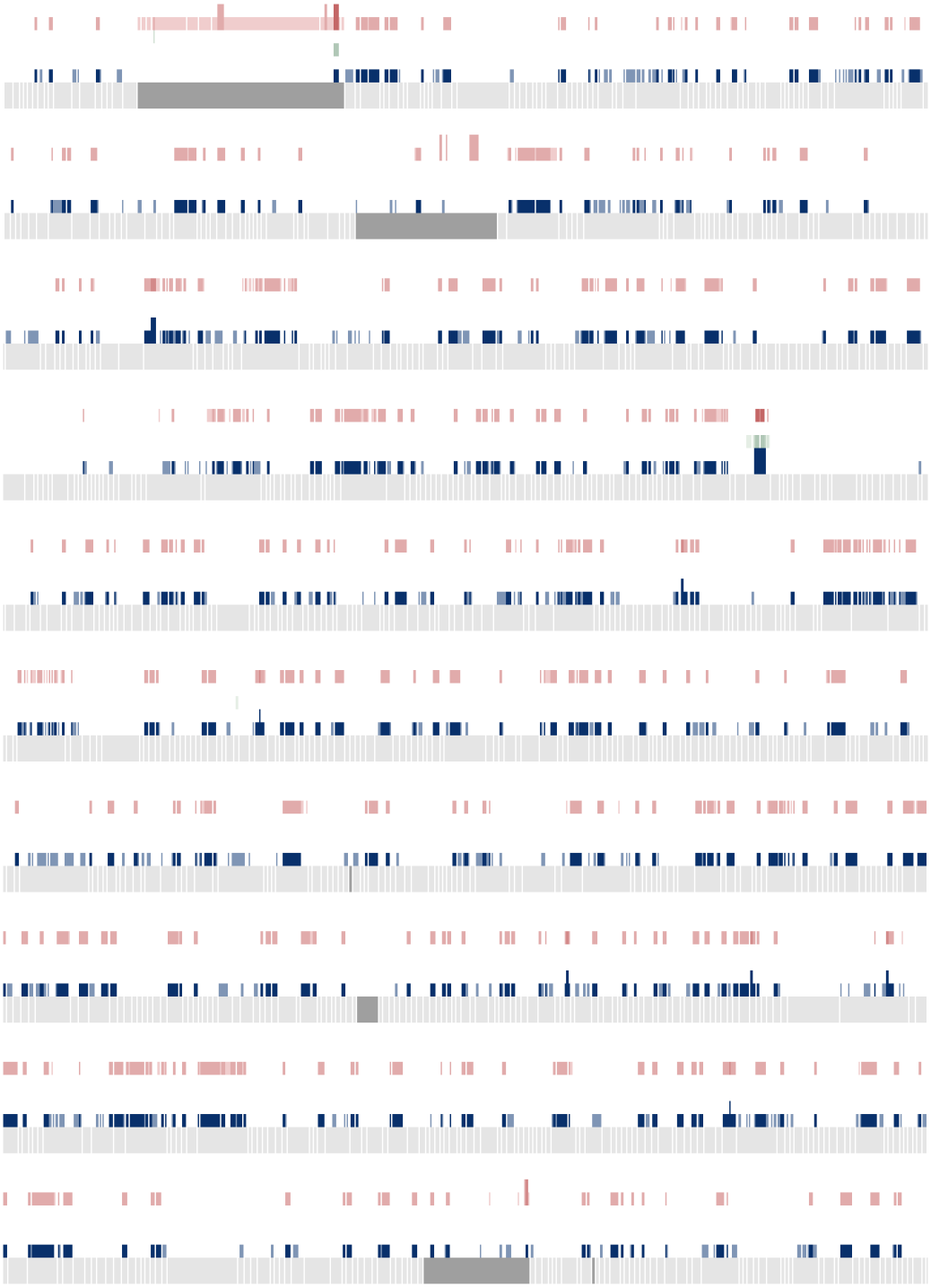

Figure D.12: *Picea abies contig plot #5.*

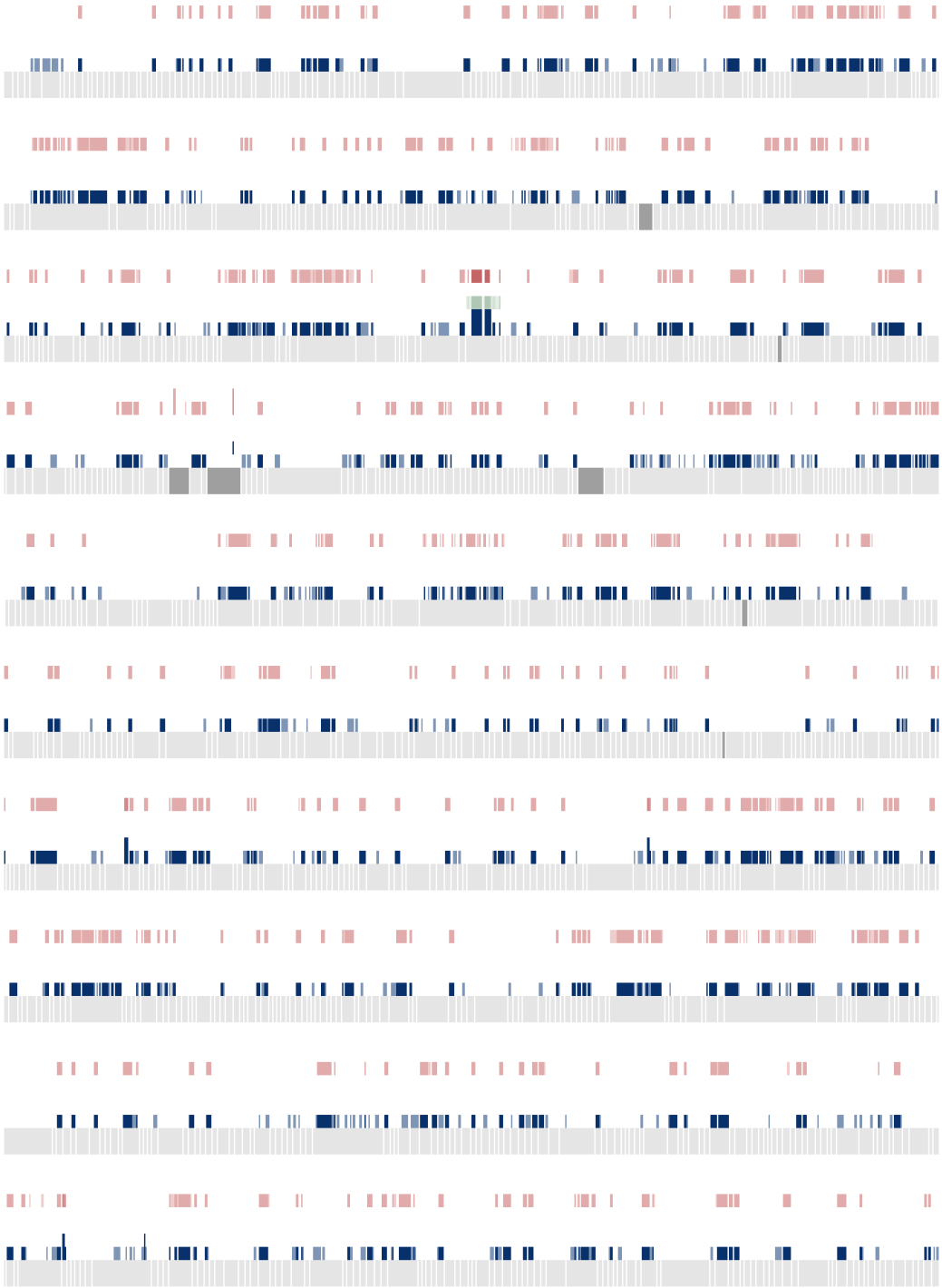

Figure D.13: *Picea abies contig plot #6.*

Figure D.14: *Picea abies contig plot #7.*

Figure D.15: *Picea abies contig plot #8.*

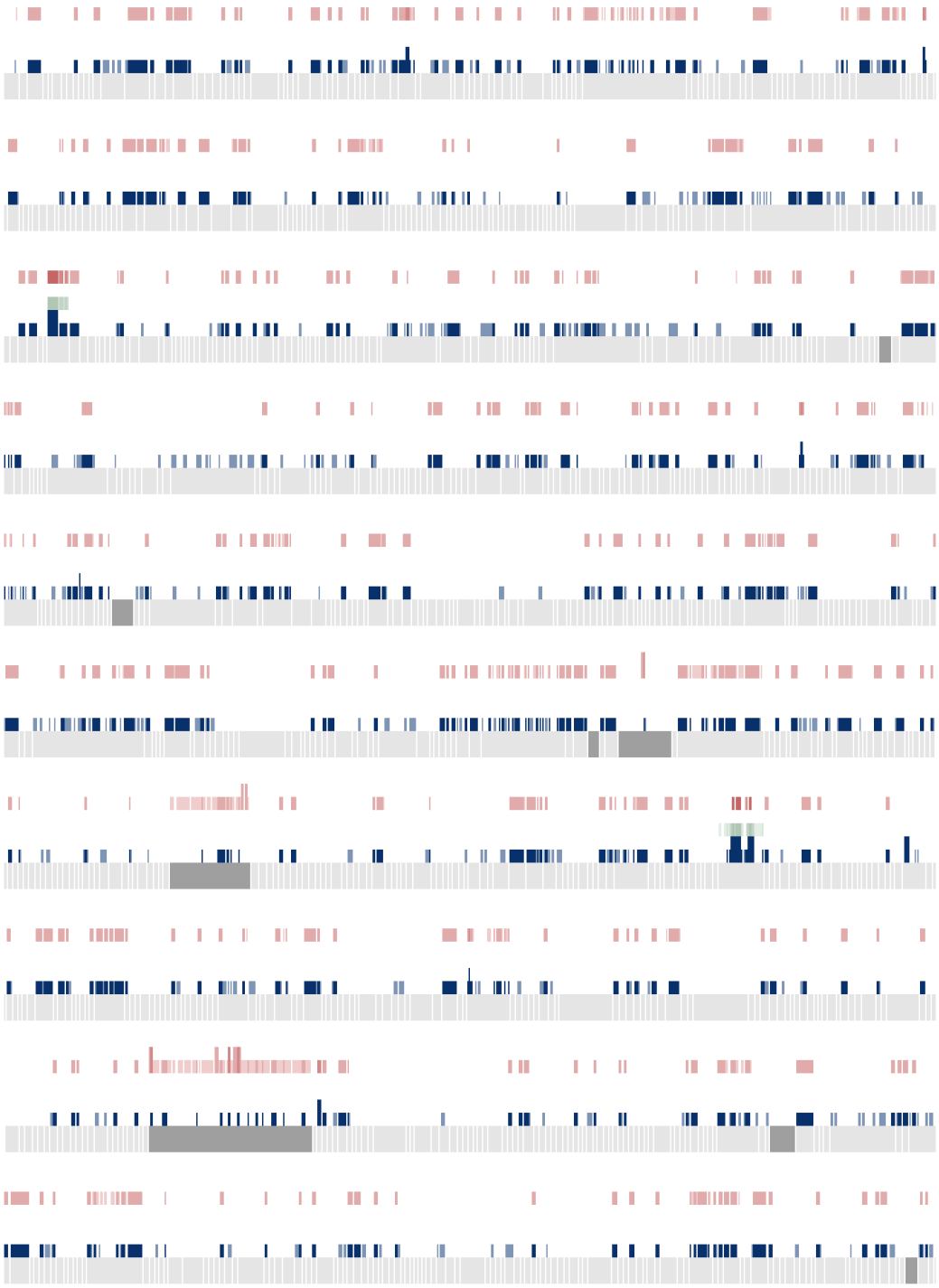

Figure D.16: *Picea abies contig plot #9.*

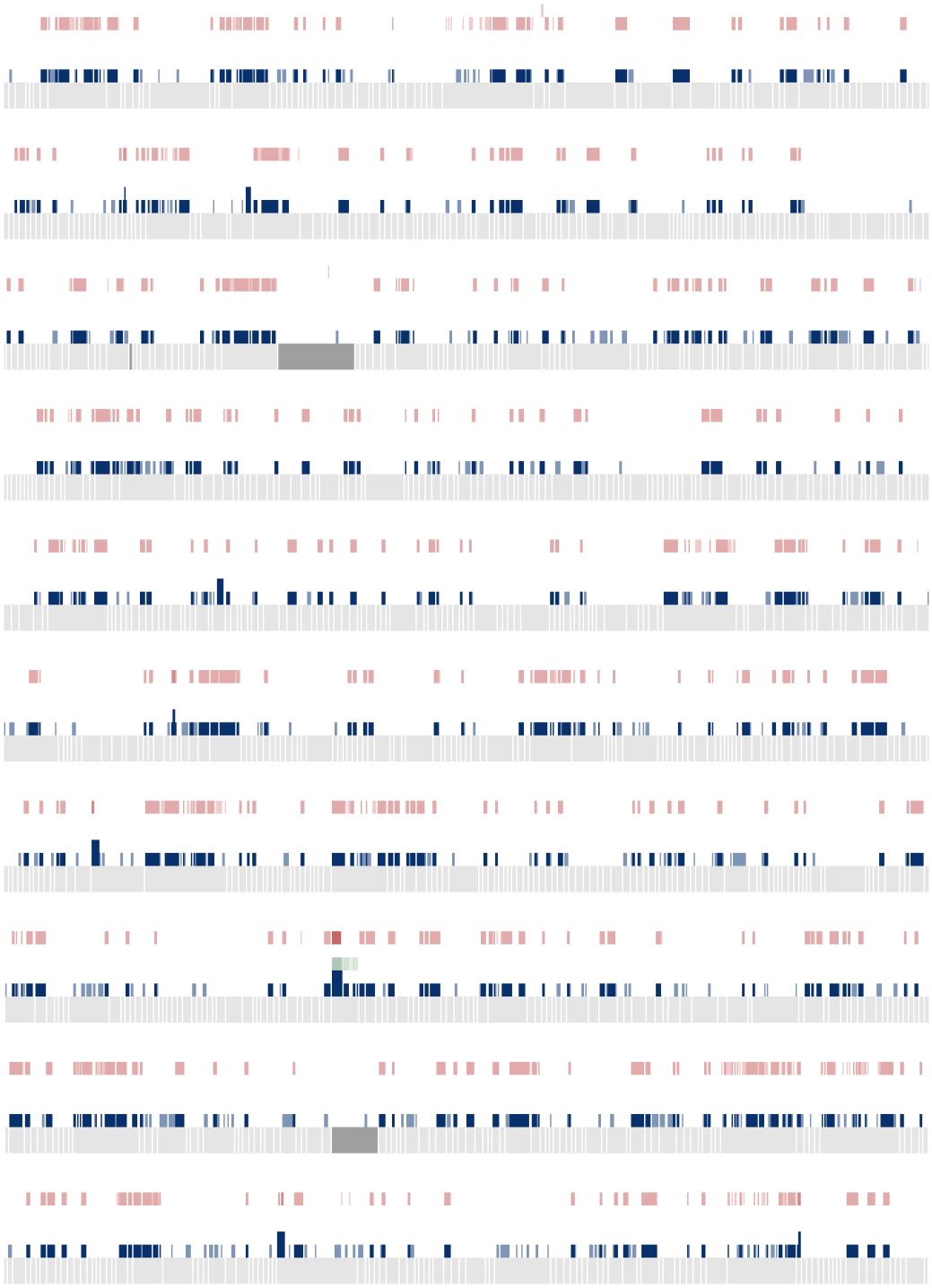

Figure D.17: *Picea abies contig plot #10.*

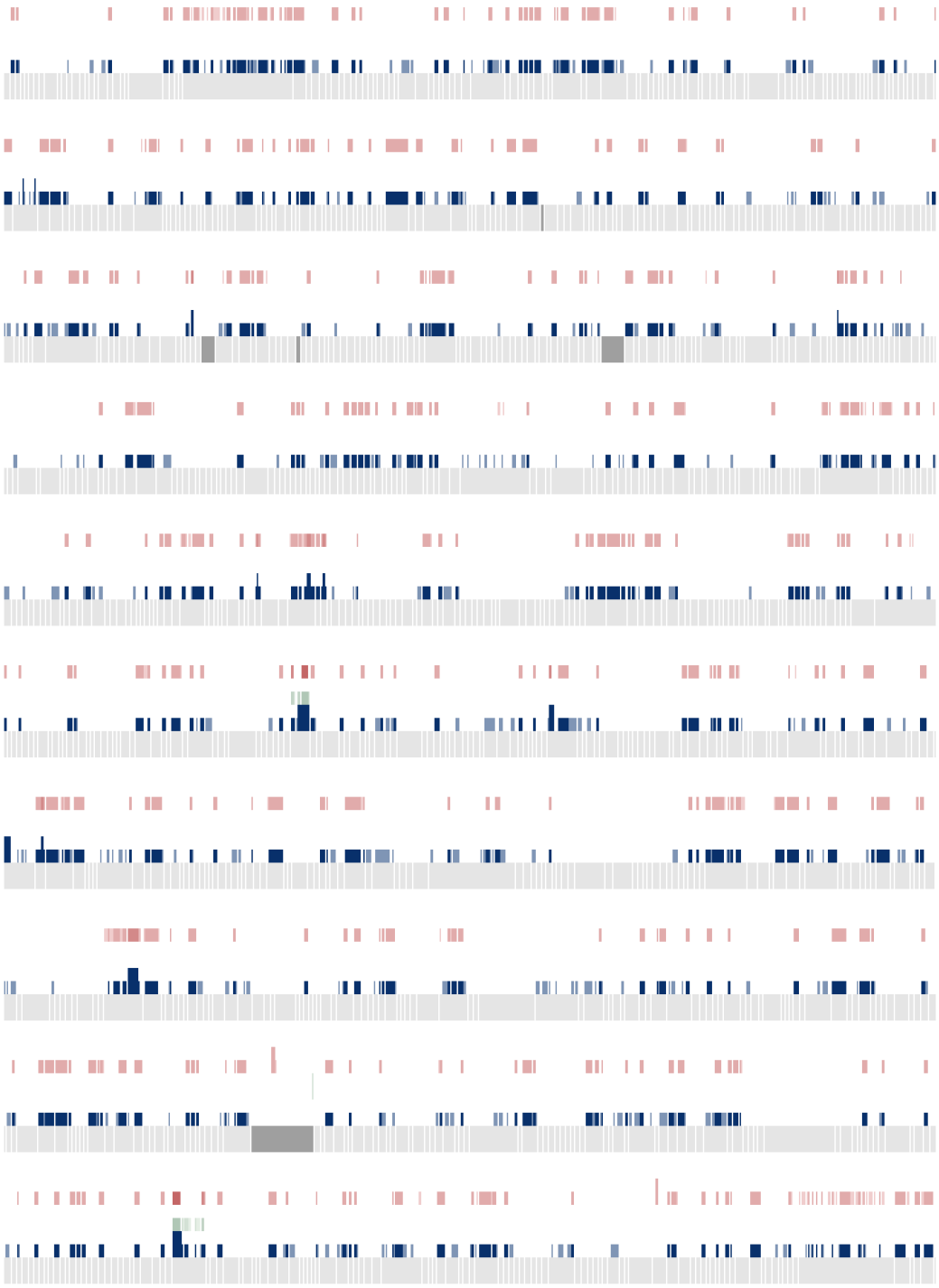

Figure D.18: *Picea abies contig plot #11.*

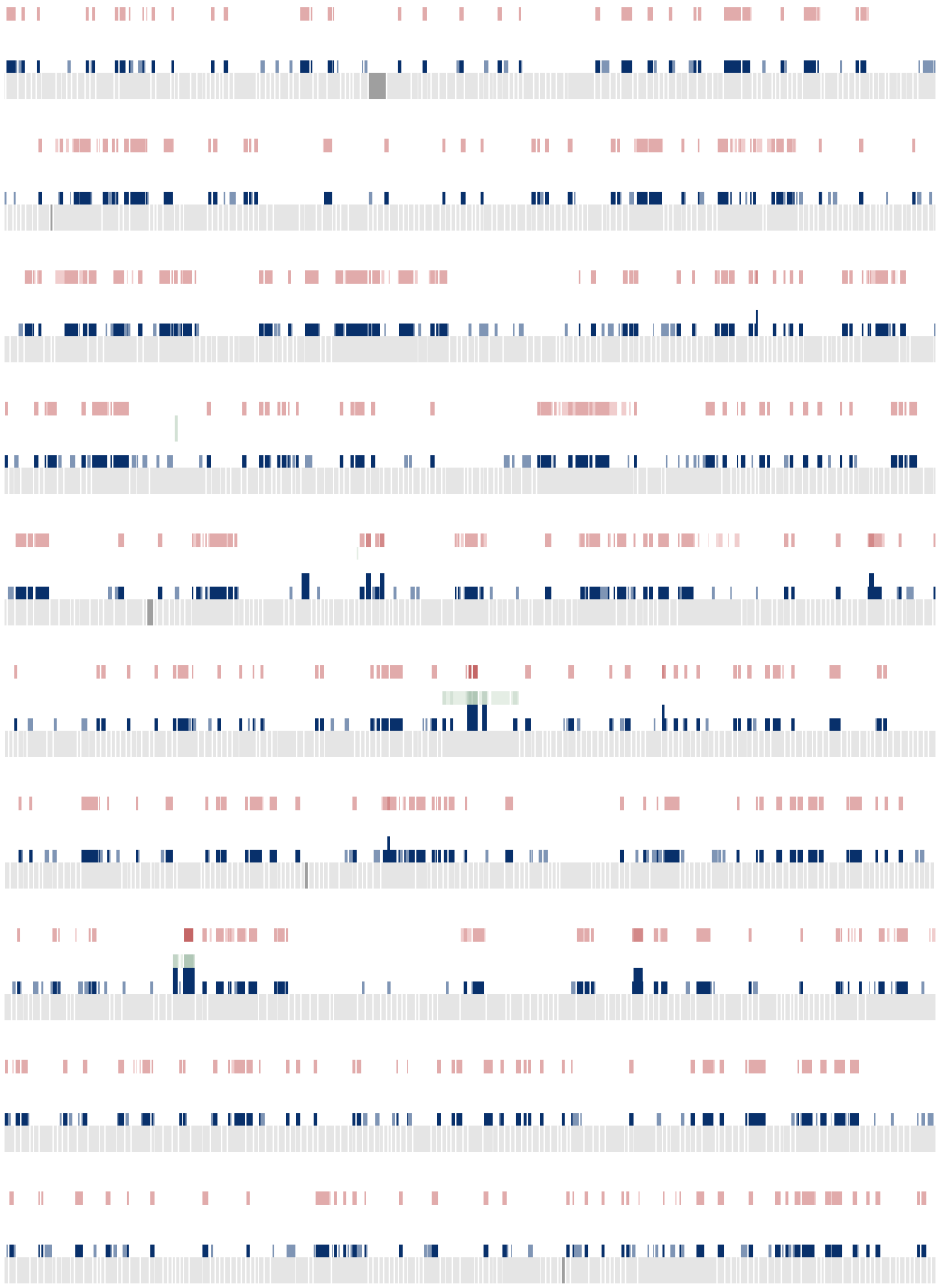

Figure D.19: *Picea abies contig plot #12.*

Figure D.20: *Picea abies contig plot #13.*

Figure D.21: *Picea abies contig plot #14.*

Figure D.22: *Picea abies contig plot #15.*

Figure D.23: *Picea abies contig plot #16.*

Figure D.24: *Picea abies contig plot #17.*

Figure D.25: *Picea abies contig plot #18.*

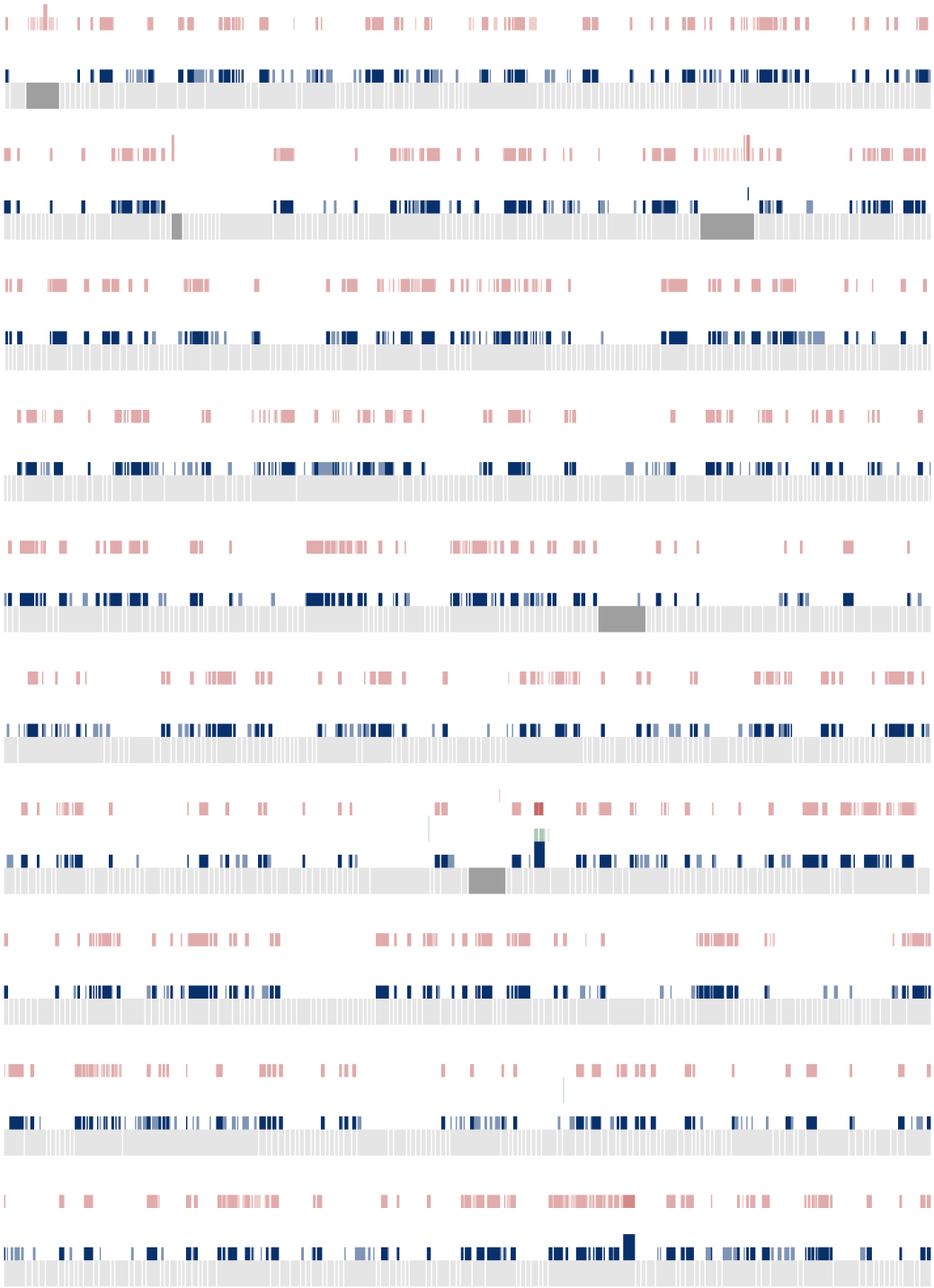

Figure D.26: *Picea abies contig plot #19.*

Figure D.27: *Picea abies contig plot #20.*

Figure D.28: *Picea abies contig plot #21.*

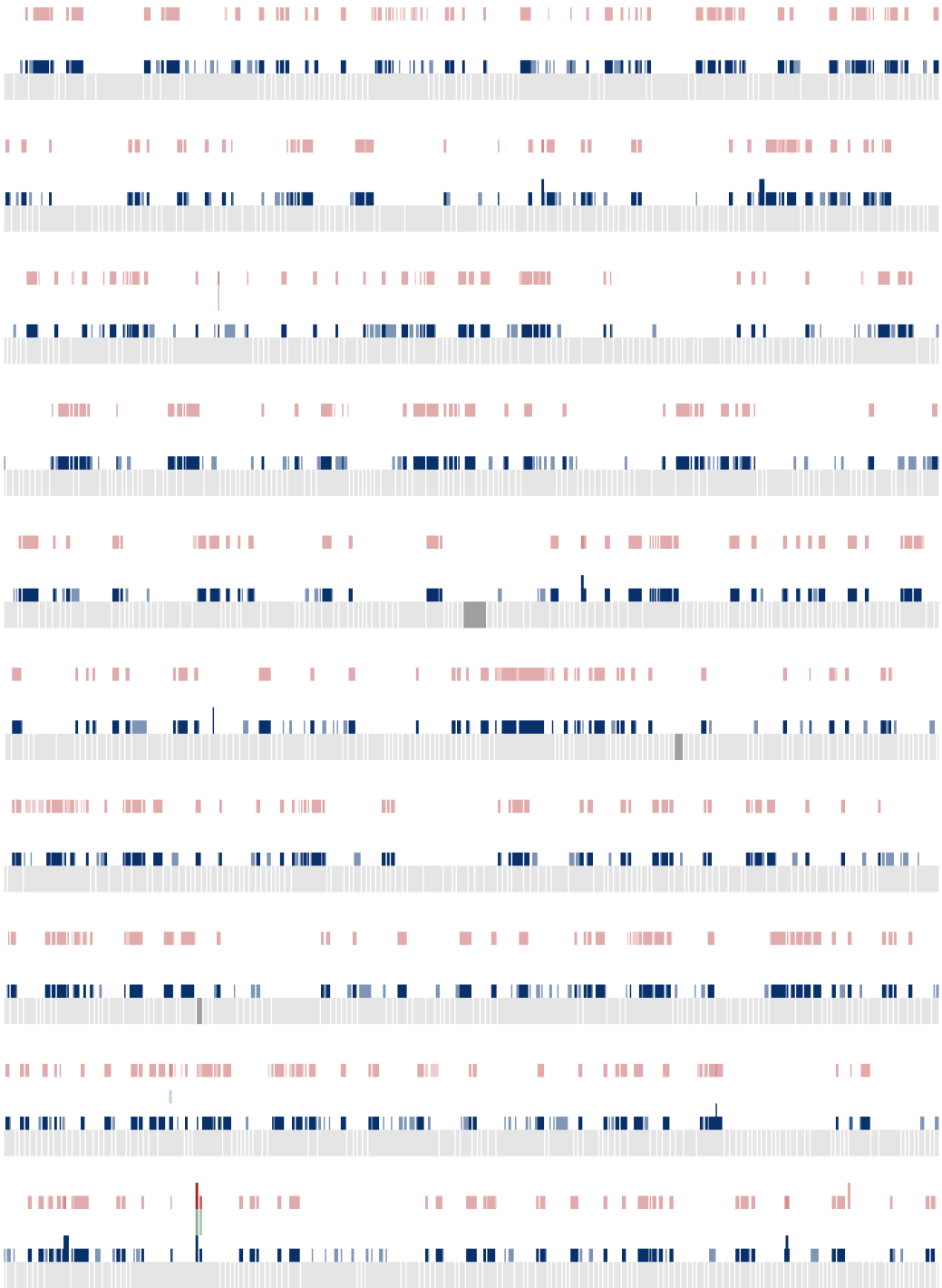

Figure D.29: *Picea abies contig plot #22.*

Figure D.30: *Picea abies contig plot #23.*

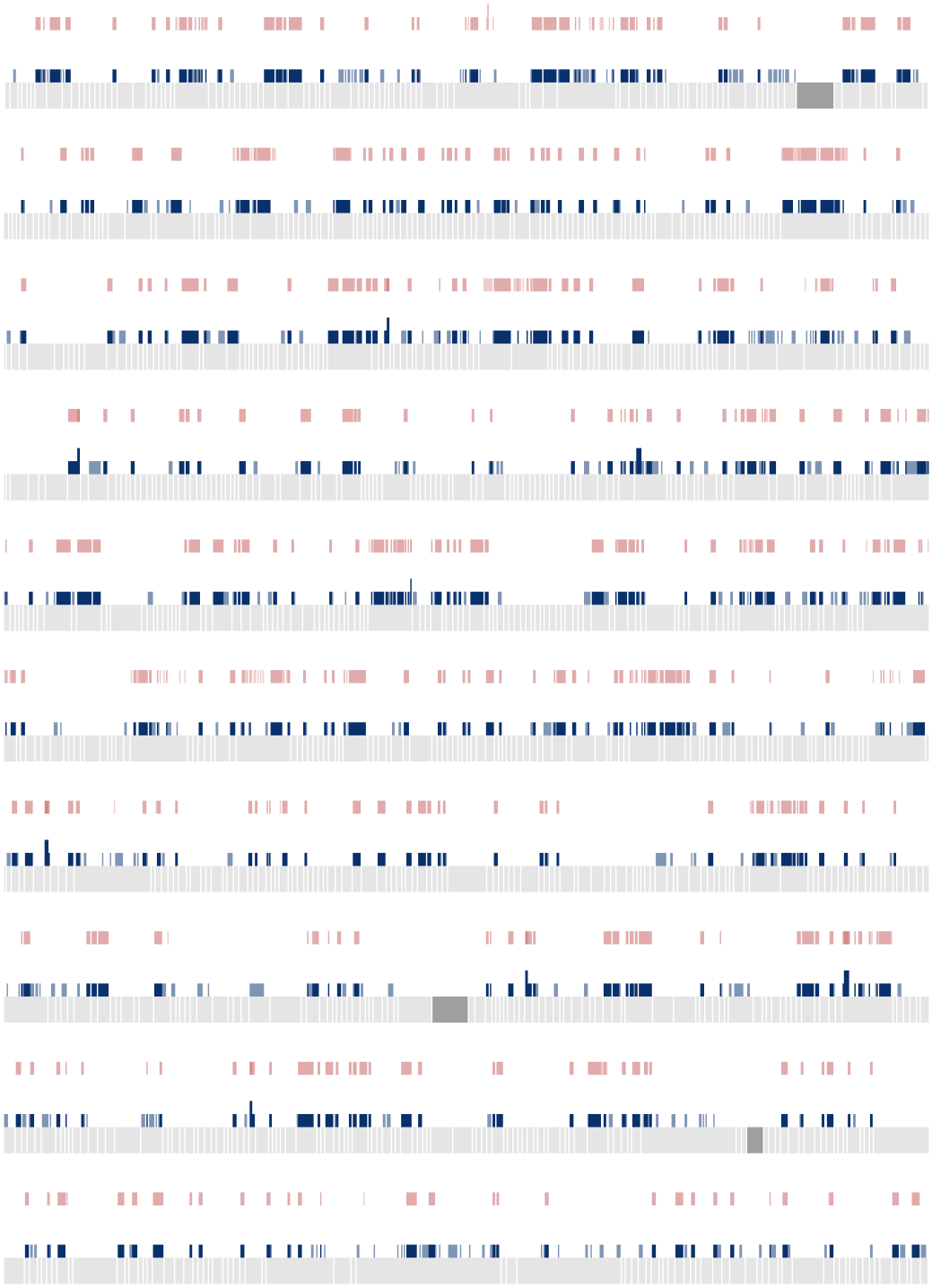

Figure D.31: *Picea abies contig plot #24.*

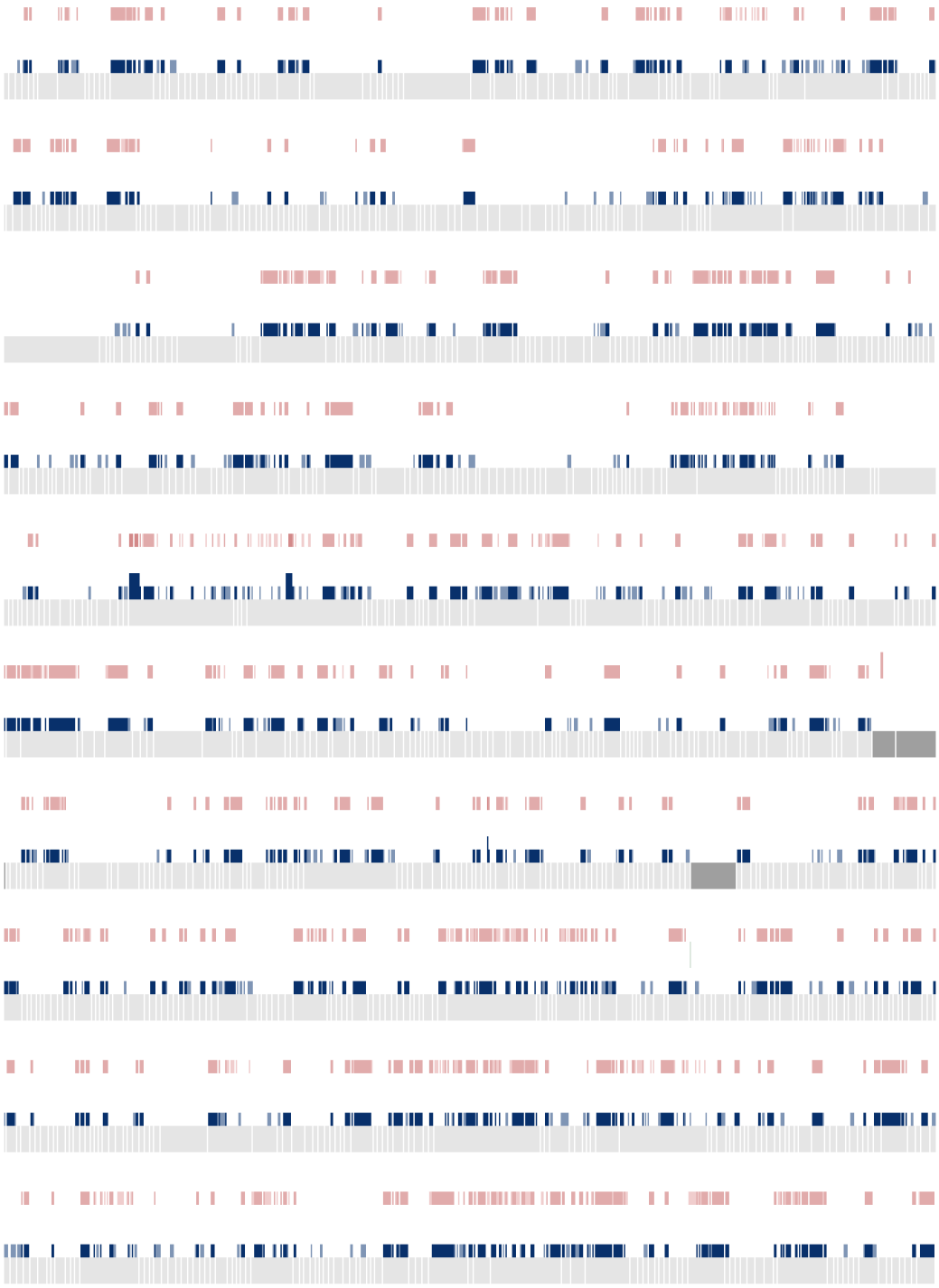

Figure D.32: *Picea abies contig plot #25.*

Figure D.33: *Picea abies contig plot #26.*

Figure D.34: *Picea abies contig plot #27.*

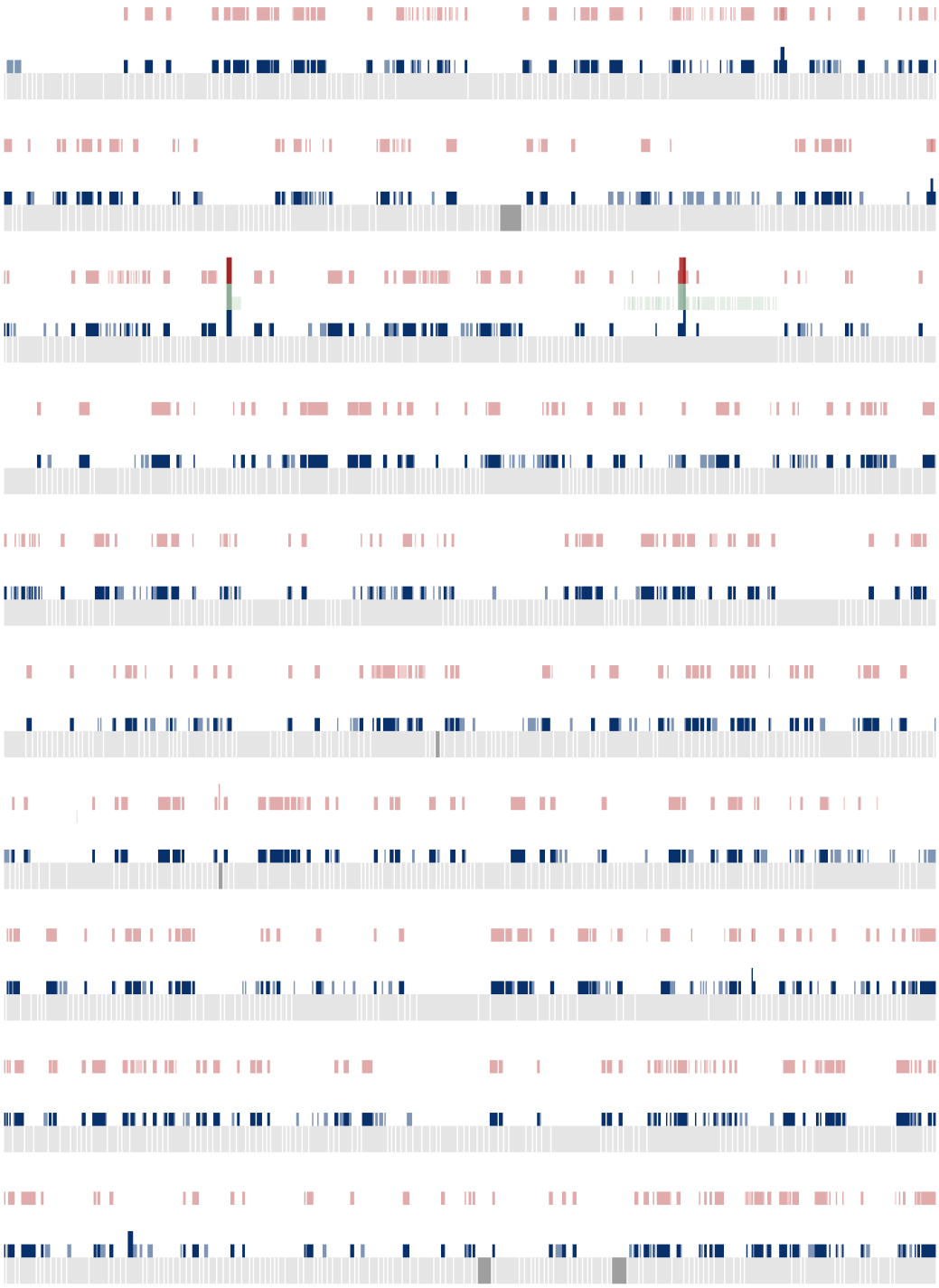

Figure D.35: *Picea abies contig plot #28.*

Figure D.36: *Picea abies contig plot #29.*

Figure D.37: *Picea abies contig plot #30.*

Figure D.38: *Picea abies contig plot #31.*

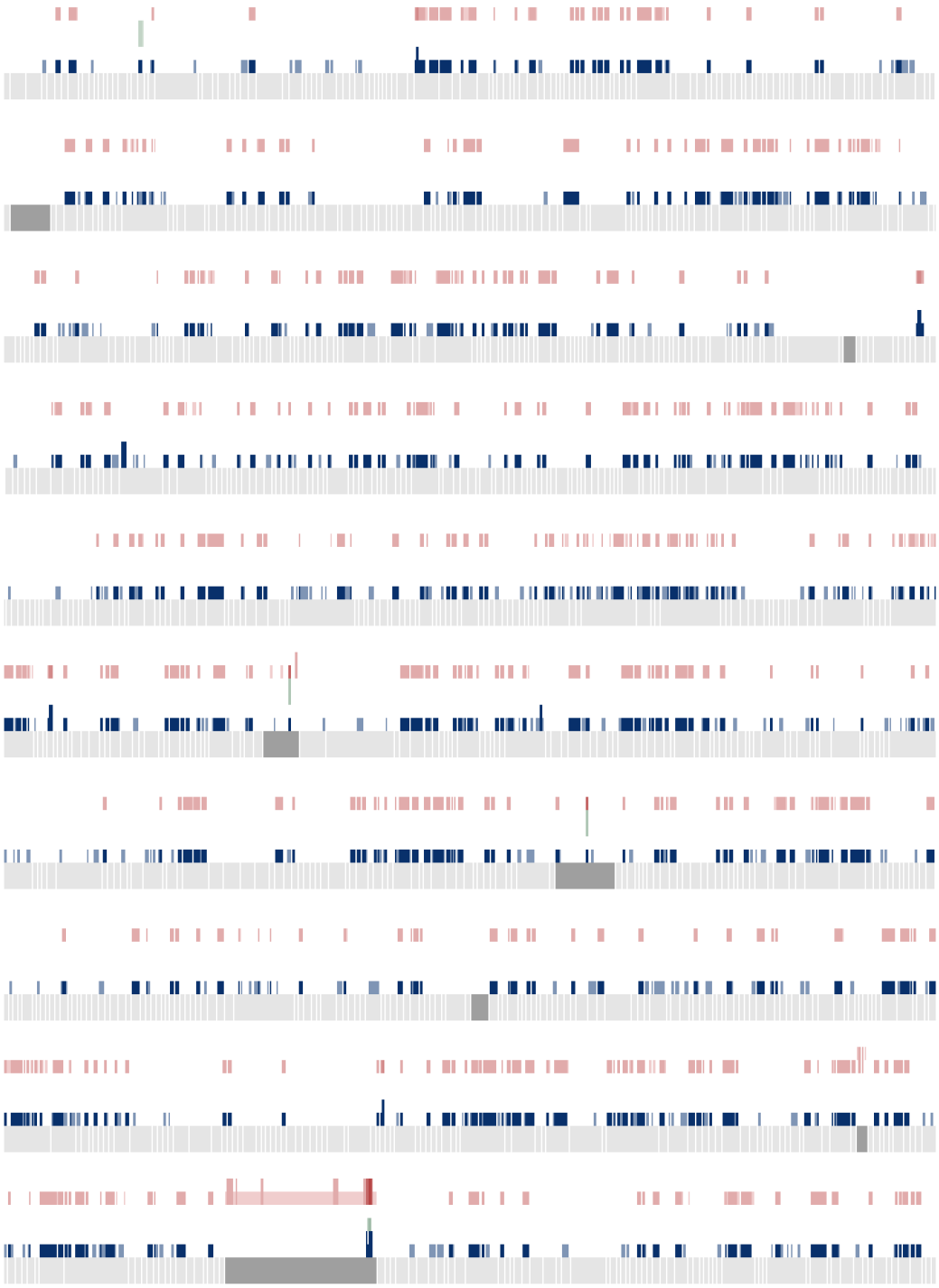

Figure D.39: *Picea abies contig plot #32.*

Figure D.40: *Picea abies contig plot #33.*

Figure D.41: *Picea abies contig plot #34.*

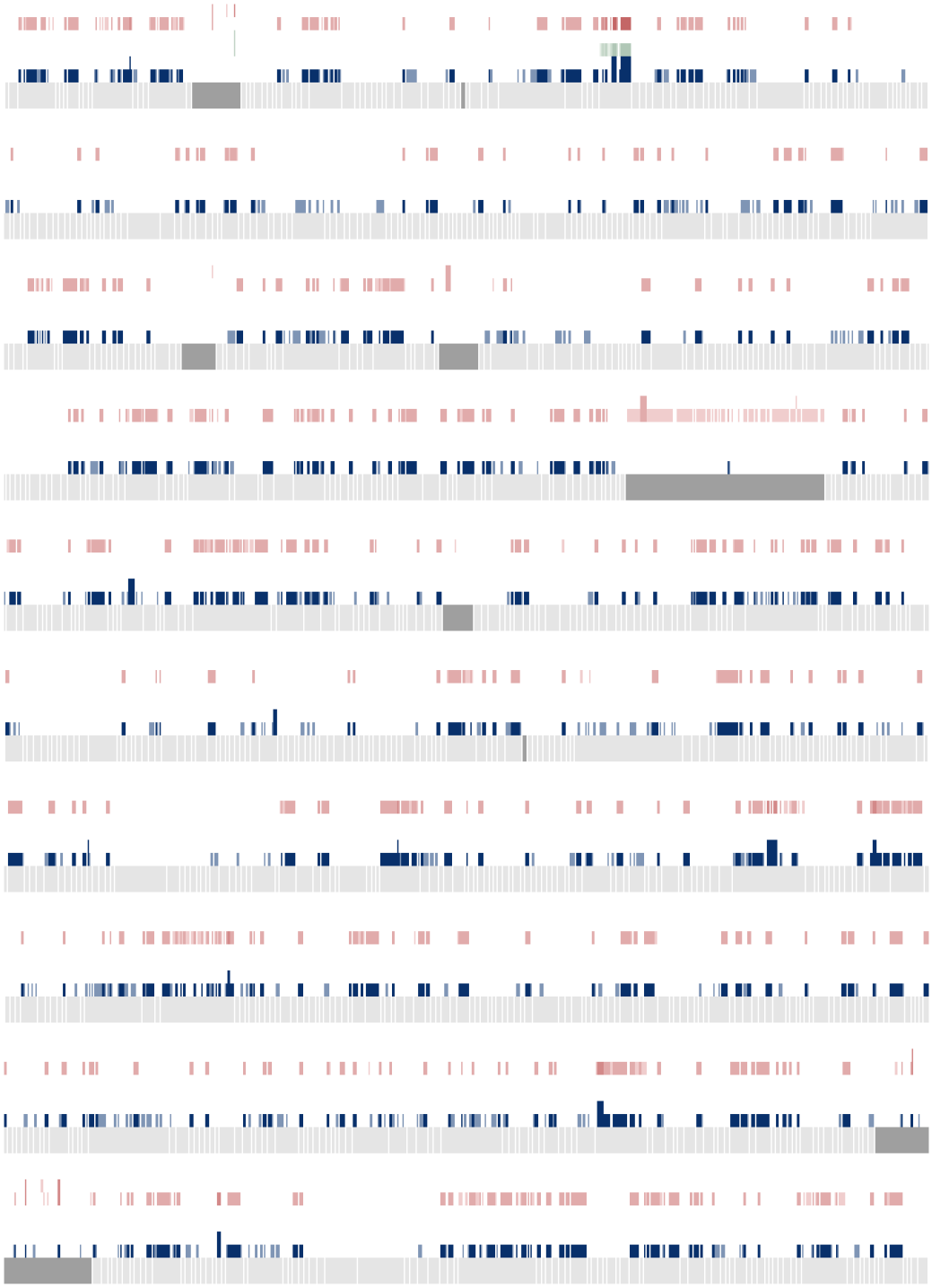

Figure D.42: *Picea abies contig plot #35.*

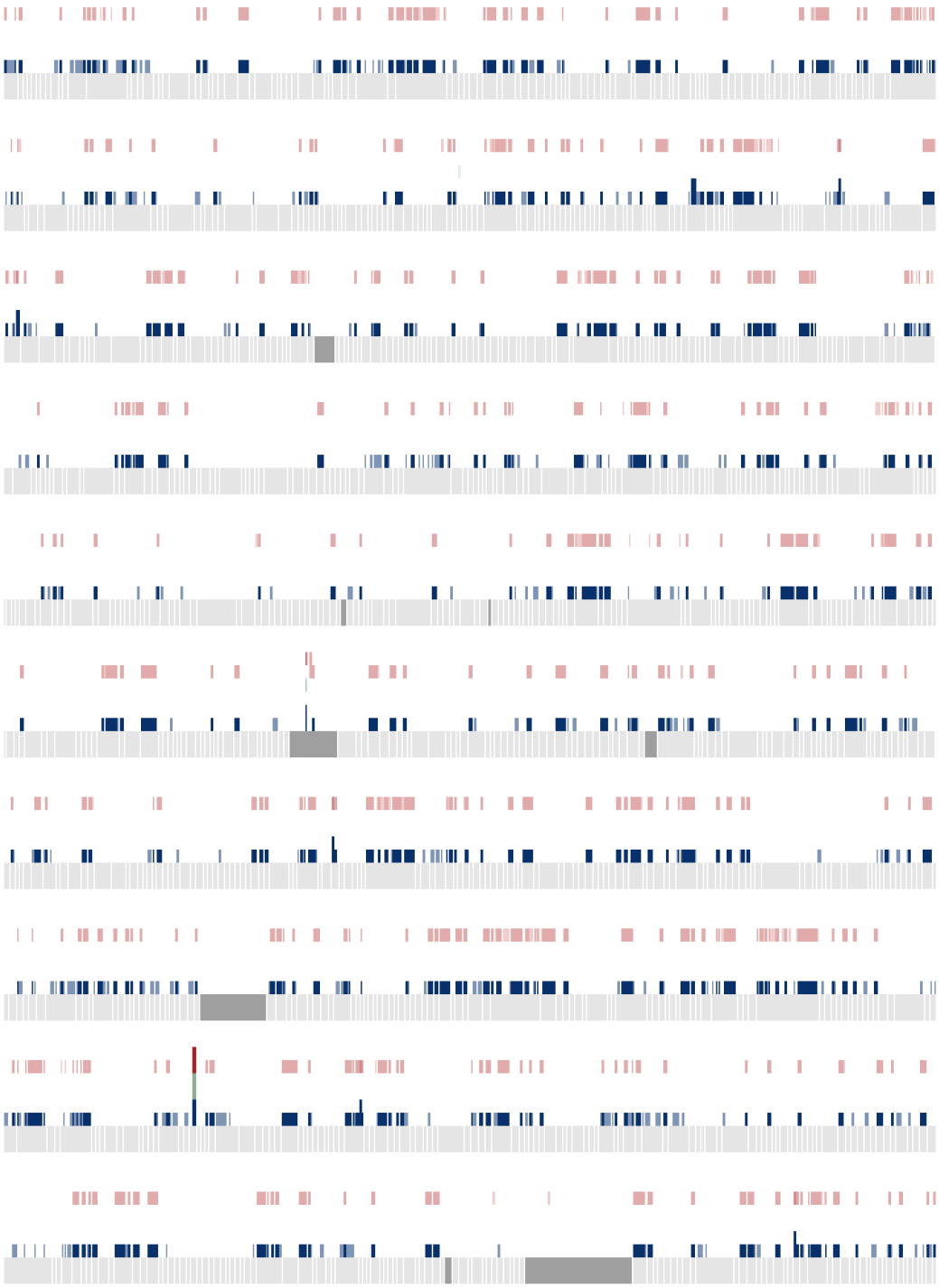

Figure D.43: *Picea abies contig plot #36.*

Figure D.44: *Picea abies contig plot #37.*

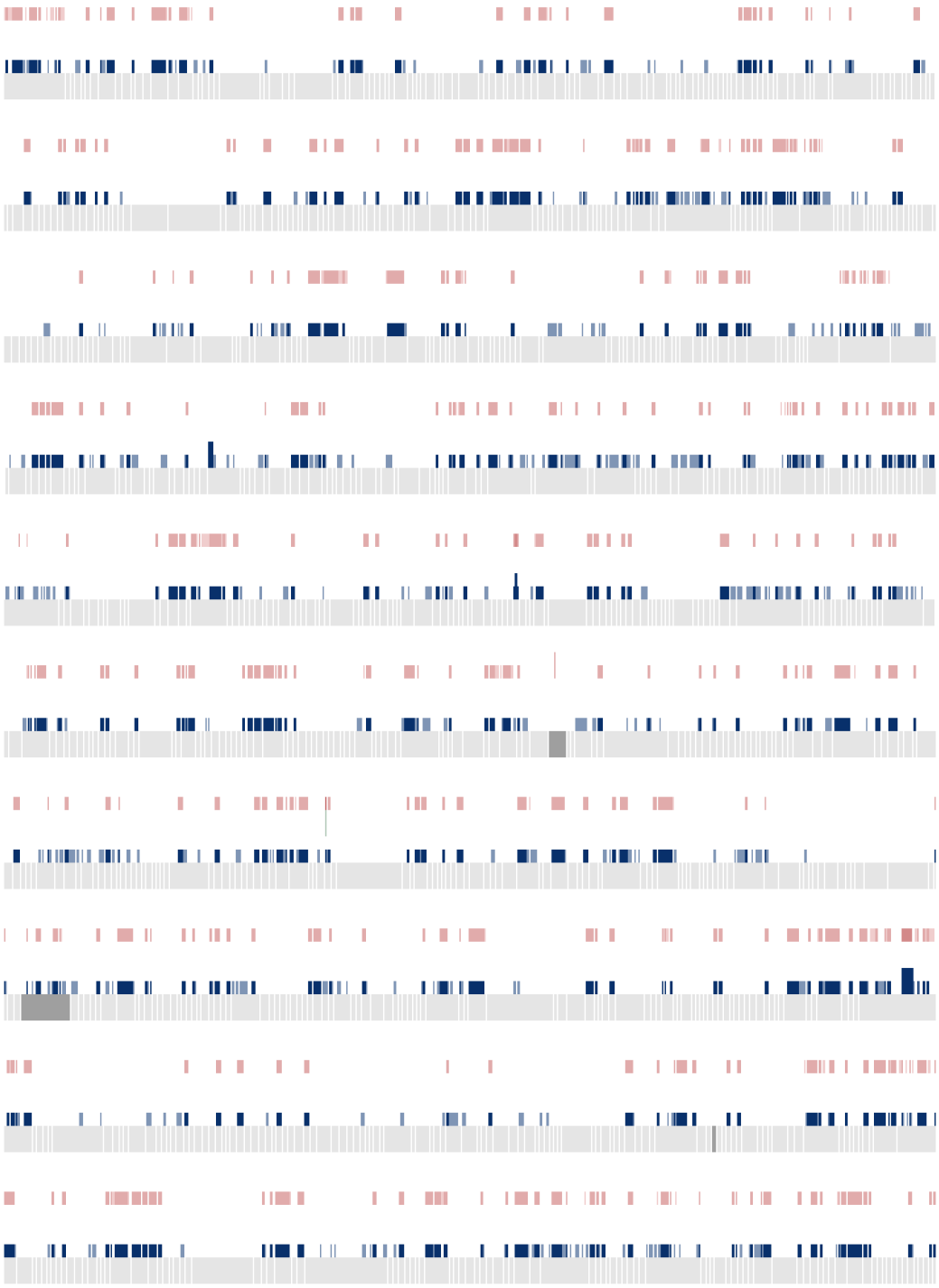

Figure D.45: *Picea abies contig plot #38.*

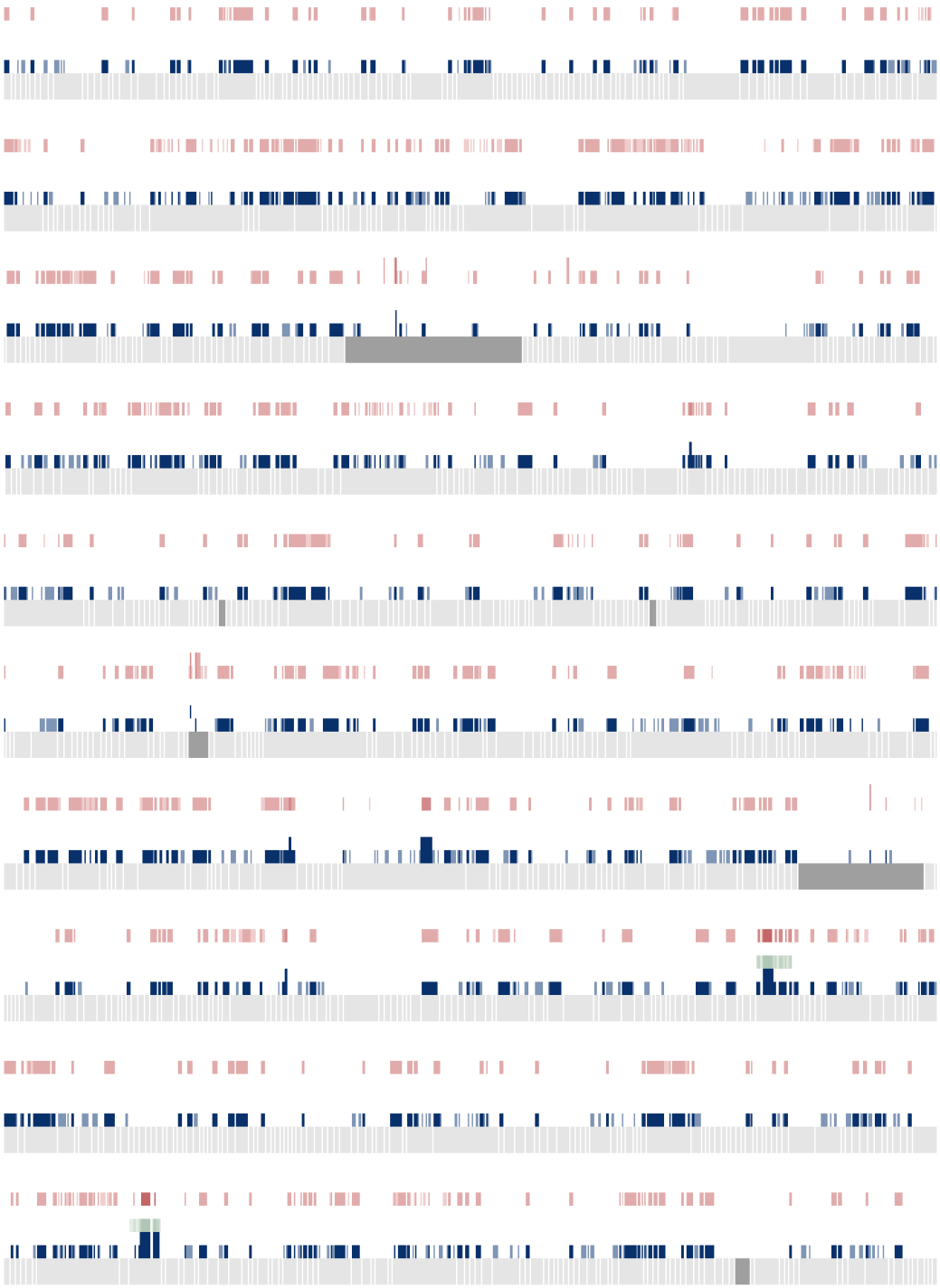

Figure D.46: *Picea abies contig plot #39.*

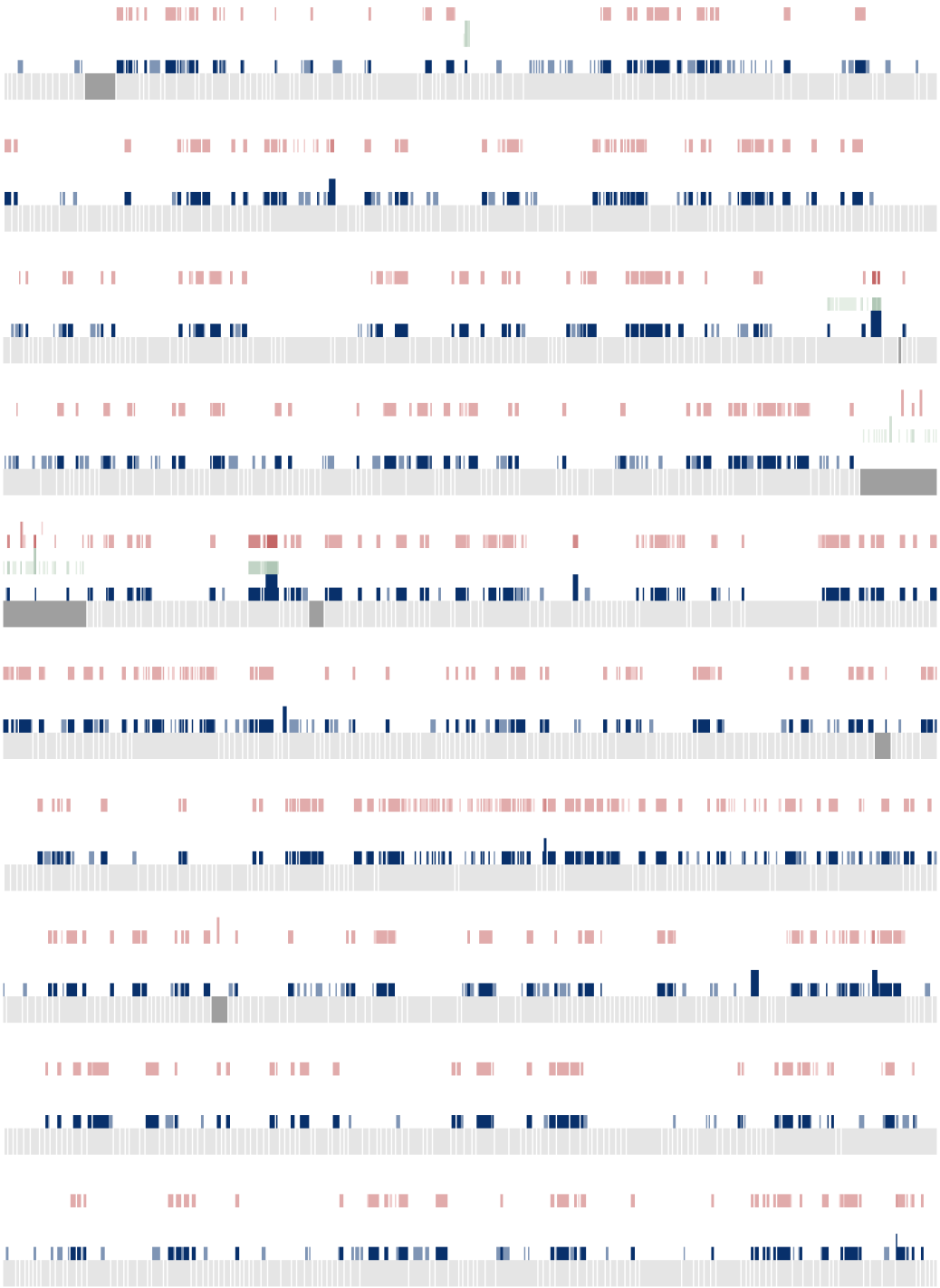

Figure D.47: *Picea abies contig plot #40.*

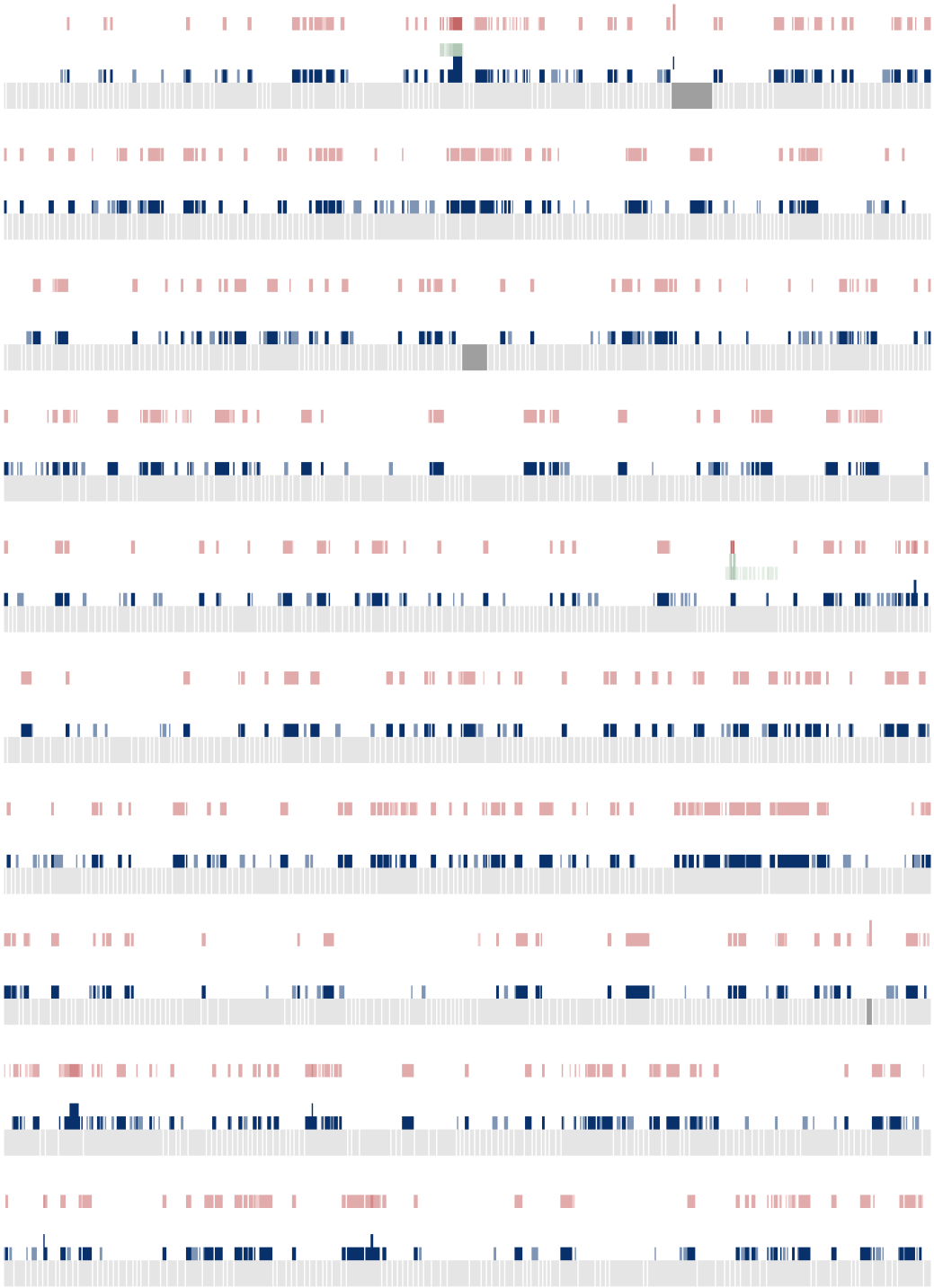

Figure D.48: *Picea abies contig plot #41.*

Figure D.49: *Picea abies contig plot #42.*

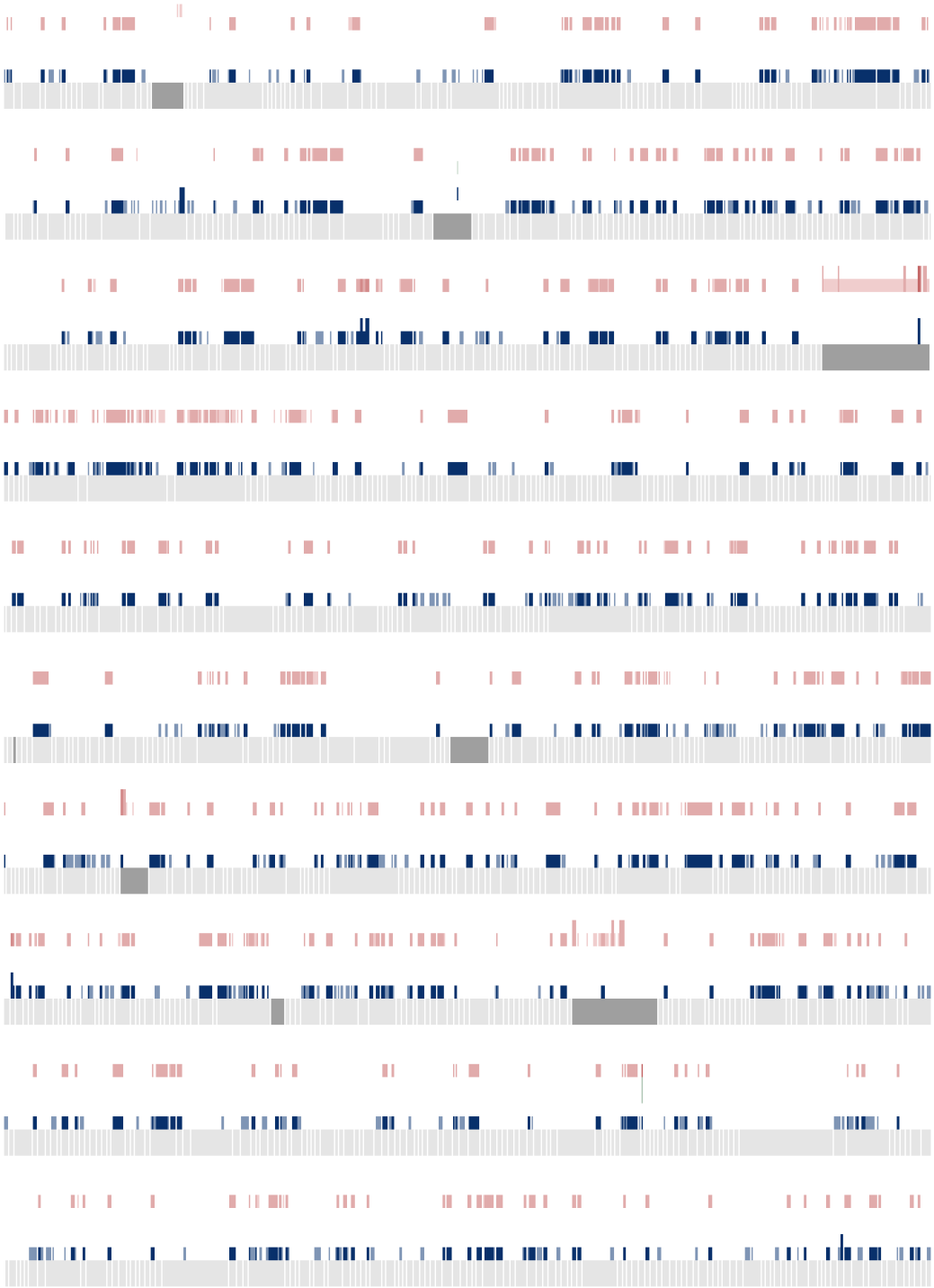

Figure D.50: *Picea abies contig plot #43.*

Figure D.51: *Picea abies contig plot #44.*

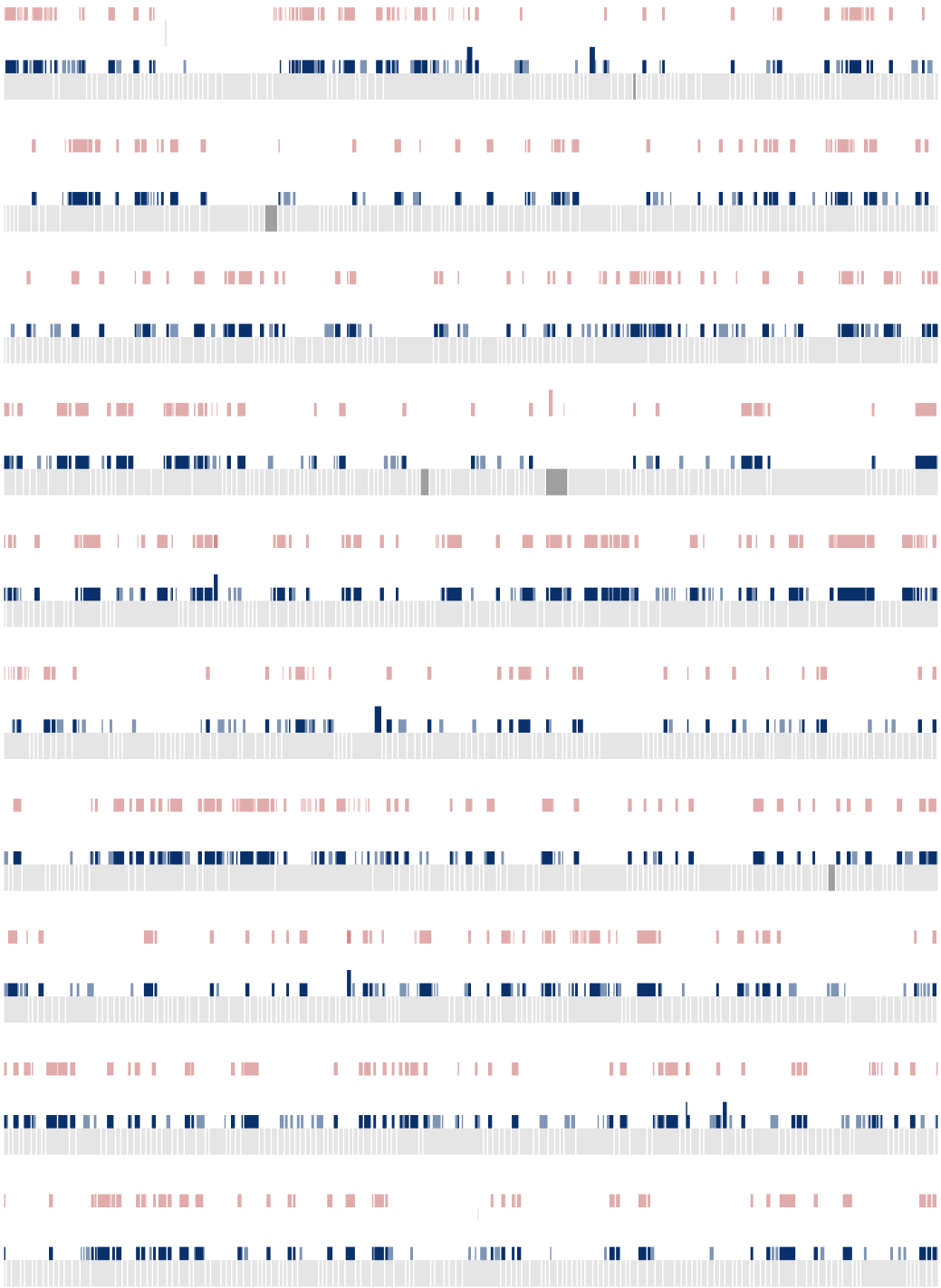

Figure D.52: *Picea abies contig plot #45.*

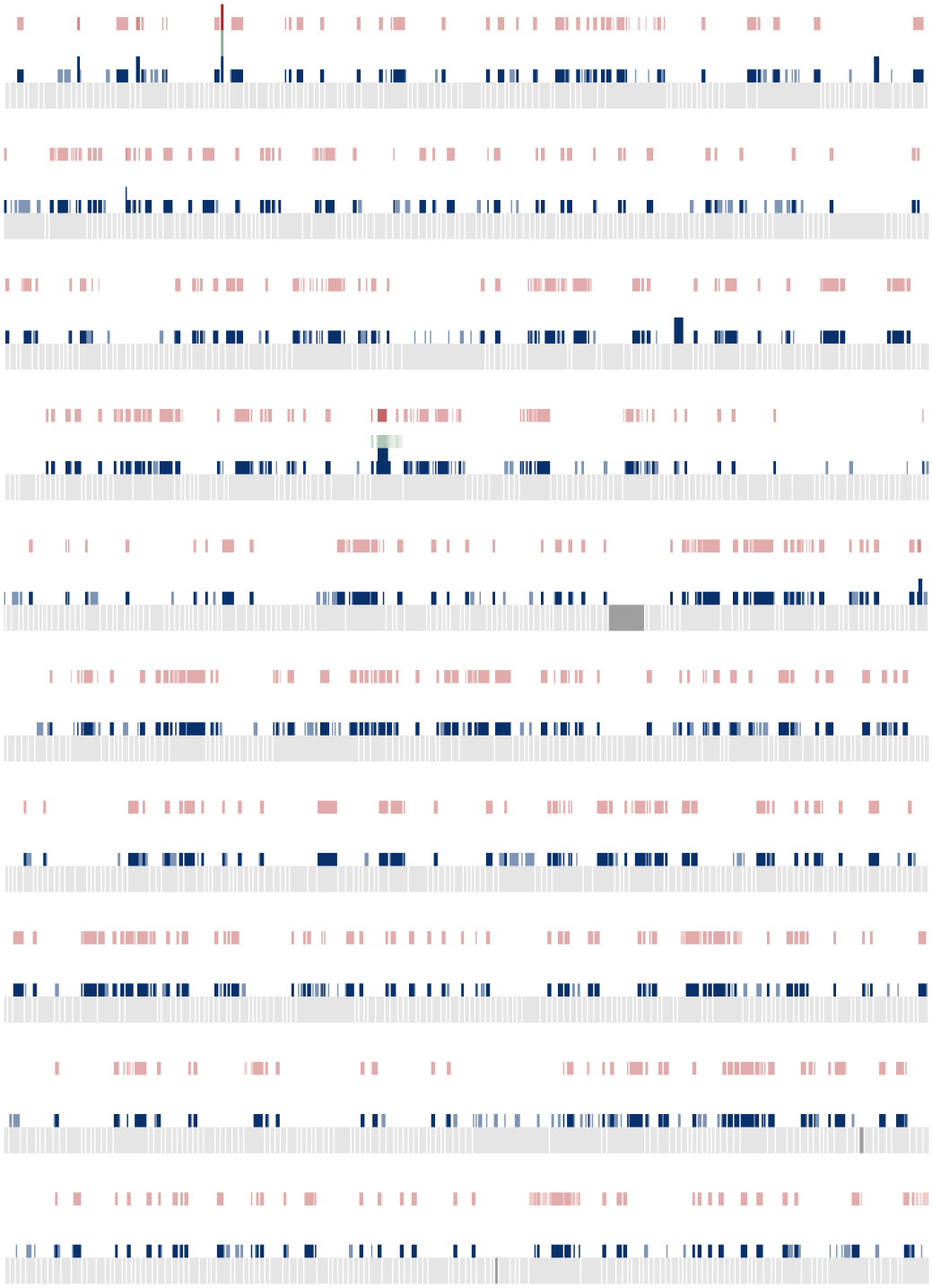

Figure D.53: *Picea abies contig plot #46.*

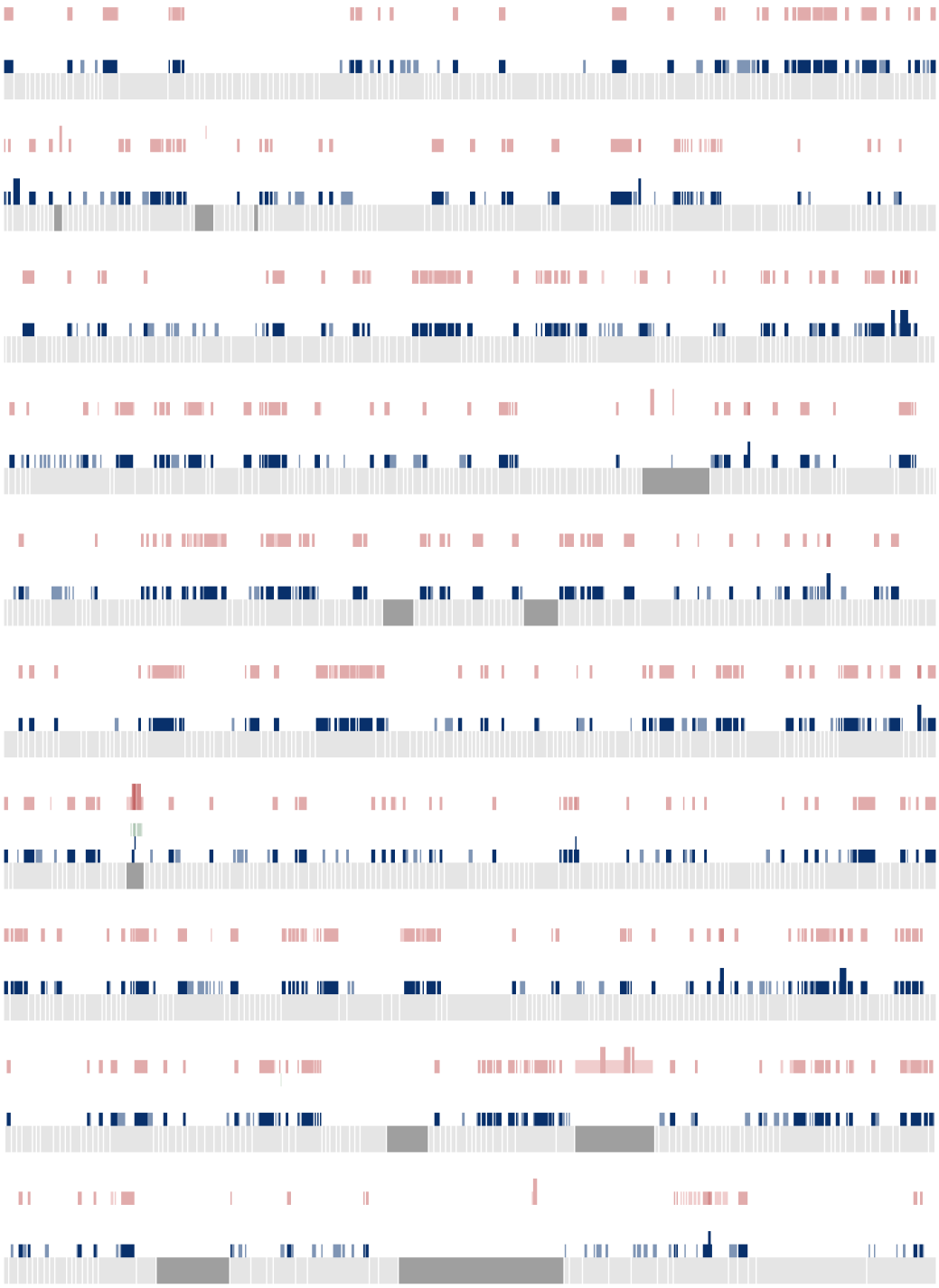

Figure D.54: *Picea abies contig plot #47.*

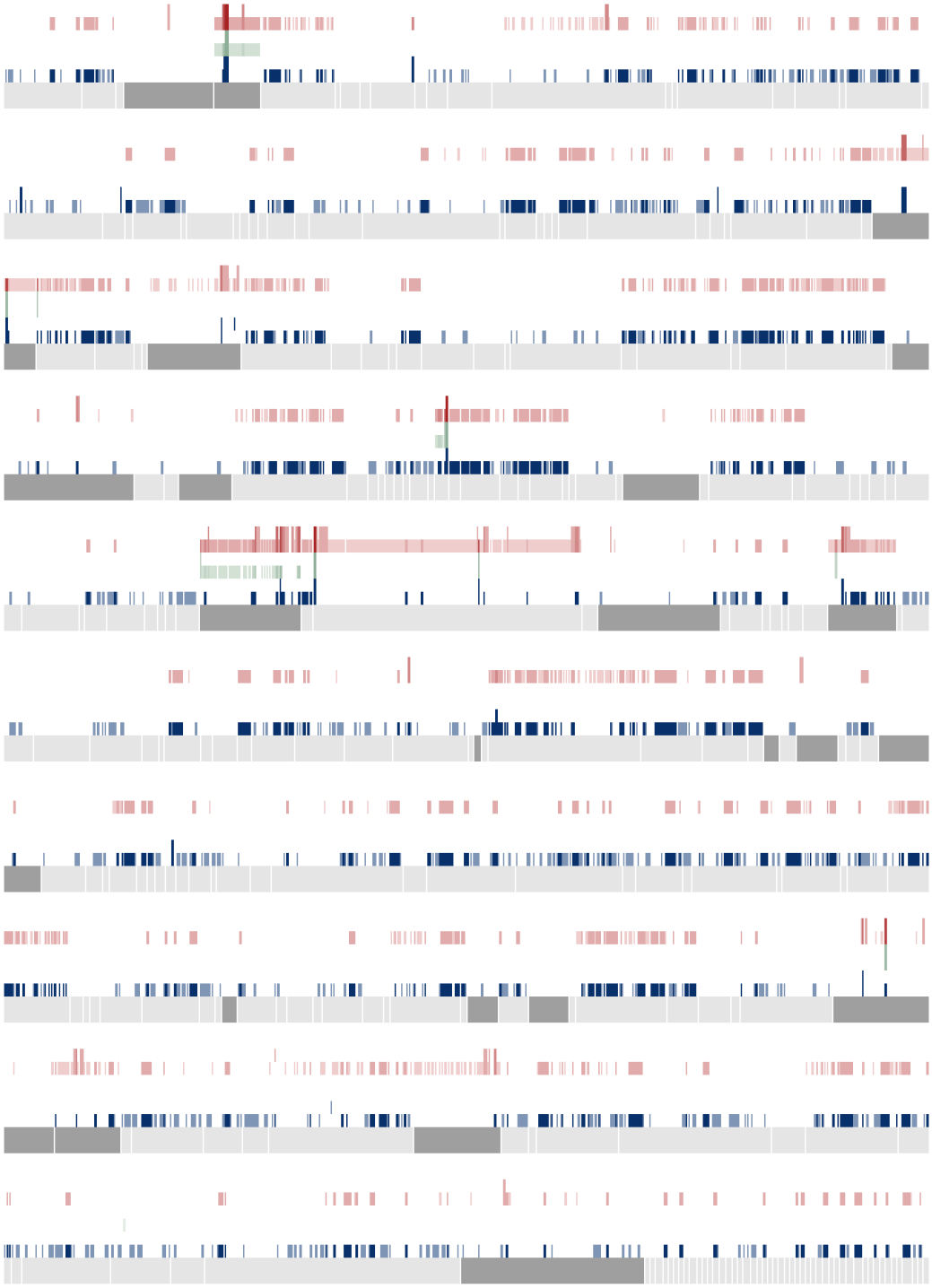

Figure D.55: *Picea abies contig plot #48.*

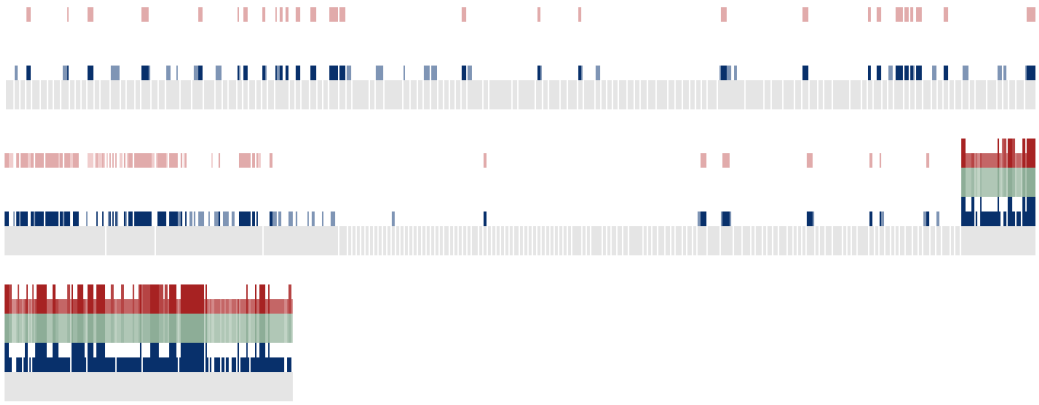

Figure D.56: *Picea abies contig plot #49.*

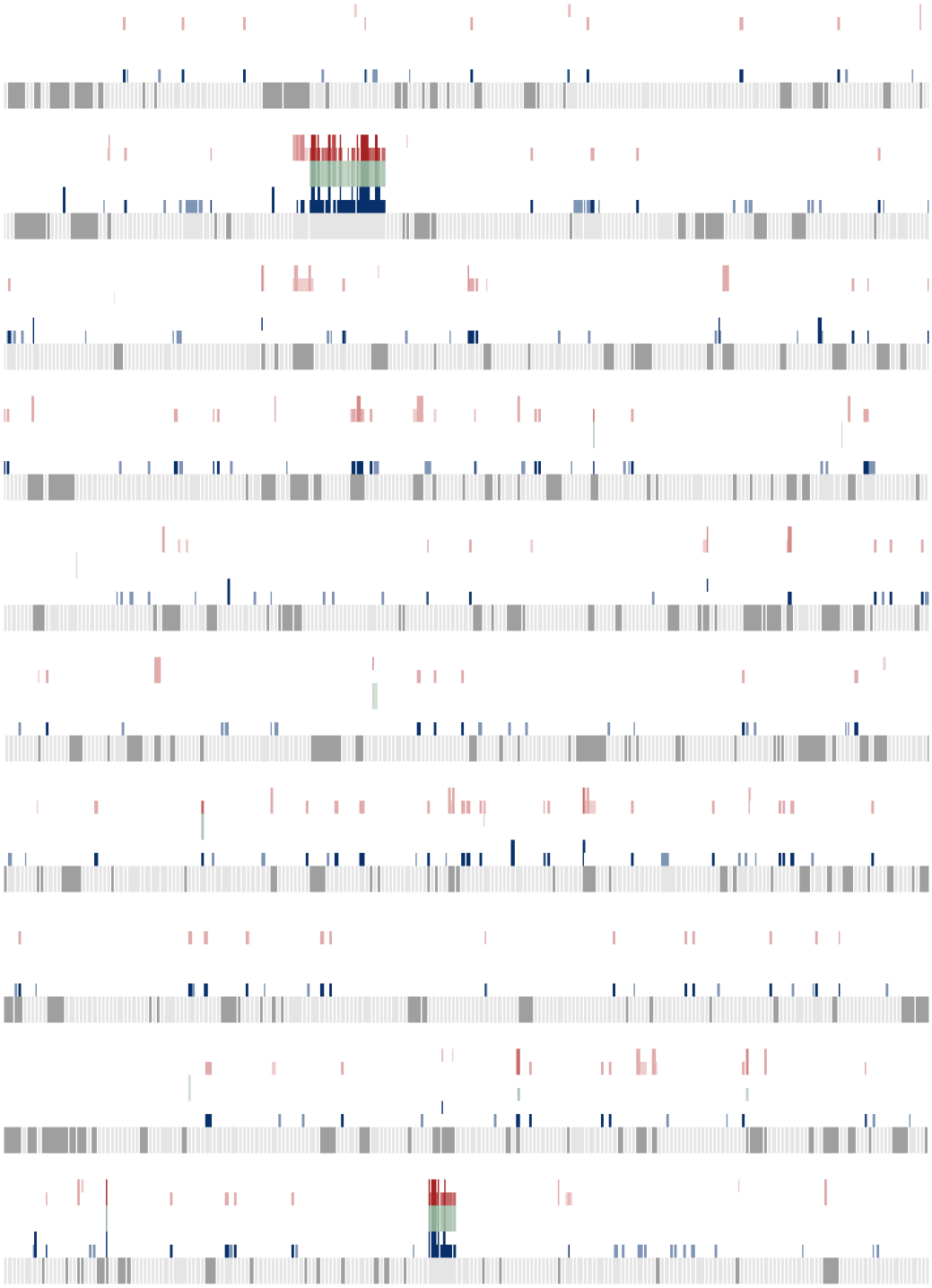

Figure D.57: *Pinus sylvestris contig plot #1.*

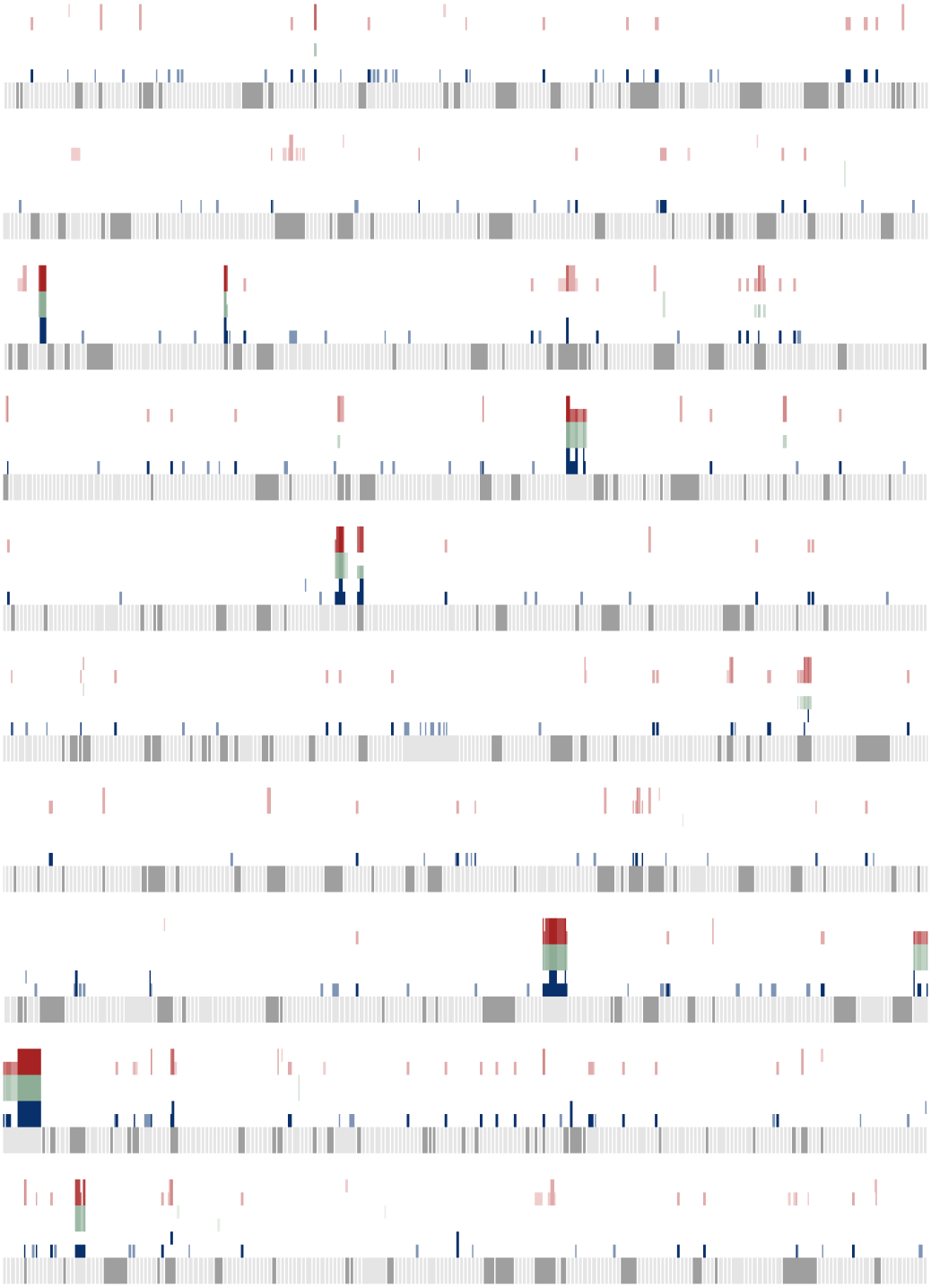

Figure D.58: *Pinus sylvestris contig plot #2.*

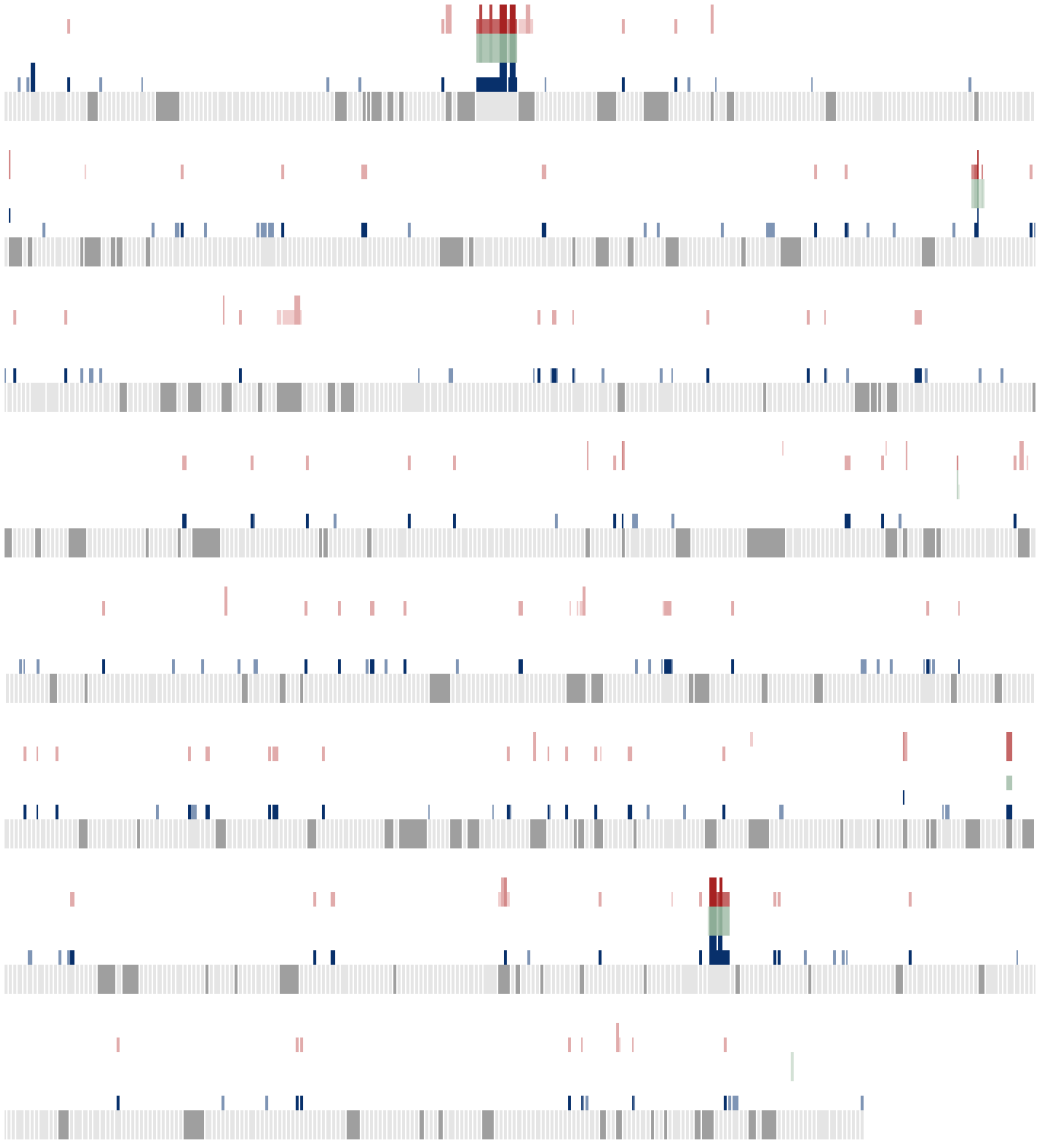

Figure D.59: *Pinus sylvestris contig plot #3.*

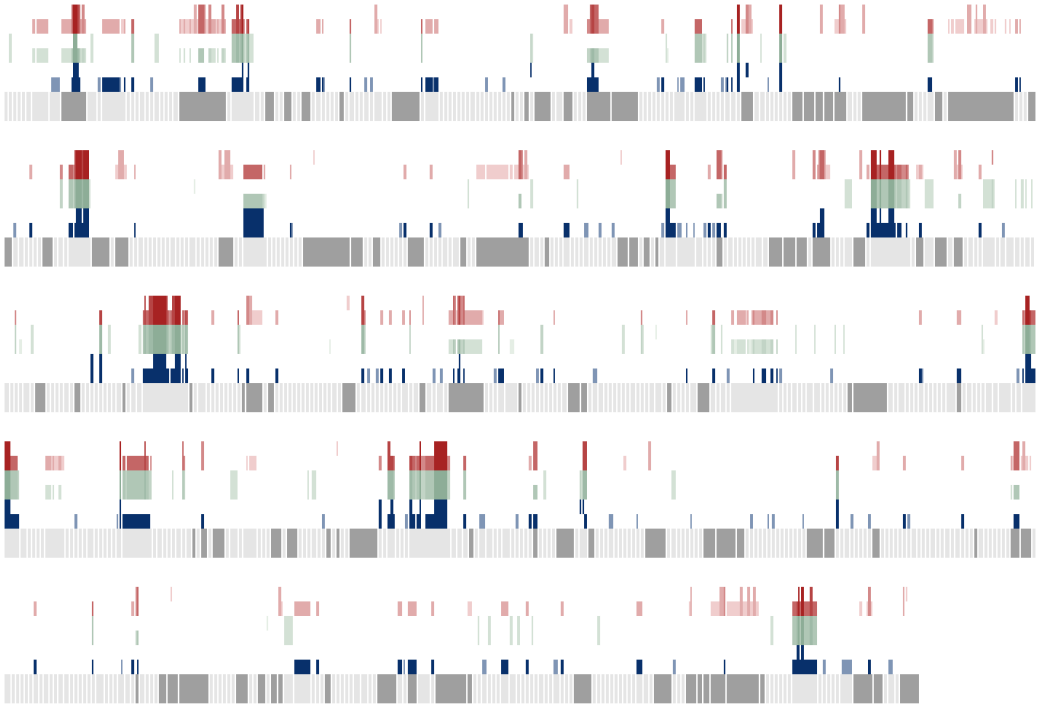

Figure D.60: *Taxus baccata contig plot #1.*