Whole-genome sequencing of river lamprey (*Lampetra fluviatilis*) and brook lamprey (*Lampetra planeri*): the first glimpse into comparative genomic divergences and similarities

Benedicte Garmann-Johnsen

Thesis submitted for the degree of
Master of bioscience (Ecology and evolution)
60 credits

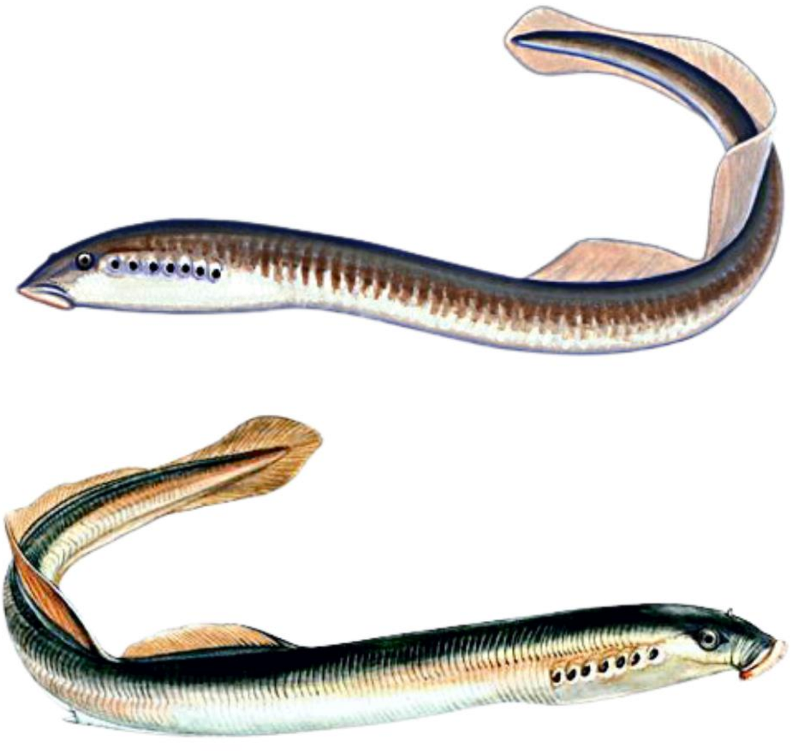
Centre for Ecological and Evolutionary Synthesis
Department of biosciences
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

June 2022

Whole-genome sequencing of river lamprey (*Lampetra fluviatilis*) and brook lamprey (*Lampetra planeri*): the first glimpse into comparative genomic divergences and similarities

Benedicte Garmann-Johnsen



Whole-genome sequencing of river lamprey (*Lampetra fluviatilis*) and brook lamprey (*Lampetra planeri*): The first glimpse into comparative genomic divergences and similarities

Benedicte Garmann-Johnsen

https://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

Front page illustration credit: Zsoldos Márton, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons. Illustration of Ukrainian brook lamprey and European river lamprey.

Abstract

Loss of biodiversity is one of the most pressing issues facing the health of our planet. In recent years, whole-genome sequencing has become a viable tool in understanding Earth's species composition. This study aims to conduct the first whole-genome sequencing of the species pair, the river lamprey (*Lampetra fluviatilis*) and brook lamprey (*Lampetra planeri*) to i) generate chromosome-level genome assemblies, for ii) initial comparative studies on genomic divergence and similarities, and iii) discuss my results in a broader evolutionary context on the species complex considering past genetic and genomic findings.

A combination of HiFi- and Hi-C-sequencing was used to generate long reads of >99% accuracy, while capturing chromatin interactions. The genome assemblies were generated by using the hifiasm-Hi-C-integrated assembler for the assembly process and the scaffolding was conducted using Pin_hic. A combination of MetaEuk and InterProScan was used for annotation. The mitochondrial genome of the brook lamprey was generated using MitoHiFi, for further phylogenetic inference of the species complex using published data of mitogenomes including river, Arctic, and sea lamprey.

The assembly process resulted in two fully haplotype-resolved chromosome-level assemblies (n = 82 chromosomes) per species. For the river lamprey, the genome size for haplotype 1 (RL1) was 963Mb, while haplotype 2 (RL2) was 945Mb. For the brook lamprey, haplotype 1 (BL1) was 894Mb, and haplotype 2 (BL2) was 996Mb. Comparative genomic analyses uncovered that the sequence similarities between the two species were as high as between the haplotypes of the same species, with 98.6% vs. 98.9% alignment-block-similarity on a whole-genome level, respectively. Additionally, 99.3% pairwise identity was found on an interspecies mitochondrial level. However, two large inversions, and significant differences in the number of structural variants between brook and river lamprey were identified, indicating that genomic variation (and possible genetic divergence) between the two species does occur.

Taken together, the genome assemblies generated in this study are of Earth BioGenome Project-standards and have given us new insight into whether the species complex, i.e. the river- and brook lamprey, should be classified as sub-species or ecotypes rather than two different species. Nevertheless, population genomic data is needed to fully characterise the genetic differentiation between the ecotypes.

Acknowledgements

To my supervisors: Thank you Asbjørn, for being a rock throughout this process. Not only have you provided me with invaluable input on everything related to lampreys, but you have also backed me up and made me feel confident in my work. I could not have asked for a better main supervisor. To Kjetill, thank you for your enthusiasm and great ideas, you are a fountain of knowledge, and I have learned so much from you. To Sissel, your expertise in genomics and passion in our meetings has made me so enthusiastic in my work and has given me so much inspiration throughout this process. To Siv, thank you for always being available to talk, and for giving me guidance on everything from writing to bioinformatics. Your input has helped me in so many ways, throughout my degree.

Ole Kristian Tørresen, it has been an absolute joy working with you. In this last year, I have learned more about the assembly process than I could have ever imagined, and I will be forever grateful for it. Thank you for your patience and taking the time to explain everything to me, in the most pedagogical way possible. Anders Krabberød, thank you for answering so many of my bioinformatic questions, and always making time for me when I drop by your office. You have really helped me out when I was in a pinch.

Thank you, Mikael Svensson, for providing me with all the brook lampreys, and thank you Eivind Schartrum for sending me the live river lamprey.

To all my friends and family, thank you for always being there for me, and for cheering me on. To my office mates at KB 3124, you have made my time at Blindern one of the greatest times of my life. Thank you for supporting me, for our picnics, ski trips, and our Monday-cake-tradition. You are the best.

And lastly, thank you to my fiancé Ola. You are the constant in my life, and I cannot wait to spend the rest of my life with you, and Tiger.

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities". Sequencing costs were covered by the Department of Biosciences (UiO) and Earth BioGenome Norway (EBP-Nor).

Table of contents

1. Introduction

We are currently in Earth's sixth mass extinction, and human-induced climate change is one of the main drivers behind this critical reduction in biodiversity (Pörtner et al., 2022). To conserve the current species richness, it is essential that we gain knowledge about the species that inhabit our globe. However, defining a species can be difficult, and in biology there is no definitive answer as to what constitutes a species (Futuyma, 2018; Ravinet & Sætre, 2019). This question becomes even more challenging when faced with what is known as cryptic species, i.e. species which can be genetically similar, whilst sharing similar morphologies (Futuyma, 2018; Ravinet & Sætre, 2019). Cryptic species can be hard to resolve into their own taxa, but this categorization is essential to understand and preserve biodiversity.

In 2021, 394 new fish species were described in Eschemeyer's Catalog of Fishes (Fricke et al., 2022), and of these 211 were freshwater species (Fricke et al., 2022). Although freshwater fish reside in lakes, rivers and streams, which are more easily accessible study areas than the ocean, it is often harder to determine whether they are their own distinct taxa. This is due to allopatry, i.e. geographical isolation which inhibits gene flow (Futuyma, 2018). Because of the environmental constraints of their habitats, freshwater fish can develop different life-histories and morphologies at a more rapid pace than their marine counterparts, with no natural means to determine whether they can hybridise to create offspring with higher fitness than their parent species. This type of hybridisation is more easily observed in marine habitats, where there is potential for interbreeding and gene flow without the limitations of geographical isolation.

One of the most studied freshwater species which has posed such taxonomic descriptive challenges is the brown trout (*Salmo trutta*) (Jonsson & Jonsson, 2011). The grouping of this species has been the topic of discussion in various different studies (Freyhof & Kottelat, 2007; McKeown et al., 2010; Webb et al., 2007), and as of today more than 60 different phenotypes have been described (Jonsson & Jonsson, 2011). This is also the case for the Arctic charr (*Salvelinus alpinus*) living in Lake Tinnsjøen. In this polymorphic species complex, four morphs reside on different depths of the lake (Østbye et al., 2020). Here, they inhabit different niches, and have adapted different morphologies in accordance with their varying life-histories (Østbye et al., 2020). The four charr morphs not only display characteristic morphologies and life-histories, but through mitochondrial analyses they form distinct genetic clusters implying that the morphs form four genetic populations (Østbye et al., 2020).

Gauging what constitutes a species has been difficult in the lamprey family as well since many taxa form paired species. The pairs often consist of a freshwater-resident species, which matures early, and at a smaller size, and an anadromous, migratory, and parasitic species. An example of this type of species pair is the migratory and parasitic European river lamprey (*Lampetra fluviatilis*) and the non-migratory and non-parasitic brook lamprey (*Lampetra planeri*). These species have been the subject of several previous genetic studies, using mtDNA (mitochondrial DNA) (Bracken et al., 2015; De Cahsan et al., 2020), RAD-sequencing (restriction-site associated DNA sequencing) (Hume et al., 2018; Mateus et al., 2013; Rougemont et al., 2017), and microsatellite markers (Rougemont et al., 2015). Based on these, and many other studies, there is no definitive consensus as to if these two taxa are separate species, or merely ecotypes, with different life-history traits, within same species.

While the river lamprey and brook lamprey are morphologically and behaviourally similar in their larval stages, sustaining themselves through filter feeding at the bottom of freshwater streams for the first five to seven years of their lives (Potter et al., 2015; Rougemont et al., 2015), they differ greatly upon entering maturity. Here, the brook lamprey develops eyes and the characteristic lamprey sucker mouth, and stops feeding, only to mate and die in the freshwater where it has spent its entire life (Rougemont et al., 2015). The river lamprey, however, following a metamorphosis, enters a migratory, and often anadromous, parasitic juvenile life stage, where it migrates (to lakes or saltwater), to feed on larger fish. For up to three years, the juvenile river lamprey lives as a parasite (Kelly, 2001; Rougemont et al., 2016). When entering sexual maturity, the river lamprey returns to running water to mate, and die (Kelly, 2001; Rougemont et al., 2016).

It is not known whether the differences between the two species are due to genetics, or if the driver behind these types of morphological and life-history differences is due to phenotypic plasticity. Phenotypic plasticity is defined as differences in individual phenotype due to the environment (Futuyma, 2018). Examples of plastic responses are phenological shifts, i.e. changes in reproductive timing and ontological shifts, i.e. changes in developmental timing and various morphological changes such as size, colouration and structure. For example, the presence of migratory and stationary phenotypes of male Atlantic salmon (*Salmo salar*) is clearly due to phenotypic plasticity (Glover et al., 2018).

Whether the differences observed are a result of plasticity, or genetic differences, having accurate ways to measure within-species variation are extremely important in conserving biodiversity. To assess this kind of diversity, such as the differences between paired species like the river lamprey, and brook lamprey, whole-genome sequencing is an essential tool for generating state-of-the-art reference genomes and doing population genomics. With high-throughput sequencing methods becoming cheaper, faster, and more accurate, this is a viable solution in the fight against loss of biodiversity. The aim of the “moonshot for biology”, the Earth Biogenome Project, is to sequence all eukaryotic life to drive new solutions for preserving biodiversity (Lewin et al., 2022). The project also aims to create a better understanding of biology and evolution, and to facilitate biotechnology innovations benefiting human society (Lewin et al., 2022). As a part of this project, I am conducting the first whole-genome sequencing of both the river lamprey and brook lamprey, using a combination of HiFi- and Hi-C-sequencing.

HiFi, or High Fidelity-sequencing uses isolated DNA or RNA to create a circularized SMRTbell library, SMRT being the sequencing system developed by Pacific Biociences (PacBio) to create accurate long reads (Pacific Biosciences of California, 2020). The advantage of this type of sequencing is that it provides longer fragments for assembly scaffolding, without compromising accuracy (Giani et al., 2020). In addition to this, I will capture the chromatin’s three-dimensional structure using Hi-C sequencing. Here, the DNA is cross-linked to preserve genomic and chromosomal interactions (Ghurye & Pop, 2019; Ghurye et al., 2017). In this study, I combine these sequencing methods, with the aim of creating accurate, annotated genome assemblies for both the river lamprey and the brook lamprey. In addition to creating these assemblies, I am comparing the two species genomes, to assess their degree of genomic differentiation. In light of previous genomic research, I address whether they are two different species, or merely ecotypes of the same species, with different plastic responses to their environment.

2. Materials and methods

2.1 Collection of samples and shipping

River lamprey: The river lamprey was caught in Ådalsåa in Telemark, Norway 21.04.2021 using electrofishing (Bohlin et al., 1987). The specimen was transported alive in a water-filled plastic container to the University of Oslo.

Brook lamprey: The brook lamprey was caught in Hunserödsbäcken in Skåne, Sweden 27.10.2020. The Hunserödbäcken-brook is an obstacle to migration, meaning no migratory species (e.g. the river lamprey) can pass. The specimen was sampled using electrofishing (Bohlin et al., 1987), euthanised on site, and shipped to the University of Oslo in 96% ethanol.

2.2 Dissection and storage

River lamprey: The adult river lamprey arrived at the University of Oslo 23.04.2021 and was euthanised using tricaine methanesulfonate (MS-222) fish anaesthetic. The fish measured 17 centimetres from snout to tail fin. Shortly after euthanasia a blood sample was collected using a syringe. In addition to the blood sample, gonad-, heart-, head kidney-, muscle-, fin-, liver-, gill-, gut-, and mouth tissues were extracted (see Figure 1). All the extracted tissues were snap-frozen in individual Eppendorf tubes using liquid nitrogen. The muscle- and heart tissue was transferred immediately after the dissection to the Norwegian Sequencing Centre for library preparation. All samples, including the rest of the fish body, was stored at minus 80 degrees Celsius.

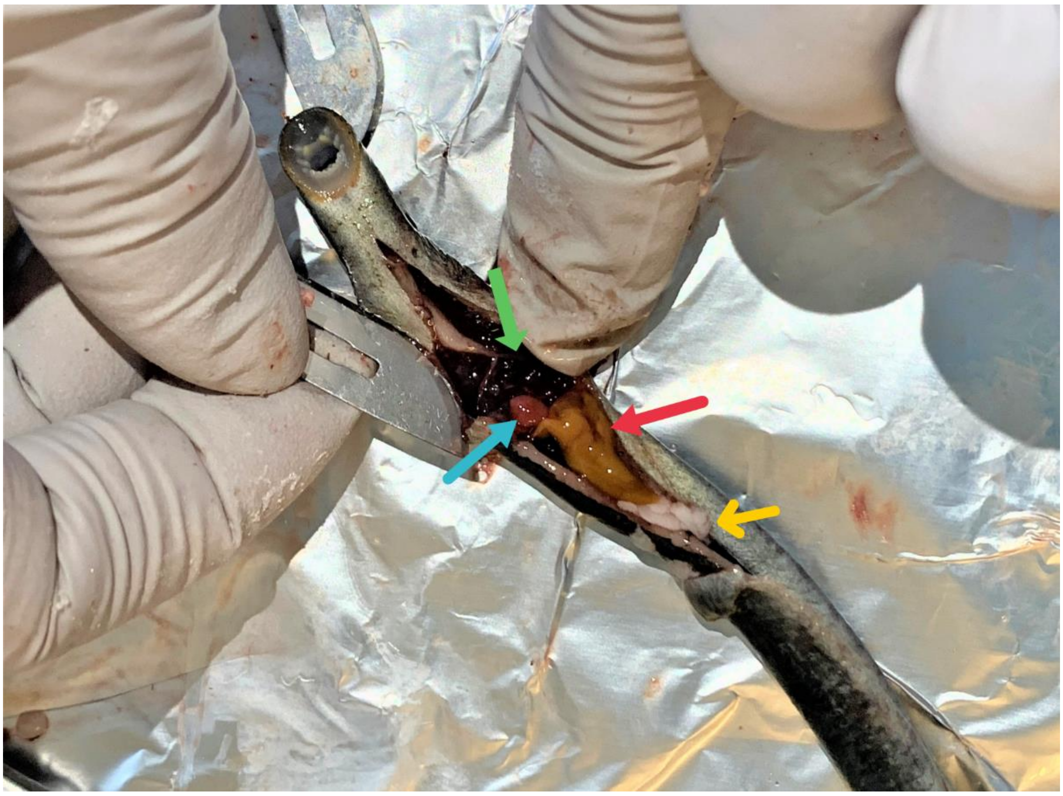


Figure 1: Picture of the river lamprey after the first incision during the dissection. In the image the heart is marked with a blue arrow, the gonads are marked with a red arrow, the intestines are marked with a yellow arrow, and the liver is marked with a green arrow.

Brook lamprey: In total, five brook lampreys were shipped to the University of Oslo, arriving at October 28. 2020. Of the five individuals, two larvae and one adult were dissected. The adult measured 12.2 centimetres from the tip of the snout to the end of the tail fin (see Figure 2). From the adult, muscle- and skin tissue, gill filaments, and the entire heart was dissected. After dissection, all tissues from the adult individual were transferred to the sequencing centre for library preparation, and the rest of the fish bodies were stored in individual containers at minus 80 degrees Celsius.

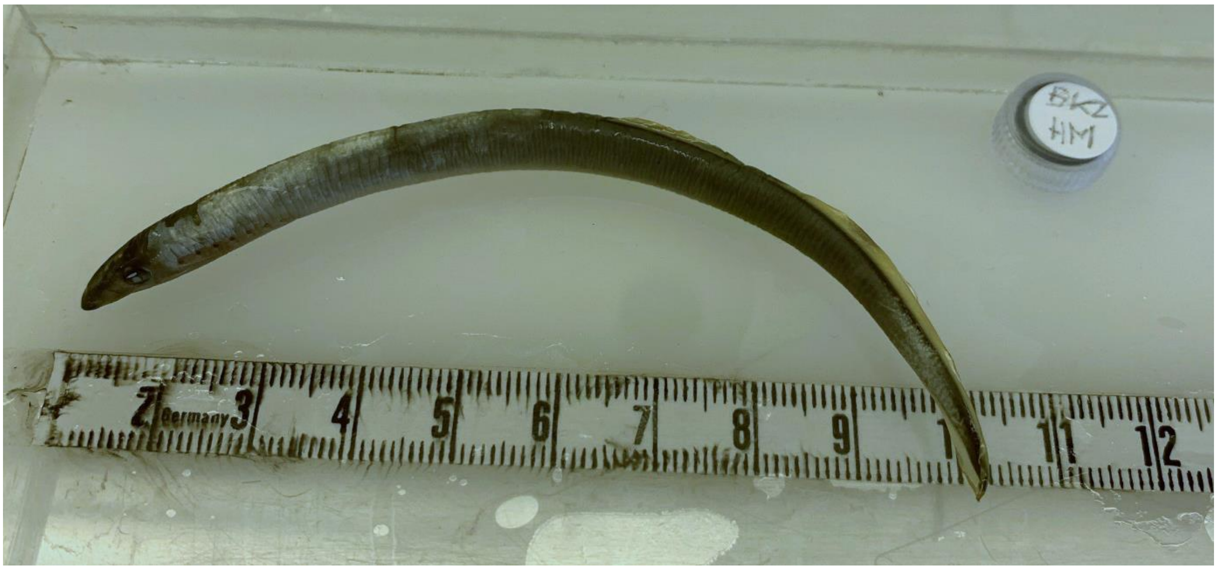


Figure 2: Picture of the adult brook lamprey individual before dissection. The individual measured 12.2 centimetres from the tip of the snout to the end of the tail fin when stretched.

2.3 Sequencing

2.3.1 HiFi sequencing

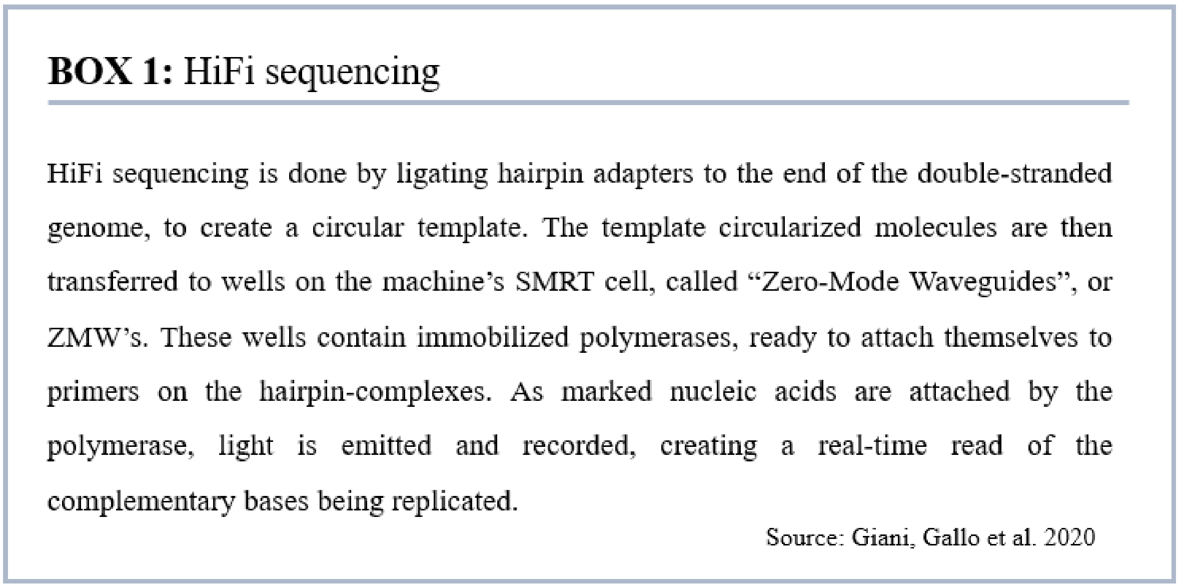
River lamprey: HiFi sequencing was chosen for its ability to generate high fidelity reads, without compromising read length (see Box 1) (Giani et al., 2020; Wenger et al., 2019). The following protocols were conducted by the Norwegian Sequencing Centre. In preparation for HiFi sequencing, two libraries were made from the muscle tissue, following the Pacific Biosciences protocol "Preparing HiFi SMRTbell® Libraries using the SMRTbell Express Template Prep Kit 2.0" (See Appendix A). The size selection for the final libraries, i.e., the process where suboptimal nucleic fragments are removed, was determined using BluePippin with an 11 kb cut-off (Wang et al., 2021).

The SMRTbell libraries were transferred to three 8M SMRT cells in the PacBio Sequel II System and placed into the Zero-Mode Waveguide (ZMW) wells (Korlach et al., 2010), to bind to polymerases and generate circularised consensus reads (Giani et al., 2020).

The first river lamprey SMRT cell was sequenced using the Sequel II Binding kit v2.0 and Sequencing chemistry v2.0, with loading performed by diffusion. Between the first and second sequencing runs, the Sequencing chemistry v2.2 was launched, therefore the second SMRT cell was sequenced using this kit combination, with adaptive loading. However, the v2.2 polymerase did not work well with the sample, thus the v2.0 Binding Kit was used for the third run. Pre-extension for SMRT cells one and three was two hours, and the movie time for all three SMRT cells was 30 hours. The mean number of passes for the circularized template were 10 for SMRT cell one, 9 for SMRT cell two, and 8 for SMRT cell three.

Brook lamprey: Two libraries were prepared using the muscle- and skin tissue, following the Pacific Biosciences protocol "Preparing HiFi SMRTbell® Libraries using the SMRTbell Express Template Prep Kit 2.0" (See Appendix A). The size selection for the final libraries were determined using BluePippin with an 11 kb cut-off for all samples prepared (Wang et al., 2021).

The SMRTbell libraries were transferred to three 8M SMRT cells in the PacBio Sequel II and placed into the ZMW-wells (Korlach et al., 2010), to bind to polymerases and generate circularised consensus reads (Giani et al., 2020). The Sequel II Binding kit v2.0 was used, in combination with Sequencing chemistry v2.0, and loading was performed by diffusion, with a pre-extension of two hours, and movie time of 30 hours. The mean number of passes for the circularized template were 9 for all three SMRT cells.



BOX 1: HiFi sequencing

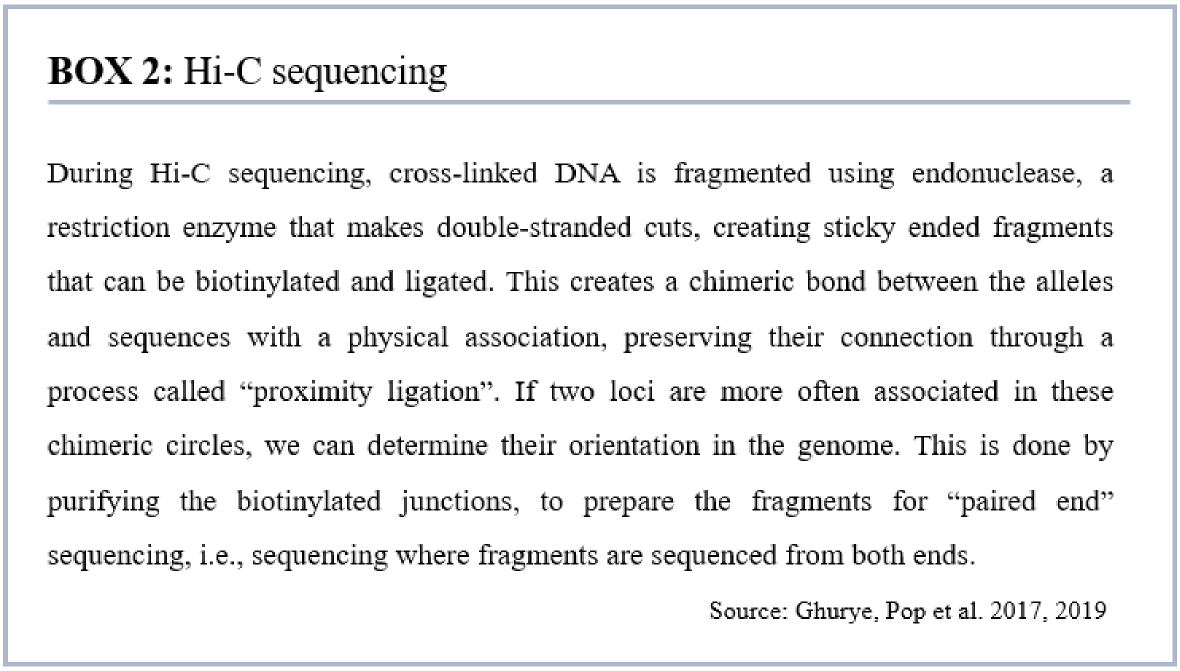
HiFi sequencing is done by ligating hairpin adapters to the end of the double-stranded genome, to create a circular template. The template circularized molecules are then transferred to wells on the machine's SMRT cell, called "Zero-Mode Waveguides", or ZMW's. These wells contain immobilized polymerases, ready to attach themselves to primers on the hairpin-complexes. As marked nucleic acids are attached by the polymerase, light is emitted and recorded, creating a real-time read of the complementary bases being replicated.

Source: Giani, Gallo et al. 2020

2.3.2 Hi-C sequencing

River lamprey: Hi-C sequencing was used to capture the three-dimensional chromatin structure, through Illumina short-reads (see Box 2) (Ghurye & Pop, 2019). The protocol used for library preparation of the river lamprey was the "Omni-C Proximity Ligation assay for Non-mammalian samples, version 1.0" (see Appendix A). Here, 20 mg of fresh, snap-frozen, heart tissue was ground to a fine powder, lysed, and proximity ligated in preparation for sequencing on the NovaSeq 6000 Sequencing System at the Norwegian Sequencing Centre. Following library preparation, one full S Prime NovaSeq Flow Cell was used for 2 x 150 bp paired end sequencing.

Brook lamprey: 100 mg of gill tissue, which had been stored in ethanol, blotted on paper, and weighed, was prepared using the "Arima Genome-Wide HiC+ Kit". Following this, the "Arima-HiC 2.0 kit standard user guide for Animal tissues"-protocol (see Appendix A) was used for library preparation. The library was sequenced on the NovaSeq 6000 at the Norwegian Sequencing Centre. One quarter of a NovaSeq Flow Cell was used for 2 x 150 bp paired end sequencing.



BOX 2: Hi-C sequencing

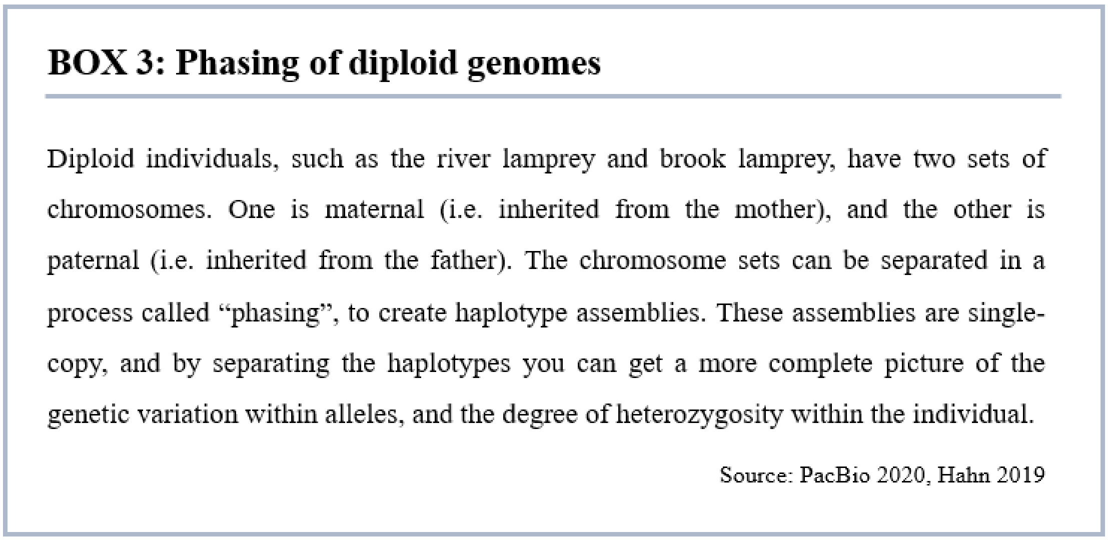
During Hi-C sequencing, cross-linked DNA is fragmented using endonuclease, a restriction enzyme that makes double-stranded cuts, creating sticky ended fragments that can be biotinylated and ligated. This creates a chimeric bond between the alleles and sequences with a physical association, preserving their connection through a process called "proximity ligation". If two loci are more often associated in these chimeric circles, we can determine their orientation in the genome. This is done by purifying the biotinylated junctions, to prepare the fragments for "paired end" sequencing, i.e., sequencing where fragments are sequenced from both ends.

Source: Ghurye, Pop et al. 2017, 2019

2.4 Assembly

The data from the HiFi sequencing runs were received as HiFi reads in FASTQ format. All reads delivered had at least 99% accuracy. To find the most suitable assembly pipeline, the reads were assembled using multiple different programs, such as hifiasm v0.15.2 (r334) (Cheng et al., 2021) via Anaconda v4.12.0 (Anaconda Sofware Distribution, 2020), Flye v2.9 (Kolmogorov et al., 2019), and HiCanu v2.1 (Nurk et al., 2020). The hifiasm assemblies were made using the default settings (Cheng et al., 2021). For Flye, two different assemblies were made for each species; one with the default settings, i.e. a minimum overlap of 10 000 bp, and the other with a minimum overlap of 3000 bp (Kolmogorov et al., 2019), which was the recommended setting for long reads. The three HiCanu assemblies for each species were made using three different minimum overlap settings, 200 bp, 500 bp and 700 bp (Nurk et al., 2020). The HiFi reads were also assembled by the Norwegian Sequencing Centre using the “Genome Assembly pipeline (SMRT Link 10.1.0.119588)” with default settings (see Appendix A). After completing the various assemblies, comparisons were performed using QUAST v5.0.2 (Mikheenko et al., 2018) and BUSCO v5.0.0 (Manni et al., 2021). These two comparative tools provide assembly metrics such as number of contigs, total assembly length, NG50, and GC-content (QUAST), and the assembly completeness based on orthologs (BUSCO). Based on the results of these comparisons, the hifiasm assemblies were chosen for Hi-C integration.

The data from the Hi-C sequencing runs were received as paired end reads in FASTQ format. For the Hi-C data, the hifiasm assembler had a built-in integration feature, and this was used to create haplotype-resolved de novo assemblies for both species, without having to run additional programs (Cheng et al., 2022). Haplotype resolution, also known as “phasing”, sorts the diploid genome into two distinct assemblies (haplotypes) based on their maternal or paternal origin (see Box 3) (Hahn, 2019). After running the assembler with the data from the HiFi and Hi-C sequencings runs, three GFA files were created for both species; one for each haplotype, and one combination of the two, where the assembler switches between each haplotype to create the longest possible contigs. These were converted to six FASTA files (RL1=river lamprey haplotype 1, RL2=river lamprey haplotype 2, RLP=river lamprey longest contigs, BL1=brook lamprey haplotype 1, BL2=brook lamprey haplotype 2, and BLP=brook lamprey longest contigs) and quality assessed using BUSCO v5.0.0 (Manni et al., 2021).



BOX 3: Phasing of diploid genomes

Diploid individuals, such as the river lamprey and brook lamprey, have two sets of chromosomes. One is maternal (i.e. inherited from the mother), and the other is paternal (i.e. inherited from the father). The chromosome sets can be separated in a process called “phasing”, to create haplotype assemblies. These assemblies are single-copy, and by separating the haplotypes you can get a more complete picture of the genetic variation within alleles, and the degree of heterozygosity within the individual.

Source: PacBio 2020, Hahn 2019

2.5 Scaffolding and manual curation

2.5.1 Scaffolding with Pin_hic

For both species, the two scaffolding tools Salsa v2.3 (Ghurye et al., 2017) and Pin_hic v3.0.0 (Guan et al., 2021) were used for scaffolding the FASTA files containing RL1, RL2, RLP, BL1, BL2, and BLP. BUSCO v5.0.0 (Manni et al., 2021) and Assemblathon_stats (Earl et al., 2011), a genome statistics tool which presents metrics such as number of scaffolds and mean scaffold size, were used to compare the scaffolds. The scaffolds with the highest complete BUSCOs (i.e. the highest number of complete- and single-copy orthologs), and lowest number of scaffolds, were the ones generated using Pin_hic. These were determined to be best suited for further curation.

2.5.2 The Rapid curation suite

Picard v2.22.1 (Broad Institute, 2019) was used to convert the FASTQ files generated during Hi-C sequencing to one BAM file for each species, containing the unmapped paired end reads. The unmapped reads were then mapped against the Pin_hic-scaffolds in a Singularity image (Kurtzer et al., 2017) using the "Rapid curation suite" (GRIT: Genome Reference Informatics Team, 2022). This software suite consisted of five individual steps; 1. "HiC", which generated HiGlass-files and a .pretext-map suitable for manual curation in the PretextView v0.2.4 desktop application (see Figure 3 a) and b)) (Harry, 2021), 2. "Coverage Track", which was used to assess the depth of the HiFi reads, 3. "Repeat Track", which identified the repeated areas in the assembly, 4. "Gap Track", which identified the gaps in the assembly, and 5. "Telomere Track", which identified the telomeres (GRIT: Genome Reference Informatics Team, 2022). The outputs from steps 2-5 created bigwig- and bedgraph-files which were used to create graph overlays for HiGlass and PretextView, and these were used as guides during the manual curation.

During curation in PretextView, a log of the gaps filled was kept in TPF-files created from the Pin_hic-scaffolded FASTA files – one for each haplotype. Upon completion, each of the two manually curated phased scaffolds were painted, i.e. given a scaffold name and number. They were then exported as AGP files and fitted against the TPF files to see if the changes made in the AGP matched the gaps filled in the TPF. For this fitting, the rapid_pretext2tpf_XL.py-script was used (GRIT: Genome Reference Informatics Team, 2022). After assessing the fit quality, the rapid_join.pl-script was used to create new FASTA files for each haplotype (GRIT: Genome Reference Informatics Team, 2022).

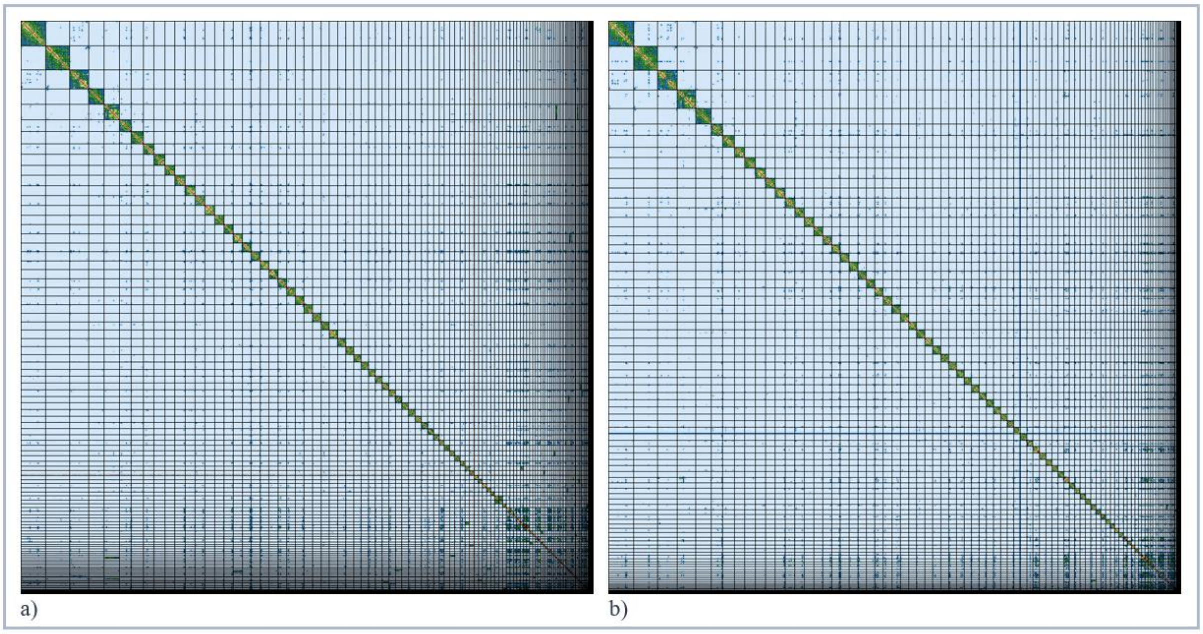


Figure 3: Image of the scaffold contact maps based on Hi-C data of RL1 before (a) and after (b) manual curation in the PretextView desktop program. There is a large number of small, unplaced scaffolds in the lower right corner of the before image (a), as well as some darker blue and yellow contact points in the lower- and right perimeter of the contact grid (a). The number of small, unplaced scaffolds are highly reduced in the after image (b). Images were generated using the PretextSnapshot-command from the pretext-suite.

2.5.3 Filtering of non-target DNA with BlobToolKit and BlobToolViewer

The BlobToolKit-suite was used to filter non-target DNA from the scaffolds (Challis et al., 2020). For each haplotype, BlobToolKit Specification was used to create a directory in a JSON-format, where metadata and information about the scaffolds could be accessed without loading the full dataset. Minimap and blastn from the BlobToolKit pipeline was used (Challis et al., 2020), in combination with MMseqs2 (Mirdita et al., 2021), to generate the datasets required for identifying non-target DNA in the BlobToolKit Viewer (Challis et al., 2020). MMseq2 was used instead of BLAST, is because it ran 10 000 times faster (Mirdita et al., 2021), and enabled blasting the DNA against several different nucleotide databases, such as UniProt's protein knowledgebase, NCBI's nucleotide and protein databases, and UniProt's reference proteomes database (Mirdita et al., 2021). Furthermore, BUSCO's for the actinopterygii-, metazoa-, vertebrata-, eukaryota-, bacteria-, proteobacteria-, archaea-, mammalia-, aves-, fungi-, insecta-, arthropoda- and viridiplantae- lineages were included as part of the BlobToolKit analyses, to determine the degree of BUSCO-completeness for each individual haplotype dataset and scaffold (Manni et al., 2021).

Since mean GC-content will vary between taxa, this can be used as a filtration parameter when searching for contaminant DNA (Challis et al., 2020). Another filtration parameter which should have similar values across most of the genome is the read coverage, therefore, these two were used in conjunction to determine which scaffolds to remove (Challis et al., 2020).

River lamprey: The filtration settings for RL1 were 1; coverage between 0 and 4.07 and GC-content between 0.608 and 0.7498, and 2; coverage between 51.1 and 1005.74 and GC-content between 0.609 and 0.7498. These values were determined by examining the datasets in blob- and kite-mode, and marking the outliers. These outliers were viewed in Table view, and were all marked as “No-hit”, meaning they lacked a taxonomic annotation (Challis et al., 2020).

The same process was repeated for RL2, and here the filtration settings were 1; coverage between 0 and 9.41 and GC-content between 0.605 and 0.748, and 2; coverage between 57 and 1160 and GC-content between 0.605 and 0.748. All outliers were marked as “No-hit” (see Figure 4 a) and b)).

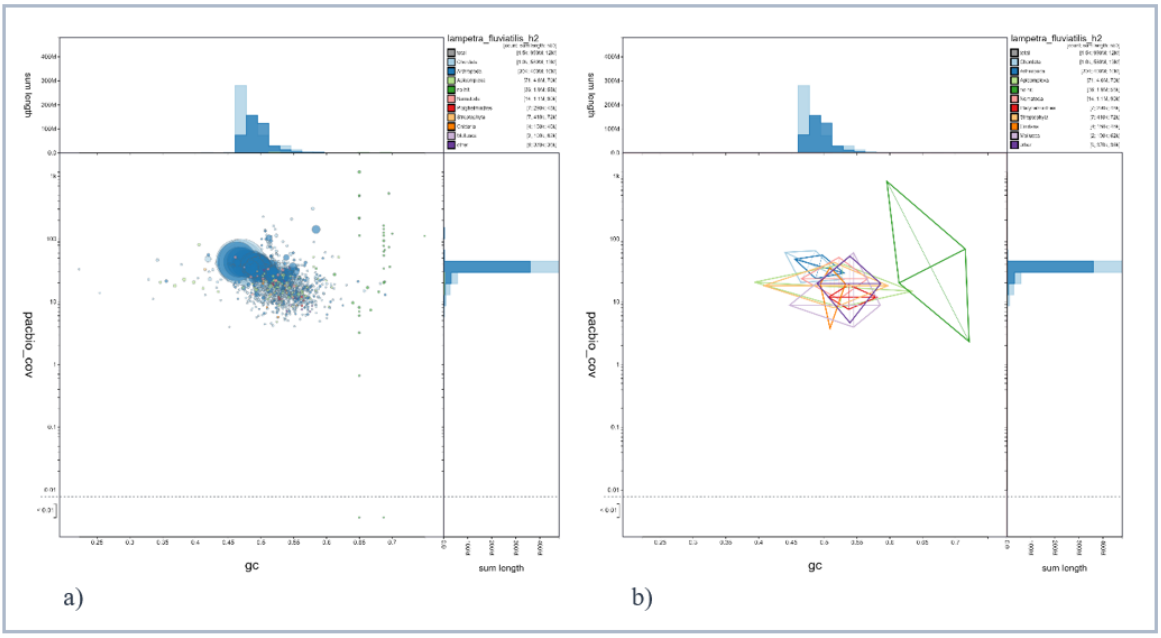


Figure 4: Blob- (a) and kite-view (b) of RL2 from BlobToolKitViewer. Marked in green (a and b) we see that the no-hits form their own distinct clusters, with far higher GC-content than the other overlapping groups, such as the light blue chordata-hits and the dark blue arthopoda-hits.

Brook lamprey: For BL1, all "No-hit"-outliers were filtered away, as they created their own, distinct group in the blob and kite-representations of the dataset (Challis et al., 2020). For BL2, the filtration settings were 1; coverage between 0 and 4 and GC-content between 0.605 and 0.719, and 2; coverage between 38,5 and 741 and GC-content between 0.605 and 0.719. All outliers for haplotype 2 were marked as "No-hit".

Following this filtration process, four TXT files containing a list of the contaminant scaffolds were created. Using the samtools faidx- (SAMtools v1.11) (Danecek et al., 2021) and cut-commands, two lists of all scaffolds in the curated FASTA files were made. Following this step, the grep-command was used to remove the list of the contaminant scaffolds from the list containing all scaffolds, and by combining the original curated FASTA files the new lists without the contaminants using the seqtk subseq-command (seqtk v1.3) (Li, 2018), two new FASTA files were created, without the contaminant DNA. In all, 77 contaminant scaffolds were removed from RL1, 25 contaminant scaffolds were removed from the RL2, 4 contaminant scaffolds were removed from the BL1, and 15 contaminant scaffolds were removed from the BL2.

2.6 Annotation

2.6.1 MetaEuk and InterProScan

To create a fast annotation, a combination of MetaEuk (Levy Karin et al., 2020) and InterProScan (Jones et al., 2014) was used for large-scale gene discovery. Vertebrate proteins from OrthoDB v10 (Kriventseva et al., 2019), and all proteins from UniProtKB/Swiss-Prot was aligned against the genome with MetaEuk and predicted genes were output as FASTA and GFF3 files (Bateman et al., 2021). Since some of the transposable elements could be registered as proteins within these databases, some transposable elements may have been annotated as genes. However, given the time constraints, this method was deemed as a fairly accurate way to create a genome annotation to be used in further analyses.

MetaEuk was installed using Anaconda v4.12.0 (Anaconda Software Distribution, 2020), and ran using the easy-predict setting, which predicted proteins from contigs based on target similarities (Levy Karin et al., 2020).

To prepare for annotation with InterProScan (Jones et al., 2014), the agat_convert_sp_gxf2gxf.pl-script from NBIS (NBIS, 2019) was used to create a protein FASTA-file from the MetaEuk GFF file. After this, the interproscan.sh-script was ran with the optional GOterms-setting, which provided mapping to Gene Ontology, which were based on the manually curated homologous superfamily-, family-, domain-, repeat- or important site- InterPro entries provided in the InterPro member databases (Jones et al., 2014). Parallel to this process, the diamond blastp-command (Buchfink et al., 2015) was used to query the amino acid sequence created using MetaEuk against the UniProtKB/Swiss-Prot database (Bateman et al., 2021). As the final step, the agat_sp_manage_functional_annotation.pl and agat_sp_extract_sequences.pl scripts from NBIS (NBIS, 2019) were used to create the final GFF containing the complete annotation, and the sequence FASTA files containing the proteins and mRNA.

2.7 Genomic comparisons

In this section genomic comparisons were made between RL1, RL2, BL1, and BL2. In some instances, the sea lamprey (*Petromyzon marinus*) somatic assembly (PMS) from the Vertebrate Genomes Project (Smith, Kuraku et al. 2013), the sea lamprey germline assembly (PMG) (Smith, Timoshevskaya et al. 2018) from NCBI, and the Arctic lamprey (*Lethenteron camtschaticum*) mitochondrial reference genome from NCBI (Lee, 2013) were used as comparative outgroups. This is specified within the text.

2.7.1 Assemblathon stats and BUSCO

Before and after the manual curation and removal of contaminant scaffolds, the curated FASTA files were compared using the aforementioned genome statistics tools Assemblathon_stats (Bradnam et al., 2013) and BUSCO v5.0.0 (Manni et al., 2021). This was to verify that no large scaffolds had been split during the manual curation, and that the number of fragmented and missing BUSCOs had not gone up.

2.7.2 D-GENIES dotplot

To get a syntenic similarity comparison between the same-species-haplotypes, inter-species-haplotypes, and the sea lamprey outgroup, i.e. the somatic and germline assemblies, D-GENIES v1.3.0 dot plots were used to gain a quick overview of duplications, breaks and inversions within, and between, the genomes (Cabanettes & Klopp, 2018). Since the assemblies exceeded the plot alignments size limit of 1024 Mb, the FASTA files were converted to PAF- and IDX-files using minimap2 (Li 2021) and the samtools faidx- (Danecek et al., 2021) and cut-commands. From these assembly combinations, 22 dot plots were created.

2.7.3 Assemblytics

To run Assemblytics (Nattestad & Schatz, 2016), DELTA files were created for 20 different alignments using the four different haplotype resolved assemblies, and the sea lamprey somatic and germline assemblies as either the reference or assembly alignment file. These 20 DELTA files were created using the nucmer-script from the MUMmer v4.0.0 utility suite (Marçais et al., 2018), and uploaded to the Assemblytics website for plotting and assembly statistics, such as type, number and size of structural variants (Nattestad & Schatz, 2016).

To get a summary of the alignment data generated using the nucmer-script, the wrapper script DNAdiff was used on the DELTA-files (Marçais et al., 2018). To calculate whether the interspecies difference in structural variants were significant, a χ-square test was ran using RStudio v1.4.1106 (RStudio Team, 2021).

2.7.4 OrthoFinder

To determine the phylogenetic relationship between the four haplotypes and the sea lamprey outgroup, OrthoFinder v2.5.4 (Emms & Kelly, 2019) was used to create a rooted species tree. This species tree was generated using the STAG-algorithm, which stands for "Species Tree from All Genes" (Emms & Kelly, 2018). The program's dependencies were downloaded using Anaconda v4.12.0 (Anaconda Software Distribution, 2020).

2.7.5 MitoHiFi and mitochondrial comparisons

To be able to create mitochondrial phylogenies and alignments, the brook lamprey's mitochondrial genome was assembled using MitoHiFi v2.2 (Allio et al., 2020). The program's dependencies were downloaded using Docker v4.7.0 (Merkel, 2014) and ran through a Singularity image (Kurtzer et al., 2017). The assembly was made from the assembled contigs created using the hifiasm Hi-C-integrated assembler (Cheng et al., 2021), with the Vertebrate Mitochondrial Code from NCBI (Elzanowski, 2019), and was referenced against the river lamprey mitochondrial FASTA and genbank-files from NCBI (Gachelin, 2000).

Several attempts were made to assemble the river lamprey's mitochondrial genome but were all unsuccessful. Therefore, since the European river, Arctic, and sea lamprey mitochondrial reference genomes were already publicly available on NCBI, these FASTA files were used in conjunction with the brook lamprey mitochondrial assembly to create genome alignments with MAFFT v7.490 (Rozewicki et al., 2019) and phylogenies in IQ-TREE v2.2.0 (see Appendix B, Figure 34B) (Nguyen et al., 2015). The mitochondrial genome alignments were also compared using PhyKIT (Steenwyk et al., 2021) to calculate the assemblies average pairwise identities.

2.7.5 BLAST searches for specific genes

To search for, and align, specific genes, the FASTA sequence for the vasotocin gene in the sea lamprey was downloaded from NCBI (Mayasich, 2016), and blasted against the RL1, RL2, BL1 and BL2-assemblies, using the blastn command. Following the BLAST-search, the gene-IDs were retrieved from the annotation GFF files and used to look up the nucleotide sequences in the mRNA FASTAs generated during annotation with InterProScan. The mRNA sequences were then aligned using MAFFT v7.490 (Rozewicki et al., 2019), and visualised and inspected in AliView v1.28 (Larsson, 2014).

3. Results

3.1 Sequencing results

River lamprey: Following sequencing on the Sequel II instrument, the number of polymerase reads generated for all three SMRT cells were 16 614 300. The average polymerase read length was 34-57 kb, and the total number of polymerase bases were 708.7 Gb. The number of HiFi reads generated from the Circular Consensus Sequences pipeline were 2 470 187, with a HiFi Yield of 38 546 515 411 bp. The mean HiFi read length of 14-17 kb. The HiFi-read quality was Q29-Q30, and the number of HiFi reads below Q20 was 803 050.

During Hi-C sequencing 333 956 869 total paired Illumina reads were generated, which added up to 100,2 Gb of data, with an 83.5X coverage.

Brook lamprey: Following sequencing on the Sequel II instrument, the number of polymerase reads generated for all three SMRT cells were 15 223 936. The average polymerase read length was 49-50 kb, and the total number of polymerase bases were 760 Gb. The number of HiFi reads generated from the Circular Consensus Sequences pipeline were 3 255 356, with a HiFi Yield of 43 995 045 444 bp. The mean HiFi read length of 13.5 kb. The HiFi-read quality was Q28-Q29, and the number of HiFi reads below Q20 was 1 248 337.

During Hi-C sequencing 406 990 761 total paired Illumina reads were generated, which added up to 122 Gbp of data with a 110X coverage.

3.2 Assembly results

3.2.1 QUAST-results

After doing initial comparisons of the Flye, IPA, HiCanu and hifiasm-assemblies, the hifiasm assemblies for both species were chosen for Hi-C integration. This was because the hifiasm assemblies had the lowest number of contigs, and the longest N50 length (Table 1). Graphical representations of these statistics, as well as plots for GC-content are available in Appendix B, Figures 1B-8B.

Table 1. Summary of assembly statistics from QUAST, showing the assembly lengths, number of contigs and N50 contig lengths following assembly of the HiFi reads using Flye, IPA, HiCanu with minimum overlaps of 200, 500 and 700, and hifiasm. The Flye results for the river lamprey are not included in the table.

	River lamprey			**Brook lamprey**		
	Assembly length (bp)	**Number of contigs**	**N50 length (bp)**	**Assembly length (bp)**	**Number of contigs**	**N50 length (bp)**
Flye	NA	NA	NA	1 443 886 436	12 205	191 049
IPA	1 046 254 573	3 539	1 816 707	1 049 760 163	4 886	1 583 620
HiCanu min overlap 200	1 986 698 009	11 803	667 952	1 817 974 844	15 304	566 689
HiCanu min overlap 500	1 983 817 156	11 750	646 484	1 815 505 054	15 290	562 493
HiCanu min overlap 700	1 981 918 288	11 756	636 886	1 811 447 818	15 177	557 607
hifiasm	1 101 718 900	2 928	3 647 336	1 110 143 197	4 213	3 899 937

3.2.2 Hifiasm assembly

Following hifiasm assembly, the river lamprey assembly consisted of 2928 contigs, while the brook lamprey assembly consisted of 4009 contigs (Table 2). Regardless of their differing number of assembled contigs, the total size of all contigs, in nucleotides, was around the same for both species, with 1 101 718 900 and 1 101 576 128 nucleotides for the river lamprey and brook lamprey, respectively (Table 2). Both assemblies had 232 complete BUSCOs of a total of 255 BUSCO groups searched, and only 9 (river lamprey) and 11 (brook lamprey) missing BUSCOs (Table 3).

Table 2. Summary of assembly statistics from running Assemblathon_stats on the hifiasm-assemblies for the river lamprey and brook lamprey.

	Number of contigs	**Total contig size (nt)**	**Longest contig (nt)**	**Mean contig size (nt)**	**Median contig size (nt)**	**N50 contig length (nt)**
River lamprey	2928	1 101 718 900	22 593 410	376 270	54 316	3 647 336
Brook lamprey	4009	1 101 576 128	26 278 604	274 776	37 558	3 904 746

Table 3. Summary of BUSCO-scores for the river lamprey and brook lamprey after running BUSCO in mode genome, with the gene predictor MetaEuk against the lineage dataset eukaryota_odb10.

	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total BUSCO groups searched
River lamprey	232	224	8	14	9	255
Brook lamprey	232	215	17	12	11	255

3.2.3 HiFiasm Hi-C-integrated assembly

After Hi-C integration, the RL2 and BL2-assemblies had the lowest number of contigs, with 2421 and 2972, respectively (Table 4). This was reflected in their mean scaffold sizes, which were the largest out of all six assemblies (Table 4). However, when assessing the assemblies BUSCO-scores, the assemblies with the highest number of complete BUSCOs were RLP and BLP (Table 5). These assemblies were also the ones with the lowest number of missing BUSCOs (Table 5).

Table 4. Summary of assembly statistics from running Assemblathon_stats on the hifiasm Hi-C-integrated assemblies both haplotypes, and the primary contig for the river and brook lamprey.

	Number of contigs	Total contig size (nt)	Longest contig (nt)	Mean contig size (nt)	Median contig size (nt)	N50 contig length (nt)
RL1	3069	1 008 746 736	13 267 019	328 689	55 738	2 244 160
RL2	2421	990 190 384	13 784 290	409 001	82 859	2 135 772
RLP	3115	1 091 797 163	24 907 613	350 497	54 523	3 238 440
BL1	3150	936 941 293	21 679 006	297 442	37 690	2 636 808
BL2	2972	1 044 089 067	18 336 540	351 309	51 952	3 099 685
BLP	4071	1 103 272 259	26 262 476	271 008	37 442	3 875 619

Table 5. Summary of BUSCO-scores for the river lamprey and brook lamprey after running BUSCO in mode genome, with the gene predictor MetaEuk against the lineage dataset metazoa_odb10.

	Complete BUSCOs	**Complete and single-copy BUSCOs**	**Complete and duplicated BUSCOs**	**Fragmented BUSCOs**	**Missing BUSCOs**	**Total BUSCO groups searched**
RL1	816	777	39	43	95	954
RL2	789	763	26	48	117	954
RLP	859	829	30	50	45	954
BL1	775	721	54	37	142	954
BL2	848	819	29	46	60	954
BLP	861	794	67	43	50	954

3.3 Scaffolding results

Of the scaffolded assemblies, RL2 and BL2 had both the lowest number of scaffolds, and the highest mean and median scaffold sizes for each species (Table 6). While BL2 had the highest number of complete BUSCOs, and the lowest number of missing BUSCOs overall, RL2 had more missing, and fewer complete BUSCOs than its haplotype counterpart, RL1 (Table 7).

Table 6. Summary of assembly statistics from running Assemblathon_stats on the scaffolded haplotype assemblies for the river lamprey and brook lamprey.

	Number of scaffolds	**Total scaffold size (nt)**	**Longest scaffold (nt)**	**Mean scaffold size (nt)**	**Median scaffold size (nt)**	**N50 scaffold length (nt)**
RL1	2144	1 008 931 736	37 972 987	470 584	42 626	12 138 619
RL2	1476	990 379 384	38 427 461	670 989	57 541	11 781 077
BL1	2216	937 128 093	39 697 215	422 892	32 141	12 263 195
BL2	1988	1 044 285 867	40 814 456	525 295	39 984	12 651 338

Table 7. Summary of BUSCO-scores for the river lamprey and brook lamprey after running BUSCO in mode genome, with the gene predictor MetaEuk against the lineage dataset metazoa_odb10.

	Complete BUSCOs	**Complete and single-copy BUSCOs**	**Complete and duplicated BUSCOs**	**Fragmented BUSCOs**	**Missing BUSCOs**	**Total BUSCO groups searched**
RL1	817	779	38	42	95	954
RL2	789	765	24	49	116	954
BL1	777	723	54	36	141	954
BL2	850	821	29	43	61	954

3.4 Manual curation results

Following manual curation and contaminant DNA-removal, RL2 and BL2 still had the lowest number of scaffolds, and longest mean and median scaffold lengths (Table 8). However, RL1 had a higher N50 scaffold length (Table 8), and a higher number of complete BUSCOs compared to RL2 (Table 9). After painting, i.e. giving the scaffolds names and numbers corresponding to the river lamprey and brook lamprey's karyotype of 82 chromosomes (Ishijima et al., 2016), the resulting number of super scaffolds of chromosome size was 82 for all assemblies.

Table 8. Summary of assembly statistics from running Assemblathon_stats on the manually curated and contaminant filtered river- and brook lamprey haplotype assemblies.

	Number of super scaffolds	**Number of scaffolds**	**Total scaffold size (nt)**	**Longest scaffold (nt)**	**Mean scaffold size (nt)**	**Median scaffold size (nt)**	**N50 scaffold length (nt)**
RL1	82	2 073	1 008 945 936	40 070 597	486 708	41 871	12 603 230
RL2	82	1 386	990 397 384	41 122 491	714 572	54 835	12 457 738
BL1	82	2 108	937 149 693	40 490 199	444 568	31 520	12 211 417
BL2	82	1 867	1 044 310 067	41 440 432	559 352	38 059	12 800 873

Table 9. Summary of BUSCO-scores for the river lamprey and brook lamprey after running BUSCO in mode genome, with the gene predictor MetaEuk against the lineage dataset metazoa_odb10.

	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total BUSCO groups searched
RL1	817	779	38	42	95	954
RL2	789	765	24	49	116	954
BL1	777	724	53	36	141	954
BL2	850	821	29	44	60	954

3.5 Annotation results

After annotation, all four assemblies had a total number of genes in the region between 48 916 and 42 293, with BL1 having the lowest number of annotated genes overall, and BL2 having the highest (Table 10). Between the river lamprey haplotype assemblies, RL1 has the highest number of annotated genes, with 47 410 genes annotated (Table 10).

Table 10. Number of genes in found in the RL1, RL2, BL1 and BL2-assemblies following annotation with MetaEuk and InterProScan.

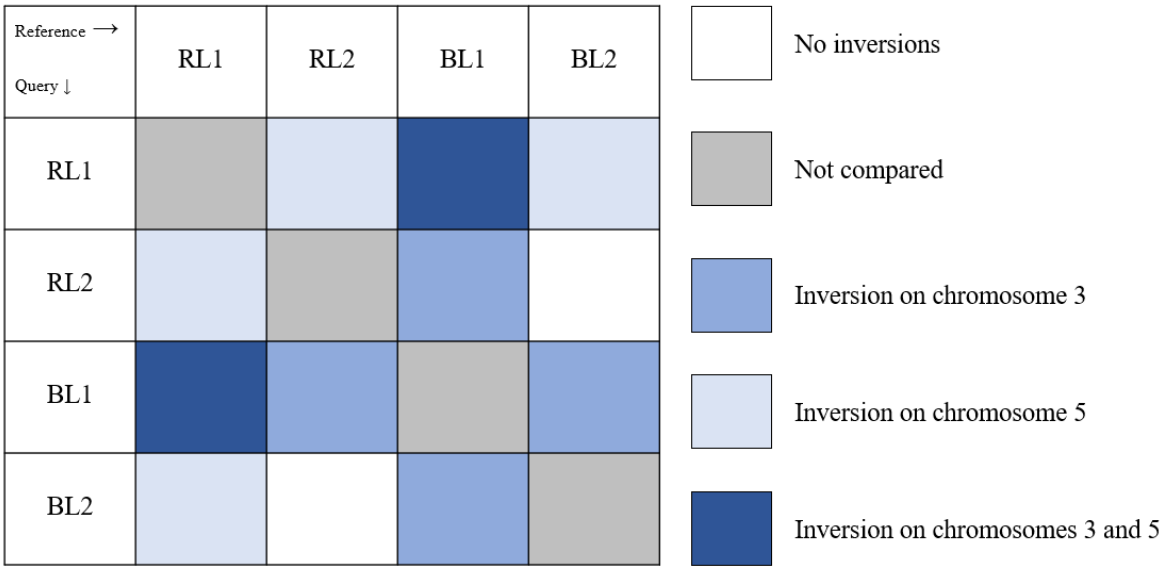
Assembly	Number of genes
RL1	47 410
RL2	45 490
BL1	42 293
BL2	48 916

3.6 Results of comparative analyses

3.6.1 D-GENIES Dotplots

When examining the dotplots created using D-GENIES (Cabanettes & Klopp, 2018), there was a great degree of synteny between each haplotype within each species. As visualised in the matrix in Figure 5, I found two inversions: one on chromosome 3, and one on chromosome 5. When a genome sequence is inverted, it means that while the chromosome sequence exists in both assemblies, part of the chromosome is not in the same orientation. Between RL2 and BL1 (Appendix B, Figure 13B and 17B), and between BL1 and BL2 (Figure 6 b) and 7 b)), there was only an inversion on chromosome 3. Between RL1 and RL2 (Figure 6 a) and 7 a)), there was only an inversion on chromosome 5. Between RL1 and BL2 (Figure 8, and Appendix B, Figure 10B and 16B) there were two inversions, on chromosomes

3 and 5, and between RL2 and BL2, there were no inversions at all (Figure 9). Between RL1 and BL2, there was an inversion on chromosome 5 (see Appendix B, Figure 20B), however, in Appendix B, Figure 11B, when BL2 was the reference and RL1 was the query, chromosome 5 was displaced (likely due to a conversion error in the PAF file), resulting in the inversion being observable in the top right corner. To compare the results to an outgroup, all four haplotypes were plotted against the sea lamprey somatic- and germline assemblies, and in the dotplots in Figure 10 a) and b), and in Appendix B, Figure 12B, 14B, 15B, 19B, and 21B-24B, there were far more gaps (i.e. sequences that only existed in one of the assemblies) and inversions along the entire length of the genomes.



Reference → Query ↓	RL1	RL2	BL1	BL2
RL1	▨	░	■	░
RL2	░	▨	▤	□
BL1	■	▤	▨	▤
BL2	░	□	▤	▨

□	No inversions
▨	Not compared
▤	Inversion on chromosome 3
░	Inversion on chromosome 5
■	Inversion on chromosomes 3 and 5

Figure 5: Comparative matrix showing which haplotype assembly combinations has which inversions. Between RL2 and BL1, and between BL1 and BL2, there is only an inversion on chromosome 3. Between RL1 and RL2, and between RL1 and BL2, there is only an inversion on chromosome 5. Between RL1 and BL2 there are two inversions, on chromosomes 3 and 5, and between RL2 and BL2, there are no inversions at all.

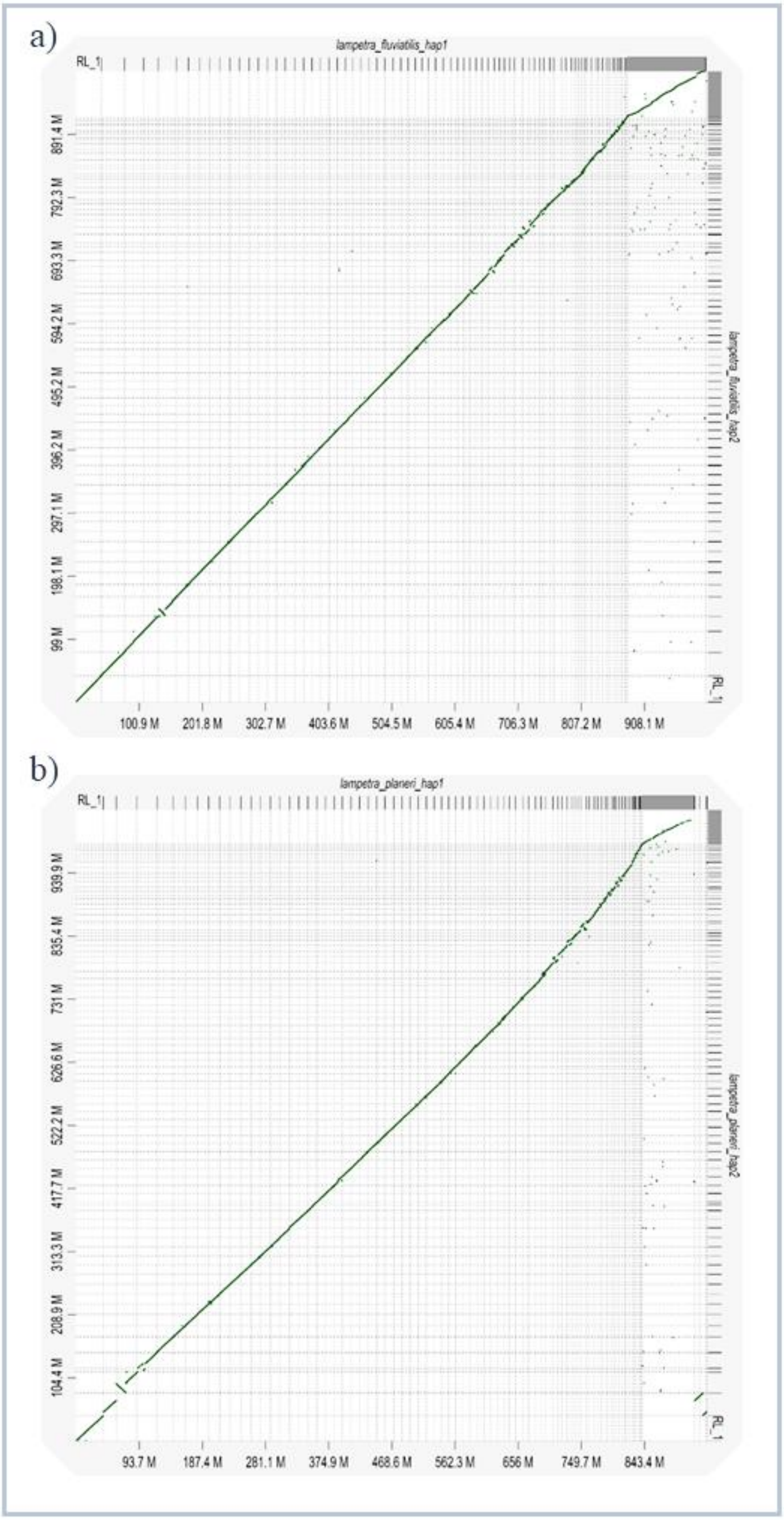


Figure 6: Dotplot of RL1 (x-axis) and RL2 (y-axis) (a) and BL1 (x-axis) and BL2 (y-axis) (b) generated using D-GENIES dotplot. In 11 a) there was a large inversion on chromosome 5, and in 11 b) there was a large inversion on chromosome 3.

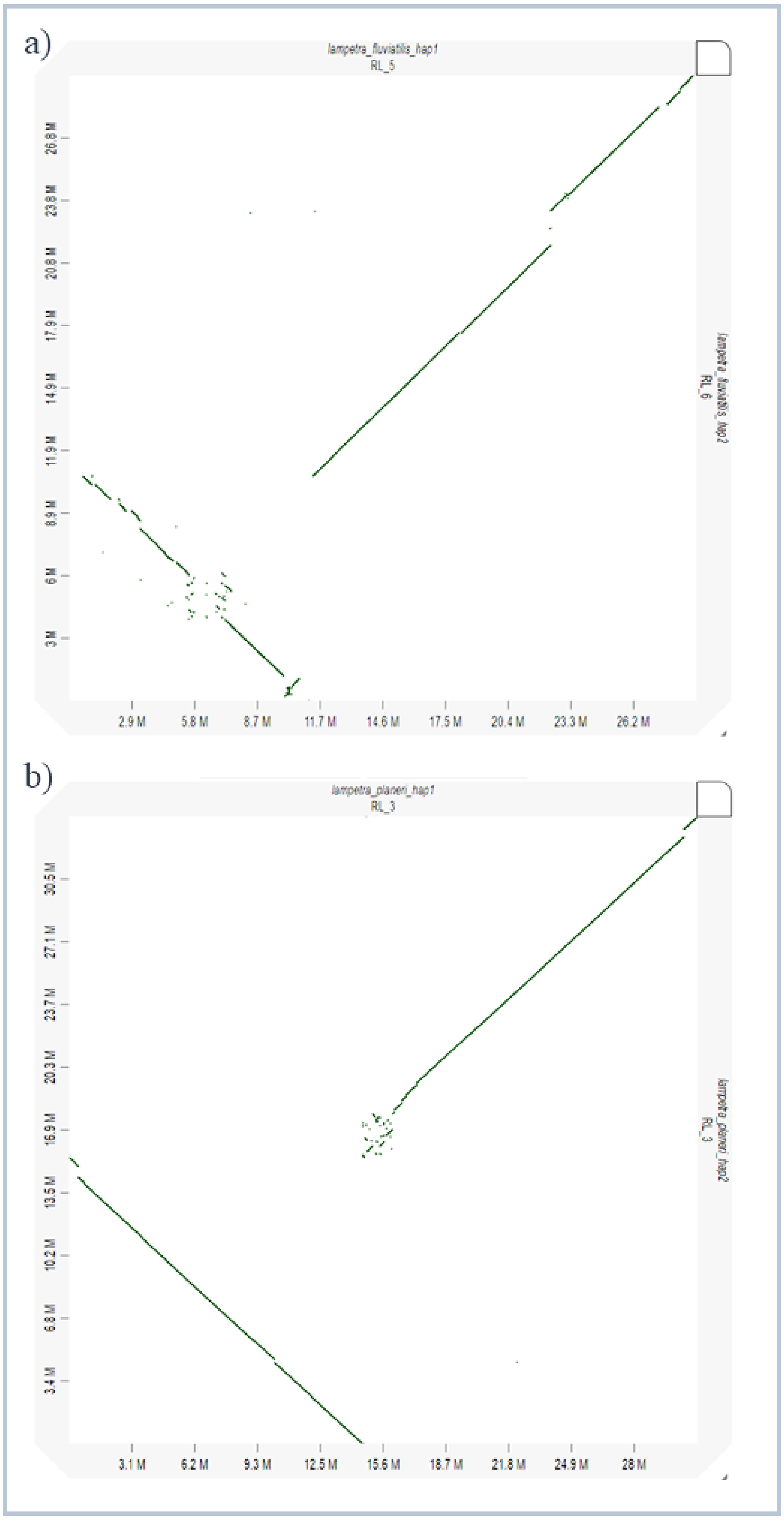


Figure 7: Zoomed in image of the inversion on chromosome 5 between RL1 (x-axis) and RL2 (y-axis) (a) and the inversion in chromosome 3 between BL1 (x-axis) and BL2 (y-axis) (b).

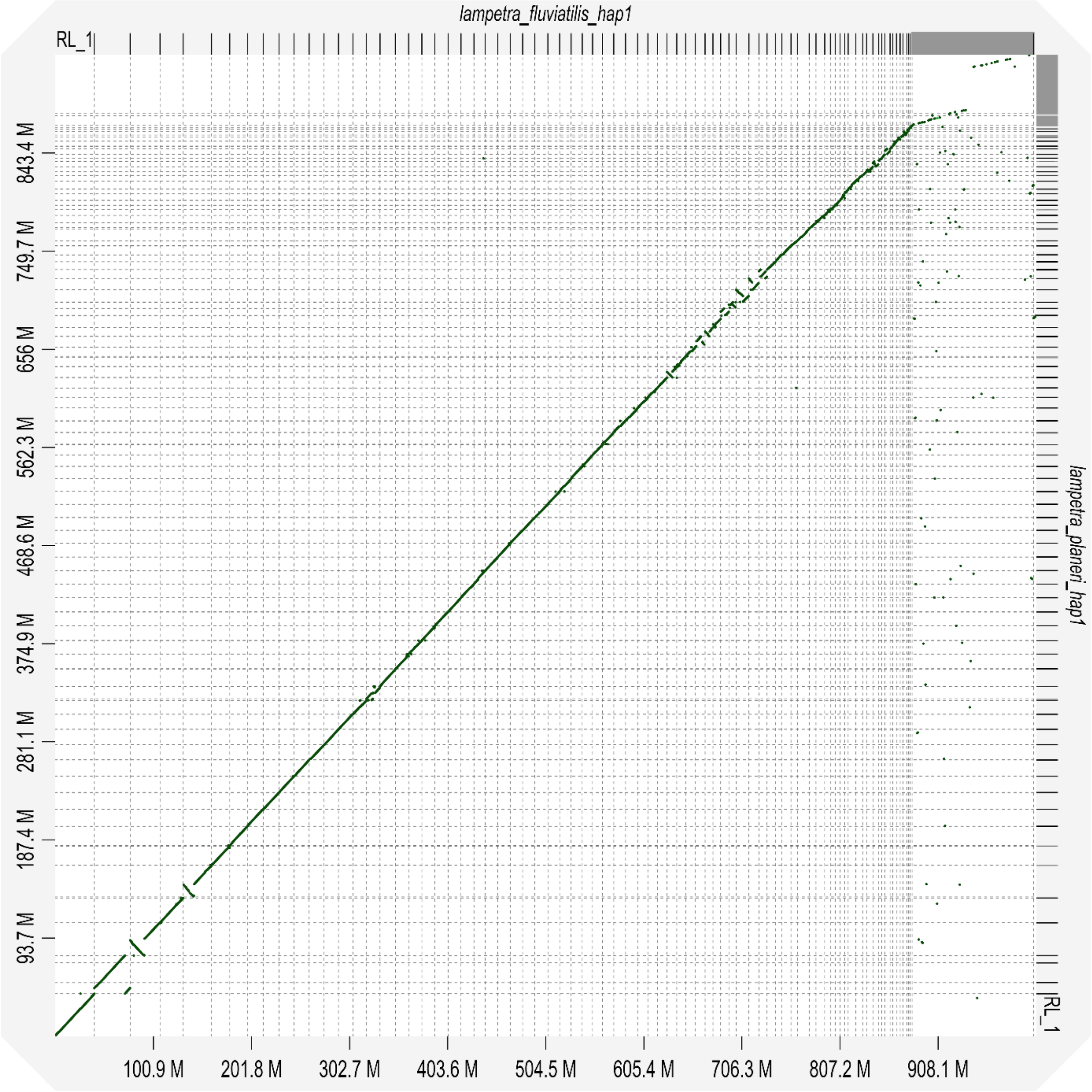


Figure 8: Dotplot generated using D-GENIES dotplot, showing the syntenic relationship between RL1 and BL1. Between these two haplotype assemblies, both inversions on chromosomes 3 and 5 are visible.

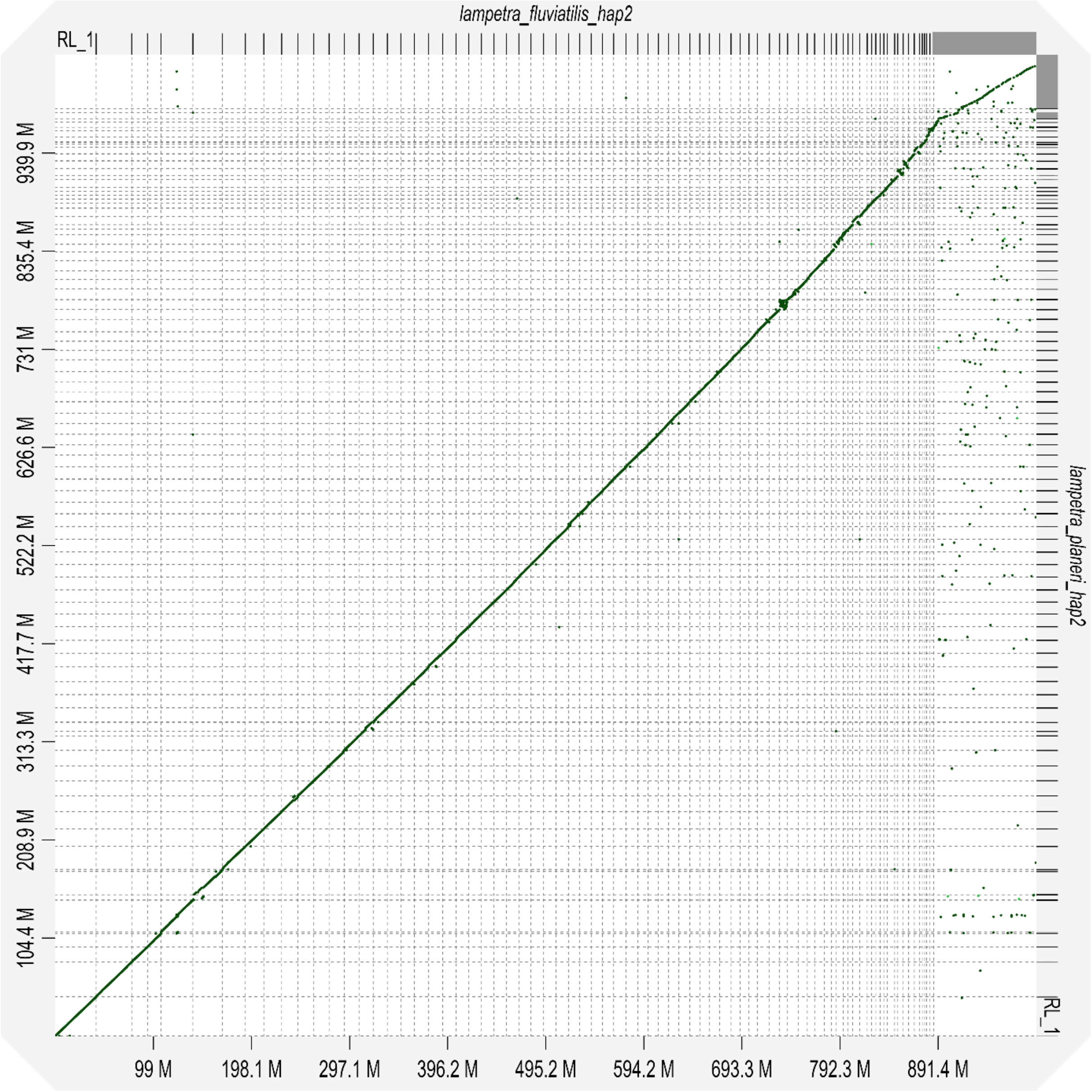


Figure 9: Dotplot generated using D-GENIES dotplot, showing the syntenic relationship between RL2 and BL2. Between these two haplotype assemblies, no inversions on chromosome 3 or 5 occurred.

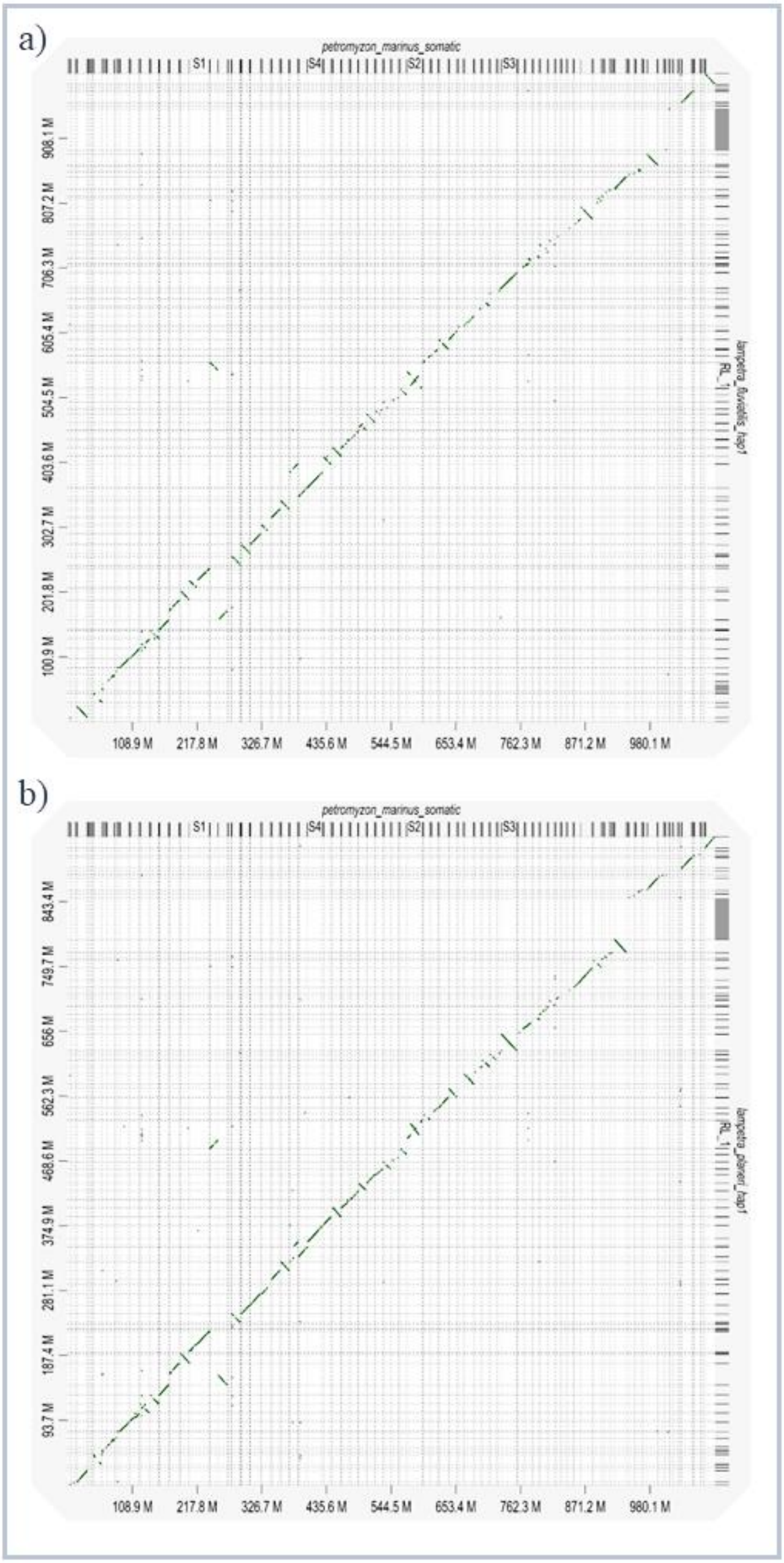


Figure 10: Dotplot showing the degree of synteny between the sea lamprey (PMS, x-axis) and RL1 (y-axis) (a) and the sea lamprey (PMS, x-axis) and BL1 (y-axis) (b). In both dotplots there are several inversions and gaps scattered throughout the lengths of the assemblies.

3.6.2 Assemblytics

The structural variant detector nucmer from the MUMmer-program suite (Marçais et al., 2018) was used on all assemblies to detect insertions, deletions, repeat expansions, repeat contractions, tandem expansions and tandem contractions in each haplotype hifiasm-Hi-C-integrated assembly (Cheng et al., 2022) relative to the curated FASTAs created using the Rapid curation suite (GRIT: Genome Reference Informatics Team, 2022).

The number of structural variants was larger between species than between the haplotypes within the same species (Table 11), and the number of structural differences correlated with the total number of bases affected by these structural differences (Table 12). The most common structural difference between all haplotypes were insertions, followed by repeat contractions and deletions. This was visualised in the plots generated with Assemblytics (not included), where all the plots had significant peaks of insertions and deletions of around 240 bp and 7000 bp and repeat expansions and contractions of around 500-1500 bp (see Figure 11). To test whether the number of structural variant differences between species were significantly higher than the within-species haplotypes, a χ-square test was conducted, using RStudio v1.4.1106 (RStudio Team, 2021). From this I found a p-value of 2.16 x 10^{-16}, a χ^2-statistical value of 626.14 with 1 degree of freedom. This confirmed that the number of structural variants were significantly higher on an interspecies level, compared to the observed intraspecies differences.

Table 11. The total number of structural variants between the RL1, RL2, BL1 and BL2-assemblies.

Reference → **Query ↓**	**RL1**	**RL2**	**BL1**	**BL2**
RL1		27 553	31 555	36 378
RL2	27 523		31 566	36 256
BL1	30 036	30 239		22 034
BL2	34 967	35 276	22 022	

Table 12. The total number of bases affected by structural variants between each haplotype.

Reference → **Query ↓**	**RL1**	**RL2**	**BL1**	**BL2**
RL1		33.12 Mbp	38.04 Mbp	43.52 Mbp
RL2	32.65 Mbp		37.88 Mbp	43.96 Mbp
BL1	36.3 Mbp	36.31 Mbp		26.33 Mbp
BL2	42.21 Mbp	42.82 Mbp	26.73 Mbp	

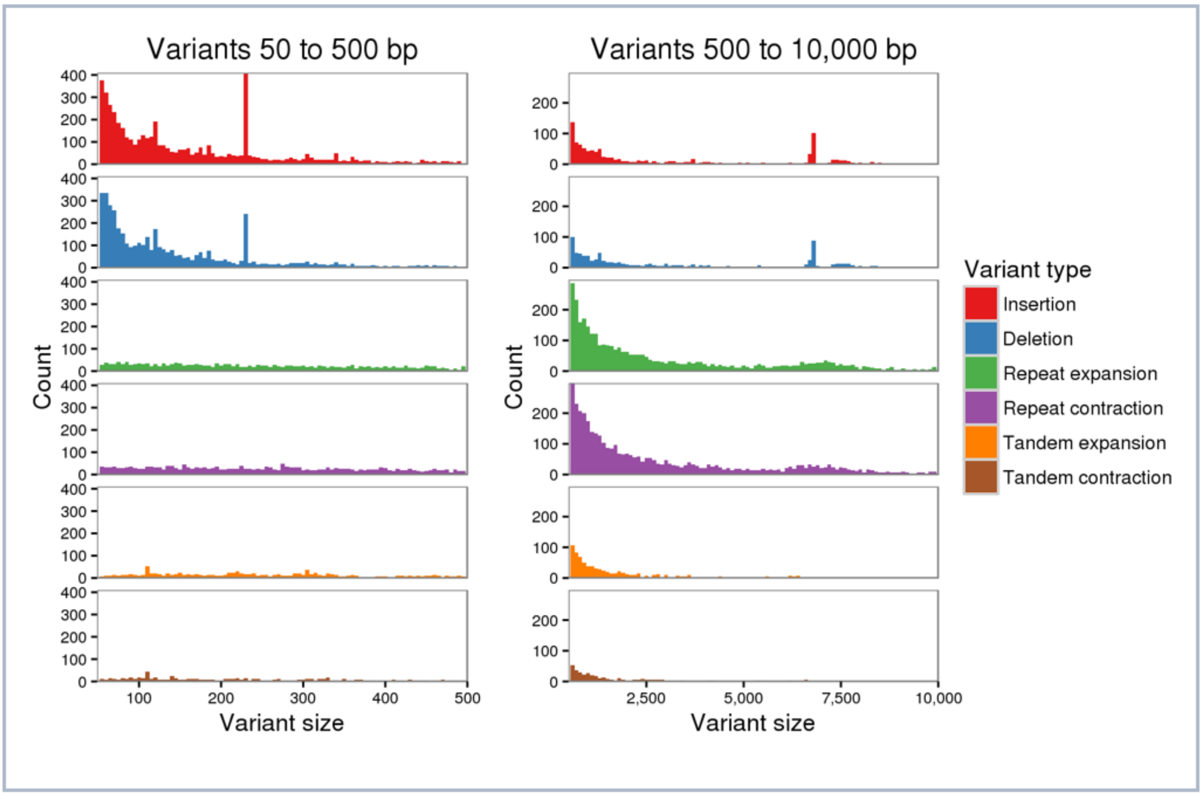


Figure 11: Graphs generated using Assemblytics, showing the size distribution of variant types between RL1 and RL2. Here, RL1 is the reference sequence and RL2 is the query sequence.

The percentage of aligned bases were consistently higher between intra-species haplotypes than inter-species haplotypes, with one notable exception, namely when the brook lamprey haplotype 2 contigs query FASTA-file is referenced against the brook lamprey haplotype 1 curated reference assembly (Table 13). Moreover, when the brook lamprey's haplotype 2 was used as a reference, all other haplotypes had >91% aligned bases, which was higher than all the other alignment percentages.

Table 13. The percentage of aligned bases between the query and reference FASTA files when using the wrapper script DNAdiff around nucmer from the MUMmer suite.

Reference → **Query ↓**	**RL1**	**RL2**	**BL1**	**BL2**
RL1		87.27%	84.35%	91.31%
RL2	88.02%		84.95%	91.74%
BL1	87.83%	87.73%		92.28%
BL2	86.55%	86.23%	83.89%	

The percentage of alignment blocks comprising the 1-to-1 mapping of the reference assemblies to the query assemblies was consistently above 98.6% between all assemblies, regardless of species (Table 14). Notably, the percentage of alignment blocks were higher between BL1 and the river lamprey haplotypes than between BL2 and the river lamprey haplotypes.

Table 14. The percentage of alignment blocks comprising the 1-to-1 mapping of the reference assemblies to the query assemblies, calculated using the DNAdiff-wrapper script around nucmer from the MUMmer suite.

Reference → **Query ↓**	**RL1**	**RL2**	**BL1**	**BL2**
RL1		98.99%	98.79%	98.62%
RL2	98.99%		98.79%	98.61%
BL1	98.78%	98.78%		99.19%
BL2	98.62%	98.62%	99.19%	

3.6.3 OrthoFinder

In the rooted species tree (Figure 12) the sea lamprey outgroup formed its own branch, with a bootstrap value of 1, or 100%. This means that for every phylogeny created using the gene trees from OrthoFinder, this branch was observed 100% of the time. In the next node in the species tree, the BL2 branched off and created a sister taxon to the node containing BL1 on one branch, and both RL1 and RL2 on the other. The clustering of the final branches had bootstrap supports of 46.6% and 59.1% respectively. Although these bootstrap values were not statistically significant, the clustering of BL1 with RL1 and RL2 still indicated that BL1 was as closely related to BL2 as both river lamprey haplotypes were.

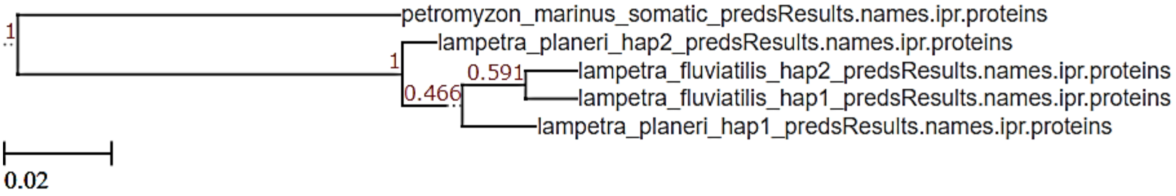

petromyzon_marinus_somatic_predsResults.names.ipr.proteins
lampetra_planeri_hap2_predsResults.names.ipr.proteins
lampetra_fluviatilis_hap2_predsResults.names.ipr.proteins
lampetra_fluviatilis_hap1_predsResults.names.ipr.proteins
lampetra_planeri_hap1_predsResults.names.ipr.proteins
1
1
0.591
0.466
0.02

Figure 12: Species tree generated after running OrthoFinder v2.5.4. Here, the STAG-algorithm was used to infer the phylogenetic relationship between the sea lamprey, RL1, RL2, BL1 and BL2 from 22 133 gene trees. The numbers on the branches represent the branch bootstrap support.

3.6.4 Mitochondrial comparisons

After running PhyKIT (Steenwyk et al., 2021) on the aligned mitochondrial FASTAs a >99.3% assembly similarity between the brook lamprey and river lamprey mitochondrial assemblies was detected (Table 15). Furthermore, the river lamprey and brook lamprey mitochondria were found to be >90% similar to the Arctic lamprey mitochondria (with 90.28% and 90.63% similarity respectively), and >87% similar to the sea lamprey mitochondria (87.37% for the river lamprey, and 87.51% for the brook lamprey) (Table 15). The pairwise identity values between the brook lamprey and river lamprey could be even higher, considering that there was a gap in the brook lamprey D-loop, which was not removed. Although the river lamprey mitochondrial assembly was fetched from NCBI and was not assembled from the river lamprey sampled from Sweden, every other region than the D-loop aligned perfectly.

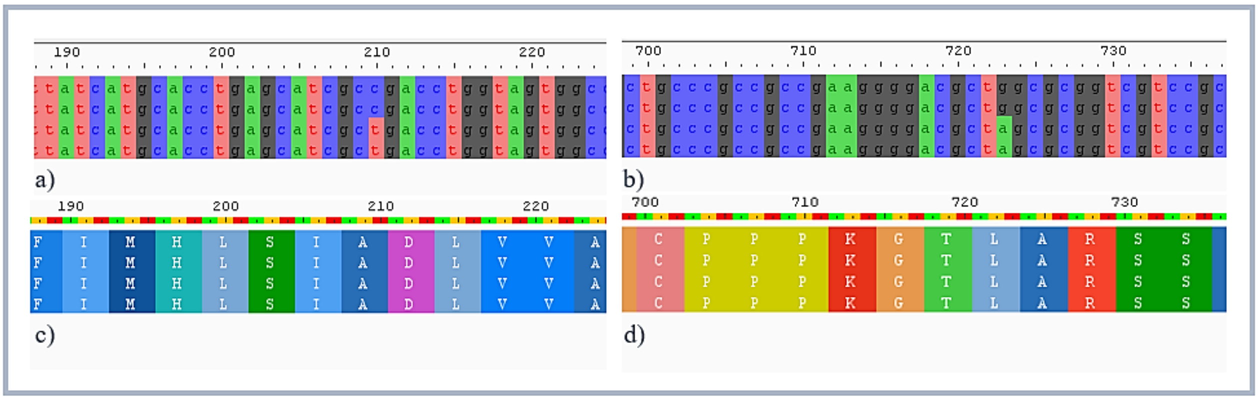
Table 15. Pairwise identities calculated after aligning the MitoHiFi-assembled brook lamprey mitochondria FASTA-file to the NCBI mitochondrial reference FASTAs for the river lamprey, sea lamprey and Arctic lamprey.

Species	**Pairwise identity**
brook lamprey and river lamprey	99.31%
brook lamprey and sea lamprey	87.51%
brook lamprey and Arctic lamprey	90.63%
river lamprey and sea lamprey	87.64%
river lamprey and Arctic lamprey	90.28%
sea lamprey and Arctic lamprey	87.37%

3.6.5 BLAST searches for specific genes

When researching background literature for my thesis, I came across Mateus and colleagues' study from 2013, where they found 12 genes with signals of strong genomic divergence between the river lamprey and brook lamprey (Mateus et al., 2013). The vasotocin gene, which is important for osmoregulation, was one of the genes with the most significant differences (Mateus et al., 2013). Moreover, one of the most distinct life-history differences between the river lamprey and brook lamprey is that the river lamprey can migrate to saltwater. Thus, this gene was of interest when looking into their life-history differences. For all four assemblies, I found sequences matching the vasotocin-BLAST-search query with total lengths of 1042 nucleotides. When aligning the sequences, I found two single nucleotide polymorphisms (SNPs); one at position 210 and another at position 723.

At position one, RL1 and RL2 have the codon GCC, and BL1 and BL2 have the codon GCT (see Figure 13 a). At position two RL1 and RL2 have the codon CTG, and BL1 and BL2 have the codon CTA (see Figure 13 b).



```
t a t c a t g c a c c t g a g c a t c g c c g a c c t g g t a g t g g c
t a t c a t g c a c c t g a g c a t c g c c g a c c t g g t a g t g g c
t a t c a t g c a c c t g a g c a t c g c t g a c c t g g t a g t g g c
t a t c a t g c a c c t g a g c a t c g c t g a c c t g g t a g t g g c
```

```
c t g c c c g c c g c c g a a g g g g a c g c t g g c g c g g t c g t c c g c
c t g c c c g c c g c c g a a g g g g a c g c t g g c g c g g t c g t c c g c
c t g c c c g c c g c c g a a g g g g a c g c t a g c g c g g t c g t c c g c
c t g c c c g c c g c c g a a g g g g a c g c t a g c g c g g t c g t c c g c
```

```
F  I  M  H  L  S  I  A  D  L  V  V  A
F  I  M  H  L  S  I  A  D  L  V  V  A
F  I  M  H  L  S  I  A  D  L  V  V  A
F  I  M  H  L  S  I  A  D  L  V  V  A
```

```
C  P  P  P  K  G  T  L  A  R  S  S
C  P  P  P  K  G  T  L  A  R  S  S
C  P  P  P  K  G  T  L  A  R  S  S
C  P  P  P  K  G  T  L  A  R  S  S
```

Figure 13: Snippets of the alignments of the vasotocin genes from RL1, RL2, BL1 and BL2, aligned in descending order. The nucleotide view from the alignment visualisation tool Aliview, shows the position of the first SNP at position 210 in the alignment (a), and the second SNP at position 723 (b). In the amino acid translation view, the order of the amino acid chain in the vasotocin gene is unchanged (c and d), regardless of the SNPs.

When translated into amino acids, we GCC and GCT both code for alanine (see Figure 13 c), and CTG and CTA both code for leucine (see Figure 13 d), thus the chain of amino acids is unchanged by the SNPs, rendering all four assemblies functionally identical when coding for the vasotocin gene.

4. Discussion

4.1 Main findings and introduction of discussion topics

For my master's thesis I conducted whole-genome sequencing and assembly of the river lamprey and the brook lamprey, to create accurate. annotated assemblies, from HiFi and Hi-C reads. I used the hifiasm-Hi-C-integrated assembler to create haplotype resolved assemblies (Cheng et al., 2022), and following curation in the Rapid curation suite (GRIT: Genome Reference Informatics Team, 2022), plus the manual removal of contaminant DNA with the BlobToolKit suite (Challis et al., 2020), I annotated the assemblies using MetaEuk (Levy Karin et al., 2020) and InterProScan (Jones et al., 2014). This resulted in two full haplotype-resolved chromosome level genome assemblies per species, which were used for some initial comparative studies.

The syntenic analyses uncovered a high degree of global synteny both at the intra- and interspecies level, with a low degree of genomic rearrangements. Additionally, I detected a two larger genomic inversion on chromosome 3 and 5 – found to be in a heterozygote state in the brook and river lamprey, respectively, meaning that one of the haplotype genome assemblies for both species (RL1 and BL1) harbour the inverted variants on chromosomes 3 and 5, whereas the other haplotype genome assemblies (RL2 and BL2) do not. I also found a large number of total structural variants on an interspecies level, which was found to be significant after running a χ-square test. Furthermore, for the whole-genome alignment comparisons I found 98.6% alignment-block-similarity between the two species, almost as high as the alignment-block-similarity between the two haplotype genome assemblies (98.9%), indicating that the two genomes are not that divergent and that the separation into two species could be questioned. This was further supported by the interspecies mitochondrial analyses conducted, where 99.31% pairwise identity between the river lamprey and brook lamprey (not accounting for the observed gap in the D-loop of the brook lamprey mitochondrial assembly), were identified. When comparing these mitogenomes to the mitogenomes of sea lamprey and Arctic lamprey, however, the pairwise identities, ranged from around 87.3-90.6%, with the Arctic lamprey being the most similar to both species.

In the BLAST-search I found two SNPs in the vasotocin gene, which were synonymous, i.e. they did not alter the amino acid sequence between the two species. In the rooted species tree generated using OrthoFinder (Emms & Kelly, 2019), the BL2-assembly formed its own sister branch to the branch cluster containing the BL1, RL1 and RL2-assemblies. Although bootstrap support was low for the BL1, RL1, RL2-cluster, the bootstrap value of the branch dividing the BL1-assembly from the other haplotype assemblies was 1, meaning that for every gene tree generated using OrthoFinder, this branch was observed 100% of the time.

In the next sections I want to discuss above-mentioned findings in a broader context. First, I will open with a discussion on the value and validity of whole-genome assemblies, and why this tool is essential in the fight against loss of biodiversity. Moving on, I will examine whole-genome alignments as a measure of phylogeny and evolutionary history and reflect on previous attempts to compare the river lamprey and brook lamprey on a genomic level. I will also discuss gene flow between the species pair and discuss how this affects the river lamprey and brook lamprey's genetic differentiation in sympatric and parapatric populations in Europe. Furthermore, I will return to the hypothesis of phenotypic plasticity, and discuss the two large genomic inversions, and structural variants, in relation to ecotype differences in other species. Finally, I will discuss the weaknesses of my study, and provide suggestions as to how to improve my assemblies and supplement my comparative analyses.

4.2 The value and validity of high-quality whole-genome assemblies

As mentioned in the introduction, my assemblies will be a part of the Earth Biogenome Project, and what defines an accurate assembly is specified within the project's guidelines. Among the list of criteria is that over 90% of the sequences within the assemblies should be assigned to a candidate chromosome, and that there should be a higher than 90% BUSCO-completeness (Lewin et al., 2022; Lewin et al., 2018). Following manual curation using the Rapid curation suite (GRIT: Genome Reference Informatics Team, 2022), and filtration of the suspected contaminant DNA with BlobToolKit (Challis et al., 2020), I was able to assign most of the scaffolds in the assemblies to a chromosome, matching the karyotype of both species, of 82 chromosomes per haplotype. Furthermore, since I was able to do continuous quality controls using BUSCO v5.0.0 (Manni et al., 2021) following each step in the assembly pipeline, I was able to verify that no coding sequences were fragmented during the scaffolding and curation process, as the number of fragmented BUSCOs went down, and the number of complete BUSCOs went up across all assemblies.

I chose to use the hifiasm assembler due to its high performance in the QUAST-comparisons, and because of its ability to explicitly identify haplotypes (Cheng et al., 2021; Cheng et al., 2022). This is made possible by the use of a combination of HiFi and Hi-C reads, which ensures a well-connected assembly graph, while retaining the contiguity of the long, accurate reads (Cheng et al., 2021; Cheng et al., 2022). Building on this robust assembly foundation, with the curation suite developed by the researchers at the Wellcome Sanger Institute (Darwin Tree of Life) (GRIT: Genome Reference Informatics Team, 2022), I manually removed assembly errors, and identified chromosome-sized units, with a high degree of reliability (Howe et al., 2021). The value of high-quality, haplotype resolved, chromosome level assemblies is immensely important for the preservation of biodiversity. Although there are an estimated 10-15 million eukaryotic species in the world, we still only have less than 15 000 complete or partially assembled genomes (Lewin et al., 2022; Lewin et al., 2018). With 23 000-80 000 species approaching extinction, we lack knowledge on how this loss of life will affect the planet's complex ecosystems (Lewin et al., 2022; Lewin et al., 2018). Within the genomes of un-assembled species may lie the secrets behind the ecosystem services provided to us by the flora and fauna around us. Moreover, in understanding the phylogenetic relationships between closely related species, such as the river lamprey and brook lamprey, we can gain insight into their evolutionary histories, and how life on Earth has changed over time.

4.3 Whole-genome alignment as a measure of phylogeny and evolutionary history

Based on the assemblies generated during the assembly, curation, and annotation process, I was able to compare the species pair on both a whole-genome level to show their degree of sequence similarity. Here, I found a 98.6% alignment block similarity on a whole-genome interspecies level, and by aligning the brook lamprey mitochondria to the river lamprey mitochondrial assembly from NCBI, I found a 99.3% pairwise identity on an interspecies mitochondrial level. Whole-genome alignments, like the one I have provided in this study, can be used to infer phylogeny and evolutionary history. If segments, or blocks, of the genome sequences align, we consider them paralogous, i.e. homologous genes which arise due to duplication (Ravinet & Sætre, 2019), or orthologous, i.e. evolutionary related sequences that diverged from their most common recent ancestor (Ravinet & Sætre, 2019). Orthologs have been used in several previous phylogenetic studies, as these are some of the best current tools for estimating evolutionary history (Dewey, 2011). When orthologous sequences are positionally preserved, they are termed toporthologous, and toporthologous regions are likely to share a common genomic function (Dewey, 2011). Moreover, the concept of toporthology is important on a whole-genome level because genes that are found in close proximity are more likely to interact (Dewey, 2011). This interaction affects gene expression, and thus the species' phenotypes.

My findings also show that the percentage of alignment differences between the haplotypes of each species (i.e. the intraspecies differences) are almost the same as the percentage of interspecies alignment differences (see Table 13 and 14). These findings are consistent with the findings from my dotplots (see section 3.6.1), where the haplotypes that show the highest level of synteny are not the within-species assemblies, but rather between RL2, and BL2, which are homozygote for the inversions observed in chromosome 3 in the brook lamprey and chromosome 5 in the river lamprey (see Figure 6, 7 and 8). This is also reflected in the species tree I created using OrthoFinder (see Figure 12), where when assessing all orthologous genes, the BL1 formed its sister group to the monophyletic group containing RL1, RL2 and BL1. When considering that these individuals originated in different freshwater systems, with a high degree of geographical separation, this is a particularly interesting finding.

In previous comparative genomic studies, researchers have relied on microsatellite data and RADseq-technology to compare the two species on a whole-genome level (Hume et al., 2018; Mateus et al., 2013; Rougemont et al., 2017). In 2013, Mateus and colleagues sampled 37 river- and brook lamprey from the Sorraia River in Portugal and created a pseudo-reference genome from one of the sampled individuals spanning 39 865 RAD loci (Mateus et al., 2013). All the sampled individuals were aligned to this reference genome, and from this they recovered 8 826 polymorphic RAD loci, which yielded 14 691 informative SNPs (Mateus et al., 2013). Overall, they found a global F_{ST} of 0.37, and concluded that this suggested strong genome-wide divergence between the two sympatric populations (Mateus et al., 2013). Of these RAD loci, 12 were linked to genes connected to adaption to migratory versus resident life-histories, with the vasotocin gene being named as a major contributor to saltwater-freshwater osmoregulation (Mateus et al., 2013). These findings were echoed by Bracken and colleagues in 2015, when they used a combination of mitochondrial DNA and microsatellite nuclear DNA markers to investigate whether the postglacial expansion of the river lamprey in England, Belgium, Wales and Ireland during the Holocene prompted the establishment of multiple differentiated brook lamprey populations (Bracken et al., 2015). While they failed to find any differentiation between the two species on a mitochondrial level, they found considerable population structure and divergence at microsatellite DNA loci (Bracken et al., 2015). This was especially evident in the brook lamprey populations, but much less so between the migratory river lamprey populations (Bracken et al., 2015).

Although these findings can be used to infer the degree of genetic differentiation, there are pitfalls to relying on microsatellite- and RADseq-data when inferring whole-genome divergence. On the one hand, microsatellites are codominant, highly polymorphic and Mendelian inherited, which makes them suitable for studies of population structure, as well as a tool for measuring differences between closely related species (Mateus et al., 2021; Putman & Carbone, 2014). On the other hand, microsatellites are highly species-specific, and different microsatellite alleles may be obscured due to insertions or deletions within the flanking loci (Mateus et al., 2021; Putman & Carbone, 2014). Also, homoplasy may go undetected among individuals with identical microsatellite lengths due to hidden point mutations (Putman & Carbone, 2014). Genome-wide markers, such as RADseq markers, are a cost-effective solution when there is no a priori whole-genome sequence data (Cerca et al., 2021). However, if the sequence coverage is too low (due to for instance library preparation errors, or computational errors), there can be artificial allelic dropout (Cerca et al., 2021). This missing data can cause limitations, which would not be the case if the species had been compared using whole-genome assemblies.

Moreover, relying solely on mtDNA as a measure of genome divergence also has its drawbacks, as it is difficult to detect recent speciation events when there is incomplete lineage sorting, due to for instance hybridisation (Mateus et al., 2021). Here, the whole-genome assemblies I have created can provide a new tool to make further qualitative assessments, and inspect regions previously deemed to be the subject of strong genomic divergence. For instance, when aligning the sequences found through a BLAST-search using the sea lamprey’s vasotocin annotation, I found synonymous mutations for SNPs at two positions in the region matching the BLAST-search (see Figure 18). Although synonymous mutations can affect genome regulation, through codon-bias and mRNA-stability, these could be examples of neutral mutations (i.e. mutations with no impact on the individual’s fitness). When looking into Mateus and colleagues’ findings, I was unable to find which SNP-differences they observed in their population study, as this information was not provided by the authors. While my data indicates that the differences between the two species have no functional meaning, without population data to back these findings up, I have no way of knowing whether these synonymous SNPs occurred by chance, or if they are the same/fixated for most or all individuals in the sampled populations. For future research, gathering population data can provide further insight into the mechanisms of these synonymous point mutations.

4.4 Evidence of gene flow between the river lamprey and brook lamprey

Another way to assess whether the river lamprey and brook lamprey are paired species or ecotypes of the same species is to investigate their demographic history. Although the river lamprey and brook lamprey in my study were sampled from geographically isolated sites from one another, they still show a high degree of sequence similarity on a whole-genome level. Moreover, between RL2 and BL2 there is a high degree of global synteny, and no inversions between the two haplotypes on chromosomes 3 or 5. The brook lamprey is thought to have originated from the adaptive radiation, and subsequent resident establishment of river lamprey in freshwater following the post-glacial period. In southern European populations, although the distribution records are scarce (Mateus et al., 2012), we know this establishment cannot have happened until around 20 000 years ago (Patton et al., 2017), as most of Southern Europe was still covered in ice. This means the brook lamprey populations observed in Scandinavia (which was covered in ice until around 11 700 years ago (Patton et al., 2017)) have had less time to diverge from the river lamprey, and other brook lamprey populations, than populations in, for instance, France and Portugal. This, or balancing selection, and ongoing gene flow between the paired species at the sample sites, may be a contributing factor to the high degree of observed sequence similarity in this study.

Through Rougemont and colleagues' study from 2015, they found evidence that hybrid offspring with high fitness can occur under semi-natural conditions (Rougemont et al., 2015). In their study, they used a combination of mating trials, experimental crosses, and population genomic analyses to investigate whether the two species were reproductively isolated (Rougemont et al., 2015). Through their crossing experiments, the researchers found that the brook lamprey males could reproduce with river lamprey females, despite their size differences (Rougemont et al., 2015). They also found through analyses of microsatellite data from both sympatric and parapatric populations that there was a continuum of gene flow between the paired species (Rougemont et al., 2015). In some of the sympatric populations there were patterns of panmixia (i.e. random mating), whilst in other sympatric populations, there were patterns of moderate differentiation. In parapatric populations separated by anthropogenic barriers they observed a strongly reduced gene flow, and these findings, in combination with the findings from the sympatric populations, are echoed throughout other population studies of river and brook lamprey.

Rougemont and colleagues hypothesised that this continuum of genetic differentiation could be due to either ecologically based speciation with gene flow, or varying introgression following secondary contract following a period of allopatric divergence (Rougemont et al., 2015; Rougemont et al., 2016). This was further explored in their 2016 study, where they aimed to reconstruct the demographic history of divergence between river lamprey and brook lamprey populations in the Oir, Bethune and Bresle rivers (Rougemont et al., 2016). Again, they found evidence of ongoing gene flow in some instances, while in other sympatric scenarios they were not able to distinguish between the model of ongoing gene flow or whether the degree of genetic differentiation was a result of secondary contact following allopatric divergence (Rougemont et al., 2016). They concluded that to fully be able to distinguish between the primary differentiation and allopatric divergence-hypotheses, a combination of genome-wide analyses and modelling of complex historical processes was necessary (Rougemont et al., 2017; Rougemont et al., 2015; Rougemont et al., 2016). Now, with the availability of my whole-genome assemblies of both species, this, and other analyses will be possible in future research.

4.5 Structural variants as drivers of heterochrony

Heterogenous environments can cause two genetically similar individuals to have very different life-histories. This phenomenon, known as phenotypic plasticity, can not only impact morphological expression, such as colouration and size, but also change individual phenology, and even impact migration behaviour (Lafuente & Beldade, 2019). When some, or all, somatic features are accelerated or delayed relative to sexual maturation, this shift is referred to as heterochrony (Futuyma, 2018). Because of heterochrony, the cascading effects of differing developmental timings can create highly different life-history traits between paired- or cryptic species, regardless of their genetic differences, or similarities.

Both individuals in my study were adults who had reached sexual maturity. This was evident in the presence of gonadal tissue in both individuals during dissection. From previous life-history studies of river lampreys and brook lampreys, we know that brook lampreys reach sexual maturity at a much younger age, and smaller size, than its migratory counterpart (Spice & Docker, 2014). Although I found a high degree of genetic similarity between the two species, on a mitochondrial, whole-genome and syntenic level, to properly assess whether phenotypic plasticity is the driver behind the heterochronic shift observed between them, we need to design studies which isolates the life-history traits for which the river lamprey and brook lamprey diverge. This is because heterochronic traits can occur as a mosaic, meaning that the development of phenotypical traits can differ between organs (Mitteroecker et al., 2005).

Moreover, within parasitic and migratory, and non-parasitic and resident lamprey species pairs, gradients of behavioural differences have been observed, with resident individuals, such as the Western brook lamprey and the American brook lamprey, displaying parasitic behaviour, and migratory species, like the Arctic lamprey, having entirely non-parasitic populations (Neave et al., 2019). Therefore, if we want to investigate feeding strategies, migration, or morphology, environmental- and population data is needed to make proper inferences about whether phenotypic plasticity is the driver behind the developmental and phenological shifts that we observe. By supplementing studies like these with annotated whole-genome assemblies, we can further investigate genes of interest, and hopefully gain new insight into the genome regulatory mechanisms behind traits that are possibly affected by plasticity.

However, as we know from previous studies into the Atlantic salmon genome, structural variants are a major source of phenotypic variation (Bertolotti et al., 2020). In Bertolotti and colleagues' study from 2020, a comparison of structural variants between wild and farmed Atlantic salmon uncovered 15 483 high-confidence structural variants, linked to everything from synaptic networks to life-history traits such as fertility and metabolism. They also uncovered an inversion on chromosome 7, containing 16 genes, which was absent in a large portion of the sampled wild salmon. The impact of inversions has also been under investigation in the Atlantic cod (*Gadus morhua*) genome. In studies from 2016 and 2017, Berg and colleagues found three chromosomal inversions determined to be key contributors to genomic and life-history divergence between migratory and resident ecotypes (Berg et al., 2017; Berg et al., 2016). These chromosomal rearrangements had strong linkage patterns, distinct F_{ST} patterns, and population specific distributions, which in conjunction with the low levels of genomic divergence in the rest of the genomes, indicated that they played a key role in facilitating adaptive genomic divergence (Berg et al., 2016).

From my analysis of structural variations between the river lamprey and brook lamprey assemblies, I found two large inversions on chromosomes 3 and 5 (Figure 6, 7 and 8), and a significant number of structural variant differences between the haplotypes from each species (Table 11). However, similar to the cod comparisons, on a whole-genome level, there was a high degree of alignment-block similarity and synteny. This could mean that the observed inversions and structural variants could act as genomic islands of divergence, i.e. regions that show greater genetic divergence than the rest of the genome, and may over time contribute to reproductive isolation (Futuyma, 2018). Although I lack population data to investigate whether the changes in structural variant allele frequencies are the same across most individuals from each population, or that the observed inversions are present in most individuals from the sampled populations, these findings indicate that the phenotypic differences I have observed in my study may be the result of genomic differences, rather than plasticity.

4.6 Weaknesses of my study, and suggestions on how to improve them

Although I have ensured that the whole-genome assemblies I have created are of Earth BioGenome Project-standards, through various genome-statistical analyses throughout my assembly, scaffolding, curation, and annotation processes, there are some weaknesses to my study. As mentioned in section 4.3 and 4.4, the river lamprey and brook lamprey in this study were collected in different freshwater systems. For optimum comparative conditions, collection of species pairs living in sympatry would be preferred, as there would be a possibility of gene flow. Furthermore, the sequencing results could have been impacted by the differing degrees of tissue freshness between the two species, as the river lamprey was euthanised and immediately dissected and snap-frozen, while the brook lamprey was shipped in ethanol. This is evident from the sequencing results (see section 3.1), where the mean HiFi read length was higher for the river lamprey, and the number of Hifi reads below Q20 were higher in the brook lamprey. This is also visible in the contact maps generated in PretextView (see Appendix B, Figures 25B-32B) where there are far more ambiguous contact signals and unplaced scaffolds in the brook lamprey assemblies than in the river lamprey assemblies.

Due to time constraints, there were also weaknesses in the curation process, as I was only able to perform manual curation in PretextView, and not in HiGlass, which had a higher resolution, as recommended by the researchers at the Wellcome Sanger Institute. Because of this, some of the smaller scaffolds were harder to place. For time management reasons, I also chose the automated annotation pipelines of MetaEuk (Levy Karin et al., 2020) and InterProScan (Jones et al., 2014), as this was a way to do a de novo annotation, combining the ab initio annotation of MetaEuk, which uses algorithms to identify coding regions, with InterProScan, which uses predictive information about protein function from several databases, such as UniProt and OrthoDB (Bateman et al., 2021; Kriventseva et al., 2019).

While this is a time-effective annotation method, it skips a crucial step in the annotation process, namely repeat masking. During the repeat masking step, homologous sequence data is used to identify areas with known repeats, and transform them to “N”’s (i.e. hard masking) or lower case “a”-, “t”-, “c”- or “g”’s (i.e. soft masking). This signals to the annotation software that these areas are repeats, and should not be used in gene database-alignments, or be identified as exons via the transposon’s open reading frames (Yandell & Ence, 2012). This means that the estimated number of genes in Table 10 is likely lower than predicted. Although automated annotation is viewed as less reliable than manual annotation methods, such as Apollo (Dunn et al., 2019), these methods are often time-intensive, and thus only used for smaller genomes. An alternative route, which is still time-effective, is to use the Funannotate v1.5.3 annotation pipeline (Palmer, 2019). Within this software, cleaning and repeat-masking options are available, and you also have the option of providing further protein-based evidence, such as closely related species annotations, to add further reliability to your annotation (Palmer, 2019). Regardless of these constraints, I was able to conduct several BLAST-searches and alignments, such as the included vasotocin-example in section 3.6.5, however, to ensure that no repeat regions were annotated falsely, the annotation process could be repeated using different software, such as Funannotate, before future comparisons.

Another weakness to my study is the lack of population data. To truly be able to make inferences about the relatedness of two paired species, you need more than one sampled individual from each species population. As mentioned in section 4.2, I have no way of knowing whether the differences observed between the two sampled individuals are fixated across the entire population, or whether the number of structural variations between the assemblies detected using Assemblytics (see section 3.6.2) are the same in, for instance, sympatric populations. To validate these findings, more data is needed, and preferably between both geographically isolated populations, and populations with possible gene flow.

5 Future perspectives

5.1 Programmed DNA-elimination

With whole-genome data from the river lamprey and brook lamprey available, we can unlock a world of secrets hidden within their genomes. By backing up my comparative findings with population data, we will be even better equipped to infer whether they are two species, or plastic ecotypes of the same species. Previous studies of the sea lamprey genome have found a reduction in genome size of 20% between the germline genome and the somatic genome (Jeramiah J. Smith et al., 2018). The loss of genomic information is thought to be related to polycomb-group proteins, which functions as Hox-gene silencers (J. J. Smith et al., 2018). During the dissection process, I was able to extract gonadal tissue from the river lamprey, and by sequencing, assembling, and annotating the germline genome, I can compare it to the river lamprey's somatic genome, to verify whether the genomic regions lost during the blastulation process are the same as in the sea lamprey. This is just the first step in further understanding one of the most primitive branches of the vertebrate phylogenetic tree, but in creating more whole-genome assemblies throughout the eukaryotic kingdoms, we will be better prepared in the fight against loss of biodiversity.

6 Conclusion

Through this study, I was able to conduct the first whole-genome sequencing and assembly of the river lamprey and brook lamprey and create four fully haplotype-resolved chromosome-level assemblies of Earth BioGenome Project-standards. My comparative genomic analyses uncovered two large inversions on chromosomes 3 and 5 which occurred between the haplotypes of the within-species assemblies but was absent in the syntenic alignment between RL2 and BL2. Further comparisons of the assemblies also uncovered that the sequence similarities between the two species were as high as between the haplotypes of the same species, with 98.6% vs. 98.9% alignment-block-similarity on a whole-genome level, respectively. Moreover, 99.3% pairwise identity was found on an interspecies mitochondrial level, and significant differences in the number of structural variants between the river and brook lamprey were identified, indicating that genomic variation, and possible genetic divergence, between the two species occurs. Although comparisons between two individuals are insufficient to make a conclusion about their species status, I have created a tool for further comparisons, and a steppingstone for future population-genomic analyses.

7 References

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources, 20*(4), 892-905. https://doi.org/10.1111/1755-0998.13160

Anaconda Software Distribution. (2020). *Anaconda Documentation.* Anaconda Inc. Retrieved 30.05.2022 from https://docs.anaconda.com/.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., … Teodoro, D. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research, 49*(D1), D480-D489. https://doi.org/10.1093/nar/gkaa1100

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., Jentoft, S., & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity, 119*(6), 418-428. https://doi.org/10.1038/hdy.2017.54

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., Jakobsen, K. S., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports, 6*(1), 23246. https://doi.org/10.1038/srep23246

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., Mulugeta, T. D., Ashton, T. J., Hindar, K., Sægrov, H., Florø-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin, S. A. M., … Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications, 11*(1). https://doi.org/10.1038/s41467-020-18972-x

Bohlin, T., Hamrin, S., Heggberget, T. G., Rasmussen, G., & Saltveit, S. J. (1987). Electrofishing - Theory and practice with special emphasis on salmonids. *Hydrobiologia, 173*, 9-43.

Bracken, F. S. A., Hoelzel, A. R., Hume, J. B., & Lucas, M. C. (2015). Contrasting population genetic structure among freshwater-resident and anadromous lampreys: the role of demographic history, differential dispersal and anthropogenic barriers to movement. *Molecular Ecology, 24*(6), 1188-1204. https://doi.org/10.1111/mec.13112

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience, 2*(1), 10. https://doi.org/10.1186/2047-217x-2-10

Broad Institute. (2019). *Picard toolkit*. Retrieved 30.05.2022 from https://broadinstitute.github.io/picard

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ, 6*, e4958. https://doi.org/10.7717/peerj.4958

Cerca, J., Maurstad, M. F., Rochette, N. C., Rivera-Colón, A. G., Rayamajhi, N., Catchen, J. M., & Struck, T. H. (2021). Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. *Methods in Ecology and Evolution, 12*(5), 805-817. https://doi.org/10.1111/2041-210x.13562

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics, 10*(4), 1361-1374. https://doi.org/10.1534/g3.119.400908

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods, 18*(2), 170-175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N. J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol*. https://doi.org/10.1038/s41587-022-01261-x

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). https://doi.org/10.1093/gigascience/giab008

De Cahsan, B., Nagel, R., Schedina, I. M., King, J. J., Bianco, P. G., Tiedemann, R., & Ketmaier, V. (2020). Phylogeography of the European brook lamprey (*Lampetra planeri*) and the European river lamprey (*Lampetra fluviatilis*) species pair based on mitochondrial data. *Journal of Fish Biology, 96*(4), 905-912. https://doi.org/10.1111/jfb.14279

Dewey, C. N. (2011). Positional orthology: putting genomic evolutionary relationships into context. *Briefings in Bioinformatics, 12*(5), 401-412. https://doi.org/10.1093/bib/bbr040

Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2019). Apollo: Democratizing genome annotation. *PLOS Computational Biology, 15*(2), e1006790. https://doi.org/10.1371/journal.pcbi.1006790

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, İ., Docking, T. R., … Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research, 21*(12), 2224-2241. https://doi.org/10.1101/gr.126599.111

Elzanowski, A., & Ostell, J. (2019). *The Genetic Codes. National Center for Biotechnology Information (NCBI).* Retrieved 30.05.2022 from https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Emms, D. M., & Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *bioRxiv preprint*. https://doi.org/10.1101/267914

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology, 20*(1). https://doi.org/10.1186/s13059-019-1832-y

Freyhof, J., & Kottelat, M. (2007). *Handbook of European freshwater fishes*. ICUN.

Fricke, R., Eschmeyer, W., & Fong, J. D. (2022). *Eschmeyer's Catalog of Fishes*. Retrieved 30.05.2022 from https://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp

Futuyma, D. K., Mark. (2018). *Evolution* (Fourth edition ed.). Oxford University Press.

Gachelin, G. (2000). *Lampetra fluviatilis mitochondrion, complete genome*. Retrieved 30.05.2022 from https://www.ncbi.nlm.nih.gov/nucleotide/NC_001131.1

Ghurye, J., & Pop, M. (2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Comput Biol, 15*(6), e1006994. https://doi.org/10.1371/journal.pcbi.1006994

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017, Jul 12). Scaffolding of long read assemblies using long range contact information. *BMC Genomics, 18*(1), 527. https://doi.org/10.1186/s12864-017-3879-z

Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J, 18*, 9-19. https://doi.org/10.1016/j.csbj.2019.11.002

Glover, K. A., Solberg, M. F., Besnier, F., & Skaala, Ø. (2018). Cryptic introgression: evidence that selection and plasticity mask the full phenotypic potential of domesticated Atlantic salmon in the wild. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-32467-2

GRIT: Genome Reference Informatics Team. (2022). *Rapid curation / rapid_hic_software*. Retrieved 25.01.2022 from https://gitlab.com/wtsi-grit/rapid-curation/-/tree/main/rapid_hic_software

Guan, D., McCarthy, S. A., Ning, Z., Wang, G., Wang, Y., & Durbin, R. (2021, Nov 27). Efficient iterative Hi-C scaffolder based on N-best neighbors. *BMC Bioinformatics, 22*(1), 569. https://doi.org/10.1186/s12859-021-04453-5

Hahn, M. W. (2019). *Molecular Population Genetics*. Oxford University Press.

Harry, E. (2021). *PretextView* (Version 0.2.5) [Computer software]. https://github.com/wtsi-hpag/PretextView

Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., & Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. *GigaScience, 10*(1). https://doi.org/10.1093/gigascience/giaa153

Hume, J. B., Recknagel, H., Bean, C. W., Adams, C. E., & Mable, B. K. (2018). RADseq and mate choice assays reveal unidirectional gene flow among three lamprey ecotypes despite weak assortative mating: Insights into the formation and stability of multiple ecotypes in sympatry. *Molecular Ecology, 27*(22), 4572-4590. https://doi.org/10.1111/mec.14881

Ishijima, J., Uno, Y., Nunome, M., Nishida, C., Kuraku, S., & Matsuda, Y. (2016). Molecular cytogenetic characterization of chromosome site-specific repetitive sequences in the Arctic lamprey (Lethenteron camtschaticum, Petromyzontidae). *DNA Research*, dsw053. https://doi.org/10.1093/dnares/dsw053

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240. https://doi.org/10.1093/bioinformatics/btu031

Jonsson, B., & Jonsson, N. (2011). *Ecology of Atlantic Salmon and Brown Trout* (Vol. 33). Springer. https://doi.org/10.1007/978-94-007-1189-1

Kelly, F., & K., James. (2001). A review of the ecology and distribution of three lamprey species, Lampetra fluviatilis (L.), Lampetra planeri (Bloch) and Petromyzon marinus (L.): A context for conservation and biodiversity considerations in Ireland. *Biology & Environment Proceedings of the Royal Irish Academy, 101*(3).

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019, May). Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol, 37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena, R., Wegener, J., & Turner, S. W. (2010). Real-Time DNA Sequencing from Single Polymerase Molecules. In *Single Molecule Tools: Fluorescence Based Approaches, Part A* (pp. 431-455). https://doi.org/10.1016/s0076-6879(10)72001-2

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research, 47*(D1), D807-D811. https://doi.org/10.1093/nar/gky1053

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS One, 12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Lafuente, E., & Beldade, P. (2019). Genomics of Developmental Plasticity in Animals. *Front Genet, 10*, 720. https://doi.org/10.3389/fgene.2019.00720

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics, 30*(22), 3276-3278. https://doi.org/10.1093/bioinformatics/btu531

Lee, J.-S. (2013). *Lethenteron camtschaticum mitochondrion, complete genome*. Retrieved 30.05.2022 from https://www.ncbi.nlm.nih.gov/nucleotide/NC_020465.1

Levy Karin, E., Mirdita, M., & Soding, J. (2020). MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome, 8*(1), 48. https://doi.org/10.1186/s40168-020-00808-x

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Balint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S. M., Chen, L., Childers, A. K., Coddington, J. A., Conde, D. A., … Zhang, G. (2022, Jan 25). The Earth BioGenome Project 2020: Starting the clock. *Proc Natl Acad Sci U S A, 119*(4). https://doi.org/10.1073/pnas.2115635118

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A, 115*(17), 4325-4333. https://doi.org/10.1073/pnas.1720115115

Li, H. (2018). *seqtk - Toolkit for processing sequences in FASTA/Q formats*. Retrieved 30.05.2022 from https://github.com/lh3/seqtk

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol, 38*(10), 4647-4654. https://doi.org/10.1093/molbev/msab199

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology, 14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Mateus, C. S., Docker, M. F., Evanno, G., Hess, J. E., Hume, J. B., Oliveira, I. C., Souissi, A., & Sutton, T. M. (2021). Population structure in anadromous lampreys: Patterns and processes. *Journal of Great Lakes Research, 47*, S38-S58. https://doi.org/10.1016/j.jglr.2021.08.024

Mateus, C. S., Rodríguez-Muñoz, R., Quintella, B. R., Alves, M. J., & Almeida, P. R. (2012). Lampreys of the Iberian Peninsula: distribution, population status and conservation. *Endangered Species Research, 16*(2), 183-198. https://doi.org/10.3354/esr00405

Mateus, C. S., Stange, M., Berner, D., Roesti, M., Quintella, B. R., Alves, M. J., Almeida, P. R., & Salzburger, W. (2013). Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology, 23*(15), R649-R650. https://doi.org/10.1016/j.cub.2013.06.026

Mayasich, S. A. (2016). *Petromyzon marinus vasotocin V1 hormone receptor 2 mRNA, complete cds*. Retrieved 30.05.2022 from https://www.ncbi.nlm.nih.gov/nuccore/KJ813004.1?report=FASTA

McKeown, N. J., Hynes, R. A., Duguid, R. A., Ferguson, A., & Prodöhl, P. A. (2010). Phylogeographic structure of brown troutSalmo truttain Britain and Ireland: glacial refugia, postglacial colonization and origins of sympatric populations. *Journal of Fish Biology*, *76*(2), 319-347. https://doi.org/10.1111/j.1095-8649.2009.02490.x

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux J., 2014*(239), Article 2.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics, 34*(13), i142-i150. https://doi.org/10.1093/bioinformatics/bty266

Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., & Levy Karin, E. (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, *37*(18), 3029-3031. https://doi.org/10.1093/bioinformatics/btab184

Mitteroecker, P., Gunz, P., & Bookstein, F. L. (2005). Heterochrony and geometric morphometrics: a comparison of cranial growth in *Pan paniscus* versus *Pan troglodytes*. *Evolution & Development*, *7*(3), 244-258. https://doi.org/10.1111/j.1525-142x.2005.05027.x

Nattestad, M., & Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, *32*(19), 3021-3023. https://doi.org/10.1093/bioinformatics/btw369

NBIS. (2019). *Genome Annotation Workshop*. Retrieved 30.05.2022 from https://github.com/NBISweden/workshop-genome_annotation

Neave, F. B., Steeves, T. B., Pratt, T. C., McLaughlin, R. L., Adams, J. V., & Docker, M. F. (2019). Stream characteristics associated with feeding type in silver (Ichthyomyzon unicuspis) and northern brook (I. fossor) lampreys and tests for phenotypic plasticity. *Environmental Biology of Fishes, 102*(4), 615-627. https://doi.org/10.1007/s10641-019-00857-8

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution, 32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *biorXiv*. https://doi.org/10.1101/2020.03.14.992248

Pacific Biosciences of California. (2020). *Sequence With Confidence Using PacBio Highly Accurate Long Reads*. Retrieved 05.10.2021 from https://www.pacb.com/

Palmer, J. a. S. J. (2019). *nextgenusfs/funannotate: funannotate v1.5.3*. Retrieved 30.05.2022 from https://zenodo.org/record/2604804#.Yo6TNKhBy3A

Patton, H., Hubbard, A., Andreassen, K., Auriac, A., Whitehouse, P. L., Stroeven, A. P., Shackleton, C., Winsborrow, M., Heyman, J., & Hall, A. M. (2017). Deglaciation of the Eurasian ice sheet complex. *Quaternary Science Reviews, 169*, 148-172. https://doi.org/10.1016/j.quascirev.2017.05.019

Potter, I. C., Gill, H. S., Renaud, C. B., & Haoucher, D. (2015). The Taxonomy, Phylogeny, and Distribution of Lampreys. In *Lampreys: Biology, Conservation and Control* (pp. 35-73). https://doi.org/10.1007/978-94-017-9306-3_2

Putman, A. I., & Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol Evol, 4*(22), 4399-4428. https://doi.org/10.1002/ece3.1305

Pörtner, H.-O., Roberts, D. C., Tignor, M., Poloczanska, E. S., & al., e. (2022). *IPCC, 2022: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Ravinet, M., & Sætre, G.-P. (2019). *Evolutionary Genetics: Concepts, Analysis, and Practice*.

Rougemont, Q., Gagnaire, P. A., Perrier, C., Genthon, C., Besnard, A. L., Launey, S., & Evanno, G. (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Mol Ecol, 26*(1), 142-162. https://doi.org/10.1111/mec.13664

Rougemont, Q., Gaigher, A., Lasne, E., Côte, J., Coke, M., Besnard, A. L., Launey, S., & Evanno, G. (2015). Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *Journal of Evolutionary Biology, 28*(12), 2248-2263. https://doi.org/10.1111/jeb.12750

Rougemont, Q., Roux, C., Neuenschwander, S., Goudet, J., Launey, S., & Evanno, G. (2016). Reconstructing the demographic history of divergence between European river and brook lampreys using approximate Bayesian computations. *PeerJ, 4*, e1910. https://doi.org/10.7717/peerj.1910

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz342

RStudio Team. (2021). *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston, MA. Retrieved 30.05.2022 from http://www.rstudio.com/

Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., Cook, M. E., Hess, J. E., Narum, S. R., Lamanna, F., Kaessmann, H., Timoshevskiy, V. A., Waterbury, C. K. M., Saraceno, C., Wiedemann, L. M., Robb, S. M. C., Baker, C., Eichler, E. E., Hockman, … Amemiya, C. T. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature Genetics, 50*(2), 270-277. https://doi.org/10.1038/s41588-017-0036-1

Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., Cook, M. E., Hess, J. E., Narum, S. R., Lamanna, F., Kaessmann, H., Timoshevskiy, V. A., Waterbury, C. K. M., Saraceno, C., Wiedemann, L. M., Robb, S. M. C., Baker, C., Eichler, E. E., Hockman, D., … Amemiya, C. T. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat Genet, 50*(2), 270-277. https://doi.org/10.1038/s41588-017-0036-1

Spice, E. K., & Docker, M. F. (2014). Reduced fecundity in non-parasitic lampreys may not be due to heterochronic shift in ovarian differentiation. *Journal of Zoology, 294*(1), 49-57. https://doi.org/10.1111/jzo.12150

Steenwyk, J. L., Buida, T. J., III, Labella, A. L., Li, Y., Shen, X.-X., & Rokas, A. (2021). PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics, 37*(16), 2325-2331. https://doi.org/10.1093/bioinformatics/btab096

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., Dong, Q., & Ye, K. (2021, Sep 3). High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. *Genomics Proteomics Bioinformatics*. https://doi.org/10.1016/j.gpb.2021.08.003

Webb, J., Verspoor, E., Aubin-Horth, N., & al, e. (2007). *The Atlantic Salmon: Genetics, Conservation and Management*. Blackwell publishing.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol, 37*(10), 1155-1162. https://doi.org/10.1038/s41587-019-0217-9

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics, 13*(5), 329-342. https://doi.org/10.1038/nrg3174

Østbye, K., Hagen Hassve, M., Peris Tamayo, A. M., Hagenlund, M., Vogler, T., & Præbel, K. (2020). "And if you gaze long into an abyss, the abyss gazes also into thee": four morphs of Arctic charr adapting to a depth gradient in Lake Tinnsjøen. *Evolutionary Applications, 13*(6), 1240-1261. https://doi.org/10.1111/eva.12983

8 List of abbreviations, figures, and tables

8.1 List of abbreviations

RL1 = River lamprey haplotype 1

RL2 = River lamprey haplotype 2

RLP = River lamprey primary contigs

BL1 = Brook lamprey haplotype 1

BL2 = Brook lamprey haplotype 2

BLP = Brook lamprey primary contigs

PMS = Sea lamprey (*Petromyzon marinus*) somatic assembly

PMG = Sea lamprey (*Petromyzon marinus*) germline assembly

8.2 List of figures in main text

8.3 List of tables in main text

Table 11: The total number of structural variants between the RL1, RL2, BL1 and BL2-assemblies.

Table 12: The total number of bases affected by structural variants between each haplotype.

Table 13: The percentage of aligned bases between the query and reference FASTA files when using the wrapper script DNAdiff around nucmer from the MUMmer suite.

Table 14: The percentage of alignment blocks comprising the 1-to-1 mapping of the reference assemblies to the query assemblies.

Table 15: Mitochondrial pairwise identities.

9 Appendices

9.1 Appendix A – Methods

Arima Genomics, Inc. (2019). Arima-HiC 2.0 kit standard user guide for Animal tissues. San Diego, USA.

Dovetail Genomics. (2019). Omni-C Proximity Ligation assay for Non-mammalian samples, version 1.0. Chicago, USA.

Pacific Biosciences of California. (2021). Preparing HiFi SMRTbell® Libraries using the SMRTbell Express Template Prep Kit 2.0. California, USA.

Pacific Biosciences of California. (2021). Genome Assembly pipeline (SMRT Link 10.1.0.119588). California, USA.

9.2 Appendix B - Results

9.2.1 QUAST-results river lamprey

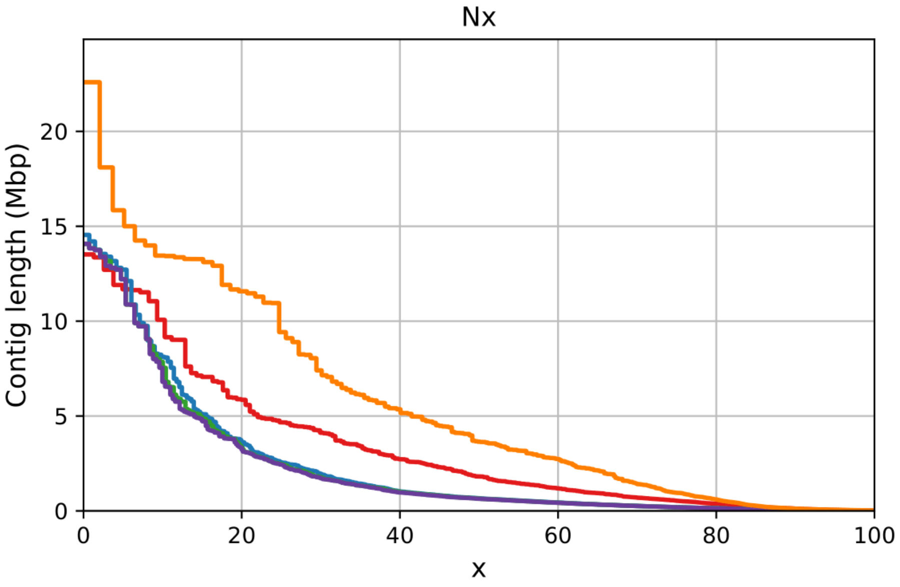


Figure 1B: Plot of the first river lamprey assemblies, showing the contig length in Mbp on the y-axis, and the percentage of contigs of that length on the x-axis. The Nx-metric is defined as the length of the shortest contig which covers x% of the assembly. Here, the hifiasm assembly is indicated by the yellow line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the blue line, the HiCanu assembly with minimum overlap of 500 is marked by the green line, and the HiCanu assembly with minimum overlap of 700 is marked with the purple line.

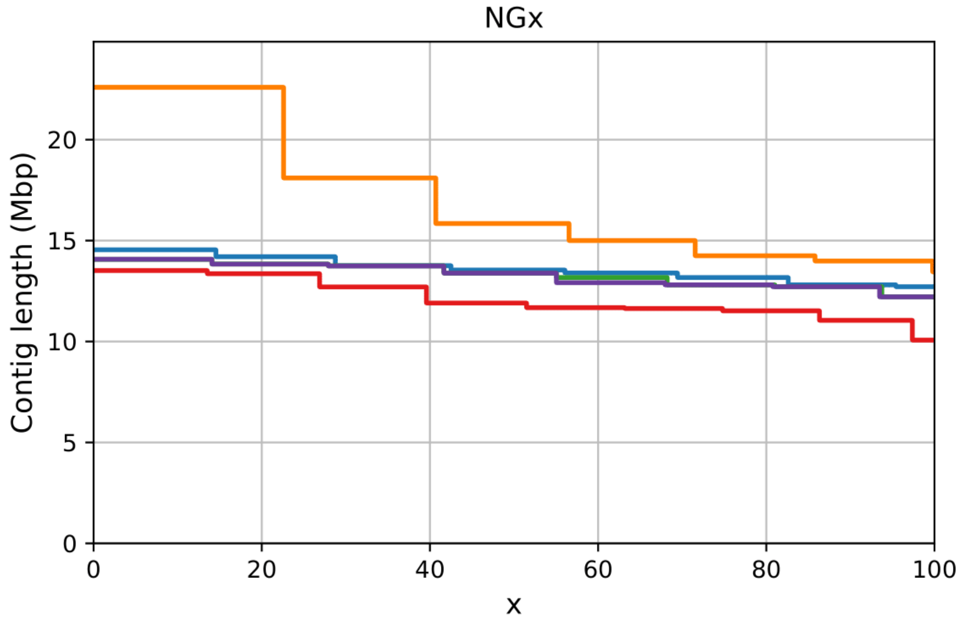


Figure 2B: Plot of the first river lamprey assemblies, showing the contig length in Mbp on the y-axis, and the percentage of contigs of that length on the x-axis. Unlike the Nx-metric, the NGx-metric relates to the genome size, rather than the assembly size. Here, the hifiasm assembly is indicated by the yellow line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the blue line, the HiCanu assembly with minimum overlap of 500 is marked by the green line, and the HiCanu assembly with minimum overlap of 700 is marked with the purple line.

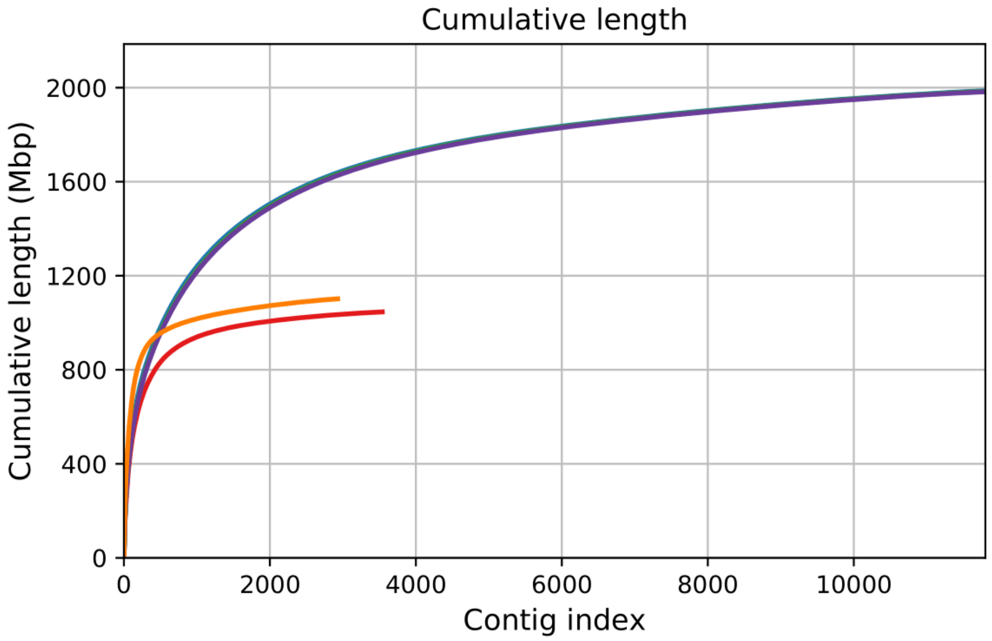


Figure 3B: Plot showing the cumulative length of all contigs assembled. Here contigs are ordered from largest to smallest, with the y-axis showing the number of contigs, and the x-axis showing the length of all the ordered contigs. Here, the hifiasm assembly is indicated by the yellow line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the blue line, the HiCanu assembly with minimum overlap of 500 is marked by the green line, and the HiCanu assembly with minimum overlap of 700 is marked with the purple line. The lines for the HiCanu assemblies are overlapping, with only the purple and blue lines visible.

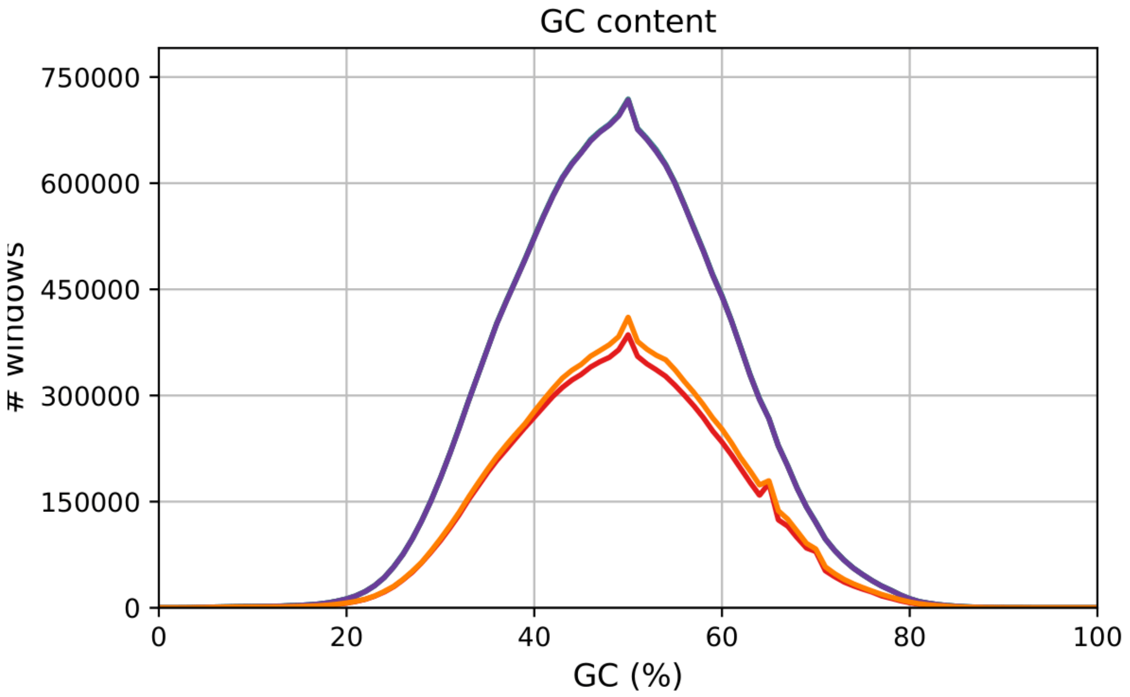


Figure 4B: Plot showing the GC-content of the assemblies, where the contigs are broken into non-overlapping 100 bp windows. The plot shows the number of windows for each GC-percentage. The y-axis shows the number of windows, and the x-axis shows the percentage of GC-content within the window. Here, the hifiasm assembly is indicated by the yellow line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the blue line, the HiCanu assembly with minimum overlap of 500 is marked by the green line, and the HiCanu assembly with minimum overlap of 700 is marked with the purple line. The lines for the HiCanu assemblies are overlapping, with only the purple line visible.

9.2.2 QUAST-results brook lamprey

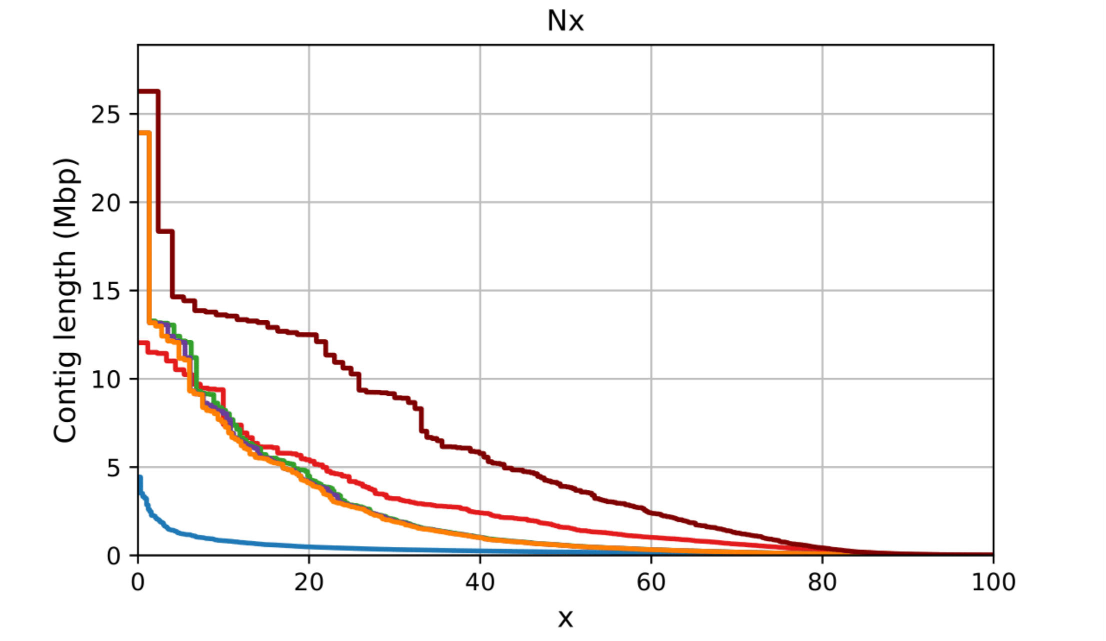


Figure 5B: Plot of the first brook lamprey assemblies, showing the contig length in Mbp on the y-axis, and the percentage of contigs of that length on the x-axis. The Nx-metric is defined as the length of the shortest contig which covers x% of the assembly. Here, the hifiasm assembly is indicated by the maroon line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the green line, the HiCanu assembly with minimum overlap of 500 is marked by the purple line, the HiCanu assembly with minimum overlap of 700 is marked with the yellow line and the Flye assembly is marked by the blue line.

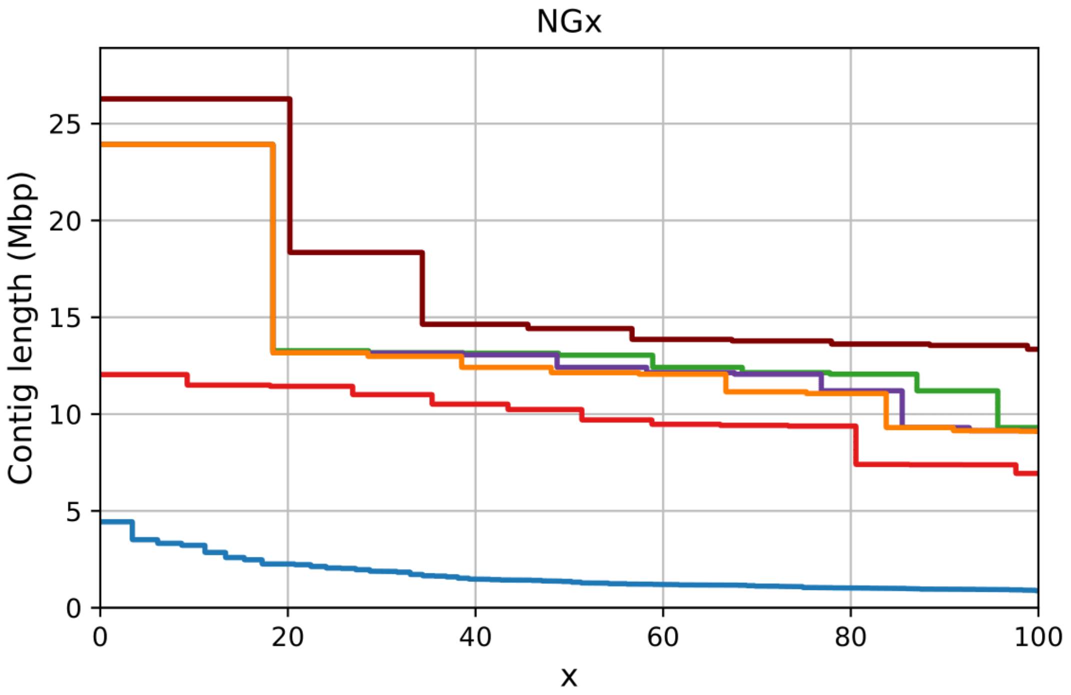


Figure 6B: Plot of the first brook lamprey assemblies, showing the contig length in Mbp on the y-axis, and the percentage of contigs of that length on the x-axis. Unlike the Nx-metric, the NGx-metric relates to the genome size, rather than the assembly size. Here, the hifiasm assembly is indicated by the maroon line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the green line, the HiCanu assembly with minimum overlap of 500 is marked by the purple line, the HiCanu assembly with minimum overlap of 700 is marked with the yellow line and the Flye assembly is marked by the blue line.

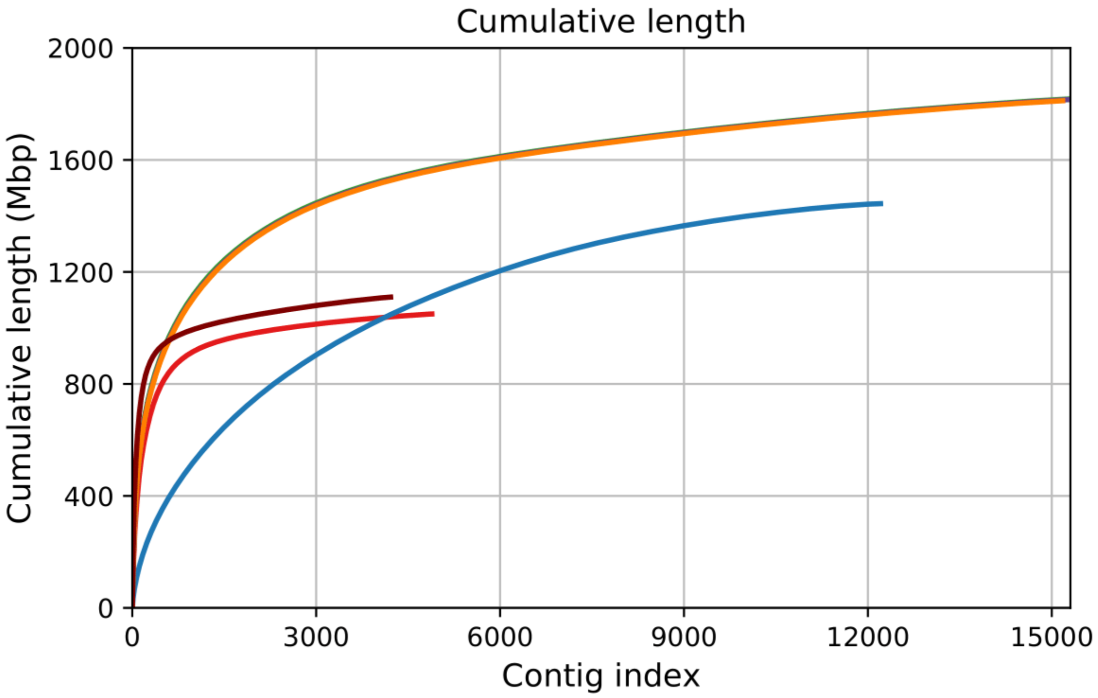


Figure 7B: Plot showing the cumulative length of all contigs assembled. Here contigs are ordered from largest to smallest, with the y-axis showing the number of contigs, and the x-axis showing the length of all the ordered contigs. Here, the hifiasm assembly is indicated by the maroon line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the green line, the HiCanu assembly with minimum overlap of 500 is marked by the purple line, the HiCanu assembly with minimum overlap of 700 is marked with the yellow line and the Flye assembly is marked by the blue line. The lines showing the HiCanu assemblies are overlapping, and only the yellow line is clearly visible.

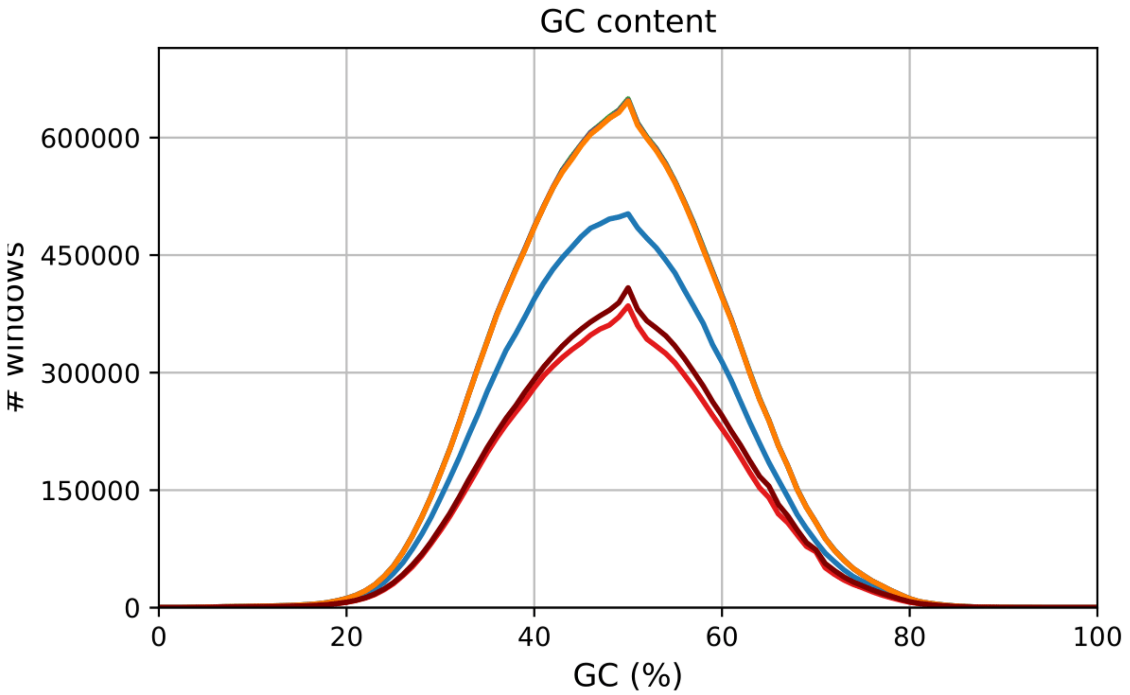


Figure 8B: Plot showing the GC-content of the assemblies, where the contigs are broken into non-overlapping 100 bp windows. The plot shows the number of windows for each GC-percentage. The y-axis shows the number of windows, and the x-axis shows the percentage of GC-content within the window. Here, the hifiasm assembly is indicated by the maroon line, the IPA assembly is indicated by the red line, the HiCanu assembly with minimum overlap of 200 bp is marked by the green line, the HiCanu assembly with minimum overlap of 500 is marked by the purple line, the HiCanu assembly with minimum overlap of 700 is marked with the yellow line and the Flye assembly is marked by the blue line. The lines showing the HiCanu assemblies are overlapping, and only the yellow line is clearly visible.

9.2.3 D-GENIES dotplots

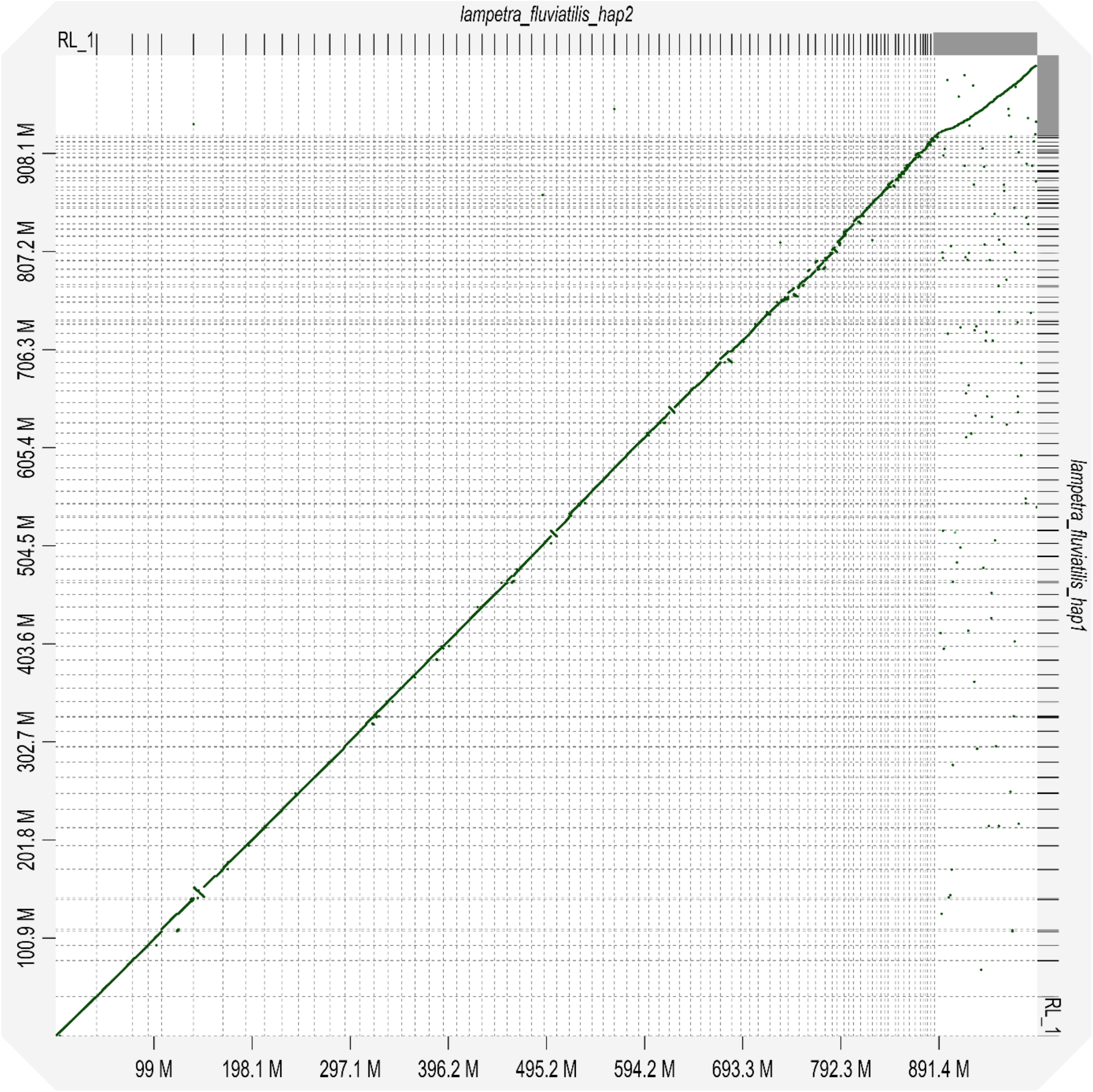


Figure 9B: Plot showing the synteny between RL2 (x-axis) and RL1 (y-axis).

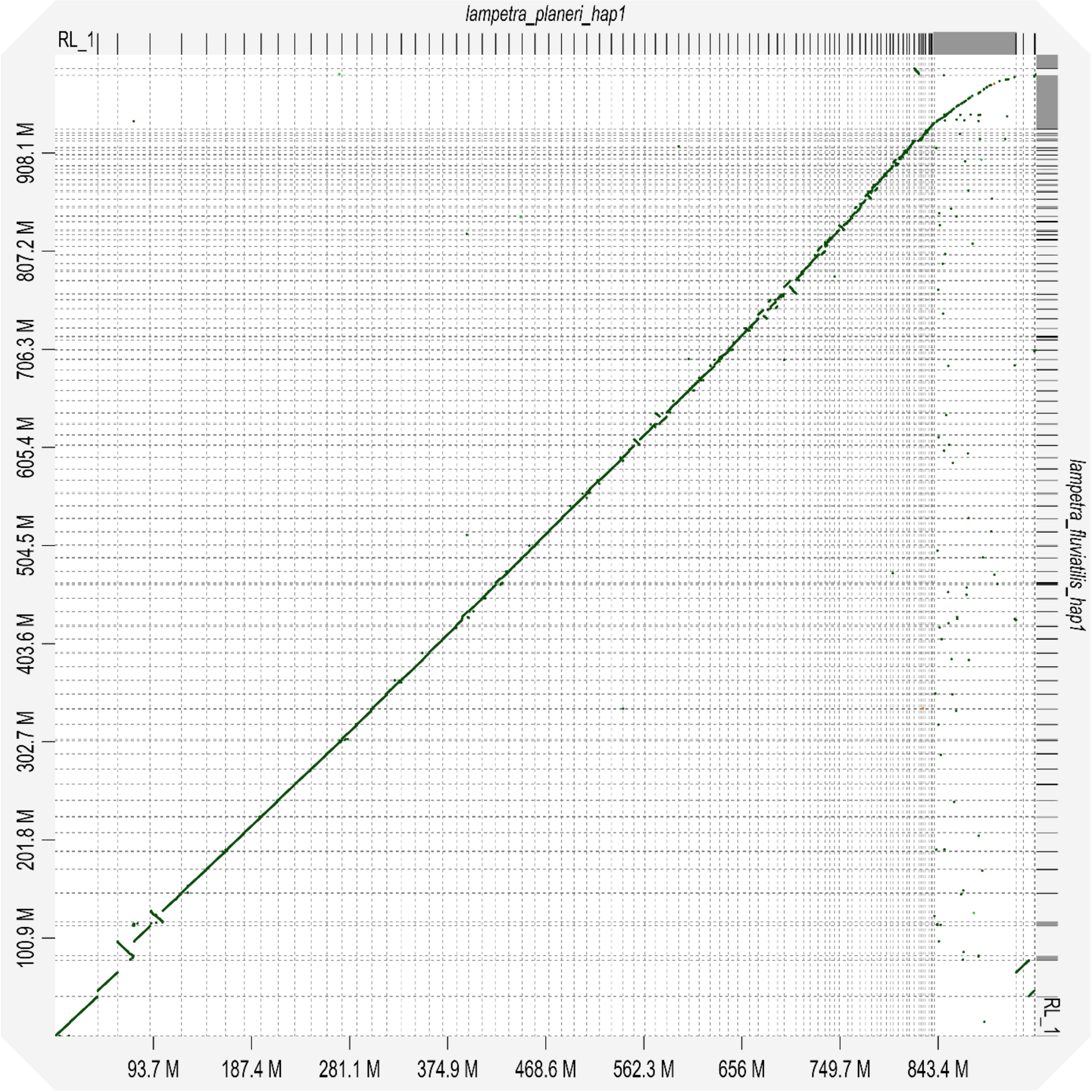


Figure 10B: Plot showing the synteny between BL1 (x-axis) and RL1 (y-axis).

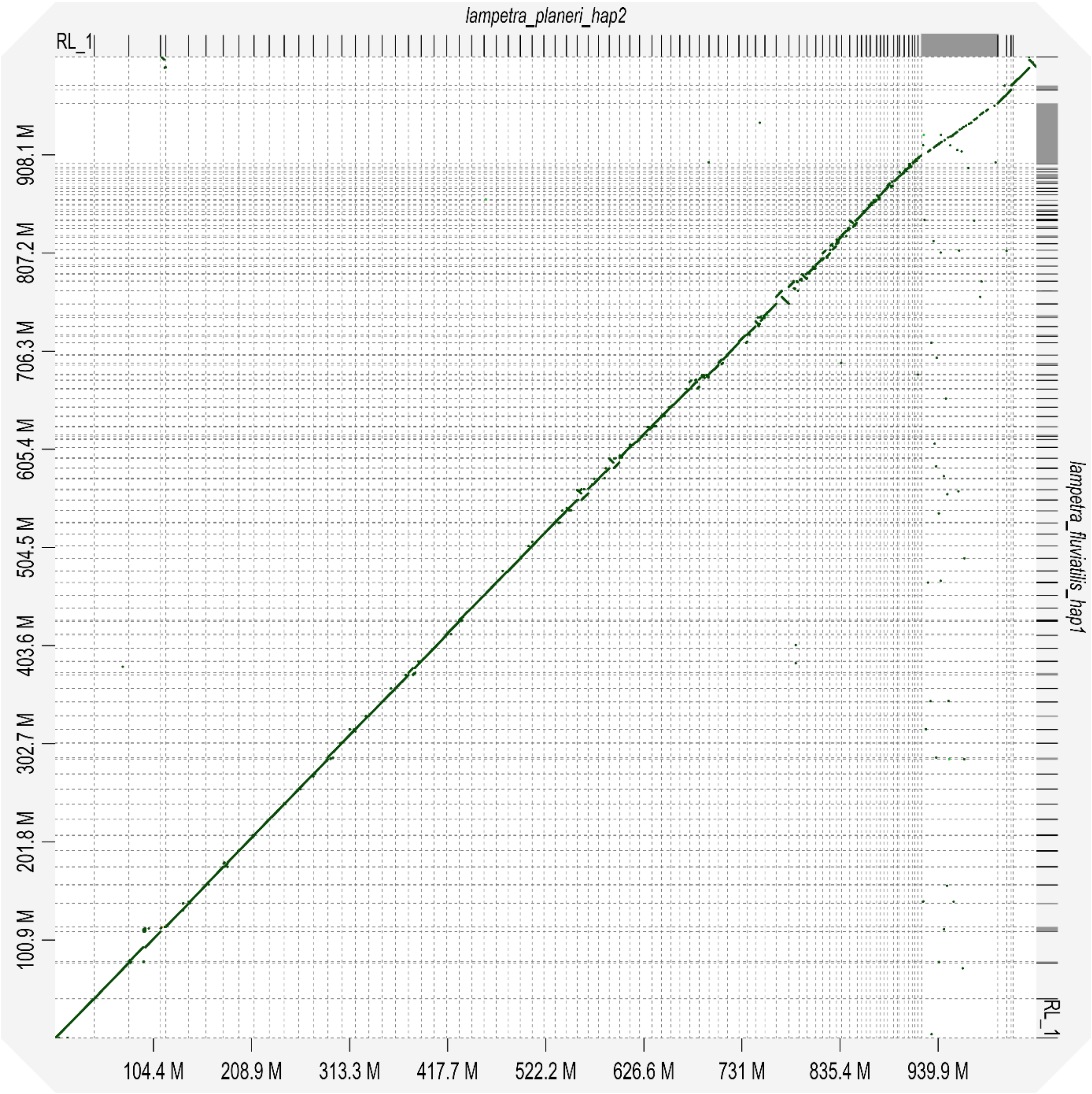


Figure 11B: Plot showing the synteny between BL2 (x-axis) and RL1 (y-axis). Chromosome 5 is displaced, and the inversion can be observed in the top right corner.

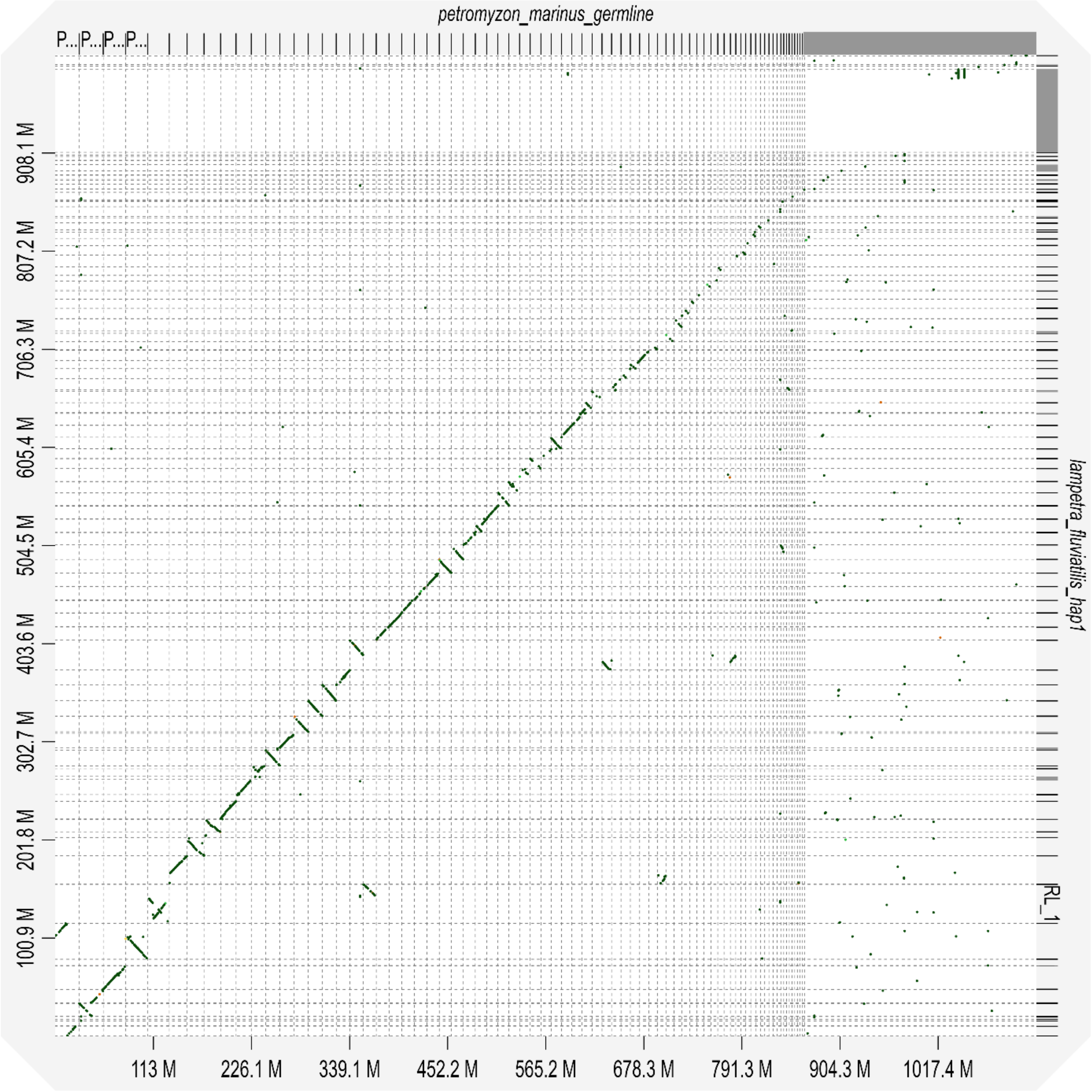


Figure 12B: Plot showing the synteny between PMG (x-axis) and RL1 (y-axis).

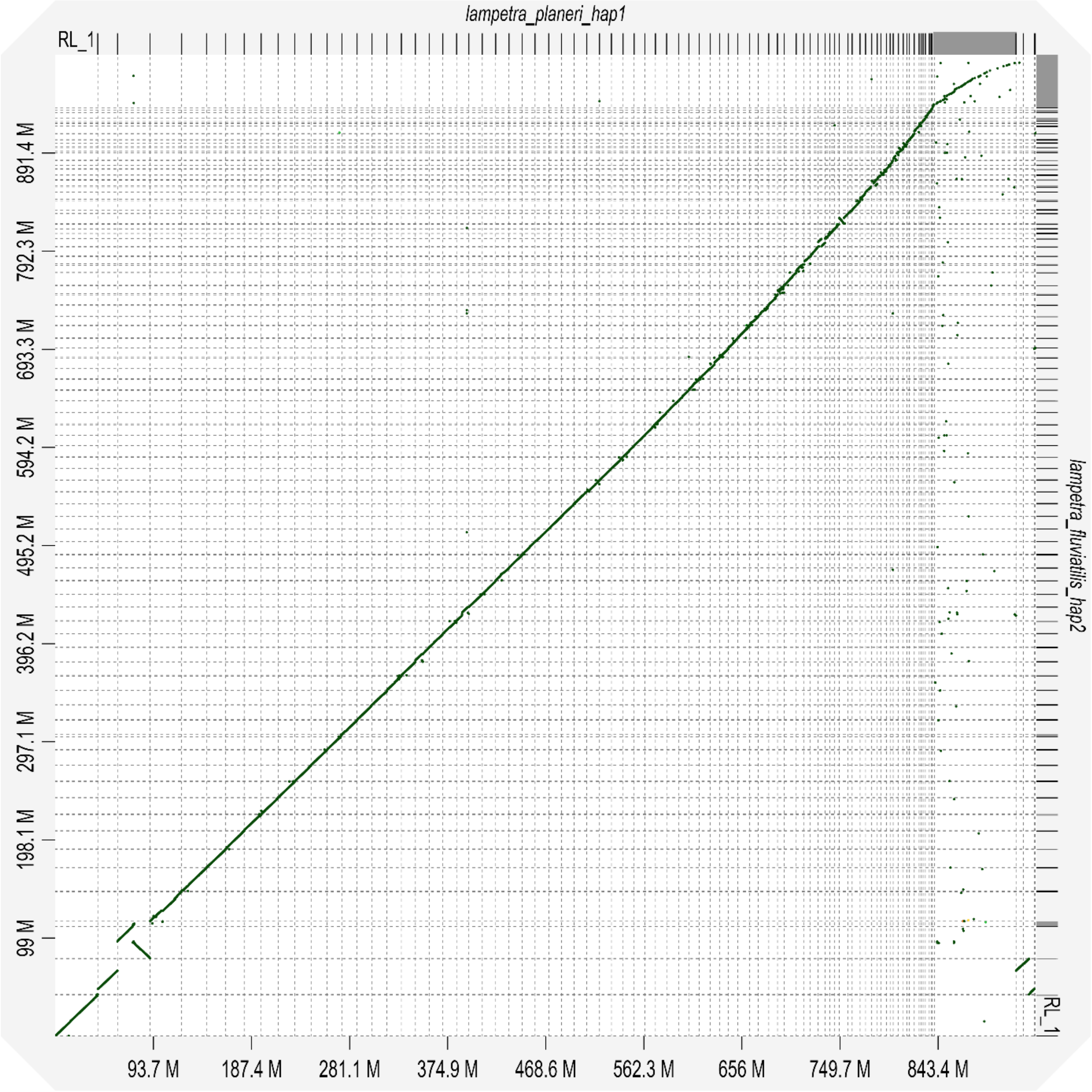


Figure 13B: Plot showing the synteny between BL1 (x-axis) and RL2 (y-axis).

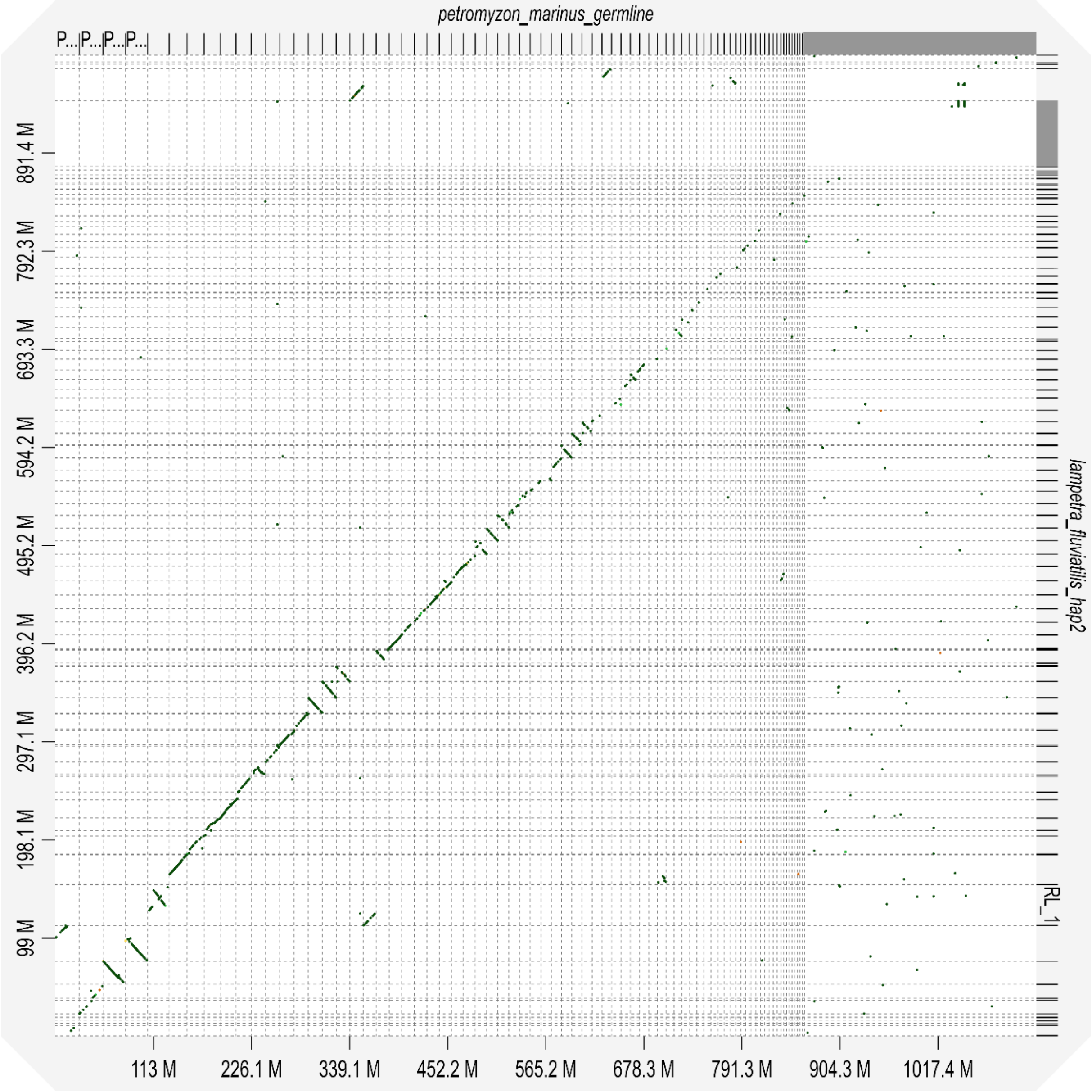


Figure 14B: Plot showing the synteny between PMG (x-axis) and RL2 (y-axis).

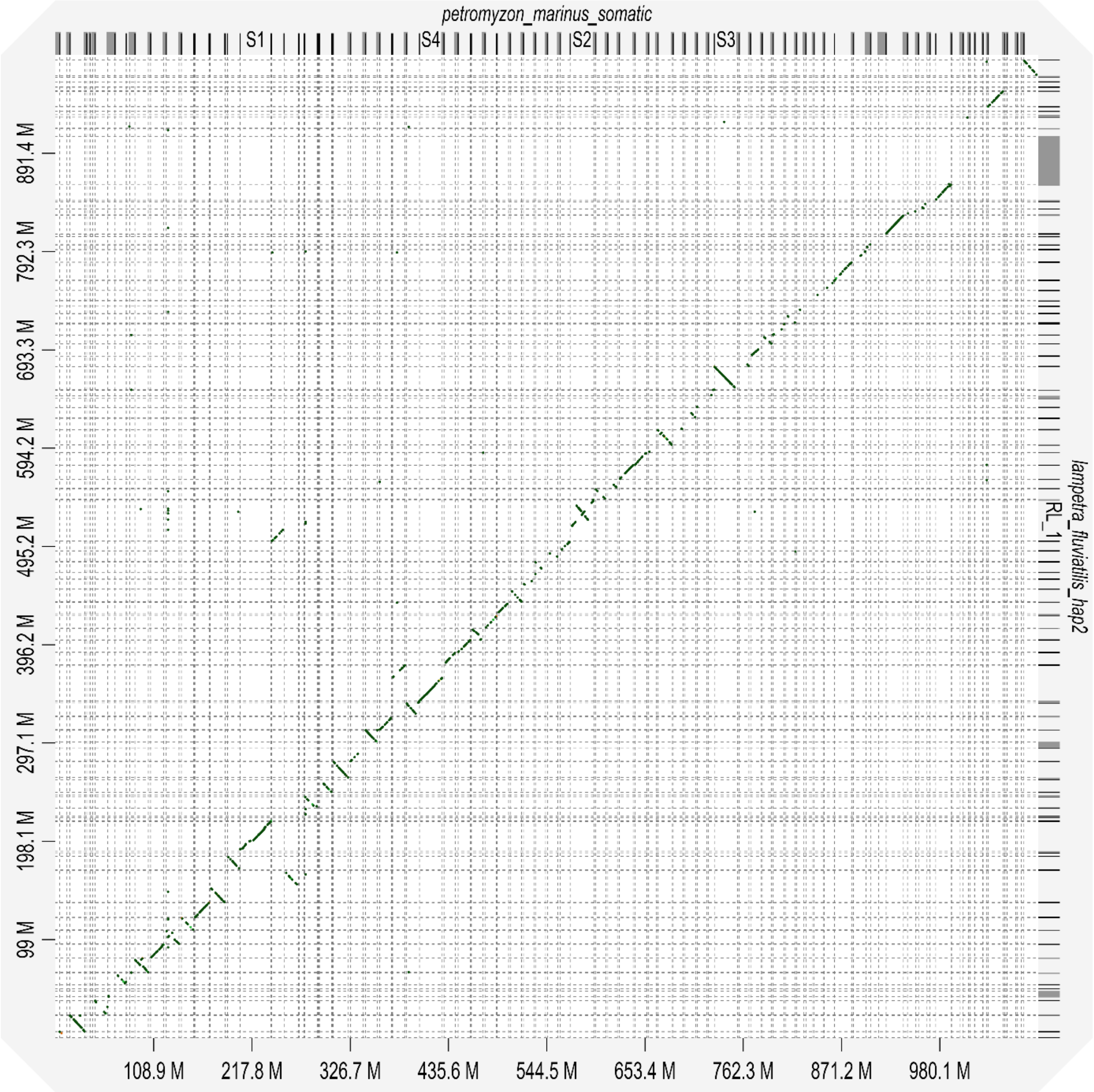


Figure 15B: Plot showing the synteny between PMS (x-axis) and RL2 (y-axis).

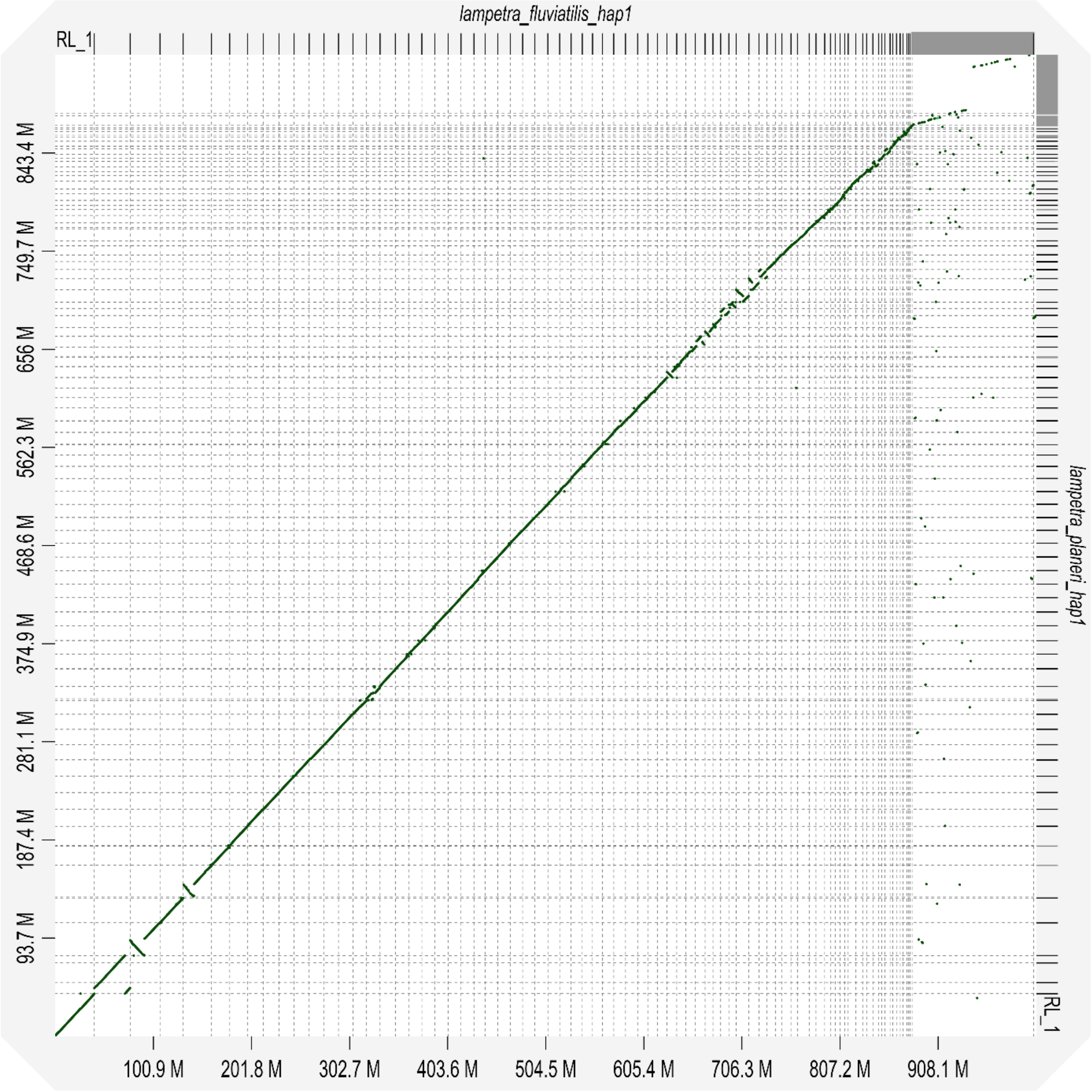


Figure 16B: Plot showing the synteny between RL1 (x-axis) and BL1 (y-axis).

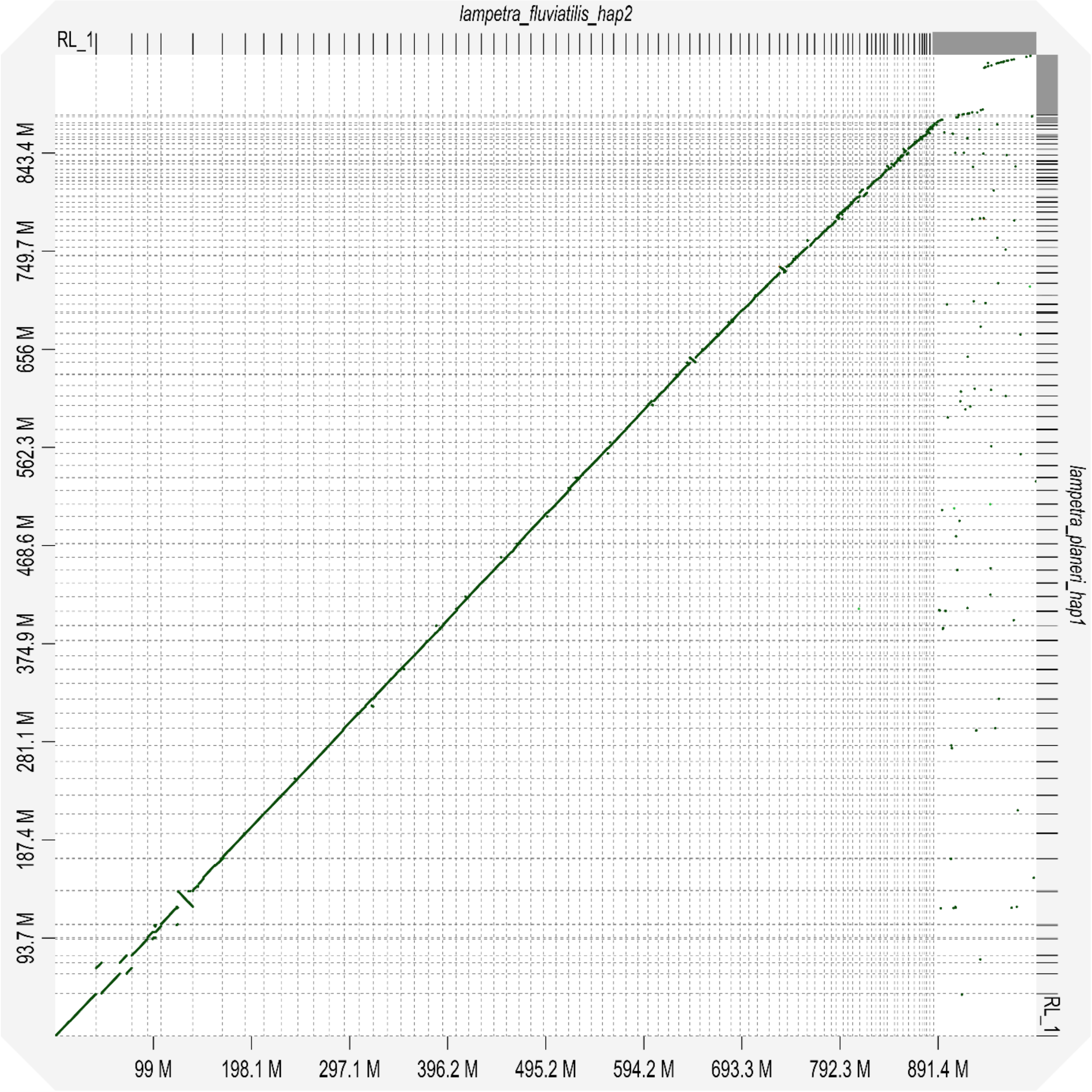


Figure 17B: Plot showing the synteny between RL2 (x-axis) and BL1 (y-axis).

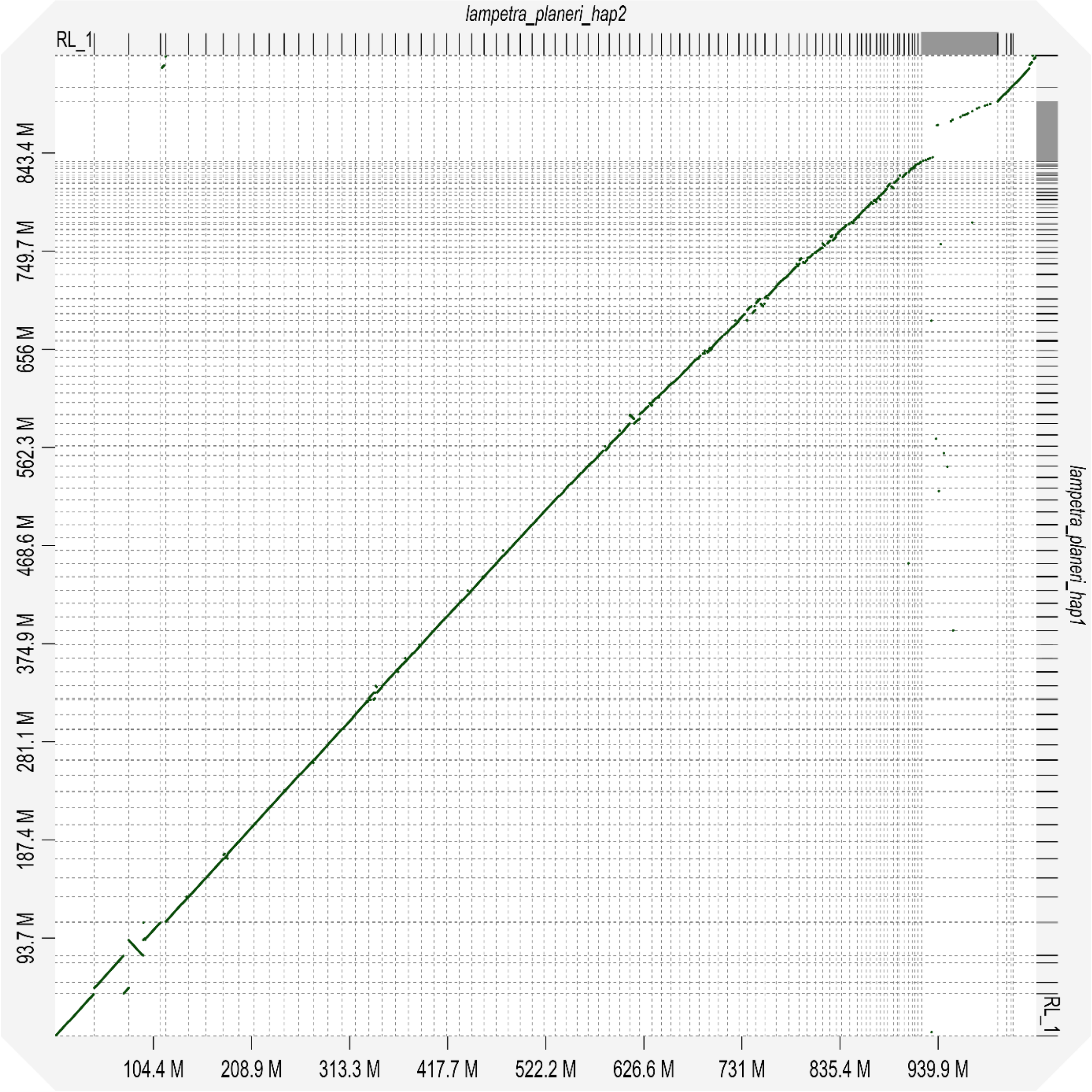


Figure 18B: Plot showing the synteny between BL2 (x-axis) and BL1 (y-axis).

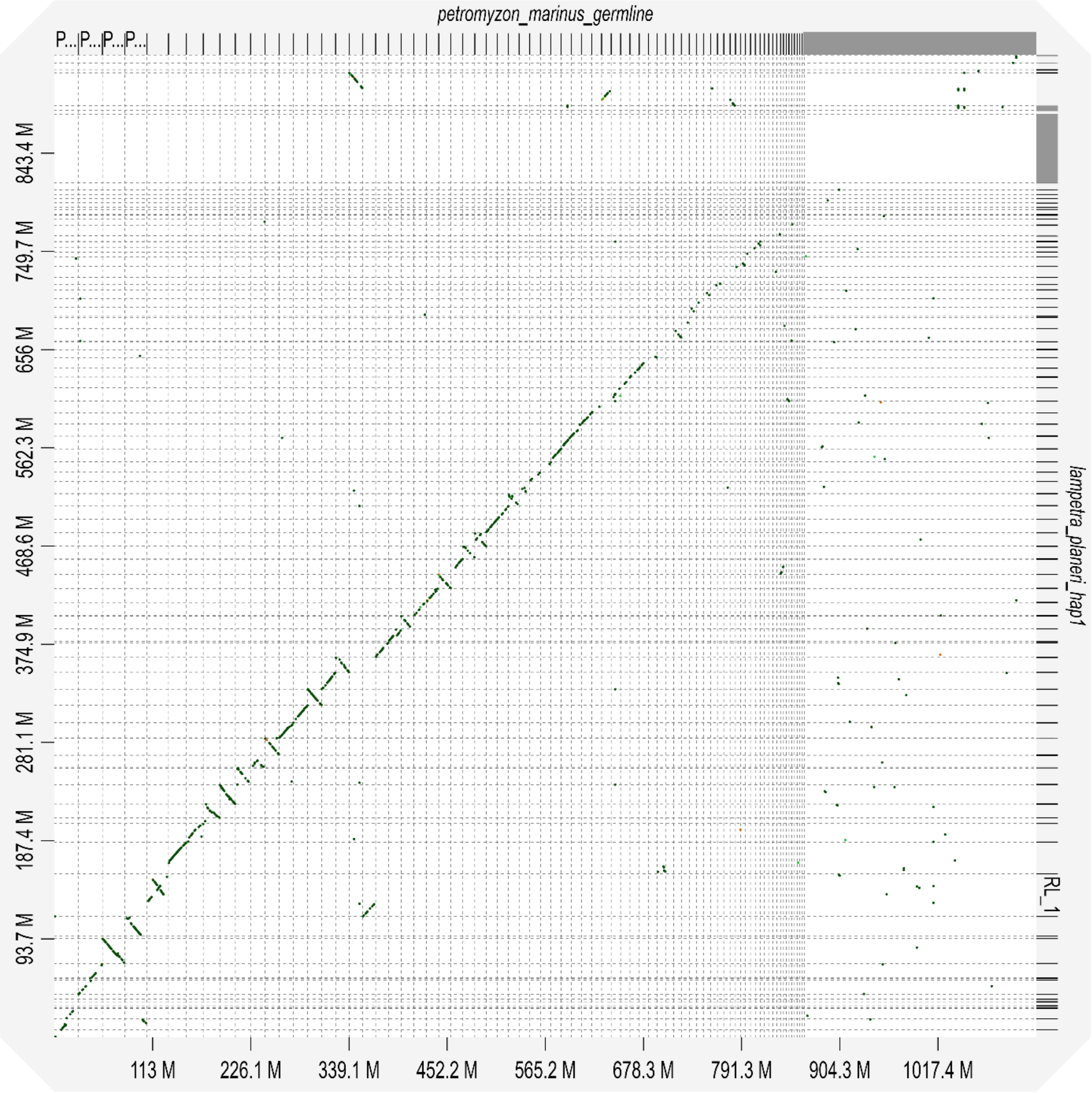


Figure 19B: Plot showing the synteny between PMG (x-axis) and BL1 (y-axis).

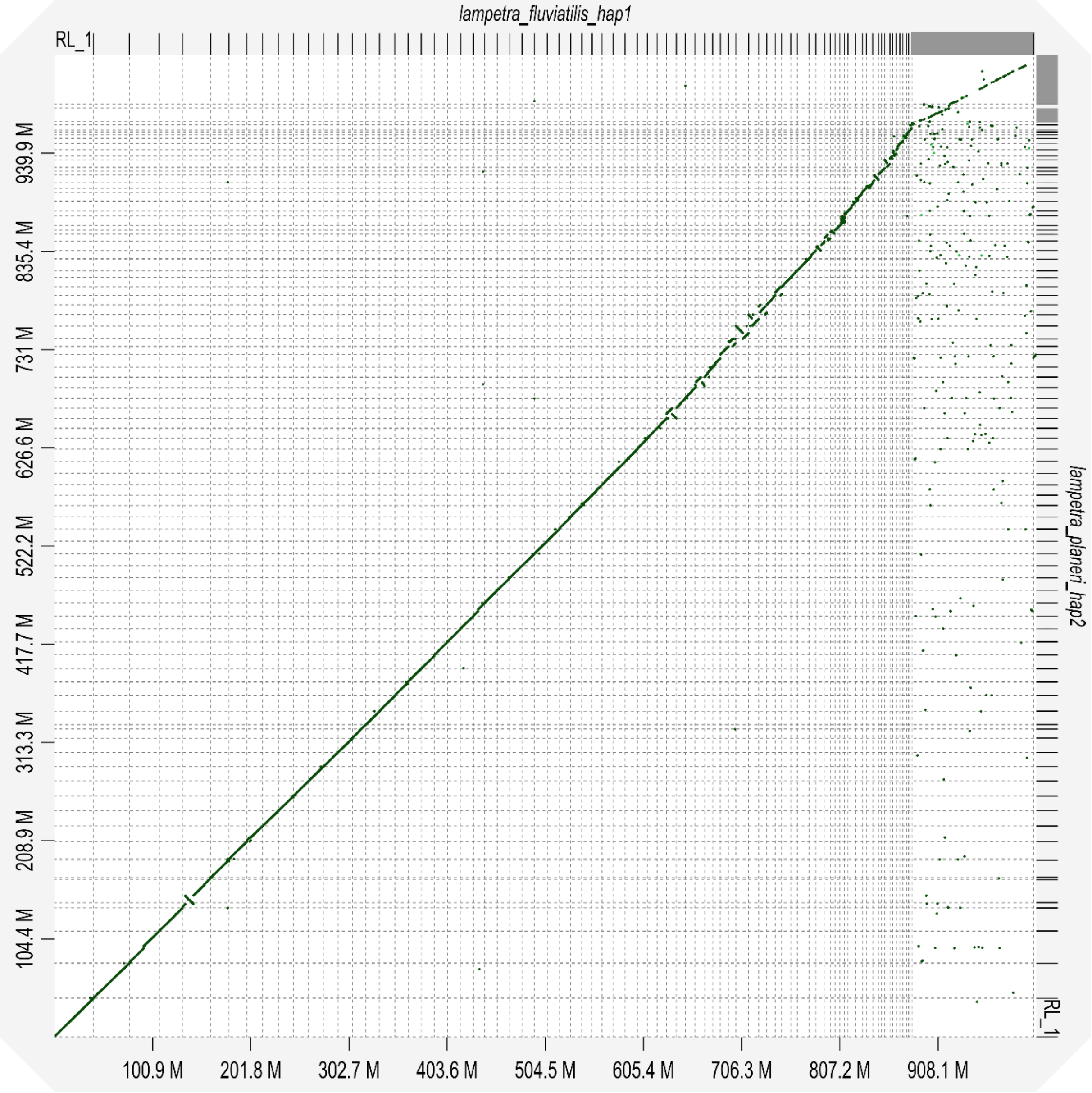


Figure 20B: Plot showing the synteny between RL1 (x-axis) and BL2 (y-axis).

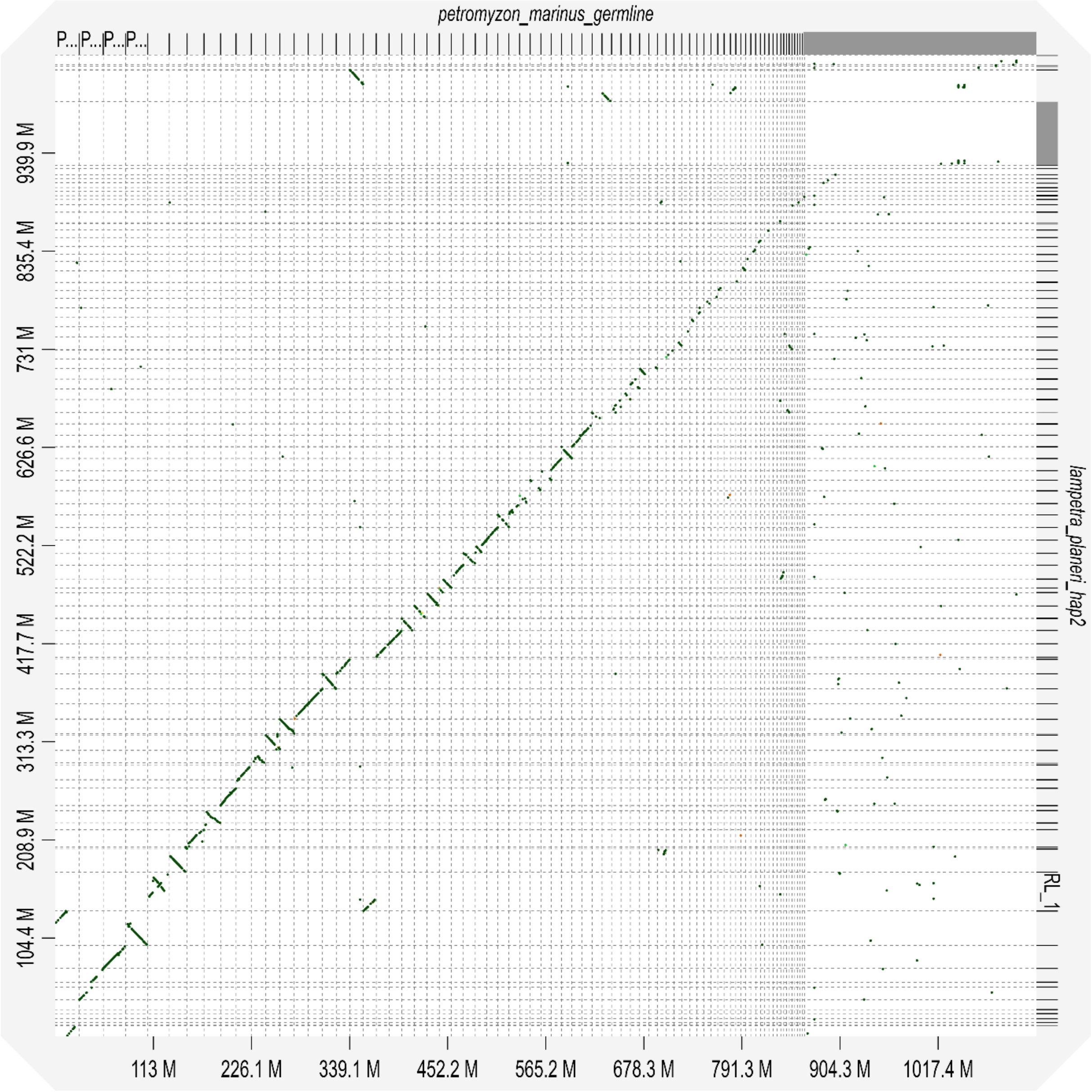


Figure 21B: Plot showing the synteny between PMG (x-axis) and BL2 (y-axis).

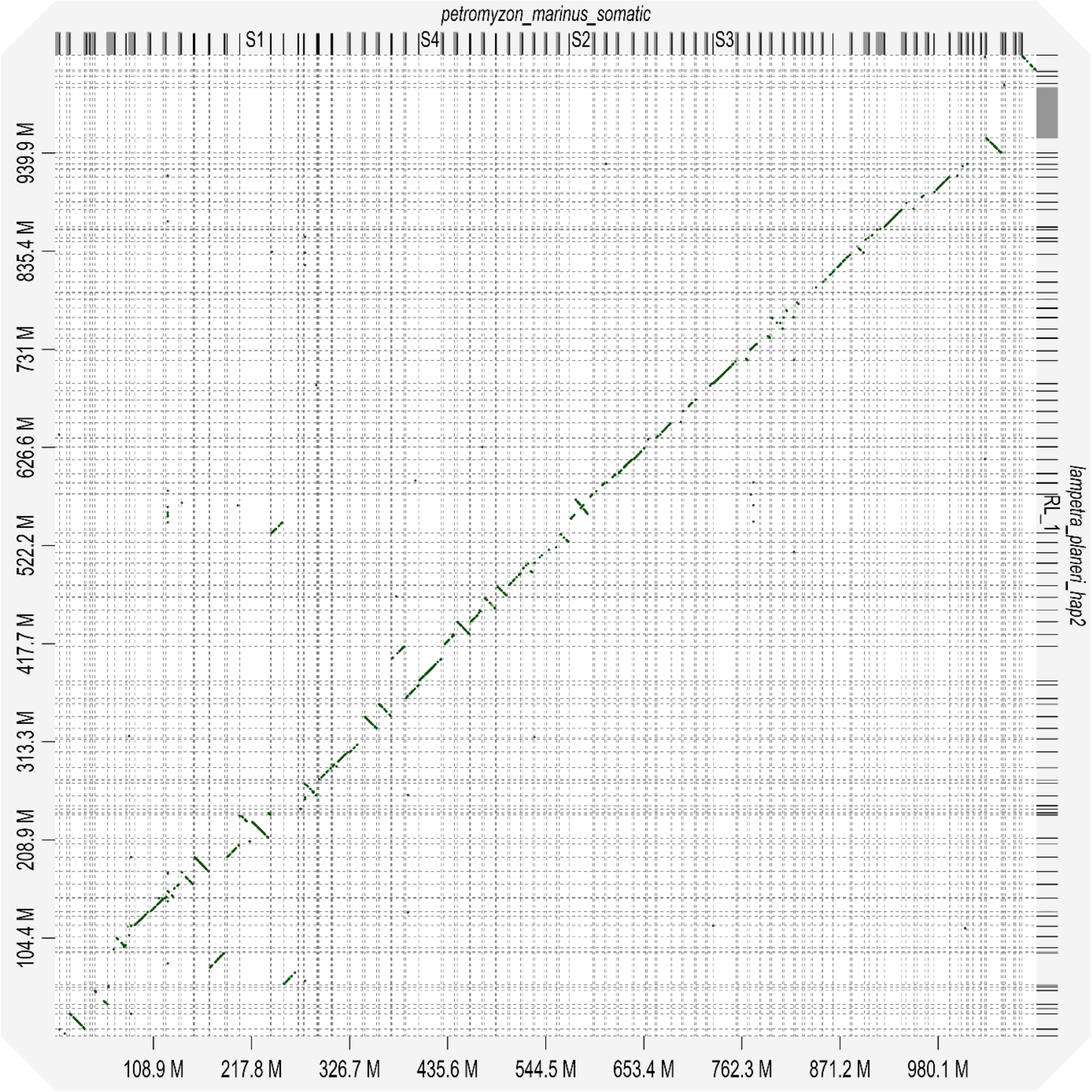


Figure 22B: Plot showing the synteny between PMS (x-axis) and BL2 (y-axis).

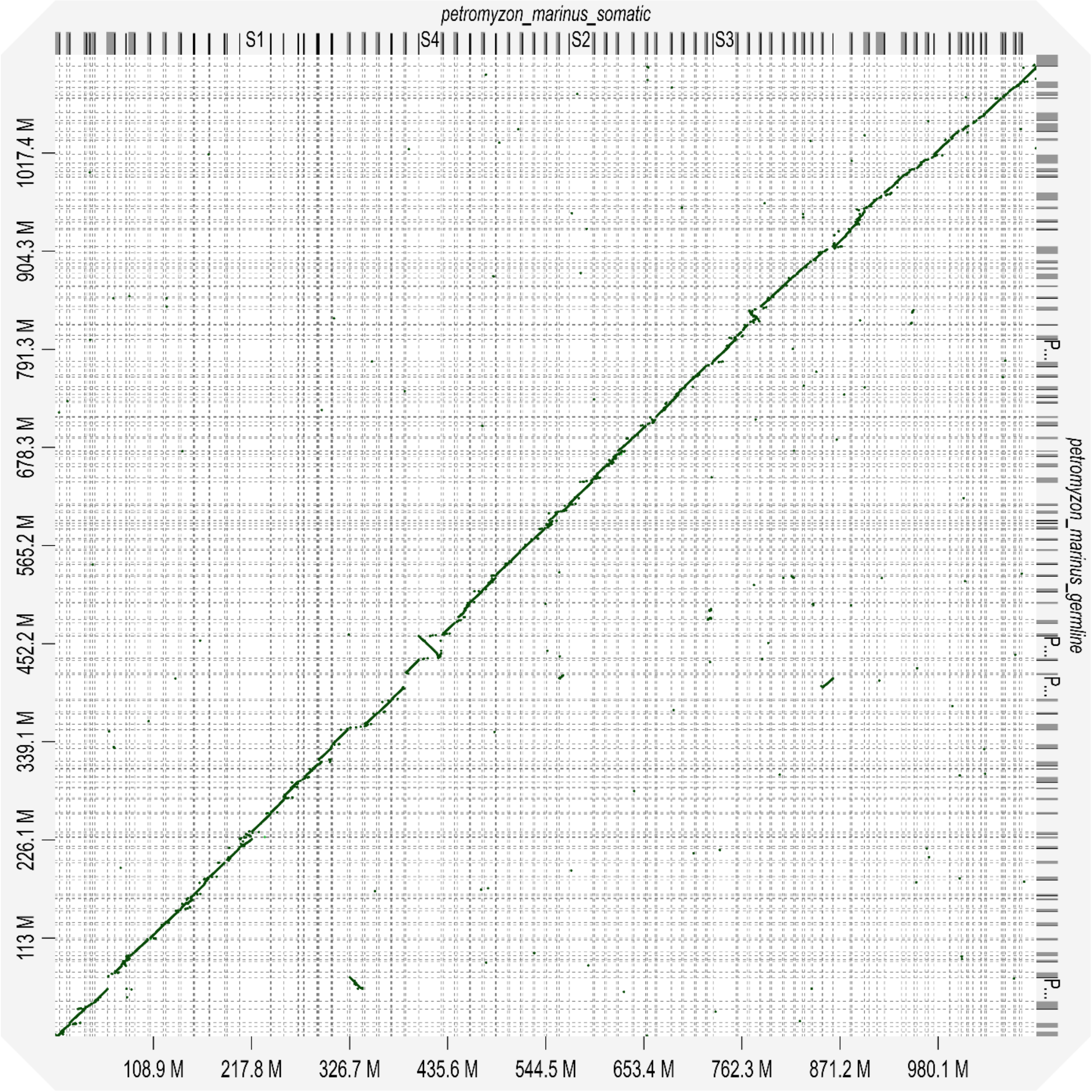


Figure 23B: Plot showing the synteny between PMS (x-axis) and PMG (y-axis).

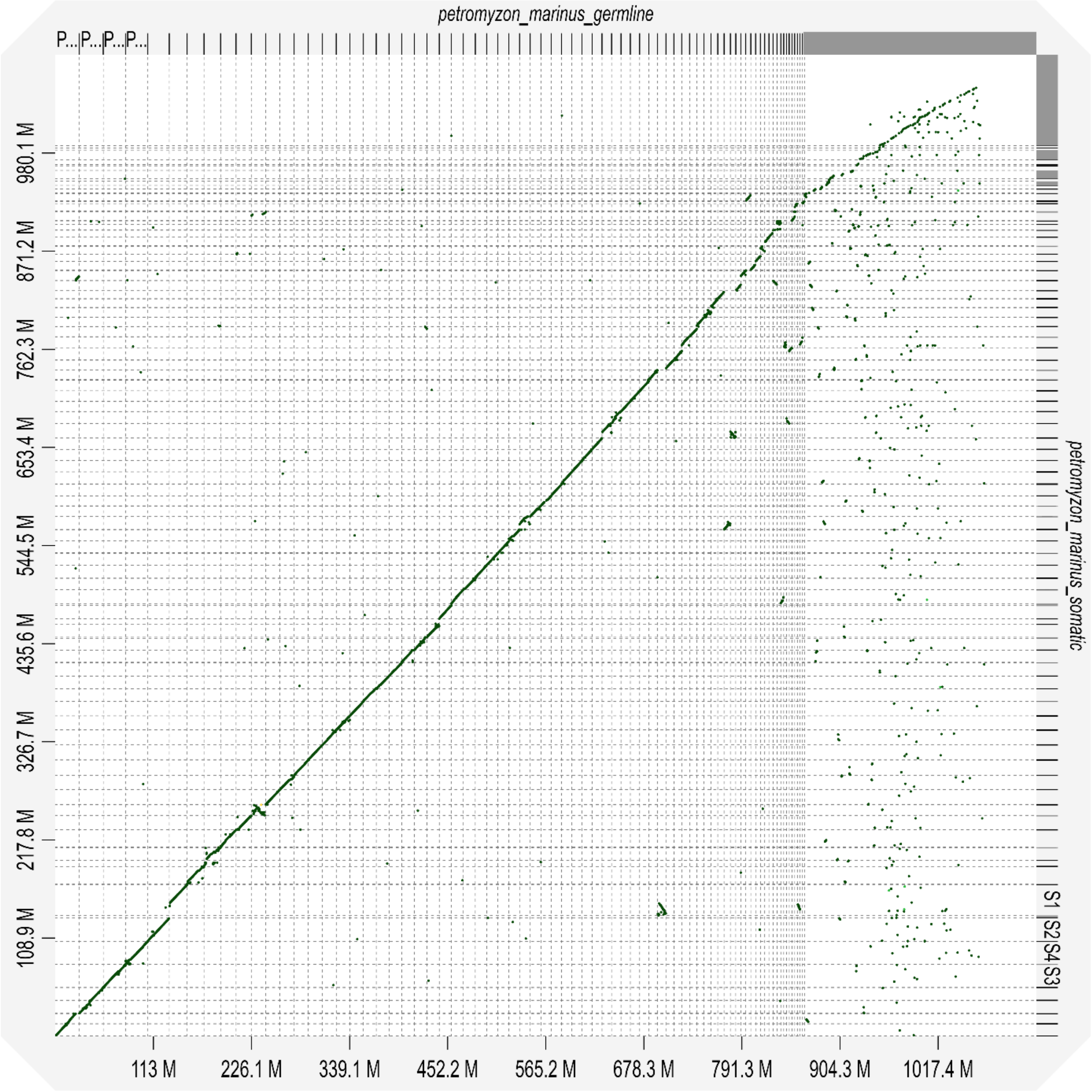


Figure 24B: Plot showing the synteny between PMS (x-axis) and PMG (y-axis).

9.2.4 Assemblytics statistics

Table 1B: The number of structural variations between all river lamprey and brook lamprey assemblies, partitioned into number of insertions, deletions, repeat expansions, repeat contractions, tandem expansions, and tandem contractions. Here, the manually curated reference FASTAs are compared to the query FASTAs, which consists of the assembled contigs generated using the hifiasm-Hi-C-integrated assembler.

	RL1 – reference			**RL2 – reference**			**BL1 – reference**			**BL2 – reference**		
Query sequence	**RL2**	**BL1**	**BL2**	**RL1**	**BL1**	**BL2**	**RL1**	**RL2**	**BL2**	**RL1**	**RL2**	**BL1**
Insertions	6546	7408	8194	6527	7480	8314	7716	7770	5323	8476	8362	5202
Deletions	5662	6409	7157	5652	6391	7057	6675	6636	4568	7393	7394	4681
Repeat expansions	5701	5951	7568	5809	5999	7598	6438	6351	4538	7775	7728	4367
Repeat contractions	6382	6764	8188	6402	6834	8366	7036	7070	4872	8717	8637	5015
Tandem expansions	2055	2245	2525	2013	2308	2576	2325	2341	1675	2596	2683	1788
Tandem contractions	1177	1259	1335	1150	1227	1365	1365	1398	1046	1421	1452	981

9.2.5 PretextView Snapshots

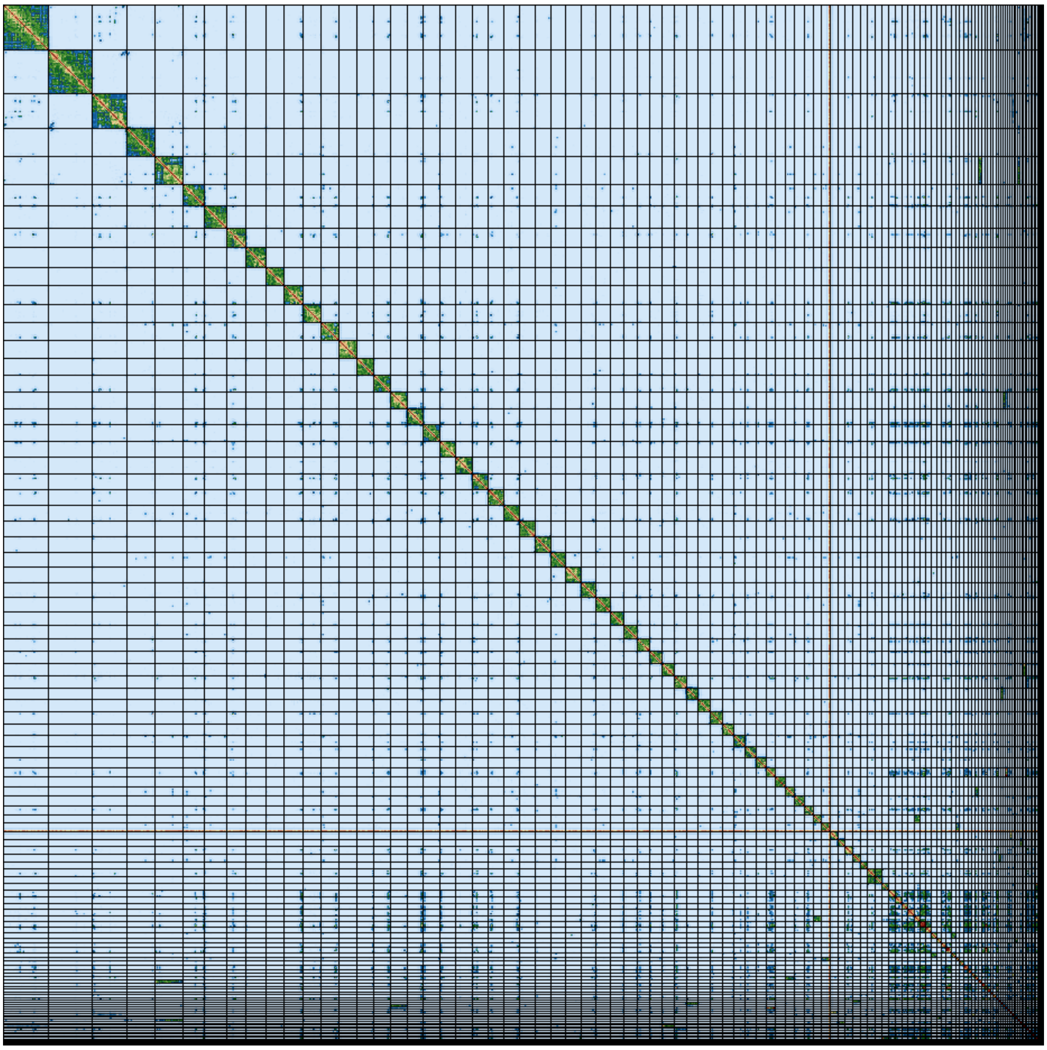


Figure 25B: RL1 before manual curation.

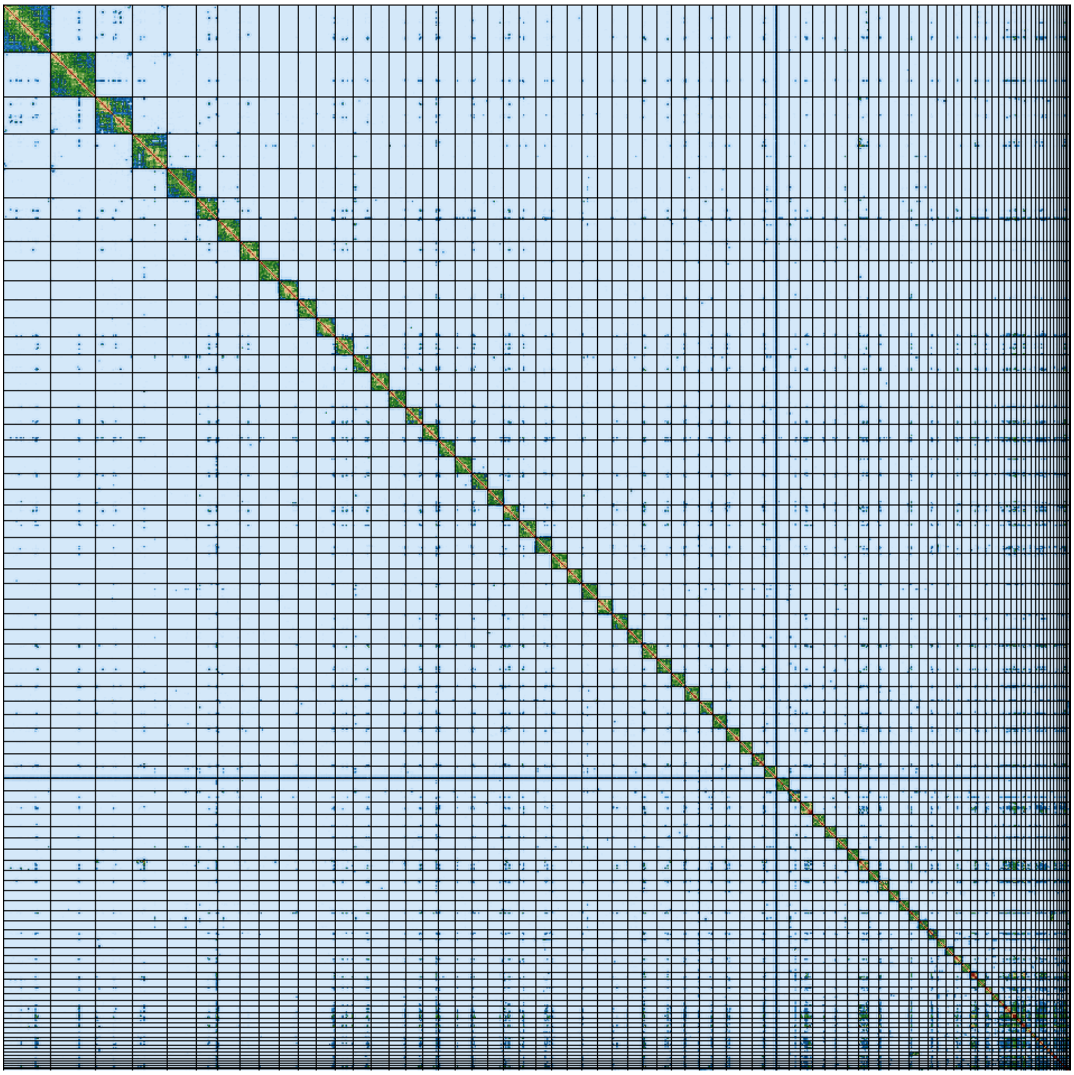


Figure 26B: RL1 after manual curation.

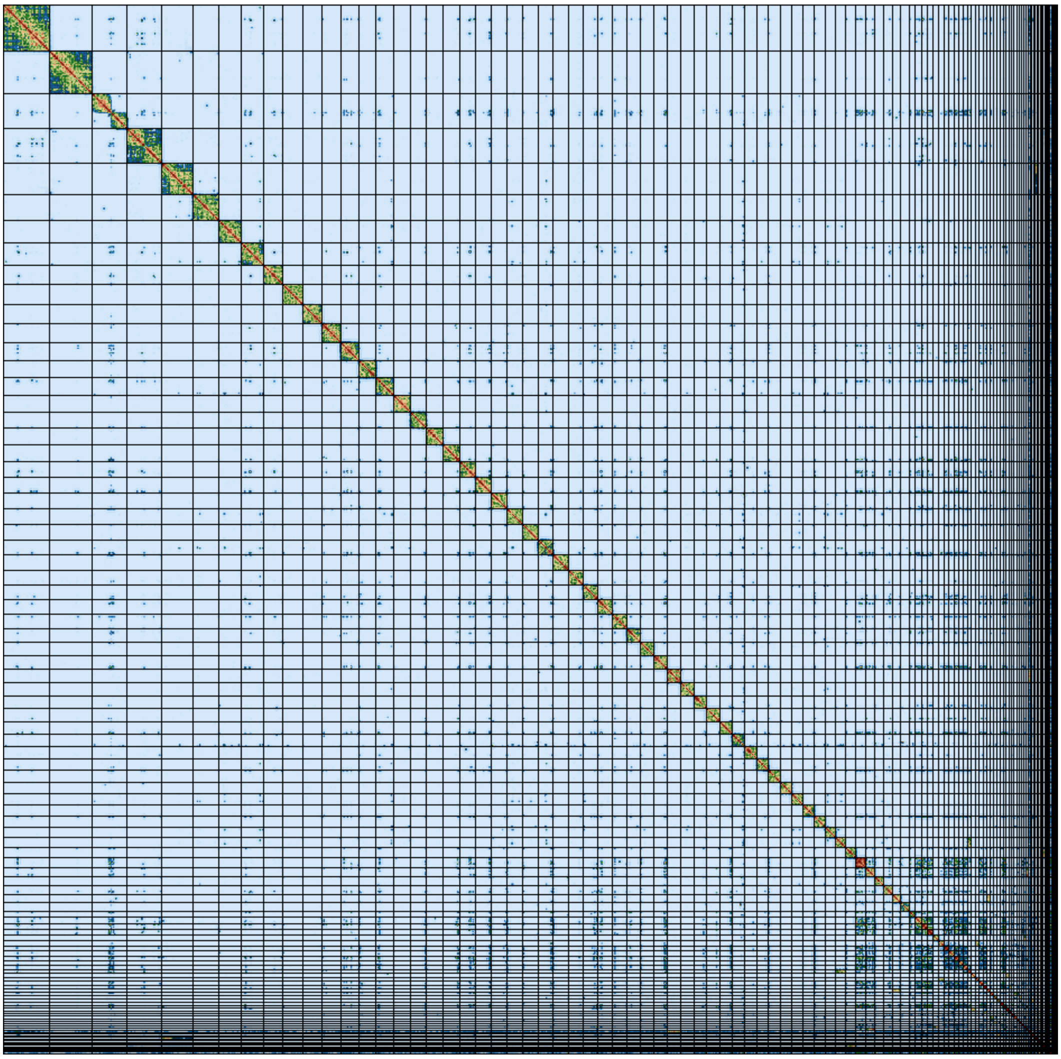


Figure 27B: RL2 before manual curation.

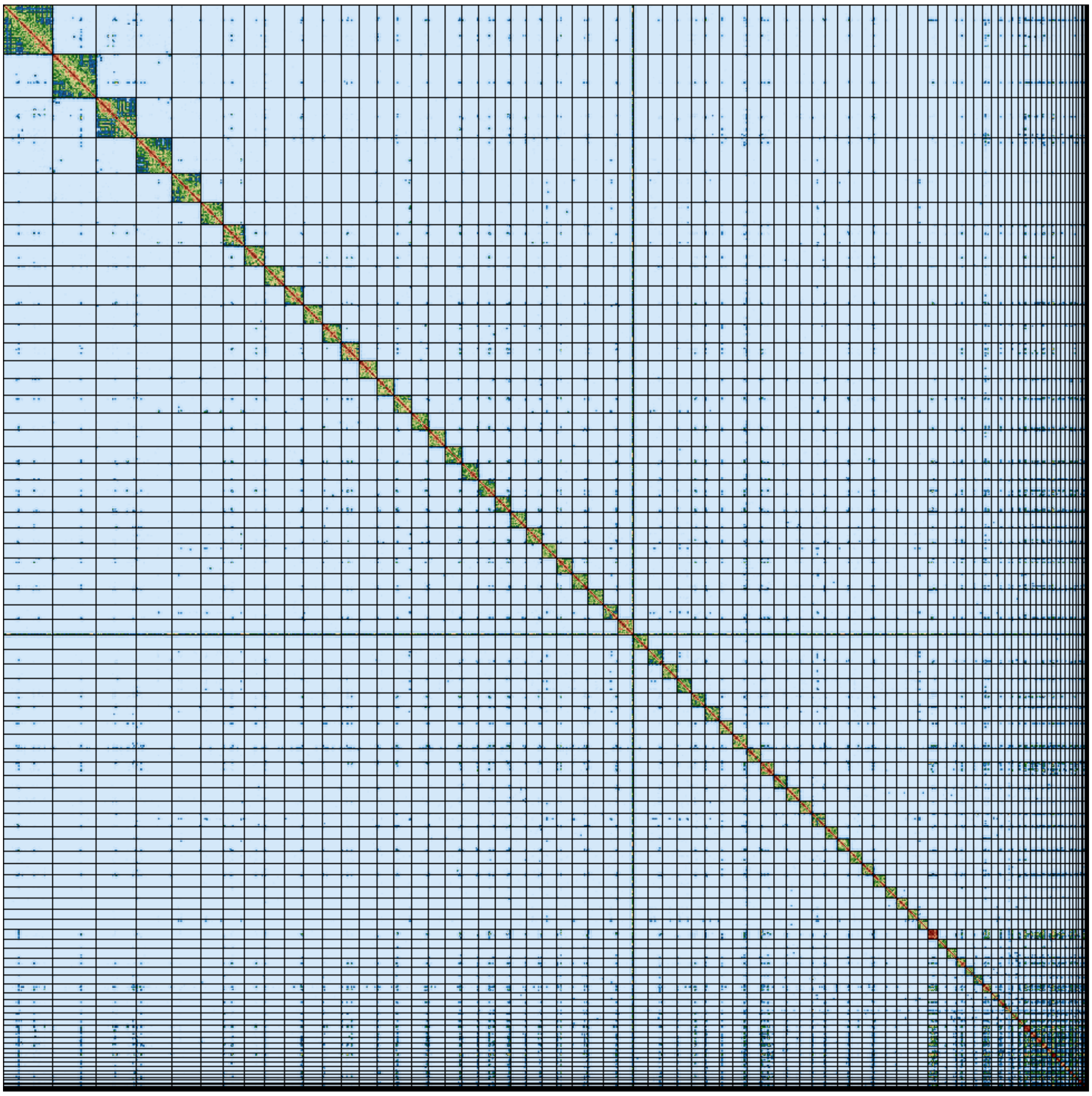


Figure 28B: RL2 after manual curation.

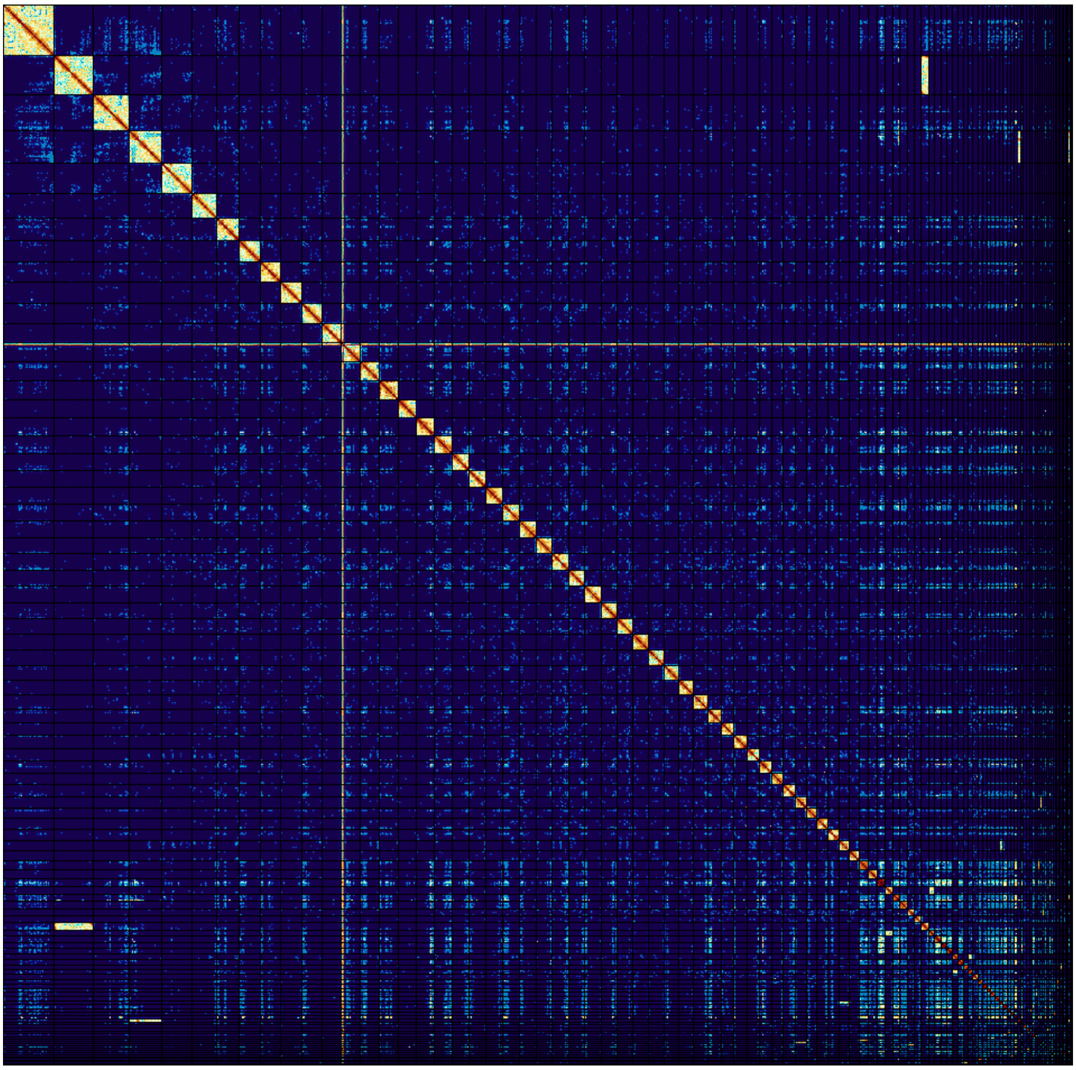


Figure 29B: BL1 before manual curation.

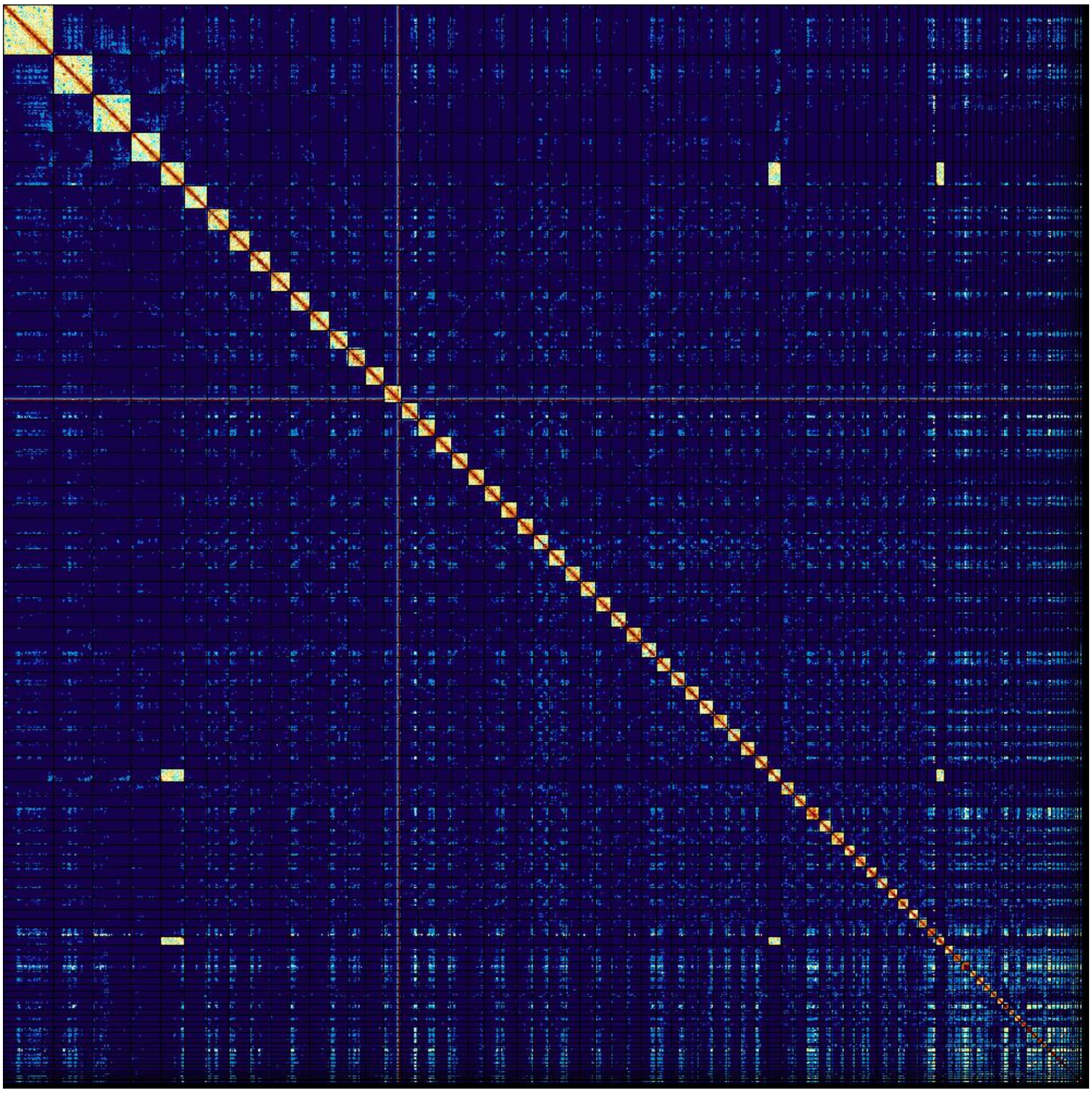


Figure 30B: BL1 after manual curation.

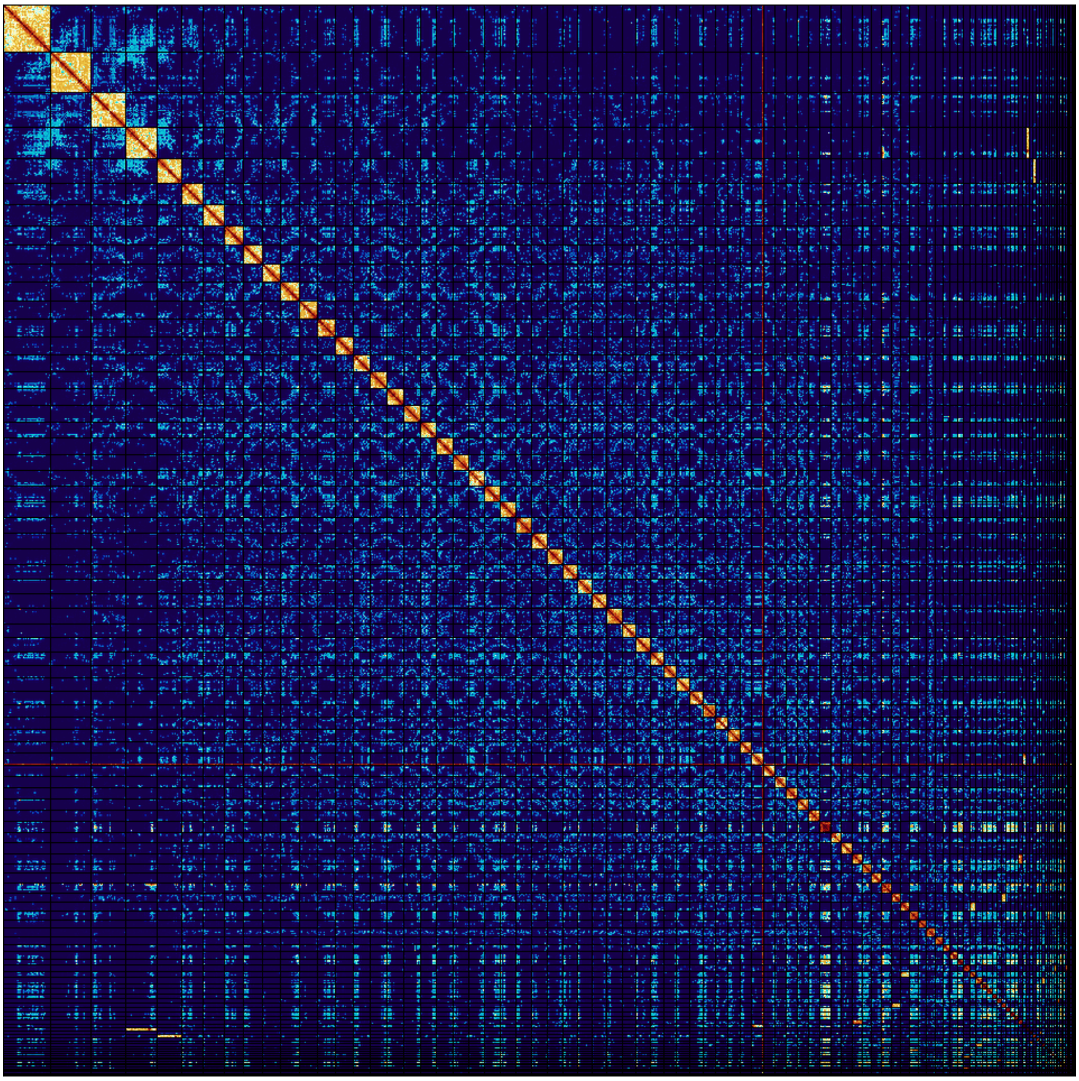


Figure 31B: BL2 before manual curation.

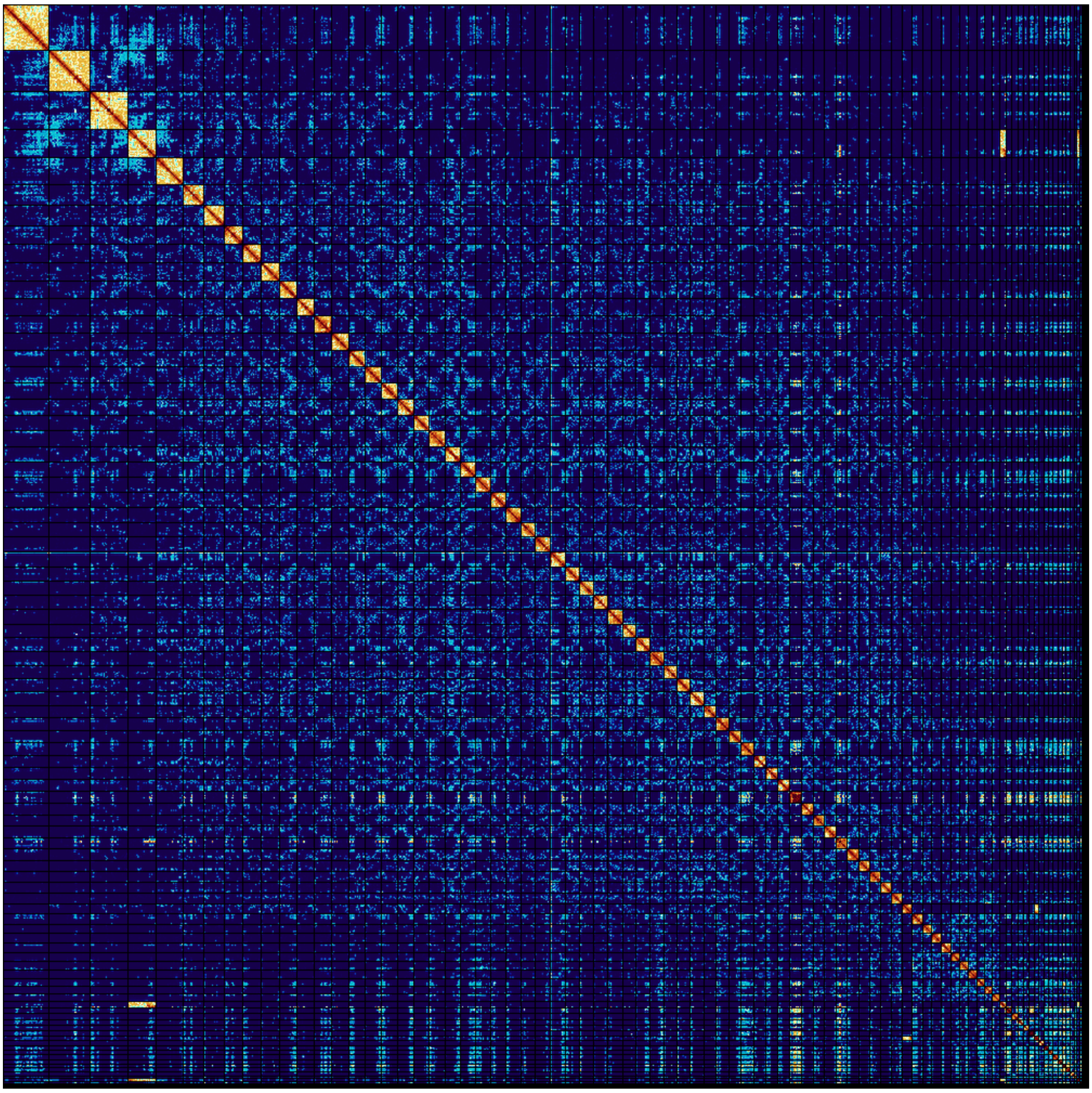


Figure 32B: BL2 after manual curation.

9.2.6 Mitochondrial species tree generated using IQ-TREE

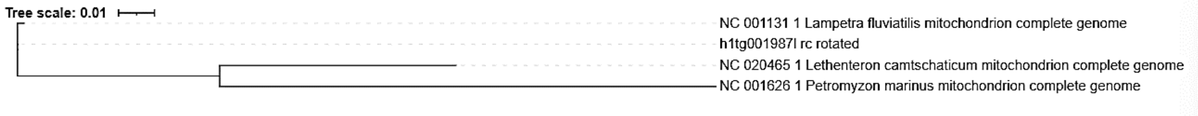

Tree scale: 0.01
NC 001131 1 Lampetra fluviatilis mitochondrion complete genome
h1tg001987l rc rotated
NC 020465 1 Lethenteron camtschaticum mitochondrion complete genome
NC 001626 1 Petromyzon marinus mitochondrion complete genome

Figure 34B: Species tree generated using the mitochondrial alignments of the brook lamprey (named rc rotated), river lamprey, Arctic lamprey, and sea lamprey. Here we see that the branches containing the river lamprey and brook lamprey mitochondria have almost no branch length.