UNIVERSITÀ POLITECNICA DELLE MARCHE
DIPARTIMENTO SCIENZE DELLA VITA E DELL'AMBIENTE

di Laurea Magistrale
Biologia Molecolare e applicata – Curriculum Biotecnologie

ASSEMBLAGGIO DEL GENOMA DI UNA FARFALLA ENDEMICA DELLE ISOLE PONTINE, *HIPPARCHIA SBORDONII* E STIMA DI ALCUNE STATISTICHE DI DIVERSITÀ NELLA POPOLAZIONE

GENOME ASSEMBLY OF A BUTTERFLY ENDEMIC TO THE PONTINE ISLANDS, *HIPPARCHIA SBORDONII* AND ESTIMATION OF SOME POPULATION DIVERSITY STATISTICS

Tesi di Laurea Magistrale
di:
Sebastiano Fava

Relatore :
Chiar.mo Prof.
Emiliano Trucchi

Sessione straordinaria
Anno Accademico 2020/2021

Assemblaggio del genoma di una farfalla endemica delle Isole Pontine, *Hipparchia sbordonii* e stima di alcune statistiche di diversità genetica nella popolazione

Riassunto esteso della tesi di laurea di Sebastiano Fava

1. INTRODUZIONE

La biodiversità in tutto il mondo è minacciata su molti fronti, tali minacce dovute alla distruzione e/o alterazione di habitat da parte dell'uomo, ad una velocità tale da impedire alle specie che li popolano di adattarsi a questi cambiamenti. Questi rapidi ed estremi cambiamenti globali stanno portando un numero sempre maggiore di specie all'estinzione, con la conseguente riduzione della biodiversità negli habitat nei quali venendo mancare l'equilibrio ecologico a loro volta collassano facendoci avvicinare alla sesta estinzione di massa. Per evitare questo destino funesto sono state adottate delle misure per salvaguardare la biodiversità. Questi approcci principalmente si basano sull'istituzione di regolamentazioni che limitano il consumo di habitat, risorse e la produzione di gas serra.

Tuttavia vi è la necessità di definire la pressione a cui è sottoposta una specie in pericolo di estinzione andando a effettuare delle stime di genetica di popolazione. Tali stime, per avere una maggior accuratezza e veridicità si devono effettuare avendo a disposizione il genoma di riferimento della specie in studio. Produrre un genoma di riferimento rappresenta una sfida economica e bioinformatica non indifferente, per superare tali sfide sempre più consorzi convergono per collaborare alla produzione di dati genomici. Questi consorzi collaborano con laboratori privati, progetti universitari e startup per produrre genomi di alta qualità. Uno di questi progetti universitari su cui si affidano grandi consorzi come l'Earth Biogenome Project (EBP) o anche il Vertebrate Genome Project è ENDEMIXIT. ENDEMIXIT ha come obiettivo quello di realizzare i genomi di riferimento di cinque specie endemiche a rischio di estinzione, per poterne determinare lo stato di "salute genomica" delle popolazioni ed eventualmente adottare delle misure di salvaguardia. Il mio contributo a questo progetto è stato la produzione dell'assemblaggio del genoma di *Hipparchia sbordonii*, che è uno dei passi fondamentali per le future analisi genomiche a livello di popolazione.

2. OBIETTIVI E SCOPI

Lo scopo principale di questo lavoro è quello di produrre un assemblaggio del genoma di riferimento di *H. sbordonii* e di definire una pipeline di software che ci permetterà di ottenere un assemblaggio del genoma di riferimento con l'alta qualità stabilita dagli standard internazionali.

Una volta ottenuto il genoma assemblato, un ulteriore scopo è quello di effettuare delle analisi di diversità genetica e della demografia passata nonché della connettività tra le due popolazioni di *H. sbordonii* e *H. semele* per le quali abbiamo dati di ri-sequenziamento.

3. MATERIALI E METODI

Per la produzione del genoma assemblato abbiamo impiegato diversi software bioinformatici che impiegano diversi dati di input con lo scopo di effettuare l'assemblaggio, incrementare la qualità, ottenere delle stime che diano indicazioni di quale passaggio effettuare per avere un maggiore incremento di qualità e come determinare i vari parametri su cui si basano gli algoritmi bioinformatici.

Di seguito vi è una lista dei software impiegati nell'assemblaggio ed il corrispettivo passaggio:

1. Stima delle dimensioni del genoma (Jellyfish + Genomescope)
2. Assemblaggio (Canu)
3. Valutazione delle qualità e della completezza (Meryl/Merqury, Busco)
4. Valutazione della contaminazione e qualità (Blobtools)
5. Polishing con l'utilizzo di long reads PacBio (Arrow)
6. Polishing con l'utilizzo di short reads Illumina (Polca)
7. Purging per la rimozione dei duplicati (Purge_dups)
8. Scaffolding (LRscaff)

9. Gap filling (TGS-Gap Closer)
10. Final Polishig con l'utilizzo di short reads Illumina (Polca)
11. Final purging per la rimozione di duplicati inseriti nelle precedenti fasi (Purge_dups)

Ad ognuno dei passaggi siamo andati a valutare le statistiche di continuità (N50, L50, numero di contig ecc), le statistiche di qualità e completezza ed il numero di geni predetti e le varie categorie in cui ricadono per poter determinare i passaggi successivi che avremmo dovuto applicare per incrementare la qualità dell'assemblaggio.

Mentre nell'impiego del genoma assemblato per analisi di genetica di popolazione abbiamo effettuato:

1. Variant Calling (Freebayes)
2. Filtraggio per qualità e separazione MNPs (Vcflib tools)
3. Ulteriore filtraggio e rimozione SNP in regioni ripetute (VCFtools)
4. Stima della diversità genetica inter- e intra-popolazione (VCFtools)

4. RISULTATI

4.1 Assemblaggio genomico

In seguito all'assemblaggio con CANU avevamo un genoma altamente frammentato, con un elevato numero di possibili errori di assemblaggio e un elevato numero di geni duplicati.

Andando ad analizzare se vi fosse un qualche tipo di contaminazione esterna non abbiamo riscontrato nessun organismo contaminante ed abbiamo ottenuto una rappresentazione grafica dell'alta frammentazione dell'assemblato.

Dopo i vari passaggi di polishing e purging abbiamo assistito ad incremento notevole nella qualità, nella completezza e nella continuità del genoma con valori di QV di 41 (il che ci indica che abbiamo circa 1 errore ogni circa 10000 basi), completezza del 98.5 e valori di N50 di 403 Kb, mentre i duplicati erano stati eliminati.

In seguito allo Scaffolding e Gap filling che hanno incrementato ulteriormente la qualità in termini di continuità con valori di N50 superiori 9 Mb ma in compenso abbiamo avuto un decremento della qualità, poiché nella sequenza erano stati inseriti numerosi Gap.

Negli ultimi step di polishing e purging abbiamo ripristinato la qualità dell'assembly raggiungendo valori di completezza del 98.6% un QV di circa 40 ed un N50 di circa 9.1 Mb raggiungendo gli alti standard di qualità imposti dal consorzio dell'EBP.

4.2 Stime di genetica di popolazione

Andando a misurare la diversità nucleotidica di *H. sbordonii* e *H. semele* abbiamo visto che entrambe le specie presentano una discreta diversità intrapopolazione (attesa nei Lepidotteri) con *H. sbordonii* caratterizzata da valori di π medio (diversità nucleotidica media lungo il genoma) di 0.0048 mentre *H. semele* ha un valore di 0.00783 suggerendo la minore dimensione della nostra specie endemica.

Andando poi a misurare il Tajima's D abbiamo notato che entrambe le popolazioni non sono stazionarie e che stanno subendo: nel caso *H. sbordoni* un restringimento del numero di individui deducibile da un Tajima's D>0 mentre *H. semele* che ha un Tajima's D<0 sta subendo un'espansione in seguito ad una riduzione del numero degli individui.

L'ultima stima rappresentata dall'F_{ST} o indice di fissazione, ha un valore di 0.190 indicando che le due specie: *H. semele* e *H. sbordonii* sono abbastanza simili tra loro geneticamente, suggerendo un certo grado di flusso genico tra le due popolazioni.

5. DISCUSSIONE E CONCLUSIONI

In seguito all'aumento del numero delle pubblicazioni e deposizioni di genomi di riferimento in database internazionali è stato necessario proporre e definire degli standard di qualità ai quali è necessario adeguarsi per poter depositare la sequenza del genoma di riferimento. Dalle precedenti stime delle statistiche di qualità del genoma confrontate con gli standard promossi dai maggiori consorzi siamo riusciti ad ottenere un genoma dalle ottime qualità. Tuttavia nonostante le statistiche di qualità siano buone vi è la possibilità di incrementarla ulteriormente applicando approcci che si basano sulla comparazione di genomi in scala cromosomica appartenenti a specie affini ad *H. sbordonii.*

Nell'ambito della genomica di popolazione, anche se abbiamo delle stime che ci danno un'idea sulla struttura delle due popolazioni delle due specie di *H. sbordonii* e *H. semele* vi è la necessità di effettuare altre analisi per avere un quadro completo della caratteristiche della popolazione in modo da poter adottare una strategia di conservazione il più adeguata possibile.

SUMMARY

ABSTRACT

In conservation biology, an increasing number of projects are taking advantage of the information available in whole genomes of endangered or threatened species in order to understand, by means of various analyses, their population characteristics and health status, their distribution in the habitats, including local adaptation, and how the intervention of biotic and abiotic elements affects their population size.

The newly emerging field of conservation genomics needs, however, the data to carry out such analyses, that is the reference genomes of the species to be studied.The aim of our work and study was therefore to produce a genome assembly of an endemic Italian species at risk of extinction, the butterfly *Hipparchia Sbordoni,* and then to analyze a set of population-level genomic data to estimate molecular summary statistics describing population diversity and structure of the species under study.

The genome assembly and the study of population characteristics was carried out using state-of-the-art bioinformatics tools according to the gold standard set by international genome assembly consortia.

1. INTRODUCTION

1.1 - Biodiversity threats

The impact of human activity on our planet is threatening biodiversity across habitats. Probably one of the greatest threats to biodiversity across the planet is the destruction of habitats (Haddad et al. 2015). As the human population increases, we need to change the landscape to meet our growing need for resources to support the modern lifestyle. Coupled with this is an increase in energy consumption that is driving climate change worldwide. The rapid pace of climate change will outstrip the natural ability of some species to respond (Hoffmann et al. 2011). The temporal analysis of biodiversity loss indicates that we are on a trajectory for the sixth mass extinction of the Earth biodiversity (Barnosky et al. 2011), with the rate of extinction in the last century which has been conservatively estimated to be 22 times faster than the baseline historical rate (Ceballos et al 2015). The picture is even bleaker if instead of looking at the complete loss of a species we look at population decline, with 32% of known vertebrate species showing substantial population decline (Ceballos et al 2017).

Efforts to stop mass extinctions and declining populations include: creating protected areas such as marine protected areas where fishing is prohibited or highly regulated; creating international agreements to limit the production of greenhouse gasses to slow climate change such as the Kyoto Protocol and the Paris Agreement; and establishing legal frameworks to protect endangered species, for example, the Convention on International Trade in Endangered Species of Wild Fauna and Flora and the US Endangered Species Act (Supple et al 2018). Genomic technologies can help these efforts by identifying biodiversity 'hotspots' to be prioritized for protection, using predictive systems and models to help build natural communities that are resilient to environmental change, and guiding management actions that seek to mitigate threats to endangered species.

For example, Italy is a biodiversity hotspot, but several endemic species, representing a unique biological heritage, are endangered. The main threats are linked to human activities: over-exploitation, persecution, habitat modification and reduction. The risks of extinction can be reduced by improving knowledge about genetic variability and developing conservation strategies to prevent its erosion. Genetic variability is essential to allow adaptation to new environmental conditions. In addition, small populations

(consisting of a small number of individuals) are subject to radical changes in their genetic diversity, as natural selection is less efficient and genetic drift becomes the main factor in determining the fate of naturally occurring mutations (Masel 2011). This process is random and can lead to the accumulation of deleterious mutations, i.e. the so-called genetic load, which affects the fitness of the individual and the population, leading to a further reduction in population size (Lynch et al. 1995). The population then enters the 'extinction vortex' which can end in its demise. Understanding the dynamics of this process at the genomic level can help in the implementation of strategies to reduce the risk of extinction.

The possibility of sequencing DNA at reduced costs, the enormous development of DNA sequencing techniques (such as Next Generation Sequencing or third generation sequencing), the implementation of new bioinformatic resources and statistical methods in recent years allow us to study the complete genomes of any non-model species. The genomes of different individuals can be screened to predict the deleterious effects of different types of mutation and, therefore, to estimate the genetic load of individuals and populations and predict its impact on fitness (Bertorelle et al. 2022) . Therefore, genomic analyses, which were previously limited to model

organisms, can now be used to estimate recent demographic events, genetic variation and population structure in endangered species through population genomic approaches.

1.2 - Conservation Genomics

One of the main objectives of conservation genomics is to implement targeted and effective strategies to safeguard endangered species, seeking to maintain genetic variability within a species and its constituent populations (Stange et al. 2021). Ensuring that populations (and therefore species) are as genetically diverse as possible is crucial, because it makes them more resilient and resistant to stress (Barrett et al. 2007), whether external or also due to evolutionary phenomena that manifest themselves with greater intensity in small populations such as gene drift or inbreeding depression. In addition, greater genetic and therefore adaptive variability is essential, although not always sufficient, to cope with recent and intense environmental changes due to anthropogenic activities and to adapt effectively to new environmental conditions. These changes include global warming, habitat fragmentation, urbanization, the introduction of alien and invasive species, the presence of

contaminants, and overhunting and overfishing. Another objective of conservation genomics is to study whether and how species evolve in response to new stresses (and the temporal dynamics of these processes). In both cases (maintaining genetic diversity and assessing the response to external stimuli), the power of genomic data makes it possible to deal with the problem more accurately.

The evaluation and quantification of the effects of neutral and adaptive evolutionary processes (genetic drift, inbreeding, hybridisation, outbreeding, migration, etc.) and thus the decision to adopt correct strategies for the conservation of populations (and thus species), can be carried out using conservation genetics, which, however, due to the limitations of using a small number genetic markers, presents limitations that do not allow us to answer all the questions necessary to adopt a suitable conservation strategy (Ouborg et al. 2010).

The most direct contribution of genomics to conservation is to greatly increase the precision and accuracy of estimating parameters that require neutral loci (e.g. effective population size (N_e) and migration rate (m)) by genotyping hundreds to thousands of neutral loci in numerous individuals.

There are problems or questions that cannot be solved accurately using genetics, but can be answered using genomics.

Primary problem	Possible genomic solution
Estimation of N_e, m and s	Increasing the number of markers, reconstructing pedigrees and using haplotype information will provide greater power to estimate and monitor N_e and m, as well as to identify migrants, estimate the direction of migration and estimate s for individual loci within a population
Reducing the amount of admixture in hybrid populations	Genome scanning of many markers will help to identify individuals with greater amounts of admixture so that they can be removed from the breeding pool
Identification of units of conservation: species, evolutionarily significant units and management units	The incorporation of adaptive genes and gene expression will augment our understanding of conservation units based on neutral genes. The use of individual-based landscape genetics will help to identify boundaries between conservation units more precisely
Minimizing adaptation to captivity	Numerous markers throughout the genome could be monitored to detect whether populations are becoming adapted to captivity
Predicting harmful effects of inbreeding depression	Understanding the genetic basis of inbreeding depression will facilitate the prediction of the effectiveness of purging. Genotyping of individuals at loci associated with inbreeding depression will allow the selection of individuals as founders or mates in captive populations. Pedigree reconstruction will allow more powerful tests of inbreeding depression
Predicting the intensity of outbreeding depression	Understanding the divergence of populations at adaptive genes will help to predict effects on fitness when these genes are combined. Detecting chromosomal rearrangements will help to predict outbreeding depression
Predicting the viability of local populations	Incorporating genotypes that affect vital rates and the genetic architecture of inbreeding depression will improve population viability models
Predicting the ability of populations to adapt to climate change and other anthropogenic challenges	Understanding adaptive genetic variation will help to predict the response to a rapidly changing environment or to harvesting by humans and allow the selection of individuals for assisted migration

Table 1**: Primary genetic problems in conservation and how genomics can contribute to their solution** (Allendorf et al. 2010)

1.3 - Conservation genomics and reference genomes

Advances in genome sequencing and reduced sequencing costs now allow the generation of genomic data based on the whole genome, which enable us to:

- understand the dynamics of deleterious mutation accumulation in small populations and its impact on individual fitness and extinction risks
- estimate the genomic susceptibility to extinction due to mutation load, predict the consequences of a genetic rescue strategy and propose conservation actions;

These genomic data are based on reference genomes. “Reference genomes provide a view of the architecture of the genome, comprising both genic and intergenic regions and serve as a genomic resource as they provide a comprehensive and fundamental framework against which genomic variation can be mapped and quantified, to characterize and ultimately help preserve genetic diversity” (Formenti et al. 2021).

The creation of a reference genome is based on several steps:

- collecting the sample that will be used for sequencing
- sequencing using systems such as NGS (next generation sequencing) or third generation systems that produce long reads

- De-novo assembly of reads produced by sequencing
- sequencing and transcriptome assembly
- genome annotation
- SNPs, genotypes and structural variants calling
- upload resources in public Database

1.4 - Genomic conservation projects

The creation of reference genomes is a challenge from an economic and bioinformatics point of view. Until recently, reference genomes were available for a small number of model organisms. However, thanks to recent international efforts on conservation genomics-based initiatives to save species, the trend has changed.

Thanks to the collaboration of numerous consortia and genome projects, the reduction of sequencing costs, the increasing quality of sequencing technology, combined with improved bioinformatics algorithms and huge advances in computing power, the creation of reference genomes across the entire spectrum of biodiversity has been facilitated in an impressive manner (Formenti et al. 2021). In recent years, numerous national and international

consortia and genomic projects have been set up with the aim of sequencing and creating high-quality reference genomes for species spanning the phylogenetic tree of life, including: the Earth Biogenome Project (EBP), the Vertebrate Genomes Project (VGP) and the Global Invertebrate Genomics Alliance (GIGA) to name a few. These projects and consortia, having to manage a huge amount of work, ranging from sampling, sequencing, assembly and annotation, have created a network of sub-projects, communities and laboratories that work together with the main project to produce reference genomes.

One of the most interesting and ambitious projects with the largest number of sub-projects is the Earth Biogenome Project (EBP), which aims to sequence and annotate the approximately 1.5 million known eukaryotic species in three phases. With Phase I it seeks to create "annotated chromosome-scale reference assemblies for at least one representative species from each of ~9,000 eukaryotic taxonomic families" (Lewin et al. 2018). In Phase II, the EBP will seek to sequence a representative species for each eukaryotic genus, currently estimated at around 150,000 taxa, including the 9,000 families from Phase I. In this Phase through comparative genomics approaches, the reference genomes produced in Phase I will be used as the scaffold on the

assembly drafts to give a reasonable approximation of the order and orientation of the scaffolds on chromosomes for Phase II genomes. Phase III will involve sequencing and assembly to obtain the reference genome of the remaining ~1.35 million eukaryotic species, diversity sequencing for endangered species, plus the remaining new species identified in bio-observatories (~100,000 total) (Lewin et al. 2018).

The Earth Biogenome Project, as mentioned earlier, faces enormous economic, computational and logistical challenges, and one of the main challenges is the development of a global strategy for the collection of voucher specimens that are preserved adequately to enable production of high-quality genome assemblies. For these reasons, in order to succeed in its aims, it is essential that the EBP involves institutions, laboratories and various projects whose mission is to find, study and conserve the world's biodiversity. Over the years, the EBP has formed a large network of partnerships and affiliated projects to address the challenges presented by the project. Some of these affiliated projects focus on the production of genomic resources from specific taxa such as 1,000 fungal genomes (1KFG), 10,000 bird genomes (B10K), 10,000 plant genomes (10KP), 5,000 insect genomes (i5K) to name a few. Other projects focus more on geographical regions, such as Africa

BioGenome Project (AfricaBP), California Conservation Genomics Project (CCGP), Darwin Tree of Life, BRIDGE Colombia and ENDEMIXIT.

1.5 - ENDEMIXIT project

In particular, the latter: ENDEMIXIT full name "Genomic susceptibility to extinction: a whole-genome approach to study and protect endangered Italian endemics" aims to study the actual genomic health of the small populations of five Italian endemic species: *Podarcis raffonei* (the Aeolian or Raffone's wall lizard), *Bombina pachypus* (italian endemic Apennine yellow-bellied toad, once considered a subspecies of *Bombina variegata*), *Ursus arctos marsicanus* (the Marsican or Apennine bear), *Acipenser naccarii* (the Adriatic sturgeon) and *Hipparchia sbordonii* (Ponza butterfly). All five of these species are "Endangered" o "Critically Endangered" by the IUCN (International Union for Conservation of Nature) (https://www.iucn.org). To understand the status and dynamics of their genomic diversity, the first aim of the project is to assemble de novo genomes for each endemic species and then to resequence ten individuals each from a small and a larger population of each species to allow for comparative analyses

My contribution to this project was the production of the *Hipparchia sbordonii* genome assembly, which is one of the fundamental steps for the downstream population-level genomic analyses.

1.6 - Hipparchia sbordonii

Hipparchia sbordonii, also known as the "Ponza butterfly", is a lepidopteran species belonging to the Nymphalidae family (subfamily Satyrinae). It is an Italian endemic species confined to the Ponziane Islands, where there are no other similar species. It has a very restricted distribution range encompassing few isolated populations, so the IUCN Red List of Italian Butterflies considers this species as "endangered".

Its discovery dates back to the 1960s. Almost twenty years passed from the discovery to the formal description of the species by Otakar Kudrna. Past studies on allozyme-based clustering analysis have shown little genetic differentiation between *H. sbordonii* and *H. semele*, the most closely related species on the Italian mainland coast nearby. Nevertheless, differences in morphology, such as wing pattern and shape, suggested that *H. sbordonii* should be retained as a valid species (Cesarono et al. 1994; Valerio et al.

2018). The wingspan is 5/6 cm. The adult butterfly's wings are coloured in various shades of light and dark brown, up to orange and yellow, enabling it to blend in with its surroundings (http://www.farfalleitalia.it/sito/910/index.php). This butterfly shows sexual dimorphism, in fact on the front wing there are also two well-developed ocelli, larger in the female than in the male; the lower page of the hind wings is marbled so as to make the individual very cryptic when resting on a tree trunk or on the ground. The species also has other traits, apart from the size of the ocelli, which are subject to sexual dimorphism: the two sexes can be easily distinguished because the female is usually slightly larger and has more extensive yellow-orange spots on the forewings (Kudrna et al. 1984).

Hipparchia sbordonii has a very restricted areal, in fact it covers an area actually occupied by less than 500 km2 as it lives only on the Ponziane Islands where there are few isolated populations (https://www.iucnredlist.org/species/173231/64640021). At the time of its discovery it was observed on all the islands of the archipelago (Ponza, Palmarola, Gavi, Ventotene, Santo Stefano and Zannone). It has recently experienced a marked demographic decline and there are no recent data on its presence on islands other than Ponza and Palmarola.

The populations of *Hipparchia sbordonii* have suffered a strong demographic decline in recent decades. It is a species considered "endangered" by the IUCN Red List due to the strong pressure of tourism, illegal harvesting, excessive urbanization and improper land management with the implementation of new agricultural practices in spite of the traditional ones that favoured the survival of *H. sbordonii* populations (Bonelli et al. 2018) (https://www.iucnredlist.org/species/173231/64640021). The species has also undergone large fluctuations in numbers, probably due to the consequences of fires. In addition, reduced hunting activity and poaching of sparrows has led to an increase in the number of birds such as *Muscicapa striata*, an insectivorous bird specialised in preying on insects in flight, which prey on *Hipparchia sbordonii*, leading to a decline in numbers (Sbordoni 2018).

Figure 1: **male specimen of Hipparchia sbordonii seen from the reverse side** (http://www.farfalleitalia.it/sito/910/index.php)

2. AIM AND OBJECTIVES

The main aim of this work is to produce an assemblage of the reference genome of *H. sbordonii* that will later be used by other researchers to make population genetics estimations.

A more technical aim of this project is to define a software pipeline that will allow us to obtain a reference genome assembly with the high quality set by international standards.

We can summarize the objectives of this work in four main points:

1. Obtain an accurate estimate of the size of the genome of *H. sbordonii* so that this can be used in the assembly and subsequent quality statistics.

2. Assemble the *H. sbordonii* genome with assembly software using PacBio sequencing data, i.e. long reads, in order to obtain a high-quality assembly product

3. Increase the quality of the assembly by carrying out various bioinformatics steps, using PacBio and short-read Illumina sequencing data, in order to refine our genome assembly to the required quality that allows us to deposit the genome in the international databases (EBP standards).

4. Use the assembled genome to begin with the analyses of population genetic diversity, past population demography, and connectivity between the two populations of *H. sbordonii* and *H. semele* for which we have re-sequencing data.

3. MATERIALS AND METHODS

Genome sequencing has become an integral part of modern molecular biology. The majority of the available analysis methods, however, are designed for established model organisms with chromosome-level reference genomes and detailed annotation readily available (Ranallo-Benavidez et al. 2020). Genome assembly has an extra layer of complexity when the basic genomic features of the species are unknown (e.g., genome size, heterozygosity, and even ploidy).

3.1 - Genome size estimation

In this initial phase, the genome size was estimated, GenomeScope was used (Vurture et al. 2017). GenomeScope is a tool that allows you to quickly estimate the characteristics of the genome under study, such as total and haploid genome length, percentage of repetitive content and heterozygosity, as well as overall sequenced read characteristics: read coverage, read duplication and error rate. In our case we are interested in estimating the size

of the genome which is important in downstream assembly steps to decide how many reads to correct and how sensitive the overlapping step should be.

The estimates do not require a reference genome and they can be automatically inferred via a statistical analysis of the k-mer profile of sequenced Illumina short reads (Chor et al. 2009). The k-mer profile (sometimes called k-mer spectrum) measures how often k-mers, substrings of length k, occur in the sequencing reads and can be computed using different tools.

For our needs the tool used to compute the k-mer profile was Jellyfish (Marçais et al. 2011). In order to produce a k-mer spectrum, Jellyfish needs to count all k-mer of length K in all the short reads (Illumina reads). To count the k-mer present in all the reads we used the Jellyfish count command, and we set 31 as the value of the k-mer length obtaining a binary file .jf .

Once we got a .jf file, we used it to run the Jellyfish histo command which calculates the histogram with the number of k-mer having a given count.

The Jellyfish histo command creates a .histo file that contains a histogram with the occurrence of the various k-mer. Once we obtained the .histo file,

which represents the k-mer profile, we used it as the input of Genomescope web tool, obtaining :

- K-mer graph, that estimates coverage depth of raw DNA reads for a genome using the number of times a K-mer is observed (coverage) by number of K-mers with that coverage (frequency).
- Statistics regarding: the max and min genome haploid length, the percentage of heterozygosity, the genome repeat length, the genome unique length and the read error rate.

The statistics obtained from these analyses were then used as a comparison with those obtained from subsequent analyzes and as parameters to optimize the software run. In particular, the main statistic we were looking for in the analyzes, with Jellyfish and later with Genomescope was the estimate of the size of the genome, which we then used as an additional parameter in the assembly of the genome.

3.1.1 - What is a Kmer?

In bioinformatics, the term K-mer is used in computational genomics and sequence analysis where it represents a substring of length K in a DNA base string.

“For example, all 2-mers of the sequence AATTGGCCG are AA, AT, TT, TG, GG, GC, CC, CG. Similarly, all 3-mers of the sequence AATTGGCCG are AAT, ATT, TTG, TGG, GGC, GCC, CCG. There is an exponentially increasing number of possible K-mers for an increasing number of K. There are 16 possible 2-mers for DNA if we assume that there are only 4 types of bases (A,T,G,C).

The equation for the number of possible K-mers for a given K is therefore 4^K” (https://ucdavis-bioinformatics-training.github.io/2020-Genome_Assembly_Workshop/kmers/kmers).

Bases	K-mer size	Total possible kmers
4	1	4
4	2	16
4	3	64
4	4	256
4	5	1,024
4	6	4,096
4	7	16,384
4	8	65,536
4	9	262,144
4	10	1,048,576
4	…	…
4	21	4.4e+12
4	27	1.8e+16
4	31	4.6e+18

Table 2: **possible number of K-mer combinations depending on the length of the K-mer**

As can be seen in the table above, there are 64 possibilities for a 3-mer and over 4 Trillion possibilities for a 21-mer. For a given sequence of length L, and a K-mer size of K, the total k-mer's possible will be given by (L – k) + 1

e.g. For the sequence of length of 14 , and a K-mer length of 8, the number of K-mer's generated will be:

GATCCTACTGATGC

n = (L - K) + 1 = (14 - 8) + 1 = 7

GATCCTAC, ATCCTACT, TCCTACTG, CCTACTGA, CTACTGAT, TACTGATG, ACTGATGC

For shorter fragments, as in the above example, the total number of K-mers estimated is n = 7, which is not close to the actual fragment size of L which is 14 bps. If we consider larger fragments, the total number of K-mer (N) provides a good approximation to the actual size of the genome. In fact, in our case we have used this system to estimate the size of the genome.

In fact, as can be seen in the following table, increasing the size of the genome decreases the percentage of error in the estimate of the genome.

Genome Sizes	**Total K-mers of k=18**	**% error in genome estimation**
L	N=(L-K)+1	((L-N)/L)*100
100	83	17
1000	983	1.7
10000	9983	0.17
100000	99983	0.017
1000000	999983	0.0017
Genome Sizes	**Total K-mers of k=31**	**% error in genome estimation**
360000000	359999970	0.0000083

Table 3: **percentage of error in genome estimation based on genome size:** as the size of the genome increases, the error rate decreases

So for a genome size of 360 Mb and K-mer size of 31, the error between estimation and reality is only $8.3*10^{-5}$%. Which is a very good approximation of actual size. In choosing a K-mer size, it should be large enough to allow the K-mer to map uniquely to the genome. So the total available K-mers should be sufficiently larger than the genome size and therefore has the ability to store all the K-mers in the genome. However, too large K-mers leads to a need for substantial computational resources, as well as producing more erroneous K-mers caused by sequencing errors. In other cases, large k-mers are used in extremely large genomes and/or in very repetitive genomes, as considering long k-mers increases the number of unique k-mers that help us solve these genomes with greater reliability (https://ucdavis-bioinformatics-training.github.io/2020-Genome_Assembly_Workshop/kmers/kmers) .

K-mer are exploited in various uses ranging from genome assembly (Compeau et al. 2017), to genome size estimation (Michal Hozza et al. 2015), predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein (Zhu et al. 2018), identify species in metagenomic samples (Perry and Beiko 2010) identification of biomarkers for diseases from samples (Wang et al. 2018) and many other applications.

3.2 - Genome Assembly

One of the most complex and computationally intensive tasks of genome sequence analysis is genome assembly. Long-read single-molecule sequencing has revolutionized de novo genome assembly and enabled the automated reconstruction of reference-quality genomes (Mihai et al. 2004). However, given the relatively high error rates of such technologies, efficient and accurate assembly of large repeats and closely related haplotypes remains challenging. To address these problems related to genome assembly we used Canu (Koren et al. 2017), a software that uses noisy single-molecule long reads.

Canu can be run on a single computer or multi-node compute cluster. In our case we used an approach that uses multiple CPUs in a single cluster node. A full Canu run includes three stages: correction, trimming, and assembly (Koren et al. 2017). In all stages, the first step constructs an indexed store of input sequences, generates a k-mer histogram, constructs an indexed store of all-vs-all overlaps, and collates summary statistics. The correction stage selects the best overlaps to use for correction, estimates corrected read lengths, and generates corrected reads. The trimming stage identifies unsupported regions in the input and trims or splits reads to their longest

supported range (Koren et al. 2017). The assembly stage makes a final pass to identify sequencing errors; constructs the best overlap graph; and outputs contigs and summary statistics.

From the input reads,which in our case are PacBio reads (Jhon 2009), the correction stage generates corrected reads; the trimming stage trims unsupported bases and detects hairpin adapters, chimeric sequences, and other anomalies; and the assembly stage constructs the contigs.

3.3 - Evaluation of quality and completeness statistics

At each step of the assembly, we carried out analyses to:

1. assess the statistics of the various fasta files produced, using software that calculates assembly statistics for FASTA files (https://github.com/b-brankovics/fasta_tools). In particular the estimates that interested us the most and had a greater informative character were:

- N50 (bp): Half of the genome sequence is covered by contigs larger than or equal to the N50 contig size and therefore the sum of the

lengths of all contigs of N50 size or larger contains at least 50% of the total genome sequence (Earl et al. 2011).

So the larger this value, the more contiguous the genome, indicating that the assembly is of high quality.

- L50 is defined as the smallest number of contigs whose sum of their lengths represents half the size of the genome.

  Therefore the smaller this index, the more contiguous the assembly and therefore the higher the quality

- With regard to N90 and L90, the estimate is the same, except that in these two cases, the length of the contig is taken into account, whereby the sum of the contigs of greater or equal length cover 90% of the genome and the smaller number of contigs, which when added together represent 90% of the genome respectively.
- The number of contigs gives us an idea of the fragmentation of the assembly because the more contigs there are, the more fragmented the genome is, so the objective of the various steps of the assembly was to reduce their number in response to the increase in contig length.

- The total size (bp) in this case gives us an indication of the length of the assembly.

2. Calculate the completeness and quality of the output produced through statistical analysis of the fasta file, with Meryl/Merqury (Rhie et al. 2020).
3. Measure completeness and the presence of duplicates or missing of BUSCO conserved genes (Seppey et al. 2019).

In this way it was possible, at each step, to decide which was the next step of the assembly in order to have the best quality of the final product.

3.4 - Evaluating assemblies

Genome assembly software combines the reads into larger regions called *contigs*. However, current sequencing technologies and software face many complications that impede reconstruction of full chromosomes, including errors in reads and large repeats in the genome (Rhie et al. 2020). After obtaining the various files produced by Canu we proceeded to evaluate the quality of the assembly using various tools, which carry out different analyzes using various inputs, giving us the results represented by plots and statistics

regarding the quality of the assembly or even if there is contamination of the sample used for sequencing.

3.5 - Evaluating assemblies with Meryl/Merqury

In this first step of evaluation of the genome assembly we used a tool that is based on the use of two software Meryl and Merqury that carry out analyzes on the assembled genome (Rhie et al. 2020). In particular Meryl, the first software to perform this analysis, is a tool for counting and working with sets of k-mers. Meryl counts the number of times a k-mer occurs in the short reads, similar to Jellyfish, and places them in a k-mer database.

Therefore, initially we have counted the k-mer of size 22 bp, present in the pair end reads illumina with the command Meryl count, obtaining 2 databases with the count of k-mer and later we have joined them obtaining a single database with the command Meryl union-sum. After obtaining the database with the k-mers count, we performed the actual analysis with Merqury using the database produced by Meryl (Rhie et al. 2020).

Merqury's analysis produces two plots: copy number spectrum (spectra-cn plot), assembly spectrum (spectra-asm plot) and two statistics: K-mer completeness and consensus quality estimate (QV).

The spectra-cn plot (copy number spectrum) represents the count of the canonical k-mers observed in the assembly and in the accurate, whole-genome read set. In the spectra-cn graph there are represented with different colors the spectra of the k-mer present in the reads set, according to the number of copies.

The spectra-asm plot similarly to the spectra-cn plot represents each k-mer in the read set by the assembly in which it is found; this becomes useful when two haploid assemblies are evaluated. This way, we can detect k-mers that are present only in one assembly, k-mers shared in both assemblies, and k-mers not present in the assembly and only found in the read set. As for statistics, Merqury calculates the k-mer completeness, which indicates the percentage of k-mer in the read set that are also found in the assembly.

In the end Merqury calculates the Consensus quality (QV) estimation which is an estimate of consensus errors in the assembly: in brief, Merqury estimates the probability P:

$P = (K_{shared} / K_{total})^{1/k}$

that a basis in the set is correct, then we derive the error probability E = (1 - P) and finally using this formula, the widely used Phred quality score (often denoted as QV) can be computed by treating the E as base error probability that a base in the assembly is correct the value of QV is given as base error probability:

$QV = -10 \log_{10} E$

3.6 - Evaluating assemblies with BUSCO

In this phase, we employed the software BUSCO: OrthoDB's (Ortholog Databases) sets of Benchmarking Universal Single-Copy Orthologs that quantify the completeness of genomic data sets in terms of the expected gene content based on evolutionary principles (Mathieu Seppey et al. 2019). BUSCO uses sequence profiles embedded in lineage-specific datasets, which in our case specified the Lepidoptera dataset, to assess the orthologous status of predicted genes in the species under analysis. These consensus sequences are derived from Hidden Markov Model (HMM) profiles (Eddy 2004) built

from multiple sequence alignments of orthologs selected from OrthoDB and capture the conserved alignable amino acids across the species set.

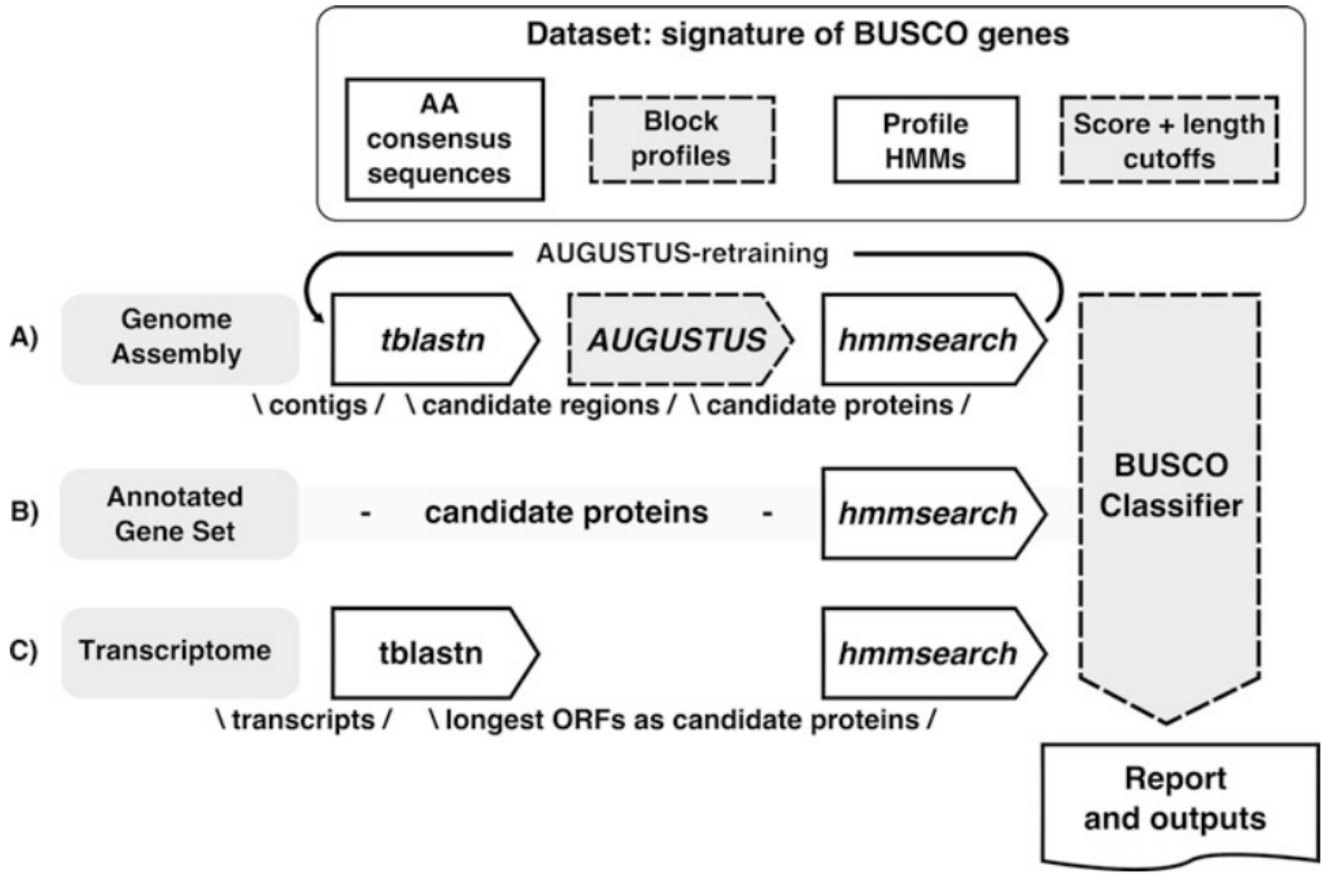

Figure 2: **Description of the BUSCO workflow for the three types of sequence input**, genome (a), gene set (b), and transcriptome (c). The same dataset is used in all modes, although not all information embedded is utilized in each situation. The genome mode includes two phases in which the three main steps are run, with the second pass only targeting the missing and fragmented BUSCO genes using additional consensus sequences and retrained AUGUSTUS parameters (Seppey et al. 2019)

Based on the input, BUSCO defines which steps need to be executed, for our needs we used as input Genome Assembly. In this way BUSCO predicts the possible sequences of genes in the genome assembly and gives us back several output folders that contain: sequences of predicted genes, missing genes, extracted genes and their coordinates in the genome or even results of the various steps carried out by Busco and in particular those that interest us in particular from the file short_summary.*.txt that contains a plain text summary of the results in BUSCO notation. This file contains a report in which there is the number of predicted genes and the percentages of the different categories in which the genes fall such as C:complete [S:single-copy, D:duplicated], F:fragmented, and M:missing.

3.7 - Quality control and taxonomic partitioning of genome datasets with Blobtools

In this phase we used the tool Blobtools: a multimodule software that allows us to evaluate the quality of the assembly and visualize possible contamination of the sample, sequenced and assembled, by other organisms different from the one under study (Laetsch et al 2017).

The first step to use Blobtools that we did was to build the BlobDB (Blob DataBase) which needs three input files:

1. Hits file [TSV]
2. Genome Assembly [FASTA file]
3. Mapping file [BAM file]

3.7.1 - Generation of the Hits file

A hits file is a TAB-separated-value (TSV) file which links sequence IDs in an assembly to NCBI TaxIDs, with a given score. These can be the results of sequence similarity searches of the assembly against a sequence database (e.g. BLASTn output files). The required format is TSV and is composed of three columns

- 1st column: sequenceID (must be part of the assembly)
- 2nd column: TaxID (a NCBI TaxID)
- 3rd column: score (a numerical score)

To generate the TSV file, we used the software BLAST: Basic Local Alignment Search Tool.

BLAST is a tool that performs a search for homologous sequences in nucleotides in a target database (Ladunga, I. et al. 2009), BLASTn searches for local alignments between the genome assembly and known sequences contained in the database.

In this way BLASTn, by letting us know the sequences that align with the assembly and the organisms to which these sequences belong, allows us to extrapolate information regarding the homologous sequences contained in the genome, and therefore if they have homology with organisms that could indicate contamination of the samples sequenced.

In our case we did an analysis with BLASTn specifying that the output should contain the following information: qseqid (Query Seq-id), staxids (Subject Taxonomy ID), Bit-score (indicates the sequence similarity), "6" (output format: tabular), in this way we got the Hist file needed by Blobtools to do the taxonomy assignment.

3.7.2 - Generation of the Mapping file

The Mapping file or Sequence Alignment Map (SAM) is a text-based format that contains alignment information of short reads mapped against reference sequences. The SAM file usually starts with a header section, followed by alignment information as tab separated lines for each read. This Mapping file analyzed by Blobtools is necessary because it contains information regarding the base/read coverage of each sequence in an assembly file.

The coverage information parsed by BlobTools in a Mapping file is:

- Base coverage
- Total number of reads and number of mapped reads
- Read coverage

Blobtools in order to use the mapping file needs that the file is in BAM format, which is the compressed binary version of a SAM file. First then we had to produce the SAM file, and then transform it into a BAM file.

3.7.2.1 - Production of SAM file

To get the SAM file we used the BWA tool (Burrows-Wheeler Alignment tool)(Li and Durbin 2009). First of all, in order to produce a SAM file, BWA needs to construct the index for the reference genome and this was done using the bwa index command which makes an Index database sequence. The index database is then used in the SAM file realization phase. Second, we used the BWA mem command, which searches for local or end-to-end alignments, between reads (in our case Illumina short reads) and contigs produced by the assembly, and once found, extends them.

As we have already said, the BWA mem command to produce the SAM file, uses Illumina reads, but the raw reads, at the ends, have the sequence of the adapters which are used to bind the DNA fragment to be sequenced to the flow cell where the amplification and sequencing takes place and act as primers for the amplification reaction first and then for the sequencing reaction, only that these adapters sequences are not useful for the construction of the SAM file but rather create problems in the phase of research of the alignments by BWA. Therefore, before launching the BWA mem command, we have done a trimming step of the Illumina reads.

3.7.2.2 - Trimming Illumina Reads

For this trimming step we have used the Trimmomatic software (Bolger et al. 2014), which uses the sequence of the adapters provided in the command, to recognize it and remove it from the Illumina Reads, it also performs a scan of the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15.

Following the trimming step, we evaluated the quality of the trimmed reads with the FastQC software, to assess whether the adapters had been removed from the reads sequence and whether the quality of the reads had increased.

Once the reads had been trimmed, we proceeded to build the SAM file with the BWA mem command.

3.7.2.3 - BAM file generation

Since Blobtools only supports BAM files (compressed binary version of a SAM file) as coverage input and we have the SAM file, we had to transform the SAM file into the input supported by Blobtools. We then used the SAMtools tool (Li et al. 2009) to obtain the BAM file.

As a first step to obtain a BAM file that can be used by Blobtools, we run the command SAMtools view which performs a conversion from SAM file to BAM file. However, this BAM file generated by the previous command is not usable by Blobtools because it needs a sorting and indexing step. We then run the SAMtools sort command which performs a sort alignments by leftmost coordinates and is used to streamline data processing and to avoid loading extra alignments into memory (Li et al. 2009). Finally, to obtain a BAM file that can be used by Blobtools, we run the SAMtool index command which indexes a coordinate-sorted BAM file.

3.7.3 - Construction of the BlobDB, visualization of assembly and generation of tabular output

After having obtained the input files (Hits file, Genome Assembly and Mapping file) needed by Blobtools to build the DataBase, we launched the Blobtools create command which does the parsing of input files and creates the BlobTools (JSON) data structure, i. e. BlobDB. After running the Blobtools Create command we get the JSON file which we then used as input to run the Blobtools view command which generates a tabular output for manual inspection and subsequent partitioning of sequences in the assembly (Laetsch et al.). The tabular file, generated by the Blobtools view command, contains information on the contig being analysed concerning: contig length, GC content, N content, coverage, phylum found in the contig (extrapolated from Hits file).

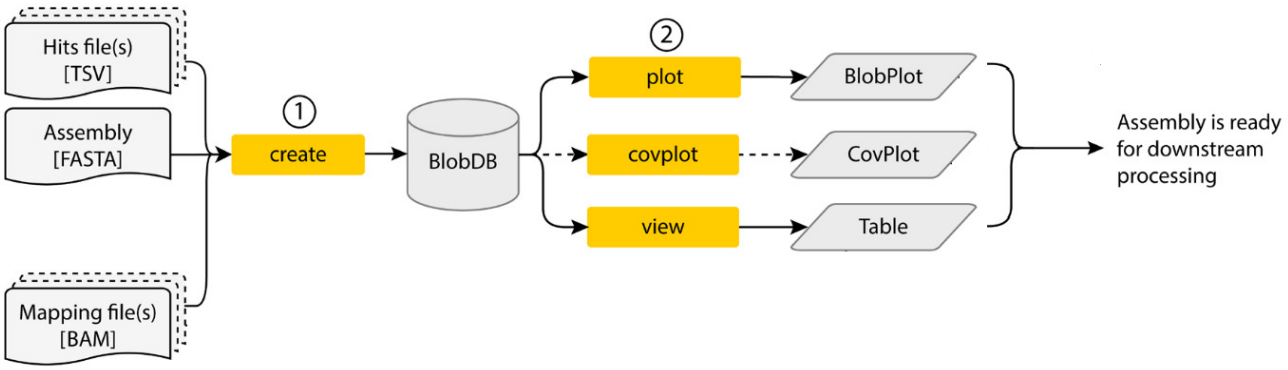

Figure 3:**BlobTools workflows for taxonomic interrogation.** Targeted at de novo genome assembly projects in the absence of a reference genome. 1: Creation of a BlobDB data structure based on input files. 2: Visualisation of assembly and generation of tabular output.(Laetsch et al 2017)

To have a graphical display of the tabular result allowing us to graphically evaluate the assembly, we have used the Blobtools plot command which produces a BlobPlot in which the sequences in the assembly are depicted as circles, with diameter scaled proportional to sequence length and coloured by taxonomic annotation based on BLASTn similarity search results provided in this order and using taxrule 'bestsumorder. Circles are positioned on the X-axis based on their GC proportion and on the Y-axis based on the sum of coverage across the library (Laetsch et al.). The command to create the BlobPlot also creates a second plot called ReadCovPlot in which the percentage of mapped and unmapped reads in the assembly is represented,

and the percentage of mapped reads according to the taxonomic group to which they belong.

3.8 - Polishing of the Genome Assembly

Following the initial analysis to assess the quality of the assembly, we carried out polishing steps to increase the quality. In this phase we used several software and different inputs, so that we could exploit the various polishing modes of the respective software according to the inputs.

3.8.1 - Polishing with GCpp (Arrow's algorithm)

In this phase, we used tools that are part of the software module produced by Pacific Bioscience called SMRT Analysis Software (https://github.com/PacificBiosciences/pbbioconda), which includes several packages to perform different analyses and operations on genomic files. In our case, we used the GenomicConsensus (GCpp) package (https://github.com/pacificbiosciences/genomicconsensus/) which is based on

an algorithm called Arrow to carry out polishing. To carry out polishing using Arrow's algorithm, we needed to have as input a sorted file of reference-aligned reads in Pacific Biosciences standard BAM format and a FASTA file, which in our case is the output produced by the Canu assembly.

So to obtain the sorted file of reference-aligned reads in BAM format we used the pbmm2 package (https://github.com/PacificBiosciences/pbmm2/) that is part of the SMRT Analysis Software module.

The pbmm2 package represents a version of minimap2 (Li 2018) that supports PacBio reads (long reads). Pbmm2 basically uses the long reads to align them to the reference genome to build a sorted BAM file.

In practice we ran the command pbmm2 align --sort providing as input: the reads PacBio.BAM, via a file of file name (fofn) whose content was a list of the reads' paths, and the reference genome represented by the output contig.fasta produced by Canu, resulting in the file pacbio.bam.

We then used the pacbio.bam file produced by pbmm2 and the genome assembly file (contig.fasta) produced by Canu for polishing with the

GenomicConsensus package (GCpp), so the Arrow algorithm will use the reads mapped to the genome to improve assembly quality.

3.8.2 - Polishing with Polca

Following polishing with GCpp, we carried out two rounds of polishing with POLCA (POLishing by Calling Alternatives) that is a polishing tool aimed at improving the consensus accuracy in genome assemblies produced from long high error sequencing data generated by PacBio SMRT or Oxford Nanopore sequencing technologies. POLCA utilizes Illumina or PacBio HIFI reads for the same genome for improving the consensus quality of the assembly (Zimin et al. 2020).

Its inputs are the genome sequence and a fasta or fastq file (or files) of Illumina reads and its outputs are the polished genome. Polca uses an approach based on the alignment of reads Illumina on the consensus, then identifies any locations where the reads indicate a possible error, and then to fix those errors using the read sequences (Zimin et al. 2020).

The basic outline of the script is to align the Illumina reads to the genome, using the BWA tool (Li and Durbin 2009) and then call short variants from the alignments. A variant call is treated as a putative error in the consensus if the count of the alternative allele observations is greater than 1 and at least twice the count of the reference allele. Each error is fixed by replacing the error variant with the highest scoring alternative allele suggested by the Illumina reads (Zimin et al. 2020).

3.9 - Purging with Purge_Dups

Canu in its assembly splits haplotypes into separate contigs producing the FASTA file in which there are so-called haplotype bubbles. Canu basically produces a diploid genome in which the homozygous regions are found in a single copy (Homotigs), while the heterozygous regions (Heterotigs) are both represented in the final assembly file.

This splitting results in an assembly size larger than the haploid genome size (Koren et al. 2017).

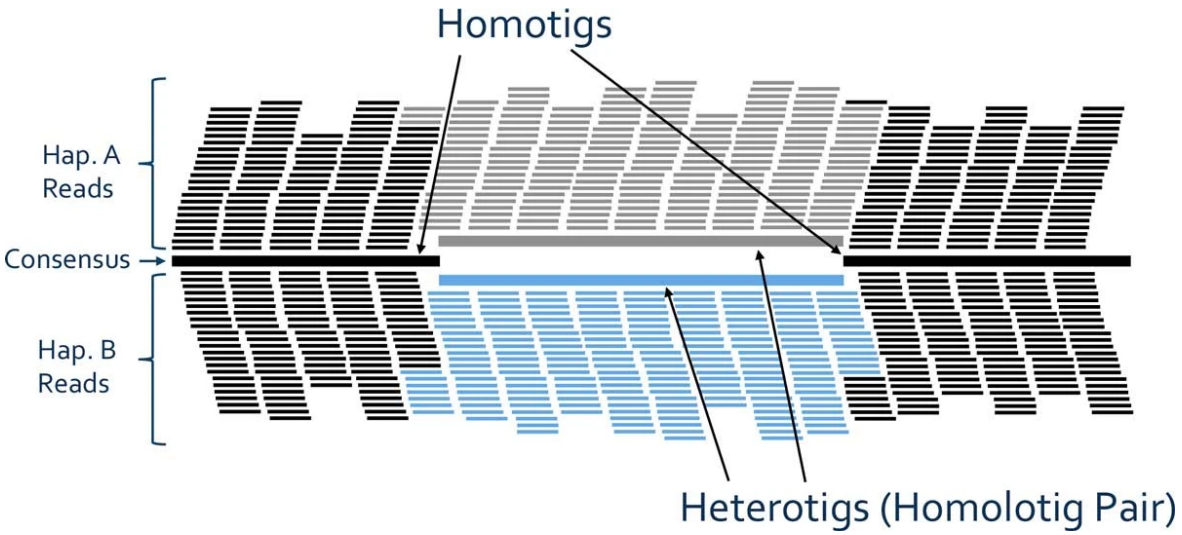

Figure 4: **Structures of a haplotype bubble**. Homotigs are formed from homozygous sequence, where read pileups from both haplotypes have the same consensus sequence. Heterotigs are formed from heterozygous sequences, where read pileups from each haplotype have a unique consensus due to variation. Inasmuch as two heterotigs are homologous, we say that they form a homolotig pair (Bodily et al. 2015).

The presence of these haplotype bubbles in the assembly leads to a large percentage of duplicates (BUSCO evaluation) and this creates problems with the subsequent assembly steps. Therefore, to remove one of the two regions in the haplotype bubbles, we did a purging round with Purge_Dups. To purge, Purge_dups first needs a configuration file to be generated with the pd_config.py script by providing the directory address of the short reads (trimmed Illumina reads), the long reads (Pacbio reads) and the name of the configuration file to be generated.

After generating the configuration file, which specifies the computational resources dedicated to each purging step and the directories on which the outputs are produced (these parameters can be modified manually), we launched the run_purge_dups.py command, which performs purging by automatically carrying out the following steps:

1. “Use minimap2 (Li 2018) to map long read sequencing data onto the assembly and collect read depth at each base position in the assembly. The software then uses the read depth histogram to select a cutoff to separate haploid from diploid coverage depths, allowing for various scenarios where the total assembly is dominated by haploid or diploid sequence.
2. The software segments the input draft assembly into contigs by cutting at blocks ‘N’s, and uses minimap2 to generate an all by all self-alignment.
3. Purge_dups next recognize and remove haplotigs, and remove all matches associated with haplotigs from the self-alignment set.
4. Finally the software chains consistent matches in the remainder to find overlaps, then calculate the average coverage of the matching intervals for each overlap, and mark an unambiguous overlap as heterozygous

when the average coverage on both contigs is less than the read depth cutoff found in step 1, removing the sequence corresponding to the matching interval in the shorter contig." (Guan et al. 2019)

We have therefore obtained as purging output a purged.fasta file containing the haploid draft of the *H. sbordonii* genome with the primary contigs and a hap.fasta file representing the alternative genome of *H. sbordonii*, i.e. it contains: the haplotigs (or heterotigs) representing: the haplotype alternative to that contained in the alternative genome, the contig sequences that overlap with each other and the junk sequences (over-represented or under-represented contigs)

3.10 - Scaffolding with LRscaf

Following purging, we have obtained a draft of the primary genome with a fair number of contigs. These contigs can be merged to give longer contigs or also called scaffolds. This step called Scaffolding uses the long reads produced by third generation sequencers (TGS) to join contigs. We used the LRscaf software to perform this step (Qin et al. 2019).

LRscaf needs the long reads to be mapped to the draft assembly so it can use these alignments to join contigs in the actual scaffolding step. So first we made an alignment of the reads on the draft assembly (produced after purging) using minimap2 (Li 2018) setting as output of the alignment a .paf file (appropriate input for the scaffolding step).

Following the alignment, we launched the scaffolding command that needs as input: the draft genome (obtained after purging) and the alignment file (.paf). The software is able to merge more contigs because using the alignment file (.paf) it evaluates if there are separate contigs on which reads are aligned to each other, in this way it can understand that in reality they can be merged. by doing so it will reduce the number of contigs because they will be merged. In the last scaffolding step, due to the presence of "complex regions" in the contigs (repeated regions or regions of low coverage) where it is not possible to join them with good reliability, gaps are inserted represented by the insertion of one or more "N" nucleotides depending on the size of the gap.

3.11 - Gap Filling with TGS-GapCloser

After the scaffolding step we had a draft genome with gaps in the sequence and therefore 'N' nucleotides that would interfere with the subsequent annotation steps. In this phase, therefore, we used the TGS-GapCloser software to perform the gap filling between the newly formed scaffolds and reduce the number of gaps and therefore of the "N" nucleotides (Xu et al. 2020). TGS-GapCloser uses as input the draft of the genome after scaffolding and the long reads (TGS reads: reads produced by third generation sequencers) to fill the gaps. TGS-GapCloser software uses two software integrated in its gapfilling process. The software used by TGS-GapCloser are minimap2 (Li 2018) and Racon (Vaser et al. 2017). We then launched the TGS-GapCloser.sh script which performed the following steps:

1. TGS-GapCloser splits the scaffolds in the presence of the "N" nucleotides and marks the two neighboring scaftigs (split scaffolds) as "gap to fill".
2. After recognizing the gaps and considering the two neighboring scaftigs, use minimap2 to align long reads against each gap to obtain the corresponding “candidate fragments”, which represents a segment

produced by the alignment of the long reads in the gap region, between two neighboring scaftigs plus 2-kb-long of aligned sequence on both sides of the gap.

Since there are more reads that can be aligned with the gap region, more "candidate fragments" can be produced, however the software only considers the first ten "candidate fragments" (based on a quality-based scoring system).

3. Once it has identified the 10 best "candidate fragments", in order to reduce the computational load, it combines them and makes corrections to the "candidate fragments" themselves using the Racon software, so that we only get one "correct candidate fragment".
4. This corrected candidate is realigned to the gap region and those 2-kb sequences aligning to the scaftigs on either side of the gap were removed and only the bases filling the gap from the corrected candidate were retained.

Following this gapfilling step, the number of contigs remained unchanged as did almost all the length statistics; however, all the "N "s and gaps inserted in the scaffolding step had been removed

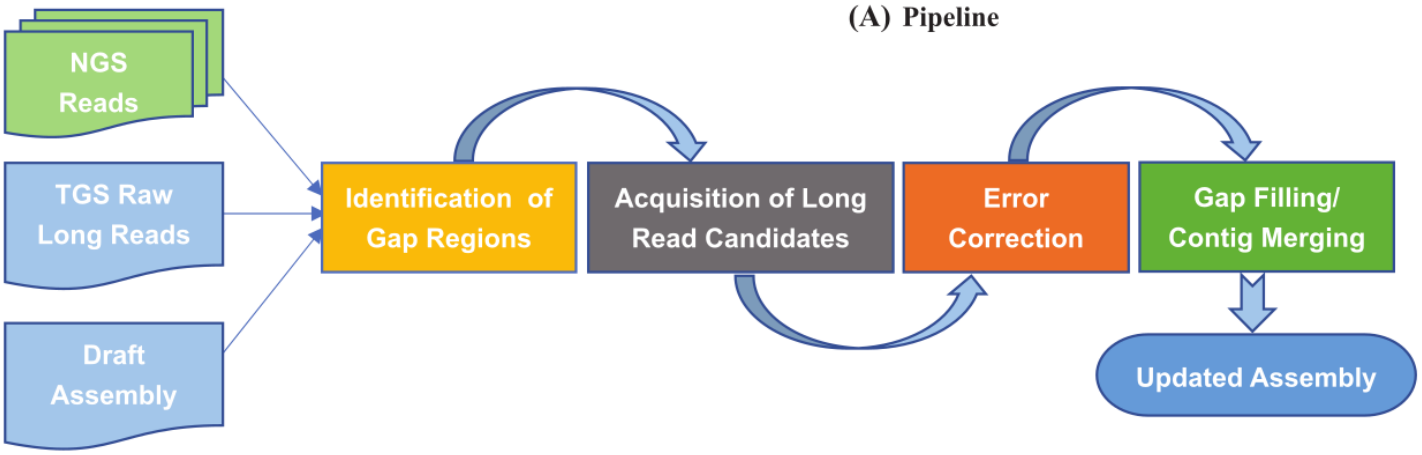

Figure 5: **A schematic of TGS-GapCloser workflow. A flow chart of the overall algorithm** (Xu et al. 2020)

3.12 - Final polishing with Polca

After gap filling, we performed a final polishing step to correct any errors made in the scaffolding and gap filling steps.

For this final step of the assembly we used the Polca software, which employs (as in the previous polishing steps with Polca) Illumina reads to carry out polishing to correct any errors inserted during scaffolding and gap filling.

3.13 - Final purging with purge_dups and merging of hap.fasta

In this last step of the genome assembly, we re-used Purge_dups. In practice, we ran Purge_dups on the fasta file obtained from Polca polishing to eliminate any duplicates that may have been added during the last scaffolding and gap filling steps.

After this second purging step, we have obtained a purged.fasta file representing the haploid primary genome of *H. sbordonii* and a hap.fasta file representing duplicates in the genome.

Finally, we merged the two hap.fasta files obtained from the two purges to obtain a draft of the alternative genome, which has a total length roughly similar to the primary genome but has more contigs with lower quality length statistics (fragmented contigs).

3.14 - Use of genomic data for population genetics estimation

After completing the genome assembly of *H. sbordonii* we used the assembly as the reference genome to align multiple individuals sequencing data and to estimate population genetics statistics of *H. sbordonii* and *H. semele* (a closely-related species to *H. sbordonii*). We analyzed 10 Illumina sequenced samples for both species plus 2 sequenced samples derived from two different species of the genus *Hipparchia*. All samples were mapped to the reference genome and the mapped reads used for the variant calling.

3.14.1 Variant calling with Freebayes

"Freebayes (Garrison and Marth 2012) is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events)" (Garrison and Marth 2012).

"Freebayes uses short-read alignments (BAM files) for any number of individuals from a population and a reference genome (in FASTA format) to

determine the most-likely combination of genotypes for the population at each position in the reference. It reports positions which it finds putatively polymorphic in variant call file (VCF) format." (Garrison and Marth 2012).

Freebayes therefore needs a BAM file for each individual, so 22 BAM files had to be generated. To generate the BAM files needed for variant calling with freebayes, I performed the following steps for the 22 sequenced individuals:

1. We ran the BWA index command which produces an index of the reference genome (I used the primary assembly file) which is then used as input in the following steps.

   This step was carried out only once.

2. We then ran the BWA mem command which needs: the short reads of the sequenced individual and the index file produced in the previous step.

   In this step, as in the following ones, we used a loop that performed the operation in question on all 22 samples, using the corresponding reads

3. We used the output from the previous step (22 SAM files) as input for the SAMTOOLS (Li et al. 2009) view command, resulting in 22 unsorted BAM files.
4. We then sorted the 22 BAM files obtained from the previous step with the SAMTOOLS sort command.
5. Once we had the 22 sorted BAM files, we had to add the read group and sample name required by freebayes to assign the sample name to each variant found in the variant calling phase.

   We used the Bamaddrg software to perform this operation (https://github.com/ekg/bamaddrg)

6. After generating the 22 BAM files we had to remove duplicates that would lead to variant misinterpretation. The duplicate reads are produced during the preparation of the library (PCR stage) and when a large amplification cluster is mistakenly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplication artifacts are referred to as optical duplicates (Auwera and O'Connor 2020). To remove these optical and PCR duplicates we used Picard on the BAM files produced after adding the read group and sample name

to each sample according to their origin (https://broadinstitute.github.io/picard/).

Once we had the 22 BAM files with read group, sample name and no duplicates, we launched the actual variant calling with Freebayes specifying all 22 samples that we divided by population according to the sample origin: population 1 for the 10 *Hipparchia sbordonii* samples, , population 2 for the 10 *Hipparchia semele* samples and population 3 and 4 for the two outgroup samplesand, Freebayes produced a VCF (variant calling format) file, which is a tabular file containing the various DNA polymorphisms such as SNPs (single nucleotide polymorphism), insertions, deletions and structural variants (compared to the reference genome) together with annotations (Petr Danecek et al 2011).

3.14.2 Filtering by quality and split MNPs

The VCF file produced by Freebayes, however, contains not only SNPs (with respect to the reference genome) but also MNPs (multiple nucleotide polymorphism), the latter are not indicative for making population genetics

estimates. MNPs represent regions where there are multiple polymorphisms and are therefore retained by the software as polymorphisms.

However, in order to be used by software that estimates the genetic diversity of a population, these MNPs must be split into several SNPs. We therefore used the Vcflib tool (Garrison et al. 2021), specifying the vcfallelicprimitives command to split the MNPs into SNPs. In conjunction with this step we wanted to keep only those SNPs that had a certain quality, which stands for phred-scaled quality score associated with the inference of the given alleles.

We therefore performed filtering for SNP quality above 30 which means that only SNPs with a call accuracy >99.9% or even that have a call error probability < 1 in 1000 are retained (Brent Ewing and Phil Green 1998; Ewing et al. 1998). Filtering by quality is done with the vcffilter command which is part of the Vcflib tool library.

3.14.3 - Further filtering and removal of SNPs in repeated regions

After filtering by quality we have a further filtering to consider only biallelic SNPs, which have a maximum average coverage of reads on the SNPs of 60x and a minimum coverage of reads on the SNPs of 5x. In this case to carry out this filtering step we have used the VCFtools program which has a number of packages that allow us to carry out different operations and extract information from the VCF files (Danecek et al. 2011) .

At this point we carried out what is known as masking, i.e. we removed the SNPs that fell within the repeated sequences from the VCF file. In doing so, we retained only those SNPs that fell on non-repeated regions previously identified by our collaborators on this project. To perform this step, we used the VCFtools tool, specifying to exclude from the VCF file the SNPs contained in the scaffold sequences specified via .bed files.

3.14.4 - Estimates of inter- and intra-population genetic diversity

After removing the SNPs in the repeated regions, we used h VCFtools to estimate population genetic diversity statistics, like nucleotide diversity (π) and Tajima's D in the population of *H. sbordonii* and *H. semele*. We first collected the SNPs corresponding to the two different populations and built two VCF files (one for each species) in which the respective SNPs were contained. Then from the VCF file filtered and masked by the repeated regions, in one case we removed all the SNPs of *H. semele* and the 2 outgroups to get a VCF file with only the SNPs of *H. sbordonii* and in the other case we removed all the SNPs of *H. sbordonii* and the 2 outgroups to get a VCF file with only the SNPs of *H. semele*. This step of retaining SNPs in two VFC files for the two species was also carried out on the VCF file before masking, so we had 4 VCF files at the end of these operations:

- 2 VCF files with only the SNPs of *H. sbordonii* of which one without SNPs in the repeated regions (masked VCF file)
- 2 VCF files with only the SNPs of *H. semele* of which one without SNPs in the repeated regions (masked VCF file)

On all 4 files we calculated: the nucleotide diversity (π) which represents the average of the nucleotide differences per site between two DNA sequences in all possible sample pairs in the population (Nei and Li 1979) and Tajima's D which represents the normalized difference between two measures of genetic diversity: the mean number of pairwise differences (π) and the expected diversity (θ) (Tajima 1989). Tajimas's D allows us to understand whether the study population is undergoing changes in the number of individuals or whether there is some kind of selection taking place (Tajima 1989). As a final index of population genetics, we used the filtered VCF file in which there were SNPs for *H. sbordonii* and *H. semele*, to calculate Fst in both the case of the VCF file with SNPs in repeated regions and in the case of the masked VCF file (without SNPs in repeated regions). The Fst between the two populations, in this case between *H. sbordoni* and *H. semele*, is a measure of differentiation of populations due to genetic structure, in practice it gives us the comparison of genetic variability within and between populations and therefore tells us how much the populations are different from each other and therefore if the two populations cross each other or if there is isolation of the two populations from a genetic point of view (Holsinger et al. 2009).

All previous population genetics estimates have been made by dividing the entire genome and thus each scaffold into 100 Kb windows.

4. RESULTS

The main objective of this work was to obtain an assembled genome that met high quality standards so that it could be used for further analysis. In order to assess the quality of the genome and, in particular, of the products obtained at each stage of the assembly, we used various instruments that allow us to measure parameters that give us an indication of quality.

4.1 - Genome size estimates

Using Genomescope and Jellyfish we obtained a genome size estimate (Table 4) of around 355 Mb together with heterozygosity estimates of 1.79% and homozygosity at 98.2%.

p = 2
k = 31

property	min	max
Homozygous (aa)	98.178%	98.2396%
Heterozygous (ab)	1.76036%	1.82201%
Genome Haploid Length	353,019,378 bp	355,030,047 bp
Genome Repeat Length	92,321,566 bp	92,847,396 bp
Genome Unique Length	260,697,812 bp	262,182,651 bp

Table 4: **Results of genomescope estimates.** Genomescope from K-mer profiles obtained in the analysis gives us maximum and minimum estimates of the percentage of the genome in homozygosity and heterozygosity, haploid genome length, extent of repeated elements and length of unique regions.

4.2 - Estimation of completeness and QV

After running Canu and producing the first draft of the H. sbordonii genome assembly, we performed a first evaluation of the assembly measuring completeness and QV using Merqury and obtained a completeness of 98.594% which indicated that almost all the K-mers present in the reads were in the assembly and therefore almost all the reads had been used in the assembly. In addition to completeness, Merqury gave us a quality estimate called QV which was 36.858. This value indicates the probability that the nucleotide in the assembly was incorrect by 0.000206131 or that the percentage of correctness of the nucleotide in the assembly was 99.979% (Rhie et al. 2020).

Both of these two indices give us a good perspective on the quality of the assembly carried out by CANU, but for further confirmation we used other systems to measure and evaluate the quality of the assembled genome.

4.3 - FASTA file statistics (assembly statistics)

After the evaluation with Merqury we went to evaluate the statistics of the FASTA file (genomic assemblate) produced by CANU (Table 4). In particular, we calculated length statistics: indicating whether the assembled product was fragmented into several contigs of different lengths, i.e. giving us an indication of the contiguity of the genome, and statistics on the presence of gaps in the contigs and the presence of 'N' nucleotides.

Number of contigs:	**1693**
Total size (bp):	**746923840**
N50 (bp):	**1053838**
L50:	**154**
N90 (bp):	195442
L90:	831
Mean contig size (bp):	441183
Longest contig (bp):	14608371
Third quartile (bp):	439083
Median (bp):	195442
First quartile (bp):	72330
Shortest contig (bp):	1718
Number of Ns:	0
Number of gaps (/N+/):	0

Table 5: **statistics of FASTA files after the genome assembly with CANU**

From the Merqury estimates we learned that CANU produced a good quality assembly (see QV and completeness) however looking at the high number of contigs produced (1693), the N50 of about 1Mb and the high L50 value which indicated high fragmentation, this indicated that the genome could be further improved in terms of quality and contiguity. Another value that caused us concern was the genome size, which was more than twice the predicted value. This high value was explainable because CANU in the assembly divides haplotypes into separate contigs when the allelic divergence is greater than the post-correction overlap error rate. This splitting results in an assembly size larger than the haploid genome size (Koren et al. 2017).

4.4 - BUSCO genome assembly evaluation

To further estimate the completeness of the assembly and assess the number of genes predicted and to see in which of the following categories C:complete [S:single-copy, D:duplicated], F:fragmented, and M:missing we used the BUSCO tool which gave us a table with: the list of predicted genes with description, which categories they fell into, their location in the contig and the code to access the gene in the OrthoDB database.

What we were interested in, however, was the summary of the statistics of the predicted genes and precisely which category they fell into.

C:98.8% [S:25.5%, D:73.3%], F:0.3%, M:0.9%,n:5286	
5224	**Complete BUSCOs (C)**
1349	**Complete and single-copy BUSCOs (S)**
3875	**Complete and duplicated BUSCOs (D)**
14	**Fragmented BUSCOs (F)**
48	**Missing BUSCOs (M)**
5286	**Total BUSCO groups searched**

Table 6: **BUSCO summary statistics** in the genome assembly

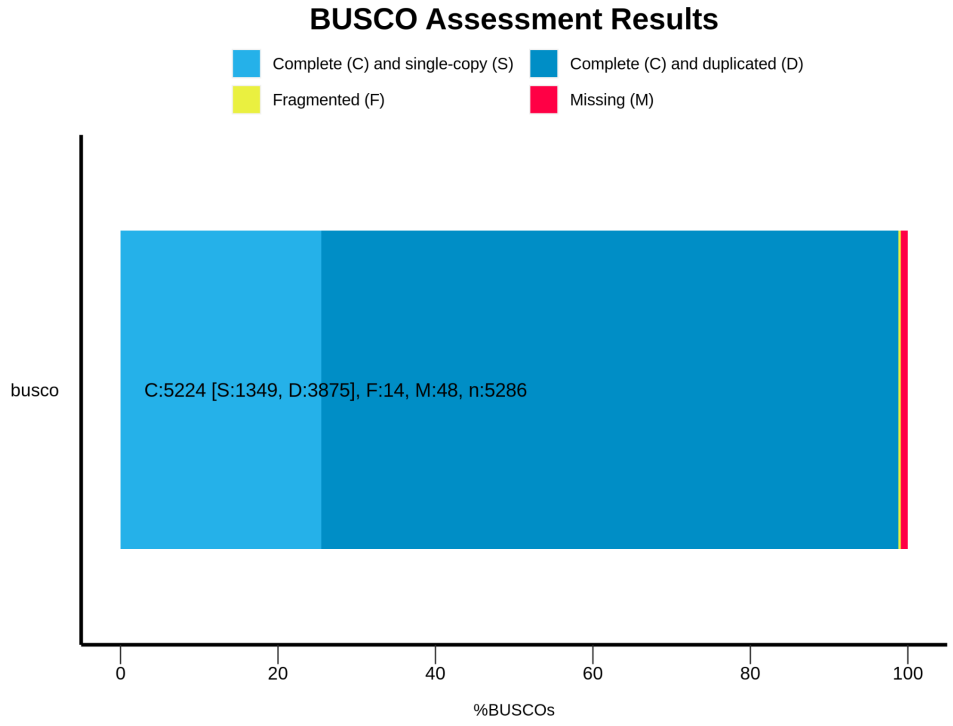

Figure 5: **BUSCO plot** graphical representation of the distribution in the various categories of the predicted genes in the genome assembly produced by CANU

As can be seen in the table above and in the graphical representation of the results, there was evidence of a high number of duplicated genes, indicating that different haplotypes of the same gene had been retained in the assembly, while the number of Missing and Fragmented were small compared to the total number of predicted genes.

4.5 - Contamination and quality assessment with Blobtools

With this investigation we wanted to assess whether there were any contaminants in the sequenced and assembled sample and, if so, which organisms they belonged to and which contigs contained traces of genetic material foreign to *H. sbordonii*.

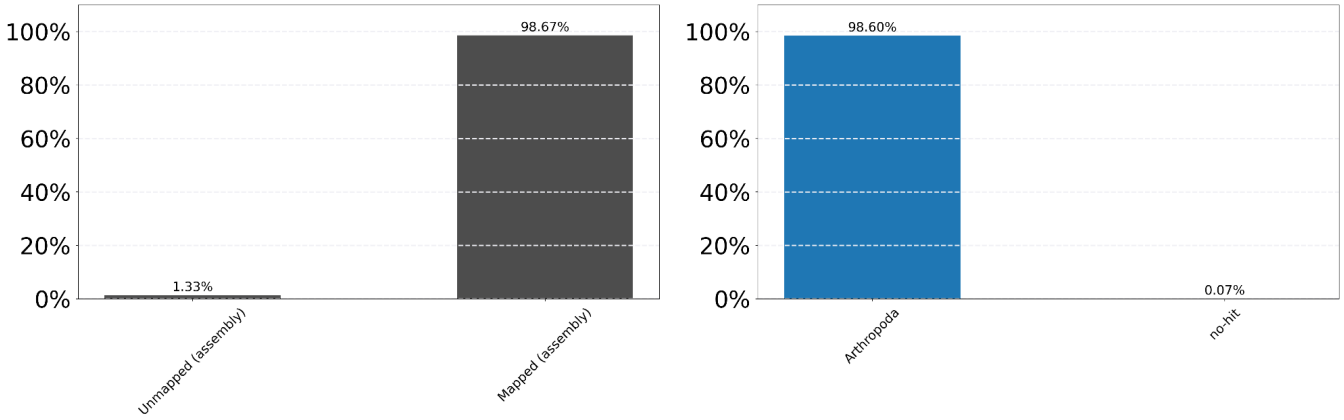

Figure 6: **ReadCovPlot**: The plot on the left shows a measure of the percentage of reads used (mapped) and reads not used in the assembly (unmapped) and this gives us an estimate of completeness. The plot on the right shows the percentage of mapped reads according to the taxonomic group to which they belong.

The ReadCovPlot (in line with previous estimates of completeness) generated by Blobtools shows that almost all reads (98.67%) were used in the assembly and that 99.92% of these reads (90.60% of total reads) have a sequence found in arthropods, which are precisely the phylum to which *H. sbordonii* belongs. This suggested that there were no contaminants in the genomic assemblage and that it was therefore not necessary to proceed with the removal of sequences belonging to other organisms, since only 0.071% of the assembled reads (0.07% of the total reads) were not attributed to any sequence (no-hit) and, being a small value compared to the total reads, was negligible in terms of the quality of the assemblage.

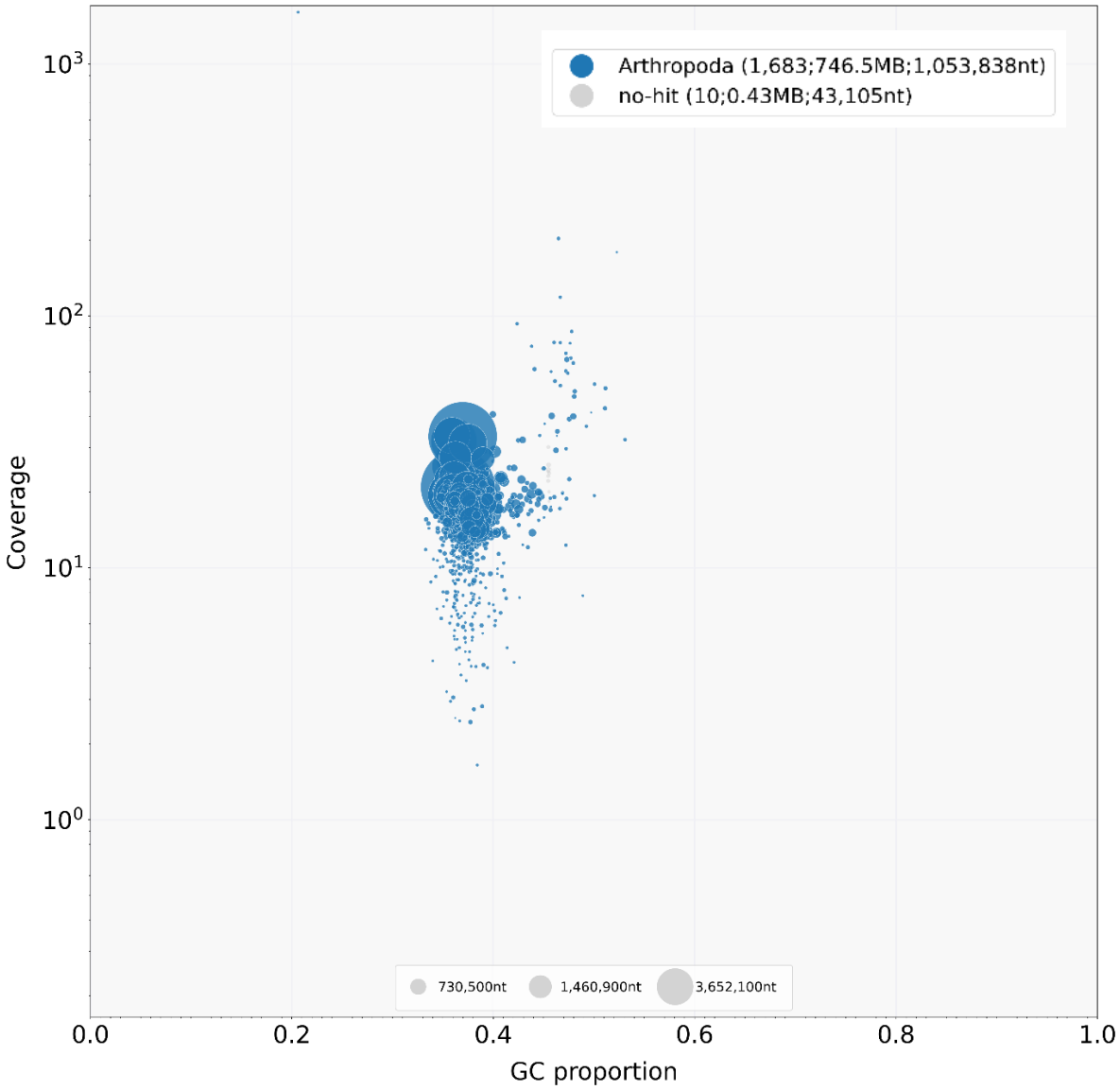

Figure 7: **BlobPlot** shows us the contigs (sequences in the assembly) represented in circles, with diameter scaled proportional to the length of the sequence and coloured according to the taxonomic annotation based on the results of the similarity search: BLASTn. The circles are positioned on the X-axis according to their GC proportion and on the Y-axis according to the sum of coverage across the library (Laetsch et al.).

In the Blobplot produced by Blobtools (Figure 7) we can appreciate the distribution of the larger contigs and how they are distributed in proportion to the GC content and coverage of reads in the contigs. In this way it was possible to graphically evaluate the fragmentation of the assembly and whether there were contigs with anomalous GC contents or too low or high coverage.

What we see from the graph is that the assembly obtained has a high amount of small contigs that have relatively low coverage, compared to large contigs (larger diameter circles) and this could indicate that the fragmented contigs contained sequences that were not assembled together with good confidence due to low coverage, but this suggested that these fragments could be joined to give larger contigs by remapping reads in the genome and making corrections with polishing rounds.

4.6 - Polishing results

Previous results showed that the genome produced had a high number of duplicates and that the quality and completeness could be improved.

In the subsequent steps of the assembly we carried out polishing rounds with the aim of increasing the quality and completeness of the genome. These polishing rounds consisted of mapping the reads used in the assembly onto the genome produced so that this operation could highlight errors in the assembly in order to correct them and have a product as reliable as possible, in order to avoid errors due to the assembly that could distort the subsequent operations carried out.

In the various polishing steps we found that the quality in terms of contiguity, represented by the values of N50, L50 and number of contigs, and ancillary evaluations (such as N90, L90, average length of contigs etc.), and in terms of completeness and probability of error (QV), increased slightly in the various steps.

After polishing with Arrow		**After the first round of polishing with Polca**		**After the second round of polishing with Polca**	
Number of contigs:	**1693**	**Number of contigs:**	**1693**	**Number of contigs:**	**1693**
Total size (bp):	**747200956**	**Total size (bp):**	**747193022**	**Total size (bp):**	**747192726**
N50 (bp):	**1054137**	**N50 (bp):**	**1054145**	**N50 (bp):**	**1054145**
L50:	**154**	**L50:**	**154**	**L50:**	**154**
N90 (bp):	199261	**N90 (bp):**	199264	**N90 (bp):**	199264
L90:	831	**L90:**	831	**L90:**	831
Mean contig size (bp):	441347	**Mean contig size (bp):**	441342	**Mean contig size (bp):**	441342
Longest contig (bp):	14612106	**Longest contig (bp):**	14611971	**Longest contig (bp):**	14611966
Third quartile (bp):	439207	**Third quartile (bp):**	439201	**Third quartile (bp):**	439201
Median (bp):	195526	**Median (bp):**	195525	**Median (bp):**	195523
First quartile (bp):	72361	**First quartile (bp):**	72364	**First quartile (bp):**	72364
Shortest contig (bp):	1723	**Shortest contig (bp):**	1723	**Shortest contig (bp):**	1723
Number of Ns:	0	**Number of Ns:**	0	**Number of Ns:**	0
Number of gaps (/N+/):	0	**Number of gaps (/N+/):**	0	**Number of gaps (/N+/):**	0
Merqury statistics		**Merqury statistics**		**Merqury statistics**	
completeness	**98.97**	**completeness**	**99.02**	**completeness**	**99.022**
QV	**35.09**	**QV**	**35.30**	**QV**	**35.3075**

Table 7: **statistics of FASTA files after the various polishing steps,** in bold are highlighted the statistics most indicative of quality

Looking at the estimate of the total length of the genome, it can be seen that the genome is still more than twice as long as the haploid genome; this, together with the high number of duplicates predicted by BUSCO and CANU's characteristic of assembling and retaining different haplotypes, has

led us to carry out purging, which had considerable effects in terms of improving quality.

4.7 - Evaluation of Purging

After purging duplicated haplotigs with Purge_dups, we obtained a primary genome (primary.fasta) and an alternative genome (hap.fasta).

The major change after this step is the total length of the genome which is now 398.56 Mb, quite close to the estimate obtained with Jellyfish, thus suggesting that we obtained a first draft of the haploid genome.

The primary genome (haploid) contains the primary contigs which, as can be seen in the following table, have a higher quality in terms of contiguity, reaching N50 values greater than 2.6 Mb and L50 of 44, while the alternative genome has lower contiguity values (higher number of contigs, N50 of 403KB and L50 of 253). This is due to the fact that the alternative genome contains more fragmented contigs, contigs represented only by the overlapping of regions in common between two contigs and junk contigs.

After Purging with Purge_Dups			
Purged.fasta		**Hap.fasta**	
Number of contigs:	**324**	**Number of contigs:**	**1580**
Total size (bp):	**398564366**	**Total size (bp):**	**348628360**
N50 (bp):	**2592164**	**N50 (bp):**	**403674**
L50:	**44**	**L50:**	**253**
N90 (bp):	652184	N90 (bp):	101515
L90:	155	L90:	878
Mean contig size (bp):	1230137	Mean contig size (bp):	220651
Longest contig (bp):	14611966	Longest contig (bp):	2699606
Third quartile (bp):	1715822	Third quartile (bp):	300896.5
Median (bp):	580761.5	Median (bp):	121473.5
First quartile (bp):	174850.5	First quartile (bp):	52935.5
Shortest contig (bp):	1723	Shortest contig (bp):	12498
Number of Ns:	0	Number of Ns:	0
Number of gaps (/N+/):	0	Number of gaps (/N+/):	0
Merqury statistics Purged.fasta		**Merqury statistics Hap.fasta**	
completeness	**76.6082**	**completeness**	**63.6046**
QV	**41.4717**	**QV**	**32.5864**
Merqury statistics combined			
combined completeness		**99.0219**	
combined QV		**35.3075**	

Table 8: **statistics of FASTA files obtained from Purging,** in bold are highlighted the statistics most indicative of quality

Another interesting fact concerns the completeness and QV of the two files compared; as the primary genome has a much higher quality in terms of QV (41.47) compared to the alternative genome (32.58), this indicates that the primary genome even has a probability of base calling error about 10 times lower than the alternative genome. Considering the two genomes separately, the completeness is reduced. In fact by excluding haplotype sequences from the primary assembly, reads belonging to secondary haplotypes can no longer map to it and vice versa, leading to a reduction in completeness in both cases. However, if we consider the combined completeness and QV, the values do not differ from the values calculated after polishing.

Further confirmation of the removal of alternative haplotypes was obtained from BUSCO on the primary genome file. BUSCO results show us that the number of duplicates is reduced to values compatible with a haploid genome. Furthermore, the completeness calculated by BUSCO remains at high values (98.5%), indicating that purging has not eliminated coding regions.

BUSCO on POLISHED and PURGED files (Lepidoptera dataset)	
C:98.5% [S:97.9%, D:0.6%], F:0.3%, M:1.2%, n:5286	
5207	Complete BUSCOs (C)
5176	Complete and single-copy BUSCOs (S)
31	Complete and duplicated BUSCOs (D)
17	Fragmented BUSCOs (F)
62	Missing BUSCOs (M)
5286	Total BUSCO groups searched

Table 9: **BUSCO summary statistics on the primary genome (purged.fasta)**

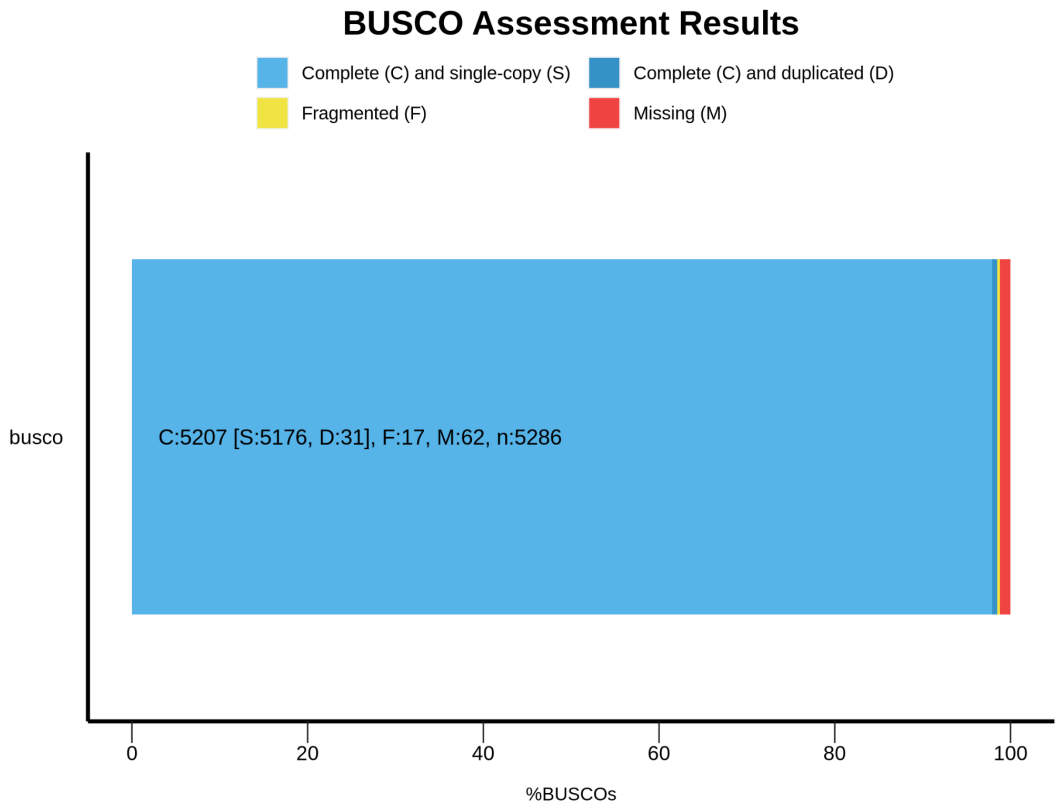

Figure 8: **BUSCO plot:** graphical representation of the distribution in the various categories of the predicted genes in the primary genome (purged.fasta).

4.8 - Results of Scaffolding and Gap Filling

In these two steps of the assembly pipeline we aimed to increase the quality, in terms of contiguity, of the haploid genome obtained by purging (purged.fasta). In particular, with the scaffolding step (with LRscaff) we tried to merge contigs that are separate but in reality belong to the same scaffold (or chromosome) and with the gap filling step (with TGS-GapCloser) we aim to fill the gaps created in the scaffolding step.

Following the scaffolding step, the quality statistics in terms of contiguity are greatly improved, halving the total number of contigs, indicating that more than half of the previous contigs had been merged (Table 10). This is reflected in the contiguity statistics, which reach values of N50 of 9.1 Mb and an L50 of 16, while the quality of the genome in terms of completeness and QV are not substantially changed as expected.

However, 31 gaps are included in the scaffolding (highlighted by the statistics below); these gaps are then eliminated by gap illing, which does not bring obvious improvements in terms of contiguity, completeness (calculated with a BUSCO and Merqury) and QV.

After Scaffolding with LRscaff		**After Scaffolding with TGS-GapCloser**	
Number of contigs:	**124**	**Number of contigs:**	**124**
Total size (bp):	**401477959**	**Total size (bp):**	**401494425**
N50 (bp):	**9145526**	**N50 (bp):**	**9145535**
L50:	**16**	**L50:**	**16**
N90 (bp):	2529149	N90 (bp):	2529149
L90:	52	L90:	52
Mean contig size (bp):	3237725	Mean contig size (bp):	3237858
Longest contig (bp):	16602411	Longest contig (bp):	16602411
Third quartile (bp):	4431463	Third quartile (bp):	4431442
Median (bp):	1616831	Median (bp):	1616831
First quartile (bp):	148819	First quartile (bp):	148819
Shortest contig (bp):	14456	Shortest contig (bp):	14456
Number of Ns:	**109953**	**Number of Ns:**	**0**
Number of gaps (/N+/):	**31**	**Number of gaps (/N+/):**	**0**
Merqury statistics		**Merqury statistics**	
completeness	**76.6034**	**completeness**	**76.6104**
QV	**41.466**	**QV**	**41.254**

Table 10: **statistics of FASTA files after Scaffolding and Gap Filling,** in bold are highlighted the statistics most indicative of quality

BUSCO on Scaffolded and Gap filled files (Lepidoptera dataset)	
C:98.5%[S:97.1%,D:1.4%],F:0.3%,M:1.2%,n:5286	
5205	Complete BUSCOs (C)
5132	Complete and single-copy BUSCOs (S)
73	Complete and duplicated BUSCOs (D)
15	Fragmented BUSCOs (F)
66	Missing BUSCOs (M)
5286	Total BUSCO groups searched

Table 11: **BUSCO summary statistics on the primary genome after scaffolding and gap filling**

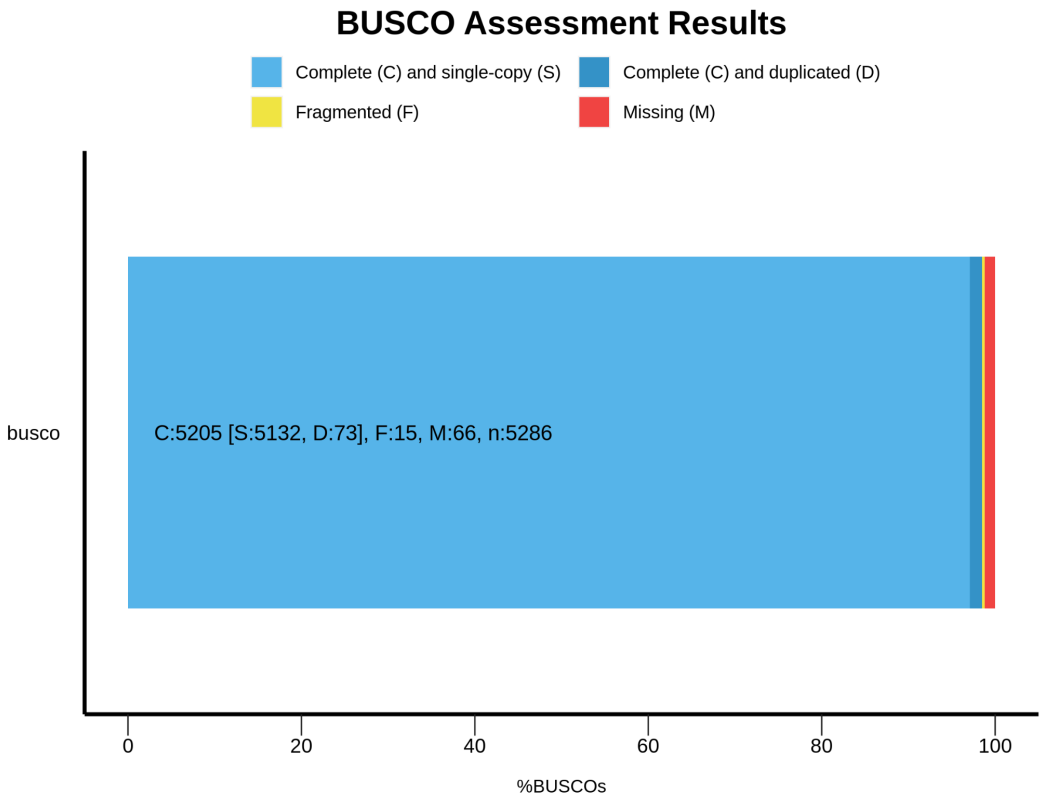

Figure 9: **BUSCO plot** graphical representation of the distribution in the various categories of the predicted genes in the primary genome after scaffolding and gap filling.

4.9 - Final steps of Polishing and Purging

In the final two steps (polishing and purging) we tried to correct any errors and possible duplicates inserted in the scaffolding and gap filling steps. In this way we obtained a further reduction in the number of contigs, which in Purged.fasta reaches 110 with an N50 of 9.1 Mb and an L50 of 16, which indicates that 50% of the genome is contained in contigs of length greater than or equal to 9.1 Mb and that there are only 16 contigs that are greater than or equal to 9.1 Mb, in other words that 50% of the genome is contained in 16 contigs and this data gives us an estimate of the good contiguity of the genome. When we evaluated the quality in terms of completeness and QV in the various steps there was not much improvement, even in some cases the QV was reduced (by decimal values).This reduction is mainly noticeable in the gap filling step, where we assumed that in this step due to the gap filling by aligning the reads on the gap there is a possible inclusion of errors and therefore the probability of error increases. However, the QV and values are not altered to the point of having to carry out other polishing steps.

After final polishing with Polca		**After final purging with Purge_Dups (Purged.fasta)**	
Number of contigs:	**124**	**Number of contigs:**	**110**
Total size (bp):	**401443506**	**Total size (bp):**	**397095341**
N50 (bp):	**9143845**	**N50 (bp):**	**9143845**
L50:	**16**	**L50:**	**16**
N90 (bp):	2528629	N90 (bp):	2526579
L90:	52	L90:	51
Mean contig size (bp):	3237447	Mean contig size (bp):	3609957
Longest contig (bp):	16600010	Longest contig (bp):	16600010
Third quartile (bp):	4431257.5	Third quartile (bp):	4788419
Median (bp):	1616641	Median (bp):	2046902
First quartile (bp):	148808	First quartile (bp):	275879
Shortest contig (bp):	14455	Shortest contig (bp):	14455
Number of Ns:	0	Number of Ns:	0
Number of gaps (/N+/):	0	Number of gaps (/N+/):	0
Merqury statistics		**Merqury statistics**	
completeness	**76.716**	**completeness**	**76.602**
QV	**40.342**	**QV**	**40.4217**

Table 12: **statistics of FASTA files after the final polishing and purging,** in bold are highlighted the statistics most indicative of quality

To see if in the last steps there were any worsening or improvement in the completeness of the genome we made an analysis with BUSCO on the FASTA file obtained after polishing and purging (final primary genome) and as in the previous statistics there were slight improvements in the completeness that reached 98.6% but more than anything else there was a reduction in the number of duplicated genes that in the product obtained from the scaffolding and gap filling steps was 73 while after the final polishing and purging we obtain a number of duplicated genes equal to 25 with a consequent increase in the number of complete genes in single copy that goes from 5132 to 5185 in the final product.

BUSCO after polishing and purging (final)(dataset Lepidoptera)	
C:98.6% [S:98.1%, D:0.5%], F:0.2%, M:1.2%, n:5286	
5210	**Complete BUSCOs (C)**
5185	**Complete and single-copy BUSCOs (S)**
25	**Complete and duplicated BUSCOs (D)**
12	**Fragmented BUSCOs (F)**
64	**Missing BUSCOs (M)**
5286	**Total BUSCO groups searched**

Table 13: **BUSCO summary statistics on the primary genome after the final polishing and purging**

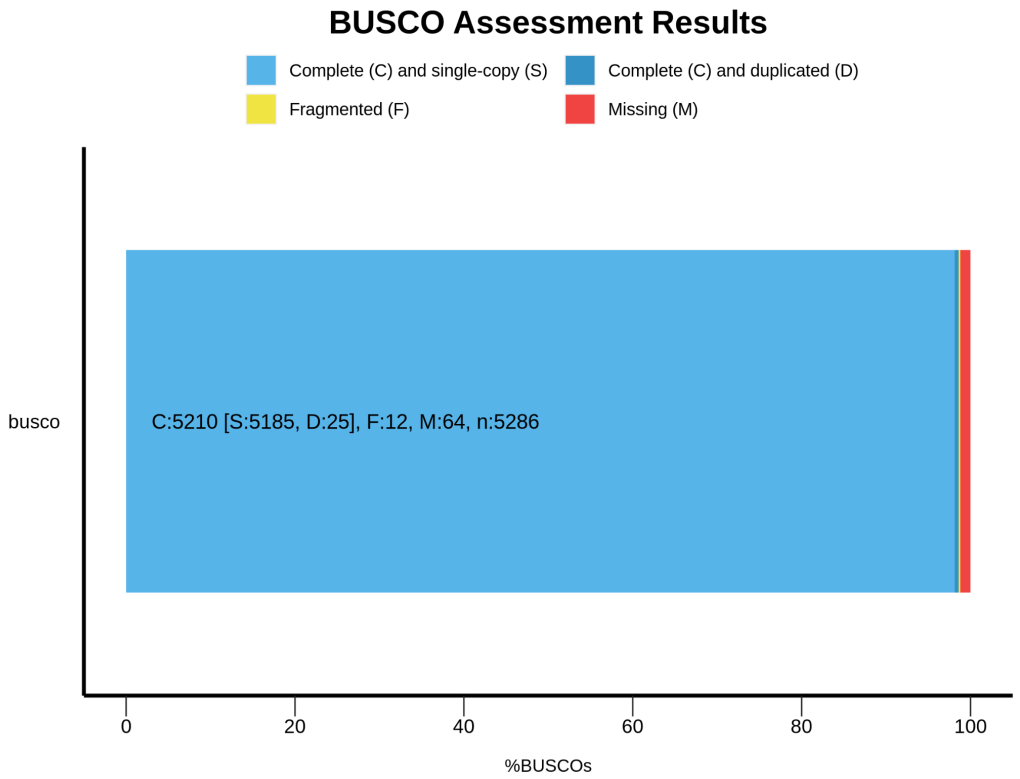

Figure 10: **BUSCO plot** graphical representation of the distribution in the various categories of the predicted genes in the primary genome after the final polishing and purging.

As a final analysis, in order to be able to graphically visualise the distribution of contigs in the primary genome, in terms of coverage and GC content, we performed a final analysis with Blobtools. The number of contigs is substantially reduced, the contigs are clustered more by coverage and GC content and these contigs are much longer than the product of the assembly with CANU (Figure 11), in line with the previous estimation.

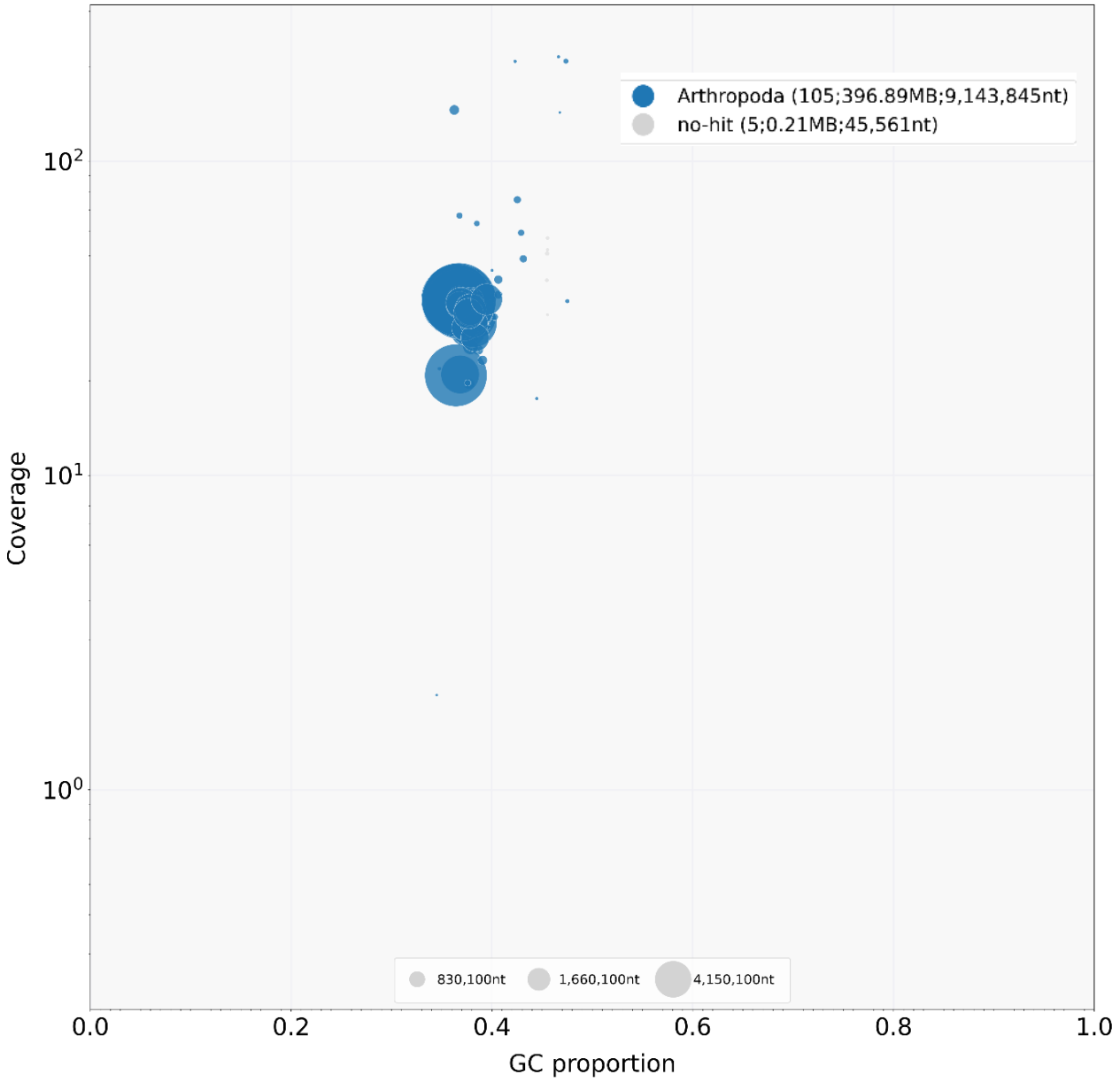

Figure 11: **BlobPlot** shows us the scaffold represented in circles, with diameter scaled proportional to the length of the sequence and coloured according to the taxonomic annotation based on the results of the similarity search: BLASTn. The circles are positioned on the X-axis according to their GC proportion and on the Y-axis according to the sum of coverage across the library (Laetsch et al.).

4.10 - Results of population genomics analyses

In this final step of the analyses, we used the genome assembly produced in the previous steps as a reference genome to map genome resequecing data from 22 specimens of *H. sbordonii*, *H. semele* and two outgroup species to evaluate:

- The nucelotide diversity (π) along the genome of the two *Hipparchia* species taken into analysis (*H. sbordonii* and *H. semele*) (Nei and Li 1979).
- Tajima's D, which allows us to understand if the populations under study underwent changes in size or if some kind of selection acted in some genomic regions (Tajima 1989).
- The Fst between the two populations, in this case between *H. sbordonii* and *H. semele*, as a measure of genetic differentiation (Holsinger et al. 2009).

The above estimates were carried out considering our assembled genome and thus of *H. sbordonii* as the reference genome. Therefore, the variant calling carried out using the reads obtained from the sequencing of 10 individuals of *H. semele* will have a higher number of SNPs than the variant calling with the

sequencing data of 10 individuals of *H. sbordonii* because, being two different species (although very close evolutionarily), they had different evolutionary histories leadingto the accumulation of different mutations and therefore different SNPs.

All the population genetics statistics were estimated following the masking of SNPs that fell into repeated regions.

4.10.1 - Nucleotide diversity (π)

With this estimate, we wanted to measure the nucleotide diversity in the two populations, which gives us an indication of the population size of *Hipparchia sbordonii* and *H. semele*, and to assess whether there were regions of the genome in the study species where there was greater or lower nucleotide diversity than the genome average.

By estimating the average nucleotide diversity of the whole genome (divided into 100Kb windows) in the two *Hipparchia* species, we obtained π values for *H. sbordonii* of 0.0048 and π values for *H. semele* of 0.00783.

In particular, if we compare the graphs of nucleotide diversity in the genome on a logarithmic scale, we see that *H. sbordonii* has a lower average logarithmic nucleotide diversity than *H. semele*.

In fact, by transforming the value of the average π on a logarithmic scale, we see that *H. sbordonii* has a logarithmic π ($-\log_{10}\pi$) of 2.32, while *H. semele* has a logarithmic π ($-\log_{10}\pi$) of 2.1, indicating a greater nucleotide diversity.

In the graphical representation of nucleotide diversity along the genome divided into 100 Kb windows it is possible to appreciate the difference in nucleotide diversity between the two populations (Figure 12 and 13).

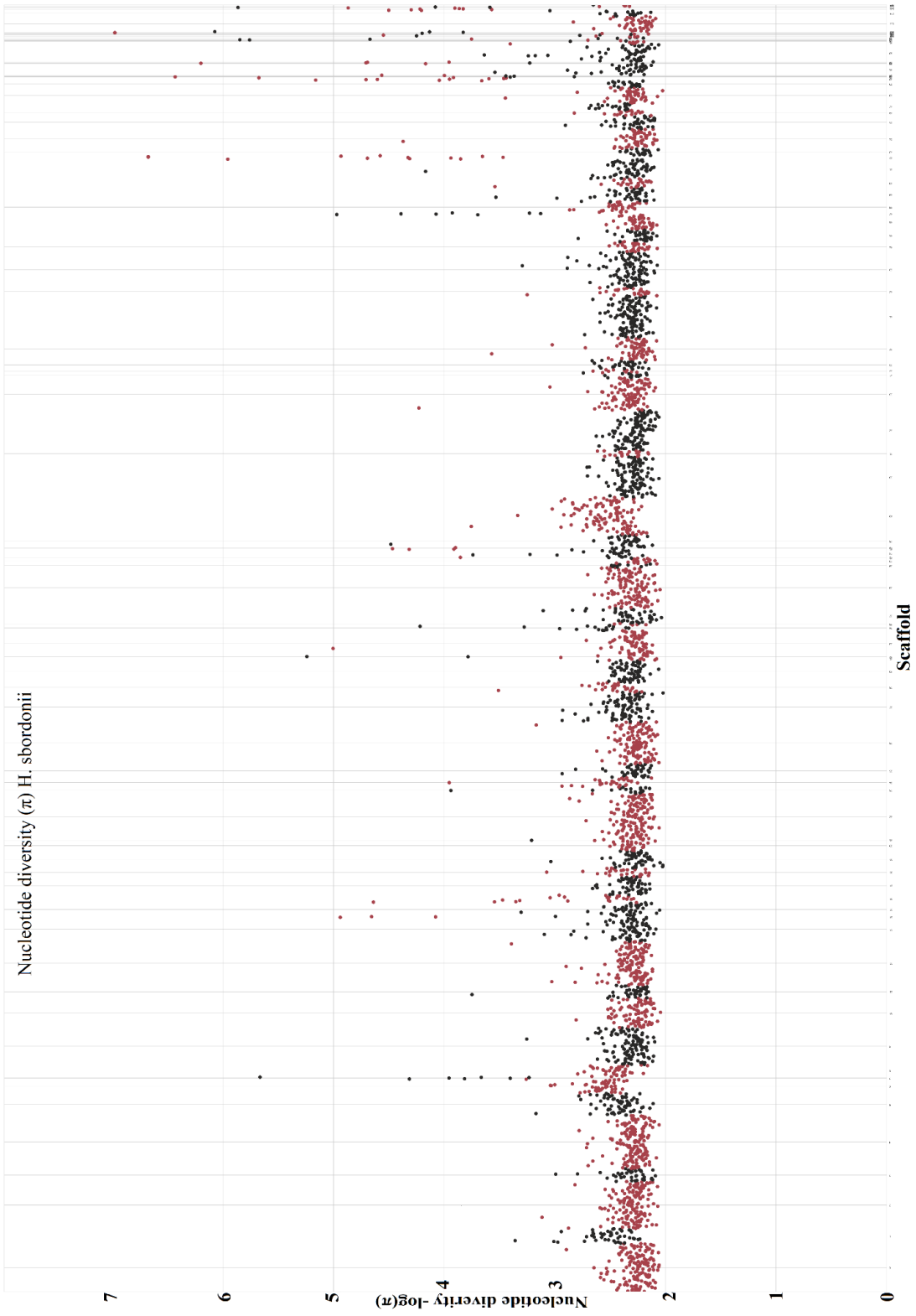

Figura 12: **nucleotide diversity (-log$_{10}$π) plot in *H. sbordonii***

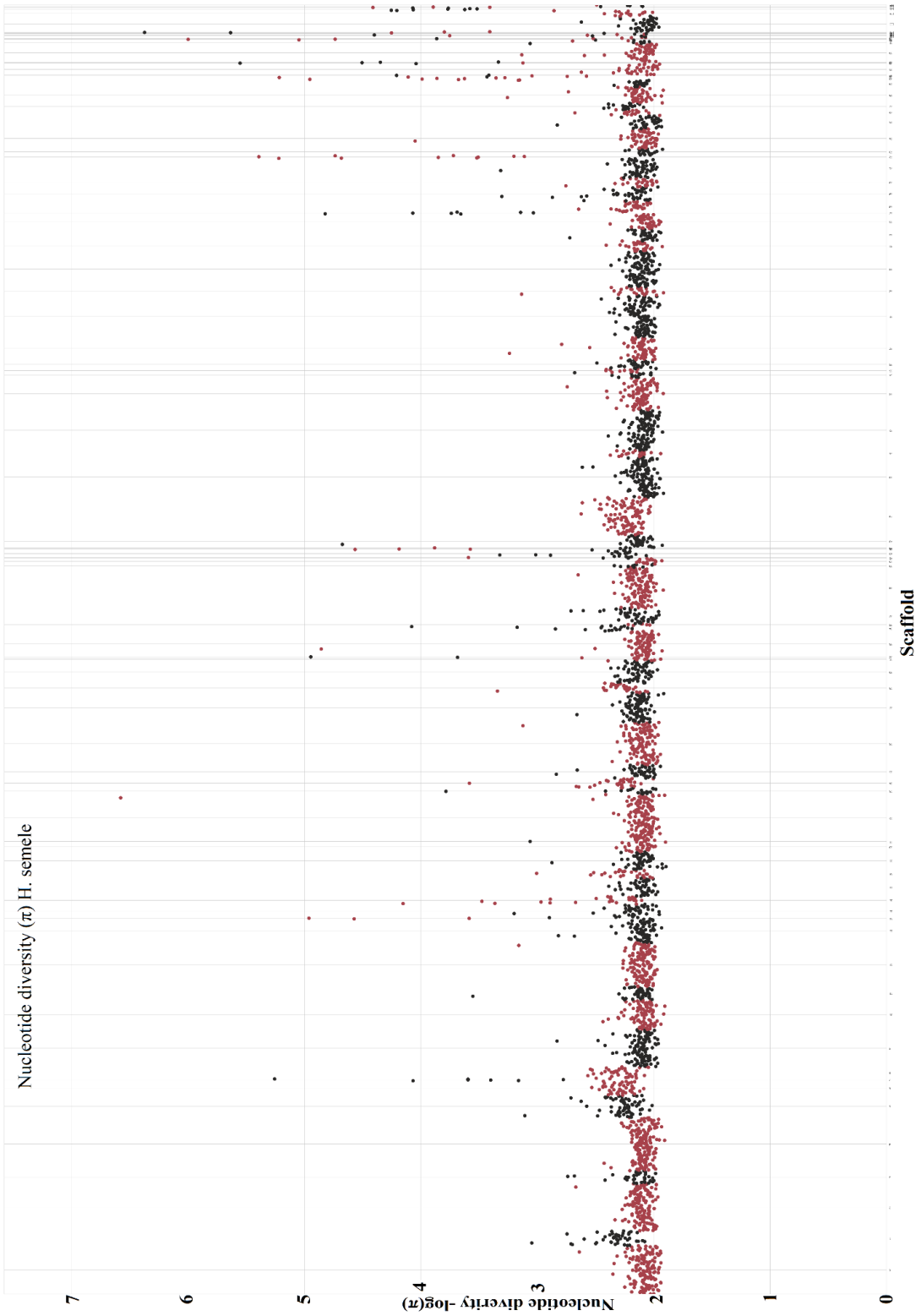

Figura 13: **nucleotide diversity (-log$_{10}$π) plot in *H. semele***

H. sbordonii has a lower nucleotide diversity than *H. semele* and this is due to the different size of the population because in a population with a reduced number of individuals it will have lower nucleotide diversity than in a population with a greater number of individuals (Mackintosh et al. 2019).

4.10.2 - Tajima's D

With this estimate we tested whether the two populations were stationary or if there has been any recent demographic change and if any region shows signature of selection taking place in the two populations of the two *Hipparchia* species. Calculating the average Tajima's D across the whole genome (divided into 100Kb windows) in the two *Hipparchia* species, we obtained a Tajima's D value for *H. sbordonii* of 0.652 and for *H. semele* a Tajima's D value of -1.142.

This different average trend can also be seen in the graphs below (Fugure 14 and 15).

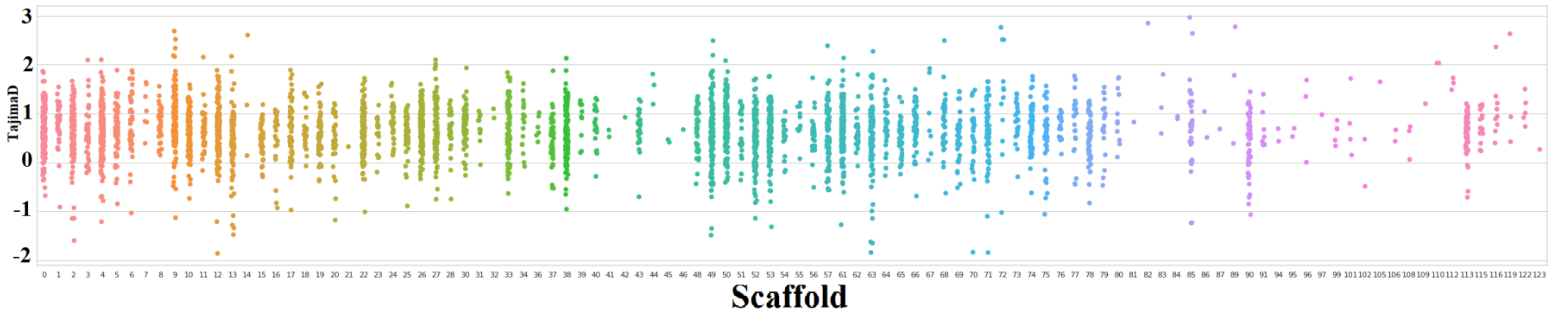

Figure 14: **Tajima's D plot of *H. sbordoniii* calculated on the genome divided into 100kb windows which is represented by a single point :** there is a positive trend in Tajima's D throughout the genome.

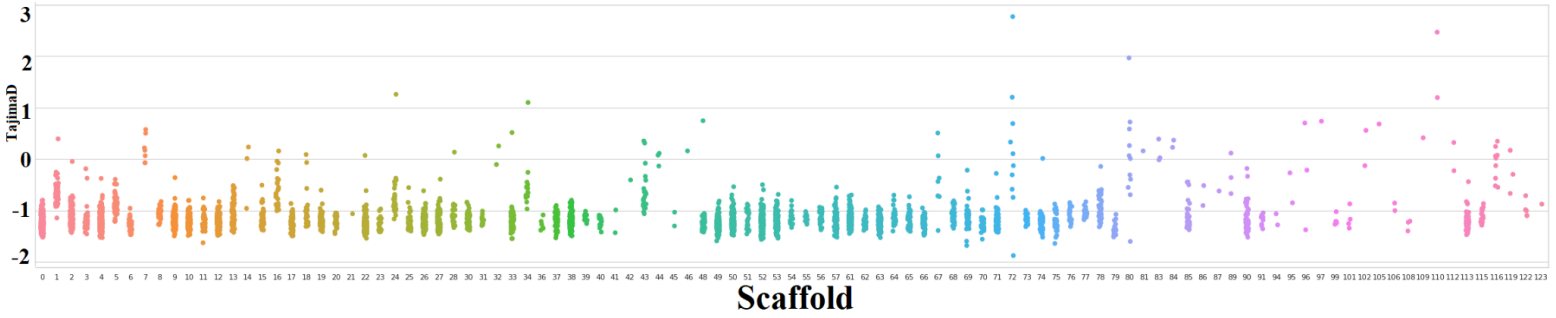

Figure 15: **Tajima's D plot of *H. semele* calculated on the genome divided into 100kb windows which is represented by a single point :** there is a negative trend in Tajima's D throughout the genome.

Tajima's D being described as the normalized difference between the observed nucleotide diversity and the expected diversity. Tajima's D can assume values

greater than, equal to or less than zero; on the basis of the value obtained, deductions can be made:

- If we have a value of Tajima's D = 0, this means that the observed diversity is equal to the expected diversity, which indicates that the population is stationary or the genomic region is not subject to any kind of selection.
- In the case of a Tajima's D > 0 this means that the observed nucleotide diversity is greater than the expected one, this is due either to a recent reduction in the number of individuals that has reduced the expected diversity (bottleneck); or it can be explained by the action of a balanced selection on a specific region.
- If Tajima's D < 0, this indicates that the observed diversity is less than the expected diversity and is due to a recent expansion of the population following a bottleneck.

In our case we have that the species *H. sbordonii* has a positive Tajima's D which indicates that it is most likely undergoing a reduction in population size, the opposite case is found in *H. semele* which has a negative Tajima's D indicating the current expansion following a bottleneck (Tajima 1989).

4.10.3 - Fixation index (Fst)

"The Fst reflects the joint effects of drift, migration, mutation and selection on the distribution of genetic variation among populations and it can be used to describe the distribution of genetic variation among any set of samples" (Holsinger et al. 2009). The Fst ranges between values 0 and 1 where a value of 0 indicates complete panmixing, i.e. the two populations interbreed freely, i.e. the genetic diversity within the two populations is equal. Conversely, a value of 1 indicates that all genetic variation is due to population structure, and that the two populations do not share any genetic diversity (Wright 1931). In our case, we obtained a value of 0.19027 as the average Fst along the genome. This value is relatively low indicating to us that the two species have not undergone a separation in the remote past, but indicates to us that they are very close evolutionarily and that there could be gene flow between the two species. The Fst calculation also makes it possible to assess whether regions in the genome are under selection. If there are regions of the genome with high Fst values, this may indicate that diversifying selection has taken place in that area, while regions with low Fst may indicate regions of stabilising

selection, which homogenized the two populations and makes them similar (Holsinger et al. 2009)

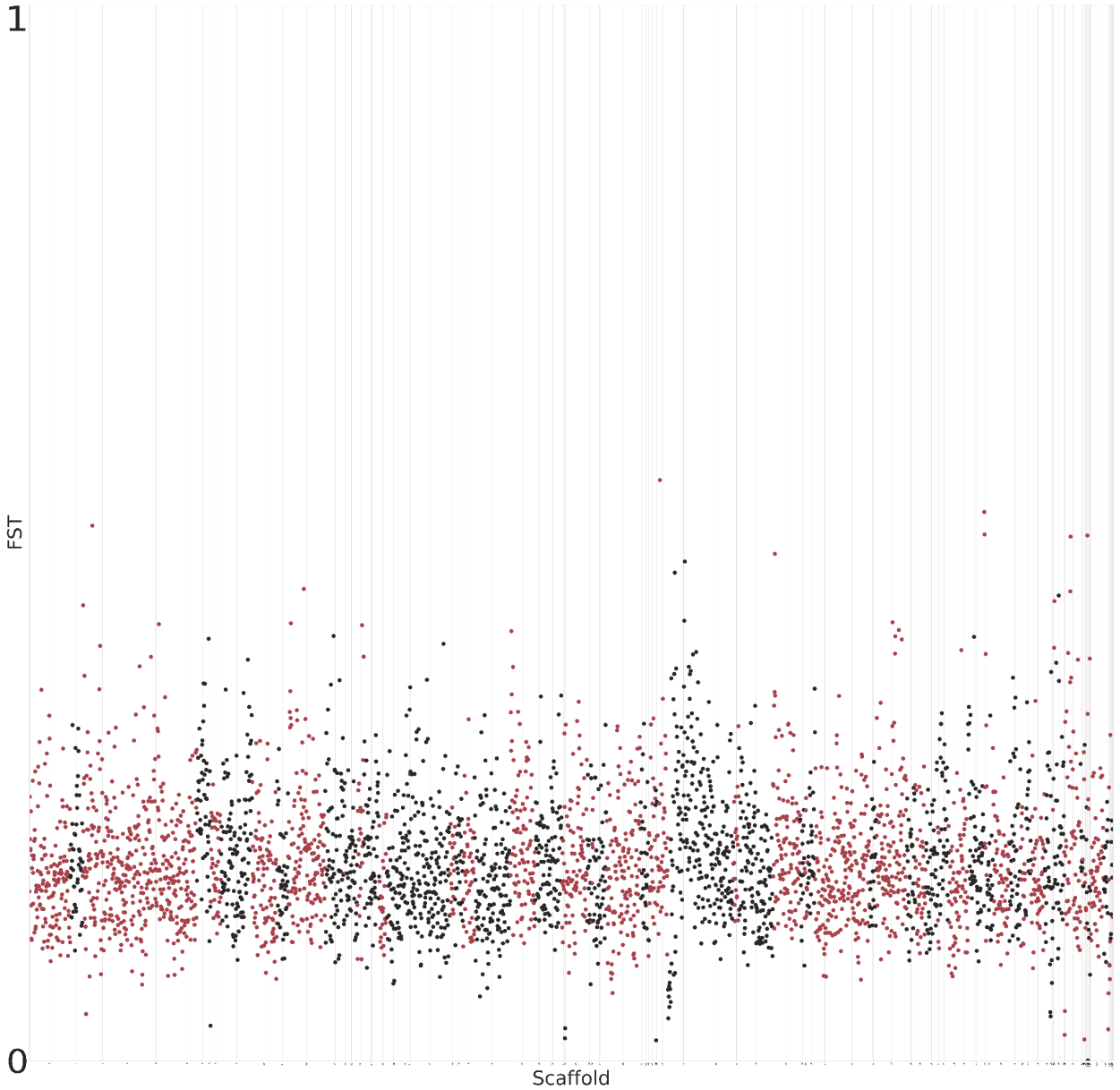

Figure 16: **Fst between H. sbodonii and *H. semele* calculated on the genome divided into 100kb windows which is represented by a single point.** We can see regions of higher Fst where it is possible that there is selection taking place in one of the two species.

5. DISCUSSION

Thanks to numerous collaborating consortia, reduced sequencing costs, reduced computational costs and increased computational capacity, more and more reference genomes are being published and deposited over the years (Hotaling et al. 2021).

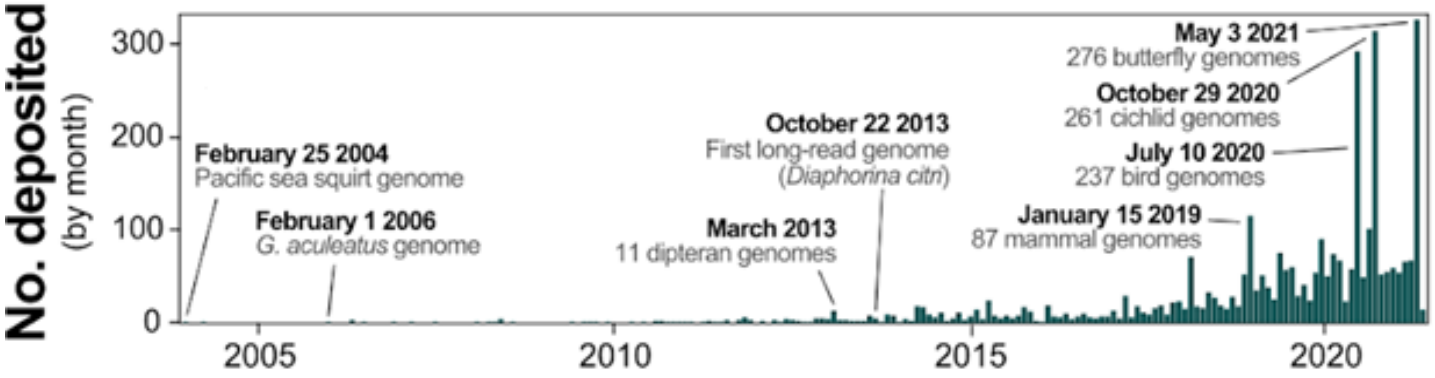

Figure 17: **The number of animal genome assemblies deposited in GenBank each month since February 2004 until August 2021** (Hotaling et al. 2021)

This trend towards increased deposition of reference genomes has increased following the emergence of third-generation sequencing. The use of long reads produced with TGS (third-generation sequencing) systems has allowed genomes to be assembled with much higher contiguity than using reads

produced with NGS (next-generation sequencing) systems of the order of up to 300-fold (Rhie et al. 2020).

The possibility of obtaining genomes of excellent quality in terms of contiguity, using long reads, without having to carry out numerous operations to improve contiguity, has enabled many laboratories, universities and consortia to deposit more and more genomes over the years (Hotaling et al. 2021).

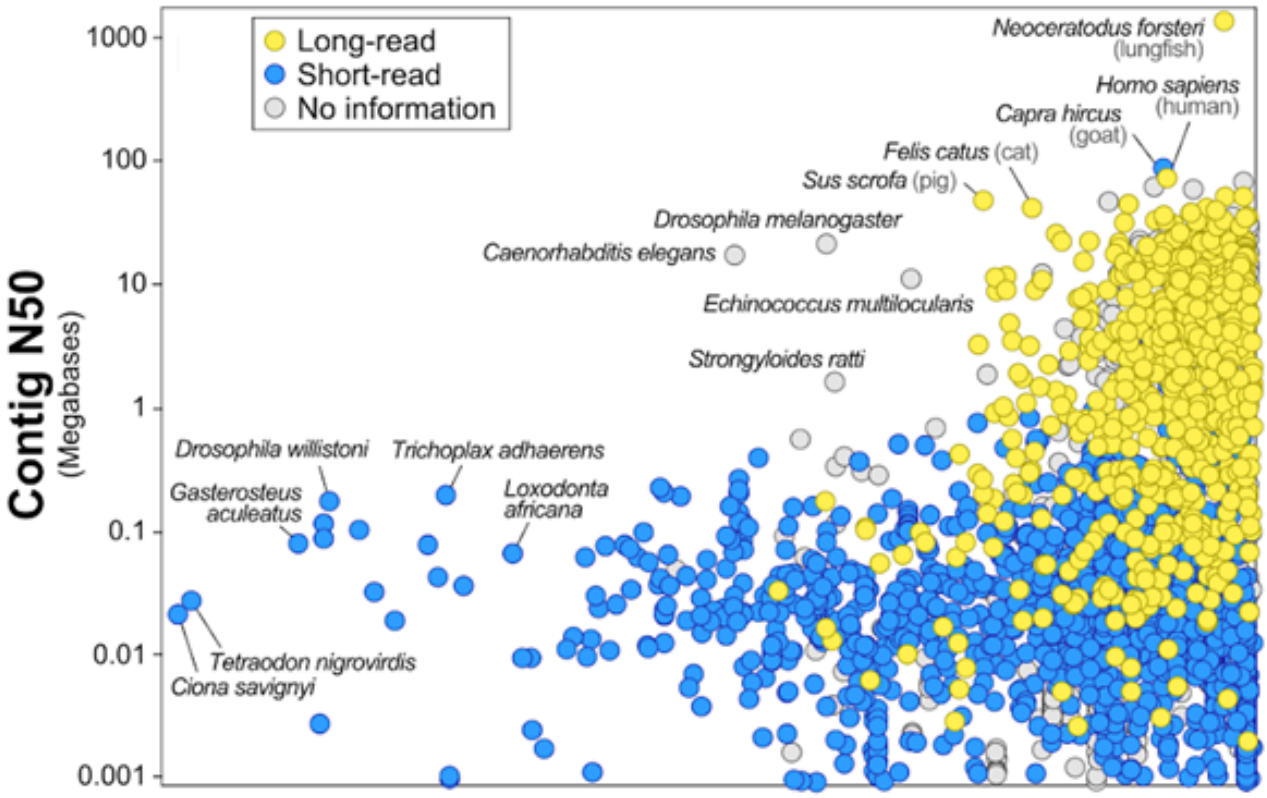

Figure 18: **The timeline of genome contiguity vs. short reads vs. long reads from February 2004 to August 2021** with the advent of the use of long reads the N50 and therefore continuity has increased (Rhie et al. 2020).

This increase in contiguity has made it possible to obtain genomes in which the N50 in the Contig reaches a size of up to 10 Mb and in the subsequent assembly steps to obtain scaffolds with a length coinciding with the chromosomes (Arang Rhie et al. 2020).

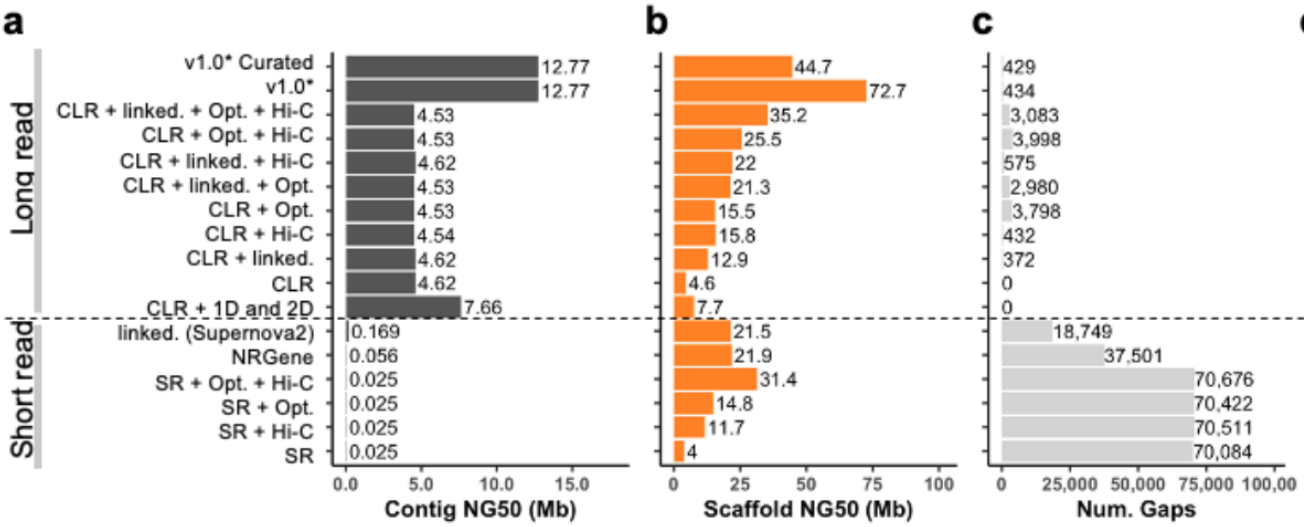

Figure 19: **Comparative analyses of genome assemblies with various data types from Anna's hummingbird** (Rhie et al. 2020). The graph shows us that the use of long reads leads to significantly larger contigs and scaffolds than the use of short reads, it can also be seen that with the use of short reads, a much larger amount of Gap is inserted than with long reads.

Therefore, the increasing deposition of reference genomes in databases together with the increasing quality of the deposited genomes has prompted various consortia, which deal with the production of reference genomes such as Vertebrate Genomes Project (VGP), Earth Biogenome Project (EBP) to name a few, to meet, determine and propose, the quality standards that must be achieved in order to deposit the genome produced. These quality standards proposed by the main consortia dealing with conservation genomics fall into several categories that allow us to quantify the quality of the assembly.

In many cases, when a large consortium is made up of numerous sub-consortia, projects and associations, it may be that the consortium itself decides and sets quality standards to be able to deposit the genome sequence in their databases. One of the consortia that have proposed these internal standards for depositing the genome in their database is the Vertebrate Genomes Project (VGP) consortium. The VGP has defined parameters divided into different quality categories: continuity, structural accuracy, base accuracy, haplotype phasing, functional completeness and chromosomal status, which in turn are divided into different quality measures. With this parameterised system, it is possible to place a genome within different classes in order of quality.

In particular, the VGP consortium has defined its own quality standards (Table 14) and a nomenclature that allows the genome to be placed, according to quality, in a class.

Quality Category	Quality Metric	Finished	7.C.Q50	6.7.Q40	4.5.Q30	VGP
Continuity	**Contig (NG50)**	= Chr. NG50	>10 Mbp	>1 Mbp	>10 kbp	1-25 Mbp
	Scaffolds (NG50)	= Chr. NG50	= Chr. NG50	>10 Mbp	>100 kbp	23-480 Mbp
	Gaps / Gbp	No gaps	<200	<1,000	<10,000	75-1500
Structural accuracy	**False duplications**	0%	<1%	<5%	<10%	0.2-5.0%
	Reliable blocks	= Chr. NG50	>90% of Scaffold NG50	>75% of Scaffold NG50	>50% of Scaffold NG50	2-75%
	Curation improvements	All conflicts resolved	Automated + Manual	Automated	No requirement	Automated + Manual
Base accuracy	**Base pair QV**	>60	>50	>40	>30	39-43
	k-mer completeness	100% complete	>95%	>90%	>80%	87-98%
Haplotype phasing	**Phased block (NG50)**	= Chr. NG50	>1 Mbp	>100 kbp	No requirement	1.6 Mbp*
Functional completeness	**Genes**	>98% complete	>95% complete	>90%	>80%	82-98%
	Transcript mappability	98%	>90%	>80%	>70%	96%
Chromosome status	**Assigned %**	98%	>90%	>80%	No requirement	94.4-99.9%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least 1 shared (e.g. X or Z)	Fragmented	At least 1 shared
	Organelles (e.g. MT)	1 Complete allele	1 Complete allele	Fragmented	No requirement	1 Complete allele

Table 14: **Proposed standards and metrics for defining genome assembly quality**

Since these categories of standards are highly descriptive of the quality of a genome, they have been adopted by several consortia, one of which has been

the Earth Biogenome Project (EBP), which includes the ENDEMIXIT project in which we have participated.

The Earth Biogenome Project (EBP) has set values of quality measures whereby if these are reached, the genome can be deposited.

The EBP propose a minimum reference standard of **6.7.Q40** (https://www.earthbiogenome.org/assembly-standards: EBP Assembly Standard version 2 - June 2020)**.** This code tells us that the genome must have the characteristic sequences:

- Contig with N50 > 1Mb **(6**.7.Q40**)**
- Scaffold with N50 > 10Mb **(**6.**7**.Q40**)**
- QV > 40 (less than 1/10,000 error rate) (6.7.**Q40)**
- > 90% kmer completeness
- sequence assigned to candidate chromosomal sequences (data not available)
- 98.1% single copy conserved genes (e.g. BUSCO) complete and single copy
- transcripts from the same organism mappable (data not available)

Our genome has values in line with the statistics required by the EBP because:

- Contig with N50 of 9.1Mb
- Scaffold with NG50 of 10 Mb
- QV 41
- 76.6% kmer completeness (if both primary and secondary are considered we have a Kmer completeness > 99%.)
- > 90% sequence assigned to candidate chromosomal sequences
- > 90% single copy conserved genes (e.g. BUSCO) complete and single copy

Our *H. sbordonii* genome assembly can then be considered of high quality.

If we then go to see the trend of continuity statistics in other assemblies by comparing, the continuity statistics (N50) obtained in our assembled genome with genome assemblies of other organisms we can see (Figure 20) that the results obtained with our pipeline defy the best statistics, both in the general comparison with genomes of different phylum and in the comparison with the

quality of Lepidoptera genomes that represent one of the genomes with greater continuity (Hotaling et al. 2021).

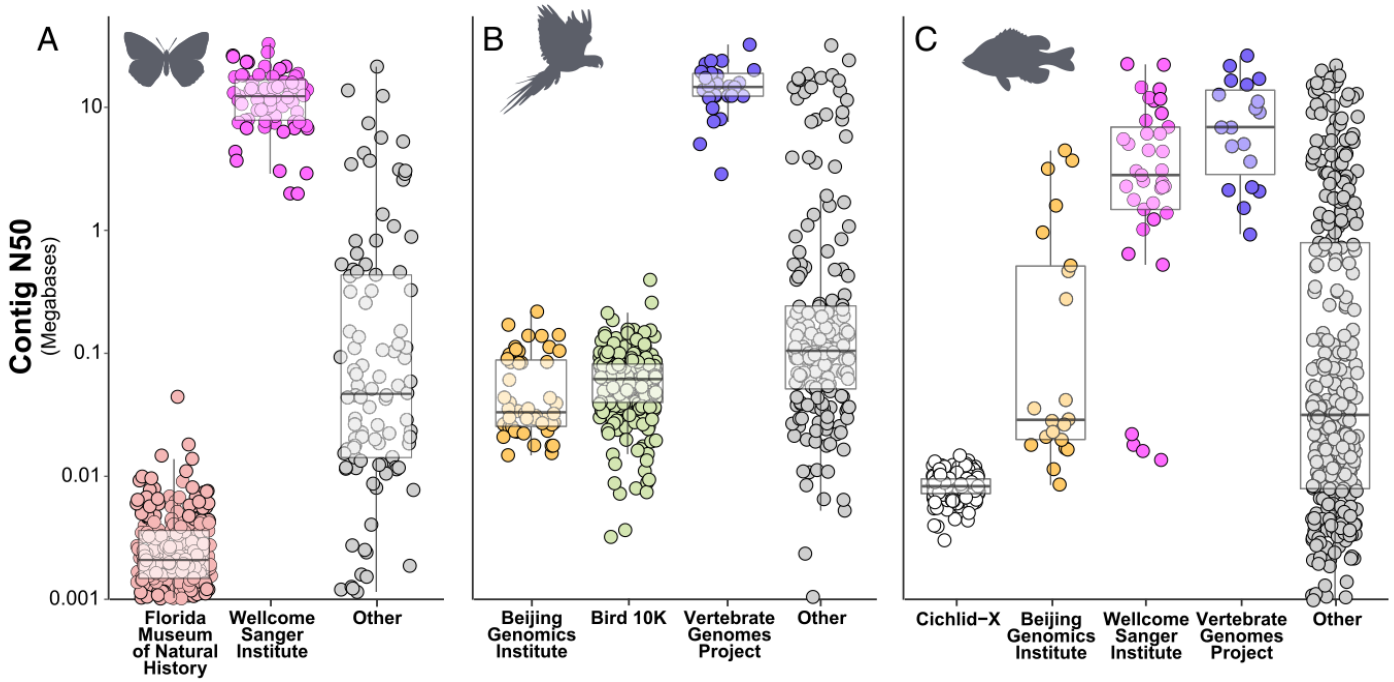

Figure 20: **continuity of genomes obtained from major contributors of genome assemblies for (A) butterflies (order Lepidoptera), (B) birds (class Aves), and (C) fish (mainly class Actinopterygii),** the order Lepidoptera represents one of the orders in which there are species with genomes assembled with better continuity statistics (Hotaling et al. 2021)

Despite the fact that our genome assembly has a continuity comparable with the best lepidopteran genome assemblies, there is the possibility of increasing it even further. As discussed in the additional results obtained by UniTS (**Supplementary information**), there are scaffolds that are separated in the *H. sbordinii* assemblage but that in some of the species studied, namely *Pararge aegeria, Maniola jurtina* and *Maniola hyperantus*, are found united in the same chromosome.

This information could help us to further increase the quality of the assemblage because if these *H. sbordoniii* scaffolds were really joined in the same chromosome in reality we could merge them and obtain an assemblage on a chromosome scale.

There are several possible approaches that can be used to assess the effective union of scaffolds in a single chromosome.

- Using a Hi-C technique (High Chromosome Conformation Capture) which allows us to reconstruct the contact zones within chromosomes and thus assess the proximity of sequences in the genome and thus assess whether two or more scaffolds are actually part of the same chromosome (Dekker 2005; Belton et al 2012)

- Carry out FISH (Fluorescence In Situ Hybridization) (in support of Hi-C data or even on its own) in which fluorescent probes are made from the sequences of the scaffolds and it is ascertained where the probes and therefore the scaffolds are located, e.g. if two probes made on the sequence of two scaffolds with a good length show a chromosome, it is very likely that those two scaffolds can be joined.

- Perform an ultralongPCR by constructing the primer pair on the basis of orthologous gene sequences that straddle two scaffolds, which in synteny analysis are instead united in one chromosome. If PCR produces an amplificate, the two scaffolds are located on the same chromosome. However, with the application of this technique, there is a high probability of false negatives and the possibility of carrying out only one PCR at a time, thus greatly increasing the workload when evaluating the union of several scaffolds (Shevchuk et al. 2004).

6. CONCLUSIONS

The main result we can deduce from the data produced by the various steps of the assembly is that we have obtained a reference genome in which the various quality statistics are in line with the standards proposed by the main consortia producing genome assemblies. The excellent quality of the assembly can also be deduced by comparing our genome with the assembled genomes deposited in the databases, where it can be seen that with regard to quality statistics based on continuity, our product is among the best deposited genomes.

In this work we have also determined a well-defined assembly pipeline using various bioinformatics tools that allowed us to obtain very good results. As already mentioned, however, the bioinformatics software used in some cases requires large computational and time resources, making the genome assembly work relatively expensive. However, as a fundamental resource for conservation biology, there is an increasing production and deposition of

assembled genomes thanks to the implementation of algorithms that improve the performance of software while requiring fewer computational resources.

Although we obtained a genome with good qualitative characteristics, the genome can be further increased in quality by carrying out comparative analyses with genomes of species that are evolutionarily close, so as to extrapolate information on the organization of the genome of these species and apply it to improve our genome and obtain a chromosome-scale assemblage. In addition, another fundamental step that could be carried out in further future assembly steps would be to obtain the mitochondrial genome sequence so that a complete genome could be deposited including the accessory sequences.

As far as the scope of the study of the *H. sbordonii* population is concerned, we have produced some preliminary results on the status of the *H. sbordonii* population. These estimates indicate, as is already known, that the population under study has a relatively small number of individuals, and in addition to this the population is undergoing a decline that could bring it towards the brink of extinction. However, with our estimates, we only began to scratch the surface of the information that can be deduced from an assembled genome

and that is fundamental to determining the actual status of the *H. sbordinii* population.

Therefore, with further study and investigation of the population structure, it will be possible to confirm or disprove the deductions we have made based on a limited number of population genetics estimates.

BIBLIOGRAPHY

1. Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., ... & Townshend, J. R. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. *Science advances*, *1*(2), e1500052.
2. Hoffmann, A. A., & Sgrò, C. M. (2011). Climate change and evolutionary adaptation. *Nature*, *470*(7335), 479-485.
3. Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., ... & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived?. *Nature*, *471*(7336), 51-57.
4. Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, *1*(5), e1400253.
5. Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by

vertebrate population losses and declines. *Proceedings of the national academy of sciences*, *114*(30), E6089-E6096.

6. Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome biology*, *19*(1), 1-12.
7. Masel, J. (2011). Genetic drift. *Current Biology*, *21*(20), R837-R838.
8. Lynch, M., Conery, J., & Burger, R. (1995). Mutation accumulation and the extinction of small populations. *The American Naturalist*, *146*(4), 489-518.
9. Bertorelle, G., Raffini, F., Bosse, M. *et al.* Genetic load: genomic estimates and applications in non-model animals. *Nat Rev Genet* (2022). https://doi.org/10.1038/s41576-022-00448-x
10. Stange, M., Barrett, R. D., & Hendry, A. P. (2021). The importance of genomic variation for biodiversity, ecosystems and people. *Nature Reviews Genetics*, *22*(2), 89-105.
11. Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in ecology & evolution*, *23*(1), 38-44.

12. Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in genetics*, *26*(4), 177-187.

13. Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature reviews genetics*, *11*(10), 697-709.

14. Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., ... & Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*.

15. Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., ... & Flicek, P. (2022). Ensembl 2022. *Nucleic acids research*, *50*(D1), D988-D995.

16. Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., ... & Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325-4333.

17. https://endemixit.com/

18. Cesaroni, D., Lucarelli, M., Allori, P., Russo, F., & Sbordoni, V. (1994). Patterns of evolution and multidimensional systematics in graylings (Lepidoptera: Hipparchia). *Biological Journal of the Linnean Society*, *52*(2), 101-119.

19. Sbordoni, V., Cesaroni, D., Coutsis, J., & Bozano, G. (2018). *GUIDE TO THE BUTTERFLIES OF THE PALEARCTIC REGION. Satyrinae part V.(Tribe Satyrini. Genera Satyrus, Minois, Hipparchia)*. Omnes Artes sas di M. Scala Minardi & C..

20. http://www.farfalleitalia.it/sito/910/index.php

21. Kudrna, O. (1984). On the taxonomy of the genus Hipparchia Fabricius, 1807, with descriptions of two new species from Italy (Lepidoptera, Satyridae). *Fragmenta entomologica*, *17*(2), 229-243.

22. https://www.iucnredlist.org/species/173231/64640021

23. Bonelli, S., Casacci, L. P., Barbero, F., Cerrato, C., Dapporto, L., Sbordoni, V., ... & Balletto, E. (2018). The first red list of

Italian butterflies. *Insect Conservation and Diversity*, *11*(5), 506-521.

24. Sbordoni, V., *Aspetti genetici ed ecologici del declino di popolazioni di farfalle e altri insetti*, in *Atti Accademia Nazionale Italiana di Entomologia*, LXVI, 2018, pp. 159-168.

25. Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). Why are de Bruijn graphs useful for genome assembly?. *Nature biotechnology*, *29*(11), 987.

26. Hozza, M., Vinař, T., & Brejová, B. (2015, September). How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. In *International Symposium on String Processing and Information Retrieval* (pp. 199-209). Springer, Cham.

27. Zhu, Z., Zhang, Z., Chen, W., Cai, Z., Ge, X., Zhu, H., ... & Peng, Y. (2018). Predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein. *Infection, Genetics and Evolution*, *61*, 183.

28. Perry, S. C., & Beiko, R. G. (2010). Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome biology and evolution*, *2*, 117-131.

29. Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in bioinformatics*, *5*(3), 237-248.

30. Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202-2204.

31. Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1-10.

32. Chor, B., Horn, D., Goldman, N., Levy, Y., & Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. *Genome biology*, *10*(10), R108.

33. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, *27*(5), 722-736.

34. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... & Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133-138.

35. Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology*, *21*(1), 1-27.

36. Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. In *Gene prediction* (pp. 227-245). Humana, New York, NY.

37. Eddy, S. R. (2004). What is a hidden Markov model?. *Nature biotechnology*, *22*(10), 1315-1316.

38. Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies [version 1; peer.

39. Ladunga, I. (2009). Finding homologs in amino acid sequences using network BLAST searches. *Current protocols in bioinformatics*, *25*(1), 3-4.

40. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.

41. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

42. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

44. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100.

45. Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896-2898.

46. Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., & Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. *BMC genomics*, *20*(1), 1-12.

47. Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, *27*(5), 737-746.

48. Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., ... & Zhang, Y. (2020). TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience*, *9*(9), giaa094.

49. Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

50. https://github.com/ekg/bamaddrg

51. “Picard Toolkit.” 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute

52. Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* O'Reilly Media.

53. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

54. Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, *8*(3), 186-194.

55. Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research*, *8*(3), 175-185.

56. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. *BioRxiv*.

57. https://vcftools.github.io/

58. Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*(10), 5269-5273.

59. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585-595.

60. Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, *10*(9), 639-650.

61. Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., ... & Paten, B. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, *21*(12), 2224-2241.

62. Simakov, O., Marlétaz, F., Yue, J. X., O'Connell, B., Jenkins, J., Brandt, A., ... & Rokhsar, D. S. (2020). Deeply conserved

synteny resolves early events in vertebrate evolution. *Nature ecology & evolution*, *4*(6), 820-830.

63. Zhu, B. H., Xiao, J., Xue, W., Xu, G. C., Sun, M. Y., & Li, J. T. (2018). P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC genomics*, *19*(1), 1-13.

64. Tatarinova, T. V., Chekalin, E., Nikolsky, Y., Bruskin, S., Chebotarov, D., McNally, K. L., & Alexandrov, N. (2016). Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific reports*, *6*(1), 1-12.

65. Subramanian, S. (2019). Population size influences the type of nucleotide variations in humans. *BMC genetics*, *20*(1), 1-12.

66. Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., ... & Jiggins, C. D. (2015). Estimation of the spontaneous mutation rate in Heliconius melpomene. *Molecular biology and evolution*, *32*(1), 239-243.

67. Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now?. *Proceedings of the National Academy of Sciences*, *118*(52).

68. Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737-746.

69. https://www.earthbiogenome.org/assembly-standards (EBP Assembly Standard version 2 - June 2020).

70. Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268-276.

71. Dekker, J. (2006). The three'C's of chromosome conformation capture: controls, controls, controls. *Nature methods*, *3*(1), 17-21.

72. Shevchuk, N. A., Bryksin, A. V., Nusinovich, Y. A., Cabello, F. C., Sutherland, M., & Ladisch, S. (2004). Construction of long DNA molecules using long PCR‐based fusion of several fragments simultaneously. *Nucleic acids research*, *32*(2), e19-e19.

SUPPLEMENTARY INFORMATION

As part of the ENDEMIXIT project, we have worked alongside several teams from different Italian universities. One of the interesting results produced by the University of Trieste concerns synteny analysis between the genome of *Hipparchia sbordonii* that we produced and the genome of butterfly species evolutionarily close to *H. sbordonii*.

Synteny analysis consists in searching for and assessing how much the distribution and orientation of gene loci in one species differs from another.

In this case the synteny analysis was carried out by comparing the genome of *H. sbordonii* that we obtained with neighboring species of butterflies.

In particular, these butterflies whose genomes had already been assembled on a chromosome scale and therefore had excellent contiguity, allowed us to see how the orthologous *H. sbordonii* gene loci contained in the assembled scaffolds were distributed along the chromosomes of these evolutionarily close butterflies.

The butterflies taken in the studio for the synteny comparison were:

- ***Pararge aegeria***

Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Parargina; Pararge

- ***Maniola jurtina***

Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Maniolina; Maniola; jurtina species complex

- ***Maniola hyperantus***

Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Maniolina; Maniola; Aphantopus

Of these butterflies, the closest species evolutionarily was *Pararge aegeria* while the others were more distant evolutionarily.

Synteny allows us to assess possible evolutionary relationships between organisms through comparative genomics studies. By comparing how genes are distributed along chromosomes in two species and making further comparisons with other species, it is possible to infer the relative evolutionary distance between the compared species (Simakov et al. 2020)

In this particular application by the team from the University of Trieste as part of the ENDEMIXIT project, the synteny between *H. sbordonii* and the above-mentioned species was used as an evaluation of the assemblage.

The very interesting approach employed by the UniTS team (University of trieste) was to first carry out a round of scaffolding with the P_RNA_scaffolder software that performs scaffolding using the transcriptomic data (RNA seq) of *H. sbordonii* so as to obtain a further increase in the contig of the genome (Zhu et al. 2018).

In fact as can be seen from the table below the contigs/scaffolds have been reduced to 99 and have N50 values of more than 10Mb and L50 of 15.

after scaffolding with P_RNA scaffolder	
Number of contigs:	**99**
Total size (bp):	**397096441**
N50 (bp):	**10019938**
L50:	**15**
N90 (bp):	2692159
L90:	44
Mean contig size (bp):	4011075
Longest contig (bp):	20430989
Third quartile (bp):	5906264
Median (bp):	1992107
First quartile (bp):	222519
Shortest contig (bp):	23432
Number of Ns:	1100
Number of gaps (/N+/):	11

Table S1: **statistics of FASTA files after Scaffolding with transcriptomic data (RNA data)**

Once they had obtained scaffolds with good contiguity, the UniTS team searched the genome of the above-mentioned species for orthologous genes with *Hipparchia sbordonii* (using software to search for orthologous genes such as BUSCO), i.e. genes present in different organisms that code for the same protein product and are evolutionarily derived from common ancestral lines.

After extrapolating the orthologous genes from the different species, they carried out a comparative analysis between some chromosomes of the above-mentioned butterfly species and the genome of *H. sbordonii* to see where the orthologous genes of these butterflies were distributed in the scaffolds obtained from the genome construction.

Knowing that there were fairly close evolutionary relationships between the examined butterflies (particularly *Pararge aegeria*) and *H. sbordoni*, they inferred that there were close syntenic relationships, i.e. that the order of genes in the examined butterflies was very similar, with a good probability, in *H. sbordonii*.

In practice, they wanted to see if there were any chromosome regions that in the species under analysis were gathered in a single chromosome but in *H. sbordoni* were not on a single scaffold but were fragmented into several scaffolds so that they could eventually join scaffolds (mapped on the same chromosome) and obtain an assembled genome on a chromosome scale.

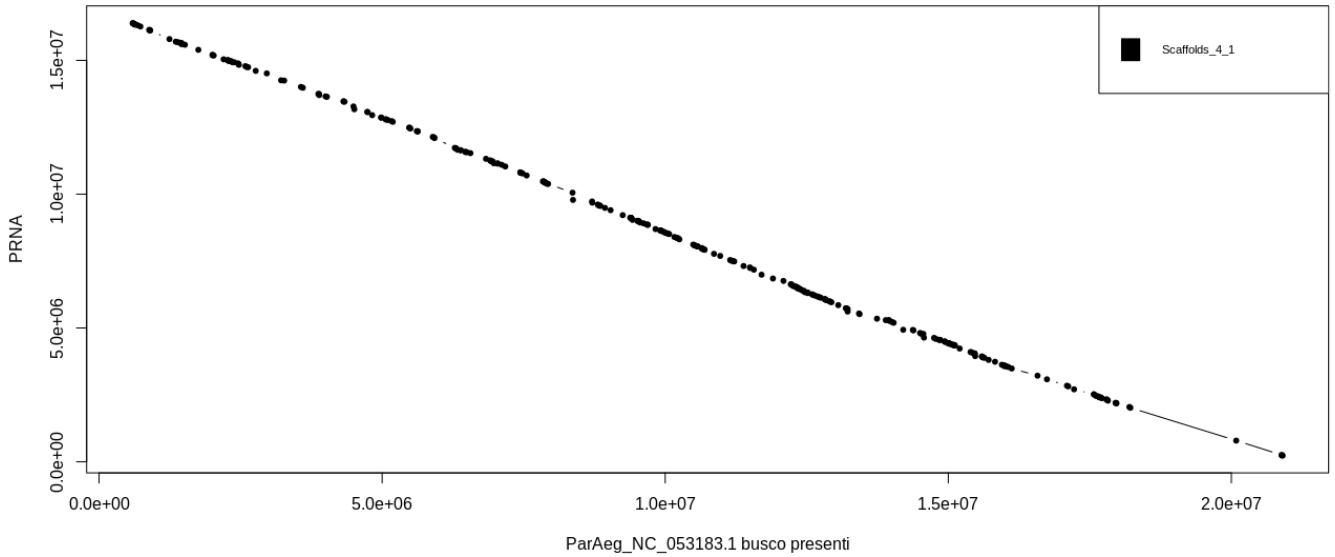

Figure S1**: synteny between chromosome 4 of *Pararge aegeria* (X-axis) and scaffold 4 of *H. sbordoni* (Y-axis)** Although the graph has a negative slope (due to the different assembly modes and resolved by implementing reverse-complement on the H. sbordinii scaffold) there is an almost perfect synteny which tells us that scaffold 4 of *H. sbordonii* is chromosomally scaled and most likely represents a single chromosome.

In the previous case (Figure S1) we have a complete correspondence of the scaffold with the chromosome of the species under analysis, in other cases we have that the chromosome is divided between 2 or more scaffolds while in others it is possible to see chromosomal rearrangements which are evidenced by stretches of the line of the graph with opposite slope (inversions) or in other cases we can see regions that deviate from the line of the graph attributable to possible translocations (Figure S2).

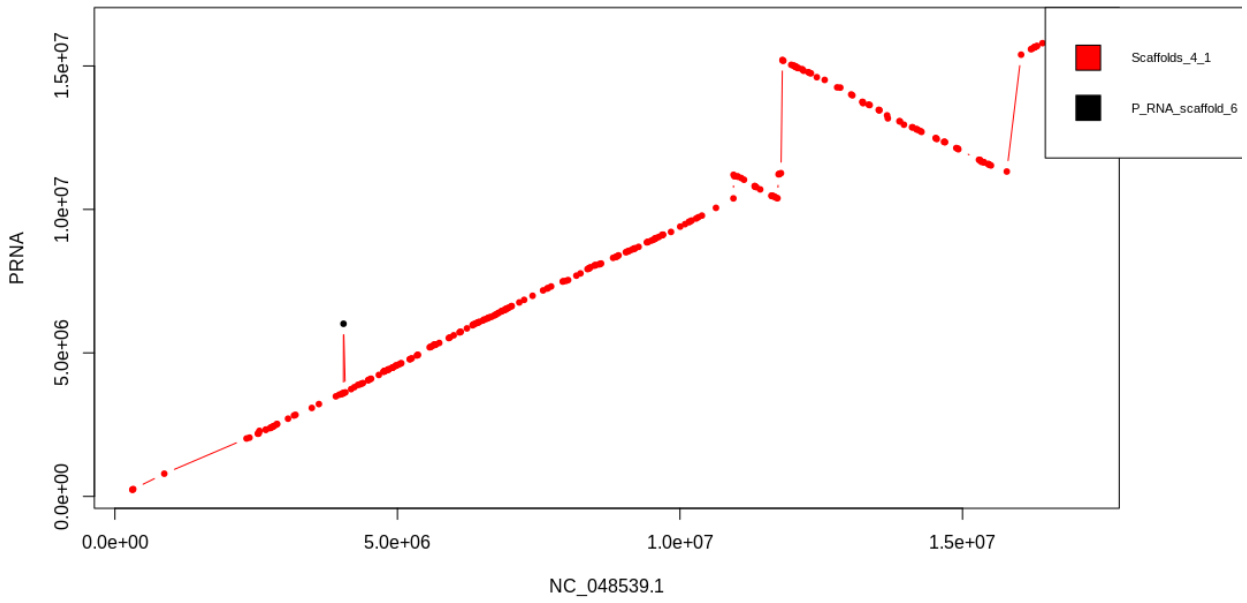

Figure S2: **synteny between chromosome 4 of *Maniola hyperantus* (X-axis) and scaffold 4 and scaffold 6 of *H. sbordonii* (Y-axis):** The graph shows two regions that have undergone inversion in *H. sbordonii* and a small region (in black) of scaffold 6 that is found in chromosome 4 of *Maniola hyperantus*, indicating a possible translocation between chromosomes.

With this type of analysis it is also possible to reconstruct with good reliability any rearrangements on a chromosomal scale, since by assessing the synteny between two species it is possible to see whether there have been any translocations, inversions or deletions by observing the distribution of genes in the species under analysis.

In my opinion, the most interesting cases that allow us to devise practical approaches to improving the quality, in terms of contiguity, of the genome occur when the chromosome of the species under analysis is represented by the union of 2 or more scaffolds of *H. sbordonii*.

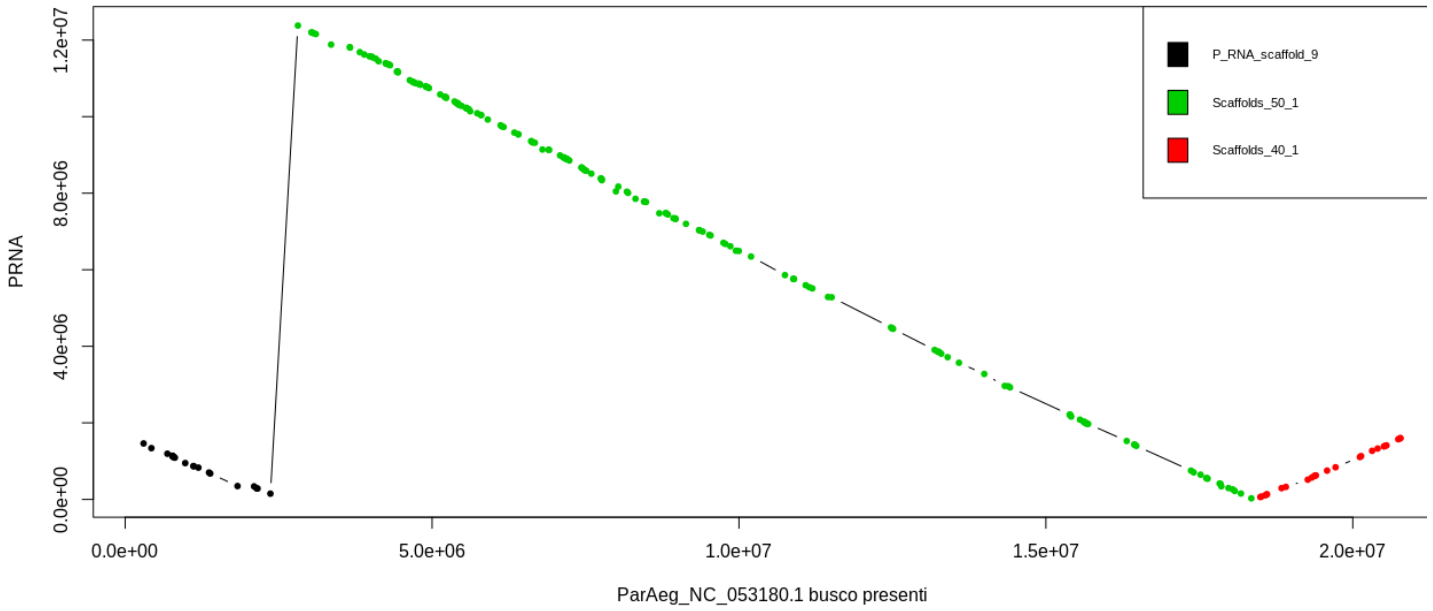

Figure S3: **Synteny between chromosome 1 of *Pararge aegeria* (X-axis) and scaffold 40, scaffold 50 and scaffold 9 of *H. sbordonii* (Y-axis):** in this case we see that chromosome 1 of *Pararge aegeria* can be described by the union of several scaffolds of *H. sbordonii* and this suggests that there may have been a chromosomal rearrangement in one of the two species or that most likely the scaffolds of *H. sbordonii* may be joined together to give a longer scaffold representing an entire chromosome.

In the last case, the UniTS team proposed different approaches to assessing whether such scaffolds can actually be joined together, which are based on assessing the actual joining of scaffolds in a chromosome in reality.

RINGRAZIAMENTI

Arrivato alla fine di questo cammino vorrei ringraziare chi mi ha aiutato e accompagnato in questo viaggio. Per iniziare vorrei ringraziare la mia famiglia che mi ha sempre incoraggiato, supportato, spronato e mi ha aiutato nei momenti bui ed in particolare vorrei ringraziare i miei genitori che mi hanno dato innanzitutto la possibilità di iniziare e proseguire fino alla fine questo percorso ma anche mi hanno aiutato a capire che i limiti che ho sono quelli che mi impongo aiutandomi a superarli. Un ringraziamento speciale va alla mia ragazza Gioia che in questi anni della magistrale mi è stata sempre vicino, mi ha aiutato a credere in me stesso e nelle mie capacità aiutandomi a superare ostacoli all'apparenza insormontabili. Sono molto felice di aver condiviso questo viaggio con una persona speciale come te. Per quanto riguarda il progetto di tesi non posso che ringraziare il Professor Emiliano che si è sempre messo a disposizione dandomi consigli e suggerimenti con un enorme dose di pazienza nei vari momenti della tesi. Vorrei anche rivolgere un enorme ringraziamento a tutto il gruppo di dottorande, dottorandi e tesisti che mi hanno aiutato ad approcciarmi alla bioinformatica, mi hanno aiutato nella realizzazione di plot ed hanno contribuito alla produzione di alcuni dati

ma soprattutto per aver trovato in questo gruppo delle persone stupende sempre disponibili all'aiuto senza mai volere nulla in cambio, vi ringrazio moltissimo per questo per questa esperienza e ve ne sarò sempre grato. Ringrazio anche tutti i Professori, dottorandi e studenti del progetto ENDEMIXIT per avermi dato questa opportunità di poter contribuire direttamente al progetto, in particolare vorrei ringraziare il Professor Giorgio Bertorelle per avermi fatto partecipare al progetto, il Professor Andrea Benazzo che mi ha suggerito degli approcci e delle indicazioni per svolgere determinate analisi come anche tutti i Professori e dottorandi dell'UniTS che partecipano al progetto per aver condiviso i dati delle loro analisi. Ringrazio anche moltissimo i miei nonni che mi hanno sempre ascoltato mentre parlavo di scienza ascoltando come dei bambini stupiti aiutandomi a comprendere il ruolo della scienza in questa società e quanto essa possa essere affascinante. Un grande grazie va a mia zia Alessandra per esserle stato sempre nei suoi pensieri e per aver speso sempre una preghiera prima degli esami e dei momenti critici del percorso universitario, Grazie Zia. Ringrazio infine i miei amici che sono stati sempre presenti e per avermi fatto svagare e sfogare nei momenti difficili ed aver condiviso e costruito insieme momenti felici che resteranno per sempre nel mio cuore. Un ringraziamento infine va a tutti

coloro che direttamente e indirettamente hanno contribuito a rendere questo viaggio indimenticabile e per avermi aiutato a crescere come persona. Mi sento estremamente fortunato ad aver potuto fare questo incredibile viaggio con tante persone così speciali e spero di poter festeggiare insieme tanti altri traguardi importanti della mia vita.

Sebastiano