UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

New bioinformatic pipeline to expand the knowledge of the chloroplast genome of Mortiño (*Vaccinium floribundum* Kunth.)

.

Sebastián Jordán Dávalos

Ingeniería en Biotecnología

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Biotecnología

Quito, 20 de diciembre de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

New bioinformatic pipeline to expand the knowledge of the chloroplast genome of Mortiño (*Vaccinium floribundum* Kunth.)

Sebastián Jordán Dávalos

Nombre del profesor, Título académico María de Lourdes Torres, PhD.

Quito, 20 de diciembre de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: Sebastián Jordán Dávalos

Código: 00212828

Cédula de identidad: 1722161989

Lugar y fecha: Quito, 20 de diciembre de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

ABSTRACT

Vaccinium floribundum (Mortiño) is an herbaceous species belonging to the *Vaccinium* genus and the Ericaceae family. The chloroplast genome of this species was previously published by Rojas et al (2023). Yet, modifications in the bioinformatic pipeline and the amount of *Vaccinium* chloroplast sequences used in the phylogenetic analysis could help elucidate new information regarding the chloroplast genome of *V. floribundum.* Therefore, the project's objectives were to implement a new bioinformatic pipeline to reassemble the chloroplast genome of mortiño and incorporate 3 new *Vaccinium* chloroplast sequences to verify the taxonomic relationships between *V. floribundum* and other *Vaccinium* species. Genome quality checks (QUAST, depth coverage graphs and assembly graphs) demonstrated that the ptGAUL bioinformatic pipeline generated assembly was the most suitable compared to the other 3 assemblers used (Canu, Wtdbg2, and Flye). Annotation and manual curation of the ptGAUL assembly identified the same 134 genes and 6 pseudogenes obtained by Rojas et al (2023). The phylogenetic analysis verified the relationship between *V. floribundum* and *V. myrtillus* with an increased boot strap (BS) support of 85%. Structural analysis showed similar patterns among all analyzed sequences, region sizes identified in this study varied slightly from the ones described by Rojas et al (2023) due to a difference in obtained reads. In conclusion, this study verifies the results obtained by Rojas et al (2023) and increased quality parameters such as a higher BS support in the phylogeny. Moreover, this study is important as it provides a useful genomic tool to continue studying *V. floribundum* as well as other *Vaccinium* species.

Key Words: *Vaccinium floribundum*, chloroplast genome, bioinformatic pipeline, ptGAUL

RESUMEN

Vaccinium floribundum (Mortiño) es una especie herbácea perteneciente al género *Vaccinium* y a la familia Ericácea. El genoma del cloroplasto de esta especie fue publicado por Rojas et al (2023). Sin embargo, Modificaciones en pipeline bioinformático y la cantidad de secuencias de cloroplasto de *Vaccinium* utilizadas en el análisis filogenético podrían ayudar a elucidar nueva información sobre el mismo. Por lo tanto, los objetivos del proyecto fueron: implementar un nuevo pipeline para volver a ensamblar el genoma del cloroplasto del Mortiño e incorporar 3 nuevas secuencias de *Vaccinium* para verificar las relaciones taxonómicas entre *V. floribundum* y otros *Vaccinium*. Las verificaciones de calidad de los genomas (QUAST, gráficos de cobertura y gráficos de ensamblaje) demostraron que el ensamblaje obtenido del pipeline de ptGAUL fue el más adecuado en comparación a los 3 ensambladores utilizados (Canu, Wtdbg2 y Flye). La anotación y la curación manual del ensamblaje identificaron los mismos 134 genes y 6 pseudogenes obtenidos por Rojas et al (2023). Además, el análisis filogenético verificó la relación entre *V. floribundum* y *V. myrtillus* e incremento el soporte de rama a 85%. El análisis estructural mostró patrones similares entre todas las secuencias analizadas. Sin embargo, los tamaños de las regiones en este estudio variaron ligeramente a los descritos por Rojas et al (2023), probablemente debido a una diferencia en las lecturas obtenidas. En conclusión, este estudio verifica los resultados obtenidos por Rojas et al (2023) y demuestra un mayor soporte en la filogenia. Además, esta investigación es de gran importancia ya que provee una herramienta genómica útil para seguir estudiando *V. floribundum* y a otra especies de *Vaccinium.*

Palabras clave: *Vaccinium floribundum*, genoma del cloroplasto, pipeline bioinformático, ptGAUL.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

1.1 General Information

1.1.1 Distribution, environment, and characteristics

Vaccinium floribundum Kunth, commonly known as Mortiño, is an herbaceous shrub species that belongs to the Ericaceae family, which has an estimated number of 4,250 species, characterized by their berry-like fruits (Christenhusz & Byng, 2016; Martău et al., 2023). *V. floribundum* is distributed along the Andean Mountain range specifically in countries like Ecuador, Venezuela, Colombia, Bolivia and Perú (Coba et al., 2012). In Ecuador, *V. floribundum* naturally occurs in the paramo ecosystem which is characterized by an altitude range of 3,000-5,000 m.a.s.l, cold and humid climates (3-17°C) and a small, shrub-like, vegetation (Mena et al., 2011; Meléndez-Jácome et al, 2021).

Morphologically, *V. floribundum* is characterized as a densely branched shrub that can grow up to 2.5 m tall. It has small, serrated, oval-shaped leaves that are pink-colored when emerging and turn green as they mature (Llivisaca-Contreras, 2022). Mortiño plants contain racemes of small purple flowers that are approximately 1 cm in diameter (Luteyn & Pedraza-Peñalosa, 2012). These flowers produce a round berry with a ranging diameter of 5 to 8 mm of blue or lilac color, containing 15-60 seeds and high contents of sugars, antioxidants, minerals, and vitamins (Meléndez-Jácome et al, 2021; Coba et al., 2012; Llivisaca-Contreras, 2022).

1.1.2 Social, Economic and Ecological Importance

V. floribundum holds an important social and economic value in the Ecuadorian culture since its berry is used for the elaboration of traditional foods and beverages (Torres et al., 2010; Aguilar, 2009). However, collection and distribution are fully dependent on natural occurrence, as the species has not been domesticated (Llivisaca-Contreras, 2022; Aguilar, 2009). Besides *V. floribundum* socio-economic importance, it also fulfills important ecological roles. For example, *V. floribundum* shrubs are one of the first species to recover from anthropogenic

activities (such as slash and burn agriculture) in the Ecuadorian paramo. Therefore, it plays a key role in the remediation of its environment (Aguilar, 2009; Llivisaca-Contreras, 2022; Caranqui-Aldaz et al., 2022).

1.2 Genome sequencing, annotation, and analysis

Genome sequencing technologies have become essential for the study of living organisms, as they can be used to elucidate knowledge about their taxonomy, evolution, genetic structure, among others (Henry, 2022). Although multiple technologies can be applied for a variety of sequencing projects, Oxford Nanopore Sequencing Technology (ONT), a third-generation sequencing technology, is commonly used for genome sequencing. This is due to its relatively simple procedure, reduced time for library construction and improvement of *de novo* assemblies due to the use of long reads (Athanasopoulou, 2022).

It relies on the extraction of long DNA chains, known as high molecular weight DNA, from the subject of interest (Wang et al., 2021). This sample is loaded on a flow cell that contains pores that are going to do the sequencing. As the DNA migrates through the pore, nucleotide base pairs are identified with a technique called “base calling”, which detects small changes in the ionic which then matches the change to a specific DNA base. (Zhang et al., 2020; Oxford Nanopore Technologies, n.d.).

When coupled with bioinformatic tools such as a genome assembler, the raw reads can be assembled into a coherent genome (Choudhuri, 2014). These technologies have made the annotation and assembly of multiple species possible, including the chloroplast genome of *V. floribundum* (Abril & Castellanos, 2019; Rojas et al., 2023).

1.3 The chloroplast

1.3.1 General characteristics

The chloroplast is a photosynthetic organelle that houses biochemical reactions that turn light energy into chemical energy. Therefore, it can be found in photosynthetic organisms

such as plants and algae (Roston et al., 2018). The chloroplast also contains its own genome that produces key proteins that regulate and drive photosynthesis among other functions. It has been hypothesized that the chloroplast has its own genome because it was once an independent unicellular organism (Campbell & Reece, 2007).

The chloroplast genome size can vary between species but are generally small with base pair (bp) counts ranging from 150,000-220,000 bp. Moreover, it is organized into 4 characteristic regions: Long single copy region (LSC), Short single copy region (SSC) and 2 Inverted repeats regions (IRA and IRB) (Contreras-Díaz et al., 2022).

1.3.2 Importance in genomic studies

Chloroplast DNA sequences have been widely utilized in evolutionary studies replacing nuclear DNA due to its unique characteristics. Nuclear DNA is large, complex, and more difficult to assemble. In contrast, chloroplast genomes (chloroplast DNA) are smaller, simpler and contain regions that mutate at different rates (Provan et al., 2001). Regions that accumulate mutations more rapidly can be used for the comparison between closely related organisms. Conversely, regions that accumulate mutations slowly serve for the comparison between distantly related organisms (Provan et al., 2001). The different rates of change of the previously mentioned regions are influenced by a number of factors such as lack of recombination due to maternal inheritance of the chloroplast, content of minimal genes and the presence of repeated regions (Provan et al., 2001; Zhang & Ren, 2015). These considerations make the chloroplast genome a more dynamic and versatile tool for better understanding taxonomic and evolutionary relationships of *V. floribundum* (Rogalski et al., 2015).

1.4 Previous studies on the chloroplast genome of *V. floribundum*

The chloroplast genome of *V. floribundum* was recently published in 2023. This study described a 187, 966 bp genome that contained 134 genes. Their taxonomic analysis compared

87 orthologous genes with other *Vaccinium* species and revealed that *V. myrtillus* could be a sister group for *V. floribundum* (Rojas, et al., 2023).

1.5 Current study on the chloroplast genome of *V. floribundum*

Despite the existing study describing the assembly, annotation, and analysis on the chloroplast genome of *V. floribundum*, different methodological approaches, such as the bioinformatic pipeline of the chloroplast genome assembly and the assembler used, can generate new results. Moreover, it would be interesting to analyze whether or not the new information of three new *Vaccinium* chloroplast genomes published recently modifies the taxonomic relationships described in the study by Rojas et al (2023).Therefore, the present study had 2 main objectives: first, utilize a new approach to assemble a new *V. floribundum* chloroplast genome to determine whether a new assembly can provide additional information to what is already known and secondly, incorporate 3 new *Vaccinium* chloroplast genomes in a phylogeny to verify the existing taxonomic relationships between *V. floribundum* and other species of *Vaccinium*.

2. METHODS

2.1 Sample collection, genomic DNA extraction, and long read sequencing

The same samples were used as the ones obtained and processed by Rojas et al (2023) which included: collection young leaves from *V. floribundum* in Lloa, Pichincha under the MAE-DNB-CM-2016-0046-M-0002 permit, high molecular weight DNA extraction and library. Long read sequencing was done with two R.9.4.1 flow cells in the MinION sequencers (Rojas et al., 2023).

2.2 Long read processing

Removal of adaptor sequences (with Porechop v.0.2.4) (Wick, 2017) and low-quality reads (with Nanofilt v.2.8.0) (De Coster et al., 2018) was executed as described by Rojas et al (2023).

2.3 Chloroplast genome assembly bioinformatic pipeline

2.3.1 Chloroplast reads extraction

The proposed bioinformatics pipeline (Figure 1) was developed based on the *Acacia pycnantha* chloroplast genome assembly methodology described by Syme et al (2021) (Appendix 3). Blasr v.5.3.5 (Chaisson & Tesler, 2012) was used to extract chloroplast reads by aligning them to nine *Vaccinium* chloroplast genomes (Appendix 1) used as baits to isolate *V. floribundum* chloroplast genome-associated reads.

2.3.2 Assembly and polishing

Chloroplast reads were then assembled using 3 different assemblers, Flye v.2.9.1 (Kolmogorov et al., 2019), Canu v.2.2 (Koren et al., 2017), and Wtdbg2 v.2.5 (Ruan & Li, 2020), as well as the ptGAUL v.1.0.5 bioinformatic pipeline (Zhou et al., 2023). The assemblers and the ptGAUL pipeline were run with standard parameters and with an estimated chloroplast genome size of 180 Kb, based on the results obtained by Rojas et al (2023). Polishing was done with the Apollo v.2.0 software. A single run was used as the program

developers suggest only a single run is necessary to obtain a polished assembly (Firtina et al., 2020).

The resulting assemblies were used as baits in a second run of the proposed bioinformatic pipeline to repeat chloroplast read extraction. This was done to extract chloroplast reads from the initial dataset with increased accuracy (Syme et al., 2021). The obtained reads were assembled and polished as mentioned in the previous paragraph, providing four final chloroplast assemblies. QUAST v.5.2.0 (Mikheenko et al., 2018) was used to obtain quality statistics from each assembly for both runs of the proposed bioinformatic pipeline. The *V. macrocarpon* chloroplast genome was used as a reference as it is one of the most studied *Vaccinium* species, therefore, more reliable information is available for this species than other *Vaccinium* (Fahrenkrog et al., 2022). Statistics of interest included: Genome size, NGA50 and indels per 100,000 bp. These statistics helped elucidate assembly characteristics such as the length of the genome and structural variations (Mikheenko et al., 2018).

2.4 Genome evaluation

The 4 assembled genomes were assessed to identify the most reliable assembly to use in further analysis following the steps described by Rojas et al (2023). To illustrate the support provided by chloroplast reads, a coverage graph was generated for each assembly following the script developed by Rojas et al., (2023). Assembly graphs were generated for the assemblies that provided gfa files (Flye and ptGAUL assemblies). Bandage v.0.9.0 (Wick et al., 2015) was used to generate assembly graphs and analyze how the genomes were organized and whether they had a defined structure (Formenti et al., 2022). The ptGAUL assembly was chosen for further analysis due to its QUAST statistics, resolved structure and consistent read coverage.

2.5 Annotation and manual curation

The selected ptGAUL assembled genome was annotated with the online organelle annotator GeSeq v.2.0.3 (Tillich et al., 2017). Next, the manual curation of the genome

annotation was executed using Geneious bioinformatics software v2022.2.2 (Kearse et al., 2012). All annotated genes were mapped with genes extracted from reference *Vaccinium* chloroplast genomes to verify proper annotation. Reannotated genes were analyzed based on their coding sequence (CDS). Genes that presented incomplete CDS, indicated by the presence of random stop codons, were removed. Finally, pseudogenes were annotated by the alignment of pseudogenes obtained from reference chloroplast genomes. OGDRAW v.1.3.1 was used to visualize the final genome annotation and structure (Greiner et al., 2019).

2.6 Phylogenetic analysis

For the phylogenetic analysis, the same 15 *Vaccinium* chloroplast genome sequences used by Rojas et al (2023), were obtained from the National Center for Biotechnological Information (NCBI) and incorporated in the analysis. In addition, the recently published chloroplast genomes of *V. ashei* (NC_071868.1), *V. microcarpum* (OQ865186.1), and *V. oxycoccos* (OQ865185.1) were included (last accessed august 20, 2023). Finally, two *Actinidia* chloroplast genomes were used as outgroups to root the phylogeny. All 21 chloroplast genomes (Appendix 2) were analyzed using PhyloHerb v1.1.2 (Cai & Davis, 2022) which extracted and concatenated 87 orthologous genes, then a maximum likelihood phylogeny was generated using RAxML v8.2.11 (Stamatakis, 2014) as described by Rojas et al (2023). The resulting phylogeny was visualized using FigTree v.1.4.4 (Rambaut, n.d.).

2.7 Structural analysis

IRscope (Amiryousefi et al., 2018) was used to compare genome structures between *V. floribundum* and other *Vaccinium* sequences. Only 10 *Vaccinium* species were used (Appendix 2), including the published *V. floribundum* (NC_073583.1) chloroplast genome because IRscope had a maximum capacity of 10 chloroplast genomes for analysis. The selected genomes were those most closely related to *V. floribundum*.

3. RESULTS

3.1 *V. floribundum* chloroplast genome assembly and annotation

3.1.1 Genome evaluation

QUAST results from the first run of the proposed bioinformatic pipeline (Table 1) for the Canu, Flye and Wtdbg2 assemblies demonstrated genome sizes of 199,051 bp, 189,001 bp, and 183,271 bp, respectively. These results differed from the expected genome size (187,966 bp) obtained by Rojas et al (2023). NGA50 alignments varied considerably among assemblies with the shortest alignment corresponding to the Wtdbg2 assembly (103,963 bp) and the longest to the Flye assembly (104,585 bp). The highest insertion and deletion (indel) rates were found in the Canu assembly with 783 events per 100,000 bp, while the Flye assembly had the lowest rate with 501.5 events per 100,000 bp. In comparison, QUAST statistics for the ptGAUL bioinformatic pipeline assembled genome demonstrated a higher quality sequence compared to the assemblers, as its genome size (187,879 bp) better approximated the expected genome size (187,966 bp) obtained by Rojas et al (2023). Moreover, its NGA50 alignment (104,617) was the longest of all assemblies. Yet, the indel rates for the ptGAUL assembly were the highest with 783 events per 100,000 bp.

After the second run of the proposed bioinformatic pipeline, QUAST results for each assembler (Canu, Wtdbg2, and Flye) (Table 2) mostly improved, except for the indel events that increased considerably. Genome sizes varied from 172,053 bp to 187,797 bp. Of these, the Flye assembly (187,797 bp) came closest to the expected genome size (187,966 bp) obtained by Rojas et al (2023). NGA50 alignments showed various alignment lengths, with the shortest alignment corresponding to the Wtdbg2 assembly (176,258 bp) and the longest corresponding to the Flye assembly (104,426 bp). Indel rates were highest in the Flye assembly with 997 events per 100,000 bp, meanwhile the Canu assembly had the lowest rate with 613.21 events per 100,000 bp. In contrast, QUAST results for the second run of the bioinformatic pipeline

for the ptGAUL assembled genome showed the closest genome size (187,892 bp) to the expected genome (187,966 bp) obtained by Rojas et al (2023). Moreover, the NGA50 alignment was the longest of all assemblies (104,600 bp) and the indel rates were the second lowest (764).

Depth coverage graphs depicted how the obtained reads supported each genome assembly and were compared to the coverage graph obtained by Rojas et al (2023) (Figure 2). The depth coverage graph obtained for the Canu assembly (Figure 2a) demonstrated low read support at the beginning of the assembly, a homogenous coverage ranging from 1,000 bp to 115,000 bp, a sudden increase in coverage from 115,000 bp to 130,000 bp, and finally a considerable decrease in coverage from 130,000 bp to the end of the assembly. The Wtdbg2 depth coverage graph (Figure 2b) showed higher support for the beginning of the assembly. Yet, a sudden increase at 110,000 bp and considerable decrease from 135,000 bp onward revealed a similar trend as the Canu depth coverage graph. On the contrary, the Flye depth coverage graph (Figure 2c) demonstrated less consistency at the beginning of the assembly, shown by the presence of peaks. Yet, the rest of the assembly showed a more uniform coverage until the end of the assembly. The coverage graphs for the 3 assemblers were not comparable to the one obtained by Rojas et al (2023) (Figure 2e) as they showed low support regions and an inconsistent coverage of the genome. On the other hand, the ptGAUL bioinformatic pipeline depth coverage graph (Figure 2d) indicated a mostly consistent and uniform genome coverage, with little change compared to the published genome depth coverage graph (Figure 2e).

Assembly graphs (Figure 3) were generated for the ptGAUL and Flye assemblies (as only these programs provided the required files) and compared to assembly graph from the published sequence (Rojas et al., 2023). The Flye assembly graph (Figure 3a) suggested an unresolved genome structure due to the presence of genome segments that could not be integrated in the genome. This result was visualized by the presence of small genome pieces

connected by black lines that represent different arrangement possibilities (paths) to form a coherent genome structure. A similar trend can be seen for the published sequence (Figure 3c) as the genome was divided into three distinct regions that could be arranged in different configurations. Unlike the previous sequences, the ptGAUL assembly graph (Figure 3b) indicated a resolved genome structure as a single arrangement of the genome regions was identified, therefore showing a more defined genome structure compared to both the Flye assembly and the sequence obtained by Rojas et al (2023). Based on all quality checks, the ptGAUL bioinformatic pipeline assembly was used for further analysis as it provided the genome with the best characteristics.

3.1.2 Annotation and manual curation

Online genome annotation through the GeSeq platform resulted in the identification of 175 genes and 19 pseudogenes. Of these, 60 genes were removed, as they did not align with reference gene sequences or contained random stop codons, and 47 were reannotated based on the manual curation of the genome. This resulted in the identification of 134 genes and 6 pseudogenes (Table 3), same as the ones identified by Rojas et al (2023). The visualization of the genome annotation and structure (Figure 4) displayed the 4 characteristic regions of the chloroplast genome (LSC, SSC, IRA, and IRB).

3.2 Phylogenetic analysis

The results of the phylogenetic analysis (Figure 5) showed a robust bootstrap support (BS) for most clades, ranging from 99% to 100%. Nevertheless, low BS branches were identified with corresponding support values of 71% and 61%. The three newly published *Vaccinium* chloroplast genomes incorporated into the analysis were included into the following sections: *V. ashei* in *Cyanococcus* (99% BS) and *V. microcarpum* and *V. oxycoccos* in *Oxycoccus* (100% BS). Finally, the phylogeny suggested that *V. floribundum* was most closely related to *V. myrtillus* with 85% bootstrap support value.

3.3 Structural analysis

The IRscope structural analysis (Figure 6) compared the chloroplast structural regions (LSC, SSC, IRA, and IRB) of the 10 most closely related *Vaccinium* species to *V. floribundum.* Seven genes (*trnV*, *ccsA*, *trnL*, *rpl32*, *ndhF*, *psbA*, and *trnK*) bordering the mentioned regions were located in each sequence. Clear gene patterns were identified in the chloroplast regions of all compared *Vaccinium* species. Patterns identified in the LSC regions showed the *trnV* gene was located 61 bp before the IRB. Moreover, in all *Vaccinium* sequences the *ccsA*, *trnL* and *rpl32* genes were all located in the IRB regions bordering the SSC in a forward direction, and the same 3 genes could be found in the IRA in a reverse direction. Next, *ndhF* genes were found within the SSC and were coded forwards or reverse, depending on the species. Finally, the LSC contained the *psbA* gene and *trnK* gene in all *Vaccinium* sequences, except for *V. ashei* that did not contain a *trnK* gene.

When compared, the *V. floribundum* chloroplast genomes showed a *trnV* gene further away than 61 bp from the IRB border and the lack of a *rpl32* gene in the same IRB region. The single *rpl32* gene identified in each *V. floribundum* sequence was located within the SSC instead of the IRA as seen in the other analyzed *Vaccinium* sequences (Figure 6).

All region sizes of the analyzed chloroplast genomes followed similar trends with the LSC being approximately 105,000 bp, the IRB and IRA being approximately 32,000-42,000 bp and the SSC being approximately 3,000 bp. The only discrepancies found between the published *V. floribundum* chloroplast sequence and the one obtained in the current study were the region sizes. In the sequence published by Rojas et al (2023), LSC, IRB, SSC and IRA regions had a length of 107,279 bp, 38,421 bp, 3,839 bp and 38,421 bp, respectively. On the other hand, LSC, IRB, SSC and IRA regions obtained in this study had a length of 107,801 bp, 38,485 bp, 3,121 bp and 38,485 bp, respectively.

4. DISCUSSION

4.1 Genome evaluation

Quality checks (QUAST statistics, depth coverage graphs, and assembly graphs) were used to evaluate genome assemblies as they provided useful metrics to determine assembly quality (Sims et al., 2014; Mikheenko et al., 2018; Jin et al., 2020).

The results obtained for the mentioned quality checks demonstrated favorable statistics for the ptGAUL bioinformatic pipeline assembly compared to the other assemblers used. The reason why the ptGAUL assembly demonstrated a better assembled genome than the other assemblers is most likely due to the differences in bioinformatic pipelines that yield the final assemblies. The ptGAUL bioinformatic pipeline is designed specifically for the assembly of plastid genomes using long reads and functions in the following steps. This pipeline separates chloroplast reads and filters them so only reads greater than 3000 bp remain. The obtained reads are then analyzed to ensure a 50x coverage subset of reads. Lastly, the reads are assembled into a genome using the Flye assembler (Appendix 4) (Zhou et al., 2023). This pipeline is clearly designed and intended to provide high quality assemblies from long reads, meanwhile the adapted Syme et al (2021) pipeline is an adaptation of a more complex pipeline that included the use of both long and short reads to provide multiple final assemblies (Zhou et al., 2023). Therefore, the use of a bioinformatic pipeline specifically orientated to long read plastid assemblies is most likely why the ptGAUL assembly outperformed other assemblies.

All quality checks were used to assess the four generated genomes except for assembly graphs. Only the Flye and ptGAUL assemblies provided the required gfa files for the visualization of the genome structure. This is due to the different assembly algorithms used by each assembler. Wtdbg2 and Canu assemblers use an Overlap-Layout-Consensus algorithm (OLC) which works by overlapping found reads, then performs a layout and graph of the overlapped reads, and finally infers a consensus sequence (Li et al., 2012; Koren et al., 2017;

Ruan & Li, 2020). On the other hand, the Flye assembler, also used in the ptGAUL bioinformatic pipeline (Appendix 3), employs an assembly algorithm known as De-Bruijn-Graph algorithm (DBG). This method is based on the formation of random contigs that are built into possible assembly graphs which are then compared and used to infer a final assembly graph and genome sequence (Li et al., 2012; Kolmogorov et al., 2019; Zhou et al., 2023). Unlike the OLC algorithm, DBGs inferred final assembly graph file (gfa) can then be used to visualize the genome structure (Figure 3).

4.2 *V. floribundum* chloroplast genome compared to previous studies

The proposed bioinformatic pipeline (Figure 1) developed based on the methodology described by Syme et al. (2021) had the novel feature of using the obtained assemblies as baits to attempt to extract plastid reads from all the obtained reads with more accuracy (Syme et al., 2021). Yet, the proposed pipeline had no effect on the ptGAUL pipeline as it is specifically designed for long read assembly by using its own programs, parameters, and steps (Zhang et al., 2023).

The resulting *V. floribundum* chloroplast genomes from the Rojas et al (2023) study and the current study had similar sizes and contained the same genes and pseudogenes. Despite their similarities, factors such as the quantity of plastid reads obtained could explain the key differences in both projects. The current study used nine *Vaccinium* chloroplast sequences, whereas Rojas et al (2023) only used the genome of *V. macrocarpon* as a reference. This caused a variation in the amount of chloroplast reads extracted for each project, which lead to different initial read files, subsequently causing differences in both assemblies (Zhang et al., 2020).

4.2.1 Genome annotation and manual curation

Genome annotation and subsequent manual curation led to the identification of the same 134 genes and 6 pseudogenes described by Rojas et al (2023). Yet, many SNPs (Single

Nucleotide Polymorphisms) were identified in comparison to the published *V. floribundum* chloroplast genome. This is most likely caused by long read sequencing as it tends to have high error rates of 10 to 12% compared to short read sequences (<0.5%) (Morisse et al., 2021; Chen et al., 2021; Ruiz et al., 2023). A potential solution could include the implementation of a polisher that uses short reads to correct SNPs (Chen et al., 2021; Zhou, 2023; Syme et al., 2021). Newly developed long read assembly polishing software tend to incorporate short reads as part of the polishing algorithm, this can be seen in novel polishing tools such as NextPolish and Polypolish that seem to outperform non hybrid long read polishing software (Zhang et al., 2020; Chen et al., 2021; Wick & Holt, 2022

4.2.2 Phylogenetic analysis

The phylogenetic analysis of *V. floribundum* used the same genomic data published by Rojas et al (2023), except for the addition of three newly published *Vaccinium* chloroplast genomes (*V. ashei*, *V. microcarpum* and *V. oxycoccos*) (Qiao et al., 2023; Yang et al., 2023). Both of the phylogenies obtained in the study by Rojas et al (2023) and in the present study (Figure 5) suggest that *V. myrtillus* is the closest relative to *V. floribundum*. However, the addition of the three new *Vaccinium* chloroplast genomes increased the bootstrap support from 78% to 85%, further supporting this phylogenetic relationship.

4.2.3 Structural analysis

The IRscope structural analysis revealed conserved patterns in the regions of all *Vaccinium* sequences analyzed (Figure 6) (Fahrenkrog et al., 2022). Both *V. floribundum* structural analysis demonstrated slight differences in these patterns which included a lack of a *rpl32* gene (1 copy instead of 2) in the IRB and one less copy of the *ndhF* gene in the SSC. Moreover, the identified *rpl32* gene was located in the SSC, contrary to other *Vaccinium* sequences where it was located at the IRA (Figure 6).

Despite their similarities, small differences were identified between the sequence obtained in this study and the one developed by Rojas et al (2023). These differences were identified in the sizes of the four characteristic chloroplast genome regions. The IR regions differed by 268 bp, 38,421 bp in Rojas et al (2023) study and 38,153 bp in the current one. The SSC regions varied by 54 bp, 3,839 bp in Rojas et al (2023) study and 3,785 bp in the current one. Finally, the LSC regions varied by 522 bp (107,279 bp in Rojas et al (2023) study and 107,801 bp in the current one).

The variation in region sizes is most likely due to differences in the procedure for the isolation of the chloroplast reads. As mentioned before, the current study incorporated 9 *Vaccinium* chloroplast genomes as references (Appendix 1) meanwhile, in the Rojas et al (2023) study only one reference genome (*V. macrocarpon* chloroplast genome) was used for chloroplast read extraction. The difference in the amount of chloroplast reference sequences used led to different amounts of reads obtained in each study, 33,280 reads in the Rojas et al (2023) study and 21,120 reads in the current study. These results suggest that increasing the number of references makes the extraction process more selective, therefore reducing the total amount of extracted reads (Twyford & Ness, 2017; Zhang et al., 2020). This means that each project started with a different data set which could be the reason why the chloroplast region sizes in each study vary.

5. CONCLUSIONS

The *V. floribundum* chloroplast genome sequence assembled in this study, through the ptGAUL bioinformatic pipeline, confirmed the results obtained by Rojas et al (2023) as the same genes, pseudogenes and phylogenetic relationships were identified. Nevertheless, differences in the number of references used for chloroplast read extraction led to different reads being extracted which generated differences between both sequences including small differences in chloroplast region sizes. New information regarding the chloroplast genome of *V. floribundum* was obtained as the addition of three new *Vaccinium* sequences in the phylogenetic analysis increased the bootstrap support (from 78% to 85%) for the clade that identified *V. myrtillus* as the closest relative of *V. floribundum.* Finally, this study is important as it provides a useful genomic tool that can be used for further studies regarding *V. floribundum* as well as other *Vaccinium* species.

6. TABLES

Table 1. QUAST statistics obtained from the first run of the bioinformatic pipeline for each assembly.

Polisher	Assembler	Size (bp)	NGA50	Indels per 100 000 bp
Apollo	**Canu**	199051	104280	658.27
	Flye*	189001	104585	501.5
	Wtdbg2	183271	103963	621.61
	ptGAUL*	187879	104617	783

* Assembly that showed the best parameters

Table 2. QUAST statistics obtained from the second run of the bioinformatic pipeline for each assembly

Polisher	Assembler	Size (bp)	NGA50	Indels per 100 000 bp
Apollo	**Canu**	172053	104152	613.21
	Flye	187797	104426	997
	Wtdbg2	176258	103532	818.97
	ptGAUL*	187892	104600	764

* Assembly that showed the best parameters

Table 3. Pseudogenes identified in the chloroplast genome of *V. floribundum*

Pseudogene	Initial position (bp)	Final position (bp)	Length (bp)
accD	99,361	99,655	285
clpP1	61238	61309	72
infA	57231	57692	462
ycf1	106558	107053	496
ycf2	48139	48327	189
ndhF	186811	187269	459

7. FIGURES

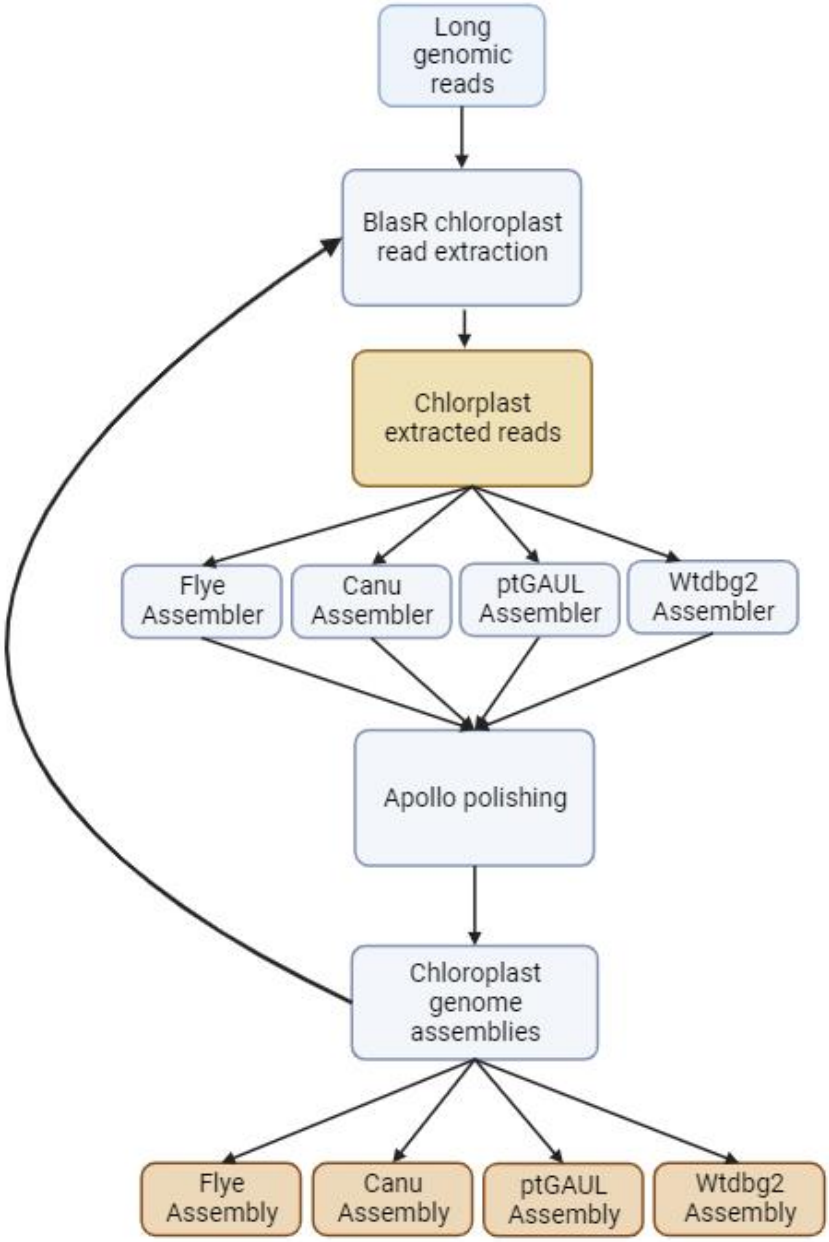

Figure 1. *V. floribundum* bioinformatic pipeline

Proposed bioinformatic pipeline for the chloroplast genome assembly of *V. floribundum* (Created with BioRender.com) based on the methodology described by Syme et al. (2021) for the assembly of the chloroplast genome of *Acacia pycnantha* using long reads.

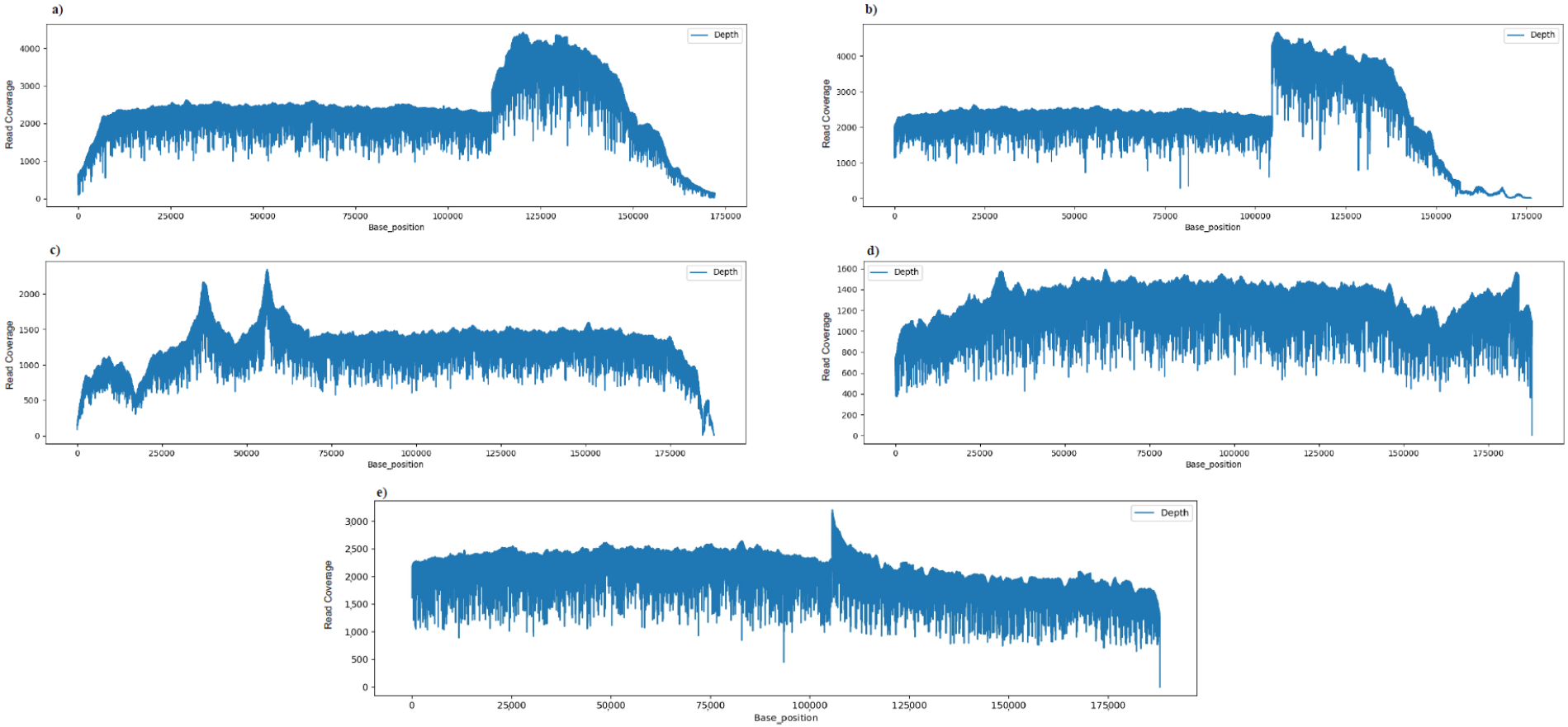

Figure 2. Depth coverage graphs

The depth coverage graphs demonstrated how the assemblies were supported by the obtained reads; the y-axis represented the read coverage while the x-axis represented the base position in the genome. a) Canu assembly depth coverage graph demonstrated an initial stable coverage of the genome. A considerable increase in coverage can be seen around 10,000 pb and finally a significant decrease in the coverage towards the end of the graph. b) Wtdbg2 depth coverage graph initially demonstrated a stable pattern until an increase in

coverage around 10,000 bp and an equally drastic decrease around 140,000 bp. c) Flye depth coverage graph showed an inconsistent pattern towards the beginning which then stabilized and subtly dropped towards the end of the graph. d) ptGAUL depth coverage graph showed the most consistent coverage pattern out of all assemblers, showing no extreme changes in the coverage as the graph progressed. e) The coverage graph of the published sequence demonstrated stable and uniform coverage throughout the entire sequence.

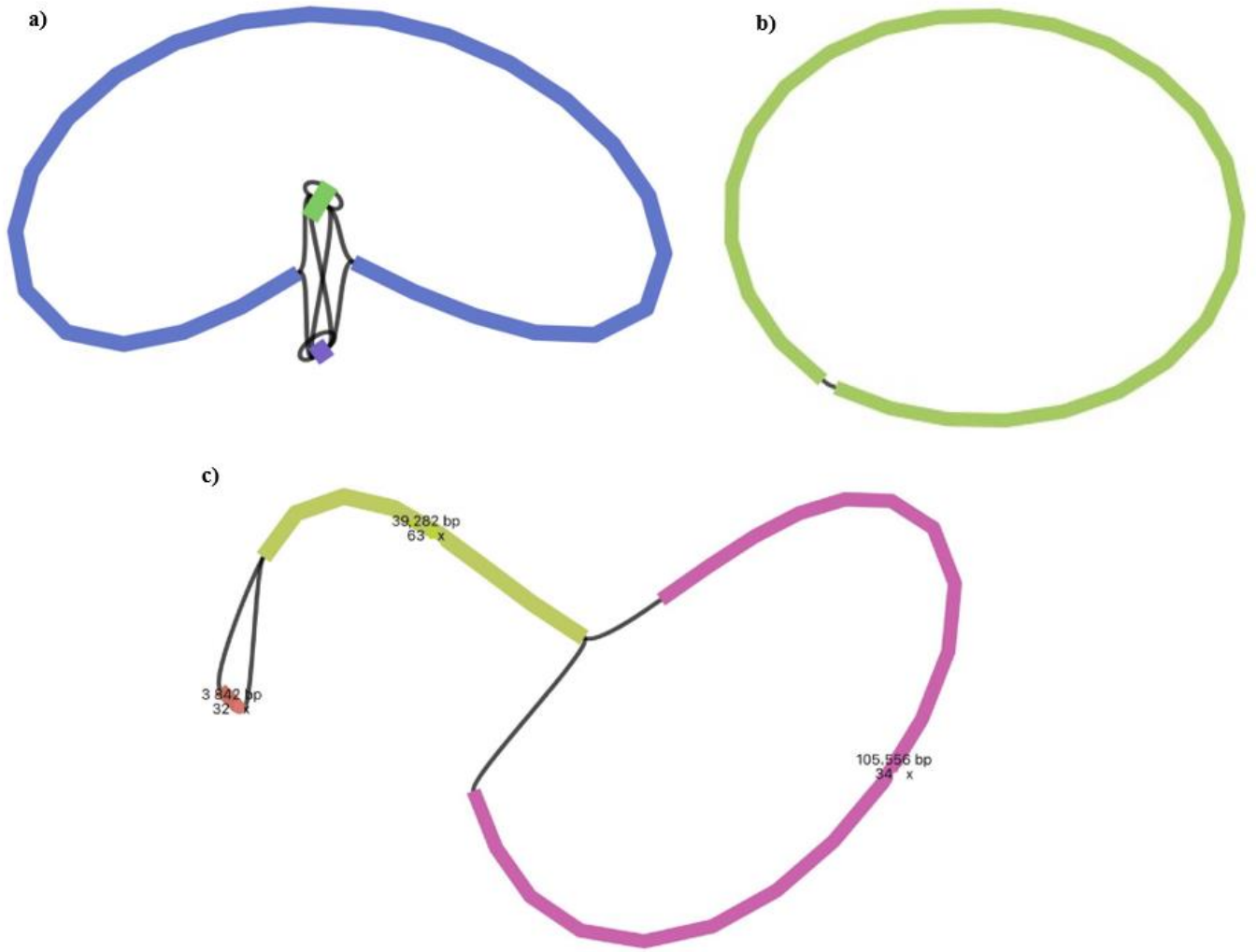

Figure 3. Flye, ptGAUL, and *V. floribundum* published sequence assembly graphs

Bandage v.0.9.0 software was used to visualize chloroplast genome structures from gfa files provided by Flye and ptGAUL assemblers. a) The Flye assembly graph demonstrated an unresolved genome due to the presence of ambiguous structures represented by segmented

figures. Each colored section represents a different part of the genome that could not be arranged into a coherent structure, meanwhile possible arrangements of these parts are represented by black connecting lines. b) The ptGAUL assembly graph demonstrated a resolved genome structure with no ambiguous segments or alternative arrangements. c) The published *V. floribundum* assembly graph demonstrated an unresolved structure with 3 main segments that can be ordered in different combinations.

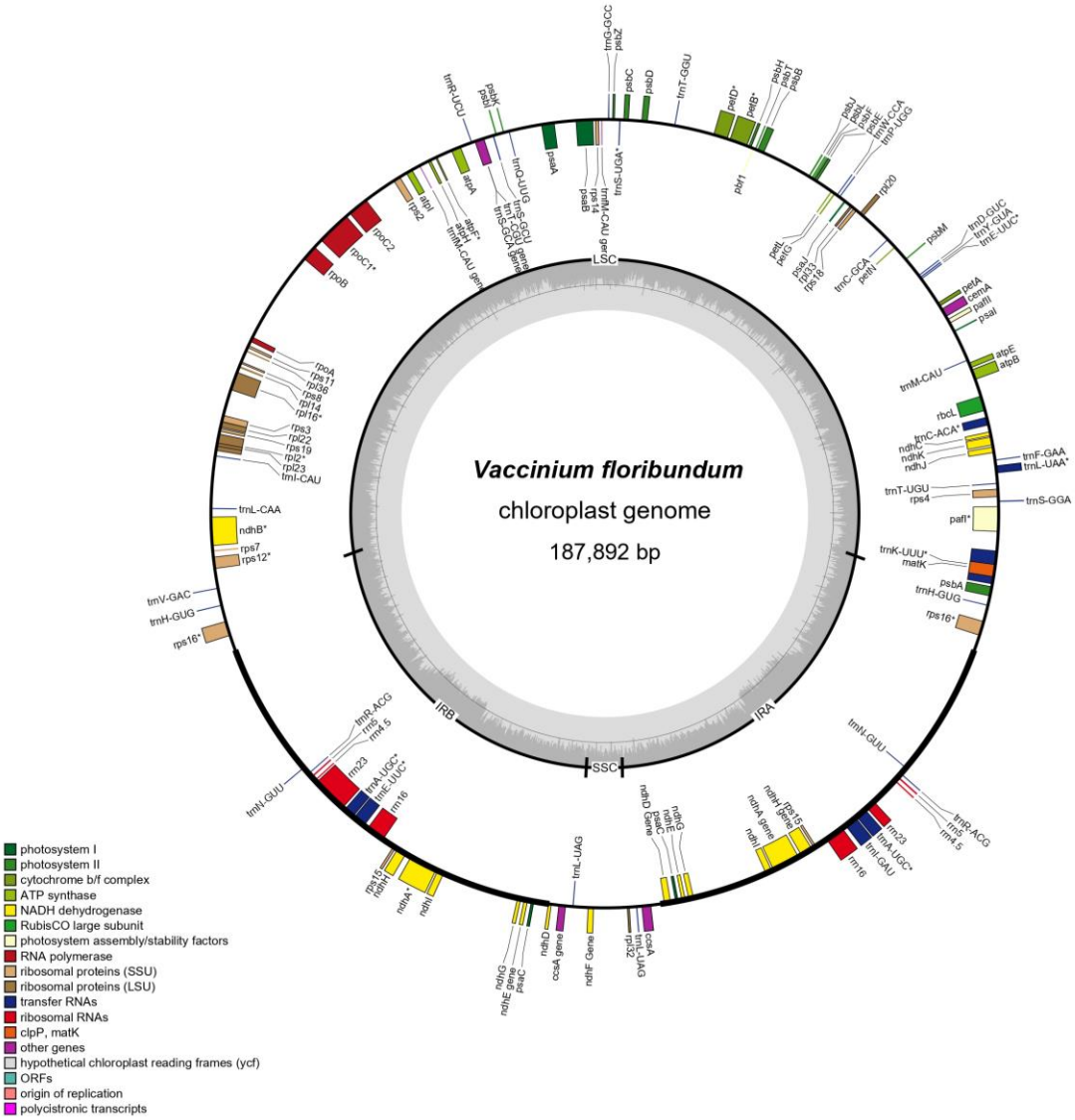

Figure 4. *V. floribundum* chloroplast genome structure

The following figure demonstrates the complete chloroplast genome structure of *V. floribundum* including all annotated genes and characteristic structural regions (LCS, SSC, IRA, and IRB). Genes represented on the inside of the genome are transcribed in a forward direction, while those on the inside are transcribed in the reverse direction. Moreover, on the bottom left corner, a list of gene functional groups is provided as a legend for the figure. Finally, AT and GC contents are visualized by a light grey line in the inner circle and a dark grey area surrounding it.

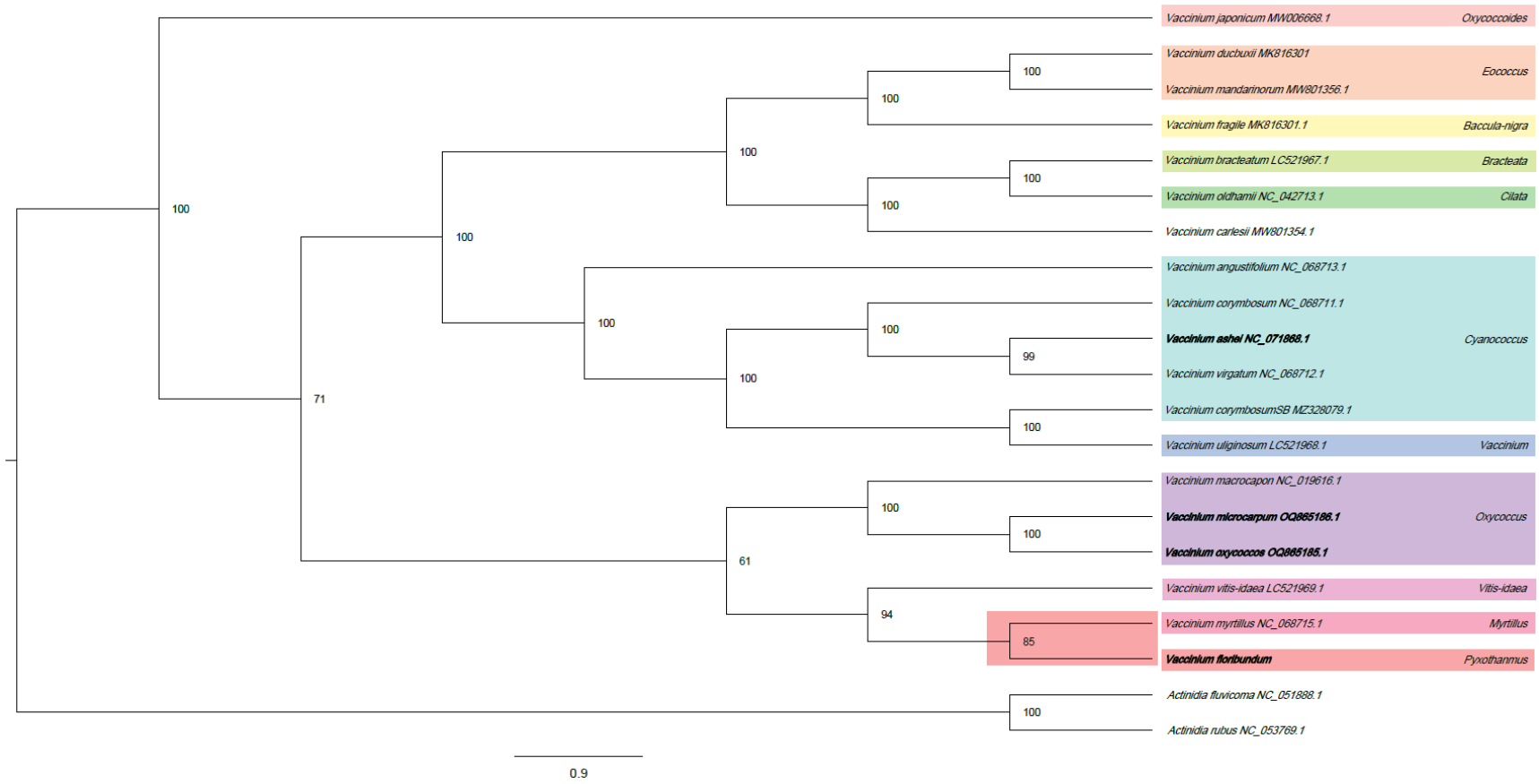

Figure 5. *V. floribundum* phylogeny where species are grouped into *Vaccinium* sections represented by distinct colors

A total of 19 *Vaccinium* chloroplast genome sequences were compared using 2 *Actinidia* chloroplast genomes as outgroups to root the phylogeny. The 3 newly published Vaccinium chloroplast genomes incorporated into the analysis were included into the following sections: *V. ashei* in *Cyanococcus* (99% BS) and *V. microcarpum* and *V. oxycoccos* in *Oxycoccus* (100% BS). Finally, the phylogenetic analysis suggests that *V. floribundum* is most closely related to *V. myrtillus* with a clade resolution of 85%.

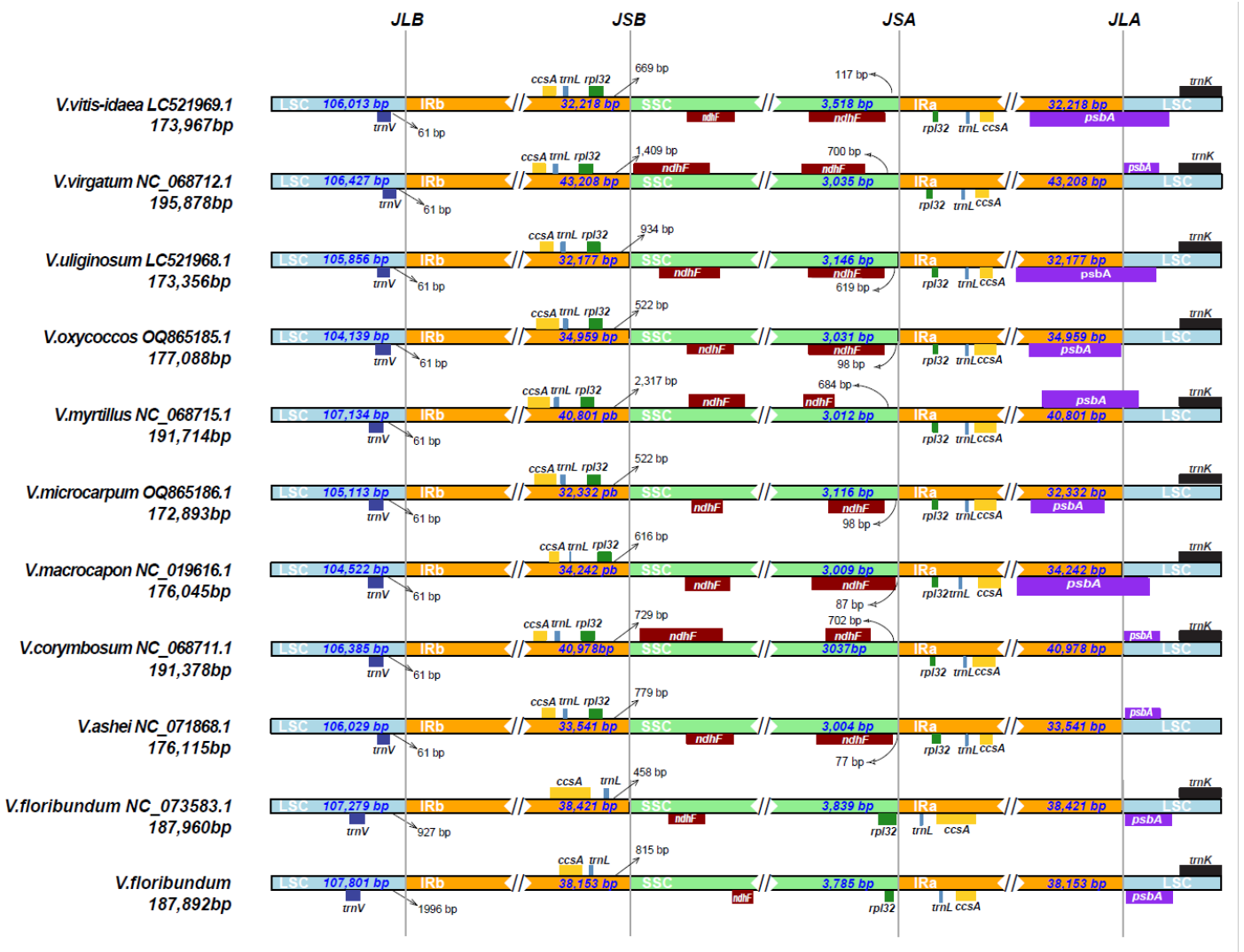

Figure 6. IR-Scope structural analysis of the chloroplast genome of *V. floribundum* compared to other *Vaccinium* chloroplast genomes

The IRscope structural analysis compared structural regions (LSC, IRB, SSC, and IRA) of 11 *Vaccinium* chloroplast sequences as well as 7 genes (*trnV, ccsA, trnL, rpl32, ndhF, psbA,* and *trnK*) bordering the chloroplast regions. Genes located above the genome are coded

in a forward direction, while genes located below the genome a transcribed in a reverse direction. A clear pattern was identified for all *Vaccinium* chloroplast genomes. The main differences were found in both *V. floribundum* structures that showed a *trnV* gene further than 61 bp from the IRB border and the lack of a *rpl32* gene in the IRB. In both *V. floribundum* sequences the identified *rpl32* gene was found within the SSC unlike other *Vaccinium* sequences. Moreover, small size variations of the IR and SSC regions were found between both *V. floribundum* sequences.

REFERENCES

Abril, J. F., & Castellano, S. (2019). Genome Annotation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 195–209). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20226-4

Aguilar, Z. 2009. Guía de plantas útiles de los páramos de Zuleta, Ecuador. *EcoCiencia, Proyecto Páramo Andino*. Programa de Apoyo a la Gestión Descentralizada de los Recursos Naturales en las Tres Provincias del Norte del Ecuador.

Amiryousefi, A., Hyvönen, J., & Poczai, P. (2018). IRscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*, *34*(17), 3030–3031. https://doi.org/10.1093/bioinformatics/bty220

Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., & Scorilas, A. (2022). Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life*, *12*(1), Article 1. https://doi.org/10.3390/life12010030

Cai, L., Zhang, H., & Davis, C. C. (2022). PhyloHerb: A high-throughput phylogenomic pipeline for processing genome skimming data. *Applications in Plant Sciences*, *10*(3), e11475. https://doi.org/10.1002/aps3.11475

Campbell, N. and Reece, J. (2007). *Biology 7th edition.* Panamerican Medical Editorial. Madrid: Spain.

Caranqui-Aldaz, J. M., Romero-Saltos, H., Hernández, F., & Martínez, R. (2022). Reproductive phenology of *Vaccinium floribundum Kunth* (Ericaceae) and codification according to the BBCH scale based on evidence from the volcano Chimborazo paramo (Ecuador). *Scientia Horticulturae*, *303*, 111207. https://doi.org/10.1016/j.scienta.2022.111207

Chaisson, M. J. & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory | *BMC Bioinformatics* | *Full Text*. Retrieved October 20, 2023, from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-238

Chen, Z., Erickson, D. L., & Meng, J. (2021). Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics*, *113*(3), 1366–1377. https://doi.org/10.1016/j.ygeno.2021.03.018

Choudhuri, S. (2014). Chapter 7—Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences**The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government. In S. Choudhuri (Ed.), *Bioinformatics for Beginners* (pp. 157–181). Academic Press. https://doi.org/10.1016/B978-0-12-410471-6.00007-4

Christenhusz, M. J., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3), 201-217. https://phytotaxa.mapress.com/pt/article/view/phytotaxa.261.3.1#:~:text=Abstract,74%2C273%3B%20eudicots%3A%20210%2C008).

Coba Santamaría, P., Coronel, D., Verdugo, K., Paredes, M. F., Yugsi, E., & Huachi, L. (2012). Estudio etnobotánico del mortiño (vaccinium floribundum) como alimento ancestral y potencial alimento funcional. *La Granja*, *16*(2), 5. https://doi.org/10.17163/lgr.n16.2012.01

Contreras-Díaz, R., Carevic, F. S., Huanca-Mamani, W., Oses, R., Arias-Aburto, M., & Navarrete-Fuentes, M. (2022). Chloroplast genome structure and phylogeny of Geoffroea decorticans, a native tree from Atacama Desert. *Electronic Journal of Biotechnology*, 60, 19–25. https://doi.org/10.1016/j.ejbt.2022.09.005

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Fahrenkrog, A. M., Matsumoto, G. O., Toth, K., Jokipii-Lukkari, S., Salo, H. M., Häggman, H., Benevenuto, J., & Munoz, P. R. (2022). Chloroplast genome assemblies and comparative analyses of commercially important Vaccinium berry crops. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-25434-5

Firtina, C., Kim, J. S., Alser, M., Senol Cali, D., Cicek, A. E., Alkan, C., & Mutlu, O. (2020). Apollo: A sequencing-technology-independent, scalable and accurate assembly polishing algorithm. *Bioinformatics*, *36*(12), 3669–3679. https://doi.org/10.1093/bioinformatics/btaa179

Formenti, G., Abueg, L., Brajuka, A., Brajuka, N., Gallardo-Alba, C., Giani, A., Fedrigo, O., & Jarvis, E. D. (2022). Gfastats: Conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics*, *38*(17), 4214–4216. https://doi.org/10.1093/bioinformatics/btac460

Greiner, S., Lehwark, P., & Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research*, *47*(W1), W59–W64. https://doi.org/10.1093/nar/gkz238

Henry, R. J. (2022). Progress in Plant Genome Sequencing. *Applied Biosciences*, *1*(2), https://doi.org/10.3390/applbiosci1020008

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. https://doi.org/10.1186/s13059-020-02154-5

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs | *Nature Biotechnology*. https://www.nature.com/articles/s41587-019-0072-8

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. https://doi.org/10.1101/gr.215087.116

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: Overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/bfgp/elr035

Llivisaca-Contreras, S. A., León-Tamariz, F., Manzano-Santana, P., Ruales, J., Naranjo-Morán, J., Serrano-Mena, L., Chica-Martínez, E., & Cevallos-Cevallos, J. M. (2022). Mortiño (Vaccinium floribundum Kunth): An Underutilized Superplant from the Andes. *Horticulturae*, *8*(5), https://doi.org/10.3390/horticulturae8050358

Luteyn, J. L. & Pedraza-Peñalosa, P. (2012). Blueberry relatives of the New World tropics (Ericaceae). *The New York Botanical Garden*, Bronx, New York. https://sweetgum.nybg.org/ericaceae/index.php

Martău, G. A., Bernadette-Emőke, T., Odocheanu, R., Soporan, D. A., Bochiș, M., Simon, E., & Vodnar, D. C. (2023). Vaccinium Species (Ericaceae): Phytochemistry and

Biological Properties of Medicinal Plants. *Molecules*, *28*(4), https://doi.org/10.3390/molecules28041533

Meléndez-Jácome, M. R., Flor-Romero, L. E., Sandoval-Pacheco, M. E., Vasquez-Castillo, W. A., Racines-Oliva, M. A., Meléndez-Jácome, M. R., Flor-Romero, L. E., Sandoval-Pacheco, M. E., Vasquez-Castillo, W. A., & Racines-Oliva, M. A. (2021). Vaccinium spp.: Características cariotípicas y filogenéticas, composición nutricional, condiciones edafoclimáticas, factores bióticos y microorganismos benéficos en la rizosfera. *Scientia Agropecuaria*, *12*(1), 109–120. https://doi.org/10.17268/sci.agropecu.2021.013

Mena Vásconez, P., & Grupo de Trabajo en Páramos del Ecuador (Eds.). (2011). *Páramo: Paisaje estudiado, habitado, manejado e institucionalizado ; selección de textos de la Serie Páramo, órgano de difusión del Grupo de Trabajo en Páramos del Ecuador (GTP)*. Ed. Univ. Abya-Yala [u.a.].

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

Morisse, P., Marchet, C., Limasset, A., Lecroq, T., & Lefebvre, A. (2021). Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-020-80757-5

Oxford Nanopore Technologies. (n.d.). *How basecalling works*. Oxford Nanopore Technologies. Retrieved September 22, 2023, from https://nanoporetech.com/how-it-works/basecalling

Provan, J., Powell, W., & Hollingsworth, P. M. (2001). Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution, 16(3), 142–147.* doi:10.1016/s0169-5347(00)02097-8

Qiao, X., Gu, Q., Ye, R., Cai, J., & Zhu, N. (2023). The complete chloroplast genome of *Vaccinium oxycoccos* (Ericaceae). *Mitochondrial DNA. Part B, Resources*, *8*(9), 942–947. https://doi.org/10.1080/23802359.2023.2252943

Rambaut, A. FigTree v1.3.1. Available online: http://tree.bio.ed.ac.uk/software/figtree/

Rogalski, M., do Nascimento Vieira, L., Fraga, H. P., & Guerra, M. P. (2015). Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. *Frontiers in plant science*, *6*, 586. https://doi.org/10.3389/fpls.2015.00586

Rojas, K. E. L., Armijos, C. E., Parra, M., & Torres, M. de L. (2023). The First Complete Chloroplast Genome Sequence of Mortiño (Vaccinium floribundum) and Comparative Analyses with Other Vaccinium Species. *Horticulturae*, *9*(3), Article 3. https://doi.org/10.3390/horticulturae9030302

Roston, R. L., Jouhet, J., Yu, F., & Gao, H. (2018). Editorial: Structure and Function of Chloroplasts. *Frontiers in Plant Science*, *9*, 1656. https://doi.org/10.3389/fpls.2018.01656

Ruan, J. & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 17, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Ruiz, J. L., Reimering, S., Escobar-Prieto, J. D., Brancucci, N. M. B., Echeverry, D. F., Abdi, A. I., Marti, M., Gómez-Díaz, E., & Otto, T. D. (2023). From contigs towards chromosomes: Automatic improvement of long read assemblies (ILRA). *Briefings in Bioinformatics*, *24*(4), bbad248. https://doi.org/10.1093/bib/bbad248

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), Article 2. https://doi.org/10.1038/nrg3642

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Syme, A. E., McLay, T. G. B., Udovicic, F., Cantrill, D. J., & Murphy, D. J. (2021). Long-read assemblies reveal structural diversity in genomes of organelles—An example with Acacia pycnantha. *GigaByte (Hong Kong, China)*, *2021*, gigabyte36. https://doi.org/10.46471/gigabyte.36

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Torres, M.L., Trujillo, D., Arahana, V. (2010). Cultivo *in vitro* del mortiño (*Vaccinium floribundum* Kunth). *ACI Av. Cienc. Ing. 2, B9–B15*. 10.18272/aci.v2i2.27.

Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*, *17*(5), 858–868. https://doi.org/10.1111/1755-0998.12626

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., Fai Au, K. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 39, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wick, R. (2017). *Porechop*. Retrieved from, https://github.com/rrwick/Porechop

Wick, R. R., & Holt, K. E. (2022). Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Computational Biology*, *18*(1), e1009802. https://doi.org/10.1371/journal.pcbi.1009802

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Yang, H., Zhang, C., Wu, Y., Wu, W., Lyu, L., & Li, W. (2023). The complete chloroplast genome of rabbiteye blueberry (Vaccinium ashei) and comparison with other Vaccinium species. *Brazilian Journal of Botany*. https://doi.org/10.1007/s40415-023-00954-0

Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. *BMC Genomics*, *21*(6), 889. https://doi.org/10.1186/s12864-020-07227-0

Zhang, Y., Akdemir, A., Tremmel, G., Imoto, S., Miyano, S., Shibuya, T., & Yamaguchi, R. (2020). Nanopore basecalling from a perspective of instance segmentation. *BMC Bioinformatics*, *21*(3), 136. https://doi.org/10.1186/s12859-020-3459-0

Zhang, Z., & Ren, Q. (2015). Why are essential genes essential? - The essentiality of *Saccharomyces* genes. *Microbial cell (Graz, Austria)*, *2*(8), 280–287. https://doi.org/10.15698/mic2015.08.218

Zhou, W. (2023). *PlasTid Genome Assembly Using Long reads data (ptGAUL)* [Shell]. https://github.com/Bean061/ptgaul (Original work published 2022)

Zhou, W. Armijos, C.E. Lee, C. Lu, R. Wang, J. Tracey, A. Jansen, R.K. Jones, A.M. Jones, C.D. (2023). Plastid genome assembly using long-read data (PtGAUL). *bioRxiv*, 1–39. 10.1111/1755-0998.13787

APPENDICES

Appendix 1. *Vaccinium* chloroplast genomes used as baits for Blasr plastid read extraction.

Organism	Accession Code
Vaccinium duclouxii	MK816300.1
Vaccinium fragile	MK816301.1
Vaccinium bracteatum	LC521967.1
Vaccinium macrocarpon	NC_019616.1
Vaccinium oldhamii	MK049537.1
Vaccinium uliginosum	LC521968.1
Vaccinium vitis-idaea	LC521969.1
Vaccinium japonicum	MW006668.1
Vaccinium corymbosum SB	MZ328079.1

Appendix 2. Chloroplast genomes used for the phylogenetic and IR-scope structural analysis of *V. floribundum*

	Organism	Accession Code
Chloroplast genomes used by Rojas et al (2023)	*Vaccinium japonicum*	MW006668.1
	Vaccinium ducbuxii	MK816301
	Vaccinium mandarinorum	MW801356.1
	Vaccinium fragile	MK816301.1
	Vaccinium bracteatum	LC521967.1
	Vaccinium oldhamii	NC_042713.1
	Vaccinium carlesii	MW801354.1
	Vaccinium angustifolium	NC_068713.1
	*Vaccinium corymbosum**	NC_068711.1
	*Vaccinium virgatum**	NC_068712.1
	Vaccinium corymbosumSB	MZ328079.1
	*Vaccinium uliginosum**	LC521968.1
	*Vaccinium macrocapon**	NC_019616.1
	*Vaccinium vitis-idaea**	LC521969.1
	*Vaccinium myrtillus**	NC_068715.1
Newly incorporated Chloroplast genomes	*Vaccinium ashei**	NC_071868.1
	*Vaccinium microcarpum**	OQ865186.1
	*Vaccinium oxycoccos**	OQ865185.1
Chloroplast genomes used to root the phylogeny	*Actinidia fluvicoma*	NC_051888.1
	Actinidia rubus	NC_053769.1

* Chloroplast genomes used for *V. floribundum's* IRscope structural analysis

Appendix 3. ptGAUL bioinformatic pipeline for long read plastid genome assembly (Zhou et al., 2023)