Genome analysis of the diploid wild potato *Solanum bukasovii*

Ilayda Bozan

Department of Plant Science
Faculty of Agricultural and Environmental Sciences

McGill University
Montreal, Quebec, Canada

June 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Potato (*Solanum tuberosum L.*) originated in the South American Andes and is an economically important staple crop that can be successfully grown in various conditions and altitudes. The potato genome is complex, with a large gene pool drawn from numerous wild species of varying ploidy levels ranging from diploids to hexaploids. Potato breeding efforts in North America and Europe have traditionally focused on vegetative propagation of tetraploid potato because of its higher yield. However, due to the high heterozygosity levels and difficulties of tetraploid breeding, new improvement efforts are increasingly looking to diploid species as a means of introgressing traits into crop varieties.

Genome sequence data and means of analyzing and visualizing the data, are crucial to achieve this goal. Significant work has previously been done by sequencing and publishing different potato reference genomes, a double monoploid (*S. tuberosum Group Phureja* – DM), two wild reference genomes (*S. commersonii* and *S. chacoense* clone M6), a landrace genome (*S. stenotomum* subsp. *goniocalyx*), and two diploid *S. tuberosum* genomes (*Solanum tuberosum* group Tuberosum RH89-039-16, *Solanum tuberosum*, Solyntus).

The present study focuses on expanding the genomic resources for potato genome analyses. First, this study presents a newly sequenced and assembled diploid genome of *S. bukasovii*, which is thought to be one of the wild species that is most closely related to the cultivated potato. This genome sequence is compared with the available potato reference genomes and the results show that Copy Number Variation (CNV) affect genes have important functions such as disease resistance and metabolite biosynthesis. Second, a web portal – the Potato Genome Diversity Portal (PGDP) - was developed and implemented to visualize the published potato genomes using JBrowse and to provide a tool to investigate the genome alignments along with aiding the structural variation analysis PGDP also enables researchers to conduct research and share data.

Résumé

La pomme de terre (*Solanum tuberosum* L.) a son origine des Andes en Amérique du Sud. Elle est une culture de base économiquement importante qui peut être cultivée avec succès dans diverses conditions et altitudes. Le génome de la pomme de terre est complexe, avec un vaste pool de gènes provenant de nombreuses espèces sauvages de niveaux de ploïdie variables allant des diploïdes aux hexaploïdes. Les efforts de sélection de la pomme de terre se sont traditionnellement concentrés sur la multiplication végétative de la pomme de terre tétraploïde en raison de sa popularité en tant que culture. Cependant, en raison des niveaux élevés d'hétérozygotie et des difficultés de la sélection tétraploïde, de nouveaux efforts d'amélioration se tournent de plus en plus vers les espèces et variétés diploïdes comme moyen d'introgresser les caractères dans les variétés de cultures. Les données de séquence du génome et les moyens d'analyser et de visualiser ainsi les données sont des outils cruciaux pour y parvenir.

Un travail important a déjà été effectué en séquençant et en publiant différents génomes de référence de la pomme de terre; un double monoploïde (*S. tuberosum* Group Phureja - DM), deux génomes sauvages de référence (*S. commersonii* et *S. chacoense* clone M6), un génome de landrance (*S. stenotomum* subsp. *goniocalyx*) et deux vrais génomes diploïdes de *S. tuberosum* (*S. tuberosum* groupe Tuberosum RH89-039-16, *S. tuberosum*, Solyntus).

La présente étude se concentre sur l'expansion des ressources génomiques pour les analyses du génome de la pomme de terre. Premièrement, l'étude présente une séquence du génome diploïde nouvellement séquencée et assemblée de *S. bukasovii*, que l'on pense être l'une des espèces sauvages les plus étroitement liées à la pomme de terre cultivée. Cette séquence génomique est comparée aux génomes de référence disponibles de la pomme de terre et les résultats montrent que la variation du nombre de copies (CNV) affecte les gènes qui ont des fonctions importantes telles que la résistance aux maladies et la biosynthèse des métabolites. Deuxièmement, un portail Web - le Potato Genome Diversity Portal (PGDP) - a été développé et mis en œuvre pour visualiser les génomes de pommes de terre publiés à l'aide de JBrowse et pour fournir un outil permettant d'étudier les alignements des génomes. En plus d'aider l'analyse des variations structurelles, le PGDP permet également aux chercheurs de mener des recherches et de partager des données.

Acknowledgements

I would like to acknowledge my supervisor Dr. Martina Strömvik for her understanding, guidance and advice throughout my studies at McGill University. It has been a great honour for me to be included in her lab and research. Her graciousness has repeatedly provided me with a safe space to conduct my research. I am grateful for her confidence in me. I have learned so much from her not only as a researcher and a student but also as a human-being. My most sincere thank goes to my supervising committee; Dr. Jianguo Xia, for his insights he provided to this research.

And I would like to acknowledge and thank to our collaborators; Dr. Helen Tai in the Agriculture and Agri-Food Canada and Dr. Noelle Anglin, Dr. Dave Ellis from the International Potato Center (CIP) for giving me the opportunity to work together and for their valuable contributions to this research.

To the members of the Strömvik Lab and my friends at McGill, you made my McGill experience wonderful. Thank you, Sai, Juan, Maria and Alice it is a pleasure to get to work with you guys.

I want to thank my wonderful family for their constant support they have given me not only during my master's degree but every single moment of my life. My fascinating parents, Ülkü and Mehmet Bozan, without your support and faith in me I would have been a completely different person. To my astonishing partner, Arsen Hovhanissian, your perspective of life and tireless support has been the supporting beam to my life. My life sources: my sister Alya Bozan without you, my life would be incredibly dull you are my joy, my cousin Burcu Ölgen, I cannot imagine a day I haven't looked up to you and that I haven't received your emotional support, you're Fred to my George and *I solemnly swear that...*

Table of Contents

List of Figures

List of Tables

List of Appendices

Abbreviations

- 2OG, 2-oxoglutarate
- BAC, Bacterial Artificial Chromosome
- BAM, Binary Sequence Alignment/Map format
- BLAST, Basic Local Alignment Search Tool
- bHLH, basic helix-loop-helix
- bp, base pair(s)
- BUK1, Solanum bukasovii
- BUK2, Solanum bukasovii
- BUSCO, Benchmarking Universal Single-Copy Orthologs, tool
- CDS, coding sequence
- CIP, Centro International de la Papa, International Potato Center
- CNV, Copy Number Variation
- COM, Solanum commersonii
- del, deletion
- DM v6.01, Solanum tuberosum group Phureja DM1-3 516 R44 double monoploid version 6.01
- DM1-3, Solanum tuberosum group Phureja DM1-3 516 R44 (Double Monoploid)
- DNA, Deoxyribonucleic acid
- dup, duplication
- EST, Expressed Sequence Tag
- FISH, Fluorescence in situ hybridization
- GA, Gibberellin
- Gb, Giga bases
- GFF, General-feature format
- GON1, Solanum stenotomum subsp. goniocalyx
- GWAS, Genome Wide Association Studies
- INDEL, Insertions Deletions
- kb, kilo base pairs
- LSR, Long Synthetic Reads
- LTR, Long Terminal Repeat
- M6, M6 clone of Solanum chacoense
- MADs box, MCM1, AGAMOUS, DEFICIENS, SRF
- Mb, mega base pairs
- MFS, Major Faciliator Superfamily
- MQM, Mean mapping quality of observed alternate alleles
- MQMR, Mean mapping quality of observed reference alleles
- NB-ARC, The core nucleotide-binding fold in NB-LRR proteins
- NBS-LRR, Nucleotide Binding Site Leucine Rich Repeat
- NCBI SRA, NCBI Sequence Read Archive
- NCBI, National Center for Biotechnology Information
- NGS, Next Generation Sequencing
- PacBio, Pacific Biosciences

- PAE, Preferential Allele Expression
- PAV, Presence/Absence Variation
- PCA, Principal Component Analysis
- PE, Paired-end
- PGDP, Potato Genome Diversity Portal
- PGSC, Potato Genome Sequencing Consortium
- PM, Pseudomolecule
- QUAST, Quality Assessment Tool for genome assemblies
- RNA-Sequencing/RNA-Seq, RNA sequencing
- RNA, Ribonucleic acid
- SAF, Number of alternate observations on the forward strand
- SAM, Sequence Alignment/Map format
- SAR, Number of alternate observations on the reverse strand
- SAUR, Small Auxin Up-RNA
- SC, Self-Compatibility
- SCF, Skp1- Cullin1 – F-box
- SCF, Skp1-Cullin1-F-box
- SE, Single End
- SI, Self-incompatibility
- SLFs, S-locus F-box proteins
- SMRT, Single Molecule Real-Time
- SNP, Single Nucleotide Polymorphism
- SOL, Solyntus
- SSR: Simple Sequence Repeats
- SV, Structural Variation
- TE: Transposable Element
- TGS, Third Generation Sequencing
- TIR, Toll/interleukin-1 receptor-like
- TMV, Tobacco Mosaic Virus
- ToMV, Tomato Mosaic Virus
- UV-B, Ultraviolet B
- WGS, Whole Genome Shotgu

Contribution of the Authors

The contents of this thesis will be reformatted into a manuscript that is co-authored by Ilayda Bozan, Sai Reddy Achakkagari, Drs. Maria Kyriakidou, Helen Tai, Noelle Anglin and Martina Strömvik. Ilayda Bozan performed the research and wrote the initial draft, completed the tables and made the figures under supervision of Dr. Martina Strömvik. Dr. Noelle Anglin provided the raw potato genome sequencing data and helped design the project together with Martina Strömvik, Helen Tai, and Ilayda Bozan. Sai Reddy Achakkagari performed the SNP based phylogeny analysis for published reference genomes and added the potato plastome sequences to the Potato genome Diversity Portal (PGDP). Dr. Noelle Anglin provided the SNP array data. Dr. Maria Kyriakidou contributed to the bioinformatics methods.

All the computations were conducted on the supercomputers Arbutus, Graham, Cedar and Béluga managed by Compute Canada, thanks to Compute / Calcul Canada Resource Allocations for Research Portals and Platforms (Potato Genome Diversity Portal) (awarded to Drs. Martina Strömvik (PI), Helen Tai and Noelle Anglin) and Resources for Research Groups (awarded to Dr. Martina Strömvik (PI)).

Thesis Format

This thesis is presented in traditional monograph format including the chapters Introduction, Literature Review, Materials and Methods, Results, Discussion and Future Work.

1 INTRODUCTION

Potato (*Solanum* sp., family Solanaceae) is the economically most important non-cereal crop globally (FAO, 2018). Though the ploidy level varies among potato species, the most common commercial potato cultivars (*Solanum tuberosum* L.) are autotetraploid (2n=4x=48) and they are vegetatively propagated. Its genome is highly heterozygous and suffers from acute inbreeding depression making it vulnerable to pests and pathogens (Gebhardt *et al.*, 2004). Because of the polyploidy and high heterozygosity, potato crop improvement is challenging with conventional methods and genome resources are sorely needed. Currently, there are six published genome sequences of diploid potato (Potato Genome Sequencing *et al.*, 2011; Aversano *et al.*, 2015b; Leisner *et al.*, 2018; Kyriakidou *et al.*, 2020; van Lieshout *et al.*, 2020; Zhou *et al.*, 2020) of which one is a wild species (Aversano *et al.*, 2015b). There is a great need to focus current research on improving potato cultivars that can be sustainable under changing climatic conditions. Related wild potato species may hold key structural variations and novel genes for this purpose.

In the present project genome variations such as Copy Number Variations (CNV's) and Single Nucleotide Polymorphisms (SNP's) are investigated in a novel *S. bukasovii* genome sequence. This genome sequence comes from a separate individual genotype stemming from the same accession as that of a previously studied genome sequence in the Strömvik lab, *S. bukasovii* (BUK1 - CIP 761748 DOI: 10.18730/E3AC). This wild species comes from the International Potato Center (CIP, Lima, Peru), and is compared to *S. bukasovii* (BUK1) as well as to the doubled monoploid *S. tuberosum* group phureja DM1-3 516 R44 (DM) (Hardigan *et al.*, 2016), *S. chacoense* (M6) (Leisner *et al.*, 2018), *S. stenotomum subsp. goniocalyx* (GON1 - CIP 702472) (Kyriakidou, 2020), *S. tuberosum* group Tuberosum RH89-039-16 (RH) (Zhou *et al.*, 2020), *S. tuberosum* (SOL) (van Lieshout *et al.*, 2020). For this purpose, a bioinformatics tool has been developed and implemented to visualize the newly assembled genomes and the structural variance discovered.

1.1 Hypotheses

1. Two genomes, BUK1 and BUK2, from the same GenBank accession of the potato species, *S. bukasovii*, have single nucleotide polymorphisms (SNPs) compared to each other.
2. The BUK2 genome has structural variation (e.g., copy number variation CNV's) compared to the potato reference genomes and to the potato diploid pan-genome.

1.2 Objectives

1. Obtain DNA genome and transcriptome (RNA-Seq) sequence data of the BUK2 potato accession (BUK1 is already available).
2. Preprocessing the genome data and *de novo* assembly of the BUK2 genome.
3. Genome to genome comparison between BUK1 and BUK2 to identify the SNPs.
4. Identification of core and accessory genes in BUK2 compared to the pan-genome, and comparison to BUK1.
5. Mapping and CNV analysis of BUK2 to DM, M6, GON1 separately, and comparison with previous BUK1 results.
6. Construction of a JBrowse based genome browser at the Potato Genome Diversity Portal to facilitate the processing and analyses.

2 LITERATURE REVIEW

2.1 Overview

Potato (*Solanum tuberosum L.*) is the third most important crop and the most important non-cereal crop for food security in the world (FAO, 2018). Due to climate change, the natural habitat of Solanum species is being lost. This loss directly effects the wild species which are the source of genes for biotic and abiotic stress resistance. Biotic and abiotic stress resistance genes can be used by breeders for improving crops. Genebanks play a key role in food security through preserving the germplasm, thus conservation in genebanks becomes crucial. Genome sequencing efforts in genebanks allows prioritising germplasm for conservation.

The family Solanaceae also includes important crops such as, tomato, pepper, aubergine, and tobacco along with potato. Potato was domesticated 10,000 years ago in Andean highlands of southern Peru, where landrace potatoes are grown at 3,000 – 4,000 m elevation (Spooner *et al.*, 2005; Ovchinnikova *et al.*, 2011). It is estimated that there are more than 4,500 varieties of native potato (CIP, 2020). The basic (haploid) chromosome number for potato is n=12, though ploidy in potato ranges between diploid to hexaploid, with the majority being diploid (Watanabe 2015). Most of the cultivated potato is autotetraploid (2n=4x=48). Due to its high heterozygosity levels and outcrossing nature, to ensure a consistency and uniformity of traits, it is generally vegetatively propagated (Bradeen & Kole 2016). While North America and Europe heavily relies on tetraploid species, in the developing world species that are grown in their respective environments are preferred.

2.2 Genomic Characteristics of Potato

Occasionally the cultivated and wild potatoes produce 2n gametes. Autopolyploidization of the offspring of these resulted in the Andean cultivated tetraploids [*S. tuberosum* group *Andigena*; 2n=4x=48] (Watanabe & Peloquin 1989). Many modern cultivars are propagated vegetatively and since they are highly related to each other they are different by only a couple meiotic generations (Gebhardt *et al.*, 2004; Simko *et al.*, 2006). As a result of a narrow genetic base (Love, 1999), it is very difficult to improve potato with classical breeding approaches (Potato Genome Sequencing *et al.*, 2011). Large populations of progeny are required to screen and select the desired individuals.

The 2n gametes hybridizing with n gametes are also the source of many of the triploid potato species.

2.3 DNA Sequence Assembly

Several DNA sequencing techniques exist, though after the sequence reads are generated, they all need to be processed and assembled. The shotgun genome assembly approach consists of taking genome sequences and assembling them according to their overlaps. For big and repetitive genomes this is quite challenging, resulting contigs possibly containing gaps and scaffolding or super-scaffolding may be impossible due to small "islands" that are being formed instead of continuous strands (Hamilton & Robin Buell 2012).

2.3.1 De novo and Reference Genome Guided Assembly Techniques

Depending on the purpose for genome assembly two different basic methods can be applied, or hybrids between them (Kyriakidou *et al.*, 2018). Reference-guided genome assembly methods are followed when there is an available reference genome of closely related species. A reference genome is a digital assembled nucleic acid sequence data. This method is applied for genome re-sequencing to support genome assembly or for determination of structural variation. A *de novo* assembly method is followed when a reference genome is not available. *De novo* assembly is placing short and/or long reads together using overlapping methods without a guide sequence.

Different strategies for both methods can be chosen depending on the sequencing data type. Reference guided genome assembly can be achieved utilizing both short and long reads alone or together. Strategies to map short and long reads on reference genomes are used to detect structural variation and polymorphism (Kyriakidou *et al.*, 2018), whereas *de novo* assembling reads before mapping on a reference genome can be applied for performing a better assembly and correct misassembled regions (Lischer & Shimizu 2017).

Several approaches are used to achieve a *de novo* assembly. This method also depends on the read lengths. Short and long reads can be used to assemble *de novo* directly, or short reads can first be *de novo* assembled and the resulting contigs will later be *de novo* assembled with long reads (Kyriakidou *et al.*, 2018).

2.3.2 10X Genomics Assembly

Short read technologies are preferred due to their high frequency reads, however because of the repetitive nature of genomes, it is difficult to construct whole genome *de novo* assemblies using short reads. Genomes with high number of repetitive regions, such as plant genomes, present a particular challenge to assemble with short reads, and downstream analyses, such as structural variation analyses become difficult.

10X linked-read technology (10X Genomics – www.10xgenomic.com) takes advantage of the high number of short reads. In addition, it barcodes the reads that originate from the same long thread with the same molecular barcode. These barcodes help narrow down from which physical DNA molecule (chromosome) each read originates and thus provide synthetic long-read information and haplotype information from the short reads.

The assembly process of 10X linked reads is done with a pipeline, Supernova, provided by the company. Supernova is a program developed by 10X Genomics that uses bcl2fastq (developed by Illumina (www.illumina.com)) in the back (Weisenfeld *et al.*, 2017).

The Supernova pipeline consists of three levels.

1- Converting Base Call Files (BCLs) to FASTQ files. This is done with *supernova mkfastq* command to convert the 10X Chromium reads into barcoded FASTQ files. Additional CSV file is needed to specify Lanes, Samples and Index.
2- *De novo* assembly stage. The *Supernova run* command takes the barcoded reads and creates a whole genome assembly.
3- The last stage is to convert *supernova* output into a FASTA file. *Supernova mkoutput* generates various FASTA files upon demand.

2.4 Annotation

Genome sequences need to be annotated in order to be assigned biological meaning. Genome annotation is the prediction of protein coding regions; exons, introns, regulatory sequences, alternative splicing regions, transcription factor binding sites and non-coding RNAs etc. Genome annotation reveals molecular function and elucidates evolution.

The annotation process often starts with masking the low complexity regions or repeats using the tool RepeatMasker (Smit *et al.*, 2015). Furthermore, RNA sequences, gene models from related genomes, or other known sequences are also used to identify gene regions and exon-intron boundaries by aligning to the assembled genome using BLAST (Boratyn *et al.*, 2012). The alignments are further filtered according to identity or similarity percentages to remove low matching alignments. The last stage of annotation is to annotate remaining novel sequences. Annotation is also done through *ab initio* gene prediction. Following the step of repeat masking, *ab initio* gene prediction is done with software such as AUGUSTUS (Stanke *et al.*, 2004) and Genscan (Burge & Karlin 1997) with an algorithm trained with the same or similar organisms.

2.5 Whole Genome Analysis of Plants

2.5.1 Comparative Genomics and Structural Variation

Current technology has enabled a multitude of genome sequencing projects with detailed sequence information on large numbers of genomes. Single nucleotide polymorphisms (SNPs) are the differences in one nucleotide at a specific position in the organism's genome, often used as a measure of variation of portions of the population (Marth *et al.*, 1999). With advances in sequencing technologies, it is faster and easier to detect SNPs, although SNPs alone are not enough to represent the structural variation in the genome (Springer *et al.*, 2009a). Polymorphisms that are larger than 1 kb are called structural variants (SVs) (Feuk *et al.*, 2006). Structural variation can be insertions or deletions (In/Del), translocation or inversions and copy number variations. Copy number variants (CNVs) are a measure of a structural variation where specific regions of DNA are copied, often leading to altered gene copy numbers. The CNVs can range in size and are typically measured between 1Kb to several Mbs (Thapar *et al.*, 2016). Presence-absence variation (PAVs) are CNVs that are present in some genomes of the species while absent in others.

Gene expression levels may be affected by CNVs and PAVs: it may be elevated due to tandem gene duplication, interspersed gene duplication or duplication of enhancer sequence; and gene expression levels may be decreased due to complete gene deletion, partial gene deletion and insertion of duplicated sequence.

The mechanisms behind CNVs may not be fully known, but nonallelic homologous recombination (NAHR) (Gu *et al.*, 2008) and fork stalling and template switching (FoSTeS) caused by DNA replication errors may be involved (Lee *et al.*, 2007; Zhang *et al.*, 2009).

CNVs are often affiliated with genetic disorders in mammals and have furthermore been shown to be connected with human phenotypes and several diseases (Weischenfeldt *et al.*, 2013; Zmienko *et al.*, 2014). It has also been shown to include adaptive traits in plants, such as flowering time in wheat (Díaz *et al.*, 2012). In potato, it has been shown that many genes associated with CNVs are related to pathogen resistance and abiotic tolerance (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020). Furthermore, SNPs can be used for predicting genetic relationships (Cao *et al.*, 2011; Hardigan *et al.*, 2016).

There are two main methods to detect CNVs, array-based comparative genome hybridization (CGH) and reference genome based NGS (Zmienko *et al.*, 2014). Also, pan-genome studies are able to capture the structural variations between genotypes within a species.

2.5.2 Structural Variation Detection with 10X Genomics Reads

10X Genomics synthetic long-read technology offers a high physical coverage for genome assemblies, and this makes 10X a proper platform for SV detection while allowing for haplotype phasing (Ho *et al.*, 2018). There are two main methods of SV detection with the 10X platform: the read-cloud method and split alignment within synthetic long-reads analysis.

Read-clouds are clustered short reads with identical barcodes. The read-cloud method investigates the overlapping barcode density changes and distant genomic loci that share more than average barcode overlap. Long Ranger (Zheng *et al.*, 2016) and GROC-SVs (Spies *et al.*, 2017) methods uses read clouds to detect SVs (Table 1).

Another approach investigates split alignments within synthetic long-reads that are constructed with barcode information. LinkedSV (Fang *et al.*, 2019), NAIBR (Elyanow *et al.*, 2018) and VALOR2 (Karaoğlanoğlu *et al.*, 2020) programs uses this method to call SVs. Novel-X (Meleshko *et al.*, 2019) focuses on calling insertions at the size of 300bp.

Table 1: 10X Genomics SV callers and their descriptions.

SV CALLERS	DESCRIPTION
Long-ranger	Long-ranger uses linked read data information to detect breakpoints in large-scale SVs. In the regions with large number of overlapping barcodes the algorithm searches for all pairs of genomic loci (Zheng *et al.*, 2016).
GROC-SVS	GROC-SVs performs local reassembly of the breakpoints to detect SVs in sizes of 10-100 kb (Spies *et al.*, 2017).
LINKEDSV	LinkedSV investigates overlapping barcode and enriched fragment endpoints to detect SVs (Fang *et al.*, 2019).
VALOR2	VALOR2 uses split molecule and read pair signatures to detect SVs (Karaoğlanoğlu *et al.*, 2020).
NOVEL-X	Focuses on the insertions that cannot be found with other programs through reassembling the unmapped reads (Meleshko *et al.*, 2019).
NAIBR	NAIBR uses a probabilistic model combining multiple signals in barcoded reads (Elyanow *et al.*, 2018).
ZOOMX	ZoomX traces the coverage in linked read molecules (Xia *et al.*, 2018).

2.6 Pan-genomics

The pan-genome of a species is defined as the sum of the core and the accessory genome of a species. Discovering new genes in newly sequenced species of *Streptococcus agalactiae* led to introduction of a new concept in the literature (Tettelin *et al.*, 2005). The pan-genome by definition means whole genome, it represents the sum of the core genome, which is the set of genes that all individuals of the species share, and the accessory genome, which is the set of genes that are present in one to many, but not all of the individuals within the species (Medini *et al.*, 2005; Tettelin *et al.*, 2005).

Once the pan-genome was defined, the size of the pan-genome of each species became relevant. A series of analyses were made in order to estimate the sizes of the pan-core and -accessory genomes of different species. The concepts of open and closed pan-genome were also introduced (Tettelin *et al.*, 2005). An open pan-genome is a pan-genome of a species for which the rate of new gene discovery does not converge to zero. A closed pan-genome is a pan-genome of a species for which the new gene discovery rate converges to zero (Tettelin *et al.*, 2005; Snipen *et*

al., 2009). In an early pan-genome study, it was found that adaptation to different habitats increases the size of the core genome of *Streptococcus* species, thus this species has an open pan-genome (Lefebure & Stanhope 2007).

There have been numerous studies on traits of the accessory genomes as well. The first evidence of impact of an accessory genome was shown in a study suggesting half of the maize genome consists of transposable elements (TEs) and arguing that the accessory genome provides regulatory elements (Morgante *et al.*, 2007). Some dispensable genes proved to be affecting biotic stress in soybean species (Li *et al.*, 2014).

2.7 Pan-genome Analyses Techniques

There have been two approaches to construct a pan-genome. The first approach is the one suggested by Tettelin *et al.*, (2005). This approach is based on annotation of whole-genome assemblies. Once annotation of the genomes is complete, the pan-genome construction is done via comparing each genome annotation. The second approach is based on iterative mapping of the reads to the available reference genome of the species (Golicz *et al.*, 2016; Montenegro *et al.*, 2017; Hurgobin *et al.*, 2018; Ou *et al.*, 2018). Also, *de novo* assembly of all the species can be performed prior to pan-genome creation (Schatz *et al.*, 2014; Zhao *et al.*, 2018; Gao *et al.*, 2019). Various tools used during pan-genome analysis and construction are presented in the Table 2.

2.7.1 Pan-genome Construction

Pan-genomes consist of two or more genome assemblies and their alignments with functional annotation data. In order to construct a pan-genome one should select a pan-genome assembly approach and an annotation pipeline. Annotation of the genomes is important in order to assess the functions of core and accessory genomes of the species. If all the genomes are well annotated, all the novel or orthologous genes a species is bringing into the pan-genome can be traced. The two main approaches to construct a pan-genome are iterative mapping on an available reference genome and *de novo* assembly of all the species and constructing the pan-genome with the assemblies. A previously constructed pan-genome can be extended and improved through RNA sequencing and aligning RNA-seq reads on the pan-genome. Other Genome Wide Association Study (GWAS) techniques such as SNPs and Linkage Disequilibrium can also be employed to improve pan-genome assembly (Hirsch *et al.*, 2014).

Table 2: Pan-genome tools.

Publication	Name	Platform	Phyla	Orthologous /Sequence	Features
(Laing *et al.*, 2010)	Panseq	Web	-	Sequence - Alignment	Panseq depends on sequence alignments such as Local, Pairwise and Multiple Sequence alignments
(Brittnacher *et al.*, 2011)	PGAT	Web	Prokaryotes	Orthologous	Database creation with input sequences and orthologues gene analysis.
(Zhao *et al.*, 2011)	PGAP	Stand-alone app – Linux	-	Orthologous	App offers pan-genome analysis from sequence to downstream. Including construction of the pan-genome, size estimation, genetic variation detection phylogenetic analysis and annotation.
(Benedict *et al.*, 2014)	ITEP	Tool	Microbial	Orthologous	SQLite database of all-to-all BLASTP and BLASTN comparison between sequences is produced, and Markov Cluster algorithm is used to cluster. Numerous paths can be followed by database construction such as, extracting a subset from the alignments such as core genes.
(Contreras-Moreira & Vinuesa 2013)	GET_ HOMOLOGUES	Software package	Microbial	Orthologous	Three different ortholog clustering algorithms, OrthoMCL(**ref**), COGtriangles(**ref**) and bidirectional best hit algorithm, all using BLAST. Protein domain con
(Zhao *et al.*, 2014b)	PanGP	Web	Bacterial		This tool is created in order to sample when studying a large dataset. Sampling can be done using one of the two methods which are totally random, and distance guided.
(Blom et al., 2009; Blom et al., 2016)	EDGAR	Web	Bacterial	Orthologous	Database containing bacterial pan-genomes. Core genome estimation by orthology and core genome analysis with multiple sequence alignment, incorporated comparative view, synteny plots, phylogeny trees.
(Fouts et al., 2012)	PanOCT	Tool (PERL)	Prokaryotes	Orthologous	Ortholog Clustering Tool of closely related prokaryote species, micro synteny and conserved gene neighborhood is used to detect orthologs.
(Lukjancenko *et al.*, 2013)	PanFunPro	Tool / Web / Stand-alone app	-	Orthologous	Pan-genome analysis tool, identifying ORFs, finding homologous proteins from domains such as Pfam-a, TIGRFAM (**ref**) and Superfamily (**ref**), unmatched proteins are defined with MHH-based algorithm. Resulting pan-genome includes functional profiles of the core genome.
(Paul *et al.*, 2015)	PanCoreGen	Stand-alone app	Microbial	Orthologous	Pan-genome is created through picking different reference-genomes out of the input. Excel file is created as an output containing a list of all gene groups such as core, strain-specific, mosaic.

2.7.1.1 Iterative Mapping

The iterative mapping method is one of the most used methods in order to construct a pan-genome. Read sequences are mapped on an already assembled reference genome of the species and the pan-genome is obtained from the assembly.

2.7.1.2 *De novo* Assembly

The *de novo* assembly method is an independent assembly of the genomes of each species without a reference genome. These genomes will be included in the pan-genome, which is created by multiple sequence alignment of these assemblies. A series of different assemblers can be used to *de novo* assemble genomes different methods are discussed later.

2.7.2 Pan-genome Annotation

Annotation of the constructed core and accessory genome is needed to make meaning out of the data that is compiled. Annotating the accessory genome is important for discovery of stress tolerance genes or their orthologs.

An annotated pan-genome allows researchers to see what functions and genes are shared amongst which groups or individuals within the species. This becomes very important while studying species that are genetically and geographically dispersed and diverse. Researchers can see the impact of a region on the accessory and the core genome and the function these changes are offering.

2.8 Plant Pan-genomes

Plants have highly complex and repetitive genomes. Genomic studies of plants are computationally challenging and time consuming. Thus, there are currently fewer pan-genome studies of plants compared with prokaryotes. Table 3 shows the plant pan-genomes published to date.

*2.8.1 Maize (*Zea mays*)*

Maize was the first plant to be considered within the pan-genome concept. In a preliminary study, four randomly selected regions from inbred lines B73 and Mo17 shared on average 50 % of the sequences (Morgante *et al.*, 2007). In a follow-up study comparing the same lines 180 sequences were annotated as present in one line while absent in the other and it is suggested that this difference causes transcription differences between two lines, which in turn lead to phenotypic diversity and heterosis (Springer *et al.*, 2009b). When the B73 reference genome was compared with six inbred lines 570 genes were found to be absent in the reference while present in one to many of the six lines (Zheng58, 5003, 478, 178, Chang7-2 and Mo17), additionally, 296 genes were present in the B73 but were absent in at least one of the inbred lines (Lai *et al.*, 2010). In a transcription-based study including 503 diverse maize inbred lines, the genes represented by 8681 representative transcripts assemblies (RTAs) were found to be absent from B73 reference genome (Hirsch *et al.*, 2014). When *de novo* assembly of the F2 European inbred line was compared to the B73 reference genome, 395 new genes were revealed (Darracq *et al.*, 2018).

2.8.2 Rice (Oryza sativa)

Three rice subpopulations *aus*, *indica* and *temperate japonica* were sequenced, and *de novo* assembled for the first rice pan-genome study. Subpopulation selection was done based on their properties such as disease resistance, previous high-quality assembly and genetic variation. All three genomes were aligned and compared for genome specific and shared regions. The core-genome of these three species was found to be 302.9 Mbp while the accessory genome for each genome ranges from 4.8 Mbp to 8.2 Mbp (Schatz *et al.*, 2014). In the 3K Rice Genome project, 3,024 rice accessions were sequenced and assembled, and the ‘map-to-pan’ approach was used to construct the pan-genome using Nipponbare reference genome (Kawahara *et al.*, 2013; Zhao *et al.*, 2018). This approach includes first *de novo* assembly of each accession then mapping these alignments onto the Nipponbare reference genome to determine unaligned contigs. Unaligned reads are then cleaned from contamination, and non-redundant sequences were merged with the Nipponbare reference genome to create the pan-genome.

2.8.3 Cabbage (Brassica oleracea)

The *Brassica* genus has several diploid and tetraploid species from three genomes of A, B and C (Figure 1) (Nagaharu & Nagaharu 1935).

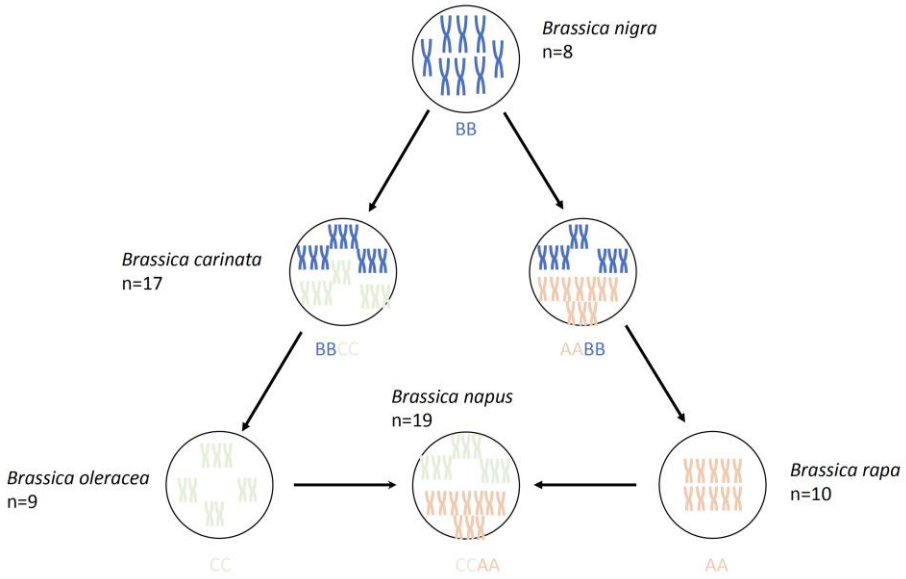

Figure 1: Triangle of U, Brassica genomes (Kumar et al., 2015).

Triangle of U is a model of *Brassica* genus. The model consists of three diploid plant species in the corners (*Brassica rapa*, *Brassica oleracea*, and *Brassica nigra*) and three amphidiploid species, result of hybridization of the diploid species, on the bases of the triangle (*Brassica napus*, *Brassica juncea*, and *Brassica carinata*) [Modified and used with permission].

The *Brassica oleracea* species is diploid and has nine chromosomes. The pan-genome is built with nine *B. oleracea* varieties and the wild relative *Brassica macrocarpa*. This pan-genome can also be viewed as the *Brassica* C pan-genome because *B. oleracea* and *B. macrocarpa* are both diploid species with a CC genome. The resulting pan-genome assembly was 587 Mbp in size and contained 59,225 gene models, 81.3 % of which are considered the core genome, while 2.2 % of the genes are only present in one species. It was found that variable genes were shorter and had higher transposable elements (TE) density around them. Variable genes were further analyzed based on their functions, and they were mostly involved in processes like disease resistance, defence response and water homeostasis.

It was also found that *B. macrocarpa* comprises the most variable genes, indicating that because a wild relative contains more variable genes, it may be that domestication caused some gene losses (Golicz *et al.*, 2016). This is also consistent with findings in other plant species such

as tomato (Gao *et al.*, 2019). Association between the PAV genes and disease resistance were also found in other species like rice and soybean (McHale *et al.*, 2012; Xu *et al.*, 2012).

2.8.4 Wheat (Triticum aestivum)

Wheat is an allohexaploid and due to its very repetitive and large genome, is challenging to study. The first *de novo* assembly of the Chinese Spring cultivar (International Wheat Genome Sequencing 2014) was performed with Velvet (Zerbino & Birney 2008) a program using non-parallel *de-bruijin* graphs. A pan-genome analysis was done with 18 other cultivars using the reference genome read mapping approach (Montenegro *et al.*, 2017). The pan-genome contains 128,656 predicted genes with 82,725 genes making up the core-genome. Amongst all the genomes Chinese Spring has the most sequence differences of the cultivars. An estimated average of 49 unique genes are introduced per cultivar and added to the pan-genome (Montenegro *et al.*, 2017).

2.8.5 Pepper (Capsicum annuum)

Pepper is a diploid crop with 12 chromosomes. The *Capsicum* pan-genome was constructed with 383 cultivars (Ou *et al.*, 2018). The genome of the Zunla-1 cultivar (Qin *et al.*, 2014) was used as a reference genome. Reads were aligned to the 12 chromosomes of the reference genome and scaffolds that are unordered collected under Chr00. PAV analysis was done with only high quality (HQ) genes based on their AED scores and on their relationship with TEs. A total of 28,840 (55.7%) genes were found in all four species. A website (http://www.pepperpan.org:8012/) was developed to visualize reference genome and the cultivars making up the pan-genome (Ou *et al.*, 2018).

2.8.6 Tomato (Solanum lycopersicum L.)

Tomato is a diploid crop with 12 chromosomes. The tomato pan-genome comprises the genome sequences of 725 tomato accessions from different selected botanical types (Gao *et al.*, 2019). Each accession was individually *de novo* assembled. Novel sequences were found through a reference genome comparison. In total, 4,873 protein-coding genes were found that are missing from the previous tomato reference genome. The pan-genome was found to contain 40,369 protein-coding sequences. Low expression rate of the non-reference genes agreed with the rice pan-genome analysis (Zhao *et al.*, 2018).

The PAV detection was done with a smaller group, 586 accessions. Categorization of the genes was done with a similar technique to Gordon *et al.*, (2017), PAVs were classified according to frequency of the genes, and core genes, softcore genes, shell genes and cloud genes were identified. Core genes are defined as shared between 100% of the accession, the softcore genes are the ones that are present in over 99% of the accession, shell genes are the genes that are present in 1-99% of the accessions while cloud genes are only present in less than 1% of the accessions. The core genome of the tomato was found to be 74.2 % of the gene content.

This study of 725 accessions of tomato showed that during domestication and improvement, the genetic diversity of the tomato cultivars has been reduced. The authors speculate that this is due to negative selection of nonutilized defense genes or random loss due to lack of positive selection (Gao *et al.*, 2019).

2.8.7 Potato (Solanum sp.)

Diploid species of potato has a high genetic diversity that can be utilized to improve crop yield in light of an increasing world population while enhancing resistance to biotic and abiotic stresses (Kyriakidou, 2020). A diploid potato pan-genome was constructed from eight potato species from the International Potato Center (CIP) germplasm collection (Kyriakidou, 2020) which included four cultivated diploid landraces; *S. stenotomum* subsp. *goniocalyx* (two accessions; GON1 and GON2), *S. phureja*, *S. xajanhuiri*, *S. stenotomum* subsp. *stenotomum*, and the wild species *S. bukasovii* (Hawkes, 1990) along with previously published reference genomes, *S. commersonii* (Aversano *et al.*, 2015b) and *S. chacoense* (M6) (Leisner *et al.*, 2018). First, all the new genomes were *de novo* assembled and a pan-genome was constructed using map-to-pan approach (Wang *et al.*, 2019). All the *de novo* assemblies were aligned to DM1-3 v 4.04 potato reference genome (Hardigan *et al.*, 2016) and unaligned contigs were extracted into a single FASTA file. Constructed FASTA file was cleaned from contamination and redundancy and finally added to the DM1- 3 v4.04 reference genome to make up the final pan-genome. Pan-genome annotation was performed through annotating all the individual genomes. PAV of the genes were predicted by aligning Illumina reads against the pan-genome and estimated with gene body coverage and CDS coverage according to (Sun *et al.*, 2017) only genes with gene body coverage > 80% and CDS coverage > 95% were called present in the genome.

Diploid potato pan-genome size was found to be 921,447,870 Mbs and included 39,751 genes. The core genome contained 28,208 genes while the accessory genome had 11,543 genes. A total of 723 newly annotated genes were not found in the initial DM1-3 reference genome. DM1-3 reference genome found to have 555 unique genes while the rest of the genomes had their 547 unique genes combined. The newly annotated 723 genes were found to be highly important for improving crop resistance such as adaptive processes, such as fertility, flowering timing, fruit and tuber development and shape, and pest and pathogen defense.

Table 3: Pipelines followed to construct major crops' pan-genomes.

Publication	**Crop**	**Assembly Method**
(Kyriakidou et al., tbp)	Diploid Potato *(Solanum Tuberosum)*	*De novo* assembly Map-to-pan
(Gao et al., 2019)	Tomato (*Solanum lycopersicum L.*)	*De novo* assembly
(Zhao et al., 2018)	Rice (*Oryza)*	*De novo* assembly
(Ou et al., 2018)	Pepper (*Capsicum annuum*)	Iterative mapping
(Montenegro *et al.*, 2017)	Wheat (*Triticum aestivum*)	Iterative mapping
(Golicz et al., 2016)	*Brassica oleracea*	Iterative mapping
(Cao *et al.*, 2011)	*Arabidopsis thaliana*	Iterative mapping
(Yu *et al.*, 2019)	Sesame (Sesamum indicum L.)	Iterative mapping
(Hirsch *et al.*, 2014)	Maize (*Zea mays)*	RNA-seq based
(Darracq *et al.*, 2018)	Maize (*Zea mays)*	*De novo* assembly

2.9 Potato Reference Genomes: DM, M6, COM, GON1, SOL and RH

De novo assembly of a genome is a long and sophisticated process. Having a reference genome to map the sequence reads onto makes the process easier. The first potato reference genome was published in 2011 by the Potato Genome Sequencing Consortium (Potato Genome Sequencing *et al.*, 2011). To overcome the challenges of the highly heterozygous diploid potato genome, a homozygous doubled monoploid of *S. tuberosum* group Phureja DM1-3 516 R44 (DM) was used. *S. tuberosum* group *tuberosum* RH89-039-16 (RH), a heterozygous diploid cultivar, was sequenced and used to map onto the anchored DM genome (Leisner *et al.*, 2018).

Genomic DNA from DM was sequenced using Whole-Genome Shotgun sequencing approach with Sanger, Illumina Genome Analyzer 2 (GA2) and Roche 454 platforms. BAC library and fosmid libraries were constructed and sequenced using the Sanger platform. To assemble the whole genome, Illumina GA2 paired-end short reads were assembled into contigs using the SOAPdenovo (Luo *et al.*, 2012) short read assembly software. To generate scaffolds from contigs, paired-end relationships, mate-paired reads, fosmid ends and BAC ends were used. Then gaps were further filled using 454 data. DNA sequencing of the RH was done on Illumina GA2 and 454 platforms. Anchoring of the contigs onto chromosomes was done with a *de novo* developed genetic map. This map consisted of sequence-tagged-sites (STS), simple sequence repeats (SSR), SNPs and diversity array technology data (DArT). Two approaches were taken during anchoring of the genome. The first approach used a genetic map to anchor the contigs and the second approach used the RH ultra-high-density linkage map (Potato Genome Sequencing *et al.*, 2011). The AFLP's of RH genetic map was linked to DM with BLAST alignments. Overall, their assembly was 727 Mb (93.9% non-gapped), which is 117 Mb less than the estimated genome size of 844 Mb (Bennet & Leitch 1997). It was found that 62.2% of this assembled genome is repetitive sequences while 29.4% is transposable elements. They were able to anchor 86% of the assembled genome, which includes 90.3% of the predicted genes (39,031) (Potato Genome Sequencing *et al.*, 2011).

Two years after the first potato reference genome was published, a newer version was published (Sharma *et al.*, 2013). In this version of the DM potato genome, DM was crossed with a diploid *Solanum tuberosum* Andigenum Group Goniocalyx (D) and one of the offspring was used to backcross with DM. The resulting population (DMDD) was used as a genetic source.

Genotyping was done with DArTs, SSRs, SNPs and AFLPs, and a linkage map was constructed. DM and tomato BAC- and Fosmid-end libraries with RH BAC-end libraries were aligned to DM and then aligned with Roche 454 paired end reads. Manual scaffolding was done using "link-peak" strategy. Their result was a 727 Mb net sequence assembly, which is 117 Mb less than estimated genome size. 94% of the genome is non-gapped, and the pseudomolecules contain 96% of the predicted genes (Sharma *et al.*, 2013).

The latest version of the DM reference genome (v.4.04) was published in 2016 (Hardigan *et al.*, 2016). In that study the authors present a new version of the potato reference genome that includes 55.7 Mb more than the previous assembly. This additional sequence is not part of the 12 pseudomolecules but unaligned with them and being called "chrUn" (Hardigan *et al.*, 2016).

A second diploid potato reference genome was released in 2018 (Leisner *et al.*, 2018). The diploid *S. chacoense* M6 clone was chosen for its traits such as self-compatibility, disease resistance and desirable market quality. Paired end and mate paired libraries were constructed and sequenced on the Illumina HiSeq 2500 platform. Both libraries were cleaned and assembled using ALLPATHS-LG assembler (Gnerre *et al.*, 2011). Gaps were filled with GapCloser (Luo *et al.*, 2012) with paired end libraries that were not used in the initial assembly. Transcriptomic data was assembled with genome guidance. Two different maps (Sanford *et al.*, 1996; Endelman & Jansky 2016) that were containing M6 as a parent were chosen to anchor the scaffolds. Pseudomolecule construction was done using SNPs as anchors. They were able to anchor 508 Mb of the 825 Mb assembly (the estimated size for the *S. chacoense* is 882Mb) (Leisner *et al.*, 2018).

The first *de novo* genome assembly of a wild potato species was for the *S. commersonii* clone cmm1t (COM), a tuber bearing wild potato species that is not sexually compatible with *S. tuberosum* but is resistant to several diseases (Micheletto *et al.*, 2000). The genome was sequenced using Illumina HiSeq 1000, and SOAPdenovo (Luo *et al.*, 2012) was used for assembly aided by the potato DM reference genome. Gaps were closed using GapCloser (Luo *et al.*, 2012). A total of 830 Mb was anchored to 12 pseudomolecules and 39,290 protein-coding genes were identified ab-initio using RNA-seq data. Interestingly, 126 cold-related genes were identified that are apparently missing from the DM reference genome (Potato Genome Sequencing *et al.*, 2011; Aversano *et al.*, 2015a).

A diploid potato landrace species, *S. stenotomum subsp. goniocalyx* (GON1) (CIP 702472 DOI: 10.18730/9DM*) is being published with the first diploid potato pan-genome study (Kyriakidou, 2020). GON1 draft genome *de novo* assembly was done using 10X Linked reads and PacBio long reads using hybrid genome assembly. Final assembly of GON1 was 855,795,280 bp with 6,424 scaffolds. Pseudomolecules were constructed against DM1-3 reference genome (Potato Genome Sequencing *et al.*, 2011) 468,652,731 bp long scaffolds were able to be anchored to the pseudo molecules while 387,432,960 bp was unanchored.

Two more diploid potato genomes were published in 2020. A draft assembly of a diploid *S. tuberosum,* Solyntus (SOL) was published (van Lieshout *et al.*, 2020). Solyntus is developed to be a self-compatible, vigorous and highly homozygous diploid potato line. Oxford Nanopore long-read sequencing reads (Oxford Nanopore Technologies; Oxford, UK) and Illumina TruSeq short reads (Illumina Inc; San Diego, USA) were used to assemble Solyntus draft genome assembly. First, long reads were assembled with software for Oxford Nanopore Sequencing, Canu v1.8 (Koren *et al.*, 2017), and TruSeq short reads were used to polish the resulting contigs with Pilon v1.23 (Walker *et al.*, 2014). Scaffolding was done with the aid of the reference genome available at the time, DM v4.03 (Potato Genome Sequencing *et al.*, 2011). Gene annotation was done using available annotations from previous potato and tomato annotations (Potato Genome Sequencing *et al.*, 2011; Sharma *et al.*, 2013) and (Hosmani *et al.*, 2019), respectively with the software GeMoMa v1.6.1 (Keilwagen *et al.*, 2016). Final assembly metrics for the 116 contigs were 13.3Mbs of N50 and total length of 716.1Mbs. 116 scaffolds were placed in 12 pseudomolecules with the same total length but N50 of 63.7 Mbs.

The first haplotype resolved diploid potato assembly was published for *S. tuberosum* group Tuberosum RH89-039-16 (RH). RH's pedigree includes dihaploidized tetraploid commercial varieties. In order to achieve haplotype resolution 10X Genomics reads were used with regular Illumina WGS reads. The achieved assembly was scaffolded, Oxford Nanopore Technologies and Hi-C data was used as long read technologies. Using mixed sequencing technologies with circular consensus sequencing (CCS) helped the assembly and variant detection. RH genome was assembled into 1,53 Gbs unitigs with N50 of 2.19 Mbs. RH-3 assembly was formed using the haplotypes and Hi-C data to create a more contiguous assembly of 1.62 Gbs and 24 pseudomolecules.

All reference genomes that have been published to date are interesting, but they have their own restrictions and shortcomings. Firstly, the DM reference genome is a doubled monoploid used to overcome the high heterozygosity of the potato genome. M6, on the other hand, is highly inbred for seven generations to reduce heterozygosity. To fully investigate the genetic diversity and heterozygosity of the *Solanum* genus it is important to have multiple reference genomes, and ideally a very complete pan-genome.

2.10 Solanum bukasovii (CIP 761748)

S. bukasovii is a self-compatible cold resistant diploid wild potato species originating from Central Peru (Hawkes, 1990) and it has been presumed to be one of the progenitors to the cultivated potato (Ugent 1970; Hosaka 1995; Spooner *et al.*, 2005; Hardigan *et al.*, 2015).

A single genotype of *S. bukasovii* (BUK1 - CIP 761748 DOI: 10.18730/E3AC) from the CIP germplasm collection (International Potato Center, Lima, Peru) has been sequenced and assembled previously (Kyriakidou *et al.*, 2020). In the same study, a CNV based PCA analysis of 14 potato species revealed that BUK was more distant from the rest of the clusters including other wild genomes such as COM and M6. In another study with a diverse panel of 13 potato species from CIP collection including *S. bukasovii* (BUK1 - CIP 761748 DOI: 10.18730/E3AC) and a second diploid individual from the same accession: BUK2 (GenBank accession number of the plastome: MT120867) was used to investigate the structural variation in their plastome. In this study, it was found that BUK2 had the highest number of SNPs with 458 SNP sites. Chloroplast types of these two *S. bukasovii* individuals were also found to differ, BUK1 has a S-type chloroplast DNA type while BUK2 has a W2-type chloroplast DNA. In the most recent study on the 13 diverse potato taxa panel from CIP, mitogenome assemblies of these species were published (Achakkagari *et al.*, 2021). It was found that while all the other 12 mitogenomes have three independent circular molecules, BUK2 has a single circular mitogenome which does not include a molecule 1 and only has molecule 2 and 3. Furthermore, a phylogeny based on the mitogenomes of this study showed that BUK2 did not cluster with the rest of the panel including BUK1 (Achakkagari *et al.*, 2021).

2.11 Genome Browsers

As in every aspect of research, visualisation brings new perspectives. Genome browsers are used to visualize genome sequences, RNA sequences and genome annotation data. Sequencing information is presented with coordinates, and corresponding RNA-seq (transcripts), annotation and other information is displayed on the tracks parallel to one another.

Web-based genome browsers can be implemented on custom websites. For example, JBrowse (Skinner *et al.*, 2009) and UCSC (Karolchik *et al.*, 2009) allow users to implement the browser applications (Wang *et al.*, 2019).

2.11.1 JBrowse

JBrowse is a genome browser that is built with JavaScript and HTML5 with supporting Perl scripts (Buels *et al.*, 2016). It is fast and easy to implement and can be used both as a stand-alone website and as a plug-in. It can be used to visualize a reference genome as a FASTA file displaying the bases with colors and peptides with a sliding window. Later, various files can be visualized using the reference genome as a backbone. These files are: Variant Calling Files (VCF), which contain information of the variants in specific positions; BAM files, which are binary versions of SAM files and contain information about sequence alignment; and Generic Feature Format (GFF files), which contain annotation data and the location of it in the reference genome.

In the present thesis, a study on a wild diploid potato genome, *S. bukasovii*, which is thought to be one of the nearest wild relatives to cultivated potato, is presented. It was hypothesized that two *S. bukasovii* genomes from the same GenBank accession have SNPs compared to each other and BUK2 has structural variance when compared to potato reference genomes. Whole genome analysis of new individuals is important for gene conservation. As wild relatives of potato may harbour valuable resources genes for traits such as biotic and abiotic stress resistance.

It has been sequenced, assembled, and compared with the current reference genomes, and visualized in a genome browser at the Potato Genome Diversity Portal (https://potatogenomeportal.org).

3 MATERIALS AND METHODS

3.1 *De Novo* Assembly of *Solanum bukasovii* And Structural Variation Analysis

3.1.1 Plant material and Sequencing

Solanum bukasovii (BUK2 – CIP 761748 - BioSample: SAMN12730757) is a Peruvian potato accession from the germplasm collection at the International Potato Center (CIP) in Lima, Peru. Genomic material was extracted from young leave samples and sequenced with 10X Genomics' GemCode technology (https://www.10xgenomics.com/) at Novogene (China).

3.1.2 10X Supernova Assembly

10X Genomics reads were assembled using Supernova assembler (Weisenfeld *et al.*, 2017) with `*--maxreads='all'*` parameter. Haplotypes were extracted from the assembly using Supernova '*mkoutput –style=pseudohap2*' arguments.

3.1.3 Decontamination of the Assembly, Scaffolding and Quality Assessment

Prior to scaffolding the pseudohaplotype contigs, BUK2 assembly was filtered using the BUK2 chloroplast (Achakkagari *et al.*, 2020) and a collection of available mitogenomes from *Solanum* species. A BLAST database was created to remove anything that matched with prokaryotic genomes. Human genome and UniVec databases were used to filter the additional contamination sequences. Tigmint with '*-arks*' parameter was used to scaffold the filtered contigs (Jackman *et al.*, 2018). Quality assessments were done with Quast and BUSCO (Gurevich *et al.*, 2013; Seppey *et al.*, 2019).

3.1.4 Alignment

Five publicly available genomes; DM v6.1 (Pham *et al.*, 2020), M6 v4.1 (Leisner *et al.*, 2018), GON1 (Kyriakidou, 2020), RH (Zhou *et al.*, 2020), SOL (van Lieshout *et al.*, 2020) and the final assembly of BUK2 was used to determine SNP and CNV events with BUK2 and BUK1 (Kyriakidou *et al.*, 2020) respectively. DM v6.1 and M6 v4.1 reference genome were downloaded from the Spud DB Potato Genomics Resource website (http://solanaceae.plantbiology.msu.edu/ taken on 1 October 2020). All the reference genomes were indexed using BWA MEM v. 0.7.17 (Li 2013). 10X Genomics sequences of BUK2 were run through LongRanger tool from 10X Genomics, with '*basic*' parameter. The sequencing reads of BUK1 were trimmed with parameters:

TruSeq3-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50 using Trimmomatic v0.36 (Bolger *et al.*, 2014) and the quality of the reads were checked with FastQC (Andrews 2010). Reads were aligned to the reference genomes using BWA MEM (Li 2013). Alignments then sorted and indexed using SAMtools v. 1.9 sort and index parameters (Li *et al.*, 2009). Duplicates were marked using MarkDuplicates by Picard tools v.2.18.9 (Institute 2016). SAMtools was used to remove unproperly oriented reads and unaligned reads with view parameter.

3.1.5 Alignment Summary Results

SAMtools depth was used to calculate the average dept of coverage. Bedtools genomecov was used to calculate the base pairs covering the reference genome (Quinlan & Hall 2010).

3.1.6 SNP Analysis

Freebayes was used to call and detect the SNPs in the alignments (Garrison & Marth 2012). The called SNPs were filtered using VCFlib's vcffilter with following criteria; mapping quality < 20, MQM < 20, MQMR < 20 and SAF && SAR < 0. SNPs that passed these criteria were annotated using snpEff tool (Cingolani *et al.*, 2012).

3.1.7 SNP phylogeny based on GBS

1078 SNPs from 447 Wild potato individuals was used to construct a phylogeny tree with the available reference genomes and BUK1 and BUK2 (SNP array from 447 wild individuals are kindly provided by Dr. Noelle Anglin at CIP). Since the GBS Array is based on DM v.4.03 (Hardigan *et al.*, 2016), the coordinates of the SNPs were transferred to DM v.6.1 (Pham *et al.*, 2020) coordinates using the FLO pipeline (Pracana *et al.*, 2017) and CrossMap (Zhao *et al.*, 2014a) to obtain a new vcf file with DM v.6.1 coordinates. Previously mentioned BWA (Li 2013) pipeline was used to map the reads from M6 v4.1 (Leisner *et al.*, 2018), GON1 (Kyriakidou, 2020), RH (Zhou *et al.*, 2020), SOL (van Lieshout *et al.*, 2020), BUK1 (Kyriakidou *et al.*, 2020) and BUK2 genomes. After the duplicates were marked, only positions in SolCap array were kept. All the BAM files were merged and Freebayes (Garrison & Marth 2012) was used to call variants and a multisample VCF file was obtained. VCFtools (Danecek *et al.*, 2011) was used to extract the SolCap array regions. Variant calling file format (VCF) was converted to PHYLIP file with a

Python script vcf2phylip.py (Ortiz, 2019). Phylogenetic tree was constructed using RAXML v.8 (Stamatakis *et al.*, 2008) with GTRGAMMA substitution model and 1000 bootstrap replicates. Figtree (Robinson *et al.*, 2016) was used to view the phylogenetic tree.

3.1.8 CNV Analysis

CNVs were calculated using CNVnator v. 0.4.1 tool (Abyzov *et al.*, 2011) with adjusted window bin size to keep the RD and standard deviation at 4-5 folds. CNVs were filtered to keep the calls with longer than 1000bp's, with q0 quality < 0.5 and cutoff p-value of 0.01. Filtered CNV calls were annotated using the GFF files of DM v.6.1, M6 v.4.1 and GON1 with intansv v. 1.12.0 (Yao 2015) a package in R v. 3.6.3 (Team 2013).

3.1.9 Significantly Enriched Gene Clusters

Significantly enriched gene clusters were calculated first with dividing the reference genomes with BEDTOOLS v2.26.0 (Quinlan & Hall 2010) into overlapping 200 kb bins with intermediate step size of 10 kb and calculating the number of CNVs that are in each bin (Hardigan *et al.*, 2016). Bins that have a higher number than the mean of all windows and additional three standard deviations is determined as significantly enriched and further analyzed (Hardigan *et al.*, 2016).

3.1.10 10X Genomics LongRanger WGS Pipeline

After analyzing the results of the initial pipeline, due to the shallow coverage and its variability across the genome, the average RD value was not determined correctly and in order to remedy this bigger bin sizes were used as suggested (Abyzov *et al.*, 2011). With the bin size adjusted to keep average RD and standard deviation 4 to 5 folds, CNVnator (Abyzov *et al.*, 2011) was unable to detect the CNVs correctly and reported only deletions around centromeric regions.

Longranger WGS pipeline was used to detect CNVs with 10X Genomics reads as described (Hulse-Kemp *et al.*, 2018). First, the reference genome was processed with LongRanger v.0.2.2 mkref parameter. Picard tools CreateSequenceDictionary was used to create the dictionary of the reference sequence (Institute 2016). Then, LongRanger wgs pipeline was used. Significant enriched gene clusters that are affected by CNVs were found with using the same process described.

3.2 Potato Genome Diversity Portal

The Potato Genome Diversity Portal (PGDP) is a website that hosts the potato genome assemblies and annotations on the JBrowse genome browser. Enables researchers, breeders and genebank managers to access and interpret genome sequence information with a Graphical User Interface (GUI). PGDP is hosted on the Arbutus Cloud Resource from Compute Canada Servers.

3.2.1 Setting Up the Portal on Arbutus

Arbutus is a Cloud Resource (CC-Cloud) (arbutus.cloud.computecanada.ca) of Compute Canada Servers (www.computecanada.ca). CC-Cloud is used for hardware virtualization, to run Virtual Machines. OpenStack is used on CC-Clouds to control computers, storage and networking (Sefraoui *et al.*, 2012). OpenStack documentation is very handy throughout the setup of VMs. According to our allocations we picked VM due to our excessive needs of both space and CPU power (https://www.openstack.org/).

3.2.1.1 Setting up Virtual Machine

Navigate to Instances page under Compute in the Arbutus cloud $arbutus.cloud.computecanada.ca/project/instances/$ and click the Launch Instance button. The first step is to name the instance, add a description and select an availability zone, for this project Any Availability Zone was selected. Second step is to choose an operating system for the VM, latest version of the Ubuntu was selected. Next step is the step where the flavor of the instance is selected, and this is the step where we decide on the hardware specifications of the VM. For the VM specifications see Appendix 4.

Once the flavor and operating system are selected in order to control the traffic in the VM Security Group created prior with the mentioned specifications below is selected. Security is one of the most important things when it comes to setting up VMs and especially if the VM will include a running web server. In order to avoid man-in-the-middle attack during SSH connections with the machine, the SSH-key of the user's computer (e.g your laptop) is added in the Key Pair settings. Last setting in the instance configuration is the metadata configuration. This VM will be used as a Web Server, thus NGINX configuration under Web Server was chosen and added. NGINX is a web server and reverse proxy (Sysoev, 2004). Once the VM is launched an IP address should be

assigned to it in order to remotely connect. Click the check button next to the instance created in the Compute/ Instances list, from the dropdown menu at the end of the row select associate Floating IP. Select the IP address created for the project.

3.2.1.2 Creating Security Group Settings

A security group is created under Webserver name from Security Groups under Network menu. This security group controls the traffic to (Ingress) and from (Egress) the VM. This security group allows VM to communicate with any port range with Ether Type IPv4 and IPv6. While it only allows Ingress traffic from ports 22 (SSH), 80 (HTTP) and 443 (HTTPS). These settings are very important to secure the VM from attacks from other ports.

3.2.1.3 Ephemeral Disk Usage

All the VMs in the cloud come with 20Gb of disk space. Additional storage space is available through ephemeral disks. In order to utilize the ephemeral disk, one must locate the mount position in the file system. In our case, disk was mounted in /mnt directory.

3.2.1.4 Key Pair Settings

Using SSH-key is an important step to ensure security between your local machine and the remote connection through SSH. SSH key is generated with following command:
$ssh - keygen\ -t\ rsa\ -C$. This command creates two files id_rsa and id_rsa.pub it is very important to know to use your public SSH key during configuration. id_rsa should not be shared. Copy your public key and paste it in the Create Key Pair option after giving it a name.

3.2.1.5 Creating IP address

An IP address is a numerical label given to devices that are connected to a computer network. A Floating IP is an IP address that can be instantly moved from one Droplet to another Droplet in the same datacenter this allows for CC-Cloud users to move around the same IP address between different projects. Under Network settings select Floating IPs. Click to %Allocate IP to Project button and IP address will be created with the given description.

3.2.1.6 Connecting to the VM

Once VM is launched and associated with a Floating IP address connection can be established through SSH from the local machine that is assigned in the Key Pair settings. Use the following command to connect: $ssh\ ubuntu@your.ip.address$.

3.2.2 Setting Up NGINX and Configuring the Server

The instance was launched with NGINX preinstalled. If this step is unsuccessful the latest version of NGINX can be downloaded from https://nginx.org/en/download.html. Location of NGINX configurations can be found on $/etc/nginx$. Starting the NGINX is very easy, it is with one command $sudo\ service\ nginx\ start$ if this process is successful www.your.ip.address should give you a message saying, "Welcome to NGINX". Most important step is to change the configuration files in the NGINX folder. These files can be found under many folders depending on the version is being used. Usually it is under conf.d/ folder but in this version (nginx version: nginx/1.14.0 (Ubuntu)) configuration files were under sites-enabled/ folder. In this folder the file default must be changed in order to assign a new root directory of the website. The location of the root is changed to /mnt/www/jbrowse because that is where JBrowse will be setup. Sample default.conf used for PGDP can be found in Appendix 5.

3.2.3 Setting up JBrowse

Setting up a software in a remote machine is more challenging than setting up in your local computer because not all the dependencies (libraries, software, etc.) are preinstalled in the new VM that is launched. Another bottleneck is to efficiently find the dependencies of the software, usually software dependencies are listed in the corresponding GitHub pages of the open-source software (Github 2016). For all the instructions to download and install JBrowse see Appendix 1, Appendix 2, Appendix 3.

3.2.4 Installing Docker

Docker is a container (https://docs.docker.com/engine/docker-overview/) and can be downloaded and installed according to the official docs on their website for Ubuntu (https://docs.docker.com/install/linux/docker-ce/ubuntu/). Instructions to install Docker can be found in Appendix 7.

3.2.5 Running Bcgsc/Orca Container with a Mounted Volume

After installing Docker on the instance ORCA container (Jackman *et al.*, 2019) can be installed and run. ORCA container includes all the bioinformatics related dependencies, and it is used to create an environment to run all the bioinformatic computations. Command to run ORCA container with a mount of volume can be found in Appendix 8.

4 RESULTS

4.1 Results of *De Novo* Assembly of *Solanum bukasovii* And Structural Variation Analysis

S. bukasovii 10X Genomics reads were used to perform a *de novo* assembly of the genome and assess the structural variation between *S. bukasovii* and the available reference genomes of DM, M6, GON1, SOL and RH-1.

4.1.1 10X Genomics de novo *Assembly*

A draft genome sequence of *S. bukasovii* (BUK2 – CIP 761748 - BioSample: SAMN12730757) was assembled into two pseudohaplotypes using 10X Genomics Linked Reads. The total resulting assemblies of both pseudohaplotypes had 11,821 scaffolds and an assembly size of 617.16 Mb with an N50 of 1,869,570 bp (Table 4).

Table 4: Assembly metrics for 10X Genomics Linked Reads assemblies of pseudohaplotypes of BUK2.

Assembly metrics	*Pseudohap 1 and Pseudohap 2*
Assembly size	617,165,355 bp
Number of scaffolds	11,821
N50	1,869,570 bp
NG50	1,457,313 bp
Largest scaffold	12,226,441 bp
Average scaffolds size	52,209 bp

4.1.2 BUSCO Results

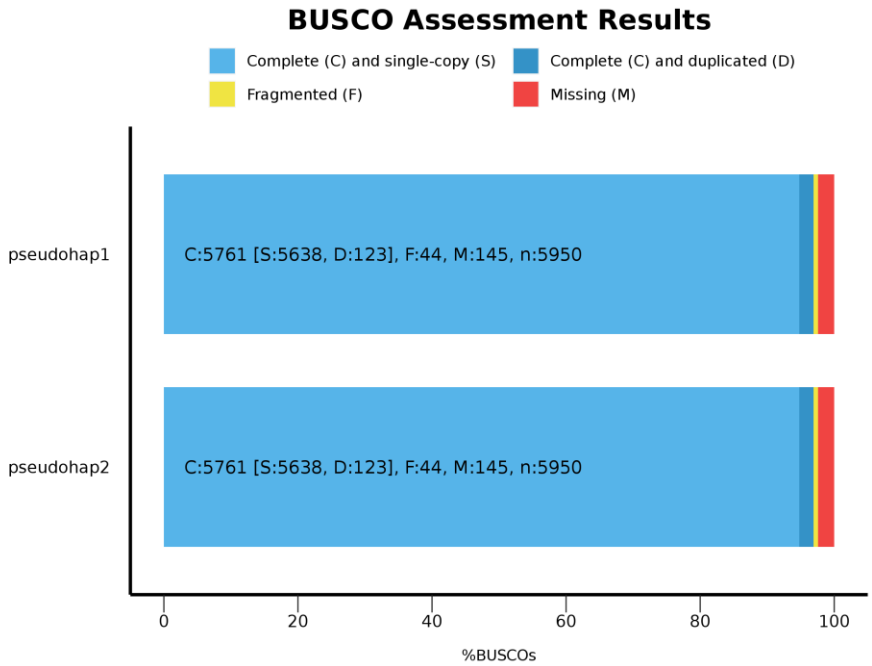

Figure 2: BUSCO results of Pseudohap 1 and Pseudohap 2.

Draft assemblies of *Solanum bukasovii* pseudohaplotypes were evaluated for completeness using BUSCO software. The results show that 96.9% (5761 genes) of the BUSCO core Plantae ortholog genes are presented in both of the pseudo assemblies. 0.7% (44 genes) of the genes are fragmented while 2.4% (145 genes) of the genes were missing.

The BUSCO results shows that both assemblies have the same gene content compared to *solanales* family genes. The subsequent results indicate that the Supernova 10X Genome assembly pipeline outputs two pseudo haplotypes that differ only in single nucleotides (SNPs) and that there are no large structural variations. Each of the assemblies have 5761 complete, 123 duplicated, 44 fragmented, 145 missing BUSCOs (C:96.9%[S:94.8%,D:2.1%],F:0.7%,M:2.4%,n:5950) (Figure 2). The Mummer, Nucmer aligner was used to assess the match of the BUK2 pseudohaplotypes with the DM v.6.01. It was found that 67.55% of the BUK2 Pseudohaplotype 1 aligns to the DM v.6.01 (Figure 3).

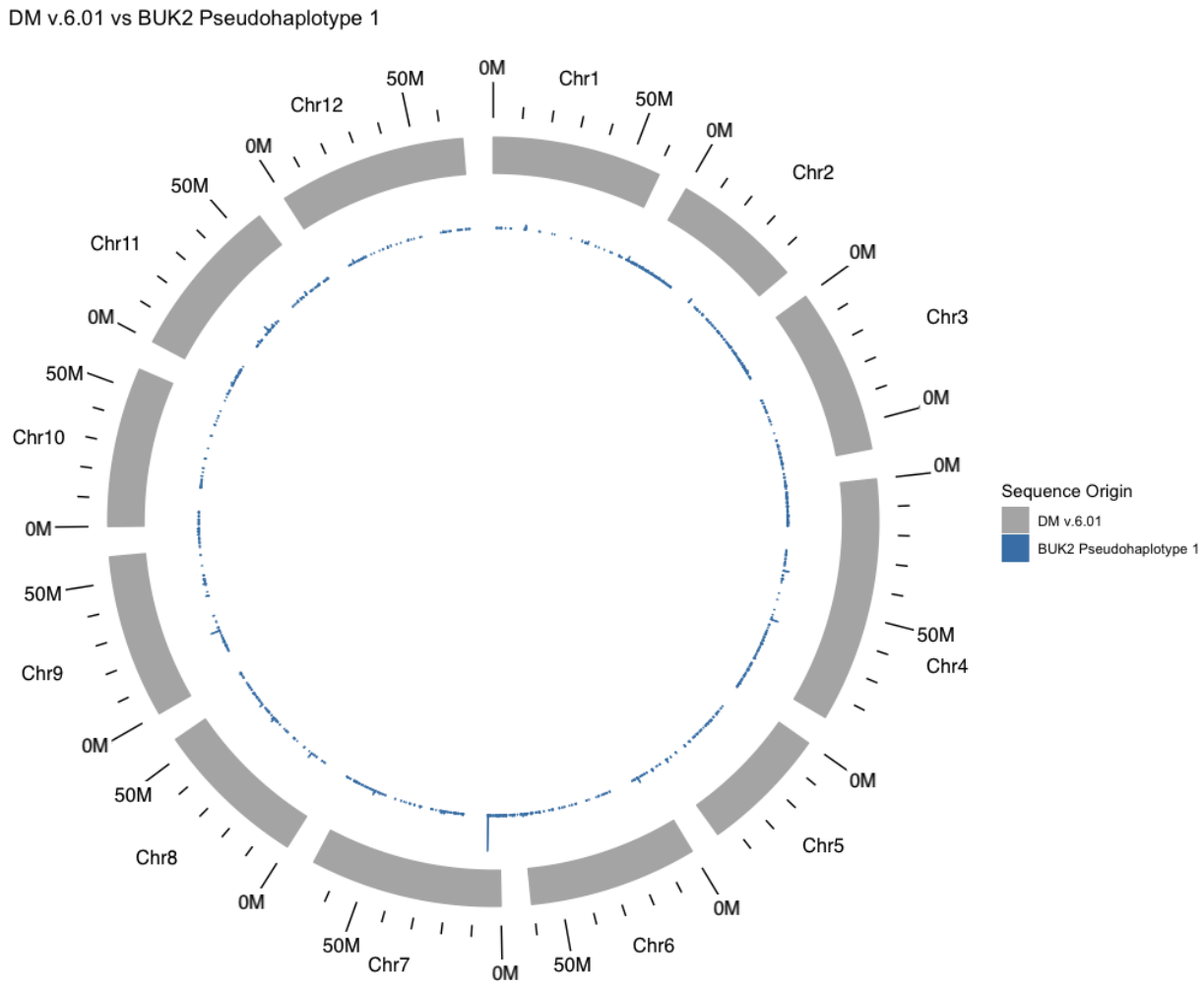

Figure 3: BUK2 Pseudohaplotype 1 alignment to DM v.6.01.

67.55% of the BUK1 assembly was found to be aligning to DM v 6.01 genome with 97% identity. Alignments were highly fragmented in chromosomes; 1, 9, 10 ,11 and 12 while alignments in chromosomes; 2, 3 and 7 were highly contiguous. High read depth is detected in the end of chromosome 6 and chromosome 9.

4.1.3 Comparing the Results of BWA MEM and Longranger Alignment and Analysis

To find variation, *S. bukasovii* 10X Genomics (www.10xgenomic.com) reads were used to align to five different reference genomes, DM v6.1 (Pham *et al.*, 2020), M6 v4.1 (Leisner *et al.*, 2018), GON1 (Kyriakidou, 2020), RH (Zhou *et al.*, 2020), SOL (van Lieshout *et al.*, 2020). In addition, the BUK1 Illumina reads were used to align to the 10X Genomics genome assembly of BUK2. BWA MEM was used for all the genome alignments and additionally LongRanger WGS pipeline was used to align BUK2 10X reads to the available five reference genomes to compare. The LongRanger pipeline increases both reference genome covered and average depth of coverage

(Figure 4). The DM-BUK2 alignment genome coverage was 352 Mbs with the BWA pipeline and the coverage increased to 531 Mbs when using the LongRanger pipeline, while the average depth of coverage improved from 32x to 43x. A similar improvement was seen for the rest of the alignments - genome coverages were improved from 326Mbs to 400Mbs in M6-BUK2, 342Mbs to 362Mbs in GON1-BUK2, and the average depth of genome coverages were increased from 35x to 49x in M6-BUK2 and 45x to 50x in GON1-BUK2 alignments with the BWA and LongRanger pipeline, respectively. For the rest of the alignments the BWA pipeline was not used due to poor performance with the 10X Genomics reads. The genome coverage was 521 Mbs for the SOL-BUK2 alignment and 496 Mbs for the RH-BUK2 alignment with the LongRanger pipeline, while average depth of coverage was 44x and 50x for the respective alignments. The results of the BUK2-BUK1 alignment showed 476 Mbs of genome coverage and 54x average depth of coverage with the BWA pipeline (Figure 4).

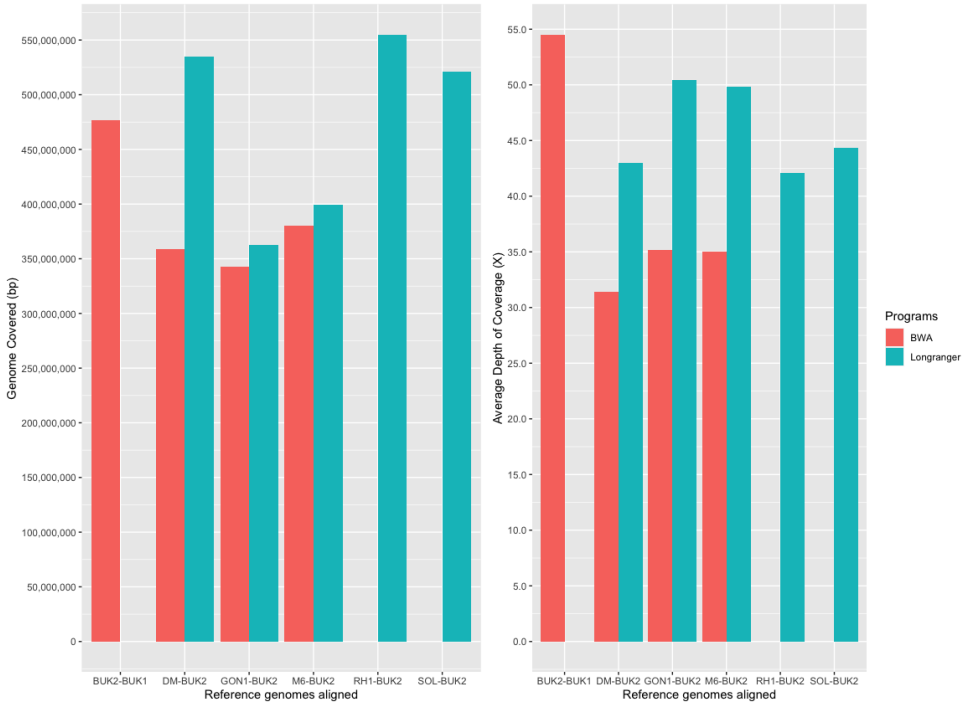

Figure 4: Alignment results for LongRanger and BWA pipelines.

Two pipelines were used to align *Solanum bukasovii* 10X Genomics reads to the published reference genomes; DM v6.1 (Pham *et al.*, 2020), M6 v4.1 (Leisner *et al.*, 2018), GON1 (Kyriakidou, 2020), RH (Zhou *et al.*, 2020), SOL (van Lieshout *et al.*, 2020) and BUK1 Illumina reads were aligned to BUK2 pseudohap1 draft assembly. Genome coverage was the highest in RH and SOL followed by BUK1. Depth of coverage was found to be the highest in BUK1, GON1 and M6. Results were both the pipelines available; DM-BUK2, GON1-BUK2 and M6-BUK2 shows that Longranger pipeline increased both the genome coverage and the average depth of genome.

The genome coverage percentage was calculated from the ratio of the size of the reference genome used and the alignment size. BUK2 aligns to 77% of GON1, BUK1, M6 (the landrace and wild potato genomes); 72% of the DM and SOL and 68% of RH-1 (Figure 5).

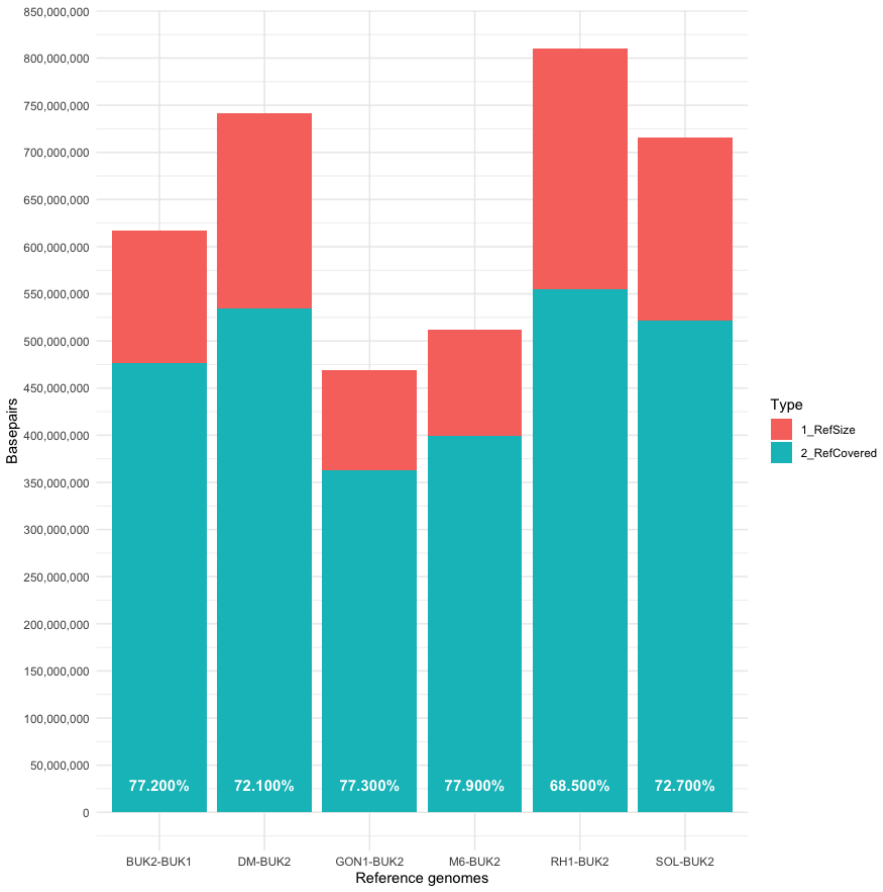

Figure 5: Genome coverage and percentage of the BUK2 alignments against reference genomes.

All the reference genomes used has different sizes, in order to compare the genome coverages, percentage of the genome covered was used. GON1, M6 and BUK1 genomes has the highest coverage percentage with 77% while DM, SOL and RH-1 has the lowest genome coverage percentage with 72%, 72% and 68% respectively.

4.1.4 SNP Results of BUK2 Alignment to Available Reference Genomes and BUK1

The number of SNPs detected in BUK2 compared to the refence genomes DM, M6, GON1, SOL RH and BUK1 ranged from 9.9 million in SOL to 3.4 million in M6 (Figure 6). 6.7 million SNPs were detected in BUK2 when compared to BUK1 (from the same CIP Genebank accession as BUK2), while 3.6 million and 4.3 million SNPs were detected in BUK2 when compared to GON1

and DM, respectively. In all genomes, SNPs affected intergenic regions the most while the lowest number of SNPs were found in the exonic regions (Figure 6). SNPs that cause missense mutations ranged from 60% to 51% in comparison to RH and GON1 respectively, while the nonsense ranged 3.2% to 1.4% in comparison to SOL and DM where SNPs that led to silent mutations ranged from 46% to 36% in comparison to GON1 and SOL respectively.

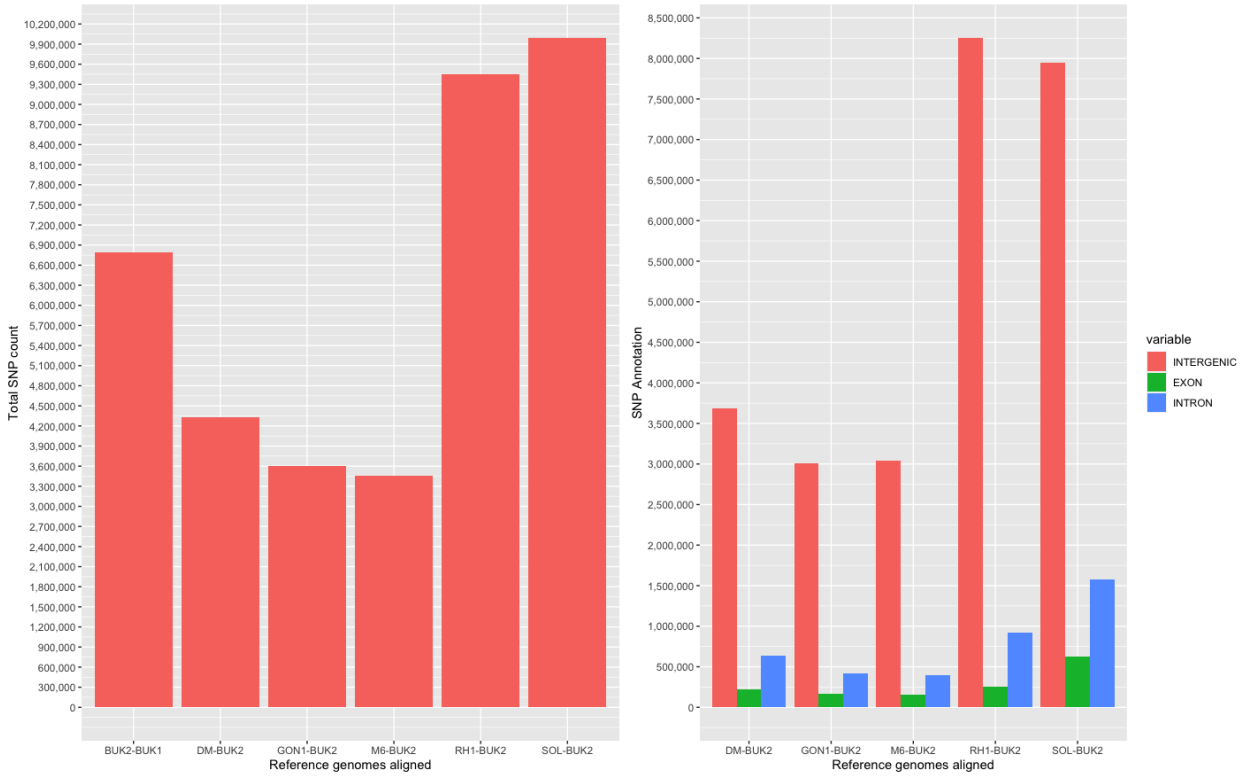

Figure 6: Total SNP count and annotation of the small variants; indels and SNPs identified in BUK2 compared to reference genomes.

SNP analysis showed that highest number of SNPs were found in SOL and RH-1 genomes. Overall, the greatest number of SNPs found to affect the intergenic regions while the fewest number of SNPs are found in exonic regions.

4.1.5 GBS SNP array phylogeny results

GBS SNP array was used to create a computational SNP based phylogeny from 1,078 SNPs with the available reference genomes BUK2, BUK1 and another 447 *bukasovii* genomes that are provided by CIP. After the alignments only 1,034 SNPs were present in all the genomes and were used to construct the phylogenetic tree.

The phylogeny showed that BUK2 clustered with *S. sparsipilum* (Bitt.) Juz. et Buk. and *S. raphanifolium* Card. et Hawkes while BUK1 clustered with other *bukasovii* genomes. Both *S. sparsipilum* (Bitt.) Juz. et Buk. and *S. raphanifolium* Card. et Hawkes species are self-compatible (Cipar *et al.*, 1964) along with the self-compatible M6 genome (Leisner *et al.*, 2018). Clusters also showed that the BUK1 and BUK2 clusters were located from the furthest end while RH and SOL clustered at the other end. This information is showing that BUK2 is closer to the wild species than is BUK1 and is likely a hybrid, possibly of *S. bukasovii* and either of *S. sparsipilum* or *S. raphanifolium* (Figure 7).

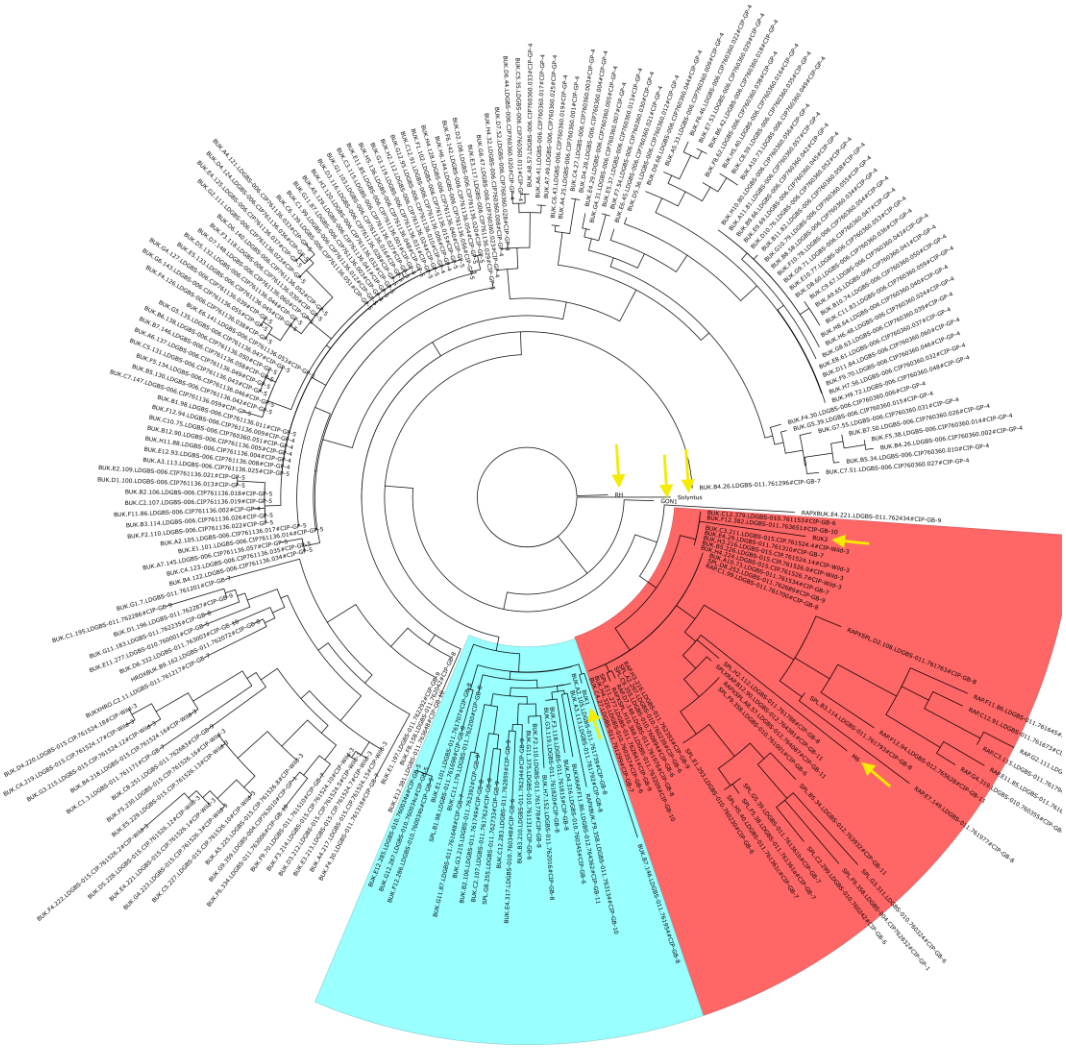

Figure 7: Phylogeny Tree based on SolCap Array.

A phylogeny analysis based on the SNPs from SolCap Array shows that BUK1 and BUK2 are not in the same cluster while BUK2 and M6 clustered together with species that are known to be self-compatible such as *S. sparsipilum* (Bitt.) Juz. et Buk. and *S. raphanifolium* Card. et Hawkes.

4.1.6 CNV Results of BUK2 Alignment to Available Reference Genomes

S. bukasovii 10X Genomics reads were used to align to five different reference genomes, DM, M6, RH, SOL and GON1 with two different pipelines to detect copy number variation (CNVs). The BWA alignment results were used to call CNV's with CNVnator (Abyzov *et al.*, 2011). For each of the alignments, various bin sizes from 100 to 100,000 was used for optimization to reach the average RD and standard deviation (SD) ratio of 4 according to the instructions by CNVnator. The bin size of 100,000 for all the alignments resulted in the favorable ratio of average RD and SD but using these larger bin sizes decreased the sensitivity to detect CNVs. As a result, zero CNVs were detected using this method with previously mentioned parameters. Upon further investigation of the causes, it was found that larger bin sizes are required when the genome coverage is too shallow, or genome coverage is non-uniform (https://github.com/abyzovlab/CNVnator/issues/206 - similar results with 10X Genomics reads.).

Due to the inconclusive results of the previous pipeline, instead the LongRanger pipeline was used to call CNVs between *Solanum bukasovii* and five available reference genomes, DM, M6, RH, SOL and GON1. As previously discussed, the LongRanger pipeline yielded improved results in both depth of coverage and genome coverage. As a result, CNVs were successfully called between *S. bukasovii* (BUK2) and the five available reference genomes.

Table 5: Number of genes affected by CNVs in each chromosome.

Significantly Enriched Gene Clusters affected by CNVs	*DM*	*M6*	*GON1*	*Solyntus*	*RH-1*
Total Number of genes effected	246	67	166	814	634
Chr1	71	26	4	153	54
Chr2	7	0	5	312	16

Table 5 continued: Number of genes affected by CNVs in each chromosome.

Significantly Enriched Gene Clusters affected by CNVs	*DM*	*M6*	*GON1*	*Solyntus*	*RH-1*
Chr3	33	7	10	24	28
Chr4	71	3	17	136	157
Chr5	28	15	35	461	101
Chr6	50	21	15	251	36
Chr7	46	28	46	243	31
Chr8	37	0	11	215	20
Chr9	33	19	2	332	63
Chr10	69	2	12	472	62
Chr11	27	0	3	88	7
Chr12	47	16	6	20	59

The data presented in Table 5 was used to create a clustered heatmap in order to view the chromosomes most affected by CNV events (Figure 8). Due to high numbers of the CNVs detected in SOL, the heatmap was skewed. Therefore, the results from SOL were designated as outliers and removed from the cluster. The results show that Chromosome 5 is the chromosome the most affected by CNV events compared to the rest of the chromosomes and chromosome 1 is the least affected. M6 and GON1 clustered together and DM and RH clustered together. This is not surprising as M6 and GON1 are less related to DM and RH, which are cultivated species.

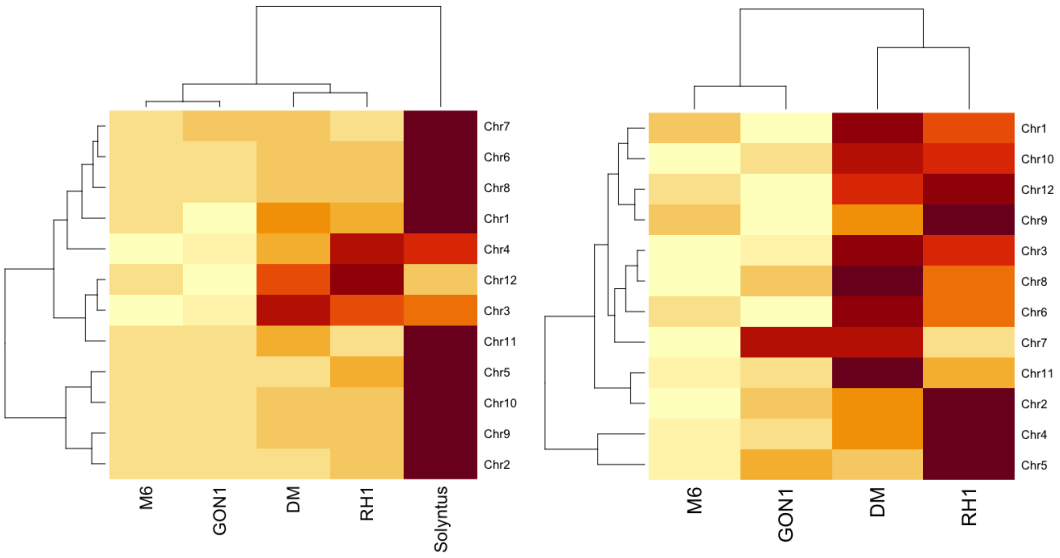

Figure 8: Clusters of number of genes affected in each chromosome.

A clustered heatmap of the number of genes affected in each chromosome was conducted to visualize the common chromosomes most affected by CNV events. In the right image, Solyntus was omitted due to very high numbers of CNV events. It was showed that chromosome 5 was the most affected while chromosome 1 was the least affected. When compared M6 and GON1 clustered together and the least affected overall, DM and RH-1 clustered together and more affected. Solyntus was the most affected by CNVs.

4.1.7 CNVs Compared to DM 6.1

Solanum bukasovii 10X Genomics linked reads were aligned to a newer version of the DM assembly, DM v.6.1 (Pham *et al.*, 2020). All CNVs were called and filtered with a 'quality > 100' parameter. Deletions were found to be more numerous than duplications and other CNVs. There were 602 deletions, 10 duplications and 57 inversions in the alignment. The average size of deletions was larger than the average size of duplications, where the average sizes found to be 1,165,095bp and 81,465bp, respectively. The largest deletion was 26,810,000bp and largest duplication was 81,465bp (Table 6). The GO enrichment results of these CNVs can be found in Table 7.

Table 6: CNV results DMv6.1 and BUK2.

DM 6.1 BUK2	*Deletion*	*Duplication*
#'s found	602	10
Max size	1,165,095bp	81,465bp
Average	26,810,000bp	81,465bp

Table 7: GO Enrichment analysis of the CNVs detected from DMv6.1 and BUK2.

	GO.ID	*Term*
1	GO:0006952	defense response
2	GO:0009733	response to auxin
3	GO:0009611	response to wounding
4	GO:0006542	glutamine biosynthetic process
5	GO:0044030	regulation of DNA methylation
6	GO:0015074	DNA integration
7	GO:0006412	translation
8	GO:0016579	protein deubiquitination
9	GO:0006032	chitin catabolic process
10	GO:0016998	cell wall macromolecule catabolic processes

Duplicated Genes Detected in BUK2 Compared to DM V.6.1

Annotation of the 10 highly enriched CNV clusters revealed 32 genes affected by duplications in BUK2. Duplication events were found in the following regions: Chr04 (~6.23Mbs - ~6.29Mbs, ~41.76Mbs - ~41.82Mbs), Chr05 (~1.25Mbs - ~1.28Mbs, ~7.68Mbs - ~7.72Mbs, ~27.68Mbs - ~27.71Mbs), Chr07 (~39.86Mbs - ~39.91Mbs, ~30.68Mbs - ~30.74Mbs), Chr09 (~30.39Mbs - ~30.42Mbs, ~58.76Mbs - ~58.80Mbs), Chr12 (~57.48Mbs - ~57.56Mbs). Twelve of the duplicated genes are hypothetical genes, and the annotation for the other 20 genes are listed in Table 8.

Table 8: Annotation of duplicated genes in BUK2 when compared to DM v6.1.

DM Gene ID	*Functional annotation of the gene*
Soltu.DM.04G006020.1 *Soltu.DM.04G006040.1*	KIX domain containing protein
Soltu.DM.04G006030.1 *Soltu.DM.04G006050.1*	Mediator of RNA polymerase II transcription subunit 15a
Soltu.DM.04G018460.1 *Soltu.DM.04G018490.1* *Soltu.DM.04G018500.1* *Soltu.DM.04G018510.1*	Basic-leucine zipper (bZIP) transcription factor family protein
Soltu.DM.05G007820.1	F-box family protein
Soltu.DM.05G007830.1 *Soltu.DM.05G007840.1*	Major facilitator superfamily protein
Soltu.DM.12G027510.1	anthranilate synthase beta subunit

Table 8 continued: Annotation of duplicated genes in BUK2 when compared to DM v6.1.

DM Gene ID	*Functional annotation of the gene*
Soltu.DM.12G027520.1 *Soltu.DM.12G027520.2* *Soltu.DM.12G027520.3*	galacturonosyltransferase
Soltu.DM.12G027530.1	Translation initiation factor IF6
Soltu.DM.12G027560.1	non-intrinsic ABC protein
Soltu.DM.12G027570.1 *Soltu.DM.12G027570.2*	Calcineurin-like metallo-phosphoesterase superfamily protein

4.1.8 CNVs Compared to M6

Solanum bukasovii 10X Genomics linked reads were aligned to M6 (Leisner *et al.*, 2018). All CNVs were called and filtered with 'quality > 100' parameter. Deletions were found to be more than duplications and other CNVs. There were 231 deletions, 9 duplications and 3 inversions in the alignment. The average deletion size was bigger than average size of duplications; average sizes found to be 2,109,031bp and 50,380bps respectively. The largest deletion was 26,370,000bp and largest duplication was 92,969bp (Table 9). The GO enrichment results of these CNVs can be found in Table 10.

Table 9: CNV results of M6 and BUK2.

M6 BUK2	*Deletion*	*Duplication*
#'s found	231	9
Max size	26,370,000bp	92,969bp
Average	2,109,031bp	50,380bp

Table 10: GO Enrichment analysis of the CNVs detected from M6 and BUK2.

	GO.ID	*Term*
1	GO:0070940	dephosphorylation of RNA polymerase II C...
2	GO:0005987	sucrose catabolic process
3	GO:0005982	starch metabolic process
4	GO:0018298	protein-chromophore linkage
5	GO:0048544	recognition of pollen
6	GO:0009649	entrainment of circadian clock
7	GO:0010569	regulation of double-strand break repair...
8	GO:0010076	maintenance of floral meristem identity
9	GO:0000422	autophagy of mitochondrion
10	GO:0006897	endocytosis

Duplicated Genes Detected in BUK2 Compared to M6

Annotation of the 9 highly enriched CNV clusters revealed 41 genes affected by duplications in BUK2. Duplication events were found in the following regions: Chr01 (~13.9Mbs - ~14.0Mbs, ~16.26Mbs - ~16.33Mbs, ~30.0Mbs - ~30.07Mbs), Chr05 (~58.01Mbs - ~58.4Mbs), Chr07 (~9446bs - ~102415bs, ~24.59Mbs - ~24.61Mbs), Chr08 (~27.91Mbs - ~27.96Mbs), Chr12 (~41.56Mbs - ~41.59Mbs, ~48.85Mbs - ~48.88Mbs). 12 of the duplicated genes are hypothetical genes, other 29 genes can be found in the Table 11.

Table 11: Genes duplicated in the BUK2 genome compared to M6.

GENE ID	FUNCTION
G3491.T1 *G3490.T1*	GDSL-like Lipase/Acylhydrolase superfamily protein
G3494.T1	RNA-binding protein
G3493.T1	uridine-ribohydrolase
G3489.T1 *G3489.T1.1.57A38F05* *G3489.T1.2.57A38F05* *G3489.T1* *G3489.T1.1.57A38F05*	Poly (ADP-ribose) glycohydrolase (PARG)
G27816.T1 *G27816.T1*	F-box/RNI-like superfamily protein
G19640.T1.1.57A38F05 *G19640.T2* *G19640.T1*	Peptidase C13 family
G19639.T1.1.57A38F05 *G19639.T1*	ent-kaurenoic acid hydroxylase
G21888.T1.1.57A38F06	homeobox protein
G21888.T1	Homeobox-leucine zipper protein family
G21885.T1 *G21887.T1*	Disease resistance protein (CC-NBS-LRR class) family
G41086.T1	SOS3-interacting protein
G39140.T1.1.57A38F04 *G39140.T1*	MLP-like protein
G39137.T1	Cytochrome c oxidase, subunit Vib family protein

Table 11 continued: Genes duplicated in the BUK2 genome compared to M6.

GENE ID	*FUNCTION*
G39139.T1	Aldehyde dehydrogenase 11A3
G39141.T1	Polyketide cyclase/dehydrase and lipid transport superfamily protein
G39138.T1.1.57A38F04	LJRHL1-like
G39138.T1	
G3493.T1	uridine-ribohydrolase

4.1.9 CNVs Compared to GON1

Solanum bukasovii 10X Genomics linked reads were aligned to GON1 (Kyriakidou, 2020). All the CNVs were called and filtered with 'quality > 100' parameter. Deletions were found to be more than duplications and other CNVs. There were 260 deletions, 4 duplications and 1 inversion in the alignment. Average deletion size was bigger than average size of duplications; average sizes found to be 16,540,000bp and 59,699bps respectively. Largest deletion was 1,649,700bp and largest duplication was 38,699bp (Table 12). The GO enrichment results of these CNVs can be found in Table 13.

Table 12: CNV results of GON1 and BUK2.

GON1 BUK2	*Deletion*	*Duplication*
#'s found	260	4
Max size	16,540,000bp	59,699bp
Average	1,649,700bp	38,699bp

Table 13: GO Enrichment analysis of the CNVs detected from GON1 and BUK2.

	GO.ID	*Term*
1	GO:0006399	tRNA metabolic process
2	GO:0006281	DNA repair
3	GO:0006952	defense response
4	GO:0006886	intracellular protein transport
5	GO:0050789	regulation of biological process
6	GO:0051726	regulation of cell cycle
7	GO:0065007	biological regulation
8	GO:0050794	regulation of cellular process
9	GO:0005991	trehalose metabolic process
10	GO:0005992	trehalose biosynthetic process

Duplicated Genes Detected in BUK2 Compared to GON1

Annotation of the four highly enriched CNV clusters revealed seven genes affected by duplications in BUK2. Duplication events were found in the following regions: gon1_pseudo_01 (~6.23Mbs - ~6.29Mbs), gon1_pseudo_03 (~1.25Mbs - ~1.28Mbs) gon1_pseudo_08 (~39.86Mbs - ~39.91Mbs), gon1_pseudo_12 (~30.39Mbs - ~30.42Mbs). Five of the duplicated genes are hypothetical genes, the other two genes are *8HGO_CATRO 8-hydroxygeraniol oxidoreductase* genes located in gon1_pseudo_03.

4.1.10 CNVs Compared to SOL

The S. bukasovii 10X Genomics linked reads were aligned to SOL (van Lieshout *et al.*, 2020). All the CNVs were called and filtered with 'quality > 100' parameter. There were more deletions than duplications and other CNVs. There were 469 deletions, seven duplications and 11 inversions in the alignment. The average deletion size was bigger than average size of duplications; average sizes found to be 24,930,000bp and 57,225bp respectively. The largest deletion was 1,462,063bp and the largest duplication was 39,376bp (Table 14). The GO enrichment results of these CNVs can be found in Table 15.

Table 14: CNV results of SOL and BUK2.

SOL BUK2	*Deletion*	*Duplication*
#'s found	469	7
Max size	24,930,000bp	57,225bp
Average	1,462,063bp	39,376bp

Table 15: GO Enrichment analysis of the CNVs detected from SOL and BUK2

	GO.ID	*Term*
1	GO:0008150	biological process
2	GO:0008152	metabolic process
3	GO:0009987	cellular process
4	GO:0044238	primary metabolic process
5	GO:0006807	nitrogen compound metabolic process

Table 15 continued: GO Enrichment analysis of the CNVs detected from SOL and BUK2

	GO.ID	*Term*
6	GO:0044260	cellular macromolecule metabolic process
7	GO:0043170	macromolecule metabolic process
8	GO:0071704	organic substance metabolic process
9	GO:0044237	cellular metabolic process
10	GO:1901576	organic substance biosynthetic process

Duplicated Genes Detected in BUK2 Compared to SOL

Annotation of the seven highly enriched CNV clusters revealed 21 genes affected by duplications in BUK2. These duplication events were found in the following regions StSOLv1.1ch04 (~50.75Mbs - ~50.79Mbs), StSOLv1.1ch05 (~5.92Mbs - ~5.96Mbs) StSOLv1.1ch09 (~46.76Mbs - ~46.79Mbs, ~48.52Mbs - ~48.54Mbs, ~61.70Mbs - ~61.75Mbs), StSOLv1.1ch10 (~7.60Mbs - ~7.64Mbs), StSOLv1.1ch11 (~32.23Mbs - ~32.29Mbs). Eleven of the duplicated genes are hypothetical genes and genes of unknown function, the other ten genes are: two disease resistance genes, late blight resistance gene, Cytochrome c1-2, heme protein, mitochondrial, Homeobox-leucine zipper protein HAT, Gag-pol protein, 'chromo' domain containing protein and Glutaredoxin.

4.1.11 CNVs Compared to RH-1

The S. bukasovii 10X Genomics linked reads were aligned to RH (Zhou *et al.*, 2020). All the CNV's were called and filtered with 'quality > 100' parameter. There were 1067 CNV events. Deletions were found to be more than duplications and other CNVs. There were 680 deletions, 10 duplications and 12 inversions in the alignment. Average deletion size was bigger than average size of duplications with average sizes found to be 1,146,795bp and 51,114bp, respectively. The largest deletion was 15,090,000bp and the largest duplication was 87,473bp (Table 16). The GO enrichment results of these CNVs can be found in Table 17.

Table 16: CNV results of RH-1 and BUK2.

SOL BUK2	*Deletion*	*Duplication*
#'s found	680	10
Max size	15,090,000	87,473
Average	1,146,795	51,114

Table 17: GO Enrichment analysis of the CNVs detected from RH-1 and BUK2.

GO.ID	*Term*	*Annotated*
1	GO:0009733	response to auxin
2	GO:0009415	response to water
3	GO:0006508	proteolysis
4	GO:0006979	response to oxidative stress
5	GO:0044267	cellular protein metabolic process
6	GO:0055114	oxidation-reduction process
7	GO:0006468	protein phosphorylation
8	GO:0009058	biosynthetic process
9	GO:0044271	cellular nitrogen compound biosynthetic process
10	GO:0006518	peptide metabolic process

Duplicated Genes Detected in BUK2 Compared to RH-1

Annotation of the 10 highly enriched CNV clusters revealed 30 genes affected by duplications in BUK2. Duplication events were found in the following regions chr1_1 (~54.48Mbs - ~54.52Mbs), chr2_1 (~9.61Mbs - ~9.66Mbs) chr7_1 (~9.44Mbs - ~9.47Mbs, ~14.54Mbs - ~14.59Mbs), chr8_1 (~34.47Mbs - ~34.56Mbs), chr9_1 (~36.81Mbs - ~36.84Mbs), chr10_1 (~44.48Mbs - ~44.51Mbs), chr12_1 (~13.31Mbs - ~13.39Mbs, ~25.88Mbs - ~25.91Mbs, ~40.05Mbs - ~40.11Mbs). Ten of the duplicated genes are hypothetical genes and genes of unknown function, the annotation for the other 20 genes is listed in Table 18.

Table 18: Gene IDs and Functions of duplicated genes in BUK2 compared to RH-1.

GENE ID	*FUNCTION*
RHC01H1G1995.2	Glycoside hydrolase superfamily
RHC02H1G0200.2 RHC09H1G1355.2 RHC10H1G1428.2	Retrotransposon Ty1/copia-like
RHC07H1G0480.2 RHC07H1G0481.2 RHC07H1G0482.2	transferase activity transferring acyl groups other than amino-acyl groups
RHC07H1G0483.2	antiporter activity xenobiotic transporter activity
RHC07H1G0484.2	Chloramphenicol acetyltransferase-like domain superfamily
RHC07H1G0683.2	protein binding
RHC08H1G1151.2	CheY-like superfamily signal transduction response regulator, receiver domain
RHC08H1G1153.2	YlmG homolog protein 2, chloroplastic

Table 18 continued: Gene IDs and Functions of duplicated genes in BUK2 compared to RH-1.

GENE ID	*FUNCTION*
RHC08H1G1155.2	Leucine-rich repeat domain superfamily
RHC12H1G0312.2	Metallo-dependent phosphatase-like
RHC12H1G0313.2	AAA+ ATPase domain ABC transporter, haem export
RHC12H1G0317.2	mature ribosome assembly
RHC12H1G0318.2	transferase activity, transferring glycosyl groups
RHC12H1G0319.2	Class I glutamine amidotransferase-like
RHC12H1G1094.2	Pentatricopeptide repeat

4.1.12 Significant Gene CNV Clusters in BUK2 Compared to All Reference Genomes

CNV enriched clusters were detected using 200kb sequence windows. In order to detect the most CNV enriched region 200kb windows were analyzed, and the largest region with the highest number of CNVs was designated as the most significant CNV clusters in the BUK2 genome when compared to reference genomes: DM, M6, GON1, RH-1 and SOL. Annotations of the reference genomes were used to determine the genes affected by these CNV events.

- **DM: chr 5 26000000:26400000**

The most enriched gene cluster in BUK2 compared to the DM genome encodes 4 hypothetical proteins.

- **M6 chr 5 28800000:29600000**

The most enriched cluster in BUK2 compared to the M6 genome is located on Chromosome 5 and there are six genes that are affected by this CNV event which include hypothetical protein, Methyl-

Cpg-Binding Domain Protein, Cytochrome P450, Family 716, Subfamily A, Polypeptide, Zinc-Finger Protein, Ribosomal L27e Protein Family, Reverse Transcriptase-Like Domain Containing Protein.

- **GON1 gon1_pseudo_02 400000:600000**

The most enriched gene cluster in BUK2 compared to the GON1 genome encodes 2 hypothetical proteins.

- **SOL StSOLv1.1ch10 7400000:8400000**

There are 62 genes affected in the most enriched gene cluster BUK2 compared to SOL, 32 genes that are enriched are hypothetical genes, and the remaining 30 genes represent 19 different functions that are highly enriched (Table 19).

Table 19: Highly enriched genes CNVs in Solyntus genome.

#	*Genes That Are Affected by CNVs*
1	Pre-Mrna Branch Site Protein P14
2	Rnase H Family Protein
3	Ta11-Like Non-Ltr Retroelement Protein
4	Non-Ltr Retroelement Reverse Transcriptase
5	Bifunctional Endo-1,4-Beta-Xylanase Xyla
6	Zinc Knuckle Family Protein
7	Zinc Finger Protein
8	Duf614 Containing Protein
9	Gag-Pol Polyprotein

Table 19 continued: Highly enriched genes CNVs in Solyntus genome.

#	*Genes That Are Affected by CNVs*
10	Class S F-Box Protein
11	'Chromo' Domain Containing Protein
12	Ribosomal Protein S10, Eukaryotic and Archaeal Form
13	Cg15040
14	Rna-Directed Dna Polymerase (Reverse Transcriptase); Ribonuclease H
15	Chromo Domain Protein Lhp1
16	Retrotransposon Protein, Unclassified
17	Structural Molecule
18	Line-Type Retrotransposon Lib Dna, Complete Sequence, Insertion at The S14 Site
19	Non-Ltr Reverse Transcriptase

- **RH-1 chr5_1 35000000: 37600000**

The most enriched cluster in BUK2 compared to RH is the 2.6Mbs window on RH-1 Chromosome 5, this window encodes 54 genes, and the only gene of known function is protein peptidyl-prolyl isomerization. The rest of the 53 genes are genes of unknown function.

4.2 Results of Potato Genome Diversity Portal

The Potato Genome Diversity Portal (PGDP) is implemented on the Arbutus cloud resources managed by Compute Canada. All the servers that the portal is using are in the Arbutus cloud resources. The platform can both run bioinformatics jobs and visualize the results on the JBrowse genome browser. At the present, there are three instances running on our allocations on Arbutus, two of them serves as processing power while the other instance is used to host the website and the JBrowse genome browser (Figure 9).

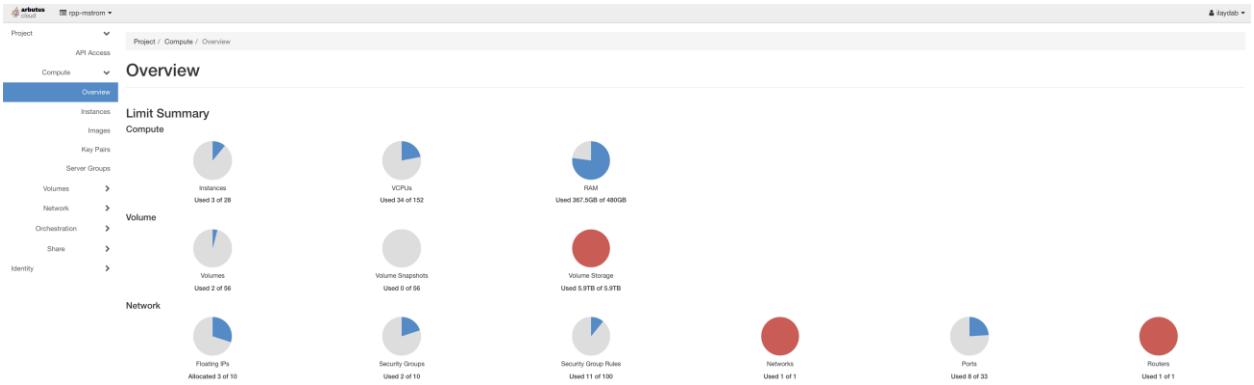

Figure 9: Potato Genome Diversity Portal allocation usage overview.

5.9TB of volume storage was used to create the PGDP. While over 80% of the allocated RAM was used. 34 VCPUs were allocated across the instances to support both genome browser and the calculations carried in the PGDP.

The landing page of PGDP, potatogenomeportal.org can help the user navigate directly to the Genome Browser to view the genome sequences of the potato genomes that are assembled and published by our lab (Figure 10). Clicking on the genome browser of choice the list of genomes available can be viewed [potato genomes (Figure 11) and their plastome (Figure 12)]. Once the genome of interest is selected, the JBrowse window will come up and the researcher can select the tracks to be visualized. The available tracks are the assembly, alignment and annotation (Figure 13, 14).

Figure 10: Potato Genome Diversity Portal landing page.

The PGDP landing page allows users to navigate through the whole website. Users can reach the Genome Browser and Chloroplast Genome Browser. Users can navigate to the Collaborators page in order to visualize our affiliations and collaborators.

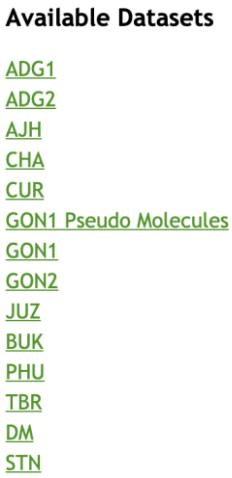

Figure 11: Genome assemblies that are available on the portal.

Available published nucleic genomes can be visualized here, as more genomes are added to the portal new datasets will be available. Currently users can visualize ADG1, ADG2, AJH, CHA, CUR, GON1, GON2, JUZ, BUK, PHU, TBR, DM and STN genomes.

Figure 12: Plastome assemblies that are available on the portal.

Available published plastome can be visualized here, as more genomes are added to the portal new datasets will be available. Currently users can visualize ADG1, ADG2, AJH, CHA, CUR, GON1, GON2, JUZ, BUK1, BUK2, PHU, TBR, and STN genomes.

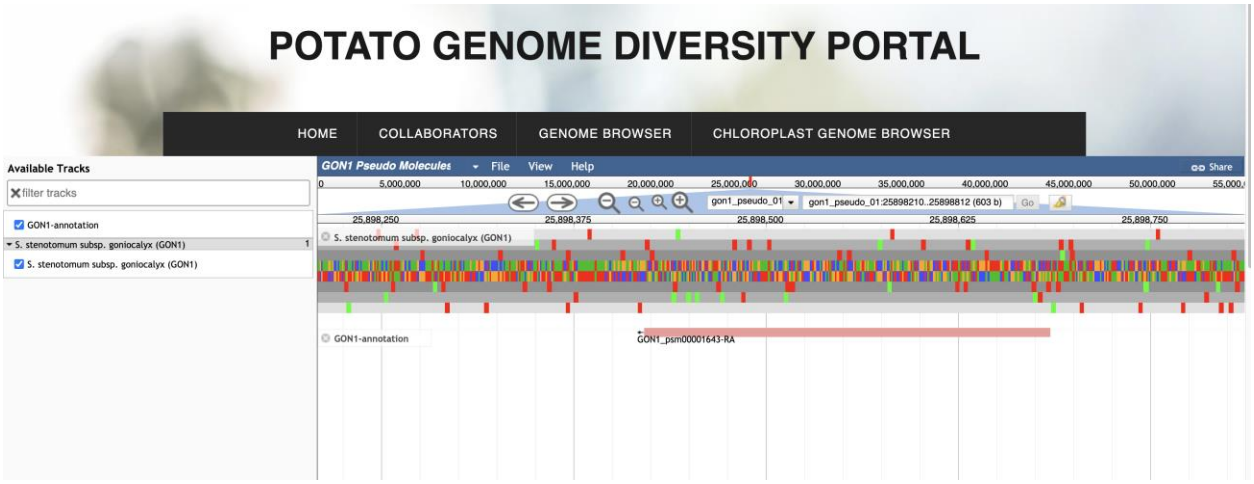

Figure 13: GON1 reference genome and its available tracks visualized in JBrowse genome browser.

Users can visualize the GON1 reference genome and its annotation and navigate through different chromosomes. Also search field is available to search for genes with the corresponding names.

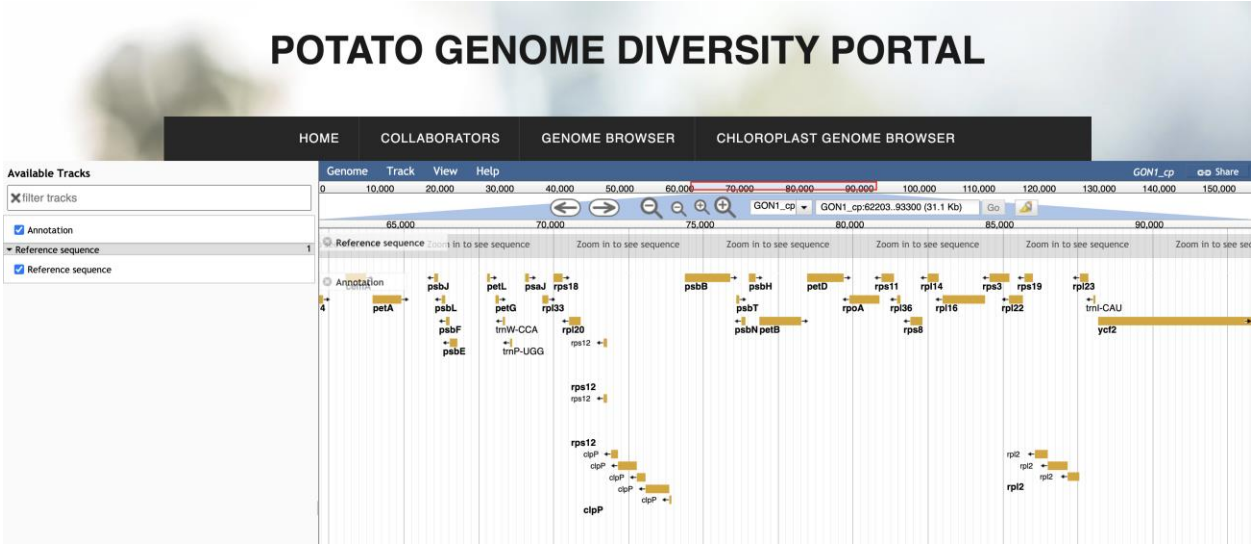

Figure 14: GON1 chloroplast and its available tracks visualized in JBrowse genome browser.

Users can visualize the GON1 chloroplast genome and its annotation. Also search field is available to search for genes with the corresponding names.

5 DISCUSSION

The results in the current study describe the 10X Genomics *de novo* genome sequence assembly of the wild, diploid potato species *Solanum bukasovii* (BUK2), and the structural variation of BUK2 compared with five published potato reference genomes (DM, M6, GON1, SOL, RH-1/RH-2) and another *S. bukasovii* genome (BUK1) from the same genebank accession from CIP. Copy number variation and Single Nucleotide Polymorphism analysis were used to reveal genome specific variation that is not captured in the reference genomes.

5.1 10X Genomics *De Novo* Assembly of *Solanum bukasovii*

The BUK2 assembly was accomplished using 10X Genomics linked-read (artificial long-reads) technology. 10X Genomics linked-read technology allowed the genome assembly to be a less fragmented genome, 617Mbs for BUK2 assembly. The power of leveraging 10X Genomics linked-read technology is demonstrated by the BUSCO analysis results; 96.9% Complete BUSCOs. 67.55% of the BUK2 assembly aligns to 54.69% of the DMv.6.1 genome assembly.

When compared to five published reference genomes, the genome that was best covered (highest number of bases) by BUK2 was the RH-1 reference genome with 555Mbs mapped, although this overlap corresponds to only 68% of the RH-1 reference genome (Figure 5). The size of the assembly must be considered while using this metric. Overall, BUK2 covered up to 77% of the reference genomes M6, GON1 and BUK1, while 72% of DMv6.1 and SOL is covered by BUK2 and 68% of RH-1 is covered with BUK2 (Figure 5). In line with the findings of previous studies with diploid potato species, deletions were found to be more prevalent than duplications (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020). Contrary to previous findings however, the deletions in BUK2 were found to be larger in size when BUK2 was compared to all the reference genomes (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020). The results show that CNVs majorly affect the intergenic regions, and more than 40% of the genes affected by CNVs across all the genomes are genes of unknown function. The most CNV affected cluster contain genes involved in biosynthesis such as zinc-finger proteins.

5.2 SNP Analysis Uncovers Phylogeny of the Wild *Solanum bukasovii* Accession

Whole-genome SNP analysis between BUK2 and the other potato genomes, BUK1, DM, M6, GON1, SOL, RH-1 revealed a higher number of SNPs with domestication: the lowest number of SNPs were found when BUK2 was compared to the genome of another wild potato species, M6, and second lowest was found with the GON1, the landrace genome. The highest levels were found when compared with the genomes of cultivated potato varieties; RH-1 and SOL, with one exception – a higher number of SNPs was found in the BUK2-BUK1 alignment when compared to other alignments (BUK2-GON1, BUK2-DM, BUK2-M6) (Figure 6). The landrace genome, GON1, showed a lower level of heterozygosity compared to the wild BUK2 genome. The lowest number of heterozygosity was found in M6, due to its pseudomolecules contains only 60% of the genome and probably because it is inbred and wild. Further phylogeny analysis with GBS array showed that M6 and BUK2 clustered together, hence lower levels of heterozygosity were found in the comparison (Figure 7). There are more intergenic SNPs than exonic or intronic SNPs, which is in accordance with previous study (Kyriakidou *et al.*, 2020).

Overall, SNP analysis revealed that in all genome comparisons missense to silent mutation ratio was 1.2±0.4, while SNPs that caused missense mutations were also found to be around 50%. These findings are in concordance with previous studies on potato (Pham *et al.*, 2017; Kyriakidou *et al.*, 2020).

The SNPs count ranged from 9.9 million to 3.4 million in SOL and M6, respectively. Considering the BUK2 genome to represent a wild potato species, the highest number of SNPs were found when compared to cultivated varieties such as RH and SOL. The lowest numbers of SNPs were found with other wild and landrace genomes such as GON1 and M6, 3.6 million and 3.4 million, respectively. These numbers were in concordance with the previous work with a 12-genome panel including genomes from potato landraces and wild species (Kyriakidou *et al.*, 2020). Surprisingly, a very high number of SNPs, 6.7 million were found when compared with BUK1, which has the same CIP Genebank accession as BUK2 – CIP 761748 - BioSample: SAMN12730757. Though the exact reason for this is currently unknown, it could be due to a hybridization done to achieve a virus free sample, or alternatively the two individuals come from an accession that was collected as seeds from one locality.

5.3 Disease Resistance Gene Clusters Affected by CNV Events in BUK2 Genome

Four disease resistance genes were affected by significant CNV clusters in BUK2 compared to DM. One disease resistance cluster is located on Chromosome 6 between 47.6 ~ 47.7Mbs. This cluster includes genes such as LRR family proteins, late blight resistance proteins and NB-ARC domain-containing disease resistance proteins. On Chromosome 8, 37.15 ~ 37.16 Mbs and on Chromosome 12, 20.3Mbs several F-box proteins were affected by large CNVs. While these four types of proteins were affected by significant CNV clusters in comparison to all the genomes, they were not significantly enriched in M6. Leucine-rich repeat domain superfamily proteins are CNV affected in the RH-1 genome (Chr 4 14.5 ~ 20.7 Mbs). Late blight genes were affected by significant CNVs in BUK2 genome compared to SOL. Previous studies showed that many disease resistance genes including NBS-LRR and late blight resistance genes were effected by CNVs (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020). It has been shown that these regions with disease resistant genes that are found in clusters, are prone to structural variation resulting in rapid evolution in other plant species as well (Bergelson *et al.*, 2001).

5.4 CNV Affected Gene Clusters are Involved in Metabolite Biosynthesis

The analysis showed that genes that oversee metabolite biosynthesis were affected by significant CNVs clusters when BUK2 was compared to all the genomes. C2H2 & C2HC Zinc finger family proteins and terpene synthase are shown to be affected by CNVs in all the genomes. The terpene synthases synthesize terpene molecules, such as isoprene, monoterpenes and sesquiterpenes (Chen *et al.*, 2011). Zinc finger proteins are also involved in disease resistance pathways (Emerson & Thomas 2009), and in a study with 12 potato genomes it was also previously shown that zinc finger family proteins and terpene synthases are affected by CNVs (Kyriakidou *et al.*, 2020).

5.5 Biotic and Abiotic Stress Responding Genes Affected by CNV Clusters

Several biotic and abiotic genes were found to be affected by large CNVs in all the genomes. Calcineurin-like metallo-phosphoesterase superfamily proteins and Pleiotropic drug resistance were found to be affected in DM and RH-1. Heat shock protein (DNAJ) was found to be affected in DM, RH-1 and SOL. Calcineurin-like metallo-phosphoesterase superfamily proteins hydrolyse phosphoesters in a metal-dependent manner (Matange *et al.*, 2015). Pleiotropic drug resistance genes are involved in transporting antimicrobial secondary metabolites to the cell surface (Crouzet

et al., 2006). Heat shock protein (DNAJ) is a protein responsive to stress, such as, infection, heat, NaCl (Zhichang *et al.*, 2010).

5.6 CNV Affected Plant Development Related Genes

Small Auxin-up RNA (SAUR) genes were found to be affected by CNVs when compared to all genomes except M6. SAUR genes participate in the auxin signaling pathways that are responsive to auxin, they also regulate a wide range of cellular and developmental processes (Ren & Gray 2015). Previous studies on potato also found SAUR genes to be affected by CNVs (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020). Cell wall related genes were found to be CNV affected when compared to BUK2. Galacturonosyltransferase, is found to be CNV affected in BUK2 when compared to DM, M6 and GON1. Pentatricopeptide repeat-containing genes, involved in organelle biogenesis and plant development (Saha *et al.*, 2007), were found to be affected by CNVs in BUK2 compared to DM, GON1, RH and SOL. Galacturonosyltransferases synthase pectic polysaccharide, which is a structural component of the cell wall (Atmodjo *et al.*, 2011).

5.7 CNV-affected Common Clusters in Potato and Plant Genomes

Annotation of the CNV analysis results of BUK2 genome compared to the available reference genomes revealed that many CNV-affected genes are hypothetical or conserved hypothetical proteins. This finding is a common finding in the previous studies with potato (Kyriakidou *et al.*, 2020) and in other plant species such as *Arabidopsis thaliana* (Cao *et al.*, 2011) and rice (Xu *et al.*, 2012), where a significant portion of the genes affected by CNV events also code for hypothetical or unknown proteins.

5.8 Conclusion

S. bukasovii was selected to investigate structural variation against all the published reference genomes. In accordance with previous studies in potato and plants in general, genes coding for disease resistance were identified as affected by variation, e.g., NBS-LRR, SAURs, zinc finger proteins, terpenes and genes of unknown function. However, unlike the previous studies, genes involved in biotic and abiotic stresses, such as metallo-phosphoesterase superfamily proteins, pleiotropic drug resistance and Heat shock protein (DNAJ), genes involved in plant development such as galacturonosyltransferase and pentatricopeptide repeat-containing genes were also identified to be impacted by CNVs.

Potato and its wild relatives are a challenging group of species to study due to its nature of polyploidization, cross hybridization, high heterozygosity, self-compatibility and incompatibility and disease resistance capabilities. Further genetic improvement of the potato can be unlocked with natural variation of CNVs. Identification of the traits of CNV affected genes will provide a great tool for potato breeding.

The recently published new potato reference genomes such as *S. stenotomum subsp. goniocalyx* (GON1), *S. tuberosum*, Solyntus (SOL) and *S. tuberosum* group Tuberosum RH89-039-16 (RH) allowed this research to be conducted with five reference genomes. Prior structural variation and CNV analyses in potato (Hardigan *et al.*, 2016; Kyriakidou *et al.*, 2020) contributed to capturing the diversity in different potato taxa (Kyriakidou *et al.*, 2020). The draft genome of the wild potato species *S. bukasovii* assembled in this study will contribute to the potato genetic diversity available in the literature. This work highlighted the structural variation amongst the published potato reference genomes and the new draft genome assembly of *S. bukasovii* identifying the CNV affected genes in disease resistance and stress tolerance.

Addressing the hypotheses of this thesis

The first hypothesis of the study was that the two *S. bukasovii* genomes have SNPs between each other. Even though the two sequenced individuals are from the same genebank accession at CIP, there are indeed a significant number of SNPs between genotypes of the same accession. In fact, more SNPs than expected were seen, and it has been concluded that the two derive from different

maternal individuals. Discussions have been had with the genebank as to the reasons for this and the most likely explanation is that the accession was collected as a population of individuals which likely consisted of a heterogenous group. Consequently, the number of SNPs between BUK2 and BUK1 was found to be the third highest number of SNPs found when BUK2 was compared to available reference genomes.

The second hypothesis was structural variations, such as Copy Number Variation, would be found when BUK2 was compared to the available reference genomes; number of SNPs found from highest to lowest were, SOL, RH-1, BUK1, DM, GON1, M6 while number of CNVs found from highest to lowest were, Solyntus, RH-1, DM, GON1 and M6. Unfortunately, we were not able to call CNV between the diploid potato pan-genome and BUK2 due to the highly fragmented nature of the diploid potato pan-genome. In order to use the diploid potato pan-genome to call CNVs the pan-genome must be in more contiguous scaffolds or in pseudo molecules.

This thesis is an effort towards the conservation of the potato germplasm in the potato genebank. As climate change diminishes the natural habitat of Solanum species, wild species which are the source of genes for biotic and abiotic stress resistance gets directly affected. Biotic and abiotic stress resistance genes can be used by breeders for improving crops. This work will allow the genebank to have more information on S. bukasovii and will facilitate in prioritizing germplasm for conservation. The identification of the difference between BUK1 and BUK2 has shown the importance of genome sequencing in managing germplasm conservation. Development of the Potato Genome Diversity Portal enables breeders and genebank managers to access and interpret genome sequencing information. This work is an advancement for potato improvement for food security.

5.9 Future Research Directions

The resulting *S. bukasovii de novo* assembly was successful thanks to the 10X Genomics linked-read technology. Although the assembly was good, it can be improved with other long-read technologies. This will introduce a new wild genome assembly that can be used as a reference genome for future research. 10X Genomics reads were not sufficient to call structural variation using them as Illumina short reads and detecting SVs through read-depth analysis. Providing more

short reads will allow researchers to call SVs not only through 10X Genomics Longranger pipeline but with read-depth SV detectors that will improve the SV results.

Pan-genomics allow researchers to compare vast number of genomes with core- and alternative- genomes. Having a good genome assembly of a wild potato species included in the diploid potato pan-genome will be useful for discovery of what unique and alternative genomes *Solanum bukasovii* introduces to the potato gene pool.

The Potato Genome Diversity Portal can be improved by adding a functionality that allows users to follow pipelines through a graphical user interface. The Galaxy platform (Afgan *et al.*, 2018) can be used to provide this functionality. Furthermore, it will be useful if researchers can implement motif finding software like Seeder (Fauteux *et al.*, 2008) as a JBrowse plugin to find motives in the published genomes. Additionally, it would also be interesting to add the diploid potato pan-genome to the genome browser and enable functionalities to update the pan-genome through the browser.

6 REFERENCES

1. Abyzov A., Urban A.E., Snyder M. & Gerstein M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21, 974-84.
2. Achakkagari S.R., Bozan I., Anglin N.L., Ellis D., Tai H.H. & Strömvik M.V. (2021) Complete mitogenome assemblies from a panel of 13 diverse potato taxa. Mitochondrial DNA part B 6, 894-7.
3. Achakkagari S.R., Kyriakidou M., Tai H.H., Anglin N.L., Ellis D. & Strömvik M.V. (2020) Complete plastome assemblies from a panel of 13 diverse potato taxa. PLoS One 15, e0240124.
4. Afgan E., Baker D., Batut B., Van Den Beek M., Bouvier D., Čech M., Chilton J., Clements D., Coraor N. & Grüning B.A. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, W537-W44.
5. Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
6. Atmodjo M.A., Sakuragi Y., Zhu X., Burrell A.J., Mohanty S.S., Atwood J.A., Orlando R., Scheller H.V. & Mohnen D. (2011) Galacturonosyltransferase (GAUT) 1 and GAUT7 are the core of a plant cell wall pectin biosynthetic homogalacturonan: galacturonosyltransferase complex. Proceedings of the National Academy of Sciences 108, 20225-30.
7. Aversano R., Contaldi F., Ercolano M.R., Grosso V., Iorizzo M., Tatino F., Xumerle L., Dal Molin A., Avanzato C., Ferrarini A., Delledonne M., Sanseverino W., Aiese Cigliano R., Capella-Gutierrez S., Gabaldon T., Frusciante L., Bradeen J.M. & Carputo D. (2015a) The Solanum commersonii Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. Plant Cell 27, 954-68.
8. Aversano R., Contaldi F., Ercolano M.R., Grosso V., Iorizzo M., Tatino F., Xumerle L., Dal Molin A., Avanzato C., Ferrarini A., Delledonne M., Sanseverino W., Cigliano R.A., Capella-Gutierrez S., Gabaldon T., Frusciante L., Bradeen J.M. & Carputo D. (2015b) The Solanum commersonii Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. Plant Cell 27, 954-68.
9. Benedict M.N., Henriksen J.R., Metcalf W.W., Whitaker R.J. & Price N.D. (2014) ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15, 8.
10. Bennet M.D. & Leitch I.J. (1997) Nuclear DNA amounts in angiosperms—583 new estimates. Ann Bot 80, 169-96.
11. Bergelson J., Kreitman M., Stahl E.A. & Tian D. (2001) Evolutionary dynamics of plant R-genes. Science 292, 2281-5.
12. Blom J., Albaum S.P., Doppmeier D., Pühler A., Vorhölter F.-J., Zakrzewski M. & Goesmann A. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10, 154.

13. Blom J., Kreis J., Spänig S., Juhre T., Bertelli C., Ernst C. & Goesmann A. (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Research 44, W22-W8.
14. Bolger A.M., Lohse M. & Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-20.
15. Boratyn G.M., Schäffer A.A., Agarwala R., Altschul S.F., Lipman D.J. & Madden T.L. (2012) Domain enhanced lookup time accelerated BLAST. Biology direct 7, 12.
16. Bradeen J.M. & Kole C. (2016) *Genetics, genomics and breeding of potato*. CRC Press.
17. Brittnacher M.J., Fong C., Hayden H., Jacobs M., Radey M. & Rohmer L. (2011) PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics 27, 2429-30.
18. Buels R., Yao E., Diesh C.M., Hayes R.D., Munoz-Torres M., Helt G., Goodstein D.M., Elsik C.G., Lewis S.E., Stein L. & Holmes I.H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 17, 66.
19. Burge C. & Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. Journal of molecular biology 268, 78-94.
20. Cao J., Schneeberger K., Ossowski S., Gunther T., Bender S., Fitz J., Koenig D., Lanz C., Stegle O., Lippert C., Wang X., Ott F., Muller J., Alonso-Blanco C., Borgwardt K., Schmid K.J. & Weigel D. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43, 956-63.
21. Chen F., Tholl D., Bohlmann J. & Pichersky E. (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. The Plant Journal 66, 212-29.
22. Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L., Land S.J., Lu X. & Ruden D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80-92.
23. Cipar M., Peloquin S. & Hougas R. (1964) Variability in the expression of self-incompatibility in tuber-bearing diploid Solanum species. American potato journal 41, 155-62.
24. Contreras-Moreira B. & Vinuesa P. (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl. Environ. Microbiol. 79, 7696-701.
25. Crouzet J., Trombik T., Fraysse Å.S. & Boutry M. (2006) Organization and function of the plant pleiotropic drug resistance ABC transporter family. Febs Letters 580, 1123-30.
26. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T. & Sherry S.T. (2011) The variant call format and VCFtools. Bioinformatics 27, 2156-8.
27. Darracq A., Vitte C., Nicolas S., Duarte J., Pichon J.P., Mary-Huard T., Chevalier C., Berard A., Le Paslier M.C., Rogowsky P., Charcosset A. & Joets J. (2018) Sequence analysis of

European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics 19, 119.
28. Díaz A., Zikhali M., Turner A.S., Isaac P. & Laurie D.A. (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS One 7, e33234.
29. Elyanow R., Wu H.T. & Raphael B.J. (2018) Identifying structural variants using linked-read sequencing data. Bioinformatics 34, 353-60.
30. Emerson R.O. & Thomas J.H. (2009) Adaptive evolution in zinc finger transcription factors. PLoS Genet 5, e1000325.
31. Endelman J.B. & Jansky S.H. (2016) Genetic mapping with an inbred line-derived F2 population in potato. Theoretical and Applied Genetics 129, 935-43.
32. Fang L., Kao C., Gonzalez M.V., Mafra F.A., Pellegrino da Silva R., Li M., Wenzel S.S., Wimmer K., Hakonarson H. & Wang K. (2019) LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. Nat Commun 10, 5585.
33. Fauteux F., Blanchette M. & Strömvik M.V. (2008) Seeder: discriminative seeding DNA motif discovery. Bioinformatics 24, 2303-7.
34. Feuk L., Carson A.R. & Scherer S.W. (2006) Structural variation in the human genome. Nature Reviews Genetics 7, 85-97.
35. Fouts D.E., Brinkac L., Beck E., Inman J. & Sutton G. (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Research 40, e172-e.
36. Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D.M., Burzynski-Chang E.A., Fish T.L., Stromberg K.A., Sacks G.L., Thannhauser T.W., Foolad M.R., Diez M.J., Blanca J., Canizares J., Xu Y., van der Knaap E., Huang S., Klee H.J., Giovannoni J.J. & Fei Z. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet.
37. Garrison E. & Marth G. (2012) FreeBayes. arXiv preprint1207. 3907 [q-bio. GN][Internet].
38. Gebhardt C., Ballvora A., Walkemeier B., Oberhagemann P. & Schüler K. (2004) Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. Molecular Breeding 13, 93-102.
39. Github I. (2016) GitHub.
40. Gnerre S., MacCallum I., Przybylski D., Ribeiro F.J., Burton J.N., Walker B.J., Sharpe T., Hall G., Shea T.P. & Sykes S. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences 108, 1513-8.
41. Golicz A.A., Bayer P.E., Barker G.C., Edger P.P., Kim H., Martinez P.A., Chan C.K., Severn-Ellis A., McCombie W.R., Parkin I.A., Paterson A.H., Pires J.C., Sharpe A.G., Tang H., Teakle G.R., Town C.D., Batley J. & Edwards D. (2016) The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun 7, 13390.

42. Gu W., Zhang F. & Lupski J.R. (2008) Mechanisms for human genomic rearrangements. Pathogenetics 1, 4.
43. Gurevich A., Saveliev V., Vyahhi N. & Tesler G. (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072-5.
44. Hamilton J.P. & Robin Buell C. (2012) Advances in plant genome sequencing. The Plant Journal 70, 177-90.
45. Hardigan M.A., Bamberg J., Buell C.R. & Douches D.S. (2015) Taxonomy and Genetic Differentiation among Wild and Cultivated Germplasm of sect. The Plant Genome 8.
46. Hardigan M.A., Crisovan E., Hamilton J.P., Kim J., Laimbeer P., Leisner C.P., Manrique-Carpintero N.C., Newton L., Pham G.M., Vaillancourt B., Yang X., Zeng Z., Douches D.S., Jiang J., Veilleux R.E. & Buell C.R. (2016) Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. Plant Cell 28, 388-405.
47. Hawkes J.G. (1990) *The potato: evolution, biodiversity and genetic resources*. Belhaven Press.
48. Hirsch C.N., Foerster J.M., Johnson J.M., Sekhon R.S., Muttoni G., Vaillancourt B., Penagaricano F., Lindquist E., Pedraza M.A., Barry K., de Leon N., Kaeppler S.M. & Buell C.R. (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121-35.
49. Ho Y.J., Anaparthy N., Molik D., Mathew G., Aicher T., Patel A., Hicks J. & Hammell M.G. (2018) Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. Genome Research 28, 1353-63.
50. Hosaka K. (1995) Successive domestication and evolution of the Andean potatoes as revealed by chloroplast DNA restriction endonuclease analysis. Theoretical and Applied Genetics 90, 356-63.
51. Hosmani P.S., Flores-Gonzalez M., van de Geest H., Maumus F., Bakker L.V., Schijlen E., van Haarst J., Cordewener J., Sanchez-Perez G. & Peters S. (2019) An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv, 767764.
52. Hulse-Kemp A.M., Maheshwari S., Stoffel K., Hill T.A., Jaffe D., Williams S.R., Weisenfeld N., Ramakrishnan S., Kumar V. & Shah P. (2018) Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Horticulture research 5, 1-13.
53. Hurgobin B., Golicz A.A., Bayer P.E., Chan C.K., Tirnaz S., Dolatabadian A., Schiessl S.V., Samans B., Montenegro J.D., Parkin I.A.P., Pires J.C., Chalhoub B., King G.J., Snowdon R., Batley J. & Edwards D. (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol J 16, 1265-74.
54. Institute B. (2016) Picard tools. Broad Institute, GitHub repository.

55. International Wheat Genome Sequencing C. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.
56. Jackman S.D., Coombe L., Chu J., Warren R.L., Vandervalk B.P., Yeo S., Xue Z., Mohamadi H., Bohlmann J. & Jones S.J. (2018) Tigmint: correcting assembly errors using linked reads from large molecules. BMC Bioinformatics 19, 1-10.
57. Jackman S.D., Mozgacheva T., Chen S., O'Huiginn B., Bailey L., Birol I. & Jones S.J. (2019) ORCA: a comprehensive bioinformatics container environment for education and research. Bioinformatics 35, 4448-50.
58. Karaoğlanoğlu F., Ricketts C., Ebren E., Rasekh M.E., Hajirasouliha I. & Alkan C. (2020) VALOR2: characterization of large-scale structural variants using linked-reads. Genome Biology 21, 1-11.
59. Karolchik D., Hinrichs A.S. & Kent W.J. (2009) The UCSC Genome Browser. Curr Protoc Bioinformatics Chapter 1, Unit1 4.
60. Kawahara Y., De La Bastide M., Hamilton J.P., Kanamori H., McCombie W.R., Ouyang S., Schwartz D.C., Tanaka T., Wu J., Zhou S., Childs K.L., Davidson R.M., Lin H., Quesada-Ocampo L., Vaillancourt B., Sakai H., Lee S.S., Kim J., Numa H., Itoh T., Buell C.R. & Matsumoto T. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6, 4.
61. Keilwagen J., Wenk M., Erickson J.L., Schattat M.H., Grau J. & Hartung F. (2016) Using intron position conservation for homology-based gene prediction. Nucleic Acids Research 44, e89-e.
62. 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, 722-36.
63. Kumar M., Choi J.Y., Kumari N., Pareek A. & Kim S.R. (2015) Molecular breeding in Brassica for salt tolerance: importance of microsatellite (SSR) markers for molecular breeding in Brassica. Frontiers in plant science 6, 688.
64. Kyriakidou M., Achakkagari S.R., López J.H.G., Zhu X., Tang C.Y., Tai H.H., Anglin N.L., Ellis D. & Strömvik M.V. (2020) Structural genome analysis in cultivated potato taxa. Theoretical and Applied Genetics 133, 951-66.
65. Kyriakidou M., Tai H.H., Anglin N.L., Ellis D. & Stromvik M.V. (2018) Current Strategies of Polyploid Plant Genome Sequence Assembly. Front Plant Sci 9, 1660.
66. Kyriakidou, M. (2020). Genome assembly and discovery of structural variation in cultivated potato taxa. [Montreal], McGill University Libraries.
67. Lai J., Li R., Xu X., Jin W., Xu M., Zhao H., Xiang Z., Song W., Ying K., Zhang M., Jiao Y., Ni P., Zhang J., Li D., Guo X., Ye K., Jian M., Wang B., Zheng H., Liang H., Zhang X., Wang S., Chen S., Li J., Fu Y., Springer N.M., Yang H., Wang J., Dai J., Schnable P.S. & Wang J. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42, 1027-30.

68. Laing C., Buchanan C., Taboada E.N., Zhang Y., Kropinski A., Villegas A., Thomas J.E. & Gannon V.P. (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 11, 461.
69. Lee J.A., Carvalho C.M. & Lupski J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. cell 131, 1235-47.
70. Lefebure T. & Stanhope M.J. (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol 8, R71.
71. Leisner C.P., Hamilton J.P., Crisovan E., Manrique-Carpintero N.C., Marand A.P., Newton L., Pham G.M., Jiang J., Douches D.S., Jansky S.H. & Buell C.R. (2018) Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. Plant J 94, 562-70.
72. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
73. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. & Durbin R. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-9.
74. Li Y.H., Zhou G., Ma J., Jiang W., Jin L.G., Zhang Z., Guo Y., Zhang J., Sui Y., Zheng L., Zhang S.S., Zuo Q., Shi X.H., Li Y.F., Zhang W.K., Hu Y., Kong G., Hong H.L., Tan B., Song J., Liu Z.X., Wang Y., Ruan H., Yeung C.K., Liu J., Wang H., Zhang L.J., Guan R.X., Wang K.J., Li W.B., Chen S.Y., Chang R.Z., Jiang Z., Jackson S.A., Li R. & Qiu L.J. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32, 1045-52.
75. Lischer H.E.L. & Shimizu K.K. (2017) Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18, 474.
76. Love S.L. (1999) Founding clones, major contributing ancestors, and exotic progenitors of prominent North American potato cultivars. American journal of potato research 76, 263-72.
77. Lukjancenko O., Thomsen M.C., Larsen M.V. & Ussery D.W. (2013) PanFunPro: PAN-genome analysis based on FUNctional PROfiles. F1000Research 2.
78. Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q. & Liu Y. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 2047-217X-1-18.
79. Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.-Y. & Gish W.R. (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23, 452-6.
80. Matange N., Podobnik M. & Visweswariah S.S. (2015) Metallophosphoesterases: structural fidelity with functional promiscuity. Biochemical Journal 467, 201-16.

81. McHale L.K., Haun W.J., Xu W.W., Bhaskar P.B., Anderson J.E., Hyten D.L., Gerhardt D.J., Jeddeloh J.A. & Stupar R.M. (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant physiology 159, 1295-308.
82. Medini D., Donati C., Tettelin H., Masignani V. & Rappuoli R. (2005) The microbial pan-genome. Curr Opin Genet Dev 15, 589-94.
83. Meleshko D., Marks P., Williams S. & Hajirasouliha I. (2019) Detection and assembly of novel sequence insertions using Linked-Read technology. bioRxiv, 551028.
84. Merkel D. (2014) Docker: lightweight linux containers for consistent development and deployment. Linux journal 2014, 2.
85. Micheletto S., Boland R. & Huarte M. (2000) Argentinian wild diploid Solanum species as sources of quantitative late blight resistance. Theoretical and Applied Genetics 101, 902-6.
86. Montenegro J.D., Golicz A.A., Bayer P.E., Hurgobin B., Lee H., Chan C.K., Visendi P., Lai K., Dolezel J., Batley J. & Edwards D. (2017) The pangenome of hexaploid bread wheat. Plant J 90, 1007-13.
87. Morgante M., De Paoli E. & Radovic S. (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10, 149-55.
88. Nagaharu U. & Nagaharu N. (1935) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization.
89. Ortiz E. (2019) vcf2phylip v2. 0: convert a VCF matrix into several matrix formats for phylogenetic analysis. URL https://doi org/105281/zenodo 2540861.
90. Ou L., Li D., Lv J., Chen W., Zhang Z., Li X., Yang B., Zhou S., Yang S., Li W., Gao H., Zeng Q., Yu H., Ouyang B., Li F., Liu F., Zheng J., Liu Y., Wang J., Wang B., Dai X., Ma Y. & Zou X. (2018) Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. New Phytol 220, 360-3.
91. Ovchinnikova A., Krylova E., Gavrilenko T., Smekalova T., Zhuk M., Knapp S. & Spooner D.M. (2011) Taxonomy of cultivated potatoes (Solanum section Petota: Solanaceae). Botanical Journal of the Linnean Society 165, 107-55.
92. Paul S., Bhardwaj A., Bag S.K., Sokurenko E.V. & Chattopadhyay S. (2015) PanCoreGen—Profiling, detecting, annotating protein-coding genes in microbial genomes. Genomics 106, 367-72.
93. Pham G.M., Hamilton J.P., Wood J.C., Burke J.T., Zhao H., Vaillancourt B., Ou S., Jiang J. & Buell C.R. (2020) Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience 9.
94. Pham G.M., Newton L., Wiegert-Rininger K., Vaillancourt B., Douches D.S. & Buell C.R. (2017) Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. The Plant Journal 92, 624-37.
95. Potato Genome Sequencing C., Xu X., Pan S., Cheng S., Zhang B., Mu D., Ni P., Zhang G., Yang S., Li R., Wang J., Orjeda G., Guzman F., Torres M., Lozano R., Ponce O., Martinez D., De la Cruz G., Chakrabarti S.K., Patil V.U., Skryabin K.G., Kuznetsov B.B., Ravin N.V., Kolganova T.V., Beletsky A.V., Mardanov A.V., Di Genova A., Bolser D.M., Martin D.M.,

Li G., Yang Y., Kuang H., Hu Q., Xiong X., Bishop G.J., Sagredo B., Mejia N., Zagorski W., Gromadka R., Gawor J., Szczesny P., Huang S., Zhang Z., Liang C., He J., Li Y., He Y., Xu J., Zhang Y., Xie B., Du Y., Qu D., Bonierbale M., Ghislain M., Herrera Mdel R., Giuliano G., Pietrella M., Perrotta G., Facella P., O'Brien K., Feingold S.E., Barreiro L.E., Massa G.A., Diambra L., Whitty B.R., Vaillancourt B., Lin H., Massa A.N., Geoffroy M., Lundback S., DellaPenna D., Buell C.R., Sharma S.K., Marshall D.F., Waugh R., Bryan G.J., Destefanis M., Nagy I., Milbourne D., Thomson S.J., Fiers M., Jacobs J.M., Nielsen K.L., Sonderkaer M., Iovene M., Torres G.A., Jiang J., Veilleux R.E., Bachem C.W., de Boer J., Borm T., Kloosterman B., van Eck H., Datema E., Hekkert B., Goverse A., van Ham R.C. & Visser R.G. (2011) Genome sequence and analysis of the tuber crop potato. Nature 475, 189-95.

96. Pracana R., Priyam A., Levantis I., Nichols R.A. & Wurm Y. (2017) The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. Molecular ecology 26, 2864-79.
97. Qin C., Yu C., Shen Y., Fang X., Chen L., Min J., Cheng J., Zhao S., Xu M. & Luo Y. (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proceedings of the National Academy of Sciences 111, 5135-40.
98. Quinlan A.R. & Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-2.
99. Ren H. & Gray W.M. (2015) SAUR proteins as effectors of hormonal and environmental signals in plant growth. Mol Plant 8, 1153-64.
100. Robinson O., Dylus D. & Dessimoz C. (2016) Phylo. io: interactive viewing and comparison of large phylogenetic trees on the web. Molecular Biology and Evolution 33, 2163-6.
101. Saha D., Prasad A. & Srinivasan R. (2007) Pentatricopeptide repeat proteins and their emerging roles in plants. Plant Physiology and Biochemistry 45, 521-34.
102. Sanford L., Kobayashi R., Deahl K. & Sinden S. (1996) Segregation of leptines and other glycoalkaloids inSolanum tuberosum (4x)× S. chacoense (4x) crosses. American potato journal 73, 21.
103. Schatz M.C., Maron L.G., Stein J.C., Wences A.H., Gurtowski J., Biggers E., Lee H., Kramer M., Antoniou E. & Ghiban E. (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome Biology 15, 506.
104. Sefraoui O., Aissaoui M. & Eleuldj M. (2012) OpenStack: toward an open-source solution for cloud computing. International Journal of Computer Applications 55, 38-42.
105. Seppey M., Manni M. & Zdobnov E.M. (2019) BUSCO: assessing genome assembly and annotation completeness. In: Gene Prediction (pp. 227-45. Springer.
106. Sharma S.K., Bolser D., de Boer J., Sonderkaer M., Amoros W., Carboni M.F., D'Ambrosio J.M., de la Cruz G., Di Genova A., Douches D.S., Eguiluz M., Guo X., Guzman F., Hackett C.A., Hamilton J.P., Li G., Li Y., Lozano R., Maass A., Marshall D., Martinez D., McLean

K., Mejia N., Milne L., Munive S., Nagy I., Ponce O., Ramirez M., Simon R., Thomson S.J., Torres Y., Waugh R., Zhang Z., Huang S., Visser R.G., Bachem C.W., Sagredo B., Feingold S.E., Orjeda G., Veilleux R.E., Bonierbale M., Jacobs J.M., Milbourne D., Martin D.M. & Bryan G.J. (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 (Bethesda) 3, 2031-47.

107. Simko I., Haynes K.G. & Jones R.W. (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173, 2237-45.
108. Skinner M.E., Uzilov A.V., Stein L.D., Mungall C.J. & Holmes I.H. (2009) JBrowse: a next-generation genome browser. Genome Research 19, 1630-8.
109. Smit A., Hubley R. & Green P. (2015) RepeatMasker Open-4.0. 2013–2015.
110. Snipen L., Almoy T. & Ussery D.W. (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10, 385.
111. Spies N., Weng Z.M., Bishara A., McDaniel J., Catoe D., Zook J.M., Salit M., West R.B., Batzoglou S. & Sidow A. (2017) Genome-wide reconstruction of complex structural variants using read clouds. Nature Methods 14, 915-+.
112. Spooner D.M., McLean K., Ramsay G., Waugh R. & Bryan G.J. (2005) A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc Natl Acad Sci U S A 102, 14694-9.
113. Springer N.M., Ying K., Fu Y., Ji T., Yeh C.-T., Jia Y., Wu W., Richmond T., Kitzman J. & Rosenbaum H. (2009a) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5, e1000734.
114. Springer N.M., Ying K., Fu Y., Ji T., Yeh C.T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A.L., Barbazuk W.B., Jeddeloh J.A., Nettleton D. & Schnable P.S. (2009b) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5, e1000734.
115. Stamatakis A., Hoover P. & Rougemont J. (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic biology 57, 758-71.
116. Stanke M., Steinkamp R., Waack S. & Morgenstern B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32, W309-W12.
117. Sun C., Hu Z., Zheng T., Lu K., Zhao Y., Wang W., Shi J., Wang C., Lu J., Zhang D., Li Z. & Wei C. (2017) RPAN: rice pan-genome browser for approximately 3000 rice genomes. Nucleic Acids Res 45, 597-605.
118. Sysoev I. (2004) Nginx. Inc.,“nginx,”
119. Team R.C. (2013) R: A language and environment for statistical computing. Vienna, Austria.
120. Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., Deboy R.T., Davidsen T.M., Mora M., Scarselli M., Margarit y Ros I., Peterson J.D., Hauser C.R., Sundaram J.P., Nelson W.C., Madupu R., Brinkac L.M., Dodson R.J., Rosovitz M.J., Sullivan S.A., Daugherty S.C., Haft D.H., Selengut J., Gwinn M.L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K.J., Smith S., Utterback T.R., White O., Rubens C.E., Grandi G., Madoff

L.C., Kasper D.L., Telford J.L., Wessels M.R., Rappuoli R. & Fraser C.M. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102, 13950-5.

121. Thapar A., Martin J., Mick E., Vásquez A.A., Langley K., Scherer S.W., Schachar R., Crosbie J., Williams N. & Franke B. (2016) Psychiatric gene discoveries shape evidence on ADHD's biology. Molecular psychiatry 21, 1202-7.
122. Ugent D. (1970) The potato. Science 170, 1161-6.
123. van Lieshout N., van der Burgt A., de Vries M.E., Ter Maat M., Eickholt D., Esselink D., van Kaauwen M.P.W., Kodde L.P., Visser R.G.F., Lindhout P. & Finkers R. (2020) Solyntus, the New Highly Contiguous Reference Genome for Potato (Solanum tuberosum). G3 (Bethesda) 10, 3489-95.
124. Walker B.J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C.A., Zeng Q., Wortman J. & Young S.K. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963.
125. Wang W., Yan H.J., Chen S.Y., Li Z.Z., Yi J., Niu L.L., Deng J.P., Chen W.G., Pu Y., Jia X.B., Qu Y., Chen A., Zhong Y., Yu X.M., Pang S., Huang W.L., Han Y., Liu G.J. & Yu J.Q. (2019) The sequence and de novo assembly of hog deer genome. Scientific Data 6.
126. Watanabe K. (2015) Potato genetics, genomics, and applications. Breed Sci 65, 53-68.
127. Watanabe K. & Peloquin S. (1989) Occurrence of 2n pollen and ps gene frequencies in cultivated groups and their related wild species in tuber-bearing Solanums. Theoretical and Applied Genetics 78, 329-36.
128. Weischenfeldt J., Symmons O., Spitz F. & Korbel J.O. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics 14, 125-38.
129. Weisenfeld N.I., Kumar V., Shah P., Church D.M. & Jaffe D.B. (2017) Direct determination of diploid genome sequences. Genome Research 27, 757-67.
130. Xia L.C., Bell J.M., Wood-Bouwens C., Chen J.M.J., Zhang N.R. & Ji H.L.P. (2018) Identification of large rearrangements in cancer genomes with barcode linked reads. Nucleic Acids Research 46.
131. Xu X., Liu X., Ge S., Jensen J.D., Hu F., Li X., Dong Y., Gutenkunst R.N., Fang L. & Huang L. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology 30, 105.
132. Yao W. (2015) An Introduction to intansv.
133. Yu J., Golicz A.A., Lu K., Dossa K., Zhang Y., Chen J., Wang L., You J., Fan D., Edwards D. & Zhang X. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol J 17, 881-92.
134. Zerbino D.R. & Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-9.

135. Zhang F., Khajavi M., Connolly A.M., Towne C.F., Batish S.D. & Lupski J.R. (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nature Genetics 41, 849-53.
136. Zhao H., Sun Z., Wang J., Huang H., Kocher J.-P. & Wang L. (2014a) CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics 30, 1006-7.
137. Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y., Fan D., Zhao Y., Wang Z., Zhou C., Chen J., Zhu C., Li W., Weng Q., Xu Q., Wang Z.X., Wei X., Han B. & Huang X. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50, 278-84.
138. Zhao Y., Jia X., Yang J., Ling Y., Zhang Z., Yu J., Wu J. & Xiao J. (2014b) PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics 30, 1297-9.
139. Zhao Y., Wu J., Yang J., Sun S., Xiao J. & Yu J. (2011) PGAP: pan-genomes analysis pipeline. Bioinformatics 28, 416-8.
140. Zheng G.X., Lau B.T., Schnall-Levin M., Jarosz M., Bell J.M., Hindson C.M., Kyriazopoulou-Panagiotopoulou S., Masquelier D.A., Merrill L. & Terry J.M. (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology 34, 303-11.
141. Zhichang Z., Wanrong Z., Jinping Y., Jianjun Z., Xufeng L.Z.L. & Yang Y. (2010) Over-expression of Arabidopsis DnaJ (Hsp40) contributes to NaCl-stress tolerance. African Journal of Biotechnology 9, 972-8.
142. Zhou Q., Tang D., Huang W., Yang Z., Zhang Y., Hamilton J.P., Visser R.G.F., Bachem C.W.B., Robin Buell C., Zhang Z., Zhang C. & Huang S. (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Genet 52, 1018-23.
143. Zmienko A., Samelak A., Kozlowski P. & Figlerowicz M. (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127, 1-18.

7 APPENDICES

Appendix 1: Installing JBrowse

Downloading JBrowse

JBrowse was downloaded and setup according to documentations on the website (http://jbrowse.org/) (Skinner *et al.*, 2009). As mentioned, hard drive allocation for this VM is under /mnt folder. JBrowse was downloaded in the www/ folder. git clone https://github.com/GMOD/jbrowse is the command to download the JBrowse repository from GitHub. This command will fetch all the repository and create a new folder called jbrowse/. (Taken on 1 March 2020)

Downloading Dependencies

During software development various libraries and software are used. In order to run the software VMs must be equipped with these libraries and software to run the program. Some of the dependencies are listed on the JBrowse website but not all of them. Listed libraries and software can be downloaded with the instructions found on the website. Some of the commands for downloading the dependencies are $sudo\ apt\ install\ build-essential\ zlib1g-dev$ and downloading Node.js from https://nodejs.org/en/download/. For the rest of the dependencies the list of the libraries is taken from the following Docker Containers' Dockerfiles (Merkel, 2014). https://github.com/GMOD/jbrowse-docker and https://hub.docker.com/r/biocontainers/biocontainers/dockerfile these Dockerfiles include all the dependencies for general bioinformatics software.

Setting up JBrowse

Once all the dependencies are loaded to the VM, JBrowse can be setup. It is very easy to set it up after navigating to the jbrowse/ folder. Running the shell script ./setup.sh does all the configuration. If the webserver is showing a message "Congratulations, JBrowse is on the web" that means the configuration has been successful.

Appendix 2: JBrowse file system

Understanding Jbrowse Configuration File System

JBrowse has two separate file systems .conf files are human readable and easy to edit, .json files on the other hand are not human readable and must be filled using the available executables.

Jbrowse.conf

The main file in the JBrowse system is jbrowse.conf. First thing in the configuration file is to add the .conf and .json configuration files in the folders of the genomes that will be displayed.

$$include \ = \{dataRoot\}/trackList.json$$

$$include \mathrel{+}= \{dataRoot\}/tracks.conf$$

These lines specifies that include the /trackList.json /tracks.conf whenever you are in a dataRoot folder. All the track information in these two files will be read with these lines.

trackList.json

This file appears in the data/ folder under each directory for different genomes. This file is created with the processing the reference genome FASTA file. This process will be described in the Adding a Reference Genome Track section. This file is not updated with hand as it is in JSON language, adding new tracks with executable files available such as; flatfile-to-json.pl and prepare-refseqs.pl. trackList.json file has information about location of the file that will be displayed as a track and the type of this file.

tracks.conf

This file appears in the data/ folder under each directory for different genomes. This file is created with processing the reference genome FASTA file. Tracks.conf includes information about the type of the data and the location of the files that will be displayed as tracks. Although this file is human readable, and it is easy to edit. If a track will be added without using executables information must be added to tracks.conf.

Adding a Reference Genome Track

Reference genome track is usually the first track that is displayed from a genome in the genome browser. Create a new directory with the name of the genome. Move genome to the directory and index the FASTA file. $samtools\ faidx\ genome.fa$. This process will produce a file called genome.fa.fai that includes information about the chromosome locations and length. The script prepare-refseqs.pl uses this file while indexing the FASTA file to create a track.

$ubuntu@potato-diversity-portal:/mnt/www/jbrowse/potato-DM\$\ ../bin/prepare-refseqs.pl\ --indexed\_fasta\ genome.fa$ will create a folder called data/ and it will place the necessary information in trackList.json. Since reference genome track is created using executable trackList.json configuration file is updated automatically.

Once this process is successful the reference genome can be viewed on the genome browser.

Adding an Annotation File Track (GFF file)

The most important point to pay attention to while working with additional files after loading the reference genome is to make sure the names of the chromosomes are the same in all file formats. This is a big issue with any bioinformatics software, and it causes combability problems in the process. Usually there are different versions to the files published (.fasta, .gff, .vcf etc.) and in each version the name of the chromosomes might be updated. In one of the cases, .gff file chromosome headers had to be changed, which was accomplished using vim's :%s/*search*/*replace*/g function. If the headers are matching, the .gff file can be uploaded using the following command: $bin/flatfile-to-json.pl\ --gff\ path/to/my.gff3\ --trackType\ CanvasFeatures\ --trackLabel\ mygff.$

Adding an Alignment File Track (BAM file)

Adding a .bam file track is different than adding the other file formats as it is not going to be uploaded using an executable. The BAM file has to be sorted and indexed in order to be displayed in the genome browser. $samtools\ sort\ genome-alignment.bam\ -o\ genome-alignment-sorted.bam$ this command will sort the alignment and create the genome-alignment-sorted.bam file. $samtools\ index\ genome-alignment-sorted.bam$ will create

genome-alignment-sorted.bam.bai which is the indexed file format for .bam files. Adding the following lines to the tracks.conf file will add the alignment track to the genome browser.

$$[tracks.GON1 - DM - Alignment]$$
$$urlTemplate = gon1 - dm - sorted.bam$$
$$storeClass = JBrowse/Store/SeqFeature/BAM$$
$$type = Alignments2$$

Appendix 4: PGDP instance server specifications

Ubuntu-18.04.3-Bionic-x64-2020-01

RAM 180GB

VCPUs 16 VCPU

Disk 20GB

Ephemeral Disk 392GB

Appendix 5: NGINX default.conf configuration

default.conf

```
location / {
root /mnt/www/jbrowse;
index  index.html index.htm;
# First attempt to serve request as file, then
# as directory, then fall back to displaying a 404.
try_files $uri $uri/ = 404;
}
```

Appendix 6: File system of the PGDP after the configurations

```
├── bin/
│   ├── prepare-refseqs.pl
│   ├── flatfile-to-json.pl
├── browser.html
├── jbrowse.conf
├ potato-DM/
├── data/
│   ├── gon1-dm-sorted.bam
│   ├── gon1-dm-sorted.bam.bai
│   ├── seq/
│   ├── trackList.json
│   ├── tracks/
│   └── tracks.conf
├── PGSC_DM_V403_genes.gff
├── potato_dm_v404_all_pm_un.fasta
└── potato_dm_v404_all_pm_un.fasta.fai
```

Appendix 7: Specific instructions to install Docker

[Downloading Docker]

$$sudo\ apt-get\ update$$

$$sudo\ apt-get\ install\ \backslash$$

$$apt-transport-https\ \backslash$$

$$ca-certificates\ \backslash$$

$$curl\ \backslash$$

$$gnupg-agent\ \backslash$$

$$software-properties-common$$

$$curl\ -fsSL\ https://download.docker.com/linux/ubuntu/gpg\ |\ sudo\ apt-key\ add\ -$$

$$sudo\ apt-key\ fingerprint\ 0EBFCD88$$

$$sudo\ add-apt-repository\ \backslash$$

$$"deb\ [arch=amd64]\ https://download.docker.com/linux/ubuntu\ \backslash$$

$$\$(lsb\_release\ -cs)\ \backslash$$

$$stable"$$

$$sudo\ apt-get\ install\ docker-ce\ docker-ce-cli\ containerd.io$$

Appendix 8: Docker command to launch an ORCA instance

$$docker\ pull\ bcgsc/orca$$

$$docker\ run\ -it\ --name\ buk2\ --mount\ type=bind,source=/mnt/www/data-transfer/buk2,target=/buk2/\ bcgsc/orca$$