Genome assembly and discovery of structural variation in cultivated potato taxa

Maria Kyriakidou

Department of Plant Science

Faculty of Agricultural and Environmental Sciences

McGill University

Montreal, Canada

April, 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Contents

List of Figures

List of Tables

Abstract

The common potato (*Solanum tuberosum* L.) is an important staple crop, with a highly complex, heterozygous, tetraploid genome. It can grow in a wide range of altitudes from sea level up to 4,700 meters above the sea level, contributing to its success as a crop. It has its origins in South America, where potato has a large secondary gene pool consisting of wild relatives of diverse ploidy levels. Genetic resources such as landraces and wild relatives are increasingly crucial for developing climate change resilient cultivars with biotic and abiotic stress tolerance. Significant efforts have previously been made to sequence and construct a double monoploid (*S. tuberosum* Group Phureja – DM1-3) reference genome as well as two wild reference genomes (*S. commersonii* and *S. chacoense* clone M6). However, it is uncertain how well the potato genome diversity is actually captured in these three potato genomes, as the genetic riches of the South America taxa are not represented. This doctoral dissertation focuses on the genomic analyses of sequenced data from twelve native South American potato genomes (ten taxa) of various ploidy (2n – 5x): *S. tuberosum* subsp. *goniocalyx* (2n), *S. stenotomum* subsp. *stenotomum* (2n), *S. phureja* (2n), *S. xajanhuiri* (2n), *S. bukasovii* (2n), *S. chaucha* (3x), *S. juzepczukii* (3x), *S. tuberosum* subsp. *andigena* (4x), *S. tuberosum* subsp. *tuberosum* (4x) and *S. curtilobum* (5x). Their comparisons with two reference genomes (DM1-3, M6) unraveled a great number of copy number variation (CNV) impacted genes, including disease resistance and abiotic stress genes. Additionally, these genomes have been assembled *de novo*. The draft genomes of the diploid *S. stenotomum* subsp. *goniocalyx* and of the tetraploid *S. tuberoum* subsp. *andigena* have been assembled using Third Generation Sequencing data, while the rest of the genomes were assembled using Next Generation Sequencing data. The diploid potato genomes have been used for the construction of a diploid potato pan-genome sequence of nine genomes, including three publicly available reference genomes. Within the pan-genome, there are self-incompatibility and disease resistant genes that are absent from the DM1-3 genome. This work reflects only a part of the tremendous variability of the South American potato taxa.

Abrégé

La pomme de terre (*Solanum tuberosum* L.) est une grande culture très importante, dotée d'un génome tétraploïde extrêmement complexe et hétérozygote. Il peut pousser dans une large gamme d'altitudes allant du niveau de la mer jusqu'à 4 700 mètres d'altitude, contribuant ainsi à son succès en tant que culture. Elle a ses origines en Amérique du Sud, où la pomme de terre possède un important pool de gènes secondaires constitué de parents sauvages de divers niveaux de ploïdie. Les ressources génétiques telles que les races locales et les espèces sauvages apparentées sont de plus en plus indispensables au développement de cultivars résistants au changement climatique et tolérants au stress biotique et abiotique. Des efforts importants ont déjà été déployés pour séquencer et construire un génome de référence double monoploïde (*S. tuberosum* Group Phureja - DM1-3) ainsi que deux génomes de référence sauvages (clone M6 de *S. commersonii* et de *S. chacoense*). Cependant, il est difficile de savoir dans quelle mesure la diversité du génome de la pomme de terre est réellement capturée dans ces trois génomes de pomme de terre, car la richesse génétique des taxons d'Amérique du Sud n'est pas représentée. Cette thèse de doctorat porte sur les analyses génomiques de données séquencées de douze génomes de pomme de terre indigènes d'Amérique du Sud (dix taxons) de diverses ploïdies (2n – 5x): *S. tuberosum* subsp. *goniocalyx* (2n), *S. stenotomum* subsp. *stenotomum* (2n), *S. phureja* (2n), *S. xajanhuiri* (2n), *S. bukasovii* (2n), *S. chaucha* (3x), *S. juzepczukii* (3x), *S. tuberosum* subsp. andigena (4x), *S. tuberosum* subsp. *tuberosum* (4x) et *S. curtilobum* (5x). Leurs comparaisons avec deux génomes de référence (DM1-3, M6) ont révélé un grand nombre de gènes impactés par la variation du nombre de copies (VCN), y compris des gènes de résistance aux maladies et de stress abiotique. De plus, ces génomes ont été assemblés *de novo*. Les génomes du diploïde *S. stenotomum* subsp. *goniocalyx* et du tétraploïde *S. tuberoum* subsp. *andigena* ont été assemblés à l'aide de données de séquençage de troisième génération, tandis que les autres génomes ont été assemblés à l'aide de données de séquençage de nouvelle génération. Les génomes diploïdes de pomme de terre ont été utilisés pour la construction d'une séquence pan-génomique diploïde de pomme de terre de neuf

génomes, dont trois génomes de référence accessibles au public. Dans le pan-génome, il existe des gènes d'auto-incompatibilité et de résistance aux maladies qui sont absents du génome de DM1-3. Ces travaux ne reflètent qu'une partie de la très grande variabilité des taxons de pomme de terre sud-américains.

Acknowledgements

My deep and genuine gratitude goes first to my supervisor, Dr. Martina Strömvik, for her faith and confidence in me. It has been an honor and a privilege for me being a PhD student in her lab. Her motivation, patience and her vast knowledge have guided me through these years during my research and writing my thesis. I greatly acknowledge all her contributions to keep me productive and I am grateful for the excellent example she has provided as a researcher and a mentor. I want to thank my beloved family and especially my parents, Irene and Christos and my sister Athena for their encouragement and support during my PhD studies. I feel bottomless gratitude for their sacrifices they made to ensure I will be able to achieve my goals. I dedicate this dissertation to Nikos and with his patience and commitment, I am sure he will be a great scientist. I dedicate it to Athena too, wishing her to be a great doctor. I would also want to express my sincere appreciation to our collaborators Dr. Noelle Anglin, Dr. Dave Ellis in the International Potato Center (CIP), at Lima, Peru and Dr. Helen Tai in the Agriculture and Agri-Food Canada for giving the opportunity to work on their potato genomics data, advising me and guiding me through my research. Their contribution was very important for the completion of my work. I would also like to thank all the past and present lab members for their support and enjoyable discussions. My great gratitude to the McGill Department of Plant Science Graduate Excellence Fund; and Margaret A. Gilliam for the Fellowship in Food Security and Schulich Scholarships. Without their financial support I would not be able to complete my studies. To my supervising committee, Prof. Jean-Benoit Charron, Prof. Jeff Xia and Prof. Olivia Wilkins, I express my sincere thankfulness for their help and guidance through this project. The authors acknowledge funding through a Nouvelles Initiatives (Project International) grant from the Centre SÈVE (Fonds de recherche du Québec - Nature et technologies (FRQ-NT) to M.V.S., N.A., D.E., and H.H.T.; the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. 283303) to M.V.S.; A-base funding from Agriculture and Agri-Food Canada to H.H.T.; the McGill Department of Plant Science Graduate Excellence Fund; and Margaret A. Gilliam for the Fellowship in

Food Security and the Schulich Scholarship. The authors also gratefully acknowledge the support of the CGIAR Genebank Platform and appreciatively express thanks to the financial support for the sequencing by GIZ on behalf of the Federal Ministry of Economic Cooperation and Development, Germany (to N.A. and D. E.) and Compute/Calcul Canada Resource Allocations for Research Portals and Platforms (The Potato Genome Diversity Portal) and Resources for Research Groups to M.V.S. Finally, the authors would like to express their appreciation to Rene Gomez for his support and expertise in selecting the type accessions of each taxa to sequence.

List of Abbreviations

2OG	2-oxoglutarate
ADG1	*Solanum tuberosum* subsp. *andigenum*
ADG2	*Solanum tuberosum* subsp. *andigenum*
AJH	*Solanum xajanhuiri*
BAC	Bacterial Artificial Chromosome
bHLH	basic helix-loop-helix
bp	base pair(s)
BUK	*Solanum bukasovii*
BUSCO	Benchmarking Universal Single-Copy Orthologs, tool
CDS	coding sequence
CHA	Solanum chaucha
CIP	Centro International de la Papa, International Potato Center
CNV	Copy Number Variation
COM	*Solanum commersonii*
CUR	*Solanum curtilobum*
del	deletion
DM1-3	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*
GON2	*Solanum stenotomum* subsp. *goniocalyx*
GWAS	Genome Wide Association Studies
INDEL	Insertions Deletions
JUZ	*Solanum juzepczukii*
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	National Center for Biotechnology Information
NCBI SRA	NCBI Sequence Read Archive
NGS	Next Generation Sequencing

OLC	Overlap-Layout-Consensus
PacBio	Pacific Biosciences
PAE	Preferential Allele Expression
PAV	Presence/Absence Variation
PCA	Principal Component Analysis
PE	Paired-end
PGSC	Potato Genome Sequencing Consortium
PHU	*Solanum phureja*
PPR	Pentatricopeptide repeat
RLK	Receptor-like kinases
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
SAF	Number of alternate observations on the forward strand
SAR	Number of alternate observations on the reverse strand
SAUR	Small Auxin Up-RNA
SC	Self-Compatibility
SCF	Skp1-Cullin1-F-box
SE	Single End
SERPIN	Serine protease inhibitor
SI	Self-incompatibility
SCF	Skp1- Cullin1 – F-box
SMRT	Single Molecule Real-Time

SLFs	S-locus F-box proteins
SNP	Single Nucleotide Polymorphism
STN	*Solanum stenotomum* subsp. *stenotomum*
SV	Structural Variation
TGS	Third Generation Sequencing
TIR	Toll/interleukin-1 receptor-like
TBR	*Solanum tuberosum* subsp. *tuberosum*
TMV	Tobacco Mosaic Virus
ToMV	Tomato Mosaic Virus
UDP	Uridine Diphosphate
UV-B	Ultraviolet B
WGS	Whole Genome Shotgun

Thesis Format

This thesis is presented in manuscript–based format, consisting of four manuscripts included as chapters. Chapter 2 (Literature Review), entitled ”Current strategies of polyploid plant genome assembly” has been published in Frontiers in Plant Science on November 21, 2018 (Kyriakidou et al., 2018). Chapter 3 ”Structural genome analysis in cultivated potato taxa” is published in Theoretical and Applied Genetics (Kyriakidou et al., 2020a). Chapter 4: “A pan-genome model for diploid potato” is under revisions for resubmission to Theoritical and Applied Genetics (April 2020). Finally, Chapter 5: ”A genome assembly of six polyploid potato genomes” is published in Scientific Data (Kyriakidou et al., 2020b). The aforementioned manuscripts have been reformatted for thesis consistency.

Contribution of Authors

Chapter 2 is co-authored by Maria Kyriakidou, Dr. Helen Tai, Dr. Noelle Anglin, Dr. David Ellis and Dr. Martina Strömvik. Maria Kyriakidou drafted the manuscript, completed the tables, and made the figure under the supervision of Dr. Martina Strömvik. Maria Kyriakidou, Dr. Helen Tai, Dr. Noelle Anglin, Dr. David Ellis and Dr. Martina Strömvik designed the outline, content and edited the manuscript. A version of this chapter is published in Frontiers in Plant Science on November 21, 2018 (Kyriakidou *et al.*, 2018). Chapter 3 is co-authored by Maria Kyriakidou, Sai Reddy Achakkagari, Jose Hector Galvez Lopez, Xinyi Zhu, Chen Yu Tang, Dr. Helen Tai, Dr. Noelle Anglin, Dr. David Ellis and Dr. Martina Strömvik. Dr. Noelle Anglin and Dr. David Ellis provided the raw potato genome sequencing data and helped design the project. Maria Kyriakidou drafted the manuscript and performed the research under the supervision of Dr. Martina Strömvik. Sai Reddy Achakkagari performed the analysis for the comparison of the diploid genomes against the M6 genome, Xinyi Zhu performed the analysis for ADG1 against DM1-3, while Chen Yu Tang of GON1 against DM1-3. Jose Hector Galvez contributed to the bioinformatics methods. Dr. Helen Tai, Dr. Noelle Anglin, Dr. David Ellis and Dr. Martina Strömvik edited the manuscript. Chapter 4 is co-authored by Maria Kyriakidou, Dr. Noelle Anglin, Dr. David Ellis and Dr. Martina Strömvik. The project was designed by Maria Kyriakidou and Dr. Martina Strömvik. Maria Kyriakidou performed the research under the supervision of Dr. Martina Strömvik and wrote the initial manuscript. Chapter 5 is co-authored by Maria Kyriakidou, Dr. Noelle Anglin, Dr. David Ellis and Dr. Matrina Strömvik. The project was designed by Maria Kyriakidou and Dr. Martina Strömvik. Maria Kyriakidou performed the research under the supervision of Dr. Martina Strömvik and wrote the initial manuscript. Computations were made on the supercomputers Graham, Cedar and Béluga managed by Compute Canada, thanks to Compute / Calcul Canada Resource Allocations for Research Portals and Platforms (The Potato Genome Diversity Portal) and Resources for Research Groups awarded to M.V. S..

The sequencing data of this study are found at (Sequence Read Archive) SRA archive of NCBI, under the BioProject PRJNA556263.

Introduction

Sequencing plant genomes and discovering novel genes can potentially be used for plant breeding programs. For instance, novel genes involved in stress tolerance (e.g. heat, cold) can be used as a pool of potential genes for new breeding strategies to ensure food security in a rapid growing global population and in climate change. Whole genome sequencing, gene expression profiling, sequence polymorphisms, and genome-association studies will give a broader view of how multiple genes are working together, which is crucial for the success of a breeding program, as many important agronomic traits are polygenic (e.g. yield, height and nutrient content).

1.1 Potato Importance

Potato (*Solanum tuberosum* L.) is native to South America. Archeological evidence suggests that it originated in the Peruvian Andean highlands, while others have hypothesized that it originated in an area of Chile, Argentina and Bolivia in southern and central South America (Bradeen et al., 2011). Potato consumption began about 12,000 years ago (Rodr\'\iguez and Spooner, 2009).

Potato can grow at high altitudes and in conditions where many other crops cannot grow. It is a short growing season plant and it does not require a lot of advanced agricultural practises to be cultivated (Smith, 2012). According to International Potato Center's/CIP's website (CIP, 2018), potatoes can grow from sea level up to 4,700 meters above the sea level; from southern Chile to Greenland. Figure 1.1 shows a map of potato production quantities per country in 2017, showing its widespread success as a crop (FAO, 2013), retrieved on November 29, 2019).

A medium sized potato contains around 100 calories and it is a good source of vitamins C and B6 and of other minerals like iron, potassium and zinc. Its skin is a great source of dietary fibre, it does not contain any cholesterol and it is low in sodium. Potato can be stored for months under proper conditions and it is easily adaptable into a variety of

dishes. In summary, potatoes are important for human consumption and for meeting the challenges of world hunger, because of their nutritional content and high productivity rates. However, potato crop improvement is an ongoing effort to create climate change resilient cultivars. For example, since potato is a cool season crop it is crucial that breeding efforts include resilience to heat stress, in the face the rising global temperatures. In addition, potato production is growing rapidly worldwide, with the biggest expansions in production in subtropical and tropical areas (Birch et al., 2012). Temperatures above 25^oC delay the start of tuberization, leading to the enhanced development of the shoots, decreasing the yield (Levy and Veilleux, 2007). Hence, to overcome this challenge wild potatoes are introduced in breeding programs since they have a diverse range of habitats as well as various levels of stress tolerance (Smillie et al., 1983). Some examples of wild species that show heat tolerance and that have been used in breeding programs are: *S. berthaultii*, *S. chacoense*, and *S. stoloniferum* (Reynolds and Ewing, 1989).

1.2 The potato genome

Potato, (*Solanum tuberosum* L.) belongs to the Solanaceae (nightshade) family. Different potato species are naturally distributed from the southwestern United States to south-central Chile and adjacent Argentina (Hijmans and Spooner, 2001). Cultivated potato landraces are restricted to South America (from western Venezuela to south-central Chile) (Gavrilenko et al., 2013). The most common cultivated potato cultivars are autotetraploids (2n = 4x = 48), the basic chromosome number is 12, but other cultivated species can range from diploids (2n = 2x = 24) to pentaploids (2n = 5x = 60) (Watanabe, 2015). Wild potato species grown in the United States, Mexico and central America also include hexaploid species (Lara-Cabrera and Spooner, 2004). The potato genome is characterized by a great heterozygosity, which probably derives from the fact that most of the diploid potato species are self-incompatible and outcrossing is reinforced (Bradshaw, 2007; Watanabe, 2015).

Since potato has an extremely large secondary gene pool consisting of related wild species,

its taxonomy has been subject to study for many years (Machida-Hirano, 2015). Different researchers have proposed different taxonomic classifications of wild and cultivated potatoes and the debate is ongoing (Hijmans and Spooner, 2001; Huamán and Spooner, 2002; Ovchinnikova et al., 2011; Schmiediche et al., 1980). This difficulty of potato taxonomy is caused by introgression, interspecific hybridization, auto- or allopolyploidy, sexual compatibility among many species, a mixture of sexual and asexual reproduction, recent species divergence, phenotypic plasticity and high morphological similarity among species (Spooner, 2009; Spooner and Bamberg, 1994).

Potato is typically asexually (vegetatively) propagated through its tubers (Watanabe, 2015). The cultivated landraces are genetically narrower than the wild ones in South America as a result of the limited amount of introduced germplasm (Kloppenburg and Kleinman, 1987). Throughout history, there has been a need for introducing genetic variations into the current cultivars as was demonstrated by the disastrous Irish potato famine. During the mid-nineteenth century, the Irish cultivated potatoes were infected by the oomycete Phytopthora infestans, the plant pathogen that causes late blight disease. This resulted in a failed crop, million deaths and forced people to emigrate from Ireland in large numbers (Ristaino, 2002). Currently, climate change is threatening the potato by decreasing diversity in South America (Machida-Hirano, 2015). The need is great to safeguard genetic diversity in potato for current and future potato cultivars, but also for conserving the wild potato genetic diversity in South America.

For over a century, wild potatoes have been used for disease resistance in breeding programs (Hawkes, 1958). Wild relatives and ancient cultivars of potato consist of a great genetic resource for breeding programs for disease resistance, environmental tolerance and other qualities of interest (Machida-Hirano, 2015). Currently, the breeding programs are using recurrent parents to produce cultivars with desired traits, leading to inbreeding depression (Bradshaw et al., 2006). Additionally, inbreeding programs are restricted because of the autotetraploidy, heterozygosity and the susceptibility of potato to pathogens (Bradshaw et al., 2006; PGSC, 2011). Next generation sequencing (NGS) is a very powerful tool for breeding programs and for this reason the annotation of the potato genome

is essential in order to better understand its complex genome and create better breeding programs (PGSC, 2011). Transcriptomic analyses contributed to understanding of molecular mechanism of white and purple potato development and uncovered genes involved in anthocyanin biosynthesis (Liu et al., 2015). Moreover, other transcriptomic studies revealed novel, candidate genes involved in photoperiodic tuberization in potato (Shan et al., 2013) and also provided evidence for regulatory coordination of N sufficiency responses at the gene level in potato (Gálvez et al., 2016).

In 2011, the first potato genome was assembled and published by the (Potato Genome Sequencing Consortium) PGSC using a doubled monoploid homozygous potato S. tuberosum group Phureja DM1-3 516 R44 (referred as DM or DM1-3) (PGSC, 2011). The assembly was performed using a whole-genome shotgun (WGS) approach and it is consisted of 727 Mb. Non-gapped sequences made up 93.3% and repetitive content 62.2% (PGSC, 2011). By combining RNA-Sequencing (RNA-Seq) reads, ab initio gene prediction, protein and Expressed Sequence Tags (ESTs) alignments, 39,031 coding genes were identified, with 9875 alternatively spliced genes (PGSC, 2011). A couple of years after the first potato genome assembly, an updated version was published, containing additional sequence data for a better reference genome: version 4.03 (Sharma et al., 2013). This version contains 95% genome super scaffolds of which the 90% have been assigned to an absolute or relative orientation with the pseudomolecules. The most up to date version (v4.04) of the potato reference genome was released previously, in February 2016 (Hardigan et al., 2016). The assembly was performed using additional data from the stem of a potato, which was cloned from the original DM1-3 reference. This data added 55.7 Mb of novel sequences in the form of $>$ 200 bp contigs, including new genes that did not map to the v4.03 reference genome. These contigs were concatenated into an unanchored pseudomolecule called "chUn", which was also annotated (Hardigan et al., 2016).

Within the significant efforts for enriching the wild potato genomic resources, the genomes of two wild species have been sequenced and assembled; *Solanum commersonii* (Aversano et al., 2015) and *Solanum chacoense* (Leisner et al., 2018). *S. commersonii* is a diploid tuber-bearing wild potato species, native to Central and South America. It is resistant

to root knot nematode, soft rot and blackleg, bacterial verticillium wilt, Potato virus X, tobacco etch virus, common scab, late blight and frost tolerance and good acclimation capacity (GRUNDT et al., 2005; Hawkes and Others, 1990; Micheletto et al., 2000). Its genome sequence was released earlier in 2015, it consists of 830 Mb and it contains 39,290 protein-coding genes (Aversano *et al.,* 2015). *S. chacoense* is another closely related tuber-bearing wild species with potential significant agronomic traits, such as high dry matter, good chip-processing qualities, and resistance to cold-induced sweetening, (Leisner et al., 2018). The genome assembly is 825 Mb long and it has 37,749 protein-coding genes (Leisner et al., 2018). Recently, there has been a report that questions whether the genome of the M6 clone is a pure *S. chacoense* (Corentin Clot, 2020).

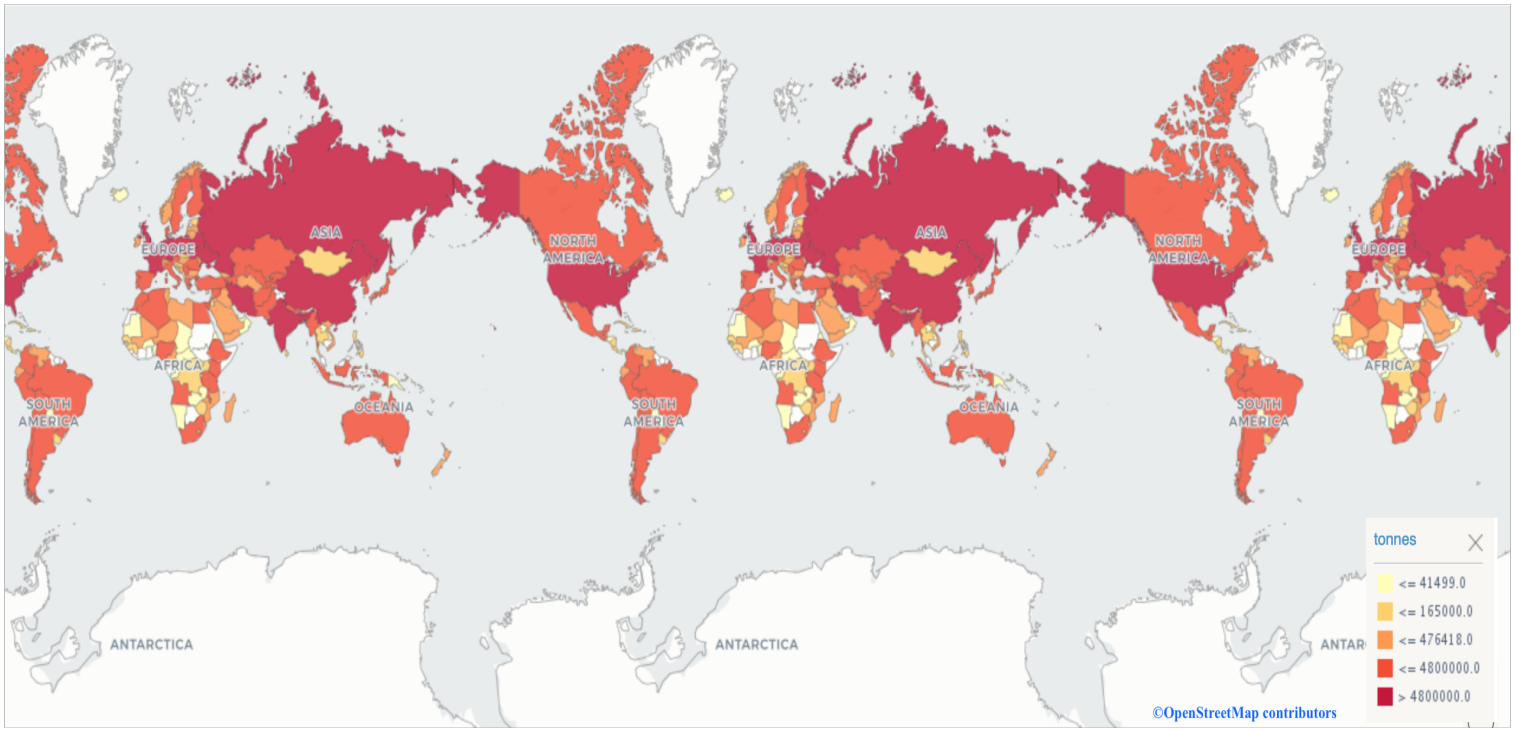


Figure 1.1: Potato production per country.
The map shows the potato production per country, worldwide in the year 2017. The yellow indicates production equal or less than 41,499 tonnes and dark red, production more than 4,800,000 tonnes. Potato is a highly adapted crop that can grow in various climates and altitudes. The map was retrieved from FAO and it is available under the Open Database License, the cartography is licenced as CC BY-SA (https://www.openstreetmap.org/copyright).

Hypotheses and Objectives

1.3 Hypothesis for Chapter 3

Genes in regions of structural variations in six diploid, two triploid, three tetraploid and a pentaploid potato genomes are primarily involved in defense mechanisms and tuber formation.

1.4 Objectives for Chapter 3

1. Illumina PE reads of the potato genomes will be aligned to the two publicly available reference genomes (DM1-3 and *S. chacoense* (M6)).

2. The aligned sequences will be used to annotate genes and to perform structural variation analyses to identify copy number variations (CNVs).

3. The regions discovered after the CNV analyses will be examined for gene content and function.

1.5 Hypothesis for Chapter 4

Will having a diploid pan-genome for potato available reveal important genes that can improve prospects for breeding cultivars with disease and climate change resilience.

1.6 Objectives for Chapter 4

1. Longer and Linked sequencing reads will be used to assemble *de novo* the genome of the diploid landrace *S. stenotomum* subsp. *goniocalyx*.

2. A pan-genome will be assembled from five sequenced diploid potato landrace genomes, three diploid wild potato genomes and the DM1-3 reference genome, and a presence-absence variation analysis will be used to determine a core and accessory genome.

3. The contigs that do not align to the DM1-3 reference genome will be identified and annotated using related genomes and included in the pan-genome.

4. The contigs will be annotated to discover genes absent from the DM1-3.

5. The Illumina reads of the diploid genomes will be aligned to the pan-genome to identify presence/absence variation.

1.7 Hypothesis for Chapter 5

Third Generation Sequencing Technology data will improve the de novo genome assembly of the polyploid potato genomes.

1.8 Objectives for Chapter 5

1. Longer and Linked sequencing reads will be used to assemble *de novo* one of the newly sequenced diploid genomes.

2. The five newly sequenced diploid genomes will be assembled *de novo* with only Illumina PE reads.

3. The *de novo* genome assemblies of the six potatoes will be compared.

Preface to Chapter 2

Chapter 2 constitutes a literature review on complex plant genome sequence assembly. It was published in Frontiers in Plant Science in November 2018 and has to date (17 months after publication) been cited 30 times.

Literature Review

Current strategies of polyploid plant genome sequence assembly

Maria Kyriakidou[1], Helen H. Tai[2], Noelle L. Anglin[3], David Ellis[3], Martina V. Stromvik[1*]

[1] Department of Plant Science, McGill University, Montreal, Canada

[2] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, Canada

[3] International Potato Center, Lima, Peru

[*] Correspondence: martina.stromvik@mcgill.ca

Keywords: polyploidy, plant genomics, genome assembly, third generation sequencing, reference genome

2.1 Abstract

Polyploidy or duplication of an entire genome occurs in the majority of angiosperms. The understanding of polyploid genomes is important for the improvement of those crops, which humans rely on for sustenance and basic nutrition. As climate change continues to pose a potential threat to agricultural production, there will increasingly be a demand for plant cultivars that can resist biotic and abiotic stresses and also provide needed and improved nutrition. In the past decade, Next Generation Sequencing (NGS) has fundamentally changed the genomics landscape by providing tools for the exploration of polyploid genomes. Here, we review the challenges of the assembly of polyploid plant genomes, and also present recent advances in genomic resources and functional tools in molecular genetics and breeding. As genomes of diploid and less heterozygous progenitor species are increasingly available, we discuss the lack of complexity of these currently available reference genomes as they relate to polyploid crops. Finally, we review recent approaches

of haplotyping by phasing and the impact of third generation technologies on polyploid plant genome assembly.

2.2 Introduction to polyploidy

The fusion of two or more genomes within one nucleus results in polyploidy, resulting in each cell containing more than two pairs of homologous chromosomes. Polyploidy occurs in the majority of angiosperms and is important in agricultural crops that humans depend on for survival. Examples of important polyploid plants used for human food include, *Triticum aestivum* (wheat), *Arachis hypogaea* (peanut), *Avena sativa* (oat), *Musa sp.* (banana), many agricultural *Brassica* species, *Solanum tuberosum* (potato), *Fragaria ananassa* (strawberry), and *Coffea arabica* (coffee). Autopolyploidy results from whole genome duplication, while an allopolyploid is characterized by interspecific or intergeneric hybridizations followed by chromosome doubling (Chen, 2010; Doyle et al., 2008). Genome duplication (autoployploidy) can be a source of genes with novel functions leading to new phenotypes and novel mechanisms for adaptation (Crow and Wagner, 2005). Autopolyploids typically suffer from reduced fertility whereas allopolyploids have potential for heterosis or hybrid vigor (Ramsey and Schemske, 1998). Polyploidy generates great genetic, genomic, and phenotypic novelty (Soltis et al., 2016); however, the higher complexity between genotype and phenotype in polyploids compared to diploid plants makes linking genotype to phenotype a challenging task. For example, allopolyploid plant cells have complex regulatory mechanisms in order to unify gene expression between the homeologues and define their relative contributions to the final phenotype. Hence, polyploidization is one of the major forces of plant evolution and is intimately linked to speciation and diversity (Bento et al., 2011). It is estimated that around 80% of all living plants are polyploids (Meyers and Levin, 2006), while many plant lineages including monocots (i.e *Oryza*) and eudicots (*Arabidopsis*) have at least one paleo-polyploidy event in their history.

Table 2.1: Sequenced plant polyploid genomes through May 2019.
Light blue shows the % of complete and single copy genes, the darker blue the % of complete and the duplicated genes, the yellow the % of fragmented genes and finally the red shows the % of missing genes in the assemblies.

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Arabidopsis lyrata subsp lyrata	206.823	Scaffold	2009-11-30	Tetraploid	(Hu et al., 2011)
Glycine max	978.972	Chromosome	2010-01-05	Allotetraploid	(Schmutz et al., 2010)
Triticum aestivum	15344.7	Chromosome 3B	2010-07-15	Allohexaploid	(Choulet et al., 2010)
Solanum tuberosum	705.934	Scaffold	2011-05-24	Autotetraploid	(PGSC, 2011)
Actinidia chinensis	604.217	Contig	2013-09-16	Tetraploid	(Huang et al., 2013)
Fragaria orientalis	214.356	Scaffold	2013-11-27	Tetraploid	(Tanaka et al., 2016)
Fragaria x ananassa	697.762	Scaffold	2013-11-27	Allooctaploid	(Tanaka et al., 2016)
Beta vulgaris	566.55	Chromosome	2013-12-18	2n, 4n (Beyaz et al., 2013; Dohm et al., 2014)	

Continued on next page

Table 2.1 – *Continued from previous page*

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Oryza minuta	45.1659	Chromosome	2014-04-16	Tetraploid	(Oryza Chr3 Short Arm Comparative Sequencing Project, 2014)
Camelina sativa	641.356	Chromosome	2014-04-17	Hexaploid	(Kagale et al., 2014)
Brassica napus	976.191	Chromosome	2014-05-05	Allotetraploid	(Chalhoub et al., 2014)
Brassica oleracea var. oleracea	488.954	Chromosome	2014-05-22	Hexaploid	(Parkin et al., 2014)
Nicotiana tabacum	3643.47	Scaffold	2014-05-29	Allotetraploid	(Sierro et al., 2013)
Eragrostis tef	607.318	Scaffold	2015-04-08	Allotetraploid	(Sierro et al., 2013)
Gossypium hirsutum	2189.14	Chromosome	2015-04-29	Allotetraploid	(Li et al., 2015)
Zoysia japonica	334.384	Scaffold	2016-03-15	Tetraploid	(Li et al., 2015)
Zoysia matrella	563.439	Scaffold	2016-03-15	Allotetraploid	(Li et al., 2015)
Zoysia pacifica	397.01	Scaffold	2016-03-15	Allotetraploid	(Li et al., 2015)

Continued on next page

Table 2.1 – *Continued from previous page*

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Musa itinerans	455.349	Scaffold	2016-05-21	2n, 3n hybrids (Jarvis et al., 2017)	
Rosa x damascena	711.72	Scaffold	2016-06-13	Tetraploid	BIO-FD & C CO., LTD
Chenopodium quinoa	1333.55	Scaffold	2016-07-11	Tetraploid	(Jarvis et al., 2017)
Brassica juncea var. tumida	954.861	Chromosome	2016-07-19	Allotetraploid	(Jarvis et al., 2017)
Hibiscus syriacus	1748.25	Scaffold	2016-07-29	2n, 3n, 4n (Kim et al., 2019)	
Gossypium barbadense	2566.74	Scaffold	2016-10-28	Tetraploid	Huazhong Agricultural University
Momordica charantia	285.614	Scaffold	2016-12-27	2n to 6n (Urasaki et al., 2016)	
Drosera capensis	263.788	Scaffold	2016-12-30	Tetraploid (Butts et al., 2016)	

Continued on next page

Table 2.1 – *Continued from previous page*

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Capsella bursa-pastoris	268.431	Scaffold	2017-01-29	Tetraploid	Lomonosov Moscow State University
Saccharum hybrid cultivar	1169.95	Contig	2017-03-03	It varies (Augusto Corrêa dos Santos et al., 2017)	
Xerophyta viscosa	295.462	Scaffold	2017-03-31	Hexaploid	(Costa et al., 2017)
Triticum dicoccoides	10495	Chromosome	2017-05-18	Tetraploid	WEWseq consortium
Utricularia gibba	100.689	Chromosome	2017-05-31	16-ploid	(Lan et al., 2017)
Eleusine coracana	1195.99	Scaffold	2017-06-08	Allotetraploid	(Hittalmani et al., 2017)
Dioscorea rotundata	456.675	Chromosome	2017-07-28	Tetraploid	Iwate Biotechnology Research Center
Ipomoea batatas	837.013	Contig	2017-08-26	Autohexaploid	(Yang et al., 2017)
Echinochloa crus-galli	1486.61	Scaffold	2017-10-23	Hexaploid	(Qiu, 2017)

Continued on next page

Table 2.1 – *Continued from previous page*

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Pachycereus pringlei	629.656	Scaffold	2017-10-31	Autotetraploid	(Zhou et al., 2017)
Olea europaea	1141.15	Chromosome	2017-11-01	2n, 4n, 6n (Ming et al., 2008; Unver et al., 2017)	
Monotropa hypopitys	2197.49	Contig	2018-01-03	Hexaploid	(Gruzdev, E.V., Beletsky, A.V., Mardanov, A.V., Kochieva and Ravin, N.V. and Skryabin, 2018)
Dactylis glomerata	839.915	Scaffold	2018-01-19	Autotetraploid	Sichuan Agricultural University
Panicum miliaceum	848.309	Scaffold	2018-01-23	Allotetraploid	(Shi, 2018)
Euphorbia esula	1124.89	Scaffold	2018-02-06	Hexaploid	(Sato et al., 2010)

Continued on next page

Table 2.1 – *Continued from previous page*

Organism Name	**Genome Size (Mb)**	**Current Status**	**1st Release date in NCBI**	**Ploidy level**	**Reference/Center**
Santalum album	220.961	Scaffold	2018-02-12	2n, 4n etc (Xie, S.-Q., Zhang, X.-M., Han, Y. and Ling, 2018)	Centre for Cellular and Molecular Platforms
Avena sativa	67.3266	Contig	2018-02-26	Hexaploid	The Sainsbury Laboratory
Panicum miliaceum	850.677	Chromosome	2018-04-09	Tetraploid	(Shi, 2018)
Arachis monticola	2618.65	Chromosome	2018-04-23	Tetraploid	(Yin, 2018)
Arachis hypogaea	2538.28	Chromosome	2018-05-02	Allotetraploid	(Schmutz, J., Jenkins, J., Grimwood, J., Bertioli et al., 2018)
Artemisia annua	1792.86	Scaffold	2018-05-08	Tetraploid	(Shen et al., 2018)

2.3 Overview of the sequencing techniques and their applications in polyploid plant genomes

Genome sequencing was initiated in the mid 1970's with alternative methods to determine the composition of DNA in a target cell or organism (Maxam and Gilbert, 1977; Sanger and Coulson, 1975) . The first whole genome to be sequenced was that of a bacteriophage PhiX (Sanger et al., 1977) with a genome size at 5.3Kb. However, the revolution in sequencing technology came about when Sanger developed the chain termination or dideoxy method (Sanger et al., 1977) . This technique, now known as Sanger sequencing, was adopted by most molecular biology laboratories and was the primary method of sequencing for 30+ years allowing sequencing of fragments of approximately 800-1000bp.
It took over 20 years from the time the first genome of a bacteriophage was sequenced until plant biologists had a draft genome of a flowering plant. First to be sequenced was the genome of *Arabidopsis thaliana*, a small weedy plant (Initiative, 2000) . After the release of the *Arabidopsis* genome sequence, economically important crops such as *Oryza sativa* (rice), *Carica papaya* (papaya), and *Zea mays* (maize) were sequenced using Sanger sequencing (Ming et al., 2008,?; Sasaki and Project, 2005) . Yet, of these plant genomes, only rice and Arabidopsis were sequenced using the Bacterial Artificial Chromosome (BAC) approach, and thus, are more complete genomes, whereas the others are drafts in a less completed stage (Ming et al., 2008) .
The diploidized tetraploid genome of *Glycine max* (soybean) was the first polyploid plant genome released; publicly available in early 2008, (Schmutz et al., 2010), followed by the tetraploid *Arabidopsis lyrata* (Ming et al., 2008) **Table 2.1**. The soybean project was very costly, and the resulting assembly consisted of the largest published plant genome performed using the Sanger Whole Genome Sequencing (WGS) method. In 2011, the genome of *Jatropha curcas* (an oil-bearing tree) that has variable ploidy levels (**Table 2.1**), was also sequenced using the Sanger method (Dart et al., 2004) . The assembly of the complex tetraploid genome of cultivated cotton - *Gossypium arboretum* (Li et al., 2015) was fol-

lowed by the reference genome of wheat, derived from the assembly of the large complex genome of *Aegilops tauschii*, one of the three diploid progenitors of bread wheat (Zimin et al., 2017) .

Next Generation Sequencing (NGS) technologies became commercially available in 2004 (Mardis, 2008) reducing sequencing costs and increasing massively sequencing throughputs, but also expanding the complexity of fragment assembly due to its short-sequence read output. NGS allows genome sequencing to be performed with lower DNA concentrations and thus, has applications in genome sequencing and re-sequencing, metagenomics, transcriptomics (RNA-sequencing) and even in personal genomics (personal medicine). These techniques can reduce the gap between genotype and phenotype by combining for example genomics and transcriptomics data. Some of the NGS platforms that have been employed in recent years include: 454 or pyro-sequencing (by Roche, Basel, Switzerland, with read lengths up to 700 bp), SOLiD (by Life Technologies, Carlsbad, California, 50 bp), HiSeq (by Illumina, San Diego, California, 2 x 250 bp), MiSeq (by Illumina, 2 x 300 bp) and Ion Torrent/Proton (by Life Technologies, 200 bp). NGS technologies are advantageous because, unlike Sanger sequencing, DNA cloning is not required making the process simpler, with greater adaption for a broad range of biological phenomena, and massive parallelization at decreased costs. However, NGS does suffer from some disadvantages: the short sequence length requires unique assembly algorithms, base calling is less accurate than Sanger sequencing, and the quality of NGS assemblies is lower than those made from Sanger sequence (Claros et al., 2012) . Examples of polyploid plant genomes sequenced using Illumina technology are the first assembly of the hexaploid *T. aestivum* (wheat) genome (Choulet et al., 2010) , and the genome of *G. hirsutum* (cotton) (Li et al., 2015) . The genomes of *Brassica oleracea* (cabbage) and *B. napus* (rapeseed) (Chalhoub et al., 2014) were sequenced with a combination of 454 and Illumina technologies. A genome assembly service using only high-quality short Illumina reads is offered by NRGene's DenovoMAGIC platform (http://www.nrgene.com/technology/denovomagic/). The recently annotated allohexaploid wheat genome was constructed using DenovoMAGIC2 (Pfeifer et al., 2014) . The latest version; DenovoMAGIC v 3.0 promises production of

long, phased scaffolds using only NGS.

The emergence of the Third Generation Sequencing technologies consists of the most recent genome sequencing approaches, characterized by long reads. These methods have further reduced sequencing costs, simplified preparatory and sequencing methods (Appels et al., 2018) , while providing longer read lengths, typically measured in kilo bases (Kb) rather than bases (bp). While there are many upsides to this new technology, caveats include high error rates and a requirement for very high-quality DNA. However, these approaches currently look promising in meeting the challenges of sequencing and assembling large, repetitive, and complex plant genomes by the production of large quantities of long reads to help bridge difficult regions in the genome. There are currently two types of technologies included in the Third-Generation sequencing approaches: long-read sequencing and long-range scaffolding technologies (Jiao and Schneeberger, 2017) .

Among the long-read sequencing technologies, the most widely used technology is the Pacific Biosciences' Single Molecule Real-Time (SMRT), with an average read length 20 Kb. For the assembly of the *Chenopodium quinoa* genome, a read length of 12 Kb was reported using this technology (Jarvis et al., 2017) . Additionally, Illumina introduced another long-read technology, the Synthetic Long-Reads (SLR) from short-read sequencing data, with a median length of 8 – 10Kb (**Table 2.2**). However, a maximum length of 21 Kb was achieved in a sugarcane hybrid sequencing project (Riaño-Pachón and Mattiello, 2017) . SLR can be used to resolve the haplotype of individuals, which is highly desired in the case of polyploid plant genomes. Finally, Nanopore, introduced by Oxford Nanopore Technologies, can generate a median length greater than 5 Kb, however a 12 Kb median length was reported while sequencing the wild **Solanum pennellii** genome (Schmidt et al., 2017) .

Even with the rapid progress and improvement of long-read technologies, it is still not possible to assemble a complete diploid plant genome using only NGS sequencing reads (Jiao and Schneeberger, 2017) . Hence, long-range scaffolding technologies are essential for improving the contiguity of an assembly, which requires the extension of the contigs into scaffolds and eventually their alignment into chromosomes. Based on currently

available sequencing technologies, additional genetic and physical maps are required. An alternative approach is based on chromosome conformation capture sequencing (Hi-C) provided by Dovetail Genomics (https://dovetailgenomics.com/) and PhaseGenomics (https://phasegenomics.com/), which creates long-range mate pair data for NGS (AU - van Berkum et al., 2010; Lieberman-Aiden et al., 2009) . The generated data can be used for phasing and scaffolding, which captures the entire eukaryotic chromosomes when they are combined with high quality draft assemblies (Sedlazeck et al., 2018) . Genome phasing is the identification of the alleles in each of the chromosomes. The most recent announcement of the PhaseGenomics Biotechnology company is its collaboration with Pacific Biosciences for the release of FALCON-Phase (Sedlazeck et al., 2018) . FALCON-Phase tool promises to solve the haplotyping problem in diploids, by enabling the construction of fully-phased chromosome-scale assemblies by combining SMRT long reads and Hi-C data. The latest technology is from GemCode, introduced by 10X Genomics in 2015 (`www.10xgenomics.com`). This approach is similar to the SLR protocol of Illumina, but it can process longer fragments and it does not require as much read depth as the SLR. The average read length captured with this approach can be greater than 100 Kb (**Table 2.2**).

Table 2.2: Third Generation Sequencing Platforms

Technology	Reads	Drawbacks	Plant assembly
PacBio	Single molecule long-reads, average length 10 - 18 Kb	False insertions in the raw reads, high error rate. Error correction algorithms are required	*Chenopodium quinoa* (Jarvis *et al.,* 2017)
Oxford Nanopore	Single molecule long-reads, average length 10 Kb, max 100Kb	Raw reads with false deletions and homopolymer errors. Requirement for error correction algorithms	*S. pennellii*, *A. thaliana*, *O. coaectata* (Mondal *et al.,* 2017; Schmidt *et al.,* 2017; Michael *et al.,* 2018)
Illumina Synthetic Long reads	Synthetic long-reads derived from the short sequencing reads, average length 100 Kb	High rate false indels (insertions, deletions). They require good trimming, correction algorithms	*Saccharum sp.* (Riaño-Pachón & Mattiello, 2017)
10X Genomics	Linked reads derived from short-read sequences, average length 100 Kb*	Needs designed algorithms and aligners, poor resolution of locally repetitive sequences. Sparse sequencing	*Capsicum annuum* (Hulse-Kemp *et al.,* 2018)
BioNano Genomics	Optical mapping of long, fluorescently labelled DNA fragments, average length 250 Kb	Not many algorithms available for a reliable alignment between the optical map and the genome assembly	*Brassica juncea* (Jinghua Yang *et al.,* 2016)

Continued on next page

Table 2.2 – *Continued from previous page*

Technology	Reads	Drawbacks	Plant assembly
Hi-C	Pairs short reads with an average length 100 bp, method originally developed to study the 3D folding of the genome	Scattered sequencing with variable genomic distance between pairs	*Triticum aestivum* (I. W. G. S. Consortium, 2014)

*10X Genomics is very similar to Illumina's SLR, with the difference that 10X Genomics can process more and larger fragments and the assemble of the different fragments does not necessarily depend on the sequencing coverage. Illumina's SLR system synthesizes the sequences of DNA fragment in contrast to 10x Genomics where the reads show only a part of DNA fragments.

2.4 Challenges of polyploid genome assembly

A reference genome is a digital, linear nucleic acid sequence containing only a single set of chromosomes plus any unanchored heterozygous contigs and/or scaffolds. A reference genome is used to observe the variations across different individuals within a species, to study evolution and to aid genome assembly. In the case of a polyploid genome, things become more complicated. For an allopolyploid organism, a reference genome contains the assembled DNA sequences of the ancestors' subgenomes (i.e *F. ananassa*, *B. napus*, *A. hypogaea*, *G. hirsutum* and *T. aestivum*) in addition to any unanchored sequences that are kept in additional pseudochromosome(s) (e.g., *T. aestivim*, *S. tuberosum*), and for an autopolyploid organism the genome that went through the duplication event(s) (e.g., *S tuberosum* in addition to any unanchored sequences. It does not represent any allelic variation present in the individuals. When high throughput sequencing reads are mapped to a reference genome, alternate alleles can be retrieved from each genomic region, based

on the sequencing coverage and diversity in the individual compared to the reference. These alternate alleles for an organism can be detected and used for haplotype assembly for each of the present haplotypes. Polyploid assembly is similar to the sum of a number of problems of haplotype reconstruction (Sedlazeck et al., 2018) ; hence, the computational complexity increases with higher ploidy. This means that the genome assembly of an n-ploid organism will result in the construction of n numbers of haplotypes. This is not an easy task as the knowledge of one haplotype does not automatically determine how to phase others (Sedlazeck et al., 2018) .

Whole-genome duplication events have also been associated with genome rearrangement, atypical recombination, transposable element activation, meiotic/mitotic defects, and intron expansions and DNA deletion (Sedlazeck et al., 2018) . The assembly of autopolyploid genomes is extremely challenging as fragments of a subgenome might be assigned to the wrong subgenome, which results in misassembled false genomes. Allopolyploids may present the same challenge, but given the greater genetic distance, resolving their subgenomes is likely less problematic during assembly. These events multiply the regular challenges of plant genome sequence assembly, such as repeat content, transposable elements, high heterozygosity, gene content and gene families of non-coding RNAs due to their repetitiveness after duplication events and the fact that their detection is crucial for proper genome annotation.

Polyploidization can lead to higher levels of heterozygosity, which can be confounded in asexually propagated plants such as potato causing greater difficulties in the identification of haplotypes. This is due to multiple alleles from the same locus being mistaken as sequences from different loci (Huang et al., 2014) . This is especially problematic when using short sequence reads for genotyping or genome assembly, because the results will be highly fragmented assemblies with a total assembly size longer than expected. In addition, contigs can break at polymorphic regions or misassemblies can occur between large-scale duplications (Claros et al., 2012) . This assembly problem is not unique to polyploid plants, however and can also occur in plants with segmental genome duplications.

The ploidy level of the plant genome must be carefully considered when choosing the appropriate assembly algorithm. The presence of two or more sets of genes within the same nucleus can affect the accuracy of the assembly, making it difficult to differentiate between homologues or homeologues (Claros et al., 2012) . (Glover et al., 2016) define homeologues as pairs of genes or chromosomes in the same species, derived by speciation but brought back to the same genome after a polyploidization event(s). Identifying functionally conserved homeologues however, provides important genetic material for crop improvement in many crops, including *Musa acuminata* (banana), *S. tuberosum* (potato), *Gossypium hirsutum* (cotton) and *T. aestivum* (wheat) (Chen and Dubcovsky, 2012; Glover et al., 2016) . Examples of how polyploids also confer emergent properties are seed oil accumulation in *Brassica napus* (canola), spinnable fibers in cotton, and grain composition in wheat (Michael and VanBuren, 2015) .

As mentioned above, several complex polyploid plant genomes have been sequenced. The decreasing costs of NGS technologies led to the sequencing and assembly of a number of polyploid plant genomes using these technologies (**Table 2.1**). Based on NCBI database (data retrieved on the 4th of July 2018: `https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/`), 320 land plants, 47 of which are polyploid, have been sequenced (as of 4th of July of 2018). Of the 72 assembled in 2017, 19 are polyploid, and three were released in January 2018. Only 16 polyploid plant genomes have been assembled into chromosomes, 26 assembled into scaffolds, and the rest (5) are still contigs (**Table 2.1**).

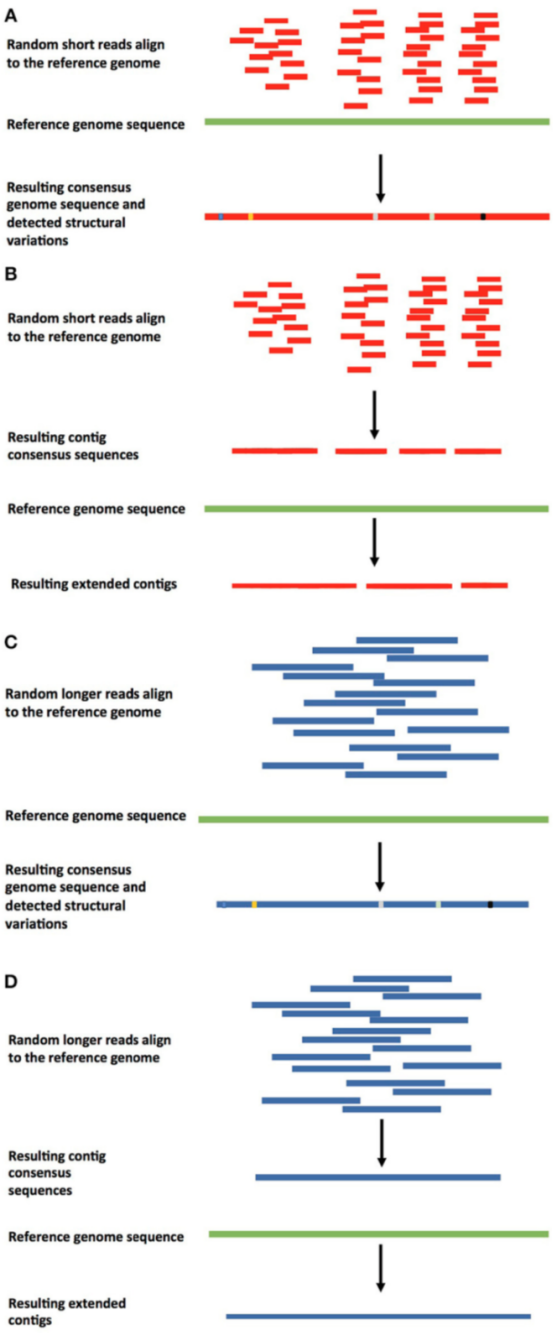


Figure 2.1: Approaches for reference - based genome assembly.
A. Shorter-read guided assembly. In this method, shorter reads are aligned against the reference genome, a consensus assembly is generated, and structural variations are detected. It can also be used to detect contamination in the sequenced reads. this approach is used when genomes are re- sequences to detect polymorphisms in individuals. **B.** Guided de novo genome assembly of shorter reads. Previously de novo assembled shorter reads are aligned against the reference or a closely related genome to extend the existing contigs. **C.** Longer-read guided assembly. Longer reads are aligned against the reference genome, a consensus genome assembly is constructed, and structural variations are detected. **D.** Guided de novo genome assembly of longer reads. Longer reads are de novo assembled into contigs, which are aligned against the reference or a closely related genome to be extended.

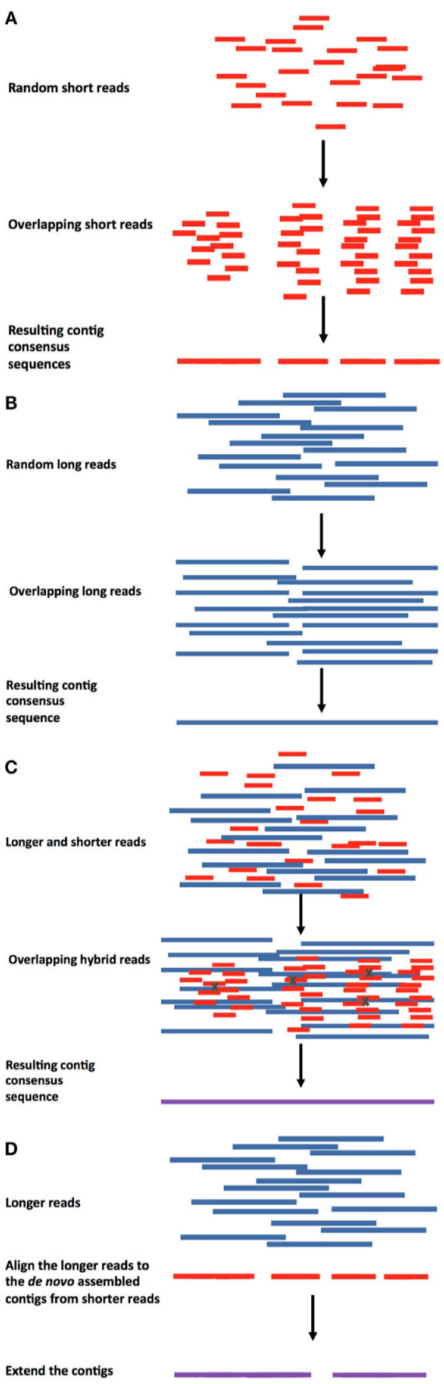


Figure 2.2: Approaches for *de novo* genome assembly.
A. Short read assembly. Genome assembly using only shorter read and any assembly tool to construct contiguous sequences/contigs. **B.** Longer reads assembly. Contig (red) assembly using longer reads (long, linked reads, optical maps) followed by scaffold assembly and gap filling. **C.** Hybrid genome assembly. In this method, shorter reads can be assembled into contigs and the longer reads can be used for error correction (errors represented by Xs), then the corrected contigs can be assembled into scaffolds and the gaps filled. **D.** Hybrid genome assembly using pre-assembled contigs. Longer reads are aligned against de novo pre-assembled contigs from shorter reads, followed by contig extension.

2.5 Technology-related challenges

There are two basic approaches to genome assembly. Comparative assembly is a reference guided method that uses the sequences of already assembled related organisms, a reference genome, for guidance. *De novo* assembly targets organisms that have not been sequenced before (Pop, 2009) , putting together the pieces without guidance from a prior reference genome. The two approaches are not completely mutually exclusive, because even in cases where reference genomes are available, regions that varied in the newly sequenced target genome need to be assembled de novo. Different approaches of guided and de novo genome assemblies can be found in **Figure 2.1** and **Figure 2.2**. The reference guided comparative assembly approach (**Figure 2.1**) can be performed in two ways: mapping short or long reads against the reference to construct a consensus (**Figure 2.1A** and **Figure 2.1C**) or assembling the reads de novo and then use the reference genome to orientate the resulting contigs or scaffolds in an alignment and identify misassembled regions (**Figure 2.1B** and **Figure 2.1D**) (Lischer and Shimizu, 2017) .

The reference-based comparative assembly approach is usually used when genomes are newly sequenced, or to correct misassemblies or extend existing contigs of already assembled genomes (**Figure 2.1B** and **Figure 2.1D**), and also for variant detection (**Figure 2.1A** and **Figure 2.1C**) and haplotype construction. An assembled genome sequence is used as a reference and the sequenced reads are independently aligned against this sequence. Dynamic programming is used to identify the optimal alignment for the candidate positions that match the best. Structural variations (such as insertions or deletions) in the newly sequenced genome(s) tend to increase the complexity of the alignment. The resulting alignment allows the extraction of the structural variants and construction of the haplotypes.

The *de novo* genome assembly method is applied when a reference genome sequence does not exist for a closely related species. In this case, the genome sequence is constructed through overlapping sequenced reads, usually using graph-based algorithms. It is difficult to perform *de novo* genome assembly, especially when only shorter reads are avail-

able. Both single end (SE) and paired-end (PE) reads are difficult to assemble de novo, with SE reads being slightly more challenging (i.e Illumina, **Figure 2.2A**). Long range reads can be used (**Figure 2.2B**), or a hybrid approach can be applied, where shorter and longer reads can be used together for a better assembly (**Figure 2.2C** and **Figure 2.2D**). As for the assessment, there are currently no unified assembly quality metrics to assess the quality of the *de novo* generated assembly, although one value that is commonly used is the N50. The value of N50 is a weighted median for when at least 50% of the assembly is contained in contigs or scaffolds of equal or greater length.

In general, the comparative method requires less computation as the sequenced reads are aligned to a reference genome. However, significant bias can occur in the comparative genome approach, as divergent (duplicated) regions of the genome may not get reconstructed properly, and thus, may completely miss the diversity present in the newly assembled genome (Lischer and Shimizu, 2017) . In contrast, the *de novo* genome assembly even for a diploid genome is classified as an "NP-hard" (non-deterministic polynomial-time) problem meaning it does not have an optimal, known solution. The genome assemblers must assemble a jigsaw puzzle of very small pieces. These pieces are the short reads (75bp-300bp) and different assembly tools are used to resolve a best-fit assembly. However, given that it is a NP-hard problem, most assemblies are likely only an approximation of the true genome order.

The assemblers also face the challenge of the repetitive nature of plant genomes along with heterozygosity and haplotype ambiguity that frequently splits these regions into multiple contigs. A number of algorithms are used for this computation. Some of the most well-known are the overlap computation, the Greedy algorithm (Huson et al., 2002) , the Eulerian path (Pevzner et al., 2001) , and two classes of assembly algorithms: Overlap-Layout-Consensus (OLC) and de Brujin graph. The overlap computation within an assembly tool requires a great deal of computational time, which can be easily reduced by parallelizing the computations using multi-processor machines or servers (Pop, 2009) . The complexity of the overlap computation is affected by the number of the input sequencing reads. Furthermore, the assemblers based on the Greedy algorithm give the

simplest (Pop, 2009) , most intuitive solution to the assembly problem, yet it is harder to prove the correctness of the algorithm even if the algorithm is correct (Pop, 2009) .

The OLC, which can effectively assemble very short reads, has been one of the most successful assembly strategies. The Eulerian approach was proposed as an alternative to the OLC for the assembly of Sanger data; however, because of its sensitivity to sequencing errors it has not been extensively used (Pop, 2009) . Overall, the short sequence reads need to be assembled into contigs, then the contigs need to be placed into bigger scaffolds, and finally chromosomes. Examples of tools that use the OLC algorithm in combination with other techniques is MASURCA that uses de Brujin graphs to construct mega-reads for a better assembly (Zimin et al., 2017) and BAUM that uses adaptive unique mapping to reconstruct repetitive regions (Wang et al., 2018) .

De novo genome assembly is essential to capture the biological diversity within newly sequenced genomes. Yet, this task is near impossible without the use of mate-pairs, longer reads, or linked reads to provide information that can bridge these difficult repetitive regions. Currently, there is a lack of genome assembly and mapping algorithms specialized for polyploid genomes. These would need to be optimized for using more computational power (resources) to handle the challenge of the increased complexity and size of the data sets. Polyploid genome assemblies made from only short reads fail to capture haplotyping variation and present only a single consensus sequence of several chromosome sets. Better algorithms are necessary to minimize misassembly of paralogous and orthologous regions in polyploid plant genomes.

Sequencing errors, read length, quality values, number of reads, and coverage are important factors in assembling genomes and there is little difference in these factors/variables between diploid and polyploid plant genomes. However, because of the complex nature of polyploid genomes, there is not "a best fit" for the main assembling pipeline and not every approach is reproducible for other polyploid plant genomes. Different results can be obtained from the various algorithms used for alignment and assemblers and often genome assemblies are only an estimate of the true biological genome. It often takes a decade or longer to make improvements and corrections to the original draft release.

For example, the human genome released in 2000 has gone through multiple revisions to correct errors. Furthermore, the metrics used to make comparisons tend to only focus on size which does not capture contig quality nor accuracy, and thus, there are no commonly accepted standardized methods for validation of the assemblers, which means most genomes are accepted as ”draft” assemblies (Narzisi and Mishra, 2011) . BUSCO (Simão et al., 2015) and QUAST (Gurevich et al., 2013) are two examples of tools that have been created in an attempt to validate the quality of an assembly.

2.6 How to estimate ploidy level in plants

The ploidy level in plants is normally estimated by measuring the C-value (amount of DNA in the unreplicated gametic nucleus) using flow cytometry (Clarindo et al., 2008; Dart et al., 2004; Eaton et al., 2004; GRUNDT et al., 2005; Harbaugh, 2008; Yang et al., 2011) . For example, flow cytometry was used to estimate genome content and ploidy in over 300 accessions of the Magnoliaceae family (Parris et al., 2010) , in six *Olea europea* (olive) subspecies (Besnard et al., 2007) , and in *B. napus* leaf tissue samples (Cousin and Nelson, 2009) . Public databases exist to capture C-value and ploidy levels in plants (e.g. `http://data.kew.org/cvalues/`). Recent tools have also been developed to infer the ploidy level using NGS data, such as ploidyNGS (Augusto Corrêa dos Santos et al., 2017) , ConPADE (Margarido and Heckerman, 2015) , and a pipeline using single nucleotide polymorphism (SNP) counts that was reported earlier by (Yoshida et al., 2013) for the estimation of ploidy level in the plant pathogen Phytophthora infestans. A general approach to estimate ploidy levels using NGS is by mapping the sequenced reads to the reference genome and then counting the number of mapped reads, representing the different alleles at each position. PloidyNGS (Augusto Corrêa dos Santos et al., 2017) was implemented by automating the process of observing the frequency of the alleles by generating a histogram. It was tested on diploid and haploid Saccharomyces cerevisiae datasets. ConPADE (Margarido and Heckerman, 2015) was specifically designed to estimate the ploidy levels of highly polyploid plant genomes and has been tested on wheat.

A weakness is its sensitivity to the quality of the mapping step as this can bias the ploidy estimation (Augusto Corrêa dos Santos et al., 2017) . Finally, the pipeline by (Yoshida et al., 2013) is similar in the sense that the distribution of read counts at biallelic SNPs is observed, which allowed the identification of diploid, triploid, and tetraploid *P. infestans* strains. Another recent statistical tool for ploidy estimation is nQuire (Weiß et al., 2018) , which uses NGS data to distinguish between diploids, triploids and tetraploids.

Ploidy estimation tools have been reported such as EAGLE (Loh et al., 2016) and Read-Sim (Schmid et al., 2006) . More recent tools for the haploid assembly consist of Hap-Compass (Aguiar and Istrail, 2013) , HaploSim (Bastiaansen et al., 2012) , HapCut (Bansal and Bafna, 2008) , and HapCUT2 (Edge et al., 2017) . Real and simulated data were analyzed with HapCUT2 (Edge et al., 2017) and it was shown that it is more accurate and can use not only WGS, but also SMRT (`www.pacb.com//smrt-science`) and Hi-C data (Lieberman-Aiden et al., 2009) for haplotype assembly. SWEEP (Clevenger and Ozias-Akins, 2015) is a tool designed to filter SNPs detected in newly sequenced autopolyploid and allopolyploid crops using NGS approaches. The detected SNPs can be further used for the haplotype construction. Another NGS tool is HANDS (Mithani et al., 2013) , which also can be used for auto- and allopolyploids and by aligning the sequenced reads to the reference genome(s) it can detect the subgenomes in polyploids. Longranger software by 10X Genomics can be used for phasing. It can determine which barcodes are associated with each heterozygous locus and while phasing, it can construct the organism's haplotypes. Simply, it aligns the raw reads to the sequence of both alleles to determine which allele each read represents.

2.7 How to "resolve" the ploidy issue (how to reduce the complexity of the problem)

2.7.1 Genome-related approach

Several strategies have been adopted for the sequencing and assembly of large polyploid genomes of crop plants (Bevan et al., 2017) . One approach involves the reduction of genome complexity using a natural or in vitro generated haploid. An example is the sequencing of the potato genome by the Potato Genome Sequencing Consortium (PGSC, 2011) . This genome was produced from a doubled monoploid that was homozygous for a single set of 12 chromosomes to generate a reference (PGSC, 2011) . A similar approach was used for the genome assembly of the hexaploid bread wheat, *T. aestivum*. Aneuploid bread wheat lines derived from double ditelosomic stocks of a hexaploid wheat cultivar were used to sequence each individual chromosome arm (except 3B) using Illumina short-reads technology (Pfeifer et al., 2014) . The chromosomes were assembled *de novo*, which reduced the complexity of assembling this highly redundant genome, aiding the differentiation of genes present in multiple copies and of highly conserved homologs.

A second approach involves sequencing a diploid progenitor species to aid in the assembly of the cultivated form. Care must be taken to choose the diploid progenitors most similar to the cultivated form. The diploid genomes of progenitor species can be used to determine the origin and structure of contigs when assembling large polyploid genomes. For example, strawberry (*Fragaria x ananassa*) is an octoploid (2n= 8x =56) whose origin remains controversial. One theory suggests that it was formed from a natural hybridization between two octoploids- *F. virginiana* and *F. chiloensis* (Darrow, 1966) . According to (Davis et al., 2007) , *F. vesca*, *F. nubicola*, and *F. orientalis* are possible progenitors. To access the genetic diversity of this valuable crop, one diploid variety of *F. vesca* (2n= 2x =14) (*F. vesca spp. vesca* accession Hawaii 4) was sequenced (Shulaev et al., 2011) .

Oilseed rape or canola (*B. napus*) is an allopolyploid derived from two diploid species of Brassica that are triplicated versions of an ancestral diploid. Genome assemblies of *B. na-*

pus were assigned to these two subgenomes using sequence assemblies from each diploid progenitor, but many sequence scaffolds showed ambiguous assignment to homeologous groups, owing to homeologue exchange and frequent gene loss (Chalhoub et al., 2014) . A similar strategy was used to characterize the allotetraploid genome of peanut (*Arachis hypogaea*), which formed from two diploid species *A. duranensis* (A genome) and *A. ipaensis* (B genome). Essentially complete assemblies of the genomes of the progenitor species *A. duranensis* and *A. ipaensis* were generated and shown to directly align with the genetic map of a cultivated tetraploid peanut (Bertioli et al., 2016) . In the same study, synthetic long-read sequencing of the tetraploid peanut genome showed that it was 98-99% identical to the diploid genomes, with differences due to recombination of polyploid genomes involved from the sequencing of DNA from purified chromosome arms (Bertioli et al., 2016) . Some of the challenges in assembling the cultivated peanut genome have been the high similarity between the two-progenitor species, a high number of transposable elements, and recent evidence of tetrasomic recombination in this allotetraploid (Bertioli et al., 2016) . Lastly, upland cotton (*G. hirsutum*) is an allotetraploid that formed 1-2 Myr (million years) ago from two unknown diploid progenitor species. The genome complexity of upland cotton was reduced by sequencing highly homozygous allohaploid lines to a coverage depth of 245x with Illumina short-read sequencing reads (Li et al., 2015) . A dense genetic map was used to align and correct scaffolds, which covered 96% of the estimated 2.5 Gb genome, and fluorescence in situ hybridization (FISH) was used to confirm a successful allotetraploid assembly.

2.7.2 Genome sequencing and algorithmic (pipeline) approach

There are several examples of successful *de novo* sequencing and assembly of large allopolyploid genomes of crops that use long-range alignments of sequence scaffolds to generate extended haplotypes to form distinctive homeologous pseudomolecules. Tobacco (*Nicotiana tabacum*; 2n=4x=48) is an allotetraploid that is derived from the diploid genomes of *N. sylvestris* and *N. tomentosiformis*. Whole-genome shotgun assemblies were

aligned to physical maps to create longer super scaffolds that could be assigned directly to the progenitor genomes (Sierro et al., 2013) . The polyploid genome of Indian mustard (*B. juncea*) (Yang, 2016) has been assembled using a combination of Illumina short reads, PacBio single molecule, real-time long sequence reads and optical maps from BioNano Genomics. The short and long reads were aligned to the maps, which directly helped in the determination of the individual molecules of tagged DNA, and dense genetic maps. The genome was almost fully represented in the assembly, which was assigned to the A genome (402 Megabase (Mb)) and the B genome (547 Mb).

Furthermore, an alternative approach to resolve polyploid complexity is by haplotyping. The process of assigning variants to a particular chromosome or defining which alleles appear together (corresponding haplotypes), is called phasing and haplotyping, respectively (Huang et al., 2014) . Haplotypes can provide more information than unphased genotypes in diverse fields, such as identifying genotype-phenotype associations and exploring genetic resistance to plant diseases. An example of this approach is the recent assembly of the hexaploid genome of sweetpotato (*Ipomoea batatas*). The authors describe haplotype construction by applying a novel approach (Yang et al., 2017) where paired reads and mate pairs were initially used for *de novo* assembly, then haplotypes were phased. Overlapping haplotypes were merged into larger haplotypes, mapping all the raw reads against the phased haplotypes. Finally, scaffolds were constructed based on the haplotypes and a consensus sequence was generated (Yang et al., 2017) . This method, called "Ranbow", can be downloaded at `https://www.molgen.mpg.de/ranbow`. A number of algorithms/tools to resolve the haplotype of polyploid genomes exist. Some examples are HANDS (Mithani et al., 2013) , SDhaP (Das and Vikalo, 2015) , and HapTree (Berger et al., 2014) . Haplotype construction depends on the read depth or coverage as it is necessary to have a high coverage for each homologue (5-20x per homologue), as well as an insert size of 600 – 800 bp (Motazedi et al., 2017) . It is also important to know the nature of the plant genome and ploidy before performing haplotyping in order to select the most appropriate tool. If available, it may be better to combine various individuals or parental information for haplotyping analysis (Motazedi et al., 2017) . From an algorith-

mic point of view, haplotyping requires a lot of memory and computation time.

Another solution is the construction of a pan-genome, which shows the variation and commonality between individuals. A pan-genome includes ”completeness” as it contains the core genome shared by all the individuals sequenced, but also the genes that are absent/present in some of the newly sequenced genomes. Generally, it is a very helpful approach for breeding applications as it anchors all the known variations and phenotype information and can include wild relatives of the cultivated crop lines. It also aids in the identification of novel genes from the available germplasm that are not found in the reference genome (TCP-G., 2018) . Additionally, it represents the polyploid genomes and in the case of the allopolyploids, it allows the quantification of allele dosage between germplasm samples (TCP-G., 2018) . Pan-genome construction is even more computationally challenging in the case of polyploid plant genomes as the corresponding genotype needs to be determined by variant calling and identifying novel variants for all the haploids. Previously, a pan-genome was constructed from 18 wheat cultivars and it was shown that a large number of variable genes affected by presence/absence and variation between the genes could be associated with important agronomic traits (Montenegro et al., 2017) . NRGene's (`www.nrgene.com`) PanMAGIC platform can be used for pangenome analysis and was applied to analyze six maize genomes (Lu et al., 2015) .

2.8 Third Generation Genomic Technologies come to the rescue

Genome assembly and scaffolding can be performed using shorter reads (Illumina data), or longer reads from either PacBio (`www.pacb.com`) or Oxford Nanapore (`https://nanoporetech.com/`), or a combination of both short and long reads. Another alternative is the assembly of linked reads from 10X genomics. Additionally, for higher contiguity, longer-range scaffolders from Dovetail (`dovetailgenomics.com`) and BioNano Genomics (`bionanogenomics.com`) can be used for the construction of physical maps using very large DNA fragments. A hybrid scaffolding approach can also be applied

where longer reads are used to improve assemblies generated using short-reads or even combined with longer-range scaffolding data.

Even though the hexaploid wheat genome was assembled from only short reads, it is very challenging to assemble such a large and highly repetitive genome using this approach. A less complicated assembly strategy is to use long-reads to aid in the assembly of difficult portions of the genome. The most widely used long-read sequencing technology is Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing. Recently, a few polyploid plant genomes were assembled using PacBio long reads including three allotetraploid plant genomes *C. quinoa* (quinoa) (Jarvis et al., 2017) , *Eleusine coracana* (finger millet) (Hatakeyama et al., 2017) and *Coffea arabica* (Arabica coffee) (Cheng et al., 2017) .

As mentioned earlier, another solution to the read length issue is the ultra-long and real-time data sequencing approach by Oxford Nanopore Technologies (`www.nanoporetech.com`). Currently three plant genomes have been sequenced with Nanopore, a wild tomato genome *Solanum pennellii* (Schmidt et al., 2017) , the genome of *A. thaliana* (Mondal et al., 2017) , and most recently the genome of *Oryza coarctata* (Michael et al., 2018) . Illumina's SLR technology on the other hand, has already been applied for the estimation of the haploid draft genome of the polyploid sugarcane hybrid SP80-3280 (Riaño-Pachón and Mattiello, 2017) .

The long-reads can also be combined with existing short-reads for genome assembly, called hybrid genome assembly. The resulting genome assembly from short-reads needs improvement in its contiguity because the contigs need to be assembled into scaffolds. Initially, the contigs are ordered using alignments from paired-end reads, read pairs from (Bacterial Artificial Chromosome) BAC or fosmid ends, which are powerful ways to increase the contiguity and help bridge the repeats - the main reason generally for breaks in the genome assemblies. In addition, genetic and physical maps are also essential for polyploid plant genome assembly (i.e - a physical map was used in the case of the tetraploid cotton genome). Optical mapping enables the fingerprinting of large genome fragments and can be used to improve highly fragmented genome assemblies. This technology promises the improvement of scaffolding and eventually lessens the need for genetic and

physical mapping (Jiao and Schneeberger, 2017) .

Another new promising technology that can potentially be applied to complex, polyploid plant genomes is the 10X genomics approach. There is only one scientific report on plant research using this technology to date on a diploid pepper genome (*Capsicum annuum*) (Hulse-Kemp et al., 2018) . The haplotype construction was generated to karyotype aneuploidy in a cancer study (Bell et al., 2017) and it was also used in the generation of a protocol for haplotyping human genome (Porubsky et al., 2017) , making it a promising technique for polyploidy genome data. Additional techniques used by polyploid plant projects include Hi-C and chromosome-scale assembly. For example, a study is underway to detect large chromosomal rearrangements in wheat genomes (Monat et al., 2019) and another project uses chromosome scale scaffolding on the allotetraploid coffee genome (Zimin et al., 2017) .

2.9 Advances in genomic resources and functional tools in molecular genetics and breeding

The advance of NGS technologies has immensely impacted the field of plant genomics in model and non-model crops alike, and it is continuously contributing to bridging the gap between genotype and phenotype. The genotype can be linked to the phenotype by Genome Wide Association studies (GWAS) and the advent of NGS has revolutionized genomic, as well as, transcriptomic (RNA-Sequencing) approaches to biology including plant genomics in model and non-model crops. Modern breeding programs combine various approaches for more efficient breeding, in parallel with the reduction of the whole breeding period (Varshney et al., 2013) . These approaches include the traditional phenotype-based selection, marker-assisted selection, and genome-assisted breeding (Varshney et al., 2013) . The continuous effort in improving major crops has resulted in great genetic and genomic resources for crop traits. Some instances of databases that host these resources can be found in **2.3**.

2.10 Lack of complexity of the currently available reference genomes of polyploid crops

High quality reference genomes, gene discovery, and comparative genomics depend on the construction of a high quality *de novo* genome assembly. These assemblies are more feasible, but still not perfect using haploid and inbred species. Despite their importance to reflect the genetic information within an organism, most of the currently available polyploid and diploid plant genome assemblies do not capture the heterozygosity present. The majority of the currently available reference genomes, especially those of the polyploids, lack variation and characteristics of other individuals that are not captured or presented. This happens because the simpler genomes are sequenced first, but also due to the sequencing of diploid and less heterozygous progenitor species for the reduction of the intricacy of the polyploid assembly problem. In reality, the assembled genome is a flat DNA sequence, which shows neither the variation between homologous chromosomes, nor allelic variations, or structural variations. The resulting ”model” reference genome is more distant than the majority of the other individuals in a species. Furthermore, genes may be missing or not annotated. A solution to this problem is the construction of pan-genomes (as described above), which show the core and the variable regions of a genome between individuals. An example of a pan-genome application is in the hexaploid bread wheat (Montenegro et al., 2017) .

Even in the case of the smaller, ”simpler” bacterial genomes, the submitted genomes are not complete. Despite the exponential generation of NGS data, the majority of the submitted genomes represent only draft or in scaffold format, incomplete genomes. The higher ploidy levels of the polyploid plant genomes make the situation even more difficult to handle. This leads to highly fragmented genome assemblies, with disconnected contigs of repetitive sequences. As discussed, better tools are needed that allow automatic contig assembly of (plant) genomes with many repeats and that are sensitive to ploidy levels and can handle haplotype construction. Also, to date allopolyploid plant genomes cannot be

represented in an integrated assembly, rather the subgenomes are found in separate assemblies.

2.11 Conclusions

Improving genome sequencing and assembly of polyploid plant crops will have a fundamental impact on genetic research and on plant breeding by better understanding the genomes, identifying genomic variants and relating them to economic, physiological, and morphological agronomic traits, such as higher yield, abiotic/biotic tolerance, root structure etc. Better polyploid plant genome assemblies will also aid in the study of the genotype-phenotype-environment relationship. For this, more plant polyploid-oriented algorithmic and technological (sequencing) advances are necessary. High quality reference subgenomes in polyploid crops in addition to multiple reference genomes or a pangenome per crop species are necessary to capture variation and to better understand these economically important genomes.

Table 2.3: Host-databases of various plant genetic and genomic resources.

DB name	Resources	Plants	URL
Genbank	Genomic	Various plant species	https://www.ncbi.nlm.nih.gov/genbank/
EMBL	Genomic	Various plant species	https://www.ebi.ac.uk/
DDBJ	Genomic	Various plant species	http://www.ddbj.nig.ac.jp/
UniProt	Protein and functional	Various plant species	http://www.uniprot.org/
NCBI	Genomic	Various plant species	https://www.ncbi.nlm.nih.gov/
GOLD	Genomic, metagenomics, transcriptomic	Various plant species	https://gold.jgi.doe.gov/cgi-bin/GOLD/bin/gold.cgi

Continued on next page

Table 2.3 – *Continued from previous page*

DB name	**Resources**	**Plants**	**URL**
Phytozome	Genomic	92 assembled and annotated plant species	`https://phytozome.jgi.doe.gov/pz/portal.html`
Plantgdb	Genomic, transcriptomic	27 assembled and annotated plant species	`http://www.plantgdb.org/`
Sol	Genomic	11 Solanaceae species	`https://solgenomics.net/`
Gramene	Genomic, genetic markers, QTLs	53 plant species	`http://www.gramene.org/`
MaizeGCB	Genomic, annotations, tool host	*Zea mays*	`https://www.maizegdb.org/`
Tair	Genetic and molecular biology data	*Arabidopsis thaliana*	https://www.arabidopsis.org/
CottonGEN	Genomic, Genetic and breeding resources	49 *Gossypium* species	`https://www.arabidopsis.org/`

Continued on next page

Table 2.3 – *Continued from previous page*

DB name	**Resources**	**Plants**	**URL**
PLEXdb	Gene expression	14 plant species	`http://www.plexdb.org/`
RicePro	Gene expression	*Oryza sativa*	`http://ricexpro.dna.affrc.go.jp/`
CerealsDB	Genetic markers	*Triticum aestivum*	`http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php`
PeanutBase	Genome, MAS, QTLs, Germplasm	*Arachis hypogaea*	`https://peanutbase.org/`
SoyKb	Genetic markers, genomic resources	*Glycine max*	http://soykb.org/

Continued on next page

Table 2.3 – *Continued from previous page*

DB name	**Resources**	**Plants**	**URL**
SoyBase	Genetic markers, QTLs, genomic resources	*G. max*	`https://soybase.org/`
PGDBj	Genetic markers, QTLs, genomic resources	80 plant species	`http://pgdbj.jp/`
SNP-Seek	Genotype, Phenotype and Variety information	*O. sativa*	`http://snp-seek.irri.org/`
GrainGenes	Genome, Genetic markers, QTLs, genomic resources	*T. aestivum*, *Hordeum vulgare*, *Secale cereale*, *Avena sativa* etc	`https://wheat.pw.usda.gov/GG3/`
ASRP	small RNA	*A. thaliana*	`http://asrp.danforthcenter.org/`

Continued on next page

Table 2.3 – *Continued from previous page*

DB name	**Resources**	**Plants**	**URL**
CSRDB	small RNA	*Z. mays*	`http://sundarlab.ucdavis.edu/smrnas/`
BrassicaInfo	Genomic	7 *Brassica* species	`http://brassica.info/`
BRAD	Genomics, Genetic Markers and Maps	*Brassica*	`http://brassicadb.org/brad/`
Ensembl Plants	Genomic	45 plant species	`http://plants.ensembl.org/index.html`
Ipomoea Genome Hub	Genomic, EST	*Ipomoea batatas*	`https://ipomoea-genome.org/`

Continued on next page

Table 2.3 – *Continued from previous page*

DB name	**Resources**	**Plants**	**URL**
PGSC	Genomic, annotation	*S. tuberosum, S.chacoense*	`http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml`
GDR	Genomics, Genetics, breeding	Rosaceae	`https://www.rosaceae.org/analysis/266`
HWG	Genomics, Transcriptomics, Genetic Markers	Forest trees and woody plants	`https://www.hardwoodgenomics.org/`

Bibliography of Chapter 2

Aguiar, D. and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–i360.

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., Himmelbach, A., Balfourier, F., Gutierrez-Gonzalez, J., Hayden, M., Koh, C., Muehlbauer, G., Pasam, R. K., Paux, E., Rigault, P., Tibbits, J., Tiwari, V., Spannagl, M., Lang, D., Gundlach, H., Haberer, G., Mayer, K. F. X., Ormanbekova, D., Prade, V., Šimková, H., Wicker, T., Swarbreck, D., Rimbert, H., Felder, M., Guilhot, N., Kaithakottil, G., Keilwagen, J., Leroy, P., Lux, T., Twardziok, S., Venturini, L., Juhász, A., Abrouk, M., Fischer, I., Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., Arnaud, D., Chalabi, S., Chalhoub, B., Cory, A., Datla, R., Davey, M. W., Jacobs, J., Robinson, S. J., Steuernagel, B., van Ex, F., Wulff, B. B. H., Benhamed, M., Bendahmane, A., Concia, L., Latrasse, D., Bartoš, J., Bellec, A., Berges, H., Doležel, J., Frenkel, Z., Gill, B., Korol, A., Letellier, T., Olsen, O.-A., Singh, K., Valárik, M., van der Vossen, E., Vautrin, S., Weining, S., Fahima, T., Glikson, V., Raats, D., Č\'\ihal\'\iková, J., Toegelová, H., Vrána, J., Sourdille, P., Darrier, B., Barabaschi, D., Cattivelli, L., Hernandez, P., Galvez, S., Budak, H., Jones, J. D. G., Witek, K., Yu, G., Small, I., Melonek, J., Zhou, R., Belova, T., Kanyuka, K., King, R., Nilsen, K., Walkowiak, S., Cuthbert, R., Knox, R., Wiebe, K., Xiang, D., Rohde, A., Golds, T., Č\'\ižková, J., Akpinar, B. A., Biyiklioglu, S., Gao, L., N\textquoterightDaiye, A., Kubaláková, M., Šafá\v r, J., Alfama, F., Adam-Blondon, A.-F., Flores, R., Guerche, C., Loaec, M., Quesneville, H., Condie, J., Ens, J., Maclachlan, R., Tan, Y., Alberti, A., Aury, J.-M., Barbe, V., Couloux, A., Cruaud, C., Labadie, K., Mangenot, S., Wincker, P., Kaur, G., Luo, M., Sehgal, S., Chhuneja, P., Gupta, O. P., Jindal, S., Kaur, P., Malik, P., Sharma, P., Yadav, B., Singh, N. K., Khurana, J. P., Chaudhary, C., Khurana, P., Kumar, V., Mahato, A., Mathur, S., Sevanthi, A., Sharma, N., Tomar, R. S., rina Holušová, K., rej Pl\'\ihal, O., Clark, M. D., Heavens, D., Kettleborough, G., Wright, J., Balcárková, B., Hu, Y., Salina,

E., Ravin, N., Skryabin, K., Beletsky, A., Kadnikov, V., Mardanov, A., Nesterov, M., Rakitin, A., Sergeeva, E., Handa, H., Kanamori, H., Katagiri, S., Kobayashi, F., Nasuda, S., Tanaka, T., Wu, J., Cattonaro, F., Jiumeng, M., Kugler, K., Pfeifer, M., Sandve, S., Xun, X., Zhan, B., Batley, J., Bayer, P. E., Edwards, D., Hayashi, S., Tulpová, Z., Visendi, P., Cui, L., Du, X., Feng, K., Nie, X., Tong, W., and Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403).

AU - van Berkum, N. L., AU - Lieberman-Aiden, E., AU - Williams, L., AU - Imakaev, M., AU - Gnirke, A., AU - Mirny, L. A., AU - Dekker, J., and AU - Lander, E. S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *JoVE*, (39):e1869.

Augusto Corrêa dos Santos, R., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics*, 33(16):2575–2576.

Bansal, V. and Bafna, V. (2008). HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159.

Bastiaansen, J. W. M., Coster, A., Calus, M. P. L., van Arendonk, J. A. M., and Bovenhuis, H. (2012). Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution*, 44(1):3.

Bell, J. M., Lau, B. T., Greer, S. U., Wood-Bouwens, C., Xia, L. C., Connolly, I. D., Gephart, M. H., and Ji, H. P. (2017). Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Research*, 45(19):e162–e162.

Bento, M., Gustafson, J. P., Viegas, W., and Silva, M. (2011). Size matters in Triticeae polyploids: larger genomes have higher remodeling. *Genome*, 54(3):175–183.

Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. In *International Conference on Research in Computational Molecular Biology*, pages 18–19. Springer.

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., Araújo, A. C. G., Kozik, A., Do Kim, K., Burow, M. D., Varshney, R. K., Wang, X., Zhang, X., Barkley, N., Guimarães, P. M., Isobe, S., Guo, B., Liao, B., Stalker, H. T., Schmitz, R. J., Scheffler, B. E., Leal-Bertioli, S. C. M., Xun, X., Jackson, S. A., Michelmore, R., and Ozias-Akins, P. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics*, 48(4):438–446.

Besnard, G., Rubio de Casas, R., and Vargas, P. (2007). Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (Olea europaea). *Journal of Biogeography*, 34(4):736–752.

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature*, 543(7645):346–354.

Beyaz, R., Alizadeh, B., Gürel, S., Fatih Özcan, S., and Yildiz, M. (2013). Sugar beet (Beta vulgaris L.) growth at different ploidy levels. *Caryologia*, 66(1):90–95.

Butts, C. T., Bierma, J. C., and Martin, R. W. (2016). Novel proteases from the genome of the carnivorous plant Drosera capensis: Structural prediction and comparative analysis. *Proteins: Structure, Function, and Bioinformatics*, 84(10):1517–1533.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Da Silva, C., Just, J., Falentin, C., Koh, C. S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., Labadie, K., Alberti, A., Charles, M., Arnaud, D., Guo, H., Daviaud, C., Alamery, S., Jabbari, K., Zhao, M., Edger, P. P., Chelaifa, H., Tack, D., Lassalle, G., Mestiri, I., Schnel, N., Le Paslier, M.-C., Fan, G., Renault, V., Bayer, P. E., Golicz, A. A., Manoli, S., Lee, T.-H., Thi, V. H. D., Chalabi, S., Hu, Q., Fan, C., Tollenaere, R., Lu, Y., Battail, C., Shen, J., Sidebottom, C. H. D., Wang, X., Canaguier, A., Chauveau, A., Bérard, A., Deniot, G., Guan, M., Liu, Z., Sun, F., Lim, Y. P., Lyons, E., Town, C. D., Bancroft, I., Wang, X., Meng, J., Ma, J., Pires, J. C., King, G. J., Brunel, D.,

Delourme, R., Renard, M., Aury, J.-M., Adams, K. L., Batley, J., Snowdon, R. J., Tost, J., Edwards, D., Zhou, Y., Hua, W., Sharpe, A. G., Paterson, A. H., Guan, C., and Wincker, P. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, 345(6199):950–953.

Chen, A. and Dubcovsky, J. (2012). Wheat TILLING mutants show that the vernalization gene VRN1 down-regulates the flowering repressor VRN2 in leaves but is not essential for flowering. *PLoS genetics*, 8(12):e1003134–e1003134.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in plant science*, 15(2):57–71.

Cheng, B., Furtado, A., and Henry, R. J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *GigaScience*, 6(11).

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.-C., Magdelenat, G., Gonthier, C., and Others (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell*, 22(6):1686–1701.

Clarindo, W. R., de Carvalho, C. R., Araújo, F. S., de Abreu, I. S., and Otoni, W. C. (2008). Recovering polyploid papaya in vitro regenerants as screened by flow cytometry. *Plant Cell, Tissue and Organ Culture*, 92(2):207–214.

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, 1(2):439–459.

Clevenger, J. P. and Ozias-Akins, P. (2015). SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. *G3: Genes, Genomes, Genetics*, 5(9):1797–1803.

Costa, M.-C. D., Artur, M. A. S., Maia, J., Jonkheer, E., Derks, M. F. L., Nijveen, H., Williams, B., Mundree, S. G., Jiménez-Gómez, J. M., Hesselink, T., and Others (2017).

A footprint of desiccation tolerance in the genome of Xerophyta viscosa. *Nature plants*, 3(4):1–10.

Cousin, A. and Nelson, M. N. (2009). Twinned microspore-derived embryos of canola (Brassica napus L.) are genetically identical. *Plant Cell Reports*, 28(5):831–835.

Crow, K. D. and Wagner, G. P. (2005). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular biology and evolution*, 23(5):887–892.

Darrow, G. M. (1966). *The Strawberry.* Holt, Rinehart and Winston New York.

Dart, S., Kron, P., and Mable, B. K. (2004). Characterizing polyploidy in Arabidopsis lyrata using chromosome counts and flow cytometry. *Canadian Journal of Botany*, 82(2):185–197.

Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, 16(1):260.

Davis, T. M., Denoyes-Rothan, B., and Lerceteau-Köhler, E. (2007). Strawberry. In *Fruits and nuts*, pages 189–205. Springer.

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T. R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, P. F., Schmidt, T., Gabaldón, T., Lehrach, H., Weisshaar, B., and Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*, 505(7484):546–549.

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., and Wendel, J. F. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annual review of genetics*, 42:443–461.

Eaton, T. D., Curley, J., Williamson, R. C., and Jung, G. (2004). Determination of the Level of Variation in Polyploidy among Kentucky Bluegrass Cultivars by Means of Flow Cytometry Research funded by a grant from the United States Golf Association. *Crop Science*, 44:2168–2174.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research*, 27(5):801–812.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? *Trends in Plant Science*, 21(7):609–621.

GRUNDT, H. H., OBERMAYER, R., and BORGEN, L. I. V. (2005). Ploidal levels in the arctic-alpine polyploid Draba lactea (Brassicaceae) and its low-ploid relatives. *Botanical Journal of the Linnean Society*, 147(3):333–347.

Gruzdev, E.V., Beletsky, A.V., Mardanov, A.V., Kochieva, E. and Ravin, N.V. and Skryabin, K. (2018). Genome of Monotropa hypopitys. *Unpublished*.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Harbaugh, D. (2008). Polyploid and Hybrid Origins of Pacific Island Sandalwoods (Santalum, Santalaceae) Inferred from Low-Copy Nuclear and Flow Cytometry Data. *International Journal of Plant Sciences*, 169(5):677–685.

Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., Poveda, L., Shimizu-Inatsugi, R., Baeten, J., Francoijs, K.-J., Nataraja, K. N., Reddy, Y. A. N., Phadnis, S., Ravikumar, R. L., Schlapbach, R., Sreeman, S. M., and Shimizu, K. K. (2017). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Research*, 25(1):39–47.

Hittalmani, S., Mahesh, H. B., Shirke, M. D., Biradar, H., Uday, G., Aruna, Y. R., Lohithaswa, H. C., and Mohanrao, A. (2017). Genome and Transcriptome sequence of Finger millet (Eleusine coracana (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC Genomics*, 18(1):465.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., and Others (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, 43(5):476.

Huang, K., Ritland, K., Guo, S., Shattuck, M., and Li, B. (2014). A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, 14(4):734–744.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., Zhang, L., Niu, X., Zhang, X., Meng, M., and Others (2013). Draft genome of the kiwifruit Actinidia chinensis. *Nature communications*, 4(1):1–9.

Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., Hill, T. A., Jaffe, D., Williams, S. R., Weisenfeld, N., Ramakrishnan, S., Kumar, V., Shah, P., Schatz, M. C., Church, D. M., and Van Deynze, A. (2018). Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Horticulture Research*, 5(1):4.

Huson, D. H., Reinert, K., and Myers, E. W. (2002). The greedy path-merging algorithm for contig scaffolding. *Journal of the ACM (JACM)*, 49(5):603–615.

Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815.

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., Kharbatia, N. M., Rupper, R. R., Sharp, A. R., Dally, N., Boughton, B. A., Woo, Y. H., Gao, G., Schijlen, E. G. W. M., Guo, X., Momin, A. A., Negrão, S., Al-Babili, S., Gehring, C., Roessner, U., Jung, C., Murphy, K., Arold, S. T., Gojobori, T., van der Linden, C. G., van Loo, E. N., Jellen, E. N., Maughan, P. J., and Tester, M. (2017). The genome of Chenopodium quinoa. *Nature*, 542(7641):307–312.

Jiao, W.-B. and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36:64–70.

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., Spillane, C., Robinson, S. J., Links, M. G., Clarke, C., Higgins, E. E., Huebert, T., Sharpe, A. G., and Parkin, I. A. P. (2014). The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, 5(1):3706.

Kim, Y., Oh, Y. J., Han, K. Y., Kim, G. H., Ko, J., and Park, J. (2019). The complete chloroplast genome sequence of Hibiscus syriacus L. 'Mamonde' (Malvaceae). *Mitochondrial DNA Part B*, 4(1):558–559.

Lan, T., Renner, T., Ibarra-Laclette, E., Farr, K. M., Chang, T.-H., Cervantes-Pérez, S. A., Zheng, C., Sankoff, D., Tang, H., Purbojati, R. W., and Others (2017). Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proceedings of the National Academy of Sciences*, 114(22):E4435—-E4441.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., Ma, Z., Shang, H., Ma, X., Wu, J., Liang, X., Huang, G., Percy, R. G., Liu, K., Yang, W., Chen, W., Du, X., Shi, C., Yuan, Y., Ye, W., Liu, X., Zhang, X., Liu, W., Wei, H., Wei, S., Huang, G., Zhang, X., Zhu, S., Zhang, H., Sun, F., Wang, X., Liang, J., Wang, J., He, Q., Huang, L., Wang, J., Cui, J., Song, G., Wang, K., Xu, X., Yu, J. Z., Zhu, Y., and Yu, S. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology*, 33(5):524–530.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293.

Lischer, H. E. L. and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18(1):474.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., and L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448.

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., Hernandez, A. G., Mikel, M. A., Soifer, I., Barad, O., and Buckler,

E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, 6(1):6914.

Margarido, G. R. A. and Heckerman, D. (2015). ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS computational biology*, 11(4):e1004229–e1004229.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564.

Meyers, L. A. and Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206.

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., Loudet, O., Weigel, D., and Ecker, J. R. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, 9(1):541.

Michael, T. P. and VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, 24:71–81.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M.-L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.-K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M.-C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., DePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang,

L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, 452(7190):991–996.

Mithani, A., Belfield, E. J., Brown, C., Jiang, C., Leach, L. J., and Harberd, N. P. (2013). HANDS: a tool for genome-wide discovery of subgenome- specific base-identity in polyploids. *BMC Genomics*, 14(1):653.

Monat, C., Schreiber, M., Stein, N., and Mascher, M. (2019). Prospects of pan-genomics in barley. *Theoretical and Applied Genetics*, 132(3):785–796.

Mondal, T. K., Rawal, H. C., Gaikwad, K., Sharma, T. R., and Singh, N. K. (2017). First de novo draft genome sequence of Oryza coarctata, the only halophytic species in the genus Oryza. *F1000Research*, 6:1750.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., and Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013.

Motazedi, E., de Ridder, D., Finkers, R., and Maliepaard, C. (2017). TriPoly: a haplotype estimation approach for polyploids using sequencing data of related individuals. *bioRxiv*.

Narzisi, G. and Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS one*, 6(4):e19175–e19175.

Oryza Chr3 Short Arm Comparative Sequencing Project (2014). Genome sequencing of Oryza minuta. Technical report.

Parkin, I. A. P., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Pires, J. C., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B., and Sharpe, A. G. (2014). Transcriptome and

methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology*, 15(6):R77.

Parris, J. K., Ranney, T. G., Knap, H. T., and Baird, W. V. (2010). Ploidy levels, relative genome sizes, and base pair composition in Magnolia. *Journal of the American Society for Horticultural Science*, 135(6):533–547.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., and Hvidsten, T. R. (2014). Consortium IWGS, Mayer KFX, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, 345:1250091.

PGSC (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4):354–366.

Porubsky, D., Garg, S., Sanders, A. D., Korbel, J. O., Guryev, V., Lansdorp, P. M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications*, 8(1):1293.

Qiu, J. (2017). The Echinochloa crus-galli whole genome shotgun (WGS) project. *Unpublished*.

Ramsey, J. and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual review of ecology and systematics*, 29(1):467–501.

Riaño-Pachón, D. M. and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research*, 6:861.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Sasaki, T. and Project, I. R. G. S. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052):793–800.

Sato, S., Hirakawa, H., Isobe, S., Fukai, E., Watanabe, A., Kato, M., Kawashima, K., Minami, C., Muraki, A., Nakazaki, N., Takahashi, C., Nakayama, S., Kishida, Y., Kohara, M., Yamada, M., Tsuruoka, H., Sasamoto, S., Tabata, S., Aizu, T., Toyoda, A., Shin-i, T., Minakuchi, Y., Kohara, Y., Fujiyama, A., Tsuchimoto, S., Kajiyama, S., Makigano, E., Ohmido, N., Shibagaki, N., Cartagena, J. A., Wada, N., Kohinata, T., Atefeh, A., Yuasa, S., Matsunaga, S., and Fukui, K. (2010). Sequence Analysis of the Genome of an Oil-Bearing Tree, Jatropha curcas L. *DNA Research*, 18(1):65–76.

Schmid, R., Schuster, S., Steel, M., and Huson, D. (2006). Readsim-a simulator for sanger and 454 sequencing. *View Article PubMed/NCBI*.

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., Bolger, M. E., Alseekh, S., Maß, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., Aury, J.-M., Koren, S., Fernie, A. R., Zamir, D., Bolger, A. M., and Usadel, B. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell*, 29(10):2336–2348.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., and Others (2010). Genome sequence of the palaeopolyploid soybean. *nature*, 463(7278):178–183.

Schmutz, J., Jenkins, J., Grimwood, J., Bertioli, D., Leal-Bertioli, S., Clevenger, J., Michelmore, R., Froenke, L., , Cannon, S.B., Varshney, R., Schleffler, B., Jackson, S., and Ozias-Akins, P. (2018). Genome sequence of Arachis hypogaea, cultivar Tifrunner. *Unpublished*.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346.

Shen, Q., Zhang, L., Liao, Z., Wang, S., Yan, T., Shi, P., Liu, M., Fu, X., Pan, Q., Wang, Y., and Others (2018). The genome of Artemisia annua provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Molecular plant*, 11(6):776–788.

Shi, J. (2018). Chromosome conformation capture resolved near complete genome assembly of proso millet (Panicum miliaceum L.). *Unpublished*.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J.-M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, 43(2):109–116.

Sierro, N., Battey, J. N. D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M. C., and Ivanov, N. V. (2013). Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biology*, 14(6):R60.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany*, 103(7):1146–1166.

Tanaka, H., Hirakawa, H., Kosugi, S., Nakayama, S., Ono, A., Watanabe, A., Hashiguchi, M., Gondo, T., Ishigaki, G., Muguerza, M., Shimizu, K., Sawamura, N., Inoue, T., Shigeki, Y., Ohno, N., Tabata, S., Akashi, R., and Sato, S. (2016). Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research*, 23(2):171–180.

TCP-G. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135.

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F. J., and Others (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences*, 114(44):E9413—-E9422.

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., Fukushima, M., Suzuki, S., Tarora, K., Tamaki, M., Sakamoto, M., Terauchi, R., and Matsumura, H. (2016). Draft genome sequence of bitter gourd (Momordica charantia), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Research*, 24(1):51–58.

Varshney, R. K., Roorkiwal, M., and Nguyen, H. T. (2013). Legume Genomics: From Genomic Resources to Molecular Breeding. *The Plant Genome*, 6(3):plantgenome2013.12.0002in.

Wang, A., Wang, Z., Li, Z., and Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, 34(12):2019–2028.

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., and Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, 19(1):122.

Xie, S.-Q., Zhang, X.-M., Han, Y. and Ling, P. (2018). No Santalum album genome assembly using oxford nanopore sequencing technologyTitle. *Unpublished*.

Yang, J. (2016). The genome of allopolyploid Brassica juncea and evidence for homoeolog expression dominance of potential agricultural significance. *Unpublished*.

Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., Liu, G., Zheng, J., Sun, Z., Fan, W., and Others (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature plants*, 3(9):696–703.

Yang, X., Ye, C.-Y., Cheng, Z.-M., Tschaplinski, T. J., Wullschleger, S. D., Yin, W., Xia, X., and Tuskan, G. A. (2011). Genomic aspects of research involving polyploid plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 104(3):387–397.

Yin, D. (2018). Genome of an allotetraploid wild peanut Arachis monticola: a de novo assembly. *Unpublished*.

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F. N., Kamoun, S., Krause, J., and Others (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife*, 2:e00731.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvo\vrák, J., and Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research*, 27(5):787–792.

Preface to Chapter 3

As presented in Chapter 2, *de novo* genome assembly of polyploid plant genomes is challenging due to the complexity of the genome to be sequenced, the technologies used and the algorithms for the assembly of the sequence reads. In Chapter 3, a multi-reference-based genome comparison of potato genomes of various ploidy levels is presented. The potato species used represent the old potato taxonomy (eleven landraces and one wild potato genomes). The results give an idea of how different these genomes are compared to the potato reference genome and give more information about their traits. Finally, the importance of the availability of multiple reference genomes for diversity exploration is highlighted. These results are important for the exploration of the potato genome. This manuscript is currently in review for Theoretical and Applied Genetics. Since there was a limit on the number of figures and tables for this journal, we selected a number of them for the body of the manuscript, the rest of them are found in the Appendices and for this reason, this section is extended. The genomes have been sequenced using Illumina PE technology and the resulting data were used for the analysis of this chapter. The genomes of S. stenotomum subsp. goniocalyx 1 and S. tuberosum subsp. andigena 1 have been sequenced with long PacBio and linked 10X Genomics technologies, in addition to Illumina PE compared to the rest of the genomes. However, the long and linked data were not used for the analysis in this chapter.

Structural genome analysis in cultivated potato taxa

Maria Kyriakidou[1], Sai Reddy Achakkagari[1], Jose Hector Galvez Lopez[1+], Xinyi Zhu[1+], Chen Yu Tang[1+], Helen H. Tai[2], Noelle L. Anglin[3], David Ellis[3], Martina V. Stromvik[1*] [1]

Department of Plant Science, McGill University, Montreal, Canada

[2] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, Canada

[3] International Potato Center, Lima, Peru

[+] Authors contributed equally

[*] Correspondence:

Corresponding Author

martina.stromvik@mcgill.ca

3.1 Abstract

Polyploidy or duplication of an entire genome occurs in the majority of angiosperms. The understanding of polyploid genomes is important for the improvement of those crops, which humans rely on for sustenance and basic nutrition. As climate change continues to pose a potential threat to agricultural production, there will increasingly be a demand for plant cultivars that can resist biotic and abiotic stresses and also provide needed and improved nutrition. In the past decade, Next Generation Sequencing (NGS) has fundamentally changed the genomics landscape by providing tools for the exploration of polyploid genomes. Here, we review the challenges of the assembly of polyploid plant genomes, and also present recent advances in genomic resources and functional tools in molecular genetics and breeding. As genomes of diploid and less heterozygous progenitor species are increasingly available, we discuss the lack of complexity of these currently available reference genomes as they relate to polyploid crops. Finally, we review recent approaches

of haplotyping by phasing and the impact of third generation technologies on polyploid plant genome assembly.

3.2 Introduction

Cultivated potato (*Solanum tuberosum* L.) originated in the Andean highlands of southern Peru. Whereas potato was not cultivated in Europe and other parts of the world until the 16th century, archeological evidence suggests that the potato has been used for human consumption in Peru for at least 10,000 years (Engel, 1970) . Since ancient times, potato has been adopted into the human diet and is today the third most important food crop for direct human consumption globally (FAO, 2013) .

This worldwide success of potato as a crop is in part due to the tubers being highly nutritious and providing a good source of fiber, minerals, proteins and vitamins C and B6. Important in the adoption of potato as a human food is its wide adaptability to varying environmental conditions and climates – it is grown from the Americas, to Africa, Eurasia and Oceania, and in a broad range of conditions (Bradeen et al., 2011) . However, genetic improvement of existing cultivars is necessary to meet the global food and nutritional demands from a changing climate and the growing human population. The great diversity in potato species and landraces, in particular the South American potato taxa, which are a hugely rich source of valuable agronomic traits, offers insights into the genetic diversity behind the adaptability of the common cultivated potato. Insights into the genomic variation of the diversity of cultivated potato taxa is crucial to crop improvement to help combat future famines and ensure food security.

A significant amount of baseline work has previously been done to aid the advance of potato genomics, reviewed by (Gálvez Helen H., Barkley, Noelle A., Gardner, Kyle, Ellis, David, Strömvik, Martina V., José Héctor, 2017) . The first publicly available potato reference genome was derived from a doubled monoploid clone of *S. tuberosum* group Phureja (DM1-3), which was sequenced and assembled by the Potato Genome Sequencing Consortium (PGSC, 2011) . The DM1-3 genome assembly consists of 12 pseudomolecules

with a total assembly length of 844 Mb. DM1-3 was soon followed by the reference genome of a *S. chacoense* clone, M6 (Leisner et al., 2018) . Additionally, a gene expression atlas of 32 developmental and stress conditions of DM1-3 is available (Massa et al., 2013, 2011) as are several studies on transcriptomes (Barandalla et al., 2018; Fogelman et al., 2019; Gálvez et al., 2016; Moon et al., 2004) . The availability of the two potato reference genomes, along with expression data, has facilitated genetic profiling of different potato varieties, particularly in the identification of structural variatants such as single nucleotide polymorphisms (SNPs) and larger copy number variations (CNVs). A comparison of 12 monoploid and doubled monoploid clones derived from *S. tuberosum* accessions, to the DM1-3 reference genome, showed great heterogeneity in the genomes and that a large portion of their genomes are affected by CNVs (Hardigan et al., 2016) . Potato genome studies have revealed that CNVs play a major role in developing or contributing to adaptive traits (Hardigan et al., 2016, 2017; Iovene et al., 2013; Pham et al., 2017) . This is in agreement with studies in other crop plants, e.g. the response to stress in *Oryza species* (Bai et al., 2016) ; and disease resistance in maize (Beló et al., 2009) , sorghum (Zheng et al., 2011) and soybean (McHale et al., 2012) . Furthermore, a SNP analysis of six potato cultivars showed that large allelic variation correlated with preferential allele expression, was significantly associated with evolutionary conserved genes (Pham et al., 2017) .

Solanum commersonii is a diploid tuber-bearing wild potato species native to Central and South America, is thought to be the first wild potato collected on a scientific expedition ((Hawkes and Others, 1990) and is phylogenetically distinct from cultivated potato (*S. tuberosum*) (Rodr\'\iguez and Spooner, 2009) . *S. commersonii* has desirable agricultural traits not commonly found in the cultivated potato, such as resistance to root knot nematode, soft rot and blackleg, bacterial and verticillium wilt, Potato virus X, tobacco etch virus, common scab and late blight as well as frost tolerance and good capacity for cold acclimation (Bamberg et al., 1986; Hawkes and Others, 1990; Micheletto et al., 2000) . Breeders have overcome the sexual incompatibility of *S. commersonii* and *S. tuberosum* (Johnston and Hanneman, 1980) yet unfortunately with no significant new varieties have

yet to be released (Bamberg et al., 1986; Cardi et al., 1993; Carputo et al., 1997) . The 2015 genome assembly of *S. commersonii* consists of 830 Mb, with 39,290 protein-coding genes, including 126 cold-related genes without orthologs in *S. tuberosum* (Aversano et al., 2015) . The heterozygosity in *S. commersonii* reaches 1.5% based on aligning the raw reads to its genome assembly and estimating the heterozygosity by estimating the total number of heterozygous calls over the total number of callable reads (Aversano et al., 2015) , in contrast to the *S. tuberosum*, where the present heterozygosity was estimating with only 6,373 SNP markers measured against the DM1-3 and resulted in a measure of 53 – 59% heterozygosity (Hirsch et al., 2014a) .

S. chacoense is another closely related tuber-bearing wild species with desirable breeding traits – e.g. disease resistance and resistance to cold-induced sweetening (Leisner et al., 2018) . Its high levels of toxic steroidal glycoalkaloids in the tubers, however, is a great disadvantage and further breeding is required to reduce the glycoalkaloid levels (McCue, 2009) . The inbred M6 *S. chacoense* clone, developed in 2014 (Jansky et al., 2014) , is highly heterozygous and is associated with important agronomic traits like high dry mater, good chip-processing qualities and disease resistance. M6 has also been sequenced and assembled (Leisner et al., 2018) resulting in a genome assembly of 825 Mb, of which, 508 Mb has been anchored into 12 pseudomolecules with an estimated 37,740 genes.

In the present study, we carried out comparisons of 12 potato genomes, of which 10 represent native Peruvian landraces, one a wild species and another one represents a native Chilean landrace. The *S. chacoense* M6 clone and *S. commersonii* public genomes (Aversano et al., 2015; Leisner et al., 2018) were included in the study to explore and identify important potential agronomic traits for the future of potato from closely related tuber-bearing potato species. All genomes were compared to the DM1-3 and *S. chacoense* M6 clone to highlight the variation in our 11 landraces and one close wild relative genome.

Significant work has been previously done to show the CNV – impact on potato (Hardigan et al., 2016) . The current study provides further evidence for the importance of CNVs underlying the potato genome sequence and advances the genome comparison to those that do not belong only to the Phureja and Stenotomum groups, including ploidy levels

(2X, 3X, 4X and 5X). Moreover, since some of the species analyzed are sexually compatible with the reference genomes and important traits can therefore be transferred to the cultivated potato through introgression, this study is also interesting to breeders and growers. Finally, this is the first report investigating structural variation and polymorphism in potato using more than one reference genome.

3.3 Materials and Methods

3.3.1 Plant Materials and Sequencing

The germplasm of eleven Peruvian potato accessions and one Chilean accession (TBR), namely *S. stenotomum* subsp *goniocalyx* (GON1 - CIP 702472 DOI: 10.18730/9DM*), *S. stenotomum* subsp *goniocalyx* (GON2 - CIP 704393 DOI: 10.18730/AGC$), *S. phureja* (PHU - CIP 703654 DOI: 10.18730/9W7J), *S. xajanhuiri* (AJH - CIP 703810 DOI: 10.18730/A0J9), *Solanum stenotomum* subsp. *stenotomum* (STN - CIP 705834 DOI: 10.18730/BTDA), *S. bukasovii* (BUK - CIP 761748 DOI: 10.18730/E3AC), *S. tuberosum* subsp. *andigena* (ADG1 - CIP 700921 DOI: 10.18730/91RP), *S. tuberosum* subsp. *andigena* (ADG2 - CIP 702853 DOI: 10.18730/9GB8), *Solanum curtilobum* (CUR – CIP 702937 DOI: 10.18730/9H1Y), *S. tuberosum* subsp. *tuberosum* (TBR – CIP 705053 DOI: 10.18730/B3MN), *S. juzepczukii* (JUZ – CIP 706050 DOI: 10.18730/C09D) and *S. chaucha* (CHA – CIP 707129 DOI: 10.18730/CS5*), are part of the in vitro potato collection at the International Potato Center (CIP) in Lima, Peru. Genomic DNA was extracted from the leaves of the in vitro plants using E.Z.N.A. Plant DNA Kit (Omega Bio-tek, Inc.), following the manufacturer's instructions. The DNA quality assessment was followed by library preparation and DNA sequencing by Novogene™ Corporation (Beijing, China). Genomic DNA libraries were prepared using the TruSeq Library Construction Kit (Illumina, Inc.) following the manufacturer's instructions. After the libraries were size-selected and purified, they were sequenced using an Illumina HiSeq sequencer (Illumina, Inc.) in paired-end mode (2 x 150 bp). The genomes of GON1 and ADG1 was also sequenced with PacBio's Single Molecule RS II system technology (`https://www.pacb.com/`) and with 10X Genomics' GemCode technology

(https://www.10xgenomics.com/). The Illumina paired-end DNA sequencing reads of *S. commersonii* (COM) were obtained from NCBI Sequence Read Archive (SRA) with the SRP050408 identifier, and the Illumina paired-end reads for *S. chacoense* (M6) with the SRP097632 identifier. The data is available in NCBI, under the BioProject PRJNA556263; the SRA accessions for the diploid genomes are SRR10244436 – SRR10244441 and those for the polyploid genomes are SRR10248510 – SRR10248515.

3.3.2 Alignment against the potato reference genome

The two publicly available potato reference genomes DM1-3 (PGSC, 2011) and M6 (Leisner et al., 2018) were used for the detection of Copy Number Variation events (both deletions and duplications) across the 12 accessions. Version 4.04 of the DM1-3 and the v4.1 of M6 reference genomes were retrieved from SpudDB – Potato Genomics Resource database (http://solanaceae.plantbiology.msu.edu/). The pseudomolecules were indexed using BWA MEM v 0.7.17 (Li, 2013) . The sequencing reads were trimmed using Trimmomatic v0.36 (Bolger et al., 2014) using the following parameters: TruSeq3-PE.fa 2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50. The resulting alignments were manipulated using SAMTOOLS v1.9 (Li et al., 2009). Duplicates were marked using Picard v 2.18.9 (Pic, 2019) and only the properly oriented reads were kept for the structural variation (SV) analyses.

3.3.3 Single Nucleotide Polymorphism (SNP) Analysis

SNPs were detected/called from the processed alignments using Freebayes v1.2.0-2 (Garrison and Marth, 2012) with the following criteria: requiring minimum 4x coverage in diploids, 6x coverage in the triploids, 8x coverage in the tetraploids and 10x coverage in the pentaploid genomes. Furthermore, SNPs with mapping quality < 20, MQM < 20, MQMR < 20 and SAF && SAR < 0 were removed. The SNPs were annotated with snpEff tool (Cingolani et al., 2012) .

3.3.4 Copy Number Variation (CNV) Analysis

Genome-wide CNVs were calculated by comparison of median read coverage in 100 bp windows using CNVnator v0.3.3 (Abyzov et al., 2011) . The resulting raw CNV calls were filtered in order to keep only the SVs larger than 1000 bp, with a cutoff p-value of 0.01 and only reads with q0 quality < 0.5. Significant CNVs were annotated with intansv v1.12.0 (Yao, 2018) package in R v3.3.3 (R Core Team, 2018) , using the GFF file with the annotation of the DM1-3 and M6 reference genomes, respectively, to identify which genes were affected by deletions and duplications.

3.3.5 Significantly Enriched Gene Clusters

Genes with 50% or more of the gene body affected by CNVs were compared to the DM1-3 and M6 reference genomes. CNV gene enriched clusters were identified by dividing the two reference genomes into overlapping 200 kb bins with an intermediate step size of 10 kb (Hardigan et al., 2016) . The number of genes affected by CNVs was calculated in each bin using overlapping bins produced by BEDTOOLS v2.26.0 (Quinlan and Hall, 2010) . Significant bins were determined using a minimum threshold based on the mean of all genomic windows with addition to three standard deviations ((Hardigan et al., 2016) . The clusters with the highest number of genes affected by CNVs were further analyzed.

3.3.6 Principal Component Analysis of CNV-status

CNV affected genes as defined above were used for clustering analysis. A tertiary matrix with 39,028 genes compared to the DM1-3 was generated along with the genes affected and not affected by CNVs in each of the twelve genomes (3 for duplications, 2 for deletions and 1 for non-CNV impacted genes). A Principal Component Analysis (PCA) plot was generated using R (R Core Team, 2018) , based on Euclidean distance. Additionally, based on the CNV status of the genes in each of the genomes, two phylogenetic trees were built using PHYLIP v.3.695 (Felsenstein, 1993) using the PARS algorithm, which accepts

multi-state input was used for the construction of the phylogenetic tree.

3.4 Results

3.4.1 Alignment of 12 potato landrace and wild genomes against two reference genomes shows greater overall match with DM1-3 than with M6

To detect structural variation in the genomes of potato landraces from the genebank at the International Potato Center (CIP, 2018) , genomic-DNA was sequenced from a panel of 12 accessions. These accessions were chosen to include representative individuals from each of the seven species, nine taxa, proposed by (Hawkes and Others, 1990) . Six are diploids: *Solanum stenotomum* subsp. *goniocalyx* (GON1), *S. stenotomum* subsp. *goniocalyx* (GON2), *S. phureja* (PHU), *S. xajanhuiri* (AJH), *S. stenotomum* subsp. *stenotomum* (STN), *S. bukasovii* (BUK); two triploids: *S. juzepczukii* (JUZ), *S. chaucha* (CHA); three tetraploids: *S. tuberosum* subsp. *andigenum* (ADG1), *S. tuberosum* subsp. *andigenum* (ADG2), *S. tuberosum* subsp. *tuberosum* (TBR); and one pentaploid: *S. curtilobum* (CUR). The genomic DNA reads from the twelve genomes were aligned against the DM1-3 potato reference genome v.4.04 (Hardigan et al., 2016) and against the pseudomolecules of the *S. chacoense* M6 potato reference genome (Leisner et al., 2018). DNA reads from *S. chacoense* (M6) and *S. commersonii* (retrieved from NCBI SRA: SRP097632 and SRP050408, respectively) were also aligned against the DM1-3, and DNA reads from *S. commersonii* (Aversano et al., 2015) were aligned against M6. The *S. commersonii* genome was not used as a reference as the scaffolds were not long enough. Unaligned, unpaired reads and aligned positions with low quality scores were removed. Overall, more reads aligned with DM1-3 than with M6 (only pseudomolecules were used for the alignment), with the average reference genome covered being 643 Mb and 436 Mb for the DM1-3 and M6 genomes, respectively 3.1.
The genome alignments against DM1-3 and M6 were used for the identification of sequence-

level variations such as SNPs and structural variations, like CNV. The average read depths for each genome ranged from 35.6X (in BUK) up to 50.3X (in GON2). The percentage of the reference genome covered by each of the sequenced genomes is shown in Figure 3 1. The panel of 12 sequenced genomes covered at minimum 604 Mb and 416 Mb of the DM1-3 and the M6 reference genomes, respectively. Within the 604 Mb of the DM1-3 genome covered, there are 37,395 genes (97% of the total number of genes). Between the newly sequenced genomes, an average size of 328 Mb in the diploids and 285 Mb in the polyploids were aligned in common to the DM1-3. When compared to M6 genome, the average genome body alignment was reduced to 119 Mb in the diploids and to 107 Mb in the polyploids.

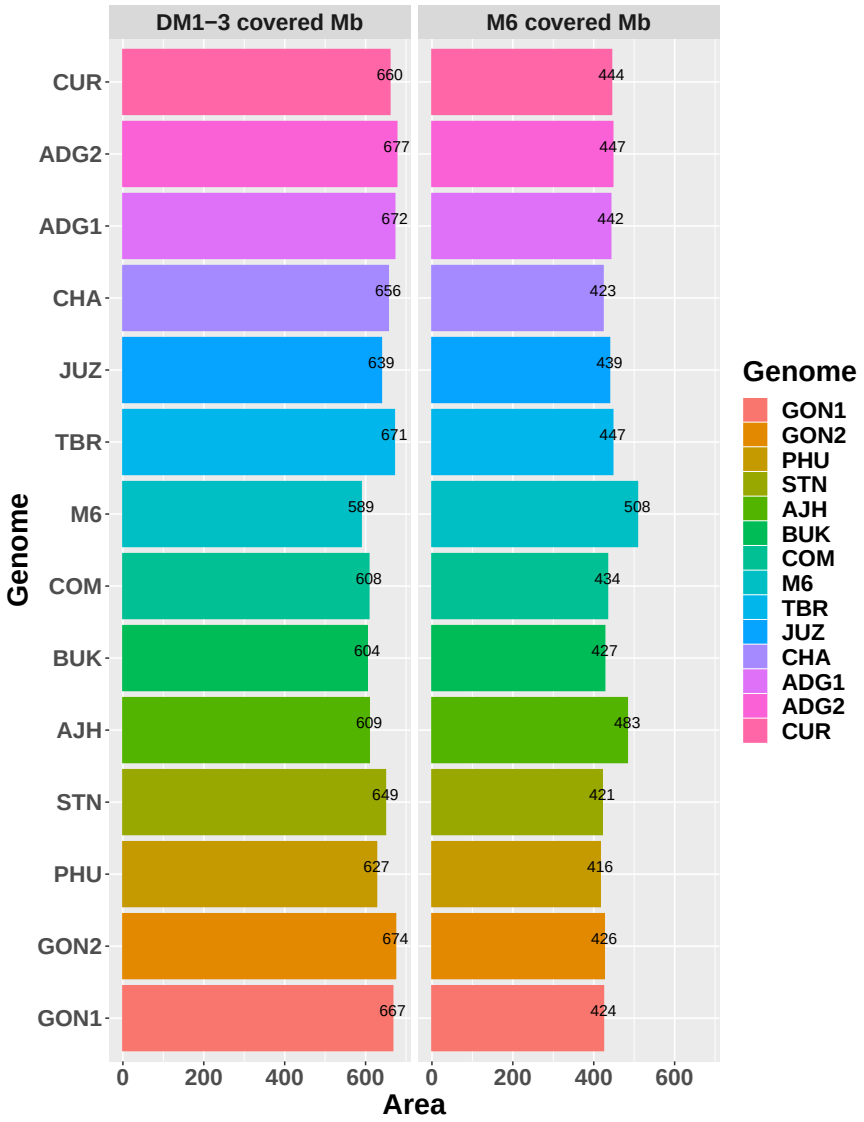


Figure 3.1: Total amount of the reference genomes: DM1-3 (left) and M6 (right) covered by the aligned reads of 14 potato genomes.
The genomes of 12 potato landraces were sequenced and the reads were aligned against the pseudomolecules of two potato reference genomes, DM1-3 (884 Mb) (Hardigan *et al.,* 2016) and M6 (508 Mb) (Leisner *et al.,* 2018) to show the coverage of each. The sequence reads from the published Solanum commersonii (Aversano *et al.,* 2015) were also used in the analysis. GON1 – *S. stenotomum* subsp *goniocalyx*; GON2 - *S. stenotomum* subsp *goniocalyx*; PHU - *S. phureja*; AJH - *S. xajanhuiri*; STN - *S. stenotomum* subsp. *stenotomum*; BUK - *S. bukasovii*; ADG1 - *S. tuberosum* subsp. *andigena*; ADG2 - *S. tuberosum* subsp. *andigena*; CUR - *S. curtilobum*; TBR - *S. tuberosum* subsp. *tuberosum*; JUZ - *S. juzepczukii*; CHA - *S. chaucha*; COM – *S. commersonii*; and M6 – *S. chacoense*.

High levels of CNVs are observed in the 12 sequenced genomes. Some of the regions of CNVs are identical, this, conserved among these genomes, sharing identical regions. The comparison of the diploids to the DM1-3 showed that in the majority of the diploids (with AJH and BUK and in addition to the publicly available COM and M6 genomes being the exceptions), the number of genes impacted by deletions is greater than the number of genes impacted by duplications **Supplementary Figure 8.1** (Appendix 1). Interestingly, in AJH, BUK, COM and M6 the number of deletions is greater than the duplications, but the duplications were larger, and thus impacting a higher number of genes. Additionally, the polyploids also have fewer, but larger duplications resulting in more genes impacted by duplications than by deletions Supplementary **Supplementary Figure 8.1** (Appendix 1). Furthermore, the comparison of the diploids and the polyploids with the M6, showed that the number of deletions and duplications are similar in number, but the duplications were again found to be larger, resulting in more genes impacted by duplications **Supplementary Figure 8.1** (Appendix 1). Not unexpected, the number of genes impacted by duplications is greater in the polyploids than the diploids. In general, both reference genome comparisons show that the majority of the deletions occur in the intergenic regions and thus duplications impact more genes than the deletions (CNVs were more common in the intergenic regions). Finally, there are many more SNPs in the 12 genomes compared to the DM1-3 than compared with the M6, probably because a smaller portion of the M6 genome was available for alignment. Overall, 275 (109 and 166 impacted by duplication and deletion, respectively) CNV impacted genes were in common across the panel of 12 sequenced genomes.

The average size of the genomic regions impacted by CNVs in the diploids is approximately 311 Mb and 314 Mb compared to DM1-3 and M6, respectively. AJH and BUK have the largest CNV-impacted genome region when compared to DM1-3, however when compared to M6, AJH and PHU have the two largest CNV-impacted regions. For the polyploid genomes, an average of 378 Mb and 333 Mb of CNV-impacted regions are observed when compared to DM1-3 and M6, respectively. JUZ had the largest CNV-impacted region when compared to DM1-3, followed by CUR. When compared to M6, CUR has the

largest CNV-impacted region, followed by JUZ.

The average size of the reference genomes that was aligned to the diploids is 328 Mb and 119 Mb when compared to DM1-3 and M6, respectively. Due to the higher heterozygosity of the polyploids, the average aligned reference genome size is 285 Mb and 107 Mb compared to DM1-3 and M6, respectively.

The % heterozygosity of each of the genomes was estimated in percent using the trimmed Illumina reads. As it is shown in Table 3 1, the heterozygosity of the diploids ranges between 1.73 % (in GON2) and 4.48 % (in AJH). The heterozygosity of the polyploids ranges between 3.52 % (in ADG1) and 12.02 % (in CUR) (Table 3 1). This indicates that the higher the ploidy, the higher the heterozygosity and that the heterozygosity is greater outside the Stenotomum and Phureja potato groups.

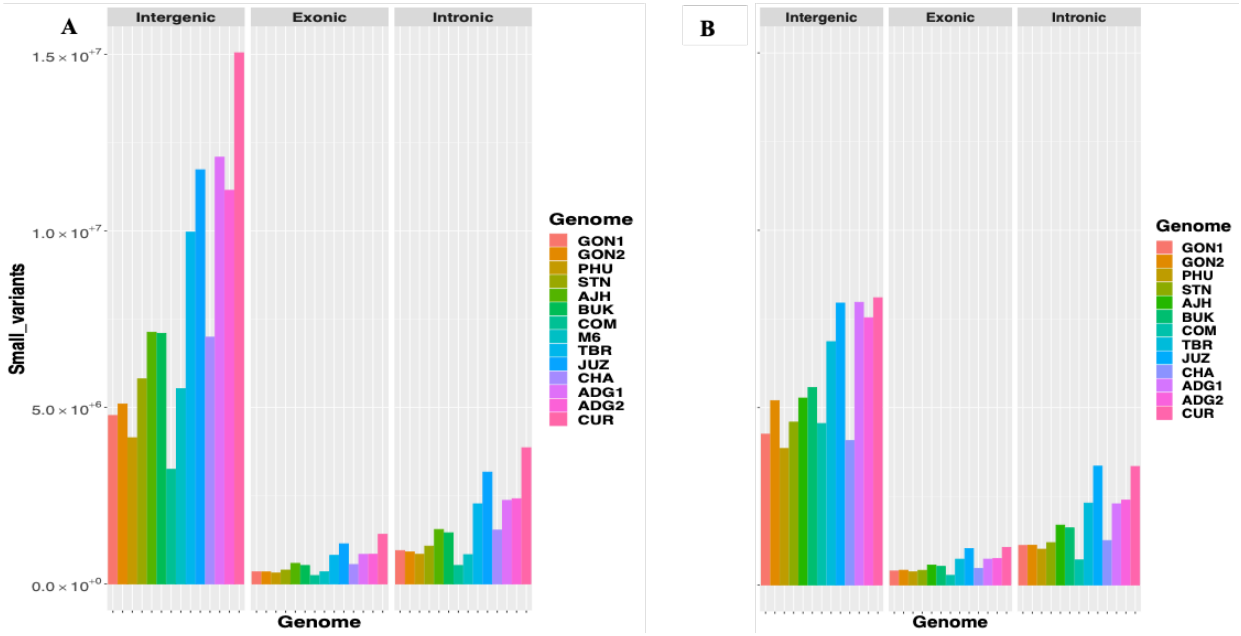


Figure 3.2: Summary of the total number of small variants (SNPs, indels) identified in 13 potato genomes in intergenic, exonic and intronic regions compared to the A) DM1-3 and B) M6 reference genomes.

Overall, more SNPs are present in the intergenic regions of the landrace genomes compared with the both reference genomes (DM1-3 on the left and M6 on the right of the figure). Not surprisingly, there are fewest SNPs in exonic regions, and most SNPs are found in the intergenic region.

3.4.2 Distribution of Single Nucleotide Polymorphisms detected in the genomes compared to the DM1-3 and M6 reference genomes

The number of SNPs detected compared to the DM1-3 genome ranges from 3.8 million in diploid PHU to 12.9 million in the pentaploid CUR genome **Table 3.1**. The largest number of SNPs detected in the diploids was found in BUK - a wild potato genome - with 7 million SNPs. In the triploids, 6.6 million SNPs were detected in CHA and 10.5 million in JUZ, while the number of SNPs detected in the tetraploids ranged between 7.9 million in ADG1 (7.7 million in ADG2) to 7.1 million in TBR. Moreover, in the comparison to M6 the number of SNPs varies between 3.8 million in the diploid PHU up to 8.8 million in the pentaploid CUR. The largest number of SNPs identified in the diploids compared to M6 was 5.6 million in BUK, in the triploids 8.6 million in JUZ and finally in the tetraploids 7.9 million in ADG1. In summary, the number of SNPs varies between 3.8 million to 10.5 million when compared with DM1-3, and between 3.8 million and 8.6 million when compared with M6 **Figure 3.2**.

Table 3.1: Potato genomes sequenced for this study. The table shows their ploidy level and the number of SNPs identified when they were compared to the two reference genomes.

Genome	Ploidy	SNPs VS DM1-3	1 variant per x bases	SNPs VS M6	1 variant per x bases	% Heterozygosity
GON1	2x	4,452,845	133	4,259,520	95	1.75
GON2	2x	4,637,259	126	4,960,736	80	1.73
PHU	2x	3,885,936	152	3,862,547	104	1.84
STN	2x	5,366,637	110	4,607,143	88	2.06
AJH	2x	6,738,160	87	5,503,098	73	4.48
BUK	2x	6,962,470	83	5,695,484	71	3.06
JUZ	3x	10,584,983	48	8,631,219	46	3.52
CHA	3x	6,614,894	83	5,350,001	76	7.75
ADG1	4x	10,488,244	49	7,978,402	50	8.43
ADG2	4x	9,998,123	52	7,763,459	51	7.3
TBR	4x	9,089,933	58	7,188,156	56	3.7
CUR	5x	12,968,439	37	8,873,871	45	12.02

A total of 96,690 and 373,932 small polymorphisms (SNPs and indels) were found in common between the panel of the 12 genomes: diploids and the polyploids, respectively, while 32,959 are shared among all the ploidy levels. From these, about 65% were in the conserved genome, which was not impacted by any CNVs, and the rest of them in the CNV-impacted genome.

The identified SNPs were annotated with snpEff (Cingolani *et al.,* 2012) and **Figure 3.2** shows the total number of small structural variations (SNPs, indels) in the intergenic, exonic and intronic regions, respectively. Based on the results of both reference genome comparisons, the majority of the SNPs are found in the intergenic regions representing 44% of the SNPs (about 22% upstream and 22% downstream). As for mutations, about 51% and 48% of the SNPs consist of missense and silent mutations, respectively, while the remaining 2% were nonsense mutations. The number of indels is smaller than the number of SNPs, with a larger amount of smaller deletions than small insertions in both comparisons.

To identify the most heterozygous regions, biallelic loci were identified in the diploid genomes. Sites that had one or more alternate alleles compared to the reference genome were counted as heterozygous sites. **Supplementary Table 9.1** (Appendix 2) shows that the heterozygosity in the genomes is not spread evenly over the genomes, and that some chromosomes are more heterozygous than others based on alternate allele frequency. The most heterozygous regions in the M6 genome compared to the DM1-3 are found on chromosomes 4, 8 and 9 (Leisner et al., 2018), which was also found in our analysis. This confirms the validity of the pipeline used in the present study (assaying a total of 589 Mb in contrast to the 298 Mb that was previously used). When the landrace genomes are compared to DM1-3, most heterozygous regions are found on chromosomes 1 (an average of 11% heterozygous SNPs) (not in M6) and 4 (an average of 10% heterozygous SNPs), even though some genomes also contained heterozygous regions on chromosomes 3, 6, 8, 9, 10 and 12 **Supplementary Table 9.1** (Appendix 2). Specifically, GON1, GON2 and PHU are highly heterozygous in chromosome 9, while AJH and M6 in chromosome 4. Chromosome 1 is the most heterozygous for the polyploids.

The same approach was also used for the identification of the highly heterozygous regions in the genomes compared to the M6 genome. Chromosomes 1 and 12 are consistently the most heterozygous for all the genomes regardless of ploidy level **Supplementary Table 9.1**. Additionally, GON1, GON2, PHU and CHA were highly heterozygous on chromosome 6, while AJH, ADG1, TBR and CUR in chromosome 5, then BUK and JUZ in chromosome 3, STN and COM in chromosome 11 and finally, ADG2 in chromosome 7 **Supplementary Table 9.1**. The highly heterozygous SNPs (compared to both reference genomes) are found predominantly in the intergenic regions based on the annotation by snpEff (Cingolani et al., 2012).

The majority of the SNPs identified across both the diploid and polyploid genomes against both reference genomes are biallelic, with the largest proportion in the ADG1 and CUR genomes (98%). Moreover, most of the biallelic SNPs are of type B (biallelic sites with at least one reference allele and at least one alternate allele). Type B constituted up to 97% of the biallelic alleles in the ADG1 and CUR genomes.

3.4.3 Distribution of Structural Variations in the landrace genomes compared to the DM1-3 and M6 references shows both polymorphism and synergy

3.4.4.1 Size of the CNVs detected.

The length of the CNVs detected in the genomes, compared to both DM1-3 and M6 reference genomes, varied in size. However, in general, when compared to the M6 genome, the CNVs were larger than those detected against the DM1-3 genome. For the DM1-3, the average median CNV size is larger in the polyploids compared to the diploids **Supplementary Tables 10.1 10.2** . The comparison against the M6 follows a similar pattern, although the size of the CNVs are much larger than the one detected when compared to the DM1-3 **Supplementary Tables 11.1 11.2**.

Duplications are generally larger than deletions for both diploids and polyploids against both the reference genomes. However, the largest CNVs detected in the genomes com-

pared to DM1-3 are deletions, even though in general the duplications tended to be larger **Supplementary Tables 10.1 10.2**. In contrast, when the genomes were compared to M6, the largest CNVs detected are duplications **Supplementary Tables 11.1 11.2**.

3.4.3.2 Significant gene CNV clusters compared to DM1-3 and M6 reference genomes.

To investigate whether large gene clusters were affected with CNVs, the reference genome was split into overlapping bins of 200 kb with a step size of 10 kb, as per (Hardigan *et al.*, 2016). The top three CNV-bins identified per genome **Supplementary Tables 12.1 13.1** are not all the same. They involve both duplications and deletions and generally affected disease resistance genes, including those coding for the nucleotide binding site leucine-rich repeat (NBS-LRR) disease resistance proteins. Other CNV-enriched loci contain genes encoding for auxin-induced SAURs (small auxin-up RNA), endo-1,4-β-mannosidase and genes of unknown function.

3.4.3.3 Significant gene CNV clusters in the diploids compared to DM1-3.

When compared to the DM1-3 reference genome, the CNV-impacted regions in common between the diploid genomes were mostly impacted by deletions (Supplementary Table 3- 6; Appendix 7). Genes coding for proteins of unknown function were found across the regions impacted in common by CNVs. Deletions on chromosome 1 affect genes such as methylketone synthase enzyme, involved in the biosynthesis of the methylketones, produced as plant defense against various herbivorous insects by the trichome glands of wild tomato species (Antonious, 2001; Fridman et al., 2005; WILLIAMS et al., 1980). Additionally, disease resistance genes impacted by deletions are found in chromosomes 4 and 11 **Supplementary Tables 12.1 14.1**. The region in chromosome 4 contains the *R2* gene, responsible for the resistance against the pathogen Phytophthora infestans, (Gebhardt and Valkonen, 2001). A cluster of genes coding for Leucine Rich Repeat (NBS-LRR) disease resistance protein, along with others coding for Tobacco Mosaic Virus (TMV) protein are impacted by deletions in chromosome 11 **Supplementary Table 14.1**. Finally, genes responsible for biotic and abiotic tolerance are impacted by deletions in chromo-

somes 9 and 12 **Supplementary Tables 12.1 14.1**. Within these genes, some of them code for UDP-glycosyltransferase that glycosylate phytormones and metabolites as a response to biotic and abiotic stresses (Rehman et al., 2018). In tobacco they found to play a significant role during the TMV infection (Chong et al., 2002; Le Roy et al., 2016) and resistance against Potato Virus Y (PVY) in tobacco (Matros and Mock, 2004). In chromosome 12, deletions impact genes coding for important immunity proteins, such as ubiquitin conjugating enzyme, RNf5, fiber protein Fb34 and others.

3.4.3.6 Significant gene CNV clusters in the diploids compared to M6.

Similar to the results from the comparison of the diploid genomes to DM1-3, the chromosomes CNV impacted genes in common between all the diploid genomes against the M6 genome are chromosomes 1, 4, 9 and 11 **Supplementary Table 14.1**. Though, the majority of the CNV-impacted genes in common are impacted by duplications. Genes involved in stress tolerance are found to be duplicated in chromosomes 1, 4, and 9 **Supplementary Table 14.1**. In Arabidopsis, Major Facilitator Superfamily (MFS) protein is responsible for drought tolerance (Remy et al., 2013). The gene coding for this protein is duplicated in all the diploids when compared to the M6 reference. Similarly, the DNAJ genes are duplicated in the diploids, suggesting a possible abiotic tolerance as it was previously found to enhance heat tolerance in transgenic tomatoes (Wang et al., 2019) and in pepper, they are involved in growth development, and also induced by heat stress (Fan et al., 2017). Moreover, genes coding for pentatricopeptide repeat proteins (PPR) are duplicated in the diploid genomes. Previously in petunia, it was found that the PPRs have various functions, including the restorage of fertility to cytoplasmatic male sterility (CMS) lines (Bentolila et al., 2002) and in Arabidopsis they are involved in salt and drought stress tolerance (Lv et al., 2014; Zhu et al., 2014, 2012). Duplications in genes coding for serine protease inhibitor (SERPIN) may indicate a defense against insect pests (Jamal et al., 2013). Finally, genes coding for various plant metabolic functions; like 2-Oxogluterate/FE (II) dependent oxygenase proteins (2OGDs) (Kawai et al., 2014) and others involved in auxin

signaling (*SAUR* genes) (Ren and Gray, 2015) are duplicated compared to M6 **Supplementary Table 14.1**.

3.4.3.7 Significant gene CNV clusters in the polyploids compared to DM1-3.

The top CNV-enriched gene clusters in the polyploids also included genes coding for SAURs as well as clusters of genes for tolerance to abiotic stress **Supplementary Tables 13.1**. Moreover, significant CNV gene clusters in common between the polyploid genomes against the DM1-3 genome were identified **Supplementary Tables 15.1**. Within these genes, there were found genes with unknown function, as it was found before in the diploid comparison. Interestingly, significant CNV-gene clusters were found in common between the tetraploid genomes only in chromosome 1 and 9 **Supplementary Tables 15.1**. In the tetraploid genomes, the regions on chromosome 1 coding for S2 self-incompatibility locus 3.2 protein and F-box protein are duplicated. In addition, on chromosome 1 in all the polyploid genomes, genes coding for male sterility proteins are impacted by duplications compared to DM1-3 **Supplementary Tables 15.1**. Genes coding for heat shock protein, verticillium wilt resistance protein, and TMV resistance protein are also duplicated in the polyploids.

3.4.3.8 Significant gene CNV clusters in the polyploids compared to M6

Significantly CNV-enriched gene clusters were detected across all the genomes compared to M6 on chromosomes 1 (64.64 – 64.82 Mb), chromosome 9 (29.23 – 29.46 Mb) and chromosome 11 (0.88 – 1.11 Mb) **Supplementary Figures 17.1 18.1 19.1**. Two of the three regions (those on chromosome 1 and 11; **Supplementary Figures 17.1 19.1** contain *SAUR* gene clusters. The region on chromosome 9 contains 30 genes coding for 2-oxyglutarate (2OG) and Fe (II) dependent oxygenase superfamily **Supplementary Figure 18.1**. All the genomes have at least 21 of these genes duplicated, with almost all of them (29) being duplicated in the pentaploid CUR genome.

3.4.3.9 CNV-based classification of 14 potato genomes

To investigate whether the CNVs have an actual impact on the distance or relatedness of the panel of 12 genomes, M6 and COM, a principal component analysis using the CNV-status (duplicated, deleted or non-affected) genes was performed. **Figure 3.3** captures that three clusters and two outliers are apparent: ADG1, ADG2, PHU, GON1, GON2. STN and CHA cluster close together, M6 and TBR make one cluster and AJH, CUR and JUZ, the bitter potatoes, cluster together, while the two wild species, COM and BUK, are outliers on opposite sides of the graph as expected. Since this largely reflects current taxonomy views, and since a SNP-based phylogenetic analysis was not trivial (because of ploidy and heterozygosity), a phylogenetic analysis was performed with the same CNV-affected gene data as was used for the PCA. **Figure 3.4C** shows the CNV-status-based phylogenetic tree constructed with discrete characters indicating the three statuses of the genes (copy number deleted, duplicated and not impacted). As with the PCA, the GON, PHU, STN and ADG genomes cluster together with CHA close. The BUK and COM are the outliers yet interesting that they map between the bitter genomes (AJH, JUZ, CUR), and the other cultivated taxa.

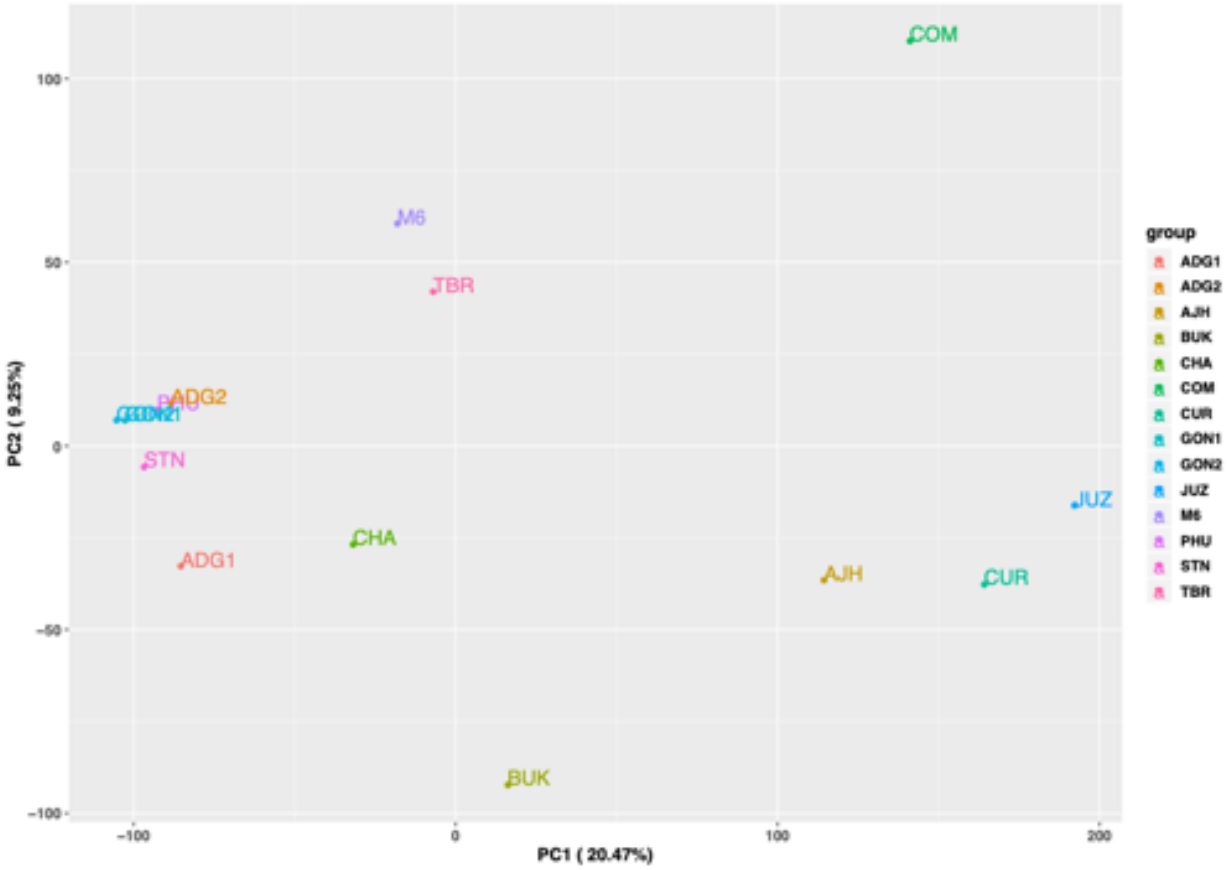


Figure 3.3: Principal Component Analysis (PCA) based on the CNV impacted genes found in the 14 potato genomes compared to the DM1-3 genome, based on Czekanowski genetic distance (also known as Manhattan).

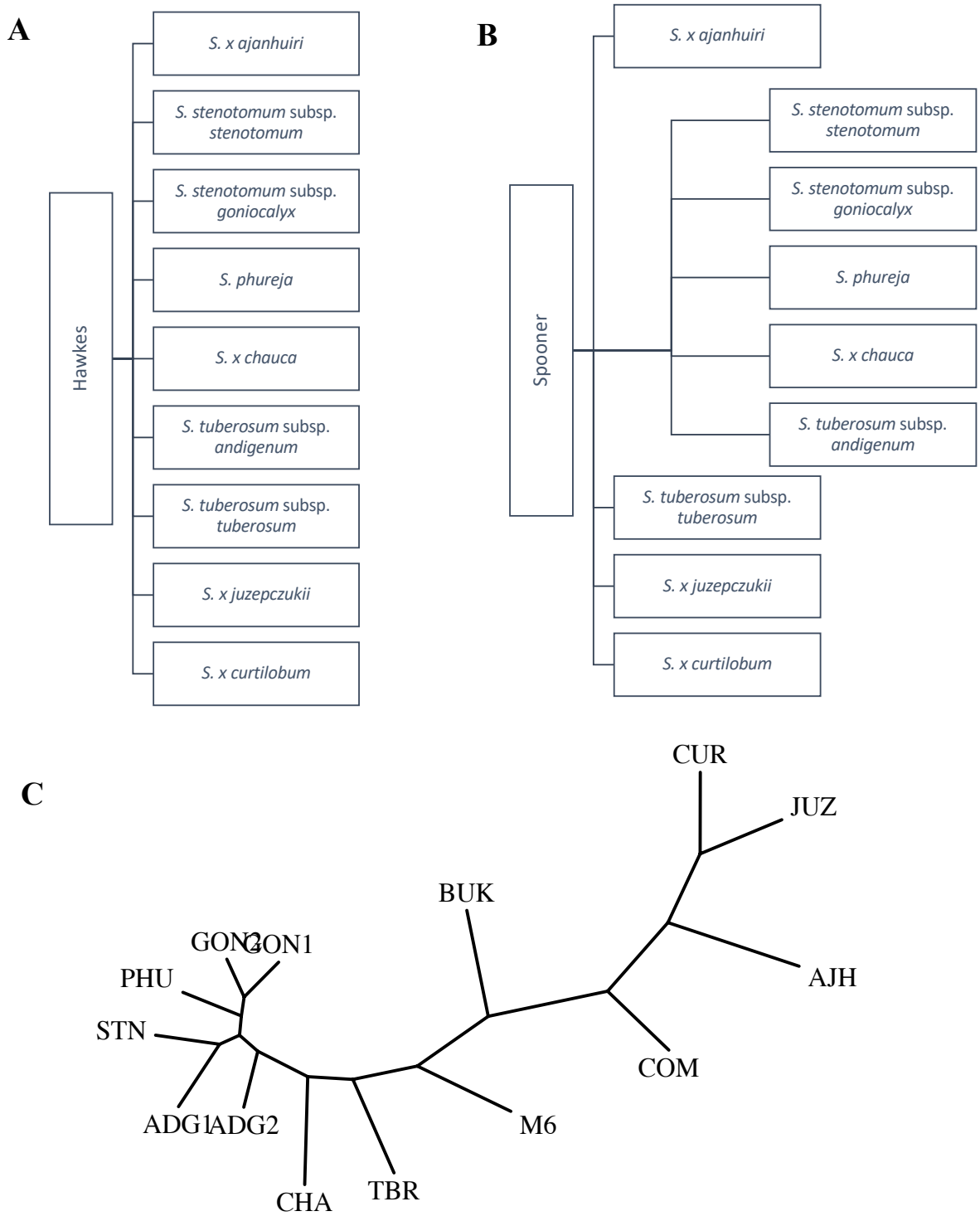


Figure 3.4: Species taxonomy based on A) (Hawkes, 1990) and B) Spooner *et al.* (2007) classifications. C) Shows the genomes' distances based on the CNV-status of the genes (this study, the same data used for the PCA). Similarly, as the PCA plot, we CUR, JUZ and AJH genomes cluster closer and they cluster closer to the wild COM genome compared to the other genomes. Moreover, the other wild genome; BUK is more distant than the other genomes. M6 and TBR genomes are close, while CHA is close to the GON1, GON2, PHU, STN ADG1 and ADG2 cluster.

3.5 Discussion

The results from the current study describe structural variation of 12 potato (*Solanum* sp.) genomes of varying ploidy levels compared with three published reference genomes, DM1-3, M6 and *S. commersonii*. There is a great variation across the landraces and the wild genomes, not captured in the reference assembly. The copy number variation of the genomes was used to classify them.

3.5.1 Comparison of the analysis with previous studies

Overall, the panel of 12 genomes matches better with the DM1-3 reference than with the M6 reference genome. The 12 sequenced genomes from this study have more deletions than duplications when compared to the two reference genomes. However, the duplications span larger regions than do the deletions, which was also previously observed in a double monoploid potato panel (Hardigan et al., 2016). The number of deleted genes is greater than the duplicated ones (with the exceptions of AJH, COM, JUZ and CUR) when compared to the DM1-3. This was also found in a previous study of six autotetraploid cultivated potato genomes (Pham et al., 2017). On the contrary, this was not the case when the genomes are compared to the M6 reference, where either the number of duplicated and deleted genes was similar, or duplicated genes were more numerous. In general, the genomes from the wild species (BUK, COM) and bitter cultivated species (AJH, JUZ, CUR) had more genes impacted by duplications than by deletions. The majority of the CNVs impact intergenic regions and 45%-50% of the CNV-impacted genes are of unknown function. The top CNV clusters include genes related to disease resistance, response to stimuli, and stress tolerance (heat, frost), which are all important traits in breeding programs.

3.5.2 Genome comparisons

Potato taxonomy is a topic of active discussion, and the current study makes no claim of authority on that topic. However, of note is that our analyses partially support both major

schools of thought (Hawkes and Spooner). In our PCA cluster analysis (**Figure 3.3**), the ADG1, ADG2, PHU, GON1, GON2, STN and CHA cluster support the view that these genomes belong be lumped into a single taxon such as a *S. tuberosum* Andigenum group (Spooner et al., 2007). However, in this cluster CHA could be an outlier and could very well be seen as a different species, as has been suggested (Hawkes, 1990). The TBR, classified as its own species, is distant from the ADG, PHU, GON, STN and CHA genomes, which is in agreement with previous literature (Hardigan et al., 2017). The TBR is joined by M6 (*S. chacoense*) in an unusual cluster. After studying wild and cultivated potato, it was previously reported that the genetic distances between cultivated tetraploids and other wild species are smaller than their diploid progenitors due to unequal wild introgression (Hardigan et al., 2017). This might explain the pairing of TBR and M6. Also observed is that the three bitter species AJH, CUR and JUZ are clustered together more distantly from the other genomes. BUK and COM, the two wild species, interestingly do not cluster with each other or anyone else. COM however is closer to the bitter species, than to any of the others.

Based on the PCA results, the AJH, CUR and JUZ genomes form one cluster. It was expected that STN and AJH would be closer together as AJH is considered a hybrid between *S. megistacrolobum* and *S. stenotomum* (Johns et al., 1987) and the cultivated *S. x ajanhuiri* is closer to *S. stenotomum* than to the wild *S x. ajanhuiri* (Johns et al., 1987). However, the AJH accession used in our study clustered with JUZ and CUR, which could be explained by these species being in the bitter potato group and are frost resistant(Johns et al., 1987; Schmiediche et al., 1980). Specifically, JUZ and CUR species are called ”papas amargas” or ”bitter potatoes” and the bitter taste is due to the high concentrations of particular combination of glycoalkaloids (Schmiediche et al., 1980). The JUZ and CUR species are also resistant to some potato cyst-nematode pathotypes (Christiansen, 1977; Dunnett, 1957), and CUR is also resistant to bacterial wilt (Martin and French, 1977). These genomes have similar characteristics and group together, but do not overlap, which is in agreement with both taxonomic treatments (Hawkes and Others, 1990) and (Spooner et al., 2007).

The wild BUK and COM place at opposite ends in the graph, and away from the rest of

the genomes. This is not surprising as COM is phylogenetically distant from cultivated potato (Rodr\'\iguez and Spooner, 2009). BUK on the other hand, is a potential potato landrace progenitor (Hardigan et al., 2015; Spooner et al., 2014; Hosaka, 1995).
TBR and M6 form a cluster in the PCA analysis. Since TBR is a tetraploid, one would have expected it to be closer to the rest of the tetraploids (ADG) along with the other genomes that belong to the Andigenum Group. However, it has been previously reported that the genetic distances between the cultivated tetraploids and the wild species are lower than their diploid progenitors due to unequal introgression (Hardigan et al., 2017).
Using two reference genomes instead of one facilitated the CNV analysis of the remarkable genetic diversity of potato. The CNV-based clustering analyses picture this diversity, the relatedness, and the uniqueness of these genomes. Specifically, the two public genomes COM, and M6 appear in two distinguishable clusters, underlining their differences. They add natural diversity and additional genomic regions, not present in the DM1-3, to the panel, which increases the proof of genetic diversity in this study compared to previous studies on structural variation in potato (Hardigan et al., 2016; Pham et al., 2017).

3.5.3 A SNP analysis uncovers regions of heterozygosity

The whole genome sequence analysis using trimmed reads showed that the genomes inside of the Phureja and Stenotomum groups have the lowest level of heterozygosity (**Table 3.1**), and our whole-genome SNP analysis unraveled an increasing number of variations and greater heterozygosity with increasing ploidy levels, in agreement with previous studies (Hardigan et al., 2017; Hirsch et al., 2013; Pham et al., 2017). The landrace genomes are highly heterozygous and contain specific regions of higher heterozygosity quite unique to the North American doubled monoploids DM1-3. A non-even distribution of heterozygous regions in potato is supported by previous research (Leisner et al., 2018). Additionally, while around 51% of the SNPs cause missense mutations in both comparisons to DM1-3 and M6, 47% are silent and around 1.2% are nonsense mutations. Similar numbers were previously reported (Pham et al., 2017). Also, in the comparison

with the M6 genome, we identified fewer small variations likely due to the fact that the pseudomolecules used in the analysis constitutes only 60% of the genome.

A SNP analysis of six autotetraploid potato cultivars (Atlantic, Kalkaska, Missaukee, Russet Norkotah, Snowden and Superior) identified about 8.4 million SNPs compared to the DM1-3 reference genome (Pham et al., 2017). The number of the SNPs identified in the three newly sequenced tetraploid genomes in our study (TBR, ADG1, ADG2) ranged slightly higher, from 9 to 10.4 million SNPs, probably because the six commercial cultivars are inbred, while TBR, ADG1, ADG2 in the present study are landraces and are therefore more likely to be heterozygous, and because a larger region of the genome was used in our analysis. Additionally, the ADG taxa has the greatest admixture (CIP's marker data unpublished) Furthermore, the SNPs in our diploid genomes ranged between 3.8 up to 6.9 million in the wild BUK genome, while a SNP analysis on a doubled monoploid panel had a lower range, from 0.8 up to 4.7 million (Hardigan et al., 2016)s.

3.5.4 Several CNV-affected gene clusters are common among potato genomes

In the genomes studied, the number of intergenic CNV events was greater than the intragenic ones. This is consistent with previous CNV studies in other organisms. In the human genome for example, it was shown that CNVs are mostly located outside of gene coding sequences and often affect important regulatory elements (Redon et al., 2006). Comparing the genomes to both potato reference genomes, in addition to identifying CNVs affecting functionally annotated genes, many CNV-affected genes were hypothetical or conserved hypothetical proteins. This is a common find based on previous population sequencing studies (Cao et al., 2011; Xu et al., 2011), where it was found that a great number of genes affected by CNVs code for hypothetical or unknown proteins.

3.5.5 *SAUR* gene clusters are affected by CNV events in all genomes studied

The most enriched CNV-impacted gene clusters in all genomes compared were those containing auxin-induced *SAURs* (small auxin-up RNA). These are located on chromosomes 1 (86.97-87.17 Mb), 4 (54.17-54.37 Mb), 6 (56.29-56.49 Mb) and 11 (0.87-1.11 Mb) in the DM1-3 genome and on chromosomes 1 (64.64-64.82 Mb) and 11 (0.88-1.14 Mb) in the M6 genome. In our study, the *SAUR* genes in comparison to both reference genomes were impacted mostly by duplications (i.e. the SAUR genes are duplicated compared to the *SAURs* in the reference genomes). The SAURs are a family of auxin responsive genes that are involved in auxin signaling pathways, regulating a wide range of cellular and developmental processes in plants (Ren and Gray, 2015). Various genomic studies have revealed that SAURs are commonly found in clusters or tandem arrays and that there are 134 *SAURs* in potato, 99 *SAURs* in tomato, 81 *SAURs* in Arabidopsis and 79 in maize (Chen et al., 2014; Hagen and Guilfoyle, 2002; Wu et al., 2012). Interestingly, the study on monoploid potato species also found highly CNV enriched regions on chromosome 1 and 11 containing *SAURs* ((Hardigan et al., 2016). A phylogenetic analysis has revealed that CNVs play an important role in *SAUR* gene family expansion in closely related populations of cultivated potato (Hardigan et al., 2016). *SAURs* are also involved in abiotic stress response and it has been shown that auxin signaling transduction interacts with other stress signaling pathways in rice (Jain and Khurana, 2009).

3.5.6 Disease Resistance gene clusters

Disease resistance genes are another category of genes highly enriched by CNVs compared to both eference genomes. In comparison to DM1-3, all landrace genomes except JUZ have disease resistance genes impacted by deletions on chromosome 4 (4.6-4.8 Mb). This region contains a gene cluster of *R2*, late blight resistance genes (Li et al., 1998), which was directly affected by deletions. Furthermore, the genomes contained CNVs impacting genes coding for nucleotide binding site leucine-rich repeat (NBS - LRR) disease resis-

tance proteins on chromosomes 8 (47.66-47.86 Mb), 11 (42.72-42. 92 Mb) and 12 (0.6-0.8 Mb). The region on chromosome 11 was previously identified and shown to be impacted by CNVs in a panel of 12 doubled-monoploid potato genomes (Hardigan et al., 2016). Regions with disease resistance gene clusters in the 14-genome panel compared to the M6 genome were found on chromosome 1 (0.39-0.59 Mb), 2 (41.24-41.44 Mb) and 5 (0.2-0.22 Mb). These regions contain *NBS - LRR* genes. Disease resistance genes are known to be found in clusters in the genomes of many plant species, hence they are known to undergo rapid evolution as a result of local structural variations (Bergelson et al., 2001) and have been also selected during domestication.

3.5.7 2-Oxogluterate/ Fe (II) dependent oxygenase superfamily proteins (2OGDs)

The cluster of genes coding for 2OGD type proteins was affected by duplication events in the 12 genome panel compared to the M6 reference. The 29.22-29.46 Mb region of chromosome 9 contains a 30-gene cluster coding for 2OGDs. All the potato genomes regardless of ploidy level had at least 11 (GON2, COM) or maximum 21 (JUZ) of these genes impacted by duplications. Proteins in this gene family catalyze various oxidative reactions in plant metabolism, for example DNA repair, biosynthesis of gibberellins (GA), flavonoids, histone demethylation, biosynthesis of plant hormones, and various other metabolites (Kawai et al., 2014). Gibberellins are important for many growth and developmental processes in plants and biosynthesis of GAs include several 2OG dependent reaction steps (Kawai et al., 2014). Flavonoids have diverse functions in plants ranging from plant coloration, protection against UV-B irradiation, nitrogen fixation, and adaptation to environmental conditions during periods of abiotic stresses where the biosynthesis of different flavonoid subclasses are catalyzed by various 2OGDs (Farrow and Facchini, 2014).

3.5.8 Genes involved in metabolite biosynthesis

After comparing the 12-genome panel to the M6 reference genome, multiple regions containing genes impacting metabolite biosynthesis were identified as impacted by CNV events. Other highly enriched regions present on chromosome 1, 3 and 11 contain CNV affected genes that are involved in terpene synthase, C2H2 & C2HC zinc finger family proteins and tetraspanins respectively. The plant terpene synthases are responsible for the synthesis of terpene molecules such as isoprenes (tolerance against heat flecks), monoterpenes, sesquiterpenes, and diterpenes (Chen et al., 2011). C2H2 and C2HC are the zinc finger domains that are reported to be involved with disease resistance (Thomas and Emerson, 2009). Comparative analysis of nine crops revealed zinc finger domains along with NBS-LRR domains in R proteins (Gupta et al., 2012). Tetraspanins are transmembrane proteins that interact with other membrane proteins to form tetraspanin-enriched microdomains, which are involved in various cellular and biological processes that play major roles in pathogenesis and immune response (Wang et al., 2012).

The 32.12 Mb-32.37 Mb region of chromosome 3 contains a 35 gene cluster, of which 34 genes are affected by CNVs in *S. bukasovii* and 28 in *S. stenotomum* subsp. *stenotomum*. GO enrichment analysis has revealed these genes to be involved in the molecular function "transmembrane transporter activity". Similarly, enrichment analysis of CNV affected genes in the 1.24 Mb-1.44 Mb region of chromosome 10, revealed genes associated with endoribonuclease activity and protein binding and that the 40.54 Mb- 40.74 Mb region of chromosome 12 has CNV-affected genes associated with NADH dehydrogenase activity.

3.6 Conclusion

The genomes of a selected set of 12 potato species covering past and current cultivated potato taxa, plus two selected wild species, were studied for structural variation. Similarly, to previous studies in other plants, and potato in particular, genes coding for SAUR, methylketones, mannan endo-1,4 – β – mannosidase, resistance against Phytopthora infestans, NBS-LRR and others of unknown function were found to be impacted by CNVs.

However, unlike previous potato studies, we identified other genes, such as those coding for fiber proteins and those involved in self-incompatibility, to be impacted by CNVs in our panel. Genetic diversity through cross hybridization, polyploidization and speciation makes potato a challenging, but exciting group of species to study. The CNVs represent a source of natural variation that can be tapped for genetic improvement of potato. An important aspect for utilizing CNVs in breeding will be an understanding of the functional impacts of varying copy numbers and an ability to quantify copy numbers with precision and accuracy in high throughput assays. There is increasing availability of resources for detection of CNVs that will facilitate development of applications for selection and breeding.

This study contains a very diverse genome panel that was not used before for the exploration of CNV in the potato genome. Specifically, a previous comprehensive study of CNV in potato (Hardigan et al., 2016) consisted of significant work in the era, although the panel used was not diverse enough to capture the diversity among different potato taxa. In addition, some of the genomes in the present study are sexually compatible with the cultivated species, and so can be used to introduce new desirable traits. Finally, this is the first study in potato exploring CNVs using more than one reference genome. This highlighted the diversity across this panel of potato genomes and identified CNVs in genes implicated in disease resistance and stress tolerance among others.

Bibliography of Chapter 3

(2019). Picard toolkit. \url{http://broadinstitute.github.io/picard/}.

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984.

Antonious, G. F. (2001). PRODUCTION AND QUANTIFICATION OF METHYL KETONES IN WILD TOMATO ACCESSIONS. *Journal of Environmental Science and Health, Part B*, 36(6):835–848.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Dal Molin, A., Avanzato, C., Ferrarini, A., and Others (2015). The Solanum commersonii genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *The Plant Cell*, 27(4):954–968.

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R. A., and Chen, M. (2016). The impact and origin of copy number variations in the Oryza species. *BMC Genomics*, 17(1):261.

Bamberg, J. B., Hanneman, R. E., and Towill, L. E. (1986). Use of activated charcoal to enhance the germination of botanical seeds of potato. *American potato journal*, 63(4):181–189.

Barandalla, L., Álvarez, A., de Galarreta, J. I. R., and Ritter, E. (2018). Identification of candidate genes involved in the response to different abiotic stresses in potato (Solanum tuberosum L.). *Revista Latinoamericana de la Papa*, 22(2):33–38.

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2009). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics*, 120(2):355.

Bentolila, S., Alfonso, A. A., and Hanson, M. R. (2002). A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences*, 99(16):10887–10892.

Bergelson, J., Kreitman, M., Stahl, E. A., and Tian, D. (2001). Evolutionary Dynamics of Plant R-Genes. *Science*, 292(5525):2281–2285.

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

Bradeen, J. M., Haynes, K. G., and Kole, C. (2011). Introduction to potato. *Genetics, Genomics and Breeding of Potatoes. Eds. JM Bradeen, KG Haynes. Enfield, NH: Sci. Publ*, pages 1–19.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–963.

Cardi, T., D'Ambrosio, E., Consoli, D., Puite, K. J., and Ramulu, K. S. (1993). Production of somatic hybrids between frost-tolerant Solanum commersonii and S. tuberosum: characterization of hybrid plants. *Theoretical and Applied Genetics*, 87(1):193–200.

Carputo, D., Barone, A., Cardi, T., Sebastiano, A., Frusciante, L., and Peloquin, S. J. (1997). Endosperm balance number manipulation for direct in vivo germplasm introgression to potato from a sexually isolated relative (Solanum commersonii Dun.). *Proceedings of the National Academy of Sciences*, 94(22):12013–12017.

Chen, F., Tholl, D., Bohlmann, J., and 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(1):212–229.

Chen, Y., Hao, X., and Cao, J. (2014). Small auxin upregulated RNA (SAUR) gene family in maize: Identification, evolution, and its phylogenetic comparison with Arabidopsis, rice, and sorghum. *Journal of Integrative Plant Biology*, 56(2):133–150.

Chong, J., Baltz, R., Schmitt, C., Beffa, R., Fritig, B., and Saindrenan, P. (2002). Downregulation of a Pathogen-Responsive Tobacco UDP-Glc:Phenylpropanoid Glucosyltransferase Reduces Scopoletin Glucoside Accumulation, Enhances Oxidative Stress, and Weakens Virus Resistance. *The Plant Cell*, 14(5):1093–1107.

Christiansen, J. A. (1977). *The utilization of bitter potatoes to improve food production in the high altitude of the tropics*. Cornell University, Jan.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92.

CIP (2018). International Potato Center.

Dunnett, J. M. (1957). Variation in pathogenicity of the potato root eelworm (Heterodera rostochiensis woll.,) and its significance in potato breeding. *Euphytica*, 6(1):77–89.

Engel, F. (1970). Exploration of the Chilca Canyon, Peru. *Current Anthropology*, 11(1):55–58.

Fan, F., Yang, X., Cheng, Y., Kang, Y., and Chai, X. (2017). The DnaJ Gene Family in Pepper (Capsicum annuum L.): Comprehensive Identification, Characterization and Expression Profiles. *Frontiers in Plant Science*, 8:689.

FAO (2013). No Title.

Farrow, S. C. and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Frontiers in Plant Science*, 5:524.

Felsenstein, J. (1993). *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

Fogelman, E., Oren-Shamir, M., Hirschberg, J., Mandolino, G., Parisi, B., Ovadia, R., Tanami, Z., Faigenboim, A., and Ginzberg, I. (2019). Nutritional value of potato (Solanum tuberosum) in hot climates: anthocyanins, carotenoids, and steroidal glycoalkaloids. *Planta*, 249(4):1143–1155.

Fridman, E., Wang, J., Iijima, Y., Froehlich, J. E., Gang, D. R., Ohlrogge, J., and Pichersky, E. (2005). Metabolic, Genomic, and Biochemical Analyses of Glandular Trichomes from the Wild Tomato Species Lycopersicon hirsutum Identify a Key Enzyme in the Biosynthesis of Methylketones. *The Plant Cell*, 17(4):1252–1267.

Gálvez, J. H., Tai, H. H., Lagüe, M., Zebarth, B. J., and Strömvik, M. V. (2016). The nitrogen responsive transcriptome in potato (Solanum tuberosum L.) reveals significant gene regulatory motifs. *Scientific reports*, 6:26090.

Gálvez Helen H., Barkley, Noelle A., Gardner, Kyle, Ellis, David, Strömvik, Martina V., José Héctor, T. (2017). Understanding potato with the help of genomics. *AIMS Agriculture and Food*, 2(1):16–39.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.

Gebhardt, C. and Valkonen, J. P. T. (2001). ORGANIZATION OF GENES CONTROLLING DISEASE RESISTANCE IN THE POTATO GENOME. *Annual Review of Phytopathology*, 39(1):79–102.

Gupta, S. K., Rai, A. K., Kanwar, S. S., and Sharma, T. R. (2012). Comparative analysis of zinc finger proteins involved in plant disease resistance. *PloS one*, 7(8):e42578–e42578.

Hagen, G. and Guilfoyle, T. (2002). Auxin-responsive gene expression: genes, promoters and regulatory factors. *Plant Molecular Biology*, 49(3):373–385.

Hardigan, M. A., Bamberg, J., Buell, C. R., and Douches, D. S. (2015). Taxonomy and Genetic Differentiation among Wild and Cultivated Germplasm of Solanum sect. Petota. *The Plant Genome*, 8.

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., and Others (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, 28(2):388–405.

Hardigan, M. A., Laimbeer, F. P. E., Newton, L., Crisovan, E., Hamilton, J. P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J. C., Douches, D. S., Farré, E. M., Veilleux, R. E., and Buell, C. R. (2017). Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, 114(46):E9999—-E10008.

Hawkes, J. G. and Others (1990). *The potato: evolution, biodiversity and genetic resources.* Belhaven Press.

Hirsch, C. D., Hamilton, J. P., Childs, K. L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C. N., Habermann, M., Neal, B., and Buell, C. R. (2014). Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding. *The Plant Genome*, 7.

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R. E., Jansky, S., Bethke, P., Douches, D. S., and Buell, C. R. (2013). Retrospective View of North American Potato (Solanum tuberosum L.) Breeding in the 20th and 21st Centuries. *G3: Genes, Genomes, Genetics*, 3(6):1003–1013.

Hosaka, K. (1995). Successive domestication and evolution of the Andean potatoes as revealed by chloroplast DNA restriction endonuclease analysis. *Theoretical and Applied Genetics*, 90(3):356–363.

Iovene, M., Zhang, T., Lou, Q., Buell, C. R., and Jiang, J. (2013). Copy number variation in potato – an asexually propagated autotetraploid species. *The Plant Journal*, 75(1):80–89.

Jain, M. and Khurana, J. P. (2009). Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. *The FEBS Journal*, 276(11):3148–3162.

Jamal, F., Pandey, P. K., Singh, D., and Khan, M. Y. (2013). Serine protease inhibitors in plants: nature's arsenal crafted for insect predators. *Phytochemistry Reviews*, 12(1):1–34.

Jansky, S. H., Chung, Y. S., and Kittipadukal, P. (2014). M6: A Diploid Potato Inbred Line for Use in Breeding and Genetics Research. *Journal of Plant Registrations*, 8:195–199.

Johns, T., Huaman, Z., Ochoa, C., and Schmiediche, P. E. (1987). Relationships among Wild, Weed, and Cultivated Potatoes in the Solanum x Ajanhuiri Complex. *Systematic Botany*, 12(4):541–552.

Johnston, S. A. and Hanneman, R. E. (1980). Support of the endosperm balance number hypothesis utilizing some tuber-bearingSolanum species. *American Potato Journal*, 57(1):7–14.

Kawai, Y., Ono, E., and Mizutani, M. (2014). Evolution and diversity of the 2–oxoglutarate-dependent dioxygenase superfamily in plants. *The Plant Journal*, 78(2):328–343.

Le Roy, J., Huss, B., Creach, A., Hawkins, S., and Neutelings, G. (2016). Glycosylation Is a Major Regulator of Phenylpropanoid Availability and Biological Activity in Plants. *Frontiers in Plant Science*, 7:735.

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., and Others (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. *The Plant Journal*, 94(3):562–570.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, X., van Eck, H. J., Rouppe van der Voort, J. N. A. M., Huigen, D.-J., Stam, P., and Jacobsen, E. (1998). Autotetraploids and genetic mapping using common AFLP markers: the R2 allele conferring resistance to Phytophthora infestans mapped on potato chromosome 4. *Theoretical and Applied Genetics*, 96(8):1121–1128.

Lv, H.-X., Huang, C., Guo, G.-Q., and Yang, Z.-N. (2014). Roles of the nuclear-encoded chloroplast SMR domain-containing PPR protein SVR7 in photosynthesis and oxidative stress tolerance in Arabidopsis. *Journal of Plant Biology*, 57(5):291–301.

Martin, C. and French, E. R. (1977). Reaction of some tuberbearing Solanum species to Pseudomonas solanacearum. In *Proc. Amer. Phytopathol. Soc*, volume 4, page 139.

Massa, A. N., Childs, K. L., and Buell, C. R. (2013). Abiotic and Biotic Stress Responses in Solanum tuberosum Group Phureja DM1-3 516 R44 as Measured through Whole Transcriptome Sequencing. *The Plant Genome*, 6.

Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., and Buell, C. R. (2011). The transcriptome of the reference potato genome Solanum tuberosum Group Phureja clone DM1-3 516R44. *PloS one*, 6(10):e26801–e26801.

Matros, A. and Mock, H.-P. (2004). Ectopic Expression of a UDP-Glucose:phenylpropanoid Glucosyltransferase Leads to Increased Resistance of Transgenic Tobacco Plants Against Infection with Potato Virus Y. *Plant and Cell Physiology*, 45(9):1185–1193.

McCue, K. F. (2009). Potato Glycoalkaloids, Past Present and Future.

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., and Stupar, R. M. (2012). Structural Variants in the Soy-

bean Genome Localize to Clusters of Biotic Stress-Response Genes. *Plant Physiology*, 159(4):1295–1308.

Micheletto, S., Boland, R., and Huarte, M. (2000). Argentinian wild diploid Solanum species as sources of quantitative late blight resistance. *Theoretical and Applied Genetics*, 101(5-6):902–906.

Moon, J., Parry, G., and Estelle, M. (2004). The Ubiquitin-Proteasome Pathway and Plant Development. *The Plant Cell*, 16(12):3181–3195.

PGSC (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189.

Pham, G. M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D. S., and Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *The Plant Journal*, 92(4):624–637.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Rehman, H. M., Nawaz, M. A., Shah, Z. H., Ludwig-Müller, J., Chung, G., Ahmad, M. Q., Yang, S. H., and Lee, S. I. (2018). Comparative genomic and transcriptomic analyses of Family-1 UDP glycosyltransferase in three Brassica species and Arabidopsis indicates stress-responsive regulation. *Scientific Reports*, 8(1):1875.

Remy, E., Cabrito, T. R., Baster, P., Batista, R. A., Teixeira, M. C., Friml, J., Sá-Correia, I., and Duque, P. (2013). A Major Facilitator Superfamily Transporter Plays a Dual Role in Polar Auxin Transport and Drought Stress Tolerance in Arabidopsis. *The Plant Cell*, 25(3):901–926.

Ren, H. and Gray, W. M. (2015). SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. *Molecular Plant*, 8(8):1153–1164.

Rodr\'\iguez, F. and Spooner, D. M. (2009). Nitrate reductase phylogeny of potato (Solanum sect. Petota) genomes with emphasis on the origins of the polyploid species. *Systematic Botany*, 34(1):207–219.

Schmiediche, P. E., Hawkes, J. G., and Ochoa, C. M. (1980). Breeding of the cultivated potato species Solanum x juzepczukii Buk. and Solanum x curtilobum Juz. etBuk. *Euphytica*, 29(3):685–704.

Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H., and Gavrilenko, T. (2014). Systematics, Diversity, Genetics, and Evolution of Wild and Cultivated Potatoes. *The Botanical Review*, 80(4):283–383.

Spooner, D. M., Núñez, J., Trujillo, G., del Rosario Herrera, M., Guzmán, F., and Ghislain, M. (2007). Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. *Proceedings of the National Academy of Sciences*, 104(49):19398–19403.

Thomas, J. H. and Emerson, R. O. (2009). Evolution of C2H2-zinc finger genes revisited. *BMC Evolutionary Biology*, 9(1):51.

Wang, F., Vandepoele, K., and Van Lijsebettens, M. (2012). Tetraspanin genes in plants. *Plant Science*, 190:9–15.

Wang, G., Cai, G., Xu, N., Zhang, L., Sun, X., Guan, J., and Meng, Q. (2019). Novel DnaJ Protein Facilitates Thermotolerance of Transgenic Tomatoes. *International Journal of Molecular Sciences*, 20(2):367.

WILLIAMS, W. G., KENNEDY, G. G., YAMAMOTO, R. T., THACKER, J. D., and BORDNER, J. O. N. (1980). 2-Tridecanone: A Naturally Occurring Insecticide from the Wild Tomato Lycopersicon hirsutum f.glabratum. *Science*, 207(4433):888–889.

Wu, J., Liu, S., He, Y., Guan, X., Zhu, X., Cheng, L., Wang, J., and Lu, G. (2012). Genome-wide analysis of SAUR gene family in Solanaceae species. *Gene*, 509(1):38–50.

Xu, J.-H., Bennetzen, J. L., and Messing, J. (2011). Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes. *Molecular Biology and Evolution*, 29(2):861–871.

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M., and Jing, H.-C. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, 12(11):R114.

Zhu, Q., Dugardeyn, J., Zhang, C., Mühlenbock, P., Eastmond, P. J., Valcke, R., De Coninck, B., Öden, S., Karampelias, M., Cammue, B. P. A., Prinsen, E., and Van Der Straeten, D. (2014). The Arabidopsis thaliana RNA Editing Factor SLO2, which Affects the Mitochondrial Electron Transport Chain, Participates in Multiple Stress and Hormone Responses. *Molecular Plant*, 7(2):290–310.

Zhu, Q., Dugardeyn, J., Zhang, C., Takenaka, M., Kühn, K., Craddock, C., Smalle, J., Karampelias, M., Denecke, J., Peters, J., Gerats, T., Brennicke, A., Eastmond, P., Meyer, E. H., and Van Der Straeten, D. (2012). SLO2, a mitochondrial pentatricopeptide repeat protein affecting several RNA editing sites, is required for energy metabolism. *The Plant Journal*, 71(5):836–849.

Preface to Chapter 4

In Chapter 3, the genomes of a selected set of 12 potato species representing past (Hawkes and Others, 1990) and current (Spooner et al., 2014) cultivated potato taxonomy along with two publicly available wild species were studied for structural variation. In general, the duplications are greater than deletions in all 12 genomes. Disease resistance NBS-LRRs, SAURs, and others, are particularly variable in numbers also in the studied set of genomes. Here, in Chapter 4, a deep genome comparison of the eight diploid potato genomes to the DM1-3 reference genome is presented, through a draft genome assembly of S. stenotomum subsp. goniocalyx, and a pan-genome assembly showing the core and accessory genes. Newly predicted coding genes, not found in the reference genome, are also presented. The manuscript is under revision to be resubmitted to Theoretical and Applied Genetics (March 2020).

Constructing a pan-genome using a reference genome of a diploid potato landrace reveals key genes for plant adaptive traits

Maria Kyriakidou[1], Helen H. Tai[2], Noelle L. Anglin[3], David Ellis[3], Martina V. Stromvik[1*]

[1] Department of Plant Science, McGill University, Montreal, Canada

[2] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, Canada

[3] International Potato Center, Lima, Peru

[*] Corresponding Author: martina.stromvik@mcgill.ca

4.1 Abstract

Climate change and the need for higher crop yield to satisfy the growing world population demand cultivars resistant to biotic and abiotic stresses, while also providing needed and improved nutrition. Potato diploids (*Solanum* sp.) are a rich source of genetic diversity, and recent developments in diploid F1 hybrid breeding are promising. Expansion of available genomic resources is however necessary to explore novel traits for potato improvement. The current study used genomic sequencing data from four diploid potato landraces and one wild species together with previously published diploid potato genomes to construct a basic pan-genome for potato. The pan-genome has a total of 28,208 core genes and 11,543 accessory genes for a total of 39,751 genes. This includes 723 newly annotated genes, involved in adaptive traits such as self-incompatibility and defense. The study also presents a first draft genome of a cultivated diploid potato landrace, *S. stenotomum* subsp. *goniocalyx*.

Keywords: *Solanum* sp., pan-genome, self-incompatibility

4.2 Introduction

A reference genome is a digital nucleic acid sequence that contains a single set of chromosomes along with any unanchored heterozygous contigs and/or scaffolds (Kyriakidou et al., 2018). The potato reference genome from a homozygous double-monoploid potato clone (DM1-3) was sequences and assembled by the International Potato Genome Sequencing Consortium (PGSC) (PGSC, 2011). The reference genome has been an important resource in understanding genes underlying potato traits and advancing tools for potato breeding (Gálvez et al., 2016). To date the genomes of two diploid tuber-bearing wild potato species native to Central and South America have also been sequenced: M6, which is an inbred clone of *Solanum chacoense* (Leisner et al., 2018), and *S. commersonii*, (Aversano et al., 2015). Despite great improvement on genetic resources for potato, there is still a need for genomic resources that can represent various taxa, contain the genetic diversity responsible for different agronomic traits, and can enable comparison between different accessions.

There are roughly 4,500 different kinds of cultivatable potatoes with differing agronomic traits (`https://cipotato.org/crops/potato/`). Ancient and wild taxa consist of numerous genes for important traits including nutrient content, resistance to diseases and pests, and tolerance to stress (Bethke et al., 2019). It is impossible to capture this level of diversity in a single reference genome and the information is better retained in a pan-genome (Tettelin et al., 2005). A pan-genome can be divided into two parts - the core genome, which contains the basic biology genes shared by all the individuals investigated, and the accessory genome, which contains genes shared by subsets of individuals and which can encode specialized biochemical pathways and functions related to adaptation (Vernikos et al., 2015). Several pan-genomic analyses have been published for plants including rice (Schatz et al., 2014; Yao et al., 2015; Zhao et al., 2018), soybean (Li et al., 2014), maize (Hirsch et al., 2014a), species of *Brassica oleracea* (Golicz et al., 2016) and sunflower (Hübner et al., 2019). These studies show that the core genome is greater than the accessory one, although the latter is more variable. Consistently, several genes involved

in important traits such as flowering time and disease resistance are found in the accessory genomes but not found in the respective reference genomes.

Native South American potato taxa are highly diverse and are a rich source of genes for potato breeding that are unique compared to the North American and European germplasm (Bethke et al., 2019). Many of the tuber-bearing South American Solanum species are diploids compared to the tetraploid *S. tuberosum*, which is the most widely cultivated species globally. Recently developments in diploid F1 hybrid breeding for potato increase the potential for introgression of diploid Solanum with *S. tuberosum* (Jansky et al., 2016; Lindhout et al., 2011).

The International Potato Center (CIP) in Lima, Peru maintains one of the largest potato germplasm collections in the world, with more than 7,500 accessions of native landraces, wild species, and improved varieties. Previously we have performed a structural variation study on 12 genomes of various ploidy from the CIP collection. In the present study, we have used the four cultivated diploid landraces *S. stenotomum* subsp. *goniocalyx* (2 accessions), *S. phureja*, *S. xajanhuiri*, *S. stenotomum* subsp. *stenotomum*, and the wild species *S. bukasovii* (Hawkes and Others, 1990) to construct the first diploid pan-genome for potato. A core genome was identified, as well as an accessory genome, containing several genes of importance to crop improvement. In addition, we are presenting a draft genome of the diploid *S. stenotomum* subsp. goniocalyx, a first for a cultivated landrace.

4.3 Materials and Methods

4.3.1 Plant materials and genome sequencing

Genomic DNA from germplasm of Potato Research Center (CIP), Lima Peru, accessions *Solanum stenotomum* subsp. *goniocalyx*, (GON1 - CIP702472), *Solanum stenotomum* subsp. *goniocalyx* (GON2 - CIP 704393), *Solanum phureja* (PHU - CIP703654), *Solanum xajanhuiri* (AJH - CIP703810), *Solanum stenotomum* subsp. *stenotomum* (STN - CIP705834) and *Solanum bukasovii* (BUK - CIP761748) was extracted and sequenced as described in (Kyriakidou et al., 2020a). Specifically, genomic DNA was extracted from the leaves of the in vitro

plants using E.Z.N.A. Plant DNA Kit (Omega Bio-tek, Inc.), following the manufacturer's instructions. The DNA quality assessment was followed by library preparation and DNA sequencing by Novogene™ Corporation (Beijing, China). Genomic DNA libraries were prepared using the TruSeq Library Construction Kit (Illumina, Inc.) following the manufacturer's instructions. After the libraries were size-selected and purified, they were sequenced using an Illumina HiSeq sequencer (Illumina, Inc.) in paired-end mode (2 x 150 bp). The sequencing data are available through SRA on NCBI, under the BioProject PRJNA556263.

4.3.2 *De novo* genome assemblies

4.3.2.1 GON1 Genome Assembly

10X Genomics Linked data were assembled using Supernova™ assembler (Weisenfeld et al., 2017) v2.0.0 to produce pseudohap assembly outputs. The first pseudo haplotype assembly was used for the following analysis. Simultaneously, PacBio Long Reads were assembled with the Canu assembler v1.5 (Koren et al., 2017). Tigmint (Jackman et al., 2018) v0.9 was used to correct misassemblies from the PacBio assembly with the aid of 10X Linked Reads. ARCS v1.0.2 (Yeo et al., 2017) was used for assembling the contigs into scaffolds, which were further assembled, and the gaps filled using RAILS v1.4.1 (Warren, 2016). Prior to the scaffolding step, Tigmint v0.9 (Jackman et al., 2018) was used to correct any mis-assemblies, then Arcs v1.0.2 (Yeo et al., 2017) was used for the initial scaffolding step which was followed by the final scaffolding step with RAILS v5.2.22 (Warren, 2016). The gap – filling of the assembly was performed with Cobbler.pl within the RAILS package. As it was previously described. The final genome assembly was evaluated by aligning to the public potato reference genome DM1-3 (PGSC, 2011) using nucmer from MUMmer v4 (Kurtz et al., 2004). The final assembly was used for anchoring pseudomolecules with chromosomer v0.1.3 (Tamazian et al., 2016), based on the DM1-3 reference genome. To evaluate the resulting pseudomolecules, we used BUSCO v3.2.0 (Simão et al., 2015) to observe the gene content. For the best assembly, multiple approaches were

applied by combining different assembly algorithms and data produced (Supplementary Table 21.1), but the conclusion was that this approach was the best for the data quality and for the nature of the GON1 genome. Transposable elements and repeats were detected in the GON1 genome using RepeatModeler v1.0.11 (Smit and Hubley, 2019).

4.3.2.2 Genome assembly of the GON2, PHU, STN, AJH and BUK genomes

The Illumina reads of the genomes (besides those of GON1, COM and M6) were assembled *de novo* using MaSuRCA assembler v 3.2.4 (Zimin et al., 2013) with default settings. From the final assemblies, contigs shorter than 200 bp were removed as per NCBI's standards.

The *de novo* assemblies resulted from both Second and Third Generation technologies were assessed using BUSCO v3.0.2 (Simão et al., 2015) and QUAST v5.0.0 (Gurevich et al., 2013) to estimate the gene distribution and the assembly statistics, respectively for the assemblies of the six genomes: GON1, GON2, PHU, STN, AJH and BUK.

4.3.3 Estimating the percentage of whole genome heterozygosity

The trimmed Illumina sequencing data of the eight genomes (GON1, GON2, PHU, STN, AJH, BUK, COM and M6) were used for the estimation of the % heterozygosity in the genomes. Jellyfish v2.2.10 (Marçais and Kingsford, 2011) was firstly used to compute the histogram of the k-mer frequencies and the final k-mer count histogram per genome was used within the GenomeScope 2.0 online tool (Ranallo-Benavidez et al., 2020). Each of the genome sizes was also estimated.

4.3.4 Pan-Genome construction and annotation

The pan-genome was constructed using the "map-to-pan" approach (Wang et al., 2018). The newly *de novo* assembled genomes, along with the assemblies of *S. commersonii* (COM) (Aversano et al., 2015) and *S. chacoense* (M6) (Leisner et al., 2018), were aligned separately using nucmer aligner against the DM1-3 v 4.04 potato reference genome (Hardigan et al., 2016), along with the chloroplast and mitochondrial genomes of *S. tuberosum* Group

Phureja and *S. tuberosum* Group Tuberosum (retrieved from `http:/solanaceae.plantbiology.msu.edu/pgsc_download.shtml`). The unaligned contigs from all the genomes were extracted and concatenated into a single FASTA file. Any heterozygous contigs present were removed using CD-HIT (Fu et al., 2012; Li and Godzik, 2006) with identity 90%. Any contaminant contigs (matching any other genome than green plants) were identified using BLAST+ v2.7.1 (Camacho et al., 2009) (-qcov_hsp_perc 80 -perc_identity 90) and removed from the rest of the contigs. The unaligned contigs were aligned against the UniVec database (Kitts et al., 2016) to remove other potential contaminants. The final, cleaned, non-redundant contigs were concatenated into a single FASTA file and added to the DM1-3 v4.04 reference genome and all of them consisted of our pan-genome.

MAKER v.2.31.10 (Campbell et al., 2014) was used for the pan-genome annotation and repeat masking. To train the gene models, RNA-Sequencing data from young leaves of GON1, GON2, PHU, STN and AJH were used, along with the *S. lycopersicum* and *S. pennellii* genome annotation. RNA Seq data from BUK was not usable. The RNA-Seq data of the five genomes can be found at GEO database, under the accession GSE137781. The RNA-Seq reads were trimmed and assembled using Trans-ABySS (Robertson et al., 2010). The gene functions were retrieved by downloading the Swissprot database for all green plants (Viridiplantae; retrieved on the 2nd of July 2019). *Ab initio* gene prediction was performed using Augustus (Stanke et al., 2006) and SNAP (Korf, 2004). The "tomato" model was selected for Augustus prediction, and SNAP was trained twice based on the RNA-Seq evidence, according to MAKER's manual. Finally, a set of high confidence, supported by transcript and/or protein evidence gene models were generated by MAKER. **Supplementary Figure 27.1** (Appendix 20) shows in summary the overall pipeline from the de novo genome assembly of the diploid genomes until the pan-genome construction.

4.3.5 Gene presence/absence variation analysis

Gene presence/absence variation was performed by aligning the trimmed PE Illumina reads of all the genomes separately against the pan-genome using BWA MEM (Li, 2013) v0.7.15 and only reads mapping in proper pairs were kept. Furthermore, the coverage

of each of the contigs was obtained and the coordinates with the coverage were kept in a .BED file. Using the pan-genome covered by each genome along with its annotation, we calculated the gene body and CDS coverage for each gene. Finally, the PAV was estimated based on the gene body and CDS coverages. We followed the approach by (Sun et al., 2016); hence genes with gene body coverage $> 80\%$ and CDS coverage $> 95\%$ were considered to be present in the genome.

4.3.6 Phylogenetic analysis

For the phylogenetic analysis, the genetic distances were estimated from PAV (gene status; present or absent from each genome) using PARS program within PHYLIP (`http://evolution.genetics.washington.edu/phylip/programs.html`). The tree was plotted with FigTree v1.4.2 (`http://tree.bio.ed.ac.uk/software/figtree/`).

4.3.7 Data availability

All the raw genome sequencing data are stored in the Sequence Read Archive (SRA) under BioProject PRJNA556263. The list of all the accession numbers is found in **Supplementary Table 20.1**. The final genome assemblies are deposited into NCBI GenBank database under the following Accession Numbers: WBHW00000000, WBHX00000000, WBHY00000000, WBHZ00000000, WBIA00000000 and WBIB00000000. The RNA-Sequencing data used for the pan-genome annotation have been deposited in GEO database, under the accession GSE137781. The data submitted at GEO databases will be available after the manuscript's publication.

4.4 Results

4.4.1 Genome assembly of GON1

A draft genome assembly of the S. stenotomum subsp. goniocalyx genome – GON1 (CIP 702472) was generated using a hybrid technology approach of 10X Genomics Linked and

PacBio Long Reads data (Supplementary 20.1; Appendix 13; sequencing technologies used). The Linked Reads were assembled in order to get a pseudohaplotype genome. The Long Reads were assembled and corrected with the Linked Reads to a final scaffold assembly of 855.80 Mb, consisting of 6,424 scaffolds and an N50 of 326,785 bp (**Table 4.1**). The genome size 892.83 Mb was estimated by a 10X Genomics Chromium library. The completeness of the assembly was evaluated with BUSCO (Simão *et al.,* 2015). The results show that 93.6% of the BUSCO core Plantae ortholog genes are represented in the assembly and another 2.9% are present as partial sequences (C:93.6% [S:82.8%, D:10.8%], F:2.9%, M:3.5%, n:1375).

To evaluate the hybrid assembly of PacBio and 10X Genomics data, it was used as a reference to which short Illumina PE, Linked 10X Genomics and Long PacBio reads of the GON1 genome were aligned. Most of the Illumina reads, 96.94%, aligned to the assembled genome, with 92.30% of the reads being properly paired. Furthermore, 97.02% of the Linked and 95.03% of the Long reads were aligned to the final assembly.

Table 4.1: Genome Assembly metrics of the Solanum tuberosum subsp. goniocalyx - GON1 (CIP 702472) genome.

Quality Metric	GON1 assembly
Total scaffold size	855,795,280 bp
No. of scaffolds	6,424
N50 scaffold size	326,785 bp
NG50 scaffold size	307,273 bp
Maximum length of scaffolds	9,325,854 bp
Minimum length of scaffolds	913 bp
Average scaffold size	133,218 bp
Total anchored scaffold length	468,652,731 bp
Total unanchored scaffold length	387,432,960 bp

4.4.2 Construction of the GON1 pseudomolecules

Pseudomolecules were constructed for GON1 using Chromosomer (Tamazian et al., 2016) to align the GON1 scaffolds to the doubled monoploid reference genome DM1-3 (PGSC, 2011). Based on the results, 469 Mb, or 54.8%, out of the assembled 856 Mb of GON1 were placed in 12 pseudomolecules. Overall, the sizes of the GON1 pseudomolecules are smaller than those of DM1-3, but they are comparable to the M6 pseudomolecules **Supplementary Table 22.1** (Appendix 15). The pseudomolecule construction was evaluated with BUSCO (Simão et al., 2015) and within these 12 newly constructed pseudomolecules (54.8% of the genome), 56.1% of the BUSCO core Plantae ortholog genes were represented as full length and another 5.2% were present as partial sequences (C:56.1% [S:54.1%, D:2.0%], F:5.2%, M:38.7%, n:1375).

Figure 4.1 shows the alignment of the newly constructed pseudomolecule 1 (GON1.v01ch01) against pseudomolecule 1 (ST4.03ch01) of the DM1-3 reference genome, the *S. chacoense* M6 clone, the *S. lycopersicum* (SL v2.4; tomato) (Consortium, T.G., 2012) and *S. pennellii* (Spenn v1) (Bolger et al., 2014). The best alignment is against the ST4.03ch01 (pseudomolecule 1 of DM1-3). In the alignment against the M6v4.1chr01, inversions appear along the pseudomolecule, and in the alignments against chromosome 1 of *S. lycopersicum* and *S. pennellii*, the best alignment is in the beginning (around 0 – 5,000,000 bp) and in the end (around 80,000,000 – 90,000,000 bp). Overall, the alignments of the 12 pseudomolecules of the GON1 genome showed concordance with the 12 chromosomes of DM1-3 and the 12 pseudomolecules of M6. The alignments against the S. lycopersicum and to the close wild relative of tomato, *S. pennellii*, were less in concordance, but still very close. The alignments for the rest of the pseudomolecules are found in the Supplementary Material (Supplementary Figures 28.1 – 38.1; Appendix 21 - 31). Classes of repetitive elements were identified in GON1 through a repeat library generated from the scaffolds with RepeatModeler (Smit and Hubley, 2019) and repeat masked assembly was constructed with RepeatMasker v4.0.7 (Smit et al., 2015). In total, 60.2% of the genome was masked. Supplementary 23.1 (Appendix 16) shows the repeat content of GON1 genome. the majority

of the repeats (28.4%) are unclassified, while LRT elements are most abundant (27.3%) of the element types.

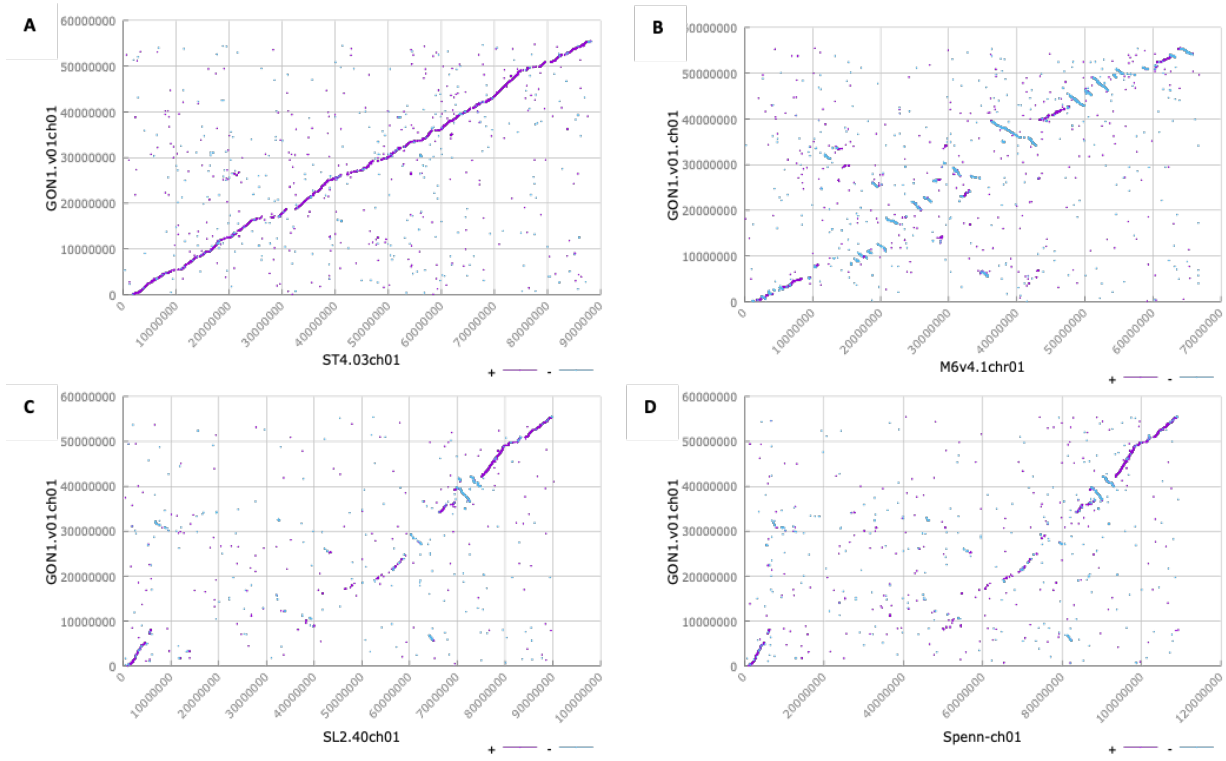


Figure 4.1: Alignment of Chromosome 1 from S. stenotumum subsp. goniocalyx with Chromosome 1 from other *Solanum* sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 01 of Solanum stenotomum subsp goniocalyx - GON1 are shown for alignment lengths of greater than 100 base pairs at greater than 90% sequence similarity against the chromosome 1 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch01), **B.** *S. chacoense* (M6v4.1chr01), **C.** *S. lycopersicum* (SL2.40ch01) and **D.** *S. pennellii* (Spenn-ch01). The purple lines show forward matches, while reverse matches (inversions) are shown in blue. The best match is found in the comparison of the GON1 with the DM1-3. The alignment with *S. chacoense* contains inversions between 10 – 60 Mb. Overall, the alignments between the ST4.03ch01 and *S. chacoense* showed concordance against both chromosomes, even though there are more inversions in the alignment with *S. chacoense*. On the other hand, there is agreement in the alignments between GON1 and *S. lycopersicum* (between 50 – 90Mb) and GON1 and *S. pennellii* (between 60 – 110 Mb) towards the end of the two chromosomes.

4.4.3 Genome assemblies of GON2, PHU, STN AJH and BUK

Five additional genomes - *S. stenotomum* subsp. *goniocalyx* (GON2), *S. phureja* (PHU), *S. xajanhuiri* (AJH), *S. stenotomum* subsp. *stenotomum* (STN), and *S. bukasovii* (BUK) - were sequenced (Illumina PE) and assembled into contigs. Due to the high heterozygosity of the sequenced diploid genomes, the assemblies generated using MaSuRCA v3.1.3 (Zimin et al., 2013), were larger than the reference genome Supplementary Table 24.1 (Appendix 17). CD-HIT (Fu et al., 2012; Li and Godzik, 2006) was therefore used to remove the heterozygous contigs from the assemblies in order to keep only unique contigs. **Table 4.2** shows the quality metrics of the de novo genome assemblies of the additional five sequenced genomes after removing the redundant contigs, along with BUSCO results with the portion of the present gene content. The size of the assembled genomes reached up to 1.6 times the size of the DM1-3 reference genome (Supplementary Table 4.1; AJH). STN (945 Mb) genome has the largest size, while GON2 (771 Mb) the smallest among the five genomes. The GC % content of the genomes as it was expected, it is consistent before and after the removal of the redundant contigs, reaching up to 35.76% (Table 4 2, Supplementary 24.1; Appendix 17).

4.4.4 Comparison of the GON1 against the GON2, PHU, STN, AJH and BUK genome assemblies

The genome assembly of the GON1 potato consists of almost 34 times fewer contigs than the other assemblies, and reached an N50 of 0.32 Mb compared with a maximum N50 of 0.09 Mb for the other genomes (which were assembled with only Illumina PE sequences) (Supplementary Table 20.1; Appendix 13, 4.1). The GON2 genome assembly has the smallest N50, the most contigs, and contains the smallest percentage (%) of BUSCO present genes at 45.7% (Table 4.1). Between the GON2, PHU, STN, AJH and BUK genomes, PHU had the biggest % of BUSCO present genes: 74.1%, while from all the assemblies, GON1 had the most BUSCO present genes; 93.6% (Table 4.2).

Table 4.2: Quality metric of the de novo genome assemblies after removing redundant contigs. * For GON1 genome it refers to the number of the scaffolds.

Quality metric	GON2	PHU	STN	AJH	BUK
# of contigs	284,724	229,366	250,113	239,686	134,715
Contigs > 1000 bp	218,574	184,167	209,411	171,898	121,820
Length of assembly	771 Mb	873 Mb	945 Mb	829 Mb	799 Mb
GC %	34.66	35.19	35.04	35.76	34.61
Largest contig length (Kb)	42	156	164	133.2	183.7
Contig N50	3,929	6,439	5,839	6,475	9,511
Contig L50	59,367	33,133	40,704	30,770	21,379
% BUSCO present genes	45.7	74.1	61.2	59.8	67.5
% BUSCO partial genes	21.6	10.8	13.8	16.7	15.4
% BUSCO duplicated genes	5.1	8.8	10	14.1	16.5

4.4.5 Pan-Genome Construction

The diploid Solanum pan-genome was built using a hybrid assembly approach or "map-to-pan" approach (Wang *et al.*, 2018), including both reference-based and de novo genome assembly approaches (Supplementary Figure 20.1; Appendix 20). In this approach, the contigs from the genome assemblies (GON1, GON2, PHU, STN, AJH, BUK, COM, M6 – 4.5Gb total in length) were first aligned to the DM1-3 reference genome. Those that did not align with the reference genome were concatenated and redundancies (heterozygous contigs), and contaminants (Univec database NCBI, 2016; human and microbial sequences) were removed. The final sequences (37.3Mb in total length) were masked, annotated and added to the DM1-3 reference genome to build the pan-genome. That is, the pan-genome is the sum of the concatenated DM1-3 pseudomolecules and the non-redundant contigs of the eight genomes that did not align with the DM1-3 reference. While the DM1-3 genome contains 39,028 genes (which we consider the reference genes or gene models) and the M6 genome contains 37,740 genes, the diploid potato pan-genome proves to contain 39,751 genes (**Table 4.3**). The number of genes present in the pan-

Table 4.3: Genome size and gene number comparison between the DM1-3, M6 reference genomes and the diploid pan-genome. *Including ST4.03ch00 and ST4.03chUn

	DM1-3 (v4.04)	M6	Diploid Pan-Genome
Genome size (Mb)	884,108,296*	825,767,562	921,447,870
Number of genes	39,028	37,740	39,751

genome may be underestimated due to the lack of diverse RNA-Seq libraries to fully annotate all the genomes.

A presence/absence variation (PAV) analysis was carried out to identify the core and accessory genomes within the pan-genome. The core genome was determined to contain a total number of 28,208 genes (or 71% of the pan-genome) that are present in all of the genomes. The remaining 11,543 genes (29 % of the pan-genome, 10,881 of the DM1-3 reference genes) constitutes the accessory genome, including 723 newly annotated genes that are not present in the DM1-3 reference, 555 genes that are present only in the DM1-3, and an additional 547 genes that are genome specific, i.e. only present in one genome or another (other than DM1-3). Figure 4.2 represents a heatmap of the of the PAV genes, explaining the concept of the pan-genome. The core genes are found at the upper part of the graph while the variable genes are found in the bottom.

A phylogeny analysis based on the PAV resulted in four distinct genome clusters. **Figure 4.2** shows the unrooted phylogenetic tree of the eight genomes used for the pan-genome construction. The figure contains four clusters where the wild *S. chacoense* and *S. commersonii* are clustered together, more distant from the cultivated species. Moreover, *S. bukasovii*, which is another wild potato genome and a potential landrace progenitor consist of another cluster. The bitter *S. xajanhuiri* is also by itself, opposite of *S. bukasovii*. Finally, the cultivated species *S. stenotomum* subsp. *stenotomum*, *S. phureja*, *S. stenotomum* subsp. *goniocalyx* 1 and 2 are clustered together. Modeling the pan-genome size by iteratively randomly sampling genomes, suggests a restricted pan-genome with a finite number of core and pan genes (Supplementary Figure 39.1). While the size of the pan-genome is increasing, the core genome size is decreasing.

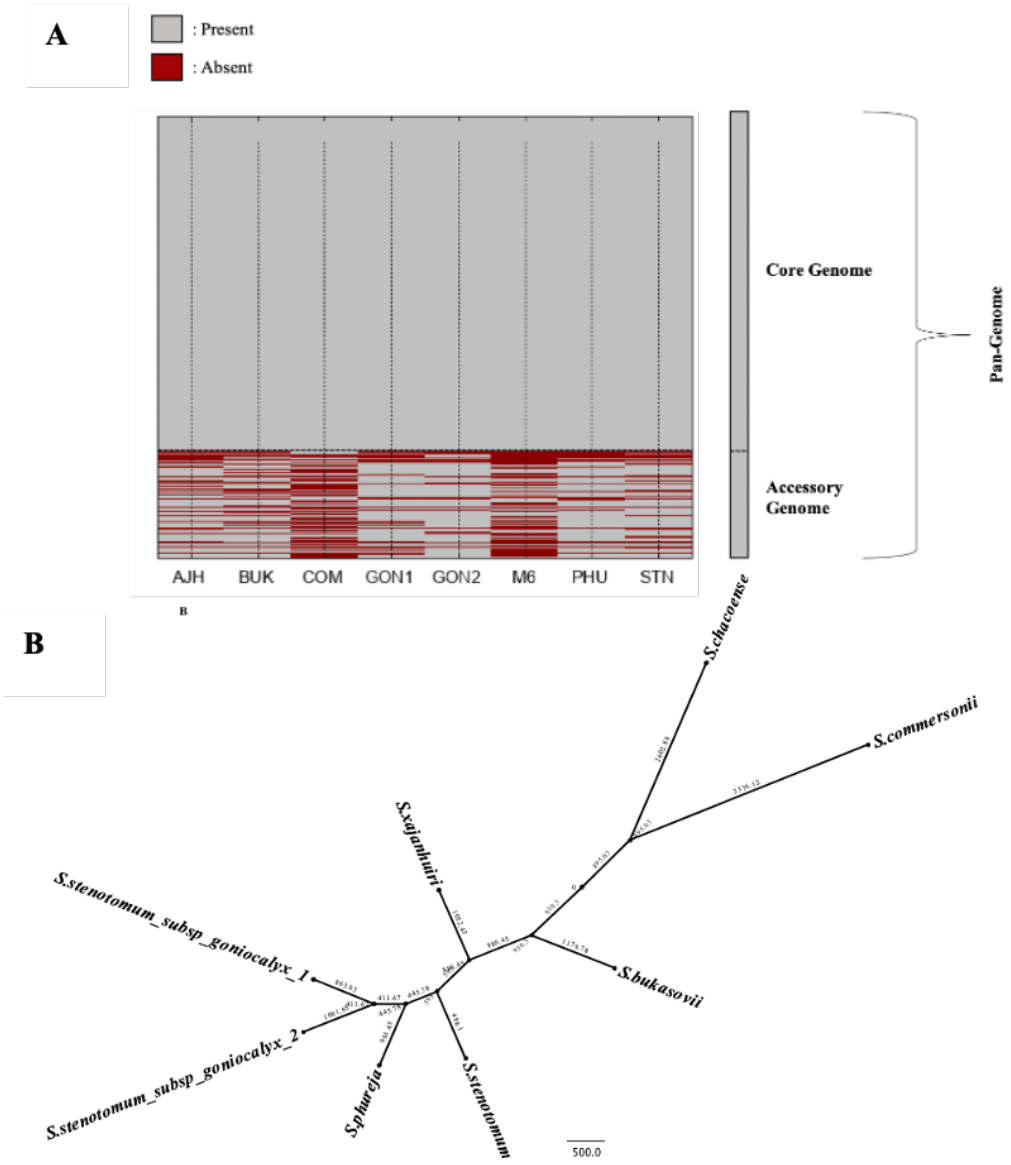


Figure 4.2: Relationship of the pan-genome species. **A.** Heatmap of the Presence/Absence Variable (PAV) genes in the diploid potato pan-genome. Genes present in all the genomes consist of the core genome while those that are absent from some or all is the accessory genome. The core and the accessory genome together consist of the pan-genome. In y-axis, the genes in grey are present in all the genomes, while the genes in maroon are absent from some of the genomes. The x-axis shows the genomes used in this study; *S. xajanhuiri* (AJH), *S. bukasovii* (BUK), *S. commersonii* (COM), *S. stenotomum* subsp. *goniocalyx* (GON1), *S. stenotomum* subsp. *goniocalyx* (GON2), *S. chacoense* (M6), *S. phureja* (PHU), *S. stenotomum* subsp. *stenotomum* (STN) and *S. tuberosum* Group Phureja (DM1-3). **B.** Unrooted Phylogenetic Tree of eight genomes used for the potato pan-genome construction, based on PAV. There are four distinct clusters; one with the wild *S. chacoense* – *S. commersonii* potatoes, another with *S. bukasovii*; another wild species, potential landrace progenitor, the bitter *S. xajanhuiri* makes a cluster itself and finally, the four *S. stenotomum* subsp. *stenotomum*, *S. phureja* and *S. stenotomum* subsp *goniocalyx* 1 and 2 consist of the final cluster.

4.4.6 Functional analysis of the variable genes

Protein families and domains were predicted using Interproscan (Quevillon et al., 2005) within MAKER (Campbell et al., 2014) for a functional analysis of the accessory genome. This uncovered gene families involved in disease resistance, e.g. Leucine Rich Repeats (LRR), putative late blight resistance proteins, NL27 protein, Hcr9-NLOD protein (from *S. lycopersicum*), Hcr2-p7.8 (from the wild *S. pimpinellifolium*). Genes coding for proteins associated with plant defense, including defensins, germ-like protein superfamily 1 member 17 (from *Nicotiana tobacum*), ankyrin repeat-containing protein (from N. attenuata), transcription repressor MYB6-like, putative deacetoxyvindoline 4-hydroxylase-like and others were also found. Additionally, genes involved in biosynthetic processes were identified, such as riboflavin biosynthesis protein RibA, putative 2-oxoglutarate-dependent oxygenase AOP1-like protein coding gene (involved in glucosinolate biosynthesis), 3-isopropylmalate dehydrogenase (involved in glucosinolate biosynthesis), deacetylvindoline O-acetyltransferase – like coding genes (involved in alkaloid biosynthesis), NADH-dependent glutamate synthase, and ribulose-phosphate 3-epimerase. Moreover, genes coding for the putative transcription repressor (Ovate Family Protein) OFP1-like protein were found, and twenty-five genes that code for putative ovule protein - an uncharacterized protein. A Gene Ontology (GO) analysis show that the genes of the accessory genome are involved in various biological functions, including tRNA processing, oxidation-reduction process, response to biotic stimulus, gluconeogenesis and photosynthesis (light reaction) as well as recognition of pollen and response to wounding **Supplementary Table 25.1** (Appendix 18).

The 723 newly identified non-reference genes include various disease resistance and stress-associated genes, and genes involved in tricarboxyl acid cycle and histidine biosynthetic processes. As shown in **Figure 4.3**, genes involved in self-incompatibility, such as those coding for S19-locus linked F-box proteins (**Figure 4.3A**) and flowering locus T (**Figure 4.3B**) were found in the newly identified genes. Based on the PAV analysis, the gene coding for S19-locus linked F-box proteins (PPAN_00000620) is only present in the AJH and

M6 genomes (**Figure 4.3A**), whereas the gene for flowering locus T (PPAN_00001393) is only present in the BUK genome. Several related genes, such as those coding for flowering promoting factor 1-like and *S-locus* were also found in this group (**Supplementary Table 25.1**). Interestingly, 57.2% (413) of the newly predicted protein-coding genes code for uncharacterized proteins mostly identified in other Solanaceae species such as *Nicotiana sylvestris*, *N. tabacum*, *S. demissum*, *S. lycopersicum*, *N. attenuata*, *Capsicum annuum*, *C. baccatum*, and *C. chinense*, but also in species from other families such as *Daucus carota*, *Triticum aestivum*, *Coffea canephora* and others. The number of newly identified genes contributed by each genome in the pan-genome differs from genome to genome, with GON2 contributing the most (366 genes) (**Figure 4.4**).

An analysis of transposable element density shows the majority of the repeats to be unclassified (56.7%). The most abundant class found was the retrotransposons (17.73 % LTR elements and 0.75% LINEs), while the DNA elements reached only the 0.01%. A total of 58.7% of the bases (21,918,247 bp) of the unaligned sequences used for the pan-genome was masked.

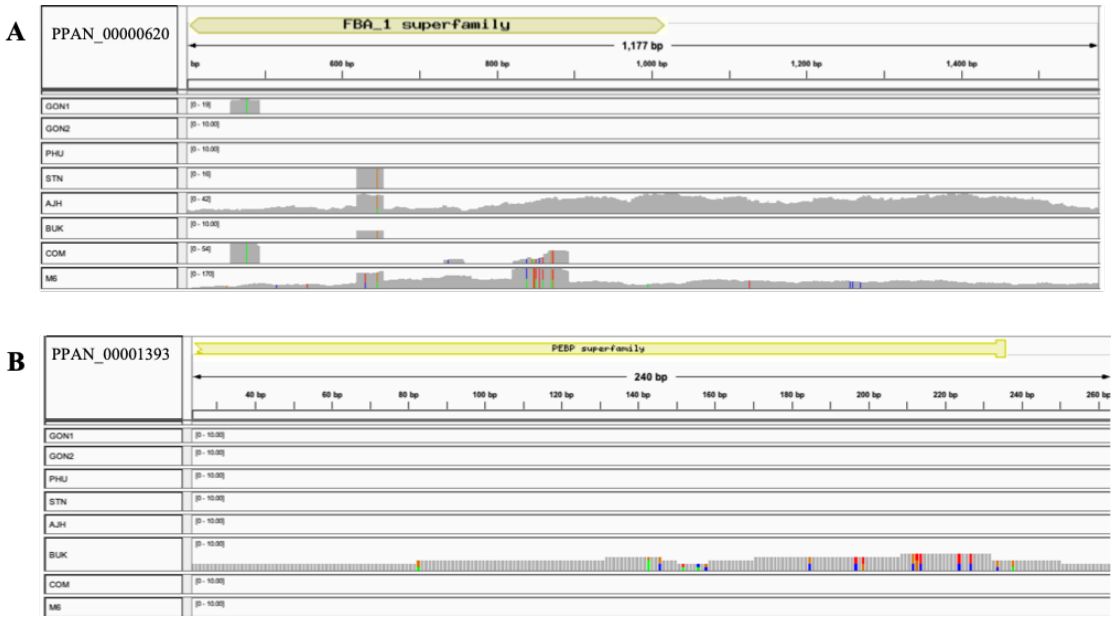


Figure 4.3: Self-incompatibility related genes are part of newly discovered genes in the accessory genome of the diploid potato pan-genome. A. The genomic variation of the newly predicted PPAN_00000620 gene, coding for S19-locus linked F-box protein. It is matched to the ajh_contig140419 (400 – 1,578 bp) of the pan-genome. Based on the PAV analysis, this gene is present only in the AJH and M6 genomes. The conserved domain identified is the F-box associated (322 – 1,010 bp). **B.** The genomic variation of the newly predicted PPAN_00001393 gene, coding for Flowering Locus T. It is located on the buk_contig6862 (24 – 263 bp) of the pan-genome. Based on the PAV analysis, this gene is present only on the BUK genome. The conserved domain identified is Phosphatidyl Ethanolamine-Binding Protein (PEBP) domain (5 - 237). *S. xajanhuiri* (AJH), *S. bukasovii* (BUK), *S. commersonii* (COM), *S. stenotomum* subsp. *goniocalyx* (GON1), *S. stenotomum* subsp. *goniocalyx* (GON2), *S. chacoense* (M6), *S. phureja* (PHU), and *S. stenotomum* subsp. *stenotomum* (STN).

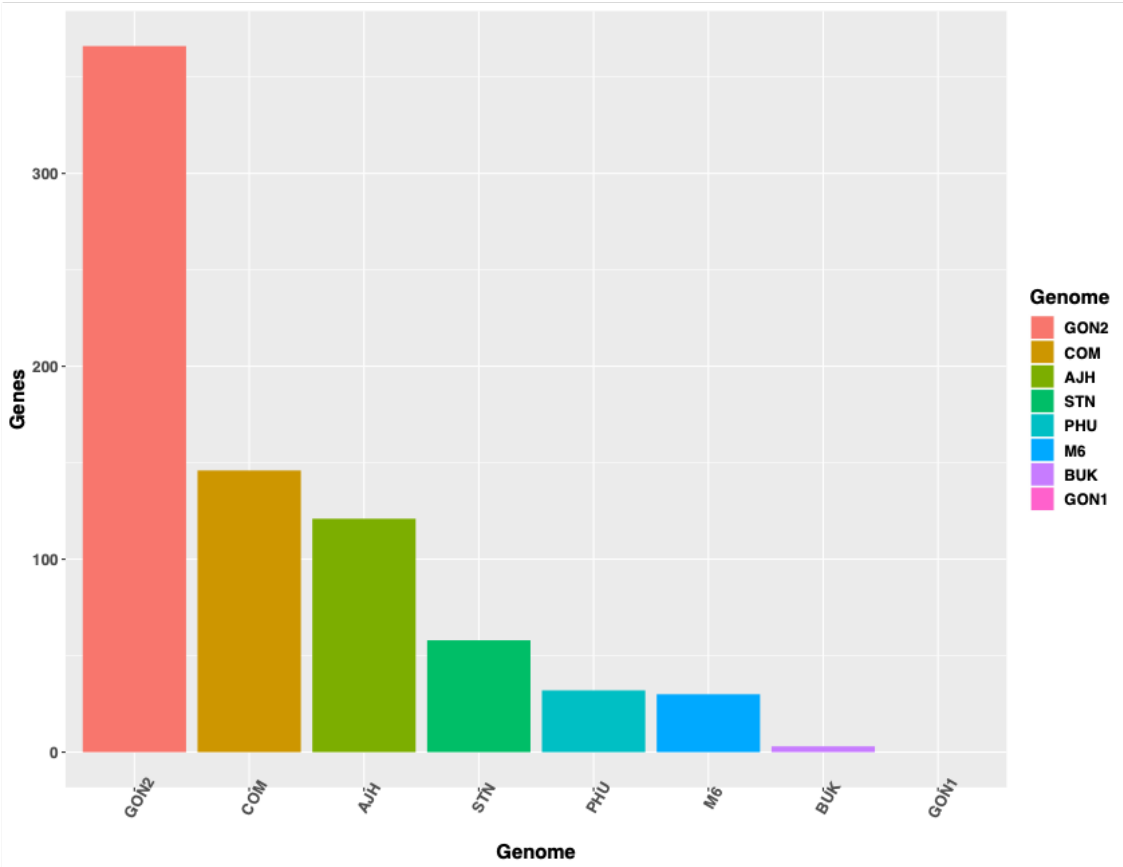


Figure 4.4: Number of newly identified genes, per genome in the pan-genome. The bar plot shows the number of genes identified per genome and contributed to the final newly predicted protein-coding genes. GON2 contributed the most with a total number of 366 genes, while GON1 was analyzed in the end, hence it is shown that it contributed no genes, *S. stenotomum* subsp. *goniocalyx* (GON2), *S. commersonii* (COM), *S. xajanhuiri* (AJH), *S. stenotomum* subsp. *stenotomum* (STN), *S. phureja* (PHU), *S. chacoense* (M6), *S. bukasovii* (BUK) and *S. stenotomum* subsp. *goniocalyx* (GON1).

4.5 Discussion

In this study, we have assembled six diploid potato genomes representing various taxonomical groups. One genome was further placed into pseudomolecules with the help of Linked and Long reads making it comparable to previously published genomes of wild diploid potato (Aversano et al., 2015; Leisner et al., 2018), and making it the first cultivated diploid potato draft genome. To gain a better view of potato core - and accessory genes, we built a pan-genome from these six genomes together with three publicly available potato genomes and analyzed the contents.

4.5.1 Diploid potato genome and pan-genome assemblies

The genomes of *S. stenotomum* subsp. *goniocalyx* (GON1 and GON2), *S. phureja* (PHU), *S. xajanhuiri* (AJH), *S. stenotomum* subsp. *stenotomum* (STN), and *S. bukasovii* (BUK) used in this study are highly heterozygous (taxonomy based on Spooner *et al.*, 2014). The assemblies using only Illumina PE reads (all except GON1) are therefore more fragmented and redundant, though still similar in size to other publicly available sequenced potato genomes (Aversano et al., 2015; Leisner et al., 2018; PGSC, 2011).

The availability of 10X Genomics Linked and PacBio Long reads significantly aided the construction of a less fragmented genome of 856Mb for *S. stenotomum* subsp. *goniocalyx* GON1. The combination of the data from these technologies was a powerful tool for the construction of the GON1 genome, as demonstrated by the BUSCO analysis and from the alignment to other *Solanum* species. About 60% of the GON1 genome is repetitive and 55% (469 Mb) of the genome assembly was placed into 12 pseudomolecules based on the alignment to the DM1-3 genome.

In our pan-genome assembly, 70% (28,208) of the pan-genome genes constitute the core genome and are shared by all the genomes. This is close to the tomato pan-genome, which had 74.2% core genes based on a construction using 725 wild and cultivated tomatoes (Gao et al., 2019). The Brassica oleracea pan-genome contains the highest core gene content described to date: 81.3% (Golicz et al., 2016), followed by the sunflower pan-genome

with 72.7% of core genes (Hübner et al., 2019), *Arabidopsis thaliana* 70% (Contreras-Moreira et al., 2017), *B. napus* 62% (Hurgobin et al., 2018), hexaploidy bread wheat 64.3% (Montenegro et al., 2017), rice 54% (Wang et al., 2018b), wild soybean 49% (Li et al., 2014) and *Brachypodium distachyon* with 35% (Gordon et al., 2017).

4.5.2 Functional analysis of the variable genes

4.5.2.1 Self-incompatibility – related genes

Genes coding for proteins involved in self-fertility were found to be variable genes (not found in the core genome). Self-incompatibility (SI) in the *Solanum* species and other Solanaceae genomes is the S-RNase-based, gametophyic type, where S-specificity is determined by S-RNases in the pistil (McClure et al., 1989) and by S-locus F-box proteins (SLFs) in the pollen (Sijacic *et al.*, 2004). SI prevents self-fertilization and reduces inbreeding. The F-box proteins are components of the SCF (Skp1-Cullin1-F-box) type ubiquitin E3 ligases, which targets specific proteins for degeneration by the 26S proteasome (Zheng et al., 2002; Moon et al., 2004). Transitions to self-compatibility (SC) occur frequently in plants (Goldberg and Igić, 2012; Igic et al., 2008). These genes are of a great importance for the future of potato breeding as the majority of the diploid tuber bearing potato species is gametophytically self-incompatible.

4.5.2.2 Ovate Family Proteins (OFP)

The *OVATE* gene was identified as the cause of pear-shaped tomato and fruit elongation (Liu et al., 2002). Moreover, *OVATE* is mainly expressed in reproductive organs and it is a negative regulator of plant growth as its over-expression reduces the size of floral organs. As for Arabidopsis, a study (Hackbusch et al., 2005) showed that the *OFP1* gene is crucial for male gamete and pollen function. Between others, the banana *OFP1* was found responsible for the fruit's ripening ((Liu et al., 2015). An *OVATE* family member also controls tuber shape (Wu et al., 2018). The results suggest that the *OVATE* genes contribute to the wide variation in tuber shapes among cultivated landraces.

4.5.2.3 FLOWERING LOCUS T genes

FLOWERING LOCUS T has a central role in regulating floral timing in plants (Pin & Nilsson, 2012). Plants carry extensive duplication of *FLOWERING LOCUS T* genes, which contributed to the evolution of multiple functions including tuberization in potato (Navarro et al., 2011). The adaptation of potato tuberization from short to long- days enabled potato production in North America and Europe and is associated with variation in the *CYCLING DOF FACTOR 1* (*StCDF1*) gene that controls the expression of the potato FLOWERING LOCUS T homolog *StSP6A* (Gutaker et al., 2019; Kloosterman et al., 2013). Furthermore, *StSP6A* was also found to be involved in heat and nutrient regulation of tuberization demonstrating the important role it plays in adaptation of potato to the environment (Lehretz et al., 2019; Gálvez et al., 2016). The presence of FLOWERING LOCUS T in the accessory genome concurs with the diversification in the function of this gene and points to a role for the accessory genome in environmental adaption of potato.

4.5.2.4 Disease resistant and plant defense genes

Not surprising was the finding of disease resistance genes in the accessory genome. Previous pan genome studies showed also that disease resistance genes are found in the accessory genome: *S. lycopersicum* ((Gao et al., 2019), sunflower (Hübner et al., 2019), rice (Zhao et al., 2018) and *B. oleracea* (Golicz et al., 2016). These are key genes for potato breeding. Introduction of new disease resistance mechanisms will be of benefit for the modern cultivars.

Genes coding for hero resistance proteins are necessary for the resistance of *Solanum* species against parasitic nematodes. It has been reported that the Hero gene is a potato cyst nematode (two are attacking the potatoes: *Globodera pallida* and *G. rostochiensis*) resistance gene and it is the only member of the *NBS-LRR* genes with unusual amino acid repeat in the LRR region (Ernst et al., 2002). As both nematode species are causing considerable economic losses in potato fields ((FAO, 2013), it is crucial to breed to protect modern cultivars with resistance genes such as this. Xylanase inhibitor proteins (XIP) were firstly discovered in wheat (*T. aestivum*) and genes for these were also found in the

accessory genome. Specifically, wheat XIP inhibits the expression of family 10 and 11 xylanases of Aspergillus nidulans and *A. niger* (FLATMAN et al., 2002).

4.6 Conclusion

In this study, we present a first draft genome of a cultivated diploid landrace potato, *S. stenotomum* subsp. *goniocalyx*, and a first diploid potato pan-genome constructed as a resource for potato genomics and breeding. A total of nine diverse taxa are included: five Andean landraces (*S. stenotomum* subsp. *goniocalyx* (two accessions), *S. phureja*, *S. xajanhuiri*, *S. stenotomum* subsp. *stenotomum*) an Andean wild species (*S. bukasovii*), and three potato genomes previously available (DM1-3/*S. tuberosum* Group Phureja, *S. commersonii*, and M6/*S. chacoense*). The pan-genome consists of 39,751 predicted genes of which 723 are newly identified and include key genes for adaptive processes, such as fertility, flowering timing, fruit and tuber development and shape, and pest and pathogen defense. This resource is a crucial step for potato improvement in the face of climate change.

Bibliography of Chapter 4

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Dal Molin, A., Avanzato, C., Ferrarini, A., and Others (2015). The Solanum commersonii genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *The Plant Cell*, 27(4):954–968.

Bethke, P. C., Halterman, D. A., and Jansky, S. H. (2019). Potato Germplasm Enhancement Enters the Genomics Era. *Agronomy*, 9(10):575.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sørensen, I., Lichtenstein, G., Fich, E. A., Conte, M., Keller, H., Schneeberger, K., Schwacke, R., Ofner, I., Vrebalov, J., Xu, Y., Osorio, S., Aflitos, S. A., Schijlen, E., Jiménez-Goméz, J. M., Ryngajllo, M., Kimura, S., Kumar, R., Koenig, D., Headland, L. R., Maloof, J. N., Sinha, N., van Ham, R. C. H. J., Lankhorst, R. K., Mao, L., Vogel, A., Arsova, B., Panstruga, R., Fei, Z., Rose, J. K. C., Zamir, D., Carrari, F., Giovannoni, J. J., Weigel, D., Usadel, B., and Fernie, A. R. (2014). The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics*, 46(9):1034–1038.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, 48(1):4.11.1–4.11.39.

Contreras-Moreira, B., Cantalapiedra, C. P., Garc\'\ia-Pereira, M. J., Yruela, I., Gordon, S. P., Vogel, J. P., Catalan, P., Igartua, E., Casas, A. M., and Vinuesa, P. (2017). Pan-genomes: estimating the true genomic diversity of plant species.

Ernst, K., Kumar, A., Kriseleit, D., Kloos, D.-U., Phillips, M. S., and Ganal, M. W. (2002). The broad-spectrum potato cyst nematode resistance gene (Hero) from tomato is the

only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. *The Plant Journal*, 31(2):127–136.

FAO (2013). No Title.

FLATMAN, R., McLAUCHLAN, R. W., JUGE, N., FURNISS, C., BERRIN, J.-G., HUGHES, R. K., MANZANARES, P., LADBURY, J. E., O'BRIEN, R., and WILLIAMSON, G. (2002). Interactions defining the specificity between fungal xylanases and the xylanase-inhibiting protein XIP-I from wheat. *Biochemical Journal*, 365(3):773–781.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Gálvez, J. H., Tai, H. H., Lagüe, M., Zebarth, B. J., and Strömvik, M. V. (2016). The nitrogen responsive transcriptome in potato (Solanum tuberosum L.) reveals significant gene regulatory motifs. *Scientific reports*, 6:26090.

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., and Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, 51(6):1044–1051.

Goldberg, E. E. and Igić, B. (2012). TEMPO AND MODE IN PLANT BREEDING SYSTEM EVOLUTION. *Evolution*, 66(12):3701–3709.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., and Edwards, D. (2016). The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, 7(1):13390.

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., Stritt, C., Roulin, A. C., Schackwitz, W., Tyler, L., Martin, J., Lipzen, A., Dochy, N.,

Phillips, J., Barry, K., Geuten, K., Budak, H., Juenger, T. E., Amasino, R., Caicedo, A. L., Goodstein, D., Davidson, P., Mur, L. A. J., Figueroa, M., Freeling, M., Catalan, P., and Vogel, J. P. (2017). Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature Communications*, 8(1):2184.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Gutaker, R. M., Zaidem, M., Fu, Y.-B., Diederichsen, A., Smith, O., Ware, R., and Allaby, R. G. (2019). Flax latitudinal adaptation at LuTFL1 altered architecture and promoted fiber production. *Scientific Reports*, 9(1):976.

Hackbusch, J., Richter, K., Müller, J., Salamini, F., and Uhrig, J. F. (2005). A central role of Arabidopsis thaliana ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(13):4908 LP – 4912.

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., and Others (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, 28(2):388–405.

Hawkes, J. G. and Others (1990). *The potato: evolution, biodiversity and genetic resources.* Belhaven Press.

Hirsch, C. D., Hamilton, J. P., Childs, K. L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C. N., Habermann, M., Neal, B., and Buell, C. R. (2014). Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding. *The Plant Genome*, 7.

Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., Lee, J. S., Baute, G. J., Owens, G. L., Grassa, C. J., Ebert, D. P., Ostevik, K. L., Moyers, B. T.,

Yakimowski, S., Masalia, R. R., Gao, L., Ćalić, I., Bowers, J. E., Kane, N. C., Swanevelder, D. Z. H., Kubach, T., Muños, S., Langlade, N. B., Burke, J. M., and Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, 5(1):54–62.

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. 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., and Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnology Journal*, 16(7):1265–1274.

Igic, B., Lande, R., and Kohn, J. (2008). Loss of Self-Incompatibility and Its Evolutionary Consequences. *International Journal of Plant Sciences*, 169(1):93–104.

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., Xue, Z., Mohamadi, H., Bohlmann, J., Jones, S. J. M., and Birol, I. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, 19(1):393.

Jansky, S. H., Charkowski, A. O., Douches, D. S., Gusmini, G., Richael, C., Bethke, P. C., Spooner, D. M., Novy, R. G., De Jong, H., De Jong, W. S., Bamberg, J. B., Thompson, A. L., Bizimungu, B., Holm, D. G., Brown, C. R., Haynes, K. G., Sathuvalli, V. R., Veilleux, R. E., Miller, J. C., Bradeen, J. M., and Jiang, J. (2016). Reinventing Potato as a Diploid Inbred Line–Based Crop. *Crop Science*, 56:1412–1422.

Kitts, P. A., Madden, T. L., Sicotte, H., Black, L., and Ostell, J. A. (2016). UniVec Database. 2016. *Available from: ncbi. nlm. nih. gov/VecScreen/UniVec. html*.

Kloosterman, B., Abelenda, J. A., Gomez, M. d. M. C., Oortwijn, M., de Boer, J. M., Kowitwanich, K., Horvath, B. M., van Eck, H. J., Smaczniak, C., Prat, S., Visser, R. G. F., and Bachem, C. W. B. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*, 495(7440):246–250.

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

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, 5(1):59.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12.

Kyriakidou, M., Achakkagari, S. R., López, J. H. G., Zhu, X., Tang, C. Y., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2020). Structural genome analysis in cultivated potato taxa. *Theoretical and Applied Genetics*, 133(3):951–966.

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, 9:1660.

Lehretz, G. G., Sonnewald, S., Hornyik, C., Corral, J. M., and Sonnewald, U. (2019). Post-transcriptional Regulation of FLOWERING LOCUS T Modulates Heat-Dependent Source-Sink Development in Potato. *Current Biology*, 29(10):1614–1624.e3.

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., and Others (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. *The Plant Journal*, 94(3):562–570.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

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. L., 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., and Qiu, L.-j. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology*, 32(10):1045–1052.

Lindhout, P., Meijer, D., Schotte, T., Hutten, R. C. B., Visser, R. G. F., and van Eck, H. J. (2011). Towards F1 Hybrid Seed Potato Breeding. *Potato Research*, 54(4):301–312.

Liu, J., Van Eck, J., Cong, B., and Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences*, 99(20):13302–13306.

Liu, J., Zhang, J., Hu, W., Miao, H., Zhang, J., Jia, C., Wang, Z., Xu, B., and Jin, Z. (2015). Banana Ovate family protein MaOFP1 and MADS-box protein MuMADS1 antagonistically regulated banana fruit ripening. *PloS one*, 10(4):e0123870–e0123870.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

McClure, B. A., Haring, V., Ebert, P. R., Anderson, M. A., Simpson, R. J., Sakiyama, F., and Clarke, A. E. (1989). Style self-incompatibility gene products of Nicotlana alata are ribonucleases. *Nature*, 342(6252):955–957.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., and Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013.

Moon, J., Parry, G., and Estelle, M. (2004). The Ubiquitin-Proteasome Pathway and Plant Development. *The Plant Cell*, 16(12):3181–3195.

Navarro, C., Abelenda, J. A., Cruz-Oró, E., Cuéllar, C. A., Tamaki, S., Silva, J., Shimamoto, K., and Prat, S. (2011). Control of flowering and storage organ formation in potato by FLOWERING LOCUS T. *Nature*, 478(7367):119–122.

PGSC (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, 33(suppl_2):W116–W120.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912.

Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., Wright, M. H., Chia, J.-m., Ware, D., McCouch, S. R., and McCombie, W. R. (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(11):506.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Smit, A. and Hubley, R. (2019). RepeatModeler-1.0. 11. *Institute for Sys-tems Biology. http://www. repeatmasker. org/RepeatModeler/. Accessed*, 15.

Smit, A. F. A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. 2013–2015.

Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H., and Gavrilenko, T. (2014). Systematics, Diversity, Genetics, and Evolution of Wild and Cultivated Potatoes. *The Botanical Review*, 80(4):283–383.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts . *Nucleic Acids Research*, 34(suppl_2):W435–W439.

Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., Shi, J., Wang, C., Lu, J., Zhang, D., Li, Z., and Wei, C. (2016). RPAN: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Research*, 45(2):597–605.

Tamazian, G., Dobrynin, P., Krasheninnikova, K., Komissarov, A., Koepfli, K.-P., and O'Brien, S. J. (2016). Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience*, 5(1).

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., 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\textquoterightConnor, K. J. B., 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., and Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial {\textquotedblleft}pan-genome{\textquotedblright}. *Proceedings of the National Academy of Sciences*, 102(39):13950–13955.

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23:148–154.

Wang, A., Wang, Z., Li, Z., and Li, L. M. (2018a). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, 34(12):2019–2028.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., Zhao, X., Shen, F., Cui, X., Yu, H., Li, Z., Chen, M., Detras, J., Zhou, Y., Zhang, X., Zhao, Y., Kudrna, D., Wang, C., Li, R., Jia, B., Lu, J., He, X., Dong, Z., Xu, J., Li, Y., Wang, M., Shi, J., Li, J., Zhang, D., Lee, S., Hu, W., Poliakov, A., Dubchak, I., Ulat, V. J., Borja, F. N., Mendoza, J. R., Ali, J., Li, J., Gao, Q., Niu, Y., Yue, Z., Naredo, M. E. B., Talag, J., Wang, X., Li, J., Fang, X., Yin, Y., Glaszmann, J.-C., Zhang, J., Li, J., Hamilton, R. S., Wing, R. A., Ruan, J., Zhang, G., Wei, C., Alexandrov, N., McNally, K. L., Li, Z., and Leung, H. (2018b). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 557(7703):43–49.

Warren, R. (2016). RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software*, 1(7):116.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome research*, 27(5):757–767.

Wu, S., Zhang, B., Keyhaninejad, N., Rodríguez, G. R., Kim, H. J., Chakrabarti, M., Illa-Berenguer, E., Taitano, N. K., Gonzalo, M. J., Díaz, A., Pan, Y., Leisner, C. P., Halterman, D., Buell, C. R., Weng, Y., Jansky, S. H., van Eck, H., Willemsen, J., Monforte, A. J., Meulia, T., and van der Knaap, E. (2018). A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nature Communications*, 9(1):4734.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., and Xie, W. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology*, 16(1):187.

Yeo, S., Coombe, L., Warren, R. L., Chu, J., and Birol, I. (2017). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34(5):725–731.

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., and Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, 50(2):278–284.

Zheng, J., Yang, X., Harrell, J. M., Ryzhikov, S., Shim, E.-H., Lykke-Andersen, K., Wei, N., Sun, H., Kobayashi, R., and Zhang, H. (2002). CAND1 Binds to Unneddylated CUL1 and Regulates the Formation of SCF Ubiquitin E3 Ligase Complex. *Molecular Cell*, 10(6):1519–1526.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677.

Preface to Chapter 5

In chapter 4, a pan-genome model for diploid potato was presented, including six newly sequenced and three previously published diploid potato genomes. Chapter 5 focuses on genome assembly in polyploids and presents the de novo genome assembly of two triploid, three tetraploid and one pentaploid potato genomes. For the de novo construction of the tetraploid ADG1 genome, Third Generation Sequencing Technologies are used (long and linked reads). For the rest of the genomes, only Illumina PE reads were used. The findings show that indeed the availability of longer reads helps the big problem of the de novo genome assembly of polyploid plants, however, it is concluded that there is a need for new sequencing technologies and new assembly algorithms. The manuscript was published in Nature Scientific Data .

Genome assembly of six polyploid potato genomes

Maria Kyriakidou[1], Helen H. Tai[2], Noelle L. Anglin[3], David Ellis[3], Martina V. Stromvik[1*]

[1] Department of Plant Science, McGill University, Montreal, Canada

[2] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, Canada

[3] International Potato Center, Lima, Peru

[*] Correspondence:

Corresponding Author

martina.stromvik@mcgill.ca

5.1 Abstract

Genome assembly of polyploid plant genomes is a laborious task as they contain more than two copies of the genome, are often highly heterozygous with a high level of repetitive DNA. Next Generation genome sequencing data representing one Chilean and five Peruvian polyploid potato (Solanum spp.) landrace genomes was used to construct genome assemblies comprising five taxa. Third Generation sequencing data (Linked and Long-read data) was used to improve the assembly for one of the genomes. Native landraces are valuable genetic resources for traits such as disease and pest resistance, environmental tolerance and other qualities of interest such as nutrition and fiber for breeding programs. The need for conservation and enhanced understanding of genetic diversity of cultivated potato from South America is also crucial to North American and European cultivars. Here, we report draft genomes from six polyploid potato landraces representing five taxa, illustrating how Third Generation Sequencing can aid in assembling polyploid genomes.

5.2 Introduction

Native potato species are distributed from the southwestern United States to Argentina (Hijmans and Spooner, 2001). The most commonly cultivated potato varieties are autotetraploids (2n=4x=48) with a base chromosome number of 12. However, cultivated potato landraces can range from diploids (2n=2x=24) to pentaploids (2n=5x=60) (Watanabe, 2015) and wild potato species from the United States, Mexico and central America also include hexaploid species (Lara-Cabrera and Spooner, 2004). The potato genome is characterized by great heterozygosity, due likely to the fact that most of the diploid potato species are self-incompatible (Bradshaw, 2007; Watanabe, 2015).

A significant amount of work has previously been performed to aid the advance of potato genomics (Gálvez Helen H., Barkley, Noelle A., Gardner, Kyle, Ellis, David, Strömvik, Martina V., José Héctor, 2017). Currently, the publicly available potato reference genomes are from the doubled monoploid *Solanum tuberosum* Group phureja DM1-3 (PGSC, 2011), the wild diploid *S. commersonii* (Aversano et al., 2015) and the diploid, inbred clone of *S. chacoense* - M6 (Leisner et al., 2018). S. tuberosum is an autotetraploid, and evidence suggests the polyploid nature resulted through duplication events. Hence, a single reference genome cannot capture the great diversity found across different potato genomes, especially in the case of polyploids since they are more heterozygous than the diploids (Hirsch et al., 2014b). Improvement of current algorithms and of current sequencing technologies are fundamental to improving the assembly of polyploid genomes such as those found in diverse potato species (Kyriakidou et al., 2018). Next Generation Sequencing (NGS) made a revolution in approaches to genome sequencing, due to reduced costs and faster sequencing compared with Sanger sequencing technology. However, NGS does have drawbacks, especially when sequencing polyploid genomes, where their short length can lead to misassemblies and extremely fragmented genome assemblies. The most recent evolution in the era of genome sequencing is the Third Generation (or Long-read) Sequencing (TGS) technologies, which can produce high quality genome assemblies with high resolution due to the longer length of the reads. TGS technologies can reduce the

problem of assembling polyploid plant genomes (Kyriakidou et al., 2018). Various complicated polyploid plant genomes have been sequenced with TGS technologies including *Chenopodium quinoa* (3x) (Jarvis et al., 2017) and *Saccharum* sp (varying ploidy levels) (Riaño-Pachón and Mattiello, 2017), *Fragaria x ananassa* (8x) ((Edger et al., 2017) and others.

Twelve potato genomes of various ploidy levels were recently sequenced (Kyriakidou et al., 2020a). These genomes, which were selected based on the Hawkes taxonomy (Hawkes and Others, 1990), in addition to the *S. commersonii* genome (Aversano et al., 2015) were compared to the two publicly available reference genomes *S. tuberosum* Group Phureja (DM1-3) (PGSC, 2011) and *S. chacoense* M6 clone (Leisner et al., 2018) for copy number variation (CNV) and SNP analyses. The study showed the great diversity across this panel of potato genomes and identified a number of CNVs in genes implicated in disease resistance and stress, among other processes.

In the present study, we have focused on assembling the reads for the six polyploid genomes from the previously sequenced cultivated potato landraces covering five taxa (based on (Hawkes and Others, 1990) (Kyriakidou et al., 2020a): *Solanum chaucha* (3x: CHA), *S. juzepczukii* (3x: JUZ), two genomes of *S. tuberosum* subsp. *andigena* (4x: ADG1 and ADG2), *S. tuberosum* subsp. *tuberosum* (4x: TBR) and *S. curtilobum* (5x: CUR). One of the genomes, ADG1 – a tetraploid, is assembled with TGS and has therefore a higher quality assembly, while NGS data is used for the others.

5.3 Materials and Methods

5.3.1 Genomic Data

S. chaucha (CHA – CIP 707129 doi:10.18730/CS5*), *S. juzepczukii* (3x: JUZ – CIP 706050 doi:10.18730/C09D), two genomes of *S. tuberosum* subsp. *andigena* (4x: ADG1 - CIP 700921 doi:10.18730/91RP; ADG2 - CIP 702853 doi:10.18730/9GB8), *S. tuberosum* subsp. *tuberosum* (4x: TBR – CIP 705053 doi:10.18730/B3MN) and *S. curtilobum* (5x: CUR – CIP 702937 doi:10.18730/9H1Y), from the *in vitro* potato germplasm collection at the International

Potato Center (CIP) in Lima, Peru (Kyriakidou et al., 2020a). Genomic DNA was extracted and sequenced using an Illumina HiSeq sequencer (Illumina, Inc.) in paired-end mode (2 x 150 bp) as described (Kyriakidou et al., 2020a). The genome of ADG1 was also sequenced (50 X) with PacBio's Single Molecule RS II system technology (Eid et al., 2009) and with 10X Genomics' GemCode technology (134 X) (Weisenfeld et al., 2017) by Novogene™.

5.3.2 Determining the whole genome heterozygosity

Trimmed sequencing reads were used for the calculation of the percentage of heterozygosity in the genomes (Kyriakidou et al., 2020a). For this, jellyfish v2.2.10 (Marçais and Kingsford, 2011) was first used to compute the histogram of the k-mer frequencies. The final k-mer count histogram per genome was used within the GenomeScope 2.0 online platform (Ranallo-Benavidez et al., 2020).

5.3.3 *De novo* genome assemblies

5.3.3.1 ADG1 assembly

Because of the availability of Linked and Long Reads, the genome of ADG1 genome was assembled following a hybrid-read method. Multiple approaches were tried but the best assembly possible was obtained using a combination of Long and Linked Reads with Canu (Koren et al., 2017) and Supernova™ assemblers (Weisenfeld et al., 2017). For the following analyses, pseudohap1 was used as suggested in the genome assembly of Capsicum annuum (Hulse-Kemp et al., 2018) with 10X Genomics reads. Moreover, the Long Reads from PacBio were assembled with Canu v1.5 assembler (Koren et al., 2017), then Tigmint v0.9 (Jackman et al., 2018) was used to correct PacBio misassemblies using the parameters from 10X Genomics. The contigs were assembled into scaffolds with ARCS v1.0.2 (Yeo et al., 2017). The final genome assembly was aligned to the DM1-3 v4.04 (Hardigan et al., 2016), and BUSCO v3.2.0 (Simão et al., 2015) and QUAST (Gurevich et al., 2013) v5.0.0 were used for the evaluation of the assembly. Transposable elements

and repeat masking was performed with RepeatModeler v1.0.11 (Smit and Hubley, 2019) and RepeatMasker v4.0.7 (Smit et al., 2015).

5.3.3.2 CHA, JUZ, ADG2, TBR and CUR assemblies

The Illumina PE reads of the CHA, JUZ, ADG2, TBR, and CUR genomes were assembled using MaSuRCA v3.2.4 (Zimin et al., 2013). Redundant contigs were removed from the assembly using CD-HIT v4.8.1 (Fu et al., 2012; Li and Godzik, 2006) with identity $> 90\%$. The resulting assemblies were evaluated using BUSCO v3.2.0 (Simão et al., 2015) and QUAST (Gurevich et al., 2013) v5.0.0. From all the genome assemblies (ADG1, ADG2, TBR, JUZ, CHA and CUR), any mitochondrial and chloroplast genome has been removed, along with the contigs with length smaller than 200 bp.

5.3.4 Data Records

The reads data is available as BioProject PRJNA556263 (SRA accessions SRR10237766, SRR10242927, SRR10248510 – SRR10248515) at NCBI. The final genome assemblies are deposited into NCBI Assembly database under the following Accession Numbers: GCA_009849705.1, GCA_009849725.1, GCA_009849745.1, GCA_009849685.1, GCA_009849625.1, and GCA_009849645.1 (NCBI Assembly GCA_009849705.1, GCA_009849725.1, GCA_009849745.1, GCA_009849685.1, GCA_009849625.1, GCA_009849645.1).

5.4 Results

5.4.1 Quality of the sequenced genomes – Whole genome heterozygosity

The read coverage ranged between 36 X in the pentaploid CUR and 44.4 X in the triploid CHA for the Illumina reads (**Table 5.1**). The read coverage for the ADG1 genome was calculated with linked and long reads and it had an average read coverage of 50 X (**Table 5.1**). The k-mer frequencies were calculated for each of the genomes (Appendix 33-38:

Supplementary Figure 40.1 – 45.1). In general, there is a tendency towards bimodal distributions. In addition, the heterozygosity of the genomes ranges between 3.52 % (in ADG1) and 12.02 % (in CUR) (**Table 5.1**). The heterozygosity is confirmed by the k-mer frequency of the genomes and the bimodal distributions, which has previously been reported for polyploid genomes (Ranallo-Benavidez et al., 2020).

5.4.2 Genome assembly of ADG1

A draft genome assembly of the *S. tuberosum* subsp. *andigena* (CIP 700921 DOI: 10.18730/91RP) – ADG1 was generated using a hybrid assembly approach of Third Generation Sequencing Data: Linked and Long reads (**Table 5.1**). This methodology was applied as it was previously tested in the group and was found to be the best approach for the data available. The initial assembly contains 87,194 contigs, with an N50 of 62,124 bp (**Table 5.2**). The final assembly, after removing redundancy, consists of 35,961 scaffolds and an N50 of 122,016 bp (**Table 5.2**). The genome size was estimated with a 10X Genomics Chromium library at 896.84 Mb, which is close to the size of other potato genomes (Aversano et al., 2015; Leisner et al., 2018; PGSC, 2011). The size of the assembly including only scaffolds longer than 10 kb, reaches 713.51 Mb. For the evaluation of the genome completeness of ADG1, BUSCO (Simão et al., 2015) was used, finding 85.8% of BUSCO's core Plantae ortholog genes present in the assembly and another 8.5% present as partial sequences (C:85.8% [S:76.3% ,D:9.5%], F:8.5%, M:5.7%, n:1375).

To identify and mask the repetitive elements in the ADG1 assembly, RepeatModeler (Smit and Hubley, 2019) was used to construct a repetitive library, followed by RepeatMasker (Smit et al., 2015). About 60% of the assembly was masked. **Table 5.3** shows the repetitive content of the ADG1 genome.

Table 5.1: Assembled genomes, along with the technologies used for sequencing and their references.

Solanum full taxon name[1]	Ploidy	Accession	Code Name	Technology	NCBI record	Coverage (X)	Heter %
Solanum chaucha	3x	CIP 707129 doi:10.18730/CS5*	CHA	Illumina PE	GCA_009849625.1	44.4	3.7
S. juzepczukii	3x	CIP 706050 doi:10.18730/C09D	JUZ	Illumina PE	GCA_009849685.1	37.7	7.3
S. tuberosum subsp. andigena	4x	CIP 700921 doi:10.18730/91RP	ADG1	10X Genomics PacBio	GCA_009849705.1	50	3.52
S. tuberosum subsp. andigena	4x	CIP 702853 doi:10.18730/9GB8	ADG2	Illumina PE	GCA_009849725.1	43.6	7.75
S. tuberosum subsp. tuberosum	4x	CIP 705053 doi:10.18730/B3MN	TBR	Illumina PE	CA_009849745.1	40.3	8.43
S. curtilobum	5x	CIP 702937 doi:10.18730/9H1Y	CUR	Illumina PE	GCA_009849645.1	35.8	12.02

Table 5.2: Genome assembly statistics of the ADG1, ADG2, TBR, JUZ, CHA and CUR genomes. (Values in the parentheses before removing the redundant contigs)

Quality metric	ADG1	ADG2	TBR	JUZ	CHA	CUR
# of contigs	35,961 *[87,194]*	310,723 *[826,888]*	1,272,956 *[433,4576]*	249,222 *[692,839]*	259,834 *[608,922]*	578,826 *[1,348,978]*
Contigs > 1Kb	35,744 -	248,064 *[456,177]*	271,542 *[271,558]*	194,864 *[436,731]*	194,390 *[344,939]*	364,379 *[657,215]*
Length (Mb)	841.4 *[842]*	991.1 *[1,611]*	1,032 *[1,598]*	1,002 *[1,800]*	790.3 *[1,251]*	1,208 *[2,067]*
GC %	34.83 -	34.84 *[35.2]*	35.69 *[36.05]*	35.41 *[35.63]*	35.32 s*[35.88]*	36.27 *[36.7]*
Largest contig length (Kb)	3,384	102	73	112	105	118
Contig N50	122,016 *[62,124]*	4,721 *[3,154]*	1,193 *[267]*	7,359 *[4,598]*	4,795 *[3,335]*	3,176 *[2,221]*
Contig L50	1,312	59,398	207,326	84,278	42,633	109,841
% BUSCO present genes	85.8	53	18.8	58.6	54.1	45.8
% BUSCO partial genes	8.5	28.4	35.6	27.2	24.7	34.6
% BUSCO duplicated genes	9.5	5.7	3.5	6.9	4.1	6

Table 5.3: Repeat Content of the ADG1 assembly. Data generated with RepeatMasker (Smit *et* al., 2015)

Element	Number of Elements	Length Occupied (bp)	Percentage of sequence
LINEs	43,676	18,473,203	1.7
LTR elements	219,424	261,037,853	23.98
DNA elements	28,736	14,413,138	1.32
Simple repeats	171,025	9,049,100	0.83
Low complexity	36,517	2,254,450	0.21
Unclassified	1,109,924	333,262,813	30.61
Total bases masked		515,341,644	60.2

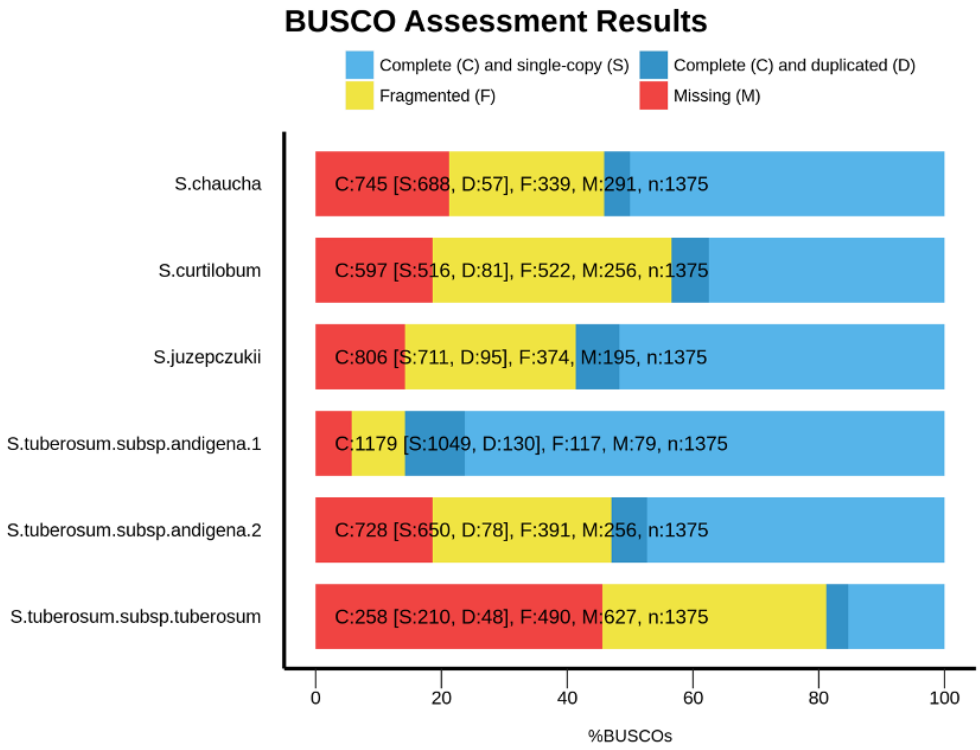


Figure 5.1: Bar chart with BUSCO's summary assessment results for the assembled six polyploid genomes. Light blue shows the % of complete and single copy genes, the darker blue the % of complete and the duplicated genes, the yellow the % of fragmented genes and finally the red shows the % of missing genes in the assemblies.

5.4.3 Genome assembly of CHA, JUZ, ADG2, TBR and CUR genomes

The initial genome assemblies were longer than the size of other reported potato genomes (Aversano et al., 2015; Leisner et al., 2018; PGSC, 2011) (**Table 5.2**). For instance, the CUR genome assembly was about 2.4 times longer than the potato reference genomes, which had genome sizes equal to 884.1 Mb (DM1-3), 830 Mb (*S. commersonii*) and 825.7 Mb (*S. chacoense*). The JUZ, ADG2, TBR genome assemblies were at least double the length the reference genomes, while CHA was shorter than the rest of the polyploid genomes (**Table 5.2**). These differences are likely due to the high heterozygosity in these polyploid genomes. Therefore CD-HIT ((Fu et al., 2012; Li and Godzik, 2006) was used to remove the redundant contigs that were present in each of the assemblies. After removing the redundant contigs from the genomes, the final contig number was reduced to almost a third of the initial number, while the genome size is 0.66% smaller compared to the initial assembly (**Table 5.2**). The assembly statistics improved after removing the redundant contigs.

Even though the removal of the redundant contigs improved the genome assemblies, the assemblies are still very heterozygous and very fragmented (**Table 5.2**). Based on the gene content, the TBR assembly is the most fragmented. **Figure 5.1** shows that presence of BUSCO's core Plantae ortholog genes in TBR almost reached 18.8%, while the majority (35.6%) are partial genes. For the rest of the genomes, the amount of orthologous genes did not exceed 58.6% (**Figure 5.1**, **Table 5.2**; JUZ), with an average amount of fragmented genes at 27.7%. The quality of the Illumina PE genome assemblies was similar among the genomes, with TBR being the exception.

5.4.4 Comparison of the genome assemblies of ADG1 and ADG2

Table 5.2 shows that the genome assembly of ADG1 using Linked and Long reads yielded 35,961 contigs, compared with the ADG2 assembly using only Illumina reads that yielded 310,723 contigs – almost one order of magnitude difference. Moreover, almost all the contigs of ADG1 are greater than 1,000 bp in length, while only 248,064 contigs (80%) of

the ADG2 have lengths greater than 1,000 bp. The N50 for ADG1 is 25.8 times larger than that of ADG2. Finally, in ADG1 85.8% of the BUSCO genes were present, in contrast to ADG2, where only 53% of BUSCO genes were detected. The GC% content was very close for both genomes; 34.83% and 34.84% for ADG1 and ADG2, respectively.

5.4.5 Comparison of the genome assemblies of ADG1 and ADG2, TBR, JUZ, CHA and CUR

As shown in **Table 5.2**, the largest genome assembly is that of the pentaploid CUR genome (1.2 Gb), while the shortest is the triploid CHA genome (790.4 Mb). The TBR assembly was the most fragmented (1,272,956 contigs) compared to the rest of the genomes. Additionally, in TBR, only 21.3% (271,542) of the total number of contigs have length more than 1,000 bp, while 78.16% (194,864) of the JUZ's contigs and all the contigs of ADG1 are larger than 1,000 bp. The GC% content ranged between 34.83% (in ADG1) and 36.27% (in CUR). ADG1 had the largest contig (3.4Mb), followed by CUR (117.7 kb) and JUZ (112 kb). The N50 is dramatically improved in the ADG1 compared to the others. TBR has the smallest N50 (1,193), showing once again the very fragmented assembly due to the high heterozygosity of this genome. Finally, all the genomes had more than 43% of BUSCO's genes present, except TBR, in which only 18.8% of the total BUSCO genes were found.

5.5 Discussion

5.5.1 Highly fragmented genome assemblies due to the heterozygous nature of the polyploid potato genomes

The high ploidy level can lead to higher heterozygosity, causing difficulties in haplotype identification in assemblies without Long range or Long read data (Kyriakidou et al., 2018). In the current study, the CHA, JUZ, TBR, and CUR assembled polyploid genomes are highly fragmented, while the ADG1 assembly, which included Long Range data, resulted in the construction of a less fragmented genome, less redundant and with

fewer contigs. This demonstrates the benefit and need for Long range data for complex genomes. Additionally, there has been innovation in novel assembly algorithms and new assembly strategies using Long range data for the genome assembly of polyploid genomes (Kyriakidou et al., 2018). Moreover, the repetitiveness of the potato genome makes its assembly even more difficult. It appears that 60.2% of the ADG1 genome accounts for repetitive sequences, which is also in agreement with previous contents of repetitive sequences in other potato species; 62.2% in the *S. tuberosum* DM1-3 genome and 60.7% in the M6 clone of the *S. chacoense* (Leisner et al., 2018).

Among the six assembled genomes, the triploid CHA is the shortest. In previous studies using copy number variation analysis and SNP detection analysis of this genome (compared to the DM1-3 genome), it appears less heterozygous than JUZ, which is also a triploid, but also less heterozygous than the rest of the polyploids (Ellis et al., 2018; Kyriakidou et al., 2020a). The most challenging genome to assemble was the tetraploid TBR and not the pentaploid CUR, as would have been expected. It may be that the greater heterozygosity in TBR lead to it being the most fragmented genome assembly. This is supported by a previous study using the Infinium 12K V2 Potato Array in a subset of the CIP potato collection – TBR were among the species with the highest amount of admixture (Ellis et al., 2018). Even in relation to other tetraploids, TBR appears to be the most heterozygous when compared to the DM1-3 v4.04 reference (Kyriakidou et al., 2020a). High levels of heterozygosity were observed from the sequencing data of the cultivated clones in the study. The clonal propagation of potato over thousands of years limited genetic recombination and led to high levels of heterozygosity. Polyploidy and self-incompatibility may also have contributed.

5.6 Conclusion

The genome assembly of plant genomes, and especially polyploid plant genomes, is very complex and challenging. The genome assemblies of two triploid (3x), three tetraploid (4x) and one pentaploid (5x) potato were constructed. Even though the majority of the

assemblies are fragmented, these genomes provide a great resource to enhance potato breeding. It is known that the polyploid genomes contain more genes, hence these potato genomes can be explored for their genetic content. Moreover, as predicted, the availability of Third Generation Sequencing data greatly reduces the genome assembly problem.

Bibliography of Chapter 5

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Dal Molin, A., Avanzato, C., Ferrarini, A., and Others (2015). The Solanum commersonii genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *The Plant Cell*, 27(4):954–968.

Bradshaw, J. E. (2007). Potato-breeding strategy. In *Potato biology and biotechnology*, pages 157–177. Elsevier.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., Callow, P., McKain, M. R., Shi, J., Collier, C., Xiong, Z., Mower, J. P., Slovin, J. P., Hytönen, T., Jiang, N., Childs, K. L., and Knapp, S. J. (2017). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *GigaScience*, 7(2).

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133 LP – 138.

Ellis, D., Chavez, O., Coombs, J., Soto, J., Gomez, R., Douches, D., Panta, A., Silvestre, R., and Anglin, N. L. (2018). Genetic identity in genebanks: application of the SolCAP 12K SNP array in fingerprinting and diversity analysis in the global in trust potato collection. *Genome*, 61(7):523–537.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Gálvez Helen H., Barkley, Noelle A., Gardner, Kyle, Ellis, David, Strömvik, Martina V., José Héctor, T. (2017). Understanding potato with the help of genomics. *AIMS Agriculture and Food*, 2(1):16–39.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

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., and Others (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, 28(2):388–405.

Hawkes, J. G. and Others (1990). *The potato: evolution, biodiversity and genetic resources.* Belhaven Press.

Hijmans, R. J. and Spooner, D. M. (2001). Geographic distribution of wild potato species. *American Journal of Botany*, 88(11):2101–2112.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, 26(1):121–135.

Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., Hill, T. A., Jaffe, D., Williams, S. R., Weisenfeld, N., Ramakrishnan, S., Kumar, V., Shah, P., Schatz, M. C., Church, D. M., and Van Deynze, A. (2018). Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Horticulture Research*, 5(1):4.

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., Xue, Z., Mohamadi, H., Bohlmann, J., Jones, S. J. M., and Birol, I. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, 19(1):393.

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., Kharbatia, N. M., Rupper, R. R., Sharp, A. R., Dally, N., Boughton, B. A., Woo, Y. H., Gao, G., Schijlen, E. G. W. M., Guo, X., Momin, A. A., Negrão, S., Al-Babili, S., Gehring, C., Roessner, U., Jung, C., Murphy, K., Arold, S. T., Gojobori, T., van der Linden, C. G., van Loo, E. N., Jellen, E. N., Maughan, P. J., and Tester, M. (2017). The genome of Chenopodium quinoa. *Nature*, 542(7641):307–312.

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

Kyriakidou, M., Achakkagari, S. R., López, J. H. G., Zhu, X., Tang, C. Y., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2020). Structural genome analysis in cultivated potato taxa. *Theoretical and Applied Genetics*, 133(3):951–966.

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, 9:1660.

Lara-Cabrera, S. I. and Spooner, D. M. (2004). Taxonomy of North and Central American diploid wild potato (Solanum sect. Petota) species: AFLP data. *Plant Systematics and Evolution*, 248(1-4):129–142.

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., and Others (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. *The Plant Journal*, 94(3):562–570.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

PGSC (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432.

Riaño-Pachón, D. M. and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research*, 6:861.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Smit, A. and Hubley, R. (2019). RepeatModeler-1.0. 11. *Institute for Sys-tems Biology. http://www. repeatmasker. org/RepeatModeler/. Accessed*, 15.

Smit, A. F. A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. 2013–2015.

Watanabe, K. (2015). Potato genetics, genomics, and applications. *Breeding science*, 65(1):53–68.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome research*, 27(5):757–767.

Yeo, S., Coombe, L., Warren, R. L., Chu, J., and Birol, I. (2017). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34(5):725–731.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677.

Contribution to knowledge

6.1 Contributions from Chapter 2

This was the first published review summarizing the current available strategies/approaches for genome sequence assembly of polyploid plant genomes. First it gives an overview of the sequencing techniques used in polyploid plant genomes, mentioning all the polyploid plant genomes sequenced until November 2018. Then, it goes through the challenges of polyploid genome assembly, it describes approaches on how to estimate ploidy levels and mentions advances in genomic resources and functional tools.

6.2 Contributions from Chapter 3

The hypothesis for Chapter 3 was that the genes in regions of structural variations (SVs) in six diploid, two triploid, three tetraploid and a pentaploid potato genomes are primarily involved in defense mechanisms and tuber formation and the genomic regions of variation and higher heterozygosity are conserved between genomes. Regarding the first part of the hypothesis, hypothesis, we found genes with unknown function, and genes involved in disease resistance and abiotic stress tolerance in regions of SVs. Moreover, some potato genomes contain genes impacted with SVs that are involved in Self Incompatibility. As for the 2nd part of the hypothesis,, we found that several genomes share numerous of genes in common that are impacted by SVs and some of the genomes show greater heterozygosity in the same chromosomes; i.e. chromosomes 1 and 12. This study used 12 selected potato accessions from CGIAR's potato germplasm collection at the International Potato Center (CIP) in Lima, Peru, to represent the two principal views on potato taxonomy. The structural variation of these eleven native Peruvian and one native Chilean landraces was detected when compared to the two publicly available potato reference genomes; DM1-3 and M6. The ploidy in these genomes varies and previously no study of this nature was reported in potatoes of various ploidy levels. A phylogeny

was performed using the CNV status of the genes, and the results support both older and more recent potato taxonomy classifications. Moreover, specific CNV – impacted regions with high gene density were identified. Among others, abiotic stress and disease resistance genes are located in these conserved CNV-impacted regions across the twelve potato genomes.

6.3 Contribution from Chapter 4

For Chapter 4, we hypothesized that a complete diploid potato pan-genome available will reveal disease important genes to improve prospects for breeding cultivars with disease and climate change resistance. When examining our 1st hypothesis, we found that Long and Link sequencing data improved dramatically the de novo genome assembly of the GON1 genome, that has an N50 3.5 times bigger than the genomes assembled only with shorter Illumina reads and it contains 34 times fewer contigs than the other assemblies. Furthermore, regarding the 2nd hypothesis, we uncovered a diploid potato genome that consists of 39,751 genes, of which the 723 are newly identified and they are absent from the current DM1-3 genome annotation. Within these genes, we discovered a core genome of 28,208 genes and an accessory genome of 11,543 genes. Finally, for the 3rd hypothesis, we found genes with important agronomically important traits displaying presence/absence variation. These genes are involved in disease resistance, biotic and abiotic tolerance, self-incompatibility, tuber shape, secondary metabolite biosynthesis, tuberization and others.

The tuber-producing Solanum clade is large and there is incredible depth in the variation that can be tapped for genetic improvement of potato, which is of increasing importance in the face of climate change. Potato genomes are complicated due to high heterozygosity and autopolyploidy in many species, which is why most work has been carried out in lab clones. This manuscript describes the de novo sequencing and pan-genome assembly of diploid potato genomes from five cultivated landraces and a wild species that are part of the landrace collection at the genebank in the International Potato Research Cen-

ter (CIP) in Lima, Peru. The study describes the genomes of taxonomically diverse, real world potato landraces/species and by so doing can circumvent important shortcomings of previous studies/reference genome, such as important genes not included in reference genomes. The study is the first to describe a draft genome of a cultivated diploid potato (*S. stenotomum* subsp. *goniocalyx*) and the first to apply a pan-genome assembly in potato. This is an important advance for potato genomics. The pan-genome assembly enabled a genome-wide comparative analysis between diploid, cultivated, tuber-bearing species and wild relatives. The assembly of the accessory genome led to identification of significant variation in genomic regions carrying key genes controlling self-incompatibility, biotic stress resistance, secondary metabolite biosynthesis, tuberization, and flowering and tuber shape, all functions of high interest to potato breeding programs. The current study has taken potato genomics beyond the single reference genome for the first time and demonstrated the power of the pan-genome in uncovering novel regions of variation in critical genes of interest to researchers and breeders.

6.4 Contribution from Chapter 5

Finally, for Chapter 5 we hypothesized that the Third Generation Sequencing technology data will improve the de novo genome assembly of the polyploid potato genomes. We found that indeed the availability of longer sequencing data improves the genome assembly in the tetraploid ADG1 genome, where the N50 is 16.5 times larger than the best assembled genome with only shorter Illumina reads and that the percentage of fragmented genes is at least 3 times smaller than the other genomes. This manuscript describes the sequencing and genome assembly of six taxonomically diverse polyploid potato genomes from the genebank at the International Potato Research Center (CIP) in Lima, Peru. Genome assemblies of six potato genomes of various ploidy (two triploids, three tetraploids and a pentaploid genome) were constructed. For one of the genomes, (ADG1), Long and Linked information data was used. This study shows how complicated the polyploid genome assembly problem is with highly heterozygous species and how the availability of Third

Generation Sequencing data aids the process. These genomes have not been previously sequenced, although the S. tuberosum is intensely studied (sequencing is underway). The study concludes that these technologies are not enough, and that new polyploid-specific algorithms and new technologies are needed.

Future Research Directions

7.1 Conclusions

The genome comparison of the diploid and polyploid genomes uncovered conserved regions with variation (CNVs) related to plant defense and disease resistance, as well as variation in other important traits. The diploid pan-genome construction unraveled a core and an accessory genome, including genes absent from the DM1-3 reference genome. These genes are key to important breeding traits such as disease resistance and defense, organ development, and fertility. The long and linked reads improved the de novo genome assembly of one of the polyploids compared to the ones assembled with only short reads (Illumina PE). There is nonetheless a major need for improved sequencing technologies and algorithms.

The following directions could be taken to go further with the research reported in this thesis:

7.2 Chapter 3

The results of the copy number variation analysis revealed the CNV-impacted genes in all the genomes compared to the DM1-3 and M6 genomes. Previous studies have shown that CNV events can alter gene expression, for example, they result in nematode resistance in soybean (Cook et al., 2014), in maize they confer tolerance to aluminium (Maron et al., 2013) and they can have an impact on the growth and development of potato plants (Iovene et al., 2013). Based on the findings in this chapter, it would be interesting to study the effects of CNV on transcript abundance, how the CNVs affect gene expression in these native South American and whether the top CNV-impacted gene clusters have altered

gene expression. Moreover, a great heterozygosity was observed in the genomes compared to the references, which also causes differential allele expression. Preferential allele expression (PAE) can cause phenotypic variations. Previously, PAE has been detected in barley in relation to developmental variation and drought stress (Von Korff et al., 2009), a genome-dependent allele-specific expression detected in rice genes (Song et al., 2013) and in maize (Springer and Stupar, 2007). Hence, taking this great heterozygosity found in the twelve genomes analyzed in this study into account, it would be of a great importance to study if there is any preferential allele expression and what are the potential phenotypes.

7.3 Chapter 4

A main problem in eukaryotic pan-genome studies is the lack of visualization tools. Although many of them have been developed for prokaryotes (Clarke et al., 2018; Ding et al., 2017; Peng et al., 2018), there are no eukaryote pan-genome visualization tools to date. A tool for plants is necessary as it can aid evolutionary studies, but also the comparison not only between different species, but also between different haplotypes of the same genome. Additionally, it would also be interesting to add genomes of higher ploidy levels in the pan-genome to observe the gene repertoire.

7.4 Chapter 5

For this chapter, the genomes of six polyploid genomes were assembled. However, only one of the assemblies, ADG1, was sufficiently assembled to map to pseudomolecules, as this was the only one for which we had Linked and Long sequencing reads. The remaining five genomes could be potentially sequenced using even newer technologies, such as longer scaffolders for the construction of physical maps, using very large DNA fragments for higher contiguity. Hopefully this will reduce the polyploidy problem.

Master List of References

(2019). Picard toolkit. \url{http://broadinstitute.github.io/picard/}.

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984.

Aguiar, D. and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–i360.

Antonious, G. F. (2001). PRODUCTION AND QUANTIFICATION OF METHYL KETONES IN WILD TOMATO ACCESSIONS. *Journal of Environmental Science and Health, Part B*, 36(6):835–848.

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., Himmelbach, A., Balfourier, F., Gutierrez-Gonzalez, J., Hayden, M., Koh, C., Muehlbauer, G., Pasam, R. K., Paux, E., Rigault, P., Tibbits, J., Tiwari, V., Spannagl, M., Lang, D., Gundlach, H., Haberer, G., Mayer, K. F. X., Ormanbekova, D., Prade, V., Šimková, H., Wicker, T., Swarbreck, D., Rimbert, H., Felder, M., Guilhot, N., Kaithakottil, G., Keilwagen, J., Leroy, P., Lux, T., Twardziok, S., Venturini, L., Juhász, A., Abrouk, M., Fischer, I., Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., Arnaud, D., Chalabi, S., Chalhoub, B., Cory, A., Datla, R., Davey, M. W., Jacobs, J., Robinson, S. J., Steuernagel, B., van Ex, F., Wulff, B. B. H., Benhamed, M., Bendahmane, A., Concia, L., Latrasse, D., Bartoš, J., Bellec, A., Berges, H., Doležel, J., Frenkel, Z., Gill, B., Korol, A., Letellier, T., Olsen, O.-A., Singh, K., Valárik, M., van der Vossen, E., Vautrin, S., Weining, S., Fahima, T., Glikson, V., Raats, D., Č\'\ihal\'\iková, J., Toegelová, H., Vrána, J., Sourdille, P., Darrier, B., Barabaschi, D., Cattivelli, L., Hernandez, P., Galvez, S., Budak, H., Jones, J. D. G., Witek, K., Yu, G., Small, I., Melonek, J., Zhou, R., Belova, T., Kanyuka, K., King, R., Nilsen, K., Walkowiak,

S., Cuthbert, R., Knox, R., Wiebe, K., Xiang, D., Rohde, A., Golds, T., Č\'\ižková, J., Akpinar, B. A., Biyiklioglu, S., Gao, L., N\textquoterightDaiye, A., Kubaláková, M., Šafá\v r, J., Alfama, F., Adam-Blondon, A.-F., Flores, R., Guerche, C., Loaec, M., Quesneville, H., Condie, J., Ens, J., Maclachlan, R., Tan, Y., Alberti, A., Aury, J.-M., Barbe, V., Couloux, A., Cruaud, C., Labadie, K., Mangenot, S., Wincker, P., Kaur, G., Luo, M., Sehgal, S., Chhuneja, P., Gupta, O. P., Jindal, S., Kaur, P., Malik, P., Sharma, P., Yadav, B., Singh, N. K., Khurana, J. P., Chaudhary, C., Khurana, P., Kumar, V., Mahato, A., Mathur, S., Sevanthi, A., Sharma, N., Tomar, R. S., rina Holušová, K., rej Pl\'\ihal, O., Clark, M. D., Heavens, D., Kettleborough, G., Wright, J., Balcárková, B., Hu, Y., Salina, E., Ravin, N., Skryabin, K., Beletsky, A., Kadnikov, V., Mardanov, A., Nesterov, M., Rakitin, A., Sergeeva, E., Handa, H., Kanamori, H., Katagiri, S., Kobayashi, F., Nasuda, S., Tanaka, T., Wu, J., Cattonaro, F., Jiumeng, M., Kugler, K., Pfeifer, M., Sandve, S., Xun, X., Zhan, B., Batley, J., Bayer, P. E., Edwards, D., Hayashi, S., Tulpová, Z., Visendi, P., Cui, L., Du, X., Feng, K., Nie, X., Tong, W., and Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403).

AU - van Berkum, N. L., AU - Lieberman-Aiden, E., AU - Williams, L., AU - Imakaev, M., AU - Gnirke, A., AU - Mirny, L. A., AU - Dekker, J., and AU - Lander, E. S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *JoVE*, (39):e1869.

Augusto Corrêa dos Santos, R., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics*, 33(16):2575–2576.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Dal Molin, A., Avanzato, C., Ferrarini, A., and Others (2015). The Solanum commersonii genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *The Plant Cell*, 27(4):954–968.

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R. A., and Chen, M. (2016). The impact and origin of copy number variations in the Oryza species. *BMC Genomics*, 17(1):261.

Bamberg, J. B., Hanneman, R. E., and Towill, L. E. (1986). Use of activated charcoal to enhance the germination of botanical seeds of potato. *American potato journal*, 63(4):181–189.

Bansal, V. and Bafna, V. (2008). HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159.

Barandalla, L., Álvarez, A., de Galarreta, J. I. R., and Ritter, E. (2018). Identification of candidate genes involved in the response to different abiotic stresses in potato (Solanum tuberosum L.). *Revista Latinoamericana de la Papa*, 22(2):33–38.

Bastiaansen, J. W. M., Coster, A., Calus, M. P. L., van Arendonk, J. A. M., and Bovenhuis, H. (2012). Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution*, 44(1):3.

Bell, J. M., Lau, B. T., Greer, S. U., Wood-Bouwens, C., Xia, L. C., Connolly, I. D., Gephart, M. H., and Ji, H. P. (2017). Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Research*, 45(19):e162–e162.

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2009). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics*, 120(2):355.

Bento, M., Gustafson, J. P., Viegas, W., and Silva, M. (2011). Size matters in Triticeae polyploids: larger genomes have higher remodeling. *Genome*, 54(3):175–183.

Bentolila, S., Alfonso, A. A., and Hanson, M. R. (2002). A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences*, 99(16):10887–10892.

Bergelson, J., Kreitman, M., Stahl, E. A., and Tian, D. (2001). Evolutionary Dynamics of Plant R-Genes. *Science*, 292(5525):2281–2285.

Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. In *International Conference on Research in Computational Molecular Biology*, pages 18–19. Springer.

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., Araújo, A. C. G., Kozik, A., Do Kim, K., Burow, M. D., Varshney, R. K., Wang, X., Zhang, X., Barkley, N., Guimarães, P. M., Isobe, S., Guo, B., Liao, B., Stalker, H. T., Schmitz, R. J., Scheffler, B. E., Leal-Bertioli, S. C. M., Xun, X., Jackson, S. A., Michelmore, R., and Ozias-Akins, P. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics*, 48(4):438–446.

Besnard, G., Rubio de Casas, R., and Vargas, P. (2007). Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (Olea europaea). *Journal of Biogeography*, 34(4):736–752.

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature*, 543(7645):346–354.

Beyaz, R., Alizadeh, B., Gürel, S., Fatih Özcan, S., and Yildiz, M. (2013). Sugar beet (Beta vulgaris L.) growth at different ploidy levels. *Caryologia*, 66(1):90–95.

Birch, P. R. J., Bryan, G., Fenton, B., Gilroy, E. M., Hein, I., Jones, J. T., Prashar, A., Taylor, M. A., Torrance, L., and Toth, I. K. (2012). Crops that feed the world 8: Potato: are the trends of increased global production sustainable? *Food Security*, 4(4):477–508.

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

Bradeen, J. M., Haynes, K. G., and Kole, C. (2011). Introduction to potato. *Genetics, Genomics and Breeding of Potatoes. Eds. JM Bradeen, KG Haynes. Enfield, NH: Sci. Publ*, pages 1–19.

Bradshaw, J. E. (2007). Potato-breeding strategy. In *Potato biology and biotechnology*, pages 157–177. Elsevier.

Bradshaw, J. E., Bryan, G. J., and Ramsay, G. (2006). Genetic resources (including wild and cultivated Solanum species) and progress in their utilisation in potato breeding. *Potato Research*, 49(1):49–65.

Butts, C. T., Bierma, J. C., and Martin, R. W. (2016). Novel proteases from the genome of the carnivorous plant Drosera capensis: Structural prediction and comparative analysis. *Proteins: Structure, Function, and Bioinformatics*, 84(10):1517–1533.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10):956–963.

Cardi, T., D'Ambrosio, E., Consoli, D., Puite, K. J., and Ramulu, K. S. (1993). Production of somatic hybrids between frost-tolerant Solanum commersonii and S. tuberosum: characterization of hybrid plants. *Theoretical and Applied Genetics*, 87(1):193–200.

Carputo, D., Barone, A., Cardi, T., Sebastiano, A., Frusciante, L., and Peloquin, S. J. (1997). Endosperm balance number manipulation for direct in vivo germplasm introgression to potato from a sexually isolated relative (Solanum commersonii Dun.). *Proceedings of the National Academy of Sciences*, 94(22):12013–12017.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., Corréa, M., Da Silva, C., Just, J., Falentin, C., Koh, C. S., Le Clainche, I., Bernard, M., Bento, P., Noel, B., Labadie, K., Alberti, A., Charles, M., Arnaud, D., Guo, H., Daviaud, C., Alamery, S., Jabbari, K., Zhao, M., Edger, P. P., Chelaifa,

H., Tack, D., Lassalle, G., Mestiri, I., Schnel, N., Le Paslier, M.-C., Fan, G., Renault, V., Bayer, P. E., Golicz, A. A., Manoli, S., Lee, T.-H., Thi, V. H. D., Chalabi, S., Hu, Q., Fan, C., Tollenaere, R., Lu, Y., Battail, C., Shen, J., Sidebottom, C. H. D., Wang, X., Canaguier, A., Chauveau, A., Bérard, A., Deniot, G., Guan, M., Liu, Z., Sun, F., Lim, Y. P., Lyons, E., Town, C. D., Bancroft, I., Wang, X., Meng, J., Ma, J., Pires, J. C., King, G. J., Brunel, D., Delourme, R., Renard, M., Aury, J.-M., Adams, K. L., Batley, J., Snowdon, R. J., Tost, J., Edwards, D., Zhou, Y., Hua, W., Sharpe, A. G., Paterson, A. H., Guan, C., and Wincker, P. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, 345(6199):950–953.

Chen, A. and Dubcovsky, J. (2012). Wheat TILLING mutants show that the vernalization gene VRN1 down-regulates the flowering repressor VRN2 in leaves but is not essential for flowering. *PLoS genetics*, 8(12):e1003134–e1003134.

Chen, Y., Hao, X., and Cao, J. (2014). Small auxin upregulated RNA (SAUR) gene family in maize: Identification, evolution, and its phylogenetic comparison with Arabidopsis, rice, and sorghum. *Journal of Integrative Plant Biology*, 56(2):133–150.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in plant science*, 15(2):57–71.

Cheng, B., Furtado, A., and Henry, R. J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *GigaScience*, 6(11).

Chong, J., Baltz, R., Schmitt, C., Beffa, R., Fritig, B., and Saindrenan, P. (2002). Downregulation of a Pathogen-Responsive Tobacco UDP-Glc:Phenylpropanoid Glucosyltransferase Reduces Scopoletin Glucoside Accumulation, Enhances Oxidative Stress, and Weakens Virus Resistance. *The Plant Cell*, 14(5):1093–1107.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.-C., Magdelenat, G., Gonthier, C., and Others (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell*, 22(6):1686–1701.

Christiansen, J. A. (1977). *The utilization of bitter potatoes to improve food production in the high altitude of the tropics*. Cornell University, Jan.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92.

CIP (2018). International Potato Center.

Clarindo, W. R., de Carvalho, C. R., Araújo, F. S., de Abreu, I. S., and Otoni, W. C. (2008). Recovering polyploid papaya in vitro regenerants as screened by flow cytometry. *Plant Cell, Tissue and Organ Culture*, 92(2):207–214.

Clarke, T. H., Brinkac, L. M., Inman, J. M., Sutton, G., and Fouts, D. E. (2018). PanACEA: a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. *BMC Bioinformatics*, 19(1):246.

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, 1(2):439–459.

Clevenger, J. P. and Ozias-Akins, P. (2015). SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. *G3: Genes, Genomes, Genetics*, 5(9):1797–1803.

Cook, D. E., Bayless, A. M., Wang, K., Guo, X., Song, Q., Jiang, J., and Bent, A. F. (2014). Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode. *Plant Physiology*, 165(2):630–647.

Corentin Clot (2020). The Origin and Widespread Occurrence of Sli based Self-Compatibility in Potato. San Diego.

Costa, M.-C. D., Artur, M. A. S., Maia, J., Jonkheer, E., Derks, M. F. L., Nijveen, H., Williams, B., Mundree, S. G., Jiménez-Gómez, J. M., Hesselink, T., and Others (2017).

A footprint of desiccation tolerance in the genome of Xerophyta viscosa. *Nature plants*, 3(4):1–10.

Cousin, A. and Nelson, M. N. (2009). Twinned microspore-derived embryos of canola (Brassica napus L.) are genetically identical. *Plant Cell Reports*, 28(5):831–835.

Crow, K. D. and Wagner, G. P. (2005). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular biology and evolution*, 23(5):887–892.

Darrow, G. M. (1966). *The Strawberry.* Holt, Rinehart and Winston New York.

Dart, S., Kron, P., and Mable, B. K. (2004). Characterizing polyploidy in Arabidopsis lyrata using chromosome counts and flow cytometry. *Canadian Journal of Botany*, 82(2):185–197.

Das, S. and Vikalo, H. (2015). SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, 16(1):260.

Davis, T. M., Denoyes-Rothan, B., and Lerceteau-Köhler, E. (2007). Strawberry. In *Fruits and nuts*, pages 189–205. Springer.

Ding, W., Baumdicker, F., and Neher, R. A. (2017). panX: pan-genome analysis and exploration. *Nucleic Acids Research*, 46(1):e5–e5.

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T. R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, P. F., Schmidt, T., Gabaldón, T., Lehrach, H., Weisshaar, B., and Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*, 505(7484):546–549.

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., and Wendel, J. F. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annual review of genetics*, 42:443–461.

Dunnett, J. M. (1957). Variation in pathogenicity of the potato root eelworm (Heterodera rostochiensis woll.,) and its significance in potato breeding. *Euphytica*, 6(1):77–89.

Eaton, T. D., Curley, J., Williamson, R. C., and Jung, G. (2004). Determination of the Level of Variation in Polyploidy among Kentucky Bluegrass Cultivars by Means of Flow Cytometry Research funded by a grant from the United States Golf Association. *Crop Science*, 44:2168–2174.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research*, 27(5):801–812.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., Callow, P., McKain, M. R., Shi, J., Collier, C., Xiong, Z., Mower, J. P., Slovin, J. P., Hytönen, T., Jiang, N., Childs, K. L., and Knapp, S. J. (2017). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *GigaScience*, 7(2).

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133 LP – 138.

Ellis, D., Chavez, O., Coombs, J., Soto, J., Gomez, R., Douches, D., Panta, A., Silvestre, R., and Anglin, N. L. (2018). Genetic identity in genebanks: application of the SolCAP 12K SNP array in fingerprinting and diversity analysis in the global in trust potato collection. *Genome*, 61(7):523–537.

Engel, F. (1970). Exploration of the Chilca Canyon, Peru. *Current Anthropology*, 11(1):55–58.

FAO (2013). No Title.

Felsenstein, J. (1993). *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

Fogelman, E., Oren-Shamir, M., Hirschberg, J., Mandolino, G., Parisi, B., Ovadia, R., Tanami, Z., Faigenboim, A., and Ginzberg, I. (2019). Nutritional value of potato (Solanum tuberosum) in hot climates: anthocyanins, carotenoids, and steroidal glycoalkaloids. *Planta*, 249(4):1143–1155.

Fridman, E., Wang, J., Iijima, Y., Froehlich, J. E., Gang, D. R., Ohlrogge, J., and Pichersky, E. (2005). Metabolic, Genomic, and Biochemical Analyses of Glandular Trichomes from the Wild Tomato Species Lycopersicon hirsutum Identify a Key Enzyme in the Biosynthesis of Methylketones. *The Plant Cell*, 17(4):1252–1267.

Gálvez, J. H., Tai, H. H., Lagüe, M., Zebarth, B. J., and Strömvik, M. V. (2016). The nitrogen responsive transcriptome in potato (Solanum tuberosum L.) reveals significant gene regulatory motifs. *Scientific reports*, 6:26090.

Gálvez Helen H., Barkley, Noelle A., Gardner, Kyle, Ellis, David, Strömvik, Martina V., José Héctor, T. (2017). Understanding potato with the help of genomics. *AIMS Agriculture and Food*, 2(1):16–39.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.

Gavrilenko, T., Antonova, O., Shuvalova, A., Krylova, E., Alpatyeva, N., Spooner, D. M., and Novikova, L. (2013). Genetic diversity and origin of cultivated potatoes based on plastid microsatellite polymorphism. *Genetic Resources and Crop Evolution*, 60(7):1997–2015.

Gebhardt, C. and Valkonen, J. P. T. (2001). ORGANIZATION OF GENES CONTROLLING DISEASE RESISTANCE IN THE POTATO GENOME. *Annual Review of Phytopathology*, 39(1):79–102.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? *Trends in Plant Science*, 21(7):609–621.

GRUNDT, H. H., OBERMAYER, R., and BORGEN, L. I. V. (2005). Ploidal levels in the arctic-alpine polyploid Draba lactea (Brassicaceae) and its low-ploid relatives. *Botanical Journal of the Linnean Society*, 147(3):333–347.

Gruzdev, E.V., Beletsky, A.V., Mardanov, A.V., Kochieva, E. and Ravin, N.V. and Skryabin, K. (2018). Genome of Monotropa hypopitys. *Unpublished*.

Gupta, S. K., Rai, A. K., Kanwar, S. S., and Sharma, T. R. (2012). Comparative analysis of zinc finger proteins involved in plant disease resistance. *PloS one*, 7(8):e42578–e42578.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Hagen, G. and Guilfoyle, T. (2002). Auxin-responsive gene expression: genes, promoters and regulatory factors. *Plant Molecular Biology*, 49(3):373–385.

Harbaugh, D. (2008). Polyploid and Hybrid Origins of Pacific Island Sandalwoods (Santalum, Santalaceae) Inferred from Low-Copy Nuclear and Flow Cytometry Data. *International Journal of Plant Sciences*, 169(5):677–685.

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., and Others (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. *The Plant Cell*, 28(2):388–405.

Hardigan, M. A., Laimbeer, F. P. E., Newton, L., Crisovan, E., Hamilton, J. P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J. C., Douches, D. S., Farré, E. M., Veilleux, R. E., and

Buell, C. R. (2017). Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, 114(46):E9999—-E10008.

Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., Poveda, L., Shimizu-Inatsugi, R., Baeten, J., Francoijs, K.-J., Nataraja, K. N., Reddy, Y. A. N., Phadnis, S., Ravikumar, R. L., Schlapbach, R., Sreeman, S. M., and Shimizu, K. K. (2017). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Research*, 25(1):39–47.

Hawkes, J. G. (1958). Significance of wild species and primitive forms for potato breeding. *Euphytica*, 7(3):257–270.

Hawkes, J. G. and Others (1990). *The potato: evolution, biodiversity and genetic resources.* Belhaven Press.

Hijmans, R. J. and Spooner, D. M. (2001). Geographic distribution of wild potato species. *American Journal of Botany*, 88(11):2101–2112.

Hirsch, C. D., Hamilton, J. P., Childs, K. L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C. N., Habermann, M., Neal, B., and Buell, C. R. (2014a). Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding. *The Plant Genome*, 7.

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., and Buell, C. R. (2014b). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, 26(1):121–135.

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R. E., Jansky, S., Bethke, P., Douches, D. S., and Buell, C. R. (2013). Retrospective View of North American Potato (Solanum tuberosum L.) Breeding in the 20th and 21st Centuries. *G3: Genes, Genomes, Genetics*, 3(6):1003–1013.

Hittalmani, S., Mahesh, H. B., Shirke, M. D., Biradar, H., Uday, G., Aruna, Y. R., Lohithaswa, H. C., and Mohanrao, A. (2017). Genome and Transcriptome sequence of Finger millet (Eleusine coracana (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC Genomics*, 18(1):465.

Hosaka, K. (1995). Successive domestication and evolution of the Andean potatoes as revealed by chloroplast DNA restriction endonuclease analysis. *Theoretical and Applied Genetics*, 90(3):356–363.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., and Others (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, 43(5):476.

Huamán, Z. and Spooner, D. M. (2002). Reclassification of landrace populations of cultivated potatoes (Solanum sect. Petota). *American Journal of Botany*, 89(6):947–965.

Huang, K., Ritland, K., Guo, S., Shattuck, M., and Li, B. (2014). A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, 14(4):734–744.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., Zhang, L., Niu, X., Zhang, X., Meng, M., and Others (2013). Draft genome of the kiwifruit Actinidia chinensis. *Nature communications*, 4(1):1–9.

Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., Hill, T. A., Jaffe, D., Williams, S. R., Weisenfeld, N., Ramakrishnan, S., Kumar, V., Shah, P., Schatz, M. C., Church, D. M., and Van Deynze, A. (2018). Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Horticulture Research*, 5(1):4.

Huson, D. H., Reinert, K., and Myers, E. W. (2002). The greedy path-merging algorithm for contig scaffolding. *Journal of the ACM (JACM)*, 49(5):603–615.

Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815.

Iovene, M., Zhang, T., Lou, Q., Buell, C. R., and Jiang, J. (2013). Copy number variation in potato – an asexually propagated autotetraploid species. *The Plant Journal*, 75(1):80–89.

Jain, M. and Khurana, J. P. (2009). Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. *The FEBS Journal*, 276(11):3148–3162.

Jamal, F., Pandey, P. K., Singh, D., and Khan, M. Y. (2013). Serine protease inhibitors in plants: nature's arsenal crafted for insect predators. *Phytochemistry Reviews*, 12(1):1–34.

Jansky, S. H., Chung, Y. S., and Kittipadukal, P. (2014). M6: A Diploid Potato Inbred Line for Use in Breeding and Genetics Research. *Journal of Plant Registrations*, 8:195–199.

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., Kharbatia, N. M., Rupper, R. R., Sharp, A. R., Dally, N., Boughton, B. A., Woo, Y. H., Gao, G., Schijlen, E. G. W. M., Guo, X., Momin, A. A., Negrão, S., Al-Babili, S., Gehring, C., Roessner, U., Jung, C., Murphy, K., Arold, S. T., Gojobori, T., van der Linden, C. G., van Loo, E. N., Jellen, E. N., Maughan, P. J., and Tester, M. (2017). The genome of Chenopodium quinoa. *Nature*, 542(7641):307–312.

Jiao, W.-B. and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36:64–70.

Johns, T., Huaman, Z., Ochoa, C., and Schmiediche, P. E. (1987). Relationships among Wild, Weed, and Cultivated Potatoes in the Solanum x Ajanhuiri Complex. *Systematic Botany*, 12(4):541–552.

Johnston, S. A. and Hanneman, R. E. (1980). Support of the endosperm balance number hypothesis utilizing some tuber-bearingSolanum species. *American Potato Journal*, 57(1):7–14.

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., Spillane, C., Robinson, S. J., Links, M. G., Clarke, C., Higgins, E. E., Huebert, T., Sharpe, A. G., and Parkin, I.

A. P. (2014). The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, 5(1):3706.

Kawai, Y., Ono, E., and Mizutani, M. (2014). Evolution and diversity of the 2–oxoglutarate-dependent dioxygenase superfamily in plants. *The Plant Journal*, 78(2):328–343.

Kim, Y., Oh, Y. J., Han, K. Y., Kim, G. H., Ko, J., and Park, J. (2019). The complete chloroplast genome sequence of Hibiscus syriacus L. 'Mamonde' (Malvaceae). *Mitochondrial DNA Part B*, 4(1):558–559.

Kloppenburg, J. and Kleinman, D. L. (1987). The plant germplasm controversy. *Bioscience*, 37(3):190–198.

Kyriakidou, M., Achakkagari, S. R., López, J. H. G., Zhu, X., Tang, C. Y., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2020a). Structural genome analysis in cultivated potato taxa. *Theoretical and Applied Genetics*, 133(3):951–966.

Kyriakidou, M., Anglin, N. L., Ellis, D., Tai, H. H., and Strömvik, M. V. (2020b). Genome assembly of six polyploid potato genomes. *Scientific Data*, 7(1):1–6.

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, 9:1660.

Lan, T., Renner, T., Ibarra-Laclette, E., Farr, K. M., Chang, T.-H., Cervantes-Pérez, S. A., Zheng, C., Sankoff, D., Tang, H., Purbojati, R. W., and Others (2017). Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proceedings of the National Academy of Sciences*, 114(22):E4435—-E4441.

Lara-Cabrera, S. I. and Spooner, D. M. (2004). Taxonomy of North and Central American diploid wild potato (Solanum sect. Petota) species: AFLP data. *Plant Systematics and Evolution*, 248(1-4):129–142.

Le Roy, J., Huss, B., Creach, A., Hawkins, S., and Neutelings, G. (2016). Glycosylation Is a Major Regulator of Phenylpropanoid Availability and Biological Activity in Plants. *Frontiers in Plant Science*, 7:735.

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., and Others (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. *The Plant Journal*, 94(3):562–570.

Levy, D. and Veilleux, R. E. (2007). Adaptation of potato to high temperatures and salinity-a review. *American Journal of Potato Research*, 84(6):487–506.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., Ma, Z., Shang, H., Ma, X., Wu, J., Liang, X., Huang, G., Percy, R. G., Liu, K., Yang, W., Chen, W., Du, X., Shi, C., Yuan, Y., Ye, W., Liu, X., Zhang, X., Liu, W., Wei, H., Wei, S., Huang, G., Zhang, X., Zhu, S., Zhang, H., Sun, F., Wang, X., Liang, J., Wang, J., He, Q., Huang, L., Wang, J., Cui, J., Song, G., Wang, K., Xu, X., Yu, J. Z., Zhu, Y., and Yu, S. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology*, 33(5):524–530.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, X., van Eck, H. J., Rouppe van der Voort, J. N. A. M., Huigen, D.-J., Stam, P., and Jacobsen, E. (1998). Autotetraploids and genetic mapping using common AFLP markers: the R2 allele conferring resistance to Phytophthora infestans mapped on potato chromosome 4. *Theoretical and Applied Genetics*, 96(8):1121–1128.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293.

Lischer, H. E. L. and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18(1):474.

Liu, Y., Lin-Wang, K., Deng, C., Warran, B., Wang, L., Yu, B., Yang, H., Wang, J., Espley, R. V., Zhang, J., and Others (2015). Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. *PloS one*, 10(6).

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., and L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448.

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., Hernandez, A. G., Mikel, M. A., Soifer, I., Barad, O., and Buckler, E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, 6(1):6914.

Machida-Hirano, R. (2015). Diversity of potato genetic resources. *Breeding science*, 65(1):26–40.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

Margarido, G. R. A. and Heckerman, D. (2015). ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS computational biology*, 11(4):e1004229–e1004229.

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., Buckler, E. S., Coluccio, A. E., Danilova, T. V., Kudrna, D., Magalhaes, J. V., Piñeros, M. A., Schatz, M. C., Wing, R. A., and Kochian, L. V. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proceedings of the National Academy of Sciences*, 110(13):5241–5246.

Martin, C. and French, E. R. (1977). Reaction of some tuberbearing Solanum species to Pseudomonas solanacearum. In *Proc. Amer. Phytopathol. Soc*, volume 4, page 139.

Massa, A. N., Childs, K. L., and Buell, C. R. (2013). Abiotic and Biotic Stress Responses in Solanum tuberosum Group Phureja DM1-3 516 R44 as Measured through Whole Transcriptome Sequencing. *The Plant Genome*, 6.

Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., and Buell, C. R. (2011). The transcriptome of the reference potato genome Solanum tuberosum Group Phureja clone DM1-3 516R44. *PloS one*, 6(10):e26801–e26801.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564.

McCue, K. F. (2009). Potato Glycoalkaloids, Past Present and Future.

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., and Stupar, R. M. (2012). Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. *Plant Physiology*, 159(4):1295–1308.

Meyers, L. A. and Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206.

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., Loudet, O., Weigel, D., and Ecker, J. R. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, 9(1):541.

Michael, T. P. and VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, 24:71–81.

Micheletto, S., Boland, R., and Huarte, M. (2000). Argentinian wild diploid Solanum species as sources of quantitative late blight resistance. *Theoretical and Applied Genetics*, 101(5-6):902–906.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M.-L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.-K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M.-C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., DePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, 452(7190):991–996.

Mithani, A., Belfield, E. J., Brown, C., Jiang, C., Leach, L. J., and Harberd, N. P. (2013). HANDS: a tool for genome-wide discovery of subgenome- specific base-identity in polyploids. *BMC Genomics*, 14(1):653.

Monat, C., Schreiber, M., Stein, N., and Mascher, M. (2019). Prospects of pan-genomics in barley. *Theoretical and Applied Genetics*, 132(3):785–796.

Mondal, T. K., Rawal, H. C., Gaikwad, K., Sharma, T. R., and Singh, N. K. (2017). First de novo draft genome sequence of Oryza coarctata, the only halophytic species in the genus Oryza. *F1000Research*, 6:1750.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., and Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013.

Moon, J., Parry, G., and Estelle, M. (2004). The Ubiquitin-Proteasome Pathway and Plant Development. *The Plant Cell*, 16(12):3181–3195.

Motazedi, E., de Ridder, D., Finkers, R., and Maliepaard, C. (2017). TriPoly: a haplotype estimation approach for polyploids using sequencing data of related individuals. *bioRxiv*.

Narzisi, G. and Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS one*, 6(4):e19175–e19175.

Oryza Chr3 Short Arm Comparative Sequencing Project (2014). Genome sequencing of Oryza minuta. Technical report.

Ovchinnikova, A., Krylova, E., Gavrilenko, T., Smekalova, T., Zhuk, M., Knapp, S., and Spooner, D. M. (2011). Taxonomy of cultivated potatoes (Solanum section Petota: Solanaceae). *Botanical Journal of the Linnean Society*, 165(2):107–155.

Parkin, I. A. P., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Pires, J. C., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B., and Sharpe, A. G. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology*, 15(6):R77.

Parris, J. K., Ranney, T. G., Knap, H. T., and Baird, W. V. (2010). Ploidy levels, relative genome sizes, and base pair composition in Magnolia. *Journal of the American Society for Horticultural Science*, 135(6):533–547.

Peng, Y., Tang, S., Wang, D., Zhong, H., Jia, H., Cai, X., Zhang, Z., Xiao, M., Yang, H., Wang, J., Kristiansen, K., Xu, X., and Li, J. (2018). MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks. *GigaScience*, 7(11).

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., and Hvidsten, T. R. (2014). Consortium IWGS, Mayer KFX, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, 345:1250091.

PGSC (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189.

Pham, G. M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D. S., and Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *The Plant Journal*, 92(4):624–637.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4):354–366.

Porubsky, D., Garg, S., Sanders, A. D., Korbel, J. O., Guryev, V., Lansdorp, P. M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications*, 8(1):1293.

Qiu, J. (2017). The Echinochloa crus-galli whole genome shotgun (WGS) project. *Unpublished*.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J. and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual review of ecology and systematics*, 29(1):467–501.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Rehman, H. M., Nawaz, M. A., Shah, Z. H., Ludwig-Müller, J., Chung, G., Ahmad, M. Q., Yang, S. H., and Lee, S. I. (2018). Comparative genomic and transcriptomic analyses of Family-1 UDP glycosyltransferase in three Brassica species and Arabidopsis indicates stress-responsive regulation. *Scientific Reports*, 8(1):1875.

Ren, H. and Gray, W. M. (2015). SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. *Molecular Plant*, 8(8):1153–1164.

Reynolds, M. P. and Ewing, E. E. (1989). Heat tolerance in tuber bearingSolanum species: A protocol for screening. *American Potato Journal*, 66(2):63–74.

Riaño-Pachón, D. M. and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid SP80-3280. *F1000Research*, 6:861.

Ristaino, J. B. (2002). Tracking historic migrations of the Irish potato famine pathogen, Phytophthora infestans. *Microbes and infection*, 4(13):1369–1377.

Rodr\'\iguez, F. and Spooner, D. M. (2009). Nitrate reductase phylogeny of potato (Solanum sect. Petota) genomes with emphasis on the origins of the polyploid species. *Systematic Botany*, 34(1):207–219.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Sasaki, T. and Project, I. R. G. S. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052):793–800.

Sato, S., Hirakawa, H., Isobe, S., Fukai, E., Watanabe, A., Kato, M., Kawashima, K., Minami, C., Muraki, A., Nakazaki, N., Takahashi, C., Nakayama, S., Kishida, Y., Kohara, M., Yamada, M., Tsuruoka, H., Sasamoto, S., Tabata, S., Aizu, T., Toyoda, A., Shin-i, T., Minakuchi, Y., Kohara, Y., Fujiyama, A., Tsuchimoto, S., Kajiyama, S., Makigano, E., Ohmido, N., Shibagaki, N., Cartagena, J. A., Wada, N., Kohinata, T., Atefeh, A., Yuasa, S., Matsunaga, S., and Fukui, K. (2010). Sequence Analysis of the Genome of an Oil-Bearing Tree, Jatropha curcas L. *DNA Research*, 18(1):65–76.

Schmid, R., Schuster, S., Steel, M., and Huson, D. (2006). Readsim-a simulator for sanger and 454 sequencing. *View Article PubMed/NCBI*.

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., Bolger, M. E., Alseekh, S., Maß, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., Aury, J.-M., Koren, S., Fernie, A. R., Zamir, D., Bolger, A. M., and Usadel, B. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell*, 29(10):2336–2348.

Schmiediche, P. E., Hawkes, J. G., and Ochoa, C. M. (1980). Breeding of the cultivated potato species Solanum x juzepczukii Buk. and Solanum x curtilobum Juz. etBuk. *Euphytica*, 29(3):685–704.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., and Others (2010). Genome sequence of the palaeopolyploid soybean. *nature*, 463(7278):178–183.

Schmutz, J., Jenkins, J., Grimwood, J., Bertioli, D., Leal-Bertioli, S., Clevenger, J., Michelmore, R., Froenke, L., , Cannon, S.B., Varshney, R., Schleffler, B., Jackson, S., and Ozias-Akins, P. (2018). Genome sequence of Arachis hypogaea, cultivar Tifrunner. *Unpublished*.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346.

Shan, J., Song, W., Zhou, J., Wang, X., Xie, C., Gao, X., Xie, T., and Liu, J. (2013). Transcriptome analysis reveals novel genes potentially involved in photoperiodic tuberization in potato. *Genomics*, 102(4):388–396.

Sharma, S. K., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., Carboni, M. F., D'Ambrosio, J. M., de la Cruz, G., Di Genova, A., Douches, D. S., and Others (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3: Genes, Genomes, Genetics*, 3(11):2031–2047.

Shen, Q., Zhang, L., Liao, Z., Wang, S., Yan, T., Shi, P., Liu, M., Fu, X., Pan, Q., Wang, Y., and Others (2018). The genome of Artemisia annua provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Molecular plant*, 11(6):776–788.

Shi, J. (2018). Chromosome conformation capture resolved near complete genome assembly of proso millet (Panicum miliaceum L.). *Unpublished*.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R.,

Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J.-M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, 43(2):109–116.

Sierro, N., Battey, J. N. D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M. C., and Ivanov, N. V. (2013). Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biology*, 14(6):R60.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Smillie, R. M., Hetherington, S. E., Ochoa, C., and Malagamba, P. (1983). Tolerances of wild potato species from different altitudes to cold and heat. *Planta*, 159(2):112–118.

Smith, A. F. (2012). *Potato: A global history*. Reaktion Books.

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany*, 103(7):1146–1166.

Song, G., Guo, Z., Liu, Z., Cheng, Q., Qu, X., Chen, R., Jiang, D., Liu, C., Wang, W., Sun, Y., Zhang, L., Zhu, Y., and Yang, D. (2013). Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F1 hybrids. *BMC Plant Biology*, 13(1):221.

Spooner, D. M. (2009). DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. *American journal of botany*, 96(6):1177–1189.

Spooner, D. M. and Bamberg, J. B. (1994). Potato genetic resources: sources of resistance and systematics. *American Potato Journal*, 71(5):325–337.

Spooner, D. M., Núñez, J., Trujillo, G., del Rosario Herrera, M., Guzmán, F., and Ghislain, M. (2007). Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. *Proceedings of the National Academy of Sciences*, 104(49):19398–19403.

Springer, N. M. and Stupar, R. M. (2007). Allele-Specific Expression Patterns Reveal Biases and Embryo-Specific Parent-of-Origin Effects in Hybrid Maize. *The Plant Cell*, 19(8):2391–2402.

Tanaka, H., Hirakawa, H., Kosugi, S., Nakayama, S., Ono, A., Watanabe, A., Hashiguchi, M., Gondo, T., Ishigaki, G., Muguerza, M., Shimizu, K., Sawamura, N., Inoue, T., Shigeki, Y., Ohno, N., Tabata, S., Akashi, R., and Sato, S. (2016). Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research*, 23(2):171–180.

TCP-G. (2018). Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135.

Thomas, J. H. and Emerson, R. O. (2009). Evolution of C2H2-zinc finger genes revisited. *BMC Evolutionary Biology*, 9(1):51.

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F. J., and Others (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences*, 114(44):E9413—-E9422.

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., Fukushima, M., Suzuki, S., Tarora, K., Tamaki, M., Sakamoto, M., Terauchi, R., and Matsumura, H. (2016). Draft genome sequence of bitter gourd (Momordica charantia), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Research*, 24(1):51–58.

Varshney, R. K., Roorkiwal, M., and Nguyen, H. T. (2013). Legume Genomics: From Genomic Resources to Molecular Breeding. *The Plant Genome*, 6(3):plantgenome2013.12.0002in.

Von Korff, M., Radovic, S., Choumane, W., Stamati, K., Udupa, S. M., Grando, S., Ceccarelli, S., Mackay, I., Powell, W., Baum, M., and Morgante, M. (2009). Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *The Plant Journal*, 59(1):14–26.

Wang, A., Wang, Z., Li, Z., and Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, 34(12):2019–2028.

Wang, F., Vandepoele, K., and Van Lijsebettens, M. (2012). Tetraspanin genes in plants. *Plant Science*, 190:9–15.

Watanabe, K. (2015). Potato genetics, genomics, and applications. *Breeding science*, 65(1):53–68.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome research*, 27(5):757–767.

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., and Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, 19(1):122.

WILLIAMS, W. G., KENNEDY, G. G., YAMAMOTO, R. T., THACKER, J. D., and BORDNER, J. O. N. (1980). 2-Tridecanone: A Naturally Occurring Insecticide from the Wild Tomato Lycopersicon hirsutum f.glabratum. *Science*, 207(4433):888–889.

Wu, J., Liu, S., He, Y., Guan, X., Zhu, X., Cheng, L., Wang, J., and Lu, G. (2012). Genome-wide analysis of SAUR gene family in Solanaceae species. *Gene*, 509(1):38–50.

Xie, S.-Q., Zhang, X.-M., Han, Y. and Ling, P. (2018). No Santalum album genome assembly using oxford nanopore sequencing technologyTitle. *Unpublished*.

Xu, J.-H., Bennetzen, J. L., and Messing, J. (2011). Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes. *Molecular Biology and Evolution*, 29(2):861–871.

Yang, J. (2016). The genome of allopolyploid Brassica juncea and evidence for homoeolog expression dominance of potential agricultural significance. *Unpublished*.

Yang, J., Moeinzadeh, M.-H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., Liu, G., Zheng, J., Sun, Z., Fan, W., and Others (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature plants*, 3(9):696–703.

Yang, X., Ye, C.-Y., Cheng, Z.-M., Tschaplinski, T. J., Wullschleger, S. D., Yin, W., Xia, X., and Tuskan, G. A. (2011). Genomic aspects of research involving polyploid plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 104(3):387–397.

Yin, D. (2018). Genome of an allotetraploid wild peanut Arachis monticola: a de novo assembly. *Unpublished*.

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F. N., Kamoun, S., Krause, J., and Others (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife*, 2:e00731.

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M., and Jing, H.-C. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, 12(11):R114.

Zhu, Q., Dugardeyn, J., Zhang, C., Takenaka, M., Kühn, K., Craddock, C., Smalle, J., Karampelias, M., Denecke, J., Peters, J., Gerats, T., Brennicke, A., Eastmond, P., Meyer, E. H., and Van Der Straeten, D. (2012). SLO2, a mitochondrial pentatricopeptide repeat protein affecting several RNA editing sites, is required for energy metabolism. *The Plant Journal*, 71(5):836–849.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvo\vrák, J., and Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome

of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research*, 27(5):787–792.

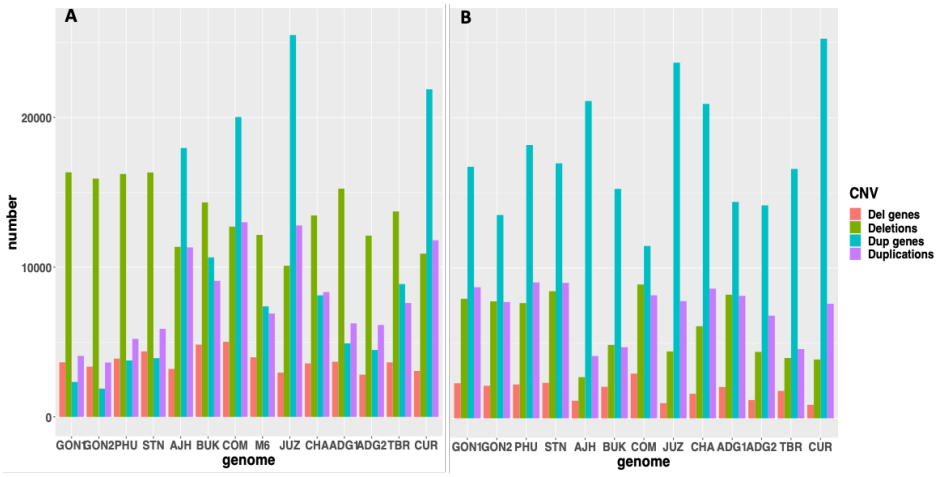


Figure 8.1: Duplications and deletions relative to duplicated and deleted genes in 14 potato genomes. A. The number of genes (of which equal or more than 50% of the gene body was) affected by deletions (red) and duplications (blue) across the 14 genomes, along with the total number of deletions (green) and duplications (purple) against the DM1-3 reference genome. In diploids in general, the number of deleted genes was greater than those affected by duplications with AJH and BUK being the exceptions. In contrast, in the polyploid genomes the number of duplicated genes was greater than the deleted ones, with an exception in ADG2 genome. **B.** The number of the genes affected by deletions (red) and duplications (blue) across the 13 genomes (M6 was not analyzed against M6), along with the total number of deletions (green) and duplications (purple) against the M6 reference genome.

Supplementary Table 9.1: The most heterozygous chromosomes of the genomes when compared to the DM1-3 and M6 genomes.

Genome	Top 3 Heterozygous chromosomes VS DM1-3	Top 3 Heterozygous chromosomes VS M6
GON1	Chr04, Chr01, Chr10	Chr01, Chr12, Chr06
GON2	Chr01, Chr09, Chr07	Chr01, Chr12, Chr06
PHU	Chr01, Chr04, Chr06	Chr01, Chr12, Chr06
STN	Chr01, Chr04, Chr06	Chr12, Chr01, Chr11
AJH	Chr04, Chr09, Chr01	Chr12, Chr01, Chr05
BUK	Chr01, Chr03, Chr04	Chr12, Chr01, Chr03
COM	Chr01, Chr04, Chr10	Chr01, Chr12, Chr11
M6	Chr09, Chr04, Chr08	———————
ADG1	Chr01, Chr05, Chr10	Chr12, Chr01, Chr05
ADG2	Chr01, Chr04, Chr10	Chr01, Chr12, Chr07
TBR	Chr01, Chr07, Chr09	Chr01, Chr12, Chr05
JUZ	Chr01, Chr04, Chr03	Chr01, Chr12, Chr03
CHA	Chr01, Chr12, Chr04	Chr12, Chr01, Chr06
CUR	Chr01, Chr04, Chr09	Chr01, Chr12, Chr05

Supplementary Table 10.1: Summary of the CNVs (deletions and duplications) detected in the A) diploids and B) polyploid genomes against the DM1-3 genome, using CNVnator.

A	**GON1**	**GON2**	**STN**	**PHU**	**AJH**	**BUK**	**COM**	**M6**
Total CNVs	**20,404**	**19,543**	**22,200**	**21,431**	**22,675**	**23,406**	**25,302**	**19,059**
Total deletions	16,331	15,914	16,322	16,221	11,356	14,322	12,705	12,153
Total duplications	4,073	3,629	5,878	5,210	11,319	9,084	12,997	6,906
Genic CNVs (%)	**20.20%**	**17.80%**	**26.70%**	**25%**	**59.60%**	**45.90%**	**71.20%**	**33.70%**
Mean CNV length	**10.5 kb**	**10.1 kb**	**12.2 kb**	**11.6 kb**	**18.4 kb**	**15.8 kb**	**21.1 kb**	**15.5 kb**
Median CNV length	**4 kb**	**3.7 kb**	**5 kb**	**4.9 kb**	**10.2 kb**	**7.2 kb**	**9.8 kb**	**6.4 kb**
Median deletion length	3.6 kb	3.3 kb	4.2 kb	4.1 kb	6.1 kb	5.1 kb	6.2 kb	4.4 kb
Median duplication length	5.5 kb	4.8 kb	7.3 kb	7.4 kb	15.7 kb	11.2 kb	14.3 kb	10.8 kb
Total large CNVs*	**160**	**144**	**333**	**187**	**283**	**345**	**440**	**344**
Size of the largest CNV	**515.2 kb**	**730.1 kb**	**506.8 kb**	**629.9 kb**	**529.5 kb**	**582.5 kb**	**621.2 kb**	**501.1 kb**
Genes affected by deletions	5,202	4,782	6,043	5,616	4,575	6,431	6,690	5,028
Genes affected by duplications	2,684	2,175	4,385	4,134	18,711	11,488	21,099	8,115
Total CNV-affected $genes^{\$}$	**7,886**	**6,957**	**10,428**	**9,750**	**23,286**	**17,919**	**27,789**	**13,143**

Supplementary Table 10.2: Summary of the CNVs (deletions and duplications) detected in the A) diploids and B) polyploid genomes against the DM1-3 genome, using CNVnator.

2B	ADG1	ADG2	TBR	JUZ	CHA	CUR
Total CNVs	**21,489**	**18,243**	**21,323**	**22,875**	**21,790**	**22,694**
Total deletions	15,243	12,105	13,719	10,091	13,452	10,900
Total duplications	6,246	6,138	7,604	12,784	8,338	11,794
Genic CNVs (%)	**26.80%**	**22.40%**	**37.50%**	**77.80%**	**35.30%**	**69%**
Mean CNV length	**11.8 kb**	**12.1 kb**	**14.3 kb**	**21.5 kb**	**15.2 kb**	**19.2 kb**
Median CNV length	**4.7 kb**	**5.1 kb**	**6.7 kb**	**11.6 kb**	**7.6 kb**	**10.4 kb**
Median deletion length	3.8 kb	3.7 kb	4.9 kb	6.4 kb	5.2 kb	6 kb
Median duplication length	7.3 kb	8.1 kb	10.8 kb	17.4 kb	12.3 kb	16.1 kb
Total large CNVs*	**195**	**148**	**221**	**465**	**226**	**317**
Size of the largest CNV	**900.3 kb**	**670.8 kb**	**1.1 Mb**	**594 kb**	**900.3 kb**	**646.8 kb**
Genes affected by deletions	5,003	3,741	5,105	3,972	5,095	4,238
Genes affected by duplications	5,453	5,012	9,537	26,411	8,681	22,684
Total CNV-affected $genes^{\$}$	**10,456**	**8,753**	**14,642**	**30,383**	**13,776**	**26,922**

* Large CNVs are defined as having a length $>$ 100 kb. The length of the CNVs found in ST4.03ch00 are not counted in the length metrics, only the impacted genes are counted in the table.

$^{\$}$ This total includes genes that are affected by both deletions and duplications

Supplementary Table 11.1: Summary of the CNVs (deletions and duplications) detected in the A) diploids and B) polyploid genomes against the M6.

A	**GON1**	**GON2**	**STN**	**PHU**	**AJH**	**BUK**	**COM**
Total CNVs	**16,679**	**16,378**	**17,471**	**16,713**	**6,889**	**9,617**	**17,096**
Total deletions	7,956	8,227	8,458	7,667	2,740	4,886	8.908
Total duplications	8,723	8,151	9,013	9,046	4,149	4,731	8,188
Genic CNVs (%)	**50.50%**	**43.70%**	**51.20%**	**54.20%**	**59%**	**46%**	**38.30%**
Mean CNV length	**19.1 kb**	**17.7 kb**	**18.5 kb**	**20 kb**	**49.7 kb**	**30 kb**	**16.8 kb**
Median CNV length	**10 kb**	**9 kb**	**9.8 kb**	**10.9 kb**	**23.6 kb**	**15.6 kb**	**8.7 kb**
Median deletion length	5.7 kb	5.2 kb	5.4 kb	5.9 kb	9.4 kb	8.4 kb	5.7 kb
Median duplication length	16.5 kb	15.5 kb	16.3 kb	17.3 kb	45.4 kb	28.6 kb	13.9 kb
Total large CNVs*	**328**	**283**	**308**	**373**	**948**	**546**	**245**
Size of the largest CNV	**439 kb**	**584.4 kb**	**568.6 kb**	**434.9 kb**	**745.8 kb**	**1 Mb**	**358 kb**
Genes affected by deletions	2,335	2,085	2,361	2,253	1,172	2,095	2,975
Genes affected by duplications	16,737	14,402	16,967	18,188	21,116	15,267	11,468
Total CNV-affected $genes^{\$}$	**19,072**	**16,487**	**19,328**	**20,441**	**22,288**	**17,362**	**14,443**

Supplementary Table 11.2: Summary of the CNVs (deletions and duplications) detected in the A) diploids and B) polyploid genomes against the M6.

3B	ADG1	ADG2	TBR	JUZ	CHA	CUR
Total CNVs	**16,378**	**11,250**	**8,624**	**12,250**	**14,759**	**11,538**
Total deletions	8,227	4,419	4,016	4,450	6,128	3,915
Total duplications	8,151	6,831	4,608	7,800	8,631	7,623
Genic CNVs (%)	**43.70%**	**40.70%**	**48.80%**	**65.40%**	**59.80%**	**69.30%**
Mean CNV length	**17.3 kb**	**23.3 kb**	**34.7 kb**	**31.9 kb**	**24.5 kb**	**34.8 kb**
Median CNV length	**8.6 kb**	**10.7 kb**	**17.2 kb**	**14.8 kb**	**12.6 kb**	**16.9 kb**
Median deletion length	5.2 kb	5.3 kb	8.8 kb	6.2 kb	6.2 kb	6.4 kb
Median duplication length	14.7 kb	18 kb	31.6 kb	24.4 kb	20.1 kb	28.6 kb
Total large CNVs*	**294**	**437**	**676**	**764**	**547**	**927**
Size of the largest CNV	**371.8 kb**	**560.8 kb**	**838.6 kb**	**5 Mb**	**505.5 kb**	**511.8 kb**
Genes affected by deletions	2,085	1,217	1,826	1,008	1,637	901
Genes affected by duplications	14,402	14,173	16,594	23,665	20,923	25,262
Total CNV-affected genes$	**16,487**	**15,390**	**18,420**	**24,673**	**22,560**	**26,163**

* Large CNVs are defined as having a length ¿ 100 kb. The length of the CNVs found in ST4.03ch00 are not counted in the length metrics, only the impacted genes are counted in the table.

$ This total includes genes that are affected by both deletions and duplications

Supplementary Table 12.1: Top 3 gene enriched CNV bins in the 8 diploid genomes against the DM1-3 and the M6 reference genomes.

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
GON1	• ST4.03ch12[$]: mannan - endo - 1,4, - β mannosidase genes (del) • ST4.03ch04: disease resistance genes (del + dup) • ST4.03ch05: receptor kinase, wall associate kinase coding genes etc. (del)	• M6_v4.1ch01[*]: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch05[$]: genes coding for proteins of unknown function, NB-ARC coding genes, chaperone subunit etc. (dup) • M6_v4.1ch09: 2 – oxoglutarate genes, bHLH etc. (dup) • M6_v4.1ch09: 2 – oxoglutarate genes, bHLH etc. (dup)

Continued on next page

Supplementary Table 12.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
GON2	• ST4.03ch12[$]: mannan – endo – 1,4, - β mannosidase genes (del) • ST4.03ch00: genes of various functions i.e matrix metalloprotease coding gene, MYC1 (dup) • ST4.03ch08: disease resistance genes, R2 (del)	• M6_v4.1ch05: genes coding for flavin – binding proteins, NAD(P), terpenesythase etc. (dup) • M6_v4.1ch07: various genes; including D – mannose binding lectin coding gene, cytochrome P450, etc. (del + dup) • M6_v4.1ch01: Auxin – induced SAUR gene cluster (dup)
PHU	• ST4.03ch11: disease resistance gene cluster, R2 (del + dup) • ST4.03ch02: conserved gene cluster of unknown function (dup) • ST4.03ch04: Auxin - induced SAUR gene cluster (dup)	• M6_v4.1ch11[!]: Auxin – induced SAUR gene cluster (del + dup) • M6_v4.1ch01: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch05[*]: genes coding for hypothetical protein, Rubisco methyltransferase, NB-AC domain containing etc. (dup)

Continued on next page

Supplementary Table 12.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
STN	• ST4.03ch11: disease resistance gene cluster (del) • ST4.03ch04: genes coding for LRR containing proteins, R2 (del + dup) • ST4.03ch12: genes of various functions (del + dup)	• M6_v4.1ch01*: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch07: genes coding for gibberellin 3 – oxidase, pyridoxine biosynthesis, α–β – hydrolases, lectin protein kinase family etc. (dup) • M6_v4.1ch11!: Auxin induced SAUR gene cluster, F-box etc. (dup)
AJH	• ST4.03ch02: genes involved in the carbohydrate metabolic process (dup) • ST4.03ch05: flavonol 4' – sulfotransferase coding gene, late blight resistance etc. (dup) • ST4.03ch06: male sterility MS5, nodulin – 26, auxin regulated coding gene etc. (dup)	• M6_v4.1ch11!: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch01*: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch01: genes coding for NB-ARC, LRR, flavin – binding kelch repeat, F-box etc. (del)

Continued on next page

Supplementary Table 12.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
BUK	• ST4.03ch03: flavonol synthase/ flavone 3 – hydroxylase coding genesetc (dup) • ST4.03ch04: genes coding for alcohol dehydroxygenase ADH, auxin responsive family protein, lysine/histidine transporter etc. (dup) • ST4.03ch10: non-structural maintenance of chromosome element, conserved gene of unknown function etc. (dup)	• M6_v4.1ch11[!]: Auxin – induced SAUR gene cluster (del + dup) • M6_v4.1ch01: genes coding for NB-ARC, hypothetical proteins, bHLH, LRR, F-box etc. (dup) • M6_v4.1ch01[*]: Auxin – induced SAUR gene cluster (dup)

Continued on next page

Supplementary Table 12.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
COM	• ST4.03ch01*: Auxin – induced SAUR gene cluster (dup) • ST4.03ch11: Auxin – induced SAUR gene cluster (dup) • ST4.03ch06: Auxin – induced SAUR gene cluster (dup)	• M6_v4.1ch01: genes coding for NB - ARC domain, LRR, 3 – flavin – binding protein etc. (del) • M6_v4.1ch05$: genes coding for EXC family protein, Rubisco, disease resistance, LRR etc. (dup) • M6_v4.1ch12: genes coding for bHLH, fatty acid hydroxylase, tetraspanins, knotted-1 like etc. (dup)
M6	• ST4.03ch01*: Auxin – induced SAUR gene cluster (dup) • ST4.03ch04: genes coding for lipid binding protein, transducing family protein etc. (dup) • ST4.03ch04: conserved genes of unknown function, ethylene – inducing xylanase, 1,4 – α – glucan branching etc. (dup)	———

Continued on next page

Supplementary Table 12.1 – *Continued from previous page*

Genome	**Top 3 CNV bins VS DM1-3**	**Top 3 CNV bins VS M6**

*, $, ! These regions are the same in the genomes where the according symbol is found.

dup – duplication event

del – deletion event

Supplementary Table 13.1: Top 3 gene enriched CNV bins in the 6 polyploid genomes against the DM1-3 and the M6 reference genomes. *, $, ! These regions are the same in the genomes where the according symbol is found. dup – duplication event
del – deletion event

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
ADG1	• ST4.03ch12$^{\$}$: genes coding for extension Ext1, stress – associated proteins etc. (dup) • ST4.03ch04*: genes coding for late blight resistance protein, retroelement, ribulose – 1.5 bisphosphate, carboxylase/oxygenase etc. (dup) • ST4.03ch12: genes coding for thioredoxin domain – containing protein, fertility restorer etc. (dup)	• M6_v4.1ch01$^{\&}$: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch11^{+}: Auxin – induced SAUR gene cluster, F-box domain coding genes, etc. (del + dup) • M6_v4.1ch02: SINE3-like coding gene, RING/U – box, peroxidase, transferases etc. (dup)

Continued on next page

Supplementary Table 13.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
ADG2	• ST4.03ch07: geranyl geranyl pyrophosphate synthase, conserved gene of unknown function etc. (dup) • ST4.03ch07: ATNMNAT, RNA – binding protein, zinic transporter etc. (dup) • ST4.03ch03: genes coding for LRR, F-box, TPR domains, genes of unknown functions, pentatricopeptide, etc. (dup)	• M6_v4.1ch05^{+}: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch11: Auxin – induced SAUR gene cluster (dup) • M6_v4.1ch01$^{\&}$: Auxin – induced SAUR gene cluster (dup)

Continued on next page

Supplementary Table 13.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
TBR	• ST4.03ch11[$]: Auxin – induced SAUR gene cluster (dup) • ST4.03ch05[*]: ribulose – 1,5 bisphosphate, carboxylase/oxygenase, conserved genes of unknown function etc. (dup) • ST4.03ch05:signal transducer, nodulin family, glycine – rich cell wall structural protein 1, etc. (dup)	• M6_v4.1ch01[&]: Auxin – induced SAUR gene cluster, PPR, calcium binding EF- hand coding genes etc. (dup) • M6_v4.1ch07: late embryogenesis abundant (LEA) coding gene, methylotransferase family protein, lectin protein kinase, gibberellin 3 -oxidase etc. (dup) • M6_v4.1ch09: genes coding for 2 – oxoglutarate, bHLH, nodulin MtN21/EaMA like transporter family, etc. (dup)

Continued on next page

Supplementary Table 13.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
JUZ	• ST4.03ch01: Auxin – induced SAUR gene cluster (dup) • ST4.03ch03: conserved genes of unknown function, protein HVA22, etc. (dup) • ST4.03ch05*: ribulose – 1,5 bisphosphate, carboxylase/oxygenase, conserved genes of unkown function etc. (dup)	• M6_v4.1ch11+: Auxin – induced SAUR gene cluster, F- box, TPR, oxidoreductase etc. (el + dup) • M6_v4.1ch01&: Auxin – induced SAUR gene cluster, calcium – binding, methyl adenosine nucleoside etc. (dup) • M6_v4.1ch05: genes coding for F-box domain containing proteins, LAG1 longevity assurance homolog, ECA1 gametogenesis related, FAD/NAD(P) – binding oxidoreductase family protein, etc. (dup)

Continued on next page

Supplementary Table 13.1 – *Continued from previous page*

Genome	Top 3 CNV bins VS DM1-3	Top 3 CNV bins VS M6
CHA	• ST4.03ch05*: Auxin – induced SAUR gene cluster (dup) • ST4.03ch11$: Auxin induced SAUR gene cluster (del + dup) • ST4.03ch12: genes coding for zinic finger, CYP82C4, ubiquitin carrier, gens of unknown functions, etc. (dup)	• M6_v4.1ch01&: Auxin – induced SAUR gene cluster, PPr, CCHC – type zinc finger, etc. (dup) • M6_v4.1ch11+: Auxin – induced gene cluster (del + dup) • M6_v4.1ch02: genes coding for TPX2 (targeting protein for Xklp2), hydrolases, glucose – 6 – phosphate-dehydrogenase, etc. (dup)
CUR	• ST4.03ch11$: Auxin – induced SAUR gene cluster(dup) • ST4.03ch06: genes coding for male sterility MS5, nodulin 26, endo 1,4 beta xylanase (xylA), etc. (dup) • ST4.03ch05*: Auxin – induced SAUR gene cluster (dup)	• M6_v4.1ch11+: Auxin – induced SAUR gene cluster, F-box coding genes, etc. (dup)

Supplementary Table 14.1: Significant CNV gene clusters in common between the diploid genomes when compared to DM1—3 and M6 reference genomes along with the variation status; duplicated or deleted.

	Compared to the DM1-3	Compared to the M6
Chr 01	**85 – 85.21 Mb – mostly impacted by deletions** Cluster of 18 genes coding for methylketone synthase enzyme.	**54.44 – 54.64 Mb – mostly impacted by duplications** A 55 – gene cluster, including genes coding for Major Facilitator Superfamily (MFS) protein of membrane transport secondary carriers, for argirine-rich cyclin, chitin elicitor receptor kinase, for DNAJ heat shock N terminal domain-containing protein and for the pentatricopeptide repeat (PPR) superfamily protein.

Continued on next page

Supplementary Table 14.1 – *Continued from previous page*

	Compared to the DM1-3	**TCompared to the M6**
Chr 04	**4.6 – 4.8 Mb – impacted by deletions (in GON1, GON2, PHU and STN), impacted by duplications (in AJH, BUK, COM and M6)** Cluster of 32 genes, specifically disease resistance genes, including *R2* gene.	**35.41 – 35.61 Mb – mostly impacted by deletions** A cluster of 33 genes coding for serine protease inhibitor (SERPIN), for UDP-Glycotransferase superfamily proteins, MFS proteins and disease resistance ADG1- like proteins. **36.47 – 36.67 Mb – mostly impacted by duplications** A cluster of 32 genes coding for ribonucleases, leucine carboxyl methyltransferases, mannose-binding lectin family protein, PPRs and a NmrA-like negative transcriptional regulator family protein.

Continued on next page

Supplementary Table 14.1 – *Continued from previous page*

	Compared to the DM1-3	**TCompared to the M6**
Chr 09		
	59.86 – 60.06 Mb – mostly impacted by deletions	**29.23 – 29.46 Mb – mostly impacted by duplications**
	A cluster of 26 genes, including those coding for glycotransferase, cembratrienol synthase, a gene coding for a xyloglucan endotrabskycosylase, a hypothetical salt-inducible protein, another for F-box protein, Cytochrome P450 and other conserved genes of unknown function.	A cluster of 43 genes; 30 of these code for 2-oxoglutarate (2OG) and FE (II)-dependent oxygenase superfamily protein, in addition to others coding for a nodulin MtN21/EamA-like transporter family protein, an electron transfer flavoprotein, bHLH, and the UDP-glycosyltrasnferase superfamily.
Chr 11		
	1.33 – 1.53 Mb – impacted by deletions (duplications in M6)	**0.89 – 1.11 Mb – mostly impacted by deletions (except AJH)**
	A cluster of 17 genes, including genes coding for Leucine Rich Repeat family proteins, TMV resistance protein N.	A 54 gene cluster, including 32 genes coding for SAUR-like auxin responsive protein family, along with ethylene-responsive element coding genes.

Continued on next page

Supplementary Table 14.1 – *Continued from previous page*

	Compared to the DM1-3	**TCompared to the M6**
Chr 12	**0.6 – 0.8 Mb – impacted by deletions (PHU and COM did not show CV)** A 29 – gene cluster, including 7 mannan endo-1,3-beta-mannosidase 1 coding genes, others coding for important plant immunity proteins, such as Ubiquitin conjugating enzyme, Rnf5 and stress tolerance, like Fiber protein Fb34, metallothionein.	**No significant CNV gene clusters in common between the diploids.**

Supplementary Table 15.1: Significant gene clusters in common between the polyploids against the DM1-3 and M6 genomes along with the variation status.

	Compared to the DM1-3	Compared to the M6
Chr 01	**4.4 – 5.38 Mb – duplicated only in the tetraploids, no CNV-impacted genes in other genomes** Cluster of 18 genes coding for methylketone synthase enzyme. **54.44 – 54.64 Mb – mostly impacted by duplications** A 66-gene cluster of unknown function, others coding for Cyt. P450, S2 SI locus-linked pollen 3.2 protein, mannan endo-1,4-beta-mannosidase 4, F-box family protein and ethylene-responsive TF. **75.8 – 85 Mb – impacted by duplications** Various genes; like *FIONA* coding for *FRIGIDA* and Type I MADS box TF, others coding for male sterility protein, low temperature and salt responsive protein, heat shock proteins sugar transporters, verticillium wilt disease resistance protein, glycosyl transferase, auxin response protein, TMV resistance protein N. **86.98 – 87.2 Mb – impacted by duplications but no CNV-impacted genes in ADG1** A 43-gene cluster, including 17 auxin induced *SAUR* genes.	**35.41 – 35.61 Mb – mostly impacted by deletions** A cluster of 33 genes coding for serine protease inhibitor (SERPIN), for UDP-Glycotransferase superfamily proteins, MFS proteins and disease resistance ADG1- like proteins. **36.47 – 36.67 Mb – mostly impacted by duplications** A cluster of 32 genes coding for ribonucleases, leucine carboxyl methyltransferases, mannose-binding lectin family protein, PPRs and a NmrA-like negative transcriptional regulator family protein. **64.64 – 64.82 Mb – mostly impacted by duplications** A *SAUR*-auxin induced coding gene cluster.

Continued on next page

Supplementary Table 15.1 – *Continued from previous page*

	Compared to the DM1-3	**Compared to the M6**
Chr 03	**No significant CNV gene clusters in common between the polyploids.**	**41.32 – 41.53 Mb – mostly impacted by duplications** A 35-gene cluster of genes coding for tetraspanin8, tetraspanin12 and a calcium-dependent phosphotriesterase protein.
Chr 05	**0 – 0.21 Mb – mostly impacted by duplications** A 33 disease resistance gene cluster as well as other genes involved in cell metabolic processes.	**No significant CNV gene clusters in common between the polyploids.**

Continued on next page

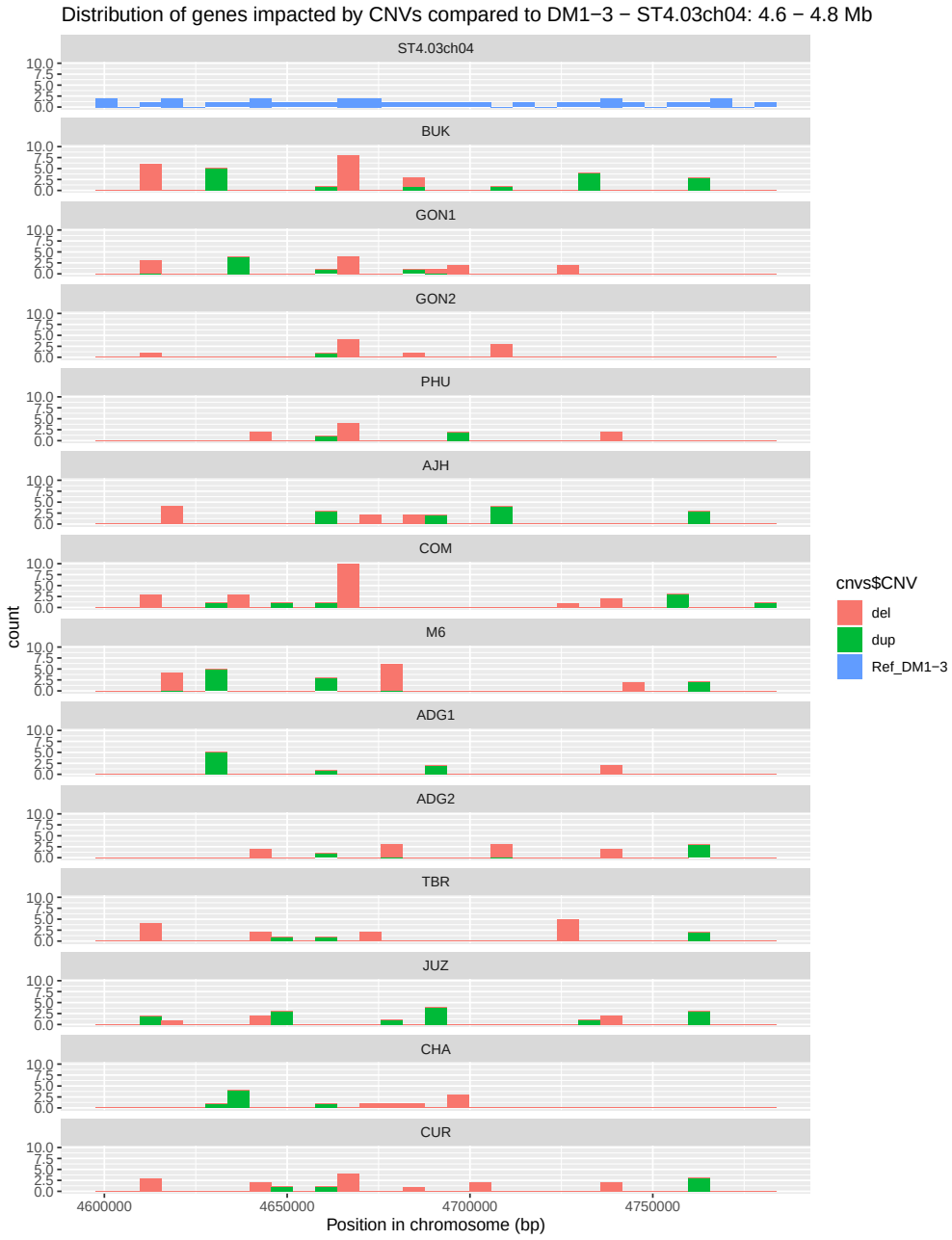
Supplementary Table 15.1 – *Continued from previous page*

	Compared to the DM1-3	**Compared to the M6**
Chr 07		
	55.36 – 55.57 Mb – duplications in all polyploids but no CNV-impacted genes in ADG1 A cluster of 186 genes coding for enoyl-CoA hydratase, mitochondrial, zinic transporter, lipid binding protein and a fascilin-like arbinogalactan protein 10.	**32.12 – 32.37 Mb – mostly impacted by duplications** A cluster of 43 genes coding for gibberellin 3-oxidase, TPR HCO3
Chr 09		
	59.4 – 61 Mb – deleted in ADG1, ADG2, duplicated only in TBR, no CNV-impacted genes in other genomes A 61-gene cluster containing Tospovirus resistance genes.	**29.2 – 29.46 Mb – mostly impacted by duplications** A cluster of 43 genes, 30 of which code for 2OGD

Continued on next page

Supplementary Table 15.1 – *Continued from previous page*

	Compared to the DM1-3	**Compared to the M6**
Chr 10	**No significant CNV gene clusters in common between the polyploids.**	**1.24 – 1.44 Mb – only impacted by duplications** A cluster of 40 genes involved in various functions, like metabolic processes and response to stimulus.
Chr 11	**No significant CNV gene clusters in common between the polyploids.**	**0.88 – 1.14 Mb – mostly impacted by duplications** A *SAUR*-auxin induced gene cluster.
Chr 12	**58.12 – 58.34 Mb – impacted by duplications** A 34 gene cluster involved in cellular metabolic processes.	**40.54 – 40.74 Mb – mostly impacted by duplications** A cluster of 28 genes, 18 of which code for hypothetical proteins.



Supplementary Figure 16.1: Overview of CNVs over fourteen potato genomes compared with the DM1-3 reference genome, in chromosome 4; 4.6 – 4.8 Mb. Distribution of genes of which 50% or more of their gene body was affected by CNVs: with pink the genes impacted by deletions, green with duplications and blue the gene distribution of the DM1-3 in this region.

Supplementary Figure 17.1: A): Overview of CNVs over fourteen potato genomes compared with the M6 reference genome, in chromosome 01: 64.64 – 64.82 Mb. Distribution of genes of which 50% or more of their gene body was affected by CNVs: with pink the genes impacted by deletions, blue with duplications and green the gene distribution of the DM1-3 in this region.

Supplementary Figure 18.1: B): Overview of CNVs over fourteen potato genomes compared with the M6 reference genome, in chromosome 09: 29.23 – 29.46 Mb. Distribution of genes of which 50% or more of their gene body was affected by CNVs: with pink the genes impacted by deletions, blue with duplications and green the gene distribution of the DM1-3 in this region.

Supplementary Figure 19.1: B) Overview of CNVs over fourteen potato genomes compared with the M6 reference genome, in chromosome: 11: 0.88 – 1.11 Mb. Distribution of genes of which 50% or more of their gene body was affected by CNVs: with pink the genes impacted by deletions, blue with duplications and green the gene distribution of the DM1-3 in this region.

Supplementary Table 20.1: Genomes used for the pan-genome construction, along with the technologies used for sequencing and their references. The % heterozygosity shows was calculated from the Illumina PE reads of the genomes using GenomeScope 2.0.

Solanum full taxon name	Code Name	Technology	SRA Accession	Reference
		Illumina PE	SRR10244441	
S. tuberosum subsp. goniocalyx	GON1	10X Genomics	SRR10237767	(Kyriakidou et al., 2019)
		PacBio	SRR10242928	
S. tuberosum subsp. goniocalyx	GON2	Illumina PE	SRR10244440	
S. phureja	PHU	Illumina PE	SRR10244439	(Kyriakidou et al., 2019)
S. stenotomum subsp. stenotomum	STN	Illumina PE	SRR10244438	
S. xajanhuiri	AJH	Illumina PE	SRR10244437	(Kyriakidou et al., 2019)
S. bukasovii	BUK	Illumina PE	SRR10244436	(Kyriakidou et al., 2019)
S. commersonii	COM	Illumina PE	SRS5775810	(Aversano et al., 2015)
S. chacoense	M6	Illumina PE	SRR5264013 SRR5264022	(Leisner et al., 2018)

Supplementary Table 21.1: Multiple approaches used to assemble the GON1 genome.

Assembler	Total Size (Mb)	No. of Scaf.	Scaf. N50 (bp)	Longest Scaf. (bp)	BUSCO C	BUSCO C + F	BUSCO C + D
Supernova (meg)[1]	680	13,135	291,434	5,021,079	95%	96.70%	4.70%
Supernova (pseu2[1]	948	65,244	240,670	5,021,079	95.90%	97.30%	6.50%
ABySS[2,3]	1100	2,248,566	3062	117,013	70.90%	78.50%	63.20%
MaSuRCA[2,3]	710	66,960	28,218	309,361	86.40%	90.40%	7.60%

*The reads from all the technologies were filtered before the assemblies — [1]10X Genomics, [2]PacBio, [3]Illumina PE — **BUSCO**: **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs. — C: Complete genes — F: Fragmented genes — D: Duplicated genes

Supplementary Table 22.1: Lengths (bp) of the newly generated pseudomolecules of the GON1 genome compared to those of DM1-3 and M6 reference genomes.

	GON1	DM1-3	M6
Pseudomolecule 1	55,497,550	88,663,952	66,849,660
Pseudomolecule 2	30,369,252	48,614,681	42,376,642
Pseudomolecule 3	39,977,587	62,290,286	47,120,749
Pseudomolecule 4	49,211,853	72,208,621	38,708,216
Pseudomolecule 5	31,032,647	52,070,158	42,711,019
Pseudomolecule 6	40,549,446	59,532,096	41,715,865
Pseudomolecule 7	36,469,541	56,760,843	41,467,633
Pseudomolecule 8	38,732,972	56,938,457	31,643,503
Pseudomolecule 9	41,688,201	61,540,751	35,384,898
Pseudomolecule 10	43,340,457	59,756,223	35,706,090
Pseudomolecule 11	25,574,028	45,475,667	38,261,750
Pseudomolecule 12	36,209,197	61,165,649	49,944,156

Supplementary Table 23.1: Repeats identified in GON1 genome.

Element	Number of Elements	Length Occupied (bp)	Percentage of sequence
LINEs	34,304	16,929,511	1.98
LTR elements	182,602	233,716,809	27.3
DNA elements	22,158	12,722,938	1.49
Simple repeats	152,453	10,546,496	1.23
Low complexity	30,772	1,762,467	0.21
Unclassified	785,690	243,142,892	28.4
Total bases masked		515,341,644	60.2

Supplementary Table 24.1: Quality metrics of the de novo genome assemblies before removing redundant contigs.

Quality metric	GON1	GON2	PHU	STN	AJH	BUK
# of contigs	6,490*	625,456	352,009	393,883	478,866	379,875
Contigs > 1000 bp	6,426	348,952	223,503	260,925	287,261	256,278
Length of assembly (Mb)	856	1014	994	1100	1200	1300
GC %	34.88	35.12	34.66	35.28	36.02	35.03
Largest contig length (Kb)	9.3 Mb	42	156	163	133	183.7
Contig N50	326,785	2,872	5,500	4,933	4,699	6,600
Contig L50	531	110,819	42,907	53,919	59,125	47,514

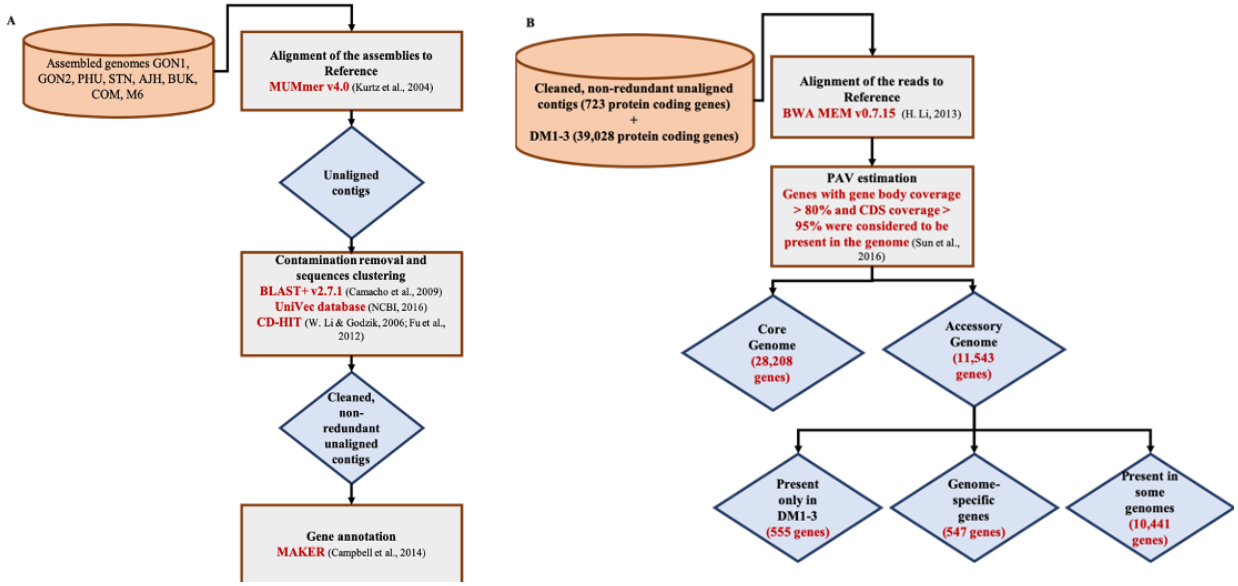
* For GON1 genome it refers to the number of the scaffolds.

Supplementary Table 25.1: Significant GO terms of the 11,542 accessory genome.

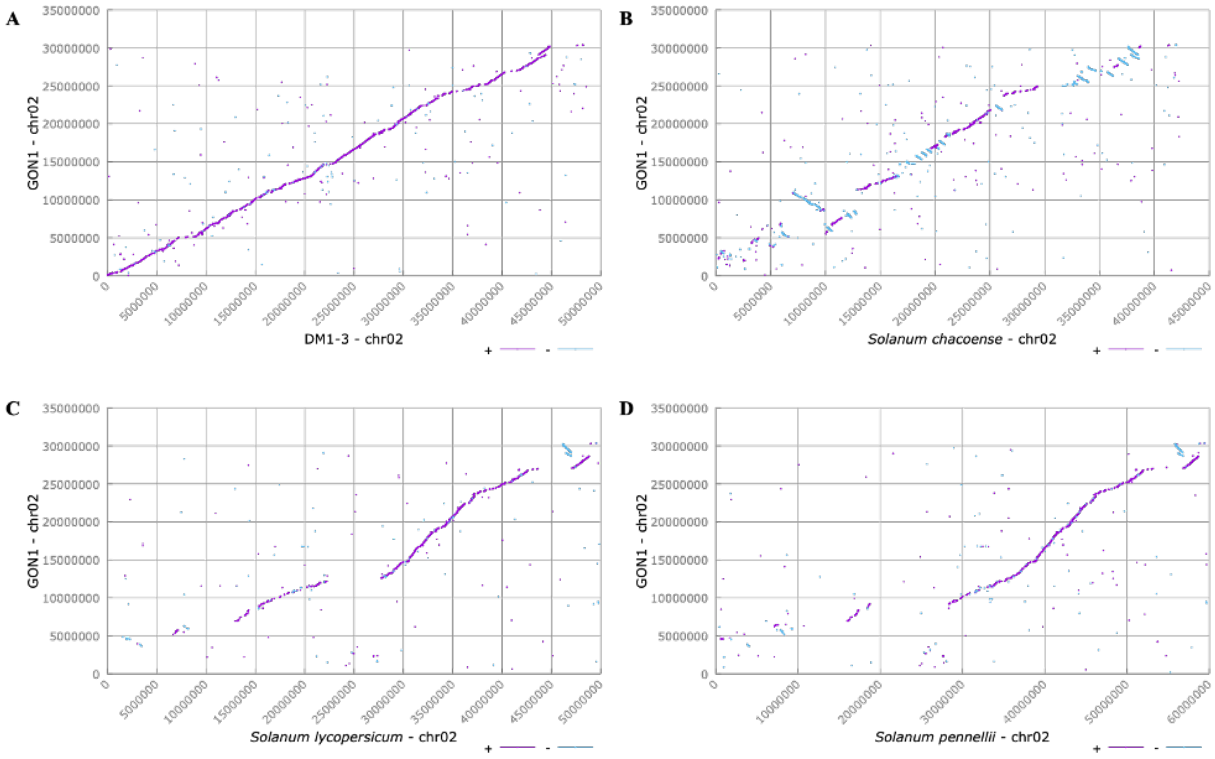
GO.ID	Term	Annotated	Significant
GO:0008033	tRNA processing	71	49
GO:0055114	oxidation-reduction process	1402	308
GO:0009607	response to biotic stimulus	74	31
GO:0006468	protein phosphorylation	791	148
GO:0007165	signal transduction	218	48
GO:0006094	gluconeogenesis	4	3
GO:0045040	protein import into mitochondrial outer membrane	4	3
GO:0009664	plant-type cell wall organization	8	4
GO:0019684	photosynthesis, light reaction	10	4

Supplementary Table 26.1: Significant GO terms of newly predicted protein-coding gene found in the pan-genome.

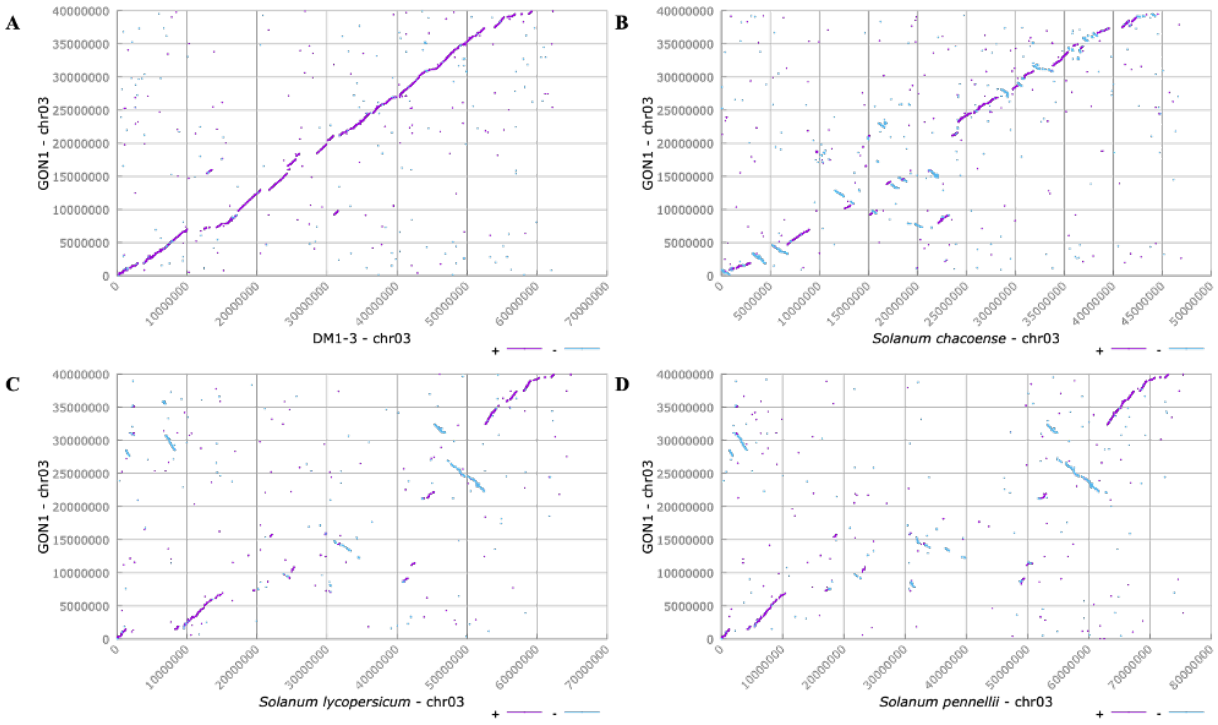
GO.ID	Term	Annotated	Significant
GO:0006952	defense response	9	7
GO:0006099	tricarboxylic acid cycle	2	2
GO:0000105	histidine biosynthetic process	4	2
GO:0000160	phosphorelay signal transduction system	35	3
GO:0015986	ATP synthesis coupled proton transport	12	2
GO:0000413	protein peptidyl-prolyl isomerization	1	1
GO:0000723	telomere maintenance	1	1
GO:0002100	tRNA wobble adenosine to inosine editing	1	1
GO:0006537	glutamate biosynthetic process	1	1
GO:0009772	photosynthetic electron transport in pho...	1	1
GO:0008610	lipid biosynthetic process	39	2
GO:0006400	tRNA modification	3	2
GO:0006450	regulation of translational fidelity	2	1
GO:0016579	protein deubiquitination	2	1
GO:0006779	porphyrin-containing compound biosynthet...	4	1
GO:0016226	iron-sulfur cluster assembly	5	1
GO:0015074	DNA integration	14	1
GO:0034220	ion transmembrane transport	57	3



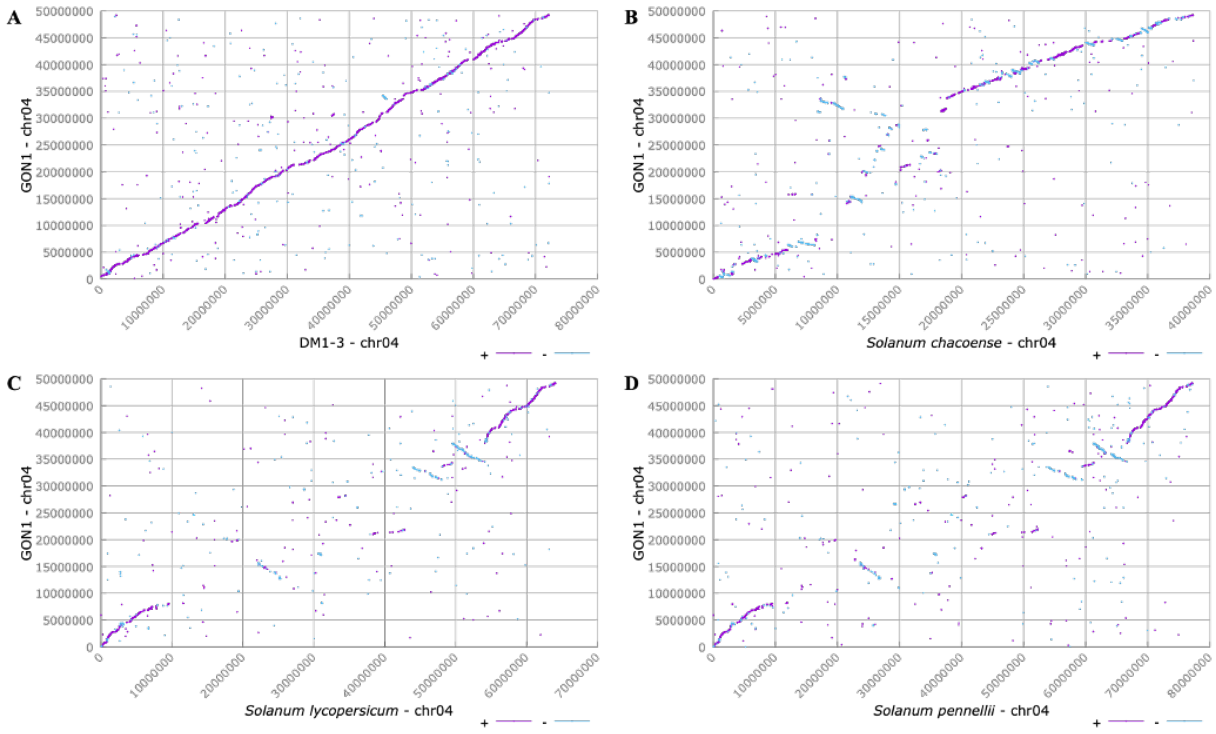
Supplementary Figure 27.1: Diploid pan-genome pipeline followed. A) The genome assemblies of the GON1, PHU, STN, AJH, BUK, COM and M6 genomes were aligned to the DM1-3, mitochondrial and chloroplast genomes. From the unaligned contigs, any contaminants and the overlapping sequences were removed to avoid redundancy. The final, cleaned, unaligned contigs were annotated. **B)** The cleaned, non-redundant unaligned contigs along with the DM1-3 pseudomolecules consist of the pan-genome, which contains the 723 newly predicted coding genes and the 39,028 protein coding genes find in the DM1-3. The sequencing reads of the eight genomes were aligned to the pan-genome (unaligned contigs and DM1-3 pseudomolecules) for the presence/absence (pav) analysis. Based on the results, the core genome (genes found in the eight genomes) consist of 28,208 genes, while the accessory genome (genes found in some of the genomes, or not at all) consists of 11,543 genes. Within the accessory genome, there are 555 genes found only in the DM1-3 and not in the rest of the genomes, 547 genome-specific genes, and 10,441 genes present in some genomes and absent from the others.



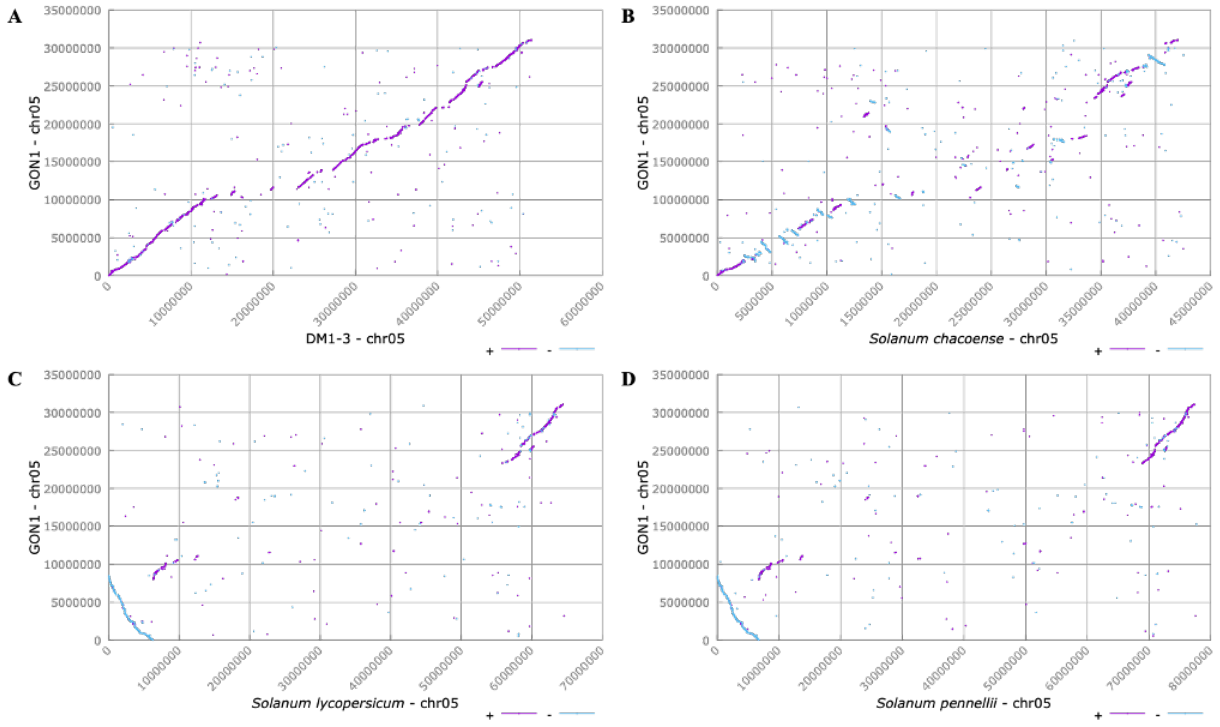
Supplementary Figure 28.1: Alignment of the chromosome 2 from the S. stenotomum subsp. goniocalyx with chromosome 2 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 02 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 2 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch02), **B.** *S. chacoense* (M6v4.1chr02), **C.** *S. lycopersicum* (SL2.40ch02) and **D.** *S. pennellii* (Spenn-ch02).



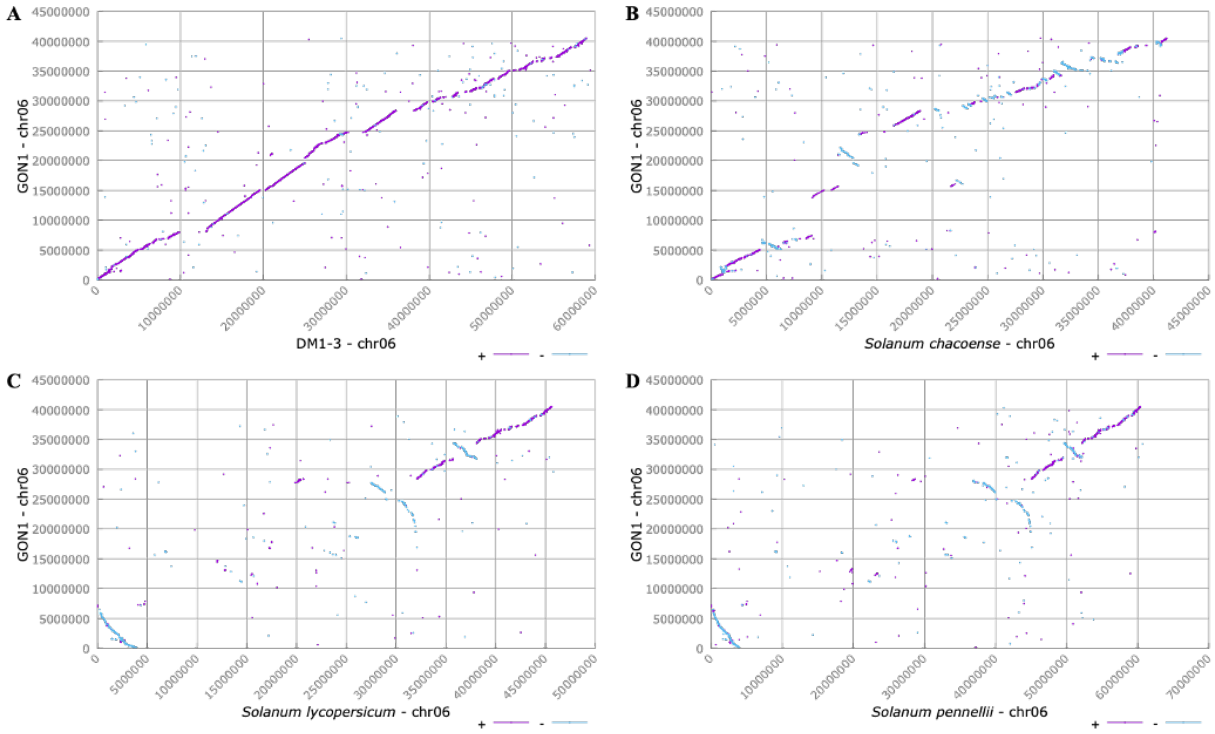
Supplementary Figure 29.1: Alignment of the chromosome 3 from the S. stenotomum subsp. goniocalyx with chromosome 3 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 03 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 3 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch03), **B.** *S. chacoense* (M6v4.1chr03), **C.** *S. lycopersicum* (SL2.40ch03) and **D.** *S. pennellii* (Spenn-ch03).



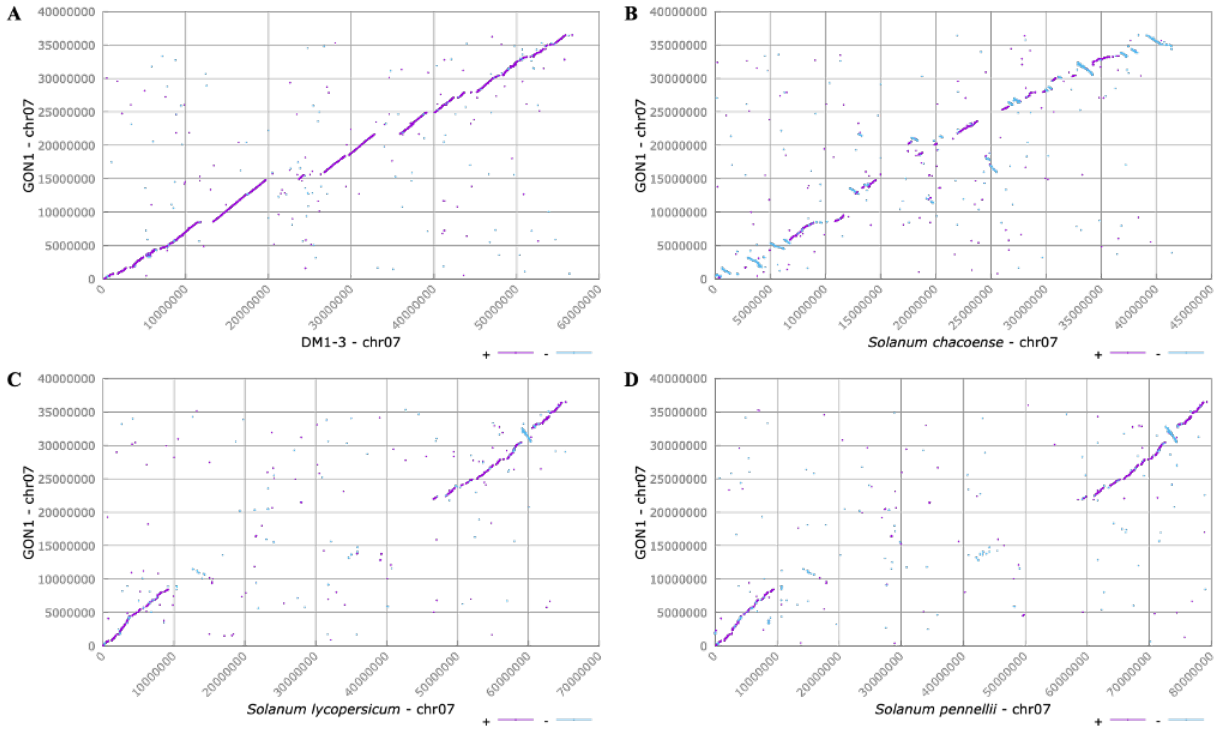
Supplementary Figure 30.1: Alignment of the chromosome 4 from the S. stenotomum subsp. goniocalyx with chromosome 4 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 04 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 4 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch04), **B.** *S. chacoense* (M6v4.1chr04), **C.** *S. lycopersicum* (SL2.40ch04) and **D.** *S. pennellii* (Spenn-ch04).



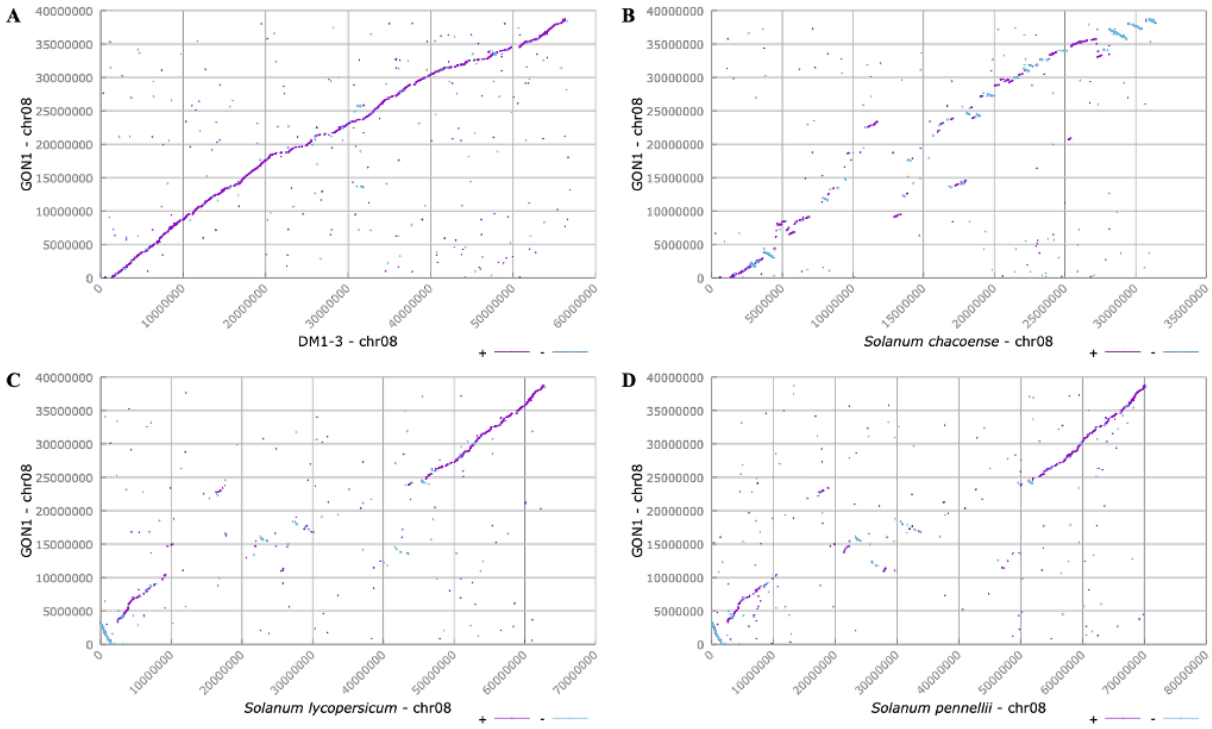
Supplementary Figure 31.1: Alignment of the chromosome 5 from the S. stenotomum subsp. goniocalyx with chromosome 5 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 05 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 5 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch05), **B.** *S. chacoense* (M6v4.1chr05), **C.** *S. lycopersicum* (SL2.40ch05) and **D.** *S. pennellii* (Spenn-ch05).



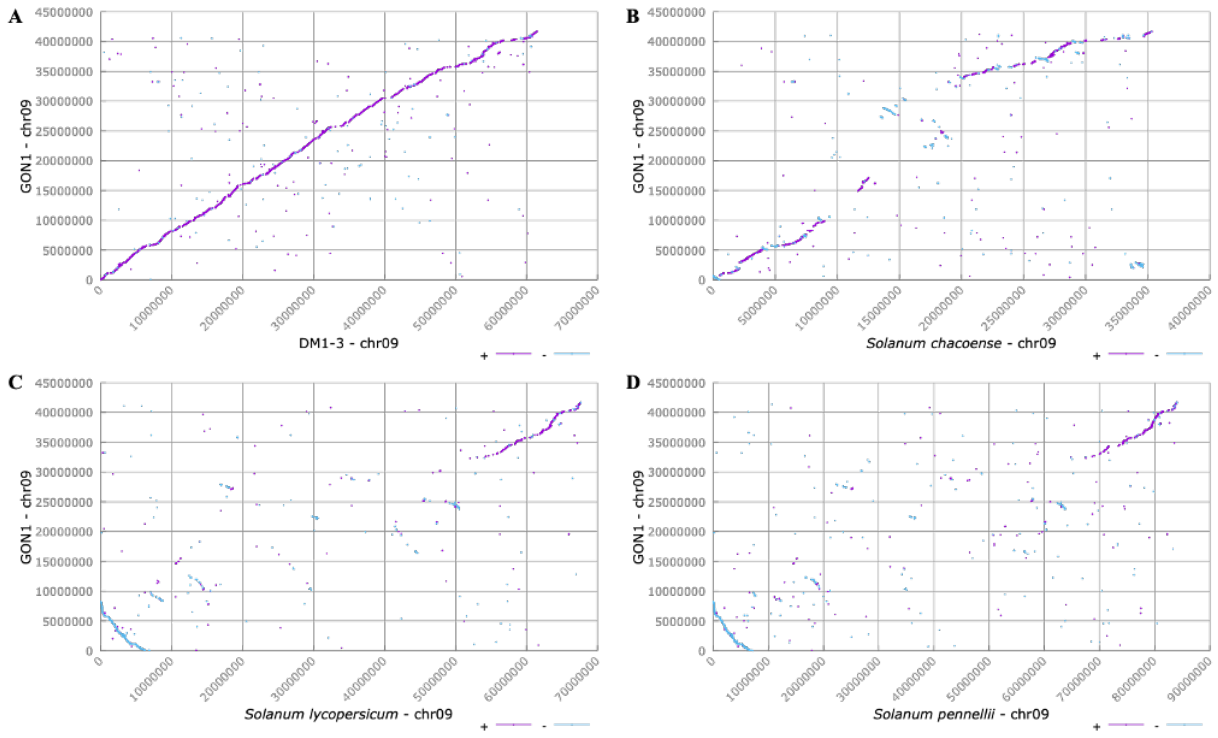
Supplementary Figure 32.1: Alignment of the chromosome 6 from the S. stenotomum subsp. goniocalyx with chromosome 6 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 05 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 6 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch06), **B.** *S. chacoense* (M6v4.1chr06), **C.** *S. lycopersicum* (SL2.40ch06) and **D.** *S. pennellii* (Spenn-ch06).



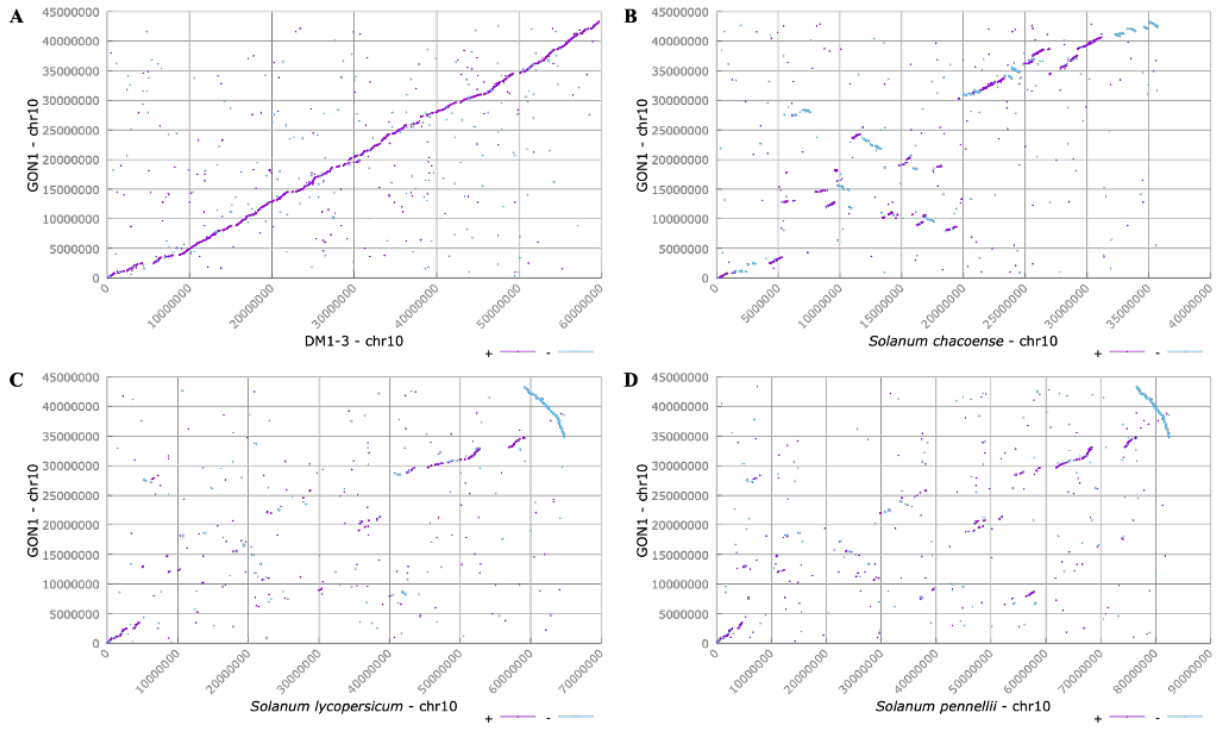
Supplementary Figure 33.1: Alignment of the chromosome 7 from the S. stenotomum subsp. goniocalyx with chromosome 7 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 07 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 7 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch07), **B.** *S. chacoense* (M6v4.1chr07), **C.** *S. lycopersicum* (SL2.40ch07) and **D.** *S. pennellii* (Spenn-ch07).



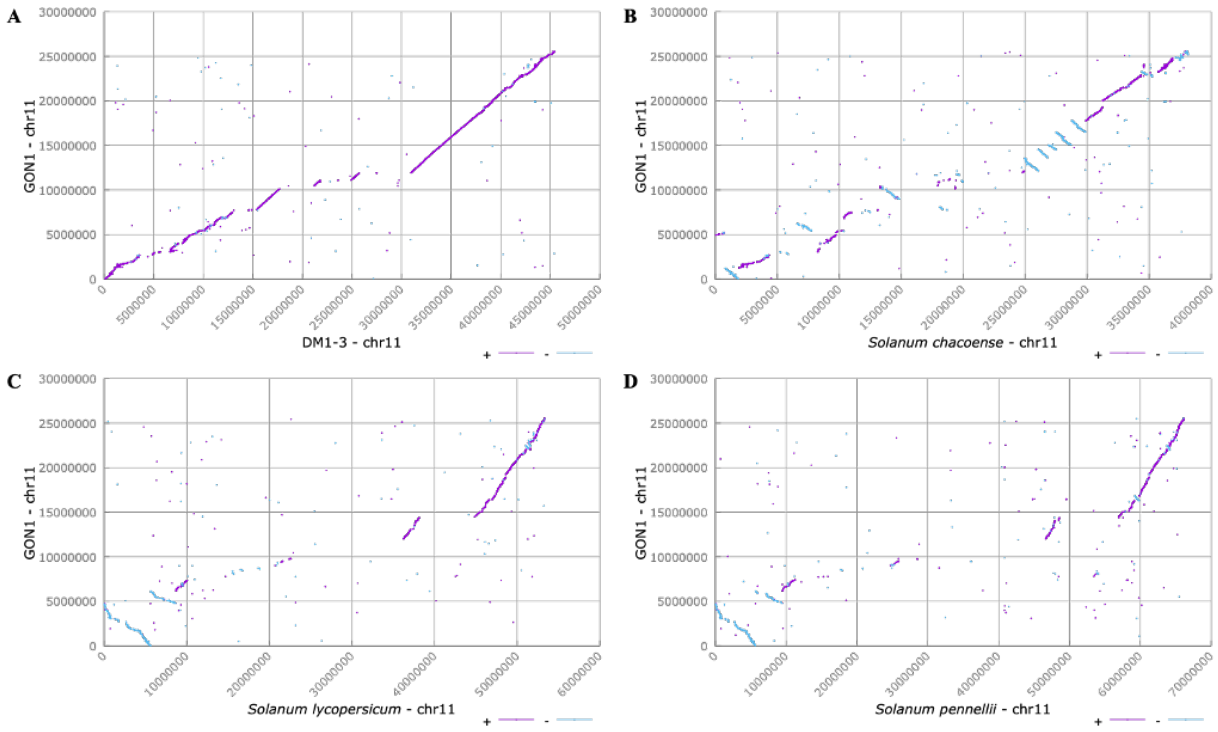
Supplementary Figure 34.1: Alignment of the chromosome 8 from the S. stenotomum subsp. goniocalyx with chromosome 8 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 08 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 8 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch08), **B.** *S. chacoense* (M6v4.1chr08), **C.** *S. lycopersicum* (SL2.40ch08) and **D.** *S. pennellii* (Spenn-ch08).



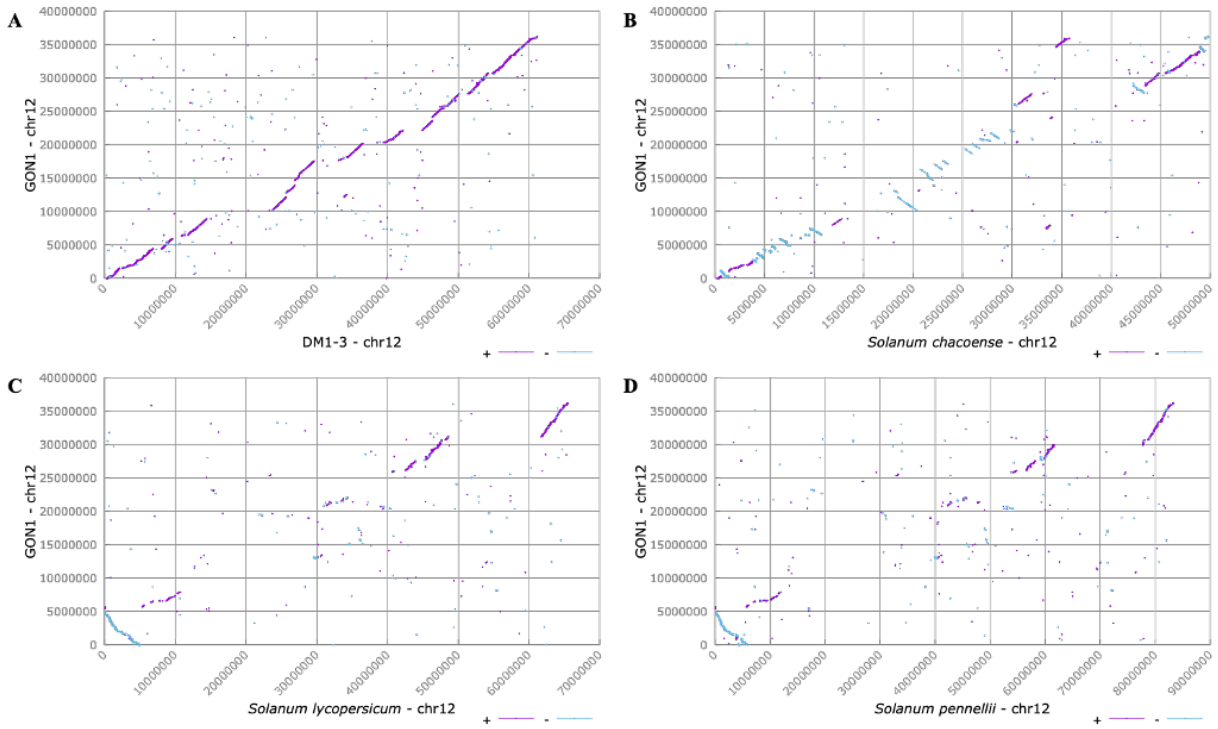
Supplementary Figure 35.1: Alignment of the chromosome 9 from the S. stenotomum subsp. goniocalyx with chromosome 9 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 09 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 9 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch09), **B.** *S. chacoense* (M6v4.1chr09), **C.** *S. lycopersicum* (SL2.40ch09) and **D.** *S. pennellii* (Spenn-ch09).



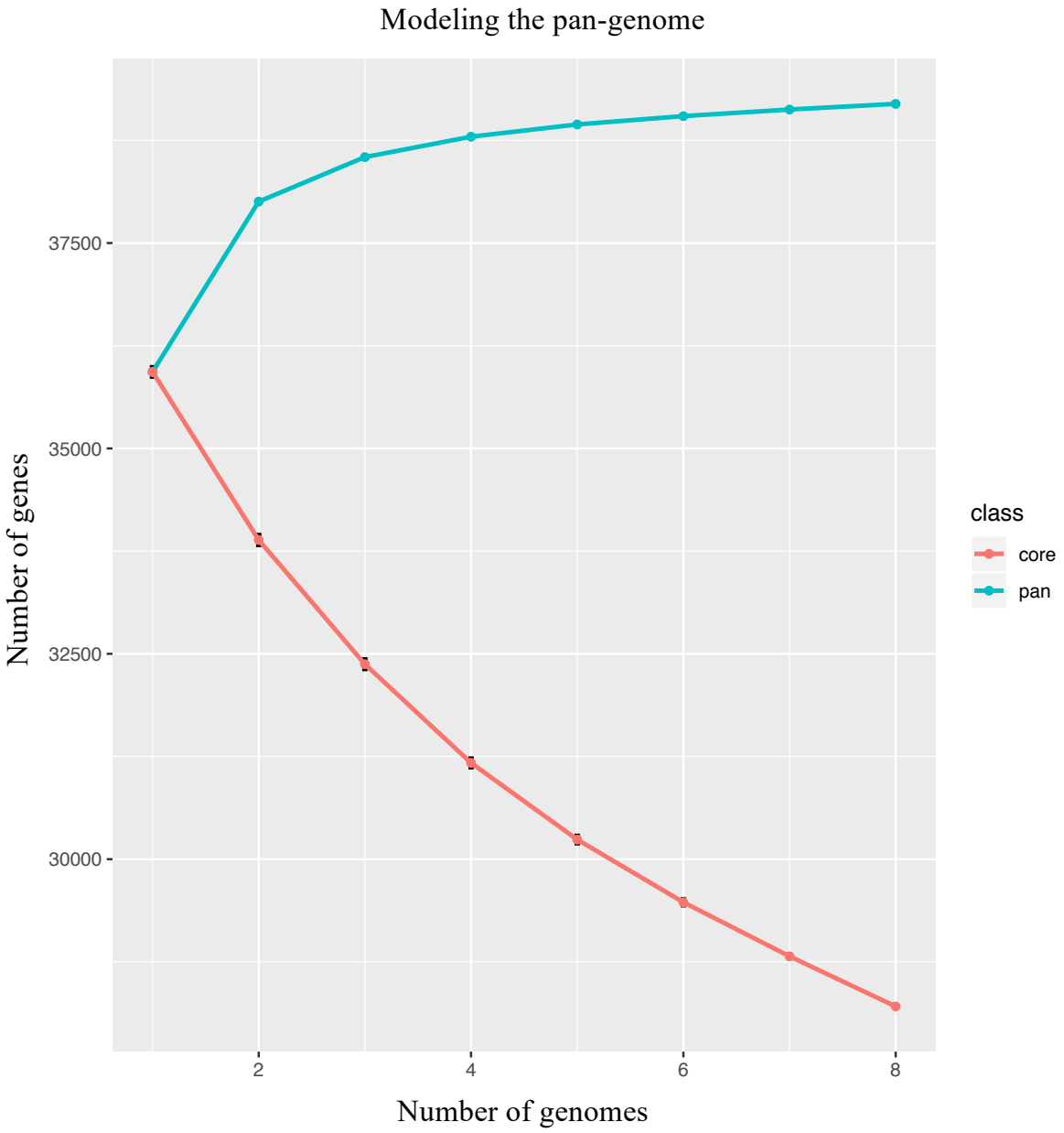
Supplementary Figure 36.1: Alignment of the chromosome 9 from the S. stenotomum subsp. goniocalyx with chromosome 10 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 10 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 10 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch10), **B.** *S. chacoense* (M6v4.1chr10), **C.** *S. lycopersicum* (SL2.40ch10) and **D.** *S. pennellii* (Spenn-ch10).



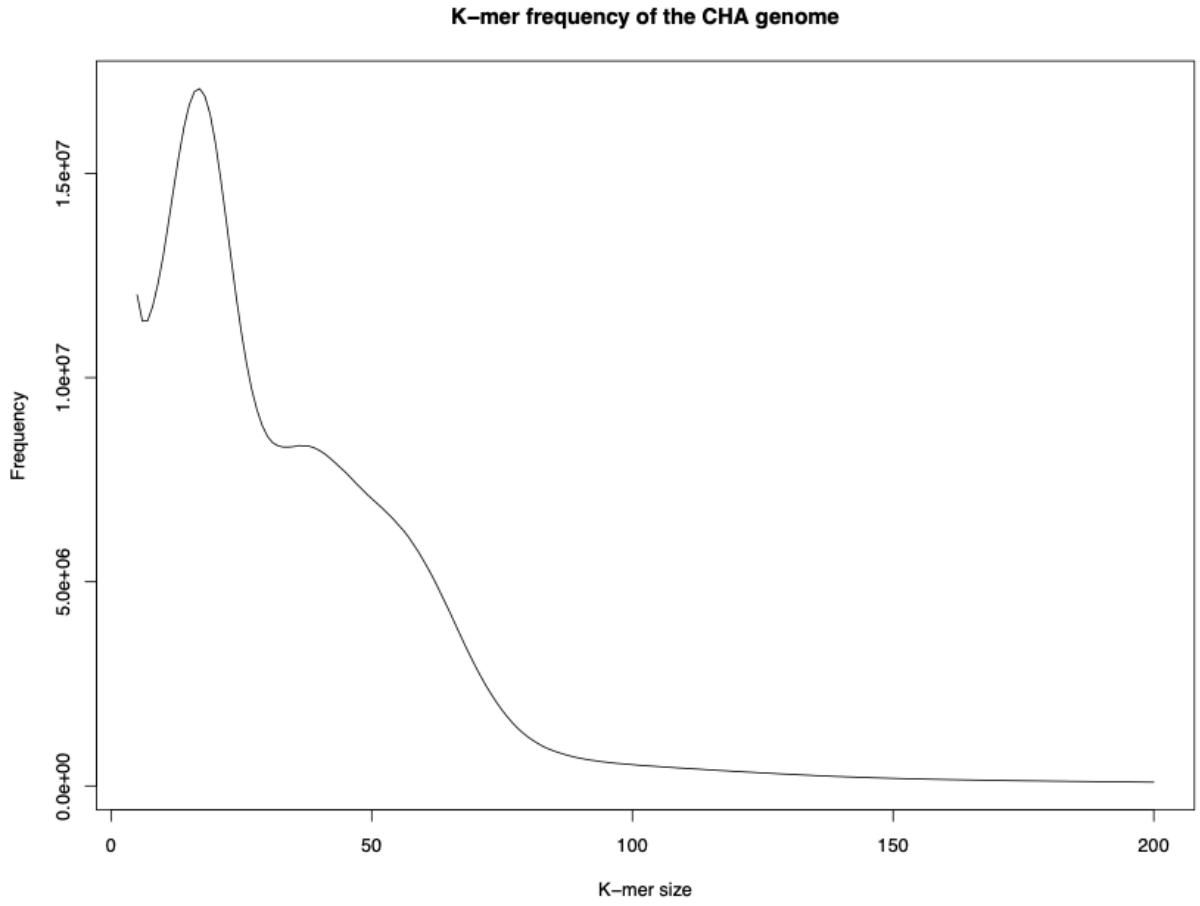
Supplementary Figure 37.1: Alignment of the chromosome 9 from the S. stenotomum subsp. goniocalyx with chromosome 11 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 11 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 11 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch11), **B.** *S. chacoense* (M6v4.1chr11), **C.** *S. lycopersicum* (SL2.40ch11) and **D.** *S. pennellii* (Spenn-ch11).



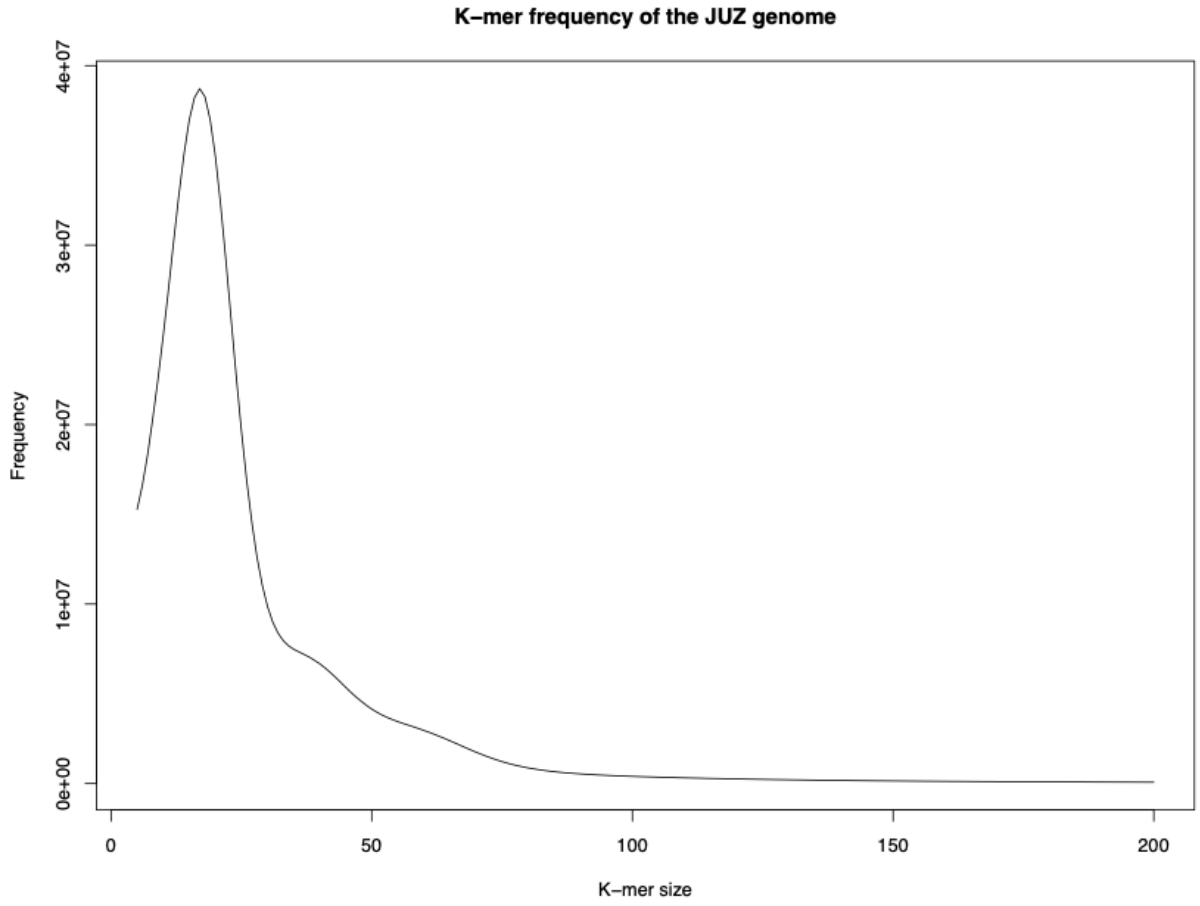
Supplementary Figure 38.1: Alignment of the chromosome 9 from the S. stenotomum subsp. goniocalyx with chromosome 12 from other Solanum sp. Filtered nucmer (aligner for standard DNA sequence alignment) pairwise alignments extracted for Chromosome 12 of Solanum stenotomum subsp goniocalyx - GON1 against the chromosome 12 of the following: **A.** *S. tuberosum*/DM1-3 (ST4.03ch12), **B.** *S. chacoense* (M6v4.1chr12), **C.** *S. lycopersicum* (SL2.40ch12) and **D.** *S. pennellii* (Spenn-ch12).



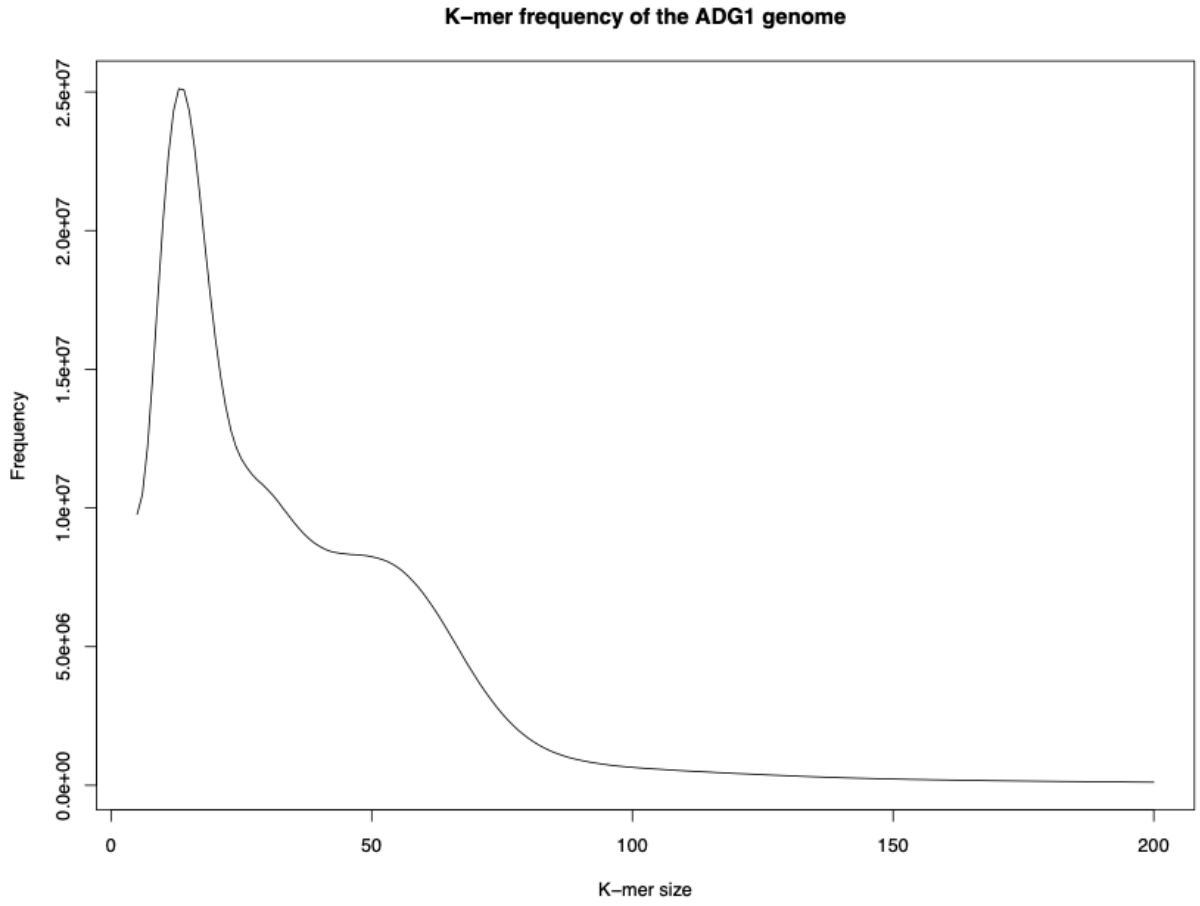
Supplementary Figure 39.1: Modeling the size of the potato pan-genome and core genome. While the size of the pan-genome is increasing, the core genome size is decreasing. The pan-genome consists of a total of 39,751 genes. 100 random combinations of the eight genomes were used for the modeling. Upper and lower blue and pink solid lines correspond to the maximum and minimum number of genes, respectively. The pan-genome increases when we add more genomes. While the core genome decreases, the accessory genome (the difference between the pan and core) increases.



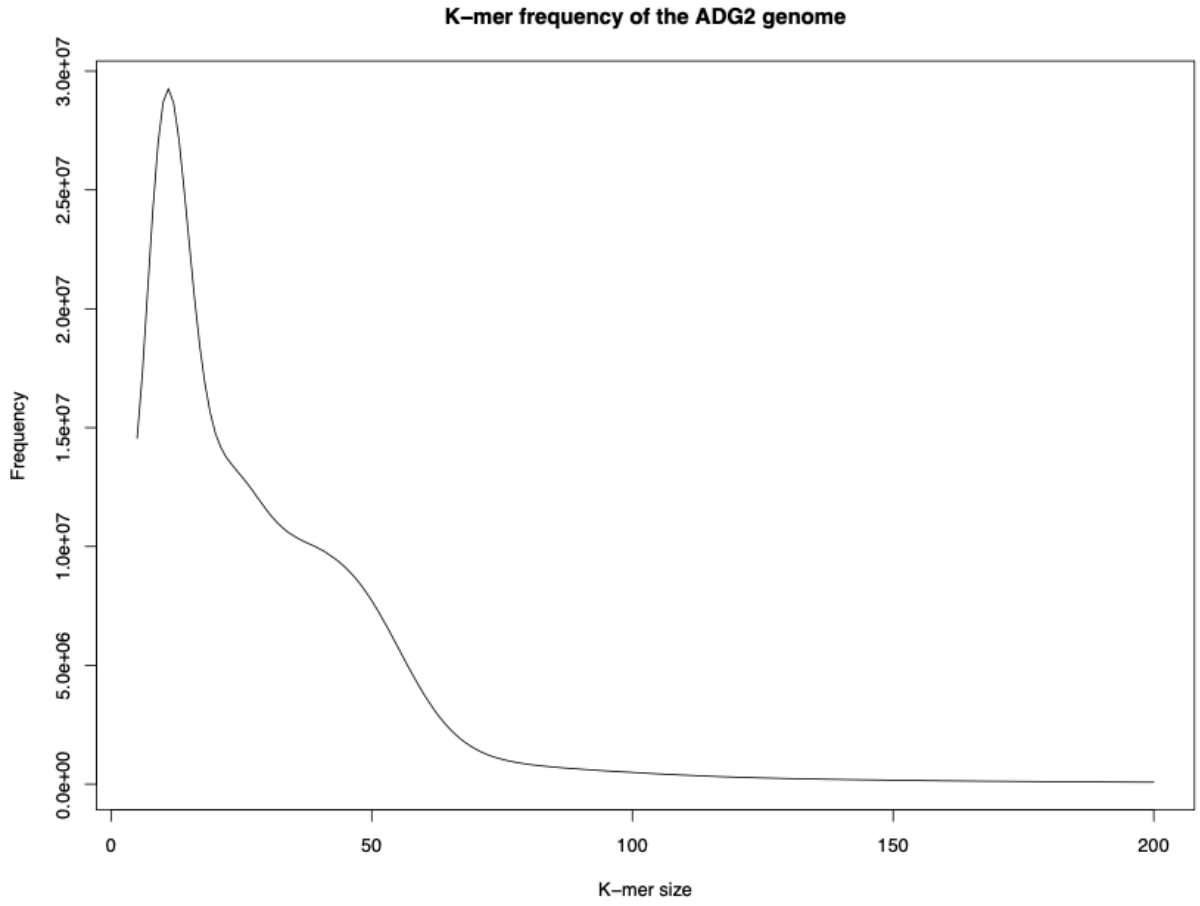
Supplementary Figure 40.1: The k-mer frequency of the CHA genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.



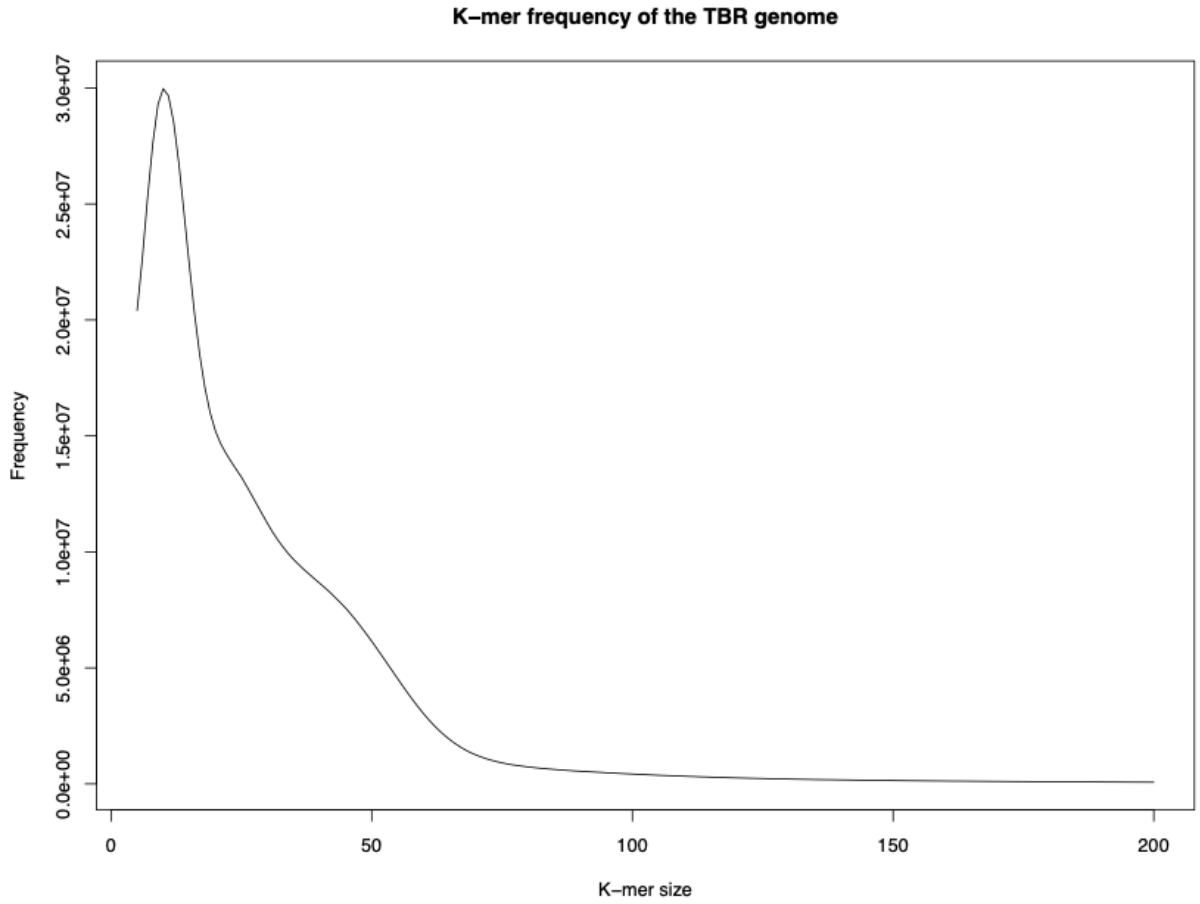
Supplementary Figure 41.1: The k-mer frequency of the JUZ genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.



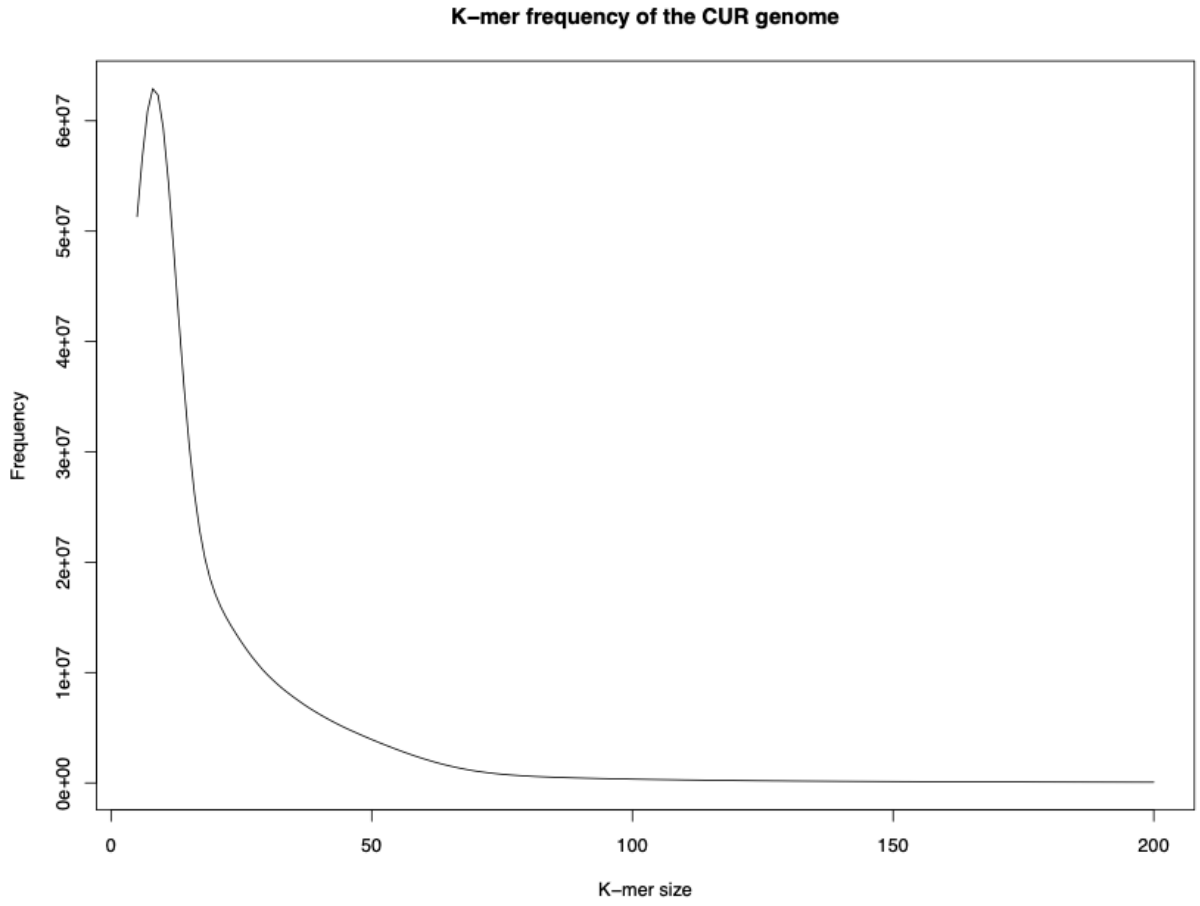
Supplementary Figure 42.1: The k-mer frequency of the ADG1 genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.



Supplementary Figure 43.1: The k-mer frequency of the ADG2 genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.



Supplementary Figure 44.1: The k-mer frequency of the TBR genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.



Supplementary Figure 45.1: The k-mer frequency of the CUR genome. The increased heterozygosity of the genome is validated by the tendency towards bimodal distribution of the k-mer frequency.