GENOME BIOLOGY OF THE CULTIVATED POTATO, SOLANUM TUBEROSUM

By

Gina Mai Pham

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology – Doctor of Philosophy

2018

PUBLIC ABSTRACT

GENOME BIOLOGY OF THE CULTIVATED POTATO, SOLANUM TUBEROSUM

By

Gina Mai Pham

In the United States, potato is the top produced vegetable crop. Worldwide, they are the fourth most consumed food crop. Despite their cultural importance, high production value, and nutritional profile, genetic advancement in potato has been extremely slow in comparison to crops like rice, maize, and wheat. This is due to factors such as clonal propagation and genome complexity. The research described in this dissertation addresses components of these challenges – namely the genome complexity of potato and how this manifests in molecular traits, and how genome complexity can be reduced using a plant breeding technique called interspecific crossing. Finally, I present a valuable community resource for genetic studies in potato, an improved version of the potato genome.

ABSTRACT

GENOME BIOLOGY OF THE CULTIVATED POTATO, SOLANUM TUBEROSUM

By

Gina Mai Pham

Cultivated potato is a highly heterozygous, clonally propagated autotetraploid. These traits make it a difficult crop to study and make genetic improvements. In the following dissertation, I present studies that aim to improve our knowledge of genetic complexity in potato and potato breeding strategies. First, I show that several thousand genes in cultivated potato varieties show evidence of preferential allele expression, a characteristic not expected for autotetraploids. This trend was observed in evolutionarily conserved genes, suggesting that cultivated potato may have preferential expression of functional alleles. Cultivated potato also has excessive copy number variation. The results indicate that ~16-18,000 genes are copy number variable, and are evolutionarily recent and related to adaptation to biotic and abiotic stress. They are also lowly expressed, with only 528 genes showing correlation between copy number and gene expression. Second, a common method of genome reduction in potato, interspecific crossing, is explored to determine possible mechanisms by which genome elimination occurs and somaclonal variation which arises during the process. The results show that haploid inducer line, IVP101, produces <1% somatic translocation event frequency in the Superior dihaploid population studied. The translocation events occurred in regions of open chromatin, suggesting that they may be driven by transcription-coupled DNA repair. Finally, I present an improved potato genome assembly and annotation using a combination of long-read sequencing methods. The new assembly, DM v.5, is 727 Mb, of which 91% is

contained in 12 chromosome-scale scaffolds. DM v.5 presents a new opportunity for studies in comparative genomics and potato biology.

ACKNOWLEDGEMENTS

I am grateful for the support and mentoring of my advisor, Dr. Robin Buell. In her laboratory I operated as an independent scientist, but my efforts were bolstered by the community efforts of Robin and others in her group. I would also like to thank other members of my graduate committee, including Dr. Dave Douches, who has provided valuable aid and advice in potato research during my time at Michigan State University. Drs. Shin-Han Shiu and Ning Jiang have provided helpful comments and criticism, which have contributed to the quality of my research and my approach to science.

Drs. Richard Veilleux and Jiming Jiang (and members of their labs) have undoubtedly improved the quality of my work and provided substantial contributions. Additionally, members of the Buell lab have lent me their advice and expertise, and I am grateful for their efforts. I have had many great professors during my coursework, whose contributions are immeasurable. Thus, I am thankful to have had the opportunity to be a student at MSU.

I would like to acknowledge my parents, who have tried their best to understand and support my personal and professional decisions. I am also exceedingly grateful for my lifelong friendship with my cousin, Jennifer Truong. Finally, I am thankful for my partner Jonathan Flowers for his love, support, and friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

Cas9	Caspase 9
CNV	Copy number variation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic Acid
DM	Doubled Monoploid
FDR	First Division Restitution
IVP	In Vitro Pollinator
QTL	Quantitative Trait Locus
RNA	Ribonucleic acid
SDR	Second Division Restitution
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
Sli	S-locus Inhibitor
TPS	True Potato Seed

CHAPTER 1

INTRODUCTION

Challenges in potato breeding

Potato (*Solanum tuberosum*) is the fourth most consumed food crop worldwide and the top produced vegetable in the United States (http://www.ers.usda.gov). In 2016, 407,810 hectares were harvested in the United States, producing 19,990,950 tons (http://www.fao.org). Today, over 50% of potatoes produced in the U.S. are sold to companies for the production of processed potato foods, such as french fries and potato chips (http://www.ers.usda.gov). Part of potato's widespread success, in addition to being a palatable, versatile, and nutritious crop, is its storage capability. Although approximately 90% of U.S. potatoes are harvested in the fall, they can actually be sold well into spring due to their ability to be kept in climate-controlled storage for months at a time. From an economic standpoint, the United States has had a fruitful trade relationship with Mexico, Canada, and Japan from 2005 through recent years, exporting a surplus of potato in the form of frozen, processed products. This generates approximately $180 million in revenue yearly.

Despite its obvious importance, potato breeding efforts are dwarfed by efforts in other crops such as maize and wheat. This is in part due to its complex autotetraploid genetics, which will be described in the subsequent section, and several other factors. First, potato is propagated asexually via tuber cuttings called "seed potatoes." Potato is not propagated by seed (called "true potato seed," or TPS) because it is highly

heterozygous. Thus, any TPS produced will not breed true. Furthermore, inbreeding potato to move towards TPS production is currently untenable due to the genetic load in cultivated germplasm. Efforts to convert potato into a TPS-based crop had been intensively researched at the International Potato Center (CIP) for many years (Jansky *et al.* 2016).

Male sterility is also a limiting factor in potato breeding. Many wild diploid *Solanum* species, which are often used in potato breeding to provide genetic diversity in the breeding germplasm, can be crossed as males with dihaploids of cultivated potato and produce vigorous, male fertile offspring (Jansky and Peloquin 2006). However, some wild hybrid crosses and crosses using Groups Phureja and Stenotomum as male plants produce male sterile families, making it useless to carry through with phenotyping. Cytoplasmic genome types must be characterized and tracked in breeding programs to manage male sterility in the germplasm (Sanetomo and Gebhardt 2015). A study of European potato varieties found that male sterility is often found in D and W/γ-type cytoplasms that are derived from *Solanum demissum* and commonly found in breeding clones. These cytoplasm types correlate positively with late blight resistance and increased tuber starch content, respectively. Thus, through the introduction of positive traits, breeding programs may unintentionally have also introduced higher levels of male sterility into breeding populations.

Gametophytic self incompatibility in diploids limits development of inbred diploid populations, which could theoretically supplant tetraploid cultivars (Spooner *et al.* 2014). Recent efforts have focused on introgression of the S-locus inhibitor (*Sli)* gene from M6, a self-compatible individual of *Solanum chacoense* with a newly sequenced

genome (Leisner *et al.* 2018). Additionally, self-compatible diploid potatoes have also been generated by CRISPR-Cas9 knockout of S-RNAse, the stylar RNAse which digests pollen tubes of the same genotype in gametophytic self-incompatible plants (Ye *et al.* 2018).

Because of the heterozygosity of potato, its marked inbreeding depression, polyploidy, and the other abovementioned factors, potato breeding programs have progressed much more slowly than breeding programs in other crops. New cultivars are selected from heterozygous F1 progeny, which are vegetatively propagated and subjected to selection based on important agronomic traits. Unlike diploids, tetraploid potato can overcome self-incompatibility due to the diploid status of pollen (Spooner, et al. 2014). However, inbred tetraploid potatoes are quickly eliminated from any breeding programs due to their low vigor, reflecting the high genetic load in cultivated potato.

Genetic modification is one strategy for improving agronomic traits in potato while working around its heterozygosity. However, because of the heterozygosity, stacking traits through the traditional method of crossing is impractical (Davies *et al.* 2008). Thus, genetic modification relies on co-transformation. Monsanto Company released the first genetically modified potato varieties, called NewLeaf potatoes, which featured resistance to Colorado potato beetle and several viruses. However, it was discontinued due to consumer reaction and low support from processors and fast food companies. Amflora, a potato modified for production of amylopectin for industrial applications and animal feed, was created by BASF and approved for growth in Europe in 2010 (www.basf.com). Simplot Plant Sciences recently released two varieties of their cisgenic potatoes called Innate® potatoes, which contain several valuable traits, including

bruising reduction, low asparagine content, and late blight resistance, that were conferred via the introduction of genes from wild species and cultivated russet potatoes (www.innatepotatoes.com).

Genome editing using zinc finger nucleases and transcription activator-like effector nucleases are newer techniques relative to traditional genetic modification that have created possibilities for precise targeting of genes of interest (Jaganathan *et al.* 2018). However, these techniques require protein engineering and thus can be difficult to execute. The advent of CRISPR-Cas9 facilitates genome editing possibilities in potato and other crops. CRISPR-Cas9 typically operates by inducing non-homologous end joining at target sites (Butler *et al.* 2016). In addition to the abovementioned knockout of S-RNAse to create self-compatible diploid potato, CRISPR-Cas9 has been used in conjunction with a geminivirus delivery system to increase incidences of homology-directed repair in potato, a method that, in conjunction with more precise Cas9 enzymes, has the potential to truly tailor the effects of genome editing in potato.

Potato genetics: polyploidy and inheritance

Cultivated North American and European potato varieties have four copies each of 12 chromosomes and are thus autotetraploid (2n = 4x = 48 chromosomes). The autotetraploid state of potato most likely arose as a result of 2n gamete formation, which is common in *Solanum* species (Watanabe *et al.* 1991). 2n gametes occur as a result of one of three different mechanisms in *Solanum*, parallel spindles, first division restitution (FDR), or second division restitution (SDR) (Mok and Peloquin 1975). Analysis by

Watanabe, et al. (1991) indicate that potato tetraploids most likely arose from SDR x FDR or FDR x FDR 2n gamete hybridization.

Conventional breeding is difficult in autotetraploid potato because of the gametic configurations resulting from tetrasomic inheritance (Bradshaw 2007). In most cases during meiosis, potato chromosomes form bivalents that result in expected Mendelian inheritance. However, quadrivalents can also occur, wherein four homologous chromosomes associate during meiosis. In these cases, it is possible to observe double reduction, a case where two sister chromatids can be passed on to the same gamete, using molecular markers (Bourke *et al.* 2015). This special case further complicates genetic studies in potato.

Even in cases of simple Mendelian inheritance, the greater number of possible allele combination between loci creates greater complexity in phasing of alleles for linkage and QTL mapping (Manrique-Carpintero *et al.* 2018). Furthermore, even at single loci, there are three different types of intralocus interactions (dominance effects) that can result in non-additive effects. Finally, additive effects at any given loci may be modified by the dosage of alleles. The situation becomes even more complex when considered for quantitative traits. Altogether, breeding and genetic studies of autotetraploid potato are quite difficult as a result of tetrasomic inheritance.

It is possible to generate haploid (2n = 2x = 24 chromosomes) progeny from autotetraploid potato, referred to as dihaploid potato, with relatively easy effort using intraspecific crosses with inducer lines called in vitro pollinators (IVP). This reduces the number of possible combinations of intra- and interlocus interactions, simplifying genetic studies and expediting potato breeding. Furthermore, dihaploid lines can be used in

crosses with wild diploid potato species to introduce new genetic variation to the cultivated potato germplasm, which has been domesticated for growth in agricultural settings. Breeding at the diploid level, in tandem with other important characteristics for modern breeding programs (self-compatibility, elimination of inbreeding depression) present the newest vessel for rapid genetic gains in potato.

Potato in the era of genome biology

Genome sequencing has changed dramatically over the last two decades. The first plant genome to be sequenced, *Arabidopsis thaliana*, was generated from overlapping bacterial artificial chromosome (BAC) sequenced using the Sanger method, the first-generation sequencing technology (Initiative 2000). At present, plant genomes are being published using combinations of so-called next-generation sequencing (now sometimes termed second-generation sequencing), and third-generation sequencing technologies including single-molecule real-time sequencing (SMRT) and nanopore sequencing (Peterson 2018). Genome quality has vastly improved with the incorporation of newer sequencing and assembly techniques, and chromosome-level assemblies have nearly become the new standard in publication.

The potato genome was published in 2011, a time-frame that places it in the pre-third-generation sequencing era, using Illumina Genome Analyzer and Sanger sequencing (Potato Genome Sequencing Consortium 2011). To facilitate assembly, heterozygous, autotetraploid cultivars were eschewed in favor of a doubled monoploid genotype derived from a South American cultivar, *S. tuberosum* group Phureja DM1-3 516 R44 (DM). Although the third-generation sequencing technologies provide promising avenues

towards deconvolution of heterozygous genomes with longer read lengths, performing heterozygous assemblies still remains quite challenging in practice and robust bioinformatics pipelines for doing so have yet to surface. Thus, DM1-3 still stands as the standard genome assembly in the potato research community and has been used to develop genotyping arrays for use in potato breeding programs and other genetic studies of potato (Ellis *et al.* 2018, Hamilton *et al.* 2011, Vos *et al.* 2015).

In recent years, several other potato genome sequences have been published. The draft genome of *Solanum commersonii*, a wild tuber-bearing potato species critical for imparting stress-tolerance in potato breeding, was generated using paired-end and mate pair Illumina whole-shotgun sequencing (Aversano *et al.* 2015). The *S. chacoense* line M6 was sequenced using paired-end and mate pair Illumina sequencing (Leisner, et al. 2018). As described above, M6 is particularly important because of its ability to confer self-compatibility via the *Sli* gene (Jansky *et al.* 2014).

The advent of genome sequencing in the potato research community has contributed tremendously to our understanding of potato traits and potato genome biology. The DM reference genome, for example, has been used to identify new resistance genes using resistance gene enrichment sequencing (RenSeq), a technique in which nucleotide binding-site leucine-rich repeats (NLRs) are enriched using a target library based on genome annotations and sequenced using the Illumina platform (Jupe *et al.* 2013). RenSeq can be extended past studies in DM and has been used to rapidly clone the late blight resistance gene *Rpi-amr3i* from the *Solanum americanum* accession (Witek *et al.* 2016). This study was further accelerated by its use of SMRT sequencing on the PacBio RSII platform, which yielded full-length NLRs. Another study in potato

employed RenSeq and an additional technique, generic-mapping enrichment sequencing (GenSeq) to rapidly map the location of a late blight resistance gene in *Solanum verrucosum*.

In addition to mapping traits, genomic techniques and analyses have allowed researchers to study genome biology in potato at a more basic level. Potato contains a mixture of archetypal satellite repeat-based centromeres and repeat-less centromeres resembling neocentromeres (Gong *et al.* 2012). This was a surprising finding, especially given that other closely related *Solanum* species did not show the same characteristic. Using the DM sequence as a reference, it has also been determined that in addition to high heterozygosity at the SNP level, potato shows a startling amount of copy number variation (CNV) (Hardigan *et al.* 2016, Iovene *et al.* 2013). In fact, Hardigan, et al. (2016) demonstrated that approximately 30% of the genome was impacted by CNV in a study of 12 related *S. tuberosum* diploid clones. The work by Iovene, et al. (2013) shows that CNV in autotetraploid cultivars can span more than 100 kilobases of DNA and affect the expression of genes within CNV regions. Using comparative genomics methods, the inferred targets of potato domestication were recently described by Hardigan *et al.* (2017). This study demonstrated that genes under selection during domestication included pathways controlling circadian rhythm, endoreduplication, sexual reproduction, steroidal glycoalkaloid biosynthesis, carbohydrate metabolism, and disease resistance. Altogether, these genome-guided studies provide critical context in our understanding of plant genome biology as it relates to domesticated, vegetatively propagated crops like potato.

Dissertation outline and significance

This dissertation expands on several of the themes described above. Using computational approaches, I have characterized a set of autotetraploid North American cultivars to study the effects of CNV on gene expression and to explore the role of preferential allele expression in functional traits in potato. Next, I examined the prevalence of somatic translocation during dihaploid production using in vitro pollinators to produce gynogenetic progeny. Finally, I discuss work towards the assembly of a new reference potato genome using a combination of second- and third-generation sequencing technologies. The dissertation concludes with remarks on future work that can be pursued to provide a more complete understanding of potato genome biology and generation of dihaploid potatoes.

The biological conclusions drawn from these studies can inform studies in other vegetatively propagated, heterozygous crops. Additionally, other crops that utilize haploid production via inter- and intra-specific crosses will benefit from the conclusions drawn in this work. Finally, the potato research community, and other plant genome biologists, will greatly benefit from the availability of a new reference genome, which will provide a more accurate and contiguous representation of the potato genome.

REFERENCES

REFERENCES

Aversano, R., Contaldi, F., Ercolano, M.R., Grosso, V., Iorizzo, M., Tatino, F., Xumerle, L., Dal Molin, A., Avanzato, C., Ferrarini, A., Delledonne, M., Sanseverino, W., Cigliano, R.A., Capella-Gutierrez, S., Gabaldon, T., Frusciante, L., Bradeen, J.M. and Carputo, D. (2015) The Solanum commersonii Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. *The Plant cell*, **27**, 954-968.

Bourke, P.M., Voorrips, R.E., Visser, R.G.F. and Maliepaard, C. (2015) The Double-Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map. *Genetics*, **201**, 853-U894.

Bradshaw, J.E. (2007) The Canon of Potato Science: 4. Tetrasomic Inheritance. *Potato Research*, **50**, 219-222.

Butler, N.M., Baltes, N.J., Voytas, D.F. and Douches, D.S. (2016) Geminivirus-Mediated Genome Editing in Potato (Solanum tuberosum L.) Using Sequence-Specific Nucleases. *Front Plant Sci*, **7**, 1045.

Davies, H., Bryan, G.J. and Taylor, M. (2008) Advances in Functional Genomics and Genetic Modification of Potato. *Potato Research*, **51**, 283.

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**, 523-537.

Gong, Z., Wu, Y., Koblizkova, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novak, P., Buell, C.R., Macas, J. and Jiang, J. (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *The Plant cell*, **24**, 3559-3574.

Hamilton, J.P., Hansey, C.N., Whitty, B.R., Stoffel, K., Massa, A.N. and Deynze, A. (2011) Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics*, **12**.

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., Yang, X., Zeng, Z., Douches, D.S., Jiang, J., Veilleux, R.E. and Buell, C.R. (2016) Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. *The Plant cell*, **28**, 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., Farre, 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 of the United States of America*, **114**, E9999-E10008.

Initiative, A.G. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 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 : for cell and molecular biology*, **75**, 80-89.

Jaganathan, D., Ramasamy, K., Sellamuthu, G., Jayabalan, S. and Venkataraman, G. (2018) CRISPR for Crop Improvement: An Update Review. *Front Plant Sci*, **9**, 985.

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.

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

Jansky, S.H. and Peloquin, S.J. (2006) Advantages of Wild Diploid Solanum Species Over Cultivated Diploid Relatives in Potato Breeding Programs. *Genetic Resources and Crop Evolution*, **53**, 669-674.

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G.J., Maclean, D., Cock, P.J., Leggett, R.M., Bryan, G.J., Cardle, L., Hein, I. and Jones, J.D.G. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant Journal*, **76**, 530-544.

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 Buell, C.R. (2018) Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. *The Plant journal : for cell and molecular biology*, **94**, 562-570.

Manrique-Carpintero, N.C., Coombs, J.J., Pham, G.M., Laimbeer, F.P.E., Braz, G.T., Jiang, J., Veilleux, R.E., Buell, C.R. and Douches, D.S. (2018) Genome Reduction in Tetraploid Potato Reveals Genetic Load, Haplotype Variation, and Loci Associated With Agronomic Traits. *Front Plant Sci*, **9**, 944.

Mok, D.W.S. and Peloquin, S.J. (1975) THREE MECHANISMS OF 2n POLLEN FORMATION IN DIPLOID POTATOES. *Canadian Journal of Genetics and Cytology*, **17**, 217-225.

Peterson, D.G., Arick, M. (2018) Sequencing Plant Genomes. In *Progress in Botany*: Springer, Berlin, Heidelberg.

Potato Genome Sequencing Consortium, T. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189-195.

Sanetomo, R. and Gebhardt, C. (2015) Cytoplasmic genome types of European potatoes and their effects on complex agronomic traits. *Bmc Plant Biol*, **15**, 162.

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**, 283+.

Vos, P.G., Uitdewilligen, J.G.A.M.L., Voorrips, R.E., Visser, R.G.F. and van Eck, H.J. (2015) Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. *Theoretical and Applied Genetics*, **128**, 2387-2401.

Watanabe, K., Peloquin, S.J. and Endo, M. (1991) Genetic significance of mode of polyploidization: somatic doubling or 2n gametes? *Genome*, **34**, 28-34.

Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D. and Jones, J.D.G. (2016) Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nature biotechnology*, **34**, 656-660.

Ye, M., Peng, Z., Tang, D., Yang, Z., Li, D., Xu, Y., Zhang, C. and Huang, S. (2018) Generation of self-compatible diploid potato by knockout of S-RNase. *Nat Plants*, **4**, 651-654.

CHAPTER 2

EXTENSIVE GENOME HETEROGENEITY LEADS TO PREFERENTIAL ALLELE EXPRESSION AND COPY NUMBER-DEPENDENT EXPRESSION IN CULTIVATED POTATO

This chapter was published in the following manuscript:

Gina M. Pham, Linsey Newton, Krystle Wiegert-Rininger, Brieanne Vaillancourt, David S. Douches, C. Robin Buell (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number dependent expression in cultivated potato. *The Plant Journal* 92(4): 624-637.

CHAPTER 3

GENOME-WIDE INFERENCE OF SOMATIC TRANSLOCATION DURING POTATO DIHAPLOID PRODUCTION

This chapter was published in the following manuscript:

Gina M. Pham, Guilherme T. Braz, Megan Conway, Emily Crisovan, F. Parker E. Laimbeer, Norma Manrique-Carpintero, Linsey Newton, Davis S. Douches, Jiming Jiang, Richard E. Veilleux, C. Robin Buell (2019). Genome-wide inference of somatic translocation during potato dihaploid production. *The Plant Genome* doi: 10.3835/plantgenome2018.10.0079.

CHAPTER 4

HYBRID GENOME ASSEMBLY OF *SOLANUM TUBEROSUM*

This chapter is in preparation for submission as a Data Note to *GigaScience*.

ABSTRACT

Worldwide, the cultivated potato, *Solanum tuberosum* L., is the number one vegetable crop. The genome sequence of DM1-3 516 R44, a doubled monoploid clone of S. *tuberosum* Group Phureja, was published in 2011 using a whole-genome shotgun sequencing approach using short read sequence data. Here, we present an updated version of the genome sequence using a hybrid assembly approach combining sequence data from the 10X Chromium and PacBio SMRT sequencing platforms with Hi-C reads generated using the Illumina platform. The new (v5) 727 Mb assembly represents 77.6% of the estimated 844 Mb genome encoding 39,900 protein-coding genes. The new sequence improves upon the previous version in terms of contiguity with a 13.1-fold reduction in numbers of contigs, a 14-fold increase in N50 contig size, a 41.1-fold increase in N50 scaffold size and a reduction in percentage of gap sequence, providing an improved resource for the potato research community.

INTRODUCTION

The genome of the vegetable crop potato (*Solanum tuberosum* L.) was published in 2011 by the Potato Genome Sequencing Consortium using a whole-genome shotgun (WGS) sequencing approach (The Potato Genome Consortium 2011). At that time, Illumina sequencing was a newly available approach with high accuracy and throughput relative to previously available technologies, yet short read lengths. To create a high quality genome assembly, it was necessary to use a doubled monoploid clone, DM1-3 516 R44 (hereafter referred to as DM) (Figure 4.1), to reduce assembly difficulties due to the heterozygous and polyploid nature of tetraploid potato. The PGSC DM genome was assembled using a combination of 36 nucleotide (nt) reads from the Illumina Genome Analyzer platform and longer end sequence reads from fosmid and bacterial artificial chromosome clones generated using Sanger sequencing technology. This resulted in a fragmented genome assembly, with 90% of the assembly contained in 443 super-scaffolds. The assembly (DM v.4.03) was anchored genetically to the 12 chromosomes of potato using 2,603 markers although a substantial fraction of the super scaffolds were ordered relative to the genetic map, not all super-scaffolds were oriented relative to the genetic map (Sharma *et al.* 2013). Since the initial publications, a new version of the genome was constructed (DM v.4.04) which includes the unscaffolded contigs (Hardigan *et al.* 2016).

The published sequence has undoubtedly served as a valuable resource in the plant genomics and potato genetics community as indicated by numerous publications that utilized the sequence (Kloosterman *et al.* 2013, Manrique-Carpintero *et al.* 2018,

Uitdewilligen *et al.* 2013, Witek *et al.* 2016). However, its quality and potential is limited by the technology that was available at the time of its publication and new technologies and approaches for genome sequencing and assembly, including long-read technologies and chromatin contact map-based strategies (Jiao and Schneeberger 2017), present new opportunities to improve upon the sequence of the potato genome. In this data note, the doubled monoploid clone DM was sequenced using a combination of 10X Genomics Chromium and PacBio SMRT sequencing technologies and assembled into a highly contiguous pseudochromosomes using Hi-C scaffolding data.

Using this approach, 91% of the 731 Mb assembly was contained in 12 chromosome-scale scaffolds that encodes 39,900 genes. The improved assembly, DM v.5, improves upon contiguity in comparison to DM v.4.04, with fewer gap sequence and longer contigs, allowing for more accuracy in future studies on potato genome biology, especially those requiring accurate intergenic sequence.

DATA DESCRIPTION

DNA isolation, library construction, and sequencing

DNA for Chromium and PacBio sequencing was isolated from young leaf tissue using the CTAB method with RNAse A digestion (Saghai-Maroof et al. 1984). Chromium library construction was completed by the Van Andel Research Institute (Grand Rapids, MI, USA) and sequenced on the Illumina HiSeq 4000 in paired-end mode with 150 nt read lengths, producing 301,117,288 purity filtered pairs and 90.3 Gb of sequence. PacBio library preparation and sequencing was completed by University of Minnesota Genome Center, producing 105.6 Gb of sequence. Tissue for Hi-C scaffolding

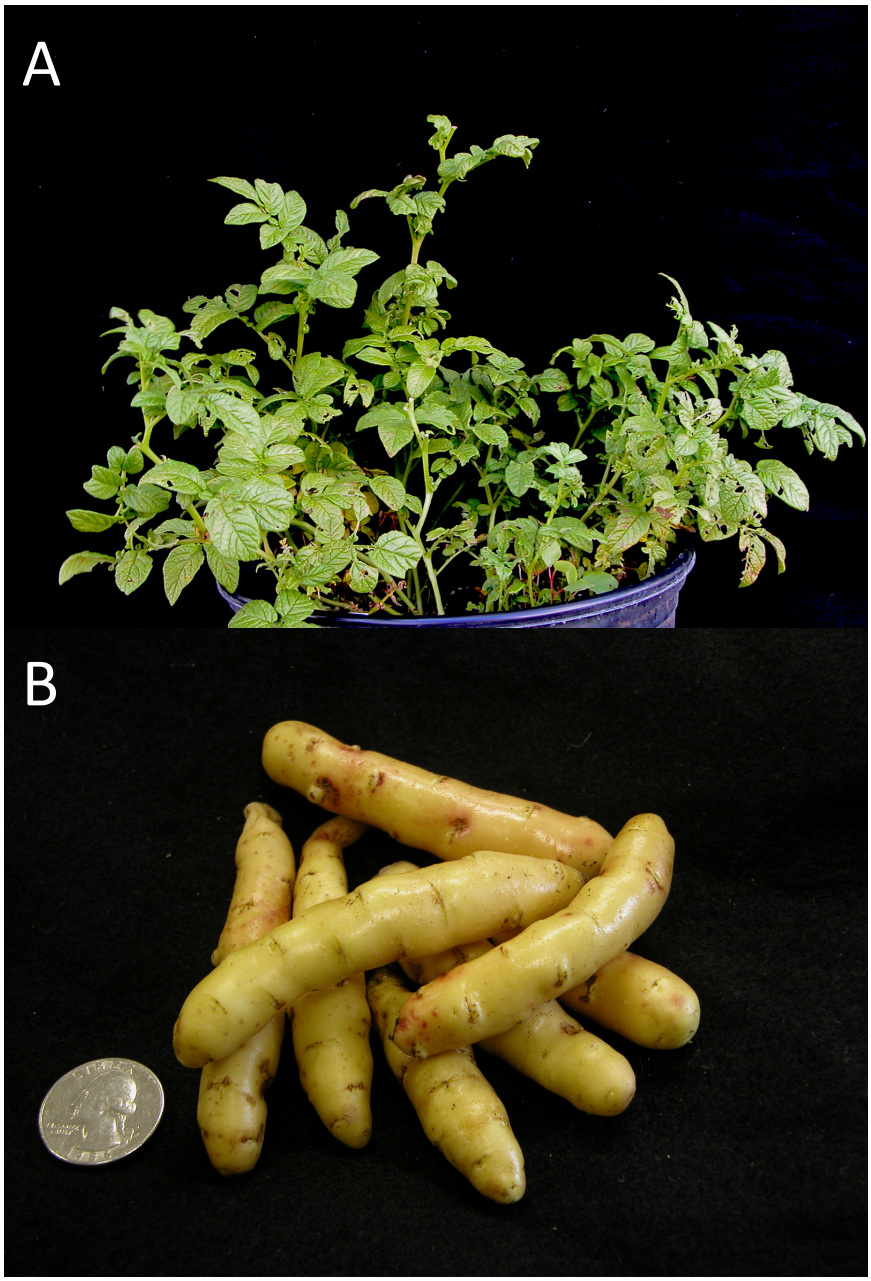

Figure 4.1 Doubled monoploid potato clone, DM1-3.

A) Aboveground tissues and B) tubers from the doubled monoploid potato clone, DM1-3 516 R44.

was processed by Phase Genomics (Seattle, WA) for DNA isolation, Hi-C library preparation, sequencing, and analysis.

Hybrid assembly using Chromium, Hi-C, and PacBio sequencing

Chromium reads were assembled using Supernova v. 2.0.1 (Weisenfeld *et al.* 2017) with 326 million reads (approximately 56X theoretical coverage of the potato genome). This initial assembly consisted of 18,168 contigs with an N50 contig size of 233.91 Kb and N50 scaffold size of 1.90 M. The total assembly size, including only scaffolds greater than or equal to 10 Kb, was 654.65 Mb. Sequencing and scaffolding of a Hi-C library yielded additional contigs after splitting chimeric contigs produced by Supernova. As a consequence, the N50 contig size decreased slightly to 233.16 Kb, while the N50 scaffold size increased to 54.02 Mb in 12 chromosome-scale scaffolds. The Hi-C scaffolding placed 91.36% of the initial Chromium assembly onto 12 pseudochromosomes (Figure 4.2) representing 664,302,635 bp and a total assembly size

Table 4.1 Assembly metrics

	DM v4.04	10X Assembly	Hi-C Assembly	Gap-filled Hi-C (DM v.5)
No. of Ctgs	199,527	18,168	18,194	3,267
Ctgs >1K nt	44,251 (22.2%)	17,555 (96.6%)	17,578 (96.6%)	3,161 (96.8%)
Ctgs>10K nt	19,098 (9.6%)	5,592 (30.8%)	5,605 (30.8%)	2,903 (88.9%)
Ctgs>100Knt	334 (0.2%)	2,043 (11.2%)	2,048 (11.3%)	1,715 (52.5%)
Ctgs > 1M nt	0	17 (0.1%)	16 (0.1%)	76 (2.3%)
Ctgs N50	30,171	233,906	233,158	456,753

of 727,087,237 bp; the unscaffolded contigs showed Hi-C signals indicative of short repetitive sequences.

The initial Hi-C scaffolded assembly contained 23,951,420 Ns. To reduce this number, PacBio SMRT reads were used for gap filling. The reads were first error-corrected with ~456 million reads using FMLRC (Wang *et al.* 2018) and the assembly was subsequently gap-filled using PBSuite (English *et al.* 2012), reducing the number of Ns in the anchored sequences to 19,055,399, or approximately 2.3% of the theoretical genome size of 844 Mb (Table 4.2). This is an improvement upon the PGSC genome version 4.03, in comparison, which contains 90,916,687 N bases in the scaffolded assembly (chromosomes 1–12). Gap filling improved the DM v.5 contig N50 and scaffold N50 to 456.75 Kb and 55.45 Mb, respectively. DM v.4.04, in comparison, has a contig N50 of 30.17 Kb and scaffold N50 of 1.32 Mb (The Potato Genome Consortium 2011). The improved assembly represents a 13.1-fold reduction in numbers of contigs, a 15-fold increase in N50 contig size and a 42-fold increase in N50 scaffold size.

Total pseudomolecule lengths decreased for all chromosomes in DM v.5 versus DM v.4.04. The mean N content for each chromosome decreased from 7.58 Mb to 1.56 Mb, while the mean chromosome length decreased from 55.32 Mb to 50.86 Mb suggesting that the discrepancy in chromosome lengths is due to the decreased percentage of gap sequences (Table 4.2).

The software Benchmarking Universal Single-Copy Orthologs (BUSCO) v.3 using the Embryophyta *odb9* database was used to estimate completeness of the DM v.5 genome (Simao *et al.* 2015). Of 1,440 total BUSCOs in the database, the DM v.4.04 assembly yielded 1,392 complete BUSCOs (96.7% completeness), of which, 1,358 were

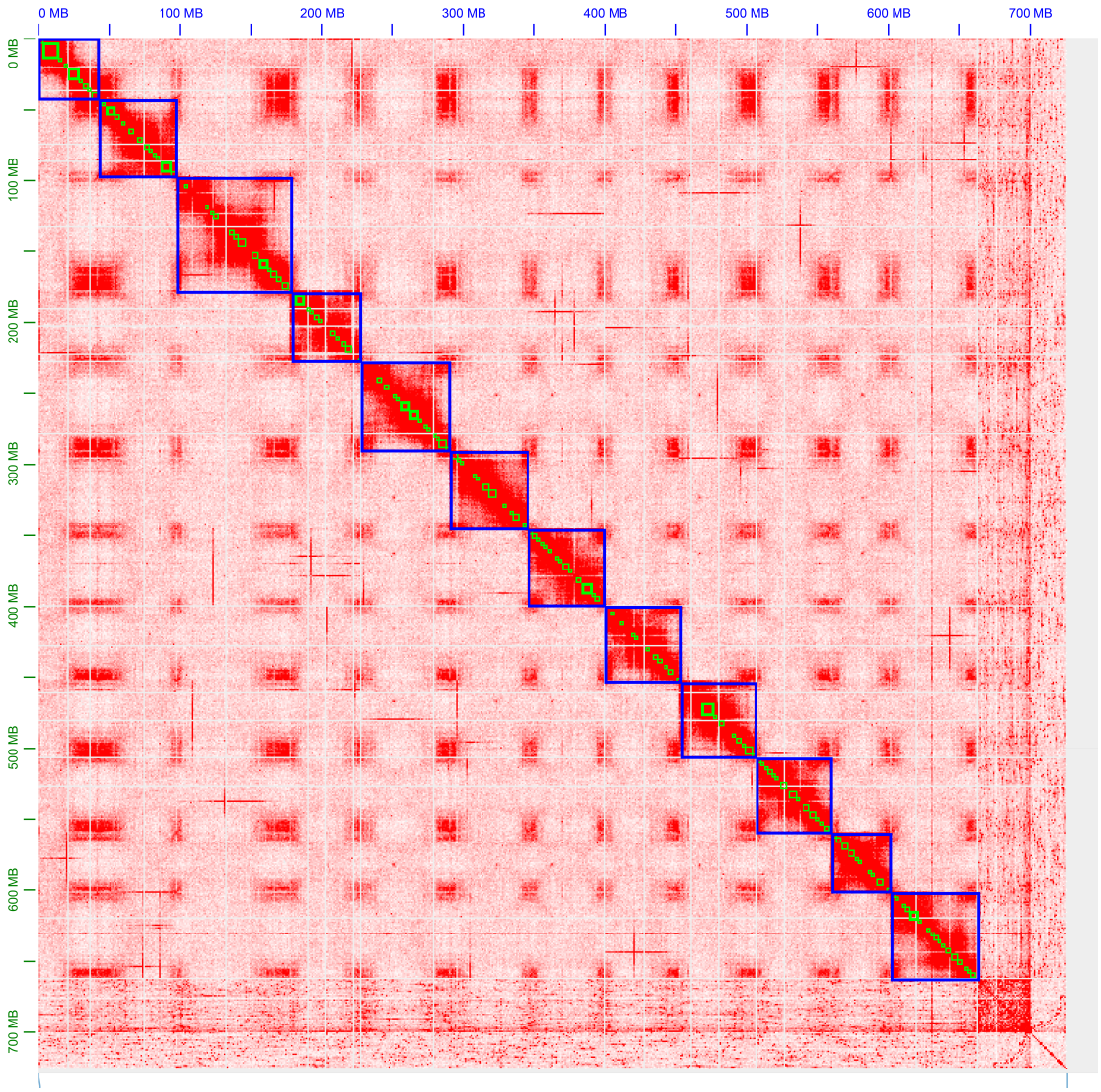

Figure 4.2. Hi-C contact map showing the inter- and intra-chromosomal chromatin interactions.

Interchromosomal chromatin interactions are off the diagonal axis and intra-chromosomal chromatin interactions are within the blue boxes. Each pixel represents the degree of interaction between each 1 Mb locus, with a dark red color indicating a greater number of reads involved in the interaction.

single copy and 34 were duplicated (Table 4.3). The initial 10X Genomics Chromium assembly yielded identical results, while the gap-filled Hi-C assembly showed a slight decrease in complete BUSCOs (96.6% completeness). The results show that even though the DM v.4.04 assembly was generated using short-read technologies, it provided an accurate and complete representation of the genic space. While DM v.5 does not improve upon DM v.4.04 in terms of completeness of the gene space, the contig number and length demonstrate that it exceeds the quality of DM v.4.04 in terms of contiguity.

Validation of the v5 assembly using a genetic map

A genetic map constructed from a DM x RH F1 population consisting of 190 individuals was used to validate the order and orientation of scaffolds placed within the

Table 4.2 Chromosome lengths and gap (N) content

Chromosome	DM v.4.04	DM v.4.04 'N' bases	% 'N' bases	DM v.5	DM v.5 'N' bases
1	88,663,952	10,769,373	12.14	80,968,998	2,370,935
2	48,614,681	5,917,869	12.17	43,009,707	1,083,567
3	62,290,286	8,361,451	13.42	55,377,174	1,319,613
4	72,208,621	10,005,054	13.85	62,853,833	2,281,823
5	52,070,158	5,459,790	10.48	49,521,837	1,462,773
6	59,532,096	7,887,319	13.24	53,775,810	1,615,519
7	56,760,843	7,210,540	12.70	52,326,458	1,422,440
8	56,938,457	7,638,281	13.41	53,351,194	1,643,656
9	61,540,751	7,649,184	12.42	61,253,832	1,551,746
10	59,756,223	7,406,736	12.39	55,110,485	1,623,607
11	45,475,667	5,347,499	11.75	42,736,230	1,128,533
12	61,165,649	7,263,591	11.87	54,017,077	1,213,975

DM v.5 pseudomolecules (Figure S 4.1) (Manrique-Carpintero *et al.* 2015). The map was generated using 2,621 SNP markers placed within 654 recombination bins and manually adjusted to eliminate incorrect bins. Vmatch (Abouelhoda *et al.* 2004) with 200 bp of flanking sequence around each marker was used in alignments against DM v.5 (Figure 4.3); the alignments demonstrate a high degree of congruity between the physical and genetic distances, with the exception of a large inversion on chromosome 3 (Figure S4.1).

To correct the misassembled region on chromosome 3, we identified nine scaffolds that spanned the entire inverted sequence. Two scaffolds were identified by aligning scaffolds to the assembly and identifying those which spanned both the inverted region and the candidate breakpoint regions. The gaps between these two scaffolds and their adjacent, properly placed scaffolds, were used as breakpoints to correct the misassembly via reverse complementation of the inverted scaffold.

Comparison of short-read alignments to DM v.4.04 and DM v.5

As an additional measure of quality, ~459 million paired-end reads from whole-genome

Table 4.3 BUSCO scores from genome assemblies

	DM v.4.04	Chromium Assembly	Hi-C Assembly
Complete BUSCOs	96.7%	96.7%	96.6%
Complete and single-copy BUSCOs	94.3%	94.3%	94.0%
Complete and duplicated BUSCOs	2.4%	2.4%	2.6%
Fragmented BUSCOs	0.8%	0.8%	0.8%
Missing BUSCOs	2.5%	2.5%	2.6%

Illumina sequencing and ~61 million reads from two different RNA-seq libraries prepared from leaf and tuber tissue were mapped to the DM v.5 and DM v.4.04 genomes, and the read mapping statistics were compared. Before read mapping, the adapters were removed and low quality bases (Q < 20) were trimmed using Cutadapt (Martin 2011). The WGS reads were mapped using BWA-MEM (Li 2013) and the RNA-seq reads were mapped using HISAT2 (Kim *et al.* 2015). WGS read mapping rates to both genomes are excellent (99.75% in DM v.4.04 and 99.67% in DM v.5), though more reads map with the correct paired orientation in DM v.5 compared to DM v.4.04 (98.35% and 96.84%, respectively) (Table S 4.1). Slightly more RNA-seq reads align to DM v.4.04 (Table S 4.2). This is due to a slightly greater number of multi-mapping reads and single reads that did not map with their pairs. However, more RNA-seq reads mapped concordantly (with properly paired orientation) using the DM v.5 genome. The low read mapping rate (~65-67%) in the RNA-seq libraries is due to the presence of potato virus X in the sample.

Evaluation of genome quality using LTR assembly index

The genome metric LTR Assembly Index (LAI) (Ou *et al.* 2018) was used to evaluate assembly continuity. This method identifies intact LTR retrotransposons in the genome to generate an overall LAI score for the assembly. Higher LAI scores correspond more complete genome assemblies, as a greater number of intact LTR retrotransposons are identified in these cases. The LAI analysis was applied to DM v.4.04 and DM v.5. The DM v.4.04 genome was found to have an LAI score of 8.26, a score that characterizes it as a draft genome assembly. In comparison, DM v.5 was found to have an

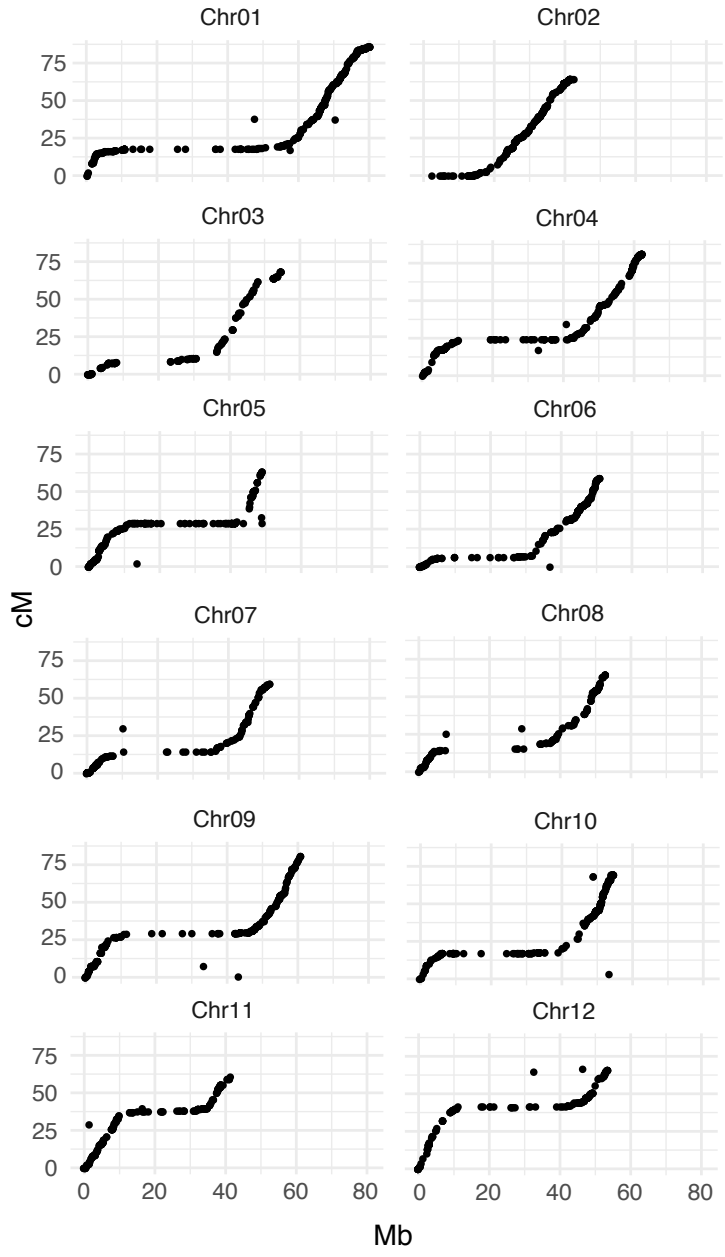

Figure 4.3 Mapping of the DM x RH F1 population markers to the DM v.5 assembly.

200 base pairs of flanking sequence around the markers were used for sequence alignments to the assembly using Vmatch (Abouelhoda, et al. 2004). The y-axis shows the map location in centimorgans and the x-axis shows the physical location in megabases.

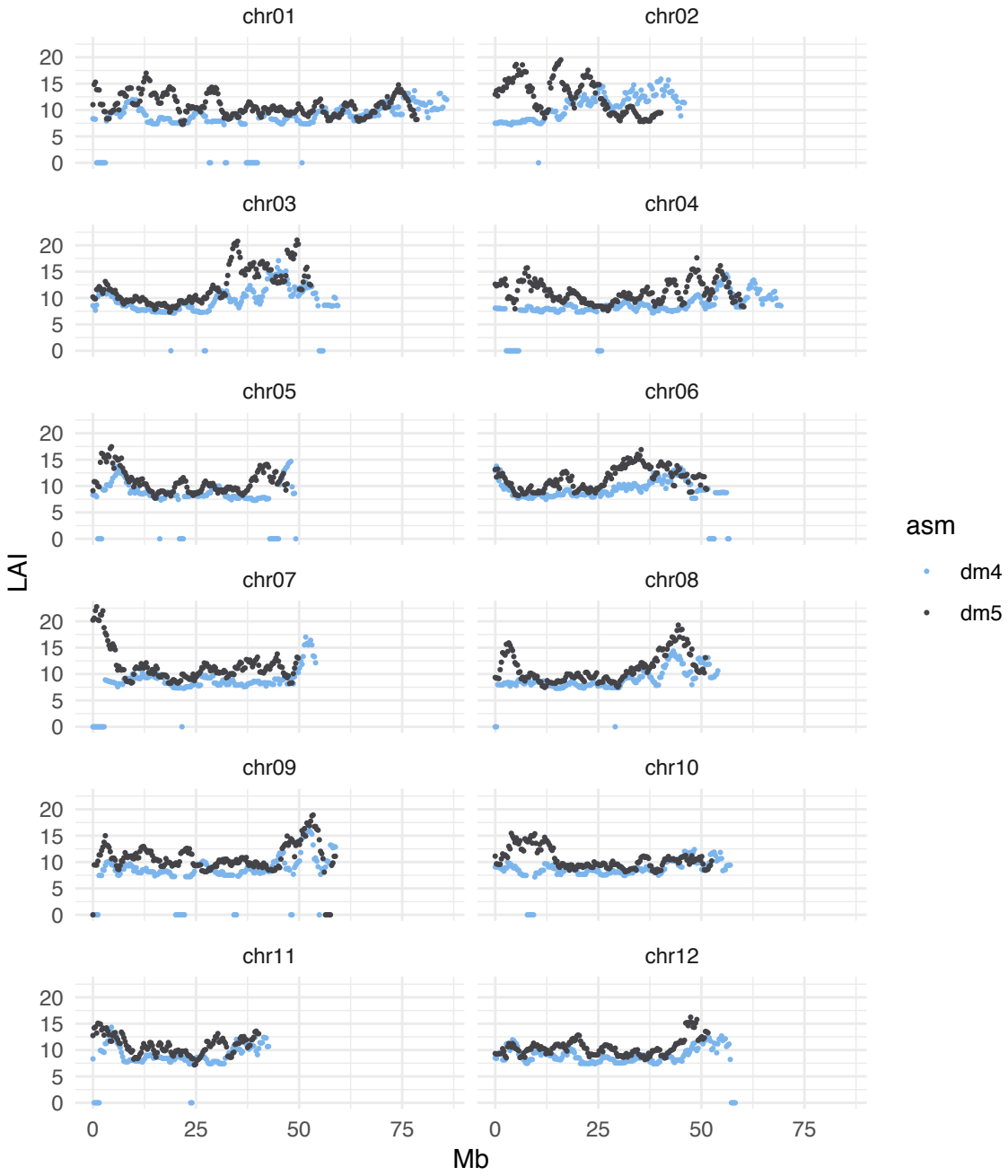

Figure 4.4. Genome-wide LAI scores for DM assembly v.4.04 (dm4) and v.5 (dm5).

LAI was calculated for 3 Mb sliding windows with a 300 Kb step size.

improved LAI score of 10.24, placing it in the category of reference genome quality. The LAI score was also calculated in sliding 300 Kb windows, showing noticeably higher scores in DM v.5 with far fewer cases of windows with an LAI score of 0 which indicates cases where no intact LTRs were found in the window (Figure 4.4). This further demonstrates the improved completeness and accuracy of the DM v.5 assembly.

Repeat masking of the v5 genome assembly

A new custom repeat library was generated by running RepeatModeler v.1.0.8 (Smit *et al.,* 2015) on the new assembly and combining the output with RepBase (Jurka 1998) repeats downloaded on September 13, 2018. Non-transposable element encoding proteins were removed from the combined file by identifying BLASTX (BLAST+ v.2.6.0) hits to a curated set of UniRef proteins. Repeat masking of the v.5 genome assembly was then performed using RepeatMasker v.4.0.6m using the custom repeat library. In total, 56.82% of the genome was repeat masked. The majority of the repeats (38.23%) were unclassified, and 17.11% of the repeats were long terminal repeats (Table S 4.3).

Initial gene prediction results

The gene prediction software Augustus v.3.3 (Stanke and Waack 2003) was trained using the DM v.5 genome and a training set comprised of 150 nt paired-end RNA-seq reads from leaf tissue. To prepare the training set, Illumina adapters and bases with quality scores lower than 20 were removed from the reads using Cutadapt v.1.14

(Martin 2011) and the cleaned reads were aligned to DM v.5 using HISAT2 v.2.1.0 (Kim, et al. 2015). Potential transcripts were assembled from the alignments using Stringtie v.1.3.4d (Pertea *et al.* 2015) with a minimum assembled transcript length of 200 nt. Coding sequences in the predicted transcripts were predicted using TransDecoder v.3.0.1 (Haas *et al.* 2013) and used to train Augustus. Genes were predicted *ab initio* from the hardmasked assembly, producing a set of 39,900 genes. The completeness of the gene set was evaluated using BUSCO in "protein" mode (Table S 4.4). The initial gene predictions included 76.1% complete BUSCOs, which is somewhat lower than the complete BUSCOs in DM v.4.04 (85.5%). This lower number seems to be due to a higher number of fragmented BUSCOs (14.6% in DM v.5 compared to 6.5% in DM v.4.04). To address this issue, a refined gene set will be produced to join fragmented gene models using PASA (Haas *et al.* 2003).

CONCLUSION

Using a hybrid sequencing approach, the potato genome sequence was vastly improved in contiguity relative to the previous release, DM v. 4.04. The N50 contig size improved nearly 14-fold in DM v.5 while the N50 scaffold size was increased 41.1-fold and several megabases of gap sequence were removed from each chromosome. BUSCO analyses demonstrate that the gene space in v.5 is equivalent to v4.04. Thus, the introduction of new sequencing methods largely improved upon the intergenic regions in the genome and on sequence contiguity. This information can be used in future studies to improve our understanding of regulatory sequences in the potato genome.

Acknowledgements

This work was supported by a grant awarded from PepsiCo to CRB and a USDA NIFA Predoctoral Fellowship (2017-67011-26038) awarded to GMP. The authors acknowledge their colleague Mandy Waters from PepsiCo who aided in the organization of the project, Kayla Young from University of Minnesota who aided in sample preparation and sequencing, and Shawn Sullivan from Phase Genomics for work on Hi-C scaffolding.

Supplemental materials in appendix

Figure S 4.1. Corrected misassembly on chromosome 3

Table S 4.1. WGS read mapping statistics

Table S 4.2. RNA-seq read mapping statistics

Table S 4.3. Repeat content in the genome

Table S 4.4. BUSCO scores from protein sets

Availability of supporting information

The raw genomic sequences will be made available in the NCBI Sequence Read Archive database in PRJNAXXXXXX. The genome assembly and annotation will be made available in Dryad Digital Repository and on Spud DB.

APPENDIX

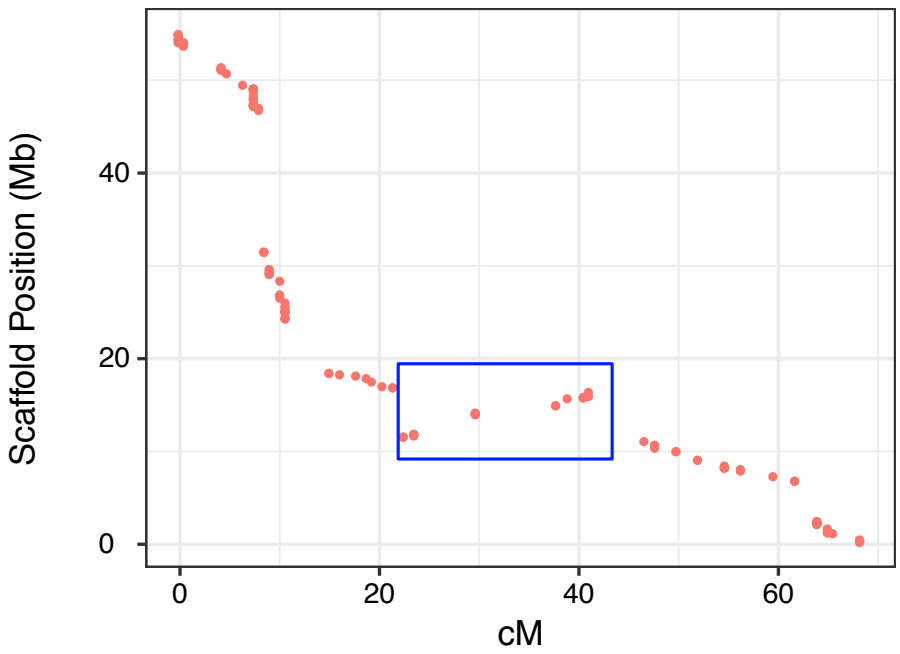

Figure S 4.1. Physical position of genetic markers from the DM x RH F1 population mapped to scaffold dm52 (corresponding to chromosome 3).

The blue box highlights a large inversion in the assembly, which was corrected by reverse complementing the sequence between gaps that flanked the misassembled region.

Table S 4.1 WGS read mapping statistics

	DM v.5	DM v.4.04
Mapped WGS	456,750,455 (99.67%)	458,040,388 (99.75%)
Properly paired WGS	449,062,316 (98.35%)	442,148,000 (96.84%)

Table S 4.2 RNA-seq read mapping statistics

	Leaf DM v.5	Leaf DM v.4.04	Tuber DM v.5	Tuber DM v.4.04
total reads	35,629,918	35,629,918	26,447,522	26,447,522
aligned concordantly 1 time	22,113,314 (62.06%)	22,056,183 (61.90%)	16,123,287 (60.96%)	15,837,406 (59.88%)
aligned concordantly >1 time	838,276 (2.35%)	1,344,100 (3.77%)	956,089 (3.62%)	1,451,062 (5.49%)
aligned discordantly 1 time	366,454 (2.89%)	361,126 (2.95%)	207,169 (2.21%)	194,304 (2.12%)
did not align concordantly or discordantly (includes single reads mapped without pair)	12311874 (34.55%)	11868509 (33.31%)	9160977 (34.64%)	8964750 (33.90%)
overall alignment rate	66.25%	67.48%	65.97%	66.76%

Table S 4.3 Repeat content in the genome

	Number of elements	Length occupied	Percentage of sequence
LINEs	15,852	6,020,091	0.82%
LTR elements	120,843	125,088,123	17.11%
DNA elements	9,493	4,770,192	0.65%
Unclassified	491,724	279,499,800	38.23%
Total interspersed repeats	637,912	415,378,206	56.82%

Table S 4.4 BUSCO scores from protein sets

	DM v.4.04	DM v.5
Complete BUSCOs	85.5%	76.1%
Complete and single-copy BUSCOs	83.5%	73.7%
Complete and duplicated BUSCOs	2.0%	2.4%
Fragmented BUSCOs	6.5%	14.6%
Missing BUSCOs	8.0%	9.3%

REFERENCES

REFERENCES

Abouelhoda, M.I., Kurtz, S. and Ohlebusch, E. (2004) Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, **2**, 53-86.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C. and Gibbs, R.A. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, **7**, e47768.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L. and White, O. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, **31**, 5654-5666.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N. and Regev, A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**, 1494.

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., Yang, X., Zeng, Z., Douches, D.S., Jiang, J., Veilleux, R.E. and Buell, C.R. (2016) Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. *The Plant cell*, **28**, 388-405.

Jiao, W.B. and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Curr Opin Plant Biol*, **36**, 64-70.

Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol*, **8**, 333-337.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, **12**, 357-360.

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

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997 [q-bio.GN].

Manrique-Carpintero, N.C., Coombs, J.J., Cui, Y., Veilleux, R.E., Buell, C.R. and Douches, D. (2015) Genetic Map and QTL Analysis of Agronomic Traits in a Diploid Potato Population using Single Nucleotide Polymorphism Markers. *Crop Science*, **55**, 2566-2579.

Manrique-Carpintero, N.C., Coombs, J.J., Pham, G.M., Laimbeer, F.P.E., Braz, G.T., Jiang, J., Veilleux, R.E., Buell, C.R. and Douches, D.S. (2018) Genome Reduction in Tetraploid Potato Reveals Genetic Load, Haplotype Variation, and Loci Associated With Agronomic Traits. *Front Plant Sci*, **9**, 944.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**.

Ou, S.J., Chen, J.F. and Jiang, N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*, **46**.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, **33**, 290-295.

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., Eguiluz, M., Guo, X., Guzman, F., Hackett, C.A., Hamilton, J.P., Li, G., Li, Y., Lozano, R., Maass, A., Marshall, D., Martinez, D., McLean, K., Mejía, N., Milne, L., Munive, S., Nagy, I., Ponce, O., Ramirez, M., Simon, R., Thomson, S.J., Torres, Y., Waugh, R., Zhang, Z., Huang, S., Visser, R.G.F., Bachem, C.W.B., Sagredo, B., Feingold, S.E., Orjeda, G., Veilleux, R.E., Bonierbale, M., Jacobs, J.M.E., Milbourne, D., Martin, D.M.A. and Bryan, G.J. (2013) Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps. *G3: Genes|Genomes|Genetics*, **3**, 2031-2047.

Simao, 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**, 3210-3212.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19 Suppl 2**, ii215-225.

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189-195.

Uitdewilligen, J.G., Wolters, A.M., D'hoop, B.B., Borm, T.J., Visser, R.G. and Eck, H.J. (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS One*, **8**.

Wang, J.R., Holt, J., McMillan, L. and Jones, C.D. (2018) FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, **19**, 50.

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

Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D. and Jones, J.D.G. (2016) Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nature biotechnology*, **34**, 656-660.

CHAPTER 5

CONCLUDING REMARKS

Preferential expression of alleles in potato

Cultivated potato is highly heterozygous due to vegetative propagation, outcrossing, and the preservation of heterozygosity is thought to be due to the maintenance of deleterious alleles in the germplasm. Recently, inbreeding depression in potato was characterized by Zhang *et al.* (2019) in 151 diploid potatoes. The study found 344,831 predicted deleterious substitutions in this population, and these mutations were enriched in pericentromeric regions. The study supported the theory of complementation of deleterious alleles as a driving factor in the maintenance of high heterozygosity in potato in that it found that there were 64.46% fewer homozygous mutations than heterozygous mutations.

The results from this dissertation demonstrate that thousands of genes show skewing in expression of alleles in potato (Pham *et al.* 2017). However, the functional mechanism of this observation has not been fully characterized. Though the genes showing preferential allele expression are involved in a diverse array of molecular functions, there was statistical enrichment in iron-sulfur cluster binding and metal-cluster binding in the potato cultivars Atlantic, Kalkaska, and Missaukee leaf samples. Iron-sulfur proteins are critical for many functional processes in plants, especially in chloroplasts and mitochondria (Couturier *et al.* 2013). The enrichment of iron-sulfur cluster binding and metal-cluster binding proteins in leaves corresponds well with the metabolic pathways containing genes with preferential allele expression, which included components from photosystems II and I, where iron-sulfur proteins participate in electron-transfer reactions. Undoubtedly, the presence of deleterious alleles in these critical genes would contribute to inbreeding depression if present in a homozygous state.

A future study characterizing the effect of SNPs in genes showing preferential allele expression could provide more evidence to support the theory that complementation of deleterious alleles plays an important role in the maintenance of heterozygosity in cultivated potato. For example, it would be informative to determine if the highly expressed allele is associated with a functional copy of the gene. This would provide more direct evidence that plant productivity is associated with preferential expression of functional alleles. Additionally, the mechanism of regulation could be further studied to determine if *cis*- or *trans*-regulation is responsible for the expression abundance of alleles. This can be achieved by quantifying the abundance of alleles expressed for any particular gene in the parents of an individual and comparing these abundances to those from the F1 individual (Springer and Stupar 2007).

Mechanisms of somatic translocation using in vitro pollinators

The production of dihaploids is a critical component of potato breeding programs for the purposes of creating genetic compatibility with diploid potato varieties and species. The incidental translocation of alleles from inducer lines during haploid production is regarded as a negative outcome. However, outside of their utility in breeding programs, haploid-inducing plants have not been studied in the context of plant evolution. In this dissertation, evidence was presented showing low levels of somatic translocation where IVP-alleles replaced Superior alleles in a dihaploid population. However, the trends in allele conversion were not further characterized to determine if there was any bias in GC content. Total GC content in angiosperms varies greatly between different species (Glemin *et al.* 2014). The cause of GC content variation has not

been well defined in angiosperm genomes. In mammalian genomes, research has shown that GC content is largely driven by GC-biased gene conversion. This trend appears to be true for some plants, but more information is needed to confirm the role of GC-biased gene conversion in plant genome evolution (Pessia *et al.* 2012). GC-biased gene conversion is believed to occur during meiosis during crossover and non-crossover resolution of double-stranded breaks. A recent study by Liu *et al.* (2018), however, shows that this assumption does not apply in their study of *Saccharomyces*, *Neurospora*, *Chlamydomonas*, and *Arabidopsis.* The organisms showed that GC content was correlated with recombination rate, but that there was no significant GC conversion bias in these species. In previous studies, the correlation between GC content and recombination rate was an indicator of GC-biased gene conversion.

Somatic translocation during haploid production, in contrast to meiosis-coupled GC-biased gene conversion, occurs in a non-meiotic context. The findings from this dissertation indirectly show that translocation is correlated with recombination rate and open chromatin. The translocation events may be coupled to homologous recombination during double-stranded break (DSB) repair. In plants, DSBs can occur in the DNA as a result of oxidative damage (Hu *et al.* 2016). Additionally, it has been suggested that a prevalent cause of DSBs is transcription (Mehta and Haber 2014). In the current research, the GC content of regions in the genome was not characterized. It would be informative to determine if there was a bias favoring the conversion of AT to GC in the dihaploids, as this would suggest that haploid production may influence the GC content of genomes. In a non-agricultural context, such matings may occur in the wild, not only influencing

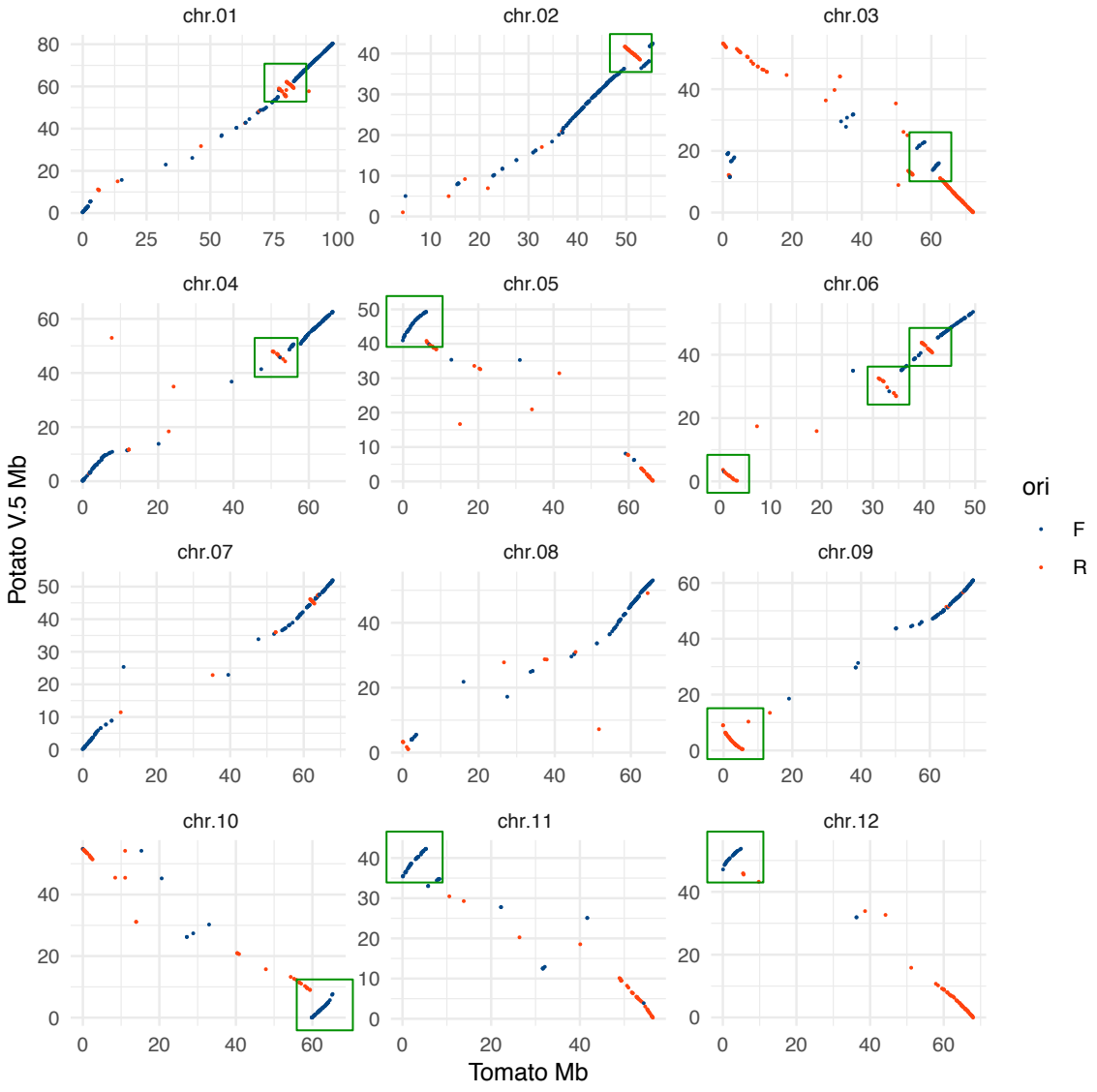

Figure 5.1. Alignment of DM v.5 potato chromosomes to tomato (SL 3.0) using PROmer (Kurtz *et al.* 2004).

Blue points show forward orientation and red points show reverse orientation. The green boxes highlight several large inversions.

nucleotide content in genomes but introducing the occurrence of different ploidy levels in populations.

Improvements to the potato reference genome sequence

The hybrid potato genome assembly DM v.5, though already an improvement on contiguity relative to DM v.4.04, requires additional work to rectify errors in the assembly. The mapping of genetic markers from the DM x RH F1 population showed a large inversion on chromosome 3. This error will be corrected by searching for the breakpoints and reverse complementing the sequence between the breakpoints. To accurately identify the breakpoints, previously collected bacterial artificial chromosome clone-end sequences spanning the region will be utilized.

The initial gene predictions require refinement using transcript alignments. This will be performed using PASA (Haas *et al.* 2003). After the gene models have been revised, functional annotations will be assigned. This will be performed using a combination of methods, including *ab initio* prediction of functions using InterProScan (Jones *et al.* 2014). Furthermore, the functions inferred in DM v.4.04 will be transferred using a reciprocal best-hit approach.

Comparative work to compare synteny between DM v.5, DM v.4.04, and the close relative of potato, *Solanum lycopersicum*, will be explored to characterize large-scale inversions between tomato and potato. Initial analysis using whole-genome alignments show evidence of several large inversions between DM and tomato (Figure 5.1). Using MCScanX (Wang *et al.* 2012), a more detailed comparison between the structure of genomes will be achieved.

CONCLUSION

The genome of cultivated potato has been shaped by selection during its domestication and breeding, a process that has resulted in the retention of heterozygosity and many deleterious alleles. The work presented in these studies offer insights into the way that heterozygosity affects gene expression in potato. In addition, it is demonstrated that the haploid inducer line, IVP101, is an effective line that does not substantially alter the DNA content of dihaploid progeny. Finally, a new potato genome assembly with more accurate and contiguous sequence will serve as a better reference for research on potato genome biology and comparative genome biology.

REFERENCES

REFERENCES

Couturier, J., Touraine, B., Briat, J.F., Gaymard, F. and Rouhier, N. (2013) The iron-sulfur cluster assembly machineries in plants: current knowledge and open questions. *Front Plant Sci*, **4**, 259.

Glemin, S., Clement, Y., David, J. and Ressayre, A. (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet*, **30**, 263-270.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L. and White, O. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, **31**, 5654-5666.

Hu, Z.B., Cools, T. and De Veylder, L. (2016) Mechanisms Used by Plants to Cope with DNA Damage. *Annual Review of Plant Biology, Vol 67*, **67**, 439-462.

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

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

Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., Liti, G., Tian, D., Hurst, L.D. and Yang, S. (2018) Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat Ecol Evol*, **2**, 164-173.

Mehta, A. and Haber, J.E. (2014) Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol*, **6**, a016428.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L. and Marais, G.A. (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol*, **4**, 675-682.

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 : for cell and molecular biology*, **92**, 624-637.

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**, 2391-2402.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C. and Paterson, A.H. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*, **40**, e49.

Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N.R., Shang, Y., Li, C. and Huang, S. (2019) The genetic basis of inbreeding depression in potato. *Nature genetics*.