Bioinformatics Analyses for *De Novo* Regulatory Motif Discovery, Structural Variant Analysis, and Genome Assembly in Potato (*Solanum tuberosum* L.)

José Héctor Gálvez López

Department of Plant Science

Faculty of Agricultural and Environmental Sciences

McGill University

Montreal, Quebec, Canada

February 2017

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

Abstract

Potato (*Solanum tuberosum* L.) is an important staple crop with a highly heterozygous and complex genome. Potato improvement efforts have been held back by the relative lack of genetic resources available to producers and breeders. This work has focused on expanding the available genomic and transcriptomic resources for potato. Specifically, by predicting gene regulatory mechanisms as a response to nitrogen (N) supplementation and through the assembly of two draft genomes for potato landraces *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx.*

The response to N supplementation is important for potato production because insufficient N can have negative impacts on yield and tuber quality while excessive N can be harmful to the environment. In total, thirty genes were found to be consistently over-expressed and nine genes were found to be consistently under-expressed in potatoes from three different cultivars (Shepody, Russet Burbank, and Atlantic) grown in fields with supplemented N. The 1000 nt upstream flanking regions of N responsive genes were analyzed and nine overrepresented motifs were found using three motif discovery algorithms (Seeder, Weeder and MEME). These putative regulatory motifs could be key to understanding the genetic response to N supplementation in commercial potato cultivars.

Genome re-sequencing data from two potato landraces (*S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*) was used to identify structural variation when compared to the potato reference genome. Using copy number variation (CNV) detection software, a significant number of deletions and duplications were identified in both landraces, affecting genes with functions ranging from carbohydrate metabolism to disease resistance. Additionally, draft genomes were assembled *de novo* for each variety, providing evidence for large-scale structural variation between each subspecies. A number of putative novel sequences that are currently not included in the potato reference genome were also discovered in these two potato varieties. While significant work remains to improve the assembled genomes for subsp. *andigena* and *goniocalyx*, this study provides evidence that structural variation in these wild potato species merits further analysis.

Résumé

La pomme de terre (*Solanum tuberosum* L.) est une culture de base importante avec un génome complexe et hétérozygote. Les efforts d'amélioration de la pomme de terre ont été freinés par le manque de ressources génétiques à la disposition des producteurs et des généticiens-sélectionneurs. Ce travail se concentre sur l'expansion des ressources génomiques et transcriptomiques disponibles pour la pomme de terre. Plus précisément, on y prédit des mécanismes de régulation des gènes en réponse à la supplémentation d'azote (N) et on y fait l'assemblage de deux ébauches de génomes pour les variétés sauvages de pomme de terre *S. tuberosum* sous-espèce *andigena* et *S. stenotomum* sous-espèce *goniocalyx*.

La réponse à la supplémentation en N est importante pour la production de pommes de terre parce que l'insuffisance en N peut avoir des impacts négatifs sur le rendement et la qualité des tubercules, tandis qu'un apport excessif en N peut être nocif pour l'environnement. Au total, trente gènes ont été constamment surexprimés et neuf gènes ont été constamment sous-exprimés dans les pommes de terre de trois cultivars différents (Shepody, Russet Burbank et Atlantique) cultivés dans les champs supplémentés en N. On a analysé les 1000 paires de bases qui flanquent les gènes sensibles à l'N et on a trouvé neuf motifs surreprésentés en utilisant trois algorithmes différents pour la découverte de motifs (Seeder, Weeder et MEME). Ces motifs de régulation putatifs pourraient être la clé pour comprendre la réponse génétique à la supplémentation en N dans les cultivars commerciaux de pomme de terre.

Des données de re-séquençage génomique de deux variétés sauvages de pommes de terre (les sous-espèces *andigena* et *goniocalyx*) ont été utilisées pour identifier leur variation structurelle par rapport au génome de référence de la pomme de terre. Un nombre significatif de délétions et de duplications ont été identifiées dans ces deux variétés avec un logiciel de détection de variation du nombre de copies (VNC). Les gènes affectés par la variation ont fonctions diverses, par exemple le métabolisme des glucides et la résistance aux maladies. De plus, une ébauche de génome a été assemblée *de novo* pour chaque variété, fournissant des preuves additionnelles de variation structurelle à grande échelle entre chaque sous-espèce. Un certain nombre de nouvelles séquences qui ne sont pas actuellement incluses dans le génome de référence de la pomme de terre ont aussi été découvertes dans ces deux variétés. Cette étude fournit des preuves que ces variétés sauvages méritent une analyse plus poussée, même s'il reste beaucoup de travaux importants à faire.

Acknowledgments

First of all, I would like to acknowledge my supervisor Prof. Martina Strömvik for her patience, guidance and advice throughout my Master's degree. I am thankful for the opportunity you gave me to join your laboratory, I have learned so much from you these last two and a half years. To my supervising committee, Dr. Helen Tai and Prof. Jean-Benoit Charron, my most sincere thank you for your help and the knowledge you contributed to this project.

My acknowledgments to the rest of our collaborators: Dr. Bernie Zebarth, Dr. David Ellis, Dr. Noelle Barkley and Dr. Kyle Gardner as well for their valuable contributions to this work. Additional thanks to the McGill University faculty members and administrative staff that have been there for me throughout the years, I am sure I have learned something of value from every single one of you and I thank you.

To all the members, past and present, of the Strömvik laboratory, thank you for your companionship, support and advice throughout the times we spent together. Know that this project would not have been possible without you. To the rest of my friends at the Department of Plant Science and throughout McGill University and Montreal, thank you for making this experience enriching beyond academics and for introducing me to the many wonders of this amazing city.

Last but not least, thank you to my Mexican friends, relatives, and, most of all, to my family: Mom, Dad, Ana Gaby, Michelle, Milton and Candy. Your support was felt every minute of every day. This is the culmination of yet another step and, as always, it could not have been possible without the education and opportunities you have worked so hard to give me. Thank you.

Table of contents

List of Tables

List of Figures

List of Abbreviations

- **AFLP:** Amplified Fragment Length Polymorphism
- **AMT:** Ammonium Transporters
- **ANOVA:** Analysis of Variance
- **BAC:** Bacterial Artificial Chromosome
- **BAM:** Binary Alignment/Map format
- **BLAST:** Basic Local Alignment Search Tool
- **BSA:** Bulk Segregant Analysis
- **C:** Carbon
- **ChIP-seq:** Chromatin Immuno-Precipitation sequencing
- **CIP:** International Potato Center (*Centro Internacional de la Papa*)
- **CNV:** Copy Number Variation
- **CNS:** Conserved Noncoding Sequence
- **DM:** *Solanum tuberosum* group Phureja DM1-3 516 R44 (Double Monoploid)
- **DNA:** Deoxyribonucleic Acid
- **E.C.:** Enzyme Commission
- **EST:** Expressed Sequence Tag
- **FAO:** Food and Agriculture Organization of the United Nations
- **FISH:** Fluorescence *in situ* Hybridization
- **FPKM:** Fragments Per Kilobase per Million mapped reads
- **GABA:** γ-Aminobutyric Acid
- **GEO:** Gene Expression Omnibus (NCBI)
- **GFF:** General Feature Format (also GFF3)
- **GO:** Gene Ontology
- **GTF:** General Transfer Format
- **GUS:** β-Glucuronidase
- **GWAS:** Genome Wide Association Study
- **HPC:** High-Performance Computing
- **ITAG:** International Tomato Annotation Group
- **KEGG:** Kyoto Encyclopedia of Genes and Genomes
- **LTR-RT:** Long Terminal Repeat Retro-Transposons
- **MAS:** Marker Assisted Selection
- **MAGIC:** Multiparent Advanced Generation Intercross
- **Mb:** Mega Base Pair
- **N:** Nitrogen
- **NAM:** Nested Association Mapping
- **NaCl:** Sodium Chloride (common salt)
- **NCBI:** National Center for Biotechnology Information
- **NCD:** North Carolina design
- **NiR:** Nitrite Reductase
- **NLP:** Ninein-Like Protein
- **NR:** Nitrate Reductase
- **NRE:** Nitrate Responsive *cis*-Element
- **N50:** A value that represents the minimum length of contigs/scaffolds covering 50% of the genome
- **PacBio:** *Pacific Biosciences of California, Inc.*; developer of long-read sequencing technology
- **PCR:** Polymerase Chain Reaction
- **PGSC:** The Potato Genome Sequencing Consortium
- **PLACE:** Plant Cis-acting Regulatory DNA Elements Database
- **PM:** Pseudomolecule
- **PVY:** Potato virus Y
- **QTL:** Quantitative Trait Loci
- **QUAST:** Quality Assessment Tool for genome assemblies
- **RAD-seq**: Restriction site Associated DNA sequencing
- **RNA:** Ribonucleic Acid
- **RNA-seq:** RNA Sequencing
- **SAM:** Sequence Alignment/Map format
- **SOAP:** Short Oligonucleotide Alignment Program
- **SPAD:** Special Products Analysis Division
- **SSR:** Simple Sequence Repeats
- **TAMO:** Tools for Analysis of Motifs
- **TE:** Transposable Element
- **TSS:** Transcription Start Site
- **WGS:** Whole-Genome Shotgun
- **Y1H:** Yeast-One Hybrid
- **bp:** Base Pair
- **cDNA:** Complementary DNA
- ***cox1*:** Potato mitochondrial gene
- **h:** hour
- **ha:** Hectare (10,000 square meters)
- **kg:** Kilogram
- **nt:** Nucleotide

- **qPCR:** Quantitative PCR
- **sRNA:** small RNA
- **subsp.:** subspecies
- **µl:** microLiter

Contribution of Authors

José Héctor Gálvez designed this project and wrote the main text for this thesis under the supervision of Martina V. Strömvik.

Additional contributors to this work: Helen H. Tai (co-supervisor), David Ellis and Noelle Barkley (for Chapters 2 and 4), Kyle Gardner (for Chapter 2), Martin Lagüe, Bernie Zebarth and Rui Li (for Chapter 3), as well as Chen Yu Tang and Xinyi Zhu (for Chapter 4). The individual contributions of every author are specified in detail before each chapter.

Chapter 1: Introduction

Potato (*Solanum tuberosum* L.) is widely recognized as the most important non-grain staple crop worldwide. The latest FAO statistics indicate that over 380 million tonnes of potatoes were produced in 2014 alone (Food and Agriculture Organization 2016), illustrating its international economic and agricultural importance. Potato is a member of the Solanaceae family, which includes other significant agricultural species such as tomato, pepper, and tobacco. The cultivated forms of potato are vegetatively propagated and are predominantly autotetraploids (2n = 4x = 48). However, ploidy ranges from diploid to hexaploid in cultivated potato (Hawkes 1990), (for a review on potato genetic diversity, see Machida-Hirano 2015). Potatoes were domesticated in the Andes approximately 10,000 years ago and the landraces have a wide variety of shapes, skin and tuber colors, often not seen in modern varieties (Ovchinnikova *et al.* 2011). It is fairly common in the Andes that landraces of all ploidy levels are grown in the same field and are also grown near wild relatives facilitating cross hybridization and gene flow (Huamán & Spooner 2002).

Potatoes are valued for their nutritious properties and their wide eco-geographical range. However, due to their high heterozygosity, complex polysomic inheritances, and narrow genetic base, they are difficult to improve through classical breeding methods. Because they are typically vegetatively propagated, many modern cultivars are only separated by a few meiotic generations (Gebhardt *et al.* 2004; Simko *et al.* 2006) making the genetic diversity among cultivars really low. They are quite susceptible to many pests and also suffer from acute inbreeding depression.

The scientific and economic importance of potato is not new. While other crops such as maize and wheat have seen great increases in yield as a consequence of genetic improvement in the last few decades, this has not been the case with potato. Instead, evidence suggests that yield increases are mostly due to improved agricultural practices. The majority of cultivated potato still comes from a narrow group of cultivars, including Russet Burbank, which was originally released in 1874 (Douches *et al.* 1996; Iovene *et al.* 2013). While many more recent cultivars have been released since the late 1800s, these have been bred mostly based on phenotypic selection, not genetic information, and they have been developed with a very particular use in mind, such as processing for the potato chip or the French fry industries (Hirsch *et al.* 2013). Worldwide demand for potato is increasing; therefore, scientists have begun to study potato genetics with the hope that

it can provide breeders with more tools to aid crop improvement in terms of yield and disease resistance.

Until recently, the genomic understanding of this crop was held back by its relatively complex genome. The challenges associated with potato improvement have prompted a number of significant genomic and transcriptomic studies in this species and its close relatives, which will provide tools for breeders and additionally shed light into mechanisms behind important molecular processes. In 2011, the first potato reference genome and transcriptome were published (Massa *et al.* 2011; The Potato Genome Sequencing Consortium 2011), and two years later, an update was released substantially improving the scaffolds and pseudomolecules of the initial reference (Sharma *et al.* 2013). Recently, the first draft genome of a wild potato species, *Solanum commersonii*, was also released (Aversano *et al.* 2015), in addition to many other genome sequencing efforts in related species, such as tomato (*Solanum lycopersicum*; The Tomato Genome Consortium 2012), chili pepper (*Capsicum annuum*; Kim *et al.* 2014), tobacco (*Nicotiana tabacum*; Sierro *et al.* 2014) and the parental genomes of petunia (*Petunia axillaris* and *Petunia integrifolia*; Bombarely *et al.* 2016) which collectively have also provided valuable information on potato.

The potato reference genome is also a starting point for the exploration of biodiversity between potato cultivars and subspecies. Using genome re-sequencing, it is now possible to assemble separate genomes as a reference for specific varieties. These new assemblies can provide useful information about the structural differences between different potato subspecies (*Solanum tuberosum* subsp. *andigena, S. stenotomum* subsp. *goniocalyx*, *S. stenotomum* subsp. *stenotomum* and 36 *S. tuberosum* subsp. tuberosum; species definition from Hawkes 1990) from Single Nucleotide Polymorphisms (SNPs) and Copy Number Variation (CNV) to large-scale structural variation. Indeed, recent research is already pointing to significant differences in gene copy number between different potato populations (Hardigan *et al.* 2016).

The main focus of this work is to continue to build upon the current foundation of potato genomics and transcriptomics studies by exploring the potential regulatory mechanisms behind the long-term response to N supplementation in field-grown potatoes, as well as the genomic differences between the potato reference genome and two potato landraces.

1.1 Hypotheses

1. Three potato cultivars (Shepody, Russet Burbank, and Atlantic) share a group of common genes that are responsive to differences in N supplementation.
2. Three potato cultivars (Shepody, Russet Burbank, and Atlantic) share overrepresented motifs in the upstream flanking regions of N responsive genes.
3. The genome of the potato landrace *S. tuberosum* subsp. *andigena* has significant CNVs, novel sequences and structural variants when compared to the potato reference genome.
4. The genome of the potato landrace *S. stenotomum* subsp. *goniocalyx* has significant CNVs, novel sequences and structural variants when compared to the potato reference genome.

1.2 Objectives

1. Analyze RNA-seq data obtained from three potato cultivars (Shepody, Russet Burbank, and Atlantic) treated with different amounts of N supplementation to detect common N responsive genes.
2. Analyze the available gene annotation data to find overrepresented metabolic pathways and Gene Ontology (GO) terms associated with N responsive genes in Shepody, Russet Burbank and Atlantic.
3. Analyze the upstream regions of N responsive genes in three potato cultivars (Shepody, Russet Burbank, and Atlantic) with different bioinformatics algorithms (Seeder, MEME and Weeder) to detect common overrepresented motifs.
4. Make adjustments to the Seeder program to improve its use within a High Performance Computing (HPC) environment.
5. Develop a strategy to deal with redundancy in motif finding results, particularly to identify instances where the same motif is reported more than once by the motif discovery software.
6. Using genome re-sequencing data, assemble new reference genomes for two potato landraces (*S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*).
7. Compare the new assemblies to each other and to the reference genome to identify potential structural and genetic differences such as CNVs and novel sequences.

Preface and Author Contributions to Chapter 2

A version of this chapter was published in *AIMS Agriculture and Food* in January 6, 2017 (Gálvez *et al.* 2017). José Héctor Gálvez wrote the main text, with contributions from Martina V. Strömvik, Helen H. Tai, Noelle Barkley, Kyle Gardner and David Ellis. All authors reviewed the manuscript.

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 Martina V. Strömvik, Noelle Barkley, David Ellis, and Helen H. Tai; the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. 283303) to Martina V. Strömvik; A-base funding from Agriculture and Agri-Food Canada to Helen H. Tai and Kyle Gardner; the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología; CONACYT) (Scholarship No. 381158) to José Héctor Gálvez; and the McGill Department of Plant Science Graduate Excellence Fund.

Chapter 2: Literature Review

2.1 Overview

Despite its cultural and economic significance, potato improvement efforts have been held back by the relative lack of genetic resources available to producers and breeders. The publication of the potato reference genome and advances in high-throughput sequencing technologies have led to the development of a wide range of genomic and transcriptomic resources. An overview these new tools, from the updated versions of the potato reference genome and transcriptome, to more recent gene expression, regulatory motif, re-sequencing and SNP genotyping analyses, paints a picture of modern potato research and how it will change our understanding of potato as well as other tuber producing Solanaceae.

2.2 The Potato Reference Genome

Because of the high degree of heterozygosity normally found in *S. tuberosum* a homozygous clone of the plant needed to be created in order to produce a high-quality draft genome. This was achieved through the duplication of a monoploid (1n = 1x = 12) specimen that had been previously derived from a heterozygous clone of the *Phureja* group of cultivated potato. This doubled monoploid, named DM1-3 516 R44 (hereafter referred to as DM) was the only source of sequencing data for the potato draft genome. The genome scaffolds assembled from DM were then used to integrate data from a heterozygous diploid breeding line that was a cross between a *S. tuberosum* "dihaploid" (SUH2293) and a diploid clone (BC1034) generated from two *S. tuberosum* group *Phureja* hybrids. Both DM and the heterozygous breeding line were chosen as the sources of data for the reference genome because collectively they represent a sample of the wider potato diversity with DM, a member of the *Phureja* group being closer to wild potato relatives, while the breeding line having more in common with cultivated potato varieties such as Russet Burbank (The Potato Genome Sequencing Consortium 2011).

The first potato reference genome was completed in 2011 by the Potato Genome Sequencing Consortium (PGSC) using a whole-genome shotgun (WGS) approach (The Potato Genome Sequencing Consortium 2011). Previous data had already determined that the potato genome is composed of 12 chromosomes and has a total size of approximately 840 Mb. Contigs were assembled *de novo* using the SOAPdenovo (Luo *et al.* 2012) assembly program. The

assembly consisted of 727 Mb, 93.3% of which were non-gapped sequences. Analysis of the DM assembly revealed that 62.2% of the assembled genome consisted of repetitive content. To assess the quality of the draft genome, Sanger-derived phase-2 BAC sequences, which amounted to approximately 1 Mb, were aligned to the assembly. No gross assembly errors were detected in the aligned data (The Potato Genome Sequencing Consortium 2011). A reference transcriptome was produced to annotate the genome and it contained around 21,000 high-confidence transcripts (Massa *et al.* 2011).

Two years after the publication of the first reference genome, a new assembly of the DM clone was released with a more accurate arrangement of scaffolds and pseudomolecules (Sharma *et al.* 2013). This updated assembly of the potato reference genome (version 4.03) was created by integrating linkage data from a segregating diploid potato population derived from the reference sequence clone (DM). This new dataset was used to revise and improve the genome pseudomolecules (PMs) of the original assembly (Sharma *et al.* 2013). The new build contains 951 genome superscaffolds of which 90% (655 Mb) have been assigned to an absolute or relative orientation within the PMs. Also, a small number of superscaffolds (about 3%) have been assigned to a random orientation. The exact chromosome position and absolute orientation of the remaining 279 Mb of superscaffold sequences found in the heterochromatin could not be determined. This means that a total of 93% of the assembled genome, comprising a total of 674 Mb, are contained in the chromosome scale PMs of the 4.03 version of the assembly. A total of around 96% of the predicted genes in potato are found in these PMs (Sharma *et al.* 2013).

A more recent update of the potato reference genome (version 4.04) was released in 2016 (Hardigan *et al.* 2016). It was built with additional genomic data obtained from foliar and stem tissue of a potato cloned from the original DM reference. It adds 55.7 Mb of novel sequences in the form of >200 bp contigs, including several new genes, that did not map to the v4.03 reference. These contigs were concatenated into an unanchored pseudomolecule called “chrUn”, which was then annotated using a standardized pipeline (Hardigan *et al.* 2016). However, since this new assembly does not anchor the new data into any chromosome, or incorporate new linkage data in any way, it is only useful as further reference for potato sequences that do not align to any of the established pseudomolecules found in v4.03. A summary of the available reference genomes for potato and its close relatives can be found in **Table 2.1**.

Table 2.1: Summary of the different versions of the potato reference genome and the reference genome of its close relatives *Solanum commersonii,* and tomato *(S. lycopersicum)*.

	***S. tuberosum* reference genome**			***S. commersonii* reference genome**	***S. lycopersicum* reference genome**
	v3	v4.03	v4.04		
Source of Genetic Material	*S. tuberosum* group Phureja DM1-3 516 R44.	Same as v3, with additional linkage data from DMDD† mapping population.	Same as v3 with additional DNA from DM1-3 516 R44 stem and leaf tissue.	*S. commersonii* accession PI 243503.	'Heinz 1706' inbred line (Heinz Corporation, Pittsburgh, PA).
Total Length [Mb]	727	723	779	730	760
Scaffold N_{50} [kb]*	1,340	4,100	4,100	44	16,470
GC content	34.80%	34.80%	N/A	34.50%	35.71%
Predicted Number of Genes	39,031	39,031	N/A	37,662	34,727
Comments	Most recent version available in the NCBI Genome database.	No new novel sequences compared with v3.	Only difference with v4.03 is the addition of 55.7 Mb of novel genes and sequences.	Also available in the NCBI Genome database.	Most recent version available in the NCBI Genome database.
Reference	(The Potato Genome Sequencing Consortium 2011)	(Sharma *et al.* 2013)	(Hardigan *et al.* 2016)	(Aversano *et al.* 2015)	(The Tomato Genome Consortium 2012)

* Minimum size in which 50% of the assembly can be found

† Mapping population of 180 backcross progeny clones derived from an initial cross (DM × D where DM=DM1-3 516 R44 and D=CIP703825) (Sharma *et al*. 2013)

2.3 The Potato Reference Transcriptome and Gene Expression Studies

For many years, the main transcriptomic resources available to potato breeders were public EST libraries containing a total of more than 200,000 tags (Crookshanks *et al.* 2001; Ronning *et al.* 2003; Flinn *et al.* 2005; Rensink, Hart, *et al.* 2005). Additionally, EST libraries from other closely related Solanaceae species (such as tomato, eggplant, pepper, tobacco and petunia) also proved to be relevant for potato because many of the genes were shared across the species and genera in this family (Rensink, Lee, *et al.* 2005). The EST sequence data was used to develop microarrays for analysis of gene expression including cDNA arrays (Rensink, Iobst, *et al.* 2005; Kloosterman *et al.* 2008) and a 44k oligo array using the Agilent platform (Potato Oligo Chip Initiative: POCI; Kloosterman *et al.* 2008). Recently, another Agilent oligo array (JHI *Solanum tuberosum* 60k array), was developed based on the predicted transcripts of potato reference genome v3.4 (see **Table 2.2**; Bengtsson *et al.* 2014). Collectively, these resources were behind significant discoveries in the gene expression profiles of potato under different conditions such as flowering and tuber development (Bachem *et al.* 2000; Campbell *et al.* 2008; Navarro *et al.* 2011), biotic (Restrepo *et al.* 2005; Tai *et al.* 2013; Bengtsson *et al.* 2014) and abiotic stress (Schafleitner *et al.* 2007; Ginzberg *et al.* 2009; Evers *et al.* 2010; Hammond *et al.* 2011; Hancock *et al.* 2014).

After the publication of the potato reference genome, it became possible to design potato transcriptomics studies using a molecular technique known as RNA-seq. Briefly, RNA-seq consists of converting any given RNA sample into a cDNA library using reverse-transcription, amplification and DNA fragmentation. Then, the library is sequenced using a massively parallel sequencer producing a number of short reads which can be used to determine the sequences of the original RNA molecules (Wang *et al.* 2009). One of the most prevalent applications of RNA-seq has been to estimate and compare gene expression between full transcriptome samples obtained from different organisms, individuals or tissues. However, the success of this technique depends on many factors, including a good reference transcriptome or genome to which reads can be mapped (Li, Ruotti, *et al.* 2009; Pachter 2011; Trapnell *et al.* 2012; Conesa *et al.* 2016). Which is why, the development of a full reference transcriptome was crucial advance gene expression studies in potato.

In order to assemble the gene models that make up the potato reference transcriptome, the PGSC collected data from 32 different tissues of the same *Phureja* DM clone used for the

sequencing of the reference genome. The tissues were selected to represent all the major plant organs, including flower, fruit, leaf, tuber and roots at different developmental stages and stress conditions (Massa *et al.* 2011). Over 550 million reads were obtained from all the tissue samples. Petal tissue yielded the lowest amount of reads with only 5.4 million, while the mature whole fruit library had the greatest number of reads, around 30 million. In terms of high- confidence transcripts, one sample of tuber tissue had the lowest amount (11,394), while the highest number (16,276) was found in plants treated with salt (NaCl). Since libraries with the lowest and highest number of reads produced roughly the same number of high-confidence transcripts, it seems there is no significant bias against transcript detection depending on sequencing depth (Massa *et al.* 2011).

A transcript was considered as expressed if its abundance, as calculated using the Cufflinks software package (Trapnell *et al.* 2010), had a fragment per kilobase of exon per million fragments mapped (FPKM) value ≥0.001 and the lower bound of the 95% confidence interval was above zero. Using these criteria, a total of 22,704 unique high-confidence transcripts were identified in all 32 libraries. The *S. tuberosum* reference genome contains a total of 39,031 protein-coding genes. If a single transcript is chosen to represent each gene also found in the genome, around 60% of the genes found in the genome are also included in the reference transcriptome. Out of all these transcripts, only 17% have no known function and a total of 1,680 (around 8%) were only found in tissues under some type of biotic or abiotic stress (Massa *et al.* 2011).

With the goal of facilitating comparative analyses between potato and tomato, the international Tomato Annotation Group (iTAG) used their annotation pipeline to re-annotate the potato reference genome. This newer potato gene annotation, referred to as iTAG, contains less total genes than the original annotation published by the PGSC (35,004 and 39,031 genes, respectively) (The Tomato Genome Consortium 2012). However, when compared to an external standard (TAIR10) (Swarbreck *et al.* 2008), 92% of the genes models in the iTAG annotation had a corresponding match, whereas more than 30% of the PGSC genes had no match at all (The Tomato Genome Consortium 2012). The iTAG annotation is therefore another valuable resource for potato research, especially in studies that involve comparisons with tomato or other members of the Solanaceae family.

Gene expression studies have proven to be a useful tool for investigating plant molecular response to different environmental stimuli (Hazen *et al.* 2003). There has been a recent increase in the application of RNA-seq to understand potato biology and the genes underlying complex traits. *Phytophthora infestans* defense response, tuberization under the control of photoperiod, drought response, tuber pigmentation, PVY resistance, response to nitrogen fertilizer and an activation-tagged mutant with altered growth habit have all been examined using RNA-seq to quantify gene expression (Gao *et al.* 2013; Shan *et al.* 2013; Zhang *et al.* 2014; Frades *et al.* 2015; Goyer *et al.* 2015; Liu *et al.* 2015; Cho *et al.* 2016; Gálvez *et al.* 2016). RNA-seq has also been used to identify genes that are predictive of cold-induced sweetening in tubers (Neilson *et al.* 2016, in preparation). It has also been used to quantify gene expression in a wild potato species, *S. commersonii*, which is resistant to bacterial wilt; this was achieved using the potato reference genome for sequence alignment (Zuluaga *et al.* 2015). Finally, the National Centre for Biotechnology (NCBI) Gene Expression Omnibus (GEO) (Wheeler *et al.* 2007) currently lists more than 1,500 *S. tuberosum* samples across 102 series of experiments performed using high throughput sequencing, this includes studies on miRNA and other non-coding RNA, in addition to gene expression.

As previously mentioned, expression profiling using RNA-seq is dependent on many factors, including the accurate alignment of short sequence reads to the reference genome or transcriptome. However, it has been noted that some genes are recalcitrant to RNA-seq analysis (Hirsch *et al.* 2015). The underlying reasoning behind gene expression studies using RNA-seq is that the population of a particular species of mRNA can be accurately used to estimate the expression of the gene from which the mRNA was transcribed. In reality, gene expression depends not just on mRNA abundance but also on the ability of an mRNA molecule to be available for translation in the ribosomes. This could potentially be estimated with additional data such as the number of cells in a sample, the volume of each cell and the physical location of transcripts. However, this data is often unavailable, which can increase the uncertainty of RNA-seq studies especially in situations where mRNA abundance correlates poorly with gene expression (Wagner *et al.* 2012). This may be due to small transcripts that are excluded during the construction of cDNA libraries and/or sequence overlap with other transcripts. Improvements to RNA-seq methodology and bioinformatic data processing are needed to increase confidence in RNA-seq results (Rajkumar *et al.* 2015). Nevertheless, RNA-seq studies have already produced an

abundance of gene expression data and contributed to the understanding of several traits in potato. As RNA-seq is improved longer reads and better software, newer datasets can expand upon this knowledge and enable the discovery of additional information, including mechanisms of gene regulation.

Biological interpretation of RNA-seq and other gene expression analyses require functional annotations of genes. Gene Ontology (GO) is frequently used to look for biological processes, molecular functions, and cellular compartments that are enriched in the dataset (Ashburner *et al.* 2000). Recent efforts have substantially improved the functional annotation of the potato genome using a structure-based pipeline that integrates the results of several different functional annotation software (Amar *et al.* 2014). Despite this improvement, manual curation is still required to further refine functional annotations, especially in plants since they have several unique pathways and processes that are not found in other animals or microorganisms. GoMapMan, a recently developed resource for manual curation, consolidation and visualization of functional annotation in plants, has already been used for several crops including potato (Ramšak *et al.* 2014).

2.4 Regulatory Motif Discovery in Potato

The availability of a high-quality potato reference genome and transcriptome have, in turn, enabled the development of techniques that allow an accurate quantification of gene transcripts that will aid in the understanding of the complexities potato genetics. This includes the analysis of the *cis*-regulatory elements that are flanking genes, which are important because many polymorphisms associated with crop domestication are found in these regions (Swinnen *et al.* 2016). Studies performed in *Arabidopsis thaliana* and maize have shown that flanking regions can contain potential binding sites for elements regulating important phenotypic characteristics such as nitrogen (N) response and assimilation (Konishi & Yanagisawa 2011; Liseron-Monfils *et al.* 2013). Therefore, a greater understanding of gene regulatory mechanisms in potato will provide important information for breeding and genetic modification.

The identification and characterization of regulatory elements has remained a challenge. Techniques such as ChIP-sequencing can reveal the binding sites of regulatory elements, such as transcription factors, by taking advantage of chromatin immuno-precipitation (ChIP) along with DNA sequencing. It has long been known that transcription factors bind to the DNA molecule at specific sites and this interaction is fundamental for the regulation of transcription. In addition,

regulation at the translational level also involves sequence motifs and proteins binding to them (RNA binding proteins). The sequences these molecules bind to are usually short (6-15 bp) and conserved both among genes and species, and they are referred to as DNA or RNA motifs (Pavesi *et al.* 2007). Gene regulatory regions also contribute to phenotypic variation. Studies in *Arabidopsis thaliana* show higher densities of SNPs in environmental response and signaling genes compared to housekeeping genes (Korkuc *et al.* 2014). Regulatory motif discovery will be important in understanding the impact of genetic variation in regulatory regions.

Throughout the years there have been numerous experimental studies linking specific DNA and RNA motifs with certain regulatory mechanisms, as well as specific phenotypes in plants and other organisms. Collectively, these studies offer great value because they can be used to annotate sequences and they can be mined for potential genetic modification targets. There are several curated databases containing experimentally validated DNA and RNA regulatory motifs of which two of the largest are JASPAR (Sandelin *et al.* 2004; Mathelier *et al.* 2014) and PLACE (Higo *et al.* 1999), the latter is specifically focused on motifs found in plants.

There are a limited number of studies on potato gene regulation. Recent approaches have leveraged increasing amounts of sequencing data to characterize not just a promoter region, but also individual motifs and their binding regulatory elements to provide a better understanding of how gene regulation is carried out at a molecular scale. An example of the regulatory importance of the 5' flanking region of genes can be found in the promoters for the Class I patatin family of genes, which encodes several isoforms of patatin and is the most abundant family of proteins found in the tuber of potato. Putative *cis*-regulatory motifs were identified in the 5'-flanking regions, using alignments of previously reported sequence data and searches in the PLACE database (Aminedi & Das 2014). Several conserved occurrences of previously validated motifs were identified and they had associations with plant functions such as light and sucrose responsive transcriptional regulation, transcription enhancers, and response to abiotic stress. Additionally, by artificially adding the upstream flanking region of these genes to other transgenic genes, it is possible to replicate similar sucrose-induced transcription in other tissues of the plant that are not the tuber (Aminedi & Das 2014). The promoters for the pathogenesis-related *PR-10a*, *chitinase C*, stolon-specific *Stgan, snakin-1, granule-bound starch synthase (GBSS1),* and *chalcone isomerase* have also been characterized in a similar fashion (Despres *et al.* 1995; Ancillo *et al.* 2003; Trindade *et al.* 2003; Almasia *et al.* 2010; Bansal *et al.* 2012; Chen *et al.* 2015).

In crops outside the Solanaceae family, there have been studies specifically linking N metabolism with certain regulatory motifs. A good example is the nitrate-responsive *cis*-element (NRE) that was identified in the *Arabidopsis* NiR1 gene by aligning the upstream flanking region of the gene the same region in other plants. The NRE consists of a highly-conserved 43 bp sequence and there is evidence this regulatory element is sufficient and necessary for nitrate responsive transcription (Konishi & Yanagisawa 2010). The NRE can be found in the upstream flanking regions of NiR genes in several crops (e.g. maize, wheat, bean, tobacco). This seems to confirm that the mechanism for transcriptional regulation in response to nitrate may be conserved in many higher plant species (Konishi & Yanagisawa 2011). A Yeast 1 Hybrid (Y1H) screening revealed a Ninein-Like Protein (NLP) that binds to the NRE and activates the nitrate-responsive transcription, indicating that NLPs have a regulatory function in nitrate response (Konishi & Yanagisawa 2013).

The identification and experimental validation of NRE, as well as the characterization of its interaction with NLPs are a good example of the potential knowledge that can be gained from research focused on the discovery of regulatory mechanisms in plants. Which is why several new approaches have been developed to enable the discovery of additional regulatory motifs in plants. One successful method that is available for families with extensive genomic information across several species, is to detect conserved noncoding sequences (CNSs). Because non-functional sequences are expected to diverge faster than sequences under selective constraint, it is likely that CNSs contain important functional elements (Haudry *et al.* 2013). This approach has already been applied to crucifers where 90,000 CNSs were identified, several of them containing overrepresented motifs and displaying evidence of regulatory function (Haudry *et al.* 2013). However, the exclusive use of sequence alignment to identify conserved targets limits the discovery of regulatory motifs to well-annotated datasets spanning many different plant species. Which is why other approaches have been developed for species without a wide range of genomic datasets available. One approach is *de novo* motif discovery, which has also been favored due to lower costs compared to other *in vitro* and *in vivo* techniques (López *et al.* 2013; Gálvez *et al.* 2016).

Regulatory motifs have been found to be overrepresented in the genome in two ways: they are often found in conjunction with the genes they regulate, and motifs acting on similar genes tend tend cluster locally (Harbison *et al.* 2004). Modern algorithms designed for the purpose of

de novo motif discovery have leveraged this information using different approaches, each with their own set of advantages and disadvantages. Three software packages that have been used successfully in plants are: Weeder (Pavesi *et al.* 2007), MEME (Bailey *et al.* 2009) and Seeder (Fauteux *et al.* 2008). To increase the probability of discovering and predicting regulatory elements, it is common to analyze a single dataset using different software, which compensates for the strengths and weaknesses of each algorithm (Gordon *et al.* 2005; López *et al.* 2013; Zolotarov & Strömvik 2015). The aggregated results obtained from these tools can then be used to search curated motif databases. If regulatory motifs with no previous experimental validation are found, targeted studies can be designed to determine the biological function of these motifs either *in vitro* or *in vivo*.

A recent study conducted using a *de novo* motif discovery approach was able to identify nine putative *cis*-regulatory motifs in the upstream flanking region of nitrogen responsive genes in three potato cultivars (Gálvez *et al.* 2016). The nine motifs had close matches to experimentally validated regulatory motif entries in both PLACE and JASPAR. These sequences could be targeted in experimental studies analyzing steady-state nitrogen response and regulation in field-grown potato, which is a pressing concern for potato producers because of the dependence of the crop on nitrogen supplementation. However, future research on motifs and regulatory elements must also take into account the diversity of potato cultivars and varieties, which requires a deeper knowledge of the genetic differences between them.

2.5 Genome Re-sequencing and Genetic Diversity

The assembly of the potato reference genome and transcriptome was possible thanks to the development of the double monoploid derived from the *Phureja* group (The Potato Genome Sequencing Consortium 2011). However, most cultivated potatoes are polyploid and highly heterozygous and until recently (Spooner *et al.* 2014) were originally classified into seven species and nine taxa (Hawkes 1990) which could mean that the genomes of potato landraces and native cultivars might differ significantly from the potato reference genome. The complexity of the potato genome has made the genetic differences between these populations difficult to discern. For example, the taxonomy of the group *Solanum* sect. *Petota* (wild potatoes), as well as the appropriate classification of different potato varieties have been a point of debate among

specialists for several years (Huamán & Spooner 2002; Spooner *et al.* 2007, 2014; Spooner 2009; Ovchinnikova *et al.* 2011; Machida-Hirano 2015).

Although different approaches have been employed to classify potato germplasm (morphological, molecular, cytometric), taxonomy remains challenging due to varying ploidy levels, sexual and asexual reproduction, the ease of interspecific hybridization, and introgressions from various wild species. One example of a characteristic that caused some confusion in taxonomic classification in the past is that potato germplasm was frequently classified as a particular species based on ploidy level or ploidy level was assumed based on classification in a particular species. However, molecular studies have demonstrated that ploidy is not a good indicator of taxonomic classification because potato species have been found with mixed ploidy levels within the species (Ghislain *et al.* 2006; Barkley et al., in preparation). Current research programs on genetic resources are working on sorting potato taxonomy and making modifications as needed.

Since the release of the potato reference genome, significant amounts of data have been collected using high-throughput sequencing and SNP arrays. These new datasets have mostly supported, with some exceptions, the current taxonomic tree of tuber-bearing *Solanaceae* and provided a general overview of the genetic diversity of these species (Hirsch *et al.* 2013; Hardigan *et al.* 2015); however, these tools are just starting to be used to discover specific differences in the genome of potato varieties. For example, SNP arrays have been shown to reveal complex relationships, such as, inter- and intraspecific diversity of the wild species (Hardigan *et al.* 2015). Evaluation of wild genotypes across loci can also potentially reveal valuable information on genes that differentiate primitive and cultivated germplasm, as well as, determine key loci involved in domestication or enhanced agronomic performance of modern varieties (Hardigan *et al.* 2015). Identification of novel alleles and their potential utilization is a key factor to assist breeding programs in developing improved varieties in order to advance this important crop.

Important structural variations between different varieties of potato have also been uncovered. A study using a Fluorescence *in situ* Hybridization (FISH) based approach concluded that CNVs were highly abundant in potatoes. However, the limitations of that technology made it

impossible to accurately determine the distribution and prevalence of CNVs throughout the genome (Iovene *et al.* 2013).

High-throughput sequencing data was recently used to identify CNVs within a panel of 12 potato monoploids containing diverse genetic backgrounds (Hardigan *et al.* 2016). Using CNVnator (Abyzov *et al.* 2011), a program developed to detect CNVs by comparing sequencing read depth to a reference genome, the prevalence of CNVs in potato varieties was confirmed. Results show that CNVs cover approximately 30% of the genome and more than 11,500 individual genes, making them one of the major components of genetic diversity. Genes found exclusively in potato, including disease resistance genes, as well as genes previously identified as dispensable were more likely to be affected by CNVs than genes that were highly conserved among angiosperms. Finally, several large scale CNVs (with sizes above 100 kb) were detected, mostly affecting the heterochromatic or peri-centromeric regions of chromosomes, especially chromosomes 5 and 7 (Hardigan *et al.* 2016).

While CNVs provide useful information about the genetic diversity of potato, another promising approach is to assemble different reference genomes for each potato variety using re-sequencing data. Research in humans has shown there are a number of complex structural variants that are difficult to discover without new assemblies, especially when the heterozygosity found in diploid genomes is taken into account (Chaisson *et al.* 2015; Pendleton *et al.* 2015). A recent de novo assembly of a diploid wild potato species (*Solanum commersonii*) revealed significant differences in the distribution of SNPs, a lower degree of heterozygosity, fewer zones of repetitive DNA, and novel genes, when compared to the potato reference genome (see below *;Aversano et al.* 2015). This study, along with additional experiments performed in other Solanaceae crops such as tomato (Aflitos *et al.* 2014), highlight the potential benefits to further sequence, assemble and analyze close potato varieties and close relatives.

However, genome assembly in potato and other plants remains a complex problem and usually requires more data and computational resources than assemblies for microorganisms or even the human genome. The main challenges for genome assembly in plants have to do with resolving repetitive regions and dealing with polyploidy. Repetitive sequences can be difficult to resolve using only small-read DNA sequencing because is not enough information to distinguish between several repeated sequences (Kosugi *et al.* 2015). A similar problem arises when resolving

different isoforms or alleles of the same gene, especially in diploids and polyploids (Liang *et al.* 2016). Without the development of new sequencing technologies that can produce longer reads, these challenges would remain difficult to overcome, always requiring more and more data which is not only expensive but also complicating the data analysis process because more computational power is also required.

That is why the recent development and refinement of long-read sequencing technologies is expected to have a great impact in the field genomics. As their name implies, long-read sequencers produce reads that can range from 5 kb to 50kb, much more than what is currently achievable using second-generation sequencers such as those produced by *Illumina* (Goodwin *et al.* 2015; Rhoads & Au 2015). Also, because these long reads come from a single molecule, long-read sequencers provide data that permits the differentiation between alleles and haplotypes, something that is very challenging using just short reads (Rhoads & Au 2015; Liang *et al.* 2016). Recent efforts to assemble a *de novo* genome in non-model plant species such as *indica* rice, carrot, and pineapple (Ming *et al.* 2015; Iorizzo *et al.* 2016; Mahesh *et al.* 2016), relied on new strategies that combine different types of sequencing data, including long reads, to produce higher quality genomes. There have even been examples of plant genomes assembled using long-read data exclusively, such as the desiccation-tolerant grass *Oropetium thomaeum* (VanBuren *et al.* 2015).

Another way of overcoming the current limitations of short read sequencing is by improving the DNA library preparation process. Recent advances in library preparation methods, such as those developed by 10X Genomics and Dovetail Genomics, have made it possible to approximate the information of long-sequence reads while still using short-read sequencers to generate the data (Eisenstein 2015; Putnam *et al.* 2016). These new technologies have also started to be used in plant genome assembly efforts including at least one wild potato species (Bredeson *et al.* 2016; Michelmore *et al.* 2016; Paajanen *et al.* 2016; Reyes Chin-Wo *et al.* 2016).

One final approach researchers have used to overcome the challenges associated with assembling new plant genomes is to take advantage of the reference genomes that are currently available as a way to reduce the need for more data. A study in *Arabidopsis* used this approach to assemble four new genome sequences for divergent strains. By doing a whole genome alignment of the sequencing reads to the reference genome, the initial dataset was divided into smaller groups

of well-aligned, contiguous reads that were then used as input for assembly software. Unaligned reads were assembled separately and integrated at a later stage in the scaffolding process (Schneeberger *et al.* 2011). The results show that using a reference-guided approach effectively increases the coverage of the resulting assemblies. However, reference-guided assemblies had comparatively worse statistics that those produced *de novo,* including a lower N_{50}. The reference-based assemblies have enabled the discovery of previously unknown variants, including several large-scale variations and experiments, such as those involving small RNA (sRNA), produce better results when aligned to a strain-specific genome than to the generic *Arabidopsis* reference genome (Schneeberger *et al.* 2011).

Conversely, if the purpose of a genome re-sequencing study is to identify all the non-redundant DNA sequences in a particular population, another approach has been developed that utilizes a metagenome-like assembly strategy. Briefly, the procedure consists of sequencing all the individuals of the population at a very low coverage, and then using this data in addition to a well annotated reference to identify unique sequences that are present in at least two of the individuals. The effectiveness of this technique has been shown in rice (*Oryza sativa*) where 1,483 accessions were sequenced enabling the assembly and mapping of most of the known agronomically important genes that were previously absent from the Nipponbare rice reference genome (Yao *et al.* 2015). In future studies where the detection of large-scale structural variants is not important, this approach can reduce the amount of sequencing data required while still enabling the discovery of novel sequences found in a sub-set of individuals in a population.

2.6 Genome Assembly of a Wild Potato Species

Solanum commersonii is a wild potato species that is sexually incompatible with *S. tuberosum* due to different endosperm balance numbers (Johnston *et al.* 1980). Breeding efforts have allowed introgression of alleles from this species into cultivated potato by overcoming the reproductive barriers; however, little progress has been made on the release of new varieties originating from *S. commersonii* (Aversano *et al.* 2015). This species has also been shown to be genetically distinct from cultivated potato by chloroplast restriction sites and nitrate reductase gene sequences (Rodríguez & Spooner 2009). *S. commersonii* has generated interest because it contains important agronomic traits such as resistance to root knot nematode, soft rot and blackleg, bacterial and *Verticillium* wilt, potato virus X, tobacco etch virus, common scab, late blight, and the ability to

acclimate to the cold/freezing conditions (Hanneman & Bamberg 1986; Hawkes 1990; Micheletto *et al.* 2000). Genomic efforts in this species may help reveal important genes or the molecular pathways for specific traits which could be further utilized to improve cultivated potato.

In 2015, the first draft of a wild potato genome was released using a whole genome shotgun sequence and assembly approach based on size selected, paired end and mate pair libraries ranging from 400 bp to 10 kb (Aversano *et al.* 2015). The genome size was slightly smaller (830 Mb) than the cultivated form which was mainly due to variations in intergenic sequences. After filtering the data, a total of 278,460 contigs with an N_{50} of 6506 nt were assembled. In total, 64,665 scaffolds greater than 1 kb were produced with a mean scaffold length of 13,543 nt. The potato reference genome was utilized to map *S. commersonii* scaffolds and anchor them on each chromosome, producing 12 pseudomolecules (Aversano *et al.* 2015).

Even though *S. commersonii* is known to be an allogamous species, it had a low rate of heterozygosity compared to the reference genome of the cultivated variety (1.5% versus 53-59%), which could be due to the maintenance of this germplasm *ex situ,* artificially reducing the diversity level. The repeated sequences were also reduced (44.5% versus 55%) in *S. commersonii* compared to cultivated reference genome *S. tuberosum*. Ty3-gypsy type long terminal repeat retrotransposons (LTR-RTs) were the predominant transposable elements (TEs) identified in the genome, but the lower frequency of TEs found relative to cultivated potato and tomato may also contribute to its smaller genome size. An evaluation of the diversity demonstrated that the majority of SNPs had a distance of < 50 bp to their nearest neighbor. The divergence time between cultivated potato and *S. commersonii* was estimated to be approximately 2.3 million years ago (Aversano *et al.* 2015).

Transcriptome data was produced from leaf, flower, stolon, and tubers, from which a total of 37,662 genes were predicted (Aversano *et al.* 2015). The annotated genes for *S. commersonii* were evaluated and compared to cultivated potato and tomato. Pathogen resistance (R) genes were compared between *S. commersonii*, *S. tuberosum*, and *S. lycopersicum*. The wild potato had fewer putative R genes than the cultivated form, but more than the tomato genome. Factors such as genome size, natural and artificial selection, polyploidization, breeding, and gene family interactions can all contribute to pathogen resistance gene evolution. It is possible that the R genes in these three species vary due to different pathogenic pressures (Andolfo *et al.* 2014;

Aversano *et al.* 2015). However, it could also be an artifact due to the disparity in the quality of each assembly. Further evidence will be required to reach a conclusion.

Cold response genes were also compared (Aversano *et al.* 2015), resulting in 5853 predicted protein sequences revealed in *S. commersonii* and 8666 in *S. tuberosum*. These predicted proteins were similar to cold responsive genes found in the annotation of the *Arabidopsis thaliana* genome. The expression profiles of *S. commersonii* were further investigated to identify the genes involved in freezing and cold acclimation response. A total of 855 genes were determined to be differentially expressed in plants acclimated to frost stress and non-acclimated plants. A total of 11 transcription factors were negatively correlated and 25 were positively correlated to acclimated and non-acclimated tolerance. Collectively, these results show how comparing related genomes can aid scientists in revealing differences in gene function and regulatory elements. Generally conserved sequences across distant species are likely constrained implying similar biological function (Alföldi & Lindblad-Toh 2013).

2.7 Genomics and Genotyping

Whole genome re-sequencing can reveal important differences between cultivated potato varieties and related wild species, especially at a large scale. Traditionally, the cost of resequencing entire populations of samples has been prohibitive, and thus, there has been a need for novel solutions to genotype large collections of potato germplasm. The recent and tremendous reduction in costs associated with high-throughput sequencing have enabled the development of genetic markers with a single nucleotide resolution that can be rapidly assayed on hundreds to thousands of individuals. These molecular markers can be used in applications such as marker-assisted breeding, quantitative trait loci (QTL) determination, genome-wide association analyses (GWAS), as well as, evolutionary and diversity studies (Uitdewilligen *et al.* 2013).

Genotyping arrays have been the most common tool for high-throughput SNP genotyping in the last decade. Arrays have been developed for multiple platforms (including Infinium and Axiom) and offer many advantages over low-throughput gel-based genotyping platforms: a relatively low cost per sample, automation and standardization that makes it easy to analyze and compare the results of many individual samples. However, regardless of platform, array development is costly, time consuming, and requires extensive knowledge of the target genome.

Additionally, researchers that use arrays are limited to the genes or sequences that are included in the platform (De Donato *et al.* 2013).

There have been several SNP arrays developed for potato. Currently, one of the most popular is the Infinium 8303 Potato Array (Felcher *et al.* 2012) which was developed using SNPs discovered in two previous studies: one that mined markers from potato EST databases (Anithakumari *et al.* 2010) and a second that analyzed cDNA sequences from six elite potato germplasm accessions (Hamilton *et al.* 2011). As its name suggests, this array contains 8,303 SNP markers chosen to provide roughly even distribution across all 12 potato chromosomes. Out of the total number of markers, 536 were previously used genetic markers, 3,018 were selected from candidate genes of interest, and 4,749 were selected for maximum genome coverage (Felcher *et al.* 2012). This platform has proven useful in a number of studies, including genetic mapping of important agricultural traits (Manrique-Carpintero *et al.* 2015, 2016; Massa *et al.* 2015; Endelman & Jansky 2016), retrospective analysis of potato breeding (Hirsch *et al.* 2013) and taxonomic studies (Hardigan *et al.* 2015).

A second recently developed SNP platform is the SolSTW array. It includes a total of 14,530 SNP markers, the majority of which were selected from a previous sequence based genotyping experiment (Vos *et al.* 2015). The design of this array was focused on expanding the genetic sources of the markers, reducing biases and making it more useful for applications such as marker-assisted breeding. As opposed to the Infinium array that used the transcriptome of only six elite cultivars as the main source for markers, the majority of the markers in the SolSTW array are derived from a broad sequencing study (see below) that included 84 unique individuals and included chloroplastic DNA (Uitdewilligen *et al.* 2013; Vos *et al.* 2015).

As an alternative to genotyping arrays, several new sequencing based genotyping methods have emerged, leveraging high-throughput, short read sequencing to genotype hundreds of individuals simultaneously at thousands of genetic loci. The two most common methods are: genotyping-by-sequencing (GBS; Elshire *et al.* 2011) and RAD-seq (Baird *et al.* 2008). Both techniques have become popular in the agricultural genomics and ecological genetics communities respectively. In each case, a small subset of the genome is sequenced at low coverage, providing a relatively cheap tool to identify molecular markers. This reduced representation of the genome is constructed by digesting the genome with restriction enzymes (GBS) or digestion in

combination with physical shearing (RAD-seq). The reduced representation libraries from many individuals can be DNA barcoded, pooled, and then sequenced in the same experiment, greatly reducing the cost per sample. Post sequencing analyses can be performed using available software packages and tools, including TASSEL-GBS (Bradbury *et al.* 2007; Glaubitz *et al.* 2014), UNEAK (Lu *et al.* 2013), Stacks (Catchen *et al.* 2011), Haplotag (Tinker *et al.* 2016) and GBS-SNP-CROP (Melo *et al.* 2016).

While there are many benefits to using GBS or RAD-seq over genotyping arrays, including no requirement of a complete reference genome, no array ascertainment bias, and the ability to identify multiple types of genetic markers, significant challenges remain. The main obstacle is the sparse genotype matrix that is missing genotype calls, produced during the computational step that calls and filters SNPs and indels. This is due to the finite amount of sequencing data produced in one experiment, which is spread across many sequenced individuals, in other words, the tradeoff of sequencing coverage and depth among multiplexed DNA samples. It is not uncommon to see sequence-based genotyping studies tolerate between 20–50% missing genotype data (Elshire *et al.* 2011). Despite this hurdle, GBS has been successfully implemented in genetic mapping studies of diploid crops such as maize (producing 200,000 markers; Elshire *et al.* 2011), wheat and barley (producing 20,000 and 34,000 SNPs, respectively; Poland *et al.* 2012), and polyploid crops such as alfalfa (11,694 SNPs; Rocher *et al.* 2015).

In potato, there has been limited application of GBS for molecular marker development perhaps due to the highly heterozygous, tetraploid genome. In one instance, however, a modified GBS approach has been successfully used in marker discovery as part of a study that genotyped a panel of 83 tetraploid potato varieties chosen to represent the most important commercial cultivars and landraces worldwide (Vos *et al.* 2015). This study also included a monoploid clone related to the variety used to develop the potato reference genome. In total 12.4 Gb of sequence data were produced, which resulted in the identification of 129,156 markers. Out of that total, ~111k corresponded to SNPs, ~13k were insertions or deletions, and ~5k were multi-nucleotide polymorphisms. These markers were then successfully used in analyses to determine population structure, sequence diversity, chloroplast type and genetic association (Vos *et al.* 2015).

The successful use of GBS in tetraploid potato cultivars opens the door to future studies exploring the wider diversity of commercial and non-commercial potato varieties. Similar studies

in other Solanaceae, such as tomato, show the potential benefits of using this technique to explore wild species diversity (Labate *et al.* 2014). Additionally, it has been recently reported that GBS can be used to aid in the analysis of diploid potato mapping populations (Endelman 2015). A summary of the different SNP genotyping tools discussed in this section can be found in **Table 2.2**.

Marker assisted selection (MAS) increases the efficiency of breeding (Barone 2004). Markers are identified using genetic mapping, which is hampered in potato by complex tetraploid genetics and heterozygosity. To date most MAS studies in potato have relied on low-throughput molecular markers, including amplified fragment length polymorphism (AFLP) and simple sequence repeats (SSRs) that have been associated with traits with relatively simple genetic basis such as disease resistance. For example, several studies have identified loci associated with resistance to late blight (Tiwari *et al.* 2013), potato virus Y (Song *et al.* 2005; Gebhardt *et al.* 2006; Fulladolsa *et al.* 2015; Nie *et al.* 2016), potato virus X (Ritter *et al.* 1991; Gebhardt *et al.* 2006) and *Verticillium* wilt (Simko *et al.* 2004; Uribe *et al.* 2014). In contrast, there are markedly fewer studies focusing on polygenic (i.e. quantitative) traits such as tuber quality (Li *et al.* 2013), and tuber starch and yield (Schönhals *et al.* 2016). Regardless of trait, many of these low-throughput, gel-based, markers in their current form are not suitable for large scale screening of progenies, which would be required for application in a breeding program. One option would be to convert the gel based markers to a more efficient platform, as has been recently done for potato virus Y resistance markers (Simko *et al.* 2004). A second option would be to validate the existing marker-trait associations with the array- or sequence-based genotyping platforms, and identify SNP markers linked to the trait of interest. The latter option would be preferable, as it could be done as a byproduct of generating genome-wide marker information, which could in turn be used in future QTL mapping or genome wide association for novel traits. The successful use of high-throughput genotyping platforms (**Table 2.2**), in potato, opens the door to exploring the wider diversity of potato genetic variation, and the practical application of MAS in breeding programs. Ultimately, the genome-wide marker information could be used to go beyond MAS at a few loci, to being able to predict the phenotype solely from marker genotypes at all marker loci using whole genome selection methods (Slater *et al.* 2016).

Bulked segregant analysis (BSA) is emerging as a method for genetic mapping that has a particularly good compatibility with genome re-sequencing. BSA is an approach for gene mapping

where pooled DNA from individuals is genotyped as a single bulked sample. The method was originally applied in lettuce using individuals from a single biparental cross that segregated for a downy mildew resistance (Michelmore *et al.* 2016), but it can also be used for three-way, four-way and multiparental crosses, including those developed with special designs such as diallel design, North Carolina design (NCD), multiparent advanced generation intercross (MAGIC) and nested association mapping (NAM; Zou *et al.* 2016). Traits are quantified for all individuals in the population. Most commonly, individuals at the two extremes ends of the trait distribution are identified and their DNA is pooled, however other pooling strategies have also been used. Genome re-sequencing of the two pools plus two parents is a cost-effective way of getting high density genotyping data. Sequence data are mapped to a reference sequence and base distribution across the genome is analyzed. Detection of trait-associated variants in pooled sequence data involves use of statistical analysis to compare observed base distributions in the pools with that predicted by parental base distributions (Bansal *et al.* 2012; Kaminski *et al.* 2016). The selection of individuals for pools, genetic architecture of the trait and population size are other factors affecting the power of BSA (Zou *et al.* 2016). BSA was successfully used in potato to map steroidal glycoalkaloid content in tetraploids (Kaminski *et al.* 2016). As sequencing costs drop the use of whole genome sequencing for genotyping will become more widespread.

Table 2.2: Summary of popular array and GBS platforms in potato.

	Gene Expression Arrays		***SNP-Arrays***		***Genotyping-by-Sequencing (GBS)***
	POCI 44k	JHI *Solanum tuberosum* 60k	Infinium 8303	SolSTW	
SNPs Markers	N/A	N/A	8303	17,987	111,212
Expression Markers	42,034	52,998	N/A	N/A	N/A
Additional Markers*	N/A	N/A	N/A	N/A	17,944
Total Number of Markers	42,034	52,998	8303	17,987	129,156
Year	2005	2013	2012	2015	2013
Source of Genetic Information	Previous data on differentially expressed transcripts and a custom text mining approach for conserved sequences.	Predicted transcripts from the potato reference genome v3.4	Transcriptomic data from previous experiments, selected for representation of genes of interest and maximum genome coverage.	A combination of GBS derived markers and previously included markers in the Infinium 8303 array.	A panel of 83 tetraploid potato cultivars selected to represent the global gene pool of commercial potato, mostly covering accessions with high breeding value.
Comments	Using the Agilent 60-mer oligo platform.	Using the Agilent 60-mer oligo platform.	Some markers were mapped to the unanchored superscaffold of the potato reference genome	Includes a small portion of chloroplast markers.	-
Reference	(Kloosterman *et al.* 2008)	(Bengtsson *et al.* 2014)	(Felcher *et al.* 2012)	(Vos *et al.* 2015)	(Uitdewilligen *et al.* 2013)

* Including insertions, deletions and other multinucleotide polymorphisms.

2.8 Conclusion

Overall, modern sequencing technologies have fundamentally changed the field of plant genomics. It is now possible to identify large structural variations among closely related species, something that was extremely challenging just few years ago. These new resources provide scientists and producers with better tools to continue working on the discovery of new genes and regulatory mechanisms. In turn, knowledge generated this way can inform future crop improvement efforts. In the case of tuber bearing Solanaceae, there is already a fair amount of evidence pointing to important genetic differences within these species. A summary of additional genomics resources for potato and related species can be found in **Table 2.3**. However, more research is required especially in wild relatives of commercial potato, which could be important sources of genetic diversity but have remained relatively unexplored so far.

Table 2.3. Summary of genomics resources available for potato and related species.

Name of resource	Description	Web Address	Reference
Potato Genomics Resources			
Spud DB: Potato Genomics Resource	Latest versions of the potato reference genome, as well as a genome browser and several other potato genomics resources.	http://solanaceae.plantbiology.msu.edu/index.shtml	(Hirsch *et al.* 2014)
NCBI Genome (Potato)	The reference genome listed for S. tuberosum in the NCBI Genome database.	https://www.ncbi.nlm.nih.gov/genome/400	(Wheeler *et al.* 2007)
NCBI GEO (Potato)	Gene expression datasets for *S. tuberosum*	https://www.ncbi.nlm.nih.gov/gds/?term=Solanum+tuberosum	(Wheeler *et al.* 2007)
ArrayExpress (Potato)	Array-based gene expression datasets for *S. tuberosum*	https://www.ebi.ac.uk/arrayexpress/search.html?query=Solanum+tuberosum	(Brazma *et al.* 2003)
PoMaMo Database	Database containing potato genomic maps and sequences.	http://www.gabipd.org/projects/Pomamo/#Tools	(Meyer *et al.* 2005)
The NSF Potato Genome Project	Portal containing several potato genomics resources including SSR and microarrays.	http://potatogenome.berkeley.edu/nsf5/	N/A
Potato Variety Databases			
The Potato Association of America Variety Database	Catalogue of potato varieties in the US.	http://potatoassociation.org/industry/varieties#Breeding	N/A
Canadian Potato Varieties Database	Catalogue of potato varieties in Canada.	http://www.inspection.gc.ca/plants/potatoes/potatovarieties/eng/1299172436155/1299172577580	N/A
European Cultivated Potato Database	Catalogue of European potato varieties.	https://www.europotato.org/menu.php	N/A
AHDB Potato Variety Database	Agriculture & Horticulture Development Board Catalogue of British potato varieties.	http://varieties.ahdb.org.uk/	N/A
Potato Germplasm Banks			
International Potato Center (CIP) Genebank	Worldwide collection of potato and sweet potato varieties and wild relatives.	http://cipotato.org/genebank/	N/A
NRSP-6 - United States Potato Genebank	Collection of germplasm of cultivated potato varieties and wild	http://www.ars-grin.gov/ars/MidWest/NR6/	N/A
Centre for Genetic Resources, The Netherlands (CGN)	Dutch-German collection of wild and Andean cultivated species.	http://www.wur.nl/en/Expertise-Services/Statutory-research-tasks/Centrefor-Genetic-Resources-the-Netherlands-1/Centre-for-Genetic-Resourcesthe-Netherlands-1/Expertise-areas/Plant-Genetic-Resources/CGN-cropcollections/Potato.htm	N/A
N. I Vavilov Institute of Plant Genetic Resources (VIR)	Wild *Solanum* species, cultivated species and indigenous Chilean cultivars, breeding varieties, hybrids and dihaploids.	http://vir.nw.ru	N/A
Canadian Potato Genetic Resources	Collection of Canadian and international potato germplasm that is part of Plant Gene Resources of Canada.	http://pgrc3.agr.gc.ca/index_e.html	N/A

Name of resource	Description	Web Address	Reference
Commonwealth Potato Collection	United Kingdom genebank of landrace and wild potatoes	http://germinate.hutton.ac.uk/germinate_cpc/app/	N/A
Other Solanaceae Resources			
Sol Genomics Network	A variety of genomics resources for several of the most important Solanaceae species.	https://solgenomics.net/	(Fernandez-Pozo *et al.* 2014)
Solanaceae Coordinated Agricultural Project (SolCAP)	A collection of germplasm, phenotype and genotype data on several Solanaceae species.	http://solcap.msu.edu/index.shtml	(Felcher *et al.* 2012)
GoMapMan	Open-source for manual gene functional annotations in plants, including potato, tomato and tobacco.	http://www.gomapman.org/	(Ramšak *et al.* 2014)

Preface and Author Contributions to Chapter 3

Gene regulatory mechanisms in potato and other crops remain an interesting subject of study, especially when they can be associated with important phenotypic characteristics (Swinnen *et al.* 2016). Because technologies such as RNA-seq have now made it possible to identify differentially expressed genes, the flanking regions of these genes can also be examined for potential regulatory motifs. As mentioned in the previous chapter, it is possible to predict regulatory motifs *de novo* using specialized bioinformatics algorithms. These putative motifs can be used to query experimentally annotated motif databases such as JASPAR (Sandelin *et al.* 2004) and PLACE (Higo *et al.* 1999).

In this chapter, the transcriptomes of field-grown potatoes from three different cultivars (Shepody, Russet Burbank and Atlantic) were compared to identify N responsive genes. Then, the flanking regions of these genes were analyzed with motif prediction software to discover putative regulatory motifs associated with steady state N response in these cultivars. Understanding the gene regulatory mechanisms behind N response in potato is key for producers and breeders because N constitutes one of the most important macronutrients for the growth of potato tubers.

A version of this chapter was published in *Scientific Reports* on May 19, 2016 (Gálvez *et al.* 2016). José Héctor Gálvez wrote the main manuscript text, with contributions from Helen H. Tai and Martina V. Strömvik. Under the supervision of Martina V. Strömvik, José Héctor Gálvez designed and carried out the bioinformatics and computational analyses, with additional contributions from Helen H. Tai, Martin Lagüe, and Rui Li. Helen H. Tai and Bernie J. Zebarth designed and carried out the field experiment. All authors reviewed the manuscript.

Computations were made on the supercomputer Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec-Nature et technologies (FRQ-NT). Sequencing data for this work was uploaded to the NCBI Gene Expression Omnibus (GEO) accession series GSE75926 (GSM1970293-1970340).

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. 283303) to Martina V. Strömvik; A-base funding from Agriculture and Agri-Food Canada to Helen H. Tai and Bernie J. Zebarth; by the Mexican National

Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología; CONACYT) (Scholarship No. 381158) to J.H.G.; and by the McGill Department of Plant Science Graduate Excellence Fund. The authors also acknowledge Centre SÈVE for funding. The authors thank Jeff Xia for helpful discussion on ANOVA analyses, Charlotte Davidson and Lana Nolan for technical assistance with gene expression analysis and Karen Terry for technical assistance with field experiments.

Chapter 3: The Nitrogen Responsive Transcriptome in Potato (*Solanum tuberosum* L.) Reveals Significant Gene Regulatory Motifs

3.1 Introduction

Potatoes (*Solanum tuberosum* L.), constituting the third most grown staple crop worldwide, usually have a sparse and shallow root system, and are therefore particularly sensitive to abiotic factors such as water and nutrient availability (Birch *et al.* 2012). The macronutrient nitrogen (N) positively impacts potato biomass, tuber yield and quality, especially in fields with a limited natural supply (Westermann 2005; Zebarth & Rosen 2007). However, excessive application of N can have two main undesirable effects: 1) decreased quality of the tubers which can render them less suitable for industrial food production (Long *et al.* 2004) and 2) leaching of nitrate into water supply systems and the emission of nitrous oxide, both of which can cause environmental damage (Zebarth & Rosen 2007). Therefore, long-standing goals within the potato production sector are to increase plant N use efficiency as well as develop sustainable N management systems to optimize N supplementation to the amount required to maintain plant growth and achieve target yields (Westermann 2005; Birch *et al.* 2012).

Whole transcriptome analyses using RNA-seq to examine genes involved in N deficiency responses have been done in maize (Humbert *et al.* 2013), *Arabidopsis* (Vidal *et al.* 2013), cucumber (Zhao *et al.* 2015), and rice (Yang *et al.* 2015). Both a well-annotated reference genome and a reference transcriptome (gene models) are needed to carry out differential gene expression analysis using RNA-seq. The potato reference genome and transcriptome were initially published in 2011 (Massa *et al.* 2011; The Potato Genome Sequencing Consortium 2011). In 2012, a new annotation system (ITAG1.0) with updated gene models generated using new data from the tomato reference genome, was made available for both potato and tomato (The Tomato Genome Consortium 2012; Fernandez-Pozo *et al.* 2014). These resources have fuelled a renewed effort to analyse the molecular response of potato under different biotic and abiotic conditions (Gong *et al.* 2015).

N related regulatory motifs have been identified in maize and *Arabidopsis thaliana* genes (Konishi & Yanagisawa 2011; Liseron-Monfils *et al.* 2013) and point to coordinated responses to

nitrate at the transcriptional level in plants. One of these motifs, the Nitrate Related *cis*-Element (NRE) identified in *Arabidopsis*, has also been found in the promoter region of the Nitrite Reductase (*NIR*) gene of several monocotyledonous and dicotyledonous plants such as spinach, tobacco, rice, maize and sorghum (Konishi & Yanagisawa 2011). However, few studies have been done on regulatory motifs in the upstream regions of genes in potato. A study analysing the level of expression of a transgenic patatin class-I and β-glucuronidase (GUS) chimeric gene in field-grown potato found significant changes in expression depending on the promoter used in the construct and the regulatory motifs it contained (Aminedi & Das 2014). These results highlight the need to further study and understand potential gene regulation mechanisms in potato, especially in response to critical abiotic factors such as N sufficiency.

Additionally, transcriptome analysis can be used for the discovery and annotation of overrepresented motifs. Since regulatory motifs tend to be overrepresented in the genome and those acting on similar genes often cluster locally, genomic information in coupled with differential gene analysis can be used to predict motifs *de novo* (Harbison *et al.* 2004). Specialized algorithms such as Seeder (Fauteux *et al.* 2008) have been used before to predict the binding sites of regulatory elements in the upstream flanking regions of genes in other plant species (López *et al.* 2013; Zolotarov & Strömvik 2015). Differences in the 5'-upstream flanking regions of potato genes, including variations in the number and types of regulatory motifs, have also been correlated with changes in gene expression (Aminedi & Das 2014).

Transcriptome analysis can also be applied as an alternative method for quantifying N sufficiency (Li *et al.* 2010; Yang *et al.* 2011; Zebarth *et al.* 2011). Expression profiles associated with N sufficiency can be used to guide decisions on N fertilizer application in potato fields. Other technologies proposed for nutrient monitoring in crops include biosentinel plants that use promoters from nutrient responsive genes to drive reporter genes (Hammond *et al.* 2011). Both of these approaches can be enhanced through transcriptome analysis.

The current study uses RNA-seq data generated from three commercial potato cultivars (Shepody, Russet Burbank and Atlantic) to examine the steady state transcriptome response of potato to N supplementation. Genes with expression that was affected by the rate of supplemented N were further analysed for overrepresented DNA motifs in their upstream flanking regions

through *de novo* motif discovery analysis. In all, 39 genes were differentially expressed in all three cultivars, and in total, nine potential nitrogen responsive motifs were identified.

3.2 Results

3.2.1 Effects of N Supplementation on Potato Plants

The availability of N in the soil is known to cause measurable changes in certain characteristics of potato plants including dry biomass at harvest, fresh tuber yield, and chlorophyll content (Zebarth & Rosen 2007). To determine the effects of N sufficiency, two contrasting rates of N supplementation were applied (0 kg N ha^{-1} and the recommended rate of 180 kg N ha^{-1}) in a randomized complete block design (**Table 3.1**). Four replicated blocks were used, and every sample consisted of pooled tissue from 15 randomly selected plants from a single block. All trait measurements were statistically tested with a two-factor Analysis of Variance to determine the significance of the observed changes among plants from different cultivars grown under different N supplementation rates (**Table 3.2, *Supplementary tables 3.1*** and ***3.2***).

The chlorophyll content index was measured in foliar tissue samples collected from the field grown plants using Special Products Analysis Division (SPAD) readings. Plants without N supplementation had significantly lower SPAD readings than those grown with supplemented N (**Figure 3.1a**). This result indicates that plants grown without added N had lower concentrations of chlorophyll in their foliar tissue, which is indicative of reduced N sufficiency. The SPAD readings among plants of different cultivars were also significantly different.

Petioles were collected from the same leaves used for the SPAD readings and the concentration of petiole nitrates was chemically determined for each biological replicate. Petioles collected from plants without supplemented N had significantly lower concentrations of petiole nitrates than did the plants grown with the addition of N (**Figure 3.1b**). There were no significant differences in the petiole nitrate concentrations among plants of different cultivars.

The effects of N on biomass were measured at harvest, before vine desiccation. To calculate the effect of N supplementation on biomass, whole plants were sampled and partitioned into vines, tubers, stolons and readily recoverable roots. Using information on the spatial distribution of plants in the field as well as the dry matter weight of the sampled plants, total biomass per hectare was calculated (**Figure 3.1c**). The results show that biomass significantly increased in the groups

grown with supplemented N. Dry matter weight in plants of different cultivars also differed significantly. However, the differences in biomass between cultivars were less pronounced than those observed due to N supplementation.

Finally, fresh tuber yield was also measured at harvest and was analysed to determine the effect of N supplementation. Tubers from the plants collected at harvest for biomass determination were also washed and weighed before drying. The total fresh yield for tubers was calculated using additional information on the spatial distribution of plants in the field (**Figure 3.1d**). Tuber yield varied significantly among the three cultivars. N supplementation also had an observable effect on fresh tuber yield, although not as significant as the cultivar effect.

Table 3.1: Experimental design for sampling the potato experiment at the Fredericton Research and Development Centre of Agriculture and Agri-Food Canada, Fredericton NB.

	Control (N deficient) [0 kg N ha^{-1}]		**Treatment (N sufficient) [180 kg N ha^{-1}]**	
***S. tuberosum* cultivar**	Time-point 1 July 25, 2012	Time-point 2 Aug. 8, 2012	Time-point 1 July 25, 2012	Time-point 2 Aug. 8, 2012
Shepody	R1, R2, R3, R4*	R1, R2, R3, R4	R1, R2, R3, R4	R1, R2, R3, R4
Russet Burbank	R1, R2, R3, R4	R1, R2, R3, R4	R1, R2, R3, R4	R1, R2, R3, R4
Atlantic	R1, R2, R3, R4	R1, R2, R3, R4	R1, R2, R3, R4	R1, R2, R3, R4

***** R1, R2, R3 and R4: Are biological replicates, each consisting of a pool of 15 randomly selected plants from each plot collected at 0800 h, 1100 h, 1400h and 1700 h respectively.

Table 3.2: Two-factor Analysis of Variance for phenotypic changes in potato grown under different N supplementation treatments.

		SPAD reading [a]		**Petiole nitrate concentration [a]**		**Plant dry matter accumulation [b]**		**Fresh tuber yield [b]**	
N treatment [kg N ha^{-1}]		***0***	***180***	***0***	***180***	***0***	***180***	***0***	***180***
Russet Burbank	**s1**	37.8	38.1	4.1	25.5	7.72	9.02	34.3	38.4
	s2	36.8	38.6	1.7	21.9				
Shepody	**s1**	33.4	37.1	3.2	24.0	8.07	9.24	31.1	36.1
	s2	29.3	35.6	1.4	23.3				
Atlantic	**s1**	35.9	36.0	1.4	20.1	9.32	10.58	40.6	43.4
	s2	33.0	35.8	0.3	24.0				
All Cultivars		**34.3**	**36.9**	**2.0**	**23.1**	**8.37**	**9.61**	**35.3**	**39.3**
Statistical Significance									
N treatment [N]	***df*=1**	< 0.0001***		< 0.0001***		0.009**		0.016*	
Cultivar [C]	***df*=2**	< 0.0001***		0.066		0.016*		< 0.001***	
N × C	***df*=2**	0.01*		0.92		0.99		0.85	

a = Average of four measurements (n=4) made at each sampling date (s1=2012-07-25, s2=2012-08-08) in [mg g^{-1}]
b = Average of four measurements (n=4) made at harvest (s1=2012-09) in [t ha^{-1}]
$ = Two-factor ANOVA. Significance codes: *** <0.001 ** <0.01 * <0.05

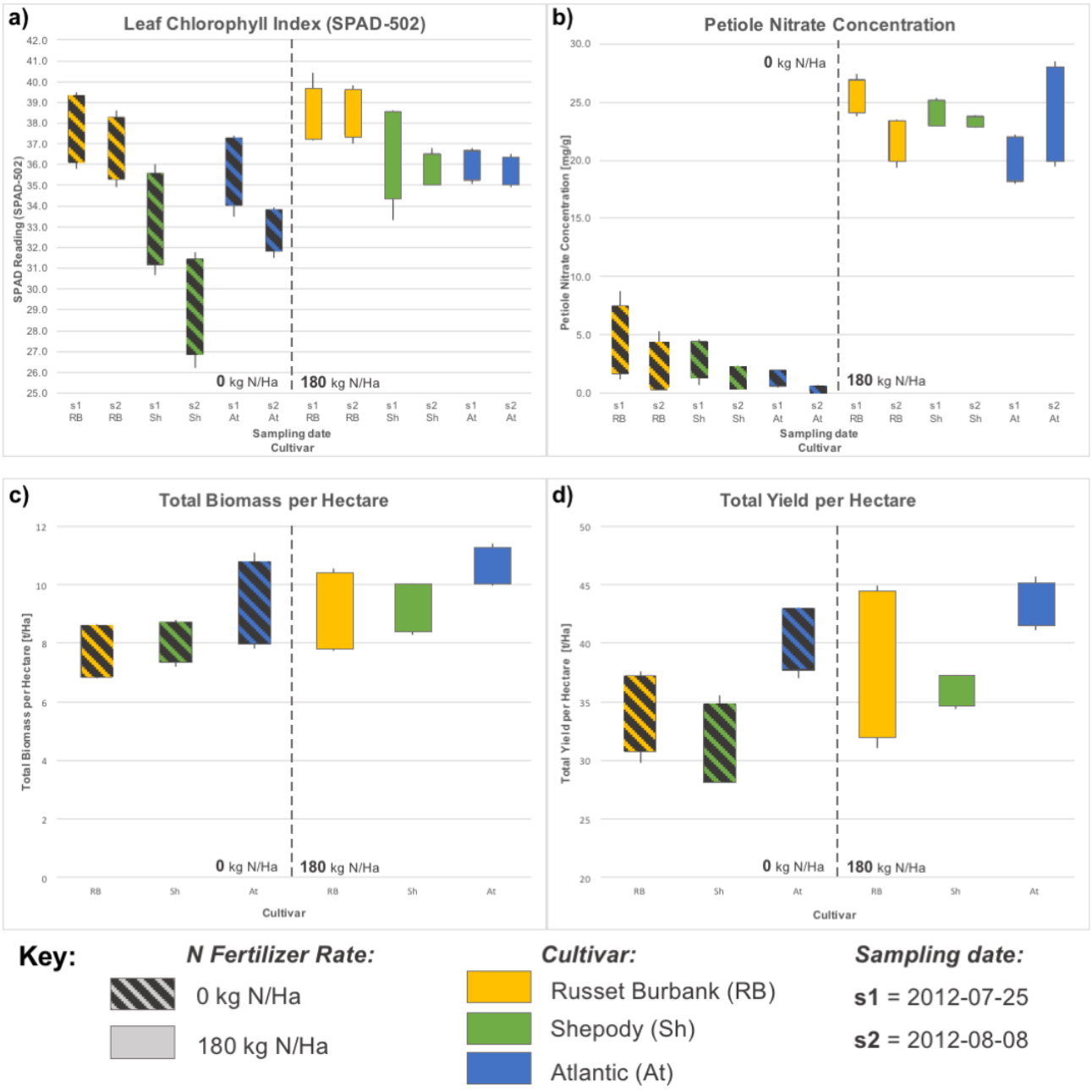


Figure 3.1: Comparison of four phenotypic traits in potato plants grown at two different N supplementation rates.

Plots showing different phenotypic measurements in potato plants from three cultivars (Shepody, Russet Burbank and Atlantic) grown at two rates of N supplementation (0 kg N ha^{-1} and 180 kg N ha^{-1}). In all cases, plants with no N supplementation display signs of N deficiency and early senescence. **a)** Relative leaf chlorophyll content measured by light transmittance using a SPAD-502 meter. **b)** Petiole nitrate concentration measured colorimetrically. **c)** Total plant biomass, measured from plant components (tubers, vines, stolons plus readily recoverable roots) for a representative sample of plants. **d)** Total fresh tuber yield in the field from a representative sample of plants.

3.2.2 Transcriptome Sequencing Reveals N Responsive Genes

Sequencing of RNA samples was carried out to determine the differences in the transcriptomes of plants grown under two N treatments, deficient and sufficient, to obtain a list of N responsive differentially expressed genes. Total RNA was extracted from the foliar tissue samples collected from the same apical leaflet used for SPAD readings and petiole sampling. Paired-end sequencing (2x100 cycles) of the prepared RNA libraries was performed using a HiSeq 2000 [Illumina]. Sequencing results were trimmed and filtered for quality, and aligned to the potato reference genome (The Potato Genome Sequencing Consortium 2011) using TopHat (Trapnell *et al.* 2009).

The transcriptomes of plants grown with and without supplemented N at 180 kg N ha^{-1} were compared to find differences in gene expression that were highly specific to N status and that were consistent across cultivars and developmental time points. Samples were collected from three cultivars (Shepody, Russet Burbank and Atlantic) at two developmental time points: eight and ten weeks after planting. Sampling of the four replicated blocks was done over the course of a day (i.e. 0800 h, 1100 h, 1400 h, and 1700 h for blocks 1 to 4, respectively). This allowed for the identification and removal of genes with significant time of day variation in expression in the analysis (as explained in: Tai & Zebarth 2015); focusing the analysis on genes whose expression level correlates only with the N supplementation treatment.

Aligned reads were analysed using CuffDiff (Kim *et al.* 2013) and lists of differentially expressed genes in plants grown under different N treatments were made for each cultivar and developmental time point. Gene lists were compared to find genes with similar expression patterns across all cultivars and developmental time points. The experimental design was focused on identifying genes involved in steady state responses to N supplementation. Differentially expressed genes that were common among the cultivars and development time points were divided into two groups: those over-expressed in plants grown with supplemented N and those under-expressed in plants with supplemented N. A summary of the number of genes found to be differentially expressed in each analysis can be found in ***Supplementary Table 3.3***.

In all, 30 genes were consistently over-expressed (**Table 3.3**) and nine genes under-expressed (**Table 3.4)** with N supplementation, across cultivars and over developmental time points. Differences were found in gene expression among cultivars and developmental time points, but they were not the focus of the current study. For example, 51 genes were over-expressed in

only one of the two time-points and another four genes were under-expressed in only one of the two time-points (***Supplementary Table 3.4***). Alternative splicing analysis was also performed on the raw RNA-seq data, but did not reveal alternate splicing events in the 39 genes found to be N responsive in all cultivars (data not shown).

Table 3.3: Genes found to be consistently over-expressed in plants grown with supplemental N across three cultivars and two sampling dates.

GeneID and Coordinates		Description and InterPro Domains[$]	GO Terms[°]	E.C. Numbers and KEGG Pathways
Sotub12g027600 chr12:64270939-64272220	-	Whole genome shotgun assembly reference scaffold set scaffold scaffold_4 IPR012336 Thioredoxin-like fold	**CC**: GO:0044444, GO:0043231	
Sotub02g033060 chr02:66207353-66223198	+	NAD-dependent epimerase/dehydratase IPR016040 NAD(P)-binding domain	**BP**: GO:0044237 **CC**: GO:0016021, GO:0044444, GO:0043231 **MF**: GO:0050662, GO:0003824	
Sotub10g014450 chr10:26464815-26466682	-	Phenylcoumaran benzylic ether reductase 3 IPR008030 NmrA-like	**BP**: GO:0055114, GO:0046686, GO:0044237, GO:0006694 **CC**: GO:0005737 **MF**: GO:0000166, GO:0050662, GO:0003854	**E.C.:** 1.3.1.45
Sotub04g026530 chr04:53765574-53768766	-	Peroxidase IPR002016 Haem peroxidase, plant/fungal/bacterial	**BP**: GO:0055114, GO:0042744 **CC**: GO:0009506, GO:0009505, **GO**:0005773, GO:0005576 MF: GO:0004601, GO:0020037	**E.C.:** 1.11.1.7
Sotub08g024220 chr08:39427772-39430543	+	Inositol 2-dehydrogenase like protein IPR016040 NAD(P)-binding domain	**BP**: GO:0055114 **CC**: GO:0005576 **MF**: GO:0050112	
Sotub08g007240 chr08:2517348-2519038	+	Cation transport regulator-like protein 2 IPR006840 ChaC-like protein	**BP**: GO:0046686, GO:0010288	
Sotub12g011100 chr12:5784211-5789157	+	Aminotransferase-like protein IPR005814 Aminotransferase class-III	**BP**: GO:0010154, GO:0046686, GO:0006979, GO:0010183, GO:0010033, GO:0009865, GO:0009450, GO:0006540, GO:0019484 **CC**: GO:0005886, GO:0005739, GO:0009507 **MF**: GO:0030170, GO:0008270, GO:0050897, GO:0003992, GO:0034387	
Sotub10g018540 chr10:43448319-43452146	+	Aminotransferase like protein IPR005814 Aminotransferase class-III	**BP**: GO:0050896, GO:0009853 **CC**: GO:0005739 **MF**: GO:0030170, GO:0008453	**E.C.:** 2.6.1.18 **Pathways**: pantothenate and coenzyme A biosynthesis II, β-alanine biosynthesis II **E.C.:** 2.6.1.44

GeneID and Coordinates		Description and InterPro Domains[$]	GO Terms[°]	E.C. Numbers and KEGG Pathways
				Pathways: glycine biosynthesis III
Sotub08g014020 chr08:22405555-22412221	+	Chalcone isomerase IPR016087 Chalcone isomerase	**CC**: GO:0009570 **MF**: GO:0016872, GO:0005504	
Sotub01g005580 chr01:427688-431987	-	Glutamate decarboxylase IPR010107 Glutamate decarboxylase	**BP**: GO:0006536, GO:0046686 **CC**: GO:0005634, GO:0005829 **MF**: GO:0030170, GO:0004351, GO:0005516	**E.C.:** 4.1.1.15 **Pathways**: glutamate degradation IV , glutamate degradation IX (via 4-aminobutyrate) , glutamate dependent acid resistance
Sotub10g024560 chr10:49073980-49076488	+	Glutathione S-transferase IPR004046 Glutathione S-transferase, C-terminal	**BP**: GO:0046686, GO:0009636 **CC**: GO:0005829 **MF**: GO:0005515, GO:0004364	**E.C.:** 2.5.1.18 **Pathways:** glutathione-mediated detoxification II
Sotub06g008080 chr06:4763763-4768618	-	Male sterility 5 family protein (Fragment) IPR011990 Tetratricopeptide-like helical	**MF**: GO:0005515	
Sotub09g018850 chr09:37905080-37908059	+	Male sterility 5 family protein (Fragment) IPR011990 Tetratricopeptide-like helical	**MF**: GO:0005515	
Sotub01g022620 chr01:69840666-69841253	-	Peptide methionine sulfoxide reductase msrB IPR002579 Methionine sulphoxide reductase B	**BP**: GO:0022900 **CC**: GO:0005829 **MF**: GO:0046872, GO:0033743, GO:0008113	**E.C.:** 1.8.4.11
Sotub05g024960 chr05:56803278-56805488	-	Amino acid transporter IPR013057 Amino acid transporter, transmembrane	**BP**: GO:0006865 **CC**: GO:0016021 **MF**: GO:0003674	
Sotub11g012150 chr11:6925093-6928544	-	Amino acid transporter IPR013057 Amino acid transporter, transmembrane	**CC**: GO:0005886, GO:0016021, GO:0005774 **MF**: GO:0015171	
Sotub02g036900 chr02:69051466-69053337	+	Cystine transporter Cystinosin IPR005282 Lysosomal cystine transporter	**BP**: GO:0006810 **CC**: GO:0005886, GO:0016021, GO:0005765	
Sotub09g024290 chr09:46805115-46809993	-	Sulfate adenylyltransferase IPR002650 ATP-sulfurylase	**BP**: GO:0046686, GO:0001887, GO:0000103, GO:0070814 **CC**: GO:0005886, GO:0005739, GO:0009570 **MF**: GO:000552, GO:0004781	**E.C.:** 2.7.7.4 **Pathways:** selenate reduction, sulfate reduction II (assimilatory), sulfate activation for sulfonation
Sotub04g021910 chr04:45794993-45803770	+	Sulfate transporter IPR001902 Sulphate anion transporter	**BP**: GO:0055085, GO:0030003, GO:0019344, GO:0009684, GO:0019761, GO:0070838, GO:0008272 **CC**: GO:0016021 **MF**: GO:0015293, GO:0008271	
Sotub04g027100 chr04:54585161-54588369	-	High affinity sulfate transporter 2 PR001902 Sulphate anion transporter	**BP**: GO:0055085, GO:0009970, GO:0008272, GO:0080160 **CC**: GO:0005886, GO:0016021	

GeneID and Coordinates		Description and InterPro Domains[$]	GO Terms[°]	E.C. Numbers and KEGG Pathways
			MF: GO:0015293, GO:0008271	
Sotub10g013960 chr10:24455891-24459003	-	High affinity sulfate transporter 2 IPR001902 Sulphate anion transporter	**BP**: GO:0055085, GO:0009970, GO:0008272, GO:0080160 **CC**: GO:0005886, GO:0016021 **MF**: GO:0008271	
Sotub08g005390 chr08:414965-419119	+	Nitrate transporter IPR000109 TGF-beta receptor, type I/II extracellular region	**BP**: GO:0009414, GO:0010167, GO:0009734, GO:0009635, GO:0006857, GO:0042128 **CC**: GO:0005886, GO:0016021 **MF**: GO:0015112, GO:0015293	
Sotub07g009860 chr07:6361828-6364633	+	Peptide transporter IPR000109 TGF-beta receptor, type I/II extracellular region	**BP**: GO:0009987, GO:0015031, GO:0006807, GO:0006857 **CC**: GO:0016021, GO:0009506 **MF**: GO:0042937, GO:0042936	
Sotub01g049270 chr01:96560230-96568031	-	Tyrosine-protein kinase transforming protein Src IPR015783 ATMRK serine/threonine protein kinase-like	**BP**: GO:0006468 **CC**: GO:0016597, GO:0005524, GO:0004674, GO:0004715	
Sotub03g018720 chr03:24381058-24384772	+	Alpha-glucosidase-like IPR000322 Glycoside hydrolase, family 31	**BP**: GO:0005975 **CC**: GO:0030246, GO:0032450	
Sotub01g023000 chr01:70600749-70601999	+	Xylanase inhibitor (Fragment) IPR001461 Peptidase A1	**BP**: GO:0006508 **CC**: GO:0004190	
Sotub11g007110 chr11:2283766-2284773	+	Plant-specific domain TIGR01615 family protein IPR006502 Protein of unknown function DUF506, plant		
Sotub11g007090 chr11:2266700-2267713	+	Plant-specific domain TIGR01615 family protein IPR006502 Protein of unknown function DUF506, plant		
Sotub04g023170 chr04:48412542-48413816	+	Unknown Protein		
Sotub03g017290 chr03:22560550-22560888	-	Unknown Protein		

[$] Gene descriptions (including InterPro domains) obtained from the ITAG1.0 annotation system. Strand: plus (+) or minus (-).

° **BP:** Biological Process; **CC:** Cell Component; **MF:** Molecular function

Table 3.4: Genes found to be consistently under-expressed in plants grown with supplemental N across three cultivars and two sampling dates.

GeneID and Coordinates		Description[$]	GO Terms[°]	E.C. Numbers and KEGG Pathways
Sotub12g012740 ** chr12:7720521-7725325	+	Chloroplast lipocalin IPR000566 Lipocalin-related protein and Bos/Can/Equ allergen	**BP**: GO:0006979 **CC**: GO:0009535, GO:0005576, GO:0031977 **MF**: GO:0005488	
Sotub02g033320 chr02:66437615-66439427	+	Proline dehydrogenase IPR015659 Proline oxidase	**BP**: GO:0055114, GO:0006979, GO:0009414, GO:0006970, GO:0006537, GO:0010133 **CC**: GO:0005739 **MF**: GO:0004657	**E.C.:** 1.5.1.2 **Pathways:** proline biosynthesis I, II (from argininte), and III, arginine degradation VI (arginase 2 pathway) **E.C.:** 1.5.99.8 **Pathways:** L-Nδ-acetylornithine biosynthesis, proline degradation
Sotub08g025870 * chr08:40745121-40749856	+	Primary amine oxidase IPR000269 Copper amine oxidase	**BP:** GO:0055114, GO:0009738, GO:0009308, GO:0006809 **CC:** GO:0005768, GO:0005802, GO:0005773 **MF:** GO:0005507, GO:0048038, GO:0008131, GO:0052596, GO:0052595, GO:0052594, GO:0052593	**E.C.:** 1.4.3.22, 1.4.3.21 **Pathways:** phenylethanol biosynthesis
Sotub09g009440 chr09:6333813-6338909	+	Cation/H+ antiporter IPR006153 Cation/H+ exchanger	**BP:** GO:0055085, GO:0006623, GO:0006813, GO:0006885, GO:0030007, GO:0030104 **CC:** GO:0016021, GO:0005783, GO:0009507, GO:0012505 **MF:** GO:0015385	
Sotub09g023510 chr09:45964253-45969159	+	High affinity sulfate transporter 2 IPR001902 Sulphate anion transporter	**BP:** GO:0055085, GO:0006950, GO:0008272 **CC:** GO:0016021, GO:0009507 **MF:** GO:0015293, GO:0008271	
Sotub02g017430 * chr02:51710764-51712899	-	Purine permease family protein IPR004853 Protein of unknown function DUF250	**BP:** GO:0016021	
Sotub05g028860 chr05:60345721-60347082	-	Flowering locus T protein IPR008914 Phosphatidylethanolamine-binding protein PEBP	**BP:** GO:0009909, GO:0048510, GO:0010229, GO:0030154, GO:0048575 **CC:** GO:0005737, GO:0005634 **MF:** GO:0008429	
Sotub12g031130 chr12:67026510-67029847	+	Poly(A) polymerase IPR007012 Poly(A) polymerase, central region	**BP:** GO:0006351, GO:0043631 **CC**: GO:0005737, GO:0005634 **MF**: GO:0003723, GO:0004652	
Sotub01g049920 chr01:97188627-97192358	-	Nodule inception protein (Fragment) IPR003035 Plant regulator RWP-RK	**BP:** GO:0006355 **CC:** GO:0005634	

GeneID and Coordinates		Description[$]	GO Terms°	E.C. Numbers and KEGG Pathways
			MF: GO:0003677	
Sotub12g020880 chr12:54796545-54800381	-	Ubiquinone/menaquinone biosynthesis methyltransferase ubiE IPR013216 Methyltransferase type 11	BP: GO:0032259, GO:0009860, GO:0009877, GO:0009555, GO:0048528, GO:0010183, GO:0009312, GO:0006656, GO:0042425 **CC:** GO:0005829 **MF:** GO:0000234	**E.C.:** 2.1.1.103 **Pathways:** superpathway of choline biosynthesis, phosphatidylcholine biosynthesis II, choline biosynthesis I
Sotub05g012720 chr05:7174173-7177870	-	Nodulin MtN21 family protein IPR000620 Protein of unknown function DUF6, transmembrane	**CC:** GO:0016020	
Sotub09g029950 chr09:52340673-52341549	+	Cell wall protein IPR010800 Glycine rich		
Sotub09g010630 * chr09:7495968-7498403	+	Hydrolase alpha/beta fold family protein IPR000073 Alpha/beta hydrolase fold-1		

[$] Gene descriptions (including InterPro domains) obtained from the ITAG1.0 annotation system. Strand: plus (+) or minus (-).
° **BP:** Biological Process; **CC**: Cell Component; **MF:** Molecular function
***** Genes found to be significantly under-expressed only in the first sampling date (2012-07-25)
****** Genes found to be significantly under-expressed only in the second sampling date (2012-08-08)

To validate the differential expression of the shared genes, the RNA samples were tested again using an nCounter Digital Analyzer (Nanostring Technologies, Inc.). Probes were designed for the 39 differentially expressed genes previously identified through RNA-seq. The nCounter reads were normalized using the expression of five housekeeping genes as a reference (Luo *et al.* 2011; de Almeida *et al.* 2015). A Spearman Rank correlation was used to compare the measurements obtained from the nCounter Digital Analyzer with previous data generated through RNA-seq (**Figure 3.2**). A positive correlation (r = 0.79) was found between both methods.

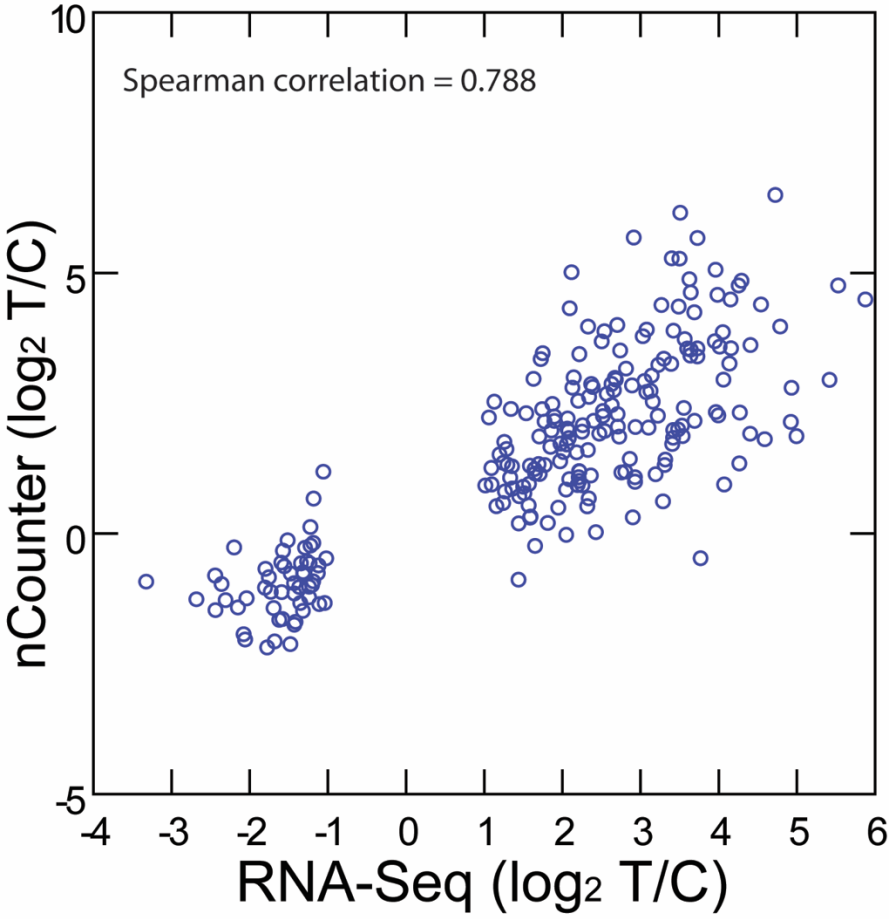


Figure 3.2: Correlation between gene expression measured using RNA-seq and an nCounter Digital Analyzer for 39 differentially expressed genes.

Spearman rank correlation (Morey *et al.* 2006) of the $\log_2$ differences between RNA-seq reads (FPKM) and nCounter Digital Analyzer measurements for the 39 genes that were found to be differentially expressed in three potato cultivars (Shepody, Russet Burbank and Atlantic). Two groups are formed: the group in the bottom left represents the measurements that correspond to the 9 genes found to be under-expressed in plants with N supplementation; the group in the top right corresponds to the 30 genes found to be over-expressed in plants with N supplementation.

The N responsive genes were annotated using Gene Ontology (GO) terms and KEGG pathway information to determine if they have known functions in common. The GO information was obtained from the collected results of five GO term analysis pipelines (Trinotate (HMM and BLAST), OrthoMCL-UniProt, BLAST2GO, Phytozome and InterPro2GO) on ITAG1.0 genes (Amar *et al.* 2014). GO terms were found for all but six of the N responsive genes. The KEGG pathway information and Enzyme Commission (E.C.) numbers of the genes were obtained from the SolCyc database in the Sol Genomics Network (Fernandez-Pozo *et al.* 2014). Out of the ten genes with associated E.C. numbers, seven were part of well-defined metabolic pathways (**Tables**

3.3 and **3.4**): *Alanine-glyoxylate aminotransferase* (Sotub10g018540), *Glutamate decarboxylase* (Sotub01g005580), *Glutathione S-transferase* (Sotub10g024560), *Sulfate adenylyltransferase* (Sotub09g024290), *Proline dehydrogenase* (Sotub02g033320), *Primary amine oxidase* (Sotub08g025870), and *Ubiquinone/menaquinone biosynthesis methyltransferase* (Sotub12g020880).

Most of the KEGG pathways associated with the differentially expressed genes are involved in amino acid metabolism. In addition, one gene is directly associated with sulphate reduction pathways (*Sulfate adenylyltransferase* Sotub09g024290), as well as three other sulphate-related genes (*Sulfate transporter* Sotub04g021910; *High affinity sulfate transporter 2* Sotub04g027100; *High affinity sulfate transporter 2* Sotub10g013960) that have been previously reported to have an N response in *Arabidopsis* and tobacco (Reuveny *et al.* 1980; Bao *et al.* 2011). The Gene Ontology analysis revealed a very wide variety of GO-terms associated with the differentially expressed genes. The most commonly found GO term was related to an integral component of the cell membrane (GO: 0016021). Other GO terms associated with the 39 differentially expressed genes included response to cadmium ion (GO:0046686), oxidation-reduction process (GO:0055114) and transmembrane transport (GO:0055085).

3.2.3 Overrepresented Sequence Motifs Are Present in the Upstream Flanking Regions of N Responsive Genes

To predict potential N responsive regulatory mechanisms in the 39 differentially expressed genes, the 1000 bp upstream flanking regions were analysed for putative regulatory DNA motifs. *De novo* motif discovery tools Seeder (Fauteux *et al.* 2008), Weeder (Zambelli *et al.* 2014) and MEME (Bailey *et al.* 2009) were used to identify overrepresented motifs in the upstream flanking region of the differentially expressed genes. Each program works by implementing a completely different motif prediction algorithm, and analysing the sequences using all three tools increases the probability of finding more overrepresented motifs.

To discover overrepresented motifs, both Weeder and Seeder require a background file that contains information on k-mer frequencies in the relevant region or regions of the genome. Because motif discovery was focused on the upstream flanking region of N responsive genes, the reference background was computed from the sequences of the upstream flanking regions of all the genes in the Potato Reference Genome (The Potato Genome Sequencing Consortium 2011).

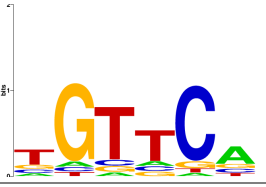
These sequences were obtained using the Transcription Start Site (TSS) and strand information for all the genes found in the ITAG1.0 annotation system (The Tomato Genome Consortium 2012; Fernandez-Pozo *et al.* 2014).

Seeder and MEME assume that overrepresented motifs will appear in all or most of the input sequences. This means that if one or several sequences do not contain an overrepresented motif, it might not be discovered. By carrying out multiple runs of motif discovery separately in smaller random subsets, there is an increased probability of finding motifs that are not present in all the sequences of the original list. Therefore, the two lists of upstream flanking regions of N responsive genes were sub-divided into smaller random sub-groups of 10 sequences and these smaller subsets were used as input for MEME and Seeder. Due to the nature of its algorithm, it is not necessary to use random sub-groups with Weeder and therefore a single list of upstream flanking regions was used as input for this program.

It is common for motif discovery programs to predict motifs that are very similar to each other, making it difficult to distinguish redundant and unique hits. This problem is further complicated when using random sub-groups as input, because the same motif might be present in several different sub-groups and will therefore appear multiple times in the final output. To overcome this issue, predicted motifs were clustered using the k-medoids clustering algorithm found in the TAMO package (Gordon *et al.* 2005) and the average of every cluster was then used as a representation of that cluster. The final motifs discovered in both lists of differentially expressed genes, after clustering, are summarized in **Table 3.5.**

Table 3.5: Nine regulatory motifs discovered in the upstream flanking region of N responsive genes.

	Weblogo	Algorithm	PLACE°	JASPAR$
1		MEME	OCTAMERMOTIFTAH3H4 1.6531e-07 CGCGCATCMG CGCGGATC-- histone; Oct; S-phase; CaMV 35S; NOS; meristem;	MA0069.1_Pax6 4.5480e-05 --CKGATGCGCG-- MANTSAWGCGTGAA
2		Seeder	SITEIOSPCNA 1.5356e-06 TGMACCTGGA -CCACCTGG- PCNA; Site I; G-box; meristem;	MA0086.1_sna 5.4427e-06 TGMACCTGGA --CACCTG--
3		Weeder	SURE2STPAT21 4.3508e-05 -ATACTC-- AATACTAAT SURE; SURE 2; patatin; sucrose; tuber; root;	MA0124.1_NKX3-1 7.2137e-06 -GAGTAT TAAGTAT
4		Seeder	OCTAMOTIF2 ATGCGG--- ATGCCGCGG octamer; histone; meristem; 1.8999e-04	MA0242.1_run_Bgb ATGCGG--- TTGCGGTTW 4.8665e-05
5		Seeder	3AF1BOXPSRBCS3 9.8670e-06 ANATAGAC---------- AAATAGATAAATAAAAACATT 3AF1 box; promoter; AT-rich sequences; GATA; rbcs; rbcs-3; leaf; shoot;	MA0011.1_br_Z2 1.0048e-06 GTCTATNT WNCTATTT
6		Seeder	TBOXATGAPB 1.2105e-05 CTAAGT CAAAGT GAPB; glyceraldehyde-3-phosphate dehydrogenase; light-activated transcription;	MA0211.1_bap 6.2836e-06 CTAAGT- TTAAGTG

	Weblogo	Algorithm	PLACE°	JASPAR[$]
7		Seeder	CATATGGMSAUR 1.3245e-05 CATAGG CATATG SAUR; NDE; auxin;	MA0423.1_YER130C 1.0885e-04 CATAGG--- NAWAGGGGN
8		Weeder	SB3NPABC1 2.7174e-07 ------TGTTCA--- AATTACTGTTCATAA sclareol; ABC; transporter; SB3;	MA0261.1_lin-14 7.0364e-06 TGAACA- -GAACRN
9*		Seeder	S1FBOXSORPS1L21 8.9090e-07 TACCAC TACCAT S1F; S1F box; S1F-box; S1; plastid protein; RPS1; RPL21; leaf; negative;	MA0002.1_RUNX1 1.0012e-04 -TACCAC---- NAACCACARWW

* Motif discovered in the upstream regions of under-expressed genes.

° Best match in PLACE database (Higo *et al.* 1999): motif accession code in **bold**; E-value of alignment (underlined); alignment (top: predicted motif, bottom: motif found in database); keywords associated with the motif (in *italics*).

[$] Best match in JASPAR database (Mathelier *et al.* 2014): motif accession code in **bold**; E-value of alignment (underlined); alignment (top: predicted motif, bottom: motif found in database).

Two databases of experimentally validated DNA-motifs, JASPAR (Mathelier *et al.* 2014) and PLACE (Higo *et al.* 1999), were consulted to determine whether any of the predicted motifs had previously reported functions that might have any relationship with N response. All discovered motifs returned at least one significant result from each of these databases, however none of those results matched the computationally discovered motif identically. Most of the results found in the databases differed with the discovered motifs by one nucleotide in the aligned region.

The results obtained from searching for motifs in the PLACE database can be used to predict regulatory mechanisms in potato because this database aggregates only experimentally validated motifs in plants. The matches from PLACE had several associated biological functions including the regulation of histone, auxin, and amylase genes (**Table 3.5**). A regulatory motif similar to one discovered by Weeder (Motif 3) has been previously shown to regulate patatin production in potato tuber, possibly through modulation with exogenous sucrose (Grierson *et al.* 1994). Also, three motifs discovered by Seeder (Motifs 5, 6, and 9) matched with PLACE motifs that have been associated with light-responsive promoters: 3AF1 (Gilmartin *et al.* 1990), S1F (Villain *et al.* 1994) and GAPB (Chan *et al.* 2001).

Because the NRE motif was described in previous studies (Konishi & Yanagisawa 2011) but was not found with our motif discovery process, an attempt was made to find instances of this motif in the upstream regions of our 39 N responsive genes. None were found. The NRE motif was identified in the promoters of *NIR* genes in plants, however no *NIR* genes were identified as N responsive in the current study, because developmental stage and time of day variable genes were removed. The upstream flanking regions of three potato *NIR* genes (Sotub01g045870, Sotub08g028180, and Sotub10g014930) were therefore inspected and the NRE motif was indeed found in two of those regions. This indicates that the NRE motif is conserved in the upstream regions of the *NIR* genes in potato and therefore possibly involved in N response specific to developmental stage or time of day. However, there is no evidence in the current study to suggest that it regulates the steady state N responsive genes identified here.

Every discovered motif was mapped back to all upstream flanking regions in the N responsive genes. This served two purposes: it enabled the identification of additional redundant motifs, which mostly mapped to the same positions within the flanking regions, and it also permitted the comparison of upstream flanking regions from genes with similar reported functions.

Figure 3.3 contains diagrams showing the position of all motif instances in the 5'-upstream flanking regions of all N responsive genes. In general, the motifs that appear the most times in the upstream-flanking regions of all N responsive genes are Motif 4 [ATGCGG] (62 times), Motif 5 [A.ATrGAC] (56 times) and Motif 7 [CATAGG] (40 times).

3.3 Discussion

The effect of N on the growth and yield of potato has been the subject of several studies, including those focused on the phenotypical effects of N-deficiency (Zebarth *et al.* 2003, 2004) and more recently, those that have analysed the effects of N on the expression of a small subset of genes (Li *et al.* 2010; Luo *et al.* 2011; Zebarth *et al.* 2011). However, a modern tool like RNA-seq offers a new way to compare the whole transcriptome of potato plants grown under differing N status. Our experiment is one of the first to use RNA-seq to find differentially expressed genes in plants grown in the field, with and without N supplementation.

The current study found that the three potato cultivars examined responded in the same way to supplemented N. The results were increased tuber yield, significantly greater N uptake and dry biomass, as well as greater leaf chlorophyll content. Gene expression analysis identified genes responding to N sufficiency similarly across three cultivars and two developmental time points. Previous studies have shown that individual genes can have different responses to N supplementation depending on the time of day (Tai & Zebarth 2015), therefore the current study has removed genes with time of day variation leaving only genes that consistently show steady-state differential expression in response to N. The final set of 39 identified genes were those that showed a consistent response to N supplementation in all three cultivars throughout the day, both at eight and ten weeks after planting.

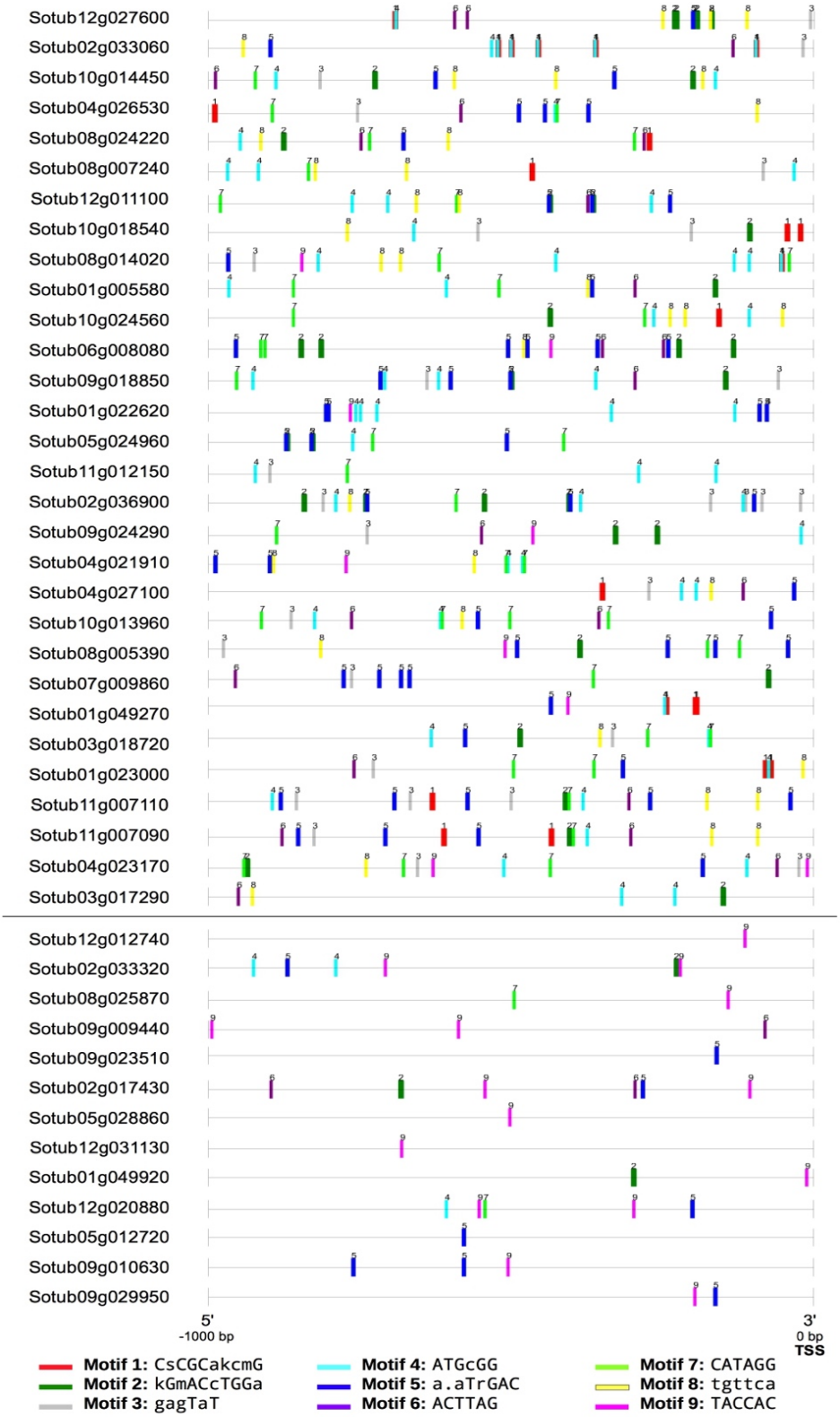


Figure 3.3: Motif locations in the 5'-upstream flanking region of N responsive genes.

Diagrams representing the 1000 nt 5'-upstream region of the 30 over-expressed (top section) and 13 under-expressed (bottom section) N responsive genes. Coloured rectangles indicate an instance of a discovered motif in that position; one-letter representations of every motif are found in the diagram key at the bottom. The Transcription Start Site (TSS) of every gene is located at the right end of each upstream flanking region.

Our previous studies examined expression of potato genes involved in N uptake, assimilation and transport, and demonstrated that an ammonium transporter gene, *AMT1*, was responsive to N rates over different developmental time points (Zebarth *et al.* 2011). However, this gene was found to have variation in expression at different times of the day (Tai & Zebarth 2015) and therefore did not meet criteria for screening in the current study. Functional analysis of the differentially expressed genes in the current study indicated association with KEGG pathways involved in amino acid metabolism. Two over-expressed genes included the aminotransaminases *Aminotransferase-like protein* (Sotub12g011100) and *Alanine-glyoxylate aminotransferase* (Sotub10g018540). Additionally, *Proline dehydrogenase* (Sotub02g033320) was found to be under-expressed with N supplementation. Decreased proline dehydrogenase activity was also found under conditions of N supply in French bean (Sánchez *et al.* 2002), which concurs with the results of this study. The action of proline dehydrogenase and pyrroline-5-carboxylate dehydrogenase lead to degradation of proline to glutamate.

There is also evidence for regulation of C:N balance with N supplementation. Over-expressed genes encoded enzymes functioning in both C and N metabolic pathways including two enzymes in the GABA shunt: *Aminotransferase-like protein* (Sotub12g01110) and *Glutamate decarboxylase* (Sotub01g005580). The GABA shunt is involved in the regulation of C:N balance in plants. GABA may also have roles related to stress response and as a signalling molecule (Michaeli & Fromm 2015). Another gene involved in both C and N metabolism was *Alanine-glyoxylate aminotransferase* (Sotub10g018540), which converts alanine and glyoxylate to glycine and pyruvate. This enzyme is involved in photorespiration in *Arabidopsis* (Liepman & Olsen 2001) and high levels of photorespiration are associated with low alanine and high glycine (Novitskaya *et al.* 2002). Interestingly, photorespiration is also linked to increased nitrate assimilation (Bloom 2015).

Additionally, four sulfate-related genes were found to be N responsive, one of which was part of the sulfate reduction and sulfate activation pathways, indicating a potential relationship between the sulfate and nitrate metabolic pathways. This type of relationship has been observed before in tobacco where N availability has been shown to regulate ATP sulfurylase (Reuveny *et al.* 1980) and, more recently, in *Arabidopsis* where sulfur transporter *SULTR1;1* is found to be down-regulated in conditions of N insufficiency (Bao *et al.* 2011).

GO-terms associated with N responsive genes were found to be related to transmembrane transport. These genes were involved in transport of sulfate, nitrate, amino acids and peptides; which correspond with the differential expression observed in amino acid and sulfate metabolism genes. Under-expression of a cation transporter and over-expression of a cation transport regulator were also found. These results indicate that proton movement may be involved in responses to N supplementation. Cation transport can affect the activity of the GABA shunt enzyme Glutamate decarboxylase, which is controlled by pH and Ca^{2+}-calmodulin (Shelp *et al.* 1999).

The Flowering locus T protein gene (Sotub05g028860) in potato plays a role in controlling maturity and tuberization (Navarro *et al.* 2011). This gene was under-expressed in plants with supplemented N, suggesting that increased N sufficiency can delay maturity. Two genes that were also under-expressed are similar to *Arabidopsis* genes involved in N response: a nodule inception protein similar to the *Arabidopsis* NIN-like transcription factor (Sotub01g049920) and Nodulin (Sotub05g012720). The former is a nitrate responsive transcription factor (Konishi & Yanagisawa 2013) and the latter participates in nodule formation in legumes, and in amino acid transport in non-leguminous plants (Denancé *et al.* 2014).

The N responsive genes were shown to have shared motifs in the upstream promoter regions, which supports the idea of coordinated regulation at the transcriptional level of genes responding to N supplementation. Motif discovery was carried out based on a previously described strategy that sub-divides genes into random subgroups to increase the probability of finding overrepresented motifs that are not present in all flanking regions (Munusamy, *et al.*, unpublished). Initial results of the motif discovery algorithms produced many redundant motifs, which was expected because different sub-groups may contain instances of the same motif. Therefore, a k-medoids clustering strategy was used to reduce the incidence of redundant motifs. In the end, the average motif of each cluster was taken as the representative motif and used in the subsequent annotation and mapping analyses.

Annotation of discovered motifs using experimentally validated motif databases, especially PLACE, revealed several intriguing putative regulatory mechanisms. Three of the overrepresented motifs in the upstream regions of N-responsive genes, including Motif 5 which was one of the motifs most frequently found in this dataset, were similar to motifs previously found to be associated with light-responsive genes (Gilmartin *et al.* 1990). Additionally, Motif 3

[GAGTAT/ATACTC] is very similar to a previously reported motif [AATACTAAT] involved in the regulation of patatin production in tubers as a response to exogenous sucrose (Grierson *et al.* 1994). These results are very similar to a previous study focusing on the regulation of patatin genes in potato (Aminedi & Das 2014), which further suggests that C:N balance regulation is involved in the response to N supplementation.

The motif discovered by Seeder in the 5´-flanking region of under-expressed genes (Motif 9 TACCAC) is very similar to the binding site of transcription factor S1F [TACCAT], a *cis*-regulatory element associated in the down-regulation of plastid related genes such as *rbc*S, *cab*, and *rp1*21.The introduction of this binding site into transgenic tobacco plants has been experimentally shown to cause the differential repression of the *rps*1 plastid gene in non-photosynthetic tissue (Villain *et al.* 1994). The similarity of the predicted motif and the experimentally validated one suggests that lower concentrations of available N could be potentially triggering a repression of plastids in the foliar tissue.

Mapping of predicted motifs in the upstream flanking regions of N responsive genes revealed few obvious patterns in the incidence of motifs with relation to the Transcription Start Site (TSS). Motifs seem to have no defined position within the upstream flank, and there are several instances of motifs appearing multiple times within the same region, which is not uncommon in other organisms. The upstream flanking regions of two over-expressed genes (Sotub11g007110 and Sotub11g007090) stand out due to their similarity, both in number and location of predicted motifs. Interestingly, both genes have the exact same annotation in the ITAG1.0 system (protein of unknown function) and are located close to each other on the same chromosome. The similarity of the positions of the discovered motifs in the upstream-flanking regions of these genes indicates a similar molecular mechanism regulates the expression of both, and also suggests that they share very similar, if not identical function. By analysing these genes and searching for a similar distribution of motifs in the upstream flanking region of at least one other gene, it might be possible to finally identify the function of these unknown proteins as well as predict their relationship with N supply in potato.

Upstream regulatory motifs in N responsive genes have previously been found in other species, such as maize (Liseron-Monfils *et al.* 2013). These genes included nitrate, nitrite and ammonium transporters; nitrate and nitrite reductases; as well as glutamate and glutamine

synthases. Nitrate transporter was the only gene in common between the maize study, which identified genes expressed at 4 h after N treatment, and the present study, which identified genes expressed at steady state nitrogen supplementation. Therefore, it is not surprising that the putative regulatory motifs found in our study are different than those in the study in maize.

There was no observed similarity between the motifs found in the upstream regions of the 39 N responsive genes in potato and the NRE motif found in the upstream region of the nitrite reductase gene *NIR1* of *Arabidopsis* and several other plants (Konishi & Yanagisawa 2011). The *NIR* genes were not among the 39 genes that were differentially expressed in the current study. Previously, the authors have shown that a potato *NIR* gene had time of day (Tai & Zebarth 2015) and developmental stage (Zebarth *et al.* 2011) variations and this is likely why it is not among the 39 genes. The current study focused on genes without time of day variation, and of longer term responses to N supplementation, with gene expression measurements at eight and ten weeks after planting. Additionally, because the experiment was done in the field, plants that did not receive N supplementation were not completely starved of N. The results demonstrate that longer term, steady state responses to N involve a different set of genes than shorter term, time of day variable responses, hence, regulatory motifs are also different. However, the presence of NRE in the upstream region of *NIR* genes in several different plant species (including potato) raises the possibility that the putative motifs found in this study could also be present in the N responsive genes of other plants.

Finally, epigenetic studies carried out in potato have been mostly focused on methylation of certain genomic regions and have shown that these are also affected by cell culture techniques, which has limited the ability to consider any potential effects they may have on gene expression (Law & Suttle 2005; Dann & Wilson 2011). In particular, the chromatin state of N responsive genes and any epigenetic changes in response to N supplementation are currently unknown, which is why they could not be considered for this study. Recently it has been shown that the modification of the upstream flanking regions of potato genes using RNA-directed DNA methylation induces heritable transcriptional gene silencing (Kasai *et al.* 2016), highlighting the importance of further analysing the impact of epigenetics in future studies of potato gene regulation.

In conclusion, our study provides evidence for regulatory coordination of steady state responses to N sufficiency at the level of gene expression in potato. These results have many potential applications including development of N status monitoring systems.

3.4 Materials and Methods

3.4.1 Plant Materials and Growth Conditions

Potato plants were propagated in the field at the Fredericton Research and Development Centre of Agriculture and Agri-Food Canada, Fredericton NB, Canada in 2012. The experiment included two fertilizer N rates (0 and 180 kg N ha^{-1}) in a randomized complete block design with four blocks. Fertilizer N was banded at planting as ammonium nitrate (34-0-0). All plots also received 150 kg ha^{-1} of P_2O_5 and K_2O banded at planting. Plots were six rows (5.46 m) by 8 m in size where the outer rows were guard rows.

The experiment was planted on May 23 using 0.91 m row spacing and 0.3 m within-row spacing. A modified planter was used to band the fertilizer treatments and open the rows. Hand-cut 50 g seed-pieces each of cultivar: Atlantic (U.S. Department of Agriculture, 1978), Russet Burbank (L. Burbank, approx. 1880) and Shepody (Agriculture Canada-New Brunswick, 1969) were hand-planted and imidacloprid was applied to control for Colorado potato beetle. The seed-pieces were covered using discs. One hill of cv. Chieftain was planted at the end of each row to avoid edge effects.

Sampling was done on July 25 and August 8 at four different time points (0800 h, 1100 h, 1400h and 1700 h). At each sampling date and time point, one plot of each treatment and cultivar was sampled by taking the apical leaflet of the last fully expanded leaf (usually the fourth leaf from the top of the plant) of 15 randomly selected plants and pooling them into a single sample in a 50 ml Falcon tube. In other words, the tissue sample collected for every block consisted of the pooled tissue of 15 random individuals in that plot (see **Table 3.1**). The tubes were then immediately placed in liquid N, and stored at -80 ˚C until RNA extraction. Petioles were then collected from the same leaf for determination of petiole nitrate concentration.

3.4.2 Leaf Chlorophyll Measurement and Petiole Nitrate Determination

Leaf chlorophyll index (LCI-S) was measured on the apical leaflet of the last fully expanded leaf, which was also used to measure gene expression, using a SPAD-502 reader (Konika Minolta). The

LCI-S was determined in the section of the leaf midway from the mid-rib to the leaf margin (Zebarth *et al.* 2003). Petioles were dried at 55 ˚C and ground to pass a 2 mm screen. A 0.2 g subsample of petiole tissue was extracted with 40 ml distilled water and a 15 min shaking time. The concentration of NO_3-N in the extract was determined colorimetrically using a Quikchem 8500 flow injection analyzer (Lachet) using QuikChem method 90-107-04-2-A (Zebarth *et al.* 2003). Two-factor ANOVA of petiole nitrate determination and SPAD readings between the groups was calculated using the R statistical language v. 3.1.1.

3.4.3 Dry Biomass and Fresh Yield Determination

Whole plants were sampled before vine desiccation to determine their dry biomass content and at harvest to measure fresh tuber yield. Each plant was separated into vines, tubers and stolons as well as any recoverable roots. Tubers with a diameter below 0.5 cm were left as part of the stolons and roots. Vines, stolons and roots were washed, weighed and then oven-dried. After drying, they were weighed again and the dry matter of each sample was determined. Tubers were washed and weighed to determine fresh yield. Finally, tuber samples were taken from every experimental group and quartered along the long axis. One quarter of every tuber was randomly selected and sliced into 1 × 1 cm strips, and then weighed before and after oven-drying to determine dry matter content (Zebarth & Milburn 2003). Two-factor ANOVA of dry biomass and fresh tuber yield between the groups was calculated using the R statistical language v. 3.1.1.

3.4.4 RNA Extraction

Leaf tissue was ground to a powder in liquid N using a mortar and pestle. Samples were pre-extracted in 1 ml Hot Borate Buffer (Luo *et al.* 2011). The lysate supernatant (200 µl) was used for RNA extraction with the Biomek NXP Laboratory Automation Workstation (Beckman Coulter) using the RNAdvance Tissue Kit (Agencourt) according to the instructions from the manufacturer for liquid samples. The RNA concentration and quality were determined using a NanoDrop 1000 Spectrophotometer (Thermo Scientific) and 2100 Bioanalyzer (Aglient Technologies) respectively.

3.4.5 Library Preparation and Sequencing

Libraries were generated using the TruSeq RNA kit (Illumina). Messenger RNA was purified from 1 µg of total RNA using oligo-dT beads. The mRNA enriched fraction was reverse transcribed to generate cDNA fragments that were sheared to yield ~200 bp fragments. Following end-repair, 3’

end adenylation steps, and index ligation, a PCR amplification step was performed. A separate index was used for each N treatment-variety-replicate combination for a total of 24 indices.

Two lanes of sequencing were done for each 24 index multiplex, one for each time point. The quality of the library was assessed on a DNA 1000 chip and quantified by qPCR. Libraries were subjected to 100 bases of sequencing on a HiSeq 2000 (Illumina) instrument in paired-end mode. Initial quality control of the data was performed using the software included with the sequencer.

3.4.6 Genome Alignment and Differential Expression Analysis

Output from the sequencer was aligned to the *S. tuberosum* reference genome v3_2.1.10 (The Potato Genome Sequencing Consortium 2011) using the TopHat software suite v. 2.0.9 (Kim *et al.* 2013) with the mode for "fr-unstranded" library types. The quality of the alignments was verified using the 'flagstat' tool from the SAMtools software suite v. 0.1.19 (Li, Handsaker, *et al.* 2009).

Reads were assembled into transcripts with CuffLinks v. 2.1.1 (Trapnell *et al.* 2012), using the *S. tuberosum* ITAG1.0 annotation file obtained from the *Sol Genomics Network* (The Tomato Genome Consortium 2012; Fernandez-Pozo *et al.* 2014). Transcriptome assembly was performed with the 'multi read correct' and 'fragment bias correct' modes activated. Finally, assembled transcripts from different replicates and treatments were merged into a single reference transcriptome for each variety using the 'CuffMerge' tool included in CuffLinks.

Differentially expressed genes were identified for each time-point and each cultivar using CuffDiff (Trapnell *et al.* 2013). The same *S. tuberosum* reference genome as well as the single, merged transcriptome were used as reference for differential gene expression. Finally, genes found to be differentially expressed in each cultivar, were compared using custom perl scripts and a single list of over-expressed and under-expressed genes found in all cultivars at both time-points was produced.

3.4.7 Expression Analysis using nCounter Digital Analyzer

The 39 genes with significant differences in expression from the CuffDiff analysis were selected and used to validate the gene expression results. The same RNA samples indicated above were prepared using the reagents and method described in Geiss *et al.* 2008 for the nCounter (Nanostring Technologies) multiplex gene expression analysis. The nCounter data was adjusted according to

the manufacturer's instructions using the manufacturer-provided spiked positive and negative controls. Gene expression of five housekeeping genes 18S rRNA, actin, cyclophilin (Tai *et al.* 2009), elongation factor 1-α (EF-1-alpha) (Nakane *et al.* 2003) and cox1-B (Li *et al.* 2010) was also measured and the geometric mean of their expression was used to normalize gene expression values for the 39 test genes (Luo *et al.* 2011; de Almeida *et al.* 2015). Spearman rank correlation was performed using SYSTAT v. 13 (Systat Software) and used to compare expression data from nCounter and transcriptome sequencing for the 39 genes.

3.4.8 De novo Motif Discovery

A FASTA file containing the 1000 nt upstream flanking regions upstream of the transcriptional start sites for all the genes in the *S. tuberosum* reference genome v3_2.1.10 (The Potato Genome Sequencing Consortium 2011) was generated using the 'faidx' tool of the SAMtools software suite v. 0.1.19 (Li, Handsaker, *et al.* 2009). The transcription start-site and strand information for every gene were obtained from the ITAG1.0 annotation file (The Tomato Genome Consortium 2012; Fernandez-Pozo *et al.* 2014). From this general file, two subsets were created, each containing only the upstream flanking regions of the genes that were found to be significantly over-expressed or under-expressed in response to supplemented N. Motif discovery was performed separately for each set of differentially expressed genes.

Three different programs were used for *de novo* motif discovery: Seeder v. 0.01 (Fauteux *et al.* 2008), MEME v. 4.10.0 (Bailey *et al.* 2009) and Weeder v. 2.0 (Zambelli *et al.* 2014). All three motif discovery programs were run simultaneously on *Guillimin*, McGill University's high-performance computing server (http://www.hpc.mcgill.ca/), using high-memory computing nodes. A series of 5000 random subsets containing 10 random promoters each were generated for each differentially expressed gene list these random subsets were then used as input for the motif discovery programs Seeder and MEME.

The FASTA file containing all 1000 nt upstream flanking regions in the genome was used to generate the background files in Seeder and Weeder. All motif discovery programs were run to find motifs with a minimum length of 6 nt in both the forward and reverse-complement strands.

Significance of predicted motifs was determined differently for each algorithm, based on the available parameters reported by the program: in Seeder, a maximum q-value of 0.05 was

allowed; in MEME, a maximum e-value of 0.001 was allowed; finally, only the top five results of each Weeder run were considered significant.

3.4.9 Motif Annotation and Mapping

All significant motifs were converted into the same format for comparison and annotation using the TAMO software suite (Gordon *et al.* 2005). Redundant motifs were clustered together using k-medoids algorithm, as implemented in TAMO. Cluster averages were uploaded to STAMP (Mahony & Benos 2007) for visualization and to search in the PLACE (Higo *et al.* 1999) and JASPAR (Mathelier *et al.* 2014) databases for potential matches.

Motif cluster averages were mapped to the promoters of differentially expressed genes using the 'Sitemap' tool provided in TAMO. The same approach was used to map the previously reported NRE motif. Mapping results were visualized using the 'GenomeDiagram' tool included in BioPython v.1.61 (Cock *et al.* 2009). Diagrams for visualization of nucleotide frequencies in motifs were all created using Weblogo v.2.8 (Crooks *et al.* 2004).

To facilitate the comparison of the promoters of genes with similar biological function, differentially expressed genes were annotated using GO terms based on the results obtained by Amar *et al.* (2014). Additional KEGG pathway information for differentially expressed genes, when available, was retrieved from the 'SolCyc' database for *S. tuberosum* in the Sol Genomics Network (Fernandez-Pozo *et al.* 2014).

3.4.10 Alternative Splicing Analysis

Individual RNA-seq reads were aligned to a BowTie v. 2.2.3 (Langmead & Salzberg 2012) index of the *S. tuberosum* reference genome (The Potato Genome Sequencing Consortium 2011) using TopHat v. 2.0.9 (Trapnell *et al.* 2009; Kim *et al.* 2013). Aligned reads were imported into an array in R v.3.1.1 (R Core Team 2015) using the DEXseq v. 1.12.2 package (Anders *et al.* 2012). In order to carry out differential exon usage analysis, two preparation steps were required: first, the generation of a modified GTF annotation file with no overlapping exons, and then the creation of a single counts-per-exon file for each RNA-seq sample. The GTF annotation file with no overlapped exons was created using the 'gffread' tool in Cufflinks 2.2.1 (Trapnell *et al.* 2010) to parse the original GFF3 annotation file, and then the HTSeq python library v. 0.6.1 (Anders *et al.* 2014) to remove overlapping exons. The counts-per-exon reads were calculated using the 'dexseq_count' tool included in the DEXseq package.

Preface and Author Contributions to Chapter 4

As shown in the previous chapter, RNA-seq data can be used to identify differentially expressed genes with a similar response in different potato cultivars. That information can then lead the prediction of putative gene regulatory mechanisms found in the upstream flanking region of those genes. These results are significant because practical applications derived from this data can be potentially used on plants of several different potato varieties. However, one limitation of this methodology is that it can be challenging to identify variety-specific regulatory mechanisms using this approach. Since *cis*-regulatory elements are found in the genome of an organism, and are therefore not directly detectable using RNA-seq data, to predict variety-specific regulatory mechanisms more information is required about the diversity and structural variations between different potato subspecies.

The rapidly decreasing costs of DNA sequencing have made it feasible to undertake large-scale re-sequencing experiments. These new datasets can be coupled with the publicly available potato reference genome and analyzed to provide useful insights on the main structural variations between the genomes of multiple potato varieties. Wild potato landraces in particular are of great interest for this purpose, because they have been previously shown to have traits of great interest to breeders (Pavek & Corsini 2001; Hirsch *et al.* 2013) and could therefore contain structural variants that have an impact on gene regulation and expression. Understanding these variants and developing reference genomes for wild species would therefore be a valuable resource for potato breeders and researchers.

José Héctor Gálvez wrote the main chapter text, with contributions from Martina V. Strömvik, Chen Yu Tang and Xinyi Zhu. Under the supervision of Martina V. Strömvik, José Héctor Gálvez designed and carried out the *de novo* genome assembly procedure. Under the supervision of Martina V. Strömvik, José Héctor Gálvez designed the CNV and novel sequence analysis, which was then carried out by Chen Yu Tang and Xinyi Zhu. Noelle Barkley and David Ellis provided raw genome re-sequencing data, and contributed to the overall analysis of these samples.

Computations were made on the supercomputer Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded

by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec-Nature et technologies (FRQ-NT).

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 Martina V. Strömvik, Noelle Barkley, David Ellis, and Helen H. Tai; the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. 283303) to Martina V. Strömvik; the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología; CONACYT) (Scholarship No. 381158) to José Héctor Gálvez; and the McGill Department of Plant Science Graduate Excellence Fund.

Chapter 4: Draft Genome Assembly and Copy Number Variation Analysis of Two Potato Landraces Reveal Significant Structural Variation

4.1 Introduction

Potato (*Solanum tuberosum* L.) is widely considered to be one of the most important staple crops worldwide. Because of its cultural and economic significance, there have been several efforts to develop high quality genetic resources to aid in the study and production of this crop and its related varieties (Hirsch *et al.* 2016). However, the high degree of heterozygosity, repetitive sequences and polyploidy complicate the development of resources for potato genomics.

Despite these challenges, there have been several important advances in the field of potato genomics in recent years. The most significant has been the *de novo* assembly of a potato reference genome (The Potato Genome Sequencing Consortium 2011). This was achieved through to the development of a unique doubled monoploid clone of the *S. tuberosum* group *Phureja* (DM1-3 516 R44, also referred to as DM) which was necessary in order to overcome the high heterozygosity of other potato mapping varieties (The Potato Genome Sequencing Consortium 2011). Subsequent studies have further improved the reference genome by anchoring the majority of contigs into chromosome-scale pseudomolecules (Sharma *et al.* 2013) and adding novel sequences (Hardigan *et al.* 2016). The availability of a reference genome has also enabled the development of numerous additional resources for potato, including a full reference transcriptome (Massa *et al.* 2011), functional and structural annotation systems (The Tomato Genome Consortium 2012), and several genetic markers on different platforms (Hirsch *et al.* 2016).

Potato landraces and wild species (*Solanum* sect. *Petota*) are an important source of genetic diversity and the basis of many breeding and improvement efforts (Ovchinnikova *et al.* 2011). However, there is a limited amount of information on the genomic differences between them and the potato reference genome. Recently, a draft genome for the potato wild species *Solanum commersonii* was assembled using deep paired-end and mate pair sequencing data (Aversano *et al.* 2015). A number of important differences between this wild species and the potato reference genome were uncovered; namely *S. commersonii* had significantly lower heterozygosity, repetitive sequences and genes than the reference. Additionally, the genome of *S. commersonii* has enabled

a thorough annotation of disease resistance and cold tolerance loci, highlighting important differences in this species that could explain some of its phenotypic characteristics (Aversano *et al.* 2015).

However, *S. commersonii* is currently the only potato wild species with a complete reference genome. That is not to say a reference genome is always necessary to discover genomic differences between potato varieties. Other potato genetic diversity studies have relied on other types of genomic resources such as fluorescence *in situ* hybridization (FISH) (Iovene *et al.* 2013), molecular markers (Gavrilenko *et al.* 2013; Hirsch *et al.* 2013; Hardigan *et al.* 2015), or genome re-sequencing data from landrace-derived monoploid potatoes (Hardigan *et al.* 2016). These tools have already enabled the development of a large number of well-annotated genetic markers, Single Nuncleotide Polymorphisms (SNPs) and structural differences such as Copy Number Variation (CNV ,Hirsch *et al.* 2016). However, large-scale re-arrangements and structural differences are still difficult to discover without new genome assemblies, especially when the heterozygosity found in diploid and polyploid genomes is taken into account (Chaisson *et al.* 2015; Pendleton *et al.* 2015).

In recent years, the decreasing costs of DNA sequencing has made it feasible to generate whole-genome sequencing data with high coverage and depth for many individual potato landraces and wild species. New computational methods have also made it possible to quickly identify and predict differences between these genomes, from SNPs and CNVs to large structural variations, through the assembly of a complete genome for each variety. This study is focused on the analysis and the *de novo* assembly of a draft genome of two potato landraces from the International Potato Center (CIP) germplasm bank: *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*.

4.2 Results

4.2.1 Genome Re-Sequencing Reveals Copy Number Variation in Potato Landraces

Genome re-sequencing reads were initially filtered for low-quality and adaptor sequences using Trimmomatic (Bolger *et al.* 2014) and then aligned to the potato reference genome (v4.03) using BowTie2 (Langmead & Salzberg 2012). A total of 183.0 and 129.7 million paired reads remained after filtering and trimming, for *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*, respectively. In the case of subsp. *andigena*, 84% of filtered reads aligned to the potato

reference genome, covering 86.4% of the reference genome with an average depth of 62.2×. Conversely, in subsp. *goniocalyx*, 87.8% of reads aligned to the potato reference genome and covered 85.5% of the reference genome with an average depth of 45.4×. A summary of the filtering and alignment statistics can be found in ***Supplementary table 4.1.***

The aligned reads for each genome were used separately as input for CNV prediction with CNVnator (Abyzov *et al.* 2011), a software tool that uses read depth to estimate the copy number of sequences throughout the genome. For the purposes of this analysis CNVs are defined as follows (Hardigan *et al.* 2016): duplications are sequences found in greater amounts in the genome of each individual landrace than in the reference genome; deletions are defined as the opposite (i.e. regions that are found in fewer or no instances in the genome each individual landrace). After filtering the raw CNV predictions ($P > 0.05$), a total of 13,425 deletions and 4,985 duplications were detected in *S. tuberosum* subsp. *andigena*, affecting a total of 7,423 genes throughout the genome. In *S. stenotomum* subsp. *goniocalyx,* 13,585 deletions and 3,401 duplications were identified, and they affected 6,040 genes in total. A summary of the CNVs detected in each landrace can be found in **Table 4.1**.

Table 4.1: Summary of CNVs (deletions and duplications) detected in *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx* using CNVnator.

	S. tuberosum* subsp. *andigena	***S. stenotomum* subsp. *goniocalyx***
Total CNVs	18,410	16,986
Total deletions	13,425	13,585
Total duplications	4,985	3,401
Genic CNVs (%)	26.9 %	24.2 %
Mean CNV length	11.8 kb	10.6 kb
Median CNV length	4.9 kb	4.3 kb
Median deletion length	3.8 kb	3.8 kb
Median duplication length	8.0 kb	6.2 kb
Total large CNVs*	142	120
Genes affected by deletions	4,490	4,490
Genes affected by duplications	3,062	1,695
Total CNV-affected genes$	7,423	6,040

* Large CNVs are defined as having a length > 100 kb.
$ This total includes genes that are affected by both deletions and duplications.

In both landraces CNVs were evenly distributed throughout the genome, with no significant increase in CNV incidence in regions with higher gene density. However, there were regions of the genome with higher numbers of CNV-affected genes in close proximity. In both

landraces, a region close to the 5' end of chromosome 12 contained a significant number of genes affected mainly by deletions (albeit some duplications as well in subsp. *andigena*). The genes in that region are related mainly to carbohydrate metabolic processes and many are code for leucine-rich proteins. However, that is the only instance in which a similar region of CNV-affected genes was found for both landraces.

In the case of *S. tuberosum* subsp. *andigena*, there were several other instances of regions with a high density of CNV-affected genes, including regions found in chromosomes 3, 4 and 7. However, the second largest cluster of CNV-affected genes was found towards the 3' end of chromosome 10, where 18 genes mostly associated with late embryogenesis were affected by CNVs. The second largest CNV-affected gene cluster in *S. stenotomum* subsp. *goniocalyx* is found near the 5' end of chromosome 4. This region of the genome contains many disease resistance genes; including several late blight resistance proteins, but not all the disease resistance genes in this cluster specify the disease they are associated with. A visual summary of the distribution of all CNVs in the genome as well as CNV-affected genes is found in **Figure 4.1**. A more detailed description of CNV-affected genes and their function is found in ***Supplementary Table 4.2***.

4.2.2 De novo Assembly of a Draft Genome Reveals Structural Differences in Potato Landraces

A draft genome for each potato landrace was assembled using the raw re-sequencing reads as input for the MaSuRCA *de novo* genome assembler (Zimin *et al.* 2013). Raw reads were used, as opposed to filtered and trimmed reads, because MaSuRCA performs its own filtering and trimming steps and requires raw reads to perform optimally. Each landrace genome was assembled separately and, in general, for every step of the process the software took twice as long to assemble the genome of *S. tuberosum* subsp. *andigena* than to assemble that of *S. stenotomum* subsp. *goniocalyx.*

A total of 724,895 scaffolds were assembled for subsp. *andigena*, and 54.5% of them had a length exceeding 1000 nt. Conversely, 241,818 scaffolds were assembled for subsp. *goniocalyx*, of which 62.5% had lengths exceeding 1000 nt. After assembly, the scaffolds were analyzed using QUAST (Gurevich *et al.* 2013), a tool developed to measure a number of quality metrics for genome assemblies. In every measured metric, the draft genome of subsp. *goniocalyx*

outperformed that of subsp. *andigena*. A brief summary of the obtained metrics for each draft genome is found in **Table 4.2**.

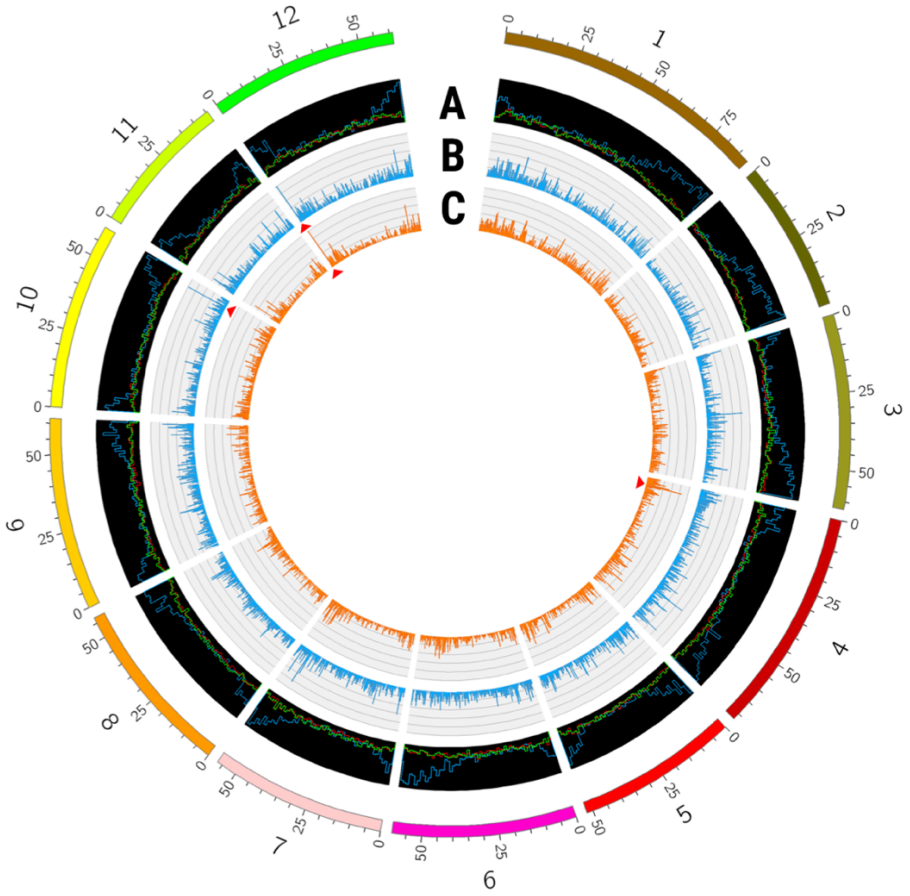


Figure 4.1: Chromosomal CNV distribution and CNV-enriched gene clusters in *S. stenotomum* subsp. *goniocalyx* and *S. tuberosum* subsp. *andigena*.

A) Gene distribution (blue) and CNVs (red: *goniocalyx*, green: *andigena*). **B)** Distribution of *andigena* copy number variable genes (200 kb step). **C)** Distribution of *goniocalyx* copy number variable genes (200 kb step). Red arrows point to the top two CNV-enriched clusters in each landrace. Generated using Circos (Krzywinski, et al., 2009).

Table 4.2: Quality metrics of *de novo* genome assemblies for *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*.

Quality Metric	***S. tuberosum* subsp. *andigena***	***S. stenotomum* subsp. *goniocalyx***
Total number of scaffolds	724,895	241,818
Total number of scaffolds ≥ 1000 nt	394,716	151,127
Total number of contigs	755,406	255,938
Total length of assembly *	1600 Mb	978 Mb
GC content **	35.25 %	35.22 %
Largest contig length	129,264 nt	250,740 nt
Scaffold N_{50}	3,789 nt	9,806 nt
Scaffold L_{50}	109,528	24,533

* Total Length of Reference Genome = 725 Mb ** GC content of Reference Genome = 34.75 %

After assessing the quality of both *de novo* assemblies, each draft genome was aligned to the potato reference genome v4.03 using NUCmer (Kurtz *et al.* 2004). In this step, once again, subsp. *andigena* took twice as long to process. The great majority of contigs from both assemblies had at least one significant match in the reference genome, either in the forward sense or in the reverse sense. Additionally, in terms of coverage, both assemblies nearly covered the whole reference genome with at least one matching contig. All of this is more clearly observed when visualizing the alignment results as a dot plot graph, as shown in **Figures 4.2** and **4.3**.

4.2.3 Potato Landraces Have a Significant Number of Putative Novel Sequences

After the initial filtering and alignment to the potato reference genome (v4.03), there were still a significant number of high-quality sequencing reads that did not align to the reference. These un-aligned reads amounted to approximately 60 million and 30 million in *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*, respectively. While these reads were included in the *de novo* draft genome assembly because raw sequencing results were used as input for MaSuRCA, the fact that un-aligned reads may consist of novel sequences not currently included in the latest versions of the potato reference genome warrants a separate analysis.

To facilitate the analysis of un-aligned reads, a separate set of contigs and scaffolds were created using only these reads as input for the ABySS *de novo* assembler (Simpson *et al.* 2009). The purpose of these new contigs and scaffolds was not to produce a full draft assembly, instead, they were generated to create longer sequences from the un-aligned reads in order to improve the chances of significant results when analyzing potential novel sequences. To that end, not much attention was placed on the length and quality of these contigs relative to those obtained in the full draft genome assembly, although it was noted that on average they were much shorter than the contigs assembled by MaSuRCA. A total of approximately 1.2 million and 427,000 contigs were produced for subsp. *andigena* and subsp. *goniocalyx* respectively, from their un-aligned reads.

During the course of these experiments, a newer version (v4.04) of the potato reference genome was released (Hardigan *et al.* 2016). The only difference between this updated version and the version of the potato reference genome used for the initial analysis (v4.03) was the inclusion of a separate, unanchored "chromosome" called 'chrUn'. This new addition consisted of 55.7 Mb novel sequences discovered in a recent re-sequencing study of a monoploid panel with 12 potato clones containing limited introgression mainly from other *Phureja* landraces (Hardigan

et al. 2016). However, the other 12 regular chromosomes in the reference remained unchanged. To account for these updates to the reference, a new alignment to 'chrUn' was made with the un-aligned sequences and as many as 21.3% of the subsp. *andigena* contigs and 29.28% of the subsp. *goniocalyx* contigs had a significant match.

The remaining contigs were then used to search the nucleotide database of NCBI (Wheeler *et al.* 2007; Benson *et al.* 2013) using the BLASTn aligner (Altschul *et al.* 1997). All contigs had at least one significant hit (e-value <0.0001) from the database, however the results came from many different taxa and species. In the case of subsp. *andigena*, the hits spanned a total of 1,183 taxa, of which 54.6% were eudicots, and 12.1% were from the Solanaceae family. Similarly, in subsp. *goniocalyx* a total of 501 taxa were represented, most of which were eudicots (72.8%), and 19.6% of the species were members of the Soalanceae family. In both cases, a few non-plant organisms were also detected in the results of the alignments, including 95 bacterial and 8 viral sequences in subsp. *andigena* and 18 bacterial and 3 viral sequences in subsp. *goniocalyx.*

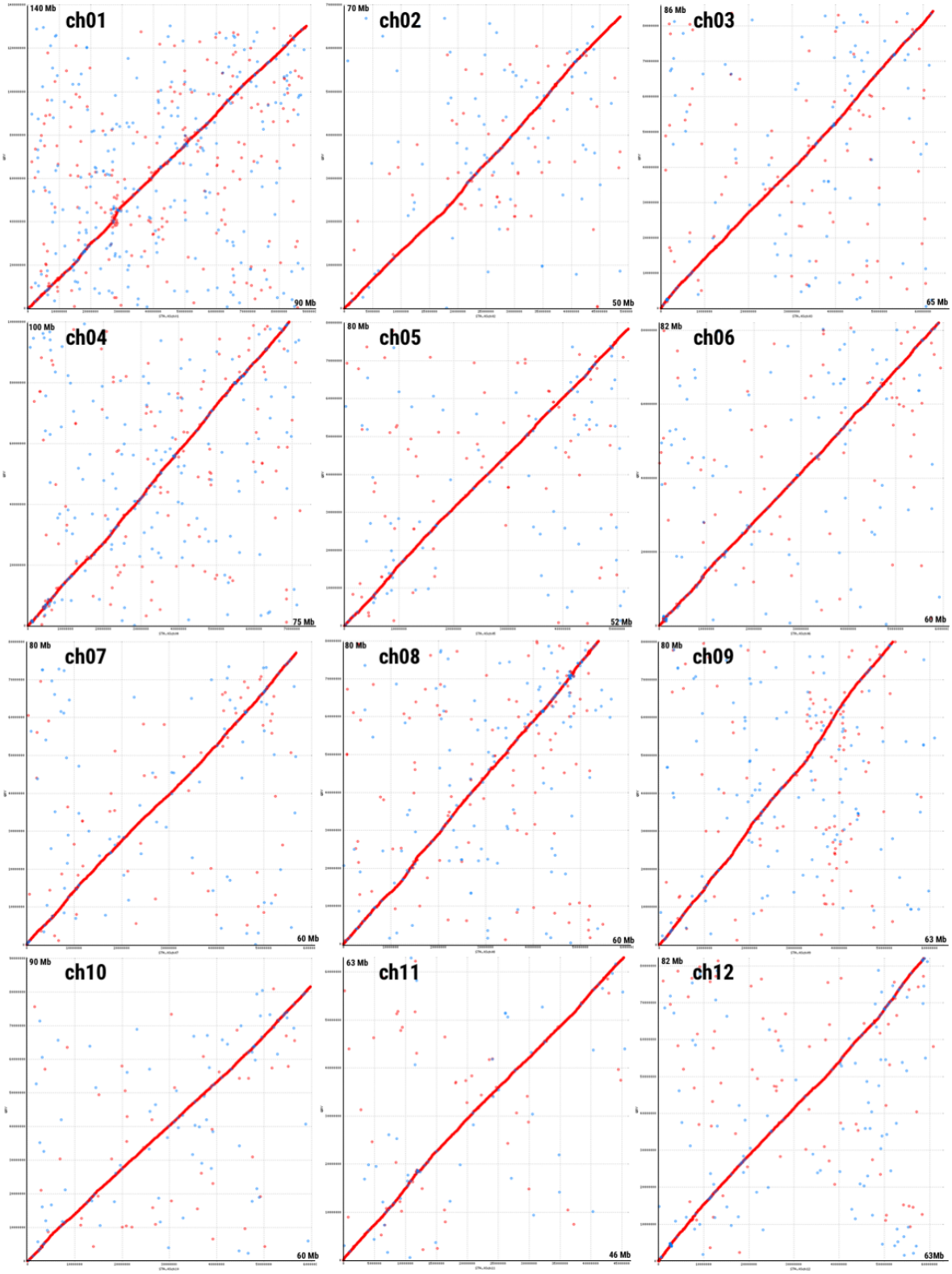


Figure 4.2: Dot plot of whole genome alignment between contigs and scaffolds assembled for *S. tuberosum* subsp. *andigena* and the potato reference genome v4.03.

Each individual plot corresponds to a single chromosome. The *x* axes represent the reference chromosome, the *y* axes the assembled contigs and scaffolds. Red lines and dots represent alignments in the forward sense, blue lines and dots represent reverse alignments. Only the top alignment for each contig and scaffold was plotted. Plots produced using the 'mummerplot' tool included in MUMmer (Kurtz *et al.* 2004).

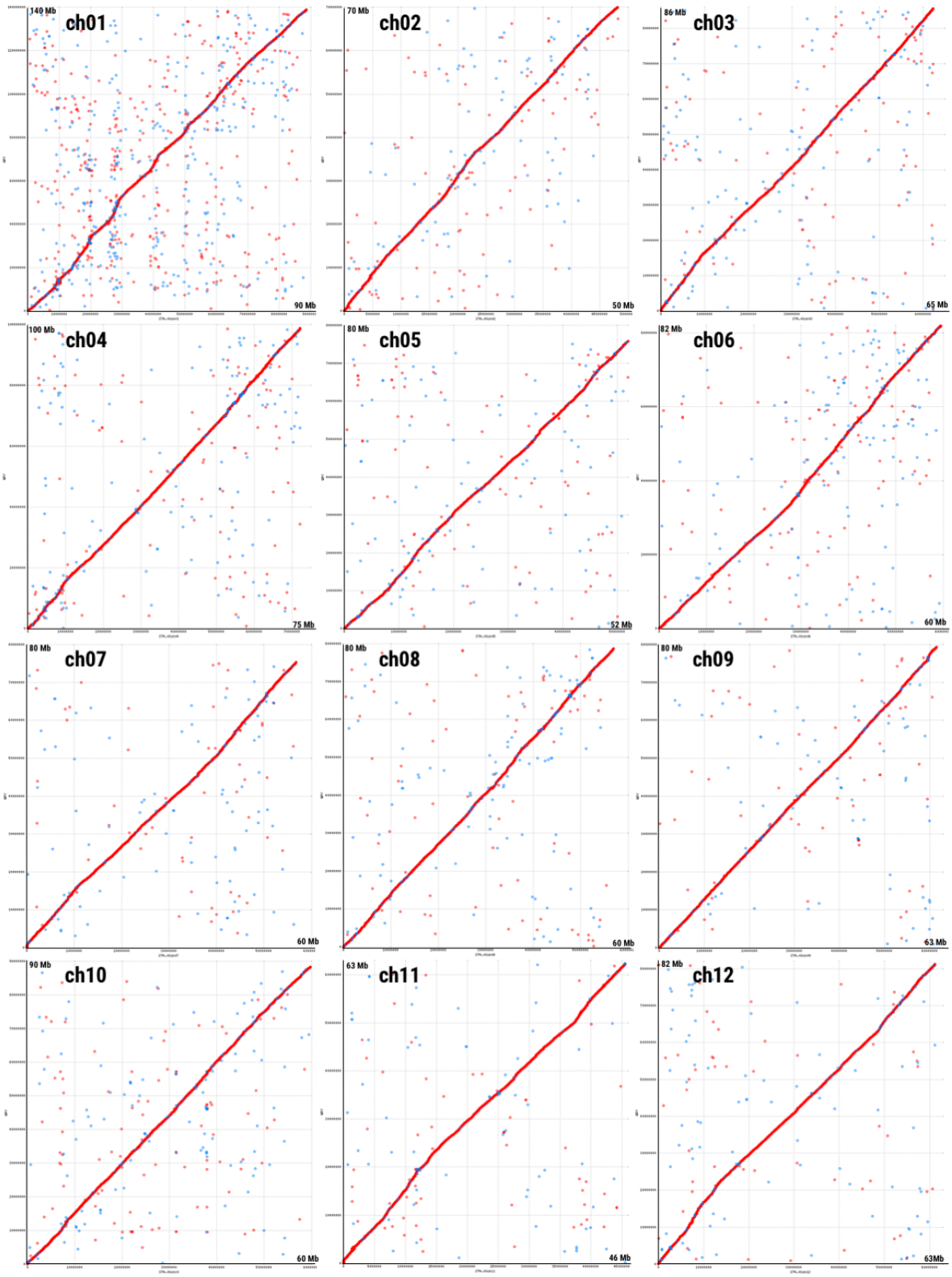


Figure 4.3: Dot plot of whole genome alignment between contigs and scaffolds assembled for *S. stenotomum* subsp. *goniocalyx* and the potato reference genome v4.03.

Each individual plot corresponds to a single chromosome. The *x* axes represent the reference chromosome, the *y* axes the assembled contigs and scaffolds. Red lines and dots represent alignments in the forward sense, blue lines and dots represent reverse alignments. Only the top alignment for each contig and scaffold was plotted. Plots produced using the 'mummerplot' tool included in MUMmer (Kurtz *et al.* 2004).

4.3. Discussion

Increases in potato yield over the last century have been mostly attributed to improvements in production practices and pest management, and not necessarily as a result of large-scale genetic improvement (Douches *et al.* 1996). While there have been several recent efforts in potato breeding (Hirsch *et al.* 2013), the relative lack of available high-quality genetic resources for potato has slowed progress. This began to change with the publication of the potato reference genome, however, more information is required on the genomic differences between potato landraces and the commercial varieties in order to successfully incorporate them into breeding programs.

Recent studies have already provided some evidence that there are important differences between the genomes of wild potato species and the reference genome. A FISH-based analysis of 16 potato cultivars from different countries has shown that CNVs are a major contributor to the genetic variation observed between these varieties (Iovene *et al.* 2013). Another study using genome re-sequencing on a panel of potato monoploids with introgression from *Phureja* landraces also found evidence of widespread CNVs (Hardigan *et al.* 2016). Our results agree with these previous findings and provide additional data on the particular cases of *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*, two Peruvian landraces that haven not been analyzed for large scale structural variants until now.

Using sequencing read depth as an estimate for copy number, CNVnator has predicted a considerable amount of structural variation between subsp. *andigena* and subsp. *goniocalyx*, when compared to the potato reference genome. CNVs appear to be evenly distributed throughout the genome, with roughly a quarter of CNVs affecting protein-coding genes in both landraces. A number of CNVs with a size exceeding 100 kb were also found in both landraces, most of them corresponding to large deletions, similar to what has been observed in other potato studies (Iovene *et al.* 2013; Hardigan *et al.* 2016). While the majority of large CNVs were deletions, the median size of duplications was actually higher in both subspecies, which again confirms previous findings and suggests a difference in the mechanisms that produce large and medium scale variants (Hardigan *et al.* 2016).

In both subspecies, the largest cluster of CNV-affected genes was found in the same region of chromosome 12 and consists of mostly deletions that affect carbohydrate metabolic genes, including several glycoside hydrolases. Several conserved genes with unknown function were also

affected in this gene cluster, as well as pathogenesis related transcription factors. A second cluster of CNV-affected genes found in chromosome 10 of subsp. *andigena* included mostly duplications affecting genes associated with late embryogenesis, but also a number of genes without any known function. Conversely, in subsp. *goniocalyx* the second largest cluster of CNV-affected genes was found in chromosome 4 and mainly consists of both deletions and duplications affecting disease resistance proteins. In the majority of cases, the annotation does not specify the disease the gene is associated with, however three of these genes are explicitly stated to provide resistance to late blight, but all three are affected by a deletion (see ***Suplementary Table 4.2***). Collectively, these CNV-affected gene clusters seem to indicate important differences between these two potato subspecies and the reference genome, and could guide future experiments that focus on exploring these differences to determine gene function in these regions as well as any potential phenotypic differences that may be a result of the detected deletions and duplications.

None of the N responsive genes identified in the previous chapter (see **Table 3.3** and **3.4**) were found to be affected by CNVs in either of these landraces, which is not surprising considering they were genes that had been explicitly filtered for a conserved response across several different cultivars. The fact that these genes are not affected by deletions or duplications provides additional evidence that they are indeed highly conserved across potato cultivars and landraces. However further evidence needs to be collected to fully support this claim, specifically re-sequencing data from every cultivar is required and then submitted to a CNV analysis.

While CNV prediction using bioinformatic tools such as CNVnator has provided interesting targets for further analysis, larger rearrangements are difficult to detect without assembling a full reference genome for each subspecies. The benefits of a new reference could extend beyond just variant detection, individual reference genomes for wild potato species could prove to be an important resources for future studies in comparative genomics and diversity (Aversano *et al.* 2015). The *de novo* genome assemblies produced in this study constitute a significant first step towards achieving the goal of producing high-quality genome assemblies for subps. *andigena* and *goniocalyx.*

As the results of the whole genome alignment show, the current assemblies already cover the majority of the reference genome and show putative large-scale variation between each subspecies and the reference (see **Figures 4.2** and **4.3**). However, these results also indicate a large

number of potential misassemblies, with several contigs mapping to loci that do not correspond with the location of that sequence in the potato reference genome, leading to several "outlier points" in the dot plots. Additionally, comparing the quality metrics of the current draft assemblies and the most recent potato reference genome shows that the assemblies produced in this study are of much lower quality than the potato reference. This is not surprising considering the difference in the data; the potato genome sequencing consortium was a multi-national effort that generated considerable amounts data for the assembly of the first potato reference genome, while our re-sequencing effort lacks the depth to produce such a high-quality assembly. Indeed, the fact that the lengths of both draft assemblies exceed the estimated length of the full potato genome seems to indicate that there are still a number of misassemblies that need to be corrected. Future studies using mate-pair or long read sequencing data could help resolve these issues and produce better anchoring and scaffolding, as well as fill the existing gaps in the draft assemblies.

Another factor that should be taken into account in any future assembly efforts is the difference in ploidy levels between landraces. In this case, *S. tuberosum* subsp. *andigena* is a tetraploid and *S. stenotomum* subsp. *goniocalyx* is a diploid, which could explain the significant disparity in quality metrics observed between both genomes (see **Table 4.2)** as well as the difference in computing time between both samples. In general, it seems that polyploid genomes take longer to assemble and, in general, have worse quality metrics than diploid (and monoploid) genomes. This could be a consequence of the high degree of heterozygosity that has been observed in potato, which could interfere with the *de novo* assembly algorithms which have been designed to work with data from organisms that are much not polyploid or are less heterozygous than potato.

Our results also indicate the potential to further improve the currently available potato reference genomes with the addition of a several putative novel sequences that were found in each landrace. Most of these new sequences had significant hits to sequencing data from other eudicots (many from the Solanaceae family), which is evidence that these wild potato species could contain genes that have are not found in current commercial potato varieties. Additional transcriptome data coupled with gene model prediction software could be used to annotate most of these novel sequences and could guide future experiments to validate the function of these putative genes, both in the laboratory and the field. Any results derived from these could be of great interest to breeders wishing to incorporate new genetic material from wild potatoes into their improvement programs, as well as for diversity conservation efforts in potato.

In conclusion, the genome re-sequencing of two potato landraces, *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*, has provided additional evidence of significant structural variation between wild potato species and the potato reference genome. Further experimentation and validation is required to produce high-quality reference genomes for each of these varieties, however these initial results indicate numerous potential benefits that could result from the continuation of this work. Additionally, the bioinformatics methods developed and tested with these samples could be applied to additional wild potato species as well as other crops with complex genomes.

4.4 Future Work

The draft genomes of potato landraces *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx* still require significant refinement in order to produce a high-quality, well annotated reference. Future efforts to improve these assemblies are already underway, and include data generation to further anchor current contigs and fragments into chromosome-length pseudomolecules, as well as to correct misassemblies present in the current drafts. Some strategies to accomplish this include the introduction of mate pair or single-molecule sequencing data (such as PacBio) data that can guide genome assembly algorithms, especially in the final scaffolding and anchoring steps. Recent genome assembly efforts in other crops with complex genomes such as carrot (Iorizzo *et al.* 2016) and rice (Schatz *et al.* 2014; Mahesh *et al.* 2016) have shown that this combination of re-sequencing data with long reads to is a powerful tool that produces high-quality reference genomes.

Another different approach would be to use the currently available potato reference genomes to guide the genome assembly process and improve the contigs and scaffolds. There have been multiple methodologies developed for reference-guided genome assembly (Wajid *et al.* 2012, 2013; Bao *et al.* 2014), however they all take advantage of the fact that closely related genomes tend to have a similar overall architecture. For example, a reference-guided genome assembly approach has been shown to be effective for the production of genome references in different strains of *Arabidopsis thaliana* (Schneeberger *et al.* 2011). Additionally, an in-depth, targeted analysis of N responsive genes and regulatory mechanisms with the goal of detecting any structural variations across could be of great interest, especially if any significant differences are found which could account for differences in N response in either of these landraces.

Finally, any future efforts to improve the current draft genome assemblies for subsp. *andigena* and *goniocalyx* must include a transcriptome analysis in addition to gene prediction in order to fully annotate the landrace assemblies. The simplest way to achieve this is by obtaining RNA-seq data from one or more tissues and then using it to assemble gene models which can be annotated and added to the draft reference (Souvorov *et al.* 2010).

4.5 Materials and Methods

4.5.1 Plant Materials, DNA Library Preparation and Re-Sequencing

Potato plants were propagated *in vitro* the International Potato Center in Lima, Peru, from the germplasm of two Peruvian potato landraces: *Solanum tuberosum* subsp. *andigena* (CIP accession number 700921) and *Solanum stenotomum* subsp. *goniocalyx* (CIP accession number 702472). Genomic DNA was extracted from the leaves of the *in vitro* plantlets using the E.Z.N.A. Plant DNA Kit [Omega Bio-tek, Inc.], following kit instructions and shipped for sequencing.

After an initial DNA quality assessment, library preparation and DNA sequencing were performed by Novogene Corporation (Beijing, China). Genomic DNA libraries were prepared using the TruSeq Library Construction Kit [Illumina, Inc.] following kit instructions. After libraries were size-selected and purified, they were pooled together and sequenced using an Illumina HiSeq sequencer [Illumina, Inc.] in paired-end mode (2 ×150 bp).

4.5.2 Initial Data Filtering and Alignment

Raw sequencing data from both samples were initially processed using Trimmomatic v0.36 (Bolger *et al.* 2014). The program was configured to trim standard TruSeq3 Paired-End Illumina adapter sequences and low quality bases. Any reads in which one of the two paired reads had less than 60 high-quality base pairs were also discarded.

After filtering, reads from each genome were aligned separately to the Potato Reference Genome v4.03 using BowTie2 v2.2.3 (Langmead & Salzberg 2012). The program was configured to perform an end-to-end alignment on paired-end reads in ‘sensitive’ mode, and to create a file with all the reads that did not align to the genome. These unaligned reads were analyzed separately (see below). Aligned paired reads for each landrace genome were saved as a BAM file and then sorted using SamTools v1.3.1 (Li, Handsaker, *et al.* 2009) to facilitate further analysis. Depth and

coverage statistics were computed using the 'genomecov' tool included in BedTools v2.17.0 (Quinlan & Hall 2010).

4.5.3 CNV Detection and Analysis

Sorted aligned reads were used as input for the CNV discovery using CNVnator v0.3.2 (Abyzov *et al.* 2011). The CNV detection protocol was based on previously reported studies using the same software in potato (Hardigan *et al.* 2016). In summary, CNVnator was set to call duplications and deletions based on sequencing depth, with sliding genomic windows of 100 nt length. Raw CNV results were filtered using a P-value cutoff of 0.05.

After filtering, CNVs results were analyzed with the built-in tools in CNVnator to determine the number of genes affected by CNVs. To do this, the entirety of the potato reference genome was subdivided into 200 kb bins and the number of CNV-affected genes were estimated for every bin. The results of these analyses were collectively visualized using the Circos visualization software (Krzywinski *et al.* 2009).

The two genomic bins with the most CNV-affected genes in each landrace were further analyzed to determine the function of affected genes. This was accomplished by searching for the annotation of each of the affected gene, in particular any associated InterPro and/or GO terms, using curated potato databases (Amar *et al.* 2014; Fernandez-Pozo *et al.* 2014; Hirsch *et al.* 2014).

4.5.4 De novo Draft Genome Assembly and Evaluation

After testing numerous *de novo* genome assembly tools, most of which were not able to scale to the size and complexity of the dataset, two draft *de novo* genome assemblies were generated using ABySS-denovo v1.3.3 (Simpson et al. 2009) and MaSuRCA v.3.1.3 (Zimin et al. 2013). The ABySS assembly was run using a k-value of 64 for both samples. The input consisted of the sequencing data after filtering and trimming. Initial quality control of the contigs and scaffolds produced by ABySS indicated they were relatively short and the software logs showed that the algorithm discarded a great number of sequencing reads during the assembly process (data not shown).

Raw sequencing reads were used as input for the MaSuRCA *de novo* genome assembler, as specified by the software developer (Zimin *et al.* 2013). In both cases, the graph kmer size was set to automatic, the cgwErrorRate parameter was set to 0.15, and the jellyfish hash size was set to 50,600,000,000 (based on the reference genome size and the re-sequencing depth). Because

preliminary analyses showed that the contigs and scaffolds assembled by MaSuRCA were much longer than those produced by ABySS, from this point forward, only MaSuRCA contigs were used for analysis.

The quality metrics of the contigs and scaffolds produced by MaSuRCA were calculated using the QUAST genome assessment tool v4.3 (Gurevich *et al.* 2013). To facilitate the analysis of a dataset of this size, QUAST was set to perform in the 'memory-efficient' mode and to discard any contigs and scaffolds of a length below 250 nt. Finally, whole-genome alignment between each draft genome and the reference was done using the NUCmer tool included in the MUMmer suite v3.23 (Kurtz *et al.* 2004). The alignment data was used to generate a dot plot for every chromosome in each assembly using 'mummerplot' (another tool included in MUMmer). To improve the legibility of the dot plots, only the best alignment was displayed for each contig and scaffold.

4.5.5 Analysis of Un-Aligned Sequences and Contigs

Reads that did not align to version 4.03 of the potato reference genome were assembled separately using ABySS (Simpson *et al.* 2009) to produce contigs that could then be used to analyze potential novel sequences. Since the goal of these contigs was only to facilitate novel sequence analysis, but not to be a part of a full draft genome, their quality metrics were not fully determined and no attention was paid to their length in comparison to the other assembled contigs using MaSuRCA.

Using BLASTn v2.4 (Altschul *et al.* 1997), these contigs were first aligned to the un-anchored, novel sequences (labeled 'crhUn') published in version 4.04 of the potato reference genome (Hardigan *et al.* 2016). All remaining unaligned sequences were then used to query the nucleotide database from NCBI (Wheeler *et al.* 2007; Benson *et al.* 2013), again using BLASTn v2.4 (Altschul *et al.* 1997). The top hit from each was extracted from the results and the taxonomic data of the species was included in the final report.

Chapter 5: Conclusions and Summary

Food security is becoming a pressing issue worldwide: increasing population coupled with environmental challenges such as climate change have created a long-term pressure on current food production systems. Scientists and producers need to collaborate to develop crops with better yields and resistance to stresses, both biotic and abiotic. Potato is a crop that has been frequently targeted for improvement, in part because evidence suggests that potato yield increases have been mostly the result of better agricultural practices as opposed to breeding or genetic modification (Douches *et al.* 1996). Wild potato relatives contain untapped genetic resources that may be key to the adaptation of this crop to current and future stresses. However, potato genomics and transcriptomics are complex and there is still much that is unknown about this species and its close relatives.

The main goal of this work has been to improve our understanding of potato genomics and transcriptomics in two specific areas: the regulatory response to N supplementation and the large-scale variations between the potato reference genome and two landraces. Together, these objectives build upon the current body of research available to potato producers and researchers by providing novel putative regulatory elements and new draft reference genomes for *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx.*

Using RNA-seq data from field grown potatoes, a total 39 genes were found to have a steady state response to N supplementation across three cultivars (Shepody, Russet Burbank and Atlantic). Thirty genes were consistently over-expressed and nine genes were under-expressed in plants with added N. A significant portion of these genes belonged to pathways involved in sulfate and amino-acid metabolism and were associated with GO terms related to integral components of the cell membrane, response to cadmium ion, oxidation-reduction processes, and transmembrane transport.

Nine putative regulatory motifs were discovered in the upstream flanking regions of the N responsive genes using a combination of three motif discovery programs: Seeder, Weeder and MEME. All motifs had significant hits on curated regulatory motif databases, suggesting a potential mechanism that regulates steady state gene response to differences in N supplementation in potato. However, additional experimental work is required to fully validate their function in potato.

Both of these results improve our understanding of N metabolism and regulation in potato and could have a short and long term impact in future potato breeding programs. Breeders can now screen for mutations in any of the identified N responsive genes, especially when developing potato cultivars with higher or lower sensitivity to N supplementation. These genes have been shown to have a similar change in expression across different cultivars and any changes could produce significant phenotypic variation. Additionally, by selecting and screening for these genes, potato researchers can collect more data on their impact in N response, which in turn could lead to the development of models for a better prediction of N state in field grown crops.

In a longer term, these results could serve as a foundation for the development of molecular sensors for N state in potato and related crops. In the future, it could be possible to determine whether field grown plants are being supplemented with enough N by measuring changes in the expression of these N responsive genes. In a similar vein, this study could serve as a template for additional experiments with other important nutrients such as potassium and phosphorous, increasing the value of a molecular monitoring system because it would be able to measure several different factors at once.

The genome re-sequencing data has shown evidence of significant structural variation in the genomes of potato the landraces *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx*. When compared to the potato reference genome, they were found to have large numbers of deletions and duplications (CNVs) spread throughout the genome. These CNVs had variable lengths, with duplications generally being of greater size than deletions. A total of 7,423 genes were affected in subsp. *andigena* while 6,040 genes were affected in subsp. *goniocalyx*. The most significant CNV-affected gene clusters in both varieties affected mainly carbohydrate metabolic genes.

While none of the identified N responsive genes were found to be affected by CNVs in this study, a future experiment could include a targeted analysis of re-sequencing data collected from several potato landraces to specifically detect differences in these and other N responsive genes. Identifying the potato landraces and wild relatives with significant deletions or duplications in N responsive genes could be useful to determine how these genes have evolved and how N metabolism can be altered in commercial potato varieties.

A draft reference genome was assembled *de novo* for each landrace. While both genomes still require considerable work, including the anchoring of contigs and scaffolds into a single pseudomolecule, they have already highlighted some key differences between each landrace and the potato reference genome. The pipeline used to assemble these two genomes can now be applied to other potato landraces and cultivars to expand our understanding of the structural diversity in potato and all its related species.

Additionally, when working with each landrace several technical differences between the datasets were uncovered. For example, the tetraploid variety (subsp. *andigena*) took almost twice as long to assemble and, on average, produced shorter contigs and scaffolds, which seem to indicate that this is a more complex genome. It is also likely that additional data or a deeper analysis will be required to resolve the potential misassemblies that arise due to haplotype phasing. However, the benefits of developing bioinformatics methods for polyploid genome assembly could extend to applications well beyond potato genomics: many cancerous tumor cells are polyploids and there has been a lot of interest in the development of robust methods to deal with polyploid genome assembly and haplotype detection in this field as well (Aguiar *et al.* 2014).

The analyzed re-sequencing data in both landraces also contained a number of sequences that did not align to the most recent version of the potato reference genome. These putative novel regions could contain valuable information about the genetic differences between each individual landrace, which in turn has interesting implications for the fields of potato biodiversity and evolution. When these novel sequences were used to query the NCBI nucleotide database, the majority of significant hits were for genes and sequences from other Solanaceae, but a reduced number of sequences also had hits to known bacterial and viral sequences. Deeper analysis is required to determine whether these sequences constitute actual bacterial and viral sequences that have been incorporated into the genome of these landraces, but the possibility itself is interesting due to the fact that a similar phenomenon was recently observed in a species of sweet potato (Kyndt *et al.* 2015).

In recent years there has been an increasing awareness that the future of plant science and crop improvement is highly dependent on the ability to develop appropriate computational tools and capacities. Taken together, these studies have provided a bioinformatics framework for genomics and transcriptomic studies in potato and other polyploid crops. Due to the fact that most

bioinformatics tools are developed to deal with bacterial or human genomic data, special care is required to adequately validate and interpret results in non-model organisms, especially in plants which often have complicating factors such as highly repetitive genomic regions, complex regulatory mechanisms, and polyploidy. This work has taken special care to test different software alternatives for every step and select the tools that produce the most consistent and reproducible results.

In summary, the results of these studies have expanded the current knowledge of potato genomics and transcriptomics. By uncovering putative gene regulatory mechanisms and large-scale structural variations in potato genomes of different varieties, future researchers have new targets for experimental validation and incorporation into existing improvement and breeding efforts. The genomes of *S. tuberosum* subsp. *andigena* and *S. stenotomum* subsp. *goniocalyx* have uncovered interesting structural variation in the genomes of potato relatives. And finally, the bioinformatics methods developed for these experiments could also be useful for the study of other non-model organisms with repetitive, complex genomes, helping to expand our molecular understanding of plants beyond Solanaceae with the ultimate goal of developing better crops that can ensure a sustainable food production today and in the future.

References

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

Aflitos S., Schijlen E., De Jong H., De Ridder D., Smit S., Finkers R., Wang J., Zhang G., Li N., Mao L., Bakker F., Dirks R., Breit T., Gravendeel B., Huits H., Struss D., Swanson-Wagner R., Van Leeuwen H., Van Ham R.C.H.J., Fito L., Guignier L., Sevilla M., Ellul P., Ganko E., Kapur A., Reclus E., De Geus B., Van De Geest H., Te Lintel Hekkert B., Van Haarst J., Smits L., Koops A., Sanchez-Perez G., Van Heusden A.W., Visser R., Quan Z., Min J., Liao L., Wang X., Wang G., Yue Z., Yang X., Xu N., Schranz E., Smets E., Vos R., Rauwerda J., Ursem R., Schuit C., Kerns M., Van Den Berg J., Vriezen W., Janssen A., Datema E., Jahrman T., Moquet F., Bonnet J., Peters S. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant Journal **80**:136–148.

Aguiar D., Wong W.S.W., Istrail S. (2014) Tumor haplotype assembly algorithms for cancer genomics. In: Pacific Symposium on Biocomputing.pp 3–14.

Alföldi J., Lindblad-Toh K. (2013) Comparative genomics as a tool to understand evolution and disease. Genome Research **23**:1063–1068.

Almasia N.I., Narhirñak V., Hopp H.E., Vazquez-Rovere C. (2010) Isolation and characterization of the tissue and development-specific potato snakin-1 promoter inducible by temperature and wounding. Electronic Journal of Biotechnology **13**:1–21.

de Almeida M.R., de Bastiani D., Gaeta M.L., de Araújo Mariath J.E., de Costa F., Retallick J., Nolan L., Tai H.H., Strömvik M. V., Fett-Neto A.G. (2015) Comparative transcriptional analysis provides new insights into the molecular basis of adventitious rooting recalcitrance in Eucalyptus. Plant Science **239**:155–165.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**:3389–3402.

Amar D., Frades I., Danek A., Goldberg T., Sharma S.K., Hedley P.E., Proux-Wera E., Andreasson

E., Shamir R., Tzfadia O., Alexandersson E. (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biology **14**:1–14.

Aminedi R., Das N. (2014) Class I patatin genes from potato (Solanum tuberosum L .) cultivars: molecular cloning, sequence comparison, prediction of diverse cis-regulatory motifs, and assessment of the promoter activities under field and in vitro conditions. In Vitro Cellular & Developmental Biology - Plant **50**:673–687.

Ancillo G., Hoegen E., Kombrink E. (2003) The promoter of the potato chitinase C gene directs expression to epidermal cells. Planta **217**:566–576.

Anders S., Pyl P.T., Huber W. (2014) HTSeq A Python framework to work with high-throughput sequencing data. Bioinformatics **31**:166–169.

Anders S., Reyes A., Huber W. (2012) Detecting differential usage of exons from RNA-seq data. Genome Research **22**:2008–2017.

Andolfo G., Jupe F., Witek K., Etherington G.J., Ercolano M.R., Jones J.D.G. (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC plant biology **14**:120.

Anithakumari A.M., Tang J., van Eck H.J., Visser R.G.F., Leunissen J.A.M., Vosman B., van der Linden C.G. (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Molecular Breeding **26**:65–75.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G., Consortium G.O. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics **25**:25–29.

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., Gabaldón T., Frusciante L., Bradeen J.M., 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**:954–68.

Bachem C., Van Der Hoeven R., Lucker J., Oomen R., Casarini E., Jacobsen E., Visser R. (2000)

Functional genomic analysis of potato tuber life-cycle. Potato Research **43**:297–312.

Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., Noble W.S. (2009) MEME Suite: Tools for motif discovery and searching. Nucleic Acids Research **37**:202–208.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z. a, Selker E.U., Cresko W. a, Johnson E. a (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one **3**:e3376.

Bansal A., Kumari V., Taneja D., Sayal R., Das N. (2012) Molecular cloning and characterization of granule-bound starch synthase I (GBSSI) alleles from potato and sequence analysis for detection of cis-regulatory motifs. Plant Cell, Tissue and Organ Culture (PCTOC) **109**:247–261.

Bao S., An L., Su S., Zhou Z., Gan Y. (2011) Expression patterns of nitrate, phosphate, and sulfate transporters in Arabidopsis roots exposed to different nutritional regimes. Botany **89**:647–653.

Bao E., Jiang T., Girke T. (2014) AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. Bioinformatics **30**:319–328.

Barone A. (2004) Molecular marker-assisted selection for potato breeding. American Journal of Potato Research **81**:111–117.

Bengtsson T., Weighill D., Proux-Wéra E., Levander F., Resjö S., Burra D.D., Moushib L.I., Hedley P.E., Liljeroth E., Jacobson D., Alexandersson E., Andreasson E. (2014) Proteomics and transcriptomics of the BABA-induced resistance response in potato using a novel functional annotation approach. BMC Genomics **15**:315.

Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2013) GenBank. Nucleic Acids Research **41**:36–42.

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

Bloom A.J. (2015) Photorespiration and nitrate assimilation: a major intersection between plant

carbon and nitrogen. Photosynthesis Research **123**:117–128.

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

Bombarely A., Moser M., Amrad A., Bao M., Bapaume L., Barry C.S., Bliek M., Boersma M.R., Borghi L., Bruggmann R., Bucher M., D'Agostino N., Davies K., Druege U., Dudareva N., Egea-Cortines M., Delledonne M., Fernandez-Pozo N., Franken P., Grandont L., Heslop-Harrison J.S., Hintzsche J., Johns M., Koes R., Lv X., Lyons E., Malla D., Martinoia E., Mattson N.S., Morel P., Mueller L.A., Muhlemann J., Nouri E., Passeri V., Pezzotti M., Qi Q., Reinhardt D., Rich M., Richert-Pöggeler K.R., Robbins T.P., Schatz M.C., Schranz M.E., Schuurink R.C., Schwarzacher T., Spelt K., Tang H., Urbanus S.L., Vandenbussche M., Vijverberg K., Villarino G.H., Warner R.M., Weiss J., Yue Z., Zethof J., Quattrocchio F., Sims T.L., Kuhlemeier C. (2016) Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. Nature plants **2**:16074.

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics **23**:2633–2635.

Brazma A., Parkinson H., Sarkans U., Shojatalab M., Vilo J., Abeygunawardena N., Holloway E., Kapushesky M., Kemmeren P., Lara G.G., Oezcimen A., Rocca-Serra P., Sansone S.A. (2003) ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Research **31**:68–71.

Bredeson J. V, Lyons J.B., Prochnik S.E., Wu G.A., Ha C.M., Edsinger-Gonzales E., Grimwood J., Schmutz J., Rabbi I.Y., Egesi C., Nauluvula P., Lebot V., Ndunguru J., Mkamilo G., Bart R.S., Setter T.L., Gleadow R.M., Kulakow P., Ferguson M.E., Rounsley S., Rokhsar D.S. (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nature Biotechnology **34**:562–570.

Campbell M., Segear E., Beers L., Knauber D., Suttle J. (2008) Dormancy in potato tuber meristems: Chemically induced cessation in dormancy matches the natural process based on transcript profiles. Functional and Integrative Genomics **8**:317–328.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. (2011) Stacks: building and

genotyping Loci de novo from short-read sequences. G3: Genes|Genomes|Genetics **1**:171–82.

Chaisson M.J.P., Wilson R.K., Eichler E.E. (2015) Genetic variation and the de novo assembly of human genomes. Nature Reviews Genetics **16**:627–640.

Chan C., Guo L., Shih M. (2001) Promoter analysis of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase B subunit of Arabidopsis thaliana. Plant molecular biology **46**:131–141.

Chen M., Zhu W.J., You X., Liu Y.D., Kaleri G.M., Yang Q. (2015) Isolation and characterization of a chalcone isomerase gene promoter from potato cultivars. Genetics and Molecular Research **14**:18872–18885.

Cho K., Cho K.S., Sohn H.B., Ha I.J., Hong S.Y., Lee H., Kim Y.M., Nam M.H. (2016) Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. Journal of Experimental Botany **67**:1519–1533.

Cock P.J.A., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., De Hoon M.J.L. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**:1422–1423.

Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szcześniak M.W., Gaffney D.J., Elo L.L., Zhang X., Mortazavi A. (2016) A survey of best practices for RNA-seq data analysis. Genome Biology **17**:13.

Crooks G., Hon G., Chandonia J., Brenner S. (2004) WebLogo: a sequence logo generator. Genome Research **14**:1188–1190.

Crookshanks M., Emmersen J., Welinder K.G., Nielsen K.L. (2001) The potato tuber transcriptome: analysis of 6077 expressed sequence tags. FEBS letters **506**:123–6.

Dann A.L., Wilson C.R. (2011) Comparative assessment of genetic and epigenetic variation among regenerants of potato (Solanum tuberosum) derived from long-term nodal tissue-culture and cell selection. Plant Cell Reports **30**:631–639.

Denancé N., Szurek B., Noël L.D. (2014) Emerging Functions of Nodulin-Like Proteins in Non-Nodulating Plant Species. Plant and Cell Physiology **55**:469–474.

Despres C., Subramaniam R., Matton D.P., Brisson N. (1995) The Activation of the Potato Pr-Loa Gene Requires the Phosphorylation of the Nuclear Factor Pbf-1. Plant Cell **7**:589–598.

De Donato M., Peters S.O., Mitchell S.E., Hussain T., Imumorin I.G. (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PloS one **8**:e62137.

Douches D.S., Jastrzebski K., Maas D., Chase R.W. (1996) Assessment of potato breeding over the past century. Crop Science **36**:1544–1552.

Eisenstein M. (2015) Startups use short-read data to expand long-read sequencing market. Nature Biotechnology **33**:433–435.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J. a, Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one **6**:e19379.

Endelman J. (2015) Genotyping-By-Sequencing of a Diploid Potato F2 Population. In: Plant and Animal Genome XXIII. San Diego [online] URL: https://pag.confex.com/pag/xxiii/webprogram/Paper15683.html

Endelman J.B., Jansky S.H. (2016) Genetic mapping with an inbred line-derived F2 population in potato. Theoretical and Applied Genetics **129**:935–943.

Evers D., Lefèvre I., Legay S., Lamoureux D., Hausman J.F., Rosales R.O.G., Marca L.R.T., Hoffmann L., Bonierbale M., Schafleitner R. (2010) Identification of drought-responsive compounds in potato through a combined transcriptomic and targeted metabolite approach. Journal of Experimental Botany **61**:2327–2343.

Fauteux F., Blanchette M., Strömvik M. V. (2008) Seeder: Discriminative seeding DNA motif discovery. Bioinformatics **24**:2303–2307.

Felcher K.J., Coombs J.J., Massa A.N., Hansey C.N., Hamilton J.P., Veilleux R.E., Buell C.R., Douches D.S. (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS ONE **7**:1–11.

Fernandez-Pozo N., Menda N., Edwards J.D., Saha S., Tecle I.Y., Strickler S.R., Bombarely A., Fisher-york T., Pujar A., Foerster H., Yan A., Mueller L.A. (2014) The Sol Genomics

Network (SGN)–from genotype to phenotype to breeding. Nucleic Acids Research **43**:1–6.

Flinn B., Rothwell C., Griffiths R., Lägue M., DeKoeyer D., Sardana R., Audy P., Goyer C., Li X.Q., Wang-Pruski G., Regan S. (2005) Potato expressed sequence tag generation and analysis using standard and unique cDNA libraries. Plant Molecular Biology **59**:407–433.

Food and Agriculture Organization (2016) Food and Agricultural commodities production / Commodities by regions. FAOSTAT [online] URL: http://faostat3.fao.org/browse/rankings/commodities_by_regions/E

Frades I., Abreha K.B., Proux-Wéra E., Lankinen Å., Andreasson E., Alexandersson E. (2015) A novel workflow correlating RNA-seq data to Phythophthora infestans resistance levels in wild Solanum species and potato clones. Frontiers in plant science **6**:718.

Fulladolsa A.C., Navarro F.M., Kota R., Severson K., Palta J.P., Charkowski A.O. (2015) Application of Marker Assisted Selection for Potato Virus Y Resistance in the University of Wisconsin Potato Breeding Program. American Journal of Potato Research **92**:444–450.

Gálvez J.H., Tai H.H., Barkley N.A., Gardner K., Ellis D., Strömvik M. V. (2017) Understanding potato with the help of genomics. AIMS Agriculture and Food **2**:16–39.

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

Gao L., Tu Z.J., Millett B.P., Bradeen J.M. (2013) Insights into organ-specific pathogen defense responses in plants: RNA-seq analysis of potato tuber-Phytophthora infestans interactions. BMC genomics **14**:340.

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

Gebhardt C., Ballvora A., Walkemeier B., Oberhagemann P., Schüler K. (2004) Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. Molecular Breeding **13**:93–102.

Gebhardt C., Bellin D., Henselewski H., Lehmann W., Schwarzfischer J., Valkonen J.P.T. (2006) Marker-assisted combination of major genes for pathogen resistance in potato. Theoretical and Applied Genetics **112**:1458–1464.

Ghislain M., Andrade D., Rodríguez F., Hijmans R.J., Spooner D.M. (2006) Genetic analysis of the cultivated potato Solanum tuberosum L. Phureja Group using RAPDs and nuclear SSRs. Theoretical and Applied Genetics **113**:1515–1527.

Gilmartin P., Niner B., Chua N. (1990) A Metal-Dependent DNA-Binding Protein lnteracts with a Constitutive Element of a Light-Responsive Promoter. Society **2**:857–866.

Ginzberg I., Barel G., Ophir R., Tzin E., Tanami Z., Muddarangappa T., De Jong W., Fogelman E. (2009) Transcriptomic profiling of heat-stress response in potato periderm. Journal of Experimental Botany **60**:4411–4421.

Glaubitz J.C., Casstevens T.M., Lu F., Harriman J., Elshire R.J., Sun Q., Buckler E.S. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PloS one **9**:e90346.

Gong L., Zhang H., Gan X., Zhang L., Chen Y., Nie F., Shi L., Li M., Guo Z., Zhang G., Song Y. (2015) Transcriptome Profiling of the Potato (Solanum tuberosum L.) Plant under Drought Stress and Water-Stimulus Conditions. Plos One **10**:1–20.

Goodwin S., Gurtowski J., Ethe-Sayers S., Deshpande P., Schatz M.C., McCombie W.R. (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome research **25**:1750–1756.

Gordon D.B., Nekludova L., McCallum S., Fraenkel E. (2005) TAMO: A flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics **21**:3164–3165.

Goyer A., Hamlin L., Crosslin J.M., Buchanan A., Chang J.H. (2015) RNA-Seq analysis of resistant and susceptible potato varieties during the early stages of potato virus Y infection. BMC genomics **16**:472.

Grierson C., Du J.S., de Torres Zabala M., Beggs K., Smith C., Holdsworth M., Bevan M. (1994) Separate cis sequences and trans factors direct metabolic and developmental regulation of a potato tuber storage protein gene. Plant Journal **5**:815–826.

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

Hamilton J.P., Hansey C.N., Whitty B.R., Stoffel K., Massa A.N., Deynze A. Van, Jong W.S. De, Douches D.S., Buell C.R. (2011) Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC genomics **12**:302.

Hammond J.P., Broadley M.R., Bowen H.C., Spracklen W.P., Hayden R.M., White P.J. (2011) Gene expression changes in phosphorus deficient potato (Solanum tuberosum L.) leaves and the potential for diagnostic gene expression markers. PloS one **6**:e24606.

Hancock R.D., Morris W.L., Ducreux L.J.M., Morris J. a, Usman M., Verrall S.R., Fuller J., Simpson C.G., Zhang R., Hedley P.E., Taylor M. a (2014) Physiological, biochemical and molecular responses of the potato (Solanum tuberosum L.) plant to moderately elevated temperature. Plant, cell & environment **37**:439–50.

Hanneman R.E., Bamberg J.B. (1986) *Inventory of tuber-bearing Solanum species*. University of Wisconsin.

Harbison C.T., Gordon D.B., Lee T.I., Rinaldi N.J., Macisaac K.D., Danford T.W., Hannett N.M., Tagne J.B., Reynolds D.B., Yoo J., Jennings E.G., Zeitlinger J., Pokholok D.K., Kellis M., Rolfe P.A., Takusagawa K.T., Lander E.S., Gifford D.K., Fraenkel E., Young R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature **431**:1–5. [online] URL: http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature02800.html

Hardigan M.A., Bamberg J., Buell C.R., Douches D.S. (2015) Taxonomy and Genetic Differentiation among Wild and Cultivated Germplasm of sect. Petota. The Plant Genome **8**:1–16.

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., Jiang J., Veilleux R.E., Buell C.R. (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. The Plant Cell **28**:388–405.

Haudry A., Platts A.E., Vello E., Hoen D.R., Leclercq M., Williamson R.J., Forczek E., Joly-Lopez Z., Steffen J.G., Hazzouri K.M., Dewar K., Stinchcombe J.R., Schoen D.J., Wang X.,

Schmutz J., Town C.D., Edger P.P., Pires J.C., Schumaker K.S., Jarvis D.E., Mandáková T., Lysak M. a, van den Bergh E., Schranz M.E., Harrison P.M., Moses A.M., Bureau T.E., Wright S.I., Blanchette M. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature genetics **45**:891–8.

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

Hazen S.P., Wu Y., Kreps J.A. (2003) Gene expression profiling of plant responses to abiotic stress. Functional & integrative genomics **3**:105–111.

Higo K., Ugawa Y., Iwamoto M., Korenaga T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research **27**:297–300.

Hirsch C.D., Buell C.R., Hirsch C.N. (2016) A Toolbox of Potato Genetic and Genomic Resources. American Journal of Potato Research **93**:21–32.

Hirsch C.D., Hamilton J.P., Childs K.L., Cepela J., Crisovan E., Vaillancourt B., Hirsch C.N., Habermann M., Neal B., 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., Buell C.R. (2013) Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries. G3 **3**:1003–13.

Hirsch C.D., Springer N.M., Hirsch C.N. (2015) Genomic Limitations to RNAseq Expression Profiling. The Plant Journal **84**:491–503.

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

Humbert S., Subedi S., Cohn J., Zeng B., Bi Y.-M., Chen X., Zhu T., McNicholas P.D., Rothstein S.J. (2013) Genome-wide expression profiling of maize in response to individual and combined water and nitrogen stresses. BMC genomics **14**:3.

Iorizzo M., Ellison S., Senalik D., Zeng P., Satapoomin P., Huang J., Bowman M., Iovene M., Sanseverino W., Cavagnaro P., Yildiz M., Macko-Podgórni A., Moranska E., Grzebelus E., Grzebelus D., Ashrafi H., Zheng Z., Cheng S., Spooner D., Van Deynze A., Simon P. (2016)

A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nature Genetics **48**:657–666.

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

Johnston S.A., den Nijs T.P.M., Peloquin S.J., Hanneman R.E. (1980) The significance of genic balance to endosperm development in interspecific crosses. Theoretical and Applied Genetics **57**:5–9.

Kaminski K.P., Kørup K., Andersen M.N., Sønderkær M., Andersen M.S., Kirk H.G., Nielsen K.L. (2016) Next Generation Sequencing Bulk Segregant Analysis of Potato Support that Differential Flux into the Cholesterol and Stigmasterol Metabolite Pools Is Important for Steroidal Glycoalkaloid Content. Potato Research **59**:81–97.

Kasai A., Bai S., Hojo H., Harada T. (2016) Epigenome Editing of Potato by Grafting Using Transgenic Tobacco as siRNA Donor. PLoS ONE **11**:e0161729.

Kim S., Park M., Yeom S.-I., Kim Y.-M., Lee J.M., Lee H.-A., Seo E., Choi J., Cheong K., Kim K.-T., Jung K., Lee G.-W., Oh S.-K., Bae C., Kim S.-B., Lee H.-Y., Kim S.-Y., Kim M.-S., Kang B.-C., Jo Y.D., Yang H.-B., Jeong H.-J., Kang W.-H., Kwon J.-K., Shin C., Lim J.Y., Park J.H., Huh J.H., Kim J.-S., Kim B.-D., Cohen O., Paran I., Suh M.C., Lee S.B., Kim Y.-K., Shin Y., Noh S.-J., Park J., Seo Y.S., Kwon S.-Y., Kim H.A., Park J.M., Kim H.-J., Choi S.-B., Bosland P.W., Reeves G., Jo S.-H., Lee B.-W., Cho H.-T., Choi H.-S., Lee M.-S., Yu Y., Do Choi Y., Park B.-S., van Deynze A., Ashrafi H., Hill T., Kim W.T., Pai H.-S., Ahn H.K., Yeam I., Giovannoni J.J., Rose J.K.C., Sørensen I., Lee S.-J., Kim R.W., Choi I.-Y., Choi B.-S., Lim J.-S., Lee Y.-H., Choi D. (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nature Genetics **46**:270–278.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology **14**:R36.

Kloosterman B., De Koeyer D., Griffiths R., Flinn B., Steuernagel B., Scholz U., Sonnewald S., Sonnewald U., Bryan G.J., Prat S., Bánfalvi Z., Hammond J.P., Geigenberger P., Nielsen K.L., Visser R.G.F., Bachem C.W.B. (2008) Genes driving potato tuber initiation and growth:

Identification based on transcriptional changes using the POCI array. Functional and Integrative Genomics **8**:329–340.

Konishi M., Yanagisawa S. (2010) Identification of a nitrate-responsive cis-element in the Arabidopsis NIR1 promoter defines the presence of multiple cis-regulatory elements for nitrogen response. The Plant Journal **63**:269–282.

Konishi M., Yanagisawa S. (2011) Roles of the transcriptional regulation mediated by the nitrate-responsive cis-element in higher plants. Biochemical and Biophysical Research Communications **411**:708–713.

Konishi M., Yanagisawa S. (2013) Arabidopsis NIN-like transcription factors have a central role in nitrate signalling. Nature Communications **4**:1617.

Korkuc P., Schippers J.H.M., Walther D. (2014) Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information. Plant Physiology **164**:181–200.

Kosugi S., Hirakawa H., Tabata S. (2015) GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics **31**:3733–3741.

Krzywinski M., Schein J., Birol I., Connors J., Gascoyne R., Jones S.J., Marra M.A. (2009) Circos: an Information Aesthetic for Comparative Genomics. Genome Research **19**:1639–1645.

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

Kyndt T., Quispe D., Zhai H., Jarret R., Ghislain M., Liu Q., Gheysen G. (2015) The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes : An example of a naturally transgenic food crop. Proceedings of the National Academy of Sciences **112**:5844–5849.

Labate J.A., Robertson L.D., Strickler S.R., Mueller L.A. (2014) Genetic structure of the four wild tomato species in the Solanum peruvianum s.l. species complex. Genome **57**:169–80.

Langmead B., Salzberg S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nature methods **9**:357–359.

Law R.D., Suttle J.C. (2005) Chromatin remodeling in plant cell culture: patterns of DNA methylation and histone H3 and H4 acetylation vary during growth of asynchronous potato cell suspensions. Plant Physiology and Biochemistry **43**:527–534.

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

Liang M., Raley C., Zheng X., Kutty G., Gogineni E., Sherman B.T., Sun Q., Chen X., Skelly T., Jones K., Stephens R., Zhou B., Lau W., Johnson C., Imamichi T., Jiang M., Dewar R., Lempicki R.A., Tran B., Kovacs J.A., Huang D.W. (2016) Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. BioData Mining **9**:13.

Li B., Ruotti V., Stewart R.M., Thomson J. a., Dewey C.N. (2009) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**:493–500.

Li X.Q., Sveshnikov D., Zebarth B.J., Tai H., de Koeyer D., Millard P., Haroon M., Singh M. (2010) Detection of nitrogen sufficiency in potato plants using gene expression markers. American Journal of Potato Research **87**:50–59.

Li L., Tacke E., Hofferbert H.-R., Lübeck J., Strahwald J., Draffehn A.M., Walkemeier B., Gebhardt C. (2013) Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality. Theoretical and Applied Genetics **126**:1039–1052.

Liepman A.H., Olsen L.J. (2001) Peroxisomal alanine: Glyoxylate aminotransferase (AGT1) is a photorespiratory enzyme with multiple substrates in Arabidopsis thaliana. Plant Journal **25**:487–498.

Liseron-Monfils C., Bi Y.-M., Downs G.S., Wu W., Signorelli T., Lu G., Chen X., Bondo E., Zhu T., Lukens L.N., Colasanti J., Rothstein S.J., Raizada M.N. (2013) Nitrogen transporter and assimilation genes exhibit developmental stage-selective expression in maize (Zea mays L.) associated with distinct cis-acting promoter motifs. Plant signaling & behavior **8**:1–14.

Liu B., Zhang N., Wen Y., Jin X., Yang J., Si H., Wang D. (2015) Transcriptomic changes during tuber dormancy release process revealed by RNA sequencing in potato. Journal of

Biotechnology **198**:17–30.

Long C., Snapp S., Douches D., Chase R. (2004) Tuber yield, storability, and quality of Michigan cultivars in response to nitrogen management and seedpiece spacing. American Journal of Potato Research **81**:347–357.

López Y., Patil A., Nakai K. (2013) Identification of novel motif patterns to decipher the promoter architecture of co-expressed genes in Arabidopsis thaliana. BMC Systems Biology **7**:S10.

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS genetics **9**:e1003215.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.-W., Wang J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**:18.

Luo S., Tai H., Zebarth B., Li X., Millard P., Koeyer D. De, Xiong X. (2011) Sample Collection Protocol Effects on Quantification of Gene Expression in Potato Leaf Tissue. Plant Molecular Biology Reporter **29**:369–378.

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

Mahesh H.B., Shirke M.D., Singh S., Rajamani A., Hittalmani S., Wang G.-L., Gowda M. (2016) Indica rice genome assembly, annotation and mining of blast disease resistance genes. BMC genomics **17**:242.

Mahony S., Benos P. V. (2007) STAMP: A web tool for exploring DNA-binding motif similarities. Nucleic Acids Research **35**:253–258.

Manrique-Carpintero N.C., Coombs J.J., Cui Y., Veilleux R.E., Robin Buell C., 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., Veilleux R.E., Buell C.R., Douches D.S. (2016) Comparative Analysis of Regions with Distorted Segregation in Three Diploid Populations

of Potato. G3: Genes|Genomes|Genetics **6**:2617–2628.

Massa A.N., Childs K.L., Lin H., Bryan G.J., Giuliano G., Buell C.R. (2011) The Transcriptome of the Reference Potato Genome Solanum tuberosum Group Phureja Clone DM1-3 516R44. PLoS ONE **6**:1–8.

Massa A.N., Manrique-Carpintero N.C., Coombs J.J., Zarka D.G., Boone A.E., Kirk W.W., Hackett C.A., Bryan G.J., Douches D.S. (2015) Genetic Linkage Mapping of Economically Important Traits in Cultivated Tetraploid Potato (Solanum tuberosum L.). G3: Genes|Genomes|Genetics **5**:2357–2364.

Mathelier A., Zhao X., Zhang A.W., Parcy F., Worsley-Hunt R., Arenillas D.J., Buchman S., Chen C.Y., Chou A., Ienasescu H., Lim J., Shyr C., Tan G., Zhou M., Lenhard B., Sandelin A., Wasserman W.W. (2014) JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research **42**:142–147.

Melo A.T.O., Bartaula R., Hale I. (2016) GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC bioinformatics **17**:29.

Meyer S., Nagel A., Gebhardt C. (2005) PoMaMo — a comprehensive database for potato genome data. Nucleic Acids Research **33**:666–670.

Michaeli S., Fromm H. (2015) Closing the loop on the GABA shunt in plants: are GABA metabolism and signaling entwined? Frontiers in Plant Science **6**:1–7.

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

Michelmore R., Reyes Chin-Wo S., Kozik A., Lavelle D., Maria Jose Truco (2016) Improvement of the Genome Assembly of Lettuce (Lactuca sativa) Using Dovetail/in vitro Proximity Ligation. In: Plant and Animal Genome XXIV Conference. [online] URL: https://pag.confex.com/pag/xxiv/webprogram/Paper22314.html

Ming R., VanBuren R., Wai C.M., Tang H., Schatz M.C., Bowers J.E., Lyons E., Wang M.-L., Chen J., Biggers E., Zhang J., Huang L., Zhang L., Miao W., Zhang J., Ye Z., Miao C., Lin Z., Wang H., Zhou H., Yim W.C., Priest H.D., Zheng C., Woodhouse M., Edger P.P., Guyot R., Guo H.-B., Guo H., Zheng G., Singh R., Sharma A., Min X., Zheng Y., Lee H., Gurtowski

J., Sedlazeck F.J., Harkess A., McKain M.R., Liao Z., Fang J., Liu J., Zhang X., Zhang Q., Hu W., Qin Y., Wang K., Chen L.-Y., Shirley N., Lin Y.-R., Liu L.-Y., Hernandez A.G., Wright C.L., Bulone V., Tuskan G.A., Heath K., Zee F., Moore P.H., Sunkar R., Leebens-Mack J.H., Mockler T., Bennetzen J.L., Freeling M., Sankoff D., Paterson A.H., Zhu X., Yang X., Smith J.A.C., Cushman J.C., Paull R.E., Yu Q. (2015) The pineapple genome and the evolution of CAM photosynthesis. Nature Genetics **47**:1435–1442.

Nakane E., Kawakita K., Doke N., Yoshioka H. (2003) Elicitation of primary and secondary metabolism during defense in the potato. Journal of General Plant Pathology **69**:378–384.

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

Nie X., Sutherland D., Dickison V., Singh M., Murphy A.M., Koeyer D. De (2016) Development and Validation of High-Resolution Melting Markers Derived from Ry sto STS Markers for High-Throughput Marker-Assisted Selection of Potato Carrying Ry sto. Phytopathology **106**:1366–1375.

Novitskaya L., Trevanion S.J., Driscoll S., Foyer C.H., Noctor G. (2002) How does photorespiration modulate leaf amino acid contents? A dual approach through modelling and metabolite analysis. Plant, Cell and Environment **25**:821–835.

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

Paajanen P.M., Giolai M., Verweij W., Baker D., Clavijo B., Garcia G., Girona E.L., Baker K., Hein I., Bryan G., Clark M. (2016) S. verrucosum, a Wild Mexican Potato As a Model Species for a Plant Genome Assembly Project. In: Plant and Animal Genome XXIV Conference. [online] URL: https://pag.confex.com/pag/xxiv/webprogram/Paper20356.html

Pachter L. (2011) Models for transcript quantification from RNA-Seq. **1**:1–28. [online] URL: http://arxiv.org/abs/1104.3889

Pavek J.J., Corsini D.L. (2001) Utilization of potato genetic resources in variety development. American Journal of Potato Research **78**:433–441.

Pavesi G., Zambelli F., Pesole G. (2007) WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. BMC bioinformatics **8**:46.

Pendleton M., Sebra R., Pang A.W.C., Ummat A., Franzen O., Rausch T., Stütz A.M., Stedman W., Anantharaman T., Hastie A., Dai H., Fritz M.H.-Y., Cao H., Cohain A., Deikus G., Durrett R.E., Blanchard S.C., Altman R., Chin C.-S., Guo Y., Paxinos E.E., Korbel J.O., Darnell R.B., McCombie W.R., Kwok P.-Y., Mason C.E., Schadt E.E., Bashir A. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods **12**:780–786.

Poland J.A., Brown P.J., Sorrells M.E., Jannink J.-L. (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach (T. Yin, Ed.). PLoS ONE **7**:e32253.

Putnam N.H., Connell B.O., Stites J.C., Rice B.J., Hartley P.D., Sugnet C.W., Haussler D., Rokhsar D.S. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Research **26**:342–350.

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

R Core Team (2015) R: A Language and Environment for Statistical Computing. [online] URL: https://www.r-project.org

Rajkumar A.P., Qvist P., Lazarus R., Lescai F., Ju J., Nyegaard M., Mors O., Børglum A.D., Li Q., Christensen J.H. (2015) Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genomics **16**:548.

Ramšak Ž., Baebler Š., Rotter A., Korbar M., Mozetič I., Usadel B., Gruden K. (2014) GoMapMan: Integration, consolidation and visualization of plant gene annotations within the MapMan ontology. Nucleic Acids Research **42**:1167–1175.

Rensink W., Hart A., Liu J., Ouyang S., Zismann V., Buell C.R. (2005) Analyzing the potato abiotic stress transcriptome using expressed sequence tags. Genome **48**:598–605.

Rensink W.A., Iobst S., Hart A., Stegalkina S., Liu J., Buell C.R. (2005) Gene expression profiling of potato responses to cold, heat, and salt stress. Functional and Integrative Genomics **5**:201–207.

Rensink W.A., Lee Y., Liu J., Iobst S., Ouyang S., Buell C.R. (2005) Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. BMC genomics **6**:124.

Restrepo S., Myers K.L., del Pozo O., Martin G.B., Hart a L., Buell C.R., Fry W.E., Smart C.D. (2005) Gene profiling of a compatible interaction between Phytophthora infestans and Solanum tuberosum suggests a role for carbonic anhydrase. Molecular plant-microbe interactions : MPMI **18**:913–922.

Reuveny Z., Dougall D.K., Trinity P.M. (1980) Regulatory coupling of nitrate and sulfate assimilation pathways in cultured tobacco cells. Proceedings of the National Academy of Sciences of the United States of America **77**:6670–6672.

Reyes Chin-Wo S., Lavelle D., Truco M.J., Kozik A., Michelmore R. (2016) Dovetail/in vitro Proximity Ligation Data Facilitates Analysis of an Ancient Whole Genome Triplication Event in Lactuca sativa. In: Plant and Animal Genome XXIV Conference. [online] URL: https://pag.confex.com/pag/xxiv/webprogram/Paper19305.html

Rhoads A., Au K.F. (2015) PacBio Sequencing and Its Applications. Genomics, Proteomics and Bioinformatics **13**:278–289.

Ritter E., Debener T., Barone A., Salamini F., Gebhardt C. (1991) RFLP mapping on potato chromosomes of two genes controlling extreme resistance to potato virus X (PVX). Molecular and General Genetics MGG **227**:81–85.

Rocher S., Jean M., Castonguay Y., Belzile F. (2015) Validation of genotyping-by-sequencing analysis in populations of tetraploid alfalfa by 454 sequencing. PLoS ONE **10**:e0131918.

Rodríguez F., 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**:207–219.

Ronning C.M., Stegalkina S.S., Ascenzi R.A., Bougri O., Hart A.L., Utterbach T.R., Vanaken S.E., Riedmuller S.B., White J.A., Cho J., Pertea G.M., Lee Y., Karamycheva S., Sultana R., Tsai J., Quackenbush J., Griffiths H.M., Restrepo S., Smart C.D., Fry W.E., Van Der Hoeven R., Tanksley S., Zhang P., Jin H., Yamamoto M.L., Baker B.J., Buell C.R. (2003) Comparative analyses of potato expressed sequence tag libraries. Plant physiology **131**:419–429.

Sánchez E., Garcia P.C., López-Lefebre L.R., Rivero R.M., Ruiz J.M., Romero L. (2002) Proline metabolism in response to nitrogen deficiency in French Bean plants (Phaseolus vulgaris L . cv Strike). Plant Growth Regulation **36**:261–265.

Sandelin A., Alkema W., Engström P., Wasserman W.W., Lenhard B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research **32**:D91–D94.

Schafleitner R., Gutierrez Rosales R.O., Gaudin A., Alvarado Aliaga C.A., Martinez G.N., Tincopa Marca L.R., Bolivar L.A., Delgado F.M., Simon R., Bonierbale M. (2007) Capturing candidate drought tolerance traits in two native Andean potato clones by transcription profiling of field grown plants under water stress. Plant Physiology and Biochemistry **45**:673–690.

Schatz M.C., Maron L.G., Stein J.C., Hernandez Wences A., Gurtowski J., Biggers E., Lee H., Kramer M., Antoniou E., Ghiban E., Wright M.H., Chia J., Ware D., McCouch S.R., 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**:506.

Schneeberger K., Ossowski S., Ott F., Klein J.D., Wang X., Lanz C., Smith L.M., Cao J., Fitz J., Warthmann N., Henz S.R., Huson D.H., Weigel D. (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proceedings of the National Academy of Sciences of the United States of America **108**:10249–10254.

Schönhals E.M., Ortega F., Barandalla L., Aragones A., Ruiz de Galarreta J.I., Liao J.C., Sanetomo R., Walkemeier B., Tacke E., Ritter E., Gebhardt C. (2016) Identification and reproducibility of diagnostic DNA markers for tuber starch and yield optimization in a novel association mapping population of potato (Solanum tuberosum L.). Theoretical and Applied Genetics **129**:767–785.

Shan J., Song W., Zhou J., Wang X., Xie C., Gao X., Xie T., Liu J. (2013) Transcriptome analysis reveals novel genes potentially involved in photoperiodic tuberization in potato. Genomics **102**: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., 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., 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**:2031–47.

Shelp B.J., Bown A.W., McLean M.D. (1999) Gamma-Aminobutyric Acid. Trends in Plant Science **4**:446–452.

Sierro N., Battey J.N.D., Ouadi S., Bakaher N., Bovet L., Willig A., Goepfert S., Peitsch M.C., Ivanov N. V (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nature communications **5**:3833.

Simko I., Haynes K.G., Ewing E.E., Costanzo S., Christ B.J., Jones R.W. (2004) Mapping genes for resistance to Verticillium albo-atrum in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. Molecular Genetics and Genomics **271**:522–531.

Simko I., Haynes K.G., Jones R.W. (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics **173**:2237–2245.

Simpson J.T., Wong K., Jackman S.D., Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M. (2009) ABySS : A parallel assembler for short read sequence data. Genome **19**:1117–1123.

Slater A.T., Cogan N.O.I., Forster J.W., Hayes B.J., Daetwyler H.D. (2016) Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. Plant Genome **9**:1–15.

Song Y.-S., Hepting L., Schweizer G., Hartl L., Wenzel G., Schwarzfischer A. (2005) Mapping of extreme resistance to PVY (Ry sto) on chromosome XII using anther-culture-derived primary dihaploid potato lines. Theoretical and Applied Genetics **111**:879–887.

Souvorov A., Kapustin Y., Kiryutin B., Chetvernin V., Tatusova T., Lipman D. (2010) *Gnomon--NCBI eukaryotic gene prediction tool*.

Spooner D.M. (2009) DNA barcoding will frequently fail in complicated groups: An example in

wild potatoes. American Journal of Botany **96**:1177–1189.

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

Spooner D.M., Núñez J., Trujillo G., Herrera M.D.R., Guzmán F., 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 of the United States of America **104**:19398–19403.

Swarbreck D., Wilks C., Lamesch P., Berardini T.Z., Garcia-Hernandez M., Foerster H., Li D., Meyer T., Muller R., Ploetz L., Radenbaugh A., Singh S., Swing V., Tissier C., Zhang P., Huala E. (2008) The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Research **36**:1009–1014.

Swinnen G., Goossens A., Pauwels L. (2016) Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. Trends in Plant Science **21**:506–515.

Tai H.H., Conn G., Davidson C., Platt H.W. (Bud) (2009) Arbitrary multi-gene reference for normalization of real-time PCR gene expression data. Plant Molecular Biology Reporter **27**:315–320.

Tai H.H., Goyer C., Platt H.W., De Koeyer D., Murphy A., Uribe P., Halterman D. (2013) Decreased defense gene expression in tolerance versus resistance to Verticillium dahliae in potato. Functional and Integrative Genomics **13**:367–378.

Tai H.H., Zebarth B.J. (2015) Effect of Time of Day of Sampling on Potato Foliar Gene Expression Used to Assess Crop Nitrogen Status. American Journal of Potato Research **92**:284–293.

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

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**:635–641.

Tinker N.A., Bekele W.A., Hattori J. (2016) Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. G3: Genes|Genomes|Genetics **6**:857–863.

Tiwari J.K., Siddappa S., Singh B.P., Kaushik S.K., Chakrabarti S.K., Bhardwaj V., Chandel P.

(2013) Molecular markers for late blight resistance breeding of potato: an update. Plant Breeding **132**:237–245.

Trapnell C., Hendrickson D.G., Sauvageau M., Goff L., Rinn J.L., Pachter L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology **31**:46–53.

Trapnell C., Pachter L., Salzberg S.L. (2009) TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105–1111.

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., Pimentel H., Salzberg S.L., Rinn J.L., Pachter L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols **7**:562–78.

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J., Pachter L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology **28**:511–515.

Trindade L.M., Horvath B., Bachem C., Jacobsen E., Visser R.G. (2003) Isolation and functional characterization of a stolon specific promoter from potato (Solanum tuberosum L.). Gene **303**:77–87.

Uitdewilligen J.G.A.M.L., Wolters A.M.A., D'hoop B.B., Borm T.J.A., Visser R.G.F., van Eck H.J. (2013) A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. PLoS ONE **8**:10–14.

Uribe P., Jansky S., Halterman D. (2014) Two CAPS markers predict Verticillium wilt resistance in wild Solanum species. Molecular Breeding **33**:465–476.

VanBuren R., Bryant D., Edger P.P., Tang H., Burgess D., Challabathula D., Spittle K., Hall R., Gu J., Lyons E., Freeling M., Bartels D., Ten Hallers B., Hastie A., Michael T.P., Mockler T.C. (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature **527**:508–11.

Vidal E.A., Moyano T.C., Krouk G., Katari M.S., Tanurdzic M., McCombie W., Coruzzi G.M., Gutiérrez R.A. (2013) Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in Arabidopsis thaliana roots. BMC Genomics **14**:701.

Villain P., Clabault G., Mache R., Zhou D.X. (1994) S1F Binding-Site Is Related To But Different From the Light-Responsive Gt-1 Binding-Site and Differentially Represses the Spinach Rps1 Promoter in Transgenic Tobacco. Journal of Biological Chemistry **269**:16626–16630.

Vos P.G., Uitdewilligen J.G.A.M.L., Voorrips R.E., Visser R.G.F., 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.

Wagner G.P., Kin K., Lynch V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences **131**:281–285.

Wajid B., Ekti A.R., Noor A., Serpedin E., Naeem Ayyaz M., Nounou H., Nounou M. (2013) Supersonic mib. Proceedings - IEEE International Workshop on Genomic Signal Processing and Statistics:86–87.

Wajid B., Serpedin E., Nounou M., Nounou H. (2012) MiB: A comparative assembly processing pipeline. Proceedings - IEEE International Workshop on Genomic Signal Processing and Statistics:86–89.

Wang Z., Gerstein M., Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics **10**:57–63.

Westermann D.T. (2005) Nutritional requirements of potatoes. American Journal of Potato Research **82**:301–307.

Wheeler D.L., Barrett T., Benson D.A., Bryant S.H., Canese K., Chetvernin V., Church D.M., DiCuccio M., Edgar R., Federhen S., Geer L.Y., Kapustin Y., Khovayko O., Landsman D., Lipman D.J., Madden T.L., Maglott D.R., Ostell J., Miller V., Pruitt K.D., Schuler G.D., Sequeira E., Sherry S.T., Sirotkin K., Souvorov A., Starchenko G., Tatusov R.L., Tatusova T.A., Wagner L., Yaschenko E. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research **35**:5–12.

Yang X.S., Wu J., Ziegler T.E., Yang X., Zayed a., Rajani M.S., Zhou D., Basra a. S., Schachtman D.P., Peng M., Armstrong C.L., Caldo R. a., Morrell J. a., Lacy M., Staub J.M. (2011) Gene Expression Biomarkers Provide Sensitive Indicators of in Planta Nitrogen Status in Maize. Plant Physiology **157**:1841–1852.

Yang W., Yoon J., Choi H., Fan Y., Chen R., An G. (2015) Transcriptome analysis of nitrogen-

starvation-responsive genes in rice. BMC Plant Biology **15**:31.

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

Zambelli F., Pesole G., Pavesi G. (2014) UNIT 2.11 Using Weeder, Pscan, and PscanChIP for the Discovery of Enriched Transcription Factor Binding Site Motifs in Nucleotide Sequences. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc., p 2.11.1-2.11.31.

Zebarth B.J., Milburn P.H. (2003) Spatial and temporal distribution of soil inorganic nitrogen concentration in potato hills. Canadian Journal of Soil Science **7**:183–195.

Zebarth B.J., Rosen C.J. (2007) Research Perspective On Nitrogen BMP Development for Potato. American Journal Of Potato Research **84**:3–18.

Zebarth B.J., Tai H., Luo S., Millard P., Koeyer D. De, Li X.-Q., Xion X. (2011) Differential gene expression as an indicator of nitrogen sufficiency in field-grown potato plants. Plant Soil **345**:387–400.

Zebarth B.J., Tai G., Tarn R., de Jong H., Milburn P.H. (2004) Nitrogen use efficiency characteristics of commercial potato cultivars. Canadian Journal of Plant Science **84**:589–598.

Zebarth B.J., Tremblay N., Fournier P., Leblon B., Rees H. (2003) Mapping Spatial Variation in Potato Nitrogen Status Using the " N Sensor ." Acta Horticulturae **627**:267–273.

Zhang N., Yang J., Wang Z., Wen Y., Wang J., He W., Liu B., Si H., Wang D. (2014) Identification of novel and conserved microRNAs related to drought stress in potato by deep sequencing. PLoS ONE **9**

Zhao W., Yang X., Yu H., Jiang W., Sun N., Liu X., Liu X., Zhang X., Wang Y., Gu X. (2015) RNA-Seq-Based Transcriptome Profiling of Early Nitrogen Deficiency Response in Cucumber Seedlings Provides New Insight into the Putative Nitrogen Regulatory Network. Plant and Cell Physiology **56**:455–467.

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

Zolotarov Y., Strömvik M. (2015) De Novo Regulatory Motif Discovery Identifies Significant

Motifs in Promoters of Five Classes of Plant Dehydrin Genes. Plos One **10**:1–19.

Zou C., Wang P., Xu Y. (2016) Bulked sample analysis in genetics, genomics and crop improvement. Plant Biotechnology Journal **14**:1941–1955.

Zuluaga A.P., Solé M., Lu H., Góngora-Castillo E., Vaillancourt B., Coll N., Buell C.R., Valls M. (2015) Transcriptome responses to Ralstonia solanacearum infection in the roots of the wild potato Solanum commersonii. BMC genomics **16**:246.

Appendices

Supplementary Table 3.1 Petiole nitrate concentrations and SPAD measurements for all cultivars and both time-points.

Date	Cultivar	N rate [kg N ha^{-1}]	Replicate Number	Petiole Nitrate Concentration [mg/g]	SPAD readings
2012-07-25	Russet Burbank	0	R1	1.2	35.8
2012-07-25	Russet Burbank	0	R2	3.4	38.7
2012-07-25	Russet Burbank	0	R3	3.0	37.1
2012-07-25	Russet Burbank	0	R4	8.8	39.5
2012-07-25	Russet Burbank	180	R1	25.5	37.4
2012-07-25	Russet Burbank	180	R2	25.1	37.3
2012-07-25	Russet Burbank	180	R3	23.8	37.2
2012-07-25	Russet Burbank	180	R4	27.4	40.4
2012-07-25	Shepody	0	R1	3.3	36.0
2012-07-25	Shepody	0	R2	4.1	32.5
2012-07-25	Shepody	0	R3	0.7	30.7
2012-07-25	Shepody	0	R4	4.6	34.2
2012-07-25	Shepody	180	R1	24.4	33.3
2012-07-25	Shepody	180	R2	25.4	37.8
2012-07-25	Shepody	180	R3	22.9	38.5
2012-07-25	Shepody	180	R4	23.3	38.6
2012-07-25	Atlantic	0	R1	0.5	33.5
2012-07-25	Atlantic	0	R2	1.9	37.0
2012-07-25	Atlantic	0	R3	1.0	35.7
2012-07-25	Atlantic	0	R4	2.0	37.4
2012-07-25	Atlantic	180	R1	22.2	35.7
2012-07-25	Atlantic	180	R2	18.0	35.1
2012-07-25	Atlantic	180	R3	21.5	36.4
2012-07-25	Atlantic	180	R4	18.8	36.8
2012-08-08	Russet Burbank	0	R1	0.3	34.9
2012-08-08	Russet Burbank	0	R2	0.4	37.2
2012-08-08	Russet Burbank	0	R3	0.9	36.4
2012-08-08	Russet Burbank	0	R4	5.3	38.6
2012-08-08	Russet Burbank	180	R1	23.1	37.0
2012-08-08	Russet Burbank	180	R2	23.5	38.8
2012-08-08	Russet Burbank	180	R3	21.6	38.8
2012-08-08	Russet Burbank	180	R4	19.4	39.8
2012-08-08	Shepody	0	R1	2.4	30.3
2012-08-08	Shepody	0	R2	1.8	31.8
2012-08-08	Shepody	0	R3	0.4	26.2
2012-08-08	Shepody	0	R4	0.8	28.8
2012-08-08	Shepody	180	R1	22.8	35.0
2012-08-08	Shepody	180	R2	23.2	35.6
2012-08-08	Shepody	180	R3	23.3	36.8
2012-08-08	Shepody	180	R4	23.9	35.1
2012-08-08	Atlantic	0	R1	0.2	31.5
2012-08-08	Atlantic	0	R2	0.2	33.6

Date	Cultivar	N rate [kg N ha^{-1}]	Replicate Number	Petiole Nitrate Concentration [mg/g]	SPAD readings
2012-08-08	Atlantic	0	R3	0.2	33.9
2012-08-08	Atlantic	0	R4	0.7	32.8
2012-08-08	Atlantic	180	R1	21.4	36.0
2012-08-08	Atlantic	180	R2	26.5	34.9
2012-08-08	Atlantic	180	R3	28.5	35.7
2012-08-08	Atlantic	180	R4	19.5	36.5

***Supplementary Table 3.2*:** Plant dry biomass and fresh tuber yields for all cultivars at harvest.

Cultivar	N rate [kg N ha^{-1}]	Replicate Number	Plant Dry Biomass [t/ha]	Fresh Tuber Yield [t/ha]
Russet Burbank	0	R1	6.82	36.0
Russet Burbank	0	R2	8.64	33.7
Russet Burbank	0	R3	6.98	37.6
Russet Burbank	0	R4	8.45	29.8
Russet Burbank	180	R1	7.75	34.8
Russet Burbank	180	R2	8.03	44.9
Russet Burbank	180	R3	9.74	31.1
Russet Burbank	180	R4	10.54	42.9
Shepody	0	R1	8.81	32.6
Shepody	0	R2	7.86	28.3
Shepody	0	R3	7.18	28.1
Shepody	0	R4	8.42	35.5
Shepody	180	R1	8.29	35.6
Shepody	180	R2	10.03	37.3
Shepody	180	R3	9.90	36.9
Shepody	180	R4	8.74	34.4
Atlantic	0	R1	7.81	37.0
Atlantic	0	R2	11.11	43.1
Atlantic	0	R3	9.81	39.8
Atlantic	0	R4	8.53	42.4
Atlantic	180	R1	9.95	45.7
Atlantic	180	R2	11.42	43.2
Atlantic	180	R3	10.64	41.1
Atlantic	180	R4	10.31	43.4

Supplementary Table 3.3: Total number of differentially expressed genes for different potato cultivars.

***S. tuberosum* cultivar**	**Time-point 1 July 25, 2012**		**Time-point 2 Aug. 8, 2012**	
	Over-expressed	*Under-expressed*	*Over-expressed*	*Under-expressed*
Shepody	182	35	218	52
Russet Burbank	64	47	116	18
Atlantic	393	33	149	40

Supplementary Table 3.4: Genes that were differentially expressed in only one of the two time points.

Time point	***Gene ID***	***Gene Description and Interpro Domain[$]***
Over expressed genes		
2012-07-25	Sotub01g007180	AMP-dependent synthetase and ligase; IPR011614 Catalase, N-terminal
2012-07-25	Sotub02g012390	Coiled-coil domain-containing protein 109A; IPR006769 Protein of unknown function DUF607
2012-07-25	Sotub03g018730	Glutamate dehydrogenase; IPR014362 Glutamate dehydrogenase
2012-07-25	Sotub03g023340	BTB/POZ domain-containing protein; IPR000197 Zinc finger, TAZ-type
2012-07-25	Sotub03g031100	Heat shock protein; IPR013126 Heat shock protein 70
2012-07-25	Sotub04g025700	ASR4 protein (Fragment); IPR003496 ABA/WDS induced protein
2012-07-25	Sotub04g035440	Cellular retinaldehyde-binding/triple function C-terminal; IPR001251 Cellular retinaldehyde-binding/triple function, C-terminal
2012-07-25	Sotub05g007580	Myb family transcription factor; IPR006447 Myb-like DNA-binding region, SHAQKYF class
2012-07-25	Sotub05g008150	3-ketoacyl-CoA synthase; IPR012392 Very-long-chain 3-ketoacyl-CoA synthase
2012-07-25	Sotub06g006790	Plant-specific domain TIGR01615 family protein; IPR006502 Protein of unknown function DUF506, plant
2012-07-25	Sotub06g011740	Gamma-glutamyl phosphate reductase; IPR005766 Delta l-pyrroline-5-carboxylate synthetase
2012-07-25	Sotub06g023090	Solute carrier family 2, facilitated glucose transporter member 3; IPR003663 Sugar/inositol transporter
2012-07-25	Sotub06g027390	Cysteine proteinase inhibitor; IPR006043 Xanthine/uracil/vitamin C permease
2012-07-25	Sotub08g005550	PII uridylyl-transferase; IPR002912 Amino acid-binding ACT
2012-07-25	Sotub11g007070	Plant-specific domain TIGR01615 family protein; IPR006502 Protein of unknown function DUF506, plant
2012-07-25	Sotub11g007100	Plant-specific domain TIGR01615 family protein; IPR006502 Protein of unknown function DUF506, plant
2012-07-25	Sotub11g020550	Hexokinase 6; IPR001312 Hexokinase
2012-08-08	Sotub02g015720	FAD binding domain-containing protein; IPR006094 FAD linked oxidase, N-terminal
2012-08-08	Sotub02g021000	Plant-specific domain TIGR01615 family protein; IPR006502 Protein of unknown function DUF506, plant
2012-08-08	Sotub02g031260	Arabinogalactan
2012-08-08	Sotub03g012290	Kunitz-type protease inhibitor; IPR002160 Proteinase inhibitor I3, Kunitz legume
2012-08-08	Sotub03g012340	Kunitz-type protease inhibitor; IPR002160 Proteinase inhibitor I3, Kunitz legume
2012-08-08	Sotub03g012360	Proteinase inhibitor II; IPR003465 Proteinase inhibitor I20, Pin2
2012-08-08	Sotub03g015880	Kunitz trypsin inhibitor 4; IPR011065 Kunitz inhibitor ST1-like
2012-08-08	Sotub03g015970	Aspartic protease inhibitor 1; PR002160 Proteinase inhibitor I3, Kunitz legume

Time point	*Gene ID*	*Gene Description and Interpro Domain*[$]
2012-08-08	Sotub03g023330	Peptide methionine sulfoxide reductase MsrA; IPR002569 Methionine sulphoxide reductase A
2012-08-08	Sotub03g035920	Taurine catabolism dioxygenase TauD/TfdA; IPR003819 Taurine catabolism dioxygenase TauD/TfdA
2012-08-08	Sotub04g028270	Phospho-2-dehydro-3-deoxyheptonate aldolase 1; IPR002480 DAHP synthetase, class II
2012-08-08	Sotub05g018510	GDSL esterase/lipase At2g04570; IPR001087 Lipase, GDSL
2012-08-08	Sotub05g021450	Glucose-6-phosphate/phosphate translocator 2; IPR004696 Tpt phosphate/phosphoenolpyruvate translocator
2012-08-08	Sotub05g027780	High affinity sulfate transporter 2; IPR001902 Sulphate anion transporter
2012-08-08	Sotub06g025010	Cortical cell-delineating protein; IPR013770 Plant lipid transfer protein and hydrophobic protein, helical
2012-08-08	Sotub06g026740	Chlorophyll a-b binding protein 4, chloroplastic; IPR001344 Chlorophyll A-B binding protein
2012-08-08	Sotub06g027990	Unknown Protein
2012-08-08	Sotub06g030410	Beta-D-glucosidase; IPR001764 Glycoside hydrolase, family 3, N-terminal
2012-08-08	Sotub06g030610	Cysteine proteinase inhibitor; IPR000010 Proteinase inhibitor I25, cystatin
2012-08-08	Sotub07g016530	Cellulose synthase-like protein H1; IPR005150 Cellulose synthase
2012-08-08	Sotub07g016550	UDP glucosyltransferase; IPR002213 UDP-glucuronosyl/UDP-glucosyltransferase
2012-08-08	Sotub07g016570	1-aminocyclopropane-1-carboxylate oxidase; IPR005123 Oxoglutarate and iron-dependent oxygenase
2012-08-08	Sotub08g024210	Exostosin family protein; IPR004263 Exostosin-like
2012-08-08	Sotub08g028270	Methanol inducible protein
2012-08-08	Sotub09g008430	Threonine dehydratase biosynthetic; IPR005787 Threonine dehydratase I
2012-08-08	Sotub09g023600	Homocysteine s-methyltransferase; IPR003726 Homocysteine S-methyltransferase
2012-08-08	Sotub09g026640	Proteinase inhibitor I; IPR000864 Proteinase inhibitor I13, potato inhibitor I
2012-08-08	Sotub09g028690	Selenium binding protein; IPR008826 Selenium-binding protein
2012-08-08	Sotub09g031120	Unknown Protein; IPR006706 Extensin-like region
2012-08-08	Sotub10g021050	UDP glucosyltransferase; IPR002213 UDP-glucuronosyl/UDP-glucosyltransferase
2012-08-08	Sotub11g024220	Superoxide dismutase; IPR001424 Superoxide dismutase, copper/zinc binding
2012-08-08	Sotub12g007850	Cytosol aminopeptidase family protein; IPR011356 Peptidase M17, leucyl aminopeptidase
2012-08-08	Sotub12g008260	Unknown Protein
2012-08-08	Sotub12g028670	Cation transport regulator-like protein 2; IPR006840 ChaC-like protein
Under expressed genes		
2012-07-25	Sotub02g017430	Purine permease family protein; IPR004853 Protein of unknown function DUF250
2012-07-25	Sotub08g025870	Primary amine oxidase; IPR000269 Copper amine oxidase
2012-07-25	Sotub09g010630	Hydrolase alpha/beta fold family protein; IPR000073 Alpha/beta hydrolase fold-1
2012-08-08	Sotub12g012740	Chloroplast lipocalin; IPR000566 Lipocalin-related protein and Bos/Can/Equ allergen

$ Gene descriptions (including InterPro domains) obtained from the ITAG1.0 annotation system (The Tomato Genome Consortium 2012)

Supplementary Table 4.1: Summary of re-sequencing trimming, filtering and alignment statistics.

Mapped stats	*S. tuberosum* subsp. *andigena*		*S. stenotomum* subsp. *goniocalyx*	
Total raw unpaired sequences	388,360,400		284,592,584	
Total unpaired filtered sequences	366,096,862		259,326,224	
Reads mapped *	307,680,121	(84.0%)	227,800,316	(87.8%)
Reads mapped and paired *	291,389,134	(79.6%)	218,793,596	(84.4%)
Reads unmapped *	58,416,741	(16.0%)	31,525,908	(12.2%)
Reads properly paired *	282,442,614	(77.1%)	212,076,888	(81.8%)
Total paired filtered sequences	183,048,431		129,663,112	
Inward oriented pairs †	141,562,983	(77.3%)	106,392,901	(82.1%)
Outward oriented pairs †	702,943	(0.4%)	287,192	(0.2%)
Pairs on different chromosome †	3,132,263	(1.7%)	2,527,666	(1.9%)
Pairs with other orientation †	206,422	(0.1%)	165,661	(0.1%)
Average insert size	281		309	
Total filtered and trimmed seq. length	52,838,079,576 nt		37,347,721,648 nt	
Bases mapped [nt] ††	45,114,571,684	(85.4%)	32,900,789,436	(88.1%)
Average length	144		144	
Average depth	62.23×		45.38×	
Coverage and depth statistics				
Reference genome nt with 0 depth°	98,594,721 nt	(13.6%)	104,993,531 nt	(14.5%)
Reference genome nt with >1× depth°	626,422,663 nt	(86.4%)	620,023,853 nt	(85.5%)
Reference genome nt with >4× depth°	616,341,343 nt	(85.0%)	605,513,003 nt	(83.5%)

* As a percentage of the total unpaired filtered sequences
† As a percentage of total paired filtered sequences
†† As a percentage of total filtered and trimmed sequence length
° As a percentage of the total length of the reference genome (v4.03)

Supplementary Table 4.2: Summary of genes and function of the top two CNV-affected gene clusters in both landraces.

		PGSC Gene ID	*Gene Name*	*Functional Annotation*
***S. tuberosum* subsp. *andigena* CNV-enriched gene clusters**				
Chromosome 12 [600000 - 800000]				
1	DL	PGSC0003DMG400015312	AP2 domain-containing transcription factor	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding.
2	DL	PGSC0003DMG400015313	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process.
3	DL	PGSC0003DMG400015316	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
4	DL	PGSC0003DMG400015345	Conserved gene of unknown function	IP: Leucine-rich repeat.
5	DL	PGSC0003DMG400015347	Conserved gene of unknown function	IP: Leucine-rich repeat.
6	DL	PGSC0003DMG400015348	Conserved gene of unknown function	IP: Leucine-rich repeat.
7	DL	PGSC0003DMG400015349	Disease resistance protein	IP: Leucine-rich repeat.
8	DL	PGSC0003DMG400015350	Monooxygenase	IP: Monooxygenase.
9	DL	PGSC0003DMG400015351	Conserved gene of unknown function	IP: Leucine-rich repeat.
10	DL	PGSC0003DMG400015352	Serine/threonine-protein kinase bri1	IP: Leucine-rich repeat; serine-threonine protein kinase; lycopene beta/epsilon cyclase.
11	DL	PGSC0003DMG400015353	Hcr2-p3	IP: Leucine-rich repeat; serine-threonine protein kinase. GO: receptor-like protein kinase.

		PGSC Gene ID	*Gene Name*	*Functional Annotation*
12	DP	PGSC0003DMG400015354	Conserved gene of unknown function	IP: Leucine-rich repeat; serine-threonine protein kinase. GO: receptor-like protein kinase.
13	DL	PGSC0003DMG400015397	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
14	DP	PGSC0003DMG400037662	Conserved gene of unknown function	IP: Serine/threonine-protein kinase; Leucine-rich repeat, LRR receptor-like kinase.
15	DL	PGSC0003DMG400041562	AP2 domain-containing transcription factor	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
16	DL	PGSC0003DMG400042371	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, family 5. GO: carbohydrate metabolic process, extracellular region, binding.
17	DL	PGSC0003DMG400043330	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
18	DL	PGSC0003DMG401015315	Dehydration responsive element binding protein	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
19	DL	PGSC0003DMG402015315	Fiber protein Fb34	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
Chromosome 10 [56000000 - 56200000]				
1	DL	PGSC0003DMG400028147	Protein kinase	IP: Leucine-rich repeat; serine-threonine protein kinase. GO: receptor-like protein kinase.
2	DP	PGSC0003DMG400028148	Conserved gene of unknown function	
3	DP	PGSC0003DMG400028149	ATEXO70A2 (exocyst subunit EXO70 family protein A2), protein binding	IP: Exo70 exocyst complex. GO: exocyst complex component 7-like.
4	DP	PGSC0003DMG400028150	Conserved gene of unknown function	
5	DP	PGSC0003DMG400028151	VAMP protein SEC22	IP: Late embryogenesis abundant protein, group 2. GO: uncharacterized protein at1g08160-like.
6	DP	PGSC0003DMG400028152	Hin1	IP: Late embryogenesis abundant protein, group 2; Syntaxin.
7	DP	PGSC0003DMG400028153	Hin1	IP: Late embryogenesis abundant protein, group 2; Syntaxin.
8	DP	PGSC0003DMG400028154	Cytokinin riboside 5'-monophosphate phosphoribohydrolase LOG7	IP: Conserved hypothetical protein CHP00730. GO: cytokinin riboside 5 -monophosphate phosphoribohydrolase log7-like.
9	DP	PGSC0003DMG400028155	Conserved gene of unknown function	IP: Protein of unknown function DUF623. GO: ovate family protein 6.
10	DL/ DP	PGSC0003DMG400028234	Coatomer subunit beta-1	IP: Coatomer; Clathrin/coatomer adapter, adaptin-like; Armadillo-like helical. GO: coatomer subunit beta-1-like.
11	DP	PGSC0003DMG400028235	Hin1	IP: Late embryogenesis abundant protein, group 2; Syntaxin.
12	DP	PGSC0003DMG400028236	Conserved gene of unknown function	IP: ATP-depenent Clp protease-related.
13	DP	PGSC0003DMG400028237	S locus F-box (SLF)-S5 protein	IP: F-box domain, cyclin-like; F-box domain, Skp2-like; F-box associated domain.
14	DP	PGSC0003DMG400028238	Thioredoxin peroxidase	IP: Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen; Peroxiredoxin; Thioredoxin fold. GO: 2-cys peroxiredoxin.
15	DP	PGSC0003DMG400028239	HVA22 i	IP: TB2/DP1/HVA22 related protein; HVA22-like protein.
16	DP	PGSC0003DMG400040827	Plant-specific domain TIGR01568 family protein	IP: Protein of unknown function DUF623.
17	DP	PGSC0003DMG400043673	Conserved gene of unknown function	IP: Protein of unknown function DUF623.
18	DP	PGSC0003DMG400045504	Conserved gene of unknown function	IP: Protein of unknown function DUF623.

		PGSC Gene ID	*Gene Name*	*Functional Annotation*
***S. stenotomum* subsp. *goniocalyx* CNV-enriched gene clusters**				
Chromosome 12 [600000-800000]				
1	DL	PGSC0003DMG400015312	AP2 domain-containing transcription factor	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding.
2	DL	PGSC0003DMG400015313	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process.
3	DL	PGSC0003DMG400015316	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
4	DL	PGSC0003DMG400015345	Conserved gene of unknown function	IP: Leucine-rich repeat.
5	DL	PGSC0003DMG400015347	Conserved gene of unknown function	IP: Leucine-rich repeat.
6	DL	PGSC0003DMG400015348	Conserved gene of unknown function	IP: Leucine-rich repeat.
7	DL	PGSC0003DMG400015349	Disease resistance protein	IP: Leucine-rich repeat.
8	DL	PGSC0003DMG400015350	Monooxygenase	IP: Monoxygenase.
9	DL	PGSC0003DMG400015351	Conserved gene of unknown function	IP: Leucine-rich repeat.
10	DL	PGSC0003DMG400015352	Serine/threonine-protein kinase bri1	IP: Leucine-rich repeat; serine-threonine protein kinase; lycopene beta/epsilon cyclase.
11	DL	PGSC0003DMG400015353	Hcr2-p3	IP: Leucine-rich repeat; serine-threonine protein kinase. GO: receptor-like protein kinase.
12	DL	PGSC0003DMG400015355	Monooxygenase	IP: Monooxygenase.
13	DL	PGSC0003DMG400015397	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
14	DL	PGSC0003DMG400037662	Conserved gene of unknown function	IP: Serine/threonine-protein kinase; Leucine-rich repeat, LRR receptor-like kinase.
15	DL	PGSC0003DMG400041562	AP2 domain-containing transcription factor	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
16	DL	PGSC0003DMG400042189	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
17	DL	PGSC0003DMG400042371	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, family 5. GO: carbohydrate metabolic process, extracellular region, binding.
18	DL	PGSC0003DMG400043330	Mannan endo-1,4-beta-mannosidase 1	IP: Glycoside hydrolase, catalytic core. GO: carbohydrate metabolic process, extracellular region, binding.
19	DL	PGSC0003DMG401015315	Dehydration responsive element binding protein	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
20	DL	PGSC0003DMG402015315	Fiber protein Fb34	IP: Pathogenesis-related transcriptional factor/ERF, DNA-binding; DNA-binding, integrase type.
Chromosome 4 [4600000- 4800000]				
1	DL	PGSC0003DMG400011529	R2	IP: Disease resistance protein.
2	DL	PGSC0003DMG400011517	HJTR2GH1 protein	IP: Disease resistance protein.
3	DL	PGSC0003DMG400011518	R2 late blight resistance protein	IP: Triosephosphate isomerase.
4	DL	PGSC0003DMG400011521	EDNR2GH5 protein	IP: Disease resistance protein.
5	DP	PGSC0003DMG400011523	HJTR2GH1 protein	IP: Disease resistance protein.
6	DP	PGSC0003DMG400011525	EDNR2GH4 protein	IP: Disease resistance protein.
7	DP	PGSC0003DMG400011527	SNKR2GH5 protein	IP: Disease resistance protein.
8	DL	PGSC0003DMG400011533	Late blight resistance protein	
9	DL	PGSC0003DMG400032548	DECOY	IP: Ribosomal protein L46. GO: ribosome.
10	DL	PGSC0003DMG400032578	HJTR2GH1 protein	IP: Disease resistance protein.
11	DL	PGSC0003DMG400032584	R2 late blight resistance protein	IP: Disease resistance protein.
12	DL	PGSC0003DMG401011506	Gene of unknown function	
13	DL	PGSC0003DMG401011522	HJTR2GH1 protein	IP: NB-ARC.
14	DP	PGSC0003DMG401011526	SNKR2GH2 protein	IP: NB-ARC.

		PGSC Gene ID	*Gene Name*	*Functional Annotation*
15	DL	PGSC0003DMG402011506	Myosin head, motor region	IP: Myosin. GO: vacuole.
16	DL	PGSC0003DMG402011522	EDNR2GH4 protein	IP: NB-ARC.
17	DP	PGSC0003DMG402011526	EDNR2GH8 protein	IP: Disease resistance protein.

Abbreviation key:
DL=Deletion, DP=Duplication, IP=InterPro, GO=Gene Ontology, PGSC=Potato Genome Sequencing Consortium