EXPLORATION OF THE GENETIC DIVERSITY OF CULTIVATED POTATO AND ITS WILD PROGENITORS (*SOLANUM* SECT. *PETOTA*) WITH INSIGHTS INTO POTATO DOMESTICATION AND GENOME EVOLUTION

By

Michael Alan Hardigan

A DISSERTATION

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

Crop and Soil Sciences - Doctor of Philosophy

2016

ABSTRACT

EXPLORATION OF THE GENETIC DIVERSITY OF CULTIVATED POTATO AND ITS WILD PROGENITORS (*SOLANUM* SECT. *PETOTA*) WITH INSIGHTS INTO POTATO DOMESTICATION AND GENOME EVOLUTION

By

Michael Alan Hardigan

Cultivated potato (*Solanum tuberosum* L.) is a clonally propagated autotetraploid (2n=4x=48) with high potential for genetic improvement due to the expansive genetic diversity of its related germplasm. This diversity includes high levels of heterozygosity but also deleterious genetic load in cultivated clones, multiple groups of genetically distinct landrace sub-species, and over 100 related tuber-bearing species that can introduce heterosis and adaptation. To generate a better understanding of the landscape of potato genetic diversity, studies were conducted to (1) explore the genetic relationships between tuber-bearing *Solanum* species, landraces, and cultivars, (2) investigate the role of structural variation in cultivated potato genome evolution and diversity, and (3) identify key loci that were selected during domestication of potato. Phylogenetic analysis of tuber-bearing *Solanum* species was examined in the first application of genome-wide nuclear single nucleotide polymorphism (SNP) markers with a diversity panel of potato species from the USDA Potato Gene Bank core collection. Genotyping a panel of 75 accessions representing 25 species and 213 cultivated tetraploid clones using 5,023 polymorphic SNPs supported previous taxonomic placement of wild potato germplasm, while reinforcing the theory of a Peruvian domestication origin for potato. Several genes involved in carbohydrate metabolism or with potential roles in tuber development were found to contain diverging allele frequencies in the wild and cultivated potato groups suggesting they may have been key loci involved in domestication.

A comprehensive analysis of genomic variation (sequence and structural variation) using 30-70x sequence coverage in a panel of 12 monoploid/doubled monoploid potatoes derived from primitive South American diploid landraces was performed. Extensive copy number variation (CNV) was found to impact over 30% of the potato genome and nearly 30% of genes, with widespread deletion affecting lowly expressed sequences, and enriched duplication within gene families that function in adaptation and environmental response. This study revealed CNV is equally relevant to sequence-level variation (SNPs, insertions/deletions) in its contribution to deleterious mutation load and gene evolution in the potato genome. The study also demonstrated that the extensive structural heterogeneity present in tetraploid genotypes is present in their diploid progenitors, and thus, is not the result of polyploidy alone. A diversity panel was sequenced at 8x to 16x coverage to enable genome-wide comparison of allele composition in 23 tetraploid potato cultivars, 20 wild species and 20 landrace progenitors. Analysis of selection signatures including F_{ST}, Tajima's D, reduced nucleotide diversity and allele frequency differences yielded a set of domestication candidate genes including several regulating carbohydrate pathways, glycoalkaloid biosynthesis and stress responses. Comparable levels of total sequence diversity were present in elite cultivars and South American landraces, yet significant genetic variance between landrace populations was re-partitioned as single-clone heterozygosity in cultivars, confirming the hypothesis that elite varieties represent a highly heterozygous group lacking significant population structure. These findings may assist the ability of potato breeders to exploit key performance-related loci and heterosis in the future.

Dedicated to my best friend and brother, Christopher J. Hardigan.

ACKNOWLEDGEMENTS

I thank Dr. C. Robin Buell as a mentor during my post-graduate research.

I thank the members of my graduate committee, Drs. C. Robin Buell, David Douches, James Hancock, Wayne Loescher, and Amy Iezzoni for dedicating their time and experience toward my development as a graduate student and young scientist.

I thank my NSF potato vigor project collaborators Parker Laimbeer, Dr. Richard Veilleux and Dr. Jiming Jiang for their substantial contributions to my own published research, and members of the Buell lab for lending their expertise in support of these studies.

I thank Kevin Childs and John Hamilton for providing mentorship during my early months of bioinformatics training.

I thank my girlfriend Jane Ames for her patience during the last three years of my graduate research.

I thank my parents for their life-long support of my education and happiness.

I owe my greatest debt to the many great scientists whose knowledge and hard work laid the foundation upon which my own research was built. It is my hope that the work herein helps to raise this foundation a bit higher.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Tuber Bearing *Solanum* Species

Tuber-bearing *Solanum* species (*Solanum* section *Petota*) represent a large cross-section of an agriculturally important genus. They include over 100 species (Spooner, 2009), of which approximately 70% are diploid and the remainder include triploids, tetraploids, pentaploids, and hexaploids. The cultivated potato (*Solanum tuberosum* L.) itself contains multiple diverse sub-species, or cultivated groups within its native range that includes diploids, triploids, autotetraploids, and hybrid allopolyploids (Ovchinnikova et al., 2011). Although they possess a wide distribution, the majority of potato species, including putative progenitors of *S. tuberosum,* are concentrated in the Andes Mountains region of Argentina, Bolivia and Peru (Spooner et al., 2005), with a second center of diversity in Mexico containing more phylogenetically primitive species (Hijmans and Spooner, 2001; Hijmans, 2002). Studies of the systematics of this group, and the ability to resolve conclusively the evolutionary history of its cultivated lineage have historically been challenged by overlapping morphologies, sparse molecular data, and hybrid interactions (Hijmans and Spooner, 2001; Spooner, 2009). Despite these issues, wild relatives of potato have shown promise in contributing useful diversity to the cultivated gene pool (Jansky et al., 2013), most notably providing durable late blight resistance to *Phytophthora infestans* following the Irish potato famine in the mid-1800s (Ross, 1966). Most species can cross freely with cultivated potato (*S. tuberosum*), or by circumventing hybridization barriers (Camadro, 2010), allowing direct access to a large pool of diversity. Although they possess small, inedible tubers, wild potatoes have been observed to produce strong heterotic effects when crossed with cultivated clones (Bradshaw et al., 2006; Hermundstad and Peloquin, 1986), suggesting they can be a valuable resource not only

for introducing adaptive genes, but also in overcoming the genetic bottlenecks that have resulted from potato's domestication and subsequent breeding.

Origins of Cultivated Potatoes

Cultivated potato of North America (*S. tuberosum* L.) is a clonally propagated autotetraploid (2*n*=4*x*=48) species from the New World, widely grown for its production of underground storage tubers in which the plant partitions large quantities of carbohydrates, as well as vitamins and protein. Potatoes were first domesticated as diploid landraces (*S. tuberosum Stenotomum* group; 2*n*=2*x*=24, now collapsed into the *Andigenum* group) approximately 8,000-10,000 years ago from wild species native to the Andean highlands of southern Peru (Brush et al., 1981; Spooner et al, 2005). Potato evolutionary history included at least one autopolyploidization event yielding Andean cultivated tetraploids (*S. tuberosum Andigena* group; 2*n*=4*x*=48, now collapsed into the *Andigenum* group) and a proposed migration of a landrace branch to the coasts of central and southern Chile. During this migration, cross-species hybridization occurred as shown by a near universal representation of wild Bolivian *S. tarijense* cytoplasm in Chilean landraces and their derived cultivars (Hosaka, 2003; Spooner et al, 2005). In the coastal southern latitudes of Chile, potatoes adapted to cultivation under long-day growing conditions and formed the smaller, secondary tetraploid genetic group (*S. tuberosum Chilotanum* group; 2n=4x=48). These long-day conditions are more similar to those found in North America and Northern Europe, and the *S. tuberosum Chilotanum* group currently represents the bulk of the cultivated potato gene pool on both of these continents (Hosaka and Hanneman, 1988) (Table 0.1, Figure 0.1).

Table 0.1. ***Overview of cultivated potato species, their sub-groups, ploidy and geographic distributions within South America.***

Species	Subspecies Group	Former Groups	Ploidy	Distribution	Elevation (masl)[b]
Solanum tuberosum	Andigenum (upland)	Phureja	2x	Venezuela, Colombia, Ecuador, Peru, Bolivia, northern Argentina	2000-4600
		Stenotomum	2x		
		Chaucha	3x		
		Andigena	4x		
	Chilotanum[a] (lowland)	Tuberosum	4x	Chonos and Guaitecas Archipelagos (southern Chile), adjacent mainland	0-500
Solanum ajanhuiri	-	-	2x	Peru, Bolivia	3600-4100
Solanum juzepczukii	-	-	3x	Peru, Bolivia	3600-4400
Solanum curtilobum	-	-	5x	Peru, Bolivia	3600-4300

[a] North American and European cultivars are derived primarily from a Chilean tetraploid landrace background (*Chilotanum*), which includes long-day adapted varieties compared to the primarily short-day adapted *Andigenum* group.
[b] masl, abbreviation for meters above sea level.

Potential Limitations of the North American Cultivar Genetic Base

Following introduction to Europe in the 16th century, potato cultivation has become globally widespread. It is now the third most important food crop worldwide in terms of direct human consumption after wheat and rice, and the most important vegetable species (Birch et al., 2012). Stabilized production levels support a major processing industry in the United States that surpassed the fresh consumption market several decades ago (Lin et al., 2001). Compared to Europe and North America, potato cultivation has witnessed substantial growth in the developing world, and is now a key species supporting food security in Asia and sub-Saharan Africa where yields still tend to be a fraction of that in developed countries (Birch et al., 2012; Scott and Suarez, 2012). Major concerns for ensuring sustainable improvements in global crop yields include adapting varieties to withstand multiple biotic and abiotic stresses (Chakraborty et al., 2011; Newton et al., 2011; Wang and Frei, 2011) and expanding the narrow genetic basis within the cultivated gene pool of potato that will be required to make further genetic gains in both yield and quality (Hirsch et al., 2013; Mendoza and Haynes, 1974). Genetic diversity constraints in cultivated U.S. potato germplasm are attributed to (i) the limited set of European varieties that survived the late blight epidemic of the 1840s (Irish potato famine) which imposed a genetic bottleneck and (ii) the small number Chilean landraces that served as the foundation of germplasm in most U.S. breeding programs during the 19th century (Love, 1999).

Despite possessing what is considered to be a narrow genetic base (Mendoza and Haynes, 1974; Plaisted and Hoopes, 1989), North American potato cultivars have access to a large pool of diversity. Indeed, individual cultivars possess high levels of intra-genome allelic

variation, i.e., heterozygosity (Potato Genome Sequencing Consortium, 2011; Hirsch et al., 2013), which is thought to be correlated with potato plant vigor and yield (Mendoza and Haynes, 1974). Derived primarily from autotetraploid Chilean landraces (*S. tuberosum Chilotanum* group), cultivars represent only one of several South American *S. tuberosum* sub-groups, now diverged from their counterparts growing in the Andes Mountains of Peru and Bolivia (*S. tuberosum Andigenum* group), which include diploids (former *Phureja* and *Stenotomum* groups), triploids (former *Chaucha* group) and tetraploids (former *Andigena* group), and allopolyploid hybrids (Ovchinnikova et al., 2011). The availability of compatible genetic groups in distinct geographic centers of diversification, each with substantial diversity among their respective populations, offers ample raw material for exploiting heterosis in breeding vigorous potato hybrids. However, the most significant pool of diversity available for direct potato breeding stems from its large body of wild tuber-bearing relatives, most of which can be introgressed directly into cultivated tetraploids, or using strategies that circumvent hybridization barriers (Jansky et al., 2013). The complete group of tuber-bearing *Solanum* species (*Solanum* sect. *Petota*) contains over 100 species (Spooner, 2009) with habitats distributed from the southwestern United States to southern Chile (Spooner and Salas, 2006), giving rise to populations adapted to a broad assortment of abiotic stress factors, pests and pathogens (Spooner et al., 1994). Hence, there is abundant allelic diversity for the expansion of cultivated heterotic potential including incorporating novel adaptive traits into North American breeding populations.

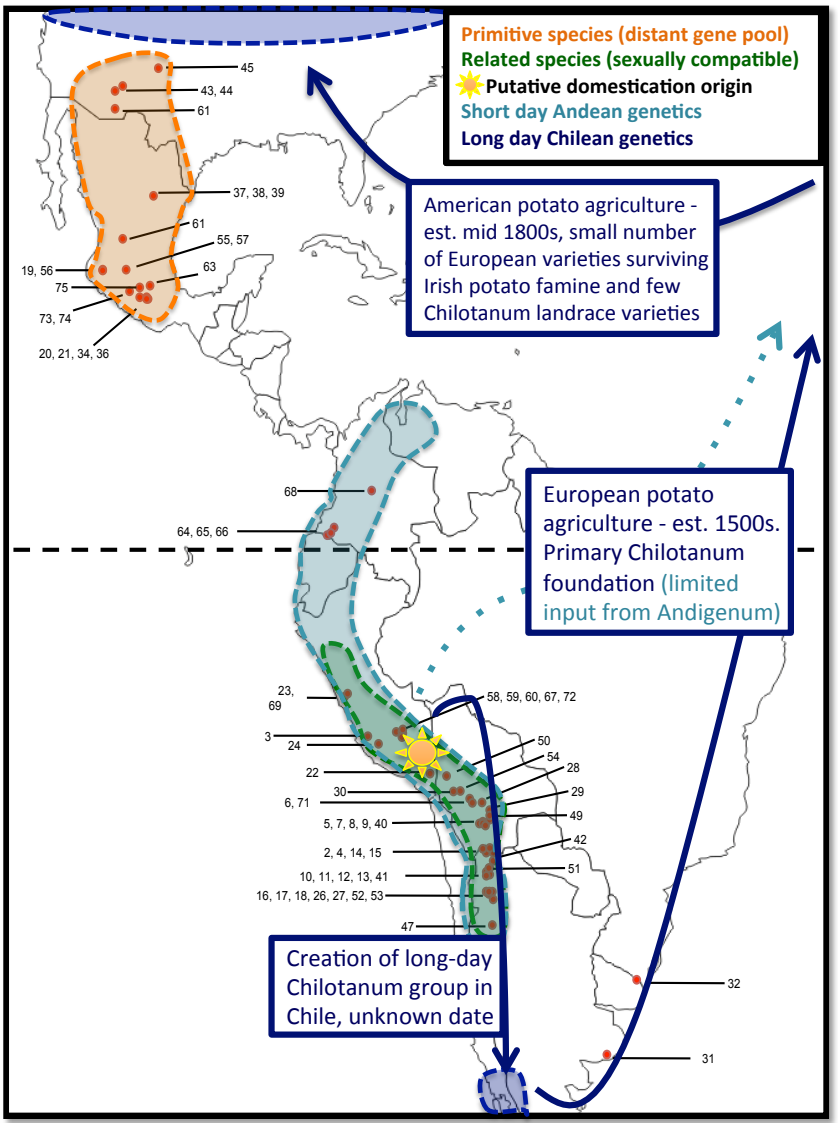

Figure 0.1. ***Overview of wild and cultivated potato geographic distributions, including primitive North American wild species (orange), wild species related to cultivated potatoes (green), Andean landraces (turquoise), and the foundation of European potato agriculture.*** *The figure does not depict Chilean landrace founders contributing to 19th century North American breeding populations, or the subsequent reintroduction of elite North American cultivars into South American potato agriculture (common in Chile). Numbers indicate locations of core collection accessions reported in Chapter 1 (Figure 1.1).*

Potato Breeding Challenges

Even with this broad array of genetic resources, North American potato breeders have struggled to improve productivity of new varieties beyond the nationally important cultivars of the early 1900s. Contrary to other major field crops, potato did not experience major yield gains due to genetic improvement in the last century (Douches et al., 1996; Feil, 1992). This trend can be partially attributed to a shifting focus that emphasized selection for a large number of market-specific quality traits and industry requirements during the 20th century, a situation that likely discouraged expansion of the genetic base with other germplasm (Douches et al., 1996). Restricted improvements in productivity are ultimately tied to inherent limitations imposed by a high degree of relatedness among cultivars (Mendoza and Haynes, 1974; Pavek and Corsini, 2001). Difficulties in producing F1 hybrid seedlings capable of exceeding the performance of either parent remain as a continued barrier to genetic gains (Simmonds, 1976; Tai, 1974). This problem is best exemplified in the fact that North America's most widely grown variety today is Russet Burbank which is a somatic mutant of a cultivar released in 1876 (Bethke et al., 2014). The state of progress in breeding potato for yield begs the question: Why have available cultivated heterotic groups and wild relatives not been more widely implemented in broadening the genetic base of potato and driving improvements in yield?

While the potato community recognizes potential value in landraces and crop wild relatives, their representation in elite germplasm remains limited (Bradshaw et al., 2006; Jansky et al., 2013). Difficulties in their use have historically prevented direct integration into breeding populations, instead leading potato programs engage in a lengthy process of "pre-breeding"

(Pavek and Corsini, 2001). Tetraploid Andean landraces (*S. tuberosum Andigenum* group) are a source of heterosis for North American cultivars, which primarily contain *Chilotanum* group genetics from long-day adapted Chilean landraces, but can introduce problems of cytoplasmic male sterility, late maturity, off-type tubers, and dominance of the Andigenum plant morphology (Tarn and Tai, 1983). Furthermore, both *Andigenum* and *Chilotanum* tetraploid germplasm are rife with deleterious mutations (Hougas et al., 1964; Kotch et al., 1992), likely facilitated by a combination of long-term clonal propagation and polyploid genetic redundancy. Genetic load is manifested as low fertility among elite germplasm. Severe inbreeding depression resulting from the high tetraploid genetic load (De Jong and Rowe, 1971; Mendoza and Haynes, 1973), compounded by complex tetrasomic inheritance in cultivars, opposes strategic development of heterotic groups to maximize heterozygosity of F1 populations, a method that has produced large gains in other outcrossing species such as maize (Duvick, 1984). This inability to perform inbreeding is also an impediment to fixation of desirable alleles within single genotypes, requiring that elite cultivars be maintained as clones in tissue culture, a process that in itself may also contribute to the build-up of deleterious alleles within cultivated populations (Cassells et al., 2001; Larkin and Snowcroft, 1981).

In addition to primitive landraces, wild species are also valuable sources of pest and disease resistance but frequently introduce unwanted traits, or "wildness" into breeding populations (Jansky et al., 2013; Spooner and Bamberg; 1994). These can be difficult to eliminate in a species with severe inbreeding depression at the tetraploid level, and self-incompatibility at the diploid level. Targeting the most desirable wild genes with more recent genome editing

tools also requires distinguishing their most useful alleles from a larger pool of unwanted wild genetics. Fundamentally, a core aspect of the genetic hurdle impeding breeders from successfully exploiting broad germplasm diversity lies in their ability to uncouple the "good genetics," or a subset of desirable alleles, from the large pool of unwanted alleles containing deleterious mutations or conferring a "wild" plant phenotype.

New Technologies to Improve Potato Breeding

The last several years have witnessed an emergence of potentially transformative approaches for enabling breeders to better utilize the broad extent of potato germplasm diversity. Foremost among these after the publication of a potato reference genome sequence to facilitate molecular breeding (Potato Genome Sequencing Consortium, 2011) are introduction of genetic self-compatibility in diploid cultivated potatoes (Jansky et al., 2016), and development of genome editing techniques, in particular transcription activator-like (TAL) effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR)/Cas systems, that continue to increase our capability for targeted addition, deletion or replacement of alleles (Butler et al., 2015; Wang et al., 2015). These resources have considerable promise with regards to bringing cultivated potato into an era of genomic breeding.

Diploid Breeding

The isolation of multiple genetic sources for overcoming self-incompatibility in diploid potatoes (Arnold, 2013; Jansky et al., 2014) represents a major breakthrough that could eventually reshape the process of potato breeding (Jansky et al., 2016). Used in combination

with day-length adapted diploid populations from the *Andigenum* group and vigorous dihaploids derived from elite cultivars, this approach may ultimately prove an invaluable tool for rebuilding a more diverse potato germplasm base absent the genetic load that has historically stood as a barrier to its improvement. Diploid breeding also offers the advantage of faster progress in selection across generations due to the greater simplicity of disomic inheritance. Indeed, diploid genotypes with plant vigor comparable to some tetraploids have already been observed by the Michigan State University potato breeding program (Douches, D., pers. comm.). The ability to self-fertilize hybrid progeny will allow potato breeders to introduce broad allelic variation and adaptive traits from wild populations while purging unwanted variation from the wild background, and enable development of near-isogenic cultivated lines (NILs) which have proven to be a highly useful tool in the tomato community (Martin et al., 1991). Lastly, the ability to develop heterotic inbreds could eventually open up potato to yield benefits similar to what drove hybrid maize productivity for much of the last century (Duvick, 1984).

Genome Editing

Genome editing offers further advantages for introducing novel variation from wild populations. Introducing genes without disrupting the genetic background of elite varieties, or imparting wildness into breeding populations will reduce the time and effort devoted to pre-breeding, ultimately accelerating variety development. This could prove particularly useful for insect, nematode and pathogen resistance breeding. Genome editing also has potential to increase the accessibility of alleles in the secondary and tertiary potato gene pools. Historically, a number of critical disease resistance loci were introduced to cultivated potato

either from allopolyploid species of hybrid origins that required extensive subsequent backcrossing, or from evolutionarily diverged species with sexual reproductive barriers (e.g. different endosperm balance numbers) that required protoplast fusion or embryo rescue followed by backcrossing (Colon et al., 1995; Helgeson et al., 1998).

In the case of both diploid breeding and genome editing, there is significant potential in a genomics-based approach in future potato breeding efforts. Sequence analysis of genome-wide and gene level diversity in North American cultivars, landraces and their wild relatives will inform breeders in making selections that diversify breeding populations, and to track alleles from these populations that most strongly benefit cultivated phenotypes during the selection process. Likewise, effective use of gene editing requires sequence-level knowledge of candidate loci controlling traits of interest to breeders. Whether the goal is tracking genetic diversity to maximize the heterotic potential of breeding populations, or identifying the best candidate gene haplotypes for transformation, a genomics-enabled approach represents a powerful strategy in potato breeding and biotechnology in the 21st century.

A Path Toward Deploying Germplasm Diversity in the Future Improvement of Potato

With the exception of a handful of genetic markers for disease resistance, potato breeding strategies have not progressed far beyond the phenotypic selection employed a century ago (Gebhardt, 2013; Hirsch et al., 2014). This has almost certainly limited the ability of breeders to effectively utilize untapped sources of genetic diversity. However, the last five years have witnessed major efforts toward developing diverse self-compatible diploid genotypes (Jansky et al., 2016), ground-breaking advances in genome editing technology (Andersson et al.,

2016; Cong et al., 2013; Wang et al., 2015), and the publication of research tools including a potato reference genome assembly (Potato Genome Sequencing Consortium, 2011) and large collections of genome-anchored molecular markers (Hamilton et al., 2011). Given the combined potential of these genetic resources to accelerate varietal development, and the inherent complexity of breeding an outcrossing polyploid with heterogeneous genome structure, a genomics-informed approach to potato breeding in the 21st century could help to harness novel genetic diversity more effectively than in past decades (Hirsch and Buell, 2013). A more complete understanding of the genetic diversity in cultivated varieties and their tuber-bearing progenitors will be crucial to maximizing potato's genetic potential through breeding and biotechnology in the genomics era. This will first require deeper investigation of the molecular basis of potato sequence diversity, its landscape in wild and cultivated species, and perhaps most importantly, how this landscape has been shaped by human selection in promoting agricultural traits. Exploring these areas may eventually improve our understanding of the key factors controlling performance in potato, including heterozygosity and inbreeding depression, impacts of polyploidy and clonal propagation on plant genome evolution, and the subset of allelic diversity required to produce key traits underlying potato's value as a food source. The knowledge derived from this research will hopefully serve as a foundation to inform breeders in their future management of cultivated potato genetic diversity, and yield candidate loci underlying traits of interest.

Problem Definition

Resources are now available to accelerate the use of untapped germplasm diversity for the improvement of cultivated potatoes by using a genomics-enabled approach to breeding. Implementing a genome-informed breeding strategy to reach this objective would benefit from updated, comprehensive evaluations of landrace and wild genetic diversity using datasets generated with genomics-era technologies. Until recently, published studies of potato landrace or wild species germplasm diversity have employed limited numbers of outdated genetic markers or morphological traits. These types of data can only approximate levels of genetic diversity that are directly quantifiable by studying sequence-level variation, and do not allow analysis of the full extent of genome-wide variation in an individual. Molecular marker applications in non-cultivar potato germplasm have been mainly limited to systematics and linkage mapping. Few studies have attempted to analyze the full extent of sequence diversity housed in samples derived from landrace and wild populations, and more importantly, how this diversity has been selected as a result of domestication and breeding.

Objective

To generate genomics-era datasets for the exploration of genome-wide diversity in cultivated potatoes and their tuber-bearing relatives, with a focus on analysis of species relationships, genomic variation in primitive diploid progenitors, and identification of loci selected during domestication of potato.

Dissertation Outline

Chapter 1 is an introduction to potato germplasm diversity, genetic considerations and genome biology that outlines challenges currently faced by breeders.

Chapter 2 is a study of diversity and systematics in a core collection of 75 accessions representing 25 tuber-bearing *Solanum* species, and 213 European and North American tetraploid cultivars.

Chapter 3 is a study using deep whole genome sequence coverage sequence data to quantify the extent of sequence variation and structural variation in a panel of 12 monoploid/doubled monoploid clones derived from diploid landraces with a focus on the impact on genome diversity, quantification of deleterious genome content, and the role of copy number variation in genome evolution in diploid progenitors of cultivated potato.

Chapter 4 is a whole genome re-sequencing study comparing the genome-wide allelic composition of 23 tetraploid cultivars, 20 primitive South American landraces, and 20 diploid wild species progenitors to measure the genetic diversity of the respective genetic groups, and identify candidate loci selected during potato's domestication.

Chapter 5 contains General Conclusions offering insights into the potential of advances in genome sequencing to assist in germplasm utilization, and proposes relevant topics for potato diversity studies in need of future investigation.

LITERATURE CITED

LITERATURE CITED

Andersson, M., Turesson, H., Nicolia, A., Fält, A., Samuelsson, M., and Hofvander, P. (2016). Efficient targeted multiallelic mutagenesis in tetraploid potato (*Solanum tuberosum*) by transient CRISPR-Cas9 expression in protoplasts. Plant Cell Rep. 1-12.

Arnold, B.E. (2013). Identification of candidate genes for self-compatability in a diploid population of potato derived from parents used in genome sequencing (Doctoral dissertation, Virginia Tech).

Bethke, P.C., Nassar, A.M., Kubow, S., Leclerc, Y.N., Li, X.-Q., Haroon, M., Molen, T., Bamberg, J., Martin, M., and Donnelly, D.J. (2014). History and origin of Russet Burbank (Netted Gem) a sport of Burbank. Am. J. Potato Res. 91, 594-609.

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

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

Brush, S.B., Carney, H.J., and Humán, Z. (1981). Dynamics of Andean potato agriculture. Econ. Bot. 35, 70-88.

Butler, N.M., Atkins, P.A., Voytas, D.F., and Douches, D.S. (2015). Generation and inheritance of targeted mutations in potato (*Solanum tuberosum* L.) using the CRISPR/Cas system. PloS ONE 10, e0144591.

Camadro, E.L. (2010). Characterization of the natural genetic diversity of Argentinian potato species and manipulations for its effective use in breeding. Am. J. Plant. Sci. Biotechnol. 3 (Special Issue 1), 65-71.

Cassells, A.C., and Curry, R.F. (2001). Oxidative stress and physiological, epigenetic and genetic variability in plant tissue culture: implications for micropropagators and genetic engineers. Plant Cell Tiss. Org. 64, 145-157.

Chakraborty, S., and Newton, A.C. (2011). Climate change, plant diseases and food security: an overview. Plant Pathol. 60, 2-14.

Colon, L., Turkensteen, L., Prummel, W., Budding, D., and Hoogendoorn, J. (1995). Durable resistance to late blight (*Phytophthora infestans*) in old potato cultivars. Eur. J. Plant Pathol. 101, 387-397.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., and Marraffini, L.A. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823.

De Jong, H., and Rowe, P. (1971). Inbreeding in cultivated diploid potatoes. Potato Res. 14, 74-83.

Douches, D., Maas, D., Jastrzebski, K., and Chase, R. (1996). Assessment of potato breeding progress in the USA over the last century. Crop Sci. 36, 1544-1552.

Duvick, D.N. (1984). Genetic contributions to yield gains of US hybrid maize, 1930 to 1980. Genetic contributions to yield gains of five major crop plants, 15-47.

Feil, B. (1992). Breeding progress in small grain cereals—A comparison of old and modern cultivars. Plant Breeding 108, 1-11.

Gebhardt, C. (2013). Bridging the gap between genome analysis and precision breeding in potato. Trends Genet. 29, 248-256.

Hamilton, J.P., Hansey, C.N., Whitty, B.R., Stoffel, K., Massa, A.N., Van Deynze, A., De Jong, W.S., Douches, D.S., and Buell, C.R. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics 12, 1.

Helgeson, J., Pohlman, J., Austin, S., Haberlach, G., Wielgus, S., Ronis, D., Zambolim, L., Tooley, P., McGrath, J., and James, R. (1998). Somatic hybrids between *Solanum bulbocastanum* and potato: a new source of resistance to late blight. Theor. Appl. Genet. 96, 738-742.

Hermundstad, S., and Peloquin, S.J. (1986) Tuber yield and tuber traits of haploid-wild species F1 hybrids. Potato Res. 29, 289-297.

Hijmans, R.J. (2002). Atlas of wild potatoes. Vol. 10. Bioversity International.

Hijmans, R.J., and Spooner, D.M. (2001). Geographic distribution of wild potato species. Am. J. Bot. 88, 2101-2112.

Hirsch, C.D., Hamilton, J.P., Childs, K.L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C.N., Habermann, M., Neal, B., and Buell, C.R. (2014). Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. Plant Genome 7(1).

Hirsch, C.N., and Robin Buell, C. (2013). Tapping the promise of genomics in species with complex, nonmodel genomes. Ann. Rev. Plant Biol. 64, 89-110.

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R.E., Jansky, S., and Bethke, P. (2013). Retrospective view of North American

potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3 Genes Genom. Genet. 3, 1003-1013.

Hosaka, K. (2003). T-type chloroplast DNA in *Solanum tuberosum* L. ssp. *tuberosum* was conferred from some populations of *S. tarijense* Hawkes. Am. J. Potato Res. 80, 21-32.

Hosaka, K., and Hanneman Jr, R. (1988). The origin of the cultivated tetraploid potato based on chloroplast DNA. Theor. Appl. Genet. 76, 172-176.

Hougas, R., Peloquin, S., and Gabert, A. (1964). Effect of seed-parent and pollinator on frequency of haploids in *Solanum tuberosum*. Crop Sci. 4, 593-595.

Jansky, S., Chung, Y.S., and Kittipadukal, P. (2014). M6: A diploid potato inbred line for use in breeding and genetics research. J. Plant Regist. 8, 195-199.

Jansky, S., Dempewolf, H., Camadro, E., Simon, R., Zimnoch-Guzowska, E., Bisognin, D., and Bonierbale, M. (2013). A case for crop wild relative preservation and use in potato. Crop Sci. 53, 746-754.

Jansky, S.H., Charkowski, A.O., Douches, D.S., Gusmini, G., Richael, C., Bethke, P.C., Spooner, D.M., Novy, R.G., De Jong, H., and De Jong, W.S. (2016). Reinventing Potato as a Diploid Inbred Line–Based Crop. Crop Sci. 56, 1-11.

Kotch, G.P., Ortiz, R., and Peloquin, S. (1992). Genetic analysis by use of potato haploid populations. Genome 35, 103-108.

Larkin, P.J., and Scowcroft, W.R. (1981). Somaclonal variation—a novel source of variability from cell cultures for plant improvement. Theor. Appl. Genet. 60, 197-214.

Lin, B.-H., Lucier, G., Allshouse, J., and Kantor, L.S. (2001). Market distribution of potato products in the United States. Journal of Food Products Marketing 6, 63-78.

Love, S.L. (1999). Founding clones, major contributing ancestors, and exotic progenitors of prominent North American potato cultivars. Am. J. Potato Res. 76, 263-272.

Martin, G.B., Williams, J., and Tanksley, S.D. (1991). Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. Proc. Natl. Acad. Sci. U. S. A. 88, 2336-2340.

Mendoza, H., and Haynes, F. (1973). Some aspects of breeding and inbreeding in potatoes. Am. Potato J. 50, 216-222.

Mendoza, H., and Haynes, F. (1974). Genetic relationship among potato cultivars grown in the United States. HortSci.

Newton, A.C., Johnson, S.N., and Gregory, P.J. (2011). Implications of climate change for diseases, crop yields and food security. Euphytica 179, 3-18.

Ovchinnikova, A., Krylova, E., Gavrilenko, T., Smekalova, T., Zhuk, M., Knapp, S., and Spooner, D.M. (2011). Taxonomy of cultivated potatoes (*Solanum* section *Petota*: *Solanaceae*). Bot. J. Linn. Soc. 165, 107-155.

Pavek, J., and Corsini, D. (2001). Utilization of potato genetic resources in variety development. Am. J. Potato Res. 78, 433-441.

Plaisted, R., and Hoopes, R. (1989). The past record and future prospects for the use of exotic potato germplasm. Am. Potato J. 66, 603-627.

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

Ross, H. (1966). The use of wild *Solanum* species in German potato breeding of the past and today. Am. J. Potato Res. 43, 63-80.

Scott, G., and Suarez, V. (2012). The rise of Asia as the centre of global potato production and some implications for industry. Potato J. 1, 1-22.

Simmonds, N. (1976). Neotuberosum and the genetic base in potato breeding. ARC Res. Rev. 2, 9-11.

Spooner, D.M. (2009). DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. Am. J. Bot. 96, 1177-1189.

Spooner, D.M., and Bamberg, J.B. (1994). Potato genetic resources: sources of resistance and systematics. Am. Potato J. 71, 325-337.

Spooner, D.M., and Salas, A. (2006). Structure, biosystematics, and genetic resources. Handbook of potato production, improvement, and postharvest management. J. Gopal, S.M. Khurana, editors. New York: Food Products Press.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. U. S. A. 102, 14694-14699.

Tai, G. (1974). A method for quantitative genetic analysis of early clonal generation seedlings of an asexual crop with special application to a breeding population of the potato (*Solanum tuberosum* L.). Theor. Appl. Genet. 45, 150-156.

Wang, S., Zhang, S., Wang, W., Xiong, X., Meng, F., and Cui, X. (2015). Efficient targeted mutagenesis in potato by the CRISPR/Cas9 system. Plant Cell Rep. 34, 1473-1476.

Wang, Y., and Frei, M. (2011). Stressed food–The impact of abiotic environmental stresses on crop quality. Agric. Ecosyst. Environ. 141, 271-286.

CHAPTER 2

TAXONOMY AND GENETIC DIFFERENTIATION AMONG WILD AND CULTIVATED GERMPLASM OF *SOLANUM* SECT. *PETOTA*

[Published in: The Plant Genome 8 (1): 1-16]

Michael A. Hardigan[1], John Bamberg[2], C. Robin Buell[1], and David S. Douches[3]*

[1]Department of Plant Biology, Michigan State University, East Lansing, MI 48824-1312

[2]USDA Agricultural Research Service, Sturgeon Bay, WI 54235-9620

[3]Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI 48824-1312

*Corresponding author (douchesd@msu.edu).

Abstract

Because of their adaptation to a diverse set of habitats and stresses, wild species of cultivated crops offer new sources of genetic diversity for germplasm improvement. Using an Infinium array representing a genome-wide set of 8303 single nucleotide polymorphisms (SNPs), we evaluated phylogenetic relationships and allele diversity within a diversity panel of germplasm from *Solanum* sect. *Petota*, the group containing tuber-bearing species and landraces of potato as well as cultivated potato (*Solanum tuberosum* L.). This diversity panel consists of 75 plant introductions (PIs) representing 25 species and provides a diverse representation of tuber-bearing *Solanum* germplasm. To determine the relatedness between current species classifications and SNP-based genetic distances, we generated a phylogeny based upon random individuals from each PI that, with few exceptions, revealed general agreement with taxonomic grouping of species in *Solanum* sect. *Petota*. Genotype comparisons between our *Solanum* sect. *Petota* diversity panel and a panel of 213 tetraploid cultivars and breeding lines revealed that the average genetic distance between landraces was higher than between cultivated clones, indicating a greater extent of diversity between populations of native Andean landraces than among modern cultivars and breeding lines. Analysis of allele frequencies at individual SNP loci between the *Solanum* sect. *Petota* diversity panel and tetraploid cultivars revealed loci with extreme divergence between cultivated potato and its tuber-bearing relatives. Interestingly, a number of these loci are associated with genes related to carbohydrate metabolism and tuber development, suggesting potential roles in domestication of potato. Our SNP data offers a new taxonomic view of potato germplasm, while further identifying candidate alleles that likely differentiate wild germplasm and cultivated potato and underlie key agronomic traits.

Abbreviations: DM, DM1–3 516 R44 reference genome assembly; EBN, endosperm balance number; GO, Gene Ontology; PI, plant introductions; SNP, single nucleotide polymorphism.

Introduction

The ability to exploit genetic diversity is critical to modern crop breeding, as it enables the introduction of new, useful genes and allelic variants into existing germplasm. Germplasm diversity offers a source of novel traits of agronomic value and can increase the variety of alleles within breeding populations (Bamberg and Del Rio 2005; Bradshaw et al., 2006; Lam et al., 2010; Pavek and Corsini, 2001; Tester and Langridge, 2010). A large extent of genetic diversity can be found in wild and landrace relatives of cultivated crop species (Bradshaw et al., 2006; Pavek and Corsini, 2001) and evaluation of the diversity in these wild species and landraces at the sequence level facilitates its use in developing new cultivars. Surveying wild genotypes across broad sets of loci can provide clues to genes that differentiate primitive and cultivated germplasm, suggesting key loci involved in the domestication and enhanced agronomic performance of modern varieties (Huang et al., 2012; Hufford et al., 2012; Olsen and Wendel, 2013; Qi et al., 2013). Such genes offer candidates for breeders in selecting elite cultivars and have implications for the expression of important field traits. Understanding the taxonomy of wild relatives of cultivated crops is also important, as it increases the efficiency of germplasm conservation and may lend predictive power to the implementation of diverse plant species in breeding programs (Daly et al., 2001, Spooner and Salas, 2006).

Potato taxonomy has changed throughout recent history, depending on whether intermediate interspecific "hybrid" populations are tolerated or whether, in contrast, the greater variation of all similar populations should be subsumed under a single, broader taxon. The current taxonomy of tuber-bearing potatoes (*Solanum* sect. *Petota*) includes approximately 110 *Solanum* species including wild species, landraces, and modern *Solanum* tuberosum L. cultivars (Spooner, 2009). Their natural habitats span an immense geographic range from the southwestern United States to central Chile and Argentina, displaying significant variation in temperature, moisture, soil quality, and biological stresses (Hijmans and Spooner, 2001; Spooner and Salas, 2006). Adaptation to such a broad assortment of ecological pressures has generated extensive diversity in wild and landrace populations (Spooner and Bamberg, 1994). The value of wild relatives and landraces for germplasm enhancement has been well established in potato and they have been exploited for the introduction or improvement of numerous traits including resistance to bacterial pathogens, viral pathogens, and insect pests; tolerance of abiotic stresses such as salinity and frost; and traits related to the market quality of harvested tubers (Pavek and Corsini, 2001). Andean landraces are also thought to have potential for broadening the genetic base of North American and European breeding populations, which are thought to house a small portion of the available diversity (Bradshaw et al., 2006; Jansky et al., 2013; Mendoza, 1989; Pavek and Corsini, 2001). Previous estimates of genetic relationships and diversity in *Solanum* sect. *Petota* have used plant morphology (Ames et al., 2008), chloroplast DNA (Spooner and Castillo, 1997), and molecular marker data from randomly amplified polymorphic DNA (Demeke et al., 1996), restriction fragment length polymorphisms (Debener et al., 1990), amplified fragment length polymorphisms (AFLPs; Spooner et al., 2005), and simple sequence repeats (Van den Berg et al., 2002).

Despite a variety of approaches in classifying potato germplasm, taxonomic treatment of these species remains a challenge complicated by interspecific hybridization, introgressions, sexual and asexual reproduction, and polyploidy. No single method of evaluating genetic relationships has proven effective for this group and there remains disagreement in the designation of species boundaries (Spooner, 2009; Spooner and Salas, 2006). For the purpose of discussing germplasm, this study recognizes those species defined by Spooner and Salas (2006) with regard to further unpublished reductions (Spooner, 2009).

In this study, high-throughput SNP analysis with the Infinium 8303 potato SNP array (Felcher et al., 2012) was used to assess taxonomic relationships in a diversity panel of wild species and landraces from *Solanum* sect. *Petota* and compare the genetic composition of these loci with that of a diversity panel of cultivated tetraploid clones representing the major market classes of potato (Hirsch et al., 2013). The *Solanum* sect. *Petota* diversity panel was derived from the US Potato Genebank (http://www.ars-grin.gov/nr6/, accessed 5 Nov. 2014) and contains 74 PIs obtained from populations in a range of habitats across North and South America (Figure 1.1, Table 1.1). These germplasm demonstrate considerable morphological variation even within species and represent several classes of ploidy and endosperm balance number (EBN, also referred to as "effective ploidy"). Furthermore, many of these species have been used in potato breeding including *Solanum* tuberosum group Andigenum landraces, *S. acaule* Bitter, *S. berthaultii* Hawkes, *S. bulbocastanum* Dunal, *S. chacoense* Bitter, *S. demissum* Lindl., *S. kurtzianum* Bitter & Wittm., *S. raphanifolium* Cardenas & Hawkes, and *S. stoloniferum* Schltdl. & Bouche (Bradshaw et al., 2006; Jansky et al., 2013; Pavek and Corsini, 2001). Phylogenetic analysis of these accessions revealed, with limited

exception, that SNP-based genetic distances largely support the current taxonomy of these tuber-bearing *Solanum* species. Analysis of diversity in the *Solanum* sect. *Petota* diversity panel and a large panel of cultivated potato varieties (Hirsch et al., 2013) revealed that a greater degree of genetic diversity can be found within the sampled landraces compared with modern breeding lines. Comparison of SNP genotypes between the groups identified numerous genes with highly divergent allele frequencies, particularly those involved in carbohydrate metabolism and tuber development, which are most likely to underlie important agronomic traits.

Materials and Methods

Plant Materials

The species and landraces used for this study were assembled at the US Potato Genebank at Sturgeon Bay, WI. With over 5000 populations of over 100 species, a comprehensive evaluation was impractical. A panel representing the broad diversity of *Solanum* section *Petota* was constructed containing 74 total PIs and, in the majority of instances, three populations from 25 different wild species and primitive cultivated forms (based on prior classifications, some species have been collapsed) collected across North and South America, hereafter referred to as the "*Solanum* sect. *Petota* diversity panel".

The *Solanum* sect. *Petota* diversity panel intends to capture a representative sample of the tuber-bearing *Solanum* diversity, but with a bias toward germplasm that is likely to be practical for evaluation and use in breeding. Thus species were selected to represent different ploidies, breeding systems, crossability groups, geographic origins, and reputation for

Table 1.1. ***Summary of wild and landrace plant introductions in the Solanum sect. Petota diversity panel.***

Solanum species	Accession[†]	Ploidy	Chr.[††]	EBN[‡]	Clade[§]	Series[¶]	Previous name[#]
S. acaule	PI175395	4x	48	2	4	Acaulia	–
S. acaule	PI472661	4x	48	2	4	Acaulia	–
S. acaule	PI473481	4x	48	2	4	Acaulia	–
S. berthaultii	PI458365	2x	24	2	4	Tuberosa	*S. tarijense*
S. berthaultii	PI498141	2x	24	2	4	Tuberosa	*S. tarijense*
S. berthaultii	PI310927	2x	24	2	4	Tuberosa	–
S. boliviense	PI265873	2x	24	2	4	Megistacroloba	*S. megistacrolobum*
S. boliviense	PI545964	2x	24	2	4	Megistacroloba	–
S. boliviense	PI597736	2x	24	2	4	Megistacroloba	–
S. brevicaule	PI265579	4x	48	4	4	Tuberosa	*S. gourlayi*
S. brevicaule	PI473011	2x	24	2	4	Tuberosa	*S. gourlayi*
S. brevicaule	PI473062	2x	24	2	4	Tuberosa	*S. gourlayi*
S. brevicaule	PI435079	2x	24	2	4	Tuberosa	*S. oplocense*
S. brevicaule	PI473185	6x	72	4	4	Tuberosa	*S. oplocense*
S. brevicaule	PI473190	6x	72	4	4	Tuberosa	*S. oplocense*
S. brevicaule	PI205407	2x	24	2	4	Tuberosa	*S. spegazzinii*
S. brevicaule	PI472978	2x	24	2	4	Tuberosa	*S. spegazzinii*
S. brevicaule	PI500053	2x	24	2	4	Tuberosa	*S. spegazzinii*
S. bulbocastanum	PI545751	2x	24	1	2	Bulbocastana	–
S. bulbocastanum	PI243510	2x	24	1	2	Bulbocastana	–
S. bulbocastanum	PI275188	2x	24	1	2	Bulbocastana	–
S. candolleanum	PI265863	2x	24	2	4	Tuberosa	*S. bukasovii*
S. candolleanum	PI365321	2x	24	2	4	Tuberosa	*S. bukasovii*
S. candolleanum	PI458379	2x	24	2	4	Tuberosa	*S. bukasovii*
S. chacoense	PI197760	2x	24	2	4	Yungasensa	–
S. chacoense	PI275139	2x	24	2	4	Yungasensa	–
S. chacoense	PI320293	2x	24	2	4	Yungasensa	–
S. circaeifolium	PI498116	2x	24	1	4	Circaeifolia	*S. capsicibaccatum*
S. circaeifolium	PI498120	2x	24	1	4	Circaeifolia	*S. capsicibaccatum*
S. commersonii	PI472837	2x	24	1	4	Commersoniana	–
S. commersonii	PI473411	2x	24	1	4	Commersoniana	–
S. commersonii	PI558050	2x	24	1	4	Commersoniana	–
S. demissum	PI160208	6x	72	4	4	Demissa	–
S. demissum	PI230589	6x	72	4	4	Demissa	–
S. demissum	PI498232	6x	72	4	4	Demissa	–
S. hjertingii	PI251065	4x	48	2	4	Longipedicellata	–
S. hjertingii	PI283103	4x	48	2	4	Longipedicellata	–
S. hjertingii	PI545715	4x	48	2	4	Longipedicellata	–
S. infundibuliforme	PI265867	2x	24	2	4	Cuneolata	–
S. infundibuliforme	PI458324	2x	24	2	4	Cuneolata	–
S. infundibuliforme	PI472894	2x	24	2	4	Cuneolata	–
S. jamesii	PI458425	2x	24	1	1	Pinnatisecta	S. michoacanum
S. jamesii	PI592422	2x	24	1	1	Pinnatisecta	–
S. jamesii	PI605370	2x	24	1	1	Pinnatisecta	–
S. kurtzianum	PI472923	2x	24	2	4	Tuberosa	–
S. kurtzianum	PI472941	2x	24	2	4	Tuberosa	–

Table 1.1 (cont'd)

S. kurtzianum	PI498359	2x	24	2	4	Tuberosa	–
S. microdontum	PI498123	2x	24	2	4	Tuberosa	–
S. microdontum	PI310979	2x	24	2	4	Tuberosa	–
S. microdontum	PI458355	2x	24	2	4	Tuberosa	–
S. okadae	PI458368	2x	24	unk	4	Tuberosa	S. venturii
S. okadae	PI320327	2x	24	unk	4	Tuberosa	–
S. okadae	PI498130	2x	24	unk	4	Tuberosa	–
S. pinnatisectum	PI184774	2x	24	1	1	Pinnatisecta	–
S. pinnatisectum	PI275236	2x	24	1	1	Pinnatisecta	–
S. pinnatisectum	PI347766	2x	24	1	1	Pinnatisecta	–
S. raphanifolium	PI296126	2x	24	2	4	Megistacroloba	–
S. raphanifolium	PI310953	2x	24	2	4	Megistacroloba	–
S. raphanifolium	PI473369	2x	24	2	4	Megistacroloba	–
S. stoloniferum	PI655250	4x	48	2	4	Longipedicellata	*S. fendleri*
S. stoloniferum	PI184770	4x	48	2	4	Longipedicellata	*S. polytrichon*
S. stoloniferum	PI161170	4x	48	2	4	Longipedicellata	–
S. tuberosum spp.	PI225710	2x	24	2	4	Tuberosa	*S. phureja*
S. tuberosum spp.	PI320355	2x	24	2	4	Tuberosa	*S. phureja*
S. tuberosum spp.	PI320377	2x	24	2	4	Tuberosa	*S. phureja*
S. tuberosum spp.	PI195204	4x	48	2	4	Tuberosa	*S. stenotomum*
S. tuberosum spp.	PI283141	2x	24	2	4	Tuberosa	*S. stenotomum*
S. tuberosum spp.	PI292110	2x	24	2	4	Tuberosa	*S. stenotomum*
S. tuberosum spp.	PI281034	4x	48	4	4	Tuberosa	–
S. tuberosum spp.	PI546023	4x	48	4	4	Tuberosa	–
S. tuberosum spp.	PI607886	4x	48	4	4	Tuberosa	–
S. verrucosum	PI161173	2x	24	2	4	Tuberosa	–
S. verrucosum	PI275255	2x	24	2	4	Tuberosa	–
S. verrucosum	PI498062	2x	24	2	4	Tuberosa	–

† Identifier associated with plant introductions in the US Potato Genebank.
‡ Endosperm balance number (EBN), considered the effective ploidy of an individual and explaining the reproductive barriers between species.
§ Potato species taxonomic clade described by Spooner and Castillo (1997).
¶ Potato species taxonomic series described by Hawkes (1990).
# Previous species names for certain populations and taxonomic treatments have changed over time for some groups.
†† Chr., chromosome number; unk, unknown.

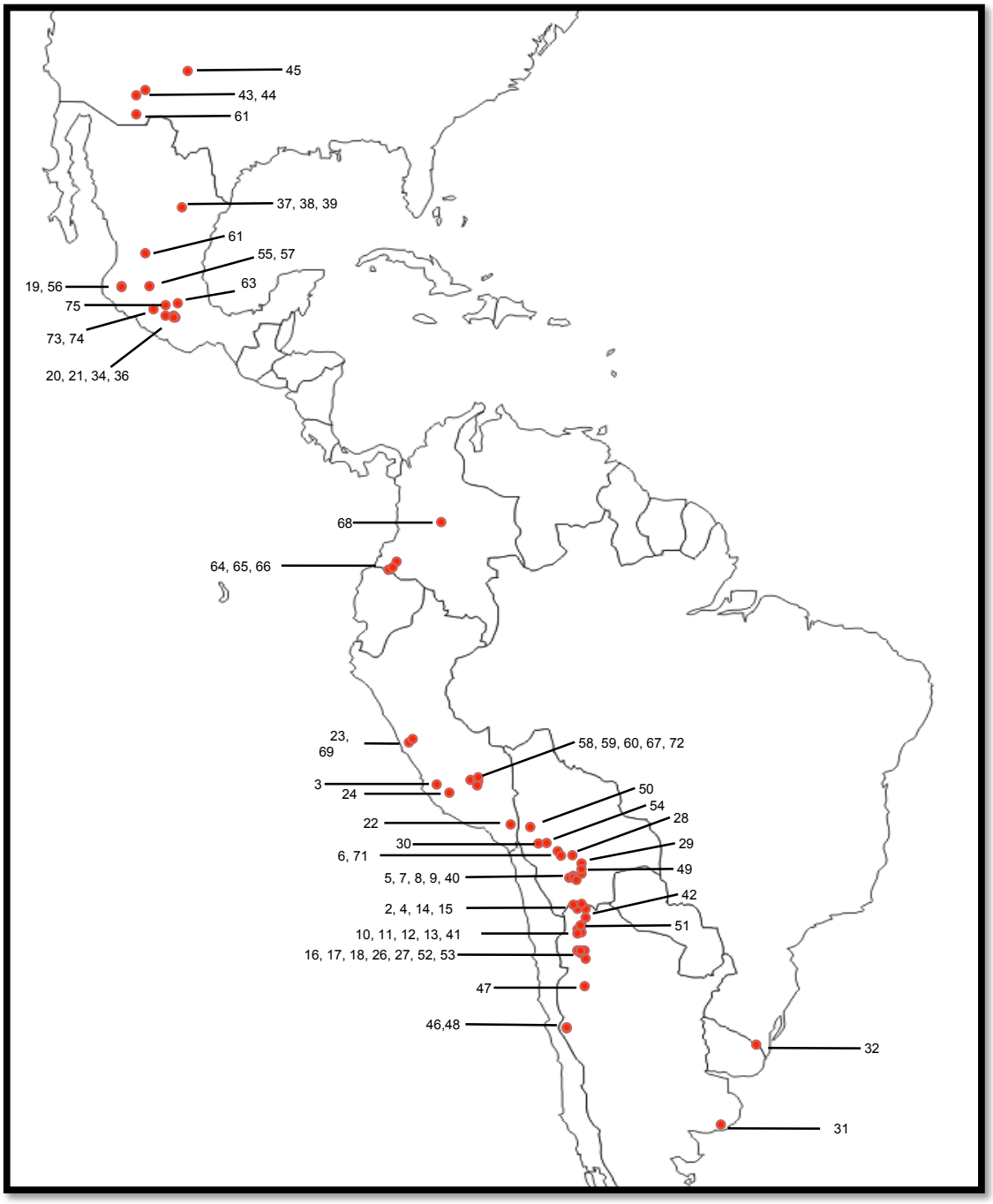

Figure 1.1. ***Map displaying geographic origins of wild species and landrace populations found in the Solanum sect. Petota diversity panel.*** *(**Population lacks record of collection location). (1) S. acaule (PI175395)**, (2) S. acaule (PI472661), (3) S. acaule (PI473481), (4) S. berthaultii (PI458365), (5) S. berthaultii (PI498141), (6) S. berthaultii (PI310927), (7) S.*

Figure 1.1 (cont'd)

*boliviense (PI265873), (8) S. boliviense (PI545964), (9) S. boliviense (PI597736), (10) S. brevicaule (PI265579), (11) S. brevicaule (PI473011), (12) S. brevicaule (PI473062), (13) S. brevicaule (PI435079), (14) S. brevicaule (PI473185), (15) S. brevicaule (PI473190), (16) S. brevicaule (PI205407), (17) S. brevicaule (PI472978), (18) S. brevicaule (PI500053), (19) S. bulbocastanum (PI545751), (20) S. bulbocastanum (PI243510, (21) S. bulbocastanum (PI275188, (22) S. candolleanum (PI265863), (23) S. candolleanum (PI365321), (24) S. candolleanum (PI458379), (25) S. chacoense (PI197760)**, (26) S. chacoense (PI275139), (27) S. chacoense (PI320293), (28) S. circaeifolium (PI498116), (29) S. circaeifolium (PI498120), (30) S. circaeifolium (PI545974), (31) S. commersonii (PI472837), (32) S. commersonii (PI473411), (33) S. commersonii (PI558050)**, (34) S. demissum (PI160208), (35) S. demissum (PI230589)**, (36) S. demissum (PI498232), (37) S. hjertingii (PI251065), (38) S. hjertingii (PI283103), (39) S. hjertingii (PI545715), (40) S. infundibuliforme (PI265867), (41) S. infundibuliforme (PI458324), (42) S. infundibuliforme (PI472894), (43) S. jamesii (PI458425), (44) S. jamesii (PI592422), (45) S. jamesii (PI605370), (46) S. kurtzianum (PI472923), (47) S. kurtzianum (PI472941), (48) S. kurtzianum (PI498359), (49) S. microdontum (PI498123), (50) S. microdontum (PI310979), (51) S. microdontum (PI458355), (52) S. okadae (PI458368), (53) S. okadae (PI320327),*

(54) S. okadae (PI498130), (55) S. pinnatisectum (PI184774), (56) S. pinnatisectum (PI275236), (57) S. pinnatisectum (PI347766), (58) S. raphanifolium (PI296126), (59) S. raphanifolium (PI310953), (60) S. raphanifolium (PI473369), (61) S. stoloniferum (PI655250), (62) S. stoloniferum (PI184770), (63) S. stoloniferum (PI161170), (64) S. tuberosum spp. (PI225710), (65) S. tuberosum spp. (PI320355), (66) S. tuberosum spp.

Figure 1.1 (cont'd)

*(PI320377), (67) S. tuberosum spp. (PI195204), (68) S. tuberosum spp. (PI283141), (69) S. tuberosum spp. (PI292110), (70) S. tuberosum spp. (PI281034)**, (71) S. tuberosum spp. (PI546023), (72) S. tuberosum spp. (PI607886), (73) S. verrucosum (PI161173), (74) S. verrucosum (PI275255), and (75) S. verrucosum (PI498062).*

particular useful traits. Highly exotic species difficult to grow, tuberize, hybridize, or maintain as populations of botanical seed were excluded. Species accessions that could not be maintained as populations of botanical seed were excluded. Both diploid (*S. phureja* and *S. stenotomum* Juz. et Buk.) and tetraploid (*S. tuberosum* group Andigenum) primitive cultivated forms were included.

Seeds from each PI were soaked for 24 h in 1500 mg L^{-1} gibberellic acid, washed in 10% bleach solution, rinsed with sterile water, and germinated on filter paper under ambient light. After 1–2 wk., sprouted seedlings were transferred under sterile conditions to Murashige and Skoog media and maintained in vitro (16/8h photoperiod at 24°C). To assess diversity within the panel, several seeds were germinated from each PI. Based on previous taxonomic studies that tested single genotypes with dominant amplified fragment length polymorphism markers for taxonomic purposes (Jacobs et al., 2011), single random seedlings were selected from each PI for genotyping. For analysis of within-population diversity, 10 random individuals were selected from four populations PI243510 (*S. bulbocastanum*), PI545964 (*S. boliviense* Dunal), PI458365 (*S. berthaultii*), and PI320355 (*S. phureja*).

Cultivated diversity was represented by 213 tetraploid lines from the SolCAP Diversity panel (Hirsch et al., 2013; Table S1.1). This cultivated tetraploid diversity panel, hereafter referred to as the "cultivated tetraploid panel", contains primarily North American germplasm (all but eight lines), including 134 cultivars and 79 advanced breeding lines that represent the major market classes of potato. Each of the wild accessions used was sampled from the *Solanum* sect. *Petota* diversity panel.

Single Nucleotide Polymorphism Genotyping

DNA was purified from leaf tissue of each individual using the Qiagen DNeasy Plant Mini kit (Qiagen, Germantown, MD) and assayed on the Infinium 8303 potato SNP array (Felcher et al., 2012) with the Illumina iScan Reader (Illumina, San Diego, CA). The array surveys 8303 biallelic SNP loci designed from transcribed gene sequences (Felcher et al., 2012; Hamilton et al., 2011). Single nucleotide polymorphism genotypes were manually scored in GenomeStudio (Illumina, San Diego, CA) under a diploid model supporting AA, AB, and BB biallelic calls, such that nonhomozygous calls in polyploid accessions were rated as AB heterozygous genotypes (Hirsch et al., 2013). Single nucleotide polymorphisms were filtered to exclude any loci with 10% missing data in the wild species, disagreeing replicate calls across the 24-sample arrays, or loci that mapped to multiple locations in the DM1–3 516 R44 (hereafter referred to as DM) reference genome assembly (Hirsch et al., 2013; Xu et al., 2011). A total of 5023 high confidence SNPs with reliable diploid calling in both the *Solanum* sect. *Petota* and the cultivated tetraploid diversity panels were selected for downstream analyses (Table S1.2).

Single Nucleotide Polymorphism Functional Annotation

Functional annotations of SNP loci were predicted using Annovar (Wang et al., 2010). A variant table containing SNP location and allele data was generated, in addition to refGene database files containing location, structure, and coding information for all potato genes based on version 4.03 annotations from the Potato Genome Sequencing Consortium (Sharma et al., 2013). Based on this data, Annovar predicted the location of variant sites (i.e., exon, intron, or

intergenic) and the functional impact of alternative alleles at each SNP. Results were used to determine exonic versus non-exonic SNPs and which variants alter coding function.

Phylogenetic Analysis and Estimates of Genetic Diversity

Pairwise genetic distances between *Solanum* sect. *Petota* diversity panel genotypes were estimated using a set of 3275 SNPs (Table S1.3; from the 5023 high-confidence SNPs) excluding SNPs located within candidate market genes to avoid bias due to a higher selection of SNPs at these sites. The sequenced DM potato clone (Lightbourn and Veilleux, 2007; Xu et al., 2011), the cultivar Atlantic, and two diploid genetic stocks (RH89–039–16 and SH83–92–488) (Park et al., 2007; Van Os et al., 2006) were included with the *Solanum* sect. *Petota* diversity panel for generation of a phylogenetic tree. Distance-based phylogenetic trees were generated for the core collection with separate subsets of exonic SNPs predicted to have synonymous (2464 markers) and non-synonymous (1307 markers) effects on gene coding sequences, without the exclusion of markers present in the candidate genes. Relative distances within and among four diploid PIs (see the Plant Materials section) were evaluated using SNP data generated for 10 randomly selected individuals from each population. For these populations, SNPs were filtered to remove loci containing more than 20% missing calls in any PI, with a final subset containing 4828 of 5023 total SNPs. Distances were also estimated between clones in the cultivated tetraploid panel as a measure of diversity. Pairwise distance estimates and tree files were generated using PowerMarker v3.25 (Liu and Muse, 2005). Genetic distances are based on Nei's (1972) distance estimate. Trees were estimated for 1000 bootstrapped datasets and used to generate a consensus tree with bootstrap support in PHYLIP (Felsenstein, 1993). A final distance-based tree was then generated based on clustering in the

consensus tree. Tree graphics were generated with FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/, accessed 20 Oct. 2014). PowerMarker also provided estimates of heterozygosity and allele frequency for high-confidence SNP markers. Comparison of allele frequencies between *Solanum* species and cultivated tetraploid potato clones was conducted in the dosage model (heterozygous allelic dosage accounted for in polyploids) using only the 50 diploid *Solanum* sect. *Petota* PIs and a subset of 3041 SNPs with reliable dosage genotype calling (Hirsch et al., 2013). This produced an unbiased comparison of allelic representation across both populations. The significance of allele and genotype frequency differences at each SNP marker was determined with a chi-square test.

Functional Analysis of Single Nucleotide Polymorphisms and Associated Genes

Existing gene annotations of the DM genes, along with the best *Arabidopsis thaliana* (L.) Heynh. gene hit, were used to determine the function of genes associated with the assayed SNPs. Gene Ontology (GO; Ashburner et al., 2000) terms for the annotated DM genes were used for enrichment analyses for loci displaying significant divergence between the *Solanum* sect. *Petota* diversity panel and the cultivated tetraploid panel using a Fisher's exact test of term count data. Testing was conducted at the gene level, where annotated genes containing one or more divergent SNPs were considered to be a single divergent locus, regardless of how many divergent or conserved SNP markers were present.

Results and Discussion

Phylogeny Results

A distance-based phylogeny was generated using allele frequencies of the Infinium 8303 potato SNP loci to assess the relationships between species present within the *Solanum* sect. *Petota* diversity panel. This analysis showed, with limited exceptions, that SNP-based genetic distances largely support the current taxonomy of tuber-bearing *Solanum* species. The resulting phylogenetic tree based on Nei's (1972) distance appeared to distinguish three major groups (Figure 1.2). Group I contains all landrace PIs from the *Solanum* sect. *Petota* diversity panel, along with wild *S. candolleanum* P. Berthault (formerly *S. bukasovii*). The cultivar Atlantic, along with DM (*S. phureja* Hawkes landrace origin) and two diploid genetic stocks, RH89–039–16 and SH83–92–488, also grouped with the landrace accessions. Group II contains all North American accessions, with the exception of *S. verrucosum* Schltdl., and includes South American *S. acaule*, *S. raphanifolium*, and *S. circaeifolium* Bitter. Group III is comprised primarily of South American wild (non-landrace) species. An interesting observation from Group III is its separation of accessions within the *S. brevicaule* species complex into three distinct clades, each of which was more closely related to accessions outside the complex than others within the complex. This split in the *S. brevicaule* complex largely follows the species boundaries defined by previous classification of *S. brevicaule* accessions at the US Potato Genebank, with three separate clades containing (formerly) *S. gourlayi* Hawkes, hexaploid *S. oplocense* Hawkes, and *S. spegazzinii* Bitter with a *diploid S. oplocense*, respectively (Table 1.1).

An important feature of the phylogeny is that, in most cases, accessions within species formed a single clade, being more closely related to one another than to other species. A greater degree of genetic distance among species than among accessions within species indicates that SNP-based distances primarily agree with current taxonomic boundaries of *Solanum* germplasm. The exceptions were *S. boliviense*, *S. candolleanum* and *S. okadae* Hawkes & Hjert.; each included single accessions that failed to cluster with the others and appeared elsewhere in the tree structure (Figure 1.2), suggesting possible errors in classification. In some cases, different species also appeared to be intermingled. Accessions of *S. stoloniferum*, previously three distinct species (*S. stoloniferum*, *S. fendleri* A. Gray, and *S. polytrichon*), mixed with those of *S. hjertingii* Hawkes, suggesting a close and complex relationship between germplasm in series Longipedicellata. These species can freely cross with one another in nature and their reproductive barriers provide isolation from outside groups (Van den Berg et al., 2002). The SNP data suggest a possible distinction between the germplasm of *S. stoloniferum* and those previously classified as *S. polytrichon* Rydb. and *S. fendleri*. Similarly, accessions of the diploid species *S. infundibuliforme* Phil. intermingled with hexaploid *S. brevicaule* accessions (previously *S. oplocense*).

In several cases, genetic relationships among accessions did not agree with their geographical distribution. However, these observations largely agree with previous studies of wild *Solanum* species. The close relationship observed between *S. acaule* germplasm (collected in Argentina and Peru) and Mexican *S. demissum* has been previously described (Debener et al., 1990; Kardolus, 1998) and supports its proposed role as a progenitor species (Spooner et al., 1995). South American diploids *S. circaeifolium* and *S. raphanifolium* grouped more closely to North

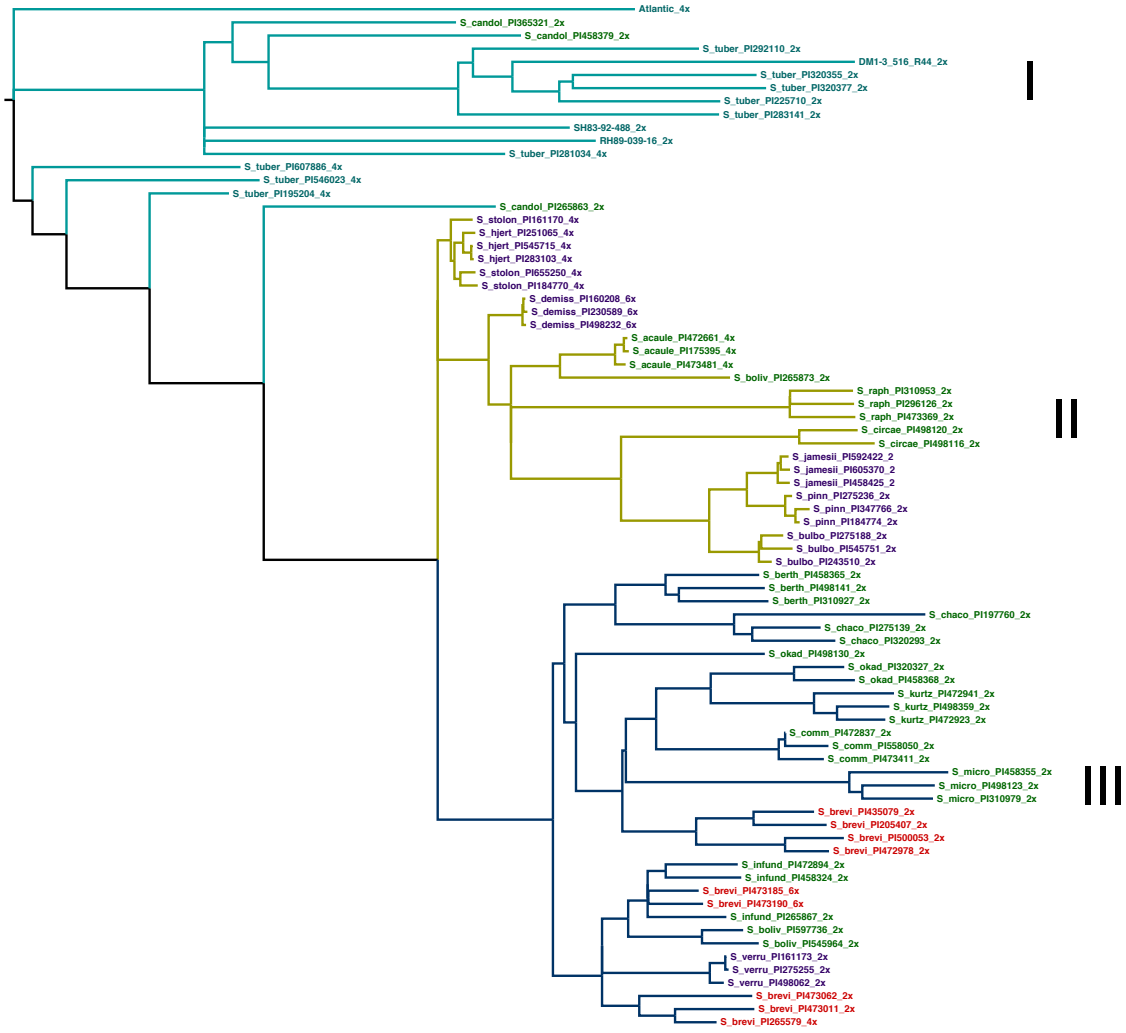

Figure 1.2. ***Phylogenetic tree of Solanum sect. Petota diversity panel genotypes based on Nei's (1972) genetic distance.*** *Groups: (I) cultivated group; (II) primarily North American species group; (III) primarily South American group. Taxon colors: green, South American species; purple, North American species; red, S. brevicaule complex species (South American); turquoise, cultivated germplasm.*

American species. A previous amplified fragment length polymorphism marker study by Kardolus et al. (1998) showed that *S. circaeifolium* is only distantly related to other South American species, possibly explaining its failure to group closely with South American germplasm in this analysis. *S. circaeifolium* is relatively uncommon among South American species in its 1EBN status, a condition prevalent among North American species (Hanneman, 1993; Ortiz and Ehlenfeldt, 1992), which supports its clustering alongside North American 1EBN diploids and suggests they may be distantly related. Furthermore, *S. verrucosum*, a Mexican species, grouped only with South American germplasm, consistent with chloroplast DNA evidence which placed *S. verrucosum* in a clade containing primarily South American diploids (Spooner and Castillo, 1997). *S. verrucosum* is the only North American bridge potato species, and is more reproductively compatible with 2EBN South American diploids (Hanneman, 1993; Ortiz and Ehlenfeldt, 1992). These findings suggest historical migrations of wild potato germplasm away from their regions of origin.

The general concordance of results from this analysis and previous marker studies supports the accuracy of using several thousand conserved SNP markers from cultivated potato for detecting species boundaries and conducting taxonomic surveys in *Solanum* sect. *Petota*, despite the limitation of using single random plant samples to represent populations. The resulting phylogeny largely confirms the current taxonomy of wild potato germplasm, while also indicating a small number of cases in which SNP-based evaluation does not agree with other molecular or morphological approaches. Most notably, this analysis revealed new trends by clearly distinguishing subsets of germplasm within the *S. brevicaule* complex, a large set of wild and semi-cultivated species that were more recently collapsed under a single name

(Miller and Spooner, 1999; Vandenberg et al., 1998). This decision was previously supported by both morphological and molecular data (Alvarez et al., 2008; Miller and Spooner, 1999; Vandenberg et al., 1998). Separation of the species in the *S. brevicaule* complex is an important feature of this phylogeny, particularly because previous studies have had difficulty in differentiating members of this group (Spooner, 2009). This indicates that genome-wide SNP markers can be highly effective for resolving complex species boundaries among germplasm with close genetic relationships. The division of *S. brevicaule* accessions into groups primarily representing previous species classifications (*S. spegazzinii* plus diploid *S. oplocense*, hexaploid *S. oplocense*, and *S. gourlayi*) further supports this distinction. It is also unexpected that each respective subset of the *S. brevicaule* group appears more closely related to other wild *Solanum* species than to each another. This degree of separation is not typical of existing potato studies (Miller and Spooner, 1999; Vandenberg et al., 1998) and suggests that the southern *S. brevicaule* species may not be as similar at the gene sequence level as previously believed.

This phylogeny also contains patterns relevant to potato domestication. The grouping of Peruvian *S. candolleanum* (formerly *S. bukasovii*) with all cultivated accessions supports their status as potential potato landrace progenitors (Hosaka, 1995; Spooner et al., 2005). Both *S. candolleanum* and *S. bukasovii* Juz. are considered "northern brevicaule" species and it has been proposed that this subsection of the *S. brevicaule* complex gave rise to Andean landraces (Spooner et al., 2005). Notably, the "southern" *S. brevicaule* accessions from Argentina do not show as close a relationship with landrace germplasm. The grouping of the cultivar Atlantic in this clade also suggests that landraces are more genetically similar to modern

breeding lines than to most wild species, despite their close geographic origin. The average genetic distance between landrace accessions (including closely related *S. candolleanum*) and the entire panel of 213 cultivars and breeding lines was 0.2081, slightly less than that observed between the landraces and wild species using a diploid genotype model (0.2552; Table S1.4), suggesting that distinct forms of selection by Andean farmers and modern breeders maintain some similarities in their impact on the sequences of various genes in cultivated potato.

To study the prediction of genetic relationships based on markers that are more or less likely to be under selective pressure, separate phylogenetic trees were generated using subsets of SNPs located in the coding sequence and predicted to be either synonymous or non-synonymous variants. The estimated distance relationships among core collection PIs were similar and a Mantel test showed a correlation of 0.9804 between the genetic distance matrices calculated based on either SNP type. The consensus trees generated using synonymous (Figure S1.2) and non-synonymous (Figure S1.3) SNPs were similar and both reflected the species grouping from the overall SNP analysis, with some differences. In both trees, nearly all PIs grouped within their species and cultivated samples, along with the putative progenitor *S. candolleanum*, primarily clustered together. The primary difference was that the synonymous SNP set predicted *S. raphanifolium* to be more closely related to cultivated genotypes, indicating that synonymous SNPs were probably less accurate in this case. The synonymous SNP data also placed a *S. candolleanum* PI (PI265863) in the South American species clade, whereas the non-synonymous SNPs placed it in the same clade as the landraces and other *S. candolleanum* PIs. It is interesting that the predicted distances among

samples were highly correlated across datasets, particularly because neutral markers are typically preferred for studying species-level phylogenetics (Via, 2009). However, estimates of heterozygosity and gene diversity (defined as the likelihood that any two alleles selected from a population will be different) across loci in the two datasets were very similar. Synonymous SNPs had an average gene diversity of 0.194 and a heterozygosity of 9.6% across core collection PIs, while non-synonymous SNPs had an average gene diversity of 0.171 and a heterozygosity of 8.2%. Diversity was slightly lower among non-synonymous SNPs, but the results do not suggest significantly stronger selective pressure on these loci relative to synonymous sites. The Infinium SNP markers were derived from transcribed regions of the genome and the set of 8303 was selected for representation across the potato genome. It is likely that a large number of SNPs are in linkage disequilibrium with other loci under selection in wild populations or cause coding changes without impacting gene function. In many cases, predictions of individual SNP coding effects probably do not fully reflect whether they are truly neutral or under selection. Despite the presence of non-neutral loci in the Infinium 8303 potato SNP array, the full set of 3275 markers, excluding highly selected candidate gene SNPs, provides an accurate representation of species relationships in *Solanum* sect. *Petota*.

Genetic Diversity Within Solanum sect. Petota Accessions

Most PIs consist of multiple genotypes sampled from natural or field populations. These PIs are maintained by genebanks through random intermating to retain the diversity of their source materials (Del Rio et al., 1997). For studying diversity across a broad array of *Solanum* germplasm in this study, single genotypes were used to represent the populations within the

panel due to fiscal limitations. To assess genetic diversity within and among PIs, 10 random individuals from four diploid populations (*S. bulbocastanum* (PI243510), *S. boliviense* (PI545964), *S. berthaultii* (PI458365), and *S. phureja* (PI320355)) were selected and assayed on the Infinium 8303 potato SNP array. Genetic distances within populations were small compared to those among populations (Figure 1.3, Table S1.5). The average pairwise distance among individuals within the *S. phureja* population was highest at 0.0468, whereas the mean distances among individuals within the *S. bulbocastanum*, *S. boliviense*, and *S. berthaultii* populations were 0.0049, 0.0139, and 0.0193, respectively. The lowest mean pairwise distance between individuals from two different populations was found between the two South American wild diploids at 0.0673, several times higher than the average distance among individuals within either population. As shown for these four PIs, a single representative genotype can accurately represent the population and be used to examine interspecies boundaries in the *Solanum* sect. *Petota* diversity panel.

These results may also have implications for the use of wild germplasm in potato breeding, as the loci used to generate distance estimates were selected based on expressed genes in cultivated potato (Hamilton et al., 2011). They suggest many SNPs that are polymorphic in cultivated potato lines contain significantly less variation for individuals within highly divergent wild populations. As a result, introduction of multiple individuals from any single genebank population will be unlikely to substantially increase the diversity of a breeding program at these SNP loci compared with the use of individuals from multiple diverse populations.

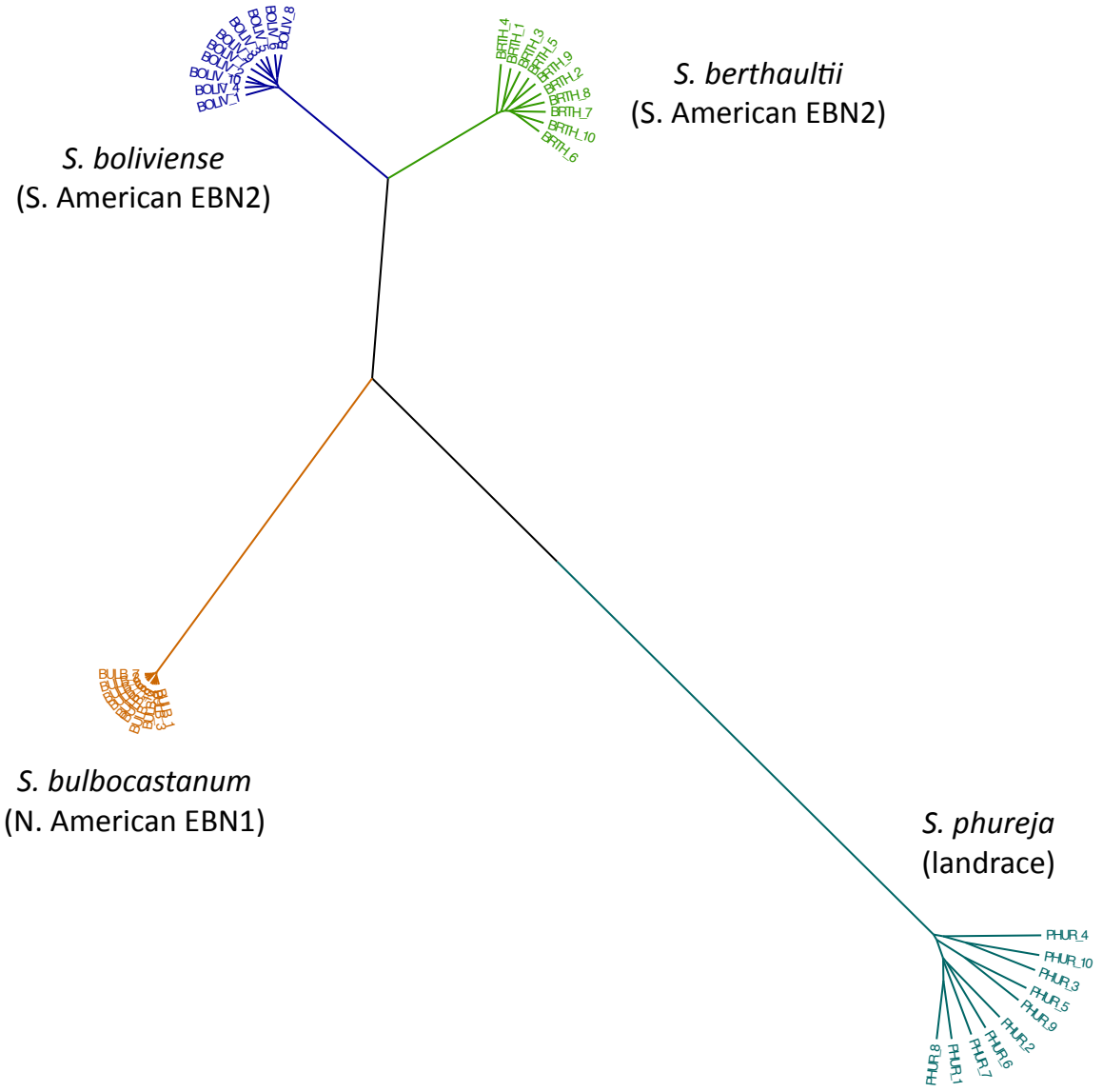

Figure 1.3. ***Phylogenetic tree of four diploid Solanum sect. Petota diversity panel populations (PI243510 (S. bulbocastanum), PI545964 (S. boliviense), PI458365 (S. berthaultii), and PI320355 (S. phureja)) based on Nei's (1972) genetic distance, demonstrating relative diversity within and between populations across the Infinium 8303 potato SNP array loci.*** *Ten random individuals were sampled from each population.*

Heterozygosity and Genetic Diversity in Solanum Germplasm and Cultivated Potato

Relatives of modern cultivated potato have been a source of diversity for improving agronomic performance and introducing novel traits in North American germplasm (Bradshaw et al., 2006; Pavek and Corsini, 2001; Spooner and Bamberg, 1994; Tarn and Tai, 1977). The Infinium 8303 potato SNP array is designed from biallelic SNPs identified in transcribed sequences from cultivated germplasm (Hamilton et al., 2011), limiting its ability to estimate true heterozygosity and genetic diversity across species due to ascertainment bias and design limitations of the Infinium assay. However, a total of 5023 high-confidence SNP loci could be called in both the *Solanum* sect. *Petota* and the cultivated tetraploid diversity panels, suggesting that for these loci, the Infinium 8303 potato SNP array provides an accurate representation of allelic composition in the sampled wild species, landraces, and cultivated lines. Estimates of diversity for species in the *Solanum* sect. *Petota* diversity panel strongly reflected their evolutionary distance from elite cultivated tetraploid germplasm. For 213 cultivated tetraploid clones, average heterozygosity across 5023 high-confidence Infinium SNPs was 57%. For the *Solanum* sect. *Petota* diversity panel, individuals displayed a range of heterozygosity from 0.67 to 37.2%, with values indicating an effect from both ploidy and relationship to cultivated tetraploid germplasm (Figure 1.4). Notably, heterozygosity among North American 1EBN species, hypothesized to be highly diverged from cultivated potato (Hawkes and Jackson, 1992), was consistently the lowest among the *Solanum* sect. *Petota* diversity panel (Figure 1.4), with a minimum of 0.67% for PI592422 (*Solanum jamesii* Torr.). Mexican allopolyploid species were significantly more heterozygous than their diploid counterparts, reflecting allelic diversity among their distinct wild sub-genomes, despite the ascertainment bias. The landrace accessions displayed the highest heterozygosity levels, with

the highest value of 37.2% for PI281034, a tetraploid *S. tuberosum* group Andigenum landrace (Figure 1.4). Further, accessions of *S. candolleanum*, whose status as a close relative and progenitor of diploid landrace populations was supported in this study, were unique as a wild diploid species in displaying heterozygosity that was more comparable to the levels observed in South American diploid landrace germplasm, with 11.1% and 18.0% SNP heterozygosity observed for PI365321 and PI265863, respectively (Figure 1.4). This bias may prevent estimation of true genome-wide diversity across potato germplasm, but provides insight to the distribution of polymorphism within individual wild and cultivated genomes. Assuming wild species possess a higher degree of variation than reported in this study, the Infinium SNP data suggest that selection pressures within natural populations and breeding programs support allelic diversity at discrete loci and that SNPs for which a heterozygous state confers agricultural benefit in cultivars have limited overlap with those for which heterozygosity is selected in wild populations. Analysis of within-population diversity (see the Genetic Diversity Within *Solanum* sect. *Petota* Accessions section) supports this possibility. Notably, *S. phureja*, a diploid cultivated landrace species, showed over twice the degree of within population diversity compared to any other species (Figure 1.3, Table S1.5), whereas South American species were intermediate. *S. bulbocastanum*, a distantly related, sexually incompatible 1EBN species, lacked significant diversity, with an average genetic distance between individuals that was less than half of either South American species. Whole-genome sequence data are still needed to determine the true extent of allelic diversity at conserved genes within wild and cultivated individuals.

An analysis of diversity based on Nei's (1972) genetic distance between individuals revealed

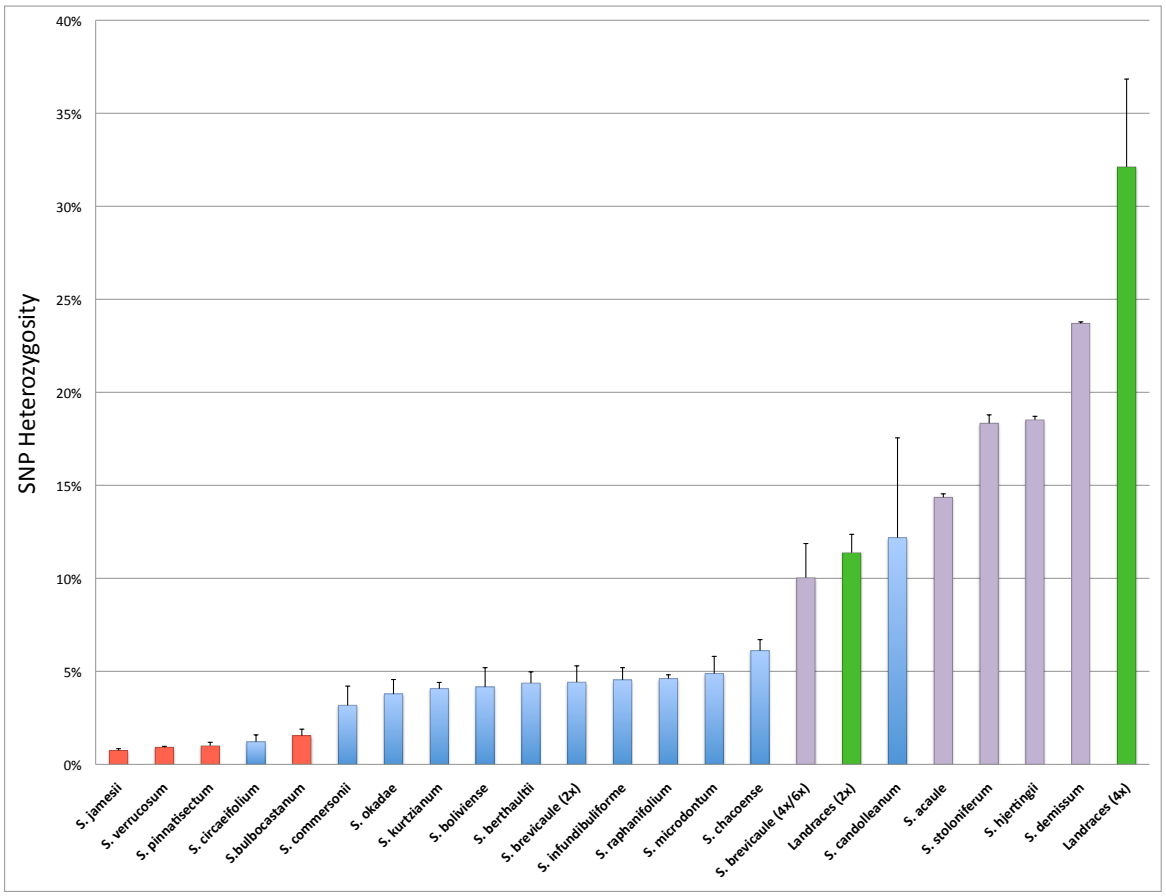

Figure 1.4. ***Mean heterozygosity within Solanum sect. Petota diversity panel species groups across 5,023 cultivated single nucleotide polymorphism markers.*** *Groups: red, North American wild diploid species; blue, South American wild diploid species; purple, wild polyploid species; green, landraces.*

a slightly greater extent of diversity among *Solanum* sect. *Petota* landrace populations compared with cultivars and advanced breeding lines (Table S1.4). The average distance among all wild and landrace accessions in the *Solanum* sect. *Petota* diversity panel using a diploid genotype model was 0.152, higher than the average genetic distance among cultivated tetraploid genotypes (0.141). Within the Infinium 8303 potato SNP set, diversity among populations of Andean landraces and their relatives was the highest. Separation of individuals from the phylogenetic group containing landraces and *S. candolleanum* (see the results for Group I earlier) removed a substantial amount of diversity from the overall *Solanum* sect. *Petota* diversity panel. The average genetic distance among the remaining wild species (see the results for Groups II and III earlier) was 0.111, less than that observed among individuals in the cultivated tetraploid panel, whereas the distance among individuals in the landrace group (Group I) remained higher at 0.156.

Due to the bias associated with this set of SNPs, these results most probably underestimate the heterozygosity and diversity among different landraces as well as species. Studies using sequence hybridization-based marker techniques have demonstrated the presence of greater allelic heterogeneity within evolutionarily divergent potato species such as *S. bulbocastanum* (Traini et al., 2013). In addition, a substantially smaller number of landraces (n = 9) were evaluated compared with cultivars and breeding lines (n = 213) and therefore, the full extent of landrace diversity as measured by the Infinium 8303 potato SNP set may not have been captured. Although further analysis using unbiased techniques for calling sequence variants, such as low-coverage sequencing, will be needed to reveal the true extent of heterozygosity and genetic diversity in wild species, these findings in which Andean landraces possess more

diversity within a few individuals than the entire cultivated tetraploid panel emphasizes the value of these landrace varieties as a source of allelic diversity for breeding.

Allele Frequency Divergence in Solanum sect. Petota Diversity Panel vs. Cultivated Tetraploid Panel Clones

Domestication and subsequent improvement by plant breeders is a major feature differentiating germplasm found in genebanks from modern cultivated varieties, as the different selective pressures associated with agriculture systems and natural plant habitats have resulted in divergence of the loci associated with cultivation or survival in native habitats. The majority of PIs in the *Solanum* sect. *Petota* diversity panel are wild species, although South American landrace PIs are considered relatively primitive compared to the cultivated tetraploid panel. As such, germplasm in the *Solanum* sect. *Petota* diversity panel have undergone lower levels of domestication relative to cultivars and breeding lines. Allele frequencies were assessed across 3041 high confidence SNPs with reliable dosage genotype calling for diploids in the *Solanum* sect. *Petota* panel (50 PIs) and all individuals in the cultivated tetraploid panel.

Allele frequency differences were then used to identify loci exhibiting major genetic divergence between primitive *Solanum* species and modern cultivated germplasm (Table S1.6). The average difference in allele frequency between the *Solanum* sect. *Petota* panel and the cultivated panel for all 3041 dosage-quality SNPs was 23.8% (SD 18.2%). Few Infinium SNP loci displayed exclusive selection of different alleles in the *Solanum* sect. *Petota* and cultivated tetraploid panel; no allele was found to be entirely specific to either population

(Figure 1.5). To assess loci showing divergence between the *Solanum* sect. *Petota* panel diploids and North American cultivated tetraploids, a chi-square test was used to determine significant differences in allelic composition between the two groups. At a Type I error rate of 0.05, 2480 SNPs (81.5%) were considered divergent between the germplasm panels, while 561 (18.5%) SNPs were conserved. For assessing gene functions associated with divergent SNP loci, 322 markers (~10.6%) showing a difference of at least 50% in allele frequency were evaluated to focus on loci under strong selection in cultivated germplasm.

The most highly divergent Infinium SNP loci were associated with a diverse set of genes (Tables S1.6, S1.7, S1.8), some of which have implications for the divergence of primitive and modern potato germplasm. An example is the KAKTUS gene, an E3 ubiquitin ligase, which contained four SNPs showing major allelic divergence, two with 88.1% and 91.5% differences in allele frequency (dosage model) between the *Solanum* sect. *Petota* panel and the cultivated tetraploids. This gene regulates endoreduplication, a process by which cells undergo chromosomal replication without mitosis (Sugimoto-Shirasu and Roberts, 2003). Endoreduplication is used by many plants to increase the size or metabolic activity of specific cell types or organs during development (El Refy et al., 2004; Sugimoto-Shirasu and Roberts,

2003). Endoreduplication commonly occurs in plant storage organs such as potato tubers and maize endosperm and is associated with establishing a nutrient sink status in these tissues (Chen and Setter, 2003; Larkins et al., 2001). It plays an active role in the development of potato tubers: Chen and Setter (2003) showed that *S. tuberosum* L. cv. Katahdin tubers contain nearly 50% cells with ≥8C nuclear DNA content. The extensive differentiation

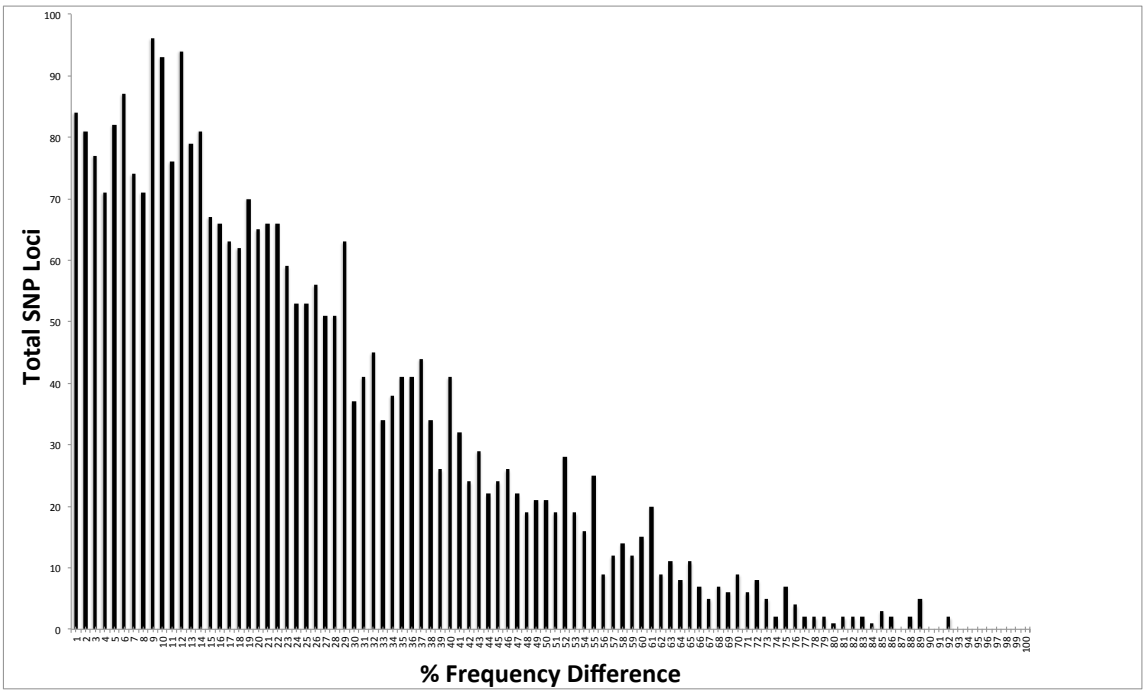

Figure 1.5. ***Distribution of Infinium 8303 potato single nucleotide polymorphism allele frequency differences between 50 Solanum sect. Petota diversity panel wild diploid potatoes and 213 tetraploid potato cultivars.***

observed at multiple loci in the KAKTUS gene may play a role in the distinct tuber characteristics observed for primitive germplasm and high-yielding cultivars (e.g., cell size, yield).

In addition to the KAKTUS gene, several SNP loci with the highest levels of divergence were associated with major regulators of gene expression or cell and organ development (Table S1.6). A homolog of the Arabidopsis curly-leaf (CLF) gene (PGSC0003DMG400034096) contained a SNP with 88.9% allele frequency difference. CLF is part of a family encoding polycomb group complexes, a class of proteins that function in both the plant and animal kingdoms to epigenetically suppress expression of specific genes within cells that have committed to differentiation (Köhler and Villar, 2008). Interference with genes in this family leads to de-repression of regulatory genes controlling apical dominance, leaf and flower development, and flowering time (Katz et al., 2004). *A. thaliana* seedlings lacking a functioning CLF gene and one of its homologs exhibited de-differentiation of cells and formation of callus on differentiated tissues, as well as swollen roots producing shoot-like stem tissues (Chanvivattana et al., 2004). A homolog of the A. thaliana HAT9 histone acetyltransferase transcription factor (PGSC0003DMG400021423) contained a SNP with 91.3% allele frequency difference. Histone acetyltransferases also have the potential for major developmental impacts, as they are able to globally regulate gene expression through the modification of chromatin epigenetic states (Pandey et al., 2002). A bZIP transcription factor homologous to *A. thaliana* NPY2 (PGSC0003DMG400014877) contained a SNP with 88.2% allele frequency difference. NPY genes are key regulators of organ development, controlling auxin-mediated organogenesis via the regulation of genes involved in auxin biosynthesis

(Cheng et al., 2008). Although it is difficult to associate these genes with specific impacts on traits found in cultivated potato, identification of highly divergent SNPs within large-effect regulatory genes underscores the major changes in plant development that potato experienced during its domestication. Furthermore, the presence of divergent SNPs in genes with regulatory functions supports their role in the domestication of potato, consistent with the identification of transcription factors as major domestication genes in crops (Doebley et al., 2006; Olsen and Wendel, 2013).

A variety of genes known to influence tuber development via carbohydrate metabolism were also found to contain SNPs with highly divergent allele frequencies between wild diploids and cultivated tetraploids. These included a sucrose transporter, a sucrose synthase (a SUS1 homolog) with three divergent loci, multiple invertases, uridine diphosphate-glucose pyrophosphorylase, fructokinase, phosphofructokinase, multiple starch synthases, and a starch branching enzyme (Tables S1.6, S1.8). These genes are all closely linked with carbohydrate biosynthesis and accumulation in potato tubers (Fernie et al., 2002; Geigenberger, 2003; Haake et al., 1998; Sturm, 1999). Selection for different tuber characteristics was a major component in the domestication of potato from wild germplasm. One of the most important changes was increased transport and partitioning of carbohydrates to underground tubers and higher biomass accumulation leading to greater size and yield. The SNP results suggest that domestication imposed selective pressure on the majority of genes found in the primary carbohydrate metabolic pathway. The fixation of different alleles within these genes is likely to be a major factor contributing to the larger sized tubers produced by modern cultivated varieties compared with wild species. The presence of three divergent SNP loci within the

SUS1 sucrose synthase homolog offers an ideal example. Sucrose synthase catalyzes conversion of sucrose to glucose and fructose in the cytosol of developing tubers, which can then be converted to starch. Transgenic studies have shown that increased expression of sucrose synthase in potato results in greater starch accumulation and final yield (Baroja Fernandez et al., 2009).

The results of this study point to carbohydrate pathway genes as key players underlying the tuber characteristics of high-yielding potato cultivars. Unfortunately, the Infinium 8303 potato array contains only a small subset of the potential variants found in most potato genes (Hamilton et al., 2011). Efforts should be made to sequence these genes in a variety of species and cultivars to summarize their full complement of alleles. This will better provide clues to the functional impact of "cultivated" alleles on potato carbohydrate metabolism, and identify those associated with desirable tuber qualities and may ultimately provide useful genetic markers in selecting for yield.

Allele frequency differences were also compared between the *Solanum* sect. *Petota* panel diploids and cultivated tetraploid panel separately based on subsets of SNPs with different gene locations and predicted functions. For variants in coding regions, no major disparity in allele frequency differences was found between neutral variants and those altering protein sequences. The average allele frequency difference for 1550 synonymous SNP loci was 23.61% and was only slightly higher for 757 non-synonymous SNP loci at 24.23%. As the average allele frequency difference between the wild species and cultivars is similar for synonymous and non-synonymous SNPs, it does not appear that significantly greater selective

pressure is being placed on the Infinium SNP loci impacting coding function. This is unexpected, because non-synonymous SNPs are more likely to affect gene function and therefore to come under selection. However, resequencing efforts in *A. thaliana* showed that projection of sequence variants on a single reference annotation could be misleading because of alternative splice forms that are not represented in the reference annotation (Gan et al., 2011).

To determine whether GO terms associated with divergent loci support the functional role of genes containing SNPs under strong selection, an enrichment test was conducted with subsets of divergent and conserved loci (Table S1.9). As a disproportionately large number of markers were considered divergent at a Type I error rate of 0.05 (81.5%), to permit better comparison of functional enrichment, the error rate for considering a locus as divergent was made more stringent and set to 0.001, which resulted in 1962 divergent SNP loci (64.5%) and 1079 conserved SNP loci (35.5%). Testing was performed at the gene level, in which terms were counted once for a gene regardless of the number of divergent SNP loci it contained.

Gene ontology term enrichment confirmed the results, suggesting selection on carbohydrate-related loci. The group of genes containing one or more divergent SNP loci (chi-square P-value 0.001) was enriched for the term "carbohydrate metabolic process" (P-value 0.0187). Several other biological process GO terms also showed enrichment in this group, including "reproduction" (P-value 0.0007), "lipid metabolic process" (P-value 0.0013), and "generation of precursor metabolites and energy" (P-value 0.0465). Strong enrichment for reproductive genes also suggests major divergence in this process. Diploid species are known to sexually

propagate in wild populations, whereas modern tetraploid breeding lines are exclusively maintained by asexual reproduction. It appears that mutations differentiating these groups have accumulated in reproductive genes and become common among cultivated germplasm. The relatively low number of meiotic events required in asexually propagated crop species could play a role by reducing selection against mutations in the genes required for flowering. Tuber production is also closely tied to flowering and the maturity response (Abelenda et al., 2011;Kloosterman et al., 2013;Martin et al., 2009). It may be that selection for tuber traits, favoring increased vegetative reproduction and nutrient partitioning at maturity in particular, has indirectly affected a number of genes involved in sexual reproduction.

Conclusions

Use of an 8303 Infinium SNP array designed using polymorphisms from cultivated potato resulted in a phylogeny of *Solanum* sect. *Petota* that was consistent with existing taxonomic classifications, suggesting that polymorphisms within conserved genes provide a robust representation of key genetic differences in cultivated and wild species of potato. Applying the array to a diverse set of both species and cultivars identified multiple SNP loci and genes with extreme divergence between primitive and modern cultivated germplasm. Some of these may be critical in differentiating high-yielding cultivars from their wild relatives. The presence of divergent SNP loci within transcriptional regulators and genes that encode proteins involved in carbohydrate metabolism and tuber development supports this hypothesis, as they play major roles in crop domestication and potato breeding. These genes represent ideal candidates for marker-assisted selection in breeding programs. With the cost of generating genome sequence data becoming more affordable, efforts should be made to

sequence a larger number of wild species and cultivated varieties to permit interrogation of all genomic variants and identify the full complement of loci contributing to potato domestication.

LITERATURE CITED

LITERATURE CITED

Abelenda, J.A., Navarro, C., and Prat, S. (2011). From the model to the crop: genes controlling tuber formation in potato. Curr. Opin. Biotechnol. 22, 287-292.

Alvarez, N.M., Peralta, I., Salas, A., and Spooner, D.M. (2008). A morphological study of species boundaries of the wild potato *Solanum brevicaule* complex: replicated field trials in Peru. Plant Syst. Evol. 274, 37-45.

Ames, M., Salas, A., and Spooner, D.M. (2008). A morphometric study of species boundaries of the wild potato *Solanum* series Piurana (*Solanaceae*) and putatively related species from seven other series in *Solanum* sect. *Petota*. Syst. Bot. 33, 566-578.

Ashburner, M., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

Bamberg, J., and Del Rio, A. (2005). Conservation of genetic resources. In Book Chapter, pp. 451.

Baroja-Fernandez, E., Munoz, F.J., Montero, M., Etxeberria, E., Sesma, M.T., Ovecka, M., Bahaji, A., Ezquer, I., Li, J., Prat, S., and Pozueta-Romero, J. (2009). Enhancing sucrose synthase activity in transgenic potato (*Solanum tuberosum* L.) tubers results in increased levels of starch, ADP-glucose and UDP-glucose and total yield. Plant Cell Physiol. 50, 1651-1662.

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

Chanvivattana, Y., Bishopp, A., Schubert, D., Stock, C., Moon, Y.-H., Sung, Z.R., and Goodrich, J. (2004). Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. Development 131, 5263-5276.

Chen, C.T., and Setter, T.L. (2003). Response of potato tuber cell division and growth to shade and elevated CO2. Ann. Bot. 91, 373-381.

Cheng, Y., Qin, G., Dai, X., and Zhao, Y. (2008). NPY genes and AGC kinases define two key steps in auxin-mediated organogenesis in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 105, 21017-21022.

Daly, D.C., Cameron, K.M., and Stevenson, D.W. (2001). Plant systematics in the age of genomics. Plant Physiol. 127, 1328-1333.

Debener, T., Salamini, F., and Gebhardt, C. (1990). Phylogeny of wild and cultivated *Solanum* species based on nuclear restriction fragment length polymorphisms (RFLPs). Theor. Appl. Genet. 79, 360-368.

Del Rio, A., Bamberg, J., and Huaman, Z. (1997). Assessing changes in the genetic diversity of potato gene banks. 1. Effects of seed increase. Theor. Appl. Genet. 95, 191-198.

Demeke, T., Lynch, D., Kawchuk, L., Kozub, G., and Armstrong, J. (1996). Genetic diversity of potato determined by random amplified polymorphic DNA analysis. Plant Cell Rep. 15, 662-667.

Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell 127, 1309-1321.

El Refy, A., Perazza, D., Zekraoui, L., Valay, J., Bechtold, N., Brown, S., Hülskamp, M., Herzog, M., and Bonneville, J.-M. (2004). The Arabidopsis KAKTUS gene encodes a HECT protein and controls the number of endoreduplication cycles. Mol. Genet. Genomics 270, 403-414.

Felcher, K.J., Coombs, J.J., Massa, A.N., Hansey, C.N., Hamilton, J.P., Veilleux, R.E., Buell, C.R., and Douches, D.S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. PLoS ONE 7, e36347.

Felsenstein, J. (1993). PHYLIP: phylogenetic inference package. Dep. of Genetics, Univ. Washington, Seattle. http://evolution.genetics.washingtion.edu/phylip.html (accessed 5 Nov. 2014).

Fernie, A.R., Willmitzer, L., and Trethewey, R.N. (2002). Sucrose to starch: a transition in molecular plant physiology. Trends Plant Sci. 7, 35-41.

Gan, X., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477, 419-423.

Geigenberger, P. (2003). Regulation of sucrose to starch conversion in growing potato tubers. J. Exp. Bot. 54, 457-465.

Haake, V., Zrenner, R., Sonnewald, U., and Stitt, M. (1998). A moderate decrease of plastid aldolase activity inhibits photosynthesis, alters the levels of sugars and starch, and inhibits growth of potato plants. Plant J. 14, 147-157.

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

Hanneman Jr, R. (1993). Assignment of endosperm balance numbers to the tuber-bearing *Solanums* and their close non-tuber-bearing relatives. Euphytica 74, 19-25.

Hawkes, J., and Jackson, M. (1992). Taxonomic and evolutionary implications of the Endosperm Balance Number hypothesis in potatoes. Theor. Appl. Genet. 84, 180-185.

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

Hijmans, R.J., and Spooner, D.M. (2001). Geographic distribution of wild potato species. Am. J. Bot. 88, 2101-2112.

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R.E., Jansky, S., and Bethke, P. (2013). Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3 Genes Genom. Genet. 3, 1003-1013.

Hosaka, K. (1995). Successive domestication and evolution of the Andean potatoes as revealed by chloroplast DNA restriction endonuclease analysis. Theor. Appl. Genet. 90, 356-363.

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., and Li, W. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature 490, 497-501.

Hufford, M.B., et al. (2012). Comparative population genomics of maize domestication and improvement. Nat. Genet. 44, 808-811.

Jacobs, M.M., Smulders, M.J., van den Berg, R.G., and Vosman, B. (2011). What's in a name; Genetic structure in *Solanum* section *Petota* studied using population-genetic tools. BMC Evol. Biol. 11, 42.

Jansky, S., Dempewolf, H., Camadro, E., Simon, R., Zimnoch-Guzowska, E., Bisognin, D., and Bonierbale, M. (2013). A Case for Crop Wild Relative Preservation and Use in Potato. Crop Sci. 53, 746-754.

Kardolus, J.P. (1998). A biosystematic analysis of *Solanum acaule* (Doctoral dissertation, Landbouwuniversiteit Wageningen).

Kardolus, J.P., van Eck, H.J., and van den Berg, R.G. (1998). The potential of AFLPs in biosystematics: a first application in *Solanum* taxonomy (*Solanaceae*). Plant Syst. Evol. 210, 87-103.

Katz, A., Oliva, M., Mosquna, A., Hakim, O., and Ohad, N. (2004). FIE and CURLY LEAF polycomb proteins interact in the regulation of homeobox gene expression during sporophyte development. Plant J. 37, 707-719.

Kloosterman, B., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495, 246-250.

Köhler, C., and Villar, C.B. (2008). Programming of gene expression by Polycomb group proteins. Trends Cell Biol. 18, 236-243.

Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.L., Li, M.-W., He, W., Qin, N., and Wang, B. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42, 1053-1059.

Larkins, B.A., Dilkes, B.P., Dante, R.A., Coelho, C.M., Woo, Y.M., and Liu, Y. (2001). Investigating the hows and whys of DNA endoreduplication. J. Exp. Bot. 52, 183-192.

Lightbourn, G.J., and Veilleux, R.E. (2007). Production and evaluation of somatic hybrids derived from monoploid potato. Am. J. Potato Res. 84, 425-435.

Liu, K., and Muse, S.V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21, 2128-2129.

Martin, A., Adam, H., Díaz-Mendoza, M., Żurczak, M., González-Schain, N.D., and Suárez-López, P. (2009). Graft-transmissible induction of potato tuberization by the microRNA miR172. Development 136, 2873-2881.

Mendoza, H. (1989). Population breeding as a tool for germplasm enhancement. Am. Potato J. 66, 639-653.

Miller, J.T., and Spooner, D.M. (1999). Collapse of species boundaries in the wild potato *Solanum brevicaule* complex (*Solanaceae*, *S.* sect. *Petota*): molecular data. Plant Syst. Evol. 214, 103-130.

Nei, M. (1972). Genetic distance between populations. Am. Nat., 283-292.

Olsen, K.M., and Wendel, J.F. (2013). A bountiful harvest: genomic insights into crop domestication phenotypes. Annu. Rev. Plant Biol. 64, 47-70.

Ortiz, R., and Ehlenfeldt, M.K. (1992). The importance of endosperm balance number in potato breeding and the evolution of tuber-bearing *Solanum* species. Euphytica 60, 105-113.

Pandey, R., Müller, A., Napoli, C.A., Selinger, D.A., Pikaard, C.S., Richards, E.J., Bender, J., Mount, D.W., and Jorgensen, R.A. (2002). Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes. Nucleic Acids Res. 30, 5036-5055.

Park, T.H., Kim, J.B., Hutten, R.C., van Eck, H.J., Jacobsen, E., and Visser, R.G. (2007). Genetic positioning of centromeres using half-tetrad analysis in a 4x–2x cross population of potato. Genetics 176, 85-94.

Pavek, J., and Corsini, D. (2001). Utilization of potato genetic resources in variety development. Am. J. Potato Res. 78, 433-441.

Qi, J., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat. Genet. 45, 1510-1515.

Sharma, K.S., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 Genes Genom. Genet. 3, 2031-2047.

Spooner, D.M., and Castillo, R. (1997). Reexamination of series relationships of South American wild potatoes (*Solanaceae*: *Solanum* sect. *Petota*): evidence from chloroplast DNA restriction site variation. Am. J. Bot. 84, 671-671.

Spooner, D.M., and Salas, A. (2006). Structure, biosystematics, and genetic resources. Handbook of potato production, improvement, and postharvest management/J. Gopal, SM Paul Khurana, editors.

Spooner, D.M. (2009). DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. Am. J. Bot. 96, 1177-1189.

Spooner, D.M., and Bamberg, J.B. (1994). Potato genetic resources: sources of resistance and systematics. Am. Potato J. 71, 325-337.

Spooner, D.M., van den Berg, R.G., and Bamberg, J.B. (1995). Examination of species boundaries of *Solanum* series Demissa and potentially related species in series Acaulia and series Tuberosa (sect. *Petota*). Syst. Bot. 20, 295-314.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. U. S. A. 102, 14694-14699.

Sturm, A. (1999). Invertases. Primary structures, functions, and roles in plant development and sucrose partitioning. Plant Physiol. 121, 1-8.

Sugimoto-Shirasu, K., and Roberts, K. (2003). "Big it up": endoreduplication and cell-size control in plants. Curr. Opin. Plant Biol. 6, 544-553.

Tarn, T., and Tai, G. (1977). Heterosis and variation of yield components in F1 hybrids between Group Tuberosum and Group Andigena potatoes. Crop Sci. 17, 517-521.

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science 327, 818-822.

Traini, A., Iorizzo, M., Mann, H., Bradeen, J.M., Carputo, D., Frusciante, L., and Chiusano, M.L. (2013). Genome Microscale Heterogeneity among Wild Potatoes Revealed by Diversity Arrays Technology Marker Sequences. Int. J. Genomics 2013, 9.

Van den Berg, R., Bryan, G., Del Rio, A., and Spooner, D.M. (2002). Reduction of species in the wild potato *Solanum* section *Petota* series Longipedicellata: AFLP, RAPD and chloroplast SSR data. Theor. Appl. Genet. 105, 1109-1114.

Van den Berg, R., Miller, J., Ugarte, M., Kardolus, J., Villand, J., Nienhuis, J., and Spooner, D.M. (1998). Collapse of morphological species in the wild potato *Solanum brevicaule* complex (*Solanaceae*: sect. *Petota*). Am. J. Bot. 85, 92-92.

Van Os, H., Andrzejewski, S., Bakker, E., Barrena, I., Bryan, G.J., Caromel, B., Ghareeb, B., Isidore, E., de Jong, W., and Van Koert, P. (2006). Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173, 1075-1087.

Via, S. (2009). Natural selection in action during speciation. Proc. Natl. Acad. Sci. U. S. A. 106, 9939-9946.

Wang, K., Li, M., Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 85, e164.

Xu, X., et al.; Potato Genome Sequence Consortium (2011). Genome sequence and analysis of the tuber crop potato. Nature. 475, 189-195.

CHAPTER 3

GENOME REDUCTION UNCOVERS A LARGE DISPENSABLE GENOME AND ADAPTIVE ROLE FOR COPY NUMBER VARIATION IN ASEXUALLY PROPAGATED *SOLANUM TUBEROSUM*

[Published in: The Plant Cell 28 (2): 388-405]

Michael A. Hardigan[1], Emily Crisovan[1], John P. Hamilton[1], Jeongwoon Kim[1], Parker Laimbeer[2], Courtney P. Leisner[1], Norma C. Manrique-Carpintero[3], Linsey Newton[1], Gina M. Pham[1], Brieanne Vaillancourt[1], Xueming Yang[4,5], Zixian Zeng[4], David S. Douches[3], Jiming Jiang[4], Richard E. Veilleux[2], and C. Robin Buell[1,*]

[1]Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824

[2]Department of Horticulture, Virginia Tech, Blacksburg, Virginia 24061

[3]Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, Michigan 48824

[4]Department of Horticulture, University of Wisconsin, Madison, Wisconsin 53706

[5]Institute of Biotechnology, Provincial Key Laboratory of Agrobiology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

*Corresponding author (buell@msu.edu).

Abstract

Clonally reproducing plants have the potential to bear a significantly greater mutational load than sexually reproducing species. To investigate this possibility, we examined the breadth of genome-wide structural variation in a panel of monoploid/doubled monoploid clones generated from native populations of diploid potato (*Solanum tuberosum*), a highly heterozygous asexually propagated plant. As rare instances of purely homozygous clones, they provided an ideal set for determining the degree of structural variation tolerated by this species and deriving its minimal gene complement. Extensive copy number variation (CNV) was uncovered, impacting 219.8 Mb (30.2%) of the potato genome with nearly 30% of genes subject to at least partial duplication or deletion, revealing the highly heterogeneous nature of the potato genome. Dispensable genes (>7,000) were associated with limited transcription and/or a recent evolutionary history, with lower deletion frequency observed in genes conserved across angiosperms. Association of CNV with plant adaptation was highlighted by enrichment in gene clusters encoding functions for environmental stress response, with gene duplication playing a part in species-specific expansions of stress-related gene families. This study revealed unique impacts of CNV in a species with asexual reproductive habits and how CNV may drive adaption through evolution of key stress pathways.

Abbreviations: BAC, bacterial artificial chromosome; CNV, copy number variation; DM, DM1–3 516 R44 reference genome assembly; FISH, fluorescence in situ hybridization; FPKM, fragments per kilobase per million mapped reads; PAV, presence absence variation; SNP, single nucleotide polymorphism.

Introduction

Cultivated potato (*Solanum tuberosum*) comprises a unique plant species (Gavrilenko et al., 2013; Hirsch et al., 2013; Uitdewilligen et al., 2013), consisting primarily of diverse diploid and tetraploid subspecies that can harbor introgressions from various wild populations (Hawkes, 1990; Spooner et al., 2007). Varieties and landraces are maintained as clones in vitro or by collection and planting of seed tubers, yielding significant potential for accumulating somatic mutations in the genome. The most widely grown variety in North America, Russet Burbank, has been maintained clonally for over 100 years and was itself selected as a somatic mutant of an older variety. The asexual and highly heterozygous nature of potato offers a unique model to examine genome variation compared with homozygous, or seed-propagated, plants, such as *Arabidopsis thaliana*, soybean (*Glycine max*), and maize (*Zea mays*). Without routine meiotic events imposing purifying selection at each generation (Simko et al., 2006), mutations have the potential to be retained at higher levels than in species tolerant of inbreeding and are more likely mitotic in origin. The mutation load in cultivated backgrounds is extremely high (Xu et al., 2011), demonstrated by low fertility in elite clones and severe inbreeding depression observed during selfing (De Jong and Rowe, 1971).

Sequence-level mutations, including single nucleotide polymorphisms (SNPs) and small insertions/deletions, have been widely investigated in several plant species (Morrell et al., 2011). With respect to structural variation, recent genome-wide surveys using array and sequencing technologies have revealed copy number variants and presence/absence variants from hundreds to millions of bases in length are prevalent in plants and animals (Abecasis et

al., 2012; Żmieńko et al., 2014), supporting their importance as components of genome diversity in eukaryotes. A growing body of evidence now suggests they play a key role underlying phenotypic diversity. While often associated with likelihood of genetic disorders in mammals (Weischenfeldt et al., 2013), copy number variation (CNV) has been shown to benefit adaptive traits in plants, such as day length neutrality in wheat (*Triticum aestivum*; Díaz et al., 2012), and is speculated to be an underlying component of hybrid vigor (Lai et al., 2010). At the functional level, CNV has also been linked to genes involved in stress responses, such as submergence tolerance in rice (*Oryza sativa*; Xu et al., 2006; Hattori et al., 2009), nematode resistance in soybean (Cook et al., 2012), and aluminum tolerance in maize (Maron et al., 2013). While genome-wide structural variation studies in maize (Chia et al., 2012), soybean (Lam et al., 2010), and Arabidopsis (Cao et al., 2011) have shown that CNV patterns are widespread and exhibit different frequency among sexually reproducing plant species, the impact of structural variation on genome and phenotypic diversity has yet to be explored in any clonally propagated plant.

The richest source of genomic variation for *S. tuberosum* exists among its native South American progenitors (Ortiz, 2001). SNPs derived from elite North American cultivars show greater variation among South American landraces than modern clones and their wild relatives (Hardigan et al., 2015), demonstrating the diversity in native populations of cultivated potato. Unlike sequence level mutation, the contribution of structural variation to this diversity remains undetermined in this clonally propagated plant species. Limited CNV analyses performed at the cytogenetic level (Iovene et al., 2013) with select BAC-sized regions showed large tracts of the potato genome (>100 kb) are commonly absent from

multiple homologous chromosomes of autotetraploids, supporting extensive genome plasticity.

We present an analysis of structural variation in diploid *S. tuberosum*, an asexually reproducing and obligate outcrossing species, based on next-generation sequencing. This study examined a panel of 12 monoploid/doubled monoploid clones derived from native South American landrace populations, selected for their rare, nonlethal introduction of full homozygosity into this highly heterozygous genome. This panel reflected more structural variation within 12 related *S. tuberosum* clones than previous plant studies encompassing much larger data sets, suggesting greater tolerance of mutation in populations of asexually reproducing species. The underlying causes could be masking of dysfunctional and deleterious alleles in a heterozygous state and an inability to purge deleterious alleles via meiosis. Thousands of CNVs including duplications, deletions, and presence/absence variation (PAV) were identified in all clones, including those closely related to the reference genotype, with variants larger than 100 kb frequently observed in pericentromeric regions. As these homozygous clones were capable of growth and development *ex vitro*, we were able to annotate many dispensable genes and estimate the core gene set required for survival. While we observed a low frequency of deletions in genes encoding functions conserved across angiosperms, CNV was shown to be closely associated with loci involved in stress tolerance, supporting the concept of an adaptive role for gene duplication in diversification of plant environmental responses. Finding that nearly half the genes specific to the potato lineage were impacted by duplication or deletion reinforced the connection between CNV and evolution of novel genes at the species level.

Materials and Methods

Germplasm

The potato clones in this study were anther-culture generated monoploids and doubled monoploids derived primarily from three accessions of a long photoperiod adapted population of diploid *Solanum tuberosum* Group Phureja landraces (Haynes, 1972) with limited introgression from wild *Solanum chacoense* and dihaploids of cultivated *S. tuberosum* Group Tuberosum (Figure S 2.1). All but M6 were able to grow under normal greenhouse conditions and produced both flowers and tubers. Ploidy is reported based on original flow cytometry analysis. Several monoploids (M1, M5, M7, M8, and M9) underwent spontaneous chromosome doubling in tissue culture since initial ploidy confirmation and are now doubled monoploids.

Improved Assembly of the Potato Reference Genome Sequence (DM v4.04)

Genomic DNA was isolated from DM stem and leaf tissue using the cetyltrimethyl ammonium bromide method, sheared to 300 bp using a Covaris ultrasonicator, end repaired, A-tailed, ligated to Illumina compatible adaptors, and PCR amplified for eight cycles. Cleaned DM genomic reads that did not map to the DM v4.03 assembly (31.5 million pairs and 1.4 million singletons; Sharma et al., 2013) were assembled into contigs using Velvet (v1.2.10) (Zerbino and Birney, 2008) using a k-mer size of 61 and minimum contig length of 200 bp. Contigs were searched against the v4.03 assembly using WUBLAST and excluded if they aligned with ≥97% identity and ≥30% coverage. Remaining contigs represented novel DM sequences absent in the v4.03 assembly (Sharma et al., 2013). These were searched using BLAST against the NCBI nr database to remove contaminants. The final, filtered contigs

represent 55.7 Mb of novel DM sequence and were concatenated by order of length into a pseudomolecule "chrUn" with 500 bp gaps. The new DM v4.04 assembly is the addition of the chrUn pseudomolecule to the existing v4.03 genome assembly (Sharma et al., 2013). Contigs were annotated using the MAKER pipeline (r112) (Cantarel et al., 2008).

Monoploid and Doubled Monoploid Genomic, Transcriptomic, Epigenomic Datasets

DNA was isolated from monoploid and doubled monoploid leaf tissue using the Qiagen DNeasy Plant Mini kit, sheared to ~200 bp and 600 to 700 bp using a Covaris ultrasonicator, and Illumina TruSeq libraries were constructed. For M6, Illumina compatible libraries were constructed as described above for DM. Libraries were sequenced in paired-end mode generating 100 nucleotide reads on the Illumina HiSeq platform, yielding a combined coverage of ~30 to 69X for each clone (Table S2.1). Total RNA was extracted from monoploid and doubled monoploid leaf and tuber tissues using the Qiagen RNeasy Plant Mini kit, and RNA-seq libraries were prepared using the TruSeq mRNA kit. RNA-seq libraries were sequenced in the single-end mode on the Illumina HiSeq platform generating 50-nucleotide reads, yielding 26 to 57 M reads per clone. ChIP-seq data were generated from the DM reference genotype using antibodies for two histone marks associated with transcribed genes, H3K4me2 and H4K5,8,12,16ac as previously described (Yan et al., 2008). Immunoprecipitated DNA samples from mature leaf and tuber tissue were used for library construction with the same steps as other DNA libraries (with the exception of 13 PCR cycles) and sequenced on an Illumina HiSeq in paired-end mode with 100 nucleotide reads.

Variant Calling

Whole-genome sequence and RNA-seq reads were cleaned using Cutadapt (v1.2.1) (Martin, 2011), using a minimum base quality of 10 and a minimum read length of 30 bp after trimming. The first 10 bases were trimmed from the 59 ends of genomic DNA reads and the first base from the 59 ends of RNA-seq to remove sequence bias. Genomic reads were mapped to the DM v4.04 potato genome assembly in paired-end mode using BWA-MEM (v0.7.8) (Li, 2013) with default parameters. Duplicates were marked using PicardTools (v1.106; http://broadinstitute.github.io/picard). GATK IndelRealigner (v2.8.1) (McKenna et al., 2010) was used to refine alignments, and SAMTools (v0.1.19) (Li et al., 2009) was used to merge the 200 and 600 bp library BAM files for downstream SNP and CNV calling. RNA-seq reads were mapped to the DM v4.04 assembly using TopHat (v1.4.1) (Trapnell et al., 2009) with minimum and maximum intron lengths of 10 and 15,000 bp, respectively, allowing for up to three mismatches in the seed alignment.

SNP calls were generated with SAMTools mpileup and converted to VCF format with bcftools (v0.1.19; http://samtools.github.io/bcftools/); calls were filtered in VCFtools (v0.1.11) (Danecek et al., 2011) using criteria D=100/Q=20/q=10/d=5/r and refiltered on a per-sample basis with maximum SNP read coverage set to each sample's theoretical coverage. A custom script was used to select homozygous calls with a minimum SNP quality of 100 and minimum genotype quality of 80. SNP function was predicted using Annovar (Wang et al., 2010). SNP calls were compared with allele calls on the same clones using the Infinium 8303 potato array (Felcher et al., 2012).

CNVs were called from genomic BAM files based on read depth using CNVnator (Abyzov et al., 2011) with a window size of 100 bp. Raw CNV calls were filtered using quality scores generated by the software with a cutoff P-value of 0.05, removing many small deletions (<500 bp) with low support. As quality scores were much lower for small intergenic CNVs, those below 500 bp were removed. CNV regions containing an N-content above 10% in the reference sequence were also removed. To account for mapping bias and errors in the reference assembly, we generated CNV calls by mapping reads from the DM reference genotype to its own assembly. In total, 139 genes were found missing based on DM self-CNV analysis and excluded as annotation artifacts. Copy number estimates generated from the DM reference genotype that were above or below a single copy were considered as mapping bias or errors in the reference assembly, and custom scripts were used to adjust copy number estimates in the monoploid panel based on these values. To limit analysis of variants to a set of high confidence calls, we considered regions with a copy number estimate between 0.8 and 1.4 indistinguishable from single copy regions and excluded from further analysis. BEDTools (Quinlan and Hall, 2010) and custom scripts were used to determine CNV-gene overlaps and assign gene copy number. For confident association, a CNV had to span at least half the gene model. Genes for which a CNV covered at least half an exon but less than half the gene model were considered partially duplicated or deleted.

To assess the sensitivity and specificity of CNVnator to identify structural variants, we performed a custom read depth analysis. Median read depths were calculated in 100 bp windows and divided by whole genome median coverage to obtain relative window coverage. Window estimates were then normalized based on DM mapping bias. Adjacent windows with

high or low coverage were concatenated to form CNV blocks, merging nearby blocks within 200 bp. Genotypes were calculated as the mean of all individual window estimates within a block. CNV blocks were removed if they contained 10% N-content, were shorter than 500 bp, and if they occurred in regions where >80% of samples were called as CNVs (regions with significant mapping bias). For validation, CNVnator calls were required to have at least 50% coverage by CNVs of the same class from the read depth method. To experimentally validate structural variant calls, deletions were randomly assessed using PCR with multiple computationally predicted single-copy and variant (duplicate or deletion) clones (Dataset S2.12). Reaction conditions were 10 ng template DNA, 0.2 mM each primer, 0.2 mM deoxynucleotide triphosphate, and 0.625 units Taq DNA polymerase (New England Biolabs) in 13X reaction buffer [20 mM Tris-HCl, pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, and 0.1% Triton X-100]. Duplications were cycled at 95°C for 4 min, 25 cycles of 95°C 30 s, 53°C 45 s, 68°C 1 min, with a final extension of 68°C for 5 min. For deletions, the reactions were at 95°C for 4 min, 30 cycles of 95°C for 30 s, 55°C for 45 s, 68°C for 1 min, with a final extension of 68°C for 5 min. Reactions were run on a 1.2% agarose gel.

Unmapped RNA-seq reads from each clone were pooled to generate *de novo* transcript assemblies using Trinity (Grabherr et al., 2011). Contigs were aligned to the DMv4.04 assembly with GMAP (Wu and Watanabe, 2005) and excluded if they had greater than 85% coverage and sequence identity to the reference genome. Sequences below 500 bp were also excluded. Transcripts were then aligned with BLASTX to the Uniref100 database to remove contaminants and the remaining set aligned to NCBI nr protein database for functional annotation. To validate putative PAV transcripts, we mapped genomic DNA sequences from

DM to both the reference and PAV transcripts, filtered for high-quality alignments (MapQ ≥ 20), and removed PAVs with median read depth above half their theoretical coverage (30X).

FISH Analysis

Root tips for FISH analysis were obtained from greenhouse-grown plants. Chromosome preparation and FISH were performed following published protocols (Cheng et al., 2002). PCR-amplified DNA fragments (Dataset S2.3) were pooled and labeled with digoxigenin-11-dUTP (Roche Diagnostics) using a standard nick translation reaction. Chromosomes were counterstained with 4',6-diamidino-2-phenylindole in VectaShield antifade solution (Vector Laboratories). FISH images were processed using Meta Imaging Series 7.5 software, and the final contrast of the images was processed using Adobe Photoshop CC 2014 software.

Epigenetic Peak Calling

Chromatin immunoprecipitation sequencing reads were cleaned using Cutadapt (Martin, 2011) with minimum base quality 10 and minimum read length of 10 nucleotides. Reads were mapped to the DM v4.04 assembly in paired-end mode using Bowtie (v1.0.0) (Langmead, 2010). Peaks were called with HOMER (v4.3) (Heinz et al., 2010) using default parameters with minimum peak size of 150 bp and minimum peak distance of 300 bp.

Phylogenetic Analysis

Genetic distances were estimated from SNP and gene level CNV data using PHYLIP (http://evolution.genetics.washington.edu/phylip.html). For each type, 1000 bootstrap data sets were used to generate a consensus tree. Distances from the original data sets were used to

add branch lengths to consensus trees. Tree diagrams were generated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). CNV-based relationships were determined using copy status (duplicated, deleted, and non-CNV) as allele states for potato genes. SAUR and MKS1 trees were created using PHYLIP with multiple-protein alignments generated using ClustalW (Thompson et al., 2002). Alignments are available as Supplemental Data Sets 2.13 and 2.14.

Gene Lineage and Functional Analysis

Gene lineage was determined based on ortholog clustering of the predicted proteomes of nine species (http://phytozome.jgi.doe.gov; *Aquilegia coerulea* v1.1, *Arabidopsis thaliana* TAIR10, *Mimulus guttatus* v2.0, *Oryza sativa* v7.0, *Populus trichocarpa* v3.0, *Solanum lycopersicum* iTAG2.3, *Solanum tuberosum* v3.4, *Vitis vinifera* 12x; *Amborella trichopoda* v1.0; http://amborella.huck.psu.edu/data) using OrthoMCL (v1) (Li et al., 2003). TE-related genes were identified based on the existing DM functional annotations, PFAM domains (Bateman et al., 2004) associated with repetitive DNA, and alignment against the RepBase gene database (Jurka et al., 2005) (cutoff 1E-10), finding 2,886 TE genes in the DM gene set. Gene Ontology assignments were obtained from SpudDB (ftp://ftp.plantbiology.msu.edu/pub/data/SGR/GO_annotations/) and a Fisher's exact test was used to test enrichment in CNV duplicates and deletions.

Copy Number Variable Enriched Gene Clusters

To determine regions of the genome with high frequency of copy number variable genes, we split the reference assembly into overlapping 200 kb bins with a step size of 10 kb and counted the number of genes showing CNV in each bin. Bins containing significant numbers

of CNV genes were determined using a minimum threshold based on the mean of all genomic windows plus three standard deviations. Consecutive bins showing enrichment were combined into single regions and ranked by average number of CNV genes per bin.

Recombination Frequency

Recombination rates were estimated using SNPs from an F1 potato mapping population that used the DM reference genotype as a parent (Manrique-Carpintero et al., 2015). Marey maps were generated by plotting genetic positions of markers against their physical position (Chakravarti, 1991) and then a 0.1 cubic spline interpolation fitted curve was calculated. The slope of the line connecting adjacent markers was used as a local estimate of recombination rate (cM/Mb).

Data Access

Sequence data from this article can be found in the National Center for Biotechnology Information Sequence Read Archive under the BioProject accession number PRJNA287005. The updated assembly of the reference genome can be downloaded from SpudDB (http://potato.plantbiology.msu.edu/pgsc_download.shtml) or from the DRYAD repository (http://dx.doi.org/10.5061/dryad.vm142). The high-confidence SNP variant calls and the transcript-derived PAVs are available for download from the DRYAD repository under accession number http://dx.doi.org/10.5061/dryad.vm142.

Results and Discussion

Generation of a Monoploid Panel

Diploid potato landraces are the progenitors of modern tetraploids, being native to the Andes Mountains of South America and existing as heterozygous populations used in breeding new varieties (Ortiz, 2001; Spooner et al., 2007). A panel of 12 monoploid and doubled monoploid clones (referred to as "monoploids" for simplicity) (Table 2.1) were generated via anther culture using germplasm primarily composed of *S. tuberosum* Group Phureja landraces with limited introgression of Group Stenotomum, Group Tuberosum, and *Solanum chacoense* backgrounds. Clones were derived from three maternal landrace populations randomly pollinated by diploids from a photoperiod adapted research population (Figure S2.1) (Haynes, 1972). Four clones (M1, M9, M10, and M11) were direct products of landrace family crosses, while others (M2, M3, M6, M7, and M8) were subsequently generated in combination with heterogeneous breeding stocks harboring limited introgression from dihaploids of cultivated tetraploid potato (*S. tuberosum* Group Tuberosum) or wild *S. chacoense*. M13 alone was an interspecific hybrid, with introgressions from *S. chacoense*. Three clones (M2, M3, and M7) were derived from backcross (BC1) progeny of the doubled monoploid Group Phureja clone DM1-3 516 R44 (hereafter referred to as DM) used to generate the potato reference genome (Xu et al., 2011), offering reference points as closely related germplasm. These clones were selected for introduction of full homozygosity into a naturally heterozygous genome, without lethality and with limited floral or tuber developmental defects (Figure 2.1). Floral phenotype was affected in several clones; M2 and M10 displayed fused stamen and carpel whorls and M13 lacked stamens entirely. M3 and M5 showed premature abortion of flower buds, although occasionally wild-type flowers were produced. M6 alone did not flower, rarely

Table 2.1. ***Summary of genetic background composition, sequencing data, and variant calls associated with clones in the monoploid panel.***

	Genetic Background (%)				Variant Counts			
Clone	Phureja[a]	Tuberosum[b]	Wild[c]	Ploidy[d]	CNVs (Total)	Duplication	Deletion	SNP
DM	100	0	0	2x	0	0	0	0
M1	100	0	0	1x	8,837	2,577	6,260	3,433,063
M2	92	5	3	1x	4,996	1,565	3,431	1,557,476
M3	92	5	3	1x	2,978	897	2,081	800,333
M4[e]	>50	–	–	1x	8,424	2,572	5,852	3,242,070
M5[e]	>50	–	–	1x	9,194	2,887	6,307	3,664,157
M6	85	9	6	1x	8,627	2,864	5,763	3,632,667
M7	92	8	0	1x	4,062	1,222	2,840	1,186,135
M8	92	8	0	1x	8,716	2,617	6,099	3,625,031
M9	100	0	0	1x	8,496	2,703	5,793	3,989,158
M10	100	0	0	2x	8,640	2,645	5,995	3,718,500
M11	100	0	0	2x	8,962	2,639	6,323	3,648,940
M13[e]	~40–50	~0–10	50	1x	10,532	3,468	7,064	4,764,182

[a] Genetic input from diploid South American landrace populations of S. tuberosum Groups Phureja and Stenotomum.
[b] Genetic input from dihaploids of S. tuberosum Group Tuberosum (tetraploid cultivated potato).
[c] Genetic input from S. chacoense, a diploid wild species sexually compatible with cultivated potato species.
[d] Ploidy is reported from initial flow cytometry results; several clones spontaneously doubled in culture (M1, M5, M7, M8, and M9).
[e] Direct or indirect product of somatic fusions from diverse germplasm with primarily diploid landrace background.

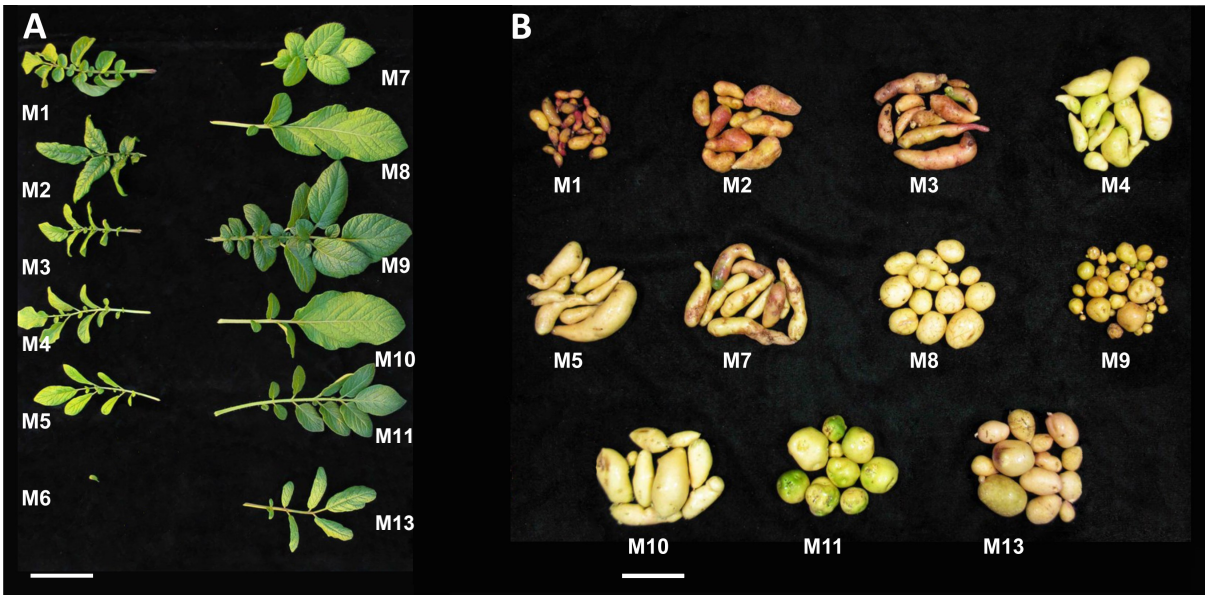

Figure 2.1. ***Phenotypic variation in a homozygous potato panel.*** *Leaf (A) and tuber (B) variation observed in the monoploid panel. M6 tubers are not available. Bars = 5 cm.*

produced a few small tubers (<0.5cm) with no plant yielding more than 0.5 g, and showed dramatic reduction in whole plant vigor, suggesting deleterious mutation of core genes. Hence, while several clones demonstrated morphological defects as a result of significant mutation load, all but M6 were able to mature and initiate tuber and floral development and therefore represent the minimal gene set required for development and reproduction of cultivated potato.

Sequencing and Variant Detection

Genome resequencing was conducted to provide coverage of 30-69X for comprehensive SNP and CNV analysis in the monoploid panel (Table S2.1). We aligned reads to an improved version of the DM potato reference genome (v4.04; see Methods) that includes 55.7 Mb of previously unassembled sequence. The DM v4.04 assembly was repeat-masked to limit analysis of structural variation to low-copy sequence. The number of SNPs relative to DM ranged from 800,333 in M3 to 4,764,182 in M13 (Table 2.1), reflective of the pedigree relationships between the clones and reference genotype (Figure S2.1). To confirm SNP calling accuracy, we compared variant calls from read alignments of 10 clones to variant calls generated using the Infinium 8303 potato array (Felcher et al., 2012), resulting in 98.5% concordance. Of the SNPs, 2.4 to 4.4% were located in coding regions and 70.1 to 75.7% were intergenic, with 0.67 to 0.84 ratios of synonymous to nonsynonymous changes in coding SNPs (Table S2.2). A SNP phylogeny measuring genetic distance between the monoploids closely supported their known pedigrees (Figure 2.2A).

Copy number variant detection was implemented in 100 bp genomic windows using

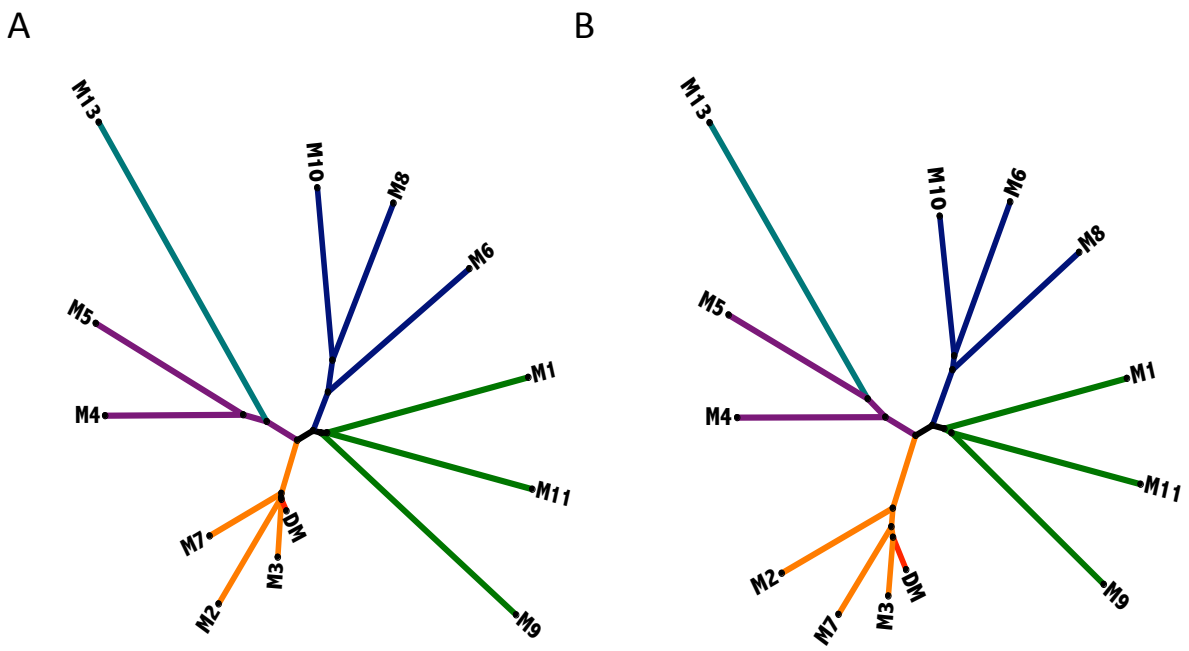

Figure 2.2. ***Phylogenetic trees of monoploid panel clones including the DM reference genotype.*** *Branch colors indicate genetic background of clones; DM reference genotype (red; DM), backcross progeny of DM (orange; M2, M3, and M7), direct progeny of non-reference landrace populations (green; M1, M9, and M11), landraces containing introgressions from non-landrace germplasm (blue; M6, M8, and M10), descended from intercrossed somatic hybrids (purple; M4 and M5), and wild/landrace interspecific hybrid (turquoise; M13). (A) Tree based on 12 million genome-wide SNP markers. (B) Tree based on copy number status of potato genes relative to the DM reference annotation.*

CNVnator (Abyzov et al., 2011). With read depth coverage of 30-69X per clone (Table S2.1), CNV detection, breakpoint precision, and copy number accuracy were well supported. For this analysis, CNVs were defined as duplications when exhibiting more copies relative to the reference genome or deletions if containing fewer copies than the reference. Several thousand CNVs were called in each monoploid ranging from 500 bp (minimum length) to 575 kb, with total CNV calls per individual varying from 2,978 to 10,532 (Table 2.1, Figure 2.3A; Table S2.3), indicating a wide range of structural variation among the clones and the reference genome. We compared CNVnator calls to those derived using a read depth method similar to other published plant CNV studies (Cao et al., 2011; Xu et al., 2012). For the 12 clones, we observed 95 and 84% support of total CNVnator deletion and duplication calls, respectively, by the read depth method (Table S2.4). CNVnator was significantly more conservative in calling CNVs; few calls were unique to CNVnator (range of 0.6 to 1 Mb for deletions and 1.3 to 2.4 Mb for duplications), whereas the read depth method generated substantially more unique variant calls (range of 79 to 151 Mb for deletions and 37 to 120 Mb for duplications). PCR validation supported 100 and 74% of the predicted copy number variants (46 target deletions and 42 target duplications) for primer pairs in which a single product of the predicted size was observed in both the reference genotype DM and at least one clone predicted to be single copy at that locus (Figure S2.2). The lack of full concordance between the computational predictions and the experimental validation results are due in part to technical limitations including sequence divergence in the primer binding sites between the clones as indicated by an inability to amplify the target locus in all variant and non-variant clones and insertions/deletions within the target amplification regions observed across the panel (Figure S2.2). Based on the concordance observed both with read depth estimations and

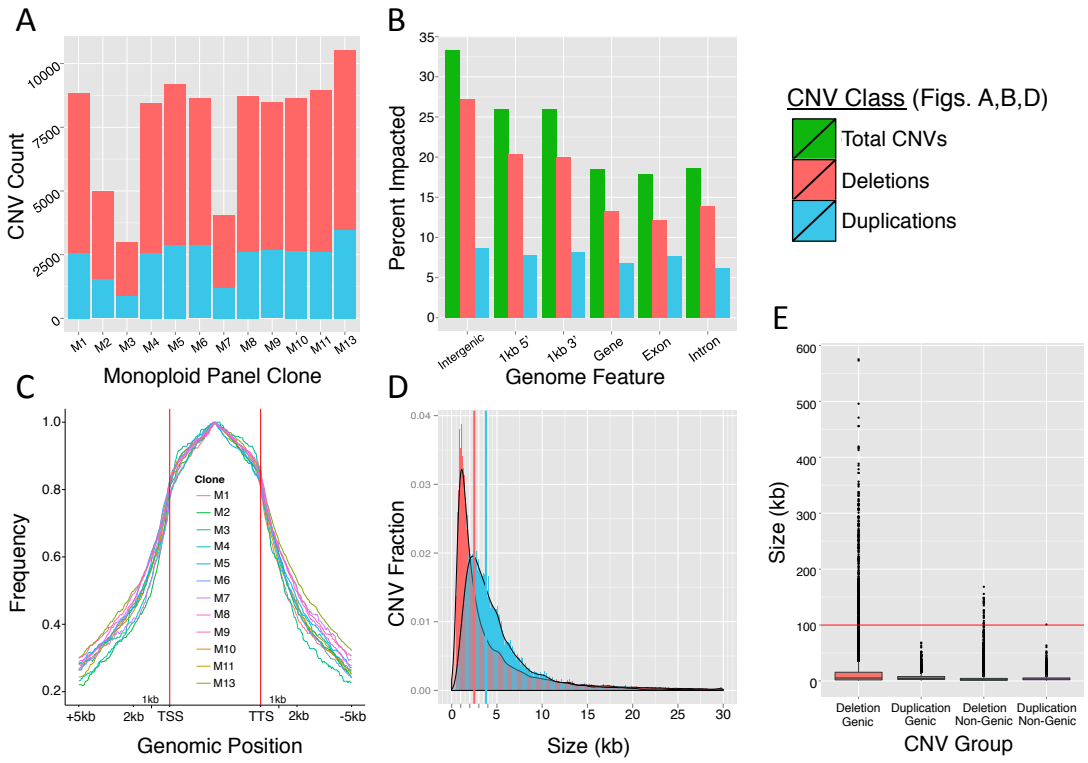

Figure 2.3. ***Summary statistics of monoploid panel copy number variation.*** *(A) Frequency of CNV per clone. The total number of filtered duplications (blue) and deletions (red) for each clone. (B) CNV representation within potato genome features. The percentage of sequence classes impacted by duplication and deletion in the monoploid panel. (C) Distribution of CNV frequency (per clone) relative to position of all duplicated genes (required minimum 50% gene model overlap with duplicated sequence). (D) CNV size distribution. Relative frequency of all CNV sizes up to 30 kb. Solid lines indicate median size for duplications and deletions. (E) Box plot of size of CNV for genic and non-genic duplications and deletions.*

experimental results, we feel that CNVnator provides a robust assessment of structural variation within our panel.

Like SNPs, CNV rates reflected the expected divergence of clones from the DM reference genotype. The greatest extent of CNV was observed in M13, a hybrid of landrace diploids and wild *S. chacoense*, and therefore was most likely to show different patterns of genome evolution. By contrast, backcross progeny of the DM reference genotype (M2, M3, and M7) exhibited lower CNV frequencies, although several thousand CNVs were found in each clone. To assess the ability of the CNV calls to reflect genetic relationships in the monoploid panel, we generated a second phylogeny based on gene level CNV (see below) using copy status (duplicated, deleted, and non-CNV) as allelic states for annotated reference genes. The resulting CNV tree closely reflected relationships estimated using SNPs (Figure 2.2B). This demonstrated the CNV calls were accurate at the gene level and that, like SNPs, they can effectively predict genetic relationships, supporting previous findings that CNVs are shared across accessions and reflect natural population structure (Cao et al., 2011).

Extent and Distribution of CNV in the Diploid Potato Genome

A total of 92,464 CNVs were identified in the panel (Dataset S2.1), collectively impacting 30.2% of non-gap sequence in the DM v4.04 reference genome. Many CNVs were conserved among the clones, sharing close breakpoints or corresponding to identical regions. Ratios of duplication and deletion were highly conserved, with duplications comprising 29.2 to 33.2% of total CNVs per clone. Similar bias in detection of deletions has been observed in previous comparative genomic hybridization and next generation sequencing-based studies (Żmieńko

et al., 2014). Structural variation was most common in intergenic sequence and on a genome scale was often more prevalent in pericentromeric regions with lower frequency observed in the gene-dense euchromatic arms, particularly in regions with high rates of recombination (Figure 2.4). This is consistent with a comprehensive examination of CNV in humans where CNV was enriched within pericentromeric regions (Lu et al., 2015; Zarrei et al., 2015). In maize, as shown using genotyping-by-sequencing, PAVs were enriched in the pericentromere (Lu et al., 2015) and negatively correlated with recombination rate, whereas a transcript-based PAV study (Hirsch et al., 2014) revealed PAVs were distributed throughout the maize genome with a lower frequency in pericentromeric regions. Thus, structural variation may differ for genic versus non-genic segments of a genome and our detection of CNV enrichment in the pericentromere reflects the use of whole genome resequencing data to assess structural variation.

The frequency of bases impacted by duplication was only slightly reduced (~1.8%) in genes compared with intergenic space (Figure 2.3B; Dataset S2.2). By comparison, rates of deletion were reduced in gene flanking sequence and 15% lower in coding sequence, suggesting a degree of selection against deleterious impacts on gene function (Figure 2.3B). While total gene sequence displayed similar rates of duplication and less deletion than whole-genome sequence, genes that were impacted by CNV (minimum 50% gene model overlap) showed signs of nonrandom targeting by CNV mechanisms. These genes displayed peak CNV frequencies within their gene bodies and a marked decrease of CNV frequency in the sequences bordering their 5’ and 3’ ends (Figure 2.3C; Figure S2.3). The reduced impact of CNV on overall coding sequence may result from selection against deleterious effects on

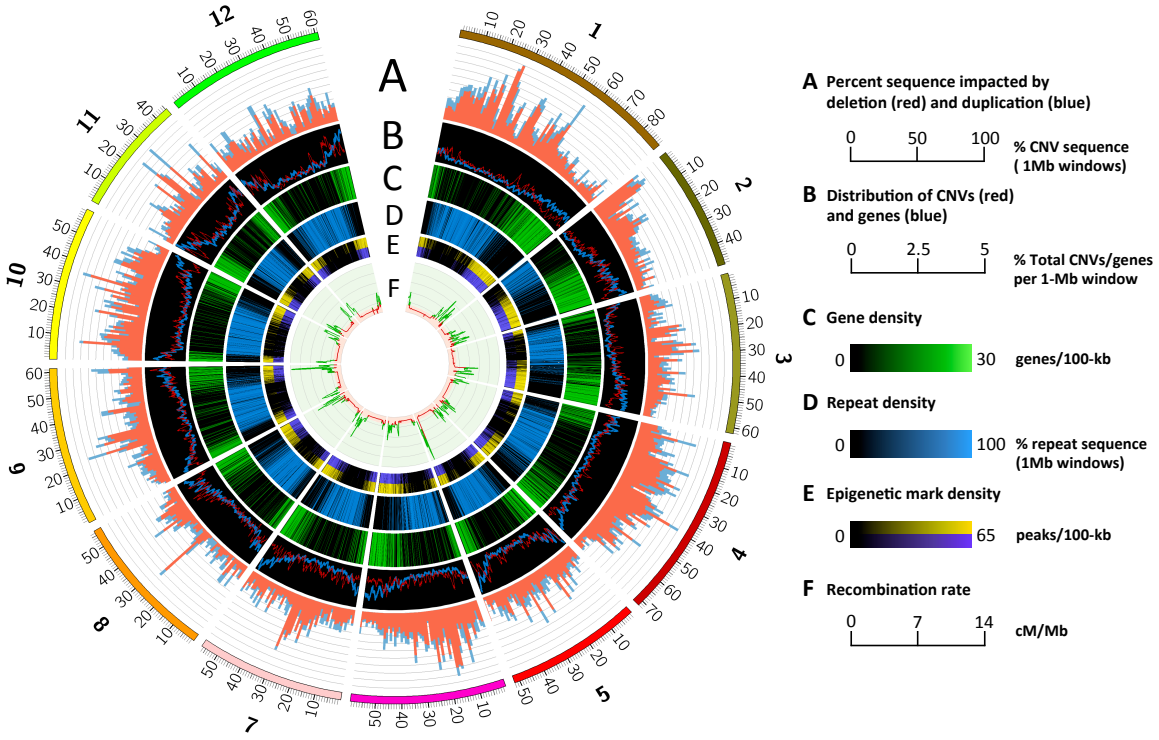

Figure 2.4. ***Chromosomal distribution of copy number variation, genes, repetitive sequence, and recombination rates in the diploid potato genome.*** *(A) Percentage of total non-gap sequence (0 to 100%) impacted by deletion (red) and duplication (blue) in 1 Mb non-overlapping windows. (B) Distribution of CNV counts (red) and gene counts (blue) (% total chromosome count in 1 Mb bins, 0.2 Mb step size). (C) Gene density (genes per 1 Mb window, 0.2 Mb step size). (D) Repeat density (% repetitive sequence in 1 Mb windows, 0.2 Mb step size). (E) Heat map of gene activating histone mark density (peaks per 1 Mb window, 0.2 Mb step size; yellow = H3K4me2 and purple = H4K5ac). (F) Recombination rate (0 to 14 cM/Mb) based on a biparental F1 mapping population (Manrique-Carpintero at al., 2015).*

expression of core gene functions, supported by a more substantial disparity in deletion compared with duplication rates with duplications being less likely to impair gene function.

Large Structural Variants are Common in Potato

Copy number variants were typically several kilobases or smaller, with a 3.0 kb median size in the panel (Figure 2.3D). Duplications (median 3.8 kb) tended to be larger than deletions (median 2.5 kb), although the fraction of CNVs represented by duplication diminished at larger size ranges (Figure S2.4). Size distribution was highly conserved among clones in the panel, suggesting similar patterns of formation and retention in the population (Figure S2.5).

Large-scale structural variation was also found to impact the diploid potato genome. A subset of variants was greater than 100 kb in length, the largest reaching 575 kb and present in clones M2 and M8, which lacked a known relationship. These CNVs (619 corresponding to 233 distinct regions) comprised 0.67% of total calls and were almost exclusively deletions (99.8%), which accounted for the majority of outlier CNV sizes (Figure 2.3E). Large CNVs may arise from different mechanisms than smaller, more common variants. Most CNVs are several kilobases or less, potentially resulting from non-allelic homologous recombination in regions containing segmental homology (Lu et al., 2012) or in regions without low-copy repeats as a result of microhomology and replication errors (Stankiewicz and Lupski, 2010; Arlt et al., 2012). Other CNVs may arise from retrotransposon activity, a common driver of structural variation in grass genomes (Morgante et al., 2007). However, a study of BAC-level (100 kb+) CNV in potato showed CNVs of this size are not segmental variants (Iovene et al., 2013), instead showing presence/absence across clones or between homologous chromosomes

within a clone. BAC-sized regions were commonly found to be missing on one to three homologous chromosomes of autotetraploids (Iovene et al., 2013). These variants likely correspond to the large CNVs identified in this study based on read depth, supporting the near exclusive detection of large CNVs as deletions in the monoploid panel. Large regions of the reference genome absent in the panel appear as deletions, while clone specific regions not present in the DM v4.04 assembly are undetectable by read depth, requiring independent assembly as PAVs.

To confirm the computational identification of these large CNVs, we performed fluorescence in situ hybridization (FISH) of three selected large CNVs (Seq26, Seq27, and Seq30), which span 105, 137.6, and 102.9 kb, respectively. Seq26 and Seq27 are at 28,282,100 to 28,387,100 bp and 30,733,700 to 30,871,300 bp on chromosome 7, respectively, and Seq30 is located on 22,656,700 to 22,759,600 bp on chromosome 9. Primers were designed to amplify four to five single copy DNA fragments for each CNV locus (Dataset S2.3), and DNA fragments amplified from the same CNV locus were pooled and labeled as a FISH probe. All three probes generated consistent FISH signals on a pair of DM chromosomes (Figure 2.5). The signals from the Seq26 and Seq27 probes were located close to the centromere of the target chromosome. In fact, most of the FISH signals overlapped with the primary constriction of the chromosome. Seq30 mapped to the middle of the long arm of its target chromosome. We then performed FISH using each probe on four monoploid/doubled monoploid clones selected based on computational prediction of presence/absence. The presence/absence of the FISH signals were concordant with the computational analysis (Figure 2.5) supporting our computational CNV calling method.

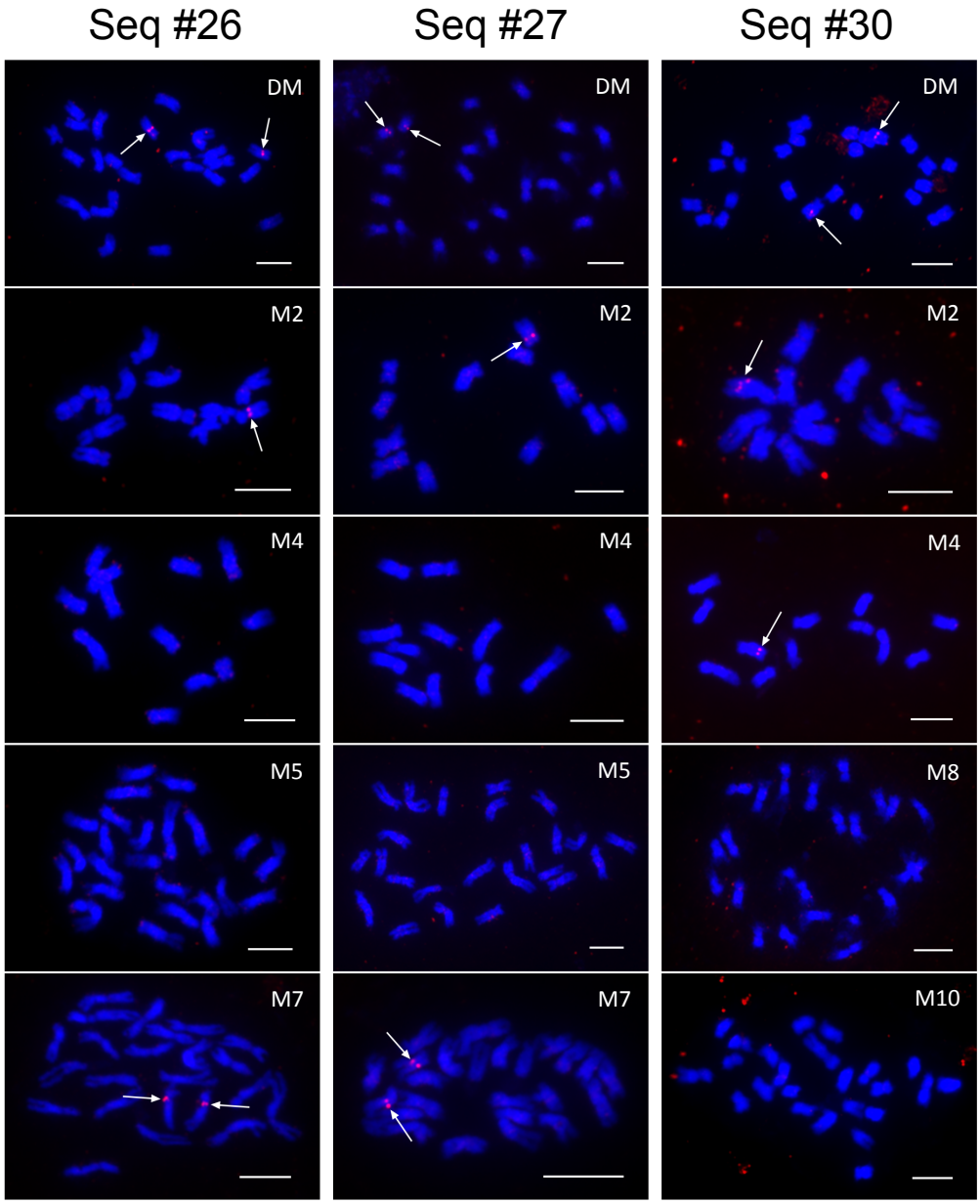

Figure 2.5. ***Fluorescent in situ hybridization (FISH) of the reference genotype DM and monoploid/doubled monoploid clones using probes targeting copy number variant regions.*** *Probes designed to multiple segments within three 100 kb+ computationally predicted CNV regions (Sequence 26 [~28.2 Mb Chromosome 7], Sequence 27 [~30.7 Mb Chromosome 7], Sequence 30 [~22.7 Mb Chromosome 9]) were labeled with digoxigenin-11-dUTP (red; arrows) and hybridized to chromosomes from the reference genotype (DM) and a subset of*

Figure 2.5 (cont'd)

the monoploid/doubled monoploid (M2, M4, M5, M7, M8, and M10). Chromosomes were prepared from root tip cells and were counterstained with 4',6-diamidino-2-phenylindole (blue). Perfect concordance between the computational prediction of CNV and the FISH signals was observed. Bars = 5 μm.

Large CNVs tended to be heterochromatic or located in the pericentromeres (Figure 2.6), underscoring the deleterious effects they can introduce to critical genes enriched in the euchromatic arms. Many corresponded to similar regions in different clones, with highly conserved breakpoints (Dataset S2.4). Chromosomes 5 and 7 contained numerous large CNVs shared by clones lacking a recent common ancestor, with a CNV on chromosome 5 reflecting deletion of a 100 kb sequence in all clones except M3 (BC1 progeny of DM) and breakpoints conserved to within 100 bp in most clones. Such conservation in germplasm from distinct progenitors suggests these variants descend from shared ancestral CNV events. Patterns of large-scale CNV also differed among chromosomes. Chromosomes 2 and 8 contained few large deletions, most being clone specific. More than half the large CNVs on chromosome 10 were specific to the hybrid M13, reflecting greater structural variation between cultivated potato and its wild relative *S. chacoense* on this chromosome. Notably, the only duplication larger than 100 kb was a 6x increase of repeats in the sub-telomeric region on the short arm of chromosome 12 in the hybridM13, indicating large-scale differences in genome structure between sexually compatible wild and landrace potato species.

Although large CNVs were uncommon in the euchromatic arms (Figure 2.6), the majority of these variants encompassed genes; 1,110 genes were deleted by large CNVs, while 875 (~81%) encoded proteins of unknown function or were associated with transposable elements (TEs). Few overlapped regulatory genes with the exception of F-box proteins, for which CNV is common in plants (Xu et al., 2009). Despite low rates of CNV impacting core gene functions, many potato genes were in fact subject to structural variation in the monoploid panel.

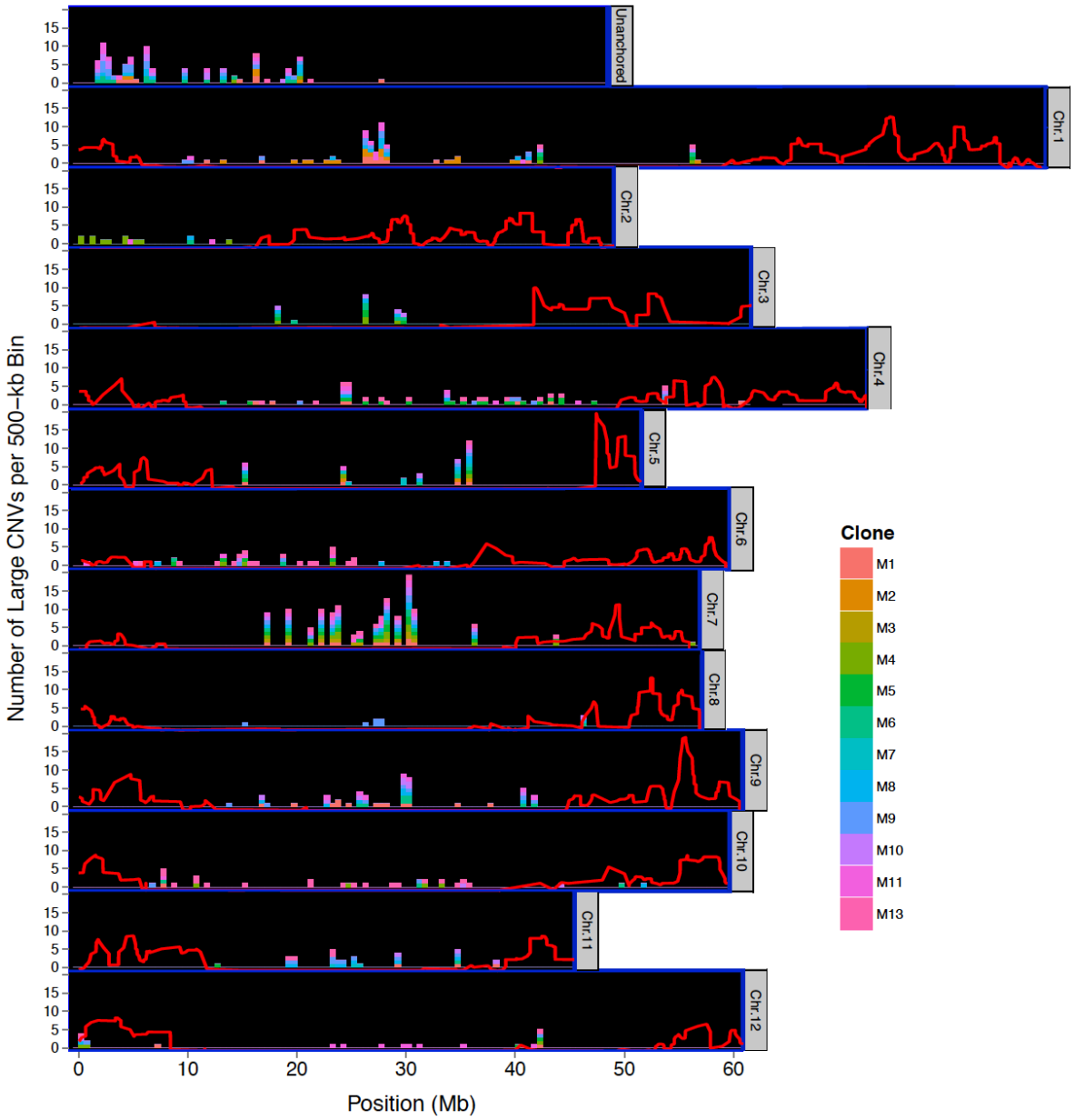

Figure 2.6. ***Positions of large (>100 kb) copy number variants in the potato reference genome assembly by counts per clone in non-overlapping 500 kb bins.*** *Variants are color coded for each clone. Red lines show chromosome-wide estimates of recombination frequency (cM/Mb) indicating the euchromatic arms (scale = 0 to 14 cM/Mb) (Manrique-Carpintero et al., 2015). "Unanchored" track represents all scaffolds that could not be anchored to the 12 main chromosomes.*

Role for CNV in Potato Adaptation

In total, 11,656 potato genes (29.7%) overlapped CNV calls, with 9,001 genes (22.9%) affected in at least half their annotated gene model (Dataset S2.5). To limit functional analysis to genes confidently affected by CNV, we used this second group to define the CNV gene set. Within the CNV gene set, ~11% consisted of TEs, ribosomal DNA, or nuclear organellar insertions, while 48% encoded proteins of unknown function, supporting association of CNV with genes that may be dispensable. Many CNV-impacted genes were also linked to pathogen resistance and abiotic stress tolerance. Gene Ontology (Ashburner et al., 2000) associations revealed several functions significantly enriched in the CNV gene set (Supplemental Data Sets 2.6 and 2.7), and many related directly (defense response, hypersensitive response, and response to UV-B) or indirectly (flavonol and trehalose biosynthesis and calcium transport) to stress tolerance, consistent with reports of CNV impacting stress-related pathways in other plant species. CNVs have been shown to influence phenotypes including modified reproductive habits and acquired tolerance to a range of harmful environmental factors, with gene duplication conferring herbicide resistance (Gaines et al., 2010), nematode resistance (Cook et al., 2012), as well as tolerance of frost (Knox et al., 2010), submergence (Xu et al., 2006), and aluminum and boron toxicity (Sutton et al., 2007; Maron et al., 2013).

To investigate if this relationship was supported in regions of the potato genome enriched in CNV activity, we counted copy number variable genes in 200 kb windows to identify regions containing high rates of gene level CNV (Dataset S2.8). Gene annotations in the 10 most highly enriched regions were examined in detail to determine functional relationship. Each

contained tandem clusters of genes with conserved functions related to stress response, supporting the role of CNV in potato adaptation.

SAURs

The region most enriched for CNV genes was located on chromosome 11 at 0.83 to 1.23 Mb, containing 19 auxin-induced SAURs (small auxin-up RNA) located in tandem arrays, with 17 of 19 duplicated in at least one clone. Additional CNV-enriched clusters were found on chromosomes 1, 4, and 12. SAURs comprise a large family of auxin-induced genes that exhibit species-specific expansion in both monocots and dicots (Jain et al., 2006). A study of this gene family in *Solanum* identified 99 SAURs in tomato (*Solanum lycopersicum*) and 134 in potato, showing greater expansion in *Solanum* species relative to Arabidopsis, rice, and sorghum (*Sorghum bicolor*; Wu et al., 2012). Phylogenetic analysis revealed expansion of multiple *Solanaceae*-specific subgroups, with upstream regulatory sequences containing cis-elements related to auxin signaling, light signaling, drought stress, salt stress, heat shock, and calcium response, while most tomato SAURs were induced by auxin and regulated by abiotic stress (Wu et al., 2012). Diploid potato contains more SAURs than several well-annotated monocot and dicot species, including its close relative tomato. To determine if recent duplications within diploid populations contributed to the *Solanum*-specific expansion of SAURs seen in potato, we generated a phylogenetic tree using protein sequences of SAURs identified by Wu et al. (2012) in rice, Arabidopsis, tomato, and potato (Figure S2.6). Potato SAURs displaying CNV were enriched in two large clades reflecting the most significant *Solanum*-specific expansions of this gene family, offering evidence for the impact of duplication on gene family diversification in these species. Our results suggest that SAURs

continue to undergo duplication within closely related populations of diploid cultivated potato, highlighting the role of CNV in the rapid evolution of a gene family involved in abiotic stress response. The large number of potato genes compared with tomato in these clades, along with high rates of CNV within related Group Phureja clones, support ongoing SAUR gene expansion in potato.

Disease Resistance

The second highest density of CNV genes was found on chromosome 11 at 42.59 to 43.05 Mb, containing a cluster of 16 genes encoding nucleotide binding site leucine-rich repeat (NBS-LRR) disease resistance proteins, of which, 14 showed variation in copy number. This is consistent with previous studies conducting genetic mapping of potato resistance quantitative trait loci, showing they are often clustered in the genome (Gebhardt and Valkonen, 2001). Resistance genes are typically found in clusters or hot spots in the genomes of many plant species and are known to be fast evolving as a result of local gene duplications (Bergelson et al., 2001). Three genes conferring race-specific resistance to *Phytophthora infestans* (R3, R6, and R7) and a root cyst nematode resistance gene (Gro1.3) were previously mapped to this locus (Gebhardt and Valkonen, 2001). Notably, three other regions among the 10 most highly enriched for CNV genes were also disease resistance clusters, highlighting the rapid evolution of gene families required for response to changing disease pressure. These were located on chromosomes 4, 7, and 9, with the cluster on chromosome 4 corresponding to the R2 locus for late blight resistance (Gebhardt and Valkonen, 2001).

Secondary Metabolites

A third locus at ~85 Mb on chromosome 1 contained 21 Methylketone Synthase 1 (*MKS1*) genes, 18 showing CNV in the panel. Methylketones are secondary metabolites produced in the glandular trichomes of solanaceous species such as tomato and potato and, in particular, their wild relatives (Bonierbale et al., 1994; Antonious, 2001). In response to insects, these compounds are secreted onto the leaf surface, conferring resistance to a variety of pests. *MKS1* expression has been directly correlated with methylketone levels and leaf gland density (Fridman et al., 2005), confirming their role in defense against herbivory. Studies of its function suggest *MKS1* emerged recently in its gene family and may be *Solanum* specific (Yu et al., 2010). Similar to patterns observed in microbial resistance genes, plant genes offering defense against insect attack may be fast evolving in order to generate new sources of genetic resistance. Their tandem clustering reflects grouping of other insect defense pathway genes in the *Solanaceae*, including steroidal glycoalkaloid biosynthesis (Itkin et al., 2013). Phylogenetic clustering of genes with sequence homology to the five tomato *MKS1* genes showed they fall within a *Solanum*-specific clade containing only potato and tomato orthologs (Figure S2.7). Other plants, including the asterid *Mimulus guttatus*, lacked close orthologs, confirming the likelihood that *MKS1* function emerged recently in the genus *Solanum*. The *Solanum*-specific clade containing *MKS1* also showed greater diversification in the diploid potato genome than tomato, with over twice as many potato homologs. Almost all potato *MKS1* genes showed CNV in the monoploid panel, supporting a role of duplication in species specific expansion of gene families involved in plant stress pathways.

Chromosome 9 contained 10 copies of the gene encoding desacetoxyvindoline 4'-hydroxylase (*D4H*), the indole alkaloid biosynthetic pathway enzyme used in synthesis of vindoline. Indole alkaloids have been associated with response to fungal elicitors, insect herbivory, and UV light exposure (St-Pierre et al., 2013), and vindoline acts as a primary substrate to form the cytotoxic chemotherapeutic vinblastine in *Catharanthus roseus* (Vazquez-Flota and De Luca, 1998). While this enzymatic function is not likely conserved in potato, its diversification may result in production of other defensive compounds. Another CNV-enriched locus on chromosome 5 contained a cluster of eight flavonol 4'-sulfotransferases. Flavonols, one of the most abundant classes of flavonoids in plants, have antioxidant properties and play a major role in plant response to abiotic stress, particularly UV light damage (Gill and Tuteja, 2010), and sulfate conjugation of secondary metabolites can affect their function within plant systems (Varin et al., 1997; Klein and Papenbrock, 2004). The remaining clusters contained duplicated genes encoding mannan endo-1,4-b-mannosidase and GH3 indole-3-acetic acid-amido synthetase, respectively, each with roles in cell wall modification already implicated in pathogen response (Ding et al., 2008; Westfall et al., 2010).

Association of CNV with disease resistance genes is well established in plants (Ellis et al., 2000). The extensive CNV observed in SAURs, *MKS1*, and other gene families in closely related germplasm suggests these are also rapidly evolving, supported by their lineage-specific expansions (Supplemental Figures 2.6 and 2.7). Whole-genome duplication is proposed to be a mechanism supporting adaptive evolution and speciation (De Bodt et al., 2005). It appears local gene duplication introduces similar potential for diversification and

sub-functionalization in potato. Our finding that the most highly enriched CNV clusters harbor genes implicated in biotic and abiotic stress response furthers the hypothesis that evolution through local gene duplication can be adaptive, allowing plants to develop genetic resistance to changing environmental pressure from pests, disease, and abiotic stress such as drought.

Gene Expression as a Predictor of CNV

Gene-level CNV revealed an association with stress-related functions, as well as TEs and proteins of unknown function, some of which may not be essential for development. We investigated whether gene expression patterns support this connection, using an atlas of RNA-seq libraries representing a tissue series, as well as abiotic and biotic stress treatments for the DM reference genotype (Xu et al., 2011), to categorize the potato gene set into expression classes (Table S2.5). The frequency of genes in each expression class was compared in the duplicated and deleted versus non-CNV gene sets on a per clone basis to determine how gene expression relates to likelihood of CNV. Classes included confidently expressed genes (fragments per kilobase per million mapped reads (FPKM) ≥10 for multiple tissue types), lowly expressed genes (FPKM < 1 in all tissues), and genes showing response to hormone or stress treatments (5-fold FPKM induction). Abiotic stress treatments included salt, mannitol, drought, abscisic acid (ABA), and heat, while biotic stress treatments included *P. infestans*, benzothiadiazole (salicylic acid analog), and b-aminobutyric acid (jasmonic acid analog). Hormone treatments included auxin, cytokinin, ABA, and gibberellic acid.

Genes with expression induced by at least one form of abiotic stress or hormone treatment were significantly enriched among duplications ($P \leq 0.05$; Figure 2.7), supporting the relationship of duplication with genes involved in environmental response and adaptation. Individual abiotic stress treatments were unequally represented; salt-induced genes were most prevalent in the duplicated gene set, followed by drought-induced genes (Figure S2.8). Mannitol, heat, and ABA responsive genes were more common among duplicated genes, but less significantly ($P \leq 0.05$). For hormone-responsive genes, those induced by cytokinin were more significantly duplicated than any other stress or hormone induced class. Biotic stress response classes (induced by *P. infestans*, benzothiadiazole, and b-aminobutyric acid) were not significantly enriched or underrepresented in either CNV group (Figure 2.7). While plant defense genes are known to be fast-evolving (Ellis et al., 2000), classic NBS-LRR disease resistance genes are lowly expressed and not typically induced by pathogen or elicitor treatment. Genes induced by wounding that mimic herbivory were significantly underrepresented among deletions in most clones (Figure 2.7), suggesting selection against loss of genes required for response to physical stress.

Expression analysis further supported the association of CNV with dispensable genes and selection against impacting core functions. Genes with low expression in all tissues were highly enriched in the deleted gene set and to a lesser extent in duplicated genes (Figure 2.7), suggesting low selection against mutation. The mean representation of lowly expressed genes in the deleted set was 56.4% per clone, higher in non-CNV genes (29.1%), or the frequency of weakly expressed genes in the DM reference genome (30.4%). Genes with high expression levels in any major tissue category (aboveground vegetative, reproductive, root, and tuber)

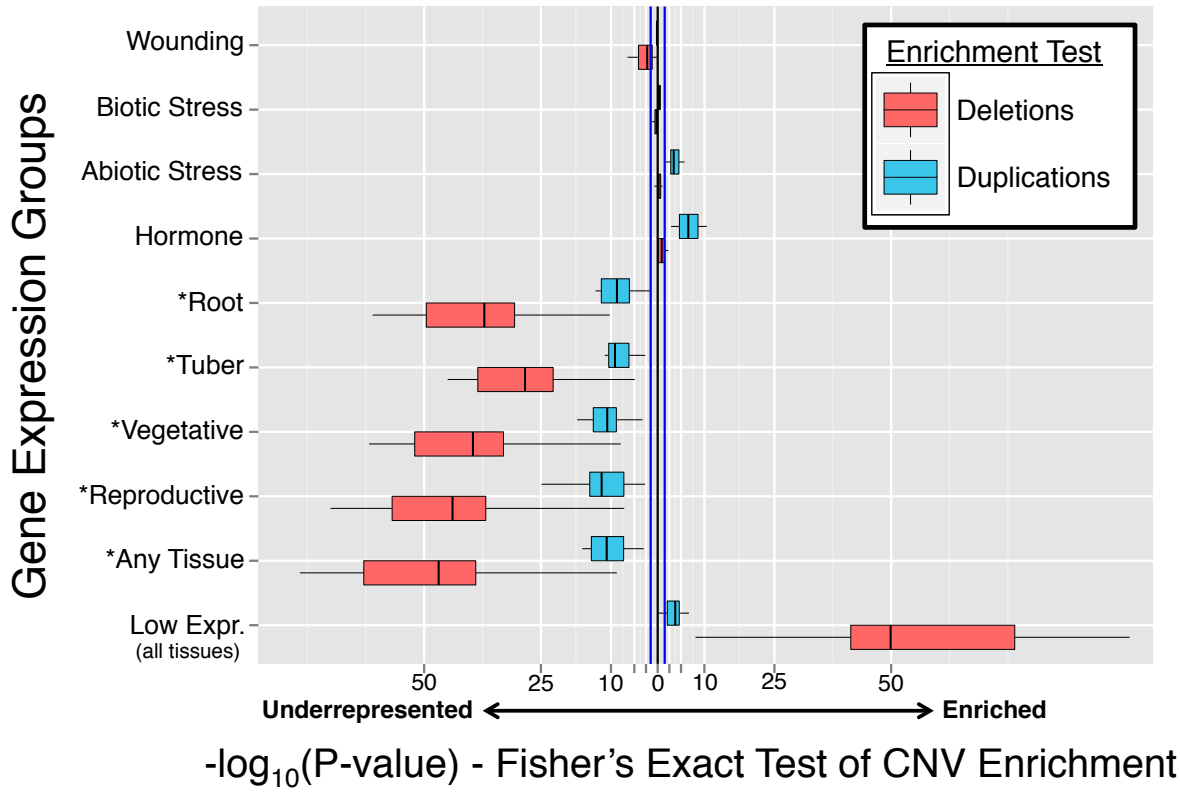

Figure 2.7. ***Representation of genes from various expression groups in the duplicated and deleted gene sets relative to genes not impacted by copy number variation.*** *Scores are based on negative log-10 transformation of P-values from a Fisher's exact test of count data, with enrichment indicating increased representation in the copy number variant gene sets and underrepresentation indicating lower prevalence in the CNV gene sets. Blue lines indicate significance threshold (P = 0.05). An asterisk denotes confidently expressed genes as defined as having a FPKM value > 10.*

were strongly underrepresented among duplications and deletions, reflecting the greater likelihood of highly expressed genes serving core functions (Figure 2.7). These genes were less likely to experience deletion than duplication, reinforcing its greater potential for deleterious effect. For each major tissue type (leaves, flowers, roots, tubers, and whole in vitro plant) CNV rates became lower at increasing FPKM levels, with strong correlation across tissues (Figure S2.9). Consistent with expression data, we observed that two histone marks associated with permissive transcription (H3K4me2 and H4K5,8,12,16ac) in DM leaves and tubers were preferentially associated with genes not impacted by CNV (Holoch and Moazed, 2015), while CNV frequency was increased in genes lacking one or both activating marks (Table 2.2).

Core and Dispensable Gene Set

Genome resequencing studies have revealed plant and animal species contain core sets of genes required for growth and development, as well as dispensable genes that are missing in individuals (Li et al., 2010; Hirsch et al., 2014), leading to the concept of the pan-genome. Dispensable genes have been speculated to be involved in heterosis in outcrossing species (Lai et al., 2010; Ding et al., 2012) and stress adaptation (DeBolt, 2010; Żmieńko et al., 2014) and are thought to contribute to species diversification and development of novel gene functions (Wang et al., 2006). Thousands of deleted genes were identified in the monoploid panel. Despite an abundance of missing genes, each homozygous clone (except M6) was able to flower and tuberize (Figure 2.1), suggesting they possessed the core gene set required for development and reproduction. Dispensable genes were defined as those affected by deletion in at least one flowering and tuberizing clone, with the CNV spanning at least half an exon

Table 2.2. ***Extent of copy number variation in genes associated with transcription activating epigenetic marks.***

		Percentage of Genes Impacted by CNV[a]			
DM Histone Mark[b]	Total Genes	Non-CNV	Total CNV	Duplicated	Deleted
H3K4me2 – leaf	24,637	86.1	13.9	6.4	8.7
H3K4me2 – tuber	6,206	75.4	24.6	10.8	16.3
H4K5ac – leaf	11,974	90.9	9.1	4.4	5.6
H4K5ac – tuber	22,344	87.6	12.4	5.8	7.8
No Leaf Mark	14,316	61.4	38.6	11.8	30.3
No Tuber Mark	14,531	62.1	37.9	11.6	29.6
No Activating Mark	11,975	59.0	41.0	11.8	32.7

[a] Values indicate the percent of genes in the DM reference affected by CNV as observed in the monoploid panel.
[b] Genes were required to share 50% gene model overlap with a histone mark for association.

within the gene. Of 8,888 (22.6%) genes overlapping deletions among these clones, 7,183 were classified as dispensable. An additional 1,429 non-deleted genes were predicted to contain SNPs encoding premature stop codons, indicating at least 8,612 (21.9%) genes in DM may be dispensable. We defined the core potato gene set of 30,401 genes (77.4%), as all annotated DM genes not impacted by deletion or premature stop in the study panel. As each monoploid/doubled monoploid clone had to survive the monoploid sieve (Wenzel et al., 1979) to be included in this study, we have most likely underestimated the number of haplotypes containing deleterious/dysfunctional alleles and deletions present in the progenitor diploid clones. Improvements in the cost and ease of whole-genome sequencing and assembly of heterozygous diploid and tetraploid genomes will permit refinement of the composition of the core genome of potato in the future.

M6 displayed heavily restricted vegetative growth and rare tuberization and was unable to flower, indicating clone-specific mutation(s) in the core potato gene set. We examined CNV and SNP alleles unique in M6 to identify putative genes essential for development and flowering in potato (Dataset S2.9). One candidate gene was a partial deletion of the putative homolog (78% amino acid sequence identity) of Arabidopsis RADICAL INDUCED CELL DEATH1 (PGSC0003DMG400014419), which encodes a protein that interacts with over 20 transcription factors and is required for development (Jaspers et al., 2009). In Arabidopsis, *rcd1* mutants had extremely stunted phenotypes with deformed leaves, developmental defects, and inhibited flowering (Jaspers et al., 2009), similar to the M6 phenotype. M6 harbored additional clone-specific deletion of genes encoding an inhibitor of growth protein

(PGSC0003DMG400011588) and a kinetochore protein involved in cell division (PGSC0003DMG400010002).

PAV represents a form of CNV in which genes lack copies in the reference but are present in non-reference individuals. To estimate the contribution of transcript-level PAV to the dispensable gene set, unmapped RNA sequences from the monoploids were pooled and assembled into putative PAV transcripts, yielding 1,169 sequences with 1,263 isoforms. DM genomic sequence reads were aligned to the genome and PAV transcripts to identify potential unassembled reference sequences missing from the DM v4.04 assembly. In total 1,256 putative PAVs lacking high quality read coverage from DM were classified as true PAVs (Dataset S2.10). Only 224 PAVs could be assigned a protein function. As with genes affected by CNV, many were related to TEs, resistance proteins, and proteins of unknown function (Dataset S2.11). This is likely a significant underrepresentation of gene level PAV in potato, as it was based on transcripts derived from only two tissues and will fail to capture PAV transcripts expressed in other tissues or transcripts that are weakly expressed.

Evolution of Dispensable Genes

We evaluated CNV in genes arising at different levels of the potato lineage to study the origin of its dispensable genome. Orthologous gene clusters were generated for nine angiosperm species, including closely related tomato (*S. lycopersicum*), non-*Solanaceae* asterid *M. guttatus*, core eudicot *Aquilegia coerulea*, monocot rice, and the basal angiosperm *Amborella trichopoda*. Based on ortholog clustering, genes were classified as lineage specific in potato (3,584), *Solanum* (11,604), asterids (12,205), and eudicots (14,892) or conserved in flowering

plants (10,392) (Figure S2.10). Relatively few genes (601) in potato seem to have appeared in asterids prior to separation of the genus *Solanum* from its other species, after which many (11,604) appeared in the *Solanum* lineage. Most of these genes (8,020) arose before speciation of potato, whereas 3584 are potato specific. This suggests major gene diversification occurred after *Solanum* separated from other asterids, with further expansion at the species level in potato, possibly due to an increase in rapidly evolving genes with high rates of sequence divergence and/or a high birth/death rate in *Solanum*-lineage specific genes. This may explain their lack of similarity with genes of known function. CNV frequency, particularly deletion, was progressively higher in more recent lineages (Figure 2.8), supporting the association of dispensable genomes with recently evolved genes observed in species such as maize (Morgante et al., 2007). Genes arising in the Solanum lineage were more likely to be dispensable and 32% of potato species-specific genes were missing in at least one monoploid, whereas genes with conserved orthologs in angiosperms had extremely low rates of CNV. It is important to note the genomes used in our evolutionary analyses were annotated separately, such that genes associated with CNV may not be equally represented within the annotated proteome of each genome. However, this bias is unlikely to be large enough to explain the observed differences in variation, particularly in light of the relatively few clones needed to observe such genome variation in potato. Overall, these results support a relationship of CNV with gene diversification at the species level and highlight the potentially disruptive force of deletion, and to a lesser extent duplication, on genes serving core functions in flowering plants.

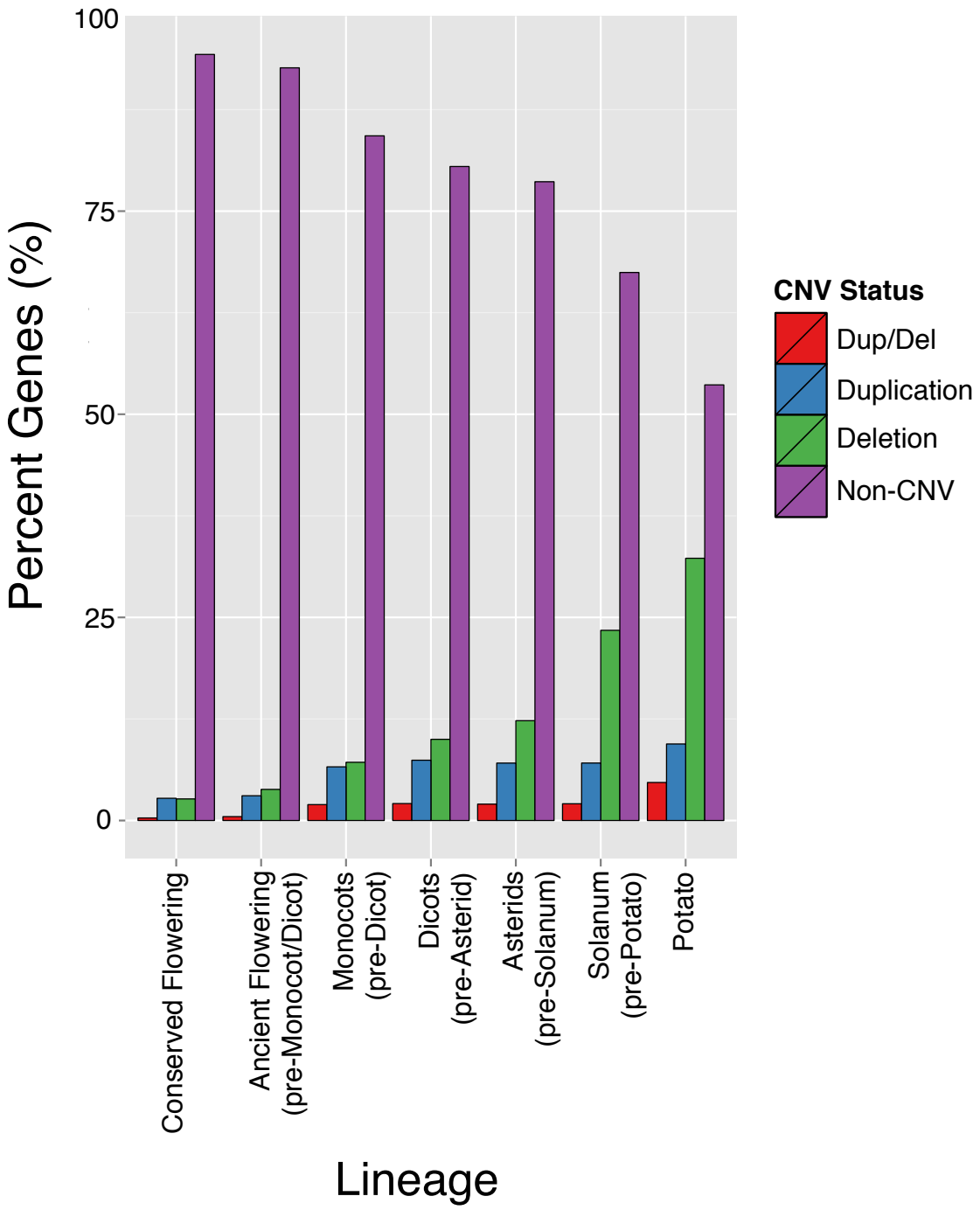

Figure 2.8. ***Copy number variation frequency among potato genes arising at different levels of the green plant lineage.*** *"Potato" contains S. tuberosum Group Phureja species-specific genes. "Solanum" contains Solanum-specific potato genes predating potato speciation. "Asterid" contains Asterid-specific potato genes predating Solanum. "Dicots" contains*

Figure 2.8 (cont'd)

eudicot-specific potato genes predating asterids. "Monocots" contains potato genes found in monocots and eudicots predating the differentiation of eudicots. "Ancient Flowering" includes all potato genes that arose before monocots. "Core Flowering" includes potato genes with orthologs in all flowering plant species.

Conclusions

The extent of CNV in the monoploid panel supports diploid potato possessing a greater degree of structural variation than reported in several sexually reproducing species. Overall, CNV impacted 30% of the genome and 11,656 genes, underscoring the heterogeneous nature of haplotypes within diploid potato compared with most sexually reproducing diploids. In contrast, a study on the core and dispensable gene set of soybean (*G. max*) explored the genomes of seven wild *Glycine soja* ecotypes (Li et al., 2014) with read-depth analysis, identifying only 1,978 of 54,175 soybean genes (3.7%) impacted by CNV, significantly fewer than in our study. Other primarily inbreeding species, including Arabidopsis, cucumber (*Cucumis sativus*), and rice, also show limited structural variation relative to potato. Cao et al. (2011) resequenced 80 Arabidopsis lines from eight geographically distinct populations across Europe and Central Asia. Using a read-depth approach, 1,059 CNVs (minimum length 1 kb) were identified across all lines, impacting ~500 protein coding genes (<2%) and 2.2 Mb (~1.6%) of the assembled genome. In a recent study including a panel of 115 cucumber accessions, fewer structural variants were discovered than in Arabidopsis (Zhang et al., 2015). A similar analysis of 50 rice accessions, including 10 wild species, detected 1,327 gene loss events (2.4%) and 865 gene-associated duplications (Xu et al., 2012).

This study shows CNV is a major component of the significant genomic diversity of clonally propagated potato. Like potato, maize is another outcrossing heterozygote containing significant diversity at a structural level (Żmieńko et al., 2014),with breeders relying on heterosis as an essential component of plant vigor. Extensive CNV and PAV between maize inbreds have been speculated as components of heterosis, in which the CNV and PAVs permit

complementation of missing genes and greater phenotypic diversity (Lai et al., 2010; Hansey et al., 2012). Maize contains a large pan-genome contributing to its diversity, and it is estimated that the B73 maize reference contains 74% of the low copy gene fraction present in all inbreds (Lu et al., 2015). Chia et al. (2012) resequenced 103 maize lines, including a mixture of wild, pre-domesticated, and elite germplasm and concluded 32% of genes in the B73 reference were affected by CNV. In this study of 12 related clones derived from only a few native populations, ~30% of potato genes overlapped CNVs, with ~23% affected in over half their gene model, suggesting clonally propagated potato tolerates greater rates of mutation than many sexually reproducing species. Passage through the monoploid sieve (via anther culture) freed the panel of lethal alleles and structural variants present in their heterozygous diploid progenitors, with the clones representing rare combinations of nonlethal alleles. In comparison to maize inbreds selected for vigor and fertility, we applied much less pressure as our only selective criteria were surviving the monoploid sieve and capacity for growth in vitro. As a consequence, the spectrum of dispensable genes identified in this study may not be directly comparable with dispensable genes identified in species such as maize. However, the abundance of variants able to be retained and identified in this study implies that CNVs and other somatic mutations may be less likely to be removed from the genomes of cultivated clones.

It was observed that CNV is more likely to impact species specific gene groups and dispensable genes, suggesting recent genome expansions in species will influence their degree of structural variation. Plants with whole-genome duplications, or genomes enlarged by TE activity such as maize (Fu and Dooner, 2002; Brunner et al., 2005), have greater potential for

genes to be impacted by CNV, whether by reduced selection on duplicated coding sequences (Tang et al., 2008; Mun et al., 2009; Schnable et al., 2009) or targeting by mobile elements (Kidwell and Lisch, 1997; Slotkin and Martienssen, 2007). Low rates of sexual reproduction may also contribute to distinct patterns of structural variation, with fewer non-allelic homologous recombination events occurring during meiosis and a higher rate of non-recurrent mitotic CNVs formed during DNA replication. This may explain the negative relationship between structural variation and recombination frequency observed on the arms of several potato chromosomes, a feature separating it from the distribution of CNV in maize (Springer et al., 2009). Gene density is also greater in the arms of potato chromosomes, such that selection against deleterious mutation in these regions could result in lower retention. Comparing structural variation within wild potato populations with higher rates of sexual reproduction and asexually propagated clones may help to elucidate the long term impacts of asexual reproduction on plant genome variation. This study supports earlier observations of large-scale CNV in potato (Iovene et al., 2013). We can now speculate that the structural variation observed in tetraploid potato is not due to polyploidy alone because substantial genome heterogeneity is also present in diploid potato. Overall, this study adds a new dimension to our understanding of intraspecies genome variation. In contrast to sexually reproducing species such Arabidopsis and maize, where meiotic events routinely purge recessive deleterious alleles in successive generations and in which inbreeding and outcrossing may affect CNV frequency, diploid and tetraploid potato retain a heavy genetic load that remains masked due to asexual reproduction and heterozygosity.

LITERATURE CITED

LITERATURE CITED

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. 491, 56-65.

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

Antonious, G.F. (2001). Production and quantification of methyl ketones in wild tomato accessions. J. Environ. Sci. Heal. B 36, 835-848.

Arlt, M.F., Wilson, T.E., and Glover, T.W. (2012). Replication stress and mechanisms of CNV formation. Curr. Opin. Genet. Dev. 22, 204-210.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., and Sonnhammer, E.L. (2004). The Pfam protein families database. Nucleic Acids Res. 32, D138-D141.

Bergelson, J., Kreitman, M., Stahl, E.A., and Tian, D. (2001). Evolutionary dynamics of plant R-genes. Science 292, 2281-2285.

Bonierbale, M.W., Plaisted, R.L., Pineda, O., and Tanksley, S. (1994). QTL analysis of trichome-mediated insect resistance in potato. Theor. Appl. Genet. 87, 973-987.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., and Rafalski, A. (2005). Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17, 343-360.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188-196.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., and Lippert, C. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43, 956-963.

Chakravarti, A. (1991). A graphical representation of genetic and physical maps: the Marey map. Genomics 11, 219-222.

Cheng, Z., Buell, C.R., Wing, R.A., and Jiang, J. (2002). Resolution of fluorescence in-situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. Chromosome Res. 10, 379-387.

Chia, J.-M., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803-807.

Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J., Hughes, T.J., Willis, D.K., and Clemente, T.E. (2012). Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science 338, 1206-1209.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., and Sherry, S.T. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. Trends. Ecol. Evol. 20, 591-597.

De Jong, H., and Rowe, P.R. (1971). Inbreeding in cultivated diploid potatoes. Potato Res. 14, 74-83.

DeBolt, S. (2010). Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. Genome Biol. Evol. 2, 441-453.

Díaz, A., Zikhali, M., Turner, A.S., Isaac, P., and Laurie, D.A. (2012). Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). Plos One 7, e33234.

Ding, X., Cao, Y., Huang, L., Zhao, J., Xu, C., Li, X., and Wang, S. (2008). Activation of the indole-3-acetic acid–amido synthetase GH3-8 suppresses expansin expression and promotes salicylate- and jasmonate-independent basal immunity in rice. Plant Cell 20, 228-240.

Ellis, J., Dodds, P., and Pryor, T. (2000). Structure, function and evolution of plant disease resistance genes. Curr. Opin. Plant. Biol. 3, 278-284.

Felcher, K.J., Coombs, J.J., Massa, A.N., Hansey, C.N., Hamilton, J.P., Veilleux, R.E., Buell, C.R., and Douches, D.S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. Plos One 7, e36347.

Fridman, E., Wang, J., Iijima, Y., Froehlich, J.E., Gang, D.R., Ohlrogge, J., and Pichersky, E. (2005). Metabolic, genomic, and biochemical analyses of glandular trichomes from the wild tomato species *Lycopersicon hirsutum* identify a key enzyme in the biosynthesis of methylketones. Plant Cell 17, 1252-1267.

Fu, H., and Dooner, H.K. (2002). Intraspecific violation of genetic colinearity and its implications in maize. Proc. Natl. Acad. Sci. 99, 9573-9578.

Gaines, T.A., Zhang, W., Wang, D., Bukun, B., Chisholm, S.T., Shaner, D.L., Nissen, S.J., Patzoldt, W.L., Tranel, P.J., and Culpepper, A.S. (2010). Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc. Natl. Acad. Sci. 107, 1029-1034.

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

Gebhardt, C., and Valkonen, J.P. (2001). Organization of genes controlling disease resistance in the potato genome. Annu. Rev. Phytopathol. 39, 79-102.

Gill, S.S., and Tuteja, N. (2010). Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. Plant Physiol. Bioch. 48, 909-930.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., and Zeng, Q. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644-652.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., De Leon, N., Kaeppler, S.M., and Buell, C.R. (2012). Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. Plos One 7, e33071.

Hardigan, M.A., Bamberg, J., Buell, C.R., and Douches, D.S. (2015). Taxonomy and genetic differentiation among wild and cultivated germplasm of *Solanum* sect. *Petota*. Plant Genome. 8.1.

Hattori, Y., Nagai, K., Furukawa, S., Song, X.-J., Kawano, R., Sakakibara, H., Wu, J., Matsumoto, T., Yoshimura, A., and Kitano, H. (2009). The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. Nature 460, 1026-1030.

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

Haynes, F.L. (1972). The use of cultivated diploid *Solanum* species in potato breeding. In Prospects for the potato in the developing world: an international symposium on key problems and potentials for greater use of the potato in the developing world, Lima, Peru. Edited by ER French. International Potato Center (CIP), pp. 100-110.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576-589.

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R.E., Jansky, S., and Bethke, P. (2013). Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3 Genes Genom. Genet. 3, 1003-1013.

Hirsch, C.N., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121-135.

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. 16, 71-84.

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

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A., Shinde, B., Cardenas, P., Bocobza, S., Unger, T., Malitsky, S., and Finkers, R. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science 341, 175-179.

Jain, M., Tyagi, A.K., and Khurana, J.P. (2006). Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*). Genomics 88, 360-371.

Jaspers, P., et al. (2009). Unequally redundant RCD1 and SRO1 mediate stress and developmental responses and interact with transcription factors. Plant Journal 60, 268-279.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462-467.

Kidwell, M.G., and Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. 94, 7704-7711.

Klein, M., and Papenbrock, J. (2004). The multi-protein family of Arabidopsis sulphotransferases and their relatives in other plant species. J. Exp. Bot. 55, 1809-1820.

Knox, A.K., Dhillon, T., Cheng, H., Tondelli, A., Pecchioni, N., and Stockinger, E.J. (2010). CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. Theor. Appl. Genet. 121, 21-35.

Lai, J., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. 42, 1027-1030.

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., and Wang, B. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42, 1053-1059.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics., 11.17. 11-11.17. 14.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In arXiv preprint arXiv:1303.3997.

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

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189.

Li, R., et al. (2010). Building the sequence map of the human pan-genome. Nat Biotech 28, 57-63.

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., and Zheng, L. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32, 1045-1052.

Lu, F., et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun. 6, 6914.

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T., and Ma, H. (2012). Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. 22, 508-518.

Manrique-Carpintero, N., Coombs, J., Cui, Y., Veilleux, R.E., Buell, C.R., and Douches, D.S. (2015). Genetic map and quantitative trait locus analysis of agronomic traits in a diploid potato population using single nucleotide polymorphism markers. Crop Sci. 55, 2566-2579.

Maron, L.G., Guimarães, C.T., Kirst, M., Albert, P.S., Birchler, J.A., Bradbury, P.J., Buckler, E.S., Coluccio, A.E., Danilova, T.V., and Kudrna, D. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. 110, 5241-5246.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, pp. 10-12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303.

Morgante, M., De Paoli, E., and Radovic, S. (2007). Transposable elements and the plant pan-genomes. Curr. Opin. Plant. Biol. 10, 149-155.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2012). Crop genomics: advances and applications. Nat. Rev. Genet. 13, 85-96.

Mun, J.-H., Kwon, S.-J., Yang, T.-J., Seol, Y.-J., Jin, M., Kim, J.-A., Lim, M.-H., Kim, J.S., Baek, S., and Choi, B.-S. (2009). Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biol. 10, R111.

Ortiz, R. (2001). The state of the use of potato genetic diversity. Broadening the genetic base of crop production. CABI Publishing, Wallingford, 181-200.

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

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., and Graves, T.A. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326, 1112-1115.

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., and Douches, D.S. (2013). Construction of reference chromosome-scale pseudomolecules for potato: Integrating the potato genome with genetic and physical maps. G3 Genes Genom. Genet. 3, 2031-2047.

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

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8, 272-285.

Spooner, D.M., Núñez, J., Trujillo, G., del Rosario Herrera, M., Guzmán, F., and Ghislain, M. (2007). Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. Proc. Natl. Acad. Sci. 104, 19398-19403.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., and Rosenbaum, H. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PloS Genet. 5, e1000734.

St-Pierre, B., Besseau, S., Clastre, M., Courdavault, V., Courtois, M., Creche, J., Ducos, E., de Bernonville, T.D., Dutilleul, C., and Glevarec, G. (2013). Deciphering the evolution, cell biology and regulation of monoterpene indole alkaloids. Adv. Bot. Res 68, 73-109.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. 61, 437-455.

Sutton, T., Baumann, U., Hayes, J., Collins, N.C., Shi, B.-J., Schnurbusch, T., Hay, A., Mayo, G., Pallotta, M., and Tester, M. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318, 1446-1449.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science 320, 486-488.

Thompson, J.D., Gibson, T., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics., 2.3. 1-2.3. 22.

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

Uitdewilligen, J.G., Wolters, A.-M.A., Bjorn, B., Borm, T.J., Visser, R.G., and van Eck, H.J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. Plos One 8, e62355.

Varin, L., Marsolais, F., Richard, M., and Rouleau, M. (1997). Sulfation and sulfotransferases 6: Biochemistry and molecular biology of plant sulfotransferases. FASEB J. 11, 517-525.

Vazquez-Flota, F.A., and De Luca, V. (1998). Developmental and light regulation of desacetoxyvindoline 4-hydroxylase in *Catharanthus roseus* (L.) G. Don. Evidence of a multilevel regulatory mechanism. Plant Physiol. 117, 1351-1361.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, 164-164.

Wang, W., et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18, 1791-1802.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. 14, 125-138.

Wenzel, G., Schieder, O., Przewozny, T., Sopory, S., and Melchers, G. (1979). Comparison of single cell culture derived *Solanum tuberosum* L. plants and a model for their application in breeding programs. Theor. Appl. Genet. 55, 49-55.

Westfall, C.S., Herrmann, J., Chen, Q., Wang, S., and Jez, J.M. (2010). Modulating plant hormones by enzyme action. Plant Signaling & Behavior 5, 1607-1612.

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

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859-1875.

Xu, G., Ma, H., Nei, M., and Kong, H. (2009). Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc. Natl. Acad. Sci. 106, 835-840.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M., Bailey-Serres, J., Ronald, P.C., and Mackill, D.J. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442, 705-708.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., and Huang, L. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30, 105-111.

Xu, X., et al.; Potato Genome Sequence Consortium (2011). Genome sequence and analysis of the tuber crop potato. Nature. 475, 189-195.

Yan, H., Talbert, P.B., Lee, H.-R., Jett, J., Henikoff, S., Chen, F., and Jiang, J. (2008). Intergenic locations of rice centromeric chromatin. Plos Biol. 6, e286.

Yu, G., Nguyen, T.T., Guo, Y., Schauvinhold, I., Auldridge, M.E., Bhuiyan, N., Ben-Israel, I., Iijima, Y., Fridman, E., and Noel, J.P. (2010). Enzymatic functions of wild tomato methylketone synthases 1 and 2. Plant Physiol. 154, 67-77.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. 16, 172–183.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821-829.

Zhang, Z., Mao, L., Chen, H., Bu, F., Li, G., Sun, J., Li, S., Sun, H., Jiao, C., and Blakely, R. (2015). Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. Plant Cell, tpc. 114.135848.

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy Number polymorphisms in plant genomes. Theor. Appl. Genet. 127, 1-18.

CHAPTER 4

GENOME DIVERSITY OF TUBER BEARING *SOLANUM* SPECIES UNCOVERS TARGETS OF SELECTION DURING POTATO DOMESTICATION

Abstract

Cultivated potatoes were derived from wild *Solanum* species native to the Andes Mountains of South America, possessing a diverse gene pool of over 100 tuber-bearing relatives (*Solanum* section *Petota*). Evaluating the diversity of this group and its key domestication loci is critical to support genetic improvement of potato, the third most important food crop for direct human consumption. This study describes a survey of genome variation in tuber-bearing *Solanum* accessions including 20 wild diploid progenitor species, 20 primitive South American landraces, and 23 North American cultivars that identified ~44 million sequence variants in 390 Mb of the conserved potato genome. Despite past assertions that North American potato contains a narrow genetic base, the cultivated lineage at both the diploid and tetraploid level was found to possess higher sequence diversity than any major crop species reported to date. Allelic introgression from multiple wild species appears to have boosted cultivated potato diversity, with over 50% of the alleles in the coding sequence of wild species shared with cultivated genotypes. Selection analyses comparing South American landraces and North American cultivars to wild species identified a putative set of 2,612 potato genes (6.6%) separating the cultivated groups from wild species that includes a 'core' set of genes with demonstrated roles in domestication traits including steroidal glycoalkaloid biosynthesis , tuber carbohydrate metabolism and the cell cycle. This study offers an updated perspective of potato genome diversity for breeding in the 21st century.

Introduction

The cultivated potato (*Solanum tuberosum* L.) was domesticated from wild diploid species (2n=2x=24) 8,000 - 10,000 years ago by South Americans native to the Andes Mountains of

southern Peru (Spooner et al., 2005). Potatoes were eventually adopted and grown across the Andean highlands of Peru, Bolivia, and Ecuador, forming a keystone of the region's cultural heritage (Brush et al., 1981; Pearsall, 2008) and later spreading from the central Andes to the equatorial regions of modern Colombia and Venezuela and the coastal lowlands of southern Chile (Hosaka, 2004; Raker and Spooner, 2002; Spooner et al., 2012). Early diploid landraces (*S. tuberosum Stenotomum* and *Phureja* groups; 2n=2x=24) experienced at least one autopolyploidization event yielding Andean cultivated tetraploids (*S. tuberosum Andigena* group; 2n=4x=48), each falling under a multi-ploidy *Andigenum* subspecies housing all Andean cultivated potatoes not derived from wild hybridization (Ovchinnikova et al., 2011). The proposed migration of a landrace branch from the Andes to coastal Chile (Hosaka, 2003; Spooner et al, 2005) resulted in a long-day adapted subspecies distinct from its *Andigenum* counterparts (*S. tuberosum Chilotanum* group; 2n=4x=48), today representing the primary genetic background of modern potato cultivars outside the Andes. From these ancient origins, cultivated potato has since been widely adopted into the global diet and is now the third most important food crop worldwide in terms of direct human consumption (faostat3.fao.org), providing food security in South America, Africa, and Asia (Birch et al., 2012; Scott and Suarez, 2012). Expanding beyond global popularity, this highly adaptable crop has been selected by the National Aeronautics and Space Administration (NASA) in collaboration with the International Potato Center (CIP) as the first crop to undergo testing for cultivation in the harsh terrains of Mars (cipotato.org/press-room), highlighting the importance of the species in efforts to derive nutrition in new environments.

The adaptability of potato for diverse growing conditions stems from a large germplasm base; it possesses over 100 related tuber-bearing species (*Solanum* section *Petota*) (Spooner, 2009) with natural habitats ranging from the southwestern United States to southern Chile, offering exposure to diverse biotic and abiotic stresses (Hawkes, 1990; Ochoa, 1990; Spooner and Salas, 2006). Potato is unique among major crops, being first domesticated and grown at high altitudes in the Andean highlands (3,000-4,500 meters above sea level) (Zimmerer, 1998), an arid region characterized by cold temperatures and high solar radiation. The primary organ of selection was an underground storage tuber rather than aboveground reproductive structure, accumulating nutrients to lie dormant and sprout vigorously under favorable conditions. Recent evidence from chloroplast and nuclear DNA markers show the primitive diploid landrace progenitors of Andean tetraploids (*Phureja* and *Stenotomum* groups, collapsed into *Andigenum*) were likely derived from wild species in the northern distribution of the *S. brevicaule* complex (southern Peru), including *S. bukasovii*, *S. canasense* and *S. candolleanum* (Spooner et al., 2005; Sukhotu and Hosaka, 2006). This was later supported by genic sequence variation within a potato species core collection generated by the US Potato Genebank (Hardigan et al., 2015). These studies upheld a single domestication origin for cultivated potato as defined by a monophyletic group containing all cultivated genotypes descended from one or a few closely related wild populations located in southern Peru, upsetting a hypotheses based on starch grain morphology that Chilean landraces (*Chilotanum* group) were independently domesticated (Ugent et al., 1987). Existing hypotheses of potato's origins are derived from pre-genomics era datasets, and newer sequencing tools have the potential to resolve this controversy (Salazar and Estrada, 2008). This study represents the

first assessment of broad potato germplasm evolution using whole-genome sequencing that employs several million genome-linked neutral markers to address evolutionary history.

The historic importance and versatility of potato in deriving nutrition from diverse environments make its improvement a priority to meet global food demand. In the era of genomics-enabled breeding, evaluating the available pool of genetic diversity within tuber-bearing *Solanum*, and the impacts of domestication and breeding efforts on this diversity is critical to support effective breeding and germplasm utilization (Jansky et al., 2013). In addition to studying genetic diversity, identifying gene pathways that have undergone selection for desirable traits in modern varieties is vital to ensure continued genetic gains. Domestication and improvement of potatoes has involved selection on a variety of processes, above- and below-ground. Modification for human consumption required loss of toxic glycoalkaloids (Friedman et al., 1997; Johns and Alonso, 1990), and increased synthesis and transport of carbohydrates to the tuber (Bradshaw et al., 2006; Jansen et al., 2001). In this study, the genome diversity of a highly heterozygous crop and its progenitor species was assessed by resequencing, reporting signatures of selection for genes related to glycoalkaloid biosynthesis and carbohydrate pathway regulation.

Materials and Methods

Sample Preparation and Sequencing

Plant materials were obtained from the USDA potato gene bank (Sturgeon Bay, Wisconsin), and included germplasm from South American wild diploid species, diploid and tetraploid landraces, and North American cultivars (Table S3.1). Wild species and landraces were

germinated from botanical seed, and single healthy individuals were selected to represent populations. Cultivars were obtained as *in vitro* clones. DNA was purified from leaf tissues using the Qiagen DNeasy Plant Tissue Kit (Valencia, CA). Paired-end sequencing libraries (500-nt fragment size) were prepared as described previously (Hardigan et al. 2016) and sequenced in the paired end mode (125-nt length) to 8x theoretical sequence coverage (diploids) and 16x theoretical sequence coverage (tetraploids) on an Illumina HiSeq 2500 (San Diego, CA) at the Michigan State University Genomics Core.

Read Alignment and Variant Calling

Sequence reads were processed to remove low quality bases and adapters/primers using Trimmomatic (v0.32) (Bolger et al., 2014). Cleaned reads were aligned to the *Solanum tuberosum* Group Phureja DM potato reference genome (v4.04; Hardigan et al. 2016) using BWA-mem (v0.7.11) (Li, 2013). Alignments were processed using Picard tools (v2.1.1) (broadinstitute.github.io/picard) to mark duplicates, and locally realigned around insertion/deletion sites using the Genome Analysis Toolkit (GATK v3.3.0) (DePristo et al., 2011). Processed alignments were used to identify sequence variants genotypes with FreeBayes (v0.9.21.19), requiring minimum 4x depth for diploids and 8x for tetraploids. Alignments with MapQ score < 20 and bases qualities < 20 were excluded, and only properly oriented read pairs mapping to the same chromosome were used. Variants were filtered to remove low quality sites (QUAL < 20), mean mapping quality of reference (MQMR) or alternate (MQM) alleles < 20, and sites for which the mean reference allele quality differed from alternate allele quality by 10. Variants sites with 80% strand bias for any allele were excluded. Variant calling excluded regions within 150 bp of gaps to avoid false positives in

poorly resolved sequence. Singleton variant sites lacking multiple alternative alleles in the full population were removed from downstream analyses. Sample genotypes were filtered for genotype quality (GQ) > 20. Genome-wide copy number variation (CNV) was calculated by comparison of median read coverage within 5 kb windows and within genes to the genome-wide median coverage, with copy number reported as copies per monoploid genome (CN = [region med./genome med.] / ploidy). Sequences were classified as homozygous deletion/absent (CN < 0.1), partial deletion or heterozygous deletion (0.1 < CN < 0.6), conserved (0.6 < CN < 2.0), or duplicated (CN > 2). Analyses of conserved sequences in the study were limited to regions with a maximum 10% deletion rate in the respective cultivar, landrace and wild species groups.

Population Analysis and Phylogenetics

Population structure was calculated using FastStructure (v1.0) (Raj et al., 2014), testing K=1-10 with a set of 50,000 biallelic SNPs selected from a larger pool of SNPs randomly pulled from genome-wide 5 kb windows. Phylogenetic analysis and estimates of genetic distance were performed using the PHYLIP software package (evolution.genetics.washington.edu/phylip.html). A coding sequence SNP phylogeny was generated using 687,172 four-fold degenerate sites from conserved genes in the DM v4.03 genome annotation. Non-coding sequence phylogenies were generated using randomly sampled SNPs from conserved intergenic sequences equally derived from sliding 5 kb blocks to account for linkage bias and requiring a minimum of 5 kb from any annotated gene sequence. Relative genome-wide similarity of the landrace and cultivar populations to specific wild species was determined by calculating the average percent sequence identity of samples

in each group to specific species in 5 kb windows, and measuring the ratio of average cultivar identity to average landrace identity.

Selection Analysis

Several population statistics were employed as indicators of selection, including F_{ST}, Tajima's D, and nucleotide diversity relative to the founder population ($\pi_{wild}/\pi_{cultivated}$). F_{ST} was calculated using Hudson's estimator (Hudson et al., 1992) for all biallelic sites with minor allele frequency > 0.05. Tajima's D and population nucleotide diversity were calculated using BioPerl. F_{ST} values were calculated in 20-kb windows (5-kb step) using a combining approach described by Bhatia et al. (2013). Selective sweeps were generated by merging contiguous windows having an F_{ST} in the top 5% of all values. Genes were categorized under three levels of selection: 'putative', 'confident', and 'core' selection. Genes under putative selection were identified as those intersecting genomic windows ranked in the top 5% of F_{ST} window values, and ranking among the top 5% single-gene values for maximum single variant site F_{ST}, Tajima's D, or reduced nucleotide diversity ($\pi_{wild}/\pi_{cultivated}$). Genes within the top 5% of all three single-gene metric values were considered selected regardless of F_{ST} window overlap to allow for erosion of local linkage disequilibrium. Genes under 'confident' and 'core selection were identified by the same criteria using the top 2% and 1% of calculated values, respectively.

Results and Discussion

Potato Sequence Variation

The germplasm assembled for this study included 20 wild diploid species, 20 South American

landraces, 23 elite North American cultivars, and four outgroups (Table S3.1). Wild accessions represent the species diversity of South American diploids sexually compatible with cultivated potato, excluding allopolyploids or species of hybrid origin. Landraces were represented by 10 diploid and 10 autotetraploid accessions whose regional habitats range from equatorial Colombia to southern Chile. Four outgroup species included a distantly related South American series *Piurana* potato (*Solanum chomatophilum*), two primitive North American species (*Solanum ehrenbergii, Solanum jamesii*), and a non-tuber bearing relative *Solanum etuberosum*. Approximately 68.9 million SNPs were identified in total by alignment of the sequencing reads to the DM v4.04 potato reference genome, of which, ~44.0 million were within 390 Mb of conserved genome sequence containing 63% of genes (Table 3.1).

Population Analysis

Phylogenetic analysis of 687,172 four-fold degenerate sites from coding sequences (Figure 3.1) and an analysis of population structure (Figure 3.2) sub-divided the panel into three primary populations: (1) South American wild species, (2) South American *Andigenum* group diploid and autotetraploid landraces, and (3) the *Chilotanum* group containing Chilean landraces and their derived North American cultivars. Several Peruvian species (*Solanum medians*, *Solanum megistacrolobum*, *Solanum raphanifolium*) appeared to form a sub-population of the wild species, but remained phylogenetically closest to the larger wild species group. Populations from this coding sequence-derived SNP phylogeny were used as groups for downstream selection analyses due to its representation of diversity within genic regions.

Table 3.1. ***Single nucleotide polymorphism variant and allele counts in tuber bearing Solanum species.***

Population[a]	Sequence Type[b]	SNP Sites[c]	SNP Alleles
Landrace (2x) (10)	conserved CDS	641,128	1,290,226
	conserved genome	16,751,844	33,832,883
	full genome	26,560,638	53,596,632
Landrace (4x) (10)	conserved CDS	1,058,697	2,143,394
	conserved genome	25,945,403	52,829,361
	full genome	40,602,059	82,564,585
Cultivar (23)	conserved CDS	1,093,256	2,216,701
	conserved genome	27,765,783	56,690,615
	full genome	44,127,329	89,965,718
Wild Species (20)	conserved CDS	1,110,047	2,254,788
	conserved genome	31,105,696	63,763,279
	full genome	46,797,252	95,820,132
Outgroup (4)	conserved CDS	493,028	996,222
	conserved genome	15,290,197	30,959,823
	full genome	23,279,525	47,112,132
Full Panel (67)	conserved CDS	1,783,737	3,648,046
	conserved genome	44,001,844	90,943,700
	full genome	68,914,903	142,150,253

[a] Numbers indicate group sample size.
[b] Conserved CDS (coding sequence) and conserved genome sequences were filtered to allow maximum of 10% deletion based on copy number variation data.
[c] Single nucleotide polymorphism sites excluded singletons and variants with >20% missing data.

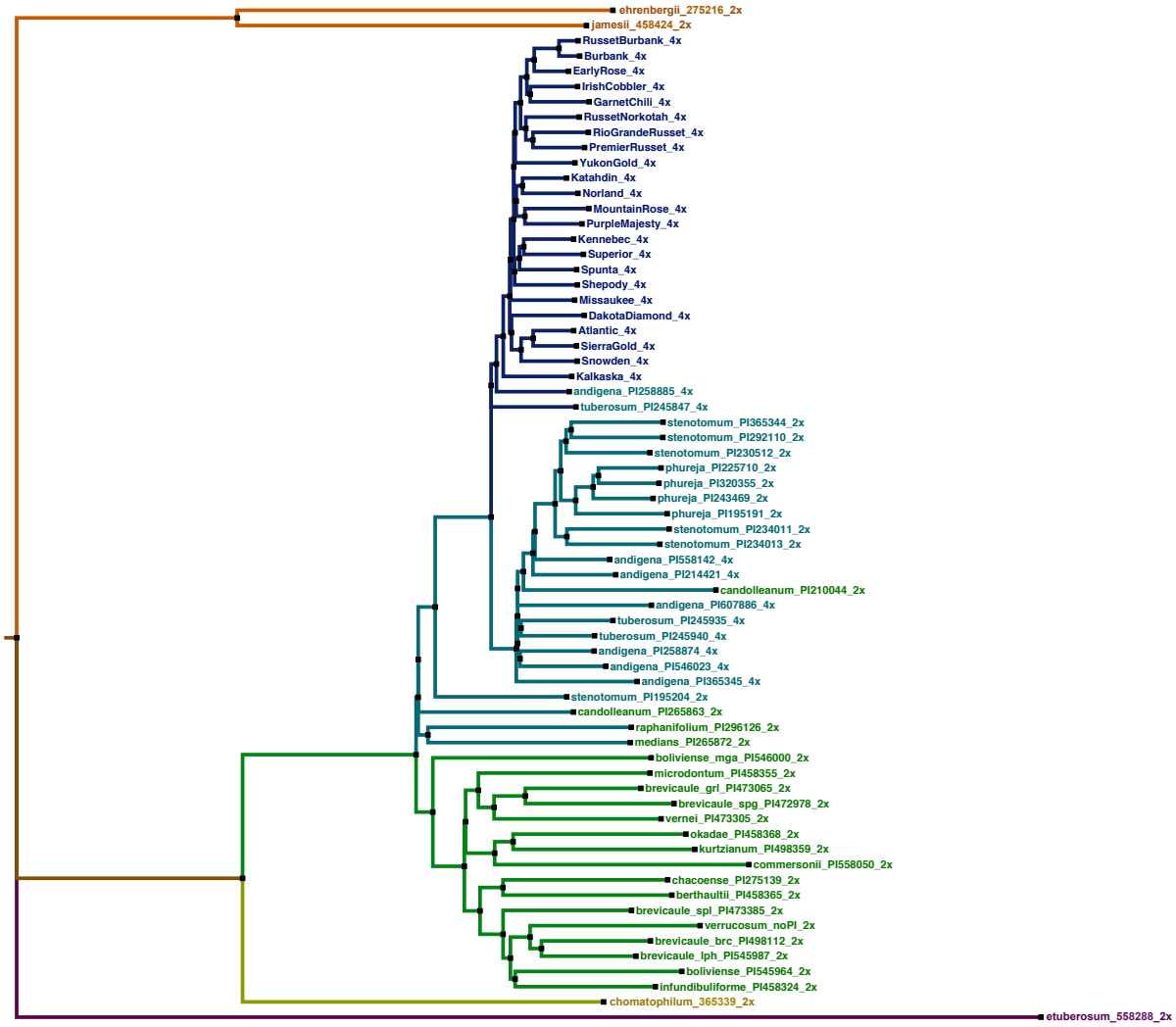

Figure 3.1. ***Phylogenetic tree of domestication panel samples based on 687,172 four-fold degenerate single nucleotide polymorphism sites.*** *Branches and taxa are colored to represent wild South American diploids (green), S. tuberosum Andigenum group landraces (turquoise), S. tuberosum Chilotanum group landraces and cultivars (blue), Mexican outgroup species (orange) and S. etuberosum (purple).*

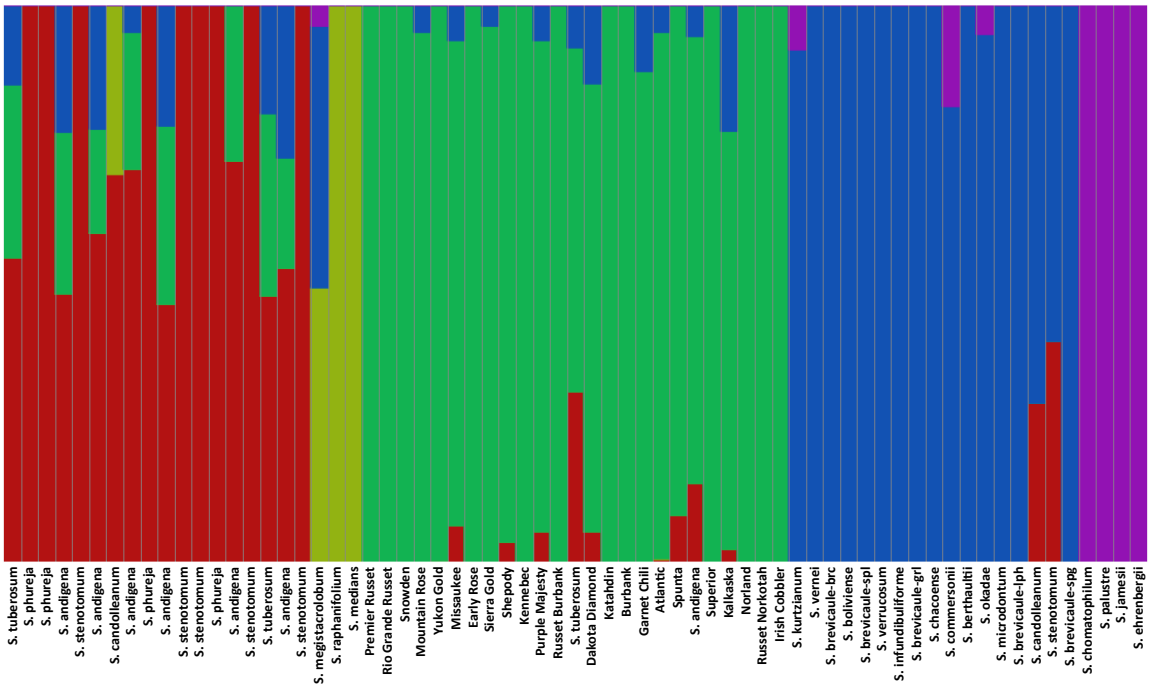

Figure 3.2. ***Population structure of domestication panel samples (K=5), including the Andigenum landrace group (red), Chilotanum landrace and cultivar group (green), primary wild South American diploids (blue) group, secondary wild South American diploid group (yellow), and outgroup species (purple).***

Origin of Potato Domestication

Potato systematics and hypotheses on the origins of its cultivated lineages have been challenged by a diverse species pool with cryptic hybridization and sparse molecular data (Spooner, 2016; Spooner et al., 2014). To more appropriately address the origin of domesticated potato, the phylogeny of *Solanum* sect. *Petota* was examined using genome-wide neutral SNPs derived from conserved intergenic sequences and selected to avoid local genome bias (Figure 3.3). These data produced a tree supporting the hypothesis of the domestication of potato in southern Peru from wild diploid species of the *S. brevicaule* species complex (Spooner et al., 2005). A member of the northern "brevicaule complex" (*S. candolleanum* [PI265863]) collected near Lake Titicaca was basal to a clade containing the entire cultivated lineage, with SNP data supporting a monophyletic origin of the Andean (*Andigenum*) and Chilean (cultivar and *Chilotanum* landrace) genetic groups. These diploid species still grow in and around South American fields forming crop-weed complexes, resulting in misidentified hybrids (PI195204, PI210044) when relying on morphology (Rabinowitz et al., 1990; Vandenberg et al., 1998). Resolution of a monophyletic cultivated lineage with basal species native to Peru advances the theory of a single Andean domestication origin, failing to support an independent domestication of the Chilean tetraploid branch that gave rise to modern cultivars.

Potato Genome Diversity

Potatoes suffer from severe inbreeding depression and have long been assumed to harbor significant heterozygosity thereby benefitting from heterosis (De Jong and Rowe, 1971; Mendoza and Haynes, 1973). Though diverse at the single-clone level, North American

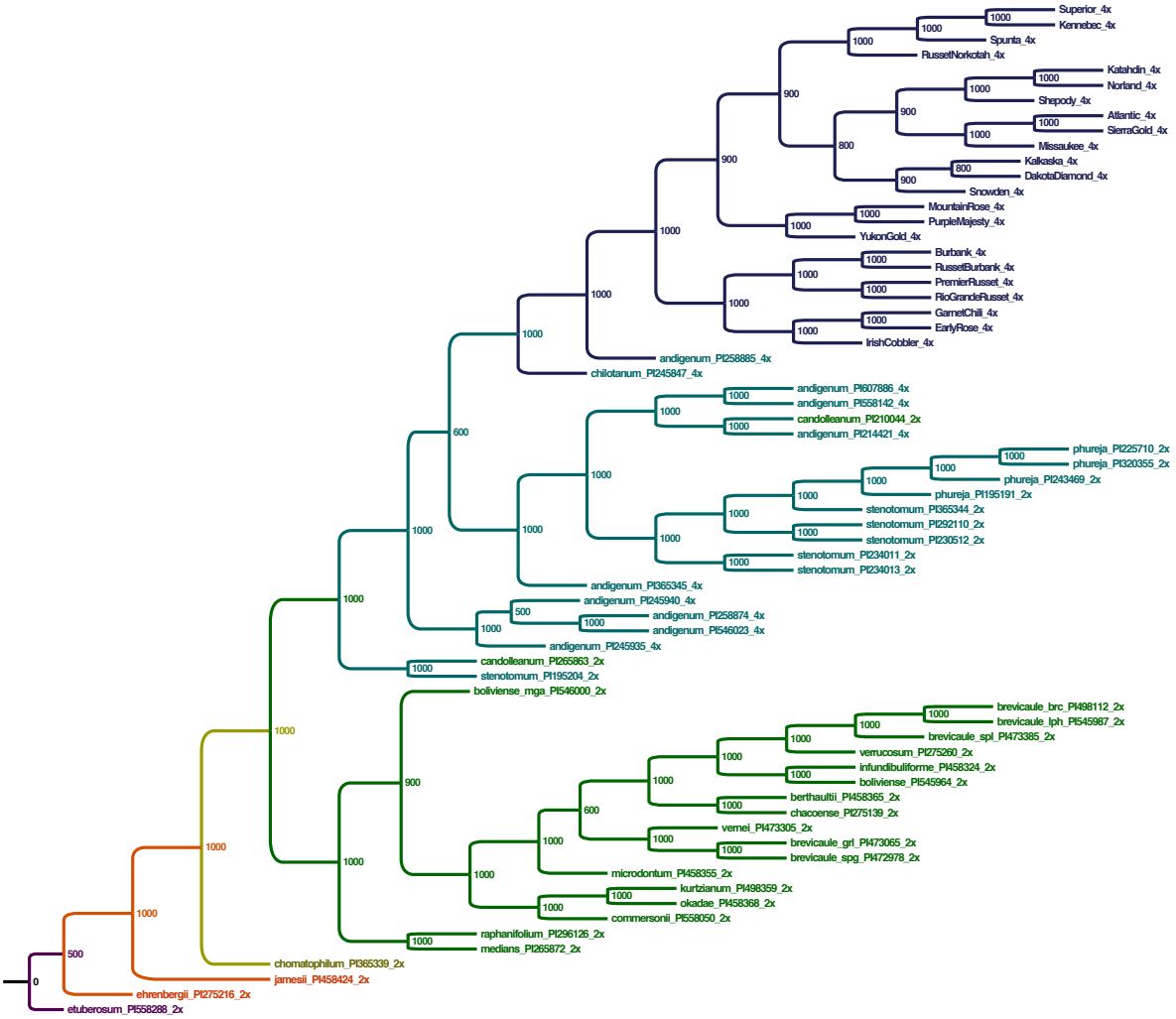

Figure 3.3. ***Phylogenetic reconstruction of domestication panel using 6.4 million genome-wide neutral intergenic single nucleotide polymorphisms.*** *Branches and taxa are colored to represent wild South American diploids (green), S. tuberosum Andigenum group landraces (turquoise), S. tuberosum Chilotanum group landraces and cultivars (blue), S. chomatophilum (olive), Mexican outgroup species (orange) and S. etuberosum (purple). Numbers indicate accession ID and ploidy.*

cultivars are regarded as derived from a narrow base of only a few Chilean landraces and European varieties that survived the genetic bottleneck imposed by a *Phytophthora infestans* epidemic in the 1800s (Irish potato famine) (Haas et al., 2009; Love, 1999; Plaisted and Hoopes, 1989). Sequence diversity (π) was assessed in each population using SNPs from ~390 Mb of conserved genome sequence (requiring ≤10% deletion in each population to be included) to estimate genome diversity and potential genetic bottlenecks. Cultivated potatoes harbor striking levels of diversity compared to estimates from previous crop resequencing studies (Table 3.2), with $\pi_C = 0.0105$ (N.A. cultivars), $\pi_L = 0.0097$ (S.A. landraces) and $\pi_{C+L} = 0.0111$ (cultivated pool), challenging theories of low founder diversity (Table S3.2). Despite lacking heterozygosity of their tetraploid counterparts ($\pi_{L\text{-}4x} = 0.0109$), diploid landrace diversity ($\pi_{L\text{-}2x} = 0.0087$) exceeded values found in landrace populations of maize (Hufford et al., 2012), supporting reports that potato harbored atypical levels of genome variation before undergoing autopolyploidy (Hardigan et al., 2016). Observing roughly similar genome diversity in the combined *Andigenum* and *Chilotanum* lineages and wild germplasm ($\pi_W/\pi_{C+L} = 1.101$, [1.188 in coding sequence]) is remarkable given they represent sub-groups within a single *S. tuberosum* species, while twenty distinct species constituted the wild group. This shows *S. tuberosum* acquired a level of diversity in the several millennia since its domestication that required potentially millions of years of diversification among the tuber-bearing *Solanum* species of South America.

Autopolyploidy more than doubled the genome heterozygosity of diploid potatoes (Figure 3.4). The mean frequency of heterozygous nucleotide sites rose from 1.05% (2x) to 2.33% (4x) in South American clones. Given an assumption of similar backgrounds producing

Table 3.2. ***Population genetic diversity and domestication bottlenecks from multiple crop resequencing studies.***

Species[a]	Population[b]	Samples	Diversity (π)	Wild Diversity Ratio
Cucumber	W	30	0.0045	1.96
	C	85	0.0023	
Watermelon	W	10	0.0076	5.43
	C	10	0.0014	
Tomato	W	16	0.0042	2.63
	C	23	0.0016	
Rice	W	446	0.0030	1.25
	C	1083	0.0024	
Maize	W	17	0.0059	1.23
	C	23	0.0048	
Soybean	W	17	0.0030	1.58
	C	24	0.0019	
Potato	W	20	0.0122	1.10
	C	43	0.0111	

[a] Species data for cucumber and tomato from Qi et al., 2013; watermelon data from Guo et al., 2013; rice data from Huang et al., 2012; maize data includes landrace values from Hufford et al., 2012; soybean data from Lam et al., 2010.
[b] W, wild group; C, cultivated group.

autotetraploids (by sexual polyploidization, genome doubling), this supports wild species contributing additional alleles to those derived from diploid founders of potato tetraploids. High levels of allele sharing within coding sequences were confirmed in wild species, landraces, and North American cultivars, with 73% of the identified wild alleles present in cultivated germplasm (53% when rare alleles are included) (Figure S3.1). Allele dosage of tetraploid potatoes has shown elite varieties contain a 2:1 bias of single-dose (simplex-AAAB/triplex-ABBB) sites to duplex (AABB) sites, whereas breeding stocks containing wild genetics are nearer a 1:1 balance, suggesting that recent introgressions cause a temporary shift from the predominance of single dose allelic variation associated with populations under tetrasomic inheritance (Hirsch et al., 2013). High rates of the simplex allele dose did not support recent introgressions in most tetraploids; however several landraces (PI607886, PI365345, PI245935) and the 19th century cultivar Garnet Chili demonstrated chromosome-specific reduction of simplex dominance attributed primarily to homozygosity of non-reference alleles in these locations (Figure S3.2). Cultivars averaged 2.73% heterozygous nucleotide sites (max. 3.04%), suggestive of selection for heterosis and gene interaction (Mendoza and Haynes, 1974). Despite a near comparable level of genome diversity in cultivars and wild species ($\pi_W/\pi_C = 1.157$) and high heterozygosity, mean pairwise genetic distance among cultivars based on all variant sites was 0.026, 3.27-fold lower than among wild accessions (0.085), showing F1-hybrid selection has re-partitioned genetic diversity between potato populations into allelic diversity (heterozygosity) that is fixed clonally within the individual. These results agree with previous examinations of tetraploid cultivars that showed potato breeding populations are without significant structure, and breeders are engaged in a process of re-shuffling heterozygosity to clonally isolate allele combinations for

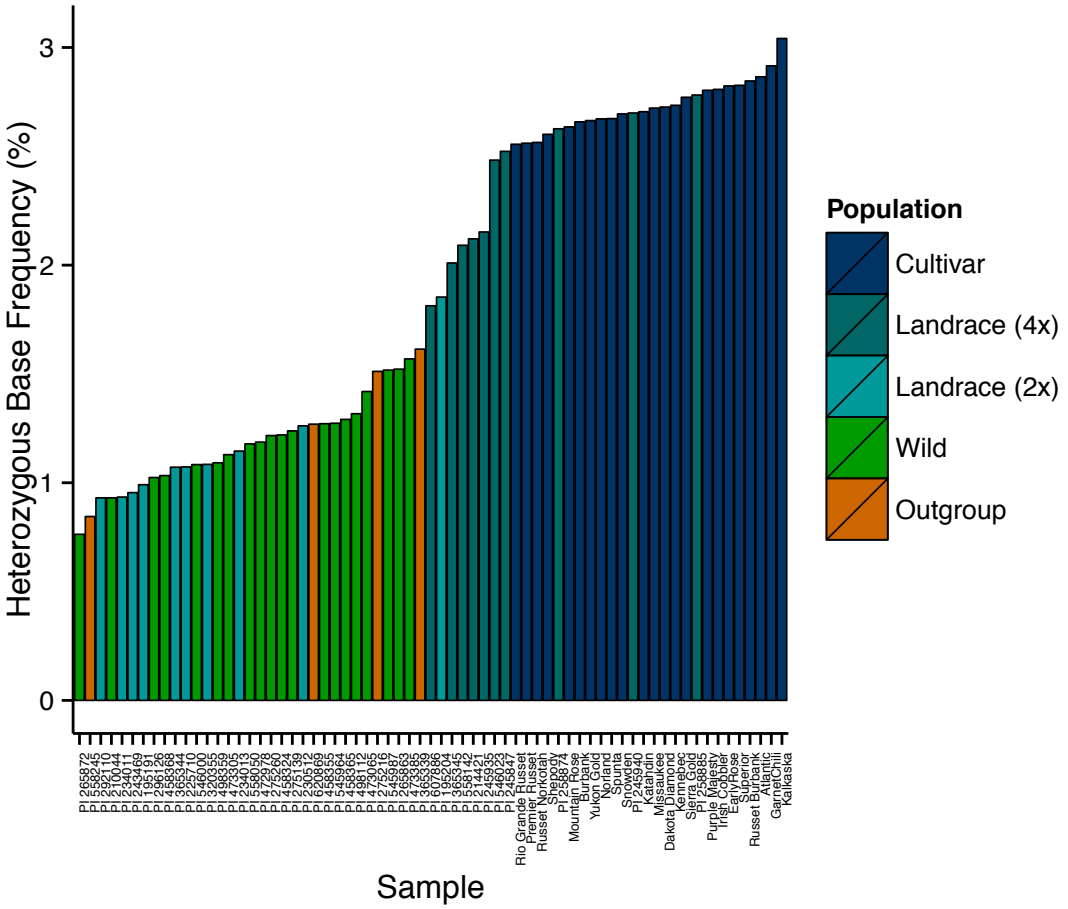

Figure 3.4. ***Frequency of heterozygous nucleotides in the genomes of domestication panel samples.***

optimizing gene interaction and masking deleterious mutations (Hirsch et al., 2013).

Identifying Genes Under Selection in Cultivated Potato

Allele frequencies from the three populations (cultivar, landrace, wild) were used to scan the potato genome for signatures of selection. Genes were classified within three selection categories by divergence between population groups: North American *Chilotanum*-derived cultivars versus wild diploid species (CVR-selected), South American *Andigenum* landraces versus wild diploid species (LND-selected), and northern versus southern hemisphere cultivated germplasm (HEM-selected). Domestication gene candidates were regarded as the intersection of genes under selection in landraces (LND) and cultivars (CVR) relative to wild species, implying function in agricultural performance regardless of hemisphere or local adaptation. Genomic sweeps in linkage with selected loci were identified using 25 kb window F_{ST} values (Figure 3.5), then calculating single-gene statistics (Tajima's D, reduced nucleotide diversity, maximum single-variant F_{ST}) of genes within sweeps. Genes were classified under three thresholds of selection (putative > confident > core) to enable analysis of gene function enrichment in a broader set, and manual evaluation of a smaller core group. The broader set of 'putative' selected loci contained genes (plus 1 kb upstream region) within the top 5% F_{ST} windows and the top 5% values for at least one single-gene selection metric. Genes in the top 5% of all three single-gene metrics were included regardless of F_{ST} window scores to allow for erosion of local linkage disequilibrium. 'Confident' and 'core' selected loci were identified with the same criteria as putative selected loci but using 2% and 1% cutoffs, respectively. With these criteria, 3,900 potato genes (9.89%) were regarded as putatively selected in at least one population, 1,278 (3.24%) confidently selected, and 494 (1.25%)

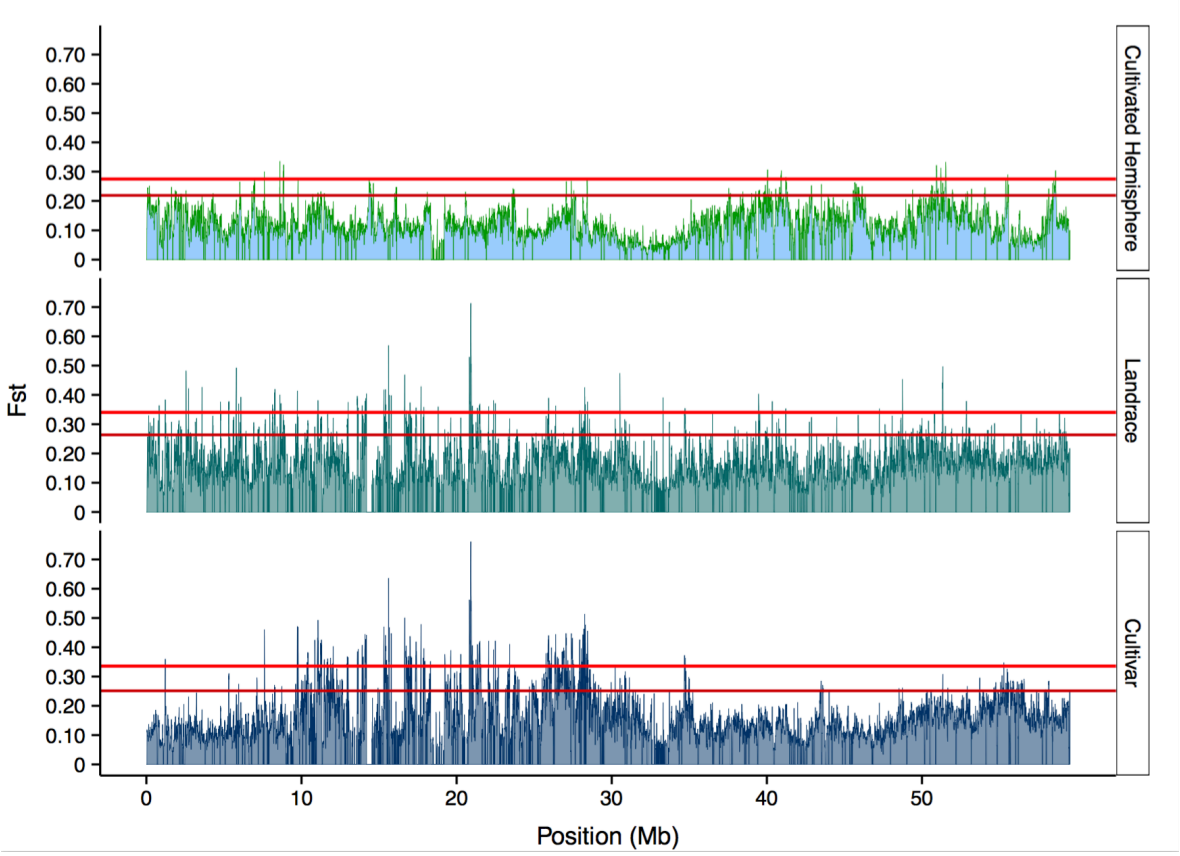

Figure 3.5. ***Genome-wide patterns of selection based on F_{ST} values calculated in 25 kb windows (5 kb step) on potato chromosome 6.*** *Horizontal lines represent the 5% and 1% value cutoffs for each population comparison.*

within the core selected set (Figure 3.6; Figure S3.3; Table S3.3). Excluding the set of genes distinguishing *Andigenum* landraces and *Chilotanum* cultivars (HEM-selected), 2,612 genes (6.62%) were identified as candidate genes separating cultivated potatoes from the wild species (CVR + LND).

It is important to note that these selection cutoffs do not imply an absence of selection pressure on other genes in the potato genome, as all analyzed population statistics exhibited a broad continuum of values. Rather, the CVR, LND and HEM-selected gene sets should be regarded as ranked genes under a degree of selection that exceed a reasonable cutoff to permit statistical analyses of functional enrichment and representation of selected genes within candidate pathways associated agricultural performance.

Selected Loci Within Candidate Domestication Pathways

Representation of loci from the CVR, LND and HEM-selected gene classes was evaluated within gene pathways having known roles in potato domestication based on the key phenotypes distinguishing cultivated potatoes from their wild relatives. Foremost among these were the analysis of selection signatures in the steroidal glycoalkaloid biosynthesis and carbohydrate metabolic pathways, pertaining to reduced toxicity, increased carbohydrate partitioning within cultivated tubers and cell cycle regulation.

Steroidal Glycoalkaloid Biosynthesis

A variety of steroidal glycoalkaloids (SGAs) give wild potatoes above- and belowground resistance to insects and pathogens, but are also toxic to humans (Friedman, 2006). Reducing

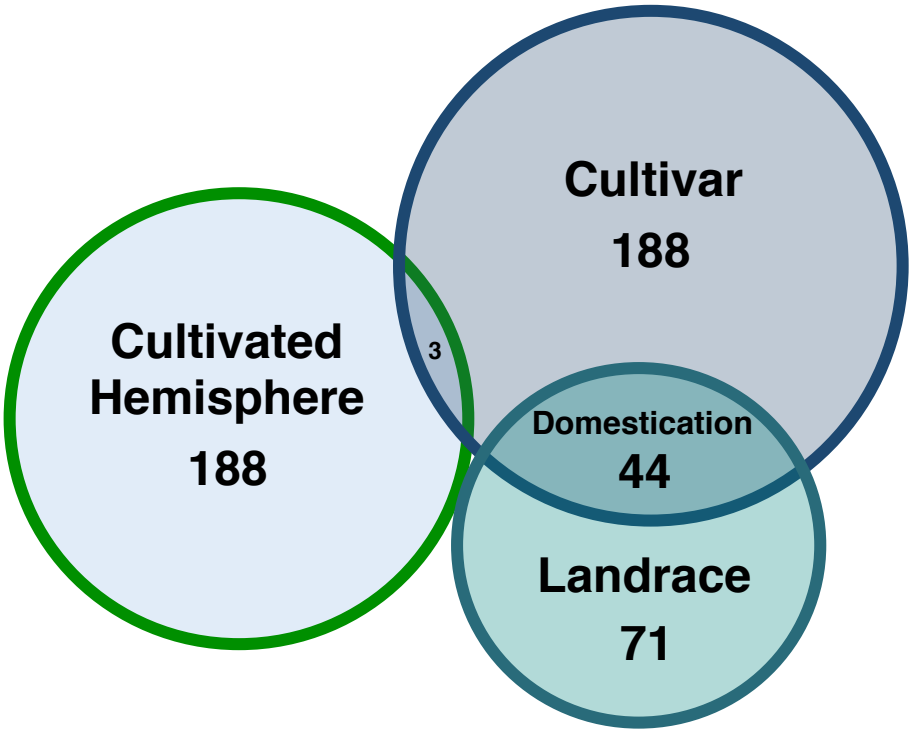

Figure 3.6. ***Venn diagram demonstrating overlap of genes in the core selection group for each population test.*** *Numbers represent gene counts in each category.*

tuber glycoalkaloids was necessary for direct consumption or processing (Maga and Fitzpatrick, 1980; Sinden et al., 1984). Genes encoding key enzymes of the SGA-specific portion of the primary metabolic pathway lacked selection. The strongest signal observed for any SGA-related gene was identified in squalene synthase (*SQS*, PGSC0003DMG400003408), the enzyme that synthesizes 2,3-oxidosqualene, guiding terpenes into the sterol biosynthetic pathway (Ginzberg et al., 2009; Ginzberg et al., 2012). *SQS* contains fixed allele differences from wild potato in landraces and cultivars. A cluster of APETELA2/ethylene response factor (ERF) genes on potato and tomato chromosome 1 was previously shown to contain *GLYCOALKALOID METABOLISM 9* (*GAME9)* (Cárdenas et al., 2016), a key gene in regulation of the glycoalkaloid biosynthetic pathway, including genes encoding enzymes within the SGA-specific pathway (Itkin et al., 2013). Strong signatures of selection for landraces and cultivars corresponding to potato *GAME9* (PGSC0003DMG400025989) were identified, and potential local signals for *GAME9-like7* (PGSC0003DMG400040573) and additional *GAME9-like* homologs in the cluster (Figure 3.7). Allelic variation within the *GAME9* locus suggests major diversification among wild species, while only three primary genotype groups present in cultivated genotypes that lack signatures of wild introgression (Figure 3.8). Interestingly, *SQS* and upstream enzymes are not regulated by *GAME9* (Cárdenas et al., 2016), suggesting the possibility of selection at multiple levels of the SGA pathway including regulation of sterol substrate flux at the top level (via *SQS*) and activity of downstream genes involved in hydroxylation, transamination and glycosylation of cholesterol to generate SGAs (via *GAME9*).

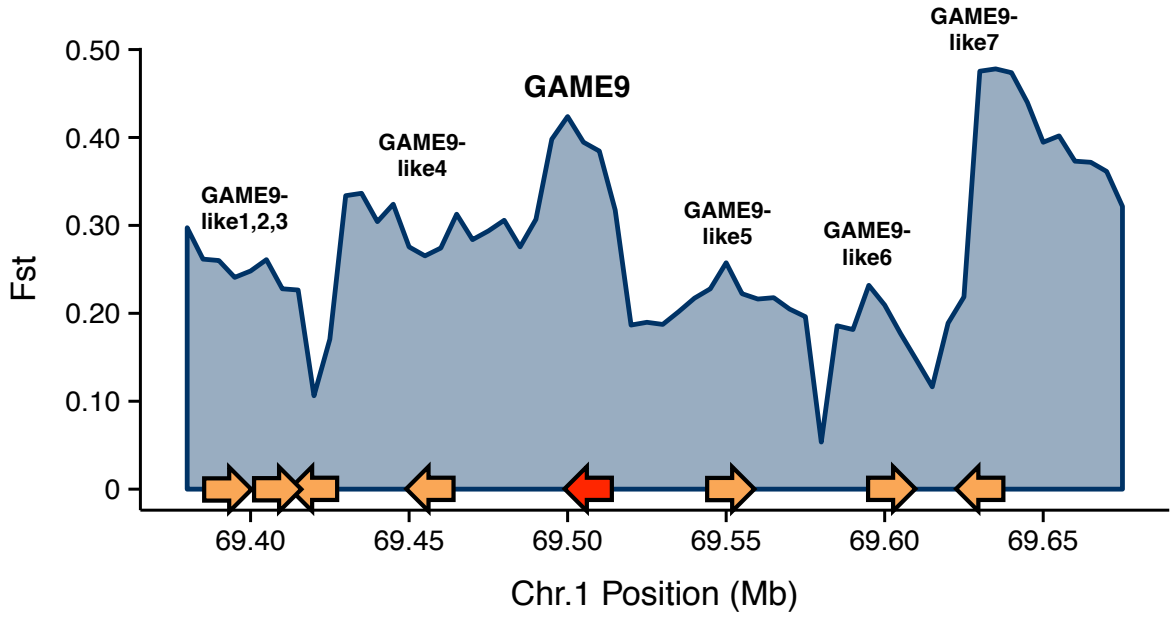

Figure 3.7. ***Selective sweep patterns in potato chromosome 1 region containing the GAME9 locus (red arrow) regulating steroidal glycoalkaloid pathway enzymes and surrounding gene cluster of GAME9-like genes (orange).***

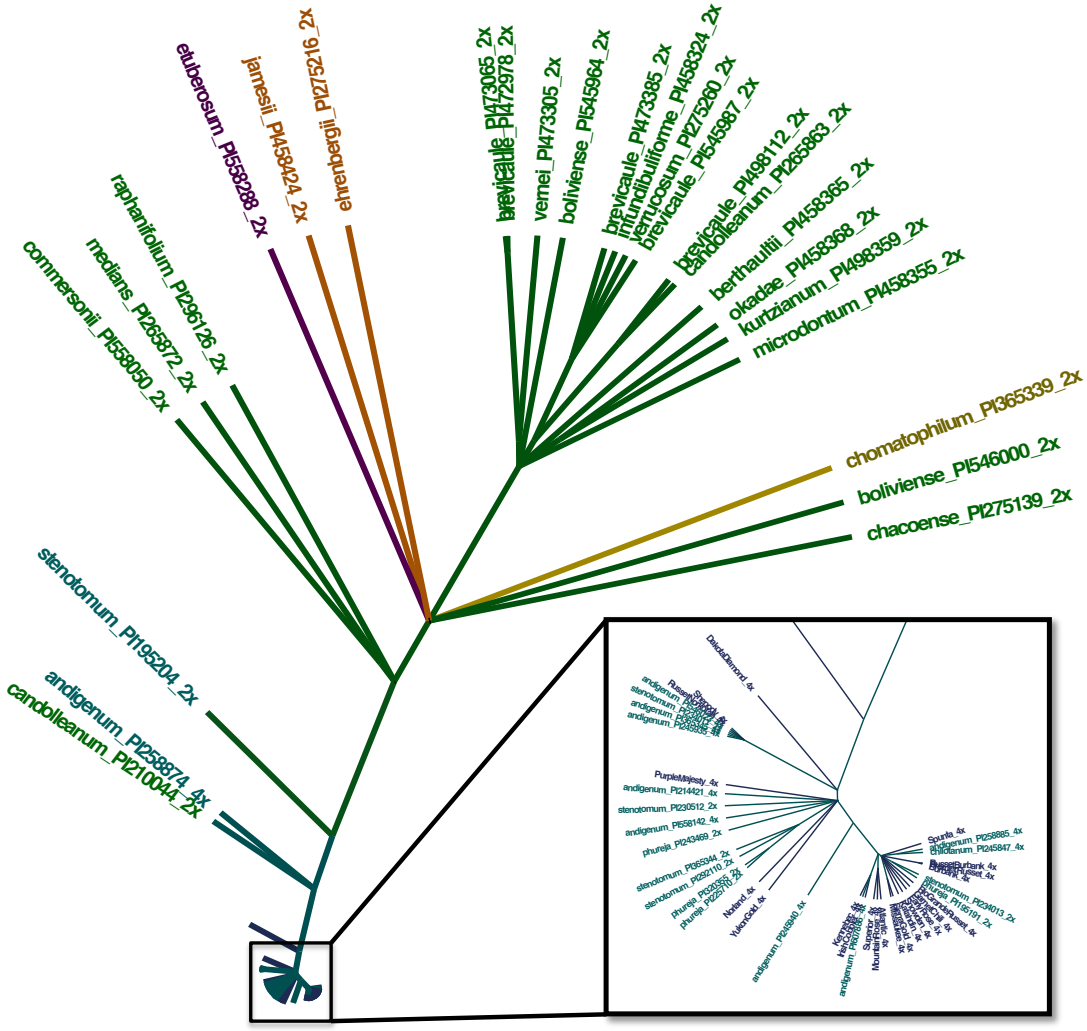

Figure 3.8. ***Gene tree based on DNA sequence diversity at the GAME9 locus regulating glycoalkaloid biosynthesis.*** *Branches and taxa are colored to represent wild South American diploids (green), S. tuberosum Andigenum group landraces (turquoise), S. tuberosum Chilotanum group landraces and cultivars (blue), Mexican outgroup species (orange) and S. etuberosum (purple).*

Carbohydrate Biosynthesis

In addition to vitamins and protein, potato tubers are valued for their high starch content derived from leaf sugars unloaded in the tuber via the phloem. Specifically, cytosolic sucrose (Suc) levels are a key determinant in establishing tuber sink status, with Suc mobilization primarily regulated by invertases during early tuber initiation (Fernie and Willmitzer, 2001), having a strong impact on tuber respiration and glycolysis, and downstream effects on tuber size and number (Sonnewald et al., 1997). Genes functioning in sugar transport/metabolism and starch biosynthesis were screened for signatures of selection, yielding little evidence for a conserved impact of domestication on non-regulatory proteins and enzymes in the carbohydrate pathway (Table S3.5). Of 232 carbohydrate transport and metabolism genes, a single invertase inhibitor (PGSC0003DMG400001276) showed selection in both landraces (core set) and cultivars (confident set), demonstrating a predominance of landrace and cultivar-specific selection in genes controlling potato carbohydrate metabolism.

Andigenum landraces exhibited further selection within genes modulating Suc flux, containing their strongest signatures in sucrose-phosphate synthase (*SPS*; PGSC0003DMG402019060), a sugar transport protein (PGSC0003DMG400006636) and fructokinase (PGSC0003DMG400024246) (Table S3.5), all genes encoding proteins that function in sugar metabolism, upstream of starch synthesis (Geigenberger et al., 2004; Huber and Huber, 1996; Williams et al., 2000). Tuber growth and starch synthesis are diurnally regulated by Suc, and reduced *SPS* activity results in attenuated diurnal changes in leaf sugars and Suc flux to tubers (Geigenberger and Stitt, 2000). Selection on *SPS* and sugar transport/metabolism genes point in part to partial regulation of tuber starch accumulation via sucrose supply in landrace

potatoes. Cultivar-specific selection in the carbohydrate pathway was strongest within two genes encoding inorganic pyrophosphatase proteins (PGSC0003DMG400025085, PGSC0003DMG400026784) (Table S3.5). During the tuber bulking process, the primary source of Suc mobilization for starch synthesis shifts from invertase to sucrose synthase (*Susy*) (Appeldoorn et al., 1997), relying on inorganic pyrophosphate (PPi) for enzymatic cleavage. It was proposed that inorganic pyrophosphatases could play a role in recycling PPi byproducts of starch synthesis in the amyloplast, exporting PPi to the cytosol to be made available for *Susy* (Farré et al, 2000). Geigenberger et al. (1998) showed that over-expression of inorganic pyrophosphatase in potato resulted in up to 30% increases in tuber starch due to the increased activity of *Susy* and ADP-glucose pyrophosphorylase (AGPase). Signatures of selection in the carbohydrate pathway demonstrate domestication-related, and additionally, cultivar and landrace-specific selection for distinct improvement genes controlling sucrose flux into starch biosynthesis. Confirming diversification of the pathway in the *Andigenum* and *Chilotanum* lineages, a number of carbohydrate pathway genes contained selection signatures that set the hemispheric groups apart (Table S3.5), including a sucrose transporter (PGSC0003DMG400009213), starch synthase (PGSC0003DMG401013540), sucrose synthase (PGSC0003DMG400013546), sucrose phosphatase (PGSC0003DMG400028134) and sucrose phosphate synthase (PGSC0003DMG400027936).

In addition to the genes that encode structural proteins in carbohydrate biosynthesis, domestication may also affect key regulators of carbohydrate pathway proteins such as the sucrose non-fermenting-1 (SNF1) related kinase (*SnRK1*) protein complex. SNF1-related protein kinases are global regulators of energy and carbon metabolism conserved in all eukaryotes (Halford et al., 2000), linking carbon metabolism to developmental shifts and plant

stress response (Halford et al., 2003; Polge and Thomas, 2007). The plant complex (*SnRK1*) directly regulates sucrose-phosphate synthase (*SPS*) (Sugden et al., 1999), is required for Suc mediated-induction of sucrose synthase (*Susy*; sucrose mobilization) (Purcell et al., 1998), impacts α-amylase activity (starch degradation) (Laurie et al., 2003), and stimulates ADP-glucose pyrophosphorylase (AGPase), the key regulatory enzyme of starch biosynthesis, in response to Suc flux (Polge and Thomas, 2007). In potato, it was found that *SnRK1* expression peaks in stolons as they undergo tuber initiation, gradually declining in mature tubers (Man et al., 1997), and is required for sucrose-induction of sucrose synthase (Purcell et al., 1998), demonstrating a regulatory role in establishing a Suc-mediated sink status in young tubers. A gene encoding the plant-specific β3 subunit of the *SnRK1* complex (PGSC0003DMG400015733; Table S3.3) was under selection in both landraces and cultivars (core selected set in cultivars, putatively selected set in landraces), suggesting that this master regulator of plant energy homeostasis (Halford and Hardie, 1998) may be a key factor in the impact of potato domestication and breeding on tuber development.

Cell Cycle Regulation

Potato tuber development contains two major aspects: biochemical changes leading to starch accumulation (discussed above), and morphological development of a tuber organ by initiation of cell division in the stolon tip. At the onset of a mobile tuberization signal, cells in the stolon pith and cortex divide longitudinally swelling the stolon tip until reaching a diameter of ~1 cm. Subsequent enlargement of the pith and cortex along with the onset of division and expansion in the perimedullary cells until harvest account for the majority of the final size of the tuber (Xu et al., 1998). It was previously hypothesized that cell cycle

regulation and endoreduplication could be processes that were modified in elite potato varieties to alter tuber initiation and tuber cell size, based on observations of SNPs within the KAKTUS gene (a regulator of endoreduplication and cell size) (El Refy et al., 2004) which contained ~90% allele frequency differences between tetraploid cultivars and South American germplasm (Hardigan et al., 2015). Both KAKTUS homologs in potato (PGSC0003DMG401011946, PGSC0003DMG402011946) were identified as putatively selected in cultivars (Table S3.6), but were not selected in landraces. Four selection candidates for domestication were identified among genes regulating cell cycle and endoreduplication: cell cycle switch 52B (CCS52B; PGSC0003DMG400029675), S phase kinase-associated F-box protein 2 (SKP2; PGSC0003DMG400027774), a CDK-activating kinase (CAK) assembly factor (PGSC0003DMG400024579), and cyclin dependent kinase C (CDKC; PGSC0003DMG400001882). Each gene was under putative selection in landraces, and confident (CCS52B, SKP2, CAK) or core (CDKC) selection in cultivars (Table S3.6). Cultivars demonstrated CVR-specific selection within cyclin A2 (PGSC0003DMG400045458), cyclin A3 (PGSC0003DMG400022017), and cyclin D3 (PGSC0003DMG400012007), cell division protease ftsH (PGSC0003DMG400018774), cell division protein kinase 7 (PGSC0003DMG400025069), and an additional SKP2 homolog (PGSC0003DMG400027776). Landraces contained LND-specific selection signatures in a CDK5 subunit (PGSC0003DMG400003535), an unknown CDK (PGSC0003DMG400017127), and cell division cycle 45 (CDC45; PGSC0003DMG400027946) (Table S3.6). Landraces revealed no confident or core selection candidates in cell cycle genes, whereas cultivars contained six confidently selected genes and one core selected gene (CDKC), suggesting broader and stronger selection for cell cycle

regulation in the cultivated group with larger tuber size and higher yield. Identification of two SKP2 F-box protein homologs confidently selected in cultivars (one putatively selected in landraces) is interesting given our findings related to *SnRK1* selection in the carbohydrate pathway; SKP2 guides the plant SCF ubiquitin ligase complex to degrade E2F-1, a transcription factor for numerous S-phase transition genes (Marti et al., 1999), and *SnRK* interacts with the SCF complex S-phase kinase associated protein (Skp) subunit to mediate proteosomal binding (Farrás et al., 2001), demonstrating a potential regulatory link between the carbohydrate pathway and tuber cell division.

Enrichment of Biological Functions in Genes Under Selection

Consistent with the predominance of *Andigenum* landrace and *Chilotanum* cultivar-specific selection signatures observed in the carbohydrate biosynthetic pathway, the overall numbers of domestication loci which were defined as the intersection of the CVR and LND selection sets, were consistently lower than the numbers of CVR or LND-specific genes identified under each of the putative, confident and core selection thresholds (Figure 3.6). The numbers of domestication candidates intersecting CVR and LND genes were 413 (10.6% selected genes) in the putative set, 116 (9.0% selected genes) in the confident set, and 44 (8.9% selected genes) in the core set (Figure S3.3). The limited overlap of genes under selection between cultivars and landraces suggests that domestication may have involved relatively few loci while distinct sets of improvement genes are further responsible for driving performance of the distinct *Andigenum* and *Chilotanum* cultivated groups.

The full complement of selected potato genes was used to explore broader functional

enrichment within selection groups. To determine the biological processes associated with genes exhibiting selection in the CVR and LND groups, we analyzed Gene Ontology (GO) annotations associated with cultivars (CVR-exclusive genes), landraces (LND-exclusive genes), and domestication (CVR-LND intersection) at each of the three confidence thresholds. This approach identified 199 significantly enriched biological terms, with 49 GO terms enriched in both CVR and LND compared to 66 GO terms in landraces only (LND-exclusive) and 88 GO terms in cultivars only (CVR-exclusive) (Table S3.4). Surveying terms specific to either cultivated group revealed enrichment for selected gene functions in common processes controlling agricultural performance including abiotic stress, nitrogen uptake, and cell development, but via distinct mechanisms. This included enrichment in landraces for nitrate assimilation (P=0.0163; confident selected, LND-specific), and enrichment in cultivars for ammonium transport (P=0.0132; confident selected, CVR) and sulfate transport (P=0.0347; confident selected, CVR-specific), and cellular response to phosphate starvation (P=0.0078; confident selected, CVR-specific). *Andigenum* landraces had enriched selection in heat response (P=0.0204; confident selected, LND-specific) and heat shock proteins (P=0.0396; putative selected, LND-specific), and *Chilotanum* cultivars showed enrichment in response to freezing (P=0.0093; putative selected, CVR-specific) and oxidative stress (P=0.0238; putative selected, CVR-specific). *Andigenum* varieties appear to have selected genes for regulation of cell growth and expansion via *Andigenum*-specific selection in genes associated with the TORC1 complex (P=0.0244; confident selected, LND-specific) (Krizek, 2009; Wullschleger et al., 2006) and actin filament organization (P=0.0123; confident selected, LND-specific), while cultivars showed selection in genes associated with the Arp2/3 complex regulating actin (P=0.0302; core selected, CVR-specific) (Machesky et al., 1999; Mullins et al., 1998).

Overall, these data suggest that the *Andigenum* landrace and *Chilotanum* cultivar groups developed population-specific strategies for controlling stress response and plant development.

Potato cultivars showed *Chilotanum* (CVR)-specific enrichment for selection of genes controlling multiple levels of the shikimate pathway including activity of 3-dehydroquinate dehydratase, chorismate synthase, and shikimate 3-dehydrogenase enzymes (Table S3.4). The *Solanaceae* shikimate pathway is induced by wounding (Dyer et al., 1989) and generates aromatic amino acids including tryptophan (Herrmann, 1995), a relevant process as creation of a tryptophan metabolic sink alters phenylpropanoid pathway activity and susceptibility to *Phytophthora infestans* (Yao et al., 1995), the most economically destructive pathogen of potato worldwide (Haas et al., 2009). Supporting a sink status via tryptophan breakdown at the end of the shikimate pathway, the most highly enriched biological process selected in cultivars, was tryptophan catabolic process [to kynurenine] ($P=4.83x10^{-8}$; confident selected genes, cultivars). Potatoes were recently discovered to be a strong source of kynurenic acid (Turski et al., 2012), a tryptophan derivative with proven neuroprotective and anticonvulsant qualities, and possibly an anti-inflammatory/antiproliferative compound (Turski et al., 2009; Turski et al., 2012). These results point to possible indirect modification of the *Solanum* shikimate pathway resulting from historic cultivar selection for *P. infestans* resistance in the last two centuries since the Irish potato famine, possibly yielding unintended benefits to tuber nutritional value.

To report on gene functions most strongly associated with loci selected for potato

domestication, the intersection of CVR and LND-selected genes within the 494-gene core selection group were evaluated. These included 44 total genes, of which 33 contained functional InterPro protein domain matches and were not associated with transposable elements. Core domestication genes indicated functions related to circadian clock regulation and photomorphogenesis, cell growth and elongation, drought response, glycoalkaloid and terpene biosynthesis, Phytophthora resistance, and programmed cell death-related resistance (Table 3.3).

Conclusions

Resequencing and phylogenetic analysis of potato cultivated groups and their wild progenitors revealed a complex evolutionary history marked by hybridization and reintroduction of wild genetics into cultivated populations since the divergence of the tetraploid lineages from a diploid Peruvian base. Perhaps as a result, the genetic diversity contained within both South American landraces and current cultivars is strikingly high among major crop species, despite assertions of a narrow genetic base for North American germplasm. The combined cultivated lineages contain nearly as much genome diversity as a pool of 20 distinct species and a large portion of wild alleles, supporting the likelihood of ongoing wild species introgressions in regional potato groups. Analysis of selected gene functions separating *Andigenum* landraces and *Chilotanum*-derived cultivars from wild populations shows a limited set of core genes support potato domestication, and a larger number of genes are under selective pressure specific to either cultivated linage. Strong signatures of selection were identified for genes regulating glycoalkaloid biosynthesis and carbohydrate metabolism, key pathways differentiating cultivars from their wild progenitors.

Table 3.3. ***Functions of core selected potato genes shared by cultivars and landraces.***

Gene Id[a]	Locus	PGSC Functional Annotation	Arabidopsis Homolog[b]	Biological Process
1026044	chr01:69646627-69650467	Palmitoyl-protein thioesterase	alpha/beta hydrolase	lipid metabolism
0000138	chr01:71505970-71516942	JHL05D22.13 protein	*URT1* - UTP:RNA URIDYLYLTRANSFERASE 1	microRNA regulation
0000166	chr01:72061560-72069245	Palmitoyl-CoA hydrolase	acyl-CoA thioesterase	lipid metabolism
0000167	chr01:72070530-72074869	Small nuclear ribonucleoprotein-associated protein	*SMB* - SMALL NUCLEAR RIBONUCLEOPROTEIN ASSOCIATED PROTEIN B	mRNA splicing
0000189	chr01:72598427-72607003	MYBR transcription factor	*RVE6* - REVEILLE 6	plant circadian clock
0029176	chr03:9963931-9966386	Transcription factor	protein of unknown function (DUF607)	unknown
0003363	chr03:10241613-10249175	Nuclear matrix constituent protein 1	*CRWN1* - CROWDED NUCLEI 1	cell nucleus size control, heterochromatin organization
0012955	chr04:53700493-53708019	Receptor protein kinase	*SIRK1* - SUCROSE-INDUCED RECEPTOR KINASE 1	sucrose regulation of membrane aquaporins
0037817	chr06:9714992-9724950	DNA helicase domain-containing protein	P-loop containing nucleoside triphosphate hydrolase	unknown
0015738	chr06:12029604-12032257	Pentatricopeptide repeat protein	*RPF7* - RNA PROCESSING FACTOR 7	mitochrondrial nad2 mRNA processing
0041216	chr06:13651070-13652589	IDS4	*SPX2* - SPX DOMAIN GENE 2	phosphate nutrient stress
0028078	chr06:15614192-15619943	Metalloendopeptidase	*EGY2* - ETHYLENE-DEPENDENT GRAVITROPISM-DEFICIENT AND YELLOW-GREEN-LIKE 2	lipid metabolism, hypocotyl elongation, chloroplast development

Table 3.3 (cont'd)

2026767	chr06:15775836-15786259	Delta 1-pyrroline-5-carboxylate synthetase	*P5CS1* - DELTA1-PYRROLINE-5-CARBOXYLATE SYNTHASE 1	drought stress, proline biosynthesis
1026767	chr06:15775949-15784535	Delta 1-pyrroline-5-carboxylate synthetase	*P5CS1* - DELTA1-PYRROLINE-5-CARBOXYLATE SYNTHASE 1	drought stress, proline biosynthesis
0006739	chr06:16990722-17001511	Serine/threonine-protein phosphatase PP1 isozyme 3	*TOPP4* - TYPE ONE SERINE/THREONINE PROTEIN PHOSPHATASE 4	photomorphogensis via regulation of PIF5
0031213	chr06:17375964-17382524	Protein gar2	*PHIP1* - PHRAGMOPLASTIN INTERACTING PROTEIN 1	cytokinesis, cell plate formation
0031212	chr06:17389767-17394217	Nucellin-like aspartyl protease	eukaryotic aspartyl protease	programmed cell death, BAG6 cleavage
0002466	chr06:17893260-17894828	Pentatricopeptide repeat-containing protein	---	unknown
0002464	chr06:17894172-17898867	Pentatricopeptide repeat-containing protein	---	unknown
2016046	chr06:20815006-20825649	Scythe/bat3	*BAG6* - BCL-2 ASSOCIATED ANTHOGENE 6	programmed cell death mediated basal pathogen resistance
1016046	chr06:20828525-20831678	Scythe/bat3	*BAG6* - BCL-2 ASSOCIATED ANTHOGENE 6	programmed cell death mediated basal pathogen resistance
0032534	chr06:20951855-20953501	Early nodulin 75 protein	*PRP10* - PROLINE-RICH PROTEIN 10	positive regulator of germination
0028063	chr06:21004758-21010491	Coatomer	---	unknown

Table 3.3 (cont'd)

0006622	chr06:26387100-26395014	Phosphoinositide-specific phospholipase C	*PLC7 - PHOSPHOINOSITIDE PHOSPHOLIPASE C 7*	regulation of stomatal aperture, drought/salt stress response
0022088	chr10:20333031-20337415	Transketolase, chloroplastic	*TKL1 - TRANSKETOLASE 1*	photosynthesis, phenylpropanoid biosynthesis, isoprenoid biosynthesis
0020928	chr10:22941479-22943957	Beta-expansin	*EXPB4 - EXPANSIN B4*	cell growth, cell elongation
1020927	chr10:23002594-23004749	bHLH transcription factor	*PIF8 - PHYTOCHROME INTERACTING FACTOR 8/ UNE10 - UNFERTILIZED EMBRYO SAC 10*	circadian clock, photomorphogenesis, stem elongation
0004393	chr10:31087063-31093161	Kinase	*LECRK-IX.2 - L-TYPE LECTIN RECEPTOR KINASE IX.2*	Phytophthora resistance; innate immunity
1006223	chr10:42517914-42522413	Fucosyltransferase	*FUT12 - FUCOSYLTRANSFERASE 12, FUCT2, FUCTB*	cell wall biosynthesis, cell elongation
2006223	chr10:42534055-42537892	Fucosyltransferase	*FUT12 - FUCOSYLTRANSFERASE 12, FUCT2, FUCTB*	cell wall biosynthesis, cell elongation
0008243	chr10:43225482-43232321	sodium channel modifier 1-like isoform 2	---	sodium transport
0016905	chr10:45100729-45104700	Cembratrienol synthase 2b	*TPS6 - TERPENE SYNTHASE 6*	terpene biosynthesis, insect and pathogen defense
0003408	chr10:45381857-45389649	Squalene synthase	*SQS1 - SQUALENE SYNTHASE 1*	glycoalkaloid biosynthesis, insect and pathogen defense

Table 3.3 (cont'd)

[a] Unique Potato Genome Sequencing Consortium ID cod (trailing 'PGSC0003DMG40' in PGSC annotation).
[b] Arabidopsis protein homolog obtained from SpudDB website, capitalized names indicate experimentally characterized gene functions.

LITERATURE CITED

LITERATURE CITED

Ahn, C.S., Ahn, H.-K., and Pai, H.-S. (2014). Overexpression of the PP2A regulatory subunit Tap46 leads to enhanced plant growth through stimulation of the TOR signalling pathway. J. Exp. Bot., eru438.

Appeldoorn, N.J., de Bruijn, S.M., Koot-Gronsveld, E.A., Visser, R.G., Vreugdenhil, D., and van der Plas, L.H. (1997). Developmental changes of enzymes involved in conversion of sucrose to hexose-phosphate during early tuberisation of potato. Planta 202, 220-226.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. 23, 1514-1521.

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

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170.

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

Brush, S.B., Carney, H.J., and Humán, Z. (1981). Dynamics of Andean potato agriculture. Econ. Bot. 35, 70-88.

Cárdenas, P.D., Sonawane, P.D., Pollier, J., Bossche, R.V., Dewangan, V., Weithorn, E., Tal, L., Meir, S., Rogachev, I., and Malitsky, S. (2016). GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. Nat. Commun. 7, 1-16.

De Jong, H., and Rowe, P. (1971). Inbreeding in cultivated diploid potatoes. Potato Res. 14, 74-83.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet. 43, 491-498.

Dyer, W.E., Henstrand, J.M., Handa, A.K., and Herrmann, K.M. (1989). Wounding induces the first enzyme of the shikimate pathway in *Solanaceae*. Proc. Natl. Acad. Sci. U.S.A. 86, 7370-7373.

El Refy, A., Perazza, D., Zekraoui, L., Valay, J., Bechtold, N., Brown, S., Hülskamp, M., Herzog, M., and Bonneville, J.-M. (2004). The Arabidopsis KAKTUS gene encodes a HECT protein and controls the number of endoreduplication cycles. Mol. Genet. Genomics 270, 403-414.

Farrás, R., Ferrando, A., Jásik, J., Kleinow, T., Ökrész, L., Tiburcio, A., Salchert, K., del Pozo, C., Schell, J., Koncz, C. (2001). SKP1-SnRK protein kinase interactions mediate proteosomal binding of a plant SCF ubiquitin ligase. EMBO J. 20, 2742-2756.

Farré, E.M., Geigenberger, P., Willmitzer, L., and Trethewey, R.N. (2000). A possible role for pyrophosphate in the coordination of cytosolic and plastidial carbon metabolism within the potato tuber. Plant Physiol. 123, 681-688.

Fernie, A.R., and Willmitzer, L. (2001). Molecular and biochemical triggers of potato tuber development. Plant Physiol. 127, 1459-1465.

Friedman, M. (2006). Potato glycoalkaloids and metabolites: roles in the plant and in the diet. J. Agr. Food Chem. 54, 8655-8681.

Friedman, M., McDonald, G.M., and Filadelfi-Keszi, M. (1997). Potato glycoalkaloids: chemistry, analysis, safety, and plant physiology. Crc. Cr. Rev. Plant Sci. 16, 55-132.

Geigenberger, P., and Stitt, M. (2000). Diurnal changes in sucrose, nucleotides, starch synthesis and AGPS transcript in growing potato tubers that are suppressed by decreased expression of sucrose phosphate synthase. Plant J. 23, 795.

Geigenberger, P., Stitt, M., and Fernie, A. (2004). Metabolic control analysis and regulation of the conversion of sucrose to starch in growing potato tubers. Plant, Cell & Environment 27, 655-673.

Geigenberger, P., Hajirezaei, M., Geiger, M., Deiting, U., Sonnewald, U., and Stitt, M. (1998). Overexpression of pyrophosphatase leads to increased sucrose degradation and starch synthesis, increased activities of enzymes for sucrose-starch interconversions, and increased levels of nucleotides in growing potato tubers. Planta 205, 428-437.

Ginzberg, I., Tokuhisa, J.G., and Veilleux, R.E. (2009). Potato steroidal glycoalkaloids: biosynthesis and genetic manipulation. Potato Res. 52, 1-15.

Ginzberg, I., Thippeswamy, M., Fogelman, E., Demirel, U., Mweetwa, A.M., Tokuhisa, J., and Veilleux, R.E. (2012). Induction of potato steroidal glycoalkaloid biosynthetic pathway by overexpression of cDNA encoding primary metabolism HMG-CoA reductase and squalene synthase. Planta 235, 1341-1353.

Gopal, J., and Khurana, S. (2006). Handbook of potato production, improvement, and postharvest management. (Food Products Press).

Halford, N., Boulyz, J.-P., and Thomas, M. (2000). SNF1-related protein kinases (SnRKs)—Regulators at the heart of the control of carbon metabolism and partitioning. Adv. Bot. Res. 32, 405-434.

Halford, N.G., and Hardie, D.G. (1998). SNF1-related protein kinases: global regulators of carbon metabolism in plants? Plant Mol. Biol. 37, 735-748.

Halford, N.G., Hey, S., Jhurreea, D., Laurie, S., McKibbin, R.S., Paul, M., and Zhang, Y. (2003). Metabolic signalling and carbon partitioning: role of Snf1-related (SnRK1) protein kinase. J. Exp. Bot. 54, 467-475.

Hardigan, M.A., Bamberg, J., Buell, C.R., and Douches, D.S. (2015). Taxonomy and genetic differentiation among wild and cultivated germplasm of *Solanum* sect. *Petota*. Plant Genome 8.1, 388-405.

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., et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. Plant Cell 28, 388-405.

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

Herrmann, K.M. (1995). The shikimate pathway: early steps in the biosynthesis of aromatic compounds. Plant Cell 7, 907.

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R.E., Jansky, S., and Bethke, P. (2013). Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3 Genes Genom. Genet. 3, 1003-1013.

Hosaka, K. (2004). Evolutionary pathway of T-type chloroplast DNA in potato. Am. J. Potato Res. 81, 153-158.

Huber, S.C., and Huber, J.L. (1996). Role and regulation of sucrose-phosphate synthase in higher plants. Ann. Rev. Plant Biol. 47, 431-444.

Hudson, R.R., Slatkin, M., and Maddison, W. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583-589.

Hufford, M.B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., and Kaeppler, S.M. (2012). Comparative population genomics of maize domestication and improvement. Nature Genet. 44, 808-811.

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A., Shinde, B., Cardenas, P., Bocobza, S., Unger, T., Malitsky, S., and Finkers, R. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science 341, 175-179.

Jansen, G., Flamme, W., Schüler, K., and Vandrey, M. (2001). Tuber and starch quality of wild and cultivated potato species and cultivars. Potato Res. 44, 137-146.

Jansky, S., Dempewolf, H., Camadro, E., Simon, R., Zimnoch-Guzowska, E., Bisognin, D., and Bonierbale, M. (2013). A case for crop wild relative preservation and use in potato. Crop Sci. 53, 746-754.

Johns, T., and Alonso, J.G. (1990). Glycoalkaloid change during the domestication of the potato, *Solanum* Section *Petota*. Euphytica 50, 203-210.

Krizek, B.A. (2009). Making bigger plants: key regulators of final organ size. Curr. Op. Plant Biol. 12, 17-22.

Laurie, S., McKibbin, R.S., and Halford, N.G. (2003). Antisense SNF1-related (SnRK1) protein kinase gene represses transient activity of an α-amylase (α-Amy2) gene promoter in cultured wheat embryos. J. Exp. Bot. 54, 739-747.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Love, S.L. (1999). Founding clones, major contributing ancestors, and exotic progenitors of prominent North American potato cultivars. Am. J. Potato Res. 76, 263-272.

Machesky, L.M., Mullins, R.D., Higgs, H.N., Kaiser, D.A., Blanchoin, L., May, R.C., Hall, M.E., and Pollard, T.D. (1999). Scar, a WASp-related protein, activates nucleation of actin filaments by the Arp2/3 complex. Proc. Natl. Acad. Sci. U.S.A. 96, 3739-3744.

Maga, J.A., and Fitzpatrick, T.J. (1980). Potato glycoalkaloids. Crc. Cr. Rev. Food Sci. 12, 371-405.

Man, A.L., Purcell, P.C., Hannappel, U., and Halford, N.G. (1997). Potato SNF1-related protein kinase: molecular cloning, expression analysis and peptide kinase activity measurements. Plant Mol. Biol. 34, 31-43.

Marti, A., Wirbelauer, C., Scheffner, M., and Krek, W. (1999). Interaction between SCF^{SKP2} ubiquitin protein ligase and E2F-1 underlies regulation of E2F-1 degradation. Nat. Cell. Biol. 1, 14-19.

Mendoza, H., and Haynes, F. (1973). Some aspects of breeding and inbreeding in potatoes. Am. Potato J. 50, 216-222.

Mendoza, H., and Haynes, F. (1974). Genetic basis of heterosis for yield in the autotetraploid potato. Theor. Appl. Genet. 45, 21-25.

Mullins, R.D., Heuser, J.A., and Pollard, T.D. (1998). The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. Proc. Natl. Acad. Sci. U.S.A. 95, 6181-6186.

Ochoa, C.M. (1990). The potatoes of South America: Bolivia. (Cambridge University Press).

Ovchinnikova, A., Krylova, E., Gavrilenko, T., Smekalova, T., Zhuk, M., Knapp, S., and Spooner, D.M. (2011). Taxonomy of cultivated potatoes (*Solanum* section *Petota*: *Solanaceae*). Bot. J. Linn. Soc. 165, 107-155.

Pearsall, D.M. (2008). Plant domestication and the shift to agriculture in the Andes. In The handbook of South American archaeology (Springer), pp. 105-120.

Perrin, R.M., DeRocher, A.E., Bar-Peled, M., Zeng, W., Norambuena, L., Orellana, A., Raikhel, N.V., and Keegstra, K. (1999). Xyloglucan fucosyltransferase, an enzyme involved in plant cell ball biosynthesis. Science. 284, 1976-1979.

Plaisted, R., and Hoopes, R. (1989). The past record and future prospects for the use of exotic potato germplasm. Am. Potato J. 66, 603-627.

Purcell, P.C., Smith, A.M., and Halford, N.G. (1998). Antisense expression of a sucrose non-fermenting-1-related-protein kinase sequence in potato results in decreased expression of sucrose synthase in tubers and loss of sucrose-inducibility of sucrose synthase transcripts in leaves. Plant J. 14,195-202.

Rabinowitz, D., Linder, C., Ortega, R., Begazo, D., Murguia, H., Douches, D., and Quiros, C. (1990). High levels of interspecific hybridization between *Solanum sparsipilum* and *S. stenotomum* in experimental plots in the Andes. Am. Potato J. 67, 73-81.

Raker, C.M., and Spooner, D.M. (2002). Chilean tetraploid cultivated potato, is distinct from the Andean populations. Crop Sci. 42, 1451-1458.

Raj, A., Stephens, M., Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. 197, 573-589.

Salazar, M., and Estrada, D. (2008). CHILE-PERU: Preserving potatoes more important than age-old dispute. Retrieved November 1, 2016, from http://www.ipsnews.net.

Scott, G., and Suarez, V. (2012). The rise of Asia as the centre of global potato production and some implications for industry. Potato J. 39, 1-22.

Sinden, S., Sanford, L., and Webb, R. (1984). Genetic and environmental control of potato glycoalkaloids. Am. Potato J. 61, 141-156.

Sonnewald, U., Hajirezaei, M.-R., Kossmann, J., Heyer, A., Trethewey, R.N., and Willmitzer, L. (1997). Expression of a yeast invertase in the apoplast of potato tubers increases tuber size. Nat. Biotechnol. 15, 794-798.

Spooner, D.M. (2016) Species delimitations in plants: lessons learned from potato taxonomy by a practicing taxonomist. J. Syst. Evol. 15, 794-798.

Spooner, D.M., Ghislain, M., Simon, R., Jansky, S.H., and Gavrilenko, T. (2014) Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. Bot. Rev. 80, 283-383.

Spooner, D., Jansky, S., Clausen, A., del Rosario Herrera, M., and Ghislain, M. (2012). The Enigma of *Solanum maglia* in the Origin of the Chilean Cultivated Potato, *Solanum tuberosum Chilotanum* Group. Econ. Bot. 54, 191-203.

Spooner, D.M. (2009). DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. Am. J. Bot. 96, 1177-1189.

Spooner, D.M., and Salas, A. (2006). Structure, biosystematics, and genetic resources. Handbook of potato production, improvement, and postharvest management/J. Gopal, SM Paul Khurana, editors.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. U.S.A. 102, 14694-14699.

Sukhotu, T., and Hosaka, K. (2006) Origin and evolution of Andigena potatoes revealed by chloroplast and nuclear DNA markers. Genome. 49, 636-647.

Sugden, C., Donaghy, P.G., Halford, N.G., and Hardie, D.G. (1999). Two SNF1-related protein kinases from spinach leaf phosphorylate and inactivate 3-hydroxy-3-methylglutaryl-coenzyme A reductase, nitrate reductase, and sucrose phosphate synthase in vitro. Plant Physiol. 120, 257-274.

Turski, M.P., Turska, M., Zgrajka, W., Kuc, D., and Turski, W.A. (2009). Presence of kynurenic acid in food and honeybee products. Amino Acids 36, 75-80.

Turski, M.P., Kamiński, P., Zgrajka, W., Turska, M., and Turski, W.A. (2012). Potato-an important source of nutritional kynurenic acid. Plant foods for human nutrition 67, 17-23.

Ugent, D., Dillehay, T., Ramirez, C. (1987). Potato remains from a Late Pleistocene settlement in southcentral chile. Econ. Bot. 41, 17-27.

Vandenberg, R., Miller, J., Ugarte, M., Kardolus, J., Villand, J., Nienhuis, J., and Spooner, D. (1998). Collapse of morphological species in the wild potato *Solanum brevicaule* complex (*Solanaceae*: sect. *Petota*). Am. J. Bot. 85, 92-92.

Vanzin, G.F., Madson, M., Carpita, N.C., Raikhel, N.V., Keegstra, K., and Reiter, W.-D. (2001). The *mur2* mutant of *Arabidopsis thaliana* lacks fucosylated xyloglucan because of a lesion in fucosyltransferase AtFUT1. Proc. Natl. Acad. Sci. U.S.A. 99, 3340-3345.

Williams, L.E., Lemoine, R., and Sauer, N. (2000). Sugar transporters in higher plants–a diversity of roles and complex regulation. Trends Plant Sci. 5, 283-290.

Wullschleger, S., Loewith, R., and Hall, M.N. (2006). TOR signaling in growth and metabolism. Cell 124, 471-484.

Xu, X., Vreugdenhil, D., van Lammeren, A.A.M. (1998). Cell division and cell enlargement during potato tuber formation. J. Exp. Bot. 49, 573-582.

Yao, K., De Luca, V., and Brisson, N. (1995). Creation of a metabolic sink for tryptophan alters the phenylpropanoid pathway and the susceptibility of potato to *Phytophthora infestans*. Plant Cell 7, 1787-1799.

Zimmerer, K.S. (1998). The ecogeography of Andean potatoes. BioScience 48, 445-454.

CHAPTER 5

GENERAL CONCLUSIONS

Overview of Dissertation Research

Re-examination of the taxonomic relationships of wild and cultivated potato species in *Solanum* sect. *Petota* using a high density single nucleotide polymorphism (SNP) array to a core collection of 25 potato species supported existing taxonomic classifications of potato. In addition, several useful patterns were observed in the SNP phylogeny, including the ability to effectively distinguish species of the *Solanum brevicaule* species complex, which have historically challenged attempts at classification due to hybridization and similar morphological features (Vandenberg, et al., 1998). The phylogeny also placed the proposed wild species progenitors of cultivated potatoes from the northern distribution of the *S. brevicaule* complex, *S. candolleanum* and *S. bukasovii*, in the same clade as cultivated genotypes, supporting the theory of potato's Peruvian domestication (Spooner et al., 2005).

Comparison of SNP allele frequencies between wild species and cultivars identified a number of carbohydrate pathway genes with divergent allelic composition between the two populations, providing new gene candidates for selection that may have undergone changes during domestication to increase mobilization of sucrose and starch synthesis in developing tubers. One gene, KAKTUS, controlling endoreduplication in plant cells, was found to contain nearly fixed allele differences between the wild and cultivar populations. Endoreduplication is a process in which cells replicate their nuclear genome without undergoing cytokinesis (El Refy et al., 2004; Sugimoto-Shirasu and Roberts, 2003) and is associated with increased in cell size and establishment of a nutrient sink status or specialized metabolic activity (Chen and Setter, 2003; Larkins et al., 2001). These findings show that

increasing cell size or nutrient sink status by altered cell cycle regulation may also have been a pathway selected in potato domestication.

Large-scale structural variation has been shown to impact the genomes of tetraploid potato cultivars, with regions greater than 100 kb found to be absent on 1-3 homologous chromosomes of elite varieties (Iovene et al., 2013), a form of copy number variation (CNV). To further investigate whether structural variation constitutes a significant component of deleterious mutation load inherited by cultivars from their progenitors, the full extent of copy number variation (CNV) and SNP variation was evaluated in a "monoploid panel" of homozygous monoploids and doubled monoploids derived from primitive diploid landrace potatoes (*S. tuberosum Andigenum* group, *Phureja* diploids). Structural variation was shown to be a key factor in the overall genome variation of tetraploid cultivars' diploid progenitors, with 30% of the potato genome affected by CNV in a panel of only 12 clones, demonstrating the significant structural heterogeneity of cultivated potato populations. Large structural variants (>100 kb, up to 575 kb) were found to exist in the monoploid panel, confirming these mutations are not exclusively a result of potato polyploidy. The majority of CNVs were smaller with median sizes of 2.5 kb (deletions) and 3.8 kb (duplications). Copy number variants demonstrated several patterns in connection with gene function and activity. Gene deletion and duplication levels were found to be directly related to evolutionary age and conservation, with genes conserved in all angiosperms reflecting low levels of duplication or deletion, versus higher levels for genes arising within *Solanum* and nearly 50% of potato lineage-specific genes. Deletion was preferentially associated with lowly expressed genes, while duplication was preferentially associated with genes responsive to hormone and abiotic

stress treatment. Analysis of CNV-enriched gene clusters showed duplication was directly related to several tandem duplicated gene families related to stress response, including resistance genes, small auxin-up RNAs (SUARs) (Wu et al., 2012) and methylketone synthase I (MKS1) (Antonius et al., 2001; Fridman et al., 2005). These results collectively demonstrate that structural variants play an important role alongside sequence variation in contributing to genome heterogeneity and deleterious load in potato cultivars, and were a factor pre-dating the advent of autopolyploidy in the cultivated lineage. Beyond the potato mutation load, CNVs and duplications in particular also demonstrated a major contribution to the expansion of gene families functioning in stress tolerance and environmental response.

A genome re-sequencing study was performed for 23 potato cultivars, 20 primitive landraces, and 20 wild diploid species from South American, designed to ascertain the genome diversity and gene targets of domestication in cultivated potatoes relative to their wild species progenitors. Phylogenetic analysis of the diversity panel yielded further support for the Peruvian domestication origin of cultivated potato (Spooner et al., 2005). Cultivated potatoes demonstrated striking levels of genome sequence diversity within conserved regions, showing higher sequence diversity than any crop re-sequencing study to date at both the diploid and tetraploid level. Tetraploid cultivars are highly heterozygous (on average 2.73% nucleotide sites, maximum 3%) but also relatively similar to one another compared with landraces and species, indicating that breeding has re-partitioned the diversity among potato populations as intra-genomic allele variation within individual clones, reflecting the limited population structure of North American elite varieties and reliance on heterosis (Mendoza and Haynes, 1973; Mendoza and Haynes, 1974). A high level of allele sharing between the wild and

cultivated populations, with over 50% of wild alleles present in landraces or cultivars, suggests frequent historic hybridization between wild potatoes and populations of cultivated landraces contributed to the substantial genetic diversity in *Solanum tuberosum*. Comparing allele frequencies in the cultivar, landrace and wild species populations identified genes putatively under selection in differentiating the cultivated groups from their wild counterparts. These included genes functioning in steroidal glycoalkaloid biosynthesis and carbohydrate metabolism, particularly sucrose mobilization and starch biosynthesis. Although a core set of "domestication genes" was uncovered in the intersection between cultivar- and landrace-selected loci, a greater number of selected "improvement" genes were found to be specific to the *Andigenum* landrace and *Chilotanum* cultivar lineages, demonstrating cultivated potato groups from different hemispheres have achieved genetic improvement by independent strategies.

The collective findings of these studies offer an updated view of the landscape of genetic diversity in cultivated potatoes and their wild relatives to assist efforts at improvement through breeding and biotechnology in the post-genomics era. Understanding the broad extent of sequence and structural variation within cultivated populations, and key loci associated with their agricultural performance will hopefully inform breeders in their efforts toward broadening the genetic base of breeding populations, eliminating harmful deleterious mutations through diploid breeding, utilizing wild crop relatives to enhance adaptability, and improving existing varieties through genome editing of loci associated with potato quality and yield QTLs.

APPENDIX

Figure S1.1. ***Consensus tree based on 1001 bootstrapped datasets. The topology of this tree was used to generate the distance-based phylogenetic tree (Figure 2).*** *Branches display bootstrap values showing number of times each grouping appeared across the 1000 replicates.*

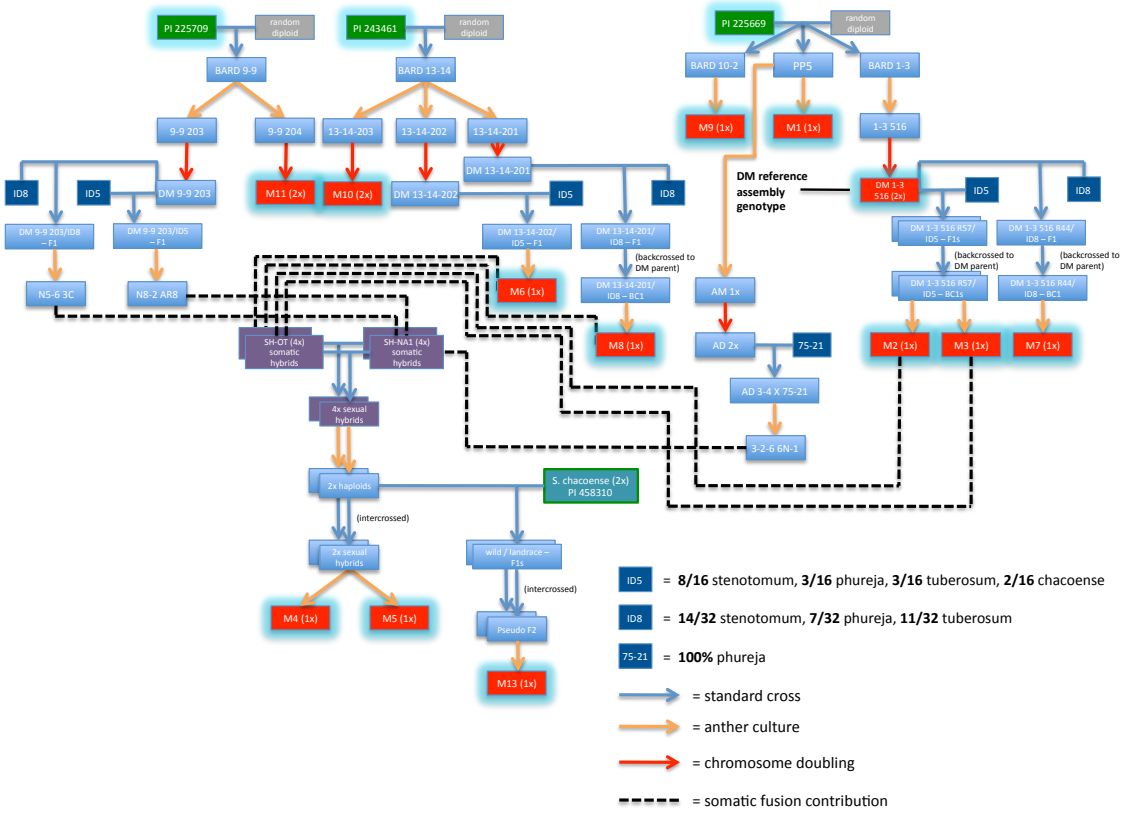

Figure S2.1. ***Pedigree information for the monoploid panel clones.***

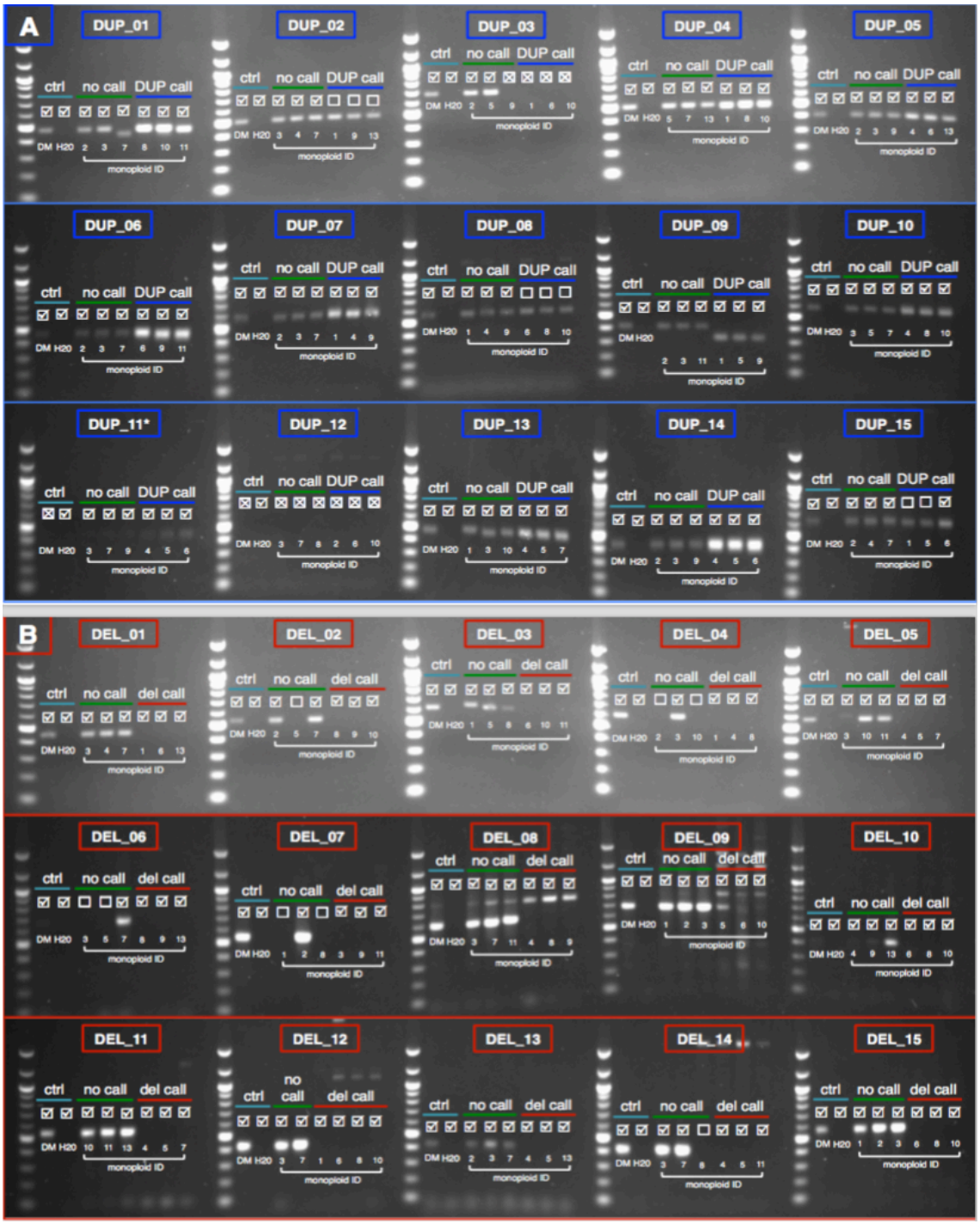

Figure S2.2. ***Experimental PCR validation of 15 randomly selected duplication and deletion loci.***

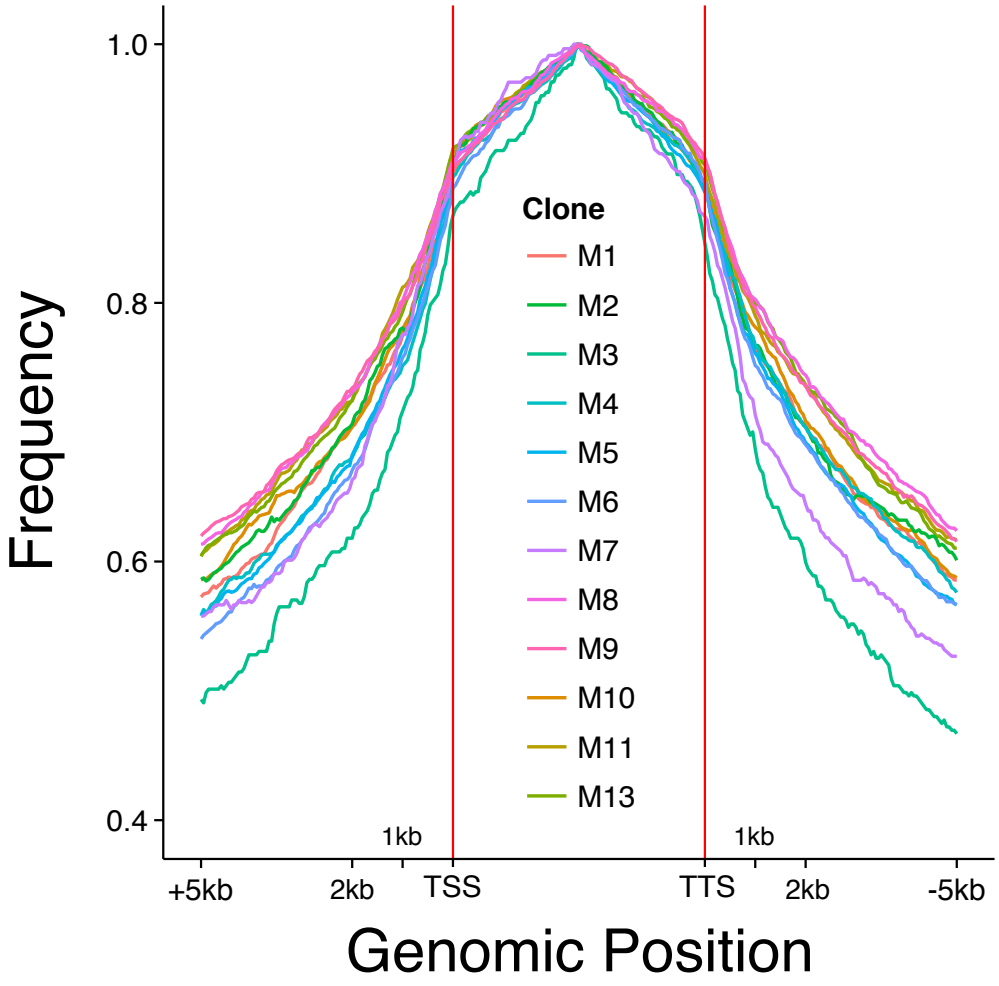

Figure S2.3. ***Distribution of copy number variation frequency (per clone) relative to the position of all genes impacted by deletion.***

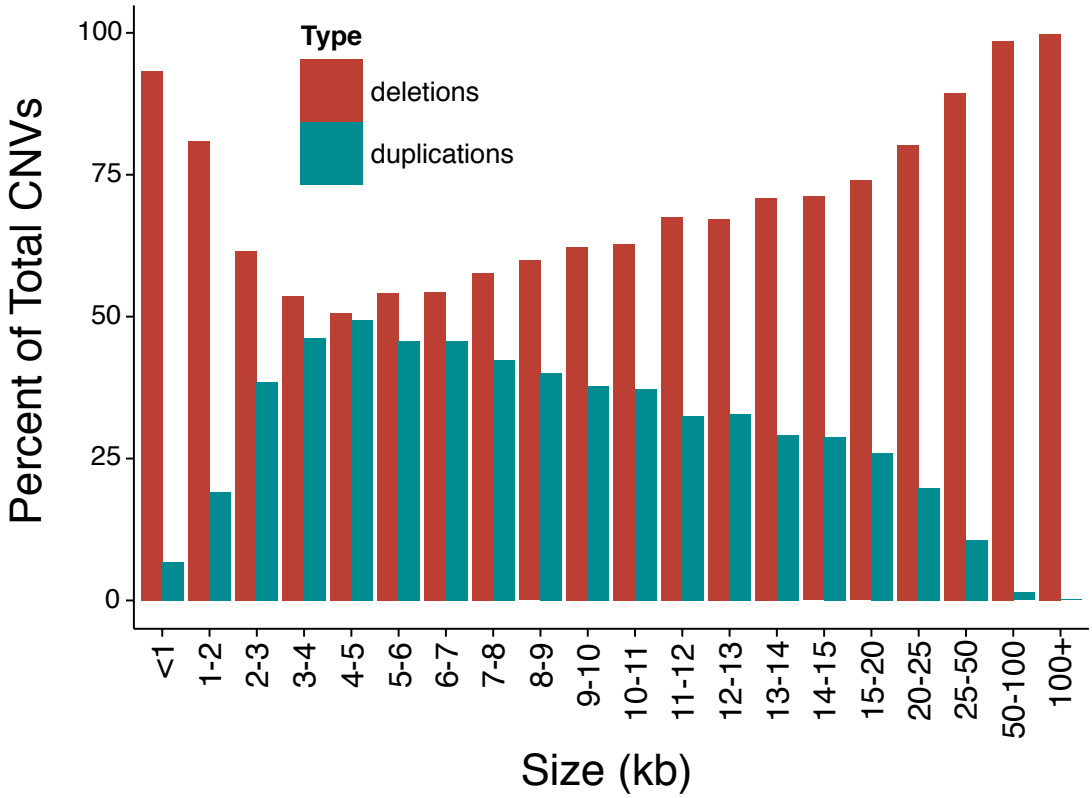

Figure S2.4. ***Fraction of copy number variants represented by duplication and deletion binned by size.***

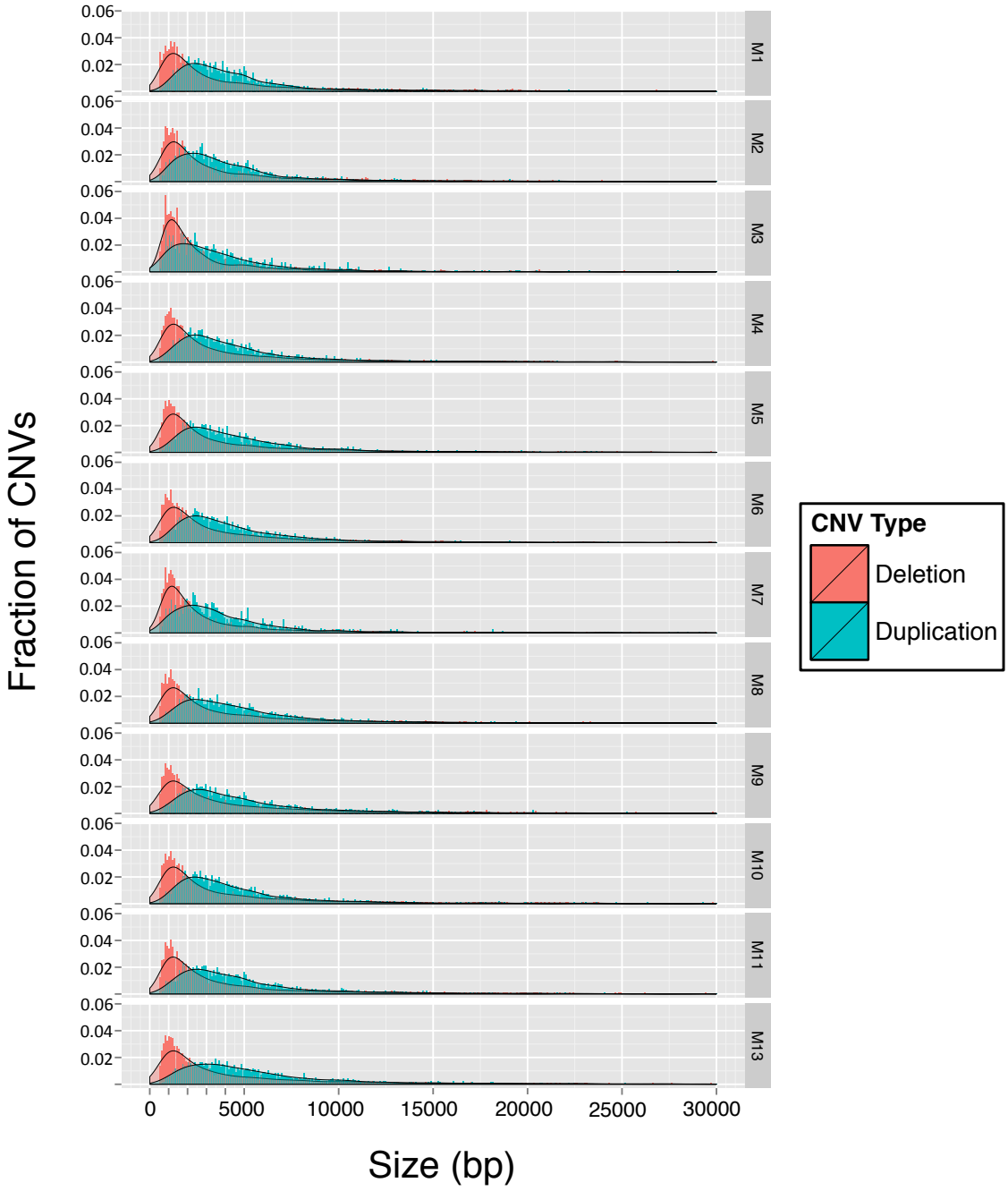

Figure S2.5. ***Copy number variation size distribution by clone.***

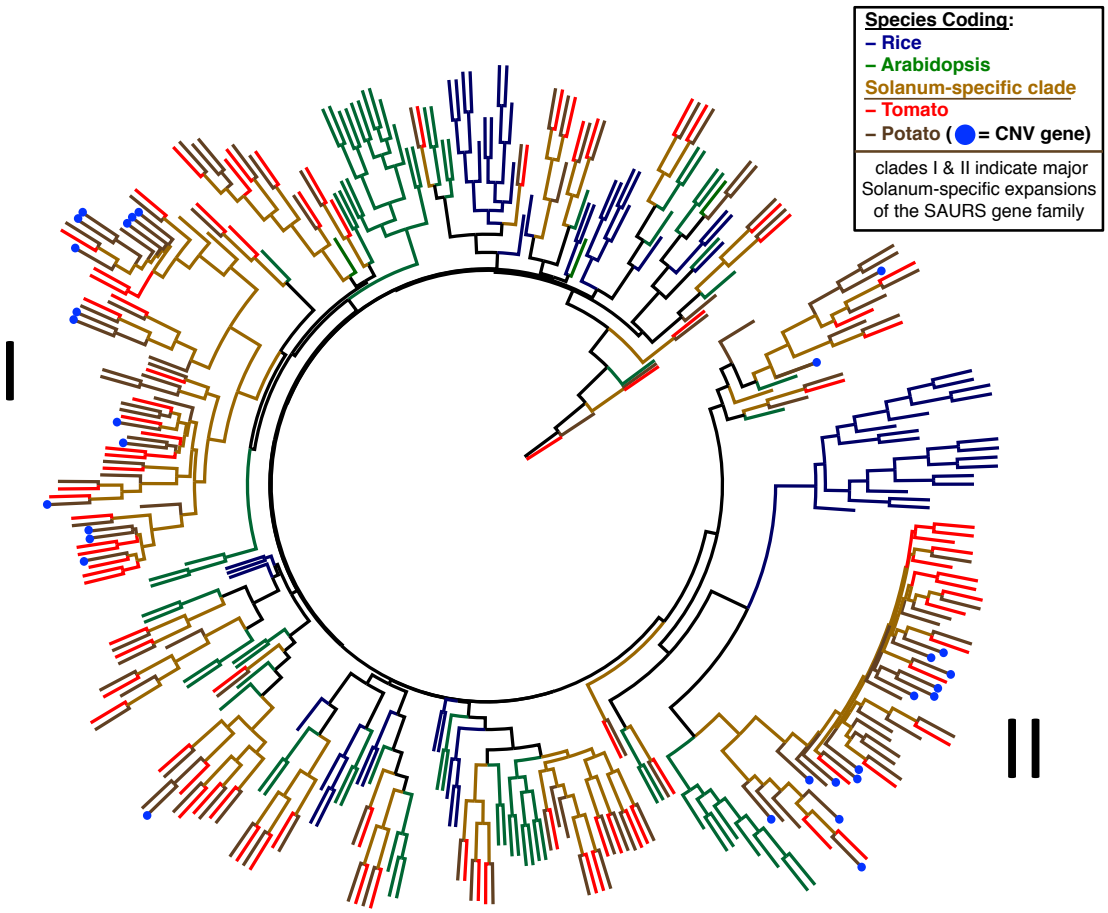

Figure S2.6. ***Phylogenetic tree based on protein alignment of annotated small auxin up-regulated RNA (SAUR) genes from rice, Arabidopsis, tomato, and potato proteomes.*** *Clades I and II indicate major Solanum-specific expansion of SAURs.*

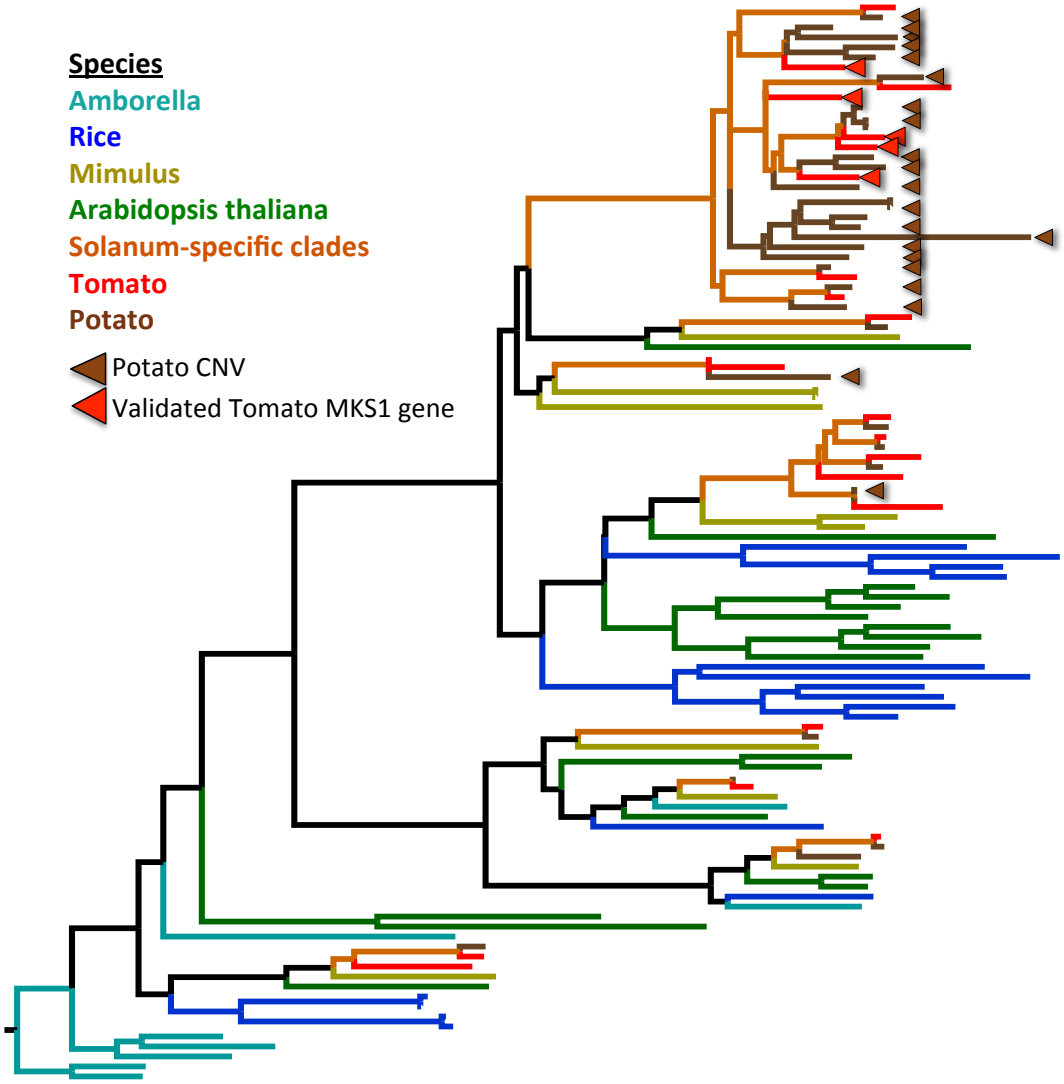

Figure S2.7. ***Phylogenetic tree based on protein alignment of genes with sequence homology to five tomato methylketone synthase 1 (MKS1) genes from Amborella, rice, Arabidopsis, Mimulus guttatus, tomato, and potato.***

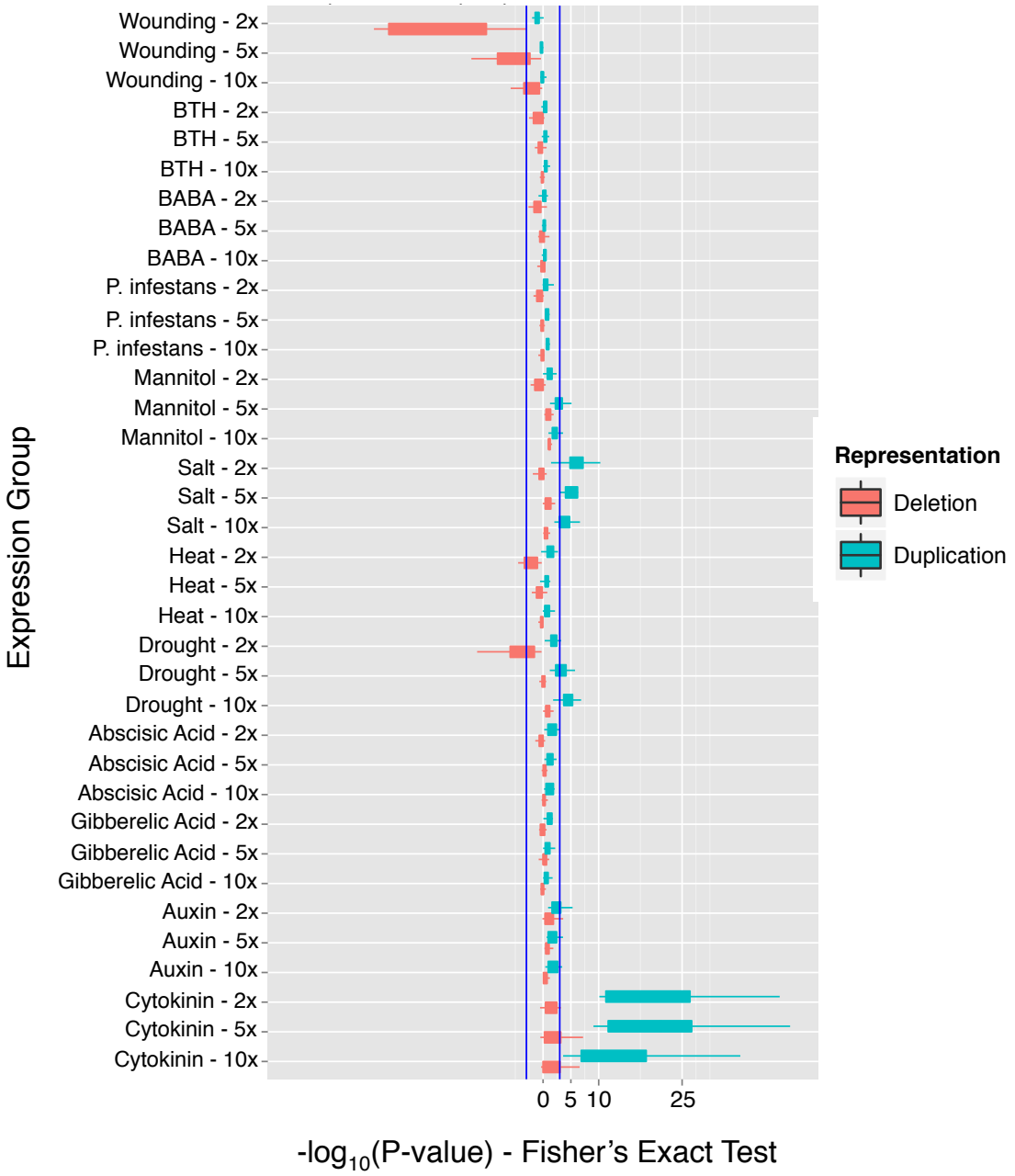

Figure S2.8. ***Box plot of copy number variation enrichment for individual stress and hormone response expression classes.***

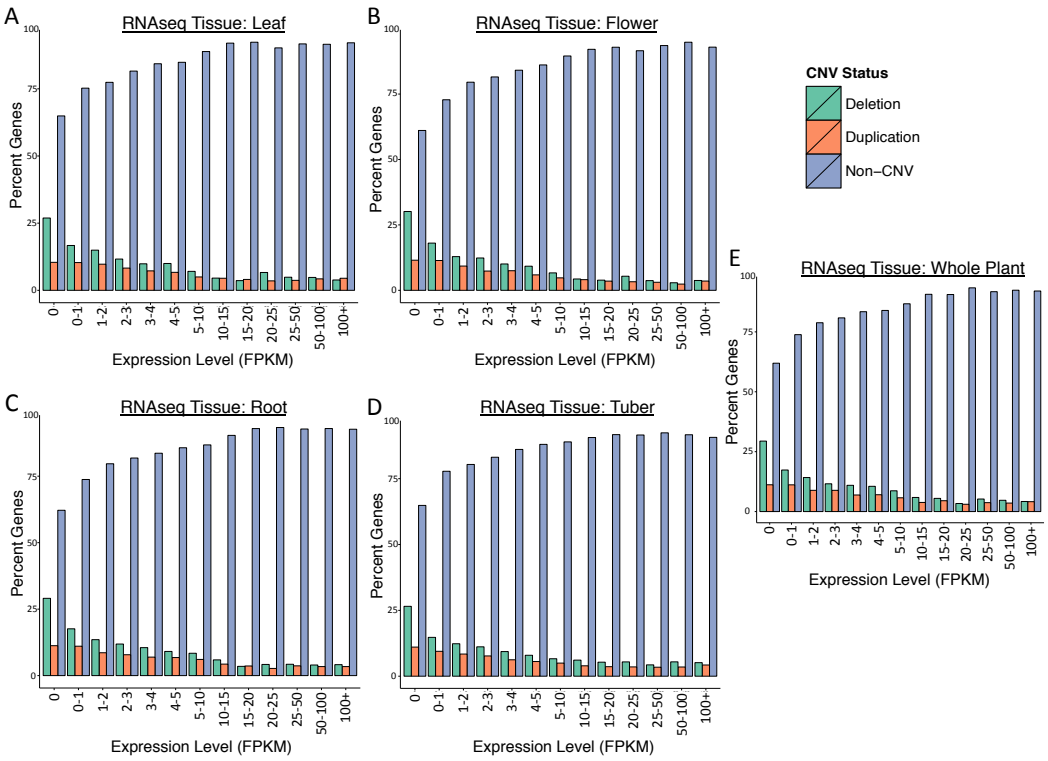

Figure S2.9. ***Summary of copy number variation rates in genes with different expression levels based on fragments per kilobase per million mapped reads values from leaf, flower, root, tuber, and whole in vitro plant tissues.***

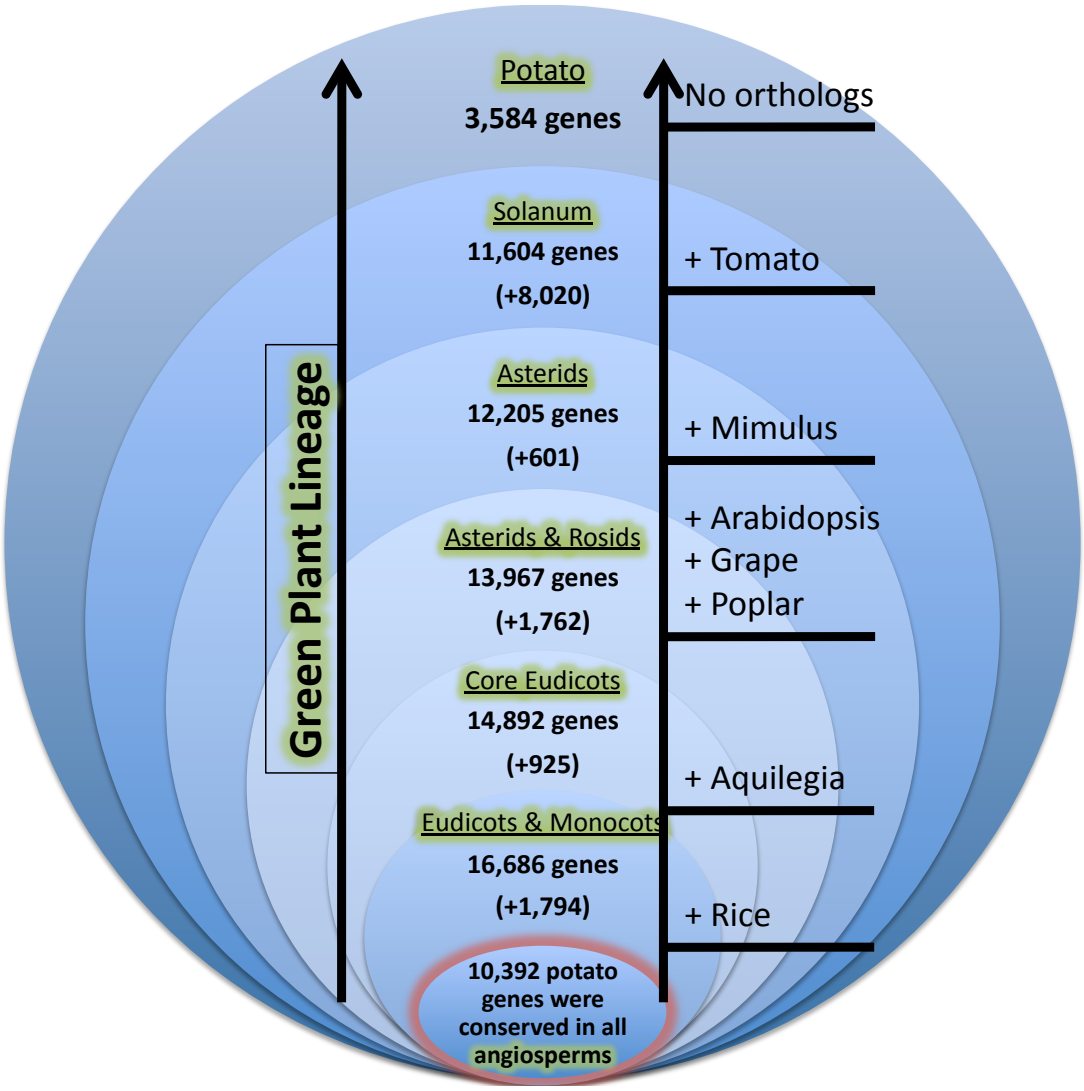

Figure S2.10. ***Overview of potato gene lineage categories generated based on orthologous gene clustering.***

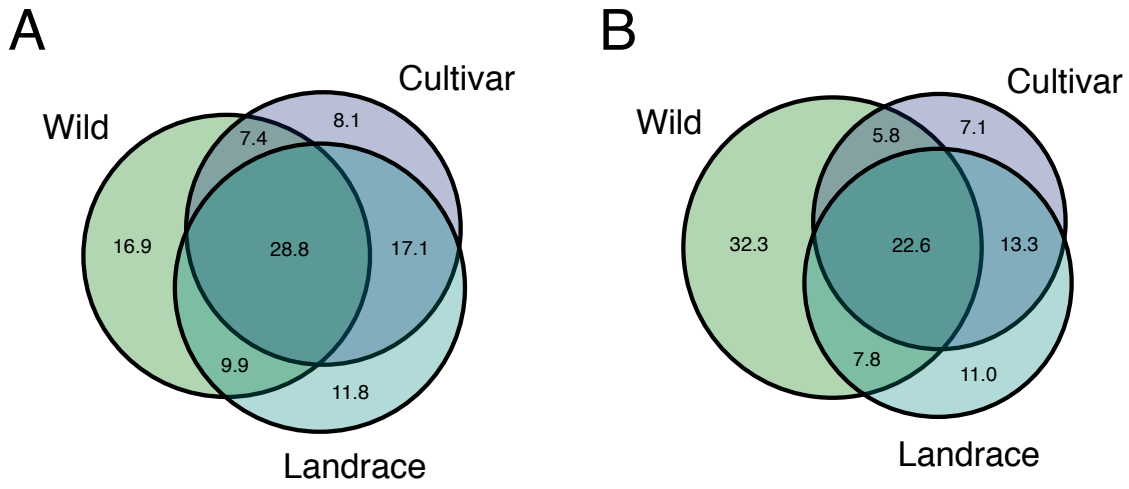

Figure S3.1. ***Venn diagrams of the fraction of identified SNP alleles shared between cultivars (blue), landraces (turquoise) and wild species (green).***

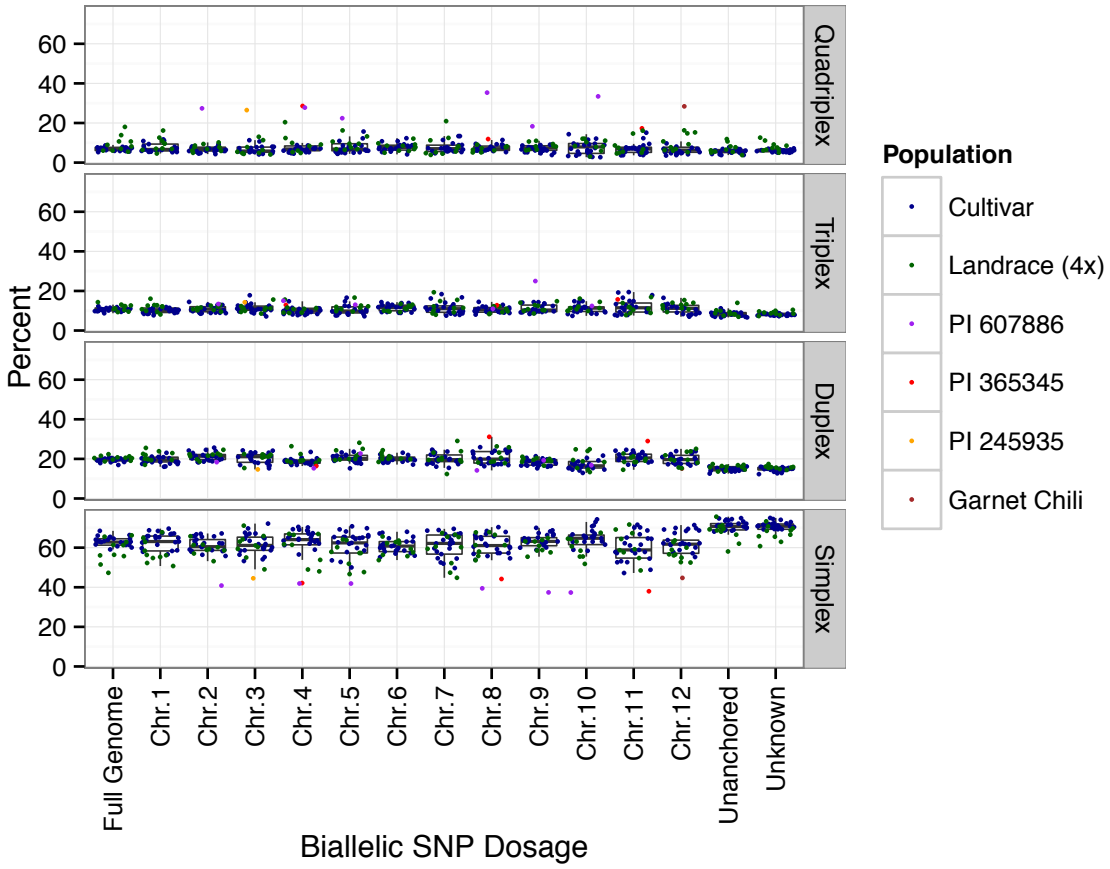

Figure S3.2. ***Alternate (non-reference) allele dosages for biallelic SNP sites in tetraploid landrace (green) and cultivar (blue) potato genotypes.***

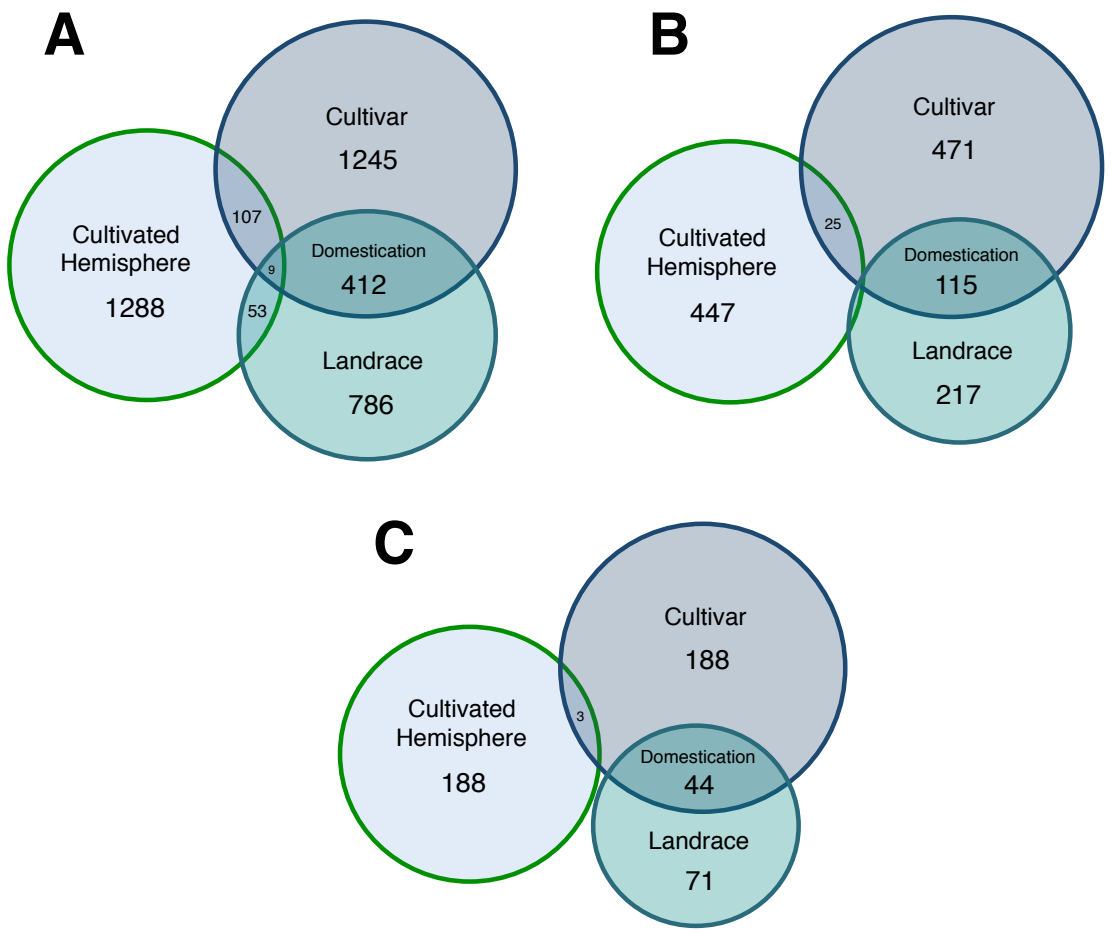

Figure S3.3. ***Venn diagrams showing the fraction of selected gene candidates shared by the three cross-population selection comparisons, including cultivar (CVR) selected (dark blue), landrace (LND) selected (turquoise), and cultivated hemisphere (HEM) selected (light blue, green border).***

LITERATURE CITED

LITERATURE CITED

Antonious, G.F. (2001). Production and quantification of methyl ketones in wild tomato accessions. J. Environ. Sci. Heal. B 36, 835-848.

Chen, C.T., and Setter, T.L. (2003). Response of potato tuber cell division and growth to shade and elevated CO2. Ann. Bot. 91, 373-381.

El Refy, A., Perazza, D., Zekraoui, L., Valay, J., Bechtold, N., Brown, S., Hülskamp, M., Herzog, M., and Bonneville, J.-M. (2004). The Arabidopsis KAKTUS gene encodes a HECT protein and controls the number of endoreduplication cycles. Mol. Genet. Genomics 270, 403-414.

Fridman, E., Wang, J., Iijima, Y., Froehlich, J.E., Gang, D.R., Ohlrogge, J., and Pichersky, E. (2005). Metabolic, genomic, and biochemical analyses of glandular trichomes from the wild tomato species *Lycopersicon hirsutum* identify a key enzyme in the biosynthesis of methylketones. Plant Cell 17, 1252-1267.

Iovene, M., Zhang, T., Lou, Q., Buell, C.R., and Jiang, J. (2013). Copy number variation in potato - an asexually propagated autotetraploid species. Plant J. 10, 1915-1925.

Larkins, B.A., Dilkes, B.P., Dante, R.A., Coelho, C.M., Woo, Y.M., and Liu, Y. (2001). Investigating the hows and whys of DNA endoreduplication. J. Exp. Bot. 52, 183-192.

Mendoza, H., and Haynes, F. (1973). Some aspects of breeding and inbreeding in potatoes. Am. Potato J. 50, 216-222.

Mendoza, H., and Haynes, F. (1974). Genetic relationship among potato cultivars grown in the United States. HortSci.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. U. S. A. 102, 14694-14699.

Sugimoto-Shirasu, K., and Roberts, K. (2003). "Big it up": endoreduplication and cell-size control in plants. Curr. Opin. Plant Biol. 6, 544-553.

Vandenberg, R., Miller, J., Ugarte, M., Kardolus, J., Villand, J., Nienhuis, J., and Spooner, D. (1998). Collapse of morphological species in the wild potato *Solanum brevicaule* complex (Solanaceae: sect. Petota). Am. J. Bot. 85, 92-109.

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