MISS GINA MAI PHAM (Orcid ID : 0000-0002-8058-7862)

Article type : Original Article

Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato

Gina M. Pham[*], Linsey Newton[*], Krystle Wiegert-Rininger[*], Brieanne Vaillancourt[*], David S. Douches[§], C. Robin Buell[*]

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

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

Corresponding author: C. Robin Buell

612 Wilson Road, Room 166

East Lansing, MI 48824-1312

(517)-353-5597

buell@msu.edu

Author email addresses

Gina M. Pham (phamgina@msu.edu), Linsey Newton (newtonl2@msu.edu), Krystle Wiegert-Rininger (wiegertk@msu.edu), Brieanne Vaillancourt (vaillan6@msu.edu), David S. Douches (douchesd@msu.edu), C. Robin Buell (buell@msu.edu)

Running title: Preferential allele and CNV expression in potato

Key words: copy number variation, preferential allele expression, gene expression, *Solanum tuberosum*, genetic variation

SUMMARY

Relative to homozygous diploids, the presence of multiple homologs or homeologs in polyploids affords greater tolerance to mutations that can impact genome evolution. In this study, we describe sequence and structural variation in the genomes of six accessions of cultivated potato (*Solanum tuberosum* L.), a vegetatively propagated autotetraploid, and their impact on the transcriptome. Sequence diversity was high with a mean SNP rate of approximately 1 per 50 bases suggestive of high levels of allelic diversity. Additive gene expression was observed in leaves (3,605 genes) and tubers (6,156 genes) that contrasted the preferential allele expression of between 2,180 and 3,502 and 3,367 and 5,270 genes in the leaf and tuber transcriptome, respectively. Preferential allele expression was significantly associated with evolutionarily conserved genes suggesting selection of specific alleles of genes responsible for biological processes common to angiosperms during the breeding selection process. Copy number variation was rampant with between 16,098 and 18,921 genes in each cultivar exhibiting duplication or deletion. Copy number variable genes tended to be evolutionarily recent, lowly expressed, and enriched in genes that show increased expression in response to biotic and abiotic stress treatments suggestive of a role in adaptation. Gene copy number impacts on gene expression were detected with 528 genes having correlations between copy number and gene expression. Collectively, these data suggest that in addition to allelic variation of coding sequence, the heterogenous nature of the tetraploid potato genome contributes to a highly dynamic transcriptome impacted by allele preferential and copy number-dependent expression effects.

INTRODUCTION

Whole genome duplications have been inferred in the evolutionary history of plants and animals. In angiosperms, it is estimated that 60 – 70% of species have experienced ancestral polyploidy through allopolyploidy, autopolyploidy, or a combination of both mechanisms (Van de Peer *et al.* 2009). It is hypothesized that polyploidization events confer evolutionary advantages through improved adaptability to changing environments, as the additional genetic material generated by polyploidization provides greater tolerance of mutations. This can lead to increased rates of evolution, with paralogs undergoing neofunctionalization and subfunctionalization. Additionally, heterotic effects may result from the heterozygosity introduced during polyploidization and subsequent sequence variation that occurs over evolutionary time. These genetic and phenotypic changes contribute to a rapid potential for adaptation and speciation. Indeed, an estimated 15% of angiosperm speciation events are accompanied by increases in ploidy level (Wood *et al.* 2009). Although polyploidy has been associated with improved survival and speciation, it is not clear if it is responsible for increased species diversification.

Cultivated potato, *Solanum tuberosum* L., is an autopolyploid, meaning that polyploidy may have arisen from a relatively recent whole-genome duplication event or hybridization of two individuals of the same species. As a consequence of autopolyploidy, homologous chromosomes can be observed forming multivalent groups during meiosis I (Swaminathan 1954). Understanding the mechanism by which intra-genome variation contributes to transcriptomic variation in potato would improve our understanding of how genomic diversity contributes to phenotypic diversity. To date, there are several genome datasets available for potato, including a high-quality assembly of a doubled monoploid clone of *S. tuberosum* Group Phureja*,* DM1-3 516R44 (DM 1-3) (Potato Genome Sequencing Consortium 2011) and a gene expression atlas of DM 1-3 that includes 32 developmental and stress conditions (Massa *et al.* 2013, Massa *et al.* 2011). Whole-genome shotgun sequencing and targeted bacterial artificial chromosome (BAC) sequencing of *S. tuberosum* Group Tuberosum RH89-039-16 (RH), a heterozygous diploid breeding line, revealed that the two haplotypes within RH are more distant from one another than each was relative to the DM 1-3 reference genome (Potato Genome Sequencing Consortium 2011) and that large regions of chromosome 5 show loss of collinearity between homologous chromosomes, highlighting the high genome heterogeneity within one set of homologous chromosomes (de Boer *et al.* 2015). A study that examined the genome heterogeneity in a panel of 12 monoploid/doubled monoploid clones derived from diploid *S. tuberosum* accessions (Hardigan *et al.* 2016a) revealed that the diploid potato genome is highly heterogeneous with rampant copy number variation that affected 30% of the annotated gene complement in this relatively small diversity panel.

In cultivated tetraploid potato, both a high degree of heterozygosity and copy number variation have been reported (Hirsch *et al.* 2013, Iovene *et al.* 2013). Using an 8303 Infinium SNP array, 56% of assayed sites were heterozygous in a panel of ~250 cultivated tetraploids (Hirsch *et al.* 2013). With respect to copy number variation in tetraploid potato, CNVs spanning more than 100 kilobases have been observed that result in copy number effects in gene expression that scale linearly with copy number for a small subset of genes analyzed in CNV regions via quantitative PCR (Iovene *et al.* 2013). This variation has the potential to contribute to between-cultivar phenotypic variation as evidenced in multiple *Glycine* polyploid species which demonstrate that gene copy number is particularly important in genes that produce subunits in complexes (Coate *et al.* 2016).

Potato is known for high genetic load and severe inbreeding depression when selfed. Thus, in contrast to sexually reproducing polyploid species such as allo-hexaploid wheat, there are few opportunities to purge deleterious duplications and deletions in cultivated potato as it is clonally propagated and selection of elite cultivars occurs at the F1 level with no backcrossing or selfing due to the desire to maintain maximal heterozygosity (Chase 1963). Thus, deleterious CNVs may be maintained in potato if they complement deletions elsewhere in the genome or if the effect of retaining deleterious CNVs is not agronomically relevant. Autopolyploidy may also provide a buffering effect for deleterious mutations in the form of SNPs and insertion-deletion events (indels); as a consequence, deleterious mutations accumulate and contribute to genetic load.

Copy number variation can affect gene expression (Cook *et al.* 2012, Iovene *et al.* 2013) and potentially alter phenotype, especially that of adaptative traits. For example, CNV has been shown to confer resistance to nematodes in soybean (Cook *et al.* 2012), tolerance to aluminum in maize (Maron *et al.* 2013), and submergence tolerance in rice (Hattori *et al.* 2009, Xu *et al.* 2006). Additionally, high heterozygosity in the form of SNPs and indels can result in differential expression of alleles. Preferential allele expression (PAE) has been detected in F1 hybrids of *Arabidopsis thaliana*, *Cirsium arvense* (creeping thistle), *Hordeum vulgare* (barley), *Oryza sativa* (rice) and *Zea mays* (maize) (Bell *et al.* 2013, Song *et al.* 2013, Springer and Stupar 2007, von Korff *et al.* 2009, Zhang and Borevitz 2009). In maize, *cis*-regulatory variation is responsible for ~70% of PAE in reciprocal hybrids in which ~50% of the genes assayed exhibited biased allelic expression (Springer and Stupar 2007). The extent of PAE and mode of action was observed in different F1 hybrids regardless of heterotic group with *cis*-acting gene regulation being the dominant mechanism of PAE in all cases. This data suggests that variation linked to PAE sites, such as local chromatin structure or promoter sequence variation, is the primary driver of imbalanced allelic expression in maize. Additionally, Guo *et al.* (2004) demonstrated that there is an environment-specific effect on PAE. Preferential allele expression has been recently shown in many animal systems including mouse and human in context of epigenetic regulation (Buckberry *et al.* 2012, Buckland 2004, Crowley *et al.* 2015, Gregg *et al.* 2010). In mice, PAE may be regulated via the proximity of a gene to DNase I hypersensitive sites (Hasin-Brumshtein *et al.* 2014).

In this study, we focus on genomic and transcriptomic variation within six commercial, elite tetraploid cultivars of *S. tuberosum*. Using a combination of whole genome DNA and RNA sequencing (RNA-seq), the genomic diversity of four round white chip processing potatoes (Atlantic, Kalkaska, Missaukee, and Snowden), a round white fresh market potato (Superior), and a fresh market russet potato (Russet Norkotah) was characterized relative to the reference potato genome (Xu *et al.* 2011). Variation in the form of SNPs, indels, and CNV was determined on a genome-wide scale in each cultivar that revealed preferential allele expression and the effects of copy number on transcript abundance. Functional enrichment, membership in orthologous groups across the angiosperms, and developmental and stress expression patterns were used to infer the function of genes with preferential allele expression or copy number variation. This genome-scale assessment of genome heterogeneity in tetraploid potato revealed a high genetic load and potential mechanisms for phenotypic variation through intra-genome variation in gene expression due to copy number-dependent and preferential allele expression.

RESULTS AND DISCUSSION

Polymorphism in autotetraploid potato

Genomic DNA from six potato cultivars was sequenced in the 100 nucleotide paired-end mode on the Illumina HiSeq 2000 and aligned to the DM1-3 reference potato genome version 4.04 (Hardigan *et al.* 2016a), yielding coverage depths ranging from 47X to 62X.

Approximately 390 million sites in the potato genome passed the quality filtering criteria for variant detection in each cultivar (Figure 1a; less confident regions of the genome were excluded due to high heterozygosity or the inability to align reads uniquely due to repetitive sequence. Of the callable sites, approximately 8.4 million SNPs and 320,000 to 424,000 indels were identified in each cultivar (Figure 1b). To confirm the accuracy of our whole genome resequencing read alignment-based genotype calls, we examined the concordance of allele and dosage calls (see below) with genotypes obtained using the 8303 Potato Infinium SNP Array (Hirsch *et al.* (2013). In total, between 1,600 and 1,900 sites overlapped between the two datasets; of these, between 94.3% and 96.3% of the sites were concordant in not only allele calls but also at the dosage level (Table S1) demonstrating that our computational pipeline used to identify allelic diversity using read-alignments and variant calling in tetraploid potato is robust.

SNPs were annotated with respect to allelic variants (Figure S1). The most abundant SNPs discovered were Type B SNPs, which are bi-allelic sites with at least one reference allele and at least one alternate allele; between 5.5 and 6 million SNPs identified were classified as Type B SNPs in each cultivar. Type A SNPs, which are homozygous for a non-reference allele were the next most abundant class of SNPs, with 2.2 to 2.6 million identified per cultivar. Type C and D SNPs, which consist of three alleles per site, were far less abundant (205,324 – 289,670 loci), and Type E SNPs, which consist of 4 alleles per site, were even less abundant (939 – 1,972 loci). Overall, biallelic loci represented 66-71% of all surveyed sites; in total, one SNP per 46 nucleotides in the ~390 Mb of the genome assayed was observed. Dosage of biallelic SNP loci was examined using the alternate allele as the dominant marker yielding simplex (AAAB), duplex (AABB), and triplex (ABBB) with type A SNPs or all reference allele calls with no invariant sites characterized as nulliplex (AAAA or BBBB) sites (Figure S2). The combined number of simplex and triplex sites was substantially higher in all six cultivars compared to the number of duplex sites, consistent with the nature of potato breeding in which F1 progeny are selected with no backcrossing or additional selection that can amplify copy numbers of specific alleles (Hirsch *et al.* 2013).

Indels showed a similar distribution of genotypes to SNPs, although there were substantially less indels detected than SNPs (Figure 1b, Figure S1). We annotated indels similar to the classification used for SNPs. Relative to the DM 1-3 reference sequence, nearly half of the indels were homozygous across all four alleles (Type A). Mirroring the observations with SNPs, the majority of indels shared one allele with the DM 1-3 reference with the second allele being novel to the cultivar (Type B). Consistent with observations with the SNPs, limited tri- and tetra-allelic indels were observed.

Intergenic regions represented approximately 77% of the genomic space sampled and contained ~43% of the SNPs and ~40% of the indels discovered among the six cultivars sampled. In regions of the genome 5,000 nucleotides upstream or downstream of genes, ~45% of all SNPs and ~51% of indels were observed (Table S2). Using all polymorphic sites, the impact of SNPs and indels on coding sequence was predicted on a per-variant basis. Among SNPs in the genic space of all cultivars, 417,413 (52.3%) SNPs were predicted to cause missense mutations, 370,420 SNPs (46.46%) introduced putative silent mutations and 9,504 SNPs (1.2%) introduced putative nonsense mutations when projected onto the reference potato genome annotation. For indels, 4,929 (0.25% of indels called) in total were predicted to cause frame-shift mutations. The lack of SNP phasing information across the entire gene length generates uncertainty in the interpretation of inferred effects on gene function when multiple SNPs are present within the coding region. The collective allelic variation in tetraploid potato suggests that single base substitutions of a single allele (simplex, triplex SNP sites) which result in multiple alleles at each locus are the dominant type of intra-genome variation in cultivated potato. When coupled with biallelic duplex sites and indels, multi-allelism is prevalent in the genic regions of the potato genome and consistent with breeder’s efforts to obtain maximal heterozygosity in which maximal heterosis can be achieved through by increasing intra-locus diversity (Chase 1963).

To visualize the genetic relatedness of the potato cultivars in this study, an unrooted neighbor-joining tree was constructed using biallelic SNPs derived from our SNP calls and genetic distance calculated with the formula derived from Gronau *et al.* (2011). The result shows a star-like phylogeny with Atlantic and Snowden forming a closer relationship due to a shared parentage of the cultivar Lenape (Figure S3). The remaining cultivars do not form any hierarchical structure despite their varied phenotypes and market classes, reflecting the general lack of population substructure shown previously by Hirsch *et al.* (2013) in analysis of a larger panel of North American cultivated potato.

Preferential allele expression is abundant in potato

In an autotetraploid species, the null expectation for gene expression of haplotypes is that the alleles contained within each haplotype will be expressed according to their dosage. This is in contrast to allopolyploids in which genome dominance occurs and one subgenome may be preferentially expressed over the other (Yang *et al.* 2016). In this study, PAE was defined as the statistical deviation of alleles in transcript abundance from their genotype ratios as determined by DNA genotyping. Using replicated RNA-seq data from leaf and tuber tissues (Figure S4; Data S1), non-copy number variable genes (see below) containing biallelic sites with a minimum RNA-seq read depth $\geq$ 50 were examined to determine the extent of PAE in cultivated potato. The mean number of SNPs evaluated per gene with these criteria varied from 6.9 – 8.4 SNPs per gene, with 4.1 – 6.7 SNPs per gene showing significant PAE using a two-sided binomial exact test with a false discovery rate (FDR) cutoff of 0.01 adjusted using Fisher’s Method. Using a cutoff of four significant SNPs per gene, between 2,180 and 3,502 genes showed PAE in leaves, and between 3,367 and 5,270 showed PAE in tubers (Figure 2a, Data S2). Tubers consistently showed a higher proportion of genes with PAE compared to leaves in all cultivars, which is not surprising as tubers are the primary focus of selection by breeders.

The dosage of allele expression at significant sites tended to be skewed towards the reference allele in comparison to non-significant sites in both replicates (Figure 3; Figure S5). This bias is not attributable to a mapping bias as allele ratios determined by aligning genomic DNA reads at these sites showed limited bias towards the reference allele (**Figure S6**) suggestive of a biological basis rather than a technical reason for the bias in expression of the reference allele. The reference genome was derived from a doubled monoploid of an adapted diploid *S. tuberosum* Group Phureja clone (Lightbourn and Veilleux 2007). Due to severe inbreeding depression caused by high genetic load, creation of a monoploid or doubled monoploid is extremely difficult in not only tetraploid but also diploid potato. Thus, we postulate that the alleles present in DM, which have been "selected" via the monoploid sieve, provide high fitness and therefore may be preferentially expressed compared to any alternative allele, especially alleles which may themselves cause functional changes in proteins. Additionally, recent analyses of a diversity panel revealed numerous introgressions from wild potato species in cultivated potato (Hardigan *et al.* 2016b) and when coupled to the common practice of breeders of incorporating traits such as disease resistance from wild potato species in cultivar development, the skewed PAE may be reflective of preferential expression of the reference allele in comparison to the alternate allele which may be derived from a wild species relative.

To provide a broad view of PAE at the molecular level, the function of genes with significant PAE in leaf and tuber was analyzed separately using enrichment analysis of Gene Ontology (GO) processes with an FDR cutoff of 0.01 (Data S3). Interestingly, in leaf tissues there was enrichment of the molecular functions "iron-sulfur cluster binding" and "metal cluster binding" in Atlantic, Kalkaska, and Missaukee. "Ligase activity" was enriched in PAE in both Kalkaska and Snowden leaf, and enriched in Atlantic, Kalkaska, Missaukee, Russet Norkotah, and Superior tuber. Preferential allele expression in tubers also was enriched in translation-related processes or components, such as "translation factor activity," "translation initiation factor activity," and "structural constituent of ribosome" in Atlantic, Missaukee, and

Superior. “Translation” was an enriched biology process in Superior leaf genes with PAE. Intriguingly, genes with PAE in tubers but not leaves showed enrichment in transport-related processes in all six cultivars suggesting that the expression of some alleles may contribute to improved agronomic performance as manifested through tuber production and tuber quality and as a consequence, were selected during the breeding process. The most shared GO terms were “organonitrogen compound biosynthetic process”, “organonitrogen compound metabolic process”, and “small molecule metabolic process”, all of which were significantly enriched in leaf and tuber PAE from all cultivars (Figure S7).

Tuber-specific and leaf-specific sharing of genes among the cultivars exhibiting allelic imbalance were also analyzed in the context of metabolic pathways (Figure 4a). Leaf samples showed many shared genes exhibiting allelic imbalance in expression in photosystem II components (chlorophyll a-b binding proteins; polypeptide subunits) and other photosynthesis-related pathway components including photosystem I (polypeptide subunits), ATP synthase, and components of the Calvin-Benson cycle (Rubisco small subunit and interacting proteins, glyceraldehyde-3-phosphate dehydrogenase B, fructose-bisphosphate aldolase, fructose-1-6-bisphosphatase, phosphoribulokinase) (Figure 4b). Thus, expression of certain alleles in photosynthesic pathways may be a key component in cultivated potato productivity and been targets of selection by breeders as a component of yield as carbon fixation in leaves is directly related to transport of sucrose to the tubers and conversion to starch.

Interestingly many genes in tuber samples showed PAE in pathways related to the tricarboxylic acid cycle, especially in pyruvate dehydrogenase and in components of the mitochondrial electron transport chain (Figure 5). Seven genes encoding subunits of ATP synthase, in particular, exhibit PAE and were specifically found in tuber samples only. Allelic imbalance in such highly expressed genes suggests that high plant productivity in potato may be attributed to the retention of functional haplotypes that are expressed.

Our panel includes five round white chip-processing cultivars and we examined PAE in genes relevant to tuber agronomic traits as previous work with chip processing potatoes had revealed selection by breeders over the last 50 years for specific alleles in genes involved in carbohydrate biosynthesis, metabolism, and transport (Hirsch *et al.* 2013). Using a set of 21 carbohydrate pathway genes that includes genes encoding enzymes involved in starch and sucrose synthesis and degradation with prior evidence of selection (Hirsch *et al.* 2013), we observed significantly more genes with PAE in tubers (p = 0.0001672; Welch two sample t-test). To evaluate if tubers were especially enriched in PAE in genes involved in starch biosynthesis, a list of starch synthesis-related genes (Van Harsselaar *et al.* 2017) was examined within the set of leaf and tuber PAE genes. Starch pathway genes were slightly enriched in tubers relative to leaves, although the results were only significant in Superior tuber without false discovery correction (p = 0.01088; Odds Ratio = 1.84; Fisher's Exact Test) (Figure S8). This gene set, similar to the gene set from Hirsch et al. (2013), includes genes involved in both synthesis and degradation of sucrose and starch and that optimal production of starch in not only tubers but also is affected by allelic variation.

Orthologous gene clusters were inferred from *Amborella trichopoda, Arabidopsis thaliana, Manihot esculenta, Oryza sativa, Solanum lycopersicum, Solanum tuberosum,* and *Vitis viniferis* to determine if there was an enrichment of PAE in genes conserved in all angiosperms relative to evolutionarily recent and lineage-specific genes as represented by *Solanum*-specific genes and potato-specific genes. Core angiosperm genes were significantly enriched in PAE in all samples (Table S3; Figure 2b), suggesting that the alleles of conserved angiosperm genes which encode core plant processes are differentially expressed in potato leaves and tubers. This is consistent with our metabolic pathway analyses which shows that enrichment of photosystem I reaction center components and genes related to photosynthesis including RuBisCO were among the most significant genes identified showing PAE in all leaf samples. As RuBisCO is abundantly expressed and a conserved gene among the angiosperms, PAE and translation of the optimal allele may be critical in efficient carbon fixation in potato. These results are consistent with observations in the natural allotetraploid *Glycine dolichocarpa,* in which photosynthetic genes and genes involved in transcription and translation show stronger biases in allelic expression in comparison to the genome-wide average (Coate *et al.* 2014).

Copy number variants affect transcript abundance and are enriched in genes responding to biotic and abiotic stress

A previous analysis of CNV in tetraploid potato using fluorescent in situ hybridization revealed that 10-15% of BAC clones showed CNV in a panel of 16 tetraploid potato clones (Iovene et al. 2013) while in a panel of 12 monoploid/doubled monoploid clones derived from diploid *S. tuberosum* Group Phureja (Hardigan *et al.* 2016a), 22.9% of genes had at least half of their annotated gene length overlap with CNV regions. With access to the entire potato genome for six tetraploid cultivars, we estimated CNV relative to the DM 1-3 reference genome using read depth variation over genic space (Figure 6a; Table 1). Genes on chromosomes 1 through 12 and unanchored scaffolds in “chromosome 00” were included in this analysis (Data S4). Using a threshold of >1.25X or <0.75X median genome depth to infer duplications and deletions, respectively, between 20.1% and 22.2% of genes were estimated to have copy number deletions in each cultivar, and between 5.8% and 8.2% of genes were estimated to contain duplications (Figure 6b; Table S4). The number of shared duplications and shared deletions was estimated in all cultivars and in total, 4,351 genes shared deletions among all six cultivars in the study (Data S5), and 848 genes shared duplications (Data S6), representing 13% of all annotated genes within this small panel.

Using a more conservative estimation of >1.5X or <0.5, a substantial number of deletions and duplications were still detected (Table S4). These numbers represent a very large portion of the gene space in comparison to humans, for which copy number variation is estimated to impact 4.8-9.5% of the genome (Zarrei *et al.* 2015). CNV has been investigated in other major plant species including *A. thaliana* (Cao *et al.* 2011), *Cucumis sativus* (Zhang *et al.* 2015), *O. sativa* (Xu *et al.* 2012), and *Z. mays* (Chia *et al.* 2012). This high extent of structural variation is nearly double that reported previously for tetraploids of potato using FISH (Iovene et al. 2013) which may be attributable to the limited number of FISH probes examined. The discovery of nearly twice the extent of CNV in tetraploid potato vs. monoploids derived from diploid *S. tuberosum* Group Phureja clones suggests that tetraploid potato can sustain a higher genetic load than diploid potato due to the presence of four rather than two homologous chromosomes. Also, monoploids are developed through anther culture and as a consequence of the meiotic sieve and selection of the most vigorous individuals, a higher proportion of deleterious/dysfunctional alleles are expected to be purged in monoploids and their derived doubled monoploids. The overall high degree of CNV is consistent with the genetic load as observed by inbreeding depression upon selfing in tetraploid potato.

The reference potato genome expression atlas, a collection of RNA-seq expression abundances determined from a developmental series as well as biotic and abiotic treatments (Massa *et al.* 2013, Massa *et al.* 2011), were obtained from Hardigan *et al.* (2016a), and provide a robust set of functional gene annotations that can provide insight into classes of genes subject to copy number variation. Genes annotated as constitutively expressed in all tissue types based on the DM 1-3 gene expression atlas were significantly under-enriched in both deleted (Two-tailed Fisher's exact test; Q = 3.93 x 10^{-241}) and duplicated genes (Q = 7.54 x 10^{-34}) suggesting that genes that are expressed in many tissues and most likely function as core housekeeping genes are not subject to copy number variation (Figure S9; Figure S10; Data S7; Data S8). In contrast, genes annotated as lowly expression in all tissues were significantly enriched in deleted genes (Q = 1.93 x 10^{-216}) and duplicated genes (Q = 2.94 x 10^{-11}). Genes responsive to abiotic stress based on five-fold induction relative to control conditions in the DM 1-3 expression atlas were enriched in both duplications (Q = 4.59 x 10^{-10}) and deletions (Q = 8.08 x 10^{-05}). Genes induced five-fold by biotic stress were significantly enriched in duplicated genes (*Q* = 0.0068, and slightly enriched but non-significant in deleted genes (*Q* = 0.13, odds ratio = 1.11). These results are consistent with that observed in Hardigan *et al.* (2016a) and are highly suggestive that CNV may be genes that are in the process of pseudogenization or may have a role in adaptive traits that are critical to agronomic production of potato. Copy number variable genes have been shown to have a role in disease resistance in soybean (Cook *et al.* 2012), submergence tolerance in rice (Xu *et al.* 2006), and aluminum tolerance in maize (Maron *et al.* 2013). With regards to the role of adaptive traits unrelated to agronomic production, gene copy number changes have also been found to occur in the extremophile species *Arabidopsis halleri*, which shows enrichment of gene copy number changes in genes related to biotic stress response and, most notably, genes related to transition metal homeostasis (Suryawanshi *et al.* 2016).

Orthologous groups constructed with *A. thaliana, A. trichopoda, M. esculenta, O. sativa, S. lycopersicum, V. viniferis* and DM 1-3 were analyzed with the putative copy number deletions and duplications to determine any association between evolutionary relationship and CNV (Data S9). In contrast to genes with evidence of PAE, which were significantly enriched in the conserved angiosperm gene set, CNVs were significantly under-enriched in core plant genes, suggesting that there may be fitness costs associated with CNV in genes highly conserved among the angiosperms. *Solanum*-specific orthologs, potato-specific paralogs and potato singletons were highly over-enriched in CNVs, suggesting that genes that have evolved more recently are not stringently retained in all four homologous chromosomes.

To explore the impact of CNV on transcript levels, genes were first annotated as expressed in leaf and tuber tissue based on expression abundances in the reference genotype DM 1-3 (Potato Genome Sequencing Consortium 2011) to filter out genes not anticipated to be expressed regardless of CNV state. As shown in Figure 6c, reduced gene copy number resulted in an overall copy number-dependent affect at the transcript level in both leaves and tubers, consistent with previous reports of copy number-dependent impacts of CNV on transcript abundance (Iovene *et al.* 2013). Gene expression was correlated with copy number using Pearson's correlation with an R^2 cutoff of 0.6 to define copy number-dependent gene expression (Figure 6d). In total, 528 genes were copy number-dependent in their expression values measured in leaf and tuber across the six genotypes; 302 genes were identified in leaf RNA-seq (Data S10), and 160 genes were identified in tuber RNA-seq (Data S10). Leaf and tuber shared 66 genes that showed copy number-dependent responses in both tissues as exemplified in Figure 6d in which four genes with dosage response in gene expression relative to copy number are shown. The examples shown are 1) a glucosyltransferase, which may have zeatin O-glucosyltransferase activity related to cytokinin production (Gong *et al.* 2015), 2) an extensin precursor, which constitutes part of the plant cell wall (Lamport *et al.* 2011), 3) a protein binding protein, and 4) a basic helix-loop-helix protein BHLH5, which is likely the potato ortholog of the *A. thaliana* transcription factor *BEE 1* (Brassinosteroid-Enhanced Expression 1) based on sequence identity (66.4%) and coverage (83%). *BEE1* is known to regulate flavonoid accumulation (Petridis *et al.* 2016), and shade avoidance syndrome (Cifuentes-Esquivel *et al.* 2013) in *A. thaliana.*

Conclusion

This study represents a genome analysis of cultivated autotetraploid potato examining the relationships between genome variation and gene expression. Intra-genome variation within tetraploid potato is exceptionally high with extensive heterogeneity manifested through multi-allelic sites diverged at the sequence and copy number level. This genome level variation is manifested at the transcript level in which not only additive but also preferential allele expression was observed. Thus, although we did not observe an overall "genome dominance" as has been observed in allopolyploids, there is abundant preferential allele gene expression that results in allelic imbalance at the transcriptome level. From examination of only six cultivars, we revealed an exceptionally high degree of copy number variation, particularly in the form of deletions of homologous gene copies that can be masked at the tetraploid level due to complementation attributable to polyploidy. Using the DM 1-3 reference potato genome expression atlas, duplicated and deleted genes were found to be enriched in lowly expressed genes, genes responsive to biotic and abiotic stress, and under-enriched in constitutively expressed genes suggesting that one component of breeder's selection has been for significant structural variation of loci involved in adaptation, consistent with the highly adaptive nature of potato as a commercial crop.

EXPERIMENTAL PROCEDURES (METHODS)

Library preparation, sequencing, and post-processing

Six elite cultivars that represent three market classes, chip-processing, round white fresh market, and fresh market russet potato, and released over the last ~70 years of breeding were selected for this study to provide representation of various market classes and impacts of artificial selection over the last ~70 years (Table S5). Genomic DNA was extracted from leaves of each cultivar using the Qiagen DNeasy Plant Mini Kit. Illumina compatible DNA libraries were constructed and approximately 240 million 100-nucleotide paired-end reads were generated on the Illumina HiSeq platform; the Superior library was sequenced in paired-end mode generating 150 nt reads. Raw reads were end-trimmed to remove bases with assigned qualities lower than 10 and adapters were removed using CutAdapt version 1.4.1 (Martin 2011); reads with a minimum length of 30 were retained for downstream analyses.

RNA was extracted from frozen ground leaf and tuber from each cultivar using hot phenol. Libraries were prepared and sequenced by Novogene (https://en.novogene.com) on the Illumina HiSeq 4000 generating 150-nucleotide paired-end reads for each sample. Raw reads generated from paired-end sequencing were cleaned in paired-end mode with CutAdapt version 1.8.1 (Martin 2011) to remove adapters and trim bases with base quality lower than 30. A minimum read length of 100 nucleotides was required for reads to be retained.

Mapping of genomic DNA reads to reference genome and variant calling

Cleaned genomic DNA reads were aligned to the reference *S. tuberosum* Group Phureja DM 1-3 assembly version 4.04 (Hardigan *et al.* 2016a) using the BWA-MEM algorithm within the Burrows-Wheeler Aligner (BWA) software version 0.7.7.r441 (Li and Durbin 2009, Li *et al.* 2009) in paired-end mode and saved in binary alignment format (BAM) using SAMTools version 0.1.19 (Li *et al.* 2009). This output was sorted, deduplicated, and indexed using Picard version 1.89 within the Genome Analysis Toolkit (GATK) version 3.1-1 framework (McKenna *et al.* 2009; DePristo *et al.* 2011; Van der Auwera *et al.* 2013). Subsequently, the reads were subjected to indel realignment around target regions using Picard version 1.86. Single nucleotide polymorphisms and indels were called in all samples jointly using the GATK's UnifiedGenotyper in GATK 3.1-1 (DePristo *et al.* 2011) with a standard minimum confidence threshold for calling of 30.0, the genotype likelihoods model "BOTH", sample ploidy of four, heterozygosity of .01, indel heterozygosity of .001, maximum deletion fraction of 0.1, maximum minimum base quality score of 17, minimum indel count for genotyping of five, and minimum indel fraction per sample of 0.10.

The called variants were hard filtered to select the final set of variants using the following parameters in GATK 3.1-1's VariantFiltration tool to remove low quality SNPs: quality by depth less than 10.0, mapping quality less than 40, strand bias estimated by Fisher's Exact Test greater than 60.0, Haplotype Score > 13.0, mapping quality rank sum test less than -12.5, read position rank sum test less than -8.0, depth greater than 400, depth less than 60, and genotype quality less than 10. The filtered variants were parsed according to genotype calls into the 14 possible combinations of alleles. Variant impacts were estimated using SNPEff version 4.0e (Cingolani *et al.* 2012).

To determine putative CNV based on read depth, median depth was calculated for each gene and normalized by dividing by the median read depth of all genes in each cultivar in R 3.0.14 (R Core Team, 2014). Genes with copy number-dependent gene expression were identified by calculating Pearson's correlation coefficients for genes with at least two different copy numbers among the cultivars using rLog-transformed expression values from DESeq2 from leaf and tuber RNA-seq as the response variable and copy number as the explanatory variable. Genes were considered dosage-sensitive if they met the R^2 cutoff of 0.6.

Enrichment of genes regulated by stress in CNV genes

Annotation of the potato gene complement relative to expression patterns and levels were obtained from (Hardigan *et al.* 2016a) that included constitutively expressed genes (across 32 tissue and treatment series), lowly expressed genes, and upregulated following biotic or abiotic-stress. Two-sided Fisher's exact tests were used to determine enrichment of shared duplicated genes between all six cultivars in upregulated gene sets in these categories. Results were considered significant at FDR < 0.01.

RNA-seq read mapping and calculation of expression values

RNA-seq reads were mapped to the reference genome version 4.04 in paired-end mode for corresponding libraries using Tophat2 version 2.1.0 (Trapnell *et al.* 2009). Reads aligning to annotated genes were counted using HTseq version 0.6.1p1 in stranded, union mode with minimum mapping quality of 17 (Anders *et al.* 2014). Counts from plant tissues were analyzed using DESeq2 to determine normalized expression values for analysis of tissue types (Love *et al.* 2014). FPKMs were determined using Cufflinks version 1.3.0 (Trapnell *et al.* 2012).

Determination of preferential allele expression and gene ontology enrichment analysis

The output from Tophat alignments was used to call SNPs in the RNA-seq data using the mpileup function in SAMTools version 0.1.19 (Li *et al.* 2009). Only uniquely mapping reads (MAPQ = 255) were used for PAE analysis and genes with evidence of CNV were removed. mRNA-seq reads were counted for each allele and sites were intersected with whole genome DNA resequencing read-based calls (i.e., genotype) to determine the expected ratio of expression for each allele at biallelic sites with one reference and one alternate allele. Outliers were identified using a two-sided binomial exact test with False Discovery Rate (FDR) (Benjamini and Yekutieli 2001) correction to determine Q-values, treating the number of reference allele reads as the successes and the total number of reads as the number of trials in an expected proportion determined by the genomic DNA-derived genotype (0.25, 0.50, and 0.75 for simplex, duplex, and triplex alternate alleles, respectively). The null hypothesis was set according to the dosage of alleles in genomic DNA. Fisher's Method was implemented using the R package metap version 0.8 (Dewey 2017).

Gene ontology terms were determined using Interproscan version 5.8.49.0 (Jones *et al.* 2014) on the version 3.4 potato gene annotation protein sequences. Gene ontology terms were subsequently used to perform GO enrichment analysis using Fisher's Exact Test implemented in the R package TopGO (Alexa and Rahnenfuhrer 2010). Tests were performed using gene IDs of biallelic sites with $Q < 0.01$ for non-additive expression to determine enrichment in cellular components, biological processes, and molecular functions. The contingency tables consisted of genes with PAE and belonging to the focal GO category, genes without PAE and belonging to the focal GO category, genes with PAE not belonging to the focal GO category, and genes without PAE not belonging to the focal GO category. The p-values were adjusted using FDR (Benjamini and Yekutieli 2001).

Identification of orthologous groups

Representative proteins from *A. trichopoda* (v1.0), *A. thaliana* (TAIR10), *M. esculenta* (v.6.1), *O. sativa* (MSU 7.0), *S. lycopersicum* (iTAG2.3), and *V.vinifera* (v.12X) were downloaded from Phytozome 11 (https://phytozome.jgi.doe.gov/) and orthologous groups were inferred using Orthofinder version 0.7.1 (Emms and Kelly 2015) with the default inflation parameter. Genes were characterized as “core plant genes” if they belonged to orthologous groups that contained every species in the analysis. *Solanum*-specific genes contained at least one gene from both potato and tomato, potato paralogous groups contained two or more genes unique to potato, and potato-specific singletons were in groups where there was a single potato gene only with no inferred orthology with any other plants in the analysis.

Accession numbers

The raw genomic DNA sequencing reads are available in the NCBI Sequence Read Archive (SRA) with the project numbers PRJNA386512 and PRJNA287438. The raw RNA-seq reads are available in project number PRJNA386512.

Accession numbers of raw sequencing data in NCBI Sequence Read Archive: PRJNA386512, PRJNA287438, PRJNA378971

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Science Foundation (ISO-1237969) to CRB and DSD, an award to DSD and CRB from USDA NIFA (2009-85606-05673), and through funds from Michigan State University. GP was supported in part by USDA NIFA (2017-67011-26038). GMP, LN, and KWR, performed the experiments. BV submitted data to online repository and managed data during the project. GMP did the analysis. GMP and CRB wrote the manuscript. The authors declare no conflict of interest.

SHORT SUPPORTING INFORMATION LEGENDS

Figure S1. Classification of single nucleotide polymorphisms and insertion/deletions into Type A through Type E.

Figure S2. Dosage (simplex, duplex, or triplex) of biallelic sites.

Figure S3. Neighbor-joining tree constructed from biallelic single nucleotide polymorphisms.

Figure S4. Pearson's correlation coefficients of tuber and leaf RNA-seq normalized gene counts.

Figure S5. Expression of alleles at biallelic sites in the second RNA-seq replicates.

Figure S6. Boxplots of DNA read counts and RNA read counts at biallelic sites.

Figure S7. Gene Ontology terms significantly enriched in PAE in multiple samples.

Figure S8. Preferential allele expression enrichment in starch pathway genes.

Figure S9. Over- or under-enrichment of deletions in DM1-3 expression atlas tissues and treatments.

Figure S10. Over- or under-enrichment of duplications in DM1-3 expression atlas tissues and treatments.

Table S1. Validation of genotypes using the Infinium 8303 array.

Table S2. Single nucleotide polymorphisms and insertion/deletion locations in genomic regions.

Table S3. Enrichment of preferential allele expression in orthogroups.

Table S4. Genes classified into conservative/non-conservative deletions and duplications

Table S5. Cultivar usage descriptions.

Data S1. Pearson's correlation coefficients of RNA-seq rLog-transformed values.

Data S2. Genes with preferential allele expression in leaves and tubers.

Data S3. Gene ontology terms enriched in preferential allele expression genes.

Data S4. Genes with copy number variation.

Data S5. Shared deletions among cultivars.

Data S6. Shared duplications among cultivars.

Data S7. Enrichment of DM expression atlas genes in shared deleted genes.

Data S8. Enrichment of DM expression atlas genes in shared duplicated genes.

Data S9. Enrichment of copy number variation in orthogroups.

Data S10. Gene expression responding to copy number variation in leaf and tuber.

REFERENCES

Alexa, A. and Rahnenfuhrer, J. (2010) topGO: Enrichment analysis for Gene Ontology. In *R package*. Bioconductor.

Anders, S., Pyl, P.T. and Huber, W. (2014) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*.

Bell, G.D.M., Kane, N.C., Rieseberg, L.H. and Adams, K.L. (2013) RNA-Seq Analysis of Allele-Specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. *Genome Biol Evol*, **5**, 1309-1323.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, **29**, 1165-1188.

Buckberry, S., Bianco-Miotto, T., Hiendleder, S. and Roberts, C.T. (2012) Quantitative allele-specific expression and DNA methylation analysis of H19, IGF2 and IGF2R in the human placenta across gestation reveals H19 imprinting plasticity. *PLoS One*, **7**, e51210.

Buckland, P.R. (2004) Allele-specific gene expression differences in humans. *Hum Mol Genet*, **13 Spec No 2**, R255-260.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J. and Weigel, D. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics*, **43**, 956-U960.

Chase, S.S. (1963) Analytic breeding in Solanum tuberosum L.—a scheme utilizing parthenotes and other diploid stocks. *Can. J. Genet. Cytol.*, **5**, 359–363.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J.S., Li, M., Liu, X., Lu, Y.L., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhajarvi, T., Rong, T.Z., Sekhon, R.S., Sun, Q., Tenaillon, M.I., Tian, F., Wang, J., Xu, X., Zhang, Z.W., Kaeppler, S.M., Ross-Ibarra, J., McMullen, M.D., Buckler, E.S., Zhang, G.Y., Xu, Y.B. and Ware, D. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*, **44**, 803-U238.

Cifuentes-Esquivel, N., Bou-Torrent, J., Galstyan, A., Gallemi, M., Sessa, G., Salla Martret, M., Roig-Villanova, I., Ruberti, I. and Martinez-Garcia, J.F. (2013) The bHLH proteins BEE and BIM positively modulate the shade avoidance syndrome in Arabidopsis seedlings. *The Plant journal : for cell and molecular biology*, **75**, 989-1002.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80-92.

Coate, J.E., Bar, H. and Doyle, J.J. (2014) Extensive translational regulation of gene expression in an allopolyploid (Glycine dolichocarpa). *The Plant cell*, **26**, 136-150.

Coate, J.E., Song, M.J., Bombarely, A. and Doyle, J.J. (2016) Expression-level support for gene dosage sensitivity in three Glycine subgenus Glycine polyploids and their diploid progenitors. *New Phytol*.

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

Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L., Yun, Z., Bell, T.A., Buus, R.J., Calaway, M.E., Didion, J.P., Gooch, T.J., Hansen, S.D., Robinson, N.N., Shaw, G.D., Spence, J.S., Quackenbush, C.R., Barrick, C.J., Nonneman, R.J., Kim, K., Xenakis, J., Xie, Y., Valdar, W., Lenarcic, A.B., Wang, W., Welsh, C.E., Fu, C.P., Zhang, Z., Holt, J., Guo, Z., Threadgill, D.W., Tarantino, L.M., Miller, D.R., Zou, F., McMillan, L., Sullivan, P.F. and Pardo-Manuel de Villena, F. (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics*, **47**, 353-360.

de Boer, J.M., Datema, E., Tang, X.M., Borm, T.J.A., Bakker, E.H., van Eck, H.J., van Ham, R.C.H.J., de Jong, H., Visser, R.G.F. and Bachem, C.W.B. (2015) Homologues of potato chromosome 5 show variable collinearity in the euchromatin, but dramatic absence of sequence similarity in the pericentromeric heterochromatin. *Bmc Genomics*, **16**.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. and Daly, M.J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491-498.

Dewey, M. (2017) metap: meta-analysis of significance values.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**, 157.

Gong, L., Zhang, H., Gan, X., Zhang, L., Chen, Y., Nie, F., Shi, L., Li, M., Guo, Z., Zhang, G. and Song, Y. (2015) Transcriptome Profiling of the Potato (Solanum tuberosum L.) Plant under Drought Stress and Water-Stimulus Conditions. *PLoS One*, **10**, e0128041.

Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D. and Dulac, C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643-648.

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A. (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, **43**, 1031-1034.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004) Allelic variation of gene expression in maize hybrids. *The Plant cell*, **16**, 1707-1716.

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

Hardigan, M.A., Crisovan, E., Wiegert-Rininger, K., Laimbeer, P., Douches, D.S., Veilleux, R.E. and Buell, C.R. (2016b) Analysis of sequence diversity in Solanum sect. Petota species identifies loci under selection during domestication of cultivated Solanum tuberosum. In *13th Annual Solanaceae Conference*. Davis, CA, USA.

Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A.J. and Drake, T.A. (2014) Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics*, **15**, 471.

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

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux,

R.E., Jansky, S., Bethke, P., Douches, D.S. and Buell, C.R. (2013) Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries. *G3*, **3**, 1003-1013.

Iovene, M., Zhang, T., Lou, Q., Buell, C.R. and Jiang, J. (2013) Copy number variation in potato - an asexually propagated autotetraploid species. *The Plant journal : for cell and molecular biology*, **75**, 80-89.

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

Lamport, D.T., Kieliszewski, M.J., Chen, Y. and Cannon, M.C. (2011) Role of the extensin superfamily in primary cell wall architecture. *Plant Physiol*, **156**, 11-19.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

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

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.

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

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

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

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

Petridis, A., Doll, S., Nichelmann, L., Bilger, W. and Mock, H.P. (2016) Arabidopsis thaliana G2-LIKE FLAVONOID REGULATOR and BRASSINOSTEROID ENHANCED EXPRESSION1 are low-temperature regulators of flavonoid accumulation. *New Phytol*, **211**, 912-925.

Potato Genome Sequencing Consortium, T., Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S.K., Patil, V.U., Skryabin, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin, D.M., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G.J., Sagredo, B., Mejia, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., Herrera Mdel, R., Giuliano, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H., Massa, A.N., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C.R., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan, G.J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S.J., Fiers, M., Jacobs, J.M.,

Nielsen, K.L., Sonderkaer, M., Iovene, M., Torres, G.A., Jiang, J., Veilleux, R.E., Bachem, C.W., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B., Goverse, A., van Ham, R.C. and Visser, R.G. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189-195.

Song, G.Y., Guo, Z.B., Liu, Z.W., Cheng, Q., Qu, X.F., Chen, R., Jiang, D.M., Liu, C., Wang, W., Sun, Y.F., Zhang, L.P., Zhu, Y.G. and Yang, D.C. (2013) Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F1 hybrids. *Bmc Plant Biol*, **13**.

Springer, N.M. and Stupar, R.M. (2007) Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *The Plant cell*, **19**, 2391-2402.

Suryawanshi, V., Talke, I.N., Weber, M., Eils, R., Brors, B., Clemens, S. and Kramer, U. (2016) Between-species differences in gene copy number are enriched among functions critical for adaptive evolution in Arabidopsis halleri. *BMC Genomics*, **17**, 1034.

Swaminathan, M.S. (1954) Nature of Polyploidy in Some 48-Chromosome Species of the Genus Solanum, Section Tuberarium. *Genetics*, **39**, 59-76.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**, 562-578.

Van de Peer, Y., Maere, S. and Meyer, A. (2009) The evolutionary significance of ancient genome duplications. *Nature reviews. Genetics*, **10**, 725-732.

Van Harsselaar, J.K., Lorenz, J., Senning, M., Sonnewald, U. and Sonnewald, S. (2017) Genome-wide analysis of starch metabolism genes in potato (Solanum tuberosum L.). *Bmc Genomics*, **18**.

von Korff, M., Radovic, S., Choumane, W., Stamati, K., Udupa, S.M., Grando, S., Ceccarelli, S., Mackay, I., Powell, W., Baum, M. and Morgante, M. (2009) Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *Plant Journal*, **59**, 14-26.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. and Rieseberg, L.H. (2009) The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 13875-13879.

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+.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J. and Wang, W. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature biotechnology*, **30**, 105-111.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S.K., Patil, V.U., Skryabin, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin, D.M., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G.J., Sagredo, B., Mejia, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., Herrera Mdel, R., Giuliano, G., Pietrella, M., Perrotta,

G., Facella, P., O'Brien, K., Feingold, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H., Massa, A.N., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C.R., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan, G.J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S.J., Fiers, M., Jacobs, J.M., Nielsen, K.L., Sonderkaer, M., Iovene, M., Torres, G.A., Jiang, J., Veilleux, R.E., Bachem, C.W., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B., Goverse, A., van Ham, R.C. and Visser, R.G. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189-195.

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D., Ju, Y., Yao, P., Li, X., Xie, K., Zhang, J., Wang, J., Liu, F., Ma, W., Shopan, J., Zheng, H., Mackenzie, S.A. and Zhang, M. (2016) The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature genetics*, **48**, 1225-1232.

Zarrei, M., MacDonald, J.R., Merico, D. and Scherer, S.W. (2015) A copy number variation map of the human genome. *Nature reviews. Genetics*, **16**, 172-183.

Zhang, X. and Borevitz, J.O. (2009) Global Analysis of Allele-Specific Expression in Arabidopsis thaliana. *Genetics*, **182**, 943-954.

Zhang, Z., Mao, L., Chen, H., Bu, F., Li, G., Sun, J., Li, S., Sun, H., Jiao, C., Blakely, R., Pan, J., Cai, R., Luo, R., Van de Peer, Y., Jacobsen, E., Fei, Z. and Huang, S. (2015) Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. *The Plant cell*, **27**, 1595-1604.

TABLES

Table 1. Genes impacted by putative copy number variation.

FIGURE LEGENDS

Figure 1. a) Summary of total bases surveyed in the potato genome and b) location of insertions/deletions (indels) and single nucleotide polymorphisms (SNPs) in genic regions of the potato genome. SNPs and indels were identified from callable sites in v4.04 of the potato genome (Hardigan *et al.* 2016a) and binned based on annotated gene features using the Potato Genome Sequencing Consortium annotation (Potato Genome Sequencing Consortium 2011). Counts of SNPs and indels were obtained using SNPeff (Cingolani *et al.* 2012).

Figure 2. a) Total genes showing evidence of preferential allele expression (PAE) using the two-sided binomial exact test with the null ratio defined according to the dosage of the reference allele. Biallelic sites with a minimum of 50 allele counts were considered and only genes with no copy number variation were included in the analysis. Genes with overlapping gene models were excluded from the analyses. A threshold of four significant SNPs per gene was used to determine if a gene showed PAE. b) Enrichment of PAE in genes belonging to lineage-specific groups (Potato specific singletons, Solanum-specific genes, Core plant genes) based on orthologous and paralogous clustering of the predicted proteomes from seven angiosperm genomes. Potato-specific paralogs were too lowly expressed or had multi-mapping reads and were not used for evaluation of PAE. Only genes with at least 50 counts at evaluated SNP sites and at least four SNPs with PAE were included in the significant results.

Figure 3. Distribution of allele dosage expression at significant sites (top) and non-significant sites (bottom) evaluated for preferential allele expression. The proportion of reference genotype allele relative to alternative genotype allele is presented as boxplots. The genotype "0" refers to the reference allele while "1" refers to the alternate allele. Data shown is from replicate one; similar results were seen in replicate two (Figure S5).

Figure 4. a) Overview of genes showing preferential allele expression (PAE) in metabolic pathways. The blue boxes indicate genes with PAE in at least one sample of leaf or tuber. b) PAE in leaf samples in photosynthesis-related pathways. The green boxes indicate genes with PAE in at least one sample. The figures were generated using Mapman software.

Figure 5. PAE in tuber samples in the TCA cycle. The red boxes indicate genes with PAE in at least one sample of tuber. The figure was generated using Mapman software.

Figure 6. a) Estimations of copy number variation in the genome. Copy numbers were estimated for 5 kb windows by dividing the median read depth of the window by the overall genome median read depth and multiplying by four to get the tetraploid copy numbers. Local polynomial regression fitting (LOESS) was used to generate the fitted lines b) Summary of copy number variation of genes in each genome estimated by dividing gene median read depth by genome median read depth. c) Expression of all genes in leaves (blue) and tubers (red) binned into estimated copy numbers based based on read depth variation. Estimated copy numbers greater than 5 are not shown. d) Example of copy number-dependent gene expression in four genes with copy number variation between the six cultivars in this study. The genes have Pearson's correlation coefficients of at least 0.6 and their expression values were transformed from gene counts of RNA-seq data using DESeq2 (Love et al., 2014). The genes shown (labeled in bottom-right) are 1: PGSC0003DMG400000432 (glucosyltransferase), 2: PGSC0003DMG400000776 (extensin precursor), 3: PGSC0003DMG400004753 (protein binding protein), 4: PGSC0003DMG400005388 (basic helix-loop-helix protein BHLH5).

Table 1. Genes impacted by putative copy number variation

Estimated copy number	Atlantic	Kalkaska	Missaukee	Russet Norkotah	Snowden	Superior
0	1,816	1,953	1,740	1,761	1,837	1,493
1	2,035	2,089	1,603	1,945	2,162	2,281
2	2,910	2,459	2,699	2,539	2,961	2,801
3	4,483	5,362	4,869	4,388	4,752	4,165
4	22,582	20,126	22,949	21,288	21,888	21,475
5-8	4,776	6,607	4,778	6,647	5,075	6,243
9-12	273	280	236	305	233	360
13-16	66	69	67	54	46	91
17-20	22	23	20	29	16	25
21-24	8	7	10	9	8	14
25-28	7	5	5	7	2	8
29-32	6	2	4	4	67	9
33+	63	65	67	71	0	82
Total putative deletions	11,244	11,863	10,911	10,633	11,712	10,740
Total putative duplications	5,221	7,058	5,187	7,126	5,447	6,832

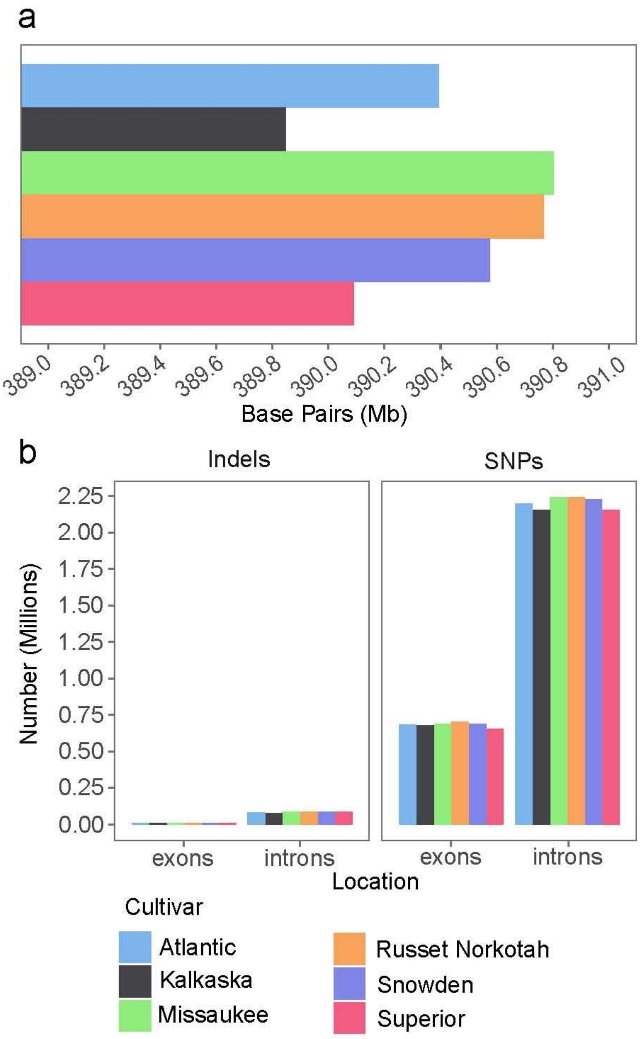

Figure 1. a) Summary of total bases surveyed in the potato genome and b) location of insertions/deletions (indels) and single nucleotide polymorphisms (SNPs) in genic regions of the potato genome. SNPs and indels were identified from callable sites in v4.04 of the potato genome (Hardigan et al. 2016a) and binned based on annotated gene features using the Potato Genome Sequencing Consortium annotation (PGSC, 2011). Counts of SNPs and indels were obtained using SNPeff (Cingolani et al., 2012).

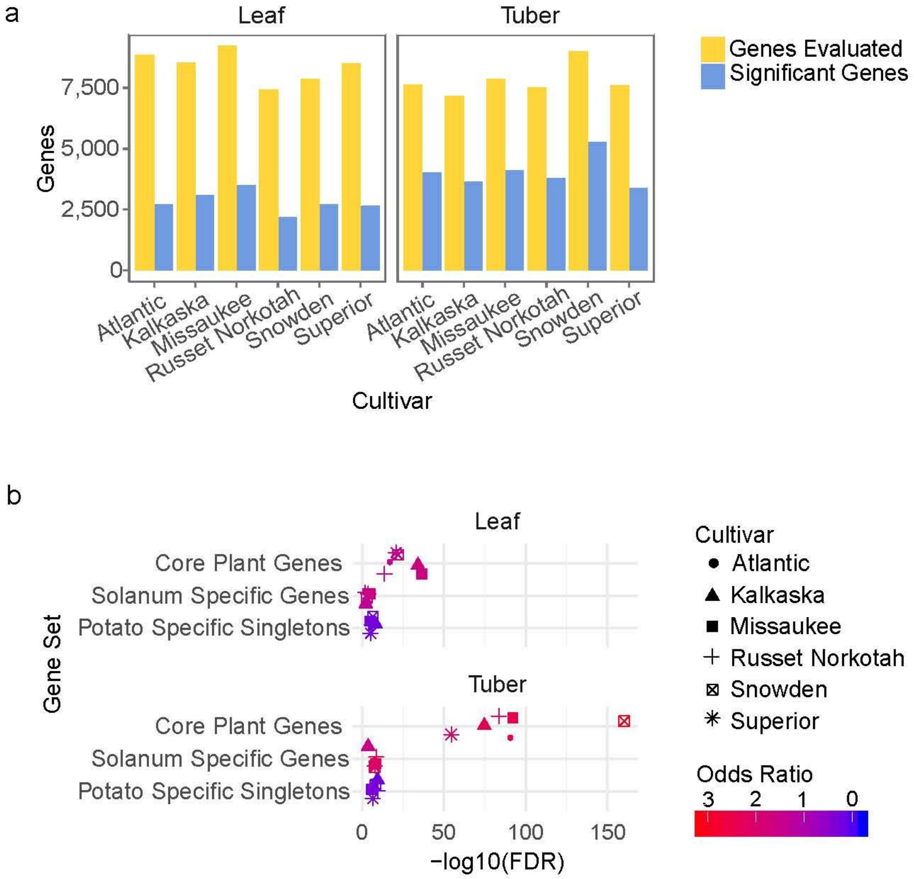

Figure 2. a) Total genes showing evidence of preferential allele expression (PAE) using the two-sided binomial exact test with the null ratio defined according to the dosage of the reference allele. Biallelic sites with at least 50 allele counts were considered and only genes with no copy number variation were included in the analysis. Genes with overlapping gene models were excluded from the analyses. A threshold of four significant SNPs per gene was used to determine if a gene showed PAE. b) Enrichment of PAE in potato genes belonging to potato- specific singletons, Solanum-specific orthogroups, and conserved plant orthogroups that contained genes from seven angiosperm genomes. Potato-specific paralogs were too lowly expressed or had multi-mapping reads that were not used for evaluation of PAE. Only genes with at least 50 counts at evaluated SNP sites and greater than three SNPs with PAE were included in the Fisher's exact test.

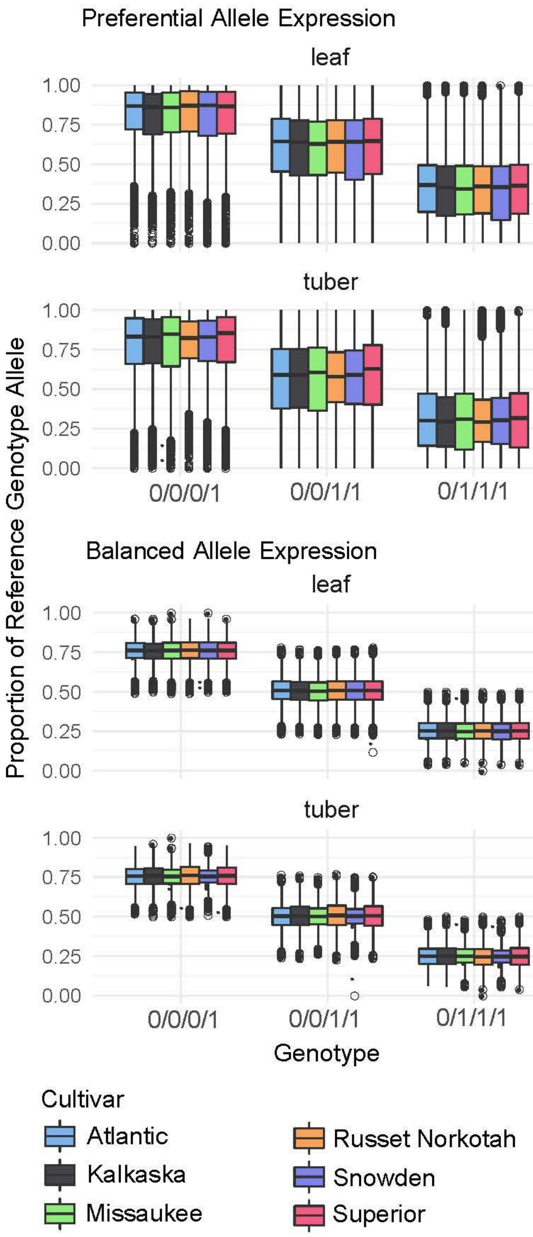

Figure 3. Distribution of allele dosage expression at significant sites (top) and non-significant sites (bottom) evaluated for preferential allele expression. The proportion of reference genotype allele relative to alternative genotype allele is presented as boxplots. The genotype “0” refers to the reference allele while “1” refers to the alternate allele. Data shown is from replicate one; similar results were seen in replicate two (Figure S5).

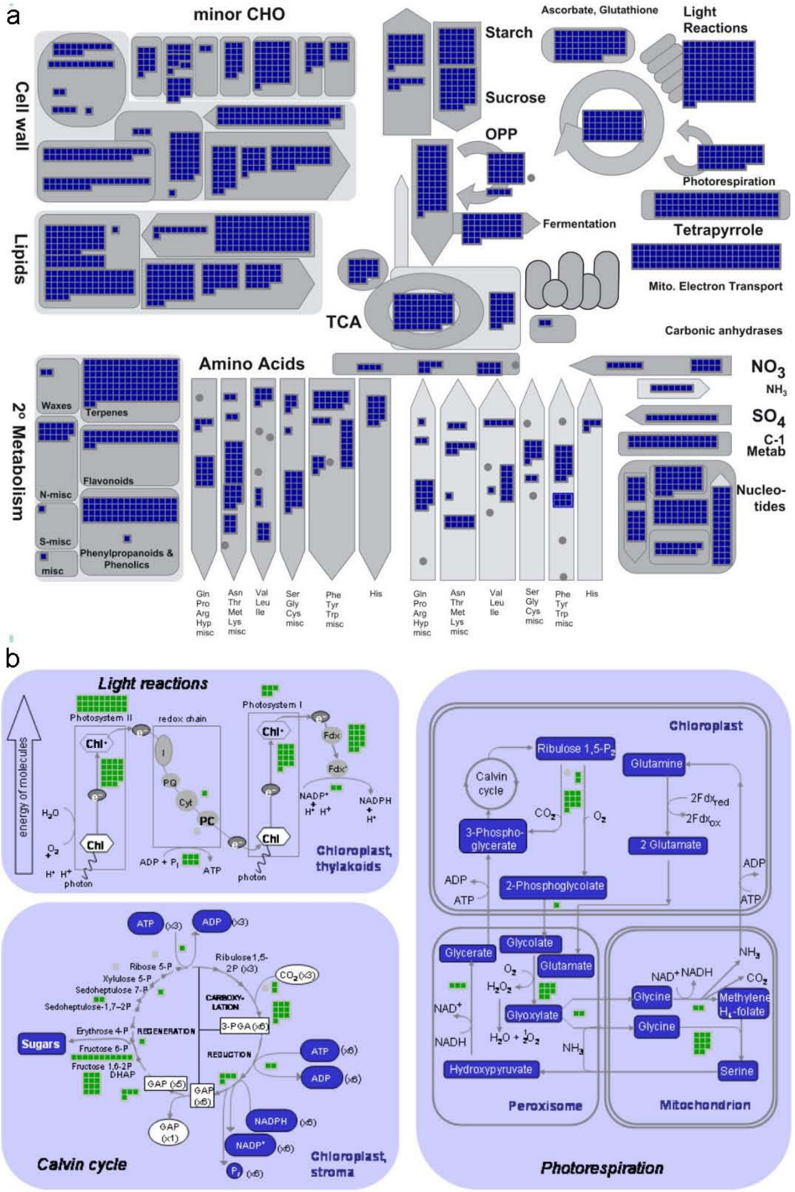

Figure 4. a) Overview of genes showing preferential allele expression (PAE) in metabolic pathways. The blue boxes indicate genes with PAE in at least one sample of leaf or tuber. b) PAE in leaf samples in photosynthesis-related pathways. The green boxes indicate genes with PAE in at least one sample. The figures were generated using Mapman software.

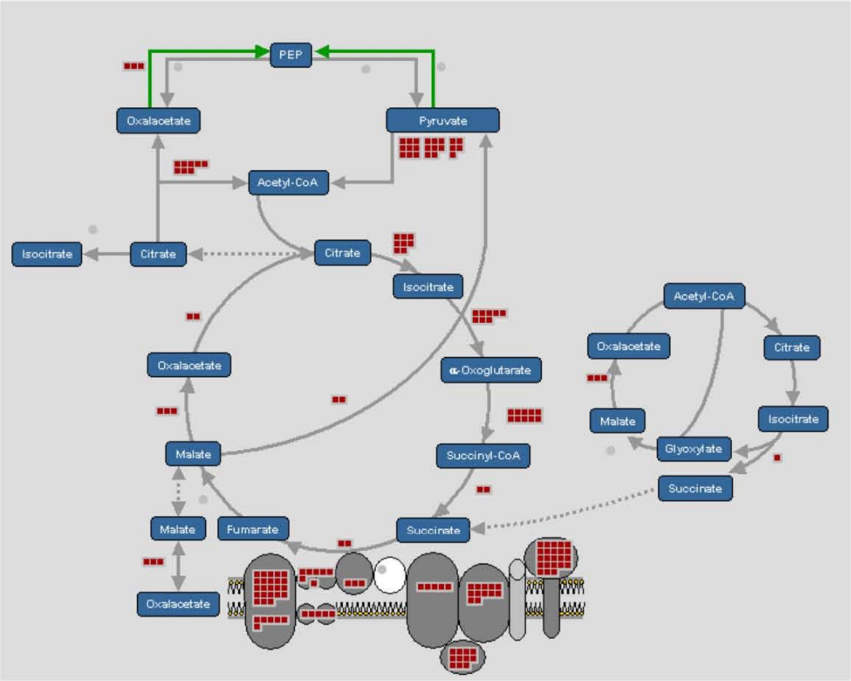

Figure 5. PAE in tuber samples in the TCA cycle. The red boxes indicate genes with PAE in at least one sample of tuber. The figure was generated using Mapman software.

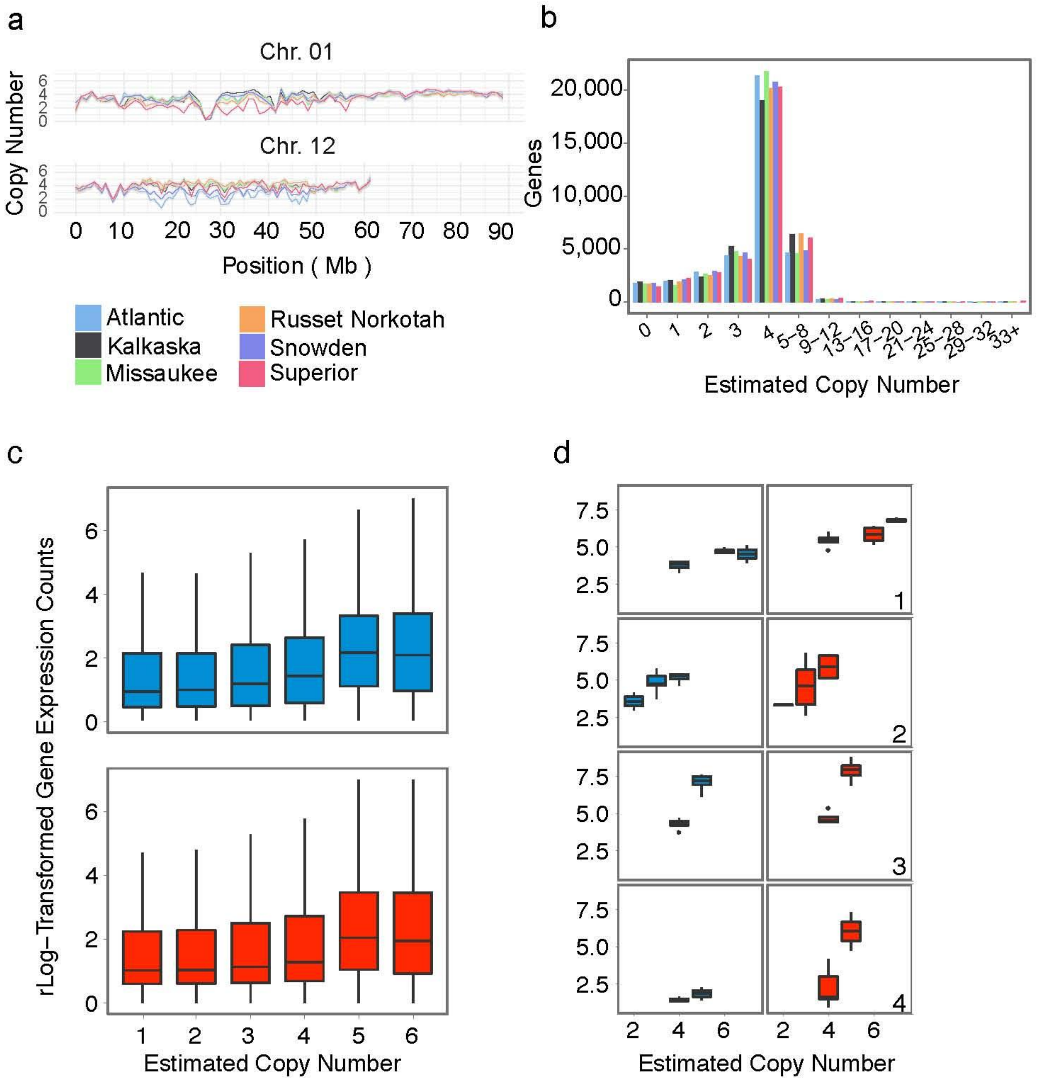

Figure 6. a) Estimations of copy number variation in the genome. Copy numbers were estimated for 5 kb windows by dividing the median read depth of the window by the overall genome median read depth and multiplying by four to get the tetraploid copy numbers. Local polynomial regression fitting (LOESS) was used to generate the fitted lines. The legend colors apply for figures a and b. b) Summary of copy number variation of genes in each genome estimated by dividing gene median read depth by genome median read depth. c) Expression of all genes in leaves (blue) and tubers (red) binned into estimated copy numbers based based on read depth variation. Estimated copy numbers greater than 5 are not shown. d) Example of copy number-dependent gene expression in four genes with copy number variation between the six cultivars in this study. The genes have Pearson's correlation coefficients of at least 0.6 and their expression values were transformed from gene counts of RNA-seq data using DESeq2 (Love et al., 2014). The genes shown (labeled in bottom-right) are 1: PGSC0003DMG400000432 (glucosyltransferase), 2: PGSC0003DMG400000776 (extensin precursor), 3: PGSC0003DMG400004753 (protein binding protein), 4: PGSC0003DMG400005388 (basic helix-loop-helix protein BHLH5).