CellPress
Partner Journal

Molecular Plant
Research Article

Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity

Genevieve Hoopes[1], Xiaoxi Meng[2], John P. Hamilton[1], Sai Reddy Achakkagari[3], Fernanda de Alves Freitas Guesdes[4], Marie E. Bolger[5], Joseph J. Coombs[6], Danny Esselink[4], Natalie R. Kaiser[6,7], Linda Kodde[4], Maria Kyriakidou[3], Brian Lavrijssen[4], Natascha van Lieshout[4], Rachel Shereda[1], Heather K. Tuttle[2], Brieanne Vaillancourt[1], Joshua C. Wood[1], Jan M. de Boer[8], Nolan Bornowski[1], Peter Bourke[4], David Douches[6], Herman J. van Eck[4], Dave Ellis[9], Max J. Feldman[10], Kyle M. Gardner[11], Johannes C.P. Hopman[8], Jiming Jiang[1,22], Walter S. De Jong[12], Joseph C. Kuhl[13], Richard G. Novy[14], Stan Oome[15], Vidyasagar Sathuvalli[16], Ek Han Tan[17], Remco A. Ursum[15], M. Isabel Vales[18], Kelly Vining[19], Richard G.F. Visser[4], Jack Vossen[4], G. Craig Yencho[20], Noelle L. Anglin[9,14], Christian W.B. Bachem[4], Jeffrey B. Endelman[21], Laura M. Shannon[2], Martina V. Strömvik[3], Helen H. Tai[11], Björn Usadel[5,23], C. Robin Buell[1,24,*] and Richard Finkers[4,25,*]

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

[2]Department of Horticultural Science, University of Minnesota, St. Paul, MN 55108, USA

[3]Department of Plant Science, McGill University, Sainte-Anne-de-Bellevue, QC H9X 3V9, Canada

[4]Plant Breeding, Wageningen University & Research, Plant Breeding, 6708 PB Wageningen, the Netherlands

[5]IBG-4 Bioinformatics, Forschungszentrum Jülich, Wilhelm Johnen Str, 52428 Jülich, Germany

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

[7]Bayer Crop Science, Woodland, CA 95695, USA

[8]Averis Seeds B.V., 9640 AA Veendam, the Netherlands

[9]International Potato Center, 1895 Avenida La Molina, Lima, Peru

[10]USDA-ARS, Prosser, WA 99350, USA

[11]Agriculture and Agri-Food Canada Fredericton Research and Development Centre, Fredericton, NB E3B 4Z7, Canada

[12]School of Integrative Plant Science, Cornell University, Ithaca, NY 14853-1901, USA

[13]Department of Plant Sciences, University of Idaho, Moscow, ID 83844, USA

[14]USDA-ARS, Small Grains and Potato Germplasm Research, Aberdeen, ID 83210, USA

[15]HZPC Research B.V., Edisonweg 5, 8501 XG Joure, the Netherlands

[16]Department of Crop and Soil Science, Oregon State University, Hermiston, OR 97838, USA

[17]School of Biology and Ecology, University of Maine, 5735 Hitchner Hall Orono, ME 04469, USA

[18]Department of Horticultural Sciences, Texas A&M University, College Station, TX 77843-2133, USA

[19]Department of Horticulture, Oregon State University, Corvallis, OR 97331, USA

[20]Department of Horticultural Science, North Carolina State University, Raleigh, NC 27695-7609, USA

[21]Department of Horticulture, University of Wisconsin-Madison, Madison, WI 53706, USA

[22]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA

[23]Institute for Biological Data Science, Heinrich Heine University, Düsseldorf, 40225 Düsseldorf, Germany

[24]Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA

[25]Present address: GenNovation B.V., Wageningen, the Netherlands

***Correspondence: C. Robin Buell (robin.buell@uga.edu), Richard Finkers (richard.finkers@wur.nl)**

https://doi.org/10.1016/j.molp.2022.01.003

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

ABSTRACT

Cultivated potato is a clonally propagated autotetraploid species with a highly heterogeneous genome. Phased assemblies of six cultivars including two chromosome-scale phased genome assemblies revealed extensive allelic diversity, including altered coding and transcript sequences, preferential allele expression, and structural variation that collectively result in a highly complex transcriptome and predicted proteome, which are distributed across the homologous chromosomes. Wild species contribute to the extensive allelic diversity in tetraploid cultivars, demonstrating ancestral introgressions predating modern breeding efforts. As a clonally propagated autotetraploid that undergoes limited meiosis, dysfunctional and deleterious alleles are not purged in tetraploid potato. Nearly a quarter of the loci bore mutations are predicted to have a high negative impact on protein function, complicating breeder's efforts to reduce genetic load. The *StCDF1* locus controls maturity, and analysis of six tetraploid genomes revealed that 12 allelic variants of *StCDF1* are correlated with maturity in a dosage-dependent manner. Knowledge of the complexity of the tetraploid potato genome with its rampant structural variation and embedded deleterious and dysfunctional alleles will be key not only to implementing precision breeding of tetraploid cultivars but also to the construction of homozygous, diploid potato germplasm containing favorable alleles to capitalize on heterosis in F1 hybrids.

Key words: potato, allele diversity, phased assembly, polyploid, pan-genome, wild introgressions

Hoopes G., Meng X., Hamilton J.P., Achakkagari S.R., de Alves Freitas Guesdes F., Bolger M.E., Coombs J.J., Esselink D., Kaiser N.R., Kodde L., Kyriakidou M., Lavrijssen B., van Lieshout N., Shereda R., Tuttle H.K., Vaillancourt B., Wood J.C., de Boer J.M., Bornowski N., Bourke P., Douches D., van Eck H.J., Ellis D., Feldman M.J., Gardner K.M., Hopman J.C.P., Jiang J., De Jong W.S., Kuhl J.C., Novy R.G., Oome S., Sathuvalli V., Tan E.H., Ursum R.A., Vales M.I., Vining K., Visser R.G.F., Vossen J., Yencho G.C., Anglin N.L., Bachem C.W.B., Endelman J.B., Shannon L.M., Strömvik M.V., Tai H.H., Usadel B., Buell C.R., and Finkers R. (2022). Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. Mol. Plant. **15**, 520–536.

INTRODUCTION

Potato (*Solanum tuberosum*) is the world's third most consumed food crop (http://www.fao.org/) and serves as a key food security crop in developing countries. There has been an increase in potato consumption in developing countries due to its wide adaptive range, ease of cultivation, and high nutritional value, all of which has led to significant increases in potato production and demand in Asia, Africa, and Latin America (Devaux et al., 2014). Potato, whose origin and diversity resides in the Andes, was domesticated 8000–10 000 years ago (Spooner et al., 2005), and has since been dispersed throughout the world (more than 160 countries; http://www.fao.org/). Key features of domestication and subsequent improvement included an increase in tuber size, a reduction in total tuber glycoalkaloids, and the ability to tuberize under long days as it originated in a short-day environment near the equator (Johns and Alonso, 1990; Jansen et al., 2001; Bradshaw et al., 2006; Friedman, 2006). Cultivated tetraploid potato exhibits signatures of admixture from a number of wild species which are hypothesized to have been sympatric during its migration from Peru and Bolivia to the southern latitudes of Chile (Hardigan et al., 2017).

Current cultivated potato varieties can be classified into various market classes, including processing (chips, frozen fries), fresh market, and starch. Modern breeding efforts are focused on improving quality traits, yield, and incorporating disease resistance, for which wild potato species remain a powerful resource for adaptive traits. As a highly heterozygous, autotetraploid, outcrossing crop, breeders have focused on creating optimal combinations of alleles in F1 progeny, resulting in the highest level of allelic diversity reported in a major crop (Hardigan et al., 2017). Allelic diversity in potato includes single nucleotide polymorphisms (SNPs), small insertions/deletions, and structural variants including deletions of up to 575 kb (Hardigan et al., 2017; Pham et al., 2017). Due to the clonal propagation nature of potato, deleterious and dysfunctional alleles are not purged through meiosis, resulting in high genetic load and inbreeding depression when selfed. Genome-scale studies of a heterozygous diploid clone revealed a substantial number of dysfunctional alleles that are fixed in repulsion and contribute to phenotypic diversity (Potato Genome Sequencing Consortium, 2011; Zhou et al., 2020). With an estimated 3.1 alleles per locus within tetraploid cultivars (Willemsen, 2018), coupled with rampant structural variation, the high degree of intra-genome heterogeneity has complicated the generation of a high-quality genome assembly of tetraploid potato.

Genome plasticity is well known in the plant kingdom, with pan-genomes available for several species, cementing the concept of core genes coupled with shell and cloud genes (dispensable or accessory genes) composing the pan-genome (for review see Bayer et al., 2020; Della Coletta et al., 2021). Features of core and accessory genes differ, mirroring their evolutionary history. Core genes tend to have a broader expression breadth and are more highly expressed than accessory genes, which

	Altus	Atlantic	Avenger	Castle Russet	Colomba	Spunta
Number of scaffolds	9762	12 835	10 396	16 227	9923	8725
Total size of scaffolds (bp)	2 140 119 115	2 447 835 705	2 345 613 167	2 217 488 234	1 990 703 545	1 992 411 514
Maximum scaffold (bp)	18 346 635	36 222 390	19 651 727	21 151 798	20 005 369	13 855 216
Number of scaffolds >10 000 nt	9761	12 835	10 395	16 227	9922	8724
Number of scaffolds > 100 000 nt	1767	2241	1954	2453	2166	2331
N50 scaffold length (bp)	2 276 075	2 565 074	1 954 301	1 448 277	1 434 362	1 244 616
L50 scaffold count	207	190	271	295	310	383
scaffold %N	1.78	1.18	1.74	0.75	1.07	3.07

Table 1. Metrics of initial DeNovoMagic phased genome assemblies

are lowly expressed, if at all. Core genes also tend to have conserved functions, whereas accessory genes are more likely to have no known function (Hirsch et al., 2014, 2016; Gordon et al., 2017). Collectively, a majority of accessory genes are on the path to pseudogenization, but a subset have been shown to function in adaptive responses, including response to pathogens, abiotic stress tolerance, and photoperiod adaptation (Hattori et al., 2009; Cook et al., 2012; Diaz et al., 2012). An initial study of the potato pan-genome using a small number of monoploid and doubled monoploid clones revealed a higher degree of structural variation than that observed in diploid, sexually reproducing species, suggesting that the size of the pan-genome in a species is affected by the species' reproductive mode (i.e., outcrossing, inbreeding, or clonal propagation), ploidy, and the diversity sampled (Hardigan et al., 2016).

To date, genome assemblies of potato include a reference genome sequence of the doubled monoploid *S. tuberosum* Gp Phureja DM 1-3 R44 (hereafter DM) (Potato Genome Sequencing Consortium, 2011; Pham et al., 2020), an inbred diploid *S. tuberosum* Gp Tuberosum (van Lieshout et al., 2020), multiple diploid and dihaploid *S. tuberosum* Gp. Tuberosum lines (Zhou et al., 2020; Freire et al., 2021; Zhang et al., 2021), and several wild diploid species (Aversano et al., 2015; Leisner et al., 2018). Available tetraploid genome sequences are limited to whole-genome shotgun sequencing data (Hardigan et al., 2017; Pham et al., 2017) or highly fragmented assemblies (Kyriakidou et al., 2020), all lacking phased haplotype data for the four alleles, thereby limiting our understanding of the number and role of deleterious and dysfunctional alleles in each haplotype. To accelerate our understanding of intra-genome variation and to begin to define the pan-genome of tetraploid potato at the gene and allele level, we generated phased genome assemblies of six potato cultivars from North America and Europe, catalogued the gene complement (including the identification of dysfunctional and deleterious alleles and introgressions), identified allelic variation associated with key agronomic traits, and constructed the first pan-genome for tetraploid potato.

RESULTS

Genome assembly of six tetraploid potato cultivars

Haplotype-phased assemblies were generated for six tetraploid cultivars (fresh market [Colomba, Spunta], chip processing [Atlantic], frozen processing [Castle Russet], starch [Altus, Avenger]) using the DeNovoMagic assembly approach (Table 1). Flow cytometry estimates the tetraploid potato genome at 3.1 Gb (Willemsen, 2018), suggesting that the DeNovoMagic assemblies either under-represent the full assembly and/or that some of the haplotypes are collapsed. Comparisons of read depth across the six DeNovoMagic assemblies revealed that a subset of scaffolds represent collapsed haplotypes (Supplemental Figure 1). BUSCO scores (Simão et al., 2015) indicated robust representation of conserved orthologs with a significant number of duplicated orthologs in all six genome assemblies (Supplemental Table 1).

Using phased genetic markers, long Oxford Nanopore Technologies (ONT) genomic reads, and Hi-C data, phased chromosome-scale pseudomolecules were constructed for Atlantic and Castle Russet in which collapsed haplotypes were dosage corrected and chimeric scaffolds resolved. Scaffolds were then ordered and oriented using the phased genetic markers and the DM reference genome (Supplemental Table 2) and Hi-C was used to confirm the scaffolding and correct misassemblies (Supplemental Figures 2 and 3); in total, 54 regions, 41 in Castle Russet, were modified. The chromosome-scale Atlantic assembly represents 2717 Mb with a scaffold N50 length of 36.91 Mb, while the Castle Russet assembly represents 2529 Mb with a scaffold N50 length of 26.00 Mb (Supplemental Tables 2 and 3). Alignment of ONT genomic DNA reads (>10 kb) to the final Atlantic and Castle Russet assemblies revealed a high degree of coverage (99.99% for both assemblies), with 97.8% and 97.5% coverage (Supplemental Table 4). Alignment of the pseudomolecules with the reference genome DM assembly revealed strong collinearity across all of the euchromatic arms except chromosome 12, in which all four homologous chromosomes contained an inversion relative to the DM reference genome (Figure 1A); hereafter, all analyses with Atlantic and Castle Russet were performed with the phased, haplotype-resolved pseudomolecules.

The plastome sequences of these six cultivars assembled into complete circular sequences with the typical quadripartite structure of potato plastomes with sizes ranging from 155 296 bp to 155 562 bp (Supplemental Table 5); pairwise identity between the plastomes was approximately 99.9%. Typing of the plastomes revealed that the Altus, Avenger, and Castle Russet plastomes are of the W type (Hosaka, 1986), while the plastomes of Atlantic, Colomba, and Spunta are of the T type (Hosaka, 1986). Each of the mitogenomes assembled into three individual molecules with total sizes ranging from 453 552 bp to 474 600 bp (Supplemental Table 6), differing in repeat structure

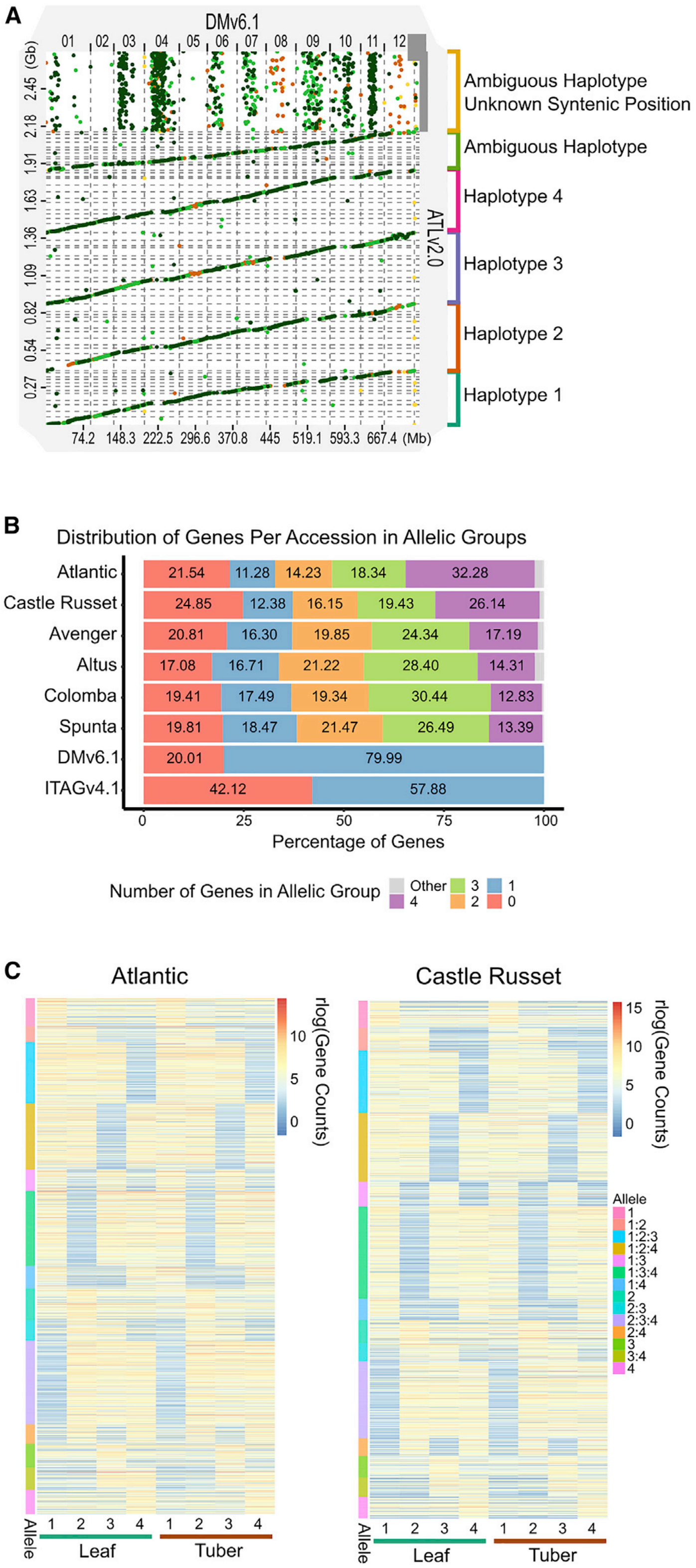

Figure 1. Haplotype and allelic diversity in tetraploid potato.
(A) Whole-genome alignment of the phased Atlantic genome assembly to the *S. tuberosum* Gp Phureja DM 1-3 R144 reference genome.
(B) Distribution of genes within allelic groups in six tetraploid cultivars, *S. tuberosum* Group Phureja DM 1-3, and *S. lycopersicum* Heinz 1706 v4.1. Genes were binned into allelic groups using orthology, paralogy, and synteny using Orthofinder and McScanX.
(C) The DESeq2 rlog-transformed gene expression counts are plotted in a heatmap grouped by the allele(s) preferentially expressed. The columns are the samples, which correspond to the allele number by tissue type, and the rows are the allelic groups. The colors along the side indicate which alleles were identified as being preferentially expressed in the leaf tissue.

(Supplemental Table 7). Each mitochondrial genome encodes 62 non-redundant genes (Supplemental Table 8), except that of Castle Russet, which lacks the *cob-fragment* pseudogene. Phylogenetic reconstruction of both assembled organellar genomes (Supplemental Figure 4) consistently grouped Altus, Avenger, and Castle Russet together, while Atlantic, Colomba, and Spunta formed another group.

Gene content and relatedness among *S. tuberosum* clones

Repetitive sequences in the six tetraploid genome assemblies were identified using a custom repeat library resulting in masking of 65%–68% of the assembly. With respect to transposable elements (TEs), the most abundant classes were retrotransposons (26%) and DNA transposons (1.48%; Supplemental Table 9). Masked assemblies were annotated for protein-coding genes using a combination of *ab initio* gene prediction coupled with transcript evidence from mRNA sequencing (mRNA-seq) and full-length cDNA sequences. Predicted gene and gene model numbers across the six assemblies mirrored the total assembly size reflective of deduplication of collapsed scaffolds in Atlantic and Castle Russet (Supplemental Table 10).

The cultivars represent diverse market classes and multiple geographical regions, enabling investigations into the genetic basis of market class traits and the underlying similarity between them. Homologous groups were identified among the cultivars, DM (Pham et al., 2020), and *S. lycopersicum* Heinz 1706 v4.1 (Hosmani et al., 2019); *Arabidopsis thaliana* TAIR10 (Lamesch et al., 2011) and *Amborella trichopoda* (Amborella Genome Project, 2013) were used as outgroups. A total of 43 475 orthologous and paralogous groups were identified, containing 95.8% of all genes. Between 95.9% and 98.4% of the genes from each cultivar were included in a homologous group, with 0.2%–2.7% of the genes from each cultivar in an accession-specific paralogous group. A phylogenetic tree constructed using orthologous groups containing all accessions (Supplemental Figure 5) corresponded to previously characterized evolutionary relationships. DM is separated from the tetraploid cultivars, reflecting the species sub-structure as DM is classified as Gp Phureja, whereas the tetraploid cultivars are classified Gp Tuberosum (Spooner et al., 2014). The four European cultivars were separated from the North American cultivars and were grouped by market class, with Altus and Avenger (starch), and Colomba and Spunta (fresh market), grouping together. Atlantic (chip) and Castle Russet (frozen fry) grouped together as the North American processing cultivars.

We compared the genomes of our six phased tetraploids with three publicly available high-quality *S. tuberosum* genomes, that of the reference doubled monoploid genome DM (741.6 Mb) (Pham et al., 2020), the phased diploid RH89-039-16 v3 (hereafter RH; 1.67 Gb) (Zhou et al., 2020), and the 12 pseudomolecule consensus dihaploid dAg1 (812 Mb) (Freire et al., 2021) assembly derived from the elite tetraploid cultivar Agria. As expected, representation of BUSCO orthologs in high-confidence gene models mirrored the ploidy and phased state of the assemblies with the diploid phased RH and the six tetraploid assemblies containing a significant number of duplicated BUSCO orthologs (75.5%–90.6%; Supplemental Table 11). In contrast, the doubled monoploid DM contained a mere 1.6% duplicated BUSCO orthologs. Likewise, the increased ploidy in the tetraploid assemblies, particularly in the phased 48-pseudomolecule assemblies of Atlantic and Castle Russet, greatly reduced the number of single-copy BUSCO orthologs to 5.3% and 8.9%, respectively, relative to the diploid RH (22.1%) and the DM reference assembly (91.3%). As all nine assemblies are derived from *S. tuberosum* clones, gene model metrics were highly similar (Supplemental Table 12) with the exception of high-confidence gene number, which was positively correlated with assembly size and extent of resolution of haplotypes within the assembly. Collinearity of genes between the tetraploids and the reference genome DM was higher than that observed between the diploid dAg1 and RH (Supplemental Table 13). This increased degree of collinearity is most likely not due to a closer phylogenetic relationship with DM as dAg1 is derived from an elite tetraploid cultivar; more likely it is due to the presence of additional phased haplotypes in the six tetraploid assemblies that enabled detection of more syntenic blocks than in the 12-pseudomolecule dAg1 and the 24-pseudomolecule phased RH assemblies.

Certainly, the presence of four homologous chromosomes in autotetraploids provides a template for mutation and genetic diversity. To examine the core proteome of potato and how ploidy affects allelic representation, we clustered the predicted proteomes of nine *S. tuberosum* clones (the monoploid [DM], two diploids [RH, dAg1], and six tetraploids [this study]) to define alleles and paralogs; 661 137 genes (95.9%) from 689 385 total genes clustered into 52 720 groups leaving 28 248 genes unassigned to any cluster. The core potato proteome, as defined as clusters containing a minimum of at least one gene/allele from each *S. tuberosum* clone regardless of ploidy, contained 366 319 genes in 16 592 clusters. Within these 16 592 clusters, a mere 2226 had one DM to two RH to four Atlantic and four Castle Russet alleles, consistent with retention of all allelic copies regardless of ploidy. In contrast, 4699 clusters contained one DM to two RH to less than four Atlantic or less than four Castle Russet alleles, suggesting dispensability of alleles in tetraploids but not diploids, while 1925 clusters contained one DM to one RH to less than four Atlantic or less than four Castle Russet alleles ratio, suggesting reduced dispensability in both diploid and tetraploid genomes. This plasticity in allele dosage is due in part to lack of meiosis that would ensure allelic integrity in the gametes coupled with the vegetative propagation of potato, which permits complementation of gene function by the remaining alleles.

Preferential allele expression

Using a combination of orthology, paralogy, and synteny, alleles corresponding to the same gene were identified on each homologous pseudomolecule forming allelic groups; DM was used as the reference genome along with tomato as an outgroup. A total of 55 528 allelic groups containing 469 925 genes (78.0% of all genes) were identified, with 13 157 of the allelic groups containing representation from all eight genomes (six tetraploid cultivars, DM, tomato). With respect to potato, 80% of the DM genes were orthologous and syntenic to at least one of the six cultivars or to tomato. In Atlantic, 32.3% of the genes were syntenic to four genes, thereby representing all four alleles, 18.3% represented the tri-allelic state, 14.2% represented the bi-allelic state, and 11.3% represented the mono-allelic state (Figure 1B). Nearly 22% of the Atlantic genes were not present in an allelic group, attributable to either introgressions that disrupted synteny, diverged coding sequence that limited clustering in orthologous and paralogous groups, and/or structural variation. While a similar distribution of genes within allelic groups was observed for Castle Russet, more tri-allelic states were observed in the other four genomes, reflective of the collapsed nature of these assemblies. Among Atlantic tetra-allelic groups, 56.2% and 82.7% have four distinct CDS and cDNA sequences, respectively (Supplemental Figure 6). Another 36.1% and 14.8% have three distinct CDS and cDNA sequences, respectively. Similar percentages were observed for Castle Russet, highlighting the importance of untranslated regions in allelic diversity. Atlantic and Castle Russet bi-, mono-, and non-allelic groups were enriched for Gene Ontology (GO) 0009607 (response to biotic stimulus) (adjusted p value < 4e-3) and GO:0009605 (response to external stimulus) (adjusted p value < 4e-3), while tri- and tetra-allelic groups were enriched for GO:1901360 (cellular nitrogen compound metabolic process) (adjusted p value < 3e-6) and GO:0044237 (cellular metabolic process) (adjusted p value < 0.01).

Using mRNA-seq data from leaf and tuber tissues, a total of 3728 and 2915 tetra-allelic groups, accounting for 44% and 47% of the tetra-allelic genes, were preferentially expressed in Atlantic and Castle Russet, respectively. Tetra-allelic groups with four distinct cDNA alleles were enriched for preferential allele expression (PAE; adjusted p value < 0.005) while groups with less than four distinct cDNA alleles were significantly depleted of PAE (adjusted p value < 0.007), suggesting increased sub- and neo-functionalization when four distinct cDNA alleles are present. A total of 54% and 59% of the groups with PAE were shared between leaf and tuber tissues, with between 70% and 74% of the shared groups containing the same alleles preferentially expressed in both tissues, respectively (Figure 1C). Groups with PAE in both leaf and tuber tissues were enriched for primary plant pathways including GO:0034641 (cellular nitrogen compound metabolic process) (adjusted p value < 8e-21), GO:0010467 (gene expression) (adjusted p value < 5e-14), and GO:0006412 (translation) (adjusted p value < 2e-10), while groups with PAE unique to the leaf or tuber tissues were enriched for environmental responses. To characterize the genomic context and determine if specific regions are enriched for PAE, the total gene density, repeat coverage, and PAE density were obtained for all 48 Atlantic and Castle Russet phased pseudomolecules (Figure 2; Supplemental Figure 7).

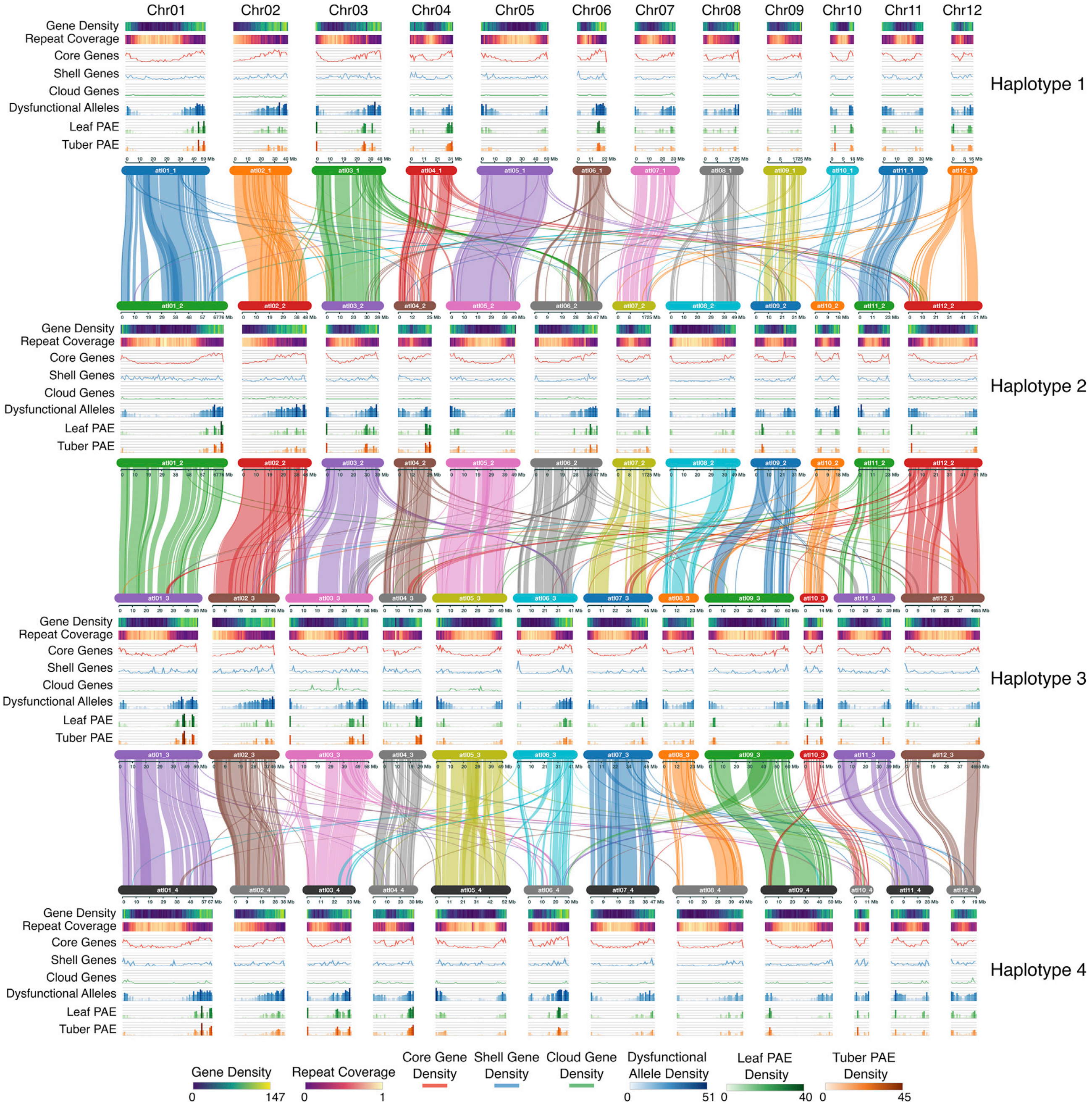

Figure 2. Genomic context of PAE, the pan-genome, and dysfunctional alleles.

Syntenic blocks between the Atlantic phased pseudomolecules are displayed with the density of all loci, repeat coverage, core genes present in all six cultivars, shell genes present in at least two cultivars, cloud genes present in only one cultivar, dysfunctional alleles, PAE in the leaf, and PAE in the tuber displayed as separate tracks.

PAE occurred in discrete genomic positions among the gene-rich regions of the chromosomal arms.

Wild *Solanum* introgression in cultivated potato

Each of the six assemblies exhibited introgressions from wild species (Figure 3). Enrichment of expressed and highly expressed genes in introgression regions was observed in all six genomes; these genes were associated with GO:0033554 (cellular response to stress), GO:0050896 (response to stimulus), and GO:0009914 (hormone transport) (Supplemental Figure 8). Taken together, this evidence points to a major functional role for introgressions in potato, especially related to stress responses, as shown previously (Hardigan et al., 2017). In particular, an introgression of potato virus Y (PVY) resistance from *S. stoloniferum* (Song et al., 2005) is present in Castle Russet. As the genome sequence of *S. stoloniferum* is unavailable, the *PVY* introgression could not be identified using the genetic distance method; however, sequence analysis of the recently cloned Ry_{sto} gene reveals 99.5% and 94.6%

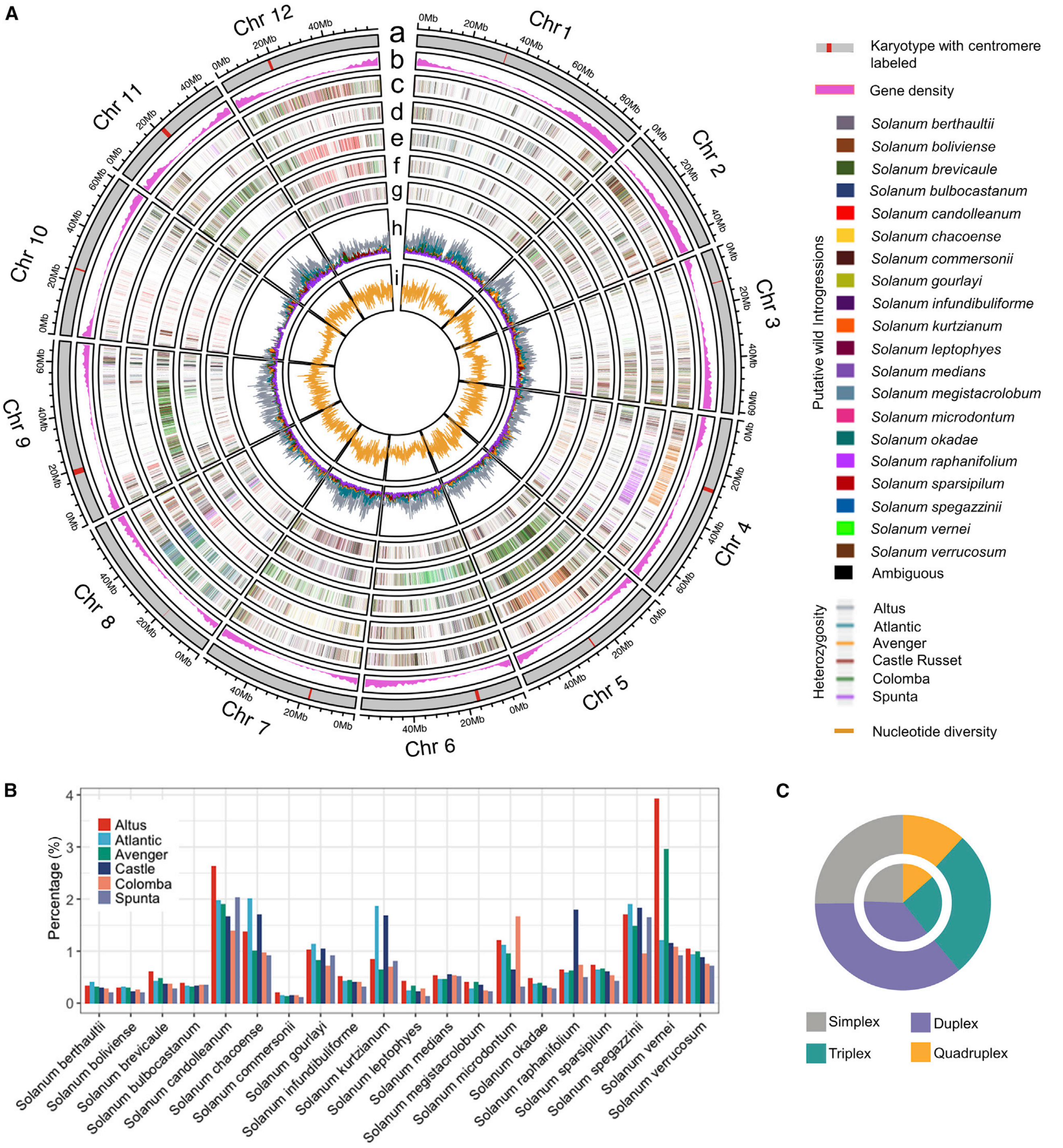

Figure 3. Wild species introgressions shape the cultivated potato genome.

(A) Genetic diversity of the six genome assemblies. Tracks displayed are as follows from outer track to inner track: (a) the potato karyotype with centromere labeled based on DMv6.1 genome annotation; (b) gene density across the genome presented by 1-Mb window size; (c) the distribution of introgressed regions from 20 wild potato species in Atlantic, where introgressions were defined as regions at least 10 kb in length that were more closely related to one of the wild species than a cultivated consensus sequence; (d)–(g) the same graphs as (c) for Castle Russet, Atlus, Avenger, Colomba, and Spunta in order; (h) distribution of heterozygosity for the six genome assemblies in 100-kb window size; (i) nucleotide diversity in 100-kb window size across the six genome assemblies.

(B) The abundance of introgressions (at least 10 kb in length) from each of the 20 wild species in the six cultivars.

(C) Homozygosity of introgressions in two fully phased cultivars. Inner: Atlantic. Outer: Castle Russet.

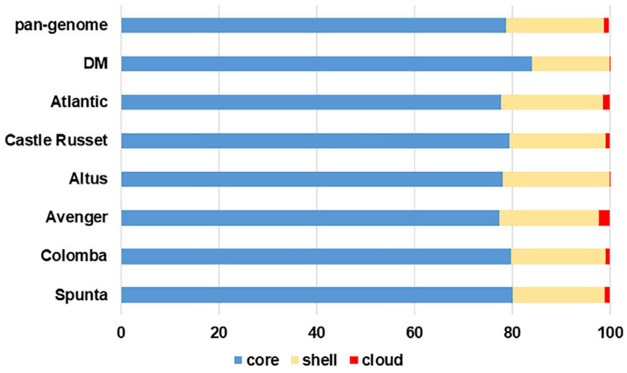

Figure 4. Proportion of core, shell, and cloud genes in the pan-genome and individual cultivar genomes.

Core genes had functions in the biosynthesis of fundamental plant cell structures (Supplemental Figure 10A; Supplemental Table 14). Core and the shell genes were enriched for terms associated with cell wall biogenesis, xyloglucan metabolic process, and hydrotropism (Supplemental Figure 10; Supplemental Table 15). Unlike core genes, shell and cloud genes are considered accessory and were enriched for GO terms involving biological processes in the mitochondria and chloroplasts (Supplemental Figure 10B and 10C; Supplemental Table 14). Of the 908 nuclear pan-genes annotated with functions in these organelles, 25% of them were in the shell and cloud, and 21% had homologs in the organellar genomes (Supplemental Table 14). The shell genes were also enriched for GO terms related to response to auxin, and, together with high-CNV genes, were enriched for defense responses (Supplemental Figure 10B and 10D), consistent with pan-genome analyses in soybean and *Brachypodium* (Gordon et al., 2017; Liu et al., 2020). Variability in genome regions involved in defense was also observed for resistance (*R*) gene clusters, as described below. Among high-CNV and 100+ copies categories, enrichment was found for the hydroxy methylglutaryl (HMG) CoA reductase gene controlling production of glycoalkaloids (Friesen and Rodwell, 2004; Ginzberg et al., 2012).

similarity on chromosomes 12-2 and 12-3 of Castle Russet respectively, chromosomes that show extensive introgressions from *S. chacoense* and *S. vernei* to *S. tuberosum*.

Our results both replicate previously described post-1945 introgressions (Supplemental Figure 8) and point to a longer history of admixture with sympatric species in South America. *Solanum candolleanum* introgressions accounted for more than 1% of all windows tested across all six cultivars, and was the only wild relative to be consistently abundant across all six genomes (Figure 3). *Solanum vernei* and *Solanum spegazzinii* accounted for more than 1% of the windows in five of the six clones. Most introgressions appeared in more than two genome assemblies, and 6.8% were shared across all six genomes (Supplemental Figure 8). Overall, 16.2%, 30.9%, and 50.5% of the introgressions were shared by at least five, four, and three cultivars, respectively. The Atlantic and Castle Russet phased pseudomolecule assemblies were used to determine dosage for the introgressions (Supplemental Figure 8), revealing only a quarter of the detected introgressions present as simplex (Figure 3). Introgressions that appeared in all six assemblies were less likely to be in simplex, especially in Atlantic (Supplemental Figure 8), and the length of an introgression was negatively correlated with dosage and number of assemblies in which an introgression appeared.

The tetraploid potato pan-genome

Potato exhibits intra- as well as inter-genome structural variation. Analyses of the six tetraploid genomes revealed extensive structural variation not only within a single tetraploid genome (i.e., variation with the haplotypes and subgenomes) but also between tetraploid genomes showing a large accessory genome for tetraploid potato. Of the 713 568 total genes annotated in the six genomes, 562 550 were core (genes present in all haplotypes of the six cultivars), 142 225 shell (genes absent in at least one haplotype), and 8793 cloud (genes present in only a single haplotype). Genes in the pan-genome had on average 0.664 copies/genome or 2.65 copies/tetraploid (Supplemental Figure 9), demonstrating the prevalence of missing alleles. Core genes had a higher average of 0.756 copies/genome, with shell and cloud genes showing lower averages of 0.548 and 0.042 copies/genome, respectively. Core genes constitute the majority of genes in the pan-genome for all cultivars and are located in regions of high gene density (Figures 2 and 4). Shell genes are less frequent, followed by cloud, with localization of these classes both within and outside of gene-dense regions (Figures 2 and 4).

Dysfunctional and deleterious alleles

Potato suffers from significant inbreeding depression when selfed (Lindhout et al., 2011; Jansky et al., 2016; Zhang et al., 2019) and we examined the phased Atlantic and Castle Russet genomes for deleterious and dysfunctional alleles. Using tomato as the reference, a total of 24.2% and 22.8% of the Atlantic and Castle Russet loci, respectively, were predicted to contain a deleterious or dysfunctional variant causing a disruptive impact on protein function using SNPeff (e.g., premature stop codon, frame-shift variants). While this is slightly lower than double previous reports based on diploid sequence (Zhou et al., 2020), it is still likely an overestimate. When variants were polarized using *Physalis floridana* (Lu et al., 2021), the potato allele was ancestral 47% of the time. For variants where potato carried the derived allele, 15% were fixed in potato, suggesting a beneficial rather than deleterious mutation (Supplemental Figure 11). However, the majority of variants identified are low frequency, as expected for deleterious mutations, and, in general, patterns of distribution and expression were consistent with the expectations for deleterious alleles. The distribution of putative deleterious/dysfunctional alleles generally mirrored gene content, although the density was variable among the homologous chromosomes (Figure 2; Supplemental Figures 7 and 12). Deleterious/dysfunctional alleles were depleted among genes that had one copy present in the assembly (p value < 3E-16) and enriched among genes that had three or four copies present (p value < 2E-6). Deleterious/dysfunctional alleles were also enriched among core genes present in all six cultivars (p value < 3E-16), and depleted from shell and cloud genes (p value < 4E-7; Figure 2, Supplemental Figure 7). Deleterious/dysfunctional alleles were more lowly expressed compared with all other genes (Wilcoxon test; adjusted p value < 3E-6), with a median $\log_2$ Transcripts Per Kilobase Million (TPM) expression level of 0.57 and 0.31 in Atlantic leaf and tuber tissues, respectively, with similar expression levels observed in Castle Russet. A total of 42.8% and 41.2% of the allelic groups displaying preferential expression patterns in Atlantic leaf and tuber tissue, respectively, contained deleterious/dysfunctional alleles.

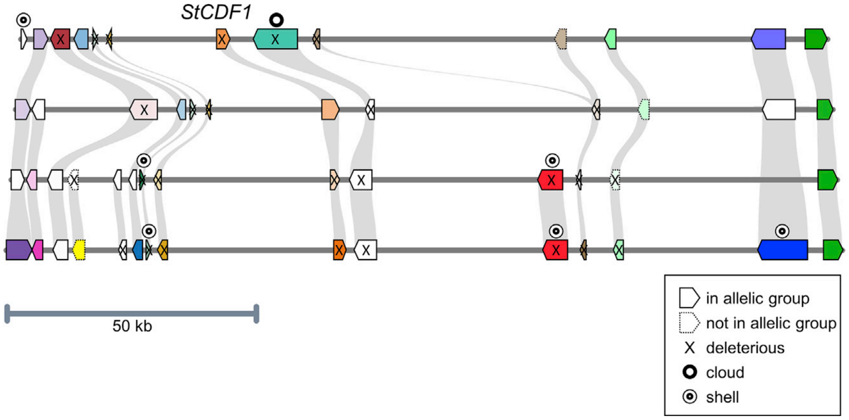

Figure 5. Exemplar locus in tetraploid potato reveal the complexity of gene content, expression, and function.

Genes are outlined in solid boxes if they are in an allelic group and dotted boxes if they are not in an allelic group. Genes with a predicted deleterious/dysfunctional allele are embedded with an X; note that, for *StCDF1*, truncated alleles have an adaptive function. Pan-gene classification is depicted with an open circle (cloud), a bullseye (shell), or no symbol (core). Syntenic genes between haplotypes are shown with gray lines. The Atlantic *StCDF1* locus on chromosome 5 is depicted (orange) on all four haplotypes with adjacent genes. Each allelic group is in a different color with intensity reflecting gene expression in leaf; genes in white boxes have a TPM equal to zero.

Enrichment of deleterious/dysfunctional alleles spanned a range of biological processes with the most significant enrichment in both Atlantic and Castle Russet in genes involved in GO:0006355 (regulation of transcription) (Supplemental Figure 13). Atlantic and Castle Russet are processing potatoes with high specific gravity being a key quality trait. Both cultivars showed enrichment of deleterious/dysfunctional SNPs in genes involved in GO:0010581 (regulation of starch biosynthetic process), suggesting that selection of alleles, including dysfunctional alleles, may have occurred during the breeding process. Tetraploid potato has poor fertility and, in both Atlantic and Castle Russet, an enrichment of dysfunctional and deleterious alleles was observed for genes involved in GO:0048544 (recognition of pollen). In Castle Russet, enrichment was observed for GO:0048235 (pollen sperm cell differentiation), GO:0045691 (regulation of embryo sac central cell differentiation), and GO:0045697 (regulation of synergid differentiation).

We examined the impact of dysfunctional/deleterious mutations on the core tetraploid potato proteome. Within the Atlantic core proteome as defined by OrthoFinder clustering of nine *S. tuberosum* genomes, 61 422 genes were present within 16 592 clusters; of these, 17 583 genes (28.6%) present within 7052 (42.5%) clusters were predicted to have a dysfunctional/deleterious variant, consistent with the overall frequency of dysfunctional/deleterious alleles in the entire Atlantic genome. However, a mere 272 (0.4%) of the genes in the Atlantic core proteome were predicted to be dysfunctional/deleterious and single copy, suggestive that, while a high genetic load is tolerated in tetraploid potato, mutation of all alleles of essential genes is rare.

Insights into potato biology

Maturity locus

The *StCDF1* locus is a master regulator of potato maturity, affecting the timing of tuberization and leaf senescence, which in turn affects yield, tuber quality traits, and field resistance to some diseases (Kloosterman et al., 2013). We identified 12 unique alleles in the 24 sequenced *StCDF1* haplotypes. Nine are classified as wild-type *StCDF1.1* alleles because they encode full-length StCDF1 proteins (Supplemental Table 16), and the remaining three alleles (*StCDF1.2*, *StCDF1.3*, and *StCDF1.4*) encode truncated proteins due to structural variation in the second exon (Supplemental Table 17). Previous research in diploid potato has shown that genotypes with truncated variants have earlier maturity (Kloosterman et al., 2013), and the allelic composition of the tetraploid genotypes in this study suggests a quantitative effect related to allele dosage. Two cultivars, Altus and Avenger, contain only *StCDF1.1* alleles and show very late maturity at temperate latitudes. On the other extreme, Atlantic and Colomba have truncated alleles in three of the four homologs and are well known for their early maturity. The two other cultivars have intermediate maturity and intermediate dosage of truncated alleles (Spunta, 2 copies; Castle Russet, 1 copy; Supplemental Table 16). From a segregating population of 967 F1 progeny of Altus and Colomba, we estimated the effect of *StCDF1.3* on early maturity was stronger than *StCDF1.2* (Supplemental Figure 14). This result is consistent with the recent discovery of a long non-coding RNA (lncRNA) antisense transcript at the *StCDF1* locus (Ramírez Gonzales et al., 2021), which is disrupted in the *StCDF1.3* allele.

Steroidal glycoalkaloids

Selection against total glycoalkaloid content in potato tubers during the domestication process can be observed in both regulatory and functional *GLYCOALKALOID METABOLISM* (*GAME*) genes (Hardigan et al., 2017), and sequence diversity was observed in all six cultivars at *GAME9*, and other members of the *GAME9* pathway, including squalene synthase 1 (*SQS1*), *Δ(7)-STEROL-C5(6)-DESATURASE* (*C5-SD*), and *HMG CoA reductase 1* (*HMGR1*) (Supplemental Table 18, Supplemental Figures 15 and 16). In contrast, the greatest sequence diversity within and across cultivars was observed for *MYC2*, *squalene epoxidase* (*SQE-SQO1*), and *sterol side chain reductase 2* (SSR2; Supplemental Table 18). SQE-SQO1 and SSR2 are key enzymes involved in production of the glycoalkaloid precursors cycloartenol and cholesterol, respectively (Ginzberg et al., 2012; Sawai et al., 2014). Atlantic and Castle Russet each contain four unique SQE-SQO1 protein sequences (eight total), suggesting reduced selection in genes involved in biosynthesis of glycoalkaloid precursors.

P *locus: Production of the purple pigment delphinidin*

The potato *P* locus gene *StPu* (Soltu.DM.11G020990) encodes for a flavonoid 3′,5′-hydroxylase (F3′5′H) (Jung et al., 2005), and, without a functional F3′5′H, potato is not able to produce purple color in tuber skin or any other tissue (e.g., flowers, flesh, light sprout). Analysis of the *P* locus revealed two alleles in Altus, three in Atlantic and Colomba, and four in Spunta, Avenger, and Castle Russet (Supplemental Table 19). In total, nine different alleles were identified in these six cultivars (Supplemental Table 20). Six out of these nine F3′5′H alleles encode truncated proteins compared with its tomato homolog (NP_001234840) or the *Petunia* homolog (P48419), suggesting that these are null alleles (Supplemental Table 21). Furthermore, several of the alleles found in the six cultivars are also present in red-skinned cultivars, including Alouette, Asterix, and Bildstar (Supplemental Table 20), which are known to not contain functional F3’5’H, suggesting that they are indeed null alleles. An allele identified in Spunta (allele 7) was shown to produce purple-skinned tubers in the tetraploid cultivar Bintje (Jung et al. (2009), suggesting that allele 7 is functional.

R1 *gene cluster*

Resistance (*R*) gene clusters are complex loci as they frequently contain paralogs of canonical nucleotide-binding domain and leucine-rich repeat (NLR) genes and TEs. Access to long and haplotype-resolved assemblies permitted a deeper interrogation of the structural organization of the late blight resistance gene *R1* cluster (Ballvora et al., 2002). The presence of *R1* in Atlantic was confirmed using diagnostic markers and agroinfiltration of the cognate *Avr1* gene from *Phytophthora infestans* (data not shown). Atlantic long nanopore genomic reads confirmed correct assembly of the NRGene scaffolds 1390 (*R1*) and 7366 (*r1*). Two other *r1* scaffolds (491 and 1389) were incomplete and chimeric. The chimeric parts were trimmed, and the longest parts of these two assemblies, supported by nanopore reads, were retained. The two complete and two partial *R1* and *r1* clusters from Atlantic were aligned to the previously assembled *R1* cluster (EF514212) and the *r1* cluster from DM and Solyntus (van Lieshout et al., 2020); NLR and TE annotations were included (Supplemental Figure 17). The organization of the sequences flanking the *R1* and *r1* clusters was conserved. In the central portion of the *R1* and *r1* clusters, significant rearrangements of NLRs and TEs were observed. The *R1* clusters were quite similar but had several large indels. Interestingly, the *R1* clusters were almost 100 kb smaller than the completely assembled *r1* cluster from Atlantic, Solyntus, and DM. Moreover, the *R1* clusters exhibited an inversion relative to the *r1* clusters. The *R1* gene was at the edge of this inversion. This context allowed us to pinpoint the *r1* alleles from Atlantic, Solyntus, and DM, which were all annotated as pseudogenes. We hypothesize that a common *R1* ancestor existed that lost its function in the *r1* haplotypes or that it acquired novel activity in *R1*, potentially resulting from the inversion event.

Integrated perspective on the tetraploid potato genome

Deleterious alleles were enriched among core genes and dispersed throughout the genome, similar to results reported in the RH diploid genome (Zhou et al., 2020) and a set of diploid clones (Zhang et al., 2021), which complicate strategies to purge deleterious alleles in both diploid and tetraploid breeding efforts. Intriguingly, the top of chromosome 5, which contains *CDF1*, the locus controlling maturity and yield (Kloosterman et al., 2013), was enriched for deleterious alleles in all four haplotypes in both Atlantic and Castle Russet (adjusted p value < 2.2e-3) (Supplemental Table 22; Figure 5), suggesting an inability to purge deleterious alleles in this region potentially due to selection by breeders for specific maturity (*CDF1*) alleles.

Allele copy number and differential allele expression contribute to dosage of transcripts and encoded proteins. Interestingly, both shell and cloud genes tend to be haplotype-specific genes and are depleted among allelic groups with more than two alleles (adjusted p value < 2.2e-16). Shell genes were also enriched for leaf PAE (adjusted p value < 1.4e-3) and pseudomolecule arms which were enriched for both leaf and tuber PAE were also enriched for shell and cloud genes (Supplemental Table 22). The large overlap of PAE in leaf and tuber tissues, 71.9% overlap in Atlantic and 75.0% overlap in Castle Russet, suggests a common genetic factor regulating allele expression patterns. Among tetra-allelic pairs displaying PAE, no haplotype was consistently preferentially expressed among all allelic pairs, in line with the autopolyploid nature of potato. However, four haplotypes in Atlantic had regions that were significantly reduced in PAE (Supplemental Table 23); three of these four regions were also enriched in cloud genes. Similarly, in Castle Russet, four haplotypes had regions that were significantly reduced in PAE; of these, three were enriched in dysfunctional/deleterious alleles, while one was enriched in cloud genes (Supplemental Table 23). Interestingly, in Castle Russet, there were two regions that were enriched for PAE, chr01_3b and chr05_1b, with 33% and 36% of genes in this haplotype exhibiting PAE in the leaf, and 33% and 37% exhibiting PAE in the tuber, respectively. For chr01_3b, there was no enrichment for deleterious/dysfunctional alleles or cloud genes; in chr05_1b, while there was no enrichment for deleterious/dysfunctional alleles, it was enriched in cloud genes. These data suggest that deleterious/dysfunctional alleles and cloud genes are drivers of localized preferential expression of a haplotype.

DISCUSSION

The plastic and dynamic tetraploid potato genome

The reference genome for potato, the doubled monoploid DM, encodes 32 917 high-confidence genes (Pham et al., 2020), yet only 80.1%–86.7% of the theoretical tetraploid gene complement was present in the phased Atlantic and Castle Russet genome assemblies. This affected dosage as only one-third of the genes in Atlantic were present in a tetra-allelic group due to introgressions and intra-genome structural variation that disrupted colinearity among the four homologs. Intra-genome allelic variation was also high within Atlantic tetra-allelic groups having four distinct coding and cDNA sequences, respectively. Thus, unlike homozygous diploid species in which the two alleles encode for an identical protein, or in the case of hybrid maize where there may be up to two protein isoforms, tetraploid potato has a highly complex transcriptome and predicted proteome.

Unlike allopolyploids, which frequently exhibit subgenome dominance, autopolyploids do not exhibit global preferential expression for a particular haplotype. However, gene-level PAE has been observed in potato (Pham et al., 2017; Zhou et al., 2020), which has critical implications for breeding and gene editing-enabled variety improvements. Interestingly, genes within tetra-allelic groups were enriched in PAE consistent with sub- and neo-functionalization at the expression level. Genes with PAE in both leaf and tuber tissues were enriched in primary plant pathways, potentially reflecting selection pressures imposed by breeders for optimal alleles functioning in potato productivity and quality traits. Not all PAE was shared among leaves and tubers, and genes with PAE in just the leaf or tuber were enriched for functions in environmental response pathways, potentially reflective of wild species introgressions, which are often utilized to incorporate disease resistance into cultivated potato (Hermsen and Ramanna, 1973; Helgeson et al., 1998; Rakosy-Tican et al., 2020).

As a vegetatively propagated autopolyploid with limited meiosis, deleterious and dysfunctional alleles are not purged from the genome and therefore accumulate, contributing to substantial genetic load and inbreeding depression. An increase in the frequency of dysfunctional and deleterious alleles in the tetraploids Atlantic and Castle Russet relative to the diploid RH clone (Zhou et al., 2020) suggests ploidy negatively affects the mutation load. Deleterious/dysfunctional mutations, like preferentially expressed alleles, were more frequent in genes within multi-allelic groups versus single-copy groups, suggesting that functional alleles may be rescuing dysfunctional alleles. One biological process enriched in deleterious/dysfunctional alleles was transcriptional regulation. While divergence in promoter sequences and transcription factors themselves can contribute to differential transcription rates among alleles, differential doses of functional transcription factors can also affect gene expression as well as co-expression networks. An enrichment in variants with a high negative impact in genes involved in transcriptional regulation is consistent with the substantial variation in PAE in tetraploid potato shown in this study and previously (Pham et al., 2017).

Collectively, this intra- and inter-genome variation reinforces that potato is one of the most complex genomes of any crop species analyzed to date. As a consequence, unlike in homozygous diploid crop species such as maize, eight distinct alleles per locus are possible in a tetraploid cross, making optimal combinations of alleles challenging.

Cultivated potato genome is a mix of introgressions from many wild relatives

Wild potato species are typically found from 2000 to 4000 m in altitude throughout the Andes, with Peru having not only the most wild species but also the rarest. While many wild potato species are endemically narrow and located within a distance of less than 100 km, areas of high species richness in the Andes have been observed (Spooner and Hijmans, 2001). Geographically overlapping wild species have been documented in Peru, Bolivia, and Argentina with up to 15 species within a 50 × 50 km grid (cell). In Ecuador and Mexico, nine or more species were present within a single cell (Spooner and Hijmans, 2001). Overlapping geographic distributions between wild and cultivated species, along with the occurrence of the production of 2n pollen and 2n eggs (Watanabe and Peloquin, 1991), facilitate exchange of alleles from wild species to cultivated taxa. Introgression patterns, observed in the pan-genome, specifically the prevalence of small introgressions from a variety of sources, the sharing of small introgressions across cultivars, the high dosage of introgressions and shared introgressions in particular, and the tendency of introgressed genes to be highly expressed, all point to a much longer history of admixture with sympatric species in South America following domestication. Indeed, the tetraploid *S. tuberosum* has previously been described as ‘‘a genetic sponge, absorbing genes from closely related wild species or other cultivated populations with which it hybridizes’’ (Grun, 1990). Introgression regions were enriched in genes that function in adaptation to abiotic and biotic stress, consistent with selection of increased fitness in diverse environments, suggesting that the history of adaptive introgression in potatoes predates the 1940s, before introduction of disease resistance alleles from wild potatoes became an essential breeding tool.

Phased assemblies reveal the allelic complexity underlying important agronomic traits

The neofunctionalized photoperiodic pathway co-opted to regulate tuber formation contains many genes sequentially interdependent on each other's expression at both transcriptional and post-transcriptional levels. The large variation in the upstream transcriptional regulator *StCDF1* not only affects tuberization and plant maturity but is also a major determinant of yield and affects numerous other traits, such as abiotic stress tolerance and disease resistance. This broad effect is likely to be linked to the multiple downstream genes *StCDF1* regulates, and due to the complexity of the locus with its two divergently expressed RNAs, one protein-coding (*StCDF1*) and the other non-coding (*StFLORE*). The allelic variation in the *StCDF1* locus points to an evolutionary history involving a wild-type progenitor, which is disrupted by the insertion of a transposon (leading to a protein truncation and a non-functional *StFLORE*) followed by excision of the transposon, leaving a lesion that maintains the truncation but restores the transcription of the lncRNA. The likelihood of multiple transposition events in the critical site in *StCDF1* between the *StGI* and *StFKF1* binding sites is small. The lack of similar insertions in related species, such as the tomato homolog *SlCDF4*, indicates that this site is not likely to be a transposon hot spot. However, it is intriguing that *StCDF1.4* significantly differs from both *StCDF1.3* (the likely progenitor) and other *StCDF1.2* excision alleles, suggesting a separate insertion event. A wider screen of allelic variants will definitely establish whether the *StCDF1.3* transposon insertion event was unique or occurred multiple times.

Breeding for disease resistance strongly relies on the introgression of *R* genes from wild *Solanum* species. To date, a number of *R* genes have been identified, including genes that confer resistance to *P. infestans*, the causal agent of late blight, which remains a significant worldwide disease of potato. Even though 17 functionally distinct *R* genes conferring resistance to late blight, located in 11 different NLR clusters, have been identified (Monino-Lopez et al., 2021), it remains a challenge to combine these *R* genes into a single genotype. In tetraploid potato, four different *R* alleles from the same locus can be stacked, yet, with diploid breeding, only two alleles can be stacked. Thus, there is an emerging need to combine *R* genes from different sources into one haplotype. To achieve this, meiotic recombinations between the *R* genes must be selected, and therefore knowledge of *R* gene order within haplotype-specific assemblies is essential. As shown in this study, haplotypes of the *R1* cluster, which has a limited number of paralogs, were only partly resolved due to limitations of the NRGene assembly approach. Improvements in sequencing technologies, which traverse the repetitive *R* gene clusters, will reveal the structure of *R* gene clusters not only in tetraploid cultivars but also in key wild *Solanum* species that carry disease resistance traits.

Harnessing tetraploid genome knowledge for diploid, inbred/F1 hybrid breeding

Since the release of the draft potato genome in 2011, significant advances in our understanding of the potato genome have altered not only breeding methods but also the paradigm for breeding. Genomic analyses of wild species, landraces, and modern breeding cultivars have revealed allelic diversity within potato germplasm, loci and alleles associated with important agronomic traits, the genetic basis of inbreeding depression, and genes associated with domestication (Hardigan et al., 2017; Li et al., 2018; Zhang et al., 2019; Zhou et al., 2020). This knowledge has permitted application of genomic selection, marker-assisted selection of key disease resistance loci, and gene editing of loci associated with agronomic traits (Li et al., 2013; Clasen et al., 2016; Sverrisdóttir et al., 2017, 2018; Enciso-Rodriguez et al., 2018, 2019; Endelman et al., 2018; Stich and Van Inghelandt, 2018; Ye et al., 2018; Byrne et al., 2020).

The complexity of haplotypes within tetraploid potato as revealed through access to two phased, chromosome-scale assemblies has revealed the challenges faced by breeders in purging non-functional alleles due to the restricted number of meioses that limit recombination, the presence of deleterious alleles in repulsion, and the complication of tetravalent pairing, all which limit not only the selection of optimal haplotypes but also purging of deleterious/dysfunctional alleles. As shown by Zhang et al. (2021), it is possible to design near-inbred diploid potato lines that, when crossed, yield heterosis. A key component of this success was the use of near-inbred diploid germplasm that reduces allelic diversity, access to fertile lines to permit hybridization, and knowledge of dysfunctional/deleterious alleles of genes controlling key agronomic traits. The complexity of the tetraploid potato genome as revealed through access to the six phased genome assemblies in this study indicates that performing genome-enabled design of improved tetraploid cultivars will be a significant challenge. However, access to phased tetraploid potato genomes with highly divergent haplotypes will aid in efforts to select dihaploids with the maximal combinations of good alleles, which, when coupled with gene editing to restore gene function, provides a path for harnessing genomics in the development of potato as a diploid inbred/F1 hybrid crop (Jansky et al., 2016; Zhang et al., 2021).

METHODS

Germplasm description

The selection of cultivars used for sequencing reflects tetraploid genome diversity bearing tubers for different market classes (fresh market [Colomba, Spunta], chip processing [Atlantic], frozen processing [Castle Russet], starch [Altus, Avenger]), geographic breeding program origins (European and North American), along with modern and heirloom cultivars (twentieth and twenty-first century). Five of the six cultivars (excluding Spunta) were developed using introgression breeding to combat pathogens (Supplemental Table 24; Supplemental Figure 18).

Genome sequencing, assembly, and pseudomolecule construction

Phased assemblies were generated for six tetraploid cultivars using the DeNovoMagic (NRGene, Nes Ziona, Israel) approach (see supplemental information for more detail). For Atlantic and Castle Russet, scaffolds greater than 10 kb from each phased assembly were aligned to the DM reference genome (Pham et al., 2020) and WGS read depth (supplemental information; Supplemental Table 25; Supplemental Figure 19) was used to infer collapsed haplotypes and to detect and correct chimeric scaffolds. For Atlantic and Castle Russet, Oxford nanopore genomic DNA reads were used with LRScaf (Qin et al., 2019) to group scaffolds and confirm syntenic positions and haplotype assignments. LRScaf groups, synteny data, and haplotype assignments were combined to infer preliminary scaffold ordering and orientation, generating 48 phased pseudomolecules and 12 unphased pseudomolecules (v0.9). Hi-C was used with Juicebox v1.11.08 (Robinson et al., 2018) to curate the v0.9 assembly, fixing scaffold order and orientation in the phased pseudomolecules to generate assembly v2.0. BUSCO (Simão et al., 2015) was used to determine the completeness of the gene content.

Transposable element analyses

De novo repetitive sequences were identified in the phased assemblies using RepeatModeler v1.0.11 (https://www.repeatmasker.org/RepeatModeler/) and MITE-Hunter v.2010 (Han and Wessler, 2010). The resulting libraries were concatenated with known potato and tomato (Mueller et al., 2005) and Viridiplantae (RepeatMasker), and used for annotation with RepeatMasker v4.7.0 (repeatmasker.org). Sequences matching the transposase database (www.hrt.msu.edu/uploads/535/78637/Tpases020812.gz) were considered transposons and classified into families. The final repeat library was screened for gene fragments by searching against the plant protein database (retrieved from www.hrt.msu.edu/uploads/535/78637/alluniRefprexp070416.gz); significant hits to genes were removed, along with 100 bp upstream and downstream of the blast hit.

Genome annotation

Transcript evidence to support annotation was generated from diverse tissues using mRNA-seq, Oxford Nanopore Technologies cDNA sequencing, and PacBio Iso-Sequencing (Supplemental Table 25). The six tetraploid assemblies were annotated using a previously established pipeline for potato (Pham et al., 2020). Functional annotation was assigned using alignments to the *Arabidopsis* proteome, Pfam domains, and expression data as described previously (Pham et al., 2020). The working gene model protein sequences were searched against the TAIR10 proteome using BLASTP v2.9.0 (Camacho et al., 2009) with an e-value cutoff of 1e-5. GO terms were assigned using the TAIR GO annotations (Berardini et al., 2004) using the top scoring match for each gene model. GO terms were then slimmed using Plant GOSlim (Gene Ontology Consortium, 2021) and map2slim, part of the go-perl package (v0.15).

Organellar genome assembly, annotation, and phylogenetic analyses

Both mitochondrial and plastid genomes were assembled from WGS reads and annotated using pipelines described previously (Achakkagari et al., 2020, 2021). Coding sequences of 34 protein-coding mitochondrial genes were extracted and concatenated from each mitogenome, while complete plastome sequences were used to construct the phylogenetic trees. Chloroplast DNA types were identified by looking for specific polymorphisms previously reported (Achakkagari et al., 2020, 2021).

Diversity and introgression analyses

Nucleotide diversity, heterozygosity, and runs of homozygosity were measured. Introgressions were identified by comparison with a hypothetical ancestor constructed from nine published diploid landraces (PRJNA378971) and with published sequence data from wild *Solanum* species (PRJNA378971 and PRJNA394943; Hardigan et al., 2017; Li et al., 2018) following methods described previously (Hardigan et al., 2017) (supplemental information). Consecutive windows at least 10 kb in length and of the shortest distance to a single wild species were considered as putative introgressions from the corresponding species. Expressed genes were defined as those with a fragments per kilobase exon model per million mapped reads (FPKM) ≥ 1 and highly expressed genes defined as those with an FPKM ≥ 10.

Pan-genome analyses

Pan-genomes were analyzed for variation in gene copy number using the annotated representative CDS sequences from each of the six tetraploid genomes as well as the reference genome DM v6.1. Genes were categorized as core, shell, and cloud as follows: core genes had at least one copy in each cultivar, shell genes were absent in at least one cultivar, and cloud genes were present in a single cultivar only. Functional analysis of gene groupings was done using topGO v2.42.0 (Alexa et al., 2006). The weighted topGOFisher analysis was used to identify enriched GO terms with *p* values $\leq$ 0.0001 for core, shell, and high copy number variant (CNV) genes. GO enrichment for cloud genes and 100+ genes were filtered at *p* value $\leq$0.01. Further reduction and visualization of GO terms was done using Revigo (Supek et al., 2011).

Deleterious and dysfunctional allele identification

The haplotype-resolved pseudomolecules of Atlantic and Castle Russet were aligned individually to the tomato SL4.0 genome assembly using nucmer v4.0.0rc1 with the default parameters (Marçais et al., 2018). Variants in the VCF files were annotated using SNPeff v4.3.1 (Cingolani et al., 2012) using a custom database with the ITAG v4.1 tomato annotation; only variants with high impact were retained. For GO enrichment tests, the *weight01* algorithm and Fisher's exact test were implemented in the R package topGO v2.34.0 (Alexa et al., 2006) using all genes with GO terms as background. Allele frequency of the Atlantic and Castle Russet high-impact variants were examined across all six phased tetraploid genomes, by aligning the Altus, Avenger, Colomba, and Spunta scaffolds to tomato using nucmer as described above. The alignments were converted to BAM format and the alleles for each clone were extracted from the BAM file only at positions determined as high-impact substitutions in Atlantic and Castle Russet due to the high sequence diversity across the clones. *Physalis floridana* was aligned to the tomato genome as above and used to determine the ancestral and derived alleles for the identified variants for which there was a single clear alignment.

Preferential allele expression

OrthoFinder v2.5.1 (Emms and Kelly, 2015, 2019) was run with default parameters and the representative protein models from the six tetraploid cultivars, DM (v6.1) (Pham et al., 2020), *S. lycopersicum* (ITAG4.1) (Hosmani et al., 2019), *A. thaliana* (TAIR10) (Lamesch et al., 2011), and *A. trichopoda* (Amborella Genome Project, 2013) to determine the paralogous and orthologous groups. The accession phylogenetic tree was generated using OrthoFinder. Using both the OrthoFinder and MCScanX results, allelic groups were defined as genes that were present in the same orthologous group and were syntelogs.

The mRNA-seq reads from leaf and tuber tissue for Atlantic and Castle Russet were processed with Cutadapt v2.10 (Martin, 2011) and high-confidence cDNA models for Atlantic and Castle Russet were used to determine expression abundances. PAE was identified using DESEq2 v1.24.0 (Love et al., 2014) with the Kallisto v0.46.0 (Bray et al., 2016) estimated count values and a likelihood ratio test. The four alleles were then contrasted in a pairwise fashion to obtain the log-2 fold changes (LFCs) using an alpha level of 0.01. An allele was characterized as preferentially expressed if it had an LFC > 2 compared with another allele and an adjusted *p* value < 0.01. GO enrichment tests were performed using topGO (v2.36.0) and a Fisher's exact test with the classic method. The *p* values were adjusted using a false discovery rate (FDR) correction and GO terms with an adjusted *p* value <0.01 were retained. All other categorial enrichment tests were performed using a chi-square test in R (v3.6.2) and *p* values were adjusted using an FDR correction. Heatmaps were generated using pheatmap (v1.0.12). Total gene, TE, core gene, shell gene, cloud gene, dysfunctional allele, and PAE density were calculated in 1 Mb windows using karyoploteR (v1.10.5) (Gel and Serra, 2017) and the results were visualized as tracks in SynVisio (https://synvisio.github.io/#/) with the MCScanX results obtained previously.

Glycoalkaloid pathway analysis

Orthologous gene clusters were identified with OrthoFinder v2.2.6 (Emms and Kelly, 2019) using high-confidence representative protein sequences generated for the six cultivars, protein sequences for *S. lycopersicon* cultivar Heinz, as well as protein sequences for DM (v6.1) (Pham et al., 2020). Extensive manual curation was performed to mine orthologous groups for known regulatory and functional genes in the glycoalkaloid pathway (Supplemental Table 18). Amino acid sequences were aligned using Clustal Omega (Sievers and Higgins, 2018). Alignments were visualized in Jalview v2.11.1.3 (Waterhouse et al., 2009).

Allele mining: maturity, *P* locus, and *R1* gene cluster

Homologs of *StCDF1* and flavonoid 3′,5′-hydroxylase in the six assemblies were identified using BLAST searches; the resulting sequences were aligned using Clustal Omega (Sievers and Higgins, 2018) and manually curated. Alleles were translated into proteins and aligned using Clustal Omega (Sievers and Higgins, 2018). Signal peptides (*P* locus) were determined using SignalP 5.0 (Almagro Armenteros et al., 2019). For *StCDF1*, edited sequences from the six cultivars were aligned, excluding introns and indels larger than 100 bp. The phylogenetic tree was constructed with Clustal Omega (Sievers and Higgins, 2018). Microsynteny at the Atlantic *StCDF1* locus was visualized using the python implementation of mcscan in the jcvi v1.1.18 toolkit (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)).Representative gene model CDS were aligned with default settings to identify syntenic genes across haplotypes. Expression estimates were the average of the three leaf tissue Kallisto-generated TPMs.

Presence of the *R1* functional genes in the six cultivars was identified by checking the Illumina reads (Supplemental Table 25) for the presence of the diagnostic *R1* forward primer. *R1* and *r1* clusters were identified with BLAST+ v2.8.1 using the R1 containing the bacterial artificial chromosome clone sequence from *S. demissum* (NCBI: EF514212) and 50-kb conserved flanking sequences extracted from DM v4.04 (Hardigan et al., 2016) as a query and aligned using Mauve as embedded in CLC Genomics Workbench v20.0.4. The NRGENE assembly of Atlantic *R1* and r1 scaffolds was validated using Atlantic Oxford Nanopore reads >10 kb. Scaffolds were annotated using the NLR annotator pipeline (Steuernagel et al., 2020).

DATA AVAILABILITY

The raw sequences described in this study are available via the National Center for Biotechnology Information under the BioProject: PRJNA718240 (see Supplemental Table 25 for all accession numbers). Large data files are available in the Dryad Digital Repository (https://doi.org/10.5061/dryad.3n5tb2rhw), including (1) GO enrichment analysis for genes in introgression regions shared by at least two genome assemblies; (2) Atlantic and Castle Russet genes with high-impact mutations; (3) list of wild species introgressions; and (4) genome assemblies, annotation, expression matrices, and gff files.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

We acknowledge financial support from AAFC-Génome Québec GQ-AAC-2019-2 to M.V.S., Agriculture and Agri-Food Canada Genomics Research and Development Initiative grant number J-002367 to H.H.T. and K.M.G., Compute Canada, Research Portals and Platforms (RPP) award to M.V.S, Compute Canada, Resources for Research Groups (RRG) award to M.V.S., DFG Germany's Excellence Strategy (EXC-2048/1–Project 390686111) to B.U., Dutch TKI top-sector project Novel genetic and genomic tools for polyploid crops (project numbers BO-26.03-009-004 and BO-50-002-022) to P.M.B., European Union's Horizon 2020 research and innovation programme under grant agreement no 862858 (ADAPT) to C.W.B.B., Germany Ministry of Education and Research BMBF FKZ031A536C to B.U., Germany Ministry of Education and Research BMBF FKZ031A536C to M.E.B., GIZ on behalf of the of the Federal Ministry for Economic Cooperation and Development, Germany to D. Ellis and N.L.A., National Science Foundation (IOS 2140176) to C.R.B., National Science Foundation NRT-IMPACTS fellowship (1828149) to N.B., NC Agricultural Research Service to G.C.Y., NC State University to G.C.Y., NSF IOS-1929982 to C.R.B., NWO-domein Toegepaste en Technische Wetenschappen MAMY project ID 16889 to C.W.B.B. and N.L., Potato Variety Management Institute to K.V. and V.S., State of Minnesota, Minnesota Department of Agriculture to L.M.S., the United States-Israel Binational Agricultural Research and Development Funds IS-5038-17C and IS-5317-20C to J.J., Texas A & M University to M.I.V., The Clen P. and Emma L. Atchley Potato Research Faculty Excellence Endowment to J.C.K., University of Maine to E.H.T., USDA AFRI NIFA Pre-doctoral Fellowship project 2019-07160 to N.R.K., USDA Multistate Research Funds accession 1004958 to W.S.D.J., USDA Hatch Act 2019-03162 to C.R.B., USDA NIFA 2020-67034-31731 to G.H., USDA-NIFA 2016-34141-25707 to L.M.S., USDA-NIFA-SCRI 2019-51181-30021 to L.M.S., Dutch TKI top-sector project Genetics Assisted Assembly of Complex Genomes (project number BO-68-001-033-WPR), LWV20.112 Application of sequence-based multi-allelic markers in genetics and breeding of polyploids (project number BO-68-001-042-WPR), and Wageningen UR Plant Breeding to R.F., Wageningen UR Plant Breeding to R.G.F.V., and Washington State Potato Commission to M.J.F. We would like to thank Emily McCoy and Martin Lague for assistance with data analysis.

AUTHOR CONTRIBUTIONS

B.U., C.R.B., D.D., D.El., E.H.T., G.C.Y., G.H., H.J.E., J.B.E., J.C.K., J.V., M.I.V., M.V.S., N.L.A., R.F., R.G.F.V., R.G.N., and S.R.A. designed the experiments. B.V., G.H., H.H.T., J.B.E., J.C.P.H., J.C.W., J.J.C., K.M.G., K.V., L.K., L.M.S., M.J.F., R.F., R.S., R.A.U., S.R.A., and V.S. performed experiments and/or provided data. B.L., B.U., C.W.B.B., C.R.B., D.Es., F.A.F.G., G.H., H.H.T., H.K.T., J.B.E., J.C.W., J.J., J.M.B., J.P.H., L.M.S., M.E.B., M.K., N.L., N.R.K., P.M.B., R.F., R.A.U., S.O., S.R.A., N.B., and X.M. analyzed data. C.R.B., D.El., J.B.E., L.M.S., M.V.S., N.L.A., and R.F. supervised experiments. B.U., C.R.B., C.W.B.B., D.D., D.El., G.C.Y., G.H., H.H.T., J.B.E., J.C.K., J.P.H., J.V., L.M.S., M.E.B., M.K., M.V.S., N.L.A., N.R.K., R.F., R.G.F.V., R.G.N., S.O., S.R.A., V.S., and W.S.D.J. prepared the manuscript.

ACKNOWLEDGMENTS

No conflict of interest is declared.

Received: June 19, 2021
Revised: October 19, 2021
Accepted: January 7, 2022
Published: January 10, 2022

REFERENCES

Achakkagari, S.R., Kyriakidou, M., Tai, H.H., Anglin, N.L., Ellis, D., and Strömvik, M.V. (2020). Complete plastome assemblies from a panel of 13 diverse potato taxa. PLoS One **15**:e0240124.

Achakkagari, S.R., Bozan, I., Anglin, N.L., Ellis, D., Tai, H.H., and Strömvik, M.V. (2021). Complete mitogenome assemblies from a panel of 13 diverse potato taxa. Mitochondrial DNA B Resour. **6**:894–897.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics **22**:1600–1607.

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. **37**:420–423.

Amborella Genome Project. (2013). The *Amborella* genome and the evolution of flowering plants. Science **342**:1241089.

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

Ballvora, A., Ercolano, M.R., Weiss, J., Meksem, K., Bormann, C.A., Oberhagemann, P., Salamini, F., and Gebhardt, C. (2002). The R1 gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. Plant J. **30**:361–371.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. Nat. Plants **6**:914–920.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., et al. (2004). Functional annotation of the *Arabidopsis* genome using controlled vocabularies. Plant Physiol. **135**:745–755.

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

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**:525–527.

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., and Milbourne, D. (2020). Genome-wide association and genomic prediction for fry color in potato. Agronomy **10**:90.

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

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

Clasen, B.M., Stoddard, T.J., Luo, S., Demorest, Z.L., Li, J., Cedrone, F., Tibebu, R., Davison, S., Ray, E.E., Daulhac, A., et al. (2016). Improving cold storage and processing traits in potato through targeted gene knockout. Plant Biotechnol. J. **14**:169–176.

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., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science **338**:1206–1209.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., and Hirsch, C.N. (2021). How the pan-genome is changing crop genomics and improvement. Genome Biol. **22**:3.

Devaux, A., Kromann, P., and Ortiz, O. (2014). Potatoes for sustainable global food security. Potato Res. **57**:185–199.

Diaz, 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.

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

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**:238.

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., and de los Campos, G. (2018). Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). G3 Genes|Genomes|Genetics **8**:2471–2481.

Enciso-Rodriguez, F., Manrique-Carpintero, N.C., Nadakuduti, S.S., Buell, C.R., Zarka, D., and Douches, D. (2019). Overcoming self-incompatibility in diploid potato using CRISPR-Cas9. Front. Plant Sci. **10**:376.

Endelman, J.B., Carley, C.A.S., Bethke, P.C., Coombs, J.J., Clough, M.E., da Silva, W.L., De Jong, W.S., Douches, D.S., Frederick, C.M., Haynes, K.G., et al. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. Genetics **209**:77–87.

Freire, R., Weisweiler, M., Guerreiro, R., Baig, N., Hüttel, B., Obeng-Hinneh, E., Renner, J., Hartje, S., Muders, K., Truberg, B., et al. (2021). Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. G3 **11**:jkab330. https://doi.org/10.1093/g3journal/jkab330.

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

Friesen, J.A., and Rodwell, V.W. (2004). The 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductases. Genome Biol. **5**:248.

Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics **33**:3088–3090.

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. **49**:D325–D334.

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.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., et al. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat. Commun. **8**:2184.

Grun, P. (1990). The evolution of cultivated potatoes. Econ. Bot. **44**:39–55.

Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. **38**:e199.

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.

Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J.C., Douches, D.S., Farré, E.M., et al. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc. Natl. Acad. Sci. U S A **114**:E9999–E10008.

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

Helgeson, J.P., Pohlman, J.D., Austin, S., Haberlach, G.T., Wielgus, S.M., Ronis, D., Zambolim, L., Tooley, P., McGrath, J.M., James, R.V., et al. (1998). Somatic hybrids between *Solanum bulbocastanum* and potato: a new source of resistance to late blight. Züchter Genet. Breed. Res. **96**:738–742.

Hermsen, J.G.T.H., and Ramanna, M.S. (1973). Double-bridge hybrids of *Solanum bulbocastanum* and cultivars of *Solanum tuberosum*. Euphytica **22**:457–466.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell **26**:121–135.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell **28**:2700–2714.

Hosaka, K. (1986). Who is the mother of the potato? - restriction endonuclease analysis of chloroplast DNA of cultivated potatoes. Theor. Appl. Genet. **72**:606–618.

Hosmani, P.S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L.V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., et al. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv, 767764. https://doi.org/10.1101/767764.

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.H., Charkowski, A.O., Douches, D.S., Gusmini, G., Richael, C., Bethke, P.C., Spooner, D.M., Novy, R.G., De Jong, H., De Jong, W.S., et al. (2016). Reinventing potato as a diploid inbred line-based crop. Crop Sci. **56**:1412–1422.

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

Jung, C.S., Griffiths, H.M., De Jong, D.M., Cheng, S., Bodis, M., and De Jong, W.S. (2005). The potato P locus codes for flavonoid 3′,5′-hydroxylase. Theor. Appl. Genet. **111**:184.

Jung, C.S., Griffiths, H.M., De Jong, D.M., Cheng, S., Bodis, M., Kim, T.S., and De Jong, W.S. (2009). The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. Theor. Appl. Genet. **120**:45–57.

Kloosterman, B., Abelenda, J.A., Gomez Mdel, M., Oortwijn, M., de Boer, J.M., Kowitwanich, K., Horvath, B.M., van Eck, H.J., Smaczniak, C., Prat, S., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature **495**:246–250.

Kyriakidou, M., Anglin, N.L., Ellis, D., Tai, H.H., and Strömvik, M.V. (2020). Genome assembly of six polyploid potato genomes. Sci. Data **7**:88.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al. (2011). The Arabidopsis Information Resource

(TAIR): improved gene annotation and new tools. Nucleic Acids Res. **40**:D1202–D1210.

Leisner, C.P., Hamilton, J.P., Crisovan, E., Manrique-Carpintero, N.C., Marand, A.P., Newton, L., Pham, G.M., Jiang, J., Douches, D.S., Jansky, S.H., et al. (2018). Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense*, reveals residual heterozygosity. Plant J. **94**:562–570.

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

Li, Y., Colleoni, C., Zhang, J., Liang, Q., Hu, Y., Ruess, H., Simon, R., Liu, Y., Liu, H., Yu, G., et al. (2018). Genomic analyses yield markers for identifying agronomically important genes in potato. Mol. Plant **11**:473–484.

Lindhout, P., Meijer, D., Schotte, T., Hutten, R.C.B., Visser, R.G.F., and van Eck, H.J. (2011). Towards F1 hybrid seed potato breeding. Potato Res. **54**:301–312.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell **182**:162–176.e13.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**:1–21.

Lu, J., Luo, M., Wang, L., Li, K., Yu, Y., Yang, W., Gong, P., Gao, H., Li, Q., Zhao, J., Wu, L., et al. (2021). The *Physalis floridana* genome provides insights into the biochemical and morphological evolution of *Physalis* fruits. Hortic Res. **8**:244. https://doi.org/10.1038/s41438-021-00705-w.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. **14**:e1005944.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. **17**:10–12.

Monino-Lopez, D., Nijenhuis, M., Kodde, L., Kamoun, S., Salehian, H., Schentsnyi, K., Stam, R., Lokossou, A., Abd-El-Haliem, A., Visser, R.G.F., et al. (2021). Allelic variants of the NLR protein Rpi-chc1 differentially recognize members of the *Phytophthora infestans* PexRD12/31 effector superfamily through the leucine-rich repeat domain. Plant J. https://doi.org/10.1111/tpj.15284.

Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C.W., Wright, M.H., Ahrens, R., Wang, Y., et al. (2005). The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. Plant Physiol. **138**:1310–1317.

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

Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., and Buell, C.R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience **9**:giaa100.

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

Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., and Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. BMC Genomics **20**:955.

Rakosy-Tican, E., Thieme, R., König, J., Nachtigall, M., Hammann, T., Denes, T.-E., Kruppa, K., and Molnár-Láng, M. (2020). Introgression of two broad-spectrum late blight resistance genes, *Rpi-Blb1* and *Rpi-Blb3*, from *Solanum bulbocastanum* Dun Plus Race-specific R genes into potato pre-breeding lines. Front. Plant Sci. **11**:699.

Ramírez Gonzales, L., Shi, L., Bergonzi, S.B., Oortwijn, M., Franco-Zorrilla, J.M., Solano-Tavira, R., Visser, R.G.F., Abelenda, J.A., and Bachem, C.W.B. (2021). Potato CYCLING DOF FACTOR 1 and its lncRNA counterpart StFLORE link tuber development and drought response. Plant J. **105**:855–869.

Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P., and Aiden, E.L. (2018). Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst. **6**:256–258.e1.

Sawai, S., Ohyama, K., Yasumoto, S., Seki, H., Sakuma, T., Yamamoto, T., Takebayashi, Y., Kojima, M., Sakakibara, H., Aoki, T., et al. (2014). Sterol side chain reductase 2 is a key enzyme in the biosynthesis of cholesterol, the common precursor of toxic steroidal glycoalkaloids in potato. Plant Cell **26**:3763–3774.

Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. **27**:135–145.

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

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

Spooner, D.M., and Hijmans, R.J. (2001). Potato systematics and germplasm collecting, 1989–2000. Am. J. Potato Res. **78**:237–268.

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.

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.

Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.-J., Yu, G., Baggs, E., Witek, A.I., Yadav, I., Krasileva, K.V., et al. (2020). The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiol. **183**:468–482.

Stich, B., and Van Inghelandt, D. (2018). Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. Front. Plant Sci. **9**:159.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One **6**:e21800.

Sverrisdóttir, E., Byrne, S., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., Janss, L., and Nielsen, K.L. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theor. Appl. Genet. **130**:2091–2108.

Sverrisdóttir, E., Sundmark, E.H.R., Johnsen, H.Ø., Kirk, H.G., Asp, T., Janss, L., Bryan, G., and Nielsen, K.L. (2018). The value of expanding the training population to improve genomic selection models in tetraploid potato. Front. Plant Sci. **9**:1118.

van Lieshout, N., van der Burgt, A., de Vries, M.E., Ter Maat, M., Eickholt, D., Esselink, D., van Kaauwen, M.P.W., Kodde, L.P., Visser, R.G.F., Lindhout, P., et al. (2020). Solyntus, the new highly contiguous reference genome for potato (*Solanum tuberosum*). G3 **10**:3489–3495.

Watanabe, K., and Peloquin, S.J. (1991). The occurrence and frequency of 2n pollen in 2x, 4x, and 6x wild, tuber-bearing *Solanum* species from

Mexico, and Central and South America. Theor. Appl. Genet. **82**:621–626.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**:1189–1191.

Willemsen, J. (2018). The identification of allelic variation in potato. https://doi.org/10.18174/459655.

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

Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N.R., Shang, Y., Li, C., and Huang, S. (2019). The genetic basis of inbreeding depression in potato. Nat. Genet. **51**:374–378.

Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., Zhu, G., Xiong, X., Shang, Y., Li, C., et al. (2021). Genome design of hybrid potato. Cell **184**:3873–3883.e12.

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G.F., Bachem, C.W.B., Robin Buell, C., Zhang, Z., et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat. Genet. **52**:1018–1023.