the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) **94,** 562–570 doi: 10.1111/tpj.13857

RESOURCE

Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense*, reveals residual heterozygosity

Courtney P. Leisner[1], John P. Hamilton[1], Emily Crisovan[1], Norma C. Manrique-Carpintero[1,2], Alexandre P. Marand[3], Linsey Newton[1], Gina M. Pham[1], Jiming Jiang[3], David S. Douches[2], Shelley H. Jansky[3,4] and C. Robin Buell[1,5,*]

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

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

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

[4]*United States Department of Agriculture-Agricultural Research Service, Vegetable Crops Research Unit, Madison, WI 53706, USA, and*

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

Received 30 October 2017; revised 25 January 2018; accepted 29 January 2018; published online 5 February 2018.
*For correspondence (e-mail buell@msu.edu).

SUMMARY

Cultivated potato (*Solanum tuberosum* L.) is a highly heterozygous autotetraploid that presents challenges in genome analyses and breeding. Wild potato species serve as a resource for the introgression of important agronomic traits into cultivated potato. One key species is *Solanum chacoense* and the diploid, inbred clone M6, which is self-compatible and has desirable tuber market quality and disease resistance traits. Sequencing and assembly of the genome of the M6 clone of *S. chacoense* generated an assembly of 825 767 562 bp in 8260 scaffolds with an N50 scaffold size of 713 602 bp. Pseudomolecule construction anchored 508 Mb of the genome assembly into 12 chromosomes. Genome annotation yielded 49 124 high-confidence gene models representing 37 740 genes. Comparative analyses of the M6 genome with six other Solanaceae species revealed a core set of 158 367 Solanaceae genes and 1897 genes unique to three potato species. Analysis of single nucleotide polymorphisms across the M6 genome revealed enhanced residual heterozygosity on chromosomes 4, 8 and 9 relative to the other chromosomes. Access to the M6 genome provides a resource for identification of key genes for important agronomic traits and aids in genome-enabled development of inbred diploid potatoes with the potential to accelerate potato breeding.

Keywords: *Solanum chacoense*, diploid potato, M6, PRJNA362370, genome assembly, genome annotation, self-compatibility.

INTRODUCTION

Cultivated potato, *Solanum tuberosum* L., grown for the production of below-ground storage tubers, was domesticated nearly 8000 years ago and is currently the world's most important vegetable crop and the fourth most important food crop after rice, wheat and maize (FAOSTAT, 2016). Current breeding foci in potato include disease resistance and quality traits in which breeders incorporate genetic diversity available within extant potato cultivars, as well as from closely related tuber-bearing species. One species, *Solanum chacoense* (Figure 1), is an excellent source of disease resistance and resistance to cold-induced sweetening, a physiological response to cold storage in the tuber in which starch is converted into fructose and glucose that increase fry and chip color, as well as the formation of acrylamide upon frying at high temperatures. However, *S. chacoense* accessions have high levels of toxic steroidal glycoalkaloids in the tubers, thereby requiring further crossing to remove high glycoalkaloid levels in tubers in initial breeder selections (McCue, 2009). Indeed, Lenape, a high-quality chip-processing cultivar released in 1967 (Akeley *et al.*, 1968), was removed from the market in 1970 due to high glycoalkaloid content in the tubers,

Figure 1. Morphology of the plant, leaves and flowers of *Solanum chacoense* M6.
Whole plant (left), leaves (top right) and flowers (bottom right).

probably contributed by *S. chacoense* in its parentage (Anonymous, 1970).

Unlike major cereal and oilseed crops, cultivated potato is a vegetatively propagated heterozygous autotetraploid. Breeding gains in potato are limited by polyploidy, inbreeding depression, low sexual fertility, low recombination and poor adaptation of wild germplasm (The Potato Genome Sequencing Consortium, 2011; Jansky *et al.*, 2014; Hardigan *et al.*, 2017). The creation of diploid inbred lines represents a new strategy for overcoming the challenges with historical tetraploid potato breeding approaches (Birhman and Hosaka, 2000; Phumichai *et al.*, 2005; Jansky *et al.*, 2016). Previous work focusing on the creation of homozygous inbred lines in diploid potato was limited by self-incompatibility in diploid germplasm (Hawkes, 1958; Cipar *et al.*, 1964). However, the presence of a dominant allele for the S-locus inhibitor gene *Sli* (Hosaka and Hanneman, 1998a) enabled inbreeding and the development of the inbred M6 *S. chacoense* clone (Jansky *et al.*, 2014). Interestingly, although the M6 clone has been inbred for seven generations it still exhibits considerable residual heterozygosity, suggesting the possibility that lethal or deleterious alleles could be maintained in repulsion with beneficial alleles (Jansky *et al.*, 2014). The M6 clone has desirable agronomic traits such as high dry matter, good chip-processing qualities and disease resistance. M6 was crossed as a male to the doubled monoploid *S. tuberosum* Group Phureja DM1-3 516 R44 (DM1-3) clone and the F_1 population was self-pollinated to create an F_2 population that permitted identification of a number of quantitative trait loci (QTL) associated with agronomic traits, including skin pigmentation, tuber flesh color, tuber length-to-width ratio, presence of eye tubers, jelly end and anther length (Endelman and Jansky, 2016). Access to a diploid, selfing and inbred clone for potato provides a major resource for shifting potato breeding from the more challenging tetraploid genetic system to one with simple diploid, inbred genetics (Lindhout *et al.*, 2011; Jansky *et al.*, 2016).

Currently, an assembled genome sequence is available for two species of potato, the reference genome of the doubled monoploid *S. tuberosum* Group Phureja DM1-3 (The Potato Genome Sequencing Consortium, 2011) and the diploid wild species *Solanum commersonii* (Aversano *et al.*, 2015). The DM1-3 genotype was selected as the focus of the Potato Genome Sequencing Consortium due to its homozygosity (The Potato Genome Sequencing Consortium, 2011), which significantly reduced the genome complexity permitting assembly with the short-read Illumina technologies available at the time (The Potato Genome Sequencing Consortium, 2011). As a doubled monoploid derived from an adapted diploid *S. tuberosum* Group Phureja clone (Lightbourn and Veilleux, 2007) it provides a reference for cultivated potato that includes *S. tuberosum* Group Phureja, Group Tuberosum, Group Andigena and Group Chilotanum clones. *Solanum commersonii*, a diploid species with disease resistance and cold tolerance, is heterozygous and is enriched in gene families that confer abiotic and biotic stress tolerance (Aversano *et al.*, 2015). To date, no genome sequence is available for tetraploid potato due to its highly heterogeneous genome that is rich in sequence and structural variation (Pham *et al.*, 2017), making it intractable with current genome assembly methods.

In this study we generated the genome sequence of M6, assembled pseudomolecules, identified the fraction of the genome remaining heterozygous following seven generations of selfing, annotated the gene complement of M6, compared the genomes of three potato species and identified key genes and gene clusters important in specialized metabolism in M6, including the glycoalkaloid pathway. These data, along with access to the M6 genome sequence and annotation, will facilitate our understanding of the genes responsible for key agronomic traits in potato, a species critical for world food security.

RESULTS AND DISCUSSION

Genome assembly and assessment

We generated a draft genome assembly of M6 using a suite of paired-end and mate pair libraries and the *de novo* genome assembler ALLPATHS-LG (Gnerre *et al.*, 2011) (Table 1). Following assembly, a total of 16 704 gaps were filled using GapCloser (Luo *et al.*, 2012) with paired-end sequences from three additional paired-end Illumina-compatible libraries not used in the initial assembly (Table 1). Following gap filling, an assembly of 825 767 562 bp in

Table 1 Metrics of libraries used for *de novo* assembly of the *Solanum chacoense* M6 genome

SRA run number	Type of library	No. of input read pairs (trimmed)	Read length (nt)	Fragment size (bp)	Use
SRR5264021	Paired end	128 785 528	100	166	ALL-PATHS
SRR5264022	Paired end	19 961 316	150	186	ALL-PATHS
SRR5264017	Paired end	215 772 234	160	250	ALL-PATHS
SRR5264020	Paired end	144 095 988	150	258	GapCloser
SRR5264019	Paired end	97 729 335	150	422	GapCloser
SRR5264018	Paired end	85 590 735	150	530	GapCloser
SRR5264016	Mate pair	69 401 744	160	3506	ALL-PATHS
SRR5264013	Mate pair	19 617 641	150	6177	ALL-PATHS
SRR5264015	Mate pair	56 697 303	160	6591	ALL-PATHS
SRR5264014	Mate pair	19 880 300	150	9621	ALL-PATHS

SRA, Sequence Read Archive; nt, nucleotides.

8260 scaffolds with an N50 scaffold size of 713 601 bp was generated (Table 2). Flow cytometry estimated the genome size of *S. chacoense* M6 as 882 Mb, similar in size to other estimations of potato genome size (The Potato Genome Sequencing Consortium, 2011; Aversano *et al.*, 2015). The completeness of the genome assembly was assessed by alignment of paired-end reads to the genome assembly using BWA-MEM (Li, 2013) and BUSCO, which identifies the presence of curated plant single-copy orthologs in the genome assembly (Simao *et al.*, 2015). More than 98% of the genomic paired-end read sequences aligned to the assembly, of which, between 92.9 and 98.1% aligned in the proper orientation (Table S1). With respect to the representation of genic sequences in the genome assembly, 96% of the BUSCO core Plantae ortholog genes were represented as full length in the genome assembly, with another 1% present as partial sequences (C:96.0%[S:91.7%,D:4.3%], F:1.0%,M:3.0%,n:1440). While RNA sequencing (RNA-Seq) reads were generated from six tissues of *S. chacoense* M6, we observed significant contamination with Potato Virus X in a subset of the libraries and did not use them for quality assessment of the assembly. Collectively, these metrics suggest that the M6 assembly is highly representative of the M6 genome, especially for genic regions.

Genome annotation

Genome annotation for the draft genome assembly resulted in a set of 53 570 working models, 49 124 high-confidence models representing 37 740 loci and 4446 low-confidence models. With respect to the representation of genic sequences in the annotated gene set, 95.4% of the BUSCO core Plantae ortholog genes were represented as full length with another 2.6% present as partial sequences (C:95.4%[S:66.9%,D:28.5%],F:2.6%,M:2.0%,n:1440).

Construction of pseudomolecules

Scaffolds were anchored to the 12 chromosomes using two genetic maps; one generated from a cross of *S. chacoense* clone *chc*80-1 (USDA 8380-1 from accession 458310) (Sanford *et al.*, 1996) with *S. chacoense* M6 (referred to as chc80-1 × M6) and a second map generated from a cross of DM1-3 by *S. chacoense* M6 (referred to as DM1-3 × M6) (Endelman and Jansky, 2016). Comparison of common markers from these two genetic maps with the DM reference genome sequence revealed high concordance. Across the 12 chromosomes, the placement of scaffolds using genetic map positions in chc80-1 × M6 were, on average, 68% concordant with the placement based on the DM1-3 × M6 positions. Overall, 508 150 181 Mb (62%) of the sequence was anchored to the 12 chromosomes, representing 748 scaffolds and 29 989 high-confidence genes mapping at 99% coverage and identity using GMAP (Wu and Watanabe, 2005). Alignment of the 12 masked *S. chacoense* M6 pseudomolecules to the 12 DM1-3 chromosomes (v.4.04) (Hardigan *et al.*, 2016) showed concordance with the DM1-3 potato genome across all 12 chromosomes (Figure S1 in the online Supporting Information).

Evaluation of the genome landscape

The M6 clone used in this study was an S7 individual, and thus it is possible that some loci remain heterozygous and both haplotypes are present in our assembly. Indeed, analysis of heterozygosity of M6 using a single nucleotide polymorphism (SNP) array revealed 892 (4.8%) heterozygous loci in this S7 generation individual. However, the 8303 loci represented in the SolCAP SNP array (Hamilton *et al.*, 2011) were pre-selected to be polymorphic across potato accessions and thus may be biased in their representation of overall genome heterozygosity. To assess residual heterozygosity on a whole-genome scale we used a stringent set of parameters to limit false positives and identified 1 414 890 biallelic SNPs from a total of 208 Mb of assayable nucleotides (Figure S2), yielding a SNP frequency of 0.68% across the whole genome. Interestingly, heterozygosity was not evenly distributed throughout the genome and enrichment of biallelic loci was evident on

Table 2 Metrics of the *Solanum chacoense* M6 genome assembly

Metric	M6 assembly
Total scaffold size	825 767 562 bp
No. of scaffolds	8260
N50 scaffold size	713 601 bp
NG50 scaffold size[a]	650 452 bp
Maximum length of scaffolds	7 385 816 bp
Minimum length of scaffolds	909 bp
Average scaffold size	99 972 bp
Total anchored scaffold length	508 150 181 bp
Total unanchored scaffold length	317 617 381 bp
No. of gaps in the assembly	29 122

[a]Estimated genome size is 882 Mb.

three chromosomes, chromosome 4 (1.73% heterozygous positions), chromosome 8 (2.37% heterozygous positions) and chromosome 9 (2.10% heterozygous positions), compared with a frequency of 0.26–0.69% heterozygous positions across the other nine chromosomes (Figures 2 and S2). The regions of elevated heterozygosity corresponded to areas of low gene density (genes per Mb), high repeat coverage (% repeats per Mb) and low recombination rates (Figures 2 and S2). Earlier studies in maize and rice reported higher than expected levels of heterozygosity in inbreds, and it was suggested that there may be a selective advantage to heterozygosity in some regions (Cho *et al.*, 1998; McMullen *et al.*, 2009). However, Gore *et al.* (2009) suggested that residual heterozygosity is more likely to be the product of low levels of recombination. It is possible the heterozygous regions have been retained in the *S. chacoense* M6 genome due to beneficial alleles being linked to deleterious alleles and maintained in repulsion. Deleterious alleles would be more abundant in genomic regions of reduced recombination, requiring increased recombination through sexual propagation to purge deleterious alleles and increase homozygosity.

The highly reduced heterozygosity in M6 contrasts with other reports of genome heterozygosity in diploid and tetraploid potato. The first genome-wide survey of heterozygosity in potato utilized the 8303 SolCAP SNP chip, revealing heterozygosity rates of 56% in tetraploid cultivars (Hirsch *et al.*, 2013) and 0.67–37.2% in a *Solanum* sect Petota diversity panel that included wild species relatives and landraces (Hardigan *et al.*, 2014). In the *Solanum* sect Petota diversity panel, the majority of wild species relatives had SNP heterozygosity rates of less than 5%, including *Solanum jamesii*, a highly diverged species from cultivated potato, which had a SNP heterozygosity rate of 0.67%. However, as noted above, these studies utilized a SNP array in which features were pre-selected to be highly polymorphic in potato and thus are biased in representation of true genome heterozygosity. Whole-genome assembly and analysis of heterozygosity of the diploid wild species *S. commersonii* (Aversano *et al.*, 2015) revealed a SNP frequency of 1.49%, a slightly elevated rate of heterozygosity compared with M6. With the advent of inexpensive whole-genome sequencing, less biased estimations of genome heterozygosity are possible via resequencing and read alignments to the *S. tuberosum* group Phureja DM v4.04 reference genome sequence, although technical challenges in read alignments in repetitive or highly diverged regions of the genome limit determination of the true heterozygosity rate. Using a panel of 63 accessions that represent the diversity of potato, including wild species, landraces and cultivars, Hardigan *et al.* (2017) revealed mean heterozygous nucleotide frequencies of 1.05% in diploid landraces and 2.73% in tetraploid cultivars. Thus, while the M6 genome has localized regions of elevated heterozygosity on a subset of chromosomes it has limited overall heterozygosity and provides a robust germplasm for developing inbred lines of potato.

Comparative analyses of three potato genomes

To assess relatedness and identify lineage-specific genes in *S. chacoense*, we identified orthologs and close paralogs with OrthoFinder (Emms and Kelly, 2015) using the predicted proteomes of *Capsicum annum* (pepper; Kim *et al.*, 2014), *Nicotiana benthamiana* (tobacco; Bombarely *et al.*, 2012), *S. chacoense* M6 (wild potato), *S. commersonii* (wild potato; Aversano *et al.*, 2015), *Solanum lycopersicum* (tomato; The Tomato Genome Consortium, 2012), *Solanum melongena* (eggplant; Hirakawa *et al.*, 2014) and *S. tuberosum* Group Phureja DM1-3 (cultivated potato; Hardigan *et al.*, 2016) (Table S2). A total of 262 882 genes (79.8% of the total) were assigned to 23 261 orthogroups, of which 11 398 had representation from all seven species (Table S3, Figure 3). With this set of four non-tuberizing Solanaceae species and three tuber-bearing potato species we were able to identify genes unique to these three tuber-bearing species, with a total of 384 orthologous groups representing a total of 1897 genes that provide candidates for studying the process of tuberization.

Uses for the *S. chacoense* M6 genome sequence and annotation

Glycoalkaloid metabolism is a prevalent pathway in tuber-bearing species, resulting in toxic products in fruit and tubers (Friedman, 2006). Since *S. chacoense* accessions have been shown to contain high levels of toxic steroidal glycoalkaloids in tubers, gene occupancy analysis of orthologous groups containing glycoalkaloid metabolism genes was performed for all seven species. Occupancy analysis of orthologous groups containing genes involved in sesquiterpene synthase revealed increased gene occupancy in the tuber-bearing species, *S. tuberosum* Group *Phureja* (DM1-3), *S. chacoense* M6 and *S. commersonii* relative to other solanaceous species (Figure 4a). Next,

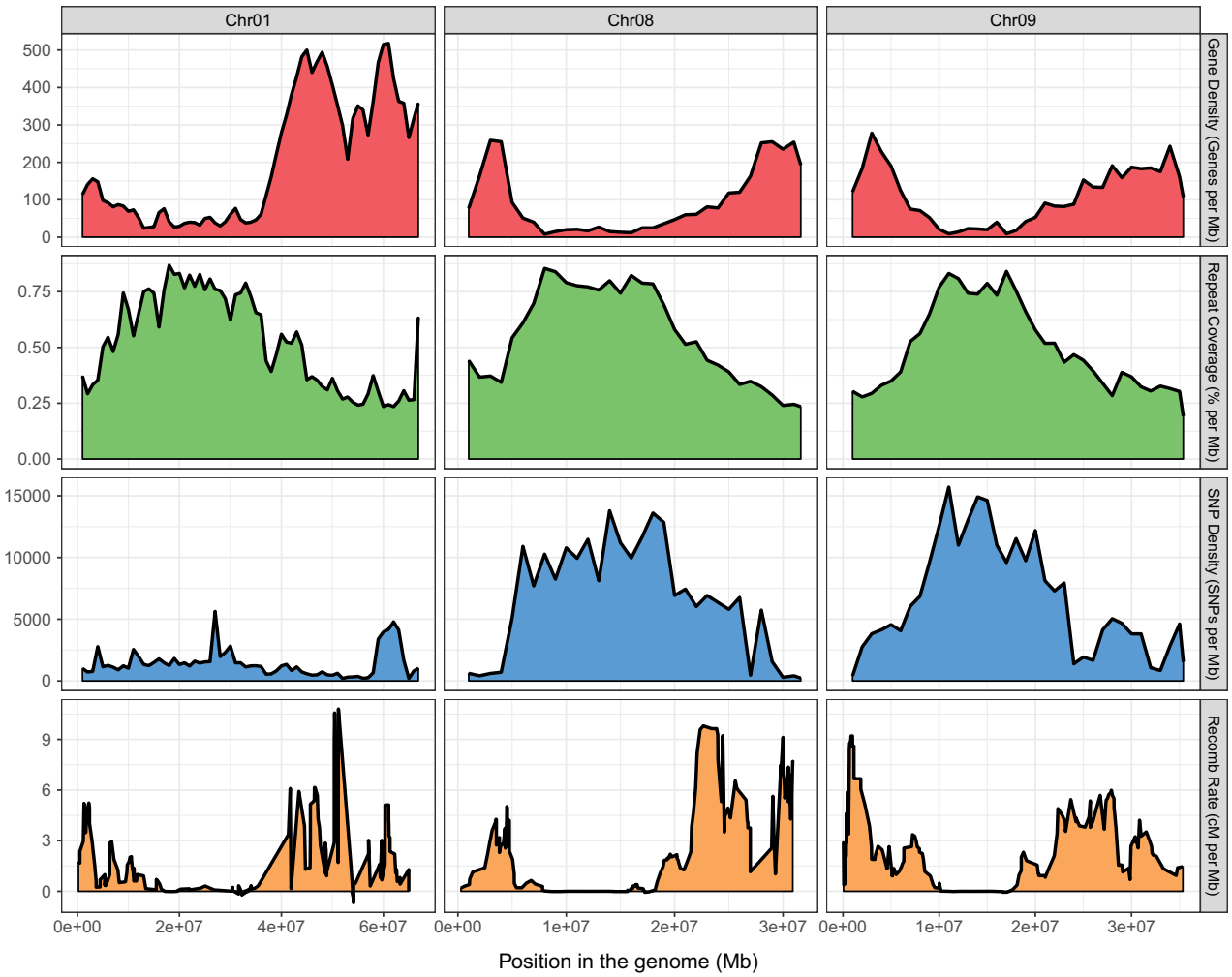

Figure 2. The *Solanum chacoense* M6 genome.
Gene density (genes per MB; top), repeat coverage (% per Mb; middle), single nucleotide polymorphism (SNP) density (SNPs per Mb; middle) and recombination rate (cM per Mb) for chromosomes 1, 8 and 9 in the M6 genome.

identification of gene clusters of secondary metabolites was performed using plantiSMASH (Kautsar *et al.*, 2017). Gene cluster analysis identified five clusters of sesquiterpene synthase genes in the *S. chacoense* M6 genome on the bottom arm of both chromosomes 6 and 9, syntenic to sesquiterpene synthase genes located in DM1-3 and tomato. Expression analyses across a developmental tissue panel in *S. chacoense* M6 revealed that glycoalkaloid genes showed different transcript abundances across development, with differences in transcript abundance of a subset of glycoalkaloid genes apparent in tubers of M6 relative to the tetraploid potato cultivar Missaukee (Figure 4b).

The *S. chacoense* M6 genome is also a useful resource for understanding self-compatibility in diploid potato. *Solanum chacoense* M6 possesses self-compatibility due to a single dominant S-locus inhibitor gene *Sli*, which prevents gametophytic self-compatibility by the stylar S gene(s) (Hosaka and Hanneman, 1998a) and is thus a valuable resource for identifying important agronomic traits required to facilitate diploid-based breeding of potato. Linkage analysis mapped the *Sli* gene to the end of chromosome 12 (Hosaka and Hanneman, 1998b), while the self-incompatibility (S) locus is located on chromosome 1 in diploid potatoes (Gebhardt *et al.*, 1991; Jacobs *et al.*, 1995; Rivard *et al.*, 1996; Hosaka and Hanneman, 1998b). Access to the *S. chacoense* M6 genome can provide additional resources for cloning and functional characterization of the *Sli* gene.

Overall, the genome assembly of the diploid wild potato species *S. chacoense* M6 provides a high-quality representation of the estimated 882-Mb genome, with a large percentage of the gene space and scaffolds anchored into 12 pseudomolecules. Single nucleotide polymorphism analysis revealed regions of residual heterozygosity, especially on chromosomes 4, 8 and 9. The genome annotation

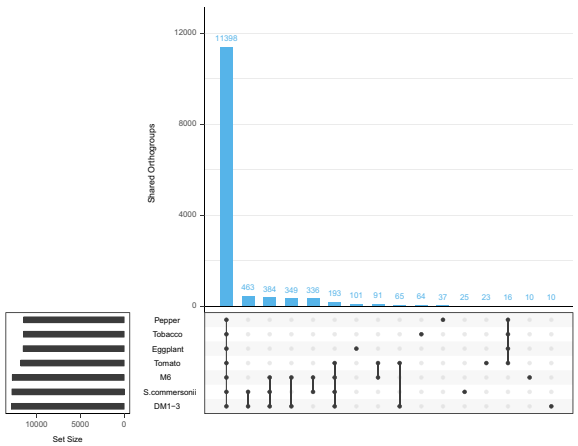

Figure 3. UpSet plot of orthologous groups among seven solanaceous species. Orthology analysis completed with Orthofinder (Emms and Kelly, 2015). Blue bars represent the number of genes present in shared orthogroups, with species occupancy per orthogroup represented by the black dot plot below each bar. Plot visualized with UpSetR (Conway *et al.*, 2017).

yielded 37 740 functionally annotated genes and orthology analysis revealed a core set of 158 367 Solanaceae genes and 1897 genes unique to potato species. Taken together, access to the *S. chacoense* M6 genome provides a resource to accelerate potato breeding through the development of inbred diploid potatoes.

EXPERIMENTAL PROCEDURES

Genome assembly and assessment

Tissue culture plantlets were grown using axillary buds and apical shoots in Magenta boxes (PhytoTechnology Laboratories, https://phytotechlab.com/) with MS media (PhytoTechnology Laboratories, product no. M516) on light racks set to a 16-h/8-h day/night photoperiod at 22°C and was DNA isolated from young leaves using the cetyltrimethyl ammonium bromide method (Saghai-Maroof *et al.*, 1984). Illumina-compatible paired-end libraries (estimated insert sizes of 166, 186 and 250 bp) and Nextera Mate Pair libraries (estimated fragment sizes of 3.5, 6.2, 6.6 and 9.6 kb) were constructed and sequenced on an Illumina HiSeq 2500. Paired-end reads were assessed for quality using FASTQC (v.0.11.5) (FASTQC, 2017) and cleaned using Cutadapt (v.1.8) (Martin, 2011). Mate pair libraries were processed with NextClip (v.1.3.1) (Leggett *et al.*, 2014) retaining trimmed reads containing the junction adapter (type A, B and C). The filtered mate pairs were then cleaned with Cutadapt (v.1.8) to ensure that the junction adapters were removed. Reads (364 519 078 paired-end and 165 596 988 mate pairs) were used with the ALLPATHS-LG assembler (v.51828) to generate an initial assembly. Gaps were filled using GapCloser (v.1.12r6) with paired-end sequences from three additional paired-end Illumina-compatible libraries not used in the initial assembly (estimated insert sizes of 258, 422 and 530 bp; Table 1).

Genome annotation

To provide transcript evidence for gene annotation, total RNA was isolated from greenhouse-grown plants for six core developmental tissues (Table S4) using an RNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com/). Paired-end libraries for RNA-Seq were constructed using the Illumina TruSeq RNA-Seq Kit (Illumina, Inc., https://www.illumina.com/). A single-stranded paired-end library for RNA-Seq was constructed from young leaves using the KAPA Stranded RNA-Seq library kit (Kapa Biosystems, https://www.kapabiosystems.com/). All RNA-Seq libraries were sequenced on an Illumina HiSeq 2000 platform, with the exception of the stranded paired-end library which was sequenced on an Illumina HiSeq 2500, assessed for quality using FASTQC (v.0.11.5) (FASTQC, 2017), and cleaned using Cutadapt (v.1.8). The cleaned RNA-Seq reads from the strand-specific leaf library were aligned to the genome assembly with TopHat2 (v.2.0.12) (Kim *et al.*, 2013) using the following parameters: -i 20 –I 20 000 –library-type fr-first-strand. Genome-guided transcriptome assembly was performed for each alignment file for each RNA-Seq library using Trinity (v.2.2.0) (Haas *et al.*, 2013) with a maximum intron size of 5 kb and a minimum contig length of 500 bp. Expression abundance for each gene model (Table S5) was determined for the RNA-Seq libraries and alignments described above using Cufflinks (v.1.3.0) (Trapnell *et al.*, 2010) with a maximum intron length of 10 kb.

To annotate the M6 genome, we first generated a repeat library with RepeatModeler (v.1.0.8) (Smit and Hubley, 2015) on scaffolds greater than 100 kb. Additionally, a repeat library of miniature inverted-repeat transposable element (MITE) sequences was created using MITEHunter (v.2011) (Han and Wessler, 2010) using the default options. The repeat libraries were searched against a curated library of plant protein-coding genes and sequences with matches were trimmed or removed with ProtExcluder (v.1.1) (Campbell *et al.*, 2014). The cleaned repeat libraries were then combined with the Viridiplantae repeats in Repbase (v.20150807) (Jurka, 1998) to create a final custom repeat library (CRL). A repeat masked assembly was generated using RepeatMasker (v.4.0.6) (Smit *et al.*, 2015) and the CRL using the –s and –nolow options. In total, 60.7% of the M6 genome assembly was repeat masked (Table S6).

Gene models were generated by first training AUGUSTUS (v.3.1) (Stanke *et al.*, 2006) with the genome-guided strand-specific leaf RNA-Seq Trinity transcript assemblies, followed by running AUGUSTUS (Stanke and Morgenstern, 2005) on the hard-masked genome to generate gene predictions. The gene models were refined using PASA2 (v.2.0.2) (Haas *et al.*, 2003; PASA, 2017) using the additional genome-guided transcriptome assemblies (Table S4) as transcript evidence. The resulting working set of gene models was categorized into high- and low-confidence gene model sets based on expression evidence from the RNA-Seq libraries, PFAM domain evidence and the absence of a full coding sequence (CDS) in the gene model. Functional annotation was assigned using custom pipeline using searches against the Arabidopsis proteome (TAIR10) (Berardini *et al.*, 2015; TAIR10, 2017), Swiss-Prot (Bairoch and Apweiler, 2000) and PFAM (v.29) (Finn *et al.*, 2014).

Construction of pseudomolecules

Two independent mapping populations containing *S. chacoense* M6 as a parent were used to anchor the scaffolds to the 12 chromosomes and construct pseudomolecules. The first population used was an F_2 population derived from the cross of DM1-3 by *S. chacoense* M6 (referred to as DM1-3 × M6) (Endelman and Jansky, 2016). For this population, 178 F_2 progeny were genotyped. The second population used was an F_2 population of 99 self-compatible progeny derived from a cross between the *S. chacoense* clone *chc*80-1 (USDA 8380-1 from accession 458310) (Sanford *et al.*, 1996) with *S. chacoense* M6 (referred to as chc80-

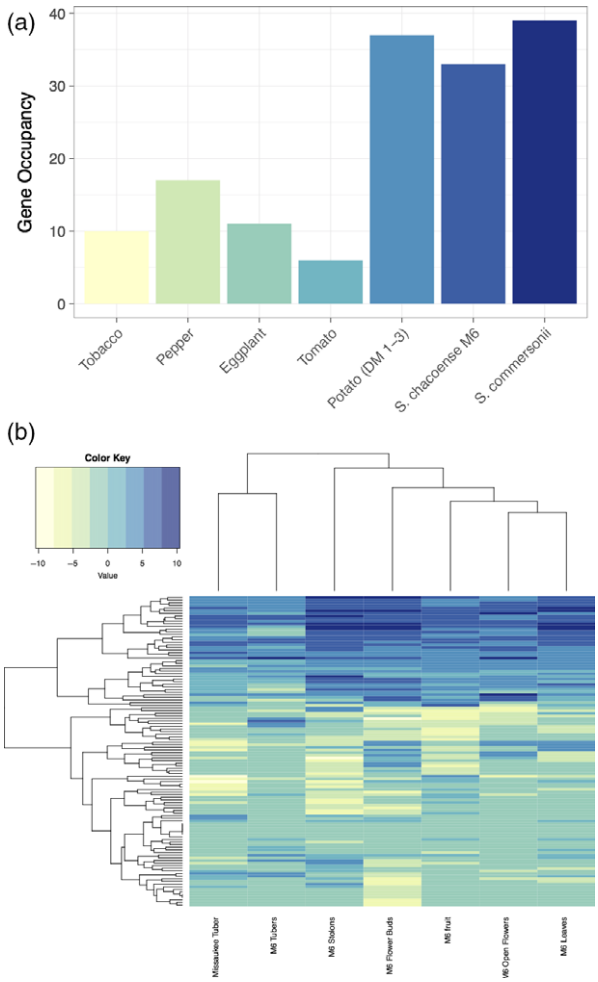

Figure 4. Glycoalkaloid expression in seven species in the Solanaceae.
(a) Gene occupancy analysis of orthologous groups containing genes involved in sesquiterpene synthase in seven Solanaceae species.
(b) Expression abundance [fragments per kilobase of transcript per million mapped reads (FPKM), $\log_2$-transformed] of glycoalkaloid metabolism genes in six tissues in diploid *S. chacoense* M6 and tetraploid tuber tissue (Missaukee).

1 × M6). The genetic map of the DM1-3 × M6 population was generated using SNPs from the potato Infinium 8303 array (Endelman and Jansky, 2016). For the chc80-1 × M6 population, the parents and progeny were genotyped using the new Illumina® Infinium 22K V3 Potato Array. This array contains the SNP from the Infinium 8303 Potato Array with additional markers from the Infinium high-confidence SNPs (69K) (Hamilton *et al.*, 2011), selected for genome coverage, candidate genes and regions with resistance genes and SNPs that best performed from the SolSTW 20K array (Vos *et al.*, 2015). A quality control and filtering process was carried out to identify segregating markers. The genetic map was constructed with 791 segregating markers using JoinMap (v.4.1) (van Ooijen, 2006) with a minimum LOD score of 6 used to define linkage groups. The Monte Carlo maximum likelihood mapping algorithm was used to calculate the linkage maps.

The first round of pseudomolecule construction was done using 2202 SNPs from the DM1-3 × M6 population that matched to scaffold positions in the *S. chacoense* M6 genome. Scaffolds were then ordered using the known position of markers in the DM1-3 genome (v.4.03) (Sharma *et al.*, 2013) to create an initial set of ordered scaffolds on the 12 chromosomes. Additional scaffolds were anchored and validated using genetic markers from the chc80-1 × M6 population. Markers matched a total of 922 positions to scaffolds in the *S. chacoense* M6 genome and scaffolds were again ordered using the known position of markers on the DM1-3 reference genome. Vmatch (v.2.2.5) (Abouelhoda *et al.*, 2004) was used to identify the corresponding location of each SNP sequence in the *S. chacoense* M6 genome. For the DM1-3 × M6 population, SNP positions were replaced with 'N' and matched against the masked *S. chacoense* M6 assembly, with a minimum match length of 90, an edit distance of 3, and computing both direct and reverse complement matches. For the chc80-1 × M6 SNP positions were replaced with 'N' and matched against the masked *S. chacoense* M6 assembly, with a minimum match length of 60, an edit distance of 4 and computing both direct and reverse complement matches. Comparison of the *S. chacoense* M6 pseudomolecules was completed by performing whole-genome sequence alignment of the 12 masked *S. chacoense* M6 pseudomolecules to the 12 DM1-3 chromosomes (v.4.04) (Hardigan *et al.*, 2016) using MUMmer (v.3.23) (Kurtz *et al.*, 2004). PROmer was run using -mincluster 500 –maxgap 20, followed by delta filtering using –l 1000 –i 90 -1 (one-to-one alignment option).

Evaluation of the genome landscape

Gene density per Mb was calculated for annotated gene sequences in the *S. chacoense* M6 genome in non-overlapping windows using BEDTools makewindows and BEDTools coverage (v.2.25.0) (Quinlan and Hall, 2010). Gene model coordinates from the original *S. chacoense* M6 assembly were lifted over to the new pseudomolecules using the pseudomolecule tiling path. Repeat content was derived using the GFF file generated from RepeatMasker (v.4.0.6) (Smit *et al.*, 2015) and percent repeat content per Mb was calculated in non-overlapping windows as described above. Gene content and repeat coverage were visualized using R (v.3.2.3) (R Core Team, 2017).

Genomic reads from a single paired-end genomic DNA library (SRR5264017) were aligned to *S. chacoense* M6 pseudomolecules using BWA MEM (v.0.7.11r1034) and the alignment filtered with SAMTools (v.0.1.19) (Li *et al.*, 2009) retaining alignments that are properly paired and have a mapping quality (MAPQ) > 30. The BAM file was then sorted and duplicates marked with Picard (v.2.1.1) (Picard, 2017). Reads were realigned around insertions/deletions (InDels) using GATK IndelRealigner (v.3.7.0) (McKenna *et al.*, 2010). Variant calling was performed using SAMTools mpileup (v.0.1.19) with the –E option and VCFtools (v.0.1.12b) (Danecek *et al.*, 2011). The variants were hard filtered using vcfannotate (VCFtools, v0.1.12b), requiring a maximum read depth of 120, a minimum read depth of 80, a minimum variant quality of 20 and a minimum RMS mapping quality of 10. The SNP density was calculated as the number of SNPs per Mb in non-overlapping windows using BEDTools makewindows and BEDTools coverage (v.2.25.0) and visualized with R (v.3.2.3). Recombination rates (cM per Mb) were calculated using the genetic and physical positions on the *S. chacoense* M6 pseudomolecules of the SNPs mapped in the DM1-3 × M6 population. A 0.1 cubic spline interpolation curved was fitted on Marey maps (Chakravarti, 1991) generated for SNPs with genetic and physical concordant map positions. Recombination rates were calculated as the derivative of the polynomial curve generated from predicted cM positions and the corresponding Mb positions (Yu *et al.*, 2001). Calculations were made using JMP® 10 SAS Institute Inc. (https://www.sas.com/).

Availability of supporting information

Raw reads for the genomic DNA and RNA-Seq libraries have been deposited in the National Center for Biotechnology Information Sequence Read Archive (SRA) under BioProject PRJNA362370. The assembled genome and associated annotation are available via the Dryad Digital Repository (https://datadryad.org//resource/doi:10.5061/dryad.kc835) and via Spud DB (http://potato.plantbiology.msu.edu/).

ACKNOWLEDGEMENTS

This work was supported in part by funds from the US Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative Plant Breeding, Genetics, and Genome grant 2009-85606-05673 to CRB and DSD, and US Department of Agriculture, National Institute of Food and Agriculture, Biotechnology Risk Assessment Grant Program award 2013-33522-21090 to CRB and DSD. We would like to acknowledge Wenli Zhang for help with RNA isolations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Whole genome sequence alignment of the 12 masked *Solanum chacoense* M6 pseudomolecules to the 12 DM1-3 chromosomes (v.4.04).

Figure S2. Single nucleotide polymorphism density and recombination rate for all 12 chromosomes in *Solanum chacoense* M6.

Table S1. Alignment metrics used to validate the *Solanum chacoense* M6 genome assembly.

Table S2. Species and sources of proteomes for orthologous clustering.

Table S3. Orthologous clusters between seven Solanaceae species.

Table S4. RNA sequencing libraries used in this study.

Table S5. Expression abundance for all RNA sequencing libraries.

Table S6. Repeat content of the *Solanum chacoense* M6 genome.

REFERENCES

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

Akeley, R., Mills, W., Cunningham, C. and Watts, J. (1968) Lenape: a new potato variety high in solids and chipping quality. *Am. J. Potato Res.* **45**, 142–145.

Anonymous (1970) Name of potato variety Lenape withdrawn. *Am. J. Potato Res.* **47**, 103.

Aversano, R., Contaldi, F., Ercolano, M.R. *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.

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*, **53**, 474–485.

Birhman, R.K. and Hosaka, K. (2000) Producction of inbred progenies of diploid potatoes using an S-locus inhibitor (*Sli)* gene, and their characterization. *Genome*, **43**, 495–502.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A. and Martin, G.B. (2012) A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Mol. Plant Microbe Interact.* **25**, 1523–1530.

Campbell, M.S., Law, M., Holt, C. *et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524.

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

Cho, J., McCouch, S., Kuiper, M., Kang, M.-R., Pot, J., Groenen, J. and Eun, M. (1998) Integrated map of AFLP, SSLP and RFLP markers using a recombinant inbred population of rice (*Orysa sativa* L.). *Theor. Appl. Genet.* **97**, 370–380.

Cipar, M.S., Peloquin, S.J. and Hougas, R.W. (1964) Variability in the expression of self-incompatibility in tuber-bearing diploid *Solanum* species. *Amer. Potato J.* **41**, 155–162.

Conway, J.R., Lex, A. and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *bioRxiv*, **33**, 2938–2940. https://doi.org/10.1101/120600

Danecek, P., Auton, A., Abecasis, G. *et al.*; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

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

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

FAOSTAT (2016) Food and agriculture organization of the United Nations statistics division. http://faostat3.fao.org/. Accessed 1 March 2018.

FASTQC (2017) http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 26 October 2017.

Finn, R.D., Bateman, A., Clements, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230.

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

Gebhardt, C., Ritter, E., Barone, A. *et al.* (1991) RFPL maps of potato and their alignment with the homoeologous tomato genome. *Theor. Appl. Genet.* **83**, 49–57.

Gnerre, S., Maccallum, I., Przybylski, D. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Gore, M., Chi, J.-M., Elshire, R. *et al.* (2009) A first-generation haplotpye mape of maize. *Science*, **326**, 1115–1117.

Haas, B.J., Delcher, A.L., Mount, S.M. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666.

Haas, B.J., Papanicolaou, A., Yassour, M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512.

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

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., Bamberg, J., Buell, C.R. and Douches, D.S. (2014) Taxonomy and genetic differentiation among wild and cultivated germplasm of *Solanum* sect. *Petota*. *The Plant Genome*, **8**, 1–16.

Hardigan, M.A., Crisovan, E., Hamilton, J.P. *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. *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. USA*, **114**, E9999–E10008.

Hawkes, J.G. (1958) Significance of wild species and primitive forms for potato breeding. *Euphytica*, **7**, 257–270.

Hirakawa, H., Shirasawa, K., Miyatake, K. *et al.* (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative

solanum species indigenous to the old world. *DNA Res.* **21**, 649–660.

Hirsch, C.N., Hirsch, C.D., Felcher, K. *et al.* (2013) Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. *Gene, Genomes and Genetics.* **3**(6), 1003–1013.

Hosaka, K. and Hanneman, R.E. (1998a) Genetics of self-compatibility in a self-incompatible wild diploid potato species *Solanum chacoense*. 1. Detection of an S locus inhibitor (Sli) gene. *Euphytica*, **99**, 191–197.

Hosaka, K. and Hanneman, R.E. (1998b) Genetics of self-compatibility in a self-incompatible wild diploid potato species *Solanum chacoense*. 2. Localization of an S locus inhibitor (Sli) gene on the potato genome using DNA markers. *Euphytica*, **103**, 265–271.

Jacobs, J.M.E., Van Eck, J.H., Arens, P., Verker-Bakker, B., Te Lintel Hekkert, B., Bastiaanssen, H.J.M., El-Kharbotly, A., Pereira, A., Jacobsen, E., Stiekema, W.J. (1995) A genetic map of potato (*Solanum tubersoum*) integrating molecular markers, including transposons, and classical markers. *Theor. Appl. Genet.* **91**, 289–300.

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

Jansky, S.H., Charkowski, A.O., Douches, D.S. *et al.* (2016) Reinventing potato as a diploid inbred line–based crop. *Crop Sci.* **56**, 1412.

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

Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A. and Medema, M.H. (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63.

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

Kim, S., Park, M., Yeom, S.I. *et al.* (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270–278.

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

Leggett, R.M., Clavijo, B.J., Clissold, L., Clark, M.D. and Caccamo, M. (2014) NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, **30**, 566–568.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997v2.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

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.

Luo, R., Liu, B., Xie, Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 1–10.

McCue, K. (2009) Fruit, vegetable and cereal science and biotechnology. *Am. Potato J.* **3**, 65–71.

McKenna, A., Hanna, M., Banks, E. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

McMullen, M., Kresovich, S., Villeda, H. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.

van Ooijen, J. (2006) *JoinMap 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations* (Kyazma, B., ed.) Wageningen, Netherlands: Kyazma B.V.

PASA (2017). PASA2. http://pasapipeline.github.io/. Accessed 29 August 2017.

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.

Phumichai, C., Mori, M., Kobayashi, A., Kamijima, O. and Hosaka, K. (2005) Toward the development of highly homozygous diploid potato lines using the self-compatibility controlling *Sli* gene. *Genome*, **48**, 977–984.

Picard (2017) https://github.com/broadinstitute/picard/releases/tag/2.12.1. Accessed 31 May 2017.

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

R Core Team (2017) https://www.R-project.org. Accessed 20 October 2017.

Rivard, S.R., Cappadocia, R. and Landry, B.S. (1996) A comparison of RFLP maps based on anther culture derived, selfed and hybrid progenies of *Solanum chacoense*. *Genome*, **39**, 611–621.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley - Mendelian inheritance, chromosomal location, and population-dynamics. *Proc. Natl Acad. Sci. USA*, **81**, 8014–8018.

Sanford, L., Kobayashi, R., Deahl, K. and Sinden, S. (1996) Segregation of leptines and other glycoalkaloids in *Solanum tuberosum* (4X) x *S. chacoense* (4X) crosses. *Am. Potato J.* **73**, 21–33.

Sharma, S.K., Bolser, D., de Boer, J. *et al.* (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3*, **3**, 2031–2047.

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

Smit, A.F.A. and Hubley, R. (2015) RepeatModeler Open-1.0. Repeat Modeler. http://www.repeatmasker.org/. Accessed 7 June 2017.

Smit, A.F.A., Hubley, R. and Green, P. (2015) RepeatMasker Open-4.0. RepeatMasker. http://www.repeatmasker.org/. Accessed 26 March 2017.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467.

Stanke, M., Schoffmann, O., Morgenstern, B. and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

TAIR10 (2017). The Arabidopsis Information Resource. Arabidopsis.org. Accessed 29 August 2017.

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

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

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

Vos, P.G., Uitdewilligen, J.G., Voorrips, R.E., Visser, R.G. and van Eck, H.J. (2015) Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* **128**, 2387–2401.

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

Yu, A., Zhao, C., Fan, Y. *et al.* (2001) Comparison of human genetic and sequence-based physical maps. *Nature*, **409**, 951–953.